Subgradient Convergence Implies Subdifferential Convergence on Weakly Convex Functions: With Uniform Rates Guarantees

Feng Ruan

*Department of Statistics and Data Science, Northwestern University, Evanston, 60202, Illinois, United States of America.

Corresponding author(s). E-mail(s): fengruan@northwestern.edu;

Abstract

In nonsmooth, nonconvex stochastic optimization, understanding the uniform convergence of subdifferential mappings is crucial for analyzing stationary points of sample average approximations of risk as they approach the population risk. Yet, characterizing this convergence remains a fundamental challenge. This work introduces a novel perspective by connecting the uniform convergence of subdifferential mappings to that of subgradient mappings as empirical risk converges to the population risk. We prove that, for stochastic weakly-convex objectives, and within any open set, a uniform bound on the convergence of subgradients—chosen arbitrarily from the corresponding subdifferential sets—translates to a uniform bound on the convergence of the subdifferential sets itself, measured by the Hausdorff metric. Using this technique, we derive uniform convergence rates for subdifferential sets of stochastic convex-composite objectives. Our results do not rely on key distributional assumptions in the literature, which require the population and finite sample subdifferentials to be continuous in the Hausdorff metric, yet still provide tight convergence rates. These guarantees lead to new insights into the nonsmooth landscapes of such objectives within finite samples.

Keywords: Subdifferential, Subgradient, Uniform Convergence, Population Risk, Sample Average Approximation, Weak Convexity, Hausdorff Metric.

1 Introduction

At the core of nonsmooth, nonconvex stochastic optimization lies a fundamental yet underexplored challenge—characterizing the statistical uniform convergence behavior of set-valued subdifferential mappings [1]. This convergence is essential for understanding how the empirical risk, derived from sample average approximation, approximates the true population risk. While results on the uniform convergence of objective values provide important insights, they do not directly characterize the convergence behaviors of stationary points formulated via subdifferentials. Analyzing the uniform convergence of these set-valued subdifferential mappings of the empirical risk is crucial for a comprehensive understanding of statistical estimation and the behaviors of numerical algorithms in stochastic programming [2, Section 7.2]. Despite its importance, this characterization of subdifferential set convergence remains elusive, warranting further investigation (see Section 1.1 for related works).

This work addresses this gap by introducing a new perspective to understand the uniform convergence behavior of subdifferential sets, specifically for stochastic weakly convex objectives—a prominent class of nonsmooth, nonconvex objectives with widespread applications [3–5]. Our approach hinges on a fundamental property of weakly convex functions. Specifically, for any pair of weakly convex functions, we establish that over any arbitrary open set within their domains, the uniform bound on the distance between their respective set-valued subdifferential mappings, measured via the Hausdorff metric, is upper bounded by the uniform bound on the distance between any corresponding vector-valued subgradient mappings of these functions (Theorem 1). This deterministic result immediately implies a crucial connection between the uniform convergence of subdifferentials and the uniform convergence of subgradient mappings in sample average approximations, enabling the analysis of the former through the latter (Theorem 2).

We further apply our technique to an important subclass of stochastic weakly-convex objectives called stochastic convex-composite objectives [6, 7], which include robust phase retrieval [8], blind deconvolution [9], matrix sensing [10, 11], conditional Value-at-Risk [12], and nonlinear quantile regressions [13], commonly encountered in the field of machine learning, statistics, imaging, and risk management (Theorem 3). We establish uniform convergence of subdifferential sets of stochastic weakly-convex objectives by analyzing the convergence behavior of stochastic subgradient mappings (Theorem 4 and Theorem 5). Our result, which does not rely on a common distributional assumption in the literature requiring the subdiffrential mappings to be continuous under Hausdorff metric [1, Theorem 3], provides tight convergence rates in the context of many existing studies (see Section 1.1 for a detailed explanation). For instance, in robust phase retrieval [8, 14], our results (Theorem 6) achieve a tight $\sqrt{d/m}$ rate (modulo logarithmic factors on d, m) for uniform convergence of subdifferential mappings under the Hausdorff distance where d is the dimension and m is the sample size. This is complementary to the state-of-the-art $\sqrt[4]{d/m}$ rate, which is established under the graphical distance of subdifferentials, through a different technique based on Attouch's epigraphical convergence theorem [15, 16]. This leads to new understanding of the nonsmooth landscape of these objectives, e.g., the landscape of the finite sample robust phase retrieval initially studied in [14], including the locations and convergence properties of stationary points.

1.1 Related Work and Contributions

This work contributes to the extensive literature on characterizing the behavior of uniform convergence of subdifferentials in sample average approximation settings, for which we review. Notably, we saturate our contributions to existing work focused on settings where the objective functions defining the risks are simultaneously nonconvex and nonsmooth.

A significant body of literature has utilized set-valued variational analysis [17] and random set theory [18] to study the uniform convergence of subdifferentials from empirical to population risks [1]. This technique is popular because subdifferential mappings of many nonsmooth and nonconvex functions—including weakly convex functions—often exhibit structural properties such as outer semicontinuity, and the subdifferentials themselves form convex sets, unlike discontinuous subgradient mappings [17]. This technique discretizes the parameter space, achieving initial subdifferential convergence at discrete points via Artstein and Vitale's strong law of large numbers for random convex sets [19]. The core assumption is then the subdifferential mapping of both population and the finite sample risks are *continuous* under the Hausdorff metric. This continuity—not merely outer semicontinuity—is crucial, see [1, Remark 2], as it enables the extension of the initial convergence from discrete points to uniform convergence of subdifferential sets over the entire parameter space. Without continuity, using outer-semicontinuity this approach can still ensure uniform convergence of *enlarged* empirical subdifferentials to *enlarged* population subdifferentials; however, this does not apply to the *original* subdifferentials [1].

Therefore, a common and core assumption of this approach to attain uniform convergence is the continuity needed for both the population and empirical subdifferentials; while it facilitates many applications, it is sometimes restrictive and does not capture the full picture of subdifferential convergence [1, Remark 2]. For achieving uniform convergence with a rate guarantee, stronger assumptions, such as Hölder continuity of subdifferentials or related H-calmness conditions, are often necessary [20]. Some research aims to relax the requirement to piecewise continuity of subdifferential sets over disjoint regions, but this still necessitates the underlying probability measure being nonatomic to ensure that samples almost surely avoid the discontinuity boundaries of the subdifferentials [21].

This work presents a novel perspective to establish uniform convergence of subdifferential mappings that distinguishes it apart from the existing literature. Specifically, our work targets the domain of stochastic weakly-convex classes, pivotal to numerous data-driven applications [5]. The fundamental principle of our approach is that, for stochastic weakly convex objectives, achieving uniform convergence—and the associated rates—of subdifferential sets using sample average approximation requires only the uniform convergence and rates of any chosen pair of the subgradient mappings of the corresponding empirical and population risk functions. As a result, we can derive uniform convergence (and also rates) for a specific subclass termed stochastic convex-composite objectives, without the need for continuity assumptions on the subdifferential mappings of either population or finite sample objectives often required in the existing approach. Importantly, obtaining uniform convergence for discontinuous subgradient mappings in stochastic convex-composite objectives still demands sophisticated tools from statistical learning theory, such as chaining, the concept of shattering, and the notion of Vapnik–Chervonenkis dimension [22].

Notably, there is a recent advance in the literature that also targets at establishment of uniform subdifferential convergence within the domain of stochastic weakly convex classes [14, 16]. However, there are significant differences between our results and their results. First, there is a subtle difference in terms of the criteria for convergence. Our notion of uniform convergence on any open domain implies their notion of graphical convergence of subdifferential mappings on the same domain, but the reverse is not generally true (see Remark of Theorem 1). Therefore, our results are established under a topologically weaker notion of convergence. Second, there is an important difference in the approaches and the established rates. Their approach, grounded in the utilization of Moreau envelope smoothing [23], and Attouch's epigraphical convergence theorem [15], achieves a convergence rate whose dependence on the sample size m is often at $m^{-1/4}$. The reason is that the core principle behind their approach, which aims to derive graphical subdifferential distance bounds through the closeness of objective values, often inherently results in suboptimal guarantees in terms of the sample size [16, Section 5]. Our results are notably precise regarding the sample size $m^{-1/2}$, although they apply mainly to stochastic convex-composite objectives and introduce additional logarithmic factors. This indicates that further research is essential to fully delineate the advantages and limitations of both approaches.

Finally, there are many interesting alternative ideas in the literature based on directly smoothing the nonsmooth objective to attain uniform convergence of gradients of the smoothed objectives [24], also used extensively in the statistics literature, e.g., in the context of quantile regressions [25]. These results, however, do not directly characterize the convergence behavior of the original subdifferential sets.

1.2 Roadmap and Paper Organizations

Section 2 provides the basic notation and definitions. Section 3 describes the general principle for analyzing the uniform convergence behavior for the subdifferential sets for weakly convex functions. Section 4 characterizes the explicit uniform convergence rates of the subdifferential sets for stochastic convex-composite minimizations. Section 5 gives concrete applications illustrating our techniques on the problem of robust phase retrieval. The remainders are proofs and discussions.

2 Notation and Basic Definitions

For $x, y \in \mathbb{R}$, we let $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. In $\mathbb{R}^d$, we use $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean norm and inner product respectively. We let $\mathbb{B}(x_0; r) = \{x : \|x - x_0\| < r\}$ denote the open ℓ_2 ball with center at $x_0 \in \mathbb{R}^d$ and radius r. For matrices in $\mathbb{R}^{d\times d}$, we identify each such matrix with a vector in $\mathbb{R}^{d^2}$ and define $\langle A, B\rangle$ as $\mathrm{tr}(AB^T)$, where tr denotes the trace. For a set X, we use cl, int to denote the closure and interior respectively for the set X. For a closed convex set X, we use ι_X to denote the $+\infty$-valued indicator for the set X, that is $\iota_X(x) = 0$ if $x \in X$ and $+\infty$ otherwise. The normal cone to X at x is $\mathcal{N}_X(x) := \{v \in \mathbb{R}^d : \langle v, y - x\rangle \le 0 \text{ for all } y \in X\}$.

For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, its epigraph is the set $\{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} \mid \alpha \ge f(x)\}$. It is called closed if its epigraph is closed in $\mathbb{R}^d \times \mathbb{R}$. It is called subdifferentially regular at x if $f(x)$ is finite and its epigraph is Clarke regular at $(x, f(x))$ as a subset of $\mathbb{R}^d \times \mathbb{R}$ [17, Definition 7.25]. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be $\mathcal{C}^1$ smooth on $\mathbb{R}^d$ if it has continuous gradients $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$.

2.1 Weakly Convex Functions and Subdifferentials

We say $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is λ locally weakly-convex near x (also known as lower-$\mathcal{C}^2$ [17] or semiconvex [26]) if there exists $\epsilon > 0$ such that

$$y \mapsto f(y) + \frac{\lambda}{2}\|y\|^2, \quad y \in \mathbb{B}(x; \epsilon)$$

is convex [17, Chapter 10.G]. We say a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a locally weakly convex function on $\mathcal{O} \subseteq \mathbb{R}^d$ if at every $x \in \mathcal{O}$, there is $\lambda < \infty$ such that f is λ locally weakly-convex near x. For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and a point x with $f(x)$ finite, we let $\partial f(x)$ denote the Fréchet subdifferential (or regular [17, Chapter 8.B]) of f at the point x,

$$\partial f(x) = \left\{g \in \mathbb{R}^d : f(y) \ge f(x) + \langle g, y - x\rangle + o(\|y - x\|) \ \text{ as } \ y \to x\right\}.$$

Here, any $g \in \partial f(x)$ is referred to as a subgradient of f at the point x. We say v is a *horizon subgradient* of f at x with $f(x)$ finite, written as $v \in \partial^\infty f(x)$ if there exist sequences $x_i, v_i \in \partial f(x_i)$ and $\tau_i \to 0$ satisfying $(x_i, f(x_i), \tau_i v_i) \to (x, f(x), v)$ [17, Chapter 8.B].

Notably, when f is smooth, the subdifferential $\partial f(x)$ consists only of the gradient $\{\nabla f(x)\}$, while for convex function, it coincides with the subdifferential in convex analysis [27]. For a weakly convex function f, the subdifferential $\partial f(x)$ is non-empty, compact, and convex for every x in the interior of $\mathrm{dom} f$. Notably, these follow from the corresponding results for convex functions [17, Chapter 8].

2.2 Set-valued Analysis

Our definitions follow the references of Rockafellar and Wet [17], and Aubin and Frankowska [28]. For a set $A \subseteq \mathbb{R}^d$, we denote $\|A\| = \sup_{a\in A}\|a\|$.

For a set $A \subseteq \mathbb{R}^d$ and $y \in \mathbb{R}^d$, we denote by $\mathrm{dist}(y, A) = \inf_{z\in A}\|y - z\|$ the distance from y to A with respect to the Euclidean norm $\|\cdot\|$. For two sets $A_1, A_2 \subseteq \mathbb{R}^d$, we denote by

$$\mathbb{D}(A_1, A_2) = \sup_{x\in A_1} \mathrm{dist}(x, A_2)$$

the deviation of the set A_1 from the set A_2, by

$$\mathbb{H}(A_1, A_2) = \max\{\mathbb{D}(A_1, A_2), \mathbb{D}(A_2, A_1)\}$$

the Hausdorff distance between A_1 and A_2. Given a sequence of sets $A_n \subseteq \mathbb{R}^d$, the limit supremum of the sets consists of limit points of subsequences $y_{n_k} \in A_{n_k}$, that is,

$$\limsup_n A_n = \{y : \exists n_k, \exists y_{n_k} \in A_{n_k} \text{ s.t. } y_{n_k} \to y \text{ as } k \to \infty\}.$$

The limit infimum of the sets consists of limit points of sequences $y_n \in A_n$, that is:

$$\liminf_n A_n = \{y : \exists y_n \in A_n \text{ s.t. } y_n \to y \text{ as } n \to \infty\}.$$

We let $G : X \rightrightarrows \mathbb{R}^d$ denote a set-valued mapping from X to $\mathbb{R}^d$, and $\text{dom}(G) := \{x : G(x) \neq \emptyset\}$. A function $g : X \to \mathbb{R}^d$ is said to be a selection of $G : X \rightrightarrows \mathbb{R}^d$ if $g(x) \in G(x)$ for every $x \in X$. We say G is outersemicontinuous if for any sequence $x_n \to x \in \text{dom}(G)$, we have $\limsup_n G(x_n) \subseteq G(x)$. We say G is innersemicontinuous if for any $x_n \to x \in \text{dom}(G)$, we have $\liminf_n G(x_n) \supseteq G(x)$. We say $G : X \rightrightarrows \mathbb{R}^d$ is continuous if it is both outersemicontinuous and innersemicontinuous.

For every weakly convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the subgradient mapping $\partial f : \text{int dom}(f) \rightrightarrows \mathbb{R}^d$ is outersemicontinuous [17, Chapter 8].

2.3 Probability Space and Random Set-Valued Map

Let $(\Xi, \mathscr{G}, \mathbb{P})$ denote a probability space. We frequently consider random set-valued mapping of the form $\mathcal{A} : X \times \Xi \rightrightarrows \mathbb{R}^d$. Let $\mathscr{B}$ denote the space of nonempty, compact subsets of $\mathbb{R}^d$. By [17, Theorem 14.4], we say the mapping $\xi \mapsto \mathcal{A}(x, \xi)$ is measurable if and only if for every $\mathcal{S} \in \mathscr{B}$, $\mathcal{A}(x, \cdot)^{-1}\mathcal{S} = \{\xi : A(x, \xi) \in \mathcal{S}\}$ is a measurable set under $(\Xi, \mathscr{G}, \mathbb{P})$.

By a selection of the random set $\mathcal{A}(x, \xi)$, we refer to a random vector $a(x, \xi) \in \mathcal{A}(x, \xi)$, meaning that $\xi \mapsto a(x, \xi)$ is measurable. Note that such selection exists if $\xi \mapsto \mathcal{A}(x, \xi)$ is closed-valued and measurable [29], see also [28, Theorem 8.4]. For a map $\mathcal{A} : X \times \Xi \rightrightarrows \mathbb{R}^d$ and a probability measure $\mathbb{P}$, we define its expectation, following Aumann [29]. At every $x \in X$:

$$\int_\Xi \mathcal{A}(x, \xi) P(d\xi) := \left\{ \int_\Xi a(x, \xi) P(d\xi) \mid a(x, \xi) \in \mathcal{A}(x, \xi) \text{ for } \xi \in \Xi,\ a(x, \cdot) \text{ integrable} \right\}.$$

We say $f : \mathbb{R}^d \times \Xi \to \mathbb{R} \cup \{+\infty\}$ is a random function when $\xi \mapsto f(x, \xi)$ is measurable at every $x \in X$ [2, Chapter 7]. Suppose a random function f also satisfies $x \mapsto f(x, \xi)$ is locally weakly convex, and real-valued near a point $x \in \mathbb{R}^d$, then the subdifferential $\partial f(x, \xi)$ satisfies $\xi \mapsto \partial f(x, \xi)$ is measurable at that $x \in \mathbb{R}^d$, following [2, Section 7.2.6] and [24, Proposition 2.1].

2.4 Measurability Issues

This subsection can be mostly skipped on a first read, but it is essential for addressing measurability issues that arise when considering the supremum of random functions over an uncountable index set [30]. This is important because we are interested in the supremum of the Hausdorff distance between stochastic subdifferential mappings over a given domain.

To address these measurability issues, we use the concept of outer measure $\mathbb{P}^*$ and inner measure $\mathbb{P}_*$ frequently adopted in empirical process theory [31, Section 1.2-5]. Given a probability space $(\Xi, \mathscr{G}, \mathbb{P})$, every subset $B \subseteq \Xi$ is assigned with an outer measure:

$$\mathbb{P}^*(B) = \inf\{\mathbb{P}(A) : B \subseteq A, A \in \mathscr{G}\}.$$

Notably, $\mathbb{P}^*$ is subadditive: $\mathbb{P}^*(B_1 \cup B_2) \le \mathbb{P}^*(B_1) + \mathbb{P}^*(B_2)$, following the union bound. Correspondingly, the inner measure is $\mathbb{P}_*(B) = \sup\{\mathbb{P}(A) : B \supseteq A, A \in \mathscr{G}\}$. Clearly, $\mathbb{P}_*(B) = 1 - \mathbb{P}^*(\Xi \backslash B)$. Finally, when $B \in \mathscr{G}$ is measurable, then the inner and outer measure agree: $\mathbb{P}^*(B) = \mathbb{P}_*(B) = \mathbb{P}(B)$.

3 General Theory

3.1 Uniform Subdifferential Bounds via Selections

In this section, we introduce Theorem 1, a general technique for obtaining uniform bounds on the subdifferentials between two real-valued, locally weakly convex functions. The theorem states that the supremum of the Hausdorff distance between their subdifferentials over any open set is upper bounded by the supremum of the norm differences between *any* selected subgradient mappings.

Theorem 1 might initially seem surprising because at any single point x, the Hausdorff distance between the subdifferentials at x can exceed the norm difference between the selected subgradients. To further appreciate this result, we will provide counterexamples in the remark to highlight why the assumption of $\mathcal{O}$ being an open set is crucial.

