Testing Approximate Stationarity Concepts for Piecewise Affine Functions and Extensions

Lai Tian TIANLAI@SE.CUHK.EDU.HK
Anthony Man-Cho So MANCHOSO@SE.CUHK.EDU.HK
The Chinese University of Hong Kong

Abstract

We study various aspects of the fundamental computational problem of detecting approximate stationary points for piecewise affine (PA) functions, including computational complexity, regularity conditions, and robustness in implementation. Specifically, for a PA function, we show that testing first-order approximate stationarity concepts in terms of three commonly used subdifferential constructions is computationally intractable unless $\mathsf{P} = \mathsf{NP}$. To facilitate computability, we establish the first necessary and sufficient condition for the validity of an equality-type (Clarke) subdifferential sum rule for a certain representation of arbitrary PA functions. Our main tools are nonsmooth analysis and polytope theory. Moreover, to address an important implementation issue, we introduce the first oracle-polynomial-time algorithm to test near-approximate stationarity for PA functions. We complement our results with extensions to other subdifferentials and applications to a series of structured piecewise smooth functions, including ρ-margin-loss SVM, piecewise affine regression, and neural networks with nonsmooth activation functions.

Overview of Main Results[1]

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 f is nondifferentiable at $\boldsymbol{x}$. Indeed, there are many different stationarity concepts for nonsmooth functions [2, 4]. 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 differentiable function. When specialized to the two-layer ReLU neural networks, this is a task already considered by Yun et al. [11]. We emphasize that "detecting" and "finding" are two very 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 [5] in 1987, the complexity of "finding" a local minimizer was an open question proposed by Pardalos and Vavasis [7] since 1992, and is recently settled negatively by Ahmadi and Zhang [1].

With finite computational resources, an algorithm usually computes an approximate solution rather than an exact one,[2] though the precision can sometimes be arbitrarily high. When the objective function f is smooth, a point $\boldsymbol{x}$ is called ε-stationary if $\|\nabla f(\boldsymbol{x})\| \leqslant \varepsilon$. If the function f is nonsmooth, the notion of approximation becomes very subtle because different approximation schemes have varying degrees of computability. Specifically, we are interested in testing the following two solution notions:

[1] The full version of the paper appears in arXiv with the same title.

[2] With algorithms for linear programming as a remarkable exception.

© L. Tian & A.M.-C. So.

(ε-Stationary Point) We call a point $\boldsymbol{x}$ a (Clarke) ε-stationary point of f if $\mathbf{0} \in \partial f(\boldsymbol{x}) + \varepsilon\mathbb{B}$.

((ε, δ)-Near-Approximately Stationary (NAS) Point) If there exists a point $\boldsymbol{y} \in \mathbb{B}_\delta(\boldsymbol{x})$ such that $\boldsymbol{y}$ is ε-stationary, we call $\boldsymbol{x}$ an (ε, δ)-Near-Approximate Stationary (NAS) point. In other words, a point $\boldsymbol{x}$ is an (ε, δ)-NAS point if $\mathbf{0} \in \partial f(\boldsymbol{x} + \delta\mathbb{B}) + \varepsilon\mathbb{B}$.

A key motivation for the approximate stationarity testing problem is its role as a universal stopping rule. This is especially significant as nearly all convergence results for a general ReLU network are asymptotic, lacking a finite-time algorithm with theoretical guarantee. Our NAS testing approach offers an algorithm-independent stopping rule, effectively transforming asymptotically convergent algorithms into finite-time ones. We also mention that, for a black-box function in the sense of [6], robust testing is impossible to implement in general [10, Theorem 2.7].

