BENIGN POPULATION LANDSCAPES DO NOT TRANSFER TO EMPIRICAL RISK MINIMIZATION IN NONSMOOTH OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benign landscapes provide a geometric explanation for the success of iterative methods on nonconvex problems arising in data science. The landscape is benign if every local minimizer is global and every other critical point is a strict saddle. For methods that avoid strict saddles, convergence to a critical point therefore implies convergence to a global minimizer. For smooth stochastic optimization, a common strategy for establishing such landscapes proceeds in two steps: (i) establish benign geometry for the population objective—which is often easier to analyze—and (ii) transfer this geometry to the empirical objective using concentration of gradients and Hessians. Recent work has extended saddle-point avoidance guarantees to weakly convex nonsmooth optimization through the notion of active strict saddles. We show, however, that benign population geometry need not transfer to empirical objectives in this setting. Specifically, for robust phase retrieval with Gaussian measurements—a well-studied and highly structured problem—the population objective is benign. Nevertheless, the empirical objective has sharp spurious local minima with probability bounded away from zero as the sample size tends to infinity. This asymptotic probability lower bound is independent of the dimension.

1 INTRODUCTION

Simple iterative methods, such as gradient descent, have proven to be highly effective at solving nonconvex optimization problems stemming from data science, machine learning, and signal processing. A working geometric explanation for this success is the benign landscape of the objective: all local minima are global, and every other critical point is a strict saddle—meaning that its Hessian has a strictly negative eigenvalue. Under this geometry, many suitably tuned first-order methods escape strict saddles and consequently recover global minimizers whenever they converge [40, 48]. Further, stochastically perturbed variants of gradient descent enjoy strong nonasymptotic guarantees [35]. This geometric explanation is supported by a number of high-impact problems that exhibit benign landscapes under suitable assumptions [56, 32, 7, 59, 62, 31].

Most of these examples arise from statistical models and admit both population and empirical objectives. Formally, for samples $z_1, \ldots, z_n$ from some distribution, these objectives are defined as

$$F(x) := \mathbb{E}_z f(x, z), \qquad \text{and} \qquad F_n(x) := \frac{1}{n}\sum_{i=1}^{n} f(x, z_i),$$

where f is problem specific. Although the objective being minimized is often F_n, the population objective F tends to be simpler to analyze. This motivates a common two-stage approach to establish benign geometry: first, show that it holds for the population objective F and then prove that this geometry persists for the empirical objective F_n whenever n is sufficiently large. The second step is carried out by exploiting the concentration of gradients and Hessians of F_n around those of F.

The seminal work of Davis and Drusvyatskiy [14] extended saddle-avoidance guarantees to nonsmooth weakly convex optimization—a broad class encompassing many problems arising in data science [13]. Unlike in the smooth setting, a classical Hessian is not available for these functions.

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.

To overcome this obstacle, they introduce the notion of an active strict saddle. In a nutshell, these are critical points lying on a manifold along which the objective is smooth—and has a direction of strictly negative curvature—and away from which the function grows sharply; see Figure 1 for an illustration and Section 2 for a formal definition. Just as in the smooth setting, several first-order methods are guaranteed to avoid these active strict saddles [14, 21, 4, 19, 50]. These results provide a reason to seek nonsmooth objectives with benign landscapes and raise a natural question:

Does benign population geometry transfer to the empirical landscape in nonsmooth optimization?

We answer this question negatively: for general weakly convex problems, this is not the case. To do so, we consider phase retrieval, an inverse problem with a long history in physics and signal processing [28, 55]. Although phase retrieval is commonly formulated over the complex numbers, we study a stylized real version in which the goal is to recover a vector $x^\star \in \mathbb{R}^d$, up to sign, from n nonlinear measurements of the form $\langle a_i, x^\star\rangle^2$, where the vectors a_i are i.i.d. standard Gaussians. To solve this problem, we consider the robust formulation

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \left| \langle a_i, x\rangle^2 - \langle a_i, x^\star\rangle^2 \right|. \qquad (1)$$

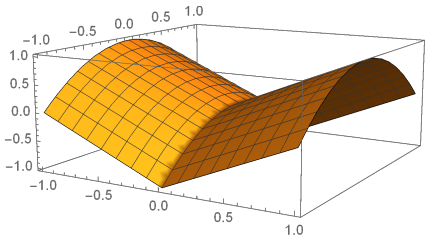


Figure 1: The function $(x, y) \mapsto |x| - y^2$ has an active strict saddle at zero, with the y-axis as the active manifold.

This objective is highly structured: it is weakly convex, a single-index model, and exhibits composite structure, i.e., it is a composition of a convex function with a smooth map [25]. Further, this formulation and related low-rank matrix recovery formulations have been studied extensively and serve as standard testbeds for theory and algorithms in weakly convex optimization [18, 17, 13, 22, 12, 11, 52].

Population objective. The population counterpart of (1) and its first-order landscape were characterized in [18]. For a unit-norm $x^\star$, the critical points of $F(x) = \mathbb{E}\left|\langle a, x\rangle^2 - \langle a, x^\star\rangle^2\right|$ are

$$\{\pm x^\star\} \cup \{0\} \cup \mathcal{S} \qquad \text{with} \qquad \mathcal{S} \coloneqq \{\tau u \mid \|u\| = 1 \text{ and } u \perp x^\star\},$$

where $\tau \approx 0.4416$ is the unique solution of a certain equation. That is, the critical points correspond to the true underlying solutions $\pm x^\star$, zero, and a $(d-2)$-dimensional sphere orthogonal to the solution. We show here that F is twice continuously differentiable on $\mathbb{R}^d \setminus \{\pm x^\star\}$. Further, zero and the vectors on the sphere $\mathcal{S}$ are all standard saddles. Close to the ground-truth solutions, the population objective is sharp, i.e., the objective can be bounded below by a constant multiple of the distance to the minimizer set. In turn, the landscape is benign: the only local minimizers are the ground-truth solutions, and all other critical points are (standard) strict saddles—which are consequently active strict saddles with an active manifold corresponding to the full domain.

Empirical landscape. The geometry of F_n differs substantially from that of its population counterpart. We establish three complementary results.

1. (**Active strict saddles**) Suppose $n \geq 2d$. Then, almost surely,

   every critical point of F_n is either a strict local minimizer or an active strict saddle.

   This dichotomy classifies the critical points, but leaves open whether every local minimizer is global. Our next result shows that this need not be the case.

