Testing Approximate Stationarity Concepts for Piecewise Affine Functions*

Lai Tian† Anthony Man-Cho So‡

Abstract

We study the basic computational problem of detecting approximate stationary points for continuous piecewise affine (PA) functions. Our contributions span multiple aspects, including complexity, regularity, and algorithms. Specifically, we show that testing first-order approximate stationarity concepts, as defined by commonly used generalized subdifferentials, is computationally intractable unless $\mathsf{P} = \mathsf{NP}$. To facilitate computability, we consider a polynomial-time solvable relaxation by abusing the convex subdifferential sum rule and establish a tight characterization of its exactness. Furthermore, addressing an open issue motivated by the need to terminate the subgradient method in finite time, we introduce the first oracle-polynomial-time algorithm to detect so-called near-approximate stationary points for PA functions.

A notable byproduct of our development in regularity is the first necessary and sufficient condition for the validity of an equality-type (Clarke) subdifferential sum rule. Our techniques revolve around two new geometric notions for convex polytopes and may be of independent interest in nonsmooth analysis. Moreover, some corollaries of our work on complexity and algorithms for stationarity testing address open questions in the literature. To demonstrate the versatility of our results, we complement our findings with applications to a series of structured piecewise smooth functions, including ρ-margin-loss SVM, piecewise affine regression, and nonsmooth neural networks.

1 Introduction

For a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, a point $\boldsymbol{x} \in \mathbb{R}^d$ is called stationary (or critical) if $\nabla f(\boldsymbol{x}) = \mathbf{0}$. However, the situation is much more complicated when the function is nondifferentiable. Indeed, there are different stationarity concepts for nonsmooth functions; we refer the reader to [29, 9] for a gentle introduction. For a locally Lipschitz function f, a classic notion of (Clarke) generalized subdifferential of f at $\boldsymbol{x}$ can be defined as

$$\partial f(\boldsymbol{x}) := \operatorname{conv}\left\{\boldsymbol{s} : \exists \boldsymbol{x}_n \to \boldsymbol{x}, \nabla f(\boldsymbol{x}_n) \text{ exists}, \nabla f(\boldsymbol{x}_n) \to \boldsymbol{s}\right\}.$$

In this paper, we consider the complexity of and robust algorithms for checking whether a given point is approximately stationary with respect to a piecewise affine (PA) function. The study of PA functions forms the foundation for investigating the analytic approximation of more general piecewise differentiable functions. Specifically, given a PA function f and a point $\boldsymbol{x}$, we want to test whether $\mathbf{0} \in \partial f(\boldsymbol{x})$ and its various approximate versions. We emphasize that "detecting" (or "testing") and "finding" (or "searching") are two different computational problems. While the co-NP-hardness of detecting the local optimality of a given point in nonconvex optimization was shown by Murty and Kabadi [33], the complexity of finding a local minimizer was an open question proposed by Pardalos and Vavasis [37], and is recently settled by Ahmadi and Zhang [2]. Furthermore, in contrast to the high efficiency of detecting stationary points for smooth function, finding such points can be difficult in general [14, 22].

1.1 Motivation The gradient method and its variants have been the workhorse in nonconvex smooth optimization. For a lower-bounded function f with a Lipschitz gradient, the standard descent lemma shows

*The full version of the paper appears with the same title on arXiv.

†Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Sha Tin, N.T., Hong Kong SAR. E-mail: tianlai.cs@gmail.com

‡Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Sha Tin, N.T., Hong Kong SAR. E-mail: manchoso@se.cuhk.edu.hk

solution concepts	degree-p polynomials		piecewise affine (PA)
	$p \leqslant 3$	$p \geqslant 4$	
local minimum	P [1, Theorem 3.3]	strongly co-NP-hard[3] [33, Theorem 2]	strongly co-NP-hard
stationary point	P (folklore)	P (folklore)	strongly NP-hard

Table 1: Complexity of deciding whether a given point belongs to a certain solution type.

that the gradient method computes a point $\boldsymbol{x}$ satisfying $\|\nabla f(\boldsymbol{x})\| \leqslant \varepsilon$ in $O(1/\varepsilon^2)$ steps. While convexity and differentiability have long been considered desirable, problems lacking both properties have recently emerged in machine learning, operations research, and statistics. For such applications, the subgradient method is the *de facto* approach employed in practice. However, its convergence behavior was not clear until very recently.

Fact (cf. [11, Corollary 5.9]). *Let a* PA *function $f : \mathbb{R}^d \to \mathbb{R}$ be given. Consider the iterates $\{\boldsymbol{x}_n\}_n$ produced by the subgradient method. Under mild assumptions on the step-size and boundedness of $\{\boldsymbol{x}_n\}_n$, every limit point, say $\boldsymbol{x}^*$, of the iterates $\{\boldsymbol{x}_n\}_n$ satisfies $\mathbf{0} \in \partial f(\boldsymbol{x}^*)$.*

These types of results [3, 30, 11] are asymptotic in nature, without any *a priori* complexity guarantees. This stands in stark contrast to the smooth and convex cases, where problem complexity is central to evaluating method efficiency. It turns out that such absence is fundamental in the sense that any *a priori* complexity guarantee for the subgradient method is impossible.

Fact (cf. [44, Theorem 2]). *For any $T \in \mathbb{N}_+$ and any chosen step-size and initial point, there exists a 6-Lipschitz, lower-bounded* PA *function $f : \mathbb{R}^2 \to \mathbb{R}$ such that the iterates $\{\boldsymbol{x}_n\}_n$ produced by the subgradient method satisfy*

$$\operatorname{dist}\Big(\mathbf{0}, \operatorname{conv} \partial f(\boldsymbol{x}_n + 0.25\mathbb{B})\Big) > 0.12, \qquad \text{for all } n \leqslant T.^{1}$$

Because proving an *a priori* guarantee is impossible, we focus on *a posteriori* analysis for the subgradient method. Given a lower-bounded PA function $f : \mathbb{R}^d \to \mathbb{R}$, by the asymptotic convergence of the subgradient method, for any $\varepsilon \geqslant 0, \delta > 0$, there exists a finite $T \in \mathbb{N}_+$ such that

$$\mathbf{0} \in \partial f(\boldsymbol{x}_T + \delta\mathbb{B}) + \varepsilon\mathbb{B}, \tag{1.1}$$

ensuring that the subgradient method will eventually bypass approximate stationary points in finite time.[2] This naturally raises the following basic question:

How can we confidently stop the subgradient method?

