Moroccan Journal of Algebra and Geometry with Applications

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 1-231

MOROCCAN JOURNAL OF ALGEBRA AND GEOMETRY WITH APPLICATIONS

MJAGA

ISSN: 2820-7114

Supported by:

Sidi Mohamed Ben Abdellah University, Fez, Morocco

جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⵎⴰⴷ ⴱⵏ ⵄⴰⴱⴷⵍⵍⴰⵀ ⵏ ⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

Moroccan Journal of Algebra and Geometry with Applications
Volume 3 (1) (2024)

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

ISSN: 2820-7114

Moroccan Journal of Algebra
and Geometry with Applications
Volume 3 (1) (2024)

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco ISSN: 2820-7114

The Moroccan Journal of Algebra and Geometry with Applications publishes two issues per year composed of high-quality original research papers with significant contributions in the fields of algebra, geometry and related fields. It publishes also invited survey articles on hot topics, from distinguished mathematicians from across the world. All published research articles are subject to rigorous peer review, based on initial screening by editors, refereeing by independent expert referees, and consequent revision by article authors when required. The published article constitutes the final and definitive version of the work. All manuscripts submitted to the journal must be original contributions, and must not be under consideration for publication with another journal, nor have been previously published, in part or whole.

Author Guidelines / Submissions

Manuscript Submission :

Submission of a manuscript implies that

- the work described has not been published before;
- it is not under consideration for publication anywhere else;
- its publication has been approved by all co-authors, if any, as well as by the responsible authorities – tacitly or explicitly – at the institute where the work has been carried out.

Articles must be written in English or, exceptionally, in French and prepared in Latex at MJAGA format. The format can be found here (https://ced.fst-usmba.ac.ma/p/mjaga/author-guidelines-submissions). Articles may be submitted directly via email as a PDF-file to any of the Journal Editors.

The manuscript should contain :

- A concise and informative title.
- The name(s) of the author(s), affiliation(s) of the author(s), a clear indication and an active e-mail address of the corresponding author, and the 16-digit ORCID of the author(s), if available.
- An abstract of 150 to 250 words.
- 4 to 6 keywords which can be used for indexing purposes.
- Appropriate numbers of MSC codes.

Peer review :

The submission will be initially assessed by the editor for suitability for the journal. Papers deemed suitable are then typically sent to a minimum of one independent expert reviewer to assess the scientific quality of the paper. The Editor is responsible for the final decision regarding acceptance or rejection of articles. The Editor’s decision is final.

After acceptance :

After the paper is accepted for publication, the corresponding author will be required to submit the LaTeX file. Final formatting matters will follow, and publication will then be immediate.

Publication Charges :

There are no submission fees, publication fees or page charges for this journal.

Moroccan Journal of Algebra
and Geometry with Applications
Volume 3, Issue 1, June 2024

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

ISSN: 2820-7114

Contents :

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 1-22

ISSN: 2820-7114

Title :

An accessible calculation of the stalks of the structure sheaf of the affine scheme of an integral domain

Author(s):

David E. Dobbs

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 1–22

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco ISSN: 2820-7114

An accessible calculation of the stalks of the structure sheaf of the affine scheme of an integral domain

David E. Dobbs
Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1320
e-mail: *ddobbs1@utk.edu*

Communicated by Najib Mahdou
(Received 26 June 2023, Revised 16 September 2023, Accepted 19 Septembrer 2023)

Abstract. Assuming minimal background in algebra and topology, we give a proof that for a domain A, the stalk of the structure sheaf of the affine scheme Spec(A) at a point P is A_P. While being more accessible than the standard proof, the proof that is given here leaves few or no ambiguities or questions concerning the foundations of mathematics. Such ambiguities arise inevitably in the standard proof which considers, more generally, A to be an arbitrary commutative ring with 1. An appendix surveys some of the history involving such ambiguities in the mathematical and philosophical literature of the past 100 years.

Key Words: Integral domain, Zariski topology, localization, stalk, sheaf, Hilbert symbol, direct limit, commutative ring.
2010 MSC: Primary 13G05, 13A15, 13B30, 14A05; Secondary 18A30, 03A05, 14A15.

1 Introduction

All rings considered here are assumed to be associative and unital; except in Appendix II and in comments about Appendix II in this Introduction, all rings are also considered to be commutative. All inclusions of rings, ring extensions, subrings, algebras and ring/algebra homomorphisms will be assumed unital. Proper inclusions will be denoted by $\subset$. In connection with any commutative ring A, we will use the following standard notation: U(A) denotes the set of units of A; Spec(A) denotes the set of all prime ideals of A; and if $c \in A$, then A_c denotes the localization of A at the multiplicatively closed set generated by c (that is, at $\{c^n \mid n \geq 0\}$, where $c^0 := 1$). It will be convenient to refer to a (commutative) integral domain as a *domain*.

Let A be a ring and let $X = \text{Spec}(A)$ endowed with the Zariski topology. Recall that a basic open set in that topology is of the form X_a (more often nowadays denoted by $D(a)$), which for any element $a \in A$, is defined by

$$X_a := \{P \in X \mid a \notin P\}.$$

Now, let P be a point of the topological space X (that is, let P be a prime ideal of A). For more than 60 years, the fundamental fact that has allowed objects isomorphic to X (along with certain morphisms in some category) to constitute the affine foundations of modern algebraic geometry is that X can be given the structure of a local ringed space whose structure sheaf has its stalk at the point P given by the direct limit

$$\varinjlim_{P \in X_a} A_a = \varinjlim_{a \in A \setminus P} A_a$$

which is canonically isomorphic to A_P (as A-algebras).

The isomorphism that was just mentioned presents challenges in virtually every classroom where it is taught. (The same can be said of the implicit assumption in the preceding paragraph that student readers are familiar with terms such as “local ringed space", “structure sheaf", “stalk" and “direct limit".) The challenges to the students can be overwhelming. In Appendix I, I list 31 specific questions that can arise (and, in my experience, have often arisen) when an instructor presents a proof of the above isomorphism *verbatim* as it had been given in a well-respected, time-honored textbook. These questions are part of a blizzard of queries that many students (and their teachers) encounter when trying to understand the standard proof of the above isomorphism for the general context that was given above. The reality of the situation is that, except for the unusual class populated by students whose undergraduate studies included much of what most universities consider graduate-level material, the typical student in a beginning graduate-level course on modern algebraic geometry is simply not ready for a presentation emulating the austere sophistication of Grothendieck and Dieudonné (as in [15]). Put simply, in my experience, many students in such a course simply do not have the background to appreciate (that is, to understand) the above isomorphism in the generality that I have stated it. For instance, those students may not yet have heard of the notions of a direct limit or a sheaf (or the stalk of a sheaf or the “germ" of a function at a point). Most instructors should probably not simply assume that their students have already taken some relevant courses on subjects such as algebraic topology or differential geometry or high-dimensional real or complex analysis. Bearing in mind that in any lecture or conversation, a teacher should expect their audience to be able to carry away at most two or three of the most salient facts from that interaction, the following question naturally arises in the mind of someone planning to teach the above isomorphism. (I am now addressing some of the challenges that instructors must decide how to face.) How should a teacher (dare I say/insert, “best") first acquaint students with the just-mentioned isomorphism **if** those students have (essentially only) the following mathematical background: apart from fields (and possibly polynomial rings), the only rings that they have studied are domains; they are comfortable with fractions in the context of a fixed quotient field of a domain; they are familiar with prime ideals in the context of $\mathbb{Z}$ (perhaps also in the context of polynomial rings in one indeterminate over a field, perhaps more generally in the context of Euclidean domains, perhaps more generally in the context of principal ideal domains) and they *have* seen the definition of a prime ideal for some class of domains broader than the singleton set $\{\mathbb{Z}\}$? In short, while students in such a course have had some exposure to point-set topology (also known as general topology), it is often the case they have not studied algebraic topology or graduate-level analysis (so, to repeat, they typically have no knowledge of topics such as sheaves, direct limits, inverse limits, germs of functions, etc.).

Section 2 contains my suggested answer to the above question of how an instructor should/could/may best plan their first presentation of the isomorphism $\varinjlim_{P \in X_a} A_a \cong A_P$. That answer has worked well in classes populated with a majority of students having the kind of background described in the preceding paragraph. The detailed approach in Section 2 is occasionally presented in an informal, conversational style, somewhat as one may expect from time to time during a lecture, and readers should feel free to alter that specific content in accordance with their teaching style (and the composition and the perceived needs of their audience). As mentioned above, Appendix I mentions some of the ambiguities that can distract students who are trying to understand the standard proof for the general context. In my opinion, the proof in Section 2 avoids essentially all of those ambiguities. Of course, those ambiguities must be addressed at some time, but let us remember that “sufficient unto the moment is the complexity thereof". That maxim which I just “recalled" (honestly, I really just invented it) *is* part of the time-honored “cyclic method" approach to learning which we have all experienced and which most good teachers instinctively use in teaching most classes. Among teachers of calculus and analysis, there is general agreement that one should first learn about limits, continuity and ε-δ arguments for real-valued functions of one real variable, *then* cycle back to a deeper study (with teachers expecting deeper understanding from students) of these topics in subsequent courses (for instance, on advanced calculus) while studying real-valued functions of “several" (finitely many real) variables, and *only then* cycle back to yet deeper studies of these topics in a variety of courses (on subjects such as complex variables, metric spaces, differentiable manifolds, etc.).

Similarly, among teachers of topology, there is general agreement that students should have some of the just-mentioned experience before being placed into a course on point-set topology (or some deeper topic). There was a time, essentially when Birkhoff and MacLane wrote their 1941 text introducing the axiomatic "modern algebra" movement from Europe to an English-reading audience in North America, that algebraists were similarly devoted to the cyclic method of teaching. Indeed, even in the 1953 revised edition of their textbook, Birkhoff and Mac Lane scarcely speak of "rings", while emphasizing instead the study of $\mathbb{Z}$ and integral domains. When and why, I must earnestly ask, did teachers of algebra decide to emphasize almost-maximal generality in beginning courses? You may protest and say that some textbooks on abstract algebra nowadays still adopt a "domains first" approach – and you would be correct to assert that. But, now that we seem to have agreed on the usefulness and appropriateness of such an approach, why should we not also agree that there should be a time and place to implement it at the beginning of a course on modern algebraic geometry? How, a busy and harried teacher may well ask, can I do that –*where* should I look for advice on *how* to do that? I humbly suggest that Section 2 gives what is at least a start to the answer to such honorable questions.

Two other appendices should be mentioned here. In my experience while doing research on domains, I have encountered a significant number of workers in the field whose work avoids using any categorical or homological methods or references. In several cases, I have found these workers to be very intelligent and inspiringly creative, especially in constructing elaborate examples, but often without their being aware of some useful methods to generalize such constructions or their contexts. Sometimes, workers of this kind prefer ideal-theoretic, rather than module-theoretic, methods. Sometimes, they prefer their "domains" to be rngs (that is "domains which need not have a multiplicative identity"). Because I believe that workers such as these could offer more to the mathematical mainstream by adopting module-theoretic methods and the appropriateness of assuming that domains *should* have a multiplicative identity, I have written Appendix II. As I believe that "De gustibus non est disputandum," I cannot hope to *prove* that the just-mentioned colleagues have misplaced priorities or values. I can only hope that Appendix II will give food for thought to many. If any reader feels that my comments in this paragraph have insulted you or your mathematical heritage, please accept my sincere apology. My intent is honorable, even if you may conclude that my actual efforts have been clumsy or unseemly. The path to self-improvement can be strewn with reversals, misunderstanding and suspicion. I mean well and I wish you well.

Finally, let me say a few words about Appendix III. This has to do with a theme that underlies many of the above-mentioned 31 questions that often arise when students are shown the traditional proof that $\varinjlim_{P \in X_a} A_a \cong A_P$. For more than 100 years, serious scholars of (meta)mathematics have striven to find an appropriate universe of discourse and to understand how to arrange and access the objects of that universe. Many working algebraists are familiar with some of the history involving the Axiom of Choice and the Well-Ordering Principle, but I would expect that few readers of this article know much about Hilbert's attempt in 1923 to sidestep such topics by introducing what he called the operators ϵ and τ. I would also not expect that many readers would know that there is, to this day, ongoing research extending Hilbert's work and forming a school of "epsilontic calculus?. Appendix III gives a brief account of some current work of that school of thought, along with contributions due to Hilbert, Bourbaki and Grothendieck in regard to what I have described as "the above-mentioned ambiguities".

As usual, $|\mathcal{U}|$ denotes the cardinal number of a set $\mathcal{U}$. Any unexplained material is standard, as in [4], [12], [16].

2 A proof for integral domains

Good day, students. Today, you will begin to understand what is perhaps the most important isomorphism at the heart of the "local" aspects of modern algebraic geometry. As you have often heard me say, this material, although it will be new to almost all of you, is being brought to you by the people who arranged the curricula for your earlier studies. So, in order to anticipate at least a part of what you should expect, let us begin with a special case, a very familiar context, where A is a (commutative unital integral) domain with quotient field K. As you know, $K = \{c/d \mid c \in A \text{ and } 0 \neq d \in A\}$. Recall that K is really a ring of fractions $A_{A \setminus 0}$. So, the elements of K are really equivalence classes. But, since domains do not have any interesting zero divisors, the underlying equivalence relation is especially simple, namely: if $c_1, d_1, c_2, d_2 \in A$ with both d_1 and d_2 being nonzero, then (c_1, d_1) is equivalent to (c_2, d_2) if and only if $c_1 d_2 = c_2 d_1$ in A (in which case, we have the same equivalence classes, $c_1/d_1 = c_2/d_2$). Now – and this is important if today's special case is going to be easily understood – I am going to ask you to forget about thinking of these fractions as equivalence classes. After all, you have been working with fractions (albeit, of integers) since elementary school. And I believe that you are very comfortable in working with them, without having to worry about where such fractions may "live". Inside that "home" where they live – which is the quotient field K that we fixed above – we will establish the kind of isomorphism that we want by building a special kind of union, called a directed union, of certain rings of fractions that are each subsets of that "home", K. In a later class, you will learn that when A is only a commutative ring, it is not so intuitively easy to understand where the various relevant rings of fractions live and the directed unions that we will see today will be generalized to "direct limit" processes by which these rings are somehow combined. Suffice it to say here that understanding direct limits will require you to do some additional foundational work. But, fortunately, none of that additional work will be necessary here today, where all of our rings of interest will be domains.

So, we're back to considering a domain A with quotient field K. Can you think of a way to build K as a union of some interesting rings that contain A? No? Well, let me suggest trying the rings of the form A_a. Recall that if $0 \neq a \in A$, then $A_a := \{c/a^n \in K \mid c \in A, n \geq 1\}$. Isn't it clear that $A \subseteq A_a \subseteq K$ for all such a, and also that $\cup_{0 \neq a \in A} A_a = K$? Yes? Yes! Good! What? Oh, you'd like to see an example. Sure! Let's consider $A := \mathbb{Z}$, so $K := \mathbb{Q}$, and let's take $a := 2$. Then in this example, $A_a = \mathbb{Z}_2 = \{c/2^n \in \mathbb{Q} \mid c \in \mathbb{Z}, n \geq 1\}$. And in this example, $3/4 \in A_a$ but $4/3 \notin A_a$. Is that all clear now? Good! Let's move on.

It would be nice if the "building blocks" A_a were all "comparable", in the sense that whenever a_1 and a_2 are nonzero elements of the domain A, then either $A_{a_1} \subseteq A_{a_2}$ or $A_{a_2} \subseteq A_{a_1}$. If that happens, then the set of the rings A_a is linearly ordered (some people call that sort of thing "totally ordered") and the building blocks would "line up" neatly. What a terrific way that would be to visualize K! Unfortunately, most familiar domains do not have those building blocks line up linearly. For instance, if $A = \mathbb{Z}$, then A_2 and A_3 are not comparable, since $3/2 \in A_2 \setminus A_3$ and $2/3 \in A_3 \setminus A_2$. But, for any domain A, the union of the building blocks is an example of what is called a "directed union", in the following sense: if a_1 and a_2 are any nonzero elements of A, there there exists some nonzero element $a \in A$ such that $A_{a_1} \subseteq A_a$ and $A_{a_2} \subseteq A_a$. Can anyone suggest how to find such an element a? What? Yes, taking $a := a_1 a_2$ *does* work. Thank you for that input. Do you all see why both A_{a_1} and A_{a_2} are contained in $A_{a_1 a_2}$? Some of you are shaking your heads. Well, please consider this: if $c \in A$ and $n \geq 1$, then $c/a_1^n = c a_2^n/(a_1 a_2)^n$. Right? Good – you're all nodding your heads. Isn't it great when we can use some old familiar algebra, even arithmetic, to validate a conjecture? Well, I'm glad that you're still with me.

Let's summarize what we've done so far. If A is a domain with quotient field K, then K is the directed union of the domains of the form A_a as a runs through the set $A \setminus \{0\}$. More formally, $K = \cup_{a \in A \setminus 0} A_a$. Let's spend some time explaining what it means for that index set to be "directed".

Most folks agree that a set I, equipped with a binary relation $\leq$ on I, is called a *directed set* if the following three conditions hold: $\leq$ is reflexive (you know that this means that $i \leq i$ for all $i \in I$); $\leq$ is transitive (you know that this means that if $i, j, k \in I$ satisfy $i \leq j$ and $j \leq k$, then $i \leq k$); and $\leq$ is directed (for most of you, this may be a new concept: this means that if $i, j \in I$, then there exists $k \in I$ such that $i \leq j$ and $j \leq k$). Isn't it clear that we have shown that K is a directed union of the domains A_a where $0 \neq a \in A$. What's that? Oh, you want to know how to define the relation $\leq$ in this case, right? Well, as in most cases involving sets with enriched structures (some people call these "concrete categories"), the relevant relation is either inclusion or reverse inclusion. These two kinds of relations are often directed because, if U and V are subsets of W, then $U \cap V$ is a subset of both U and V, while both U and V are subsets of $U \cup V$. Of course, the theory of an a enriched structure is often richer than set theory, since $U \cap V$ and/or $U \cup V$ may not share the same kind of enriched structure that U and V shared. In our example, if a_1 and a_a are nonzero elements of a domain A, then $A_{a_1} \cap A_{a_2}$ is a domain, but it may not be of the form A_a for some $a \in A$. Moreover, $A_{a_1} \cup A_{a_2}$ may not even be a domain. In fact, it may not be closed under addition – for homework, please construct an example showing this fact. Fortunately, our example does not need to use intersections or unions to establish the "directed" property. Do you recall that both A_{a_1} and A_{a_2} are subsets of $A_{a_1 a_2}$? Good! *That* is why we were able to view K as being a directed union of the rings A_a. What's that? Yes, I only verified the third axiom for a directed set. You see, the other two axioms are about reflexivity and transitivity, and those properties always hold because of basic set theory for any relation $\leq$ which has been induced by either inclusion or reverse inclusion. I apologize for not mentioning that earlier. Please keep it in mind for the future, because I probably won't remember to say it again!

You may be wondering if the above relation $\leq$ could have been described, perhaps using some equations, in terms of the "arithmetic" of the domain A. Yes, that can – and should – be done. We will do it below, in Proposition 2.1 (d).

Now, let's begin to generalize the above result to the context that really matters here: A is still a domain, but another piece of data is a prime ideal P of A. (Remember that can be summarized by writing $P \in \mathrm{Spec}(A)$.) You will come to see that what we did above really treated the case $P = 0$ (which *is* a prime ideal of A because A *is* a domain). The general fact that we are aiming for is the following:

$$\cup_{a \in A \setminus P} A_a = A_P$$

describes A_P as a directed union of the domains A_a as a ranges over the directed index set $A \setminus P$. You can easily modify the above reasoning to see that A_P is the just-displayed union. And that union is directed, once again because both A_{a_1} and A_{a_2} are contained in $A_{a_1 a_2}$. But this time, where P may not be 0, it may be less obvious why $a_1 a_2$ is admissible. Earlier (when $P = 0$), we just used the fact that A was assumed to be a domain to conclude that $a_1 a_2$, being the product of two nonzero elements of a domain, must be nonzero. Why, in the present situation, is $a_1 a_2$ admissible? In other words, if both a_1 and a_2 are elements of $A \setminus P$, why is $a_1 a_2$ also an element of $A \setminus P$? Thank you for that answer. It is absolutely right. The answer is: precisely because P is a prime ideal of A! And do you know what that suggests? That last fact did not use the "domain" property of A. Maybe some of this analysis could carry over more generally, to arbitrary commutative rings. Let's spend some time looking into that possibility. Don't worry – we will return to the context of domains long before any blizzard of ambiguities has been forecast by your local mathematical weatherperson.

Let's ease into the general case with a short paragraph involving some review and some topology, then get "radical" (sorry for the bad pun) in the following paragraph, and then get the result (Proposition 2.1) which holds the key to a better understanding of the index set for the above directed union(s).

Let A be a commutative (unital) ring. Consider the set $X := \mathrm{Spec}(A)$. For each $c \in A$, let $X_c := \{P \in X \mid c \notin P\}$. (So, for instance, $X_0 = \emptyset$ and $X_1 = X$.) Recall (cf. [4, Exercises 15 and 17, page 127]) that X can be given the structure of a topological space via the *Zariski topology*, by taking the sets of the form X_c (as c runs though the elements of A) as a basis for the open sets. Indeed, given the above information about X_0 and X_1, one gets this topological conclusion directly from the definition of a prime ideal of a commutative ring, as that easily gives that $X_a \cap X_b = X_{ab}$ for all $a, b \in A$.

It is well known (cf. [4, Proposition 1.14], [12, Corollary 2.10], [16, Theorem 26]) that if I is an ideal of a (commutative unital) ring A, then the *radical of I (in A)* is the following ideal of A:

$$\sqrt{I} := \{u \in A \mid \text{there exists an integer } n \geq 1 \text{ such that } u^n \in I\} =$$

$$\cap\{P \in \mathrm{Spec}(A) \mid I \subseteq P\}.$$

Part (b) of the next result shows that the above open basis of the Zariski topology can be described in terms of radicals of principal ideals. Part (d) of the next result shows that if the ambient commutative (unital) ring A is a domain, then the above open basis of the Zariski topology can also be described in terms of rings of fractions of the form A_a (with $a \in A$).

Proposition 2.1. (*a*) *Let A be a* (*commutative unital*) *ring. Let $a, b \in A$. Then $X_a \subseteq X_b$ if and only if $\sqrt{Aa} \subseteq \sqrt{Ab}$.*

(*b*) *Let A be a* (*commutative unital*) *ring. Let $a, b \in A$. Then $X_a = X_b$ if and only if $\sqrt{Ab} = \sqrt{Aa}$.*

(*c*) *Let A be a* (*commutative unital*) *domain, with quotient field K. Let $a, b \in A$ such that $a \neq 0$. Then $\sqrt{Aa} \subseteq \sqrt{Ab}$* (*that is, $X_a \subseteq X_b$*) *if and only if $A_b \subseteq A_a$* (*that is, if and only if A_b is a* (*unital*) *subring of A_a inside K*).

(*d*) *Let A be a* (*commutative unital*) *domain, with quotient field K. Let $a, b \in A$. Then $\sqrt{Aa} = \sqrt{Ab}$* (*that is, $X_a = X_b$*) *if and only if $A_a = A_b$* (*that is, if and only if A_a and A_b are* (*unital, but possibly zero*) *subrings of each other*).

Proof. (a) We have the following equivalences and implications: $X_a \subseteq X_b \Leftrightarrow X \setminus X_a \supseteq X \setminus X_b \Leftrightarrow \{P \in X \mid a \in P\} \supseteq \{P \in X \mid b \in P\} \Rightarrow \cap\{P \in X \mid a \in P\} \subseteq \cap\{P \in X \mid b \in P\} \Leftrightarrow \sqrt{Aa} \subseteq \sqrt{Ab} \Leftrightarrow a \in \sqrt{Ab} \Leftrightarrow$ there exists an integer $n \geq 1$ and an element $\alpha \in A$ such that $a^n = \alpha b$. This (more than) proves the "only if" assertion. To prove the converse, suppose that $\sqrt{Aa} \subseteq \sqrt{Ab}$. Our task is to prove that $X_a \subseteq X_b$; equivalently, that if P is a prime ideal of A such that $a \notin P$, then $b \notin P$. This, in turn, follows easily from P being a prime ideal of A, since the above reasoning gives an equation $a^n = \alpha b$ with $n \geq 1$ and $\alpha \in A$.

(b) It suffices to combine (a) with the assertion obtained by reversing the roles of a and b in (a).

(c) The first parenthetical comment follows from (a); the second parenthetical comment follows from the fact that the operations of addition and multiplication in both A_a and A_b are induced by the corresponding operations in K.

Let us first prove the "only if" assertion. Since $a \in \sqrt{Ab}$, there is an equation $a^n = \alpha b$ for some $n \geq 1$ and $\alpha \in A$. As $a \neq 0$ by hypothesis, then neither α nor b is 0 (since A is a domain). Therefore, as $A_{c^k} = A_c$ (as subsets of K) for all nonzero elements $c \in A$ and all integers $k \geq 1$, we have, in view of the assumption that $a \neq 0$, that $1/b = \alpha/(\alpha b) = \alpha/a^n$ in K, whence $1/b \in A_{a^n} = A_a$, and then it follows easily that $A_b \subseteq A_a$ (as subsets of K).

For the converse, suppose that $A_b \subseteq A_a$. Then, working in the quotient field K of A, we have $1/b = \alpha/a^n$ for some $\alpha \in A$ and some integer $n \geq 1$. Thus $a^n = \alpha b$, whence $a \in \sqrt{Ab}$, whence $\sqrt{Aa} \subseteq \sqrt{Ab}$, as desired.

(d) In view of (b), it suffices, if neither a nor b is 0, to apply (c).

It remains to consider the cases(s) where either $a = 0$ or $b = 0$ (or both). This situation requires separate treatment because of the existence of nilpotent elements. Indeed, notice that if A were only assumed to be a commutative (unital) ring, then $c \in A$ satisfies $X_c = \emptyset$ if and only if c is nilpotent; and, still assuming only that A is a commutative ring, notice that $c \in A$ satisfies $c \in \sqrt{A \cdot 0}$ if and only if c is nilpotent. As the present A is assumed to be a (commutative unital) domain, the assumption that $\sqrt{Aa} = \sqrt{Ab}$ (equivalently, $X_a = X_b$), when coupled with the assumption (of the prevailing case) that either $a = 0$ or $b = 0$, ensures that *both* a and b equal 0 in the domain A. Similarly, while working with the domain A, we have that the assumption that $A_a = A_b$, when coupled with the assumption that either $a = 0$ or $b = 0$, ensures that both A_a and A_b are zero rings, whence *both* a and b equal 0 in the domain A. Thus, under the assumption that either $a = 0$ or $b = 0$ (or both), we have:

$$\sqrt{Aa} = \sqrt{Ab} \Leftrightarrow a = 0 = b \Leftrightarrow A_a = A_b.$$

It is perhaps worth pointing out that when a and b are each equal to the same element $0 \in A$, the use of that element $0 \in A$ in the construction of both of the relevant rings of fractions, A_a and A_b, gives that $A_a = A_0 = A_b$, whence A_a and A_b *are* equal as rings, although *that* ring is a zero ring and not a (unital) subring of K. □

The hypothesis that A is a domain allows the conclusion, via Proposition 2.1 (d), that $\sqrt{Aa} = \sqrt{Ab} \subseteq A$ implies that A_a and A_b are equal rings, to be unambiguous. However, if A had been assumed only to be a commutative (unital) ring, we could hope to (at best) conclude that A_a and A_b are isomorphic rings. Consequently, if one attempts to apply a functor to an unspecified one of the pertinent rings that is isomorphic to A_a, it becomes unclear (that is, ambiguous; that is, known only up to isomorphism) as to what is meant by *the* alleged result of such an application. Yet, *that* is exactly the sort of thing that the literature does, many times over, in this general area when working with commutative (unital) rings A. I believe that during your *initial* exposure to the ring-theoretic foundations of modern algebraic geometry, there is no urgent reason for you to be bombarded with a blizzard of ambiguities. The term “blizzard" is not mere hyperbole here, as you will see if you read my critique in Appendix I of two well-respected expositions of the general case. Also, you will see, if you read Appendix III, that worries concerning the meaning and well-definedness of such applications of functions or functors to unspecified isomorphic copies of a known object have been the topic of ongoing studies for more than 100 years. To temporarily avoid (that is, to forestall) the ambiguities which arise in the general case, we will usually assume for the rest of this section that the ambient (commutative unital) ring A is a domain. Occasionally, we may pause to explain where/how that restriction to domains has simplified matters and avoided ambiguity, but typically we will leave it to you, the reader, to be alert to such instances. I believe that the following is a sound principle, both for students and for researchers: while reading each step of a proof, ask yourself if the step follows as indicated *and* also ask yourself if the conclusion of the step would have been possible under weaker assumptions.

Remark 2.2. Consider the form of the statement that $\varinjlim_{P \in X_a} A_a \cong A_P$. How could one come to understand this statement if it were expressed in its most efficient form? If the ring A is “far" from being a domain then, even if a and b are elements of A such that $X_a = X_b$, it is by no means clear that A_a and A_b are the same mathematical object, because there is no obvious universe containing both A_a and A_b within which one could compare A_a and A_b (in order to see if they are the same). As one can quickly see by tweaking the proof of Proposition 2.1, if $X_a = X_b$, then $A_a \cong A_b$. But that is palpably *not the same* as saying that $A_a = A_b$! Fortunately, we have seen in Proposition 2.1 (d) that if A is a domain, then any quotient field of A is the desirable kind of universe, as we showed that if $X_a = X_b$ for nonzero elements a and b of a domain A (with quotient field K), then we *do* have $A_a = A_b$ (as subsets of K). This suggests that a more efficient (or economical or elegant) description of $\varinjlim_{P \in X_a} A_a$ should be possible, especially if A is a domain, if one were to impose an appropriate equivalence relation of the index set. That is what we will do five paragraphs hence. This completes the remark.

For a fixed domain A (with given quotient field K) and a fixed prime ideal P of A, a reading of Proposition 2.1 (b) suggests (correctly) that it would be useful to define the following equivalence relation $\sim$ on $A \setminus P$. If $a, b \in A \setminus P$, we say that $a \sim b$ if and only if $\sqrt{Aa} = \sqrt{Ab}$; equivalently, if and only if $X_a = X_b$; equivalently (by Proposition 2.1 (c)), that $A_a = A_b$ (as A-subalgebras of K). Note also that by defining $\sim$ in this way, $c \in A \setminus P$ ensures that $c \neq 0$ (for the more general context where A is a commutative ring, $c \in A \setminus P$ would ensure that c is not nilpotent), so that the fussiness involving "such that $a \neq 0$" in the statement of Proposition 2.1 (c) will usually not be a concern as we work with (A and) P.

With A, K and P fixed as above, it may occasionally be necessary to denote the above equivalence relation $\sim$ by $\sim_{A\setminus P}$. If $a \in A \setminus P$, the $\sim$-equivalence class represented by a will be denoted by

$$[a] \text{ or } [a]_\sim \text{ or } [a]_{\sim_{A\setminus P}}$$

with the appropriate notation to be chosen in any given situation as simply as possible, solely in order to avoid ambiguity.

Let us examine more carefully our earlier description of A_P as the directed union $\cup_{a\in A\setminus P} A_a$. How, more precisely, can this directed union be understood to have been expressed in the form $\cup_{i\in I} R_i$ for some directed union of rings R_i indexed by some directed set I? Obviously, one should take $I := A \setminus P$, with the "dummy index" i being replaced by the dummy index a, and with the ring R_i, or rather R_a, then being taken to be A_a. But what is the precise order relation $\leq$ that is underlying this directed union? In other words, if a and b are elements of $A \setminus P$, what does/should it mean to say that $a \leq b$? The answer to this question comes from Proposition 2.1 (d). Indeed, if we say that for $a, b \in A \setminus P$, the definition of $a \leq b$ is that $A_a \subseteq A_b$ (in the quotient field K), then everything falls into place rigorously as desired, because *this* relation $\leq$ is, indeed, reflexive, transitive and directed (with the last of these properties holding since both A_a and A_b are subsets of A_{ab}). Notice also that if $a, b \in A \setminus P$ as above, then we have the following additional formulations of the above equivalence relation, thanks to Proposition 2.1: $a \leq b \Leftrightarrow X_b \subseteq X_a \Leftrightarrow \sqrt{Ab} \subseteq \sqrt{Aa}$.

The above understanding of A_P as the directed union $\cup_{a\in A\setminus P} A_a$ can be made "crisper" (some would say, "sharper" or "more economical" or "more elegant") by using the above equivalence relation $\sim = \sim_{A\setminus P}$. In a moment, I will explain how to do that. When that has been accomplished, I hope that you will agree that we will have a new description of A_P as a new directed union which merits the just-mentioned laudatory adjectives. But my main reason for getting to that new description has to do with some ambiguities in the literature. You see, the literature is not entirely uniform as to the definition of a directed index set. Of course, this fact affects the definition of a directed union (and it also affects, more generally, the definition of a direct limit). While the literature *does* agree that the binary relation $\leq$ on a directed set should be reflexive, transitive and directed (as in the definition that we have been working with here), a noticeable minority of the literature also requires $\leq$ to be antisymmetric (in the usual sense, namely, that if $i, j \in I$ satisfy $i \leq j$ and $j \leq i$, then $i = j$). Unfortunately, requiring the above relation $\leq$ on $A \setminus P$ to be antisymmetric would mean that whenever elements a and b of $A \setminus P$ satisfy $A_a = A_b$, then one would need to have $a = b$. *That* sad situation, for the prime ideal $P = 0$, would imply that $a^2 = a$ for each nonzero element of the domain A. And *that* would imply that $A \cong \mathbb{F}_2$, which is not all what we wanted in this attempt to say something interesting and useful about *all* domains A. So, to placate the above-mentioned minority, the promised "moment" has passed/come, and it is now time to introduce an equivalence relation $\preceq$ which will allow us to replace the index set $A \setminus P$ with the set of $\sim$-equivalence classes from $A \setminus P$. That will be done in the next paragraph.

Given a domain A and a prime ideal P of A, we can define a binary relation on the equivalence classes of the equivalence relation $\sim = \sim_{A\setminus P}$ as follows. If $[a]$ and $[b]$ are such equivalence classes, let us say that $[a] \preceq [b]$ if and only if $a \leq b$. (To avoid ambiguity, you may occasionally prefer to use the notation "$\preceq_{A\setminus P}$" instead of "$\preceq$".) Notice that the binary relation $\preceq$ has been well defined (for if $[a_1] = [a_2]$ and $[b_1] = [b_2]$ with $a_1 \sim b_1$, then we have $A_{a_1} = A_{a_2}$ and $A_{b_1} = A_{b_2}$, along with $A_{a_1} \subseteq A_{b_1}$, whence $A_{a_2} \subseteq A_{b_2}$.) Moreover, it is easy to see (please check this, but do not hand it in as homework, as it really is very easy) that $\preceq$ inherits each of the properties of reflexivity, transitivity and directedness from $\leq$, and so $\preceq$ endows the set of $\sim_{A \setminus P}$-equivalence classes with the structure of a directed set. Furthermore, this structure has the additional property that is cherished by the "noticeable minority", namely, that $\preceq$ is antisymmetric. Indeed, if $[a] := [a]_{\sim_{A\setminus P}}$ and $[b] := [b]_{\sim_{A\setminus P}}$ satisfy $[a] \preceq [b]$ and $[b] \preceq [a]$, then $a \leq b$ and $b \leq a$, whence $A_a \subseteq A_b$ and $A_b \subseteq A_a$, whence $A_a = A_b$, whence $a \sim_{A\setminus P} b$, whence $[a] = [b]$, as desired. Thus, we now have what *everyone* can agree is a description of A_P as a "directed union" (and, by taking $P := 0$, one would get a description of the quotient field of A as "a directed union"), namely,

$$A_P = \bigcup_{[a] \text{ is a } \sim_{A\setminus P}\text{-equivalence class}} A_a.$$

Ignoring the mini-controversy concerning the definition of a directed set, let us consider a claim to the effect that the last display is more "elegant" than our earlier result that (if P is a prime ideal of a domain A, then) $A_P = \cup_{a \in A\setminus P} A_a$. By sending each a to its equivalence class $[a]$, one obtains a surjection from the second index set to the first index set. (One could, instead, have noted that the Axiom of Choice gives an injection from the first index set to the second index set.) However, it would be wrong to conclude that, in general, the first index set is "smaller than" the second index set. While the cardinal number of the first index set *is* less than or equal to the cardinal number of the second index set, those cardinal numbers could be, depending on A and P, equal infinite cardinal numbers. Consider, for instance, $A := \mathbb{Z}$ and $P := 2A$. Since the set of odd integers is denumerable, this example satisfies $|A \setminus P| = \aleph_0$. In other words, the second index set in this example has cardinal number $\aleph_0$. So, in view of the above-mentioned injection, the first index set in this example is either finite or denumerable. In fact, that first index set is denumerable, since it has a fairly prominent denumerable subset. Let's pause a moment. Did you find or guess what that denumerable subset is? No? Well, thanks for trying. The subset that I noticed is the set of $\preceq_{A\setminus P}$-equivalence classes represented by odd prime numbers. The underlying fact is a gem from elementary number theory: if q and r are distinct odd prime numbers, then $\mathbb{Z}_q \neq \mathbb{Z}_r$. (You can check that this follows from the Fundamental Theorem of Arithmetic.) Since any subset of a finite set is finite, we have proved that the first index set in this example is denumerable; that is, it has cardinal number $\aleph_0$. I will leave this example by asking you to ponder the following question: should you call the first index set (in this example) "more elegant" than the second index set (in this example) *even though* these sets have the same cardinal number?

In looking at various books for the main result that we proved today, you may have come across statements such as

$$\varinjlim_{P \in X_a} \mathcal{F}(X_a) \cong A_P \text{ or } \varinjlim_{a \in A\setminus P} \mathcal{F}(X_a) \cong A_P.$$

So, you know that "$\varinjlim$" is a standard notation for direct limit, and *that* is a generalization of directed union. You may have realized that $\mathcal{F}$ is what is usually called the structure sheaf of the affine scheme $X := \text{Spec}(A)$. (Most algebraic geometers denote $\mathcal{F}$ by $\underline{O}_X$.) Given that we have focused on the result that $\cup_{a \in A\setminus P} A_a = A_P$ (when P is a prime ideal of a domain A), you have probably also surmised that $\mathcal{F}(X_a) = A_a$ (although it may not yet be clear to you whether that equation is a definition or a proven fact). It would be natural for you to wonder what sort of binary relation is being imposed on the index set $A \setminus P$ in the just-displayed statements from the literature. (Let's skim over the technical but important difference between a direct limit and a directed set, and agree that there is something like an underlying ordering on $A \setminus P$ going on in those statements from the literature.) Remember (cf. Proposition 2.1) that when P is a prime ideal of a domain A, $a \leq b$ used to mean that $A_a \subseteq A_b$, and for that context, that this condition was equivalent to $X_b \subseteq X_a$. If b is "later than" a in the relevant directed union or direct limit process (that is, if a is "less than or equal to" b in some sense), it is traditional to have $X_b \subseteq X_a$, so that "later" indexes give "smaller" neighborhoods, and the "functorial" behavior of the sheaf $\mathcal{F}$ gives a “restriction map" from $\mathcal{F}(X_a)$ to $\mathcal{F}(X_b)$, that is, from A_a to A_b. For our familiar domain-theoretic context, this is just the inclusion map $A_a \hookrightarrow A_b$. For more general ring-theoretic contexts, you will eventually develop enough intuition, based on repeated exposure and familiarity with examples, to appreciate whether/when/why restriction maps can or should be regarded as inclusion maps. It will take a while, but I am confident that you can do it. In time, the general setting will seem as natural to you as it was today when you worked with fractions inside a fixed quotient field.

I would like to point out that it is possible to extend the above reasoning in today's main result by generalizing from the multiplicatively closed set $A \setminus P$ (where P is a prime ideal of a domain A) to the case where S is an arbitrary (nonempty) multiplicatively closed subset of a domain A such that $0 \notin S$. As homework, please do the following. Show, under these conditions, that the ring of fractions A_S is a directed union of the domains A_a as the elements a run through the set S. It will be part of the assignment for you to decide what the ordering is on the relevant directed set I (remember to identify I and its ordering and to prove that I *is* directed!). You will also need to explain how, if a and b are suitably related by that ordering, one has that $A_a \subseteq A_b$ (inside the given quotient field K of A). If you wish, for extra credit, you may also try to find conditions under which you can reduce the size of the index set by setting up a suitable equivalence relation on S and letting the “new" indices be the corresponding equivalence classes (instead of the “old" indices which were elements of S).

Next, as an additional step in preparing you for some of the “wrinkles" that may arise when A is a commutative (unital) ring, but not necessarily a domain, let us notice some of what may be gained by slightly tweaking the proof of Proposition 2.1. Suppose that $a, b \in A$. Then: $X_b \subseteq X_a \Rightarrow$ there exists an integer $n \geq 1$ and an element $\alpha \in A$ such that $b^n = \alpha a$. Now suppose that, in fact, $X_b \subseteq X_a$. It will be desirable for the resulting A-algebra homomorphism $f : A_a \to A_b$ (which sends c/a^m to $c\alpha^m/b^{nm}$, for all $c \in A$ and integers $m \geq 1$) to be injective. (Notice that such f exists, thanks to the universal mapping property of rings of fractions, since $a/1$ is a unit in A_b; indeed, it has multiplicative inverse α/b^n there.) To accomplish this injectivity, some fine-tuning will be necessary in regard to the ring elements a and b that will come under consideration in the general case (when A is not necessarily a domain). Rather than delving into that fine-tuning here, let us simply notice why there was no such difficulty in case A is a domain (with quotient field K). In *that* context, with a and b each nonzero elements of a domain A such that $b^n = \alpha a$ for some integer $n \geq 1$ and some (necessarily nonzero) element $\alpha \in A$, we were actually working with a (directed) union in the above argument. Indeed, one gets that $A_a \subseteq A_b$ there since, if $c \in A$ and m is a positive integer, then $c/a^m = c\alpha^m/b^{nm}$ in K and, hence, in A_b.

There will be more fine-tuning as you continue to study the affine scheme Spec(A) (for an arbitrary commutative ring A) and its role in the “local" part of modern algebraic geometry. In addition to what we have just seen here, you will learn new machinery, involving things called direct limits, sheaves and stalks. You will also learn why $\mathcal{F}(U)$ is called the “sections of the sheaf $\mathcal{F}$ over the open set U". That should help you to understand better or more easily some material that you have seen or will see in some courses on topology or analysis (especially, in regard to the “germs" of functions at a point). You will certainly learn that if $\mathcal{F}$ is the structure sheaf of the affine scheme $X = \text{Spec}(A)$, then “the ring of global sections" $\mathcal{F}(X)$ is isomorphic to A. (Here's one final exercise: give a quick proof of this fact, using only information from today's class.) That may lead you to study other historically important representation theorems where a given ring is realized (up to isomorphism) as the ring of global sections of some sheaf on some topological space.

Congratulations! You are about to dive into the really geometric part of algebraic geometry. But that's enough for today.

3 Appendix I: Comparison with the traditional proof

For many years during the 1960s, graduate students in an algebraic geometry course learned the modern approach to the basics of that subject by reading some notes [19] (with red front and back covers) that were affectionately known as "the red Mumford". (The only alternative at that time was to read the variety of French language material that was being produced by Grothendieck and his followers.) In retrospect, Mumford did a good job at getting to the basics, while providing background and examples that were sufficient for his intended audience. I would like to begin this appendix by reviewing Mumford's five-line proof in [19, page 40] that (to use the notation of our Section 2 but now for an arbitrary commutative ring A), the stalk of the structure sheaf (I will denote that sheaf by $\mathcal{F}$) at a point P of $X := \mathrm{Spec}(A)$ (that is, at a prime ideal P of A) is A_P. Apart from some notational changes, the following is only a slight rewording of Mumford's proof:
"Since the sets of the form X_a give a basis of the Zariski topology of $\mathrm{Spec}(A)$, we have that the stalk of the sheaf $\mathcal{F}$ [of course, Mumford denotes $\mathcal{F}$ by $\underline{O}_X$] at the point P [of course, Mumford denotes P by x] is

$$\varinjlim_{P\in U} \mathcal{F}(U) = \varinjlim_{P\in X_a} \mathcal{F}(X_a) = \varinjlim_{a(P)\neq 0} A_a.$$

Since all restriction maps in our sheaf are injective, this is just $\cup_{a(P)\neq 0} A_a$, which is clearly A_P."

An objective reading of the above argument raises, in order, 31 questions. (The Introduction promised a "blizzard of queries"!) These questions are collected, together with some comments, as the following seven items:
(i) What does the first equal sign in the display really mean? If two objects (such as the partners in that asserted equality) are each only defined up to isomorphism, what sense does it make to say that those objects are equal? Would it not make more sense to say, instead, that those objects are isomorphic? Or, given that those objects are each only defined up to isomorphism, would an assertion of isomorphism here mean the same thing as your assertion of equality here? With respect, I must ask: is your assertion of equality even meaningful? Was it intentional or was it a typo?
(ii) I suppose that the answer to last part of the above set of questions is that you intended to use that equal sign and there was no typo, as I have now seen that you have continued to use equal signs two more times. Let me ask next about a quantity on the right-hand side of the first equal sign in the display. What does "$\mathcal{F}(X_a)$" really mean? I understand that for each relevant element a, the set X_a is well defined and so is $\mathcal{F}(X_a)$. But in reading about direct limits over directed sets, it was not clear to me if a directed set must be asymmetric. (I know that it must be reflexive, transitive and directed.) Looking online, I see that many people are confused, like I am, about how direct limits are defined, asking whether the index set that we are discussing is what they call an "order" or what they call a "preorder". Since P is given and a is somehow varying, should the subscript of "$\varinjlim$" on the right-hand side of the second equal sign in the display be, instead, a specific statement about the behavior of a, or perhaps, about the behavior of an equivalence class (for some equivalence relation that you have not mentioned) represented by a? If the answer to the last question is "Yes", what is that equivalence relation and what sense does it then make to speak of "$\mathcal{F}(X_a)$"? After all, if elements $a_1, a_2 \in A$ are such that $X_{a_1} = X_{a_2}$, I can probably believe that $\mathcal{F}(X_{a_1}) \cong \mathcal{F}(X_{a_2})$ – would you please give or assign a proof of this fact? – but I would need to be convinced that $\mathcal{F}(X_{a_1}) = \mathcal{F}(X_{a_2})$. Was that a hidden part of the message that you were trying to convey here? Is this question somehow linked to the questions listed under (i)? Will some or all of these concerns in (ii) dissipate if we just decide to not worry whether the index set is asymmetric? I wish that I had asked this part of my question earlier when you covered direct limits, but it only occurred to me now when I saw how you were using them. Maybe no one has ever asked you this before, so perhaps your lesson plans did not anticipate such a question from the audience. If so, please excuse me and I'll wait a bit longer for you to think about this before you answer.

(iii) I have a comment about the index set on the right-hand side of the second equal sign in the display. This is the kind of "specific statement about the behavior of a" that I mentioned in (ii). Would it have been better – or possible –to go directly from the first direct limit to the third direct limit in the display? If so, that would eliminate some (maybe all) of my concerns in (ii) above. By the way, thank you for explaining the notation "$a(P)$" earlier (on the preceding page of [19]), when you said/wrote that "The elements of $\mathcal{F}(U)$ can be viewed as functions on U." I was certainly ready to understand that part of the display!

(iv) I am going to have several questions about your phrase "This is just ...." First, what did you intend the word "This" to refer to?

(v) Next, still about your phrase "This is just", I would like to ask: did you intend the word "is" to refer to "is equal to" or "is isomorphic to"?

(vi) Here are my final questions about your phrase "This is just". What did you mean by the word "just" there? Did you mean "equal to" or "isomorphic to"? I remember that when we covered direct limits, you mentioned that every directed union is isomorphic to a direct limit over a directed set. But is the converse true? In particular, are there some instances of "$\varinjlim_{P\in X_a} \mathcal{F}(X_a)$" that should not be viewed as being isomorphic to directed unions? Would your answer to this last question depend on whether we had grouped the various elements a into equivalence classes as suggested above? If so, what is the relevant equivalence relation?

(vii) Why is your union "$\cup_{a(P)\neq 0} A_a$" well defined? I was always taught that a meaningful union of the form $\cup_{j\in J} W_j$ requires that the objects W_j be well understood sets and that there exists a universe which contains each of these sets W_j as a subset. Is that really the case here? I do not know whether you intended the elements a to range over a subset of A or over certain equivalences classes (again: if so, what is the relevant equivalence relation?), but regardless of your answer to that question, I do not see how there could exist some universe containing each of the relevant sets A_a because each of these "sets" is only defined up to isomorphism. What sense doers it make to talk about a union of things that are only defined up to isomorphism? And what sense would it make for such a union, if it *were* well defined, to be equal to something like A_P, which is itself only defined up to isomorphism? Are these kinds of questions related to what I was asking about in (i) (and occasionally later)? Has some group of mathematical leaders somehow agreed that mathematicians are working in an ideal Platonic world where all isomorphic objects are equal? If so, I did not get that memo and, unlike Leibnitz, I do not necessarily believe in a "pre-established harmony"! Who or what has somehow ordered everything to work so well together? Is some kind of well ordering being supposed and used? Excuse me, I am only trying to learn and understand, but I must add, with respect: your saying "which is clearly" did not seem at all clear to me.

As I recall, the courses in modern algebraic geometry that I took (which were taught by Professors George Rinehart and Stephen Lichtenbaum) presented the above proof (and prepared the class for it) almost exactly as in [19].

Let us next review how, a few years later, Atiyah-Macdonald approached the above result. One should note that although Mumford's notes may have been written under time pressure and were compiled into a "Preliminary version" of the first three chapters of a foreseen book, Atiyah and Macdonald had the advantage of the passage of time and their book was not explicitly a "preliminary version". Also, because of the intentionally small size of [4], much of the substance of that book is to be found in its exercises. Prior to the actual exercise stating, in effect, that $\varinjlim_{P\in U} \mathcal{F}(U) \cong A_P$, Atiyah-Macdonald did a good job of covering the Zariski topology (having the sets of the form X_a as an open basis), direct limits and rings of fractions. We next essentially reproduce the five parts of [4, Exercise 23, page 47]. As before, in order to assure uniformity of notation for comparison purposes, the following summary is the result of only a light editing of what Atiyah and Macdonald wrote in that exercise. As before, we are considering a commutative ring A, $X := \text{Spec}(A)$ has been equipped with the Zariski topology, and the structure sheaf of this affine scheme is being denoted by $\mathcal{F}$. Here,

then, are the five parts of the pertinent exercise from [4]:
(i) If $U = X_a$ for some $a \in A$, show that $\mathcal{F}(U) := A_a$ depends only on U and not on a.
(ii) Let $U^{'} = X_b$ be another basic open set in X such that $U^{'} \subseteq U$ $(= X_a)$. Show that there is an equation of the form $b^n = ua$ for some integer $n > 0$ and some $u \in A$, and use this to define a homomorphism $\rho : \mathcal{F}(U) \to \mathcal{F}(U^{'})$ (that is, $A_a \to A_b$) by $c/a^m \mapsto cu^m/b^{mn}$. Show that ρ depends only on U and $U^{'}$. This homomorphism is called the *restriction* homomorphism.
(iii) If $U = U^{'}$, then ρ is the identity map.
(iv) If $U \supseteq U^{'} \supseteq U^{''}$ are basic open sets in X, show that the composite of the restriction homomorphisms $\mathcal{F}(U) \to \mathcal{F}(U^{'})$ and $\mathcal{F}(U^{'}) \to \mathcal{F}(U^{''})$ is the restriction homomorphism $\mathcal{F}(U) \to \mathcal{F}(U^{''})$.
(v) If P is a prime ideal of A, show that $\varinjlim_{P \in U} \mathcal{F}(U) \cong A_P$.

In commentary after (v), Atiyah-Macdonald went on to state (again, I will lightly edit their notation) the following: "The assignment sending each basic open set U to the ring $\mathcal{F}(U)$, together with the restriction homomorphisms ρ satisfying the conditions in (iii) and (iv), constitutes a *presheaf of rings* on the basis of open sets $\{X_a \mid a \in A\}$" and that "(v) says that the stalk of this presheaf at P is the (quasi-)local ring A_P."

For the most part, I would prefer to let the reader decide the following three things: which, if any, of the earlier 31 questions about the presentation in [19] applies to the presentation in [4]; whether any new questions arise as a result of that presentation in [4]; and whether the presentations in [19] and/or [4] would be preferable to the presentation that I gave (for domains A) in Section 2, when the reader is considering how to present the "stalk" result to his/her/their class. Before leaving instructor/readers with such weighty matters (after all, you surely know your students, their background and their needs better than I do!), I would like to close this appendix by making three sets of points.

First, in teaching graduate courses on commutative ring theory several times at two state universities, I have often given a fuller treatment, than in either [19] or [4], of the identification (of A_P as) the stalk of the structure sheaf of Spec(A) at a prime ideal P of the commutative ring A. Occasionally, because of time pressure in such a course, I have covered only the special case for domains A, as in Section 2 above. But in *all* those courses, I took/found the time to explain what a sheaf is and why the construction at hand actually produces a sheaf. I did so, in part, because I have found the categorical concepts of an equalizer and a coequalizer to be helpful and illuminating, both for research and in teaching, on several occasions. Also, graduate students specializing in analysis, topology and differential geometry have told me that my comments along those lines had been helpful to them in their research. Speaking of teaching, I am uncomfortable in speaking of "the stalk of a presheaf at a point" (as Atiyah-Macdonald did), but perhaps this sort of worry is a personal one that the reader need not be concerned about.

Second, while exercises do not have the same purpose as lecture material, each should be clearly stated, and so, if only for the sake of completeness and fairness, I would like to raise a few questions and/or comments in regard to the presentation in [4, Exercise 23, page 47]. (Yes, that process did begin in the preceding paragraph – thank you for noticing that.) It seems natural to me to ask what Atiyah-Macdonald intended to mean by the phrase "depends on" in (i)? One could ask a similar question about Atiyah-Macdonald's (ii). In regard to their (ii), one could also ask if the name "*restriction* homomorphism" should be appended by something like "from $\mathcal{F}(U)$ to $\mathcal{F}(U^{'})$". Given the previous sentence, I must admit that I found it heartening to see the plural in "restriction homomorphisms" in Atiyah-Macdonald's commentary after (v).

Third, the following advice/principle will, I hope, meet with universal acceptance. Any graduate course on modern algebraic geometry should cover in detail the "stalk" result for the general context of an arbitrary commutative ring A. Whether or not that coverage should be preceded (either in class or as homework) with the special case where A is a domain (as given in Section 2) is a decision that should be up to the instructor (or instructors) who is (are) responsible for such a course. While someone in my position can offer advice, please let me repeat: you surely know your students, their background and their needs better than I do!

4 Appendix II: Should a domain have a multiplicative identity element?

Let R be an rng. (Some authors write “a rng" instead of “an rng". Which option is appropriate grammatically depends on how one pronounces “rng" – should it be like “urng" or like “rung"? – and there is no universal agreement about that pronunciation.) Then with respect to addition, R is an abelian group, that is, a $\mathbb{Z}$-module. Consider the mathematical object $\mathcal{R} := \mathbb{Z} \oplus R$, the external direct sum of $\mathbb{Z}$ and R as an abelian group under addition, but also equipped with a multiplication, given by

$$(n_1, r_1)(n_2, r_2) = (n_1 n_2, n_1 r_2 + n_2 r_1 + r_1 r_2) \text{for all}\, n_1, n_2 \in \mathbb{Z} \text{and}\, r_1, r_2 \in R.$$

It is straightforward to check that $\mathcal{R}$ is a (unital) ring under the above operations, with multiplicative identity element $(1,0)$. Also, there is an injective rng homomorphism $\theta : R \to \mathcal{R}$, given by $r \mapsto (0,r)$ for all $r \in R$. It is customary to use θ to view R as a subrng of the (unital) ring $\mathcal{R}$. Of course, there would be no practical reason to use the above construction if the given rng R is known to be a ring (that is, if it is known to have a multiplicative identity element, say 1_R). Indeed, in that case, the above rng homomorphism θ would not be unital, the point being that $\theta(1_R) = (0, 1_R) \in \mathcal{R}$ is distinct from the multiplicative identity element $(1,0)$ of $\mathcal{R}$.

At one time, many writers of textbooks on “ring theory" appreciated that one should introduce students early to constructions like the one given in the preceding paragraph. It seemed to take until the late 1960s, or perhaps even the mid-1970s, until a sizable majority of the community coalesced around the idea that a “ring" should have a multiplicative identity element, whereas an “rng" possesses all the properties of a ring except possibly that of having a multiplicative identity element. So, when one is reading textbooks on ring theory that were written long ago, one must take care to understand which definition of “ring" the author is using. Yet, there is often much to be gained from reading old textbooks. One such book that comes to mind is [18]. After I finished the master’s degree and just before I moved to the United States to study for the doctorate, a former professor suggested that I should read [18] to learn something about rings. (He was aware that although I had extensive knowledge of group theory and matrix theory, I had never heard the word “ring" uttered in a classroom.) In reading [18], I developed a quick respect for ring theory. Even the second section of the first chapter of [18] had a couple of challenging homework problems. That same section (to be precise, page 8 of that book) contained McCoy’s version of the construction in the preceding paragraph. (To be accurate, I should point out the following ultimately insignificant difference: where the above construction used the external direct sum $\mathbb{Z} \oplus R$, McCoy had used the external direct sum $R \oplus \mathbb{Z}$, with the necessary concomitant changes in the definition of multiplication.) McCoy’s version of “embedding a ring [I would say “an rng"] in a ring with unity" [I would omit “with unity"] was actually better tailored to the ring R at hand. Indeed, it used essentially what we did in the above paragraph if R has characteristic 0, but replaced $\mathbb{Z}$ with the ring of integers modulo n if the characteristic of R is some positive integer n. (By definition, the characteristic of an rng is the smallest positive integer n such that the sum of n copies of each element of R is 0, if such an n exists; and the characteristic of an rng R is taken to be 0 if no such n exists. Notice that this definition of the characteristic of an rng is, in the case of positive finite characteristic, really talking about the exponent of the additive group of R. Notice also that the characteristic of a/the zero rng is 1, a situation that has led some workers to argue that the notations $\mathbb{Z}/1\mathbb{Z}$ or $\mathbb{Z}_1$ should be used for “the" zero ring, since one could then say that the characteristic of $\mathbb{Z}/n\mathbb{Z}$ is n for all positive integers n. It is worth recording that no one has been foolish enough to suggest extending this practice by letting the notations $\mathbb{Z}/0\mathbb{Z}$ or $\mathbb{Z}_0$ stand for the ring of integers, although $\mathbb{Z}$ *does* have characteristic 0.) It is interesting to note that the definition of “characteristic" had not been finalized as recently as 1953, as that is when the revised edition of the classic textbook, “A survey of modern algebra," by Birkhoff and Mac Lane defined the characteristic of $\mathbb{Z}$ to be ∞ (while acknowledging in a footnote that “most" authors had decided to say that the usage “characteristic ∞" should be replaced by “characteristic 0").

When I was a pre-doctoral student trying to learn about rings and linear transformations on my own, an appealing aspect of the approach to modern algebra in the above-mentioned textbook by Birkhoff and Mac Lane is that (as opposed to the “groups first" approach in the popular textbook, “Topics in algebra," by Herstein) they defined “characteristic" only for domains. That “domains first" approach suited me well, as I was quite comfortable with fields and I was becoming comfortable with some domains (notably, $\mathbb{Z}$ and polynomials rings in one indeterminate over a field). I wonder if the researcher in commutative algebra that I became (spending many years studying many classes of domains) would agree with his younger self that the embedding of an rng in a ring is worth teaching to today’s students.

Following [12], a “domain" is defined to be a nonzero commutative rng with no nonzero zero-divisors. With that usage, a “domain" need not be unital, that is, it need not be a ring. For me (and, I suspect, for most readers of this article), a “domain" is unital, that is, it is a ring. That has also been my belief ever since I learned about domains. As my doctoral research had been on what is nowadays called arithmetic algebraic geometry, it was natural for me to see a “domain" as an example of a (necessarily unital) commutative ring. I had not heard of multiplicative ideal theory (or [12] or Robert Gilmer) until almost the end of my postdoctoral year at UCLA (and it was a noncommutative ring theorist, Julius Zelmanowitz, who informed me of the area and who suggested that I familiarize myself with [12]). The fact that a significant number of commutative ring theorists and other mathematicians do not believe that a “domain" needs to be unital was brought to my attention in an anecdote that I relate in the next paragraph. (That anecdote does not reflect me in the highest moral light, but I do find it amusing and instructive – I hope that you will, too.)

One weekday around noon several decades ago, I left my office and went to the mathematics department’s mail room to see if the daily mail had been delivered. Discovering that the current issue of the MAA’s Monthly magazine had just arrived, I took my copy of that magazine back to my office and quickly turned to the problem section. I found a problem that seemed to be in commutative algebra (that was a rare occurrence in that section of the Monthly during that period of history) and I set to work on it. The problem was about domains, and within moments, I had solved the problem by using a standard tool, the ring-theoretic generalization of the classical result on extending valuations (as in [12, Theorem 19.6] or [16, Theorem 56]). I quickly printed (by hand) my solution, handed it to a secretary to be typed appropriately (professors did not have typewriters or computers at that time, but we did have secretaries to type for us), received the typed copy for proofreading just a few minutes later, found the typing to be perfect (that is, accurate), and managed to get my submitted solution mailed to the MAA before the departmental mail was picked up that day. Surely, I thought, with the MAA office just one or two days away by normal mail, my solution had a good chance of being the first to be received. I eagerly awaited the eventual issue of the Monthly magazine, expecting to see my name next to the published solution. (I promised you that this anecdote would make me look all too human.) To my dismay, my name only appeared in the alphabetic list under the heading “also solved by". The published solution did not look familiar to me. But when I saw that the solution was due to Robert Gilmer, my dismay disappeared. At that point in time, Gilmer was indisputably the world leader in multiplicative ideal theory (and perhaps, more generally, on the topic of domains). My feelings were further assuaged when I read Gilmer’s solution and realized that it differed significantly from my solution. In fact, Gilmer’s solution seemed slightly longer than mine. (Yes, more human frailty is on exhibit here, but the story is nearly over.) More importantly, Gilmer’s solution did not use the assumption that the ambient domain was unital. (Remember that the definition of a “domain" in [12] does not require the unital property.) So, quite likely, I had achieved a Pyrrhic victory, in that my solution was probably the first to arrive at the office of the MAA Monthly, but Gilmer's solution was eventually deemed to be better by the powers that be, presumably because his solution assumed less (and, therefore by the standards that most mathematics would use for such matters, was the "better", more elegant solution). This experience taught me the following valuable lesson: although some things are easier to do when an rng happens to be unital, one should always be alert to the possibility that a result that has just been proved could be susceptible to a different line of reasoning, perhaps coming from a different mathematical genre, leading to a more elegant/economical proof.

The last few paragraphs may have caused some readers who are *aficionados* of domains to wonder if "domains" really should be unital. At this point, I cannot claim that the above material has convincingly presented the case for an affirmative answer. I do think, however, that the next paragraph will help to make that case (especially in the minds of any of the just-mentioned *aficionados* who are not yet convinced about this matter). I also think that the development of algebra during the past 60 years will also help to make that case. In that regard, following the next paragraph, please see the subsequent seven paragraphs. There, you will find what I consider to be the most convincing reasons why domains should be unital. Those seven paragraphs give what was, in my experience as a student and a young professional, the beginning of a series of critical observations about modules. The material in the initial five of those seven paragraphs comes from an exercise (that I recall working nearly 60 years ago) from van der Waerden's classic textbook "Modern Algebra."

First, recall that the comments at the beginning of this section embedded any rng R as a subrng of some (unital) ring $\mathcal{R}$. However, if R happened to be a domain, then that construction could not be guaranteed to produce a ring $\mathcal{R}$ which is a domain. Indeed, if the characteristic of R is some prime number p, then that ring $\mathcal{R}$ is definitely *not* a domain, the point being that if r is any nonzero element of R, then $(0,r)\cdot(p,0) = (0,pr) = (0,0) = 0$ in $\mathcal{R}$. However, we show next that a more suitable embedding is available. For clarity, let us change notation and begin with a domain D (in the sense of [12]) which is definitely not unital. To avoid trivialities, one supposes that $D \neq \{0\}$, since the/a zero ring cannot be a unital subring of any unital domain. According to [12], D has a quotient field, say K. (More generally, I learned from a seminar talk by Kaplansky at UCLA in the spring quarter of 1970 that special cases of what we now call rings of fractions R_S were anticipated (long ago, before I was born) by workers such as Grell, with the role of 1 in R_S being played by the fraction s/s for any element $s \in S$. For more about this, see [12] and [18, pages 138-139].) I will next show that D can be embedded as a subrng of some domain $\mathcal{D}$ such that $\mathcal{D}$ is a unital domain and $\mathcal{D}$ also has K as its quotient field. (As [12] emphasizes the importance of such an "overring" extension in multiplicative ideal theory, I find this result, whose proof will follow next, to be especially persuasive.) Observe that K is a ring, with multiplicative identity element $1 = s/s$ for any nonzero element s of D. Take $\mathcal{D}$ to be the subring of K that is generated by D and 1. (In other words, take $\mathcal{D}$ to be the intersection of all the subrings of K which contain D and, necessarily, 1.) Then $\mathcal{D}$ is a (unital) subring of K (since $\mathcal{D}$ is a subrng of K such that $1 \in \mathcal{D}$), so $\mathcal{D}$ is a "domain with 1". Of course, we also have that D is a subrng of $\mathcal{D}$ and that K is a quotient field of $\mathcal{D}$.

The benefits of changing from predominantly ideal-theoretic reasoning to module-theoretic reasoning in commutative ring theory were widely recognized and took hold during the late 1950s and 1960s, producing many useful generalizations and new methods. Prior to that, in part because of the embedding result discussed in the first paragraph of this section, there was natural interest in deciding whether a "module" over a (unital) ring should, by definition, be required to be unital. Many mathematicians were convinced that this question should be answered in the affirmative (and I concur with them) because of the following result from van der Waerden's textbook. Let R be a not necessarily commutative (but unital) ring and let M be a left module over R. Then M can be uniquely expressed as an internal direct sum of (not necessarily unital left) R-modules, $M = M_1 \oplus M_2$, where M_1 *is* a unital (left) R-module and the action of R on M_2 is like the action of a zero ring on M_2 (in the sense that $r \cdot m = 0$ for all $r \in R$ and all $m \in M_2$).

Proof of uniqueness: Suppose that $M = M_1 \oplus M_2 = N_1 \oplus N_2$, where M_1 and N_1 are each unital (left) R-modules and R acts as a zero ring on both M_2 and N_2. We will prove that $M_1 = N_1$ and $M_2 = N_2$. Suppose first that $u \in M_1$. By hypothesis, $u = v + w$ for some uniquely determined $v \in N_1$ and $w \in N_2$. Then $u - v = w$ satisfies

$$u - v = 1 \cdot u - 1 \cdot v = 1 \cdot (u - v) = 1 \cdot w = 0 \cdot w = 0,$$

whence $u = v$. Hence $M_1 \subseteq N_1$. Similarly, $N_1 \subseteq M_1$. Thus $M_1 = N_1$.

Suppose next that $x \in M_2$. Then $x = y + z$ for some uniquely determined $y \in N_1$ and $z \in N_2$. Then $x - z = y$ satisfies

$$x - z = y = 1 \cdot y = 1 \cdot (x - z) = 1 \cdot x - 1 \cdot z = 0 \cdot x - 0 \cdot z = 0 - 0 = 0,$$

whence $x = z$. Hence $M_2 \subseteq N_2$. Similarly, $N_2 \subseteq M_2$. Thus $M_2 = N_2$. This completes the proof of the uniqueness assertion.

Proof of existence: Let $M_1 := \{x \in M \mid 1 \cdot x = x\}$. It is straightforward to check that M_1 contains 0 and is closed under scalar multiplication from R, sums and differences, and so M_1 is a not necessarily unital R-submodule of M. But M_1 is then also clearly a unital R-module. Next, let $M_2 := \{y \in M \mid 1 \cdot y = 0\}$. It is straightforward to check that M_2 contains 0 and is closed under scalar multiplication from R, sums and differences, and so M_2 is a not necessarily unital R-submodule of M. In fact, R acts as a zero ring on M_2 since, if $r \in R$ and $y \in M_2$, then $r \cdot y = (r \cdot 1) \cdot y = r \cdot (1 \cdot y) = r \cdot 0 = 0$. It remains only to prove that M is the internal direct sum of M_1 and M_2, that is, that $M_1 + M_2 = M$ and $M_1 \cap M_2 = 0$.

Let $u \in M$. Put $v := 1 \cdot u$ and $w := u - v$. Observe that $1 \cdot v = 1 \cdot (1 \cdot u) = (1 \cdot 1) \cdot u = 1 \cdot u = v$, whence $v \in M_1$; and, since we have just noted that $1 \cdot u = v = 1 \cdot v$, we have $1 \cdot w = 1 \cdot u - 1 \cdot v = v - v = 0$, whence $1 \cdot w = 0$, whence $w \in M_2$. Hence $u = v + w \in M_1 + M_2$, and so $M \subseteq M_1 + M_2$. The reverse inclusion is obvious, and so $M_1 + M_2 = M$. Finally, we need only show that if $z \in M_1 \cap M_2$, then $z = 0$. This, in turn, holds since $0 = 1 \cdot z$ (as $z \in M_2$) and $1 \cdot z = z$ (as $z \in M_1$). This completes the proof of the existence assertion. This completes the proof.

I would suggest that the main point to be gleaned from the result in the past five paragraphs is this. Because of the nature of the direct summands in the direct sum decomposition $M = M_1 \oplus M_2$, that result has reduced the study of non necessarily unital modules to the following two studies: the study of unital modules and the study of abelian groups (because a not necessarily unital module on which the ambient ring acts as a zero ring is nothing more than an abelian group). Hence, from the point of view of a ring-theorist, “modules" should be unital, as *other* considerations involving “not necessarily unital modules" have been reduced to (abelian) group theory. If a reader believes that my conclusion is outlandish, I can assure you that it is torn from the pages of history. Specifically (yes, here comes another anecdote): each academic year during the late 1960s and early 1970s, UCLA’s mathematics department hosted promising postdocs, some folks on sabbatical, some mid-career specialists and senior leaders in a particular field of mathematics (the field varied annually). The field in 1969-70 was “Algebra", and I was lucky enough to be invited to participate as a Visiting Professor for the entire year. Many of the visitors were present for only three weeks, during which such visitors were obliged to give three lectures per week. One of the year-long visitors, S. A. Amitsur, gave three lectures a week for the entire academic year. More than half of those lectures were devoted to a theorem that he had only recently proved. The statement of the theorem could be given in many formulations, some of which involved noncommutative ring theory (and were thus of interest to many of those present for the “Algebra year") and one of which involved classical geometries (and hence was of interest to me, largely because of my masters studies in 1964-65 in Canada). At the end of his last lecture, Amitsur declared that, from the point of view of a ring theorist, he had just completed the solution of the overall problem that his lectures had been devoted to. There was a stunned silence in the crowd, as none of us in attendance could “connect the dots". Amitsur sensed our confusion (perhaps he had anticipated it) and then, with a twinkle in his eye, he added a fuller explanation. His analysis had reduced the overall problem at hand to a problem in group theory and so, he concluded, our interest in it, as ring-theorists, was now at an end. One by one, the audience members grinned as the wisdom of Amitsur's comment sank into their understanding, and we rose in applause of Amitsur's great accomplishment. While the preceding five paragraphs concern much, much lower-level mathematics, I suggest that they have made a similar point, hopefully as convincingly as Amitsur did in 1970.

The late 1950s and 1960s witnessed what has been called an “invasion" (I would prefer the term “infusion") of homological algebra into many areas of algebra. This use of homological and categorical predilections has continued and, in my opinion, has enriched much of algebra and its applications. A principal effect has been that there is now widespread agreement that ring homomorphisms and algebra homomorphisms should be unital. Of course, one would argue, algebras should be unital since rings should be unital and, after all, an extension involving commutative rings is an example of an algebra, is it not? More generally, given a commutative ring R and an R-algebra S (for a commutative ring R, this means that there is a ring homomorphsim f from R to the center of S), it has long been traditional to view S as a (left) R-module via $r \cdot s := f(r)s$ for all $r \in R$ and all $s \in S$. By taking $s := 1$, we see that the *only* way for this module to be unital (and we have been arguing that modules *should* be unital) is for f to be unital. Once one agrees that algebra homomorphisms should be unital, one must agree that ring homomorphisms should be unital (the point being that every ring is a $\mathbb{Z}$-algebra).

I hope that this section has given the reader some food for thought. When it comes to a discussion of values, one cannot hope to *prove* that one's views are “correct" and that others' views are “wrong." I can only hope that this section will be of help to anyone who is hesitating as to whether their rings (or their modules or their homomorphisms) should be unital. I will have accomplished my goal for this section if such readers understand better what they may expect to gain or lose as a result of any particular decision they may make about such matters.

5 Appendix III: Some professional preoccupations with beginners' angst

Some of the questions that were raised in Appendix I indicate that many beginning students of category theory and/or algebraic geometry express concerns about the use of the definite article “the" instead of the indefinite articles “a" or “an" in describing a mathematical object that is only well defined up to isomorphism. (Such angst is often manifested in regard to constructions such as A_S or $\varinjlim_{i \in \mathbf{I}} A_i$, and it is only compounded by the use of notation such as $\varinjlim_{P \in X_a} A_a$, which contains at least two such stimuli for concern.) As a beginning graduate student and later in doing my doctoral research, such worries arose naturally in the course of my reading and my research. For instance, the n^{th} piece of Amitsur's cochain complex (cf. [3]) is obtained by applying the units functor U (also known as G_m) to the tensor product, over a given field K, of $n+1$ copies of a field extension L of K. It is natural to ask what it means to apply a functor to something that is only defined up to isomorphism, and so I had some concern about the well-definedness of Amitsur's cochains. That concern compounded when I needed to address the (co)homology groups inferred from Amitsur's cochain complex, since *the* n^{th} such group was defined as *the* quotient group of *the* group of n^{th} cocycles modulo *the* group of n^{th} coboundaries. It is natural to ask what it means to be *the* factor group G/N when a group G and its normal subgroup N are each only defined up to isomorphism. Such concerns intensified during the first week of my doctoral research, as part of my assignment for that week was to read [7] where, *inter alia*, Amitsur's field extension $K \subset L$ was generalized to any (perhaps one should add “faithful") commutative algebra (over a commutative ring) and the units functor U was generalized to any abelian group-valued functor on a suitable category of algebras. Of course, the relevant cohomology groups were generalized. For an R-algebra S and a functor F, *the* associated n^{th} cohomology group was denoted by $\mathrm{H}^n(S/R,F)$. That was quite a first week of work, as my assignment also included reading a book about profinite groups. (That actually was not as difficult, even though it mixed algebra with topology, because the relevant inverse limit defining a profinte completion was truly *the* inverse limit of some unambiguous things indexed by an unambiguous directed set.) My unease was triply compounded, even that first week, because I knew that my area of doctoral research was not going to be Amitsur cohomology – it was going to be Cech cohomology and “*the*" n^{th} Cech cohomology group of a given object R and a given functor/presheaf F (in something like a Grothendieck topology $\mathcal{T}$ – yes, I also had to quickly absorb M. Artin's 1962 Harvard notes on Grothendieck topologies) is *the* direct limit of *the* corresponding Amitsur cohomology groups $\mathrm{H}^n(S/R,F)$ as S ranges over some appropriate directed set of objects drawn from $\mathcal{T}$. I quickly realized that my advisor should not be bothered with my triply-provoked concerns, but I resolved to identify the secret by which mathematicians had decided that some super version of the Axiom of Choice could be used to turn all of those occurrences of what should perhaps have been “*a*" or *an*" into occurrences of “*the*".

The semester before beginning my doctoral research, I took a very stimulating course on homological algebra. It was taught by Professor Len Silver and its official textbook was the classic work by Cartan and Eilenberg. As that work was already 10 years old by then, I realized that it would be advisable for me to try to understand many of the ideas in Professor Silver's class in a more general categorical setting. Fortunately, one of the sources that I chose to read in order to learn more about category theory during my “spare time" (what graduate student ever has any spare time?) was Grothendieck's classic paper [14], which was then widely known as “Tohoku". Fortunately, in reading (and re-reading) [14], I came across a passage that stuck in my memory. It is on page 133 of [14] and it is quoted in the next paragraph. By remembering that passage, I was able (a few months later, when I began my doctoral research) to unlock the “secret" that I had resolved to identify. It turns out that the “super Axiom of Choice" that I supposed must lie at the crux of the secret has to do with a well ordered set-theoretic universe. The availability of that universe is due (depending on one's point of view) to one or both of the following: Hilbert's desire to have the benefits of a rather strong Axiom of Choice, without explicitly committing himself to such an axiom, but instead introducing (c. 1923) certain operators, dubbed τ and ϵ, which had certain desirable properties; and Gödel's construction (barely 10 years later) of the model V for ZFC set theory which featured a well -ordered universe. In the next two paragraphs, I will say a little more about the first of these matters, having to do with Grothendieck's use of the Hilbert symbol τ. The final three paragraphs will discuss, *inter alia*, well-ordered universes.

In [14, page 133], Grothendieck addressed and dismissed some concerns similar to the ones that were mentioned in the first sentence of this appendix. He focused on the well-definedness of direct limits in the following passage (the rather literal translation is mine, but the usage of italics is from the original): “In particular, two direct limits of the same directed system are canonically isomorphic (in an evident sense), also it is natural to choose, for each directed system that admits a direct limit, one such direct limit (for example by means of Hilbert's symbol τ), which we will then denote by $\varinjlim \mathbf{A}$ or $\varinjlim_{i\in\mathbf{I}} A_i$ and which we will call *the* direct limit of the given directed system. If $\mathbf{I}$ and $\mathbf{C}$ are such that $\varinjlim \mathbf{A}$ exists for *every* directed system $\mathbf{A}$ indexed by $\mathbf{I}$ with values in $\mathbf{C}$, it follows from the above that $\varinjlim \mathbf{A}$ is a *covariant functor* defined on the category of directed systems indexed by $\mathbf{I}$ in $\mathbf{C}$, with values in $\mathbf{C}$." I can only suppose that in referring to “Hilbert's symbol τ", Grothendieck was assuming familiarity with an earlier (French language) edition of the appropriate chapter of [6].

My online searches in April 2023 indicated that this year (2023) marks the centennial of Hilbert's introduction of the operator τ. In this regard, I would like to mention some recent work of M. Abrusci and his collaborators having to do with some philosophical/mathematical questions concerning quantification and proof. First, one should acknowledge that there seems to be a widespread impression online to the effect that "Hilbert's symbol τ" had really originally been Hilbert's symbol "ϵ" and that various workers had decided to change the notation "ϵ" to "τ" some time before the original French edition of the relevant chapter of [6], presumably in order to avoid confusion between "ϵ" and the set-theoretic symbol "$\in$". However, this widespread belief seems to have been refuted by Abrusci in [1], as can be seen from the following beginning of the author's (that is, Abrusci's) summary of that work: "In section 1, I expose in an informal way the rules – and the logical rules – on the proofs of the universal statements and existential statements, and the rules – and the logical rules – on the deductions from these statements. In section 2, I show how Hilbert's operators τ and ϵ allow a representation of the universal statements and existential statements which is strictly related to the logical rules on the proofs of these statements and to the logical rules on the deductions from these statements, so that we may say that Hilbert in the introduction of the operators τ and ϵ aimed to propose a kind of proof-theoretical representation of the universal statements and existential statements." In joint work [2] with Pasquali and Retoré two years earlier, Abrusci makes clear that the set-theoretic foundational concerns of the late 19th century which mathematicians typically associate with people such as Cantor and Frege (concerns which were only heightened by Hilbert's formalist pronouncement in Königsberg in 1930 that "Wir müssen wiesen. Wir werden wissen" – a belief that was shattered by Gödel's incompleteness results shortly afterward) are shared and are still being examined further to this day in some serious research (however remote such research may seem to be from our daily activities as mathematicians). A sense of the flavor and scope of [2] can be gotten from its Math. Review by B. H. Mayoh: "Quantifiers are ubiquitous in natural language. This paper presents many approaches to capturing the complexity of natural language quantification and suggests a new proof-theoretic approach. First, the authors discuss the classical universal and existential quantifiers and why G. F. L. Frege rejected the appealing idea of domain restriction. Next they present individual concepts, second-order logic and various Hilbert operators. Finally, they present a section on generalized quantifiers. Many problems remain." If there are any readers who wish to learn more about some serious, current, professional studies related to the τ and ϵ operators, I would encourage them to look into the extensive literature on what is nowadays called the "epsilontic calculus?.

In my experience, a working algebraist can occasionally benefit by attention to foundational matters. Consider, for example, the following result in category theory: a functor is a categorical equivalence if (and only if) it is fully faithful and essentially surjective. As a doctoral student, I first came across this result when I read its use by Bass in [5, Chapter II, 1.2] for some work on algebraic K-theory. Although Bass did not mention any foundational issues that may arise when using that categorical result, the only proof that I know of that result requires that some well ordering be applicable to the domain category of the given functor (certainly a well ordering of the class of objects of that category, perhaps also something like – or more than – the well ordering of each set of morphisms with a given domain and a given codomain in that category). Thanks to a famous result of Gödel [13], there *is* a model satisfying the ZFC (Zermelo-Frankel and the Axiom of Choice) foundations whose universe is well ordered.

Some mathematicians have occasionally used the above characterization of a categorical equivalence to conclude that every category is equivalent to a skeletal category, that is, to a category in which any two isomorphic objects are equal. While this would be acceptable (assuming ZFC) for a *small* category (that is, a category whose class of objects is a *set*), the famous paradoxes of intuitive set theory have led several mathematicians to conclude that many important categories are not small. In reading authors such as Grothendieck or Mac Lane (see, especially, [17, pages 23-24 and 30]), I have often had the impression that they preferred the meaning of "set" to be placed on a "sliding scale", that is, to be adjusted in accordance with the data for the problem at hand. It has been said that although most mathematicians profess to be formalists in their official pronouncements on foundational matters, we tend to think, create and act like Platonists, as though the objects of our professional attention are "ideal" things, in the spirit of "The Republic of Plato." Is there a better way to guarantee access to such ideal things than to have a well-ordered universe?

The fact that having a well ordered universe is consistent with ZFC allowed me to access and use what I called "chosen fields" to construct a functor in [8, Definition 3.8, page 24] which had several cohomologically useful applications (cf. [8, Chapter I, Theorems 3.10, 3.13 and 5.9]). The "chosen fields" were also instrumental in my proof of a very useful result [9, Theorem 2.2] stating that for any field k, in the étale toplogy for Spec(k), there is a left adjoint functor sending presheaves to what may be called "additive presheaves" in a way that is analogous to the "sheafification" functor that sends presheaves to sheaves (in a more general context, of course). My research has perhaps had only two other noteworthy interactions with mathematical logic: in [11, Proposition 2.5 (a)], A. Hetzel and I worked with countable models to prove the "lifting" result that a ring homomorphism is a chain morphism if (and only if) it is an n-chain morphism for every positive integer n; and in [10], R. C. Heitmann and I showed that the answer to a certain question depends on which model of ZFC is being used. That question asked to determine those infinite cardinal numbers $\aleph_\alpha$ for which there exists a field extension $K \subset L$ such that $\aleph_\alpha$ is the supremum of the set of cardinalities that arise as lengths of chains of intermediate fields contained between K and L. Regardless of whether the reader has found my anecdotes to be interesting or merely self-indulgent, I should close by pointing out that there have been several (I would add "other") interesting questions in algebra whose answers depend on the model of ZFC that is being used. To be brief, let me mention just two of them (in chronological order). In [20], B. L. Osofsky proved that the global dimension of a countable direct product of fields is $k+1$ if and only if $2^{\aleph_0} = \aleph_k$. In [21], S. Shelah proved that the Whitehead Problem is undecidable; that is, he proved that there are two axioms, each of which is consistent with ZFC, that give different answers to the question which asks whether an abelian group A such that $\mathrm{Ext}^1_{\mathbb{Z}}(A,\mathbb{Z}) = 0$ must be a free abelian group.

References

[1] V. M. Abrusci, Hilbert's τ and ϵ in proof theory: a proof-theoretical representation of universal and existential statements. *From arithmetic to metaphysics*, 1?21, Philos. Anal., 73, De Gruyter, Berlin, 2018.

[2] M. Abrusci, F. Pasquali and C. Retoré, Quantification in ordinary language and proof theory, Philos. Sci. (Paris) 20 (1) (2016), 185–205.

[3] S. A. Amitsur, Simple algebras and cohomology groups of arbitrary fields, Trans. Amer. Math. Soc. 90 (1959), 73–112.

[4] M. F. Atiyah and I. G. Macdonald, Introduction to Commutative Algebra, Addison-Wesley, Reading, MA, 1969.

[5] H. Bass, Lectures on Topics in Algebraic K-theory, Tata Institute of Fundamental Research, Bombay, 1967.

[6] N. Bourbaki, Elements of mathematics – Theory of sets, English translation of Théorie des ensembles (Hermann, Publishers in Arts and Science, Paris), Addison-Wesley, Reading (MA)-London-Don Mills (Ont.), 1968.

[7] S. U. Chase, D. K. Harrison and A. Rosenberg, Galois theory and Galois cohomology of commutative rings, Mem. Amer. Math. Soc., Volume 52 (1965).

[8] D. E. Dobbs, Cech cohomological dimensions for commutative rings, Lecture Notes in Math., Volume 147, Springer-Verlag, Berlin-Heidelberg-New York, 1970.

[9] D. E. Dobbs, Amitsur cohomology in additive functors, Can. Math. Bull. 16 (1973), 417–426.

[10] D. E. Dobbs and R. C. Heitmann, Realizing infinite cardinal numbers via maximal chains of intermediate fields, Rocky Mountain J. Math. 44 (5) (2014), 1471–1503.

[11] D. E. Dobbs and A. J. Hetzel, On chain morphisms of commutative rings, Rend. Circ. Mat. Palermo (2), 53 (1) (2004), 71?-84.

[12] R. Gilmer, Multiplicative Ideal Theory, Dekker, New York, 1972.

[13] K. Gödel, The consistency of the axiom of choice and of the generalized-continuum hypothesis with the axioms of set theory, Princeton University Press, Princeton, 1940.

[14] A. Grothendieck, Sur quelques points d'algèbre homologique, Tohoku Math. J. 9 (2) (1957), 119–221.

[15] A. Grothendieck and J. A. Dieudonné, Éléments de Géométrie Algébrique, I, Springer-Verlag, 1971.

[16] I. Kaplansky, Commutative Rings, rev. ed., Univ. of Chicago Press, Chicago, 1974.

[17] S. Mac Lane, Categories for the Working Mathematician, second edition, Graduate Texts in Math., Volume 5, Springer-Verlag, New York, 1998.

[18] N. H. McCoy, The Theory of Rings, Macmillan, New York, 1964.

[19] D. Mumford, Introduction to Algebraic Geometry (preliminary version of first 3 chapters), undated notes from Harvard University [my copy was obtained in January 1967].

[20] B. L. Osofsky, Homological dimension and cardinality, Trans. Amer. Math. Soc. 151 (1970), 641–649.

[21] S. Shelah, Infinite abelian groups – Whitehead problem and some constructions, Israel J. Math. 18 (1974), 243–256.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 23-35

ISSN: 2820-7114

Title :

On semi radical ideals of noncommutative rings

Author(s):

Nico Groenewald

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 23–35

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco
ISSN: 2820-7114

On semi radical ideals of noncommutative rings

Nico Groenewald
Department of Mathematics, Nelson Mandela University, Port Elizabeth, South Africa
e-mail: *nico.groenewald@mandela.ac.za*

Communicated by Najib Mahdou
(Received 01 July 2023, Revised 18 September 2023, Accepted 21 September 2023)

Abstract. Since the introduction of n-ideals and J-ideals in commutative rings many different aspects of these ideals have been investigated. As a generalization the notion of weakly n-ideals and weakly J-ideals was introduced and studied. Recently it was proved that many of the results are also true for noncommutative rings as a special case of a more general situation. In a recent paper Khashan et. al introduced the notion of semi n-ideals as a generalization of n-ideals where n is the prime radical and studied this generalization. In this note we show that these results are special cases of a more general situation. If ρ is a special radical and R a noncommutative ring then the ideal I of R is a semi ρ-ideal if $aRa \subseteq I$, then $a \in \rho(R)$ or $a \in I$. This covers a wide spectrum of semi ideals and if ρ is the prime radical we have the notion of semi n-ideals for noncommutative rings. In this note we prove that most of the results for the semi n-ideals are satisfied for noncommutative rings as a special case.

Key Words: Special radical, semi n-ideal, semi ρ-ideal, semi ρ-submodule.
2010 MSC: 16N20, 16N40, 16N80, 16L30.

1 Introduction

Throughout this paper, all rings are assumed to be noncommutative with nonzero identity. We recall that a proper ideal I of a ring R is called semiprime if whenever $a \in R$ is such that $aRa \subseteq I$, then $a \in I$. In 2017, Tekir, Koc and Oral in [10] introduced the concept of n-ideals of commutative rings. A proper ideal I of a commutative ring R is called an n-ideal if whenever $a, b \in R$ are such that $ab \in I$ and $a \notin \mathcal{P}(R)$, then $b \in I$ where $\mathcal{P}(R)$ is the prime radical of the ring R. Recently, Khashan and Bani-Ata in [8] generalized n-ideals by defining and studying the class of J-ideals. A proper ideal I of R is called a J-ideal if $ab \in I$ and $a \notin J(R)$ imply $b \in I$ for $a, b \in R$, where $J(R)$ denotes the Jacobson radical of R. In [5] Groenewald introduce the notion of ρ-ideals for a noncommutative ring and a special radical ρ. An ideal I of a noncommutative ring R is a ρ-ideal if for $a, b \in R$ such that $aRb \subseteq I$ and $a \notin \rho(R)$, then $b \in I$. In [1] the notion of a semi n-ideal is introduced as a new generalization of the concept of n-ideals by defining a proper ideal I of a commutative ring R to be a semi n-ideal if whenever $a \in R$ is such that $a^2 \in I$, then $a \in \mathcal{P}(R)$ or $a \in I$. Some examples of semi n-ideals are given and semi n-ideals are investicated under various contexts. In this paper we introduce the notion of semi ρ-ideals for a special radical ρ and a noncommutative ring R as new generalization of the concept of ρ-ideals. If I is an ideal of the noncommutative ring R and ρ is a special radical, then I is a semi ρ-ideal if $aRa \subseteq I$ and $a \notin \rho(R)$, then $a \in I$. The class of semi ρ-ideals is a generalization of semiprime and n-ideals. We start Section 2 by giving some examples (see Example 2.4) to show that this generalization is proper. Next, we determine several characterizations of semi ρ-ideals for a special radical ρ. In the rest of the paper ρ will always be a special radical. We investigate semi ρ-ideals under various contexts of constructions such as homomorphic images and idealizations, see Propositions 5.1 and 5.2. Moreover, for a direct product of rings $R = R_1 \times R_2 \times \ldots \times R_k$, we determine all semi ρ-ideals of R, see Theorems 3.2 and 3.3.

In 1978, the concept of semiprime submodules is presented. A proper submodule is said to be semiprime if whenever $r \in R, m \in M$ and $rRrm \subseteq N$, then $rm \in N$. See [3] for properties of semiprime submodules. Afterwards, the notions of ρ-submodules are introduced and studied in [5]. A proper submodule N is called an ρ-submodule of M if whenever $rRm \subseteq N$ and $r \notin (\rho(R)M : M)$, then $m \in N$. As a new generalization of the above structures, in Section 4, we define a proper submodule N of M to be a semi ρ-submodule if whenever $rRrm \subseteq N$ and $r \notin (\rho(R)M : M)$, then $rm \in N$. We illustrate (see Example 4.7) that this generalization of ρ-submodules is proper.

In what follows, R is a ring (associative, not necessarily commutative and not necessarily with identity) and M is an $R-R$-bimodule. The idealization of M is the ring $R \boxplus M$ with $(R \boxplus M, +) = (R, +) \oplus (M, +)$ and the multiplication is given by $(r, m)(s, n) = (rs, rn + ms)$. $R \boxplus M$ itself is, in a canonical way, an $R-R$-bimodule and $M \simeq 0 \boxplus M$ is a nilpotent ideal of $R \boxplus M$ of index 2. We also have $R \simeq R \boxplus 0$ and the latter is a subring of $R \boxplus M$. If I is an ideal of R and N is an $R-R$-bi-submodule of M, then $I \boxplus N$ is an ideal of $R \boxplus M$ if and only if $IM + MI \subseteq N$. If ρ is a special radical, it follows from [11] that if R is any ring, then $\rho(R \boxplus M) = \rho(R) \boxplus M$ for all $R-R$-bimodules M. In Proposition 5.1, we clarify the relation between semi ρ-ideals of the idealization ring $R \boxplus M$ and those of R. For the following definitions of special radicals and related results we refer the reader to [12].

A class ρ of rings forms a radical class in the sense of Amitsur-Kurosh if ρ has the following three properties

1. The class ρ is closed under homomorphism, that is, if $R \in \rho$, then $R/I \in \rho$ for every $I \lhd R$.
2. Let R be any ring. If we define $\rho(R) = \sum\{I \lhd R : I \in \rho\}$, then $\rho(R) \in \rho$.
3. For any ring R the factor ring $R/\rho(R)$ has no nonzero ideal in ρ i.e. $\rho(R/\rho(R)) = 0$.

A class $\mathcal{M}$ of rings is a **special class** if it is hereditary, consists of prime rings and satisfies the following condition $(*)$ if $0 \neq I \lhd R$, $I \in \mathcal{M}$ and R a prime ring, then $R \in \mathcal{M}$.

Let $\mathcal{M}$ be any special class of rings. The class $\mathcal{U}(\mathcal{M}) = \{R : R$ has no nonzero homomorphic image in $\mathcal{M}\}$ of rings forms a radical class of rings and the upper radical class $\mathcal{U}(\mathcal{M})$ is called a special radical class.

Let ρ be a special radical with special class $\mathcal{M}$ i.e. $\rho = \mathcal{U}(\mathcal{M})$. Now let $\mathcal{S}_\rho = \{R : \rho(R) = 0\}$. If $\mathcal{P}$ denotes the class of prime rings, then for the special radical ρ it follows from [12] that $\rho = \mathcal{U}(\mathcal{P} \cap \mathcal{S}_\rho)$. For a ring R we have $\rho(R) = \cap\{I \lhd R : R/I \in \mathcal{P} \cap \mathcal{S}_\rho\}$ i.e. ρ has the intersection property relative to the class $\mathcal{P} \cap \mathcal{S}_\rho$.

Let $I \lhd R$, then $\rho(R/I) = \rho^*(I)/I$ for some uniquely determined ideal $\rho^*(I)$ of R with $\rho(I) \subseteq I \subseteq \rho^*(I)$ and $\rho^*(I)$ is called the radical of the ideal I while $\rho(I)$ is the radical of the ring I.

We also have $\rho^*(I) = \rho(R)$ if and only if $I \subseteq \rho(R)$. Also $I = \rho^*(I)$ if and only if $R/I \in \mathcal{S}_\rho$.

In what follows let ρ be a special radical with special class $\mathcal{M}$. Hence $\rho = \mathcal{U}(\mathcal{P} \cap \mathcal{S}_\rho)$.

The following are some of the well known special radicals which are defined in [12], prime radical β, Levitski radical $\mathcal{L}$, Kőthe's nil radical $\mathcal{N}$, Jacobson radical $\mathcal{J}$ and the Brown McCoy radical $\mathcal{G}$.

Definition 1.1. Let ρ be a special radical. A proper ideal I of the ring R is called a ρ-ideal if whenever $a, b \in R$ and $aRb \subseteq I$ and $a \notin \rho(R)$, then $b \in I$.

In [10] and [8] the notions of n-ideals and J-ideals were introduced for commutative rings.

Definition 1.2. [10, Definition 2.1] and [8, Definition 2.1] If ρ is the prime radical or the Jacobson radical of a commutative ring, then a proper ideal I of R is a ρ-ideal if whenever $a, b \in R$ with $ab \in I$ and $a \notin \rho(R)$, then $b \in I$.

Remark 1.3. Let R be a commutative ring and I a proper ideal of R. I is a ρ-ideal if and only if $a, b \in R$ with $ab \in I$ and $a \notin \rho(R)$, then $b \in I$.

2 Semi-ρ-ideals

Definition 2.1. Let ρ be a special radical. A proper ideal I of the ring R is called a semi ρ-ideal if whenever $a \in R$ and $aRa \subseteq I$, then $a \in I$ or $a \in \rho(R)$.

Proposition 2.2. *If ρ is a special radical, then I is a semi ρ-ideal if $I = \rho^*(I)$ or $\rho(R) = \rho^*(I)$.*

Proof. Since $I = \rho^*(I)$ if and only if $R/I \in \mathcal{S}_\rho$, it is clear that I is a semiprime ideal and hence a semi ρ-ideal. Now, if $\rho(R) = \rho^*(I)$ we have that $I \subseteq \rho(R)$ and if $aRa \subseteq I$, then $aRa \subseteq \rho(R)$ and since $\rho(R)$ is a semiprime ideal, we have $a \in \rho(R)$ and hence I is a semi ρ-ideal. □

It is known that if R is a commutative ring and ρ is the prime radical then if I is a semi ρ-ideal then $I = \rho^*(I)$ or $\rho(R) = \rho^*(I)$ (see [1, Proposition 2.2]). It is not clear if this is also the case for noncommutative rings.

Since for any special radical ρ and a ring R, $\rho(R)$ is a semiprime ideal, the following properties of semi ρ-ideals can be easily observed.

Proposition 2.3. *For a special radical ρ and a ring R, the following statements hold.*

1. *Every ρ-ideal is a semi ρ-ideal.*
2. *Every (weakly) semiprime ideal I is a semi ρ-ideal. The converse also holds if $\rho(R) \subseteq I$.*
3. *For every proper ideal I of R, $\rho^*(I)$ is a (semiprime) semi ρ-ideal. In particular, $\rho(R)$ is a semi ρ-ideal of R.*
4. *If I is an ideal such that $I \subseteq \rho(R)$, then I is a semi ρ-ideal.*
5. *If ρ is a special radical and $R \in \mathcal{S}_\rho$, then an ideal I of R is a semi ρ-ideal if and only if it is a semi-prime ideal.*

However, the converses of 1. and 2. in Proposition 2.3 are not true in general.

Example 2.4.

1. Let ρ be a special radical and $R \in \mathcal{S}_\rho$. If I is a nonzero ideal of R then I is a semi ρ-ideal which is not a ρ-ideal. This follows from [5, Proposition 1.5] since $I \neq \rho(R) = \{0\}$.
2. Let $\rho = \mathcal{P}$ and $R = M_2(\mathbb{Z}_{32})$. $I = M_2(\langle\overline{16}\rangle)$ is a semi ρ-ideal which is not a semi prime ideal.

Remark 2.5. If R is an Artinian ring, then since $\beta(R) = \mathcal{L}(R) = \mathcal{N}(R) = \mathcal{J}(R)$= $\mathcal{G}(R)$ the notions of $\beta, \mathcal{L}, \mathcal{N}, \mathcal{J}$ and semi $\mathcal{G}$-ideals are the same. For a commutative ring R, we have $\beta(R) = \mathcal{L}(R) = \mathcal{N}(R)$. Hence for commutative rings the notions semi β, semi $\mathcal{L}$ and semi $\mathcal{N}$-ideals are the same.

Next, we give some equivalent conditions that characterize semi ρ-ideals for a special radical ρ.

Theorem 2.6. Let ρ be a special radical and let I be a proper ideal of a ring R. The following statements are equivalent.

1. I is a semi ρ-ideal of R.
2. Whenever $a \in R$ with $0 \neq aRa \subseteq I$, then $a \in \rho(R)$ or $a \in I$.
3. Whenever $a \in R$ with $\langle a\rangle^2 \subseteq I$, then $\langle a\rangle \subseteq \rho(R)$ or $\langle a\rangle \subseteq I$.
4. If A is an ideal of R such that $A^2 \subseteq I$, then $A \subseteq \rho(R)$ or $A \subseteq I$.
5. If A is an ideal of R such that $A^n \subseteq I$ for some positive integer n, then $A \subseteq \rho(R)$ or $A \subseteq I$.
6. If A is a left ideal (right ideal) of R such that $A^2 \subseteq I$, then $A \subseteq \rho(R)$ or $A \subseteq I$.

Proof. $(1) \Rightarrow (2)$ This is clear.

$(2) \Rightarrow (1)$ Let $a \in R$ such that $aRa \subseteq I$. If $aRa = \{0\}$, then $aRa \subseteq \rho(R)$ and since $\rho(R)$ is a semiprime ideal, we have $a \in \rho(R)$. If $0 \neq aRa \subseteq I$ the result follows from (2).

$(1) \Rightarrow (3)$ Let $a \in R$ with $\langle a\rangle^2 \subseteq I$. Now $aRa \subseteq \langle a\rangle^2 \subseteq I$ and we have $a \in I$ or $a \in \rho(R)$ and hence $\langle a\rangle \subseteq I$ or $\langle a\rangle \subseteq \rho(R)$.

$(3) \Rightarrow (4)$ Let A be an ideal of R such that $A^2 \subseteq I$. Suppose $A \not\subseteq \rho(R)$, then $A^2 \not\subseteq \rho(R)$ since $\rho(R)$ is a semiprime ideal of R. We show that $A \subseteq I$. Suppose $a \in A^2$ and $a \notin \rho(R)$. Let b be any element of A. Now $\langle b\rangle^2 \subseteq A^2 \subseteq I$. If $b \notin \rho(R)$, then $b \in I$ from (3). Suppose $b \in \rho(R)$. We have $(\langle a+b\rangle)^2 \subseteq (\langle a\rangle + \langle b\rangle)^2 \subseteq \langle a\rangle\langle a\rangle + \langle a\rangle\langle b\rangle + \langle b\rangle\langle a\rangle + \langle b\rangle\langle b\rangle \subseteq A^2 \subseteq I$. Hence $\langle a+b\rangle \subseteq I$ or $\langle a+b\rangle \subseteq \rho(R)$. $\langle a+b\rangle \not\subseteq \rho(R)$ for if $\langle a+b\rangle \subseteq \rho(R)$, then $a \in \rho(R)$ a contradiction. Hence $\langle a+b\rangle \subseteq I$. Since $a \in I$, we have $b \in I$ and hence $A \subseteq I$.

$(4) \Rightarrow (5)$ Let $A^n \subseteq I$ for some positive integer n. To prove the argument, we use mathematical induction. If $n \leqslant 2$ the result follows from (4). Assume that the claim of (4) holds for all $2 < k < n$. We show that it is also true for n. Suppose n is even, say, $n = 2t$ for some positive integer t. Now, $A^n = (A^t)^2 \subseteq I$. From (4) we have $A^t \subseteq I$ or $A^t \subseteq \rho(R)$. If $A^t \subseteq \rho(R)$, then $A \subseteq \rho(R)$ since $\rho(R)$ is a semi prime ideal of R. If $A^t \subseteq I$, then by the induction hypothesis, we conclude that $A \subseteq I$. Now, suppose n is odd. Then $n+1 = 2s$ for some $s < n$. Similarly, since $(A^s)^2 \subseteq I$, $(A^s) \subseteq I$ or $A^s \subseteq \rho(R)$. If $A^s \subseteq \rho(R)$, then $A \subseteq \rho(R)$ since $\rho(R)$ is a semi prime ideal of R. If $A^t \subseteq I$, then by the induction hypothesis, we conclude that $A \subseteq I$, so we are done.

$(5) \Rightarrow (4)$ is clear.

$(4) \Rightarrow (6)$ Let T be a left ideal of R such that $T^2 \subseteq I$. Now $TRTR \subseteq T^2R \subseteq I$. From (4) $TR \subseteq I$ or $TR \subseteq \rho(R)$. Since R has an identity, we have $T \subseteq I$ or $T \subseteq \rho(R)$ and we are done.

$(6) \Rightarrow (4)$ is clear.

$(4) \Rightarrow (1)$ Let $a \in R$ such that $aRa \subseteq I$. Now $RaRRaR \subseteq I$ and from (4) we have that $a \in RaR \subseteq I$ or $a \in RaR \subseteq \rho(R)$ and we are done. □

Lemma 2.7. *Let ρ be a special radical and I and J be ideals of R with $I, J \not\subseteq \rho(R)$. Then*

1. *If I and J are semi ρ-ideals with $I^2 = J^2$, then $I = J$.*

2. *If I^2 is a semi ρ-ideal, then $I^2 = I$.*

Proof. 1. Since $I^2 \subseteq J$ and $I \not\subseteq \rho(R)$, then by Theorem 2.3, we have $I \subseteq J$. Similarly, since $J^2 \subseteq I$ and $J \not\subseteq \rho(R)$, we have $J \subseteq I$. Thus, we have the equality.

2. Since $I^2 \subseteq I^2$, $I \not\subseteq \rho(R)$ and I^2 is a semi ρ-ideal, we have $I \subseteq I^2$ and so $I^2 = I$. □

Proposition 2.8. *Let ρ_1 and ρ_2 be two special radicals such that $\rho_1 \leq \rho_2$, then every semi ρ_1-ideal is a semi ρ_2-ideal.*

Proof. Let I be a semi ρ_1-ideal of the ring R and suppose $aRa \subseteq I$ and $a \notin \rho_2(R)$. Since $\rho_1 \leq \rho_2$, we have $\rho_1(R) \subseteq \rho_2(R)$ and therefore $a \notin \rho_1(R)$. Since I is a semi ρ_1-ideal, we have $a \in I$ and we are done. □

Remark 2.9. The converse of Proposition 2.8 is not true in general as can be seen from the following example. Consider the local ring $R = \mathbb{Z}_{\langle 2\rangle} = \{\frac{a}{b} : a, b \in \mathbb{Z}, 2 \nmid b\}$ and let $I = \langle 4\rangle_{\langle 2\rangle} = \{\frac{a}{b} : a \in \langle 4\rangle, 2 \nmid b\}$. Since R is a local ring, I is a $\mathcal{J}$-ideal and hence also a semi $\mathcal{J}$-ideal. I is not a semi $\mathcal{P}$-ideal of R. For example, $\left(\frac{2}{3}\right)^2 \in I$ but $\frac{2}{3} \notin \mathcal{P}(R) = \{0\}$ and $\frac{2}{3} \notin I$.

Proposition 2.10. *Let $\{I_i\}_{i\in\Delta}$ be a family of semi ρ-ideals of R, then $\bigcap\limits_{i\in\Delta} I_i$ is a semi ρ-ideal of R.*

Proof. Let $aRa \subseteq \bigcap\limits_{i\in\Delta} I_i$ with $a \notin \rho(R)$ for $a \in R$. Then $aRa \subseteq I_i$ for every $i \in \Delta$. Since I_i is a semi ρ-ideal of R and $a \notin \rho(R)$, we get $a \in I_i$ for every $i \in \Delta$. Hence $a \in \bigcap\limits_{i\in\Delta} I_i$. □

Theorem 2.11. Let R and S be rings and $f : R \to S$ be a surjective ring-homomorphism. If ρ is a special radical, then the following statements hold:

1. If I is a semi ρ-ideal of R and $\ker(f) \subseteq I$, then $f(I)$ is a semi ρ-ideal of S.

2. If J is a semi ρ-ideal of S and $\ker(f) \subseteq \rho(R)$, then $f^{-1}(J)$ is a semi ρ-ideal of R.

Proof. **1.** Let $c \in S$ such that $cSc \subseteq f(I)$ and $c \notin \rho(S)$. Since f is surjective we can choose $a \in R$ such that $f(a) = c$. Now, $cSc = f(a)f(R)f(a) = f(aRa) \subseteq f(I)$ and since $\ker(f) \subseteq I$, we have $aRa \subseteq I$. Because $c \notin \rho(S)$ we have $a \notin \rho(R)$ for if $a \in \rho(R)$, then $c = f(a) \in f(\rho(R) \subseteq \rho(S)$ since ρ is a special radical. Thus $a \notin \rho(R)$ and since $aRa \subseteq I$ and a semi ρ-ideal of R, we get $a \in I$. Hence $c = f(a) \in f(I)$ and therefore $f(I)$ is a semi ρ-ideal of S.

2. Let $a \in R$ such that $aRa \subseteq f^{-1}(J)$ and $a \notin \rho(R)$. Now, $f(a)Sf(a) = f(aRa) \subseteq J$. We show that $f(a) \notin \rho(S)$. Suppose $f(a) \in \rho(S)$ and $M \lhd R$ such that $R/M \in \mathcal{S}_\rho \cap \mathcal{P}$. Since f is a surjective homomorphism and $\ker(f) \subseteq \rho(R) \subseteq M$, we have $f(R)/f(M) \simeq R/\ker(f)/M/\ker(f) \simeq R/M$. Hence $f(R)/f(M) \in \mathcal{S}_\rho \cap \mathcal{P}$ and therefore $f(a) \in f(M)$. Hence $a \in M$ since $\ker(f) \subseteq M$ and therefore $a \in \cap\{I \lhd R : R/I \in \mathcal{P} \cap \mathcal{S}_\rho\}$ $=\rho(R)$ which is a contradiction. Since J is a semi ρ-ideal, we have $f(a) \in J$ and so $a \in f^{-1}(J)$. It follows that $f^{-1}(J)$ is a semi ρ-ideal of R. □

Corollary 2.12. *Let ρ be a special radical and let R be a ring and let I, K be two ideals of R with $K \subseteq I$. Then the following hold.*

1. *If I is a semi ρ-ideal of R, then I/K is a semi ρ-ideal of R/K.*

2. *If I/K is a semi ρ-ideal of R/K and $K \subseteq \rho(R)$, then I is a semi ρ-ideal of R.*

3. *If I/K is a semi ρ-ideal of R/K and K is a semi ρ-ideal of R, then I is a semi ρ-ideal of R.*

Proof. **1.** Assume that I is a semi ρ-ideal of R with $K \subseteq I$. Let $\pi : R \to R/K$ be the natural epimorphism defined by $\pi(R) = r + K$. Note that $\ker(\pi) = K \subseteq I$. Thus, by Theorem 2.11 1., it follows that $\pi(I) = I/K$ is a semi ρ-ideal of R/K.