We note that in the (weakly) convex settings, graphical distance between subdifferentials under the Hausdorff metric is typically established through controlling the epi-distance of functions due to Attouch's celebrated theorem [15, 16, 32]. See, e.g., [17, Theorem 12.35]. We shall make comparisons to clarify two distance metrics of subdifferential mappings in the second remark: the supremum of the Hausdorff distance between subdifferential mappings over x in an open set $\mathcal{O}$, and the graphical distance between subdifferential mappings over the same open set $\mathcal{O}$.

Theorem 1. *Let f_1 and f_2 be locally weakly convex functions from $\mathcal{O}$ to $\mathbb{R}$, where $\mathcal{O}$ is an open set in $\mathbb{R}^d$. Let g_1 and g_2 be selections of the subdifferentials ∂f_1 and ∂f_2 on $\mathcal{O}$, respectively, i.e., obeying $g_1(x) \in \partial f_1(x)$ and $g_2(x) \in \partial f_2(x)$ for all $x \in \mathcal{O}$. Then the following inequality holds:*

$$\sup_{x\in\mathcal{O}} \mathbb{H}(\partial f_1(x), \partial f_2(x)) \le \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| . \tag{3.1}$$

Proof. Our proof is structured into two parts. In the first part, we demonstrate that Theorem 1 holds under the assumption that $\mathcal{O}$ is open and *convex*, and that both f_1 and f_2 are *convex* on $\mathcal{O}$.

In the second part, we address the general case where f_1 and f_2 are *locally weakly convex* on an open set $\mathcal{O}$. We show that this scenario can be reduced to the convex case, thus concluding our proof of Theorem 1.

Part I. In this part, we address the case where f_1 and f_2 are both real-valued convex functions on an open and convex set $\mathcal{O}$. Recall the definition of Hausdorff distance:

$$\mathbb{H}(\partial f_1(x), \partial f_2(x)) = \max\{\mathbb{D}(\partial f_1(x), \partial f_2(x)), \mathbb{D}(\partial f_2(x), \partial f_1(x))\}.$$

By symmetry, it suffices to prove $\sup_{x\in\mathcal{O}} \mathbb{D}(\partial f_1(x), \partial f_2(x)) \le \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\|$, or equivalently,

$$\sup_{x\in\mathcal{O}} \sup_{y_1\in\partial f_1(x)} \inf_{y_2\in\partial f_2(x)} \|y_1 - y_2\| \le \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| . \tag{3.2}$$

By applying convex duality, we obtain:

$$\inf_{y_2\in\partial f_2(x)} \|y_1 - y_2\| = \inf_{y_2\in\partial f_2(x)} \sup_{v:\|v\|\le 1} \langle y_1 - y_2, v\rangle = \sup_{v:\|v\|\le 1} \inf_{y_2\in\partial f_2(x)} \langle y_1 - y_2, v\rangle, \tag{3.3}$$

where the interchangeability between inf and sup in the second identity is due to Sion's minimax theorem [33]. By substituting this identity back into equation (3.2), our goal reduces to prove:

$$\sup_{x\in\mathcal{O}} \sup_{y_1\in\partial f_1(x)} \sup_{v:\|v\|\le 1} \inf_{y_2\in\partial f_2(x)} \langle y_1 - y_2, v\rangle \le \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| .$$

This ultimately reduces to proving for any $x \in \mathcal{O}$ and $y_1 \in \partial f_1(x)$, and any v with $\|v\| \leq 1$:

$$\inf_{y_2 \in \partial f_2(x)} \langle y_1 - y_2, v \rangle \leq \sup_{x \in \mathcal{O}} \|g_1(x) - g_2(x)\| . \tag{3.4}$$

To prove this inequality, we invoke a perturbation argument standard in convex analysis. We start by taking a sequence $t_n \to 0$ as $n \to \infty$ with $t_n > 0$ and $x + t_n v \in \mathcal{O}$ per $n \in \mathbb{N}$. Such a sequence t_n exists because $\mathcal{O}$ is open. Next, we pick any cluster point $\tilde{y}_2$ of the sequence $\{g_2(x + t_n v)\}_{n \in \mathbb{N}}$. Given that $g_2(x+t_n v) \in \partial f_2(x+t_n v)$, and the subdifferential map $z \mapsto \partial f_2(z)$ is locally bounded [34, Proposition 6.2.2], a cluster point $\tilde{y}_2$ necessarily exists in $\mathbb{R}^d$. Furthermore, because the subdifferential map $z \mapsto \partial f_2(z)$ is outer semicontinuous [34, Theorem 6.2.4], $\tilde{y}_2$ must belong to the set $\partial f_2(x)$. Finally, by selecting a subsequence of t_n if necessary, we can assume W.L.O.G. that the following limit holds:

$$\lim_{n \to \infty} g_2(x + t_n v) = \tilde{y}_2.$$

Given convexity of f_1, $y_1 \in \partial f_1(x)$ and $g_1(x + t_n v) \in \partial f_1(x + t_n v)$, we obtain for every $n \in \mathbb{N}$:

$$\langle y_1, v \rangle \leq \frac{1}{t_n}(f_1(x + t_n v) - f_1(x)) \leq \langle g_1(x + t_n v), v \rangle.$$

Thus, we have for every v with $\|v\| \leq 1$:

$$\inf_{y_2 \in \partial f_2(x)} \langle y_1 - y_2, v \rangle \leq \langle y_1 - \tilde{y}_2, v \rangle \leq \liminf_{n \to \infty} \langle g_1(x + t_n v) - g_2(x + t_n v), v \rangle,$$

which by an application of Cauchy-Schwartz inequality to the RHS, yields:

$$\inf_{y_2 \in \partial f_2(x)} \langle y_1 - y_2, v \rangle \leq \liminf_{n \to \infty} \|g_1(x + t_n v) - g_2(x + t_n v)\| \leq \sup_{x \in \mathcal{O}} \|g_1(x) - g_2(x)\| .$$

This shows that the desired inequality (3.4) holds for any $x \in \mathcal{O}$ and $y_1 \in \partial f_1(x)$, and any v with $\|v\| \leq 1$. Thus, we successfully establish the assertion of Theorem 1 in the convex settings.

Part II. In the second part, we study the general case where f_1 and f_2 are both *locally weakly convex* on an open set $\mathcal{O}$, extending the established results for convex functions in Part I.

Let $z \in \mathcal{O}$ be arbitrary. Then there is a convex, open neighborhood of $U_z \subseteq \mathcal{O}$ such that f_1 and f_2 are weakly convex on U_z. To utilize the established results for convex functions in Part I, we transform f_1 and f_2 into convex functions on U_z by adding a quadratic regularization. Specifically, we consider:

$$\tilde{f}_1 = f_1 + \frac{L}{2}\|\cdot\|^2 \quad \text{and} \quad \tilde{f}_2 = f_2 + \frac{L}{2}\|\cdot\|^2 .$$

where L is a large constant. This addition renders both $\tilde{f}_1$ and $\tilde{f}_2$ convex on U_z.

As a result of this transformation, we modify the selections from the subdifferentials accordingly:

$$\tilde{g}_1(x) = g_1(x) + Lx \quad \text{and} \quad \tilde{g}_2(x) = g_2(x) + Lx.$$

Importantly, this modification does not affect the Hausdorff distance between the subdifferentials nor the norm of the difference between selections from the subdifferentials, i.e., for every $x \in U_z$:

$$\mathbb{H}(\partial f_1(x), \partial f_2(x)) = \mathbb{H}(\partial \tilde{f}_1(x), \partial \tilde{f}_2(x)) \quad \text{and} \quad \|g_1(x) - g_2(x)\| = \|\tilde{g}_1(x) - \tilde{g}(x)\| .$$

Since $\tilde{f}_1$ and $\tilde{f}_2$ are convex on the open, convex set U_z, and both $\tilde{g}_1$ and $\tilde{g}_2$ are selections from $\partial\tilde{f}_1$ and $\partial\tilde{f}_2$, we can apply the result as proved in Part I to these transformed functions to conclude:

$$\begin{aligned}\sup_{x\in U_z} \mathbb{H}(\partial f_1(x), \partial f_2(x)) &= \sup_{x\in U_z} \mathbb{H}(\partial\tilde{f}_1(x), \partial\tilde{f}_2(x)) \\ &\leq \sup_{x\in U_z} \|\tilde{g}_1(x) - \tilde{g}_2(x)\| = \sup_{x\in U_z} \|g_1(x) - g_2(x)\| \,.\end{aligned}$$

Since $z \in U_z$, and $U_z \subseteq \mathcal{O}$, it follows that for every $z \in \mathcal{O}$:

$$\mathbb{H}(\partial f_1(z), \partial f_2(z)) \leq \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| \,.$$

Taking supremum over $z \in \mathcal{O}$, this observation directly leads to the conclusion of Theorem 1:

$$\sup_{x\in\mathcal{O}} \mathbb{H}(\partial f_1(x), \partial f_2(x)) \leq \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| \,.$$

□

Remark A (Necessity of $\mathcal{O}$ being an open set)**.** Theorem 1 would not hold if $\mathcal{O}$ were not an open set. Consider the convex functions $f_1(x) = |x|$ and $f_2(x) = 2|x|$. Notably, the subdifferential sets are

$$\partial f_1(x) = \begin{cases} 1 & x \geq 0 \\ [-1, 1] & x = 0 \\ -1 & x \leq 0 \end{cases}, \qquad \partial f_2(x) = \begin{cases} 2 & x \geq 0 \\ [-2, 2] & x = 0 \\ -2 & x \leq 0 \end{cases}.$$

Let the subgradient selections g_1 and g_2 of ∂f_1 and ∂f_2 respectively be as follows:

$$g_1(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}, \qquad g_2(x) = \begin{cases} 2 & x > 0 \\ 0 & x = 0 \\ -2 & x < 0 \end{cases}.$$

Consider the case where $\mathcal{O} = \{0\}$, which is a singleton, and thus a closed set. Then

$$\sup_{x\in\mathcal{O}} \mathbb{H}(\partial f_1(x), \partial f_2(x)) = \mathbb{H}(\partial f_1(0), \partial f_2(0)) = 1 > 0 = |g_1(0) - g_2(0)| = \sup_{x\in\mathcal{O}} \|g_1(x) - g_2(x)\| \,.$$

This illustrates the necessity for $\mathcal{O}$ to be an open set for Theorem 1 to hold.

Remark B (Comparison of Metrics on Subdifferential Mappings)**.** Let $\mathcal{O} \subseteq \mathbb{R}^d$ be open. For a pair of locally weakly convex functions $f_i : \mathcal{O} \to \mathbb{R}$ for $i = 1, 2$, we define:

$$\mathrm{d}_1(\partial f_1, \partial f_2) := \sup_{x\in\mathcal{O}} \mathbb{H}(\partial f_1(x), \partial f_2(x)), \quad \mathrm{d}_2(\partial f_1, \partial f_2) := \mathbb{H}(\mathrm{gph}_{\mathcal{O}}\partial f_1, \mathrm{gph}_{\mathcal{O}}\partial f_2). \tag{3.5}$$

In defining d_2, for a locally weakly convex $f : \mathcal{O} \to \mathbb{R}$, we recall the graph of its subdifferential:

$$\mathrm{gph}_{\mathcal{O}}\partial f = \{(x, y) \in \mathcal{O} \times \mathbb{R}^d : y \in \partial f(x)\},$$

where the Cartesian product $\mathcal{O} \times \mathbb{R}^d \subseteq \mathbb{R}^d \times \mathbb{R}^d \cong \mathbb{R}^{2d}$ is equipped with the Euclidean distance. By definition, $\mathrm{d}_1, \mathrm{d}_2$ are metrics on $\mathbb{S}_{\mathcal{O}} = \{g : \mathcal{O} \rightrightarrows \mathbb{R}^d : g = \partial f$ for some locally weakly convex $f : \mathcal{O} \to \mathbb{R}\}$. The metric d_2 is known as the graphical distance, aligning well with applications of Attouch's epigraphical convergence theorem [15], see also [16, Theorem 5.1].

We document relations between d_1 and d_2. For every open set $\mathcal{O}$, and every pair of locally weakly convex $f_i : \mathcal{O} \to \mathbb{R}$ for $i = 1, 2$, the following bound holds from the definition:

$$\mathrm{d}_2(\partial f_1, \partial f_2) \leq \mathrm{d}_1(\partial f_1, \partial f_2). \tag{3.6}$$

In other words, the metric d_1 is topologically stronger than the metric d_2 on the space $\mathbb{S}_{\mathcal{O}}$. On the other hand, when $\mathcal{O} = \mathbb{R}$, we can construct locally weakly convex functions $f_{1,n}$ and $f_{2,n}$ such that

$$\lim_{n\to\infty} \mathrm{d}_2(\partial f_{1,n}, \partial f_{2,n}) = 0 \quad \text{while} \quad \lim_{n\to\infty} \mathrm{d}_1(\partial f_{1,n}, \partial f_{2,n}) = 1, \tag{3.7}$$

meaning the metrics d_1 and d_2 are not equivalent in general. To construct such a sequence, we can take convex functions $f_{1,n}(x) = |x|$, and $f_{2,n}(x) = \frac{1}{2}\left|x - \frac{1}{n}\right| + \frac{1}{2}\left|x + \frac{1}{n}\right|$, which yields:

$$\partial f_{1,n}(x) = \begin{cases} 1 & x \geq 0 \\ [-1,1] & x = 0 \\ -1 & x \leq 0 \end{cases}, \quad \text{and} \quad \partial f_{2,n}(x) = \begin{cases} 1 & x > 1/n \\ [0,1] & x = 1/n \\ 0 & x \in (-1/n, 1/n) \\ [-1,0] & x = -1/n \\ -1 & x < -1/n \end{cases}.$$

A direct verifcation gives $\mathrm{d}_1(\partial f_{1,n}, \partial f_{2,n}) = 1$ and $\mathrm{d}_2(\partial f_{1,n}, \partial f_{2,n}) = 1/n$ for every $n \in \mathbb{N}$.

To conclude, the metric d_1 is topologically stronger than the metric d_2 on $\mathbb{S}_{\mathcal{O}}$ for every open set $\mathcal{O}$. An upper bound on $\mathrm{d}_1(\partial f_1, \partial f_2)$ induces an upper bound on $\mathrm{d}_2(\partial f_1, \partial f_2)$. Nevertheless, the two metrics are not equivalent in general.

3.2 Implications to Stochastic Weakly Convex Minimizations

Theorem 1 provides key insights into the uniform convergence of subdifferentials in stochastic minimizations. Theorem 2, which builds on top of Theorem 1, shows that for stochastic minimizing objectives with certain weak convexity requirements, the uniform convergence of subdifferential sets can be effectively understood by analyzing the uniform convergence of any pair of selections from the subdifferentials of both population and empirical objectives.

To formalize our results, we begin by introducing our setup. Let $f(\cdot,\cdot) : \mathbb{R}^d \times \Xi \to \mathbb{R}$ denote a real-valued random function. We are sampling $\xi \sim \mathbb{P}$ where $\mathbb{P}$ is supported on Ξ, assumed to be a Euclidean space.

Under this setup, the *population risk* is given by:

$$\min_{x\in\mathbb{R}^d} \phi(x) = f(x) + R(x) + \iota_X(x) \quad \text{where} \quad f(x) = \mathbb{E}_{\xi\sim\mathbb{P}}[f(x,\xi)] = \int_{\Xi} f(x,\xi)P(d\xi),$$

whereas its associated *empirical risk* is given by:

$$\min_{x\in\mathbb{R}^d} \phi_S(x) = f_S(x) + R(x) + \iota_X(x) \quad \text{where} \quad f_S(x) = \mathbb{E}_{\xi\sim\mathbb{P}_m}[f(x,\xi)] = \frac{1}{m}\sum_{i=1}^{m} f(x,\xi_i).$$

In the above, $\xi_1, \xi_2, \ldots, \xi_m$ are i.i.d. samples drawn from the distribution $\mathbb{P}$, where $\mathbb{P}_m$ denotes the empirical distribution of $\xi_1, \xi_2, \ldots, \xi_m$. Formally, we start with a single random variable ξ defined on a probability space $(\Xi, \mathcal{G}, \mathbb{P})$. Utilizing this space, we can construct the countable product space $(\Xi, \mathcal{G}, \mathbb{P})^{\mathbb{N}}$ where the sequence of i.i.d. random variables $(\xi_1, \xi_2, \ldots, \xi_m, \ldots)$ can be properly defined on this product probability space [30, Section 36], with each ξ_i acting as a coordinate projection. In this product space, we then extend the sigma-algebra and the probability measure to obtain a complete measure, following standard procedures in measure theory [30], to avoid some cumbersome measurability issues [2, Section 7]. The function $R : \mathbb{R}^d \to \mathbb{R}$ is a real-valued, closed and convex function. The constraint set X is a nonempty, closed, and convex set.

We are interested in establishing an upper bound for the rate of convergence of subdifferentials of the empirical objective ϕ_S to the population objective ϕ over some subset of X, frequently modeled as $X \cap \mathcal{O}$ where $\mathcal{O}$ is an open set in $\mathbb{R}^d$. This openness requirement for $\mathcal{O}$ originates from Theorem 1 and, despite this condition, our result is relevant to a broad range of interesting applications.

To ease our discussions, we assume:

Assumption A. *$f(x) = \int_\Xi f(x,\xi)P(d\xi) < \infty$ for all $x \in \mathcal{O}$ where $\mathcal{O}$ is an open set.*

Assumption A implies, with probability one, f and f_S are real-valued functions on $\mathcal{O}$. In many practical scenarios, the objective f has its domain $\mathbb{R}^d$, for which Assumption A naturally holds.

Our results also require the following locally weak convexity assumption on $f(\cdot, \xi)$.

Assumption B. *For all x in $\mathcal{O}$, there exists $\epsilon(x) > 0$ and $\lambda(x,\xi) \geq 0$ such that*

$$y \mapsto f(y,\xi) + \frac{\lambda(x,\xi)}{2}\|y\|^2$$

is convex on the open ball $\mathbb{B}(x;\epsilon(x))$ and $\mathbb{E}_{\xi\sim\mathbb{P}}[\lambda(x,\xi)] < \infty$.

