High-Probability Bounds for Robust Stochastic Frank-Wolfe Algorithm

Abstract

We develop and analyze robust Stochastic Frank-Wolfe type algorithms for projection-free stochastic convex optimization problems with heavy-tailed stochastic gradients. Existing works on the oracle complexity of such algorithms require a uniformly bounded variance assumption, and hold only in expectation. We develop tight high-probability bounds for robust versions of Stochastic Frank-Wolfe type algorithm under heavy-tailed assumptions, including infinite variance, on the stochastic gradient. Our methodological construction of the robust Stochastic Frank-Wolfe type algorithms leverage techniques from the robust statistic literature. Our theoretical analysis highlights the need to utilize robust versions of Stochastic Frank-Wolfe type algorithm for dealing with heavy-tailed data arising in practice.

1 INTRODUCTION

In this paper, we consider constrained stochastic optimization problem of the form

$$\arg\min_{x\in\mathcal{X}} \ \{f(x) = \mathbb{E}_\xi[F(x,\xi)] = \int F(x,\xi)\, d\mathbb{P}(\xi)\}, \quad (1)$$

where f is a smooth convex function, $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$, ξ is a random vector defined according to a distribution $\mathbb{P}$ on the domain of ξ. We denote by x^*, the minimizer of $f(x)$ in (1). Problems of the form in (1) arise frequently in modern machine learning, including matrix completion [15], structured linear inverse problems [7], multi-task learning [43] and particle filtering [29]. Stochastic Frank-Wolfe (SFW) and Stochastic Conditional Gradient Sliding (SCGS) algorithms are widely used iterative first-order algorithm for solving (1). Each iteration of SFW/SCGS involves performing linear minimization over the constraint set $\mathcal{X}$ based on stochastic first-order (or gradient) information, $\nabla F(\cdot,\xi) \in \mathbb{R}^d$. Depending on the geometry of the constraint set, SFW/SCGS is widely used in practice due to its projection-free nature. Despite its widespread usage, our understanding of the oracle complexity of SFW/SCGS algorithm is limited.

As each iteration of the SFW/SCGS algorithm requires access to a Stochastic First-order Oracle (SFO) and a Linear Minimization Oracle (LMO), the oracle complexity is measured by the number of calls to SFO and LMO to obtain an ϵ-optimal solution – that is, a point $\bar{x}$ such that $f(\bar{x}) - f(x_*) \leq \epsilon$. A majority of the existing results on the oracle complexity of SFW/SCGS algorithm are available only in expectation, i.e., on the metric $\mathbb{E}_\xi[f(\bar{x}) - f(x_*)] \leq \epsilon$. Furthermore, to obtain such oracle complexity results in expectation, it is assumed that stochastic gradient used has uniformly bounded variance (i.e., $\mathbb{E}_\xi[\|\nabla F(x,\xi) - \nabla f(x)\|^2] \leq \sigma^2$), where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. This characterization of the oracle complexity provides very little understanding regarding the behaviour of the SFW/SCGS algorithms. In particular, the effect of the properties of the distribution $\mathbb{P}$, on the heavy-tail nature of the stochastic gradient, and consequently on the oracle complexity is *not characterized* by the oracle complexity results in expectation. Furthermore, the oracle complexity of SFW/SCGS algorithm or its robust variants under infinite variance assumption is not known.

To provide a concrete motivating example, consider the problem of sparse linear regression: Given $(a, y) \in \mathbb{R}^d \times \mathbb{R}$, consider the linear regression model $y = \langle a, \bar{x}\rangle + \epsilon$, where for two vectors $c, d \in \mathbb{R}^d$, $\langle c, d\rangle$ represents the Euclidean inner product. Here, we let $\epsilon \sim N(0,1)$ and the true parameter $\bar{x}$ is assumed to be s-sparse (i.e., it has only s non-zero coordinates out of the d coordinates). The L_1-constrained least-squares estimator is then given by

$$\arg\min_{x\in\mathcal{X}_1(s)} \ \mathbb{E}[(y - \langle a, x\rangle)^2], \quad (2)$$

where $\mathcal{X}_1(s) := \{x \in \mathbb{R}^d : \sum_{j=1}^d |x_j| \leq s\}$ is the L_1 ball

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

of radius s. This problem fits in the setup of (1) with

$$\xi := (a, y) \quad \text{and} \quad F(x, \xi) := (y - a^\top x)^2.$$

Hence the stochastic gradient is given by

$$G(x, \xi) := \nabla F(x, \xi) = 2(aa^\top x - ya) \in \mathbb{R}^d.$$

Note that as the iterates of SFW/SCGS algorithms (see, for example, Algorithm 2 for the description of SCGS) are in the set $\mathcal{X}_1(s)$, we have $\|x\|$ is to be always bounded for all x along the trajectory of the algorithm. Hence, the $(1+\alpha)$-th moment of the stochastic gradient, i.e., $\mathbb{E}[\|G(x, \xi)\|^{(1+\alpha)}]$, is controlled by the order of $\mathbb{E}[\|a\|^{2(1+\alpha)}]$. When the covariate a is a zero-mean multivariate t-distribution with degrees of freedom in the interval $[2, 4)$, or is a zero-mean multivariate Pareto distribution with parameter in the interval $[2, 4)$, the stochastic gradients have infinite variance. Hence, the existing oracle complexity results for SFW/SCGS are not applicable.

Focusing on the SCGS algorithm [31], in this work we develop and analyze robust versions of it, and establish high-probability SFO and LMO complexity bounds. First, under a sub-Gaussian tail assumption on the stochastic gradient, we establish high-probability bounds for the standard SCGS algorithm. However, when the stochastic gradient is heavy-tailed or has infinite variance, the standard SCGS algorithm is sub-optimal. This is due to the fact that each iteration of SCGS algorithm requires a sample average of mini-batch stochastic gradients. When the stochastic gradients are heavy-tailed, it is well-known from classical robust statistics literature that sample averages are poor estimates of the true expectation. Hence, to deal with the case when the stochastic gradient is heavy-tailed or has infinite variance, we construct robust versions of mini-batch stochastic gradient estimates which are subsequently used in the SCGS algorithm. Our motivations for developing robust SCGS algorithms are based on the fact that data arising from various real-world problems (for example, finance, networks, linguistics) are modeled efficiently using heavy-tailed distribution [42, 11, 48]. We establish high-probability bounds on the oracle complexity results on robust SCGS algorithm by developing tight concentration inequalities for heavy-tailed martingales, which might be of independent interest. The established high-probability bounds on the SFO and LMO complexity of SCGS, provide a fine-grained characterization of the effect of the moments/tails of the distribution $\mathbb{P}$ on the performance of SCGS algorithms in terms of both the level of solution accuracy and the confidence specified by the practitioner.

1.1 RELATED WORKS

Robust statistics: Robust statistics is a classical topic with too large a literature to summarize completely. We refer the reader to [25] for an overview. We emphasize that a majority of the robust estimators developed in the statistics literature are invariably computationally intractable. The revival of robust statistics in modern mathematical statistics and learning theory communities arguably started with the work of [6]. Since then, there has been intense work on robust mean and covariance estimation and robust empirical risk minimization; see [33] for a survey. However, such results are mainly statistical in nature and are not directly applicable for stochastic optimization with heavy-tailed gradients. In this work, we leverage the classical trick of *trimmed estimator* from the robust statistics literature, and the recently proposed optimal mean-estimation technique by [9] in the context of projection-free stochastic optimization.

Robust stochastic optimization: Early works on robust stochastic optimization include [27, 8, 37]. In modern stochastic optimization, several works consider the effect of heavy-tails on the performance of the algorithm [36, 50, 10]. There has also been overwhelming theoretical and empirical evidence in the modern machine learning literature that shows the noise in stochastic gradient algorithm could easily turn out to have heavy-tails due to the composite or product nature of the random vectors/matrices in the stochastic gradient [24, 39, 45, 46, 18, 5, 44]. Furthermore, recently robust mean estimation based and trimming based SGD algorithms were analyzed in [40] and [4, 17, 53, 52] respectively. However the above works do not deal with projection-free robust stochastic optimization, which is the main focus of our work.

Frank-Wolfe Algorithms: The Frank-Wolfe method first proposed by [14], has had a renewed interest in the past decade. We refer the reader to [26, 19, 28, 2, 16], for a partial list of recent works in the deterministic setting. Considering the stochastic convex setup, [21, 22] provided expected oracle complexity results for SFW algorithm. The complexities were further improved by a sliding procedure in [31], based on a modified Frank-Wolfe method by Nesterov's acceleration. It is common in SFW/SCGS analysis to require an increasing batch-size in each step, to obtain oracle complexity results. Recently, [35, 20, 54] addressed this issue of increasing batch size. But these works require restrictive assumptions, and have sub-optimal LMO complexity. In section 4 we empirically compare against the 1-sample SFW method from [35] in the heavy-tailed setting. We also highlight that several works [41, 51, 22] considered variance reduced versions of SFW and provided expected oracle complexities. In comparison to the above works, we focus on high-probability bounds on the oracle complexity in both the light-tailed and heavy-tailed setting (including infinite variance).

1.2 PRELIMINARIES AND OUR CONTRIBUTIONS

We now provide a list of theoretical contributions to the literature on oracle complexity of SFW algorithm. To do so,

Mean-Estimator	Tails	Theorem	SFO	LMO
Average	Asmp. 1.1	Thm. 3.1	$\mathcal{O}\left(\left(\frac{\log(1/\delta)}{\epsilon}\right)^2\right)$	$\mathcal{O}\left(\left(\frac{\log(1/\delta)}{\epsilon}\right)\right)$
Clipped Grad.	Asmp. 1.2	Thm. 3.2 (a)	$\mathcal{O}\left(\left(\frac{(\log(1/\delta))^{\frac{\alpha}{1+\alpha}}}{\epsilon}\right)^{\frac{3\alpha+2}{2\alpha}}\right)$	$\mathcal{O}\left(\frac{(\log(1/\delta))^{\frac{\alpha}{1+\alpha}}}{\epsilon}\right)$
Clipped Grad.	Asmp. 1.3	Thm. 3.2 (b)	$\mathcal{O}\left(\left(\frac{\sqrt{d}(\log(1/\delta))^{\frac{\beta}{1+\beta}}}{\epsilon}\right)^{\frac{3\beta+2}{2\beta}}\right)$	$\mathcal{O}\left(\frac{\sqrt{d}(\log(1/\delta))^{\frac{\beta}{1+\beta}}}{\epsilon}\right)$
CTBJ [9]	Asmp. 1.3	Thm. 3.3	$\mathcal{O}\left(\left(\frac{(\log(1/\delta))^{\frac{\beta}{1+\beta}}}{\epsilon}\right)^{\frac{3\beta+2}{2\beta}}\right)$	$\mathcal{O}\left(\frac{(\log(1/\delta))^{\frac{\beta}{1+\beta}}}{\epsilon}\right)$
BC Clipped Grad.	Asmp. 1.2	Thm. 3.4 (a)	$\mathcal{O}\left(\left(\frac{C(d,\alpha,\delta)}{\epsilon}\right)^{\frac{5\alpha+3}{4\alpha}}\right)$	$\mathcal{O}\left(\frac{C(d,\alpha,\delta)}{\epsilon}\right)$
BC Clipped Grad.	Asmp. 1.3	Thm. 3.4 (b)	$\mathcal{O}\left(\left(\frac{C(d,\beta,\delta)}{\epsilon}\right)^{\frac{5\beta+3}{4\beta}}\right)$	$\mathcal{O}\left(\frac{C(d,\beta,\delta)}{\epsilon}\right)$

Table 1: A summary of the obtained high-probability bounds. All results corresponds to the notion of (ϵ, δ)-optimal solution introduced in Definition 1.1. The parameters $C(d, \alpha, \delta)$ and $C(d, \beta, \delta)$ are defined in Theorem 3.4. BC stands for Bias-Corrected. The last two rows also require the symmetric condition described in Assumption 3.2.

we first introduce the notion of optimality that we consider for solving (1).

Definition 1.1 ((ϵ, δ)-optimal solution)**.** We call $\bar{x}$ an (ϵ, δ)-optimal solution of the stochastic optimization problem (1), if we have $\mathbb{P}\left(f(\bar{x}) - f(x_*) \leq \epsilon(\delta)\right) \geq 1-\delta$. Here, the function $\epsilon(\delta)$ stands for the target accuracy and the parameter δ stands for the required level of confidence.

In contrast to the oracle complexity results obtained in the literature which are only in expectation, the above definition takes into account an user-specified level of confidence δ, with which we could obtain an ϵ-optimal solution of the stochastic optimization problem (1). This helps obtain a fine-grained characterization of the moments/tails of $\mathbb{P}$ on the oracle complexity of SFW algorithms. We next list the specific notion of light and heavy-tailed assumption that we make on the stochastic first-order oracle.

Assumption 1.1 (Sub-Gaussian)**.** For any $x \in \mathbb{R}^d$, the SFO outputs an estimator $G(x, \xi)$ of $\nabla f(x)$ such that $\mathbb{E}[G(x, \xi)] = \nabla f(x)$, and $\mathbb{E}[\exp\{\|G(x, \xi) - \nabla f(x)\|^2/\sigma^2\}] \leq \exp\{1\}$, for some $\sigma^2 > 0$.

Remark 1.1. The sub-Gaussian tail assumption is satisfied by light-tailed distributions arising in practice, and hence has been used widely in statistics and stochastic optimization [49]. It should be note Assumption 1.1 implies by Jensen's inequality that $\mathbb{E}[\|G(x, \xi) - \nabla f(x)\|^2] \leq \sigma^2$, which is used to obtain oracle complexity results for SFW in expectation. However, the converse is not true.

