Optimal Rates for Robust Stochastic Convex Optimization

Changyu Gao ✉ 🏠
University of Wisconsin-Madison, Madison, WI, USA

Andrew Lowy ✉
University of Wisconsin-Madison, Madison, WI, USA

Xingyu Zhou ✉
Wayne State University, Detroit, MI, USA

Stephen J. Wright ✉
University of Wisconsin-Madison, Madison, WI, USA

Abstract

Machine learning algorithms in high-dimensional settings are highly susceptible to the influence of even a small fraction of structured outliers, making robust optimization techniques essential. In particular, within the ϵ-contamination model, where an adversary can inspect and replace up to an ϵ-fraction of the samples, a fundamental open problem is determining the optimal rates for robust stochastic convex optimization (SCO) under such contamination. We develop novel algorithms that achieve *minimax-optimal* excess risk (up to logarithmic factors) under the ϵ-contamination model. Our approach improves over existing algorithms, which are not only suboptimal but also require stringent assumptions, including Lipschitz continuity and smoothness of individual sample functions. By contrast, our optimal algorithms do not require these stringent assumptions, assuming only population-level smoothness of the loss. Moreover, our algorithms can be adapted to handle the case in which the covariance parameter is unknown, and can be extended to nonsmooth population risks via convolutional smoothing. We complement our algorithmic developments with a *tight information-theoretic lower bound* for robust SCO.

2012 ACM Subject Classification Theory of computation → Mathematical optimization

Keywords and phrases Adversarial Robustness, Machine Learning, Optimization Algorithms, Robust Optimization, Stochastic Convex Optimization

Digital Object Identifier 10.4230/LIPIcs.FORC.2025.9

Related Version *Full Version*: `https://arxiv.org/abs/2412.11003`

Funding Research of CG, AL, and SW was supported by NSF Awards 2023239 and 2224213 and AFOSR Award FA9550-21-1-0084. XZ is supported in part by NSF CNS-2153220 and CNS-2312835.

Acknowledgements Changyu would like to thank Shuyao Li for helpful discussions.

1 Introduction

Machine learning models are increasingly deployed in security-critical applications, yet they remain vulnerable to data manipulation. A particular threat is data poisoning, where adversaries deliberately insert malicious points into training data to degrade model performance [1]. Even in non-adversarial settings, naturally occurring outliers can significantly impact learning algorithms, especially in high-dimensional settings. These challenges motivate our study of optimization algorithms for training machine learning models in the presence of outliers, both natural and adversarial.

Motivation for our work traces to Tukey's pioneering research on robust estimation [17]. Recent breakthroughs have produced efficient algorithms for high-dimensional robust estimation under the ϵ-contamination model, where an adversary can arbitrarily replace up to an

6th Symposium on Foundations of Responsible Computing (FORC 2025).
Editor: Mark Bun; Article No. 9; pp. 9:1–9:21
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

ϵ-fraction of the samples. Notable advances include polynomial-time algorithms for robust mean estimation in high dimensions [5, 3]. See [6] for a comprehensive survey of recent developments in high-dimensional robust estimation.

These developments in robust estimation naturally lead to a fundamental question: *Can we solve* stochastic optimization *problems, under the ϵ-contamination model?* Stochastic optimization is used in machine learning to find the parameter that minimizes the population risk using training samples. We focus specifically on robust stochastic optimization with convex objective functions whose gradients exhibit bounded covariance, a standard assumption in robust mean estimation [7]. While our goal aligns with the classical use of stochastic convex optimization in minimizing population risk, the presence of adversarial contamination introduces significant new challenges.

Prior research in robust optimization has concentrated primarily on narrow domains. One line of work focuses on robust linear regression [13, 8, 2]. While [12, 16] have explored general problems, they focus on robust *regression*. To our best knowledge, SEVER [4] is the only work that considers general stochastic optimization problems. However, this approach has several limitations that restrict the applicability of SEVER. First, it focuses only on achieving dimension-independent error due to corruption, with only a suboptimal sample complexity. Second, the results for SEVER depend on several stringent assumptions, including Lipschitzness and smoothness conditions on *individual sample* functions. *Because of these limitations, optimal excess risk bounds for robust stochastic convex optimization, and under what conditions they can be achieved, remain unknown.*

In this work, we develop efficient algorithms for robust stochastic convex optimization that achieve **optimal excess risk bounds** (up to logarithmic factors) under the ϵ-contamination model. Notably, Algorithm 1 assumes only the smoothness of the population risk. Moreover, we prove a matching lower bound to show the minimax-optimality of our algorithms.

Due to space limitations, we omit some details and proofs. These appear in the full version, available at `https://arxiv.org/abs/2412.11003`.

1.1 Problem Setup and Motivation

Notation. For a vector $v \in \mathbb{R}^d$, $\|v\|$ denotes the ℓ_2 norm of v. For a matrix $A \in \mathbb{R}^{d\times d}$, $\|A\|$ denotes the spectral norm of A. For symmetric matrices A and B, we write $A \preceq B$ if $B - A$ is positive semidefinite (PSD). We use $\tilde{O}$ and $\tilde{\Omega}$ to hide logarithmic factors in our bounds.

Let $\mathcal{W} \subset \mathbb{R}^d$ be a closed convex set. Consider a distribution p^* over functions $f : \mathcal{W} \to \mathbb{R}$. Stochastic optimization aims to find a parameter vector $w^* \in \mathcal{W}$ minimizing the population risk $\overline{f}(w) := \mathbb{E}_{f\sim p^*}[f(w)]$. For example, function f can take the form of a loss function $f_x(w)$ dependent on the data point x, and the data distribution on x induces the function distribution p^*. In robust stochastic optimization, some data samples may be corrupted. Following [4], we adopt the strong ϵ-contamination model, which allows the adversary to replace up to ϵ fraction of samples.

▶ **Definition 1** (ϵ-contamination model). *Given $\epsilon > 0$ and a distribution p^* over functions $f : \mathcal{W} \to \mathbb{R}$, data is generated as follows: first, n clean samples $f_1, \dots, f_n$ are drawn from p^*. An* adversary *is then permitted to examine the samples and replace up to ϵn of them with arbitrary samples. The algorithm is subsequently provided with this modified set of functions, which we refer to as* ϵ-corrupted *samples (with respect to p^*).*

This model is strictly stronger than the Huber contamination model [11], in which the samples are drawn from a mixture of the clean and adversarial distributions of the form $p^* = (1 - \epsilon)p + \epsilon q$, where p is the clean distribution and q is the adversarial distribution.

Our objective is to develop an efficient algorithm that minimizes the population risk $\overline{f}(w)$, even when the data is ϵ-corrupted. The following is assumed throughout the paper.

▶ **Assumption 2.**

1. *$\mathcal{W} \subset \mathbb{R}^d$ is a compact convex set with diameter D, that is, $\sup_{w,w' \in \mathcal{W}} \|w - w'\| \leq D$.*
2. *f is differentiable almost surely. The population risk $\overline{f}(w)$ is convex.*
3. *The regularity condition holds $\mathbb{E}_{f \sim p^*}[\nabla f(w)] = \nabla \overline{f}(w)$.*[1]

We also assume in most results that the gradients of the functions have bounded covariance as in [4], which is a typical assumption used in robust mean estimation.

▶ **Assumption 3.** *There is $\sigma > 0$ such that for all $w \in \mathcal{W}$ and all unit vectors v, we have $\mathbf{E}_{f \sim p^*}[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2] \leq \sigma^2$.*

An equivalent form of this assumption is that for every $w \in \mathcal{W}$, the covariance matrix of the gradients, defined by $\Sigma_w := \mathbb{E}_{f \sim p^*}[(\nabla f(w) - \nabla \overline{f}(w))(\nabla f(w) - \nabla \overline{f}(w))^T]$ satisfies $\Sigma_w \preceq \sigma^2 I$. (See Section E for a proof.)

We will additionally assume that the population risk $\overline{f}(w)$ satisfies certain properties, or that certain properties are satisfied almost surely for functions f from distribution p^*, as needed.

To our best knowledge, SEVER [4] is the only work that studies robust stochastic optimization for general convex losses. While SEVER focuses on finding approximate critical points, our work focuses on minimizing the population risk $\overline{f}(w)$, and we measure the performance of our algorithm in terms of the excess risk $\overline{f}(\hat{w}) - \min_w \overline{f}(w)$, where $\hat{w}$ is the output of the algorithm.

We remark that SEVER also derives excess risk bounds. To contrast with SEVER, we decompose the excess risk of a stochastic optimization algorithm as follows[2]:

$$\text{Excess risk} = \text{Error due to corruption} + \text{Statistical error},$$

where "error due to corruption" refers to the error due to the presence of corruption in the data, while "statistical error" denotes the error that accrues even when there is no corruption. SEVER [4] focuses only on the error due to corruption. The statistical error term is implicit in their requirement on the sample complexity n, that is,

$$\text{Excess risk} = \text{Error due to corruption, if } n \geq \text{[sample complexity]}.$$

Specifically, they design a polynomial-time algorithm that achieves $O(D\sigma\sqrt{\epsilon})$ error due to corruption term for $n = \tilde{\Omega}\left(\frac{dL^2}{\epsilon\sigma^2} + \frac{dL^4}{\sigma^4}\right)$, provided that $f - \overline{f}$ is L-Lipschitz and β-smooth almost surely for $f \in p^*$, and that f is smooth almost surely. (Their analysis has an incorrect sample complexity result, which we fix in the appendix of the full version of this paper.) This sample complexity can be huge (even infinite), as some functions in the distribution may have a very large (possibly unbounded) Lipschitz constant. Moreover, SEVER implicitly requires f to be smooth almost surely.