2. (**Expected number of spurious minimizers**) Let $M_{d,n}$ denote the number of sharp spurious local minimizers. Here, sharpness means that the objective grows at least linearly with distance from each such minimizer in a sufficiently small neighborhood. For every fixed dimension $d \geq 4$, we show that

   $$\mathbb{E}M_{d,n} = \Omega\left(n^{d/2-1}\right) \quad \text{as } n \to \infty.$$

   Thus, the expected number of spurious minimizers grows at least polynomially with the sample size. However, this alone does not rule out their absence with high probability: the growing expectation could be driven by rare events with very many spurious minimizers.

3. (**Probability of having a spurious minimizer**) We therefore establish a stronger guarantee:

$$\liminf_{n\to\infty} \mathbb{P}(F_n \text{ has a spurious sharp local minimum}) > 0.136, \quad \text{for any fixed } d \geq 4.$$

This lower bound is independent of dimension. Moreover, the spurious minimizers constructed in our proof approach the population saddle set $\mathcal{S}$ as $n \to \infty$. Thus, sharp empirical local minima can accumulate at population strict saddles, even as the sample size grows.

Intuition. Our results may initially seem counterintuitive in light of the population-to-empirical landscape transfer familiar from smooth optimization. The key difference is that nonsmooth empirical perturbations can introduce linear growth that dominates the negative curvature of the population objective. To illustrate this point, consider the one-dimensional empirical objective

$$H_n(t) = -t^2 + \frac{1}{n}\sum_{i=1}^{n} \alpha_i |t|,$$

where $\alpha_1, \ldots, \alpha_n$ are i.i.d. standard Gaussian random variables. The population objective is $\mathbb{E}H_n(t) = -t^2$, so the origin is a strict maximizer and hence a strict saddle. Yet, for every finite n, the empirical coefficient $n^{-1}\sum_{i=1}^{n} \alpha_i$ is positive with probability $1/2$. On this event, the linear term dominates the negative quadratic term sufficiently close to the origin, creating a sharp local minimum. Although the perturbation vanishes as n grows, the probability of creating this local minimum remains unchanged. This illustrates the mechanism behind the sharp local minima that arise near the population saddle set $\mathcal{S}$.

Outline. The remainder of this section covers related work. Section 2 introduces the necessary background for the rest of the paper. Section 3 presents our results on the population and empirical landscapes, and Section 4 outlines the proof strategy. Section 5 studies algorithmic recovery from perturbations of the population saddle sphere. Finally, Section 6 summarizes our conclusions and discusses limitations and future directions. All technical proofs are deferred to the appendix.

Related work

Population-to-empirical transfer. For smooth losses, [30, 46] established general results for the uniform concentration of gradients and Hessians under mild assumptions. For weakly convex optimization, [15] showed the graphical convergence of subdifferentials for weakly convex stochastic problems. Similarly, [52] studied uniform convergence of the subdifferentials for a more structured family of weakly convex functions and used it to analyze robust phase retrieval [18]. To our knowledge, our result is the first to identify a second-order obstruction in this setting. Beyond weakly convex optimization, [54] established uniform laws of large numbers for random set-valued mappings and applied them to the consistency of stationary points in nonsmooth stochastic programming. On the other hand, [57, 58] provided negative and positive results for uniform convergence of Clarke subdifferentials of Lipschitz objectives.

Nonsmooth landscapes. For symmetric rank-one matrix factorization with entrywise ℓ_1 loss, the recent works [33, 38] established that every second-order stationary point, defined through the second subderivative, is globally optimal. Moreover, [33] showed that nonglobal stationary points may lack active manifolds and that arbitrarily small linear perturbations can lead to spurious local minima. The asymmetric analogue of this matrix factorization formulation was studied in [37], which characterized precisely when it has no spurious local minima. For orthogonal dictionary learning, [3] used uniform concentration of subdifferentials to establish recovery guarantees for a randomly initialized Riemannian subgradient method applied to a sphere-constrained ℓ_1 objective. The authors of [61] established the benign empirical landscape of a smoothed version of (1) with a sufficiently small fixed positive smoothing parameter. Intuitively, this landscape seems to smooth out the sharp kinks we identify in the purely nonsmooth objective. The second-order analysis and possible existence of spurious local minima for robust phase retrieval (1) were explicitly identified as relevant unresolved problems in [61, 33].

Nonsmooth algorithmic guarantees. Sharpness, together with Lipschitzness and weak convexity, yields fast local convergence of subgradient and prox-linear methods [49, 17, 25, 11]. These

convergence results require initialization near the solution set and do not deal with stationary points outside that region. Beyond local convergence, [36] developed a smoothing algorithm to find approximate second-order stationary points of a smooth population risk from a sufficiently accurate, potentially nonsmooth uniform approximation. To our knowledge, all general-purpose guarantees for convergence to second-order critical points of nonsmooth objectives rely on active strict saddles [14, 4, 21, 19, 50]. Our results suggest that we need to reevaluate this algorithmic explanation if we wish to explain the empirical success of first-order methods on statistical recovery problems.

2 PRELIMINARIES

In this section, we introduce the notation and background required for the rest of the paper.

Basic notation. We use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ for the Euclidean inner product and norm, and $\|\cdot\|_{\mathrm{op}}$ for the operator norm. Set addition and scalar multiplication are understood elementwise: $A+B = \{a+b : a \in A, b \in B\}$ and $Sv = \{sv : s \in S\}$ for $S \subset \mathbb{R}$. Given a differentiable map $g\colon \mathbb{R}^d \to \mathbb{R}^m$, we use ∇g to denote its Jacobian. When $m = 1$, this corresponds to the gradient map, and, further, when g is a twice differentiable function, the symbol $\nabla^2 g$ denotes its Hessian map.

Manifolds. Our treatment of manifolds here is rather brief since we will only focus on Euclidean manifolds, i.e., affine subspaces. We refer the interested reader to the textbooks by Boumal [8] and Lee [41]. A C^p manifold $\mathcal{M} \subset \mathbb{R}^d$ is locally a level set $G^{-1}(0)$ of a C^p map whose Jacobian has full row rank. Its tangent space at x is $T_{\mathcal{M}}(x) = \ker \nabla G(x)$. When $\mathcal{M}$ is an affine subspace, $T_{\mathcal{M}}(x)$ is equal to the same affine subspace shifted to the origin. We endow $\mathcal{M}$ with a Riemannian inner product induced by the Euclidean inner product. Given a C^2 function $g\colon \mathcal{M} \to \mathbb{R}$, we use the symbol $\nabla^2_{\mathcal{M}} g(x)$ to denote the Riemannian Hessian at x. When $\mathcal{M}$ is affine, $\nabla^2_{\mathcal{M}} g(x)$ is represented by $V^\top \nabla^2 \widetilde{g}(x) V$, where the columns of V form an orthonormal basis of $T_{\mathcal{M}}(x)$ and $\widetilde{g}$ is any C^2 extension of g to a neighborhood of x.

