Defending Against Saddle Point Attack in Byzantine-Robust Distributed Learning Supplementary Material

Dong Yin [*1], Yudong Chen [†3], Kannan Ramchandran [‡1], and Peter Bartlett [§1,2]

[1]Department of Electrical Engineering and Computer Sciences, UC Berkeley
[2]Department of Statistics, UC Berkeley
[3]School of Operations Research and Information Engineering, Cornell University

A Additional Related Work

Outlier-robust estimation is a classical topic in statistics [8]. The coordinate-wise median aggregation subroutine that we consider is related to the median-of-means estimator [17, 9], which has been applied to various robust inference problems [15, 14, 16]. A recent line of work develops efficient robust estimation algorithms in high-dimensional settings [2, 5, 11, 3, 18, 12, 1, 10, 13]. In the centralized setting, the recent work [7] proposes a scheme, similar to the iterative filtering procedure, that iteratively removes outliers for gradient-based optimization.

B Challenges of Escaping Saddle Points in the Adversarial Setting

We provide two examples showing that in non-convex setting with saddle points, inexact oracle can lead to much worse sub-optimal solutions than in the convex setting, and that in the adversarial setting, escaping saddle points can be inherently harder than the adversary-free case.

Consider standard gradient descent using exact or Δ-inexact gradients. Our first example shows that Byzantine machines have a more severe impact in the non-convex case than in the convex case.

Example 1. Let $d = 1$ and consider the functions $F^{(1)}(w) = (w-1)^2$ and $F^{(2)}(w) = (w^2-1)^2/4$. Here $F^{(1)}$ is strongly convex with a unique local minimizer $w^* = 1$, whereas $F^{(2)}$ has two local (in fact, global) minimizers $w^* = \pm 1$ and a saddle point (in fact, a local maximum) $w = 0$. Claim 1 below shows the following: for the convex $F^{(1)}$, gradient descent (GD) finds a near-optimal solution with sub-optimality proportional to Δ, regardless of initialization; for the nonconvex $F^{(2)}$, GD initialized near the saddle point $w = 0$ suffers from an $\Omega(1)$ sub-optimality gap.

Claim 1. *Suppose that $\Delta \le 1/2$. Under the setting above, the following holds.*
(i) For $F^{(1)}$, starting from any w_0, GD using a Δ-inexact gradient oracle finds w with $F^{(1)}(w) - F^{(1)}(w^*) \le \mathcal{O}(\Delta)$.
(ii) For $F^{(2)}$, there exists an adversarial strategy such that starting from a w_0 sampled uniformly from $[-r, r]$, GD with a Δ-inexact gradient oracle outputs w with $F^{(2)}(w) - F^{(2)}(w^*) \ge \frac{9}{64}, \forall w^* = \pm 1$, with probability $\min\{1, \frac{\Delta}{r}\}$.

[*]dongyin@berkeley.edu
[†]yudong.chen@cornell.edu
[‡]kannanr@berkeley.edu
[§]peter@berkeley.edu

Proof. Since $F^{(2)}(w) = \frac{1}{4}(w^2-1)^2$, we have $\nabla F^{(2)}(w) = w^3 - w$. For any $w \in [-\Delta, \Delta]$, $|\nabla F^{(2)}(w)| \le \Delta$ (since $\Delta \le 1/2$). Thus, the adversarial oracle can always output $\widehat{g}(w) = 0$ when $w \in [-\Delta, \Delta]$, and we have $|\widehat{g}(w) - \nabla F^{(2)}(w)| \le \Delta$. Thus, if $w \in [-\Delta, \Delta]$, the iterate can no longer move with this adversarial strategy. Then, we have $F^{(2)}(w) - F^{(2)}(w^*) \ge F^{(2)}(\Delta) - 0 \ge \frac{9}{64}$ (since $\Delta \le 1/2$). The result for the convex function $F^{(1)}$ is a direct corollary of Theorem 1 in [21]. □

Our second example shows that escaping saddle points is much harder in the Byzantine setting than in the non-Byzantine setting.

Example 2. Let $d = 2$, and assume that in the neighborhood $\mathbb{B}_0(b)$ of the origin, F takes the quadratic form $F(\mathbf{w}) \equiv \frac{1}{2}w_1^2 - \frac{\lambda}{2}w_2^2$, with $\lambda > \epsilon_H$.[1] The origin $\mathbf{w}_0 = 0$ is not an (ϵ_g, ϵ_H)-second-order stationary point, but rather a saddle point. Claim 2 below shows that exact GD escapes the saddle point almost surely, while GD with an inexact oracle fails to do so.

Claim 2. *Under the setting above, if one chooses $r < b$ and sample $\mathbf{w}$ from $\mathbb{B}_0(r)$ uniformly at random, then:*
(i) Using exact gradient descent, with probability 1, the iterate $\mathbf{w}$ eventually leaves $\mathbb{B}_0(r)$.
(ii) There exists an adversarial strategy such that, when we update $\mathbf{w}$ using Δ-inexact gradient oracle, if $\Delta \ge \lambda r$, with probability 1, the iterate $\mathbf{w}$ cannot leave $\mathbb{B}_0(r)$; otherwise with probability $\frac{2}{\pi}\Big(\arcsin\big(\frac{\Delta}{\lambda r}\big) + \frac{\Delta}{\lambda r}\sqrt{1-(\frac{\Delta}{\lambda r})^2}\Big)$ the iterate $\mathbf{w}$ cannot leave $\mathbb{B}_0(r)$.

Proof. Since $F(\mathbf{w}) = \frac{1}{2}w_1^2 - \frac{1}{2}\lambda w_2^2$, $\forall\ \mathbf{w} \in \mathbb{B}_0(r)$, we have $\nabla F(\mathbf{w}) = [w_1,\ -\lambda w_2]^\top$. Sample $\mathbf{w}_0$ uniformly at random from $\mathbb{B}_0(r)$, and we know that with probability 1, $w_{0,2} \neq 0$. Then, by running exact gradient descent $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla F(\mathbf{w}_t)$, we can see that the second coordinate of $\mathbf{w}_t$ is $w_{t,2} = (1+\eta\lambda)^t w_{0,2}$. When $w_{0,2}$, we know that as t gets large, we eventually have $w_{t,2} > r$, which implies that the iterate leaves $\mathbb{B}_0(r)$.

On the other hand, suppose that we run Δ-inexact gradient descent, i.e., $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\widehat{\mathbf{g}}(\mathbf{w}_t)$ with $\|\widehat{\mathbf{g}}(\mathbf{w}_t) - \nabla F(\mathbf{w}_t)\|_2 \le \Delta$. In the first step, if $|w_{0,2}| \le \frac{\Delta}{\lambda}$, the adversary can simply replace $\nabla F(\mathbf{w}_0)$ with $\widehat{\mathbf{g}}(\mathbf{w}_0) = [w_{0,1},\ 0]^\top$ (one can check that here we have $\|\widehat{\mathbf{g}}(\mathbf{w}_0) - \nabla F(\mathbf{w}_0)\|_2 \le \Delta$), and then the second coordinate of $\mathbf{w}_1$ does not change, i.e., $w_{1,2} = w_{0,2}$. In the following iterations, the adversary can keep using the same strategy and the second coordinate of $\mathbf{w}$ never changes, and then the iterates cannot escape $\mathbb{B}_0(r)$, since $F(\mathbf{w})$ is a strongly convex function in its first coordinate. To compute the probability of getting stuck at the saddle point, we only need to compute the area of the region $\{\mathbf{w} \in \mathbb{B}_0(r) : |w_2| \le \frac{\Delta}{\lambda}\}$, which can be done via simple geometry. □

Remark. Even if we choose the largest possible perturbation in $\mathbb{B}_0(r)$, i.e., sample $\mathbf{w}$ from the circle $\{\mathbf{w} \in \mathbb{R}^2 : \|\mathbf{w}\|_2 = r\}$, the stuck region still exists. We can compute the length of the arc $\{\|\mathbf{w}\|_2 = r : |w_2| \le \frac{\Delta}{\lambda}\}$ and find the probability of stuck. One can find that when $\Delta \ge \lambda r$, the probability of being stuck in $\mathbb{B}_0(r)$ is still 1, otherwise, the probability of being stuck is $\frac{2}{\pi}(\arcsin(\frac{\Delta}{\lambda r}))$.

The above examples show that the adversary can significantly alter the landscape of the function near a saddle point. We counter this by exerting a large perturbation on the iterate so that it escapes this bad region. The amount of perturbation is carefully calibrated to ensure that the algorithm finds a descent direction "steep" enough to be preserved under Δ-corruption, while not compromising the accuracy. Multiple rounds of perturbation are performed, boosting the escape probability exponentially.

C Proof of Theorem 3

We first analyze the gradient descent step with Δ-inexact gradient oracle.

