Parameterized Complexity of Stationarity Testing for Piecewise-Affine Functions and Shallow CNN Losses

Yuhan Ye
MIT
yyh03@mit.edu

Abstract

We study the parameterized complexity of testing approximate first-order stationarity at a prescribed point for continuous piecewise-affine (PA) functions, a basic task in nonsmooth optimization. PA functions form a canonical model for nonsmooth stationarity testing and capture the local polyhedral geometry that appears in ReLU-type training losses. Recent work of Tian and So [2025] shows that testing approximate stationarity notions for PA functions is computationally intractable in the worst case, and identifies fixed-dimensional tractability as an open direction. We address this direction from the viewpoint of parameterized complexity, with the ambient dimension d as the parameter. In this paper, we give XP algorithms in fixed dimension for the tractable sides, and prove W[1]-hardness for the complementary sides. Moreover, lower bounds under the parameterized k-Clique hypothesis rule out deterministic algorithms running in time $\rho(d)\mathsf{s}^{o(d)}$ for any computable function ρ, where s denotes the total binary encoding length of the instance. As a further consequence, our results yield the corresponding parameterized complexity picture for testing local minimality of continuous PA functions. We also extend our hardness results to a family of shallow ReLU CNN training losses, tested in the trainable weight space. Thus, the same parameterized-complexity picture also appears for simple CNN training losses. Notably, our lower bounds assume that k-Clique on N-vertex graphs admits no deterministic $\rho(k)N^{o(k)}$-time algorithm for any computable function ρ. The recent refutations of the Exponential Time Hypothesis (ETH) [OpenAI, 2026] and the stronger fine-grained k-Clique hypothesis [Vassilevska Williams and Zamir, 2026] do not contradict this assumption.

1 Introduction

For a continuously differentiable objective $f : \mathbb{R}^d \to \mathbb{R}$, a point is stationary precisely when its gradient vanishes. Testing stationarity at a prescribed point is then straightforward. For nonsmooth objectives, the situation is more subtle: several generalized subdifferentials coexist, and the resulting stationarity notions can differ substantially [Clarke, 1990, Rockafellar and Wets, 1998, Mordukhovich, 2006]. Moreover, deciding whether a prescribed point satisfies a given nonsmooth stationarity notion can itself be computationally nontrivial. This issue is intrinsic to ReLU-type training losses, where nondifferentiability is an architectural feature rather than a degenerate exception [Tian and So, 2025, Yun et al., 2019].

We study the computational complexity of testing approximate first-order stationarity at a prescribed point, which is different from the search problem of finding a stationary point. As a foundation for understanding more general piecewise differentiable objectives, we focus on continuous piecewise-affine (PA) functions. The polyhedral local geometry of PA objectives is also expressive enough to capture local directional models arising in ReLU and max-pooling training losses [Rister and Rubin, 2017, Yun et al., 2019, Tian and So, 2025]. The two stationarity notions considered throughout are

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Fréchet and Clarke stationarity.[1] For continuous PA objectives, exact Fréchet stationarity is equivalent to local minimality. Recent work of Tian and So [2025] showed that approximate PA stationarity testing is already computationally intractable in the worst-case sense, and raised fixed-dimensional tractability as a natural open direction.

Against this background, we ask how the complexity changes when the ambient dimension d is treated as a parameter. Let s denote the total binary encoding length of the PA input formula, query point, and the stationarity approximation tolerance. The most direct enumeration algorithms give running times of the form $\mathsf{s}^{O(d)}$ for explicit max–min input and $\mathsf{s}^{\Gamma(d)}$ for fixed-depth n-MC DC input[2], where the computable exponent Γ may depend on the fixed depth. In parameterized-complexity terminology, such bounds place the corresponding fixed-depth problems in XP. The central question is whether one can substantially beat this direct enumeration dependence—in particular, obtain fixed-parameter tractable (FPT) algorithms with running time $f(d)\mathsf{s}^{O(1)}$—or whether a $\mathsf{s}^{\Theta(d)}$ running time is unavoidable.

This fixed-dimension viewpoint is closely connected to recent work on ReLU training and verification. For two-layer ReLU training, fixed data dimension yields natural enumeration algorithms, but this does not remove hardness [Arora et al., 2018, Froese et al., 2022, Froese and Hertrich, 2023]. For ReLU verification, practical verification tools and complexity results have motivated a detailed study of the functions computed by ReLU networks [Katz et al., 2017, 2019, Froese et al., 2025b]. A COLT 2025 open-problem note by Froese et al. [2025a] asks whether several zonotope problems arising from ReLU verification are fixed-parameter tractable with respect to the input dimension. The follow-up work of Froese et al. [2026] proves W[1]-hardness and ETH-tight lower bounds for zonotope containment and related ReLU verification tasks. Our results show that an analogous parameterized-complexity picture arises for stationarity testing of PA functions, and we further transfer this picture to a family of shallow ReLU convolutional neural network (CNN) training losses.

1.1 Our contributions

We use the following convention for the stationarity tests of PA functions f. Fréchet-ε-YES asks whether $0 \in \partial_{\mathrm{F}} f(w) + \varepsilon\mathbb{B}$, and Clarke-$\varepsilon$-NO asks whether $\operatorname{dist}(0, \partial_{\mathrm{C}} f(w)) > \varepsilon$; the other two sides are the complementary Fréchet-ε-NO and Clarke-ε-YES tests. Our contributions are as follows.

(i) **Fixed-dimensional algorithms.** For every $\varepsilon \geq 0$, Fréchet-ε-YES and Clarke-ε-NO are in XP. For explicit max–min input, both tests are decidable in time $\mathsf{s}^{O(d)}$. For fixed-depth explicit n-MC DC input, both tests are decidable in time $\mathsf{s}^{\Gamma(d)}$, where Γ is a computable function that may depend on the fixed depth.

(ii) **Parameterized hardness and lower bounds.** For the complementary sides, d-Fréchet-ε-NO is W[1]-hard for every fixed $\varepsilon \geq 0$, and d-Clarke-ε-YES is W[1]-hard for every fixed $\varepsilon \in [0, 1/2)$. The reductions are parameter-preserving from k-Clique. Under Hypothesis 2.1, neither hard side admits a deterministic algorithm running in time $\rho(d)\mathsf{s}^{o(d)}$ for any computable function ρ.

(iii) **Consequences for local minimality and CNN losses.** Our results give the corresponding dimension-parameterized picture for testing whether a prescribed point is a local minimizer of f. We also construct a shallow ReLU CNN training loss family whose Fréchet nonstationarity and Clarke stationarity test are W[1]-hard with respect to the trainable parameter dimension. Hence, even this basic stationarity and local-minimality test can be fixed-parameter intractable for shallow CNN losses.

1.2 Related work

Stationarity testing for PA functions and ReLU network losses. The complexity of testing whether a prescribed point satisfies a local solution concept has a long history. Classical results establish hardness for local optimality in quadratic and nonlinear optimization, and later work studies local-minimum testing and related local solution concepts for polynomial and PA objectives [Murty and Kabadi, 1987, Nesterov, 2013, Ahmadi and Zhang, 2022a,b]. Tian and So [2025] developed a PA framework for stationarity testing that isolates the representation-level polyhedral core of the problem,

[1] Formal definitions are given in Section 2.3.
[2] Formal definitions are given in Section 2.1.

Table 1: Summary of the complexity landscape.

Problem	Description	Complexity	Reference
Fréchet-ε-YES	$0 \in \partial_{\mathrm{F}} f(w) + \varepsilon\mathbb{B}$	coNP-hard	[Tian and So, 2025]
Clarke-ε-YES	$0 \in \partial_{\mathrm{C}} f(w) + \varepsilon\mathbb{B}$	NP-hard	[Tian and So, 2025]
d-Fréchet-ε-YES	$0 \in \partial_{\mathrm{F}} f(w) + \varepsilon\mathbb{B}$	in XP	Theorem 3.1
d-Clarke-ε-NO	$\mathrm{dist}(0, \partial_{\mathrm{C}} f(w)) > \varepsilon$	in XP	Theorem 3.1
d-Fréchet-ε-NO	$\mathrm{dist}(0, \partial_{\mathrm{F}} f(w)) > \varepsilon$	W[1]-hard	Theorem 4.1
d-Clarke-ε-YES	$0 \in \partial_{\mathrm{C}} f(w) + \varepsilon\mathbb{B}$	W[1]-hard	Theorem 4.1
CNN training losses L: shallow ReLU CNN with max-pooling; parameter $p = \dim(\theta)$.			
Fréchet-ε-NO	$\mathrm{dist}(0, \partial_{\mathrm{F}} L(\theta_0)) > \varepsilon$	W[1]-hard in p	Theorem 5.1
Clarke-ε-YES	$0 \in \partial_{\mathrm{C}} L(\theta_0) + \varepsilon\mathbb{B}$	W[1]-hard in p	Theorem 5.1

proves coNP-hardness for approximate Fréchet stationarity and NP-hardness for approximate Clarke stationarity, and develops near-stationarity algorithms and regularity conditions. An earlier version appeared at the NeurIPS 2023 OPT workshop [Tian and So, 2023]. Their results also include applications to structured nonsmooth neural-network losses, including shallow ReLU-type networks and CNNs. Local optimality and saddle phenomena for shallow ReLU losses at nondifferentiable points have also been studied through explicit polyhedral analysis [Yun et al., 2019].

Approximate stationarity in nonsmooth optimization. Approximate stationarity in nonsmooth nonconvex optimization has been studied from several algorithmic and oracle-complexity perspectives. Representative directions include subdifferential-based stationarity notions [Li et al., 2020], algorithms for finding approximate or Goldstein-type stationary points [Zhang et al., 2020, Tian et al., 2022, Jordan et al., 2023], stochastic and online algorithms for nonsmooth DC optimization [Ye et al., 2025], descent-oriented subgradient methods for nonsmooth optimization [Li and Cui, 2025], and lower bounds or dimension-free impossibility results [Kornowski and Shamir, 2021, Tian and So, 2024]. These works mostly concern oracle or algorithmic models for finding approximate stationary points. In contrast, our task is to decide a stationarity test with dimension as the hardness parameter.

Parameterized complexity of fixed-dimensional ReLU training and verification. Fixed-dimensional ReLU training and verification provide a closely related parameterized-complexity backdrop. For two-layer ReLU training, fixed data dimension yields natural enumeration algorithms, but this does not eliminate hardness [Arora et al., 2018, Froese et al., 2022, Froese and Hertrich, 2023]. A complementary line studies the polyhedral partition structure of ReLU networks through their linear regions: classical work gives expressivity bounds for such regions, while recent work studies the complexity of counting them [Montúfar et al., 2014, Stargalla et al., 2025]. For ReLU verification, algorithmic systems such as Reluplex and Marabou helped establish neural-network verification as a central computational task [Katz et al., 2017, 2019], while recent complexity work connects ReLU positivity, surjectivity, and verification to zonotope containment [Froese et al., 2025b, Kulmburg and Althoff, 2021]. The COLT 2025 open-problem of Froese et al. [2025a] formulated the fixed-parameter tractability of several zonotope problems as a central question. Subsequent work proved W[1]-hardness and ETH-tight lower bounds for zonotope containment and related ReLU verification tasks [Froese et al., 2026]. These works concern global training or verification problems in fixed input dimension, whereas our focus is local stationarity testing at a prescribed point.

2 Preliminaries

All numerical input data are rationally encoded. We write $[m] := \{1, \dots, m\}$. All norms are Euclidean, and $\mathbb{B} := \{u \in \mathbb{R}^d : \|u\|_2 \le 1\}$ denotes the closed unit ball in the ambient space. For a set $S \subseteq \mathbb{R}^d$, $\mathrm{conv}(S)$, $\mathrm{cl}(S)$, $\mathrm{ri}(S)$, and $\mathrm{cone}(S)$ denote its convex hull, closure, relative interior, and conic hull, respectively. For a polytope P, $\mathrm{vert}(P)$ denotes its vertex set. For a nonempty compact convex set $K \subseteq \mathbb{R}^d$, its support function is $\sigma_K(u) := \max_{z \in K} \langle z, u \rangle$. For a polytope $P \subseteq \mathbb{R}^d$ and a point $v \in P$, we write $N_P(v) := \{u \in \mathbb{R}^d : \langle v, u \rangle = \sigma_P(u)\}$ for the normal cone of P at v; $\mathrm{ri}\, N_P(v)$ denotes the relative interior of this cone. We use the convention $\mathrm{dist}(0, \emptyset) = +\infty$.

For upper bounds, s denotes the total binary encoding length of the relevant PA input formula together with the query point and, when it is part of the input, the stationarity approximation tolerance $\varepsilon \geq 0$. When the dimension dependence is displayed explicitly, $\mathsf{s}^{O(d)}$ means s^{cd+c_0} for absolute constants c, c_0 independent of other parameters of the instance. For every fixed d, the exponent $cd + c_0$ is therefore a constant depending only on d, so such a bound is of the form $\mathsf{s}^{O_d(1)}$. For hardness reductions, N denotes the number of vertices in the input graph and also the total encoding size when no confusion can arise.

2.1 Piecewise-affine representations

A continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is piecewise-affine (PA) if $\mathbb{R}^d$ admits a finite polyhedral partition such that f is affine on every cell. Continuous PA functions admit several equivalent finite descriptions, including max–min and DC representations; see, for example, Ovchinnikov [2002] and Scholtes [2012]. For complexity-theoretic purposes, the representation and its encoding length are part of the input model. We use two compact input models. The first exposes max–min structure explicitly.

Definition 2.1 (Explicit max–min representation)**.** A PA function $f : \mathbb{R}^d \to \mathbb{R}$ is given in explicit max–min form if there are rational affine functions $\alpha_j(w) = \langle x_j, w\rangle + a_j$ for $j \in [k]$, with $(x_j, a_j) \in \mathbb{Q}^{d+1}$, and nonempty index sets $M_1, \ldots, M_\ell \subseteq [k]$ such that

$$f(w) = \max_{1\leq i\leq \ell} \min_{j\in M_i} \alpha_j(w).$$

For example, $\max\{\min(\alpha_1, \alpha_2), \alpha_3\}$ is an explicit max–min formula with $M_1 = \{1, 2\}$ and $M_2 = \{3\}$. The second model is a DC input model. We follow the nested max/sum syntax of Tian and So [2025] for the convex PA components.

Definition 2.2 (n-MC representation)**.** A convex PA function $h : \mathbb{R}^d \to \mathbb{R}$ is given in explicit n-MC form if there are integers $n \in \mathbb{N}$ and positive integers (I_p, J_p), $p \in [n]$, together with rational affine leaf data

$$\left(x_{i_1,j_1,\ldots,i_n,j_n}, a_{i_1,j_1,\ldots,i_n,j_n}\right) \in \mathbb{Q}^{d+1}$$

indexed by $(i_1, j_1, \ldots, i_n, j_n) \in \prod_{p=1}^{n}([I_p] \times [J_p])$ such that

$$h(w) = \sum_{1\leq j_n\leq J_n} \max_{1\leq i_n\leq I_n} \cdots \sum_{1\leq j_1\leq J_1} \max_{1\leq i_1\leq I_1} \left(\langle w, x_{i_1,j_1,\ldots,i_n,j_n}\rangle + a_{i_1,j_1,\ldots,i_n,j_n}\right).$$

For example, $h(w) = \sum_{j=1}^{m} \max\{\alpha_j(w), 0\}$ is a simple 1-MC formula. A DC representation of a PA function is an expression $f = h - g$ in which h and g are convex PA functions given in n_h-MC and n_g-MC form, respectively. The finite expression tree is given as part of the input: the ranges (I_p, J_p), the alternating max/sum structure, and all affine leaf data are listed. The depths n_h, n_g contribute to the encoding length but are not parameters; in fixed-depth families they are bounded by an input-independent constant.

V-representations. We also use the following polyhedral representation convention. A V-representation of a polytope $P \subseteq \mathbb{R}^d$ is a finite rational set $V \subseteq \mathbb{Q}^d$ such that $P = \mathrm{conv}(V)$; the set V need not be irredundant. In the algorithmic arguments below, whenever we write $\mathrm{vert}(P)$, we mean the reduced vertex list of P. Given a finite rational V-representation V, this reduction can be performed by testing, for each $v \in V$, whether $v \in \mathrm{conv}(V \setminus \{v\})$, which is a rational linear feasibility problem.

