BROXIMAL ALIGNMENT FOR GLOBAL NON-CONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most non-convex optimization theory is built around gradient dynamics, leaving global convergence largely unexplored. The dominant paradigm focuses on stationarity, certifying only that the gradient norm vanishes, which is often a weak proxy for actual optimization success. In practice, gradient norms can stagnate or even increase during training, and stationary points may be far from global solutions. In this work, we propose a new framework for global non-convex optimization that avoids gradient-based reasoning altogether. We revisit the Ball Proximal Point Method (BPM), a trust-region–style algorithm introduced by Gruntkowska et al. (2025), and propose a novel structural condition—*Broximal Alignment*—under which BPM provably converges to a global minimizer. Our condition requires no convexity, smoothness, or Lipschitz assumptions, and it permits multiple and disconnected global minima as well as non-optimal local minima. We show that this class generalizes standard non-convex frameworks such as quasi-convexity, star convexity, quasar convexity, and the aiming condition. Our results provide a new conceptual foundation for global non-convex optimization beyond stationarity.

1 INTRODUCTION

Non-convex optimization lies at the heart of modern machine learning (ML). Matrix factorization (Koren et al., 2009), phase retrieval (Patterson, 1934), representation learning, and most prominently deep neural networks (LeCun et al., 2015; Goodfellow et al., 2016) all lead to objective functions with highly non-convex landscapes. Although such functions provide the expressive capacity needed to model complex learning tasks, they often pose significant challenges for optimizer designers. The underlying landscape can be highly irregular, with many poor local minima and saddle points that can trap optimization algorithms (Pascanu et al., 2014; Safran & Shamir, 2018; Jin et al., 2021).

Unlike convex optimization, where global minimizers can be efficiently found (Nemirovski & Yudin, 1983; Nemirovski & Nesterov, 1985; Nesterov, 2003), we lack a comparable toolkit for non-convex problems. Indeed, many non-convex optimization tasks are NP-hard to solve (Murty & Kabadi, 1987), and a number of them are NP-hard even to approximate (Meka et al., 2008; Jain & Kar, 2017). As a result, rather than aiming for (approximate) global solutions, most theoretical results for non-convex optimization, especially those for gradient-based methods, are framed in terms of *stationarity*: convergence to points where the gradient norm is small. Although mathematically convenient, such guarantees provide limited insight into *global* optimization behavior.

Limitations of Stationarity-Based Non-convex Optimization Theory. To fix notation, we consider solving the general (potentially non-convex) optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{1}$$

where $f:\mathbb{R}^d \mapsto \mathbb{R}\cup\{+\infty\}$ is proper (that is, $\operatorname{dom} f := \{x\in\mathbb{R}^d : f(x) < +\infty\} \neq \emptyset$), closed, and has at least one minimizer. We denote the set of all minimizers by $\mathcal{X}_f$ and write $f_\star := \min_x f(x)$.

The dominant theoretical paradigm for such tasks is built around *stationarity*. Convergence guarantees are typically stated in terms of the norm of the gradient, asserting that an algorithm reaches a point where $\|\nabla f(x)\|$ is small. This framework underlies much of the modern theory of stochastic

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gradient methods. From the modeling perspective, however, stationarity-based guarantees are fundamentally misaligned with the goals of learning. A stationary point can be a global minimizer, a poor local minimizer, or a saddle point; first-order information alone does not distinguish among these possibilities. Therefore, a small gradient provides little guidance on whether an algorithm meaningfully solves the underlying problem. Indeed, in highly non-convex landscapes, large regions of near-zero gradient may exist far from any global solution (Pascanu et al., 2014). Algorithms that follow gradient-based dynamics can stagnate in such regions, producing iterates that are theoretically certified as "successful" while being arbitrarily suboptimal in objective value (Jin et al., 2017). In this sense, stationarity is a weak proxy for optimization progress.

Empirical evidence further challenges the relevance of gradient norms as a convergence signal. In long training runs of modern ML models, the gradient norm has been observed not to decrease, and in some cases to *increase*, toward the end of training (Defazio, 2025). This behavior has been attributed to interactions between weight decay, learning rate schedules, and normalization layers.

In practice, escaping stationary points is usually left to stochasticity and noise, either explicitly (e.g., perturbed gradient methods (Jin et al., 2017; Kleinberg et al., 2018; Jin et al., 2021; Ahn et al., 2023)) or implicitly (e.g., minibatch sampling (Smith & Le, 2017)). Even when successful, the resulting guarantees of such methods remain fundamentally local. At best, they ensure convergence to approximate second-order stationary points, which still need not correspond to globally good solutions. Thus, much of the existing non-convex optimization theory addresses the question: *what local geometric properties characterize the iterates produced by an algorithm?* rather than the more fundamental question: *when does an algorithm find a good solution?* This state of affairs motivates the central question of this work:

How to obtain global convergence guarantees for non-convex problems without relying on gradient-based dynamics and gradient-norm metrics?

Addressing it requires rethinking the algorithmic primitives and structural assumptions of non-convex optimization, moving toward mechanisms that are not driven solely by local first-order information.

Ball Proximal Point Method. To study the questions raised above, we revisit the Ball Proximal Point Method (BPM) – a conceptual framework for global optimization introduced by Gruntkowska et al. (2025).[1] BPM is a trust-region–type method in which each iterate is obtained by minimizing the objective over a ball of radius $t_k > 0$ centered at the current point:

$$\boxed{x_{k+1} \in \operatorname{brox}_f^{t_k}(x_k)} \tag{2}$$

(in the remainder of this work, we restrict attention to the constant-radius setting and assume $t_k \equiv t > 0 \; \forall k$), where the *ball-proximal ("broximal") operator* is defined as

$$\operatorname{brox}_f^t(x) := \operatorname*{arg\,min}_{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)} f(z).$$

Here $\mathcal{B}_{\boldsymbol{X}}(x,t) = \left\{z \in \mathbb{R}^d : \|z - x\|_{\boldsymbol{X}} \leq t\right\}$ and $\|\cdot\|_{\boldsymbol{X}}$ is the norm induced by a symmetric positive definite matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, defined by $\|x\|_{\boldsymbol{X}} := \sqrt{x^\top \boldsymbol{X} x}$.[2] We denote $\langle x, y\rangle_{\boldsymbol{X}} := \langle \boldsymbol{X}x, y\rangle = x^\top \boldsymbol{X} y$. When $\boldsymbol{X} = \boldsymbol{I}$, we omit the subscript and simply write $\mathcal{B}(x,t)$, $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$. Like the classical Proximal Point Method (PPM) (Rockafellar, 1976), BPM solves an auxiliary optimization problem at each iteration, but with a hard trust-region constraint instead of a quadratic regularizer. The operator is set-valued in general, which is why we write "$\in$" rather than "$=$" in equation 2. When multiple minimizers exist within the ball, the algorithm is free to choose any of them.

Motivated by challenges in non-convex optimization, the method was originally developed as an oracle capable of escaping local minima and saddle points. Along the way, it was found to enjoy unusually strong convergence guarantees: it converges *linearly* without the strong convexity assumptions required by methods such as PPM. Moreover, while PPM with finite steps can only

[1] Ball oracles have also been studied in several prior works (Carmon et al., 2020; 2021; Asi et al., 2021); see Gruntkowska et al. (2025) for a review.

[2] The original BPM paper (Gruntkowska et al., 2025) considered the simplified setting where $\boldsymbol{X} = \boldsymbol{I}$ is the identity matrix. Other norms were considered in Gruntkowska & Richtárik (2025).

produce approximate solutions, BPM can attain the *exact global minimum* in a *finite number of iterations* using *finite radii*. Despite this promise, and despite evidence that the method can handle non-convexity (see Figure 2), existing analyses focus almost exclusively on convex objectives, with only limited non-convex treatment (Gruntkowska et al., 2025).

Our aim is to understand what properties of the objective enable BPM to achieve global convergence, and to identify the structural assumptions under which such guarantees can be established in the *non-convex* setting. Crucially, we do *not* view BPM as a practical algorithm in its current form, nor do we claim efficient implementations for the problems we study. Indeed, the broximal oracle itself is very powerful and generally intractable: if the radius is sufficiently large ($t_0 \geq \|x_0 - x_\star\|_{\boldsymbol{X}}$, where x_0 is a starting point and $x_\star$ is an optimal point), then $x_\star \in \operatorname{brox}_f^{t_0}(x_0)$ and the algorithm finds a global solution in 1 step. Thus, solving the subproblem can be as difficult as solving the original problem itself. While practical implementations are possible (see Section 6), they are not the focus of this work. Rather, we treat BPM as a conceptual tool for isolating and studying a novel class of non-convex functions for which global convergence can be established. This class stands on its own and may be viewed as an analytic tool for studying optimization algorithms beyond BPM itself.

2 Contributions

This paper proposes a new framework for global non-convex optimization beyond stationarity. We introduce a novel function class under which BPM converges globally, and show that this condition strictly subsumes several established non-convex frameworks. Our main contributions are:

▷ **New assumptions for non-convex optimization.** We introduce a novel class of non-convex functions, termed *Broximal Aligned Functions*, defined by Assumptions 3.1 and 3.2. The framework enables global optimization guarantees beyond gradient-based reasoning.

▷ **Global convergence of BPM.** We prove that BPM converges to a global minimizer within this class, without reliance on gradients or stochastic noise (Theorem 4.1).

▷ **Interpretation and examples.** We provide intuitive interpretations, comparisons to existing function classes, and concrete examples demonstrating that our assumptions are weaker than several standard conditions such as pseudoconvexity, quasiconvexity, quasar convexity, and the aiming condition (Section 5).

The remainder of the paper is organized as follows. In Section 3, we introduce the new function class via Assumptions 3.1 and 3.2 and motivate its origins. Building on this abstract framework, Section 4 presents global convergence guarantees for BPM. Section 5 then clarifies the assumptions by providing examples and comparisons to other non-convex function classes. We provide comments on practical implementations in Section 6 and conclude in Section 7. All proofs are deferred to the Appendix, and a complete list of notation is provided in Table 2.

We emphasize that this work is theoretical and foundational in nature. Our goal is not to propose a new practical optimizer, but to broaden the conceptual foundations of non-convex optimization by identifying mechanisms that enable global guarantees beyond gradient-based paradigms. We hope that this perspective will stimulate further research into practical approximations and connections to modern ML optimization methods.

3 Optimization Beyond Convexity

Non-convex optimization is generally challenging, yet algorithmic developments have yielded both heuristic and provably convergent approaches. One pragmatic approach exploits the local nature of many optimization methods by running them in parallel from multiple initializations and aggregating the resulting solutions. Another line of work studies particle-based approaches (Fornasier et al., 2020; 2024; Fornasier & Sun, 2025), which address non-convex problems by reformulating them as PDE-governed dynamics over measures. Although originally heuristic, these methods now admit provable global solutions. The main drawback is computational: approximating the PDE typically requires an exponentially large or even infinite number of particles. In addition, certain problems become tractable when the objective has some special structure. For example, one-layer neural networks have no spurious local minima (Feizi et al., 2017; Haeffele & Vidal, 2017; Soltanolkotabi

et al., 2018), and SGD can find global minima in linear and sufficiently wide over-parameterized networks under appropriate assumptions (Danilova et al., 2022; Shin, 2022; Allen-Zhu et al., 2019). Similarly, with suitable assumptions on input data, a range of non-convex problems, including matrix completion, some deep learning tasks, phase retrieval, and eigenvalue problems, display "convexity-like" behavior (Ge et al., 2016; 2017; Bartlett et al., 2018; Kangal et al., 2018).

Despite the strong theoretical guarantees offered by convex optimization, its assumptions are often too restrictive for modern large-scale models. This has motivated the development of intermediate frameworks that retain meaningful convergence properties while being applicable to broader problem classes. Examples include star convexity (Nesterov & Polyak, 2006), quasar convexity (Hardt et al., 2018; Hinder et al., 2019), pseudoconvexity (Mangasarian, 1975), quasiconvexity (Arrow & Enthoven, 1961), invexity (Craven & Glover, 1985), variational coherence (Zhou et al., 2017), the Polyak-Łojasiewicz (PŁ) condition (Polyak, 1963; Łojasiewicz, 1963), and aiming condition (Liu et al., 2023). Each of these offers certain benefits, but also comes with inherent limitations.

Here, we move beyond settings in which global convergence follows from the absence of suboptimal local minima, and instead pursue the more ambitious goal of achieving global optimality without relying on such favorable geometric properties. This shift requires abandoning first-order methods in favor of more powerful ball oracles. Drawing on the theory of BPM, we reverse-engineer the underlying conditions necessary to achieve global convergence under this broader framework.