Nonsmooth analysis. A proper closed function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is ρ-weakly convex if $x \mapsto f(x) + \frac{\rho}{2}\|x\|^2$ is convex, where $\rho \geq 0$. For $x \in \mathrm{dom} f$, its subdifferential $\partial f(x)$ consists of the vectors $v \in \mathbb{R}^d$ satisfying

$$f(y) \geq f(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2 \qquad \text{for all } y \in \mathbb{R}^d.$$

We call x a critical point if $0 \in \partial f(x)$. This class includes compositions $h \circ G$ with h convex and L-Lipschitz and G having a β-Lipschitz Jacobian: such functions are $L\beta$-weakly convex [17]. For the empirical objective F_n in (1), we show in Theorem 3.2 that F_n is $2\left\|\frac{1}{n}\sum_{i=1}^n a_i a_i^\top\right\|_{\mathrm{op}}$-weakly convex for every realization of the measurements.

Nonsmooth optimization admits several different notions of second order information [51]. We adopt the notion of an active strict saddle introduced by Davis and Drusvyatskiy [14], which underlies saddle-avoidance guarantees for several first-order methods [4, 21]. To introduce this notion, we require an auxiliary definition.

Definition 2.1 (Active manifold)**.** *Consider a closed, weakly convex function f, and let $\bar{x}$ be a critical point. A set $\mathcal{M}$ containing $\bar{x}$ is an* active C^p manifold *at $\bar{x}$ if there is some neighborhood U of $\bar{x}$ for which the following holds.*

1. *(**Smoothness**) The set $\mathcal{M} \cap U$ is a C^p manifold and the restriction $f|_{\mathcal{M}\cap U}$ is C^p-smooth.*
2. *(**Sharpness**) Subgradients are bounded away from zero off the manifold,*

$$\inf\{\|v\| : v \in \partial f(x),\ x \in U \setminus \mathcal{M}\} > 0.$$

The notion of "active sets" has a long history in nonsmooth optimization [1, 9, 10, 26, 27, 29, 53, 23]. Beyond the definition we follow here, this notion has been modeled in a variety of ways, including identifiable surfaces [60], partial smoothness [43], $\mathcal{VU}$-decompositions [42], and cone reducibility [5]. Furthermore, these notions have led to a range of algorithmic developments [47, 45, 44, 20, 16, 39]. Armed with Definition 2.1 we can now introduce active strict saddles.

Definition 2.2 (Active strict saddles)**.** *A critical point $\bar{x}$ of a weakly convex function f is an* active strict saddle *if it admits an active C^2 manifold $\mathcal{M}$ such that $\nabla^2_{\mathcal{M}}(f|_{\mathcal{M}})(\bar{x})$ has a negative eigenvalue.*[1] *A function f exhibits the* active strict saddle property *if every critical point is either a local minimizer or an active strict saddle.*

Intuitively, at an active strict saddle, the function has a direction of negative curvature along the active manifold and grows sharply away from it. The existence of an active manifold appears crucial for saddle avoidance by simple iterative methods: without one, subgradient methods may converge to critical points that are not local minimizers [14]. Finally, a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ with a minimizer set $\mathcal{X}_\star$ is sharp if there exists a constant $\mu > 0$ such that

$$f(x) - \inf f \geq \mu \operatorname{dist}(x, \mathcal{X}_\star) \quad \text{for all } x \in \mathbb{R}^d.$$

A point $\bar{x} \in \operatorname{dom} f$ is a *sharp local minimizer* if there exist constants $\mu, r > 0$ such that

$$f(x) - f(\bar{x}) \geq \mu \|x - \bar{x}\| \qquad \text{whenever } \|x - \bar{x}\| < r.$$

Whenever f is ρ-weakly convex, μ-sharp and L-Lipschitz, then several first order methods converge fast towards a minimum provided they are initialized near enough [17, 24].

3 The second-order nonsmooth landscape of phase retrieval

In this section, we formally describe our results on the second-order landscape of robust phase retrieval. The first subsection covers the population objective, while the second subsection covers the empirical objective. Throughout, we make the blanket assumption that the signal we wish to recover has norm one, i.e., $\|x^\star\| = 1$. The assumption is made only for convenience, as it simplifies expressions that would otherwise depend on its norm.

3.1 Population landscape

We start by showing that F is C^2-smooth everywhere but at the solutions and that each critical point is either a global minimum or a saddle. The proof of this proposition appears in Appendix A.

Proposition 3.1. *Fix a unit-norm vector $x^\star \in \mathbb{R}^d$ and consider the population objective $F(x) = \mathbb{E}|\langle a, x\rangle^2 - \langle a, x^\star\rangle^2|$ with $a \sim N(0, I_d)$. Then, $F(\cdot)$ is a closed, weakly convex function, and its critical points correspond to*

$$\{\pm x^\star\} \cup \{0\} \cup \mathcal{S} \qquad \text{with} \qquad \mathcal{S} := \big\{\tau u \mid \|u\| = 1 \text{ and } u \perp x^\star\big\}, \tag{2}$$

where τ is the unique solution of the equation $\arctan\tau + \tau/(1+\tau^2) = \pi/4$. The function F is C^2 on $\mathbb{R}^d \setminus \{\pm x^\star\}$, and its Hessian evaluates to

$$\nabla^2 F(0) = -2I_d, \quad \text{and} \quad \nabla^2 F(\tau u) = \frac{16}{\pi} \cdot \frac{\tau}{(1+\tau^2)^2} \cdot (uu^\top - x^\star x^{\star\top}) \quad \text{for all } \tau u \in \mathcal{S}.$$

Further, F is $2/\pi$-sharp with global minimizers $\pm x^\star$.