Assumption 1.2 (Weak-exponential)**.** For any $x \in \mathbb{R}^d$, the first-order oracle outputs an estimator $G(x, \xi)$ of $\nabla f(x)$ such that $\mathbb{E}[G(x, \xi)] = \nabla f(x)$, and $\mathbb{E}[\|G(x, \xi)\|^{1+\alpha}] \leq \sigma^{1+\alpha}$ where $\alpha \in (0, 1]$.

Assumption 1.3 (Weak-moment)**.** For any $x \in \mathbb{R}^d$, the first-order oracle outputs an estimator $G(x, \xi)$ of $\nabla f(x)$ such that $\mathbb{E}[G(x, \xi)] = \nabla f(x)$, and for all $\|v\| = 1$, we have $\mathbb{E}[|\langle G(x, \xi) - \nabla f(x), v\rangle|^{1+\beta}] \leq 1$ for some $\beta \in (0, 1]$

Remark 1.2. Firstly, we point out that both Assumptions 1.2 and 1.3 allow for the stochastic gradient to have infinite variance (when $\alpha, \beta < 1$). Furthermore, even when $\alpha, \beta = 1$, Assumptions 1.2 and 1.3 allow for tails heavier than the sub-Gaussian tail in Assumption 1.1. For canonical problems like structured linear or multi-class logistic regression, this assumption allows for the covariate to be heavy-tailed including Pareto or t-distribution, which do not satisfy Assumption 1.1. Furthermore, Assumption 1.2 is stronger than Assumption 1.3. Indeed if $\zeta \in \mathbb{R}^d$ is a random vector that satisfies Assumption 1.3, then we have $\mathbb{E}[\|\zeta\|^{1+\beta}] \leq (\pi/2)\, d^{\frac{1+\beta}{2}}$.

With the above preliminaries, we now present our **main contributions** in this work.

- We first characterize the SFO and LMO complexity, in high-probability, when the standard mini-batch average of the stochastic gradients is used, in Theorem 3.1 under the condition that the distribution of the stochastic gradients follows the sub-Gaussian tail Assumption 1.1.
- We next establish the SFO and LMO complexity, in high-probability, when using the clipped gradient estimators, under the heavy-tailed Assumptions 1.2 and 1.3 in Theorem 3.2 (a) and (b) respectively.

- Next, in Theorem 3.3, we show that the high-probability oracle complexity results under Assumption 1.3 could be further improved by using a gradient estimator based on a recently proposed optimal mean-estimator procedure in [9], at the cost of increased per-iteration complexity.
- Finally, we propose a computationally efficient biased-corrected clipped gradient procedure. We show in Theorem 3.4 that this approach obtains improved high-probability oracle complexities (over Theorem 3.2) results under an additional symmetry condition (see Assumption 3.2) on the distribution of the stochastic gradients, for certain regimes of ϵ and d.

A summary of our oracle complexity results is provided in Table 1 and visual illustrations are provided in Section B. To our knowledge, our work provides the *first comprehensive high-probability oracle complexity results* for SCGS algorithm with light and heavy-tailed stochastic gradients (including infinite variance). We discuss the consequences of our theoretical results for practice by reporting simulations for heavy-tailed sparse linear regression (Section 4) and multi-class logistic regression (Appendix–Section A).

2 STOCHASTIC FRANK-WOLFE ALGORITHM WITH ROBUST MEAN ESTIMATORS

Recall that our task is to solve constrained stochastic convex optimization problems of the form in (1) using projection-free Frank-Wolfe type algorithm, which involves two main steps: the gradient evaluation step and the linear optimization step. The gradient evaluation step is typically based on averaging a mini-batch of gradients [22, 31, 1]. It is well-known from classical robust statistics that the sample average is not an accurate estimate of true expectation in the heavy-tailed setting. Hence, a natural strategy is to replace the sample average with appropriate robust versions of mean estimators, and incorporate such estimators in Frank-Wolfe type algorithms. Specifically, we consider the version of stochastic conditional gradient sliding algorithm in [31], also analyzed in [1] for the case of biased gradients, as it allows one to skip the gradient evaluation from time to time. We first state a subroutine in Algorithm 1 that we subsequently use.

Note that Algorithm 1 is indeed the SFW algorithm for inexactly solving the following quadratic program

$$P_{\mathcal{X}}(x, g, \gamma) = \arg\min_{u \in \mathcal{X}} \left\{ \langle g, u \rangle + \frac{\gamma}{2} \|u - x\|^2 \right\}, \quad (3)$$

which is the standard subproblem of stochastic first-order methods applied to a minimization problem when g is an unbiased stochastic gradient of the objective function at x. We now present Algorithm 2 which applies the Frank-Wolfe method to inexactly solve subproblems of the stochastic accelerated gradient method. It is well known that the above approach can significantly reduce the total number of calls to the stochastic oracle [31, 1].

Algorithm 1 Inexact Conditional Gradient (ICG) method

Input: (x, g, γ, μ).
Set $\bar{y}_0 = x$, $t = 1$, and $\kappa = 0$..
while $\kappa = 0$ **do**
$y_t = \arg\min_{u \in \mathcal{X}} \{h_\gamma(u) := \langle g + \gamma(\bar{y}_{t-1} - \xi), u - \bar{y}_{t-1} \rangle\}$
If $h_\gamma(y_t) \geq -\mu$, set $\kappa = 1$. **Else** $\bar{y}_t = \frac{t-1}{t+1}\bar{y}_{t-1} + \frac{2}{t+1} y_t$ and $t = t + 1$.
end while
Output $\bar{y}_t$.

Algorithm 2 Robust Stochastic Accelerated Gradient Method with Inexact Updates

Input: $x_0 \in \mathcal{X}$, positive integer sequence m_k, and sequences α_k, γ_k, μ_k and iteration limit $N \geq 1$.
for $k = 1, \ldots, N$ **do**
1. Set $w_k = (1 - \alpha_k) z_{k-1} + \alpha_k x_{k-1}$
2. Call the stochastic oracle m_k times, and compute (robust) stochastic gradients $\bar{G}_k$ as given by (4), (5), (6) or (10).
3. Set $x_k = ICG(x_{k-1}, \bar{G}_k, \gamma_k, \mu_k)$, where $ICG(\cdot)$ is the output of Algorithm 1 with input $(x_{k-1}, \bar{G}_k, \gamma_k, \mu_k)$.
4. Set $z_k = (1 - \alpha_k) z_{k-1} + \alpha_k x_k$.
end for
Output: z_N

2.1 ROBUST MEAN ESTIMATORS

Note that step 2 in Algorithm 2 requires the gradient estimator denoted by $\bar{G}_k$. We now elaborate the robust stochastic gradient estimators that we propose to use in this work. Recall that standard analysis of SCGS algorithm takes the sample average of the mini-batch of i.i.d. stochastic gradient (obtained by querying the SFO) in each iteration. In this case, the gradient estimator is given by

$$\bar{G}_k := \frac{1}{m_k} \sum_{j=1}^{m_k} G(w_k, \xi_{k,j}). \quad (4)$$

As we will see from our analysis, the above gradient estimator is not robust to heavy-tails, i.e., when the vectors $G(w_k, \xi_{k,j})$ are heavy-tailed. Our first robust stochastic gradient estimator is based on the idea of trimmed or clipped estimators, generalized to the multivariate setting [3, 47]. More recently such ideas have been used in the context of bandit optimization [4] and stochastic gradient descent algorithm [17, 52, 53]. Formally, in our setting, given i.i.d. stochastic gradients $\{G(w_k, \xi_{k,j})\}_{j=1}^{m_k}$, the *clipped gradient*

estimator is defined, for some $\delta \in (0,1)$,

$$\bar{G}_k := \frac{1}{m_k}\sum_{j=1}^{m_k}\left[G(w_k,\xi_{k,j})\ \mathbb{1}\{A\}\right] \tag{5}$$

$$\text{with} \quad A := \|G(w_k,\xi_{k,j})\| \le \left(\frac{j\sigma^{1+\alpha}}{\log(1/\delta)}\right)^{\frac{1}{1+\alpha}}.$$

Here, $\mathbb{1}\{A\}$, for the event A is defined as taking value 1 when the event A is true and taking value 0 otherwise. While the above estimator handles robust stochastic gradients (including ones with potential infinite variance condition in Assumption 1.2), it turns out that the oracle complexities under the even weaker condition in Assumption 1.3 with the above clipped gradient estimator is sub-optimal. To improve the oracle complexity under Assumption 1.3, we leverage the recent optimal robust mean estimation procedure proposed in [9], which we call as the CTBJ procedure. Given i.i.d. stochastic gradients $\{G(w_k,\xi_{k,j})\}_{j=1}^{m_k}$, the CTBJ estimator is given by (see Algorithm 3)

$$\bar{G}_k := \textsc{optimalmeanest}\left(\{G(w_k,\xi_{k,j})\}_{j=1}^{m_k}\right). \tag{6}$$

Roughly, the idea of [9] is to use filtering to remove outliers and then compute the median by gradient descent procedures. A full description of the procedure is provided in Appendix–Section E for the sake of completeness. We also emphasize that while the estimator for the light-tailed case in (4) is unbiased, the robust gradient estimators in (5) and (6) are biased. This is another challenge that we handle in our analysis. Finally, we also remark that in Section 3.4, we introduce a bias-corrected clipped gradient procedure which achieves improved oracle complexities under an additional symmetry assumption on the distribution on stochastic gradients.

3 HIGH-PROBABILITY BOUND FOR ROBUST STOCHASTIC FRANK-WOLFE

We now provide our main results on high-probability bounds on the oracle complexity of Algorithm 2. To do so, we also make the following standard smoothness assumption about the stochastic gradient, which is common in the literature of smooth convex optimization [38, 31, 1].

Assumption 3.1. The objective function f has Lipschitz continuous gradient with constant $L > 0$, i.e., $\|\nabla f(y) - \nabla f(x)\| \le L\|y - x\|$ for all $x, y \in \mathbb{R}^d$.

We first state a preliminary result about the iterates of Algorithm 2, under Assumption 3.1.

Lemma 3.1. Let $\{z_k\}_{k\ge1}$ be generated by Algorithm 2 and the function f be convex. Let $\bar{\Delta}_k := \bar{G}_k - \nabla f(w_k)$, $\hat{\Gamma}_k := \prod_{i=2}^{k}(1-\alpha_i)$, $\hat{\Gamma}_1 := 1$ and $D_0 = \|x_0 - x_*\|^2$. Then under Assumption 3.1, we have

$$\frac{f(z_N) - f(x_*)}{\hat{\Gamma}_N} \le \frac{\gamma_1}{2}\|x_0 - x_*\|^2 + \sum_{i=1}^{N}\frac{\alpha_k\mu_k}{\hat{\Gamma}_k} + \sum_{i=1}^{N}\frac{\alpha_k}{\hat{\Gamma}_k}\langle\bar{\Delta}_k, x_* - x_{k-1}\rangle + \sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{2L\hat{\Gamma}_k}. \tag{7}$$

For our subsequent analysis, we set

$$\alpha_k = \frac{2}{k+1}, \quad \gamma_k = \frac{4L}{k}, \quad \text{and} \quad \mu_k = \frac{LD_0}{kN}. \tag{8}$$

The proof of Lemma 3.1 is provided in Appendix–Section D. Our high-probability bounds are now based on developing concentration inequalities for the various gradient estimators $\bar{G}_k$ and bounding the terms appearing in right hand side of (7) in high-probability. To do so, we prove novel user-friendly concentration inequalities for (scalar-valued) martingales with heavy-tails that are discussed in detail in Section C. It is worth mentioning that [32] and [12] also consider tail bounds for heavy-tailed martingales. However, they only provide deviation inequalities and their assumptions do no cover the regimes of heavy-tails that we are interested in.

3.1 ORACLE COMPLEXITY WITH SAMPLE AVERAGE ESTIMATOR

We now provide oracle complexity results for Algorithm 2 with the sample average gradient estimator in (4), that hold in high-probability.

Theorem 3.1. *Let Algorithm 2 be run with $\bar{G}_k$ as in* (4)*, and with parameters α_k, γ_k and μ_k as in* (8)*. If the stochastic gradients $G(x,\xi)$ satisfy Assumption 1.1 and 3.1, and $m_k = \mathcal{O}(N^3)$, then*

$$\mathbb{P}\left(f(z_N) - f(x_*) \le \frac{D_0\sigma^2\log(1/\delta)}{N(N+1)}\right) \ge 1 - \delta,$$

and the SFO and LMO complexity are respectively bounded by

$$\mathcal{O}\left(\left(\frac{\log(1/\delta)}{\epsilon}\right)^2\right) \quad \textit{and} \quad \mathcal{O}\left(\frac{\log(1/\delta)}{\epsilon}\right).$$

Remark 3.1. Note that [31] shows that the SFO and LMO oracle complexity for Algorithm 2 with the sample average gradient estimator are of order $\mathcal{O}(1/\epsilon^2)$ and $\mathcal{O}(1/\epsilon)$ in expectation. Our results in Theorem 3.1 generalize this to the high-probability setting quantifying the effect of the allowed confidence level δ precisely.

3.2 ORACLE COMPLEXITY WITH CLIPPED GRADIENT ESTIMATOR

We first provide results on the bias, tail and moment bounds on the clipped gradient estimator $\bar{G}_k$ defined in (5). Then, we provide oracle complexity results for Algorithm 2 with the clipped gradient estimator, that hold with high-probability.

Lemma 3.2. Let $\delta \in (0,1)$ and C be a positive universal constant. Let $\bar{G}_k$ be as defined in (5).

(a) If the stochastic gradients $G(x,\xi)$ satisfy Assumption 1.2, then we have

$$\left\|\mathbb{E}[\bar{G}_k] - \nabla f(w_k)\right\| \leq \sigma \left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\alpha}{1+\alpha}}$$

$$\text{and} \quad \mathbb{P}\left(\left\|\bar{\Delta}_k\right\| \geq 4\sigma \left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\alpha}{1+\alpha}}\right) \leq \delta.$$

Consequently, we also have the following moment bound $\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}\right] \leq C$.