Unlike the smooth case, where verifying $\|\nabla f(\boldsymbol{x}_n)\| \leqslant \varepsilon$ is rarely a problem, deciding whether a point $\boldsymbol{x}_n$ is approximately stationary in the sense of (1.1) is highly nontrivial and barely explored. The main goal of this paper is to initiate the study of testing concepts of approximate stationarity for nonconvex, nonsmooth functions.

1.2 Our Results

We start with a classic fact about the representation of PA functions.

Fact (DC form; cf. [32, Proposition 4] and [26]). *Any* PA *function $f : \mathbb{R}^d \to \mathbb{R}$ can be written as the difference of two convex* PA *functions $h, g : \mathbb{R}^d \to \mathbb{R}$, i.e., $f = h - g$.*

In this paper, for a given point $\boldsymbol{x} \in \mathbb{R}^d$ and two convex PA functions $h, g : \mathbb{R}^d \to \mathbb{R}$, our main goal is to check $\mathbf{0} \in \partial(h - g)(\boldsymbol{x})$ in both exact and approximate senses.

[1] For a set S, we write $\partial f(S)$ for the set $\cup_{\boldsymbol{y} \in S} \partial f(\boldsymbol{y})$.

[2] Notably, when considering algorithms beyond subgradient method, Kornowski and Shamir [25] establish an exponential oracle complexity lower bound for any local algorithm that computes approximate stationary points in the sense of (1.1).

[3] See the footnote of [1, p. 4, Table 1].

1.2.1 Negative Results

Hardness. Given a sum or composition of smooth elemental functions, testing the (approximate) stationarity of a point relies on the applicability of classic gradient calculus rules. In modern computational environments, this can be implemented efficiently by using Algorithmic Differentiation (AD) [16] software, such as PyTorch and TensorFlow. A natural question that arises is whether testing stationarity concepts for a piecewise smooth function (e.g., the loss of a ReLU network) can be handled as efficiently as its smooth counterpart. In sharp contrast, we show that such testing for even PA functions is already computationally intractable unless $\mathsf{P} = \mathsf{NP}$.

Theorem (Informal)**.** *Fix any $\varepsilon \in [0, 1/2)$. Let two convex* PA *functions $h, g : \mathbb{R}^d \to \mathbb{R}$ and a point $\boldsymbol{x} \in \mathbb{R}^d$ be given. The following hold:*

- *Checking whether $\mathbf{0} \in \widehat{\partial}(h - g)(\boldsymbol{x}) + \varepsilon\mathbb{B}$ is strongly co-*NP*-hard.*
- *Checking whether $\mathbf{0} \in \partial(h - g)(\boldsymbol{x}) + \varepsilon\mathbb{B}$ is strongly* NP*-hard.*

In the above theorem, we use the notation $\widehat{\partial} f(\boldsymbol{x})$ to denote the Fréchet subdifferential of the function f at the point $\boldsymbol{x}$. Both $\widehat{\partial} f(\boldsymbol{x})$ and $\partial f(\boldsymbol{x})$ coincide with $\{\nabla f(\boldsymbol{x})\}$ when f is continuously differentiable at $\boldsymbol{x}$. Notably, for a PA function f, a point $\boldsymbol{x}$ satisfies $\mathbf{0} \in \widehat{\partial} f(\boldsymbol{x})$ if and only if $\boldsymbol{x}$ is a local minimum. To put the above hardness results in perspective, let us make the following remarks:

- The strong co-NP-hardness of detecting local minima for degree-p polynomials, where $p \geqslant 4$, has been established in [33, Theorem 2]. This result is confirmed to be tight in terms of degree p by [1, Theorem 3.3]. We show that for piecewise degree-p polynomials, the strong co-NP-hardness of checking local minimality emerges even when $p = 1$, that is, for piecewise affine functions; see Table 1.
- Nesterov [35] shows that deciding whether a given point is a local minimizer for a PA function is weakly co-NP-hard. However, the result in [35] only applies to the exact testing of local minimizers (i.e., $\varepsilon = 0$), and it is unclear whether the construction in [35] is DC-representable with a constant-layer composition of elemental functions (e.g., pointwise maximum, absolute value, and summation). Moreover, the reduction in [35] is from the weakly NP-complete problem of subset sum, leaving open the possibility of the existence of a pseudo-polynomial time algorithm and/or a fully-polynomial time approximation scheme (FPTAS). By contrast, our construction is a constant-layer composition of simple convex functions. Our strongly co-NP-hardness result holds with approximations and rules out the aforementioned possibilities.
- For smooth functions, there is little doubt about the high efficiency in detecting stationary points. We demonstrate that for a relatively simple class of nonsmooth functions, even approximately testing for (Clarke) stationary points is already NP-hard. Note that a Clarke stationary point of a PA function is not necessarily local minimal. To the best of our knowledge, this is the first hardness result concerning the testing of a non-minimizing first-order optimality condition.
- For PA functions, we highlight the complexity distinction between testing for a local minimal point (i.e., co-NP-hardness for $\mathbf{0} \in \widehat{\partial}(h - g)(\boldsymbol{x})$) and for a non-minimizing stationary point (i.e., NP-hardness for $\mathbf{0} \in \partial(h - g)(\boldsymbol{x})$). This distinction appears fundamental. In fact, we show that detecting a Clarke stationary point (resp. a local minimum) cannot be co-NP-hard (resp. NP-hard) unless $\mathsf{NP} = \text{co-}\mathsf{NP}$ and the Polynomial Hierarchy (PH) collapses to the second level.

We mention three notable corollaries here.

