J. Nonlinear Var. Anal. 10 (2026), No. 2, pp. 201-230
Available online at http://jnva.biemdas.com
https://doi.org/10.23952/jnva.10.2026.2.02

OUTLIER-ROBUST NONSMOOTH STOCHASTIC OPTIMIZATION

SHUYAO LI*, STEPHEN J. WRIGHT, JELENA DIAKONIKOLAS

Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

Abstract. We study nonsmooth stochastic optimization under adversarial data contamination, which models outliers that are often unavoidable in modern machine learning tasks. While robust methods for such settings with *smooth* objectives have been developed, *nonsmooth* models remain largely unexplored despite their central role in machine learning, including regression with ℓ_1 losses, support vector machines, and distributionally robust optimization. We introduce a general framework for outlier-robust nonsmooth optimization, combining robust mean estimation with projected subgradient methods. Our analysis establishes the first polynomial-time algorithms with provable guarantees for nonsmooth (weakly) convex objectives under adversarial corruptions. As a key application, we resolve an open problem in outlier-robust distributionally robust optimization, obtaining polynomial-time algorithms with bounded errors for Conditional Value-at-Risk and f-divergence–based formulations. These results advance the theory of robust nonsmooth optimization and highlight new directions for robust learning with corrupted data.

Keywords. Distributionally robust optimization; Data contamination; Nonsmooth optimization.

1. Introduction

The challenge of learning from contaminated data is a central problem in modern machine learning and statistics. Real-world datasets are frequently corrupted by outliers, which can arise from measurement errors (a central motivation in classical treatments of outliers and robust statistics [1, 2, 28, 43]), as well as out-of-distribution examples [48], adversarial attacks [3, 6, 25, 45, 46], and inherent data heterogeneity, as found in biological datasets, for example [32, 37, 42]. Outliers can severely degrade the performance of standard optimization algorithms, motivating the need for methods that are robust to data corruption. While significant progress has been made in developing outlier-robust algorithms for *smooth* optimization problems [15, 38], many important models in machine learning rely on *nonsmooth* objectives. These include robust regression with nonsmooth losses such as the ℓ_1 loss, low-rank matrix sensing with ℓ_1 loss, support vector machines with the hinge loss, and modern risk-averse optimization arising in, e.g., distributionally robust optimization (DRO) [5, 12, 24, 34, 39, 40, 44].

[*] Corresponding author.

E-mail address: shuyaoli.madison@gmail.com (S. Li), swright@cs.wisc.edu (S. Wright), jelena@cs.wisc.edu (J. Diakonikolas).

Received 30 September 2025; Accepted 22 November 2025; Published online 1 January 2026.

In this work, we address the challenge of solving nonsmooth stochastic optimization problems in the presence of adversarial data corruption. We consider problems of the form

$$\min_{\boldsymbol{x}\in\mathscr{C}} f(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}[\varphi(\boldsymbol{x};\boldsymbol{\xi})], \tag{1.1}$$

where $\boldsymbol{x}\in\mathbb{R}^d$ is the decision variable, $\mathscr{C}\subset\mathbb{R}^d$ is a closed convex set for which we assume access to a projection operator, and $f(\boldsymbol{x})$ is the population objective defined over an underlying (inlier) data distribution $\mathbb{P}$. The per-sample function $\varphi(\boldsymbol{x};\boldsymbol{\xi})$ may be nonsmooth. We make the following standard assumptions: (1) the population objective $f(\boldsymbol{x})$ is ρ-weakly convex and L-Lipschitz continuous; and (2) for each $\boldsymbol{\xi}$, the function $\varphi(\cdot;\boldsymbol{\xi})$ is subdifferentiable with respect to its first argument. We consider the subdifferential in the sense of variational analysis, as formally defined in Section 2. The outliers in this stochastic optimization context arise when an adversary corrupts the data samples $\{\boldsymbol{\xi}_i\}$ drawn from the clean distribution $\mathbb{P}$.

This paper develops a general-purpose, provably robust framework for this broad and important class of nonsmooth optimization problems. We study optimization under the classical, strong adversarial contamination model [16, 26], where an adversary can arbitrarily corrupt a constant fraction of the training data.

Definition 1.1 (Strong Contamination Model)**.** Given a parameter $0<\varepsilon<1/2$ and an underlying (inlier) data distribution $\mathbb{P}$, an algorithm requests N samples drawn i.i.d. from $\mathbb{P}$. An omniscient adversary may then inspect the N clean samples and replace an ε-fraction of them with arbitrary points. The resulting set of N points, termed an *ε-corrupted set*, is then provided to the algorithm.

In this paper, we show that techniques from robust statistics can be effectively integrated with classical nonsmooth optimization. The central idea is to show that a standard algorithm—the projected subgradient method—can be made robust to contamination by equipping it with a *robust subgradient oracle*.

1.1. Overview of Results.

We develop a principled algorithmic framework for outlier-robust nonsmooth optimization. We leverage results in robust statistics, which establish the existence of a robust gradient estimator with bounded error, under mild distributional assumptions on stochastic subgradients (unbiasedness and bounded covariance under outlier-free data). This gives rise to optimization problems with inexact gradients, which we show can be addressed by the subgradient method. We establish required distributional conditions for robust subgradient estimation for a range of tasks in f-divergence-based DRO and, applying our general results for outlier-robust nonsmooth optimization, obtain the first polynomial-time guarantees for addressing such DRO tasks with outliers in the training data while achieving optimality gap error that appears unimprovable. Specifically, we establish the following results.

Convergence Guarantees for the Robustified Algorithm. We provide a rigorous analysis for the inexact projected subgradient method when coupled with an outlier-robust gradient estimation oracle, applying to both convex and weakly convex nonsmooth objectives. To do so, we first observe that, under bounded covariance of stochastic subgradient, existing robust mean estimation algorithms lead to high-probability estimates of the subgradient with error bounded by $\delta=O(\sigma\sqrt{\varepsilon})$, where σ is the bound on the operator norm of the covariance and $\varepsilon\in(0,1/2)$ denotes the fraction of outliers (Section 3.1). Error of this order is unimprovable in general [17].

While first-order optimization methods with inexact oracle pioneered by [13] can be used to address such problems, their blackbox application to our setting would lead to an error (as measured by the optimality gap) scaling with the iterate distance to an optimum, which is not a priori guaranteed to be even bounded, unless the problem is defined on a compact set. Instead, our strategy for the convex setting (Section 3.3) is inspired by the varying regularization technique introduced by Nesterov for the analysis of dual averaging methods [36]. In our context, this strategy enables us to establish a bound on the optimality gap that scales with the *initial* distance to optima D_0; namely, the bound is of the order $O(\delta D_0)$. The scaling of the error with each of these parameters cannot be improved in general; see the discussion in [33]. We further establish a tighter bound $O(\delta^2/\mu)$ for the optimality gap and $O(\delta/\mu)$ for the distance to optimum when minimizing the sum of a convex objective and a simple quadratic $\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2$, where $\boldsymbol{z}$ is a fixed reference point. This stronger bound plays a role in addressing the weakly convex case.

For weakly convex objectives (Section 3.4), we first show, for completeness, that the analysis of [11] can be generalized to the setting with inexact subgradients. Unfortunately, the direct application of this result is insufficient for our purposes, as it only guarantees the existence of one iterate $\boldsymbol{x}_k$ in the entire algorithm run that has a small gradient norm of the Moreau envelope of the objective f (a standard stationarity guarantee for this setting; see further discussion in Section 2). We cannot tell which iterate the algorithm should output, as it is unclear a priori how to certify a small subgradient of the Moreau envelope under inexact subgradient access. Our key idea is to leverage the results for minimizing the sum of a convex function and a quadratic with inexact subgradient access (described in the previous paragraph) to approximate the minimizer of the Moreau envelope, which allows us to estimate its gradient norm for each iterate. This leads to the $O(\delta)$ bound on the Moreau envelope gradient norm, which is unimprovable, due to standard lower bounds for robust mean estimation; see, e.g., [15, 17].

Finally, our error guarantee in all considered settings improves upon prior work [22] by ensuring the error vanishes completely in the absence of data variation; see Section 1.2 for further context.

Solving an Open Problem in Outlier-Robust DRO. We apply our framework to provide the first efficient, provably correct algorithms for outlier-robust Distributionally Robust Optimization with CVaR and, under additional assumptions, Cressie-Read f-divergence ambiguity sets (Section 4). This resolves the algorithmic shortcomings of prior work [49], which relied on heuristics that have been shown to fail [33]. The obtained bounds on the optimality gap appear unimprovable; see the relevant discussion in Section 4.

The key technical component of this result is an oracle that computes an accurate subgradient estimate from corrupted data. Specifically, an application of robust mean estimation algorithms, as mentioned earlier, requires bounded covariance of the stochastic subgradient oracle under the clean inlier distribution. We derive bounds on the operator norm of the covariance matrix under mild assumptions about the loss function (convexity and bounded moment of the subgradient, of the appropriate order) and argue that the results from Section 3.3 can be applied, obtaining the first polynomial-time algorithm with bounded-error guarantees for this problem in the presence of outliers. While not directly comparable to the error lower bounds from [49] due to different assumptions (bounded loss moments vs bounded loss subgradient moments), the error we establish has the same scaling with the key problem parameters as the lower bound from [49], and we conjecture it is optimal.

1.2. **Further Related work.** The problem of designing optimization algorithms that are resilient to data contamination has a rich history and is a major focus of machine learning research.

Robust Smooth Optimization: Parameter (e.g., mean) estimation based on data containing a constant fraction of outliers is a central problem in robust statistics, a classical area initiated in the 1960s [26, 47]. Recent literature in this domain has established the existence of sample and computationally-efficient algorithms for such high-dimensional tasks; see [14, 29] and the recent book [17]. First-order stochastic optimization methods are compatible with robust mean estimation algorithms, as the estimation of the stochastic gradient is fundamentally a high-dimensional mean estimation task. Hence, a significant line of work has addressed outlier-robust stochastic optimization for *smooth* loss functions [15, 38].

A landmark in this area is SEVER [15], a meta-algorithm that wraps an inner-loop optimization solver in an outer-loop filtering procedure. In each outer step, SEVER runs the optimization solver (like gradient descent) to the target accuracy and then uses the resulting model to identify and filter potential outliers by inspecting their empirical gradients. Its theoretical guarantees rely on the property that for a clean dataset near a stationary point, the mean of the true stochastic gradient has a small norm. However, extending this gradient-based filtering to nonsmooth loss functions encounters a fundamental obstacle. For a nonsmooth objective, approximate stationarity is characterized by the existence of a *specific* set of stochastic subgradients whose average is small. Standard nonsmooth solvers (e.g., subgradient descent) select an arbitrary computable subgradient at each step, so they do not provide access to the particular collection of subgradients required for SEVER's analysis to apply. In other words, unlike for smooth problems, the *certification* of approximate stationarity is not possible based on the (sub)gradient norm, which is essential to the outer filtering procedure. Consequently, the suitability of SEVER for problems where the data-dependent loss is itself nonsmooth—the primary focus of our work—is not established, motivating our search for alternative methodologies.

Robust Nonsmooth Convex Optimization: More recently, [22] studied robust stochastic convex optimization and provided rates of convergence. While their framework covers nonsmooth objectives, their suboptimality guarantee (see [22, Proposition 14]) has a significant shortcoming: the error bound has a residual dependence on the problem's Lipschitz constant L, even as the variance of samples $\|\mathrm{Cov}(\boldsymbol{\xi})\|_{\mathrm{op}}$ approaches zero. This implies that even with clean, deterministic data, the algorithm may not converge to the true minimizer. Our proposed framework rectifies this issue, providing guarantees that vanish as $\|\mathrm{Cov}(\boldsymbol{\xi})\|_{\mathrm{op}} \to 0$.

Distributionally Robust Optimization (DRO): DRO is a framework for managing uncertainty by optimizing for worst-case performance over an ambiguity set of distributions. We show that our general robust optimization framework can be used to address f-divergence-based DRO problems with outliers.

DRO is typically used to handle *post-decision* uncertainty, such as distributional shifts between training and testing data [7]. This contrasts with our focus on *pre-decision* uncertainty arising from training data contamination; see also the discussion in [33, Section 3]. The role of DRO depends on the choice of the ambiguity set's radius: A vanishing radius is often used to improve generalization and mitigate statistical overfitting [23], while a constant radius (as used in the DRO applications of this paper) guards against persistent shifts [30].

The nature of the robustness in DRO is determined by the geometry of the ambiguity set. Our applications focus on ambiguity sets defined by the *Cressie-Read family of f-divergences* [10].

This class of ambiguity sets is well suited for learning models that achieve uniformly good performance on the tails of a distribution, or on minority subgroups where standard empirical risk minimization can fail [21], a critical goal in fairness- and safety-conscious applications. The dual representation of the f-divergence-robust objective reveals that it is equivalent to minimizing a risk measure that penalizes high-moment deviations of the loss. A particularly important special case is Conditional Value-at-Risk (CVaR), which arises as a limit of the Cressie-Read family. Minimizing CVaR corresponds to minimizing the expected loss over the worst-performing α-fraction of the data distribution [40], providing a direct mechanism for improving tail performance.

While these DRO formulations are powerful tools, efficient algorithms for optimizing them with rigorous error guarantees in the presence of training data outliers were previously unknown. Our work provides the first such algorithms, applicable to both CVaR and the broader Cressie-Read family. Concurrent work [33] addresses the related problem of outlier-robust DRO for Wasserstein ambiguity sets in the convex setting.

Modeling of Contamination and Shift: The framework that is perhaps most closely related in spirit to our applications is Distributionally Robust Outlier-aware Optimization (DORO) [49], which simultaneously considers training data contamination and test-time distributional shifts. DORO makes the valuable contribution of defining a statistical estimator—the minimizer of the "DORO risk"—that has desirable robustness properties. However, it remains an open question whether this estimator can be computed or even approximated in polynomial time. The algorithm proposed in [49] to optimize this objective is a heuristic that lacks formal convergence guarantees. As shown in a counterexample by [33], this algorithm, which iteratively trims samples with the highest current loss, can fail catastrophically by converging to the outliers rather than the true parameter. This algorithmic failure left open the problem of designing a provably correct method for the setting that DORO aims to address. Our work provides a positive answer to this open problem by developing the first efficient and provably robust algorithm for this task.

2. Preliminaries

This section introduces the necessary concepts for our analysis. We begin by defining the classes of functions to be considered, then formalize the notion of approximate stationarity that is the goal of our optimization algorithm in weakly convex settings. Additional background on the Cressie-Read family of f-divergences, related risk measures, and the DRO formulation is provided in Section 4, as it is the only section where they are used.

2.1. Function Properties and Subdifferentials.

We work with weakly convex functions, which are a broad class of nonsmooth functions that include all convex functions and many objectives found in modern machine learning (see, e.g., [11] for illustrative examples). Our analysis relies on the standard notion of the (Fréchet) subdifferential from variational analysis.

Definition 2.1 (Subdifferentials [41, Definition 8.3])**.** For a function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the *subdifferential* of f at a point $\boldsymbol{x}$ in its domain, denoted $\partial f(\boldsymbol{x})$, is the set of all vectors $\boldsymbol{g} \in \mathbb{R}^d$ satisfying:

$$\liminf_{\boldsymbol{y}\to\boldsymbol{x}} \frac{f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{g}, \boldsymbol{y}-\boldsymbol{x}\rangle}{\|\boldsymbol{y}-\boldsymbol{x}\|_2} \geq 0.$$

If f is convex, this definition coincides with the standard subdifferential from convex analysis. If f is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$.

Definition 2.2 (Weakly Convex Function)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is called ρ*-weakly convex* for $\rho \geq 0$ if the function $\boldsymbol{x} \mapsto f(\boldsymbol{x}) + \frac{\rho}{2}\|\boldsymbol{x}\|_2^2$ is convex. For any $\boldsymbol{g} \in \partial f(\boldsymbol{x})$, this definition implies the inequality:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle - \frac{\rho}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

2.2. **Approximate Stationarity.** For a general nonsmooth, weakly convex function f, a point $\boldsymbol{x}^*$ is stationary if $\boldsymbol{0} \in \partial f(\boldsymbol{x}^*)$. Our goal is to find an iterate $\boldsymbol{x}$ that is approximately stationary. Following modern analyses of nonsmooth optimization, we measure proximity to stationarity using the gradient of the Moreau envelope, which we define next.

Definition 2.3 (Moreau Envelope and Proximal Operator)**.** For a function f and a parameter $\lambda > 0$, the *Moreau envelope* is defined by

$$f_\lambda(\boldsymbol{x}) = \min_{\boldsymbol{u} \in \mathbb{R}^d} \Big\{ f(\boldsymbol{u}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \Big\}.$$

The minimizer in this definition defines the *proximal operator*:

$$\mathrm{prox}_{\lambda f}(\boldsymbol{x}) = \arg\min_{\boldsymbol{u} \in \mathbb{R}^d} \Big\{ f(\boldsymbol{u}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{u}\|_2^2 \Big\}.$$

When f is ρ-weakly convex and $\lambda < 1/\rho$, the objective in the minimization above is strongly convex, ensuring the proximal operator is single-valued and well-defined.