Lemma 1. *Let Assumptions A and B hold. Then with probability one, the function $f_S(x)$ is locally weakly convex on the open set $\mathcal{O}$. Also, the function $f(x) = \mathbb{E}[f_S(x)]$ is locally weakly convex on $\mathcal{O}$.*

Proof. By Assumption B, for each $x \in \mathcal{O}$, there exists a radius $\epsilon(x) > 0$ and $\lambda(x,\xi) \geq 0$ such that

$$y \mapsto f(y,\xi) + \frac{\lambda(x,\xi)}{2}\|y\|^2$$

is convex on the open ball $\mathbb{B}(x;\epsilon(x))$. The collection of these balls, $\{\mathbb{B}(x;\epsilon(x))\}_{x\in\mathcal{O}}$ forms an open cover of $\mathcal{O}$. By Lindelöf's covering Theorem, see, e.g., [35, Page 50], and noting that $\mathbb{R}^d$ with its norm topology is second countable, there is a countable subcover $\{\mathbb{B}(x;\epsilon(x))\}_{x\in\mathcal{O}^o}$ of $\mathcal{O}$, where $\mathcal{O}^o$ is a countable subset of $\mathcal{O}$. In particular, $\mathcal{O}$ obeys the inclusion:

$$\mathcal{O} \subseteq \bigcup_{x\in\mathcal{O}^o} \mathbb{B}(x;\epsilon(x)).$$

For each $x \in \mathcal{O}^o$, the function

$$y \mapsto f_S(y) + \frac{\mathbb{E}_{\xi\sim\mathbb{P}_m}[\lambda(x,\xi)]}{2}\|y\|^2 = \frac{1}{m}\sum_{i=1}^m \left(f(y,\xi_i) + \frac{\lambda(x,\xi_i)}{2}\|y\|^2\right)$$

is convex on the open neighborhood $\mathbb{B}(x;\epsilon(x))$. Given that $\mathbb{E}_{\xi\sim\mathbb{P}}[\lambda(x,\xi)] < \infty$ holds, for each $x \in \mathcal{O}^o$ it follows that $\mathbb{E}_{\xi\sim\mathbb{P}_m}[\lambda(x,\xi)] < \infty$ with probability one. Using a union bound over the countable set $\mathcal{O}^o$ this implies that with probability one, $\mathbb{E}_{\xi\sim\mathbb{P}_m}[\lambda(x,\xi)] < \infty$ simultaneously holds for every $x \in \mathcal{O}^o$. Thus, with probability one $f_S(x)$ is locally weakly convex on the set $\mathcal{O}$, as $\mathcal{O}$ is included in the countable union of these open balls: $\mathcal{O} \subseteq \bigcup_{x\in\mathcal{O}^o}\mathbb{B}(x;\epsilon(x))$. A similar reasoning yields that the function $f(x) = \mathbb{E}[f_S(x)]$ is locally weakly convex within the open set $\mathcal{O}$. □

Building on Theorem 1 and Lemma 1 then Theorem 2 naturally follows.

Theorem 2. *Let Assumptions A and B hold. Let G and G_S be selections of the subdifferentials ∂f and ∂f_S over $x \in \mathcal{O}$ respectively, i.e., obeying the inclusions $G(x) \in \partial f(x)$ and $G_S(x) \in \partial f_S(x)$ for every $x \in \mathcal{O}$. Then, the following inequality holds with probability one:*

$$\sup_{x \in \mathcal{O} \cap X} \mathbb{H}(\partial \phi(x), \partial \phi_S(x)) \leq \sup_{x \in \mathcal{O}} \|G(x) - G_S(x)\| .$$

Proof. Given that X is closed and convex, ι_X is subdifferentially regular at any $x \in X$ [17, Exercise 8.14]. The functions $f(x)$ and $f_S(x)$ are real-valued, closed and weakly convex on $\mathcal{O}$ by Assumptions A and B. Thus they are subdifferentially regular on $\mathcal{O}$ [17, Corollary 8.11], with $\partial^\infty f(x) = \partial^\infty f_S(x) = \{0\}$ for all $x \in \mathcal{O}$. Similarly, the real-valued, closed, and convex function r, is subdifferentially regular at $x \in \mathcal{O}$ with $\partial^\infty R(x) = \{0\}$.

These regularity properties of the functions enable the application of basic subdifferential calculus [17, Corollary 10.9]. Accordingly, the subdifferentials at $x \in X$ are given by:

$$\partial \phi(x) = \partial f(x) + \partial R(x) + \mathcal{N}_X(x), \qquad \partial \phi_S(x) = \partial f_S(x) + \partial R(x) + \mathcal{N}_X(x).$$

Here, the addition is the Minkowski sum. A fundamental property of Hausdorff distance is that

$$\mathbb{H}(A_1 + A_3, A_2 + A_3) \leq \mathbb{H}(A_1, A_2)$$

holds for any sets A_1, A_2, A_3. Consequently, this implies for every $x \in X$:

$$\mathbb{H}(\partial \phi(x), \partial \phi_S(x)) \leq \mathbb{H}(\partial f(x), \partial f_S(x)).$$

Since both functions $f(x)$ and $f_S(x)$ are real-valued and locally weakly convex functions on $\mathcal{O}$ due to Assumptions A and B, the following chain of inequalities then follows:

$$\begin{aligned}\sup_{x \in \mathcal{O} \cap X} \mathbb{H}(\partial \phi(x), \partial \phi_S(x)) &\leq \sup_{x \in \mathcal{O} \cap X} \mathbb{H}(\partial f(x), \partial f_S(x)) \\ &\leq \sup_{x \in \mathcal{O}} \mathbb{H}(\partial f(x), \partial f_S(x)) \leq \sup_{x \in \mathcal{O}} \|G(x) - G_S(x)\| .\end{aligned}$$

The last inequality is due to Theorem 1, acknowledging that, with probability one, f and f_S are real-valued and weakly-convex on the open set $\mathcal{O}$ by Lemma 1. This inequality, which involves the supremum over uncountable random variables, represents a measurable event with probability one due to the completeness of the probability measure on the product space (see the setup on the probability space at the beginning of this subsection). □

3.3 A Class of Stochastic Weakly Convex Minimizations

One important example of stochastic weakly convex minimizations corresponds to the setting where

$$f(x, \xi) = h(c(x; \xi)). \tag{3.8}$$

Here, $h : \mathbb{R}^k \to \mathbb{R}$ is a closed, convex and real-valued function, and $c(\cdot, \xi) : \mathbb{R}^d \to \mathbb{R}^k$ is $\mathcal{C}^1$ smooth on $\mathbb{R}^d$ for every ξ. Under this scenario, the population and empirical objectives become

$$\begin{aligned}\min_{x \in \mathbb{R}^d} \phi(x) = f(x) + R(x) + \iota_X(x) \quad &\text{where} \quad f(x) = \mathbb{E}_{\xi \sim \mathbb{P}}[h(c(x; \xi))] = \int_\Xi h(c(x; \xi)) P(d\xi) \\ \min_{x \in \mathbb{R}^d} \phi_S(x) = f_S(x) + R(x) + \iota_X(x) \quad &\text{where} \quad f_S(x) = \mathbb{E}_{\xi \sim \mathbb{P}_m}[h(c(x; \xi))] = \frac{1}{m}\sum_{i=1}^m h(c(x; \xi_i))\end{aligned} . \tag{3.9}$$

In the literature, it is known that under mild conditions, the convex composition $f(\cdot, \xi) = h(c(\cdot;\xi))$ satisfies weak-convexity requirements in Assumption B, resulting in locally weakly convex objectives $f(x)$ and $f_S(x)$ that fit into our diagrams in Section 3.2, see, e.g., [7, Claim 1] and [36]. We will revisit these conditions subsequently before we present our main results in Theorem 3.

We give examples of stochastic convex-composite objectives that appear in statistics, machine learning, imaging, and risk management. More examples can be found in [36, Section 2.1].

Example 1 (Robust Phase Retrieval)**.** Phase retrieval is a computational problem with applications across various fields, including imaging, X-ray crystallography, and speech processing. The (real-valued) phase retrieval seeks to detect a point x satisfying $|\langle a_i, x\rangle|^2 = b_i$ where $a_i \in \mathbb{R}^d$, and $b_i \in \mathbb{R}$ for $i = 1, 2, \ldots, m$. We can choose $\xi = (a, b) \in \mathbb{R}^d \times \mathbb{R}$, $h(z) = |z|$ and $c(x;\xi) = (a^T x) - b$, and $X = \mathbb{R}^d$, in which case the form (3.9) gives an exact penalty formulation for solving the collection of quadratic equations, which yields strong statistical recovery and robustness guarantees [8, 14, 37], among other nonconvex formulations [38–40].

Example 2 (Robust Matrix Sensing)**.** This problem can be viewed as a variant of phase retrieval. Let $A_1, A_2, \ldots, A_m \in \mathbb{R}^{D\times D}$ be measurement matrices. Given the measurement $b_i = \langle A_i, M_\sharp\rangle + \eta_i \in \mathbb{R}$, where $M_\sharp \in \mathbb{R}^{D\times D}$ is the true matrix, and $\eta_i \in \mathbb{R}$ is the noise corruption, the goal is to recover a low-rank approximation of $M_\sharp$, which is modeled through XX^T where $X \in \mathbb{R}^{D\times r_0}$ where $1 \le r_0 \le D$. Recently, there has been a series of efforts showing that the following potential function has strong stability guarantees under certain assumptions on A and η [10, 11, 41]:

$$\min_{X\in\mathbb{R}^{d\times r}} \frac{1}{m}\sum_{i=1}^{m} |\langle A_i, XX^T\rangle - b_i|.$$

This falls into the form (3.9) by simply setting $x = X \in \mathbb{R}^{D\times r_0}$, $\xi = (A, b)$ where $A \in \mathbb{R}^{D\times D}, b \in \mathbb{R}$, $h(z) = |z|$ and $c(x;\xi) = c(x;(A, b)) = \langle A, XX^T\rangle - b$.

Example 3 (Conditional Value-at-Risk)**.** Let $\ell(w, \xi)$ represent a decision rule parameterized by $w \in \mathbb{R}^d$ for a data point ξ, where $\xi \sim \mathbb{P}$. Instead of minimizing the expected value $\mathbb{E}_{\xi\sim\mathbb{P}}[\ell(w, \xi)]$, it is often preferable to minimize the conditional expectation of the random variable $\ell(w, \cdot)$ over its α-tail, where $\alpha \in (0, 1)$ is given. This quantity is termed the Conditional Value-at-Risk (cVaR) [2, Section 6]. Remarkably, a seminal work shows that minimizing cVaR can be formulated as [12]:

$$\min_{\gamma\in\mathbb{R}, w\in\mathbb{R}^d} \mathbb{E}_{\xi\sim P}[(1-\alpha)\gamma + (\ell(w, \xi) - \gamma)_+].$$

Now we suppose $\ell(\cdot, \xi)$ is $\mathcal{C}^1$ smooth. This falls into the form (3.9) by setting $x = (\gamma, w) \in \mathbb{R} \times \mathbb{R}^d$, $h(z) = (z)_+$, $c(x;\xi) = c((\gamma, w);\xi) = \ell(w, \xi) - \gamma$, and $R(x) = R((\gamma, w)) = (1-\alpha)\gamma$.

We now revisit the conditions on the convex function h and smooth function $c(x, \xi)$ that guarantees the weak-convexity assumptions for the composition $f(x, \xi) = h(c(x, \xi))$, ensuring the regularity required for subdifferential calculus rules. We take the following conditions—local Lipschitzian and integrability conditions on h, c—from [7, Section 2.2].

Assumption A'. *$f(x) = \int_\Xi h(c(x;\xi))P(d\xi) < \infty$ for every $x \in \mathcal{O}$ where $\mathcal{O}$ is an open set.*

Assumption B'. *For every $x \in \mathcal{O}$, there is $\epsilon(x) > 0$ such that*

- $\sup_{y\in\mathbb{B}(x;\epsilon(x))} \|\nabla c(y;\xi)\partial h(c(y;\xi))\|$ *is integrable with respect to $\xi \sim \mathbb{P}$.*
- $\|\nabla c(y;\xi) - \nabla c(y';\xi)\| \le \beta_\epsilon(x, \xi)\|y - y'\|$ *for $y, y' \in \mathbb{B}(x;\epsilon(x))$, where $\mathbb{E}[\beta_\epsilon(x, \xi)] < \infty$.*

Given Assumptions A' and B', the following basic property on the stochastic convex-composite objectives follows from [7, Claim 1+Lemma 3.6]. We use $\nabla c(x;\xi)$ to denote the gradient of a smooth function c with respect to x for a given ξ.

Lemma 2. *Let Assumptions A' and B' hold. Then $f(x;\xi) = h(c(x;\xi))$ satisfies Assumption B. Moreover,*

$$\partial f(x;\xi) = \partial h(c(x;\xi))\nabla c(x;\xi),$$

with

$$\partial f(x) = \mathbb{E}_{\xi\sim\mathbb{P}}[\partial f(x;\xi)], \quad \text{and} \quad \partial f_S(x) = \mathbb{E}_{\xi\sim\mathbb{P}_m}[\partial f(x;\xi)].$$

For ease of reference in future discussions, we establish a result that stems directly from Theorem 2 and Lemma 2 for stochastic convex-composite minimizations.

Theorem 3. *Let Assumptions A' and B' hold for $f(x,\xi) = h(c(x;\xi))$. Suppose at every $x \in \mathcal{O}$, the functions $G(x)$ and $G_S(x)$ obey:*

$$G(x) \in \mathbb{E}_{\xi\sim\mathbb{P}}[\partial f(x;\xi)] \quad \text{and} \quad G_S(x) \in \mathbb{E}_{\xi\sim\mathbb{P}_m}[\partial f(x;\xi)].$$

Then, the following inequality holds with probability one:

$$\sup_{x\in X\cap\mathcal{O}} \mathbb{H}(\partial\phi(x), \partial\phi_S(x)) \leq \sup_{x\in\mathcal{O}} \|G(x) - G_S(x)\|.$$

In the above, $\phi(x)$ and $\phi_S(x)$ refer to the convex-composite objectives in equation (3.9).

4 Explicit Uniform Convergence Rates

In this section, we will delve deep into the convergence rate of the subdifferentials

$$\sup_{x\in X\cap\mathcal{O}} \mathbb{H}(\partial\phi(x), \partial\phi_S(x)) \tag{4.1}$$

for a class of stochastic convex-composite objectives where

$$\phi(x) = \mathbb{E}_{\xi\sim\mathbb{P}}[h(c(x;\xi))] + R(x) + \iota_X(x), \qquad \phi_S(x) = \mathbb{E}_{\xi\sim\mathbb{P}_m}[h(c(x;\xi))] + R(x) + \iota_X(x). \tag{4.2}$$

Here, the crucial assumption is that $h : \mathbb{R} \to \mathbb{R}$ is a one-dimensional convex function. The function $c(\cdot;\xi) : \mathbb{R}^d \to \mathbb{R}$ is $\mathcal{C}^1$ smooth on $\mathbb{R}^d$ for every ξ, aligning with the framework in Section 3.3. As before, $R : \mathbb{R}^d \to \mathbb{R}$ is closed and convex, and X is nonempty, closed and convex. This setting is particularly relevant as it covers a broad spectrum of practical objectives, including all examples previously discussed in Section 3.3. In our considerations, the set $\mathcal{O}$ is usually an open ball $\mathbb{B}(x_0;r)$ in the Euclidean space where $x_0 \in \mathbb{R}^d$ is the center and $r > 0$ is the radius.

Despite their wide applications, tight bounds on the subdifferential difference of these objectives remain largely open in the literature [16, Section 5], which motivates our investigations. We will show in Section 5 how our main result in this section, Theorem 5, yields tight bounds on the subdifferential differences, and leads to sharp quantitative characterizations on the nonsmooth landscape of stochastic convex-composite formulations in finite samples.

4.1 Subgradient Selections

Theorem 3 shows that, under regularity Assumptions A' and B', there is the bound:

$$\sup_{x\in X\cap\mathcal{O}} \mathbb{H}(\partial\phi(x), \partial\phi_S(x)) \leq \sup_{x\in\mathcal{O}} \|G(x) - G_S(x)\| \tag{4.3}$$

where $\mathcal{O}$ is an open set. Remarkably, G and G_S can be any pair of selections from the subdifferentials (cf. Lemma 2 and Theorem 3):

$$G(x) \in \mathbb{E}_{\xi \sim P}[\partial h(c(x;\xi))\nabla c(x;\xi)], \quad G_S(x) \in \mathbb{E}_{\xi \sim P_m}[\partial h(c(x;\xi))\nabla c(x;\xi)]. \tag{4.4}$$

The main goal of this subsection is to select a pair of subgradients $G(x)$ and $G_S(x)$ that are simple to analyze—from a probabilistic sense—regarding their uniform differences over an open set $\mathcal{O}$.

This selection process begins with identifying a subgradient for the nonsmooth convex function h, which we decompose into a sum $h = h^{sm} + h^{ns}$ of a smooth component h^{sm} and a nonsmooth component h^{ns}. To do so, we use h'_+ and h'_- to denote the right-hand and left-hand derivative of h, respectively. We enumerate the set of nondifferentiable points $\{t \in \mathbb{R} : h'_+(t) \neq h'_-(t)\}$ as $\{t_j\}_{j=1}^{\infty}$, given the fact that these nondifferentiable points are always countable [42, Theorem 2.1.2].

For the majority of Section 4, we will work under the assumption that $\sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) < \infty$ so that h^{ns} and h^{sm} are well-defined, although this assumption will be unnecessary in our final bound in Theorem 5.

Under this assumption, we can decompose this nonsmooth h into two distinct parts:

$$h(x) = h^{sm}(z) + h^{ns}(z). \tag{4.5}$$

Here, $h^{ns} : \mathbb{R} \mapsto \mathbb{R}$ (where "ns" stands for nonsmooth) denotes the nonsmooth component of h, which is defined by:

$$h^{ns}(x) = \sum_{j=1}^{\infty} a_j \cdot (z - t_j)_+, \tag{4.6}$$

where the coefficients are defined by $a_j = h'_+(t_j) - h'_-(t_j) > 0$. Notably, h^{ns} is well-defined because $\sum_{j=1}^{\infty} a_j = \sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) < \infty$. The smooth component $h^{sm} : \mathbb{R} \mapsto \mathbb{R}$ (where "sm" stands for smooth) is defined as $h^{sm} = h - h^{ns}$. By construction, h^{sm} is a $\mathcal{C}^1$ smooth and convex function, as one can easily verify its derivative $(h^{sm})'$ is continuous and monotonically increasing on $\mathbb{R}$.

To select a subgradient of the convex function h, we first note that the smooth component h^{sm} possesses a unique subgradient, which corresponds to its derivative:

$$g^{sm}(z) = (h^{sm})'(z). \tag{4.7}$$

For the nonsmooth component h^{ns}, a subgradient at every $z \in \mathbb{R}$ is chosen as follows:

$$g^{ns}(z) = \sum_{j=1}^{\infty} a_j \mathbf{1}(z \geq t_j) \in \partial h^{ns}(z), \tag{4.8}$$

where $z \mapsto \mathbf{1}(z \geq t_j)$ is an indicator function that is a subgradient for the mapping $z \mapsto (z - t_j)_+$, taking the value 1 when $z \geq t_j$ and 0 otherwise. Combining these, we define a subgradient g for h at every z:

$$g(z) := (g^{sm} + g^{ns})(z) \in \partial h(z).$$

With $g(z)$ defined, we can then pick the subgradients G and G_S that satisfy equation (4.4):

$$\begin{aligned} G(x) &:= \mathbb{E}_{\xi \sim P}\left[g(c(x,\xi)) \cdot \nabla c(x,\xi)\right] \\ G_S(x) &:= \mathbb{E}_{\xi \sim P_m}\left[g(c(x,\xi)) \cdot \nabla c(x,\xi)\right] \end{aligned}. \tag{4.9}$$

To summarize, G and G_S are well-defined whenever $\sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) < \infty$.

In the next subsection, we will provide a uniform convergence result that establishes a high probability upper bound on $\|G_S(x) - G(x)\|$ over x in a Euclidean ball in $\mathbb{R}^d$. We will address the following question.