(b) If the stochastic gradients $G(x,\xi)$ satisfy Assumption 1.3, then we have

$$\left\|\mathbb{E}[\bar{G}_k] - \nabla f(w_k)\right\| \leq \sqrt{d} \left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\beta}{1+\beta}}$$

$$\text{and} \quad \mathbb{P}\left(\left\|\bar{\Delta}_k\right\| \geq 4\sqrt{d} \left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\beta}{1+\beta}}\right) \leq \delta.$$

Consequently, we also have the following moment bound $\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sqrt{d}}\right\|^{\frac{1+\beta}{\beta}} m_k\right\}\right] \leq C$.

Theorem 3.2. *Let Algorithm 2 be run with $\bar{G}_k$ as defined in* (5)*, and with parameters α_k, γ_k and μ_k as defined in* (8)*.*

(a) Define $a(\delta,\omega) = (\log(1/\omega))^{\frac{\omega}{1+\omega}}$ and hence $a(\delta,\alpha) = (\log(1/\delta))^{\frac{\alpha}{1+\alpha}}$. If the stochastic gradients $G(x,\xi)$ satisfy Assumptions 1.2 and 3.1 and $m_k = \mathcal{O}(N^{\frac{2(\alpha+1)}{\alpha}})$, then

$$\mathbb{P}\left(f(z_N) - f(x_*) \leq \frac{D_0 \sigma \max\left\{a(\delta,\alpha), \frac{\sigma}{N} a(\delta,\alpha)^2\right\}}{N(N+1)}\right) \geq 1-\delta,$$

and the SFO and LMO complexity are respectively bounded by

$$\mathcal{O}\left(\left(\frac{a(\delta,\alpha)}{\epsilon}\right)^{\frac{3\alpha+2}{2\alpha}}\right) \quad \textit{and} \quad \mathcal{O}\left(\frac{a(\delta,\alpha)}{\epsilon}\right).$$

(b) If the stochastic gradients $G(x,\xi)$ satisfy Assumptions 1.3 and 3.1 and $m_k = \mathcal{O}(N^{\frac{2(\beta+1)}{\beta}})$, then

$$\mathbb{P}\left(f(z_N) - f(x_*) \leq \frac{D_0 \sqrt{d} \max\left\{a(\delta,\beta), \frac{\sqrt{d}}{N} a(\delta,\beta)^2\right\}}{N(N+1)}\right) \geq 1-\delta,$$

and the SFO and LMO complexity are, respectively, bounded by

$$\mathcal{O}\left(\left(\frac{\sqrt{d}a(\delta,\beta)}{\epsilon}\right)^{\frac{3\beta+2}{2\beta}}\right) \quad \textit{and} \quad \mathcal{O}\left(\frac{\sqrt{d}a(\delta,\beta)}{\epsilon}\right),$$

where $a(\delta,\beta) = (\log(1/\delta))^{\frac{\beta}{1+\beta}}$.

Remark 3.2. First note that the oracle complexities under the weaker condition in Assumption 1.3 has an additional dimension factor $\sqrt{d}$. Hence, for a fixed value of δ, for $\alpha = 1$ and $\beta = 1$ (i.e., finite variance case), we have the SFO complexity to be of order $\mathcal{O}(\epsilon^{-5/2})$ and $\mathcal{O}(d^{5/4}\epsilon^{-5/2})$ respectively. Furthermore, note that under our assumptions, only $(1+\alpha)$ or $(1+\beta)$ moment exists for the stochastic gradients. Consequently, as $\alpha \to 0$ or $\beta \to 0$, for a fixed value of $0 < \epsilon < 1$ and δ, the SFO complexity tends to infinity, highlighting the difficulty of the problem.

3.3 ORACLE COMPLEXITY WITH CTBJ-BASED GRADIENT ESTIMATOR

In this section, we will use the mean estimator procedure proposed recently in [9], and show that the dimension factor $\sqrt{d}$ appearing in the SFO complexity in part (b) of Theorem 3.2 could be removed, even under the weaker condition in Assumption 1.3.

Theorem 3.3. *Let Algorithm 2 be run with $\bar{G}_k$ as defined in* (6)*, and with parameters α_k, γ_k and μ_k as defined in* (8)*. If the stochastic gradients $G(x,\xi)$ satisfy Assumptions 1.3 and 3.1, and $m_k = \mathcal{O}(N^{\frac{2(\beta+1)}{\beta}})$, then with a target confidence $\delta > 2^{-\frac{m_k}{16000}}$ and $d \lesssim \log(1/\delta)$, we have*

$$\mathbb{P}\left(f(z_N) - f(x_*) \leq \frac{D_0 \max\left\{a(\delta,\beta), \frac{1}{N} a(\delta,\beta)^2\right\}}{N(N+1)}\right) \geq 1-\delta,$$

and the SFO and LMO complexity are respectively bounded by

$$\mathcal{O}\left(\left(\frac{a(\delta,\beta)}{\epsilon}\right)^{\frac{3\beta+2}{2\beta}}\right) \quad \textit{and} \quad \mathcal{O}\left(\frac{a(\delta,\beta)}{\epsilon}\right).$$

Remark 3.3. The proof of the above theorem is based on a concentration result for the gradient estimator (6), established in [9]. In comparison to part (b) of Theorem 3.2,

the SFO complexity in Theorem 3.3 under Assumption 1.3 does not have the additional dimensional factor $\sqrt{d}$, thereby demonstrating the benefits of using the mean-estimation procedure proposed in [9]. However, this improvement does not come for free, as the per-iteration complexity of using the robust mean-estimator (6) is significantly higher than that of the clipped gradient based robust mean-estimator in (5), although it has a polynomial dependency on the problem parameters. See Section E for details regarding the per-iteration computational complexity of (6).

3.4 IMPROVING THE ORACLE COMPLEXITY WITH BIAS-CORRECTED CLIPPED GRADIENTS

The above oracle complexity results based on clipped gradient based and CTBJ based robust gradient estimators have the following drawbacks. The ϵ-dependency of the SFO complexity in part (a) of Theorem 3.2 under Assumption 1.2 or Theorem 3.3 under Assumption 1.3 is $\mathcal{O}(\epsilon^{-2.5})$ when $\alpha = 1$ and $\beta = 1$. In this section, we propose a bias-corrected clipped gradient based robust gradient estimation procedure with which SFO complexity of Algorithm 2 could be improved to $\mathcal{O}(d^2\epsilon^{-2})$ under an additional symmetry assumption on the distribution of the stochastic gradient. Hence, when high-accuracy solutions in low-dimensional settings are required, the bias-corrected clipped gradient based robust SFW algorithm could be preferred. We now introduce the symmetry assumption and the clipped gradient procedure.

Assumption 3.2. Let the distribution of $G(x,\xi)$ be continuous and symmetric about $\mathbb{E}[G(x,\xi)] = \nabla f(x)$, for all $x \in \mathbb{R}^d$. Also, let the probability density function be a decreasing function with respect to $\|G(x,\xi) - \nabla f(x)\|$, for all $x \in \mathbb{R}^d$.

Proposition 3.1. *For iteration k the i.i.d. stochastic gradients $G(w_k, \xi_{k,j})$ are assumed to satisfy Assumptions 1.3 and 3.2. Let $\delta \in (0,1)$. Consider the initial estimate defined as*

$$\widehat{G}_k := \underset{G(w_k,\xi_{k,j}):j\geq \frac{m_k}{2}}{\arg\min} \min\Bigg\{ r \geq 0 : \sum_{\ell=\frac{m_k}{2}}^{m_k} \mathbb{1}\left\{\|G(w_k,\xi_{k,\ell}) - G(w_k,\xi_{k,j})\| \leq r\right\} \geq 0.3 m_k \Bigg\}, \tag{9}$$

and consider the bias-corrected clipped gradient estimator $\bar{G}_k$ defined as

$$\bar{G}_k := \frac{2}{m_k}\sum_{t=1}^{m_k/2} \min\left\{ \frac{\left[\left(\frac{t}{\log(1/\delta)}\right)^{\frac{1}{1+\beta}} + 24\right]\sqrt{d}}{\|G(w_k,\xi_{k,t}) - \widehat{G}_k\|}, 1 \right\} \times \left[G(w_k,\xi_{k,t}) - \widehat{G}_k)\right] + \widehat{G}_k. \tag{10}$$

Then, as long as $m_k \geq 2\max\{50, 24^{1+\beta}\}\log(1/\delta)$, by recalling that $\bar{\Delta}_k = \bar{G}_k - \nabla f(w_k)$, we have

$$\mathbb{E}[\bar{G}_k] = \nabla f(w_k) \quad \textit{and}$$

$$\mathbb{P}\left\{\|\bar{\Delta}_k\| \leq 8\pi\sqrt{d}\left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\beta}{1+\beta}}\right\} \geq 1-\delta.$$

Remark 3.4. The initial estimate defined in (9) is the same as that in the CTBJ estimator in (6). We show that this initial step, along with the clipped gradient procedure for a specific choice of clipping parameter (as defined in (10)) helps obtain an unbiased gradient estimator which is sufficiently concentrated.

We now leverage the result in Proposition 3.1 and show that one could obtain improved SFO complexity for certain ranges of d and ϵ when running Algorithm 2 with the robust gradient estimator (10).

Theorem 3.4. *Let Algorithm 2 be run with $\bar{G}_k$ as defined in* (10)*, and with parameters α_k, γ_k and μ_k as defined in* (8).

(a) If the stochastic gradients $G(x,\xi)$ satisfy Assumptions 1.2, 3.1 and 3.2, and $m_k = \mathcal{O}(N^{\frac{3(\alpha+1)}{2\alpha}})$, we have for $(\log(1/\delta))^{\frac{1}{1+\alpha}} \geq [\Gamma(\frac{\alpha}{1+\alpha})\frac{1+\alpha}{\alpha}]^{\frac{1}{1-\alpha}}$,

$$\mathbb{P}\left(f(z_N) - f(x_*) \geq \frac{C(d,\alpha,\delta)D_0\sigma}{N(N+1)}\right) \leq \delta, \tag{11}$$

where

$$C(d,\alpha,\delta) = \max\left\{\sqrt{d}a(\delta,\alpha)^{\frac{1}{\alpha}}, a(\delta,\alpha)^2\right\},$$

and the SFO and LMO complexity are respectively bounded by

$$\mathcal{O}\left(\left(\frac{C(d,\alpha,\delta)}{\epsilon}\right)^{\frac{5\alpha+3}{4\alpha}}\right) \quad \textit{and} \quad \mathcal{O}\left(\frac{C(d,\alpha,\delta)}{\epsilon}\right).$$

(b) If the stochastic gradients $G(x,\xi)$ satisfy Assumptions 1.2,3.1 and 3.2, and $m_k = \mathcal{O}(N^{\frac{3(\beta+1)}{2\beta}})$, the same conclusion in (11) *holds with α replaced by β, with $C(d,\beta,\delta) := C(d,\alpha=\beta,\delta)$, for $(\log(1/\delta))^{\frac{1}{1+\beta}} \geq [\Gamma(\frac{\beta}{1+\beta})\frac{1+\beta}{\beta}]^{\frac{1}{1-\beta}}$.*

Remark 3.5. Note that when $\alpha = 1$, for any fixed value of δ, $C(d,\alpha,\delta)$ is $\mathcal{O}(d)$. Hence, the SFO complexity is

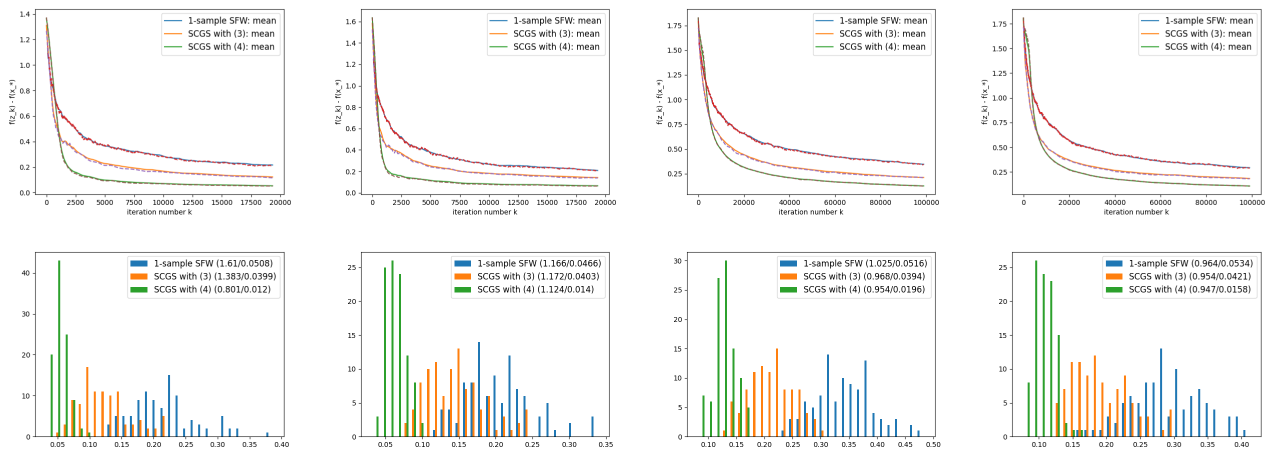

Figure 1: The two left and two right columns correspond to Pareto, Student-t distributions with $d = 100$ and $d = 500$ respectively. **Top row:** Mean (solid lines) and Median (dotted lines) over 100 trails of iterations versus $f(z_N) - f(x_*)$ for $N = 100$. **Bottom row:** Histogram of $f(z_N) - f(x_*)$ for $N = 100$. Numbers in the legend correspond to *heavy-tail index/standard deviation*.

of order $\mathcal{O}(d^2\epsilon^{-2})$. Hence, when $d^2 < o(\epsilon^{-0.5})$ the SFO complexity of part (a) of Theorem 3.4 is better than part (a) of Theorem 3.2. A similar improvement holds for part (b). In Figure 4 (Appendix–Section B), we compare Theorem 3.2, and Theorem 3.4 visually. For comparing part (a) of the respective theorems, we set $\alpha = 1$, $\epsilon = 10^{-10}$ and $\delta = 0.05$ vary d from 200 to 1000 in steps of two hundred. For comparing part (b) of the respective theorems, we set $\alpha = 1$, $\epsilon = 10^{-6}$ and $\delta = 0.05$ vary d from 2000 to 10000 in steps of two thousand.