- DC-criticality, which characterizes the points to which DCA-type algorithms converge [13], has been studied for over 30 years [27]. It is well-known that DC-criticality represents a weaker notion than Clarke stationarity. We demonstrate that determining whether a given DC-critical point is Clarke stationary is NP-hard. Therefore, distinguishing between these two solution concepts is computationally intractable.
- We prove that testing so-called first-order minimality (FOM) for the abs-normal form of piecewise differentiable functions is co-NP-complete, confirming a conjecture of Griewank and Walther [18, p. 284].
- We show that detecting a (Clarke) stationary point for the loss of Convolutional Neural Networks (CNNs) is NP-hard. CNNs are one of the most popular network architectures for image classification.

Remark 1.1 (Max-Min representation)**.** *Just like convex polytopes can be described either as the convex hull of extreme points or as the intersection of finitely many halfspaces,* PA *functions also have different representations. Besides being written as the difference between two convex* PA *functions, any* PA *function can also be written as the maximin value of finitely many affine functions. Compared with the* DC *form, which could have an exponentially large number of affine pieces, a function given in* Max-Min *form can only have polynomially many pieces. Nevertheless, we show that similar hardness results still hold for the* Max-Min *representation. The* Max-Min *form appears to be less popular in real-world applications than its* DC *counterpart, perhaps due to its limited expressive power and the inconvenience of encompassing sum and multi-composite structures.*

1.2.2 Positive Results We have just shown that even approximate stationarity testing is already NP-hard. A natural strategy to proceed is to solve a relaxation of the original computational problem and isolate conditions under which such a relaxation is tight. In this paper, we consider a very intuitive relaxation by abusing the convex subdifferential sum rule[4] to nonconvex, nonsmooth functions. Specifically, we propose to check the condition $\mathbf{0} \in \partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$, rather than the NP-hard one $\mathbf{0} \in \partial(h-g)(\boldsymbol{x})$, as follows:

Sum Rule Relaxation (SRR). Given convex PA functions h, g and a point $\boldsymbol{x}$, we check the "ε-stationarity" of a PA function $h-g$ by running the following procedure:

- Compute the shortest vector $\boldsymbol{g}$ in the polytope $\partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$.
- If $\|\boldsymbol{g}\| \leqslant \varepsilon$: return True; else return False.

For convex PA functions h and g, the convex subdifferentials $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ are both non-empty polytopes. The projection in above procedure can be efficiently performed via solving a convex QP. However, despite its efficiency, such a relaxation procedure cannot always guarantee its correctness. The reason for this turns out to be quite fundamental in the field of nonsmooth analysis.

Subdifferential Sum Rule. The problem here is on the failure of exact (equality-type) subdifferential sum rule. For a locally Lipschitz function, subdifferential sum rule is only known to hold in the weak form of set inclusions rather than equalities.[5]

Fact (Clarke; cf. [8, Proposition 2.3.3])**.** *Let $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz functions. Then, for every $\boldsymbol{x} \in \mathbb{R}^d$, we have $\partial(f_1+f_2)(\boldsymbol{x}) \subseteq \partial f_1(\boldsymbol{x}) + \partial f_2(\boldsymbol{x})$.*

This weak form may hinder one from computing the subdifferential set of the concerned function. Thus, to facilitate the tractability of stationarity testing, it is of interest to establish a condition under which an equality-type sum rule holds. This would allow for efficient characterization of the subdifferential set and guarantee the correctness of our SRR method.

For convex PA functions h, g and a given point $\boldsymbol{x}$, we establish the first necessary and sufficient condition for the validity of an equality-type subdifferential sum rule $\partial(h-g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$. Our new condition is built on a new geometric property, termed *compatibility*, concerning a pair of convex polytopes.

Definition (Compatibility)**.** *Two polytopes A and B in $\mathbb{R}^d$ are called compatible if for any vectors $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$ such that $\boldsymbol{a} - \boldsymbol{b} \in \operatorname{ext}(A-B)$, we have $\boldsymbol{a} + \boldsymbol{b} \in \operatorname{ext}(A+B)$.*[6]

One of the main technical contributions of this paper is the following full characterization of the validity of exact subdifferential sum rule for PA functions. To the best of our knowledge, despite subdifferential calculus being studied for decades in the nonsmooth analysis literature [7, 39, 40, 8], this is the first time a nontrivial condition validating the (Clarke) sum rule for the difference of convex PA functions has been identified as simultaneously necessary and sufficient.

[4]For convex functions $h, g : \mathbb{R}^d \to \mathbb{R}$ and any point $\boldsymbol{x} \in \mathbb{R}^d$, we always have $\partial(h+g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) + \partial g(\boldsymbol{x})$; see [38, Theorem 23.8].
[5]For a quick example, consider $\{0\} = \partial(|\cdot| - |\cdot|)(0) \subsetneq \partial|\cdot|(0) - \partial|\cdot|(0) = [-2, 2]$.
[6]The set $\operatorname{ext}(A)$ denotes the set of extreme points of a convex set A.

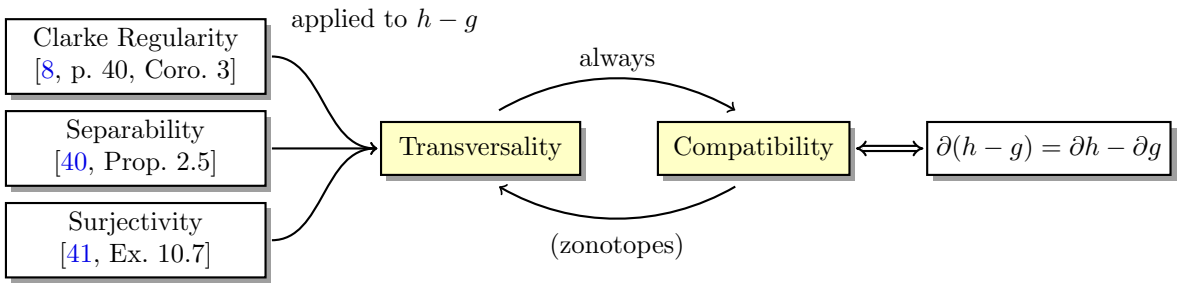

Figure 1: Interrelations of various conditions validating the exact sum rule for PA functions.