Question: Given a center $x_0 \in \mathbb{R}^d$, a radius r, a threshold $\delta \in (0, 1)$, what values of u can we choose such that

$$\mathbb{P}^* \left(\sup_{x: x \in \mathbb{B}(x_0; r)} \|G_S(x) - G(x)\| \geq u \right) \leq \delta. \tag{4.10}$$

Answers to this question will automatically yield a high probability upper bound on the subdifferential difference $\sup_{x \in X \cap \mathbb{B}(x_0; r)} \mathbb{H}(\phi(x), \phi_S(x))$ by applying equation (4.3) to the open set $\mathcal{O} = \mathbb{B}(x_0; r)$. Recall $\mathbb{P}^*$ denotes the outer probability (Section 2.4), which is required because the supremum in equation (4.10) may not be measurable under the probability space.

4.2 A Uniform Bound on Subgradient Selections

This section examines the uniform rate of convergence of G_S to G over any Euclidean ball of radius r, a question that is arised in equation (4.10). Recall our definitions:

$$\begin{aligned} G_S(x) &= \mathbb{E}_{\xi \sim P_m} \left[\sum_{j=1}^{\infty} a_j \mathbf{1}\{c(x, \xi) \geq t_j\} \cdot \nabla c(x, \xi) + (h^{sm})'(c(x, \xi)) \cdot \nabla c(x, \xi) \right] \\ G(x) &= \mathbb{E}_{\xi \sim P} \left[\sum_{j=1}^{\infty} a_j \mathbf{1}\{c(x, \xi) \geq t_j\} \cdot \nabla c(x, \xi) + (h^{sm})'(c(x, \xi)) \cdot \nabla c(x, \xi) \right] \end{aligned}. \tag{4.11}$$

A key challenge in examining the rate of uniform convergence from G_S to G is due to the non-smooth nature of G_S and G, whose definition involves indicator functions, as outlined in equation (4.11). To resolve this challenge, our approach leverages the notion of Vapnik–Chervonenkis dimension (VC dimension) [43], a concept familiar to experts in statistical learning theory. In short words, VC dimension measures the complexity of sets used in defining these indicator functions (termed as classifiers in statistical learning theory), and this complexity measure governs the uniform generalizability from empirical averages to expectations for these indicator functions.

We state the definition of VC dimension for completeness [22, Section 3.6].

Definition 1 (VC Dimension)**.** *Let $\mathcal{H}$ denote a set family, and E a set. Let $\mathcal{H} \cap E = \{H \cap E \mid H \in \mathcal{H}\}$. We say a set E is* shattered *if $\mathcal{H} \cap E$ contains all the subsets of E, i.e., $|\mathcal{H} \cap E| = 2^{|E|}$. The VC dimension $\mathrm{vc}(\mathcal{H})$ is the cardinality of the largest set that is* shattered *by $\mathcal{H}$. If arbitrarily large sets can be shattered, the VC dimension of $\mathcal{H}$ is ∞.*

The collection of sets that appear in the indicator functions will be denoted as:

$$\mathcal{F} = \{\{\xi \in \Xi : c(x, \xi) \geq t\} \mid x \in \mathbb{R}^d, t \in \mathbb{R}\}. \tag{4.12}$$

More transparently, the VC dimension of $\mathcal{F}$, denoted by $\mathrm{vc}(\mathcal{F})$, is the largest integer N such that there exist $\xi_1, \xi_2, \ldots, \xi_N \in \Xi$ so that for any binary labeling sequence $b_1, b_2, \ldots, b_N \in \{0, 1\}$, it is possible to find parameters $x \in \mathbb{R}^d, t \in \mathbb{R}$ that satisfy:

$$b_i = \mathbf{1}\{c(x; \xi_i) \geq t\} \quad \text{for every } 1 \leq i \leq N.$$

The complexity measure $\mathrm{vc}(\mathcal{F})$ characterizes generalization capabilities of $\mathcal{F}$ and bounds the uniform rate of convergence from G_S to G. For common techniques of upper bounding VC dimensions of set families, see [31, Chapter 2]. When the mapping $x \mapsto c(x; \xi)$ are polynomials in $x \in \mathbb{R}^d$ for any given ξ—a condition met in our applications—we establish a result giving a tight upper bound of the VC dimension, based on results in real algebraic geometry. See Section 6 for details.

Our main results rely on Lipschitz continuity and regularity assumptions concerning the tails of the random vectors $\nabla c(x, \xi)$ and $(h^{sm})'(c(x, \xi)) \cdot \nabla c(x, \xi)$. These random vectors appear in our subgradient selection (4.11), and our assumptions about them align with standard treatments in the literature. To state the assumptions, we first recall the concept of a subexponential tail random vector. We define this in terms of Orlicz norms, following [44, Section 3.4.4].

Definition 2. *A random vector Z is said to be σ-subexponential if for every v with $\|v\| = 1$:*

$$\mathbb{E}\left[\exp\left(\frac{|\langle Z, v\rangle|}{\sigma}\right)\right] \le 2.$$

Examples of σ-sub-exponential random vectors include a normal distribution $Z \sim \mathsf{N}(0, \Sigma)$ where the largest eigenvalue of Σ is bounded above by $c\sigma$, or distributions with bounded support, such as $\|Z\| \le c\sigma$ almost surely, with the numerical constant $c = 1/2$ [44, Section 2].

By Markov's inequality, $\mathbb{P}(|\langle v, Z\rangle| \ge \sigma \cdot u) \le 2e^{-u}$ for any unit vector v and $u \ge 0$. Furthermore, by Bernstein's inequality [44, Theorem 2.8.1] and the centering property of sub-exponential random variables [44, Exercise 2.7.10], for i.i.d. σ-subexponential random variables $Z_1, Z_2, \dots, Z_m$, there is the *exponential* tail bound for the mean $\frac{1}{m}\sum_{i=1}^m Z_i$ that holds for every v with $\|v\| = 1$, and $u \ge 0$:

$$\mathbb{P}\left(\frac{1}{m}\left|\sum_{i=1}^m \langle Z_i - \mathbb{E}[Z], v\rangle\right| \ge \sigma \cdot u\right) \le 2\exp\left(-cm\min\{u^2, u\}\right). \tag{4.13}$$

In the above, $c > 0$ is an absolute constant.

We first assume the random vectors $\nabla c(x_0, \xi)$ and $(h^{sm})'(c(x_0, \xi)) \cdot \nabla c(x_0, \xi)$ are subexponential, where we recall that x_0 is the center of the ball of interest (see equation (4.10)). This facilitates an initial *exponential* tail probability bound on the difference between $G_S(x_0)$ and $G(x_0)$.

Assumption C.1. *Each of the two random vectors*

$$\nabla c(x_0, \xi) \quad \textit{and} \quad (h^{sm})'(c(x_0, \xi)) \cdot \nabla c(x_0, \xi)$$

is σ-subexponential.

To further ensure a uniform *exponential* tail probability bound on the difference between $G_S(x)$ and $G(x)$ over all possible $x \in \mathbb{B}(x_0; r)$, we will assume that the increments of the random processes $\{\nabla c(x, \xi)\}_{x \in \mathbb{B}(x_0;r)}$ and $\{(h^{sm})'(c(x, \xi))\nabla c(x, \xi)\}_{x \in \mathbb{B}(x_0;r)}$ are subexponential in the following sense.

Assumption C.2. *For every $x_1, x_2 \in \mathbb{B}(x_0; r)$, each of the two random vectors*

$$\nabla c(x_1, \xi) - \nabla c(x_2, \xi) \quad \textit{and} \quad e(x_1, \xi) - e(x_2, \xi)$$

is $\sigma\|x_1 - x_2\|$ subexponential. In the above, $e(x, \xi) = (h^{sm})'(c(x, \xi))\nabla c(x, \xi)$ for every $x \in \mathbb{B}(x_0; r)$.

This assumption is commonly used in empirical process theory to extend high probability bounds from individual points to the uniform control of the supremum of stochastic processes, utilizing the chaining technique. Notably, this assumption naturally holds in a variety of statistical applications. For a reference, see [31, Chapter 2].

We are now ready to present our main result. The proof of Theorem 4 employs empirical process theory tools, namely, the calculus rules for outer probability measure [31, Chapter 1.2-5], the chaining argument [45], and the Sauer-Shelah Lemma—a fundamental combinatorial principle using the VC dimension to control function class complexities [46, 47]. The proof is relatively standard for experts in statistical learning theory and is detailed in Section A.

Theorem 4. *Assume Assumptions C.1 and C.2. Assume* $\sum_{j=1}^{\infty}(h'_{+}(t_j) - h'_{-}(t_j)) < \infty$. *There exists a universal constant* $C > 0$ *such that for every* $\delta \in (0, 1)$*:*

$$\mathbb{P}^* \left(\sup_{x: x \in \mathbb{B}(x_0; r)} \|G(x) - G_S(x)\| \geq Cr\sigma\zeta \cdot \max\{\Delta, \Delta^2\} \right) \leq \delta. \tag{4.14}$$

In the above,

$$\zeta = 1 + \sum_{j=1}^{\infty}(h'_{+}(t_j) - h'_{-}(t_j)) \tag{4.15}$$

where t_j *enumerate points where* h *is non-differentiable, and*

$$\Delta = \sqrt{\frac{1}{m} \cdot \Big(d + \mathrm{vc}(\mathcal{F}) \log m + \log(\frac{1}{\delta})\Big)}, \tag{4.16}$$

where $\mathrm{vc}(\mathcal{F})$ *is the VC dimension of the set family* $\mathcal{F}$ *defined in equation (4.12).*

Recall $\mathbb{P}^*$ denotes the outer probability (Section 2.4), also used to address measurability issues with supremums in empirical process theory [31]. This is technically required because the supremum in equation (4.14) may not be measurable under the probability space.

Theorem 4 establishes a uniform closeness between the selected subgradient G_S and G over the Euclidean ball $\mathbb{B}(x_0; r)$. In the final bound (4.14), ζ captures how the nondifferentiability of h could potentially deteriorate the convergence rate. Crucially, the error term Δ in equation (4.16) demonstrates how the dimensionality d and the complexity measure $\mathrm{vc}(\mathcal{F})$ of the set family $\mathcal{F}$ influence the uniform convergence rate.

The presence of $\max\{\Delta, \Delta^2\}$ in the final bound instead of just Δ stems from the sub-exponential assumption (cf. $\min\{u, u^2\}$ in the Bernstein inequality (4.13)). When $\Delta \leq 1$, which is typical for large sample size m, this term simplifies to Δ, which is $\sqrt{(d + \mathrm{vc}(\mathcal{F}))/m}$ up to logarithmic factors. Our convergence rates align well with the standard uniform convergence rate $\sqrt{d/m}$ for averages of smooth functions and $\sqrt{\mathrm{vc}(\mathcal{F})/m}$ for indicator functions in the statistics literature [48]. This suggests our bounds are in general tight up to logarithmic factors in m.

Finally, the $\log m$ term arises from the subexponential tail assumption and can be removed if a norm boundedness assumption replaces it in Assumption C.2, or by using more advanced arguments involving additional probabilistic control of the envelope function of $\nabla c(x, \xi)$ and $(h^{sm})'(c(x, \xi)) \cdot \nabla c(x, \xi)$ [31, Chapter 2]. For conciseness and clarity, this paper does not pursue such removal.

4.3 Implication: A Uniform Bound on Subdifferentials

We integrate Theorem 3 and Theorem 4 to document our ultimate result, Theorem 5, on uniform convergence of subdifferentials. At this point, the only work is to collect and streamline all the assumptions in Theorem 3 and 4. To fulfill Assumption B' required in Theorem 3, we identify an additional condition beyond Assumptions C.1—C.2. This Lipschitz-type condition is very mild, and holds in many applications in stochastic programming [2, Chapter 5] and statistics [49, Chapter 5].

Assumption C.3. *There is a measurable function* $L : \Xi \to \mathbb{R}$ *with* $\mathbb{E}_{\xi \sim \mathbb{P}}[L(\xi)] < \infty$ *such that*

$$\|\nabla c(x_1, \xi) - \nabla c(x_2, \xi)\| \leq L(\xi) \|x_1 - x_2\|$$
$$\|e(x_1, \xi) - e(x_2, \xi)\| \leq L(\xi) \|x_1 - x_2\|$$

hold for every $x_1, x_2 \in \mathbb{B}(x_0; r)$*, where* $e(x, \xi) = (\psi^{sm})'(c(x, \xi))\nabla c(x, \xi)$*.*

Theorem 5 documents the final result on uniform convergence of subdifferentials, recognizing that Assumptions C.1—C.3 imply that Assumption B' holds for $\mathcal{O} = \mathbb{B}(x_0; r)$.

Theorem 5. *Assume Assumptions C.1–C.3, and $f(x) = \mathbb{E}_{\xi \sim P}[h(c(x;\xi))] < \infty$ for $x \in \mathbb{B}(x_0; r)$. Then there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$:*

$$\mathbb{P}^* \left(\sup_{x: x \in X \cap \mathbb{B}(x_0;r)} \mathbb{H}(\partial\phi(x), \partial\phi_S(x)) \geq Cr\sigma\zeta \cdot \max\{\Delta, \Delta^2\} \right) \leq \delta. \tag{4.17}$$

In the above, ζ, Δ follows the same definition in the statement of Theorem 4.

Proof. We can assume without loss of generality that $\zeta = 1 + \sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) < \infty$. If this sum were infinite, then equation (4.17) would be trivially satisfied.

Given Theorem 3 and Theorem 4, and with Assumption A' assumed for $\mathcal{O} = \mathbb{B}(x_0; r)$ in the statement of Theorem 5, the remaining task is to show that Assumption B' is met for $\mathcal{O} = \mathbb{B}(x_0; r)$. This can be achieved by first recognizing that for $y \in \mathbb{B}(x_0; r)$:

$$\begin{aligned} \|\nabla c(y;\xi)\partial h(c(y;\xi))\| &\leq \|\nabla c(y;\xi) \cdot \partial h^{ns}(c(y;\xi))\| + \|\nabla c(y;\xi) \cdot (h^{sm})'(c(y;\xi))\| \\ &\leq \|\nabla c(y;\xi)\| \cdot \sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) + \|\nabla c(y;\xi) \cdot (h^{sm})'(c(y;\xi))\|. \end{aligned}$$

Assumptions C.1–C.2 imply that $\sup_{y \in \mathbb{B}(x_0;r)} \|\nabla c(y;\xi)\|$ and $\sup_{y \in \mathbb{B}(x_0;r)} \|(h^{sm})'(c(y;\xi))\nabla c(y;\xi)\|$ are integrable over $\xi \sim \mathbb{P}$ due to the standard chaining argument [44, Section 8]. This implies that $\sup_{y \in \mathbb{B}(x_0;r)} \|\nabla c(y;\xi)\partial h(c(y;\xi))\|$ is also integrable over ξ since $\sum_{j=1}^{\infty}(h'_+(t_j) - h'_-(t_j)) < \infty$, thus confirming the first condition of Assumption B' holds for $\mathcal{O} = \mathbb{B}(x_0; r)$. The second condition of Assumption B' follows trivially from Assumption C.3. □

Remark C. Theorem 5 establishes a rate for uniform convergence of subdifferential mappings for stochastic convex-composite objectives at $\sqrt{\max\{d, \mathrm{vc}(\mathcal{F})\}/m}$, modulo logarithmic factors. Our results are different from approaches that depend on the *continuity* of both empirical and population subdifferentials under the Hausdorff metric, or require a nonatomic probability distribution $\mathbb{P}$ to achieve a $1/\sqrt{m}$ rate [1, 21] (see Section 1.1 for a more detailed literature review). Indeed, Theorem 5 achieves this tight rate without requiring continuity in the subdifferential mappings of the objective or specific conditions on the data distribution $\mathbb{P}$.

In the literature on uniform convergence rates of subdifferential mappings for stochastic weakly-convex objectives, methods using Moreau envelope smoothing [23] and Attouch's epigraphical convergence theorem [15] achieve a convergence rate of $\sqrt[4]{1/m}$ suboptimal in terms of the sample size m [14, 16]. This $\sqrt[4]{1/m}$ convergence is measured using graphical distance, a notion which is topologically stronger than the uniform convergence criterion we consider in Theorem 5 (see Remark of Theorem 1). Our work complements these studies by showing a convergence rate's dependence on sample size m to be $\sqrt{1/m}$ (modulo logarithmic factors) under a topologically weaker notion of convergence.

In summary, for the class of nonsmooth, nonconvex, stochastic convex-composite objectives we study, our results often provide more precise convergence rates or require fewer assumptions about the distribution $\mathbb{P}$. Examples in Section 5 make this concrete.

5 Applications

In this section, we return to some of the examples previously discussed in Section 3.3 to demonstrate the application of our main Theorem 5. We will concretely evaluate the bounds in Theorem 5 for some of these objectives, compare these to existing results in the literature, and use them to derive new findings.

5.1 Illustration I: Phase Retrieval

We start with the (real-valued) robust phase retrieval objective (Example 1):

$$\begin{aligned} &\min_x \Phi(x) \quad \text{where} \quad \Phi(x) &&= \mathbb{E}_{(a,b)\sim\mathbb{P}}|\langle a,x\rangle^2 - b|. \\ &\min_x \Phi_S(x) \quad \text{where} \quad \Phi_S(x) &&= \mathbb{E}_{(a,b)\sim\mathbb{P}_m}|\langle a,x\rangle^2 - b| = \frac{1}{m}\sum_{i=1}^m |\langle a_i,x\rangle^2 - b_i|. \end{aligned} \tag{5.1}$$

This objective has been studied in the literature [8, 14, 37], yet tight bounds on the gap between the empirical and population subdifferential maps have yet to be established.

In the above, $x \in \mathbb{R}^d$ denotes the unknown signal vector, $a \in \mathbb{R}^d$ denotes the measurement vector, $b \in \mathbb{R}$ denotes the response, and $\mathbb{P}_m$ refers to the empirical distribution of $(a_1,b_1),(a_2,b_2),\ldots,(a_m,b_m)$ which are i.i.d. sampled from the distribution $\mathbb{P}$. This objective falls into the class of stochastic convex-composite objective in Section 4, as we can pick $\xi = (a,b) \in \mathbb{R}^d\times\mathbb{R}$, $h(z) = |z|$ and $c(x;\xi) = (a^Tx)^2 - b$. Given a radius $r > 0$, we are interested in a high probability bound onto the subdifferential gap for the empirical and population phase retrieval objectives over $x \in \mathrm{cl}(\mathbb{B}(0;r))$:

$$\sup_{x:\|x\|\le r} \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)).$$

We need to specify conditions on the distribution of measurements $(a,b) \sim \mathbb{P}$ that allow us to apply Theorem 5. We recall the concept of a subgaussian tail random vector [44, Section 2.5].

Definition 3. *A random vector Z is said to be σ-subgaussian if for every vector v with $\|v\| = 1$:*

$$\mathbb{E}\left[\exp\left(\frac{\langle Z,v\rangle^2}{\sigma^2}\right)\right] \le 2.$$

Examples of a σ-subgaussian distribution include normal distributions $Z \sim \mathsf{N}(0,(c\sigma)^2 I)$, and uniform distribution on the hypercube $Z \sim \mathrm{unif}\{\pm(c\sigma)\}^d$ where $c > 0$ can be any constant obeying $c \le 1/4$. Applying Theorem 5, we derive an upper bound on the subdifferential gap for the robust phase retrieval. Remarkably, our result, which achieves an almost optimal $m^{-1/2}$ rate (up to logarithmic factors), holds regardless of whether the random variables a and b are discrete or continuous. This contrasts with existing literature, which either requires a continuity assumption or attains a slower $m^{-1/4}$ rate (see Remark of Theorem 5).