Consider functions of the form $f_x(w) = -\frac{1}{2}x \cdot \|w\|^2$ for w such that $\|w\| \leq D$, where $x \sim P$ for a probability distribution P with bounded mean and variance but with unbounded values, e.g. the normal distribution. We have $\nabla f_x(w) = -x \cdot w$. Since x is unbounded,

[1] This technical assumption allows us to exchange the expectation and the gradient. See discussions in Section E.2

[2] We omit the term due to optimization error that depends on the number of iterations of the algorithm, since it will be dominated by the other terms when we run the optimization algorithm for a sufficient number of iterations.

the worst-case Lipschitz parameter and smoothness of f are both infinite. However, the population risk $\overline{f}(w) = -\frac{1}{2}\|w\|^2 \cdot \mathbf{E}[x]$ is smooth and Lipschitz. This example demonstrates that the assumptions in SEVER that assume properties uniformly for individual functions $f \sim p^*$ can be too stringent. In this paper, we aim to answer the following question:

> Can we design computationally efficient algorithms that achieve the *optimal excess risk* for robust SCO, under much milder conditions?

We give positive answers to this question and summarize our contributions below.

1.2 Our Contributions

1. Optimal Rates for Robust SCO (Section 3): We develop algorithms that achieve the following minimax-optimal (up to logarithmic factors) excess risk:

$$\overline{f}(\hat{w}_T) - \min_{w \in \mathcal{W}} \overline{f}(w) = \tilde{O}\left(D\left(\sigma\sqrt{\epsilon} + \sigma\sqrt{\frac{d\log(1/\tau)}{n}}\right)\right).$$

   Compared with SEVER, we achieve the same error due to corruption $O(D\sigma\sqrt{\epsilon})$ provided $n = \tilde{\Omega}(d/\epsilon)$, a significant improvement in sample complexity.[3]
2. Much Weaker Assumptions for Robust SCO: Algorithm 1 achieves the optimal rates while only assuming the smoothness of the population risk, which is significantly weaker than the assumptions used in SEVER. By contrast, SEVER requires $f - \overline{f}$ to have bounded worst-case Lipschitz and smoothness parameter, and that individual functions f are smooth almost surely.
3. Handling unknown σ and extensions to nonsmooth case: Simple adaptations allow our algorithm to handle the case in which the covariance parameter σ is unknown. We also extend our algorithm to nonsmooth population risks using convolutional smoothing. The resulting algorithm achieves the minimax-optimal excess risk.
4. A Matching Lower Bound for Robust SCO: We show a matching lower bound, demonstrating that our excess risk bound is *minimax-optimal* (up to logarithmic factors). Consequently, our sample complexity for achieving the error due to corruption $O(D\sigma\sqrt{\epsilon})$ is also minimax-optimal.
5. A Straightforward Algorithm for Robust SCO (Section 4): Algorithm 3 is an elementary algorithm that achieves the same optimal excess risk, with more stringent assumptions compared to Algorithm 1. Our approach builds on the "many-good-sets" assumption, which SEVER briefly introduced without providing a concrete analysis.

Our results might be surprising, as net-based approaches (e.g., uniform convergence) typically suffers from suboptimal error. Our results, however, imply that the net-based approach can indeed achieve the optimal excess risk under the ϵ-contamination model. We discuss this further in Section E.3. A high-level summary of our results appears in Table 1.

2 Revisiting SEVER

In this section, we revisit SEVER [4] to motivate our work. Below we fix the corruption parameter ϵ and the covariance boundedness parameter $\sigma > 0$. Given ϵ-corrupted function samples $f_1, \ldots, f_n$, we say a subset of functions is "good" with respect to w if their sample mean and covariance at w are close to those of the true distribution, as defined below.

[3] We remark that in excess risk bounds, $\tilde{O}$ always hides logarithmic factors only in the statistical error term, and the robust term is always $O(D\sigma\sqrt{\epsilon})$.

Table 1 Comparison of assumptions, rates, and sample complexity of SEVER and our two algorithms. The parameters β, L, etc are all assumed to be finite. All algorithms assume Assumption 2, and bounded covariance of the gradients, that is, the covariance matrix Σ_w satisfies $\Sigma_w \preceq \sigma^2 I$ for all w. Optimality is up to logarithmic factors. For the case when $\bar{f}$ is nonsmooth but Lipschitz (see Section 3.2), the excess risk is optimal (up to logarithmic factors) under the noncentral moment assumption.

Algorithm	Assumptions	Excess Risk	Sample Complexity
SEVER [4]	1. $f - \bar{f}$ is L-Lipschitz a.s. 2. f is β-smooth a.s.	suboptimal	$\tilde{\Omega}\left(\frac{dL^2}{\epsilon\sigma^2} + \frac{dL^4}{\sigma^4}\right)$
Algorithm 1	$\bar{f}$ is $\bar{\beta}$-smooth *or* $\bar{L}$-Lipschitz.	optimal	$\tilde{\Omega}\left(d/\epsilon\right)$
Algorithm 3	1. $f - \bar{f}$ is L-Lipschitz a.s. and β-smooth a.s. 2. $\bar{f}$ is $\bar{\beta}$-smooth *or* $\bar{L}$-Lipschitz.	optimal	$\tilde{\Omega}\left(d/\epsilon\right)$

▶ **Definition 4** ("Good" set). *We say a set $S_{good} \subseteq [n]$ with $|S_{good}| \geq (1-\epsilon)n$ is "good" w.r.t. w if the functions $\{f_i\}_{i \in S_{good}}$ satisfy the following,*

$$\left\|\frac{1}{|S_{good}|}\sum_{i \in S_{good}}\left(\nabla f_i(w) - \nabla \bar{f}(w)\right)\left(\nabla f_i(w) - \nabla \bar{f}(w)\right)^T\right\| \leq O(\sigma^2),$$
$$\left\|\frac{1}{|S_{good}|}\sum_{i \in S_{good}}\left(\nabla f_i(w) - \nabla \bar{f}(w)\right)\right\| \leq O(\sigma\sqrt{\epsilon}). \tag{1}$$

A "good" set w.r.t. w allows us to robustly estimate the gradient at w. SEVER requires the existence of a set that is uniformly good for all w, which we refer to as the "uniform-good-set" assumption.

▶ **Assumption 5** ("Uniform good set", [4, Assumption B.1]). *There exists a set $S_{good} \subseteq [n]$ with $|S_{good}| \geq (1-\epsilon)n$ such that S_{good} is "good" w.r.t. w, for all $w \in \mathcal{W}$.*

SEVER operates through an iterative filtering framework built around a black-box learner. Its core algorithm consists of three main steps: (1) The black-box learner processes the current set of functions to find approximate critical points. (2) A filtering mechanism identifies and removes outlier functions. (3) The algorithm updates its working set with the remaining functions. This process repeats until convergence. Crucially, SEVER's theoretical guarantees rely on its "uniform-good-set" assumption. Without this assumption (as opposed to "many-good-sets" assumption introduced later), the set of "good" functions can change at each iteration, potentially preventing the iterative filtering process from converging.

We argue that the "uniform-good-set" assumption can be too strong. Recall that SEVER requires a sample complexity of $n = \tilde{\Omega}\left(\frac{dL^2}{\epsilon\sigma^2} + \frac{dL^4}{\sigma^4}\right)$. When $n = \tilde{\Omega}(d/\epsilon)$, the "uniform-good-set" assumption can no longer be guaranteed to hold. In contrast, the "many-good-sets" assumption introduced below is weaker, and aligns with the general framework of robustly estimating gradients in each iteration.

SEVER also assumes the existence of a black box approximate learner.

▶ **Definition 6** (γ-approximate learner). *A learning algorithm $\mathcal{L}$ is called γ-approximate if, for any functions $f_1, \ldots, f_m : \mathcal{W} \to \mathbb{R}$, each bounded below on a closed domain $\mathcal{H}$, the output w of $\mathcal{L}$ is a γ-approximate critical point of $\hat{f}(x) := \frac{1}{m}\sum_{i=1}^m f_i(x)$, that is, there exists $\delta > 0$ such that for all unit vectors v where $w + \delta v \in \mathcal{W}$, we have that $v \cdot \nabla \hat{f}(w) \geq -\gamma$.*

▶ Remark 7. We remark that the existence of a γ-approximate learner implies that the learner can find a γ-approximate critical point of any function f by choosing $f_1 = \ldots = f_m = f$. To our best knowledge, any polynomial-time algorithm that finds approximate critical points requires smoothness of the objective. Therefore, SEVER does not apply to problems where some functions in the distribution are nonsmooth. For example, consider a distribution p^* consisted of two functions with equal probability, $h + g$ and $h - g$, where h is smooth but g is nonsmooth. The population risk is smooth, but the individual functions are not.

In the appendix of [4], the authors consider the "many-good-sets" assumption, an alternative weaker assumption that allows the good set to depend on the point w.

▶ **Assumption 8** ("Many good sets", [4, Assumption D.1]). *For every w, there exists a set $S_{good}(w) \subseteq [n]$ with $|S_{good}(w)| \geq (1-\epsilon)n$ such that $S_{good}(w)$ is "good" with respect to w.*

We remark that the "many-good-sets" assumption allows us to do robust gradient estimation in each iteration. The SEVER paper mentions (without going into detail) that under the "many-good-sets" assumption, projected gradient descent can be used to find a $O(\sigma\sqrt{\epsilon})$-approximate critical point. It is unclear that under what conditions "many-good-sets" assumption can be satisfied, and no excess risk bound or sample complexity is provided.