A. Contributions.

In summary, our contributions are as follows:

- We demonstrate that checking first-order approximate stationarity concepts for a piecewise affine function, in terms of any one of three commonly used subdifferential constructions, is computationally intractable. As a corollary, we prove that testing the First-Order Minimality (FOM) for the abs-normal form of piecewise differentiable functions is co-NP-complete, confirming a conjecture by Griewank and Walther [3, p. 284].
- We establish the first necessary and sufficient condition for the validity of an equality-type (Clarke) subdifferential sum rule for a certain representation of arbitrary piecewise affine functions. Our new condition is a geometric property concerning a pair of convex polytopes. Moreover, we introduce a new polynomial-time verifiable sufficient condition, which becomes simultaneously necessary and sufficient when a polytope related to the subdifferential set is a zonotope.
- To address an important implementation issue, we introduce the first oracle-polynomial-time algorithm to test near-approximate stationarity for a piecewise affine function. When specialized to the loss of two-layer neural networks with ReLU-type activation function, our new algorithm is the first practical and robust stationarity test approach, resolving an open issue in the work of Yun et al. [11].
- We complement our results with extensions to other subdifferentials and applications to a series of structured piecewise smooth functions, including ρ-margin-loss SVM, piecewise affine regression, and shallow/deep neural networks with nonsmooth activation functions.

In this extended abstract, we provide an informal overview of our main results. For formal results and additional details, please refer to the full paper.

B. Computational Hardness.

For smooth optimization, co-NP-hardness has been shown for detecting local optimality [5, Theorem 2] and checking second-order sufficient condition [5, Theorem 4]. However, in the nonsmooth case, little is known about the complexity of detecting various stationarity concepts. In the following, we demonstrate that even checking a first-order necessary condition approximately in terms of various subdifferential constructions is already computationally intractable unless $\mathsf{P} = \mathsf{NP}$.

Theorem 1 (Hardness) *Given a piecewise affine function $f : \mathbb{R}^d \to \mathbb{R}$ in the form of MAX-MIN*[3] *representation with integer data. Determining whether* $\operatorname{dist}\big(\mathbf{0}, A\big) \leqslant \varepsilon$ *is*

(a) *strongly co-*NP*-hard, when the set A is $\widehat{\partial} f(\mathbf{0})$ and $\varepsilon \in [0, +\infty) \cap \mathbb{Q}$;*

(b) *strongly* NP*-hard, when the set A is $\partial_L f(\mathbf{0})$ and $\varepsilon \in [0, 1/2) \cap \mathbb{Q}$;*

(c) *strongly* NP*-hard, when the set A is $\partial f(\mathbf{0})$ and $\varepsilon \in [0, 1/2) \cap \mathbb{Q}$.*

We compare Theorem 1 with the classic hardness result of Murty and Kabadi [5]. In [5], checking the local optimality of a point for a simply constrained indefinite quadratic problem [5, Problem 1] and for an unconstrained quartic polynomial objective [5, Problem 11] are both co-NP-complete. However, these hardness results are inapplicable for checking first-order necessary conditions. In fact, for any hard construction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ in [5] and a given point $\boldsymbol{x} \in \mathbb{Q}^n$, testing ε-stationarity can be done in polynomial time with respect to the input size. Quite the contrary, in Theorem 1, we show that for unconstrained problems with piecewise affine objective functions, even an approximate test of the first-order necessary condition for a certain point $\boldsymbol{x} = \mathbf{0}$ is already computationally intractable unless $\mathsf{P} = \mathsf{NP}$.

We present two noteworthy corollaries of Theorem 1. The first corollary of our hard construction concerns the complexity of checking an optimality condition for functions representable in abs-normal form [3, Definition 2.1]. The following result gives an affirmative answer to a conjecture of Griewank and Walther [3, p. 284]:

Corollary 2 (Abs-normal form) *Testing First-Order Minimality (FOM) for a piecewise differentiable function given in the abs-normal form is co-*NP*-complete.*

Another notable corollary is about the complexity of detecting an approximately (Clarke) stationary point for the loss of a shallow modern convolutional neural network.

Corollary 3 (Convolutional neural networks) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be the loss function of a shallow convolutional neural network with ReLU activation function and max-pooling operator. Then, for any $\varepsilon \in [0, 1/2) \cap \mathbb{Q}$, testing the (Clarke) ε-stationarity* $\operatorname{dist}\big(\mathbf{0}, \partial f(\boldsymbol{\theta})\big) \leqslant \varepsilon$ *for given network parameters $\boldsymbol{\theta} \in \mathbb{Q}^d$ is strongly* NP*-hard.*

C. Exact Subdifferential Sum Rules

The computational hardness results for testing is partly due to the failure of exact subdifferential calculus rules. For a general locally Lipschitz function, the calculus rules are only known to hold in the weak form of set inclusions rather than equalities, except in several special cases (see [8, Chapter 10]). Thus, to enhance the tractability of stationarity testing, it is of interest to find out a condition, under which an equality-type calculus rule holds, so that the subdifferential set of the function can be efficiently characterized.