4 CONSEQUENCES FOR PRACTICE AND OUTLOOK

We now demonstrate the consequences of our theoretical results for practice. Among the robust gradient estimators in (5), (6) and (10), the *most practical one* (i.e., least per-iteration complexity) is the clipped gradient estimator in (5). Hence, we compare Algorithm 2 with the mini-batch average gradient in (4) and the clipped gradient estimator in (5) via experiments on sparse linear regression and multi-class logistic regression (presented in Appendix–Section A due to space limitations) with heavy-tailed covariates.

Sparse Linear Regression: We now provide simulation results for the regression problem described in (2). For our experiments, we consider the data vector $a \in \mathbb{R}^d$ to be a t-distribution or Pareto distribution with degrees of freedom 2.2. We ran Algorithm 2 with parameters as defined in (8) for 100 trails. Here, L could be calculated analytically for our problem. For the choice of batch size, while Theorem 3.2 suggests $m_k = \mathcal{O}(N^{\frac{2(\alpha+1)}{\alpha}})$, we found that in our experiments setting $m_k = 500$ works well already. In Figure 1, we report the performance of Algorithm 2 with the clipped gradient estimator (5) and mini-batch average estimator (4). We also compare against the 1-sample SFW method from [35]. From the top row, we see that the clipped gradient estimator has faster convergence, i.e., it achieves higher accuracy with lesser iterations compared to the standard mini-batch averaging or the 1-sample SFW method. Furthermore, from the histogram in the bottom row, we see that the distribution of the last iterate has more fluctuations for the mini-batch average estimator and the 1-sample SFW method, compared to the clipped gradient estimator. We quantify this statement by reporting the *standard deviation* and also the *heavy-tailed index*, a widely used metric to quantify fluctuations [23], which is defined as

$$\tau(F) = \frac{F^{-1}(0.95) - F^{-1}(0.5)}{F^{-1}(0.75) - F^{-1}(0.5)} \Big/ \frac{\Phi^{-1}(0.95) - \Phi^{-1}(0.5)}{\Phi^{-1}(0.75) - \Phi^{-1}(0.5)},$$

where Φ is the distribution of a standard normal and F is the empirical CDF obtained from the histogram. The results in Figure 1 **confirm our theoretical results and highlight the benefits of using robust versions of SFW algorithms for dealing with heavy-tailed data arising practice**.

5 SUMMARY

We proposed and analyzed robust versions of stochastic Frank-Wolfe type algorithms and established high-probability oracle complexity results. Our theoretical results are supported by numerical experiments on the problem of sparse linear regression and multi-class logistic regression with heavy-tailed data. Developing oracle complexity results for robust projection-free algorithms under non-convexity, and developing more practical versions of robust projection-free algorithms are interesting problems that we plan to examine as future work.

REFERENCES

[1] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 2021.

[2] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 2017.

[3] P Bickel. On some robust estimates of location. *Annals of Mathematical Statistics*, 1965.

[4] Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[5] Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in gaussian noise injections. *arXiv preprint arXiv:2102.07006*, 2021.

[6] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, 2012.

[7] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[8] Han-fu Chen and Ai-Jun Gao. Robustness analysis for stochastic approximation algorithms. *Stochastics and Stochastic Reports*, 26(1):3–20, 1989.

[9] Yeshwanth Cherapanamjeri, Nilesh Tripuraneni, Peter L Bartlett, and Michael I Jordan. Optimal mean estimation without a variance. *arXiv preprint arXiv:2011.12433*, 2020.

[10] Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38, 2021.

[11] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 2014.

[12] X Fan, I Grama, and Q Liu. Deviation inequalities for martingales with applications. *Journal of Mathematical Analysis and Applications*, 2017.

[13] Xiequan Fan, Ion Grama, Quansheng Liu, et al. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 2015.

[14] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[15] R Freund, P Grigas, and R Mazumder. Extended Frank–Wolfe method with in-face directions, and its application to low-rank matrix completion. *SIAM Journal on optimization*, 2017.

[16] Dan Garber, Shoham Sabach, and Atara Kaplan. Fast generalized conditional gradient method with applications to matrix recovery problems. *arXiv preprint arXiv:1802.05581*, 2018.

[17] Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Advances in Neural Information Processing Systems*, volume 33, pages 15042–15053, 2020.

[18] Mert Gurbuzbalaban and Yuanhan Hu. Fractional moment-preserving initialization schemes for training deep neural networks. In *AISTATS*, 2021.

[19] Z Harchaoui, A Juditsky, and A Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[20] Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Zebang Shen. Stochastic conditional gradient++:(non) convex minimization and continuous submodular maximization. *SIAM Journal on Optimization*, 30(4):3315–3344, 2020.

[21] Elad Hazan and Satyen Kale. Projection-free online learning. In *ICML 2012*, pages 521–528, 2012.

[22] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *ICML*, 2016.

[23] David C Hoaglin, Frederick Mosteller, and John W Tukey. *Understanding robust and exploratory data analysis*. Wiley Classics Series, 2000.

[24] Liam Hodgkinson and Michael Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *ICML*, 2021.

[25] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[26] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

[27] Tatiana Pavlovna Krasulina. On stochastic approximation processes with infinite variance. *Theory of Probability & Its Applications*, 14(3):522–526, 1969.

[28] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504, 2015.

[29] Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552, 2015.

[30] Guanghui Lan, Arkadi Nemirovski, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

[31] Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization*, 26(2):1379–1409, 2016.

[32] Emmanuel Lesigne and Dalibor Volnỳ. Large deviations for martingales. *Stochastic processes and their applications*, 96(1):143–159, 2001.

[33] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[34] Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[35] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 2020.

[36] Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 2019.

[37] Arkadij Nemirovskij and David Yudin. Problem complexity and method efficiency in optimization. 1983.

[38] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[39] Abhishek Panigrahi, Raghav Somani, Navin Goyal, and Praneeth Netrapalli. Non-gaussianity of stochastic gradient noise. *SEDL workshop, NeurIPS*, 2019.

[40] A Prasad, A S Suggala, S Balakrishnan, and P Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B*, 2020.

[41] Sashank J Reddi, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic frank-wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016.

[42] Sidney I Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.

[43] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 525–536, 2018.

[44] U Simsekli, O Sener, G Deligiannidis, and M A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[45] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *ICML*, 2019.

[46] Umut Simsekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gurbuzbalaban. Fractional underdamped langevin dynamics: Retargeting sgd with momentum under heavy-tailed gradient noise. In *ICML*, 2020.

[47] Stephen M Stigler. The asymptotic distribution of the trimmed mean. *Annals of Statistics*, 1973.

[48] N.N. Taleb. *Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications*. Technical Incerto. 2020.

[49] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[50] Hongjian Wang, Mert Gürbüzbalaban, Lingjiong Zhu, Umut Şimşekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *arXiv preprint arXiv:2102.10346*, 2021.

[51] Alp Yurtsever, Suvrit Sra, and Volkan Cevher. Conditional gradient methods via stochastic path-integrated differential estimator. In *ICML*, 2019.

[52] Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for nonconvex optimization. In *NeurIPS*, 2020.

[53] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33, 2020.

[54] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4012–4023. PMLR, 2020.

Supplementary Material

A SIMULATION RESULTS FOR MULTICLASS LOGISTIC REGRESSION EXPERIMENTS

In this section, we consider the problem of multi-class logistic regression. Given $(a, y) \in \mathbb{R}^d \times \mathbb{R}$, consider the multi-class logistic model

$$\mathbb{P}(y|a) = \exp(a^\top \bar{x}_y) / \sum \exp(a^\top \bar{x}_l).$$

Here, the true parameter $\bar{X} = \{\bar{x}_1, \cdots, \bar{x}_L\}^\top \in \mathbb{R}^{L \times d}$ is assumed to be with bounded trace norm. The trace norm constrained estimator is then given by

$$\arg \min_{x \in \mathcal{X}_{TR}(\tau)} \mathbb{E}[\log\big(\sum_l \exp(a^\top x_l - a^\top x_y)\big)],$$

where $\mathcal{X}_{TR}(\tau) := \{X \in \mathbb{R}^{L \times d} : \sum_{j=1}^d \sigma_j(X) \leq \tau\}$ is the $\|.\|_{tr}$ ball of radius τ. This problem fits in the setup of (1) with $\xi := (a, y)$ and $F(X, \xi) := \log\big(\sum_l \exp(a^\top x_l - a^\top x_y)\big)$. Hence the stochastic gradient $G(X, \xi) = \nabla F(X, \xi) = \{\nabla_1 F(X, \xi), \cdots, \nabla_L F(X, \xi)\} \in \mathbb{R}^{L \times d}$ where

$$\nabla_l F(X, \xi) = [\exp(a^\top \bar{x}_y) / \sum \exp(a^\top \bar{x}_l)] \mathbb{1}\{y \neq l\} a.$$

Note that as the iterates of Algorithm 2 is in the set $\mathcal{X}_{tr}(\tau)$, we have $\|x\|$ is to be always bounded for all x along the trajectory of Algorithm 2. Hence, the $(1+\alpha)$-th moment of the stochastic gradient, i.e., $\mathbb{E}[\|G(x, \xi)\|^{(1+\alpha)}]$, is controlled by the order of $\mathbb{E}[\|a\|^{(1+\alpha)}]$. When the covariate a is a zero-mean multivariate t-distribution with degrees of freedom in the interval $[1, 2)$, or is a zero-mean multivariate Pareto distribution with parameter in the interval $[1, 2)$, the stochastic gradients have infinite variance but finite $(1+\alpha)$-th moment. In other words, Assumption 1.2 is satisfied, while Assumption 1.1 is not.

For our experiments, we select the degrees of freedom of t-distribution and the parameter of Pareto distribution to be 1.1. We ran Algorithm 2 with parameters as defined in (8) for 100 trails. We report the results in Figure 2. We report the performance of Algorithm 2 with the clipped gradient estimator (5) and mini-batch average estimator (4). In our experiments with multi-class logistic regression, we observe a similar performance as in the linear regression setting – clipped gradient method performed the best.

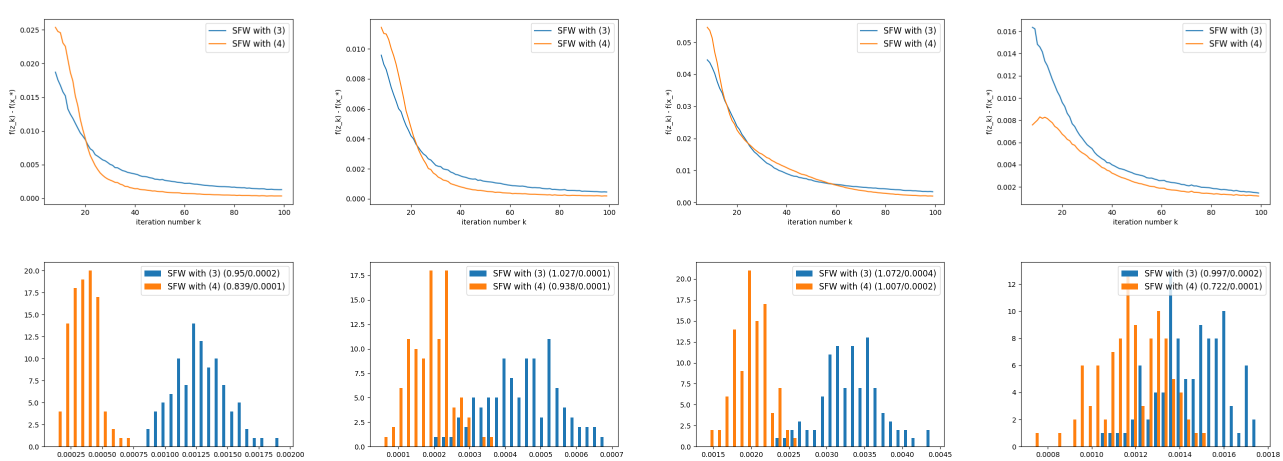

Figure 2: The two left and two right columns corresponds to Pareto, Student-t distributions with $d = 20/L = 10$ and $d = 100/L = 20$ respectively. **Top row:** Mean (solid lines) over 100 trails of iterations versus $f(z_N) - f(x_*)$ for $N = 100$. **Bottom row:** Histogram of $f(z_N) - f(x_*)$ for $N = 100$. Numbers in the legend correspond to *heavy-tail index/standard deviation*.

B PLOTS ILLUSTRATING MAIN THEORETICAL RESULTS

In Figure 3, we provide an illustration of the SFO complexity from Theorem 3.2. We set $\delta = 0.05$. We split the scale of ϵ and α from $(0.2, 0.5)$ and $(0.4, 0.9)$ for better visualization with the large difference in the scale of the vertical (SFO) axis. We set $d = 100$ for part (b).