Theorem. *For convex* PA *functions* $h, g : \mathbb{R}^d \to \mathbb{R}$ *and any* $\boldsymbol{x} \in \mathbb{R}^d$*, we have* $\partial(h-g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$ *if and only if the polytopes* $\partial h(\boldsymbol{x})$ *and* $\partial g(\boldsymbol{x})$ *are compatible.*

For the intuitive SRR method, the compatibility of convex subdifferentials provides a tight characterization of its correctness. Specifically, if the polytopes $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ are incompatible, then, by [41, Exercise 8.8(c)], there exists a vector $\boldsymbol{v} \in \mathbb{R}^d$ such that the relaxation method SRR gives an incorrect answer for the PA function $h' - g$, where $h'(\boldsymbol{x}) := \boldsymbol{x}^\top \boldsymbol{v} + h(\boldsymbol{x})$. However, the time required to verify the compatibility of $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ can be exponential. This is because the concerned polytopes may contain an exponentially large number of vertices. To alleviate computational difficulties, we introduce a new polynomial-time verifiable sufficient condition, termed transversality,[7] which, when applied to PA functions, generalizes classical notions of regularity in the nonsmooth analysis literature; see Figure 1.

Definition (Transversality)**.** *Given two convex* PA *functions* $h, g : \mathbb{R}^d \to \mathbb{R}$*, we say that the functions* h *and* g *are transversal at a point* $\boldsymbol{x} \in \mathbb{R}^d$ *if*

$$\operatorname{par}(\partial h(\boldsymbol{x})) \cap \operatorname{par}(\partial g(\boldsymbol{x})) = \{\mathbf{0}\}.^{8}$$

A remarkable property of the notion of transversality, beyond its polynomial-time verifiability and sufficiency in general, is its simultaneous necessity and sufficiency when the polytopes $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ are zonotopes.[9]

Proposition (Informal)**.** *The following hold:*

- *Transversality implies compatibility.*
- *If* $\partial h(\boldsymbol{x})$ *and* $\partial g(\boldsymbol{x})$ *are zonotopes, then compatibility implies transversality.*

We note in passing that the subdifferentials of the loss for two-layer ReLU networks [47] and ramp-loss SVMs [6] are of zonotope type. Additionally, our results on regularity, which validate the sum rule, are geometric and independent of the representation of the concerned functions.

Rounding and Finite Termination. Up to this point, our focus has been exclusively on the *exact* ε-stationarity testing problem, which involves verifying whether $\mathbf{0} \in \partial(h-g)(\boldsymbol{x}) + \varepsilon\mathbb{B}$ holds for a point $\boldsymbol{x}$. However, in practice, exact nondifferentiable points are almost impossible to reach, primarily due to randomization or finite-precision limitations. Therefore, it is desirable to have a *robust* stationarity testing algorithm that works for any point sufficiently *close* to a stationary (and possibly nondifferentiable) one. In other words, we are interested in testing so-called (ε, δ)-Near-Approximate Stationarity ((ε, δ)-NAS; see (1.1) and [45, Definition 2.5]), that is, verifying the condition $\mathbf{0} \in \partial(h-g)(\boldsymbol{x} + \delta\mathbb{B}) + \varepsilon\mathbb{B}$ for a given point $\boldsymbol{x}$. See [45] and the references therein for discussion and algorithms for computing (ε, δ)-NAS points.

In our new robust testing algorithm, we only need to call the exact ε-stationarity testing procedure, i.e., checking whether $\mathbf{0} \in \partial(h-g)(\boldsymbol{x}) + \varepsilon\mathbb{B}$, in a black-box manner. Therefore, we start by abstracting this procedure for exact testing into the following oracle.

[7] See the full version for connections with existing transversality-type conditions in the literature.
[8] The set $\operatorname{par}(S)$ denotes the subspace parallel to the set S.
[9] Zonotopes are a subclass of convex polytopes.

Definition. *Given a* PA *function* $f : \mathbb{R}^d \to \mathbb{R}$ *and point* $\boldsymbol{x} \in \mathbb{R}^d$*, for* $\varepsilon \geqslant 0$*, the oracle decides whether* $\mathbf{0} \in \partial f(\boldsymbol{x}) + \varepsilon\mathbb{B}$ *or not.*

The main algorithmic contribution of this paper is a geometric scheme to certify or refute (ε, δ)-NAS for a given point. The following result guarantees the correctness and efficiency of our new robust testing approach.

Theorem (Informal)**.** *Let convex* PA *functions* $h, g : \mathbb{R}^d \to \mathbb{R}$ *and* $\varepsilon \geqslant 0$ *be given. Suppose that* $\boldsymbol{x}_n \to \boldsymbol{x}^*$ *for some unknown* $\boldsymbol{x}^* \in \mathbb{R}^d$ *satisfying* $\mathbf{0} \in \partial(h-g)(\boldsymbol{x}^*) + \varepsilon\mathbb{B}$ *and* $\|\boldsymbol{x}_n\| \leqslant B$ *for all* $n \in \mathbb{N}$*. There exists an oracle-polynomial-time algorithm that, for any* $n \in \mathbb{N}$ *and any* $\delta > 0$*, certifies either*

$$\mathbf{0} \in \partial(h-g)(\boldsymbol{x}_n + \delta\mathbb{B}) + \varepsilon\mathbb{B}, \qquad \text{or} \qquad \|\boldsymbol{x}_n - \boldsymbol{x}^*\| > \min\{\delta, \delta^*\},$$

where $\delta^* > 0$ *is a constant only dependent on* h, g, ε*, and* B*.*

One notable application of such a NAS test is to obtain an efficient termination criterion for algorithms that only have asymptotic convergence results. For example, given a lower-bounded PA function, every limiting point of the sequence generated by the subgradient method is a Clarke stationary point. However, it is still unclear when to terminate the algorithm and how to certify the obtained point is at least close to some Clarke stationary point, as the norm of any vector in the subdifferential is almost surely lower bounded away from zero along the entire trajectory (consider running the subgradient method on $x \mapsto |x|$). Our NAS testing approach offers an algorithm-independent stopping rule, effectively transforming asymptotically convergent algorithms into finite-time ones. The following immediate corollary highlights this application.