2. Again consider the natural epimorphism $\pi : R \to R/K$. Since $K \subseteq \rho(R)$, by Theorem 2.11 2., $I = \pi^{-1}(I/K)$ is a semi ρ-ideal of R.

3. This is clear by 2. and Theorem 2.11. □

Proposition 2.13. *Let ρ be a special radical and let I and J be two semi ρ-ideals in a ring R. If $I + J$ is proper in R, then $I + J$ is a semi ρ-ideal of R.*

Proof. By (1) of Corollary 2.12, $I/I \cap J$ is a semi ρ-ideal of $R/I \cap J$. Thus, $(I+J)/J \cong I/I \cap J$ is also a semi ρ-ideal of R/J. Therefore, by (2) of Corollary 2.12, we conclude that $I + J$ is a semi ρ-ideal of R. □

However, if I and J are two semi $\mathcal{P}$-ideals in a ring R, then IJ need not be a semi $\mathcal{P}$-ideal. For example, while $M_2(\langle 2\rangle)$ is a semi $\mathcal{P}$-ideal of $M_2(\mathbb{Z})$, $(M_2(\langle 2\rangle))^2 = M_2(\langle 4\rangle)$ is not so.

Let I be a proper ideal of R, then $Z_I(R)$ denote the set $\{r \in R : sr \in I$ for some $s \in R\backslash I\}$.

Proposition 2.14. *Let ρ be a special radical and R a ring with S a non-empty subset of R where $\langle S\rangle \cap Z_{\rho(R)}(R) = \emptyset$. If I is a semi ρ-ideal of R with $S \nsubseteq I$, then $(I : \langle S\rangle)$ is a semi ρ-ideal of R.*

Proof. Let $a \in R$ such that $aRa \subseteq (I : \langle S\rangle)$ but $a \notin \rho(R)$. Then $asRas \subseteq aRa\langle S\rangle \subseteq I$ for all $s \in \langle S\rangle$. As I is a semi ρ-ideal of R, we have either $as \in \rho(R)$ or $as \in I$ for all $s \in \langle S\rangle$. If $as \in \rho(R)$, then $\langle S\rangle \cap Z_{\rho(R)}(R) \neq \emptyset$, a contradiction. Thus, $as \in I$ for all $s \in \langle S\rangle$ and so $a \in (I : \langle S\rangle)$ as required. □

Theorem 2.15. Let ρ be a special radical and R a commutative ring. If an ideal I of R is a maximal semi ρ-ideal satisfying $Z_{\rho(R)}(R) \subseteq I$, then I is semi prime in R. Additionally, if $I \subseteq \rho(R)$, then $I = \rho(R)$ is a prime ideal.

Proof. The same as [1, Theorem 3.1] by replacing $\mathcal{P}(R)$ with $\rho(R)$. □

3 Product of rings

Suppose that R_1, R_2 are two noncommutative rings with nonzero identities and $R = R_1 \times R_2$. Then R becomes a noncommutative ring with coordinate-wise addition and multiplication. Also, every ideal I of R has the form $I = I_1 \times I_2$, where I_i is an ideal of R_i for $i = 1,2$. Now, we give the following result.

Proposition 3.1. *Let R_1 and R_2 be two noncommutative rings and let ρ be a special radical such that $\rho(R) = \rho(R_1) \times \rho(R_2)$. Then $R_1 \times R_2$ has no ρ-ideals.*

Proof. Assume that $I = I_1 \times I_2$ is a ρ-ideal of $R_1 \times R_2$, where I_i is an ideal of R_i for $i = 1,2$. Since $(0,1)R_1 \times R_2(1,0) \subseteq I_1 \times I_2$, $(0,1) \notin \rho(R_1 \times R_2) = \rho(R_1) \times \rho(R_2)$ and $(1,0) \notin \rho(R_1 \times R_2) = \rho(R_1) \times \rho(R_2)$, we conclude that $(0,1),(1,0) \in I$ and so $I = R_1 \times R_2$, a contradiction.

By characterizing semi ρ-ideals of R, the next theorem allows us to build some examples for semi ρ-ideals which are not ρ-ideals. □

Theorem 3.2. Let R_1 and R_2 be two noncommutative rings and let ρ be a special radical such that $\rho(R) = \rho(R_1) \times \rho(R_2)$. Then a proper ideal $I = I_1 \times I_2$ is a semi ρ-ideal of R if and only if one of the following statements holds.

1. I is a semiprime ideal of R.
2. I_1 is a semi ρ-ideal of R_1 and $I_2 = \rho(R_2)$.
3. I_2 is a semi ρ-ideal of R_2 and $I_1 = \rho(R_1)$.

Proof. ⇒Suppose $I = I_1 \times I_2$ is a semi ρ-ideal which is not a semiprime ideal. Hence there exists $(x,y) \in R_1 \times R_2$ such that $(x,y)(R_1 \times R_2)(x,y) \subseteq I_1 \times I_2$ but $(x,y) \notin I_1 \times I_2$. We show that $I_1 = \rho(R_1)$ or $I_2 = \rho(R_2)$. Assume not. If $I_1 \neq \rho(R_1)$ and $I_2 \neq \rho(R_2)$, then there exist $a \in I_1 \backslash \rho(R_1)$ and $b \in I_2 \backslash \rho(R_2)$. Now $(x+a)R_1(x+a) = xR_1x + xR_1a + aR_1x + aR_1a \subseteq I_1$ and also $(y+b)R_2(y+b) \subseteq I_2$. From this it follows that $(x+a, y+b)(R_1 \times R_2)(x+a, y+b) \subseteq I_1 \times I_2 = I$. We have $(x,y) \notin I_1 \times I_2$, so without lost of generality we may suppose $x \notin I_1$. Hence $(x+a) \notin I_1$ and so $(x+a, y+b) \notin I$. Since $I = I_1 \times I_2$ is a semi ρ-ideal, we have $(x+a, y+b) \in \rho(R) = \rho(R_1) \times \rho(R_2)$. Hence $(x+a) \in \rho(R_1)$ and $(y+b) \in \rho(R_2)$ which implies that $(x,y) \notin \rho(R)$ since $a \notin \rho(R_1)$ and $b \notin \rho(R_2)$. This is impossible since I is a semi ρ-ideal.

Suppose without loss of generality that $I_1 \neq \rho(R_1)$ and $I_2 = \rho(R_2)$. Let $aR_1a \subseteq I_1$ and $a \notin I_1$. Now, $(a,0)R(a,0) = (aR_1a, 0) \subseteq I_1 \times I_2 = I$. Since $(a,0) \notin I$ and I a semi ρ-ideal, we have $(a,0) \in \rho(R) = \rho(R_1) \times \rho(R_2)$. Hence $a \in \rho(R_1)$ and I_1 is a semi ρ-ideal of R_1. Similarly if $I_1 = \rho(R_1)$ and $I_2 \neq \rho(R_2)$ we get I_2 is a semi ρ-ideal of R_2

⇐ If I is a semiprime ideal of R then I is a semi ρ-ideal of R by Proposition 2.6. Suppose $I = I_1 \times \rho(R_2)$ with I_1 a semi ρ-ideal of R_1. Let $(a,b) \in R = R_1 \times R_2$ such that $(a,b)(R_1 \times R_2)(a,b) \subseteq I_1 \times \rho(R_2)$ and $(a,b) \notin \rho(R) = \rho(R_1) \times \rho(R_2)$. Now, $bR_2b \subseteq \rho(R_2)$ and since $\rho(R_2)$ is a semiprime ideal, we have $b \in \rho(R_2)$. Since $(a,b) \notin \rho(R_1) \times \rho(R_2)$, it now follows that $a \notin \rho(R_1)$. Since $aR_1a \subseteq I_1$ and $a \notin \rho(R_1)$, it follows that $a \in I_1$ from the fact that I_1 is a semi ρ-ideal. Hence we have $(a,b) \in I = I_1 \times \rho(R_2)$ and therefore I is a semi ρ-ideal of R. □

Generalizing Theorem 3.2 we have the following for a special radical ρ such that $\rho(R_1 \times R_2 \times \cdots \times R_n) = \rho(R_1) \times \rho(R_2) \times \cdots \times \rho(R_n)$.

Theorem 3.3. Let $R_1, R_2, ..., R_n$ be rings and $R = R_1 \times R_2 \times \cdots \times R_n$, where $n \geqslant 2$. Then a proper ideal I of R is a semi ρ-ideal if and only if one of the following statements is satisfied.

1. I is a semiprime ideal of R.

2. $I = I_1 \times I_2 \cdots \times I_n$, where I_k is a semi ρ-ideal of R_k for some $k \in \{1,...,n\}$ and $I_j = \rho(R_j)$ for all $j \in \{1,...,n\} \backslash \{k\}$.

Proof. This follows simmilar to the proof of [1, Theorem 3.3]. □

4 Semi ρ-submodules

de la Rosa and Veldsman in [4] defined a weakly special class of modules. We follow the definition in [4] of a weakly special class of modules to define a special class of modules.

Definition 4.1. For a ring R, let $\mathcal{K}_R$ be a (possibly empty) class of R-modules. Let $\mathcal{K} = \cup\{\mathcal{K}_R : R$ a ring$\}$. $\mathcal{K}$ is a special class of modules if it satisfies:

S1 $M \in \mathcal{K}_R$ and $I \lhd R$ with $I \subseteq (0 : M)_R$ implies $M \in \mathcal{K}_{R/I}$.

S2 If $I \lhd R$ and $M \in \mathcal{K}_{R/I}$, then $M \in \mathcal{K}_R$.

S3 $M \in \mathcal{K}_R$and $I \lhd R$ with $IM \neq 0$ implies $M \in \mathcal{K}_I$.

S4 $M \in \mathcal{K}_R$ implies $RM \neq 0$ and $R/(0 : M)_R$ is a prime ring.

S5 If $I \lhd R$ and $M \in \mathcal{K}_I$, then there exists $N \in \mathcal{K}_R$ such that $(0 : N)_I \subseteq (0 : M)_I$.

Following similar techniques of [4], we get the following theorems.

Theorem 4.2. [6, Theoerem 5.1] Let $\mathcal{M} = \cup\mathcal{M}_R$ be a special class of modules. Then, $\mathcal{J} = \{R$: there exists $M \in \mathcal{M}_R$ with $(0 : M)_R = 0\} \cup \{0\}$ is a special class of rings. If ρ is the corresponding special radical, then, $\rho(R) := \cap\{(0 : M)_R : M \in \mathcal{M}\}$.

Theorem 4.3. [6, Theoerem 5.2] Let $\mathcal{J}$ be a special class of rings and for every ring R, let $\mathcal{M}_R = \{M : M$ is an R-module, $RM \neq 0$ and $R/(0 : M)_R \in \mathcal{J}\}$. If $\mathcal{M} = \cup\mathcal{M}_R$, then $\mathcal{M}$ is a special class of modules. If ρ is the corresponding special radical and M is any R-module, then
$\rho(M) := \cap\{P \leq M : M/P \in \mathcal{M}_R\}$.

Definition 4.4. [5, Definition 2.4] Let ρ be a special radical and let M be an R-module. The proper submodule N of M is a ρ-submodule if for all $a \in R$ and $m \in M$, whenever $aRm \subseteq N$ and $a \notin (\rho(R)M : M)$, then $m \in N$.

Definition 4.5. Let ρ be a special radical and let M be an R–module. The proper submodule N of M is a semi ρ-submodule if for all $a \in R$ and $m \in M$, whenever $aRam \subseteq N$ and $a \notin (\rho(R)M : M)$, then $am \in N$.

Definition 4.6. A submodule N of M is said to be semiprime if $N \neq M$ and whenever $r \in R$ and $m \in M$ are such that $rRrm \subseteq N$, then $rm \in N$. The reader clearly observe that any semi ρ-submodule of an R-module R is a semi ρ-ideal of R. The zero submodule is always a semi ρ-submodule of M. Also, see the implications:

$$\begin{tikzcd}
\rho\text{-submodule} \arrow[dr] & \\
& \text{semi } \rho\text{-submodule} \\
\text{semiprime submodule} \arrow[ur] &
\end{tikzcd}$$

However, the next examples show that these arrows are irreversible.

Example 4.7. 1. Consider the submodule $N = 6\mathbb{Z}\times(0)$ of the $\mathbb{Z}$-module $M = \mathbb{Z}\times\mathbb{Z}$. Let the special radical ρ be the prime radical. Now let $r \notin (\mathcal{P}(\mathbb{Z})M : M) = (0)$ and $m = (m_1, m_2) \in M$ such that $r^2\cdot(m_1, m_2) \in N$. Then $r^2 m_1 \in 6\mathbb{Z}$, $r^2 m_2 = 0$. Since $6\mathbb{Z}$ and (0) are semi $\mathcal{P}$-ideals of $\mathbb{Z}$, then $r\cdot(m_1, m_2) \in N$ and so N is a semi $\mathcal{P}$-submodule of M. On the other hand, we have $2\cdot(3, 0) \in N$ with $2 \notin (\mathcal{P}(\mathbb{Z})M : M)$ and $(3, 0) \notin N$ and so N is not a ρ-submodule of M.

2. Consider the submodule $N = \left\langle \overline{4} \right\rangle \times \{0\}$ of the $\mathbb{Z}$-module $M = \mathbb{Z}_8 \times \mathbb{Z}$. Let $r \notin (\mathcal{P}(\mathbb{Z})M : M)$ and $m = (m_1, m_2) \in M$ such that $r^2\cdot(m_1, m_2) \in N$. It is clear to observe that as $\left\langle \overline{4} \right\rangle$ is a semi $\mathcal{P}$-ideal of $\mathbb{Z}_8$ and $\{0\}$ is a semi $\mathcal{P}$-ideal of $\mathbb{Z}$ that $r(m_1, m_2) \in N$. Hence N is a semi $\mathcal{P}$-submodule of M However, $2^2\cdot(\overline{1}, 0) \in N$ but $2\cdot(\overline{1}, 0) \notin N$ and so N is not a semiprime submodule of M.

Proposition 4.8. *Let ρ be a special radical and let M be an R-module. For N a submodule of M and I an ideal of R. If N is a semi ρ-submodule of M and $(\rho(R)M : M) = \rho(R)$, then $(N : M) = \{r \in R : rm \in N$ for every $m \in M\}$ is a semi ρ-ideal of R.*

Proof. Let $aRa \subseteq (N : M)$ where $a \in R$ and $a \notin \rho(R)$. Then we have $aRaM \subseteq N$ and so $aRam \subseteq N$ for all $m \in M$. Since N is a semi ρ-submodule of M and $a \notin \rho(R) = (\rho(R)M : M)$, $am \in N$ for all $m \in M$. Thus, $aM \subseteq N$ and so $a \in (N : M)$. Therefore, $(N : M)$ is a semi ρ-ideal of R. □

Remark 4.9. If $(\rho(R)M : M) \nsubseteq \rho(R)$, then Proposition 4.8 need not be true. Let $\mathcal{P}$ be the prime radical. For the $\mathbb{Z}$ module $M = \mathbb{Z}_4$ we have $\mathcal{P}(\mathbb{Z}) = (0)$ and $(\mathcal{P}(\mathbb{Z})\mathbb{Z}_4 : \mathbb{Z}_4) = ((0) : \mathbb{Z}_4) = 4\mathbb{Z}$. Now, $N = (0)$ is clearly a semi $\mathcal{P}$-submodule. $(N : M) = ((0) : \mathbb{Z}_4) = 4\mathbb{Z}$ is not a semi $\mathcal{P}$-ideal of $\mathbb{Z}$. We have $2\mathbb{Z}2 \subseteq 4\mathbb{Z}$ with $2 \notin 4\mathbb{Z}$.

In the following proposition, we give a characterization of ρ-submodules for a special radical ρ.

Proposition 4.10. *Let ρ be a special radical and let M be an R-module where R is a ring with identity. Let N be a proper submodule of M. Then N is a semi ρ-submodule of M if for any $a \in R$ and every submodule K of M, we have that $aRaK \subseteq N$ with $a \notin (\rho(R)M : M)$ implies $aK \subseteq N$.*

Proof. Suppose $aRaK \subseteq N$ and $a \notin (\rho(R)M : M)$. Let $k \in K$. Since $aRak \subseteq N$ and N is a semi ρ-submodule of M, $ak \in N$. It follows that $aK \subseteq N$ as needed. □

Proposition 4.11. *Let $\varphi : M_1 \rightarrow M_2$ be an R homomorphism. Then*

1. *If φ is surjective and N is a semi ρ-submodule of $M_1$with* $\ker(\varphi) \subseteq N$, *then $\varphi(N)$ is a semi ρ-submodule of M_2.*

2. *If φ is one-to-one and K is a semi ρ-submodule of M_2, then $\varphi^{-1}(K)$ is a semi ρ-submodule of M_1.*

Proof. 1. Suppose $\varphi(N) = M_2 = \varphi(M_1)$ and $m_1 \in M_1$. Then $\varphi(m_1) = \varphi(n)$ for some $n \in N$ and so $(m_1 - n) \in \ker(\varphi) \subseteq N$. So $m_1 \in N$ and we have $N = M_1$ a contradiction. Hence $\varphi(N)$ is a proper submodule of M_2. Let $r \in R$ and $m_2 \in M_2$ such that $rRrm_2 \subseteq \varphi(N)$ and $r \notin (\rho(R)M_2 : M_2)$. Choose $m_1 \in M_1$ such that $\varphi(m_1) = m_2$. Then $rRrm_2 = rRr\varphi(m_1) = \varphi(rRrm_1) \subseteq \varphi(N)$ which implies $rRrm_1 \subseteq N$ as $\ker(\varphi) \subseteq N$. If $rM_1 \subseteq \rho(R)M_1$, then $rM_2 = r\varphi(M_1) = \varphi(rM_1) \subseteq \varphi(\rho(R)M_1) = \rho(R)\varphi(M_1) = \rho(R)M_2$. Hence $r \in (\rho(R)M_2 : M_2)$ a contradiction. Thus $r \notin (\rho(R)M_1 : M_1)$. Since N is a semi ρ-submodule, $rm_1 \in N$ and hence $rm_2 = \varphi(rm_1) \in \varphi(N)$ as required.

2. Let $r \in R$ and $m_1 \in M_1$ such that $rRrm_1 \subseteq \varphi^{-1}(K)$ and $r \notin (\rho(R)M_1 : M_1)$. Since $\ker(\varphi) = 0$, we have $\varphi(rRrm_1) = rRr\varphi(m_1) \subseteq K$. Moreover, we have $r \notin (\rho(R)M_2 : M_2)$ for if $rM_2 \subseteq \rho(R)M_2$, then $r\varphi(M_1) \subseteq \rho(R)\varphi(M_1)$ and so $\varphi(rM_1) \subseteq \varphi(\rho(R)M_1)$. Now, if $x \in rM_1$, then $\varphi(x) \in \varphi(\rho(R)M_1)$. Hence $(x - y) \in \ker(\varphi) \subseteq \rho(R)M_1$ for some $y \in \rho(R)M_1$. Hence $x \in \rho(R)M_1$ and we have $rM_1 \subseteq \rho(R)M_1$ a contradiction. Since K is a semi ρ-submodule of M_2, $r\varphi(m_1) = \varphi(rm_1) \in K$ and hence $rm_1 \in \varphi^{-1}(K)$ and we are done. □

Corollary 4.12. *Let N and L be two submodules of an R-module M with $L \subseteq N$.*

1. *If N is a semi ρ-submodule of M, then N/L is a semi ρ-submodule of M/L.*
2. *If L is a semi ρ-submodule of M and N/L is a semi ρ-submodule of M/L, then N is a semi ρ-submodule of M.*
3. *If L is a ρ-submodule of M and N/L is a semi ρ-submodule of M/L, then N is a ρ-submodule of M.*

Proof. 1. Clear by Proposition 4.11.

2. Suppose that $rRrm \subseteq N$ and $r \notin (\rho(R)M : M)$. If $rRrm \subseteq L$, then $rm \in L \subseteq N$ since L is a semi ρ-submodule of M. So assume $rRrm \nsubseteq L$. One can easily observe that $r \notin (\rho(R)M/N : M/N)$. N/L is a semi ρ-submodule of M/L and $rRr(m+L) \subseteq N/L$, then $r(m+L) \in N/L$. Therefore $rm \in N$ and N is a semi ρ-submodule of M.

3. Similar to 2. □

Proposition 4.13. *Let $\{N_i : i \in \Delta\}$ be a nonempty set of semi ρ-submodules of an R-module M. Then $\bigcap_{i\in\Delta} N_i$ is a semi ρ-submodule.*

Proof. Suppose $rRrm \in \bigcap_{i\in\Delta} N_i$ for some $r \in R-(\rho(R)M : M)$, $m \in M$. Since N_i is a semi ρ-submodule of M, for every $i \in \Delta$, we have $rm \in I_i$. Thus $rm \in \bigcap_{i\in\Delta} N_i$. □

5 Idealization

We now show how to construct ρ-ideals using the Method of Idealization. In what follows, R is a ring (associative, not necessarily commutative and not necessarily with identity) and M is an $R-R$-bimodule. The idealization of M is the ring $R \boxplus M$ with $(R \boxplus M, +) = (R, +) \oplus (M, +)$ and the multiplication is given by $(r, m)(s, n) = (rs, rn + ms)$. $R \boxplus M$ itself is, in a canonical way, an $R-R$-bimodule and $M \simeq 0 \boxplus M$ is a nilpotent ideal of $R \boxplus M$ of index 2. We also have $R \simeq R \boxplus 0$ and the latter is a subring of $R \boxplus M$. Note also that $R \boxplus M$ is a subring of the Morita ring $\begin{bmatrix} R & M \\ 0 & R \end{bmatrix}$ via the mapping $(r, m) \mapsto \begin{bmatrix} r & m \\ 0 & r \end{bmatrix}$. We will require some knowledge about the ideal structure of $R \boxplus M$. If I is an ideal of R and N is an $R-R$-bi-submodule of M, then $I \boxplus N$ is an ideal of $R \boxplus M$ if and only if $IM + MI \subseteq N$.

If ρ is a special radical, it follows from [11] that if R is any ring, then $\rho(R \boxplus M) = \rho(R) \boxplus M$ for all $R-R$-bimodules M.

Proposition 5.1. *For the special radical ρ, let I be an ideal of the ring R. I is a semi ρ-ideal of R if and only $I \boxplus M$ is a semi ρ-ideal of $R \boxplus M$.*

Proof. Let $(r_1, m_1) \in R \boxplus M$ such that $(r_1, m_1) R \boxplus M (r_1, m_1) \subseteq I \boxplus M$ and $(r_1, m_1) \notin \rho(R \boxplus M) = \rho(R) \boxplus M$. Hence $r_1 R r_1 \subseteq I$ and $r_1 \notin \rho(R)$. Since I is a semi ρ-ideal of R, we conclude that $r_1 \in I$ and so $(r_1, m_1) \in I \boxplus M$. Consequently $I \boxplus M$ is a semi ρ-ideal of $R \boxplus M$.

Conversely, suppose that $I \boxplus M$ is a semi ρ-ideal of $R \boxplus M$ and let $aRa \subseteq I$ but $a \notin I$. Then $(a, 0) R \boxplus M (a, 0) \subseteq I \boxplus M$ and $(a, 0) \notin I \boxplus M$ imply that $(a, 0) \in \rho(R \boxplus M) = \rho(R) \boxplus M$. Thus, $a \in \rho(R)$ and we are done. □

If I is a semi ρ-ideal of a ring R and N is a $R-R$-bi-submodule of M with $IM + MI \subseteq N$, then $I \boxplus N$ need not be a semi ρ-ideal of $R \boxplus M$. For example if ρ is the prime radical, $\langle 2 \rangle$ is a semi ρ-ideal of the ring $\mathbb{Z}$ and $\left\{\overline{0}\right\}$ is a submodule of the $\mathbb{Z}$-module $\mathbb{Z}_4$. But $\langle 2 \rangle \boxplus \left\{\overline{0}\right\}$ is not a semi ρ-ideal of $\mathbb{Z} \boxplus \mathbb{Z}_4$ since $(2, \overline{1}) \mathbb{Z} \boxplus \mathbb{Z}_4 (2, \overline{1}) \subseteq \langle 2 \rangle \boxplus \left\{\overline{0}\right\}$ but $(2, \overline{1}) \notin \mathcal{P}(\mathbb{Z} \boxplus \mathbb{Z}_4) = \mathcal{P}(\mathbb{Z}) \boxplus \mathbb{Z}_4$ and $(2, \overline{1}) \notin \langle 2 \rangle \boxplus \left\{\overline{0}\right\}$.

Proposition 5.2. *Let ρ is a special radical and let I be an ideal of R and N a proper $R-R$-bi-submodule of the $R-R$-bi-module M.*

1. *If $I \boxplus N$ is a semi ρ-ideal of $R \boxplus M$, then I is a semi ρ-ideal of R and N is a semi ρ-submodule of M.*
2. *If $(\rho(R)M : M) = \rho(R)$ and N is a ρ-submodule of M with $IM + MI \subseteq N$ and I a semi ρ-ideal then $I \boxplus N$ is a semi ρ-ideal of $R \boxplus M$.*

Proof. (1) Suppose that $I \boxplus N$ is a semi ρ-ideal of $R \boxplus M$. First we show I is a semi ρ-ideal. Let $aRa \subseteq I$ and $a \notin \rho(R)$.Then we have $(a,0)R\boxplus M(a,0) = (aRa,0) \subseteq I\boxplus N$. Since $I\boxplus N$ is a semi ρ-ideal of $R\boxplus M$, and $(a,0) \notin \rho(R)\boxplus M = \rho(R\boxplus M)$ we have that $(a,0) \in I\boxplus N$. Hence $a \in I$ and it follows that I is a semi ρ-ideal of R. Now, we show that N is a semi ρ-submodule of M. Let $aRam \subseteq N$ with $a \notin (\rho(R)M : M)$. Since $a \notin (\rho(R)M : M)$, we have $a \notin \rho(R)$. Then we have $(a, 0_M)R \boxplus M(a, 0_M)(0, m) = (0, aRam) \subseteq I \boxplus N$ with $(a, 0_M) \notin \rho(R\boxplus M)$. Since $I\boxplus N$ is a semi ρ-ideal of $R\boxplus M$, we conclude that $(a, 0_M)(0, m) = (0, am) \in I\boxplus N$ and so $am \in N$, as needed.

(2) Let $(r_1, m_1), (r_1, m_1) \in R \boxplus M$ such that $(r_1, m_1)R \boxplus M(r_1, m_1) \subseteq I \boxplus N$ and $(r_1, m_1) \notin \rho(R \boxplus M) = \rho(R) \boxplus M$. We have $r_1Rr_1 \subseteq I$ and $r_1 \notin \rho(R)$. Since I is a semi ρ-ideal of R and $r_1 \notin \rho(R)$, we have $r_1 \in I$. Now, $(r_1, m_1)R \boxplus M(r_1, m_1) = (r_1Rr_1, r_1Rm_1 + m_1Rr_1) \subseteq I \boxplus N$. Since $r_1Rm_1 + m_1Rr_1 \subseteq N$ and $m_1Rr_1 \subseteq N$, we have $r_1Rm_1 \subseteq N$. Since $r_1 \notin \rho(R)$ and N is a ρ-submodule of M, we have $m_1 \in N$. Hence $(r_1, m_1) \in I \boxplus N$ and $I \boxplus N$ is a semi ρ-ideal of $R \boxplus M$. □

The condition $(\rho(R)M : M) = \rho(R)$ in Proposition 5.2 2. can not be discarded. For example, consider the $\mathbb{Z}$-module $\mathbb{Z}_2$. Put $I = \langle 2\rangle$ and $N = \{\overline{0}\}$. Then I is a semi $\mathcal{P}$-ideal of $\mathbb{Z}$ and N is a $\mathcal{P}$-submodule of $\mathbb{Z}_2$. Also note that $(\mathcal{P}(\mathbb{Z})\mathbb{Z}_2 : \mathbb{Z}_2) = \langle 2\rangle \neq \mathcal{P}(\mathbb{Z}) = \{0\}$. However, $I \boxplus N$ is not a semi $\mathcal{P}$-ideal of $\mathbb{Z} \boxplus \mathbb{Z}_2$ because $(2,\overline{1})\mathbb{Z} \boxplus \mathbb{Z}_2(2,\overline{1}) \subseteq I \boxplus N$, $(2,\overline{1}) \notin \mathcal{P}(\mathbb{Z}) \boxplus \mathbb{Z}_2$ and $(2,\overline{1}) \notin I \boxplus N$.

6 Semi $\mathcal{P}$-ideals (semi n-ideals)

In this section the special radical will be the prime radical. In [1] Khashan et al. introduced the notion of semi n-ideals for commutative rings with identity element. They investigated many properties of semi n-ideals.. We show that for the prime radical many of the results proved by Khashan et al. are also true for noncommutative rings.

In what follows for the noncommutative ring R, $\mathcal{P}(R)$ will denote the prime radical of the ring R.

Throughout this section the rings are noncommutative but not necessarily assumed to have a unity unless indicated.

Definition 6.1. A proper ideal I of a ring R is a semi $\mathcal{P}$-ideal if whenever $a \in R$ such that $aRa \subseteq I$ and $a \notin \mathcal{P}(R)$, then $a \in I$.

If R is a commutative ring, then the notion of a semi $\mathcal{P}$-ideal coincides with a semi n-ideal as been defined by Khashan et al. in [1].

Proposition 6.2. *(see [1, Proposition 2.1]) For a ring R, the following statements hold.*

(1) Every $\mathcal{P}$-ideal is a semi $\mathcal{P}$-ideal.

(2) Every (weakly) semiprime ideal I is a semi $\mathcal{P}$-ideal. The converse also holds if $\mathcal{P}(R) \subseteq I$.

(3) For every proper ideal I of R, $\mathcal{P}^*(I)$ is a (semiprime) semi $\mathcal{P}$-ideal. In particular, $\mathcal{P}(R)$ is a semi $\mathcal{P}$-ideal of R.

(4) Any ideal I such that $I \subseteq \mathcal{P}(R)$ is a semi $\mathcal{P}$-ideal.

(5) If R is a semiprime ring then an ideal I of R is a semi $\mathcal{P}$-ideal if and only if is a semiprime ideal.

Example 6.3. In any semiprime ring R the a nonzero ideal I is a semi $\mathcal{P}$-ideal which is not a $\mathcal{P}$-ideal since $I \nsubseteq \mathcal{P}(R) = (0)$ see [5, Proposition 1.5].

Proposition 6.4. *(See [1, Proposition 3.2]) Let $\{I_i\}_{i\in\Delta}$ be a family of semi $\mathcal{P}$-ideals of R, then $\bigcap\limits_{i\in\Delta} I_i$ is a semi $\mathcal{P}$-ideal of R.*

Proof. This follows from Proposition 2.10 by taking ρ to be the prime radical. □

Proposition 6.5. *Let $\mathcal{P}$ be the prime radical and R a ring with S a non-empty subset of R where $\langle S\rangle \cap Z_{\rho(R)}(R) = \emptyset$. If I is a semi $\mathcal{P}$-ideal of R with $S \nsubseteq I$, then $(I : \langle S\rangle)$ is a semi $\mathcal{P}$-ideal of R.*

Proof. This follows from Proposition 2.14 by taking ρ to be the prime radical. □

Proposition 6.6. *[13, Corollary 4]For any ring R the following are equivalent:*

1. *R has an unique prime ideal.*
2. *R is a local ring and $\mathcal{J}(R) = \mathcal{P}(R)$.*
3. *Every non invertible element is nilpotent.*

Theorem 6.7. The following statements are equivalent for a ring R.

1. $\mathcal{P}(R)$ is the unique prime ideal of R.
2. Every proper ideal of R is an $\mathcal{P}$-ideal.
3. R is a local ring and every proper ideal of R is a semi $\mathcal{P}$-ideal.

Proof. (1) ⇒ (3) Let I be any ideal of R and $a \in R$ such that $aRa \subseteq I$. If $a \in \mathcal{P}(R)$, then we done. If $a \notin \mathcal{P}(R)$ then it follows from Propostion 6.7 that $a \notin \mathcal{J}(R)$ since $\mathcal{P}(R) = \mathcal{J}(R)$. Now, since we also have that R is a local ring, a is an invertible element with inverse b. Now, since $a^2 \in aRa \subseteq I$, we have $a = ba^2 \in I$ and we are done.

(3) ⇒ (1) Let R be a local ring with every proper ideal of R a semi $\mathcal{P}$-ideal. Let M be the unique maximal ideal of R and P a prime ideal of R. Assume that $P \nsubseteq \mathcal{P}(R)$. Since P^2 is a semi $\mathcal{P}$-ideal, it follows from Lemma 2.7 that $P = P^2$. From [7, Corollary 4] $P = \bigcap\limits_{n=1}^{\infty} P^n = \bigcap\limits_{n=1}^{\infty} M^n = (0)$, a contradiction. Hence $P = \mathcal{P}(R)$ and is the unique prime ideal of R.

(2) ⇒ (3) Let M be a maximal ideal right ideal of R and $x \in M$. Since $xR1 \subseteq M$ and M is a $\mathcal{P}$-ideal, then we must have $x \in \mathcal{P}(R)$ and so $M \subseteq \mathcal{P}(R) \subseteq \mathcal{J}(R) \subseteq M$. It follows that $M = \mathcal{J}(R)$ and R is a local ring. The other part of (3) follows directly by Proposition 2.3 (1).

(2) ⇒ (1) Suppose every proper ideal of R is an $\mathcal{P}$-ideal. Let P be any prime ideal. Now, since P is a $\mathcal{P}$-ideal and a prime ideal, it follows from [5, Proposition 1.13] that $P = \mathcal{P}(R)$. Hence $\mathcal{P}(R)$ is the unique prime ideal of R. □

We note that the condition "R is local" in (3) of Theorem 6.7 cannot be omitted. For example, in the ring $M_2(\mathbb{Z}_6)$ every proper ideal is a semi $\mathcal{P}$-ideal but $M_2(\mathbb{Z}_6)$ has no $\mathcal{P}$-ideals. Also it is known that in a local ring every proper ideal is a $\mathcal{J}$-ideal see [5, Theoerem 5.6]. In the following example, we see that we may find a non semi $\mathcal{P}$-ideal in a local ring. Consider the local ring $R = \mathbb{Z}_{\langle 2\rangle} = \{\frac{a}{b} : a, b \in \mathbb{Z}, 2 \nmid b\}$ and let $I = \langle 4\rangle_{\langle 2\rangle} = \{\frac{a}{b} : a \in \langle 4\rangle, 2 \nmid b\}$. R is a local ring but I is not a semi $\mathcal{P}$-ideal of R. For example, $\left(\frac{2}{3}\right)^2 \in I$ but $\frac{2}{3} \notin \mathcal{P}(R) = \{0\}$ and $\frac{2}{3} \notin I$.

Proposition 6.8. *(See [1, Proposition 3.1]) Let R and S be rings and $f : R \to S$ be a surjective ring-homomorphism. Then the following statements hold:*

1. *If I is a semi $\mathcal{P}$-ideal of R and $\ker(f) \subseteq I$, then $f(I)$ is a semi $\mathcal{P}$-ideal of S.*
2. *If J is a semi $\mathcal{P}$-ideal of S and $\ker(f) \subseteq \rho(R)$, then $f^{-1}(J)$ is a semi $\mathcal{P}$-ideal of R.*

Proof. This follows from Theorem 2.11 by taking ρ to be the prime radical. □

Corollary 6.9. *(see [1, Corollary 3.1]) Let R be a ring and let I, K be two ideals of R with $K \subseteq I$. Then the following hold.*

1. *If I is a semi $\mathcal{P}$-ideal of R, then I/K is a semi $\mathcal{P}$-ideal of R/K.*
2. *If I/K is a semi $\mathcal{P}$-ideal of R/K and $K \subseteq \rho(R)$, then I is a semi $\mathcal{P}$-ideal of R.*
3. *If I/K is a semi $\mathcal{P}$-ideal of R/K and K is a semi $\mathcal{P}$-ideal of R, then I is a semi $\mathcal{P}$-ideal of R.*

Proof. Follows from Corollary 2.12by taking ρ to be the prime radical. □

Proposition 6.10. *(see [1, Proposition 3.3]Let ρ be a special radical and let I and J be two semi ρ-ideals in a ring R. If $I + J$ is proper in R, then $I + J$ is a semi ρ-ideal of R.*

Proof. Follows from Proposition 2.13 by taking ρ to be the prime radical. □

Theorem 6.11. (see [1, Theorem 3.2]) Let R_1 and R_2 be two noncommutative rings. Then a proper ideal $I = I_1 \times I_2$ is a semi $\mathcal{P}$-ideal of R if and only if one of the following statements holds.

1. I is a semi prime-ideal of R..
2. I_1 is a semi $\mathcal{P}$-ideal of R_1 and $I_2 = \mathcal{P}(R_2)$.
3. I_2 is a semi $\mathcal{P}$-ideal of R_2 and $I_1 = \mathcal{P}(R_1)$.

Proof. Follows from Theorem 3.2 by taking ρ to be the prime radical. □

Theorem 6.12. (see [1, Theorem 3.3]Let $R_1, R_2, \ldots, R_n$ be rings and $R = R_1 \times R_2 \times \cdots \times R_n$, where $n \geqslant 2$. Then a proper ideal I of R is a semi $\mathcal{P}$-ideal if and only if one of the following statements is satisfied.

1. I is a semiprime ideal of R.
2. $I = I_1 \times I_2 \cdots \times I_n$, where I_k is a semi $\mathcal{P}$-ideal of R_k for some $k \in \{1, \ldots, n\}$ and $I_j = \mathcal{P}(R_j)$ for all $j \in \{1, \ldots, n\} \backslash \{k\}$.

Proposition 6.13. *Let I be a semi $\mathcal{P}$-ideal of R and N an $R-R$-bi-submodule of the $R-R$-bi-module M. Then*

1. *$I \boxplus N$ is a semi $\mathcal{P}$-ideal of $R \boxplus M$.*
2. *If $(\mathcal{P}(R)M : M) = \mathcal{P}(R)$ and N is a semi $\mathcal{P}$-submodule of M with $IM + MI \subseteq N$, then $I \boxplus N$ is a semi $\mathcal{P}$-ideal of $R \boxplus M$.*

Proof. Follows from Proposition 5.1 by taking ρ to be the prime radical. □

Proposition 6.14. *Let I be an ideal of R and N a proper $R-R$-bi-submodule of the $R-R$-bi-module M. If $I \boxplus N$ is a semi $\mathcal{P}$-ideal of $R \boxplus M$, then I is a semi $\mathcal{P}$-ideal of R and N is a semi ρ-submodule of M.*

Proof. Follows from Proposition 5.2 by taking ρ to be the prime radical. □

References

[1] E. Yetkin Celikel, H. A. Khashan, Semi n-ideals of commutative rings, Czechoslovak Mathematical Journal, **72**, 977–988(2022), Zbl: 7655775, https://doi.org/10.21136/CMJ.2022.0208-21.

[2] E. Yetkin Celikel, H. A. Khashan, Semi r-ideals of commutative rings, An. St. Univ. Ovidius Constantia, **31**(2), 101–126, 2023, DOI:10.2478/auom-2023-0022.

[3] J. Dauns, Prime modules, reine Angew. Math. 298 (1978), 156-181, Zbl:0365.16002, https://doi.org/10.1515/crll.1978.298.156.

[4] B. de la Rosa and S. Veldsman, A relationship between ring radicals and module radicals. Quaestiones Mathematicae. **17** (1994), 453-467, Zbl:0821.16023, https://doi.org/10.1080/16073606.1994.9631777.

[5] N. Groenewald, On radical ideals of non-commutative rings, Journal of Algebra and its Applications, 2350196, https://doi.org/10.1142/S0219498823501967.

[6] N.J. Groenewald and D. Ssevviiri, Completely prime submodules, International Electronic Journal of Algebra,**13**, 2013, 1-14, Zbl:1329.16005, https://doi.org/10.1155/2013/128064, Zbl: :1329.16005.

[7] O.A.S. Karamzadeh, On the Krull intersection theorem, Acta Mathematica Academiae Scientiarum Hungaricae 42(1):(1983) 139-141, Zbl:0526.16026 DOI-https://doi.org/10.1007/BF01960558.

[8] Hani A. Khashan and Amal B. Bani-Ata, J-ideals of commutative rings, International Electronic Journal of Algebra Volume **29** (2021) 148-164, Zbl:1467.13005, DOI: 10.24330/ieja.852139.

[9] T.Y. Lam, A First Course in Noncommutative Rings, second ed., Graduate Texts in Mathematics, vol. **131**, Springer-Verlag, New York, 2001.

[10] U. Tekir, S. Koc and K.H. Oral, n-Ideals of commutative rings, Filomat, **31**(10) (2017), 2933-2941, Zbl 1488.13016.

[11] S. Veldsman, A Note on the Radicals of Idealizations, *Southeast Asian Bulletin of Mathematics* **32**, (2008), 545-551, Zbl 1174.16006, https://doi.org/10.2217/thy.09.46.

[12] B.J. Gardner, R. Wiegandt. Radical Theory of Rings, Marcel Dekker Inc, New York, 2004.

[13] Zubayda M. Ibraheem, On local rings,Raf. J. of Comp. & Math's. , Vol. 11, No. 1, 2014, 93-97, DOI: 10.33899/CSMJ.2014.163734.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 36-44

ISSN: 2820-7114

Title :

Extension of star-operation

Author(s):

Elmakki Ahmed & Taha Eddhay

Moroccan Journal of Algebra and Geometry with Applications
Vol. 3(1) (2024), 36–44

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

Extension of star-operation

Elmakki Ahmed[1] and Taha Eddhay[2]
[1] Department of Mathematics, Faculty of Sciences, Monastir, Tunisia.
e-mail: *elmakkiahmed@gmail.com*
[2] Preparatory Institute for Engineering Studies, Gafsa, Tunisia.
e-mail: *taha.eddhay@gmail.com*

Communicated by Ahmed Hamed
(Received 19 June 2023, Revised 12 September 2023, Accepted 22 September 2023)

Abstract. Let D be an integral domain, $*$ a star operation on D and S a multiplicative subset of D. In this paper, we generalize the notion of $*$-ideals (resp, $*$-invertible) of D, by introducing the concept of S-$*$-ideals (resp, S-$*$-invertible) of D. A fractional ideal of D is called S-$*$-ideals (resp, S-$*$-invertible) if there exists an $s \in S$ such that $sI^* \subseteq I \subseteq I^*$ (resp, if there exists an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$). We investigate many proprieties and characterizations of the notion S-$*$-ideals (resp, S-$*$-invertible).

Key Words: $*$-operation, S-$*$-ideals, S-$*$-invertible.
2010 MSC: 13G05, 13A15.

1 Introduction

Throughout this paper D will be an integral domain with quotient field K. We denote by $\mathcal{F}(D)$, the set of nonzero fractional ideals of D. A $*$-operation on D is a mapping $I \longmapsto I^*$, from $\mathcal{F}(D)$ to $\mathcal{F}(D)$ which satisfies the following conditions for $a \in K\backslash\{0\}$ and $I, J \in \mathcal{F}(D)$:

1. $(a)^* = (a)$ and $(aI)^* = aI^*$,
2. $I \subseteq I^*$; if $I \subseteq J$, then $I^* \subseteq J^*$ and
3. $(I^*)^* = I^*$.

$I \in \mathcal{F}(D)$ is called a $*$-ideal if $I^* = I$. We use the notation $*$-$Max(D)$ for the set of $*$-ideals which are maximal among proper integral $*$-ideals of D. An element I of $\mathcal{F}(D)$ is called to be $*$-invertible if $(IJ)^* = D$ for some $J \in \mathcal{F}(D)$ or equivalently $(II^{-1})^* = D$, where $I^{-1} = \{x \in K \mid xI \subseteq D\}$. We can construct the $*$-operation $*_s$ defined by $I^{*_s} = \bigcup\{(I')^* \mid I' \in \mathcal{F}(D),\ I'$ is finitely generated and $I' \subseteq I\}$. We say $*_s$ that is the finite type $*$-operation induced by $*$. Also, $*$ is said to be of finite type if $* = *_s$ i.e., $I^* = I^{*_s}$ for each $I \in \mathcal{F}(D)$. For the general theory of $*$-operations, the reader is referred to [4, Sects. 32 and 34]. An important $*$-operation is the v-operation given by $I_v = (I^{-1})^{-1}$ for each $I \in \mathcal{F}(D)$. The finite type $*$-operation induced by the v-operation is called the t-operation. For $f = a_0 + \cdots + a_nX^n \in K[X]$, A_f will denote the D-submodule of K generated by $\{a_0, ..., a_n\}$. The set $N_* = \{f \in D[X] \mid (A_f)^* = D\}$ is a multiplicatively closed subset of $D[X]$ by [9, Proposition 2.1], and it is easy to see that, $N_* = N_{*_s}$.

In this paper, we generalize the notion of $*$-ideal (resp, $*$-invertible) by introducing the concept of S-$*$-ideal (resp, S-$*$-invertible). Let I be a fractional ideal of an integral domain D and S a multiplicative subset of D. We say that I is S-$*$-ideal if there exists an $s \in S$ such that $sI^* \subseteq I \subseteq I^*$. We say that I is S-$*$-invertible if there exists an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$, equivalently there exists an $s \in S$ such that $sD \subseteq (II^{-1})^* \subseteq D$ (Proposition 3.4).

In Section 2, we study basic results of S-$*$-ideal, we give an example of an S-$*$-ideal which is not

$*$-ideal. We also, show that every S-invertible ideal (recall from [6], that a fractional ideal I of D is said to be S-*invertible* if $sD \subseteq IJ \subseteq D$ for some $s \in S$ and some fractional ideal J of D) is S-$*$-ideal (Proposition 2.4). An ideal M of D disjoint with S is called S-$*$-maximal if it is maximal in the set of all integral proper S-$*$-ideals of D. We prove that every S-$*$-maximal ideal of D is a prime ideal of D (Proposition 2.8). Let D be an integral domain and S a multiplicative subset of D. We say that S is anti-Archimedean if $\cap_{n\geq 1} s^n D \cap S \neq \emptyset$ for every $s \in S$. In [2], the authors generalized this notion by introducing the concept of weakly anti-Archimedean multiplicative set. According [2], a multiplicative set S of an integral domain D is called weakly anti-Archimedean if for each family $(s_\alpha)_{\alpha\in\Lambda}$ of elements of S we have $(\cap_{\alpha\in\Lambda} s_\alpha D) \cap S \neq \emptyset$. Note that every weakly anti-Archimedean multiplicative set is anti-Archimedean. The converse is not true as was observed in [3, Example 2.7]. Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. We show that every integral proper S-$*$-ideal of D is included in an S-$*$-maximal ideal of D (Theorem 2.9). In the particular case when S consists of units of D, we get every integral proper $*$-ideal of D is included in a $*$-maximal ideal of D (Corollary 2.10). Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. We prove that for each S-$*$-ideal I of D, $I = \bigcap_{M\in S\text{-}*\text{-}Max(D)} ID_M$ (Theorem 2.12).

In section 3, we study basic propertis of S-$*$-invertible. It's easy to show that if S consists of units of D the notions $*$-invertible and S-$*$-invertible coincide. Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. Let I be a fractional ideal of D. We show that I is an S-$*$-invertible ideal of D if and only if I is S-$*$-finite and for each $M \in S$-$*$-Max(D), ID_M is a principal ideal of D_M (Theorem 3.8). In the particular case when S consists of units of D we recover the folloing known result, I is a $*$-invertible ideal of D if and only if I is of $*$-finite type and it is t-locally principal (Corollary 3.9). Let D be an integral domain and S a multiplicative subset of D. It is well-known that for each finitely generated fractional ideal I of D, $(I_S)^{-1} = (I^{-1})_S$. We extented this result to S-$*$-finite ideal of D. We show that if I is an S-$*$-finite ideal of D, then $(I_S)^{-1} = (I^{-1})_S$ (Proposition 3.10) where $*$ a finite type $*$-operation on D and I a fractional ideal of D.

2 Basic properties of S-$*$-ideals

Definition 2.1. Let D be an integral domain, S a multiplicative subset of D and $*$ a star-operation on D. A fractional ideal I of D is called S-$*$-ideal if there exists an $s \in S$ such that $sI^* \subseteq I \subseteq I^*$.

Example 2.2.

1. Every $*$-ideal is an S-$*$-ideal.

2. Let $D = \mathbb{Z}[X]$ and $I = 2\mathbb{Z} + X\mathbb{Z}[X]$. By [1, Lemma 2.1], it is easy to show that $I^{-1} = (\frac{1}{2}\mathbb{Z}) \cap \mathbb{Z} + X\mathbb{Z}[X]$; so $I_v = \mathbb{Z}[X]$ which implies that I is not a divisorial ideal of D. Now, let $S = \{2^n \mid n \in \mathbb{N} \cup \{0\}\}$. Then S is a multiplicative subset of D. Moreover,
$$2I_v = 2\mathbb{Z}[X] \subseteq I \subseteq \mathbb{Z}[X] = I_v.$$
Hence I is an S-v-ideal of D. This shows that the converse of (1) is not true in general.

3. Let D be an integral domain, S a multiplicative subset of D and $*$ a star-operation on D. If S consists of units of D, then the notions of S-$*$-ideals and $*$-ideals are coincide.

Let D be an integral domain and S a multiplicative subset of D. Recall from [8] that an ideal I of D is called S-*principal*, if $sI \subseteq J \subseteq I$ for some principal ideal J of D and some $s \in S$. The next proposition collects some properties of S-$*$-ideals of an integral domain D.

Proposition 2.3. *Let D be an integral domain, S a multiplicative subset of D and $*$ a star-operation on D.*

1. *Let $S \subseteq T$ be multiplicative subsets of D. If I is an S-$*$-ideal of D, then I is a T-$*$-ideal of D.*

2. *Let $\bar{S}$ be the saturation of S. Then I is an S-$*$-ideal of D if and only if I is an $\bar{S}$-$*$-ideal of D.*

3. *If I is S-principal, then I is an S-$*$-ideal of D.*

Proof. (1). Obvious.

(2). The "only if" part follows from (1). Now, assume that I is an $\bar{S}$-$*$-ideal of D. Then there exists an $s \in \bar{S}$ such that $sI^* \subseteq I \subseteq I^*$. Since $s \in \bar{S}$, there exists a $t \in S$ such that $t = ss'$ for some $s' \in D$. Thus

$$tI^* \subseteq sI^* \subseteq I \subseteq I^*,$$

and hence I is an S-$*$-ideal of D.

(3). Since I is S-principal, there exist an $s \in S$ and $d \in D$ such that $sI \subseteq dD \subseteq I$. This implies that

$$sI^* = (sI)^* \subseteq (dD)^* = dD \subseteq I \subseteq I^*.$$

Hence I an S-$*$-ideal of D. □

Recall from [6], that for a multiplicative set S in D, a fractional ideal I of D is said to be *S-invertible* if $sD \subseteq IJ \subseteq D$ for some $s \in S$ and some fractional ideal J of D. It is shown that I is an S-invertible ideal of D if and only if $sD \subseteq II^{-1} \subseteq D$ for some $s \in S$. It well known that every invertible ideal is a $*$-ideal. Our next Proposition generalize this result.

Proposition 2.4. *Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D. Each S-invertible ideal of D is S-$*$-ideal.*

Proof. Let I be an S-invertible ideal of D. By [6, Remark 2.4], $sJ^{-1} \subseteq I \subseteq J^{-1}$ for some $s \in S$ and some fractional ideal J of D. This implies that

$$sJ^{-1} = (sJ^{-1})^* \subseteq I^* \subseteq (J^{-1})^* = J^{-1}.$$

Thus $sI^* \subseteq sJ^{-1} \subseteq I$, and hence I is an S-$*$-ideal of D. □

Example 2.5. Let D be a Prüfer domain, $*$ a star-operation on D and S a multiplicative subset of D. Then each nonzero S-finite ideal of D is S-$*$-ideal. Indeed, let I be an S-finite ideal of D. Then there exist an $s \in S$ and a nonzero finitely generated ideal F of D such that $sI \subseteq F \subseteq I$. Thus $sF^{-1} \subseteq I^{-1}$. Since D is a Prüfer domain, $FF^{-1} = D$; so

$$sD = sFF^{-1} \subseteq FI^{-1} \subseteq II^{-1} \subseteq D$$

which implies that I is an S-invertible ideal of D. Hence by the previous Proposition, I is an S-$*$-ideal of D.

Let D be an integral domain and S a multiplicative subset of D. We say that S is *anti-Archimedean* if $\cap_{n\geq 1} s^n D \cap S \neq \emptyset$ for every $s \in S$. In [2], the authors generalized this notion by introducing the concept of weakly anti-Archimedean multiplicative set. According [2], a multiplicative set S of an integral domain D is called *weakly anti-Archimedean* if for each family $(s_\alpha)_{\alpha\in\Lambda}$ of elements of S we have $(\cap_{\alpha\in\Lambda} s_\alpha D) \cap S \neq \emptyset$. Note that every weakly anti-Archimedean multiplicative set is anti-Archimedean. The converse is not true as was observed in [3, Example 2.7].

Proposition 2.6. *Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. Let $(I_\alpha)_{\alpha\in\Lambda}$ be a totally ordered family of fractional ideals of D. If for each $\alpha \in \Lambda$, I_α is S-$*$-ideal, then $\cup_{\alpha\in\Lambda} I_\alpha$ is an S-$*$-ideal of D.*

Proof. For each $\alpha \in \Lambda$, there exists an $s_\alpha \in S$ such that $s_\alpha I_\alpha^* \subseteq I_\alpha$. Since S is weakly anti-Archimedean, $\cap_{\alpha\in\Lambda} s_\alpha D \cap S \neq \emptyset$. Let $t \in \cap_{\alpha\in\Lambda} s_\alpha D \cap S$. Note that for each $\alpha \in \Lambda$, $tI_\alpha^* \subseteq I_\alpha$. We show that $t(\cup_{\alpha\in\Lambda} I_\alpha)^* \subseteq \cup_{\alpha\in\Lambda} I_\alpha$. Let $x \in (\cup_{\alpha\in\Lambda} I_\alpha)^*$. Since $*$ is of finite character, there exists a finitely generated subideal J of $\cup_{\alpha\in\Lambda} I_\alpha$ such that $x \in J^*$. Since J is a finitely generated ideal of D, there exists a $\beta \in \Lambda$ such that $J \subseteq I_\beta$. We have $tx \in tJ^* \subseteq tI_\beta^* \subseteq I_\beta$; so $tx \in I_\beta$ for some $\beta \in \Lambda$ which implies that $t(\cup_{\alpha\in\Lambda} I_\alpha)^* \subseteq \cup_{\alpha\in\Lambda} I_\alpha$, and hence $\cup_{\alpha\in\Lambda} I_\alpha$ is an S-$*$-ideal of D. □

Notation 2.7. Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D. An ideal M of D disjoint with S is called S-$*$-maximal if it is maximal in the set of all integral proper S-$*$-ideal of D. We denote by S-$*$-$\mathrm{Max}(D)$ the set of all S-$*$-maximal ideals of D.

Proposition 2.8. *Every S-$*$-maximal ideal of D is a prime ideal of D.*

Proof. Let P be an S-$*$-maximal ideal of D. Assume that P is not prime, there exist $a, b \in D\backslash P$ such that $ab \in P$. Let $I = P + aD$ and $J = P + bD$. Since $P \subsetneq I \subseteq I^* \subseteq D$, by maximality of P in the set of all integral proper S-$*$-ideal of D, $I^* = D$. In the same way we can prove $J^* = D$. This implies that $(IJ)^* = (I^*J^*)^* = D$. But $IJ = P^2 + aP + bP + abP \subseteq P$; so $P^* = D$. Now, since P is an S-$*$-ideal of D, there exists an $s \in S$ such that $sP^* \subseteq P$ which implies that $sD \subseteq P$, a contradiction because $P \cap S = \emptyset$. Hence P is a prime ideal of D. □

Theorem 2.9. Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. Then every integral proper S-$*$-ideal of D is included in an S-$*$-maximal ideal of D.

Proof. Let $\mathcal{F}$ be the set of all integral proper S-$*$-ideals of D. Then $\mathcal{F} \neq \emptyset$, since $\mathcal{F}$ contain all integral proper S-principal ideals of D. Now, let $(I_\alpha)_{\alpha\in\Lambda}$ be a totally ordered family of elements of $\mathcal{F}$. By Proposition 2.6, $\cup_{\alpha\in\Lambda} I_\alpha$ is an element of $\mathcal{F}$; so we conclude by Zorn's Lemma our result. □

In the particular case when S consists of units of D, we regain the following well-known result.

Corollary 2.10. *Let D be an integral domain and $*$ a finite type $*$-operation on D. Then every integral proper $*$-ideal of D is included in a $*$-maximal ideal of D.*

Lemma 2.11. *Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D. Let $(I_k)_{1\leq k\leq n}$ be a finite family of fractional ideals of D such that $\cap_{1\leq k\leq n} I_k \neq (0)$. If for each $1 \leq k \leq n$, I_k is S-$*$-ideal, then $\cap_{1\leq k\leq n} I_k$ is an S-$*$-ideal of D.*

Proof. For each $1 \leq k \leq n$, there exists an $s_k \in S$ such that $s_k I_k^* \subseteq I_k$. Let $t = s_1 s_2 \cdots s_n$. Then $t \in S$ and for each $1 \leq k \leq n$, $tI_k^* \subseteq I_k$. For each $1 \leq m \leq n$, $t(\cap_{1\leq k\leq n} I_k)^* \subseteq tI_m^* \subseteq I_m$. This implies that $t(\cap_{1\leq k\leq n} I_k)^* \subseteq \cap_{1\leq k\leq n} I_k$, and hence $\cap_{1\leq k\leq n} I_k$ is an S-$*$-ideal of D. □

Theorem 2.12. Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. Then for each S-$*$-ideal I of D,

$$I = \bigcap_{M\in S\text{-}*\text{-}Max(D)} ID_M.$$

Proof. Let x be a nonzero element of $\bigcap_{M\in S\text{-}*\text{-}Max(D)} ID_M$. Then for each S-$*$-maximal ideal M of D, there exists an $s_M \in D\backslash M$ such that $s_M x \in I$. Let $J = D \cap (\frac{1}{x} I)$. Then $s_M \in J$ for each S-$*$-maximal ideal M of D. Moreover, Since I is an S-$*$-ideal of D, $\frac{1}{x} I$ is an S-$*$-ideal of D; so by Lemma 2.11, J is an S-$*$-ideal of D. Assume that $J \neq D$. Then J is an integral proper S-$*$-ideal of D; so by Theorem 2.9, there exists $M \in S\text{-}*\text{-}Max(D)$ such that $J \subseteq M$ which implies that $s_M \in J \subseteq M$, a contradiction. Thus $J = D$ which implies that $x \in I$. Hence $I \subseteq \bigcap_{M\in S\text{-}*\text{-}Max(D)} ID_M$. This completed the proof, since other inclusion is obvious. □

Corollary 2.13. *Let D be an integral domain, $*$ a finite type $*$-operation on D and I a $*$-ideal of D. Then*

$$I = \bigcap_{M \in *\text{-}Max(D)} ID_M.$$

Remark 2.14. Let I be an S-$*$-ideal of an integral domain D, where S is a multiplicative subset of D and $*$ a star-operation of finite character on D. Then there exits an $s \in S$ such that $sI^* \subseteq I$. But $I^* = \bigcap_{M \in *\text{-}Max(D)} I^* D_M$; so

$$s(\bigcap_{M \in *\text{-}Max(D)} ID_M) \subseteq s(\bigcap_{M \in *\text{-}Max(D)} I^* D_M) = sI^* \subseteq I \subseteq \bigcap_{M \in *\text{-}Max(D)} ID_M.$$

Hence there exists an $s \in S$ such that

$$s(\bigcap_{M \in *\text{-}Max(D)} ID_M) \subseteq I \subseteq \bigcap_{M \in *\text{-}Max(D)} ID_M.$$

3 S-$*$-invertible ideals

In this section we extended the notion of S-invertible using the $*$-operation and we generalize some classical results concerning the notion of $*$-invertibility. We begin this section by the following definition.

Definition 3.1. Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D. A fractional ideal I of D is called S-$*$-invertible if there exists an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$.

Example 3.2. Let $D = \mathbb{Z} + X\mathbb{Z}[i][X]$, $S = \{2^n \mid n \in \mathbb{N}\}$ and $I = 2\mathbb{Z} + (1+i)X\mathbb{Z}[i][X]$. Since $2 \in I$, then $2D \subseteq I.D \subseteq D$. Which implies that I is S-invertible. On the other part, by [1, Lemma 2.1], it is easy to show that $I^{-1} = \mathbb{Z} + X\frac{1-i}{2}\mathbb{Z}[i][X]$. Thus if $II^{-1} = D$, then $1 = P_1(0)Q_1(0) + \cdots + P_n(0)Q_n(0)$ for some $P_1, \ldots, P_n \in I$ and $Q_1, \ldots, Q_n \in I^{-1}$. But for $1 \leq j \leq n$, $P_j(0) \in 2\mathbb{Z}$ and $Q_j(0) \in \mathbb{Z}$; so $1 = 2m_1 + \cdots + 2m_n$, $m_j \in \mathbb{Z}$. A contradiction. Hence I is not invertible.

Remark 3.3. Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D.

1. Since $I^* \subseteq I_v$ for each fractional ideal I of D, every S-$*$-invertible ideal of D is S-v-invertible.

2. Note that for a fractional ideal I of D, we have I is S-$*$-invertible if and only if I^* is S-$*$-invertible. Indeed, I is S-$*$-invertible if and only if $sD \subseteq (IJ)^* = (I^*J)^* \subseteq D$ for some $s \in S$ and some fractional ideal J of D if and only if I^* is S-$*$-invertible.

3. Let I be a fractional S-$*$-invertible ideal of D, then there exist an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$. We have
$$sI^{-1} = (I^{-1}sD)^* \subseteq (I^{-1}(IJ)^*)^* = (I^{-1}(IJ))^* \subseteq J^*.$$
Moreover, since $IJ^* \subseteq (IJ)^* \subseteq D$, $J^* \subseteq I^{-1}$. Thus $sI^{-1} \subseteq J^* \subseteq I^{-1}$. Note that in the same way we can prove that $sJ^{-1} \subseteq I^* \subseteq J^{-1}$.

4. By [6, Proposition 2.7], every S-principal ideal of D is S-invertible. This implies that each S-principal ideal of D is S-$*$-invertible.

Proposition 3.4. *Let I be a fractional ideal of an integral domain D, S a multiplicative subset of D and $*$ a star-operation on D. Then I is S-$*$-invertible if and only of there exists an $s \in S$ such that $sD \subseteq (II^{-1})^* \subseteq D$. In particular, I^{-1} is also an S-$*$-invertible ideal of D.*

Proof. If I is S-$*$-invertible, then there exist an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$. But by Remark 3.3(3), $J^* \subseteq I^{-1}$; so $sD \subseteq (IJ)^* = (IJ^*)^* \subseteq (II^{-1})^* \subseteq D$. The other implication is obvious. □

Definition 3.5. Let D be an integral domain, S a multiplicative subset of D and $*$ a star-operation on D. A fractional ideal I of D is called of S-$*$-finite type if there exist an $s \in S$ and a fractional finitely generated ideal F of D such that $sI \subseteq F^* \subseteq I^*$.

Let D be an integral domain and S a multiplicative subset of D. According to [5], D is called an *S-Mori domain* if every increasing sequence of integral divisorial ideals of D is S-stationary (an increasing sequence $(I_k)_{k\in\mathbb{N}}$ of ideals of D is called *S-stationary* if there exist a positive integer n and an $s \in S$ such that for each $k \geq n$, $sI_k \subseteq I_n$ [8]). It was shown in [5], that if D is an S-Mori domain, then for each nonzero fractional ideal I of D, $sI \subseteq J_v \subseteq I_v$ for some $s \in S$ and some finitely generated fractional ideal J of D such that $J \subseteq I$. This implies that in an S-Mori domain every nonzero fractional ideal I of D is of S-v-finite type.

Remark 3.6. Let D be an integral domain, $*$ a star-operation on D and S a multiplicative subset of D. Let I be a fractional ideal of D of S-$*$-finite type. Then there exist an $s \in S$ and a fractional finitely generated ideal J of D such that $sI \subseteq J^* \subseteq I^*$. If the star-operation $*$ is of finite character, then we can suppose that $J \subseteq I$. Indeed, let $J = (a_1, ..., a_n)$, where $a_i \in I^*$. Then for each $1 \leq i \leq n$, there exist a finitely generated subideal J_i of I. Let $J' = J_1 + \cdots + J_n$. Then J' is a finitely generated subideal of I. Moreover, $J \subseteq J_1^* + \cdots + J_n^* \subseteq (J')^*$; so $sI \subseteq J^* \subseteq (J')^* \subseteq I^*$.

Let D be an integral domain and $*$ a star-operation on D. Let I and J be tow fractional ideals of D. It will known that if $*$ is of finite character, then

$$(IJ)^* = \cup\{(I'J')^* \mid I' \subseteq I, J' \subseteq J, \text{ two finitely generated fractional ideals of } D\}.$$

Our next Theorem prove a neccesary and sufficient condition for a fractional ideal to be S-$*$-invertible. This extended a result proved by Kang in [9]. To prove it we need the following Lemma.

Lemma 3.7. *Let D be an integral domain, $*$ a finite type $*$-operation on D and S a multiplicative subset of D. Every S-$*$-invertible ideal of D is an S-$*$-finite ideal of D.*

Proof. Let I be an S-$*$-invertible ideal of D. There exist an $s \in S$ and a fractional ideal J of D such that $sD \subseteq (IJ)^* \subseteq D$. Since $*$ is of finite character, there exist two finitely generated fractional ideals I' and J' of D such that $I' \subseteq I$, $J' \subseteq J$ and $s \in (I'J')^*$. This implies that $sD \subseteq (I'J')^* \subseteq D$. Now by Remark 3.3(3), $s(J')^{-1} \subseteq (I')^* \subseteq (J')^{-1}$ and $sJ^{-1} \subseteq I^* \subseteq J^{-1}$. Since $J' \subseteq J$, $J^{-1} \subseteq (J')^{-1}$; so

$$sI \subseteq sI^* \subseteq s(J')^{-1} \subseteq (I')^* \subseteq I^*.$$

Hence I is of S-$*$-finite type. □

Theorem 3.8. Let D be an integral domain, $*$ a finite type $*$-operation on D and S a weakly anti-Archimedean multiplicative subset of D. Let I be a fractional ideal of D. Then the following statements are equivalent.

1. I is an S-$*$-invertible ideal of D.

2. I is S-$*$-finite and for each $M \in S$-$*$-$\mathrm{Max}(D)$, ID_M is a principal ideal of D_M.

Proof. (1) $\Rightarrow$ (2) By Lemma 3.7, I is of S-$*$-finite type. Let M be an S-$*$-maximal ideal of D. We have $II^{-1} \nsubseteq M$, indeed, if $II^{-1} \subseteq M$, then $sD \subseteq (II^{-1})^* \subseteq M$ for some $s \in S$; so $s \in M$, a contradiction because $S \cap M = \emptyset$. This implies that $(ID_M)(I^{-1}D_M) = II^{-1}D_M = D_M$, and thus ID_M is an invertible ideal of D_M. Hence ID_M is principal since D_M is a local ring.

(2) $\Rightarrow$ (1) By hypothesis, there exist an $s \in S$ and a fractional finitely generated subideal J of I such that $sI \subseteq J^* \subseteq I^*$. Assume that I is not S-$*$-invertible. Then $(II^{-1})^* \subsetneq D$; so by Theorem 2.9, there exist an S-$*$-maximal ideal M of D such that $(II^{-1})^* \subseteq M$. By hypothesis, ID_M is principal, then $ID_M = aD_M$ for some $a \in I$. This implies that $\frac{1}{a}I \subseteq D_M$; so $\frac{1}{a}J \subseteq D_M$. Since J is finitely generated, there exists a $t \in D \backslash M$ such that $\frac{t}{a}J \subseteq D$. We have

$$\frac{st}{a}I \subseteq \frac{st}{a}I^* \subseteq \frac{t}{a}J^* \subseteq D.$$

Thus $\frac{st}{a} \in I^{-1}$ which implies that $st \in aI^{-1} \subseteq II^{-1} \subseteq M$. Since $t \notin M$, $s \in M$ because M is a prime ideal of D by Proposition 2.8. This contradict that $M \cap S = \emptyset$. Hence I is an S-$*$-invertible ideal of D. □

In the particular case when S consists of units of D we regain the following well-known result proved by B.G. Kang ([9]).

Corollary 3.9. *Let D be an integral domain, $*$ a finite type $*$-operation on D and I a fractional ideal of D. Then the following statements are equivalent.*

1. *I is a $*$-invertible ideal of D.*

2. *I is of $*$-finite type and it is t-locally principal.*

Let D be an integral domain and S a multiplicative subset of D. It is well-known that for each finitely generated fractional ideal I of D, $(I_S)^{-1} = (I^{-1})_S$. Our next Proposition improves this result.

Proposition 3.10. *Let S a multiplicative subset of an integral domain D, $*$ a finite type $*$-operation on D and I a fractional ideal of D. If I is an S-$*$-finite ideal of D, then $(I_S)^{-1} = (I^{-1})_S$.*

Proof. We have always that $(I^{-1})_S \subseteq (I_S)^{-1}$, so we must prove the converse in order to conclude. Since I is S-$*$-finite, there exist an $s \in S$ and a finitely generated ideal $J \subseteq I$ such that $sI \subseteq J^* \subseteq I^*$. Thus $J^{-1} \subseteq \frac{1}{s}I^{-1}$, and consequently $(J^{-1})_S \subseteq (I^{-1})_S$. Since J is finitely generated, $(J^{-1})_S = (J_S)^{-1}$. Moreover, $J_S \subseteq I_S$. Thus $(I_S)^{-1} \subseteq (J_S)^{-1} = (J^{-1})_S \subseteq (I^{-1})_S$, and hence $(I^{-1})_S = (I_S)^{-1}$. □

Next, we give a relation between S-t-invertible ideals of D and t-invertible ideals of the localization D_S, where t- is the t-operation.

Proposition 3.11. *Let S a multiplicative subset of an integral domain D and I a fractional ideal of D.*

1. *If I is an S-t-invertible ideal of D, then I_S is a t-invertible ideal of D_S.*

2. *Assume that for each t-finite type ideal J of D, $(J_S)_t \cap D = J_t : s$ for some $s \in S$. Then I is S-t-invertible if and only if I_S is t-invertible and I is an S-$*$-finite ideal of D.*

Proof. (1). Since I is S-t-invertible, $sD \subseteq (II^{-1})_t \subseteq D$ for some $s \in S$. This implies that $D_S = ((II^{-1})_t)_S$. But $((II^{-1})_t)_S \subseteq ((II^{-1})_S)_t$; so $D_S = ((II^{-1})_S)_t$ because $((II^{-1})_S)_t \subseteq D_S$. Thus $D_S = (I_S(I^{-1})_S)_t$, and hence I_S is a t-invertible ideal of D_S.

(2). The "only if" part follows from (1) and Lemma 3.7, since t is a finite type $*$-operation. For the "if" part, let $s \in S$ and J a finitely generated subideal of I such that $sI \subseteq J_t \subseteq I_t$. This implies that

$(I_t)_S = (J_t)_S$. First we show that J_S is t-invertible. Since I_S is t-invertible, $D_S = (I_S(I^{-1})_S)_t$. Thus

$$\begin{array}{rcl} D_S & = & (I_S(I^{-1})_S)_t \\ & \subseteq & ((I_t)_S(I^{-1})_S)_t \\ & \subseteq & ((J_t)_S(J^{-1})_S)_t \\ & = & ((J_tJ^{-1})_S)_t \\ & \subseteq & ((JJ^{-1})_S)_t \\ & \subseteq & D_S. \end{array}$$

This implies that $((J_S(J^{-1})_S))_t = ((JJ^{-1})_S)_t = D_S$, hence J_S is t-invertible. Now, since J_S is t-invertible, $(J_S)^{-1}$ is of t-finite type; so there exists a finitely generated subideal F of J^{-1} such that $(J^{-1})_S = (J_S)^{-1} = (F_S)_t$. Thus $D_S = ((JJ^{-1})_S)_t = ((FJ)_S)_t$; so $D = ((FJ)_S)_t \cap D$. By hypothesis, $D = (FJ)_t : s'$ for some $s' \in S$, which implies that $s'D \subseteq (FJ)_t$. But $F \subseteq J^{-1} \subseteq \frac{1}{s}I^{-1}$ and $J \subseteq I$, thus $ss'D \subseteq (sFJ)_t \subseteq (II^{-1})_t \subseteq D$, and hence I is an S-t-invertible ideal of D. □

Proposition 3.12. *Let I be a non zero ideal of an integral domain D. Let T be a multiplicatively closed subset of D and S be a multiplicative subset of D.*

1. *If I is an S-t-ideal of D, then $I_T \bigcap D$ is an S-t-ideal of D.*

2. *If I_T is an S-t-ideal of D_T, then $I_T \bigcap D$ is an S-t-ideal of D.*

Proof. 1. Let I be a S-t-ideal of D. Then $sI_t \subseteq I$ for some $s \in S$. We show that $s(I_T \bigcap D)_t \subseteq I_T \bigcap D$. Let $\alpha \in (I_T \bigcap D)_t$, thus there exists a finitely generated fractional ideal F of D contained in $(I_T \bigcap D)$ such that $\alpha \in F_v$. Since $F \subseteq F_T \subseteq I_T$, then $s\alpha \in s(I_T)_t$ and there exists an $r \in T$ such that $rF \subseteq I$. Then $r\alpha \in rF_v = (rF)_v \subseteq I_t \subseteq \frac{1}{s}I$. Hence $sr\alpha \subseteq I$, so $s\alpha \subseteq I_T$, then $s\alpha \subseteq I_T \bigcap D$. Therefore $s(I_T \bigcap D)_t \subseteq I_T \bigcap D$.

2. Let I_T be an S-t-ideal of D_T. Then $s(I_T)_t \subseteq I_T$ for some $s \in S$. We show that $s(I_T \bigcap D)_t \subseteq I_T \bigcap D$. Let $\alpha \in (I_T \bigcap D)_t$, thus there exists a finitely generated fractional ideal J of D contained in $(I_T \bigcap D)$ such that $\alpha \in J_v$. Since $J \subseteq J_T \subseteq I_T$, then $s\alpha \in s(I_T)_t$. Hence $s\alpha \in s(I_T)_t \bigcap D \subseteq I_T \bigcap D$. Therefore $s(I_T \bigcap D)_t \subseteq I_T \bigcap D$.

□

Let D be an integral domain with quotient field K. Let $*$ be a star operation on D. Let $f = a_0 + \cdots + a_nX^n \in K[X]$, A_f will denote the D-submodule of K generated by $\{a_0, ..., a_n\}$. The set $N_* = \{f \in D[X] \mid (A_f)^* = D\}$ is a multiplicatively closed subset of $D[X]$. We defined the ring $D[X]_{N_*}$ by $D[X]_{N_*} = \{\frac{f}{g} \mid f \in D[X], g \in N_*\}$.

Proposition 3.13. *Let $*$ be a $*$-operation on an integral domain D with quotient field K, S be a multiplicative subset of D. Let I be an ideal of D. Then :*

1. *If I is S-$*$-ideal, then there exist $s \in S$ such that $s(ID[X]_{N_*} \bigcap K) \subseteq I$.*

2. *If I is an S-v-ideal (resp., S-t-ideal) of D, then $I[X]_{N_v}$ is an S-v-ideal (resp., S-t-ideal) of $D[X]_{N_v}$.*

Proof. 1. Let I be S-$*$-ideal. Then $sI^* \subseteq I$, for some $s \in S$. We show that $s(ID[X]_{N_*} \bigcap K) \subseteq I$. Let $a \in (ID[X]_{N_*} \bigcap K)$. Then $ag = f$ for some $g \in N_*$ and $f \in I[X]$. Hence $(a) = (aA_g)^* = (A_{ag})^* = (A_f)^* \subseteq I^* \subseteq \frac{1}{s}I$. So $sa \in I$. Therefore $s(ID[X]_{N_*} \bigcap K) \subseteq I$.

2. Suppose that I is a S-v-ideal, then $sI_v \subseteq I$, for some $s \in S$. Then $s(I[X]_{N_v})_v = sI_v[X]_{N_v}$ by [9, Proposition 2.2]. Hence $s(I[X]_{N_v})_v \subseteq I[X]_{N_v}$. Therefore $I[X]_{N_v}$ is a S-v-ideal of $D[X]_{N_v}$. In the some way we can show that $I[X]_{N_v}$ is an S-t-ideal of $D[X]_{N_v}$.

□

Acknowledgments. The authors would like to thank the reviewer for many helpful comments and suggestions which improved the paper in a significant way.

References

[1] D.F. Anderson, S. E. Baghdadi and S. E. Kabbaj, On the class group of $A + XB[X]$ domains, Advances in Commutaive Ring Theory, Lecture Notes in Pure and Appl. Math. Marcel Dekker 205 (1999) 73 – 85.

[2] M. Achraf and A. Hamed, S-prime ideals of a commutative ring, Beitr*ä*ge zur Algebra und Geometrie, 61 (2020), 533 – 542.

[3] D.E. Dobbs, Ahmes Expansions of Formal Laurent Series and a Class of Nonarchimedean Integral Domains, J. Algebra, **103** (1986), 193-201.

[4] R. Gilmer, Multiplicative Ideal theory, Maecel Dekker, New York, (1972).

[5] A. Hamed, On S-Mori domains, J. Algebra Appl., 17 (09), (2018)1850171.

[6] A. Hamed, The local S-class group of an integral domain, Rocky Mountain J. Math., 48(5) (2018), 1585 – 1605.

[7] A. Hamed, A new characterization of GCD domains of formal power series, St. Petersbg. Math. J., 33 (2022), 879 – 889.

[8] A. Hamed and S. Hizem, Modules satisfying the S-Noetherian property and S-ACCR, Comm. Algebra, 44 (2016), 1941 – 1951.

[9] B.G. Kang, Prüfer v-multiplication domains and the ring $R[X]_{N_v}$, J. Algebra 123 (1989), 151 – 170.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 45-77

ISSN: 2820-7114

Title :

Reflecting on parabolas

Author(s):

David E. Dobbs

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 45–77

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco ISSN: 2820-7114

Reflecting on parabolas

David E. Dobbs
Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1320
e-mail: *ddobbs1@utk.edu*

Communicated by Najib Mahdou
(Received 15 July 2023, Revised 04 October 2023, Accepted 08 October 2023)

Abstract. Two known results are proven in this teaching note. First, a proof of the reflection property of parabolas is given; that proof would be accessible early in a calculus class or in a course that combines precalculus with an introduction to differential calculus. Second, that reflection property plays a key role in a characterization of parabolas; that proof solves an initial value problem concerning a first order ordinary differential equation, and so it would be accessible early in a course on differential equations or in some courses on integral calculus. A closing remark discusses characterization results and classification results.

Key Words: Euclidean analytic geometry, parabola, focus, directrix, tangential half-line, vector, dot product, inverse cosine function, derivative, initial value problem.

2010 MSC: Primary 51-02; Secondary 51N20, 33B10, 97G70, 26A06, 34-01.

1 Introduction

Readers who wish to avoid or defer the reading of pedagogic comments may proceed at once to Section 2 (where the reflection property of a parabola is proven) and to Section 3 (where this reflection property is used to characterize parabolas and arcs thereof).

Conic sections (that is, parabolas, ellipses and hyperbolas) have traditionally been studied for several reasons at the high school level. Some of those reasons are algebraic, some are geometric, and some are related to scientific applications. Indeed, those reasons include the following facts: conics (along with their degenerate cases) are the only possible graphs (in Euclidean plane analytic geometry) of equations of the form $f(x,y) = 0$, where f is a second-degree polynomial expression in x and y; conics (along with their degenerate cases) are the only possible intersections (in three-dimensional Euclidean geometry) of a plane with a double-napped right circular cone; relative to a given point F that is not on a given line L, each of the three basic types of conics is characterized as the set $\mathcal{S}$ of points P such that the associated “eccentricity" e (that is, the ratio of the distance between P and F to the distance between P and L) has a specific constant value, with $e = 1$ (resp., $e < 1$; resp., $e > 1$) corresponding to $\mathcal{S}$ being a parabola (resp., an ellipse; resp., a hyperbola) with F being a “focus" of $\mathcal{S}$ and L being the corresponding “directrix" of $\mathcal{S}$; and each of the three basic types of conics has a reflection property with a number of physical applications. Despite all these reasons for the study of conics to play central roles in the mathematics and science curricula in high school, the topic of conic sections has received much less coverage in recent years. Indeed, many students now leave high school with the *mistaken* impression that parabolas are *defined* as the graphs (in Euclidean plane analytic geometry) of equations of the form $y = a(x-h)^2 + k$ for suitable $a, h, k \in \mathbb{R}$ with $a \neq 0$; also with the *mistaken* impression that ellipses are *defined* as the graphs (in Euclidean plane analytic geometry) of equations of the form $(x-h)^2/a^2 + (y-k)^2/b^2 = 1$ for suitable $a, b, h, k \in \mathbb{R}$ with $a \neq 0$ and $b \neq 0$; also with similar *mistaken* impressions as to the variety of the possible equations whose graphs (in Euclidean plane analytic geometry) can be hyperbolas; and often not having learned the reflection property of *any* of the three basic types of conics. In a short article, one could not hope to suggest ways to address or to redress all of these (what I consider to be) poor pedagogic decisions by the planners of some high school curricula. Accordingly, we will focus (pun intended) here on parabolas, as they are arguably the simplest kind of conic. So, **the purpose of this article** is twofold: using the "$e = 1$" definition of a parabola, to prove that any parabola does have a certain reflection property; and to show that the just-mentioned reflection property actually characterizes parabolas (among the family of graphs of sufficiently well-behaved functions in the Euclidean plane).

Neither of the two results mentioned in the above description of the purpose of this article is new. However, our exposition here is intended to be more accessible and simpler than some other presentations of these two results, while also indicating a way to reintroduce the currently under-emphasized topics of rotation and translation of coordinate axes. The next two paragraphs will provide some details supporting the claims that were made in the preceding sentence.

In Section 2, we state the appropriate reflection property and then prove that any parabola satisfies that property. For each of the three basic types of conic, the appropriate refection property involves the notion of a tangential half-line to a graph. That, in turn, involves the notion of a derivative; *that*, in turn, requires familiarity with the notion of a limit and experience in calculating the limits of some difference quotients. While that sort of familiarity and experience is traditionally gained early in a first course on calculus, the topic of the average rate of change for a non-linear function over an interval in its domain has recently been introduced in many high schools at or below the level of a precalculus course. Moreover, several colleges now offer a course that combines precalculus and calculus. (Such a course may run for three semesters.) Thus, the time seems ripe to offer a proof that parabolas satisfy the appropriate refection property in a way that assumes only the ability to differentiate the functions given by $f(x) = x^2$ and $g(x) = x^{1/2}$ (which are always among the examples of non-linear functions that one is first taught to differentiate by "using the definition of a derivative"). The proof in Theorem 2.1 assumes only that much of a prerequisite from calculus-related material. The figures supporting that proof show horizontal rays that either (i) arrive at a parabola from "within" and then get directed toward the parabola's focus or (ii) are the result of rays that are emitted from that focus and then redirected ("inward") after meeting the parabola. The proof begins with the geometric observation that rotation and translation of axes allows us to assume that the directrix is a vertical line, the focus is on the x-axis, and the origin is on the parabola. (Some instructors may wish to spend extra time to provide the "change of variable" descriptions that pertain to rotations of axes, as in [5, page 480], and possibly follow that up with guidelines, as in [5, page 482], for graphing an equation $f(x,y) = 0$ where f is *any* second-degree polynomial expression in x and y.) Thus, the proof reduces to considering the (parabolic) graph of the equation $y^2 = 4ax$ for some nonzero real number a. Most instructors would probably not expect students at or below the level of first-year calculus to be comfortable with a situation where "x is a function of y." (On the other hand, honors calculus students would understand that $x = y^2/(4a)$, so that $\frac{dx}{dy} = \frac{y}{2a}$, and such students could be led to an intuitive version of the Inverse Function Theorem which leads to the slope of the tangent line being

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}} = \frac{2a}{y}$$

when $y \neq 0$, that is, except at the origin.) So, most instructors would probably prefer to consider the top and bottom of the parabola separately. For instance if $a > 0$, the top (resp., the bottom) of the parabola is the graph of the function given by $f_1(x) = 2\sqrt{a}x^{1/2}$ (resp., by $f_2(x) = -f_1(x)$) with domain $[0, \infty)$. With the slope of the tangent line to the parabola at a given point in hand (except when $x = 0$, that is, except at the origin), the proof then turns to a formula to measure the angles (regardless of whether acute, obtuse or right) that are formed at the intersection point of two distinct nonparallel lines in the Euclidean plane. This formula involves standard background material on bound vectors in the Euclidean plane and their dot product; for the sake of completeness, that background is recalled early in Section 2. (Instructors who would prefer to avoid the use of vectors in their classes are invited to replace the just-mentioned approach by using instead a slope-heavy approach which detects right angles via the usual "$m_1 m_2 = -1$" criterion and measures acute or obtuse angles at an intersection point of two non-vertical non-perpendicular lines in the Euclidean plane by formulas in [4, Theorem 2.2]. While our vectorial approach will require use of the inverse cosine function, note that the alternative use of the just-mentioned formulas in [4] would instead require use of the inverse tangent function. Although the alternative method that has just been sketched would perhaps seem slightly more cumbersome (involving more case analyses) than the vectorial method which is used in the proofs in Sections 2 and 3, the alternative method would, because of its appeal to the formulas in [4, Theorem 2.2], reflect (another intended pun) my long-standing interest in using the tangent function (and, as needed, the inverse tangent function) extensively in the curriculum.) Here are two more pedagogic notes about Section 2. First, if an instructor is willing to ask his/her class to differentiate the squaring function but does not want those students to differentiate the square root function or to grapple with the situation where "x is a function of y", then he/she can tweak our presentation as follows: by rotating and translating axes appropriately, reduce to the situation where the directrix is a horizontal line, the focus is on the y-axis, and the origin is on the parabola; infer that the parabola in question is the graph of the equation $x^2 = 4ay$ for some nonzero real number a, that is, the graph of the function given by $y = h(x) = x^2/(4a)$; in the spirit of Section 2, draw a graph of the function h, showing rays that are vertical (instead of the horizontal rays in the figures in Section 2); without having to consider the top and the bottom of the parabola separately, show that the slope of the tangent line to the parabola is $\frac{dy}{dx} = \frac{x}{2a}$; and, finally, adapt the proof in Theorem 2.1 to finish the proof. Second, for the details of a shorter, but somewhat less accessible, proof of Theorem 2.1 which treats all points of the parabola (except its vertex) at once, albeit at the cost of viewing x as a differentiable function of y (except at the origin), see Remark 2.2 (c).

Section 3 uses calculus and differential equations to prove that the above-mentioned reflection property characterizes parabolas (among certain graphs of differentiable functions): see Theorem 3.1 and Remark 3.2 (b)-(c). As noted in the Abstract, this is not a new result, although we do hope that the reader will find our approach to it to be especially accessible. For more general (and higher-dimensional) results, two papers of D. Drucker deserve to be mentioned. In [6], Drucker gives a unified treatment of the reflection properties for all the types of conic sections (and analogues in three dimensions). That treatment is probably too complicated for beginning students to appreciate, as many of its proofs feature, *inter alia*, points at infinity, considerations of limiting positions, the calculus of polar coordinates, ... In the spirit of a unified treatment, [6] considers a parabola as a kind of limit of a "two foci, two directrices" situation. For an ellipse or a hyperbola, that kind of situation would be appropriate, even without having to mention a kind of limit. However, to view a parabola from that standpoint, Drucker resorts to "allowing [one of the foci] to be a point at infinity" [6, page 326]. Since [6] was not especially intended as a pedagogic paper, the above comments are not intended as negative criticism. I can say only the following about [7]: although I have not been able to get a copy of [7], I can report that the Math Review of [7] (available on MathSciNet of the American Mathematical Society) reported that [7] contains the same two main theorems as in [6] with the same proofs. I will close this paragraph with two final comments about Section 3. Firstly, as one can see from the references mentioned in [6], the literature has a number of other proofs of the characterization result for parabolas, and the interested instructor is invited to compare those with the presentation in Section 3 before choosing which (if any) approach should be shown or assigned to his/her class. Lastly, Section 3 ends with a remark which recalls some glorious characterization results from four basic areas of mathematics, including an instance where the proof of a characterization result led to a change in the definition of a fundamental concept.

No doubt, one could also fashion alternative proofs of the main result in Section 2 by using parametric methods. The interested reader is invited to do so. An overview of the subject would be incomplete without mentioning applications of the reflection property of a parabola. This property has been applied in building various useful items, such as certain headlights and cable television dishes, in the shape of paraboloids of revolution. For some related worked examples and homework exercises, see [5, pages 450-452; Exercises 51, 52, 53, 55, pages 453-454; and Exercise 53, page 489]. Readers are also encouraged to find proofs of the reflection properties of an ellipse or a hyperbola in various calculus textbooks (either as worked examples or as homework exercises, occasionally with hints).

To close the Introduction, we would like to warmly thank Dr. Michael Saum for providing, at our request, the LaTeX keystroke instructions that converted our freehand drawings into the figures that appear in this paper.

2 The reflection property of a parabola

As a student in a “without calculus" physics course during my first year at university in 1960, I learned the fundamental principle of “geometric optics" (that is, optics simplified by supposing that light moves in straight lines) which states that “the angle of incidence equals the angle of reflection" (assuming also that the medium that light had been initially traveling in is essentially the same as the medium into which the light has been reflected). That wording reflects usage that is reminiscent of Euclid’s *Elements*. Nowadays, a more typical statement of that principle would be “the angle of incidence is congruent to the angle of reflection" or “the radian measure of the angle of incidence equals the radian measure of the angle of reflection." The following question arises naturally: how does one define the angle of incidence and the angle of reflection? Anticipating this question, my physics textbook produced a diagram, augmented only with the statement that these angles were to be measured “from the normal." Since this course was given in a before-calculus environment, the diagram showed rays impinging on a (straight) linear “surface" and the “normal" was labeled as *the* line (in the plane of the page) that is perpendicular to that linear surface. Unfortunately, unless two non-parallel lines in the same plane are perpendicular, they meet in a way that creates four non-right angles, which are easily seen to break naturally into two pairs of congruent angles (by using the principle that vertically opposite angles are congruent). So, perhaps a clearer formulation of the physical principle would state that an acute (resp., obtuse) angle of incidence is congruent to an acute (resp., obtuse) angle of reflection. In any case, mathematicians have decided to measure angles of incidence or reflection “from the tangent line" instead of “from the normal." To avoid ambiguities arising from the just-mentioned pairs of congruent angles when two non-perpendicular coplanar lines intersect, we will broach such topics in terms of “tangential half-lines" and “tangential vectors". Before introducing these concepts, we will review some elementary material about vectors that will likely be familiar to most readers, possibly from high school courses on precalculus or physics. (For a more comprehensive introduction to vectors at the precalculus level, see [5, pages 403-428].) This material will be used to unambiguously define (and give the radian measure of) the angle between two nonparallel nonzero vectors in the same plane, which in turn will lead to a precise statement of a “Principle of reflection". That vectorial approach (and that principle) will be featured in the proofs of all the main results in Sections 2 and 3. As mentioned in the Introduction, non-vectorial formulas are also available to measure the angles formed when two nonparallel coplanar lines intersect (cf [4]). If a reader/instructor wishes to avoid vectorial methods, he/she is invited to adapt the proofs given below by using the angle-measuring formulas of his/her choice.

Let us work in a fixed Euclidean plane. If P and Q are (possibly equal) points (in that plane), the *bound vector* $\overrightarrow{PQ}$, with *initial point* P and *terminal point* Q, is the directed line segment going from P

to Q. (Context clues can usually help the reader to determine whether the overworked symbol $\overrightarrow{PQ}$ is referring to a bound vector or a ray.) Bound vectors are used in applications to represent physical quantities that have "magnitude" and "direction". The *magnitude* (also known as the *length*) of a bound vector $\mathbf{v} = \overrightarrow{PQ}$ is denoted by $|\mathbf{v}| = |\overrightarrow{PQ}|$ and is defined as the distance from P to Q (as calculated using the standard distance formula from analytic geometry). While every bound vector has a length, only a *nonzero bound vector* (that is, $\overrightarrow{PQ}$ such that $P \neq Q$) has a direction. (At the precalculus level, the notion of "direction" is treated as being intuitive. For various rigorous approaches to vectors (both "bound" and "free") and some kindred concepts in contexts that are much more general than Euclidean spaces, see [3] and the bibliography of that thesis.) The bound vector with initial point $(0,0)$ and terminal point $(1,0)$ is denoted by $\mathbf{i}$; the bound vector with initial point $(0,0)$ and terminal point $(0,1)$ is denoted by $\mathbf{j}$. We say that two (possibly equal) bound vectors are equivalent (some would say "equal") if these bound vectors have the same length and the same direction. A *zero vector* (that is, a bound vector of the form $\overrightarrow{PP}$ for some point P) is equivalent to itself and to any other zero vector, but is not equivalent to any nonzero bound vector. The relation of equivalence on the set of bound vectors is an equivalence relation. An equivalence class resulting from this equivalence relation is classically known as a *free vector*. It is almost always the case that, in practice, one blurs the distinction between a bound vector and a free vector, calling each simply a "vector".

It is commonly observed that physical quantities with magnitude and direction (such as winds, for instance) that act at the same point can be added according to what scientists call "the parallelogram law of vector addition". It is also commonly observed that it is useful to have a "multiplication" whereby a real number r multiplies a bound vector $\mathbf{v} = \overrightarrow{PQ}$ to give the bound vector $r\mathbf{v} = \overrightarrow{PW}$ such that $|r\mathbf{v}| = |r| \cdot |\mathbf{v}|$ and the point W is chosen on the line determined by the points P and Q so that if $r > 0$ (resp., if $r < 0$), then W is on the ray $\overrightarrow{PQ}$ (resp., then W is on the ray $\overrightarrow{QP}$). As degenerate cases, we agree that $0 \cdot \overrightarrow{PQ} = \overrightarrow{PP}$ for all points P and Q, and that $r \cdot \overrightarrow{PP} = \overrightarrow{PP}$ for all $r \in \mathbb{R}$ and all points P. In short, $r\mathbf{v}$ is a zero vector if and only if either $r = 0$ or $\mathbf{v}$ is a zero vector (or both); and if $r\mathbf{v}$ is not a zero vector, then its length is $|r|$ times the length of $\mathbf{v}$, and its direction is the same as (resp., the opposite of) the direction of $\mathbf{v}$ if $r > 0$ (resp., if $r < 0$).

It can be rigorously proven by geometric reasoning (see, for instance, [3] and its bibliography) that if $P_k(x_k, y_k)$ are points (for $1 \leq k \leq 4$) and $r \in \mathbb{R}$, then: the bound vectors $\overrightarrow{P_1P_2}$ and $\overrightarrow{P_3P_4}$ are equivalent (again, some would say and write "equal") if and only if $x_4 - x_3 = x_2 - x_1$ and $y_4 - y_3 = y_2 - y_1$; and the bound vectors $r\overrightarrow{P_1P_2}$ and $\overrightarrow{P_3P_4}$ are equivalent (again, some would say and write "equal") if and only if $x_4 - x_3 = r(x_2 - x_1)$ and $y_4 - y_3 = r(y_2 - y_1)$. Blurring the distinction between "free" and "bound", we infer that one can describe equivalence/equality of vectors, magnitude of vectors, vector addition and scalar multiplication in terms of "components" as follows (for P_1, P_2 as above and for $a, b, c, d, r \in \mathbb{R}$):

$$\overrightarrow{P_1P_2} = (x_2 - x_1)\mathbf{i} + (y_2 - y_1)\mathbf{j}, \ |\overrightarrow{P_1P_2}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$

$$(a\mathbf{i} + b\mathbf{j}) + (c\mathbf{i} + d\mathbf{j}) = (a + c)\mathbf{i} + (b + d)\mathbf{j}, \text{ and } r(a\mathbf{i} + b\mathbf{j}) = ra\mathbf{i} + rb\mathbf{j}.$$

Note that $|a\mathbf{i} + b\mathbf{j}| = \sqrt{a^2 + b^2}$, by applying the first and second displayed facts. Several other useful properties of vector addition and scalar multiplication follow easily from the just-displayed properties. These include the expected behavior of the (free) vector $\mathbf{0} = 0\mathbf{i} + 0\mathbf{j}$ (which is equal to (or more precisely, is represented by) any zero bound vector). One property of scalar multiplication that will be used often below is that $r(s\mathbf{v}) = (rs)\mathbf{v}$ for all $r, s \in \mathbb{R}$ and all vectors $\mathbf{v}$. If t is a nonzero real number and $\mathbf{v}$ is a vector, it will be convenient to write $\mathbf{v}/t$ instead of $(1/t)\mathbf{v}$, and it will be convenient to often use the resulting fact that

$$r(\frac{\mathbf{v}}{t}) = (\frac{r}{t})\mathbf{v},$$

for all $r, t \in \mathbb{R}$ with $t \neq 0$ and for all vectors $\mathbf{v}$. Also, as one may expect, vectors form an abelian group under addition, with neutral element $\mathbf{0}$ and with additive inverses given by

$$-(a\mathbf{i} + b\mathbf{j}) = (-a)\mathbf{i} + (-b)\mathbf{j}.$$

We come now to a very useful product of vectors. It is commonly called the *dot product*, but it is occasionally called the *scalar product* (not to be confused with the above "scalar multiplication"!) or the *inner product*. This is defined (for $a, b, c, d \in \mathbb{R}$) as follows:

$$(a\mathbf{i} + b\mathbf{j}) \cdot (c\mathbf{i} + d\mathbf{j}) := ac + bd.$$

Notice that the dot product of two vectors is a scalar (that is, a real number). It follows from the Law of Cosines (cf. [5, page 418]) that if θ is the radian measure of "the angle between two nonzero nonparallel vectors" $\mathbf{v}$ and $\mathbf{w}$, then

$$\cos(\theta) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|}.$$

It is of some interest to note that the just-displayed fact can be used to prove the Law of Cosines. Rather than discuss *that* further, let us address the "elephant in the room": what is meant by "the angle between two nonzero nonparallel vectors"? *That* is an excellent question. Answering it will dispel the ambiguity that I mentioned earlier is often present in the literature in typical statements of a principle of reflection. That answer will be given in the next paragraph, after which we will give a pair of short definitions, then give a precise statement of the principle of reflection, and then prove the reflection property of parabolas.

Let $\mathbf{v}$ and $\mathbf{w}$ be two nonzero nonparallel (hence, distinct) vectors. There is no harm (and great benefit) in viewing these as bound vectors with the same initial point, say with $\mathbf{v} = \overrightarrow{PQ}$ and $\mathbf{w} = \overrightarrow{PR}$ for some points Q and R which are distinct from P (and from each other) such that the line L_1 passing through P and Q is not parallel to (and hence is distinct from) the line L_2 passing through P and R. If L_1 and L_2 are perpendicular, then *each* of the four angles that are formed by L_1 and L_2 and that have vertex P is a right angle, necessarily with radian measure $\pi/2$, and this situation of perpendicularity is characterized by the condition $\mathbf{v} \cdot \mathbf{w} = 0$. So, for our needs here, we can assume henceforth that L_1 and L_2 are not perpendicular. Then the four angles that were mentioned above can be organized as a pair of (vertically opposite) congruent acute angles and a pair of (vertically opposite) congruent obtuse angles. Since $\overrightarrow{PQ}$ and $\overrightarrow{PR}$ are *directed* line segments, it is absolutely obvious that exactly one of these four angles deserves to be called "the angle between $\mathbf{v}$ and $\mathbf{w}$". Since we are dealing with angles that are either acute or obtuse, it is also clear that the angle between $\mathbf{v}$ and $\mathbf{w}$ is the same as the angle between $\mathbf{w}$ and $\mathbf{v}$. Moreover, it follows from the last-displayed formula (in the preceding paragraph) that *that* angle is acute (resp., obtuse) if and only $\mathbf{v} \cdot \mathbf{w} > 0$ (resp., $\mathbf{v} \cdot \mathbf{w} < 0$), the point being that both $|\mathbf{v}|$ and $|\mathbf{w}|$ are positive real numbers. Furthermore, if θ is the radian measure of *that* angle, then $0 < \theta < \pi$, and so the fact that $\cos|_{(0,\pi)}$ is a one-to-one function ensures that

$$\theta = \cos^{-1}\left(\frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|}\right).$$

The just-displayed fact leads to the following main tool that will be used in our proofs. Let $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ be nonzero vectors, no two of which are parallel. (There is no harm in viewing these three vectors as having the same initial point.) Let $\mathrm{m}(\angle)$ denote the radian measure of an angle $\angle$. Let $\angle_1$ be the angle between $\mathbf{u}$ and $\mathbf{w}$; and let $\angle_2$ be the angle between $\mathbf{v}$ and $\mathbf{w}$. Then:

$$\angle_1 \text{ is congruent to } \angle_2 \Leftrightarrow \mathrm{m}(\angle_1) = \mathrm{m}(\angle_2) \Leftrightarrow$$

$$\cos(\mathrm{m}(\angle_1)) = \cos(\mathrm{m}(\angle_2)) \Leftrightarrow \frac{\mathbf{u} \cdot \mathbf{w}}{|\mathbf{u}| \cdot |\mathbf{w}|} = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|}.$$

The above vectorial background can be used to define the following concept. It will play fundamental roles in our statement of the principle of reflection and, hence, in the proofs of the main results in Sections 2 and 3. Let P be a point on the graph of a differentiable function $f : I \to \mathbb{R}$, for some $I \subseteq \mathbb{R}$. By a *tangential vector to f at P*, we mean a (bound) vector $\mathcal{T} = \overrightarrow{PQ}$, where Q is a point on the tangent line to the graph of f at P such that $Q \neq P$. (The corresponding ray $\overrightarrow{PQ}$ is sometimes called a *tangential half-line of f at P*.)

In this section and in Section 3, we will use the following precise form of a principle of reflection:

Principle of reflection: In a Euclidean plane, suppose C is a curve, F is a point that is not on C but is understood to be "inside" C, L is a line that does not intersect C and does not pass through F and is understood to be "outside" C, P is a point on C, and the tangent line T to C at P exists. Let $\mathcal{T} = \overrightarrow{PQ}$ be a tangential vector to f at P. Suppose S is a point that is "outside" C and the ray $\mathcal{R} := \overrightarrow{SP}$ is perpendicular to L. Then the radian measure of the angle between the bound vectors $\mathcal{T}$ and $\overrightarrow{PS}$ is equal to the radian measure of the angle between the bound vectors $\mathcal{T}$ and $\overrightarrow{PF}$; that is, the two just-mentioned angles are congruent. (If S had instead been "inside" C, a context which is more suggestive of "reflection", an equivalent conclusion is that the two angles in question are supplementary. Accordingly, I was tempted to call the above principle a "Principle of refraction.") Suppose one has a context where F is considered to be a "focus of C" and L is considered to be the "corresponding directrix of C". Then (in view of the above parenthetical comment) the physical interpretation of the above "Principle of reflection" is the following twofold assertion. If a ray $\overrightarrow{\mathfrak{R}_1}$ comes from "inside" C on a line of action which is perpendicular to L such that $\overrightarrow{\mathfrak{R}_1}$ meets C at P, then the "reflected" ray which results from that intersection stays "inside" C and passes through F. If a ray $\overrightarrow{\mathfrak{R}_2}$ is emitted from F and meets C at P, then the "reflected" ray which results from that intersection stays "inside" C, going on a line of action which is perpendicular to L.

We next prove the main result of this section.

Theorem 2.1. Let $\mathcal{P}$ be a parabola with focus F and directrix L. Let P be a point on $\mathcal{P}$. Then the following two assertions are consequences of the above "Principle of reflection":

(a) Let $\overrightarrow{R}$ be a ray, coming from "inside" $\mathcal{P}$ on a line of action which is perpendicular to L, such that $\overrightarrow{R}$ meets $\mathcal{P}$ at P. Then the "reflected" ray which results from that intersection stays "inside" $\mathcal{P}$ (at least for a while) and passes through F.

(b) Let $\overrightarrow{R}$ be a ray which is emitted from F and meets $\mathcal{P}$ at P. Then the "reflected" ray which results from that intersection stays "inside" $\mathcal{P}$, going on a line of action which is perpendicular to L.

Proof. Let T be the tangent line to $\mathcal{P}$ at P. Fix a tangential half-line corresponding to T (that is, fix a ray emanating from P which points in one of the two directions of the line T), and then pick a point Q on that tangential half-line such that $Q \neq P$. Consider the (bound) vector $\mathcal{T} := \overrightarrow{PQ}$. Let S be a point on the horizontal line that passes through P and is perpendicular to L such that $S \neq P$. Consider the bound vector $\overrightarrow{PS}$. By the above discussion involving the "Principle of reflection", dot products and the inverse cosine function, it will be enough to prove that the angle that is between $\overrightarrow{PS}$ and $\mathcal{T}$ is congruent to the angle that is between $\overrightarrow{PF}$ and $\mathcal{T}$ (equivalently, that the radian measures of these angles have equal cosines).

Our task is to show that a certain pair of angles are congruent. By fundamental principles of Euclidean geometry, the congruence class of an angle does not change when the coordinate axes are rotated and/or translated. Thus our task is not changed (but its execution will be computationally eased) if we rotate the coordinate axes so that the perpendicular from F to L is horizontal (and then, necessarily, L is vertical), then translate the x-axis vertically so that F is on the (newly-named) x-axis, and then translate the y-axis horizontally so that the (newly-named) y-axis intersects the x-axis at a point that is exactly half-way between F and L. In other words, the origin (for the newly-created coordinate system that we are about to use) is half-way between F and L. By (other) fundamental principles of Euclidean geometry, the distance between two (possibly equal) points in a Euclidean plane does not change when the coordinate axes are rotated and/or translated. It follows that (after the completion of the above rotation and/or translations of axes), $\mathcal{P}$ is (still) the set of points P in the given Euclidean plane such that the distance from P to F equals the (perpendicular) distance from P to L. We have arranged that, for some uniquely determined nonzero real number a, the focus has coordinates $(a,0)$ and the directrix has Cartesian equation $x = -a$. Consequently, the origin is on $\mathcal{P}$ (since the distance from the origin to F and the distance from the origin to L are each equal to $|a|$). It is well known that a Cartesian equation for $\mathcal{P}$ is then $y^2 = 4ax$, but for the sake of completeness, we will establish that fact in the next paragraph.

We are now considering the parabola $\mathcal{P}$ with focus $F(a,0)$ and directrix $L : x = -a$ (for some nonzero $a \in \mathbb{R}$). Let $P(x,y)$ be a point in the plane. (This minor *abus de langage* should not be alarming, as the present P will, in effect, soon be shown to be the P in the statement of this result.) Using the distance formula and bearing in mind that distance is nonnegative, we have that P is on $\mathcal{P}$ $\Leftrightarrow$ the distance from P to F equals the distance from P to L $\Leftrightarrow$

$$\sqrt{(x-a)^2 + (y-0)^2} = |x-(-a)| \Leftrightarrow (x-a)^2 + y^2 = (x+a)^2 \Leftrightarrow$$

$$x^2 - 2ax + a^2 + y^2 = x^2 + 2ax + a^2 \Leftrightarrow y^2 = 4ax, \text{ as asserted.}$$

The proofs of (a) and (b) can be carried out simultaneously by examining two cases. Case 1 examines the graph of $y^2 = 4ax$ for some $a > 0$, while Case 2 examines the graph of $y^2 = 4ax$ for some $a < 0$. The graph pertinent to Case 1 (resp., Case 2) is given in Figure 1 (resp., Figure 2). Notice that in Figure 1 (that is, for Case 1), the "top" of the parabola is the graph of $y = f_1(x) := 2\sqrt{a}x^{1/2}$ and the "bottom" of the parabola is the graph of $y = f_2(x) := -f_1(x) = -2\sqrt{a}x^{1/2}$. Similarly, in Figure 2 (that is, for Case 2), the "top" of the parabola is the graph of $y = g_1(x) := 2\sqrt{-a}(-x)^{1/2}$ and the "bottom" of the parabola is the graph of $y = g_2(x) := -g_1(x) = -2\sqrt{-a}(-x)^{1/2}$. We have just used the following two familiar facts: if $u > 0$ and $v > 0$, then $\sqrt{uv} = \sqrt{u}\sqrt{v}$; and if $u < 0$ and $v < 0$, then $\sqrt{uv} = \sqrt{-u}\sqrt{-v}$. These facts will be used often in the proofs in Section 2 and 3 without further comment.