Below, we summarize basic facts about the Moreau envelope and associated proximal operator, which can either be found in [11] or easily derived from standard results [4, Chapter 6].

The Moreau envelope is a smooth approximation of the original function f. Its gradient is

$$\nabla f_\lambda(\boldsymbol{x}) = \frac{1}{\lambda}(\boldsymbol{x} - \mathrm{prox}_{\lambda f}(\boldsymbol{x})) \tag{2.1}$$

and serves as a natural measure of stationarity, as formalized in the following fact.

Fact 2.1 (Moreau Gradient as a Stationarity Measure)**.** *If f is a proper, lower semicontinuous function and $\lambda > 0$ (such that $\mathrm{prox}_{\lambda f}(\boldsymbol{x})$ is well-defined), then a point $\boldsymbol{x}$ with a small Moreau gradient norm $\|\nabla f_\lambda(\boldsymbol{x})\|_2$ is close to a point $\hat{\boldsymbol{x}} := \mathrm{prox}_{\lambda f}(\boldsymbol{x})$ that is nearly stationary for the original problem. Specifically, we have*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = \lambda\|\nabla f_\lambda(\boldsymbol{x})\|_2 \quad \textit{and} \quad \mathrm{dist}(0, \partial f(\hat{\boldsymbol{x}})) \leq \|\nabla f_\lambda(\boldsymbol{x})\|_2.$$

It is a standard fact that the gradient of Moreau envelope is smooth.

Fact 2.2 (Smoothness of the Moreau Envelope)**.** *The gradient of the Moreau envelope $\nabla f_\lambda(\boldsymbol{x}) = \frac{1}{\lambda}(\boldsymbol{x} - \mathrm{prox}_{\lambda f}(\boldsymbol{x}))$ is $\frac{1}{\lambda(1-\lambda\rho)}$-Lipschitz continuous whenever $\lambda\rho \in (0,1)$.*

3. A Framework for Outlier-Robust Nonsmooth Optimization

In this section, we present a general framework for outlier-robust nonsmooth optimization. We argue that the existing algorithms for robust-mean estimation are compatible with standard projected gradient descent, under a suitable choice of step sizes. The oracle producing robust estimates of the mean can be used to obtain robust subgradient estimates. From the optimization perspective, this constitutes an inexact subgradient oracle with bounded but possibly adversarial error. We show how the analysis of subgradient descent can be adapted to deal with such error, leading to generic guarantees for minimizing nonsmooth (weakly) convex loss functions in the presence of outliers.

3.1. **Robust Mean Estimation.**

In this subsection we review standard results for robust mean estimation, based on material from the book [17]. For simplicity, we state the results for the basic filtering algorithm (Algorithm 1), which runs in polynomial time (in the input size). We note however that there exist algorithms with the same error guarantee and near-linear runtime in the sample size; see [19]. Any such algorithm can be used in place of Algorithm 1; we omit these variants to avoid introducing excessive notation.

To address the challenge of estimating gradients from corrupted data, we adopt the framework of robust statistics. We assume that the algorithm operates under the strong contamination model (Definition 1.1). Note here that algorithms that succeed in this model *must* rely on some structural property of the inlier distribution (such as being "clean" or "noiseless"), as it is impossible to distinguish inliers from outliers absent such properties. A useful property of this type is *stability*, which ensures that the empirical mean and covariance of the inlier data are well-behaved even when a small fraction of points are removed.

Definition 3.1 (Stability, [17, Definition 2.1])**.** Fix $0 < \varepsilon < 1/2$ and $\omega > \varepsilon$. A finite set $S \subset \mathbb{R}^k$ is (ε, ω)-stable with respect to mean $\mu \in \mathbb{R}^k$ and variance proxy σ^2 if for every subset $S' \subset S$ with $|S'| \geq (1-\varepsilon)|S|$, the following holds: (i) $\|\mu_{S'} - \mu\| \leq \sigma\omega$, and (ii) $\|\bar{\Sigma}_{S'} - \sigma^2 I\|_{\mathrm{op}} \leq \sigma^2\omega^2\varepsilon$, where $\mu_{S'}$ and $\bar{\Sigma}_{S'}$ denote the empirical mean and empirical covariance of the set S', respectively.

Crucially, stability of a sample set can be guaranteed with high probability if the sample size is sufficiently large and the underlying distribution has *bounded covariance*—a mild condition implied by standard assumptions in the literature on stochastic optimization (see, e.g., [8, 35]). This guarantee is formalized in the following fact, which connects the properties of the data-generating distribution to the stability of the empirical samples drawn from it.

Fact 3.1 (Sample Complexity for Stability [17, Proposition 3.9] and [18, Theorem 1.4])**.** *Fix $\tau \in (0,1)$. Let S be a set of N independent samples from a distribution on $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. With probability at least $1-\tau$, the subset $S' = \{\boldsymbol{x} \in S : \|\boldsymbol{x} - \boldsymbol{\mu}_S\|_2 \leq 2\sqrt{\|\boldsymbol{\Sigma}\|_{\mathrm{op}}}\sqrt{d/\varepsilon}\}$ satisfies $|S'| \geq (1-\varepsilon)|S|$ and S' is $\big(\varepsilon, O(\sqrt{\varepsilon})\big)$-stable with respect to $\boldsymbol{\mu}$ and $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$, provided the sample size $N = O\big((d\log d + \log(1/\tau))/\varepsilon\big)$ is sufficiently large.*

Given an ε-corrupted stable set, one can estimate the mean of the original data. Algorithm 1 is a deterministic robust mean estimation algorithm with the error guarantee in Fact 3.2 that does not require knowledge of the stability parameters σ or ω. State-of-the-art methods achieve this task in near-linear time, with only logarithmically many passes over the data [9, 19, 20].

Fact 3.2 (Robust Mean Estimation with Stability [18, Theorem A.3])**.** *Let $T \subset \mathbb{R}^k$ be an ε-corrupted version of a set S, where S is $(O(\varepsilon), \omega)$-stable with respect to its mean $\boldsymbol{\mu}_S$ and a variance proxy σ^2. Algorithm 1, on input ε and T, deterministically returns a vector $\hat{\boldsymbol{\mu}}$ such that $\|\boldsymbol{\mu}_S - \hat{\boldsymbol{\mu}}\|_2 = O(\sigma\omega)$, in time polynomial in the sample size.*

An immediate corollary of the previous two facts is that filtering algorithms can be used to obtain $O(\sigma\omega)$ subgradient estimates, with high probability.

Algorithm 1: RobustMeanEstimation(T,ε) with unknown covariance bound

```
Input: $0<\varepsilon<1/2$ and $T\subset\mathbb{R}^d$ is an $\varepsilon$-corrupted set of points
Output: $\hat{\boldsymbol{\mu}}$ approximating the mean of the uncorrupted points
Initialize a weight function $q : T\to\mathbb{R}_+$ with $q(\boldsymbol{z}) = 1/|T|$ for all $\boldsymbol{z}\in T$
while $\sum_{z\in T} q(z)\ge 1-2\varepsilon$ do
    $\hat{\boldsymbol{\mu}} \leftarrow \sum_{z\in T} q(\boldsymbol{z})\boldsymbol{z}/\sum_{z\in T} q(\boldsymbol{z})$
    $\hat{\boldsymbol{\Sigma}} \leftarrow \sum_{z\in T} q(\boldsymbol{z})(\boldsymbol{z}-\hat{\boldsymbol{\mu}})(\boldsymbol{z}-\hat{\boldsymbol{\mu}})^\top/\sum_{z\in T} q(\boldsymbol{z})$
    Compute the top eigenvector $\boldsymbol{v}$ of $\hat{\boldsymbol{\Sigma}}$
    $h(\boldsymbol{z}) := |\boldsymbol{v}^\top(\boldsymbol{z}-\hat{\boldsymbol{\mu}})|^2$
    Find the largest threshold $t>0$ such that $\sum_{z\in T:h(z)\ge t} q(\boldsymbol{z})\ge\varepsilon$
    $f(\boldsymbol{z}) := h(\boldsymbol{z})\mathbb{I}\{h(\boldsymbol{z})\ge t\}$
    $q(\boldsymbol{z}) \leftarrow q(\boldsymbol{z})\left(1-\frac{f(\boldsymbol{z})}{\max_{z'\in T:q(z')\neq 0} f(\boldsymbol{z}')}\right)$
return $\hat{\boldsymbol{\mu}}$
```

Corollary 3.1 (Robust Subgradient Estimation)**.** *Suppose an optimization algorithm runs for K iterations. We assume that for any point $\boldsymbol{x}\in\mathscr{C}$, there is a consistent choice of subgradient $\boldsymbol{g}(\boldsymbol{x};\cdot)$ such that the distribution of $\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi})$ over the clean distribution $\mathbb{P}$ has mean $\bar{\boldsymbol{g}}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}[\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi})]$ and a uniformly bounded covariance, i.e., $\|\mathrm{Cov}_{\boldsymbol{\xi}\sim\mathbb{P}}(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi}))\|_{\mathrm{op}}\le\sigma^2$. Let $\zeta\in(0,1)$ be a desired failure probability. For some sample size $N = O((dK\log d + K\log(K/\zeta))/\varepsilon)$, with probability at least $1-\zeta$, the following holds for all $k\in\{1,\dots,K\}$: the robust subgradient oracle, when run on the k-th batch on the point $\boldsymbol{x}_k$, returns an estimate $\hat{\boldsymbol{g}}_k$ satisfying $\|\hat{\boldsymbol{g}}_k-\bar{\boldsymbol{g}}(\boldsymbol{x}_k)\|_2 = O(\sigma\sqrt{\varepsilon})$.*

Proof. The proof relies on a data-splitting argument, which partitions a single large dataset to restore the statistical independence required for a straightforward analysis. We begin with a dataset of size N and partition it into K disjoint mini-batches, one for each iteration, with each mini-batch having size N/K.

For each iteration k, the iterate $\boldsymbol{x}_k$ is determined by the samples from the first $k-1$ mini-batches. We analyze the process sequentially by conditioning on the history up to iterate k. Since the k-th mini-batch is disjoint from and thus independent of the previous batches, the set of clean subgradients computed using $\boldsymbol{x}_k$ and the k-th mini-batch, denoted S_k, constitutes a set of (conditional) i.i.d. random vectors. This observation allows us to apply Fact 3.1.

By Fact 3.1, if we choose the size of the mini-batch to be $N/K = O((d\log d + \log(K/\zeta))/\varepsilon)$, the clean set of subgradients S_k is guaranteed to be $(O(\varepsilon), O(\sqrt{\varepsilon}))$-stable with a conditional probability of at least $1-\zeta/K$. Since this bound holds for history up to iterate k for any given k, the marginal probability of failure at step k is at most ζ/K for all k. Conditioned on the event that all applications of Fact 3.1 succeed, we can invoke the guarantee of our robust mean estimator. The corrupted subgradients from the k-th mini-batch are given to Algorithm 1, and by Fact 3.2, it produces an estimate $\hat{\boldsymbol{g}}_k$ satisfying $\|\hat{\boldsymbol{g}}_k-\bar{\boldsymbol{g}}(\boldsymbol{x}_k)\|_2 = O(\sigma\sqrt{\varepsilon})$.

To ensure this guarantee holds simultaneously for all K iterations, we use a union bound. The probability that at least one of the K applications of Fact 3.1 fail is bounded by the sum of the marginal probabilities of failure, which is at most $K(\zeta/K)=\zeta$. This implies a total sample size of $N = O((dK\log d + K\log(K/\zeta))/\varepsilon)$, as stated. □

3.2. **Subgradient Descent and Optimization Setup.** Observe that we can equivalently state (1.1) as

$$\min_{\boldsymbol{x}\in\mathscr{C}} f(\boldsymbol{x}) \equiv \min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + \iota_{\mathscr{C}}(\boldsymbol{x}) =: \bar{f}(\boldsymbol{x}), \tag{3.1}$$

where f is L-Lipschitz and ρ-weakly convex, and $\iota_{\mathscr{C}}$ is the indicator function for a closed convex set $\mathscr{C} \subset \mathbb{R}^d$. Algorithm 2 shows the projected subgradient method with an inexact subgradient oracle $\tilde{\boldsymbol{g}}_k$. We state the algorithm assuming the objective is "regularized," meaning that we are minimizing the original objective plus a quadratic term $\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2$, for a fixed vector $\boldsymbol{z}$. However, since we allow $\mu = 0$, the algorithm applies to the original objective in (3.1). The rationale for considering the $\mu > 0$ case will become clear when we discuss the weakly convex objectives in Section 3.4. Note further that when $\mu = 0$, Algorithm 2 can be stated as using standard subgradient descent updates $\boldsymbol{x}_{k+1} = \mathrm{Proj}_{\mathscr{C}}(\boldsymbol{x}_k - \eta_k\tilde{\boldsymbol{g}}_k)$ with step size $\eta_k = \beta_k/c_{k+1}$.

Algorithm 2: Projected Subgradient Method with Inexact Subgradient Oracle

```
Input: Initial point $\boldsymbol{x}_0 \in \mathscr{C}$, number of iterations $K$, parameters $\{\beta_k, c_{k+1}\}_{k=0}^{K-1}$, $\mu \geq 0$,
       $\boldsymbol{z} \in \mathscr{C}$ (if $\mu > 0$)
for $k = 0, 1, \ldots, K-1$ do
    Compute an inexact subgradient $\tilde{\boldsymbol{g}}_k$ s.t. $\|\boldsymbol{g}_k - \tilde{\boldsymbol{g}}_k\|_2 \leq \delta$ for some $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$;
    Update the iterate: $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{y}\in\mathscr{C}} \left\{\beta_k\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{y}\rangle + \frac{\beta_k\mu}{2}\|\boldsymbol{y}-\boldsymbol{z}\|_2^2 + \frac{c_{k+1}}{2}\|\boldsymbol{y}-\boldsymbol{x}_k\|_2^2\right\}$;
Output: The weighted average iterate $\bar{\boldsymbol{x}}_K = \frac{1}{B_K}\sum_{k=0}^{K-1}\beta_k\boldsymbol{x}_k$, where $B_K = \sum_{k=0}^{K-1}\beta_k$, or a
        *certified* iterate (see Corollary 3.4).
```

3.3. **Nonsmooth Convex Functions: Function Value Suboptimality and Stationarity Guarantee.** Consider first the case of a nonsmooth, L-Lipschitz, convex objective function f. Below, we provide a generic analysis of subgradient descent assuming access to an inexact subgradient oracle $\tilde{\boldsymbol{g}}_k$ with bounded error $\|\boldsymbol{g}_k - \tilde{\boldsymbol{g}}_k\|_2 \leq \delta$, where $\boldsymbol{g}_k$ is a subgradient of f at an iterate $\boldsymbol{x}_k$ for $k \geq 0$. We let $\boldsymbol{x}^\star$ denote a(ny) minimizer of f over a closed convex set $\mathscr{C}$. At the end of the subsection, we discuss the implications on outlier-robust minimization.

To set up for applying our results to weakly convex functions with a meaningful exit condition, we consider minimizing the regularized function

$$f_{\boldsymbol{z}}(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2, \tag{3.2}$$

where f is L-Lipschitz and convex, $\mu \geq 0$, and $\boldsymbol{z}$ is fixed. The result applies to arbitrary convex L-Lipschitz functions f (by setting $\mu = 0$), but the introduction of the quadratic term will allow us to utilize this result also to approximate the Moreau envelope of the objective.

The analysis in this section is for a variant of the projected subgradient method, presented in Algorithm 2, which incorporates an extra sequence of positive parameters $\{c_{k+1}\}_{k=0}^{K-1}$ and returns a weighted average of the iterates, the latter being a standard technique to ensure convergence in function value for nonsmooth problems.

Considering convergence of convex optimization methods under an inexact subgradient oracle, and, in particular, of subgradient descent, is not new to optimization literature. The well-known approach of Devolder, Glineur, and Nesterov [13] can handle such problems under the assumption that the inexact gradient oracle $\tilde{\boldsymbol{g}}_{\boldsymbol{x}}$ satisfies, for some $\delta_g \geq 0$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle\tilde{\boldsymbol{g}}_{\boldsymbol{x}}, \boldsymbol{y}-\boldsymbol{x}\rangle - \delta_g, \tag{3.3}$$

Since the analysis in [13] requires an application of (3.3) at least for $\boldsymbol{y} = \boldsymbol{x}^\star$ and $\boldsymbol{x} = \boldsymbol{x}_k$ for iterates $\boldsymbol{x}_k$, it follows that under our assumption on inexactness, it would be required that $\delta_g \geq \delta \max_{0 \leq k \leq K-1} \|\boldsymbol{x}^\star - \boldsymbol{x}_k\|_2$, which is not a priori guaranteed to be bounded. As a consequence, this approach would lead to the final error scaling with $\delta_g \geq \delta \max_{0 \leq k \leq K-1} \|\boldsymbol{x}^\star - \boldsymbol{x}_k\|_2$.