Lemma 1. *Suppose that $\eta = 1/L_F$. For any $\mathbf{w} \in \mathcal{W}$, if we run the following inexact gradient descent step:*

$$\mathbf{w}' = \mathbf{w} - \eta\widehat{\mathbf{g}}(\mathbf{w}), \tag{1}$$

with $\|\widehat{\mathbf{g}}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le \Delta$. Then, we have

$$F(\mathbf{w}') \le F(\mathbf{w}) - \frac{1}{2L_F}\|\nabla F(\mathbf{w})\|_2^2 + \frac{1}{2L_F}\Delta^2.$$

[1] $F(\mathbf{w}) \equiv \frac{1}{2}w_1^2 - \frac{\lambda}{2}w_2^2$ holds locally around the origin, not globally; otherwise $F(\mathbf{w})$ has no minimum.

Proof. Since $F(\mathbf{w})$ is L_F smooth, we know that

$$\begin{aligned}
F(\mathbf{w}') \leq & F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{L_F}{2} \|\mathbf{w}' - \mathbf{w}\|_2^2 \\
= & F(\mathbf{w}) - \langle \nabla F(\mathbf{w}), \frac{1}{L_F} (\widehat{\mathbf{g}}(\mathbf{w}) - \nabla F(\mathbf{w})) \rangle - \langle \nabla F(\mathbf{w}), \frac{1}{L_F} \nabla F(\mathbf{w}) \rangle \\
& + \frac{1}{2L_F} \|\widehat{\mathbf{g}}(\mathbf{w}) - \nabla F(\mathbf{w}) + \nabla F(\mathbf{w})\|_2^2 \\
\leq & F(\mathbf{w}) - \frac{1}{2L_F} \|\nabla F(\mathbf{w})\|_2^2 + \frac{1}{2L_F} \Delta^2.
\end{aligned}$$

□

Let ϵ be the threshold on $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2$ that the algorithm uses to determine whether or not to add perturbation. Choose $\epsilon = 3\Delta$. Suppose that at a particular iterate $\widetilde{\mathbf{w}}$, we observe $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 > \epsilon$. Then, we know that

$$\|\nabla F(\widetilde{\mathbf{w}})\|_2 \geq \|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 - \Delta \geq 2\Delta.$$

According to Lemma 1, by running one iteration of the inexact gradient descent step, the decrease in function value is at least

$$\frac{1}{2L_F} \|\nabla F(\widetilde{\mathbf{w}})\|_2^2 - \frac{1}{2L_F} \Delta^2 \geq \frac{3\Delta^2}{2L_F}. \tag{2}$$

We proceed to analyze the perturbation step, which happens when the algorithm arrives at an iterate $\widetilde{\mathbf{w}}$ with $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 \leq \epsilon$. In this proof, we slightly abuse the notation. Recall that in equation (2) in Section 3.1 , we use $\mathbf{w}_t'$ $(0 \leq t \leq T_{\mathrm{th}})$ to denote the iterates of the algorithm in the saddle point escaping process. Here, we simply use $\mathbf{w}_t$ to denote these iterates. We start with the definition of *stuck region* at $\widetilde{\mathbf{w}} \in \mathcal{W}$.

Definition (stuck region)**.** *Given $\widetilde{\mathbf{w}} \in \mathcal{W}$, and parameters r, R, and T_{th}, the stuck region $\mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}}) \subseteq \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ is a set of $\mathbf{w}_0 \in \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ which satisfies the following property: there exists an adversarial strategy such that when we start with $\mathbf{w}_0$ and run T_{th} gradient descent steps with Δ-inexact gradient oracle $\widehat{\mathbf{g}}(\mathbf{w})$:*

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta \widehat{\mathbf{g}}(\mathbf{w}_{t-1}), \ t = 1, 2, \ldots, T_{\mathrm{th}}, \tag{3}$$

we observe $\|\mathbf{w}_t - \mathbf{w}_0\|_2 < R$, $\forall\ t \leq T_{\mathrm{th}}$.

When it is clear from the context, we may simply use the terminology stuck region $\mathbb{W}_S$ at $\widetilde{\mathbf{w}}$. The following lemma shows that if $\nabla^2 F(\widetilde{\mathbf{w}})$ has a large negative eigenvalue, then the stuck region has a small width along the direction of the eigenvector associated with this negative eigenvalue.

Lemma 2. *Assume that the smallest eigenvalue of $\mathbf{H} := \nabla^2 F(\widetilde{\mathbf{w}})$ satisfies $\lambda_{\min}(\mathbf{H}) \leq -\gamma < 0$, and let the unit vector $\mathbf{e}$ be the eigenvector associated with $\lambda_{\min}(\mathbf{H})$. Let $\mathbf{u}_0, \mathbf{y}_0 \in \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ be two points such that $\mathbf{y}_0 = \mathbf{u}_0 + \mu_0 \mathbf{e}$ with some $\mu_0 \geq \mu \in (0, r)$. Choose step size $\eta = \frac{1}{L_F}$, and consider the stuck region $\mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$. Suppose that r, R, T_{th}, and μ satisfy* [2]

$$T_{\mathrm{th}} = \frac{2}{\eta\gamma} \log_{9/4}(\frac{2(R+r)}{\mu}), \tag{4}$$

$$R \geq \mu, \tag{5}$$

$$\rho_F (R+r)\mu \geq \Delta, \tag{6}$$

$$\gamma \geq 24 \rho_F (R+r) \log_{9/4}(\frac{2(R+r)}{\mu}). \tag{7}$$

Then, there must be either $\mathbf{u}_0 \notin \mathbb{W}_S$ or $\mathbf{y}_0 \notin \mathbb{W}_S$.

We prove Lemma 2 in Appendix C.1. With this lemma, we proceed to analyze the probability that the algorithm escapes the saddle points. In particular, we bound the probability that $\mathbf{w}_0 \in \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$ when $\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \leq -\gamma$ and $\mathbf{w}_0$ is drawn from $\mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ uniform at random.

[2] Without loss of generality, here we assume that $\frac{2}{\eta\gamma} \log_{9/4}(\frac{2(R+r)}{\mu})$ is an integer, so that T_{th} is an integer.

Lemma 3. *Assume that $\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \leq -\gamma < 0$, and let the unit vector $\mathbf{e}$ be the eigenvector associated with $\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}}))$. Consider the stuck region $\mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$ at $\widetilde{\mathbf{w}}$, and suppose that r, R, T_{th}, and μ satisfy the conditions in (4)-(7). Then, when we sample $\mathbf{w}_0$ from $\mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ uniformly at random, the probability that $\mathbf{w}_0 \in \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$ is at most $\frac{2\mu\sqrt{d}}{r}$.*

Proof. Since the starting point $\mathbf{w}_0$ is uniformly distributed in $\mathbb{B}_{\widetilde{\mathbf{w}}}(r)$, to bound the probability of getting stuck, it suffices to bound the volume of $\mathbb{W}_S$. Let $\mathbb{1}_{\mathbb{W}_S}(\mathbf{w})$ be the indicator function of the set $\mathbb{W}_S$. For any $\mathbf{w} \in \mathbb{R}^d$, let $w^{(1)}$ be the projection of $\mathbf{w}$ onto the $\mathbf{e}$ direction, and $\mathbf{w}^{(-1)} \in \mathbb{R}^{d-1}$ be the remaining component of $\mathbf{w}$. Then, we have

$$
\begin{aligned}
\mathrm{Vol}(\mathbb{W}_S) =& \int_{\mathbb{B}^{(d)}_{\widetilde{\mathbf{w}}}(r)} \mathbb{1}_{\mathbb{W}_S}(\mathbf{w}) \mathrm{d}\mathbf{w} \\
=& \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{w}}}(r)} \mathrm{d}\mathbf{w}^{(-1)} \int_{\widetilde{w}^{(1)} - \sqrt{r^2 - \|\widetilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|_2^2}}^{\widetilde{w}^{(1)} + \sqrt{r^2 - \|\widetilde{\mathbf{w}}^{(-1)} - \mathbf{w}^{(-1)}\|_2^2}} \mathbb{1}_{\mathbb{W}_S}(\mathbf{w}) \mathrm{d}\widetilde{w}^{(1)} \\
\leq& 2\mu \int_{\mathbb{B}^{(d-1)}_{\widetilde{\mathbf{w}}}(r)} \mathrm{d}\mathbf{w}^{(-1)} \\
=& 2\mu \mathrm{Vol}(\mathbb{B}^{(d-1)}_0(r)),
\end{aligned}
$$

where the inequality is due to Lemma 2. Then, we know that the probability of getting stuck is

$$
\frac{\mathrm{Vol}(\mathbb{W}_S)}{\mathrm{Vol}(\mathbb{B}^{(d)}_0(r))} \leq 2\mu \frac{\mathrm{Vol}(\mathbb{B}^{(d-1)}_0(r))}{\mathrm{Vol}(\mathbb{B}^{(d)}_0(r))} = \frac{2\mu}{\sqrt{\pi} r} \frac{\Gamma(\frac{d}{2}+1)}{\Gamma(\frac{d}{2}+\frac{1}{2})} \leq \frac{2\mu}{\sqrt{\pi} r} \sqrt{\frac{d}{2} + \frac{1}{2}} \leq \frac{2\mu\sqrt{d}}{r},
$$

where we use the fact that $\frac{\Gamma(x+1)}{\Gamma(x+\frac{1}{2})} < \sqrt{x + \frac{1}{2}}$ for any $x \geq 0$. □

We then analyze the decrease of value of the population loss function $F(\cdot)$ when we conduct the perturbation step. Assume that we successfully escape the saddle point, i.e., there exists $t \leq T_{\mathrm{th}}$ such that $\|\mathbf{w}_t - \mathbf{w}_0\|_2 \geq R$. The following lemma provides the decrease of $F(\cdot)$.