2.2 Parameterized complexity, ETH, and Clique hypotheses

We use standard parameterized-complexity terminology [Cygan et al., 2015]. A parameterized problem with parameter k is fixed-parameter tractable (FPT) if it is solvable in time $f(k)|x|^{O(1)}$ for some computable function f. It is in XP if it is solvable in time $|x|^{\Gamma(k)}$ for some computable function Γ. The class W[1] can be viewed as a parameterized analogue of NP for problems whose witnesses have size controlled by the parameter; the canonical W[1]-complete problem is k-Clique parameterized by the target clique size k. A parameterized reduction maps an instance (x, k) to an instance (x', k') in time $f(k)|x|^{O(1)}$, preserves yes/no answers, and ensures that k' is bounded by a computable function of k. Thus, an FPT algorithm for any W[1]-hard problem would imply

FPT algorithms for all problems in W[1], in particular for k-Clique; this is the standard reason such hardness is taken as evidence against fixed-parameter tractability.

For lower bounds, we use the following assumption.

Hypothesis 2.1 (Parameterized k-Clique)**.** *There is no deterministic algorithm that, given an N-vertex graph and k as input, detects a k-clique in time $\rho(k)N^{o(k)}$ for any computable function ρ.*

The Exponential Time Hypothesis (ETH) implies Hypothesis 2.1 [Chen et al., 2006]. However, the SAT algorithm that refutes ETH [OpenAI, 2026] does not give a $\rho(k)N^{o(k)}$-time algorithm for k-Clique.

Moreover, the stronger fine-grained k-Clique hypothesis [Abboud et al., 2018] asserted that the exponent $\omega k/3$ was optimal for fixed k divisible by three, where ω is the matrix-multiplication exponent. Vassilevska Williams and Zamir [2026, Theorem 1.1] refute this hypothesis with an $O(N^{(\omega-\delta)k/3})$-time algorithm for all sufficiently large fixed k and an absolute constant $\delta > 0$. The exponent remains linear in k. Whether a sublinear exponent in k is possible remains open [Vassilevska Williams and Zamir, 2026, Section 5]. In the reductions we construct in this paper, an N-vertex k-Clique instance is transformed into a stationarity-testing instance with dimension $d = \Theta(k)$ and encoding size $\mathsf{s} \le N^{O(1)}$. Hence, a target algorithm running in time $\rho(d)\mathsf{s}^{o(d)}$ would contradict Hypothesis 2.1. The CNN lower bounds use the same argument with trainable dimension $p = \Theta(k)$ and network encoding size $M \le N^{O(1)}$.

2.3 Stationarity notions

We write ∂_{C} and ∂_{F} for the Clarke and Fréchet subdifferentials, respectively. The general definitions follow the standard nonsmooth-analysis viewpoint; see, for example, [Clarke, 1990, Rockafellar and Wets, 1998] and the monograph of Cui and Pang [2021]. For a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, the Clarke directional derivative at x in direction u is

$$f^\circ(x;u) := \limsup_{\substack{x' \to x \\ t \downarrow 0}} \frac{f(x'+tu) - f(x')}{t}.$$

The Clarke subdifferential is the compact convex set

$$\partial_{\mathrm{C}} f(x) := \{g \in \mathbb{R}^d : \langle g, u\rangle \le f^\circ(x;u)\ \forall u \in \mathbb{R}^d\},$$

equivalently characterized by $f^\circ(x;u) = \max_{g \in \partial_{\mathrm{C}} f(x)} \langle g, u\rangle$. The Fréchet subdifferential is the set of first-order local lower-support slopes,

$$\partial_{\mathrm{F}} f(x) := \left\{ g \in \mathbb{R}^d : \liminf_{y \to x,\ y \neq x} \frac{f(y) - f(x) - \langle g, y - x\rangle}{\|y - x\|} \ge 0 \right\}.$$

For the PA functions considered in this paper, the one-sided directional derivative $f'(x;u) := \lim_{t \downarrow 0}(f(x+tu) - f(x))/t$ exists in every direction, and the preceding definition is equivalent to

$$\partial_{\mathrm{F}} f(x) = \{g \in \mathbb{R}^d : \langle g, u\rangle \le f'(x;u)\ \forall u \in \mathbb{R}^d\}.$$

Intuitively, Fréchet subgradients are first-order local lower supports, whereas for PA functions Clarke subgradients are convex combinations of gradients on nearby full-dimensional affine regions. For locally Lipschitz functions one always has $\partial_{\mathrm{F}} f(x) \subseteq \partial_{\mathrm{C}} f(x)$.

A locally Lipschitz function is *Clarke regular* at x if $f'(x;u)$ exists and equals $f^\circ(x;u)$ for every direction u. At regular points, the Fréchet and Clarke subdifferentials coincide in the directional description above. This includes C^1 functions and finite (weakly) convex functions; in particular, every convex PA component used below is regular, so its Fréchet, Clarke, and usual convex subdifferentials agree. If h is convex and PA, then $h'(x;u) = \sigma_{\partial_{\mathrm{C}} h(x)}(u)$.

Stationarity-testing polarities. A point x is Fréchet stationary if $0 \in \partial_{\mathrm{F}} f(x)$, and Clarke stationary if $0 \in \partial_{\mathrm{C}} f(x)$. Appendix A.1 gives a geometric guide to these notions. The ε-stationary versions replace membership by distance at most ε. Fix a PA input formula, a query point w, and a stationarity approximation tolerance $\varepsilon \ge 0$. For $\star \in \{\mathrm{F}, \mathrm{C}\}$, the $\star$-ε-YES predicate is $\operatorname{dist}(0, \partial_\star f(w)) \le \varepsilon$, equivalently $0 \in \partial_\star f(w) + \varepsilon\mathbb{B}$, while the $\star$-ε-NO predicate is $\operatorname{dist}(0, \partial_\star f(w)) > \varepsilon$.

2.4 Local PA calculus

The following local PA calculus lemmas will be used repeatedly; proofs are deferred to Appendix A.2.

Lemma 2.3 (Exact local model for PA functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and PA, and fix $w \in \mathbb{R}^d$. There exist $r > 0$ and a positively homogeneous PA function $\phi_w : \mathbb{R}^d \to \mathbb{R}$ such that*

$$f(w+u) = f(w) + \phi_w(u) \qquad \text{whenever } \|u\| \le r.$$

Consequently, $f'(w;u) = \phi_w(u)$, $\partial_{\mathrm{F}} f(w) = \partial_{\mathrm{F}}\phi_w(0)$, and $\partial_{\mathrm{C}} f(w) = \partial_{\mathrm{C}}\phi_w(0)$. Moreover, $0 \in \partial_{\mathrm{F}} f(w)$ if and only if w is a local minimizer of f.

For explicit max–min input, the local model can be written directly from the active indices. Let f be given as in Definition 2.1 and fix $w \in \mathbb{R}^d$. Set $m_i(w) := \min_{j\in M_i} \alpha_j(w)$, $I(w) := \arg\max_{i\in[\ell]} m_i(w)$, and $J_i(w) := \arg\min_{j\in M_i} \alpha_j(w)$ for $i \in I(w)$. Then the exact local model from Lemma 2.3 is

$$\phi_w(u) = f'(w;u) = \max_{i\in I(w)} \min_{j\in J_i(w)} \langle x_j, u\rangle. \tag{1}$$

Lemma 2.4 turns the Fréchet test into the translate-containment condition $s + Y \subseteq X$.

Lemma 2.4 (Characterization of Fréchet stationarity)**.** *Let $h, g : \mathbb{R}^d \to \mathbb{R}$ be convex PA functions, set $f = h - g$, fix $w \in \mathbb{R}^d$, and put $X := \partial_{\mathrm{C}} h(w)$ and $Y := \partial_{\mathrm{C}} g(w)$ for the usual convex subdifferential polytopes. Then $\partial_{\mathrm{F}}(h-g)(w) = \{s \in \mathbb{R}^d : \ s + Y \subseteq X\}$, and*

$$0 \in \partial_{\mathrm{F}}(h-g)(w) + \varepsilon\mathbb{B} \iff \exists s \in \varepsilon\mathbb{B} \text{ such that } s + Y \subseteq X \quad (\varepsilon \ge 0).$$

In particular, $0 \in \partial_{\mathrm{F}}(h-g)(w)$ if and only if $Y \subseteq X$.

Lemma 2.5 (Characterization of Clarke stationarity)**.** *Let $X, Y \subseteq \mathbb{R}^d$ be nonempty compact polytopes, set $\phi = \sigma_X - \sigma_Y$, and write $N_P(v) := \{u : \langle v, u\rangle = \sigma_P(u)\}$ for the normal cone of a polytope P at a vertex v. A pair $x \in \mathrm{vert}(X)$, $y \in \mathrm{vert}(Y)$ is simultaneously exposed if $\mathrm{ri}\, N_X(x) \cap \mathrm{ri}\, N_Y(y) \neq \emptyset$. If $S_e(X,Y)$ denotes the set of differences $x - y$ over all such pairs, then $\partial_{\mathrm{C}}\phi(0) = \mathrm{conv}(S_e(X,Y))$.*

The local objects above are the geometric interface used throughout the paper. For max–min input, the active formula at w gives the exact homogeneous model ϕ_w. For explicit n-MC DC input $f = h - g$, the natural objects are instead the convex subdifferential polytopes $X = \partial_{\mathrm{C}} h(w)$ and $Y = \partial_{\mathrm{C}} g(w)$. The fixed-dimensional algorithms either enumerate arrangement cones of the max–min local model or manipulate V-representations of the two convex subdifferential polytopes; the hardness reductions build these local objects so that containment or convex balancing encodes k-Clique.

Remark 2.6 (Why the input representation matters)**.** The explicit max–min and explicit n-MC DC models expose different local data, and neither is subsumed by the flat single-max DC baseline in Appendix A.3 through a polynomial-size transformation in general.

That appendix shows that, if both convex components are given in the flat form $h = \max_{a\in A} \ell_a$ and $g = \max_{b\in B} m_b$ with all affine pieces explicitly listed, then the corresponding stationarity tests can be solved in polynomial time. However, many compact PA constructions do not fit this flat input model without an exponential expansion. For example, the hard DC family of Tian and So [2025, Problems 3.10–3.11] uses, for $y_1, \dots, y_m \in \{-1, 0, 1\}^n$ and $r \in \mathbb{N}$, the simple PA function

$$f_F(\xi) = h_F(\xi) - g_F(\xi) = r\|\xi\|_\infty - \max\left\{ r\|\xi\|_\infty, \sum_{q=1}^{m} |\langle \xi, y_q\rangle| \right\}.$$

The term $\sum_{q=1}^m |\langle \xi, y_q\rangle|$ is compact in the n-MC syntax, since it is a sum of two-term maxima. In contrast, its flat single-max expansion is $\sum_{q=1}^m |\langle \xi, y_q\rangle| = \max_{\sigma\in\{\pm1\}^m} \left\langle \xi, \sum_{q=1}^m \sigma_q y_q \right\rangle$, which may have exponentially many affine pieces. Thus, the single-max DC case is a useful polynomial baseline for a much flatter representation, but it does not subsume the compact n-MC DC instances or the explicit max–min instances studied here. This is why we analyze the stationarity-testing complexity of both input models directly.

3 Fixed-dimension algorithms

We first show the enumeration upper bounds when the ambient dimension is fixed. This gives the XP baseline. Throughout, s denotes the total binary encoding length of the relevant input representation together with the query point and the stationarity approximation tolerance $\varepsilon \ge 0$.

Theorem 3.1. *Let d be the ambient dimension. Consider the following input models:*

(a) *explicit max–min representation;*

(b) *fixed-depth explicit n-MC DC representation $f = h - g$.*

For every $\varepsilon \geq 0$, the following problems are decidable by deterministic enumeration:

(i) *deciding whether* $\operatorname{dist}(0, \partial_{\mathrm{F}} f(w)) \leq \varepsilon$ *(Fréchet-ε-YES);*

(ii) *deciding whether w is a local minimizer of f;*[3]

(iii) *deciding whether* $\operatorname{dist}(0, \partial_{\mathrm{C}} f(w)) > \varepsilon$ *(Clarke-ε-NO).*

The running time is $\mathsf{s}^{O(d)}$ under model (a) *and $\mathsf{s}^{\Gamma(d)}$ under model* (b)*, where the computable function Γ may depend on the fixed depth. Hence the listed tests are in XP. Complementing the output gives the same upper bounds for Fréchet-ε-NO, the non-local-minimality test, and Clarke-ε-YES.*

Proof sketch. For fixed-depth explicit n-MC DC input, process the canonical alternating max/sum expression trees bottom-up and compute reduced vertex lists for the convex subdifferential polytopes $X = \partial_{\mathrm{C}} h(w)$, $Y = \partial_{\mathrm{C}} g(w)$. More explicitly, affine leaves give singleton polytopes, sum nodes give Minkowski sums, and max nodes give convex hulls of active children. For any fixed depth, both the recursive vertex output size and the enumeration time are bounded by $\mathsf{s}^{\Gamma(d)}$ for some computable depth-dependent exponent Γ. This gives an XP algorithm for vertex enumeration (Lemma B.1).

Lemma 2.4 reduces the Fréchet test to finding $\xi \in \mathbb{R}^d$ with $\|\xi\| \leq \varepsilon$ and $\xi + Y \subseteq X$. Write $\operatorname{vert}(X) = \{x_1, \ldots, x_N\}$ and $\operatorname{vert}(Y) = \{y_1, \ldots, y_m\}$. Since $X = \operatorname{conv}\{x_1, \ldots, x_N\}$ and $Y = \operatorname{conv}\{y_1, \ldots, y_m\}$, the containment $\xi + Y \subseteq X$ is equivalent to $\xi + y_i \in X$ for all $i \in [m]$. Equivalently, for each $i \in [m]$ there are coefficients $\lambda_{i,r} \geq 0$ such that $\xi + y_i = \sum_{r=1}^{N} \lambda_{i,r} x_r$ and $\sum_{r=1}^{N} \lambda_{i,r} = 1$. Thus Fréchet-$\varepsilon$-YES is decided by minimizing $\|\xi\|_2^2$ over this rational polyhedron in the variables (ξ, λ) and comparing the optimum with ε^2.

For Clarke-ε-NO, the local PA model is $\phi(u) = \sigma_X(u) - \sigma_Y(u)$. By Lemma 2.5, the essentially active slopes are exactly the differences $x - y$ over simultaneously exposed vertex pairs. After enumerating these slopes into S_e, Clarke-ε-NO is decided by minimizing $\| \sum_{a \in S_e} \lambda_a a \|_2^2$ over the simplex $\lambda_a \geq 0$, $\sum_{a \in S_e} \lambda_a = 1$, and comparing the optimum with ε^2.

For explicit max–min input, first compute the homogeneous local model in (1). Let A be the set of distinct active slopes appearing in that model. Form the hyperplanes $\langle a - a', u \rangle = 0$ for distinct $a, a' \in A$. These hyperplanes divide the direction space into open polyhedral cones. We enumerate the full-dimensional ones in time $\mathsf{s}^{O(d)}$ [Edelsbrunner et al., 1986]. On each such cone C, all comparisons among the numbers $\langle a, u \rangle$ are fixed, so every inner minimum and the outer maximum in (1) select fixed slopes and $\phi_w(u) = \langle a_C, u \rangle$ on C. Let $K_C = \overline{C}$ and write it as $K_C = \{u : B_C u \geq 0\}$. These closures cover $\mathbb{R}^d$, and continuity gives $\phi_w(u) = \langle a_C, u \rangle$ on each K_C. Thus a vector ξ belongs to $\partial_{\mathrm{F}} \phi_w(0)$ exactly when, for every enumerated cone C, the inequality $\langle \xi - a_C, u \rangle \leq 0$ holds for all $u \in K_C$. This is equivalent to the existence of multipliers $\lambda_C \geq 0$ satisfying $\xi = a_C - B_C^{\top} \lambda_C$. Minimizing $\|\xi\|_2^2$ over the resulting rational polyhedron and comparing the optimum with ε^2 gives a rational convex QP of size $\mathsf{s}^{O(d)}$. For Clarke-ε-NO, the same cone enumeration gives the essentially active slope set $S_e = \{a_C : C \text{ is a full-dimensional cone}\}$. Since $\partial_{\mathrm{C}} f(w) = \operatorname{conv}(S_e)$, Clarke-$\varepsilon$-NO is decided by minimizing the QP $\min\{\| \sum_{a \in S_e} \lambda_a a \|_2^2 : \lambda_a \geq 0, \sum_{a \in S_e} \lambda_a = 1\}$ and comparing the optimum with ε^2. All subproblems above have size $\mathsf{s}^{O(d)}$ under max–min input, or $\mathsf{s}^{\Gamma(d)}$ under fixed-depth n-MC DC input, and are polynomial-time solvable in that enumerated size. Hence they give XP algorithms. Full proofs are deferred to Appendix B. □

4 Parameterized hardness with respect to the input dimension

We next show the parameterized W[1]-hardness and conditional running-time lower bounds for our testing problems.