The update rule in Algorithm 2 can be interpreted as a projected subgradient step. Specifically, the update is equivalent to $\boldsymbol{x}_{k+1} = \mathrm{Proj}_{\mathscr{C}}(\boldsymbol{x}_k - \eta_k \tilde{\boldsymbol{g}}_k)$ with a step size $\eta_k = \beta_k / c_{k+1}$ when $\mu = 0$. Our analysis leverages the flexibility of having two sequences of parameters, β_k and c_k, to construct a sharp convergence guarantee. Our particular choice of $\beta_k = \beta$ and $c_{k+1} = c_k - \beta_k(\gamma - \mu)$ will cause the step size sequence η_k to *increase* with k when $\mu = 0$, a property that is crucial to our analysis. This property allows us to obtain an error guarantee that scales with the initial distance to the optimum, $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$, in place of $\max_{0 \leq k \leq K-1} \|\boldsymbol{x}^\star - \boldsymbol{x}_k\|_2$. The strategy of employing decreasing regularization $\frac{c_{k+1}}{2}\|\boldsymbol{y} - \boldsymbol{x}_k\|_2^2$ in the iterate update definition originates from the classical dual averaging technique of Nesterov [36], where it was used in a different context (to ensure that subgradients from later, more relevant iterations are not down-weighted).

Our main result in this subsection is summarized below.

Theorem 3.1 (Function Value Suboptimality Guarantee). *Let f be a convex and L-Lipschitz function, $f_{\boldsymbol{z}}(\cdot) = f(\cdot) + \frac{\mu}{2}\|\cdot - \boldsymbol{z}\|_2^2$ (see (3.2)), and let $\boldsymbol{x}^\star$ be any minimizer of $f_{\boldsymbol{z}}$ over the closed convex set $\mathscr{C}$. Let $D_k := \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2$, $F_k := \|\boldsymbol{x}_k - \boldsymbol{z}\|_2$. Consider running Algorithm 2 for K iterations with the following parameter choices:*

- *Constant step coefficient $\beta_k = \beta > 0$ for all $k = 0, \ldots, K-1$.*
- *A free parameter $\gamma > 0$, and coefficients c_k defined by the backward recurrence $c_k = c_{k+1} + \beta(\gamma - \mu)$, for some fixed terminus $c_K > 0$.*

Then the following holds, where $\bar{\boldsymbol{x}}_K = \frac{1}{K}\sum_{k=0}^{K-1} \boldsymbol{x}_k$:

$$\begin{aligned} &2\beta K(f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star)) + (c_K + \beta\mu)D_K^2 \\ &\leq (c_0 + \beta\mu)D_0^2 + \beta\mu F_0^2 + (L+\delta)^2\beta^2 \sum_{k=0}^{K-1} \frac{1}{c_{k+1}} + \beta K \frac{\delta^2}{\gamma}. \end{aligned}$$

Specifically, this scheme leads to the following bounds:

(1) If $\mu = 0$, then with the choices $\gamma = \delta / D_0$, $c_K = 1$, and constant $\beta = \frac{D_0}{(L+\delta)\sqrt{K}}$, we have the following bound on the optimality gap:

$$f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) \leq \frac{D_0(L+\delta)}{\sqrt{K}} + D_0\delta. \tag{3.4}$$

Thus, for any $K \geq (L+\delta)^2/\delta^2$, we have $f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) \leq 2D_0\delta$.

(2) If $\mu > 0$, then setting $\gamma = \mu/2$, $c_k = k\beta\gamma$, we have

$$\begin{aligned} D_K^2 &\leq \frac{2(D_0^2 + F_0^2) + 4(L+\delta)^2(\ln(K)+1)/\mu^2}{K+2} + \frac{4K}{K+2}\frac{\delta^2}{\mu^2}, \text{ and} \\ f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star) &\leq \frac{\mu(D_0^2 + F_0^2) + 2(L+\delta)^2(\ln(K)+1)/\mu}{K} + \frac{\delta^2}{\mu}. \end{aligned} \tag{3.5}$$

Thus, for any K satisfying

$$K \geq \frac{4\mu^2(D_0^2 + F_0^2)}{\delta^2} + \frac{32(L+\delta)^2}{\delta^2} \ln \frac{16(L+\delta)^2}{\delta^2}, \tag{3.6}$$

we have $D_K \leq \sqrt{5}\delta/\mu$ *and* $f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star) \leq 2\delta^2/\mu$.

Proof. Let $\boldsymbol{x}_{k+1}$ be the minimizer of the objective $h_k(\boldsymbol{y}) := \beta\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{y}\rangle + \frac{\beta\mu}{2}\|\boldsymbol{y}-\boldsymbol{z}\|_2^2 + \frac{c_{k+1}}{2}\|\boldsymbol{y}-\boldsymbol{x}_k\|_2^2$ over $\boldsymbol{y} \in \mathscr{C}$. Observe that $h_k(\boldsymbol{y})$ is $(\beta\mu + c_{k+1})$-strongly convex, and since it is minimized by $\boldsymbol{x}_{k+1}$, we have

$$h_k(\boldsymbol{x}^\star) \geq h_k(\boldsymbol{x}_{k+1}) + \frac{\beta\mu + c_{k+1}}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|_2^2.$$

A rearrangement of this formula yields our main recurrence on the iterates:

$$\begin{aligned} &\beta\Big(\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\rangle + \frac{\mu}{2}\big(\|\boldsymbol{x}_{k+1}-\boldsymbol{z}\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{z}\|_2^2\big)\Big) \\ &\leq \frac{c_{k+1}}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 - \frac{c_{k+1}}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|_2^2 - \frac{c_{k+1}+\beta\mu}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2^2. \end{aligned} \tag{3.7}$$

Next, we relate this inequality to the value of f. By the convexity of f and the definition of a subgradient $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$:

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \langle\boldsymbol{g}_k, \boldsymbol{x}_k - \boldsymbol{x}^\star\rangle.$$

Letting $\boldsymbol{e}_k = \tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k$ be the subgradient error, with $\|\boldsymbol{e}_k\|_2 \leq \delta$, we have:

$$\begin{aligned} f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) &\leq \langle\tilde{\boldsymbol{g}}_k - \boldsymbol{e}_k, \boldsymbol{x}_k - \boldsymbol{x}^\star\rangle \\ &= \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle + \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\rangle - \langle\boldsymbol{e}_k, \boldsymbol{x}_k - \boldsymbol{x}^\star\rangle \\ &\leq \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle + \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\rangle + \|\boldsymbol{e}_k\|_2\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2 \\ &\leq \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle + \langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\rangle + \frac{\gamma}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 + \frac{\delta^2}{2\gamma}, \end{aligned}$$

where the last step uses Young's inequality together with the bound $\|\boldsymbol{e}_k\|_2 \leq \delta$. Adding $\frac{\mu}{2}(\|\boldsymbol{x}_k - \boldsymbol{z}\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{z}\|_2^2)$ to both sides, multiplying by β, using the definition (3.2) of $f_{\boldsymbol{z}}$, adding and subtracting the term $(\beta\mu/2)\|\boldsymbol{x}_{k+1} - \boldsymbol{z}\|_2^2$ on the right-hand side, and rearranging, we obtain

$$\begin{aligned} \beta(f_{\boldsymbol{z}}(\boldsymbol{x}_k) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star)) \leq \beta\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle + \beta\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\rangle + \frac{\beta\mu}{2}(\|\boldsymbol{x}_{k+1} - \boldsymbol{z}\|_2^2 - \|\boldsymbol{x}^\star - \boldsymbol{z}\|_2^2) \\ + \frac{\beta\gamma}{2}\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 + \frac{\beta\delta^2}{2\gamma} + \frac{\beta\mu}{2}(\|\boldsymbol{x}_k - \boldsymbol{z}\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{z}\|_2^2). \end{aligned} \tag{3.8}$$

We bound the two inner product terms separately. For the first, Young's inequality gives:

$$2\beta\langle\tilde{\boldsymbol{g}}_k, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle \leq \frac{\beta^2}{c_{k+1}}\|\tilde{\boldsymbol{g}}_k\|_2^2 + c_{k+1}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|_2^2.$$

Since f is L-Lipschitz, we have $\|\boldsymbol{g}_k\|_2 \leq L$ and thus $\|\tilde{\boldsymbol{g}}_k\|_2 \leq \|\boldsymbol{g}_k\|_2 + \|\boldsymbol{e}_k\|_2 \leq L + \delta$. For the second inner product term in (3.8), we use our main recurrence (3.7). By making both substitutions, we obtain

$$\begin{aligned}2\beta(f_{\boldsymbol{z}}(\boldsymbol{x}_k)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star)) \le &\left(\frac{\beta^2(L+\delta)^2}{c_{k+1}}+c_{k+1}\|\boldsymbol{x}_k-\boldsymbol{x}_{k+1}\|_2^2\right)-c_{k+1}\|\boldsymbol{x}_k-\boldsymbol{x}_{k+1}\|_2^2\\ &+c_{k+1}\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|_2^2-(c_{k+1}+\beta\mu)\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^\star\|_2^2\\ &+\beta\gamma\|\boldsymbol{x}_k-\boldsymbol{x}^\star\|_2^2+\frac{\beta\delta^2}{\gamma}\\ &+\beta\mu(\|\boldsymbol{x}_k-\boldsymbol{z}\|_2^2-\|\boldsymbol{x}_{k+1}-\boldsymbol{z}\|_2^2).\end{aligned}$$

The terms involving $\|\boldsymbol{x}_k-\boldsymbol{x}_{k+1}\|_2^2$ cancel out. Recalling that $D_k^2 := \|\boldsymbol{x}_k-\boldsymbol{x}^\star\|_2^2$, $F_k^2 := \|\boldsymbol{x}_k-\boldsymbol{z}\|_2^2$ and collecting terms, we obtain

$$\begin{aligned}2\beta(f_{\boldsymbol{z}}(\boldsymbol{x}_k)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star)) \le &(c_{k+1}+\beta\gamma)D_k^2-(c_{k+1}+\beta\mu)D_{k+1}^2+\beta\mu(F_k^2-F_{k+1}^2)\\ &+\frac{\beta^2(L+\delta)^2}{c_{k+1}}+\frac{\beta\delta^2}{\gamma}.\end{aligned}\tag{3.9}$$

Recall that $c_{k+1}=c_k+\beta(\mu-\gamma)$ (so the D_k^2 terms telescope). Then, summing (3.9) over k between 0 and $K-1$, we obtain:

$$\begin{aligned}&2\beta K(f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star))+(c_K+\beta\mu)D_K^2\\ &\le (c_0+\beta\mu)D_0^2+\beta\mu F_0^2+(L+\delta)^2\beta^2\sum_{k=0}^{K-1}\frac{1}{c_{k+1}}+\beta K\frac{\delta^2}{\gamma},\end{aligned}\tag{3.10}$$

where we have used Jensen's inequality to bound $f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star)\le\frac{1}{K}\sum_{k=0}^{K-1}(f_{\boldsymbol{z}}(\boldsymbol{x}_k)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star))$ and we dropped the non-positive term $-\beta\mu F_K^2$ from the right-hand side.

To complete the proof, it remains to plug in the concrete choices of the parameters from the statement and simplify (3.10), using that both terms on the left-hand side are non-negative.

Case 1: $\mu=0$**.** In this case, the parameter choices are $c_K=1$, and $\gamma=\delta/D_0$. The relation $c_k=c_{k+1}+\beta\gamma$ implies $c_0=c_K+\gamma\beta K=1+\gamma\beta K$. Simplifying (3.10) thus leads to:

$$f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star)\le\frac{(1+\gamma\beta K)D_0^2}{2\beta K}+\frac{\beta(L+\delta)^2}{2K}\sum_{k=0}^{K-1}\frac{1}{c_{k+1}}+\frac{\delta^2}{2\gamma}.$$

Since $c_{k+1}\ge c_K=1$ for $k=0,1,\ldots,K-1$, we have $\sum_{k=0}^{K-1}\frac{1}{c_{k+1}}\le K$. Using this fact and our choice of $\gamma=\delta/D_0$, the bound simplifies to:

$$f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K)-f_{\boldsymbol{z}}(\boldsymbol{x}^\star)\le\frac{D_0^2}{2\beta K}+\frac{\gamma D_0^2}{2}+\frac{\beta(L+\delta)^2}{2}+\frac{\delta^2}{2\gamma}=\frac{D_0^2}{2\beta K}+\frac{\beta(L+\delta)^2}{2}+D_0\delta.$$

The term $D_0\delta$ represents the error floor from the inexact oracle. To minimize the diminishing part, we choose $\beta=\frac{D_0}{(L+\delta)\sqrt{K}}$ to balance the remaining two terms: Substituting this choice of β back, the diminishing part of the error is bounded by:

$$2\cdot\frac{D_0^2}{2\beta K}=\frac{D_0^2}{K}\frac{(L+\delta)\sqrt{K}}{D_0}=\frac{D_0(L+\delta)}{\sqrt{K}}.$$

Combining with the error floor leads to the claimed bound on the optimality gap.

Case 2: $\mu > 0$. In this case, the parameters are set to $\gamma = \mu/2$, $c_k = k\beta\gamma = k\beta\mu/2$. Thus, simplifying (3.10) and assuming $K \geq 3$ leads to:

$$\begin{aligned} D_K^2 &\leq \frac{\beta\mu(D_0^2+F_0^2)+2(L+\delta)^2(\beta/\mu)\sum_{k=1}^K(1/k)+2\beta K\delta^2/\mu}{K\beta\mu/2+\beta\mu} \\ &\leq \frac{2(D_0^2+F_0^2)+4(L+\delta)^2(\ln(K)+1)/\mu^2}{K+2}+\frac{4K}{K+2}\frac{\delta^2}{\mu^2} \\ &\leq \frac{2(D_0^2+F_0^2)}{K}+\frac{8(L+\delta)^2}{\mu^2}\frac{\ln(K)}{K}+4\frac{\delta^2}{\mu^2} \end{aligned}$$

For $K \geq \frac{4\mu^2(D_0^2+F_0^2)}{\delta^2}$, the first term in the right-hand side above is bounded by $\delta^2/(2\mu^2)$. When K satisfies $\ln(K)/K \leq \tau := \frac{\delta^2}{16(L+\delta)^2}$, the second term is also bounded by $\delta^2/(2\mu^2)$. We note that $\tau \in (0, 1/16)$. For such τ a sufficient condition for $\ln(K)/K \leq \tau$ is $K \geq (1/\tau)(\ln(1/\tau)+\ln\ln(1/\tau)+1)$, for which in turn a sufficient condition is $K \geq (2/\tau)\ln(1/\tau)$. We conclude that for K satisfying (3.6), we have that $D_K^2 \leq 5\delta^2/\mu^2$, giving the claimed bound on D_K.

Similarly, for the optimality gap, we have for $K \geq 3$ that:

$$\begin{aligned} f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star) &\leq \frac{\beta\mu(D_0^2+F_0^2)+2(L+\delta)^2(\beta/\mu)\sum_{k=1}^K(1/k)+2\beta K\delta^2/\mu}{\beta K} \\ &\leq \frac{\mu(D_0^2+F_0^2)+4(L+\delta)^2\ln(K)/\mu}{K}+\frac{\delta^2}{\mu}. \end{aligned}$$

For K satisfying (3.6), the first term on the right-hand side is bounded by $\delta^2/(2\mu)$, and the final claim follows. □

Outlier-robust minimization. With Theorem 3.1 on hand, we can discuss implications on outlier-robust optimization of nonsmooth convex functions and their regularized versions by combining with Corollary 3.1, and setting $\delta = \sigma\sqrt{\varepsilon}$.

Corollary 3.2 (Total Sample Complexity for Outlier-Robust Convex Optimization). *Let f be a convex, L-Lipschitz function, and let $\boldsymbol{x}^\star$ be a minimizer of f over a closed convex set $\mathscr{C}$. Assume the iterates of the algorithm remain within a bounded set of radius W. Let $D_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$. Suppose the clean distribution of single-sample subgradient map $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\xi})$ has a uniformly bounded covariance, $\|\mathrm{Cov}_{\boldsymbol{\xi}\sim\mathbb{P}}(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\xi}))\|_{\mathrm{op}} \leq \sigma^2$, for all $\boldsymbol{x} \in \mathscr{C}$.*

Consider running Algorithm 2 where the inexact subgradient at each step is computed by Algorithm 1 on an ε-corrupted batch of N/K samples. To achieve a final suboptimality of $f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) = O(D_0\sigma\sqrt{\varepsilon})$ with probability at least $1-\zeta$, it suffices to set the number of iterations to $K = O(L^2/(\sigma^2\varepsilon))$ and the total sample size to

$$N = O\Big(\frac{dL^2\log d + L^2\log(L/(\zeta\varepsilon\sigma))}{\sigma^2\varepsilon^2}\Big).$$

For the regularized function $f_{\boldsymbol{z}}$ when $\mu > 0$, we similarly run Algorithm 2 on $f_{\boldsymbol{z}}$ and let $F_0 = \|\boldsymbol{x}_0 - \boldsymbol{z}\|_2$. Then there exists an iteration complexity $K = O(\frac{\mu^2(D_0^2+F_0^2)+L^2\log(L/(\sigma\varepsilon))}{\sigma^2\varepsilon})$ and a sample complexity $N = O(\frac{d\mu^2(D_0^2+F_0^2)\log d + L^2\log(L/(\sigma\varepsilon))}{\sigma^2\varepsilon^2}\log(\frac{d\mu(D_0+F_0)+L}{\sigma\zeta\varepsilon}))$ such that with probability at least $1-\zeta$, it holds that $f_{\boldsymbol{z}}(\bar{\boldsymbol{x}}_K) - f_{\boldsymbol{z}}(\boldsymbol{x}^\star) \leq 2\sigma^2\varepsilon/\mu$ and $\|\boldsymbol{x}_K - \boldsymbol{x}^\star\|_2 \leq \sqrt{5}\sigma\sqrt{\varepsilon}/\mu$.