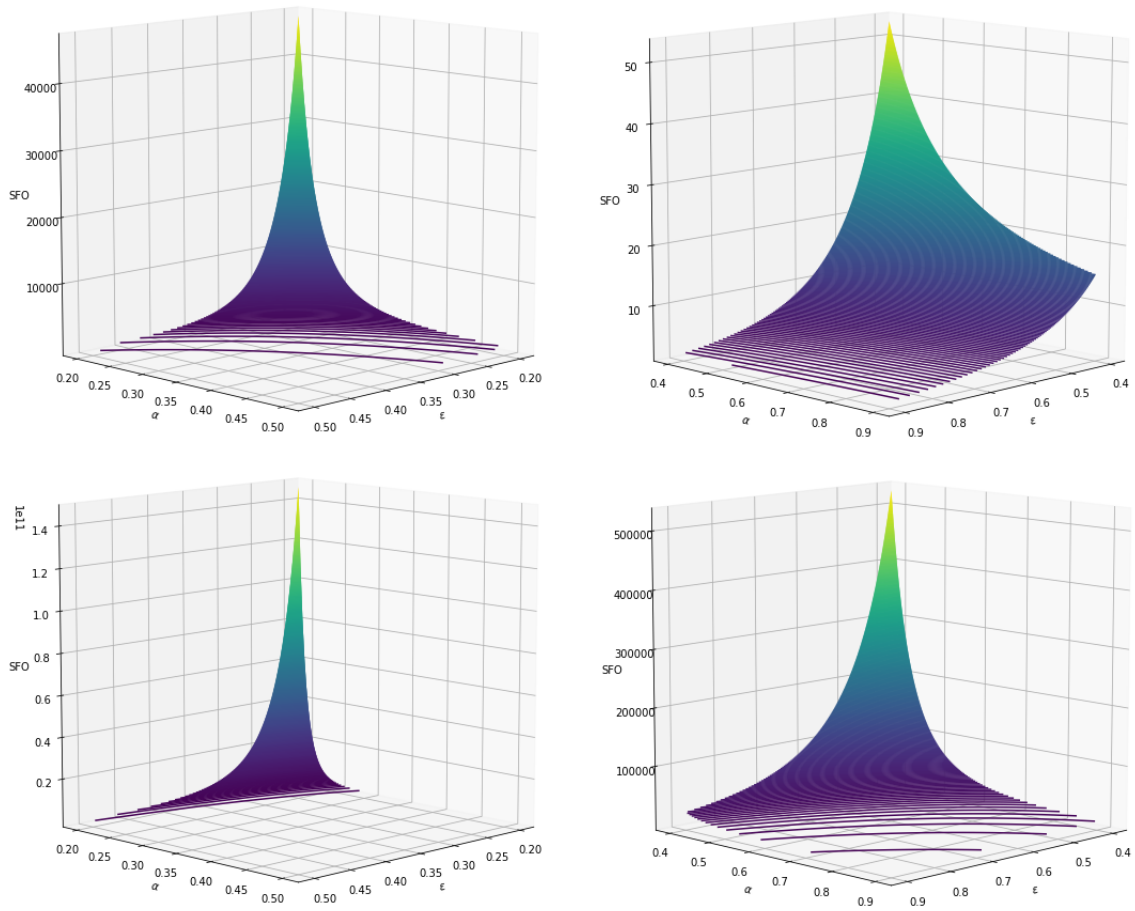

Figure 3: Visualization of the SFO complexity from Theorem 3.2 part (a) **(top row)** and part (b) **(bottom row)**. Note that for smaller values of ϵ (i.e, higher accuracy) and lower values of α (i.e., when only a smaller order moment exists for the stochastic gradients) the SFO complexity increases rapidly. Note the effect of dimension is pronounced in the bottom row corresponding to part (b).

C CONCENTRATION INEQUALITY FOR MARTINGALES WITH HEAVY-TAILS

We first start with two assumptions that turn out to be equivalent.

Assumption C.1. The random variable $X \in \mathbb{R}$ satisfies

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{C_1}{\vartheta^2}\, t^{\frac{1+\alpha}{\alpha}}\right),$$

for some $\vartheta^2, C_1 > 0$ with $\alpha \in (0, 1]$, for all $t \geq 0$.

Assumption C.2. The random variable $X \in \mathbb{R}$ satisfies

$$\mathbb{E}\left[\exp\left(C_2 |X|^{\frac{1+\alpha}{\alpha}} \frac{1}{\vartheta^2}\right)\right] \leq 2,$$

for some $\vartheta^2, C_2 > 0$ with $\alpha \in (0, 1]$.

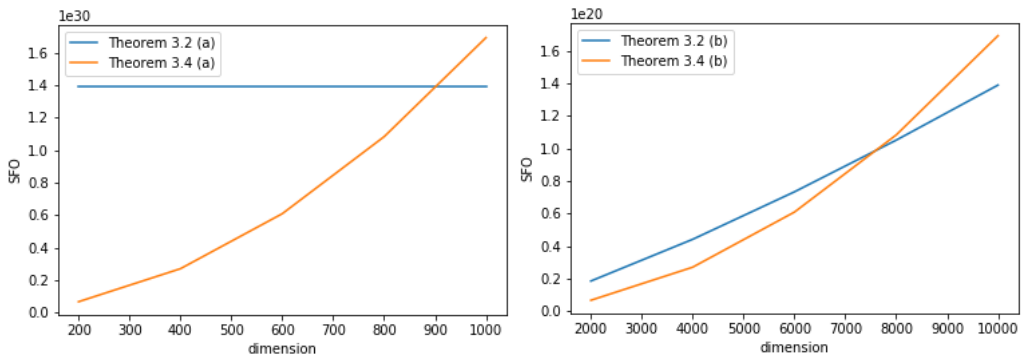

Figure 4: Comparing the SFO complexity in Theorem 3.2 and Theorem 3.4. Part (a) on the left and part (b) on the right. Note that in both cases, there is a certain threshold in dimension below which the SFO complexity of Theorem 3.4 is better than that of Theorem 3.2.

.

Lemma C.1. Assumptions C.1 and C.2 are equivalent.

Proof. Suppose X satisfies Assumption C.1 and assume $C_2 < C_1$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(C_2|X|^{\frac{1+\alpha}{\alpha}}\frac{1}{\vartheta^2}\right)\right] &\leq 1 + C_2\int_0^\infty \frac{1+\alpha}{\alpha}t^{\frac{1}{\alpha}}\frac{1}{\vartheta^2}\exp\left(C_2 t^{\frac{1+\alpha}{\alpha}}/\vartheta^2\right)\mathbb{P}(|X|>t)dt \\
&\leq 1 + 2C_2\int_0^\infty \frac{1+\alpha}{\alpha}t^{\frac{1}{\alpha}}\frac{1}{\vartheta^2}\exp\left(-(C_1-C_2)t^{\frac{1+\alpha}{\alpha}}/\vartheta^2\right)dt \\
&= 1 + 2\frac{C_2}{C_1-C_2}.
\end{aligned}$$

Then, by taking $C_2 \leq C_1/3$, we obtain $\mathbb{E}[\exp(C_2|X|^{\frac{1+\alpha}{\alpha}}\frac{1}{\vartheta^2})] \leq 2$. This completes one direction of the equivalence.

Now, suppose X satisfies Assumption C.2 and assume $C_2 = 1$, then

$$\mathbb{P}(|X|\geq t) = \mathbb{P}\left(\exp\left(|X|^{\frac{1+\alpha}{\alpha}}/\vartheta^2\right) \geq \exp\left(t^{\frac{1+\alpha}{\alpha}}/\vartheta^2\right)\right) \leq 2\exp\left(-t^{\frac{1+\alpha}{\alpha}}/\vartheta^2\right).$$

This proves the other direction of the equivalence, thereby completing the proof. □

Lemma C.2. Let $\Gamma(x)$ denote the gamma function which is defined via a convergent improper integral: $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$. For a random variable that satisfies Assumption C.1, the following properties hold:

(a) For some positive constant C_3, the moments satisfy

$$\mathbb{E}|X|^k \leq (2\vartheta^2)^{\frac{\alpha k}{\alpha+1}}\frac{\alpha k}{\alpha+1}\Gamma\left(\frac{\alpha k}{\alpha+1}\right), \quad \text{and} \quad (\mathbb{E}|X|^k)^{1/k} \leq C_3(\vartheta^{2k})^{\frac{\alpha}{1+\alpha}}, \text{for } k \geq 1.$$

(b) For some positive constant C_4, when $\alpha + 1/\alpha \in \mathbb{N}$, we have

$$\mathbb{E}[\exp(tX)] \leq \left(1 + C_4(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{\alpha}{1+\alpha}}\right)\exp\left(t^{\frac{1+\alpha}{\alpha}}\vartheta^2\right).$$

Furthermore, if $\mathbb{E}[X] = 0$, we have

$$\mathbb{E}[\exp(tX)] \leq \left(1 + C_4(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{2\alpha}{1+\alpha}}\right)\exp\left(t^{\frac{1+\alpha}{\alpha}}\vartheta^2\right).$$

Proof. For part (a), without loss of generality assume that $C_1 = 1$. Then, we have

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k \geq t)dt \\
&= \int_0^\infty \mathbb{P}(|X| \geq t^{\frac{1}{k}})dt \\
&\leq 2\int_0^\infty \exp\left(-t^{\frac{1+\alpha}{\alpha k}}/\vartheta^2\right)dt \\
&= (2\vartheta^2)^{\frac{\alpha k}{\alpha+1}} \frac{\alpha k}{\alpha+1}\int_0^\infty e^{-u}u^{\frac{\alpha k}{\alpha+1}-1}du \\
&= (2\vartheta^2)^{\frac{\alpha k}{\alpha+1}} \frac{\alpha k}{\alpha+1}\Gamma\left(\frac{\alpha k}{\alpha+1}\right).
\end{aligned}$$

Then, by the elementary facts that

$$\Gamma\big(\frac{\alpha k}{\alpha+1}\big) \leq \big(\frac{\alpha k}{\alpha+1}\big)^{\frac{\alpha k}{\alpha+1}} \quad \text{and} \quad k^{1/k} \leq e^{1/e},$$

we have that for any $k \geq 2$,

$$(\mathbb{E}[|X|^k])^{1/k} \leq (\vartheta^2)^{\frac{\alpha}{\alpha+1}}\left(\frac{\alpha k}{\alpha+1}\right)^{\frac{\alpha}{\alpha+1}} e^{1/e} \leq C(\vartheta^2 k)^{\frac{\alpha}{\alpha+1}},$$

which completes the proof of part (a). We now show that part(a) implies part (b). To do so, first note that

$$\begin{aligned}
\mathbb{E}[\exp(tX)] &\leq 1 + \sum_{k=1}^\infty \frac{t^k\mathbb{E}[|X|^k]}{k!} \\
&\leq 1 + \sum_{k=1}^\infty \frac{t^k(\vartheta^2)^{\frac{\alpha k}{\alpha+1}}\frac{\alpha k}{\alpha+1}\Gamma(\frac{\alpha k}{\alpha+1})}{k!} \\
&\leq 1 + \sum_{k=1}^\infty \frac{(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{\frac{\alpha k}{\alpha+1}}k\Gamma(\frac{\alpha k}{\alpha+1})}{k!} \\
&= 1 + \sum_{k=1}^{\frac{\alpha+1}{\alpha}-1} \frac{(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{\frac{\alpha k}{\alpha+1}}k\Gamma(\frac{\alpha k}{\alpha+1})}{k!} + \sum_{j=0}^{\frac{\alpha+1}{\alpha}-1}\sum_{k=1}^\infty \frac{(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{k+\frac{\alpha j}{1+\alpha}}(\frac{(\alpha+1)k}{\alpha}+j)\Gamma(k+\frac{\alpha j}{1+\alpha})}{(\frac{(\alpha+1)k}{\alpha}+j)!} \\
&\leq 1 + \Gamma\big(\frac{\alpha}{1+\alpha}\big)\sum_{k=1}^{\frac{\alpha+1}{\alpha}-1}(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{\frac{\alpha k}{1+\alpha}} + \sum_{j=0}^{\frac{\alpha+1}{\alpha}-1}(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{\frac{\alpha j}{1+\alpha}}\sum_{k=1}^\infty \frac{(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^k k!}{(\frac{(\alpha+1)k}{\alpha})!} \\
&\leq 1 + \Gamma\big(\frac{\alpha}{1+\alpha}\big)\frac{(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{\alpha}{1+\alpha}}(1-t^{\frac{1+\alpha}{\alpha}}\vartheta^2)}{1-(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{\alpha}{1+\alpha}}} + \frac{\alpha}{1+\alpha}\sum_{j=0}^{\frac{\alpha+1}{\alpha}-1}(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^{\frac{\alpha j}{1+\alpha}}\sum_{k=1}^\infty \frac{(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)^k}{k!} \\
&\leq 1 + \Gamma\big(\frac{\alpha}{1+\alpha}\big)\frac{1+\alpha}{\alpha}(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{\alpha}{1+\alpha}} + (\exp(t^{\frac{\alpha+1}{\alpha}}\vartheta^2)-1) \\
&\leq (1 + C_4(t^{\frac{1+\alpha}{\alpha}}\vartheta^2)^{\frac{\alpha}{1+\alpha}})\exp(t^{\frac{\alpha+1}{\alpha}}\vartheta^2),
\end{aligned}$$

where $C_4 = \Gamma\left(\frac{\alpha}{1+\alpha}\right)\frac{1+\alpha}{\alpha}$, thereby completing the proof. The second claim in part (b) follows immediately. □

We now state our concentration inequality for heavy-tailed martingales.