The characterization of critical points in (2) was first established in [18]. As direct consequences, the origin is a strict local maximizer, while each point $\tau u \in \mathcal{S}$ is a saddle with positive curvature in the radial direction u, negative curvature along $x^\star$, and zero curvature in directions tangent to $\mathcal{S}$. Since F is C^2 near these nonglobal critical points, each is also an active strict saddle, with active manifold equal to the whole space $\mathbb{R}^d$. Together with the sharpness of the global minimizers $\pm x^\star$, this establishes the benign geometry of the population landscape.

3.2 Empirical landscape

In this subsection, we state our main results for the empirical objective F_n, defined in (1). To describe these guarantees, we first define the active index set at a vector $x \in \mathbb{R}^d$ as

$$I(x) = \big\{i \in \{1, \ldots, n\} : \langle a_i, x\rangle^2 = \langle a_i, x^\star\rangle^2\big\}. \tag{3}$$

As it turns out, when $n \geq 2d$, almost surely every critical point $x \notin \{\pm x^\star\}$ has an active manifold encoded via $I(x)$. The proof of the following theorem appears in Appendix B.

[1] This is equivalent to the parabolic-subderivative definition in [14]; see [4, Definition 4 and footnote 2].

Theorem 3.2 (Active strict saddle property)**.** *Fix a unit-norm vector $x^\star \in \mathbb{R}^d$. The empirical objective F_n in (1) is $2\left\|\frac{1}{n}\sum_{i=1}^n a_i a_i^\top\right\|_{\mathrm{op}}$-weakly convex for any realization of the measurements $\{a_i\}$. If $n \geq 2d$, then almost surely every critical point $x_\sharp$ of F_n is either a strict local minimizer or an active strict saddle. In the latter case, the active manifold is*

$$\mathcal{M}_{x_\sharp} = \left\{x \in \mathbb{R}^d \mid \langle a_i, x\rangle = \langle a_i, x_\sharp\rangle \text{ for all } i \in I(x_\sharp)\right\}.$$

Let us make a few remarks. This guarantee, in tandem with the saddle-avoidance results from [4, 21], shows that certain stochastically perturbed variants of the subgradient method can converge only to local minimizers of the empirical objective F_n. Since any subset of d Gaussian vectors $\{a_i\}$ is almost surely linearly independent, at every critical point $x_\sharp \notin \{\pm x^\star\}$ the index set satisfies $|I(x_\sharp)| \leq d$, and moreover, when equality holds, the manifold reduces to a single point and $x_\sharp$ is a sharp local minimizer. This simple observation will be key in our later developments.

Theorem 3.2, together with the benign population landscape described in Proposition 3.1, might give the impression that the empirical objective exhibits benign geometry as well. The next result shows that this is not the case. Recall that $M_{d,n}$ denotes the random variable that counts the number of spurious sharp local minima of F_n. We defer the proof of this guarantee to Appendix C.

Theorem 3.3 (**Existence of spurious sharp local minima**)**.** *Fix a unit-norm vector $x^\star \in \mathbb{R}^d$, consider the empirical objective F_n in (1), and suppose that $d \geq 4$. Then, there exists a positive constant c_d depending solely on d such that the following hold true.*

1. *(**Bounded moments**) The first moment is asymptotically bounded by*

$$\liminf_{n\to\infty} \frac{\mathbb{E}M_{d,n}}{n^{d/2-1}} \geq c_d.$$

2. *(**Constant probability**) Furthermore,*

$$\liminf_{n\to\infty} \mathbb{P}(F_n \text{ has a sharp spurious local minimum}) \geq 0.136.$$

Remarkably, the limiting probability lower bound is uniform over $d \geq 4$. Our proof constructs these sharp local minima within a shrinking tube around the population saddle sphere $\mathcal{S}$, so their distance to this saddle set tends to zero as n increases. The next section outlines the proof strategy. Section 5 studies algorithmic recovery from perturbations of $\mathcal{S}$.

4 Proof strategy

In this section, we sketch our proof strategy for establishing Theorem 3.3. A full proof appears in Appendix C. Before starting, let us provide a roadmap. The key idea is to define a random variable $N_{d,n}$ that counts a selected family of sharp spurious local minima inside a tube, denoted T_n, around the sphere of saddles $\mathcal{S}$. This tube shrinks as n goes to infinity. Since $M_{d,n}$ counts all sharp spurious local minima, it follows that

$$M_{d,n} \geq N_{d,n}, \qquad \text{and so} \qquad \mathbb{P}(M_{d,n} > 0) \geq \mathbb{P}(N_{d,n} > 0). \tag{4}$$

Thus, it suffices to lower bound the probability that our selected family contains at least one minimum. The proof progresses in three steps: (i) define the family counted by $N_{d,n}$, (ii) compute the asymptotic first moment of $N_{d,n}$, (iii) compute the second moment of $N_{d,n}$ and use the Paley–Zygmund inequality to obtain the desired probability bound. Throughout, we assume that $d \geq 4$ is fixed, and let n tend to infinity.

A family of spurious sharp minimizers. The following lemma, proved in Appendix B.1, characterizes critical points and sharp local minima. Recall the active index set $I(x)$ in (3).

Lemma 4.1 (Characterization of critical points and sharp local minima)**.** *Suppose $n \geq 2d$. Then, almost surely, every $x \in \mathbb{R}^d \setminus \{\pm x^\star\}$ satisfies $|I(x)| \leq d$. Such a point is critical if, and only if, there exist scalars $\lambda_i \in (-1, 1)$ with $i \in I(x)$, satisfying*

$$0 = \sum_{i \notin I(x)} \operatorname{sign}(\langle a_i, x\rangle^2 - \langle a_i, x^\star\rangle^2)\,\langle a_i, x\rangle\, a_i + \sum_{i \in I(x)} \lambda_i \langle a_i, x\rangle\, a_i. \tag{5}$$

Moreover, a critical point x is a sharp local minimum if, and only if, it is critical and $|I(x)| = d$.

The lemma implies a simple way to count sharp spurious local minima: enumerate the candidates determined by their active indices and check which are critical. We will use this strategy to bound $N_{d,n}$, but first we will describe the family of sharp local minima that $N_{d,n}$ counts. For $I = \{i_1 < \cdots < i_d\} \subset \{1, \ldots, n\}$ and $s \in \{\pm 1\}^d$, the active equations $a_{i_j}^\top x = s_j a_{i_j}^\top x^\star$, almost surely, determine a unique candidate

$$x_{I,s} = A_I^{-1} \operatorname{diag}(s) A_I x^\star. \tag{6}$$

Here A_I is the matrix with rows $a_{i_j}^\top$. To define $N_{d,n}$, we fix the pattern $s^\circ = (1, \ldots, 1, -1, -1)$, with $d-2$ plus signs, and write $x_I = x_{I,s^\circ}$. Counting sharp minima with only this pattern suffices for a lower bound as in (4). Yet, we will count an even smaller subset of these x_I.

