High-Probability Bounds for Robust Stochastic Frank-Wolfe Algorithm (Supplementary material)

Tongyi Tang[1] **Krishnakumar Balasubramanian**[1] **Thomas C. M. Lee**[1]

[1]Department of Statistics, University of California, Davis, CA, USA.

1 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 $\mathcal{X}_{TR}(\tau) \coloneqq \{X \in \mathbb{R}^{L\times d} : \sum_{j=1}^d \sigma_j(X) \leq \tau\}$ is the $\|.\|_{tr}$ ball of radius τ. This problem fits in the setup of (1) with $\xi \coloneqq (a, y)$ and $F(X, \xi) \coloneqq \log\big(\sum_l \exp(a^\top x_l - a^\top x_y)\big)$. 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 1.4 is satisfied, while Assumption 1.2 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 1. 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.

2 PLOTS ILLUSTRATING MAIN THEORETICAL RESULTS

In Figure 2, we provide an illustration of the SFO complexity from Theorem 3.6. 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).

3 CONCENTRATION INEQUALITY FOR MARTINGALES WITH HEAVY-TAILS

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

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

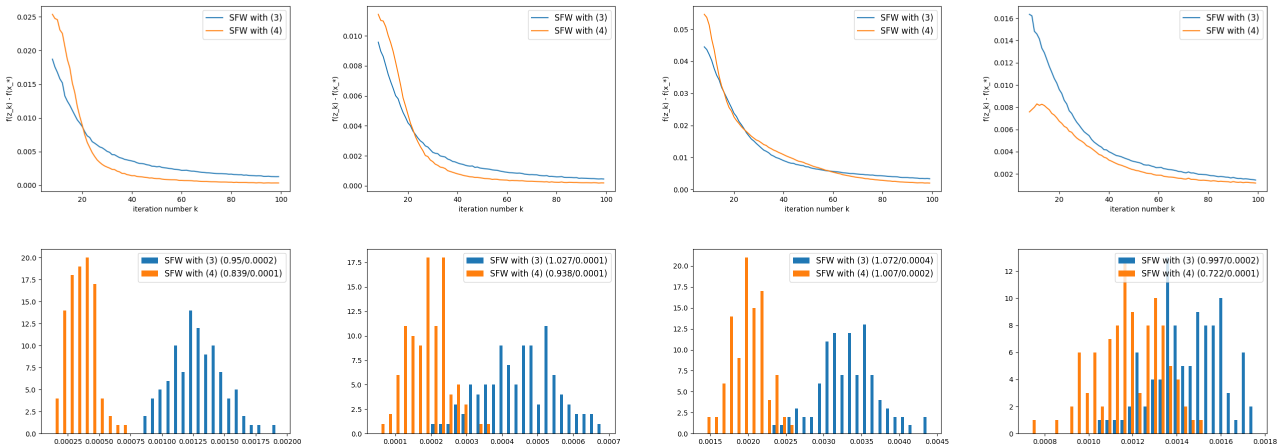

Figure 1: The two left and two right columns corresponds to Pareto, Student-t distributions with $d = 20/L = 10$ and $d = 100/L = 20$ respectively. **Top row:** Mean (solid lines) over 100 trails of iterations versus $f(z_N) - f(x_*)$ for $N = 100$. **Bottom row:** Histogram of $f(z_N) - f(x_*)$ for $N = 100$. Numbers in the legend correspond to *heavy-tail index/standard deviation*.

Assumption 3.1. 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 3.2. 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 $\alpha \in (0, 1]$.

Lemma 3.3. *Assumptions 3.1 and 3.2 are equivalent.*

Proof. Suppose X satisfies Assumption 3.1 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 3.2 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 3.4. *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 3.1, the following properties hold:*