Proposition C.1. *Suppose a sequence of random variables $\{X_k\}_{k=1}^\infty$ satisfies, for $\alpha \in (0,1]$,*

$$\mathbb{E}\left[\exp(tX_k)|X_1,\ldots,X_{k-1}\right] \leq \left(1 + C\left(t^{\frac{1+\alpha}{\alpha}}\vartheta_{k-1}^2\right)^{\frac{\alpha}{1+\alpha}}\right)\exp\left(t^{\frac{\alpha+1}{\alpha}}\vartheta_{k-1}^2\right).$$

If we assume that $\vartheta_i^2 \le n^{-\frac{\alpha+1}{\alpha}}$ *for all i, then, we have*

$$\mathbb{P}\left(\sum_{k=1}^n X_k \ge \lambda\right) \le \exp\left(-\tfrac{1}{\alpha+1}\left(\tfrac{\alpha}{\alpha+1}\right)^\alpha (\lambda - C)^{1+\alpha} n\right).$$

If we further have $\mathbb{E}[X_k|X_1,\ldots,X_{k-1}] = 0$*, and* $\vartheta_i^2 \le n^{-\frac{\alpha+1}{2\alpha}}$ *for all i, then*

$$\mathbb{P}\left(\sum_{k=1}^n X_k \ge \lambda\right) \le \exp(-C_\alpha \lambda^{1+\alpha}) \quad \textit{when} \quad \lambda \ge \left[\Gamma\left(\frac{\alpha}{1+\alpha}\right)\frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}}.$$

where C_α *is as defined in* (26).

Proof. First note that we have the following expression for the moment generating function for the sum:

$$\begin{aligned}
& \mathbb{E}_{X_1,\ldots,X_n}\left[\exp\left(t\sum_{k=1}^n X_k\right)\right] \\
= \; & \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\mathbb{E}_{X_n}\left[\exp\left(t\sum_{k=1}^n X_k\right)\middle| X_1,\ldots,X_{n-1}\right]\right] \\
= \; & \mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(t\sum_{k=1}^{n-1} X_k\right)\mathbb{E}_{X_n}\left[\exp(tX_n)|X_1,\ldots,X_{n-1}\right]\right] \\
\le \; & \left(1 + C\left(t^{\frac{1+\alpha}{\alpha}}\vartheta_n^2\right)^{\frac{\alpha}{1+\alpha}}\right)\exp\left(t^{\frac{1+\alpha}{\alpha}}/\vartheta_n^2\right)\mathbb{E}_{X_1,\ldots,X_{n-1}}\left[\exp\left(t\sum_{k=1}^{n-1} X_k\right)\right].
\end{aligned}$$

By repeatedly performing the above calculation for the term on the right hand side of the last inequality, we obtain

$$\mathbb{E}_{X_1,\ldots,X_n}\left[\exp\left(t\sum_{k=1}^n X_k\right)\right] \le \exp\left(t^{\frac{1+\alpha}{\alpha}}\sum_{k=1}^n \vartheta_n^2\right)\prod_{k=1}^n (1 + C(t^{\frac{1+\alpha}{\alpha}}\vartheta_k^2)^{\frac{\alpha}{1+\alpha}}).$$

Hence, by Markov's inequality and by our assumption that $\vartheta_i^2 \le n^{-\frac{\alpha+1}{\alpha}}$ for all i, we obtain

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^n X_k \ge \lambda\right) &= \mathbb{P}\left(\exp\left(t\sum_{k=1}^n X_k\right) \ge \exp(\lambda t)\right) \\
&\le \exp\left(t^{\frac{1+\alpha}{\alpha}}\sum_{k=1}^n \vartheta_n^2 - \lambda t\right)\prod_{k=1}^n\left(1 + Ct\vartheta_k^{\frac{2\alpha}{1+\alpha}}\right) \\
&\le \exp\left(t^{\frac{1+\alpha}{\alpha}} n^{-\frac{1}{\alpha}} - \lambda t + Ct\right) \\
&\le \exp\left(-\tfrac{1}{\alpha+1}\left(\tfrac{\alpha}{\alpha+1}\right)^\alpha(\lambda - C)^{1+\alpha} n\right)
\end{aligned}$$

where in the last step we set $t = \left(\frac{\alpha}{\alpha+1}(\lambda - C)n^{\frac{1}{\alpha}}\right)^\alpha$. This proves the first claim. Now, when $\mathbb{E}[X_k|X_1,\ldots,X_{k-1}] = 0$ and $\vartheta_i^2 \le n^{-\frac{\alpha+1}{2\alpha}}$, we have

$$\begin{aligned}
\mathbb{P}\left(\sum_{k=1}^n X_k \ge \lambda\right) &= \mathbb{P}\left(\exp\left(t\sum_{k=1}^n X_k\right) \ge \exp(\lambda t)\right) \\
&\le \exp\left(t^{\frac{1+\alpha}{\alpha}}\sum_{k=1}^n \vartheta_n^2 - \lambda t\right)\prod_{k=1}^n\left(1 + Ct^2\vartheta_k^{\frac{4\alpha}{1+\alpha}}\right) \\
&\le \exp\left(t^{\frac{1+\alpha}{\alpha}} - \lambda t + Ct^2\right) \\
&\le \exp(-C_\alpha\lambda^{1+\alpha})
\end{aligned}$$

where in the penultimate step, we set $t = \left(\frac{\alpha\lambda}{\alpha+1}\right)^{\alpha}$, and C_α is defined in (26). Clearly, the last inequality holds when $\lambda > \left[\Gamma\left(\frac{\alpha}{1+\alpha}\right)\frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}}$. □

D PROOFS FOR SECTION 3

Proof of Lemma 3.1. First, note that by Assumption 3.1, we have

$$\begin{aligned} f(z_k) &\le f(w_k) + \langle \nabla f(w_k), z_k - w_k\rangle + \frac{L}{2}\|z_k - w_k\|^2 \\ &\le (1-\alpha_k) f(z_{k-1}) + \alpha_k\left[f(w_k) + \langle \nabla f(w_k), x_k - w_k\rangle\right] + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2, \end{aligned} \tag{12}$$

where the second inequality follows from the convexity of f, and the definition of the sequence w_k and z_k from Algorithm 2. Also note that by definition of the sequence x_k from Algorithm 2 (based on Algorithm 1), we have

$$-\mu_k \le \langle \bar{G}_k + \gamma_k(x_k - x_{k-1}), u - x_k\rangle \quad \forall u \in \mathcal{X}. \tag{13}$$

Letting $u = x_*$ in the above inequality and multiplying it by α_k, summing it up with (12), and denoting $\bar{\Delta}_k = \bar{G}_k - \nabla f(w_k)$, we obtain

$$f(z_k) \le (1-\alpha_k) f(z_{k-1}) + \alpha_k f(x_*) + \alpha_k\left[\mu_k + \langle \bar{\Delta}_k + \gamma_k(x_k - x_{k-1}), x_* - x_k\rangle\right] + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2,$$

which together with the facts that

$$\begin{aligned} \|x_{k-1} - x_*\|^2 &= \|x_k - x_{k-1}\|^2 + \|x_k - x_*\|_2^2 + 2\langle x_{k-1} - x_k, x_k - x_*\rangle, \\ \alpha_k\langle \bar{\Delta}_k, x_* - x_k\rangle &\le \alpha_k\langle \bar{\Delta}_k, x_* - x_{k-1}\rangle + \frac{\|\bar{\Delta}_k\|^2}{2L} + \frac{L\alpha_k^2}{2}\|x_k - x_{k-1}\|^2, \end{aligned}$$

imply

$$\begin{aligned} f(z_k) \le (1-\alpha_k) f(z_{k-1}) + \alpha_k f(x_*) + \alpha_k &\left[\mu_k + \frac{2L\alpha_k - \gamma_k}{2}\|x_k - x_{k-1}\|^2 + \langle \bar{\Delta}_k, x_* - x_{k-1}\rangle\right] \\ &+ \frac{\alpha_k\gamma_k}{2}\left[\|x_{k-1} - x_*\|^2 - \|x_k - x_*\|^2\right] + \frac{\|\bar{\Delta}_k\|^2}{2L}. \end{aligned}$$

Recalling the definition of $\hat{\Gamma}_k$ and $\hat{\Gamma}_1$, subtracting $f(x_*)$ from both sides, dividing by $\hat{\Gamma}_k$, summing them up, we obtain

$$\frac{f(z_N) - f(x_*)}{\hat{\Gamma}_N} \le \frac{\gamma_1}{2}\|x_0 - x_*\|^2 + \sum_{i=1}^{N}\frac{\alpha_k\mu_k}{\hat{\Gamma}_k} + \sum_{i=1}^{N}\frac{\alpha_k}{\hat{\Gamma}_k}\langle \bar{\Delta}_k, x_* - x_{k-1}\rangle + \sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{2L\hat{\Gamma}_k}, \tag{14}$$

which completes the proof. □

Proof of Theorem 3.1. By Lemma 3.1, we have (14) where we recall that $\bar{\Delta}_k = \bar{G}_k - \nabla f(w_k)$ with $\bar{G}_k$ as defined in (4). Now, note that the first two terms on the right hand side of (14) are bounded by the constant $3LD_0$.

Hence, we proceed to getting a handle on the third and fourth terms in the right hand side of (14) with high probability. Considering the fourth term, note that according to Assumption 1.1, we have

$$\mathbb{E}\left[\exp\left\{\|\frac{\bar{\Delta}_k}{\sigma}\|^2 m_k\right\}\middle|\mathcal{F}_{k-1}\right] \le \exp\{1\}, \tag{15}$$

where $\mathcal{F}_{k-1} = \sigma(\xi_1, \cdots, \xi_{k-1})$ is the σ-algebra generated by the random sequence $\xi_1, \cdots, \xi_{k-1}$. Now, by defining

$$\pi_k := \frac{1}{\hat{\Gamma}_k m_k}, \quad \text{and} \quad \theta_k := \frac{\pi_k}{\sum_k \pi_k},$$

we obtain the inequality corresponding to the fourth term on the right hand side of (14):

$$\exp\left\{\sum_{k=1}^{N}\frac{\theta_k\|\bar{\Delta}_k\|^2 m_k}{\sigma^2}\right\}\leq\sum_{k=1}^{N}\theta_k\exp\left\{\frac{\|\bar{\Delta}_k\|^2 m_k}{\sigma^2}\right\}.$$

Taking expectations on both sides, and using (15) we then obtain

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\frac{\sum_{k=1}^{N}\frac{1}{\hat{\Gamma}_k m_k}\|\bar{\Delta}_k\|^2 m_k}{\left(\sigma^2\sum_{k=1}^{N}\frac{1}{\hat{\Gamma}_k m_k}\right)}\right\}\right]&\leq\sum_{k=1}^{N}\theta_k\mathbb{E}\left[\exp\left\{\frac{\|\bar{\Delta}_k\|^2 m_k}{\sigma^2}\right\}\right]\\&=\sum_{k=1}^{N}\theta_k\mathbb{E}\left[\mathbb{E}\Big[\exp\left\{\frac{\|\bar{\Delta}_k\|^2 m_k}{\sigma^2}\right\}\Big|\mathcal{F}_{k-1}\Big]\right]\\&\leq\exp\{1\}.\end{aligned}$$

It then follows by Markov's inequality that for all $\lambda\geq 0$, we have

$$\mathbb{P}\left(\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k}\geq\lambda\Big(\sigma^2\sum_{k=1}^{N}\frac{1}{\hat{\Gamma}_k m_k}\Big)\right)\leq\exp\{-\lambda\}.$$

Note that, by our choice of α_k and m_k, we have $\frac{1}{\hat{\Gamma}_k m_k}\leq\frac{D_0}{N}$, where $D_0=\|x_0-x^*\|^2$. Substituting this fact in the above bound, we hence obtain for all $\lambda\geq 0$,

$$\mathbb{P}\left(\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k}\geq\lambda\sigma^2 D_0\right)\leq\exp\{-\lambda\}. \tag{16}$$

This completes the high-probability bound for the fourth term on the right hand side of (14). In order to bound the third term on the right hand side of (14), we first let

$$\zeta_k=\frac{\alpha_k}{\hat{\Gamma}_k}\langle\bar{\Delta}_k,x_*-x_{k-1}\rangle.$$

Then Assumption 1.1 implies that,

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\frac{\zeta_k^2 m_k}{[\alpha_k\hat{\Gamma}_k^{-1}D_0\sigma]^2}\right\}\Big|\mathcal{F}_{k-1}\right]&\leq\mathbb{E}\left[\exp\left\{\frac{m_k(\|\bar{\Delta}\|\|x_{k-1}-x^*\|)^2}{[\sigma D_0]^2}\right\}\Big|\mathcal{F}_{k-1}\right]\\&\leq\exp\{1\}.\end{aligned}$$

As $\mathbb{E}[\langle\bar{\Delta}_k,x_*-x_{k-1}\rangle|\mathcal{F}_{k-1}]=0$ it follows that $\{\zeta_k\}_{k\geq 1}$ is a martingale difference sequence. Then by exponential concentration inequalities for sums of martingale difference sequence (specifically by [30, Lemma 2]), we have for all $\lambda\geq 0$

$$\mathbb{P}\left(\sum_{k=1}^{N}\zeta_k\geq\lambda\sigma D_0\Big[\sum_{k=1}^{N}(\hat{\Gamma}_k^{-1}m_k^{-\frac{1}{2}}\alpha_k)^2\Big]^{\frac{1}{2}}\right)\leq\exp\{-\lambda^2/3\}.$$

We remark that while more refined exponential inequalities exist in the literature (for example, [13]) in our above calculation, it suffices to use the version from [30]. Now, note that, by our choice of α_k and m_k we have $\hat{\Gamma}_k^{-1}m_k^{-\frac{1}{2}}\alpha_k\leq(N/D_0)^{-1/2}$. Substituting this fact in the above inequality, we obtain

$$\mathbb{P}\left(\sum_{k=1}^{N}\zeta_k\geq\lambda\sigma D_0\right)\leq\exp\{-\lambda^2/3\}. \tag{17}$$

Combine (14), (16) and (17), we get the high probability bound stated in Theorem 3.1.