The Origins. The original analysis of BPM relies on the *Second Brox Theorem* (Gruntkowska et al., 2025), exploiting the fact that if f is convex, then there exists $c_t(x) \geq 0$ such that

$$f(y) - f(u) \geq c_t(x) \langle x - u, y - u \rangle_{\boldsymbol{X}}$$

for $x \in \operatorname{dom} f$, $u \in \operatorname{brox}_f^t(x)$ and any $y \in \mathbb{R}^d$. In particular, specializing to $y = x_\star$, where $x_\star \in \mathcal{X}_f$, yields $c_t(x) \langle x - u, u - x_\star \rangle_{\boldsymbol{X}} \geq f(u) - f_\star \geq 0$. This observation motivates dropping convexity altogether and instead postulating this inequality directly, leading to the following assumption:

Assumption 3.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function, and let $t > 0$. There exists $x_\star \in \mathcal{X}_f$ such that, for every $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0, t)\}$ and for any $u \in \operatorname{brox}_f^t(x)$

$$\langle x - u, x_\star - u \rangle_{\boldsymbol{X}} \leq 0. \tag{3}$$

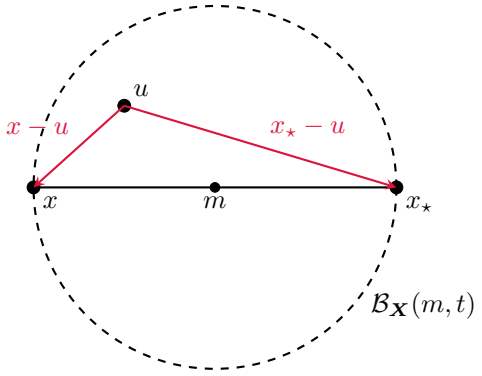

Figure 1: The inequality in Assumption 3.1 is equivalent to requiring that u lies inside the ball centered at the midpoint $m = \frac{1}{2}(x + x_\star)$ with radius $t = \frac{1}{2}\|x - x_\star\|_{\boldsymbol{X}}$. Equivalently, the angle at u in the triangle formed by x, u and $x_\star$ is ≥ 90 degrees (in the geometry induced by $\boldsymbol{X}$).

Intuitively, whenever x lies outside of the t-neighborhood of the set of global minimizers $\mathcal{X}_f$, any local minimizer u of f within the ball $\mathcal{B}_{\boldsymbol{X}}(x, t)$ cannot lie "past" some global minimizer $x_\star$, ensuring that the step from x to u is aligned toward it. The relaxation $\operatorname{dist}(x, \mathcal{X}_f) > t$ is essential: if the condition were imposed for all $x \in \operatorname{dom} f$, then Lemma B.2 would imply uniqueness of the global minimizer. This phenomenon is consistent with the convex analysis of Gruntkowska et al. (2025), where the constant $c_t(x)$ becomes 0 when $u \in \mathcal{X}_f$, making the inner product sign irrelevant.

Under Assumption 3.1, unlike in the convex case, the broximal operator need not be single-valued outside $\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0, t)$. In such cases, we can define $\operatorname{brox}_f^t(x)$ to select a maximally distant point from x among the broximal minimizers. In particular, if $x \in \operatorname{brox}_f^t(x)$ but the set contains additional elements, the algorithm must choose a point in $\operatorname{brox}_f^t(x) \setminus \{x\}$ to ensure progress.

A more problematic case arises when the only broximal minimizer is x itself, i.e., $\operatorname{brox}_f^t(x) = \{x\}$. Under Assumption 3.1, this does not necessarily imply that $x \in \mathcal{X}_f$, and we will need an additional assumption to rule out this scenario. First, however, we examine what can be deduced from

Assumption 3.1 alone. Suppose that $u \in \operatorname{brox}_f^t(x) \not\subseteq \mathcal{X}_f$ is such that $f(u) = f(x)$ and $u \neq x$. By Lemma B.2, this implies $x \notin \operatorname{brox}_f^t(u)$. Since $\operatorname{brox}_f^t(u)$ is non-empty (Theorem B.1), there exists $\bar{u} \in \operatorname{brox}_f^t(u)$ such that $f(\bar{u}) < f(u)$ (the inequality is strict because otherwise we would have $x \in \operatorname{brox}_f^t(u)$). Consequently, $\bar{u} \notin \mathcal{B}_{\boldsymbol{X}}(x,t)$, meaning that $\|x - \bar{u}\|_{\boldsymbol{X}} > t$. Therefore, under Assumption 3.1, as long as the algorithm makes progress ($x_k \neq x_{k+1}$), a non-decreasing step $f(x_k) = f(x_{k+1})$ must be followed by a strict decrease $f(x_{k+1}) > f(x_{k+2})$, and then $\|x_k - x_{k+2}\|_{\boldsymbol{X}} > t$. This behavior is formalized in Lemma B.4.

The above argument relies on the condition $x_k \neq x_{k+1}$. To guarantee this unless a global minimizer has been reached, we introduce it explicitly as an assumption.

Assumption 3.2. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function, and let $t > 0$. Then

$$\operatorname{brox}_f^t(x) = \{x\} \quad \Longrightarrow \quad x \in \mathcal{X}_f \qquad \forall x \in \operatorname{dom} f.$$

In other words, if x is a strict minimizer of f over the ball $\mathcal{B}_{\boldsymbol{X}}(x,t)$, then x must already be globally optimal, i.e., there are no "bad" local minima detectable at the scale of radius t. Note that Assumption 3.2 still permits $x \in \operatorname{brox}_f^t(x)$ with $x \notin \mathcal{X}_f$, provided that there exists another point $\bar{x} \in \operatorname{brox}_f^t(x)$ such that $\bar{x} \neq x$. To simplify references, we introduce a notation for the family of functions that satisfy our new assumptions.

Definition 3.1 (Broximal Aligned Functions)**.** A proper, closed function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ belongs to the class $\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ if it satisfies Assumptions 3.1 and 3.2 for a radius $t > 0$ with respect to the geometry induced by the symmetric positive definite matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d}$.

Importantly, this class requires no convexity, differentiability, smoothness, or Lipschitz assumptions – only the geometric alignment condition of Assumption 3.1 and the absence of isolated local minima as specified by Assumption 3.2.

4 Convergence Theory

It turns out that Assumptions 3.1 and 3.2 are enough for global convergence guarantees. Building upon them, the main convergence result relies on the fact that every *second* iterate is separated by a distance of at least t.

Theorem 4.1. Assume that $f \in \mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ and let $\{x_k\}_{k \geq 0}$ be the iterates of BPM run with any $t > 0$, where $x_0 \in \operatorname{dom} f$. Then

(i) $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) \neq \emptyset$ if and only if x_{k+1} is optimal.

(ii) The distances $\{\|x_k - x_\star\|_{\boldsymbol{X}}\}_{k \geq 0}$ are non-increasing. In particular, if $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) = \emptyset$, we have

$$\|x_{k+1} - x_\star\|_{\boldsymbol{X}}^2 \leq \|x_{k-2} - x_\star\|_{\boldsymbol{X}}^2 - \frac{t^2}{3}.$$

(iii) If $\left\lfloor \frac{K}{3} \right\rfloor \geq \frac{3}{t^2} \|x_0 - x_\star\|_{\boldsymbol{X}}^2$, then $x_K \in \mathcal{X}_f$.

Unlike the non-convex optimization literature standard, the guarantee above is not stated in terms of stationary points, but rather in terms of distances from the minimizer. We show that after a sufficient number of iterations, the method finds the exact global minimum.

5 Unpacking the Broximal Geometry

So far, we have operated with the abstract function class $\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ without clarifying which objective functions belong to it. At first glance, Assumption 3.1 may seem cryptic, as it cannot be expressed using standard smoothness or curvature conditions. This section aims to unpack the structure behind our assumptions and explain the class of functions they define. We begin with simple examples that

Table 1: **Comparison of convex and non-convex function classes. Non-Convex** = does not require f to be convex; **Non-Differentiable** = does not require f to be differentiable; **Stationary pts** = allows existence of stationary points that are not global minima; **Non-unique min** = allows existence of multiple global minima; **Disconnected $\mathcal{X}_f$**: the set of minimizers need not be connected.

Function class	Non-Convex	Non-Differentiable	Stationary pts	Non-unique min	Disconnected $\mathcal{X}_f$
Strongly convex	✗	✓	✗	✗	✗
Convex	✗	✓	✗	✓	✗
Pseudoconvex (Mangasarian, 1975)	✓	✗	✗	✓	✗
Quasiconvex (Arrow & Enthoven, 1961)	✓	✓	✓	✓	✗
Star-convex (Nesterov & Polyak, 2006)	✓	✗	✗	✓	✗
Quasar convex (Hardt et al., 2018)	✓	✗	✗	✓	✗
Aiming (Liu et al., 2023)	✓	✗	✗	✓	✗
Polyak-Łojasiewicz (Polyak, 1963; Łojasiewicz, 1963)	✓	✗	✗	✗	✗
Quadratic growth (Bonnans & Ioffe, 1995)	✓	✓	✓	✓	✓
$\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ (ours)	✓	✓	✓	✓	✓

illustrate the kinds of functions covered by our framework. We then present a study of existing notions of generalized convexity, their characterizations, and their relationship to our assumptions (summarized in Table 1 and Figure 3). Before proceeding, we emphasize an important caveat: in principle, *any* function can be cast under our framework for sufficiently large t. Once the ball defining the broximal operator contains a global minimizer $x_\star$, the assumptions hold trivially. Of course, this regime is not of practical interest: as $t \to \infty$, the optimization subroutine in equation 2 becomes equivalent to global optimization, and thus is no easier than the original problem. For this reason, we focus on settings where $t < \infty$. Importantly, for *any* fixed $t > 0$, the standard non-convex function classes discussed below are more restrictive than $\mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$. In other words, our framework is strictly more general and can capture broader non-convex behavior. While the presence of $\boldsymbol{X}$ adds additional flexibility, we restrict attention to the standard Euclidean case $\boldsymbol{X} = \boldsymbol{I}$ for the remainder of this section, unless stated otherwise.

Toy Examples. We first present a simple illustrative example that satisfies Assumptions 3.1 and 3.2 while exhibiting a high degree of non-convexity.

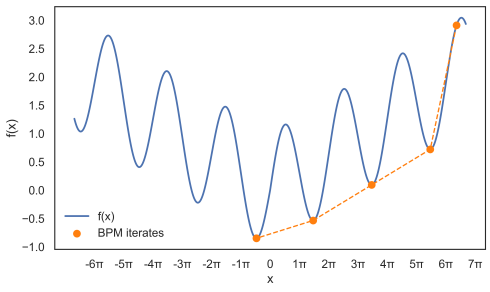

Figure 2: Trajectory of BPM with $t = 2\pi$ starting from $x_0 = 20$ applied to $f(x) = |x| + 10\sin(x)$ (Example 1).

Example 1. *Consider the 1D function $f(x) = |x| + 10\sin(x)$, which has infinitely many local minima spaced every 2π (Figure 2). Hence, for any fixed $t \geq 2\pi$, the ball $\mathcal{B}_{\boldsymbol{X}}(x, t)$ always contains at least two local minimizers whose function values differ. It is easy to check that Assumptions 3.1 and 3.2 hold for this radius, and so $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t \geq 2\pi$.*

The example shows that our setup does not preclude the existence of multiple local minima. This is in contrast to classical conditions, under which all stationary points are globally optimal (see Table 1). The function class we identify can exhibit multiple local minimizers and nontrivial geometry, while still permitting global convergence of BPM.

The next example shows that the set of minimizers $\mathcal{X}_f$ can be disconnected.

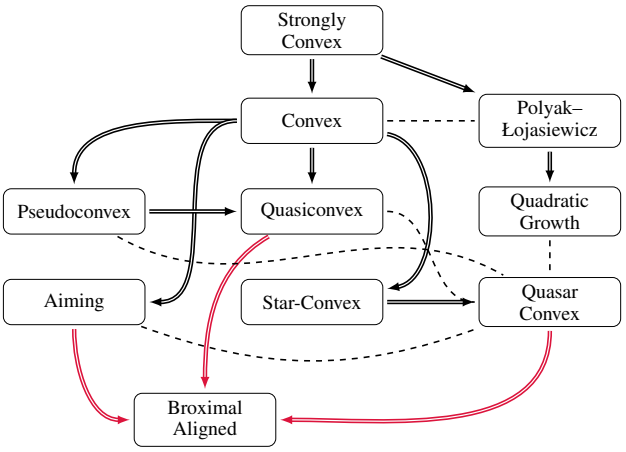

Figure 3: **Hierarchy of assumptions.** Solid arrows encode one-way implications, dashed lines denote mutual non-implications, and red arrows indicate conditions guaranteeing Broximal Alignment for any $t > 0$. For simplicity, we assume f is L-smooth (i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ $\forall x, y \in \mathbb{R}^d$) with domain $\operatorname{dom} f = \mathbb{R}^d$. See the appendix and Karimi et al. (2016); Hinder et al. (2019); Khanh & Lara (2025) for proofs.