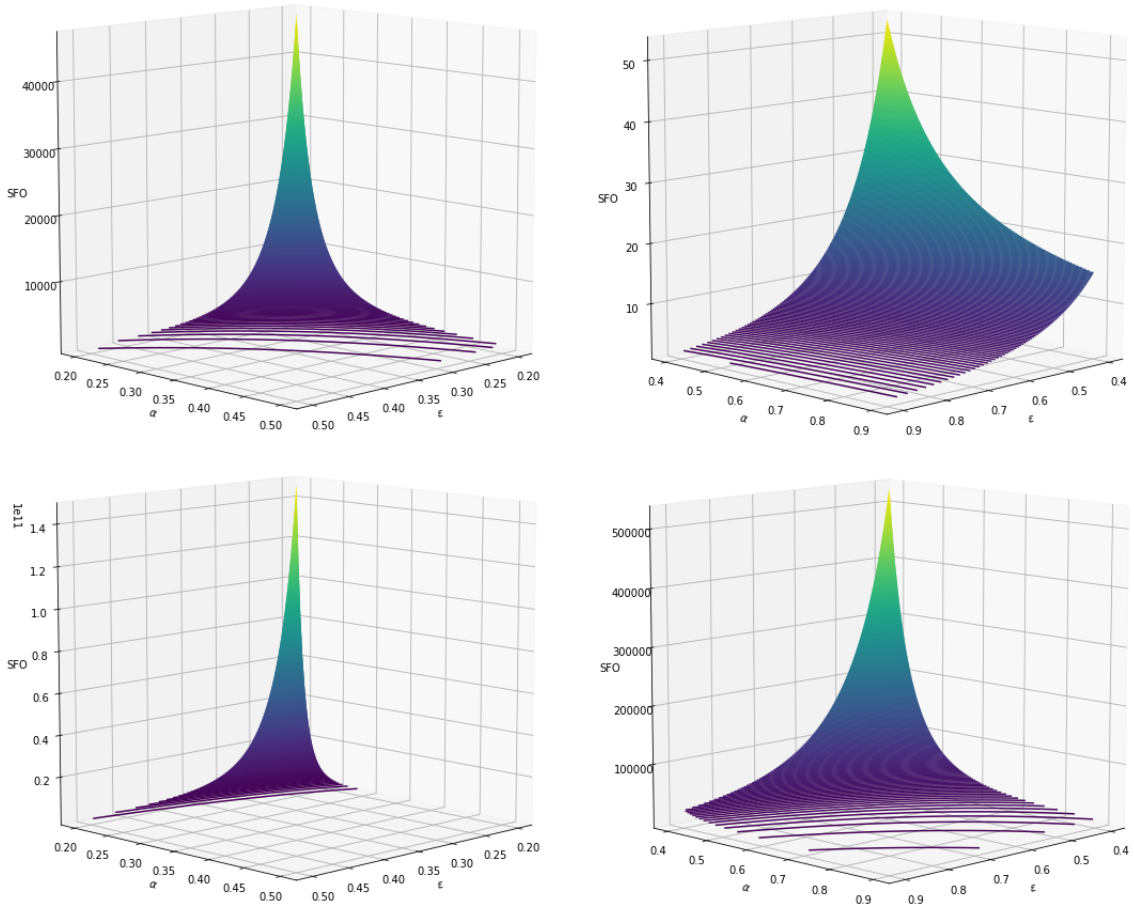

Figure 2: Visualization of the SFO complexity from Theorem 3.6 part (a) **(top row)** and part (b) **(bottom row)**. Note that for smaller values of ϵ (i.e, higher accuracy) and lower values of α (i.e., when only a smaller order moment exists for the stochastic gradients) the SFO complexity increases rapidly. Note the effect of dimension is pronounced in the bottom row corresponding to part (b).

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

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

(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

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

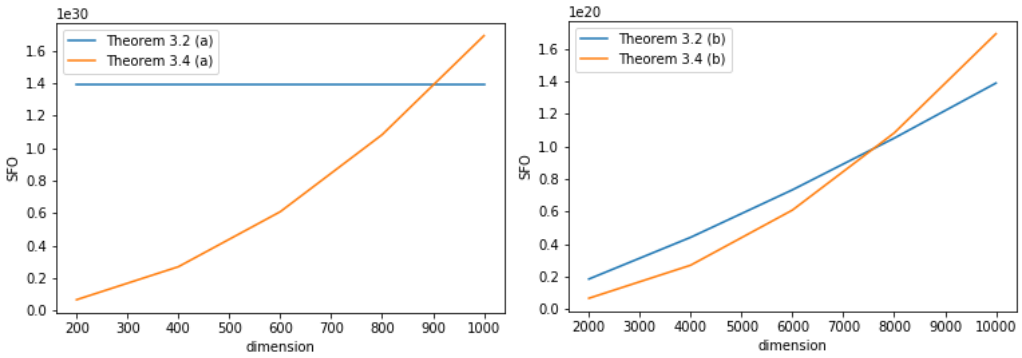

Figure 3: Comparing the SFO complexity in Theorem 3.6 and Theorem 3.13. Part (a) on the left and part (b) on the right. Note that in both cases, there is a certain threshold in dimension below which the SFO complexity of Theorem 3.13 is better than that of Theorem 3.6.