[3] By Lemma 2.3, part (ii) follows from part (i) with $\varepsilon = 0$.

Theorem 4.1. *Let d denote the input dimension. For each of the following input models*

(a) *explicit max–min representation;*

(b) *explicit n-MC DC representation of constant depth (indeed depth two),*

the following decision problems are W[1]-hard with respect to d:

(i) *for every fixed $\varepsilon \geq 0$, deciding whether $\operatorname{dist}(0, \partial_{\mathrm{F}} f(w)) > \varepsilon$ (d-Fréchet-ε-NO);*

(ii) *deciding whether w is not a local minimizer of f;*[4]

(iii) *for every fixed $\varepsilon \in [0, 1/2)$, deciding whether $\operatorname{dist}(0, \partial_{\mathrm{C}} f(w)) \leq \varepsilon$ (d-Clarke-ε-YES).*

Consequently, the corresponding Fréchet-ε-YES and Clarke-ε-NO polarities are co-W[1]-hard. Finally, under Hypothesis 2.1, none of the hard problems above admits a deterministic algorithm of running time $\rho(d)\mathsf{s}^{o(d)}$ for any computable function ρ. Combined with the $\mathsf{s}^{O(d)}$ max–min enumeration bound in Theorem 3.1, this rules out improving the exponent of the natural max–min enumeration approach to $o(d)$ on these W[1]-hard tests.[5]

Proof sketch. We reduce from k-Clique. Let $V(G) = [N]$, set $p_v = (v, v^2) \in \mathbb{Z}^2$, and put $Q = \operatorname{conv}\{p_v : v \in [N]\} \subseteq \mathbb{R}^2$. Each p_v is exposed by the direction $(2v, -1)$, since $\langle p_v - p_a, (2v, -1)\rangle = (a - v)^2 > 0$ for $a \neq v$. For ordered pairs, write $(u, v) \in \mathrm{Forb}$ when $u = v$ or $\{u, v\} \notin E(G)$; a k-tuple is a clique precisely when none of its pairs is forbidden. In block coordinates $z = (z_1, \ldots, z_k) \in (\mathbb{R}^2)^k$, define $g(z) = \sum_{i=1}^{k} \sigma_Q(z_i)$, $Y := \partial_{\mathrm{C}} g(0) = Q^k$. For $1 \leq i < j \leq k$ and $(u, v) \in \mathrm{Forb}$, define

$$\psi_{i,j,u,v}(z) := \langle p_u, z_i\rangle + \langle p_v, z_j\rangle + \sum_{r \in [k] \setminus \{i,j\}} \sigma_Q(z_r), \qquad h(z) := \max_{i<j,\ (u,v)\in \mathrm{Forb}} \psi_{i,j,u,v}(z),$$

and let $X := \partial_{\mathrm{C}} h(0)$. Each $\psi_{i,j,u,v}$ is the support function of a product face of Y with the i-th block fixed to p_u and the j-th block fixed to p_v. Thus $X \subseteq Y$, and a vertex $(p_{v_1}, \ldots, p_{v_k})$ of Y belongs to X exactly when some pair (v_i, v_j) is forbidden. It follows that $X = Y$ if and only if G has no k-clique. If G has a k-clique, then X is the convex hull of a strict subset of the vertices of the full-dimensional polytope Y; hence $\operatorname{vol}(X) < \operatorname{vol}(Y)$, and no translate of Y can be contained in X.

For the Fréchet side, set $f_F := h - g$. Lemma 2.4 gives $\partial_{\mathrm{F}} f_F(0) = \{s : s + Y \subseteq X\}$. Therefore no-clique instances satisfy $\partial_{\mathrm{F}} f_F(0) = \{0\}$, while clique instances satisfy $\partial_{\mathrm{F}} f_F(0) = \emptyset$. This proves W[1]-hardness of d-Fréchet-ε-NO for every fixed $\varepsilon \geq 0$. Since the gadget is continuous PA, Lemma 2.3 also turns the same dichotomy into W[1]-hardness of deciding non-local-minimality at the origin. The explicit max–min form for $f_F(z)$ follows from $\langle p_u, z_i\rangle - \sigma_Q(z_i) = \min_{a \in [N]} \langle p_u - p_a, z_i\rangle$, giving

$$f_F(z) = \max_{i<j,\ (u,v)\in \mathrm{Forb}} \min_{a,b \in [N]} \big(\langle p_u - p_a, z_i\rangle + \langle p_v - p_b, z_j\rangle\big).$$

The same construction has an explicit n-MC DC representation of constant depth, indeed depth two.

For the Clarke side, add one coordinate and define $f_C(z, t) := t/2 + \max\{f_F(z), -|t|/2\}$. If the graph has no k-clique, then $h = g$ and $f_F \equiv 0$, so $f_C(z, t) = t/2$ and $\operatorname{dist}(0, \partial_{\mathrm{C}} f_C(0, 0)) = 1/2$. If the graph has a k-clique $(v_1, \ldots, v_k)$, the exposing directions $r_i = (2v_i, -1)$ can be scaled so that $f_F(z) < -1$. By continuity, there is a neighborhood of $(z, 1)$ on which $t > 0$ and the branch $-|t|/2$ is strictly selected, so $f_C = 0$ there. Since f_C is positively homogeneous, it also has gradient zero at points arbitrarily close to the origin. Hence $0 \in \partial_{\mathrm{C}} f_C(0, 0)$. This proves W[1]-hardness of d-Clarke-ε-YES for every fixed $\varepsilon \in [0, 1/2)$. All constructions and representation expansions run in time $f(k)|(G, k)|^{O(1)}$ and produce dimension $d = 2k$ for Fréchet and $d = 2k + 1$ for Clarke; hence they are valid parameterized reductions. With polynomial encoding growth, the running-time lower bounds follow under Hypothesis 2.1. Full proofs are given in Appendix C. □

[4] By Lemma 2.3, part (ii) follows from part (i) with $\varepsilon = 0$.

[5] Swapping the YES and NO outputs of a deterministic algorithm preserves its running time. Hence the lower bound in Theorem 4.1 applies to both polarities of each stationarity test for every ε in its stated range, and the complementary polarities are co-W[1]-hard.

5 Extension to stationarity testing for ReLU CNN training losses

We now connect our PA hardness results to neural-network training by showing that the same parameterized-complexity picture holds for stationarity testing of a concrete family of shallow CNN training losses. Following the CNN embedding construction of Tian and So [2025], we consider a shallow ReLU/max-pooling CNN with one fixed input tensor and a single trainable no-bias 1×1 convolutional filter; all linear readouts and pooling operations are fixed and rationally encoded.

Once the architecture and the single training example are fixed, the loss is a scalar PA function of the filter parameter $\theta \in \mathbb{R}^p$. We write M for the encoding size of the fixed tensor, fixed rational weights, pooling structure, designated parameter point, and the stationarity approximation tolerance $\varepsilon \geq 0$ when it is part of the input. The formulas below define the two losses used in the reduction.[6]

Let $G = ([N], E)$ be a k-Clique instance, set $p_v := (v, v^2) \in \mathbb{Z}^2$ for $v \in [N]$, and write $\mathcal{N}_1(u) := \{a \in [N] : |a - u| = 1\}$. Define the local ReLU penalty $P_u(s) := \sum_{a \in \mathcal{N}_1(u)} \mathrm{ReLU}(\langle p_a - p_u, s\rangle)$ for $u \in [N]$ and $s \in \mathbb{R}^2$. This penalty has a separation property: if $u \in \arg\max_{a \in [N]} \langle p_a, s\rangle$, then $P_u(s) = 0$, while for $r_c = (2c, -1)$, $P_c(r_c) = 0$ and $P_u(r_c) \geq 1$ whenever $u \neq c$; see Appendix D.1. Let $\mathrm{Forb} := \{(u, v) \in [N]^2 : u = v \text{ or } \{u, v\} \notin E\}$, and write $z = (z_1, \ldots, z_k) \in (\mathbb{R}^2)^k$, where z_i is associated with the i-th clique position.

The Fréchet-side loss is

$$L_G^{\mathrm{F}}(z) := \max_{\substack{1 \leq i < j \leq k \\ (u,v) \in \mathrm{Forb}}} \big[-P_u(z_i) - P_v(z_j)\big]. \tag{2}$$

The Clarke-side loss uses one additional filter coordinate t and is

$$L_G^{\mathrm{C}}(z, t) := \frac{t}{2} + \max\left\{L_G^{\mathrm{F}}(z), -\frac{1}{2}\,\mathrm{ReLU}(t) - \frac{1}{2}\,\mathrm{ReLU}(-t)\right\} = \frac{t}{2} + \max\left\{L_G^{\mathrm{F}}(z), -\frac{|t|}{2}\right\}. \tag{3}$$

Theorem 5.1 (Parameterized hardness for shallow CNN losses)**.** *For the CNN family in (2)–(3), with stationarity tested at the origin of the trainable parameter space and parameterized by the trainable dimension $p = \dim(\theta)$, the following hold.*

(i) *For L_G^{F}, for every fixed $\varepsilon \geq 0$, Fréchet-ε-NO is W[1]-hard with respect to p.*

(ii) *For L_G^{F}, deciding whether the origin is not a local minimizer is W[1]-hard with respect to p.*

(iii) *For L_G^{C}, for every fixed $\varepsilon \in [0, 1/2)$, Clarke-ε-YES is W[1]-hard with respect to p.*

Moreover, under Hypothesis 2.1, none of these hard problems admits a deterministic algorithm running in time $\rho(p)M^{o(p)}$ for any computable function ρ, where M denotes the CNN encoding size.

Proof sketch. We reduce from k-Clique. For the Fréchet side, use L_G^{F} with trainable parameter $z \in (\mathbb{R}^2)^k$. If G has no k-clique, then for arbitrary z choose $v_i \in \arg\max_{a \in [N]} \langle p_a, z_i\rangle$ for each clique position i. Since $(v_1, \ldots, v_k)$ is not a clique, some $(v_i, v_j) \in \mathrm{Forb}$. For the term indexed by (i, j, v_i, v_j), the separation property (see Lemma D.1) gives $P_{v_i}(z_i) = P_{v_j}(z_j) = 0$, while every term in (2) is nonpositive. Hence $L_G^{\mathrm{F}} \equiv 0$ and $\partial_{\mathrm{F}} L_G^{\mathrm{F}}(0) = \{0\}$.

If G has a k-clique $(c_1, \ldots, c_k)$, take $z_i = T(2c_i, -1)$. For any term indexed by (i, j, u, v), the pair (u, v) is forbidden whereas (c_i, c_j) is allowed, so $u \neq c_i$ or $v \neq c_j$. The separation property gives $P_u(z_i) \geq T$ in the first case or $P_v(z_j) \geq T$ in the second, and therefore $L_G^{\mathrm{F}}(z) \leq -T < 0$. Since $L_G^{\mathrm{F}} \leq 0$ everywhere and is positively homogeneous, any $g \in \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)$ would satisfy $\langle g, d\rangle \leq L_G^{\mathrm{F}}(d)$ for all d. Taking $d = g$ gives $g = 0$, but $0 \in \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)$ would imply $L_G^{\mathrm{F}} \geq 0$, contradicting the negative direction above. Thus $\partial_{\mathrm{F}} L_G^{\mathrm{F}}(0) = \emptyset$, proving the Fréchet hard side. Since L_G^{F} is continuous PA, Lemma 2.3 also says that the no-clique and clique cases are exactly the local-minimum and non-local-minimum cases.

For the Clarke side, use L_G^{C} with trainable parameter (z, t). If G has no k-clique, then $L_G^{\mathrm{F}} \equiv 0$, so $L_G^{\mathrm{C}}(z, t) = t/2$ and $\mathrm{dist}(0, \partial_{\mathrm{C}} L_G^{\mathrm{C}}(0, 0)) = 1/2$. If G has a k-clique, the Fréchet construction gives $z^\star$ with $L_G^{\mathrm{F}}(z^\star) < -1$. Near $(z^\star, 1)$, the $-|t|/2$ alternative is strictly selected in (3), so

[6] Appendix D.2 describes the loss family and gives its exact CNN realizations.

$L_G^{\mathrm{C}}(z,t) = t/2 - t/2 = 0$. By positive homogeneity, this yields a full-dimensional cone whose closure contains the origin on which L_G^{C} is affine with gradient zero; hence the zero gradient is essentially active and $0 \in \partial_{\mathrm{C}} L_G^{\mathrm{C}}(0,0)$. This proves Clarke-$\varepsilon$-YES hardness for every fixed $\varepsilon \in [0, 1/2)$.

The CNN instances can be assembled in time $f(k)|(G,k)|^{O(1)}$, with $M = \mathrm{poly}(N,k)$ and $p = \Theta(k)$, so these are valid parameterized reductions. The running-time lower bounds follow from Hypothesis 2.1. Full details are in Appendix D.3. □

Corollary 5.2 (Complete parameterized picture)**.** *For the CNN losses in (2)–(3), parameterized by the trainable dimension p, all four stationarity polarities are decidable in time $M^{O(p)}$, hence in XP.*

The proof is deferred to Appendix D.4. Together with Theorem 5.1, this makes the $M^{O(p)}$ upper bound tight under Hypothesis 2.1 for Fréchet-ε-NO and Clarke-ε-YES in the stated ranges of ε.

Hence, the parameterized-complexity picture developed in Sections 3–4 carries over to stationarity testing for a simple family of CNN training losses. This reinforces the intractability of neural-network training from a local-verification perspective: even for shallow ReLU CNN losses, testing stationarity and local minimality can be W[1]-hard with respect to the number of trainable parameters.

6 Conclusion

This paper addresses a basic and foundational problem in nonsmooth optimization: deciding approximate first-order stationarity at a prescribed point for continuous PA functions. For the input models considered here, we resolve the fixed-dimensional tractability picture from a parameterized-complexity viewpoint. Fréchet-ε-YES and Clarke-ε-NO admit deterministic XP algorithms in the ambient dimension d, and all four tests are in XP by complementing the output. Conversely, Fréchet-ε-NO for $\varepsilon \geq 0$ and Clarke-ε-YES for $\varepsilon \in [0, 1/2)$ are W[1]-hard and admit no $\rho(d)\mathsf{s}^{o(d)}$-time deterministic algorithms under Hypothesis 2.1; for explicit max–min input, this matches the $\mathsf{s}^{O(d)}$ enumeration dependence.

Limitations. Our results are worst-case complexity results for explicitly represented PA functions and a shallow CNN family. The lower bounds do not rule out faster algorithms under additional geometric regularity, average-case assumptions, or more specialized neural architectures.

AI Disclosure

The author conceived the overall research approach. An AI assistant suggested the hard-instance construction for the k-Clique reduction in Section 4 and helped develop its proof. The assistant also helped formalize the author's proof ideas for the fixed-dimension algorithms in Section 3 and the shallow CNN extension in Section 5.

Acknowledgments and Disclosure of Funding

Yuhan Ye is supported by the MIT Schwarzman College of Computing PhD Fellowship. The author declares no competing financial interests.

References

Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams. If the current clique algorithms are optimal, so is Valiant's parser. *SIAM Journal on Computing*, 47(6):2527–2555, 2018. doi: 10.1137/16M1061771.

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

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, 2022b.

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

Jianer Chen, Xiuzhen Huang, Iyad A. Kanj, and Ge Xia. Strong computational lower bounds via parameterized complexity. *Journal of Computer and System Sciences*, 72(8):1346–1367, 2006. doi: 10.1016/j.jcss.2006.04.007.