Example 2. *Fix $a \in \mathbb{R}^d$ such that $a \neq 0$ and consider the function $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ defined by $f(x) = \|x\|^2$ if $x \neq a$ and $f(x) = 0$ if $x = a$. Then f has a disconnected set of minimizers $\mathcal{X}_f = \{0, a\}$, but one can show that $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ with any $t > 0$.*

The example above motivates the construction of a larger class of functions that satisfy Assumptions 3.1 and 3.2 (see Lemma D.1). We emphasize that both Examples 1 and 2 belong to $\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ for suitable choices of t, yet they fail to satisfy any of the classical conditions listed in Table 1.

Quasiconvexity. Quasiconvexity is one of the most classical and widely studied relaxations of convexity, dating back to early work by Arrow & Enthoven (1961). It plays an important role in optimization theory because it preserves several global properties of convex functions while allowing for richer geometric structure. As a result, quasiconvex optimization has found applications in many areas, including economics (Wolfstetter, 1999; Laffont & Martimort, 2002), control (Seiler & Balas, 2010), computational geometry (Eppstein, 2005), and computer vision (Ke & Kanade, 2007).

Definition 5.1 (Quasiconvexity)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *quasiconvex* if

$$f\left((1-\lambda)x + \lambda y\right) \leq \max\left\{f(x), f(y)\right\} \qquad \forall x, y \in \mathbb{R}^d, \lambda \in [0, 1].$$

If the inequality is strict for all $x \neq y$ and $\lambda \in (0, 1)$, then f is *strictly quasiconvex*.

Definition 5.1 can be viewed as a relaxation of convexity obtained by replacing the convex combination of function values by a maximum. A standard consequence is that all sublevel sets of a quasiconvex function are convex (Lemma D.2). The next result shows that strict quasiconvexity is sufficient for membership in our function class.

Theorem 5.1. Suppose that f is strictly quasiconvex. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

This inclusion holds for *every* radius $t > 0$, showing that $\mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ contains the class of strictly quasiconvex functions for any fixed t. At the same time, unlike functions in our class, quasiconvex functions remain restrictive for many modern non-convex objectives in ML. The requirement that every sublevel set be convex excludes functions with complex basin structures, multiple disconnected minima, or other nontrivial geometric features that arise in deep learning and related applications.

Pseudoconvexity. This first-order relaxation of convexity occupies the middle ground between convex and quasiconvex functions. Introduced in early work by Mangasarian (1975), it has since been studied extensively in nonlinear programming (Mangasarian, 1994; Bazaraa et al., 1993).

Definition 5.2 (Pseudoconvexity)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is *pseudoconvex* if

$$\langle \nabla f(x), y - x \rangle \geq 0 \quad \Longrightarrow \quad f(y) \geq f(x) \qquad \forall x, y \in \mathbb{R}^d.$$

Every convex differentiable function is pseudoconvex, but the converse does not hold. Nevertheless, pseudoconvexity still guarantees that every stationary point is a global minimizer, and a classical

result further shows that any differentiable pseudoconvex function is quasiconvex. This makes pseudoconvexity a natural setting in which to test our broximal alignment conditions.

Theorem 5.2. Suppose that f is pseudoconvex. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Although Definition 5.2 is stated for differentiable functions, in Section D.4.1 we extend the comparison to nondifferentiable generalizations of pseudoconvexity and show that our assumptions remain strictly weaker in that broader setting as well.

Quasar Convexity. Quasar convexity (Hardt et al., 2018; Hinder et al., 2019), like quasiconvexity and pseudoconvexity, generalizes the notion of unimodality to higher dimensions. Informally, quasar convex functions are unimodal along all lines passing through a global minimizer (Hinder et al., 2019). In dimensions greater than one, quasar convexity is neither implied by nor implies quasiconvexity or pseudoconvexity.

Definition 5.3 (Quasar convexity)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$ for some $\zeta \in (0, 1]$ if

$$f(x) - f_\star \leq \frac{1}{\zeta} \langle \nabla f(x), x - x_\star \rangle \qquad \forall x \in \mathbb{R}^d.$$

The class is parameterized by a constant $\zeta \in (0, 1]$ which controls the degree of non-convexity. As ζ becomes smaller, f becomes "more non-convex". Quasar convexity appears in several practical settings, such as learning linear dynamical systems (Hardt et al., 2018), a problem closely related to training recurrent neural networks. When $\zeta = 1$, the above condition is called *star convexity* (Nesterov & Polyak, 2006), another well-known relaxation of convexity that has been hypothesized to describe the loss geometry of neural networks in large neighborhoods around global minimizers (Kleinberg et al., 2018; Zhou et al., 2019). As the next result shows, quasar convex objectives also fall within the scope of our framework.

Theorem 5.3. Suppose that f is ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Aiming Condition. The *aiming condition* (Liu et al., 2023) links the gradient direction to the function value, while allowing the reference minimizer to vary with the current point. Following Liu et al. (2023), we assume throughout that $f_\star = 0$. We denote by $\Pi(\cdot, \mathcal{X})$ the projection onto a set $\mathcal{X}$ (i.e., $\Pi(x, \mathcal{X}) := \arg\min_{y \in \mathcal{X}} \|x - y\|_2^2$).

Definition 5.4 (Aiming Condition)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the *aiming condition* if $f_\star = 0$ and there exist $\bar{x} \in \Pi_{\mathcal{X}_f}(x)$ and $\theta > 0$ such that

$$\theta f(x) \leq \langle \nabla f(x), x - \bar{x} \rangle \qquad \forall x \in \mathbb{R}^d. \tag{4}$$

The aiming condition is closely related to quasar convexity: while quasar convexity requires equation 4 to hold with respect to a fixed minimizer $x_\star \in \mathcal{X}_f$, aiming permits the reference point $\bar{x}$ to depend on x. This yields a strictly larger class of functions, encompassing, for example, wide neural network losses (Liu et al., 2023). Even so, aiming remains stronger than our assumptions.

Theorem 5.4. Let f satisfy the aiming condition and have a unique global minimizer. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Here, we additionally require the minimum to be unique. However, Liu et al. (2023) do not rely on the aiming condition in isolation, but rather combine it with quadratic growth (Necoara et al., 2019), smoothness, and interpolation assumptions.

Operations that Preserve Assumptions 3.1 and 3.2. Lastly, we record several basic closure properties of Assumptions 3.1 and 3.2. In these results, we again work under the general conditions formulated in the geometry induced by $\boldsymbol{X}$. These properties clarify which transformations leave the geometric structure underlying our framework unchanged, and can be useful for constructing new admissible objectives from existing ones. We begin with monotone transformations of the objective.

Lemma 5.1. Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with a nonempty set of minimizers $\mathcal{X}_h$ and let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Define the function $f = g \circ h$ by

$$f(x) = \begin{cases} g(h(x)) & \text{if } h(x) \in \mathbb{R}, \\ +\infty & \text{if } h(x) = +\infty. \end{cases}$$

If h satisfies Assumptions 3.1 and 3.2 for some $t > 0$, then f also satisfies Assumptions 3.1 and 3.2 with the same t.

As a corollary of the result above, we note that simple affine transformations of the objective function also preserve the assumptions (see Corollary 1). Next, we consider changes of coordinates, showing that our assumptions are invariant under transformations that preserve the geometry induced by $\boldsymbol{X}$.

Lemma 5.2 ($\boldsymbol{X}$–orthogonal affine transformations)**.** Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with nonempty set of minimizers $\mathcal{X}_h \neq \emptyset$. Suppose that h satisfies Assumptions 3.1 and 3.2 for some $t > 0$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be defined by $\phi(x) = \boldsymbol{Q}x + b$, where $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ is invertible and satisfies $\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q} = \boldsymbol{X}$. Define $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by

$$f(x) = \begin{cases} h(\phi(x)) & x \in \phi^{-1}(\operatorname{dom} h), \\ +\infty & \text{otherwise.} \end{cases}$$

Then f is proper, has nonempty minimizer set $\mathcal{X}_f = \phi^{-1}(\mathcal{X}_h)$, and satisfies Assumptions 3.1 and 3.2 with the same radius t.

6 Practical Considerations

Our theoretical developments assume access to the broximal oracle, that is, the ability to exactly minimize f over balls of the form $\mathcal{B}_{\boldsymbol{X}}(x, t)$. For many objectives of practical interest, particularly in modern ML, such subproblems are computationally intractable. Nevertheless, the framework remains relevant in practice through approximation strategies. The first approach is to *approximately solve the broximal subproblem* itself, e.g., one may consider sampling-based approaches or apply an iterative optimization method to minimize the original objective f over the constraint set $\mathcal{B}_{\boldsymbol{X}}(x_k, t_k)$ up to a prescribed accuracy. This strategy preserves the structure of the original problem, at the cost of additional inner-loop computation. The second approach is to *replace the original objective by a simpler local model*. Typical choices include first-order or quadratic approximations, such as linearizing f around the current iterate and solving the resulting trust-region subproblem. Such model-based subproblems are often tractable and, in certain cases, admit closed-form solutions. The effectiveness of these approximations is not limited to Euclidean settings. On the contrary, their benefits often become more pronounced in non-Euclidean geometries, where the norm $\|\cdot\|_{\boldsymbol{X}}$ is replaced by a certain operator norm. This viewpoint underlies several modern optimization methods, e.g., the Muon optimizer (Jordan et al., 2024) can be interpreted as performing approximate broximal steps by minimizing linearized local models over non-Euclidean balls induced by the spectral norm. We refer to Gruntkowska & Richtárik (2025) for a more detailed discussion and further connections.

7 Summary and Outlook

This paper introduces a new analytic tool for global non-convex optimization that avoids the limitations of stationarity-based theory. Revisiting the Ball Proximal Point Method (BPM), we identify a novel geometric condition—Broximal Alignment—that guarantees global convergence without any convexity, smoothness, or Lipschitz assumptions. The resulting function class, $\mathcal{F}^{\mathrm{BA}}_{\boldsymbol{X}}(t)$, is unusually broad: it allows disconnected minimizer sets, multiple local minima, and highly non-convex landscapes, yet still ensures that BPM reaches a global minimizer in finitely many iterations. Our theory is strong in two ways. First, it provides global optimality guarantees in settings where gradient-based methods do not certify meaningful progress. Second, it strictly extends several classical non-convex frameworks, showing that these are all special cases of our geometric alignment principle. This positions our work as a conceptual foundation for global non-convex optimization. At the same time, the framework is intentionally conceptual, as the broximal oracle is generally intractable. This opens a rich direction for future work: developing practical approximations of broximal steps to bridge the gap between theory and modern ML practice.

AI USE STATEMENT

In this work, we used generative AI tools for refining the prose. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, assisting with translation, and interpreting results. Generating synthetic datasets, cleaning and reformatting datasets, implementing methods, and supporting qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Ali Jadbabaie, and Suvrit Sra. How to escape sharp minima with random perturbations. *arXiv preprint arXiv:2305.15659*, 2023. (Cited on page 2)

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pp. 242–252. PMLR, 2019. (Cited on page 4)

Kenneth J. Arrow and Alain C. Enthoven. Quasi-concave programming. *Econometrica*, 29(4): 779–800, 1961. (Cited on page 4, 6, and 7)

Hilal Asi, Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Stochastic bias-reduced gradient methods. *Advances in Neural Information Processing Systems*, 34:10810–10822, 2021. (Cited on page 2)

Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pp. 521–530. PMLR, 2018. (Cited on page 4)

Mokhtar S. Bazaraa, Hanif D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, Ltd, 1993. (Cited on page 7)

Joseph Frédéric Bonnans and Alexander Ioffe. Second-order sufficiency and quadratic growth for nonisolated minima. *Mathematics of Operations Research*, 20(4):801–817, 1995. (Cited on page 6)

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33:19052–19063, 2020. (Cited on page 2)

Yair Carmon, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Thinking inside the ball: Near-optimal minimization of the maximal loss. In *Conference on Learning Theory*, pp. 866–882. PMLR, 2021. (Cited on page 2)

B. Craven and B. Glover. Invex functions and duality. *Journal of the Australian Mathematical Society*, 39:1 – 20, 08 1985. (Cited on page 4)

Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. In *High-Dimensional Optimization and Probability: With a View Towards Data Science*, pp. 79–163. Springer, 2022. (Cited on page 4)

Aaron Defazio. Why gradients rapidly increase near the end of training. *arXiv preprint arXiv:2506.02285*, 2025. (Cited on page 2)

David Eppstein. Quasiconvex programming. *Combinatorial and Computational Geometry*, 52(287-331):3, 2005. (Cited on page 7)

Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks: (almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017. (Cited on page 3)

Massimo Fornasier and Lukang Sun. Regularity and positivity of solutions of the consensus-based optimization equation: unconditional global convergence. *arXiv preprint arXiv:2502.01434*, 2025. (Cited on page 3)

Massimo Fornasier, Hui Huang, Lorenzo Pareschi, and Philippe Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Mathematical Models and Methods in Applied Sciences*, 30(14):2725–2751, 2020. (Cited on page 3)

Massimo Fornasier, Timo Klock, and Konstantin Riedl. Consensus-based optimization methods converge globally. *SIAM Journal on Optimization*, 34(3):2973–3004, 2024. doi: 10.1137/22M1527805. (Cited on page 3)

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016. (Cited on page 4)

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017. (Cited on page 4)