We express the criticality condition (5) in a form that will help us estimate the probability that a candidate x_I is a sharp local minimum. Define $S_I(x) = \sum_{i \notin I} \operatorname{sign}((a_i^\top x)^2 - (a_i^\top x^\star)^2)(a_i^\top x) a_i$ and let $Q_I(x)$ have columns $(a_i^\top x) a_i$, $i \in I$, in increasing index order. By Lemma 4.1, x_I is a sharp local minimum precisely when

$$-S_I(x_I) \in Q_I(x_I)(-1, 1)^d, \tag{7}$$

where the set on the right consists of the vectors $Q_I(x_I)\lambda$ with $\lambda \in (-1, 1)^d$.

To make the probability of (7) easier to estimate, we impose two further restrictions. First, we require the selected measurement vectors $\{a_i\}_{i \in I}$ to have bounded norms and the determinants $|\det A_I|$ and $|\det Q_I(x_I)|$ to be bounded below by positive constants. These bounds prevent the equations defining x_I from becoming nearly dependent and ensure that the set on the right-hand side of (7) has volume bounded away from zero. The constants depend only on d and are chosen so that these bounds hold with positive probability. We do not include them here since their precise values are not particularly insightful. Second, we count only candidates in the tube

$$T_n = \left\{ n^{-1/2} z_1 x^\star + (\tau + n^{-1/2} z_2) u \mid u \perp x^\star,\ \|u\| = 1,\ \text{and } z \in [-1, 1]^2 \right\}.$$

The set T_n is a tube around the population saddle sphere $\mathcal{S}$ that shrinks to $\mathcal{S}$ as $n \to \infty$. The width $n^{-1/2}$ matches the fluctuations exhibited by critical points of the empirical objective [52].

The first moment. For any fixed set I, the probability that x_I is counted in the family described above is exactly the same, thus

$$\mathbb{E} N_{d,n} = \sum_I \mathbb{P}(x_I \text{ is counted}) = \binom{n}{d} \mathbb{P}(x_I \text{ is counted}). \tag{8}$$

We therefore fix one set I and focus on computing this probability. Our plan to do so is to integrate over the possible locations of the candidate x_I. For a proposed location x, write $X_I(x) = A_I x - \operatorname{diag}(s^\circ) A_I x^\star$, so that $X_I(x) = 0$ precisely when $x = x_I$. Let $H_I(x)$ be a boolean function that is equal to one precisely when the selected measurements a_i satisfy our bounds on their norms and associated determinants, and the condition (7) holds with x_I replaced by x, and zero otherwise. The Kac–Rice formula [2] yields the following identity

$$\mathbb{P}(x_I \text{ is counted}) = \int_{T_n} p_{X_I(x)}(0)\, \mathbb{E}[|\det A_I| H_I(x) \mid X_I(x) = 0]\, \mathrm{d}x, \tag{9}$$

where $p_{X_I(x)}(0)$ is the probability density of $X_I(x)$ evaluated at zero.

To describe the asymptotic value of (9), we establish the following two estimates, with the first holding uniformly over $x \in T_n$,

$$\mathbb{E}[|\det A_I| H_I(x) \mid X_I(x) = 0] \asymp_d n^{-d/2}, \qquad \text{and} \qquad \int_{T_n} p_{X_I(x)}(0) dx \asymp_d n^{-1}. \tag{10}$$

Here $\asymp_d$ denotes comparison up to positive constants depending only on d. Before explaining the rationale behind these two results, let us derive an immediate consequence. Recall that $\binom{n}{d} \asymp_d n^d$; combining this fact with (8), (9), and (10) yields $\mathbb{E} N_{d,n} \asymp_d n^d \cdot n^{-d/2} \cdot n^{-1} = n^{d/2-1}$. This establishes the asymptotic lower bound in Theorem 3.3. A detailed computation accounting for the hidden constants yields two positive constants γ_d and $\mathcal{I}_d$ such that

$$n^{-d/2+1} \mathbb{E} N_{d,n} \to \gamma_d \mathcal{I}_d \qquad \text{as } n \to \infty. \tag{11}$$

Here $\gamma_d > 0$ accounts for the selected measurements, while $\mathcal{I}_d > 0$ accounts for the Gaussian density across the tube. Their precise expressions and the full calculation appear in Appendix C.2.

Let us quickly summarize the arguments behind (10). For a fixed location $x \in T_n$, we condition further on the selected measurement vectors $\{a_i\}_{i\in I}$. This fixes the set $Q_I(x)(-1,1)^d$, while the remaining measurements retain their independent standard Gaussian distributions. When our norm and determinant bounds hold, the determinant factor $|\det A_I|$ is bounded above and below by positive constants independent of n, and the set $Q_I(x)(-1,1)^d$ is bounded with volume bounded away from zero. The condition $H_I(x) = 1$ then reduces to asking whether $-S_I(x)$ lies in this fixed set. Each summand in $S_I(x)$ has mean $\frac{1}{2}\nabla F(x)$. Since ∇F vanishes on $\mathcal{S}$ and is Lipschitz continuous nearby, this mean has norm $O(n^{-1/2})$ on T_n, whose points lie within distance $O(n^{-1/2})$ of $\mathcal{S}$. Summing the $n-d$ terms, the mean of $S_I(x)$ therefore has norm of order at most $\sqrt{n}$, uniformly over $x \in T_n$. We can show that its standard deviation in every unit direction is also of order $\sqrt{n}$, uniformly over $x \in T_n$. A local limit theorem implies a Gaussian approximation to the probability of entering this region, yielding a probability of order $n^{-d/2}$ uniformly over the selected measurements satisfying our bounds and the locations in T_n. The first estimate in (10) then follows.

For the second estimate in (10), we consider the volume of T_n. The tube restricts the component along $x^\star$ to within $n^{-1/2}$ of zero and the length of the perpendicular component to within $n^{-1/2}$ of τ. These two restrictions give its volume order n^{-1}. We can prove that the density $p_{X_I(x)}(0)$ is bounded above and below by positive constants on T_n, so its integral has order n^{-1}.