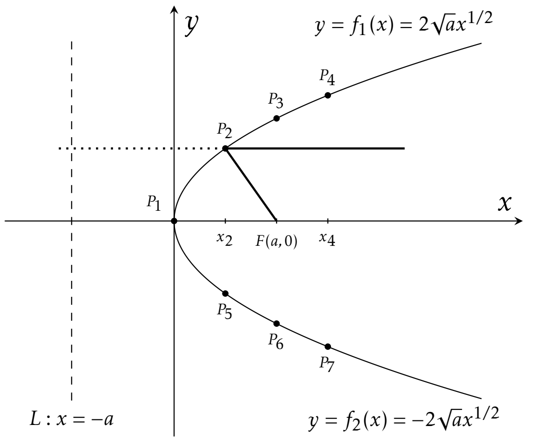
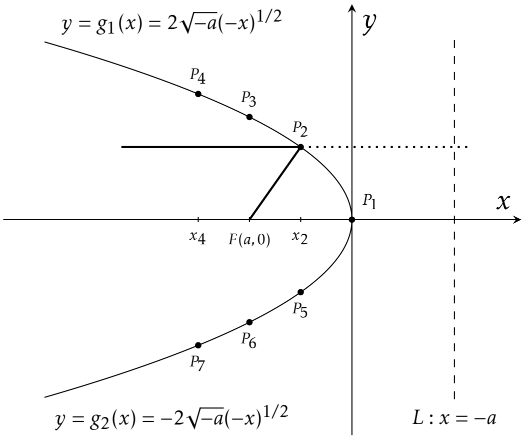


Figure 1: $y^2 = 4ax$, with fixed $a > 0$



$y = g_1(x) = 2\sqrt{-a}(-x)^{1/2}$
y
P_4
P_3
P_2
P_1
x
x_4
$F(a,0)$
x_2
P_5
P_6
P_7
$y = g_2(x) = -2\sqrt{-a}(-x)^{1/2}$
$L : x = -a$

Figure 2: $y^2 = 4ax$, with fixed $a < 0$

It is clear, for both Case 1 and Case 2 (that is, in both Figure 1 and Figure 2), that if P is a point on the parabola $\mathcal{P}$, then: the tangent line to $\mathcal{P}$ at P is vertical $\Leftrightarrow$ the line segment connecting P and F is horizontal $\Leftrightarrow$ P is the origin. Next, observe that the "Principle of reflection" (or any sensible variant of it) implies that the "angle of incidence" is a right angle if and only if the "angle of reflection" is a right angle. Since any two right angles are congruent (and so have the same radian measure), it follows that we have established the assertion for the point $P(0,0)$ in both (a) and (b).

We will next give separate (but similar) proofs for Case 1 and Case 2 for the points P_1 ($i = 2, \dots, 7$). In Case 1, these named points in Figure 1 have the following coordinates: $P_2(x_2, f_1(x_2))$ with $0 < x_2 < a$; $P_3(a, f_1(a))$; $P_4(x_4, f_1(x_4))$ with $x_4 > a$; $P_5(x_2, -f_1(x_2))$; $P_6(a, -f_1(a))$; and $P_7((x_4, -f_1(x_4))$. In Case 2, these named points in Figure 2 have the following coordinates: $P_2(x_2, g_1(x_2))$ with $a < x_2 < 0$; $P_3(a, g_1(a))$; $P_4(x_4, g_1(x_4))$ with $x_4 < a$; $P_5(x_2, -g_1(x_2))$; $P_6(a, -g_1(a))$; and, finally, $P_7((x_4, -g_1(x_4))$. It is clear that (apart from the origin P_1, which has already been treated) any point P on $\mathcal{P}$ is of the form P_i for a uniquely determined $i \in \{2, \dots, 7\}$.

For both Case 1 and Case 2, the first paragraph of this proof suggests that we should (and we will) use the following approach to study the situation at/"near" P_i (for $2 \leq i \leq 7$): since each of the functions f_1, $-f_1$, g_1 and $-g_1$ is differentiable at each of the three values of x that are relevant for it, we will be able (for each i) to find the slope, and hence a Cartesian equation of, the tangent line T to $\mathcal{P}$ at P_i, hence choose a tangential half-line of T that emanates from P_i, choose a point Q_i on that half-line which is distinct from P_i (to ease calculations, we will always take the x-coordinate of Q_i to be 0), consider the bound vector $\mathcal{T} := \overrightarrow{P_iQ_i}$, pick a point S_i that is "outside" $\mathcal{P}$ and on the horizontal line passing through P_i (hence $S_i \neq P_i$; to ease calculations, we will always take the x-coordinate of S_i to be 0), and consider the bound vector $\overrightarrow{P_iQ_i}$. As noted above: by the above discussion involving the "Principle of reflection", dot products and the inverse cosine function, it will be enough to prove (for $2 \leq i \leq 7$) that

$$\frac{\overrightarrow{P_iS_i} \cdot \mathcal{T}}{|\overrightarrow{P_iS_i}| \cdot |\mathcal{T}|} = \frac{\overrightarrow{P_iF} \cdot \mathcal{T}}{|\overrightarrow{P_iF}| \cdot |\mathcal{T}|};$$

equivalently, by multiplying through by $|\mathcal{T}| = |\overrightarrow{P_iQ_i}|$, that

$$\frac{\overrightarrow{P_iS_i} \cdot \mathcal{T}}{|\overrightarrow{P_iS_i}|} = \frac{\overrightarrow{P_iF} \cdot \mathcal{T}}{|\overrightarrow{P_iF}|};$$

equivalently, that

$$((\frac{1}{|\overrightarrow{P_iS_i}|})\overrightarrow{P_iS_i})\cdot\overrightarrow{P_iQ_i} = \frac{\overrightarrow{P_iF}\cdot\overrightarrow{P_iQ_i}}{|\overrightarrow{P_iF}|}.$$

Note that $(1/|\overrightarrow{P_iS_i}|)\overrightarrow{P_iS_i}$ is the unit vector (that is, the vector of length 1) which has the same direction as $\overrightarrow{P_iS_i}$. To further ease the calculations, this unit vector will always be either **i** or −**i**. Finally, in the proofs of Case 1 and Case 2, the proof for the subcase $P = P_2$ also works (by changing subscripts as needed) for the subcases $P = P_3$ and $P = P_4$; and the proof for the subcase $P = P_5$ can be similarly tweaked to handle the subcases $P = P_6$ and $P = P_7$. In short, we need only prove (a) and (b) for the cases $P = P_2$ and $P = P_5$; and to do *that*, we need only establish the last-displayed equation (for the appropriate value of i); and to do *that*, we will begin by choosing the above-mentioned tangential half-line of T, the points Q_i and S_i, and the above-mentioned unit vector.

Let us analyze Case 1 for the subcase $P = P_2$. The slope of the tangent line T to $\mathcal{P}$ at P is

$$m := f_1'(x_2) = 2\sqrt{a}(\frac{1}{2})(x_2)^{-1/2} = \frac{\sqrt{a}}{\sqrt{x_2}}(>0),$$

and so a Cartesian equation for T is

$$y = m(x-x_2)+f_1(x_2) = (\frac{\sqrt{a}}{\sqrt{x_2}})(x-x_2)+2\sqrt{a}\sqrt{x_2}.$$

Thus, the point $Q_2(0,\sqrt{a}\sqrt{x_2})$ is on the tangential half-line of T that emanates from P_2 and points in a somewhat south-westerly direction. Note also that $Q_2 \neq P_2$ (since $x_2 \neq 0$). Also, the point $S_2(0,2\sqrt{a}\sqrt{x_2})$ is "outside" $\mathcal{P}$ and to the left of P_2 on the horizontal line passing through P_2, and so the unit vector in the direction of $\overrightarrow{P_2S_2}$ is −**i**. Therefore, by the preceding paragraph, it will suffice to check that

$$-\mathbf{i}\cdot\overrightarrow{P_2Q_2} = \frac{\overrightarrow{P_2F}\cdot\overrightarrow{P_2Q_2}}{|\overrightarrow{P_2F}|}.$$

We have $\overrightarrow{P_2Q_2} = (0-x_2)\mathbf{i}+(\sqrt{a}\sqrt{x_2}-f_1(x_2))\mathbf{j}$, and so the left-hand side of the preceding display is $(-1)(-x_2)+0\cdot(\sqrt{a}\sqrt{x_2}-f_1(x_2)) = x_2$. Next, observe that $\overrightarrow{P_2F} = (a-x_2)\mathbf{i}+(0-f_1(x_2))\mathbf{j}$. Hence, the right-hand side of the preceding display is

$$\frac{((a-x_2)\mathbf{i}+(0-f_1(x_2))\mathbf{j})\cdot((0-x_2)\mathbf{i}+(\sqrt{a}\sqrt{x_2}-f_1(x_2))\mathbf{j})}{|(a-x_2)\mathbf{i}+(0-f_1(x_2))\mathbf{j}|} =$$

$$\frac{(a-x_2)(-x_2)+(-2\sqrt{a}\sqrt{x_2})(\sqrt{a}\sqrt{x_2}-2\sqrt{a}\sqrt{x_2})}{\sqrt{(a-x_2)^2+(-2\sqrt{a}\sqrt{x_2})^2}} =$$

$$\frac{-ax_2+x_2^2+2ax_2}{\sqrt{a^2-2ax_2+x_2^2+4ax_2}} = \frac{ax_2+x_2^2}{\sqrt{(a+x_2)^2}} = \frac{(a+x_2)x_2}{a+x_2} = x_2,$$

as desired.

To complete the analysis for Case 1, we now address its subcase $P = P_5$. Its proof will have essentially the same tempo as the above proof for the subcase $P = P_2$. The slope of the tangent line T to $\mathcal{P}$ at P $(= P_5(x_2,-2\sqrt{a}\sqrt{x_2}))$ is

$$m := f_2'(x_2) = -f_1'(x_2) = -\frac{\sqrt{a}}{\sqrt{x_2}}(<0),$$

and so a Cartesian equation for T is

$$y = m(x - x_2) + (-f_1(x_2)) = (-\frac{\sqrt{a}}{\sqrt{x_2}})x - \sqrt{a}\sqrt{x_2}.$$

Thus, the point $Q_5(0, -\sqrt{a}\sqrt{x_2})$ is on the tangential half-line of T that emanates from P_5 and points in a somewhat north-westerly direction. As in the earlier subcase, we also see that $Q_5 \neq P_5$. Also, the point $S_5(0, -2\sqrt{a}\sqrt{x_2})$ is "outside" $\mathcal{P}$ and is to the left of P_5 on the horizontal line passing through P_5, and so the unit vector in the direction of $\overrightarrow{P_5S_5}$ is $-\mathbf{i}$. Therefore, by the preceding paragraph, it will suffice to check that

$$-\mathbf{i} \cdot \overrightarrow{P_5Q_5} = \frac{\overrightarrow{P_5F} \cdot \overrightarrow{P_5Q_5}}{|\overrightarrow{P_5F}|}.$$

We have

$$\overrightarrow{P_5Q_5} = (0 - x_2)\mathbf{i} + (-\sqrt{a}\sqrt{x_2} - (-2\sqrt{a}\sqrt{x_2}))\mathbf{j} = -x_2\,\mathbf{i} + \sqrt{a}\sqrt{x_2}\,\mathbf{j},$$

and so the left-hand side of the next-to-last display is

$$(-1)(-x_2) + 0 \cdot \sqrt{a}\sqrt{x_2} = x_2.$$

Next, observe that

$$\overrightarrow{P_5F} = (a - x_2)\mathbf{i} + (0 - (-2\sqrt{a}\sqrt{x_2}))\mathbf{j} = (a - x_2)\mathbf{i} + 2\sqrt{a}\sqrt{x_2}\,\mathbf{j}.$$

So, the right-hand side of the next-to-next-to-next-to-last display is

$$\frac{((a - x_2)\mathbf{i} + 2\sqrt{a}\sqrt{x_2}\,\mathbf{j}) \cdot (-x_2\,\mathbf{i} + \sqrt{a}\sqrt{x_2}\,\mathbf{j})}{|(a - x_2)\mathbf{i} + 2\sqrt{a}\sqrt{x_2}\,\mathbf{j}|} =$$

$$\frac{(a - x_2)(-x_2) + (2\sqrt{a}\sqrt{x_2})(\sqrt{a}\sqrt{x_2})}{\sqrt{(a - x_2)^2 + (2\sqrt{a}\sqrt{x_2})^2}},$$

which simplifies to x_2 (by tweaking the corresponding step in the above proof for the subcase of P_2), as desired. This completes the proof for Case 1.

An experienced geometer or analyst may wish to argue ("conformally" while invoking symmetry) that the assertions for Case 2 follow from the corresponding assertions for Case 1, since the radian measure of an angle is preserved (up to algebraic sign) by any Euclidean reflection. However, in the interest of recording a self-contained proof for less experienced readers/students, we will next outline a rather detailed proof for Case 2. (It would be appropriate for many instructors/classes to assign some or all of the finer details of the proof that is heavily sketched below for Case 2 as homework or as questions on examinations.) This will be done by adapting the method of proof that was used above for Case 1. That overall approach will play a key role in the proof of Theorem 3.1 (which will give a partial converse for the assertions in Case 1 when P is of the form P_1, P_2, P_3 or P_4) and Remark 3.2 (giving similar partial converses for the other subcases).

We next analyze Case 2 for the subcase $P = P_2$. Recall that the relevant graph containing P_2 is that of the function given by $y = g_1(x) = 2\sqrt{-a}(-x)^{1/2}$ for $x \leq 0$. In particular, P_2 has coordinates $(x_2, 2\sqrt{-a}\sqrt{-x_2})$. For some students, the treatment of the functions g_1 and g_2 in Case 2 will seem slightly harder than the corresponding treatment of f_1 and f_2 had been in Case 1 because of the presence of "$(-x)^{1/2}$" in the formula for $g_1(x)$. Indeed, while those familiar with the standard rules from differential calculus (including the chain rule) can easily check that the derivative (with respect to x) of $\sqrt{-x}$ is $-1/(2\sqrt{-x})$ whenever $x < 0$, instructors may wish to plan to spend extra time with beginning students who are expected to calculate this derivative as an explicit limit of difference quotients. With that formula for the derivative of $\sqrt{-x}$ in hand, one sees easily that the slope of the tangent line T to $\mathcal{P}$ at P_2 is

$$m := g_1^{'}(x_2) = -\frac{\sqrt{-a}}{\sqrt{-x_2}}(< 0);$$

it follows that a Cartesian equation for T is

$$y = m(x - x_2) + (g_1(x_2)) = (-\frac{\sqrt{-a}}{\sqrt{-x_2}})x + \sqrt{-a}\sqrt{-x_2}.$$

(Note that the simplification that was used in the preceding calculation depended on the fact that $x_2/\sqrt{-x_2} = -\sqrt{-x_2}$. Similar facts will be used later in this proof.) Thus, the point $Q_2(0, \sqrt{-a}\sqrt{-x_2})$ is on the tangential half-line of T that emanates from P_2 and points in a somewhat south-easterly direction. As above, we also see that $Q_2 \neq P_2$. Also, the point $S_2(0, 2\sqrt{-a}\sqrt{-x_2})$ is “outside" $\mathcal{P}$ and to the right of P_2 on the horizontal line passing through P_2, and so the unit vector in the direction of $\overrightarrow{P_2S_2}$ is $\mathbf{i}$. Therefore, by tweaking the preceding reasoning, it will suffice to check that

$$\mathbf{i} \cdot \overrightarrow{P_2Q_2} = \frac{\overrightarrow{P_2F} \cdot \overrightarrow{P_2Q_2}}{|\overrightarrow{P_2F}|}.$$

Note $\overrightarrow{P_2Q_2} = -x_2\mathbf{i} - \sqrt{-a}\sqrt{-x_2}\,\mathbf{j}$ and $\overrightarrow{P_2F} = (a - x_2)\mathbf{i} - 2\sqrt{-a}\sqrt{-x_2}\,\mathbf{j}$. The left-hand side of the last display simplifies to $-x_2$. The right-hand side of the last display simplifies to

$$\frac{(a - x_2)(-x_2) + (-2\sqrt{-a}\sqrt{-x_2})(-\sqrt{-a}\sqrt{-x_2})}{\sqrt{(a - x_2)^2 + (-2\sqrt{-a}\sqrt{-x_2})^2}} =$$

$$\frac{-ax_2 + x_2^2 + 2ax_2}{\sqrt{a^2 - 2ax_2 + x_2^2 + 4ax_2}} = \frac{ax_2 + x_2^2}{\sqrt{(a + x_2)^2}} = \frac{(a + x_2)x_2}{|a + x_2|} = \frac{(a + x_2)x_2}{-(a + x_2)} =$$

$-x_2$, as desired.

To complete the analysis for Case 2, we now address its subcase $P = P_5$. Only minor changes will be needed in adapting the proof for the preceding subcase. The relevant graph containing P_5 is that of the function given by $y = g_2(x) = -2\sqrt{-a}(-x)^{1/2}$ for $x \leq 0$. In particular, P_5 has coordinates $(x_2, -2\sqrt{-a}\sqrt{-x_2})$. The slope of the tangent line T to $\mathcal{P}$ at P_5 is

$$m := g_2^{'}(x_2) = \frac{\sqrt{-a}}{\sqrt{-x_2}}(> 0);$$

it follows that a Cartesian equation for T is

$$y = m(x - x_2) + (g_2(x_2)) = (\frac{\sqrt{-a}}{\sqrt{-x_2}})x - \sqrt{-a}\sqrt{-x_2}.$$

Thus, the point $Q_5(0, -\sqrt{-a}\sqrt{-x_2})$ is on the tangential half-line of T that emanates from P_5 and points in a somewhat north-easterly direction. Note that $Q_5 \neq P_5$. Also, the point $S_5(0, -2\sqrt{-a}\sqrt{-x_2})$ is “outside" $\mathcal{P}$ and to the right of P_5 on the horizontal line that passes through P_5, and so the unit vector in the direction of $\overrightarrow{P_5S_5}$ is $\mathbf{i}$. Therefore, by tweaking the above reasoning, it will suffice to check that

$$\mathbf{i} \cdot \overrightarrow{P_5Q_5} = \frac{\overrightarrow{P_5F} \cdot \overrightarrow{P_5Q_5}}{|\overrightarrow{P_5F}|}.$$

Note $\overrightarrow{P_5Q_5} = -x_2\,\mathbf{i} + \sqrt{-a}\sqrt{-x_2}\,\mathbf{j}$ and $\overrightarrow{P_5F} = (a-x_2)\mathbf{i} + 2\sqrt{-a}\sqrt{-x_2}\,\mathbf{j}$. The left-hand side of the last display simplifies to $-x_2$. The right-hand side of the last display simplifies to

$$\frac{(a-x_2)(-x_2) + (2\sqrt{-a}\sqrt{-x_2})(\sqrt{-a}\sqrt{-x_2})}{\sqrt{(a-x_2)^2 + (2\sqrt{-a}\sqrt{-x_2})^2}}.$$

With *very* minor changes, the reasoning from the last subcase can be used to show that the last display simplifies to $-x_2$, as desired. The proof is complete. □

Remark 2.2. (a) Our contention that the strategy implemented in the proof of Theorem 2.1 actually gives a proof of Theorem 2.1 follows from some foundational facts about Euclidean plane geometry, including the "Plane Separation Axiom" (in short, the PSA). To see this, let π be a Euclidean plane and let T be a line in π. According to the PSA, T separates $\pi \setminus T$ into two half-planes, π_1 and π_2. (In detail, this can be done in a unique way, apart from permuting the labels "π_1" and "π_2", by requiring that π_1 and π_2 are nonempty, disjoint convex sets such that any closed line segment with one endpoint in π_1 and its other endpoint in π_2 must intersect T nontrivially. Colloquially, one views π_1 and π_2 as the subsets of $\pi \setminus T$ that lie on "opposite sides of" T.) Let P and Q be distinct points on T. Let S be a point in π_1 (and so $S \neq P$). Let F be a point in π_2 (and so $F \neq P$). Then (by appropriate facts/postulates concerning angles in Euclidean plane geometry), there exists a unique ray $\overrightarrow{R}$ emanating from P and pointing into π_2 such that F lies on $\overrightarrow{R}$ and the angle between $\overrightarrow{PQ}$ and $\overrightarrow{PS}$ is congruent to the angle between $\overrightarrow{PQ}$ and $\overrightarrow{PF}$.

(b) The statement of Theorem 2.1 referred to planar points that are either "inside" or "outside" a parabola $\mathcal{P}$. As the Jordan Curve Theorem does not apply to $\mathcal{P}$, some readers/students may wonder if some unspecified geometric intuition is being assumed in order to explain what is meant by these "sides" of a parabola. The answer is in the negative, as these sides can be defined in a geometrically rigorous way as follows, in terms of the vertex V, the focus F and the directrix L of $\mathcal{P}$. By defining the set of points "outside" $\mathcal{P}$ in a given Euclidean plane $\mathbb{R}^2$ (containing $\mathcal{P}$) as being the set-theoretic complement in that plane of the union of $\mathcal{P}$ and the set of points (in that plane) that are "inside" $\mathcal{P}$, our task is reduced to defining the set of points of $\mathbb{R}^2$ that are "inside" $\mathcal{P}$. That, in turn, can be done as follows. The *inside of* $\mathcal{P}$ consists of the points (in the given plane) of the form I which can be obtained as follows. Consider any point U on the ray $\overrightarrow{VF}$ (in the given plane) such that $U \neq V$; let L^* denote the line (in the given plane) that passes through U and is parallel to L; let V and W denote the two (necessarily distinct) points where L^* intersects $\mathcal{P}$; and then let I be *any* point on L^* that is strictly between V and W. This definition has the following analytic interpretation in case $\mathcal{P}$ is the graph of $y^2 = 4ax$ for some $a > 0$ (resp., for some $a < 0$), as the vertex of $\mathcal{P}$ is then the origin: a point (x, y) is inside $\mathcal{P}$ if and only if $x > 0$ and $-2\sqrt{a}\sqrt{x} < y < 2\sqrt{a}\sqrt{x}$ (resp., if and only if $x < 0$ and $-2\sqrt{-a}\sqrt{-x} < y < 2\sqrt{-a}\sqrt{-x}$).

(c): We could have proven Theorem 2.1 by studying parabolas $\mathcal{P}$ that arise as graphs of equations of the form $x^2 = 4ay$ (either for some $a > 0$ or for some $a < 0$), supported by the obvious analogues of Figure 1 and Figure 2 (in which the horizontal rays in Figure 1 and Figure 2 would be replaced by vertical rays). As that approach involves the graph of the differentiable function given by $y = h(x) = x^2/(4a)$, it would seem to require fewer (sub)cases than were used in the above proof of Theorem 2.1. Instructors/readers preferring such an approach are invited to carry out the obvious analogues of what we will do in the next two paragraphs (which will continue to address $y^2 = 4ax$, together with Figure 1 and Figure 2).

This paragraph begins an alternative proof of Theorem 2.1 that would be appropriate for an audience that is comfortable with "x being a function of y" and an ensuing derivative of x with respect to y. For any nonzero real number a, the graph of the equation $y^2 = 4ax$ is, of course, the graph of the function given by $x = \lambda(y) := y^2/(4a)$. As above, Figure 1 (resp., Figure 2) depicts the situation where $a > 0$ (resp., $a < 0$). The reflection behavior involving the point $P_1(0,0)$ can be handled as in the earlier proof (essentially because, at P_1, both the "angle of incidence" and the "angle of reflection" are right angles). We next address whether all the other points P_i (for $2 \leq i \leq 7$) can be handled at once (rather than considering several similar (sub)cases).

Let $P(x_0, y_0)$ be any P_i (for $2 \leq i \leq 7$). As neither of the coordinates of P is 0, we have

$$\frac{dx}{dy} = \frac{2y}{4a} = \frac{y}{2a} \text{ at } P, \text{ and so, by the Inverse Function Theorem,}$$

$$\frac{dy}{dx} = \frac{1}{(\frac{dx}{dy})} = \frac{2a}{y} \text{ at } P.$$

(Some care is needed in applying the Inverse Function Theorem. In first-year courses on calculus, that result is often stated for a strictly monotonic function that is defined on a closed interval and has a never-zero derivative over that entire interval (cf. [11, Theorem 6.2.3]). Although the given P could be handled via this form of the Inverse Function Theorem by devising a suitable closed interval, it would probably be less troublesome for students if an instructor would use a text on advanced calculus, such as [12, Theorem II, page 70], where the ambient interval is not assumed to be closed.) Thus, we find that the slope of the tangent line T to $\mathcal{P}$ at P is $2a/y_0$. Rather than giving a Cartesian equation for T, let us, instead, identify one of the tangential half-lines of T emanating from P as being the ray $\overrightarrow{PQ}$, where Q is the point $(x_0 + y_0, y_0 + 2a)$ on T. (These coordinates for Q were found by starting at P and then "running" y_0 units and "rising" $2a$ units.) Note that $Q \neq P$, since $y_0 \neq 0$ (alternatively, since $2a \neq 0$). Consider the point $S(0, y_0)$ which is "outside" $\mathcal{P}$ on the horizontal line that passes through P. Note that $S \neq P$, since $x_0 \neq 0$. We have the (bound) vectors

$$\mathcal{T} := \overrightarrow{PQ} = y_0\,\mathbf{i} + 2a\mathbf{j},\ \overrightarrow{PS} = -x_0\,\mathbf{i}, \text{ and } \overrightarrow{PF} = (a - x_0)\mathbf{i} - y_0\,\mathbf{j}.$$

As explained above (see also parts (a) and (b) of this remark), an (alternative) proof of Theorem 2.1 requires only a proof that

$$\frac{\overrightarrow{PS} \cdot \mathcal{T}}{|\overrightarrow{PS}|} = \frac{\overrightarrow{PF} \cdot \mathcal{T}}{|\overrightarrow{PF}|}; \text{ equivalently, that}$$

$$\frac{-x_0y_0 + 0(2a)}{\sqrt{(-x_0)^2 + 0^2}} = \frac{(a - x_0)y_0 + (-y_0)(2a)}{\sqrt{(a - x_0)^2 + (-y_0)^2}}; \text{ equivalently, that}$$

$$-x_0y_0\sqrt{a^2 - 2ax_0 + x_0^2 + y_0^2} = \sqrt{(-x_0)^2}[-x_0y_0 - ay_0]; \text{ equivalently, that}$$

$$(\frac{-y_0^2y_0}{4a})\sqrt{a^2 - \frac{y_0^2}{2} + \frac{y_0^4}{16a^2} + y_0^2} = (\frac{y_0^2}{4|a|})[-(\frac{y_0^2}{4a})y_0 - ay_0]; \text{ equivalently, that}$$

$$\frac{\sqrt{a^2 + \frac{y_0^2}{2} + \frac{y_0^4}{16a^2}}}{a} = \frac{\frac{y_0^2}{4a} + a}{|a|}.$$

The right-hand side of the last display simplifies to

$$(\frac{\frac{y_0^2}{4a} + a}{a})(\frac{a}{|a|}) = \frac{\frac{y_0^2}{4|a|} + \frac{a^2}{|a|}}{a} = \frac{\sqrt{(\frac{y_0^2}{4|a|} + \frac{a^2}{|a|})^2}}{a} =$$

$$\frac{\sqrt{\frac{y_0^4}{16|a|^2} + \frac{2y_0^2a^2}{4|a|^2} + \frac{a^4}{|a|^2}}}{a} = \frac{\sqrt{\frac{y_0^4}{16a^2} + \frac{y_0^2}{2} + a^2}}{a},$$

which clearly equals the corresponding left-hand side. This completes the alternative proof of Theorem 2.1.

3 The reflection property characterizes parabolas

We begin the section with a result which shows, by building on Theorem 2.1, how certain reflection properties serve to characterize what could be considered "half a parabola." The rest of the section explains how to adapt our methods in order to characterize all of, or selected arcs of, a parabola. Besides some expected uses of reflection properties, the novelty in Section 3 is the relevance of ordinary differential equations and the concomitant need to solve certain initial value problems.

Theorem 3.1. Let $0 < a \in \mathbb{R}$. Working in a fixed Euclidean plane $\mathbb{R}^2$, let F be the point with coordinates $(a, 0)$, let L be the line with Cartesian equation $x = -a$, let $f : [0, \infty) \to \mathbb{R}$ be a function, and let Γ be the graph of f. Suppose that f is differentiable on $(0, \infty)$, f is continuous at $x = 0$, $f(0) = 0$, f is strictly monotonic increasing and $f^{'}(x) \neq 0$ for all $x > 0$. Suppose also that Γ has a vertical tangent line at the origin. For each point P on Γ, let $T = T_P$ denote the tangent line to Γ at P, let $\overrightarrow{R} = \overrightarrow{R}_P$ denote a fixed tangential half-line induced by T (and emanating from P), let $Q = Q_P$ be a chosen point on $\overrightarrow{R}$ such that $Q \neq P$, and also let $S = S_P$ be a point that is "outside" Γ and on the horizontal line that passes through P and is perpendicular to L such that $S \neq P$. Then the following five conditions are equivalent:

(1) $f(x) = 2\sqrt{a}\sqrt{x}$ for all real numbers $x \geq 0$;

(2) Γ is the "top half" of the parabola with focus F and directrix L;

(3) For each point P on Γ (with T, $\overrightarrow{R}$, Q, and S associated to P, Γ and L as above), the angle that is between the bound vectors $\overrightarrow{PS}$ and $\overrightarrow{PQ}$ is congruent to the angle that is between the bound vectors $\overrightarrow{PF}$ and $\overrightarrow{PQ}$;

(4) For each point P on Γ (with T, $\overrightarrow{R}$, Q, and S associated to P, Γ and L as above),

$$((\frac{1}{|\overrightarrow{PS}|})\overrightarrow{PS}) \cdot \overrightarrow{PQ} = \frac{\overrightarrow{PF} \cdot \overrightarrow{PQ}}{|\overrightarrow{PF}|};$$

(5) Let P be a point on Γ. Then the following two reflection properties hold:

(i) If $\overrightarrow{\mathcal{L}}$ is a ray coming from "inside" Γ on a line of action which is perpendicular to L such that $\overrightarrow{\mathcal{L}}$ meets Γ at P, then the "reflected" ray which results from that intersection stays "inside" Γ (at least for a while) and passes through F;

(ii) If $\overrightarrow{\mathcal{L}}$ is a ray which is emitted from F and meets Γ at P, then the "reflected" ray which results from that intersection stays "inside" Γ (at least for a while), going on a line of action which is perpendicular to L.

Proof. Since f is strictly monotonic increasing, it follows easily from the definition of a derivative as a limit that there does not exist $\xi > 0$ such that $f^{'}(\xi) < 0$. Hence $f^{'}x) > 0$ for all $x > 0$. Note also that $f(x) > 0$ for all $x > 0$. Consider the upper half-plane $\mathcal{U} := \{(x, y) \in \mathbb{R}^2 \mid y \geq 0\}$. Clearly, $\Gamma \subseteq \mathcal{U}$. Moreover, if $x_0 \geq 0$, the horizontal line $y = f(x_0)$ intersects Γ only at the point $(x_0, f(x_0))$. As continuous functions preserve connected topological spaces, the image of f is a generalized subinterval of $[0, \infty)$. Thus, it is evident in a geometrically intuitive way that every point in the set $\{(x, y) \in \mathcal{U} \mid$ there exists $x_0 \in \mathbb{R}$ such that $x > x_0 \geq 0$ and $y = f(x_0)\}$ deserves to be viewed as being "inside Γ". Apart from these points and also apart from every point on Γ, it also seems compelling to agree to view all the *other* points of $\mathbb{R}^2$ as being "outside Γ". These considerations explain/justify the uses of "outside Γ" and "inside Γ" in the statement of this result.

Let $\mathcal{P}$ denote the parabola with focus F and directrix L. Recall from the third paragraph of the proof of Theorem 2.1 that $\mathcal{P}$ is the graph of the equation $y^2 = 4ax$. By defining the "top half" of $\mathcal{P}$ (if a definition of this term is really needed, in which case, the following definition is admittedly belated) as the set of points (x,y) on $\mathcal{P}$ such that (necessarily $x \geq 0$ and) $y \geq 0$, one sees easily that (1) $\Leftrightarrow$ (2).

Let P be a point on Γ. Although parts of the statements of conditions (3)-(5) are somewhat reminiscent of some earlier material, one may wonder if the quantifications pertaining to that earlier material align properly with the present context which is based, in part, on the fifth sentence in the statement of this result. To begin to allay such concerns, we will first show that the (possible) validity of the equation

$$((\frac{1}{|\overrightarrow{PS}|})\overrightarrow{PS}) \cdot \overrightarrow{PQ} = \frac{\overrightarrow{PF} \cdot \overrightarrow{PQ}}{|\overrightarrow{PF}|}$$

(from (4)) is not affected if one replaces S and Q with other points that meet the requirements indicated above. Indeed, this assertion concerning S is clear because changing S would not change the unit vector in the direction of $\overrightarrow{PS}$ (namely, $\overrightarrow{PS}/|\overrightarrow{PS}|$). Moreover, changing Q would simply replace the former $\overrightarrow{PQ}$ with $r\overrightarrow{PQ}$ for some $r > 0$, thus causing both the left- and right-hand sides of the last displayed equation to be multiplied by r, and that change would clearly also not affect the validity of the last displayed equation. Next, let us show that the validity of that equation is not affected if one chooses a different tangential half-line for the tangent line T to Γ at P. This new choice for $\overrightarrow{R}$ (and Q) would simply replace the former $\overrightarrow{PQ}$ with $s\overrightarrow{PQ}$ for some $s < 0$, thus causing both the left- and right-hand sides of the last displayed equation to be multiplied by s, and it is also clear that this change would not affect the validity of the last displayed equation. Finally, note that changes of the kind already discussed in this paragraph would not affect the validity of (3) (even though changing the tangential half-line $\overrightarrow{R}$ would change both of the angles mentioned in (3) to their supplements).

By the first paragraph of the proof of Theorem 2.1 (especially its sixth sentence invoking the earlier "discussion involving the Principle of reflection, dot products and the inverse cosine function"), it is now clear that (3) $\Leftrightarrow$ (5); and also that (3) $\Leftrightarrow$ (4). Moreover, Theorem 2.1 (see also Remark 2.2 (a)) gives that (2) $\Rightarrow$ (5). Therefore, it remains only to prove that (4) $\Rightarrow$ (1).

Assume (4). Let $P_0(x_0, y_0)$ be a point on Γ. We will prove that $f(x) = 2\sqrt{a}\sqrt{x}$ for all $x \geq 0$. Since we have assumed that f is continuous at $x = 0$ and $f(0) = 0$, we may assume henceforth that $x > 0$ and $x_0 > 0$. As $y_0 = f(x_0)$, the tangent line T to Γ at P_0 has Cartesian equation

$$y = f^{'}(x_0)(x - x_0) + f(x_0).$$

Thus, choosing the points $Q(0, -x_0 f^{'}(x_0) + f(x_0))$ and $S(0, f(x_0))$ is compatible with the above requirements for Q and S. Hence, by (4),

$$((\frac{1}{|\overrightarrow{P_0S}|})\overrightarrow{P_0S}) \cdot \overrightarrow{P_0Q} = \frac{\overrightarrow{P_0F} \cdot \overrightarrow{P_0Q}}{|\overrightarrow{P_0F}|}.$$

We have

$$\overrightarrow{P_0S} = -x_0\mathbf{i},\ \overrightarrow{P_0Q} = -x_0\mathbf{i} - x_0 f^{'}(x_0)\mathbf{j}, \text{ and}$$

$\overrightarrow{P_0F} = (a - x_0)\mathbf{i} - f(x_0)\mathbf{j}$. As $\overrightarrow{P_0S}/|\overrightarrow{P_0S}|$ is the unit vector in the direction of $\overrightarrow{P_0S}$, namely $-\mathbf{i}$, we get

$$-1(-x_0) + 0(-x_0 f^{'}(x_0)) = \frac{(a - x_0)(-x_0) + (-f(x_0))(-x_0 f^{'}(x_0))}{\sqrt{(a - x_0)^2 + (-f(x_0))^2}}.$$

By letting y denote the function f and doing some easy algebraic simplifications, we have thus reduced our task to showing that the only solution of the differential equation

$$\frac{dy}{dx} = \frac{\sqrt{(a-x)^2 + y^2} + a - x}{y}, \text{ for all } x > 0,$$

which is continuous at $x = 0$ and is such that $f(0) = 0$ satisfies $f(x) = 2\sqrt{a}\sqrt{x}$ for all $x > 0$. As it is easy to see that the function given by this formula *is* a solution of this initial value problem (it comes down to noticing that $|a + x| = a + x$ since $a > 0$ and $x > 0$), let us proceed to solve this differential equation.

Our methods will use a couple of changes of variable. First, consider $w := y^2$. By the chain rule,

$$\frac{dw}{dx} = 2y\frac{dy}{dx} \text{ for all } x > 0.$$

Then, by substituting the just-displayed fact into the above differential equation and doing some minor algebraic rewriting, we get

$$\frac{dw}{dx} = 2\sqrt{(a-x)^2 + w} + 2(a-x), \text{ for all } x > 0.$$

Since $y^2 = w$, it will suffice to prove that $w(x) = 4ax$ for all $x > 0$.

Next, consider $z := z(x) := w + (a-x)^2$ for all $x \geq 0$. Note that $z(0) = w(0) + a^2 = (y(0))^2 + a^2 = 0^2 + a^2 = a^2$. Also, by the usual rules of differential calculus,

$$\frac{dz}{dx} = \frac{dw}{dx} + 2(a-x)(-1) \text{ for all } x > 0.$$

Substituting the above expression for the derivative of w into the just-obtained expression for the derivative of z, we get

$$\frac{dz}{dx} = [2\sqrt{(a-x)^2 + w} + 2(a-x)] + 2(a-x)(-1) =$$

$$2\sqrt{(a-x)^2 + w} = 2\sqrt{z} \text{ for all } x > 0.$$

Note that $x > 0 \Rightarrow w = y^2 > 0 \Rightarrow z = w + (a-x)^2 \geq w + 0 > 0$. Separating variables and then performing indefinite integration(s), we get the following, for all $x > 0$:

$$\frac{dz}{\sqrt{z}} = 2\,dx \text{ and } \int \frac{dz}{\sqrt{z}} = \int 2\,dx.$$

Hence, there exists a constant of integration C such that $2\sqrt{z} = 2x + C$, for all $x > 0$. Therefore, by applying the operator $\lim_{x \to 0^+}$ (and using that z is a continuous function of x and $z(0) = a^2$), we have $2\sqrt{a^2} = 2 \cdot 0 + C$. Thus $C = 2a$. It follows that for all $x > 0$,

$$2\sqrt{z} = 2x + 2a, \text{ whence } \sqrt{z} = x + a, \text{whence}$$

$$w = z - (a-x)^2 = (\sqrt{z})^2 - (a-x)^2 = (x+a)^2 - (a-x)^2 = 4ax.$$

The proof is complete. □

Theorem 3.1 has shown that, once coordinate axes have been rotated and translated so that parabolas of interest can be assumed to be graphs of equations of the form $y^2 = 4ax$, one can use the “Principle of reflection" to obtain a function-theoretic characterization of the “top" half of any such parabola if $a > 0$. In fact, one can use the “Principle of reflection" to obtain a function-theoretic characterization of parabolas. Indeed, this can be done by combining Theorem 2.1 with three suitable analogues of Theorem 3.1 to get respective characterizations of the “bottom" half of the graph of $y^2 = 4ax$ when $a > 0$, the “top" half of the graph of $y^2 = 4ax$ when $a < 0$, and the “bottom" half of the graph of $y^2 = 4ax$ when $a < 0$. The precise statements of those analogues are given and proofs are sketched in parts (b) and (c) of Remark 3.2. Those sketches suitably adapt the proof of Theorem 3.1. Remark

3.2 (d) follows up a comment from the Introduction (and Remark 2.2 (c)), by indicating how similar characterizations of sufficiently small arcs of parabolas (viewed as graphs of equations of the form $x = y^2/(4a)$) can be obtained for readers/classes that are comfortable dealing with "x as a function of y", $\frac{dx}{dy}$, and partial derivatives. Finally, in Remark 3.2 (e), an analogous characterization of an arc of a parabola is stated and a proof of it is sketched. One should note that the above-mentioned papers of Drucker also featured a similar characterization of parabolic arcs.

Remark 3.2. (a) Although it may have been a distraction in Theorem 3.1 to point out the redundancy of its assumption that the tangent line to Γ at the origin exists and is vertical, we wish to do so here. This will be done by appealing to the following somewhat standard definition (the literature is surprisingly nonuniform about this matter!): if x_0 is in the domain of a real-valued function f of one variable, then the graph of f has a vertical tangent line at $(x_0, f(x_0))$ if f is continuous at x_0 and $\lim_{x\to x_0} f^{'}(x) = \pm\infty$. We show next that a function f satisfies this condition if it satisfies *all the other* conditions stipulated in the second and third sentences of the statement of Theorem 3.1. To see this, note first that f is assumed to be continuous at $x_0 := 0$. Moreover, if we examine the secant lines whose limiting position (if it exists) would be that of the tangent line to Γ at the origin, the corresponding limit of the slopes of those secant lines is

$$\lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} = \lim_{x\to 0} \frac{f(x)-f(0)}{x-0} = \lim_{x\to 0^+} \frac{f(x)-0}{x} = \lim_{x\to 0^+} f^{'}(x),$$

where the last step was obtained by using the general form of L'Hôpital's Rule (as formulated in [15, Theorem 1]). Next, by tweaking the *proof* that (4) ⇒ (1) in Theorem 3.1 (by now allowing x_0 to be 0 wherever needed in that proof), we can use the formula for the derivative of $y = f(x)$ in that proof to reformulate our task as seeking a proof that

$$\lim_{x\to 0^+} \frac{\sqrt{(a-x)^2+y^2}+a-x}{y} = \infty.$$

To that end, recall that $a > 0$ and that f takes only positive values when $x > 0$. Hence, working in the extended real number system ($\mathbb{R}\cup\{\infty,-\infty\}$) and using the appropriate limit theorem there (while bearing in mind that f is continuous at the origin), we get

$$\lim_{x\to 0^+} \frac{\sqrt{(a-x)^2+y^2}+a-x}{y} = \frac{\sqrt{(a-0)^2+0^2}+a-0}{0^+} = \frac{2a}{0^+} = \infty,$$

as desired. This proof should dispel any lingering worries that the proof of Theorem 3.1 may have only characterized the "open top half" of the parabola $y^2 = 4ax$ (when $a > 0$), as we have just shown that the origin is indeed part of the "top half" which was characterized in that proof (even if one had not assumed that the tangent line to Γ at the origin exists and is vertical). We will have further need to consider the extended real number system in (e) below.

(b) In the spirit of Theorem 3.1, we can characterize the "bottom" half of the graph of $y^2 = 4ax$ when $a > 0$ by modifying the statement and proof of Theorem 3.1 as follows. Assume that f is strictly monotonic *decreasing*. One can show, by tweaking the above argument in (a), that there will be no need to assume that Γ has a vertical tangent line at the origin. (Indeed, this *will follow*, as we will get that

$$\lim_{x\to 0^+} \frac{dy}{dx} = \frac{2a}{0^-} = -\infty,$$

by exploiting the fact that f will now take only *negative* values for $x > 0$.) In the statement of condition (1), change the formula for $f(x)$ to $f(x) = -2\sqrt{a}\sqrt{x}$ for all $x \geq 0$. In the statement of condition (2), change "top half" to "bottom half". There is no need to change the statements of conditions (3),

(4) or (5). In the proof, conclude that $f^{'}(x) < 0$ for all $x > 0$, and replace $\mathcal{U}$ with the *lower* half-plane, $\{(x,y) \in \mathbb{R}^2 \mid y \leq 0\}$. Also, for the proof that (4) $\Rightarrow$ (1): we do get the same formula for the derivative of y with respective to x as before; we do get that $\sqrt{z} = x + a$ and $w(x) = 4ax$ as before; and, since y is now $-\sqrt{w}$, we get $y = -2\sqrt{a}\sqrt{x}$, to complete the proof.

(c) In the spirit of Theorem 3.1 and (b), we can characterize both the "top half" and the "bottom half" of the graph of $y^2 = 4ax$ when $a < 0$ by modifying the statement and proof of Theorem 3.1. In explaining how to carry out those modifications, the next paragraph addresses both the "top half" context and the "bottom half" context.

For the "top half" (resp., "bottom half") context, the "strictly monotonic behavior" of f is assumed to be decreasing (resp., increasing), the formula for $f(x)$ in condition (1) is $f(x) = 2\sqrt{-a}\sqrt{-x}$ (resp., $f(x) = -2\sqrt{-a}\sqrt{-x}$) for all $x \leq 0$, and there is no need to change the statements of conditions (3), (4) or (5) from what they had been in Theorem 3.1. We next address the proofs for the two contexts. For the "top half" (resp., "bottom half"), conclude that $f^{'}(x) < 0$ (resp., $f^{'}(x) > 0$) for all $x < 0$ and notice that Γ is a subset of the upper (resp., lower) half-plane. Of course, since $a < 0$, a unit vector in the direction of $\overrightarrow{PS}$ is now $\mathbf{i}$. Consequently, one sign changes in the relevant differential equation; that equation simplifies algebraically to

$$\frac{dy}{dx} = \frac{-\sqrt{(a-x)^2 + y^2} + a - x}{y}, \text{ for all } x < 0.$$

In solving that new differential equation (with the same changes of variable as in the proof of Theorem 3.1), we find that the derivative of z with respect to x is $-2\sqrt{z}$ for all $x < 0$, whence $2\sqrt{z} = -2x + C$ for all $x < 0$, whence (by application of the operator $\lim_{x \to 0^-}$) we get $C = 2\sqrt{a^2} = -2a$, whence $\sqrt{z} = -x - a$ for all $x < 0$, whence

$$w = y^2 = z - (a-x)^2 = (-x-a)^2 - (a-x)^2 = 4ax, \text{ for all } x < 0.$$

Therefore, in the proof that (4) $\Rightarrow$ (1), we find that the "top half" (resp., "bottom half") is described by $y = \sqrt{w} = 2\sqrt{-a}\sqrt{-x}$ (resp., $y = -\sqrt{w} = -2\sqrt{-a}\sqrt{-x}$) for all $x < 0$, with continuity then ensuring the corresponding equality at $x = 0$. It remains only to explain why there is no need to assume that Γ has a vertical tangent line at the origin. Indeed, this will follow for the "top half" since

$$\lim_{x \to 0^-} \frac{dy}{dx} = \frac{2a}{0^+} = -\infty,$$

the relevant matter being that for the "top half", f takes only *positive* values for $x < 0$. On the other hand, for the "bottom half", f takes only *negative* values for $x < 0$, whence the "bottom half" of Γ also has a vertical tangent at the origin, since that part of that graph's data satisfies

$$\lim_{x \to 0^-} \frac{dy}{dx} = \frac{2a}{0^-} = \infty.$$

(d) In the spirit of Remark 2.2 (c), we next present an alternate approach to (what is effectively the main thrust of) Theorem 3.1 that would be appropriate for an audience that is comfortable with "x being a function of y" and partial derivatives.

Let $0 \neq a \in \mathbb{R}$. Consider the planar point $F(a,0)$ and the vertical line L with Cartesian equation $x = -a$. Assume that a smooth function (that is, a differentiable function with a continuous derivative) given by $x = \lambda(y)$ is such that the graph Γ of λ satisfies a reflection-theoretic condition that is in the spirit of conditions (3), (4) and (5) from Theorem 3.1. Assume also that $\lambda(y) = 0$ if and only if $y = 0$, and that $\lambda^{'}(0) = 0$ if and only if $y = 0$. Let $P_0(x_0, y_0)$ be a point on Γ which is not the origin. Using the motivation (from the Inverse Function Theorem) that the "run"/"rise" of a tangential vector $\mathcal{T}$ of λ at P_0 should be $\lambda^{'}(y_0)$, let us use

$$\mathcal{T} := \lambda^{'}(y_0)\mathbf{i} + \mathbf{j}.$$

Suppose that $a > 0$. Expecting that $x \geq 0$ and also anticipating the location of the "outside" and "inside" of Γ, points Q_0 and S_0 are chosen so that the bound vector $\overrightarrow{PQ} = \mathcal{T}$ and the unit vector in the direction of $\overrightarrow{P_0S_0}$ is $-\mathbf{i}$. After dropping subscripts "$_0$", the equation (from condition (4))

$$((\frac{1}{|\overrightarrow{PS}|})\overrightarrow{PS}) \cdot \overrightarrow{PQ} = \frac{\overrightarrow{PF} \cdot \overrightarrow{PQ}}{|\overrightarrow{PF}|}$$

leads to

$$-\lambda^{'}(y) = \frac{(a-x)\lambda^{'}(y) + (-y)1}{\sqrt{(a-x)^2 + (-y)^2}} \text{ for all } y \neq 0, \text{ whence}$$

$$\frac{dx}{dy} = \frac{y}{\sqrt{(a-x)^2 + y^2} + a - x} \text{ for all } y \neq 0.$$

Note that this equation also holds at $y = 0$ (where, necessarily, $x = 0$), since

$$\lambda^{'}(0) = 0 = \frac{0}{2a} = \frac{0}{\sqrt{(a-0)^2 + 0^2} + a - 0}.$$

It is easy to check that the above first-order ordinary differential equation (ODE) is satisfied by the (smooth) function given by $x = y^2/(4a)$, and that this function and its derivative each take the value 0 only at $y = 0$. I believe that many experienced readers would now find it reasonable to consider an associated initial value problem. We will discuss that two paragraphs hence, before going on to solve the problem at hand.

First, we wish to point out that the above choices were not made independently. Indeed, we show next that if the point S_0 had, instead, been chosen so that unit vector in the direction of $\overrightarrow{P_0S_0}$ is $\mathbf{i}$, then *that vector* would not have pointed "outside" Γ when we restrict attention to $0 < y$ $(< \infty)$. By the above methods, one consequence of this new choice for S_0 would be that $\lambda^{'}(y)$ would become the negative of the earlier expression for $\lambda^{'}(y)$, for all $y \geq 0$. Recall that the values of $\lambda^{'}$ (resp., of λ) are nonzero and have the same algebraic sign for all $y > 0$. We next derive a contradiction (in the next sentence) from the assumption that $x < 0$ and $y \to \infty$. Since $a > 0$, that assumption would imply that $\frac{dx}{dy} > 0$, whence x would be a strictly increasing function of y, whence $\mathbf{i}$ would point "inside" Γ, the desired contradiction. We can conclude that $x > 0$ as $y \to \infty$ (and hence that $x > 0$ for all $y > 0$). It follows that $\lambda^{'}(y) > 0$ for all $y > 0$, and so λ is a strictly increasing function of y on $[0, \infty)$. As promised, we turn next to some matters related to initial value problems.

Consider the real-valued function G of two variables given by

$$G(x,y) = \frac{y}{\sqrt{(a-x)^2 + y^2} + a - x}.$$

The (natural) domain of G consists of all the points of $\mathbb{R}^2$ except $F(a,0)$. Note that G is continuous (in the usual sense, for a function of two real variables) on that domain. In view of the literature on (existence and) uniqueness theorems for ODEs (especially, the celebrated theorem of Picard-Lipshitz), it is natural to examine the behavior of $\partial G/\partial x$. By using familiar formulas from differential calculus, one checks easily that this partial derivative of G exists on the just-mentioned domain and is continuous on that domain. So, by the just-mentioned literature, the initial value problem consisting of the ODE $dx/dy = G(x,y)$ and the initial condition $x(0) = 0$ has a unique solution $x = \nu(y)$ "locally", in the sense that there exists a closed rectangle $\mathcal{R} := [b,c] \times [d,g] \subset \mathbb{R}^2$ such that the origin is an interior point of $\mathcal{R}$ (that is, $b < 0 < c$ and $d < 0 < g$), with $\nu(0) = 0$, and also such that both $b \leq \nu(y) \leq c$ and $\nu^{'}(y) = G(\nu(y), y)$ hold whenever $d \leq y \leq g$ (and, necessarily $c < a$), and also such that $\nu|_{[d,g]}$ is the only function of y with these properties. Hence, by the above material, $\lambda(y) = x = \nu(y) = y^2/(4a)$ whenever $d \leq y \leq g$. (In essence, we have now managed to characterize a certain arc of the parabola $y^2 = 4ax$. In (e), we will be more precise, studying – and characterizing – arbitrary arcs of $y^2 = 4ax$ regardless of whether $a > 0$ (as in the present situation) or $a < 0$.) Our goal here in (d) is to prove more than "local" conclusions, namely, that $\lambda(y) = x = y^2/(4a)$ *for all* $y \in \mathbb{R}$. Unfortunately, I do not know of a "global" (existence-)uniqueness theorem for solutions of initial value problems involving ODEs that would directly give this conclusion at this point. (Perhaps, someone who is more knowledgeable than I about ODEs will be aware of such a theorem. Note that there *does* exist a "somewhat global" existence-uniqueness theorem with the above flavor, but its assumptions impose natural restrictions on the vertical extent of the associated closed rectangle $\mathcal{R}$. To see this, consider the (expected and unique) solution for $\nu(y)$ as being $y^2/(4a)$. If one uses this expression for x and if one could put $x := a$ (that is, if the "horizontal base" $[b,c]$ of $\mathcal{R}$ contains a), one would get $y = \pm\sqrt{4ax} = \pm 2a$, so that the "vertical base" $[d,g]$ of $\mathcal{R}$ contains either $2a$ or $-2a$, whence the point $(a,0)$ is in $\mathcal{R}$ but not in the domain of G (contrary to the assumptions of the known "somewhat global" theorem). One thus infers the promised restriction, namely, either $d > -2a$ or $g < 2a$.) Also, I have not been clever enough to find change(s) of variable that would give a closed-form expression for the general solution of the ODE $dx/dy = G(x,y)$ (for $-\infty < y < \infty$) which could then be used in conjunction with the initial condition $x(0) = 0$ to produce the (expected and unique) solution for $\nu(y)$ as being $y^2/(4a)$ (over $-\infty < y < \infty$). (Perhaps, someone will be more clever in that regard.) However, we will proceed to solve the problem in a "global" way. The naïve way to explain our upcoming approach is that ν has an inverse function, the derivative of that inverse function has a form that we examined in the proof of Theorem 3.1, and we solved the initial value problem associated with that ODE in that proof. It is fair for the reader to ask the following: why would it be necessary to say much more than that before concluding? My answer is that one must first determine the domain and range of that inverse function. That will be done next. The upcoming details are somewhat predictable, admittedly tedious at some points, and (in my opinion) necessary if one is to compete a proof of Theorem 3.1 having started from the "x as a function of y" point of view.

Recall that $\frac{dx}{dy} = y/(\sqrt{(a-x)^2 + y^2} + a - x)$ for all $y \geq 0$. Let us restrict y to the domain $[0,\infty)$ and, by *abus de langage*, continue to use λ to denote the restriction of λ to that domain. Then it follows (via the Mean Value Theorem) that λ (with its domain restricted as just mentioned) is a strictly increasing monotonic function and, hence, has an inverse function. Let μ denote that inverse function. Of course, the range of μ is $[0,\infty)$ (because we restricted the domain of λ to be $[0,\infty)$). One would expect the domain of μ to be $[0,\infty)$; equivalently, one would expect the range of λ to be $[0,\infty)$ (when the domain of λ has been restricted as above). To prove this, it will suffice to show that $\lim_{y\to\infty} \lambda(y) = \infty$, since continuous functions preserve connectedness.

Suppose, on the contrary, that $\lim_{y\to\infty} \lambda(y)$ is not ∞. Then, since λ is strictly monotonic increasing, $\lim_{y\to\infty} \lambda(y) = M$, for some real number $M > 0$. We will show that this leads to a contradiction.

By the inverse function theorem,

$$\mu^{'}(x) = \frac{1}{\lambda^{'}(y)} = \frac{\sqrt{(a-x)^2 + y^2} + a - x}{y} \text{ whenever } 0 < x < M.$$

So, by the *proof* of Theorem 3.1,

$$(y =)\, \mu(x) = 2\sqrt{a}\sqrt{x} \text{ whenever } 0 \leq x < M.$$

Therefore, each value of y that is in the domain of λ satisfies $y < 2\sqrt{a}\sqrt{M}$. Since the domain of λ is

$[0,\infty)$, we have found the desired contradiction. This completes the proof that $\lim_{y\to\infty}\lambda(y)=\infty$, and hence that the range of λ is $[0,\infty)$, and hence that the domain of μ is $[0,\infty)$ (if $a>0$).

Then, by the proof that (4) ⇒ (1) in Theorem 3.1, $y=\sqrt{w}=\sqrt{4ax}=2\sqrt{a}\sqrt{x}$ for all $x>0$. By continuity, this equation still holds at $x=0$. Hence, the intersection of Γ with the upper half-plane is

$$\{(x,y)\in\mathbb{R}^2 \mid y\geq 0,\, x\geq 0 \text{ and } y=2\sqrt{a}\sqrt{x}\},$$

namely, the "top half" of the parabola given by $y^2=4ax$.

Next, while still assuming that $a>0$, suppose now that $y\leq 0$. Arguing as above, one can show that

$$\frac{dx}{dy}=\frac{y}{\sqrt{(a-x)^2+y^2}+a-x} \text{ for all } y<0.$$

Tweaking the reasoning showing that (4) ⇒ (1) in the proof of Theorem 3.1, we still get that $\sqrt{z}=x+a$ and $w=4ax$. But now, as $y<0$, these facts lead to $y=-\sqrt{w}=-2\sqrt{a}\sqrt{x}$ for all $x>0$. By continuity, this equation still holds at $x=0$. Consequently, the intersection of Γ with the lower half-plane is

$$\{(x,y)\in\mathbb{R}^2 \mid y\leq 0,\, x\geq 0 \text{ and } y=-2\sqrt{a}\sqrt{x}\},$$

namely, the "bottom half" of the parabola given by $y^2=4ax$. As Γ is the union of its intersections with the upper half-plane and lower half-plane, it follows from the next-to-next-to-last display and the last display that $\Gamma=\{(x,y)\in\mathbb{R}^2 \mid y^2=4ax\}$. This completes the proof in case $a>0$. The similar details proving the analogous conclusion for the case $a<0$ are left to the reader.

(e) Drucker noted in [6] that his methods led to reflection-theoretic characterizations of arcs of a parabola. We next sketch how to modify the statement and proof of Theorem 3.1 in order to obtain characterizations (including one that is explicitly reflection-theoretic) of arcs of parabolas that are described, without loss of generality, as graphs of equations of the form $y^2=4ax$ for some given nonzero real number a. It will suffice to work with closed arcs, as the context of "open arcs" could then be handled by simply restricting the (closed) domains of relevant functions and looking at appropriate subsets of the corresponding relevant graphs. Consider the planar point $F(a,0)$ and the vertical line L with Cartesian equation $x=-a$.

Let $a>0$. Let us first consider an arc γ (hopefully taken from the graph of $y^2=4ax$), assuming that γ is a subset of the *upper* half-plane and is based on the interval $\alpha\leq x\leq\beta$, where $0\leq\alpha\in\mathbb{R}$ and $\alpha<\beta\in\mathbb{R}\cup\{\infty\}$. (By convention, $\alpha\leq x\leq\infty$ describes the closed subset $[\alpha,\infty)$ of $\mathbb{R}$.) We will modify the reasoning from (d) by sketching how the assumption that γ satisfies reflection-theoretic criteria having the flavor of conditions (3)-(5) from Theorem 3.1, but having context that is based on the interval $\alpha\leq x\leq\beta$ (rather than $[0,\infty)$), can imply that a suitable function $f:[\alpha,\beta]\to\mathbb{R}$ having γ as its graph must necessarily be given by $f(x)=2\sqrt{a}\sqrt{x}$ for all x such that $\alpha\leq x\leq\beta$.

Recall that we are assuming that f is a function $[\alpha,\beta]\to\mathbb{R}$ whose graph, γ, is in the upper half-plane and that reflection-theoretic conditions such as (4) are satisfied for all points on γ. Suppose also that f is differentiable on (α,β); also, that if $\alpha\neq 0$, then f is differentiable at $x=\alpha$; and also that if $\beta\neq\infty$, then f is differentiable at $x=\beta$. Suppose also that f is continuous at $x=\alpha$, $f(\alpha)=2\sqrt{a}\sqrt{\alpha}$, f is strictly monotonic increasing and $f^{'}(x)\neq 0$ whenever $\alpha<x\leq\beta$. Note that if $\alpha=0$ then, by tweaking the reasoning in (a), one can prove that γ has a vertical tangent line at the origin. The analysis carries on much as it did in the proof that (4) ⇒ (1) in Theorem 3.1. In particular, working with a point $P_0(x_0,y_0)$ on γ (but avoiding the possibility that $x_0=\alpha$ if $\alpha=0$), select Q_0 and S_0 as in the earlier proof, being sure that the unit vector in the direction of $\overrightarrow{P_0S}$ is $-\mathbf{i}$. As in the proof of Theorem 3.1, obtain an ordinary ODE and work to solve the corresponding initial value problem (involving the condition $y(\alpha)=2\sqrt{a}\sqrt{\alpha}$) with ($y=f(x)$ and) the changes of variable $w:=y^2$ and $z:=z(x):=w+(a-x)^2$ for all x such that $\alpha\leq x\leq\beta$. Then

$$z(\alpha)=w+(a-\alpha)^2=(f(\alpha))^2+(a-\alpha)^2=(2\sqrt{a}\sqrt{\alpha})^2+(a-\alpha)^2=(a+\alpha)^2.$$

Eventually, separate variables and perform indefinite integration, getting a constant of integration C such that $2\sqrt{z} = 2x + C$ whenever $\alpha < x \leq \beta$. By applying the operator $\lim_{x\to\alpha^+}$ (and continuity), we get

$$2(a+\alpha) = 2\sqrt{(a+\alpha)^2} = 2\sqrt{z(\alpha)} = 2\alpha + C,$$

whence $C = 2a$, whence (as before)

$$w = z - (a-x)^2 = (x+a)^2 - (a-x)^2 = 4ax.$$

Thus, $f(x) = y = \sqrt{w} = 2\sqrt{a}\sqrt{x}$ for all x such that $\alpha < x \leq \beta$. Once again applying the operator $\lim_{x\to\alpha^+}$ (and continuity), we get that this equation also holds at $x = \alpha$. This completes the proof of the characterization result for an arc in the upper half-plane in case $a > 0$.

It remains only to indicate the changes that are needed while analyzing the other (sub)cases. To save space, I will leave to the reader the details of how the assumptions on the function f should be modified in each of those situations.

While still supposing that $a > 0$, one can produce characterizations (including one that is explicitly reflection-theoretic) of an arc γ (hopefully taken from the graph of $y^2 = 4ax$), assuming that γ is a subset of the *lower* half-plane and is based on the interval $\alpha \leq x \leq \beta$ where (as above) $0 \leq \alpha \in \mathbb{R}$ and $\alpha < \beta \in \mathbb{R} \cup \{\infty\}$. A reader who has carried out the activity mentioned in the final sentence of (d) should have no trouble in modifying what we have already done in (e) in order to produce the desired characterizations.

Next, while still supposing that $a > 0$, one can produce characterizations of the desired kind for an arc γ (hopefully taken from the graph of $y^2 = 4ax$), assuming that γ is not a subset of the upper half-plane and that γ is not a subset of the lower half-plane, by proceeding as follows. Let γ_1 (resp., γ_2) denote the intersection of γ with the upper (resp., lower) half-plane. Then γ_1 is based on the interval $\alpha_1 = 0 \leq x \leq \beta_1$ where $0 < \beta_1 \in \mathbb{R}\cup\{\infty\}$; and γ_2 is based on the interval $\alpha_2 = 0 \leq x \leq \beta_2 \in \mathbb{R}\cup\{\infty\}$. Earlier in (e), we have seen in some detail how to get the desired characterizations of γ_1; and in the preceding paragraph, we have indicated how to produce the corresponding characterizations of γ_2. To obtain a characterization of γ, the reader need only combine any of the former characterizations (of γ_1) with any of the latter characterizations (of γ_2). Since one of those combinations will be reflection-theoretic (while addressing all of γ), the discussion for the case $a > 0$ is complete.

The details for the case $a < 0$ can be handled in the above spirit. To wit, the reader is advised to do the following: by tweaking the above approach, first address (hopefully parabolic) arcs γ that lie entirely in the upper half-plane, with γ based on an interval $\alpha \leq x \leq \beta$, where $\alpha \in \mathbb{R} \cup \{-\infty\}$, $\alpha < \beta \in \mathbb{R}$ and $\beta \leq 0$; next, modify your work to address (hopefully parabolic) arcs that lie in the lower half-plane (and are based on the same kind of interval); and, finally, combine the two parts of your work by combining characterizations of the intersection of γ and the upper half-plane with characterizations of the intersection of γ and the lower half-plane. Readers who carry out these steps will have obtained characterizations that are precise, detailed and complete. I would then invite those readers to judge whether characterizations with those qualities are as readily obtainable from [6].

(f) Before closing this line of inquiry, we cannot resist the temptation of raising the question of whether one could use *partial* differential equations to obtain characterizations (with the above flavor) of parabolas. Frankly, a comprehensive answer to this question would go beyond my expertise and it would probably take us too far afield (while imposing more severe prerequisites than we have assumed so far). However, we have found what may be reasons to think that such a study could be successful. The next paragraph comments further on this, but please be advised that the next paragraph is intended only to be motivational, as I am not an expert on partial differential equations.

We again warn that this paragraph is only motivational, as I am not up-to-date on partial differential equations. In seeking tools that may be useful in characterizing all of the (parabolic) graph of

$y^2 = 4ax$ (rather than giving separate characterizations of only its "top half" and its "bottom half"), it may be useful to find a formula for tangential vectors $\mathcal{T}$ that works without exception (even at the origin). Suppose, for instance, that $a > 0$ and $f = f(x,y)$ is a function of two variables that figures in the statement of condition (1) in an anticipated analogue of Theorem 3.1 that would seek to play a role in a reflection-property characterization of all (rather than just the "top half") of the above-mentioned parabola. As before, let us assume that $f(0) = 0$, let Γ denote the graph of f, and fix a point $P_0(x_0,y_0)$ on Γ. We will apply our earlier approach by considering the bound vector

$$\mathcal{T} := \frac{\partial f}{\partial y}(x_0,y_0)\,\mathbf{i} - \frac{\partial f}{\partial x}(x_0,y_0)\,\mathbf{j}.$$

To motivate our study of this bound vector, note that courses on second-year calculus or advanced calculus usually mention the following fact (cf. [12, Theorem II, page 277]): if a surface $\mathcal{S}$ in $\mathbb{R}^3$ is the graph of the equation $F(x,y,z) = 0$ for a sufficiently smooth function F of three variables, then the tangent plane at a point (x_0,y_0,z_0) on $\mathcal{S}$ has normal vector

$$\frac{\partial f}{\partial x}(x_0,y_0,z_0)\,\mathbf{i} + \frac{\partial f}{\partial y}(x_0,y_0,z_0)\,\mathbf{j} + \frac{\partial f}{\partial z}(x_0,y_0,z_0)\,\mathbf{k}.$$

Suppose, as above, that we have a reason to believe that the "outward"-pointing unit vector in the direction of a bound vector playing a role analogous to that of $\overrightarrow{PS}$ from Theorem 3.1 (and also Theorem 2.1) is $-\mathbf{i}$. Then the equation resulting from an analogue of the equation in the statement of condition (4) in Theorem 3.1 (with $\mathcal{T}$ playing an analogue of the earlier role of $\overrightarrow{PQ}$) is (after dropping subscripts "$_0$")

$$-\frac{\partial f}{\partial y}(x,y) = \frac{(a-x)\frac{\partial f}{\partial y}(x,y) + y\,\frac{\partial f}{\partial x}(x,y)}{\sqrt{(a-x)^2+y^2}}.$$

This equation can be written, equivalently, in the (possibly more familiar) standard form:

$$y\,\frac{\partial f}{\partial x}(x,y) + (a - x + \sqrt{(a-x)^2+y^2})\frac{\partial f}{\partial y}(x,y) = 0.$$

(We have *not* divided through by y, although that would have produced a somewhat simpler-looking equation, because *that* equation would fail to address the origin.) This is a first-order linear partial differential equation (PDE). It should come as no surprise that *a* solution of this PDE, together with the initial condition $f(0,0) = 0$, is given by $f(x,y) = y^2 - 4ax$. (That happens, both when $a > 0$ and also when $a < 0$. Checking all this carefully comes down to simplifying $|a+x|$ as either $a+x$ or $-(a+x)$ according as to whether $a > 0$ or $a < 0$, bearing in mind that $x \geq 0$ when $a > 0$ and that $x \leq 0$ when $a < 0$.) So, all that it would take to conclude that PDEs can be used to present the desired approach is to find an (existence and) uniqueness theorem for PDEs of the kind I have just described. Some of the literature alleges that such a uniqueness theorem exists, pointing occasionally to classic authorities such as [8] (whose first edition was published in 1885). Much of the literature indicates that any PDE of the kind that we are studying, together with its initial condition, can be solved uniquely (in some neighborhood of (x_0,y_0)) by the method of characteristic curves. Unfortunately, I have not been able to solve explicitly for the characteristic curves corresponding to the PDE that is at hand here, and I am not aware of a precise and rigorous uniqueness theorem which would pertain to it. Hopefully, some expert in the field will be able to resolve this matter easily.

(g) The reader may be moved to ask, in the spirit of Theorem 3.1, whether the classical reflection property of ellipses (resp., of hyperbolas), can be used in a characterization of that kind of planar figure. In a paper in preparation, I will answer that question affirmatively for the case of ellipses. In particular, that paper will obtain what could be considered the analogue (for an elliptic arc) of

Theorem 3.1. However, we will also show that by weakening the assumption that the function f in question and its derivative take only nonzero values at all relevant positive x-values other than the right-hand endpoint of the domain of f, one can exhibit a different, *non-elliptic*, function whose graph has the classical reflection property of ellipses (that is, whose graph satisfies the "elliptic" analogues of conditions (3)-(4) of Theorem 3.1 on some domain of the form $[0,\beta)$ for some $\beta > 0$). It will turn out that for the maximal such β, this "different" solution is unique; and, as one might expect from the literature on conic sections, the graph of this "degenerate" solution is a (straight) line segment (of positive length).

Drucker's result [6, Theorem 1] concerning the reflective properties of planar curves was stated as follows: "A smooth connected plane curve has a reflection property if and only if it is a connected subset of a circle, ellipse, hyperbola, parabola, or straight line." Since our paper in preparation will show that the linear kind of degenerate case does arise when examining the reflection property of ellipses, it seems worthwhile to discuss here whether such degenerate cases can arise from an examination of the reflective property of parabolas. We devote the next three paragraphs to that discussion. In attending to Theorem 3.1, that discussion will begin by addressing only graphs that lie in the first quadrant of the Euclidean plane. By tweaking those comments, one obtains reformulated comments that hold in regard to graphs in each of the other three quadrants.

Let us consider an analogue of Theorem 3.1 for a more general function f. Assume that f is differentiable on $(0,\beta)$ for some real number $\beta > 0$, f is continuous at both 0 and β, $f(0) = 0$, and $f^{'}(x) \geq 0$ for all $x \in (0,\beta)$. Also assume that the (equivalent) conditions (3) and (4) hold for all points P on the graph of f in the first quadrant. As in the proof of Theorem 3.1, those "reflective" properties show that the coordinates of any point $P(x,y)$ on the graph of f (with $x > 0$ and $y = f(x)$) satisfy

$$x = \frac{-ax + x^2 + xy(\frac{dy}{dx})}{\sqrt{(a-x)^2 + y^2}}.$$

However, if $y = 0$, one *cannot* then obtain the expression for $\frac{dy}{dx}$ that was inferred in the proof of Theorem 3.1. It is precisely this kind of situation (that is, where $P(x,y)$ is on the graph of the relevant f with $x > 0$ and $y = 0$) that will lead to the above-mentioned degenerate linear case for the "elliptic" reflection property in our paper in preparation. One is thus led to ask what *can* be inferred from the last displayed equation when $y = 0$ (and $x > 0$). At such a point, that displayed "ODE" is simply the algebraic equation

$$x = \frac{-ax + x^2}{\sqrt{(a-x)^2}},$$

with $x > 0$ and $x \neq a$; equivalently, $|a - x| = x - a > 0$; equivalently, $x > a$. We show next that this situation cannot actually arise. Indeed, suppose, on the contrary, that such x exists. Then $a < x < \beta$, and so

$$x_0 := \inf(\{x \in \mathbb{R} \mid x > 0 \text{ and } f(x) = 0\})$$

is well defined and satisfies $x_0 \geq a > 0$. Note that by the proof of Theorem 3.1, $f(x) = 2\sqrt{a}\sqrt{x}$ for all x such that $0 < x < x_0$. Since f is continuous at x_0, we thus have

$$f(x_0) = \lim_{x \to x_0^-} f(x) = \lim_{x \to x_0^-} 2\sqrt{a}\sqrt{x} = 2\sqrt{a}\sqrt{x_0} > 0.$$

Therefore, also by the continuity of f at x_0 (and the definition of limit), there exists $\varepsilon > 0$ such that $f(x) > f(x_0)/4 > 0$ for all x such that $x_0 \leq x \leq x_0 + \varepsilon$. Hence, $f(x) > 0$ for all $x \in (0, x_0 + \varepsilon]$. So, by the definition of x_0, we have

$$x_0 + \varepsilon \leq x_0,$$

whence $\varepsilon \leq 0$, the desired contradiction.

We have just proven that no degenerate "linear" case can arise for the reflective property of parabolas for the contexts that were studied in Theorem 3.1 and in parts (b) and (c) of this remark, *if* one assumes that the domain of the relevant function contains $[0,a]$ or $[-a,0]$ (or both). However, there is no *a priori* reason to suppose that either of those intervals would/should be a subset of such a domain. Indeed, the reasoning in the preceding paragraph shows that if a function f whose domain is an interval which is a subset of $[0,\infty)$ has any serious possibility of providing a linear degenerate case satisfying a parabola's reflection property, then the left-hand endpoint of the domain of that function *must* be greater than a. (As discussed in (e), we need only be interested here in closed arcs in the first quadrant, so that left-hand endpoint would be in the domain of f, without loss of generality.) With that motivation, fix any real number $\beta > a$ (> 0), and let $f_\beta : [\beta,\infty) \to \mathbb{R}$ be the function that is identically zero on the domain $[\beta,\infty)$. Of course, the graph Γ_β of f_β is just the interval $[\beta,\infty)$ (when viewed as a subset of $\mathbb{R}^2$). The question arises naturally whether Γ_β satisfies conditions (3)-(5) from the statement of Theorem 3.1. We will argue shortly via vectorial methods that Γ_β *does* satisfy (3) and (4). However, we would prefer not to involve condition (5) in discussing a function whose domain and graph are so different from what was assumed in Theorem 3.1. (Our unease in this regard is not merely a matter of taste. Frankly, the statement of condition (5) was not designed to accommodate such a graph. Indeed, the following two questions arise as stumbling blocks for any such enterprise. Given such a horizontal graph, what physical scientific sense can be made of the reflections referred to in parts (i) and (ii) of (5)? If the statements of (i) and (ii) are deemed to be meaningful for such a graph, is it not plain that those meanings are logically inconsistent with one another?) To explain why Γ_β satisfies (3) and (4), it will first be necessary to define the radian measure of the angle that is "between" two nonzero parallel vectors, $\overrightarrow{u}$ and $\overrightarrow{v}$, which have the same initial point. Although we had no need of defining this concept in the vectorial review in Section 2, it will be needed here. Fortunately, the community has long ago agreed on the appropriate definition, namely: that radian measure is 0 (resp., π) if $\overrightarrow{u}$ and $\overrightarrow{v}$ have the same direction (resp., if $\overrightarrow{u}$ and $\overrightarrow{v}$ have opposite directions). Now, let us see why Γ_β satisfies (3) and (4). It seems reasonable to agree that the "outside" of Γ_β is simply $\mathbb{R}^2 \setminus \Gamma_\beta$. So, with P any given point on Γ_β, one can choose points S and Q (pertinent to conditions (3) and (4)) as follows: take S to be the point $(-1,0)$ and take Q to be any point on the x-axis other than P. Then the angle between $\overrightarrow{PS}$ and $\overrightarrow{PQ}$ is the same as the angle between $\overrightarrow{PF}$ and $\overrightarrow{PQ}$ (that is, the radian measures of those angles have the same cosine values) simply because the unit vector in the direction of $\overrightarrow{PS}$ is the same as the unit vector in the direction of $\overrightarrow{PF}$ (regardless of whether the choice of Q entails the unit vector in the direction of $\overrightarrow{PQ}$ to be $\mathbf{i}$ or $-\mathbf{i}$). Accordingly, we conclude that f_β (together with Γ_β) gives a linear degenerate case example that satisfies the reflection property of a parabola. By now allowing β to run through the interval (a,∞), we thus get uncountably infinitely many pairwise distinct examples of linear degenerate cases that each satisfy the reflection property of the parabola with focus $(a,0)$ and directrix $x = -a$. It is just as easy to construct uncountably infinitely many pairwise distinct examples of linear degenerate cases, featuring an identically zero function f^*_α with domain $(-\infty,\alpha]$ and graph Γ^*_α, that each satisfy the reflection property of the parabola with focus $(a,0)$ and directrix $x = -a$, given $a < 0$: simply let α run through the interval $(-\infty,-a)$.

Motivated in part by (e), one can now ask whether there exists a function f whose domain contains both negative numbers and positive numbers such that its graph $\Gamma(f)$ satisfies the reflection property of parabolas while $\Gamma(f)$ is not at all parabolic. The answer is certainly in the affirmative. Indeed, one can produce uncountably infinitely many pairwise distinct such examples, as follows. Let $a > 0$, choose real numbers β and α such that $\beta > a$ and $\alpha < -a$, and define the function

$$f : (-\infty,\alpha] \cup [\beta,\infty) \to \mathbb{R}, \text{ (with graph } \Gamma(f),)$$

by specifying that $f|_{(-\infty,\alpha]} = f^*_\alpha$ and $f|_{[\beta,\infty)} = f_\beta$. Observe that $\Gamma(f)$ is the (disjoint) union of Γ^*_α and Γ_β.

Although we showed in the preceding paragraph that Γ_β and Γ_α^* each satisfy the reflection properties (3) and (4) of a parabola, it seems that we *cannot* conclude that $\Gamma(f)$ is the graph of a degenerate case example with the reflection property of a parabola. The difficulty, which seems insuperable, is that Γ_α^* and Γ_β satisfied the reflection property of *different parabolas* (one of which had its focus on the negative x-axis and the other of which had its focus on the positive x-axis). It seems that there is no parabola whose reflection property is shared by both Γ_α^* and Γ_β. Thus, while I find the above function f to be interesting, it seems that its graph does not present a degenerate case of the kind that should have been asked for at the start of this paragraph. (Note that the question that was asked there mentioned "of parabolas", not "of a parabola".) That is, perhaps, for the best, since the statement of [6, Theorem 1] (which was recalled above) considered only the functions with connected graphs which satisfy the reflective properties pertinent to a parabola, an ellipse or a hyperbola (and the graph of f, while being piecewise linear, is visibly not connected!). Lastly, let us consider the following definition that was given by Drucker at the start of [6]:

"Say that a smooth connected plane curve $\mathcal{C}$ has a *reflection property* if there are points F and $F^{'}$, not on $\mathcal{C}$ and not necessarily distinct, such that the tangent line at any point P of $\mathcal{C}$ bisects one of the pairs of opposite angles formed by the intersection of the lines joining P to F and $F^{'}$. The 'foci' F and $F^{'}$ are allowed to be points at infinity, provided they are not the *same* point at infinity."

I would suppose that by "opposite angles", Drucker meant a pair of angles that I was taught to call "supplementary angles" (that is, a pair of angles often nowadays said to "form a linear pair"). One may ask whether, for parabolas or linear degenerate cases of parabolas, this reflection property (as stated by Drucker) is equivalent to the reflection properties stated in conditions (3) and (4) of Theorem 3.1 (cf. also Theorem 2.1). To encourage comparison of this work with [6], I leave it to the reader to answer that question. This completes the remark.

I do not think that a serious mathematician or a serious student of mathematics can be reminded too often of the importance of characterization results in mathematics. Accordingly, I will close with some comments along those lines. Perhaps Remark 3.3 should be (sub)titled "In praise of characterization results and classification results."

Remark 3.3. It is fair to say that our modern understanding that mathematics needs to rely heavily on the axiomatic method came from the discoveries in the 19th century by Hamilton of the real quaternions $\mathbb{H}$ (a *noncommutative* division ring that contains $\mathbb{R}$!) and by Bolyai and Lobachevsky of hyperbolic (plane) geometry (a geometry that satisfies *all but one* of Euclid's axioms for plane geometry!). It was natural to ask what other discoveries of that kind were lying ahead. Put differently, one could ask the following two questions. Up to a suitable notion of isomorphism, can you list all the possible division rings that are finite-dimensional as algebras over $\mathbb{R}$? Up to a suitable notion of isomorphism, can you list all the "neutral geometries", that is, the plane geometries which satisfy all of Euclid's axioms for plane geometry except possibly what has come to be called the "parallel postulate" (as formulated by Playfair)? Frobenius gave the answer for the first of these questions: $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ (with no redundancies in the list). Several mathematicians (working over several decades during the late 19th and early 20th centuries) gave the answer for the second of these questions (cf. [1]): Euclidean (plane) geometry and hyperbolic geometry (with no redundancies in the list). These characterization results could also be thought of as classification results and, as such, one commonly uses such facts as practical tools on a daily basis. For instance, Frobenius' Theorem allows us to conclude that if we confront a noncommutative division ring Δ that is a finite-dimensional $\mathbb{R}$-algebra, then Δ must be $\mathbb{R}$-algebra isomorphic to $\mathbb{H}$. Similarly, if we think that we may have our hands on a neutral geometry $\mathcal{G}$ in which some point P and some line L not passing through P are such that there are *exactly* two distinct lines which pass through P and are parallel to L, then we should look for at least one error in our work (because no such $\mathcal{G}$ can be isomorphic to either Euclidean plane geometry or hyperbolic geometry). Readers of this journal may have seen one of my recent papers which used the fact that the upper (Euclidean) half-plane gives a model for hyperbolic geometry to produce an accessible proof of the known result that hyperbolic geometry (with angle measurement normalized as in [9, Theorem 4.3 (A)]) has a realizability theorem (stating that, for any positive real numbers α, β and γ, possibly listed with repetition, such that $\alpha + \beta + \gamma < \pi$, there exists a hyperbolic triangle (which is unique up to congruence) whose interior angles have radian measures α, β and γ).

The above examples from geometry and algebra indicate that characterization results often take the form of classification results. I believe that here is no genuine difference between the two concepts, as context often can be used to determine the appropriate designation. Consider the Fundamental Theorem of Finite Abelian Groups (in short, the FTFAG). As the reader likely knows, this result describes a couple of ways of listing, up to isomorphism and irredundantly, all the abelian groups of a specific finite cardinality. For instance, it states that, up to isomorphism, the only abelian groups of order (that is, of cardinality) 4 are $\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ and that these two groups are not isomorphic. (Since 4 is the square of a prime number, one should perhaps also note that any group of order 4 is abelian, but that fact will play no role in this discussion.) Suppose that one has at hand an abelian group G of order 4 which is known to have no element of order 4. By the FTFAG, G must be isomorphic to $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. (Of course, there are easier ways to prove that!) This proof used the fact that the "characterization result" FTFAG can be viewed as a classification result (which, as applied, supplied two candidates, and we eliminated one of the candidates on the basis of what was known about G). On the other hand, any characterization result can be thought of a classification result. Indeed, if properties $\mathcal{P}$ and $\mathcal{Q}$ are known to be equivalent for all the objects in a universe $\mathcal{U}$ (some would call that assumption a "characterization result"), then anyone who is comfortable with property $\mathcal{Q}$ would accept an irredundant list of the objects in $\mathcal{U}$ which satisfy $\mathcal{Q}$ as constituting/proving a classification result describing the objects in $\mathcal{U}$ that satisfy $\mathcal{P}$. Sometimes, it is not feasible to produce such a list. For instance, consider the (characterization) result that a metric space is compact if and only if it is complete and totally bounded. The above discussion of properties $\mathcal{P}$ and $\mathcal{Q}$ make it clear how to view this characterization result as a classification result in principle, although no one would relish the prospect of using the ZFC foundations to produce a well-ordered list of, up to isomorphism, all compact metric spaces. Nevertheless, the just-mentioned characterization result can be used to make important conclusions. For instance, it can be used to conclude that the Cantor set is a compact metric space. (Of course, there are other ways to prove that!) It could also be used to conclude that a metric space which is known to not be totally bounded cannot be compact.

By the way, not every result that characterizes or classifies deserves to be called a "characterization result" or a "classification result." For instance, one can prove that the sums of two (possibly equal) non-negative integers that add to a preassigned positive even integer $n = 2m \geq 8$ can be listed (up to order of summands) irredundantly as

$$0 + n, 1 + (n-1), 2 + (n-2), \ldots, (m-1) + (m-1), m + m,$$

that is, as $\{a + (n-a) \mid a \in \mathbb{N} \text{ and } 0 \leq a \leq m\}$, but no one would consider this triviality to be a "result", let alone a "characterization result" or a "classification result." The decision to make (or not to make) any such a designation is subject to change as our understanding of the mathematical world grows and changes. As for the wag who would say that all mathematical truth consists of mere trivialities, I will not waste your time or mine on such sophistry.

Next, turning to algebra, I recall the paper [14] and the characterization result in it that accelerated what has been called "the invasion of homological algebra into ring theory." Recall that a commutative Noetherian (quasi-)local ring (with unity) R with maximal ideal M was classically called a *regular local ring* if M could be generated as an ideal of R by a set of cardinality $\dim(R)$ (which denotes the necessarily finite Krull dimension of R). It is not especially difficult to prove that any regular local ring has finite global dimension. (In fact, that was one of the facts that I was asked to prove during the oral examination in April 1967 where my eventual success entitled me to begin my doctoral research.) The converse had been conjectured for some time but had only been proved in low dimensions. In [14], Serre proved the converse; that is, Serre proved that any commutative local Noetherian ring of finite global dimension must be a regular (local) ring. The resulting homological characterization of (commutative Noetherian) regular local rings led quickly to the result (also in [14], also something that been known only in low dimensions) that the localization of any regular local ring at any prime ideal must also be a regular local ring. It is no exaggeration to say that the characterization result in [14] was a principal impetus (along with Serre's "Géometrie algébrique et géometrie analytique" paper, Grothendieck's "Sur quelques points d'algèbre homologique" paper, ...) for the subsequent reshaping and melding of algebraic geometry and number theory that is apparent in our times.

By the way, a characterization result can lead to pedagogic innovations and new theoretical insights. I recall next a case where this happened. Serre's characterization result that was mentioned in the preceding paragraph led Kaplansky (10 years later) to ask essentially the following: would the theory surrounding that result change significantly if one were to change the definition of a regular local ring, by *defining* a commutative Noetherian local ring to be regular if it is of finite global dimension? Carrying out that investigation was a major part of Kaplansky's famous Queen Mary lecture notes [10]. Surveying the results in the preface to [10], Kaplansky modestly, and with circumspection, expressed the opinion that "The verdict here is that [the experiment of defining a Noetherian commutative local ring to be regular if it is of finite global dimension] is entirely feasible. Somewhat unpredictably certain things became easy to prove while others stubbornly resist." Note that the experience of writing [10] led Kaplansky to write (two subsequent editions of) the famous book, "Commutative Rings," which I have been pleased to use as a textbook in many courses over the years.

Although the appearance of [10] did not lead the algebraic community to change the generally accepted definition of a Noetherian local ring, I turn next to some history where a characterization result of a concept that is fundamental in topology (and, some would add, also fundamental in analysis and algebra) *did* lead the community to change the generally accepted definition of that concept. The concept is that of compactness (which is sometimes referred to as "covering compactness"). More precisely, a topological space is said to be *compact* if each open cover of this space has a finite subcover. The reader is probably aware that the subject of topology was, just like every other field of mathematics, not born overnight. Indeed, the very name of the field changed as the field went through its delivery/growing pains. Perhaps the most famous losing contender for the eventual name of the field was "analysis situs" (roughly translated as "analysis of position"), a name that correctly indicates that the early impetus for developing the field of topology came from geometry and, to some extent, analysis (and differential equations and ...). During the earliest developments of topology, the field of analysis had considerable influence on all of mathematics. (To this day, many research universities in my country find that a majority of their mathematics faculty are either analysts or workers in adjacent fields who could be considered "applied analysts.") Then, as now, society felt the need to produce mathematically literate scientists and engineers (and, to a lesser extent, actual mathematicians). A primary vehicle for providing such training has been a course (or courses) in advanced calculus offered at most universities. For many years (and to some extent, still today), the context for studying a function of several variables (with "several" meaning, in effect, "finitely many, probably more than one") in an advanced calculus course was the Euclidean space (that is, $\mathbb{R}^n$ for some $n \geq 1$). It is well known that a subset of a Euclidean space is a compact space if and only if it is closed (as a subspace of $\mathbb{R}^n$) and bounded. As topology matured and its fundamental concerns were considered in more general contexts, the connection between "compact" and "closed and bounded" began to loosen. It is easy to see that any compact subset of a metric space is closed and bounded. However, a closed and bounded subset of a metric space need *not* be compact. For instance, the closed unit ball in the Hilbert space (hence, metric space) ℓ^∞ is not compact. (This was a historically significant example because the elements of ℓ^∞ are certain sequences of well understood numbers with well understood behavior.) As it became increasingly important/necessary for certain creative mathematicians (and for some others who applied mathematics) to consider infinitely many variables and spaces that seemed intuitively to be of infinitely large "dimension", the needs of such researchers on Hilbert spaces (and, eventually, on more general kinds of topological spaces) prevailed and the community essentially agreed that 'compact" would mean "covering compact", at least in the research literature. It took longer for that consensus to take root in textbooks on intermediate calculus or advanced calculus. Indeed, it was clear to me from reading the footnotes in [2], the textbook for my first course on real analysis, that analysts in North America (perhaps worldwide) preferred to define "compact" as being "closed and bounded" for the first three or four decades of the 20th century, with arguments using the "finite open subcover" approach either absent or relegated to footnotes or appendices. (My limited experience as a student searching through libraries in the 1960s indicated that many textbooks showing the just-mentioned kind of reluctance to change their definition of "compact" also "welcomed" the emergence of Lebesgue integration as a replacement for Riemann integration in a similar grudging fashion.) Indeed, some excellent advanced calculus textbooks of the early 1960s continued to define "compact" as "closed and bounded" but, to their credit, also acknowledged/warned that the definition should not be used when discussing "compact" beyond the context of Euclidean spaces. Although it took quite a while, the community did eventually agree to use the "finite open subcover" approach for "compact." In short, progress won.

One of the textbooks that was mentioned in the preceding paragraph and one of the themes that was mentioned in that paragraph will make return appearances in the next paragraph.

I will close with an example drawn from analysis. While many of my contemporaries learned the fundamentals about real analysis by studying the (1953) first edition of Walter Rudin's textbook, "Principles of mathematical analysis," my first course in that subject (in 1962-63) was more old-fashioned, using the third edition of a text by Carslaw [2] as its official textbook, with the famous books by G. H. Hardy and R. Courant (volume 1 only) as suggested references. The course that I was taught on real analysis *defined* a real number as being a Dedekind cut in the rational numbers; every result (including homework and written answers to examination questions) had to be proven by going back to first principles and using that definition (rather than earlier results) in the proof; and the completeness property of $\mathbb{R}$ was called a "continuity" result, which essentially stated that "every Dedekind cut in the set of real numbers goes through a real number." (At the turn of the last century, some geometers were using the word "continuity" similarly while they endeavored to use real numbers to rigorize the foundations of Euclidean geometry.) In short, while readers of [13] and several other modern axiom-oriented textbooks during the period 1955-65 were encouraging students to learn that an ordered field is isomorphic to $\mathbb{R}$ if and only if it is Dedekind complete (in the sense that every nonempty subset that is bounded above has a least upper bound), students in my class and other readers of books such as [2] were being deprived of this important characterization result. In fact, our education did not even equip us to understand the very statement of that result, let alone its proof, let alone give us access to the opportunities for discovery that are made possible when one has access to that result! In surveying anecdotes such as the above from geometry and topology, I have concluded that a teacher or researcher's preference for emphasizing a specific construction instead of a property possessed by that construction can harm the students or collaborators of that individual, by restricting apparent avenues for investigation. (Algebra and number theory have also had their growing pains in this regard: for instance, while p-adic integers today are fruitfully viewed as inverse limits, I still recall being advised in 1964 to learn about them by reading C. C. MacDuffee's 1940 text, "An introduction to abstract algebra," where such p-adics were "defined" as certain formal series such that .... To be fair, I must admit that Nathan Jacobson's Math Review of MacDuffee's book was accurate in stating that "The book is very readable and the wealth of concrete examples should make it a useful text." But I would also maintain that it is incumbent on instructors and researchers to stay current in their knowledge of their field of activity.) In short, I suggest that we may suffer, as teachers and as researchers, when we fail to unleash the power of the axiomatic method. One antidote for such lethargy has been adopted in recent years by several commutative algebraists, and I recommend it to you as my final offering. Instead of defining a property $\mathcal{P}$ and then proving that it is equivalent to n other properties, why not simply prove the equivalence of $n+1$ properties (namely, what had been called $\mathcal{P}$, along with the other n properties) and then say, as a matter of *definition*, that $\mathcal{P}$ holds if any (hence, all) of these $n+1$ properties hold? That kind of approach may help a student or collaborator to feel free to pursue work from any of the (or all of the) $n+1$ points of view (rather than proceeding solely from what had been the original definition of $\mathcal{P}$). What is the worst that could happen? Some would answer: "Progress."

4 Appendix

This paper is being submitted shortly before I will finish drafting its partner, the above-mentioned sequel about ellipses. In fact, that sequel will also concern hyperbolas. For both of those contexts, that sequel will include, not only analogues of the above Theorem 3.1, but much stronger reflection-theoretic characterization results. I have decided to add a parabolic analogue of those stronger results to this paper. Rather than tamper with the organization of Sections 1-3 of this paper, I am placing that analogue into this appendix.

The gist of the main result of this appendix is that any nontrivial parabolic arc (no matter how "tiny" it may be) can reveal the unique parabola of which it is a subset and a Cartesian equation for that parabola. For simplicity (but with no loss of generality), we will address, in the spirit of Theorem 3.1, a (potentially) parabolic arc which is a subset of the first quadrant.

The main result of this appendix generalizes (Theorem 3.1 and) **Remark 3.2 (e) by** (replacing the domain $[0,\infty)$ with the interval $[\alpha,\beta]$, where $\alpha \in \mathbb{R}$ and $0 \le \alpha < \beta \le \infty$, and by also) **removing any assumption as to the value of $f(\alpha)$ (and if $\beta \in \mathbb{R}$, we also remove any assumption as to the value of $f(\beta)$). Note that we *do* retain the hypothesis that f is continuous at α (and also continuous at β if $\beta \in \mathbb{R}$). We will show that under these conditions** (that is, the hypotheses in the second and third paragraphs of Remark 3.2 (e), including restriction to the case $a > 0$, as just modified here in this paragraph), **the graph of f is a (connected) subset of the graph of the parabola given by $y^2 = 4\delta x + \gamma$ for some real numbers $\delta > 0$ and γ.** As the preceding equation can be written as

$$y^2 = 4\delta(x - \frac{-\gamma}{4}),$$

this parabola has the expected shape; that is, it "opens to the rightŤ and its vertex, $(-\gamma/4, 0)$, is on the x-axis.

To explain how "∞" can be considered the right-hand endpoint of a closed interval, recall that I am using the following definitions for such interval notation:

$$\text{For } \alpha \in \mathbb{R}, \text{ let } (\alpha,\infty] := (\alpha,\infty) \text{ and } [\alpha,\infty] := [\alpha,\infty).$$

Proof. As in the proofs of Theorem 3.1 and Remark 3.2 (e), the assumed reflection-theoretic behavior leads to an ODE which, after certain changes of variables (w and z) and some indefinite integrations, leads to a constant of integration C such that

$$2\sqrt{z} = 2x + C \text{ whenever } \alpha < x < \beta.$$

As f is continuous, the just-displayed equation also holds for $x = \alpha$ (and for $x = \beta$ if $\beta \in \mathbb{R}$). By applying the limiting process $\lim_{x\to\alpha^+}$ and using the definitions of the variables z and w, we get

$$2\sqrt{f(\alpha)^2 + (a-\alpha)^2} = 2\alpha + C.$$

Note that if $\alpha = 0 = f(\alpha)$ as in Theorem 3.1 and Remark 3.2 (e), the just-displayed equation (in conjunction with the assumption that $a > 0$) would give that $C = 2a$. However, the present assumptions *do* give that

$$C = 2\sqrt{f(\alpha)^2 + (a-\alpha)^2} - 2\alpha.$$

It follows that, whenever $\alpha < x < \beta$, we have

$$\sqrt{z} = x + \sqrt{f(\alpha)^2 + (a-\alpha)^2} - \alpha.$$

By once again using the definitions of the variables z and w, we get that if $\alpha < x < \beta$, then

$$\sqrt{f(x)^2 + (a-x)^2} = x + \sqrt{f(\alpha)^2 + (a-\alpha)^2} - \alpha.$$

Squaring both sides of the last displayed equation leads, after some routine (but, frankly, somewhat tiresome) algebraic simplification to the desired equation, $y^2 = 4\delta x + \gamma$, where

$$\delta := \frac{a - \alpha + \sqrt{f(\alpha)^2 + (a-\alpha)^2}}{2} \text{ and}$$

$$\gamma := f(\alpha)^2 + 2\alpha^2 - 2a\alpha - 2\alpha\sqrt{f(\alpha)^2 + (a-\alpha)^2}.$$

It remains only to prove that $\delta > 0$. As it is easy to see that $\delta \geq 0$, it will suffice to prove that $\delta \neq 0$.

Suppose, on the contrary, that $\delta = 0$; equivalently, that $f(\alpha) = 0$ and $\alpha > a$. It is now not difficult to conclude that $f|_{(\alpha,\beta)}$ is a constant function, but knowing the above value of γ makes this easier to see. Indeed,

$$\gamma = 0^2 + 2\alpha^2 - 2a\alpha - 2\alpha\sqrt{0^2 + (a-\alpha)^2} =$$

$$2\alpha^2 - 2a\alpha - 2\alpha \cdot |a-\alpha| = 2\alpha^2 - 2a\alpha - 2\alpha(\alpha - a) = 0.$$

Thus, for all $x \in (\alpha, \beta)$, $f(x)^2 = \sqrt{4\delta x + \gamma} = \sqrt{(4 \cdot 0)x + 0} = 0$. Therefore, $f|_{(\alpha,\beta)} : (\alpha,\beta) \to \mathbb{R}$ is an identically zero function. Hence, $f^{'}(x) = 0$ for all $x \in (\alpha, \beta)$. This contradicts (at least) two of the assumptions that were inherited from Remark 3.2 (e), namely: that $f^{'}(x) \neq 0$ for all $x \in (\alpha, \beta)$; and that f is a strictly increasing monotonic function. The proof is complete.

References

[1] K. Borsuk and W. Szmielew, Foundations of geometry, North Holland, Amsterdam, 1960.

[2] H. S. Carslaw, Introduction to the theory of Fourier's series and integrals, third edition, revised and enlarged, Macmillan, London, 1930.

[3] D. E. Dobbs, Three dimensional vector geometry, M. A. thesis, University of Manitoba, Fort Garry, Manitoba, Canada, 1965.

[4] D. E. Dobbs, Determining the angles between two lines, delta-K Math. J., 41 (2) (2004), 4–9.

[5] D. E. Dobbs and J. C. Peterson, Precalculus, Wm. C. Brown, Dubuque, 1993.

[6] D. Drucker, Euclidean hypersurfaces with reflection properties, Geom. Dedicata 33 (3) (1990), 325–329; Correction to Euclidean hypersurfaces with reflection properties, Geom. Dedicata 39 (3) (1991), 361–362.

[7] D. Drucker, Reflection properties of curves and surfaces, Math. Mag. 65 (3) (1992), 147–157.

[8] A. R. Forsyth, A treatise on differential equations, reprint of the sixth (1929) edition, Dover Publications, Inc., Mineola, NY, 1996.

[9] M. Greenberg, Euclidean and non-Euclidean geometries: development and history, second edition, W. H. Freeman, San Francisco, 1974.

[10] I. Kaplansky, Commutative rings, Queen Mary College Mathematics Notes, Queen Mary College, London, 1966.

[11] L. Leithold, The calculus with analytic geometry, fifth edition, Harper & Row, New York, 1986.

[12] J. M. H. Olmsted, Advanced calculus, Appleton-Century-Crofts, New York, 1961.

[13] W. Rudin, Principles of mathematical analysis, McGraw-Hill, New York-Toronto-London, 1953.

[14] J.-P. Serre, Sur la dimension homologique des anneaux et des modules noethériens, pp. 175-179, in: Proceedings of the international symposium on algebraic number theory, Tokyo & Nikko, 1955 (Science Council of Japan, Tokyo, 1956).

[15] A. E. Taylor, L'Hospital's rule, Amer. Math. Monthly, 59 (1) (1952), 20–24, doi:10.2307/2307183.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 78-114

ISSN: 2820-7114

Title :

$\phi$-rings from a module-theoretic point of view: a survey

Author(s):

Hwankoo Kim, Najib Mahdou & El Houssaine Oubouhou

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 78–114

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

ϕ-rings from a module-theoretic point of view: a survey

Hwankoo Kim[1], Najib Mahdou[2], and El Houssaine Oubouhou[3]
[1]Division of Computer Engineering, Hoseo University, Asan 31499, Republic of Korea
e-mail: *hkkim@hoseo.edu*
[2]Laboratory of Modelling and Mathematical Structures,
Department of Mathematics, Faculty of Science and Technology of Fez, Box 2202, University S.M. Ben Abdellah Fez, Morocco.
e-mail: *mahdou@hotmail.com*
[3]Laboratory of Modelling and Mathematical Structures,
Department of Mathematics, Faculty of Science and Technology of Fez, Box 2202, University S.M. Ben Abdellah Fez, Morocco.
e-mail: *hossineoubouhou@gmail.com*

Communicated by Ünsal Tekir
(Received 02 September 2023, Revised 10 October 2023, Accepted 13 October 2023)

Abstract. Let R be a commutative ring with a nonzero identity and $\mathrm{Nil}(R)$ be its set of nilpotent elements. Recall that a prime ideal P of R is called divided prime if $P \subset (x)$ for every $x \in R \backslash P$; thus a divided prime ideal is comparable to every ideal of R. In many articles, the authors investigated the class of rings $\mathscr{H} = \{R \mid R$ is a commutative ring and $\mathrm{Nil}(R)$ is a divided prime ideal of $R\}$ (observe that if R is an integral domain, then $R \in \mathscr{H}$). If $R \in \mathscr{H}$, then R is called a ϕ-ring. In this paper, we survey known results concerning ϕ-rings from a module-theoretic point of view.

Key Words: ϕ-ring, ϕ-Prüfer, nonnil-Noetherian, ϕ-Dedekind, ϕ-torsion, ϕ-flat, nonnil-injective, ϕ-von Neumann regular, nonnil-coherent, nonnil-commutative diagram, ϕ-(weakly) global dimension, nonnil-projective.
2020 MSC: Primary 13A15, 13C11, 13D05; Secondary 13A15, 13B02, 3C05, 13C12, 13E15, 13F05.

1 Introduction

Throughout this paper, it is assumed that all rings are commutative with nonzero identity and all modules are unitary. If R is a ring, we denote by $Nil(R)$ and $Z(R)$ the ideal of all nilpotent elements of R and the set of all zero-divisors of R respectively. A ring R is called an NP-ring if $Nil(R)$ is a prime ideal, and a ZN-ring if $Z(R) = Nil(R)$. An ideal I of R is said to be a nonnil ideal if $I \not\subseteq Nil(R)$.

Recall from [10, 22] that a prime ideal P of R is said to be divided if it is comparable to every ideal of R, equivalently if $P \subseteq (x)$ for any $x \in R \backslash P$. A ring R is called a divided ring if every prime ideal of R is divided. Recently Badawi, in [6], has studied the following class of rings: $\mathscr{H} = \{R \mid R$ is a commutative ring and $Nil(R)$ is a divided prime ideal of $R\}$. If $R \in \mathscr{H}$, then R is called a ϕ-ring. Moreover, a ZN ϕ-ring is said to be a strongly ϕ-ring. It is easy to see that every integral domain is a ϕ-ring. An ideal I of R is said to be a nonnil ideal if $I \not\subset Nil(R)$. If I is a nonnil ideal of a ϕ-ring R, then $Nil(R) \subseteq I$. Let R be a ring with its ring of quotients $T(R)$ such that $\mathrm{Nil}(R)$ is a divided prime ideal of R. As in [10], we define $\phi : T(R) \to K := R_{\mathrm{Nil}(R)}$ such that $\phi\left(\frac{a}{b}\right) = \frac{a}{b}$ for every $a \in R$ and every $b \in R \backslash Z(R)$. Then ϕ is a ring homomorphism from $T(R)$ into K, and ϕ restricted to R is also a ring homomorphism from R into K given by $\phi(x) = \frac{x}{1}$ for every $x \in R$. Observe that if $R \in \mathscr{H}$, then $\phi(R) \in \mathscr{H}$, $\mathrm{Ker}(\phi) \subseteq Nil(R)$, $Nil(T(R)) = Nil(R)$, $Nil\left(R_{\mathrm{Nil}(R)}\right) = \phi(Nil(R)) = Nil(\phi(R)) = Z(\phi(R))$, $T(\phi(R)) = R_{\mathrm{Nil}(R)}$ is local with maximal ideal $\mathrm{Nil}(\phi(R))$, and $R_{\mathrm{Nil}(R)}/\mathrm{Nil}(\phi(R)) = T(\phi(R))/\mathrm{Nil}(\phi(R))$ is the quotient field of $\phi(R)/\mathrm{Nil}(\phi(R))$. Many well-known notions of integral domains have the corresponding analogues in the class of ϕ-rings, such as valuation domains, Dedekind domains, Prüfer domains, Noetherian domains, coherent domains, Bézout domains and Krull domains. For more on ϕ-rings, see Badawi's survey article [8].

The study of ϕ-rings from a module-theoretic viewpoint was started by Yang [50], who introduced the notion of nonnil-injective modules by replacing the ideals in Baer's criterion for injective modules with nonnil ideals. Dually, Zhao et al. [56] defined the ϕ-flat modules in terms of nonnil ideals and the Tor functor. The present survey is devoted to covering most of the results about ϕ-rings and known results concerning ϕ-rings from the module-theoretic point of view.

2 ϕ-ring properties on some ring constructions

Let A and B be two rings, J an ideal of B and let $f : A \longrightarrow B$ be a ring homomorphism. In this setting, we consider the following subring of $A \times B$:

$$A \bowtie^f J = \{(a, f(a) + j) \mid a \in A \ \ and \ j \in J\}$$

is called the amalgamation of A and B along J with respect to f. This construction is a generalization of the amalgamated duplication of a ring along an ideal denoted by $A \bowtie I$ (introduced and studied by D'Anna and Fontana in [17]). The interest of amalgamation resides, partly, in its ability to cover several basic constructions in commutative algebra, pullbacks and trivial ring extensions. See for instance [18, 19, 26].

For this purpose, we will start with the following theorem which characterizes the case where the amalgamation $A \bowtie^f J$ is an NP-ring. We recall from [26, Proposition 2.20] that

$$Nil(A \bowtie^f J) = \{(a, f(a) + j) \mid a \in Nil(A), j \in J \cap Nil(B)\}.$$

Theorem 2.1. ([34, Theorem 2.32]) Let A and B be rings, J a nonzero ideal of B and $f : A \to B$ be a ring homomorphism. Then

1. Assume that $J \not\subseteq Nil(B)$. Then $A \bowtie^f J$ is an NP-ring if and only if B is an NP-ring and $a \in Nil(A)$ for every $a \in A$ such that $f(a) + j \in Nil(B)$ for some $j \in J$.

2. Assume that $J \subseteq Nil(B)$. Then $A \bowtie^f J$ is an NP-ring if and only if so is A.

Throughout this paper, we will use the technique of trivial ring extensions (idealization) of a module to construct examples. Let A be a ring and M be an A-module. Then $A \ltimes M$, the *trivial* (*ring*) *extension of A by M*, is the ring whose additive structure is that of the external direct sum $A \oplus M$ and whose multiplication is defined by $(r_1, m_1)(r_2, m_2) := (r_1r_2, r_1m_2 + r_2m_1)$ for all $r_1, r_2 \in A$ and all $m_1, m_2 \in M$. A standard notation for the trivial ring extension (idealization) is $A \ltimes M$. The basic properties of trivial ring extensions are summarized in the books [27, 30]. Mainly, trivial ring extensions have been useful for solving many open problems and conjectures in both commutative and non-commutative ring theory. See for instance [4, 32, 33].

Corollary 2.2. *Let A be a ring and E a nonzero A-module. Then, $A \ltimes E$ is an NP-ring if and only if A is an NP-ring.*

The following result characterizes when the amalgamation of rings is a ϕ-ring.

Theorem 2.3. ([25, Theorem 2.1]) Let A and B be two rings, J a nonzero ideal of B, and $f : A \longrightarrow B$ be a ring homomorphism. Set $R := A \bowtie^f J$ and $N(J) := \mathrm{Nil}(B) \cap J$. Then

1. If J is a nonnil ideal of B, then R is a ϕ-ring if and only if $f^{-1}(J) = 0$, A is an integral domain, and $N(J)$ is a divided prime ideal of $f(A)+J$.