Theorem 6. *Assume the measurement vector a is σ-subgaussian, and $\mathbb{E}[|b|^2] < \infty$.*
Then there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$ and $r > 0$:

$$\mathbb{P}^*\left(\sup_{x:\|x\|\le r} \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \ge C\sigma r\cdot\max\{\Delta_\Phi, \Delta_\Phi^2\}\right) \le \delta. \tag{5.2}$$

In the above,

$$\Delta_\Phi = \sqrt{\frac{1}{m}\cdot\left(d\log d\log m + \log(\frac{1}{\delta})\right)}.$$

Proof. We apply Theorem 5. We first check the assumptions. For the nonsmooth $h(z) = |z|$, it is decomposed as $h = h^{ns} + h^{sm}$ where $h^{sm}(z) = -z$ and $h^{ns}(z) = 2(z)_+$.

- (Assumption C.1). For the phase retrieval problem,

$$\nabla c(x,\xi) = 2\langle a,x\rangle a, \quad (h^{sm})'(c(x;\xi))\nabla c(x;\xi) = -2\langle a,x\rangle a.$$

Thus, at the center $x = 0 \in \mathbb{R}^d$, both gradients vanish: $\nabla c(0,\xi) = (h^{sm})'(c(0;\xi))\nabla c(0;\xi) = 0$.

- (Assumption C.2). It is easy to verify that

$$\begin{aligned} \nabla c(x_1,\xi) - \nabla c(x_2,\xi) &= 2\langle a, x_1 - x_2\rangle a \\ e(x_1,\xi) - e(x_2,\xi) &= 2\langle a, x_2 - x_1\rangle a \end{aligned},$$

where $e(x,\xi) = (h^{sm})'(c(x;\xi))\nabla c(x;\xi)$. Notably, for every vector v, $\langle a, v\rangle$ is $\|v\|$-subgaussian. Given the property that the product of two subgaussian random variables is subexponential [44, Lemma 2.7.7], we derive that for some universal constant $C > 0$, each of

$$\begin{aligned} \langle \nabla c(x_1,\xi) - \nabla c(x_2,\xi), v\rangle &= 2\langle a, x_1 - x_2\rangle\langle a, v\rangle \\ e(x_1,\xi) - e(x_2,\xi) &= 2\langle a, x_2 - x_1\rangle\langle a, v\rangle \end{aligned}$$

must be $C\|x_1 - x_2\|$ subexponential for every unit vector v with $\|v\| = 1$.

- (Assumption C.3). It is easy to verify that

$$\begin{aligned} \|\nabla c(x_1,\xi) - \nabla c(x_2,\xi)\| &= 2\|\langle a, x_1 - x_2\rangle a\| \le 2\|a\|^2\|x_1 - x_2\| \\ \|e(x_1,\xi) - e(x_2,\xi)\| &= 2\|\langle a, x_1 - x_2\rangle a\| \le 2\|a\|^2\|x_1 - x_2\| \end{aligned},$$

where $e(x,\xi) = (\psi^{sm})'(c(x;\xi))\nabla c(x;\xi)$. Since a is σ-subgaussian, $\mathbb{E}[L(\xi)] = 2\mathbb{E}[\|a\|^2] \le C\sigma^2 d$ for some universal constant $C > 0$. Thus Assumption C.3 is satisfied with $L(\xi) = 2\|a\|^2$.

- (Integrability). Notably $\mathbb{E}[|c(x;\xi)|^2] = \mathbb{E}[|b - \langle a, x\rangle|^2] < \infty$ for every $x \in \mathbb{R}^d$.

We then compute the bound in Theorem 5. Namely, we need to compute ζ and $\mathrm{vc}(\mathcal{F})$.

For the nonsmooth function $h(z) = |z|$, the corresponding value of ζ is given by

$$\zeta = 1 + h'_+(0) - h'_-(0) = 3,$$

as 0 is the only nondifferentiable point of h. The corresponding $\mathcal{F}$ is given by

$$\mathcal{F} = \left\{\{a \in \mathbb{R}^d, b \in \mathbb{R} : \langle a, x\rangle^2 - b \ge t\} \mid x \in \mathbb{R}^d, t \in \mathbb{R}\right\}.$$

We now bound its VC dimension using Theorem 8, whose proof is based on a beautiful theorem in real algebraic geometry [50–52]. Note $x \mapsto \langle a, x\rangle^2 - b$ is a degree 2 polynomial in $x \in \mathbb{R}^d$ for every a, b. Using Theorem 8, we obtain that for some universal constant $C > 0$:

$$\mathrm{vc}(\mathcal{F}) \le Cd\log(d).$$

Given $r > 0$, we apply Theorem 8 to the open ball $\mathbb{B}(0; 2r)$, and we obtain that for some universal constant $C > 0$:

$$\mathbb{P}^*\left(\sup_{x: x\in\mathbb{B}(0;2r)} \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \ge C\sigma r \cdot \max\{\Delta_\Phi, \Delta_\Phi^2\}\right) \le \delta. \tag{5.3}$$

Theorem 6 then follows by recognizing the set inclusion: $\mathrm{cl}(\mathbb{B}(0;r)) \subseteq \mathbb{B}(0;2r)$. □

In the literature on optimization, statistics and machine learning, establishing a uniform upper bound on the convergence of subdifferentials is important because it facilitates our understanding of the landscape of the nonsmooth, nonconvex finite sample objective ϕ_S, e.g., the location of stationary points, growth conditions, etc., through the lens of the population objective ϕ. We shall give further applications in Section 5.1.1 how Theorem 6 contributes to existing results to refine our understanding of the landscape of the nonsmooth phase retrieval objective Φ_S in finite samples.

Notably, characterization of the nonsmooth landscape of empirical risk function is an important and profound topic, and Section 5.1.1 presents only a very basic example of how to apply our general theorem. Further results on this topic will be detailed in an upcoming manuscript.

5.1.1 Noiseless Phase Retrieval

We consider the noiseless phase retrieval, where the measurement b obeys:

$$b = \langle a, \bar{x} \rangle^2$$

for some fixed vector $\bar{x} \in \mathbb{R}^d$. In this case, the objectives Φ_S and Φ become:

$$\Phi_S(x) = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x \rangle^2 - \langle a_i, \bar{x} \rangle^2|, \quad \Phi(x) = \mathbb{E}_{a \sim \mathbb{P}}[|\langle a, x \rangle^2 - \langle a, \bar{x} \rangle^2|].$$

Let us denote the set of stationary points of Φ_S and Φ to be:

$$\mathcal{Z}_S = \{x \in \mathbb{R}^d : 0 \in \partial \Phi_S(x)\} \quad \text{and} \quad \mathcal{Z} = \{x \in \mathbb{R}^d : 0 \in \partial \Phi(x)\}.$$

For standard normal measurement vectors $a \sim \mathsf{N}(0, I)$, a recent important and interesting study first characterizes $\mathcal{Z}$, the locations of stationary points of the population objective $\Phi(x)$ [14]. Remarkably, the authors further develop a quantitative version of Attouch's epi-convergence theorem to show that the stationary points of the finite samples objective Φ_S converge to those of the population objective Φ at a rate of $\sqrt[4]{d/m}$ [14]. More precisely, [14, Theorem 5.2, Corollary 6.3] prove that, there exist numerical constants $c, C > 0$ such that when $m \geq Cd$, with an inner probability at least $1 - C\exp(-cd)$ (recall the definition of inner probability from Section 2.4):

$$\mathbb{D}(\mathcal{Z}_S, \mathcal{Z}) \leq C \sqrt[4]{\frac{d}{m}} \|\bar{x}\| .$$

In words, with high probability, for any stationary point of Φ_S, there is a stationary point of Φ that is at most a distance of $C\sqrt[4]{d/m}\,\|\bar{x}\|$ in the ℓ_2 norm away from it. It is important to note that the slow rate in m, namely $1/\sqrt[4]{m}$, appears intrinsic to the approach based on the quantitative version of Attouch's epi-convergence theorem [16, Section 5].

In this paper, we show a stronger result for large sample size m: the stationary points of the finite samples objective Φ_S converge to those of the population objective Φ at a rate of $\sqrt{d/m}$ up to logarithmic factors in d, m. This rate improvement in m is due to our bound on uniform convergence of subdifferentials of Φ and Φ_S in Theorem 6, and our proof of Theorem 7 benefits from the characterization of approximate stationary points of the population objective Φ [14, Corollary 5.3]. Nonetheless, our result, Theorem 7, comes with a tradeoff of additional logarithmic factors in d and m.

Theorem 7. *Assume $a \sim \mathsf{N}(0, I)$. Then there exist numerical constants $c, C > 0$ such that when $m \geq Cd$, with an inner probability at least $1 - C\exp(-cd)$:*

$$\mathbb{D}(\mathcal{Z}_S, \mathcal{Z}) \leq C \left(\sqrt{\frac{d}{m} \log(d) \log(m)} + \frac{d}{m} \log(d) \log(m) \right) \|\bar{x}\| .$$

The formal proof of Theorem 7 is deferred to Appendix, Section B.

5.2 Illustration II: Matrix Sensing

Our second illustration demonstrates the flexibility of our framework. Consider the robust matrix sensing objective [10, 11, 41]:

$$\begin{aligned} &\min_X \bar{\Phi}(X) \quad \text{where} \quad \bar{\Phi}(X) \ = \mathbb{E}_{(a,b)\sim\mathbb{P}}|\langle A, XX^T\rangle - b|. \\ &\min_X \bar{\Phi}_S(X) \quad \text{where} \quad \bar{\Phi}_S(X) = \mathbb{E}_{(a,b)\sim\mathbb{P}_m}|\langle A, XX^T\rangle - b| = \frac{1}{m}\sum_{i=1}^m |\langle A_i, XX^T\rangle - b_i|. \end{aligned} \tag{5.4}$$

Here, $A \in \mathbb{R}^{D\times D}$ represents the measurement matrices, $b \in \mathbb{R}$ the measurements, and $X \in \mathbb{R}^{D\times r_0}$ a low-rank matrix, typically with r_0 significantly smaller than D in applications. This objective falls into the stochastic convex-composite objectives in Section 4, as we can simply set $x = X \in \mathbb{R}^{D\times r_0}$, $\xi = (A, b) \in \mathbb{R}^{D\times D} \times \mathbb{R}$, $h(z) = |z|$ and $c(x; \xi) = c(x; (A, b)) = \langle A, XX^T\rangle - b$.

Another application of Theorem 5 yields an upper bound on the subdifferential gap for the robust matrix sensing objective. Recall a matrix $A \in \mathbb{R}^{D\times D}$ is a σ-subgaussian random matrix if the vectorized matrix $\text{vec}(A) \in \mathbb{R}^{D\times D}$ is a σ-subgaussian vector in the sense of Definition 3 [44]. For clarity in notation, we use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix $A \in \mathbb{R}^{D\times D}$.

Corollary 1. *Assume the measurement matrix $A \in \mathbb{R}^{D\times D}$ is σ-subgaussian, and $\mathbb{E}[|b|^2] < \infty$. Then there exists a universal constant $C > 0$ such that for every $\delta \in (0, 1)$ and $r > 0$:*

$$\mathbb{P}^* \left(\sup_{X:\|X\|_F \le r} \mathbb{H}(\partial\bar{\Phi}(X), \partial\bar{\Phi}_S(X)) \ge C\sigma r \cdot \max\{\Delta_{\bar{\Phi}}, \Delta_{\bar{\Phi}}^2\} \right) \le \delta. \tag{5.5}$$

In the above,

$$\Delta_{\bar{\Phi}} = \sqrt{\frac{1}{m} \cdot \left(Dr_0 \log(Dr_0) \log m + \log(\frac{1}{\delta}) \right)}.$$

The proof is provided in Section C and is analogous to the phase retrieval case. The primary challenge, which comes from evaluating the VC dimension, is addressed using Theorem 8.

6 VC Dimension Bounds

We establish a result on upper bounding the VC dimension of the set $\mathcal{F}$ defined in equation (4.12) when all the functions $x \mapsto c(x; \xi)$ are polynomials. Notably, this result does not require $\xi \mapsto c(x; \xi)$ to be polynomials in the data ξ.

Theorem 8. *Consider $x \in \mathbb{R}^d$ and $\xi \in \Xi$, with each mapping $x \mapsto c(x; \xi)$ being a polynomial of degree at most K for every $\xi \in \Xi$, and $K \ge 1$. The VC dimension of the set $\mathcal{F}$ (cf. equation (4.12))*

$$\mathcal{F} = \{\{\xi \in \Xi : c(x, \xi) \ge t\} \mid x \in \mathbb{R}^d, t \in \mathbb{R}\}.$$

is bounded by $Cd \log(Kd)$, where $C > 0$ is a universal constant.

Definition 4. *Let $p_1, p_2, \dots, p_N$ be a sequence of functions where each function $p_i : \mathbb{R}^d \to \mathbb{R}$. A vector $\sigma \in \{-1, 0, +1\}^N$ is called a sign pattern of $p_1, p_2, \dots, p_N$ if there exists an $x \in \mathbb{R}^d$ such that the sign of $p_i(x)$ is σ_i for all $i = 1, 2, \dots, n$, where σ_i is the ith coordinate of σ.*

The following theorem, rooted in the field of real algebraic geometry, is referenced from Matousek's monograph on discrete geometry [52]. See [50, 51] and [53, Theorem 3].

Theorem 9 ([52, Theorem 6.2.1]). *Let $p_1, p_2, \ldots, p_N$ be d-variate real polynomials of degree at most K. The number of sign patterns of $p_1, p_2, \ldots, p_N$ is bounded by*

$$\left(\frac{50KN}{d}\right)^d.$$

Proof of Theorem 8. Define $z = (x,t) \in \mathbb{R}^d \times \mathbb{R} \cong \mathbb{R}^{d+1}$. Given a dataset $\xi_1, \xi_2, \ldots, \xi_N$, let us define the functions $q_i(z) = q_i(x,t) = c(x;\xi_i) - t$ for $1 \le i \le N$. Notably, each q_i is a $d+1$-variate polynomial with degree at most K.

Consider the set of binary patterns b defined by:

$$\mathcal{B} = \{(b_1, b_2, \ldots, b_N) \in \{0,1\}^N : \\ \text{there exists } z \in \mathbb{R}^{d+1} \text{ such that } b_i = \mathbf{1}\{q_i(z) \ge 0\} \text{ for all } 1 \le i \le N\}.$$

By definition, the number of possible binary patterns, $|\mathcal{B}|$, cannot exceed the number of sign patterns for the same polynomials $q_1, q_2, \ldots, q_N$. According to Theorem 9, the number of sign patterns is bounded by $(50KN/(d+1))^{d+1}$. Therefore, we conclude:

$$|\mathcal{B}| \le \left(\frac{50KN}{d+1}\right)^{d+1}.$$

Remarkably, this bound on $|\mathcal{B}|$ is valid for any choice of the data points $\xi_1, \xi_2, \ldots, \xi_N$. To establish an upper limit on the VC dimension of $\mathcal{F}$, it remains to analyze the set of possible integers N that satisfies the following inequality:

$$\left(\frac{50KN}{d+1}\right)^{d+1} < 2^N.$$

A straightforward algebraic manipulation reveals that there exists a universal constant $C > 0$ such that setting $N > Cd\log(Kd)$ ensures the inequality holds. The conclusion is that for any dataset $\xi_1, \xi_2, \ldots, \xi_N$ with size $N > Cd\log(Kd)$, the total number of binary patterns generated by $q_i(z) = c(x;\xi_i) - t$ with x ranging over $\mathbb{R}^d$ and t over $\mathbb{R}$ is strictly fewer than 2^N.

This proves the bound on VC dimension: $\mathrm{vc}(\mathcal{F}) \le Cd\log(Kd)$ as desired. □

7 Discussions

This paper presents a general technique for proving subdifferential convergence in weakly convex functions through the establishment of subgradient convergence (Theorem 1). We demonstrate this approach to stochastic convex-composite minimizations—an important class of weakly convex objectives—and derive concrete convergence rates for subdifferentials of empirical objectives using tools from statistical learning theories (Theorem 5). Our results achieve optimal rates with tight dependence on the sample size and dimensionality (up to logarithmic factors) and do not rely on key distributional assumptions in the literature requiring the subdifferential mappings to be continuous under the Hausdorff metric. In Section 5, we demonstrate how our results lead to complementary understanding of subdifferential convergence and its uses for landscape analysis, in more specific applications such as robust phase retrieval and matrix sensing problems.

There are many interesting future directions to extend the scope of the current work. Notable examples include:

- Can Theorem 5 be extended to other weakly-convex models, such as max-of-smooth functions [54, 55]? Extending this theorem could provide tools for solving other weakly-convex optimization problems.

- Can Theorem 1 be extended to other set-valued mappings, such as maximal monotone operators? Extending this theorem could enhance our understanding of solutions to certain stochastic variational inequalities [56].
- Theorem 5 could facilitate our understanding of the landscape of a certain class of nonsmooth empirical risk minimizations. Exploring its statistical applications, such as nonlinear quantile regressions [13, 57], may lead to new insights into data-driven models.

Acknowledgements. Feng Ruan extends his gratitude to his friends and advisor—John Duchi, Qiyang Han, X.Y. Han, Kate Yao and Yichen Zhang—for their support and patience as he tackled this once seemingly challenging problem. A shift in perspective during his time at Northwestern led him to this solution. He is also grateful to Damek Davis and Dmitriy Drusvyatskiy for their valuable feedback on the initial manuscript draft posted on arXiv.

Appendix A Proof of Theorem 4

In this section, we prove Theorem 4. The key to our proof is to establish high probability upper bounds of the following two quantities:

$$\Gamma_1 = \sup_{x \in \mathbb{B}(x_0;r), t\in\mathbb{R}} \|\mathbb{E}_{\xi\sim\mathbb{P}_m}[\mathbf{1}\{c(x;\xi) \geq t\} \cdot \nabla c(x;\xi)] - \mathbb{E}_{\xi\sim\mathbb{P}}[\mathbf{1}\{c(x;\xi) \geq t\} \cdot \nabla c(x;\xi)]\|$$

$$\Gamma_2 = \sup_{x \in \mathbb{B}(x_0;r)} \|\mathbb{E}_{\xi\sim\mathbb{P}_m}[(h^{sm})'(c(x;\xi)) \cdot \nabla c(x;\xi)] - \mathbb{E}_{\xi\sim\mathbb{P}}[(h^{sm})'(c(x;\xi)) \cdot \nabla c(x;\xi)]\|$$

Our interests on these two quantities Γ_1, Γ_2 are motivated by the following inequality:

$$\sup_{x\in\mathbb{B}(x_0;r)} \|G_S(x) - G(x)\| \leq \Big(\sum_{j=1}^{\infty} a_j\Big) \cdot \Gamma_1 + \Gamma_2. \tag{A1}$$

Below, we will establish the existence of a universal constant $C > 0$ such that for every $\delta \in (0,1)$:

$$\mathbb{P}^*(\Gamma_1 \geq C\sigma r \cdot \max\{\Delta, \Delta^2\}) \leq \delta, \quad \mathbb{P}^*(\Gamma_2 \geq C\sigma r \cdot \max\{\Delta, \Delta^2\}) \leq \delta \tag{A2}$$

where Δ has the same definition as in equation (4.16):

$$\Delta = \sqrt{\frac{1}{m} \cdot \Big(d + \mathrm{vc}(\mathcal{F}) \log m + \log(\frac{1}{\delta})\Big)}.$$

Given these high probability bounds on Γ_1, Γ_2 in equation (A2), Theorem 4 follows by plugging these high probability bounds into equation (A1).