A probability bound uniform in dimension. We use the Paley–Zygmund inequality to turn moment estimates into a lower bound on the probability that $N_{d,n} > 0$, namely

$$\liminf_{n\to\infty} \mathbb{P}(N_{d,n} > 0) \geq \lim_{n\to\infty} \frac{(\mathbb{E}N_{d,n})^2}{\mathbb{E}N_{d,n}^2}. \tag{12}$$

Thus, to derive our lower bound we also asymptotically characterize $\mathbb{E}N_{d,n}^2$. Computing the second moment follows a similar Kac–Rice approach as for the first moment, but the dependence between candidate pairs (x_I, x_J) makes the argument significantly more technical. We defer the proof to Appendix C.3 and simply state here the resulting estimate

$$n^{-d+2}\mathbb{E}N_{d,n}^2 \to \gamma_d^2 \mathcal{J}_d \qquad \text{as } n \to \infty,$$

where $\mathcal{J}_d > 0$ is finite and depends on d. Combining this estimate with (11) and (12) implies

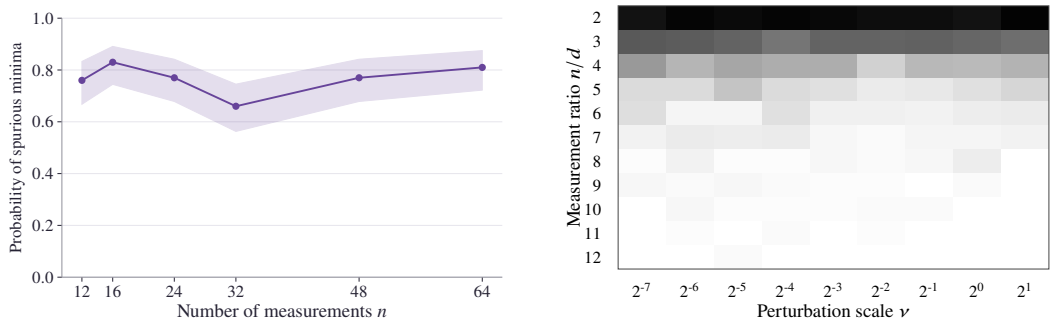
$$\liminf_{n\to\infty} \mathbb{P}(N_{d,n} > 0) \geq \frac{\mathcal{I}_d^2}{\mathcal{J}_d},$$

the γ_d factor cancels exactly. By arguing directly with the definitions of $\mathcal{I}_d$ and $\mathcal{J}_d$ we show that this ratio can be further lower bounded by an explicit quantity $c_\star$ that is independent of the dimension. Combining analytic estimates with certified numerical computations, as described in Appendix C.5, we show that the value of $c_\star$ is lower bounded by 0.1361.

5 Numerical Experiments

In this section, we numerically test our theory. The objective of our experiments is two-fold: (i) to show that for moderate d and n, we observe empirically that the probability of having a spurious sharp minimum is large and (ii) to show that despite our negative results, simple subgradient methods do not often get trapped at these minima. We present two numerical experiments supporting these claims. For both experiments we set $x^\star = e_1$, the first canonical vector. The code to reproduce these experiments can be found at https://anonymous.4open.science/r/Phase-Retrieval-B643.

Probability of having a spurious sharp minimum. We estimate the probability that F_n has a sharp spurious minimum for $d = 4$ and $n \in \{12, 16, 24, 32, 48, 64\}$. For each n, we draw 100 independent problem instances with n i.i.d. Gaussian measurement vectors. For each instance, we enumerate all pairs (I, s) with $|I| = d$ and nonconstant $s \in \{\pm 1\}^d$, and test the candidates $x_{I,s}$ defined in (6). We count an instance as positive if at least one candidate satisfies the criticality



(a) (b)

Figure 2: (a) Estimated probability of a sharp spurious minimum. (b) Recovery frequency of the perturbed subgradient method. White denotes a recovery fraction of one and black denotes zero.

condition (5) up to the numerical tolerances specified in Appendix G. Each estimate is the fraction of positive instances, and the shaded band shows 95% confidence intervals.

Figure 2(a) shows that the estimates change little with n and remain above 0.5 across all tested settings. These results are qualitatively consistent with Theorem 3.3, but suggest that its asymptotic lower bound of 0.136 is conservative. This may reflect the fact that our proof counts only a strict subset of the sharp spurious minima.

Recovery near the saddle sphere. For this experiment, we estimate how often a perturbed subgradient method recovers $x^\star$ up to sign. We fix $d = 100$ and vary the measurement ratio $n/d \in \{2, 3, \dots, 12\}$ and perturbation radius $\nu \in \{2^{-7}, \dots, 2^{1}\}$. For each measurement ratio, we conduct 100 independent trials. In each trial, we draw n i.i.d. Gaussian measurement vectors. For each radius ν, we draw fresh independent $u \sim \mathrm{Unif}(\mathcal{S})$ and $v \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, then initialize $x_0 = u + \nu v$. We run the stochastically perturbed subgradient method analyzed in [21] for at most 5×10^5 updates per initialization, with updates

$$x_{k+1} = x_k - \frac{0.05}{(k+1)^{0.6}}\big(g(x_k) + 0.2\xi_k\big),$$

where $g(x_k) \in \partial F_n(x_k)$ and the vectors $\xi_k \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ are independent. We count a run as recovered if an updated iterate satisfies both $\mathrm{dist}(x_k, \{\pm e_1\}) \le 10^{-3}$ and $F_n(x_k) \le 10^{-4}$.

Figure 2(b) suggests that the perturbation radius $\nu = \|x_0 - u\|$ has little effect on recovery frequency over the tested range. The measurement ratio n/d has a stronger effect, with a sharp increase in recovery near $n/d = 4$. Thus, sharp spurious minima near the saddle $\mathcal{S}$ do not preclude high recovery rates for this method for a sufficiently large number of samples.

6 Conclusion and Limitations

Our main result shows that even for highly structured weakly convex stochastic objectives, benign population geometry need not transfer to empirical risk minimization. As a consequence, active strict saddle theory alone cannot explain the practical success of iterative methods applied to robust phase retrieval. We conjecture that a similar phenomenon occurs in other nonsmooth formulations for low-rank matrix recovery problems. This opens two directions for future research: (i) to understand which other geometric properties enable these iterative methods to find global solutions for these nonsmooth formulations, and (ii) to characterize general geometric conditions of weakly convex stochastic problems under which population-to-empirical transfer holds. Our analysis has several limitations: it covers only noiseless real Gaussian measurements and takes $n \to \infty$ at each fixed dimension. Although the limiting probability bound is independent of d, our proof does not quantify how large n must be relative to d or characterize regimes in which both grow. Extending the analysis to noisy or non-Gaussian measurements would clarify the scope of this obstruction.