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. (Cited on page 1)

Kaja Gruntkowska and Peter Richtárik. Non-Euclidean Broximal Point Method: A blueprint for geometry-aware optimization. *arXiv preprint arXiv:2510.00823*, 2025. (Cited on page 2, 9, and 15)

Kaja Gruntkowska, Hanmin Li, Aadi Rane, and Peter Richtárik. The Ball-Proximal (="Broximal") Point Method: a new algorithm, convergence theory, and applications. *arXiv preprint arXiv:2502.02002*, 2025. (Cited on page 1, 2, 3, 4, and 15)

Benjamin D. Haeffele and René Vidal. Global optimality in neural network training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4390–4398, 2017. doi: 10.1109/CVPR.2017.467. (Cited on page 3)

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018. (Cited on page 4, 6, and 8)

Oliver Hinder, Aaron Sidford, and Nimit S Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *arXiv preprint arXiv:1906.11985*, 2019. (Cited on page 4, 7, 8, and 23)

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *arXiv preprint arXiv:1712.07897*, 2017. (Cited on page 1)

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International conference on machine learning*, pp. 1724–1732. PMLR, 2017. (Cited on page 2)

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *Journal of the ACM (JACM)*, 68(2):1–29, 2021. (Cited on page 1 and 2)

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. (Cited on page 9)

Fatih Kangal, Karl Meerbergen, Emre Mengi, and Wim Michiels. A subspace method for large-scale eigenvalue optimization. *SIAM Journal on Matrix Analysis and Applications*, 39(1):48–82, 2018. doi: 10.1137/16M1070025. (Cited on page 4)

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016. (Cited on page 7)

Qifa Ke and Takeo Kanade. Quasiconvex optimization for robust geometric reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1834–1847, 2007. doi: 10.1109/TPAMI.2007.1083. (Cited on page 7)

Phan Quoc Khanh and Felipe Lara. Star quasiconvexity: an unified approach for linear convergence of first-order methods beyond convexity. *arXiv preprint arXiv:2510.24981*, 2025. (Cited on page 7)

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018. (Cited on page 2 and 8)

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263. (Cited on page 1)

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025. (Cited on page 15)

Jean-Jacques Laffont and David Martimort. *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, 2002. ISBN 9780691091846. (Cited on page 7)

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539. (Cited on page 1)

Chaoyue Liu, Dmitriy Drusvyatskiy, Misha Belkin, Damek Davis, and Yian Ma. Aiming towards the minimizers: fast convergence of SGD for overparametrized problems. *Advances in neural information processing systems*, 36:60748–60767, 2023. (Cited on page 4, 6, and 8)

Stanisław Łojasiewicz. A topological property of real analytic subsets (in French). *Coll. du CNRS, Les équations, aux dérivées partielles*, pp. 87–89, 1963. (Cited on page 4 and 6)

Olvi L. Mangasarian. Pseudo-convex functions. In W.T. Ziemba and R.G. Vickson (eds.), *Stochastic Optimization Models in Finance*, pp. 23–32. Academic Press, 1975. ISBN 978-0-12-780850-5. (Cited on page 4, 6, and 7)

Olvi L. Mangasarian. *Nonlinear Programming: Theory and Algorithms*. SIAM, 1994. (Cited on page 7)

Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit S. Dhillon. Rank minimization via online learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML 2008, pp. 656–663, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390239. (Cited on page 1)

Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987. (Cited on page 1)

Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical programming*, 175(1):69–107, 2019. (Cited on page 8)

A.S. Nemirovski and Y.E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985. ISSN 0041-5553. (Cited on page 1)

A.S. Nemirovski and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983. (Cited on page 1)

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003. (Cited on page 1)

Yurii Nesterov and Boris Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006. (Cited on page 4, 6, and 8)

Razvan Pascanu, Yann N Dauphin, Surya Ganguli, and Yoshua Bengio. On the saddle point problem for non-convex optimization. *arXiv preprint arXiv:1405.4604*, 2014. (Cited on page 1 and 2)

A. L. Patterson. A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.*, 46:372–376, Sep 1934. doi: 10.1103/PhysRev.46.372. (Cited on page 1)

Boris Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3:864–878, 12 1963. (Cited on page 4 and 6)

R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976. (Cited on page 2)

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pp. 4433–4441. PMLR, 2018. (Cited on page 1)

Peter Seiler and Gary J. Balas. Quasiconvex sum-of-squares programming. In *49th IEEE Conference on Decision and Control (CDC)*, pp. 3337–3342, 2010. doi: 10.1109/CDC.2010.5717672. (Cited on page 7)

Yeonjong Shin. Effects of depth, width, and initialization: A convergence analysis of layer-wise training for deep linear neural networks. *Analysis and Applications*, 20(01):73–119, 2022. (Cited on page 4)

Samuel L. Smith and Quoc V. Le. A Bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017. (Cited on page 2)

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018. (Cited on page 3)

Elmar Wolfstetter. *Topics in Microeconomics: Industrial Organization, Auctions, and Incentives*. Cambridge University Press, 1999. (Cited on page 7)

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019. (Cited on page 8)

Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Stephen Boyd, and Peter W Glynn. Stochastic mirror descent in variationally coherent optimization problems. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. (Cited on page 4)

A Appendix

Contents

1 Introduction **1**

2 Contributions **3**

3 Optimization Beyond Convexity **3**

4 Convergence Theory **5**

5 Unpacking the Broximal Geometry **5**

6 Practical Considerations **9**

7 Summary and Outlook **9**

A Appendix **14**

B Facts and Lemmas **15**

C Proof of Theorem 4.1 **17**

D Proofs for Section 5 **18**

D.1 Examples 18

D.2 Operations that Preserve Assumptions 3.1 and 3.2 19

D.3 Quasiconvexity 21

D.4 Pseudoconvexity 22

D.4.1 Generalized Pseudoconvexity 22

D.5 Quasar convexity 23

D.6 Aiming Condition 24

E Monotonicity of Function Classes **26**

E.1 Uniform Broximal Alignment 27

F Notation **29**

B FACTS AND LEMMAS

We first collect key definitions and present some fundamental results needed for the proofs in the remainder of the paper. The presentation and notation largely follows that of Gruntkowska & Richtárik (2025).

For any nonempty, closed, and convex set $\mathcal{X} \subseteq \mathbb{R}^d$, we define its *indicator function* $\delta_{\mathcal{X}} : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by

$$\delta_{\mathcal{X}}(z) := \begin{cases} 0 & \text{if } z \in \mathcal{X}, \\ +\infty & \text{otherwise.} \end{cases}$$

This function is proper, closed, and convex.

Definition B.1 (Subdifferential)**.** Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be proper and let $x \in \operatorname{dom}(f)$. The *subdifferential* of f at x, denoted as $\partial f(x)$, is the set of vectors $g \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + \langle g, y - x \rangle \qquad \forall y \in \mathbb{R}^d.$$

The elements of $\partial f(x)$ are called the *subgradients* of f at x.

Fact 1. *The subdifferential of the indicator function of a ball $\mathcal{B}_{\boldsymbol{X}}(x,t)$ is*

$$\partial \delta_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(y) = \mathcal{N}_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(y) = \begin{cases} \mathbb{R}_{\geq 0}\boldsymbol{X}(y-x) & \|x-y\|_{\boldsymbol{X}} = t, \\ \{0\} & \|x-y\|_{\boldsymbol{X}} < t, \\ \emptyset & \|x-y\|_{\boldsymbol{X}} > t. \end{cases}$$

where $\mathcal{N}_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(y)$ is the normal cone of $\mathcal{B}_{\boldsymbol{X}}(x,t)$ at y.

Lemma B.1. Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be differentiable. Fix $x \in \mathbb{R}^d$ and let $u \in \arg\min_{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)} f(z)$. Then $-\nabla f(u) \in \mathcal{N}_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(u)$.

Proof. By the argument in the proof of Lemma 3 in Kovalev (2025), we have $-\nabla f(u) \in \partial \delta_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(u)$. The result then follows directly from Fact 1. □

The following is an analogue of the First Brox Theorem (Gruntkowska et al., 2025, Theorem D.1). Note that unlike in the convex case considered by Gruntkowska et al. (2025), when convexity assumption is dropped, $\operatorname{brox}_f^t(x)$ need not be a singleton lying on the boundary of $\mathcal{B}_{\boldsymbol{X}}(x,t)$ when $\mathcal{B}_{\boldsymbol{X}}(x,t) \cap \mathcal{X}_f = \emptyset$ (see Figure 2).

Theorem B.1. Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be proper and closed and choose $x \in \operatorname{dom} f$. Then $\operatorname{brox}_f^t(x) \neq \emptyset$. Moreover, if $\mathcal{B}_{\boldsymbol{X}}(x,t) \cap \mathcal{X}_f \neq \emptyset$, then $\operatorname{brox}_f^t(x)$ is a nonempty subset of $\mathcal{X}_f$.

Proof. The broximal operator is minimizing a proper, closed function over a closed set $\mathcal{B}_{\boldsymbol{X}}(x,t)$. Since such functions attain their minimum over compact sets, it follows that $\operatorname{brox}_f^t(x) \neq \emptyset$. Moreover, if $\mathcal{B}_{\boldsymbol{X}}(x,t) \cap \mathcal{X}_f \neq \emptyset$, then $\operatorname{brox}_f^t(x) \subseteq \mathcal{X}_f$ is nonempty. □

The next lemma shows that under Assumption 3.1, BPM cannot get stuck away from the global minimum and keep "jumping" between two points.

Lemma B.2. Let Assumption 3.1 hold and suppose that $u_1 \in \operatorname{brox}_f^t(u_2)$ and $u_2 \in \operatorname{brox}_f^t(u_1)$. If $u_1, u_2 \notin \mathcal{X}_f$, then $u_1 = u_2$.

Proof. Suppose that $u_1 \in \operatorname{brox}_f^t(u_2)$, $u_2 \in \operatorname{brox}_f^t(u_1)$ and $u_1, u_2 \notin \mathcal{X}_f$. Since $u_1 \in \arg\min_{z \in \mathcal{B}_{\boldsymbol{X}}(u_2,t)} f(z)$ and $u_1 \notin \mathcal{X}_f$, we have $\operatorname{dist}(u_2, \mathcal{X}_f) > t$, and likewise $\operatorname{dist}(u_1, \mathcal{X}_f) > t$. Therefore, $u_1, u_2 \notin \mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t)$. Then, by Assumption 3.1, we have

$$\langle u_2 - u_1, u_1 - x_\star \rangle_{\boldsymbol{X}} \geq 0,$$

$$\langle u_1 - u_2, u_2 - x_\star \rangle_{\boldsymbol{X}} \geq 0.$$

Adding the two inequalities gives

$$0 \leq \langle u_1 - u_2, u_2 - x_\star - u_1 + x_\star \rangle_{\boldsymbol{X}} = - \|u_1 - u_2\|_{\boldsymbol{X}}^2 ,$$

and hence $u_1 = u_2$. □

The result below is a direct consequence of Assumption 3.2.

Lemma B.3. Let Assumption 3.2 hold and let $\{x_k\}_{k\geq 0}$ be the iterates of BPM. If $x_k \notin \mathcal{X}_f$, then $x_k \neq x_{k+1}$.

Proof. Since $x_k \notin \mathcal{X}_f$, Assumption 3.2 says that $\operatorname{brox}_f^t(x_k) \neq \{x_k\}$, and hence the set $\operatorname{brox}_f^t(x_k)$ has cardinality greater than 1. By definition of the algorithm, it follows that $x_k \neq x_{k+1}$. □

Remark 1. *Note that by the reasoning above, under Assumptions 3.1 and 3.2, we know that the algorithm has already reached $\mathcal{X}_f$ when the function value does not decrease for two consecutive iterations.*

Lemma B.4. Let Assumptions 3.1 and 3.2 hold. Then, for any iteration k such that $x_{k+1} \notin \mathcal{X}_f$, the iterates of BPM satisfy either

- $\|x_k - x_{k+2}\|_{\boldsymbol{X}} > t$, or
- $\|x_k - x_{k+3}\|_{\boldsymbol{X}} > t$ and $\|x_{k+1} - x_{k+3}\|_{\boldsymbol{X}} > t$.

Therefore, $\|x_k - x_{k+3}\|_{\boldsymbol{X}} > t$ for any such k.

Proof. Since $x_{k+1} \notin \mathcal{X}_f$, we have $x_k \notin \mathcal{X}_f$, so from Lemma B.3 we know that $x_k \neq x_{k+1}$ and $x_{k+1} \neq x_{k+2}$. By design, for any j we have $f(x_j) \geq f(x_{j+1})$. Now, let us consider two cases:

Case 1: $f(x_k) = f(x_{k+1})$**.** By Lemma B.2, $x_k \notin \operatorname{brox}_f^t(x_{k+1})$. Since $\operatorname{brox}_f^t(x_{k+1})$ is non-empty (Theorem B.1), there exists $u \in \operatorname{brox}_f^t(x_{k+1})$ such that $f(u) < f(x_{k+1})$ (the inequality is strict because otherwise we would have $x_k \in \operatorname{brox}_f^t(x_{k+1})$). Consequently, $u \notin B(x_k, t)$ for any $u \in \operatorname{brox}_f^t(x_{k+1})$, meaning that $\|x_k - x_{k+2}\|_{\boldsymbol{X}} > t$. Since $f(x_{k+3}) \leq f(x_{k+2}) < f(x_k)$, clearly $\|x_k - x_{k+3}\|_{\boldsymbol{X}} > t$.