Frank H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. SIAM, 1990.

Ying Cui and Jong-Shi Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS-SIAM Series on Optimization. SIAM, 2021.

Marek Cygan, Fedor V. Fomin, Łukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michał Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015.

Herbert Edelsbrunner, Joseph O'Rourke, and Raimund Seidel. Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2):341–363, 1986.

Vincent Froese and Christoph Hertrich. Training neural networks is NP-hard in fixed dimension. In *Advances in Neural Information Processing Systems*, volume 36, pages 44039–44049, 2023.

Vincent Froese, Christoph Hertrich, and Rolf Niedermeier. The computational complexity of ReLU network training parameterized by data dimensionality. *Journal of Artificial Intelligence Research*, 74:1775–1790, 2022.

Vincent Froese, Moritz Grillo, Christoph Hertrich, and Martin Skutella. Open problem: Fixed-parameter tractability of zonotope problems. In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 6210–6214. PMLR, 2025a.

Vincent Froese, Moritz Grillo, and Martin Skutella. Complexity of injectivity and verification of ReLU neural networks (extended abstract). In *Proceedings of the 38th Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 2188–2189. PMLR, 2025b.

Vincent Froese, Moritz Grillo, Christoph Hertrich, and Moritz Stargalla. Parameterized hardness of zonotope containment and neural network verification. In *International Conference on Learning Representations*, 2026. ICLR 2026 poster.

Komei Fukuda. From the zonotope construction to the Minkowski addition of convex polytopes. *Journal of Symbolic Computation*, 38(4):1261–1272, 2004. doi: 10.1016/j.jsc.2003.08.007.

Komei Fukuda. Polyhedral computation, 2020. Lecture notes, ETH Zurich.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2 of *Algorithms and Combinatorics*. Springer, 1988.

Michael Jordan, Guy Kornowski, Tianyi Lin, Ohad Shamir, and Manolis Zampetakis. Deterministic nonsmooth nonconvex optimization. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4570–4597. PMLR, 2023.

Volker Kaibel and Marc E. Pfetsch. Some algorithmic problems in polytope theory. In *Algebra, Geometry, and Software Systems*, pages 23–47. Springer, 2003.

Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. In *Computer Aided Verification*, volume 10426 of *Lecture Notes in Computer Science*, pages 97–117. Springer, 2017.

Guy Katz, Derek A. Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljic, David L. Dill, Mykel J. Kochenderfer, and Clark Barrett. The Marabou framework for verification and analysis of deep neural networks. In *Computer Aided Verification*, volume 11561 of *Lecture Notes in Computer Science*, pages 443–452. Springer, 2019.

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 324–334, 2021.

M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan. The polynomial solvability of convex quadratic programming. *USSR Computational Mathematics and Mathematical Physics*, 20(5):223–228, 1980.

Adrian Kulmburg and Matthias Althoff. On the co-NP-completeness of the zonotope containment problem. *European Journal of Control*, 62:84–91, 2021.

Hanyang Li and Ying Cui. Subgradient regularization: A descent-oriented subgradient method for nonsmooth optimization, 2025.

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, September 2020.

Guido F. Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Boris S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*, volume 330 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 2006.

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

Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

OpenAI. The Exponential Time Hypothesis is false. OpenAI Math Release preprint, 2026. Dated October 17, revised October 25, 2026.

Sergei Ovchinnikov. Max-min representation of piecewise linear functions. *Beiträge zur Algebra und Geometrie*, 43(1):297–302, 2002.

Blaine Rister and Daniel L. Rubin. Piecewise convexity of artificial neural networks. *Neural Networks*, 94:34–45, 2017.

R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften*. Springer, 1998.

Stefan Scholtes. *Introduction to Piecewise Differentiable Equations*. SpringerBriefs in Optimization. Springer, 2012.

Moritz Stargalla, Christoph Hertrich, and Daniel Reichman. The computational complexity of counting linear regions in ReLU neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Lai Tian and Anthony Man-Cho So. Testing approximate stationarity concepts for piecewise affine functions and extensions. NeurIPS 2023 Workshop on Optimization for Machine Learning (OPT), poster, 2023. URL https://openreview.net/forum?id=5HbXROFu3p.

Lai Tian and Anthony Man-Cho So. No dimension-free deterministic algorithm computes approximate stationarities of Lipschitzians. *Mathematical Programming*, 208(1–2):51–74, 2024.

Lai Tian and Anthony Man-Cho So. Testing approximate stationarity concepts for piecewise affine functions. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2213–2224. SIAM, 2025.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21360–21379. PMLR, 2022.

Virginia Vassilevska Williams and Or Zamir. The k-Clique hypothesis is dead, for good. Author preprint, 2026.

Christophe Weibel. Maximal f-vectors of Minkowski sums of large numbers of polytopes. *Discrete & Computational Geometry*, 47:519–537, 2012. doi: 10.1007/s00454-011-9385-1.

Yuhan Ye, Ying Cui, and Jingyi Wang. An online adaptive sampling algorithm for stochastic difference-of-convex optimization with time-varying distributions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 71959–71978. PMLR, 13–19 Jul 2025.

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

Thomas Zaslavsky. Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes. *Memoirs of the American Mathematical Society*, 1(154), 1975.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11173–11182. PMLR, 2020.

Günter M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, 1995.

A Supplementary material for Section 2

A.1 A geometric guide to Clarke and Fréchet stationarity

This subsection gives a geometric reading of the two stationarity notions used in the paper. The general definitions follow the standard nonsmooth-analysis viewpoint; see, for example, Clarke [1990], Rockafellar and Wets [1998] and the monograph of Cui and Pang [2021]. The key distinction is that Fréchet stationarity is governed by first-order local lower supports centered at the base point, whereas Clarke stationarity, for locally Lipschitz functions, is governed by the convex hull of limiting gradients from nearby differentiability points.

General definitions. For an extended-real function that is finite at x, the Fréchet, or regular, subdifferential can be written as

$$\partial_{\mathrm{F}} f(x) = \left\{ v \in \mathbb{R}^d : \liminf_{y \to x,\ y \neq x} \frac{f(y) - f(x) - \langle v, y - x \rangle}{\|y - x\|} \geq 0 \right\}.$$

Thus $v \in \partial_{\mathrm{F}} f(x)$ means that the affine function $y \mapsto f(x) + \langle v, y - x \rangle$ is a local lower support up to an $o(\|y - x\|)$ error. In the convex case this reduces to the usual global supporting-hyperplane inequality. For the continuous PA functions in this paper, the exact local fan model turns this into the directional inequality $\langle v, u \rangle \leq f'(x; u)$ for every direction u.

For a locally Lipschitz function, Rademacher's theorem gives differentiability almost everywhere. If D_f denotes the set of differentiability points of f, the Clarke subdifferential has the limiting-gradient representation

$$\partial_{\mathrm{C}} f(x) = \mathrm{cl}\,\mathrm{conv} \left\{ \lim_{r \to \infty} \nabla f(x_r) : x_r \in D_f,\ x_r \to x,\ \nabla f(x_r) \text{ converges} \right\}.$$

Equivalently, it is the compact convex set whose support function is the Clarke directional derivative $f^\circ(x; u)$. For PA functions this formula becomes especially concrete (see Lemma 2.5): if $S_e(x)$ is the set of gradients of affine pieces attained on full-dimensional cells whose closures contain x, then

$$\partial_{\mathrm{C}} f(x) = \mathrm{conv}\, S_e(x).$$

This is the precise sense in which Clarke stationarity checks whether the gradients of neighboring full-dimensional regions can be convexly balanced to zero.

If f is differentiable at x, then both notions reduce to the usual gradient: $\partial_{\mathrm{F}} f(x) = \partial_{\mathrm{C}} f(x) = \{\nabla f(x)\}$. The distinction only appears at nonsmooth points, where the exact one-point directional model and the convexified nearby-gradient model can behave differently.

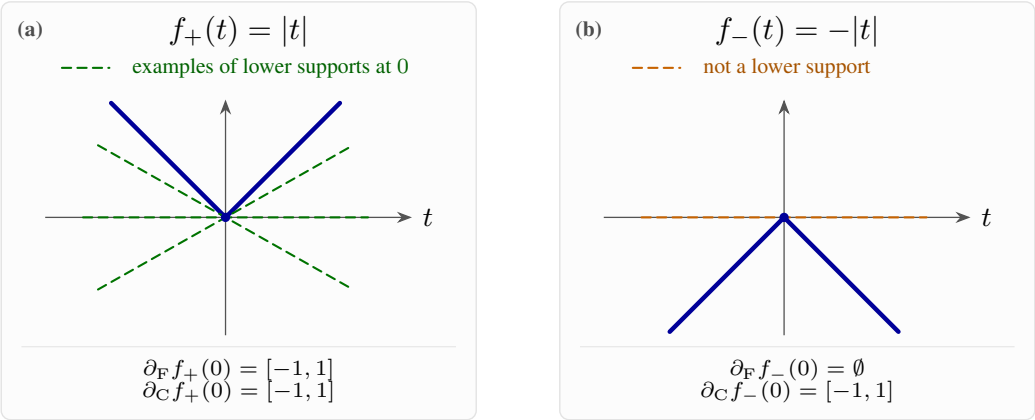

Figure 1: Fréchet stationarity is controlled by first-order local lower supports of the exact local model, whereas Clarke stationarity uses the convex hull of gradients from nearby full-dimensional differentiability regions. The functions $|t|$ and $-|t|$ have the same set of neighboring slopes, $\{-1, 1\}$, at the origin, but only $|t|$ is Fréchet stationary.

A one-dimensional example. At the origin, $f_+(t) = |t|$ has $f'_+(0;u) = |u|$, so

$$\partial_{\mathrm{F}} f_+(0) = \{g \in \mathbb{R} : gu \le |u| \ \forall u\} = [-1,1], \qquad \partial_{\mathrm{C}} f_+(0) = [-1,1].$$

Both stationarity notions hold, and the origin is a local minimizer. By contrast, $f_-(t) = -|t|$ has $f'_-(0;u) = -|u|$. The inequalities $gu \le -|u|$ would require $g \le -1$ from $u > 0$ and $g \ge 1$ from $u < 0$, so $\partial_{\mathrm{F}} f_-(0) = \emptyset$. Nevertheless the two neighboring full-dimensional gradients are still -1 and 1, hence $\partial_{\mathrm{C}} f_-(0) = \mathrm{conv}\{-1,1\} = [-1,1]$ and $0 \in \partial_{\mathrm{C}} f_-(0)$. This example is the simplest reason why Clarke stationarity may hold at a sharp maximum, while exact Fréchet stationarity for continuous PA functions is equivalent to local minimality.

A nonempty strict-inclusion example. The inclusion $\partial_{\mathrm{F}} f(x) \subseteq \partial_{\mathrm{C}} f(x)$ can be strict even when the Fréchet subdifferential is nonempty. Consider the positively homogeneous PA function

$$f(x_1, x_2) := \max\{0, \min\{x_1, x_2\}\}.$$

At the origin, the full-dimensional affine pieces have gradients 0, e_1, and e_2, where e_1, e_2 are the standard basis vectors. Hence

$$\partial_{\mathrm{C}} f(0) = \mathrm{conv}\{0, e_1, e_2\}.$$

For the Fréchet subdifferential, the directional inequality $\langle v, u\rangle \le f(u)$ for all u applies because f is positively homogeneous. Since $f(1,0) = f(-1,0) = f(0,1) = f(0,-1) = 0$, any $v = (v_1, v_2) \in \partial_{\mathrm{F}} f(0)$ must satisfy both $v_i \le 0$ and $v_i \ge 0$ for $i = 1, 2$, so $v = 0$. Conversely $0 \in \partial_{\mathrm{F}} f(0)$ because $f \ge 0$. Therefore

$$\partial_{\mathrm{F}} f(0) = \{0\} \subsetneq \mathrm{conv}\{0, e_1, e_2\} = \partial_{\mathrm{C}} f(0).$$

This example separates the two notions without relying on an empty Fréchet subdifferential.

A.2 Deferred proofs for local PA calculus

Proof of Lemma 2.3. Take a finite polyhedral partition on which f is affine. After translating the base point to the origin, each cell whose closure contains w is locally described by the inequalities of that cell that are tight at w; inequalities not tight at w remain inactive in a sufficiently small ball. Hence, for some $r > 0$, the cells incident to w induce a finite polyhedral fan in the displacement variable u on $\{u : \|u\| \le r\}$.

On each cone of this fan, $f(w+u) - f(w)$ is a linear function of u. Because f is continuous, these linear pieces agree on common faces, and therefore define a continuous positively homogeneous PA function ϕ_w satisfying

$$f(w+u) = f(w) + \phi_w(u) \qquad (\|u\| \le r).$$

Consequently,

$$f'(w;u) = \lim_{t \downarrow 0} \frac{f(w+tu) - f(w)}{t} = \lim_{t \downarrow 0} \frac{\phi_w(tu)}{t} = \phi_w(u).$$

The exact local identity also identifies the Fréchet and Clarke subdifferentials of f at w with those of ϕ_w at the origin, since both definitions depend only on the function in an arbitrarily small neighborhood of the base point.

Finally, $0 \in \partial_{\mathrm{F}} f(w)$ is equivalent to $f'(w;u) = \phi_w(u) \ge 0$ for every direction u. By the exact local identity, this is equivalent to $f(w+u) \ge f(w)$ for all sufficiently small u, namely to w being a local minimizer of f. □

Proof of Lemma 2.4. For convex polyhedral functions, directional derivatives are support functions of their subdifferentials: $h'(w;u) = \sigma_X(u) := \max_{x \in X} \langle x, u\rangle$ and $g'(w;u) = \sigma_Y(u) := \max_{y \in Y} \langle y, u\rangle$. Thus $(h-g)'(w;u) = \sigma_X(u) - \sigma_Y(u)$, and

$$s \in \partial_{\mathrm{F}}(h-g)(w) \iff \langle s, u\rangle + \sigma_Y(u) \le \sigma_X(u) \quad \forall u \in \mathbb{R}^d.$$

Because $\sigma_{s+Y}(u) = \langle s, u\rangle + \sigma_Y(u)$, this is support-function domination $\sigma_{s+Y} \le \sigma_X$. For compact convex sets, support-function domination is equivalent to containment, so $s + Y \subseteq X$. The approximate statement is obtained by requiring some $s \in \partial_{\mathrm{F}}(h-g)(w)$ with $\|s\| \le \varepsilon$, equivalently $s \in \varepsilon\mathbb{B}$. Setting $\varepsilon = 0$ gives the exact stationarity condition $Y \subseteq X$. □

Proof of Lemma 2.5. Let $S_e(X, Y)$ be the set in the statement. For a vertex v of a compact polytope P, the relative interior of its normal cone is

$$\operatorname{ri} N_P(v) = \{u : \langle v, u\rangle > \langle v', u\rangle \text{ for every } v' \in \operatorname{vert}(P) \setminus \{v\}\}.$$

Thus $x \in \operatorname{vert}(X)$ and $y \in \operatorname{vert}(Y)$ are simultaneously exposed exactly when there is a direction u satisfying

$$\langle x - x', u\rangle > 0 \quad (x' \in \operatorname{vert}(X) \setminus \{x\}), \qquad \langle y - y', u\rangle > 0 \quad (y' \in \operatorname{vert}(Y) \setminus \{y\}).$$

On such a cone, both support functions are differentiable and

$$\phi(u) = \sigma_X(u) - \sigma_Y(u) = \langle x - y, u\rangle,$$

so $x - y$ is a slope attained on a neighboring full-dimensional differentiability region of ϕ.