It remains to prove the high probability bounds on Γ_1 and Γ_2 as specified in equation (A2). To unify the proofs for these two bounds, we utilize Theorem 10, which details the uniform convergence rate involving indicator functions and smooth functions. The proof of Theorem 10 is in Section A.1.

Theorem 10. *Consider two subsets of $\mathbb{R}^d$, $\mathcal{Z}_1$ and $\mathcal{Z}_2$, where $\mathcal{Z}_2$ is a Euclidean ball centered at $z_{2,0}$ with radius r. Let $o : \mathcal{Z}_1 \times \Xi \to \mathbb{R}$ and $w : \mathcal{Z}_2 \times \Xi \to \mathbb{R}^d$. Suppose $\xi_1, \xi_2, ..., \xi_m$ are i.i.d. samples from a distribution $\mathbb{P}$ with support Ξ. For any $z_1 \in \mathcal{Z}_1$ and $z_2 \in \mathcal{Z}_2$, we define:*

$$\begin{aligned} K_m(z_1, z_2, t) &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{o(z_1, \xi_i) \geq t\}\, w(z_2, \xi_i), \\ K(z_1, z_2, t) &= \mathbb{E}_{\xi\sim\mathbb{P}}\Big[\mathbf{1}\{o(z_1, \xi) \geq t\}\, w(z_2, \xi)\Big] \end{aligned}.$$

Assume the following:
(a) The random variable $w(z_{2,0}, \xi)$ is sub-exponential with parameter σ^2.
(b) For every $z_2, z_2' \in \mathcal{Z}_2$, the random vector $w(z_2, \xi) - w(z_2', \xi)$ is $\sigma \|z_2 - z_2'\|$ subexponential.
(c) The random function $z_2 \mapsto w(z_2, \xi)$ and $z_1 \mapsto o(z_1, \xi)$ are continuous.
Under these conditions, there exists a universal constant $C > 0$ such that for every $\delta \in (0, 1)$:

$$\mathbb{P}^*\Bigg(\sup_{z_1\in\mathcal{Z}_1, z_2\in\mathcal{Z}_2, t\in\mathbb{R}} \|K_m(z_1, z_2, t) - K(z_1, z_2, t)\| \geq C\sigma r \cdot \max\{\Delta_{\mathcal{W}}, \Delta_{\mathcal{W}}^2\}\Bigg) \leq \delta \tag{A3}$$

where

$$\Delta_{\mathcal{W}} = \sqrt{\frac{1}{m} \cdot \Big(d + \mathrm{vc}(\mathcal{W}) \log(m) + \log(\frac{1}{\delta})\Big)}.$$

Here, $\mathcal{W} := \{\{\xi \in \Xi : o(z_1, \xi) \geq t\} \mid z_1 \in \mathcal{Z}_1, t \in \mathbb{R}\}$*, and* $\mathrm{vc}(\mathcal{W})$ *denotes its VC dimension.*

We will apply Theorem 10 for two different sets of functions o, w to deduce the high probability bounds on Γ_1, Γ_2 in equation (A2). Below we set $\mathcal{Z}_1 = \mathbb{R}^d$ and $\mathcal{Z}_2 = \mathbb{B}(x_0; r)$.

First, we set $o(x, \xi) = c(x, \xi)$ and $w(x', \xi) = \nabla c(x', \xi)$. Given Assumptions C.1 and C.2, the conditions of Theorem 10 are satisfied. Applying this theorem yields the following guarantee. Let

$$\Gamma_1' = \sup_{x \in \mathbb{R}^d, t \in \mathbb{R}, x' \in \mathbb{B}(x_0; r)} \|\mathbb{E}_{\xi \sim \mathbb{P}_m}[\mathbf{1}\{c(x; \xi) \geq t\} \cdot \nabla c(x'; \xi)] - \mathbb{E}_{\xi \sim \mathbb{P}}[\mathbf{1}\{c(x; \xi) \geq t\} \cdot \nabla c(x'; \xi)]\|$$

Then there exists a universal constant $C > 0$ such that for every $\delta \in (0, 1)$:

$$\mathbb{P}^*(\Gamma_1' \geq C\sigma r \cdot \max\{\Delta, \Delta^2\}) \leq \delta,$$

where Δ is defined as in equation (4.16). Since $\Gamma_1 \leq \Gamma_1'$ by definition, this establishes the first inequality of equation (A2).

Next, we set $o(x, \xi) \equiv 0$ and $w(x', \xi) = (h^{sm})'(c(x'; \xi)) \cdot \nabla c(x'; \xi)$. Given Assumptions C.1 and C.2, the conditions of Theorem 10 are satisfied. Applying this theorem provides the following guarantee. Let

$$\Gamma_2' = \sup_{t \in \mathbb{R}, x' \in \mathbb{B}(x_0; r)} \|\mathbb{E}_{\xi \sim \mathbb{P}_m}[\mathbf{1}\{0 \geq t\} \cdot (h^{sm})'(c(x'; \xi)) \nabla c(x'; \xi)] - \mathbb{E}_{\xi \sim \mathbb{P}}[\mathbf{1}\{0 \geq t\} \cdot (h^{sm})'(c(x'; \xi)) \nabla c(x'; \xi)]\|$$

There exists a universal constant $C > 0$ such that for every $\delta \in (0, 1)$:

$$\mathbb{P}^*(\Gamma_2' \geq C\sigma r \cdot \max\{\Delta, \Delta^2\}) \leq \delta,$$

where Δ is defined as in equation (4.16). Since $\Gamma_2 \leq \Gamma_2'$ by definition, this establishes the second inequality of equation (A2).

At this point, the proof of Theorem 4 is complete.

A.1 Proof of Theorem 10

Our proof of Theorem 10 combines two well-established techniques for proving uniform bounds on the supremum of a random process in the statistical learning theory: chaining, which constructs a sequence of progressively finer approximations to the continuous process, and Sauer-Shelah Lemma, a combinatorial principle using the VC dimension to control the complexity of the function class under consideration [31, Chapter 2]. The chaining argument specifically addresses the continuous parameter z_2, and the complexity of its increments is then controlled using the VC dimension technique applied to the function class specified by the parameters z_1, t. To deal with measurability issues, we use the calculus of *outer integral* [31, Chapter 1.2]. A crucial part of our proof demonstrates that our process becomes measurable after symmetrization, building on ideas in the monograph [31, Chapter 2.3].

Given the subtlety in dealing with measurability, and differences from existing literature, which typically focuses either on chaining or on VC dimension arguments applied to indicator functions, we provide detailed explanations of our combined approach to ensure clarity and completeness.

A.1.1 Chaining Argument

For notational simplicity, below we denote:

$$R_m(z_1, z_2, t) = K_m(z_1, z_2, t) - K(z_1, z_2, t).$$

The main idea in the chaining argument involves discretizing the parameter z_2 into increasingly finer subsets, constructing chains to approximate the supremum of $R_m(z_1, z_2, t)$, which allows more tractable probabilistic analysis.

Let $\epsilon_k = r2^{-k}$ for every $k \geq 0$. An ϵ_k-cover of $\mathcal{Z}_2$, denoted by $\mathcal{Z}_{2,\epsilon_k}$, enables the approximation of every point $z_2 \in \mathcal{Z}_2$ by a point $\pi_k(z_2) \in \mathcal{Z}_{2,\epsilon_k}$ such that $\|z_2 - \pi_k(z_2)\| \leq \epsilon_k$. Specifically, we take the initial cover $\mathcal{Z}_{2,\epsilon_0} = \{z_{2,0}\}$ to consist solely of the center $z_{2,0}$ of the ball $\mathcal{Z}_2$. By employing a standard volume-type argument, we can further select the cover $\mathcal{Z}_{2,\epsilon_k}$ such that $\log|\mathcal{Z}_{2,\epsilon_k}| \leq p\log(2 + r/\epsilon_k)$ for every $k \geq 1$ [44, Corollary 4.2.13].

Using these coverings, we can decompose $R_m(z_1, z_2, t)$ through a chaining argument [45]:

$$R_m(z_1, z_2, t) = R_m(z_1, z_{2,0}, t) + \sum_{k=1}^{\infty} \Big[R_m(z_1, \pi_k(z_2), t) - R_m(z_1, \pi_{k-1}(z_2), t)\Big].$$

This decomposition is valid because the limit $R_m(z_1, z_2, t) = \lim_{k\to\infty} R_m(z_1, \pi_k(z_2), t)$ holds with probability one, as $z_2 \mapsto R_m(z_1, z_2, t)$ is continuous by assumption. Applying the supremum over z_1, z_2, t after taking the norm $\|\cdot\|$, and using the triangle inequality we obtain:

$$\begin{aligned}\sup_{z_1\in\mathcal{Z}_1, z_2\in\mathcal{Z}_2, t\in\mathbb{R}} \|R_m(z_1, z_2, t)\| \leq &\sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \|R_m(z_1, z_{2,0}, t)\| \\ &+ \sum_{k=1}^{\infty} \sup_{z_1\in\mathcal{Z}_1, z_2\in\mathcal{Z}_2, t\in\mathbb{R}} \|R_m(z_1, \pi_k(z_2), t) - R_m(z_1, \pi_{k-1}(z_2), t)\|.\end{aligned} \tag{A4}$$

We will bound the tail behavior of every single term in this sum. To this end, the following lemma is useful. For every $\delta \in (0, 1)$, define:

$$\Delta(\delta) := \sqrt{\frac{1}{m} \cdot \left(d + \mathrm{vc}(\mathcal{W})\log(m) + \log(\frac{1}{\delta})\right)}. \tag{A5}$$

Lemma 3. *There exists a universal constant $C > 0$ such that for any given pair $z_2, z_2' \in \mathcal{Z}_2$, the following occurs with an inner probability at least $1 - \delta$:*

$$\sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \|R_m(z_1, z_2, t) - R_m(z_1, z_2', t)\| \leq C\sigma \cdot \|z_2 - z_2'\| \cdot \max\{\Delta(\delta), \Delta(\delta)^2\}.$$

Similarly, for the center $z_{2,0}$ of $\mathcal{Z}_2$, the following occurs with an inner probability at least $1 - \delta$:

$$\sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \|R_m(z_1, z_{2,0}, t)\| \leq C\sigma \cdot \max\{\Delta(\delta), \Delta(\delta)^2\}.$$

The proof of Lemma 3 is deferred to Section A.1.2.

Back to our bound of the tail behavior of every term on the RHS of equation (A4). Fix $\delta \in (0, 1)$. Let us denote $\delta_0 = \delta/2$ and

$$\delta_k = \delta \cdot 2^{-(k+1)} \cdot (|\mathcal{Z}_{2,\epsilon_k}||\mathcal{Z}_{2,\epsilon_{k-1}}|)^{-1}$$

for each $k \geq 1$. Below we use C to denote the universal constant that appears in Lemma 3.

First, by Lemma 3, with an inner probability at least $1 - \delta/2$:

$$\sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \|R_m(z_1, z_{2,0}, t)\| \leq C\sigma \cdot \max\{\Delta(\delta_0), \Delta(\delta_0)^2\}.$$

This controls the first term on the RHS of equation (A4).

Next, the adjustment of δ into δ_k allows the use of a union bound to extend the probability guarantees from Lemma 3 from individual pairs to all pairs within the cover sets $\mathcal{Z}_{2,\epsilon_k}$ and $\mathcal{Z}_{2,\epsilon_{k-1}}$. Specifically, for each $k \geq 1$, with an inner probability at least $1 - \delta \cdot 2^{-(k+1)}$, there is the bound:

$$\sup_{z_1 \in \mathcal{Z}_1, z_2 \in \mathcal{Z}_2, t \in \mathbb{R}} \|R_m(z_1, \pi_k(z_2), t) - R_m(z_1, \pi_{k-1}(z_2), t)\| \leq 3C\sigma\epsilon_k \cdot \max\{\Delta(\delta_k), \Delta(\delta_k)^2\}.$$

This result holds because the cardinality of pairs of $\{(\pi_k(z_2), \pi_{k-1}(z_2))\}$ over $z_2 \in \mathcal{Z}_2$ is capped by $|\mathcal{Z}_{2,\epsilon_k}||\mathcal{Z}_{2,\epsilon_{k-1}}|$, and there is the bound $\|\pi_k(z_2) - \pi_{k-1}(z_2)\| \leq 3\epsilon_k$ that holds for every $z_2 \in \mathcal{Z}_2$.

By substituting all the above high probability bounds into the RHS of equation (A4), and using the union bounds, noting $\sum_{k=0}^{\infty} \delta \cdot 2^{-(k+1)} = \delta$, we obtain with an inner probability at least $1 - \delta$:

$$\sup_{z_1 \in \mathcal{Z}_1, z_2 \in \mathcal{Z}_2, t \in \mathbb{R}} \|R_m(z_1, z_2, t)\| \leq C\sigma \cdot \max\{\Delta(\delta_0), \Delta(\delta_0)^2\} + 3C\sigma \cdot \sum_{k=1}^{\infty} \epsilon_k \cdot \max\{\Delta(\delta_k), \Delta(\delta_k)^2\}.$$

To accumulate terms on the RHS, we deduce a basic bound whose proof is deferred to Section A.1.3.

Lemma 4. *For some universal constant $c > 0$:*

$$\sum_{k=1}^{\infty} \epsilon_k \cdot \max\{\Delta(\delta_k), \Delta(\delta_k)^2\} \leq cr \max\{\Delta(\delta), \Delta(\delta)^2\}.$$

Ultimately, this proves the existence of a universal constant $\bar{C} > 0$ such that

$$\sup_{z_1 \in \mathcal{Z}_1, z_2 \in \mathcal{Z}_2, t \in \mathbb{R}} \|R_m(z_1, z_2, t)\| \leq \bar{C}\sigma r \cdot \max\{\Delta(\delta), \Delta(\delta)^2\}$$

holds with probability at least $1 - \delta$. The proof of Theorem 10 is then complete.

A.1.2 Bounding Individual Term using VC-Dimension (Proof of Lemma 3)

We give a unifying proof for the two statements in Lemma 3. To simplify the notation, we define:

$$\psi_{z_1,t}(\xi) = \mathbf{1}\{o(z_1, \xi) \geq t\}.$$

We will show that it suffices to prove the following more general statement.

Lemma 5. *There exists a universal constant $C > 0$ such that the following holds. Define:*

$$W_m(z_1, t) = \frac{1}{m} \sum_{i=1}^{m} \psi_{z_1,t}(\xi_i) u(\xi_i) - \mathbb{E}[\psi_{z_1,t}(\xi) u(\xi)]. \tag{A6}$$

where $u(\xi)$ in $\mathbb{R}^d$ is $\bar{\sigma}$ subexponential. Then it happens with an inner probability at least $1 - \delta$:

$$\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1, t)\| \leq C\bar{\sigma} \cdot \max\{\Delta(\delta), \Delta(\delta)^2\}.$$

where $\Delta(\delta)$ is defined in equation (A5).

We will first demonstrate how Lemma 5 implies Lemma 3 and defer the proof of Lemma 5 to the end. Introduce two notation:

$$u_{z_2,z_2'}(\xi) = w(z_2, \xi) - w(z_2', \xi), \quad u_{z_{2,0}}(\xi) = w(z_{2,0}, \xi).$$

These notation lead us to the following critical representation:

$$\begin{aligned} R_m(z_1, z_2, t) - R_m(z_1, z_2', t) &= \frac{1}{m}\sum_{i=1}^{m} \psi_{z_1,t}(\xi_i)u_{z_2,z_2'}(\xi_i) - \mathbb{E}\left[\psi_{z_1,t}(\xi)u_{z_2,z_2'}(\xi)\right] \\ R_m(z_1, z_{2,0}, t) &= \frac{1}{m}\sum_{i=1}^{m} \psi_{z_1,t}(\xi_i)u_{z_{2,0}}(\xi_i) - \mathbb{E}\left[\psi_{z_1,t}(\xi)u_{z_{2,0}}(\xi)\right] \end{aligned}. \tag{A7}$$

Note the similarities between the expressions in equation (A6) and in equation (A7). Because $u_{z_2,z_2'}(\xi)$ and $u_{z_{2,0}}(\xi)$ are subexponential random vectors, the conditions of Lemma 3 are met. More precisely, $u_{z_2,z_2'}(\xi)$ is $\sigma \|z_2 - z_2'\|$ subexponential for every pair of $z_2, z_2' \in \mathcal{Z}_2$, and $u_{z_{2,0}}(\xi)$ is σ subexponential. This allows us to leverage Lemma 5 to conclude that Lemma 3 follows.

In the remainder, we prove Lemma 5. Our proof exploits the fact that the VC dimension of the set family $\mathcal{W}$ controls the Rademacher complexity associated with the indicators $\psi_{z_1,t}(\xi)$. Additionally, we utilize calculus rules for *outer probability measures* [31, Section 1.2-5] to address measurability issues. Recall the notion of *outer integral* [31, Section 1.2].

Definition 5. *Given a probability space* $(\Xi, \mathscr{G}, \mathbb{P})$*, and an arbitrary map* $T : \Xi \mapsto \mathbb{R} \cup \{+\infty\}$*, we define the outer integral of* T *by:*

$$\mathbb{E}^*[T] = \inf\{\mathbb{E}[U] : U \geq T, \quad U : \Xi \mapsto \mathbb{R} \cup \{+\infty\} \text{ is measurable, and } \mathbb{E}[U] \text{ exists}\}.$$

Outer integrals exhibit monotone property: if $T_1(\xi) \leq T_2(\xi)$ for every ξ, then $\mathbb{E}^*[T_1] \leq \mathbb{E}^*[T_2]$. This is essentially all we need, as the remaining properties of outer integral will be reference from the monograph [31]. To ensure clarity when discussing the calculus rules of the *outer integral*, we will explicitly reference the relevant chapters in the monograph [31].