For the total number of iterations in Algorithm 1, from the classical analysis of the CG method, one can show that the FW-gap ($-h_\gamma$) of problem (3) is bounded by $LD_{\mathcal{X}}^2/T$ (where L is the Lipschitz constant and $\max_{x,y\in\mathcal{X}}\|y-x\|\leq D_{\mathcal{X}}$) if the CG method runs for T iterations; see, for example [1]. Since the gradient of the objective function in (3) is Lipschitz continuous with constant γ, we have

$$-h_{\gamma_k}(\bar{y}_{T_k})\leq\frac{\gamma_k D_{\mathcal{X}}^2}{T},$$

which together with the choice of μ_k and γ_k in (8), imply that at iteration k of Algorithm 2, we need to run Algorithm 1 for at most $T_k = 4D_{\mathcal{X}}^2 N/D_0$ iterations. Therefore, the total number of iterations of Algorithm 1 to find an ϵ-stationary point of problem (1) is bounded by $\sum_{k=1}^{N} T_k \leq 48LD_{\mathcal{X}}^2/\epsilon$. Hence, we obtain the oracle complexity stated in Theorem 3.1. □

Proof of Lemma 3.2. We first prove part (a). For $\bar{G}_k$ as defined in (5), we let

$$G_t := G(w_k, \xi_{k,t}) \quad \text{and} \quad B_t = \left(\frac{\sigma^{1+\alpha} t}{\log(1/\delta)}\right)^{\frac{1}{1+\alpha}}.$$

Now, by Assumption 1.2, we obtain

$$\begin{aligned}
\left\|\mathbb{E}[\bar{G}_k - \nabla f(w_k)]\right\| &= \frac{1}{m_k}\left\|\sum_{t=1}^{m_k}\left(\mathbb{E}[G_t \mathbb{1}\{\|G_t\| \leq B_t\}] - \nabla f(w_k)\right)\right\| \\
&\leq \frac{1}{m_k}\sum_{t=1}^{m_k}\mathbb{E}[\|G_t\|\mathbb{1}\{\|G_t\| \geq B_t\}] \\
&\leq \frac{1}{m_k}\sum_{t=1}^{m_k}\frac{\sigma^{1+\alpha}}{B_t^{\alpha}}. \qquad (18)
\end{aligned}$$

Now, note that we have

$$\begin{aligned}
\left\|\bar{\Delta}_k\right\| &\leq \frac{1}{m_k}\left\|\sum_{t=1}^{m_k}(\nabla f(w_k) - \mathbb{E}[G_t \mathbb{1}\{\|G_t\| \leq B_t\}])\right\| \\
&\quad + \frac{1}{m_k}\left\|\sum_{t=1}^{m_k}(\mathbb{E}[G_t \mathbb{1}\{\|G_t\| \leq B_t\}] - G_t \mathbb{1}\{\|G_t\| \leq B_t\})\right\| \\
&= \|\mathbb{E}[\bar{G}_k - \nabla f(w_k)]\| + \frac{1}{m_k}\left\|\sum_{t=1}^{m_k}\mathbb{E}[G_t \mathbb{1}\{\|G_t\| \leq B_t\}] - G_t \mathbb{1}\{\|G_t\| \leq B_t\}\right\|.
\end{aligned}$$

Furthermore, we also have that

$$\mathbb{E}(\|G_t\|^2 \mathbb{1}\{\|G_t\| \leq B_t\}) \leq \sigma^{1+\alpha} B^{1-\alpha}.$$

Hence, by (18) and by vector-valued Bernstein's inequality for bounded independent random vectors (see, for example [34, Corollary 4.1]), we have with probability at least $1-\delta$,

$$\left\|\bar{\Delta}_k\right\| \leq \frac{1}{m_k}\sum_{t=1}^{m_k}\frac{\sigma^{1+\alpha}}{B_t^{\alpha}} + \sqrt{\frac{2B_{m_k}^{1-\alpha}\sigma^{1+\alpha}\log(1/\delta)}{m_k}} + \frac{B_{m_k}\log(1/\delta)}{3m_k}.$$

Plugging in the expression for B_t concludes the proof.

The proof of part (b) follows verbatim the proof of part (a) and by noting the fact that $G_t/\sqrt{d}$ satisfies Assumption 1.2. □

Proof of Theorem 3.2. We first prove part (a). Note that by Lemma 3.1, we can obtain the inequality (14). As before, we note that the first two terms on the right hand side of (14) are bounded by the constant $3LD_0$. Hence, we proceed to bound the last two terms on the right hand side of (14) with a high probability bound.

For the last term, according to Lemma 3.2, we have

$$\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}\middle|\mathcal{F}_{k-1}\right] \leq C. \qquad (19)$$

where $\mathcal{F}_{k-1} = \sigma(\xi_1, \cdots, \xi_{k-1})$. Now, by defining

$$\pi_k := \frac{1}{\hat{\Gamma}_k m_k^{\frac{2\alpha}{\alpha+1}}}, \quad \text{and} \quad \theta_k := \frac{\pi_k}{\sum_k \pi_k},$$

we obtain the following inequality:

$$\exp\left\{\left(\sum_{k=1}^{N}\theta_k\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^2 m_k^{\frac{2\alpha}{\alpha+1}}\right)^{\frac{\alpha+1}{2\alpha}}\right\} \leq \exp\left\{\sum_{k=1}^{N}\theta_k\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\} \leq \sum_{k=1}^{N}\theta_k \exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}.$$

By taking expectation on both sides and using (19) we obtain

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\left(\frac{\sum_{k=1}^{N}\hat{\Gamma}_k^{-1}m_k^{-\frac{2\alpha}{\alpha+1}}\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^2 m_k^{-\frac{2\alpha}{\alpha+1}}}{\left(\sum_{k=1}^{N}\hat{\Gamma}_k^{-1}m_k^{-\frac{2\alpha}{\alpha+1}}\right)}\right)^{\frac{1+\alpha}{2\alpha}}\right\}\right] \\
\leq\ & \sum_{k=1}^{N}\theta_k\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}\right] \\
=\ & \sum_{k=1}^{N}\theta_k\mathbb{E}\left[\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}\middle|\mathcal{F}_{k-1}\right]\right] \leq C.
\end{aligned}$$

Hence, by Markov's inequality we have for all $\lambda \geq 0$ that

$$\mathbb{P}\left\{\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \lambda\left(\sum_{k=1}^{N}\frac{\sigma^2}{\hat{\Gamma}_k m_k^{\frac{2\alpha}{\alpha+1}}}\right)\right\} \leq C\exp\{-\lambda^{\frac{1+\alpha}{2\alpha}}\}.$$

If we set $m_k = N^{\frac{2\alpha+2}{\alpha}}$, then we obtain

$$\hat{\Gamma}_k^{-1}m_k^{-\frac{2\alpha}{\alpha+1}} \leq \frac{D_0}{N}.$$

Hence, we have for all $\lambda \geq 0$ that

$$\mathbb{P}\left(\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \lambda\sigma^2 D_0\right) \leq \exp\{-\lambda^{\frac{1+\alpha}{2\alpha}}\}$$

which equivalently leads to

$$\mathbb{P}\left(\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \sigma^2 D_0\log\left(\frac{1}{\delta}\right)^{\frac{2\alpha}{1+\alpha}}\right) \leq \delta. \tag{20}$$

corresponding to the fourth term on the right hand side of (14).

For the third term, again by setting $m_k = N^{\frac{2\alpha+2}{\alpha}}$ and applying Lemma 3.2, we have with probability $1-\delta$

$$\sum_{i=1}^{N}\frac{\alpha_k}{\hat{\Gamma}_k}\langle\bar{\Delta}_k, x_* - x_{k-1}\rangle \leq D_0\sigma\sum_{i=1}^{N}\frac{\alpha_k}{\hat{\Gamma}_k}\left(\frac{\log(1/\delta)}{m_k'}\right)^{\frac{\alpha}{\alpha+1}} \leq \sigma D_0\log\left(\frac{1}{\delta}\right)^{\frac{\alpha}{1+\alpha}} \tag{21}$$

Now, the claim in Theorem 3.2 follows by (14), (20) and (21). The oracle complexity results then follows by our choice of m_k and the argument similar to that used in the proof of Theorem 3.1.

Furthermore, the proof of part (b) follows verbatim the proof of part (a) as $G/\sqrt{d}$ satisfies Assumption 1.2. □

Proof of Theorem 3.3. We first require a concentration result from [9] for (6), which we restate below in our notation.

Lemma D.1. Suppose G satisfies Assumption 1.3 and $d \lesssim \log(1/\delta)$, then the estimator $\bar{G}_k$ in (6), with $\bar{\Delta}_k := \bar{G}_k - \nabla f(w_k)$ satisfies

$$\mathbb{P}\left(\|\bar{\Delta}_k\| \gtrsim \left(\frac{d}{n}\right)^{\frac{\beta}{1+\beta}} + \left(\frac{\log(1/\delta)}{m_k}\right)^{\frac{\beta}{1+\beta}}\right) \leq \delta, \quad \text{and} \quad \mathbb{E}\left[\exp\left\{\|\bar{\Delta}_k\|^{\frac{1+\beta}{\beta}} m_k\right\}\right] \leq C.$$

Note that by Lemma D.1, we also have

$$\mathbb{P}\left(\|\bar{\Delta}_k\| \geq \lambda\right) \leq C \exp\left(-\lambda^{\frac{1+\beta}{\beta}} m_k\right). \tag{22}$$

With (22) in hand, the proof of Theorem 3.3 follows verbatim the proof of Theorem 3.2. □

Proof of Proposition 3.1. Before proving Proposition 3.1, we introduce an intermediate result regarding the initial estimator in (9), which is essentially [9, Lemma B.1], restated in our notation.

Lemma D.2. For a given k, let $G(w_k, \xi_{k,j})$, for $t = 1, \ldots, m_k$ be $i.i.d.$ random vectors satisfying Assumption 1.3 for some $\beta \in (0, 1]$. Then the estimator $\widehat{G}_k$ as defined by (9), with probability at least $1 - e^{-\frac{m_k}{50}}$. satisfies

$$\|\widehat{G}_k - \nabla f(w_k)\| \leq 24\sqrt{d}.$$

Now we are ready to prove Proposition 3.1. First, we recall the definition of $\bar{G}_k$ from (10):

$$\bar{G}_k := \frac{2}{m_k} \sum_{t=1}^{m_k/2} \min\left\{ \frac{\left[\left(\frac{t}{\log(1/\delta)}\right)^{\frac{1}{1+\beta}} + 24\right]\sqrt{d}}{\|G(w_k, \xi_{k,t}) - \widehat{G}_k\|}, 1 \right\} \left[G(w_k, \xi_{k,t}) - \widehat{G}_k)\right] + \widehat{G}_k.$$

Now, under Assumption 3.2, it is straightforward to see that $\widehat{G}_k$ is an unbiased estimator of $\nabla f(w_k)$ and the distribution of $\widehat{G}_k$ is symmetric about $\nabla f(w_k)$.

Now, we proceed to first prove that $\bar{G}_k$ is unbiased, i.e., $\mathbb{E}[\bar{G}_k] = \nabla f(w_k)$ by showing

$$\mathbb{E}\left[\min\left\{\frac{B}{\|G(w_k, \xi) - \widehat{G}_k\|}, 1\right\} [G(w_k, \xi) - \widehat{G}_k] + \widehat{G}_k\right] = \nabla f(w_k) \tag{23}$$

for any $B \geq 0$ when the distribution of $G(w_k, \xi)$ is symmetric about $\nabla f(w_k)$. Note that without loss of generality, one can assume that $\nabla f(w_k) = 0$. As, if this is not true, we can define $U(w_k, \xi) = G(w_k, \xi) - \nabla f(w_k)$, for which we have $\mathbb{E}[U(w_k, \xi)] = 0$, and $\widehat{U}_k = \widehat{G}_k - \nabla f(w_k)$, which leads to

$$\mathbb{E}\left[\min\left\{\frac{B}{\|G(w_k, \xi) - \widehat{G}_k\|}, 1\right\} [G(w_k, \xi) - \widehat{G}_k] + \widehat{U}_k\right] = 0$$

as we have $G(w_k, \xi) - \widehat{G}_k = U(w_k, \xi) - \widehat{U}_k$, which would prove (23) for the general case.

Denoting the distribution of $G = G(w_k, \xi)$ by $g(x)$, we immediately have that $g(x) = g(-x)$ and $-\widehat{G}_k \overset{d}{\sim} \widehat{G}_k$ (i.e., $-\widehat{G}_k$ has the same distribution as $\widehat{G}_k$). Hence, we have for any $B \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[G(w_k, \xi)\mathbb{1}\{\|G(w_k, \xi) - \widehat{G}_k\| \leq B\}\right] \\
&= \mathbb{E}_{\hat{G}_k}\left[\int_{\|x - \hat{G}_k\| \leq B} x g(x) dx\right] \\
&= \mathbb{E}_{\widehat{G}_k}\left[\int_{\|x\| \leq B} (x + \widehat{G}_k) g(x + \widehat{G}_k) dx\right] \\
&= \mathbb{E}_{\widehat{G}_k}\left[\int_{\|x\| \leq B} (\widehat{G}_k - x) g(\widehat{G}_k - x) dx\right] \\
&= \mathbb{E}_{\widehat{G}_k}\left[\int_{\|x\| \leq B} (\widehat{G}_k - x) g(x - \widehat{G}_k) dx\right] \\
&= \mathbb{E}_{\widehat{G}_k}\left[\int_{\|x\| \leq B} (-\widehat{G}_k - x) g(x + \widehat{G}_k) dx\right].
\end{aligned}$$

Comparing the equation in the second and last line, we can immediately obtain that

$$\mathbb{E}[G(w_k,\xi)\mathbb{1}\{\|G(w_k,\xi)-\widehat{G}_k\|\le B\}]=0. \tag{24}$$

Similarly, we can show that

$$\mathbb{E}\left[\mathbb{1}\left\{\|G(w_k,\xi)-\widehat{G}_k\|\ge B\right\}\left[\frac{B}{\|G(w_k,\xi)-\widehat{G}_k\|}(G(w_k,\xi)-\widehat{G}_k)+\widehat{G}_k\right]\right]=0,$$

which together with (24) proves (23).