2. If $J \subseteq \mathrm{Nil}(B)$, then R is a ϕ-ring if and only if A is a ϕ-ring, and for each $i, j \in J$ and each $a \in A \setminus \mathrm{Nil}(A)$, there exist $x \in \mathrm{Nil}(A)$ and $k \in J$ such that $xa = 0$ and $j = kf(a) + i(f(x)+k)$.

Corollary 2.4. *([25, Corollary 2.5]) Let A and B be two rings, J a nonzero ideal of B, and $f : A \longrightarrow B$ be a ring homomorphism. Assume that $J = \mathrm{Nil}(B)$ and $f^{-1}(J) \subseteq \mathrm{Nil}(A)$. If A and $f(A)+J$ are ϕ-rings, then so is $A \bowtie^f J$.*

Corollary 2.5. *([25, Corollary 2.6]) Let A be a ring. Then, the polynomial ring $A[X]$ is a ϕ-ring if and only if A is an integral domain.*

Let I be a proper ideal of A. The (amalgamated) duplication of A along I is a special amalgamation given by

$$A \bowtie I := A \bowtie^{id_A} I = \{(a, a+i) \mid a \in A, i \in I\}.$$

The next corollary is an immediate consequence of Theorem 2.3 to duplications.

Corollary 2.6. *([25, Corollary 2.7]) Let A be a reduced ring and I be an ideal of A. Then, $A \bowtie I$ is a ϕ-ring if and only if A is a ϕ-ring and $I = 0$.*

Corollary 2.7. *([25, Corollary 2.8]) Let A and B be two rings such that A is a local ring with maximal ideal $M = \mathrm{Nil}(A)$. Let $J \subseteq \mathrm{Nil}(B)$ be an ideal of B. Then, $A \bowtie^f J$ is a ϕ-ring.*

Example 2.8. ([25, Example 2.9]) In general, $A \bowtie^f J$ need not be a ϕ-ring. Indeed, let $f : A \longrightarrow B$ be a surjective homomorphism of integral domains and J a nonzero ideal of B. By Theorem 2.3, $A \bowtie^f J$ is not a ϕ-ring since $f^{-1}(J) \neq 0$.

Example 2.9. ([25, Example 2.10]) Let $A = \mathbb{Z}/8\mathbb{Z}, I = 4\mathbb{Z}/8\mathbb{Z}$, and $B = A/I \cong \mathbb{Z}/4\mathbb{Z}$. Let $f : A \longrightarrow B$ be the canonical surjection and $J = 2\mathbb{Z}/4\mathbb{Z}$ an ideal of B. By using Corollary 2.4, we conclude that $A \bowtie^f J$ is a ϕ-ring.

The next corollary studies when the trivial ring extension is a ϕ-ring.

Corollary 2.10. *([25, Corollary 2.4]) Let A be a ring and E a nonzero A-module. Then, $A \ltimes E$ is a ϕ-ring if and only if A is a ϕ-ring and $E = aE$ for each $a \in A \setminus \mathrm{Nil}(A)$. In particular, If A is an integral domain, then $A \ltimes E$ is a ϕ-ring if and only if E is a divisible A-module.*

Example 2.11. Let D be an integral domain that is not a field and Q its quotient field. Then

1. $D \ltimes Q$ is a strongly ϕ-ring.

2. Set $E = \bigoplus_{i=1}^{\infty} Q/D$. Then the two rings $D \ltimes Q/D$ and $D \ltimes E$ are ϕ-rings but not strongly ϕ-rings.

Recently, Chang and Kim [15] introduced a new pullback : $R_n = D+\theta K[\theta]$ construction. Let D be a domain with K its quotient field. Let $K[x]$ be the polynomial ring in an indeterminate x over $K, n \geq 2$ be an integer and $K[\theta] = K[x]/(x^n)$, where $\theta = x + (x^n)$. Denote by $i : D \hookrightarrow K$ the natural embedding map and $\pi : K[\theta] \to K$ a ring homomorphism satisfying $\pi(f) = f(0)$. Consider the pullback of i and π as follows:

$$\begin{tikzcd} R_n := D+\theta K[\theta] \arrow[r] \arrow[d] & K[\theta] \arrow[d, "\pi"'] \\ D \arrow[r, "i"] & K. \end{tikzcd}$$

Then $R_n = D + \theta K[\theta] = \{f \in K[\theta] \mid f(0) \in D\}$ is a subring of $K[\theta]$. Note that $\mathbb{Z}(R_n) = \mathrm{Nil}(R_n) = \theta K[\theta]$. Then R_n is a strongly ϕ-ring by [15, Proposition 2.1].

3 Nonnil-Noetherian rings

The concept of Noetherian rings is one of the most important topics that is widely used in many areas including commutative algebra and algebraic geometry. The Noetherian property was originally due to the mathematician Noether, who was the first to consider a relation between the ascending chain condition on ideals and the finitely generatedness of ideals. More precisely, she showed that if R is a ring, then the ascending chain condition on ideals of R holds if and only if every ideal of R is finitely generated. This equivalence plays a significant role in simplifying the ideal structure of a ring. Due to the importance of Noetherian rings, many mathematicians have tried to use Noetherian properties in several classes of rings and attempted to generalize the notion of Noetherian rings. Nonnil-Noetherian rings and S-Noetherian rings are typical generalizations of Noetherian rings.

We say that a ring R is nonnil-Noetherian if each nonnil ideal of R is finitely generated. This notion is introduced and studied by Badawi (2003) under the strong hypothesis that the nilradical of the ring is a divided prime ideal. Yang and Liu (2009), Hizem and Benhissi (2011) have extended some properties of nonnil-Noetherian rings but without any assumption on the nilradical.

Theorem 3.1. ([29, Proposition 1.2] and [37, Proposition 2.1]) The following assertions are equivalent for a ring R:

1. R is nonnil-Noetherian;
2. R satisfies the ACC on nonnil ideals;
3. R satisfies the ACC on nonnil finitely generated ideals;
4. For every nonnil ideal I of $R, R/I$ is a Noetherian R-module;
5. Each nonempty set of nonnil ideals of R has a maximal element under set inclusion.

The nonnil-Noetherian rings have an analogue to Cohen's theorem.

Theorem 3.2. ([37, Theorem 2.2] and [29, Proposition 1.2]) A ring R is nonnil-Noetherian if and only if its nonnil prime ideals are finitely generated.

In [7, Proposition 1.10], it is stated that if a ring R is nonnil-Noetherian, then $R/Nil(R)$ is Noetherian. The converse is studied in the following theorem.

Theorem 3.3. ([11, Theorem 2.1]) Let R be a ring. Then

1. If $Nil(R) \notin \mathrm{Spec}(R)$, the following assertions are equivalent:
   (a) R is nonnil-Noetherian;
   (b) R is Noetherian;
   (c) The quotient ring $R/Nil(R)$ is Noetherian and all the minimal prime ideals of R are finitely generated.
2. If $Nil(R) \in \mathrm{Spec}(R)$, then R is nonnil-Noetherian if and only if $R/Nil(R)$ is Noetherian and all the height 1 prime ideals of R are finitely generated.

Corollary 3.4. *([11, Corollary 2.2]) Let R be a ring. Then*

1. *If $Nil(R) \notin \mathrm{Spec}(R)$, then R is Noetherian if and only if R is nonnil-Noetherian.*
2. *If $Nil(R) \in \mathrm{Spec}(R)$, then R is Noetherian if and only if R is nonnilNoetherian and $Nil(R)$ is finitely generated.*

The following theorem studies the stability of the nonnil-Noetherian property when we pass to the polynomial and formal power series ring.

Theorem 3.5. [29, Theorem 3.3] The following assertions are equivalent for a ring R:

1. R is nonnil-Noetherian and the ideal $Nil(R)$ is finitely generated;
2. R is Noetherian;
3. $R[X]$ is Noetherian;
4. $R[X]$ is nonnil-Noetherian;
5. $R[[X]]$ is Noetherian;
6. $R[[X]]$ is nonnil-Noetherian.

Let R be a nonnil-Notherian ring which is not Noetherian. Then the polynomial ring over R is never nonnil-Noetherian. However we have the following two results.

Theorem 3.6. ([11, Theorem 3.8]) A ring R is nonnil-Noetherian if and only if for every nonnil prime ideal P of R, $P[[X]] = P.R[[X]]$.

Theorem 3.7. ([37, Theorem 3.1]) Let R be a ring, n any non-negative integer. Then R is a nonnil-Noetherian ring if and only if the ring $R[x]/\left(x^{n+1}\right)$ is nonnil-Noetherian.

It is well known that $R \ltimes R \cong R[x]/\left(x^2\right)$. Hence, we conclude the following corollary.

Corollary 3.8. *([37, Corollary 3.3]) A ring R is nonnil-Noetherian if and only if the trivial extension $R \ltimes R$ is a nonnil-Noetherian ring.*

The following example shows that the direct sum of nonnil-Noetherian rings need not be nonnil-Noetherian.

Example 3.9. ([37, Example 3.3]) Let S be a Noetherian domain with quotient field K such that $\dim(S) = 1$ and S has infinitely many maximal ideals. Then $R = S \ltimes K$ is a nonnil-Noetherian ring but not Noetherian. Thus $R \oplus R$ is not a nonnil-Noetherian ring.

Let R be a ring. Recall that R is said to be decomposable if R can be written as $R = R_1 \oplus R_2$ for some nonzero rings R_1 and R_2.

Proposition 3.10. *([36, Theorem 2]) Let R be a decomposable ring with identity, and $\{\pi_i\}_{i\in\Lambda}$ the set of canonical epimorphisms from R to each component of decompositions of R. Then the following statements are equivalent.*

1. *R is a Noetherian ring;*
2. *R is a nonnil-Noetherian ring;*
3. *For each $i \in \Lambda$, $\pi_i(R)$ is a Noetherian ring;*
4. *If e is a nonzero nonunit idempotent element of R, then every ideal of R contained in (e) is finitely generated.*

In the case where $R \in \mathscr{H}$, Badawi showed the following result.

Theorem 3.11. ([7, Theorem 2.4]). Let $R \in \mathscr{H}$. The following statements are equivalent:

1. R is a nonnil-Noetherian ring;
2. $R/Nil(R)$ is a Noetherian domain;
3. $\phi(R)/\mathrm{Nil}(\phi(R))$ is a Noetherian domain;
4. $\phi(R)$ is a nonnil-Noetherian ring.

In the following theorem, Badawi showed that there is a nonnil-Noetherian ring with Krull dimension 1 that is not a Noetherian ring.

Theorem 3.12. ([7, Theorem 3.4]) Let R be a Noetherian domain with quotient field K such that $\dim(R) = 1$ and R has infinitely many maximal ideals. Then $D = R \ltimes K \in \mathscr{H}$ is a nonnil-Noetherian ring with Krull dimension one which is not a Noetherian ring. In particular, $\mathbb{Z} \ltimes \mathbb{Q}$ is a nonnil-Noetherian ring with Krull dimension one which is not a Noetherian ring (where $\mathbb{Z}$ is the set of all integer numbers with quotient field $\mathbb{Q}$).

Theorem 3.13. ([40, Corollary 2.16.]) Let A and B be two rings, J a nonzero ideal of B, and let $f : A \to B$ be a ring homomorphism such that $A \bowtie^f J$ is a ϕ-ring. Then the following statements are equivalent:

1. $A \bowtie^f J$ is a nonnil-Noetherian ring;
2. A is a nonnil-Noetherian ring and $f(A) + J$ is a nonnil-Noetherian ring.

It must be noted that the authors of [47] have been studied when $A \bowtie^f J$ is a nonnil-Noetherian ring, and it is shown that if $A \bowtie^f J$ is a ϕ-ring, then $A \bowtie^f J$ is a nonnil-Noetherian ring if and only if A and $f(A) + J$ are nonnil-Noetherian rings and $f^{-1}(J) \subseteq Nil(A)$.

Remark 3.14. Let $f : A \to B$ be a ring homomorphism and J an ideal of B. If $A \bowtie^f J$ is a ϕ-ring, then $f^{-1}(J) \subseteq Nil(A)$ by [25, Lemma 2.3]. Whence our Theorem 3.13 and [47, Theorem 2.7] are identical.

The following example shows that the condition R is a ϕ-ring is a necessary condition in Theorem 3.13.

Example 3.15. ([47, Example 2.10]) Set $A = \mathbb{Z} \ltimes \mathbb{Q}$ and consider the surjective ring homomorphism $f : A \to \mathbb{Z}/6\mathbb{Z}$; $f((n, q)) = \bar{n}$. Consider $J = 3\mathbb{Z}/6\mathbb{Z}$ as an ideal of $\mathbb{Z}/6\mathbb{Z}$. Then, R and $f(A) + J$ are nonnil-Noetherian rings. However, $A \bowtie^f J$ is not.

It is very interesting to note that, in [3], Anderson and Dumitrescu introduced the notion of S-Noetherian rings as a generalization of Noetherian rings. Let R be a ring, S be a multiplicative set of R, and M be an R-module. We say that M is S-finite if there exist a finitely generated submodule F of M and $s \in S$ such that $sM \subseteq F$. Also, we say that M is S-Noetherian if each submodule of M is S-finite. A ring R is said to be S-Noetherian if it is S-Noetherian as an R-module (i.e., if each ideal of R is S-finite). In [36], Kwon and Lim introduced the notion of nonnil-S-Noetherian rings as a generalization of both nonnil-Noetherian rings and S-Noetherian rings. Let R be a ring, S be a multiplicative set of R. Then R is said to be a nonnil-S-Noetherian ring if each nonnil ideal of R is S-finite. If S consists of units of R, then the concept of S-finite ideals is the same as that of finitely generated ideals; so if S consists of units of R, then the notion of nonnil-S-Noetherian rings is identical to that of nonnil-Noetherian rings. Moreover, if R is a reduced ring, then the concept of nonnil-S-Noetherian rings is exactly the same as that of S-Noetherian rings. Obviously, if $S_1 \subseteq S_2$ are multiplicative subsets, then any nonnil-S_1-Noetherian ring is nonnil-S_2-Noetherian; and if S^* is the saturation of S in R, then R is a nonnil-S-Noetherian ring if and only if R is a nonnil-S^*-Noetherian ring. The nonnil-S-Noetherian rings have been studied in [36, 40] using the Cohen-type theorem, the flat extension, the faithfully flat extension, the Eakin-Nagata-Formanek theorem, the polynomial ring extension, the power series ring extension and the amalgamation algebra. For more on Noetherian-like properties, see Benhissi's book [12].

4 ϕ-Prüfer and ϕ-Dedekind

Recall from [9] that a ring $R \in \mathcal{H}$ is called a ϕ-chained ring (ϕ-CR for short) if for each $x \in R_{Nil(R)} \backslash \phi(R)$ we have $x^{-1} \in \phi(R)$.

Proposition 4.1. *([9, Proposition 2.2]) A ring R is a ϕ-CR if and only if for every $a, b \in R \backslash Nil(R)$, either $a \mid b$ in R or $b \mid a$ in R. Hence, if R is a ϕ-CR and $x \in T(R) \backslash R$, then $x^{-1} \in R$.*

Recall that a non-zerodivisor element of a ring R is called a regular element and an ideal of R is said to be regular if it contains a regular element. A ring R is called a Prüfer ring, in the sense of [28], if every finitely generated regular ideal of R is invertible, i.e., if I is a finitely generated regular ideal of R and $I^{-1} = \{x \in T(R) \mid xI \subset R\}$, then $II^{-1} = R$. A Prüfer domain is a Prüfer ring and a homomorphic image of a Prüfer domain is a Prüfer ring. Note that if I is a nonnil ideal of a ϕ-ring R, then $\phi(I)$ is a regular ideal of $\phi(R)$, and so a nonnil ideal I of a ϕ-ring R is called ϕ-invertible if $\phi(I)$ is an invertble ideal of $\phi(R)$. In 2004, Anderson and Badawi [2] extended the notion of Prüfer domains to that of ϕ-Prüfer rings, which are ϕ-rings R satisfying that each finitely generated nonnil ideal is ϕ-invertible.

Theorem 4.2. ([2, Corollary 2.10]). Let $R \in \mathcal{H}$. Then the following statements are equivalent:

1. R is a ϕ-Prüfer ring;
2. $\phi(R)$ is a Prüfer ring;
3. $\phi(R)/Nil(\phi(R))$ is a Prüfer domain;
4. R_P is a ϕ-CR for each prime ideal P of R;
5. $R_P/Nil(R_P)$ is a valuation domain for each prime ideal P of R;
6. $R_M/Nil(R_M)$ is a valuation domain for each maximal ideal M of R;
7. R_M is a ϕ-CR for each maximal ideal M of R.

Recall [31] that a ring R is called an arithmetical ring if R_M is a chained ring for every maximal ideal M of R. Since a chained ring is a ϕ-chained ring, we conclude that if $R \in \mathcal{H}$ is an arithmetical ring, then R is a ϕ-Prüfer ring by Theorem 4.2. Since a ϕ-chained ring need not be a chained ring by [9], we conclude that a ϕ-Prüfer ring need not be an arithmetical ring. However, the following theorem shows that a ϕ-Prüfer ring is a Prüfer ring.

Theorem 4.3. ([2, Theorem2.14]) Let $R \in \mathcal{H}$. If R is a ϕ-Prüfer ring, then R is a Prüfer ring.

If R is a Prüfer ring and $R \notin \mathcal{H}$, then R is not a ϕ-Prüfer ring by definition. The following example shows that for each integer $n \geq 1$, there is a Prüfer ring $R \in \mathcal{H}$ with Krull dimension n which is not a ϕ-Prüfer ring.

Example 4.4. ([2, Example 2.15]) Let $n \geq 1$ and D be a non-integrally closed domain with quotient field K and Krull dimension n. Set $R = D \ltimes (K/D)$. Then $R \in \mathcal{H}$ and R is a Prüfer ring with Krull dimension n which is not a ϕ-Prüfer ring.

However, if R is a strongly ϕ-ring, then the following theorem shows that a Prüfer ring is a ϕ-Prüfer ring.

Theorem 4.5. ([2, Theorem2.16]) Let $R \in \mathcal{H}$ with $Nil(R) = Z(R)$. Then R is a Prüfer ring if and only if R is a ϕ-Prüfer ring.

Recall from [14] that a ring R is called a pre-Prüfer ring if every proper homomorphic image of R is a Prüfer ring, i.e., if R/I is a Prüfer ring for each nonzero proper ideal I of R. Note that the class of Prüfer rings and the class of pre-Prüfer rings are not comparable under set inclusion (cf. [14]).

Theorem 4.6. ([2, Theorem 2.19]) Let $R \in \mathcal{H}$ such that $Nil(R) \neq \{0\}$. Then R is a pre-Prüfer ring if and only if R is a ϕ-Prüfer ring.

The following example shows that the hypothesis $Nil(R) \neq \{0\}$ in Theorem 4.6 is crucial.

Example 4.7. ([2, Example 2.20]) Let D be a Prüfer domain with quotient field F. For indeterminates X, Y, let $K = F(Y)$ and let V be the valuation domain $K + XK[[X]]$. Then V is one-dimensional with maximal ideal $M = XK[[X]]$. Set $R = D + M$. Then $Nil(R) = \{0\}$, and R is a pre-Prüfer ring (domain) which is not a Prüfer ring (domain). Hence R is not a ϕ-Prüfer ring.

It is well-known that a valuation overring of a Prüfer domain R is of the form R_P for some prime ideal P of R. We have a similar result for ϕ-Prüfer rings. Recall that an overring of a ring R is a ring between R and $T(R)$.

Theorem 4.8. ([2, Theorem 2.11]) Let $R \in \mathcal{H}$ be a ϕ-Prüfer ring and let S be a ϕ-chained overring of R. Then $S = R_P$ for some prime ideal P of R containing $Z(R)$.

Observe that if every overring of $R \in \mathcal{H}$ is integrally closed, then R need not be a ϕ-Prüfer ring by Example 4.4. However, we have the following result.

Theorem 4.9. ([2, Theorem 2.17]) Let $R \in \mathcal{H}$. Then R is a ϕ-Prüfer ring if and only if every overring of $\phi(R)$ is integrally closed.

In the following example, Anderson and Badawi show that for each integer $n \geq 1$, there is a (non-domain) ϕ-Prüfer ring with Krull dimension n.

Example 4.10. ([2, Example 2.18]) Let $n \geq 1$ and let D be a Prüfer domain with quotient field K and Krull dimension n. Set $R = D \ltimes K$. Then $R \in \mathcal{H}$ is a (non-domain) ϕ-Prüfer ring with Krull dimension n.

Recall that a ring $R \in \mathcal{H}$ is called ϕ-integrally closed if $\phi(R)$ is integrally closed in $T(\phi(R)) = R_{Nil(R)}$.

Let $R \in \mathcal{H}$. If every nonnil ideal of R is ϕ-invertible, then we say that R is a ϕ-Dedekind ring [1]. The following characterization of ϕ-Dedekind rings resembles that of Dedekind domains.

Theorem 4.11. ([1, Theorems 2.5, 2.10 and Corollary 2.2]) Let $R \in \mathcal{H}$. Then the following statements are equivalent:

1. R is ϕ-Dedekind;
2. $\phi(R)$ is a Dedekind ring;
3. $R/Nil(R)$ is a Dedekind domain;
4. R is nonnil-Noetherian, ϕ-integrally closed, and of dimension ≤ 1;
5. R is nonnil-Noetherian and R_M is a discrete ϕ-chained ring for each maximal ideal M of R.

Theorem 4.12. ([1, Theorem 2.12]) Let $R \in \mathcal{H}$ be a ϕ-Dedekind ring. Then R is a Dedekind ring.

The following is an example of a ring $R \in \mathcal{H}$ which is a Dedekind ring but not a ϕ-Dedekind ring.

Example 4.13. ([1, Example 2.13]) Let D be a non-Dedekind domain with quotient field K. Set $R = D \ltimes K/D$. Then R is a Dedekind ring which is not a ϕ-Dedekind ring.

However, if R is a strongly ϕ-ring, then the following theorem shows that a Dedekind ring is a ϕ-Dedekind ring.

Theorem 4.14. ([1, Theorem 2.14]) Let R be a strongly ϕ-ring. Then R is a Dedekind ring if and only if R is a ϕ-Dedekind ring.

It is well-known that an integral domain R is a Dedekind domain if and only if every nonzero proper ideal of R is (uniquely) a product of prime ideals of R. We have the following result.

Theorem 4.15. ([1, Theorem 2.15]) Let $R \in \mathscr{H}$. Then R is a ϕ-Dedekind ring if and only if every nonnil proper ideal of R is (uniquely) a product of nonnil prime ideals of R.

Theorem 4.16. ([1, Theorem 2.16]) Let $R \in \mathscr{H}$. Then the following statements are equivalent:

1. R is a ϕ-Dedekind ring;
2. Each nonnil proper principal ideal aR can be written in the form $aR = Q_1 \cdots Q_n$, where each Q_i is a power of a nonnil prime ideal of R and the Q_i 's are pairwise comaximal;
3. Each nonnil proper ideal I of R can be written in the form $I = Q_1 \cdots Q_n$, where each Q_i is a power of a nonnil prime ideal of R and the Q_i's are pairwise comaximal.

Recall from [1] that a ring R is called a ZPI-ring if every nonzero proper ideal of R is uniquely a product of prime ideals of R, and R is called a general ZPI-ring if every nonzero proper ideal of R is a product of prime ideals of R. A ring $R \in \mathscr{H}$ is called a nonnil-ZPI-ring if every nonnil proper ideal of R is uniquely a product of (nonnil) prime ideals of R, and R is said to be a general nonnil-ZPI-ring if every nonnil proper ideal of R is a product of (nonnil) prime ideals of R.

Corollary 4.17. *([1, Corolary 2.17]) Let $R \in \mathscr{H}$. Then the following statements are equivalent:*

1. *R is a ϕ-Dedekind ring;*
2. *R is a nonnil-ZPI-ring;*
3. *R is a general nonnil-ZPI-ring.*

Theorem 4.18. ([1, Theorem 2.18]) Let $R \in \mathscr{H}$ be a ϕ-Dedekind ring and let I be an ideal of R. Then

1. If $I \subseteq Nil(R)$, then R/I is a ϕ-Dedekind ring.
2. If I is a nonnil ideal of R, then R/I is a general ZPI-ring.

It is well-known that an integral domain R is a Dedekind domain if and only if every nonzero prime ideal of R is invertible, if and only if R is Noetherian and every nonzero maximal ideal of R is invertible.

Theorem 4.19. ([1, Theorem 2.20]) Let $R \in \mathscr{H}$. Then the following statements are equivalent:

1. R is a ϕ-Dedekind ring;
2. Each nonnil prime ideal of R is ϕ-invertible;
3. R is a nonnil-Noetherian ring and each nonnil maximal ideal of R is ϕ-invertible.

It is well-known that an overring of a Dedekind domain is a Dedekind domain. We end this section with the following result.

Theorem 4.20. ([1, Theorem 2.23]) Let $R \in \mathscr{H}$ be a ϕ-Dedekind ring. Then every overring of R is a ϕ-Dedekind ring.

5 ϕ-torsion modules and ϕ-torsion free modules

Let R be a ring and M be an R-module. Set

$$\phi\text{-}tor(M) = \{x \in M \mid sx = 0 \text{ for some } s \in R \setminus Nil(R)\}.$$

If ϕ-$tor(M) = M$, then M is called a ϕ-torsion module, and if ϕ-$tor(M) = 0$, then M is called a ϕ-torsion free module.

If $Nil(R)$ is a prime ideal, then ϕ-tor(M) is a submodule of M, which is called the total ϕ-torsion submodule of M. Set $T = \phi$-tor (M). Then T is always ϕ-torsion and M/T is always ϕ-torsion free.

Example 5.1. Let R be a ring. Then R/I is a ϕ-torsion R-module for any nonnil ideal I of R.

Every regular ideal is a nonnil ideal, thus every torsion R-module is a ϕ-torsion R-module and every ϕ-torsion free R-module is a torsion free R-module. If R is a strongly ϕ-ring, in the sense that each zero divisor is nilpotent, then every ϕ-torsion R-module is a torsion R-module, and every torsion free R-module is a ϕ-torsion free R-module.

The following results give us a criterion for the ϕ-torsion module, and the ϕ-torsion free module.

Theorem 5.2. ([56, Theorem 2.2]) An R-module M is ϕ-torsion if and only if $Ann_R(x)$ is a nonnil ideal for every $x \in M$.

Theorem 5.3. ([56, Theorem 2.3]) The following statements are equivalent for a module M:

1. M is ϕ-torsion free;
2. $\mathrm{Hom}_R(R/J, M) = 0$ for every nonnil ideal J of R;
3. $\mathrm{Hom}_R(B, M) = 0$ for every nonnil ideal J of R and every R/J-module B.

Theorem 5.4. ([56, Theorem 2.4]) Let R be a ring. Then

1. A module M is ϕ-torsion if and only if $\mathrm{Hom}_R(M, N) = 0$ for any ϕ-torsion free module N.
2. A module N is ϕ-torsion free if and only if $\mathrm{Hom}_R(M, N) = 0$ for any ϕ-torsion module M.

Denote by $\mathcal{T}$ (resp., $\mathcal{F}$) the set of all ϕ-torsion modules (resp., ϕ-torsion free modules). Then by Theorem 5.4 $(\mathcal{T}, \mathcal{F})$ is a (hereditary) torsion theory.

Theorem 5.5. ([56, Theorem 2.6]) Let $f : R \to T$ be a monomorphism from rings R to T. If M is a ϕ-torsion R-module, then $M \otimes_R T$ is a ϕ-torsion T-module.

Corollary 5.6. *([56, Corollary 2.7]) If M is a ϕ-torsion R-module, then $M[x] = M \otimes_R R[x]$, as an $R[x]$-module, is also a ϕ-torsion module.*

Proposition 5.7. *([55, Proposition 2.4]) Let R be a ring with prime nil radical and $0 \to A \to B \to C \to 0$ be an exact sequence of R-modules. Then B is ϕ-torsion if and only if A and C are both ϕ-torsion. Moreover, $\bigoplus_{i\in\Gamma} M_i$ is a ϕ-torsion module if and only if each M_i is a ϕ-torsion module.*

A valuation domain is an integral domain such that for any two elements r and s, either r divides s or s divides r. The author in [49] showed that a finitely presented module over a valuation domain is a direct sum of cyclically presented modules. Similarly, Zhao proved the following result.

Theorem 5.8. ([55, Theorem 4.1]) A finitely presented ϕ-torsion module over a ϕ-CR is a direct sum of cyclically presented ϕ-torsion modules.

6 ϕ-flat modules

An R-module M is said to be ϕ-flat if for every monomorphism $f : A \to B$ with ϕ-torsion $\mathrm{Coker}(f)$, $f \otimes \mathbf{1} : A \otimes_R M \to B \otimes_R M$ is also a monomorphism; equivalently, if $0 \to A \to B \to C \to 0$ is an exact R-sequence where C is ϕ-torsion, then $0 \to A \otimes_R M \to B \otimes_R M \to C \otimes_R M \to 0$ is exact.

Due to Zhao, Wang and Tang we have the following characterizations.

Theorem 6.1. ([56, Theorem 3.2]) The following conditions are equivalent for an R-module M:

1. M is ϕ-flat;
2. $\mathrm{Tor}_1^R(P, M) = 0$ for every ϕ-torsion R-module P;
3. $\mathrm{Tor}_1^R(R/I, M) = 0$ for every nonnil ideal I of R;
4. $0 \to I \otimes_R M \to R \otimes_R M$ is an exact sequence for every nonnil ideal I of R;
5. $I \otimes_R M \cong IM$ for every nonnil ideal I of R;
6. $-\otimes_R M$ is exact for every exact R-sequence $0 \to N \to F \to C \to 0$, where N, F, C are finitely generated, C is a ϕ-torsion R-module, and F is free;
7. $-\otimes_R M$ is exact for every exact R-sequence $0 \to N \to F \to C \to 0$, where C is a ϕ-torsion R-module, and F is free;
8. $\mathrm{Tor}_1^R(R/I, M) = 0$ for every finitely generated nonnil ideal I of R;
9. $0 \to I \otimes_R M \to R \otimes_R M$ is an exact sequence for every finitely generated nonnil ideal I of R;
10. $I \otimes_R M \cong IM$ for every finitely generated nonnil ideal I of R;
11. $\mathrm{Ext}_R^1(I, M^+) = 0$ for any nonnil ideal I of R, where M^+denote by the character module $\mathrm{Hom}_Z(M, \mathbb{Q}/\mathbb{Z})$;
12. Let $0 \to K \to F \xrightarrow{g} M \to 0$ be an exact sequence of R-modules, where F is free. Then $K \bigcap FI = IK$ for every nonnil ideal I of R;
13. Let $0 \to K \to F \xrightarrow{g} M \to 0$ be an exact sequence of R-modules, where F is free. Then $K \cap FI = IK$ for every finite generated nonnil ideal I of R.

Every flat R-module is ϕ-flat. If R is a domain, then every ϕ-flat R-module is flat.

We know that the flatness of R-modules is a local property. The following two results imply that the ϕ-flatness is also a local property.

Theorem 6.2. ([56, Theorem 3.4]) Let M be a ϕ-flat R-module, and S be a multiplicative set in the ring R. Then M_S is a ϕ-flat R-module.

Theorem 6.3. ([56, Theorem 3.5]) Let M be an R-module. The following conditions are equivalent:

1. M is a ϕ-flat R-module;
2. M_P is a ϕ-flat R_P-module for each prime ideal P of R;
3. M_m is a ϕ-flat R_m-module for each prime ideal m of R.

We know that the inductive limit of flat modules is a flat module. The following theorem shows that we have the same result for ϕ-flat modules.

Proposition 6.4. *([39, Proposition 2.1]) The inductive limit of ϕ-flat modules is a ϕ-flat module.*

Corollary 6.5. *([39, Corollary 2.2]) Let M be an R-module. If every finitely generated submodule of M is ϕ-flat, then M is ϕ-flat.*

As in the case of flat modules, we have the following theorems.

Theorem 6.6. ([39, Proposition 2.3]) Let M be a ϕ-flat module and $0 \to A \to B \to M \to 0$ be an exact sequence. If A is ϕ-flat, then so is B. Moreover, if $\mathrm{Tor}_2^R(R/I, M) = 0$ for every nonnil ideal I of R, then the converse is true.

Theorem 6.7. ([56, Theorem 3.6]) Let $f : R \to T$ be a surjective ring homomorphism. If M is a ϕ-flat R-module, then $M \otimes_R T$ is a ϕ-flat T-module.

Corollary 6.8. *([56, Corollary 3.7]) Let M be a ϕ-flat R-module and I be an ideal of R. Then M/IM is a ϕ-flat R/I-module.*

Theorem 6.9. ([56, Theorem 3.8]) Let R be a ϕ-ring, M be an R-module and I be an ideal of R. Suppose that $I \subseteq Nil(R)$ and $I \otimes_R M \cong IM$. Then M is a ϕ-flat R-module if and only if M/IM is a ϕ-flat R/I-module.

Proposition 6.10. *([34, Proposition 2.10]) Let R be a ϕ-ring and let I be a nonnil ideal of R. Then I is ϕ-flat over R if and only if $I/Nil(R)$ is flat over $R/Nil(R)$.*

Recall that a flat module is torsion-free. If R is an NP-ring with $Nil(R) \subsetneq Z(R)$, then R is a flat R-module, which means that it is ϕ-flat but not ϕ-torsion free.

Proposition 6.11. *([39, Proposition 2.4]) Let R be a ZN-ring. Then every ϕ-flat R-module is ϕ-torsion free.*

For rings in which every ideal is ϕ-flat, we get the following result.

Theorem 6.12. ([39, Theorem 2.7]) Let R be a ring. If all ideals of R are ϕ-flat, then all ideals of $R/Nil(R)$ are flat.

The converse of the previous theorem is not always true, as the following example shows.

Example 6.13. ([39, Example 2.8]) Let $R = \mathbb{Z} \ltimes \mathbb{Z}/2\mathbb{Z}$. Then every ideal of $R/Nil(R)$ is flat, but the ideal $I = 0 \times \mathbb{Z}/2\mathbb{Z}$ is not ϕ-flat.

In 2018, Zhao [55] gave the following homological characterization of ϕ-Prüfer rings.

Theorem 6.14. ([55, Theorem 4.3]) Let R be a strongly ϕ-ring. Then the following statements are equivalent.

1. R is a ϕ-Prüfer ring;
2. All ϕ-torsion free R-modules are ϕ-flat;
3. Each submodule of a ϕ-flat R-module is ϕ-flat;
4. Each nonnil ideal of R is a ϕ-flat R-module;
5. Each finitely generated nonnil ideal of R is a ϕ-flat R-module;
6. If M is a ϕ-torsion R-module and N is a ϕ-torsion free R-module, then $\mathrm{Tor}_1^R(M, N) = 0$;

7. If M is a ϕ-torsion R-module and I is a nonnil ideal of R, then $\mathrm{Tor}_1^R(M,I)=0$;

8. If M is a ϕ-torsion R-module and I is a finitely generated nonnil ideal of R, then $\mathrm{Tor}_1^R(M,I)=0$.

It is well-known that a domain D is a Prüfer domain if and only if every ideal of D is flat. Kim, Mahdou and Oubouhou [34] provided the following characterization of a ϕ-ring in which every ideal is ϕ-flat.

Theorem 6.15. ([34, Corollary 2.8]) Let R be a ϕ-ring. Then every ideal of R is ϕ-flat if and only if R is a ϕ-Prüfer ring with $Z(R)=Nil(R)$.

Recall that an R-module M is called P-flat if, for every $(s,x)\in R\times M$ such that $sx=0, x\in(0:s)M$. In this last part of this section, we investigates the notion of ϕ-P-flat modules.

Definition 6.16. An R-module M is called ϕ-P-flat if for any $s\in R\backslash\mathrm{Nil}(R)$ and $m\in M$ such that $sx:=0, x\in(0:s)M$.

Theorem 6.17. ([39, Theorem 3.2]) Let M be an R-module. Then the following conditions are equivalent:

1. M is a ϕ-P-flat R-module;

2. The canonical map: $M\otimes_R Ra\to M\otimes_R R$ is injective for every $a\in R\backslash Nil(R)$;

3. $Ra\otimes_R M\cong aM$, for every $a\in R\backslash\mathrm{Nil}(R)$;

4. $\mathrm{Tor}_1^R(M,R/Ra)=0$ for every $a\in R\backslash\mathrm{Nil}(R)$;

5. For every $a\in R\backslash\mathrm{Nil}(R)$, every homomorphism from R/aR to M factors through a free R-module;

6. There exists an exact sequence $0\to K\to F\to M\to 0$ with F is free such that for any $a\in R\backslash\mathrm{Nil}(R)$, $Ka=K\cap Fa$;

7. For every exact sequence $0\to K\to F\to M\to 0$ with F is free, $Ka=K\cap Fa$ for every $a\in R\backslash\mathrm{Nil}(R)$.

The class of ϕ-P-flat modules is a generalization of both ϕ-flat and P-flat modules, as shown by the following proposition.

Proposition 6.18. *([39, Proposition 3.3]) Let R be a ring. Then*

1. *Every P-flat module is ϕ-P-flat.*

2. *Every ϕ-flat module is ϕ-P-flat.*

Now we give an example of a ϕ-P-flat module which is not P-flat and an example of a ϕ-P-flat module which is not ϕ-flat.

Example 6.19. ([39, Example 3.4])

1. Let $R=\mathbb{Z}/2\mathbb{Z}\ltimes\mathbb{Z}/2\mathbb{Z}$ and $I=0\ltimes\mathbb{Z}/2\mathbb{Z}$. Then I is ϕ-P-flat which is not P-flat.

2. Let $R=\mathbb{Z}\ltimes\mathbb{Z}$ and $J=0\ltimes 2\mathbb{Z}$. Then J is ϕ-P-flat which is not ϕ-flat.

Recall from [34] that a ring R is called a ϕ-PF-ring if every ideal of R is ϕ-P-flat.

Theorem 6.20. ([34, Theorem 2.2]) The following conditions are equivalent for a ring R.

1. R is a ϕ-PF-ring;

2. Every principal ideal of R is ϕ-P-flat;

3. Every submodule of every ϕ-P-flat R-module is ϕ-P-flat;

4. $\mathrm{Tor}_2^R(N, R/Ra) = 0$ for every R-module N and every $a \in R\backslash \mathrm{Nil}(R)$;

5. Every nonnil principal ideal of R is flat;

6. For every element $x \in R$ and $s \in R\backslash \mathrm{Nil}(R)$ with $sx = 0$, there exists $\alpha \in \mathrm{Ann}(x)$ such that $s = \alpha s$;

7. For every element $x \in R$ and $s \in R\backslash \mathrm{Nil}(R)$ with $sx = 0$, there exists $\alpha \in \mathrm{Ann}(s)$ such that $x = \alpha x$.

Recall that an ideal I of a ring R is said to be pure if for every $x \in I$, there exists $y \in I$ such that $xy = x$.

Corollary 6.21. *([34, Corollary 2.3]) A ring R is a ϕ-PF-ring if and only if $Ann(a)$ is a pure ideal of R for every $a \in R\backslash \mathrm{Nil}(R)$.*

We next give some examples of ϕ-PF-rings. Recall that a ring R is said to be PF-ring if every principal ideal of R is flat.

Example 6.22. ([34, Example 2.4])

1. Every PF-ring is a ϕ-PF-ring.

2. Every ring R with $Z(R) = \mathrm{Nil}(R)$ is a ϕ-PF-ring.

In general, R being a ϕ-PF-ring does not imply that $Z(R) = Nil(R)$. It suffices to consider $R := \mathbb{Z}/6\mathbb{Z}$. Then R is a ϕ-PF-ring by ([34, Remark 2.5]). But $Z(R) = \{0, 2, 3, 4\} \neq \mathrm{Nil}(R) = 0$. Recall that a ring R is said to be présimplifiable if for every $a, r \in R$, $ar = a$ implies $a = 0$ or r is a unit. It is easy to check that any local ring is présimplifiable.

The following corollary shows that if we assume that R is a présimplifiable ring or an NP-ring, we will get the equivalent between the ϕ-PF-rings and the rings R with $Z(R) = \mathrm{Nil}(R)$.

Corollary 6.23. *([34, Corollary 2.6])*

1. *If R is a NP-ring, then R is a ϕ-PF-ring if and only if $Z(R) = \mathrm{Nil}(R)$.*

2. *If R is présimplifiable, then R is a ϕ-PF-ring if and only if $Z(R) = \mathrm{Nil}(R)$.*

We denote by $U(R)$ the set of all units of a ring R. Now we give an example of an ideal which is ϕ-P-flat but which is neither ϕ-flat nor P-flat.

Example 6.24. ([34, Example 2.19]) Let D be a domain which is not a field, and set $R = D \ltimes D$. Then the ideal $J = (0, a)R$, generated by $(0, a)$ with $a \in D\backslash U(D)$, is ϕ-P-flat which is neither ϕ-flat nor P-flat.

7 Nonnil-injective modules

The notion of nonnil injectivity was introduced and studied by the authors of [53, 50]. Recall that an R-module M is said to be ϕ-torsion, if every $x \in M$, there exists $s \in R\backslash Nil(R)$ such that $sx = 0$.

Definition 7.1. ([53, Definition 1.1]) Let R be a ring. An R-module E is said to be nonnil injective if given every diagram of R-modules and homomorphisms

$$\begin{tikzcd}
0 \arrow[r] & X \arrow[r, "g"] \arrow[d, "f"] & Y \arrow[dl, dashed, "h"] \\
& E &
\end{tikzcd}$$

with the above row exact, where $Cokerg$ is a ϕ-torsion module, there is a homomorphism $h : Y \to E$ making this diagram commute (i.e., $hg = f$). In particular, every injective module is nonnil injective.

Theorem 7.2. ([53, Theorem 1.2]) Let R be a ring. The following statements are equivalent for an R-module E:

1. E is nonnil injective;
2. Every exact sequence $0 \to E \xrightarrow{f} B \to C \to 0$ is split for every ϕ-torsion R-module C;
3. If $0 \to A \to B \to C \to 0$ is an exact sequence such that C is ϕ-torsion, then the sequence $0 \to \mathrm{Hom}_R(C,E) \to \mathrm{Hom}_R(B,E) \to \mathrm{Hom}_R(A,E) \to 0$ is also exact.

The following theorem characterizes the nonnil injective module by the Bear's criterion.

Theorem 7.3. ([53, Theorem 1.4]) Let R be a ring. An R-module E is nonnil injective if and only if for every nonnil ideal I of R, every homomorphism $f : I \to E$ can be extended to R.

Corollary 7.4. *Let R be a ring and E be an R-module. Then E is nonnil injective if and only if $Ext^1_R(R/I,E) = 0$ for every nonnil ideal I of R.*

The following result is an answer to the question: When are ϕ-flat (resp., nonnil-injective) modules flat (resp., injective)?

Theorem 7.5. ([43, Theorem 1.6] and [35, Theorem 3]) Let $R \in \mathscr{H}$. Then the following assertions are equivalent:

1. R is an integral domain;
2. Every ϕ-flat module is flat;
3. Every nonnil injective module is injective.

Recall that an R-module M is called nonnil-FP-injective provided that $\mathrm{Ext}^1_R(T,M) = 0$ for any finitely presented ϕ-torsion module T. Note that the class of nonnil-FP-injective modules is closed under direct sums, direct products, extensions and pure submodules.

Proposition 7.6. *([43, Proposition 1.8]) Let R be an NP-ring. Then the following assertions are equivalent:*

1. *M is ϕ-flat;*
2. $\mathrm{Hom}_R(M,E)$ *is nonnil-injective for any injective module E;*
3. $\mathrm{Hom}_R(M,E)$ *is nonnil-FP-injective for any injective module E;*
4. *If E is an injective cogenerator, then* $\mathrm{Hom}_R(M,E)$ *is nonnil-injective;*

5. *If E is an injective cogenerator, then $\mathrm{Hom}_R(M,E)$ is nonnil-FP-injective.*

Proposition 7.7. *([43, Proposition 1.4]) Let R be a ϕ-ring and E an $R/Nil(R)$-module. Then E is injective over $R/Nil(R)$ if and only if E is nonnil-injective over R.*

Proposition 7.8. *([43, Proposition 1.5]) Let R be a ϕ-ring and M an FP-injective $R/Nil(R)$-module. Then M is nonnil-FP-injective over R.*

Obviously, any FP-injective module is nonnil-FP-injective. However, the converse characterizes integral domains.

Theorem 7.9. ([44, Theorem 1.6]) Let R be a ϕ-ring. Then the following assertions are equivalent:

1. R is an integral domain;
2. Any nonnil-FP-injective module is FP-injective.

Proposition 7.10. *([43, Proposition 1.8]) Let R be a ϕ-ring. Then R is nonnil-Noetherian if and only if any nonnil-FP-injective module is nonnil-injective.*

The well-known Cartan-Eilenberg-Bass theorem for the Noetherien ring states that a ring is Noetherian if and only if any direct sum of injective modules is injective.

Theorem 7.11. ([37, Theorem 2.5]) The following statements are equivalent:

1. R is a nonnil-Noetherian ring;
2. Every direct sum of nonnil-injective R-modules is nonnil-injective;
3. Every direct sum of injective hulls of simple R-modules is nonnil-injective;
4. For every nonnil-injective R-module E, $\bigoplus_I E$ is nonnil-injective.

Theorem 7.12. ([43, Theorem 2.8]) Let R be a ϕ-ring. Then the following statements are equivalent:

1. R is a ϕ-Dedekind ring and a strongly ϕ-ring;
2. Any divisible module is nonnil-injective;
3. Any h-divisible module is nonnil-injective;
4. Any nonnil ideal of R is projective.

Theorem 7.13. ([43, Theorem 2.13]) Let R be a ϕ-ring. Then the following statements are equivalent:

1. R is a ϕ-Prüfer ring and a strongly ϕ-ring;
2. Any divisible module is nonnil-FP-injective;
3. Any h-divisible module is nonnil-FP-injective;
4. Any finitely generated nonnil ideal of R is projective;
5. Any (finitely generated) nonnil ideal of R is flat;
6. Any (finitely generated) ideal of R is ϕ-flat;
7. Any submodule of a ϕ-flat module is ϕ-flat;
8. Any R-module has an epimorphism ϕ-flat preenvelope;
9. Any R-module has an epimorphism ϕ-flat envelope.

8 ϕ-von Neumann regular rings

Recall that a ring R is said to be von Neumann regular if every R-module is flat. A ring R is called a π-regular if for each $r \in R$ there exist a positive integer n and an element $x \in R$ such that $r^{2n}x = r^n$.

An NP-ring R is called a ϕ-von Neumann regular ring if every R-module is ϕ-flat.

Theorem 8.1. ([39, Theorem 4.3]) Let R be an NP-ring. Then the following conditions are equivalent:

1. R is a ϕ-von Neumann regular ring;
2. Every R-module is ϕ-P-flat;
3. For any non-nilpotent element $a \in R$, there is an element $x \in R$ such that $a = xa^2$;
4. R has only one prime ideal;
5. R is π-regular.

Corollary 8.2. *([39, Corollary 4.5]) Any ϕ-von Neumann regular ring is a strongly ϕ-ring.*

From the previous corollary, we have that any ϕ-von Neumann regular ring is a ϕ-ring, so the two definitions in [39] and [56] are identical.

Corollary 8.3. *([39, Corollary 4.7]) Let R be an NP-ring. Then R is a von Neumann regular ϕ-von Neumann regular ring if and only if R is a field.*

Example 8.4. ([39, Example 4.8]) Let p be a prime number. Then

1. $\mathbb{Z}/p^2\mathbb{Z}$ is a ϕ-von Neumann regular ring which is not a von Neumann regular ring.
2. $\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}$ is a von Neumann regular ring which is not a ϕ-von Neumann regular ring.

Corollary 8.5. *([39, Corollary 4.9]) Every ϕ-von Neumann regular ring is ϕ-chained.*

Corollary 8.6. *([39, Corollary 4.10]) Every ϕ-von Neumann regular ring is ϕ-Dedekind.*

Corollary 8.7. *([39, Corollary 4.11]) Every ϕ-von Neumann regular ring is nonnil-Noetherian.*

The natural question now is whether we can define an NP-ring in which any descending chain of nonnil ideals is stationary. The following theorem shows that this ring is exactly a ϕ-von Neumann regular ring.

Theorem 8.8. ([39, Theorem 4.12]) Let R be an NP-ring. Then the following conditions are equivalent:

1. Every descending chain of nonnil ideals is stationary;
2. R is a ϕ-von Neumann regular ring.

Corollary 8.9. *([39, Corollary 4.13]) Let R be a ϕ-ring. Then R is a ϕ-von Neumann regular ring if and only if so is $\phi(R)$.*

9 Nonnil-coherent and ϕ-coherent rings

Recall that a ring R is coherent if any finitely generated ideal is finitely presented. Bacem and Benhissi [5] generalized the notion of coherent rings to two classes of rings in $\mathcal{H}$: nonnil-coherent rings and ϕ-coherent rings. Let R be a ϕ-ring. Then

1. R is called nonnil-coherent provided that any finitely generated nonnil ideal of R is finitely presented.

2. R is called ϕ-coherent provided that $\phi(R)$ is nonnil-coherent.

Proposition 9.1. *([44, Proposition 1.1]) Let R be a ϕ-ring. Then the following assertions are equivalent:*

1. *R is nonnil-coherent;*

2. *$(0:_R r)$ is a finitely generated ideal for any non-nilpotent element $r \in R$, and the intersection of two finitely generated nonnil ideals of R is a finitely generated nonnil ideal of R;*

3. *$(I:_R b)$ is a finitely generated ideal for any non-nilpotent element $b \in R$ and any finitely generated ideal I of R.*

Proposition 9.2. *([44, Proposition 1.3]) A ϕ-ring R is nonnil-coherent if and only if R is ϕ-coherent and $(0:_R r)$ is a finitely generated ideal for any non-nilpotent element $r \in R$.*

The following example shows that the condition " $(0:_R r)$ is a finitely generated ideal for any non-nilpotent element $r \in R$ " in Proposition 9.2 cannot be removed.

Example 9.3. ([44, Example 1.5]) Let D be a coherent domain not a field, Q its quotient field and $E = \bigoplus_{i=1}^{\infty} Q/D$. Let $R = D \ltimes E$ be the idealization construction. Then R is ϕ-coherent which is not nonnil-coherent.

Proposition 9.4. *([5, Corollary 3.2]) Let R be a ϕ-ring. Then the following statements are equivalent:*

1. *R is a ϕ-coherent ring;*

2. *$\phi(R)$ is a nonnil-coherent ring;*

3. *$\phi(R)/Nil(\phi(R))$ is a coherent domain;*

4. *$R/Nil(R)$ is a coherent domain.*

Bacem and Benhissi [5] generalized the Chase Theorem for coherent rings to that for nonnil-coherent rings.

Theorem 9.5. ([5, Theorem 2.4]) Let R be a ϕ-ring. The following statements are equivalent:

1. R is nonnil-coherent;

2. Any direct product of ϕ-flat R-modules is ϕ-flat;

3. For any indexing set I, any R-module R^I is ϕ-flat.

In 1970, Stenström [46] obtained that a ring R is coherent if and only if any direct limit of FP-injective modules is FP-injective. In 2008, Pinzon [41] showed that if R is coherent, the class of FP-injective modules is (pre)covering. Recently, Dai and Ding [20, 21] showed that the converse of Pinzon's result also hold true. The next result generalizes these results to nonnil-coherent rings.

Theorem 9.6. ([44, Theorem 1.11]) Let R be a ϕ-ring. The following statements are equivalent:

1. R is nonnil-coherent;
2. The class of nonnil-FP-injective R-modules is closed under pure quotients;
3. The class of nonnil-FP-injective R-modules is closed under direct limits;
4. The class of nonnil-FP-injective R-modules is precovering;
5. The class of nonnil-FP-injective R-modules is covering.

In 1993, Chen and Ding in [16] showed that a ring R is coherent if and only if $\mathrm{Hom}_R(M,E)$ is flat for any absolutely pure R-module M and any injective R-module E if and only if $\mathrm{Hom}_R(M,E)$ is flat for any injective R-modules M and E. In [44] Qi and Zhang generalized this result to nonnil-coherent rings.

Theorem 9.7. ([44, Theorem 1.12]) Let R be a ϕ-ring. The following statements are equivalent:

1. R is nonnil-coherent;
2. $\mathrm{Hom}_R(M,E)$ is ϕ-flat for any nonnil-FP-injective module M and any injective module E;
3. $\mathrm{Hom}_R(M,E)$ is ϕ-flat for any nonnil-injective module M and any injective module E;
4. $\mathrm{Hom}_R(\mathrm{Hom}_R(M,E_1),E_2)$ is ϕ-flat for any ϕ-flat module M and any injective modules E_1,E_2;
5. If E_1 and E_2 are injective cogenerators, then $\mathrm{Hom}_R(\mathrm{Hom}_R(M,E_1),E_2)$ is ϕ-flat for any ϕ-flat module M.

Let $R \in \mathscr{H}$ and M be an R-module. Recall from [23] that a submodule N of M is said to be a ϕ-submodule if M/N is a ϕ-torsion module and M is said to be nonnil-coherent if M is finitely generated and every finitely generated ϕ-submodule of M is finitely presented. In particular, every coherent module over a ϕ-ring is nonnil-coherent. Note that for a ϕ-torsion R-module M, we have that M is nonnil-coherent if and only if M is coherent.

Theorem 9.8. ([23, Theorem 2.6]) The following are equivalent for a ϕ-ring R:

1. R is a nonnil-coherent ring;
2. R is a nonnil-coherent R-module;
3. Every finitely generated free R-module is nonnil-coherent;
4. Every finitely presented R-module is nonnil-coherent;
5. Every finitely generated ϕ-submodule of a finitely presented R-module is finitely presented.

The following theorem characterizes when a finitely generated submodule of a nonnil-coherent module is nonnil-coherent.

Theorem 9.9. ([23, Theorem 2.7]) Let $R \in \mathscr{H}$ and M be a nonnil-coherent R-module. If N is a finitely generated ϕ-submodule of M, then N is a nonnil-coherent module.

Corollary 9.10. *([23, Corollary 2.9]) If R is a nonnil-coherent ring, then any finitely generated nonnil ideal of R is a nonnil-coherent R-module.*

An R-module M is said to be a ϕ-divisible module if $sM = M$ for every $s \in R\backslash Nil(R)$. Now, we study the transfer of nonnil-coherent rings in the trivial ring extensions. From [26, Corollary 2.4], a trivial ring extension $R \ltimes M$ is a ϕ-ring if and only if R is a ϕ-ring and M is a ϕ-divisible module.

Theorem 9.11. ([23, Theorem 4.1]) Let $A \in \mathscr{H}$, M be a ϕ-divisible A-module, and set $R := A \ltimes M$. Then the following statements are equivalent:

1. R is a nonnil-coherent ring;
2. A is a nonnil-coherent ring and $(0 : r) \ltimes (0 :_M r)$ is a finitely generated ideal of R for each $r \in A\backslash Nil(A)$;
3. A is a nonnil-coherent ring and $R(r, 0)$ is finitely presented for all $r \in A\backslash Nil(A)$.

Corollary 9.12. *([23, Corollary 4.6]) Let $R = A \ltimes M$ be a ϕ-ring such that $Z(A) = Nil(A)$. Then R is a nonnil-coherent ring if and only if A is a nonnil-coherent ring and $(0 : Mr)$ is a finitely generated A-submodule of M for every $r \in A\backslash Nil(A)$.*

Let R be an NP-ring. Recall from [44] that an R-module M is called ϕ-copure flat provided that $\mathrm{Tor}_1^R(E, M) = 0$ for any nonnil-injective module E, and if N is a submodule of M, then $N \hookrightarrow M$ is said to be a ϕ-embedding map provided that M/N is a ϕ-torsion module.

Theorem 9.13. ([44, Theorem 2.5]) Let R be a ϕ-ring. The following statements are equivalent:

1. Every nonnil-injective R-module is ϕ-flat;
2. R is a ϕ-von Neumann regular ring;
3. R is a nonnil-coherent ring and $R_{\mathfrak{p}}$ is a ϕ-von Neumann regular ring for any $\mathfrak{p} \in \mathrm{Spec}(R)$;
4. R is a nonnil-coherent ring and $R_{\mathfrak{m}}$ is a ϕ-von Neumann regular ring for any $\mathfrak{m} \in \mathrm{Max}(R)$;
5. Any R-module can be ϕ-embedded into a ϕ-flat module;
6. Any nonnil-FP-injective module is ϕ-flat;
7. R is nonnil-coherent and any ϕ-flat module is nonnil-FP-injective;
8. An R-module M is ϕ-flat if and only if M is nonnil-FP-injective;
9. Any ϕ-torsion R-module is ϕ-copure flat;
10. $R/Nil(R)$ is a field.

It is interesting to note that, in 2018, D. Bennis and M. El Hajoui [13] introduced S-finitely presented modules and S-coherent rings, which are S-versions of finitely presented modules and coherent rings. They also gave an S-version of Chase's result to characterize S-coherent rings using ideals. After that, the authors of [45] characterized S-coherent rings in terms of S-Mittag-Leffler modules and S-flat modules (which can be seen as flat modules by localizing at S). In [38] Mahdou and Oubouhou introduced the notion of nonnil-S-coherent rings as a generalization of both nonnil-coherent rings and S-coherent rings. Let R be a ϕ-ring and S be a multiplicative set of R. Then R is said to be a *nonnil-S-coherent* ring if each nonnil ideal of R is S-finitely presented.

10 ϕ-exact sequences and ϕ-projective modules

This section is due to Zhao [54]. Throughout this section R always denotes an NP-ring.

Let M be an R-module. We set

$$NN(R) = \{I \mid I \text{ is a nonnil ideal of } R, \text{ that is, } I \not\subseteq Nil(R)\}$$

and

$$\mathrm{Ntor}(M) = \{x \in M \mid Ix = 0 \text{ for some } I \in NN(R)\}.$$

Then $\mathrm{Ntor}(M)$ is a submodule of M. If $\mathrm{Ntor}(M) = M$ (resp., $\mathrm{Ntor}(M) = 0$), M is called a nonnil-torsion R-module (resp., a nonnil-torsion-free R-module). Define the map $\phi : R \to R_{Nil(R)}$ by $\phi(r) = \frac{r}{1}$ for $r \in R$ and $\psi : M \to M_{Nil(R)}$ by $\psi(x) = \frac{x}{1}$ for $x \in M$. Then ψ is a homomorphism of R-modules. In this case, $\ker\psi = \mathrm{Ntor}(M)$ and $\psi(M) \cong M/\mathrm{Ntor}(M)$ is a $\phi(R)$-module. If M is a nonnil-torsion-free R-module, then $\psi(M) \cong M$. If $f : M \to N$ is a homomorphism of R-modules, then f induces naturally a $\phi(R)$-homomorphism $\tilde{f} : \psi(M) \to \psi(N)$ of $\phi(R)$-modules such that $\tilde{f}\left(\frac{x}{1}\right) = \frac{f(x)}{1}$ for $x \in M$.

This induced homomorphism $\widetilde{f}$ is a homomorphism of nonnil-torsion-free R-modules. If $f : A \to B$ and $g : B \to C$ are two homomorphisms of R-modules, then $\widetilde{gf} = \tilde{g}\tilde{f}$.

If $f, g : A \to B$ are homomorphisms of R-modules, then

$$\widetilde{g+f} = \widetilde{g} + \tilde{f}.$$

Therefore, ψ is an additive covariant functor from the category of R-modules to the category of nonnil-torsion-free $\phi(R)$-modules. Moreover, ψ is an additive covariant functor from the category of R-modules to the category of nonnil-torsion-free R-modules. We have $\psi(\psi(M)) = \psi(M)$ and $\psi(M \oplus N) = \psi(M) \oplus \psi(N)$ for every R-modules M and N.

Definition 10.1. A sequence of R-modules and homomorphisms

$$A \xrightarrow{f} B \xrightarrow{g} C$$

is called a ϕ-complex (resp., a ϕ-exact sequence) if

$$\psi(A) \xrightarrow{\widetilde{f}} \psi(B) \xrightarrow{\widetilde{g}} \psi(C)$$

is a complex (resp., an exact sequence) of $\phi(R)$-modules.

Let $f : A \to B$ be a homomorphism of R-modules. Set

$$\mathrm{NKer}(f) = \{a \in A \mid sf(a) = 0 \text{ for some } s \in R\backslash Nil(R)\},$$

$$\mathrm{NIm}(f) = \{b \in B \mid sb = sf(a) \text{ for some } a \in A \text{ and } s \in R\backslash Nil(R)\}.$$

Because $Nil(R)$ is prime, $\mathrm{NKer}(f)$ is a submodule of A, called the nonnil-kernel of f, and $\mathrm{NIm}(f)$ is a submodule of B, called the nonnil-image of f. We set $\mathrm{NCoker}(f) := B/\mathrm{NIm}(f)$. It is easy to verify that $\mathrm{Ker}(f) + \mathrm{Ntor}(A) \subseteq \mathrm{NKer}(f)$ and $\mathrm{Im}(f)+ \mathrm{Ntor}(B) = \mathrm{NIm}(f)$.

Theorem 10.2. ([54, Theorem 2.6]) Let $A \xrightarrow{f} B \xrightarrow{g} C$ be a sequence of R-modules and homomorphisms. Then

1. $A \xrightarrow{f} B \xrightarrow{g} C$ is a ϕ-complex if and only if $\mathrm{NIm}(f) \subseteq \mathrm{NKer}(g)$.

2. $A \xrightarrow{f} B \xrightarrow{g} C$ is a ϕ-exact sequence if and only if $\mathrm{NIm}(f) = \mathrm{NKer}(g)$.

A homomorphism $f : A \to B$ is called a ϕ-monomorphism (resp., a ϕ-epimorphism; a ϕ-isomorphism) if the induced homomorphism $\widetilde{f}$ is a monomorphism (resp., an epimorphism; an isomorphism).

Note that a ϕ-monomorphism is not always a monomorphism, a ϕ-epimorphism is not always an epimorphism (cf. [54]).

In [54] Zhao provided the following characterizations of a ϕ-monomorphism and a ϕ-epimorphism with the help of the nonnil-kernel and the nonnil-image of an R-homomorphism.

Theorem 10.3. ([54, Theorem 2.7]) Let $f : A \to B$ be a homomorphism of R-modules. Then

1. f is a ϕ-monomorphism if and only if $0 \to A \xrightarrow{f} B$ is ϕ-exact if and only if $\mathrm{NKer}(f) = \mathrm{Ntor}(A)$.

2. f is a ϕ-epimorphism if and only if $A \xrightarrow{f} B \to 0$ is ϕ-exact if and only if $\mathrm{NCoker}(f) = 0$.

3. f is a ϕ-isomorphism if and only if $0 \to A \xrightarrow{f} B \to 0$ is ϕ-exact.

Recall from [51, 54] that an R-module F is ϕ-free (resp., ϕ-projective) if $\psi(F)$ is free (resp., projective) as a $\phi(R)$-module.

Any free R-module is ϕ-free; any ϕ-free R-module is ϕ-projective; any projective R-module is ϕ-projective. If R is a ZN-ring, then a nonnil-torsion-free R-module is ϕ-free (resp., ϕ-projective) if and only if it is free (resp., projective).

Example 10.4. ([54, Example 3.1]) Let D be an integral domain and M a non torsion-free D-module. Taking $R = D \ltimes M$, we have that $\phi(R)$ is a ϕ-free R-module but not free, also not projective.

Example 10.5. ([39, Example 2.16]) Let $D = \mathbb{Z}[\sqrt{5}]$ and K its quotient field. Let $R = D \ltimes K/D$. Then the ring R contains a ϕ-projective ideal which is not ϕ-free.

Proposition 10.6. *([54, Proposition 3.2]) Let P be an R-module. Then the following conditions are equivalent:*

1. *P is a ϕ-projective R-module;*

2. $\mathrm{Ext}^1_{\phi(R)}(\psi(P), A) = 0$ *for any $\phi(R)$-module A (or any nonnil-torsion-free $\phi(R)$-module A);*

3. $\mathrm{Hom}_{\phi(R)}(\psi(P), -)$ *is an exact functor;*

4. *Every exact sequence of $\phi(R)$-modules such as $0 \to A \to B \to \psi(P) \to 0$ (or A is nonnil-torsion-free) is split;*

5. *$\psi(P) \oplus K$ is a free $\phi(R)$-module for some $\phi(R)$-module K.*

It is well known that an R-module P is projective if and only if every exact sequence $0 \to A \to B \to P \to 0$ is split.

Theorem 10.7. ([54, Theorem 3.3]) Let P be an R-module. Then P is ϕ-projective if and only if every ϕ-exact sequence of the form $0 \to A \to B \to P \to 0$ is ϕ-split, that is, the sequence of $\phi(R)$-module $0 \to \psi(A) \to \psi(B) \to \psi(P) \to 0$ is split.

Let M, N be R-modules. If $\psi(M) \cong \psi(N)$, then M is ϕ-projective if and only if N is ϕ-projective. We have that M is ϕ-free if and only if N is ϕ-free. An R-module P is projective if and only if P is a direct summand of some free R-module. The following theorem gives the relationship between ϕ-projective modules and ϕ-free R-modules.

Theorem 10.8. ([54, Theorem 3.4]) Let P be an R-module. Then P is ϕ-projective if and only if P is a direct summand of some ϕ-free module.

We know that any projective module is flat, so the natural question is whether a ϕ-projective module is ϕ-flat.

The following example shows that a ϕ-projective module in an NP-ring is not always ϕ-flat.

Example 10.9. ([39, Example 2.10]) Let $R = \mathbb{Z} \ltimes \mathbb{Z}/2\mathbb{Z}$. Then $\phi(R)$ is a ϕ-projective R-module which is not ϕ-flat R-module.

Theorem 10.10. Let R be an NP-ring. Then the following conditions are equivalent:

1. Every ϕ-projective R-module is ϕ-flat;

2. Every ϕ-free R-module is ϕ-flat.

It is known that if R is a domain and I is an ideal, then I is invertible if and only if I is projective as an R-module. Comparatively, Zhao, Wang and Zhang provided the following result.

Theorem 10.11. ([51, Theorem 4.2]) Let R be a ϕ-ring and I be a finitely generated nonnil ideal. Then I is ϕ-invertible if and only if I is ϕ-projective as an R-module.

The following theorem gives more characterizations concerning modules over ϕ-Prüfer rings.

Theorem 10.12. ([51, Corollary 4.3]) Let R be a ϕ-ring. Then R is a ϕ-Prüfer ring if and only if each finitely generated nonnil ideal of R is ϕ-projective.

Corollary 10.13. *([51, Corollary 4.4]) Let R be a nonnil-Noetherian ring. Then R is a ϕ-Dedekind ring if and only if each nonnil ideal of R is ϕ-projective.*

The following result investigated the ϕ-rings over which all R-modules are ϕ-projective. This result identifies that all semisimple domains are fields.

Theorem 10.14. ([54, Theorem 4.6]) Let R be a ϕ-ring. If all R-modules are ϕ-projective, then R is a field.

11 On nonnil-commutative diagrams and nonnil-projective modules

Let R be an NP-ring and let A, B, C, D be R-modules and $f : A \to B, g : B \to D, h : A \to C, k : C \to D$ be homomorphisms of R-modules. Then the following diagram:

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "h"'] & B \arrow[d, "g"] \\
C \arrow[r, "k"] & D
\end{tikzcd}$$

is said to be nonnil-commutative if $\mathrm{NIm}(gf - kh) = \mathrm{Ntor}(D)$; equivalently, $\mathrm{NKer}(gf - kh) = \mathrm{Ntor}(A)$. A sequence of R-modules and homomorphisms $A \xrightarrow{f} B \xrightarrow{g} C$ is called a nonnil-complex (resp., a nonnil-exact sequence) if it is ϕ-complex (resp., ϕ-exact); equivalently, $\mathrm{NIm}(f) \subseteq \mathrm{NKer}(g)$ (resp., $\mathrm{NIm}(f) = \mathrm{NKer}(g)$) according to [54, Theorem 2.6].

A homomorphism $f : A \to B$ of R-modules is called a nonnil-monomorphism if $\mathrm{NKer}(f) = \mathrm{Ntor}(A)$, equivalently $0 \to A \xrightarrow{f} B$ is a nonnil-exact sequence; f is called a nonnil-epimorphism if $\mathrm{NIm}(f) = B$

(i.e., $\mathrm{NCoker}(f) = 0$), equivalently $A \xrightarrow{f} B \to 0$ is a nonnil exact sequence. Also f is called a nonnil-isomorphism if there exists a homomorphism $g : B \to A$ such that $\mathrm{NIm}(\mathbf{1}_A - gf) = \mathrm{Ntor}(A)$ and $\mathrm{NIm}(\mathbf{1}_B - fg) = \mathrm{Ntor}(B)$. If there exists a nonnil-isomorphism $f : A \to B$, we say that A and B are nonnil-isomorphic, denoted by $A \overset{N}{\simeq} B$. Note that if $f : A \to B$ is a nonnil-isomorphism, then f is both a nonnil-monomorphism and a nonnil-epimorphism. It is interesting to note that a homomorphism f of R-modules is both a nonnil-monomorphism and a nonnil-epimorphism without being a nonnil-isomorphism (see [42]).

Definition 11.1. A homomorphism of R-modules $g : B \to C$ is said to be right nonnil-split if there exists a homomorphism $g' : C \to B$ such that $\mathrm{NIm}(\mathbf{1}_C - gg') = \mathrm{Ntor}(C)$; an R-module homomorphism $f : A \to B$ is said to be left nonnil-split if there exists a homomorphism $f' : B \to A$.

Theorem 11.2. ([42, Theorem 2.8]) Let R be a ring and $f : A \to B$ a homomorphism of R-modules. If f is left nonnil-split (resp., right nonnil-split) and a nonnil-epimorphism (resp., a nonnil-monomorphism), then f is a nonnil-isomorphism.

By Theorem 11.2, an R-module homomorphism $f : A \to B$ is a nonnil-isomorphism if and only if it is both left nonnil-split and right nonnil-split.

Theorem 11.3. ([42, Theorem 2.9]) Let $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ be a short nonnil-exact sequence.

1. If g is right nonnil-split, then there exist submodules M, L of B such that $B = M + L, L \overset{N}{\simeq} C$ and $f : A \to M$ is a nonnil-epimorphism, $M \cap L = \mathrm{Ntor}(B)$, and also $B \cong (M \oplus L)/\mathrm{Ntor}(N)$.
2. If f is left nonnil-split, then there exist submodules M, L of B such that $B = M + L, A \simeq M$ and $g|_L : L \to C$ is a nonnil-monomorphism, $M \cap L = \mathrm{Ntor}(B)$, and also $B \cong (M \oplus L)/\mathrm{Ntor}(B)$.
3. The sequence $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ is nonnil-split if and only if there exist homomorphisms $f' : B \to A, \quad g' : C \to B$ such that $\mathrm{NIm}(\mathbf{1}_A - f'f) = \mathrm{Ntor}(A)$, $\mathrm{NIm}(\mathbf{1}_C - gg') = \mathrm{Ntor}(C)$ and $\mathrm{NIm}(\mathbf{1}_B - ff' - g'g) = \mathrm{Ntor}(B)$.

Theorem 11.4. ([42, Theorem 2.10]) Let $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ be a short nonnil-exact sequence. Then

1. If g is right nonnil-split, then $B \overset{N}{\simeq} \mathrm{NIm}(f) \oplus C$.
2. If f is left nonnil-split, then $B \overset{N}{\simeq} A \oplus \mathrm{NKer}(g)$.
3. If the sequence $0 \to A \xrightarrow{f} B \xrightarrow{g} C \to 0$ is nonnil-split, then $B \overset{N}{\simeq} A \oplus C$.

Definition 11.5. 1. An R-module P is said to be nonnil-projective if given any diagram of module homomorphisms

$$\begin{tikzcd} & P \arrow[ld, dashed, "h"'] \arrow[d, "f"] & \\ B \arrow[r, "g"] & C \arrow[r] & 0 \end{tikzcd}$$

with the bottom row nonnil-exact, there is a homomorphism $h : P \to B$ making this diagram nonnil-commutative.

2. An R-module F_0 is said to be N-free if it is nonnil-isomorphic to a free module.

Theorem 11.6. ([42, Theorem 3.7]) The following statements are equivalent for an R-module P :

1. P is nonnil-projective;
2. Every nonnil-exact sequence such as $0 \to A \to B \to P \to 0$ is right nonnil-split;
3. P is a direct summand of an N-free module.

Corollary 11.7. *([42, Corollary 3.8]) Let P be an R-module. If P is nonnil-isomorphic to a projective module P_0, that is, there is a projective module P_0 such that $P \overset{N}{\simeq} P_0$, then P is nonnil-projective.*

Afterwards, Zhao, Wang, and Pu proposed an interesting problem as follows.

Problem: Is every nonnil-projective module nonnil-isomorphic to some projective module?

Theorem 11.8. ([42, Theorem 3.9]) An R-module P is nonnil-projective if and only if there exist elements $\{x_i \mid i \in \Gamma\} \subseteq P$ and R-homomorphisms $\{f_i \mid i \in \Gamma\} \subseteq \mathrm{Hom}_R(P,R)$ such that:

1. If $x \in P$, then almost all $f_i(x) = 0$.
2. If $x \in P$, then there exists an element $s \in R \backslash Nil(R)$ such that $sx = s\sum_i f_i(x)x_i$.

In this case, P is generated by $\{x_i \mid i \in \Gamma\}$, and $\{x_i, f_i \mid i \in \Gamma\}$, which is called a nonnil-projective basis of P.

Theorem 11.9. ([42, Theorem 4.4]) Let R be a ZN-ring. Then an R-module is ϕ-projective if and only if it is nonnil-projective.

In Theorem 10.14, it is shown that if all R-modules are ϕ-projective, then R is a field. Here we have the following result.

Theorem 11.10. ([42, Theorem 4.5]) Let R be a ring. If all R-modules are nonnil-projective, then R is a field.

Theorem 11.11. ([42, Theorem 4.7]) If all nonnil-torsion-free R-modules are nonnil-projective, then R is a ϕ-Dedekind ring.

Theorem 11.12. ([42, Theorem 4.8]) If all nonnil ideals of R are nonnil-projective, then R is a Dedekind ZN-ring.

12 Strongly ϕ-flat modules, strongly nonnil-injective modules and their homological dimensions

It is well-known that the notions of flat modules and injective modules have the hereditary property, that is, let $0 \to A \to B \to C \to 0$ be a short exact sequence of R-modules. Then it is easy to verify that if B and C are flat modules, so is A; and that if A and B are injective modules, so is C. And so it is ubiquitous to study modules and rings by using flats and injectives. So it is natural and worth asking that:

Do ϕ-flat modules and nonnil-injective modules have the similar hereditary property?

The following two examples show that ϕ-flat modules and nonnil-injective modules do not verify the hereditary property.

Example 12.1. ([52, Example 1.1]) Let $\mathbb{Z}$ be the ring of all integers with $\mathbb{Q}$ its quotients field, and $\mathbb{Z}(\mathfrak{p}^\infty) := \left\{\frac{n}{\mathfrak{p}^k} + \mathbb{Z} \mid \frac{n}{\mathfrak{p}^k} + \mathbb{Z} \in \mathbb{Q}/\mathbb{Z}\right\}$ the $\mathfrak{p}$-Prüfer group with $\mathfrak{p}$ a prime in $\mathbb{Z}$. Set $R = \mathbb{Z} \ltimes \mathbb{Z}(\mathfrak{p}^\infty)$ the trivial extension of $\mathbb{Z}$ with $\mathbb{Z}(\mathfrak{p}^\infty)$. Then R is a ϕ-ring with $Nil(R) = 0 \ltimes \mathbb{Z}(\mathfrak{p}^\infty)$, and so $R/Nil(R)$ is ϕ-flat. However, $Nil(R)$ is not ϕ-flat.

Example 12.2. ([52, Example 1.2]) Consider the above Example 12.1 and let $E := \mathrm{Hom}_{\mathbb{Z}}(R/Nil(R), \mathbb{Q}/\mathbb{Z})$. Then E is nonnil-injective. However, the quotient $\mathrm{Hom}_{\mathbb{Z}}(Nil(R), \mathbb{Q}/\mathbb{Z})$ of the injective module $\mathrm{Hom}_{\mathbb{Z}}(R, \mathbb{Q}/\mathbb{Z})$ by E is not nonnil-injective.

To obtain the hereditary property flatness and injectivity in NP-rings, Zhang et al.introduced the following ?strong version? of ϕ-flat modules and nonnil-injective modules using higher derived functors.

Definition 12.3. Let R be an NP-ring and M an R-module. Then

1. M is called strongly ϕ-flat if $\mathrm{Tor}_n^R(T, M) = 0$ for any ϕ-torsion module T and any $n \geq 1$.
2. M is called strongly nonnil-injective if $\mathrm{Ext}_R^n(T, M) = 0$ for any ϕ-torsion module T and any $n \geq 1$.

Proposition 12.4. *([52, Proposition 1.5]) Let R be a ϕ-ring and $0 \to A \to B \to C \to 0$ a short exact sequence of R-modules. Then the following statements hold.*

1. *The class of strongly ϕ-flat modules (resp., strongly nonnil-injective modules) is closed under direct limits (resp., direct products), direct summands, and extensions.*
2. *If B and C are strongly ϕ-flat modules, so is A.*
3. *If A and B are strongly nonnil-injective modules, so is C.*

Obviously, every strongly ϕ-flat module is ϕ-flat, and every strongly nonnil-injective module is nonnil-injective. However, Example 12.1 and Example 12.2 show that ϕ-flat modules are not always strongly ϕ-flat, and nonnil-injective modules are also not always strongly nonnil-injective. But the following result exhibits that over ZN-rings, ϕ-flat modules are exactly strongly ϕ-flat and nonnil-injective modules are exactly strongly nonnil-injective.

Proposition 12.5. *([52, Theorem 1.6]) Let R be a ZN-ring. Then the following assertions hold:*

1. *An R-module M is ϕ-flat if and only if it is strongly ϕ-flat.*
2. *An R-module M is nonnil-injective if and only if it is strongly nonnil-injective.*

Proposition 12.6. *([52, Proposition 1.8]) Let R be an NP-ring. Then the following assertions are equivalent:*

1. *M is strongly ϕ-flat;*
2. $\mathrm{Hom}_R(M, E)$ *is strongly nonnil-injective for any injective module E;*
3. *If E is an injective cogenerator, then* $\mathrm{Hom}_R(M, E)$ *is strongly nonnil-injective.*

Let R be a ring. It is well known that the flat dimension of an R-module M is defined as the length of the shortest flat resolution of M and the weak global dimension of R is the supremum of the flat dimensions of all R-modules. Zhang et al. introduced the notion of ϕ-flat dimension of an R-module as follows:

Definition 12.7. Let R be a ring and M an R-module. We write $\phi\text{-fd}_R(M) \leq n$ (ϕ-fd abbreviates ϕ-flat dimension) if there is an exact sequence of R-modules

$$0 \to F_n \to \cdots \to F_1 \to F_0 \to M \to 0 \qquad (\diamond)$$

where each F_i is strongly ϕ-flat for $i = 0, \ldots, n$. The exact sequence $(\diamond)$ is said to be a ϕ-flat resolution of length n of M. If such finite resolution does not exist, then we say $\phi\text{-fd}_R(M) = \infty$; otherwise, define $\phi\text{-fd}_R(M) = n$ if n is the length of the shortest ϕ-flat resolution of M.

It is obvious that an R-module M is strongly ϕ-flat if and only if $\mathrm{fd}_R(M) = 0$. Certainly, $\phi\text{-}\mathrm{fd}_R(M) \leq \mathrm{fd}_R(M)$. If R is an integral domain, then $\phi\text{-}\mathrm{fd}_R(M) = \mathrm{fd}_R(M)$.

Proposition 12.8. *([52, Proposition 2.2]) Let R be an NP-ring. The following statements are equivalent for an R-module M:*

1. *ϕ-$fd_R(M) \leq n$;*
2. $\mathrm{Tor}^R_{n+k}(T,M) = 0$ *for every ϕ-torsion R-module T and every positive integer k;*
3. $\mathrm{Tor}^R_{n+k}(R/I,M) = 0$ *for every nonnil ideal I and every positive integer k;*
4. $\mathrm{Tor}^R_{n+k}(R/I,M) = 0$ *for every finitely generated nonnil ideal I and every positive integer k;*
5. *If $0 \to F_n \to \cdots \to F_1 \to F_0 \to M \to 0$ is an exact sequence, where $F_0, F_1, \ldots, F_{n-1}$ are strongly ϕ-flat R-modules, then F_n is strongly ϕ-flat;*
6. *If $0 \to F_n \to \cdots \to F_1 \to F_0 \to M \to 0$ is an exact sequence, where $F_0, F_1, \ldots, F_{n-1}$ are flat R-modules, then F_n is strongly ϕ-flat;*
7. *There exists an exact sequence $0 \to F_n \to \cdots \to F_1 \to F_0 \to M \to 0$, where $F_0, F_1, \ldots, F_{n-1}$ are flat R-modules, then F_n is strongly ϕ-flat.*

Definition 12.9. The ϕ-weak global dimension of a ring R is defined by

$$\phi\text{-}w.gl.dim(R) = \sup\{\phi\text{-}fdR(M) \mid M \text{ is an } R\text{-module }\}.$$

Obviously, by definition, ϕ-$w.gl.dim(R) \leq w.gl.dim(R)$. Notice that if R is an integral domain, then ϕ-$w.gl.dim(R) = w.gl.dim(R)$.

Theorem 12.10. ([52, Theorem 2.6]) Let R be an NP-ring. The following statements are equivalent:

1. ϕ-$w.gl.dim(R) \leq n$;
2. ϕ - $fd_R(M) \leq n$ for every R-module M;
3. $\mathrm{Tor}^R_{n+k}(T,M) = 0$ for every R-module M, every ϕ-torsion T and every positive integer k;
4. $\mathrm{Tor}^R_{n+k}(R/I,M) = 0$ for every R-module M, every nonnil ideal I of R and every positive integer k;
5. $\mathrm{Tor}^R_{n+k}(R/I,M) = 0$ for every R-module M and every finitely generated nonnil ideal I;
6. $\mathrm{Tor}^R_{n+1}(T,M) = 0$ for every R-module M and every ϕ-torsion T;
7. $\mathrm{Tor}^R_{n+1}(R/I,M) = 0$ for every R-module M and every nonnil ideal I of R;
8. $\mathrm{Tor}^R_{n+1}(R/I,M) = 0$ for every R-module M and every finitely generated nonnil ideal I of R;
9. $fd_R(R/I) \leq n$ for every nonnil ideal I of R;
10. $fd_R(R/I) \leq n$ for every finitely generated nonnil ideal I of R.

Consequently, the ϕ-weak global dimension of R is determined by the formulas:

$$\begin{aligned}\phi\text{-w.gl.dim }(R) &= \sup\{fd_R(R/I) \mid I \text{ is a nonnil ideal of } R\}\\ &= \sup\{fd_R(R/I) \mid I \text{ is a finitely generated nonnil ideal of } R\}.\end{aligned}$$

The following theorem characterizes ϕ-von Neumann regular rings in terms of strongly ϕ-flat modules and ϕ-weak global dimensions.

Theorem 12.11. ([52, Theorem 2.8]) Let R be a ϕ-ring. The following statements are equivalent:

1. ϕ-w.gl.dim $(R) = 0$;
2. Every R-module is strongly ϕ-flat;
3. R is a ϕ-von Neumann regular ring.

Theorem 12.12. ([52, Theorem 2.9]) Let R be a ϕ-ring. The following statements are equivalent:

1. ϕ-w.gl.dim $(R) \leq 1$;
2. Every submodule of a flat R-module is strongly ϕ-flat;
3. Every submodule of strongly ϕ-flat R-module is strongly ϕ-flat;
4. R is a ϕ-Prüfer strongly ϕ-ring.

Let R be a ring. It is well known that the injective dimension of an R-module M is defined as the length of the shortest injective resolution of M, and the global dimension of R is the supremum of the injective dimensions of all R-modules. Comparatively, Zhang introduced the notion of ϕ-injective dimension of an R-module as follows.

Definition 12.13. Let R be a ring and M an R-module. We write $\phi\text{-}id_R(M) \leq n$ ($\phi\text{-}id$ abbreviates ϕ-injective dimension) if there is an exact sequence of R-modules

$$0 \to M \to E_0 \to E_1 \to \cdots \to E_n \to 0 \qquad (\nabla)$$

where each E_i is strongly nonnil-injective for $i = 0, \ldots, n$. The exact sequence (∇) is said to be a ϕ-injective resolution of length n of M. If such finite resolution does not exist, then we say ϕ-id $(M) = \infty$; otherwise, define $\phi\text{-id}_R(M) = n$ if n is the length of the shortest ϕ-injective resolution of M.

It is obvious that an R-module M is strongly nonnil-injective if and only if $\phi\text{-id}_R(M) = 0$. Certainly, $\phi\text{-id}_R(M) \leq \text{id}_R(M)$. If R is an integral domain, then $\phi\text{-}id_R(M) = \text{id}_R(M)$.

Theorem 12.14. ([52, Proposition 3.2]) Let R be an NP-ring. The following statements are equivalent for an R-module M:

1. $\phi - id_R(M) \leq n$;
2. $\text{Ext}_R^{n+k}(T, M) = 0$ for every ϕ-torsion R-module T and every positive integer k;
3. $\text{Ext}_R^{n+k}(R/I, M) = 0$ for every nonnil ideal I and every positive integer k;
4. If $0 \to M \to E_0 \to E_1 \to \cdots \to E_n \to 0$ is an exact sequence, where $E_0, E_1, \ldots, E_{n-1}$ are strongly nonnil-injective R-modules, then E_n is strongly nonnil-injective;
5. If $0 \to M \to E_0 \to E_1 \to \cdots \to E_n \to 0$ is an exact sequence, where $E_0, E_1, \ldots, E_{n-1}$ are injective R-modules, then E_n is strongly nonnil-injective;
6. There exists an exact sequence $0 \to M \to E_0 \to E_1 \to \cdots \to E_n \to 0$, where $E_0, E_1, \ldots, E_{n-1}$ are injective R-modules and E_n is strongly nonnil-injective.

Corollary 12.15. *([52, Corollary 3.3]) Let R be an NP-ring, M an R-module and E an injective cogenerator of R-Mod. Then $\phi\text{-}fd_R(M) = \phi\text{-}id\,(\text{Hom}_R(M, E))$*

Definition 12.16. The ϕ-global dimension of a ring R is defined by

$$\phi\text{-}gl.dim(R) = \sup\{\phi\text{-}id_R(M) \mid M \text{ is an } R\text{-module }\}.$$

Obviously, by definition, $\phi\text{-}gl.dim(R) \leq gl.dim(R)$. Notice that if R is an integral domain, then $\phi\text{-}gl.dim(R) = gl.dim(R)$.

Theorem 12.17. ([52, Theorem 3.7]) Let R be an NP-ring. The following statements are equivalent for R.

1. $\phi\text{-}gl.dim(R) \leq n$;
2. $\phi\text{-}id_R(M) \leq n$ for every R-module M;
3. $\mathrm{Ext}_R^{n+k}(T,M) = 0$ for every R-module M, every ϕ-torsion T and every positive integer k;
4. $\mathrm{Ext}_R^{n+k}(R/I,M) = 0$ for every R-module M, every nonnil ideal I of R and every positive integer k;
5. $\mathrm{Ext}_R^{n+1}(T,M) = 0$ for every R-module M and every ϕ-torsion T;
6. $\mathrm{Ext}_R^{n+1}(R/I,M) = 0$ for every R-module M and every nonnil ideal I of R.

Consequently, the ϕ-global dimension of R is determined by the formulas:

$$\phi\text{-}gl.dim(R) = \sup\{pd_R(R/I) \mid I \text{ is a nonnil ideal of } R\}.$$

Theorem 12.18. ([52, Theorem 3.9]) Let R be a ϕ-ring. The following statements are equivalent:

1. ϕ-gl.dim $(R) = 0$;
2. Every R-module is strongly nonnil-injective;
3. R is a ϕ-von Neumann regular ring.

Theorem 12.19. ([52, Theorem 3.10]) Let R be a ϕ-ring. The following statements are equivalent:

1. ϕ-gl.dim $(R) \leq 1$;
2. Every quotient module of an injective R-module is strongly nonnil-injective;
3. Every quotient module of a strongly nonnil-injective R-module is strongly nonnil-injective;
4. R is a ϕ-Dedekind strongly ϕ-ring.

13 ϕ-(weakly) global dimension

This section is due to El Haddaoui and Mahdou [24].

Definition 13.1. An R-module M is said to be ϕ-uniformly torsion (ϕ-u-torsion for short) if $sM = 0$ for some $s \in R \backslash Nil(R)$.

Example 13.2. For every nonnil ideal I of R, we get that R/I is ϕ-u-torsion.

Definition 13.3. Let R be a ring. An R-module P is said to be ϕ-uniformly projective (ϕ-u-projective for short) if $Ext_R^1(P,N) = 0$ for every ϕ-u-torsion R-module N. In particular, every projective module is ϕ-u-projective.

The following theorem characterizes ϕ-u-projective modules by short exact sequences.

Theorem 13.4. ([24, Theorem 3.5]) Let R be a ring. The following statements hold for an R-module P:

1. If P is ϕ-u-projective, then every exact sequence $0 \to A \to B \xrightarrow{g} P \to 0$ is split for every ϕ-u-torsion R-module A.

2. P is ϕ-u-projective if and only if every exact sequence $0 \to A \to B \to C \to 0$, where A is ϕ-u-torsion, induces the exact sequence

$$0 \to \mathrm{Hom}_R(P,A) \to \mathrm{Hom}_R(P,B) \to \mathrm{Hom}_R(P,C) \to 0.$$

As in the classical homology, El Haddaoui and Mahdou defined the ϕ-projective dimension of an R-module as follows:

Definition 13.5. Let R be a ring and let M be an R-module. The ϕ-projective dimension of M over R, denoted by $\phi\text{-}pd_R M$, is said to have at most $n \geq 1$ (where $n \in \mathbb{N}$) if either $M = 0$ or $M \neq 0$ which is not ϕ-u-projective and which satisfies $Ext_R^{n+1}(M,N) = 0$ for every ϕ-u-torsion module N. If n is the least non-negative integer for which $Ext_R^{n+1}(M,N) = 0$ for every ϕ-u-torsion module N, then we set $\phi\text{-}pd_R M = n$. If no such n exists, set $\phi\text{-}pd_R M = \infty$.

Let P be an R-module. Then it is easy to see that P is a ϕ-u-projective module if and only if $\phi\text{-}pd_R P = 0$.

Theorem 13.6. ([24, Theorem 3.10]) Let R be a ring and $n \in \mathbb{N}^*$. The following statements are equivalent for a non ϕ-u-projective R-module M:

1. $\phi\text{-pd}_R M \leqslant n$;

2. $\mathrm{Ext}_R^{n+1}(M,N) = 0$ for every ϕ-u-torsion R-module N;

3. If $0 \longrightarrow P_n \longrightarrow P_{n-1} \longrightarrow \cdots \longrightarrow P_1 \longrightarrow P_0 \longrightarrow M \longrightarrow 0$ is exact, where $P_0, P_1, \ldots, P_{n-1}$ are projective, then P_n is ϕ-u-projective.

Theorem 13.7. ([24, Theorem 3.12]) Let $0 \to N \to P \to M \to 0$ be an exact sequence where P is projective. If $\phi\text{-}pd_R M = n \geq 1$, then $\phi\text{-}pd_R N = n-1$.

Theorem 13.8. ([24, Theorem 3.13]) Let $\{M_i\}_{i\in I}$ be a family of R-modules. Then

$$\phi\text{-}pd_R(\bigoplus_{i\in I} M_i) = sup_{i\in I}\phi\text{-}pd_R M_i.$$

Theorem 13.9. ([24, Theorem 4.1]) The following are equivalent for a ZN-ring R:

1. $\phi\text{-pd}_R(R/I) \leqslant n$ for every nonnil ideal I of R;

2. $\phi\text{-id}_R N \leqslant n$ for every ϕ-u-torsion R-module N;

3. $\mathrm{Ext}_R^{n+1}(R/I,N) = 0$ for every nonnil ideal I of R and every ϕ-u-torsion R-module N.

Let R be a ring. It is well known that the global dimension of R is the supremum of the projective dimensions of all R-modules. Comparatively, El Haddaoui and Mahdou introduced the notion of ϕ-global dimension as follows.

Definition 13.10. Let R be a ZN-ring, define

$$\begin{aligned}\phi\text{-gldim}(R) &= \sup\{\phi\text{-pd}_R(R/I) \mid I \textit{ is a nonnil ideal of } R\}\\ &= \sup\{\phi\text{-id}_R N \mid N \textit{ is a } \phi\text{-u-}\textit{torsion } R\textit{-module}\}\end{aligned}$$

which is called the ϕ-global dimension of R.

Definition 13.11. The ϕ-global dimension of a ring R is either 0 (i.e., R is ϕ-semisimple) or the supremum of all $\phi\text{-}pd_R(R/I)$ where I is a nonnil ideal of R such that R/I is not ϕ-u-projective.

Definition 13.12. A ϕ-ring R is said to be ϕ-hereditary if every nonnil ideal of R is ϕ-u-projective. In particular, every hereditary ϕ-ring is ϕ-hereditary.

As in the classical homology, the following theorem characterizes the ϕ-hereditary rings.

Theorem 13.13. ([24, Theorem 4.3]) The following statements are equivalent for a strongly ϕ-ring R:

1. R is ϕ-hereditary;
2. Every factor E/X of an injective module E by a ϕ-u-torsion submodule X of E is nonnil-injective;
3. $\phi\text{-}gldim(R) \leq 1$.

Definition 13.14. The sequence $0 \to A \to B \to C \to 0$ is said to be ϕ-pure exact if for every finitely presented ϕ-torsion module F, we get the following exact sequence $0 \to F \otimes_R A \to F \otimes_R B \to F \otimes_R C \to 0$. In particular, every pure exact sequence is ϕ-pure.

The following theorem characterizes the ϕ-pure exact sequence as in the classical case.

Theorem 13.15. ([24, Theorem 5.3]) The following are equivalent for an exact sequence $0 \to A \to B \to C \to 0$.

1. The above exact sequence is ϕ-pure;
2. For every finitely presented ϕ-torsion R-module F, we get the following exact sequence $0 \to Hom_R(F,A) \to Hom_R(F,B) \to Hom_R(F,C) \to 0$.

Theorem 13.16. ([24, Theorem 5.4]) The following are equivalent for an R-module F:

1. F is ϕ-flat;
2. Every exact sequence $0 \to M' \to M \to F \to 0$ is ϕ-pure;
3. There exists a ϕ-pure exact sequence $0 \to M' \to M \to F \to 0$, where M is ϕ-flat.

A submodule A of B is said to be ϕ-pure if the exact sequence $0 \to A \to B \to B/A \to 0$ is ϕ-pure. The following theorem characterizes the ϕ-pure ideals of a ring R.

Theorem 13.17. ([24, Theorem 5.7]) Let I be an ideal of R. Then the following are equivalent:

1. I is a ϕ-pure ideal;
2. R/I is ϕ-flat;
3. $I \cap J = IJ$ for every nonnil ideal J of R;

4. $I \cap J = IJ$ for every finitely generated nonnil ideal J of R.

The following example gives a ϕ-pure ideal which is not pure.

Example 13.18. ([24, Example 5.2]) In the ring $R := \mathbb{Q} \ltimes \mathbb{Q}$, the $Nil(R) = 0 \ltimes \mathbb{Q}$ is a ϕ-pure ideal of R which is not pure.

In the classical homology, every projective module is flat. The following theorem shows the same result.

Theorem 13.19. ([24, Theorem 5.11]) Let R be a ring and P be an R-module. If P is a ϕ-u-projective module, then P is ϕ-flat.

In the classical homology, every finitely presented flat module is projective. The following theorem shows the analog of the result.

Theorem 13.20. ([24, Theorem 5.13]) Let R be a ring. Then every finitely presented ϕ-flat module is ϕ-u-projective.

The following example 13.21 gives a ϕ-u-projective module which isn't projective.

Example 13.21. ([24, Example 5.4]) $\mathrm{Nil}(\mathbb{Z}/2\mathbb{Z} \ltimes \mathbb{Z}/2\mathbb{Z})$ is a ϕ-u-projective ideal of $R = \mathbb{Z}/2\mathbb{Z} \ltimes \mathbb{Z}/2\mathbb{Z}$ which isn't projective.

Definition 13.22. Let R be a ring and let M be an R-module. The ϕ-flat dimension of M over R, denoted by $\phi\text{-}fd_RM$, is said to have at most $n \geq 1$ (where $n \in \mathbb{N}$) if either $M = 0$ or $M \neq 0$ which is not ϕ-flat and satisfies $\mathrm{Tor}_{n+1}^R(M,N) = 0$ for every ϕ-u-torsion module N. If n is the least non-negative integer for which $\mathrm{Tor}_{n+1}^R(M,N) = 0$ for every ϕ-u-torsion module N, then we set $\phi\text{-}fd_RM = n$. If no such n exists, set $\phi\text{-}fd_RM = \infty$.

Theorem 13.23. ([24, Theorem 5.19]) Let R be a ring and $n \in \mathbb{N}^*$. The following statements are equivalent for a non ϕ-flat R-module M:

1. $\phi\text{-fd}_RM \leqslant n$;
2. $\mathrm{Tor}_{n+1}^R(M,N) = 0$ for every ϕ-torsion module N;
3. $\mathrm{Tor}_{n+1}^R(M,N) = 0$ for every ϕ-u-torsion module N;
4. If $0 \to F_n \to F_{n-1} \to \cdots \to F_1 \to F_0 \to M \to 0$ is exact, in which $F_0, F_1, \ldots, F_{n-1}$ are flat, then F_n is ϕ-flat;
5. $\mathrm{Tor}_{n+1}^R(M,R/I) = 0$ for every nonnil ideal I of R;
6. $\mathrm{Tor}_{n+1}^R(M,R/I) = 0$ for every finitely generated nonnil ideal I of R;
7. $\mathrm{Tor}_{n+1}^R(M,X) = 0$ for every finitely presented ϕ-torsion module X.

Definition 13.24. Let R be a ring. Then the ϕ-weak global dimension of R, denoted by $\mathrm{w.gldim}(R)$, is the supremum of all $\phi\text{-}fd_R(R/I)$, where I is a nonnil ideal of R.

In the case where R is a ZN-ring, we have:

$$\begin{aligned}
\phi\text{-w.gldim}(R) &= \sup\{\phi\text{-fd}_RM \mid M\ is\ \phi\text{-}torsion\}\\
&= \sup\{\phi\text{-fd}_RM \mid M\ is\ \phi\text{-u-}torsion\}\\
&= \sup\{\phi\text{-fd}_RM \mid M\ is\ finitely\ presented\ \phi\text{-}torsion\}\\
&= \sup\{\phi\text{-fd}_RM \mid M\ is\ finitely\ presented\ \phi\text{-u-}torsion\}\\
&= \sup\{\phi\text{-fd}_R(R/I) \mid I\ is\ a\ nonnil\ ideal\ of\ R\}\\
&= \sup\{\phi\text{-fd}_R(R/I) \mid I\ is\ a\ finitely\ generated\ nonnil\ ideal\ of\ R\}.
\end{aligned}$$

Corollary 13.25. *([24, Corollary 5.27]) The following statements are equivalent for a strongly ϕ-ring R:*

1. *R is ϕ-Prüfer;*
2. *ϕ-$w.gldim(R) \leq 1$.*

Corollary 13.26. *([24, Corollary 5.34]) The following statements are equivalent for a ϕ-ring R:*

1. *R is a ϕ-von Neumann regular ring;*
2. *Every R-module is ϕ-u-projective;*
3. *Every short exact sequence: $0 \to A \to B \to C \to 0$ is split, where C is ϕ-torsion;*
4. *Every short exact sequence is ϕ-pure.*

Corollary 13.27. *([24, Corollary 5.36]) Let R be a ϕ-ring with non-maximal nilradical. Then every finitely presented ϕ-u-projective module is projective.*

In Corollary 13.25, it provided that a strongly ϕ-ring is ϕ-Prüfer if and only if its ϕ-weak global dimension is at most one. In the next example El Haddaoui and Mahdou established that the above result is not true if we assume that R is not strongly ϕ-ring.

Example 13.28. ([24, Example 5.21]) Let $R := \mathbb{Z} \ltimes \mathbb{Q}/\mathbb{Z}$. Then the following conditions hold:

1. R is both nonnil-coherent and ϕ-Prüfer;
2. ϕ-w.gldim $(R) = +\infty$.

Recall from [48] that a ring R is said to be semi-hereditary, if every finitely generated ideal is projective. The following theorem establishes that the ϕ-Prüfer rings R of $Z(R) = Nil(R)$ are the analog of semi-hereditary rings in the classical case.

Theorem 13.29. ([24, Theorem 5.41]) The following statements are equivalent for a strongly ϕ-ring R:

1. R is a ϕ-Prüfer ring;
2. Every finitely presented ϕ-torsion R-module M is isomorphic to F/N, where F is a finitely generated free R-module and N is a finitely generated ϕ-u-projective module;
3. Every finitely generated nonnil ideal of R is ϕ-u-projective.

In the rest of the section, we apply the results seen previously concerning the ϕ-(weak) global dimension of rings, we study the ϕ-(weak) global dimension of the trivial ring extensions $R \ltimes M$ which are ϕ-rings and also flat extensions of R.

Theorem 13.30. ([24, Theorem 6.1]) Let R be a ϕ-ring and M be a ϕ-divisible module. If M is a flat R-module, then ϕ-$gldim(R \ltimes M) = \phi$-$gldim(R)$.

Example 13.31. ([24, Examples 6.1, 6.2, 6.3])

1. If K is a field, then ϕ-$gldim(K \ltimes K^n) = \phi$-$gldim(K) = 0$.
2. If p is a positive prime integer and $k, n \in \mathbb{N}^*$, then

$$\phi\text{-}gldim(\mathbb{Z}/p^n\mathbb{Z} \ltimes (\mathbb{Z}/p^n\mathbb{Z})^{(k)}) = \phi\text{-}gldim(\mathbb{Z}/p^n\mathbb{Z}) = 0.$$

3. $\phi\text{-}gldim(\mathbb{Z} \ltimes \mathbb{Q}) = 1$.

Theorem 13.32. ([24, Theorem 6.6]) Let R be a ϕ-ring and M be a ϕ-divisible module. If M is a flat R-module, then $\phi\text{-}w.gldim(R \ltimes M) = \phi\text{-}w.gldim(R)$.

Corollary 13.33. *([24, Corollary 6.8]) If $R \in \mathcal{H}$ and M be a ϕ-divisible R-module such that $Z_R(M) = 0$, then $\phi\text{-}w.gldim(R \ltimes M) = \phi\text{-}w.gldim(R)$.*

Corollary 13.34. *([24, Corollary 9]) If R is an integral domain and $Q = qf(R)$, then $\phi\text{-}gldim(R \ltimes Q) = \phi\text{-}w.gldim(R) = w.gldim(R)$.*

Example 13.35. ([24, Examples 6.4, 6.5, 6.6])

1. If K is a field, then $\phi\text{-}w.gldim(K \ltimes K^n) = \phi\text{-}w.gldim(K) = 0$.

2. $\phi\text{-}w.gldim(\mathbb{Z} \ltimes \mathbb{Q}) = 1$.

3. $\phi\text{-}w.gldim(\mathbb{Z}[X_1,\dots,X_n] \ltimes qf(\mathbb{Z}[X_1,\dots,X_n]) = n+1$.

4. $\phi\text{-}w.gldim(\mathbb{Z}[X_1,\dots,X_n,\cdots] \ltimes qf(\mathbb{Z}[X_1,\cdots,X_n,\dots]) = +\infty$.

References

[1] D. F. Anderson and A. Badawi, *On ϕ-Dedekind rings and ϕ-Krull rings*, Houston J. Math, 31(4) (2005), 1007–1022.

[2] D. F. Anderson and A. Badawi, *On ϕ-Prüfer rings and ϕ-Bézout rings*, Houston J. Math, 30(2) (2004), 331–343.

[3] D. D. Anderson and T. Dumitrescu, *S-Noetherian rings*, Comm. Algebra, 30(9) (2002), 4407–4416.

[4] D. D. Anderson and M. Winders, *Idealization of a module*, J. Commut. Algebra, 1(1) (2009), 3–56.

[5] K. Bacem and A. Benhissi, *Nonnil-coherent rings*, Beitr. Algebra Geom., 57(2) (2016), 297–305.

[6] A. Badawi, *On divided commutative rings*, Comm. Algebra, 27(3) (1999), 1465–1474.

[7] A. Badawi, *On nonnil-Noetherian rings*, Comm. Algebra, 31(4) (2003), 1669–1677.

[8] A. Badawi, *On rings with divided nil ideal: a survey*. In: Fontana, M., et al. eds. Commutative Algebra And Its Applications. Berlin: Walter de Gruyter, (2009), 21–40.

[9] A. Badawi, *On ϕ-chained rings and ϕ-pseudo-valuation rings*, Houston J. Math, 27(4) (2001), 725–736.

[10] A. Badawi, *On ϕ-pseudo-valuation rings*, Lecture Notes Pure Applied Mathematics, Marcel Dekker, New York 205 (1999), 101–110.

[11] A. Benhissi, *Nonnil-Noetherian rings and formal power series*, Algebra Colloq., 27(3) (2020), 361–368.

[12] A. Benhissi, *Chain Conditions in Commutative Rings*, Springer Nature, (2022) ISBN 978-3-031-09897-7 (eBook).

[13] D. Bennis and M. El Hajoui, *On S-coherence*, J. Korean Math. Soc., 55(6) (2018), 1499–1512.

[14] M. B. Boisen, JR. and P. B. Sheldon, *Pre-Prüfer rings*, Pacific J. Math. 58 (1975), 331–344.

[15] G. W. Chang and H. Kim, *Prüfer rings in a certain pullback*, Comm. Algebra, 51 (2023), 2045–2063.

[16] J. L. Chen and N. Q. Ding, *The weak global dimension of commutative coherent rings*, Comm. Algebra, 21(10) (1993), 3521–3528.

[17] M. D'Anna and M. Fontana, *An amalgamated duplication of a ring along an ideal: the basic properties*, J. Algebra Appl, 6(3) (2007), 443–459.

[18] M. D'Anna, C. A. Finocchiaro and M. Fontana, *Amalgamated algebras along an ideal*, Comm. Algebra and Applications, Walter De Gruyter, (2009), 241–252.

[19] M. D'Anna, C. A. Finocchiaro and M. Fontana, *Properties of chains of prime ideals in amalgamated algebras along an ideal*, J. Pure Applied Algebra, 214(9) (2010), 1633–1641.

[20] G. C. Dai and N. Q. Ding, *Coherent rings and absolutely pure covers*, Comm. Algebra, 46(3) (2018), 1267–1271.

[21] G. C. Dai and N. Q. Ding, *Coherent rings and absolutely pure precovers*, Comm. Algebra, 47(11) (2019), 4743–4748.

[22] D. E. Dobbs, *Divided rings and going-down*, Pacific J. Math., 67(2) (1976), 353–363.

[23] Y. El Haddaoui, H. Kim and N. Mahdou, *On nonnil-coherent modules and nonnil-Noetherian modules*, Open Mathematics, 20(1) (2022), 1521–1537.

[24] Y. El Haddaoui and N. Mahdou, *On ϕ-(weak) global dimension*, J. Algebra Appl., to appear.

[25] A. El Khalfi, H. Kim and N. Mahdou, *Amalgamated algebras issued from ϕ-chained rings and ϕ-pseudo-valuation rings*, Bull. Iranian Math. Soc., 47(5) (2021), 1599–1609.

[26] A. El Khalfi, H. Kim and N. Mahdou, *Amalgamation extension in commutative ring theory, a survey*, Moroccan Journal of algebra and Geometry with applications, 1(1) (2022), 139–182.

[27] S. Glaz, *Commutative Coherent Rings*, *Springer-Verlag*, Lecture Notes in Mathematics, 1371 (1989).

[28] M. Griffin, *Prüfer rings with zerodivisors*, J. Reine Angew. Math., 240 (1970), 55–67.

[29] S. Hizem and A. Benhissi, *Nonnil-Noetherian rings and the SFT property*, Rocky Mountain J. Math., 41(5) (2011), 1483–1500.

[30] J. A. Huckaba, *Commutative Rings with Zero Divisors*, Dekker, New York, 1988.

[31] G. U. Jensen, *Arithmetical rings*, Acta Sci. Acad. Hungar. 17(1966), 115–123.

[32] S. Kabbaj, *Matlis' semi-regularity and semi-coherence in trivial ring extensions: a survey*, Moroccan Journal of Algebra and Geometry with Applications, 1(1) (2021), 1–17.

[33] S. Kabbaj and N. Mahdou, *Trivial extensions defined by coherent-like conditions*, Comm. Algebra, 32(1) (2004), 3937–3953.

[34] H. Kim, N. Mahdou and E. H. Oubouhou, *When every ideal is ϕ-P-flat*, Hacettepe J. of Math. Stat., 52(3) (2023), 708–720.

[35] H. Kim, N. Mahdou and E. H. Oubouhou, *On ϕ-u-S-flat modules and nonnil-u-S-injective modules*, Georgian Math. J., to appear.

[36] M. J. Kwon and J. W. Lim, *On nonnil-S-Noetherian rings*, Mathematics, 8(9) (2020), 1428.

[37] Z. K. Liu and X. Y. Yang, *On nonnil-Noetherian rings*, Southeast Asian Bull. Math., 33(6) (2009), 1215–1223.