Conversely, consider any full-dimensional cone on which ϕ is affine. Refining this cone by the common refinement of the normal fans of X and Y, we obtain a full-dimensional subcone on which both σ_X and σ_Y are affine. An interior direction of this subcone uniquely exposes some vertices $x \in \operatorname{vert}(X)$ and $y \in \operatorname{vert}(Y)$, and the slope of ϕ on the subcone is $x - y$. Hence $S_e(X, Y)$ is exactly the set of slopes attained on neighboring full-dimensional differentiability regions of the PA function ϕ. The Clarke subdifferential of a PA function is the convex hull of these essentially active slopes, so

$$\partial_{\mathrm{C}}\phi(0) = \operatorname{conv}(S_e(X, Y)).$$

□

A.3 A polynomial baseline for single-max DC input

Lemma A.1 (Polynomial tests for single-max DC input)**.** *Let*

$$h(w) = \max_{i \in I}\{\langle a_i, w\rangle + \alpha_i\}, \qquad g(w) = \max_{j \in J}\{\langle b_j, w\rangle + \beta_j\},$$

with rational data, and let $w \in \mathbb{Q}^d$ and $\varepsilon \geq 0$ be given. Then both tests $0 \in \partial_{\mathrm{F}}(h - g)(w) + \varepsilon\mathbb{B}$ and $0 \in \partial_{\mathrm{C}}(h - g)(w) + \varepsilon\mathbb{B}$ can be decided in polynomial time.

Proof. Let

$$A := \{a_i : \ i \in I, \ \langle a_i, w\rangle + \alpha_i = h(w)\}, \qquad B := \{b_j : \ j \in J, \ \langle b_j, w\rangle + \beta_j = g(w)\}$$

be the active slope sets at w, after removing duplicates. Then $\partial_{\mathrm{C}}h(w) = \operatorname{conv}(A)$ and $\partial_{\mathrm{C}}g(w) = \operatorname{conv}(B)$. For the Fréchet test, Lemma 2.4 gives $0 \in \partial_{\mathrm{F}}(h - g)(w) + \varepsilon\mathbb{B}$ if and only if there exists s with $\|s\| \leq \varepsilon$ and $s + \operatorname{conv}(B) \subseteq \operatorname{conv}(A)$. Since $\operatorname{conv}(B)$ is the convex hull of the explicitly listed points in B, the containment condition is equivalent to $s + b \in \operatorname{conv}(A)$ for every $b \in B$. Thus the test is the feasibility of the polynomial-size system

$$s + b = \sum_{a \in A} \lambda_{b,a} a, \qquad \sum_{a \in A} \lambda_{b,a} = 1, \qquad \lambda_{b,a} \geq 0 \quad (b \in B), \qquad \|s\| \leq \varepsilon.$$

which is a rational convex QP; for $\varepsilon = 0$ it reduces to a linear feasibility problem.

For the Clarke test, the local directional model is $\phi(u) = \max_{a \in A}\langle a, u\rangle - \max_{b \in B}\langle b, u\rangle$. A pair $(a, b) \in A \times B$ contributes the slope $a - b$ on a full-dimensional linearity cone precisely when the strict system

$$\langle a - a', u\rangle > 0 \quad (a' \in A \setminus \{a\}), \qquad \langle b - b', u\rangle > 0 \quad (b' \in B \setminus \{b\})$$

is feasible. By homogeneity, this can be checked by the linear feasibility system with all right-hand sides replaced by 1. Solving this LP for every pair (a, b) gives the essentially active slope set S_e in polynomial time, and $\partial_{\mathrm{C}}(h - g)(w) = \operatorname{conv}(S_e)$. Finally, $0 \in \partial_{\mathrm{C}}(h - g)(w) + \varepsilon\mathbb{B}$ if and only if the minimum of $\|\sum_{s \in S_e} \lambda_s s\|$ over $\lambda \geq 0$ and $\sum_{s \in S_e} \lambda_s = 1$ is at most ε, which is a convex quadratic program of polynomial size. □

Local geometry and tractable structure. Explicit local geometry also yields tractable cases: if the full-dimensional cones and gradients of the local PA model at the query point are explicitly listed, the two tests reduce to polynomial-size linear or convex-quadratic programs. For DC input $f = h - g$, set $X = \partial_C h(w)$ and $Y = \partial_C g(w)$. Compatibility of X, Y characterizes when $\partial_C(h-g)(w) = X - Y$; the transversality condition $\mathrm{par}(X) \cap \mathrm{par}(Y) = \{0\}$ is a polynomial-time verifiable sufficient condition in the explicit n-MC model, where $\mathrm{par}(P) = \mathrm{span}(P - P)$ [Tian and So, 2025, Theorem 4.11 and Propositions 4.16, 4.18]. When this equality holds and X, Y are explicitly available, Clarke testing reduces to convex quadratic programming over $X - Y$.

B Deferred proofs for Theorem 3.1

To prove Theorem 3.1, we first establish a vertex-enumeration bound for the convex subdifferential polytopes arising from fixed-depth n-MC DC input. At a max node, the subdifferential is the convex hull of the union of the subdifferentials of the children attaining the maximum at the query point; at a sum node, it is the Minkowski sum of all child subdifferentials. Our main task is therefore to enumerate the vertices produced by these two operations and bound both the output size and running time by a power of the input length whose exponent is computable.

Background. Candidate vertices of a convex hull can be pruned by linear programming. For Minkowski sums, vertex enumeration is available [Fukuda, 2004, Theorem 3.3]; see also Weibel [2012] for fixed-dimensional vertex bounds in terms of the total number of input vertices. The proof below gives the bound needed here by enumerating a hyperplane arrangement. An arrangement of N hyperplanes in $\mathbb{R}^d$ has $N^{O(d)}$ cells, and its cell complex can be constructed in $N^{O(d)}$ time for fixed d [Zaslavsky, 1975, Edelsbrunner et al., 1986]. For background on normal cones, exposed faces, and polyhedral computation, see Ziegler [1995], Kaibel and Pfetsch [2003], Fukuda [2020].

The following lemma makes this precise.

Lemma B.1 (Vertex enumeration for n-MC subdifferentials)**.** *Fix an ambient dimension d and an integer depth bound D. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a convex PA expression, represented by a rational expression tree of depth at most D, whose leaves are affine functions and whose internal nodes are finite sum nodes and finite max nodes. Let s denote the total binary encoding length of the expression and of the query point w. For every node ν, let F_ν be the subexpression rooted at ν and put*

$$P_\nu := \partial_C F_\nu(w).$$

Then there is a computable function $\Gamma_D(d)$, depending only on d and the fixed depth bound D, such that for every node ν the vertex set $\mathrm{vert}(P_\nu)$ has cardinality at most $\mathsf{s}^{\Gamma_D(d)}$, all listed vertices have rational bit length at most $\mathsf{s}^{\Gamma_D(d)}$, and the vertex sets $\mathrm{vert}(P_\nu)$ for all nodes can be computed in deterministic time $\mathsf{s}^{\Gamma_D(d)}$.

Proof. Let the expression be represented by a rooted tree. For a node ν, write $\mathrm{child}(\nu)$ for the set of its children, and define the height of ν as the maximum length of a downward path from ν to a leaf. Leaves have height zero. We prove the claim by induction on node height. We slightly strengthen the induction statement: for every node of height at most t, we compute both the exact value $F_\nu(w)$ and the vertex list $\mathrm{vert}(P_\nu)$, where $P_\nu = \partial_C F_\nu(w)$, with cardinalities and bit lengths bounded by $\mathsf{s}^{\Gamma_t(d)}$.

We first record the fixed-dimensional vertex-enumeration fact used in the induction. Let $P_1, \dots, P_R \subseteq \mathbb{R}^d$ be rational polytopes given by reduced vertex lists, and let N be the total number of input vertices. For each fixed d, there is a constant c_d such that the vertices of

$$\mathrm{conv}\Big(\bigcup_{r=1}^{R} P_r\Big) \qquad \text{and} \qquad \sum_{r=1}^{R} P_r$$

can be computed in time N^{c_d} times a polynomial in the input bit length.

For the convex hull of a union, every extreme point of $\mathrm{conv}(\cup_r P_r)$ is already a vertex of at least one input polytope. Thus the union of the input vertex lists is a complete candidate list. To reduce it, test each candidate v for membership in the convex hull of the other candidates; this is a rational linear feasibility problem, and exactly the candidates that fail the test are the vertices of the convex hull.

For the Minkowski sum $P := P_1 + \cdots + P_R$, we use exposed faces. For a polytope P and a direction u, define

$$F_P(u) := \arg\max_{x \in P} \langle x, u \rangle.$$

This is the face on which the linear objective $x \mapsto \langle x, u \rangle$ is maximized. The directions exposing the same face form a normal cone, and these normal cones partition direction space. We only need the following elementary consequence: for every direction u,

$$F_{P_1 + \cdots + P_R}(u) = F_{P_1}(u) + \cdots + F_{P_R}(u),$$

because maximizing a linear function over a Minkowski sum separates over the summands. Hence, if $F_{P_r}(u) = \{v_r\}$ for every r, then $F_P(u) = \{v_1 + \cdots + v_R\}$, so $v_1 + \cdots + v_R$ is a vertex of P.

To find all such generic directions, form the hyperplanes

$$\langle v - v', u \rangle = 0$$

for all pairs of distinct vertices v, v' from the same summand P_r. These hyperplanes divide the direction space into open polyhedral cones. On any one of these cones, the ordering of the values $\langle v, u \rangle$ is fixed inside each summand, so each P_r has the same unique maximizing vertex throughout the cone. The arrangement has $O(N^2)$ hyperplanes and therefore $N^{O(d)}$ full-dimensional cones in fixed dimension; by the fixed-dimensional arrangement construction, these cones can be enumerated in $N^{O(d)}$ time [Zaslavsky, 1975, Edelsbrunner et al., 1986].

For each cone, choose a rational representative direction u, compute the unique maximizers $v_r(u) \in \mathrm{vert}(P_r)$, and add $v_1(u) + \cdots + v_R(u)$ to the candidate list. Remove duplicate candidates. Then test each remaining candidate q for membership in the convex hull of the other candidates; this is a rational linear feasibility problem, and the candidates that fail this test are exactly the vertices. The candidate list is complete: if q is a vertex of P, choose u in the interior of its normal cone, so $F_P(u) = \{q\}$. Since finitely many hyperplanes have empty interior, we may choose such a u outside all hyperplanes above. Then each summand has a unique maximizer $v_r(u)$, and

$$q = v_1(u) + \cdots + v_R(u),$$

so q appears in the list. Since the candidates are sums of input vertices and the filtering steps are rational LP feasibility tests, the running time is $N^{O(d)}$ times a polynomial in the input bit length, and all intermediate bit lengths stay polynomially bounded.

For a leaf $F_\nu(x) = \langle a, x \rangle + b$, we have $P_\nu = \{a\}$, and both $F_\nu(w)$ and $\mathrm{vert}(P_\nu)$ are computed directly. Thus the claim holds for height zero.

Now assume the claim holds for all nodes of height at most $t-1$, and let ν be a node of height t. Since the expression is explicitly listed, every node has at most s children, and the total number of nodes is at most s.

If ν is a sum node, say

$$F_\nu = \sum_{\mu \in \mathrm{child}(\nu)} F_\mu,$$

then $F_\nu(w) = \sum_{\mu \in \mathrm{child}(\nu)} F_\mu(w)$, and convex subdifferential calculus gives

$$P_\nu = \sum_{\mu \in \mathrm{child}(\nu)} P_\mu.$$

By the induction hypothesis, the total number of child vertices is at most $\mathsf{s} \cdot \mathsf{s}^{\Gamma_{t-1}(d)}$. Applying the fixed-dimensional Minkowski-sum enumeration above gives

$$|\,\mathrm{vert}(P_\nu)| \le \left(\mathsf{s}^{\Gamma_{t-1}(d)+1}\right)^{c_d} \le \mathsf{s}^{\Gamma_t(d)}$$

after choosing $\Gamma_t(d)$ large enough. The bit length of $F_\nu(w)$ and of the listed vertices satisfies the same type of bound, after increasing $\Gamma_t(d)$ if necessary.

If ν is a max node, say

$$F_\nu = \max_{\mu \in \mathrm{child}(\nu)} F_\mu,$$

we first compute all child values $F_\mu(w)$ exactly and determine the active set

$$\mathrm{Act}(\nu, w) := \{\mu \in \mathrm{child}(\nu) : F_\mu(w) = F_\nu(w)\}.$$

The exact comparisons have bit length bounded by the induction hypothesis. Convex subdifferential calculus for a pointwise maximum gives

$$P_\nu = \mathrm{conv}\Big(\bigcup_{\mu \in \mathrm{Act}(\nu,w)} P_\mu \Big).$$

The candidate vertices are contained in the union of the active child vertex lists, whose total size is at most $\mathsf{s}^{\Gamma_{t-1}(d)+1}$. Removing non-extreme candidates by rational linear feasibility tests gives $\mathrm{vert}(P_\nu)$ within the same bound, again after increasing $\Gamma_t(d)$ if necessary.

Thus, for every height t, there is a computable exponent $\Gamma_t(d)$ controlling the vertex-list size, bit length, and computation time for all nodes of height at most t. Since the depth bound D is fixed independently of the input, taking $t = D$ gives a computable exponent $\Gamma_D(d)$. The total work over all nodes is absorbed into $\mathsf{s}^{\Gamma_D(d)}$. □

Remark B.2. For explicit n-MC DC input, the lifted LP descriptions of Tian and So [2025, Appendix A] give polynomial-size projected representations of the subdifferentials. This is used to obtain polynomial-time subroutines for subgradient membership, the sum-rule relaxation condition $0 \in \partial h(w) - \partial g(w) + \varepsilon B$, and transversality [Tian and So, 2025, Proposition A.1(a)–(c)]. These subroutines also support the (co)NP membership proofs for exact stationarity testing [Tian and So, 2025, Theorem 3.2]

Lemma B.1, by contrast, uses a different tool tailored to our fixed-dimensional parameterized setting. Under a fixed depth bound D, the same subdifferential polytopes are converted into explicit reduced vertex lists by an efficient vertex-enumeration algorithm routine [Edelsbrunner et al., 1986, Ziegler, 1995], within $\mathsf{s}^{\Gamma_D(d)}$ time. This use of vertex enumeration is the new algorithmic ingredient here.

We also use the classical polynomial-time solvability of rational convex quadratic programming in the Turing model [Kozlov et al., 1980]; see also Grötschel et al. [1988] for the rational-data ellipsoid framework. We record only the projection forms needed below.

Lemma B.3 (Exact polynomial-time comparison for rational quadratic subproblems)**.** *Let $Q \succeq 0$, q, r, A, b, and τ be rational input data. The threshold decision problem*

$$\inf\{z^\top Q z + 2q^\top z + r : \ Az \le b\} \le \tau$$

can be decided in polynomial time in the binary encoding length of the displayed rational data. The same algorithmic framework detects infeasibility and unboundedness below.

In particular, the following two projection comparisons are exact polynomial-time decisions in their encoded size:

$$\min_{z \in P} \|Lz + c\|_2^2 \le \tau \quad \textit{for a rational polyhedron } P,$$

and

$$\min\left\{ \left\| \sum_{a \in S} \lambda_a a \right\|_2^2 : \lambda_a \ge 0,\ \sum_{a \in S} \lambda_a = 1 \right\} \le \tau \quad \textit{for a finite nonempty rational set } S \subseteq \mathbb{R}^d.$$

Here all matrices, vectors, points, and the threshold τ are rationally encoded.

Proof. The first displayed problem is a convex quadratic program with rational data. By the polynomial-time solvability theorem for rational convex quadratic programming of Kozlov et al. [1980], one can decide the rational threshold comparison in the Turing model in time polynomial in the binary input length; infeasibility and unboundedness below are detected within the same bound.

The two projection problems are special cases. In the first one,

$$\|Lz + c\|_2^2 = z^\top L^\top L z + 2(L^\top c)^\top z + \|c\|_2^2,$$

so the objective has a rational positive semidefinite quadratic matrix $L^\top L$, and the constraints defining P are rational linear inequalities. In the second one, the variables are the simplex coefficients λ, the feasible region is the rational simplex, and the objective matrix is the rational Gram matrix

$$\big(\langle a, a' \rangle\big)_{a,a' \in S}.$$

Thus both are rational convex quadratic programs, and their comparisons with the rational threshold τ are exact polynomial-time decisions in the encoded input size. □

Proof of Theorem 3.1: fixed-depth explicit n-MC DC input, Fréchet-ε-YES. Let the fixed-depth explicit n-MC formulas for h and g be viewed as their canonical alternating max/sum expression trees. Apply Lemma B.1 to both trees. This gives reduced vertex lists for

$$X := \partial_{\mathrm{C}} h(w), \qquad Y := \partial_{\mathrm{C}} g(w)$$

in time $\mathsf{s}^{\Gamma(d)}$, where Γ depends only on the fixed depth and on d.