Lemma 4. *Suppose that $\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \leq -\gamma < 0$, and at $\widetilde{\mathbf{w}}$, we observe $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 \leq \epsilon = 3\Delta$. Assume that $\mathbf{w}_0 \in \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ and that $\mathbf{w}_0 \notin \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$. Let $t \leq T_{\mathrm{th}}$ be the step such that $\|\mathbf{w}_t - \mathbf{w}_0\|_2 \geq R$. Then, we have*

$$
F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t) \geq \frac{L_F}{4T_{\mathrm{th}}} R^2 - \frac{\Delta^2 T_{\mathrm{th}}}{L_F} - 4\Delta r - \frac{L_F}{2} r^2. \tag{8}
$$

We prove Lemma 4 in Appendix C.2.

Next, we choose the quantities μ, r, R, and γ such that (i) the conditions (4)-(7) in Lemma 2 are satisfied, (ii) the probability of escaping saddle point in Lemma 3 is at least a constant, and (iii) the decrease in function value in (8) is large enough. We first choose

$$
\mu = \Delta^{3/5} d^{-1/5} \rho_F^{-1/2}, \tag{9}
$$

$$
r = 4\Delta^{3/5} d^{3/10} \rho_F^{-1/2}, \tag{10}
$$

$$
R = \Delta^{2/5} d^{1/5} \rho_F^{-1/2}. \tag{11}
$$

One can simply check that, according to Lemma 3, when we drawn $\mathbf{w}_0$ from $\mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ uniformly at random, the probability that $\mathbf{w}_0 \in \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$ is at most $1/2$. Since we assume that $\Delta \leq 1$, one can also check that the condition (5) is satisfied. In addition, since $\rho_F R \mu = \Delta$, the condition (6) is also satisfied. According to (4), we have

$$
T_{\mathrm{th}} = \frac{2L_F}{\gamma} \log_{9/4}\Big(\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2}\Big). \tag{12}
$$

In the following, we choose

$$
\gamma = 768(\rho_F^{1/2} + L_F)(\Delta^{2/5} d^{1/5} + \Delta^{3/5} d^{3/10}) \log_{9/4}\Big(\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2}\Big), \tag{13}
$$

which implies

$$T_{\text{th}} = \frac{L_F}{384(\rho_F^{1/2} + L_F)(\Delta^{2/5}d^{1/5} + \Delta^{3/5}d^{3/10})} \tag{14}$$

Then we check condition (7) holds. We have

$$24\rho_F(R+r)\log_{9/4}(\frac{2(R+r)}{\mu}) = 24\rho_F^{1/2}(\Delta^{2/5}d^{1/5} + 4\Delta^{3/5}d^{3/10})\log_{9/4}(\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2}) \leq \gamma.$$

Next, we consider the decrease in function value in (8). Using the equations (12) and (13), we can show the following three inequalities by direct algebra manipulation.

$$\frac{L_F}{4T_{\text{th}}}R^2 \geq 6\frac{\Delta^2 T_{\text{th}}}{L_F}, \tag{15}$$

$$\frac{L_F}{4T_{\text{th}}}R^2 \geq 24\Delta r, \tag{16}$$

$$\frac{L_F}{4T_{\text{th}}}R^2 \geq 3L_F r^2. \tag{17}$$

By adding up (15), (16), and (17), we obtain

$$\frac{L_F}{4T_{\text{th}}}R^2 \geq 2\frac{\Delta^2 T_{\text{th}}}{L_F} + 8\Delta r + L_F r^2,$$

which implies that when we successfully escape the saddle point, we have

$$F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t) \geq \frac{L_F}{8T_{\text{th}}}R^2 = 48(\rho_F^{-1/2} + L_F\rho_F^{-1})(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10}). \tag{18}$$

Then, one can simply check that, the average decrease in function value during the successful round of the Escape process is

$$\frac{F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t)}{t} \geq \frac{F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t)}{T_{\text{th}}} \geq \frac{2(\Delta^{8/5}d^{4/5} + \Delta^2 d)}{L_F} > \frac{3\Delta^2}{2L_F}. \tag{19}$$

Recall that according to (2), when the algorithm is not in the Escape process, the function value is decreased by at least $\frac{3\Delta^2}{2L_F}$ in each iteration. Therefore, if the algorithm successfully escapes the saddle point, during the Escape process, the average decrease in function value is *larger* than the iterations which are not in this process.

So far, we have chosen the algorithm parameters r, R, T_{th}, as well as the final second-order convergence guarantee γ. Now we proceed to analyze the total number of iterations and the failure probability of the algorithm. According to Lemma 3 and the choice of μ and r, we know that at each point with $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 \leq \epsilon$, the algorithm can successfully escape this saddle point with probability at least $1/2$. To boost the probability of escaping saddle points, we need to repeat the process Q rounds in Escape, independently. Since for each successful round, the function value decrease is at least

$$48(\rho_F^{-1/2} + L_F\rho_F^{-1})(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10}) \geq 48L_F\rho_F^{-1}(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10}),$$

and the function value can decrease at most $F_0 - F^*$. Therefore, the total number of saddle points that we need to escape is at most

$$\frac{\rho_F(F_0 - F^*)}{48L_F(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}. \tag{20}$$

Therefore, by union bound, the failure probability of the algorithm is at most

$$\frac{\rho_F(F_0 - F^*)}{48L_F(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}(\frac{1}{2})^Q,$$

and to make the failure probability at most δ, one can choose

$$Q \geq 2\log\left(\frac{\rho_F(F_0 - F^*)}{48L_F\delta(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}\right). \tag{21}$$

Again, due to the fact that the function value decrease is at most $F_0 - F^*$, and in each *effective* iteration, the function value is decreased by at least $\frac{3\Delta^2}{2L_F}$. (Here, the effective iterations are the iterations when the algorithm is not in the Escape process and the iterations when the algorithm successfully escapes the saddle points.) The total number of effective iterations is at most

$$\frac{2(F_0 - F^*)L_F}{3\Delta^2}. \tag{22}$$

Combing with (21), we know that the total number of parallel iterations is at most

$$\frac{4(F_0 - F^*)L_F}{3\Delta^2} \log\left(\frac{\rho_F(F_0 - F^*)}{48L_F\delta(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}\right).$$

When all the algorithm terminates, and the saddle point escaping process is successful, the output of the algorithm $\widetilde{\mathbf{w}}$ satisfies $\|\widehat{\mathbf{g}}(\widetilde{\mathbf{w}})\|_2 \leq \epsilon$, which implies that $\|\nabla F(\widetilde{\mathbf{w}})\|_2 \leq 4\Delta$, and

$$\begin{aligned}\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \geq -\gamma &= -768(\rho_F^{1/2} + L_F)(\Delta^{2/5}d^{1/5} + \Delta^{3/5}d^{3/10})\log_{9/4}(\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2}) \\ &\geq -950(\rho_F^{1/2} + L_F)(\Delta^{2/5}d^{1/5} + \Delta^{3/5}d^{3/10})\log(\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2}).\end{aligned} \tag{23}$$

We next show that we can simplify the guarantee as

$$\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \geq -1900(\rho_F^{1/2} + L_F)\Delta^{2/5}d^{1/5}\log(\frac{10}{\Delta}). \tag{24}$$

We can see that if $\Delta \leq \frac{1}{\sqrt{d}}$, then $\Delta^{3/5}d^{3/10} \leq \Delta^{2/5}d^{1/5}$ and $\frac{2d^{2/5}}{\Delta^{1/5}} + 8d^{1/2} \leq \frac{10}{\Delta}$. Thus, the bound (24) holds. On the other hand, when $\Delta > \frac{1}{\sqrt{d}}$, we have $\Delta^{2/5}d^{1/5} > 1$ and thus

$$1900(\rho_F^{1/2} + L_F)\Delta^{2/5}d^{1/5}\log(\frac{10}{\Delta}) > L_F.$$

By the smoothness of $F(\cdot)$, we know that $\lambda_{\min}(\nabla^2 F(\widetilde{\mathbf{w}})) \geq -L_F$. Therefore, the bound (24) still holds, and this completes the proof.

C.1 Proof of Lemma 2

We prove by contradiction. Suppose that $\mathbf{u}_0, \mathbf{y}_0 \in \mathbb{W}_S$. Let $\{\mathbf{u}_t\}$ and $\{\mathbf{y}_t\}$ be two sequences generated by the following two iterations:

$$\mathbf{u}_t = \mathbf{u}_{t-1} - \eta\widehat{\mathbf{g}}(\mathbf{u}_{t-1}), \tag{25}$$
$$\mathbf{y}_t = \mathbf{y}_{t-1} - \eta\widehat{\mathbf{g}}(\mathbf{y}_{t-1}), \tag{26}$$

respectively, where $\|\widehat{\mathbf{g}}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \Delta$ for any $\mathbf{w} \in \mathcal{W}$. According to our assumption, we have $\forall\ t \leq T_{\text{th}}$, $\|\mathbf{u}_t - \mathbf{u}_0\|_2 < R$ and $\|\mathbf{y}_t - \mathbf{y}_0\|_2 < R$.