[38] N. Mahdou and E. H. Oubouhou, *Nonnil-S-coherent rings*, Commun. Korean Math. Soc., to appear.

[39] N. Mahdou and E. H. Oubouhou, *On ϕ-P-flat modules and ϕ-von Neumann regular rings*, J. Algebra Appl., to appear.

[40] N. Mahdou, E. H. Oubouhou and E. Yetkin Celikel, *On nonnil-S-Noetherian and nonnil-u-S-Noetherian rings*, An. S, t. Univ. Ovidius Constanta., to appear.

[41] K. R. Pinzon, *Absolutely pure covers*, Comm. Algebra, 36 (2008), 2186–2194.

[42] Y. Pub, M. Wang and W. Zhao, *On nonnil-commutative diagrams and nonnil-projective modules*, Comm. Algebra, 50(7) (2022), 2854–2867.

[43] W. Qi and X. Zhang, *Some remarks on ϕ-Dedekind rings and ϕ-Prüfer rings*, arXiv preprint arXiv, (2021), 2103.08278.

[44] W. Qi and X. Zhang, *Some remarks on nonnil-coherent rings and ϕ-IF rings*, J. Algebra Appl., 21(11), (2021), 2250211.

[45] W. Qi, X. Zhang and W. Zhao, *New characterizations of S-coherent rings*, J. Algebra Appl., 22(4) (2023), 2350078.

[46] B. Stenstrom, *Coherent rings and FP-injective modules*, J. London Math. Soc., 2 (1970), 323–329.

[47] M. Tamekkante, K. Louartiti and M. Chhiti, *Conditions in amalgamated algebras along an ideal*, Arab. J. Math., 2(4) (2013), 403–408.

[48] F. Wang and H. Kim, *Foundations of Commutative Rings and Their Modules*, Algebra and Applications, 22, Springer, Singapore, (2016).

[49] R. B. Warfield, *Decomposability of finitely presented modules*, Proc. Amer. Math. Soc., 25 (1970), 167–172.

[50] X. Y. Yang, *Generalized Noetherian property of rings and modules*, Lanzhou: Northwest Normal University Library, 2006.

[51] W. Zhao, F. Wang and X. Zhang, *On ϕ-projective modules and ϕ-Prüfer rings*, Comm. Algebra, 48(7) (2020), 3079–3090.

[52] X. Zhang, S. Q. Xing and W. Qi, *Strongly ϕ-flat modules, strongly nonnil-injective modules and their homology dimensions*, (2022), arXiv preprint arXiv:2211.14681.

[53] X. Zhang and W. Zhao, *On nonnil-injective modules*, J. Sichuan Normal Univ., 42(6) (2009), 808–815.

[54] W. Zhao, *On ϕ-exact sequences and ϕ-projective modules*, J. Korean Math. Soc., 58(6) (2021), 1513–1528.

[55] W. Zhao, *On ϕ-flat modules and ϕ-Prüfer rings*, J. Korean Math. Soc., 55(5) (2018), 1221–1233.

[56] W. Zhao, F. Wang and G. Tang, *On ϕ-von Neumann regular rings*, J. Korean Math. Soc., 50(1) (2013), 219–229.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 115-169

ISSN: 2820-7114

Title :

Reflecting on ellipses and hyperbolas

Author(s):

David E. Dobbs

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 115–169

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco ISSN: 2820-7114

Reflecting on ellipses and hyperbolas

David E. Dobbs
Department of Mathematics, University of Tennessee, Knoxville, Tennessee 37996-1320
e-mail: *ddobbs1@utk.edu*

Communicated by Najib mahdou
(Received 02 August 2023, Revised 31 October 2023, Accepted 01 November 2023)

Abstract. Variants of some known results are proven in this teaching note. Proofs of the reflection property of ellipses and the reflection properties of hyperbolas are given; those proofs would be accessible early in a calculus class or in a course that combines precalculus with an introduction to differential calculus. Also, those reflection properties play key roles in results characterizing ellipses and hyperbolas; their proofs solve initial value problems concerning certain first order ordinary differential equations, and so they would be accessible early in a course on differential equations or in some courses on integral calculus. Special attention is paid to identifying some piecewise linear, not necessarily connected, degenerate cases resulting from proofs of the characterization results.

Key Words: Euclidean analytic geometry, ellipse, hyperbola, focus, tangential half-line, vector, dot product, inverse cosine function, derivative, initial value problem.

2010 MSC: Primary 51-02; Secondary 51N20, 33B10, 97G70, 26A06, 34-01.

1 Introduction

In [3], we proved the reflection property of a parabola and also showed how to use that reflection property in a result which characterized parabolas and arcs thereof. The present paper will accomplish the same goals for the other two main kinds of conic sections, namely, ellipses and hyperbolas. Also in the spirit of [3], we will address the piecewise-linear degenerate cases which formally also satisfy the reflection property of an ellipse or the reflection properties of a hyperbola. Our results on ellipses (resp., hyperbolas) can be found in Sections 2 and 3 (resp., Sections 4 and 5).

As one may expect, the treatment of hyperbolas will be more time-consuming than the corresponding treatment of ellipses, largely owing to the disconnected nature of a hyperbolic graph. The existence of disconnected branches of a hyperbola leads naturally to the question whether an individual branch of a hyperbola can be characterized by using a reflection property, and our methods produce an affirmative answer. As was the case in [3], we have chosen accessible methods. Our proof in Theorem 2.1 that an ellipse satisfies a certain reflection property will use a description of an ellipse via a well known pair of parametric equations, thus eliminating the need to consider only one half (or one fourth) of an ellipse at a time. However, our proof in Theorem 4.1 that a hyperbola satisfies a certain (different) reflection property will use Cartesian coordinates (as was the case in [3], in contrast to the methods using polar coordinates in, for instance, [5]). The proofs in Sections 3 and 5 explaining how to use certain reflection properties to characterize certain subsets of ellipses and hyperbolas will also use Cartesian (rather than polar) equations. Essentially all our proofs use vectorial methods that are familiar from precalculus and/or an elementary physics course. Readers seeking background information on vectorial matters are invited to read the beginning of [3, Section 2] or the parts of [4] from which that vectorial background was drawn. As in [3], the vectorial methods in our proof of the characterization result for an ellipse (resp., a hyperbola) lead to an ordinary differential equation (in short, an ODE), which is then solved in a way that could be covered early in a first course on differential equations (and even earlier in some courses on integral calculus). A careful examination of the solution of that ODE reveals the above-mentioned piecewise-linear solutions.

The Introduction to [3] reviewed the numerous ways that the topic of conic sections can (and, we argued there, *should*) play a more central role in the before-calculus mathematical curriculum. Readers interested in such pedagogical matters are invited to read the Introduction of [3] and various comments elsewhere in that paper. Since we are covering the two more complicated kinds of conics here, this paper will, for the sake of brevity, make comparatively fewer pedagogic comments.

To close the Introduction, we would like to warmly thank Dr. Michael Saum for providing, at our request, the LaTeX keystroke instructions that converted our freehand drawings into the figures that appear in this paper.

2 The reflection property of an ellipse

Except in remarks having a specifically three-dimensional context, all the work in Sections 2-5 is done inside a fixed Euclidean plane, where all "points", "lines", "rays", etc., are assumed to exist. An ellipse (in this plane) is determined by two distinct points, called foci and denoted by F_1 and F_2, together with two distinct positive real numbers $c < a$, such that the distance from F_1 to F_2 is $2c$. The distance from a point P to F_1 (resp., to F_2) will be denoted by $d_1 := d_1(P)$ (resp., by $d_2 := d_2(P)$). By definition, the *ellipse* determined by F_1, F_2, c and a (with $0 < c < a$ and the distance from F_1 to F_2 being $2c$) is the set of points P such that $d_1(P)+d_2(P) = 2a$. The graph of a typical ellipse can be found in Figure 1, where it is the oval-shaped figure.

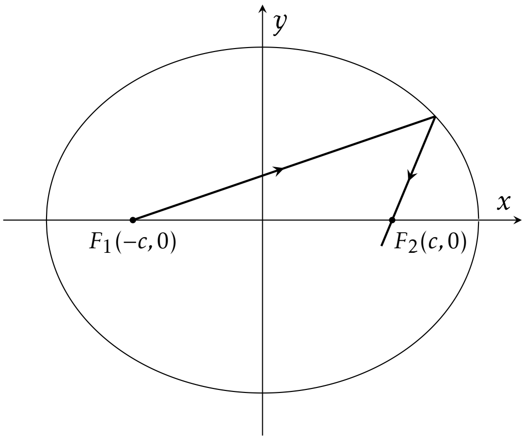


Figure 1

The various line segments or rays in Figure 1 are not part of the ellipse *per se*, but they are related to the reflection property of an ellipse, which will be discussed below.

In connection with the ellipse determined by F_1, F_2, c and a, the following usage is conventional: the line passing through F_1 and F_2 intersects the ellipse at two points which are called the *vertices* (singular: *vertex*) of the ellipse; the line segment connecting the vertices is called the *major axis* of the ellipse; it turns out that the length of the major axis is $2a$, and that is the reason that a is called the *semi-major axis* of the ellipse; the midpoint of the major axis [which turns out also to be the midpoint of the line segment connecting F_1 and F_2] is called the *center* of the ellipse (a well chosen name since any ellipse is symmetric about its center); the *minor axis* of the ellipse is the chord of the ellipse (that is, the line segment connecting two distinct points of the ellipse) which passes through the center and is perpendicular to the major axis; and the positive real number $b := \sqrt{a^2 - c^2}$ is called the *semi-minor axis* of the ellipse (a well chosen name since the length of the minor axis turns out to be $2b$).

Beginners often have difficulty in remembering the relationship between the parameters a, b and c of an ellipse, possibly because there are similarly denoted parameters of a hyperbola which satisfy a somewhat similar, but different, quadratic relationship. For an ellipse, one can recall the relationship $c^2 + b^2 = a^2$ by remembering that Pythagoras' Theorem applies to each of the four right triangles whose vertices are the center, a focus, and a point of the ellipse that is on the minor axis (the fortunate circumstance being that in each of those triangles, the hypotenuse has length a). Please notice that neither of the vertices of an ellipse is a vertex of any of those four triangles.

The notion of a tangential vector was useful in [3] and it will be useful here, too. For the sake of completeness, we next recall its definition. Let P be a point on the graph of a differentiable function f whose domain is some set $I \subseteq \mathbb{R}$. By a *tangential vector to f at P*, we mean a (bound) vector $\mathcal{T} = \overrightarrow{PQ}$, where Q is a point on the tangent line to the graph of f at P such that $Q \neq P$. (The corresponding ray $\overrightarrow{PQ}$ is sometimes called a *tangential half-line of f at P*.)

Let us consider what role the above concept can play in helping us to decide how to state a reflection property which is satisfied by an ellipse. Any such statement should be informed by the answer to the more general question of what is meant by a "reflection property". As every student of a first course in physics knows (when working in an "ideal" situation), the guiding "Principle of Reflection" asserts that "the angle of incidence is congruent to the angle of reflection". In practice, physicists measure the just-mentioned angles "from the normal" (with "normal" being perpendicular to "tangential"). However, given our familiarity with differential calculus, mathematicians tend to prefer to use tangent lines rather than normal lines. I suggest that the reader apply the "Principle of reflection" to a few positions (that seem intuitively "random/typical") for a point P on the ellipse in Figure 1 and then convert his/her conclusions about normal lines to conclusions about tangent lines. (Here are two familiar facts from plane Euclidean geometry that will help in carrying out that conversion: vertically opposite angles (nowadays often called "vertical angles") are congruent; and any two supplementary angles (nowadays often called "the members of any linear pair") have radian measures whose sum is π.) I trust that readers will agree with me that the reflection property which is depicted in Figure 1 can be formulated as follows. If P is a point on an ellipse $\mathcal{E}$ and $\mathcal{T}$ is a tangential vector (at P) to a function whose graph includes a nontrivial arc of $\mathcal{E}$ containing P, then the angle between $\mathcal{T}$ and (the bound vector) $\overrightarrow{PF_1}$ is congruent to the angle between the *opposite of* $\mathcal{T}$ and (the bound vector) $\overrightarrow{PF_2}$.

Let me elaborate on how I intuited the preceding formulation and how vectorial reasoning can reformulate it. (The rest of this paragraph includes some of the necessary background that I mentioned could be found in [3] or [4], along with what some may regard as the beginning of a proof of Theorem 2.1.) In considering what I called "a few positions (that seem intuitively 'random/typical') for a point P on the ellipse", the reader likely noticed that the angle between $\mathcal{T}$ and $\overrightarrow{PF_1}$ is acute (resp., obtuse; resp., a right angle) if and only if the angle between $\mathcal{T}$ and $\overrightarrow{PF_2}$ is obtuse (resp., acute; resp., a right angle). (We are using the following facts. Let $\mathbf{v}$ and $\mathbf{w}$ be two nonzero nonparallel (hence, distinct) bound vectors with the same initial point. Then $\mathbf{v}$ and $\mathbf{w}$ are perpendicular if and only if the dot product $\mathbf{v}\cdot\mathbf{w}$ is 0. Assume henceforth that $\mathbf{v}$ and $\mathbf{w}$ are not perpendicular (that is, assume that $\mathbf{v}\cdot\mathbf{w} \neq 0$). Since $\mathbf{v}$ and $\mathbf{w}$ are nonzero and not parallel, it is clear (up to congruence of angles) as to what is meant by "the angle between $\mathbf{v}$ and $\mathbf{w}$" [if it does not seem absolutely clear, stipulate also that the angle in question has radian measure, say θ, strictly between 0 and π]. It is known (by, and essentially equivalent to, the Law of Cosines) that

$$\cos(\theta) = \frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}| \cdot |\mathbf{w}|}.$$

Consequently, since *the opposite of* $\mathbf{v}$ is $-\mathbf{v}$ and $|-\mathbf{v}| = -|\mathbf{v}|$, it follows that if φ is the angle between $-\mathbf{v}$ and $\mathbf{w}$, then $\cos(\varphi) = -\cos(\theta)$. In view of the identity $\cos(\pi - x) = -\cos(x)$ and the fact that the function $\cos|_{[0,\pi]}$ is one-to-one, we see that $\varphi = \pi - \theta$. In particular, θ is acute (resp., obtuse) if and only if φ is obtuse (resp., acute).) The preceding parenthetical information also serves to explain that if $\angle_1$ and $\angle_2$ are angles with radian measures strictly between 0 and π, then: $\angle_1$ is congruent to $\angle_2 \Leftrightarrow \cos(\angle_1) = \cos(\angle_2)$. It is important, in my opinion, to highlight the role of the inverse cosine function in the development of this theory. Indeed, the last equivalence depended crucially on the observation (using the above notation) that

$$\theta = \cos^{-1}(\cos(\theta)) = \cos^{-1}(\frac{\mathbf{v}\cdot\mathbf{w}}{|\mathbf{v}|\cdot|\mathbf{w}|}).$$

It is now clear that (apart from possibly a few exceptional cases that will be handled individually in the proofs below) the reflection property of an ellipse is equivalent to the statement that for any point on $\mathcal{E}$, a tangential vector $\mathcal{T}$ to (a function describing) $\mathcal{E}$ at P satisfies

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\mathcal{T}|\cdot|\overrightarrow{PF_1}|} = \frac{(-\mathcal{T})\cdot\overrightarrow{PF_2}}{|-\mathcal{T}|\cdot|\overrightarrow{PF_2}|}.$$

For the purposes needed below, we can ignore the possibility that $\mathcal{T} = \mathbf{0}$. Hence, since $|-\mathcal{T}| = |\mathcal{T}| \neq 0$, the last displayed equation is equivalent to

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

To use the just-displayed equation efficiently, it will be very helpful to have a Cartesian equation for an ellipse “in standard form", and so we turn to that matter next.

By fundamental principles of Euclidean geometry, the distance between two points is unchanged when the coordinates of those points are reinterpreted in terms of new coordinate axes that have arisen as a result of a (rigid) rotation and/or a translation of the original coordinate axes. Thus, such changes of coordinate axes do not affect what it means to be *the* ellipse $\mathcal{E}$ determined by preassigned distinct foci F_1 and F_2 and preassigned positive real numbers $c < a$. Since we have seen that the statement of the reflection property of an ellipse can be formulated so as to assert that two angles of radian measure between 0 and π are congruent (that is, that the cosines of those angles are equal), it follows that a rotation and/or a translation of coordinate axes does not affect whether an ellipse $\mathcal{E}$ satisfies the reflection property in question, since it follows from (other) fundamental principles of Euclidean geometry that the measure of an angle formed at the common initial point of two rays is unchanged by rotations or translations of coordinate axes. Therefore, in proving that an ellipse *does* satisfy the relevant reflection property in Theorem 2.1, we will be able to assume that, without loss of generality, a Cartesian equation for $\mathcal{E}$ has been obtained after appropriate rotations and/or translations of coordinate axes.

Given $\mathcal{E}$ as above, consider the following transformations. Rotate the coordinate axes so that the major axis of $\mathcal{E}$ lies along a horizontal line (and, necessarily, the minor axis of $\mathcal{E}$ lies along a vertical line), then translate the x-axis vertically so that F_1 and F_2 are on the (newly-named) x-axis, and then translate the y-axis horizontally so that the (newly-named) y-axis intersects the x-axis at a point that is exactly half-way between F_1 and F_2. It is well known that (in terms of these new coordinate axes) a Cartesian equation for $\mathcal{E}$ is then

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

An ellipse having a Cartesian equation of the just-displayed kind will be said to be in *standard form*. Such an ellipse is an example of an “east-west ellipse" (also known as a “horizontal ellipse"), so-named because its major axis is horizontal (while its minor axis is vertical). Notice that if an ellipse has the just-displayed equation, then: its center is the origin; its vertices are at $(-a,0)$ and $(a,0)$; the endpoints of its minor axis are $(0,-b)$ and $(0,b)$; and its foci are classically labeled (relabeled, if necessary) as $F_1(-c,0)$ and $F_2(c,0)$.

There is a good reason why that the above discussion did not include a proof that "$d_1(P)+d_2(P) = 2a$" is equivalent (after appropriate rotations and/or translations of coordinate axes) to "$x^2/a^2 + y^2/b^2 = 1$." The sad fact (which I realized only when writing this paper) is that many (but not all) textbooks which purport to prove this equivalence actually prove only that the first condition implies the second condition. To avoid further interruptions to the presentation of Theorem 2.1, I will defer additional comments about this situation to Remark 2.2 (a).

We next prove the reflection property of an ellipse. The statement of Theorem 2.1 can be used to interpret the meaning of the segments/rays in Figure 1.

Theorem 2.1. Let $\mathcal{E}$ be an ellipse with foci F_1 and F_2. Then the following assertion is a consequence of the above "Principle of reflection". Let $\{i,j\} = \{1,2\}$. Let $\overrightarrow{R}$ be a ray which is emitted from F_i and meets $\mathcal{E}$ at a point P. Then the "reflected" ray which results from that intersection stays "inside" $\mathcal{E}$ (for a while), going on a line of action which passes through F_j.

Proof. By the above comments, the assertion is equivalent to proving that

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

We will proceed to prove this condition.

As explained above, we can assume, without loss of generality, that the coordinate axes have been suitably rotated and/or translated so that $\mathcal{E}$ is an east-west ellipse whose center is at the origin. Therefore, $\mathcal{E}$ has a Cartesian equation $x^2/a^2 + y^2/b^2 = 1$, where $a > 0$, the distance between F_1 and F_2 is $2c > 0$, and $b = \sqrt{a^2 - c^2} > 0$. It is well known that, under these conditions, $\mathcal{E}$ can be described by

$$x = a\cos(t) \text{ and } y = b\sin(t) \text{ for } 0 \le t < 2\pi.$$

(At one time, it was common to refer to the parameter t in the preceding equations as the "eccentric angle" of the ellipse (cf. [4, page 593]), but that usage seems to be much less current now.) We have the derivatives $x'(t) = -a\sin(t)$ and $y'(t) = b\cos(t)$. Thus, if $0 < t < 2\pi$ and $t \neq \pi$, it follows from the chain rule that the slope of the tangent line to $\mathcal{E}$ at a point $P(x,y)$ on $\mathcal{E}$ is

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{y'(t)}{x'(t)} = -\frac{b\cos(t)}{a\sin(t)}.$$

Hence, if $0 < t < 2\pi$ and $t \neq \pi$, we can (cf. [3, Remark 3.2 (d)]) take the tangential vector $\mathcal{T}$ to $\mathcal{E}$ at P to be any nonzero vector having the line of action of

$$\mathbf{i} + (\frac{dy}{dx})\mathbf{j} = \mathbf{i} + (\frac{b\cos(t)}{-a\sin(t)})\mathbf{j}.$$

It will be convenient to take

$$\mathcal{T} = -a\sin(t)\mathbf{i} + b\cos(t)\mathbf{j}.$$

Notice that $\mathcal{T} = x'(t)\mathbf{i}+y'(t)\mathbf{j}$. The next paragraph explains why this formula for $\mathcal{T}$ would be appropriate even in situations where $x'(t) = 0$ (that is, for $\mathcal{E}$, even when t is 0 or π, the values of the parameter t giving a point $P(x,y)$ at which the tangent line to $\mathcal{E}$ is vertical). Notice also that $\mathcal{T} \neq \mathbf{0}$. Indeed, if $\mathcal{T} = \mathbf{0}$, then $x'(t) = 0 = y'(t)$ (equivalently, $-a\sin(t) = 0 = b\cos(t)$; equivalently, $\sin(t) = 0 = \cos(t)$), whence $\sin^2(t)+\cos^2(t) = 0^2+0^2 = 0$, contradicting the Pythagorean identity that $\sin^2(t)+\cos^2(t) = 1$.

Fix any real number ε such that $0 < \varepsilon < \pi/2$. Let I be the open interval $(-\varepsilon, 2\pi + \varepsilon)$. Consider the function $f : I \to \mathbb{R}^2$ defined by

$$f(t) := (a\cos(t), b\sin(t)) \text{ for } -\varepsilon < t < 2\pi + \varepsilon.$$

According to the "modern" definition of a differentiable function (cf. [7, Definition, page 38]), the function f is differentiable on the closed interval $[0, 2\pi]$ (the point being that since f is defined on an open interval containing $[0, 2\pi]$, continuous and of class $C^{(1)}$, the "differentiable" conclusion follows from [7, Theorem 2, page 4]). The "modern" definition of a tangent vector, as given in [7, page 73], would then yield that the tangent vector to $\mathcal{E}$ at a point $P(a\cos(t), b\sin(t))$ on $\mathcal{E}$ is $-a\sin(t)\mathbf{i} + b\cos(t)\mathbf{j}$ (which is the same formula that we got above for the tangential vector $\mathcal{T}$ in case the tangent line to $\mathcal{E}$ at P is not vertical). The classical and modern approaches are compatible, as it is explained in [7, page 73] how the modern approach to a tangent vector leads to the classical notion of a tangent line (regardless of whether that line is vertical).

The just-mentioned "modern" approaches to differentiable functions and tangent vectors have been standard in most textbooks on advanced calculus for nearly 60 years. The "classical" (I prefer this term instead of "old-fashioned") approach to differentiable functions can be found in a number of textbooks on advanced calculus (cf. [11, pages 267-268]). In the interest of accessibility, we will directly verify the assertion when $t = 0$ and when $t = \pi$, as these are the values of the parameter t giving the points $P_1(a, 0)$ and $P_2(-a, 0)$ where $\mathcal{E}$ has a vertical tangent line. In that regard, recall the following somewhat standard definition (the literature is surprisingly nonuniform about this matter!): if x_0 is in the domain of a real-valued function h of one real variable, then the graph of h has a vertical tangent line at $(x_0, h(x_0))$ if h is continuous at x_0 and $\lim_{x\to x_0} h^{'}(x) = \pm\infty$. We will verify that this condition holds at P_1, leaving it to the reader to provide the similar verification for the tangential behavior at the point P_2. Let us use the fact that P_1 is on the upper half of $\mathcal{E}$, which is the graph of the function h given by

$$y = h(x) = (\frac{b}{a})\sqrt{a^2 - x^2},$$

with the point P_1 having coordinates $(a, 0) = (a, h(a))$. It is evident that h is continuous at a. As $h^{'}(x) = -bx/(a\sqrt{a^2 - x^2})$ if $|x| < a$,

$$\lim_{x\to a} h^{'}(x) = \lim_{x\to a^-} \frac{-bx}{a\sqrt{a^2 - x^2}} = -ba/0^+ = -\infty,$$

as desired.

It will be useful to know that neither F_1 nor F_2 is a point on $\mathcal{E}$. We will prove this fact for F_1, leaving the details of the similar proof about F_2 to the reader. For a proof, it suffices to use the hypothesis that $c < a$ to get that $d_1(F_1) + d_2(F_1) = 0 + 2c = 2c \neq 2a$. Hence $|\overrightarrow{PF_1}| \neq 0$ and, similarly, $|\overrightarrow{PF_2}| \neq 0$.

For $k \in \{1, 2\}$, the work two paragraphs ago lets us take $\mathcal{T} := \mathcal{T}_k$, the tangential vector to $\mathcal{E}$ at P_k, to be the bound vector that has initial point P_k and is equivalent to $\mathbf{j}$. The angle between $\mathcal{T}_k$ and $\overrightarrow{P_kF_1}$ (resp., between $\mathcal{T}_k$ and $\overrightarrow{P_kF_2}$) is a right angle. So, the angle between the *opposite* of $\mathcal{T}_k$ and $\overrightarrow{P_kF_2}$ is also a right angle. As it is a fundamental principle of Euclidean geometry that any two right angles are congruent, this completes a direct proof that the assertion holds at P_1 and P_2, that is, when t is either 0 or π. (Interested readers are invited to fashion an alternate direct proof of this assertion that uses the Principle of Reflection and measures angles "from the normal.")

It now remains to prove that

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}$$

if t is such that $0 < t < 2\pi$ and $t \neq \pi$. Under these conditions, which will be assumed for the rest of this proof, it will be convenient to denote the left- and right-hand sides of the desired equation, respectively, by

$$\mathcal{L} := \frac{\mathcal{T} \cdot \overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} \text{ and } \mathcal{R} := -\frac{\mathcal{T} \cdot \overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

We have $\mathcal{T} = -a\sin(t)\mathbf{i} + b\cos(t)\mathbf{j}$,

$$\overrightarrow{PF_1} = (-c-x)\mathbf{i} + (0-y)\mathbf{j} = (-c - a\cos(t))\mathbf{i} - b\sin(t)\mathbf{j},$$

$$|\overrightarrow{PF_1}| = \sqrt{(-c-a\cos(t))^2 + (-b\sin(t))^2} = \sqrt{(c+a\cos(t))^2 + b^2\sin^2(t)},$$

$$\overrightarrow{PF_2} = (c-x)\mathbf{i} + (0-y)\mathbf{j} = (c - a\cos(t))\mathbf{i} - b\sin(t)\mathbf{j}, \text{ and}$$

$$|\overrightarrow{PF_2}| = \sqrt{(c-a\cos(t))^2 + (-b\sin(t))^2} = \sqrt{(c-a\cos(t))^2 + b^2\sin^2(t)}.$$

We will algebraically simplify $\mathcal{L}$ and $\mathcal{M}$ separately, and then show that those simplifications agree. Using the standard formula for dot product, we get

$$\mathcal{L} = \frac{a\sin(t)[c + a\cos(t)] - b^2\cos(t)\sin(t)}{\sqrt{(c+a\cos(t))^2 + b^2\sin^2(t)}}.$$

Then, by using the identity $c^2 = a^2 - b^2$ (resp., $\cos^2(t) = 1 - \sin^2(t)$) to rewrite the numerator (resp., denominator) of the right-hand side of the last display, we get

$$\mathcal{L} = \frac{ac\sin(t) + c^2\sin(t)\cos(t)}{\sqrt{c^2 + 2ca\cos(t) + a^2\cos^2(t) + b^2\sin^2(t)}} =$$

$$\frac{c\sin(t)[a + c\cos(t)]}{\sqrt{c^2(1-\sin^2(t)) + a^2(\cos^2(t) + \sin^2(t)) + 2ca\cos(t)}}.$$

We can further rewrite the denominator of the last display as

$$\sqrt{c^2\cos^2(t) + a^2 + 2ca\cos(t)} = \sqrt{(c\cos(t) + a)^2} = |c\cos(t) + a|.$$

We thus have the following rewriting of $\mathcal{L}$:

$$\mathcal{L} = \frac{c\sin(t)[a + c\cos(t)]}{|c\cos(t) + a|}.$$

Next, let us simplify $\mathcal{M}$ by tweaking the method that was just used to rewrite $\mathcal{L}$. We get

$$\mathcal{M} = \frac{ac\sin(t) - (a^2 - b^2)\sin(t)\cos(t)}{\sqrt{c^2 - 2ca\cos(t) + a^2\cos^2(t) + (a^2 - c^2)\sin^2(t)}} =$$

$$\frac{c\sin(t)[a - c\cos(t)]}{\sqrt{c^2(1-\sin^2(t)) + a^2(\cos^2(t) + \sin^2(t)) - 2ca\cos(t)}}.$$

We can further rewrite the denominator of the last display as

$$\sqrt{c^2\cos^2(t)+a^2-2ca\cos(t)}=\sqrt{(c\cos(t)-a)^2}=|c\cos(t)-a|.$$

We thus have the following rewriting of $\mathcal{M}$:

$$\mathcal{M}=\frac{c\sin(t)[a-c\cos(t)]}{|c\cos(t)-a|}.$$

It remains only to show that the above simplifications of $\mathcal{L}$ and $\mathcal{M}$ agree. Since $c>0$ and $t\notin\{0,\pi\}$ ensures that $\sin(t)\neq 0$, we can see, by factoring $c\sin(t)$ out of both $\mathcal{L}$ and $\mathcal{M}$, that the assertion $\mathcal{L}=\mathcal{M}$ is equivalent to the following assertion:

$$\frac{a+c\cos(t)}{|c\cos(t)+a|}=\frac{a-c\cos(t)}{|c\cos(t)-a|}.$$

The equality asserted in the last display *does* hold, as both its left- and right-hand sides equal 1. Indeed, as both its left- and right-hand sides are elements of $\{-1,1\}$, it will suffice to prove that $a+c\cos(t)>0$ and $a-c\cos(t)>0$. Both of these inequalities follow from the facts that $0<c<a$ and $-1\leq\cos(t)\leq 1$, together with the familiar rules on how to transform an inequality when both sides of the inequality are multiplied by the same nonzero real number. In detail:

$$a+c\cos(t)\geq a+c\cdot(-1)=a-c>0 \text{ and}$$

$$a-c\cos(t)=a+(-c)\cdot\cos(t)\geq a+(-c)\cdot 1=a-c>0.$$

The proof is complete. □

The next remark collects a variety of observations.

Remark 2.2. (a) Suppose that $\mathcal{E}$ is the ellipse determined by the foci $F_1(-c,0)$ and $F_2(c,0)$, together with the positive parameters $c<a$. As above, define $b:=\sqrt{a^2-c^2}$. Any number of textbooks on precalculus or calculus prove that "$d_1(P)+d_2(P)=2a$" implies "$x^2/a^2+y^2/b^2=1$." For instance, see the detailed proof in [4, pages 454-455]. (I apologize for the typo on [4, page 455, line 13], where "P_2" should be "F_2".) It is not immediately clear whether two of the steps in that proof can be reversed, as both of those steps are of the kind "$A=B\Rightarrow A^2=B^2$". In preparing this paper, I was surprised to discover that [4] does not contain a proof that "$x^2/a^2+y^2/b^2=1$" implies "$d_1(P)+d_2(P)=2a$". As the first draft of [4] was completed during the period February 1986 - September 1987 and the material surrounding [4, page 455] was not altered in later drafts, the passage of some 37 years has caused me to forget whether the authors of [4] did not include the somewhat difficult proof of the latter implication for pedagogical reasons (that is, because of the perceived level of precalculus course in which [4] would typically be used) or whether the authors of [4] did not realize that a complete proof of the equivalence requires a proof of that converse. Be that as it may, some other textbooks that introduce the standard-form equation "$x^2/a^2+y^2/b^2=1$" for an ellipse (with appropriately positioned coordinate axes) also omit the proof of that converse. One might also query the authors of *those* textbooks as to whether their omission was made because of pedagogical reasons or because of a mathematical blunder, but many such authors may be as forgetful as I (and many of those authors are deceased). Sadly, I recently discovered this kind of omission in the textbook from which I had learned (what is now called) precalculus in 1960-61, in a course whose name was something like "Algebra, trigonometry and analytic geometry". In fact, that textbook was so old-fashioned that it used "distance" to mean "directed distance", a concept which could be positive, negative or zero, so that the equations which I referred to above as being of the kind "$A=B$" were written in that

textbook as being of the kind "$\pm A = B$", which *obviously* is equivalent to "$A^2 = B^2$", albeit at the cost of contradicting our current perspective that "distance" in a metric space cannot be negative. This observation leads me to caution any readers who may consult an old work on analytic geometry to ascertain whether its author permits "distance" to be negative.

I also recently discovered that in [9, page 783; and Exercises 42-43, page 791], which I regard as one of the finest textbooks for non-honors calculus courses, Leithold sketched a couple of proofs of the converse. What seems to me like a more elegant way of presenting one of those proofs of the converse was given earlier by Johnson and Kiokemeister [8, Exercise 15, page 176] in an exercise whose beautiful hint essentially comes down to noting that "$x^2/a^2+y^2/b^2 = 1$" implies that $a^2 \pm cx \geq 0$ (the point being that $\pm cx \leq c|x| \leq ca < a^2$). Interested readers may scrutinize the ending passages of our proofs of Theorems 2.1 and 3.1 to see whether some version of the insight that "$\pm cx \leq ca$" can be found there. As I proved those theorems before consulting any of [4], [9] or [8] in regard to a complete proof that "$x^2/a^2 + y^2/b^2 = 1$" implies "$d_1(P) + d_2(P) = 2a$", I should perhaps revise my description of the proof of this implication as being "somewhat difficult". In any event, let us charitably agree that all the authors (except, possibly, yours truly) who omitted a proof of the "somewhat difficult" converse did so intentionally for pedagogical reasons.

(b) Many instructors have been asked by students to explain why circles are not considered to be special cases of ellipses. Often, students advance the following argument: "The circle of radius a with center at the origin is the graph of the equation $x^2 + y^2 = a^2$; *this* equation is equivalent to the equation $x^2/a^2 + y^2/a^2 = 1$; and *that* is the equation of an ellipse for the special case where $a = b$." It is important to recognize why anyone adhering to the formal definition of an ellipse must conclude that this three-step argument is fallacious, for the following reason(s). The fault lies entirely in the third step of the argument. There simply is no special case of an ellipse where $a = b$. After all, since b was defined as $\sqrt{a^2 - c^2}$, saying that $a = b$ would be equivalent to saying that $c = 0$, and we began the study of ellipses at the beginning of this section with several stipulations, one of which was that c is a positive real number.

A stronger, more creative student has been known to reply to the above explanation with the following questions and remarks: "Why can't we assume that $c = 0$ for an ellipse? That would just force the foci F_1 and F_2 to be the same point, say F. Wouldn't that point F just be the center of a circle? After all, if $F_1 = F = F_2$, then the set of points P such that $d_1(P) + d_2(P) = 2a$ and $d_1(P) = d_2(P)$ is just the set of points P whose distance from F is a, and *that is a circle*, with center F and radius a." There is much that is correct and perceptive in such a student's comment. However, an adherent of the formal definition of an ellipse would insist that this student is also wrong, by offering the following comments. We cannot "just force the foci F_1 and F_2 to be the same point" because we stipulated at the outset that the foci F_1 and F_2 are distinct points. This rebuttal is unlikely to satisfy the student, who is apt to respond by asking, "Why do we *have to* stipulate that $F_1 \neq F_2$?" We should welcome such a creative and persistent student by providing the following comment. The equation that you have been mentioning, $x^2/a^2 + y^2/b^2 = 1$, is actually obtained logically after several steps, starting with $d_1(P) + d_2(P) = 2a$, where the next-to-last step is the inference that

$$(a^2 - c^2)x^2 + a^2y^2 = a^2(a^2 - c^2).$$

(See, for instance, [4, page 455, line 10] or [8, page 174, line 5].) So, since you have decided to assume that $c = 0$ and presumably $a \neq 0$, you have that $a^2(a^2 - c^2) = a^4 \neq 0$, and dividing through by this nonzero number *would* give you the equation $x^2/a^2 + y^2/a^2 = 1$, and that equation *is just* $x^2/a^2 + y^2/b^2 = 1$ in case $a = b$. And, checking the hint for the exercise proving the converse that we discussed in the textbook of Johnson and Kiokemeister, I can see that "$x^2/a^2 + y^2/b^2 = 1$" *does* imply "$d_I(P) + d_2(P) = 2a$" *regardless of whether $a = b$*; that is, regardless of whether $c = 0$; that is, regardless of whether $F_1 = F_2$. So, I suppose that we could agree that a circle is a special case of an ellipse where $c = 0$, that is, where the foci F_1 and F_2 are one and the same point. But I am not sure whether Johnson and Kiokemeister would agree (and, unfortunately, they are no longer available to be contacted here on Earth). When they defined an ellipse in their textbook, they did refer to F_1 and F_2 as "two points", not as "two distinct points." But if you check mathematical writing style from the 1950s and 1960s, you will find that few writers at that time were as careful as most of us are nowadays to say "n distinct things" or even "n pairwise distinct things" instead of just "n things". So, you have won me over and I agree that a circle is a special case of an ellipse.

(c) The sort of conversation that was described in (b) has happened several times in my experience as a teacher. It has never been followed up by what I will mention next, but I offer the following observation for the reader's consideration. Suppose that we decided to take $a = c$ in the last-displayed equation. Since that equation had been inferred from "$d_1(P) + d_2(P) = 2a$", that would mean that, in some sense, we can conclude that the "degenerate case" of an ellipse where $a = c$ (that is, where $b = 0$) satisfies the condition $a^2y^2 = 0$, or equivalently, $y^2 = 0$. In other words, there is a subset $\mathcal{S}$ of the Euclidean plane $\mathbb{R}^2$ consisting of some points $P(x,0)$ on the x-axis such that "$d_1(P) + d_2(P) = 2c$", that is, such that $|x+c|+|x-c| = 2c$. This shows that the closed interval $[-c,c]$, when regarded as a subset of the Euclidean plane, should be regarded as a degenerate case of an ellipse such that $a = c$. In fact, any subset of that interval could be regarded in the same way! We will have more to say about degenerate cases of ellipses in Section 3, where *those* degenerate cases will arise by examining the implications of a figure satisfying the reflection property of an ellipse.

(d) There is a certain harmony connecting the various approaches to ellipses that we are considering here. If a reader does not wish to wait until the next section to find degenerate cases of ellipses, the above experiences in (b) and (c) may suggest that it would be worthwhile to look closely at the proof of Theorem 2.1 to see if some exceptional situations in that proof needed (or, at least, received) separate treatment. Any such "close look" brings attention to the vertices, $(-a,0)$ and $(a,0)$, as these were the points on the ellipse $\mathcal{E}$, with equation $x^2/a^2 + y^2/b^2 = 1$, at which the tangent line to $\mathcal{E}$ is vertical. Recall how the proof of Theorem 2.1 handled such points. The desired reflection property was shown to hold at each vertex because the tangent line to $\mathcal{E}$ at a vertex V was vertical, so that both a preassigned vertical bound (tangential) vector emanating from the vertex V and its opposite (bound) vector each made right angles with both $\overrightarrow{VF_1}$ and $\overrightarrow{VF_2}$. (The proof then concluded by appealing to the fundamental principle of Euclidean geometry that all right angles are congruent.) One is led to ask the following question: are there any planar figures $\mathcal{G}$ *all* of whose points behave in a similar way? More precisely put: is there an accessible example of a planar figure $\mathcal{G}$ with an associated planar point F such that F is not on $\mathcal{G}$ and, for *each* point P on $\mathcal{G}$, the line connecting F to P is the normal line to $\mathcal{G}$ at P (that is, such that F is not on $\mathcal{G}$ and, for *each* point P on $\mathcal{G}$, the line connecting F to P is perpendicular to the tangent line to $\mathcal{G}$ at P)? As any student of Euclidean plane geometry knows, there is an obvious family of such figures $\mathcal{G}$, namely, circles, the relevant fact being that the radius vector from the center of a given circle $\mathcal{C}$ to any point P on $\mathcal{C}$ is perpendicular to the tangent line to $\mathcal{C}$ at P. While it may be heartening that our question in (d) has led to contact with Euclid's *Elements*, some readers may be sad that these considerations in (d) have not revealed any additional special kinds of ellipses or any additional degenerate cases of ellipses. (After all, we settled the matter of circles in (b), didn't we?) It now seems natural to ask if the characterization results in Section 3 will be more successful in revealing additional kinds of ellipses or degenerate cases of ellipses. With the next section having been thus motivated, the remark is complete.

To close this section, we mention a pair of physical applications of the reflection property of an ellipse, or more precisely, of a (three-dimensional) ellipsoid of revolution. Perhaps, the most familiar application of this property can be found in "whispering galleries". I will mention only three of the many famous examples of whispering galleries: along the circular walkway situated some 257 steps above the floor of St. Paul's Cathedral in London, England; along the balcony around the edge of the dome of the Gol Gumbaz Mausoleum in Vijayapura, India; and in the statuary hall of the Capitol in Washington, DC, USA. In such a gallery, an individual who is positioned at one of the foci (of the ellipse whose rigid rotation into three-dimensional space produced the ellipsoid of revolution) is able to overhear whatever is said by another individual who is positioned at the other focus of that ellipse. The underlying scientific principle assumes that sound is propagated along straight lines. (As we have known for more than a century, much of Newtonian physics is only approximately true in our world but it is nevertheless often very useful in everyday life.) Waves of a different sort are involved in a more recent application, the lithotripter (a machine that crushes stones in an individual's urological system), where an energy source is placed at one of the foci of the generating ellipse, an appropriate portion of an individual's anatomy is placed at the other focus of that ellipse, the energy source is activated, and the rays that are thus generated are then reflected off the surface of the ellipsoid so as to combine with a crushing effect at the other focus. (Successful uses of a modern lithotripter often require bursts of waves numbering in the thousands.) Some homework problems involving whispering galleries can be found in [4, Exercises 52-53, page 464; and Exercise 54, page 489].

3 A reflection-theoretic characterization of an ellipse

We move at once to the main result of this section. Theorem 3.1 gives a characterization of the top half of an ellipse (assumed, without loss of generality, to be of "east-west" type and in standard position) which uses the reflection properties of an ellipse that were established in Theorem 2.1. The corresponding characterization of the bottom half of such an ellipse will be given in Remark 3.2 (b).

Theorem 3.1. Let $0 < c < a$ in $\mathbb{R}$, and put $b := \sqrt{a^2 - c^2}$. Working in a fixed Euclidean plane, consider the points $F_1(-c, 0)$ and $F_2(c, 0)$. Let $f : [-a, a] \to \mathbb{R}$ be a function, and let Γ be the graph of f. Suppose that f is differentiable on $(-a, a)$, f is continuous at $x = -a$ and at $x = a$, $f(-a) = 0 = f(a)$, $f(t_1) > 0$ for some $t_1 \in (-a, 0)$, $f(t_2) > 0$ for some $t_2 \in (0, a)$, and $f^{'}(x) \neq 0$ for all $x \in (-a, 0) \cup (0, a)$. Suppose also that Γ has a vertical tangent line at both the point $W_1(-a, 0)$ and the point $W_2(a, 0)$. For each point P on Γ, let $\mathcal{T} := \mathcal{T}_P$ be a tangential vector to Γ at P. Then the following five conditions are equivalent:

(1) $f(x) = (\frac{b}{a})\sqrt{a^2 - x^2}$ for all $x \in [-a, a]$;

(2) Γ is the "top half" of the "east-west" ellipse with foci F_1 and F_2 and semi-major axis a (and necessarily with semi-minor axis b);

(3) For each point P on Γ, the angle between $\mathcal{T}_P$ and $\overrightarrow{PF_1}$ is congruent to the angle between the opposite of $\mathcal{T}_P$ and $\overrightarrow{PF_2}$;

(4) For each point P on Γ,

$$\frac{\mathcal{T}_P \cdot \overrightarrow{PF_1}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_1}|} = -\frac{\mathcal{T}_P \cdot \overrightarrow{PF_2}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_2}|};$$

(5) For each point P on Γ, the following reflection property holds: if $\overrightarrow{R}$ is a ray which is emitted from F_1 and meets Γ at P, then the "reflected" ray which results from that intersection stays "inside" Γ (for a while), going on a line of action which passes through F_2;

(6) For each point P on Γ, the following reflection property holds: if $\overrightarrow{R}$ is a ray which is emitted from F_2 and meets Γ at P, then the "reflected" ray which results from that intersection stays "inside" Γ (for a while), going on a line of action which passes through F_1.

Proof. The hypotheses have been arranged so that, when the various conditions hold, Γ will be shown to be the top half of the ellipse that is in standard position with semi-major axis a and semi-minor axis b. In particular, the assumed existence of t_1 and t_2 ensures, thanks to the Intermediate Value Theorem (for continuous functions) and Rolle's Theorem, that $f(x) > 0$ for all $x \in (-a, a)$.

Assume henceforth that P is a point on Γ, with coordinates (x,y). Let us consider first the case where x is either $-a$ or a. As explained in the previous paragraph, this situation occurs if and only if $y = 0$. Recall that the corresponding points on Γ were denoted by W_1 and W_2. By hypothesis, the tangent line to Γ at W_1 (resp., at W_2) is vertical. So, we can take the tangential vector $\mathcal{T}$ to Γ at W_1 (resp., at W_2) to be $\mathbf{j}$. We claim that all the conditions in question are satisfied if P is W_1 (resp., if P is W_2). It will suffice to show that the angle between $\mathbf{j}$ and $\overrightarrow{W_1F_1}$ (resp., the angle between $\mathbf{j}$ and $\overrightarrow{W_2F_1}$) is congruent to the angle between $-\mathbf{j}$ and $\overrightarrow{W_1F_2}$ (resp., between $-\mathbf{j}$ and $\overrightarrow{W_2F_2}$). Since $\overrightarrow{W_1F_1}$ and $\overrightarrow{W_1F_2}$ (resp., $\overrightarrow{W_2F_1}$ and $\overrightarrow{W_2F_2}$) are nonzero horizontal vectors, that is nonzero multiples of $\mathbf{i}$, all the angles in question are right angles. As it is a fundamental principle of Euclidean geometry that all right angles are congruent, the above claim has been proved. Therefore, for the rest of this proof, as we consider the behavior of points P on Γ, we can assume that P on Γ is neither W_1 nor W_2.

By the remarks early in Section 2, (1) $\Rightarrow$ (2). By Theorem 2.1, (2) implies both (5) and (6). By the vectorial background mentioned earlier (cf. also the proof of Theorem 2.1), we have that (3) $\Leftrightarrow$ (4); and also that (5) $\Leftrightarrow$ (4) $\Leftrightarrow$ (6). Hence, it remains only to prove that (4) $\Rightarrow$ (1).

Assume (4). We next focus, at least for a while, on the context $0 < x < a$. The slope, say m, of the tangent line to Γ at P is $m = f'(x)$. So, we can take the tangential vector to Γ at P to be

$$\mathcal{T} = \mathbf{i} + m\mathbf{j} = \mathbf{i} + f'(x)\mathbf{j}.$$

We also have

$$\overrightarrow{PF_1} = (-c-x)\mathbf{i} - y\,\mathbf{j} \text{ and } \overrightarrow{PF_2} = (c-x)\mathbf{i} - y\,\mathbf{j},$$

$$|\overrightarrow{PF_1}| = \sqrt{(-c-x)^2+(-y)^2} = \sqrt{(c+x)^2+y^2} \text{ and}$$

$$|\overrightarrow{PF_2}| = \sqrt{(c-x)^2+(-y)^2} = \sqrt{(c-x)^2+y^2}.$$

By hypothesis,

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\mathcal{T}|\cdot|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\mathcal{T}|\cdot|\overrightarrow{PF_2}|}, \text{ or equivalently,}$$

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

Using the standard formula for dot product, we can rewrite the last display as

$$\frac{1(-c-x)+f'(x)(-y)}{\sqrt{(c+x)^2+y^2}} = -\frac{1(c-x)+f'(x)(-y)}{\sqrt{(c-x)^2+y^2}}, \text{ or equivalently,}$$

$$\frac{c+x+y\frac{dy}{dx}}{\sqrt{(c+x)^2+y^2}} = \frac{c-x-y\frac{dy}{dx}}{\sqrt{(c-x)^2+y^2}}.$$

Cross-multiplying and solving for the derivative leads, after some minor algebraic rewriting, to

$$\frac{dy}{dx} = \frac{(c-x)\sqrt{c^2+2cx+x^2+y^2}-(c+x)\sqrt{c^2-2cx+x^2+y^2}}{y[\sqrt{c^2+2cx+x^2+y^2}+\sqrt{c^2-2cx+x^2+y^2}]}.$$

The numerator of the last display can be rewritten as the difference

$$c[\sqrt{c^2+2cx+x^2+y^2}-\sqrt{c^2-2cx+x^2+y^2}]-$$

$$x[\sqrt{c^2+2cx+x^2+y^2}+\sqrt{c^2-2cx+x^2+y^2}].$$

Hence, we can write

$$\frac{dy}{dx}=-\frac{x}{y}+\frac{c[A-B]}{y[B+A]},\text{ where}$$

$$A:=\sqrt{c^2+2cx+x^2+y^2}\text{ and }B:=\sqrt{c^2-2cx+x^2+y^2}.$$

Multiplying both the numerator and the denominator of the second term of the right-hand side of the last display by $B-A$ leads, after some routine (but time-consuming) algebraic simplification to

$$\frac{dy}{dx}=\frac{-x^2+c^2+y^2-\sqrt{(y^2+c^2+x^2)^2-4c^2x^2}}{2xy}.$$

The last display is reminiscent of (but different from) the ODE (ordinary differential equation) that was involved in our recent proof of a result that characterized parabolas (in what one could consider “standard position") [3, Theorem 3.1]. Let us next try to use here the substitution that had been useful in that earlier proof, namely,

$$w:=y^2+c^2+x^2.$$

Then

$$\frac{dw}{dx}=2y\frac{dy}{dx}+2x.$$

So, by substituting the just-displayed fact into the above ODE and doing some minor algebraic rewriting, we get

$$\frac{dw}{dx}-2x=2y\frac{dy}{dx}=\frac{-x^2+w-x^2-\sqrt{w^2-4c^2x^2}}{x}.$$

Let us try another substitution that was useful in the proof of [3, Theorem 3.1], namely, $v:=w/x$. Using the product rule of differential calculus and some algebraic rewriting, we get

$$\frac{dv}{dx}=-\frac{w}{x^2}+\frac{\frac{dw}{dx}}{x}=-\frac{v}{x}+\frac{\frac{w}{x}-\frac{\sqrt{w^2-4c^2x^2}}{x}}{x}=$$

$$-\frac{v}{x}+\frac{v-\sqrt{(\frac{w}{x})^2-4c^2}}{x},\text{ and so}$$

$$\frac{dv}{dx}=-\frac{\sqrt{v^2-4c^2}}{x}\quad(\text{if }0<x<a).$$

Thus, by separating variables and performing indefinite integration, we have (if $0<x<a$) that

$$\int\frac{dv}{\sqrt{v^2-4c^2}}=-\int\frac{dx}{x}+K,$$

with constant of integration K.

According to a table of (indefinite) integrals (specifically, formula 27 on the page opposite the inside front cover of [9]), if k is any positive real number,

$$\int\frac{dt}{\sqrt{t^2-k^2}}=\ln(|t+\sqrt{t^2-k^2}|)+C.$$

By applying the just-displayed formula (with $k := 2c$) to the last result of the preceding paragraph, we get

$$\ln(|v+\sqrt{v^2-4c^2}|) = -\ln(|x|)+K.$$

Next, exponentiate both sides of the last display, and then rewrite the resulting equation by using the property that $\ln(\lambda\nu) = \ln(\lambda)+\ln(\nu)$ for all positive λ and ν. This gives that $|x(v+\sqrt{v^2-4c^2})|$ is constant (for $0<x<a$). Therefore, there exists a constant $E>0$ such that

$$v+\sqrt{v^2-4c^2} = \frac{E}{x} \text{ whenever } 0<x<a.$$

Substituting $v = w/x$ into the last display, then multiplying through by x, and then using the definition of w leads to

$$y^2+c^2+x^2+\sqrt{(y^2+c^2+x^2)^2-4c^2x^2} = E \text{ if } 0<x<a.$$

Squaring both sides of the equation in the last display leads to

$$(y^2+c^2+x^2-E)^2 = (y^2+c^2+x^2)^2-4c^2x^2, \text{ whence}$$

$$(2E-4c^2)x^2+2Ey^2 = E^2-2Ec^2 \text{ if } 0<x<a.$$

If $E = 2c^2$, then the final statement of the preceding paragraph would imply that, for *any* real number x such that $0<x<a$, we would have $0\cdot x^2+2Ey^2 = E(E-2c^2) = 0$, which is a contradiction, since $E>0$ and ($0<x<a$ ensures that) $y>0$. Therefore, $E \neq 2c^2$. This fact will be of great importance below.

Consider the limit process $x \to a^-$. As $\lim_{x\to a^-} y = f(a) = 0$ by hypothesis, an application of standard limit theorems to the final result two paragraphs ago gives

$$(2E-4c^2)a^2+2E\cdot 0^2 = E^2-2Ec^2,$$

whence $(2E-4c^2)a^2 = E^2-2Ec^2$ (if $0<x<a$). Therefore,

$$\frac{x^2}{a^2} = \frac{x^2}{(\frac{E^2-2Ec^2}{2E-4c^2})} = \frac{(2E-4c^2)x^2}{E^2-2Ec^2} =$$

$$\frac{E^2-2Ec^2-2Ey^2}{E^2-2EC^2} = 1-(\frac{2E}{E^2-2Ec^2})y^2 = 1-(\frac{2}{E-2c^2})y^2.$$

Observe that

$$a^2 = \frac{E^2-2Ec^2}{2E-4c^2} = \frac{E}{2} = \frac{E-2c^2}{2}+c^2,$$

whence we get the important fact that

$$b^2 = a^2-c^2 = \frac{E-2c^2}{2}.$$

Consequently, if $0<x<a$, then

$$\frac{x^2}{a^2}+\frac{y^2}{b^2} = 1.$$

By the continuity of f, the last displayed equation also holds at a (where $y = f(a) = 0$). Since $y = f(x) \geq 0$, it follows that $f(x) = (b/a)\sqrt{a^2-x^2}$. This completes the proof that (4) implies the assertion in (1) in case $0 \leq x \leq a$. It will be useful, for the purposes of the next two paragraphs, to note that this proof (so far) has used only assumptions about the behavior of the function $f|_{[0,a]}$ and its graph.

It remains to prove (while still assuming (4)) the assertion in (1) in case $-a \leq x \leq 0$. The above remarks have already handled the subcases where x is either $-a$ or 0, and so we can assume henceforth, without loss of generality, that $-a < x < 0$. This case could be handled by reasoning as above, but it will be faster (and it will motivate some of the methods in Remark 3.2) to proceed as follows. Define a function $g : [0,a] \to \mathbb{R}$ by $g(x) := f(-x)$ whenever $0 \leq x \leq a$. Let γ be the graph of g. Then g is differentiable on $(0,a)$, since the chain rule gives $g^{'}(x) = f^{'}(-x)\cdot(-1) = -f^{'}(-x)$ whenever $0 < x < a$ (that is, whenever $-a < -x < 0$). Moreover, g is continuous at $x = 0$ and at a. Since

$$\lim_{x\to a^-} g^{'}(x) = \lim_{x\to a^-} -f^{'}(-x) = -\lim_{-x\to(-a)^+} f^{'}(-x) = -(\pm\infty) = \mp\infty,$$

it follows that γ has a vertical tangent line at the point $(a,0)$. Furthermore, $g(a) = 0$, $g(t_1) > 0$ for some $t_1 \in (0,a)$ and $g^{'}(x) \neq 0$ for all $x \in (0,a)$.

It will be convenient to write $Y = g(t)$, with $0 < t < a$. The point $Q(t,Y)$ on γ corresponds to the point $P(x,y)$ on Γ, where $x = -t \in (-a,0)$ and $y = f(x) = f(-t) = g(t) = Y$. Let $m := f^{'}(x)$. Then $g^{'}(t) = -f^{'}(-t) = -f^{'}(x) = -m$. For our present purposes, we can take $\mathcal{T} = \mathcal{T}_P = \mathbf{i} + m\mathbf{j}$. Similarly, we can take the tangential vector to/of γ at Q to be $\mathcal{U} := \mathcal{U}_Q := \mathbf{i} + (-m)\mathbf{j} = \mathbf{i} - m\mathbf{j}$. We claim that g (or, if you wish, γ) satisfies the analogue of condition (4) relative to the interval $[0,a]$; that is, we claim that

$$\frac{\mathcal{U}_Q\cdot\overrightarrow{QF_1}}{|\overrightarrow{QF_1}|} = -\frac{\mathcal{U}_Q\cdot\overrightarrow{QF_2}}{|\overrightarrow{QF_2}|};$$

that is, we claim that

$$\frac{1(-c-t)+(-m)(0-Y)}{\sqrt{(-c-t)^2+(0-Y)^2}} = -\frac{1(c-t)+(-m)(0-Y)}{\sqrt{(c-t)^2+(0-Y)^2}}.$$

On the other hand, because we are assuming (4), we have that

$$\frac{\mathcal{T}_P\cdot\overrightarrow{PF_1}}{|\mathcal{T}_P|\cdot|\overrightarrow{PF_1}|} = -\frac{\mathcal{T}_P\cdot\overrightarrow{PF_2}}{|\mathcal{T}_P|\cdot|\overrightarrow{PF_2}|}, \text{ or equivalently}$$

$$\frac{1(-c-x)+m(0-y)}{\sqrt{(-c-x)^2+(0-y)^2}} = -\frac{1(c-x)+m(0-y)}{\sqrt{(c-x)^2+(0-y)^2}}.$$

Since $x = -t$ and $y = Y$, it is the easiest of high school algebra exercises to confirm that our claim is equivalent to the last display. Thus, our claim has now been proved. By the final comment two paragraphs ago, it follows from the earlier part of this proof that only the behavior of f on the interval $[0,a]$ was used in determining a formula for $f|_{[0,a]}(x)$. Therefore, in view of what we have noted concerning g and γ, it follows from the earlier part of this proof that if $-a < x < 0$, then

$$f(x) = g(-x) = (\frac{b}{a})\sqrt{a^2-(-x)^2} = (\frac{b}{a})\sqrt{a^2-x^2}.$$

This completes the proof that (4) implies the assertion in (1) in case $-a \leq x \leq 0$. The proof is complete. □

In [3, Remark 2.2 (b)], we felt it necessary to explain what it had meant in the proof of [3, Theorem 2.1] for a point to be “inside" or “outside" a parabola. Comments of that kind would be of use in Sections 4 and 5 (and they will be left to the reader there). Such references to the “inside" of a potentially elliptic curve have already occurred in the statements of Theorems 2.1 and 3.1 (and will recur in Remark 3.2 (b)). Fortunately, in “potentially elliptic" contexts such as those, there is no ambiguity about the meaning of “inside", thanks to the Jordan Curve Theorem, the point being that any ellipse (unlike any parabola or any hyperbola) is a simple closed curve.

To reduce the length of the statement of Theorem 3.1, some variants or sharpenings of Theorem 3.1 will be collected in the next result. The strongest of those sharpenings is in Remark 3.2 (d). Remark 3.2 also addresses some the themes from [3], insofar as they concern ellipses, such as approaches using "x as a function of y" and degenerate cases.

Remark 3.2. (a) We begin with a comment that is in the spirit of [3, Remark 3.2 (a)]. Although it may have been a distraction in Theorem 3.1 to point out the redundancy of its assumption that the tangent lines to Γ at both $(-a,0)$ and $(a,0)$ are vertical lines, we wish to do so here. We will provide full details for the case of $(a,0)$, leaving to the reader the similar details for the case of $(-a,0)$. This will be done by appealing to the above-mentioned definition of what it means for the graph of a function to have a vertical tangent line at a given point on the graph. Assume that a function $f:[-a,a]\to\mathbb{R}$ satisfies all the conditions in the third and fourth sentences of the statement of Theorem 3.1. Since f is assumed to be continuous, our task is to show that $\lim_{x\to a^-} f^{'}(x)=\pm\infty$ (in fact, this limit will be shown to be $-\infty$).

If we examine the secant lines whose limiting position (if it exists) would be that of the tangent line to Γ at $(a,0)$, the corresponding limit of the slopes of those secant lines is

$$\lim_{x\to a}\frac{f(x)-f(a)}{x-a}=\lim_{x\to a^-}\frac{f(x)-0}{x-a}=\lim_{x\to a^-}f^{'}(x),$$

where the last step was obtained by using the general form of L'Hôpital's Rule (as formulated in [12, Theorem 1]). Next, by tweaking the first part of the proof that (4) $\Rightarrow$ (1) in Theorem 3.1, we can use the formula for the derivative of $y=f(x)$ on the interval $(0,a)$ in that proof to reformulate our task as seeking a proof that

$$\lim_{x\to a^-}\frac{-x^2+c^2+y^2-\sqrt{(y^2+c^2+x^2)^2-4c^2x^2}}{2xy}=-\infty.$$

To that end, recall that $y=f(x)>0$ for all $x\in(0,a)$, f is continuous at $x=a$, $f(a)=0$, and $0<c<a$. Hence, working in the extended real number system ($\mathbb{R}\cup\{\infty,-\infty\}$) and using the appropriate limit theorems there, we get

$$\lim_{x\to a^-}\frac{-x^2+c^2+y^2-\sqrt{(y^2+c^2+x^2)^2-4c^2x^2}}{2xy}=$$

$$\frac{-a^2+c^2+0^+-\sqrt{(0^++c^2+a^2)^2-4c^2a^2}}{2a\cdot 0^+}=\frac{-a^2+c^2-\sqrt{(c^2-a^2)^2}}{0^+}=$$

$$=\frac{-a^2+c^2-|c^2-a^2|}{0^+}=\frac{-a^2+c^2-(a^2-c^2)}{0^+}=\frac{2(c^2-a^2)}{0^+}=-\infty,$$

as desired. This proof should dispel any lingering worries that the proof of Theorem 3.1 may have only characterized the "open top half" of the ellipse $x^2/a^2+y^2/b^2=1$, as we have just shown that the point $(a,0)$ [and one can similarly show that the point $(-a,0)$] is indeed part of the "top half" which was characterized in that proof (even if one had not assumed that the tangent lines to Γ at $(a,0)$ and $(-a,0)$ exist and are vertical).

(b) As promised above, we next give the "bottom half of the ellipse" analogue of the characterization result in Theorem 3.1. To avoid possibly unclear platitudinous general comments, we will next state this new result in full before proving it.

THEOREM. Let $0<c<a$ in $\mathbb{R}$, and put $b:=\sqrt{a^2-c^2}$. Working in a fixed Euclidean plane, consider the points $F_1(-c,0)$ and $F_2(c,0)$. Let $g:[-a,a]\to\mathbb{R}$ be a function, and let γ be the graph of g. Suppose that g is differentiable on $(-a,a)$, g is continuous at $x=-a$ and at $x=a$, $g(-a)=0=g(a)$, $g(t_1)<0$ for some $t_1\in(-a,0)$, $g(t_2)<0$ for some $t_2\in(0,a)$, and $g^{'}(x)\neq 0$ for all $x\in(-a,0)\cup(0,a)$. For each point Q

on γ, let $\mathcal{U} := \mathcal{U}_Q$ be a tangential vector to γ at Q. Then the following five conditions are equivalent:

(1) $g(x) = -(\frac{b}{a})\sqrt{a^2 - x^2}$ for all $x \in [-a, a]$;

(2) γ is the "bottom half" of the "east-west" ellipse with foci F_1 and F_2 and semi-major axis a (and necessarily with semi-minor axis b);

(3) For each point Q on γ, the angle between $\mathcal{U}_Q$ and $\overrightarrow{QF_1}$ is congruent to the angle between the opposite of $\mathcal{U}_Q$ and $\overrightarrow{QF_2}$;

(4) For each point Q on γ,

$$\frac{\mathcal{U}_Q \cdot \overrightarrow{QF_1}}{|\mathcal{U}_Q| \cdot |\overrightarrow{QF_1}|} = -\frac{\mathcal{U}_Q \cdot \overrightarrow{QF_2}}{|\mathcal{U}_Q| \cdot |\overrightarrow{QF_2}|};$$

(5) For each point Q on γ, the following reflection property holds: if $\overrightarrow{R}$ is a ray which is emitted from F_1 and meets γ at Q, then the "reflected" ray which results from that intersection stays "inside" γ (for a while), going on a line of action which passes through F_2;

(6) For each point Q on γ, the following reflection property holds: if $\overrightarrow{R}$ is a ray which is emitted from F_2 and meets γ at Q, then the "reflected" ray which results from that intersection stays "inside" γ (for a while), going on a line of action which passes through F_1.

Any interested readers are encouraged to confirm that the just-stated THEOREM can be proved by tweaking the first half of the proof of Theorem 3.1. However, we will, instead, prove that THEOREM by tweaking the second half of the proof of Theorem 3.1. Several later parts of the present remark will uncover some general principles uniting these methods of proof.

We begin the proof by defining a function $f : [-a, a] \to \mathbb{R}$ by $f(x) := -g(x)$ whenever $-a \leq x \leq a$. Let Γ be the graph of f. Then f is differentiable on $(-a, a)$, since $f^{'}(x) = -g^{'}(x)$ whenever $-a < x < a$. Moreover, f is continuous at $x = -a$ and at a. Also, by tweaking the reasoning in (a), one can show that Γ has vertical tangent lines at the points $(-a, 0)$ and $(a, 0)$. Therefore, as in the second paragraph of the proof of Theorem 3.1, one can show that each of the conditions (1)-(5) holds if the point Q is either $(-a, 0)$ or $(a, 0)$. Accordingly, for the rest of this proof, we can assume, without loss of generality, that Q is neither of these points. Furthermore, $f(-a) = 0 = f(a)$, $f(t_1) > 0$ for some $t_1 \in (-a, 0)$, $f(t_2) > 0$ for some $t_2 \in (0, a)$, and $f^{'}(x) \neq 0$ for all $x \in (-a, 0) \cup (0, a)$.

By reasoning as in the third paragraph of the proof of Theorem 3.1, it remains only to prove that (4) $\Rightarrow$ (1).

It will be convenient to write $Y = f(t)$, with $-a < t < a$. The point $Q(t, Y)$ on γ corresponds to the point $P(x, y)$ on Γ, where $x = t \in (-a, a)$ and $y = f(x) = -g(x) = -g(t) = -Y$. Let $m := f^{'}(x)$. Then $g^{'}(t) = -f^{'}(t) = -f^{'}(x) = -m$. For our present purposes, we can take $\mathcal{U} = \mathcal{U}_Q = \mathbf{i} + (-m)\mathbf{j} = \mathbf{i} - m\mathbf{j}$. Similarly, we can take the tangential vector to/of γ at P to be $\mathcal{T} := \mathcal{T}_P := \mathbf{i} + m\mathbf{j}$. We claim that f (or, if you wish, Γ) satisfies the analogue of condition (4) relative to the interval $[-a, a]$; that is, we claim that

$$\frac{\mathcal{T}_P \cdot \overrightarrow{PF_1}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_1}|} = -\frac{\mathcal{T}_P \cdot \overrightarrow{PF_2}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_2}|}, \text{ or equivalently}$$

$$\frac{1(-c - x) + m(0 - y)}{\sqrt{(-c - x)^2 + (0 - y)^2}} = -\frac{1(c - x) + m(0 - y)}{\sqrt{(c - x)^2 + (0 - y)^2}}.$$

On the other hand, because we are assuming (4), we have that

$$\frac{\mathcal{U}_Q \cdot \overrightarrow{QF_1}}{|\overrightarrow{QF_1}|} = -\frac{\mathcal{U}_Q \cdot \overrightarrow{QF_2}}{|\overrightarrow{QF_2}|}, \text{ that is.}$$

$$\frac{1(-c - t) + (-m)(0 - Y)}{\sqrt{(-c - t)^2 + (0 - Y)^2}} = -\frac{1(c - t) + (-m)(0 - Y)}{\sqrt{(c - t)^2 + (0 - Y)^2}}.$$

Since $t = x$ and $Y = -y$, it is the easiest of high school algebra exercises to confirm that our claim is equivalent to the last display. Thus, our claim has now been proved. Therefore, by the implication (4) $\Rightarrow$ (1) in Theorem 3.1,

$$f(x) = (\frac{b}{a})\sqrt{a^2 - x^2} \text{ for all } x \in [-a, a].$$

Hence, $g(x) = -f(x) = -(b/a)\sqrt{a^2 - x^2}$ for all $x \in [-a, a]$. This completes the proof of the above THEOREM. This completes part (b) of this remark.

(c) The proof in (b) ended because a certain pair of equations were equivalent. We came across *the same* pair of equations in the second part of the proof that (4) $\Rightarrow$ (1) in the proof of Theorem 3.1. Perhaps this situation seems puzzling, since this pair of equations arose from *different* changes in variables. Indeed, in (b), we used the substitutions $x = t$ and $y = -Y$, whereas in the proof of Theorem 3.1, we used the substitutions $x = -t$ and $y = Y$. It seems natural to ask how/why two *different* algebraic substitutions – which were designed to address two *different* analytic/geometric situations - translated the problems at hand into exactly the *same* pair of equations. The answer concerns a certain class of rigid motions of the Euclidean plane, namely, the reflections about a line that is either vertical or horizontal. Indeed, the change of coordinates/points $P(x,y) \mapsto Q(t,Y) := Q(x,-y)$ that was used in (b) describes reflection about the (horizontal) x-axis, whereas the change of coordinates/points $(x,y) \mapsto (t,Y) := (-x,y)$ that was used in the proof of Theorem 3.1 describes reflection about the (vertical) y-axis. As is the case with any rigid motion of the Euclidean plane, these kinds of reflections preserve distance and the measures of (undirected) angles. The behavior of the measures of certain angles formed between certain bound vectors is *exactly* the sort of thing that condition (4) of Theorem 3.1 dealt with. (The same can be said about condition (4) in [3, Theorem 3.1] (concerning parabolas) and a number of conditions in Sections 4 and 5 (concerning hyperbolas).) Note that the above-mentioned change $P \mapsto Q$ in (b) (resp., in the proof of Theorem 3.1) had the effect of fixing F_1 and F_2 (resp., had the effect of interchanging F_1 and F_2). Both of these changes also had the effect of sending any vertical line to a vertical line and also replacing a tangent line having slope m with a tangent line having slope $-m$. By simply examining a few possible combinations of situations, we see that both kinds of changes preserve and reflect the property that the angle between $\mathcal{T}$ and $\overrightarrow{PF_1}$ is (apart from orientation, that is, as an undirected angle) congruent to the angle between the opposite of $\mathcal{T}$ and $\overrightarrow{PF_2}$. Thus, a “conformal" point of view (dare I say “perspective") has served to completely explain a situation that had been termed “puzzling".

There is more to be gained by considering the above changes of variables (and some other related changes). In the proof of Theorem 3.1, we saw that the change $(x,y) \mapsto (t,Y) := (-x,y)$ can be used to convert a first-quadrant arc that seems elliptic into a second-quadrant arc that seems elliptic (and is part of the same ellipse). A moment's thought reveals that the process is reversible, giving a change of variables that converts a second-quadrant arc that seems elliptic into a first-quadrant arc that seems elliptic (and is part of the same ellipse). Another moment's thought shows that the same change of variables can be used to convert a fourth- (resp., third-) quadrant arc that seems elliptic into a third- (resp., fourth) quadrant arc that seems elliptic (and is part of the same ellipse). On the other hand, let us next consider the change $(x,y) \mapsto (t,Y) := (x,-y)$ that we saw in (b). The proof of (b) showed, in essence, that this change can be used to convert an upper- (resp., lower-) half-plane arc that seems elliptic into a lower- (resp., upper-) half-plane arc that seems elliptic (and is part of the same ellipse). A closer examination of that proof reveals that this change also converts first- (resp., fourth) quadrant seemingly elliptic arcs into fourth- (resp., first-) quadrant seemingly elliptic arcs, and also converts third- (resp., second) quadrant seemingly elliptic arcs into second- (resp., third-) quadrant seemingly elliptic arcs. (Of course, by “seems elliptic" or “seemingly elliptic", I am referring to graphs of functions with certain analytic properties – which were prescribed in (b) and in Theorem 3.1 – that are then proven to be arcs of an “east-west" ellipse whose center is at the origin). This compilation of results implies the following. If one has a function h_1 satisfying certain analytic properties whose graph is the intersection of some quadrant with an "east-west" ellipse $\mathcal{E}$ with center at the origin, then it is possible to use h_1 and specific changes of variables to define functions h_2, h_3 and h_4 having analogous analytic properties and also such that the graph of each h_k is the intersection of some quadrant with $\mathcal{E}$ in such a way that the union of the graphs of h_1, h_2, h_3 and h_4 is the ellipse $\mathcal{E}$. Moreover, the construction of these functions h_2, h_3 and h_4 uses only the specific changes of variables mentioned above (and h, of course). Note also that the method does not, *per se*, produce a Cartesian equation for $\mathcal{E}$.

All the success recorded in the last two paragraphs was based on beginning with a function whose domain was a substantial portion of the projection onto the x-axis of an inferred ellipse. That success may cause one to hope to do more with less. This raises the following more specific questions. If one has any nontrivial seemingly elliptic arc γ (as the graph of a function g whose domain is any nontrivial interval), can one show that γ is the arc of only one "east-west" ellipse $\mathcal{E}$ with center at the origin and prescribed potential foci $(\pm c, 0)$ and, if so, can one then use γ (that is, can one then use g) to obtain a Cartesian equation for $\mathcal{E}$? We will give the following partial answers to these questions in (d) (recalling that $0 < c < a$ are given and $b := \sqrt{a^2 - c^2}$): at least one such $\mathcal{E}$ exists; and the ellipse described by the Cartesian equation $x^2/a^2 + y^2/b^2 = 1$ is the only such $\mathcal{E}$ if (and only if) there exists a point Q on γ such that $d(Q, F_1) + d(Q, F_2) = 2a$.

(d) In the spirit of the reflection-theoretic result characterizing some parabolic arcs in [3, Remark 3.2 (e)], it is natural to ask if there is an analogous result for elliptic arcs. We next give a result that affirmatively answers this question (and leads to partial answers to the questions raised at the end of (c)). (My reading of [5, Theorem 1] is that Drucker asserted a somewhat similar result in regard to the reflection property that he considered for ellipses.) We will give details for some seemingly elliptic arcs in the first quadrant and later state a similar result for another quadrant.

Warning: The following proof of a reflection-theoretic characterization of a first-quadrant elliptic arc is going to use the fact (which will be proved later in Theorem 4.1) that hyperbolas satisfy a certain reflection-theoretic property (which is different from the reflection property of an ellipse that was established in Theorem 2.1). I have chosen to present the two sections on ellipses (resp., on hyperbolas) without any other section intervening between them for the following reason: the reflection property of an ellipse and the reflection properties of a hyperbola are rather well known, while characterization results for these types of figures (proven in Sections 3 and 5, respectively) are much less familiar.

As in Theorem 3.1, let $0 < c < a$ in $\mathbb{R}$, put $b := \sqrt{a^2 - c^2}$, and consider the points $F_1(-c, 0)$ and $F_2(c, 0)$. Next, consider real numbers $0 \le \alpha < \beta \le a$ and a function $f : [\alpha, \beta] \to \mathbb{R}$, and let Γ be the graph of f. Suppose that f is differentiable on (α, β), f is continuous at $x = \alpha$ and at $x = \beta$, $f(\alpha) = (b/a)\sqrt{a^2 - \alpha^2}$ and $f(\beta) = (b/a)\sqrt{a^2 - \beta^2}$, $f(t_1) > 0$ for some $t_1 \in (\alpha, \beta)$, and $f'(x) \neq 0$ for all $x \in (\alpha, \beta)$. Next, suppose that the data satisfy the following condition which is analogous to the reflection-theoretic property in Theorems 2.1 and 3.1: for each point P on Γ, if $\mathcal{T}$ denotes a tangential vector to Γ at P, then

$$\frac{\mathcal{T} \cdot \overrightarrow{PF_1}}{|\mathcal{T}| \cdot |\overrightarrow{PF_1}|} = -\frac{\mathcal{T} \cdot \overrightarrow{PF_2}}{|\mathcal{T}| \cdot |\overrightarrow{PF_2}|}.$$

By Theorem 2.1, *if* Γ were an arc of an east-west ellipse in standard position with semi-major axis a, semi-minor axis $b := \sqrt{a^2 - c^2}$, and foci at F_1 and F_2, then the just-displayed equation does hold for each P on Γ.

We next prove, conversely, that if the just-displayed equation holds for each P on Γ, then Γ is the arc of some east-west ellipse (in standard position) having foci F_1 and F_2. Moreover, if one momentarily ignores the assumed values of $f(\alpha)$ and $f(\beta)$ but instead imposes the (also necessary) condition that some point Q on γ satisfies $d(Q, F_1) + d(Q, F_2) = 2a$, then Γ is an arc of *some* east-west ellipse (in standard position) having foci F_1 and F_2. (While it may seem that the ellipse which we find below must have semi-major axis a and semi-minor axis b, the actual facts are otherwise. This is being left as a "temporary exercise" here for any interested readers. Its proof can be found by adapting the proof provided below in Theorem 5.6 (see also Remark 5.7 (d)) for the similar fact about hyperbolas. My reading of [5] is that Drucker did not specifically address this matter in regard to the reflection property that he considered for ellipses.) If one does resume the assumed value of $f(\alpha)$ (or that of $f(\beta)$) and ignores the possible existence of a point Q with the above-mentioned property, then we will get the desired sharpening of Theorem 3.1, as we will prove that under those conditions, Γ is indeed an arc of *the* east-west ellipse (in standard position) having foci F_1 and F_2, semi-major axis a and semi-minor axis b.

As in the proof of Theorem 3.1, we get that there exists a real number $E > 0$ such that

$$(2E-4c^2)x^2+2Ey^2=E^2-2Ec^2 \text{ if } \alpha < x < \beta \text{ and } y=f(x).$$

For the moment, let us *not yet* use the information that was assumed about the values of $f(\alpha)$ and $f(\beta)$. (We *will* eventually use those hypotheses.) Recalling from the proof of Theorem 3.1 that $E-2c^2 \neq 0$, we can put

$$A := (E^2-2Ec^2)/(2E-4c^2) \text{ and } B := (E^2-2Ec^2)/(2E).$$

Then it follows from the displayed equation in the last paragraph that

$$\frac{x^2}{A}+\frac{y^2}{B}=1 \text{ if } \alpha < x < \beta \text{ and } y=f(x).$$

Since f is continuous at both α and β, the just-displayed equation also holds if x is either α or β (and $y := f(x)$). Observe that $A = E/2 > 0$. Hence, there exists a positive real number a^* such that $A = (a^*)^2$. Note also that $A - B = c^2$. Thus, *if* $B > 0$, there would exist a positive real number b^* such that $B = (b^*)^2$, so that the above reasoning would give that

$$\frac{x^2}{(a^*)^2}+\frac{y^2}{(b^*)^2}=1 \text{ if } \alpha \le x \le \beta \text{ and } y=f(x),$$

and we would be well on our way to proving the converse. So, let us show that $B > 0$. As $B \neq 0$, it will suffice to obtain a contradiction from an assumption that $B < 0$.

Assume that $B < 0$. Then there exists a positive real number $b^\diamond$ such that $B = -(b^\diamond)^2$. Therefore, each point $P(x,y)$ on Γ satisfies

$$\frac{x^2}{(a^*)^2}-\frac{y^2}{(b^\diamond)^2}=1,$$

which (as will be recalled in Section 4) is the Cartesian equation of a certain hyperbola, say $\mathcal{H}$, having foci F_1 and F_2. As will be shown in Theorem 4.1, the reflection properties of this hyperbola can be described by the statement that each point $P(x,y)$ on $\mathcal{H}$ satisfies

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\mathcal{T}|\cdot|\overrightarrow{PF_1}|}=\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\mathcal{T}|\cdot|\overrightarrow{PF_2}|}.$$

However, each point $P(x,y)$ on Γ satisfies

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\mathcal{T}|\cdot|\overrightarrow{PF_1}|}=-\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\mathcal{T}|\cdot|\overrightarrow{PF_2}|}.$$

As each point $P(x,y)$ on Γ satisfies *both* of the last two displayed equations, it follows that

$$\mathcal{T}\cdot\overrightarrow{PF_1} = 0 = \mathcal{T}\cdot\overrightarrow{PF_2};$$

that is, $\mathcal{T}$ is perpendicular to both $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$. Hence, $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$ are parallel vectors. Choose $x \in (\alpha,\beta)$. This choice ensures that P is not on the x-axis. Let L (resp., M) denote the line passing through P and F_1 (resp., passing through P and F_2). Then L and M are parallel lines, and these parallel lines are distinct since L (resp., M) intersects the x-axis at only the point F_1 (resp., F_2). By a fundamental principle of Euclidean geometry, distinct parallel lines must have an empty intersection. As $P \in L \cap M$, we have the desired contradiction. This completes the proof that $B > 0$.

By the above work, there exist $a^* > 0$ and $b^* > 0$ such that $A = (a^*)^2$, $B = (b^*)^2$, $A - B = c^2$, and each point $P(x,y)$ on Γ satisfies

$$\frac{x^2}{(a^*)^2} + \frac{y^2}{(b^*)^2} = 1.$$

The last display is a Cartesian equation of an ellipse, say $\mathcal{F}$, with semi-major axis a^*, semi-minor axis b^*, and foci at the points $(-c^*,0)$ and $(c^*,0)$, where $(c^*)^2 = (a^*)^2 - (b^*)^2$. Hence, $(c^*)^2 = A - B = c^2$. Thus, $c^* = \sqrt{(c^*)^2} = \sqrt{c^2} = c$. Therefore, the foci of $\mathcal{F}$ are F_1 and F_2. This completes the proof of the characterization result that Γ is an arc of some east-west ellipse (in standard position) having foci F_1 and F_2.

Next, suppose that some point Q on Γ satisfies $d(Q,F_1) + d(Q,F_2) = 2a$ (equivalently, that some point Q on Γ is also on the ellipse having Cartesian equation $x^2/a^2 + y^2/b^2 = 1$). As Q is then on the ellipse $\mathcal{F}$, it follows from the definition of an ellipse that

$$d(P,F_1) + d(P,F_2) = 2a^*.$$

Thus $2a = 2a^*$, and so $a = a^*$. It remains only to show that $b = b^*$. This, in turn, follows since $b > 0$, $b^* > 0$ and

$$b^2 = a^2 - c^2 = (a^*)^2 - (c^*)^2 = (b^*)^2.$$

Next, let us resume the assumption that $f(\alpha) = (b/a)\sqrt{a^2 - \alpha^2}$ (we will not need to use the assumed value of $f(\beta)$) and let us ignore the possible existence of a point Q with the above-mentioned property. We will now obtain the desired sharpening of Theorem 3.1 (in which the domain $[0,a]$ is replaced by the domain $[\alpha,\beta]$), by showing that $a = a^*$. That will complete the proof of the enhanced characterization result, for once this equation has been obtained, it will follow that

$$(b^*)^2 = (a^*)^2 - c^2 = a^2 - c^2 = b^2,$$

whence $b^* = b$, whence Γ is an arc of the ellipse having foci F_1 and F_2, semi-major axis a and semi-minor axis b.

Recall that there exist positive real numbers a^* and b^* such that $(a^*)^2 - (b^*)^2 = c^2$ and

$$\frac{x^2}{(a^*)^2} + \frac{y^2}{(b^*)^2} = 1 \text{ if } \alpha \le x \le \beta \text{ and } y = f(x).$$

By substituting $x = \alpha$ and $y = (b/a)\sqrt{a^2 - \alpha^2}$ into the just-displayed equation and then multiplying through by $a^2(a^*)^2(b^*)^2$, we get

$$\alpha^2 a^2 (b^*)^2 + b^2(a^2 - \alpha^2)(a^*)^2 = a^2(a^*)^2(b^*)^2.$$

Next, by substituting $(b^*)^2 = (a^*)^2 - c^2$ and $b^2 = a^2 - c^2$ into the last-displayed equation, we get

$$\alpha^2 a^2[(a^*)^2 - c^2] + (a^2 - c^2)(a^2 - \alpha^2)(a^*)^2 = a^2(a^*)^2[(a^*)^2 - c^2].$$

Then by routine uses of the distributive property and additive cancellation of like terms, we get

$$a^4(a^*)^2 = a^2(a^*)^4$$

Finally, dividing through by (the nonzero number) $a^2(a^*)^2$ gives $a^2 = (a^*)^2$, whence $a = a^*$. The proof is complete.

As promised, we next state an analogue (for the fourth quadrant) of the above "first-quadrant" characterization result. (By definition, the fourth quadrant consists of all $(u,v) \in \mathbb{R}^2$ such that $u \geq 0$ and $v \leq 0$.) Let $0 < c < a$ in $\mathbb{R}$, put $b := \sqrt{a^2 - c^2}$, and consider the points $F_1(-c, 0)$ and $F_2(c, 0)$. Next, consider real numbers $0 \leq \alpha < \beta \leq a$ and a function $f : [\alpha, \beta] \to \mathbb{R}$, and let Γ be the graph of f. Suppose f is differentiable on (α, β) and continuous at $x = \alpha$, $f(\alpha) = -(b/a)\sqrt{a^2 - \alpha^2}$, $f(t_1) < 0$ for some $t_1 \in (\alpha, \beta)$, and $f^{'}(x) \neq 0$ for all $x \in (\alpha, \beta)$. Next, suppose that the data satisfy the following condition which is analogous to the reflection-theoretic property in Theorems 2.1 and 3.1: for each point P on Γ, if $\mathcal{T}$ denotes a tangential vector to Γ at P, then

$$\frac{\mathcal{T} \cdot \overrightarrow{PF_1}}{|\mathcal{T}| \cdot |\overrightarrow{PF_1}|} = -\frac{\mathcal{T} \cdot \overrightarrow{PF_2}}{|\mathcal{T}| \cdot |\overrightarrow{PF_2}|}.$$

Then Γ is a subset of the fourth quadrant and Γ is an arc of the ellipse (in standard position) having foci F_1 and F_2, semi-major axis a and (necessarily) semi-minor axis b.

Next, in the interest of brevity, let us illustrate a general method for characterizing certain (potentially elliptic) arcs that have nontrivial intersections with more than one quadrant. Specifically, we address such an arc Γ contained in the union of the first and fourth quadrants. Because of the meaning of "arc", the point $(a, 0)$ is necessarily on Γ. Let Γ_1 (resp., Γ_4) denote the intersection of Γ with the first (resp., with the fourth) quadrant. Suppose that Γ_1 (resp., Γ_4) is the graph of a function satisfying all the hypotheses stated in the third paragraph of this part (d) (resp., stated in the preceding paragraph). Then, since $(a, 0)$ is on both Γ_1 and Γ_4, it follows (from the method of the proof that was completed two paragraphs ago) that Γ_1 and Γ_4 can be described respectively (over suitable interval domains with left-hand endpoints being some nonnegative real numbers and the same right-hand endpoint, a) by $y = (b/a)\sqrt{a^2 - x^2}$ and $y = -(b/a)\sqrt{a^2 - x^2}$. Consequently, any graph Γ satisfying the above hypotheses must be an arc of the ellipse given by $x^2/a^2 + y^2/b^2 = 1$ and Γ is not a subset of any other east-west ellipse whose center is the origin.

(e) The proofs of [3, Remarks 2.2 (c) and 3.2 (d)] used the "x as a function of y" point of view and its attendant methods. It seemed natural to include such proofs in [3] since every parabola, after suitable rigid rotation and/or translations of coordinate axes, is the graph of an equation of the form $x = y^2/(4a)$ for some nonzero $a \in \mathbb{R}$. I consider it to be less important to feature such methods in developing material about ellipses here. Some of this attitude derives from the fact that no amount of rotation or translation of coordinate axes can produce an ellipse which is the graph of an equation $y = f(x)$ or an equation $x = g(y)$, since every ellipse fails both the Vertical Line Test and the Horizontal Line Test. While the same is true of hyperbolas, I will use the "x as a function of y" point of view in proving Theorem 5.4 about hyperbolas. In that characterization result, that is the natural point of view to use in dealing with a branch of an east-west hyperbola. To be sure, one can point to naturally occurring subsets of ellipses that are graphs of equations of the form $x = g(y)$. For example, consider the left- (or the right-) hand half of an east-west ellipse (with, for simplicity, center at the origin). However, that particular subset is not of any special importance in regard to the reflection properties of an ellipse. By way of contrast, each branch of a hyperbola plays a critical role in the description of the reflection properties of a hyperbola: see, for instance, Theorem 4.1 below. Instructors who wish to use the "x as a function of y" approach to obtain alternative proofs of some of the results in this paper (for instance, a variant of Theorem 3.1) are invited to try to do so, perhaps consulting some of the above-mentioned material from [3] for any appropriate background needs that may arise for them or their students.

(f) In the spirit of [3, Remark 3.2 (g)], we next consider whether any degenerate cases of ellipses can be discovered by a careful reading of the proof of Theorem 3.1. That proof depended heavily on a differential equation which was derived in the fourth paragraph of that proof. That equation expressed the derivative of y $(= f(x))$ with respect to x as a fraction, say N/D, where $D = 2xy$. That kind of fraction was not objectionable since the context for that part of the proof of Theorem 3.1 involved $0 < x < a$ and the hypotheses of Theorem 3.1 ensured that $x \neq 0$ implies $y \neq 0$. However, the question of characterizing the curves satisfying the reflection properties of an ellipse can be cast more generally than in the hypotheses of Theorem 3.1. Indeed, since those reflection properties are perfectly encoded by condition (4) in the statement of Theorem 3.1, a more general attempt to characterize the functions f with domain a subset of $[0,a]$ and with graph satisfying the reflection properties of an ellipse, given $0 < c < a \in \mathbb{R}$ and points $F_1(-c,0)$ and $F_2(c,0)$, would ask the following: which differentiable functions f, having domain a subset of $[0,a]$, satisfy the above-mentioned condition (4)?

In order that (4) be meaningful, it must be the case that the vectors $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$ are each nonzero; that is, that neither F_1 nor F_2 is on the graph of f. Our search here for degenerate cases of ellipses that were not found in Theorem 3.1 will focus on such functions f for which the above-mentioned differential equation is meaningless because $D = 0$. As f is differentiable, it seems reasonable to assume that the domain of f is a union of (possibly denumerably many) open subintervals of $[0,a]$, together with possibly some left-hand endpoints and/or some right-hand endpoints. It would also seem reasonable to assume that f is continuous at any such endpoint. So, we will focus on certain values of x such that $0 < x \leq a$. As $D = 2xy$ and we are requiring $D = 0$ with a focus on certain x such that $0 < x \leq a$, it must be the case that $y = 0$, that is, $f(x) = 0$. Recall that the proof of Theorem 3.1 derived the above-mentioned differential equation from the following application of condition (4):

$$\frac{c + x + y\,\frac{dy}{dx}}{\sqrt{(c+x)^2 + y^2}} = \frac{c - x - y\,\frac{dy}{dx}}{\sqrt{(c-x)^2 + y^2}}.$$

Let us cast our nets more widely. Working in conjunction with the conditions $0 \neq x \in \mathbb{R}$ and $y = 0$, we see that the just-displayed equation is equivalent to the following algebraic equation:

$$\frac{c + x}{\sqrt{(c+x)^2}} = \frac{c - x}{\sqrt{(c-x)^2}}.$$

A straightforward case analysis shows that the solution set of the just-displayed equation, under the just-stated conditions, consists of the points (x,y) such that $-c < x < c$ and $y = 0$. It follows that if we focus on the universe $[0,a]$ for values of x, we can construct infinitely many degenerate cases of ellipses. Indeed, each of the following graphs Γ is of that kind: take any (possibly denumerable) nonempty set $\{I_j\}$ of open intervals I_j contained in $[0,c)$; for each j, let I_j^* result from I_j by possibly appending one or both endpoints of I_j to I_j, but do not append the element c to any I_j; let $\mathcal{D} := \cup_j I_j^*$; let $\mathcal{D}^*$ result from appending 0 to $\mathcal{D}$ if there exists j such that 0 is the left-hand endpoint of I_j; and then take $\Gamma := \{(x,y) \in \mathbb{R}^2 \mid x \in \mathcal{D}^* \text{ and } y = 0\}$.

Notice that a degenerate ellipse Γ which is constructed in the above way need not be a connected topological space – indeed, it may have infinitely many connected components. In particular, while it is commonplace for the literature to refer to some degenerate ellipses as being “piecewise-linear", observe that some of the examples Γ that we have just constructed have stretched the meaning of that term because they have infinitely many “pieces." Note also that the largest (in the obvious sense) interval that was constructed above as being a degenerate ellipse is $[0,c)$.

Having "cast our nets more widely", the above work naturally draws our attention to certain values of x such that $-c < x < 0$. One sees easily that for a nontrivial line segment γ containing the point $P(x, 0)$ for such a value of x, P satisfies the above-mentioned condition (4). Indeed, the tangential vector $\mathcal{T}$ of γ at P can be taken as $\pm\mathbf{i}$, and $\mathcal{T}$ is parallel to both $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$, while having the same direction as one of these vectors and the opposite direction of the other vector. (Of course, the same comment could have been made about the points $(x, 0)$ that arose as elements of the graphs Γ that were constructed two paragraphs ago.) Accordingly, we conclude that one can produce even more degenerate ellipses (although none that are qualitatively new in any mathematically important way) by revising the above directions for constructing degenerate ellipses Γ as follows: change the requirement "I_j contained in $[0, c)$" to "I_j contained in $(-c, c)$"; change "but do not append the element c to any I_j" to "but do not append either of the elements $-c$, c to any I_j; and change the definition of $\mathcal{D}^*$ to $\mathcal{D}^* := \mathcal{D}$.

(g) Following up on (e) and (f), let us consider any degenerate cases of ellipses that may be discovered from an examination of a proof of a variant of Theorem 3.1 that has used methods related to the "x as a function of y" point of view. It seems clear that those new degenerate cases would consist of analogues of the degenerate cases which were identified in (f). In other words, each of those new degenerate cases Δ of ellipses would be built by starting with a a union of (possibly denumerably many) open subintervals of the y-axis, each of which is a subset of the line segment going from the point $(0, -c)$ to the point $(0, c)$, with the precise rules for then building any such Δ (starting with such a union of open subintervals of the y-axis) being the obvious analogues of the corresponding rules that were given above for building the counterpart degenerate ellipses Γ. However, in my opinion, these Δ should *not* be considered as being "new" degenerate cases, for the following reason. The considerations in (f) (resp., here in (g)) have looked for degenerate cases of east-west (resp., north-south) ellipses that have center at the origin and a horizontal (resp., vertical) major axis. But *any* ellipse can be viewed that way after suitable rigid rotation and/or translation of coordinate axes. Such changes of coordinate axes do not change whether a geometric figure is an ellipse (or whether it should be considered as a degenerate ellipse in regard to satisfying certain reflection properties) because such changes of coordinate axes do not affect distance or the measure of (undirected) angles between bound vectors. Consequently, I conclude, for *any* ellipse $\mathcal{E}$, with major (resp., minor) axis falling along a line L (resp., M) in the Euclidean plane, that one can build a family of degenerate ellipses Γ^* (resp., Δ^*) that is naturally associated to $\mathcal{E}$ by taking the following steps: rigidly rotate and/or translate coordinate axes so that L is horizontal (resp., vertical) and M is vertical (resp., horizontal) in regard to the new coordinate axes [then the ellipse $\mathcal{E}$ is east-west (resp., north-south) in regard to the new coordinate axes], intersecting at the "new" origin; proceed to use the parameters a, b, c of $\mathcal{E}$ as in the final paragraph of (f) (resp., as earlier in this paragraph) to build a degenerate ellipse Γ (resp., Δ); and then perform (in reverse order) the sequence of the *inverse* operations corresponding to the above-mentioned rigid rotations and/or translations of coordinate axes. Notice that $\mathcal{E}$ has been carried back to its original position relative to the original coordinate axes. By definition, Γ^* (resp., Δ^*) is the geometric figure to which Γ (resp., Δ) has been carried. While Γ^* (resp., Δ^*) is a subset of the line L (resp, M), I would repeat the sentiment from the final sentence of (f) that these changes of coordinate axes have not produced anything that is "qualitatively new in any mathematically important way."

(h) The second paragraph of the Introduction of [3] mentioned several ways that conic sections can be introduced and viewed in the before-calculus curricula for mathematics and science. As we wrote there, "some of those [ways] are algebraic, some are geometric, and some are related to scientific applications." We went on to recall that the principal such "way" involving (real) analytic geometry is the following: "conics (along with their degenerate cases) are the only possible graphs (in Euclidean plane analytic geometry) of equations of the form $f(x, y) = 0$, where f is a second-degree polynomial expression in x and y". Among the geometric "ways" noted there, we recalled that "conics (along with their degenerate cases) are the only possible intersections (in three-dimensional Euclidean geometry) of a plane with a double-napped right circular cone"; and that "relative to a given point F that is not on a given line L, each of the three basic types of conics is characterized as the set $\mathcal{S}$ of points P such that the associated "eccentricity" e (that is, the ratio of the distance between P and F to the distance between P and L) has a specific constant value, with $e = 1$ (resp., $e < 1$; resp., $e > 1$) corresponding to $\mathcal{S}$ being a parabola (resp., an ellipse; resp., a hyperbola) with F being a "focus" of $\mathcal{S}$ and L being the corresponding "directrix" of $\mathcal{S}$". With respect to "scientific applications", we also noted that "each of the three basic types of conics has a reflection property with a number of physical applications." In the next-to-last paragraph of the Introduction of [3], we mentioned a couple of physical applications of the reflection properties of a parabola; in Section 2, we noted a couple of the physical application of the refection properties of an ellipse; and in Section 4, we will mention some applications of the reflection properties of a hyperbola.

We have noted elsewhere a couple of other algebrogeometric ways to use conic sections in the before-calculus curriculum. In [1], we used/introduced the intuitive idea of a "limit" to examine some algebraic ways to help a student to understand how certain planar sets are degenerate cases of one of the three standard conic sections. For instance, the "limit" of the equation $2x^2 + 3y^2 = d$ of an ellipse, under the limit process where the parameter $d \to 0^+$, is the equation $2x^2 + 3y^2 = 0$, whose (real) graph in the plane is the singleton set consisting of the origin. A three-dimensional analogue of this example can be found in [10, Exercise 5, pages 121-122], where J. M. H. Olmsted uses the limiting forms of Cartesian equations to explain how that "a [real quadric] cone can be thought of as a degenerate hyperboloid of either one or two sheets". In that spirit, I would suggest that if a, b and k are positive real numbers, then the cone which is the graph of the equation $x^2/a^2 + y^2/b^2 = z^2/k^2$ degenerates, when one uses the "limiting" process where the parameter $k \to 0$, to a line in $\mathbb{R}^3$ (given by the equation $z = 0$), but one could also conclude that this cone has degenerated (under this limit process) to the singleton set consisting of the origin in $\mathbb{R}^3$. For a more "geometric" way to "see" that degenerate set as the result of a "limiting" process, I would suggest the following: consider the "limiting position" of the ellipses which result by intersecting the cone $2x^2 + 3y^2 = z^2$ with the plane $z = k$ (for some nonzero parameter $k \in \mathbb{R}$).

In regard to the "eccentric" view of the main kinds of conic sections in terms of foci and directrices, let me mention another paper. In [2], I used the well known viewpoint that circles are degenerate forms of ellipses (recall that I considered the viewpoint that circles *are* ellipses in Remark 2.2 (b)) to motivate the introduction of points at infinity, as a first step toward introducing (real) projective geometry. (I still believe that a firm understanding of classical projective geometry is important for any student who intends to work in algebraic geometry in the form in which that subject is practiced today.) Let me give another reason to be interested in points at infinity. Some readers will surely be interested in comparing [3] and the present paper with the papers of Drucker ([5], [6]). In the latter papers, Drucker attempts to treat the reflection properties of the three main kinds of conic sections in a uniform way. That seems to necessitate Drucker's viewpoint that every conic has two foci. To handle parabolas, Drucker suggests that one should look at the limit of equations of more complicated conics and/or consider the "other" focus of a parabola to be a point at infinity which is subjected to additional restrictions. The reader is invited to assess the clarity, rigor and precision of Drucker's explanations/arguments that Drucker himself describes as being sketches of proofs.

4 The reflection properties of a hyperbola

We continue to work inside a fixed Euclidean plane. A hyperbola (in this plane) is determined by two distinct points, called foci and denoted by F_1 and F_2, together with two distinct positive real numbers $a < c$, such that the distance from F_1 to F_2 is $2c$. As in Sections 2-3, the distance from a point P to F_1 (resp., to F_2) will be denoted by $d_1 := d_1(P)$ (resp., by $d_2 := d_2(P)$). By definition, the *hyperbola* determined by F_1, F_2, a and c (with $0 < a < c$ and the distance from F_1 to F_2 being $2c$) is the set of points P such that $|d_1(P) - d_2(P)| = 2a$. The graph of a hyperbola consists of two disconnected subsets which are called the *branches* of the hyperbola, with one of the branches being the graph of the equation $d_1(P) - d_2(P) = 2a$ and the other branch being the graph of $d_2(P) - d_1(P) = 2a$. The graph of a typical hyperbola can be inferred from Figures 2-7.

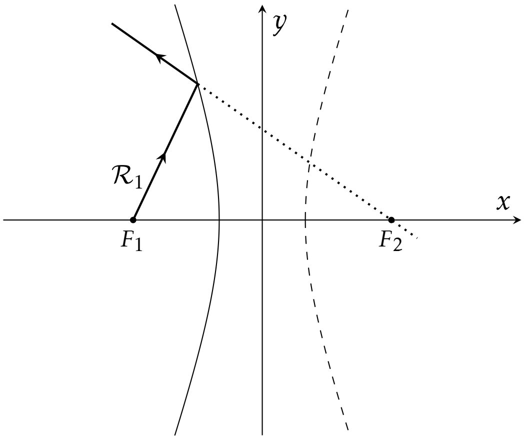


Figure 2

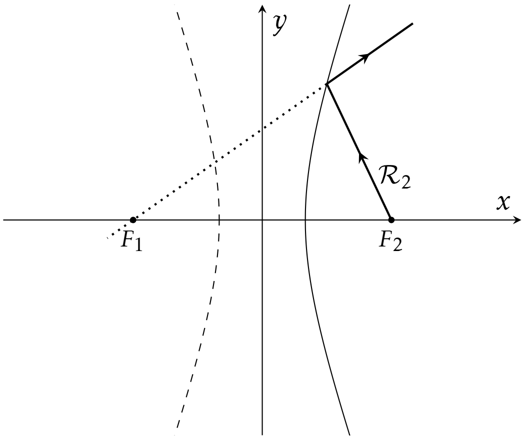


Figure 3

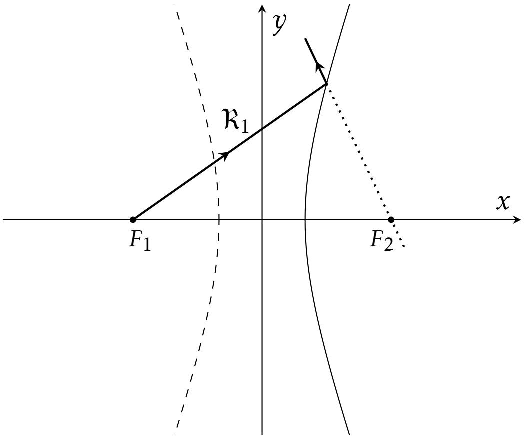


Figure 4

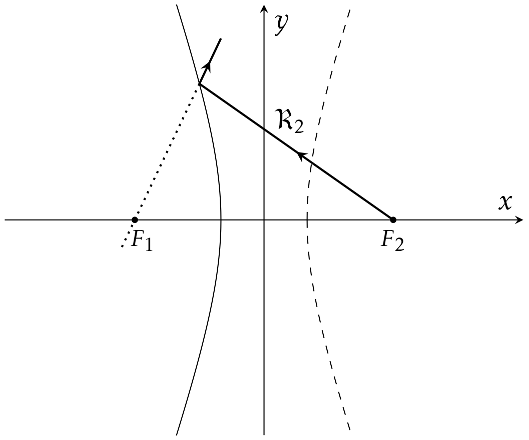


Figure 5

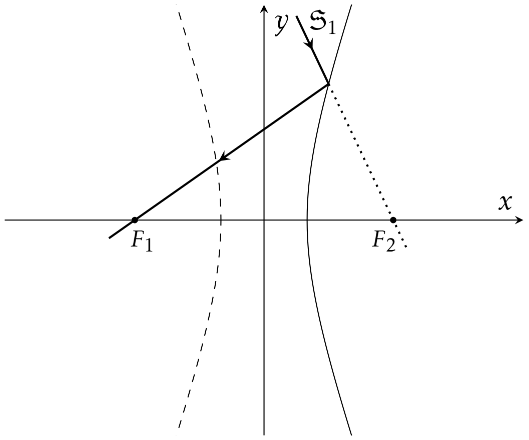


Figure 6

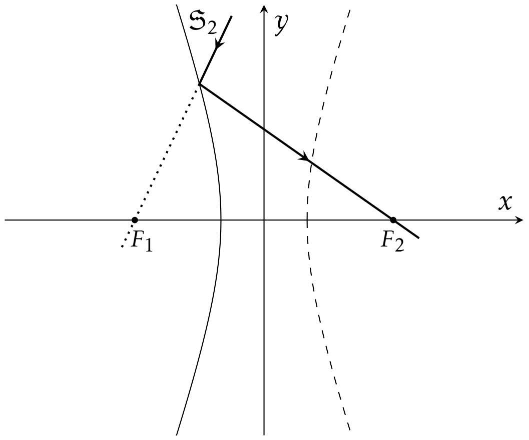


Figure 7

The branches of the hyperbola in Figures 2-7 are shown as either solid or dotted/dashed curves to facilitate the discussion (later in this section) of the reflection properties of a hyperbola. The various line segments or rays in Figures 2-7 are not part of the hyperbola *per se*, but they are related to the reflection properties of a hyperbola. If one were not considering those reflection properties, it would, of course, be appropriate to have *both* branches of a hyperbola presented as solid curves in *any* graph of the hyperbola.

In connection with the hyperbola determined by F_1, F_2, a and c, the following usage is conventional: the line passing through F_1 and F_2 intersects the hyperbola at two points which are called the *vertices* of the hyperbola; the line segment connecting the vertices is called the *transverse axis* of the hyperbola; it turns out that the length of the transverse axis is $2a$; the midpoint of the transverse axis [which turns out also to be the midpoint of the line segment connecting F_1 and F_2] is called the *center* of the hyperbola (a well chosen name since any hyperbola is symmetric about its center); a (or either of the line segments of length a from the center to a vertex) is called the (a) *semitransverse axis* of the hyperbola; the positive real number $b := \sqrt{c^2 - a^2}$ is called the *semi-conjugate axis* of the hyperbola; and the line segment that is perpendicular to the transverse axis and is bisected by the center is called the *conjugate axis* of the hyperbola (a well chosen name since the length of the conjugate axis is $2b$).

Beginners often have difficulty in remembering the relationship between the parameters a, b and c of a hyperbola, possibly because there are similarly denoted parameters of an ellipse which satisfy a different quadratic relationship. To help remember the quadratic relationship, $c^2 = a^2 + b^2$, that pertains to a hyperbola, it may help to remember that *in a hyperbola, there is no fixed inequality or equality relating* a and b. In other words, in a hyperbola, each of the following three conditions is possible: $a < b$, $a > b$, $a = b$. A hyperbola for which $a = b$ is called a *rectangular hyperbola* (also known as an *equilateral hyperbola*). A hyperbola is rectangular if and only if its asymptotes are perpendicular (to one another). We will not have occasion to mention the asymptotes of a hyperbola after this sentence, because asymptotes seem to play no important role in the study of the reflection properties of hyperbolas.

Every hyperbola has a number of interrelated reflection properties. Just as was the case for parabolas in [3] and for ellipses in Section 2, the reflection properties of a hyperbola can be scientifically justified by using the Principle of Reflection asserting that “the angle of incidence is congruent to the angle of reflection." Recall that physicists measure the just-mentioned angles “from the normal" (with “normal" being perpendicular to “tangential"), while most mathematicians prefer to use tangent lines rather than normal lines. As in Section 2, I suggest that some readers may wish to apply the “Principle of Reflection" to a few “random/typical" positions for a point P on a branch of the hyperbola in Figures 2-7 and then convert his/her conclusions about normal lines to conclusions about tangent lines. I trust that readers will agree with me that the three reflection properties of a hyperbola, with foci F_1 and F_2, can be stated as in the following paragraph.