By Lemma 2.4,

$$0 \in \partial_{\mathrm{F}}(h-g)(w) + \varepsilon\mathbb{B} \iff \exists \xi \in \mathbb{R}^d \text{ such that } \|\xi\| \le \varepsilon \text{ and } \xi + Y \subseteq X.$$

Write

$$\mathrm{vert}(X) = \{x_1, \dots, x_N\}, \qquad \mathrm{vert}(Y) = \{y_1, \dots, y_m\}.$$

Since $Y = \mathrm{conv}\{y_1, \dots, y_m\}$ and $X = \mathrm{conv}\{x_1, \dots, x_N\}$, the containment $\xi + Y \subseteq X$ is equivalent to requiring $\xi + y_i \in X$ for every $i \in [m]$. Equivalently, for every $i \in [m]$, there exist coefficients $\lambda_{i,1}, \dots, \lambda_{i,N}$ such that

$$\xi + y_i = \sum_{r=1}^{N} \lambda_{i,r} x_r, \qquad \sum_{r=1}^{N} \lambda_{i,r} = 1, \qquad \lambda_{i,r} \ge 0 \quad (r = 1, \dots, N).$$

Thus the Fréchet-ε-YES test is the exact comparison

$$\min_{\xi,\lambda} \left\{ \|\xi\|_2^2 : \xi + y_i = \sum_{r=1}^{N} \lambda_{i,r} x_r, \sum_{r=1}^{N} \lambda_{i,r} = 1, \lambda_{i,r} \ge 0 \quad (i = 1, \dots, m) \right\} \le \varepsilon^2.$$

The number of variables, constraints, and coefficient bit lengths in the displayed rational convex QP are bounded by $\mathsf{s}^{\Gamma(d)}$ after increasing Γ if necessary. By Lemma B.3, this comparison is decided exactly in time polynomial in the displayed QP size, hence in time $\mathsf{s}^{\Gamma(d)}$. For local minimality, set $\varepsilon = 0$ in the same Fréchet computation and use Lemma 2.3, which identifies $0 \in \partial_{\mathrm{F}} f(w)$ with w being a local minimizer for continuous PA functions. □

Proof of Theorem 3.1: fixed-depth explicit n-MC DC input, Clarke-ε-NO. First compute reduced vertex lists for

$$X := \partial_{\mathrm{C}} h(w), \qquad Y := \partial_{\mathrm{C}} g(w)$$

using Lemma B.1. For convex PA functions, the local directional model of $f = h - g$ at w is

$$\phi(u) := \sigma_X(u) - \sigma_Y(u),$$

and Lemma 2.3 gives $\partial_{\mathrm{C}} f(w) = \partial_{\mathrm{C}} \phi(0)$.

Let $V_X := \mathrm{vert}(X)$ and $V_Y := \mathrm{vert}(Y)$. For $x \in V_X$ and $y \in V_Y$, the pair (x, y) is simultaneously exposed precisely when the strict system

$$\langle x - x', u \rangle > 0 \qquad (x' \in V_X \setminus \{x\}),$$
$$\langle y - y', u \rangle > 0 \qquad (y' \in V_Y \setminus \{y\})$$

is feasible. If the system has at least one inequality, homogeneity makes this equivalent to feasibility of the rational linear system obtained by replacing every strict right-hand side by 1. If there are no inequalities, the pair is feasible by convention. This formulation also covers lower-dimensional polytopes and singleton polytopes: the strict vertex inequalities characterize the relative interiors of the corresponding normal cones, and singleton normal cones contribute no strict inequalities.

Define

$$S_e := \{x - y : \ x \in V_X, \ y \in V_Y, \operatorname{ri} N_X(x) \cap \operatorname{ri} N_Y(y) \neq \emptyset\}.$$

The number of vertex pairs and the size of every linear feasibility system are bounded by $\mathsf{s}^{\Gamma(d)}$. By Lemma 2.5,

$$\partial_{\mathrm{C}} f(w) = \partial_{\mathrm{C}} \phi(0) = \operatorname{conv}(S_e).$$

Therefore Clarke-ε-NO is decided by the exact comparison

$$\min \left\{ \left\| \sum_{a \in S_e} \lambda_a a \right\|_2^2 : \lambda_a \geq 0, \ \sum_{a \in S_e} \lambda_a = 1 \right\} > \varepsilon^2.$$

By Lemma B.3, this rational convex-QP comparison is exact and polynomial in the displayed QP size. The overall running time is $\mathsf{s}^{\Gamma(d)}$. □

Proof of Theorem 3.1: explicit max–min input, Fréchet-ε-YES. Let

$$\phi(u) = \max_{i \in I(w)} \min_{j \in J_i(w)} \langle x_j, u \rangle$$

be the exact local model from (1). By Lemma 2.3, $\partial_{\mathrm{F}} f(w) = \partial_{\mathrm{F}} \phi(0)$. Let A be the set of distinct slopes appearing in this local model; clearly $|A| \leq \mathsf{s}$.

Build the central hyperplane arrangement

$$\mathcal{H} := \{\langle a - a', u \rangle = 0 : \ a, a' \in A, \ a \neq a'\}.$$

If $\mathcal{H}$ is empty, it has the single full-dimensional cone $\mathbb{R}^d$. Otherwise, it has $O(|A|^2)$ hyperplanes and hence $\mathsf{s}^{O(d)}$ full-dimensional open cones in fixed dimension. With $H := |\mathcal{H}|$, their number is at most $\sum_{j=0}^{d} \binom{H}{j} \leq \mathsf{s}^{O(d)}$. These cones can be enumerated in time $\mathsf{s}^{O(d)}$ [Edelsbrunner et al., 1986].

On each full-dimensional open cone C, the ordering of all quantities $\langle a, u \rangle$, $a \in A$, is fixed and strict. Consequently, every inner minimum and the outer maximum select fixed slopes. Thus there is a slope $a_C \in A$ such that

$$\phi(u) = \langle a_C, u \rangle \qquad (u \in C).$$

Let

$$K_C := \overline{C}.$$

From the arrangement description, orient the inequalities so that

$$K_C = \{u : B_C u \geq 0\},$$

where B_C has $\mathsf{s}^{O(1)}$ rows.

By the directional-derivative definition of the Fréchet subdifferential and the positive homogeneity of ϕ, for a fixed ξ we have

$$\xi \in \partial_{\mathrm{F}} \phi(0) \quad \Longleftrightarrow \quad \langle \xi, u \rangle \leq \phi(u) \quad \text{for every } u \in \mathbb{R}^d.$$

The closures K_C of the full-dimensional arrangement cones cover $\mathbb{R}^d$, so this condition is equivalent to imposing the same inequality on every K_C. Moreover, ϕ equals $\langle a_C, u \rangle$ on the open cone C, and by continuity the same equality extends to its closure K_C. Hence

$$\partial_{\mathrm{F}} \phi(0) = \bigcap_C \{\xi : \langle \xi, u \rangle \leq \phi(u) \ \forall u \in K_C\}.$$

On K_C, this condition is

$$\langle \xi - a_C, u \rangle \leq 0 \qquad \forall u \in K_C,$$

or equivalently

$$\xi - a_C \in K_C^\circ.$$

Because $K_C = \{u : B_C u \geq 0\}$, its polar cone is

$$K_C^\circ = \{-B_C^\top \lambda_C : \lambda_C \geq 0\}.$$

Therefore $\xi \in \partial_{\mathrm{F}}\phi(0)$ if and only if, for every full-dimensional arrangement cone C, there exists $\lambda_C \geq 0$ such that

$$\xi = a_C - B_C^\top \lambda_C.$$

Thus $0 \in \partial_{\mathrm{F}} f(w) + \varepsilon\mathbb{B}$ is equivalent to feasibility of

$$\exists \xi, \{\lambda_C\}_C : \qquad \xi = a_C - B_C^\top \lambda_C \quad \forall C, \qquad \lambda_C \geq 0 \quad \forall C, \qquad \|\xi\|_2 \leq \varepsilon.$$

Equivalently, one minimizes $\|\xi\|_2^2$ over the displayed rational polyhedron in the variables ξ and $\{\lambda_C\}_C$ and compares the optimum with ε^2. The number of variables, constraints, and the bit length of all rational coefficients are bounded by $\mathsf{s}^{O(d)}$. Lemma B.3 gives an exact comparison in time polynomial in this displayed size, and hence in time $\mathsf{s}^{O(d)}$. For local minimality, set $\varepsilon = 0$ and apply Lemma 2.3; the resulting exact Fréchet test decides whether w is a local minimizer. □

Proof of Theorem 3.1: explicit max–min input, Clarke-ε-NO. Compute the local model ϕ as in the preceding proof. By Lemma 2.3,

$$\partial_{\mathrm{C}} f(w) = \partial_{\mathrm{C}}\phi(0).$$

Let A be the set of distinct slopes appearing in ϕ, and construct the central arrangement

$$\mathcal{H} = \{\langle a - a', u\rangle = 0 : \ a, a' \in A, \ a \neq a'\}.$$

If A is a singleton, the arrangement has the single full-dimensional cone $\mathbb{R}^d$. Otherwise, $\mathcal{H}$ has $\mathsf{s}^{O(d)}$ full-dimensional cones, enumerable within the same time bound.

On every full-dimensional open cone C, all slopes in A are strictly ordered. Thus each inner minimum selects a fixed slope and the outer maximum among these selected slopes also selects a fixed slope. Consequently, there is a slope $a_C \in A$ such that

$$\phi(u) = \langle a_C, u\rangle \qquad (u \in C).$$

Define

$$S_e := \{a_C : \ C \text{ is a full-dimensional cone of } \mathcal{H}\},$$

with duplicates removed. Every a_C is attained as the gradient of ϕ on the open full-dimensional cone C, so it is essentially active. Conversely, if a slope is essentially active for ϕ at the origin, then it is attained on some full-dimensional differentiability region of the local PA fan. The arrangement $\mathcal{H}$ refines that fan, so one of its full-dimensional cones lies inside this region and contributes the same slope. Hence S_e is exactly the essentially active slope set.

For a Lipschitz PA function, the Clarke subdifferential at a point is the convex hull of the limiting gradients from neighboring full-dimensional differentiability regions. Since ϕ is positively homogeneous, all full-dimensional cones of its local fan are adjacent to the origin. Therefore

$$\partial_{\mathrm{C}} f(w) = \partial_{\mathrm{C}}\phi(0) = \mathrm{conv}(S_e).$$

The comparison $\mathrm{dist}(0, \partial_{\mathrm{C}} f(w)) \leq \varepsilon$ is exactly

$$\min\left\{ \left\| \sum_{a \in S_e} \lambda_a a \right\|_2^2 : \lambda_a \geq 0, \ \sum_{a \in S_e} \lambda_a = 1 \right\} \leq \varepsilon^2.$$

The set S_e has size $\mathsf{s}^{O(d)}$, and Lemma B.3 decides the displayed comparison exactly in time polynomial in its encoded size. Clarke-ε-NO is obtained by checking whether the optimum is greater than ε^2. The total running time is $\mathsf{s}^{O(d)}$. □

C Deferred proofs for Theorem 4.1

This appendix gives the full reduction underlying Theorem 4.1. We first build, from a k-Clique instance, a pair of polytopes $X \subseteq Y$ such that the translate-containment condition $s + Y \subseteq X$ distinguishes clique from no-clique instances. Lemma 2.4 then turns this containment gadget into the Fréchet stationarity test.

Clique construction. Throughout this appendix, let $G = ([N], E)$ be an undirected graph with $N \geq 3$ and let $k \geq 2$. Put

$$p_v := (v, v^2) \in \mathbb{Z}^2, \qquad Q := \mathrm{conv}\{p_v : v \in [N]\} \subseteq \mathbb{R}^2,$$

and define the forbidden ordered pairs

$$\mathrm{Forb} := \{(u, v) \in [N]^2 : u = v \text{ or } \{u, v\} \notin E\}.$$

Thus a tuple $(v_1, \ldots, v_k)$ is a k-clique exactly when no pair (v_i, v_j), $i < j$, is in Forb. Let

$$\mathcal{A}_G := \{(i, j, u, v) : 1 \leq i < j \leq k,\ (u, v) \in \mathrm{Forb}\}.$$

For $z = (z_1, \ldots, z_k) \in (\mathbb{R}^2)^k$, define

$$g(z) := \sum_{r=1}^{k} \sigma_Q(z_r), \qquad \psi_{i,j,u,v}(z) := \langle p_u, z_i \rangle + \langle p_v, z_j \rangle + \sum_{r \in [k] \setminus \{i,j\}} \sigma_Q(z_r),$$

and $h(z) := \max_{\alpha \in \mathcal{A}_G} \psi_\alpha(z)$. Finally set

$$Y := \partial_{\mathrm{C}} g(0) = Q^k, \qquad X := \partial_{\mathrm{C}} h(0).$$

Geometric intuition. Before presenting the full proof, we first provide a geometric intuition to help understanding the high-level idea. Every p_v is an exposed vertex of Q, so the vertices of $Y = Q^k$ represent all ordered k-tuples. Fixing two tuple positions to a forbidden pair gives a product face of Y; X is the convex hull of these faces and contains precisely the non-clique tuple vertices. A missing exposed clique vertex obstructs the containment required by Fréchet stationarity. The seesaw gadget in Lemma C.3 turns that missing vertex into a neighboring zero-gradient region, giving Clarke stationarity.

Lemma C.1 (Clique polytope gadget)**.** *With the notation above, $X \subseteq Y = Q^k$. Moreover, $X = Y$ if and only if G has no k-clique. If G has a k-clique, then there is no $s \in \mathbb{R}^{2k}$ such that $s + Y \subseteq X$. If G has no k-clique, then the only $s \in \mathbb{R}^{2k}$ with $s + Y \subseteq X$ is $s = 0$.*

Proof. For $\alpha = (i, j, u, v) \in \mathcal{A}_G$, let

$$F_\alpha := Q \times \cdots \times Q \times \{p_u\} \times Q \times \cdots \times Q \times \{p_v\} \times Q \times \cdots \times Q \subseteq Q^k,$$

where the singleton factors occur in blocks i and j. Then $\psi_\alpha = \sigma_{F_\alpha}$. By the support-function calculus for polytopes [Ziegler, 1995], the maximum of support functions is the support function of the convex hull of the union, and therefore

$$h = \max_\alpha \sigma_{F_\alpha} = \sigma_{\mathrm{conv}(\cup_\alpha F_\alpha)}.$$

Therefore $X = \partial_{\mathrm{C}} h(0) = \mathrm{conv}(\cup_\alpha F_\alpha) \subseteq Q^k = Y$. Because each F_α is a face of Y, the polytope X is the convex hull of precisely those vertices of Y that lie in at least one F_α. A vertex of Y has the form $y(v_1, \ldots, v_k) := (p_{v_1}, \ldots, p_{v_k})$. It lies in $F_{i,j,u,v}$ if and only if $v_i = u$ and $v_j = v$. Thus it is covered whenever some pair (v_i, v_j) is forbidden. Conversely, if no pair is forbidden, then the tuple corresponds to a k-clique and the vertex is not covered. It also cannot lie in the convex hull of the covered vertices: by the exposed-face facts for polytopes [Ziegler, 1995], every vertex of $Y = Q^k$ is exposed. Indeed, for $p_v = (v, v^2)$ the vector $r_v := (2v, -1)$ satisfies $\langle p_v, r_v \rangle - \langle p_a, r_v \rangle = (a - v)^2 > 0$ for $a \neq v$, and the product vector $(r_{v_1}, \ldots, r_{v_k})$ exposes $y(v_1, \ldots, v_k)$ in Y. Consequently, among the vertices of Y, the polytope X contains exactly those corresponding to non-clique tuples.

If G has no k-clique, every vertex of Y is covered, so $Y = \mathrm{conv}(\mathrm{vert}(Y)) \subseteq X \subseteq Y$ and hence $X = Y$. If G has a k-clique $(v_1, \ldots, v_k)$, the exposed vertex $y^* = y(v_1, \ldots, v_k)$ is missing from X, so $X \subsetneq Y$. Moreover, the exposing vector above gives a strict gap between y^* and every covered vertex; hence X is contained in a closed halfspace that cuts off a nonempty small cap of the full-dimensional polytope Y. Thus $\mathrm{vol}(X) < \mathrm{vol}(Y)$, where volume is taken in $\mathbb{R}^{2k}$. If some translate $s + Y$ were contained in X, then $\mathrm{vol}(Y) = \mathrm{vol}(s + Y) \leq \mathrm{vol}(X) < \mathrm{vol}(Y)$, a contradiction.