In this section, we study the regularity conditions facilitating efficient ε-stationarity testing for piecewise affine functions. In particular, we focus on PA functions $f : \mathbb{R}^d \to \mathbb{R}$ represented in the following Difference of Sum-of-Max (DSM) form:

$$f(\boldsymbol{w}) := \left(\sum_{p=1}^{P} \max_{1 \leqslant i \leqslant n_p} \boldsymbol{w}^\top \boldsymbol{x}_{p,i} + a_{p,i} \right) - \left(\sum_{q=1}^{Q} \max_{1 \leqslant j \leqslant m_q} \boldsymbol{w}^\top \boldsymbol{y}_{q,j} + b_{q,j} \right). \tag{DSM}$$

[3] See [9, Proposition 2.2.2] for details.

One can immediately observe that any DSM function is a difference-of-convex function and is typically subdifferential irregular (see [8, Definition 7.25]). We aim to investigate the validity of the following equality for subdifferential of a DSM function $f:\mathbb{R}^d\to\mathbb{R}$:

$$\partial f(\boldsymbol{w})=\sum_{p=1}^{P}\operatorname{conv}\{\boldsymbol{x}_{p,i}:i\in\mathcal{A}_{1,p}(\boldsymbol{w})\}-\sum_{q=1}^{Q}\operatorname{conv}\{\boldsymbol{y}_{q,j}:j\in\mathcal{A}_{2,q}(\boldsymbol{w})\},\tag{1}$$

where the active sets $\mathcal{A}_{1,p}(\boldsymbol{w})$ and $\mathcal{A}_{2,p}(\boldsymbol{w})$ are defined by

$$\mathcal{A}_{1,p}(\boldsymbol{w}):=\operatorname*{argmax}_{1\leqslant i\leqslant n_p}\ \boldsymbol{w}^\top\boldsymbol{x}_{p,i}+a_{p,i},\qquad \mathcal{A}_{2,q}(\boldsymbol{w}):=\operatorname*{argmax}_{1\leqslant j\leqslant m_q}\ \boldsymbol{w}^\top\boldsymbol{y}_{q,j}+b_{q,j},$$

for any p and q. Given above exact sum rule in Equation (1), one can readily see that computing the quantity $\operatorname{dist}(\boldsymbol{0},\partial f(\boldsymbol{w}))\in\mathbb{R}_+$ for rational data is a special case of convex quadratic optimization problem. It is well-known that such a problem can be approximately solved in polynomial time. Therefore, efficient ε-stationarity detection for DSM functions can be guaranteed by an exact subdifferential sum rule.

The goal of this section is to establish the necessary and sufficient condition for the validity of exact subdifferential sum rule in Equation (1). To this end, we introduce a geometric condition concerning polytope pairs.

Definition 4 (Compatible polytopes) *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 $\boldsymbol{a}-\boldsymbol{b}\in\operatorname{ext}(A-B)$, it holds $\boldsymbol{a}+\boldsymbol{b}\in\operatorname{ext}(A+B)$.*

One of our main results is the following characterization of the validity of exact sum rule.

Theorem 5 (Exact sum rule for DSM) *Let a DSM function $f:\mathbb{R}^d\to\mathbb{R}$ with data $\{(\boldsymbol{x}_{p,i},a_{p,i})\}_{p,i}$ and $\{(\boldsymbol{y}_{q,j},b_{q,j})\}_{q,j}$ be given. Fix a point $\boldsymbol{w}$. Define polytopes X and Y as*

$$X:=\sum_{p=1}^{P}\operatorname{conv}\{\boldsymbol{x}_{p,i}:i\in\mathcal{A}_{1,p}(\boldsymbol{w})\},\quad Y:=\sum_{q=1}^{Q}\operatorname{conv}\{\boldsymbol{y}_{q,j}:j\in\mathcal{A}_{2,q}(\boldsymbol{w})\}.$$

The following are equivalent.