In this paper, we utilize a further relaxed assumption stated below, which only requires the existence of good sets at points in a fine net of the domain. For these purposes, we define a ξ-net of $\mathcal{W}$ (for some small $\xi > 0$) to be a set $\mathcal{C}$ such that for any $w \in \mathcal{W}$, there exists $w' \in \mathcal{C}$ with $\|w - w'\| \leq \xi$.

▶ **Assumption 9** ("Dense good sets"). *For a given $\xi > 0$, there exists a ξ-net $\mathcal{C}$ of the domain $\mathcal{W}$ such that for every $w \in \mathcal{C}$, there exists a set $S_{good}(w) \subseteq [n]$ with $|S_{good}(w)| \geq (1-\epsilon)n$ such that $S_{good}(w)$ is "good" with respect to w.*

When the "dense good sets" assumption holds, we can approximate the gradient at any point in the domain $\mathcal{W}$ by robustly estimating the gradient at the nearest point in the net. The approximation error will be small, provided that the population risk is smooth and the net is fine enough. (The parameter ξ will depend on σ, as we will see in Algorithm 1.) This relaxed assumption allows us to circumvent the technical difficulties of dealing with infinitely many w, thus removing the requirements of uniform Lipschitzness and smoothness of $f - \overline{f}$ for all f that are used in SEVER. As a consequence, we are able to achieve the same corruption error as SEVER with a *significantly reduced* sample complexity. The next section presents our algorithm that achieves this result.

3 Optimal Rates for Robust SCO under Weak Distributional Assumptions

We now present a net-based algorithm that achieves the minimax-optimal excess risk under the weak assumption that the population risk $\overline{f}$ is smooth.

▶ **Assumption 10.** *Given the distribution p^* over functions $f : \mathcal{W} \to \mathbb{R}$ with $\overline{f} = \mathbf{E}_{f\sim p^*}[f]$, we have that $\overline{f}$ is β-smooth.*

Here, the β-smoothness requirement only applies to the population risk. Each individual function f can have different smoothness parameters.

We outline our algorithm below, see Algorithm 1. The algorithm is based on projected gradient descent with a robust estimator. Here, we treat the robust gradient estimator `RobustEstimator` as a black box, which can be any deterministic stability-based algorithm.

For completeness, we provide an instantiation of the robust gradient estimator due to [7], outlined in Algorithm 2, which at a high level iteratively filters out points that are "far" from the sample mean in a large variance direction. Algorithm 2 runs in polynomial time.

The key innovation lies in its gradient estimation strategy. Rather than computing gradients at arbitrary points, it makes use a dense net of the domain $\mathcal{W}$, estimating the gradient at the current iterate w by the gradient at the nearest point w' in the net to w. The smoothness of the population risk ensures that this approximation remains accurate. As mentioned at the end of Section 2, this strategy helps us avoid the technical challenges of handling infinitely many w with a net argument, thereby achieving optimal rates under significantly weaker distributional assumptions compared to SEVER.

Algorithm 1 Net-based Projected Gradient Descent with Robust Gradient Estimator.

```
Input: $\epsilon$-corrupted set of functions $f_1, \ldots, f_n$, stepsize parameters $\{\eta_t\}_{t\in[T]}$, robust gradient estimator RobustEstimator, $(\sigma\sqrt{\epsilon}/\bar{\beta})$-net of $\mathcal{W}$ denoted by $\mathcal{C}$.
Initialize $w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ do
    Let $w'_{t-1} := \arg\min_{w\in\mathcal{C}} \|w - w_{t-1}\|$ denote the closest point to $w_{t-1}$ in the net.
    Robustly estimate gradient at $w'_{t-1}$ by running the RobustEstimator on samples $\left\{\nabla f_i(w'_{t-1})\right\}_{i=1}^{n}$, which outputs $\tilde{g}_t$.
    $w_t \leftarrow \mathrm{Proj}_{\mathcal{W}}(w_{t-1} - \eta_t \tilde{g}_t)$.
end for
Output: $\hat{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$.
```

Algorithm 2 An Instantiation of the Robust Gradient Estimator: Iterative Filtering [7].

```
Input: $\epsilon$-corrupted set $S \subset \mathbb{R}^d$ of $n$ samples.
Initialize weight function $h : S \to \mathbb{R}_{\geq 0}$ with $h(x) = 1/|S|$ for all $x \in S$.
while $\|h\|_1 \geq 1 - 2\epsilon$ do
    Compute $\mu(h) = \frac{1}{\|h\|_1}\sum_{x\in S} h(x)x$.
    Compute $\Sigma(h) = \frac{1}{\|h\|_1}\sum_{x\in S} h(x)(x - \mu(h))(x - \mu(h))'$.
    Compute approximate largest eigenvector $v$ of $\Sigma(h)$.
    Define $g(x) = |v \cdot (x - \mu(h))|^2$ for all $x \in S$.
    Find largest $t$ such that $\sum_{x\in S: g(x)\geq t} h(x) \geq \epsilon$.
    Define $g_{\geq t}(x) = \begin{cases} g(x) & \text{if } g(x) \geq t \\ 0 & \text{otherwise} \end{cases}$.
    Let $m = \max\{g_{\geq t}(x) : x \in S, h(x) \neq 0\}$.
    Update $h(x) \leftarrow h(x)(1 - g_{\geq t}(x)/m)$ for all $x \in S$.
end while
Output: $\mu(h)$.
```

Efficient Implementation. For implementation efficiency, we propose a grid-based net construction. Let $\xi = \sigma\sqrt{\epsilon}/\bar{\beta}$. We use grid points spaced $\xi/\sqrt{d}$ apart in each dimension, i.e.,

$$\left\{\frac{\xi}{\sqrt{d}} \cdot z = \left(\frac{\xi}{\sqrt{d}} \cdot z_1, \frac{\xi}{\sqrt{d}} \cdot z_2, \ldots, \frac{\xi}{\sqrt{d}} \cdot z_d\right) : z = (z_1, z_2, \ldots, z_d) \in \mathbb{Z}^d, \left\|\frac{\xi}{\sqrt{d}} \cdot z\right\|_2 \leq D\right\}$$

to construct a ξ-net.[4] Given a point w, we can find a net point within ξ distance in $O(d)$ time through: (1) Scaling: Divide w by $\xi/\sqrt{d}$. (2) Rounding: Convert to the nearest integral vector in $\mathbb{Z}^d$. (3) Rescaling: Multiply by $\xi/\sqrt{d}$.

This construction yields a net of size $|\mathcal{C}| = O\left(D\sqrt{d}/\xi\right)^d$, which is larger than the optimal covering number $O((D/\xi)^d)$. While this introduces an extra $\log d$ factor in the excess risk bound (due to union bound over net points), it offers two significant practical advantages: (1) Implicit net: No need to explicitly construct and store the net. (2) Efficient computation: $O(d)$ time for finding the nearest net point. An exponential-time algorithm that achieves the excess risk without the $\log d$ factor is described in Section F.

Polynomial runtime. The robust gradient estimator in Algorithm 2 runs in polynomial time, when used with the grid-based construction above. As can be seen in Section A, the required number of iterations is also polynomial in parameters. Therefore, the algorithm runs in polynomial time overall.

Convergence of Algorithm 1 is described in the following result.

▶ **Theorem 11.** *Suppose that Assumption 2, Assumption 3, and Assumption 10 hold. There are choices of stepsizes $\{\eta_t\}_{t=1}^T$ and T such that, with probability at least $1-\tau$, we have*

$$\overline{f}(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}(w) = \tilde{O}\left(\sigma D\sqrt{\epsilon} + \sigma D\sqrt{\frac{d\log(1/\tau)}{n}}\right).$$

As a consequence, the algorithm achieves excess risk of $O(D\sigma\sqrt{\epsilon})$ with high probability whenever $n = \tilde{\Omega}(d/\epsilon)$.

▶ Remark 12. Theorem 11 is minimax-optimal (up to logarithmic factors). Our sample complexity $n = \tilde{\Omega}(d/\epsilon)$, significant improves over the sample complexity of SEVER, which is $n = \tilde{\Omega}\left(\frac{dL^2}{\epsilon\sigma^2} + \frac{dL^4}{\sigma^4}\right)$.

The following matching lower bound can be established, showing the minimax-optimality (up to logarithmic factors) of Algorithm 1. A proof of the lower bound, drawing in part on an adaptation of [15], is provided in the appendix of the full version of the paper.

▶ **Theorem 13.** *For $d \geq 140$ and $n \geq 62500$, there exist a closed bounded set $\mathcal{W} \subset \mathbb{R}^d$ with diameter at most D and a distribution p^* over functions $f : \mathcal{W} \to \mathbb{R}$ that satisfy the following: Let $\overline{f} = \mathbf{E}_{f\sim p^*}[f]$. We have that for every $w \in \mathcal{W}$ and unit vector v that $\mathbf{E}_{f\sim p^*}[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2] \leq \sigma^2$. Both f (almost surely) and $\overline{f}$ are convex, Lipschitz and smooth. The output $\hat{w}$ of any algorithm with access to an ϵ-corrupted set of functions $f_1, \ldots, f_n$ sampled from p^* satisfies the following with probability at least $1/2$:*

$$\overline{f}(\hat{w}) - \min_{w\in\mathcal{W}} \overline{f}(w) = \Omega\left(D\sigma\sqrt{\epsilon} + D\sigma\sqrt{\frac{d}{n}}\right). \tag{2}$$

3.1 Proof Sketch of Theorem 11

We defer the full proof to Section A and sketch the proof below. In each iteration of Algorithm 1, we estimate the gradient at the current iterate w by applying the robust gradient estimator to its nearest point in the net w'. We can decompose the error as follows:

$$\|\tilde{g}(w') - \nabla\overline{f}(w)\| \leq \|\tilde{g}(w') - \nabla\overline{f}(w')\| + \|\nabla\overline{f}(w') - \nabla\overline{f}(w)\|,$$

[4] Technically, we can choose a grid spaced $\xi/\sqrt{4d}$ apart in each dimension, and add additional points to cover the boundary of the feasible set. This would reduce the size of grid points by almost a factor of 2^d.

where the first term measures the bias of the robust gradient estimator, and the second term is due to the approximation error due to the net.