Finally, in the no-clique case $X = Y$. If $s + Y \subseteq Y$, then support functions give $\langle s, u \rangle + \sigma_Y(u) \leq \sigma_Y(u)$ for all u, so $\langle s, u \rangle \leq 0$ for all u. Applying this to both u and $-u$ yields $s = 0$. □

Lemma C.2 (Fréchet gadget). *For $f_F := h - g$, the following equivalences hold:*

$$G \text{ has no } k\text{-clique} \iff \partial_{\mathrm{F}} f_F(0) = \{0\}, \qquad G \text{ has a } k\text{-clique} \iff \partial_{\mathrm{F}} f_F(0) = \emptyset.$$

In particular, for every fixed $\varepsilon \geq 0$, G has a k-clique if and only if $\operatorname{dist}(0, \partial_{\mathrm{F}} f_F(0)) > \varepsilon$. Moreover, G has a k-clique if and only if the origin is not a local minimizer of f_F.

Proof. By Lemma 2.4, $\partial_{\mathrm{F}} f_F(0) = \{s : s + Y \subseteq X\}$. The result is exactly Lemma C.1. The final stationarity statement uses the convention $\operatorname{dist}(0, \emptyset) = +\infty$ and the fact that $\operatorname{dist}(0, \{0\}) = 0$. The local-minimality statement follows from Lemma 2.3, since f_F is continuous PA and $0 \in \partial_{\mathrm{F}} f_F(0)$ is equivalent to local minimality at the origin. □

To obtain Clarke hardness, we use the seesaw gadget introduced by [Tian and So, 2025], where a scalar variable t and the branch $-|t|/2$ are added. In the no-clique case, $f_F \equiv 0$, so the gadget is just $t/2$. In the clique case, a negative direction of f_F makes the branch $-|t|/2$ active on an open cone, where $t/2 - |t|/2 = 0$. Thus, the zero slope appears in the clique case, and $0 \in \partial_{\mathrm{C}} f_C(0, 0)$.

Lemma C.3 (Seesaw gadget). *Let $f_C(z, t) = t/2 + \max\{f_F(z), -|t|/2\}$. If G has no k-clique, then $\operatorname{dist}(0, \partial_{\mathrm{C}} f_C(0, 0)) = 1/2$. If G has a k-clique, then $0 \in \partial_{\mathrm{C}} f_C(0, 0)$.*

Proof. If G has no k-clique, then Lemma C.1 gives $X = Y$. Since $h = \sigma_X$ and $g = \sigma_Y$, we have $f_F \equiv 0$. Therefore $f_C(z, t) = t/2 + \max\{0, -|t|/2\} = t/2$, and $\partial_{\mathrm{C}} f_C(0, 0) = \{(0_{2k}, 1/2)\}$. Thus $\operatorname{dist}(0, \partial_{\mathrm{C}} f_C(0, 0)) = 1/2$.

Now suppose G has a k-clique $(v_1, \ldots, v_k)$. For each block choose the exposing vector $r_i := (2v_i, -1) \in \mathbb{R}^2$. Then p_{v_i} uniquely maximizes $\langle \cdot, r_i \rangle$ over Q, and for every $a \neq v_i$,

$$\langle p_{v_i} - p_a, r_i \rangle = (a - v_i)^2 \geq 1.$$

Set $z_i := Tr_i$ and let $z = (z_1, \ldots, z_k)$. For a forbidden branch $\alpha = (i, j, u, v)$ of f_F, its value at z is

$$B_\alpha(z) = \langle p_u, z_i \rangle - \sigma_Q(z_i) + \langle p_v, z_j \rangle - \sigma_Q(z_j) = T\langle p_u - p_{v_i}, r_i \rangle + T\langle p_v - p_{v_j}, r_j \rangle.$$

Both terms are nonpositive, and at least one is at most $-T$, because (v_i, v_j) is not forbidden whereas (u, v) is forbidden. Hence $B_\alpha(z) \leq -T$ for every forbidden branch, and therefore $f_F(z) \leq -T$. Choose $T > 1$. Then $f_F(z) < -1$. At the point $(z, 1)$ we have $f_F(z) < -1 < -1/2$. By continuity, there is an open full-dimensional neighborhood U of $(z, 1)$ such that every $(\tilde{z}, \tilde{t}) \in U$ satisfies $\tilde{t} > 0$ and $f_F(\tilde{z}) < -\tilde{t}/2$. Let

$$C := \{\lambda(\tilde{z}, \tilde{t}) : \lambda > 0, (\tilde{z}, \tilde{t}) \in U\}.$$

Because f_F is positively homogeneous, every $(z', t') \in C$ satisfies $t' > 0$ and $f_F(z') < -t'/2 = -|t'|/2$. Thus, throughout the full-dimensional open cone C, $f_C(z', t') = t'/2 - t'/2 = 0$. Hence the affine piece with gradient 0 is attained on a full-dimensional region adjacent to the origin. For PA functions, the Clarke subdifferential at the origin is the convex hull of gradients on full-dimensional adjacent cells. Therefore $0 \in \partial_{\mathrm{C}} f_C(0, 0)$. □

It remains to check that the hard functions just constructed are given efficiently in the two input models studied in the paper.

Lemma C.4 (Max–min and depth-two n-MC DC representations). *The Fréchet gadget admits the explicit max–min representation*

$$f_F(z) = \max_{1 \leq i < j \leq k,\ (u,v) \in \mathrm{Forb}} \ \min_{a,b \in [N]} \big(\langle p_u - p_a, z_i \rangle + \langle p_v - p_b, z_j \rangle \big).$$

It also admits an explicit polynomial-size DC representation $f_F = h - g$ in which g is 1-MC and h is 2-MC. The Clarke gadget $f_C(z, t) = t/2 + \max\{f_F(z), -|t|/2\}$ admits both an explicit polynomial-size max–min representation and an explicit polynomial-size DC representation $f_C = H_C - G_C$ in which G_C is 1-MC and H_C is 2-MC. In particular, both gadgets have explicit n-MC DC representations of constant depth (indeed depth two after padding).

Proof. Since $\sigma_Q(z_i) = \max_{a \in [N]} \langle p_a, z_i \rangle$,

$$\langle p_u, z_i \rangle - \sigma_Q(z_i) = \min_{a \in [N]} \langle p_u - p_a, z_i \rangle, \qquad \langle p_v, z_j \rangle - \sigma_Q(z_j) = \min_{b \in [N]} \langle p_v - p_b, z_j \rangle.$$

Substituting these identities into the branch formula for $f_F = h - g$ gives the displayed max–min representation.

The convex component $g(z) = \sum_{r=1}^{k} \sigma_Q(z_r) = \sum_{r=1}^{k} \max_{a\in[N]} \langle p_a, z_r\rangle$ is a 1-MC formula. Let $\mathcal{A} := \{(i,j,u,v) : 1 \le i < j \le k,\ (u,v) \in \mathrm{Forb}\}$. For $\alpha = (i,j,u,v) \in \mathcal{A}$, define, for $r \in [k]$ and $a \in [N]$,

$$q_{\alpha,r,a} := \begin{cases} p_u, & r = i, \\ p_v, & r = j, \\ p_a, & r \notin \{i,j\}. \end{cases}$$

Then $\psi_\alpha(z) = \sum_{r=1}^{k} \max_{a\in[N]} \langle q_{\alpha,r,a}, z_r\rangle$, because the max is degenerate in the two fixed blocks. Thus

$$h(z) = \max_{\alpha\in\mathcal{A}} \sum_{r=1}^{k} \max_{a\in[N]} \langle q_{\alpha,r,a}, z_r\rangle,$$

which is a 2-MC formula: an outer maximum over α, followed by a sum over blocks and an inner maximum over a. Consequently $f_F = h - g$ is a DC representation whose convex components have depths at most two after padding the 1-MC formula for g.

For the Clarke gadget, the max–min form follows from $f_C(z,t) = \max\{f_F(z) + t/2, \min\{t, 0\}\}$ and the max–min representation of f_F. For the DC representation, define

$$G_C(z,t) := g(z) + \frac{|t|}{2} = \sum_{r=1}^{k} \max_{a\in[N]} \langle p_a, z_r\rangle + \max\left\{\frac{t}{2}, -\frac{t}{2}\right\}.$$

This is a 1-MC formula. Next set

$$H_C(z,t) := \max\left\{h(z) + \frac{|t|}{2}, g(z)\right\} + \frac{t}{2} = \max\left\{h(z) + \max\{t,0\},\ g(z) + \frac{t}{2}\right\}.$$

The first class of branches can be written as

$$\max_{\alpha\in\mathcal{A},\ \eta\in\{0,1\}} \left[\sum_{r=1}^{k} \max_{a\in[N]} \langle q_{\alpha,r,a}, z_r\rangle + \eta t\right],$$

and the remaining branch is $\sum_{r=1}^{k} \max_{a\in[N]} \langle p_a, z_r\rangle + t/2$. Therefore H_C is an outer maximum over polynomially many branches, each branch being a sum of $k+1$ max-affine groups, where the last group is a singleton affine function in t. Hence H_C is 2-MC. Finally,

$$H_C(z,t) - G_C(z,t) = \frac{t}{2} + \max\left\{h(z) - g(z), -\frac{|t|}{2}\right\} = f_C(z,t).$$

All formulas have polynomial size in N and k, and the depths are bounded by two. □

Proof of Theorem 4.1. We reduce from the k-Clique problem; its W[1]-hardness is classical [Cygan et al., 2015]. Lemma C.2 proves the Fréchet hardness equivalence and the corresponding non-local-minimality equivalence, and Lemma C.3 proves the Clarke hardness equivalence for every fixed $\varepsilon \in [0, 1/2)$. Lemma C.4 provides both required input representations for our constructed hard instances. The Fréchet construction has dimension $2k$ and the Clarke construction has dimension $2k+1$. In both constructions, the branches are indexed by at most $O(k^2N^2)$ forbidden choices together with polynomially many vertex choices, and all coefficients have polynomial bit length. Thus the instances can be output in time $f(k)|(G,k)|^{O(1)}$ and the target dimension is bounded by a function of k, so both reductions are valid parameterized reductions. Since Hypothesis 2.1 rules out algorithms for k-Clique on N-vertex graphs running in time $\rho(k)N^{o(k)}$, and since the target encoding size satisfies $\mathsf{s} \le N^{O(1)}$ with $d = \Theta(k)$, the same reductions rule out algorithms for the Fréchet, non-local-minimality, and Clarke hard polarities running in time $\rho(d)\mathsf{s}^{o(d)}$.

D Deferred proofs for Section 5

This appendix transfers the clique gadgets from Appendix C to the displayed shallow CNN family. First, we prove the local ReLU penalty that replaces the support-function terms used in the PA gadget. Second, we show that the losses in our hard instances can be realized using a shared trainable 1×1 convolution, ReLUs, fixed linear readouts, and max-pooling. Finally, we verify the Fréchet and Clarke hardness reductions and provide an XP algorithm for the stationarity tests of the constructed shallow CNN loss family.

We use the same graph notation as in Appendix C. Throughout, $G = ([N], E)$,

$$p_v = (v, v^2) \in \mathbb{Z}^2, \qquad \mathcal{N}_1(u) = \{a \in [N] : |a - u| = 1\},$$

and

$$\mathrm{Forb} = \{(u, v) \in [N]^2 : u = v \text{ or } \{u, v\} \notin E\}.$$

As before, Forb records exactly the ordered pairs that cannot appear as two positions of a valid clique. We assume $N \geq 2$; smaller instances are trivial or may be padded by isolated vertices.

D.1 The local ReLU penalty

Lemma D.1 (Separation property of the local ReLU penalty)**.** *For $u \in [N]$, define*

$$P_u(s) := \sum_{a \in \mathcal{N}_1(u)} \mathrm{ReLU}\big(\langle p_a - p_u, s\rangle\big), \qquad s \in \mathbb{R}^2.$$

If p_u maximizes $a \mapsto \langle p_a, s\rangle$ over $[N]$, then $P_u(s) = 0$. Moreover, for $r_c = (2c, -1)$, one has $P_u(r_c) = 0$ when $u = c$, and $P_u(r_c) \geq 1$ when $u \neq c$.

Proof. If p_u maximizes $a \mapsto \langle p_a, s\rangle$ over $[N]$, then $\langle p_a - p_u, s\rangle \leq 0$ for every $a \in \mathcal{N}_1(u)$, and hence every ReLU term in $P_u(s)$ vanishes.

For the second claim, note that for $r_c = (2c, -1)$,

$$\langle p_a, r_c\rangle = 2ca - a^2 = c^2 - (a - c)^2.$$

Thus $a = c$ is the unique maximizer over $[N]$, so $P_c(r_c) = 0$. If $u < c$, then $u + 1 \in \mathcal{N}_1(u)$ and

$$\langle p_{u+1} - p_u, r_c\rangle = (u - c)^2 - (u + 1 - c)^2 = 2(c - u) - 1 \geq 1.$$

If $u > c$, then $u - 1 \in \mathcal{N}_1(u)$ and

$$\langle p_{u-1} - p_u, r_c\rangle = (u - c)^2 - (u - 1 - c)^2 = 2(u - c) - 1 \geq 1.$$

In either case at least one ReLU term in $P_u(r_c)$ is at least 1, proving $P_u(r_c) \geq 1$ whenever $u \neq c$. □

D.2 Realizing the losses by shallow ReLU/max-pooling CNNs

The network class. Following the shared-filter construction of Tian and So [2025, Appendix B.3], fix one rational training tensor with p channels, and let Z be the matrix whose rows are the channel vectors at its spatial sites. A shared trainable no-bias 1×1 filter $\theta \in \mathbb{R}^p$ produces $h_\theta^{(0)} = Z\theta$. The subsequent layers satisfy $h_\theta^{(\ell)} = T_\ell(h_\theta^{(\ell-1)})$, where each T_ℓ is a fixed rational affine map, coordinatewise ReLU, or max-pooling over fixed windows. The depth is bounded by an absolute constant, and only θ is trained. The construction below realizes Equations (2)–(3) within this class; their scalar outputs can be viewed as objectives with an identity loss head.

Let

$$\mathcal{A}_G := \{(i, j, u, v) : 1 \leq i < j \leq k,\ (u, v) \in \mathrm{Forb}\}.$$

For the Fréchet-side construction, the trainable parameter is

$$\theta = z = (z_1, \ldots, z_k) \in (\mathbb{R}^2)^k = \mathbb{R}^{2k}.$$

For every $\alpha = (i, j, u, v) \in \mathcal{A}_G$ and $a \in \mathcal{N}_1(u)$, create a spatial site with fixed channel vector

$$x_{\alpha,L,a} := (0, \ldots, 0, \underbrace{p_a - p_u}_{i\text{-th block}}, 0, \ldots, 0) \in \mathbb{Q}^{2k}.$$

For every $b \in \mathcal{N}_1(v)$, create a spatial site with fixed channel vector

$$x_{\alpha,R,b} := (0, \dots, 0, \underbrace{p_b - p_v}_{j\text{-th block}}, 0, \dots, 0) \in \mathbb{Q}^{2k}.$$

Write

$$B_\alpha(z) := -P_u(z_i) - P_v(z_j).$$

The trainable 1×1 convolutional filter outputs $\langle p_a - p_u, z_i \rangle$ and $\langle p_b - p_v, z_j \rangle$ on these sites. After the ReLU layer, a fixed rational readout over the sites belonging to branch α computes

$$-\sum_{a \in \mathcal{N}_1(u)} \mathrm{ReLU}\big(\langle p_a - p_u, z_i \rangle\big) - \sum_{b \in \mathcal{N}_1(v)} \mathrm{ReLU}\big(\langle p_b - p_v, z_j \rangle\big) = B_\alpha(z).$$

This readout can be implemented as a fixed convolutional/readout window with all nonzero coefficients equal to -1. A global max-pooling layer over the branch channels computes

$$L_G^{\mathrm{F}}(z) = \max_{\alpha \in \mathcal{A}_G} B_\alpha(z).$$

The number of sites is at most $4|\mathcal{A}_G| = O(k^2 N^2)$, and all coefficients have polynomial bit length.