AI USE STATEMENT

We established Proposition 3.1 and Theorem 3.2 back in 2021. After failing to prove a population-to-empirical landscape transfer, we were confused as to whether such a result could be true. So, we conjectured the existence of spurious sharp local minima of the empirical objective. This work was not made public at the time because the conjecture remained unresolved. ChatGPT 5.6 Sol developed the core ideas to resolve this conjecture, which ultimately led to Theorem 3.3. After obtaining the initial results, we parsed them, completed and simplified the arguments, and rewrote the proof from scratch with human readability in mind. Appendix H contains a summary of the interactions we had in order to obtain these results. Beyond theorem-proving, we also used ChatGPT 6.0 Astra to proofread the paper, polish the writing, and implement numerical experiments. We take full responsibility for the content of this paper.

ETHICS STATEMENT

This work is theoretical in nature and we do not foresee it having any direct ethical or societal implications.

REPRODUCIBILITY STATEMENT

The code and instructions for reproducing the numerical experiments in Section 5 are provided at https://anonymous.4open.science/r/Phase-Retrieval-B643.

REFERENCES

[1] F. Al-Khayyal and J. Kyparisis. Finite convergence of algorithms for nonlinear programs and variational inequalities. *Journal of Optimization Theory and Applications*, 70(2):319–332, 1991. doi: 10.1007/BF00940629. URL https://doi.org/10.1007/BF00940629.

[2] Jean-Marc Azaïs and Mario Wschebor. *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons, 2009.

[3] Yu Bai, Qijia Jiang, and Ju Sun. Subgradient descent learns orthogonal dictionaries. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HklSf3CqKm.

[4] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research*, 49(3):1761–1790, 2024. doi: 10.1287/moor.2021.0194. URL https://doi.org/10.1287/moor.2021.0194.

[5] J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, New York, 2000. ISBN 978-0-387-98705-7. doi: 10.1007/978-1-4612-1394-9. URL https://doi.org/10.1007/978-1-4612-1394-9.

[6] Alexander Alekseevich Borovkov and Anatolii Al'fredovich Mogul'skii. Integro-local theorems for sums of independent random vectors in the series scheme. *Mathematical Notes*, 79 (4):468–482, 2006.

[7] Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4): 2355–2377, 2016. doi: 10.1137/16M105808X. URL https://doi.org/10.1137/16M105808X.

[8] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book.

[9] James V. Burke and Jorge J. Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988. doi: 10.1137/0725068. URL https://doi.org/10.1137/0725068.

[10] Jim Burke. On the identification of active constraints II: The nonconvex case. *SIAM Journal on Numerical Analysis*, 27(4):1081–1102, 1990. doi: 10.1137/0727064. URL https://doi.org/10.1137/0727064.

[11] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: Good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021. doi: 10.1007/s10208-020-09490-9. URL https://doi.org/10.1007/s10208-020-09490-9.

[12] Vasileios Charisopoulos, Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Composite optimization for robust rank one bilinear sensing. *Information and Inference: A Journal of the IMA*, 10(2):333–396, 2021. doi: 10.1093/imaiai/iaaa027. URL https://doi.org/10.1093/imaiai/iaaa027.

[13] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019. doi: 10.1137/18M1178244. URL https://doi.org/10.1137/18M1178244.

[14] Damek Davis and Dmitriy Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22(2):561–606, 2022. doi: 10.1007/s10208-021-09516-w. URL https://doi.org/10.1007/s10208-021-09516-w.

[15] Damek Davis and Dmitriy Drusvyatskiy. Graphical convergence of subgradients in nonconvex optimization and learning. *Mathematics of Operations Research*, 47(1):209–231, 2022. doi: 10.1287/moor.2021.1126.

[16] Damek Davis and Liwei Jiang. A local nearly linearly convergent first-order method for nonsmooth functions with quadratic growth. *Foundations of Computational Mathematics*, 25 (3):943–1024, 2025. doi: 10.1007/s10208-024-09653-y. URL https://doi.org/10.1007/s10208-024-09653-y.

[17] Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018. doi: 10.1007/s10957-018-1372-8. URL https://doi.org/10.1007/s10957-018-1372-8.

[18] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

[19] Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. *SIAM Journal on Optimization*, 32(3): 1958–1983, 2022. doi: 10.1137/21M1430868. URL https://doi.org/10.1137/21M1430868.

[20] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Asymptotic normality and optimality in nonsmooth stochastic approximation. *The Annals of Statistics*, 52(4):1485–1508, 2024. doi: 10.1214/24-AOS2401. URL https://doi.org/10.1214/24-AOS2401.

[21] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Active manifolds, stratifications, and convergence to local minima in nonsmooth optimization. *Foundations of Computational Mathematics*, 26(2):779–861, 2026. doi: 10.1007/s10208-025-09691-0.

[22] Mateo Díaz. The nonsmooth landscape of blind deconvolution. In *11th Annual Workshop on Optimization for Machine Learning*, 2019. URL https://opt-ml.org/papers/2019/paper_20.pdf.

[23] Dmitriy Drusvyatskiy and Adrian S Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147(1–2):467–498, 2014. doi: 10.1007/s10107-013-0730-4. URL https://doi.org/10.1007/s10107-013-0730-4.

[24] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019. doi: 10.1007/s10107-018-1311-3. URL https://doi.org/10.1007/s10107-018-1311-3.

[25] John C. Duchi and Feng Ruan. Solving (most) of a set of quadratic equalities: composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019. doi: 10.1093/imaiai/iay015. URL https://doi.org/10.1093/imaiai/iay015.

[26] J. C. Dunn. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications*, 55(2):203–216, 1987. doi: 10.1007/BF00939081. URL https://doi.org/10.1007/BF00939081.

[27] Michael C. Ferris. Finite termination of the proximal point algorithm. *Mathematical Programming*, 50:359–366, 1991. doi: 10.1007/BF01594944. URL https://doi.org/10.1007/BF01594944.

[28] James R. Fienup. Phase retrieval algorithms: a comparison. *Applied Optics*, 21(15):2758–2769, 1982. doi: 10.1364/AO.21.002758. URL https://doi.org/10.1364/AO.21.002758.