Let $\{i, j\} = \{1, 2\}$. Let $\mathcal{B}_i$ (resp., $\mathcal{B}_j$) be the branch of the hyperbola in question which is associated with the focus F_i (resp., F_j). Figures 2 and 3 depict the following reflection property: in the absence of $\mathcal{B}_j$, if a ray $\mathcal{R}_i$ is emitted from F_i and intersects $\mathcal{B}_i$, then $\mathcal{R}_i$ reflects off $\mathcal{B}_i$ and as a result of that reflection, the redirected ray stays “inside" $\mathcal{B}_i$ and moves along a line which appears to have originated from F_j (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass thorough F_j). On the other hand, Figures 4 and 5 depict the following reflection property: in the absence of $\mathcal{B}_i$, if a ray $\mathfrak{R}_i$ is emitted from F_i and intersects $\mathcal{B}_j$, then $\mathfrak{R}_i$ reflects off $\mathcal{B}_j$ and as a result of that reflection, the redirected ray stays “outside" $\mathcal{B}_j$ and moves along a line which appears to have originated from F_j. Lastly, Figures 6 and 7 depict the following reflection property: in the absence of $\mathcal{B}_i$, if a ray $\mathfrak{S}_i$ approaches $\mathcal{B}_j$ from “outside" $\mathcal{B}_j$ along a line of action that passes through F_j then, as a result of intersecting $\mathcal{B}_j$, the ray is reflected/diverted along a new line of action which passes through F_i.

The most prominent scientific applications of the reflection properties of a hyperbola have to do with the “$d_1 - d_2 = \pm 2a$" definition of the branches of a hyperbola. These applications use a kind of “triangulation" to locate a point as an intersection of certain branches of various hyperbolas (with the parameters for those hyperbolas and Cartesian equations for their branches being calculated by using observed data). One of these applications, LORAN C, has both civilian uses and military uses, which include locating the source of an explosion or the epicentre of an earthquake. Some homework problems in this vein can be found in [4, Exercises 59-60, page 476; and Exercise 58, page 490]. More fanciful applications of the reflection properties of a hyperbola (mostly of a "war games" nature) can be found in [4, Exercises 57-58, page 476].

The above statements of the reflection properties of a hyperbola have to do with tangent lines (to a hyperbola) and the radian measures of angles formed at the intersection of various lines. As noted in Section 2, a rigid rotation and/or translation of the coordinate axes does not affect whether a line is tangent to a curve at a given point, and it also does not affect the radian measure of an angle formed at the intersection point of two lines. Accordingly, we can assume, without loss of generality, that the coordinate axes have been suitably rotated and/or translated so that the hyperbola can be viewed in "standard position," that is, so that the hyperbola has an especially tractable Cartesian equation (relative to the new coordinates axes). It is well known that when one rotates and/or translates the coordinate axes so that the transverse axis of a given hyperbola is horizontal and the center of that hyperbola is the origin, then the hyperbola has the following especially tractable Cartesian equation:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

Any hyperbola whose transverse axis is horizontal is said to be an "east-west" hyperbola (also known as a *horizontal hyperbola*). The hyperbola in Figures 2-7 is of "east-west" type, with the additional feature that its center is the origin, and hence it does have a Cartesian equation of the just-displayed type. The left-hand (resp., right-hand) branch of this hyperbola is the graph of the equation

$$x = -\frac{a}{b}\sqrt{y^2 + b^2} \text{ (resp., } x = \frac{a}{b}\sqrt{y^2 + b^2}\text{)}.$$

We show (in this section and in Section 5) that each hyperbola satisfies the three reflection properties mentioned above and, in the presence of certain additional hypotheses, can be characterized by each of those properties. Moreover, we will also show that the same kind of conclusions hold for each branch of a hyperbola and for some associated arcs of a hyperbola.

We pause to address the derivation of the Cartesian equation, $x^2/a^2 - y^2/b^2 = 1$, for an east-west hyperbola having its center at the origin. While many textbooks introducing hyperbolas include a complete proof that the condition "$|d_1(P) - d_2(P)| = 2a$" for this hyperbola implies the asserted equation (cf. [4, page 465]), I have not seen any textbook that gives all the detail needed for a proof of the converse. To be fair, both [8] and [9] mention that their hints for a proof of the corresponding fact about ellipses can be adjusted to give a proof of that converse for hyperbolas. In the interest of brevity, I leave it to the reader to confirm at least one of those statements or to devise another proof of the converse (using, I would suggest, the hyperbolic trigonometric functions cosh and sinh).

Recall that the notion of a tangential vector was useful in [3] and in Sections 2 and 3 as well. It will also be useful in studying hyperbolas and their reflection properties. Indeed, according to the "Principle of Reflection," *each* of the above three reflection-theoretic properties of a hyperbola (whose statements we agreed about four paragraphs ago) is equivalent to the following mathematical formulation. Let P be the point at which the initial ray intersects a branch (which is not "absent") of the hyperbola, and let $\mathcal{T}$ be a tangential vector to that branch at P; then the angle between $\mathcal{T}$ and (the bound vector) $\overrightarrow{PF_1}$ is congruent to the angle between $\mathcal{T}$ and (the bound vector) $\overrightarrow{PF_2}$.

We next prove the three reflection properties of a hyperbola. The statement of Theorem 4.1 can be used to interpret the meaning of the segments/rays in Figures 2-7. Indeed, the statement of part (a) (resp., part (b); resp., part (c)) of Theorem 4.1 summarizes what is depicted in Figures 2 and 3 (resp., Figures 4 and 5; resp., Figures 6 and 7).

Theorem 4.1. Let $\mathcal{H}$ be a hyperbola with foci F_2 and F_2. Let the branches of $\mathcal{H}$ be labeled $\mathcal{B}_1$ and $\mathcal{B}_2$ in such a way that F_1 (resp., F_2) is contained "inside" $\mathcal{B}_1$ (resp., $\mathcal{B}_2$) in the obvious intuitive sense. Let $\{i,j\} = \{1,2\}$. Then:

(a) In the absence of $\mathcal{B}_j$, if a ray $\mathcal{R}_i$ is emitted from F_i and intersects $\mathcal{B}_i$, then $\mathcal{R}_i$ reflects off $\mathcal{B}_i$ and as a result of that reflection, the redirected ray stays "inside" $\mathcal{B}_i$ and moves along a line which appears to have originated from F_j (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass thorough F_j).

(b) In the absence of $\mathcal{B}_i$, if a ray $\Re_i$ is emitted from F_i and intersects $\mathcal{B}_j$, then $\Re_i$ reflects off $\mathcal{B}_j$ and as a result of that reflection, the redirected ray stays "outside" $\mathcal{B}_j$ and moves along a line which appears to have originated from F_j.

(c) In the absence of $\mathcal{B}_i$, if a ray $\mathfrak{S}_i$ approaches $\mathcal{B}_j$ from "outside" $\mathcal{B}_j$ along a line of action that passes through F_j then, as a result of intersecting $\mathcal{B}_j$, the ray is reflected/diverted along a new line of action which passes through F_i.

Proof. In the next-to-last paragraph before the statement of this result, we gave an equivalent formulation of the reflection properties of a hyperbola in terms of certain behavior of a "typical" tangential vector to the given hyperbola. That formulation implicitly assumed that for each point P on $\mathcal{H}$, the bound vectors $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$ are each nonzero, equivalently, that neither F_1 nor F_2 is on $\mathcal{H}$. We will prove this fact for F_1, leaving the details of the similar proof about F_2 to the reader. To prove that F_1 is not on $\mathcal{H}$, we need only use the hypothesis that $a < c$ to see that $|d_1(F_1) - d_2(F_1)| = |0 - 2c| = 2c \neq 2a$.

Four paragraphs before the statement of this result, we explained why we could assume during this proof that, without loss of generality, the coordinate axes have already been suitably rotated and/or translated so that $\mathcal{H}$ is in standard position, that is, that $\mathcal{H}$ is an east-west hyperbola whose foci lie on the x-axis and whose center is at the origin. Necessarily, $\mathcal{H}$ has a Cartesian equation of the form

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

where a and b are positive real numbers and $c^2 = a^2 + b^2$. Note that the vertices of $\mathcal{H}$ are the points $P_1(a,0)$ and $P_2(-a,0)$.

It will be convenient to directly verify that the reflection properties hold when the point of impact P of the incoming ray is a point at which the tangent line to $\mathcal{H}$ is vertical. The points P with this property are the vertices of $\mathcal{H}$. We will verify that the "vertical tangent line" condition holds at P_1, leaving it to the reader to provide the similar verification at P_2. Let us use the fact that P_1 is on the right-hand side of the upper half of $\mathcal{H}$, which is the graph of the function h given by

$$y = h(x) = (\frac{b}{a})\sqrt{x^2 - a^2},$$

with the point P_1 having coordinates $(a,0) = (a, h(a))$. It is evident that h is continuous at a. As $h^{'}(x) = bx/(a\sqrt{x^2-a^2})$ if $|x| > a$,

$$\lim_{x\to a} h^{'}(x) = \lim_{x\to a^+} \frac{bx}{a\sqrt{x^2-a^2}} = ba/0^+ = \infty,$$

thus proving that the tangent line to $\mathcal{H}$ at P_1 is indeed vertical. The verification of the reflection properties of $\mathcal{H}$ at P_1 and P_2 can now be done essentially as the corresponding verification for the ellipse $\mathcal{E}$ in the proof of Theorem 2.1. For the sake of completeness, we provide those details next. For $k \in \{1,2\}$, the work earlier in this paragraph lets us take $\mathcal{T} := \mathcal{T}_k$, the tangential vector to $\mathcal{H}$ at P_k, to be the bound vector that has initial point P_k and is equivalent to $\mathbf{j}$. Hence, the angle between $\mathcal{T}_k$ and $\overrightarrow{P_kF_1}$ is a right angle, and the angle between $\mathcal{T}_k$ and $\overrightarrow{P_kF_2}$ is also a right angle. As it is a fundamental principle of Euclidean geometry that any two right angles are congruent, this completes a direct proof that the assertion of the reflection properties holds at P_1 and P_2. (Interested readers are invited to fashion an alternate direct proof of this assertion that uses the Principle of Reflection and measures angles "from the normal.")

By combining the above discussion with the review of vectorial material (involving dot products and the inverse cosine function) two paragraphs before the statement of Theorem 4.1 (see also six paragraphs before that statement, as well as the vectorial background that was given five paragraphs before the statement of Theorem 2.1), we can see that (a), (b) and (c) are each equivalent to the following condition: if P is a point on a hyperbola $\mathcal{H}$ and $\mathcal{T}$ is a tangential vector (at P) to a function whose graph includes a nontrivial arc of $\mathcal{H}$ containing P, then

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\mathcal{T}|\cdot|\overrightarrow{PF_1}|}=\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\mathcal{T}|\cdot|\overrightarrow{PF_2}|},\text{ or, equivalently,}$$

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|}=\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

We will proceed to prove this condition. Given the last result in the preceding paragraph, we can assume henceforth, without loss of generality, that P has coordinates (x,y) such that $y\neq 0$. It will be convenient to denote the left- and right-hand sides of the desired equation, respectively, by

$$\mathcal{L}:=\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|}\text{ and }\mathcal{R}:=\frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

Note that this is *not* (resp., is) the same definition of the notation $\mathcal{R}$ (resp., of the notation $\mathcal{L}$) that was used in the proof of Theorem 2.1. (That difference should not come as a surprise. After all, since ellipses and hyperbolas *do* have different reflection properties, it is to be expected that the equations which encode those different properties are different equations.)

It will be helpful to have a formula for the slope, say m, of the tangent line to $\mathcal{H}$ at P. (Recall that P is not a vertex of $\mathcal{H}$ since we have reduced to the case $y\neq 0$.) By applying implicit differentiation to the equation $x^2/a^2-y^2/b^2=1$, we find that $m=y^{'}:=y^{'}(x)$ satisfies

$$\frac{2x}{a^2}-\frac{2yy^{'}}{b^2}=0,\text{ whence }m=\frac{b^2x}{a^2y}.$$

So, we can take the tangential vector to $\mathcal{H}$ at P to be $\mathcal{T}=a^2y\mathbf{i}+b^2x\mathbf{j}$. We also have

$$\overrightarrow{PF_1}=(-c-x)\mathbf{i}-y\mathbf{j}\text{ and }\overrightarrow{PF_2}=(c-x)\mathbf{i}-y\mathbf{j},$$

$$|\overrightarrow{PF_1}|=\sqrt{(-c-x)^2+(-y)^2}=\sqrt{(c+x)^2+y^2}\text{ and}$$

$$|\overrightarrow{PF_2}|=\sqrt{(c-x)^2+(-y)^2}=\sqrt{(c-x)^2+y^2}.$$

Using the standard formula for dot product, we get

$$\mathcal{L}=\frac{a^2y(-c-x)+b^2x(-y)}{\sqrt{(-c-x)^2+(-y)^2}}=\frac{-a^2y(c+x)-b^2xy}{\sqrt{(c+x)^2+y^2}}\text{ and}$$

$$\mathcal{M}=\frac{a^2y(c-x)+b^2x(-y)}{\sqrt{(c-x)^2+(-y)^2}}=\frac{a^2y(c-x)-b^2xy}{\sqrt{(c-x)^2+y^2}}.$$

Consequently, $\mathcal{L} = \mathcal{M}$ if and only if

$$\frac{-a^2y(c+x) - b^2xy}{\sqrt{(c+x)^2+y^2}} = \frac{a^2y(c-x) - b^2xy}{\sqrt{(c-x)^2+y^2}}.$$

By substituting $b^2 = c^2 - a^2$, dividing through the just-displayed equation by (the nonzero quantity) y, performing some minor algebraic rewriting, and then dividing through by (the nonzero quantity) c, we see that our task is equivalent to proving that

$$\frac{-a^2 - cx}{\sqrt{(c+x)^2+y^2}} = \frac{a^2 - cx}{\sqrt{(c-x)^2+y^2}}.$$

We claim that if $|x| \geq a$, then $-a^2 - cx$ and $a^2 - cx$ have the same algebraic sign. Since $-a^2 - cx < a^2 - cx$, the claim will be shown if we prove the following two assertions: if $x \leq -a$, then $-a^2 - cx > 0$; and if $x \geq a$, then $a^2 - cx < 0$. Recalling that $0 < a < c$ (so that $-c < -a < 0$) and using the familiar rules for multiplying the partners of an inequality by the same nonzero real number, we get the following: $x \leq -a \Rightarrow cx \leq c(-a) = (-c)a < (-a)a = -a^2 \Rightarrow -a^2 - cx > 0$; and $x \geq a \Rightarrow cx > ax \geq a \cdot a = a^2 \Rightarrow a^2 - cx < 0$. This proves the above claim.

By the above claim, the left- and right-hand sides of the last-displayed equation have the same algebraic sign. Thus, squaring both sides of that equation produces another equivalent equation. Therefore, since cross-multiplying then produces yet another equivalent equation, we get that $\mathcal{L} = \mathcal{M}$ if and only if

$$(-a^2 - cx)^2[(c-x)^2 + y^2] = (a^2 - cx)^2[(c+x)^2 + y^2], \text{ or equivalently,}$$
$$(a^4 + 2a^2cx + c^2x^2)[x^2 + y^2 - 2cx + c^2] = (a^4 - 2a^2cx + c^2x^2)[x^2 + y^2 + 2cx + c^2].$$

Next, by using the generalized distributivity property to expand both the left- and right-hand sides of the last equation and then additively canceling like terms, we get that $\mathcal{L} = \mathcal{M}$ if and only if the quantity

$$Q := -2a^4cx + 2a^2cx^3 + 2a^2cxy^2 + 2a^2c^3x - 2c^3x^3$$

satisfies $Q = -Q$, that is, if and only if $Q = 0$. Next, since $x \neq 0$ (because $\mathcal{H}$ does not have a y-intercept), we can divide Q by (the nonzero quantity) $2cx$ and thus get that $\mathcal{L} = \mathcal{M}$ if and only if

$$-a^4 + a^2x^2 + a^2y^2 + a^2c^2 - c^2x^2 = 0.$$

Next, use $x^2/a^2 - y^2/b^2 = 1$ and $b^2 = c^2 - a^2$ to get

$$a^2y^2 = b^2x^2 - a^2b^2 = (c^2 - a^2)x^2 - a^2(c^2 - a^2) = c^2x^2 - a^2x^2 - a^2c^2 + a^4,$$

whence $-a^4 + a^2x^2 + a^2y^2 + a^2c^2 - c^2x^2 = 0$. The proof is complete. □

We close the section with two remarks. The first (resp., second) of these gives some information about tangent lines to hyperbolas (resp., some "hyperbolic" analogues of Remark 2.2 (b)-(d)).

Remark 4.2. Using notation from Theorem 4.1 and the condition

$$\frac{\mathcal{T} \cdot \overrightarrow{PF_1}}{|\mathcal{T}| \cdot |\overrightarrow{PF_1}|} = \frac{\mathcal{T} \cdot \overrightarrow{PF_2}}{|\mathcal{T}| \cdot |\overrightarrow{PF_2}|}$$

from the proof of Theorem 4.1, one can show that no line passing through F_1 is the tangent line to $\mathcal{H}$ at a point P on the upper half of $\mathcal{B}_2$. A direct proof of this fact can also be given via high school algebra and analytic geometry. Similar reasoning shows that no line passing through F_1 can intersect the upper (or lower) half of $\mathcal{B}_2$ more than once. The latter fact can be used to explain why the list of Figures 2-7 could not be augmented by, for instance, considering the possible existence of a line that would pass through F_1, intersect $\mathcal{B}_2$ after approaching from "outside" $\mathcal{B}_2$, and then also intersect $\mathcal{B}_2$ after approaching from "inside" $\mathcal{B}_2$. The remark is complete.

Remark 4.3. (a) If one is considering the east-west hyperbola with foci $F_1(-c,0)$ and $F_2(c,0)$, together with the positive real numbers $a<c$ (and $b:=\sqrt{c^2-a^2}$), one is led easily to the Cartesian equation

$$(c^2-a^2)x^2-a^2y^2=a^2(c^2-a^2).$$

Rather than dividing through by $a^2(c^2-a^2)$ in order to get the equation $x^2/a^2-y^2/b^2=1$, let us instead ask, in the spirit of Remark 2.2 (c), what the just-displayed equation would imply if we set $a=c$. Since $a\neq 0$, this substitution would lead to the equation $y^2=0$, so that $P(x,y)$ is $P(0,y)$, a point on the y-axis. In hoping for a converse, let us ask: if $P(x,y)$ is $P(0,y)$ and $0<a=c$ (whence $b=0$), does $|d_1(P)-d_2(P)|=2a$? No! Unfortunately, by a classical result in Euclidean plane geometry, *every* point $P(0,y)$ on the y-axis satisfies $d_1(P)=d_2(P)$, since P is on the perpendicular bisector of the line segment connecting F_1 to F_2. Thus, setting $a=c$ (>0) has led to the uninteresting fact that the empty set is a degenerate case of a hyperbola.

Of course, data satisfying the usual definition of a hyperbola could not support the equation $a=c$ (because the definition of a hyperbola stipulated that $a<c$). However, the above substitution was considered in the hope that, as had been the case for a similar substitution in Remark 2.2 (c), it might lead to an interesting degenerate case (this time, of a hyperbola). Although that did not happen, one could ask what, if anything, can be said about the implications of the last-displayed equation if we suppose also that $b=c$. Of course, *this* equation is not possible for parameters satisfying the usual definition of a hyperbola, but with the experience from Remark 2.2 behind us, that fact will not deter us in our search for a degenerate case of a hyperbola.

Setting $b=c$ (that is, $\sqrt{c^2-a^2}=c$), whence $a=0$, in the last-displayed equation gives $c^2x^2=0$. An analysis of the most extreme subcase, where *each* of a, b and c is 0, reveals that the entire Euclidean plane is a degenerate hyperbola! This conclusion is *very* untraditional and it shows the danger of straying too far from the stipulations in the standard definition of a hyperbola.

Let us return to setting $b=c$ (whence $a=0$), but this time insisting that $c>0$. Under these conditions, the implication of the last-displayed equation would be that $x=0$. In other words, assuming that $|d_1(P)-d_2(P)|=2a$, while also assuming that $b=c\neq 0$, gives $x=0$, so that $P(x,y)$ is $P(0,y)$, with P equidistant from $F_1(-c,0)$ and $F_2(c,0)$ (since $a=0$). The above-mentioned classical result in Euclidean plane geometry gives the converse conclusion that any point on the perpendicular bisector of the line segment connecting F_1 and F_2 is (on the y-axis and) equidistant from F_1 and F_2. Thus, we can conclude that if $0=a<c$ $(=b)$, then the set of points $P(x,y)$ such that $|d_1(P)-d_2(P)|=2a$ is the y-axis. *This* degenerate case is not just piecewise linear – it is not just a line segment – it is a line!

(b) Some readers who are familiar with the approach to degenerate cases in [1] may have wondered why Remark 2.2 (b)-(d) did not address the question of which degenerate cases of an ellipse can arise by using the approach in [1]. While we did not mention that matter in Remark 2.2 in the interest of brevity, the general method was touched on later in Remark 3.2 (h). Perhaps some readers will choose to pursue it further after reading the following discussion of some of the degenerate cases of hyperbolas which can be found by using the approach from [1].

The approach from [1] does not produce exactly the same degenerate cases of a hyperbola as in (a), but there is some overlap. Consider the standard-form equation of an east-west hyperbola with center at the origin, $x^2/a^2-y^2/b^2=1$. Fixing $(x,y)\in\mathbb{R}^2$, consider next subjecting both sides of this equation to the limit process where $a\to\infty$ and $b\to\infty$. The resulting equation is $0-0=1$, which is, of course, satisfied by only $(x,y)\in\emptyset$. So, from the point of view of [1], the empty set is a degenerate case of a hyperbola. Notice that the “equate some parameters" method in the first paragraph of (a) reached the same conclusion.

It is straightforward (but somewhat time-consuming) to carry out a case analysis to investigate what happens to the equation $x^2/a^2-y^2/b^2=1$ when both sides of this equation are subjected to a limit process in which a and b each (independently) approach (possibly different) elements of $\mathbb{R}\cup\{-\infty,\infty\}$. One upshot is that no such limit process produces a Cartesian equation of a (subset of

a) line. This fact stands in contrast to the result in the final paragraph of (a), where it was shown that the "equate some parameters" method *can* produce a Cartesian equation of a line (namely, the y-axis).

On the other hand, notice that if one subjects both sides of the equation $x^2/a^2 - y^2/b^2 = 1$ to the limit process in which $a \to 1$ and $b \to \infty$, the resulting equation, $x^2 - 0 = 1$, is the equation of a line (namely, the vertical line given by $x = 1$, "counted twice"). In this way, the point of view of [1] allows us to consider a single line as a degenerate case of a hyperbola. However, another straightforward (and somewhat less time-consuming) case analysis shows that no application of the "equate some parameters" method (in which some, perhaps all, of the real numbers a, b and c are equated) is able to produce a Cartesian equation of a single line.

By combining the results of the preceding three paragraphs, one confirms the assertions in the first sentence of the second paragraph of (b). By returning to the "characterization" theme from Section 3, let us next see (in Section 5) what other (if any) degenerate cases of hyperbolas can be discovered by carefully examining the implications of a planar figure satisfying the reflection properties of a hyperbola.

5 Reflection-theoretic characterizations of a hyperbola

This section presents several results giving reflection-theoretic characterizations of various subsets of a hyperbola. This work includes a deeper analysis of the reflection properties of a hyperbola that were established in Theorem 4.1. Just as Section 3 began with a result characterizing the top half of an "east-west" ellipse in standard position (resp., just as Section 3 of [3] began with a characterization of the top half of a parabola that opens to the right and is in standard position), this section begins with reflection-theoretic characterizations of the top half of a branch of an "east-west" hyperbola in standard position.

The statement of Theorem 5.1 makes explicit that each focus of the hyperbola can play the role of "emitter" or the role of "attractor" in such a characterization. While the statement of Theorem 3.1 included a similar fact about the foci of an ellipse, the elucidation of the correspondingly detailed information about hyperbolas leads to a longer statement of Theorem 5.1. This situation is unavoidable because hyperbolas are intrinsically more complicated than ellipses. Perhaps the most striking evidence of *that* fact is the disconnectedness of the branches of a hyperbola. Moreover, although we used only one figure in Section 2 to depict the reflection properties of an ellipse, it required six figures (in Section 4) to reveal all the nuances of the reflection properties of a hyperbola.

Note that the reduction to considering hyperbola(-like graph)s in standard position that is assumed in the setting for Theorem 5.1 (and in some, but not all, of the other characterization results in this section) is done essentially without loss of generality. Indeed, this assertion can be confirmed, just as the corresponding assertion about ellipses was confirmed four paragraphs before the statement of Theorem 2.1, by appealing to some basic principles of Euclidean geometry about distance and the measure of undirected angles being invariant under rigid rotation and/or translations of coordinate axes.

Theorem 5.1. Let $0 < a < c$ in $\mathbb{R}$, and put $b := \sqrt{c^2 - a^2}$. Working in a fixed Euclidean plane, consider the points $F_1(-c, 0)$ and $F_2(c, 0)$. Let $f : [a, \infty) \to \mathbb{R}$ be a function, and let Γ be the graph of f. Suppose that f is strictly increasing on $[a, \infty)$ and differentiable on (a, ∞), $f'(x) \neq 0$ for all $x > a$, f is continuous at $x = a$, and $f(a) = 0$. For each point P on Γ, let $\mathcal{T} := \mathcal{T}_P$ be a tangential vector to Γ at P. Let $\mathcal{H}$ be the hyperbola with foci F_1 and F_2 and with semitransverse axis a (and necessarily with semiconjugate axis b). Let the branches of $\mathcal{H}$ be labeled $\mathcal{B}_1$ and $\mathcal{B}_2$ in such a way that F_1 (resp., F_2) is contained "inside" $\mathcal{B}_1$ (resp., $\mathcal{B}_2$) in the obvious intuitive sense. Then the following eight conditions are equivalent:

(1) $f(x) = (\frac{b}{a})\sqrt{x^2 - a^2}$ for all $x \in [a, \infty)$;

(2) Γ is the "top half" of the right-hand branch $\mathcal{B}_2$ of the "east-west" hyperbola $\mathcal{H}$;

(3) For each point P on Γ, the angle between $\mathcal{T}_P$ and $\overrightarrow{PF_1}$ is congruent to the angle between $\mathcal{T}_P$ and $\overrightarrow{PF_2}$;

(4) For each point P on Γ,

$$\frac{\mathcal{T}_P \cdot \overrightarrow{PF_1}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_1}|} = \frac{\mathcal{T}_P \cdot \overrightarrow{PF_2}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_2}|};$$

(5) If a ray $\mathcal{R}_1$ is emitted from F_1 and intersects Γ at a point P, then $\mathcal{R}_1$ reflects off Γ and as a result of that reflection, the redirected ray moves along a line which appears to have originated from F_2 (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass through F_2);

(6) If a ray $\mathcal{R}_2$ is emitted from F_2 and intersects Γ at a point P, then $\mathcal{R}_2$ reflects off Γ and as a result of that reflection, the redirected ray appears to have originated from F_1 (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass through F_1);

(7) If a ray $\mathfrak{S}_1$ approaches Γ from "outside" Γ along a line of action that passes through (that is, that would have passed through) F_2 then, as a result of intersecting Γ at a point P, the ray is reflected/diverted along a new line of action which passes through F_1;

(8) If a ray $\mathfrak{S}_2$ approaches Γ from "inside" Γ along a line of action that passes through (that is, that would have passed through) F_1 then, as a result of intersecting Γ at a point P, the ray is reflected/diverted along a new line of action which passes through F_2.

Proof. The hypotheses on f ensure that $f(x) > 0$ whenever $x > a$; and, hence, that if a point $P(x,y)$ is on Γ with $x > 0$, then $y > 0$. It follows that if $P(x,y)$ is on Γ with $x > 0$, then P is not F_2. Also, $P \neq F_1$, since $-a < 0 < x$. Thus, each of the bound vectors $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$ is nonzero. Therefore, the fractions appearing in the statement of condition (4) are well defined.

We will assume henceforth that $P(x,y)$ is a point on Γ, with coordinates (x,y). Let us consider first the case where $x = a$, that is, where P is the vertex $V(a,0)$ of $\mathcal{H}$. By the previous paragraph, V is the only point on Γ whose second coordinate is 0. We claim that the tangent line to Γ at V is vertical. We will prove this claim by adapting the reasoning in the second paragraph of Remark 3.2 (a). Since f is continuous, it will suffice to show that $\lim_{x \to a^+} f'(x) = \infty$. We will show this by using the following fact:

$$f'(x) = \frac{c^2 - x^2 + y^2 + \sqrt{(x^2 + y^2 + c^2)^2 - 4c^2x^2}}{2xy} \quad \text{(if } x > a).$$

The fact which was just displayed will be proved later in this proof. We assure the reader that this proof contains no "circular arguments."

As in Remark 3.2 (a), let us examine the secant lines whose limiting position (if it exists) would be that of the tangent line to Γ at V. The corresponding limit of the slopes of those secant lines is

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \lim_{x \to a^+} \frac{f(x) - 0}{x - a} = \lim_{x \to a^+} f'(x),$$

where the last step followed from the general form of L'Hôpital's Rule (as formulated in [12, Theorem 1]). Next, by appealing to the above formula for $f'(x)$ (when $x > a$), we can reformulate our task as seeking a proof that

$$\lim_{x \to a^+} \frac{c^2 - x^2 + y^2 + \sqrt{(x^2 + y^2 + c^2)^2 - 4c^2x^2}}{2xy} = \infty.$$

Working in the extended real number system and using the appropriate limit theorems there, we find this limit to be

$$\frac{c^2 - a^2 + |c^2 - a^2|}{0^+} = \frac{c^2 - a^2 + (c^2 - a^2)}{0^+} = \frac{2(c^2 - a^2)}{0^+} = \infty,$$

as desired. This completes the proof of the above claim (modulo the proof, which will be given soon, of the formula for $f^{'}(x)$ which was used above).

By the above claim, we can take the tangential vector $\mathcal{T}$ to Γ at V to be $\mathbf{j}$. Then, as in the proof of Theorem 3.1, we can use the Principle of Reflection to show that the six conditions (3)-(8) are all satisfied if P is V (by combining the fact that $\overrightarrow{VF_1}$ and $\overrightarrow{VF_2}$ are horizontal vectors with the fundamental principle of Euclidean geometry that all right angles are congruent). Therefore, for the rest of this proof, as we consider the behavior of points $P(x,y)$ on Γ, we can assume that $P \neq V$, that is, that $x > 0$ (and so $y > 0$).

Next, we apply our familiar combination of vectorial material (including dot products), the inverse cosine function and the Principle of Reflection. (A summary of where to find the details of most of that background methodology can be found in the fourth paragraph of the proof of Theorem 4.1.) The upshot is that conditions (3), (4), (5), (6), (7) and (8) are equivalent. Theorem 4.1 ensures that (2) implies some of those of those equivalent conditions, while it is classically known that (1) $\Leftrightarrow$ (2). Accordingly, it remains only to prove that if $x > 0$, then (4) (when predicated for points $P(x,y)$ on Γ with $x > 0$) implies both (1) and the above-used formula for $f^{'}(x)$.

As in the proof of Theorem 4.1, if $P(x,y)$ is on Γ (with $x > 0$ and $y > 0$), then (4) ensures (after we multiply through by $|\mathcal{T}|$) that

$$\frac{\mathcal{T}\cdot\overrightarrow{PF_1}}{|\overrightarrow{PF_1}|} = \frac{\mathcal{T}\cdot\overrightarrow{PF_2}}{|\overrightarrow{PF_2}|}.$$

Next, let m denote the slope of the tangent line to Γ at P. As $m = f^{'}(x)$, we can take $\mathcal{T} = \mathcal{T}_P = 1\,\mathbf{i} + m\,\mathbf{j} = \mathbf{i} + f^{'}(x)\,\mathbf{j}$. We also have

$$\overrightarrow{PF_1} = (-c-x)\mathbf{i} - y\,\mathbf{j},\ \ \overrightarrow{PF_2} = (c-x)\mathbf{i} - y\,\mathbf{j},$$

$$|\overrightarrow{PF_1}| = \sqrt{(c+x)^2+y^2} \text{ and } |\overrightarrow{PF_2}| = \sqrt{(c-x)^2+y^2}.$$

Using the standard formula for dot product, we can now rewrite the above consequence of (4) as

$$\frac{1(-c-x)+f^{'}(x)(-y)}{\sqrt{(c+x)^2+y^2}} = \frac{1(c-x)+f^{'}(x)(-y)}{\sqrt{(c-x)^2+y^2}}, \text{ equivalently, } f^{'}(x) =$$

$$\frac{dy}{dx} = \frac{(c-x)\sqrt{(c+x)^2+y^2}+(c+x)\sqrt{(c-x)^2+y^2}}{y[\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2}]}.$$

We next summarize how the just-obtained formula for the derivative can be algebraically simplified to give the earlier-asserted formula for $f^{'}(x)$. First, rationalize the denominator of the right-hand side of the last display; that is, multiply both the numerator and the denominator of that right-hand side by $\sqrt{(c+x)^2+y^2} + \sqrt{(c-x)^2+y^2}$. This leads to an (algebraically) equivalent ODE that can be algebraically rewritten as $(dy)/dx) = [A+B][C+D]/(y4cx)$, where

$$A := (c-x)\sqrt{(c+x)^2+y^2}, B := (c+x)\sqrt{(c-x)^2+y^2},$$

$$C := \sqrt{(c+x)^2+y^2} \text{ and } D := \sqrt{(c-x)^2+y^2}.$$

Further algebraic simplification, followed by dividing both the numerator and the denominator of the then ambient right-hand side by $2c$ gives

$$\frac{dy}{dx} = \frac{c^2+x^2+y^2+\sqrt{(c+x)^2+y^2}\sqrt{(c-x)^2+y^2}}{2xy} - \frac{x}{y}.$$

This easily simplifies to

$$\frac{dy}{dx} = \frac{c^2 - x^2 + y^2 + \sqrt{(x^2+y^2+c^2)^2 - 4c^2x^2}}{2xy},$$

thus proving the earlier-asserted formula for $f'(x)$.

The above formula for $f'(x)$ is similar to, but not the same as, the corresponding formula in the proof of Theorem 3.1. This is not surprising, as different types of conic sections could be expected to have their (different) reflection properties encoded by different ODEs. Nevertheless, since the proof of Theorem 3.1 (about ellipses) was able to solve its relevant ODE by using the same changes of variables that had been used in solving the relevant ODE in a result about parabolas [3, Theorem 3.1], let us see whether those same changes of variables will also be useful here, in our attempt to characterize the reflection properties of (a significant subset of) a hyperbola. Accordingly, we let

$$w := y^2 + c^2 + x^2 \text{ and } v := w/x\text{; then, as before,}$$

$$\frac{dw}{dx} = 2y\frac{dy}{dx} + 2x.$$

So, by substituting the just-displayed facts into the above ODE and doing some minor algebraic rewriting, we get

$$\frac{dw}{dx} = \frac{c^2 - x^2 + y^2 + \sqrt{w^2 - 4c^2x^2}}{x} + 2x = \frac{c^2 + x^2 + y^2 + \sqrt{w^2 - 4c^2x^2}}{x},$$

$$\text{whence } x\frac{dw}{dx} = w + \sqrt{w^2 - 4c^2x^2} \text{ (for all} \, x > a).$$

Therefore

$$\frac{dv}{dx} = -\frac{w}{x^2} + \frac{\frac{dw}{dx}}{x} = -\frac{w}{x^2} + \frac{\frac{w}{x} + \frac{\sqrt{w^2-4c^2x^2}}{x}}{x} = \frac{\sqrt{w^2 - 4c^2x^2}}{x^2}, \text{ and so}$$

$$\frac{dv}{dx} = \frac{\sqrt{v^2 - 4c^2}}{x} \text{ for all } x > a.$$

Hence, by separating variables and performing indefinite integration, we have (if $x > a$) that

$$\int \frac{dv}{\sqrt{v^2 - 4c^2}} = \int \frac{dx}{x} + K,$$

with constant of integration K.

According to a table of (indefinite) integrals (specifically, formula 27 on the page opposite the inside front cover of [9]), if k is any positive real number,

$$\int \frac{dt}{\sqrt{t^2 - k^2}} = \ln(|t + \sqrt{t^2 - k^2}|) + C.$$

By applying the just-displayed formula (with $k := 2c$) to the last result of the preceding paragraph, we get

$$\ln(|v + \sqrt{v^2 - 4c^2}|) = \ln(|x|) + K.$$

Next, exponentiate both sides of the last display, and then rewrite the resulting equation by using the property that $\ln(\lambda/\nu) = \ln(\lambda) - \ln(\nu)$ for all positive λ and ν. This gives that $|(v + \sqrt{v^2 - 4c^2})/x|$ is constant (for all $x > a$). Therefore, there exists a constant $E > 0$ such that

$$v + \sqrt{v^2 - 4c^2} = Ex \text{ whenever } x > a.$$

Substituting $v = w/x$ into the last display, then multiplying through by x, and then using the definition of w leads to

$$y^2 + c^2 + x^2 + \sqrt{(y^2 + c^2 + x^2)^2 - 4c^2x^2} = Ex^2 \text{ if } x > a.$$

The just-displayed equation has two useful consequences. First, since $f(a) = 0$, applying the limiting process $\lim_{x\to a^+}$ gives

$$2c^2 = c^2 + a^2 + |c^2 - a^2| = c^2 + a^2 + \sqrt{(c^2 - a^2)^2} =$$

$$c^2 + a^2 + \sqrt{(c^2 + a^2)^2 - 4c^2a^2} = f(a)^2 + c^2 + a^2 + \sqrt{[f(a)^2 + c^2 + a^2]^2 - 4c^2a^2} =$$

$$\lim_{x\to a^+}(y^2 + c^2 + x^2 + \sqrt{(y^2 + c^2 + x^2)^2 - 4c^2x^2}) = \lim_{x\to a^+} Ex^2 = Ea^2.$$

It follows that

$$E = \frac{2c^2}{a^2} = \frac{2(a^2 + b^2)}{a^2} = 2 + \frac{2b^2}{a^2}.$$

To get the second of the above-promised “useful consequences," transpose two terms of the last equation in the next-to-last paragraph and then square both sides of the resulting equation. This gives

$$(y^2 + c^2 + x^2 - Ex^2)^2 = (y^2 + c^2 + x^2)^2 - 4c^2x^2.$$

After several additive cancellations, we get the second “useful consequence":

$$E^2x^4 - 2Ex^4 - 2Ex^2y^2 - 2Ec^2x^2 = -4c^2x^2.$$

Substituting $E = 2+2b^2/a^2$ and $c^2 = a^2+b^2$ into the last display leads to an equivalent equation. As the hypotheses rule out $x = 0$, we can divide through the last-mentioned equation by $-4x^2$ and also multiply through that equation by a^2, thus producing the equivalent equation

$$a^2b^2 = -(b^2 + \frac{b^4}{a^2})x^2 + a^2y^2 + b^2y^2 + a^2(2b^2 + \frac{b^4}{a^2}).$$

Algebraic simplification then gives the following equivalent equation:

$$(\frac{b^2a^2 + b^4}{a^2})x^2 - (a^2 + b^2)y^2 = a^2b^2 + b^4.$$

Finally, dividing through by $(a^2 + b^2)b^2$ gives

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1,$$

whence $f(x) = y = (b/a)\sqrt{x^2 - a^2}$. The proof is complete. □

Since Theorem 3.1 used reflection properties to characterize the upper half of an ellipse in standard position and Theorem 5.1 used reflection properties to characterize only half of the upper half of a hyperbola $\mathcal{H}$ in standard position, it seems natural to ask if one can use reflection properties to characterize the “other half" of the upper half of $\mathcal{H}$. Next, Corollary 5.2 gives an affirmative answer to this question. As some instructors may prefer to interchange the roles of Theorem 5.1 and Corollary 5.2 by covering the latter before the former, some of the proof of Corollary 5.2 will, for the sake of accessibility (and completeness), repeat some of the material from the proof of Theorem 5.1.

Corollary 5.2. *Let $0 < a < c$ in $\mathbb{R}$, and put $b := \sqrt{c^2 - a^2}$. Working in a fixed Euclidean plane, consider the points $F_1(-c, 0)$ and $F_2(c, 0)$. Let $g : (-\infty, -a] \to \mathbb{R}$ be a function, and let γ be the graph of g. Suppose that g is strictly decreasing on $(-\infty, -a]$ and differentiable on $(-\infty, -a)$, $g'(x) \neq 0$ for all $x < -a$, g is continuous at $x = -a$, and $g(-a) = 0$. For each point Q on γ, let $\mathcal{U} := \mathcal{U}_Q$ be a tangential vector to γ at Q. Let $\mathcal{H}$ be the hyperbola with foci F_1 and F_2 and with semitransverse axis a (and necessarily with semi-conjugate axis b). Let the branches of $\mathcal{H}$ be labeled $\mathcal{B}_1$ and $\mathcal{B}_2$ in such a way that F_1 (resp., F_2) is contained "inside" $\mathcal{B}_1$ (resp., $\mathcal{B}_2$) in the obvious intuitive sense. Then the following eight conditions are equivalent:*

(1) $g(x) = (\frac{b}{a})\sqrt{x^2 - a^2}$ for all $x \in (-\infty, -a]$;

(2) γ is the "top half" of the left-hand branch $\mathcal{B}_1$ of the "east-west" hyperbola $\mathcal{H}$;

(3) For each point Q on γ, the angle between $\mathcal{U}_Q$ and $\overrightarrow{QF_1}$ is congruent to the angle between $\mathcal{U}_Q$ and $\overrightarrow{QF_2}$;

(4) For each point Q on γ,

$$\frac{\mathcal{U}_Q \cdot \overrightarrow{QF_1}}{|\mathcal{U}_Q| \cdot |\overrightarrow{QF_1}|} = \frac{\mathcal{U}_Q \cdot \overrightarrow{QF_2}}{|\mathcal{U}_Q| \cdot |\overrightarrow{QF_2}|};$$

(5) If a ray $\mathcal{R}_2$ is emitted from F_2 and intersects γ at a point Q, then $\mathcal{R}_2$ reflects off γ and as a result of that reflection, the redirected ray moves along a line which appears to have originated from F_1 (that is, after the reflection, the new direction of the ray is such that its opposite ray would pass through F_1);

(6) If a ray $\mathcal{R}_1$ is emitted from F_1 and intersects γ at a point Q, then $\mathcal{R}_1$ reflects off γ and as a result of that reflection, the redirected ray appears to have originated from F_2 (that is, after the reflection, the new direction of the ray is such that its opposite ray would pass through F_2);

(7) If a ray $\mathfrak{S}_2$ approaches γ from "outside" γ along a line of action that passes through (that is, that would have passed through) F_1 then, as a result of intersecting γ at a point Q, the ray is reflected/diverted along a new line of action which passes through F_2;

(8) If a ray $\mathfrak{S}_1$ approaches γ from "inside" γ along a line of action that passes through (that is, that would have passed through) F_2 then, as a result of intersecting γ at a point Q, the ray is reflected/diverted along a new line of action which passes through F_1.

Proof. Define a function $f : [a, \infty) \to \mathbb{R}$ by $f(x) := g(-x)$ for all $x \geq a$. As the chain rule gives $f'(x) = -g'(-x)$ for all $x > a$ (and so $g'(x) = -f'(-x)$ for all $x < -a$), it follows from the hypotheses on g that f is strictly increasing on $[a, \infty)$ and differentiable on (a, ∞), $f'(x) \neq 0$ for all $x > a$, f is continuous at $x = a$, and $f(a) = 0$. Hence, by tweaking the reasoning in the first, second and third paragraphs of the proof of Theorem 5.1, we get the following: $g'(x) < 0$ for all $x < -a$; if $Q(x, y)$ is any point on γ such that $x < -a$, then $y > 0$ and Q is neither F_1 nor F_2, so that both $\overrightarrow{QF_1}$ and $\overrightarrow{QF_2}$ are nonzero bound vectors; and $\lim_{x \to (-a)^-} g'(x) = -\infty$, whence the tangent line to γ at the point $V(-a, 0)$ is vertical. Thus, by tweaking the reasoning in the second paragraph of the proof of Theorem 3.1, we can use the Principle of Reflection to show that each of the conditions (3)-(8) holds if the point Q is V. Accordingly, for the rest of this proof, we can assume, without loss of generality, that any point Q on γ which is being considered has coordinates (x, y) with $x < -a$ and $y > 0$.

Next, note that the points $Q(x, y)$ on the graph γ of g are in one-to-one correspondence with the points $P(t, Y)$ on the graph Γ of f, via the equations $t = -x$ and $Y = y$ (equivalently, $x = -t$ and $y = Y$). In particular, if one uses the tangential vector

$$\mathcal{U}_Q = \mathbf{i} + g'(x)\mathbf{j} = \mathbf{i} + m\mathbf{j}$$

to γ at a point $Q(x, y)$ of γ where $x < -a$ and $m := g'(x)$ $(= -f'(-x))$, then it would be appropriate to use the tangential vector

$$\mathcal{T}_P = \mathbf{i} + f'(-x)\mathbf{j} = \mathbf{i} + (-m)\mathbf{j} = \mathbf{i} - m\mathbf{j}$$

at the corresponding point $P(-x,y)$ of Γ. As we also have

$$\overrightarrow{QF_1} = (-c-x)\mathbf{i} - y\,\mathbf{j},\ \overrightarrow{QF_2} = (c-x)\mathbf{i} - y\,\mathbf{j},\ \overrightarrow{PF_1} = (-c+x)\mathbf{i} - y\,\mathbf{j} \text{ and}$$

$\overrightarrow{PF_2} = (c+x)\mathbf{i} - y\,\mathbf{j}$, it follows that (4) holds if and only if

$$\frac{1(-c-x)+m(-y)}{\sqrt{1^2+m^2}\cdot\sqrt{(-c-x)^2+(-y)^2}} = \frac{1(c-x)+m(-y)}{\sqrt{1^2+m^2}\cdot\sqrt{(c-x)^2+(-y)^2}},$$

that is, if and only if

$$\frac{c+x+my}{\sqrt{(c+x)^2+y^2}} = \frac{x-c+my}{\sqrt{(c-x)^2+y^2}};$$

and, similarly, that condition (4) of Theorem 5.1 holds if and only if

$$\frac{1(-c+x)+(-m)(-y)}{\sqrt{1^2+(-m)^2}\cdot\sqrt{(-c+x)^2+(-y)^2}} = \frac{1(c+x)+(-m)(-y)}{\sqrt{1^2+(-m)^2}\cdot\sqrt{(c+x)^2+(-y)^2}},$$

that is, if and only if

$$\frac{x-c+my}{\sqrt{(c-x)^2+y^2}} = \frac{c+x+my}{\sqrt{(c+x)^2+y^2}}.$$

It is now clear that (4) holds if and only if condition (4) of Theorem 5.1 holds.

We next apply our customary combination of vectorial reasoning and the Principle of Reflection (along with the fact that $\cos|_{(0,\pi)}$ is a one-to-one function). This is to be done in the spirit of the appropriate parts of the proofs of Theorems 2.1, 3.1, 4.1 and 5.1; see, especially, the fourth paragraph of the proof of Theorem 4.1. (Some readers may find it useful to also see the following in [3]: the first paragraph of the proof of its Theorem 2.1; and the third and fourth paragraphs of the proof of its Theorem 3.1.) The upshot is that the conditions (3), (4), (5), (6), (7) and (8) are equivalent. It is also clear that condition (1) is equivalent to condition (1) of Theorem 5.1; and that condition (2) is equivalent to condition (2) of Theorem 5.1. By combining this information with the final assertion of the preceding paragraph and the fact that Theorem 5.1 has already been proven, one sees that this proof of Corollary 5.2 is complete. □

Remark 5.3. Throughout the various parts of this remark, it will be convenient to continue using the following notation from Theorem 5.1 and Corollary 5.2. Let $0 < a < c$ in $\mathbb{R}$, and put $b := \sqrt{c^2-a^2}$. Let $\mathcal{H}$ be the hyperbola with foci F_1 and F_2 and with semitransverse axis a (and necessarily with semi-conjugate axis b); that is, $\mathcal{H}$ is the hyperbola with Cartesian equation $x^2/a^2 - y^2/b^2 = 1$.

(a) We next mention two other proofs of Corollary 5.2. The first of these also uses Theorem 5.1 but, instead of using explicit expansions of dot products (as in the above proof), it uses the “conformal" point of view that was explained in the first paragraph of Remark 3.2 (c). That point of view would be applicable because, as we showed above, the slope of the tangent line to γ at Q is the negative of the slope of the tangent line to Γ at P. (It is interesting to note that the just-mentioned point of view also appeared in our reflection-theoretic study of parabolas: see [3, Remark 3.2 (c)].) The second alternative proof of Corollary 5.2 does not use Theorem 5.1 but, instead, essentially repeats that proof as adjusted for the context of Corollary 5.2. While this second alternative proof is longer than the first alternative proof, its use in a classroom would emphasize the fact that the left-hand branch of $\mathcal{H}$ is just as important as the right-hand branch of $\mathcal{H}$. Depending on the level of the audience, an instructor could then go on to obtain Theorem 5.1 as a corollary of (the independently proved) Corollary 5.2 and/or to discuss a relevant notion of “isomorphism".

(b) Let f and Γ be as in Theorem 5.1, and let g and γ be as in Corollary 5.2. Assume also that $g(x) = f(-x)$ for all $x < a$; equivalently, that $f(x) = g(-x)$ for all $x > a$. Let $\mathcal{G} := \Gamma \cup \gamma$. One can obtain 17 characterizations of the top half of $\mathcal{H}$ by stating the equivalence of the following conditions: (i) $\mathcal{G}$ is the top half of $\mathcal{H}$; (ii)-(ix) are the (equivalent) conditions (1)-(8) of Theorem 5.1; and (x)-(xvii) are the (equivalent) conditions (1)-(8) of Corollary 5.2.

(c) The changes of variables given by $(x,y) \leftrightarrow (t,Y) := (x,-y)$ can be used to convert the characterizations of the upper part of the right- (resp., left-)hand branch of $\mathcal{H}$ in Theorem 5.1 (resp., in Corollary 5.2) into characterizations of the lower part of the right- (resp., left-)hand branch of $\mathcal{H}$. Details of the proofs of these two assertions are left to the reader, as the reasoning in the proof of Remark 3.2 (b) (which was a result about a subset of an ellipse) carries over, *mutatis mutandis*, to the present context (even though that concerns results about a subset of a hyperbola).

(d) In the spirit of (b), we next point out that the two results which were given in (c) can be combined to produce 17 characterizations of the bottom half of $\mathcal{H}$. The key addition to this "combination" of results is to specify that the functions, say h_1 and h_2, whose respective graphs are related to the bottom half of the left- (resp., right-)hand branch of $\mathcal{H}$ in the two assertions in (c) are now also to be assumed to satisfy $h_1(x) = h_2(-x)$ for all $x < -a$ (equivalently, $h_2(x) = h_1(-x)$ for all $x > a$). Details are left to the reader. It is interesting that the relevant changes of variables, given by $(x,y) \leftrightarrow (t,Y) := (-x,y)$, have already been useful in the proofs of Theorem 3.1 and Corollary 5.2.

(e) In the spirit of (b) and (d), we point out that Theorem 5.1 can be combined with the first assertion in (c) to produce characterizations of the right-hand branch of $\mathcal{H}$. Details are left to the reader.

(f) In the spirit of (b), (d) and (e), we point out that Corollary 5.2 can be combined with the second assertion in (c) to produce characterizations of the left-hand branch of $\mathcal{H}$. Details are left to the reader. The remark is complete.

It seems natural to ask if there occasionally exists an alternative to the "combining of results" approach that was used in the proofs of parts (b), (d), (e) and (f) of Remark 5.3. In particular, with respect to Remark 5.3 (e)-(f), one can ask the following. Is there a "one-step" method to obtain characterizations of a branch of a hyperbola? Theorem 5.4 gives an affirmative answer, by using the "x as a function of y" point of view to derive such characterizations of the right-hand branch of an east-west hyperbola $\mathcal{H}$ that is in standard position (that is, with foci on the x-axis and center at the origin). Interested readers are invited to adapt Theorem 5.4 to give a "one-step" method to obtain characterizations of the left-hand branch of $\mathcal{H}$. Instructors who use [3, Remarks 2.2 (c) and 3.2 (d)] in their courses may find Theorem 5.4 (and the preceding sentence) to be useful for either lectures or homework/examinations.

Theorem 5.4. Let $0 < a < c$ in $\mathbb{R}$, and put $b := \sqrt{c^2 - a^2}$. Working in a fixed Euclidean plane, consider the points $F_1(-c,0)$ and $F_2(c,0)$. Let $\mathcal{H}$ be the hyperbola with foci F_1 and F_2 and with semitransverse axis a (and necessarily with semi-conjugate axis b). Let $\lambda : \mathbb{R} \to [a,\infty)$ be a function, with γ denoting the graph of the equation $x = \lambda(y)$. Suppose that λ is differentiable on $(-\infty,0) \cup (0,\infty)$, continuous at $y = 0$, strictly decreasing on $(-\infty,0)$ and strictly increasing on $(0,\infty)$. Suppose also that $\lambda(0) = a$ and $\lambda^{'}(y) \neq 0$ whenever $y \neq 0$. For each point Q on γ, let $\mathcal{U} := \mathcal{U}_Q$ be a tangential vector to γ at Q. Then the following eight conditions are equivalent:

(1) $\lambda(y) = (\frac{a}{b})\sqrt{y^2 + b^2}$ for all $y \in (-\infty,0) \cup (0,\infty)$;

(2) γ is the right-hand branch of the "east-west" hyperbola $\mathcal{H}$;

(3) For each point Q on γ, the angle between $\mathcal{U}_Q$ and $\overrightarrow{QF_1}$ is congruent to the angle between $\mathcal{U}_Q$ and $\overrightarrow{QF_2}$;

(4) For each point Q on γ,

$$\frac{\mathcal{U}_Q\cdot\overrightarrow{QF_1}}{|\mathcal{U}_Q|\cdot|\overrightarrow{QF_1}|}=\frac{\mathcal{U}_Q\cdot\overrightarrow{QF_2}}{|\mathcal{U}_Q|\cdot|\overrightarrow{QF_2}|};$$

(5) If a ray $\mathcal{R}_1$ is emitted from F_1 and intersects γ at a point Q, then $\mathcal{R}_1$ reflects off γ and as a result of that reflection, the redirected ray moves along a line which appears to have originated from F_2 (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass through F_2);

(6) If a ray $\mathcal{R}_2$ is emitted from F_2 and intersects γ at a point Q, then $\mathcal{R}_2$ reflects off γ and as a result of that reflection, the redirected ray appears to have originated from F_1 (that is, after the reflection, the new direction of the ray is such that its *opposite* ray would pass through F_1);

(7) If a ray $\mathfrak{S}_1$ approaches γ from "outside" γ along a line of action that passes through (that is, that would have passed through) F_2 then, as a result of intersecting γ at a point Q, the ray is reflected/diverted along a new line of action which passes through F_1;

(8) If a ray $\mathfrak{S}_2$ approaches γ from "inside" γ along a line of action that passes through (that is, that would have passed through) F_1 then, as a result of intersecting γ at a point Q, the ray is reflected/diverted along a new line of action which passes through F_2.

Proof. Let us first examine what can be learned if the condition (4) is assumed to hold at each point Q on γ whose second coordinate is nonzero. Consider any such point Q with coordinates $(x,y)=(\lambda(y),y)$ where $y\neq 0$. Put $m:=\lambda'(y)$. We can take the tangential vector $\mathcal{U}:=\mathcal{U}_Q$ to be $m\mathbf{i}+\mathbf{j}$. Since

$$\overrightarrow{QF_1}=(-c-x)\mathbf{i}+(0-y)\mathbf{j}\text{ and }\overrightarrow{QF_2}=(c-x)\mathbf{i}+(0-y)\mathbf{j},$$

it follows easily from (4) that

$$\frac{m(-c-x)+1(0-y)}{\sqrt{(-c-x)^2+(0-y)^2}}=\frac{m(c-x)+1(0-y)}{\sqrt{(c-x)^2+(0-y)^2}},\text{ whence}$$

$$\frac{dx}{dy}=\lambda'(y)=m=\frac{y[\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2}]}{(c+x)\sqrt{(c-x)^2+y^2}+(c-x)\sqrt{(c+x)^2+y^2}}.$$

This formula ensures that $m\neq 0$ (essentially because $c+x$ cannot be equal to either $c-x$ or $x-c$). Consequently,

$$\frac{1}{(\frac{dx}{dy})}=\frac{(c+x)\sqrt{(c-x)^2+y^2}+(c-x)\sqrt{(c+x)^2+y^2}}{y[\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2}]}\quad(\text{if }y\neq 0).$$

We have seen the right-hand side of the just-displayed formula before – in the proof of Theorem 5.1, where (something algebraically equivalent to) it was the formula for a certain derivative. To connect that work with the present context, we need to pause and consider a certain inverse function, in preparation for an application of the Inverse Function Theorem.

The hypotheses ensure that $\lambda'(y)<0$ if $y<0$; $\lambda'(y)>0$ if $y>0$; and $\lambda(y)=a$ if and only if $y=0$. Those hypotheses also imply that the function

$$\lambda_1:=\lambda|_{[0,\infty)}$$

is strictly increasing and, hence, has an inverse function. It will be convenient to consider the function $h:=\lambda_1^{-1}$. Of course, the range of h is $[0,\infty)$ (since that is the domain of λ_1). We claim, and we will prove, that the domain of h is $[a,\infty)$; equivalently, that the range of λ_1 is $[a,\infty)$. Since λ_1 is a strictly increasing continuous function and $\lambda_1(0)=a$, it follows from the Intermediate Value Theorem (for continuous functions) that this claim is also equivalent to the assertion that $\lim_{y\to\infty}\lambda(y)=\infty$. The proof of this fact will use the above formula for $\lambda^{'}(y)$ (for all $y\neq 0$).

This paragraph is the result of lightly editing a comment that was made at a similar stage of our study of parabolas in [3]. The details which are involved in proving the above claim are somewhat predictable, admittedly tedious at some points, and (in my opinion) necessary if one is to compete a proof of Theorem 5.4 having started from the "x as a function of y" point of view.

Suppose the above claim fails. Then there exists a unique least real number M such that $\lambda(y)\leq M$ for all $y\geq 0$. The assumptions ensure that $M>a$. Since λ_1 is a continuous function whose domain is the interval $[0,\infty)$, it follows from the Intermediate Value Theorem that the range of λ_1 is necessarily an interval, say J. Let J^* denote the corresponding closed interval (that is, the topological closure of J in $\mathbb{R}$). The left-hand endpoint of J^* is a, and $a\in J$. As λ_1 is strictly increasing and continuous, it is easy to see that the right-hand endpoint of J^* must be M and that $M\notin J$. Thus, the range of λ_1 is $[a,M)$. Next, since λ is strictly monotonic and differentiable on the open interval $(0,\infty)$ and $\lambda_1^{'}(y)\neq 0$ for all $y>0$, it follows from the version of the Inverse Function Theorem in [11, Theorem II, page 70] that $(\lambda_1^{-1}=)\ h$ is strictly monotonic and differentiable on the open interval (a,M) and that

$$h^{'}(x)=\frac{dy}{dx}=\frac{1}{\lambda_1^{'}(y)} \text{ for all } x\in(a,M).$$

Thus, by the above formula for $1/\lambda_1^{'}(y)$ (for all $y>0$),

$$h^{'}(x)=\frac{(c+x)\sqrt{(c-x)^2+y^2}+(c-x)\sqrt{(c+x)^2+y^2}}{y[\sqrt{(c+x)^2+y^2}-\sqrt{(c-x)^2+y^2}]} \text{ (if } a<x<M).$$

Therefore, by some algebraic simplification that was done in the proof of Theorem 5.1,

$$h^{'}(x)=\frac{c^2-x^2+y^2+\sqrt{(x^2+y^2+c^2)^2-4c^2x^2}}{2xy} \text{ if } a<x<M.$$

Hence, by the *proof* of Theorem 5.1,

$$(y=)\,h(x)=(\frac{b}{a})\sqrt{x^2-a^2} \text{ whenever } a<x<M.$$

So, each value of y in the domain of λ_1 satisfies $y<(b/a)\sqrt{M^2-a^2}$. Since the domain of $\lambda_1|_{(0,\infty)}$ is $(0,\infty)$, we have found the desired contradiction. This completes the proof of the claim that $\lim_{y\to\infty}\lambda(y)=\infty$; equivalently, that the range of λ_1 is $[a,\infty)$; equivalently, that the domain of h is $[a,\infty)$.

Under the still-prevailing assumption that (4) holds at each point on γ whose second coordinate is nonzero, we have, by the mildest tweaking of the reasoning in the preceding paragraph, that

$$(y=)\,h(x)=(\frac{b}{a})\sqrt{x^2-a^2} \text{ whenever } x\in(a,\infty).$$

Equivalently [although we will choose not to use this next fact for a while],

$$\lambda_1(y)=x=\sqrt{x^2}=(\frac{a}{b})\sqrt{y^2+b^2} \text{ whenever } y>0.$$

The first equation in this paragraph also holds at $x=a$ since $h(a)=0$ (that is, since $\lambda_1(0)=\lambda(0)=a$). It follows that h is continuous (on its domain, $[a,\infty)$). Also, more mild tweaking of the reasoning in the preceding paragraph shows that h is diffferentiable on (a,∞), with

$$h^{'}(x)=\frac{c^2-x^2+y^2+\sqrt{(x^2+y^2+c^2)^2-4c^2x^2}}{2xy} \text{ if } x\in(a,\infty);$$

and that if $y > 0$, then

$$\frac{dx}{dy} = \lambda'(y) = \lambda_1'(y) = \frac{2xy}{c^2 - x^2 + y^2 + \sqrt{(x^2+y^2+c^2)^2 - 4c^2x^2}}.$$

Next, working in the extended real number system, we get

$$\lim_{y\to 0^+} \lambda'(y) = \lim_{y\to 0^+} \lambda_1'(y) =$$

$$\lim_{y\to 0^+} \frac{2xy}{c^2 - x^2 + y^2 + \sqrt{(x^2+y^2+c^2)^2 - 4c^2x^2}} =$$

$$\frac{2a\cdot 0^+}{c^2 - a^2 + 0^2 + \sqrt{(a^2+0^2+c^2)^2 - 4c^2a^2}} =$$

$$\frac{0^+}{c^2 - a^2 + |a^2 - c^2|} = \frac{0^+}{2(c^2-a^2)} = 0^+;$$

that is, to use a common turn of phrase, this limit is reached as "0 through positive values". (Since λ_1 is continuous, one interesting consequence is that λ_1' is continuous at $y = 0$. Indeed, when one combines the definition of $\lambda_1'(0)$ with the formulation of L'Hôpital's Rule in [12], one now gets easily that $\lambda_1'(0) = \lim_{\delta\to 0^+} \lambda_1'(\delta) = 0$.) Therefore, by combining the Inverse Function Theorem and limit theorems,

$$\lim_{x\to a^+} h'(x) = \frac{1}{\lim_{y\to 0^+} \lambda'(y)} = \frac{1}{0^+} = \infty.$$

As we proved above that h is continuous, it follows that the graph of h has a vertical tangent line at the point $(a,0)$; that is, γ has a vertical tangent line at $(a,0)$.

Next, continuing to work under the above prevailing assumption, consider the function $\lambda_2 := \lambda|_{(-\infty,0)} : (-\infty,0) \to [a,\infty)$. By slightly adjusting the above reasoning to fit the present context, one can prove all of the following information. The function λ_2 is strictly decreasing and, hence, has an inverse function. Let k denote that inverse function. Of course, the range of k is $(-\infty,0]$. Moreover, the domain of k is $[a,\infty)$; equivalently, the range of λ_2 is $[a,\infty)$. The earlier formula for $\lambda_1'(y)$ (for $y > 0$) also holds for $\lambda_2'(y)$ (when predicated for $y < 0$). That formula, together with the the Intermediate Value Theorem and the facts that λ_2 is a strictly decreasing continuous function with $\lambda_2(0) = a$, can be used to prove that $\lim_{y\to-\infty} \lambda(y) = \lim_{y\to-\infty} \lambda_2(y) = \infty$. The earlier formula for $h'(x)$ (for $x > a$) also holds for $k'(x)$ (for $x > a$). As in the proof of Theorem 5.1, $y := k(x)$ satisfies $x^2/a^2 - y^2/b^2 = 1$. It follows that

$$k(x) = y = -\sqrt{y^2} = -(\frac{b}{a})\sqrt{x^2 - a^2} \text{ whenever } x \in (a,\infty), \text{ and}$$

$$\lambda_2(y) = x = \sqrt{x^2} = (\frac{a}{b})\sqrt{y^2 + b^2} \text{ whenever } y < 0.$$

The earlier formula for $\lambda_1'(y)$ (for $y > 0$) also holds for $\lambda_2'(y)$ (for $y < 0$). Working in the extended real number system, one can prove that

$$\lim_{y\to 0^-} \lambda'(y) = \lim_{y\to 0^-} \lambda_2'(y) = \frac{0^-}{2(c^2-a^2)} = 0^-;$$

that is, this limit of 0 is "reached through negative values". It then follows from the Inverse Function Theorem and limit theorems that

$$\lim_{x\to a^+} k'(x) = \frac{1}{\lim_{y\to 0^-} \lambda_2'(y)} = \frac{1}{0^-} = -\infty.$$

Thus, the graph of k has a vertical tangent line at the point $(a,0)$ (and we see once again that γ has a vertical tangent line at $(a,0)$). Also, $\lambda_2^{'}$ is continuous at $y=0$, with $\lim_{\delta\to 0^-}\lambda_2^{'}(\delta)=0$.

Continuing to work under the above prevailing assumption, we next show that each of the conditions (3)-(8) holds at the (vertex) point $Q=V(a,0)$. This can be done essentially as in the fourth paragraph of the proof of Theorem 5.1. (In detail: the above proofs that γ has a vertical tangent line at V allow us to take the tangential vector $\mathcal{U}_V$ at V to be $\mathbf{j}$. Then one uses the Principle of Reflection to show that the six conditions (3)-(8) are all satisfied when $Q=V$, by combining the fact that $\overrightarrow{VF_1}$ and $\overrightarrow{VF_2}$ are horizontal vectors with the principle that all right angles are congruent). Next, recall that the hypotheses ensured that $\lambda(y)=a$ if and only if $y=0$; and, of course, the point $(a,0)$ is on the hyperbola $\mathcal{H}$. Therefore, for the rest of this proof, as we consider the behavior of points $Q(x,y)$ on γ, we can assume that $Q\neq V$, that is, that $x>a$ (and so $y\neq 0$).

We next apply our customary combination of vectorial reasoning and the Principle of Reflection (along with the fact that $\cos|_{(0,\pi)}$ is a one-to-one function). This combination can be used to show that the conditions (3), (4), (5), (6), (7), and (8) are equivalent. Moreover, (1) $\Rightarrow$ (2) because of the classical Cartesian equation for $\mathcal{H}$; and (2) $\Rightarrow$ (3) by Theorem 4.1. Next, note what we have called “the above prevailing assumption" was simply the assumption that condition (4) holds for all points Q other than V on γ. Therefore, the reasoning which ended two paragraphs ago (which, *inter alia*, gave formulas for $\lambda_1(y)$ and $\lambda_2(y)$) has served to establish that (4) $\Rightarrow$ (1). This completes the proof. □

Remark 5.5. (a) It would be reasonable to maintain that a wholeheartedly “x as a function of y" approach to a proof of Theorem 5.4 would have included an independent solution of the ODE involving $\lambda^{'}(y)$ (at least for $y>0$). Instead, the above proof of Theorem 5.4 appealed to the Inverse Function Theorem (for real-valued functions of one real variable) and to the solution of the ODE which figured in the proof of Theorem 5.1. Any readers who would wish to have a proof of Theorem 5.4 that avoids those two appeals are invited to devise such a proof. For that enterprise, I suggest that it would to helpful to use the variables w and v (or other variables closely related to them) which figured in the solution of the ODE in the proof of Theorem 5.1 (and in the solutions of other ODEs in the proofs of Theorem 3.1 and [3, Theorem 3.1]).

(b) As an alternative to Remark 5.3 (f), we next observe that one can combine the method of proof of Corollary 5.2 with Theorem 5.4 in order to obtain reflection-theoretic characterizations of the left-hand branch of an east-west hyperbola $\mathcal{H}$ that is in standard position. Indeed, the changes of variable $(x,y)\leftrightarrow(t,Y):=(-x,y)$ can be used to convert the characterizations of the right-hand branch of $\mathcal{H}$ in Theorem 5.4 into characterizations of the left-hand branch of $\mathcal{H}$. The remark is complete.

Next, we give this section’s strongest reflection-theoretic characterization result. It is a companion for some results on parabolas [3, Remark 3.2 (e) and Appendix] and the above Remark 3.2 (d) about ellipses. The gist of Theorem 5.6 is that any nontrivial hyperbolic arc (no matter how “tiny" it may be) can reveal the unique hyperbola of which it is a subset and a Cartesian equation for that hyperbola. For simplicity (but with no loss of generality), we will address, in the spirit of Theorem 5.1, a (potentially) hyperbolic arc which is a subset of the first quadrant. To facilitate the statement of condition (2) in Theorem 5.6 (a), we will adopt the convention here that if x is a positive real number, then the point $(x,0)$ is considered to be in the first quadrant. As usual, if $\alpha\in\mathbb{R}$ and $\beta=\infty$, it will be convenient to take $[\alpha,\beta]$ (resp., $(\alpha,\beta]$; resp., $x\leq\beta$) to mean $[\alpha,\infty)$ (resp., (α,∞); resp., $x<\infty$).

Theorem 5.6. Let $0<a<c$ in $\mathbb{R}$, and put $b:=\sqrt{c^2-a^2}$. Let $a\leq\alpha<\beta\leq\infty$. Working in a fixed Euclidean plane, consider the points $F_1(-c,0)$ and $F_2(c,0)$. Let $f:[\alpha,\beta]\to\mathbb{R}$ be a function, and let Γ be the graph of f. Suppose that f is strictly increasing on $[\alpha,\beta)$ and differentiable on (α,β), $f^{'}(x)\neq 0$ for all $x\in(\alpha,\beta)$, f is continuous at $x=\alpha$ (and at $x=\beta$ if $\beta\in\mathbb{R}$), and $f(t_1)>0$ for some $t_1\in(\alpha,\beta)$. Suppose also that F_2 is not on Γ and that Γ has a tangent line at $x=\alpha$. For each point P on Γ, let $\mathcal{T} := \mathcal{T}_P$ be a tangential vector to Γ at P. Let $\mathcal{H}$ be the hyperbola with foci F_1 and F_2 and with semitransverse axis a (and necessarily with semi-conjugate axis b). Then:

(a) The following two conditions are equivalent:

(1) For each point P on Γ,

$$\frac{\mathcal{T}_P \cdot \overrightarrow{PF_1}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_1}|} = \frac{\mathcal{T}_P \cdot \overrightarrow{PF_2}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_2}|};$$

(2) Γ is a (necessarily connected) subset of a (necessarily uniquely determined) hyperbola $\mathfrak{H}$ with foci F_1 and F_2 (and, necessarily, Γ is a subset of the first quadrant).

(b) Suppose that the equivalent conditions in (a) hold. Then $\mathfrak{H}$ is a hyperbola with foci F_1 and F_2 whose right-hand x-intercept a^* satisfies $0 < a^* \le \alpha$ and

$$f(x) = (\frac{\sqrt{c^2 - (a^*)^2}}{(a^*)^2})\sqrt{x^2 - (a^*)^2} \text{ for all } x \text{ such that } \alpha \le x \le \beta.$$

(c) Suppose that the equivalent conditions in (a) hold, with $\mathfrak{H}$ and a^* as in (a) and (b). Then the following three conditions are equivalent:

(i) $\mathfrak{H} = \mathcal{H}$;

(ii) $a^* = a$;

(iii) There exists a point Q on Γ such that $d_1(Q) - d_2(Q) = 2a$.

Proof. (a) (2) $\Rightarrow$ (1): As explained in the fifth paragraph of the proof of Theorem 5.1, this implication follows from the proof of Theorem 4.1.

(1) $\Rightarrow$ (2): Assume (1). Let us next consider points $P(x,y)$ on Γ such that $\alpha < x < \beta$. We claim that it follows, as in the proof of Theorem 5.1, that there exists a constant $E > 0$ such that

$$y^2 + c^2 + x^2 + \sqrt{(y^2 + c^2 + x^2)^2 - 4c^2x^2} = Ex^2 \text{ (if } \alpha < x < \beta).$$

To prove this claim, note that the just-referenced part of the proof of Theorem 5.1 used (indefinite) integration over a closed subinterval of (α, β) of finite length. So, that earlier material does prove that if $\alpha < \beta_1 < \beta_2 < \beta$, then there exist constants $E_1 > 0$ and $E_2 > 0$ such that the last-displayed equation holds for $E = E_1$ (resp., for $E = E_2$) and all x such that $\alpha < x < \beta_1$ (resp., and all x such that $\alpha < x < \beta_2$). Since $x^* := (\alpha + \beta_1)/2$ satisfies $\alpha < x^* < \beta_1$ and $\alpha < x^* < \beta_2$, we get $E_1(x^*)^2 = E_2(x^*)^2$. As $(x^*)^2 \neq 0$, it follows that $E_1 = E_2$, and the claim now follows easily.

We have

$$y^2 + c^2 + (1-E)x^2 = -\sqrt{(y^2 + c^2 + x^2)^2 - 4c^2x^2} \text{ if } \alpha < x < \beta.$$

Taking x to be the above x^* shows that $E \neq 1$ (for otherwise, we would have the contradiction that the left-hand side of the just-displayed equation is positive while right-hand side of that equation is not positive). Next, squaring both sides of the last-displayed equation leads, after some algebraic rewriting, to

$$(1-E^2)x^4 + 2(1-E)x^2y^2 + 2y^2c^2 + 2(1-E)c^2x^2 =$$
$$x^4 + 2x^2y^2 + 2y^2c^2 + 2c^2x^2 - 4c^2x^2 \text{ if } \alpha < x < \beta.$$

Rewriting and simplifying the just-displayed equation gives (with some effort) the following equivalent equation:

$$-2Ex^2y^2 = E(2-E)x^4 + (2E-4)c^2x^2 \text{ if } \alpha < x < \beta.$$

As each of the relevant values of x is nonzero, dividing through by x^2 gives the following equivalent equation:

$$-2Ey^2 = E(2-E)x^2 + 2(E-2)c^2 \text{ if } \alpha < x < \beta.$$

Consequently $E \neq 2$, for otherwise, we would have $-2Ey^2 = 0$, whence $y^2 = 0$ for all $x \in (\alpha, \beta)$, which is a contradiction since f is not an identically zero function. (We could also have obtained a contradiction from the hypothesized existence and behavior of t_1.)

Next comes the key symbolic step in this proof. By taking

$$A := \frac{2c^2}{E} \text{ and } B := \frac{(E-2)c^2}{E},$$

we can rewrite the last displayed equation in the last paragraph as

$$\frac{x^2}{A} - \frac{y^2}{B} = 1 \text{ for all } x \text{ such that } \alpha < x < \beta.$$

Moreover, since f is assumed continuous at α (and at β, if $\beta \in \mathbb{R}$), the just-displayed equation also holds if $x = \alpha$ (and at $x = \beta$, if $\beta \in \mathbb{R}$). Also, observe that $A > 0$, $B \neq 0$ and $A + B = c^2$. So, if $B > 0$, the last display would give (2). Therefore, to get (2), we need only obtain a contradiction from the supposition that $B < 0$.

Suppose that $B < 0$. It follows that $0 < E < 2$. Moreover, by considering the positive real numbers

$$a^* := \sqrt{A} = \sqrt{\frac{2}{E}}c \text{ and } b^\diamond := \sqrt{-B} = \sqrt{(\frac{2-E}{E})}c,$$

we can rewrite the last displayed equation in the preceding paragraph as

$$\frac{x^2}{(a^*)^2} + \frac{y^2}{(b^\diamond)^2} = 1 \text{ for all } x \text{ such that } \alpha < x < \beta.$$

The last-displayed equation is (if one temporarily ignores its restrictions on x) that of an east-west ellipse, say $\mathcal{E}$, with center at the origin, semi-major axis a^* and semi-minor axis $b^\diamond$. Put

$$c^* := \sqrt{(a^*)^2 - (b^\diamond)^2}.$$

The foci of the ellipse $\mathcal{E}$ are the points $(-c^*, 0)$ and $(c^*, 0)$. Also, since

$$(c^*)^2 = (a^*)^2 - (b^\diamond)^2 = A - (-B) = A + B = c^2,$$

we have $(c^*)^2 = c^2$, whence $c^* = c$. Therefore, the foci of the ellipse $\mathcal{E}$ are F_1 and F_2. Next, consider any point $P(x, y)$ on Γ such that $\alpha < x < \beta$. As P is on $\mathcal{E}$, it follows from the seventh paragraph of Section 2 (see also the first sentence of the proof of Theorem 2.1) that

$$\frac{\mathcal{T}_P \cdot \overrightarrow{PF_1}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_1}|} = -\frac{\mathcal{T}_P \cdot \overrightarrow{PF_2}}{|\mathcal{T}_P| \cdot |\overrightarrow{PF_2}|}.$$

In conjunction with the hypothesis that (1) holds, the just-displayed equation implies that

$$\mathcal{T}_P \cdot \overrightarrow{PF_1} = 0 = \mathcal{T}_P \cdot \overrightarrow{PF_2}.$$

In other words, the nonzero vectors $\overrightarrow{PF_1}$ and $\overrightarrow{PF_1}$ are each perpendicular to the (nonzero tangential) vector $\mathcal{T}_P$. Hence, by fundamental principles of Euclidean geometry, $\overrightarrow{PF_1}$ and $\overrightarrow{PF_1}$ are parallel vectors such that the points P, F_1 and F_2 are collinear. Therefore, the point P, with coordinates $(x, f(x))$, is on the x-axis for *every* x such that $\alpha < x < \beta$. Since f was assumed to be strictly increasing on $[\alpha, \beta)$, we have found the desired contradiction. (We could also have obtained a contradiction in the following two other ways: from the hypothesis that $f'(x) \neq 0$, together with an appeal to the Mean Value Theorem; or from the hypothesized existence and behavior of t_1.) This completes the proof that (1) implies the non-parenthetical part of the statement of (2).

It remains to justify the three parenthetical pieces of the statement of condition (2). The first of these pieces is easily explained, as any continuous image of a connected topological space must be connected. We turn next to the justification of the second piece (namely, that Γ cannot be a subset of two distinct hyperbolas each of which has foci F_1 and F_2). In fact, more is true. To wit: if $\mathfrak{H}_1$ and $\mathfrak{H}_2$ are hyperbolas with the same foci $F_1(-c,0)$ and $F_2(c,0)$ and there exists a point Q that is on both $\mathfrak{H}_1$ and $\mathfrak{H}_2$, then $\mathfrak{H}_1 = \mathfrak{H}_2$. To prove this fact, let $a_1^\diamond$ (resp., $a_2^\diamond$) be the semitransverse axis of $\mathfrak{H}_1$ (resp., of $\mathfrak{H}_2$). Since Q is on $\mathfrak{H}_1$ (resp., on $\mathfrak{H}_2$),

$$|d_1(Q) - d_2(Q)| = 2a_1^\diamond \text{ (resp., } |d_1(Q) - d_2(Q)| = 2a_2^\diamond),$$

whence $2a_1^\diamond = 2a_2^\diamond$, and so $a_1^\diamond = a_2^\diamond$. Let $b_1^\diamond$ (resp., $b_2^\diamond$) be the semi-conjugate axis of $\mathfrak{H}_1$ (resp., of $\mathfrak{H}_2$). Then

$$b_1^\diamond = \sqrt{(a_1^\diamond)^2 - c^2} = \sqrt{(a_2^\diamond)^2 - c^2} = b_2^\diamond.$$

So, $\mathfrak{H}_1$ and $\mathfrak{H}_2$ have the same Cartesian equation, $x^2/(a^\diamond)^2 - y^2/(b^\diamond)^2 = 1$, where $a^\diamond := a_1^\diamond$ $(= a_2^\diamond)$ and $b^\diamond := b_1^\diamond$ $(= b_2^\diamond)$. Consequently, $\mathfrak{H}_1 = \mathfrak{H}_2$, as asserted.

It remains to justify the third parenthetical piece of the statement of (2) (namely, that Γ is a subset of the first quadrant). Recall from (2) that the graph of f (namely, Γ) is a subset of a hyperbola $\mathfrak{H}$ which has foci $F_1(-c,0)$ and $F_2(c,0)$, with $c > 0$. Since the domain of f is $[\alpha,\beta]$ and $\alpha > 0$, our knowledge of the graphs of east-west hyperbolas such as $\mathfrak{H}$ ensures that Γ is a subset of the right-hand branch of $\mathfrak{H}$, whence Γ is a subset of the union of the first quadrant and the fourth quadrant. However, the hypothesis that f is strictly increasing on $[\alpha,\beta)$ implies (again, because of our knowledge of the graphs of hyperbolas such as $\mathfrak{H}$) that the only possible fourth-quadrant point on Γ is $(\alpha,0)$, and a notational convention announced prior to the statement of the present result that any point with such coordinates is to be considered in the first quadrant. Therefore, Γ is a subset of the first quadrant. This completes the proof that (1) implies (all of) (2). This completes the proof of (a).

(b) Let $\mathfrak{H}$ be the hyperbola satisfying condition (2) in (a). Let a^* be the semitransverse axis of $\mathfrak{H}$. The assertion follows by combining the following six facts: the foci of $\mathfrak{H}$ are $(-c,0)$ and $(c,0)$ (with $c > 0$); a Cartesian equation of $\mathfrak{H}$ is

$$\frac{x^2}{(a^*)^2} - \frac{y^2}{c^2 - (a^*)^2} = 1;$$

$\mathfrak{H}$ is the only such hyperbola containing Γ as a subset; Γ is a subset of the first quadrant; the right-hand x-intercept of $\mathfrak{H}$ is a^*; and Γ is the graph of a function f whose domain is $[\alpha,\beta]$ where $a \leq \alpha < \beta$ $(\leq \infty)$.

(c) Note that $\mathfrak{H}$ and $\mathcal{H}$ are each hyperbolas with the foci $F_1(-c,0)$ and $F_2(c,0)$. (In other words, these foci are determined by the same value of the parameter $c > 0$ for both $\mathfrak{H}$ and $\mathcal{H}$.) Recall that in the Euclidean plane $\mathbb{R}^2$, any hyperbola H with this same pair of foci and semitransverse axis some positive real number $a^\diamond$ is given by

$$\mathrm{H} = \{P \in \mathbb{R}^2 \mid |d_1(P) - d_2(P)| = 2a^\diamond\}.$$

Thus, it is clear that for a fixed value of c, the correspondence described by $H \leftrightarrow a^\diamond$ is a bijection. The equivalence (i) $\Leftrightarrow$ (ii) is now immediate.

(iii) $\Rightarrow$ (i). Assume (iii). By the preceding paragraph, Q is on $\mathcal{H}$. However Q is also on Γ and, by condition (2), $\Gamma \subset \mathfrak{H}$. As Γ is nonempty, so is $\mathfrak{H} \cap \mathcal{H}$. Therefore, by the above proof of the second parenthetical piece of the statement of condition (2), $\mathfrak{H} = \mathcal{H}$, as desired.

(i) $\Rightarrow$ (iii): Assume (i); that is, $\mathfrak{H} = \mathcal{H}$. Pick $x \in (\alpha, \beta)$, and consider the point $Q(x, f(x))$. Note that Q is on Γ and (once again) that $\Gamma \subset \mathfrak{H}$. Hence, Q is on $\mathfrak{H}$. Hence, Q is on $\mathcal{H}$. So, by the first paragraph of this proof, Q satisfies $|d_1(P) - d_2(P)| = 2a$, as desired. The proof is complete. □

We close with three remarks. Remark 5.7 collects some comments stimulated by the proof of Theorem 5.6. Remark 5.8 identifies some degenerate cases of hyperbolas, in the spirit of Remark 3.2 (f)-(g) and [3, Remark 3.2 (g)]. Remark 5.9 provides some final reflections.

Remark 5.7. (a) The list of hypotheses in the statement of Theorem 5.6 was intentionally redundant. In detail, consider the following four hypotheses in that result: Γ has a tangent line at $x = \alpha$; F_2 is not on Γ; f is strictly increasing on $[\alpha, \beta)$; and $f(t_1) > 0$ for some $t_1 \in (\alpha, \beta)$. The first two of these hypotheses were included out of an abundance of caution. Indeed, the proof used a reflection-theoretic equation involving dot products and tangential vectors that was predicated for an *arbitrary* point P of Γ, and that equation would be meaningless if the tangent line does not exist at a point P on Γ if either the x-coordinate of P is α or $P = F_2$. Even though the actual proof of Theorem 5.6 did not use that equation in either of those two cases, those two hypotheses were included in order to lighten the burden of readers. I have found that reliable authors sometimes include redundant hypotheses for a similar reason, while tacitly expressing confidence that an able reader could be expected to sharpen the statement of the result (via his/her understanding of the offered proof) if and when the need to do so may arise in the future. (Besides, removing either of those two hypotheses would have created fewer parallels between the statement of Theorem 5.6 and the statements of some earlier named results, reestablishing those parallels would have then required clumsier restatements of those earlier results, and I believe that each of those outcomes would have added to a reader's burden.) The third of the above-mentioned hypotheses is *not redundant*, as it was used twice in the proof of Theorem 5.6; to be more specific, it was used at the end of the fifth paragraph of the proof of Theorem 5.6 and, in order to prove that Γ is a subset of the first quadrant, it was also used at the end of the proof of Theorem 5.6 (a). The fourth of the above hypotheses *is* logically redundant, as its only use in the proof of Theorem 5.6 was as an alternative to the above-mentioned first use of the third hypothesis. I would defend the inclusion of the fourth hypothesis because an individual seeking to apply Theorem 5.6 may find it easier to check that his/her data admit a point t_1 with the asserted behavior rather than having to check that a function under consideration is strictly increasing over an entire interval. Any readers who would prefer to have a sharper statement of Theorem 5.6 are hereby invited to delete its hypothesis concerning t_1.

On the other hand, one may wonder whether the list of hypotheses that was stated in Remark 3.2 (d) (which was a result about certain elliptic arcs) is incomplete. Here are two possible causes for concern in that regard: that list did not explicitly state that neither F_1 nor F_2 is on Γ; and that list did not explicitly state that if $\beta = a$, then there exists a tangent line to Γ at the point $(a, 0)$. These concerns are possibly only heightened by the following two facts: the statement of the relevant result in Remark 3.2 (d) used a hypothesis about a tangential vector to Γ at an *arbitrary* point P on Γ and that hypothesis involved an equation (which encoded a reflection-theoretic property) that would be meaningless if the point P under consideration were either F_1 or F_2. I would address (I do not say "counter" here) these concerns as follows. The actual *proof* that was given for that result in Remark 3.2 (d) *did not use* the equation in question if the x-coordinate of P is the endpoint at issue, and so the possible existence of a tangent line to Γ at that point was irrelevant to that proof. These observations reflect my view that one can occasionally omit a hypothesis in stating a remark (as opposed to a named theorem, corollary, etc.) if a reader can be reasonably assumed to infer that unstated hypothesis from the assumed meaningfulness of another (stated) hypothesis. I understand that some readers may object to what they may regard as a lax attitude here toward rigor in remarks (in distinction to my defense, in the preceding paragraph, of the occasional use of redundant hypotheses in the statement of named/numbered results). As I have said before, there is nothing to be gained (except, possibly, good will) in arguing about matters of taste or style. So, any readers with concerns of the kind which I just addressed are invited to add the two just-mentioned hypotheses to the list of assumptions for the relevant result in Remark 3.2 (d).

In comparing the statement of Theorem 5.6 with the statements of some of the other results here that presented a list of equivalent conditions, one is struck by how relatively few equivalent conditions are listed in the statement of Theorem 5.6 (a). That discrepancy could be easily remedied. Indeed, if one wished to emulate the statements of (implicitly or explicitly) characterization results such as Theorem 3.1, Remark 3.2 (b), Theorem 5.1, Corollary 5.2 and Theorem 5.4, one could augment the statement of Theorem 5.6 (a) with a huge list of variants, that is, of additional conditions which are equivalent to conditions (1) and (2) in Theorem 5.6 (a). To be more specific: one could add versions of condition (3) and conditions (5)-(8) from Theorem 5.1 and Corollary 5.2, as our usual vectorial methods (combined with the Principle of Reflection and the strictly monotonic behavior of the function $\cos|_{[0,\pi]}$) would show that each of those versions is equivalent to condition (1) in Theorem 5.6 (a). One could then essentially double the list of equivalent conditions by adding variants obtained by revising the just-mentioned versions by deleting any mention of any unused assumptions about the existence or behavior of tangent lines at certain points on the graph of the function f corresponding to certain endpoints of the domain of f. Moreover, one could also alter the statement of Theorem 5.6 by varying its list of hypotheses. For instance, we explained two paragraphs ago that the hypothesis concerning t_1 is redundant. The list of equivalent conditions in such an enhancement of the statement of Theorem 5.6 (a) would be much longer than the lists of 17 equivalent conditions that were mentioned in parts (b) and (d) of Remark 5.3. In such situations, an author and his/her readers must decide whether a rather long list of equivalent conditions (or of variant results) would be of greater use (for theoretical and/or applied purposes, at once or eventually) than a shorter list. My decision in this regard for Theorem 5.6 can be seen above. Any interested readers are invited to use the above comments to formulate variants of the above statement of Theorem 5.6.

Despite what may have seemed like a minimalist's comment at the end of the preceding paragraph, I would like to add the following. By providing different formulations/variants of his/her mathematical results, an author allows for the possibility that some reader(s) may perceive opportunities for applications which the author had not foreseen (perhaps because those applications are, frankly, outside the area of expertise of the author). Of course, I am referring here to “applications" in the broadest possible sense: not only to mathematics *per se*, but also to other socially beneficent areas, such as engineering, architecture, medicine, data transfer, etc.). It would be wonderful if some enterprising scientist could devise some new real-world ways to use some of the degenerate conic sections that are studied in this paper and in [3] in connection with reflection-theoretic properties.

(b) A constant denoted by $E > 0$ has played important roles in the proofs of Theorems 5.1 and 5.6 (a). Recall from the proof of Theorem 5.1 that the value of E was shown to be

$$E = 2 + \frac{2b^2}{a^2} \left(= \frac{2(a^2+b^2)}{c^2} = \frac{2c^2}{a^2}\right)$$

if condition (4) of Theorem 5.1 and the hypotheses of Theorem 5.1 all hold. Those hypotheses included $0 < a < c$ in $\mathbb{R}$, $b := \sqrt{c^2 - a^2}$, and certain conditions on a given function $f : [a, \infty) \to \mathbb{R}$. Consider the following: Theorem 5.6 is plainly a generalization of Theorem 5.1; and the hypotheses of Theorem 5.6 included the same restrictions on a and c (and the same definition of b) as in Theorem 5.1, along with a (different) given function $f : [\alpha, \beta] \to \mathbb{R}$, where $a \leq \alpha < \beta \leq \infty$. Furthermore, Theorem 5.6 featured a condition (1) which was clearly formulated by adapting the statement of condition (4) from Theorem 5.1 to the domain $[\alpha, \beta]$ in Theorem 5.6. Accordingly, it seems natural to ask the following two-part question. If one assumes condition (1) in Theorem 5.6, can one obtain a formula for the constant E (appearing in Theorem 5.6) that generalizes the last displayed formula for the constant E (appearing in Theorem 5.1) that was obtained in Theorem 5.1; and, if so, is the resulting formula for the constant E (from Theorem 5.6, under the assumption of condition (1)) more complicated because of some logical need to accommodate the role of α, β or a^*? We next answer this two-part question as follows. If condition (1) in Theorem 5.6 (a) holds, then the general formula for the constant E appearing in the proof of Theorem 5.6 (a) is $E = 2c^2/(a^*)^2$, where a^* is the semitransverse axis of the hyperbola $\mathfrak{H}$ appearing in Theorem 5.6 (b).

Our proof of the above formula will use notation from Theorem 5.6 (a)-(c) (along with the hypothesis that condition (1) holds). By Theorem 5.6 (a), we also have condition (2), and so $\Gamma \subset \mathfrak{H}$. One can thus use the familiar Cartesian equation for $\mathfrak{H}$ to get a formula for $y = f(x)$ for all $x \in (\alpha, \beta)$. As f is continuous at α, it follows that

$$f(\alpha) = \lim_{x \to \alpha^+} f(x) = (\frac{\sqrt{c^2 - (a^*)^2}}{a^*})\sqrt{\alpha^2 - (a^*)^2}.$$

(Note that if $\alpha = a = a^*$, then the just-displayed equation reduces to $f(a) = 0$.) Also, recall from the proof of Theorem 5.6 (a) that the assumption of condition (1), along with the other hypotheses in Theorem 5.6, ensures that $y = f(x)$ satisfies

$$y^2 + c^2 + x^2 + \sqrt{(y^2 + c^2 + x^2)^2 - 4c^2x^2} = Ex^2 \text{ (if } \alpha < x < \beta).$$

With the above formula for $f(\alpha)$ in hand, perform enough algebraic simplification to confirm that

$$f(\alpha)^2 + c^2 + \alpha^2 = \frac{c^2\alpha^2}{(a^*)^2} + (a^*)^2.$$

Next, apply the operator $\lim_{x \to \alpha^+}$ to the second displayed equation in this paragraph, substitute the last displayed equation, and perform enough algebraic simplification to confirm that

$$\frac{c^2\alpha^2}{(a^*)^2} + (a^*)^2 + \sqrt{[\frac{c^2\alpha^2}{(a^*)^2} - (a^*)^2]^2} = E\alpha^2.$$

Next, use $0 < a^* \leq \alpha$ and $a^* < c$ to get $c^2\alpha^2/(a^*)^2 - (a^*)^2 > 0$. It follows that the last displayed equation can be rewritten as

$$\frac{c^2\alpha^2}{(a^*)^2} + (a^*)^2 + [\frac{c^2\alpha^2}{(a^*)^2} - (a^*)^2] = E\alpha^2.$$

Then, dividing through by α^2 ($\neq 0$), we easily get $E = 2c^2/(a^*)^2$, completing the proof.

Some readers may find it appropriate to assign the algebraic simplifications in the above argument to their beginning algebra students as homework.

An alternative proof of the result that was established two paragraphs ago may have occurred to some readers. Rather than viewing the assertion $E = 2c^2/(a^*)^2$ as being *motivated by* Theorem 5.1, the alternative approach would view this assertion as being a *corollary of* the proof of Theorem 5.1. To legitimize this alternative approach, one would need to use the hypotheses of Theorem 5.6, involving a certain given function $f : [\alpha, \beta] \to \mathbb{R}$ satisfying certain conditions, to construct a useful function, let us call it $g : [a^*, \infty) \to \mathbb{R}$, that satisfies the hypotheses that were attributed to the function which was called "f" in Theorem 5.1. One may initially think that a likely candidate for such a function g would be the restriction of the function f (which was given in Theorem 5.6) to the domain $[a, \infty)$. This approach may initially seem feasible, but it has the serious drawback that the just-mentioned "restriction" need not exist, the point being that the intended domain, $[a, \infty)$, of this "restriction" need not be a subset of the domain $[\alpha, \beta]$. There are two reasons for this drawback: first, although $a^* \leq \alpha$, there is no reason to believe that $\alpha \leq a$ (we will say more about the underlying issue here in (d) below); and second, $[a, \infty)$ is certainly *not* a subset of $[\alpha, \beta]$ if $\beta \in \mathbb{R}$.

However, the alternative approach can be adjusted, as follows. Since the graph Γ of the function f is a subset of the right-hand branch of the hyperbola $\mathfrak{H}$ and a Cartesian equation $y = h(x)$ for the top half of that branch can easily be given, with h a certain function $[a^*, \infty) \to \mathbb{R}$, it *does* make sense to apply the proof of Theorem 5.1, with a^* playing the earlier role of the endpoint a and with h playing the role of the function that was called "f" in Theorem 5.1. To make the alternative proof complete, one would still have to check that this h satisfies all the hypotheses that were placed on "f" in the statement of Theorem 5.1, and that is very easy to do. Nevertheless, I frankly consider the direct proof that was given three paragraphs ago to be a better proof, primarily because it is more accessible. However, I understand that many mathematicians prefer a pedagogic approach that would advocate for "fewer theorems and more applications". I believe that a compassionate author/educator should try to anticipate (and, if possible, to accommodate) the preferences of an expected audience. So, I have offered the alternate proof here even though the difference between the two proofs is largely a matter of presentation. (In that regard, note that the introduction of the key equation which involved the constant E in the proof of Theorem 5.6 (a) was preceded by the justifying phrase "as in the proof of Theorem 5.1".) In conclusion, the alternative approach has been discussed here for the following two reasons. First, as I just explained, it has been offered as an attempt to satisfy the pedagogic tastes of some individuals. Second, it is intended to validate an experimental approach to creating rigorous mathematics. Indeed, the alternative approach could begin to develop in some minds by reasoning as follows: "The function f in Theorem 5.1 satisfies $f(a) = 0$ and the function h describing the top half of the hyperbola $\mathfrak{H}$ in Theorem 5.6 satisfies $h(a^*) = 0$, so why not try to apply the proof of Theorem 5.1 to something related to the function h, with a^* playing the earlier role of a?". Much good mathematics has resulted from such analogical flights of fanciful thought.

(c) Assume that condition (1) in Theorem 5.6 holds, along with all the hypotheses of Theorem 5.6 (including that one is given $0 < a < c$). Let E be the constant appearing in the proof of Theorem 5.6 (a), let $\mathfrak{H}$ be the hyperbola appearing in condition (2) of Theorem 5.6, and let a^* be the semitransverse axis of $\mathfrak{H}$. It was shown in (b) that under these conditions, one has $E = 2c^2/(a^*)^2$. It follows that the statement of Theorem 5.6 (c) can be enhanced, by adding the following fourth equivalent condition, (iv): If E is the constant appearing in the proof of (a), then $E = 2c^2/a^2$.

(d) Let $c > 0$ in $\mathbb{R}$. Then there exist infinitely many (in fact, uncountably many) hyperbolas whose foci are the points $(-c, 0)$ and $(c, 0)$, as the set of such hyperbolas is in one-to-one correspondence with the open interval $(0, c)$. Indeed, to establish this bijection, one need only associate each real number $a^* \in (0, c)$ with the hyperbola $\mathfrak{H}_{a^*}$ having Cartesian equation

$$\frac{x^2}{(a^*)^2} - \frac{y^2}{c^2 - (a^*)^2} = 1.$$

(A similar "elliptic" observation applies by associating any real number a^{**} in the open interval (c, ∞) with the ellipse $\mathcal{E}_{a^{**}}$ having Cartesian equation

$$\frac{x^2}{(a^{**})^2} + \frac{y^2}{(a^{**})^2 - c^2} = 1.$$

This completes the solution of the "temporary exercise" that was mentioned in Remark 3.2 (d).) The remark is complete.

Remark 5.8. In the spirit of Remark 3.2 (f)-(g) and [3, Remark 3.2 (g)], we next address degenerate cases of hyperbolas. For the sake of brevity, we restrict attention to those cases which arise from a careful examination of the proof of Theorem 5.1 (as those are the cases which can perhaps be said to satisfy the reflection properties of a hyperbola). Essentially all the themes about degenerate ellipses in Remark 3.2 (f)-(g) carry over to the "hyperbolic" context. For the sake of completeness, we will provide most of the details by suitably/simply editing Remark 3.2 (f)-(g) to produce the rest of (a).

Readers interested in identifying the geometric figures which can be viewed as being "degenerate hyperbolas" when one adopts the point of view of [1] are directed to Remark 4.3 (b) (or to [1] itself); see also Remark 4.3 (a).

The proof of Theorem 5.1 involved a differential equation that expressed the derivative of y ($= f(x)$) with respect to x as a fraction with denominator $D = 2xy$. That kind of fraction was meaningful since the context for that part of the proof of Theorem 5.1 involved $x > a$ (> 0) and the hypotheses of Theorem 5.1 then ensured that ($x \neq \pm a$ implies) $y \neq 0$. To cast the question of characterizing the curves satisfying the reflection properties of a hyperbola more generally than in the hypotheses of Theorem 5.1, recall that those reflection properties are described by condition (4) in the statement of Theorem 5.1. Let us begin with a more general attempt to characterize the functions f with domain a subset of $[a, \infty)$ and with graph satisfying the reflection properties of a hyperbola, given $0 < a < c \in \mathbb{R}$ and points $F_1(-c, 0)$ and $F_2(c, 0)$, by asking the following: which differentiable functions f, having domain a subset of $[a, \infty)$, satisfy the above-mentioned condition (4)?

For (4) to be meaningful, it must be the case that neither F_1 nor F_2 is on the graph of f. Our search here for degenerate cases of a hyperbola that were not found in Theorem 5.1 will focus on such functions f for which the above-mentioned differential equation is meaningless because $D = 0$. As f is differentiable, it seems reasonable to assume that the domain of f is a union of (possibly denumerably many) open subintervals of (a, ∞), together with possibly some left-hand endpoints and/or some right-hand endpoints. It would also seem reasonable to assume that f is continuous at any such endpoint. So, we will focus on certain values of x such that $x > a$. As $D = 2xy$ and we are now requiring $D = 0$ with a focus on certain x such that $x > 0$, it must be the case that $y = 0$, that is, $f(x) = 0$. Recall that the proof of Theorem 5.1 derived the above-mentioned differential equation from the following application of condition (4):

$$\frac{1(-c-x)+f'(x)(-y)}{\sqrt{(c+x)^2+y^2}} = \frac{1(c-x)+f'(x)(-y)}{\sqrt{(c-x)^2+y^2}}.$$

Let us cast our nets more widely. Working in conjunction with the conditions $x > a$ (> 0) in $\mathbb{R}$ and $y = 0$, we see that the just-displayed equation is equivalent to the following algebraic equation:

$$\frac{-c-x}{|c+x|} = \frac{c-x}{|c-x|}.$$

A straightforward case analysis shows that the solution set of the just-displayed equation, under the just-stated conditions, consists of the points (x, y) such that $x > c$ and $y = 0$. It follows that if we focus on the universe $[a, \infty)$ for values of x, we can construct infinitely many degenerate cases of hyperbolas. Indeed, each of the following graphs Γ is of that kind: take any (possibly denumerable) nonempty set $\{I_j\}$ of open intervals I_j contained in (c, ∞); for each j, let I_j^* result from I_j by possibly appending one or both endpoints of I_j to I_j, but do not append the element c to any I_j; let $\mathcal{D} := \cup_j I_j^*$; let $\mathcal{D}^*$ result from appending c to $\mathcal{D}$ if there exists j such that c is the left-hand endpoint of I_j; and then take $\Gamma := \{(x, y) \in \mathbb{R}^2 \mid x \in \mathcal{D}^* \text{ and } y = 0\}$.

A degenerate hyperbola Γ which is constructed in the above way need not be a connected topological space – indeed, it may have infinitely many connected components. In particular, while the literature often refers to some degenerate hyperbolas as being "piecewise-linear", observe that some of the examples Γ that we have just constructed have stretched the meaning of that term because they have infinitely many "pieces." Note also that the largest (in the obvious sense) interval that was constructed above as being a degenerate hyperbola is $[c, \infty)$.

Having "cast our nets more widely", the above work naturally draws our attention to certain values of x such that $x < -c$ (< 0). One sees easily that for a nontrivial line segment γ containing the point $P(x, 0)$ for such a value of x, P satisfies the above-mentioned condition (4). Indeed, the tangential vector $\mathcal{T}$ of γ at P can be taken as $\pm\mathbf{i}$, and $\mathcal{T}$ is parallel to, and has the same direction as, both $\overrightarrow{PF_1}$ and $\overrightarrow{PF_2}$. (Of course, the same comment could have been made about the points $(x,0)$ that arose as elements of the graphs Γ that were constructed two paragraphs ago.) Hence, one can produce even more degenerate hyperbolas (although none that are qualitatively new in any mathematically important way) by revising the above directions for constructing degenerate hyperbolas Γ as follows: change the requirement "I_j contained in (c,∞)" to "I_j contained in $(-\infty,-c)\cup(c,\infty)$"; change "but do not append the element c to any I_j" to "but do not append either of the elements $-c$, c to any I_j; and change the definition of $\mathcal{D}^*$ to $\mathcal{D}^* := \mathcal{D}$.

It is natural to seek degenerate cases of hyperbolas from the "x as a function of y" point of view by, for instance, closely examining the proof of Theorem 5.4 (or an alternative proof that was suggested in Remark 5.5 (a)). It seems clear that those new degenerate cases would consist of analogues of the degenerate cases that have already been identified here in (a). In other words, each of those new degenerate cases Λ of hyperbolas would be built by starting with a union of (possibly denumerably many) open subintervals of the y-axis, each of which is a subset of either the set of points below $(0,-c)$ or the set of points above $(0,c)$, with the precise rules for then building any such Λ (starting with such a union of open subintervals of the y-axis) being the obvious analogues of the corresponding rules that were given above for building the counterpart degenerate hyperbolas Γ. However, in my opinion, these Λ should *not* be considered as being "new" degenerate cases, for the following reason. *Any* hyperbola can be viewed as having a horizontal (resp., vertical) transverse axis after suitable rigid rotation and/or translation of coordinate axes. Such changes of coordinate axes do not change whether a geometric figure is a hyperbola (or whether it should be considered as a degenerate hyperbola in regard to satisfying certain reflection properties) because such changes of coordinate axes do not affect distance or the measure of (undirected) angles between bound vectors. I conclude, for *any* hyperbola $\mathcal{H}$, with transverse (resp., conjugate) axis falling along a line L (resp., M) in the Euclidean plane, that one can build a family of degenerate hyperbolas Γ^* (resp., Λ^*) that is naturally associated to $\mathcal{H}$ by taking the following steps: rigidly rotate and/or translate coordinate axes so that L is horizontal (resp., vertical) and M is vertical (resp., horizontal) in regard to the new coordinate axes [then $\mathcal{H}$ is east-west (resp., north-south) in regard to the new coordinate axes], intersecting at the "new" origin; proceed to use the parameters a, b, c of $\mathcal{H}$ as above to build a degenerate hyperbola Γ (resp., Λ); and then perform (in reverse order) the sequence of the *inverse* operations corresponding to the above-mentioned rigid rotations and/or translations of coordinate axes. Notice that $\mathcal{H}$ has been carried back to its original position relative to the original coordinate axes. By definition, Γ^* (resp., Λ^*) is the geometric figure to which Γ (resp., Λ) has been carried. While Γ^* (resp., Λ^*) is a subset of the line L (resp, M), I would repeat the earlier sentiment that these changes of coordinate axes have not produced anything that is "qualitatively new in any mathematically important way." The remark is complete.

Remark 5.9 offers one final perspective on this body of reflection-theoretic work.

Remark 5.9. As variants of Theorems 2.1 and 4.1 are somewhat well known, I have concluded that Theorems 3.1 and 5.1 are *more* important (because they respectively lead to converses of Theorems 2.1 and 4.1). I have also concluded that Remark 3.2 (d) and Theorem 5.6 (a) are the *most* important results in this paper (because they, respectively, generalize the two just-mentioned "*more* important" results). One reason for this conclusion is that both of these "*most* important" results involved accessible methods involving Cartesian equations (as opposed to polar, or otherwise parametric, equations), vectors, calculus of functions of one variable, and introductory differential equations. However, that reason could be applied to each of the main results in this paper. A deeper – and more appropriate – reason for our second conclusion is that the just-mentioned results from Sections 3 and 5 are characterization results. After careful examinations of their proofs, I was led to what could be considered some partial classification results, involving the degenerate ellipses in Remark 3.2 (f) and the degenerate hyperbolas in Remark 5.8. Results of this kind can be among the most important products of our work as mathematical scientists. Readers who may be interested in a more extensive expression of my views on characterization results and classification results are directed to [3, Remark 3.3].

References

[1] D. E. Dobbs, Limits of conics, Int. J. Math. Educ. Sci. Technol., 23 (5) (1992), 801–805.

[2] D. E. Dobbs, Where is the directrix of a circle?, delta-K Math. J., 43 (2) (2006), 33–36.

[3] D. E. Dobbs, Reflecting on parabolas, Moroccan J. Alg. Geom. Appl., to appear.

[4] D. E. Dobbs and J. C. Peterson, Precalculus, Wm. C. Brown, Dubuque, 1993.

[5] D. Drucker, Euclidean hypersurfaces with reflection properties, Geom. Dedicata 33 (3) (1990), 325–329; Correction to Euclidean hypersurfaces with reflection properties, Geom. Dedicata 39 (3) (1991), 361–362.

[6] D. Drucker, Reflection properties of curves and surfaces, Math. Mag. 65 (3) (1992), 147–157.

[7] W. Fleming, Functions of several variables, Addison-Wesley, Reading, Mass., 1965.

[8] R. E. Johnson and F. L. Kiokemeister, Calculus with analytic geometry, second edition, Allyn and Bacon, Boston, 1960.

[9] L. Leithold, The calculus with analytic geometry, fifth edition, Harper & Row, New York, 1986.

[10] J. M. H. Olmsted, Solid analytic geometry, Appleton-Century-Crofts, New York, 1947.

[11] J. M. H. Olmsted, Advanced calculus, Appleton-Century-Crofts, New York, 1961.

[12] A. E. Taylor, L'Hospital's rule, Amer. Math. Monthly, 59 (1) (1952), 20–24, doi:10.2307/2307183.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 170-177

ISSN: 2820-7114

Title :

Absorbing ideals of the form $I[[X]]$

Author(s):

Sana Hizem

Moroccan Journal of Algebra and Geometry with Applications
Vol. 3(1) (2024), 170–177

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

Absorbing ideals of the form $I[[X]]$

Sana Hizem
Department of Mathematics, Faculty of Sciences, University of Monastir, Tunisia
e-mail: *hizems@yahoo.fr*

Communicated by Aymane Badawi
(Received 24 August 2023, Revised 30 October 2023, Accepted 03 November 2023)

Abstract. Let R be a commutative ring with identity and n a positive integer. In [1], Anderson and Badawi define a proper ideal I of a commutative ring R to be n-absorbing if whenever $x_1 ... x_{n+1} \in I$ for $x_1, ..., x_{n+1} \in R$, then there are n of the $x_i' s$ whose product is in I. In this paper we investigate the transfer of the property n-absorbing from the ideal I of R to the ideal $I[[X]]$ of the formal power series ring $R[[X]]$.