.

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

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

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 3.5. *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 $\vartheta_i^2\leq n^{-\frac{\alpha+1}{\alpha}}$ for all i, then, we have

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

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

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

where C_α is as defined in (15).

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

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

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

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

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

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

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 $\vartheta_i^2 \leq n^{-\frac{\alpha+1}{2\alpha}}$, we have

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

where in the penultimate step, we set $t = \left(\frac{\alpha\lambda}{\alpha+1}\right)^{\alpha}$, and C_α is defined in (15). Clearly, the last inequality holds when $\lambda > \left[\Gamma\left(\frac{\alpha}{1+\alpha}\right)\frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}}$. □

4 PROOFS FOR SECTION 3

Proof of Lemma 3.2. First, note that by Assumption 3.1, we have

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

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

$$-\mu_k \leq \langle \bar{G}_k + \gamma_k(x_k - x_{k-1}), u - x_k \rangle \quad \forall u \in \mathcal{X}. \tag{2}$$

Letting $u = x_*$ in the above inequality and multiplying it by α_k, summing it up with (1), and denoting $\bar{\Delta}_k = \bar{G}_k - \nabla f(w_k)$, we obtain

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

which together with the facts that

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

imply

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

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

$$\frac{f(z_N) - f(x_*)}{\hat{\Gamma}_N} \leq \frac{\gamma_1}{2}\|x_0 - x_*\|^2 + \sum_{i=1}^{N} \frac{\alpha_k \mu_k}{\hat{\Gamma}_k} + \sum_{i=1}^{N} \frac{\alpha_k}{\hat{\Gamma}_k} \langle \bar{\Delta}_k, x_* - x_{k-1} \rangle + \sum_{k=1}^{N} \frac{\|\bar{\Delta}_k\|^2}{2L\hat{\Gamma}_k}, \tag{3}$$

which completes the proof. □

Proof of Theorem 3.3. By Lemma 3.2, we have (3) 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 (3) 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 (3) with high probability. Considering the fourth term, note that according to Assumption 1.2, we have

$$\mathbb{E}\left[\exp\left\{ \| \frac{\bar{\Delta}_k}{\sigma} \|^2 m_k \right\} \middle| \mathcal{F}_{k-1} \right] \leq \exp\{1\}, \tag{4}$$

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

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

we obtain the inequality corresponding to the fourth term on the right hand side of (3):

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

Taking expectations on both sides, and using (4) 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 $\lambda \geq 0$,

$$\mathbb{P}\left(\sum_{k=1}^{N}\frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \lambda\sigma^2 D_0\right) \leq \exp\{-\lambda\}. \tag{5}$$

This completes the high-probability bound for the fourth term on the right hand side of (3). In order to bound the third term on the right hand side of (3), we first let

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

Then Assumption 1.2 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 [**?**, 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, **?**) in our above calculation, it suffices to use the version from **?**. 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

$$\mathbb{P}\left(\sum_{k=1}^{N}\zeta_k \geq \lambda\sigma D_0\right) \leq \exp\{-\lambda^2/3\}. \tag{6}$$

Combine (3), (5) and (6), we get the high probability bound stated in Theorem 3.3.

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 Balasubramanian and Ghadimi [2021]. 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 3.3. □

Proof of Lemma 3.5. We first prove part (a). For $\bar{G}_k$ as defined in (5), we let

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

Now, by Assumption 1.4, we obtain

$$\begin{aligned}
\left\|\mathbb{E}[\bar{G}_k - \nabla f(w_k)]\right\| &= \frac{1}{m_k}\left\|\sum_{t=1}^{m_k}\left(\mathbb{E}[G_t \mathbb{1}\{\|G_t\| \le B_t\}] - \nabla f(w_k)\right)\right\| \\
&\le \frac{1}{m_k}\sum_{t=1}^{m_k}\mathbb{E}[\|G_t\|\mathbb{1}\{\|G_t\| \ge B_t\}] \\
&\le \frac{1}{m_k}\sum_{t=1}^{m_k}\frac{\sigma^{1+\alpha}}{B_t^{\alpha}}. \qquad (7)
\end{aligned}$$

Now, note that we have

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

Furthermore, we also have that

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

Hence, by (7) and by vector-valued Bernstein's inequality for bounded independent random vectors (see, for example [**?**, Corollary 4.1]), we have with probability at least $1-\delta$,

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

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 1.4. □

Proof of Theorem 3.6. We first prove part (a). Note that by Lemma 3.2, we can obtain the inequality (3). As before, we note that the first two terms on the right hand side of (3) are bounded by the constant $3LD_0$. Hence, we proceed to bound the last two terms on the right hand side of (3) with a high probability bound.