3.4. **Nonsmooth Weakly Convex Functions: Stationarity.** Suppose now that f is L-Lipschitz and ρ-weakly convex. In this case, subgradient descent can be analyzed by tracking the progress (descent) on the Moreau envelope of $\bar{f}$ (defined in (3.1)). This can be done using similar arguments as in [11], while accounting for the error induced by the inexact subgradient oracle. We provide the full analysis below, summarized in Lemma 3.1 and its proof, for completeness.

The core issue of such an analysis is that, on its own, it only guarantees that within a certain number of iterations K there exists an iterate $\boldsymbol{x}_k$, $k \in \{0,1,\dots,K-1\}$, such that $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|$ is small, which is the desired stationarity guarantee, as per the discussion from Section 2. However, it is unclear a priori how to estimate $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|$, and so we run into the issue of how to certify that any given iterate $\boldsymbol{x}_k$ has the desired error guarantee. To resolve this certification issue, we adopt a post-processing approach. We introduce a procedure to approximate the Moreau gradient for each point in the sequence of iterates by computing an estimate of the proximal operator, $\operatorname{prox}_{\lambda \bar{f}}(\boldsymbol{x}_k)$, for each $\boldsymbol{x}_k$. Computing the proximal operator is a strongly convex minimization problem of the form analyzed in Section 3.3. We can therefore use the inexact subgradient method from Algorithm 2 as an inner-loop solver to find a sufficiently accurate estimate. This two-stage procedure allows us to deterministically identify an iterate with a small Moreau gradient norm and provide a high-probability stationarity guarantee.

The lemma below provides a "sufficient decrease" property for the Moreau envelope when subgradient descent method (Algorithm 2) is applied to the original problem, stated in (3.1).

Lemma 3.1 (Approximate Sufficient Decrease)**.** *Consider the problem* (3.1) *where f is L-Lipschitz and ρ-weakly convex and $\mathscr{C}$ is closed and convex. For a parameter $\lambda > 0$ such that $\mu = 1/\lambda > \rho$, the iterates of Algorithm* 2 *with $\mu = 0$ and $c_k = 1$, $\forall k \geq 1$, satisfy*

$$\bar{f}_\lambda(\boldsymbol{x}_{k+1}) - \bar{f}_\lambda(\boldsymbol{x}_k) \leq -\beta_k \frac{\mu-\rho}{2\mu} \|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 + \mu\beta_k^2(L+\delta)^2 + \frac{\mu\beta_k\delta^2}{2(\mu-\rho)}. \tag{3.11}$$

Proof. By definition, $\bar{f}_\lambda(\boldsymbol{x}_{k+1}) = \min_{\boldsymbol{u}\in\mathscr{C}}\{f(\boldsymbol{u}) + \frac{\mu}{2}\|\boldsymbol{u}-\boldsymbol{x}_{k+1}\|_2^2\}$. Since $\hat{\boldsymbol{x}}_k = \operatorname{prox}_{\lambda\bar{f}}(\boldsymbol{x}_k)$ is in $\mathscr{C}$, we can bound $\bar{f}_\lambda(\boldsymbol{x}_{k+1})$ from above by evaluating the expression at $\boldsymbol{u} = \hat{\boldsymbol{x}}_k$:

$$\begin{aligned}\bar{f}_\lambda(\boldsymbol{x}_{k+1}) &\leq f(\hat{\boldsymbol{x}}_k) + \frac{\mu}{2}\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_{k+1}\|_2^2 \\ &= f(\hat{\boldsymbol{x}}_k) + \frac{\mu}{2}\|P_\mathscr{C}(\hat{\boldsymbol{x}}_k) - P_\mathscr{C}(\boldsymbol{x}_k - \beta_k\tilde{\boldsymbol{g}}_k)\|_2^2.\end{aligned}$$

The projection operator $P_\mathscr{C}$ is non-expansive, so by continuing, we have

$$\begin{aligned}\bar{f}_\lambda(\boldsymbol{x}_{k+1}) &\leq f(\hat{\boldsymbol{x}}_k) + \frac{\mu}{2}\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k + \beta_k\tilde{\boldsymbol{g}}_k\|_2^2 \\ &= f(\hat{\boldsymbol{x}}_k) + \frac{\mu}{2}\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_2^2 + \mu\beta_k\langle\tilde{\boldsymbol{g}}_k, \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\rangle + \frac{\mu\beta_k^2\|\tilde{\boldsymbol{g}}_k\|_2^2}{2} \\ &= \bar{f}_\lambda(\boldsymbol{x}_k) + \mu\beta_k\langle\tilde{\boldsymbol{g}}_k, \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\rangle + \frac{\mu\beta_k^2\|\tilde{\boldsymbol{g}}_k\|_2^2}{2}.\end{aligned} \tag{3.12}$$

Next, we use the ρ-weak convexity of f to bound the inner product term in (3.12). For $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$, we have:

$$\begin{aligned}&f(\hat{\boldsymbol{x}}_k) \geq f(\boldsymbol{x}_k) + \langle\boldsymbol{g}_k, \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\rangle - \frac{\rho}{2}\|\hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_2^2 \\ \Leftrightarrow\ &\langle\boldsymbol{g}_k, \hat{\boldsymbol{x}}_k - \boldsymbol{x}_k\rangle \leq f(\hat{\boldsymbol{x}}_k) - f(\boldsymbol{x}_k) + \frac{\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2.\end{aligned} \tag{3.13}$$

Recall that $\boldsymbol{e}_k = \tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k \Leftrightarrow \tilde{\boldsymbol{g}}_k = \boldsymbol{g}_k + \boldsymbol{e}_k$. From (3.12) and (3.13), we have

$$\begin{aligned}&\bar{f}_\lambda(\boldsymbol{x}_{k+1}) - \bar{f}_\lambda(\boldsymbol{x}_k)\\ &\le \mu\beta_k\Big(f(\hat{\boldsymbol{x}}_k) - f(\boldsymbol{x}_k) + \frac{\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2\Big) - \mu\beta_k\langle \boldsymbol{e}_k, \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\rangle + \frac{\mu\beta_k^2\|\tilde{\boldsymbol{g}}_k\|_2^2}{2}.\end{aligned} \tag{3.14}$$

Now we analyze the parenthetical term in the right-hand side of (3.14). By definition, $\bar{f}_\lambda(\boldsymbol{x}_k) = f(\hat{\boldsymbol{x}}_k) + \frac{\mu}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2$, so by substituting in this term, we obtain

$$\begin{aligned} f(\hat{\boldsymbol{x}}_k) - f(\boldsymbol{x}_k) + \frac{\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 &= \Big(\bar{f}_\lambda(\boldsymbol{x}_k) - \frac{\mu}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2\Big) - f(\boldsymbol{x}_k) + \frac{\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2\\ &= \Big(\bar{f}_\lambda(\boldsymbol{x}_k) - f(\boldsymbol{x}_k)\Big) - \frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2. \end{aligned} \tag{3.15}$$

The function $h(\boldsymbol{u}) := f(\boldsymbol{u}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{x}_k\|_2^2$ is $(\mu-\rho)$-strongly convex because f is ρ-weakly convex and $\mu > \rho$. Since $\hat{\boldsymbol{x}}_k$ minimizes $h(\boldsymbol{u})$ over $\mathscr{C}$ and $\boldsymbol{x}_k \in \mathscr{C}$, we have that

$$h(\boldsymbol{x}_k) \ge h(\hat{\boldsymbol{x}}_k) + \frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2.$$

Substituting the definitions of $h(\cdot)$ and $\bar{f}_\lambda(\cdot)$, this is $f(\boldsymbol{x}_k) \ge \bar{f}_\lambda(\boldsymbol{x}_k) + \frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2$, which implies

$$\bar{f}_\lambda(\boldsymbol{x}_k) - f(\boldsymbol{x}_k) \le -\frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2. \tag{3.16}$$

By substituting into (3.15), we have

$$\begin{aligned} f(\hat{\boldsymbol{x}}_k) - f(\boldsymbol{x}_k) + \frac{\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 &\le -\frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 - \frac{\mu-\rho}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2\\ &= -(\mu-\rho)\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2. \end{aligned}$$

By substituting this bound into Equation (3.14), we obtain

$$\begin{aligned} &\bar{f}_\lambda(\boldsymbol{x}_{k+1}) - \bar{f}_\lambda(\boldsymbol{x}_k) \le -\mu\beta_k(\mu-\rho)\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 - \mu\beta_k\langle \boldsymbol{e}_k, \boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\rangle + \frac{\mu\beta_k^2\|\tilde{\boldsymbol{g}}_k\|^2}{2}\\ &\le -\mu\beta_k(\mu-\rho)\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 + \frac{\mu^2\beta_k^2\|\boldsymbol{e}_k\|_2^2}{2\mu\beta_k(\mu-\rho)} + \frac{\mu\beta_k(\mu-\rho)}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 + \mu\beta_k^2(L^2 + \|\boldsymbol{e}_k\|_2^2)\\ &= -\frac{\mu\beta_k(\mu-\rho)}{2}\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 + \frac{\mu\beta_k\|\boldsymbol{e}_k\|_2^2}{2(\mu-\rho)} + \mu\beta_k^2(L^2 + \|\boldsymbol{e}_k\|_2^2), \end{aligned} \tag{3.17}$$

where the second inequality follows from Young's inequality, which states that $\langle a, b\rangle \le \beta\|a\|_2^2/2 + \|b\|_2^2/(2\beta)$ for all $\beta > 0$. Using the identity (2.1) (recalling that $\mu = 1/\lambda$), we have $\|\boldsymbol{x}_k - \hat{\boldsymbol{x}}_k\|_2^2 = \frac{1}{\mu^2}\|\nabla\bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2$. This simplifies (3.17) to

$$\bar{f}_\lambda(\boldsymbol{x}_{k+1}) - \bar{f}_\lambda(\boldsymbol{x}_k) \le -\beta_k\frac{\mu-\rho}{2\mu}\|\nabla\bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 + \mu\beta_k^2(L^2 + \|\boldsymbol{e}_k\|_2^2) + \frac{\mu\beta_k\|\boldsymbol{e}_k\|_2^2}{2(\mu-\rho)}.$$

It remains to recall that $\|\boldsymbol{e}_k\|_2 \le \delta$ and $L^2 + \delta^2 \le (L+\delta)^2$ for positive L and δ. □

It is immediate from Lemma 3.1 that we can obtain a bound on $\min_{0\le k\le K-1}\|\nabla\bar{f}_\lambda(\boldsymbol{x}_k)\|_2$ by summing the inequality from the lemma statement over $k \in \{0, 1, \ldots, K-1\}$. This guarantees the existence of an iterate with a small Moreau gradient norm. However, since we do not have an oracle for evaluating $\|\nabla\bar{f}_\lambda(\boldsymbol{x}_k)\|_2$, it is unclear which iterate to output.

Unlike the approach in [11], which outputs a randomly sampled iterate to obtain an in-expectation guarantee, we aim for a procedure with certifiable output error. The natural approach is to run the algorithm for K iterations and then, in a post-processing step, approximate $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2$ for each k and return the iterate with the smallest approximate norm. The following corollary provides the guarantee on the quality of the best iterate, and the subsequent corollary analyzes the complexity of this full procedure.

Corollary 3.3. *Let f be an L-Lipschitz and ρ-weakly convex function, and let $\bar{f} = f + \iota_{\mathscr{C}}$, where $\mathscr{C}$ is closed and convex. Consider the projected subgradient method (Algorithm 2 with $\mu = 0$ and $c_k = 1$) with a Moreau parameter λ such that $\mu := 1/\lambda > \rho$ and an inexactness parameter δ such that $\|\boldsymbol{e}_k\|_2 \le \delta$, where $\boldsymbol{e}_k = \tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k$, for all iterations k. Then the minimum squared gradient norm among all iterates is bounded by:*

$$\min_{0\le k<K} \|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 \le \frac{2\mu}{(\mu-\rho)B_K}\left(\bar{f}_\lambda(\boldsymbol{x}_0) - \inf_{\boldsymbol{x}} \bar{f}_\lambda(\boldsymbol{x})\right) + \frac{2\mu^2(L+\delta)^2}{(\mu-\rho)B_K}\sum_{k=0}^{K-1}\beta_k^2 + \frac{\mu^2\delta^2}{(\mu-\rho)^2},$$

where $B_K = \sum_{k=0}^{K-1}\beta_k$.

Proof. We begin by rearranging the inequality from Lemma 3.1:

$$\beta_k \frac{\mu-\rho}{2\mu}\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 \le \bar{f}_\lambda(\boldsymbol{x}_k) - \bar{f}_\lambda(\boldsymbol{x}_{k+1}) + \mu\beta_k^2(L+\delta)^2 + \frac{\mu\beta_k\delta^2}{2(\mu-\rho)}.$$

Summing this inequality from $k = 0$ to $K-1$, we obtain

$$\begin{aligned}\frac{\mu-\rho}{2\mu}\sum_{k=0}^{K-1}\beta_k\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 &\le \sum_{k=0}^{K-1}\left(\bar{f}_\lambda(\boldsymbol{x}_k) - \bar{f}_\lambda(\boldsymbol{x}_{k+1})\right) + \mu(L+\delta)^2\sum_{k=0}^{K-1}\beta_k^2 + \frac{\mu\delta^2}{2(\mu-\rho)}\sum_{k=0}^{K-1}\beta_k \\ &\le \bar{f}_\lambda(\boldsymbol{x}_0) - \bar{f}_\lambda(\boldsymbol{x}_K) + \mu(L+\delta)^2\sum_{k=0}^{K-1}\beta_k^2 + \frac{\mu\delta^2}{2(\mu-\rho)}\sum_{k=0}^{K-1}\beta_k.\end{aligned}$$

Let $B_K = \sum_{k=0}^{K-1}\beta_k$. Since $\min_k\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 \le \frac{1}{B_K}\sum_{k=0}^{K-1}\beta_k\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2$ and $\bar{f}_\lambda(\boldsymbol{x}_K) \ge \inf_{\boldsymbol{x}} \bar{f}_\lambda(\boldsymbol{x})$, we can write

$$\frac{\mu-\rho}{2\mu}B_K \min_{0\le k<K}\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2 \le \bar{f}_\lambda(\boldsymbol{x}_0) - \inf \bar{f}_\lambda + \mu(L+\delta)^2\sum_{k=0}^{K-1}\beta_k^2 + \frac{\mu\delta^2 B_K}{2(\mu-\rho)}.$$

Multiplying by $2\mu/((\mu-\rho)B_K)$ yields the claimed result. □

To find the iterate guaranteed by Corollary 3.3, we must approximate the proximal operator to evaluate the Moreau gradient. The following corollary formalizes this procedure and its complexity. In the statement of the corollary, we assume that the proximal operator is computed in each iteration; alternative implementations are available that admit the same upper bound on the total number of iterations without requiring computation of the full sequence of $\boldsymbol{x}_k$ or $\boldsymbol{p}_k$, e.g., by storing the iterates of the outer procedure and computing their proximal oracles in a post-processing stage.

Corollary 3.4. *Let $\Delta_0 \ge \bar{f}_\lambda(\boldsymbol{x}_0) - \inf \bar{f}_\lambda$. Set $\mu = 1/\lambda = \max\{2\rho, 1\}$, the stepsize $\beta = \sqrt{\frac{\Delta_0}{K\mu(L+\delta)^2}}$, and run Algorithm 2 with $c_k \equiv 1$ and $\beta_k \equiv \beta$ for K iterations. Then, for each iterate $\boldsymbol{x}_k$, approximate $\boldsymbol{p}_k = \mathrm{prox}_{\lambda\bar{f}}(\boldsymbol{x}_k)$ by running the inner solver from Algorithm 2 (Part 2 of Theorem 3.1) for*

$K_{\text{inner}} = \tilde{O}((L+\delta)^2/\delta^2)$ iterations to find $\tilde{\boldsymbol{p}}_k$. Let $\tilde{k} = \arg\min_{0\le k<K} \|\boldsymbol{x}_k - \tilde{\boldsymbol{p}}_k\|_2^2$. Then the output $\boldsymbol{x}_{\tilde{k}}$ satisfies

$$\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2^2 = O\Big(\sqrt{\frac{\mu^3\Delta_0(L+\delta)^2}{K(\mu-\rho)^2}} + \frac{\mu^2\delta^2}{(\mu-\rho)^2} + \delta^2\Big).$$

To achieve a final error of $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2^2 = O(\delta^2)$, the total iteration complexity of the inner and outer solvers is $\tilde{O}\Big(\frac{\mu\Delta_0(L+\delta)^4}{\delta^6}\Big)$.