(a) $\partial f(\boldsymbol{w})=X-Y$.

(b) *X and Y are compatible polytopes.*

D. Robustness in Inplementation

Up to this point, our focus has been exclusively on the *exact* ε-stationarity testing problem, which involves verifying whether $\boldsymbol{0}\in\partial f(\boldsymbol{w})+\varepsilon\mathbb{B}$ holds for a point $\boldsymbol{w}$. However, in practice, exact nondifferentiable points are almost impossible to reach, primarily due to algorithmic randomization or finite-precision limitations. Therefore, a need arises for a *robust* stationarity testing approach. This approach, when queried at a point $\boldsymbol{w}$, certifies $\boldsymbol{0}\in\partial f(\boldsymbol{w}+\delta\mathbb{B})+\varepsilon\mathbb{B}$ for given precisions $\varepsilon,\delta\geqslant 0$, if $\boldsymbol{w}$ is sufficiently close to an ε-stationary point $\boldsymbol{w}^*$. In this section, we present the main algorithmic results of this paper regarding robust testing for general DSM functions.

We begin by abstracting the procedure for exact ε-stationarity testing into the following oracle.

Definition 6 (Oracle) *A stationarity test oracle for a DSM function f is an oracle whose input are vectors $\{(\boldsymbol{x}_{p,i},a_{p,i})\}_{p,i\in[n_p]},\{(\boldsymbol{y}_{q,j},b_{q,j})\}_{q,j\in[m_q]}\subseteq\mathbb{Q}^{d+1}$, a point $\boldsymbol{w}\in\mathbb{Q}^d$, and a rational number $\varepsilon\geqslant 0$. This oracle decides* $\operatorname{dist}\big(\boldsymbol{0},\partial f(\boldsymbol{w})\big)\leqslant\varepsilon$ *or not.*

Problem Setup. We adopt a constructive viewpoint, wherein we certify the (ε, δ)-NAS status of a point $\boldsymbol{w}$ for a DSM function f only when we find a point $\boldsymbol{w}^* \in \mathbb{B}_\delta(\boldsymbol{w})$ that satisfies $\boldsymbol{0} \in \partial f(\boldsymbol{w}^*) + \varepsilon\mathbb{B}$. Note that if a point $\boldsymbol{w}^* \in \mathbb{B}_\delta(\boldsymbol{w})$ successfully passes the exact stationarity test using an implementation of the oracle in Definition 6, then $\boldsymbol{w}$ must indeed be an (ε, δ)-NAS point. In other words, there are no false positives in this test. The question that arises is whether, if $\boldsymbol{w}$ is sufficiently close to an ε-stationary point, we can *always* find a point $\boldsymbol{w}^*$ near $\boldsymbol{w}$ that is ε-stationary. In other words, we aim to control the occurrence of false negatives in our robust test.

Suppose we have an ε-stationary point $\boldsymbol{w}^* \in \mathbb{R}^d$. When a point $\boldsymbol{w}$ is sufficiently close (within a distance of δ) to $\boldsymbol{w}^*$, our objective is to certify the condition: $\boldsymbol{0} \in \partial f(\boldsymbol{w} + \delta\mathbb{B}) + \varepsilon\mathbb{B}$. To quantify the required degree of proximity, we introduce the following separation constants for DSM functions.

Definition 7 (Separation) *Let* $R := \max\{\|\boldsymbol{x}_{p,i}\|, \|\boldsymbol{y}_{q,j}\|\}_{p,q,i,j}$ *and a point* $\boldsymbol{w}$ *be given. We define functions* $v_p : \mathbb{R}^d \to \mathbb{R}$ *and* $u_q : \mathbb{R}^d \to \mathbb{R}$ *for any* $p \in [P]$ *and* $q \in [Q]$ *as*

$$v_p(\boldsymbol{w}) := \max_{1 \leqslant i \leqslant n_p} \boldsymbol{x}_{p,i}^\top \boldsymbol{w} + a_{p,i}, \quad u_q(\boldsymbol{w}) := \max_{1 \leqslant j \leqslant m_q} \boldsymbol{y}_{q,j}^\top \boldsymbol{w} + b_{q,j}.$$