We will show that there exist good sets for all net (grid) points (cf. Assumption 9) with high probability, so that we can robustly estimate gradients for all points in the net. This gives a bound for the first term in the equation above, whereas the second term can be bounded using smoothness of the population risk $\overline{f}$.

Once we establish the gradient estimation bias in each iteration (with high probability), we use the projected biased gradient descent analysis framework (Section 4) to establish an upper bound on the excess risk.

3.2 Handling Nonsmooth but Lipschitz Population Risks

We now consider a setting in which Assumption 2 and Assumption 3 hold, but $\overline{f}$ is nonsmooth in the sense that Assumption 10 is not satisfied. That is, there is no $\bar{\beta} < \infty$ such that $\overline{f}$ is $\bar{\beta}$-smooth. We assume instead that $\overline{f}$ is $\bar{L}$-Lipschitz for some finite $\bar{L}$. In this setting, we can use convolutional smoothing and run Algorithm 1 on the smoothed objective. Our algorithm works as follows:

1. For every index $i \in [n]$, we independently sample a perturbation $u_i \sim \mathcal{U}_s$, where $\mathcal{U}_s$ is the uniform distribution over the d-dimensional L^2-norm ball of radius s centered at the origin. We replace samples $\{f_i\}_{i=1}^n$ by the smoothed samples $\{f_i(\cdot + u_i)\}_{i=1}^n$.
2. We run Algorithm 1 on the smoothed samples with β replaced by $\frac{\bar{L}\sqrt{d}}{s}$ and σ replaced by $\sqrt{\sigma^2 + 4\bar{L}^2}$.

The modified algorithm has the following convergence guarantees.

▶ **Proposition 14.** *Suppose that Assumption 2 and Assumption 3 hold and that $\overline{f}$ is $\bar{L}$-Lipschitz. There are choices of algorithmic parameters such that, with probability at least $1 - \tau$, the output $\hat{w}_T$ of the modified algorithm satisfies the following excess risk bound:*

$$\overline{f}(\hat{w}_T) - \min_{w \in \mathcal{W}} \overline{f}(w) = \tilde{O}\left((\sigma + \bar{L})D\sqrt{\epsilon} + (\sigma + \bar{L})D\sqrt{\frac{d \log(1/\tau)}{n}}\right).$$

▶ Remark 15. Compared to the smooth case, the excess risk bound has an extra $\bar{L}$ term. Using this result, we can show that under the alternative noncentral moment assumption, that is, instead of Assumption 3, assume that for every $w \in \mathcal{W}$ and unit vector v, $\mathbf{E}_{f \sim p^*}[(v \cdot \nabla f(w))^2] \leq G^2$, our modified algorithm (with σ and $\bar{L}$ both replaced by G) achieves the following excess risk bound.

▶ **Theorem 16.** *Suppose that Assumption 2 holds, and that for every $w \in \mathcal{W}$ and unit vector v, we have $\mathbf{E}_{f \sim p^*}[(v \cdot \nabla f(w))^2] \leq G^2$ for some G. There are choices of algorithmic parameters such that, with probability at least $1 - \tau$, the output $\hat{w}_T$ of the modified algorithm (with σ and $\bar{L}$ both replaced by G) satisfies the following excess risk bound:*

$$\overline{f}(\hat{w}_T) - \min_{w \in \mathcal{W}} \overline{f}(w) = \tilde{O}\left(GD\sqrt{\epsilon} + GD\sqrt{\frac{d \log(1/\tau)}{n}}\right).$$

See Section B for proofs of both results above.

We can show that the excess risk bound in Theorem 16 is minimax-optimal (up to logarithmic factors) under the noncentral moment assumption, as a matching lower bound can be established.

▶ **Theorem 17.** *For $d \geq 140$ and $n \geq 62500$, there exist a closed bounded set $\mathcal{W} \subset \mathbb{R}^d$ with diameter at most D, and a distribution p^* over functions $f : \mathcal{W} \to \mathbb{R}$ that satisfy the following: Let $\overline{f} = \mathbf{E}_{f\sim p^*}[f]$. We have that for every $w \in \mathcal{W}$ and unit vector v that $\mathbf{E}_{f\sim p^*}[(v \cdot \nabla f(w))^2] \leq G^2$. Both f (almost surely) and $\overline{f}$ are convex, Lipschitz and smooth. The output $\hat{w}$ of any algorithm with access to an ϵ-corrupted set of functions $f_1, \ldots, f_n$ sampled from p^* satisfies the following with probability at least $1/2$,*

$$\overline{f}(\hat{w}) - \overline{f}^* = \Omega\left(DG\sqrt{\epsilon} + DG\sqrt{\frac{d}{n}}\right). \tag{3}$$

The proof is essentially the same as that of Theorem 13 (available in the full version of the paper), since the same hard instances that are used to establish Theorem 13 can be reused to establish Theorem 17.

3.3 Handling Unknown Covariance Parameter σ

In Algorithm 1, σ primarily affects the fineness of the net through $\xi = \sigma\sqrt{\epsilon}/\bar{\beta}$. When σ is unknown, we can adapt our algorithm with a preprocessing step to estimate σ: (1) Run iterative filtering (Algorithm 2) to obtain a lower bound $\hat{\sigma}$ of σ, and (2) Run Algorithm 1 with the modified fineness parameter $\xi = \hat{\sigma}\sqrt{\epsilon}/\bar{\beta}$. This adaptation preserves the optimal excess risk guarantees for the known σ case. Full details are provided in Section G.

4 Projected Gradient Descent with Robust Gradient Estimator

Algorithm 1 uses a net-based approach to estimate gradients robustly. A more naïve approach is to directly estimate gradients at arbitrary points using a robust gradient estimator. We will show that the simple projected gradient descent algorithm can achieve the same optimal rate as Algorithm 1 under stronger assumptions. Even so, our new assumptions are still slightly weaker than those used in SEVER [4]. Concretely, following assumptions on the distribution over functions are assumed.

▶ **Assumption 18.** *Let p^* be a distribution over functions $f : \mathcal{W} \to \mathbb{R}$ with $\overline{f} = \mathbf{E}_{f\sim p^*}[f]$ so that:*

1. *$f - \overline{f}$ is L-Lipschitz and β-smooth almost surely, where[5] $L \geq \sigma$.*
2. *$\overline{f}$ is $\bar{\beta}$-smooth or $\bar{L}$-Lipschitz.*

We use bars on the constants $\bar{\beta}$ and $\bar{L}$ to emphasize that they reflect properties of $\overline{f}$.

Algorithm 3 follows the "many-good-sets" assumption. We are able to robustly estimate the gradient of the population risk $\overline{f}$ at any point w with high probability, at the cost of requiring additional almost-sure assumptions on $f - \overline{f}$ compared to Algorithm 1.

[5] Without loss of generality, see Section E.

Algorithm 3 Projected Gradient Descent with Robust Gradient Estimator.

```
Input: $\epsilon$-corrupted set of functions $f_1, \dots, f_n$, stepsize parameters $\{\eta_t\}_{t\in[T]}$, robust gradient estimator RobustEstimator$(w)$.
Initialize $w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ do
    Apply robust gradient estimator to get $\tilde{g}_t = $ RobustEstimator$(w_{t-1})$.
    $w_t \leftarrow \mathrm{Proj}_{\mathcal{W}}(w_{t-1} - \eta_t \tilde{g}_t)$.
end for
Output: $\hat{w}_T = \frac{1}{T}\sum_{t=1}^{T} w_t$.
```

Algorithm 3 achieves the same optimal excess risk bounds as in Theorem 11.

▶ **Theorem 19.** *Suppose that Assumption 2, Assumption 3, and Assumption 18 hold. There are choices of stepsizes $\{\eta_t\}_{t=1}^T$ and T such that, with probability at least $1 - \tau$, we have*

$$\overline{f}(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}(w) = \tilde{O}\left(\sigma D\sqrt{\epsilon} + \sigma D\sqrt{\frac{d\log(1/\tau)}{n}}\right).$$

As a consequence, the algorithm achieves excess risk of $O(D\sigma\sqrt{\epsilon})$ with high probability whenever $n = \tilde{\Omega}(d/\epsilon)$. The expected excess risk is bounded by $\tilde{O}\left(\sigma D\sqrt{\epsilon} + \sigma D\sqrt{d/n}\right)$.

The proof is based on the net argument, similar to that of Lemma C.5 in [14]. The high level idea is as follows: For simplicity, we say w is "good" if there exists a good set of functions at w. We will show that with high probability, there exists a good set for all w (cf. Assumption 8), so that we can robustly estimate the gradient at all w. To show this, we employ a net argument, based on the claim that if w is "good", then all points in a small neighborhood of w are also "good." By the union bound, with high probability, all points in the net are "good". Then it follows that all w are "good". The full proof can be found in Section C.

5 Conclusion and Future Work

In this work, we have advanced robust stochastic convex optimization under the ϵ-contamination model. While the prior state of the art SEVER [4] focused on finding approximate critical points under stringent assumptions, we have developed algorithms that directly tackle population risk minimization, obtaining the optimal excess risk under more practical assumptions. Our first algorithm (Algorithm 1) achieves the minimax-optimal excess risk by leveraging our relaxed "dense-good-sets" assumption and estimating gradients only at points in a net of the domain, relaxing the stringent distributional conditions as required in SEVER. Our second algorithm (Algorithm 3) provides a simple projected gradient descent approach that achieves the same optimal excess risk, making use of the "many-good-sets" assumption briefly noted in [4]. Both of our algorithms significantly reduce sample complexity compared to SEVER.