For the last term, according to Lemma 3.5, we have

$$\mathbb{E}\left[\exp\left\{\left\|\frac{\bar{\Delta}_k}{\sigma}\right\|^{\frac{1+\alpha}{\alpha}} m_k\right\}\middle|\mathcal{F}_{k-1}\right] \le C. \qquad (8)$$

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

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

we obtain the following inequality:

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

By taking expectation on both sides and using (8) 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

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

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

$$\mathbb{P}\left(\sum_{k=1}^N \frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq \sigma^2 D_0 \log\left(\frac{1}{\delta}\right)^{\frac{2\alpha}{1+\alpha}}\right) \leq \delta. \tag{9}$$

corresponding to the fourth term on the right hand side of (3).

For the third term, again by setting $m_k = N^{\frac{2\alpha+2}{\alpha}}$ and applying Lemma 3.5, we have with probability $1-\delta$

$$\sum_{i=1}^N \frac{\alpha_k}{\hat{\Gamma}_k}\langle\bar{\Delta}_k, x_* - x_{k-1}\rangle \leq D_0\sigma \sum_{i=1}^N \frac{\alpha_k}{\hat{\Gamma}_k}\left(\frac{\log(1/\delta)}{m_k'}\right)^{\frac{\alpha}{\alpha+1}} \leq \sigma D_0 \log\left(\frac{1}{\delta}\right)^{\frac{\alpha}{1+\alpha}} \tag{10}$$

Now, the claim in Theorem 3.6 follows by (3), (9) and (10). The oracle complexity results then follows by our choice of m_k and the argument similar to that used in the proof of Theorem 3.3.

Furthermore, the proof of part (b) follows verbatim the proof of part (a) as $G/\sqrt{d}$ satisfies Assumption 1.4. □

Proof of Theorem 3.8. We first require a concentration result from Cherapanamjeri et al. [2022] for (6), which we restate below in our notation.

Lemma 4.1. *Suppose G satisfies Assumption 1.5 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*

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

Note that by Lemma 4.1, we also have

$$\mathbb{P}\left(\|\bar{\Delta}_k\| \geq \lambda\right) \leq C \exp\left(-\lambda^{\frac{1+\beta}{\beta}} m_k\right). \tag{11}$$

With (11) in hand, the proof of Theorem 3.8 follows verbatim the proof of Theorem 3.6. □

Proof of Proposition 3.11. Before proving Proposition 3.11, we introduce an intermediate result regarding the initial estimator in (9), which is essentially [Cherapanamjeri et al., 2022, Lemma B.1], restated in our notation.

Lemma 4.2. *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 1.5 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 3.11. First, we recall the definition of $\bar{G}_k$ from (10):

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

Now, under Assumption 3.10, 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

$$\mathbb{E}\left[\min\left\{\frac{B}{\|G(w_k, \xi) - \widehat{G}_k\|}, 1\right\}[G(w_k, \xi) - \widehat{G}_k] + \widehat{G}_k\right] = \nabla f(w_k) \tag{12}$$

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 (12) 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 $B \geq 0$,

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

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

$$\mathbb{E}[G(w_k,\xi)\mathbb{1}\{\|G(w_k,\xi)-\widehat{G}_k\|\le B\}]=0. \tag{13}$$

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 (13) proves (12).

Next, we will prove the concentration of $\bar{G}_k$. As $G(w_k,\xi)$ satisfies Assumption 1.5, 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 [**?**, 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 $m_k\ge100\log(1/\delta)$, 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 3.11 then follows by setting B_t and noting that the norm of $\widehat{G}_k$ is bounded, i.e., $\widehat{G}_k\le24\sqrt{d}$. □

Proof of Theorem 3.13. We first prove part (a). Note that by Lemma 3.2, we can obtain the inequality (3). As before, we note that the first two terms in the right hand side of (3) are bounded by the constant $3LD_0$. Hence, we proceed to bound the last two terms on the right hand side of (3) with a high probability bound.