Proof of Lemma 5. We define the moment-generating function, using the notion of outer integral (see Definition 5):

$$U(\tau) := \mathbb{E}^*\left[\exp\left(\tau \cdot \sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1, t)\|\right)\right]. \tag{A8}$$

We aim to find a tight upper bound for $U(\tau)$ for every $\tau > 0$, which will allow us to infer a high probability upper bound under the outer measure onto the random variable:

$$\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1, t)\|.$$

Let $V_{1/2}$ denote a 1/2 cover of unit ball in $\mathbb{R}^d$ with $\log |V_{1/2}| \leq d \log 6$. Following a standard argument (see, e.g., [44, Exercise 4.2.2]), we know that for every $\bar{v} \in \mathbb{R}^d$:

$$\|\bar{v}\| \leq 2 \max_{v \in V_{1/2}} \langle \bar{v}, v \rangle.$$

Applying this inequality to $W_m(z_1, t)$ for every $z_1 \in \mathcal{Z}_1$ and $t \in \mathbb{R}$, we obtain:

$$U(\tau) \leq \mathbb{E}^*\left[\exp\left(2\tau \cdot \sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \max_{v \in V_{1/2}} \langle W_m(z_1, t), v \rangle\right)\right] \tag{A9}$$

We employ symmetrization techniques from empirical process theory to upper bound the RHS. To simplify the notation, we introduce a function $u_v(\xi) = \langle u(\xi), v\rangle$, which allows us to express:

$$\langle W_m(z_1,t), v\rangle = \frac{1}{m}\sum_{i=1}^{m} \psi_{z_1,t}(\xi_i)u_v(\xi_i) - \mathbb{E}\left[\psi_{z_1,t}(\xi)u_v(\xi)\right].$$

Furthermore, we introduce i.i.d. Rademacher random variables $\{\varepsilon_i\}_{i=1}^m$ with $\mathbb{P}(\varepsilon_i = \pm 1) = 1/2$, independent of the data samples $\{\xi_i\}_{i=1}^m$. To simplify our notation, we use $\varepsilon_{1:m}$ to denote the vector $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ and $\xi_{1:m}$ to denote the vector $(\xi_1, \xi_2, \ldots, \xi_m)$.

Applying symmetrization techniques from empirical process theory for the outer integral [31, Lemma 2.3.1], we deduce an upper bound for $U(\tau)$:

$$U(\tau) \le \mathbb{E}^*\left[\exp\left(4\tau \cdot \sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \max_{v\in V_{1/2}} \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i)u_v(\xi_i)\right)\right] \tag{A10}$$

where $\mathbb{E}^*$ is understood as the outer integral on the product probability space involving the data samples $\xi_{1:m}$ and the independent Rademacher random variables $\varepsilon_{1:m}$. To be more formal, we follow [31, Chapter 2.3]. Recall that a single random variable ξ can be defined on the probability space $(\Xi, \mathscr{G}, \mathbb{P})$. Thus, we can define our data samples $\xi_{1:m}$ on the product probability space $(\Xi^{\otimes m}, \sigma(\mathscr{G}^{\otimes m}), \mathbb{P}^{\otimes m})$, where $\Xi^{\otimes m}$ represents the m-fold Cartesian product of Ξ, $\sigma(\mathscr{G}^{\otimes m})$ represents the σ-field generated by the m-fold product of $\mathscr{G}$, and $\mathbb{P}^{\otimes m}$ is the m-fold product of $\mathbb{P}$, where $\xi_1, \xi_2, \ldots, \xi_m$ corresponds to the coordinate projections. Then, we introduce Rademacher random variables $\varepsilon_{1:m} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$ defined on another probability space $(\Xi', \mathscr{H}, \mathbb{P}')$. We finally combine these to define the final product probability space of interest:

$$(\Xi^{\otimes m} \times \Xi', \sigma(\mathscr{G} \otimes \mathscr{H}), \mathbb{P}^{\otimes m} \otimes \mathbb{P}'). \tag{A11}$$

In this product probability space, the sample $\xi_{1:m}$ corresponds to the first m coordinates projections and the additional Rademacher random variables $\varepsilon_{1:m}$ depending only on the $(m+1)$st coordinate. The outer integral $\mathbb{E}^*$ in equation (A10) then refers to the outer integral for this probability space.

Lemma 6 is the critical component that resolves the measurability issue. From the empirical process theory, we know that measurability of certain forms is needed at this point, as Fubini's theorem is not valid for outer integrals [31, Chapter 2.3]. Lemma 6 follows the established ideas to resolve the measurability issues, see, e.g., discussions in the monograph on the measurable class [31, Definition 2.3.3]. The proof of Lemma 6 is deferred to Section A.1.4.

Lemma 6. *The mapping:*

$$(\xi_{1:m}, \varepsilon_{1:m}) \mapsto \sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \max_{v\in V_{1/2}} \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i)u_v(\xi_i)$$

is measurable on the product probability space defined in equation (A11).

Lemma 6 provides the justification for replacing the outer integral $\mathbb{E}^*$ on the RHS of equation (A10) with the standard integral $\mathbb{E}$. Indeed, equation (A10) now leads to

$$U(\tau) \le \mathbb{E}\left[\exp\left(4\tau \cdot \sup_{z_1\in\mathcal{Z}_1, t\in\mathbb{R}} \max_{v\in V_{1/2}} \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i)u_v(\xi_i)\right)\right] \tag{A12}$$

where the expectation on the RHS now is the standard expectation, taken jointly over the samples $\xi_{1:m}$ and the independent Rademacher random variables $\xi_{1:m}$.

To further upper bound the expectation on the RHS of inequality (A10), we first consider conditioning on the samples $\xi_{1:m}$. This means that we condition on the values of $\xi_1, \xi_2, \ldots, \xi_m$, and study the quantity:

$$V(\tau, \xi_{1:m}) = \mathbb{E}_{\varepsilon_{1:m}}\left[\exp\left(\tau \cdot \sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i) u_v(\xi_i)\right)\right]$$

where the symbol $\mathbb{E}_{\varepsilon_{1:m}}$ means the standard expectation taken solely over the Rademacher variables $\varepsilon_{1:m}$. As the Fubini's theorem holds for the standard expectation, under this notation, the inequality (A12) then becomes:

$$U(\tau) \leq \mathbb{E}_{\xi_{1:m}}[V(4\tau, \xi_{1:m})] \tag{A13}$$

where the expectation on the RHS is taken over the samples $\xi_{1:m}$. That is to say, to upper bound $U(\tau)$, we first focus on bounding $V(\tau, \xi_{1:m})$ for every possible value of $\xi_{1:m}$.

Fix $\xi_{1:m}$. Consider the set:

$$H(\xi_{1:m}) := \{(\psi_{z_1,t}(\xi_1), \psi_{z_1,t}(\xi_2), \ldots, \psi_{z_1,t}(\xi_m)) : z_1 \in \mathcal{Z}_1, t \in \mathbb{R}\} \subseteq \{0,1\}^m$$

which represents all binary outcomes of applying ψ_{z_1} to the data points $\xi_{1:m}$ across all $z_1 \in \mathcal{Z}_1$. Given that the VC-dimension of the set collection

$$\mathcal{W} := \{\{\xi \in \Xi : o(z_1, \xi) \geq t\} \mid z_1 \in \mathcal{Z}_1, t \in \mathbb{R}\}$$

is upper bounded by $\mathrm{vc}(\mathcal{W})$, Sauer-Shelah Lemma then constrains the size of $H(\xi_{1:m})$ to obey

$$|H(\xi_{1:m})| \leq (2m)^{|\mathrm{vc}(\mathcal{W})|}. \tag{A14}$$

For a reference of Sauer-Shelah Lemma, see, e.g., [46, 47] or [44, Theorem 8.3.16].

Notably, this definition also allows replacing the supremum over continuous z_1 values with the supremum over the finite binary labeling set $H(\xi_{1:m})$. Specifically, there is the upper bound:

$$\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i) u_v(\xi_i) \leq \max_{b \in H(\xi_{1:m})} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i b_i u_v(\xi_i).$$

On the RHS, $b \in H(\xi_{1:m}) \subseteq \{0,1\}^m$ denotes a vector whose components are $b = (b_1, b_2, \ldots, b_m)$.

This establishes an upper bound on $V(\tau, \xi_{1:m})$ that applies for every $\tau > 0$:

$$V(\tau, \xi_{1:m}) \leq \mathbb{E}_{\varepsilon_{1:m}}\left[\exp\left(\tau \cdot \max_{b \in H(\xi_{1:m})} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i b_i u_v(\xi_i)\right)\right].$$

By bounding the exponential of a supremum with the sum of exponentials across all configurations, we can further upper bound $V(\tau, \xi_{1:m})$ and derive:

$$\begin{aligned} V(\tau, \xi_{1:m}) &\leq \mathbb{E}_{\varepsilon_{1:m}}\left[\sum_{b \in H(\xi_{1:m})} \sum_{v \in V_{1/2}} \exp\left(\frac{\tau}{m} \cdot \sum_{i=1}^{m} \varepsilon_i b_i u_v(\xi_i)\right)\right] \\ &= \sum_{b \in H(\xi_{1:m})} \sum_{v \in V_{1/2}} \prod_{i=1}^{m} \mathbb{E}_{\varepsilon_i}\left[\exp\left(\frac{\tau}{m} \cdot \varepsilon_i b_i u_v(\xi_i)\right)\right] \end{aligned} \tag{A15}$$

where the symbol $\mathbb{E}_{\varepsilon_i}$ means the expectation is taken solely over the Rademacher variables ε_i.

Given that for any $b_i \in \{0,1\}$ and $q_i \in \mathbb{R}$, the expectation over ε_i obeys:

$$\mathbb{E}_{\varepsilon_i}[\exp(\varepsilon_i b_i q_i)] = \cosh(b_i q_i) \le \cosh(q_i) = \mathbb{E}_{\varepsilon_i}[\exp(\varepsilon_i q_i)]$$

this allows us to upper bound the RHS in equation (A15), removing the dependence on $b_{1:m}$:

$$\begin{aligned} V(\tau, \xi_{1:m}) &\le \sum_{v \in V_{1/2}} \sum_{b \in H(\xi_{1:m})} \prod_{i=1}^{m} \mathbb{E}_{\varepsilon_i}\left[\exp\left(\frac{\tau}{m} \cdot \varepsilon_i u_v(\xi_i)\right)\right] \\ &= |H(\xi_{1:m})| \sum_{v \in V_{1/2}} \prod_{i=1}^{m} \mathbb{E}_{\varepsilon_i}\left[\exp\left(\frac{\tau}{m} \cdot \varepsilon_i u_v(\xi_i)\right)\right]. \end{aligned}$$

Given the bound we have established for $|H(\xi_{1:m})| \le (2m)^{|\mathrm{vc}(\mathcal{W})|}$ in equation (A14), we deduce:

$$V(\tau, \xi_{1:m}) \le (2m)^{\mathrm{vc}(\mathcal{W})} \sum_{v \in V_{1/2}} \prod_{i=1}^{m} \mathbb{E}_{\varepsilon_i}\left[\exp\left(\frac{\tau}{m} \cdot \varepsilon_i u_v(\xi_i)\right)\right]. \tag{A16}$$

Notably, this bound applies to every possible $\xi_{1:m}$ value.

Finally, we substitute this bound into inequality (A13), resulting in:

$$U(\tau) \le \mathbb{E}_{\xi_{1:m}}[V(4\tau, \xi_{1:m})] \le (2m)^{\mathrm{vc}(\mathcal{W})} \sum_{v \in V_{1/2}} \prod_{i=1}^{m} \mathbb{E}_{\varepsilon_i, \xi_i}\left[\exp\left(\frac{4\tau}{m} \cdot \varepsilon_i u_v(\xi_i)\right)\right] \tag{A17}$$

where $\mathbb{E}_{\varepsilon_i, \xi_i}$ denotes the joint expectation taken over ε_i and ξ_i.

Since $u_v(\xi)$ is $\bar{\sigma}$-sub-exponential, then $\varepsilon_i u_v(\xi_i)$ is also $\bar{\sigma}$-sub-exponential for every v by definition. Using [44, Proposition 2.7.1(e)], this implies that, for every θ_2, v, $\tau \in (0, cm/\bar{\sigma})$:

$$\mathbb{E}_{\varepsilon_i, \xi_i}\left[\exp\left(\frac{4\tau}{m} \cdot \varepsilon_i u_v(\xi_i)\right)\right] \le \exp\left(C \frac{\bar{\sigma}^2 \tau^2}{m^2}\right)$$

where $c, C > 0$ are absolute constants. Because $|V_{1/2}| \le 6^d$, we further obtain for $\tau \in (0, cm/\bar{\sigma})$:

$$U(\tau) \le (2m)^{\mathrm{vc}(\mathcal{W})} 6^d \exp\left(C \frac{\bar{\sigma}^2 \tau^2}{m}\right). \tag{A18}$$

Recall the definition of $U(\tau)$, which yields:

$$\mathbb{E}^*\left[\exp\left(\tau \cdot \sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1, t)\|\right)\right] \le (2m)^{\mathrm{vc}(\mathcal{W})} 6^d \exp\left(C \frac{\bar{\sigma}^2 \tau^2}{m}\right).$$

By applying Markov's inequality under the outer integral [31] (which follows from the monotone property of *outer integral* discussed below Definition 5), we establish that there exists a universal constant $c > 0$ such that for every $u \ge 0$:

$$\mathbb{P}^*\left(\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1, t)\| \ge u\right) \le (2m)^{\mathrm{vc}(\mathcal{W})} 6^d \cdot \exp\left(-c \min\left\{\frac{mu^2}{\bar{\sigma}^2}, \frac{mu}{\bar{\sigma}}\right\}\right).$$

Equivalently, there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$, the following happens with an inner probability at least $1-\delta$:

$$\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \|W_m(z_1,t)\| \le C\bar{\sigma} \cdot \max\{\Delta(\delta), \Delta(\delta)^2\},$$

where $\Delta(\delta)$ is defined in equation (A5). □

A.1.3 A Basic Bound

This section proves Lemma 4. By definition:

$$\Delta(\delta_k) = \sqrt{\frac{1}{m} \cdot \left(d + \mathrm{vc}(\mathcal{W}) \log m + \log(\frac{1}{\delta}) + \log(2^{k+1}) + \log|\mathcal{Z}_{2,\epsilon_k}| + \log|\mathcal{Z}_{2,\epsilon_{k-1}}|\right)}.$$

Recall that $\log|\mathcal{Z}_{2,\epsilon_k}| \le d\log(2 + r/\epsilon_k) \le d\log(2+2^k)$. This yields that

$$\Delta(\delta_k) \le \sqrt{\frac{1}{m} \cdot \left(d + \mathrm{vc}(\mathcal{W}) \log m + \log(\frac{1}{\delta})\right)} + \sqrt{\frac{12kd}{m}}$$

As a result, there exists a universal constant $c > 0$ such that

$$\begin{aligned}\sum_{k=1}^{\infty} \epsilon_k \Delta(\delta_k) &\le \left(\sum_{k=1}^{\infty} \epsilon_k\right) \cdot \sqrt{\frac{1}{m} \cdot \left(d + \mathrm{vc}(\mathcal{W}) \log m + \log(\frac{1}{\delta})\right)} + \sum_{k=1}^{\infty} \epsilon_k \sqrt{\frac{12kd}{m}} \\ &\le cr \cdot \sqrt{\frac{1}{m} \cdot \left(d + \mathrm{vc}(\mathcal{W}) \log m + \log(\frac{1}{\delta})\right)} = cr\Delta(\delta).\end{aligned}$$

Similarly, one can prove the existence of a universal constant $c > 0$ such that

$$\sum_{k=1}^{\infty} \epsilon_k \Delta(\delta_k)^2 \le cr\Delta(\delta)^2.$$

This then concludes the existence of a universal constant $c > 0$ such that

$$\sum_{k=1}^{\infty} \epsilon_k \max\{\Delta(\delta_k), \Delta(\delta_k)^2\} \le \sum_{k=1}^{\infty} \epsilon_k \Delta(\delta_k) + \sum_{k=1}^{\infty} \epsilon_k \Delta(\delta_k)^2 \le 2cr \max\{\Delta(\delta), \Delta(\delta)^2\}.$$

A.1.4 A Result on Measurability

We prove Lemma 6. Let $\mathcal{Z}_1^o$ and $\mathbb{R}^o$ be any countable dense subsets of $\mathcal{Z}_1$ and $\mathbb{R}$, respectively. Our main observation is that the supremum over the uncountable sets $\mathcal{Z}_1, \mathbb{R}$ is equal to the supremum over the countable sets $\mathcal{Z}_1^o, \mathbb{R}^o$, holding for every data instances $\xi_{1:m}, \varepsilon_{1:m}$:

$$\sup_{z_1 \in \mathcal{Z}_1, t \in \mathbb{R}} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i) u_v(\xi_i) = \sup_{z_1 \in \mathcal{Z}_1^o, t \in \mathbb{R}^o} \max_{v \in V_{1/2}} \frac{1}{m} \sum_{i=1}^{m} \varepsilon_i \psi_{z_1,t}(\xi_i) u_v(\xi_i) \tag{A19}$$

Since the RHS is the pointwise supremum over *countable* measurable functions of $\xi_{1:m}, \varepsilon_{1:m}$, it is measurable. Therefore, the LHS must also be measurable under the product space. To complete the proof of Lemma 6, it suffices to prove equation (A19).

Fix the values of $\xi_{1:m}$ and $\varepsilon_{1:m}$. It suffices to show that for every $z_1 \in \mathcal{Z}_1$ and $t \in \mathbb{R}$, there exist $z_1^o \in \mathcal{Z}_1^o$ and $t^o \in \mathbb{R}^o$ such that for every $i = 1, 2, \ldots, m$:

$$\psi_{z_1,t}(\xi_i) = \psi_{z_1^o,t^o}(\xi_i). \tag{A20}$$

Recall that $\psi_{z_1,t}(\xi) = \mathbf{1}\{o(z_1, \xi) \geq t\}$. First, we note an important observation, which uses the fact that $t \mapsto \mathbf{1}\{w \geq t\}$ is left continuous for every real value w. For every z_1, there exists $t^o \in \mathbb{R}^o$ with $t^o < t$ such that for every $i = 1, 2, \ldots, m$:

$$\psi_{z_1,t}(\xi_i) = \psi_{z_1,t^o}(\xi_i) \quad \text{and} \quad o(z_1, \xi_i) - t^o \neq 0.$$

Fix this t^o. Next, since $z_1 \mapsto o(z_1, \xi)$ is continuous for every ξ, this implies the existence of $z_1^o \in \mathcal{Z}_1^o$ close to z_1 such that for every $i = 1, 2, \ldots, m$:

$$\psi_{z_1,t^o}(\xi_i) = \psi_{z_1^o,t^o}(\xi_i).$$

This proves that for every $z_1 \in \mathcal{Z}_1$ and $t \in \mathbb{R}$, there is the existence of the pair $z_1^o \in \mathcal{Z}_1^o$ and $t^o \in \mathbb{R}^o$ such that equation (A20) holds for all $i = 1, 2, \ldots, m$. This completes the proof of Lemma 6.

Appendix B Proof of Theorem 7

Our approach comprises two main steps. First, we demonstrate that every stationary point of the empirical objective Φ_S must be approximately stationary for the population objective Φ, leveraging the uniform convergence of subdifferentials in Theorem 6. Second, we utilize established results from the literature [14, Section 5] on these (approximate) stationary points of Φ to pinpoint the location of empirical stationary points.

To set up the stage, we recall the location of the stationary point and approximate stationary point of the population objective Φ formally derived in [14].

Theorem 11 ([14, Theorem 5.2]). *The stationary points of the population objective Φ are*

$$\{0\} \cup \{\pm\bar{x}\} \cup \{x : \langle x, \bar{x}\rangle = 0, \|x\| = c \cdot \|\bar{x}\|\}$$

where $c > 0$ is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

The following Theorem 12 can be easily deduced from [14, Theorem 5.2+Theorem 5.3]. Recall that $\mathcal{Z} = \{x : 0 \in \partial\Phi(x)\}$ denotes the set of stationary points of Φ.