For future work, it would be interesting to explore following directions: (1) Our excess risk bound is tight up to logarithmic factors. Can we improve the bound to remove the logarithmic factors? (2) Our lower bound is with constant probability. Is it possible to derive a lower bound that the includes $\log(1/\tau)$ term with probability τ? (3) Robustness has been shown to be closely related to differential privacy [10]. Can we design optimization algorithms that are both robust and differentially private?

References

1 Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

2 Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. Optimal robust linear regression in nearly linear time. *arXiv preprint arXiv:2007.08137*, 2020. arXiv:2007.08137.

3 Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019. doi:10.1137/17M1126680.

4 Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. SEVER: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019. URL: http://proceedings.mlr.press/v97/diakonikolas19a.html.

5 Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2017. URL: http://proceedings.mlr.press/v70/diakonikolas17a.html.

6 Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019. arXiv:1911.05911.

7 Ilias Diakonikolas, Daniel M Kane, and Ankit Pensia. Outlier robust mean estimation with subgaussian rates via stability. *Advances in Neural Information Processing Systems*, 33:1830–1840, 2020.

8 Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019. doi:10.1137/1.9781611975482.170.

9 Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *Advances in Neural Information Processing Systems*, 29, 2016.

10 Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 497–506, 2023. doi:10.1145/3564246.3585115.

11 Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

12 Arun Jambulapati, Jerry Li, Tselil Schramm, and Kevin Tian. Robust regression revisited: Acceleration and improved estimation rates. *Advances in Neural Information Processing Systems*, 34:4475–4488, 2021. URL: https://proceedings.neurips.cc/paper/2021/hash/23b023b22d0bf47626029d5961328028-Abstract.html.

13 Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018. URL: http://proceedings.mlr.press/v75/klivans18a.html.

14 Shuyao Li, Yu Cheng, Ilias Diakonikolas, Jelena Diakonikolas, Rong Ge, and Stephen Wright. Robust second-order nonconvex optimization and its application to low rank matrix sensing. *Advances in Neural Information Processing Systems*, 36, 2024.

15 Andrew Lowy and Meisam Razaviyayn. Private stochastic optimization with large worst-case lipschitz parameter: Optimal rates for (non-smooth) convex losses and extension to non-convex losses. In *International Conference on Algorithmic Learning Theory*, pages 986–1054. PMLR, 2023. URL: https://proceedings.mlr.press/v201/lowy23a.html.

16 Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627, 2020.

17 John Wilder Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

18 Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On stochastic gradient and subgradient methods with adaptive steplength sequences. *Automatica*, 48(1):56–67, 2012. doi:10.1016/J.AUTOMATICA.2011.09.043.

A Proof of Theorem 11

We start with the robust estimation result from [7], then proceed with the proof of Theorem 11.

▶ **Lemma 20** ([7, Proposition 1.5]). *Let S be an ϵ-corrupted set of n samples from a distribution in $\mathbb{R}^d$ with mean μ and covariance Σ such that $\Sigma \preceq \sigma^2 I$. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \le c$ be given, for a constant $c > 0$. Then any stability-based algorithm (e.g. Algorithm 2) on input S and ϵ', efficiently computes $\widehat{\mu}$ such that with probability at least $1 - \tau$, we have*

$$\|\widehat{\mu} - \mu\| = O(\sigma \cdot \delta(\tau)), \ \textit{where}\ \delta(\tau) = \sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}. \tag{4}$$

Proof of Theorem 11. 1. Bound the bias of the gradient estimator at w. For given w, let $w' = \arg\min_{z\in\mathcal{C}} \|z - w\|$. Applying Lemma 20 to samples $\nabla f_1(w'), \nabla f_2(w'), \ldots, \nabla f_n(w')$, we have that with probability at least $1 - \tau'$, the robust gradient estimator $\tilde{g}(w')$ satisfies

$$\|\tilde{g}(w') - \nabla \overline{f}(w')\| = \sigma \cdot \tilde{O}\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau')/n}\right).$$

We have $\|w - w'\| \le \sigma\sqrt{\epsilon}/\bar{\beta}$ by definition of the net. By $\bar{\beta}$-smoothness of the population risk $\overline{f}$, we have

$$\|\nabla \overline{f}(w) - \nabla \overline{f}(w')\| \le \bar{\beta}\|w - w'\| \le \sigma\sqrt{\epsilon}. \tag{5}$$

Combining the two bounds, we have

$$\|\tilde{g}(w') - \nabla \overline{f}(w)\| = \sigma \cdot \tilde{O}\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau')/n}\right). \tag{6}$$

2. Apply the union bound over all points in the net $\mathcal{C}$. By union bound, setting $\tau' = \tau/|\mathcal{C}|$, we have that with probability at least $1 - \tau$, (6) simultaneously holds for all $w' \in \mathcal{C}$. Recall $|\mathcal{C}| = O\big(D\sqrt{d}/\xi\big)^d$. We have $\log|\mathcal{C}| = \tilde{O}(d)$. It follows that, with probability at least $1 - \tau$, simultaneously for all $w \in \mathcal{W}$, let $w' = \arg\min_{z\in\mathcal{C}} \|z - w\|$, we have

$$\|\tilde{g}(w') - \nabla \overline{f}(w)\| = \sigma \cdot \tilde{O}\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{d\log(1/\tau)/n}\right). \tag{7}$$

Therefore, with probability at least $1 - \tau$, the bias of the gradient estimator at w is bounded by the above expression, simultaneously for all $w \in \mathcal{W}$.

3. Apply the projected biased gradient descent analysis. By Lemma 27, choosing a constant step size $\eta = 1/\bar{\beta}$, the excess risk of the algorithm is bounded by

$$\overline{f}(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}(w) = \tilde{O}\left(\frac{\bar{\beta}D^2}{T} + D \cdot \left(\sigma\sqrt{\epsilon} + \sigma\sqrt{\frac{d\log(1/\tau)}{n}}\right)\right). \tag{8}$$

Choosing $T = \tilde{\Omega}\left(\frac{\bar{\beta}D}{\sigma\sqrt{\epsilon}+\sigma\sqrt{\frac{d\log(1/\tau)}{n}}}\right)$ gives the optimal rate. ◀

B Analysis of Convolutional Smoothing for Nonsmooth but Lipschitz Population Risk

Before proving Proposition 14, we need some properties of the convolutional smoothing.

▶ **Lemma 21** ([18]). *Suppose $\{f(w)\}$ is convex and L-Lipschitz over $\mathcal{W} + B_2(0,s)$, where $B_2(0,s)$ is the d-dimensional L^2 ball of radius s centered at the origin. For $w \in \mathcal{W}$, the convolutional smoother with radius s, $\tilde{f}_s(w) := \mathbb{E}_{u\sim\mathcal{U}_s}[f(w+u)]$, where $\mathcal{U}_s$ is the uniform distribution over $B_2(0,s)$, has the following properties:*

1. *$f(w) \le \tilde{f}_s(w) \le f(w) + Ls$;*
2. *$\tilde{f}_s(w)$ is convex and L-Lipschitz;*
3. *$\tilde{f}_s(w)$ is $\frac{L\sqrt{d}}{s}$-smooth.*

Proof of Proposition 14. Let $\overline{f}_s(w) = \mathbb{E}_{u\sim\mathcal{U}_s}[\overline{f}(w+u)]$ be the smoothed population risk.

1. By properties of convolutional smoothing (part 3 of Lemma 21), we know that since $\overline{f}$ is $\bar{L}$-Lipschitz, $\overline{f}_s$ is $(\bar{L}\sqrt{d}/s)$-smooth.
   As $f_1, f_2, \dots, f_n$ are ϵ-corrupted samples from p^*, we know $f_1(\cdot+u_1), f_2(\cdot+u_2), \dots, f_n(\cdot+u_n)$ are ϵ-corrupted samples from the product distribution of p^* and $\mathcal{U}_s$. Below, we show that in expectation, perturbed gradient (clean) samples $\nabla f(w+u)$ is equal to the smoothed gradient $\nabla\overline{f}_s(w)$.
   By the law of total expectation, we have
   $$\mathbb{E}_{f\sim p^*,\, u\sim\mathcal{U}_s}\left[\nabla f(w+u)\right] = \mathbb{E}_{u\sim\mathcal{U}_s}\Big[\mathbb{E}_{f\sim p^*}\left[\nabla f(w+u)\right]\Big],$$
   and using the regularity condition to interchange the gradient with the expectation, we obtain
   $$\mathbb{E}_{f\sim p^*,\, u\sim\mathcal{U}_s}\left[\nabla f(w+u)\right] = \mathbb{E}_{u\sim\mathcal{U}_s}\Big[\nabla\overline{f}(w+u)\Big].$$
   Below we drop the distributions and write $\mathbb{E}_f$, $\mathbb{E}_u$ for simplicity. Since $\overline{f}$ is $\bar{L}$-Lipschitz, we can exchange the order of expectation and gradient, that is,
   $$\mathbb{E}_u\Big[\nabla\overline{f}(w+u)\Big] = \nabla\mathbb{E}_u\Big[\overline{f}(w+u)\Big] = \nabla\overline{f}_s(w), \tag{9}$$
   which shows that $\mathbb{E}_{f,u}\Big[\nabla f(w+u)\Big] = \nabla\overline{f}_s(w)$.
2. Next, we bound the covariance of the perturbed gradient $\mathrm{Cov}_{f,u}(\nabla f(w+u))$.
   Using law of total covariance, that is, $\mathrm{Cov}(X,Y) = \mathbb{E}(\mathrm{Cov}(X,Y \mid Z)) + \mathrm{Cov}(\mathbb{E}(X \mid Z), \mathbb{E}(Y \mid Z))$, conditioned on u, we can write the covariance of the perturbed gradient as
   $$\begin{aligned}\mathrm{Cov}_{f,u}(\nabla f(w+u)) &= \mathbb{E}_u\left[\mathrm{Cov}_f(\nabla f(w+u))\right] + \mathrm{Cov}_u\left[\mathbb{E}_f[\nabla f(w+u)]\right]\\ &= \underbrace{\mathbb{E}_u\left[\mathrm{Cov}_f(\nabla f(w+u))\right]}_{\text{Term 1}} + \underbrace{\mathrm{Cov}_u\left[\nabla\overline{f}(w+u)\right]}_{\text{Term 2}},\end{aligned} \tag{10}$$
   In the first term, for all u, the covariance inside the expectation is bounded by $\sigma^2 I$ by assumption. Therefore, we have (Term 1) $\preceq \mathbb{E}_u[\sigma^2 I] = \sigma^2 I$. We bound the second term using the boundedness of the gradient. Consider the following fact: for any random vector u, we have $\mathbb{E}[uu^\top] \le C^2 I$ if $\|u\| \le C$ almost surely, and consequently, $\mathrm{Cov}(u) \le 4C^2 I$. It follows that, (Term 2) $\preceq 4\bar{L}^2 I$. Therefore, the covariance parameter increases from σ^2 to $\sigma^2 + 4\bar{L}^2$ due to smoothing.
   We have verified the conditions to apply the original algorithm, with σ replaced by $\sqrt{\sigma^2 + 4\bar{L}^2} = O(\sigma + \bar{L})$ and $\bar{\beta} = \bar{L}\sqrt{d}/s$.