Proof. Let $k^* = \arg\min_{0\le k<K} \|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2^2$. With our choice of β, the bound from Corollary 3.3 becomes

$$\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{k^*})\|_2^2 \le \frac{2\mu\Delta_0}{K(\mu-\rho)} \frac{\sqrt{\mu}(L+\delta)}{\sqrt{\Delta_0/K}} + \frac{2\mu^2(L+\delta)^2}{\sqrt{\mu}(\mu-\rho)(L+\delta)}\sqrt{\frac{\Delta_0}{K}} + \frac{\mu^2\delta^2}{(\mu-\rho)^2}$$
$$\le 4\sqrt{\frac{\Delta_0(L+\delta)^2\mu^3}{K(\mu-\rho)^2}} + \frac{\mu^2\delta^2}{(\mu-\rho)^2} =: E_K^2.$$

For each k, we approximate $\boldsymbol{p}_k = \text{prox}_{\lambda \bar{f}}(\boldsymbol{x}_k)$ by minimizing $g_k(\boldsymbol{u}) = f(\boldsymbol{u}) + \frac{\mu}{2}\|\boldsymbol{u} - \boldsymbol{x}_k\|_2^2$. This is a μ-strongly convex problem. Using Part 2 of Theorem 3.1 with initial point $\boldsymbol{x}_k$ and reference point $\boldsymbol{z} = \boldsymbol{x}_k$ (so $F_0 = 0$), we can find an iterate $\tilde{\boldsymbol{p}}_k$ such that $\|\tilde{\boldsymbol{p}}_k - \boldsymbol{p}_k\|_2^2 \le 5\delta^2/\mu^2$ after K_{inner} iterations. The number of iterations required is $K_{\text{inner}} = \tilde{O}(\frac{\mu^2 D_0^2 + (L+\delta)^2}{\delta^2})$, where $D_0 = \|\boldsymbol{x}_k - \boldsymbol{p}_k\|_2 = \frac{1}{\mu}\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2$. Since f is L-Lipschitz, we have $\|\boldsymbol{g}\|_2 \le L$ for all $\boldsymbol{g} \in \partial f(\boldsymbol{x})$; thus, noting that $\nabla \bar{f}_\lambda(\boldsymbol{x}_k) \in \partial \bar{f}(\boldsymbol{p}_k)$, we have $D_0 \le L/\mu$. This yields $K_{\text{inner}} = \tilde{O}((L+\delta)^2/\delta^2)$.

The error in our approximation of the Moreau gradient norm is

$$\big|\|\nabla \bar{f}_\lambda(\boldsymbol{x}_k)\|_2 - \mu\|\boldsymbol{x}_k - \tilde{\boldsymbol{p}}_k\|_2\big| = \mu\,\big|\|\boldsymbol{x}_k - \boldsymbol{p}_k\|_2 - \|\boldsymbol{x}_k - \tilde{\boldsymbol{p}}_k\|_2\big| \le \mu\|\boldsymbol{p}_k - \tilde{\boldsymbol{p}}_k\|_2 \le \sqrt{5}\delta.$$

Let $\tilde{\delta}_k = \mu\|\boldsymbol{x}_k - \tilde{\boldsymbol{p}}_k\|_2$. By the triangle inequality, $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2 \le \tilde{\delta}_{\tilde{k}} + \sqrt{5}\delta$. Since $\tilde{k}$ is the minimizer of $\tilde{\delta}_k$, we have $\tilde{\delta}_{\tilde{k}} \le \tilde{\delta}_{k^*} \le \|\nabla \bar{f}_\lambda(\boldsymbol{x}_{k^*})\|_2 + \sqrt{5}\delta \le E_K + \sqrt{5}\delta$. Therefore, $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2 \le E_K + 2\sqrt{5}\delta$, which implies

$$\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2^2 = O(E_K^2 + \delta^2) = O\Big(\sqrt{\frac{\mu^3\Delta_0(L+\delta)^2}{K(\mu-\rho)^2}} + \frac{\mu^2\delta^2}{(\mu-\rho)^2} + \delta^2\Big).$$

To achieve a final squared norm of $O(\delta^2)$, we must choose the outer iteration count K such that $\sqrt{\frac{\mu^3\Delta_0(L+\delta)^2}{K(\mu-\rho)^2}} = O(\delta^2)$. This requires K to be of the order of $\frac{\mu\Delta_0(L+\delta)^2}{\delta^4}$. The total complexity is $K \times K_{\text{inner}} = \tilde{O}(\frac{\mu\Delta_0(L+\delta)^4}{\delta^6})$. □

Outlier-robust optimization. We now translate the stationarity guarantee from Corollary 3.4 into the statistical setting of outlier-robust optimization. By combining our algorithmic analysis with the guarantees for robust subgradient estimation in Corollary 3.1, we derive the total sample complexity required to find an approximate stationary point with high probability.

Corollary 3.5 (Total Sample Complexity for Outlier-Robust Stationarity)**.** *Let f be an L-Lipschitz and ρ-weakly convex function. Suppose the single-sample subgradient map $\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi})$ has a uniformly bounded covariance under the clean distribution $\mathbb{P}$, i.e., $\|\text{Cov}_{\boldsymbol{\xi}\sim\mathbb{P}}(\boldsymbol{g}(\boldsymbol{x};\boldsymbol{\xi}))\|_{\text{op}} \le \sigma^2$ for all iterates.*

Consider the procedure described in Corollary 3.4, where each inexact subgradient is computed by Algorithm 1 on an ε-corrupted batch of samples. To find an iterate $\boldsymbol{x}_{\tilde{k}}$ that, with probability at least $1-\zeta$, satisfies the stationarity condition $\|\nabla \bar{f}_\lambda(\boldsymbol{x}_{\tilde{k}})\|_2^2 = O(\sigma^2\varepsilon)$, it suffices to use a total sample size of

$$N = \tilde{O}\Big(\frac{d\mu\Delta_0(L^2+\sigma^2\varepsilon)^2}{\sigma^6\varepsilon^4}\Big),$$

where $\Delta_0 = \bar{f}_\lambda(\boldsymbol{x}_0) - \inf \bar{f}_\lambda$, $\mu = 1/\lambda = \max\{2\rho, 1\}$, and $\tilde{O}$ additionally hides logarithmic factors in ζ.

4. Distributionally Robust Optimization with Outliers

We now describe the applications of the approaches of the previous section to two problems in distributionally robust optimization (DRO). The goal of DRO is to minimize the worst-case expected loss over an ambiguity set $\mathscr{U}$ of distributions, i.e., to solve [34, 44]

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \max_{\mathbb{Q}\in\mathscr{U}} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{w},\gamma)]. \tag{DRO}$$

We consider ambiguity sets defined by an f-divergence constraint, containing all distributions $\mathbb{Q} \ll \mathbb{P}$ that satisfy $D_\phi(\mathbb{Q}\|\mathbb{P}) \le \rho$, where

$$D_\phi(\mathbb{Q}\|\mathbb{P}) := \int \phi\Big(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\Big)\mathrm{d}\mathbb{P}.$$

We focus on the **Cressie-Read family** of f-divergences [10], parameterized by $m \in (1,\infty)$, with

$$\phi_m(t) = \frac{t^m - mt + m - 1}{m(m-1)}.$$

Here, we rely upon known duality results that translate the min-max DRO problem into *equivalent* primal-only problems. Specifically, the DRO problem based on the Cressie-Read family of f-divergences has the following convenient and tractable formulation [21, Lemma 1]:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d,\gamma\in\mathbb{R}} f_m(\boldsymbol{w},\gamma) := s_m(\rho)(\mathbb{E}[(\ell(\boldsymbol{w},\boldsymbol{\xi})-\gamma)_+^{m_*}])^{1/m_*} + \gamma, \tag{4.1}$$

where $m_* = m/(m-1)$ is the conjugate exponent and $s_m(\rho) = (1+m(m-1)\rho)^{1/m}$ is a constant. This family includes, or asymptotically approaches, several important special cases:

- $m \to 1$: The Kullback-Leibler (KL) divergence.
- $m = 2$: The χ^2-divergence.
- $m \to \infty$: A translated version of Conditional Value-at-Risk (CVaR), discussed below.

Our goal is to solve (4.1) in the presence of outliers. To apply the results from Section 3, we need to argue that: (1) the objective is (weakly) convex and Lipschitz-continuous, (2) we can compute unbiased stochastic estimates of the subgradient of $f_m(\boldsymbol{w},\gamma)$ under the "clean" data distribution, and (3) the covariance of the subgradients is bounded, so robust mean estimation framework (Section 3.1) applies. Most of the technical work is carried out to obtain part (3). The bound on the estimation error then follows from results presented in Section 3.

4.1. CVaR Minimization.

We begin with Conditional Value-at-Risk (CVaR). We first discuss why it can be understood as a limiting case of the Cressie-Read family.

On the one hand, as $m \to \infty$, $\phi_m(t) \to \chi_{[0,1]}$, where we define the characteristic function $\chi_A(a) = 0$ if $a \in A$ and $\chi_A(a) = +\infty$ if $a \notin A$. On the other hand, given a loss function $\ell(\boldsymbol{w}, \xi)$, the CVaR objective is defined to be

$$\mathrm{CVaR}^{\alpha}_{\boldsymbol{\xi}\sim\mathbb{P}} := \inf_{\gamma\in\mathbb{R}} \gamma + \frac{1}{\alpha}\mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}[(\ell(\boldsymbol{w},\boldsymbol{\xi}) - \gamma)_+] \tag{4.2}$$

which equals the f-divergence DRO objective $\max_{\mathbb{Q}\in\mathscr{U}(\mathbb{P})} \mathbb{E}_{\xi\in\mathbb{Q}}[\ell(\boldsymbol{w}, \xi)]$ for $\mathscr{U}(\mathbb{P}) = \{\mathbb{Q} \ll \mathbb{P} : D_{\phi_{+\infty,\alpha}}(\mathbb{Q}\|\mathbb{P}) \leq 1\}$ if we define $\phi_{+\infty,\alpha} = \chi_{[0,1/\alpha]}$ [21, Example 3].

To minimize (4.2) with respect to $\boldsymbol{w}$, we can solve the joint minimization problem with the pair $\boldsymbol{x} := (\gamma, \boldsymbol{w}) \in \mathbb{R}^{d+1}$ as the optimization variable. We assume that the loss function $\ell(\boldsymbol{w}, \boldsymbol{\xi})$ is convex and differentiable in $\boldsymbol{w}$ for all $\boldsymbol{\xi}$ almost surely[1]. The objective in (4.2) is then jointly convex in $(\gamma, \boldsymbol{w})$ but is nonsmooth due to the positive part function.

For completeness, the following lemma provides the characterization of the subdifferential for the single-sample objective.

Lemma 4.1 (Subgradient of the CVaR Objective)**.** *Let the single-sample CVaR objective for the sample $\boldsymbol{\xi}$ be*

$$f(\gamma, \boldsymbol{w}, \boldsymbol{\xi}) = \gamma + \frac{1}{\alpha}(\ell(\boldsymbol{w},\boldsymbol{\xi}) - \gamma)_+ .$$

All elements $\tilde{\boldsymbol{g}} = (\tilde{g}_\gamma, \tilde{\boldsymbol{g}}_{\boldsymbol{w}}) \in \partial f(\gamma, \boldsymbol{w}, \boldsymbol{\xi})$ satisfy

$$\tilde{g}_\gamma = 1 - \beta/\alpha, \quad \tilde{\boldsymbol{g}}_{\boldsymbol{w}} = \frac{\beta}{\alpha}\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi}), \text{ where } \beta \begin{cases} = 1 & \text{if } \ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma \\ = 0 & \text{if } \ell(\boldsymbol{w},\boldsymbol{\xi}) < \gamma \\ \in [0,1] & \text{if } \ell(\boldsymbol{w},\boldsymbol{\xi}) = \gamma \end{cases} .$$

Proof. The gradient of γ with respect to $(\gamma, \boldsymbol{w})$ is $(1, \boldsymbol{0})$. For the second term, we apply the chain rule for subgradients to the term $\frac{1}{\alpha}(\ell(\boldsymbol{w},\boldsymbol{\xi}) - \gamma)_+$. The subgradient of the outer function $(\cdot)_+$ is $\beta \in [0,1]$ as defined in the lemma statement. The gradient of the inner function $\ell(\boldsymbol{w},\boldsymbol{\xi}) - \gamma$ with respect to $(\gamma, \boldsymbol{w})$ is $(-1, \nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi}))$. Combining these via the chain rule and summing over all samples yields the expression for the subgradient of the second term as $(-\frac{1}{\alpha}\beta, \frac{\beta}{\alpha}\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi}))$. Adding the gradient of the first term γ gives the stated result. □

A key step in our analysis is establishing the following bound on the subgradient covariance.

Proposition 4.1 (Subgradient Covariance Bound for CVaR)**.** *Consider the subgradient $\tilde{\boldsymbol{g}}$ from Lemma 4.1 where we make the consistent choice $\beta = \mathbb{I}(\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma)$. The operator norm of the single-sample covariance of the subgradient,* $\mathrm{Cov}(\tilde{\boldsymbol{g}})$*, is bounded by:*

$$\|\mathrm{Cov}(\tilde{\boldsymbol{g}})\|_{\mathrm{op}} \leq \frac{1}{\alpha^2}\big((2-p)\|\mathbb{E}[\nabla_{\boldsymbol{w}}\ell(\nabla_{\boldsymbol{w}}\ell)^\top]\|_{\mathrm{op}} + p(1-p)\big),$$

where $p = \mathbb{P}(\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma)$.

[1] The differentiability assumption can be easily relaxed under our framework by replacing $\nabla_{\boldsymbol{w}}\ell$ by a consistent choice of subgradients.

Proof. Let $\boldsymbol{v} := \nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi})$ and $\beta := \mathbb{I}(\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma)$. The components of the subgradient are $\tilde{g}_\gamma = 1-\beta/\alpha$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{w}} = \beta\boldsymbol{v}/\alpha$. The covariance of the subgradient $\tilde{\boldsymbol{g}}$ is $\operatorname{Cov}(\tilde{\boldsymbol{g}}) = \frac{1}{\alpha^2}\operatorname{Cov}(\boldsymbol{z})$, where $\boldsymbol{z}$ is the single-sample composite vector $\boldsymbol{z} := (\beta\boldsymbol{v}, -\beta)$. We proceed to bound the operator norm of $\operatorname{Cov}(\boldsymbol{z})$.