Corollary (Informal)**.** *Let two convex* PA *functions* $h, g : \mathbb{R}^d \to \mathbb{R}$ *be given. Consider the iterates* $\{\boldsymbol{x}_n\}_n$ *produced by the subgradient method on the function* $h - g$*. There exists an oracle-polynomial-time algorithm such that, for any* $\varepsilon \geqslant 0$ *and* $\delta > 0$*, the stopping criterion* $\mathbf{0} \in \partial(h-g)(\boldsymbol{x}_T + \delta\mathbb{B}) + \varepsilon\mathbb{B}$ *can be certified for a finite* $T \in \mathbb{N}_+$ *by the algorithm. Consequently, for* PA *functions, the subgradient method can be terminated confidently in finite time.*

Notably, when specialized to neural networks with ReLU activation functions, the above corollary resolves the open problem mentioned in [47, Section 5]. We also mention that, for black-box PA functions in the sense of [34], robust testing (i.e., detecting NAS points) is impossible to implement in general [44, Proposition 2].

1.3 Our Techniques

We now discuss the main challenges and new ideas in our development.

1.3.1 Computational Complexity

DC Representation. For PA functions, the set of Fréchet stationary points exactly matches the local minima. However, the existing hardness result for detecting local minima of polynomials cannot be applied here mainly due to two reasons. First, the proof for degree-p polynomials in [33, Problem 11] requires $p \geqslant 4$, which is tight according to [1, Theorem 3.3]. Despite the piecewise nature, PA functions are merely piecewise degree-1 polynomials. Second, in sharp contrast to PA functions, the set of Fréchet stationary points of a smooth function f, such as polynomials, coincides with the set $\{\boldsymbol{x} : \nabla f(\boldsymbol{x}) = \mathbf{0}\}$, whose detection is known to be in the class P [1]. As for the complexity of detecting a Clarke stationary point, let convex PA functions $h, g : \mathbb{R}^d \to \mathbb{R}$ be given. The subdifferentials $\partial h(\mathbf{0})$ and $\partial g(\mathbf{0})$ are both convex polytopes. Indeed, we can verify the condition $\mathbf{0} \in \partial h(\mathbf{0}) - \partial g(\mathbf{0})$ by solving an LP feasibility problem in polynomial time, provided a proper description of the polytopes $\partial h(\mathbf{0})$ and $\partial g(\mathbf{0})$. Therefore, the construction for proving hardness of checking whether $\mathbf{0} \in \partial(h-g)(\mathbf{0})$ must fundamentally exploit the failure of the exact subdifferential sum rule (i.e., $\partial(h-g)(\mathbf{0}) \subsetneq \partial h(\mathbf{0}) - \partial g(\mathbf{0})$). Besides, while being popular in computation, a Clarke stationary point could be non-minimizing in a meaningful sense, unlike Fréchet stationarity. To our knowledge, there is no hardness result on verifying non-minimizing first-order necessary conditions; therefore, we need some new ideas for our proof.

To obtain the desired hardness results, for the Fréchet case, we construct two simple convex PA functions h_{F} and g_{F}. Our reduction is from maximizing the ℓ_1-norm over a centered parallelotope, called PAR$\{-1,0,1\}$MAX$_1$ problem; see [4, Theorem 12]. We relate the complement problem of checking whether $\mathbf{0} \in \widehat{\partial}(h_{\mathsf{F}} - g_{\mathsf{F}})(\mathbf{0}) + \varepsilon\mathbb{B}$ to a polytope containment problem, and then to the strongly NP-hard norm-maximization PAR$\{-1,0,1\}$MAX$_1$ problem. To prove the hardness of detecting Clarke stationarity, the key construction is a seesaw-type gadget composed of two convex PA functions h_{C} and g_{C}. The gadget uses the Fréchet stationarity status of $h_{\mathsf{F}} - g_{\mathsf{F}}$ as its trigger. Specifically, when $\mathbf{0} \in \widehat{\partial}(h_{\mathsf{F}} - g_{\mathsf{F}})(\mathbf{0})$, the gadget function $h_{\mathsf{C}} - g_{\mathsf{C}}$ is affine near the point $\mathbf{0}$ with

$\|\nabla(h_{\mathsf{C}} - g_{\mathsf{C}})(\mathbf{0})\| \geqslant 1/2$. When $\mathbf{0} \notin \widehat{\partial}(h_{\mathsf{F}} - g_{\mathsf{F}})(\mathbf{0})$, the gadget will open a flat passage near the point $\mathbf{0}$ resulting $\mathbf{0} \in \partial(h_{\mathsf{C}} - g_{\mathsf{C}})(\mathbf{0})$.

The distinction in complexity between testing whether $\mathbf{0} \in \widehat{\partial}(h-g)(\mathbf{0})$ and testing whether $\mathbf{0} \in \partial(h-g)(\mathbf{0})$ can be demonstrated by showing membership in their own complexity classes. Unlike problems with explicit discrete structure, the completeness for some problems related to continuous optimization can be anything but trivial (see, e.g., [14, 22]), especially when we allow a rather rich and expressive class of convex PA components. The crux to our completeness results is a protocol to certify membership to certain index set, termed *essentially active index set* [42, p. 92]. Assuming the separation between NP and co-NP, we conclude that the distinction in computational complexity between testing Fréchet and Clarke stationarities is fundamental.

Max-Min Representation. The main challenge in proving hardness for the Max-Min form lies in its inefficiency in expressiveness. Indeed, any PA function represented in Max-Min form can have only a polynomial number of affine pieces with respect to its input size. In our proof, we build a polynomial-time reduction from 3SAT; see [15, Section 3.1.1]. For the Fréchet case, we construct a PA function f_{F} given in Max-Min form and demonstrate that the subdifferential $\widehat{\partial} f_{\mathsf{F}}(\mathbf{0})$ contains a small vector if and only if a given instance of 3SAT is unsatisfiable. For the Clarke case, we adapt the seesaw-type gadget originally developed for the DC representation to translate the hardness of testing Fréchet stationarity to that of Clarke stationarity.