Case 2: $f(x_k) > f(x_{k+1})$**.** If $\|x_k - x_{k+2}\|_{\boldsymbol{X}} > t$, then the claim holds. Now, suppose that $\|x_k - x_{k+2}\|_{\boldsymbol{X}} \leq t$. Then $x_{k+1}, x_{k+2} \in \operatorname{brox}_f^t(x_k)$, which forces $f(x_{k+1}) = f(x_{k+2})$. As in Case 1, a plateau in function values must be followed by a strict decrease, so $f(x_{k+3}) < f(x_{k+2})$. This in turn means that $x_{k+3} \notin \operatorname{brox}_f^t(x_k)$ and $x_{k+3} \notin \operatorname{brox}_f^t(x_{k+1})$, which completes the proof. □

C Proof of Theorem 4.1

Theorem 4.1. Assume that $f \in \mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$ and let $\{x_k\}_{k\geq 0}$ be the iterates of BPM run with any $t > 0$, where $x_0 \in \operatorname{dom} f$. Then

(i) $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) \neq \emptyset$ if and only if x_{k+1} is optimal.

(ii) The distances $\{\|x_k - x_\star\|_{\boldsymbol{X}}\}_{k\geq 0}$ are non-increasing. In particular, if $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) = \emptyset$, we have

$$\|x_{k+1} - x_\star\|_{\boldsymbol{X}}^2 \leq \|x_{k-2} - x_\star\|_{\boldsymbol{X}}^2 - \frac{t^2}{3}.$$

(iii) If $\left\lfloor \frac{K}{3} \right\rfloor \geq \frac{3}{t^2} \|x_0 - x_\star\|_{\boldsymbol{X}}^2$, then $x_K \in \mathcal{X}_f$.

Proof. (i) That $x_{k+1} \in \mathcal{X}_f$ if $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) \neq \emptyset$ is obvious from the definition of broximal operator. Similarly, if $\mathcal{X}_f \cap \mathcal{B}_{\boldsymbol{X}}(x_k, t) = \emptyset$, then x_{k+1} cannot belong to $\mathcal{X}_f$.

(ii) We start with the simple decomposition

$$\|x_{k+1} - x_\star\|_{\boldsymbol{X}}^2 = \|x_k - x_\star\|_{\boldsymbol{X}}^2 - 2\langle x_k - x_{k+1}, x_{k+1} - x_\star\rangle_{\boldsymbol{X}} - \|x_{k+1} - x_k\|_{\boldsymbol{X}}^2.$$

Since $x_k \notin \mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0, t)$, applying Assumption 3.1, we get

$$\begin{aligned} \|x_{k+1} - x_\star\|_{\boldsymbol{X}}^2 &\leq \|x_k - x_\star\|_{\boldsymbol{X}}^2 - \|x_{k+1} - x_k\|_{\boldsymbol{X}}^2 \\ &< \|x_k - x_\star\|_{\boldsymbol{X}}^2, \end{aligned} \tag{5}$$

so the distances $\{\|x_k - x_\star\|_{\boldsymbol{X}}\}_{k\geq 0}$ are decreasing. Now, from Lemma B.4 and triangle inequality, it follows that

$$t < \|x_{k-2} - x_{k+1}\|_{\boldsymbol{X}} \leq \|x_{k-2} - x_{k-1}\|_{\boldsymbol{X}} + \|x_{k-1} - x_k\|_{\boldsymbol{X}} + \|x_k - x_{k+1}\|_{\boldsymbol{X}},$$

and the AM-QM inequality gives

$$\begin{aligned} &\sqrt{\frac{\|x_{k-2} - x_{k-1}\|_{\boldsymbol{X}}^2 + \|x_{k-1} - x_k\|_{\boldsymbol{X}}^2 + \|x_k - x_{k+1}\|_{\boldsymbol{X}}^2}{3}} \\ &\quad\geq \frac{\|x_{k-2} - x_{k-1}\|_{\boldsymbol{X}} + \|x_{k-1} - x_k\|_{\boldsymbol{X}} + \|x_k - x_{k+1}\|_{\boldsymbol{X}}}{3} > \frac{t}{3}. \end{aligned}$$

This in turn implies that $\|x_{k-2} - x_{k-1}\|_{\boldsymbol{X}}^2 + \|x_{k-1} - x_k\|_{\boldsymbol{X}}^2 + \|x_k - x_{k+1}\|_{\boldsymbol{X}}^2 > \frac{t^2}{3}$. Hence, applying equation 5, we get

$$\begin{aligned} &\|x_{k+1} - x_\star\|_{\boldsymbol{X}}^2 \qquad (6) \\ &\leq \|x_{k-1} - x_\star\|_{\boldsymbol{X}}^2 - \left(\|x_k - x_{k-1}\|_{\boldsymbol{X}}^2 + \|x_{k+1} - x_k\|_{\boldsymbol{X}}^2\right) \\ &\leq \|x_{k-2} - x_\star\|_{\boldsymbol{X}}^2 - \left(\|x_{k-1} - x_{k-2}\|_{\boldsymbol{X}}^2 + \|x_k - x_{k-1}\|_{\boldsymbol{X}}^2 + \|x_{k+1} - x_k\|_{\boldsymbol{X}}^2\right) \\ &< \|x_{k-2} - x_\star\|_{\boldsymbol{X}}^2 - \frac{t^2}{3}. \qquad (7) \end{aligned}$$

(iii) Applying equation 6 iteratively, we get

$$\begin{aligned} \|x_K - x_\star\|_{\boldsymbol{X}}^2 &\leq \|x_{K-3} - x_\star\|_{\boldsymbol{X}}^2 - \frac{t^2}{3} \\ &\leq \left\|x_{(K \bmod 3)} - x_\star\right\|_{\boldsymbol{X}}^2 - \left\lfloor \frac{K}{3} \right\rfloor \frac{t^2}{3} \\ &\overset{\text{equation 5}}{<} \|x_0 - x_\star\|_{\boldsymbol{X}}^2 - \left\lfloor \frac{K}{3} \right\rfloor \frac{t^2}{3}, \end{aligned}$$

and the conclusion follows. □

D PROOFS FOR SECTION 5

In this section, we prove the results from Section 5 identifying the classes of potentially non-convex functions from the optimization literature under which Assumptions 3.1 and 3.2 hold. We begin by elaborating on the examples from Section 5 (Section D.1), then move on to presenting transformations of the loss under which our assumptions are preserved (Section D.2), and finish off with providing proofs of the implications presented in Section 5, connecting Assumptions 3.1 and 3.2 to existing literature (Sections D.3, D.5 and D.6).

Unless otherwise stated, we specialize to $\boldsymbol{X} = \boldsymbol{I}$; the exception is Section D.2, where the general $\boldsymbol{X}$ setting is used.

D.1 EXAMPLES

We first provide the proofs for the results in Section 5.

Example 2. *Fix $a \in \mathbb{R}^d$ such that $a \neq 0$ and consider the function $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ defined by $f(x) = \|x\|^2$ if $x \neq a$ and $f(x) = 0$ if $x = a$. Then f has a disconnected set of minimizers $\mathcal{X}_f = \{0, a\}$, but one can show that $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ with any $t > 0$.*

Indeed, fix $t > 0$ and take any $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}(0, t)\}$. Then $f(z) = \|z\|^2$ for all $z \in \mathcal{B}(x, t)$, and the broximal operator in this case will return $u = x - t\frac{x}{\|x\|}$. Therefore, choosing $x_\star = 0$,

$$\langle x - u, u - x_\star \rangle = \left\langle t\frac{x}{\|x\|}, x - t\frac{x}{\|x\|} \right\rangle = t(\|x\| - t) > 0,$$

so Assumption 3.1 holds.

Lastly, if $\operatorname{brox}_f^t(x) = \{x\}$, then x must be a local minimizer of f within the ball $\mathcal{B}(x, t)$. The only points that could satisfy this are the global minimizers 0 and a, and hence

$$\operatorname{brox}_f^t(x) = \{x\} \implies x \in \{0, a\} = \mathcal{X}_f.$$

The example above motivates the construction of a larger class of functions that satisfy Assumptions 3.1 and 3.2.

Lemma D.1. Let $g : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with a nonempty set of minimizers $\mathcal{X}_g$. Suppose further that g satisfies Assumptions 3.1 and 3.2 with some radius $t > 0$. Let $\mathcal{C} \subseteq \operatorname{dom} g$ be any closed set such that

$$\mathcal{C} \cap \mathcal{X}_g = \emptyset$$

and define $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ by

$$f(x) = \begin{cases} g_{\inf} & x \in \mathcal{C}, \\ g(x) & \text{otherwise}, \end{cases}$$

where $g_{\inf}$ is the minimum of g. Then the following statements hold:

1. f is proper and closed, and its set of global minimizers is $\mathcal{X}_f = \mathcal{X}_g \cup \mathcal{C}$.
2. f satisfies Assumption 3.1 with the same radius t and the same choice of $x_\star$.
3. f satisfies Assumption 3.2 with the same radius t.

Proof. We prove each claim in turn.

1. Since g is proper, we have $\operatorname{dom} g \neq \emptyset$ and $g > -\infty$ everywhere. By construction, $f(x) \geq g_{\inf} > -\infty$ for all x, and

   $$\operatorname{dom} f = \operatorname{dom} g \neq \emptyset.$$

   Hence f is proper. We next verify that f is closed, i.e., lower semicontinuous.

Case 1: $x \notin \mathcal{C}$**.** Since $\mathcal{C}$ is closed, there exists $\varepsilon > 0$ such that $\mathcal{B}(x, \varepsilon) \cap \mathcal{C} = \emptyset$. Therefore, for any sequence $\{x_k\}$ with $x_k \to x$, we have $x_k \notin \mathcal{C}$ for all sufficiently large k, and hence $f(x_k) = g(x_k)$. Since g is lower semicontinuous at x, it follows that f is lower-semicontinuous at x as well.

Case 2: $x \in \mathcal{C}$**.** Then $f(x) = g_{\inf}$. By construction, $f \geq g \geq g_{\inf}$ everywhere, so for any sequence $x_k \to x$,

$$\liminf_{k\to\infty} f(x_k) \geq g_{\inf} = f(x).$$

Combining the two cases, f is lower semicontinuous and hence closed. Lastly, $\mathcal{X}_f = \mathcal{X}_g \cup \mathcal{C}$ by construction.

2. Let $x \in \operatorname{dom} f$ satisfy $\operatorname{dist}(x, \mathcal{X}_f) > t$, and let $u \in \operatorname{brox}_f^t(x)$. Then the ball $\mathcal{B}(x, t)$ does not intersect $\mathcal{X}_f$, and in particular does not intersect $\mathcal{C}$. Consequently, $f = g$ on $\mathcal{B}(x, t)$, which implies

$$\operatorname{brox}_f^t(x) = \operatorname{brox}_g^t(x).$$

Since $\mathcal{X}_g \subseteq \mathcal{X}_f$, we also have $\operatorname{dist}(x, \mathcal{X}_g) > t$, so Assumption 3.1 for g applies. Therefore, there exists $x_\star \in \mathcal{X}_g$ such that

$$\langle x - u, u - x_\star \rangle \geq 0.$$

This establishes Assumption 3.1 for f with the same $x_\star$.

3. Assume that $\operatorname{brox}_f^t(x) = \{x\}$ for some $x \in \operatorname{dom} f$. If $x \in \mathcal{C}$, then $x \in \mathcal{X}_f$ and the claim holds. Now, suppose $x \notin \mathcal{C}$. If $\mathcal{B}(x, t) \cap \mathcal{C} \neq \emptyset$, choose $z \in \mathcal{B}(x, t) \cap \mathcal{C}$. Then

$$f(z) = g_{\inf} \leq f(x).$$

Since x is the unique minimizer of f over $\mathcal{B}(x, t)$, this forces $z = x$, contradicting $x \notin \mathcal{C}$. Hence $\mathcal{B}(x, t) \cap \mathcal{C} = \emptyset$. Therefore, $f = g$ on $\mathcal{B}(x, t)$ and

$$\operatorname{brox}_f^t(x) = \operatorname{brox}_g^t(x) = \{x\}.$$

By Assumption 3.2 for g, we conclude that $x \in \mathcal{X}_g \subseteq \mathcal{X}_f$. Thus Assumption 3.2 also holds for f.

□

D.2 Operations that Preserve Assumptions 3.1 and 3.2

The next three lemmas apply in the general $\boldsymbol{X}$ setting.

We first show that strictly increasing transformations preserve Assumptions 3.1 and 3.2.