Let $p := \mathbb{P}(\beta = 1)$, $\boldsymbol{\mu}_c := \mathbb{E}[\boldsymbol{v}|\beta = 1]$, and $\boldsymbol{\Sigma}_c := \operatorname{Cov}(\boldsymbol{v}|\beta = 1)$. Because the covariance matrix of $\boldsymbol{z}$ is given by $\operatorname{Cov}(z) = \mathbb{E}[zz^\top] - \mathbb{E}[z]\,\mathbb{E}[z]^\top$, a direct calculation shows that

$$\operatorname{Cov}(\boldsymbol{z}) = \begin{pmatrix} p\boldsymbol{\Sigma}_c + p(1-p)\boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top & -p(1-p)\boldsymbol{\mu}_c \\ -p(1-p)\boldsymbol{\mu}_c^\top & p(1-p) \end{pmatrix}.$$

To bound the norm of this matrix, we decompose it as $\operatorname{Cov}(\boldsymbol{z}) = \boldsymbol{C}_1 + \boldsymbol{C}_2$, where

$$\boldsymbol{C}_1 := \begin{pmatrix} p\boldsymbol{\Sigma}_c & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{pmatrix}, \qquad \boldsymbol{C}_2 := p(1-p)\begin{pmatrix} \boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top & -\boldsymbol{\mu}_c \\ -\boldsymbol{\mu}_c^\top & 1 \end{pmatrix}.$$

By the triangle inequality for operator norms, $\|\operatorname{Cov}(\boldsymbol{z})\|_{\text{op}} \le \|\boldsymbol{C}_1\|_{\text{op}} + \|\boldsymbol{C}_2\|_{\text{op}}$. The norm of the first part is $\|\boldsymbol{C}_1\|_{\text{op}} = p\|\boldsymbol{\Sigma}_c\|_{\text{op}}$. The matrix in $\boldsymbol{C}_2$ is a rank-one outer product of the vector $\boldsymbol{u} := (\boldsymbol{\mu}_c, -1)$, so its norm is $\|\boldsymbol{C}_2\|_{\text{op}} = p(1-p)\|\boldsymbol{u}\|_2^2 = p(1-p)(\|\boldsymbol{\mu}_c\|_2^2 + 1)$. This gives the bound:

$$\|\operatorname{Cov}(\boldsymbol{z})\|_{\text{op}} \le p\|\boldsymbol{\Sigma}_c\|_{\text{op}} + p(1-p)(\|\boldsymbol{\mu}_c\|_2^2 + 1). \tag{4.3}$$

We now bound the conditional quantities in Equation (4.3) using the unconditional second moment matrix $\boldsymbol{S} := \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top]$. Let $\boldsymbol{S}_c := \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top|\beta = 1]$ be the conditional second moment matrix. Since $\boldsymbol{\Sigma}_c = \boldsymbol{S}_c - \boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top$, both $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\mu}_c\boldsymbol{\mu}_c^\top$ are positive semidefinite and subordinate to $\boldsymbol{S}_c$ in the Loewner order. This implies their operator norms are bounded by that of $\boldsymbol{S}_c$:

$$\|\boldsymbol{\Sigma}_c\|_{\text{op}} \le \|\boldsymbol{S}_c\|_{\text{op}} \quad \text{and} \quad \|\boldsymbol{\mu}_c\|_2^2 \le \|\boldsymbol{S}_c\|_{\text{op}}.$$

Substituting these into Equation (4.3) yields

$$\begin{aligned} \|\operatorname{Cov}(\boldsymbol{z})\|_{\text{op}} &\le p\|\boldsymbol{S}_c\|_{\text{op}} + p(1-p)(\|\boldsymbol{S}_c\|_{\text{op}} + 1) \\ &= (2p - p^2)\|\boldsymbol{S}_c\|_{\text{op}} + p(1-p). \end{aligned}$$

Finally, the law of total expectation implies $p\boldsymbol{S}_c \preceq \boldsymbol{S}$, which gives the norm inequality $\|\boldsymbol{S}_c\|_{\text{op}} \le \frac{1}{p}\|\boldsymbol{S}\|_{\text{op}}$. Substituting this provides the final bound on $\|\operatorname{Cov}(\boldsymbol{z})\|_{\text{op}}$:

$$\|\operatorname{Cov}(\boldsymbol{z})\|_{\text{op}} \le (2p - p^2)\frac{1}{p}\|\boldsymbol{S}\|_{\text{op}} + p(1-p) = (2-p)\|\boldsymbol{S}\|_{\text{op}} + p(1-p).$$

Scaling by $1/\alpha^2$ completes the proof. □

Remark 4.1. It is generally not possible to derive a meaningful upper bound that is independent of $\|\mathbb{E}[\nabla_{\boldsymbol{w}}\ell]\|_2^2$. The mean of the loss gradient is fundamentally linked to the behavior of the CVaR subgradient, and removing it from the bound would require much stronger assumptions about the problem. The core of the issue lies in the nature of the subgradient calculation. The term $\beta = \mathbb{I}(\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma)$ acts as a "selection" or "gating" mechanism, which is a non-centered operation. It selects which gradients $\boldsymbol{v}$ will be included in the sum. $\operatorname{Cov}(\boldsymbol{v})$ tells us about the shape and spread of the cloud of possible gradient vectors, centered around their mean, while $\mathbb{E}[\boldsymbol{v}]$ tells us the location of that cloud's center in space. The variance of the subgradient depends on the properties of the *selected* part of the cloud. However, where we "slice" the cloud with the threshold γ is critical. If the entire cloud is far from the threshold (i.e., if $\mathbb{E}[\ell]$ is far from γ), the selection β might be almost always 1 or almost always 0, leading to low variance. If the threshold slices right through the densest part of the cloud, the variance will be high. Without knowing the location of the cloud (given by $\mathbb{E}[\boldsymbol{v}]$, which is related to $\mathbb{E}[\ell]$), we cannot know how the threshold γ interacts with it. The term $\|\mathbb{E}[\boldsymbol{v}]\|_2^2$ in the bound is precisely what accounts for this location information.

We now combine these tools to develop a provably robust optimization algorithm for CVaR minimization.

Theorem 4.1 (Outlier-Robust CVaR)**.** *Let f be the CVaR objective and let $\boldsymbol{x}^\star$ be a minimizer of f over a compact convex set $\mathscr{C} \subset \mathbb{R}^{d+1}$. Suppose that for any $\boldsymbol{x} = (\gamma, \boldsymbol{w})$ such that $\boldsymbol{w}$ is in the projection of $\mathscr{C}$, the loss gradients from the inlier distribution $\mathbb{P}$ have a uniformly bounded second moment matrix, i.e., $\|\mathbb{E}[\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi})\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi})^\top]\|_{\mathrm{op}} \leq G^2$ for all $\boldsymbol{w} \in \mathscr{W}$.*

Consider a procedure based on Algorithm 2, where, for the current iterate $\boldsymbol{x}_k$, the inexact subgradient $\tilde{\boldsymbol{g}}_k$ is computed by forming the ε-corrupted set of subgradients $T_k = \{\boldsymbol{z}(\boldsymbol{x}_k, \boldsymbol{\xi}_i)\}_{i=1}^N$ and then computing $\tilde{\boldsymbol{g}}_k = \mathsf{RobustMeanEstimation}(T_k, \varepsilon)$.

For some sample size $N = O((d\log d + \log(1/(\zeta\varepsilon)))/\varepsilon^2)$, running Algorithm 2 for $K = O(1/\varepsilon)$ iterations produces an output $\bar{\boldsymbol{x}}_K$ that, with probability at least $1-\zeta$, satisfies:

$$f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) = O(G\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2\sqrt{\varepsilon}/\alpha).$$

Proof. The proof strategy is to show that the robust mean estimation procedure provides an inexact subgradient with a controllable error bound, which can then be inserted into the convergence guarantee of Theorem 3.1.

For the population CVaR objective $f(\boldsymbol{x}) = \gamma + \frac{1}{\alpha}\mathbb{E}[(\ell(\boldsymbol{w},\boldsymbol{\xi}) - \gamma)_+]$, where $\boldsymbol{x} = (\gamma, \boldsymbol{w})$, its subgradient $\boldsymbol{g} = (\boldsymbol{g}_\gamma, \boldsymbol{g}_{\boldsymbol{w}}) \in \partial f(\boldsymbol{x})$ has components:

$$\boldsymbol{g}_\gamma = 1 - \mathbb{E}[\beta]/\alpha, \quad \boldsymbol{g}_{\boldsymbol{w}} = \mathbb{E}[\beta\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi})]/\alpha$$

where β is an indicator (random) variable for the event $\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma$. Let $p = \mathbb{E}[\beta] = \mathbb{P}(\ell(\boldsymbol{w},\boldsymbol{\xi}) > \gamma)$. The squared norm of this subgradient is $\|\boldsymbol{g}\|_2^2 = (1-p/\alpha)^2 + \frac{1}{\alpha^2}\|\mathbb{E}[\beta\nabla_{\boldsymbol{w}}\ell]\|_2^2 \leq 1/\alpha^2 + G^2/\alpha^2$. So the Lipschitz constant of f is $L = O(G/\alpha)$. Proposition 4.1 implies that $\sigma = O(G/\alpha)$ is an upper bound on the standard deviation of the single-sample subgradients.

Inexact subgradient oracle error. Let $\boldsymbol{x}_k \in \mathscr{C}$ be the iterate at step k. The true subgradient is $\boldsymbol{g}_k = \mathbb{E}[\boldsymbol{z}(\boldsymbol{x}_k,\boldsymbol{\xi})]$, where $\boldsymbol{z}(\boldsymbol{x}_k,\boldsymbol{\xi})$ is the single-sample subgradient defined in the proof of Proposition 4.1. Our inexact subgradient is $\tilde{\boldsymbol{g}}_k = \mathsf{RobustMeanEstimation}(T_k, \varepsilon)$. The error is $\|\tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k\|_2$.

To bound this error, we first analyze the properties of the clean subgradient distribution. From Proposition 4.1, the covariance of a single-sample subgradient $\boldsymbol{z}$ is bounded by $\|\mathrm{Cov}(\boldsymbol{z})\|_{\mathrm{op}} \leq (2-p)\|\mathbb{E}[\nabla_{\boldsymbol{w}}\ell\nabla_{\boldsymbol{w}}\ell^\top]\|_{\mathrm{op}} + p(1-p)$. Under our assumption, $\|\mathbb{E}[\nabla_{\boldsymbol{w}}\ell\nabla_{\boldsymbol{w}}\ell^\top]\|_{\mathrm{op}} \leq G^2$, and since $p(1-p) \leq 1/4$, we have a uniform bound $\|\mathrm{Cov}(\boldsymbol{z}(\boldsymbol{x},\boldsymbol{\xi}))\|_{\mathrm{op}} \leq 2G^2 + 1/4$ for all $\boldsymbol{x} \in \mathscr{C}$. Thus, we can set $\sigma^2 = O(G^2/\alpha^2)$ as a uniform upper bound on the spectral norm of subgradient covariance. With $N = O((dK\log d + K\log(K/\zeta))/\varepsilon)$, Corollary 3.1 ensures that with probability at least $1-\zeta$, the estimate $\tilde{\boldsymbol{g}}_k$ satisfies $\|\tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k\|_2 = O(\sigma\omega) = O(\sigma\sqrt{\varepsilon})$ for all k.

The optimization guarantee. Recall that the CVaR objective is convex and L-Lipschitz, where $L = O(G/\alpha)$. We have constructed an inexact subgradient oracle with error $\delta = O(\sigma\sqrt{\varepsilon})$. We can now apply Theorem 3.1 with $\mu = 0$, which implies

$$f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 (L+\delta)}{\sqrt{K}} + \delta \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2.$$

Substituting our value for δ and denoting $D_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$, we get

$$f(\bar{\boldsymbol{x}}_K) - f(\boldsymbol{x}^\star) \leq \frac{D_0(L + O(\sigma\sqrt{\varepsilon}))}{\sqrt{K}} + D_0 \cdot O(\sigma\sqrt{\varepsilon}) = O\Big(\frac{D_0 L}{\sqrt{K}} + D_0 \sigma \sqrt{\varepsilon}\Big).$$

Setting $K = 1/\varepsilon$ and substituting in $L = O(G/\alpha)$ and $\sigma = O(G/\alpha)$ completes the proof. □

4.2. **General Cressie-Read f-divergence DRO.** We now turn to the analysis of the general Cressie-Read objective for $m \in (1, \infty)$. For completeness, we first establish the convexity of this generalized objective function, which ensures that the problem is well-posed for the subgradient methods we employ.

Lemma 4.2 (Joint Convexity of the Generalized Objective)**.** *Let the loss function $\ell(\boldsymbol{w}, \boldsymbol{\xi})$ be convex in $\boldsymbol{w}$ for any realization of $\boldsymbol{\xi}$. Then for any $m \in (1, \infty)$, the generalized objective*

$$f_m(\boldsymbol{w}, \gamma) = s_m(\rho) \left(\mathbb{E}\left[(\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+^{m_*}\right]\right)^{1/m_*} + \gamma$$

is jointly convex in $\boldsymbol{x} := (\boldsymbol{w}, \gamma)$.

Proof. Since linearity preserves convexity, the joint convexity of $f_m(\boldsymbol{w}, \gamma)$ depends entirely on the joint convexity of the term $G(\boldsymbol{w}, \gamma) := \left(\mathbb{E}\left[(\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+^{m_*}\right]\right)^{1/m_*}$.

The key to the proof is to recognize that this expression is the L_p-norm of a random variable, where $p = m_*$. Specifically, let $\mathscr{X}$ be the space of random variables defined over the probability space of $\boldsymbol{\xi}$. The function $G(\boldsymbol{w}, \gamma)$ is the L_{m_*}-norm of the random variable $h(\boldsymbol{w}, \gamma; \boldsymbol{\xi}) := (\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+$. The proof proceeds by showing that the mapping from the parameters $(\boldsymbol{w}, \gamma)$ to the random variable $h(\boldsymbol{w}, \gamma; \boldsymbol{\xi})$ is convex, and then composing this with the convex L_{m_*}-norm function.

First, we establish the convexity of the mapping $(\boldsymbol{w}, \gamma) \mapsto h(\boldsymbol{w}, \gamma; \boldsymbol{\xi})$. For any fixed realization of $\boldsymbol{\xi}$, the function h is a composition of the inner function $a(\boldsymbol{w}, \gamma) = \ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma$ and the outer function $b(x) = (x)_+$. The inner function a is jointly convex because it is the sum of a function convex in $\boldsymbol{w}$ and a linear function in γ. The outer function b is convex and non-decreasing. A composition of a jointly convex function with a convex, non-decreasing function is itself jointly convex. Thus, for any pair of points $(\boldsymbol{w}_1, \gamma_1)$ and $(\boldsymbol{w}_2, \gamma_2)$ and any $\lambda \in [0, 1]$, we have the pointwise inequality:

$$h(\lambda \boldsymbol{w}_1 + (1-\lambda)\boldsymbol{w}_2, \lambda\gamma_1 + (1-\lambda)\gamma_2; \boldsymbol{\xi}) \leq \lambda h(\boldsymbol{w}_1, \gamma_1; \boldsymbol{\xi}) + (1-\lambda) h(\boldsymbol{w}_2, \gamma_2; \boldsymbol{\xi}) \quad \text{a.s.}$$

This shows that the mapping from $(\boldsymbol{w}, \gamma)$ to the random variable h is a convex mapping.

Next, we leverage the properties of the L_{m_*}-norm, which is the function that maps a random variable Z to $(\mathbb{E}[|Z|^{m_*}])^{1/m_*}$. Since $m \in (1, \infty)$, we have $m_* = m/(m-1) > 1$, and for such exponents, the L_{m_*}-norm is a convex function. Furthermore, because its argument h is always non-negative, the norm is also a non-decreasing function. The composition of a convex mapping with a convex and non-decreasing function is convex. By applying the L_{m_*}-norm to the inequality above, we find:

$$\begin{aligned} G(\lambda \boldsymbol{w}_1 + (1-\lambda)\boldsymbol{w}_2, \lambda\gamma_1 + (1-\lambda)\gamma_2) &= \|h(\lambda \boldsymbol{w}_1 + (1-\lambda)\boldsymbol{w}_2; \boldsymbol{\xi})\|_{L_{m_*}} \\ &\leq \|\lambda h(\boldsymbol{w}_1, \gamma_1; \boldsymbol{\xi}) + (1-\lambda) h(\boldsymbol{w}_2, \gamma_2; \boldsymbol{\xi})|_{L_{m_*}} \\ &\leq \lambda \|h(\boldsymbol{w}_1, \gamma_1; \boldsymbol{\xi})\|_{L_{m_*}} + (1-\lambda)\|h(\boldsymbol{w}_2, \gamma_2; \boldsymbol{\xi})\|_{L_{m_*}} \\ &= \lambda G(\boldsymbol{w}_1, \gamma_1) + (1-\lambda) G(\boldsymbol{w}_2, \gamma_2). \end{aligned}$$

The first inequality holds because the mapping to h is convex and the norm is non-decreasing. The second inequality is the triangle inequality (i.e., the convexity) of the L_{m_*}-norm itself. This confirms that $G(\boldsymbol{w}, \gamma)$ is jointly convex, and therefore the entire objective $f_m(\boldsymbol{w}, \gamma)$ is jointly convex. □

Next, we derive a bound on the spectral norm of the subgradient covariance for the generalized objective. Our approach extends the CVaR analysis, yielding a bound that depends on the second moment of the loss gradient. Notably, while [31] also provides a covariance bound for the special case of χ^2-divergence ($m = 2$), their result hinges on a technical assumption about the ambiguity set's geometry and bounded loss. It is not clear whether the broader Cressie-Read family we study here satisfies such a condition.

To proceed, we first compute the single-sample subgradient. For the objective $f_m(\boldsymbol{w}, \gamma) = s_m(\rho)\| (\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+ \|_{L_{m_*}} + \gamma$, the chain rule for subgradients implies that a valid single-sample subgradient $\boldsymbol{g}_k(\boldsymbol{x}; \boldsymbol{\xi})$ is:

$$\boldsymbol{g}_k(\boldsymbol{x}; \boldsymbol{\xi}) = \boldsymbol{e}_1 + \frac{s_m(\rho)}{A_m^{m_*-1}} \left(-(\ell-\gamma)_+^{m_*-1}, \quad (\ell-\gamma)_+^{m_*-1} \nabla_{\boldsymbol{w}} \ell \right)^\top$$

where $A_m = (\mathbb{E}[(\ell-\gamma)_+^{m_*}])^{1/m_*}$ is a deterministic scalar for a fixed $(\boldsymbol{w}, \gamma)$, and $\boldsymbol{e}_1$ is the standard basis vector for the γ component. Let us define the random weight $\beta_m = (\ell-\gamma)_+^{m_*-1}$. The single-sample subgradient then has the familiar structure:

$$\boldsymbol{g}_k(\boldsymbol{x}; \boldsymbol{\xi}) = \boldsymbol{e}_1 + s_m(\rho) \left(-\beta_m, \beta_m \nabla_{\boldsymbol{w}} \ell \right)^\top / A_m^{m_*-1}$$

Its covariance is $\operatorname{Cov}(\boldsymbol{g}_k) = s_m(\rho)^2 \cdot \operatorname{Cov}((-\beta_m, \beta_m \nabla_{\boldsymbol{w}} \ell)^\top)/A_m^{m_*-1}$. Our goal is to bound the spectral norm of this matrix.