Define $\mathbf{v}_t := \mathbf{y}_t - \mathbf{u}_t$, $\boldsymbol{\delta}_t := \widehat{\mathbf{g}}(\mathbf{u}_t) - \nabla F(\mathbf{u}_t)$, and $\boldsymbol{\delta}'_t := \widehat{\mathbf{g}}(\mathbf{y}_t) - \nabla F(\mathbf{y}_t)$. Then we have

$$\begin{aligned}\mathbf{y}_{t+1} &= \mathbf{y}_t - \eta(\nabla F(\mathbf{y}_t) + \boldsymbol{\delta}'_t) \\ &= \mathbf{u}_t + \mathbf{v}_t - \eta(\nabla F(\mathbf{u}_t + \mathbf{v}_t) + \boldsymbol{\delta}'_t) \\ &= \mathbf{u}_t + \mathbf{v}_t - \eta\nabla F(\mathbf{u}_t) - \eta\left[\int_0^1 \nabla^2 F(\mathbf{u}_t + \theta\mathbf{v}_t)\right]\mathbf{v}_t - \eta\boldsymbol{\delta}'_t \\ &= \mathbf{u}_{t+1} + \eta\boldsymbol{\delta}_t + \mathbf{v}_t - \eta\left[\int_0^1 \nabla^2 F(\mathbf{u}_t + \theta\mathbf{v}_t)\mathrm{d}\theta\right]\mathbf{v}_t - \eta\boldsymbol{\delta}'_t,\end{aligned}$$

which yields

$$\mathbf{v}_{t+1} = (\mathbf{I} - \eta\mathbf{H})\mathbf{v}_t - \eta\mathbf{Q}_t\mathbf{v}_t + \eta(\boldsymbol{\delta}_t - \boldsymbol{\delta}'_t), \tag{27}$$

where

$$\mathbf{Q}_t := \int_0^1 \nabla^2 F(\mathbf{u}_t + \theta\mathbf{v}_t)\mathrm{d}\theta - \mathbf{H}. \tag{28}$$

By the Hessian Lipschitz property, we know that

$$\begin{aligned}\|\mathbf{Q}_t\|_2 \leq & \rho_F(\|\mathbf{u}_t - \widetilde{\mathbf{w}}\|_2 + \|\mathbf{y}_t - \widetilde{\mathbf{w}}\|_2) \\ \leq & \rho_F(\|\mathbf{u}_t - \mathbf{u}_0\|_2 + \|\mathbf{u}_0 - \widetilde{\mathbf{w}}\|_2 + \|\mathbf{y}_t - \mathbf{y}_0\|_2 + \|\mathbf{y}_0 - \widetilde{\mathbf{w}}\|_2) \\ \leq & 2\rho_F(R + r).\end{aligned} \tag{29}$$

We let ψ_t be the norm of the projection of $\mathbf{v}_t$ onto the $\mathbf{e}$ direction, and ϕ_t be the norm of the projection of $\mathbf{v}_t$ onto the remaining subspace. By definition, we have $\psi_0 = \mu_0 \geq \mu > 0$ and $\phi_0 = 0$. According to (27) and (29), we have

$$\psi_{t+1} \geq (1 + \eta\gamma)\psi_t - 2\eta\rho_F(R + r)\sqrt{\psi_t^2 + \phi_t^2} - 2\eta\Delta, \tag{30}$$

$$\phi_{t+1} \leq (1 + \eta\gamma)\phi_t + 2\eta\rho_F(R + r)\sqrt{\psi_t^2 + \phi_t^2} + 2\eta\Delta. \tag{31}$$

In the following, we use induction to prove that $\forall\ t \leq T_{\text{th}}$,

$$\psi_t \geq (1 + \frac{1}{2}\eta\gamma)\psi_{t-1} \quad \text{and} \quad \phi_t \leq \frac{t}{T_{\text{th}}}\psi_t \tag{32}$$

We know that (32) holds when $t = 0$ since we have $\phi_0 = 0$. Then, assume that for some $t < T_{\text{th}}$, we have $\forall\ \tau \leq t$, $\psi_\tau \geq (1 + \frac{1}{2}\eta\gamma)\psi_{\tau-1}$ and $\phi_\tau \leq \frac{\tau}{T_{\text{th}}}\psi_\tau$. We show that (32) holds for $t + 1$.

First, we show that $\psi_{t+1} \geq (1 + \frac{1}{2}\eta\gamma)\psi_t$. Since $\forall\ \tau \leq t$, $\psi_\tau \geq \psi_{\tau-1}$, we know that $\psi_t \geq \psi_0 \geq \mu$. Therefore, according to (6), we have

$$\Delta \leq \rho_F(R + r)\mu \leq \rho_F(R + r)\psi_t. \tag{33}$$

In addition, since $t < T_{\text{th}}$, we have

$$\phi_t \leq \psi_t. \tag{34}$$

Combining (33), (34) and (30), (31), we get

$$\psi_{t+1} \geq (1 + \eta\gamma)\psi_t - 2\eta\rho_F(R + r)\sqrt{2\psi_t^2} - 2\eta\rho_F(R + r)\psi_t > (1 + \eta\gamma)\psi_t - 6\eta\rho_F(R + r)\psi_t, \tag{35}$$

$$\phi_{t+1} \leq (1 + \eta\gamma)\phi_t + 2\eta\rho_F(R + r)\sqrt{2\psi_t^2} + 2\eta\rho_F(R + r)\psi_t < (1 + \eta\gamma)\phi_t + 6\eta\rho_F(R + r)\psi_t. \tag{36}$$

According to (7), we have $\gamma \geq 24\rho_F(R + r)\log_{9/4}(\frac{2(R+r)}{\mu}) > 12\rho_F(R + r)$. Combining with (35), we know that $\psi_{t+1} \geq (1 + \frac{1}{2}\eta\gamma)\psi_t$.

Next, we show that $\phi_{t+1} \leq \frac{t+1}{T_{\text{th}}}\psi_{t+1}$. Combining with (35) and (36), we know that to show $\phi_{t+1} \leq \frac{t+1}{T_{\text{th}}}\psi_{t+1}$, it suffices to show

$$(1 + \eta\gamma)\phi_t + 6\eta\rho_F(R + r)\psi_t \leq \frac{t + 1}{T_{\text{th}}}[1 + \eta\gamma - 6\eta\rho_F(R + r)]\psi_t. \tag{37}$$

According to the induction assumption, we have $\phi_t \leq \frac{t}{T_{\text{th}}}\psi_t$. Then, to show (37), it suffices to show that

$$(1 + \eta\gamma)t + 6\eta\rho_F(R + r)T_{\text{th}} \leq (t + 1)[1 + \eta\gamma - 6\eta\rho_F(R + r)] \tag{38}$$

Since $t + 1 \leq T_{\text{th}}$, we know that to show (38), it suffices to show

$$12\eta\rho_F(R + r)T_{\text{th}} \leq 1. \tag{39}$$

Then, according to (4) and (7), we know that (39) holds, which completes the induction.

Next, according to (32), we know that

$$\begin{aligned}\|\mathbf{u}_{T_{\text{th}}} - \mathbf{y}_{T_{\text{th}}}\|_2 \geq \phi_{T_{\text{th}}} &\geq (1 + \frac{1}{2}\eta\gamma)^{T_{\text{th}}}\mu_0 \\ &\geq (1 + \frac{1}{2}\eta\gamma)^{\frac{2}{\eta\gamma}\log_{9/4}(\frac{2(R+r)}{\mu})}\mu_0 \\ &\geq \frac{2(R + r)}{\mu} \cdot \mu_0 = 2(R + r),\end{aligned}$$

where the last inequality is due to the fact that $\eta = \frac{1}{L_F}$ and thus $\eta\gamma \leq 1$. On the other hand, since we assume that $\mathbf{u}_0, \mathbf{y}_0 \in \mathbb{W}_S$, we know that

$$\|\mathbf{u}_{T_{\text{th}}} - \mathbf{y}_{T_{\text{th}}}\|_2 \leq \|\mathbf{u}_{T_{\text{th}}} - \mathbf{u}_0\|_2 + \|\mathbf{y}_{T_{\text{th}}} - \mathbf{y}_0\|_2 + \|\mathbf{u}_0 - \mathbf{y}_0\|_2 < 2(R + r),$$

which leads to contradiction and thus completes the proof.