1.4 Subdifferential Sum Rule For convex PA functions $h, g : \mathbb{R}^d \to \mathbb{R}$, the validity of the exact sum rule $\partial(h-g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$ at a given point $\boldsymbol{x} \in \mathbb{R}^d$ is central to understanding the correctness of the natural SRR method for stationarity testing in Section 1.2.2. Moreover, the equality-type sum rule has numerous applications in nonsmooth optimization and analysis; see, e.g., [41, Chapter 10]. The main obstacle and technical contribution in our study are mostly conceptual, focusing on identifying the correct regularity condition.

The calculus of (Clarke) subdifferentials is a well-developed area, pioneered by giants in convex and variational analysis. Monographs by Clarke [8] and Rockafellar and Wets [41] have collected and consolidated decades of development in this direction. After examining some known regularity conditions validating the sum rule for the PA function $h-g$, we are left with a vague sense that some form of separability between h and g seems necessary; see [32, p. 119] for similar remark. This observation leads us to consider a transversality-type condition, which can be seen as a generalized separability condition encompassing three known regularities as special cases when applied to PA functions. However, its unnecessity for the sum rule was initially unclear to us. Meanwhile, the discovery of the simultaneous necessity and sufficiency of transversality for zonotopes appears as strong evidence of its necessity for the general PA function $h-g$.

This belief is eventually disproven by a nontrivial example in $\mathbb{R}^4$, where the subdifferential sum rule holds without transversality. Consequently, the new notion of compatibility emerges as the correct regularity condition that fully characterizes the validity of the sum rule, with transversality serving as an appealing, polynomial-time verifiable sufficient condition. The main technical challenge in our investigation lies in characterizing how the geometry of the polytopes $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ affects the structure of the analytically defined polytope $\partial(h-g)(\boldsymbol{x})$. The goal is to identify a simple, practical, and nontrivial regularity condition that uses only information from $\partial h(\boldsymbol{x})$ and $\partial g(\boldsymbol{x})$ to ensure $\partial(h-g)(\boldsymbol{x}) = \partial h(\boldsymbol{x}) - \partial g(\boldsymbol{x})$. We find the new notion of compatibility, in some sense, neat and elegant. Moreover, both the concepts of transversality and compatibility are purely geometric and independent of the concrete representation of the convex functions h and g.

1.5 Rounding and Finite Termination Stopping criterion is important in the design of iterative algorithms for optimization problems. Our rounding algorithm is inspired by the finite-termination strategies for solving LP with interior-point methods (IPMs); see [19, Chapter 6] and [46, 31, 43]. However, our task differs significantly from that of terminating IPMs. LP is convex, and IPMs are guaranteed to converge to a global optimal solution. In contrast, our interest lies in terminating an algorithm that converges to a stationary point of a nonconvex objective function. Hence, neither convexity nor optimality gap can be exploited in algorithm design for finite termination. Moreover, the existing techniques in [46, 31, 43] crucially rely on the strict complementarity of the solution that many IPMs converge to (see [20]). However, in our stationarity testing setting, it is unrealistic to make any assumptions about the stationary point to which the algorithm is converging. Another closely related line of research is on *active manifold/constraints identification*; see [21, 28] and the references therein. These works typically require the concerned function to be Clarke regular or even amenable, ruling out general PA functions. Moreover, they can only guarantee identification if the sequence of subgradients evaluated along the

iterates approaches zero, which is almost impossible when applying the subgradient method to PA functions.

Similar to LP and many numerical algorithms, we need a standard algebraic representation for the convex functions h and g. According to [41, Theorem 2.49], every convex PA function can be written as the pointwise maximum of finitely many affine functions. However, this representation is somewhat inefficient, as the number of affine pieces grows only polynomially with respect to the input size. This lack of expressive power renders many computational problems trivial and cannot adequately address many real-world problems (e.g., the empirical loss of a shallow ReLU neural network), where a natural parameterization results in at least exponentially many pieces. In this paper, we consider a multi-composite (MC) form of convex PA functions. Simply put, the MC representation allows for arbitrarily deep compositions and summations of pointwise maximum over affine functions. An MC function is, by definition, convex, piecewise affine, and can have exponentially many affine pieces. As discussed in [42, Proposition 2.2.3], every PA function $f : \mathbb{R}^d \to \mathbb{R}$ defines a corresponding polyhedral subdivision of $\mathbb{R}^d$. The PA function f is actually affine on every element of the polyhedron subdivision. Given a point $\boldsymbol{w} \in \mathbb{R}^d$, there can be more than exponentially many polyhedra in the subdivision near $\boldsymbol{w}$, whose enumeration is clearly computational intractable. The key component of our algorithm is a family of convex polyhedra $\{P^\delta\}_\delta$, parameterized by a positive scalar δ, which governs the process of searching for the unknown stationary point (or certifying its absence).