As a first step, we bound $\|\operatorname{Cov}(\boldsymbol{g}_k)\|_{\text{op}}$ under the assumption that $(\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+$ is uniformly bounded. We later show (in Proposition 4.3) how to relax this assumption by instead assuming higher-order moment bounds for the stochastic subgradient.

Proposition 4.2 (Subgradient Covariance Bound for the Generalized Objective)**.** *Assume that for a given iterate $\boldsymbol{x} := (\boldsymbol{w}, \gamma)$, the residual loss is almost surely bounded, i.e., $(\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+ \leq L_0$ for some constant $L_0 > 0$. Let $\boldsymbol{S} := \mathbb{E}[\nabla_{\boldsymbol{w}} \ell (\nabla_{\boldsymbol{w}} \ell)^\top]$ be the unconditional second moment matrix of the loss gradient. Then the spectral norm of the single-sample subgradient covariance is bounded by:*

$$\|\operatorname{Cov}(\boldsymbol{g}_k(\boldsymbol{x}; \boldsymbol{\xi}))\|_{\text{op}} \leq s_m(\rho)^2 \cdot (L_0/A_m)^{2(m_*-1)} \left(1 + \sqrt{\|\boldsymbol{S}\|_{\text{op}}}\right)^2.$$

Proof. Let $\boldsymbol{v} = \nabla_{\boldsymbol{w}} \ell$. We will bound the operator norm of the block matrix

$$\boldsymbol{M} = \operatorname{Cov}((-\beta_m, \beta_m \boldsymbol{v}^\top)^\top) = \begin{pmatrix} \operatorname{Var}(\beta_m) & -\operatorname{Cov}(\beta_m, \beta_m \boldsymbol{v})^\top \\ -\operatorname{Cov}(\beta_m, \beta_m \boldsymbol{v}) & \operatorname{Cov}(\beta_m \boldsymbol{v}) \end{pmatrix}.$$

Using the property that, for a symmetric matrix $\boldsymbol{M} = \begin{pmatrix} a & \boldsymbol{b}^\top \\ \boldsymbol{b} & \boldsymbol{C} \end{pmatrix}$, $\|\boldsymbol{M}\|_{\mathrm{op}} \le |a| + \|\boldsymbol{C}\|_{\mathrm{op}} + 2\|\boldsymbol{b}\|_2$ (which can be derived from the triangle inequality), we have

$$\|\boldsymbol{M}\|_{\mathrm{op}} \le \mathrm{Var}(\beta_m) + 2\|\mathrm{Cov}(\beta_m, \beta_m \boldsymbol{v})\|_2 + \|\mathrm{Cov}(\beta_m \boldsymbol{v})\|_{\mathrm{op}}.$$

We bound each term by leveraging the assumption $(\ell - \gamma)_+ \le L_0$, which implies the random weight is bounded: $\beta_m = (\ell - \gamma)_+^{m_*-1} \le L_0^{m_*-1}$.

First, consider the main diagonal block, $\mathrm{Cov}(\beta_m \boldsymbol{v})$. Since covariance matrices are positive semidefinite, its norm is bounded by the norm of its second moment matrix: $\mathrm{Cov}(\beta_m \boldsymbol{v}) \preceq \mathbb{E}[\beta_m^2 \boldsymbol{v}\boldsymbol{v}^\top]$, which implies $\|\mathrm{Cov}(\beta_m \boldsymbol{v})\|_{\mathrm{op}} \le \|\mathbb{E}[\beta_m^2 \boldsymbol{v}\boldsymbol{v}^\top]\|_{\mathrm{op}}$. Using our boundedness assumption:

$$\mathbb{E}[\beta_m^2 \boldsymbol{v}\boldsymbol{v}^\top] = \mathbb{E}\left[(\ell - \gamma)_+^{2(m_*-1)} \boldsymbol{v}\boldsymbol{v}^\top\right] \preceq L_0^{2(m_*-1)} \mathbb{E}[\boldsymbol{v}\boldsymbol{v}^\top] = L_0^{2(m_*-1)} \boldsymbol{S}.$$

This gives the crucial bound $\|\mathrm{Cov}(\beta_m \boldsymbol{v})\|_{\mathrm{op}} \le L_0^{2(m_*-1)} \|\boldsymbol{S}\|_{\mathrm{op}}$.

Next, we bound the top-left block, $\mathrm{Var}(\beta_m)$.

$$\mathrm{Var}(\beta_m) \le \mathbb{E}[\beta_m^2] = \mathbb{E}\left[(\ell - \gamma)_+^{2(m_*-1)}\right] \le L_0^{2(m_*-1)}.$$

Finally, we bound the norm of the off-diagonal block, $\mathrm{Cov}(\beta_m, \beta_m \boldsymbol{v})$, using a standard inequality for covariance norms, $\|\mathrm{Cov}(X, \boldsymbol{Y})\|_2 \le \sqrt{\mathrm{Var}(X)}\sqrt{\|\mathrm{Cov}(\boldsymbol{Y})\|_{\mathrm{op}}}$:

$$\begin{aligned} \|\mathrm{Cov}(\beta_m, \beta_m \boldsymbol{v})\|_2 &\le \sqrt{\mathrm{Var}(\beta_m)}\sqrt{\|\mathrm{Cov}(\beta_m \boldsymbol{v})\|_{\mathrm{op}}} \\ &\le \sqrt{L_0^{2(m_*-1)}}\sqrt{L_0^{2(m_*-1)}\|\boldsymbol{S}\|_{\mathrm{op}}} = L_0^{2(m_*-1)}\sqrt{\|\boldsymbol{S}\|_{\mathrm{op}}}. \end{aligned}$$

It follows that

$$\begin{aligned} \|\boldsymbol{M}\|_{\mathrm{op}} &\le \mathrm{Var}(\beta_m) + \|\mathrm{Cov}(\beta_m \boldsymbol{v})\|_{\mathrm{op}} + 2\|\mathrm{Cov}(\beta_m, \beta_m \boldsymbol{v})\|_2 \\ &\le L_0^{2(m_*-1)} + L_0^{2(m_*-1)}\|\boldsymbol{S}\|_{\mathrm{op}} + 2L_0^{2(m_*-1)}\sqrt{\|\boldsymbol{S}\|_{\mathrm{op}}} \\ &= L_0^{2(m_*-1)}\left(1 + \|\boldsymbol{S}\|_{\mathrm{op}} + 2\sqrt{\|\boldsymbol{S}\|_{\mathrm{op}}}\right). \end{aligned}$$

Multiplying by the scalar constant $(s_m(\rho)/A_m^{m_*-1})^2$ completes the proof. □

It is challenging to establish a dimension-independent bound on the residual loss, $(\ell(\boldsymbol{w}, \boldsymbol{\xi}) - \gamma)_+ \le L_0$, as required by the preceding proposition. We therefore derive a more broadly applicable bound that relies instead on a standard higher-moment assumption on the loss gradients, a condition that is satisfied by many common models and data distributions.

Proposition 4.3 (Subgradient Covariance Bound for the Generalized Objective Under Higher Moment Bounds)**.** *Let the loss function $\ell(\boldsymbol{w}, \boldsymbol{\xi})$ be convex in $\boldsymbol{w}$. Assume that for a given iterate $\boldsymbol{x} := (\boldsymbol{w}, \gamma)$, the loss gradient $\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{\xi})$ has a bounded $2r$-th moment for some $r > 1$:*

$$\max_{\boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_2 = 1} \mathbb{E}\left[(\nabla_{\boldsymbol{w}} \ell(\boldsymbol{w}, \boldsymbol{\xi}) \cdot \boldsymbol{u})^{2r}\right] \le B^{2r},$$

for some constant $B > 0$. Let $m_* = m/(m-1)$ and $r_* = r/(r-1)$ be the respective conjugate exponents. The spectral norm of the single-sample subgradient covariance is bounded by:

$$\|\operatorname{Cov}(\boldsymbol{g}_k(\boldsymbol{x};\boldsymbol{\xi}))\|_{\mathrm{op}} \le \frac{2s_m(\rho)^2\Big(B^2(\mathbb{E}[(\ell-\gamma)_+^{2(m_*-1)r_*}])^{1/r_*} + \mathbb{E}[(\ell-\gamma)_+^{2(m_*-1)}]\Big)}{(\mathbb{E}[(\ell-\gamma)_+^{m_*}])^{2-2/m_*}}.$$

In particular, for risk measures with $m > 2$, if the moment order r is sufficiently large to satisfy $r \ge \frac{m}{m-2}$, the bound simplifies to a constant that is independent of the iterate's residual loss:

$$\|\operatorname{Cov}(\boldsymbol{g}_k(\boldsymbol{x};\boldsymbol{\xi}))\|_{\mathrm{op}} \le 2s_m(\rho)^2(B^2+1).$$

Proof. Let $\boldsymbol{v} = \nabla_{\boldsymbol{w}}\ell$ and $\beta_m = (\ell-\gamma)_+^{m_*-1}$. The single-sample subgradient is

$$\boldsymbol{g}_k = \boldsymbol{e}_1 + s_m(\rho)A_m^{-(m_*-1)}(-\beta_m, \beta_m\boldsymbol{v}^\top)^\top,$$

where $A_m = (\mathbb{E}[(\ell-\gamma)_+^{m_*}])^{1/m_*}$. Its covariance is

$$\operatorname{Cov}(\boldsymbol{g}_k) = \frac{s_m(\rho)^2}{A_m^{2(m_*-1)}}\operatorname{Cov}((-\beta_m, \beta_m\boldsymbol{v}^\top)^\top).$$

Let $\boldsymbol{M} = \operatorname{Cov}((-\beta_m, \beta_m\boldsymbol{v}^\top)^\top) = \begin{pmatrix} \operatorname{Var}(\beta_m) & -\operatorname{Cov}(\beta_m, \beta_m\boldsymbol{v})^\top \\ -\operatorname{Cov}(\beta_m, \beta_m\boldsymbol{v}) & \operatorname{Cov}(\beta_m\boldsymbol{v}) \end{pmatrix}$. The operator norm of this 2×2 block matrix can be bounded. For a random scalar X and random vector $\boldsymbol{Y}$, the Cauchy-Schwarz inequality implies $\|\operatorname{Cov}(X,\boldsymbol{Y})\|_2^2 \le \operatorname{Var}(X)\|\operatorname{Cov}(\boldsymbol{Y})\|_{\mathrm{op}}$. Using Young's inequality, $2\sqrt{\operatorname{Var}(X)\|\operatorname{Cov}(\boldsymbol{Y})\|_{\mathrm{op}}} \le \operatorname{Var}(X) + \|\operatorname{Cov}(\boldsymbol{Y})\|_{\mathrm{op}}$. This leads to the bound:

$$\begin{aligned}\|\boldsymbol{M}\|_{\mathrm{op}} &\le \operatorname{Var}(\beta_m) + \|\operatorname{Cov}(\beta_m\boldsymbol{v})\|_{\mathrm{op}} + 2\|\operatorname{Cov}(\beta_m, \beta_m\boldsymbol{v})\|_2 \\ &\le 2\operatorname{Var}(\beta_m) + 2\|\operatorname{Cov}(\beta_m\boldsymbol{v})\|_{\mathrm{op}}.\end{aligned}$$

We now bound the two terms separately. The variance is bounded by its second moment, $\operatorname{Var}(\beta_m) \le \mathbb{E}[\beta_m^2]$. The norm of the vector covariance is bounded by its second moment matrix, $\|\operatorname{Cov}(\beta_m\boldsymbol{v})\|_{\mathrm{op}} \le \|\mathbb{E}[\beta_m^2\boldsymbol{v}\boldsymbol{v}^\top]\|_{\mathrm{op}}$. To bound this, consider any unit vector $\boldsymbol{u} \in \mathbb{R}^d$:

$$\boldsymbol{u}^\top\mathbb{E}[\beta_m^2\boldsymbol{v}\boldsymbol{v}^\top]\boldsymbol{u} = \mathbb{E}[\beta_m^2(\boldsymbol{v}\cdot\boldsymbol{u})^2].$$

By Hölder's inequality with exponents r and $r_* = r/(r-1)$, we have

$$\begin{aligned}\mathbb{E}[\beta_m^2(\boldsymbol{v}\cdot\boldsymbol{u})^2] &\le (\mathbb{E}[|(\boldsymbol{v}\cdot\boldsymbol{u})|^{2r}])^{1/r}(\mathbb{E}[|\beta_m^2|^{r_*}])^{1/r_*} \\ &\le (B^{2r})^{1/r}(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*} = B^2(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*}.\end{aligned}$$

Since this holds for any unit vector $\boldsymbol{u}$, we have $\|\mathbb{E}[\beta_m^2\boldsymbol{v}\boldsymbol{v}^\top]\|_{\mathrm{op}} \le B^2(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*}$. Combining the pieces, we have

$$\|\boldsymbol{M}\|_{\mathrm{op}} \le 2\mathbb{E}[\beta_m^2] + 2B^2(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*}.$$

Substituting this into the expression for $\operatorname{Cov}(\boldsymbol{g}_k)$ and replacing β_m and A_m with their definitions in terms of $(\ell-\gamma)_+$ yields the first result of the proposition.

For the second part, we simplify the expression under additional conditions. By the property of L_p norms of random variables, since $r_* > 1$, we have $(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*} \ge \mathbb{E}[\beta_m^2]$. Thus, the bound on $\|\boldsymbol{M}\|_{\mathrm{op}}$ is dominated by the term involving B^2:

$$\|\boldsymbol{M}\|_{\mathrm{op}} \le 2(1+B^2)(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*}.$$

The full bound on the subgradient covariance is therefore:

$$\|\mathrm{Cov}(\boldsymbol{g}_k)\|_{\mathrm{op}} \le \frac{2s_m(\rho)^2(B^2+1)}{A_m^{2(m_*-1)}}(\mathbb{E}[\beta_m^{2r_*}])^{1/r_*}.$$

Let $X = (\ell - \gamma)_+$. The ratio of the moment terms can be written as:

$$\frac{(\mathbb{E}[X^{2(m_*-1)r_*}])^{1/r_*}}{(\mathbb{E}[X^{m_*}])^{2-2/m_*}} = \frac{(\mathbb{E}[X^{2(m_*-1)r_*}])^{1/r_*}}{(\mathbb{E}[X^{m_*}])^{2(m_*-1)/m_*}} = \left(\frac{(\mathbb{E}[X^{2(m_*-1)r_*}])^{1/(2(m_*-1)r_*)}}{(\mathbb{E}[X^{m_*}])^{1/m_*}}\right)^{2(m_*-1)}.$$

This is a ratio of L_p norms of random variables. If the exponent in the numerator's power mean is less than or equal to that of the denominator, the ratio is at most 1. This condition is $2(m_*-1)r_* \le m_*$. Substituting $m_* = m/(m-1)$ and $r_* = r/(r-1)$, the condition becomes:

$$2\frac{1}{m-1}\frac{r}{r-1} \le \frac{m}{m-1} \implies \frac{2r}{r-1} \le m \implies 2\left(1+\frac{1}{r-1}\right) \le m.$$

This inequality requires $m > 2$. Rearranging for r gives $r(m-2) \ge m$, or $r \ge \frac{m}{m-2}$. When this condition on m and r holds, the ratio of moments is at most 1, simplifying the overall covariance bound to:

$$\|\mathrm{Cov}(\boldsymbol{g}_k(\boldsymbol{x};\boldsymbol{\xi}))\|_{\mathrm{op}} \le 2s_m(\rho)^2(B^2+1).$$

This completes the proof. □

With the necessary machinery in place, we now present our main result for the outlier-robust minimization of the generalized risk measures based on f-divergences.