[29] Sjur D. Flåm. On finite convergence and constraint identification of subgradient projection methods. *Mathematical Programming*, 57:427–437, 1992. doi: 10.1007/BF01581092. URL https://doi.org/10.1007/BF01581092.

[30] Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/59ab3ba90ae4b4ab84fe69de7b8e3f5f-Abstract.html.

[31] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 797–842. PMLR, 2015. URL https://proceedings.mlr.press/v40/Ge15.html.

[32] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1233–1242. PMLR, 2017. URL https://proceedings.mlr.press/v70/ge17a.html.

[33] Jiewen Guan and Anthony Man-Cho So. ℓ_1-norm rank-one symmetric matrix factorization has no spurious second-order stationary points. *arXiv preprint arXiv:2410.05025*, 2024.

[34] Mourad E. H. Ismail and Jim Pitman. Algebraic evaluations of some Euler integrals, duplication formulae for Appell's hypergeometric function F_1, and Brownian variations. *Canadian Journal of Mathematics*, 52(5):961–981, 2000. doi: 10.4153/CJM-2000-040-3.

[35] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732. PMLR, 2017. URL https://proceedings.mlr.press/v70/jin17a.html.

[36] Chi Jin, Lydia T. Liu, Rong Ge, and Michael I. Jordan. On the local minima of the empirical risk. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.neurips.cc/paper_files/paper/2018/hash/da4902cb0bc38210839714ebdcf0efc3-Abstract.html.

[37] Cédric Josz and Lexiao Lai. Nonsmooth rank-one matrix factorization landscape. *Optimization Letters*, 16(6):1611–1631, 2022.

[38] Cédric Josz and Lexiao Lai. Nonsmooth rank-one symmetric matrix factorization landscape. *Optimization Letters*, 19(8):1691–1698, 2025.

[39] Siyu Kong and Adrian S. Lewis. Lipschitz minimization and the Goldstein modulus. *Mathematical Programming*, 219:315–344, 2026. doi: 10.1007/s10107-025-02262-9. URL https://doi.org/10.1007/s10107-025-02262-9.

[40] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1–2):311–337, 2019. doi: 10.1007/s10107-019-01374-3. URL https://doi.org/10.1007/s10107-019-01374-3.

[41] John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, NY, second edition, 2012. doi: 10.1007/978-1-4419-9982-5. URL https://doi.org/10.1007/978-1-4419-9982-5.

[42] Claude Lemaréchal, François Oustry, and Claudia Sagastizábal. The $\mathcal{U}$-lagrangian of a convex function. *Transactions of the American Mathematical Society*, 352(2):711–729, 2000. doi: 10.1090/S0002-9947-99-02243-6. URL https://doi.org/10.1090/S0002-9947-99-02243-6.

[43] Adrian S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2002. doi: 10.1137/S1052623401387623. URL https://doi.org/10.1137/S1052623401387623.

[44] Adrian S. Lewis and Calvin Wylie. Active-set Newton methods and partial smoothness. *Mathematics of Operations Research*, 46(2):712–725, 2021. doi: 10.1287/moor.2020.1075. URL https://doi.org/10.1287/moor.2020.1075.

[45] Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017. doi: 10.1137/16M106340X. URL https://doi.org/10.1137/16M106340X.

[46] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018. doi: 10.1214/17-AOS1637.

[47] Robert Mifflin and Claudia Sagastizábal. A $\mathcal{VU}$-algorithm for convex minimization. *Mathematical Programming*, 104(2–3):583–608, 2005. doi: 10.1007/s10107-005-0630-3. URL https://doi.org/10.1007/s10107-005-0630-3.

[48] Andreea-Alexandra Muşat and Nicolas Boumal. Gradient descent avoids strict saddles with a simple line-search method too. *arXiv preprint arXiv:2507.13804*, 2025.

[49] Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969. doi: 10.1016/0041-5553(69)90061-5.

[50] Junwen Qiu, Bohao Ma, Andre Milzarek, and Junyu Zhang. Stochastic saddle avoidance beyond unit excitation and smoothness: A pathwise Lyapunov–Perron framework. *arXiv preprint arXiv:2608.03001*, 2026. URL https://arxiv.org/abs/2608.03001.

[51] R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

[52] Feng Ruan. On the uniform convergence of subdifferentials in stochastic optimization and learning. *Mathematics of Operations Research*, 51(3):2323–2353, 2026. doi: 10.1287/moor.2024.0533. URL https://doi.org/10.1287/moor.2024.0533.

[53] Alexander Shapiro. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28(4):677–692, 2003. doi: 10.1287/moor.28.4.677.20512. URL https://doi.org/10.1287/moor.28.4.677.20512.

[54] Alexander Shapiro and Huifu Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *Journal of Mathematical Analysis and Applications*, 325 (2):1390–1399, 2007. doi: 10.1016/j.jmaa.2006.02.078.

[55] Yoav Shechtman, Yonina C. Eldar, Oren Cohen, Henry N. Chapman, Jianwei Miao, and Mordechai Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2015. doi: 10.1109/MSP.2014.2352673. URL https://doi.org/10.1109/MSP.2014.2352673.

[56] Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. doi: 10.1007/s10208-017-9365-9. URL https://doi.org/10.1007/s10208-017-9365-9.

[57] Lai Tian and Johannes O. Royset. Failure of uniform laws of large numbers for subdifferentials and beyond. arXiv preprint arXiv:2511.16568, 2025. URL https://arxiv.org/abs/2511.16568. Revised 16 March 2026.

[58] Lai Tian and Johannes O. Royset. Lipschitzian SLLNs for random functions. arXiv preprint arXiv:2607.20411, 2026. URL https://arxiv.org/abs/2607.20411.

[59] Kaizheng Wang, Yuling Yan, and Mateo Díaz. Efficient clustering for stretched mixtures: Landscape and optimality. *Advances in Neural Information Processing Systems*, 33:21309–21320, 2020.

[60] Stephen J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993. doi: 10.1137/0331048. URL https://doi.org/10.1137/0331048.

[61] Zhong Zheng and Lingzhou Xue. Smoothed robust phase retrieval. *arXiv preprint arXiv:2409.01570*, 2024.

[62] Zhihui Zhu, Daniel Soudry, Yonina C. Eldar, and Michael B. Wakin. The global optimization geometry of shallow linear neural networks. *Journal of Mathematical Imaging and Vision*, 62(3):279–292, 2020. doi: 10.1007/s10851-019-00889-w. URL https://doi.org/10.1007/s10851-019-00889-w.