C.2 Proof of Lemma 4

Recall that we have the iterations $\mathbf{w}_{\tau+1} = \mathbf{w}_\tau - \eta\widehat{\mathbf{g}}(\mathbf{w}_\tau)$ for all $\tau < t$. Let $\boldsymbol{\delta}_\tau = \nabla F(\mathbf{w}_\tau) - \widehat{\mathbf{g}}(\mathbf{w}_\tau)$, and then $\|\boldsymbol{\delta}_\tau\|_2 \leq \Delta$. By the smoothness of $F(\cdot)$ and the fact that $\eta = \frac{1}{L_F}$, we have

$$\begin{aligned}
F(\mathbf{w}_\tau) - F(\mathbf{w}_{\tau+1}) \geq & \langle \nabla F(\mathbf{w}_\tau), \mathbf{w}_\tau - \mathbf{w}_{\tau+1}\rangle - \frac{L_F}{2}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 \\
= & \left\langle \frac{\mathbf{w}_\tau - \mathbf{w}_{\tau+1}}{\eta} + \boldsymbol{\delta}_\tau, \mathbf{w}_\tau - \mathbf{w}_{\tau+1}\right\rangle - \frac{L_F}{2}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 \\
= & \frac{L_F}{2}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 + \langle \boldsymbol{\delta}_\tau, \mathbf{w}_\tau - \mathbf{w}_{\tau+1}\rangle \\
\geq & \frac{L_F}{4}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 - \frac{\|\boldsymbol{\delta}_\tau\|_2^2}{L_F} \\
\geq & \frac{L_F}{4}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 - \frac{\Delta^2}{L_F}.
\end{aligned} \tag{40}$$

By summing up (40) for $\tau = 0, 1, \ldots, t-1$, we get

$$F(\mathbf{w}_0) - F(\mathbf{w}_t) \geq \frac{L_F}{4}\sum_{\tau=0}^{t-1}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 - \frac{\Delta^2 t}{L_F}. \tag{41}$$

Consider the k-th coordinate of $\mathbf{w}_\tau$ and $\mathbf{w}_{\tau+1}$, by Cauchy-Schwarz inequality, we have

$$\sum_{\tau=0}^{t-1}(w_{\tau,k} - w_{\tau+1,k})^2 \geq \frac{1}{t}(w_{0,k} - w_{t,k})^2,$$

which implies

$$\sum_{\tau=0}^{t-1}\|\mathbf{w}_\tau - \mathbf{w}_{\tau+1}\|_2^2 \geq \frac{1}{t}\|\mathbf{w}_0 - \mathbf{w}_t\|_2^2. \tag{42}$$

Combining (41) and (42), we obtain

$$F(\mathbf{w}_0) - F(\mathbf{w}_t) \geq \frac{L_F}{4t}\|\mathbf{w}_0 - \mathbf{w}_t\|_2^2 - \frac{\Delta^2 t}{L_F} \geq \frac{L_F}{4T_{\mathrm{th}}}R^2 - \frac{\Delta^2 T_{\mathrm{th}}}{L_F}. \tag{43}$$

On the other hand, by the smoothness of $F(\cdot)$, we have

$$F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_0) \geq \langle \nabla F(\widetilde{\mathbf{w}}), \widetilde{\mathbf{w}} - \mathbf{w}_0\rangle - \frac{L_F}{2}\|\mathbf{w}_0 - \widetilde{\mathbf{w}}\|_2^2 \geq -(\epsilon + \Delta)r - \frac{L_F}{2}r^2. \tag{44}$$

Adding up (43) and (44), we obtain

$$F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t) \geq \frac{L_F}{4T_{\mathrm{th}}}R^2 - \frac{\Delta^2 T_{\mathrm{th}}}{L_F} - (\epsilon + \Delta)r - \frac{L_F}{2}r^2, \tag{45}$$

which completes the proof.

D Proof of Theorem 4

First, when we run gradient descent iterations $\mathbf{w}' = \mathbf{w} - \eta\nabla F(\mathbf{w})$, according to Lemma 1, we have

$$F(\mathbf{w}') \leq F(\mathbf{w}) - \frac{1}{2L_F}\|\nabla F(\mathbf{w})\|_2^2. \tag{46}$$

Suppose at $\widetilde{\mathbf{w}}$, we observe that $\|\nabla F(\widetilde{\mathbf{w}})\|_2 \leq \epsilon$, and then we start the Escape process. When we have exact gradient oracle, we can still define the stuck region $\mathbb{W}_S$ at $\widetilde{\mathbf{w}}$ as in the definition of stuck region in Appendix C, by simply replacing the inexact gradient oracle with the exact oracle. Then, we can analyze the size of the stuck region according to Lemma 2. Assume that the smallest eigenvalue of $\mathbf{H} := \nabla^2 F(\widetilde{\mathbf{w}})$ satisfies $\lambda_{\min}(\mathbf{H}) \leq -\gamma < 0$, and let the unit vector $\mathbf{e}$ be the eigenvector associated with $\lambda_{\min}(\mathbf{H})$. Let $\mathbf{u}_0, \mathbf{y}_0 \in \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ be two points such that $\mathbf{y}_0 = \mathbf{u}_0 + \mu_0\mathbf{e}$

with some $\mu_0 \geq \mu \in (0, r)$. Consider the stuck region $\mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$. Suppose that r, R, T_{th}, and μ satisfy

$$T_{\mathrm{th}} = \frac{2}{\eta\gamma} \log_{9/4}(\frac{2(R+r)}{\mu}), \tag{47}$$

$$R \geq \mu, \tag{48}$$

$$\gamma \geq 24\rho_F(R+r) \log_{9/4}(\frac{2(R+r)}{\mu}). \tag{49}$$

Then, there must be either $\mathbf{u}_0 \notin \mathbb{W}_S$ or $\mathbf{y}_0 \notin \mathbb{W}_S$. In addition, according to Lemma 3, if conditions (47)-(49) are satisfied, then, when we sample $\mathbf{w}_0$ from $\mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ uniformly at random, the probability that $\mathbf{w}_0 \in \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$ is at most $\frac{2\mu\sqrt{d}}{r}$. In addition, according to (45) in the proof of Lemma 4, assume that $\mathbf{w}_0 \in \mathbb{B}_{\widetilde{\mathbf{w}}}(r)$ and that $\mathbf{w}_0 \notin \mathbb{W}_S(\widetilde{\mathbf{w}}, r, R, T_{\mathrm{th}})$. Let $t \leq T_{\mathrm{th}}$ be the step such that $\|\mathbf{w}_t - \mathbf{w}_0\|_2 \geq R$. Then, we have

$$F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t) \geq \frac{L_F}{4T_{\mathrm{th}}} R^2 - \epsilon r - \frac{L_F}{2} r^2. \tag{50}$$

Combining (47) and (49), we know that the first term on the right hand side of (50) satisfies

$$\frac{L_F}{4T_{\mathrm{th}}} R^2 \geq 3\rho_F R^3. \tag{51}$$

Choose $R = \sqrt{\epsilon/\rho_F}$ and $r = \epsilon$. Then, we know that when $\epsilon \leq \min\{\frac{1}{\rho_F}, \frac{4}{L_F^2\rho_F}\}$, we have $\epsilon r \leq \rho_F R^3$ and $\frac{1}{2}L_F r^2 \leq \rho_F R^3$. Combining these facts with (50) and (51), we know that, when the algorithm successfully escapes the saddle point, the decrease in function value satisfies

$$F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t) \geq \rho_F R^3. \tag{52}$$

Therefore, the average function value decrease during the Escape process is at least

$$\frac{F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_t)}{T_{\mathrm{th}}} \geq \frac{12}{L_F}\epsilon^2. \tag{53}$$

When we have exact gradient oracle, we choose $Q = 1$. According to (46) and (53), for the iterations that are not in the Escape process, the function value decrease in each iteration is at least $\frac{1}{2L_F}\epsilon^2$; for the iterations in the Escape process, the function value decrease on average is $\frac{12}{L_F}\epsilon^2$. Since the function value can decrease at most $F_0 - F^*$, the algorithm must terminate within $\frac{2L_F(F_0-F^*)}{\epsilon^2}$ iterations.

The we proceed to analyze the failure probability. We can see that the number of saddle points that the algorithm may need to escape is at most $\frac{F_0-F^*}{\rho_F R^3}$. Then, by union bound the probability that the algorithm fails to escape one of the saddle points is at most

$$\frac{2\mu\sqrt{d}}{r} \cdot \frac{F_0 - F^*}{\rho_F R^3}$$

By letting the above probability to be δ, we obtain

$$\mu = \frac{\delta\epsilon^{5/2}}{2\sqrt{\rho_F d}(F_0 - F^*)},$$

which completes the proof.

E Proof of Proposition 1

We consider the following class of one-dimensional functions indexed by $s \in \mathbb{R}$:

$$\mathcal{F} = \{f_s(\cdot) : f_s(w) = \Delta^{3/2} \sin(\Delta^{-1/2} w + s), s \in \mathbb{R}\}.$$

Then, for each function $f_s(\cdot) \in \mathcal{F}$, we have

$$\nabla f_s(w) = \Delta \cos(\Delta^{-1/2} w + s),$$

and

$$\nabla^2 f_s(w) = -\Delta^{1/2}\sin(\Delta^{-1/2}w + s).$$

Thus, we always have $|\nabla f_s(w)| \le \Delta, \forall w$. Therefore, the Δ-inexact gradient oracle can simply output 0 all the time. In addition, we verify that for all s and w, $|\nabla^2 f_s(w)| \le \Delta^{1/2} \le 1$ and $|\nabla^3 f_s(w)| = |-\cos(\Delta^{-1/2}w + s)| \le 1$ under the assumption that $\Delta \le 1$, so all the functions in $\mathcal{F}$ are 1-smooth and 1-Hessian Lipschitz as claimed.