Key Words: absorbing ideals, strongly absorbing ideals, formal power series rings.
2010 MSC: Primary 13A15; 13F25; 13F05; Secondary 13A99.

1 Introduction

All rings considered in this paper are commutative with an identity different from zero. Let R be a commutative ring and n be a positive integer. In [1], Anderson and Badawi define a proper ideal I of a commutative ring R to be n-absorbing if whenever $x_1 ... x_{n+1} \in I$ for $x_1, ..., x_{n+1} \in R$, then there are n of the $x_i' s$ whose product is in I. They also define $\omega_R(I) = \min\{n \mid I \text{ is an } n\text{-absorbing ideal of } R\}$. The ideal I is called strongly n-absorbing if whenever $I_1 ... I_{n+1} \in I$ for ideals $I_1, ..., I_{n+1}$ of R, then there are n of the $I_i' s$ whose product is in I. They define $\omega_R^*(I) = \min\{n \mid I \text{ is a strongly } n\text{-absorbing ideal of } R\}$. It is clear that if I is strongly n-absorbing, then it is n-absorbing, so $\omega_R(I) \leq \omega_R^*(I)$. They conjecture that the converse is true (Conjecture 1). It is clear that for $n = 1$, an ideal I is (strongly) 1-absorbing if and only if I is a prime ideal so Conjecture 1 is true for $n = 1$. Note that for $n = 2$, an ideal I of R is strongly 2-absorbing if and only if I is 2-absorbing [[4], Theorem 2.13]. Note also that in Prüfer domains the two concepts of n-absorbing and strongly n-absorbing ideals are equivalent. On the other hand, they conjecture that $\omega_{R[X]}(I[X]) = \omega_R(I)$ for any ideal I of R (Conjecture 3). A 1-absorbing ideal is just a prime ideal and it is well known that I is a prime ideal if and only if $I[X]$ is a prime ideal so Conjecture 3 is true for $n = 1$. In [1], the authors proved that Conjecture 3 is true for $n = 2$. Many authors investigated this conjecture. For example in [14], the author showed that Conjecture 3 is true if one of the following conditions hold:

(1) The ring R is a Prüfer domain.

(2) The ring R is a Gaussian ring such that its additive group is torsion free.

(3) The additive group of the ring R is torsion-free and I is a radical ideal of R.

In [13], the author proved that if I is a strongly n-absorbing ideal of R and R/I is Armendariz, then $I[X]$ is n-absorbing (R is said to be Armendariz, if $c(f)c(g) = 0$ for all $f, g \in R[X]$ such that $fg = 0$). Moreover, he proved that if I is n-absorbing, then $I[X]$ is n-absorbing in each of the following cases:

(1) The ring R/I is Armendariz and $|R/M| \geq n$ for each maximal ideal M of R containing I.

(2) The ring R/I is Armendariz and is $(n-1)!$-torsion-free as an additive group.

(3) The ring R/I is torsion-free as an additive group.

(4) The ring R/I is locally Bézout.

He showed also that Conjecture 3 is true in an arithmetical ring.

In this paper, we consider n-absorbing ideals of the form $I[[X]]$ of the power series ring $R[[X]]$. More precisely we explore the transfer of the property (strongly) n-absorbing from an ideal I of R to the ideal $I[[X]]$ of $R[[X]]$. The case $n = 1$ is clear since it is well known that an ideal I of R is prime if and only if the ideal $I[[X]]$ is prime. In [10], the authors proved that for an ideal I of a commutative ring R, I is 2-absorbing if and only if $I[[X]]$ is a 2-absorbing ideal of $R[[X]]$ (see also [13]). It was also shown in [10] that if R is a Prüfer domain, then I is n-absorbing if and only if $I[[X]]$ is n-absorbing. The proof was based on the characterization of absorbing ideals in Prüfer domains. In addition, they showed that if R is a Noetherian Gaussian u-ring, then I is n-absorbing if and only if $I[[X]]$ is n-absorbing (a commutative ring R is called u-ring provided R has the property that an ideal contained in a finite union of ideals must be contained in one of those ideals). Moreover, they proved that if R is a pseudo-valuation domain and I is an ideal of R with a non maximal radical, then $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$. On the other hand, in [14], the author proved that for a Dedekind domain R, $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$ for every ideal I of R. Moreover, if R is a Noetherian ring whose additive group is torsion-free, then $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$ for every radical ideal I of R.

In this paper we prove first that if the ideal $I[[X]]$ is n-absorbing, then the ideal I is strongly n-absorbing. Conversely, we prove that if the ideal I is strongly n-absorbing, then the ideal $I[[X]]$ is n-absorbing if one of the following conditions hold:

(1) The ring R is P-gaussian.

(2)The ring R is a Krull domain and I is a divisorial ideal.

(3) The ring R is a formally integrally closed domain and I is a t-ideal.

Most of the results proved here are based on content formulas for power series.

On the other hand, we prove that if the ideal I is n-absorbing, then $I[[X]]$ is n-absorbing if one of the following conditions hold:

(1) The ideal I is radical.

(2) The ring R is a Krull domain and I is of the form $(P_1...P_n)_v$ where the P_i are height one prime ideals of R.

(3) The ideal I has exactly n minimal prime ideals which are comaximal.

(4) The ideal I is a P-primary ideal where P is a prime ideal of R.

2 Absorbing ideals of the form $I[X]$

Let R be a commutative ring, n a positive integer and I a proper ideal of R. In [13], Laradji showed that if $I[X]$ is an n-absorbing ideal of $R[X]$, then I is a strongly n–absorbing ideal of R. We present here another proof which is completely different and which may be of independent interest, so we include it below.

Proposition 2.1. *Let R be a commutative ring, n a positive integer and I a proper ideal of R such that $I[X]$ is an n-absorbing ideal of $R[X]$ then I is a strongly n–absorbing ideal of R.*

Proof. By [[6], Lemma 2.1], let $I_1,...,I_{n+1}$ $(n+1)$ finitely generated ideals of R such that $I_1...I_{n+1} \subset I$. We shall prove that there are n of the $I_i's$ whose product is in I. For $j \in \{1,...,n+1\}$, put $I_j =< a_{1,j};...;a_{k_j,j} >$ and let $f_1 = a_{1,1}X + ... + a_{k_1,1}X^{k_1} \in I_1[X]$, $f_2 = a_{1,2}X^{k_1} + a_{2,2}X^{2k_1} + ... + a_{k_2,2}X^{k_1k_2} \in I_2[X], ...,$ $f_{n+1} = a_{1,n+1}X^{k_1(k_2+1)...(k_n+1)} + ... + a_{k_{n+1},n+1}X^{k_1k_{n+1}(k_2+1)...(k_n+1)} \in I_{n+1}[X]$, then $f_1...f_{n+1} \in I_1[X]...I_{n+1}[X] \subset (I_1...I_{n+1})[X] \subset I[X]$. Hence there are n of the $f_i's$ whose product is in $I[X]$. Suppose for example that $f_1...f_n \in I[X]$, thus $a_{l_1,1}...a_{l_n,n} \in I$, for every $1 \leq l_i \leq k_i$ and $i \in \{1,...,n\}$. Hence, $I_1...I_n \subset I$. □

In the sequel, we will prove that for some class of rings, we have the equivalence: I is a strongly n-absorbing ideal of R if and only if $I[X]$ is an n-absorbing ideal of $R[X]$ and so $\omega_{R[X]}(I[X]) = \omega_R^*(I)$. Recall that a commutative ring R is called Gaussian if $c(fg) = c(f)c(g)$ for all $f, g \in R[X]$, where $c(f)$ denotes the content of the polynomial $f \in R[X]$.

Proposition 2.2. *Let R be Gaussian ring, n a positive integer and I a proper ideal of R. The ideal I is a strongly n-absorbing ideal of R if and only if $I[X]$ is an n-absorbing ideal of $R[X]$. Hence $\omega_{R[X]}(I[X]) = \omega_R^*(I)$.*

Proof. It is sufficient to prove that if I is strongly n-absorbing, then $I[X]$ is n-absorbing. Let $f_1, ..., f_{n+1} \in R[X]$ such that $f_1...f_{n+1} \in I[X]$ then $c(f_1...f_{n+1}) \subset I$. As R is a Gaussian ring then $c(f_1)...c(f_{n+1}) \subset I$. Since I is strongly n-absorbing, there are n of the $c(f_i)$'s whose product is contained in I. But $f_1...f_n \in c(f_1...f_n)[X] \subset c(f_1)...c(f_n)[X] \subset I[X]$. □

In [14], the author proved that if I is a radical n-absorbing ideal and the additive group of the ring R is torsion-free, then $I[X]$ is n-absorbing. In[11], the authors proved that if the ring R satisfies (**) (that is each proper ideal I of R with $\omega_R(I) < \infty$, $\omega_R(I) = |Min_R(I)|$, where $Min_R(I)$ denotes the set of prime ideals of R minimal over I), then if I is a radical n-absorbing ideal, then $I[X]$ is n-absorbing. Note that for a radical strongly n-absorbing ideal I, the ideal $I[X]$ is n-absorbing (without any additional assumption on the ring R) by the Dedekind-Mertens lemma. In the following proposition, we generalize the results of [14] and [11] by releasing the additional assumption on the ring R.

Proposition 2.3. *Let I be a proper radical ideal of a commutative ring R and n a positive integer. The following are equivalent:*

1. *I is a strongly n-absorbing ideal of R.*
2. *I is an n-absorbing ideal of R.*
3. *$I[X]$ is an n-absorbing ideal of $R[X]$.*
4. *$I[X]$ is a strongly n-absorbing ideal of $R[X]$.*
5. *$\forall k \in \mathbb{N}$, $I[X_1, ..., X_k]$ is an n-absorbing ideal of $R[X_1, ..., X_k]$.*
6. *$\forall k \in \mathbb{N}$, $I[X_1, ..., X_k]$ is a strongly n-absorbing ideal of $R[X_1, ..., X_k]$.*

Proof. $1 \Longrightarrow 2$ is clear.
$2 \Longrightarrow 3$ Since I is an n-absorbing ideal of R then $|Min_R(I)| \leq n$ by [[1], Theorem 2.5]. Let $P_1, ..., P_k$ the minimal prime ideals over I. Hence $I = \sqrt{I} = P_1 \cap ... \cap P_k$. Therefore $I[X] = P_1[X] \cap ... \cap P_k[X]$. By [[1], Theorem 2.1], $I[X]$ is k-absorbing so it is also n-absorbing.
$3 \Longrightarrow 1$ is clear.
The other equivalences result from the equality $\sqrt{I[X]} = \sqrt{I}[X]$, so since I is radical then $I[X]$ is also radical and then use an induction on $k \geq 1$.

□

Since every ideal of a von Neumann regular ring is radical, we get the following corollary:

Corollary 2.4. *Let R be a von Neumann regular ring, n a positive integer and I a proper ideal of R. The following are equivalent:*

1. *I is an n-absorbing ideal of R.*

2. *I is a strongly n-absorbing ideal of R.*

3. *$I[X]$ is an n-absorbing ideal of $R[X]$.*

4. *$I[X]$ is a strongly n-absorbing ideal of $R[X]$.*

5. *$\forall k \in \mathbb{N}$, $I[X_1,...,X_k]$ is an n-absorbing ideal of $R[X_1,...,X_k]$.*

6. *$\forall k \in \mathbb{N}$, $I[X_1,...,X_k]$ is a strongly n-absorbing ideal of $R[X_1,...,X_k]$.*

Recall that an ideal I of an integral domain R with quotient field K is called divisorial (or v-ideal) if $I = I_v$, where $I_v = (I^{-1})^{-1}$ and $I^{-1} = R : I = \{x \in K \mid xI \subset R\}$. In the sequel we prove that if I is a divisorial strongly n-absorbing ideal of an integrally closed domain R, then $I[X]$ is n-absorbing.

Lemma 2.5. *Let R be an integrally closed domain. For every $m \in \mathbb{N}^*$ and $f_1,...,f_m \in R[X]$, $(c(f_1...f_m))_v = (c(f_1)...c(f_m))_v$.*

Proof. By [[15], Lemme 1], if R is an integrally closed domain, then for every $f,g \in R[X]$, $(c(fg))_v = (c(f)c(g))_v$, hence the result is obtained by a simple induction on m. □

Proposition 2.6. *Let R be an integrally closed domain, I a divisorial ideal of R and n a positive integer. Then I is strongly n-absorbing if and only if $I[X]$ is n-absorbing. Hence $\omega_{R[X]}(I[X]) = \omega^*_R(I)$.*

Proof. Let $f_1,...,f_{n+1} \in R[X]$ such that $f_1...f_{n+1} \in I[X]$ then $c(f_1...f_{n+1}) \subset I$. Hence $(c(f_1...f_{n+1}))_v \subset I_v = I$. As R is integrally closed then $(c(f_1...f_{n+1}))_v = c(f_1)_v...c(f_{n+1})_v$. Therefore $c(f_1)...c(f_{n+1}) \subset I$. Since I is strongly n-absorbing then there are n of the $c(f_i)'s$ whose product is in I. Suppose for example that $c(f_1)...c(f_n) \subset I$. Consequently, $f_1...f_n \in c(f_1...f_n)[X] \subset c(f_1)...c(f_n)[X] \subset I[X]$. □

3 Absorbing ideals of the form $I[[X]]$

Let R be a commutative ring, I a proper ideal of R and n a positive integer. It is clear that if $I[[X]]$ is an n-absorbing ideal of $R[[X]]$, then $I[X]$ is an n-absorbing ideal of $R[X]$ and so I is a strongly n-absorbing ideal of R. In fact, let $f_1,...,f_{n+1} \in R[X]$ such that $f_1...f_{n+1} \in I[X]$ then $f_1...f_{n+1} \in I[[X]]$ so there are n of the $f_i's$ whose product is in $I[[X]] \cap R[X] = I[X]$.

Note that for a Noetherian ring R, if I is a strongly n-absorbing radical ideal, then $I[[X]]$ is an n-absorbing ideal. In fact, recall first that in [7], the authors established the following Dedekind-Mertens lemma for power series rings:

Proposition 3.1. *[7] Let R be a Noetherian ring and let $0 \neq g \in R[[X]]$. There exists a positive number k such that $c(f)^k c(g) = c(f)^{k-1} c(fg)$ for any $f \in R[[X]]$, where $c(f)$ is the ideal of R generated by the coefficients of f.*

Using this result, we prove that if I is a strongly n-absorbing radical ideal of a Noetherian ring R, then $I[[X]]$ is an n-absorbing ideal. Indeed, let $f_1,...,f_{n+1} \in R[[X]]$ such that $f_1...f_{n+1} \in I[[X]]$ then $c(f_1...f_{n+1}) \subset I$. By the Dedekind-Mertens lemma there exist positive integers $\alpha_1,..,\alpha_n$ such that $c(f_1)^{\alpha_1+1}c(f_2...f_{n+1}) = c(f_1)^{\alpha_1}c(f_1...f_{n+1}) \subset I$, $c(f_2)^{\alpha_2+1}c(f_3...f_{n+1}) = c(f_2)^{\alpha_2}c(f_2...f_{n+1}),..., c(f_n)^{\alpha_n+1}c(f_{n+1}) = c(f_n)^{\alpha_n}c(f_nf_{n+1})$. Now, we multiply the first equality by $c(f_2)^{\alpha_2}$, we get $c(f_1)^{\alpha_1+1}c(f_2)^{\alpha_2+1}c(f_3...f_{n+1}) \subset I$. Continuing this process, we get $c(f_1)^{\alpha_1+1}...c(f_n)^{\alpha_n+1}c(f_{n+1}) \subset I$. As I strongly n-absorbing then there exists $(k_1,...,k_{n+1}) \in \mathbb{N}^{n+1}$ such that $k_1 + ... + k_{n+1} = n$ and $c(f_1)^{k_1}...c(f_n)^{k_n}c(f_{n+1})^{k_{n+1}} \subset I$. Suppose for example that $k_{n+1} = 0$, so $c(f_1)^{k_1}...c(f_n)^{k_n} \subset I$. Since I is radical then $c(f_1)...c(f_n) \subset I$. But $f_1...f_n \in c(f_1...f_n)[[X]] \subset c(f_1)...c(f_n)[[X]] \subset I[[X]]$.

In the sequel we prove that the hypothesis R is Noetherian can be released. More precisely we show that if I is a radical n-absorbing ideal of a commutative ring R, then $I[[X]]$ is n-absorbing.

More generally, in the first part of this section, we prove that if I is an n-absorbing ideal of R, then $I[[X]]$ is an n-absorbing ideal of $R[[X]]$ if one of the following conditions hold:

1. The ideal I is radical.
2. The ring R is a Krull domain and I is of the form $(P_1...P_n)_v$ where the P_i are height one prime ideals of R.
3. The ideal I has exactly n minimal prime ideals which are comaximal.
4. The ideal I is a P-primary ideal where P is a prime ideal of R.

In the next proposition we generalize Corollary 16 of [14] for any commutative ring R.

Proposition 3.2. *Let I be a proper radical ideal of a commutative ring R and n a positive integer. The following are equivalent:*

1. *I is a strongly n-absorbing ideal of R.*
2. *I is an n-absorbing ideal of R.*
3. *$I[[X]]$ is an n-absorbing ideal of $R[[X]]$.*
4. *$I[[X]]$ is a strongly n-absorbing ideal of $R[[X]]$.*
5. *$\forall k \in \mathbb{N}$, $I[[X_1,...,X_k]]$ is an n-absorbing ideal of $R[[X_1,...,X_k]]$.*
6. *$\forall k \in \mathbb{N}$, $I[[X_1,...,X_k]]$ is a strongly n-absorbing ideal of $R[[X_1,...,X_k]]$.*

Proof. The proof is similar to the case of polynomial rings. For the sake of completeness, we include it here.
$1 \Longrightarrow 2$ is clear.
$2 \Longrightarrow 3$ Since I is an n-absorbing ideal of R then $|Min_R(I)| \leq n$ by [[1], Theorem 2.5]. Let $P_1,...,P_k$ the minimal prime ideals over I. Hence $I = \sqrt{I} = P_1 \cap ... \cap P_k$. Therefore $I[[X]] = P_1[[X]] \cap ... \cap P_k[[X]]$. By [[1], Theorem 2.1], $I[[X]]$ is k-absorbing so it is also n-absorbing.
$3 \Longrightarrow 1$ is clear.
The other equivalences result from the equality $\sqrt{I[[X]]} = \sqrt{I}[[X]]$. In fact, $\sqrt{I[[X]]} \subset \sqrt{I}[[X]]$ for any ideal I of R, since if P is a prime ideal of R containing I, then $P[[X]]$ is a prime ideal of $R[[X]]$ containing $I[[X]]$, so $\sqrt{I[[X]]} \subset P[[X]]$ for any prime ideal P containing I which implies that $\sqrt{I[[X]]} \subset \sqrt{I}[[X]]$. Conversely, if I is an n-absorbing ideal of R, then by [5], $(\sqrt{I})^n \subset I$ so $(\sqrt{I}[[X]])^n \subset (\sqrt{I})^n[[X]] \subset I[[X]]$, which implies that $\sqrt{I}[[X]] \subset \sqrt{I[[X]]}$ and then the equality$\sqrt{I[[X]]} = \sqrt{I}[[X]]$.
Now since I is radical then $I[[X]]$ is also radical and then use an induction on $k \geq 1$. □

Corollary 3.3. *Let R be a von Neumann regular ring, n a positive integer and I a proper ideal of R. The following are equivalent:*

1. *I is a strongly n-absorbing ideal of R.*
2. *I is an n-absorbing ideal of R.*
3. *$I[[X]]$ is an n-absorbing ideal of $R[[X]]$.*
4. *$I[[X]]$ is a strongly n-absorbing ideal of $R[[X]]$.*
5. *$\forall k \in \mathbb{N}$, $I[[X_1,...,X_k]]$ is an n-absorbing ideal of $R[[X_1,...,X_k]]$.*

6. $\forall k \in \mathbb{N}$, *$I[[X_1,...,X_k]]$ is a strongly n-absorbing ideal of $R[[X_1,...,X_k]]$.*

Proposition 3.4. *Let R be a Krull domain, n a positive integer and $P_1,..,P_n$ be heigt one prime ideals of R and $I = (P_1...P_n)_v$ then $I[[X]]$ is an n-absorbing ideal of $R[[X]]$. Hence $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$.*

Proof. Note that, by [[1], Corollary 4.5], the ideal I is n-absorbing. We have $I[[X]] = ((P_1...P_n)[[X]])_v = ((P_1...P_n).A[[X]])_v = ((P_1.A[[X]])...(P_n.A[[X]]))_v$. So $I[[X]] = ((P_1.A[[X]])_v...(P_n.A[[X]])_v)_v = ((P_1[[X]])_v...(P_n[[X]])_v)_v = (P_1[[X]]...P_n[[X]])_v$.
By [9], $R[[X]]$ is also a Krull domain and for each $k \in \{1,...,n\}$, $P_k[[X]]$ is a height one prime ideal of $R[[X]]$. Hence by [[1], Corollary 4.5], the ideal I is n-absorbing. □

In the following, we give two cases where the property n-absorbing is stable when passing from I to the ideal $I[[X]]$.

Proposition 3.5. *Let I be an n-absorbing ideal of a ring R such that I has exactly n minimal prime ideals which are comaximal then $I[[X]]$ is n-absorbing. Hence $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$.*

Proof. Let $\{P_1,...,P_n\}$ be the minimal prime ideals over I. By [[1], Corollary 2.15], $I = P_1...P_n = P_1 \cap ... \cap P_n$, so $P[[X]] = P_1[[X]] \cap ... \cap P_n[[X]]$. Again by Theorem 2.1 of [1], the ideal $I[[X]]$ is n-absorbing. □

Proposition 3.6. *Let P be a prime ideal of a ring R and I be a primary ideal of R such that $P^n \subset I$ then $I[[X]]$ is n-absorbing. Hence $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$.*
In particular if P^n is a P-primary ideal of R, then $P^n[[X]]$ is n-absorbing. Moreover, if M is a maximal ideal of R, then $M^n[[X]]$ is n-absorbing.

Proof. By [[1], Theorem 3.1], the ideal I is n-absorbing. By [[8], Corollary 4], $I[[X]]$ is a $P[[X]]$-primary ideal of $R[[X]]$ and $(P[[X]])^n \subset P^n[[X]] \subset I[[X]]$. So again by [[1], Theorem 3.1], the ideal $I[[X]]$ is n-absorbing. □

In the sequel, we prove that if I is a strongly n-absorbing ideal of R, then $I[[X]]$ is an n-absorbing ideal of $R[[X]]$ if one of the following conditions hold:

1. The ring R is P-Gaussian.
2. The ring R is a Krull domain and I is a divisorial ideal.
3. The ring R is a formally integrally closed domain and I is a t-ideal.

Recall from [16], that a commutative ring R is called P-Gaussian if for every $f,g \in R[[X]]$, $c(fg) = c(f)c(g)$. For example a Noetherian Gaussian ring is P-Gaussian.

Proposition 3.7. *Let R be a P-Gaussian ring, n a positive integer and I an ideal of R. Then $I[[X]]$ is n-absorbing if and only if I is strongly n-absorbing. Hence $\omega_{R[[X]]}(I[[X]]) = \omega^*_R(I)$.*

Proof. Let $f_1,...,f_{n+1} \in R[[X]]$ such that $f_1...f_{n+1} \in I[[X]]$ then $c(f_1...f_{n+1}) \subset I$. As R is a P-Gaussian ring then $c(f_1)...c(f_{n+1}) \subset I$. Since I is strongly n-absorbing then $c(f_1)...c(f_n) \subset I$ for example. But $f_1...f_n \in c(f_1...f_n)[[X]] \subset c(f_1)...c(f_n)[[X]] \subset I[[X]]$. □

Proposition 3.8. *Let R be an integral domain such that $R = \bigcap\limits_{\alpha} V_\alpha$ where $(V_\alpha)_\alpha$ is a collection of rank one valuation overrings of R and I a strongly n-absorbing ideal such that $I = \bigcap\limits_{\alpha} IV_\alpha$ then $I[[X]]$ is n-absorbing. In particular, if R is a Krull domain and I is a strongly n-absorbing divisorial ideal, then $I[[X]]$ is n-absorbing. Hence $\omega_{R[[X]]}(I[[X]]) = \omega^*_R(I)$.*

Proof. Consider the star operation $*$ defined by $E^* = \bigcap_{\alpha} IV_\alpha$, for every nonzero fractional ideal of R. By [[2], Theorem 2.5] for nonzero $f,g \in R[[X]]$, $(c(fg))^* = (c(f)c(g))^*$. Let $f_1,...,f_{n+1} \in R[[X]]$ such that $f_1...f_{n+1} \in I[[X]]$ then $c(f_1...f_{n+1}) \subset I$. Hence $c(f_1)...c(f_n) \subset (c(f_1)...c(f_n))^* = (c(f_1...f_n))^* \subset I^* = I$. Now the result follows from the fact that I is strongly n-absorbing. □

Now we can recover Corollary 11 of [14] since a Dedekind domain is a Krull domain in which every ideal is divisorial. Moreover a Dedekind domain is a Prüfer domain so by [[1], Corollary 6.9], every n-absorbing ideal is strongly n-absorbing.

Corollary 3.9. *Let R be a Dedekind domain, then $\omega_{R[[X]]}(I[[X]]) = \omega_R(I)$.*

More generally if R is a completely integrally closed domain and I is a strongly n-absorbing divisorial ideal, then $I[[X]]$ is n-absorbing by [[12], Theorem 2.11].

Recall from [3], that an integral domain R is called formally integrally closed if for nonzero $f,g \in R[[X]]$, $(c(fg))_t = (c(f)c(g))_t$, where $I_t = \cup\{J_v \mid J$ is a finitely generated non zero fractional ideal of R such that $J \subset I\}$, for every non zero fractional ideal I of R. A nonzero fractional ideal I of R is called a t-ideal if $I_t = I$. Integral domains R such that R_M is a one dimensional valuation domain for every t-maximal ideal of R are examples of formally integrally closed domains. We get then the following proposition:

Proposition 3.10. *Let R be a formally integrally closed domain, n a positive integer and I a strongly n-absorbing t-ideal then $I[[X]]$ is n-absorbing. Hence $\omega_{R[[X]]}(I[[X]]) = \omega^*_R(I)$.*

References

[1] Anderson, D.F., Badawi, A., On n-absorbing ideals of commutative rings, Comm. Algebra, 39, 1646 – 1672 (2011)

[2] Anderson, D.D., Kang, B.G., Content formulas for polynomials and power series and complete integral closure, J. Algebra, 181, 82 – 94 (1996)

[3] Anderson, D.D., Kang, B.G., Formally integrally closed domains and the rings $R((X))$ and $R\{\{X\}\}$, J. Algebra, 200, 347 – 362 (1998)

[4] Badawi, A., On 2-absorbing ideals of commutative rings, Bull. Austral. Math. Soc., 75, 417 – 429 (2007)

[5] Choi, H.S., Walker, A., The radiacl of an n-absorbing ideal, J. Commut. Algebra, 12 (2), 171 – 177 (2020)

[6] Donadze, G., The Anderson-Badawi conjecture for commutative algebras over infinite fields. Indian J. Pure Appli. Math., 47 (4), 691 – 696 (2016)

[7] Epstein, N., Shapiro, J., A Dedekind-Mertens theorem for power series rings, Proc. Am. Math. Soc., 144 (3), 917 – 924 (2016)

[8] Fields, D.E., Zero divisors and nilpotent elements in power series rings, Proc. Am. Math. Soc., 3 (27), 427 – 433 (1971)

[9] Gilmer, R. Power series rings over a Krull domain. Pac. J. Math., 29, 543 – 549 (1969)

[10] Hizem, S, Smach, S., On Anderson-Badawi conjectures, Beitr. Algebra Geom., 58 (4), 775 – 785 (2017)

[11] Issoual, M., Mahdou, N., Moutui, M.A.S., On n-absorbing prime ideals of commutative rings, Hacet. J. Math. Stat., 51 (2), 455 – 465 (2022)

[12] Kang, B.G., Park, M.H., Toan, P.T., Dedekind-Mertens lemma and content formulas in power series rings, J. Pure Appl. Algebra, 222, 2299 – 2309 (2018)

[13] Laradji, A., On n-absorbing rings and ideals. Colloq. Math., 147 (2), 265 – 273 (2017)

[14] Nasehpour, P., On the Anderson-Badawi $\omega_{R[X]}(I[X]) = \omega_R(I)$ conjecture, Arch. Math., Brno, 52 (2), 71 – 78 (2016)

[15] Querré, J., idéaux divisoriels d'un anneau de polynômes, J. Algebra, 64, 270 – 284 (1980)

[16] Tsang, H., Gauss lemma, Ph. D thesis, University of Chicago, (1965)

Moroccan Journal of Algebra
and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 178-186

ISSN: 2820-7114

Title :

On $\phi$-Prüfer and $\phi$-Bézout rings in amalgamation algebra along an ideal

Author(s):

Younes El Haddaoui

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 178–186

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco
ISSN: 2820-7114

On ϕ-Prüfer and ϕ-Bézout rings in amalgamation algebra along an ideal

Younes El Haddaoui
Department of Mathematics, Faculty of Science and Technology, Fez, Morocco.
e-mail: *younes.elhaddaoui@usmba.ac.ma*

Communicated by Moutu Abdou Salam Moutui
(Received 08 September 2023, Revised 10 December 2023, Accepted 13 December 2023)

Abstract. In this paper, all rings considered are assumed commutative with nonzero identity. A ring R is said to be ϕ-ring if its Nilradical is divided and prime ideal. The authors of [2] introduced and studied two new generalizations of Prüfer domains and Bézout domains respectively, a ring R is said to be ϕ-Prüfer ring (resp., ϕ-Bézout ring) provided that $R/Nil(R)$ is a Prüfer domain (resp., a Bézout domain). In this work, we study the notions of ϕ-Prüfer rings and ϕ-Bézout ring in different contexts of commutative rings such us trivial ring extensions and amalgamations of algebras along ideals. Our aim is to generate new families of ϕ-Prüfer rings and ϕ-Bézout rings and also to enrich the literature with such a rings. Examples illustrating the aims and scopes of our results are given.

Key Words: ϕ-Prüfer rings, ϕ-Bézout rings, nonnil coherent rings, ϕ-coherent rings.
2010 MSC: 13A15, 13A18, 13F05, 13G05, 13C20.

1 Introduction

All rings considered in this paper are assumed to be commutative with non-zero identity and prime Nilradical. We use $Nil(R)$ to denote the set of nilpotent elements of R and $Z(R)$ the set of zero-divisors of R. A ring with Nil(R) being divided prime (i.e., $Nil(R) \subset xR$ for all $x \in R \setminus Nil(R)$) is called a *ϕ-ring*. El Khalfi, Kim, and Mahdou [15], and Chhiti, Louartiti, and Tamekkante [8] studied when the amalgamation algebra along an ideal is a ϕ-ring. Let $\mathcal{H}$ be the set of all rings with divided prime Nilradical. A ring R is called a *strongly ϕ-ring* if $R \in \mathcal{H}$ and $Z(R) = Nil(R)$. Let R be a ring and M be an R-module; we define

$$\phi\text{-tor}(M) = \{x \in M \mid sx = 0 \text{ for some } s \in R \setminus Nil(R)\}.$$

If ϕ-tor$(M) = M$, then M is called a ϕ-torsion module, and if ϕ-tor$(M) = 0$, then M is called a *ϕ-torsion free module*. It is worth noting that in the language of torsion theory, the class $\mathfrak{T}$ of all ϕ-torsion modules is a (hereditary) torsion class, equivalently $\mathfrak{T}$ is closed under (submodules,) direct sums, epimorphic images and extensions. An R-module M is said to be ϕ-uniformly torsion (ϕ-u-torsion for short) if $sM = 0$ for some $s \in R \setminus Nil(R)$. An ideal I of R is said to be *nonnil* if $I \not\subseteq Nil(R)$. An R-module M is said to be *ϕ-divisible* if $M = sM$ for all $s \in R \setminus Nil(R)$. Recall that in [24], a module M is said to be Bézout if all finitely generated submodules of M are cyclic. For more on ϕ-rings from a module-theoretic point of view, see survey article [20].

Let R be a ring and E an R-module. Then $R \propto E$, the trivial ring extension of R by E, is the ring whose additive structure is that of the external direct sum $R \oplus E$ and whose multiplication is defined by $(a,e)(b,f) := (ab, af + be)$ for all $a, b \in R$ and all $e, f \in E$. (This construction is also known by other terminology and other notation, such as the idealization $R(+)E$) (see [5, 16, 17, 18, 19]).

Let A and B be two rings, let J be an ideal of B and let $f : A \longrightarrow B$ be a ring homomorphism. In this setting, we can consider the following sub-ring of $A \times B$:

$$A \bowtie^f J = \{(a, f(a) + j) \mid a \in A, j \in J\},$$

called the amalgamation of A with B along J with respect to f (introduced and studied by D'Anna et al. [10, 11]). This construction is a generalization of the amalgamated duplication of a ring along an ideal (introduced and studied by D'Anna and Fontana [9] and denoted by $A \bowtie I$).

In [2], Anderson and Badawi introduced the class of ϕ-rings which are called ϕ-Prüfer and ϕ-Bézout rings. A ϕ-ring is said to be ϕ-Prüfer if $R/Nil(R)$ is a Prüfer domain [2, Theorem 2.6]. All ϕ-Prüfer rings are Prüfer [2, Theorem 2.14], if in addition $Z(R) = Nil(R)$, then all Prüfer rings are is ϕ-Prüfer [2, Theorem 2.16]. A ϕ-ring R is said to be ϕ-Bézout if $R/Nil(R)$ is a Bézout domain [2, Theorem 3.3]. Recently, the authors of [14] gave some homological properties characterized the ϕ-Prüfer and the ϕ-Bézout rings.

From [1], Bacem and Ali introduced a new generalization of coherent rings in the class of ϕ-rings, a ϕ-ring R is called *ϕ-coherent* if $R/Nil(R)$ is a coherent domain [1, Corollary 3.1]. A ϕ-ring R is said to be *nonnil-coherent* if every finitely generated nonnil ideal is finitely presented, which is equivalent to saying that R is ϕ-coherent and $(0 : r)$ is a finitely generated ideal of R for each $r \in R \setminus Nil(R)$, where $(0 : r) = \{x \in R \mid rx = 0\}$ [22, Proposition 1.3]. In [3], Badawi introduced and studied a new class of ϕ-rings which are said to be nonnil-Noetherian. A ϕ-ring R is said to be *nonnil-Noetherian* if $R/Nil(R)$ is a Noetherian domain [3, Theorem 1.2].

Among the many recent generalizations of the notion of a coherent module and Noetherian module in the literature, we find the following, due to Y. El Haddaoui, H. Kim and N. Mahdou [13], a submodule N of an R-module M is said to be ϕ-submodule if M/N is a ϕ-torsion module [13, Definition 2.1]. For $R \in \mathcal{H}$, an R-module M is said to nonnil-coherent if M is finitely generated and every finitely generated ϕ-submodule of M is finitely presented [13, Definition 2.4], it's easy to see that every coherent module over a ϕ-ring is nonnil-coherent. They next established in [13, Theorem 2.6] the analogue of well-known behavior the relationship between the coherent rings and the finitely generated submodules of a finitely generated free module, it's shown that a ϕ-ring R is nonnil coherent if and only if every finitely generated ϕ-submodule of a finitely generated free module is finitely presented. If $R \in \mathcal{H}$, then an R-module M is said to be nonnil-Noetherian if every ϕ-submodule N of M is a finitely generated [13, Definition 3.1]. It's shown in [13, Theorem 3.15] that over a nonnil-Noetherian ring, every finitely generated ϕ-torsion module M is finitely presented.

This paper consists of three sections including introduction. In section 2, we give some results of the paper [15] which characterizes when an amalgamation of rings is a ϕ-rings and we next study the transfer of ϕ-Prüfer in the amalgamation algebra along an ideal. In section 3, we introduce and study a new class of modules over a ϕ-ring which are called the ϕ-Bézout modules. An R-module M is said to be ϕ-Bézout if every finitely generated ϕ-submodule N of M (that is a submodule N of M such that M/N is ϕ-torsion) is cyclic (see Definition 3.1). We next study the transfer of ϕ-Bézout in the amalgamation algebra along an ideal.

For any undefined terminology and notation the reader is referred to [17, 21, 23]. Throughout this paper, if S is a multiplicative subset of a ring R, then we assume that $S \cap Nil(R) = \emptyset$.

2 On transfer ϕ-Prüfer rings in amalgamation algebra along an ideal

Our first result characterizes when the amalgamation of rings is ϕ-ring. Before starting this section,we give some results of the paper [15] which characterizes when an amalgamation of rings is a ϕ-rings.

Theorem 2.1. Let $f : A \to B$ be a ring homomorphism and J be an ideal of B. Then

$$Nil(A \bowtie^f J) = \{(a, f(a) + j),\ a \in Nil(A) \text{ and } j \in J \cap Nil(B)\}.$$

Proof. Straightforward. □

Recall from [4, 12] that a prime ideal P of R is called a divided prime ideal if $P \subset Rx$ for every $x \in R \backslash P$.

The following Theorem 2.2 characterizes when the amalgamation of a ring along a nonnil ideal is a ϕ-ring.

Theorem 2.2. ([15, Theorem 2.1]) Let $f : A \to B$ be a ring homomorphism and J be a nonnil ideal of B. Set $N(J) = J \cap Nil(B)$. The following statements are equivalent:

1. $R = A \bowtie^f J \in \mathcal{H}$,
2. A is an integral domain, $f^{-1}(J) = 0$ and $N(J)$ is a divided prime ideal of $f(A) + J$.

Theorem 2.3. ([15, Corollary 2.6 – 2.7]) The following results holds for a ring A:
1) The polynomial ring $A[X]$ is a ϕ-ring if and only if A is an integral domain.
2) If A is a reduced ring and I is an ideal of A, then $A \bowtie I$ is a ϕ-ring if and only if A is a ϕ-ring and $I = 0$.

Recall that an R-module M is said to be ϕ-divisible if $sM = M$ for all $s \in R \backslash Nil(R)$. The following characterizes when the trivial ring extension is ϕ-ring.

Theorem 2.4. ([15, Corollary 2.4]) Let $A = R \propto M$ be a trivial ring extension. The following statements are equivalent:

1. $A \in \mathcal{H}$,
2. $R \in \mathcal{H}$ and M is a ϕ-divisible R-module.

Next, we study the transfer of ϕ-Prüfer in the amalgamation algebra along an ideal. Recall from [2] that a ring R is called ϕ-Prüfer if $R/Nil(R)$ is a Prüfer domain.

The following Theorem 2.5 characterizes when an amalgamation algebra along a nonnil ideal is a ϕ-Prüfer ring.

Theorem 2.5. Let A and B two rings and $f : A \longrightarrow B$ be a ring homomorphism. Let J be a nonnil ideal of B. Set, $\bar{f} : A \longrightarrow B/N(J)$ defined by: $\bar{f}(a) = f(a) + N(J)$, for all $a \in A$. If $A \bowtie^f J$ is a ϕ-ring, then the following statements are equivalent:

1. $A \bowtie^f J$ is a ϕ-Prüfer ring,
2. $A \bowtie^{\bar{f}} \frac{J}{N(J)}$ is a Prüfer domain,
3. $f^{-1}(J) = \{0\}$ and $\bar{f}(A) + J/N(J)$ is a Prüfer domain.

Before proving Theorem 2.5, we need the following lemmas.

Lemma 2.6. *([13, Lemma 5.4]) With the notations of Theorem 2.5, we get $\bar{f}^{-1}(J/N(J)) = f^{-1}(J)$.*

Lemma 2.7. *([13, Lemma 5.5]) Let $f : A \longrightarrow B$ be a ring homomorphism and J be a nonzero ideal of B. Let J' be a subideal of J and I be an ideal of A such that $f(I) \subset J'$. Define $\overline{\overline{f}} : A/I \longrightarrow B/J'$ by $\overline{\overline{f}}(\overline{a}) = \overline{f(a)}$, where $\overline{a} := a + I$ and $\overline{f(a)} := f(a) + J'$. Then we have the following ring isomorphism:*

$$\frac{A \bowtie^f J}{I \bowtie^f J'} \cong \frac{A}{I} \bowtie^{\overline{\overline{f}}} \frac{J}{J'}.$$

Proof of Theorem 2.5. 1) $\Longrightarrow$ 2) Assume that $A \bowtie^f J$ is a ϕ-Prüfer ring. Since $A \bowtie^f J \in \mathcal{H}$, A is an integral domain by Theorem 2.2. Therefore, $Nil(A \bowtie^f J) = 0 \times N(J)$. As $A \bowtie^f J$ is a ϕ-Prüfer ring, we get $\frac{A \bowtie^f J}{0 \times N(J)}$ is a Prüfer domain. Therefore, $A \bowtie^{\bar{f}} \frac{J}{N(J)}$ is a Prüfer domain, by Lemma 2.7.

2) $\Longrightarrow$ 1) Straightforward by Lemma 2.7 and [2, Theorem 2.6].

2) $\Longrightarrow$ 3) Assume that $A \bowtie^{\bar{f}} J/N(J)$ is a Prüfer domain. From [11, Proposition 5.2] and Lemma 2.6, $f^{-1}(J) = 0$ and $\bar{f}(A) + J/N(J)$ is an integral domain. From [11, Proposition 5.1] $\bar{f}(A) + J/N(J) \cong A \bowtie^{\bar{f}} J/N(J)$, as desired $\bar{f}(A) + J/N(J)$ is a Prüfer domain.

3) $\Longrightarrow$ 2) By Lemma 2.6 we have $\bar{f}^{-1}(J/N(J)) = 0$ and from [11, Proposition 5.1] we get $\bar{f}(A) + J/N(J) \cong A \bowtie^{\bar{f}} J/N(J)$ which is a Prüfer domain, as desired. □

Corollary 2.8. *Let R be an integral domain. Then $R[X]$ is a ϕ-Prüfer ring if and only if $R[X]$ is a Prüfer domain.*

Proof. By Theorem 2.3, $R[X]$ is a ϕ-ring and $R[X] \cong R \bowtie^j J$ where $J = XR[X]$ and $j : R \hookrightarrow R[X]$. Since $J \not\subset Nil(R[X])$, then $R[X]$ is a ϕ-Prüfer ring if and only if $R \bowtie^j J$ is a Prüfer domain by Theorem 2.5, as desired $R[X]$ is a Prüfer domain. □

Corollary 2.9. *Let A be a ring and J be a nonnil ideal of A. Assume that $A \bowtie J \in \mathcal{H}$. Then $A \bowtie J$ is never a ϕ-Prüfer ring.*

Proof. If $A \bowtie J$ is a ϕ-Prüfer ring, then $A \bowtie J/Nil(A)$ is a Prüfer domain and so $J = Nil(A)$ by [11, Remark 5.3]. Therefore, $J \subset Nil(A)$, a desired contradiction. □

Next, Theorem 2.10 study the transfer of being a ϕ-Prüfer ring between a ϕ-ring A and an amalgamation algebra along a nil ideal $A \bowtie^f J$.

Theorem 2.10. Let A and B two rings and $f : A \longrightarrow B$ be a ring homomorphism. Let J be a nil ideal of B. If $A \bowtie^f J$ is a ϕ-ring, then the following statements are equivalent:

1. $A \bowtie^f J$ is a ϕ-Prüfer ring,
2. A is a ϕ-Prüfer ring.

Proof. First, we have $J \subset Nil(B)$, thus $N(J) = J$ and so $Nil(A \bowtie^f J) = Nil(A) \bowtie^f J$. Therefore, $A \bowtie^f J$ is a ϕ-Prüfer ring if and only if $\frac{A \bowtie^f J}{Nil(A) \bowtie^f J}$ is a Prüfer domain, if and only if $\frac{A}{Nil(A)}$ is a Prüfer domain, if and only if A is a ϕ-Prüfer ring. □

The following Corollary 2.11 study the transfer of being a ϕ-Prüfer ring in trivial extensions.

Corollary 2.11. *Let $R \in \mathcal{H}$ and M be a ϕ-divisible R-module. The following statements are equivalent:*

1. *$R \propto M$ is a ϕ-Prüfer ring,*
2. *R is a ϕ-Prüfer ring.*

Proof. By Theorem 2.4, $R \propto M$ is a ϕ-ring. In addition $R \propto M \cong R \bowtie^j J$ where $J = 0 \propto M \subset Nil(R \propto M)$ and $j : R \hookrightarrow R \propto M$ such that $j(r) = (r, 0)$ for all $r \in R$. By Theorem 2.10, $R \propto M$ is a ϕ-Prüfer ring if and only if so is R. □

Recall from [3] that a ϕ-ring is said to be nonnil-Noetherian if $R/Nil(R)$ is a Noetherian domain. Recall from [1] that a ϕ-ring is said to be nonnil coherent if every finitely generated nonnil ideal is finitely presented, a ϕ-ring R is said to be ϕ-coherent if $R/Nil(R)$ is a coherent domain and so all ϕ-Prüfer ring is ϕ-coherent.

The following Example 2.12 gives a ϕ-Prüfer ring which isn't a nonnil- Noetherian ring.

Example 2.12. The ring $R := (\mathbb{Z} + X\mathbb{Q}[[X]]) \propto qf(\mathbb{Q}[[X]])$ is a ϕ-Prüfer ring which isn't a nonnil-Noetherian ring.

Proof. It is clear that $R/Nil(R) \cong \mathbb{Z} + X\mathbb{Q}[[X]]$ which is a Prüfer domain by [6, Theorem 2.1 (i)], but the coset $R/Nil(R)$ is never a Noetherian domain by [6, Theorem 2.1 (m)]. Therefore, R is ϕ-Prüfer ring by Corollary 2.11 which is not nonnil-Noetherian. □

The following Example 2.13 gives a nonnil-Noetherian ring which isn't a ϕ-Prüfer ring.

Example 2.13. For a field K, the ring $R := K[X,Y] \propto qf(K[X,Y])$ is a nonnil-Noetherian ring which isn't a ϕ-Prüfer ring.

Proof. First, we have $w.gldim(K[X,Y]) = 2$ and so $K[X,Y]$ isn't a Prüfer domain. Therefore, R is not ϕ-Prüfer by Corollary 2.11. The coset $R/Nil(R) \cong K[X,Y]$ is a Noetherian domain and so R is nonnil-Noetherian. □

It is natural to ask when a ϕ-Prüfer ring is nonnil-coherent. The following Theorem 2.14 responds to our question.

Theorem 2.14. The following statements are equivalent for a ϕ-Prüfer ring R:

1. R is nonnil-coherent,
2. $(0:s) := \{r \in R \mid rs = 0\}$ is a finitely generated ideal for every $s \in R \backslash Nil(R)$.

Proof. Assume that R is a nonnil-coherent. It's easy to see that every $s \in R \backslash Nil(R)$ we get $(0:s)$ is a finitely generated ideal of R by [22, Proposition 1.3].

Conversely, If we have $(0:s)$ is a finitely generated ideal of R for every $s \in R \backslash Nil(R)$, then by [22, Proposition 1.3] R is a nonnil-coherent ring since it's a ϕ-Prüfer ring. □

The following Example 2.15 gives a ϕ-Prüfer ring which isn't nonnil-coherent.

Example 2.15. The ring $R = \mathbb{Z} \propto \oplus_{i=1}^{\infty} \mathbb{Q}/\mathbb{Z}$ is a ϕ-Prüfer ring which isn't a nonnil-coherent ring.

Proof. From [13, Example 4.11], we get R is not a nonnil-coherent ring. By Corollary 2.11, we get immediately $R = \mathbb{Z} \propto \oplus_{i=1}^{\infty} \mathbb{Q}/\mathbb{Z}$ is a ϕ-Prüfer ring. □

Remark 2.16. The ring in the Example 2.13 is a nonnil-coherent ring which isn't ϕ-Prüfer.

Proof. First, it is easy to see that $R/Nil(R) = K[X,Y]$ is a coherent domain and so R is ϕ-coherent. But $Z(R) = Nil(R)$, as desired R is nonnil-coherent by [1, Remark 2.1]. □

3 On ϕ-Bézout modules and transfer of ϕ-Bézout rings in amalgamation algebra along an ideal

Recall from [24] that a module M is said to be Bézout if every finitely generated submodule of M is cyclic. Recall that an R-module M is said to be ϕ-torsion if for any $x \in M$, we get $sx = 0$ for some $s \in R \backslash Nil(R)$.

Definition 3.1. Let $R \in \mathcal{H}$. An R-module M is said to be ϕ-Bézout if every finitely generated ϕ-submodule of M (that is a submodule N of M such that M/N is ϕ-torsion) is cyclic. In particular, every Bézout module is ϕ-Bézout.

Remark 3.2. 1) Note that for any ϕ-torsion R-module M, we have

$$M \text{ is } \phi\text{-Bézout} \Longleftrightarrow M \text{ is Bézout.}$$

2) Recall from [13] that a module M over a ϕ-ring R is said to be nonnil-Noetherian if every its ϕ-submodule is finitely generated, [13, Definition 3.1]. Then it's easy to see that every ϕ-Bézout module is nonnil-Noetherian.

Theorem 3.3. Let R be a ϕ-ring. Then R is a ϕ-Bézout R-module if and only if R is a ϕ-Bézout ring.

Proof. This follows immediately from the fact that every nonnil ideal of R is a ϕ-submodule of R, the Definition 3.1 and [2, Theorem 3.2]. □

Theorem 3.4. Every ϕ-submodule of a ϕ-Bézout module is ϕ-Bézout.

Proof. Let M be a ϕ-Bézout module and N be a ϕ-submodule of M. We claim that N is a ϕ-Bézout module, let X be a finitely generated ϕ- submodule of N, by the following exact sequence $0 \to N \to M \to M/N \to 0$, we get the exact sequence $0 \to N/X \to M/X \to M/N \to 0$ and so X is a ϕ-submodule of M from [25, Proposition 2.4]. Therefore, X is cyclic, as desired N is a ϕ-Bézout module. □

Theorem 3.5. If M is a ϕ-Bézout module, then every factor of M is ϕ-Bézout.

Proof. Let M be a ϕ-Bézout module and N be a submodule of M. We claim that M/N is a ϕ-Bézout module. Let P/N be a ϕ-submodule of M/N where P is a submodule of M containing N. Since $\frac{M/N}{P/N} \cong \frac{M}{P}$ is a ϕ-torsion R-module, then P is cyclic and so P/N is a cyclic submodule of M/N, as desired M/N is ϕ-Bézout. □

If $R \in \mathcal{H}$ and S be a multiplicative subset of R, then it's easy to establish that $S^{-1}R \in \mathcal{H}$.

Next, Theorem 3.6 establishes that the ϕ-Bézout modules are closed by localization.

Theorem 3.6. Let R be a ϕ-ring and S be a multiplicative subset of R. If M is a ϕ-Bézout R-module, then $S^{-1}M$ is a ϕ-Bézout $(S^{-1}R)$-module.

Proof. Let M be a ϕ-Bézout R-module and $S^{-1}N$ be a ϕ-submodule of $S^{-1}M$ where N is a submodule of M. It is easy to establish that N is a ϕ-submodule of M, thus N is a cyclic R-module and so $S^{-1}N$ is a cyclic $(S^{-1}R)$-module, as desired $S^{-1}M$ is a ϕ-Bézout $(S^{-1}R)$-module. □

Next, attention is paid to the localization of ϕ-Bézout rings. Using Theorem 3.6, we obtain immediately.

Corollary 3.7. *If R is a ϕ-Bézout ring and S is a multiplicative subset of R, then $S^{-1}R$ is a ϕ-Bézout ring.*

Proof. Straightforward. □

Now, we study the transfer of being a ϕ-Bézout rings in the amalgamation algebra along an ideal.

The following Theorem 3.8 characterizes when an amalgamation algebra along a nonnil ideal is a ϕ-Bézout ring.

Theorem 3.8. Let A and B two rings and $f : A \longrightarrow B$ be a ring homomorphism. Let J be a nonnil ideal of B. Set, $\bar{f} : A \longrightarrow B/N(J)$ defined by: $\bar{f}(a) = f(a) + N(J)$, for all $a \in A$. If $A \bowtie^f J$ is a ϕ-ring, then the following statements are equivalent:

1. $A \bowtie^f J$ is a ϕ-Bézout ring,
2. $A \bowtie^{\bar{f}} \frac{J}{N(J)}$ is a Bézout domain,

3. $f^{-1}(J)=\{0\}$ and $\bar{f}(A)+J/N(J)$ is a Bézout domain.

Proof. 1) $\Longrightarrow$ 2) Assume that $A\bowtie^f J$ is a ϕ-Bézout ring. Since $A\bowtie^f J\in\mathcal{H}$, we get A is an integral domain by Theorem 2.2 and so $Nil(A\bowtie^f J)=0\times N(J)$. As $A\bowtie^f J$ is a ϕ-Bézout ring, we get $\frac{A\bowtie^f J}{0\times N(J)}$ is a Bézout domain. Therefore, $A\bowtie^{\bar{f}}\frac{J}{N(J)}$ is a Bézout domain by Lemma 2.7.

2) $\Longrightarrow$ 1) Straightforward by Lemma 2.7 and [2, Theorem 3.3].

2) $\Longrightarrow$ 3) Assume that $A\bowtie^{\bar{f}} J/N(J)$ is a Bézout domain. From [11, Proposition 5.2] and Lemma 2.6, $f^{-1}(J)=0$ and $\bar{f}(A)+J/N(J)$ is an integral domain. From [11, Proposition 5.1] $\bar{f}(A)+J/N(J)\cong A\bowtie^{\bar{f}} J/N(J)$, as desired $\bar{f}(A)+J/N(J)$ is a Bézout domain.

3) $\Longrightarrow$ 2) By Lemma 2.6 we have $\bar{f}^{-1}(J/N(J))=0$ and from [11, Proposition 5.1] we get $\bar{f}(A)+J/N(J)\cong A\bowtie^{\bar{f}} J/N(J)$ which is a Bézout domain, as desired. □

Corollary 3.9. *Let R be an integral domain. Then $R[X]$ is a ϕ-Bézout ring if and only if $R[X]$ is a Bézout domain.*

Proof. By Theorem 2.3, $R[X]$ is a ϕ-ring and $R[X]\cong R\bowtie^j J$ where $J=XR[X]$ and $j:R\hookrightarrow R[X]$. Since $J\not\subset Nil(R[X])$, $R[X]$ is ϕ-Bézout ring if and only if $R\bowtie^j J$ is a Bézout domain by Theorem 3.8. □

Next, Theorem 3.10 study the transfer of being a ϕ-Bézout ring between a ϕ-ring A and amalgamation algebra along a nil ideal $A\bowtie^f J$.

Theorem 3.10. Let A and B two rings and $f:A\longrightarrow B$ be a ring homomorphism. Let J be a nil ideal of B. If $A\bowtie^f J$ is a ϕ-ring, then the following statements are equivalent:

1. $A\bowtie^f J$ is a ϕ-Bézout ring,
2. A is a ϕ-Bézout ring.

Proof. Since $J\subset Nil(B)$, we get $N(J)=J$. It is easy to see that $Nil(A\bowtie^f J)=Nil(A)\bowtie^f J$. Therefore, $A\bowtie^f J$ is a ϕ-Bézout ring if and only if $\frac{A\bowtie^f J}{Nil(A)\bowtie^f J}$ is a Bézout domain if and only if $\frac{A}{Nil(A)}$ is a Bézout domain, if and only if A is a ϕ-Bézout ring. □

The following Corollary 3.11 study the transfer of being a ϕ-Bézout ring to trivial extensions.

Corollary 3.11. *Let $R\in\mathcal{H}$ and M be a ϕ-divisible R-module. The following statements are equivalent:*

1. *$R\propto M$ is a ϕ-Bézout ring,*
2. *R is a ϕ-Bézout ring.*

Proof. First $R\propto M$ is a ϕ-ring. In addition, $R\propto M\cong R\bowtie^j J$ where $J=0\propto M\subset Nil(R\propto M)$ and $j:R\hookrightarrow R\propto M$ such that $j(r)=(r,0)$ for all $r\in R$. By Theorem 3.10, $R\propto M$ is a ϕ-Bézout ring if and only if so is R. □

Theorem 3.12. Every ϕ-Bézout ring is a ϕ-Prüfer.

Proof. Straightforward since every Bézout domain is Prüfer. □

The following Example 3.13 establishes that the converse of Theorem 3.12 is not true in general.

Example 3.13. If D be a Noetherian and Prüfer domain which is not a principal ideal domain (for example set $D=\mathbb{Z}\left[\sqrt{-5}\right]$), then $R:=D\propto qf(D)$ is an example of ϕ-Prüfer ring which is not ϕ-Bézout.

Proof. First, we have $R/Nil(R) \cong D$ since $Nil(R) := 0 \propto qf(D)$. Therefore, R is a ϕ-Prüfer ring by Corollary 2.11. Denote that D is never a Bézout domain since it is not a principal ideal domain. Therefore, R is not a ϕ-Bézout ring by Corollary 3.11, as desired. □

It is easy to establish that every ϕ-Bézout ring is nonnil-Noetherian. The following Example 3.14 gives a nonnil-Noetherian ring which isn't a ϕ-Bézout ring.

Example 3.14. The ring $R = \mathbb{Z}[X] \propto qf(\mathbb{Z}[X])$ is a nonnil- Noetherian ring which isn't a ϕ-Bézout ring since $\mathbb{Z}[X]$ is a Noetherian domain which is not a principal ideal domain.

Acknowledgment. The author would like to express his sincere thanks to the referee for his/her helpful suggestions and comments, which have greatly improved this paper.

References

[1] B. Ali and K. Bacem, Nonnil-coherent rings, Beitr. Algebra Geom. 57(2) (2016), 297–305.

[2] D. F. Anderson and A. Badawi, On ϕ-Prüfer rings and ϕ-Bezout rings, Houston J. Math. 30(2) (2004) 331–343.

[3] A. Badawi, On nonnil-Noetherian rings, Comm. Algebra 31(4) (2003), 1669–1677.

[4] A. Badawi, On divided commutative rings, Comm. Algebra 27(3) (1999), 1465–1474.

[5] C. Bakkari, S. Kabbaj and N. Mahdou, Trivial extensions defined by Prüfer conditions, J. Pure Appl. Algebra 214(1) (2010), 53–60.

[6] E. Bastida and R. Gilmer, Overrings and divisorial ideals of rings of the form $D+M$, Michigan Math. J. 20(1) (1973), 79–95.

[7] H. Cartan and S. Eilenberg, Homological algebra, (Princeton 1956).

[8] M. Chhiti, K. Louartiti and M. Tamekkante, Chain conditions in amalgamated algebras along an ideal, Arab J. Math. 2(4) (2013), 403–408.

[9] M. D'Anna and M. Fontana, An amalgamated duplication of a ring along an ideal: the basic properties, J. Algebra Appl. 6(3) (2007), 443–459.

[10] M. D'Anna, C.A. Finocchiaro and M. Fontana, Properties of chains of prime ideals in amalgamated algebras along an ideal, J. Pure Appl. Algebra 214(9) (2010), 1633–1641.

[11] M. D'Anna, C.A. Finocchiaro, M. Fontana, Amalgamated algebras along an ideal. Commutative Algebra and Applications, Proceedings of the Fifth International Fez Conference on Commutative Algebra and Applications, Fez, Morocco, 2008, 155–172. W. de Gruyter Publisher, Berlin (2009).

[12] D. E. Dobbs, Divided rings and going down, Pac. J. Appl. Math. 67(2) (1976), 353–363.

[13] Y. El Haddaoui, H. Kim and N. Mahdou, On nonnil-coherent modules and nonnil-Noetherian modules, Open Math. 20(1) (2022), 1521–1537.

[14] Y. El Haddaoui and N. Mahdou, On ϕ-(weak) global dimension, J. Algebra Appl., to appear.

[15] A. El Khalfi, H. Kim, and N. Mahdou, Amalgamated Algebras Issued from ϕ-Chained Rings and ϕ-Pseudo-Valuation Rings, Bull. Iran. Math. Soc. 47(1) (2021), 1599–1609.

[16] A. El Khalfi, H. Kim and N. Mahdou, Amalgamation extension in commutative ring theory, a survey, Moroccan Journal of Algebra and Geometry with Applications 1(1) (2022), 139–182.

[17] S. Glaz, Commutative Coherent Rings, Lecture Notes Math. 1371, Springer-Verlag, Berlin, 1989.

[18] J. A. Huckaba, Commutative Rings with Zero Divisors, Monographs and Textbooks in Pure and Appl. Math., 117, Dekker, New York, 1988.

[19] S. Kabbaj, Matlis' semi-regularity and semi-coherence in trivial ring extensions: a survey, Moroccan Journal of Algebra and Geometry with Applications 1(1) (2022), 1–17.

[20] H. Kim, N. Mahdou and E. H. oubouhou, ϕ-rings from a module-theoretic point of view: a survey, Moroccan Journal of Algebra and Geometry with Applications (2023), to appear.

[21] H. Kim and F. Wang, Foundations of commutative rings and their modules, Algebra and Applications, 22, Springer, Singapore, 2016.

[22] W. Qi and X. Zhang, Some remarks on nonnil-coherent rings and ϕ-IF rings, J. Algebra Appl. 21(11) (2022), 225–211.

[23] B. Stenström, Rings of Quotients, Grundl. Math. Wiss. 217 (Springer, Berlin, 1975).

[24] A. Tuganbaev, Bezout modules and rings, J. Math. Sci. 163(5) (2009), 596–597.

[25] W. Zhao, On ϕ-flat modules and ϕ-Prüfer rings, J. Korean Math. Soc. 55(5) (2018), 1221–1233.

Moroccan Journal of Algebra
and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 187-196

ISSN: 2820-7114

Title :

On Residually (Completely) Integrally Closed Rings

Author(s):

Ali Tamoussit

Moroccan Journal of Algebra
and Geometry with Applications
Vol. 3(1) (2024), 187–196

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

On Residually (Completely) Integrally Closed Rings

Ali Tamoussit
Department of Mathematics, The Regional Center for Education and Training Professions Souss Massa,
Inezgane, Morocco.
e-mail: *a.tamoussit@crmefsm.ac.ma*

Communicated by El Hassan El Kinani
(Received 09 November 2023, Revised 06 January 2024, Accepted 11 January 2024)

Abstract. All rings considered in this paper are commutative rings with identity. A ring R is called a *residually completely integrally closed* (for short, *RCIC*) *ring* if the integral domain R/P is completely integrally closed, for all prime ideals P of R. The main goal of this paper is to study the behavior of the RCIC property in some distinguished constructions such as homomorphic image, finite direct products, Nagata ring, amalgamated duplications of rings and trivial ring extensions. Our study is then extended to the case of residually integrally closed rings.

Key Words: (Completely) integrally closed, Nagata ring, amalgamated duplication, trivial ring extension.
2010 MSC: Primary 13A15, 13B22; Secondary 13B02, 13F99.

Introduction

An integral domain D with quotient field K is said to be *completely integrally closed* (for short, *CIC*) if for any element x of K and for any nonzero element d of D such that $dx^n \in D$ for all $n \geqslant 1$, x necessarily belongs to D; or equivalently, D contains all elements x of K such that $D[x]$ is contained in a finitely generated D-module. It is well known that completely integrally closed domains are integrally closed, and valuation domains are completely integrally closed if and only if they have rank at most one as proved in [9, Theorem 17.5]. Moreover, it is worth mentioning that the intersection of any family of completely integrally closed domains with the same quotient field, or more generally, contained in some large given field, is still completely integrally closed. Consequently, Krull domains form a (proper) subclass of completely integrally closed domains. Furthermore, we note that a quotient ring of a completely integrally closed domain need not be completely integrally closed. Indeed, the ring $\mathbb{Z}[X]$ is known to be completely integrally closed but the quotient ring $\mathbb{Z}[X]/(X^2+3) \simeq \mathbb{Z}[\sqrt{-3}]$ is not (it is, *a fortiori*, not integrally closed). Motivated by this last fact, we introduce the notion of *residually completely integrally closed ring*. We say that a ring R is residually completely integrally closed (for short, RCIC) if the integral domain R/P is CIC for all prime ideals P of R; or equivalently, R/P is a CIC domain for all non-maximal prime ideals P of R. Trivially, any ring of (Krull) dimension 0 is an RCIC ring. Additionally, RCIC domains form a subclass of CIC domains, and one-dimensional CIC domains must be RCIC. On the other hand, the above example leads also to consider the notion of residually integrally closed rings. In a similar way, a ring R is said to be *residually integrally closed* (for short, *RIC*) if R/P is an integrally closed domain for all prime ideals P of R. From the fact that any completely integrally closed domain is integrally closed, it follows that the class of RIC rings includes that of RCIC rings. But, an RIC ring need not be RCIC. As a matter of fact, take any Prüfer domain that is not CIC (for example, $\mathbb{Z}+X\mathbb{Q}[X]$) because any Prüfer domain is always integrally closed and the property of being Prüfer domain is stable under homomorphic image.

In this paper, we investigate the transfer of the residually (completely) integrally closed to various contexts of constructions. Among other things, we show that any homomorphic image of an RCIC ring is always an RCIC ring (Proposition 1.4), and that the finite direct product of RCIC rings is again an RCIC ring (Proposition 1.6). Also, we characterize RCIC rings issued from Nagata ring, amalgamated duplications of rings and trivial ring extensions (Theorems 1.7 and 1.11). Next we establish some necessary and sufficient conditions for an amalgamation of rings to be an RCIC ring (Proposition 1.12). Then, we generalize the results concerning RCIC rings to the case of RIC rings (Proposition 1.15, Theorem 1.17 and Proposition 1.18). Moreover, we show that being RIC is preserved under flat overring extensions (Proposition 1.19).

Throughout this paper all rings are assumed to be commutative with identity, all modules are unitary and also all homomorphisms are unital. The symbol $\subset$ (resp., $\subseteq$) denotes the proper (resp., large) containment.

1 Residually (completely) integrally closed rings

We begin by recalling some well known facts about CIC domains.
— [9, Theorem 13.1(2)] Any CIC domain is integrally closed.
— [9, Theorem 17.5] A valuation domain is CIC if and only if it has rank at most one.
— [9, Theorem 23.4(3)] One-dimensional Prüfer domains are CIC.
— [9, Exercise 11, page 145] For any subfield L of the quotient field of a CIC domain D, the intersection $D \cap L$ is a CIC domain.
— [13, Corollary 7] For any CIC domain D, the Nagata ring $D(X)$ is also CIC.
— [1, Theorem 2.7] For any extension of integral domains $A \subseteq B$, the integral domain $A + XB[X]$ is CIC if and only if so is A and $A = B$.

We next state the principal definition of this paper as stated in the introduction.

Definition 1.1. A ring R is said to be *residually completely integrally closed* (for short, *RCIC*) if the integral domain R/P is CIC for all prime ideals P of R.

Remark 1.2. (1) It is worth noting that RCIC rings can be characterized by replacing prime ideals with non-maximal prime ideals in the previous definition, since the integral domain R/M is a field for any maximal ideal M of R.

(2) Any RCIC domain is always a CIC domain, and the converse holds for one-dimensional domains. Therefore, every one-dimensional Prüfer domain is an RCIC domain. Consequently, the class of RCIC rings contains Dedekind domains, and, moreover, the integral domain $D + XK[X]$ is never RCIC when D is an integral domain different from its quotient field K.

(3) It is clear that finite rings and 0-dimensional rings are RCIC rings but not conversely. Indeed, the ring of integers $\mathbb{Z}$ is an example of an RCIC ring that is neither finite nor 0-dimensional.

(4) As mentioned in the introduction, the ring $\mathbb{Z}[X]$ is a CIC domain that is not RCIC. Moreover, an RCIC ring is not necessarily a CIC domain. For instance, consider any 0-dimensional ring R that is not a field, such as $R := \prod_{n=0}^{\infty} \mathbb{F}_2$ (the infinite product of copies of $\mathbb{F}_2$). In this case, R is an RCIC ring that is not CIC since any 0-dimensional domain must be a field.

(5) A two-dimensional Prüfer domain is RCIC if and only if it is CIC. To see this, it suffices to prove the reverse implication. Let D be a CIC Prüfer domain of (Krull) dimension 2, and let P be a non-maximal prime ideal of D. Since the height of P is at most one, we have either $D/P \simeq D$ or D/P is a one-dimensional Prüfer domain. In both cases, the integral domain D/P is CIC, and thus, D is an RCIC domain.

The previous remarks yield the following examples:

Example 1.3. (1) For any field K, the rings $K[X]$ and $K[[X]]$ are both RCIC and CIC.

(2) Consider the ring $D := \mathrm{Int}(\mathbb{Z})$, which is the ring of integer-valued polynomials. It is well known that D is a two-dimensional Prüfer domain that is also CIC. Hence, from Remark 1.2(5), we conclude that D is an RCIC ring.

In what follows, we show that the class of RCIC rings is closed under homomorphic images.

Proposition 1.4. *Let R be a ring. Then R is an RCIC ring if and only if so is R/I for each ideal I of R.*

Proof. Assume that R is an RCIC ring, and let Q be a prime ideal of R/I. We have Q is of the form P/I, where P is a prime ideal of R containing I. Since R is an RCIC ring, R/P is a CIC domain, and hence $(R/I)/Q = (R/I)/(P/I) \simeq R/P$ is also a CIC domain. Thus, R/I is an RCIC ring. The converse is trivial. □

Let us provide some additional remarks regarding Proposition 1.4.

Remark 1.5. (1) It is important to note that in Proposition 1.4, the statement "each ideal I of R" cannot be replaced by "each nonzero ideal I of R". To illustrate this, consider a two-dimensional valuation domain (V,M) and a nonzero ideal I of V. According to [9, Theorem 17.5], V is not a CIC domain, and therefore it is not an RCIC ring (since RCIC domains are CIC). However, the prime ideals over I can be M or P, where P is the only height-one prime ideal of V. Thus, V/M is a field and V/P is a one-dimensional valuation domain, and therefore V/I is RCIC for all nonzero ideals I of V.

(2) It is worth mentioning that if $R[X]$ is an RCIC ring, Proposition 1.4 ensures that $R \simeq R[X]/XR[X]$ is also an RCIC ring. However, the converse is not true in general. For instance, $\mathbb{Z}$ is an RCIC ring, but $\mathbb{Z}[X]$ is not (as noticed before).

Next, we study the transfer of the RCIC property to finite direct product of rings.

Proposition 1.6. *Let $\{R_k\}_{1\leqslant k\leqslant n}$ be a finite set of rings. Then $\prod_{k=1}^{n} R_k$ is an RCIC ring if and only if so is each R_k.*

Proof. By induction on n it suffices to prove the result for the case of two rings, say R and S.

Assume that $R\times S$ is an RCIC ring, and let P be a prime ideal of R. Since $P\times S$ is a prime ideal of $R\times S$, we have that $(R\times S)/(P\times S)$ is a CIC domain. Then $R/P \simeq (R\times S)/(P\times S)$ is also a CIC domain, which implies that R is an RCIC ring. By a symmetry argument, we can prove that S is an RCIC ring.

Conversely, assume that R and S are both RCIC rings, and let I be a prime ideal of $R\times S$. We have I is of the form $P\times S$ or $R\times Q$, where P is a prime ideal of R and Q is a prime ideal of S. Since R and S are RCIC rings, R/P and S/Q are CIC domains, and then so are $(R\times S)/(P\times S) \simeq R/P$ and $(R\times S)/(R\times Q) \simeq S/Q$. Therefore, $(R\times S)/I$ is a CIC domain, which implies that $R\times S$ is an RCIC ring. □

The following theorem shows that the RCIC property is conserved between a ring and its Nagata ring. Recall that the *Nagata ring* of a ring R, denoted by $R(X)$, is defined as follow

$$R(X) := \left\{ \frac{f}{g};\ f,g \in R[X] \text{ and } c(g) = R \right\},$$

where X is an indeterminate over R and $c(g)$ is the ideal of R generated by the coefficients of g.

Theorem 1.7. For a ring R, the Nagata ring $R(X)$ is RCIC if and only if so is R.

To prove this result, we require the following two lemmas.

Lemma 1.8. *Let D be an integral domain. Then $D(X)$ is a CIC domain if and only if so is D.*

Proof. The direct implication follows from the equality $D = D(X) \cap K$, where K is the quotient field of D. The converse is proved in [13, Corollary 7]. □

Lemma 1.9 ([10, Theorem 14.1])**.** *Let R be a ring. Then there is a one-to-one correspondance between the prime ideals of R and the prime ideals of $R(X)$ given by $P \longleftrightarrow PR(X)$. Moreover, for each prime ideal P of R, $R(X)/PR(X) \simeq (R/P)(X)$.*

Proof of Theorem 1.7. Assume that $R(X)$ is an RCIC ring, and let P be a prime ideal of R. By Lemma 1.9, we have $PR(X)$ is a prime ideal of $R(X)$, and then $R(X)/PR(X)$ is a CIC domain because $R(X)$ is an RCIC ring. Again by Lemma 1.9, we have $R(X)/PR(X) \simeq (R/P)(X)$. Thus, $(R/P)(X)$ is a CIC domain, forces that R/P is also a CIC domain. Therefore, R is an RCIC ring.

Conversely, assume that R is an RCIC ring, and let Q be a prime ideal of $R(X)$. By Lemma 1.9, there exists a prime ideal P of R such that $Q = PR(X)$ with $R(X)/Q \simeq (R/P)(X)$. Since R is an RCIC ring, R/P is a CIC domain and then it follows from Lemma 1.8 that $(R/P)(X)$ is also a CIC domain. Hence, $R(X)/Q$ is a CIC domain, and therefore, $R(X)$ is an RCIC ring. □

Now, we will investigate the transfer of the RCIC property to trivial ring extensions and amalgamated duplications of rings. For a ring R, we recall the following two notions:

— *The trivial ring extension* of R by an R-module E is the ring denoted by $R \propto E$, whose underlying group is $R \times E$ with multiplication given by $(r,e)(s,f) = (rs, rf + se)$.
— *The amalgamated duplication* of R along an ideal I of R is a subring of $R \times R$, defined as $R \bowtie I := \{(r, r+i);\ r \in R \text{ and } i \in I\}$.

To prove our next theorem, we need to describe the prime ideals of these two constructions. First, we state the following lemma:

Lemma 1.10. *Let R be a ring, E an R-module, and I an ideal of R. We have the following:*

1. *Each prime ideal $\mathcal{P}$ of $R \propto E$ is of the form $\mathcal{P} = P \propto E$ for some prime ideal P of R. Moreover, $(R \propto E)/\mathcal{P} \simeq (R/P)$.*

2. *Each prime ideal $\mathcal{P}$ of $R \bowtie I$ is of the form $\{(i, i+r);\ i \in I \text{ and } r \in P\}$ or $\{(i+r, i);\ i \in I \text{ and } r \in P\}$, where $P = \mathcal{P} \cap R$. Moreover, in both cases, $(R \bowtie I)/\mathcal{P} \simeq (R/P)$.*

Proof. See [2, Theorem 3.2(2)] and [7, Proposition 2.2]. □

Theorem 1.11. For any ring R, the following statements are equivalent:

1. R is RCIC;

2. The trivial ring extension $R \propto E$ is an RCIC ring, for every R-module E;

3. The amalgamated duplication $R \bowtie I$ is an RCIC ring, for every ideal I of R.

Proof. The implications (2) ⇒ (1), and (3) ⇒ (1) follow directly from Proposition 1.4 and the fact that $(R \propto E)/(\{0\} \propto E) \simeq R$ and $(R \bowtie I)/(\{0\} \times I) \simeq R$.

(1) ⇒ (2) Assume that R is an RCIC ring and let Q be a prime ideal of $R \propto E$. By Lemma 1.10(1), there is a prime ideal P of R such that $Q = P \propto E$. Since R is an RCIC ring, R/P is a CIC domain and then so is $(R \propto E)/Q = (R \propto E)/(P \propto E) \simeq (R/P)$. Thus $R \propto E$ is an RCIC ring.

(1) ⇒ (3) Assume that R is an RCIC ring and let Q be a prime ideal of $R \bowtie I$. Set $P := Q \cap R$. Then it follows from Lemma 1.10(2) that Q is of the form $\{(i, i+r);\ i \in I \text{ and } r \in P\}$ or $\{(i+r, i);\ i \in I \text{ and } r \in P\}$, and so, in both cases, we have: $(R \bowtie I)/Q \simeq (R/P)$. Since R is an RCIC ring, R/P is a CIC domain and then so is $(R \bowtie I)/Q$. Therefore $R \bowtie I$ is an RCIC ring. □

To treat the transfer of the RCIC property to amalgamation construction, we need first to recall the definition of amalgamated algebras along an ideal as presented in [5].

Let R and S be two rings, $f : R \to S$ a ring homomorphism and J an ideal of S. The following subring of $R \times S$:

$$R \bowtie^f J = \{(r, f(r) + j);\ r \in R \text{ and } j \in J\},$$

is called the *amalgamation* of R with S along J with respect to f. Notably, if $R = S$, $f = \iota$ (the identity of R) and $J = I$, then $R \bowtie^f J$ corresponds precisely to the amalgamated duplication of R along I, denoted as $R \bowtie I$. For more details on amalgamated algebra, the reader may consult the survey paper [8].

We now present the following proposition, which provides necessary and sufficient conditions for an amalgamation to be an RCIC ring. Here, $\mathrm{Nil}(R)$ and $\mathrm{Jac}(R)$ denote the nilradical and the Jacobson radical of a ring R, respectively.

Proposition 1.12. *Let R and S be two rings, $f : R \to S$ a ring homomorphism, and J an ideal of S. The following statements hold:*

1. *If $R \bowtie^f J$ is an RCIC ring then R and $f(R) + J$ are RCIC rings.*
2. *If $f^{-1}(J) = \{0\}$, then $R \bowtie^f J$ is an RCIC ring if and only if so is $f(R) + J$.*
3. *If either $J \subseteq \mathrm{Nil}(S)$ or $J \subseteq \mathrm{Jac}(S)$, then $R \bowtie^f J$ is an RCIC ring if and only if so is R.*

Proof. (1) Assume that $R \bowtie^f J$ is an RCIC ring. By Proposition 1.4, we deduce that $(R \bowtie^f J)/(0 \times J) \simeq R$ and $(R \bowtie^f J)/(f^{-1}(J) \times 0) \simeq f(R) + J$ are RCIC rings.

(2) If $f^{-1}(J) = 0$, then it follows from [5, Proposition 5.1(3)] that $R \bowtie^f J \simeq f(R) + J$.

(3) Assume that $J \subseteq \mathrm{Nil}(S)$ or $J \subseteq \mathrm{Jac}(S)$. From [6, Proposition 2.6], we infer that every prime ideal of $R \bowtie^f J$ is of the form $P \bowtie^f J$ for some prime ideal P of R. Since $(R \bowtie^f J)/(P \bowtie^f J) \simeq R/P$, $(R \bowtie^f J)/(P \bowtie^f J)$ is a CIC domain if and only if so is R/P. Thus, we conclude that $R \bowtie^f J$ is an RCIC ring if and only if so is R, as desired. □

We next investigate the transfer of the residually integrally closed property to some remarkable ring extensions. First, we introduce the notion of residually integrally closed rings.

Definition 1.13. A ring R is said to be *residually integrally closed* (for short, *RIC*) if, for each prime ideal P of R, the integral domain R/P is integrally closed.

Remark 1.14. (1) Note that the class of RIC rings includes RCIC rings and Prüfer domains.

(2) It is clear that RIC domains are integrally closed. Conversely, one-dimensional integrally closed domains are RIC.

(3) Any RCIC ring is RIC, but the converse is not true in general. For instance, let $D = \mathbb{Z} + X\mathbb{Q}[X]$ which is a Prüfer domain and then it is RIC. However, D is not RCIC since it is not CIC, as asserted in [1, Theorem 2.7].

In the following, we consider transferring the RIC property to localization and finite direct product.

Proposition 1.15. *For any two rings R and S, we have:*

1. *If R is an RIC ring, then $T^{-1}R$ is an RIC ring for any multiplicative subset T of R.*
2. *The direct product $R \times S$ is an RIC ring if and only if so are R and S.*

Proof. (1) Assume that R is an RIC ring, and let T be multiplicative subset T of R and Q a prime ideal of $T^{-1}R$. Then Q is of the form $T^{-1}P$ for some prime ideal P of R with $P \cap T = \emptyset$. Since R is an RIC ring, R/P is an integrally closed domain. As the property of being integrally closed is stable under localization, $\overline{T}^{-1}(R/P)$ is integrally closed, where $\overline{T}$ denotes the natural image of T in R/P, and hence $T^{-1}R/Q$ is also integrally closed because $T^{-1}R/T^{-1}P \simeq \overline{T}^{-1}(R/P)$. Thus, $T^{-1}R$ is an RIC ring.

(2) The proof of this statement is similar to that of Proposition 1.6. □

As a quick consequence of Proposition 1.15(2), we have the following:

Corollary 1.16. *Let $\{R_k\}_{1\leqslant k\leqslant n}$ be a finite set of rings. Then $\prod_{k=1}^{n} R_k$ is an RIC ring if and only if so is each R_k.*

In the remainder of this section, we examine the transfer of the RIC property in various ring extensions.

Theorem 1.17. For any ring R, the following statements are equivalent:

1. R is an RIC ring;
2. R is an locally RIC ring (that is, R_P is an RIC ring, for every prime ideal P of R);
3. R_M is an RIC ring, for every maximal ideal M of R;
4. R/I is an RIC ring, for every ideal I of R;
5. The Nagata ring $R(X)$ is RIC;
6. The trivial ring extension $R \propto E$ is an RIC ring, for every R-module E;
7. The amalgamated duplication $R \bowtie I$ is an RIC ring, for every ideal I of R.

Proof. The proof of the equivalences (1) ⇔ (4) ⇔ (5) ⇔ (6) ⇔ (7) are similar to the case of RCIC rings.

The implication (1) ⇒ (2) follows from Proposition 1.15(1), and (2) ⇒ (3) is trivial.

To prove (3) ⇒ (1), assume that R_M is an RIC ring for every maximal ideal M of R, and let P be a non-maximal prime ideal of R. Using the fact that the integrally closed property is a local property, we need to check that $(R/P)_M$ is integrally closed for all maximal ideals M of R/P. Let M be a maximal ideal of R/P. Then $M = m/P$ for some maximal ideal m of R containing P. Since R_m is RIC, we have $(R/P)_M \simeq R_m/P_m$ is an integrally closed domain, which implies that R/P is also integrally closed. Thus, R/I is an RIC ring, and this completes the proof. □

Proposition 1.18. *Let R and S be two rings, $f : R \to S$ a ring homomorphism, and J an ideal of S. The following statements hold:*

1. *If $R \bowtie^f J$ is an RIC ring then so are the rings R and $f(R)+J$.*
2. *If $f^{-1}(J) = \{0\}$, then $R \bowtie^f J$ is an RIC ring if and only if so is $f(R)+J$.*
3. *If either $J \subseteq \mathrm{Nil}(S)$ or $J \subseteq \mathrm{Jac}(S)$, then $R \bowtie^f J$ is an RIC ring if and only if so is R.*
4. *If J is a maximal ideal and $R \bowtie^f J$ is an RIC ring, then R and S are RIC rings.*

Proof. The statements (1), (2) and (3) are similar to the case of RCIC rings.

(4) Assume that J is a maximal ideal and $R \bowtie^f J$ is an RIC ring. From statement (1), we conclude that R is RIC. Furthermore, since J is maximal, we have $\overline{Q}^f := \{(r, f(r)+j);\ r \in R,\ j \in J \text{ and } f(r)+j \in Q\}$ is a prime ideal of $R \bowtie^f J$ for each prime ideal Q of S. Using Proposition 1.15(1), we deduce that $(R \bowtie^f J)_{\overline{Q}^f} \simeq S_Q$ is an RIC ring, and therefore by Theorem 1.17, S is an RIC ring. □

Inspired by [12], we can establish that the RIC property is preserved under flat overrings. To prove this, we need to recall the notion of generalized transform of a ring with respect to a generalized multiplicative system.

Let R be a ring with the total quotient ring K, I an ideal of R, and $\mathcal{S}$ a generalized multiplicative system of R, i.e., $\mathcal{S}$ is a multiplicative set of ideals of R. The *$\mathcal{S}$-transform* of R (or the *generalized transform* of R with respect to $\mathcal{S}$) is an overring $R_{\mathcal{S}} := \{x \in K;\ xA \subseteq R \text{ for some } A \in \mathcal{S}\}$. Moreover, $I_{\mathcal{S}} := \{x \in K;\ xA \subseteq I \text{ for some } A \in \mathcal{S}\}$ is an ideal of $R_{\mathcal{S}}$ containing I. Here, an overring of a ring R refers to a subring of K that contains R.

Proposition 1.19. *Let R be a ring. Then R is an RIC ring if and only if any flat overring T of R is RIC.*

Proof. We will only prove the direct implication. Assume that R is an RIC ring, and let T be a flat overring of R and M a maximal ideal of T. Set $P := M \cap R$. By [3, Theorems 1.1 and 1.3], there exists a generalized multiplicative system $\mathcal{S}$ of R such that $T = R_{\mathcal{S}}$ and $M = P_{\mathcal{S}}$. Furthermore, we can suppose that any generalized multiplicative system of R is saturated, as stated in [4, Proposition 4.6]. Then it follows from [4, Theorem 4.12] that $T_M = (R_{\mathcal{S}})_{P_{\mathcal{S}}} \simeq R_P$. Since R is an RIC ring, R_P is also an RIC ring, and hence T_M is an RIC ring. Thus by Theorem 1.17, we conclude that T is an RIC ring, as desired. □

Remark 1.20. It is worth noting that the previous proposition cannot be extended to the general case of a flat ring extension. Indeed, consider the ring $\mathbb{Z}$, which is an RIC domain. As mentioned in the introduction, $\mathbb{Z}[X]$ is not an RIC. However, the ring extension $\mathbb{Z} \hookrightarrow \mathbb{Z}[X]$ is flat.

2 Further generalization

In this section, we aim to show that the previously established results hold in a more general context by considering a specific property $\mathcal{X}$ of integral domains instead of the notion of a (completely) integrally closed domain. To do this, we need to introduce the following definitions:
— A ring R is said to be *residually $\mathcal{X}$* if the integral domain R/P has the property $\mathcal{X}$, for all prime ideals P of R.
— A ring R is said to be *totally $\mathcal{X}$* if R_P is a residually $\mathcal{X}$ ring, for all prime ideals P of R.
— We say that $\mathcal{X}$ *has a good behavior under integral extensions* if, for any integral extension of domains $R \subseteq S$, we have R is an $\mathcal{X}$ domain if and only if so is S.
— We say that $\mathcal{X}$ *has a good behavior under Nagata ring*, if the integral domains D and $D(X)$ simultaneously have the same property $\mathcal{X}$.

Theorem 2.1. Let $\mathcal{X}$ denote a property of integral domains, and let R and S be two rings. We have:

1. If R is a residually $\mathcal{X}$ ring then so is R/I, for each ideal I of R.

2. The direct product $R \times S$ is a residually $\mathcal{X}$ ring if and only if so are R and S.

3. The following statements are equivalent:

   (a) R is a residually $\mathcal{X}$ ring;

   (b) The trivial ring extension $R \propto E$ is a residually $\mathcal{X}$ ring, for every R-module E;

   (c) The amalgamated duplication $R \bowtie I$ is a residually $\mathcal{X}$ ring, for every ideal I of R.

4. Assume that $\mathcal{X}$ is stable under localization. If R is a residually $\mathcal{X}$ ring then so is $T^{-1}R$ for each multiplicative subset T of R.

5. Assume that $\mathcal{X}$ has a good behavior under integral extensions. If $R \subseteq S$ is an integral extension of rings, then R is a residually $\mathcal{X}$ ring if and only if so is S.

6. Let $f : R \to S$ be a ring homomorphism and J an ideal of S.

   (a) If $R \bowtie^f J$ is a residually $\mathcal{X}$ ring then so are the rings R and $f(R)+J$.

   (b) If either $J \subseteq \mathrm{Nil}(S)$ or $J \subseteq \mathrm{Jac}(S)$, then $R \bowtie^f J$ is a residually $\mathcal{X}$ ring if and only if so is R.

   (c) If $\mathcal{X}$ has a good behavior under integral extensions, then $R \bowtie^f J$ is a residually $\mathcal{X}$ ring if and only if so are the rings R and $f(R)+J$.

7. Assume that $\mathcal{X}$ has a good behavior under Nagata ring. Then R is a residually $\mathcal{X}$ ring if and only if so is $R(X)$.

Proof. The proofs of (1), (2), and (3) are similar to those of Propositions 1.4 and 1.6, and Theorem 1.11.

The proof of (4) is similar to the case of RIC rings.

(5) Assume that R is a residually $\mathcal{X}$ ring and let Q be a prime ideal of S. Since $R/(Q \cap R) \subseteq S/Q$ is an integral extension and $R/(Q \cap R)$ is an $\mathcal{X}$ domain, it follows that S/Q is also an $\mathcal{X}$ domain. Thus, S is a residually $\mathcal{X}$ ring.

Conversely, assume that S is a residually $\mathcal{X}$ ring and let P be a prime ideal of R. By Lying-Over, there is a prime ideal Q of S such that $P = Q \cap R$, and so S/Q is an $\mathcal{X}$ domain. Moreover, since $R/P \subseteq S/Q$ is an integral extension, R/P is an $\mathcal{X}$ domain, and therefore R is a residually $\mathcal{X}$ ring.

(6) The statements (a) and (b) are similar to those of Proposition 1.12.

(c) It is well known that the ring $A \times (f(A)+J)$ is integral over $A \bowtie^f J$, as asserted in [6, Lemma 3.3]. Then by statement (5), $A \bowtie^f J$ is a residually $\mathcal{X}$ ring if and only if so is $A \times (f(A)+J)$, and thus the conclusion follows from statement (2).

(7) The proof of this statement is similar to that of Theorem 1.7. □

Theorem 2.2. Let $\mathcal{X}$ denote a property of integral domains which is stable under localization. For any ring R, the following statements are equivalent:

1. R is a totally $\mathcal{X}$ ring;

2. R_P is a totally $\mathcal{X}$ ring, for every prime ideal P of R;

3. R_M is a totally $\mathcal{X}$ ring, for every maximal ideal M of R;

4. Any flat overring T of R is totally $\mathcal{X}$ (in particular, every localization of a totally $\mathcal{X}$ ring is totally $\mathcal{X}$);

5. R/I is a totally $\mathcal{X}$ ring, for every ideal I of R;

6. The trivial ring extension $R \propto E$ is a totally $\mathcal{X}$ ring, for every R-module E;

7. The amalgamated duplication $R \bowtie I$ is a totally $\mathcal{X}$ ring, for every ideal I of R.

Proof. The proof of (1) ⇒ (2) is similar to that of Proposition 1.15(1).

The implications (2) ⇒ (3), (4) ⇒ (1), and (5) ⇒ (1) are straightforward.

(3) ⇒ (1) Assume that R_M is a totally $\mathcal{X}$ ring for every maximal ideal M of R and let P be a nonmaximal prime ideal of R. Then there exists a maximal ideal M of R such that $P \subset M$. So, from the fact that the localization of a residually $\mathcal{X}$ ring is still residually $\mathcal{X}$ (see Theorem 2.1(4)), it follows that $(R_M)_{PR_M} = R_P$ is a residually $\mathcal{X}$ ring, and hence R is a totally $\mathcal{X}$ ring.

(1) $\Rightarrow$ (4) Assume that R is a totally $\mathcal{X}$ ring, and let T be a flat overring of R and M a maximal ideal of T and set $P := M \cap R$. Then by [3, Theorem 1.3], there exists a generalized multiplicative system $\mathcal{S}$ of R such that $T = R_{\mathcal{S}}$ and $AT = T$ for all $A \in \mathcal{S}$. Also by [3, Theorem 1.1] we have $M = P_{\mathcal{S}}$. It follows from [4, Proposition 4.6] that we may assume that any generalized multiplicative system of R is saturated in this situation. Thus by [4, Theorem 4.12] there is an isomorphism $(R_{\mathcal{S}})_{P_{\mathcal{S}}} \simeq R_P$. Since R is totally $\mathcal{X}$, R_P is a residually $\mathcal{X}$ ring and then so is T_M. Therefore $T = R_{\mathcal{S}}$ is totally $\mathcal{X}$ by the implication (3) $\Rightarrow$ (1).

(1) $\Rightarrow$ (5) Assume that R is a totally $\mathcal{X}$ ring and let M be a maximal ideal of R/I. Then $M = m/I$ for some maximal ideal m of R containing I. Since R is totally $\mathcal{X}$ and m is a maximal ideal of R, R_m is residually $\mathcal{X}$, and then it follows from Theorem 2.1(1) that $(R/I)_M \simeq R_m/I_m$ is also a residually $\mathcal{X}$ ring. Therefore by the implication (3) $\Rightarrow$ (1), R/I is a totally $\mathcal{X}$ ring.

The implications (6) $\Rightarrow$ (1) and (7) $\Rightarrow$ (1) follow from Theorem 2.1(1) and the fact that $(R \propto E)/(\{0\} \propto E) \simeq R$ and $(R \bowtie I)/(\{0\} \times I) \simeq R$.

(1) $\Rightarrow$ (6) Assume that R is a totally $\mathcal{X}$ ring and let M be a maximal ideal of $R \propto E$. By [2, Theorem 3.2], there is a maximal ideal m of R such that $M = m \propto E$. Since R is a totally $\mathcal{X}$ ring, R_m is a residually $\mathcal{X}$ ring and then so is $(R \propto E)_M = (R \propto E)_{(m \propto E)} \simeq R_m \propto E_m$ by Theorem 2.1(4). Thus, $R \propto E$ is a totally $\mathcal{X}$ ring.

(1) $\Rightarrow$ (7) Assume that R is a totally $\mathcal{X}$ ring and let M be a maximal ideal of $R \bowtie I$. Then by [7, Proposition 2.2], M is of the form $\{(i, i+r);\ i \in I \text{ and } r \in m\}$ or $\{(i+r, i);\ i \in I \text{ and } r \in m\}$ with $m = M \cap R$. So, we discuss the following two cases:

Case 1: $I \subseteq m$. Since R is totally $\mathcal{X}$, R_m is residually $\mathcal{X}$ and then it follows from Theorem 2.1(3) and [7, Proposition 2.2] that $(R \bowtie I)_M \simeq R_m \bowtie I_m$ is residually $\mathcal{X}$.

Case 2: $I \nsubseteq m$. Then by Theorem 2.1(3) and [7, Proposition 2.2], $(R \bowtie I)_M \simeq R_m$ is residually $\mathcal{X}$ because R_m is residually $\mathcal{X}$.

Therefore, $R \bowtie I$ is a totally $\mathcal{X}$ ring. □

References

[1] D.D. Anderson, D.F. Anderson and M. Zafrullah, Rings between $D[X]$ and $K[X]$, Houston J. Math. **17** (1991), 109–129.

[2] D.D. Anderson and M. Winderes, Idealization of a module, J. Comm. Algebra **1** (2009), 3–56.

[3] J.T. Arnold and J.W. Brewer, On flat overrings, ideal transforms and generalized transforms of a commutative ring, J. Algebra **18** (1971), 254–263.

[4] H.S. Butts and C.G. Spaht, Generalized quotient rings, Math. Nachr. **53** (1972), 181–210.

[5] M. D'Anna, C.A. Finocchiaro and M. Fontana, Amalgamated algebras along an ideal, in: *Commutative Algebra and Applications*, Walter de Gruyter, Berlin, (2009), 155–172.

[6] M. D'Anna, C.A. Finocchiaro and M. Fontana, Properties of chains of prime ideals in an amalgamated algebra along an ideal, J. Pure Appl. Algebra **214** (2010), 1633–1641.

[7] M. D'Anna and M. Fontana, The amalgamated duplication of a ring along a multiplicative-canonical ideal, Ark. Mat. **45** (2007), 241–252.

[8] A. El Khalfi, H. Kim and N. Mahdou, Amalgamation extension in commutative ring theory: a survey, Moroccan Journal of Algebra and Geometry with Applications **1** (1) (2022), 139–182.

[9] R. Gilmer, *Multiplicative Ideal Theory*, Queen's Papers in Pure and Appl. Math., vol. 90, Queen's University, Kingston, Ontario, (1992).

[10] J.A. Huckaba, *Commutative Rings with Zero Divisors*, Marcel Dekker, 1988.

[11] S. Kabbaj, Matlis' semi-regularity and semi-coherence in trivial ring extensions: a survey, Moroccan Journal of Algebra and Geometry with Applications **1** (1) (2022), 1–17.

[12] H. Kim, O. Ouzzaouit and A. Tamoussit, Noetherian-like properties and zero-dimensionality in some extensions of rings, Afr. Mat. **34** (3) (2023), https://doi.org/10.1007/s13370-023-01083-3

[13] T.G. Lucas, Characterizing when $R(X)$ is completely integrally closed, in: *Factorization in Integral Domains*, Lect. Notes in Pure Appl. Math. **189**, Marcel Dekker, New York (1997), pp. 401–415.

Moroccan Journal of Algebra and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 197-218

ISSN: 2820-7114

Title :

Fast multiplication algorithm for square sparse matrices. Application to images processing

Author(s):

Ștefan-Daniel Achirei, Antonio Lasanta, Laiachi El Kaoutit, and Carlos Rodriguez Dominguez

Moroccan Journal of Algebra and Geometry with Applications
Vol. 3(1) (2024), 197–218

Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

Fast multiplication algorithm for square sparse matrices. Application to images processing

Ştefan-Daniel Achirei[1], Antonio Lasanta[2], Laiachi El Kaoutit[3] and Carlos Rodriguez Dominguez[4]

[1] "Gheorghe Asachi" Technical University of Lasi.

e-mail: *stefan.achirei@cs.tuiasi.ro*

[2] Universidad de Granada, Departamento de Álgebra. Facultad de Educación, Econonía y Tecnología de Ceuta. Cortadura del Valle, s/n. E-51001 Ceuta, Spain. Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071 Granada, Spain.

e-mail: *alasanta@ugr.es*

[3] Universidad de Granada, Departamento de Álgebra and IMAG. Facultad de Ciencias s/n. E-18071 Granada, Spain.

e-mail: *kaoutit@ugr.es*

[4] Universidad de Granada, Departamento de Lenguajes y Sistema Informáticos. Facultad de Educación, Econonía y Tecnología de Ceuta. Cortadura del Valle, s/n. E-51001 Ceuta, Spain

e-mail: *carlosrodriguez@ugr.es*

Communicated by Najib Mahdou

(Received 14 November 2023, Revised 10 January 2024, Accepted 14 January 2024)

Abstract. We provide a fast multiplication algorithm for a certain class of square sparse matrices that are commonly used in image processing. In order to compare our approach with the standard algorithms, we will use two working sets: *SuitSparse Matrix Collection* and *Anonymous MRI Brain Scan Images Database*. The evaluations show that our algorithm has up to 75 times better time-efficiency and an improvement between 21% and 96% of memory-efficiency. *

Key Words: Digital Images, Image Processing, Gray-scale, Sparse Matrices, Software Algorithms, Iterative Algorithms.
2010 MSC: 65Y20, 68W01, 68W35.

Introduction

When solving different problems from economy, technical fields, social environment, optimization, as well as in modelling or simulating industrial and specially technological processes (such as images processing), it is necessary to determine the mathematical model that describes the problem itself. This leads to finding some mathematical object that usually involves binary operations with linear algebraic equation systems in which each of the associated coefficient matrix is a *sparse matrix* (i.e., contains relatively few nonzeros entries). Sparse matrix algorithms have become a must-have in many fields for improving computational efficiency and resource optimization. They have completely changed the way numerical computations are done. These algorithms take advantage of the fact that matrices are naturally sparse to speed up operations, which has big benefits for memory use and computing speed. For a deeper and comprehensive reading, we refer to the few papers [2, 4, 8, 1, 3, 5].

In medical imaging, sparse matrix operations are often used to speed up the process of making high-resolution images from low-resolution or incomplete measurements. A lot of medical images have a sparse representation in a transform domain, like wavelets, Fourier, or dictionary learning. Sparse matrix operations take advantage of this. Sparse matrix operations can cut down on the number of measurements needed for accurate reconstruction by putting a sparsity constraint on the image coefficients. This saves time, lowers noise, and improves image quality. Along with other methods like parallel imaging, compressed sensing, and deep learning, sparse matrix operations can also be used together to make medical image reconstruction even better. For more details on this application of spares matrices, see [14, 15, 16, 12, 9, 10, 11] among other contributions.

Now, from the practical point of view the analysis of that equation systems produces very large mathematical models that may involve linear algebraic equation systems that can include thousands of equations. Consequently, the mathematical models of many real processes produce a large number of variables and constraints which contributes to the sparsity phenomenon of a matrix: Most of

*Research partially supported by the grant MTM2016-77033-P, PID2021128970OA-I00, PID2020-116567GB-C22 from the Spanish Ministerio de Economía y Competitividad MCIN/AEI/10.13039/501100011033 and the European Union FEDER.

the entries are zero and are not connected between each other. From a computational perspective, such systems require a lot of memory to represent them, and a lot of time to provide a solution to the equation system. Thus, from this point of view, a "fast implementation" of binary operations with sparse matrices seems to be the most desirable task to achieve.

Taking notice of the sparsity property of the matrix can result in a more efficient approach, implying the development of specific applications that use a special data representation/structure. This will save memory and reduce the run time.

The main purpose of this paper is to give an algebraic and implementation approaches to an efficient computer representation of square sparse matrices. This format will be referred to as *Sparse Matrix Format* (*SMF* for short). The second objective is to introduce and implement efficient algorithms for the associated binary arithmetic operations to this particular matrix format: multiplication, transposing, summation and subtraction.

Moreover, the SMF proposal is compared, in terms of memory use and run time, to the classical matrix representation and associated operations to ensure its efficiency and its practical applicability, especially in images processing. More precisely, considering the different existing approaches to find an efficient memory format for large sparse matrices, this note proposes a different representation format, that can be seen as a combination of the already existing formats in the literature. Roughly speaking, there are two classical approaches for representing sparse matrices:

- The static approach, in which the memory allocation is done in the compilation phase. This approach assumes that the programmer knows with a good precision the maximum number of non-zero entries
- The dynamic approach, in which the memory allocation is done during the execution phase of program. In this case it is not necessary to know the number of non-zero elements. This approach is, consequently, the one that we have implemented.

Usually for identifying the non-zero entries two indices are used, for *row* and for *column*. Firstly, we propose a way to use only one index which contains information for both row and column number (in some sense we "linearise" the matrix, converting it to a "one dimensional" array). For each non-zero entry it is attached an integer number, an *aggregated* index, from which both the row and the column can be determined. In this way the arithmetic binary operations on SMF matrices become more intuitive and easily handleable, specially the multiplication one.

To sum up the achievements to be presented in this work, a square sparse matrix of order n, with N number of non-zero entries, the *SMF* data structure stores in memory $2N+n$ entries in comparison with the classical matrix format that stores n^2 numbers which is greater than or equal to N. Thus for matrices with N smaller than or equal to $\frac{n(n+1)}{2}$ the SMF uses less or equal memory compared to the classical matrix format. In practice the big sparse matrices have less than 3% (see [7]) of the entries non-zero, so there is an obvious improvement in memory use. In terms of run time, in image processing we have been able to achieve up to 75 times better timings.

The paper is organized as follows. Section 1 contains basic illustrative examples of sparse matrices. In Section 2, we give the algebraic foundation behind the proposed algorithm that will be used in the forthcoming sections. The most striking idea here is perhaps the interpretation of a square matrix cells, as arrows in certain groupoid of pairs[†] (for more details, see Remarks 2.2 and 2.3 below), which leads us to introduce the above terminology of *Sparse matrix Format*. The ordinary arithmetic operations: summation, multiplication and transpose are then implemented, using C++ programming language, in Section 3. Lastly, Section 4 is devoted to the efficiency of the implemented algorithms applied to sparse matrices extracted from large existing image databases. To do so we will use two working sets: The first one is the set of *SuitSparse Matrix Collection* from *The University of Florida Sparse Matrix Set* taken from [13], and the second one is an *Anonymous MRI Brain Scan Images Database (University of Granada)* taken from public health sources.

1 Rappels on Sparse Matrices

The contains of this section is merely illustrative and perhaps folkloric, for more details we refer to [5]. All matrices handled below are matrices with entries in the rig of positive integers $\mathbb{N}$ (except for the forthcoming sections, where we use matrices with entries in the ring of integers).

1.1 Basic definitions and properties

Roughly speaking, a matrix is called a *sparse matrix* if most of its entries are zero, in comparison with a *dense matrix*, which has the majority of its entries different than zero. In order to determine if a matrix is sparse or not, it is required to introduce the notion of *sparsity* of a matrix. Following [7], the resulting rational number from the division of the number of all null-valued entries and the total number of entries, is termed the *sparsity of the matrix*. The *density* of a matrix is defined as the division between non-zero and total number of entries. In other words *sparsity* is equal to 1 minus *density* of a matrix. Using these definitions, a matrix is said to be *sparse*, provided its sparsity is greater than $\frac{1}{2}$.

In practical applications there are encountered large sparse matrices with non-zero entries between 0.15% and 3%, the sparsity varies form 0.97 to 0.9985.

[†] See [6, Definition 1.4 and Example 1.11] for the precise definition of this mathematical object.

Table 1: List or Groups

1	9	2	8	3	4	5	6	7	3	2	1	7	1	3	6	2	1
1^{st} row		2^{nd} row			3^{rd} row		4^{th} row		5^{th} row			6^{th} row		7^{th} row		8^{th} row	

Example 1.1. The 5 × 5-matrix exhibited in equation (1) below, provides an example of square matrix of order $n = 5$ with $N = 7$ non-zero entries out of 25. The *sparsity* and *density* are respectively calculated, as shown in equations (2) and (3). Since its *sparsity* value is greater than 0.5, then the matrix is considered *sparse*.

$$A = \begin{pmatrix} 0 & 0 & A_{13} & 0 & 0 \\ 0 & 0 & A_{23} & A_{24} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & A_{42} & A_{43} & 0 & 0 \\ A_{51} & 0 & 0 & 0 & A_{55} \end{pmatrix} \tag{1}$$

$$\begin{aligned} sparsity(A) &= \frac{(n^2) - N}{n^2} = \frac{(5^2) - 7}{5^2} \\ &= \frac{18}{25} = 0.72 > 0.5 \end{aligned} \tag{2}$$

$$density(A) = \frac{N}{n^2} = \frac{7}{25} = 0.28. \tag{3}$$

1.2 Some classes of sparse matrices

Following the literature there at least three classes of sparse matrices, which we briefly recall below:

1.2.1 Band matrix

Roughly speaking, in band sparse matrices the non-zero values are somehow grouped around the main diagonal (see [5, §8.2]), like the following 8 × 8-matrix A:

$$A = \begin{pmatrix} 1 & 9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 8 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 6 & 7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \end{pmatrix} \tag{4}$$

is a sparse matrix in which the non-zero elements are grouped on or near the main diagonal. A *list* (or *group*) is a chain of non-zero consecutive entries in a given row. In this way the matrix A can be represented by its groups. Thus, the groups of this matrix are written in Table 1.

1.2.2 Diagonal matrix

Diagonal matrices only have non-zero entries on the main or upper diagonal. For example the following matrix A:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7 \end{pmatrix} \tag{5}$$

1.2.3 Permutation matrix

The permutation matrix has on each row or column only one non-zero entry whose value is 1, and all the rest being zeros. This matrix is usefull in algebraic operations to permute the coordinates according to a previously established model. For instance, considering following matrix M:

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

and the vector $X = (1,2,3,4,5)$. By multiplying the two, a vector $MX = (4,3,1,2,5)$ is obtained. The entries of X were rearranged according to the values of one from matrix M. Of course, by multiplying M on the right by any square matrix, returns a matrix with the same entries with reordered columns, as before.

Remark 1.2. As the reader could realize, the information carried by sparse matrices is concentrated in few of its entries, so that global binary operations with these matrices generate quite a lot off useless partial operations that should be stored and remaining in the memory without any benefit.

1.3 Sparse matrices in Computer Science

Sparse matrices are encountered in modeling or simulating processes from different fields like: industry, economics, technology, social, etc (see [7]). Sparse matrices are the core of solving systems of linear equations. Therefore, some fields that widely use linear algebra represented by sparse matrices are:

- *modeling and simulation of large-scale systems*: described by thousands of linear algebraic equations in form of large sparse matrices
- *computer graphics*: adding and multiplying matrices is the most common operation in image processing
- *recommendation systems or search engines*: for instance links on the web are described in a sparse matrix, element *(i, j)* is non-zero if web page i has a link to web page j. Examples of this such implementations: *Google Ranking System* or *Facebook Friend Relations*.
- *machine learning*: in applied machine learning large sparse matrices are often used, for instance the correlation matrix or stochastic matrix whose edges define a relation between data points.

The last two examples are in fact based on the incidence matrix of a given directed graph.

2 Sparse matrix format and the arithmetic foundations

Next it is detailed an algebraic definition and then implementation of *Sparse Matrix Format*, regarding only square matrices (see the last paragraph of Remark 2.3 for the general case). The arithmetic operations with this new format are also treated in this section.