3. By applying Theorem 11 to the smoothed function $\overline{f}_s$, we know there are choices of $\{\eta_t\}_{t=1}^T$ and T such that, with probability at least $1-\tau$, the output $\hat{w}_T$ satisfies,

$$\overline{f}_s(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}_s(w) = \tilde{O}\left((\sigma+\bar{L})D\sqrt{\epsilon} + (\sigma+\bar{L})D\sqrt{\frac{d\log(1/\tau)}{n}}\right).$$

By properties of convolutional smoothing, we have

$$\overline{f}(\hat{w}_T) \le \overline{f}_s(\hat{w}_T) \quad \text{and} \quad \min_{w\in\mathcal{W}} \overline{f}_s(w) \le \min_{w\in\mathcal{W}} \overline{f}(w) + \bar{L}s.$$

It follows that, choosing $s = \tilde{O}\left(D(\sigma/\bar{L}+1)(\sqrt{\epsilon} + \sqrt{d\log(1/\tau)/n})\right)$, we have

$$\begin{aligned}\overline{f}(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}(w) &\le \overline{f}_s(\hat{w}_T) - \min_{w\in\mathcal{W}} \overline{f}_s(w) + \bar{L}s \\ &= \tilde{O}\left((\sigma+\bar{L})D\sqrt{\epsilon} + (\sigma+\bar{L})D\sqrt{\frac{d\log(1/\tau)}{n}}\right).\end{aligned} \tag{11}$$

◀

As a corollary, the result under the noncentral moment condition (Theorem 16) follows using the lemma below, which the relates noncentral second moment to the mean.

▶ **Lemma 22.** *For a d-dimensional random vector u that satisfies $\mathbb{E}[uu^\top] \preceq G^2 I_d$, we have that $\|\mathbb{E}[u]\| \le G$.*

Proof. Let v be any fixed unit vector in $\mathbb{R}^d$. By Jensen's inequality, we have

$$\left(v^\top\mathbb{E}[u]\right)^2 = \left(\mathbb{E}[v^\top u]\right)^2 \le \mathbb{E}[(v^\top u)^2] = v^\top\mathbb{E}[uu^\top]v \le G^2 \tag{12}$$

Since this inequality holds for any unit vector v, we can choose $v = \frac{\mathbb{E}[u]}{\|\mathbb{E}[u]\|}$ when $\mathbb{E}[u] \ne 0$, which gives $\|\mathbb{E}[u]\|^2 \le G^2$, and thus $\|\mathbb{E}[u]\| \le G$. If $\mathbb{E}[u] = 0$, inequality $\|\mathbb{E}[u]\| \le G$ holds trivially. ◀

Proof of Theorem 16. For any random vector u, we have for any unit vector v,

$$\mathbb{E}[(v^\top(u - \mathbb{E}[u]))^2] = \mathbb{E}[(v^\top u)^2] - \|v^\top\mathbb{E}[u]\|^2.$$

Substituting $u = \mathbb{E}_f[\nabla f(w)]$, we have $\mathbb{E}[u] = \nabla\overline{f}(w)$. So for every w, we have $\mathbf{E}[(v\cdot(\nabla f(w) - \nabla\overline{f}(w)))^2] \le \mathbf{E}[(v\cdot\nabla f(w))^2] \le G^2$ for any unit vector v. We know that $\mathbf{E}[\nabla f(w)\nabla f(w)^\top] \preceq G^2 I_d$ is equivalent to $\mathbf{E}[(v\cdot\nabla f(w))^2] \le G^2$ holding for every unit vector v. By Lemma 22, we have $\|\nabla\overline{f}(w)\| \le G$ for every w. Therefore, applying Proposition 14 with $\sigma = G$ and $\bar{L} = G$, we obtain the desired result. ◀

C Analysis of Algorithm 3

Before proving Theorem 19, we need some results from robust estimation literature.

C.1 Results from Robust Mean Estimation

Recall Definition 4. The "good" set property is a special case of stability, defined as follows:

▶ **Definition 23** (Stability [3]). *Fix $0 < \epsilon < 1/2$ and $\delta \ge \epsilon$. A finite set $S \subset \mathbb{R}^d$ is (ϵ,δ)-stable with respect to mean $\mu \in \mathbb{R}^d$ and σ^2 if for every $S' \subseteq S$ with $|S'| \ge (1-\epsilon)|S|$, the following conditions hold: (i) $\|\mu_{S'} - \mu\| \le \sigma\delta$, and (ii) $\|\overline{\Sigma}_{S'} - \sigma^2 I\| \le \sigma^2\delta^2/\epsilon$, where $\mu_{S'} = (1/|S'|)\sum_{x\in S'} x$ and $\overline{\Sigma}_{S'} = (1/|S'|)\sum_{x\in S'}(x-\mu)(x-\mu)^\top$.*

The following due to [7] establishes stability of samples from distributions with bounded covariance.

▶ **Lemma 24** ([7])**.** *Fix any $0 < \tau' < 1$. Let S be a multiset of n i.i.d. samples from a distribution on $\mathbb{R}^d$ with mean μ and covariance Σ such that $\Sigma \preceq \sigma^2 I$. Let $\epsilon' = \Theta(\log(1/\tau')/n + \epsilon) \leq c$, for a sufficiently small constant $c > 0$. Then, with probability at least $1 - \tau'$, there exists a subset $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon')n$ and S' is $(2\epsilon', \delta')$-stable with respect to μ and σ^2, where $\delta' = \delta(\tau')$ depends on τ' as $\delta(\tau') = O(\sqrt{(d \log d)/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau')/n})$.*

With the stability condition, we can robustly estimate the mean of a distribution with bounded covariance.

▶ **Lemma 25** (Robust Mean Estimation Under Stability [3])**.** *Let $T \subset \mathbb{R}^d$ be an ϵ-corrupted version of a set S with the following stability properties: S contains a subset $S' \subseteq S$ such that $|S'| \geq (1 - \epsilon)|S|$ and S' is $(C\epsilon, \delta)$ stable with respect to $\mu \in \mathbb{R}^d$ and σ^2, for a sufficiently large constant $C > 0$. Then there is a polynomial-time algorithm (e.g. Algorithm 2), that on input ϵ, T, computes $\widehat{\mu}$ such that $\|\widehat{\mu} - \mu\| = O(\sigma\delta)$.*

C.2 Proof of Theorem 19

As long as the stability condition holds, we can use deterministic stability-based algorithms (e.g. Algorithm 2) to robustly estimate the mean. Using union bound over the net, it suffices to argue that at a given point w, given the existence of a stable subset of the form $\{\nabla f_i(w)\}_{i \in \mathcal{I}}$, where $\mathcal{I}$ denotes the index set of the stable subset at w, such subset is also stable within a small neighborhood of w, that is, $\{\nabla f_i(w')\}_{i \in \mathcal{I}}$ is stable for all w' in a small neighborhood of w. We have the following stability result, which corresponds to "many-good-sets" Assumption 8.