In this case, the output of the algorithm does not depend on s, that is, the actual function that we aim to minimize. Consequently, for any output $\widetilde{w}$ of the algorithm, there exists $s \in \mathbb{R}$ such that $\Delta^{-1/2}\widetilde{w} + s = \pi/4$, and thus $|\nabla f_s(\widetilde{w})| = \Delta/\sqrt{2}$ and $\lambda_{\min}(\nabla^2 f_s(\widetilde{w})) = -\Delta^{1/2}/\sqrt{2}$.

F Proof of Proposition 2

Suppose that during all the iterations, the Escape process is called $E+1$ times. In the first E times, the algorithm escapes the saddle points, and in the last Escape process, the algorithm does not escape and outputs $\widetilde{\mathbf{w}}$. For the first E processes, there might be up to Q rounds of perturb-and-descent operations, and we only consider the successful descent round. We can then partition the algorithm into $E+1$ segments. We denote the starting and ending iterates of the t-th segment by $\mathbf{w}_t$ and $\widetilde{\mathbf{w}}_t$, respectively, and denote the length (number of inexact gradient descent iterations) by T_t. When the algorithm reaches $\widetilde{\mathbf{w}}_t$, we randomly perturb $\widetilde{\mathbf{w}}_t$ to $\mathbf{w}_{t+1}$, and thus we have $\|\widetilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2 \le r$ for every $t = 0, 1, \ldots, E-1$. According to (22), we know that

$$\sum_{t=0}^{E} T_t \le \frac{2(F_0 - F^*)L_F}{3\Delta^2} := \widetilde{T},$$

and according to (20), we have

$$E \le \frac{\rho_F(F_0 - F^*)}{48L_F(\Delta^{6/5}d^{3/5} + \Delta^{7/5}d^{7/10})}.$$

According to (43), we know that

$$F(\mathbf{w}_t) - F(\widetilde{\mathbf{w}}_t) \ge \frac{L_F}{4T_t}\|\mathbf{w}_t - \widetilde{\mathbf{w}}_t\|_2^2 - \frac{\Delta^2 T_t}{L_F},$$

which implies

$$\|\mathbf{w}_t - \widetilde{\mathbf{w}}_t\|_2 \le \frac{2}{\sqrt{L_F}}\sqrt{T_t(F(\mathbf{w}_t) - F(\widetilde{\mathbf{w}}_t))} + \frac{2\Delta T_t}{L_F}.$$

Then, by Cauchy-Schwarz inequality, we have

$$\sum_{t=0}^{E}\|\mathbf{w}_t - \widetilde{\mathbf{w}}_t\|_2 \le 2\sqrt{\frac{\widetilde{T}}{L_F}\sum_{t=0}^{E}(F(\mathbf{w}_t) - F(\widetilde{\mathbf{w}}_t))} + \frac{2\Delta\widetilde{T}}{L_F}. \tag{54}$$

On the other hand, we have

$$\sum_{t=0}^{E}(F(\mathbf{w}_t) - F(\widetilde{\mathbf{w}}_t)) + \sum_{t=0}^{E-1}(F(\widetilde{\mathbf{w}}_t) - F(\mathbf{w}_{t+1})) = F(\mathbf{w}_0) - F(\widetilde{\mathbf{w}}_E) \le F(\mathbf{w}_0) - F^*.$$

According to (44), we have

$$F(\widetilde{\mathbf{w}}_t) - F(\mathbf{w}_{t+1}) \ge -4\Delta r - \frac{L_F}{2}r^2,$$

and thus

$$\sum_{t=0}^{E}(F(\mathbf{w}_t) - F(\widetilde{\mathbf{w}}_t)) \le F(\mathbf{w}_0) - F^* + E(4\Delta r + \frac{L_F}{2}r^2) \tag{55}$$

Combining (54) and (56), and using the bounds for $\widetilde{T}$ and E, we obtain that

$$\sum_{t=0}^{E}\|\mathbf{w}_t - \widetilde{\mathbf{w}}_t\|_2 \le C_1\frac{F(\mathbf{w}_0) - F^*}{\Delta}, \tag{56}$$

where $C_1 > 0$ is a quantity that only depends on L_F and ρ_F. In addition, we have

$$\sum_{t=0}^{E-1} \|\widetilde{\mathbf{w}}_t - \mathbf{w}_{t+1}\|_2 \leq Er \leq C_2 \frac{F(\mathbf{w}_0) - F^*}{\Delta^{3/5} d^{3/10} + \Delta^{4/5} d^{2/5}}, \tag{57}$$

where $C_2 > 0$ is a quantity that only depends on L_F and ρ_F. Combining (56) and (57), and using triangle inequality, we know that

$$\|\widetilde{\mathbf{w}}_E - \mathbf{w}_0\|_2 \leq C_1 \frac{F(\mathbf{w}_0) - F^*}{\Delta} + C_2 \frac{F(\mathbf{w}_0) - F^*}{\Delta^{3/5} d^{3/10} + \Delta^{4/5} d^{2/5}} \leq C \frac{F(\mathbf{w}_0) - F^*}{\Delta}.$$

Here, the last inequality is due to the fact that we consider the regime where $\Delta \to 0$, and C is a quantity that only depends on L_F and ρ_F. As a final note, the analysis above also applies to any iterate prior to the final output, and thus, all the iterates during the algorithm stays in the ℓ_2 ball centered at $\mathbf{w}_0$ with radius $C\frac{F(\mathbf{w}_0)-F^*}{\Delta}$.

G Robust Estimation of Gradients

G.1 Iterative Filtering Algorithm

We describe an iterative filtering algorithm for robust mean estimation. The algorithm is originally proposed for robust mean estimation for Gaussian distribution in [5], and extended to sub-Gaussian distribution in [6]; then algorithm is reinterpreted in [18]. Here, we present the algorithm using the interpretation in [18]. Suppose that m random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^d$ are drawn i.i.d. from some distribution with mean $\boldsymbol{\mu}$. An adversary observes all these vectors and changes an α fraction of them in an arbitrary fashion, and we only have access to the corrupted data points $\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2, \ldots, \widehat{\mathbf{x}}_m$. The goal of the iterative filtering algorithm is to output an accurate estimate of the true mean $\boldsymbol{\mu}$ even when the dimension d is large. We provide the detailed procedure in Algorithm 1. Here, we note that the algorithm parameter σ needs to be chosen properly in order to achieve the best possible statistical error rate.

Algorithm 1 Iterative Filtering [5, 6, 18]

Require: corrupted data $\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2, \ldots, \widehat{\mathbf{x}}_m \in \mathbb{R}^d$, $\alpha \in [0, \frac{1}{4})$, and algorithm parameter $\sigma > 0$.
  $\mathcal{A} \leftarrow [m]$, $c_i \leftarrow 1$, and $\tau_i \leftarrow 0$, $\forall\, i \in \mathcal{A}$.
  while true **do**
    Let $\mathbf{W} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$ be a minimizer of the convex optimization problem:

$$\min_{\substack{0 \leq W_{ji} \leq \frac{3+\alpha}{(1-\alpha)(3-\alpha)m} \\ \sum_{j \in \mathcal{A}} W_{ji} = 1}} \max_{\substack{\mathbf{U} \succeq 0 \\ \mathrm{tr}(\mathbf{U}) \leq 1}} \sum_{i \in \mathcal{A}} c_i (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji})^\top \mathbf{U} (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji}),$$

    and $\mathbf{U} \in \mathbb{R}^{d \times d}$ be a maximizer of the convex optimization problem:

$$\max_{\substack{\mathbf{U} \succeq 0 \\ \mathrm{tr}(\mathbf{U}) \leq 1}} \min_{\substack{0 \leq W_{ji} \leq \frac{3+\alpha}{(1-\alpha)(3-\alpha)m} \\ \sum_{j \in \mathcal{A}} W_{ji} = 1}} \sum_{i \in \mathcal{A}} c_i (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji})^\top \mathbf{U} (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji}).$$

    $\forall\, i \in \mathcal{A},\ \tau_i \leftarrow (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji})^\top \mathbf{U} (\widehat{\mathbf{x}}_i - \sum_{j \in \mathcal{A}} \widehat{\mathbf{x}}_j W_{ji})$.
    if $\sum_{i \in \mathcal{A}} c_i \tau_i > 8m\sigma^2$ **then**
      $\forall\, i \in \mathcal{A},\ c_i \leftarrow (1 - \frac{\tau_i}{\tau_{\max}}) c_i$, where $\tau_{\max} = \max_{i \in \mathcal{A}} \tau_i$.
      $\mathcal{A} \leftarrow \mathcal{A} \setminus \{i : c_i \leq \frac{1}{2}\}$.
    else
      return $\widehat{\boldsymbol{\mu}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \widehat{\mathbf{x}}_i$
    end if
  end while

G.2 Proof of Theorem 5

To prove Theorem 5, we first state a result that bounds the error of the iterative filtering algorithm when the original data points $\{\mathbf{x}_i\}$ are deterministic. The following lemma is proved in [6, 18];

also see [19] for additional discussion.