Let us informally sketch the idea behind our new robust testing algorithm. Let convex PA functions $h, g : \mathbb{R}^d \to \mathbb{R}$ and a point $\boldsymbol{w} \in \mathbb{R}^d$ be given. Fix any desired precision $\varepsilon \geqslant 0$ and $\delta > 0$. Our goal is to identify an unknown stationary point $\boldsymbol{w}^* \in \mathbb{R}^d$ near the point $\boldsymbol{w}$ or certify its absence. To convey the intuition informally, imagine we are standing at the point $\boldsymbol{w}$, swinging a butterfly net in an attempt to capture the elusive $\boldsymbol{w}^*$. The length of the net is adjustable, set by a positive scalar $\delta_0 := \delta$, and the shape of its opening is also adjustable, polyhedral, and parameterized by δ_0. Let $P^{\delta_0} \subseteq \mathbb{R}^d$ represent the opening. We swing the net with δ_0 by projecting the point $\boldsymbol{w}$ onto the polyhedral P^{δ_0}. Suppose that the projection is $\widehat{\boldsymbol{w}} \in P^{\delta_0}$. If $\|\boldsymbol{w} - \widehat{\boldsymbol{w}}\| \leqslant \delta$ and $\mathbf{0} \in \partial(h - g)(\widehat{\boldsymbol{w}}) + \varepsilon\mathbb{B}$, which is efficiently verifiable by the ε-stationarity testing oracle, then we succeed by confirming $\boldsymbol{w}$ as an (ε, δ)-NAS point with the certification $\widehat{\boldsymbol{w}}$. If not, then we halve δ_0 to $\delta_0/2$ and repeat the capturing process. This loop will not continue indefinitely; we will stop the iteration after at most a polynomial number of steps. The crux of the correctness of our algorithm is that if an identification condition is satisfied, then we are guaranteed to capture the point $\boldsymbol{w}^* \in P^{\delta_0}$ for some positive δ_0, and the projection $\widehat{\boldsymbol{w}}$ satisfies $\partial(h - g)(\widehat{\boldsymbol{w}}) = \partial(h - g)(\boldsymbol{w}^*)$. This enables us to verify the (ε, δ)-NAS status of $\boldsymbol{w}$ without precisely pinpointing the original target $\boldsymbol{w}^*$. On the contrary, if the identification condition is not satisfied before termination, then we can certify that the candidate point $\boldsymbol{w}$ is at least $\min\{\delta, \delta^*\}$ away from any ε-stationary point, where δ^* is a constant independent of the point $\boldsymbol{w}$.

1.6 Related Work

Complexity of Testing Solutions. There have been many works on the complexity of deciding whether a given point belongs to a certain solution type. It is shown in [33, 36] that checking whether a given point is local minimum, or strict local minimum for unconstrained or simply constrained problem are both co-NP-hard. For low-degree polynomials, the work [1] shows that it is possible to efficiently determine the existence and membership of local minima. The work [5, Theorem 4] demonstrates that certifying the singleton of the Clarke subdifferential set is NP-hard. The work [35] shows that deciding whether a given point is a local minimizer for a PA function is (weakly) co-NP-hard. However, the result in [35] only applies to the exact testing of local minimizers, and it is unclear whether the construction is DC-representable with constant-layer MC components. Moreover, the reduction in [35] is from the subset sum problem, leaving open the possibility of detecting local minimizers in pseudo-polynomial time and the existence of a fully-polynomial time approximation scheme (FPTAS). For any piecewise smooth function representable by the so-called abs-normal form, the work [17] shows that a first-order optimality condition, also called first-order minimality (FOM) in [18], can be verified efficiently under a linear independence constraint qualification (LICQ)-type condition, termed linear independence kink qualification (LIKQ). When applied to PA functions, this LIKQ condition has a close relation to a surjectivity-type assumption. Another related work to ours is the one by Yun et al. [47]. They consider the empirical loss of a two-layer ReLU network and introduce a theoretical algorithm to check Clarke stationarity, which is essentially similar to the SRR method presented in Section 1.2.2. A limitation of the work [47] (discussed in [47, Section 5]) is that the algorithm therein can only perform *exact* stationarity testing. That is to say, if the objective function is $x \mapsto |x|$, then the algorithm in [47] will certify approximate stationarity if and only if the given point is exactly equal to

zero. Our results resolve the open problem on *robust* testing mentioned in [47, Section 5] for more general PA functions.

Nonsmooth Optimization. The convergence of subgradient-type methods to Clarke stationary points is primarily analyzed from a continuous-time perspective, as studied in [3, 30, 11]. These results are only asymptotic, lacking any oracle complexity guarantees. The work [44] shows that this absence is fundamental; even for PA functions, the trajectory of the subgradient method cannot bypass an approximate stationary point in any *a priori* finite time. For the nonasymptotic aspect, the development of iterative methods for solving nonconvex nonsmooth optimization problems is still in its early stages. The work [10] shows that for weakly convex, Lipschitz functions, (ε, δ)-NAS points can be obtained using a subgradient-type method. However, the requirement of weak convexity is somewhat stringent and excludes general PA functions. Recently, there has been a surge of research aimed at obtaining oracle complexity results for general Lipschitz functions; see, e.g., [48, 45, 12]. An important notion of approximate stationarity adopted in these works is the so-called Goldstein approximate stationarity (GAS). Formally, a point $\boldsymbol{x}$ is called an (ε, δ)-GAS point for a function f if $\mathbf{0} \in \operatorname{conv} \partial f(\boldsymbol{x} + \delta\mathbb{B}) + \varepsilon\mathbb{B}$. It is easy to see that (ε, δ)-NAS, in the sense of (1.1), is a more stringent, and therefore more desirable, solution concept than (ε, δ)-GAS. However, the work [25] establishes an exponential lower bound on the oracle complexity for computing (ε, δ)-NAS points by local algorithms. In comparison, the works [48, 45, 12] propose randomized algorithms that compute (ε, δ)-GAS points with dimension-free oracle complexity. Interestingly, the works [44, 23] show that no deterministic algorithm can achieve the same for computing (ε, δ)-GAS points. By restricting the function class with a so-called nonconvexity modulus, the work [24] presents a deterministic algorithm that computes (ε, δ)-GAS points with dimension-free oracle complexity.

2 Closing Remarks

We explore the computational complexity, regularity conditions, and the development of robust algorithms for testing approximate stationary points in the context of piecewise affine functions. Our findings reveal the computational hardness in testing various first-order approximate stationarity concepts. We establish the first necessary and sufficient condition for the validity of an equality-type (Clarke) subdifferential sum rule, which applies to a specific representation of arbitrary piecewise affine functions. Additionally, we introduce the first oracle-polynomial-time algorithm designed to detect near-approximate stationary points for piecewise affine functions, which offers an efficient algorithm-independent stopping rule. These results are complemented with applications to various structured piecewise smooth functions. As for open research directions, for piecewise affine functions with a fixed input dimension, it is unclear whether the hardness results still hold. It would be interesting to study the necessary and sufficient conditions for the validity of calculus rules for a broader class of nonsmooth functions. Additionally, developing an algorithm-independent, potentially non-constructive, and robust testing approach for more general piecewise smooth functions would be an intriguing direction for further exploration.