For the fourth term on the right hand side of (3), using the same approach as in the proof of Theorem 3.6 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

$$\mathbb{P}\left(\sum_{k=1}^{N} \frac{\|\bar{\Delta}_k\|^2}{\hat{\Gamma}_k} \geq D_0 d \log\left(\frac{1}{\delta}\right)^{\frac{2\alpha}{1+\alpha}}\right) \leq \delta. \tag{14}$$

For the third term in the right hand side of (3), define

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

As we have that Assumption 1.4 is stronger than Assumption 1.5, Proposition 3.11 and Lemma 3.3 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 3.2. Consequently according to Lemma 3.3 it also satisfies Assumption 3.1. 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 3.4 and Proposition 3.5 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

$$C_\alpha = \left(\frac{\alpha}{1+\alpha}\right)^{\alpha} - \left(\frac{\alpha}{1+\alpha}\right)^{1+\alpha} - \left(\frac{\alpha}{1+\alpha}\right)^{2\alpha} \geq 0. \tag{15}$$

The above probability bound, in turn leads to

$$\mathbb{P}\left(\sum_{k=1}^{N} \zeta_k \geq D_0\sqrt{d}\left[\log\left(\frac{1}{\delta}\right)^{\frac{1}{1+\alpha}}\right]\right) \leq \delta, \quad \text{when} \quad (\log(1/\delta))^{\frac{1}{1+\alpha}} \geq \left[\Gamma\left(\frac{\alpha}{1+\alpha}\right)\frac{1+\alpha}{\alpha}\right]^{\frac{1}{1-\alpha}}. \tag{16}$$

Combining (3) with the high probability bounds in (14) and (16) proves the claim in Theorem 3.13. The oracle complexity results then follow by our choice of m_k and the argument similar to that used in the proof of Theorem 3.3.

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

□

5 SUMMARY OF THE ROBUST MEAN ESTIMATION PROCEDURE FROM Cherapanamjeri et al. [2022]

For the sake of completeness, we now describe the robust mean-estimation procedure from Cherapanamjeri et al. [2022]. The main algorithm from Cherapanamjeri et al. [2022] is provided in Algorithm 1.

Algorithm 1 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 1 comprises of the following three sub-steps.

1. Data Pruning Step

The following algorithms correspond to the first sub-step. Algorithm 2 compute an initial estimate of the mean which is with $\mathcal{O}(\sqrt{d})$ of the mean and Algorithm 3 use this estimate to filter out data points which are far away from the estimate.

Algorithm 2 Initial Mean Estimate

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

Algorithm 3 Prune Data

Input: Set of Data Points $\{G_i\}_{i=1}^n$, Mean Estimate G^+
$\tau \leftarrow \max\left(100 n^{\frac{1}{1+\beta}} d^{-\frac{1-\beta}{2(1+\beta)}}, 100\sqrt{d}\right)$
$\mathcal{C} \leftarrow \{G_i : \|G_i - G^+\| \le \tau\}$
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 4.

Algorithm 4 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 5 and 6 display the estimation of distance and gradient.

Algorithm 5 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 6 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

The following polynomial and its semidefinite optimization $\mathbf{MT}(x, r, Z)$ play a key role in the subsequent analysis. Intuitively, given a test point x, the problem searches for a direction (v) such that a large proportion of the bucket estimates, Z_i, are far away from x along v. Formally, the polynomial optimization problem, parameterized by x, r and Z, is defined below:

$$\begin{aligned} \max & \sum_{i=1}^{k} b_i \\ \text{Subjectd to} \quad & b_i^2 = b_i \\ & \|v\|^2 = 1, \\ b_i(\langle v, Z_i - x \rangle - r) & \geq 0 \quad \forall i \in [k] \end{aligned}$$

The binary variables b_i indicates whether $i - th$ bucket mean Z_i is far away along v. However, the binary constraints on b_i, the restriction of v and the final constraint make this problem nonconvex and hard to optimize efficiently. Therefore, they work with the simidefinite relaxation defined as follows:

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

where $v_{b_i} = [X_{b_i,v_1}, \cdots, X_{b_i,v_d}]$. The matrix $X \in \mathcal{S}_+^{(k+d+1)}$ is symbolically indexed by 1 and the variables $b_1, \ldots, b_k$ and $v_1, \ldots, v_d$. Here, $(v, X) = \mathbf{MT}(x, r, Z)$ refers the optimal value v and solution X of the following semidefinite optimization problem initialized with x, r and Z:

The estimate above is then used in Algorithm 7 to obtain an improved estimate.

Algorithm 7 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 Cherapanamjeri et al. [2022]. However, from the perspective of implementation, especially on large scale datasets, it is perhaps not efficient.