▶ **Lemma 26.** *Under Assumption 3 and Assumption 18, let $f_1, \ldots, f_n$ denote an ϵ-corrupted set of functions sampled from p^*. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \leq c$ be given, for a constant $c > 0$. With probability at least $1 - \tau$, for all $w \in \mathcal{W}$, there exists index set $\mathcal{I} \subseteq [n]$ (here $\mathcal{I}$ depends on the choice of w) such that $|\mathcal{I}| \geq (1 - \epsilon')n$ and $\{\nabla f_i(w)\}_{i \in \mathcal{I}}$ is $(2\epsilon', \delta(\tau'))$-stable with respect to $\nabla \overline{f}(w)$ and σ^2, where $\tau' = \tau / \exp(\tilde{O}(d))$ and $\delta(\tau') = \tilde{O}\left(\sqrt{\epsilon} + \sqrt{d \log(1/\tau)/n}\right)$.*

Proof. We use a net argument to show that the stability condition holds for all w, following similar proof techniques used in [14]. For fixed w, by Lemma 24, with probability at least $1 - \tau'$, there exists a subset $\mathcal{I} \subseteq [n]$ such that $|\mathcal{I}| \geq (1 - \epsilon')n$ and $\{\nabla f_i(w)\}_{i \in \mathcal{I}}$ is $(2\epsilon', \delta')$-stable where $\delta' = \delta(\tau')$, with respect to $\nabla \overline{f}(w)$ and σ^2, that is

$$\left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f_i(w) - \nabla \overline{f}(w) \right\| \leq \sigma\delta', \tag{13a}$$

$$\left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (\nabla f_i(w) - \nabla \overline{f}(w))(\nabla f_i(w) - \nabla \overline{f}(w))^\top - \sigma^2 I \right\| \leq \sigma^2 \delta'^2 / \epsilon'. \tag{13b}$$

By β-smoothness of $f_i - \overline{f}$, we have

$$\begin{aligned} \left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f_i(w') - \nabla \overline{f}(w') \right\| &\leq \left\| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (\nabla f_i(w') - \nabla f_i(w)) \right\| + \|\nabla f_i(w) - \nabla \overline{f}(w)\| \\ &\leq \beta \|w' - w\| + \sigma\delta'. \end{aligned} \tag{14}$$

Therefore, (13a) holds (up to a constant factor) for all w' such that $\|w - w'\| \leq \sigma\delta'/\beta$.

Next, Equation (13b) is equivalent to the following: for any unit vector v, we have

$$\left| \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} (v \cdot (\nabla f_i(w) - \nabla \overline{f}(w)))^2 - \sigma^2 \right| \leq \sigma^2 \delta'^2 / \epsilon'.$$

By L-Lipschitzness and β-smoothness of $f_i - \overline{f}$, for any unit vector v, we have

$$\begin{aligned} &\left| \left\{ v \cdot \left(\nabla f_i(w) - \nabla \overline{f}(w) \right) \right\}^2 - \left\{ v \cdot \left(\nabla f_i(w') - \nabla \overline{f}(w') \right) \right\}^2 \right| \\ &\quad = \left\{ v \cdot \left(\nabla f_i(w) - \nabla \overline{f}(w) \right) + v \cdot \left(\nabla f_i(w') - \nabla \overline{f}(w') \right) \right\} \\ &\quad\quad \cdot \left\{ v \cdot \left(\nabla f_i(w) - \nabla \overline{f}(w) \right) - v \cdot \left(\nabla f_i(w') - \nabla \overline{f}(w') \right) \right\} \leq 2L \cdot \beta \|w - w'|, \end{aligned} \tag{15}$$

It follows that (13b) holds (up to a constant factor) for w' such that $\|w - w'\| \leq \sigma^2 \delta'^2 / (\epsilon' L \beta)$.

Let $\xi = \min\left(\sigma\delta'/\beta, \sigma^2 \delta'^2 / (\epsilon' L \beta) \right)$. Then, for all w' such that $\|w - w'\| \leq \xi$, $\{\nabla f_i(w')\}_{i \in \mathcal{I}}$ is $(2\epsilon', 2\delta')$-stable with respect to $\nabla \overline{f}(w')$ and σ^2. It suffices to choose a ξ-net $\mathcal{C}$ of $\mathcal{W}$, where the optimal size of the net is $|\mathcal{C}| = O((D/\xi)^d)$, and choose $\tau' = \tau / |\mathcal{C}|$. By union bound, with probability at least $1 - |\mathcal{C}|\tau'$, the stable subset exists for all $w \in \mathcal{C}$ simultaneously. Since we have argued that for fixed w, the same stable subset applies for all w' within distance ξ from w, the subset stability holds simultaneously for all w with probability at least $1 - \tau$, as claimed. ◀

Proof of Theorem 19. Combining Lemma 26 and Lemma 25, with probability at least $1 - \tau$, in each iteration, we can estimate the gradient up to a bias as follows:

$$\|\tilde{g}(w_t) - \nabla \overline{f}(w_t)\| = O(\sigma \cdot \delta(\tau')) = \tilde{O}\left(\sigma\sqrt{\epsilon} + \sigma\sqrt{d \log(1/\tau)/n} \right).$$

Note that the probability is simultaneously for all w, so it does not matter how many iterations we run. Conditioned on the gradient estimation bias bound being held, the excess risk bound then follows by applying Lemma 27 for smoothness loss, or Lemma 28 for Lipschitz loss with corresponding choices of stepsizes and large enough T. ◀

D Projected Biased Gradient Descent

In this section, we analyze the convergence of the projected gradient descent algorithm with a biased gradient estimator. We assume the loss function is convex throughout this section.

Both Algorithm 1 and Algorithm 3 can be viewed as instances of the following algorithm.

Algorithm 4 Projected Gradient Descent with Biased Gradient Estimator.

```
Input: Convex function $F$, stepsize parameters $\{\eta_t\}_{t \in [T]}$, biased gradient estimator BiasedEstimator($w$), feasible set $\mathcal{W}$
Initialize $w_0 \in \mathcal{W}$ and $t = 1$.
for $t \in [T]$ do
    Let $\tilde{g}_t$ = BiasedEstimator($w_t$).
    $w_t \leftarrow \Pi_{\mathcal{W}}(w_{t-1} - \eta_t \tilde{g}_t)$.
end for
Output: $\hat{w}_T = \frac{1}{T} \sum_{t=1}^{T} w_t$.
```

Here, $\Pi_{\mathcal{W}}(\cdot)$ denotes the projection operator onto the feasible set $\mathcal{W}$, that is,

$$\Pi_{\mathcal{W}}(y) = \arg\min_{w \in \mathcal{W}} \|w - y\|^2.$$

The projection operation ensures that the iterates w_t remain within the feasible set $\mathcal{W}$ throughout the optimization process. The projection step is crucial when the optimization problem is constrained, as it guarantees that the updates do not violate the constraints defined by $\mathcal{W}$.

Let us assume that the biased gradient estimator $\tilde{g}_t$ has a bias bounded by B, that is, $\|\tilde{g}_t - F(w_t)\| \le B$ for all t, and the diameter of the feasible set $\mathcal{W}$ is bounded by D, that is, $\|w - w'\| \le D$ for all $w, w' \in \mathcal{W}$. We have the following convergence results for the algorithm.

▶ **Lemma 27.** *Suppose F is convex and β-smooth. Running Algorithm 4 with constant step size $\eta = \frac{1}{\beta}$, we have*

$$F\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - \min_{w\in\mathcal{W}} F(w) \le \frac{\beta D^2}{2T} + BD. \tag{16}$$

Alternatively, we can consider the case where the loss function $F(w)$ is convex and L-Lipschitz. The following lemma holds.

▶ **Lemma 28.** *Suppose F is convex and L-Lipschitz. Running Algorithm 4 with constant step size $\eta = \frac{1}{\beta}$, we have*

$$F\left(\frac{1}{T}\sum_{t=1}^{T} w_t\right) - \min_{w\in\mathcal{W}} F(w) \le \frac{DL}{\sqrt{T}} + \left(\frac{1}{\sqrt{T}} + 1\right) BD. \tag{17}$$

The proof of the two convergence results above can be found in the full version.

E Discussions on the Assumptions

E.1 On the bounded covariance assumption

Without loss of generality, we can assume $\sigma \le L$. The reason is as follows: By Lipschitzness, we have $\|\nabla f(w) - \nabla \overline{f}(w)\| \le L$ almost surely. By Cauchy-Schwarz Inequality, we have $\mathbf{E}_{f\sim p^*}[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2] \le L^2$ holds for any unit vector v. On the other hand, L can be as larger as $\sqrt{d} \cdot \sigma$ (e.g. consider standard multivariate normal).

The condition $\mathbf{E}[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2] \le \sigma^2$ for every unit vector v is equivalent to requiring that the covariance matrix Σ_w of the gradients $\nabla f(w)$ satisfies $\Sigma_w \preceq \sigma^2 I$.

▶ **Proposition 29.** *Let Σ_w denote the covariance matrix of the gradients $\nabla f(w)$. For given w, the following two assumptions are equivalent:*

1. *For every unit vector v, we have $\mathbf{E}_{f\sim p^*}[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2] \le \sigma^2$.*
2. *The covariance matrix satisfies $\Sigma_w \preceq \sigma^2 I$.*

Furthermore, since Σ_w is positive semidefinite, by definition of the spectral norm, the latter assumption can be equivalently written as $\|\Sigma_w\| \le \sigma^2$.

Proof. By definition,

$$\Sigma_w = \mathbf{E}_{f\sim p^*}[(\nabla f(w) - \nabla \overline{f}(w))(\nabla f(w) - \nabla \overline{f}(w))^\top]. \tag{18}$$

We have

$$\begin{aligned}\mathbf{E}_{f\sim p^*}\left[(v \cdot (\nabla f(w) - \nabla \overline{f}(w)))^2\right] &= \mathbf{E}_{f\sim p^*}\left[v \cdot (\nabla f(w) - \nabla \overline{f}(w))(\nabla f(w) - \nabla \overline{f}(w))^\top \cdot v\right] \\ &= v^\top \cdot \mathbf{E}_{f\sim p^*}\left[(\nabla f(w) - \nabla \overline{f}(w))(\nabla f(w) - \nabla \overline{f}(w))^\top\right] \cdot v \\ &= v^\top \Sigma_w v.\end{aligned} \tag{19}$$

Therefore, the two assumptions are equivalent. ◀

E.2 On the regularity condition

For technical reasons, we need to assume regularity conditions such that we can exchange the gradient and expectation, that is, for any w,

$$\mathbb{E}_{f\sim p^*}[\nabla f(w)] = \nabla\overline{f}(w). \tag{20}$$

A necessary condition[6] due to dominated convergence theorem for the regularity condition is that there exists some functional (mapping of functions) $g(f)$ such that $\mathbb{E}_{f\sim p^*}[g(f)] < \infty$, and for all w, $\|\nabla f(w)\| \le g(f)$ almost surely.