We also define the extended-real-valued function $\delta_{\mathsf{sep}} : \mathbb{R}^d \to \mathbb{R}_{++} \cup \{+\infty\}$ *as*

$$\begin{aligned} \delta_{\mathsf{sep}}(\boldsymbol{w}) := \sup\ & \frac{\gamma}{12R} \\ \text{s.t. } & \gamma \leqslant v_p(\boldsymbol{w}) - \boldsymbol{x}_{p,i}^\top \boldsymbol{w} - a_{p,i}, \qquad \forall p \in [P], i \in [n_p] : v_p(\boldsymbol{w}) \neq \boldsymbol{x}_{p,i}^\top \boldsymbol{w} + a_{p,i}, \\ & \gamma \leqslant u_q(\boldsymbol{w}) - \boldsymbol{y}_{q,j}^\top \boldsymbol{w} - b_{q,j}, \qquad \forall q \in [Q], j \in [m_q] : u_q(\boldsymbol{w}) \neq \boldsymbol{y}_{q,j}^\top \boldsymbol{w} + b_{q,j}. \end{aligned}$$

We can now formulate the robust stationarity testing problem for DSM functions as follows.

Robust Stationarity Testing of DSM (RST-DSM)

Instance. A DSM function f with data $\{(\boldsymbol{x}_{p,i}, a_{p,i})\}_{p \in [P], i \in [n_p]}, \{(\boldsymbol{y}_{q,j}, b_{q,j})\}_{q \in [Q], j \in [m_q]} \subseteq \mathbb{Q}^{d+1}$ and $P, Q, n_p, m_q \in \mathbb{N}$. A point $\boldsymbol{w} \in \mathbb{Q}^d$. The precision parameters $0 \leqslant \varepsilon, \delta \in \mathbb{Q}$.

Question. Either

(i) find a vector $\boldsymbol{w}^* \in \mathbb{Q}^d$ such that $\|\boldsymbol{w} - \boldsymbol{w}^*\| \leqslant \delta$ and $\operatorname{dist}\big(\boldsymbol{0}, \partial f(\boldsymbol{w}^*)\big) \leqslant \varepsilon$, or

(ii) assert that any vector $\boldsymbol{w}^* \in \mathbb{R}^d$ with $\operatorname{dist}\big(\boldsymbol{0}, \partial f(\boldsymbol{w}^*)\big) \leqslant \varepsilon$ satisfies

$$\|\boldsymbol{w} - \boldsymbol{w}^*\| > \min\{\delta, \delta_{\mathsf{sep}}(\boldsymbol{w}^*)\}.$$

Our main algorithmic results in this paper are a robust testing algorithm and a "rounding" subprocedure. The following result guarantees that our new algorithms will correctly answer the RST-DSM question in oracle-polynomial time.

Theorem 8 (Polynomality) *Let vectors* $\{(\boldsymbol{x}_{p,i}, a_{p,i})\}_{p,i \in [n_p]}, \{(\boldsymbol{y}_{q,j}, b_{q,j})\}_{q,j \in [m_q]} \subseteq \mathbb{Q}^d$*, point* $\boldsymbol{w} \in \mathbb{Q}^d$*, and rational precisions* $\varepsilon, \delta \geqslant 0$ *be given. Suppose we have an implementation of the stationarity test oracle for the DSM function in Definition 6. There exists an oracle-polynomial-time algorithm that provides the correct answer to Problem RST-DSM.*

References

[1] Amir Ali Ahmadi and Jeffrey Zhang. On the complexity of finding a local minimizer of a quadratic function over a polytope. *Mathematical Programming*, 195(1-2):783–792, 2022.

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

[3] 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.

[4] 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.

[5] Katta G Murty and Santosh N Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

[6] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

[7] Panos M Pardalos and Stephen A Vavasis. Open questions in complexity theory for numerical optimization. *Mathematical Programming*, 57(1-3):337–339, 1992.

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

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

[10] Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians. *arXiv preprint arXiv:2210.06907*, 2022.

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