Lemma 5. *[6, 18] Let $\mathcal{S} := \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ be the set of original data points and $\boldsymbol{\mu}_{\mathcal{S}} := \frac{1}{m}\sum_{i=1}^m \mathbf{x}_i$ be their sample mean. Let $\widehat{\mathbf{x}}_1, \widehat{\mathbf{x}}_2, \ldots, \widehat{\mathbf{x}}_m$ be the corrupted data. If $\alpha \le \frac{1}{4}$, and the algorithm parameter σ is chosen such that*

$$\left\| \frac{1}{m}\sum_{i=1}^m (\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{S}})(\mathbf{x} - \boldsymbol{\mu}_{\mathcal{S}})^\top \right\|_2 \le \sigma^2, \tag{58}$$

then the output of the iterative filtering algorithm satisfies $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 \le \mathcal{O}(\sigma\sqrt{\alpha})$.

By triangle inequality, we have

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \le \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}_{\mathcal{S}}\|_2 + \|\boldsymbol{\mu}_{\mathcal{S}} - \boldsymbol{\mu}\|_2, \tag{59}$$

and

$$\begin{aligned}
\left\| \frac{1}{m}\sum_{i=1}^m (\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{S}})(\mathbf{x} - \boldsymbol{\mu}_{\mathcal{S}})^\top \right\|_2 =& \frac{1}{m}\left\|([\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}_{\mathcal{S}}\mathbf{1}^\top)([\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}_{\mathcal{S}}\mathbf{1}^\top)^\top\right\|_2 \\
=& \frac{1}{m}\left\|[\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}_{\mathcal{S}}\mathbf{1}^\top\right\|_2^2 \\
\le& \frac{1}{m}\Big(\|[\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}\mathbf{1}^\top\|_2 + \sqrt{m}\|\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathcal{S}}\|_2\Big)^2,
\end{aligned} \tag{60}$$

where $\mathbf{1}$ denotes the all-one vector.[3] By choosing

$$\sigma = \Theta\Big(\frac{1}{\sqrt{m}}\|[\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}\mathbf{1}^\top\|_2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathcal{S}}\|_2\Big)$$

in Lemma 5 and combining with the bounds (59) and (60), we obtain that

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \lesssim \frac{\sqrt{\alpha}}{\sqrt{m}}\|[\mathbf{x}_1, \cdots, \mathbf{x}_m] - \boldsymbol{\mu}\mathbf{1}^\top\|_2 + \|\boldsymbol{\mu} - \boldsymbol{\mu}_{\mathcal{S}}\|_2. \tag{61}$$

With the above bound in hand, we now turn to the robust gradient estimation problem, where the data points are drawn i.i.d. from some unknown distribution. Let $\widehat{\mathbf{g}}(\mathbf{w}) := \mathsf{filter}\{\widehat{\mathbf{g}}_i(\mathbf{w})\}_{i=1}^m$, where filter represents the iterative filtering algorithm. In light of (61), we know that in order to bound the gradient estimation error $\sup_{\mathbf{w}\in\mathcal{W}}\|\widehat{\mathbf{g}}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2$, it suffices to bound the quantities

$$\sup_{\mathbf{w}\in\mathcal{W}}\|[\nabla F_1(\mathbf{w}), \cdots, \nabla F_m(\mathbf{w})] - \nabla F(\mathbf{w})\mathbf{1}^\top\|_2$$

and

$$\sup_{\mathbf{w}\in\mathcal{W}}\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|_2.$$

Here, we recall that $\nabla F_i(\mathbf{w})$ is the true gradient of the empirical loss function on the i-th machine, and $\widehat{\mathbf{g}}_i(\mathbf{w})$ is the (possibly) corrupted gradient.

We first bound $\sup_{\mathbf{w}\in\mathcal{W}}\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|_2$. Note that we have $\frac{1}{m}\sum_{i=1}^m \nabla F_i(\mathbf{w}) = \frac{1}{nm}\sum_{i=1}^m\sum_{j=1}^n \nabla f(\mathbf{w}; \mathbf{z}_{i,j})$. Using the same method as in the proof of Lemma 6 in [4], we can show that for each fixed $\mathbf{w}$, with probability at least $1-\delta$,

$$\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le \frac{2\sqrt{2}\zeta}{\sqrt{nm}}\sqrt{d\log 6 + \log\Big(\frac{1}{\delta}\Big)}.$$

For some $\delta_0 > 0$ to be chosen later, let $\mathcal{W}_{\delta_0} = \{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{N_{\delta_0}}\}$ be a finite subset of $\mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$, there exists some $\mathbf{w}^\ell \in \mathcal{W}_{\delta_0}$ such that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \le \delta_0$. Standard ϵ-net results from [20] ensure that $N_{\delta_0} \le (1 + \frac{D}{\delta_0})^d$. Then, by the union bound, we have with probability $1-\delta$, for all $\mathbf{w}^\ell \in \mathcal{W}_{\delta_0}$,

$$\|\frac{1}{m}\sum_{i=1}^m \nabla F_i(\mathbf{w}^\ell) - \nabla F(\mathbf{w}^\ell)\|_2 \le \frac{2\sqrt{2}\zeta}{\sqrt{nm}}\sqrt{d\log 6 + \log\Big(\frac{N_{\delta_0}}{\delta}\Big)}. \tag{62}$$

[3] We note that similar derivation also appears in [19].

When (62) holds, by the smoothness of $f(\cdot;\mathbf{z})$ we know that for all $\mathbf{w}\in\mathcal{W}$,

$$\|\frac{1}{m}\sum_{i=1}^{m}\nabla F_i(\mathbf{w})-\nabla F(\mathbf{w})\|_2\le\frac{2\sqrt{2}\zeta}{\sqrt{nm}}\sqrt{d\log 6+\log\left(\frac{N_{\delta_0}}{\delta}\right)}+2L\delta_0.$$

By choosing $\delta_0=\frac{1}{nmL}$ and $\delta=\frac{1}{(1+mnDL)^d}$, we obtain that with probability at least $1-\frac{1}{(1+mnDL)^d}$, for all $\mathbf{w}\in\mathcal{W}$,

$$\|\frac{1}{m}\sum_{i=1}^{m}\nabla F_i(\mathbf{w})-\nabla F(\mathbf{w})\|_2\lesssim\frac{\zeta}{\sqrt{nm}}\sqrt{d\log(1+nmDL)}. \tag{63}$$

We next bound $\sup_{\mathbf{w}\in\mathcal{W}}\|[\nabla F_1(\mathbf{w}),\cdots,\nabla F_m(\mathbf{w})]-\nabla F(\mathbf{w})\mathbf{1}^\top\|_2$. We note that when the gradients are sub-Gaussian distributed, similar results for the centralized setting have been established in [3]. One can check that for every i, $\nabla F_i(\mathbf{w})-\nabla F(\mathbf{w})$ is $\frac{\zeta}{\sqrt{n}}$-sub-Gaussian. Define $\mathbf{G}(\mathbf{w}):=[\nabla F_1(\mathbf{w}),\cdots,\nabla F_m(\mathbf{w})]-\nabla F(\mathbf{w})\mathbf{1}^\top$. Using a standard concentration inequality for the norm of a matrix with independent sub-Gaussian columns [20], we obtain that for each fixed $\mathbf{w}$, with probability at least $1-\delta$,

$$\|\frac{1}{m}\mathbf{G}(\mathbf{w})\mathbf{G}(\mathbf{w})^\top-\frac{1}{n}\mathbf{\Sigma}(\mathbf{w})\|_2\lesssim\frac{\zeta^2}{n}\left(\sqrt{\frac{d}{m}}+\frac{d}{m}+\frac{1}{m}\log\left(\frac{1}{\delta}\right)+\sqrt{\frac{1}{m}\log\left(\frac{1}{\delta}\right)}\right),$$

which implies that

$$\frac{1}{\sqrt{m}}\|\mathbf{G}(\mathbf{w})\|_2\lesssim\frac{\sigma}{\sqrt{n}}+\frac{\zeta}{\sqrt{n}}\left(\sqrt{\frac{d}{m}}+\frac{d}{m}+\frac{1}{m}\log\left(\frac{1}{\delta}\right)+\sqrt{\frac{1}{m}\log\left(\frac{1}{\delta}\right)}\right)^{1/2}.$$

Recall the δ_0-net $\mathcal{W}_{\delta_0}=\{\mathbf{w}^1,\mathbf{w}^2,\ldots,\mathbf{w}^{N_{\delta_0}}\}$ as defined above. Then, we have with probability at least $1-\delta$, for all $\mathbf{w}^\ell\in\mathcal{W}_{\delta_0}$

$$\frac{1}{\sqrt{m}}\|\mathbf{G}(\mathbf{w}^\ell)\|_2\lesssim\frac{\sigma}{\sqrt{n}}+\frac{\zeta}{\sqrt{n}}\left(\sqrt{\frac{d}{m}}+\frac{d}{m}+\frac{1}{m}\log\left(\frac{N_{\delta_0}}{\delta}\right)+\sqrt{\frac{1}{m}\log\left(\frac{N_{\delta_0}}{\delta}\right)}\right)^{1/2}. \tag{64}$$