Consider the following example, where f takes the form $f_X(w) = \frac{1}{2}(X^\top w)^2$ where X is a random vector in $\mathbb{R}^d$ with distribution P. The distribution P of X induces the distribution p^* of functions f. Let P be such that $\mathbb{E}_X[\|XX^T\|_2] \le M$ for some $M > 0$, but X has unbounded support (e.g., $X \sim \mathcal{N}(0, I_d)$ is multivariate Gaussian). Note that $\nabla f_X(w) = XX^\top w$. So we have $\|\nabla f_X(w)\| \le \|XX^\top\|_2 \cdot \|w\|$. In this case, we can take $g(X) = D \cdot \|XX^\top\|$ so that $\|\nabla f_X(w)\| \le g(X)$ almost surely. We have that $\mathbb{E}_X[g(X)] \le \|XX^\top\|_2 \cdot \|w\| \le MC < \infty$. In this case, we can exchange the order of expectation and gradient.

E.3 Comparing bounded covariance assumption with bounded variance assumption

Net-based approaches (e.g. uniform convergence) often suffer from suboptimal error [9]. However, Algorithm 1 indeed achieves the minimax-optimal rate. We believe the reason is due to the bounded covariance assumption $\Sigma \preceq \sigma^2 I$. Below, we provide a discussion on the bounded covariance assumption and compare it with the bounded variance assumption.

The bounded covariance assumption $\Sigma \preceq \sigma^2 I$ is different from the bounded variance assumption $\mathbf{E}\left\|\nabla f(w) - \nabla\overline{f}(w)\right\|^2 \le \Phi^2$ as commonly used in optimization literature without corruption. Using the property $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, this is equivalent to $\operatorname{tr}(\Sigma) \le \Phi^2$.

We comment that neither assumption implies the other. For isotropic Gaussian distribution, where the covariance matrix is $\Sigma = \sigma^2 I$, we have $\operatorname{tr}(\Sigma) = d\sigma^2$. On the other hand, consider the distribution where the variance is concentrated in one direction, i.e., $\Sigma = \Phi^2 \cdot vv^\top$ for some unit vector v. We have $\operatorname{tr}(\Sigma) = \Phi^2$ and $\|\Sigma\| = \Phi^2$. In general, we only know that $\|\Sigma\| \le \operatorname{tr}(\Sigma) \le d\|\Sigma\|$.

Recall Lemma 24. The complete version of the lemma is as follows:

▶ **Lemma 30** ([7]). *Fix any $0 < \tau < 1$. Let S be a multiset of n i.i.d. samples from a distribution on $\mathbb{R}^d$ with mean μ and covariance Σ. Let $\epsilon' = \Theta(\log(1/\tau')/n + \epsilon) \le c$, for a sufficiently small constant $c > 0$. Then, with probability at least $1 - \tau$, there exists a subset $S' \subseteq S$ such that $|S'| \ge (1-\epsilon')n$ and S' is $(2\epsilon', \delta(\tau'))$-stable with respect to μ and $\|\Sigma\|$, where $\delta(\tau') = O(\sqrt{(\mathrm{r}(\Sigma)\log \mathrm{r}(\Sigma))/n} + \sqrt{\epsilon} + \sqrt{\log(1/\tau')/n})$. Here we use $\mathrm{r}(M)$ to denote the stable rank (or intrinsic dimension) of a positive semidefinite matrix M, i.e., $\mathrm{r}(M) := \operatorname{tr}(M)/\|M\|$.*

Following identical proof steps (recall proofs for our algorithms), we can express our excess risk bound in terms of the covariance matrix Σ as follows:

$$D \cdot \tilde{O}\left(\sqrt{\|\Sigma\|\epsilon} + \sqrt{\operatorname{tr}(\Sigma)/n} + \sqrt{d\|\Sigma\|\log(1/\tau)/n}\right). \tag{21}$$

[6] Technically, we need to be more precise about the distribution p^* over functions. In this paper, we follow the same convention as used by [4]. To be more concrete, we can just think of f parameterized by some random variable X and the distribution p^* is induced by the distribution of X. See the example.

In our paper, we consider the bounded covariance assumption $\Sigma \preceq \sigma^2 I$, which is a standard assumption in robust optimization literature. Otherwise, we cannot control the error term $\sqrt{\|\Sigma\|\epsilon}$ due to corruption. In the worse case (e.g. isotropic Gaussian), we have $\|\Sigma\| = \sigma^2$ and $\operatorname{tr}(\Sigma) = d\sigma^2$, so the bound reduces to

$$\sigma \cdot O\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{d\log(1/\tau)/n}\right). \tag{22}$$

We see that the second term already contains the dependence on d. Therefore, the d factor in the last term due to our net-based approach in conjunction with the use of the union bound over the net points, does not affect the rate.

F An exponential time algorithm that achieves the minimax-optimal excess risk bound without $\log d$ factor

Both of our two algorithms achieve the minimax-optimal excess risk bound up to logarithmic factors. In this section, we show that the minimax-optimal excess risk bound can be achieved without the $\log d$ factor, but at the cost of exponential time complexity. Based on Lemma 24, we can remove the $\log d$ factor when estimating the gradients, by using the following framework, as shown in [7].

1. Set $k = \lfloor \epsilon' n \rfloor$. Randomly partition n samples S into k buckets of size $\lfloor n/k \rfloor$ (remove the last bucket if n is not divisible by k).
2. Compute the empirical mean within each bucket and denote the means as $z_1, \ldots, z_k$.
3. Run stability-based robust mean estimation over the set $\{z_1, \ldots, z_k\}$.

Here, the first two steps serve as preprocessing before feeding the data into the robust mean estimation algorithm. We now restate the robust estimation result without $\log d$ factor below.

▶ **Lemma 31.** *Let S be an ϵ-corrupted set of n samples from a distribution in $\mathbb{R}^d$ with mean μ and covariance $\Sigma \preceq \sigma^2 I$. Let $\epsilon' = \Theta(\log(1/\tau)/n + \epsilon) \le c$ be given, for a constant $c > 0$. Then any stability-based algorithm on input S and ϵ', efficiently computes $\widehat{\mu}$ such that with probability at least $1 - \tau$, we have $\|\widehat{\mu} - \mu\| = \sigma \cdot O\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau)/n}\right)$.*

We recall that our efficient implementation using grid points cost a $\log d$ factor due to the suboptimal net size. Using a net with a size matching the covering number $O((D/\xi)^d)$ will remove the $\log d$ factor, but at the cost of exponential time complexity for constructing the net and finding a point within $O(\xi)$ distance for a given point.

Following the same proof steps, as in Section A, we can derive the excess risk bound without the $\log d$ factor, at the cost of exponential time complexity.

G Dealing with Unknown σ

We adapt Algorithm 1 to work without knowing σ by first getting a lower bound on σ using the filtering algorithm (Algorithm 2) and then using this lower bound to set the fineness parameter ξ of the net in Algorithm 1.

The modified algorithm is as follows: (1) Estimate σ: Choose a point w and run Algorithm 2 with input $S = \{\nabla f_i(w)\}_{i=1}^n$ to obtain a lower bound $\hat{\sigma}$. (2) Then, we run Algorithm 1 with $\xi = \hat{\sigma}\delta/\bar{\beta}$.

In Algorithm 1, we use σ only to determine the fineness of the net via $\xi = \sigma\sqrt{\epsilon}/\bar{\beta}$. A smaller ξ results in a finer net and consequently reduces the error when evaluating gradients at the net point w' instead of w, that is, (5) still holds with a smaller ξ. Since the excess risk depends on ξ only through logarithmic terms, the same analysis (see Section A) holds with a smaller ξ. It then suffices to choose $\xi = \hat{\sigma}\delta/\bar{\beta}$, where $\hat{\sigma}$ is a lower bound on σ. We also need to choose $T = \tilde{\Omega}\left(\frac{\beta D}{\hat{\sigma}\sqrt{\epsilon}+\hat{\sigma}\sqrt{\frac{d\log(1/\tau)}{n}}}\right)$ where we use $\hat{\sigma}$ in place of σ. When using smoothing to handle nonsmooth losses, we can choose β similarly by replacing σ with $\hat{\sigma}$.

Recall that Algorithm 2 works even when σ is unknown. Moreover, the output h satisfies $\|\Sigma(h)\| \leq \sigma^2(1+O(\delta^2/\epsilon))$ (see [7]). It follows that we can use $\|\Sigma(h)\|$ to obtain a lower bound on σ. Using Lemma 24, at any fixed w, we can run Algorithm 2 with input $S = \{\nabla f_i(w)\}_{i=1}^n$ to obtain a lower bound $\hat{\sigma}$ on σ. We have that (plugging in $\delta(\tau')$ in Lemma 24), with probability at least $1 - \tau'$,

$$\|\Sigma(h)\| \leq \sigma^2(1 + O(\delta^2/\epsilon)), \tag{23}$$

where $\delta = \tilde{O}\left(\sqrt{\epsilon} + \sqrt{d/n} + \sqrt{\log(1/\tau')/n}\right)$. Therefore, $\hat{\sigma} := \sqrt{\|\Sigma(h)\|}/\sqrt{1 + O(\delta^2/\epsilon)}$ is a lower bound on σ with probability at least $1 - \tau'$.