Next, we will prove the concentration of $\bar{G}_k$. As $G(w_k,\xi)$ satisfies Assumption 1.3, we have $\mathbb{E}[\|G(w_k,\xi)\|^{1+\beta}]\le\frac{\pi}{2}d^{\frac{1+\beta}{2}}$. Let

$$B_t=\left[\left(\frac{t}{\log(1/\delta)}\right)^{\frac{1}{1+\beta}}+24\right]\sqrt{d},\qquad G_t=G(w_k,\xi_{k,t}),$$

and

$$\widetilde{G}_t=\widetilde{G}(w_k,\xi_{k,t})=\min\left\{\frac{B_t}{\|G(w_k,\xi_{k,t})-\widehat{G}_k\|},1\right\}[G(w_k,\xi_{k,t})-\widehat{G}_k]+\widehat{G}_k.$$

As $\widetilde{G}_t$, for $t=1,\dots,m_k/2$ depends on $\widehat{G}_k$, they are only independent conditioned on $\widehat{G}_k$. Hence, conditioned on $\widehat{G}_k$, we have the following:

$$\begin{aligned}\left\|\mathbb{E}[\bar{G}_k-\nabla f(w_k)]\right\|&=\frac{2}{m_k}\left\|\sum_{t=1}^{m_k/2}\left(\widetilde{G}_t-\nabla f(w_k)\right)\right\|\\&\le\frac{2}{m_k}\sum_{t=1}^{m_k/2}\mathbb{E}[(\|G_t\|+\|\widehat{G}_k\|)\mathbb{1}\{\|G_t\|\ge B_t-\|\widehat{G}_k\|\}]\\&\le\frac{2}{m_k}\sum_{t=2}^{m_k/2}\frac{(\frac{\pi}{2}+24)d^{\frac{1+\beta}{2}}}{(B_t-\widehat{G}_k)^{\alpha}}.\end{aligned}$$

Now, again conditioned on $\widehat{G}_k$, note that we obtain

$$\begin{aligned}\left\|\bar{\Delta}_k\right\|&\le\left\|\mathbb{E}[\bar{G}_k-\nabla f(w_k)]\right\|+\frac{2}{m_k}\left\|\sum_{t=1}^{m_k}(\mathbb{E}[\tilde{G}(w_k,\xi_{k,t})]-\tilde{G}(w_k,\xi_{k,t}))\right\|\\&\le\frac{2}{m_k}\sum_{t=2}^{m_k/2}\frac{(\frac{\pi}{2}+24)d^{\frac{1+\beta}{2}}}{(B_t-\widehat{G}_k)^{\alpha}}+\frac{2}{m_k}\left\|\sum_{t=1}^{m_k}(\mathbb{E}[\tilde{G}(w_k,\xi_{k,t})]-\tilde{G}(w_k,\xi_{k,t}))\right\|.\end{aligned}$$

Now, by vector-valued Bernstein's inequality for bounded independent random vectors (see, for example [34, Corollary 4.1]), and by noting that conditioned on $\widehat{G}_k$ we have

$$\mathbb{E}[\|\widetilde{G}(w_k,\xi_{k,t}\|^2]\Big|\widehat{G}_k\le\frac{\pi}{2}d^{\frac{1+\beta}{2}}(B+\|\widehat{G}_k\|)^{1-\beta},$$

when $m_k\ge 100\log(1/\delta)$, we have, with probability at least $1-\delta$,

$$\left\|\bar{\Delta}_k\right\|\le\left(\frac{2}{m_k}\sum_{t=2}^{m_k/2}\frac{(\frac{\pi}{2}+24)d^{\frac{1+\beta}{2}}}{(B_t-\widehat{G}_k)^{\alpha}}+\sqrt{\frac{4(B_{m_k}+\widehat{G}_k)^{1-\beta}\frac{\pi}{2}d^{\frac{1+\beta}{2}}\log(1/\delta)}{m_k}}+\frac{2(B_{m_k}+\widehat{G}_k)\log(1/\delta)}{3m_k}\right).$$

The high-probability statement from Proposition 3.1 then follows by setting B_t and noting that the norm of $\widehat{G}_k$ is bounded, i.e., $\widehat{G}_k\le 24\sqrt{d}$. □

Proof of Theorem 3.4. We first prove part (a). Note that by Lemma 3.1, we can obtain the inequality (14). As before, we note that the first two terms in the right hand side of (14) are bounded by the constant $3LD_0$. Hence, we proceed to bound the last two terms on the right hand side of (14) with a high probability bound.

For the fourth term on the right hand side of (14), using the same approach as in the proof of Theorem 3.2 we have for all $\lambda \geq 0$ that

$$\mathbb{P}\left(\sum_{k=1}^{N} \frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \lambda D_0 d\right) \leq \exp\left\{-\lambda^{\frac{1+\alpha}{2\alpha}}\right\}.$$

Next, by setting $m_k = N^{\frac{3\alpha+3}{2\alpha}}$, we obtain

$$\hat{\Gamma}_k^{-1} m_k^{-\frac{2\alpha}{\alpha+1}} \leq \frac{D_0}{N}.$$

Hence, we have for all $\lambda \geq 0$ that

$$\mathbb{P}\left(\sum_{k=1}^{N} \frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq D_0 d \log\left(\frac{1}{\delta}\right)^{\frac{2\alpha}{1+\alpha}}\right) \leq \delta. \tag{25}$$

For the third term in the right hand side of (14), define

$$\zeta_k := \frac{\alpha_k}{\hat{\Gamma}_k} \langle \bar{\Delta}_k, x_* - x_{k-1} \rangle.$$

As we have that Assumption 1.2 is stronger than Assumption 1.3, Proposition 3.1 and Lemma C.1 imply that,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\left(\frac{\zeta_k}{[\alpha_k \hat{\Gamma}_k^{-1} D_0 \sqrt{d}]}\right)^{\frac{1+\alpha}{\alpha}} m_k\right\} \middle| \mathcal{F}_{k-1}\right] \\
&\leq \mathbb{E}\left[\exp\left\{\frac{m_k \left(\|\bar{\Delta}\| \|x_{k-1} - x^*\|\right)^{\frac{1+\alpha}{\alpha}}}{(D_0\sqrt{d})^{\frac{1+\alpha}{\alpha}}}\right\} \middle| \mathcal{F}_{k-1}\right] \leq 2.
\end{aligned}$$

which indicates that ζ_k satisfies Assumption C.2. Consequently according to Lemma C.1 it also satisfies Assumption C.1. By setting $m_k = N^{\frac{3\alpha+3}{2\alpha}}$, we have

$$\frac{1}{m_k} (\hat{\Gamma}_k^{-1} \alpha_k)^{\frac{1+\alpha}{\alpha}} \leq N^{-\frac{\alpha+1}{2\alpha}}.$$

Now, part (b) of Lemma C.2 and Proposition C.1 implies that we have

$$\mathbb{P}\left(\sum_{k=1}^{N} \zeta_k \geq \lambda D_0 \sqrt{d}\right) \leq \exp\{-C_\alpha \lambda^{1+\alpha}\}, \quad \text{for all} \quad \lambda \geq \left[\Gamma\left(\tfrac{\alpha}{1+\alpha}\right) \frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}},$$

where

$$C_\alpha = \left(\frac{\alpha}{1+\alpha}\right)^{\alpha} - \left(\frac{\alpha}{1+\alpha}\right)^{1+\alpha} - \left(\frac{\alpha}{1+\alpha}\right)^{2\alpha} \geq 0. \tag{26}$$

The above probability bound, in turn leads to

$$\mathbb{P}\left(\sum_{k=1}^{N} \zeta_k \geq D_0 \sqrt{d} \left[\log\left(\frac{1}{\delta}\right)^{\frac{1}{1+\alpha}}\right]\right) \leq \delta, \quad \text{when} \quad (\log(1/\delta))^{\frac{1}{1+\alpha}} \geq \left[\Gamma\left(\frac{\alpha}{1+\alpha}\right) \frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}}. \tag{27}$$

Combining (14) with the high probability bounds in (25) and (27) proves the claim in Theorem 3.4. The oracle complexity results then follow by our choice of m_k and the argument similar to that used in the proof of Theorem 3.1.

Furthermore, the proof of part (b) follows verbatim the proof of part (a) with α replaced by β.

□

E SUMMARY OF THE ROBUST MEAN ESTIMATION PROCEDURE FROM [9]

For the sake of completeness, we now describe the robust mean-estimation procedure from [9]. The main algorithm from [9] is provided in Algorithm 3.

Algorithm 3 OPTIMALMEANEST($\{G(w_k, \xi_{k,j})\}_{j=1}^{m_k}$)

Input: Data Points $\{G(w_k, \xi_{k,j})\}_{j=1}^{m_k} \in \mathbb{R}^d$, Target Confidence δ
$G^+ \leftarrow$ Initial Mean Estimate($\{G(w_k, \xi_{k,j})\}_{j=1}^{m_k/2}$)
$Z \leftarrow$ Produce Bucket Estimates($\{G(w_k, \xi_{k,j})\}_{j=m_k/2}^{m_k}, G^+, \delta$)
$T \leftarrow 10^6 \log dn$
$\bar{G}_k =$ Gradient Descent(Z, G^+, T)
Return: $\bar{G}_k$

Algorithm 3 comprises of the following three sub-steps.

1. Data Pruning Step

The following algorithms correspond to the first sub-step. Algorithm 4 compute an initial estimate of the mean which is with $\mathcal{O}(\sqrt{d})$ of the mean and Algorithm 5 use this estimate to filter out data points which are far away from the estimate.

Algorithm 4 Initial Mean Estimate

Input: Set of Data Points $\{G_i\}_{i=1}^n$
$\hat{\mu} \leftarrow \arg\min_{G_i : i \leq n} \min \left\{ r > 0, \sum_{j=1}^n \mathbb{1}\{\|G_j - G_i\| \leq r\} \geq 0.6n \right\}$
Return: $\hat{\mu}$

Algorithm 5 Prune Data

Input: Set of Data Points $\{G_i\}_{i=1}^n$, Mean Estimate G^+
$\tau \leftarrow \max\left(100n^{\frac{1}{1+\alpha}} d^{-\frac{1-\alpha}{2(1+\alpha)}}, 100\sqrt{d}\right)$
$\mathcal{C} \leftarrow \{G_i : \|G_i - G^+\| \leq \tau\}$
Return: $\mathcal{C}$

2. Data Batching Step

The following algorithm corresponds to the second sub-step. The data points that survive the truncation procedure in the data pruning stage are then divided into k bins and mean estimates are computed based on sample-averaging in each bin by Algorithm 6.

Algorithm 6 Produce Bucket Estimates

Input: Set of Data Points $\{G_i\}_{i=1}^n$, Mean Estimate G^+, Target Confidence δ
$Y \leftarrow$ Prune Data($\{G_i\}, G^+$)
$m \leftarrow |Y|$
$k \leftarrow 4000 \log 1/\delta$
Split data points into k buckets with bucket $\mathcal{B}_i$ consisting of the points $G_{(i-1)\frac{m}{k}+1}, \cdots, G_{i\frac{m}{k}}$
$Z_i \leftarrow$ Mean($\mathcal{B}_i$) $\forall i \in [k]$ and $Z \leftarrow (Z_i, \cdots, Z_k)$
Return: Z

3. Median Computation Step

The following algorithms correspond to the third sub-step. The bucket estimates from the previous stage are aggregated to produce the final estimate following the testing-to-estimation framework. The testing program is defined in **MT** below. Algorithms 7 and 8 display the estimation of distance and gradient.

Algorithm 7 Distance Estimation

Input: Data Points $Z \in \mathbb{R}^{k \times d}$, Current point x
$d = \arg\max_{r>0} \mathbf{MT}(x, r, Z) \geq 0.9k$
Return: d

Algorithm 8 Gradient Estimation

Input: Data Points $Z \in \mathbb{R}^{k \times d}$, Current point x
$d^* =$ Distance Estimation(Z, x)
$(v, X) = \mathbf{MT}(x, d^*, Z)$
$g \leftarrow$ Top Singular Vector(X_v)
Return: g

Here, $\mathbf{MT}(x, r, Z)$ is the solution of the following semidefinite optimization problem:

$$\begin{aligned} \max \sum_{i=1}^{k} X_{1,b_i} \\ \text{Subjectd to} \quad X_{1,b_i} &= X_{b_i,b_i} \\ \sum_{j=1}^{d} X_{v_j,v_j} &= 1, \\ \langle v_{b_i}, Z_i - x \rangle &\geq X_{b_i,b_i} r \quad \forall i \in [k] \\ X_{1,1} &= 1 \\ X &\succeq 0 \end{aligned}$$

where $v_{b_i} = [X_{b_i,v_1}, \cdots, X_{b_i,v_d}]$. The estimate above is then used in Algorithm 9 to obtain an improved estimate.

Algorithm 9 Gradient Descent

Input: Bucket Means $Z \in \mathbb{R}^{k \times d}$, Initialization G^+, Number of Iterations T
x^*, $x_0 \leftarrow G^+$ and d^*, $d_0 \leftarrow \infty$
for $t = 0, \ldots, T$ **do**
  $d_t \leftarrow$ Distance Estimation(Z, x_t)
  $g_t \leftarrow$ Gradient Estimation(Z, x_t)
  if $d_t < d^*$ **then**
    $x^* \leftarrow x_t$
    $d^* \leftarrow d_t$
  end if
  $x_{t+1} \leftarrow x_t + \frac{1}{20} d_t g_t$
end for
Output: x^*

The overall computational complexity of the algorithm is polynomial in the problem parameters [9]. However, from the perspective of implementation, especially on large scale datasets, it is perhaps not efficient.