Theorem 4.2 (Outlier-Robust f-Divergence Minimization)**.** *Let $f_m(\boldsymbol{x})$ be the generalized risk objective (4.1) for some $m > 1$ where we write $\boldsymbol{x} = (\boldsymbol{w}, \gamma)$, which is convex and L_m-Lipschitz. Let $\boldsymbol{x}^\star$ be a minimizer of f_m over a compact convex set $\mathscr{C} \subset \mathbb{R}^{d+1}$. Suppose that for any $\boldsymbol{x} \in \mathscr{C}$, the loss gradients $\nabla_{\boldsymbol{w}}\ell(\boldsymbol{w},\boldsymbol{\xi})$ from the inlier distribution $\mathbb{P}$ have a bounded $2r$-th moment for some $r > 1$, as specified in Proposition 4.3 with constant B.*

Consider the procedure from Algorithm 2, where the inexact subgradient $\tilde{\boldsymbol{g}}_k$ at each step k is computed by applying the robust mean estimator from Algorithm 1 to an ε-corrupted set of N subgradients. For a sufficiently large sample size $N = O((dL_m^2 \log d + L_m^2 \log(L_m/(\zeta\sigma_m\varepsilon)))/\sigma_m^2\varepsilon^2)$, running the algorithm for $K = O(L_m^2/(\sigma_m^2\varepsilon))$ iterations produces an output $\bar{\boldsymbol{x}}_K$ that, with probability at least $1-\zeta$, satisfies:

$$f_m(\bar{\boldsymbol{x}}_K) - f_m(\boldsymbol{x}^\star) = O\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \sigma_m \sqrt{\varepsilon}\right),$$

where σ_m^2 is the upper bound on the single-sample subgradient covariance given by the general result in Proposition 4.3.

In particular, if the risk measure satisfies $m > 2$ and the moment order of the loss gradient satisfies $r \ge m/(m-2)$, the suboptimality bound simplifies to:

$$f_m(\bar{\boldsymbol{x}}_K) - f_m(\boldsymbol{x}^\star) = O\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 s_m(\rho) B\sqrt{\varepsilon}\right).$$

Proof. The proof strategy is to instantiate the guarantee for the inexact projected subgradient method, Theorem 3.1, with a robust subgradient oracle. The quality of this oracle is determined by the stability of the single-sample subgradient distribution.

At each iteration k, our procedure computes an inexact subgradient $\tilde{\boldsymbol{g}}_k$ by applying the robust mean estimator to an ε-corrupted set of single-sample subgradients, $\{\boldsymbol{g}_k(\boldsymbol{x}_k;\boldsymbol{\xi}_i)\}_{i=1}^N$. From

Proposition 4.3, we have an upper bound σ_m^2 on the covariance of these subgradient vectors. By choosing a sufficiently large sample size N, Corollary 3.1 guarantees that the robustly estimated mean $\tilde{\boldsymbol{g}}_k$ is close to the true subgradient $\boldsymbol{g}_k = \mathbb{E}[\boldsymbol{g}_k(\boldsymbol{x}_k;\boldsymbol{\xi})]$ for all k, providing an error bound of $\|\tilde{\boldsymbol{g}}_k - \boldsymbol{g}_k\|_2 \le \delta = O(\sigma_k\sqrt{\varepsilon})$.

The objective function f_m is convex and L_m-Lipschitz. We can therefore apply the convergence guarantee from Theorem 3.1. Setting the number of iterations $K = O(L_m^2/\delta^2)$ and denoting $D_0 = \|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2$, the suboptimality is bounded by:

$$\begin{aligned} f_m(\bar{\boldsymbol{x}}_K) - f_m(\boldsymbol{x}^\star) &= O\left(\frac{D_0(L_m+\delta)}{\sqrt{K}} + D_0\delta\right) \\ &= O(D_0\delta) \\ &= O(D_0\sigma_m\sqrt{\varepsilon}). \end{aligned}$$

This establishes the first result, where σ_m is derived from the general covariance bound in Proposition 4.3.

For the particular case where $m > 2$ and $r \ge m/(m-2)$, Proposition 4.3 provides a simplified, constant bound on the covariance, yielding $\sigma_m = O(s_m(\rho)B)$. Substituting this into the general performance guarantee gives the second, simplified result and completes the proof. □

4.3. **Comparison to DORO.** Our work provides the first polynomial-time algorithms for the problem of outlier-robust f-divergence DRO, which was similarly conceptualized in the Distributionally Robust Outlier-aware Optimization (DORO) framework [49]. The DORO framework defines a statistical estimator—the minimizer of the "DORO risk"—that possesses desirable robustness properties. The ε-DORO risk is defined with respect to an f-divergence D and ambiguity radius ρ as

$$\mathscr{R}_{\rho,D,\varepsilon}(\boldsymbol{w};\mathbb{P}_{\text{train}}) = \inf_{\substack{\mathbb{P}':\exists\tilde{\mathbb{P}}\text{ s.t.}\\ \mathbb{P}_{\text{train}}=(1-\varepsilon)\mathbb{P}'+\varepsilon\tilde{\mathbb{P}}}} \sup_{\substack{\mathbb{Q}\ll\mathbb{P}':\\ D(\mathbb{Q}\|\mathbb{P}')\le\rho}} \mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{Q}}[\ell(\boldsymbol{w};\boldsymbol{\xi})], \tag{4.4}$$

where the infimum is taken over all possible decompositions of the corrupted data distribution $\mathbb{P}_{\text{train}}$ into a clean part $\mathbb{P}'$ and a corrupted part $\tilde{\mathbb{P}}$. While this formulation is statistically appealing, it remains an open question whether the minimizer of (4.4) can be computed or even approximated in polynomial time. The algorithm proposed in [49] is a heuristic that lacks convergence guarantees and has been shown to fail in certain cases [33]. Our work resolves this algorithmic gap by providing a provably correct and efficient method.

It is instructive to compare our algorithm's performance guarantees with the statistical guarantees established for the idealized DORO estimator. The key distinction lies in the underlying assumptions: the DORO guarantees are predicated on moment bounds of the *loss values*, whereas our results rely on moment bounds of the *loss gradients*.

Fact 4.1 (DORO's Guarantees, adapted from [49, Theorems 5 and 6])**.** *Let the contaminated distribution be $\mathbb{P}_{train} = (1-\varepsilon)\mathbb{P} + \varepsilon\tilde{\mathbb{P}}$, and let $\boldsymbol{w}^\star$ be the minimizer of the DORO risk (4.4). If the loss function $\ell(\boldsymbol{w};\boldsymbol{\xi})$ is non-negative and $\ell(\boldsymbol{w}^\star;\boldsymbol{\xi})$ has a bounded $(2q)$-th moment under the clean distribution $\mathbb{P}$ ($\mathbb{E}_{\boldsymbol{\xi}\sim\mathbb{P}}[\ell(\boldsymbol{w}^\star;\boldsymbol{\xi})^{2q}] = \sigma_{2q}^{2q} < +\infty$ for some $q \ge 1$), then the suboptimality of $\boldsymbol{w}^\star$ is bounded. For CVaR at level α, the bound is:*

$$\mathrm{CVaR}^{\alpha}_{\boldsymbol{\xi}\sim\mathbb{P}}(\ell(\boldsymbol{w}^\star;\boldsymbol{\xi})) - \inf_{\boldsymbol{w}}\mathrm{CVaR}^{\alpha}_{\boldsymbol{\xi}\sim\mathbb{P}}(\ell(\boldsymbol{w};\boldsymbol{\xi})) \le 4\sigma_{2q}\varepsilon^{1-\frac{1}{2q}}/\alpha.$$

For the DRO problem with χ^2-divergence and $q > 1$, the bound is:

$$\max_{\chi^2(\mathbb{Q}\|\mathbb{P})\le\rho} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{w}^\star;\boldsymbol{\xi})] - \inf_{\boldsymbol{w}} \max_{\chi^2(\mathbb{Q}\|\mathbb{P})\le\rho} \mathbb{E}_{\mathbb{Q}}[\ell(\boldsymbol{w};\boldsymbol{\xi})] = O_{\rho,q}(\sigma_{2q}\varepsilon^{1/2-1/(2q)}).$$

Moreover, these dependencies on α,ε,q, and σ_{2q} are optimal up to a constant factor.

While the assumptions are not directly comparable, our error rates are analogous to these statistically optimal rates in important special cases. For the CVaR problem, Fact 4.1 with a loss second-moment bound ($q = 1$) yields an error of $O(\sigma_2\sqrt{\varepsilon}/\alpha)$. Our result in Theorem 4.1, derived under a uniform gradient second-moment bound, achieves an error of $O(G\sqrt{\varepsilon}/\alpha)$, matching the dependence on ε and α. For the χ^2-divergence problem ($m = 2$), Fact 4.1 under a bounded loss assumption ($q \to \infty$) gives an error of $O(\sqrt{\varepsilon})$. Our result from Theorem 4.2 also yields an $O(\sqrt{\varepsilon})$ error rate under the assumptions of bounded loss and a uniformly bounded gradient second moment. The latter is a relatively mild condition in the context of bounded loss. Thus, our work not only provides the first polynomial-time algorithm for this problem but also achieves error guarantees that align with the optimal statistical rates established for the idealized (but not known to be computationally tractable) DORO estimator.

5. Conclusion

We presented a general framework for nonsmooth stochastic optimization in the presence of adversarial outliers. By integrating robust mean estimation with classical subgradient methods, we obtained error guarantees that are unimprovable in general and vanish entirely in the absence of data variation. As its main application, our framework further yields the first efficient algorithm for outlier-robust distributionally robust optimization, addressing both CVaR and f-divergence settings. These results extend the reach of robust optimization methods to a broad class of nonsmooth objectives fundamental to modern machine learning. Looking ahead, we hope our work will stimulate further research in this area. In particular, our focus in this work was on theory—establishing the existence of polynomial-time algorithms with optimal error guarantees in outlier-robust settings. While the established sample and iteration complexities are polynomial in the input size, we did not optimize their order or consider practical implementations. We leave such considerations to future research.

Acknowledgments

This article was written as a tribute to Prof. Yurii Nesterov's seminal contributions to the field of optimization algorithms, on the occasion of his 70th birthday. The first-order optimization portion of this work is directly motivated by Prof. Nesterov's work on dual averaging methods [36] and first-order optimization with inexact oracles [13].

This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-24-1-0076, by the U.S. Office of Naval Research under contract number N00014-22-1-2348, and by the NSF CAREER Award CCF-2440563. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

References

[1] F. J. Anscombe, Rejection of outliers, Technometrics, 2 (1960), 123–147.

[2] V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd edition, John Wiley & Sons, Chichester, 1994.

[3] M. Barreno, B. Nelson, A. D. Joseph, J. D. Tygar, The security of machine learning, Machine Learning, 81 (2010), 121–148.

[4] A. Beck, First-order Methods in Optimization, SIAM, Philadelphia, 2017.

[5] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, Robust Optimization, vol. 28, Princeton University Press, Princeton, 2009.

[6] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, In: Proceedings of the 29th International Coference on International Conference on Machine Learning, pp. 1467–1474, Edinburgh, 2012.

[7] J. Blanchet, J. Li, S. Lin, X. Zhang, Distributionally robust optimization and robust statistics, arXiv preprint arXiv:2401.14655, 2024.

[8] L. Bottou, F.E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Rev. 60 (2018), 223–311.

[9] Y. Cheng, I. Diakonikolas, R. Ge, High-dimensional robust mean estimation in nearly-linear time, In: Proceedings of the 30th ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 2755–2771, SIAM, 2019.

[10] N. Cressie, T.R.C. Read, Multinomial goodness-of-fit tests, J. R. Stat. Soc. Ser. B. Stat. Methodol. 46 (1984), 440–464.

[11] D. Davis, D. Drusvyatskiy, Stochastic model-based minimization of weakly convex functions, SIAM J. Optim. 29 (2019), 207–239.

[12] E. Delage, Y. Ye, Distributionally robust optimization under moment uncertainty with application to data-driven problems, Oper. Res. 58 (2010), 595–612.

[13] O. Devolder, F. Glineur, Y. Nesterov, First-order methods of smooth convex optimization with inexact oracle, Math. Program. 146 (2014), 37–75.

[14] I. Diakonikolas, G. Kamath, D. Kane, J. Li, A. Moitra, A. Stewart, Robust estimators in high-dimensions without the computational intractability, SIAM J. Comput. 48 (2019), 742–864.

[15] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, A. Stewart, Sever: A robust meta-algorithm for stochastic optimization, In: Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 1596–1606, PMLR, 2019.

[16] I. Diakonikolas, G. Kamath, D.M. Kane, J. Li, A. Moitra, A. Stewart, Robust estimators in high dimensions without the computational intractability, In: 57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016, pp. 655–664, IEEE Computer Soc., Los Alamitos, CA, 2016.

[17] I. Diakonikolas, D.M. Kane, Algorithmic high-dimensional robust statistics, Cambridge University Press, Cambridge, 2023.

[18] I. Diakonikolas, D.M. Kane, A. Pensia, Outlier robust mean estimation with subgaussian rates via stability, In: Advances in Neural Information Processing Systems, vol. 33, pp. 1830–1840, Curran Associates, Inc., 2020.

[19] I. Diakonikolas, D.M. Kane, A. Pensia, T. Pittas, Streaming algorithms for high-dimensional robust statistics, In: Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pp. 5061–5117, PMLR, 2022.

[20] Y. Dong, S. Hopkins, J. Li, Quantum entropy scoring for fast robust mean estimation and improved outlier detection, Advances in Neural Information Processing Systems 32, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, 2019.

[21] J.C. Duchi, H. Namkoong, Learning models with uniform performance via distributionally robust optimization, Ann. Statist. 49 (2021), 1378–1406.

[22] C. Gao, A. Lowy, X. Zhou, S.J. Wright, Optimal rates for robust stochastic convex optimization, In: 6th Symposium on Foundations of Responsible Computing (FORC), pp. 1–21, 2025. doi: 10.48550/arXiv.2412.11003

[23] R. Gao, Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality, Oper. Res. 71 (2023), 2291–2306.

[24] R. Gao, A. Kleywegt, Distributionally robust stochastic optimization with Wasserstein distance, Math. Oper. Res. 48 (2023), 603–655.

[25] J. Hayase, W. Kong, R. Somani, S. Oh, Spectre: defending against backdoor attacks using robust statistics, Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[26] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Statist. 35 (1964), 73–101.

[27]

[28] P.J. Huber, E.M. Ronchetti, Robust Statistics, Wiley Series in Probability and Statistics, 2nd edition, John Wiley & Sons, Hoboken, NJ, 2009.

[29] K. A. Lai, A. B. Rao, S. Vempala, Agnostic estimation of mean and covariance, In: 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pp. 665–674, IEEE, 2016.

[30] J. Lee, M. Raginsky, Minimax statistical learning with wasserstein distances, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, 2018.

[31] D. Levy, Y. Carmon, J.C. Duchi, A. Sidford, Large-scale methods for distributionally robust optimization, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, 2020.

[32] J.Z. Li, D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, M. Feldman, L.L. Cavalli-Sforza, R.M. Myers, orldwide human relationships inferred from genome-wide patterns of variation, Science, 319 (2008), 1100–1104.

[33] S. Li, I. Diakonikolas, J. Diakonikolas, Distributionally robust optimization with adversarial data contamination, arXiv preprint arXiv:2507.10718, 2025.

[34] H. Namkoong, J.C. Duchi, Stochastic gradient methods for distributionally robust optimization with f-divergences, Advances in neural information processing systems, 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, 2016.

[35] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM J. Optim. 19 (2009), 1574–1609.

[36] Y. Nesterov, Primal-dual subgradient methods for convex problems, Math. Program. 120 (2009), 221–259.

[37] P. Paschou, J. Lewis, A. Javed, P. Drineas, Ancestry informative markers for fine-scale individual assignment to worldwide populations, J. Medical Genetics, 47 (2010), 835–847.

[38] A. Prasad, S. Balakrishnan, P. Ravikumar, A robust univariate mean estimator is all you need, In: International Conference on Artificial Intelligence and Statistics, pp. 4034–4044, PMLR, 2020.

[39] H. Rahimian, S. Mehrotra, Frameworks and results in distributionally robust optimization, Open J. Math. Optim. 3 (2022), 4.

[40] R.T. Rockafellar, Stanislav Uryasev, Optimization of conditional value-at-risk, J. Risk 2 (2000), 21–42.

[41] R.T. Rockafellar, R.J.B. Wets, Variational Analysis, Springer Berlin, Heidelberg, 1998.

[42] N. Rosenberg, J. Pritchard, J. Weber, H. Cann, K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science, 298 (2002), 2381–2385.

[43] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, 1987.

[44] A. Shapiro, Distributionally robust stochastic programming, SIAM J. Optim. 27 (2017), 2258–2275.

[45] J. Steinhardt, P. W. Koh, P. Liang, Certified defenses for data poisoning attacks, In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3520–3532, 2017.

[46] B. Tran, J. Li, A. Madry, Spectral signatures in backdoor attacks, In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 8011–8021, Montréal, 2018.

[47] J. W. Tukey, A survey of sampling from contaminated distributions, Contrib. to Probab. Stat. 2 (1960), 448–485.

[48] J. Yang, K. Zhou, Y. Li, Z. Liu, Generalized out-of-distribution detection: A survey, Int. J. Comput. Vis. 132 (2024), 5635-–5662.

[49] R. Zhai, C. Dan, Z. Kolter, P. Ravikumar, Doro: Distributional and outlier robust optimization, In: Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 12345–12355, PMLR, 2021.