For each $\mathbf{w}$ with $\|\mathbf{w}^\ell-\mathbf{w}\|_2\le\delta_0$, we have

$$\begin{aligned}\|\mathbf{G}(\mathbf{w}^\ell)-\mathbf{G}(\mathbf{w})\|_2\le&\|\mathbf{G}(\mathbf{w}^\ell)-\mathbf{G}(\mathbf{w})\|_F\\&\le\left(\sum_{i=1}^{m}\|(\nabla F_i(\mathbf{w}^\ell)-\nabla F(\mathbf{w}^\ell))-(\nabla F_i(\mathbf{w})-\nabla F(\mathbf{w}))\|_2^2\right)^{1/2}\\&\le 2L\delta_0\sqrt{m}.\end{aligned}$$

This implies that when the bound (64) holds, we have for all $\mathbf{w}\in\mathcal{W}$,

$$\frac{1}{\sqrt{m}}\|\mathbf{G}(\mathbf{w})\|_2\lesssim\frac{\sigma}{\sqrt{n}}+\frac{\zeta}{\sqrt{n}}\left(\sqrt{\frac{d}{m}}+\frac{d}{m}+\frac{1}{m}\log\left(\frac{N_{\delta_0}}{\delta}\right)+\sqrt{\frac{1}{m}\log\left(\frac{N_{\delta_0}}{\delta}\right)}\right)^{1/2}+2L\delta_0. \tag{65}$$

Choose $\delta_0=\frac{1}{nmL}$, in which case the last term above is a high order term. In this case, choosing $\delta=\frac{1}{(1+mnDL)^d}$, we have with probability at least $1-\frac{1}{(1+mnDL)^d}$, for all $\mathbf{w}\in\mathcal{W}$,

$$\begin{aligned}\frac{1}{\sqrt{m}}\|\mathbf{G}(\mathbf{w})\|_2\lesssim&\frac{\sigma}{\sqrt{n}}+\frac{\zeta}{\sqrt{n}}\left(\left(\frac{d}{m}+\sqrt{\frac{d}{m}}\right)\log(1+nmDL)\right)^{1/2}\\\lesssim&\frac{\sigma}{\sqrt{n}}+\frac{\zeta}{\sqrt{n}}\left(1+\sqrt{\frac{d}{m}}\right)\sqrt{\log(1+nmDL)}.\end{aligned} \tag{66}$$

Combining the bounds (61), (63), and (66), we obtain that with probability at least $1-\frac{2}{(1+mnDL)^d}$,

$$\sup_{\mathbf{w}\in\mathcal{W}}\|\widehat{\mathbf{g}}(\mathbf{w})-\nabla F(\mathbf{w})\|_2\lesssim\left((\sigma+\zeta)\sqrt{\frac{\alpha}{n}}+\zeta\sqrt{\frac{d}{nm}}\right)\sqrt{\log(1+nmDL)},$$

which completes the proof.

G.3 Lower Bound for First-Order Guarantee

In this section we prove Observation 2. We consider the simple mean estimation problem with random vector $\mathbf{z}$ drawn from a distribution $\mathcal{D}$ with mean $\boldsymbol{\mu}$. The loss function associated with $\mathbf{z}$ is $f(\mathbf{w};\mathbf{z}) = \frac{1}{2}\|\mathbf{w}-\mathbf{z}\|_2^2$. The population loss is $F(\mathbf{w}) = \frac{1}{2}(\|\mathbf{w}\|_2^2 - 2\boldsymbol{\mu}^\top\mathbf{w} + \mathbb{E}[\|\mathbf{z}\|_2^2])$, and $\nabla F(\mathbf{w}) = \mathbf{w}-\boldsymbol{\mu}$. We first provide a lower bound for distributed mean estimation in the Byzantine setting, which is proved in [21].

Lemma 6. *[21] Suppose that $\mathbf{z}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $\sigma^2\mathbf{I}$. Then, any algorithm that outputs an estimate $\widetilde{\mathbf{w}}$ of $\boldsymbol{\mu}$ has a constant probability such that*

$$\|\widetilde{\mathbf{w}}-\boldsymbol{\mu}\|_2 = \Omega\Big(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d}{nm}}\Big).$$

Since $\nabla F(\widetilde{\mathbf{w}}) = \widetilde{\mathbf{w}}-\boldsymbol{\mu}$, the above bound directly implies the lower bound on $\|\nabla F(\widetilde{\mathbf{w}})\|_2$ in Observation 2.

G.4 Median and Trimmed Mean

In this section, we present the error bounds of median and trimmed mean operations in the Byzantine setting in [21] for completeness.

Condition 1. *For any $\mathbf{z}\in\mathcal{Z}$, the k-th partial derivative $\partial_k f(\cdot;\mathbf{z})$ is L_k-Lipschitz for each $k\in[d]$. Let $\widehat{L} := (\sum_{k=1}^d L_k^2)^{1/2}$.*

For the median-based algorithm, one needs to use the notion of the *absolute skewness* of a one-dimensional random variable X, defined as $S(X) := \mathbb{E}[|X-\mathbb{E}[X]|^3]/\mathrm{Var}(X)^{3/2}$. Define the following upper bounds on the standard deviation and absolute skewness of the gradients:

$$v := \sup_{\mathbf{w}\in\mathcal{W}} \big(\mathbb{E}[\|\nabla f(\mathbf{w};\mathbf{z}) - \nabla F(\mathbf{w})\|_2^2]\big)^{1/2}, \quad s := \sup_{\mathbf{w}\in\mathcal{W}} \max_{k\in[d]} S\big(\partial f_k(\mathbf{w};\mathbf{z})\big).$$

Then one has the following guarantee for the median-based algorithm.

Claim 3 (median)**.** *[21] Suppose that Condition 1 holds. Assume that*

$$\alpha + \left(\frac{d\log(1+nmD\widehat{L})}{m(1-\alpha)}\right)^{1/2} + c_1\frac{s}{\sqrt{n}} \le \frac{1}{2} - c_2$$

for some constant $c_1, c_2 > 0$. Then, with probability $1-o(1)$, $\mathsf{GradAGG}\equiv\mathsf{med}$ provides a Δ_{med}-inexact gradient oracle with

$$\Delta_{\mathsf{med}} \le \frac{c_3}{\sqrt{n}} v\Big(\alpha + \big(\frac{d\log(nmD\widehat{L})}{m}\big)^{1/2} + \frac{s}{\sqrt{n}}\Big) + \mathcal{O}\big(\frac{1}{nm}\big),$$

where c_3 is an absolute constant.

Therefore, the median operation provides a $\widetilde{\mathcal{O}}(v(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d}{nm}} + \frac{s}{n}))$-inexact gradient oracle. If each partial derivative is of size $\mathcal{O}(1)$, the quantity v is of the order $\mathcal{O}(\sqrt{d})$ and thus one has $\Delta_{\mathsf{med}} \lesssim \frac{\alpha\sqrt{d}}{\sqrt{n}} + \frac{d}{\sqrt{nm}} + \frac{\sqrt{d}}{n}$.

For the trimmed mean algorithm, one needs to assume that the gradients of the loss functions are sub-exponential.

Condition 2. *For any $\mathbf{w}\in\mathcal{W}$, $\nabla f(\mathbf{w};\mathbf{z})$ is ξ-sub-exponential.*

In this setting, there is the following guarantee.

Claim 4 (trimmed mean)**.** *[21] Suppose that Conditions 1 and 2 hold. Choose $\beta = c_4\alpha \le \frac{1}{2} - c_5$ with some constant $c_4 \ge 1$, $c_5 > 0$. Then, with probability $1-o(1)$, $\mathsf{GradAGG}\equiv\mathsf{trmean}_\beta$ provides a Δ_{tm}-inexact gradient oracle with*

$$\Delta_{\mathsf{tm}} \le c_6\xi d\Big(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}}\Big)\sqrt{\log(nmD\widehat{L})},$$

where c_6 is an absolute constant.

Therefore, the trimmed mean operation provides a $\widetilde{\mathcal{O}}(\xi d(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}}))$-inexact gradient oracle.

References

[1] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems*, pages 2107–2116, 2017.

[2] K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[3] M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60. ACM, 2017.

[4] Y. Chen, L. Su, and J. Xu. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *arXiv preprint arXiv:1705.05491*, 2017.

[5] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 655–664. IEEE, 2016.

[6] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Being robust (in high dimensions) can be practical. *arXiv preprint arXiv:1703.00893*, 2017.

[7] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, J. Steinhardt, and A. Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[8] P. J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[9] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

[10] A. Klivans, P. K. Kothari, and R. Meka. Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*, 2018.

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

[12] J. Li. Robust sparse estimation tasks in high dimensions. *arXiv preprint arXiv:1702.05860*, 2017.

[13] L. Liu, Y. Shen, T. Li, and C. Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

[14] G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *arXiv preprint arXiv:1608.00757*, 2016.

[15] S. Minsker et al. Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

[16] S. Minsker and N. Strawn. Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*, 2017.

[17] A. Nemirovskii, D. B. Yudin, and E. R. Dawson. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

[18] J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

[19] L. Su and J. Xu. Securing distributed machine learning in high dimensions. *arXiv preprint arXiv:1804.10140*, 2018.

[20] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[21] D. Yin, Y. Chen, R. Kannan, and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5650–5659, 2018.