For the Clarke-side construction, the trainable parameter is

$$\theta = (z, t) \in \mathbb{R}^{2k+1}.$$

Use the same branch sites, padded by one zero coordinate, and add two scalar sites with channel vectors e_{2k+1} and $-e_{2k+1}$. After ReLU, these two sites output $\mathrm{ReLU}(t)$ and $\mathrm{ReLU}(-t)$. A fixed rational readout computes

$$R_t(t) := -\frac{1}{2}\,\mathrm{ReLU}(t) - \frac{1}{2}\,\mathrm{ReLU}(-t) = -\frac{|t|}{2}.$$

The final max-pooling layer computes $\max\{L_G^{\mathrm{F}}(z), R_t(t)\}$, and the scalar output adds the fixed affine term $t/2$. Equivalently, this affine term can be written as $\mathrm{ReLU}(t)/2 - \mathrm{ReLU}(-t)/2$ using the same two scalar sites. Thus the realized loss is exactly

$$L_G^{\mathrm{C}}(z, t) = \frac{t}{2} + \max\left\{ L_G^{\mathrm{F}}(z), -\frac{|t|}{2} \right\}.$$

A standard supervised loss head. Let S_G denote either displayed scalar output L_G^{F} or L_G^{C}. Assign the fixed training example label $y = -1$ and use the standard hinge loss $\ell(y, s) = \max\{0, 1 - ys\}$, giving

$$\widehat{L}_G(\theta) = \max\{0, 1 + S_G(\theta)\}.$$

The stationarity query point is $\theta = 0$, where every ReLU penalty vanishes and $S_G(0) = 0$. Continuity gives $S_G(\theta) > -1$ near the origin, so $\widehat{L}_G = 1 + S_G$ throughout that neighborhood. Thus the Fréchet and Clarke subdifferentials and local-minimality behavior at the query point agree. The fixed affine/ReLU head adds no trainable parameters and preserves constant depth. The lower bounds in Theorem 5.1 therefore also hold for these one-example hinge losses, and for the constant-depth loss class containing them.

D.3 Proof of Theorem 5.1

Proof of Theorem 5.1. The functions L_G^{F} and L_G^{C} are continuous, PA, and positively homogeneous. Moreover $L_G^{\mathrm{F}}(z) \le 0$ for all z, because every penalty P_u is nonnegative.

Fréchet reduction: no-clique case. Assume that G has no k-clique. Fix arbitrary $z = (z_1, \dots, z_k)$. For each block $r \in [k]$, choose

$$v_r \in \arg\max_{a \in [N]} \langle p_a, z_r \rangle.$$

Since $(v_1, \dots, v_k)$ is not a clique, there exist $i < j$ such that $(v_i, v_j) \in \mathrm{Forb}$. For the branch $\alpha = (i, j, v_i, v_j)$, Lemma D.1 gives

$$P_{v_i}(z_i) = P_{v_j}(z_j) = 0,$$

and hence $B_\alpha(z) = 0$. All branch scores are nonpositive, so $L_G^{\mathrm{F}}(z) = 0$. As z was arbitrary, $L_G^{\mathrm{F}} \equiv 0$. Therefore

$$\partial_{\mathrm{F}} L_G^{\mathrm{F}}(0) = \{0\}, \qquad \mathrm{dist}(0, \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)) = 0.$$

Fréchet reduction: clique case. Assume that G has a k-clique $(c_1, \dots, c_k)$. For $T > 0$, set

$$z_i := T(2c_i, -1).$$

For any forbidden branch $\alpha = (i, j, u, v)$, the ordered pair (c_i, c_j) is allowed whereas (u, v) is forbidden, so $u \neq c_i$ or $v \neq c_j$. Lemma D.1 gives $P_u(z_i) \geq T$ in the first case and $P_v(z_j) \geq T$ in the second. Thus

$$B_\alpha(z) \leq -T$$

for every branch, and $L_G^{\mathrm{F}}(z) \leq -T < 0$.

We now show that $\partial_{\mathrm{F}} L_G^{\mathrm{F}}(0) = \emptyset$. Suppose $g \in \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)$. Since L_G^{F} is positively homogeneous, the Fréchet subgradient inequality gives

$$\langle g, d \rangle \leq L_G^{\mathrm{F}}(d) \qquad \forall d.$$

Taking $d = g$ and using $L_G^{\mathrm{F}} \leq 0$ yields $\|g\|^2 \leq 0$, so $g = 0$. But $0 \in \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)$ would imply $L_G^{\mathrm{F}}(d) \geq 0$ for all d, contradicting the negative clique direction above. Hence

$$\partial_{\mathrm{F}} L_G^{\mathrm{F}}(0) = \emptyset.$$

By the convention $\mathrm{dist}(0, \emptyset) = +\infty$, clique instances are exactly the instances with

$$\mathrm{dist}(0, \partial_{\mathrm{F}} L_G^{\mathrm{F}}(0)) > \varepsilon$$

for every fixed $\varepsilon \geq 0$. The same two cases also decide local minimality: when there is no clique, $L_G^{\mathrm{F}} \equiv 0$, so the origin is a local minimizer; when there is a clique, the negative direction above shows that the origin is not a local minimizer.

Clarke reduction: no-clique case. If G has no k-clique, then $L_G^{\mathrm{F}} \equiv 0$. Thus

$$L_G^{\mathrm{C}}(z, t) = \frac{t}{2} + \max\left\{0, -\frac{|t|}{2}\right\} = \frac{t}{2}.$$

Consequently

$$\partial_{\mathrm{C}} L_G^{\mathrm{C}}(0, 0) = \{(0, \dots, 0, 1/2)\}, \qquad \mathrm{dist}(0, \partial_{\mathrm{C}} L_G^{\mathrm{C}}(0, 0)) = \frac{1}{2}.$$

For every fixed $\varepsilon \in [0, 1/2)$, no-clique instances are therefore Clarke-ε-NO instances.

Clarke reduction: clique case. If G has a k-clique, the Fréchet part gives $z^\star$ with

$$L_G^{\mathrm{F}}(z^\star) < -1.$$

At $(z^\star, 1)$, the branch $-|t|/2$ has value $-1/2$ and is strictly larger than $L_G^{\mathrm{F}}(z^\star)$. By continuity, there is a full-dimensional open neighborhood U of $(z^\star, 1)$ on which $t > 0$ and the branch $-|t|/2$ is strictly selected. On U,

$$L_G^{\mathrm{C}}(z, t) = \frac{t}{2} - \frac{t}{2} = 0.$$

Since L_G^{C} is positively homogeneous, the cone generated by a sufficiently small open subset of U is a full-dimensional region whose closure contains the origin, and on this region the affine gradient is 0. For PA functions, the Clarke subdifferential at the origin is the convex hull of gradients attained on adjacent full-dimensional regions. Thus

$$0 \in \partial_{\mathrm{C}} L_G^{\mathrm{C}}(0, 0).$$

This proves that clique instances are exactly the instances with

$$\mathrm{dist}(0, \partial_{\mathrm{C}} L_G^{\mathrm{C}}(0, 0)) \leq \varepsilon$$

for every fixed $\varepsilon \in [0, 1/2)$.

Parameterized and conditional running-time consequences. The Fréchet construction has trainable dimension $p = 2k$; the Clarke construction has trainable dimension $p = 2k + 1$. The number of sites, fixed readout coefficients, and pooling windows is polynomial in N and k: there are $O(kN)$ local penalty channels and $O(k^2 N^2)$ forbidden-branch channels, all with rational coefficients of polynomial bit length. Thus the CNN instances can be output in time $f(k)|(G, k)|^{O(1)}$, with p bounded by a function of k, so the reductions are valid parameterized reductions. If any of the hard CNN problems above were solvable in time $\rho(p)M^{o(p)}$, then k-Clique on N vertices would be solvable in time $\rho'(k)N^{o(k)}$ for a computable function ρ', contradicting Hypothesis 2.1. □

D.4 Conversion to explicit max–min form

Proof of Corollary 5.2. For $\alpha = (i, j, u, v) \in \mathcal{A}_G$, write

$$\ell_a^{i,u}(z) := \langle p_a - p_u, z_i \rangle \quad (a \in \mathcal{N}_1(u)), \qquad \ell_b^{j,v}(z) := \langle p_b - p_v, z_j \rangle \quad (b \in \mathcal{N}_1(v)).$$

The identity

$$-\operatorname{ReLU}(\ell) = \min\{0, -\ell\} = \min_{\omega \in \{0,1\}} (-\omega \ell)$$

and distributivity of independent minima over sums give

$$B_\alpha(z) = \min_{\omega \in \{0,1\}^{\mathcal{N}_1(u)},\, \tau \in \{0,1\}^{\mathcal{N}_1(v)}} \left(- \sum_{a \in \mathcal{N}_1(u)} \omega_a \ell_a^{i,u}(z) - \sum_{b \in \mathcal{N}_1(v)} \tau_b \ell_b^{j,v}(z) \right).$$

Since

$$|\mathcal{N}_1(u)|,\ |\mathcal{N}_1(v)| \le 2,$$

each branch has at most $2^4 = 16$ affine pieces. Thus

$$L_G^{\mathrm{F}}(z) = \max_{\alpha \in \mathcal{A}_G} \min_{\omega, \tau} \left(- \sum_{a \in \mathcal{N}_1(u)} \omega_a \ell_a^{i,u}(z) - \sum_{b \in \mathcal{N}_1(v)} \tau_b \ell_b^{j,v}(z) \right)$$

is an explicit max–min formula of size $O(|\mathcal{A}_G|)$ up to a constant factor, and hence of size polynomial in M.

For the Clarke loss, use

$$\frac{t}{2} - \frac{|t|}{2} = \min\{t, 0\}.$$

Combining this identity with the preceding branch expansion yields

$$L_G^{\mathrm{C}}(z, t) = \max \left\{ \max_{\alpha \in \mathcal{A}_G} \min_{\omega, \tau} \left[\frac{t}{2} - \sum_{a \in \mathcal{N}_1(u)} \omega_a \ell_a^{i,u}(z) - \sum_{b \in \mathcal{N}_1(v)} \tau_b \ell_b^{j,v}(z) \right],\ \min\{t, 0\} \right\}.$$

This is again an explicit max–min formula of size polynomial in M and dimension $p = 2k + 1$. Applying Theorem 3.1 gives deterministic time $M^{O(p)}$ for Fréchet-ε-YES and Clarke-ε-NO, and complementing the deterministic output gives the same upper bound for the opposite polarities. Together with Theorem 5.1, this proves Corollary 5.2. □

Scope of the upper bound and network parameters. The $M^{O(p)}$ upper bound in Corollary 5.2 concerns the displayed constructed family. It does not assert an XP algorithm for the entire network class described above. In the construction, the channel dimension equals p, so the bound is XP in channel dimension and formally in the larger total tensor dimension. The output dimension is always one; an XP algorithm parameterized only by output dimension would imply $P = NP$.

NeurIPS Paper Checklist

1. **Claims**
   - **Question:** Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   - **Answer:** [Yes]
   - **Justification:** The abstract and Introduction state the paper's XP algorithms, W[1]-hardness results, conditional running-time lower bounds, and transfer to a specific one-filter ReLU CNN training family. The assumptions and scope are stated in Sections 2–5, with limitations explicitly discussed in the limitations paragraph before the appendices.
   - **Guidelines:**
     - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
     - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
     - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
     - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   - **Question:** Does the paper discuss the limitations of the work performed by the authors?
   - **Answer:** [Yes]
   - **Justification:** The paper includes a dedicated limitations paragraph before the appendices. It clarifies that the lower bounds are worst-case, that the CNN transfer is for an explicit restricted family, and that the XP algorithms are fixed-dimensional enumeration results rather than practical large-dimensional solvers.
   - **Guidelines:**
     - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
     - The authors are encouraged to create a separate "Limitations" section in their paper.
     - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
     - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
     - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
     - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
     - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
     - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that are not acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

- **Question:** For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
- **Answer:** [Yes]
- **Justification:** The paper is theoretical. Theorems and lemmas state the relevant input models, parameters, tolerances, and representations; proof sketches appear in the main text and full details are provided in Appendices A.2–D.
- **Guidelines:**
  - The answer [N/A] means that the paper does not include theoretical results.
  - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
  - All assumptions should be clearly stated or referenced in the statement of any theorems.
  - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
  - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
  - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   - **Question:** Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   - **Answer:** [N/A]
   - **Justification:** The paper does not include experiments. The reproducible objects are mathematical reductions, algorithms, and network constructions, which are specified in the theorem statements and proofs.
   - **Guidelines:**
     - The answer [N/A] means that the paper does not include experiments.
     - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
     - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
     - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
     - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example:
       * If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
       * If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
       * If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
       * We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   - **Question:** Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   - **Answer:** [N/A]
   - **Justification:** The paper has no empirical results requiring data or code. The constructions use synthetic graph instances and explicit rational formulas described in the text.
   - **Guidelines:**
     - The answer [N/A] means that paper does not include experiments requiring code.
     - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
     - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
     - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
     - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
     - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
     - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
     - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.
6. **Experimental setting/details**
   - **Question:** Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   - **Answer:** [N/A]
   - **Justification:** The paper does not include experimental training or test evaluations. The CNN section is an exact symbolic construction rather than a trained empirical model.
   - **Guidelines:**
     - The answer [N/A] means that the paper does not include experiments.
     - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
     - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   - **Question:** Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   - **Answer:** [N/A]
   - **Justification:** The paper does not include experiments or statistical estimates.
   - **Guidelines:**
     - The answer [N/A] means that the paper does not include experiments.
     - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
     - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.).
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or plots symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   - **Question:** For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   - **Answer:** [N/A]
   - **Justification:** The paper does not include experiments that consume compute resources. No empirical compute claims are made.
   - **Guidelines:**
     - The answer [N/A] means that the paper does not include experiments.
     - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
     - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
     - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that did not make it into the paper).

9. **Code of ethics**
   - **Question:** Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
   - **Answer:** [Yes]
   - **Justification:** The paper is a theoretical study of complexity and exact symbolic reductions. It does not involve human subjects, private data, deployed systems, or data collection, and the results are presented with explicit assumptions and limitations.
   - **Guidelines:**
     - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
     - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
     - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
    - **Question:** Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    - **Answer:** [N/A]
    - **Justification:** This is foundational theoretical work; it releases no model, dataset, or deployed system, and we do not identify a direct path to negative societal impact.
    - **Guidelines:**
      - The answer [N/A] means that there is no societal impact of the work performed.
      - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from intentional or unintentional misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

- **Question:** Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
- **Answer:** [N/A]
- **Justification:** The paper does not release data or models, and it does not introduce a high-risk artifact requiring release safeguards.
- **Guidelines:**
  - The answer [N/A] means that the paper poses no such risks.
  - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
  - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
  - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

- **Question:** Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
- **Answer:** [N/A]
- **Justification:** The paper does not use external code, datasets, or model assets. Prior scientific work is credited through citations.
- **Guidelines:**
  - The answer [N/A] means that the paper does not use existing assets.
  - The authors should cite the original paper that produced the code package or dataset.
  - The authors should state which version of the asset is used and, if possible, include a URL.
  - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
  - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
  - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
    - **Question:** Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    - **Answer:** [N/A]
    - **Justification:** The paper does not release new datasets, code packages, model checkpoints, or other assets.
    - **Guidelines:**
      - The answer [N/A] means that the paper does not release new assets.
      - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
      - The paper should discuss whether and how consent was obtained from people whose asset is used.
      - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
    - **Question:** For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    - **Answer:** [N/A]
    - **Justification:** The paper does not involve crowdsourcing, user studies, or research with human subjects.
    - **Guidelines:**
      - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
      - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
      - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    - **Question:** Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    - **Answer:** [N/A]
    - **Justification:** The paper does not involve crowdsourcing or research with human subjects, so IRB or equivalent approval is not applicable.
    - **Guidelines:**
      - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
      - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
      - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
      - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

- **Question:** Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
- **Answer:** [Yes]
- **Justification:** See the AI assistance statement at the end of the main text.
- **Guidelines:**
  - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
  - Please refer to the LLM policy in the NeurIPS handbook for what should or should not be described.