Lemma 5.1. Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with a nonempty set of minimizers $\mathcal{X}_h$ and let $g : \mathbb{R} \to \mathbb{R}$ be strictly increasing. Define the function $f = g \circ h$ by

$$f(x) = \begin{cases} g(h(x)) & \text{if } h(x) \in \mathbb{R}, \\ +\infty & \text{if } h(x) = +\infty. \end{cases}$$

If h satisfies Assumptions 3.1 and 3.2 for some $t > 0$, then f also satisfies Assumptions 3.1 and 3.2 with the same t.

Proof. We first note that $\operatorname{dom} f = \operatorname{dom} h$. Since h is proper, f is also proper.

Now, we claim that $\mathcal{X}_h = \mathcal{X}_f$. Let $x_\star \in \mathcal{X}_h$. Then for all $x \in \operatorname{dom} h = \operatorname{dom} f$, we have $h(x_\star) \leq h(x)$. Since g is strictly increasing, this implies

$$f(x_\star) = g(h(x_\star)) \leq g(h(x)) = f(x),$$

and hence $x_\star \in \mathcal{X}_f$. Conversely, suppose $x_\star \in \mathcal{X}_f$. Then for all $x \in \operatorname{dom} f$,

$$g(h(x_\star)) \leq g(h(x)).$$

Strict monotonicity of g implies $h(x_\star) \le h(x)$, and thus $x_\star \in \mathcal{X}_h$. Therefore, $\mathcal{X}_f = \mathcal{X}_h$.

Now, fix any $x \in \operatorname{dom} h$ and define

$$h_x := h + \delta_{\mathcal{B}_{\boldsymbol{X}}(x,t)}, \qquad f_x := g \circ h_x = f + \delta_{\mathcal{B}_{\boldsymbol{X}}(x,t)}.$$

Applying the argument above to h_x and f_x and using the definition of the broximal mapping yields

$$\operatorname{brox}_h^t(x) = \mathcal{X}_{h_x} = \mathcal{X}_{f_x} = \operatorname{brox}_f^t(x).$$

Since $\operatorname{brox}_f^t(x) = \operatorname{brox}_h^t(x)$ for all x, and since $\mathcal{X}_f = \mathcal{X}_h$, Assumption 3.1 for f follows immediately from Assumption 3.1 for h with the same choice of $x_\star$.

Lastly, if $\operatorname{brox}_f^t(x) = \{x\}$, then the equality above implies

$$\operatorname{brox}_h^t(x) = \{x\}.$$

By Assumption 3.2 for h, we conclude that $x \in \mathcal{X}_h = \mathcal{X}_f$. Hence Assumption 3.2 holds for f, so Assumption 3.2 holds for f as well. □

We next show that Assumptions 3.1 and 3.2 are invariant under Euclidean isometries.

Lemma 5.2 (**$\boldsymbol{X}$**–orthogonal affine transformations)**.** Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with nonempty set of minimizers $\mathcal{X}_h \neq \emptyset$. Suppose that h satisfies Assumptions 3.1 and 3.2 for some $t > 0$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be defined by $\phi(x) = \boldsymbol{Q}x + b$, where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ is invertible and satisfies $\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q} = \boldsymbol{X}$. Define $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ by

$$f(x) = \begin{cases} h(\phi(x)) & x \in \phi^{-1}(\operatorname{dom} h), \\ +\infty & \text{otherwise.} \end{cases}$$

Then f is proper, has nonempty minimizer set $\mathcal{X}_f = \phi^{-1}(\mathcal{X}_h)$, and satisfies Assumptions 3.1 and 3.2 with the same radius t.

Proof. Since h is proper and ϕ is bijective, f is proper. Moreover,

$$x \in \mathcal{X}_f \quad \Longleftrightarrow \quad h(\phi(x)) = \inf h \quad \Longleftrightarrow \quad \phi(x) \in \mathcal{X}_h,$$

and hence $\mathcal{X}_f = \phi^{-1}(\mathcal{X}_h) \neq \emptyset$.

Now, by the definition of $\boldsymbol{Q}$, for any $x, y \in \mathbb{R}^d$,

$$\|\phi(y) - \phi(x)\|_{\boldsymbol{X}}^2 = (y-x)^\top \boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q} (y-x) = \|y - x\|_{\boldsymbol{X}}^2,$$

so $y \in \mathcal{B}_{\boldsymbol{X}}(x,t)$ if and only if $\phi(y) \in \mathcal{B}_{\boldsymbol{X}}(\phi(x),t)$. Using the properties of the broximal operator, one can easily check that for any $x \in \operatorname{dom} f$

$$\operatorname{brox}_f^t(x) = \phi^{-1}\left(\operatorname{brox}_h^t(\phi(x))\right).$$

Next, we verify that Assumption 3.1 holds. Let $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t)\}$ and $u \in \operatorname{brox}_f^t(x)$. Then $\phi(x) \in \operatorname{dom} h \setminus \{\mathcal{X}_h + \mathcal{B}_{\boldsymbol{X}}(0,t)\}$ and $\phi(u) \in \operatorname{brox}_h^t(\phi(x))$. By Assumption 3.1 for h, there exists $x_\star \in \mathcal{X}_h$ such that

$$\langle \phi(x) - \phi(u), \phi(u) - x_\star \rangle_{\boldsymbol{X}} \ge 0.$$

Using $\boldsymbol{Q}^\top \boldsymbol{X} \boldsymbol{Q} = \boldsymbol{X}$, this is equivalent to

$$\left\langle x - u, u - \phi^{-1}(x_\star) \right\rangle_{\boldsymbol{X}} \ge 0.$$

Since $\phi^{-1}(x_\star) \in \mathcal{X}_f$, Assumption 3.1 holds for f.

Lastly, if $\operatorname{brox}_f^t(x) = \{x\}$, then $\operatorname{brox}_h^t(\phi(x)) = \{\phi(x)\}$. By Assumption 3.2 for h, $\phi(x) \in \mathcal{X}_h$, and hence $x \in \mathcal{X}_f$. □

The following lemma is, in fact, a direct corollary of Lemma 5.1. For completeness, we include a full proof.

Corollary 1 (Affine transformations of the function)**.** *Let $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function with nonempty minimizer set $\mathcal{X}_g$, satisfying Assumptions 3.1 and 3.2 for radius $t > 0$. Let $a > 0$ and $b \in \mathbb{R}$, and define*

$$f(x) = a\, g(x) + b.$$

Then f is proper, has nonempty minimizer set $\mathcal{X}_f = \mathcal{X}_g$, and satisfies Assumptions 3.1 and 3.2 with the same radius t.

Proof. Since g is proper, f is also proper. Moreover, the set of minimizers is unchanged under positive scaling and additive shifts, so $\mathcal{X}_f = \mathcal{X}_g \neq \emptyset$.

Note that for any $x \in \operatorname{dom} f$ and radius $t > 0$, by definition of the broximal operator:

$$\begin{aligned} u \in \operatorname{brox}_f^t(x) \quad &\iff \quad u \in \operatorname*{arg\,min}_{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)} f(z) = \operatorname*{arg\,min}_{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)} (ag(z) + b) \\ &\iff \quad u \in \operatorname*{arg\,min}_{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)} g(z) \\ &\iff \quad u \in \operatorname{brox}_g^t(x). \end{aligned}$$

Hence, we have $\operatorname{brox}_f^t(x) = \operatorname{brox}_g^t(x)$. Now, let $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t)\} = \operatorname{dom} g \setminus \{\mathcal{X}_g + \mathcal{B}_{\boldsymbol{X}}(0,t)\}$ and $u \in \operatorname{brox}_f^t(x) = \operatorname{brox}_g^t(x)$. It is obvious that Assumption 3.1 holds from f whenever it holds for g.

Lastly, if $\operatorname{brox}_f^t(x) = \{x\}$, then

$$\operatorname{brox}_g^t(x) = \operatorname{brox}_f^t(x) = \{x\}.$$

By Assumption 3.2 for g, we have $x \in \mathcal{X}_g = \mathcal{X}_f$, so Assumption 3.2 also holds for f with radius t. □

D.3 Quasiconvexity

Definition 5.1 (Quasiconvexity)**.** A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is *quasiconvex* if

$$f\left((1-\lambda)x + \lambda y\right) \leq \max\{f(x), f(y)\} \qquad \forall x, y \in \mathbb{R}^d, \lambda \in [0,1].$$

If the inequality is strict for all $x \neq y$ and $\lambda \in (0,1)$, then f is *strictly quasiconvex*.

The proof of the main result in this section will use the characterization of quasiconvexity via sublevel sets.

Lemma D.2. Let f be quasiconvex. Then, for any $x \in \mathbb{R}^d$, the sublevel sets $L_f^{\leq}(x) := \left\{u \in \mathbb{R}^d : f(u) \leq f(x)\right\}$ and $L_f^{<}(x) := \left\{u \in \mathbb{R}^d : f(u) < f(x)\right\}$ are convex.

Remark 2. *In fact, one can show that this is an "if and only if" statement, i.e., $L_f^{\leq}(x)$ being convex for all x implies quasiconvexity.*

Proof. Let $u_1, u_2 \in L_f^{\leq}(x)$. Then, for any $\lambda \in [0,1]$, quasiconvexity gives

$$f\left((1-\lambda)u_1 + \lambda u_2\right) \leq \max\{f(u_1), f(u_2)\} \leq f(x),$$

proving the first result. The proof of the second statement is entirely analogous. □

Theorem 5.1. Suppose that f is strictly quasiconvex. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Proof. Fix some $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}(0,t)\}$ and suppose for a contradiction that there exists $u \in \operatorname{brox}_f^t(x)$ such that

$$\langle x - u, u - x_\star \rangle < 0.$$

Let $p := \Pi_{[u,x_\star]}(x)$ be the projection of x onto the line passing through u and $x_\star$. Since we assume that $\langle x-u, u-x_\star\rangle < 0$, p must lie strictly between u and $x_\star$. But then

$$\|x-p\|^2 = \|x-u\|^2 - \|u-p\|^2 < \|x-u\|^2,$$

and hence $p \in \mathcal{B}(x,t)$, so by definition of broximal operator $f(p) \geq f(u)$. On the other hand, strict quasiconvexity implies

$$f(p) < \max\{f(u), f_\star\} = f(u).$$

This contradiction shows that Assumption 3.1 holds.

Now, suppose that $\text{brox}_f^t(x) = \{x\}$. Then x is the unique minimizer of f on $\mathcal{B}(x,t)$, and so by quasiconvexity it must be the global minimizer (otherwise $L_f^{\leq}(x)$ would not be connected, and so would not be convex). □

D.4 Pseudoconvexity

Definition 5.2 (Pseudoconvexity)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is *pseudoconvex* if

$$\langle \nabla f(x), y-x\rangle \geq 0 \quad \Longrightarrow \quad f(y) \geq f(x) \qquad \forall x, y \in \mathbb{R}^d.$$

Theorem 5.2. Suppose that f is pseudoconvex. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\text{BA}}(t)$ for any $t > 0$.

Proof. Fix some $x \in \text{dom}\, f \setminus \{\mathcal{X}_f + \mathcal{B}(0,t)\}$ and suppose for a contradiction that there exists $u \in \text{brox}_f^t(x)$ such that

$$\langle x-u, u-x_\star\rangle < 0.$$

By Lemma B.1, we have

$$-\nabla f(u) \in \mathcal{N}_{\mathcal{B}(x,t)}(u) \overset{equation\ 1}{=} \begin{cases} \mathbb{R}_{\geq 0}(u-x) & \|x-u\| = t, \\ \{0\} & \|x-u\| < t. \end{cases}$$

Since pseudoconvexity implies that every stationary point is a global minimizer, and because $x \in \text{dom}\, f \setminus \{\mathcal{X}_f + \mathcal{B}(0,t)\}$, we must have $u \notin \mathcal{X}_f$. Therefore, there exists $c \geq 0$ such that $\nabla f(u) = c(x-u)$. Moreover, $c > 0$, since otherwise $\nabla f(u) = 0$ and the same contradiction would arise. Using this representation, we obtain

$$0 > \langle x-u, u-x_\star\rangle = \frac{1}{c}\langle \nabla f(u), u-x_\star\rangle.$$

By pseudoconvexity, this implies $f(x_\star) \geq f(u)$, contradicting the fact that $u \notin \mathcal{X}_f$. This contradiction establishes Assumption 3.1.

Assumption 3.2 holds since under pseudoconvexity any stationary point is a global minimizer. □

D.4.1 Generalized Pseudoconvexity

Definition 5.2 is stated for differentiable functions. We now extend the comparison to the non-differentiable setting by considering generalized notions of pseudoconvexity, and show that our assumptions remain strictly weaker in this broader regime.

To this end, we work with an abstract notion of a generalized subdifferential. Specifically, we only require the mapping ∂f to satisfy the following properties:

- $0 \in \partial f(x)$ whenever $x \in \text{dom}(f)$ is a local minimizer of f,
- if $u \in \underset{z \in \mathcal{B}_{\boldsymbol{X}}(x,t)}{\arg\min}\ f(z)$, then $0 \in \partial f(u) + \mathcal{N}_{\mathcal{B}_{\boldsymbol{X}}(x,t)}(u)$.