2.1 The SMF format of square matrices

Let $n \in \mathbb{N} \setminus \{0\} := \mathbb{N}^*$ be a non-zero positive integer. We denote by $\mathbb{N}'_n := \{1, \cdots, n\}$ the set of integers $1 \leq j \leq n$. There are bijective maps:

$$\begin{tikzcd} \mathbb{N}'_n \times \mathbb{N}'_n \arrow[r, "\psi_n"] & \mathbb{N}'_{n^2} \\ (i,j) \arrow[r, mapsto] & (i-1)n + j \\ (c(l)+1, r(l)) & l, \arrow[l, mapsto] \end{tikzcd} \tag{7}$$

where the positive integers $c(l)$ and $r(l)$, are uniquely computed from a given number l with $1 \leq l \leq n^2$, by using as follows the Euclidean algorithm of division: (1) If $1 \leq l < n$, then the Euclidean algorithm says that $l = 0.n + l$. So we have $c(l) = 0$ and $r(l) = l$; (2) If l is a multiple of n then $l = kn$, for some k, and we take $c(l) = k - 1$ while $r(l) = n$; (3) If $n < l < n^2$ and l is not a multiple of n, then $c(l)$ and $r(l)$ are exactly the quotient and the reminder after applying the Euclidean algorithm of division of l over n.

In other words, for each of such n, the inverse of ψ_n is given by equation (8).

$$\psi_n^{-1}(l) = \begin{cases} (1, l) & \text{if } 1 \leq l \leq n \\ (c(l)+1, r(l)) & \text{if } n < l \leq n^2 \text{ and } l = c(l)n + \\ & r(l), \text{ with } 0 \neq r(l) < n \\ (c(l), n) & \text{if } n < l \leq n^2 \text{ and } l = c(l)n, \end{cases} \tag{8}$$

Table 3: Vector form of the matrix (9)

Value	0	0	5	0	1	0	4	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	8	0
Index	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

Useful to us is to consider the sets $\mathbb{N}_n := \{0, 1, \cdots, n-1\}$, for any $n \in \mathbb{N}^*$, instead of $\mathbb{N}'_n$. In this case for a given $n \in \mathbb{N}^*$ and employing the bijections (7), up to the identification of $\mathbb{N}'_n$ with $\mathbb{N}_n$, we can list the elements of the set $\mathbb{N}_{n^2}$ in form of a $n \times n$-square as follows:

$$\begin{array}{c} 0 \\ 1 \\ 2 \\ \vdots \\ n-2 \\ n-1 \end{array} \begin{pmatrix} 0 & 1 & \cdots & n-1 \\ n & n+1 & \cdots & 2n-1 \\ 2n & 2n+1 & \cdots & 3n-1 \\ \vdots & \vdots & \vdots & \vdots \\ (n-2)n & (n-2)n+1 & \cdots & (n-1)n-1 \\ (n-1)n & (n-1)n+1 & \cdots & n^2-1 \end{pmatrix}$$

The elements of $\mathbb{N}_{n^2}$ are referred to as *locations and/or indices.*

The previous matrix format, will be our starting point in sorting out square matrices. Thus, for square matrices of order n, an non-zero (i, j)-entry, where i stands for the row position and j the column one, and with $0 \leq i, j \leq n-1$, is stored in position (*location* or *index*) $i \times n + j$. We will use the notation: $index_{A_{ij}} := i \times n + j$, for a non-zero entry A_{ij} of a given n-square matrix A.

Example 2.1. For example, if $n = 5, 6$, then the general list of indices are of the form:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 & 9 \\ 10 & 11 & 12 & 13 & 14 \\ 15 & 16 & 17 & 18 & 19 \\ 20 & 21 & 22 & 23 & 24 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 10 & 11 \\ 12 & 13 & 14 & 15 & 16 & 17 \\ 18 & 19 & 20 & 21 & 22 & 23 \\ 24 & 25 & 26 & 27 & 28 & 29 \\ 30 & 31 & 32 & 33 & 34 & 35 \end{pmatrix}$$

and the following matrix:

$$A = \begin{pmatrix} 0 & 0 & 5 & 0 & 1 \\ 0 & 4 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 & 0 \\ 0 & 0 & 0 & 8 & 0 \end{pmatrix} \tag{9}$$

is stored as shown in Table 2:

Table 2: Sparse Matrix Format of the matrix (9)

Values	5	1	4	1	9	8
Cumulative index	2	4	6	10	18	23

Thus the index of the value 5 is the positive integer 2, and that of 4 is 6 and so on.

In this way, one can also write the whole matrix as a vector, for instance, the matrix of equation (9), has the vector form as indicated in Table 3:

The above process can be also reversed, that is, one can recover the row and column positions of a given entry from its index. Thus, the row position for any entry located in index $k \in \{0, 1, \cdots, n^2-1\}$, is calculated as the integer part (or *floor* function) of the ratio between the $index_{A_{ij}}$ and n the size of the given matrix, as follows:

$$row_{A_{ij}} = \left\lfloor \frac{index_{A_{ij}}}{n} \right\rfloor = floor(\frac{index_{A_{ij}}}{n}), \tag{10}$$

where the *floor function* takes as an input a real number x and gives as output the greatest integer less than or equal to x. For example $floor(3,4) = 3$ and $floor(3) = 3$.

The column index for any entry of location k is calculated as the remainder of the division between the $index_{A_{ij}}$ and n, also known as the *modulo* function:

$$col_{A_{ij}} = index_{A_{ij}} - \left\lfloor \frac{index_{A_{ij}}}{n} \right\rfloor = index_{A_{ij}} \text{ modulo } n. \tag{11}$$

From the implementation perspective, the advantage of this structure is that it uses only one index for each non-zero entry. On the other hand there are two operations executed in order to find the row and column indexes.

The total number of words needed for this format to be stored in the memory can be calculated using the formula (12) below, while the total number of words to store SMF on the disk is a little bit different and it is calculated using the subsequent formula (13):

$$DIM_{SMFmem} = 2 \times n^2 \times density(A) + n \tag{12}$$

$$DIM_{SMFdisk} = 2 \times n^2 \times density(A) + 1 \tag{13}$$

The division between the memory requirements of SMF and the classical format is:

$$r_{SMF} = 2 \times density(A) + \frac{1}{n^2} \tag{14}$$

The upper limit for the density of a matrix for which this format is still memory efficient is $density(A) = 0,5$.

2.2 Arithmetic operations with sparse matrices in SMF format

The sum of two square matrices in their SMF format is not difficult and will not be discuss from the mathematical point of view, see Section 3 below, however, for the implementation of these two arithmetic operations.

The transpose of a given matrix in it SMF format is formulated as follows. Assume we are working with square matrices of order n. Chose a position $k = i \times n + j \in \{0, \cdots, n^2 - 1\}$, where a certain matrix has a non-zero entry $A_{ij} \neq 0$, we know that this is the $(j,i)^{\text{th}}$-entry of the transpose matrix, that is, the entry in location $\overline{k} = j \times n + i$. This in fact define a bijective map $inv : \mathbb{N}_{n^2} \to \mathbb{N}_{n^2}$, $k \mapsto \overline{k}$, called the *inverse map*. For example, if $n = 5$ then the table of inverses is given in Table 4.

Table 4: The inverse map for $n = 5$

k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
$\overline{k}$	0	5	10	15	20	1	6	11	16	21	2	7	12	17	22	3	8	13	18	23	4	9	14	19	24

For a large positive integer n, the algorithm in computing the outputs of the inverse map is based on the following observations: For a position $k \in \{0, \cdots, n^2 - 1\}$, we have the following equation (15).

$$\overline{k} = \begin{cases} l & \text{if } k = l \times n, \text{ for some } l \in \mathbb{N}_n; \\ l' \times n + l & \text{if } l \times n < k \leq (l+1) \times n + 1, \text{ and } k = \\ & l \times n + l', \text{ for some } 1 \leq l' \leq n - 1 \end{cases} \tag{15}$$

The first case says that k belongs to block number l and it is occupying position 0 in this block. The second case says that k belongs to block number l and it is located in position $l' \geq 1$.

This suggests that first we should provide n lists (or blocks) each of which have n elements and mark the position of k inside one of these blocks. In this way, the inverse of k belong to the block enumerated by the position of k (inside its block) and located inside this block by the position enumerated by the block of k. In other words, in order to compute the inverse of a given location we only need to interchange the pair $(position, block)$ by $(block, position)$. For instance, if $n = 5$, the the blocks partitions are displayed as in Table 5.

Table 5: The inverse map by using blocks and positions: $n = 5$.

	0					1					2					3					4				
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
k	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
$\overline{k}$	**0**	**5**	**10**	**15**	**20**	**1**	**6**	**11**	**16**	**21**	**2**	**7**	**12**	**17**	**22**	**3**	**8**	**13**	**18**	**23**	**4**	**9**	**14**	**19**	**24**

Specifically, if we look at location $k = 8$, then it belongs to the second block and it is located in position enumerated by 3 in this block. Thus, its inverse $\overline{8}$ is the location which belongs to the third block and occupy the second position in this block so that $\overline{8} = 16$.

Table 8: The SMF format of A and B

indices	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	0	0	5	0	1	0	4	0	0	0	1	0	0	0	0	0	0	0	9	0	0	0	0	8	0
B	0	1	0	2	1	0	0	0	2	0	0	6	0	0	0	0	0	5	1	0	0	7	0	0	0

The algorithm of computing the transpose of given square matrix in its SMF format, is then deduced by restricting the permutation of Table 4, to the set of its indices. For example, given the square matrix of order 5 in (16), then, Table 6 shows how to apply this algorithm to B, leading to the SFM format of the transpose of B as shown in Table 7.

$$B = \begin{pmatrix} 0 & 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 5 & 1 & 0 \\ 0 & 7 & 0 & 0 & 0 \end{pmatrix} \tag{16}$$

Table 6: The transpose algorithm applied to the matrix (16)

Values	1	2	1	2	6	5	1	7
index	1	3	4	8	11	17	18	21
inverse index	5	15	20	16	7	13	18	2
transpose value	1	2	1	2	6	5	1	7

Table 7: The Sparse Matrix Format of the transpose matrix of the matrix (16)

value	1	2	1	2	6	5	1	7
index	5	15	20	16	7	13	18	2

Remark 2.2. The terminology *inverse map* or *inverse index*, has in fact an algebraic interpretation, which is revealed by the groupoid pair (see [6, Definition 1.4, Example 1.11] for a precise definition of this mathematical object) structure given on the set $\mathbb{N}_{n^2}$ by identifying it with the Cartesian product $\mathbb{N}_n \times \mathbb{N}_n$ via the bijection (7). More precisely, the set $\mathbb{N}_n \times \mathbb{N}_n$ can be given a structure of groupoid (of pair) over $\mathbb{N}_n$. Thus $\mathbb{N}_n \times \mathbb{N}_n$ is the set of arrows of this groupoid and $\mathbb{N}_n$ is its set of objects. Concretely, any pair (i, j) can be considered as an arrow with source $s(i, j) = i$ and target $t(i, j) = j$, and for any pair of pairs $(i, j), (i', j') \in \mathbb{N}_n \times \mathbb{N}_n$ with $t(i, j) = s(i', j')$ we define the partial multiplication (opposite to the composition) $(i, j) \star (i', j') := (i, j')$, whenever $j = i'$.

The identity arrow of a given object $i \in \mathbb{N}_n$, is the pair (i, i). Lastly the inverse arrow of a given arrow (i, j) is the pair (j, i).

Next we discuss the multiplication algorithm. In order to illustrate the steps of this algorithm, let us start, for instance, by the matrix multiplication AB, where A and B are the matrices of equations (9) and (16), respectively.

The SMF format of both matrices is given in Table 8. So we can put the values of B in horizontal form and those of A in vertical one, and the usual outcome local multiplications operation are summarised in Table 9.

As the extended algorithm is very large when put in the table form of 9, it can be reduced to Table 10 and the content being unchanged. Furthermore, it is possible to write only the lines that have either the entry of matrix A or B non-zero like in Table 11 and still organizing it in blocks.

And in a short of *compact format*, we can write the main information as in Table 11.

This compact format works as follows. In Table 11, the entries of the column labelled by **index**, **A**, **B** and **block**, are clear. Now, the first row in the mid rectangle, corresponds to the multiplication of the value of location 1 (i.e., the entry 0 of index 1 in the matrix A) by all the values (of the entries of B) allocated in block number 1 [‡]; while the second row is the resulting pairwise multiplication of the value 5 (of the matrix A) at location 2 by all the values (of matrix B) in locations situated in block number 2, and so on. In this way the first rectangle has to contains four row which corresponds exactly to the number of locations with non zero entry shared by A and B in

[‡] That is, we execute the pairwise operations: $0 \times 0 = 0$ and $0 \times 2 = 0$ and the resulting values are allocated in the first and the third positions of the first row, respectively. The rest of this row-entries are automatically fill out by zero values.

Table 9: Matrix multiplication in vector format

Index		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	**block**
	Value	0	1	0	2	1	0	0	0	2	0	0	6	0	0	0	0	0	5	1	0	0	7	0	0	0	
0	0	0	0	0	0	0																					0
1	0						0	0	0	0	0																
2	5											0	30	0	0	0											
3	0																0	0	0	0	0						
4	1																					0	7	0	0	0	
5	0	0	0	0	0	0																					1
6	4						0	0	0	8	0																
7	0											0	0	0	0	0											
8	0																0	0	0	0	0						
9	0																					0	0	0	0	0	
10	1	0	1	0	2	1																					2
11	0						0	0	0	0	0																
12	0											0	0	0	0	0											
13	0																0	0	0	0	0						
14	0																					0	0	0	0	0	
15	0	0	0	0	0	0																					3
16	0						0	0	0	0	0																
17	0											0	0	0	0	0											
18	9																0	0	45	9	0						
19	0																					0	0	0	0	0	
20	0	0	0	0	0	0																					4
21	0						0	0	0	0	0																
22	0											0	0	0	0	0											
23	8																0	0	40	8	0						
24	0																					0	0	0	0	0	

Table 10: Matrix multiplication in vector format: reduced

Index	**A**						**B**	**block**
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	
2	5	0	30	0	0	0	0	
3	0	0	0	0	0	0	2	
4	1	0	7	0	0	0	1	
		0	**37**	**0**	**0**	**0**		
5	0	0	0	0	0	0	0	1
6	4	0	0	0	8	0	0	
7	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	2	
9	0	0	0	0	0	0	0	
		0	**0**	**0**	**8**	**0**		
10	1	0	1	0	2	1	0	2
11	0	0	0	0	0	0	6	
12	0	0	0	0	0	0	0	
13	0	0	0	0	0	0	0	
14	0	0	0	0	0	0	0	
		0	**1**	**0**	**2**	**1**		
15	0	0	0	0	0	0	0	3
16	0	0	0	0	0	0	0	
17	0	0	0	0	0	0	5	
18	9	0	0	45	9	0	1	
19	0	0	0	0	0	0	0	
		0	**0**	**45**	**9**	**0**		
20	0	0	0	0	0	0	0	4
21	0	0	0	0	0	0	7	
22	0	0	0	0	0	0	0	
23	8	0	0	40	8	0	0	
24	0	0	0	0	0	0	0	
		0	**0**	**40**	**8**	**0**		

Table 11: Compact matrix multiplication AB in vector format

Index	**A**						**B**	**block**
1	0	0	0	0	0	0	1	0
2	5	0	30	0	0	0	0	
3	0	0	0	0	0	0	2	
4	1	0	7	0	0	0	1	
		$^{0}\mathbf{0}$	$^{1}\mathbf{37}$	$^{2}\mathbf{0}$	$^{3}\mathbf{0}$	$^{4}\mathbf{0}$		
6	4	0	0	0	8	0	0	1
8	0	0	0	0	0	0	2	
		$^{5}\mathbf{0}$	$^{6}\mathbf{0}$	$^{7}\mathbf{0}$	$^{8}\mathbf{8}$	$^{9}\mathbf{0}$		
10	1	0	1	0	2	1	0	2
11	0	0	0	0	0	0	6	
		$^{10}\mathbf{0}$	$^{11}\mathbf{1}$	$^{12}\mathbf{0}$	$^{13}\mathbf{2}$	$^{14}\mathbf{1}$		
17	0	0	0	0	0	0	5	3
18	9	0	0	45	9	0	1	
		$^{15}\mathbf{0}$	$^{16}\mathbf{0}$	$^{17}\mathbf{45}$	$^{18}\mathbf{9}$	$^{19}\mathbf{0}$		
21	0	0	0	0	0	0	7	4
23	8	0	0	40	8	0	0	
		$^{20}\mathbf{0}$	$^{21}\mathbf{0}$	$^{22}\mathbf{40}$	$^{23}\mathbf{8}$	$^{24}\mathbf{0}$		

block 0. The last row, in this first mid rectangle, is the result of the pairwise summations of the values of each column, and this will be, of course, the first row in the matrix AB. The second mid rectangle leads to the second row of A and so on.

For more illustrative examples, the compact matrix multiplication algorithm has been applied to the following two matrices:

$$\begin{pmatrix} 0 & 0 & 1 & 3 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 2 & 4 \\ 0 & 3 & 0 & 1 \\ 0 & 0 & 4 & 0 \\ 0 & 5 & 2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 15 & 10 & 0 \\ 0 & 11 & 2 & 2 \\ 0 & 15 & 6 & 0 \\ 0 & 0 & 4 & 0 \end{pmatrix}.$$

Then the outcome is given in Table 12.

Table 12: Compact matrix multiplication in vector format: $n = 4$

Index	**A**					**B**	**block**
2	1	0	0	4	0	2	0
3	3	0	15	6	0	4	
		$^{0}\mathbf{0}$	$^{1}\mathbf{15}$	$^{2}\mathbf{10}$	$^{3}\mathbf{0}$		
5	2	0	6	0	2	3	1
7	1	0	5	2	0	1	
		$^{4}\mathbf{0}$	$^{5}\mathbf{11}$	$^{6}\mathbf{2}$	$^{7}\mathbf{2}$		
10	0	0	0	0	0	4	2
11	3	0	15	6	0	0	
		$^{8}\mathbf{0}$	$^{9}\mathbf{15}$	$^{10}\mathbf{6}$	$^{11}\mathbf{0}$		
13	0	0	0	0	0	5	3
14	1	0	0	4	0	2	
		$^{12}\mathbf{0}$	$^{13}\mathbf{0}$	$^{14}\mathbf{4}$	$^{15}\mathbf{0}$		

Remark 2.3. As one can realizes the compact format reduces hugely the number of pairwise multiplications in matrix multiplication operation.

On the other hand, in relation with Remark 2.2, if we denote by $\mathcal{N}$ the groupoid of pair over $\mathbb{N}_n$, then the path rig[§]

$$\mathbb{N}\mathcal{N} = \oplus_{0\leq i,j\leq n-1}\mathbb{N}\cdot(i,j),$$

where $\mathbb{N}\cdot(i,j)$ stands for the free commutative monoid generated by the element (i,j) (i.e., the arrow from i to j), coincides with the rig of all square matrices $M_{n\times n}(\mathbb{N})$ with positives integers as entries. Of course, the same arguments holds true for matrices with entries in integer numbers or even in the field of fractions.

[§] In general the notion of *rig* stands for a set with two compatible internal commutative binary operations (summation and multiplication with distributive law), each of which leads to a monoid structure. A prototype example of this algebraic structure is the set of positive integer $\mathbb{N}$ with the usual sum and multiplication.

Lastly, it is noteworthy to mention that the above multiplication algorithm and the compact format, can be applied to the multiplication operations of non square matrices. This is done by enlarging the size of the column and/or the row numbers (by putting the value zero in the new entries), in order to inject the input matrices into two big square matrices, and then apply the block matrices multiplication. This approach will not be contemplated here.

3 Implementation of operations with Sparse Matrix Format

In this Section, we give in terms of C++ codes the computer implementations of the arithmetic operations on square matrices in their SMF format. Most of the illustrative example are matrices with entries in the rig of positive integer, however, the implementation works as well for matrices with entries in the ring of integers numbers.

Let us start by the following simplest situation. Consider the matrix M as in (17) below, the SMF implementation looks like the representation (18). Therein, the notation $[p]$ is for pointer to another memory zone and, for instance, the pair $(5,2)$ indicates the value 5 of the entry with index 2 (recall that we have $n = 4$ and the list of indices is $\{0,1,2,\cdots,15\}$). Each row is a dynamically adjustable array of tuples:

$$M = \begin{pmatrix} 1 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 8 & 0 & 0 & 4 \end{pmatrix} \tag{17}$$

$$\begin{array}{l} (2,[p]) \rightarrow [(1,0),(5,2)] \\ (0,[p]) \rightarrow NULL \\ (1,[p]) \rightarrow [(2,2)] \\ (2,[p]) \rightarrow [(8,0),(4,3)] \end{array} \tag{18}$$

For a more modular implementation there were defined classes used for storing data and operations on data:

Listing 1: Class definition of SMF in C++ code

```
class entry {
  public: value_type value; size_type column;
  //...
};
class SMF {
  public: index_type matrix_size;
    entry** rows; size_type* row_len;
    index_type nnz = 0;
  //...
};
```

3.1 From Standard to Sparse Matrix Format

The function *StandardtoSMF* is used to transform the Standard format to SMF. It is a member function of the SMF class and it takes as an argument a matrix as a pointer to pointer to ***value_type***. In order to find the non-zero entries of the standard matrix it is necessary to iterate through all the entries. Once a non-zero entry is found it is inserted in the SMF format along with the aggregated index. Here is the implementation of such function:

Listing 2: Creation of a SMF from a standard square matrix in C++ code

```
void StandardtoSMF(value_type** Standard){
  //iterate through rows
  for (value_type i = 0; i < matrix_size; ++i)
    //iterate through columns
    for (value_type j = 0; j < matrix_size; ++j)
      //insert to SMF value and
      //aggregated index if entry is non-zero
      if (Standard[i][j] != 0)
        this->insert(Standard[i][j],
             i*matrix_size + j);
}
```

3.2 From Sparse Matrix Format to Standard

The function *SMFtoStandard* is used to make the translation between the two formats. As SMF is implemented to only store each non-zero entry of the matrix and the composed index, for each SMF entry the column and row have to be determined. Once this operations are done the entry is ready to take its position in the standard matrix. The C++ implementation is as follows:

Listing 3: Creation of a standard square matrix from SMF in C++ code

```
value_type** SMFtoStandard() {
  value_type **Standard, val;
  index_type col;
  //allocating dynamic array of size=matrix_size
  //of pointers to element type (value_type)
  //initializing all to 0
  Standard = new value_type*[matrix_size]();
  //allocate each row
  for(index_type i = 0; i < matrix_size; ++i)
    Standard[i] = new value_type[matrix_size];
  //each i-th pointer is now pointing
  //to dynamic array (size matrix_size)
  //of actual value_type values

  //iterating through rows
  for(index_type i = 0; i < matrix_size; ++i)
    //iterating through elements
    for(index_type j = 0; j < row_len[i]; ++j) {
      //get col index and value of entry
      col = rows[i][j].getC();
      val = rows[i][j].getV();
      //store the value at the precise indexes
      Standard[i][col] = val;
    }

  return Standard;
}
```

3.3 Operation with matrices in SMF format

In order to make the SMF a viable replacement for the standard matrix format, operations with it have to be defined. As stated before, the second objective here is to implement a solution for the associated operations of SMF. Operations considered to be implemented are: multiplication, transposing, summation and subtraction.

3.3.1 Sum of two SMF matrices

The summation algorithm described below is implemented as a member function in the SMF class. It uses two position pointers to iterate through A's row respectively B's row called *apos* and *bpos*: (1) For each row of both matrices (first *while* loop); (2) Reinitialize *apos* and *bpos* to 0 and get row length in *len_rowA* and *len_rowB*; (3) While the pointers did not reach the end of the row (second *while* loop); (4) Get the 2nd element of tuple entry, the column index in *col_A* and *col_B*; (5) By comparing the column index it is decided if entries must be inserted individually in the result or added and then inserted in the result; (6) The remaining elements from A's or B's row are inserted (3rd and 4th *while* loops).

Listing 4: Sum of two matrices in C++ code

```
SMF add(SMF B) {
  SMF rez(matrix_size);
  index_type apos, bpos; index_type i = 0;
  //same row for both matrices
  while(i < matrix_size) {
    apos=0; bpos=0;
    //get row A & B len
    size_type len_rowA = row_len[i];
    size_type len_rowB = B.row_len[i];
    while (apos < len_rowA && bpos < len_rowB) {
      //get A's col
          size_type A_col = rows[i][apos].getC();
      //get B's col
      size_type B_col = B.rows[i][bpos].getC();
      if(B_col < A_col){
        //insert B's val & calc index
        rez.insert(B.rows[i][bpos].getV(),
                       i*matrix_size + B_col);
        bpos++;
      } else if(B_col > A_col){
         //insert A's val & calc index
         rez.insert(rows[i][apos].getV(),
                       i*matrix_size + A_col);
         apos++;
      } else { //else: same col -> add them
        rez.insert(rows[i][apos].getV()+
        B.rows[i][bpos].getV(),
                       i * matrix_size + A_col);
        apos++; bpos++;
      }
    }
    //insert ramaining el from A
    while(apos < len_rowA){
      rez.insert(rows[i][apos].getV(),
        i * matrix_size + rows[i][apos].getC());
      apos++;
    }
    //insert ramaining el from B
    while(bpos < len_rowB) {
      rez.insert(B.rows[i][bpos].getV(),
        i * matrix_size + B.rows[i][bpos].getC());
      bpos++;
    }
    i++;
  }
  return rez;
}
```

3.3.2 Subtraction of two SMF matrices

The subtraction algorithm described below is implemented as a member function in the SMF class. It is very similar with the summation algorithm with the difference that if the column indexes are different then the inverse number 0-*val* is inserted in the result and if there are equal subtraction of the two is done instead of summation. A description of the algorithm is as follows: (1) For each row of both matrices; (2) Reinitialize *apos* and *bpos* to 0 and get row length in *len_rowA* and *len_rowB*; (3) While the pointers did not reach the end of the row; (4) Get the 2$^{\text{nd}}$ element of tuple entry, the column index in *col_A* and *col_B*; (5) By comparing the column index it is decided if the inverse number (0-*val*) must be inserted individually in the result or subtracted and then inserted in the result; (6) The remaining elements from *A*'s or *B*'s row are inverted and inserted.

3.3.3 Transposing the SMF

The transposing algorithm is simply iterating through all rows, then through entries and inserting to the result matrix the calculated aggregated index of the transpose: (1) For each row (the first *for* loop); (2) For each entry in the row (the 2$^{\text{nd}}$ *for* loop); (3) Get the value and column index of the entry; (4) Insert to the result matrix the transposed aggregated index.

Listing 5: Matrix transposition C++ code

```
SMF transpose(){
  SMF rez(matrix_size);
  value_type val; index_type col;
  //iterating through rows
  for(index_type i = 0; i < matrix_size; ++i)
    //iterating through elements
    for(index_type j = 0; j < row_len[i]; ++j) {
      //get col index and value of entry
      col = rows[i][j].getC();
      val = rows[i][j].getV();
      //add to rez the new index for this value
      rez.insert(val, col * matrix_size + i);
    }
  return rez;
}
```

3.3.4 Multiplication of two SMF matrices

The multiplication algorithm is one key operation when dealing with matrices. Considering two matrices A and B that can be multiplied, a short description of the algorithm is as follows: (1) First transpose B for an easier iteration through columns, as SMF is row-major order, we need B in column-major order; (2) Iterate through rows of A (first *for* loop); (3) For each row in A, iterate through columns of B (second *for* loop); (4) Reinitialize the local pointer of A's row and B's column and the sum to 0; (5) While these pointers are within the range of A's row length respectively B's column length; (6) Compare A's column with B's row and skip elements in A's row respectively B's column until they are equal; (7) If A's column is equal to B's row do the multiplication and add to sum; (8) If sum is different than 0 add it to result matrix with the aggregated index.

Listing 6: Matrix multiplication in C++ code

```
SMF multiply(SMF B) {
  SMF rez(matrix_size);
  index_type col_A, row_B, apos, bpos;
  value_type sum;
  //iterating through rows of A
  for(index_type i = 0; i < matrix_size; ++i)
    //iterating through cols of B
    for(index_type j = 0; j < matrix_size; ++j) {
      //local pointers within A's row and B's col
      apos = 0; bpos = 0;
      //sum of multiplication
      sum = 0;
      //iterating through A rows & B cols.
      while(apos < row_len[i] &&
            bpos < B.col_len[j]) {
        //get col index of A's entry
        col_A = rows[i][apos].getC();
        //get row index of B's entry
        row_B = B.rows[bpos][j].getR();
        //if A's col is smaller than B's row
        //skip entry in A
        if(col_A < row_B){
          apos ++;
        } else if (col_A > row_B){
          //if B's row is smaller than A's col
          //skip entry in B
          bpos++;
        } else {
          //else both row and col are equal
          //multiply the entries and add to sum
          sum += rows[i][apos].getV() *
            B.rows[bpos][j].getV();
          apos++; bpos++;
        }
      }
      //if the sum is non-zero add to rezult
```

```
        if (sum != 0)
            rez.insert(sum, i * matrix_size + j);
    }
    return rez;
}
```

4 Efficiency of the Sparse Matrix Format algorithms

This section discuss the memory efficiency in two cases: when storing the matrix on disk and when manipulating the data in algorithms, as they are slightly different. In the second case a few more information are needed in order to easily access and manipulate the data. Besides the use of memory point of view, we also analysed the computational efficiency of the *Sparse Matrix Format*. Obviously, the validation of the Sparse Matrix Format is done by calculating the memory and computational efficiency with respect to the classical format and algorithms.

4.1 The working sets

Test data that will be analysed has two collection sources: *The University of Florida Sparse Matrix Set* taken from [13] and *Anonymous MRI Brain Scan Images Database (University of Granada)* taken from public health sources.

4.1.1 The University of Florida Sparse Matrix Collection

The SuitSparse Matrix Collection also known as *The University of Florida Sparse Matrix Collection* is a huge and continuously growing database of sparse matrices that are encountered in real applications. This collection is intensively used by the numerical linear algebra community for performance evaluation of sparse matrix algorithms. Following [13], the collection covers a large number of domains, divided in two classes, these are: Matrices resulting from problems with a 2D or 3D geometrical representation (e.g., computer graphics/vision, robotics/kinematics, model reduction, etc.) and matrices without geometrical source interpretation (e.g., optimization, mathematics and statistics, economic and financial modeling, etc.).

From the *SuitSparse Matrix Collection* (see [13]) were chosen in a random manner 97 matrices. Size of this working set varies between 5 to *5000* (see Figures 2 and 1). It is worth mentioning that in practice the size can reach even 200.000 in which case *SMF* has even higher efficiency.

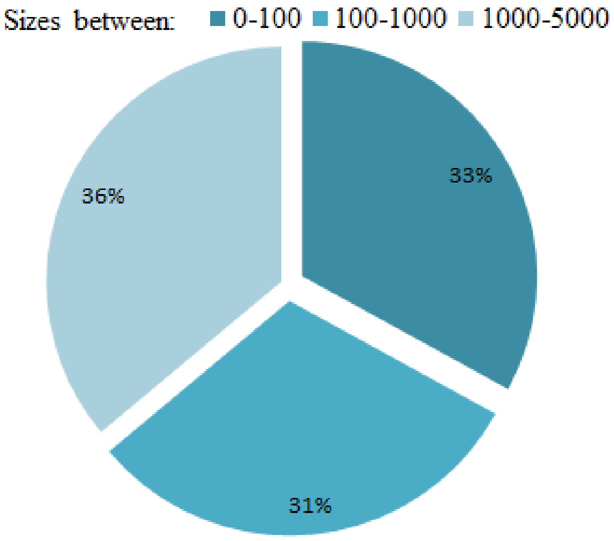


Figure 1: Intervals of sizes

The chart depicted in Figure 2 plots the matrix size against the sample number, an easy way to visualize the number of matrices in each interval of size ranges.

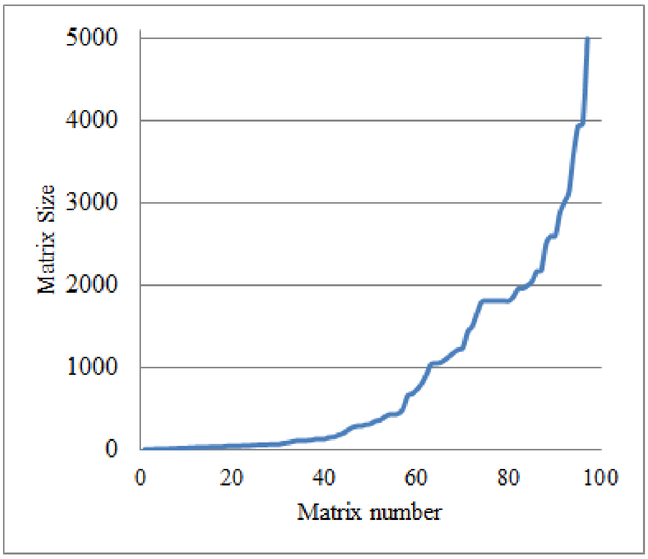


Figure 2: Matrix size range

From practical experiments the matrix density decreases as the matrix size increases. Figure 3 represents the size-density distribution. It can be easily notice how the left half of the chart with sizes smaller than 500 have a higher density in comparison with the right half in which, with a few exceptions, the density values are between 0 and 3%. Figure 3 also displays a trend line approximation of the data. It has been used a 3rd degree polynomial equation to approximate. For a chart with the matrices above the size of 500 with a linear trend see Figure 4.

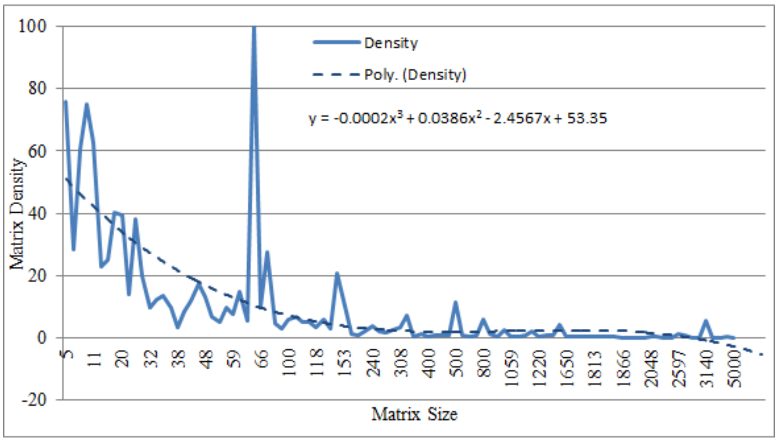


Figure 3: Matrix size-density distribution

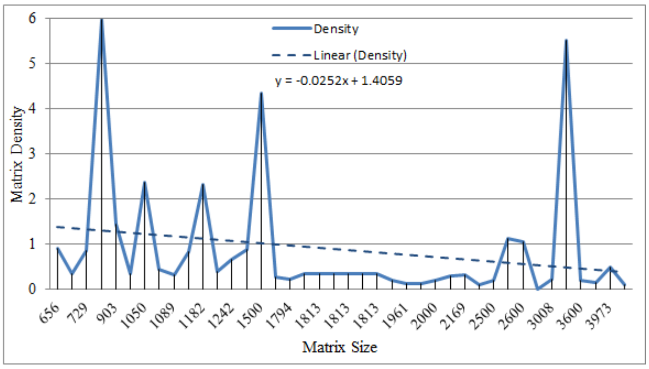
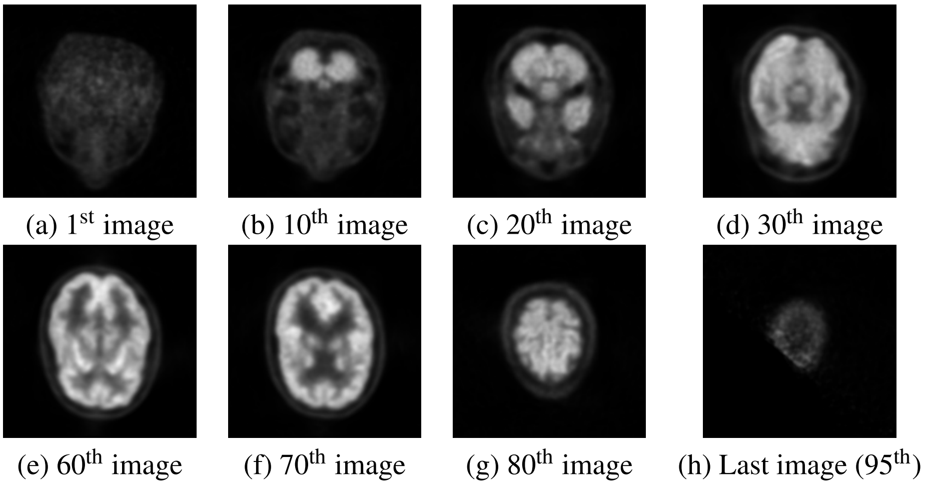


Figure 4: Zoomed matrix size-density distribution



(a) 1st image (b) 10th image (c) 20th image (d) 30th image

(e) 60th image (f) 70th image (g) 80th image (h) Last image (95th)

Figure 6: Sample images taken from a full MRI scan

4.1.2 Anonymous MRI Brain Scan Images Database

A second working set is *Anonymous MRI (Magnetic Resonance Imaging) Brain Scan Images Database*. But first it is needed to ensure that the representation of these images is a sparse matrix. For instance the image in Figure 5 can be loaded as a pixel-value 2D array. A gray-scale image has the pixel value represented on 8 bits: (1) for white the pixel has a value of 255 (all bits are 1); (2) for shades of gray the pixel has a value within the interval $[254, 1]$; (3) for black the value is 0.

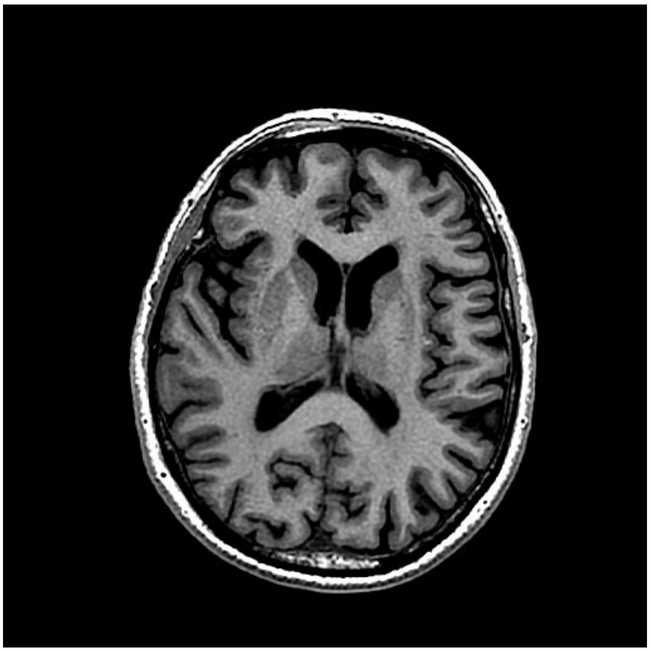


Figure 5: Brain MRI Example

As it can easily be noticed how the image below has most of the pixels black, so it can be considered a sparse matrix of pixels. The density of this image is among the higher that it can be in a MRI scan and it has a value of 25%. Therefore the compressed *SMF* brings a memory improvement.

A full MRI brain scan contains between 90 and 170 images. The scan is done as sections from one extremity of the head to the other, density of these images varies between 0% and 25% as it is illustrated in Figure 6. Below is showed every other 10th image of a MRI scan containing 95 images.

The best approximation of such evolution is a 2nd degree parabola as showed in Figure 7. The turning point of this equation is the image with the most information corresponding to the middle section of the head (Figure (6) between (e) and (f)).

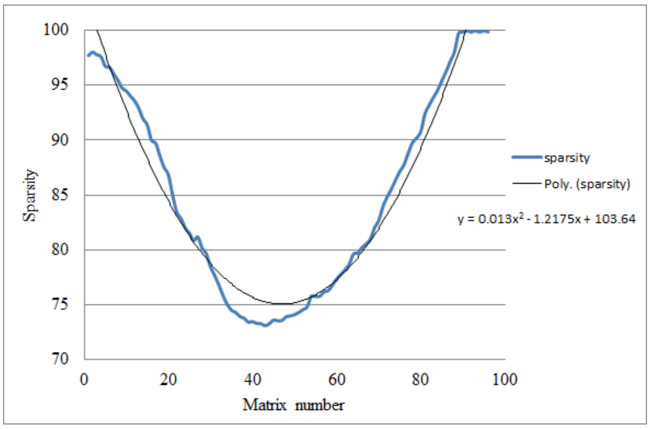


Figure 7: Sparsity of the MRI Brain Example

By plotting the sparsity-memory efficiency tuples it is observed in Figure 8 the linear growth of the memory efficiency proportional to the sparsity.

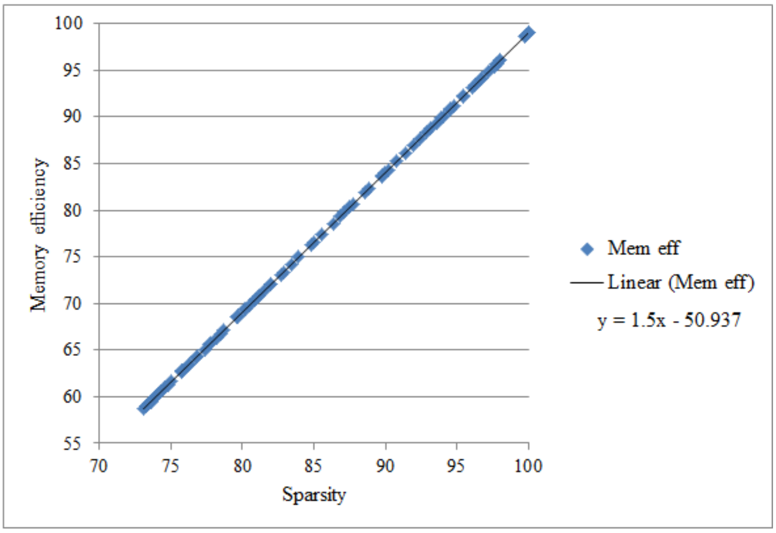


Figure 8: Memory efficiency of the MRI Eexample

4.2 Storing sparse matrices on the file system

As the sparse matrices can sometimes reach enormous sizes a method to store these matrices on the file system is needed. As described before, the SMF will be used, meaning that only the matrix size and the value-index tuples will be stored. The comparative way to store the SFM format in shown in Table 13.

4.2.1 Case study: large square matrix from the SuitSparse Matrix Collection

In order to study the efficiency of SMF when storing a big square sparse matrix on the disk, a matrix A with order $n = 20685$ was chosen from the *SuitSparse Matrix Collection*. It describes a *Structural Problem* system and it was published by Christian Damhaug (Oslo, Norway) in 2004 [13]. This matrix contains only binary values and it has a perfect *pattern and value symmetry* (see Figure 9). It has 2.454.957 non-zero values and a density of 0.5738% according to the following equation: $density(M) = \frac{N}{n^2} \times 100 = \frac{2.454.957}{20.685^2} \times 100 =$ 0.573763%.

Table 13: Matrix in the standard format (left) and the Sparse Matrix Format (right)

$n = 10$										$n = 10$											
1	0	0	1	0	0	1	0	0	1	1	0	1	3	1	6	1	9				
1	0	0	1	0	0	1	0	0	1	1	10	1	13	1	16	1	19				
1	1	1	1	0	0	1	0	0	1	1	20	1	21	1	22	1	23	1	26	1	29
0	1	1	1	0	0	1	0	0	1	1	31	1	32	1	33	1	36	1	39		
0	0	1	1	0	0	1	0	0	1	1	42	1	43	1	46	1	49				
0	0	0	1	1	1	1	0	0	1	1	53	1	54	1	55	1	56	1	59		
0	0	0	0	1	1	1	0	0	1	1	64	1	65	1	66	1	69				
0	0	0	0	0	1	1	1	1	1	1	75	1	76	1	77	1	78	1	79		
0	0	0	0	0	0	0	1	1	1	1	87	1	88	1	89						
0	0	0	0	0	0	0	0	1	1	1	98	1	99								

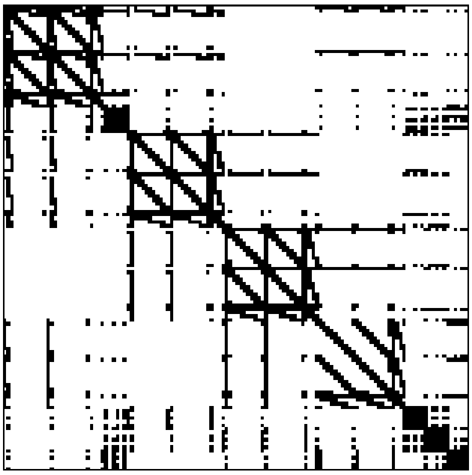


Figure 9: Pattern of the matrix A

The plain text file containing the above matrix on disk in the standard format (Table 13 - left) takes $816MB$ while the plain text file with the SMF style (Table 13 - right) of the same matrix occupies only $27,4MB$.

The second format represents an important improvement, taking 96.64% less space on the disk when compared to the standard storing format. The improvement has been determined according to the equation:

$$improvement = 100 - \frac{27.4 \times 100}{816} = 96.64\%. \tag{19}$$

4.3 SMF data structure implementation memory efficiency

For the memory efficiency, we will use the two sources of sparse matrices described in the previous subsections.

4.3.1 Working set 1: The University of Florida Collection

From the representation of the SMF memory efficiency against density (Figure 10) can easily be noticed that the smaller density has the higher memory efficiency.

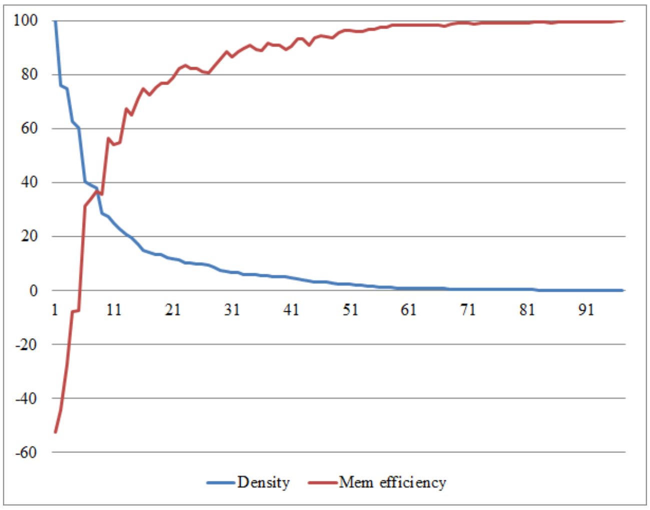


Figure 10: Memory efficiency against density

The total required memory for the new SMF data structure is calculated in several steps: (1) for each row a pointer to an array of entries and the length of the row are stored; (2) for each non-zero value it is used an entry structure (value and index), add to the total bytes number *sizeof(entry)* $\times N$.

For the SuitSparse working set the disk occupancy efficiency improvement in terms of percentages calculated using (19) plotted against the matrix size clearly shows, with a very few exceptions of small size matrices, that SMF is very efficient compared to the standard format (Figure 11). Both Sparse Matrix Format and standard are storing data on the disk in plain text.

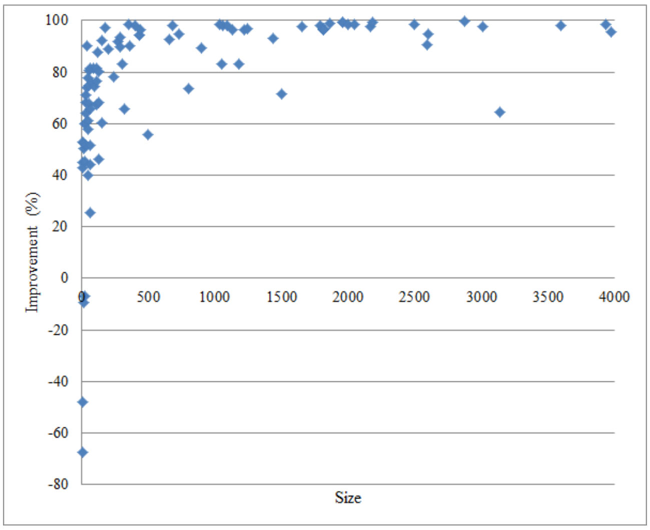


Figure 11: Memory improvement against matrix size

4.3.2 Working set 2: Anonymous MRI Brain Scan Images Database

In Figure (12) it is compared the disk space needed for the standard way of storing matrices in plain text with the new Sparse Matrix Format. Figure (13) represents the difference between standard and SMF. The difference is negative, meaning for a few matrices with a higher density the standard matrix arrangement is memory efficient. It can be easily observed how for the beginning and end image sections of the human head MRI scan the SMF is more efficient when compared with the middle section images where the information represents up to 25% of the image so less sparse.

For this full MRI scan example the total disk space taken by the 90 images in standard format is $8275KB$ compared with the SMF which takes $6596KB$, representing 79% of the original space on disk, so an improvement of 21% ($1679KB$) for each scan. Scaling this percentage up to the whole database the disk savings are considerable.

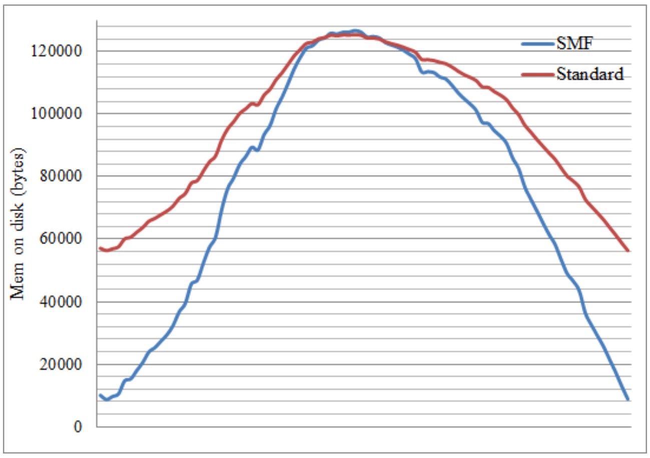


Figure 12: Memory on disk SMF and standard

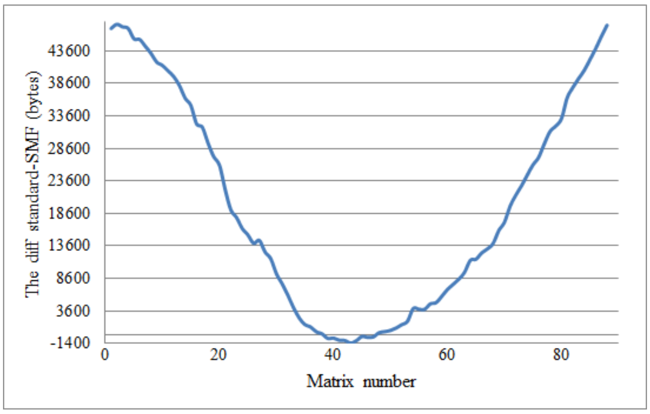


Figure 13: The difference between standard and SMF in Bytes

4.4 Computational efficiency

Computational efficiency is proven by comparing the classical format multiplication algorithm with the *SMF* multiplication algorithm detailed with examples in Subsection 2.2. A script written in Python was used to generate this simplified multiplication algorithm and compared with the classical matrix multiplication algorithm:

Listing 7: Python script generator

```
def new_mult_generator(n):
   program = '''C = [0] * (%d*%d)''' % (n, n)
   for i in range(0, n*n):
      block = int(i / n)
      program += '''
if A[%d] != 0: ''' % i
      for k in range(0, n):
         b_indx = int((i * n + k) % (n*n))
         program += '''
   if B[%d] != 0:
      C[%d] += A[%d] * B[%d]''' % (b_indx,
            ((b_indx % n) + (block * (n))), i, b_indx)
   return program
```

In Figure (14) it is presented the time evolution of the classical multiplication algorithm versus the new scripting method. The matrices used to make this comparison have a 20% density. With each matrix the operation was repeated 500 times in order to obtain a more precise average time.

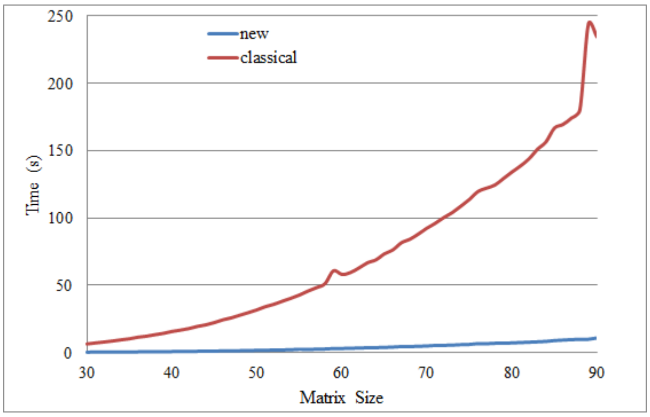


Figure 14: The time needed for 500 multiplications with the scripting vs classical method

The chart depicted in Figure (15) represents the time difference in seconds for the operation on matrices raging from 30 to 90 in size.

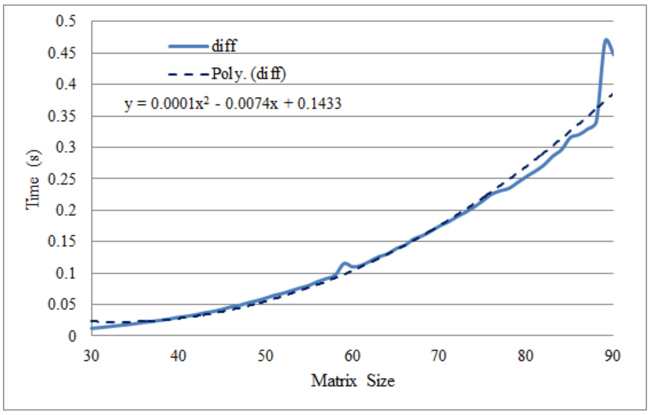


Figure 15: The time difference between average classical and scripting algorithm

The best approximation for such a trend is a 2^{nd} degree equation. For this benchmark set of sparse matrices (sizes 30 to 90, 20% density) the time difference of 500 multiplication of each matrix is 70 minutes. Therefore the script as it was generated in listing (7) is proven to be more efficient for matrices with the presented constraints.

References

[1] Beck, Amir, and Marc Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*. SIAM journal on imaging sciences 2.1 (2009): 183–202.

[2] Bioucas-Dias, José M. , and Mário AT Figueiredo, *Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing*, 2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing. IEEE, 2010.

[3] Candès, Emmanuel J. , and Michael B. Wakin, *An introduction to compressive sampling*. IEEE signal processing magazine 25. 2 (2008): 21–30.

[4] Elad, Michael, *Sparse and redundant representations: from theory to applications in signal and image processing*. Vol. **2**. No. 1. New York: springer, 2010.

[5] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices (Second Edition)*. Oxford Science Publications. Oxford University Press, 2017.

[6] A. Ibort and M. A. Rodríguez, *Introduction to Groups, Groupoids and their Representations*, Taylor & Francis Group, Boca Raton London New York (2020).

[7] I. Ivan, M. Popa, and P. Pocatilu, *Structuri de date: Tipologii de structuri de date*, 2008. [Online]. Available: http://www.ionivan.ro/2015-PUBLICATII/CARTI.htm

[8] Golub, Gene H. , and Charles F. Van Loan, *Matrix computations*. JHU press, 2013.

[9] Knopp, Tobias, and Alexander Weber, *Sparse reconstruction of the magnetic particle imaging system matrix*. IEEE transactions on medical imaging 32.8 (2013): 1473–1480. https://ieeexplore.ieee.org/document/6497631.

[10] Mohn, Fabian, et al, *System matrix based reconstruction for pulsed sequences in magnetic particle imaging*. IEEE transactions on medical imaging 41.7 (2022): 1862–1873. https://ieeexplore.ieee.org/document/9706173.

[11] Lustig, Michael, David Donoho, and John M. Pauly, *Sparse MRI: The application of compressed sensing for rapid MR imaging*. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 58.6 (2007): 1182–1195.

[12] Han, Shuguang, et al, *Application of Sparse Representation in Bioinformatics*. Frontiers in Genetics 12 (2021): 810875. https://www.frontiersin.org/articles/10.3389/fgene.2021.810875/full.

[13] *The Suite Sparse Matrix Collection*. [Online]. Available: https://sparse.tamu.edu/

[14] Woo, Jonghye, et al, *A sparse non-negative matrix factorization framework for identifying functional units of tongue behavior from MRI*. IEEE transactions on medical imaging 38.3 (2018): 730–740. https://arxiv.org/pdf/1804.05370.pdf

[15] Yang, Hujun, et al, *Deep learning in medical image super resolution: a review*. Applied Intelligence (2023): 1–26. https://link.springer.com/article/10.1007/s10489-023-04566-9

[16] Zeng, Gushan, et al. "A review on deep learning MRI reconstruction without fully sampled k-space." BMC Medical Imaging 21.1 (2021): 195. https://bmcmedimaging.biomedcentral.com/articles/10.1186/s12880-021-00727-9

Moroccan Journal of Algebra
and Geometry with Applications
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco

Volume 3, Issue 1 (2024), pp 219-231

ISSN: 2820-7114

Title :

Weakly uniformly graded-coherent rings

Author(s):

Abdelkbir Riffi

Moroccan Journal of Algebra and Geometry with Applications
Vol. 3(1) (2024), 219–231
Supported by Sidi Mohamed Ben Abdellah University, Fez, Morocco **ISSN: 2820-7114**

Weakly uniformly graded-coherent rings

Abdelkbir Riffi
Laboratory of Mathematics and applications (LMA), Department of Mathematics,
Faculty of Sciences, Ibn Zohr University, Agadir, Morocco.
e-mail: *riffiabdelkbir@gmail.com*

Communicated by Ünsal Tekir
(Received 20 October 2023, Revised 11 January 2024, Accepted 17 January 2024)

Abstract. Let $R = \bigoplus_{\alpha\in\Gamma} R_\alpha$ be a ring graded by an arbitrary grading abelian group Γ. We say that R is a weakly uniformly graded-coherent ring if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to R$ of degree 0, where $\lambda_1,\ldots,\lambda_n$ are degrees in Γ, $\ker f$ can be generated by $\phi(n)$ elements (not necessary homogeneous). In this paper, we provide the elementary properties of weakly uniformly graded-coherent rings.

Key Words: Weakly uniformly graded-coherent, uniformly graded-coherent, graded-coherent, uniformly coherent, coherent, graded modules and rings.
2010 MSC: 13A02, 13A15.

1 Introduction

We devote this section to some conventions and a review of some standard background material. All rings are commutative with unity, Γ will denote an abelian group written additively with an identity element denoted by 0 and all the graded rings and modules are graded by Γ.

Let R be a ring. An R-module M is called a *finitely presented R-module* if there is a *finite presentation of M*, that is, an exact sequence $F_1 \to F_0 \to M \to 0$ of R-modules such that both F_0 and F_1 are finitely generated free R-modules. Any finitely presented R-module is a finitely generated R-module; while the converse holds for Noetherian rings R, it is false in general. A finitely generated R-module M is said to be a *coherent R-module* if every finitely generated submodule of M is finitely presented; and a ring R is said to be a *coherent ring* if R is coherent as an R-module. A finitely generated R-module M is said to be a *uniformly coherent R-module* if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero R-module homomorphism $f : R^n \to M$, $\ker f$ can be generated by $\phi(n)$ elements; and a ring R is said to be a *uniformly coherent ring* if R is uniformly coherent as an R-module. An excellent summary of work done on coherence up to 1989 can be found in [15]. See for instance [19, 5, 8, 13, 11, 12, 14, 16, 17, 20, 24, 25].

The concept of coherence has many generalizations, see for instance [9, 10, 27, 18, 21]; among which we have the graded-coherence introduced by Cohen [9] for $\mathbb{Z}$-graded rings and modules, and then studied by Bakkari et al. [3] for Γ-graded ones. Let R be a graded ring. A finitely generated graded R-module M is said to be a *gr-coherent R-module* if every finitely generated homogeneous submodule of M is finitely presented; equivalently, if for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to M$, where $\lambda_1,\ldots,\lambda_n$ are degrees in Γ, $\ker f$ is finitely generated [3, Propodition 2.3]. The graded ring R is said to be a *gr-coherent ring* if it is gr-coherent as a graded R-module; equivalently, if $(I : a)$ is finitely generated, for every finitely generated homogeneous ideal I of R and for every homogeneous element $a \in R$; equivalently, if $(0 : a)$ is finitely generated, for every homogeneous element $a \in R$ and the intersection of two finitely generated homogeneous ideals of R

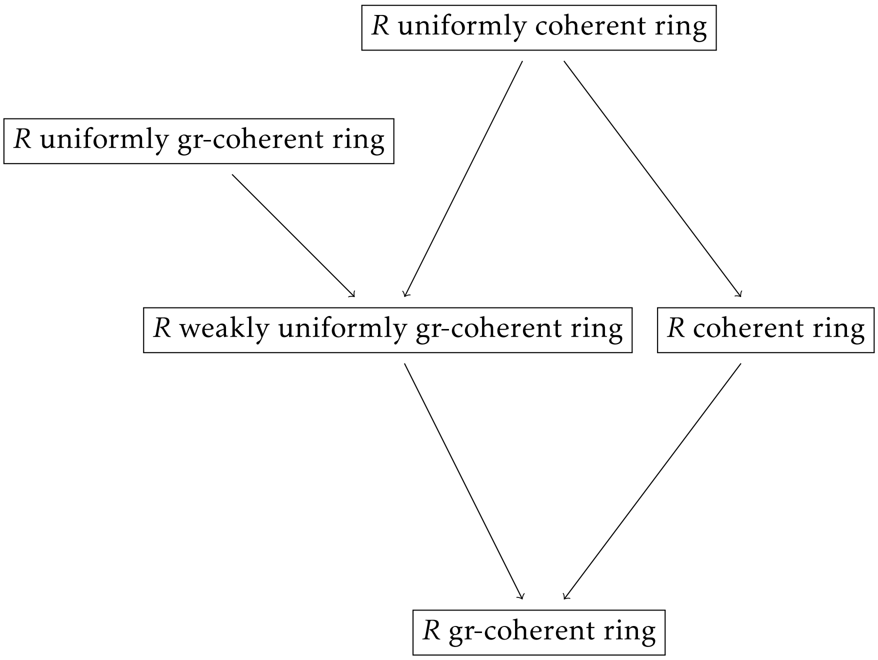

R uniformly coherent ring
R uniformly gr-coherent ring
R weakly uniformly gr-coherent ring
R coherent ring
R gr-coherent ring

Figure 1: The relations between the (graded-) coherent like notions for a graded ring R.

is finitely generated [3, Theorem 3.3]. Examples of gr-coherent rings, see [3, Examples 3.4], include graded-Noetherian rings[23], graded-valuation domains [1] and graded PrÃijfer domains [2].

In [4], Bakkari et al. generalize the concept of uniform coherence to the context of Γ-graded rings and modules, as follows. Let R be a graded ring. A finitely generated graded R-module M is said to be a *uniformly gr-coherent R-module* if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to M$ of degree 0, where $\lambda_1,...,\lambda_n$ are degrees in Γ, $\ker f$ can be generated by $\phi(n)$ homogeneous elements; the map ϕ is called a *uniformity map* of M. The graded ring R is said to be a *uniformly gr-coherent ring* if it is uniformly gr-coherent as a graded R-module.

In this paper, we investigate a particular class of uniformly gr-coherent rings (and modules) that we call weakly uniformly gr-coherent rings (and modules). Let R be a graded ring. A finitely generated graded R-module M is said to be a *weakly uniformly gr-coherent R-module* if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to M$ of degree 0, where $\lambda_1,...,\lambda_n$ are degrees in Γ, $\ker f$ can be generated by $\phi(n)$ elements (not necessary homogeneous); the map ϕ is called a *uniformity map* of M. The graded ring R is said to be a *weakly uniformly gr-coherent ring* if it is weakly uniformly gr-coherent as a graded R-module. The diagram (Figure 1), of a graded ring R, summarizes the relations between the (graded-) coherent like notions involved in this paper.

This article is organized as follows. In Section 2, we introduce the notion of weakly uniformly gr-coherent modules. Among other things, we show that for an exact sequence of graded R-modules $0 \to P \to N \to M \to 0$, the following hold: (1) If P is finitely generated and N is weakly uniformly gr-coherent, then M is weakly uniformly gr-coherent, (2) If M is uniformly gr-coherent and P is weakly uniformly gr-coherent, then N is weakly uniformly gr-coherent, and (3) If M is finitely presented and N is weakly uniformly gr-coherent, then P is weakly uniformly gr-coherent. We also show that, for a multiplicatively closed set S of homogeneous elements of R: if M is a weakly uniformly gr-coherent R-module, then $S^{-1}M$ is a weakly uniformly gr-coherent $S^{-1}R$-module. In Section 3, we study weakly uniformly gr-coherent rings. Among other things, several characterizations of weakly uniformly gr-coherent rings are given; for example, R is a weakly uniformly gr-coherent ring if and only if there is a map $\psi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any homogeneous ideal I of R generated by n homogeneous elements and any homogeneous element $a \in R$, $(I : a)$ can be generated by $\psi(n)$ elements. We also show that the group ring $A[X;\Gamma]$, over a ring A, is weakly uniformly gr-coherent if and only if it is uniformly gr-coherent if and only if A is uniformly coherent. We also show that, if R is a weakly uniformly gr-coherent ring, then so is R/I for any finitely generated homogeneous ideal I of R. We also show that, for a direct system of graded rings $(R_\lambda)_{\lambda\in S}$ such that $R := \varinjlim R_\lambda$ is a flat R_λ-module for any λ: if the R_λ's are weakly uniformly gr-coherent with the same uniformity map, then so is R.

We pause to review some definitions and preliminary results on graded modules and rings, see for instance

[7, II, §11, pp. 163–176]. Let Γ be a grading abelian group. By a *graded ring R*, we mean a ring graded by Γ, that is, a direct sum of subgroups R_α of R such that $R_\alpha R_\beta \subseteq R_{\alpha+\beta}$ for every $\alpha, \beta \in \Gamma$. An element $x \in R$ is called *homogeneous* if it belongs to one of the R_α, *homogeneous of degree α* if $x \in R_\alpha$. The element 0 is therefore homogeneous of all degrees; but if $x \neq 0$ is homogeneous, it belongs to only one of the R_α; the index α such that $x \in R_\alpha$ is then called the *degree* of x and is sometimes denoted by $deg(x)$. Every $y \in R$ may be written uniquely as a sum $\sum_\alpha y_\alpha$ of homogeneous elements with $y_\alpha \in R_\alpha$; y_α is called the *homogeneous component of degree α* of y. Clearly, R_0 is a subring of R (and in particular $1 \in R_0$).

By a *graded R-module $M = \bigoplus_{\alpha\in\Gamma} M_\alpha$*, we mean an R-module graded by Γ, that is, a direct sum of subgroups M_α of M such that $R_\alpha M_\beta \subseteq M_{\alpha+\beta}$ for every $\alpha, \beta \in \Gamma$. A graded R-module M is called a *graded-free R-module* if there exists a basis $(m_\iota)_{\iota\in I}$ of M consisting of homogeneous elements. The M_α are R_0-modules. Clearly, if R is a graded ring, R is a graded R-module. We can form a new graded R-module by *twisting* the grading on M as follows: if $\alpha_0 \in \Gamma$, define $M(\alpha_0)$ (read "M *twisted by* α_0"), to be equal to M as an R-module, but with its grading defined by $M(\alpha_0)_\alpha = M_{\alpha+\alpha_0}$.

Let R be a graded ring and $(M_\iota)_{\iota\in I}$ a family of graded R-modules, then $\bigoplus_{\iota\in I} M_\iota$ is a graded R-module, where $\left(\bigoplus_{\iota\in I} M_\iota\right)_\alpha = \bigoplus_{\iota\in I}(M_\iota)_\alpha$, for every $\alpha \in \Gamma$. Clearly, $\left(\bigoplus_{\iota\in I} M_\iota\right)(\alpha_0) = \bigoplus_{\iota\in I} M_\iota(\alpha_0)$.

Let R and R' be two graded rings, a ring homomorphism $h : R \to R'$ is called *graded* if $h(R_\alpha) \subseteq R'_\alpha$ for all $\alpha \in \Gamma$. A *graded ring isomorphism* is a bijective graded ring homomorphism. Let M and M' be two graded R-modules and let $u : M \to M'$ be an R-module homomorphism and $\delta \in \Gamma$; u is called *graded of degree δ* if $u(M_\alpha) \subseteq M'_{\alpha+\delta}$ for all $\alpha \in \Gamma$. An R-module homomorphism $u : M \to M'$ is called *graded* if there exists $\delta \in \Gamma$ such that u is graded of degree δ. Clearly, if $u : M \to M'$ is graded of degree δ, then $u : M(-\delta) \to M'$ and $u : M \to M'(\delta)$ are graded of degree 0. A *graded R-module isomorphism* is a bijective graded R-module homomorphism of degree 0. If $u \neq 0$, the degree of u is then determined uniquely. An *exact sequence of graded R-modules* is an exact sequence where the R-modules and the R-module homomorphisms in question are graded.

A submodule N of M is *homogeneous* if $N = \bigoplus_{\alpha\in\Gamma}(N \cap M_\alpha)$. It is well known that the following are equivalent for a submodule N of M: (1) N is homogeneous; (2) the homogeneous components of every element of N belong to N; (3) N is generated by homogeneous elements. A homogeneous submodule of R is called a *homogeneous ideal* of R. If N is a homogeneous submodule of a graded R-module M, then M/N is a graded R-module, where $(M/N)_\alpha := (M_\alpha + N)/N$. If I is a homogeneous ideal of a graded ring R, then R/I is a graded ring, where $(R/I)_\alpha := (R_\alpha + I)/I$.

Let R be a graded ring and M a graded R-module. If S is a multiplicatively closed set of homogeneous elements of R, then $S^{-1}R$ is a graded ring and $S^{-1}M$ is a graded $S^{-1}R$-module, where $(S^{-1}R)_i = \{\frac{r}{s} \mid r \in R_j, s \in R_k \textit{ and } j-k = i\}$ and $(S^{-1}M)_i = \{\frac{m}{s} \mid m \in M_j, s \in R_k \textit{ and } j-k = i\}$.

A *direct system $(R_\lambda, \phi_{\mu\lambda})$ of graded rings* is a direct system of rings such that each R_λ is graded and each $\phi_{\mu\lambda}$ is a homomorphism of graded rings. If $(R^\alpha_\lambda)_{\alpha\in\Gamma}$ is the graduation of R_λ and if we write $R = \varinjlim R_\lambda$, $R^\alpha = \varinjlim R^\alpha_\lambda$, then $(R^\alpha)_{\alpha\in\Gamma}$ is a graduation of R and R is a graded ring. If $\phi_\lambda : R_\lambda \to R$ is the canonical mapping, ϕ_λ is a homomorphism of graded rings.

2 Weakly uniformly gr-coherent modules

In this section, we define weakly uniformly gr-coherent modules.

Definition 2.1. Let R be a graded ring. A finitely generated graded R-module M is called a weakly uniformly gr-coherent R-module if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to M$ of degree 0, where $\lambda_1,\ldots,\lambda_n$ are degrees in Γ, $\ker f$ can be generated by $\phi(n)$ elements (not necessary homogeneous). The map ϕ is called a uniformity map of M.

Another way to look at this definition is the following. For a ring R and an R-module M, let $\mu_R(M)$ denote the minimal number of generators of M. A finitely generated graded R-module M is weakly uniformly gr-coherent $\Leftrightarrow$ for every $n \in \mathbb{N}$, $\sup_f \mu_R(\ker f) < \infty$, where f runs over the set of nonzero graded R-module homomorphisms $\bigoplus_{i=1}^{n} R(-\lambda_i) \to M$ of degree 0, with degrees $\lambda_1,\ldots,\lambda_n$ in Γ.

Remark 2.2. Every finitely generated homogeneous submodule of a weakly uniformly gr-coherent R-module is a weakly uniformly gr-coherent R-module.

Let R be a Γ-graded ring. By $sup(R) = \{\alpha \in \Gamma, R_\alpha \neq 0\}$ we denote the *support* of the graded ring R. In case $sup(R)$ is a finite set we will write $sup(R) < \infty$ and then R is said to be a Γ-graded ring *of finite support* [22]. We next collect some classes of weakly uniformly gr-coherent modules.

Proposition 2.3. *Let R be a graded ring. Then:*

1. *Every uniformly coherent graded R-module is weakly uniformly gr-coherent.*
2. (a) *Every uniformly gr-coherent R-module is weakly uniformly gr-coherent.*
   (b) *Assume that R is of finite support. Then any weakly uniformly gr-coherent R-module is uniformly gr-coherent.*
3. *Every weakly uniformly gr-coherent R-module is a gr-coherent R-module.*

Proof. (1) and (2)(a) These are straightforward.
(2)(b) and (3) Cf. (the proof of) [4, Propositions 2.3(2) and 2.4] respectively. □

We record the following Lemma.

Lemma 2.4 ([4, Lemma 2.6]). *Let R be a graded ring, N a graded R-module and $\lambda \in \Gamma$. Then, N is (weakly) uniformly gr-coherent if and only if $N(\lambda)$ is (weakly) uniformly gr-coherent.*

Theorem 2.5. Let R be a graded ring and let $0 \to P \xrightarrow{\alpha} N \xrightarrow{\beta} M \to 0$ be an exact sequence of graded R-modules. Then:

1. If P is finitely generated and N is weakly uniformly gr-coherent, then M is weakly uniformly gr-coherent.

2. If M is uniformly gr-coherent and P is weakly uniformly gr-coherent, then N is weakly uniformly gr-coherent.

3. If M is finitely presented and N is weakly uniformly gr-coherent, then P is weakly uniformly gr-coherent.

Proof. By Lemma 2.4, we may assume that $0 \to P \xrightarrow{\alpha} N \xrightarrow{\beta} M \to 0$ is an exact sequence of degree 0 graded R-module homomorphisms. Indeed, by twisting, we have an exact sequence of degree 0 graded R-module homomorphisms $0 \to P(-deg(\alpha)) \xrightarrow{\alpha} N \xrightarrow{\beta} M(deg(\beta)) \to 0$.

(1) Assume that N is weakly uniformly gr-coherent with uniformity map ϕ and that P is generated by s homogeneous elements. Then, by [7, II, p. 167, Remarque 3], we have a surjective degree 0 graded R-module homomorphism $h : \bigoplus_{i=1}^{s} R(-\lambda_i) \to P$ for some degrees $\lambda_1,\dots,\lambda_s$ in Γ. We show that M is weakly uniformly gr-coherent with uniformity map $\psi(n) := \phi(n+s)$. Since N is finitely generated and $\beta : N \to M$ is surjective, M is finitely generated. Let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n} R(-\lambda_{s+i}) \to M$ a graded R-module homomorphism of degree 0, where $\lambda_{s+1},\dots,\lambda_{s+n}$ are degrees in Γ. Consider the following commutative diagram with exact graded rows and columns:

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
0 \arrow[r] & \ker h \arrow[r, "u"] \arrow[d, hook] & \ker g \arrow[r, "v"] \arrow[d, hook] & \ker f \arrow[r] \arrow[d, hook] & 0 \\
0 \arrow[r] & \bigoplus_{i=1}^{s} R(-\lambda_i) \arrow[r, "u"] \arrow[d, "h"] & \bigoplus_{i=1}^{s+n} R(-\lambda_i) \arrow[r, "v"] \arrow[d, "g"] & \bigoplus_{i=1}^{n} R(-\lambda_{s+i}) \arrow[r] \arrow[d, "f"] & 0 \\
0 \arrow[r] & P \arrow[r, "\alpha"] \arrow[d] & \beta^{-1}(M_1) \arrow[r, "\beta"] \arrow[d] & M_1 := im(f) \arrow[r] \arrow[d] & 0 \\
 & 0 & 0 & 0 &
\end{tikzcd}$$

Since P and M_1 are generated respectively by s and n homogeneous elements, $\beta^{-1}(M_1)$ is generated by $s+n$ homogeneous elements, and so we have the surjective degree 0 graded R-module homomorphism $g : \bigoplus_{i=1}^{s+n} R(-\lambda_i) \to \beta^{-1}(M_1)$. Hence $\ker g$, and therefore $\ker f$, can be generated by $\phi(n+s)$ elements, as desired.

(2) Assume that M is uniformly gr-coherent and P is weakly uniformly gr-coherent with uniformity maps ϕ and ψ respectively. We show that N is weakly uniformly gr-coherent with uniformity map $\chi(n) = \psi(\phi(n))$. Since P and M are finitely generated, so is N. Let $n \in \mathbb{N}$ and $g : \bigoplus_{i=1}^{n} R(-\lambda_i) \to N$ a graded R-module homomorphism of degree 0, where $\lambda_1,\dots,\lambda_n$ are degrees in Γ. Consider the following commutative diagram with exact graded rows and columns:

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & 0 \arrow[d] & \\
 & \ker h \arrow[r, "u"] \arrow[d, hook] & \ker g \arrow[r] \arrow[d, hook] & 0 \arrow[d, hook] & \\
 & \bigoplus_{i=1}^{\phi(n)} R(-\mu_i) \arrow[r, "u"] \arrow[d, "h"] & \bigoplus_{i=1}^{n} R(-\lambda_i) \arrow[r, "f"] \arrow[d, "g"] & \beta(N_1) \arrow[r] \arrow[d, "id"] & 0 \\
0 \arrow[r] & \alpha^{-1}(N_1) \arrow[r, "\alpha"] & N_1 := im(g) \arrow[r, "\beta"] & \beta(N_1) \arrow[r] & 0
\end{tikzcd}$$

N_1, and so $\beta(N_1)$, are generated by n homogeneous elements of degrees $\lambda_1,\ldots,\lambda_n$. Then, we have a surjective degree 0 graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to \beta(N_1)$ such that $f = \beta \circ g$. Therefore, $\ker f$ can be generated by $\phi(n)$ homogeneous elements of $\bigoplus_{i=1}^{n} R(-\lambda_i)$ of degrees $\mu_1,\ldots, \mu_{\phi(n)}$ in Γ. Hence, we obtain a surjective degree 0 graded R-module homomorphism $u : \bigoplus_{i=1}^{\phi(n)} R(-\mu_i) \to \ker f$. Since $f \circ u = 0$, $g \circ u\left(\bigoplus_{i=1}^{\phi(n)} R(-\mu_i)\right) \subseteq N_1 \cap \ker\beta = N_1 \cap im(\alpha)$. Thus, we have a degree 0 graded R-module homomorphism $h : \bigoplus_{i=1}^{\phi(n)} R(-\mu_i) \to \alpha^{-1}(N_1)$ that makes the diagram commute. Therefore $\ker h$, and so $\ker g$, can be generated by $\psi(\phi(n))$ elements, as desired.

(3) Assume that M is finitely presented and that N is weakly uniformly gr-coherent. We show that P is weakly uniformly gr-coherent. As N is finitely generated and M is finitely presented, P is finitely generated. Now, the map $\alpha : P \to \alpha(P)$ is a graded R-module isomorphism. By Remark 2.2, $\alpha(P)$ and so P, is weakly uniformly gr-coherent R-module, as desired. □

Corollary 2.6. *If $f : M \to N$ is a graded R-module homomorphism and M and N are weakly uniformly gr-coherent, then so are* $\ker(f)$, *$im(f)$ and $coker(f)$.*

Proof. The sequences of graded R-modules:

$$0 \to \ker(f) \hookrightarrow M \xrightarrow{f} im(f) \to 0$$

$$0 \to im(f) \hookrightarrow N \twoheadrightarrow coker(f) \to 0$$

are exact. By Remark 2.2, $im(f)$ is weakly uniformly gr-coherent (as a finitely generated homogeneous submodule of N). Then, by Theorem 2.5, $\ker(f)$ and $coker(f)$ are weakly uniformly gr-coherent, as desired. □

We next clarify the situation for scalar restrictions.

Theorem 2.7. Let $\phi : R \to S$ be a graded ring homomorphism making S a finitely generated graded R-module. Let M be a graded S-module. If M is weakly uniformly gr-coherent over R, then M is weakly uniformly gr-coherent over S.

Proof. As M is finitely generated over R, M is finitely generated over S. Let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n} S(-\lambda_i) \to M$ a graded S-module homomorphism of degree 0, where $\lambda_1,\ldots,\lambda_n$ are degrees in Γ. If the graded R-module S is generated by m homogeneous elements, then the graded R-module $\bigoplus_{i=1}^{n} S(-\lambda_i)$ is generated by mn homogeneous elements. Therefore we have a surjective degree 0 graded R-module homomorphism $g : \bigoplus_{i=1}^{mn} R(-\beta_i) \twoheadrightarrow \bigoplus_{i=1}^{n} S(-\lambda_i)$ for some degrees $\beta_1,\ldots,\beta_{mn}$ in Γ. Consider the graded R-module homomorphism $f \circ g : \bigoplus_{i=1}^{mn} R(-\beta_i) \xrightarrow{g} \bigoplus_{i=1}^{n} S(-\lambda_i) \xrightarrow{f} M$. If ϕ is a uniformity map of the weakly uniformly gr-coherent R-module M, $\ker(f \circ g)$ and so $\ker f = g(\ker f \circ g)$ can be generated by $\phi(mn)$ over R. Hence $\ker f$ can be generated by $\phi(mn)$ over S. Thus M is a weakly uniformly gr-coherent S-module, as desired. □

We close this section with a result concerning localizations of weakly uniformly gr-coherent modules.

Proposition 2.8. *Let S be a multiplicatively closed set of homogeneous elements of a graded ring R. If M is a weakly uniformly gr-coherent R-module, then $S^{-1}M$ is a weakly uniformly gr-coherent $S^{-1}R$-module.*

Proof. Since M is a finitely generated R-module, $S^{-1}M$ is a finitely generated $S^{-1}R$-module. Let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n}(S^{-1}R)(-\lambda_i) \to S^{-1}M$ a graded $S^{-1}R$-module homomorphism, where $\lambda_1,...,\lambda_n$ are degrees in Γ. Denote by $(\varepsilon_i)_{i=1}^{n}$ and $(e_i)_{i=1}^{n}$ the canonical bases of R^n and $(S^{-1}R)^n$ respectively. For $1 \le i \le n$, $f(e_i) = \frac{x_i}{s}$ for some homogeneous elements $x_i \in M$ and $s \in S$ of respective degrees μ_i and λ. Consider the (degree 0) graded R-module homomorphism $g : \bigoplus_{i=1}^{n} R(-\mu_i) \to M$, $g(\varepsilon_i) = x_i$. If ϕ is a uniformity map of the weakly uniformly gr-coherent R-module M, then $\ker g$ can be generated by $\phi(n)$ elements $\left((a_1^i,...,a_n^i)\right)_{i=1}^{\phi(n)}$. Therefore, $\ker f$ can be generated by the $\phi(n)$ elements $\left((\frac{a_1^i}{1},...,\frac{a_n^i}{1})\right)_{i=1}^{\phi(n)}$, as desired. □

3 Weakly uniformly gr-coherent rings

In this section, we study weakly uniformly gr-coherent rings.

Definition 3.1. A graded ring R is called a weakly uniformly gr-coherent ring if it is weakly uniformly gr-coherent as a graded R-module, that is, if there is a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any nonzero graded R-module homomorphism $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to R$ of degree 0, where $\lambda_1,...,\lambda_n$ are degrees in Γ, $\ker f$ can be generated by $\phi(n)$ elements (not necessary homogeneous). The map ϕ is called a uniformity map of R.

We next characterize weakly uniformly gr-coherent rings.

Theorem 3.2. Let R be a graded ring. The following assertions are equivalent:

1. R is a weakly uniformly gr-coherent ring.
2. There is a map $\psi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any homogeneous ideal I of R generated by n homogeneous elements and any homogeneous element $a \in R$, $(I : a)$ can be generated by $\psi(n)$ elements.
3. (a) There is an integer $s \in \mathbb{N}$ such that for every homogeneous element $a \in R$, $(0 : a)$ can be generated by s elements, and

   (b) There is a map $\chi : \mathbb{N}^2 \to \mathbb{N}$ such that for every $(n, m) \in \mathbb{N}^2$, and any homogeneous ideals I and J of R generated respectively by n and m homogeneous elements, $I \cap J$ can be generated by $\chi(n, m)$ elements.
4. (a) There is an integer $s \in \mathbb{N}$ such that for every homogeneous element $a \in R$, $(0 : a)$ can be generated by s elements, and

   (b) There is a map $\chi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any homogeneous ideal I of R generated by n homogeneous elements and any homogeneous element $a \in R$, $I \cap aR$ can be generated by $\chi(n)$ elements.

Before proving Theorem 3.2, we record the following two lemmas.

Lemma 3.3 ([4, Lemma 3.9])**.** *Let R be a graded ring and $u_1,...,u_{n+1}$ some homogeneous elements of R of degrees $\lambda_1,...,\lambda_{n+1}$ respectively. Set $I = (u_1,...,u_n)$ and $J = I + Ru_{n+1}$. Consider the following exact sequences of graded R-modules:*

$$0 \to \ker f \hookrightarrow \bigoplus_{i=1}^{n+1} R(-\lambda_i) \xrightarrow{f} J \to 0,$$

$$0 \to \ker g \hookrightarrow \bigoplus_{i=1}^{n} R(-\lambda_i) \xrightarrow{g} I \to 0,$$

with $f(e_j) = g(e_j) = u_j$, $1 \leq j \leq n$ and $f(e_{n+1}) = u_{n+1}$ where $\left(e_j\right)_{j=1}^{n+1}$ is the canonical basis of R^{n+1}. Then there exists a graded R-module homomorphism $\alpha : \ker f \to (I : u_{n+1})$ such that the sequence of graded R-modules $0 \to \ker g \hookrightarrow \ker f \xrightarrow{\alpha} (I : u_{n+1}) \to 0$ is exact.

Lemma 3.4 ([4, Lemma 3.10]). *Let R be a graded ring and $u_1, ..., u_{n+m}$ some homogeneous elements of R of degrees $\lambda_1, ..., \lambda_{n+m}$ respectively. Set $I = (u_1, ..., u_n)$ and $J = (u_{n+1}, ..., u_{n+m})$. Consider, as in Lemma 3.3, the following exact sequences of graded R-modules:*

$$0 \to \ker f \hookrightarrow \bigoplus_{i=1}^{n} R(-\lambda_i) \xrightarrow{f} I \to 0,$$

$$0 \to \ker g \hookrightarrow \bigoplus_{i=1}^{m} R(-\lambda_{n+i}) \xrightarrow{g} J \to 0,$$

$$0 \to \ker h \hookrightarrow \bigoplus_{i=1}^{n+m} R(-\lambda_i) \xrightarrow{h} I + J \to 0.$$

Then there exists a graded R-module homomorphism $\beta : \ker h \to I \cap J$ such that the sequence of graded R-modules $0 \to \ker f \times \ker g \hookrightarrow \ker h \xrightarrow{\beta} I \cap J \to 0$ is exact.

Proof of Theorem 3.2. (1) ⇒ (2) Let $n \in \mathbb{N}$, I a homogeneous ideal of R generated by n homogeneous elements and a a homogeneous element of R. Let $0 \to \ker f \to \bigoplus_{i=1}^{n+1} R(-\lambda_i) \xrightarrow{f} J := I + Ra \to 0$ be the exact sequence as in Lemma 3.3. Let ϕ be a uniformity map of the weakly uniformly gr-coherent ring R. Then $\ker f$, so $(I : a)$, can be generated by $\phi(n+1) =: \psi(n)$ elements (by Lemma 3.3), as desired.

(2) ⇒ (1) To show that R is a weakly uniformly gr-coherent ring, let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to R$ a graded R-module homomorphism of degree 0, where $\lambda_1, ..., \lambda_n \in \Gamma$. We use induction on n to show that $\ker f$ can be generated by $\phi(n) := \sum_{0 \leq i < n} \psi(i)$ elements where ψ is given by hypothesis (2). Set $J := im(f) = \sum_{i=1}^{n} Ru_i$ for some homogeneous elements $u_1, ..., u_n$ of R of degrees $\lambda_1, ..., \lambda_n$ respectively. For $n = 1$, hypothesis (2) yields $\ker f = (0 : u_1)$ can be generated by $\psi(0) =: \phi(1)$ elements. For $n > 1$, $J = I + Ru_n$, where $I := \sum_{i=1}^{n-1} Ru_i$. Set $0 \to \ker f \to \bigoplus_{i=1}^{n} R(-\lambda_i) \xrightarrow{f} J \to 0$ and $0 \to \ker g \to \bigoplus_{i=1}^{n-1} R(-\lambda_i) \xrightarrow{g} I \to 0$ be the exact sequences as in Lemma 3.3. By induction hypothesis, $\ker g$ can be generated by $\phi(n-1) := \sum_{0 \leq i < n-1} \psi(i)$ elements and by hypothesis (2), $(I : u_n)$ can be generated by $\psi(n-1)$ elements. Therefore, by Lemma 3.3, $\ker f$ can be generated by $\phi(n-1) + \psi(n-1) = \phi(n)$ elements, as desired.

(1) ⇒ (3) Assume that R is a weakly uniformly gr-coherent ring with uniformity map ϕ. To show (a), let $a \in R$ be a homogeneous element of degree λ. Then $f : R(-\lambda) \to R$, $f(x) = ax$ is a degree 0 graded R-module homomorphism. Therefore $\ker f = (0 : a)$ can be generated by $\phi(1)$ elements, as desired. To show (b), let $(n, m) \in \mathbb{N}^2$ and I, J be homogeneous ideals of R generated respectively by n and m homogeneous elements. Consider, as in Lemma 3.4, the exact sequence $0 \to \ker h \hookrightarrow \bigoplus_{i=1}^{n+m} R(-\lambda_i) \xrightarrow{h} I + J \to 0$. Then $\ker h$, so $I \cap J$, can be generated by $\phi(n+m) =: \chi(n, m)$ elements (by Lemma 3.4), as desired.

(3) ⇒ (4) This is straightforward.

(4) ⇒ (1) To show that R is a weakly uniformly gr-coherent ring, let $n \in \mathbb{N}$ and $h : \bigoplus_{i=1}^{n} R(-\lambda_i) \to R$ be a graded R-module homomorphism of degree 0, where $\lambda_1, ..., \lambda_n \in \Gamma$. We use induction on n to

show that $\ker h$ can be generated by $\phi(n) := sn + \sum_{1\leq i<n}\chi(i)$ elements where s and χ are given by hypothesis (4). Set $J := im(h) = \sum_{i=1}^{n} Ru_i$ for some homogeneous elements $u_1,...,u_n$ of R of respective degrees $\lambda_1,...,\lambda_n$. For $n = 1$, hypothesis (4)(a) yields $\ker h = (0 : u_1)$ can be generated by $\phi(1) := s$ elements. For $n > 1$, $J = I + Ru_n$, where $I := \sum_{i=1}^{n-1} Ru_i$. Consider, as in Lemma 3.4, the exact sequences $0 \to \ker f \to \bigoplus_{i=1}^{n-1} R(-\lambda_i) \xrightarrow{f} I \to 0$, $0 \to \ker g = (0 : u_n) \to R(-\lambda_n) \xrightarrow{g} Ru_n \to 0$ and $0 \to \ker h \hookrightarrow \bigoplus_{i=1}^{n} R(-\lambda_i) \xrightarrow{h} J = I + Ru_n \to 0$. By induction hypothesis, $\ker f$ can be generated by $\phi(n-1) := s(n-1)+\sum_{1\leq i<n-1}\chi(i)$ elements and by hypothesis (4), $(0 : u_n)$ and $I\cap Ru_n$ can be generated respectively by s and $\chi(n-1)$ elements. Then, by Lemma 3.4, $\ker h$ can be generated by $\phi(n-1)+s+\chi(n-1) = \phi(n)$ elements, as desired. □

We next compare the concepts of "weak uniform gr-coherence" and "uniform gr-coherence" for graded rings.

Proposition 3.5. *Every uniformly gr-coherent ring is a weakly uniformly gr-coherent ring.*

Proof. This is straightforward by Proposition 2.3 (2)(a). □

Let R be a graded ring. Is there a map $\phi : \mathbb{N} \to \mathbb{N}$ such that for every $n \in \mathbb{N}$, and any homogeneous ideal I of R generated by n elements, I can be generated by $\phi(n)$ homogeneous elements?

E. Wofsey answered in the negative to the above question [26]. For instance, consider the $\mathbb{Z}$-graded ring R with $R_n = \mathbb{Z}/(n)$ for each n where all products of homogenenous elements of nonzero degree are 0. Given any pairwise coprime integers $n_1,\dots,n_k$, consider the element x which is 1 in degrees $n_1,\dots,n_k$ and 0 in all other degrees. Then x generates a homogeneous ideal, since each homogeneous part of x can be written as mx for some $m \in \mathbb{Z}$ (choose m which is 1 mod n_i and 0 mod n_j for all $j \neq i$). But clearly (x) cannot be generated by fewer than k homogeneous elements. So, there are principal homogeneous ideals in R which require arbitrarily large numbers of homogeneous generators.

We have been unable to determine whether the absolute converse of Proposition 3.5 is true; however, some partial converses will be given, see Proposition 3.6 bellow and Proposition 3.10.

Proposition 3.6. *Let R be a graded ring of finite support. If R is weakly uniformly gr-coherent, then R is uniformly gr-coherent.*

Proof. This is straightforward by Proposition 2.3 (2)(b). □

We now present the relation between the notion of "weak uniform gr-coherence" and that of "uniform coherence".

Proposition 3.7. *Let R be a graded ring. If R is uniformly coherent, then it is weakly uniformly gr-coherent.*

Proof. This is straightforward by Proposition 2.3 (1). □

Every ring R can be graded trivially by Γ, via $R_0 = R$ and $R_\alpha = 0$ for $\alpha \neq 0$. It is easy to see that, for trivially graded rings, the concepts of "weak uniform graded-coherence", "uniform graded-coherence" and "uniform coherence" coincide.

Let A be a ring and let $\{X_1,...,X_n\}$ be indeterminates over A. For $m = (m_1,...,m_n) \in \mathbb{Z}^n$, let $X^m = X_1^{m_1}...X_n^{m_n}$. Then the Laurent polynomial ring $R = A[X_1, X_1^{-1},...,X_n, X_n^{-1}]$ is graded by $\mathbb{Z}^n$, via $R_m = \{aX^m | a \in A\}$ for every $m \in \mathbb{Z}^n$. The converse of Proposition 3.7 fails; in fact, there exist (weakly) uniformly gr-coherent rings which are not uniformly coherent, as shown by the following example.

Example 3.8 ([4, Example 3.4]). Let $R = K[X_1, X_1^{-1},...,X_n, X_n^{-1}]$ be the Laurent polynomial ring over a field K with $n > 2$ indeterminates. Then R is a graded-field so is (weakly) uniformly gr-coherent; but R is not uniformly coherent.

Recall from the Introduction that, a graded ring R is called a *gr-coherent ring* if every finitely generated homogeneous ideal of R is finitely presented. As an immediate consequence of Proposition 2.3 (3), we compare the concepts of "weak uniform gr-coherence" and "gr-coherence".

Proposition 3.9. *Any weakly uniformly gr-coherent ring is a gr-coherent ring.*

The converse of Proposition 3.9 fails: the easiest example is any trivially graded ring which is coherent but not uniformly coherent; a nontrivially graded example is provided by Example 3.12.

We next determine when the group ring $A[X;\Gamma]$ over a ring A, is (weakly uniformly) gr-coherent.

Proposition 3.10. *Let $R = A[X;\Gamma]$ be the group ring of Γ over a ring A graded by $deg(aX^\alpha) = \alpha$ for every $0 \neq a \in A$ and $\alpha \in \Gamma$. Then:*

1. *The following statements are equivalent.*
   - *(a) R is a gr-coherent ring.*
   - *(b) A is a coherent ring.*
2. *The following statements are equivalent.*
   - *(a) R is a weakly uniformly gr-coherent ring.*
   - *(b) R is a uniformly gr-coherent ring.*
   - *(c) A is a uniformly coherent ring.*

Lemma 3.11. *Let $R = A[X;\Gamma]$ be the group ring of Γ over a ring A. Then:*

1. *The extension mapping $\mathcal{E} : I \mapsto IR$ induces a bijection from the ideals of A to the homogeneous ideals of R.*
2. *The following statements are equivalent for an ideal I of A.*
   - *(a) IR is generated by n elements of R.*
   - *(b) IR is generated by n homogeneous elements of R.*
   - *(c) I is generated by n elements of A.*

Proof. (1) and (2) $(b) \Leftrightarrow (c)$ [4, Lemma 3.15].

(2) Let I be an ideal of A. The implication $(b) \Rightarrow (a)$ is straightforward, so it remains to show $(a) \Rightarrow (c)$. Assume that IR is generated by n elements. Denote by aug the augmentation map on R, that is, the surjective ring homomorphism $\sum_{i=1}^n a_i X^{\alpha_i} \mapsto \sum_{i=1}^n a_i$ of R onto A. Then $I = aug(IR)$ is generated by n elements, as desired. □

Proof of Proposition 3.10. (1) and (2) $(b) \Leftrightarrow (c)$ [4, Proposition 3.14].

(2) We show that $(a) \Leftrightarrow (b)$. By Lemma 3.11, an ideal I of R is generated by n elements if and only if it is generated by n homogeneous elements. Then, by Theorem 3.2, R is weakly uniformly gr-coherent if and only if it is uniformly gr-coherent, as desired. □

Armed with Proposition 3.10, we have now examples of gr-coherent rings which are not weakly uniformly gr-coherent.

Example 3.12. If A is a coherent ring which is not uniformly coherent, then the group ring $A[X;\Gamma]$ is gr-coherent but not weakly uniformly gr-coherent.

The following result concerns quotients of weakly uniformly gr-coherent rings.

Theorem 3.13. Let I be a finitely generated homogeneous ideal of a graded ring R. If R is a weakly uniformly gr-coherent ring, then so is R/I.

Proof. Since $0 \to I \hookrightarrow R \twoheadrightarrow R/I \to 0$ is an exact sequence of graded R-modules, R/I is a weakly uniformly gr-coherent R-module (by Theorem 2.5(1)). Therefore, R/I is a weakly uniformly gr-coherent ring (by Theorem 2.7), as desired. □

We next examine the product of weakly uniformly gr-coherent rings.

Proposition 3.14. *Let R_1 and R_2 be graded rings. Then $R_1 \times R_2$ is a weakly uniformly gr-coherent ring if and only if so are R_1 and R_2.*

Proof. Assume that $R_1 \times R_2$ is a weakly uniformly gr-coherent ring. As $R_1 \times 0$ is a finitely generated homogeneous ideal of $R_1 \times R_2$, $R_2 \cong \frac{R_1 \times R_2}{R_1 \times 0}$ (graded ring isomorphism) is a weakly uniformly gr-coherent ring (by Theorem 3.13).

Conversely, assume that R_1 and R_2 are weakly uniformly gr-coherent rings with uniformity maps ϕ_1 and ϕ_2 respectively. We claim that $R := R_1 \times R_2$ is a weakly uniformly gr-coherent ring with uniformity map $\phi(n) = \phi_1(n) + \phi_2(n)$. Let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n} R(-\lambda_i) \to R$ a graded R-module homomorphism of degree 0, where $\lambda_1, \dots, \lambda_n \in \Gamma$. Then $f = f_1 \times f_2$ where $f_j : \bigoplus_{i=1}^{n} R_j(-\lambda_i) \to R_j$ is a graded R_j-module homomorphism of degree 0, for $j = 1, 2$. Therefore, the graded R_j-module $\ker f_j$ can be generated by $\phi_j(n)$ elements $\left((r_j^{i1}, \dots, r_j^{in})\right)_{i=1}^{\phi_j(n)}$, for $j = 1, 2$. Hence the graded R-module $\ker f = \ker f_1 \times \ker f_2$ can be generated by the $\phi_1(n) + \phi_2(n)$ elements $\left(((r_1^{i1}, 0), \dots, (r_1^{in}, 0))\right)_{i=1}^{\phi_1(n)} \cup \left(((0, r_2^{i1}), \dots, (0, r_2^{in}))\right)_{i=1}^{\phi_2(n)}$, as claimed. □

The next result clarifies the situation for direct limits of weakly uniformly gr-coherent rings (see for instance [6, I, Exercice 12, e), p. 63]).

Theorem 3.15. Let $(R_\lambda)_{\lambda \in S}$ be a direct system of graded rings and $R := \varinjlim R_\lambda$. Assume that R is a flat R_λ-module for all $\lambda \in S$. If the R_λ's are weakly uniformly gr-coherent rings with the same uniformity map χ, then so is R.

Proof. Let $n \in \mathbb{N}$ and $f : \bigoplus_{i=1}^{n} R(-\delta_i) \to R$ a graded R-module homomorphism of degree 0, where $\delta_1, \dots, \delta_n \in \Gamma$. There exist $\lambda \in S$ and homogeneous elements $x_\lambda^1, \dots, x_\lambda^n \in R_\lambda$ of respective degrees $\delta_1, \dots, \delta_n$ such that $f(e_i) = \phi_\lambda(x_\lambda^i)$ for all $i = 1, .., n$, where $(e_i)_{i=1}^{n}$ is the canonical basis of R^n and $\phi_\lambda : R_\lambda \to R$ is the canonical mapping. Consider the degree 0 graded R_λ-module homomorphism $g : \bigoplus_{i=1}^{n} R_\lambda(-\delta_i) \to R_\lambda$, $g(\varepsilon_i) = x_\lambda^i$ for all $i = 1, .., n$, where $(\varepsilon_i)_{i=1}^{n}$ is the canonical basis of R_λ^n. Then $\ker(g)$

can be generated by $\chi(n)$ elements $\left((b_\lambda^{k,1},...,b_\lambda^{k,n})\right)_{k=1}^{\chi(n)}$ of $\bigoplus_{i=1}^n R_\lambda(-\delta_i)$. Therefore $(\phi_\lambda(b_\lambda^{k,1}),...,\phi_\lambda(b_\lambda^{k,n}))\in$ $\ker(f)$ for all $k=1,..,\chi(n)$. We claim that, $\ker(f)$ can be generated by the $\chi(n)$ elements $\left((\phi_\lambda(b_\lambda^{k,1}),...,\phi_\lambda(b_\lambda^{k,n}))\right)_{k=1}^{\chi(n)}$ of $\bigoplus_{i=1}^n R(-\delta_i)$. Let $(x^1,...,x^n)\in\ker(f)$. Then $\sum_{i=1}^n x^i\phi_\lambda(x_\lambda^i)=0$. As R is a flat R_λ-module, there exist a positive integer $m\in\mathbb{N}$, a family $(y^j)_{j=1}^m$ of elements of R and a family $(a_\lambda^{ji})_{1\le i\le n,1\le j\le m}$ of elements of R_λ such that

$$\begin{cases} \sum_{i=1}^n a_\lambda^{ji}x_\lambda^i=0, & \forall j=1,...,m,\\ x^i=\sum_{j=1}^m y^j\phi_\lambda(a_\lambda^{ji}), & \forall i=1,...,n. \end{cases}$$

Hence for every $j=1,...,m$, $(a_\lambda^{j,1},...,a_\lambda^{j,n})\in\ker(g)$ so that $(a_\lambda^{j,1},...,a_\lambda^{j,n})=\sum_{k=1}^{\chi(n)}c_\lambda^{j,k}(b_\lambda^{k,1},...,b_\lambda^{k,n})$ for some $(c_\lambda^{j,k})_{k=1}^{\chi(n)}$ of elements of R_λ. Now,

$$\begin{aligned} \left(x^1,...,x^n\right) &= \left(\sum_{j=1}^m y^j\phi_\lambda(a_\lambda^{j1}),...,\sum_{j=1}^m y^j\phi_\lambda(a_\lambda^{jn})\right)\\ &= \sum_{j=1}^m y^j\left(\phi_\lambda(a_\lambda^{j1}),...,\phi_\lambda(a_\lambda^{jn})\right)\\ &= \sum_{j=1}^m y^j\sum_{k=1}^{\chi(n)}\phi_\lambda(c_\lambda^{j,k})\left(\phi_\lambda(b_\lambda^{k,1}),...,\phi_\lambda(b_\lambda^{k,n})\right)\\ &= \sum_{k=1}^{\chi(n)}\left(\sum_{j=1}^m y^j\phi_\lambda(c_\lambda^{j,k})\right)\left(\phi_\lambda(b_\lambda^{k,1}),...,\phi_\lambda(b_\lambda^{k,n})\right), \end{aligned}$$

as claimed. □

Let $\{X_1,...,X_d\}$ be indeterminates over a ring R. For $m=(m_1,...,m_d)\in\mathbb{N}^d$, let $X^m=X_1^{m_1}...X_d^{m_d}$. Then the polynomial ring $S=R[X_1,...,X_d]$ is graded by $\mathbb{Z}$, via $S_n=\left\{\sum_{m\in\mathbb{N}^d}a_mX^m\middle|a_m\in R \text{ and } \sum_{i=1}^d m_i=n\right\}$ for $n\ge 0$ and $S_n=0$ for $n<0$.

Corollary 3.16. *Let R be a ring and let $\{X_1,X_2,...\}$ be indeterminates over R. If the polynomial rings $R[X_1,...,X_d]$, d a positive integer, are weakly uniformly gr-coherent with the same uniformity map, then so is the polynomial ring $R[X_1,X_2,...]$.*

We close by a result about localisations of weakly uniformly gr-coherent rings.

Proposition 3.17. *Let S be a multiplicatively closed set of homogeneous elements of a graded ring R. If R is a weakly uniformly gr-coherent ring, then so is $S^{-1}R$.*

Proof. This is straightforward by Proposition 2.8. □

References

[1] D. D. Anderson, D. F. Anderson and G. W. Chang, Graded-valuation domains, Comm. Algebra **45** (2017), 4018–4029.

[2] D. F. Anderson, G. W. Chang and M. Zafrullah, Graded Prüfer domains, Comm. Algebra **46** (2018), 792–809.

[3] C. Bakkari, N. Mahdou and A. Riffi, Commutative graded-coherent rings, Indian J. Math. **61** (2019), 421–440.

[4] C. Bakkari, N. Mahdou and A. Riffi, Uniformly graded-coherent rings, Quaest. Math. **44** (2021), 1371–1391.

[5] V. Barucci, D. F. Anderson and D. E. Dobbs, Coherent Mori domains and the principal ideal theorem, Comm. Algebra **15** (1987), 1119–1156.

[6] N. Bourbaki, Algèbre Commutative, Chapitres 1–4, Springer-Verlag, Berlin, 2006.

[7] N. Bourbaki, Algèbre, Chapitres 1–3, Springer-Verlag, Berlin, 2007.

[8] S. U. Chase, Direct products of modules, Trans. Amer. Math. Soc. **97** (1960), 457–473.

[9] J. M. Cohen, Coherent graded rings and the non-existence of spaces of finite stable homotopy type, Comment. Math. Helv. **44** (1969), 217–228.

[10] N. Ding, Y. L. Li and L. Mao, J-coherent rings, J. Algebra Appl. **8** (2009), 139–155.

[11] D. E. Dobbs, S. Kabbaj and N. Mahdou, n-coherent rings and modules, Commutative Ring Theory, Lect. Notes Pure Appl. Math., vol. 185, Marcel Dekker, New York/Basel, 1997, pp. 269–281.

[12] D. E. Dobbs, S. Kabbaj, N. Mahdou and M. Sobrani, When is $D+M$ n-coherent and an (n,d)-domain?, Advances in Commutative Ring Theory, Lect. Notes Pure Appl. Math., vol. 205, Marcel Dekker, New York/Basel, 1999, pp. 257–270.

[13] D. E. Dobbs and I. Papick, When is $D+M$ coherent?, Proc. Amer. Math. Soc. **56** (1976), 51–54.

[14] S. Gabelli and E. Houston, Coherent like conditions in pullbacks, Michigan Math. J. **44** (1997), 99–123.

[15] S. Glaz, Commutative Coherent Rings, Lecture notes in mathematics, vol. 1371, Springer-Verlag, Berlin, 1989.

[16] M. E. Harris, Some results on coherent rings, Proc. Amer. Math. Soc. **17** (1966), 474–479.

[17] M. E. Harris, Some results on coherent rings II, Glasgow Math. J. **8** (1967), 123–126.

[18] K. Alaoui Ismaili, D. E. Dobbs and N. Mahdou, Commutative rings and modules that are Nil_*-coherent or special Nil_*-coherent, J. Algebra Appl. **16** (2017), 1750187.

[19] K. Alaoui Ismaili and N. Mahdou, Coherence in amalgamated algebra along an ideal, Bull. Iranian Math. Soc. **41** (2015), 625–632.

[20] S. Kabbaj and N. Mahdou, Trivial extensions defined by coherent-like conditions, Comm. Algebra **32** (2004), 3937–3953.

[21] R. Nagpal and A. Snowden, Gröbner-coherent rings and modules, J. Commut. Algebra **12** (2020), 107–114.

[22] C. Nastasescu and F. Van Oystaeyen, Methods of Graded Rings, Lecture Notes in Math., vol. 1836, Springer-Verlag, Berlin, 2004.

[23] D. E. Rush, Noetherian properties in monoid rings, J. Pure Appl. Algebra **185** (2003), 259–278.

[24] J.-P. Serre, Faisceaux algébriques cohérents, Ann. of Math. **61** (1955), 197–278.

[25] J. P. Soublin, Anneaux et modules cohérents, J. Algebra **15** (1970), 455–472.

[26] E. Wofsey, Can we control the number of homogeneous generators of a f.g. homogeneous ideal?, URL:https://math.stackexchange.com/q/3578204 (version: 2020-03-12).

[27] Y. Xiang and L. Ouyang, Nil_*-coherent rings, Bull. Korean Math. Soc. **51** (2014), 579–594.