Acknowledgement

LT would like to thank Jiewen Guan (CUHK) for his careful reading of a previous version of the manuscript.

References

[1] Amir Ali Ahmadi and Jeffrey Zhang. Complexity aspects of local minima and related notions. *Advances in Mathematics*, 397:108119, 2022.

[2] Amir Ali Ahmadi and Jeffrey Zhang. On the complexity of finding a local minimizer of a quadratic function over a polytope. *Mathematical Programming*, 195:783–792, 2022.

[3] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[4] Hans L. Bodlaender, Peter Gritzmann, Victor Klee, and Jan Van Leeuwen. Computational complexity of norm-maximization. *Combinatorica*, 10:203–225, 1990.

[5] Jérôme Bolte, Ryan Boustany, Edouard Pauwels, and Béatrice Pesquet-Popescu. On the complexity of nonsmooth automatic differentiation. In *International Conference on Learning Representations*, 2022.

[6] J. Paul Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2):467–479, 2011.

[7] Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.

[8] Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[9] Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. SIAM, 2021.

[10] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[11] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[12] Damek Davis, Dmitriy Drusvyatskiy, Yin Tat Lee, Swati Padmanabhan, and Guanghao Ye. A gradient sampling method with complexity guarantees for Lipschitz functions in high and low dimensions. In *Advances in Neural Information Processing Systems*, 2022.

[13] Welington de Oliveira. The ABC of DC programming. *Set-Valued and Variational Analysis*, 28:679–706, 2020.

[14] John Fearnley, Paul Goldberg, Alexandros Hollender, and Rahul Savani. The complexity of gradient descent: CLS = PPAD ∩ PLS. *Journal of the ACM*, 70(1):1–74, 2022.

[15] Michael R. Garey and David S. Johnson. *Computers and Intractability*, volume 174. 1979.

[16] Andreas Griewank and Andrea Walther. *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. SIAM, 2008.

[17] Andreas Griewank and Andrea Walther. First-and second-order optimality conditions for piecewise smooth objective functions. *Optimization Methods and Software*, 31(5):904–930, 2016.

[18] Andreas Griewank and Andrea Walther. Relaxing kink qualifications and proving convergence rates in piecewise smooth optimization. *SIAM Journal on Optimization*, 29(1):262–289, 2019.

[19] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer Science & Business Media, 2012.

[20] Osman Güler and Yinyu Ye. Convergence behavior of interior-point algorithms. *Mathematical Programming*, 60:215–228, 1993.

[21] Warren L. Hare and Adrian S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[22] Alexandros Hollender and Emmanouil Zampetakis. The computational complexity of finding stationary points in non-convex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5571–5572. PMLR, 2023.

[23] Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4570–4597. PMLR, 2023.

[24] Siyu Kong and Adrian S. Lewis. The cost of nonconvexity in deterministic nonsmooth optimization. *Mathematics of Operations Research*, 2023.

[25] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

[26] A. Kripfganz and R. Schulze. Piecewise affine functions as a difference of two convex functions. *Optimization*, 18(1):23–29, 1987.

[27] Hoai An Le Thi and Tao Pham Dinh. DC programming and DCA: Thirty years of developments. *Mathematical Programming*, 169:5–68, 2018.

[28] Adrian S. Lewis and Stephen J. Wright. Identifying activity. *SIAM Journal on Optimization*, 21(2):597–614, 2011.

[29] Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in nonsmooth optimization: A guided tour of various constructions of subdifferential for nonsmooth functions. *IEEE Signal Processing Magazine*, 37(5):18–31, 2020.

[30] Szymon Majewski, Błażej Miasojedow, and Eric Moulines. Analysis of nonsmooth stochastic approximation: The differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.

[31] Sanjay Mehrotra and Yinyu Ye. Finding an interior point in the optimal face of linear programs. *Mathematical Programming*, 62:497–515, 1993.

[32] D. Melzer. On the expressibility of piecewise-linear continuous functions as the difference of two piecewise-linear convex functions. *Mathematical Programming Studies*, 29:118–134, 1986.

[33] Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987.

[34] Arkadi Nemirovski and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[35] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140:125–161, 2013.

[36] Panos M. Pardalos and Georg Schnitger. Checking local optimality in constrained quadratic programming is NP-hard. *Operations Research Letters*, 7(1):33–35, 1988.

[37] Panos M. Pardalos and Stephen A. Vavasis. Open questions in complexity theory for numerical optimization. *Mathematical Programming*, 57:337–339, 1992.

[38] R. T. Rockafellar. *Convex Analysis*, volume 18. Princeton University Press, 1970.

[39] R. T. Rockafellar. Directionally Lipschitzian functions and subdifferential calculus. *Proceedings of the London Mathematical Society*, 3(2):331–355, 1979.

[40] R. T. Rockafellar. Extensions of subgradient calculus with applications to optimization. *Nonlinear Analysis: Theory, Methods & Applications*, 9(7):665–698, 1985.

[41] R. T. Rockafellar and R. J-B Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[42] Stefan Scholtes. *Introduction to Piecewise Differentiable Equations*. Springer Science & Business Media, 2012.

[43] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of termination of linear programming algorithms. *Mathematical Programming*, 97:375–404, 2003.

[44] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians. *Mathematical Programming*, 208:51–74, 2024.

[45] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of Lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

[46] Yinyu Ye. On the finite convergence of interior-point algorithms for linear programming. *Mathematical Programming*, 57:325–335, 1992.

[47] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Efficiently testing local optimality and escaping saddles for ReLU networks. In *International Conference on Learning Representations*, 2019.

[48] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Ali Jadbabaie, and Suvrit Sra. Complexity of finding stationary points of nonsmooth nonconvex functions. In *International Conference on Machine Learning*, pages 11173–11182, 2020.