Definition D.1 (Generalized Pseudoconvexity)**.** Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a proper function and let $\partial f \subseteq \mathbb{R}^d$ denote a generalized subdifferential defined for all all $x \in \operatorname{dom} f$. We say that f is *generalized pseudoconvex* if

$$(\exists g \in \partial f(x) \text{ s.t. } \langle g, y - x\rangle \geq 0) \quad \implies \quad f(y) \geq f(x)$$

for all $x, y \in \operatorname{dom} f$.

Remark 3. *In the differentiable case, where $\partial f(x) = \{\nabla f(x)\}$, the above condition reduces to*

$$\langle \nabla f(x), y - x\rangle \geq 0 \quad \implies \quad f(y) \geq f(x),$$

which is the standard definition of pseudoconvexity.

Lemma D.3. Suppose that f is generalized pseudoconvex. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Proof. Fix any $x \in \operatorname{dom} f \backslash \{\mathcal{X}_f + \mathcal{B}(0, t)\}$ and let $u = \operatorname{brox}_f^t(x)$. Taking the contrapositive of the condition in Definition D.1, we have

$$f(x) > f(y) \implies (\langle g, y - x\rangle < 0 \quad \forall g \in \partial f(x)),$$

and hence, choosing $x = u$ and $y = x_\star \in \mathcal{X}_f$ yields

$$\langle g, x_\star - u\rangle < 0 \quad \forall g \in \partial f(u).$$

By the assumed properties of the generalized subdifferential, there exists $g_u \in \partial f(u)$ such that $-g_u \in \mathcal{N}_{\mathcal{B}\boldsymbol{x}(x,t)}(u)$. Since under generalized pseudoconvexity every local minimizer is global and $u \notin \mathcal{X}_f$, there exists $c \geq 0$ such that $g_u = c(x - u)$ (Fact 1). Substituting this expression above gives

$$\langle c(x - u), x_\star - u\rangle < 0.$$

Thus $c > 0$ and $\langle x - u, u - x_\star\rangle > 0$, which establishes Assumption 3.1 for any choice of $x_\star \in \mathcal{X}_f$.

Finally, Assumption 3.2 follows again from the fact that generalized pseudoconvexity ensures that all local minimizers are global minimizers. □

D.5 Quasar convexity

Definition 5.3 (Quasar convexity)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$ for some $\zeta \in (0, 1]$ if

$$f(x) - f_\star \leq \frac{1}{\zeta} \langle \nabla f(x), x - x_\star\rangle \qquad \forall x \in \mathbb{R}^d.$$

Lemma D.4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$. Then

$$f\left(\lambda x_\star + (1 - \lambda)x\right) \leq \zeta\lambda f_\star + (1 - \zeta\lambda) f(x)$$

for all $x \in \mathbb{R}^d$ and $\lambda \in [0, 1]$.

Proof. The result follows from Lemma 10 by Hinder et al. (2019). □

Lemma D.5. Suppose that f is ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$ and take any $x \in \mathbb{R}^d \backslash \{\mathcal{X}_f + \mathcal{B}(0, t)\}$, where $t > 0$. Let $u = \operatorname{brox}_f^t(x)$ and let $\Pi(x)$ be the projection of x onto the line passing through u and $x_\star$. Then there exists $\lambda \in [0, 1)$ such that $u = \lambda x_\star + (1 - \lambda)\Pi(x)$.

Proof. First, note that $\|x - \Pi(x)\| \leq \|x - u\| \leq t$, so $\Pi(x) \in \mathcal{B}(x, t)$, meaning that $f(u) \leq f(\Pi(x))$ by the definition of broximal operator. Clearly, $u \neq x_\star$ and $\Pi(x) \neq x_\star$ since $x \in \mathbb{R}^d \backslash \{\mathcal{X}_f + \mathcal{B}(0, t)\}$. If $u = \Pi(x)$, then the claim of the lemma holds with $\lambda = 0$. Now, suppose that the three points u, $x_\star$ and $\Pi(x)$ are distinct. There are 3 possibilities: $\Pi(x)$ lies between u and $x_\star$, $x_\star$ lies between u and $\Pi(x)$, or u lies between $\Pi(x)$ and $x_\star$. That $x_\star$ does not lie between

u and $\Pi(x)$ is obvious, since if that was the case, then $x_\star$ would belong to the ball $\mathcal{B}(0,t)$, contradicting the fact that $x \in \mathbb{R}^d \backslash \{\mathcal{X}_f + \mathcal{B}(0,t)\}$. Assume that $\Pi(x)$ lies between u and $x_\star$, i.e., there exists $\lambda \in (0,1)$ such that $\Pi(x) = \lambda x_\star + (1-\lambda)u$. Then, Lemma D.4 gives

$$f(\Pi(x)) = f(\lambda x_\star + (1-\lambda)u) \leq \zeta\lambda f_\star + (1-\zeta\lambda)f(u) < \zeta\lambda f(\Pi(x)) + (1-\zeta\lambda)f(\Pi(x)),$$

where the last inequality follows from the fact that $f_\star < f(u) \leq f(\Pi(x))$. This contradiction finishes the proof. □

Theorem 5.3. Suppose that f is ζ-quasar convex with respect to $x_\star \in \mathcal{X}_f$. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Proof. By Lemma D.5, we have

$$\begin{aligned}\langle x-u, x_\star - u\rangle &= (1-\lambda)\langle x-u, x_\star - \Pi(x)\rangle \\ &= (1-\lambda)\langle x-\Pi(x), x_\star - \Pi(x)\rangle + (1-\lambda)\langle \Pi(x)-u, x_\star - \Pi(x)\rangle \\ &= 0 - (1-\lambda)\|\Pi(x)-u\|\,\|\Pi(x)-x_\star\| \leq 0,\end{aligned}$$

so Assumption 3.1 holds.

Now, suppose that $\mathrm{brox}_f^t(x) = \{x\}$. Then x is the unique minimizer of f on $\mathcal{B}_{\boldsymbol{X}}(x,t)$, and so it must be the global minimizer (otherwise $\mathcal{X}_f$ would not be connected, which cannot hold under quasar convexity – see Lemma D.4). □

D.6 Aiming Condition

Definition 5.4 (Aiming Condition)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies the *aiming condition* if $f_\star = 0$ and there exist $\bar{x} \in \Pi_{\mathcal{X}_f}(x)$ and $\theta > 0$ such that

$$\theta f(x) \leq \langle \nabla f(x), x - \bar{x}\rangle \qquad \forall x \in \mathbb{R}^d. \tag{4}$$

Lemma D.6. Let f satisfy the aiming condition. Then Assumption 3.2 holds with any $t > 0$ and

$$\langle x-u, u-\bar{u}\rangle \geq 0$$

for any $x \in \mathrm{dom}\, f \backslash \{\mathcal{X}_f + \mathcal{B}(0,t)\}$, where $\bar{u} \in \Pi_{\mathcal{X}_f}(u)$.

Proof. The proof follows the same structure as that of Theorem 5.2. For completeness, we present the full argument again. Let $x \in \mathrm{dom}\, f \backslash \{\mathcal{X}_f + \mathcal{B}(0,t)\}$ and $u \in \mathrm{brox}_f^t(x)$. By Lemma B.1, we have

$$-\nabla f(u) \in \mathcal{N}_{\mathcal{B}(x,t)}(u) \overset{\text{equation 1}}{=} \begin{cases} \mathbb{R}_{\geq 0}(u-x) & \|x-u\| = t, \\ \{0\} & \|x-u\| < t. \end{cases}$$

Since the aiming condition implies that every stationary point is a global minimizer, and because $x \in \mathrm{dom}\, f \backslash \{\mathcal{X}_f + \mathcal{B}(0,t)\}$, we must have $u \notin \mathcal{X}_f$. Therefore, there exists $c \geq 0$ such that $\nabla f(u) = c(x-u)$. In fact, c must be strictly positive. Indeed, if we had $c = 0$, then u would be a global minimizer, again contradicting the choice of x. Now, let $\bar{u} \in \Pi_{\mathcal{X}_f}(u)$ be the point for which equation 4 holds. Then

$$\langle x-u, u-\bar{u}\rangle = \frac{1}{c}\langle \nabla f(u), u-\bar{u}\rangle \geq \frac{\theta}{c} f(u) \geq 0,$$

proving the claim. □

Theorem 5.4. Let f satisfy the aiming condition and have a unique global minimizer. Then $f \in \mathcal{F}_{\boldsymbol{I}}^{\mathrm{BA}}(t)$ for any $t > 0$.

Proof. By Lemma D.6, for any $x \in \mathrm{dom}\, f \backslash \{\mathcal{X}_f + \mathcal{B}(0,t)\}$ and $t > 0$, we have

$$\langle x-u, u-\bar{u}\rangle \geq 0,$$

where $\bar{u} \in \Pi_{\mathcal{X}_f}(u)$. Since the minimizer is unique, $\mathcal{X}_f = \{x_\star\}$ is a singleton, and hence Assumption 3.1 holds.

Now, suppose that $\text{brox}_f^t(x) = \{x\}$. Then x is the unique minimizer of f on $\mathcal{B}(x,t)$, and so $\theta f(x) \le \langle \nabla f(x), x - \bar{x} \rangle = 0$, meaning that $x \in \mathcal{X}_f$. Hence Assumption 3.2 holds as well. □

Lastly, we establish the connectedness of$\mathcal{X}_f$ under aiming condition under the aiming condition (as used in Table 1).

Proposition 1 (Connectedness of minimizers under aiming)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and satisfy the aiming condition. Then the set of global minimizers $\mathcal{X}_f$ is connected.*

Proof. Since $\mathcal{X}_f = f^{-1}(\{0\})$ and f is continuous, $\mathcal{X}_f$ is closed. Assume for contradiction that $\mathcal{X}_f$ is disconnected. Then there exist two nonempty, disjoint, closed, connected components $C_1, C_2 \subseteq \mathcal{X}_f$ such that $\mathcal{X}_f = C_1 \cup C_2$ and $C_1 \cap C_2 = \emptyset$. Among all such pairs, choose C_1, C_2 so that the distance between them is minimal. Since C_1 and C_2 are disjoint closed sets in $\mathbb{R}^d$, the distance

$$d := \inf_{x \in C_1,\, y \in C_2} \|x - y\|$$

is strictly positive. Moreover, by closedness, there exist points $x_1 \in C_1$ and $x_2 \in C_2$ such that $\|x_1 - x_2\| = d$. Consider the line segment

$$\gamma(t) := (1-t)x_1 + t x_2, \qquad t \in [0,1],$$

and define the function $\varphi(t) := f(\gamma(t))$. Since f is differentiable, the function φ is differentiable on $[0,1]$. Moreover,

$$\varphi(0) = f(x_1) = 0, \qquad \varphi(1) = f(x_2) = 0,$$

so by Rolle's theorem, there exists $t_0 \in (0,1)$ such that $\varphi'(t_0) = 0$. Using the chain rule, we obtain

$$\varphi'(t) = \langle \nabla f(\gamma(t)), x_2 - x_1 \rangle ,$$

and hence

$$\langle \nabla f(\gamma(t_0)), x_2 - x_1 \rangle = 0. \tag{8}$$

Now, let $\bar{x} \in \Pi_{\mathcal{X}_f}(\gamma(t_0))$ be a projection point for which the aiming condition holds. We claim that $\bar{x}$ is either x_1 or x_2. Indeed, suppose there exists another point $z \in \Pi_{\mathcal{X}_f}(\gamma(t_0))$ such that $z \notin \{x_1, x_2\}$. By the minimality of d and the triangle inequality, we have

$$\|x_1 - \gamma(t_0)\| + \|\gamma(t_0) - x_2\| = d \le \|x_1 - z\| \le \|x_1 - \gamma(t_0)\| + \|\gamma(t_0) - z\| .$$

It follows that $\|\gamma(t_0) - x_2\| \le \|\gamma(t_0) - z\|$. On the other hand, since $z \in \Pi_{\mathcal{X}_f}(\gamma(t_0))$, we also have $\|\gamma(t_0) - x_2\| \ge \|\gamma(t_0) - z\|$, meaning that $\|\gamma(t_0) - x_2\| = \|\gamma(t_0) - z\|$. But then

$$\begin{aligned}
\|x_1 - z\|^2 &= \|x_1 - \gamma(t_0)\|^2 + \|z - \gamma(t_0)\|^2 - 2\langle x_1 - \gamma(t_0), z - \gamma(t_0)\rangle \\
&= \|x_1 - \gamma(t_0)\|^2 + \|z - \gamma(t_0)\|^2 - 2\|x_1 - \gamma(t_0)\|\,\|z - \gamma(t_0)\|\cos(\theta_1) \\
&< \|x_1 - \gamma(t_0)\|^2 + \|z - \gamma(t_0)\|^2 + 2\|x_1 - \gamma(t_0)\|\,\|z - \gamma(t_0)\| \\
&= \|x_1 - \gamma(t_0)\|^2 + \|x_2 - \gamma(t_0)\|^2 - 2\|x_1 - \gamma(t_0)\|\,\|x_2 - \gamma(t_0)\|\cos(\theta_2) \\
&= \|x_1 - \gamma(t_0)\|^2 + \|x_2 - \gamma(t_0)\|^2 - 2\langle x_1 - \gamma(t_0), x_2 - \gamma(t_0)\rangle \\
&= \|x_1 - x_2\|^2
\end{aligned}$$

where θ_1 is the angle between the vectors $x_1 - \gamma(t_0)$ and $z - \gamma(t_0)$ and θ_2 is the angle between the vectors $x_1 - \gamma(t_0)$ and $x_2 - \gamma(t_0)$. This contradicts the minimality of d.