Theorem 12 ([14, Theorem 5.3]). *There exist numerical constants $\gamma, C > 0$ such that the following holds. For any point $x \in \mathbb{R}^d$ with*

$$\operatorname{dist}(0; \partial\Phi(x)) \leq \gamma \|x\|,$$

it must be the case that $\|x\| \leq C\|\bar{x}\|$ and x satisfies $\operatorname{dist}(x, \mathcal{Z}) \leq C\operatorname{dist}(0, \partial\Phi(x))$.

The following result is immediate given [14, Corollary 6.3].

Proposition 13 ([14, Corollary 6.3]). *There exist a numerical constant $C > 0$ so that for $m \geq Cd$, with an inner probability at least $1 - 2\exp(-d)$[1], the set of stationary points $\mathcal{Z}_S$ of Φ_S satisfies:*

$$\mathcal{D}(\mathcal{Z}_S, \mathcal{Z}) \leq C\sqrt[4]{\frac{m}{d}}\|\bar{x}\|.$$

[1] The original expression in the paper [14], $1 - 2\exp(-\min\{c_2 m, d^2\})$, should be corrected to $1 - 2\exp(-d)$.

We shall build on top of these existing results to establish Theorem 7. First, Proposition 13 implies there exists a numerical constant $C_1 > 0$ such that:

$$\mathbb{P}_* \left(\|x\| \le C_1 \|\bar{x}\| \quad \forall x \in \mathcal{Z}_S \right) \ge 1 - 2e^{-d}. \tag{B1}$$

Corollary 2 refines Theorem 6 using a standard peeling argument to substitute the norm $\|x\|$ for the constant radius r in the original probability expression in equation (5.2). This adjustment allows varying distances x from the origin, ranging from $\underline{r}$ to r, enhancing the applicability of the probability bounds over different scales. For completeness, we give its proof in Section B.1.

Corollary 2. *Assume the measurement vector a is σ-subgaussian, and $\mathbb{E}[|b|^2] < \infty$.*
Then there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$ and $r \ge \underline{r} > 0$:

$$\mathbb{P}_* \left(\mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C\sigma \|x\| \cdot \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\} \quad \forall x \textit{ satisfying } \underline{r} \le \|x\| \le r \right) \ge 1 - \delta. \tag{B2}$$

In the above,

$$\tilde{\Delta}_\Phi = \sqrt{\frac{1}{m} \cdot \left(d \log d \log m + \log\left(\frac{\log(er/\underline{r})}{\delta}\right) \right)}. \tag{B3}$$

We are now ready to prove Theorem 7. We pick

$$\underline{r} = C_1 \sqrt{\frac{d}{m}} \|\bar{x}\|, \quad r = C_1 \|\bar{x}\|, \quad \delta = e^{-d}, \quad \sigma = 4$$

in Corollary 2. We then derive for some absolute constant C_2:

$$\begin{aligned} \mathbb{P}_* \Big(& \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C_2 \|x\| \cdot \max\{\Delta_0, \Delta_0^2\}, \\ & \forall x \text{ satisfying } C_1 \sqrt{d/m} \|\bar{x}\| \le \|x\| \le C_1 \|\bar{x}\| \Big) \ge 1 - e^{-d}. \end{aligned} \tag{B4}$$

where

$$\Delta_0 = \sqrt{\frac{d}{m} \log d \log m}.$$

We without loss of generality assume that $C_2 \max\{\Delta_0, \Delta_0^2\} \le \gamma$ where γ is the numerical constant specified in Theorem 12, since otherwise Theorem 7 trivially follows from equation (B1).

With equations (B1) and (B4), we know that with an inner probability at least $1 - 3e^{-d}$, the following simultaneously happen:

1. $\|x\| \le C_1 \|\bar{x}\|$ for every $x \in \mathcal{Z}_S$.
2. $\mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C_2 \|x\| \cdot \max\{\Delta_0, \Delta_0^2\}$ for every x with $C_1 \sqrt{d/m} \|\bar{x}\| \le \|x\| \le C_1 \|\bar{x}\|$.

Consequentially, on this event, any stationary point x^* of Φ_S must either obey:

$$\|x^*\| \le C_1 \sqrt{d/m} \|\bar{x}\|$$

or satisfies

$$\mathbb{H}(\partial\Phi(x^*), \partial\Phi_S(x^*)) \le C_2 \|x^*\| \cdot \max\{\Delta_0, \Delta_0^2\}.$$

In the first case, since $0 \in \mathcal{Z}$ by Theorem 11, we immediately obtain:

$$\operatorname{dist}(x^*, \mathcal{Z}) \le \|x^*\| \le C_1 \sqrt{d/m} \|\bar{x}\|.$$

In the second case, since $0 \in \partial\Phi_S(x^*)$, then $\mathrm{dist}(0, \partial\Phi(x^*)) \le \mathbb{H}(\partial\Phi(x^*), \partial\Phi_S(x^*))$ by definition of Hausdorff distance, and thus:

$$\mathrm{dist}(0, \partial\Phi(x^*)) \le C_2 \|x^*\| \cdot \max\{\Delta_0, \Delta_0^2\}.$$

Since we have assumed $C_2 \max\{\Delta_0, \Delta_0^2\} \le \gamma$, by Theorem 12 we obtain that x^* must obey

$$\mathrm{dist}(x^*, \mathcal{Z}) \le C_2 \|x^*\| \cdot \max\{\Delta_0, \Delta_0^2\} \le C_1 C_2 \|\bar{x}\| \cdot \max\{\Delta_0, \Delta_0^2\}.$$

where we recall that $\|x\| \le C_1 \|\bar{x}\|$ for every $x \in \mathcal{Z}_S$ on this event.

Summarizing, we have established with an inner probability at least $1 - 3e^{-d}$, any stationary point x^* of Φ_S must obey:

$$\mathrm{dist}(x^*, \mathcal{Z}) \le C \|\bar{x}\| \cdot \max\{\Delta_0, \Delta_0^2\}$$

where $C > 0$ is an absolute constant. This completes the proof of Theorem 7.

B.1 Proof of Corollary 2

We can define a sequence $\{r^{(i)}\}_{i=1}^N$ so that $r^{(0)} = \underline{r}$, $r^{(N)} = r$, $r^{(i+1)}/r^{(i)} \le e$ and $N \le \log(er/\underline{r})$. Under the condition of Corollary 2, we can apply Theorem 6 and obtain for every $1 \le i \le N$

$$\mathbb{P}^* \left(\sup_{x:\|x\| \le r^{(i)}} \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \ge C\sigma r^{(i)} \cdot \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\} \right) \le \delta/N$$

In the above, C is a universal constant. A union bound shows that:

$$\mathbb{P}_* \left(\sup_{x:\|x\| \le r^{(i)}} \mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C\sigma r^{(i)} \cdot \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\} \text{ for every } 1 \le i \le N \right) \ge 1 - \delta.$$

On this event, for every x with $\underline{r} \le \|x\| \le r$, we find $1 \le i(x) \le N$ such that $r^{(i(x)-1)} \le \|x\| \le r^{(i(x))}$, and then we obtain that

$$\mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C\sigma r^{(i(x))} \cdot \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\} \le Ce\sigma \|x\| \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\}$$

where the last inequality is due to $r^{(i(x))}/\|x\| \le r^{(i(x))}/r^{(i(x)-1)} \le e$ by construction. Hence, we have proven the existence of a numerical constant $C' > 0$ such that:

$$\mathbb{P}_* \left(\mathbb{H}(\partial\Phi(x), \partial\Phi_S(x)) \le C'\sigma \|x\| \max\{\tilde{\Delta}_\Phi, \tilde{\Delta}_\Phi^2\} \text{ for all } x \text{ such that } \underline{r} \le \|x\| \le r \right) \ge 1 - \delta.$$

This completes the proof of Corollary 2.

Appendix C Proof of Corollary 1

We apply Theorem 5. We first check the assumptions. For the nonsmooth $h(z) = |z|$, it is decomposed as $h = h^{ns} + h^{sm}$ where $h^{sm}(z) = -z$ and $h^{ns}(z) = 2(z)_+$.

- (Assumption C.1). For the matrix sensing problem,

$$\nabla c(X, \xi) = (A + A^T)X, \quad (h^{sm})'(c(X;\xi))\nabla c(X;\xi) = -(A + A^T)X.$$

  Thus, at $X = 0 \in \mathbb{R}^{d\times r}$, both gradients vanish: $\nabla c(0, \xi) = (h^{sm})'(c(0;\xi))\nabla c(0;\xi) = 0$.

- (Assumption C.2). It is easy to verify that

$$\nabla c(X_1,\xi) - \nabla c(X_2,\xi) = (A + A^T)(X_1 - X_2)$$
$$e(X_1,\xi) - e(X_2,\xi) = (A + A^T)(X_2 - X_1),$$

where $e(X,\xi) = (h^{sm})'(c(X;\xi))\nabla c(X;\xi)$. For every matrix $V \in \mathbb{R}^{D\times D}$, $\langle A, V\rangle$ is $\|V\|_F$-subgaussian. Given the property that the sum of two subgaussian random variables is still subgaussian [44, Exercise 2.5.7], we deduce that $\langle A + A^T, V\rangle$ is $C\|V\|_F$-subgaussian for every matrix V, where $C > 0$ is a universal constant. As a consequence, we derive that, each of

$$\langle V, \nabla c(X_1,\xi) - \nabla c(X_2,\xi)\rangle = \langle (A + A^T), (X_1 - X_2)^T V\rangle$$
$$\langle V, e(X_1,\xi) - e(X_2,\xi)\rangle = \langle (A + A^T), (X_2 - X_1)^T V\rangle,$$

must be $C\|X_1 - X_2\|_F$ subgaussian, and thus $C'\|X_1 - X_2\|_F$ subexponential for every matrix $V \in \mathbb{R}^{D\times D}$ with $\|V\|_F = 1$ for some universal constant $C' > 0$. Here we use the fact that every sub-gaussian distribution is subexponential [44, Section 2.7].
- (Assumption C.3). It is easy to verify that

$$\|\nabla c(X_1,\xi) - \nabla c(X_2,\xi)\|_F = \left\|(A + A^T)(X_1 - X_2)\right\|_F \le 2\|A\|_F\|X_1 - X_2\|_F$$
$$\|e(X_1,\xi) - e(X_2,\xi)\|_F = \left\|(A + A^T)(X_2 - X_1)\right\|_F \le 2\|A\|_F\|X_1 - X_2\|_F,$$

where $e(X,\xi) = (\psi^{sm})'(c(X;\xi))\nabla c(X;\xi)$. Since A is σ-subgaussian, $\mathbb{E}[\|A\|_F] < \infty$. As a result, Assumption C.3 is satisfied with $L(\xi) = 2\|A\|_F$.
- (Integrability). Notably $\mathbb{E}[|c(X;\xi)|^2] = \mathbb{E}[|b - \langle a, X\rangle|^2] < \infty$ for every $x \in \mathbb{R}^d$.

We then compute the bound in Theorem 5. Namely, we need to compute ζ and $\mathrm{vc}(\mathcal{F})$.

For the nonsmooth function $h(z) = |z|$, the corresponding value of ζ is given by $\zeta = 3$ as 0 is the only nondifferentiable point of h. The corresponding $\mathcal{F}$ is given by

$$\mathcal{F} = \left\{\{A \in \mathbb{R}^{D\times D}, b \in \mathbb{R} : \langle A, XX^T\rangle - b \ge t\} \mid X \in \mathbb{R}^{D\times r_0}, t \in \mathbb{R}\right\}.$$

Note that $X \mapsto \langle A, XX^T\rangle - b$ is a degree 2 polynomial in X for every A, b. We then bound its VC dimension using Theorem 8, which yields that for some universal constant $C > 0$:

$$\mathrm{vc}(\mathcal{F}) \le CDr_0\log(Dr_0).$$

Corollary 1 then follows by applying Theorem 8.

References

[1] Shapiro, A., Xu, H.: 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)

[2] Shapiro, A., Dentcheva, D., Ruszczynski, A.: Lectures on Stochastic Programming: Modeling and Theory. SIAM, Philadelphia (2021)

[3] Poliquin, R., Rockafellar, R.T.: Amenable functions in optimization. Nonsmooth optimization: methods and applications (Erice, 1991), 338–353 (1992)

[4] Poliquin, R., Rockafellar, R.T.: Prox-regular functions in variational analysis. Transactions of the American Mathematical Society **348**(5), 1805–1838 (1996)

[5] Davis, D., Drusvyatskiy, D.: Subgradient methods under weak convexity and tame geometry. SIAG/OPT Views and News **28**, 1–10 (2020)

[6] Burke, J.V.: Descent methods for composite nondifferentiable optimization problems. Mathematical Programming **33**, 260–279 (1985)

[7] Duchi, J.C., Ruan, F.: Stochastic methods for composite and weakly convex optimization problems. SIAM Journal on Optimization **28**(4), 3229–3259 (2018)

[8] Duchi, J.C., Ruan, F.: 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)

[9] Charisopoulos, V., Davis, D., Díaz, M., Drusvyatskiy, D.: Composite optimization for robust rank one bilinear sensing. Information and Inference: A Journal of the IMA **10**(2), 333–396 (2021)

[10] Li, X., Zhu, Z., Man-Cho So, A., Vidal, R.: Nonconvex robust low-rank matrix recovery. SIAM Journal on Optimization **30**(1), 660–686 (2020)

[11] Charisopoulos, V., Chen, Y., Davis, D., Díaz, M., Ding, L., Drusvyatskiy, D.: Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. Foundations of Computational Mathematics **21**(6), 1505–1593 (2021)

[12] Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. Journal of risk **2**, 21–42 (2000)

[13] Koenker, R., Park, B.J.: An interior point algorithm for nonlinear quantile regression. Journal of Econometrics **71**(1-2), 265–283 (1996)

[14] Davis, D., Drusvyatskiy, D., Paquette, C.: The nonsmooth landscape of phase retrieval. IMA Journal of Numerical Analysis **40**(4), 2652–2695 (2020)

[15] Attouch, H.: Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés. CR Acad. Sci. Paris **284**(539-542), 13 (1977)

[16] Davis, D., Drusvyatskiy, D.: Graphical convergence of subgradients in nonconvex optimization and learning. Mathematics of Operations Research **47**(1), 209–231 (2022)

[17] Rockafellar, R.T., Wets, R.J.: Variational Analysis. Springer, New York (1998)

[18] Molchanov, I.: Theory of Random Sets vol. 19. Springer, New York (2005)

[19] Artstein, Z., Vitale, R.A.: A strong law of large numbers for random compact sets. The Annals of Probability, 879–882 (1975)

[20] Xu, H.: Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. Journal of Mathematical Analysis and Applications **368**(2), 692–710 (2010)

[21] Ralph, D., Xu, H.: Convergence of stationary points of sample average two-stage stochastic programs: A generalized equation approach. Mathematics of Operations Research **36**(3), 568–592 (2011)

[22] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (2013)

[23] Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société mathématique de France **93**, 273–299 (1965)

[24] Xu, H., Zhang, D.: Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. Mathematical programming **119**, 371–401 (2009)

[25] Horowitz, J.L.: Bootstrap methods for median regression models. Econometrica, 1327–1351 (1998)

[26] Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. Transactions of the American Mathematical Society **362**(6), 3319–3363 (2010)

[27] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)

[28] Aubin, J.-P., Frankowska, H.: Set-valued Analysis. Springer, New York (2009)

[29] Aumann, R.J.: Integrals of set-valued functions. Journal of mathematical analysis and applications **12**(1), 1–12 (1965)

[30] Billingsley, P.: Probability and Measure, Second edn. Wiley, New York (1986)

[31] Vaart, A.W., Wellner, J.A.: Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York (1996)

[32] Attouch, H., Beer, G.: On the convergence of subdifferentials of convex functions. Archiv der Mathematik **60**(4), 389–400 (1993)

[33] Sion, M.: On general minimax theorems. Pacific J. Math. **8**(4), 171–176 (1958)

[34] Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms I: Fundamentals vol. 305. Springer, Cambridge (1996)

[35] Armstrong, M.A.: Basic Topology. Springer, New York (2013)

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

[37] Eldar, Y.C., Mendelson, S.: Phase retrieval: Stability and recovery guarantees. Applied and Computational Harmonic Analysis **36**(3), 473–494 (2014)

[38] Candes, E.J., Li, X., Soltanolkotabi, M.: Phase retrieval via wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory **61**(4), 1985–2007 (2015)

[39] Wang, G., Giannakis, G.B., Eldar, Y.C.: Solving systems of random quadratic equations via truncated amplitude flow. IEEE Transactions on Information Theory **64**(2), 773–794 (2017)

[40] Sun, J., Qu, Q., Wright, J.: A geometric analysis of phase retrieval. Foundations of Computational Mathematics **18**, 1131–1198 (2018)

[41] Ding, L., Jiang, L., Chen, Y., Qu, Q., Zhu, Z.: Rank overspecified robust matrix recovery: Subgradient method and exact recovery. arXiv preprint arXiv:2109.11154 (2021)

[42] Borwein, J.M., Vanderwerff, J.D.: Convex Functions: Constructions, Characterizations and Counterexamples vol. 109. Cambridge University Press, Cambridge (2010)

[43] Vapnik, V.N., Chervonenkis, A.Y.: On the uniform convergence of relative frequencies of events to their probabilities. vol. 16, pp. 264–280 (1971)

[44] Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science vol. 47. Cambridge university press, Cambridge (2018)

[45] Dudley, R.M.: The sizes of compact subsets of hilbert space and continuity of gaussian processes. Journal of Functional Analysis **1**(3), 290–330 (1967)

[46] Sauer, N.: On the density of families of sets. Journal of Combinatorial Theory, Series A **13**(1), 145–147 (1972)

[47] Shelah, S.: A combinatorial problem; stability and order for models and theories in infinitary languages. Pacific Journal of Mathematics **41**(1), 247–261 (1972)

[48] Wainwright, M.J.: High-dimensional Statistics: A Non-asymptotic Viewpoint vol. 48. Cambridge university press, Cambridge (2019)

[49] Vaart, A.W.: Asymptotic Statistics vol. 3. Cambridge university press, Cambridge (2000)

[50] Milnor, J.: On the betti numbers of real varieties. Proceedings of the American Mathematical Society **15**(2), 275–280 (1964)

[51] Thom, R.: On the homology of real algebraic varieties. Differential and Combinatorial Topology (1965)

[52] Matousek, J.: Lectures on Discrete Geometry vol. 212. Springer, New York (2013)

[53] Warren, H.E.: Lower bounds for approximation by nonlinear manifolds. Transactions of the American Mathematical Society **133**(1), 167–178 (1968)

[54] Rockafellar, R.T.: Favorable classes of lipschitz continuous functions in subgradient optimization (1981)

[55] Han, X., Lewis, A.S.: Survey descent: A multipoint generalization of gradient descent for nonsmooth optimization. SIAM Journal on Optimization **33**(1), 36–62 (2023)

[56] Facchinei, F., Pang, J.-S.: Finite-dimensional Variational Inequalities and Complementarity Problems. Springer, New York (2003)

[57] Koenker, R.: Quantile Regression vol. 38. Cambridge university press, Cambridge (2005)