So suppose without loss of generality that $\bar{x} = x_1$. Then the vector $\gamma(t_0) - \bar{x} = \gamma(t_0) - x_1$ is parallel to $x_2 - x_1$, and hence, by aiming condition and equation 8,

$$\theta f(\gamma(t_0)) \le \langle \nabla f(\gamma(t_0)), \gamma(t_0) - \bar{x} \rangle = 0.$$

Since $\theta > 0$, it follows that $f(\gamma(t_0)) = 0$, i.e., $\gamma(t_0) \in \mathcal{X}_f$. However, since $t_0 \in (0,1)$, we have $\|x_1 - \gamma(t_0)\| < d$, which contradicts the minimality of d. Therefore, $\mathcal{X}_f$ must be connected. □

E MONOTONICITY OF FUNCTION CLASSES

In this section, we study how the classes of functions satisfying our main assumptions behave as the radius parameter t changes. We show that functions satisfying only Assumption 3.2 form a monotone class, while functions satisfying Assumption 3.1 do not in general exhibit monotonicity in t.

That said, if monotonicity in the radius is desired, there exists a subclass of $\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$, which we term the *Uniform Broximal Alignment* class (denoted $\mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t)$; see Section E.1), that *is* monotone in t.

We first define the class of functions that satisfy Assumption 3.2 alone.

Definition E.1. For $t > 0$, define

$$\begin{aligned}\mathcal{F}_2(t) &:= \{f : f \text{ is proper and satisfies Assumption 3.2 with } t > 0\} \\ &= \left\{f \text{ proper} : \forall x \in \operatorname{dom} f,\ \operatorname{brox}_f^t(x) = \{x\} \implies x \in \mathcal{X}_f\right\}.\end{aligned}$$

The following proposition shows that $\mathcal{F}_2(t)$ is monotone with respect to t.

Proposition 2. *For $0 < t_1 \le t_2$, we have*

$$\mathcal{F}_2(t_1) \subseteq \mathcal{F}_2(t_2).$$

Proof. Take any $f \in \mathcal{F}_2(t_1)$ and fix $x \in \operatorname{dom} f$ such that $\operatorname{brox}_f^{t_2}(x) = \{x\}$. We want to show that $x \in \mathcal{X}_f$. Since $t_1 \le t_2$, we have $B_{t_1}(x) \subseteq B_{t_2}(x)$, so x is also the unique minimizer of f over $B_{t_1}(x)$, i.e.,

$$\operatorname{brox}_f^{t_1}(x) = \{x\}.$$

By definition of $\mathcal{F}_2(t_1)$, we know that $x \in \mathcal{X}_f$, and hence $f \in \mathcal{F}_2(t_2)$. □

Thus, the class $\mathcal{F}_2(t)$ grows as t increases.

Next, we consider the class of functions satisfying Assumption 3.1:

Definition E.2. For $t > 0$, define

$$\mathcal{F}_1(t) := \{f : f \text{ is proper and satisfies Assumption 3.1 with radius } t > 0\}.$$

Unlike $\mathcal{F}_2(t)$, the class $\mathcal{F}_1(t)$ is generally *not* monotone in t.

Proposition 3. *The class $\mathcal{F}_1(t)$ is in general not monotone in t: for $t_1 < t_2$, neither $\mathcal{F}_1(t_1) \subseteq \mathcal{F}_1(t_2)$ nor $\mathcal{F}_1(t_2) \subseteq \mathcal{F}_1(t_1)$ holds in general.*

Proof. We provide two concrete examples.

Example 1: $f \in \mathcal{F}_1(1)$ but $f \notin \mathcal{F}_1(2)$.

Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ be defined on

$$\operatorname{dom} f = \{(0,0), (1,0), (2,0), (3,0), (3,2)\},$$

with values

$$f(0,0) = 0, \quad f(1,0) = 1, \quad f(2,0) = 2, \quad f(3,0) = 3, \quad f(3,2) = 0.5.$$

One can easily check that for radius $t = 1$, all points outside $\operatorname{dom} f \setminus \{\mathcal{X}_f + B(0,1)\} = \{(2,0), (3,0), (3,2)\}$ satisfy Assumption 3.1:

- $x = (2,0)$: feasible points $B_1(x) \cap \operatorname{dom} f = \{(1,0), (2,0), (3,0)\}$, so $u = \operatorname{brox}_f^1(x) = \{(1,0)\}$, and

$$\langle x - u, u - x_\star \rangle = \langle (2,0) - (1,0), (1,0) - (0,0) \rangle = 1 \ge 0.$$

- $x = (3, 0)$: feasible points $\{(2, 0), (3, 0)\}$, so $u = \text{brox}_f^1(x) = \{(2, 0)\}$, and
$$\langle x - u, u - x_\star \rangle = \langle (3, 0) - (2, 0), (2, 0) - (0, 0) \rangle = 2 \geq 0.$$
- $x = (3, 2)$: feasible points $\{(3, 2)\}$, so $u = \text{brox}_f^1(x) = \{(3, 2)\}$, and
$$\langle x - u, u - x_\star \rangle = 0.$$

Hence $f \in \mathcal{F}_1(1)$. Now, we show that $f \notin \mathcal{F}_1(2)$. Indeed, consider $x = (3, 0)$. Then $\text{brox}_f^2(x) = \{(3, 2)\}$ and
$$\langle x - u, u - x_\star \rangle = \langle (3, 0) - (3, 2), (3, 2) - (0, 0) \rangle = -4 < 0,$$
which violates Assumption 3.1. Hence $f \notin \mathcal{F}_1(2)$.

Example 2: $f \in \mathcal{F}_1(3)$ but $f \notin \mathcal{F}_1(1)$.

Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ be defined on
$$\text{dom } f = \{(0, 0), (2, 0), (2, 1)\},$$
with values
$$f(0, 0) = 0, \quad f(2, 0) = 1, \quad f(2, 1) = 0.1,$$
with global minimizer $x_\star = (0, 0)$.

We first show that $f \notin \mathcal{F}_1(1)$. Consider $x = (2, 0)$. The feasible points in $B_1(x)$ are $(2, 0)$ and $(2, 1)$, and $u = \text{brox}_f^1(x) = \{(2, 1)\}$. Then
$$\langle x - u, u - x_\star \rangle = \langle (2, 0) - (2, 1), (2, 1) - (0, 0) \rangle = -1 < 0,$$
violating Assumption 3.1. Hence $f \notin \mathcal{F}_1(1)$.

Now, we show that $f \in \mathcal{F}_1(3)$. For $t = 3$, observe that
$$\|(2, 0)\| = 2 < 3, \qquad \|(2, 1)\| = \sqrt{5} < 3,$$
meaning that
$$\text{dom } f \subset \mathcal{X}_f + B_3(0).$$
Therefore, there are no points outside $\mathcal{X}_f + B_3(0)$, and Assumption 3.1 is trivially satisfied. Hence $f \in \mathcal{F}_1(3)$. □

E.1 Uniform Broximal Alignment

Definition E.3 (Uniform Broximal Alignment)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function. We say that f satisfies the *uniform broximal alignment with radius* $t > 0$ if there exists $x_\star \in \mathcal{X}_f$ such that, for every
$$x \in \text{dom } f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0, t)\}$$
and every $z \in \text{dom } f$ satisfying
$$f(z) \leq f(x),$$
we have
$$\langle x - z, x_\star - z \rangle_{\boldsymbol{X}} \leq 0.$$
If this holds, we will write $f \in \mathcal{F}_{\boldsymbol{X}}^{\text{UBA}}(t)$.

Uniform alignment is strictly stronger than broximal alignment: it requires the alignment inequality for all points z with $f(z) \leq f(x)$, whereas Assumption 3.1 from our paper only requires it for those z that are broximal minimizers of f over $\mathcal{B}_{\boldsymbol{X}}(x, t)$.

We now show that the uniform alignment condition is monotonic in the radius.

Theorem E.1. Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper, closed function and let $t' \geq t > 0$. Then

$$\mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t) \subseteq \mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t').$$

Proof. Let $f \in \mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t)$ and fix any $t' \geq t$. We will show that f satisfies the uniform broximal alignment condition with radius t'. Since $f \in \mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t)$, there exists $x_\star \in \mathcal{X}_f$ such that, for every $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t)\}$ and every $z \in \operatorname{dom} f$ satisfying $f(z) \leq f(x)$, we have

$$\langle x - z, x_\star - z \rangle_{\boldsymbol{X}} \leq 0. \tag{9}$$

Now, let $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t')\}$ and let $z \in \operatorname{dom} f$ satisfy $f(z) \leq f(x)$. Because $t' \geq t$, we have $\mathcal{B}_{\boldsymbol{X}}(0,t) \subseteq \mathcal{B}_{\boldsymbol{X}}(0,t)$, and hence

$$\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t) \subseteq \mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t').$$

Therefore,

$$x \notin \mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t') \qquad \Longrightarrow \qquad x \notin \mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t).$$

So the pair (x, z) satisfies the assumptions of equation 9, and thus

$$\langle x - z, x_\star - z \rangle_{\boldsymbol{X}} \leq 0.$$

Since this holds for every $x \in \operatorname{dom} f \setminus \{\mathcal{X}_f + \mathcal{B}_{\boldsymbol{X}}(0,t')\}$ and every $z \in \operatorname{dom} f$ with $f(z) \leq f(x)$, we conclude that $f \in \mathcal{F}_{\boldsymbol{X}}^{\mathrm{UBA}}(t')$. The proof is complete. □

F NOTATION

Notation	
d	Dimensionality of the problem
x_k	k-th iterate of an algorithm
$\nabla f(x)$	Gradient of function f at x
$\partial f(x)$	Subdifferential of function f at x
$\mathcal{X}_f$	Set of global minimizers of f
$f_\star$	Minimum of f
$\mathrm{dom}(f)$	$:= \{x \in \mathbb{R}^d : f(x) < +\infty\}$ – domain of f
$\langle \cdot, \cdot \rangle$	Standard Euclidean inner product
$\|\cdot\|$	Standard Euclidean norm
$\langle \cdot, \cdot \rangle_{\boldsymbol{X}}$	$\langle x, y \rangle_{\boldsymbol{X}} := \langle \boldsymbol{X}x, y \rangle$ for a symmetric positive definite matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d}$
$\|\cdot\|_{\boldsymbol{X}}$	$\|x\|_{\boldsymbol{X}} := \sqrt{x^\top \boldsymbol{X} x}$ for a symmetric positive definite matrix $\boldsymbol{X} \in \mathbb{R}^{d \times d}$
$\mathcal{B}(x, t)$	$:= \left\{z \in \mathbb{R}^d : \|z - x\| \leq t\right\}$
$\mathcal{B}_{\boldsymbol{X}}(x, t)$	$:= \left\{z \in \mathbb{R}^d : \|z - x\|_{\boldsymbol{X}} \leq t\right\}$
$\mathrm{brox}_f^t(x)$	Broximal operator associated with function f with radius $t > 0$
$\mathcal{F}_{\boldsymbol{X}}^{\mathrm{BA}}(t)$	The class of Broximal Aligned functions (see Definition 3.1)
$L_f^<(x)$	$:= \left\{u \in \mathbb{R}^d : f(u) < f(x)\right\}$
$L_f^\leq(x)$	$:= \left\{z \in \mathbb{R}^d : f(z) \leq f(x)\right\}$
$\Pi(\cdot, \mathcal{X})$	Projection onto a set $\mathcal{X}$
$\delta_{\mathcal{X}}(y)$	$= \begin{cases} 0, & y \in \mathcal{X} \\ +\infty, & y \notin \mathcal{X} \end{cases}$
$\mathrm{dist}(x, \mathcal{X})$	$:= \inf_{z \in \mathcal{X}} \|x - z\|_{\boldsymbol{X}}$
$\mathcal{N}_{\mathcal{X}}(x)$	$:= \left\{g \in \mathbb{R}^d : \langle g, z - x \rangle \leq 0 \, \forall z \in \mathcal{X}\right\}$ – the normal cone of $\mathcal{X}$ at x
$\mathbb{R}_{\geq 0}(z)$	$:= \{\lambda z \ : \ \lambda \geq 0\}$

Table 2: Frequently used notation.