Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates
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

Notation We denote vectors by boldface lowercase letters such as $\mathbf{w}$, and the elements in the vector are denoted by italics letters with subscripts, such as w_k. Matrices are denoted by boldface uppercase letters such as $\mathbf{H}$. For any positive integer N, we denote the set $\{1, 2, \ldots, N\}$ by $[N]$. For vectors, we denote the ℓ_2 norm and ℓ_∞ norm by $\|\cdot\|_2$ and $\|\cdot\|_\infty$, respectively. For matrices, we denote the operator norm and the Frobenius norm by $\|\cdot\|_2$ and $\|\cdot\|_F$, respectively. We denote by $\Phi(\cdot)$ the CDF of standard Gaussian distribution. For any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative with respect to the k-th argument by $\partial_k f$.

A Variance, Skewness, and Sub-exponential Property

A.1 Proof of Proposition 1

We use the simplified notation $f(\mathbf{w}) := f(\mathbf{w}; \mathbf{x}, y)$. One can directly compute the gradients:

$$\nabla f(\mathbf{w}) = \mathbf{x}(\mathbf{x}^{\mathrm{T}}\mathbf{w} - y) = \mathbf{x}\mathbf{x}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^*) - \xi\mathbf{x},$$

and thus

$$\nabla F(\mathbf{w}) = \mathbb{E}[\nabla f(\mathbf{w})] = \mathbf{w} - \mathbf{w}^*.$$

Define $\Delta(\mathbf{w}) := \nabla f(\mathbf{w}) - \nabla F(\mathbf{w})$ with its k-th element being $\Delta_k(\mathbf{w})$. We now compute the variance and absolute skewness of $\Delta_k(\mathbf{w})$.

We can see that

$$\Delta_k(\mathbf{w}) = \sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*) + (x_k^2 - 1)(w_k - w_k^*) - \xi x_k. \tag{1}$$

Thus,

$$\mathbb{E}[\Delta_k^2(\mathbf{w})] = \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_k^2 x_i^2 (w_i - w_i^*)^2 + \xi^2 x_k^2] = \|\mathbf{w} - \mathbf{w}^*\|_2^2 - (w_k - w_k^*)^2 + \sigma^2, \tag{2}$$

which yields

$$\mathrm{Var}(\nabla f(\mathbf{w})) = \mathbb{E}[\|\nabla f(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2] = (d-1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2.$$

Then we proceed to bound $\gamma(\Delta_k(\mathbf{w}))$. By Jensen's inequality, we know that

$$\gamma(\Delta_k(\mathbf{w})) = \frac{\mathbb{E}[|\Delta_k(\mathbf{w})|^3]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^{3/2}} \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \tag{3}$$

We first find a lower bound for $\mathrm{Var}(\Delta_k(\mathbf{w}))^3$. According to (2), we know that

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 = \big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2 + \sigma^2\big)^3 \ge \big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2\big)^3 + \sigma^6.$$

Define the following three quantities.

$$W_1 = \sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^6 \tag{4}$$

$$W_2 = \sum_{\substack{1\le i,j\le d\\ i,j\neq k\\ i\neq j}} (w_i - w_i^*)^4 (w_j - w_j^*)^2 \tag{5}$$

$$W_3 = \sum_{\substack{1\le i,j,\ell\le d\\ i,j,\ell\neq k\\ i\neq j, i\neq \ell, j\neq \ell}} (w_i - w_i^*)^2 (w_j - w_j^*)^2 (w_\ell - w_\ell^*)^2 \tag{6}$$

By simple algebra, one can check that

$$\big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2\big)^3 = W_1 + 3W_2 + W_3, \tag{7}$$

and thus

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 \ge W_1 + 3W_2 + W_3 + \sigma^6. \tag{8}$$

Then, we find an upper bound on $\mathbb{E}[\Delta_k^6(\mathbf{w})]$. According to (1), and Hölder's inequality, we know that

$$\begin{aligned}\mathbb{E}[\Delta_k^6(\mathbf{w})] &= \mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*) - \xi x_k)^6] \le 32\big(\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*))^6] + \mathbb{E}[\xi^6 x_k^6]\big)\\ &= 32\big(\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i (w_i - w_i^*))^6] + 15\sigma^6\big),\end{aligned} \tag{9}$$

where in the last inequality we use the moments of Gaussian random variables. Then, we compute the first term in (9). By algebra, one can obtain

$$\begin{aligned}\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i (w_i - w_i^*))^6] =& \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_i^6 (w_i - w_i^*)^6] + 15\mathbb{E}[\sum_{\substack{1\le i,j\le d\\ i,j\neq k\\ i\neq j}} x_i^4 x_j^2 (w_i - w_i^*)^4 (w_j - w_j^*)^2]\\ &+ 15\mathbb{E}[\sum_{\substack{1\le i,j,\ell\le d\\ i,j,\ell\neq k\\ i\neq j, i\neq \ell, j\neq \ell}} x_i^2 x_j^2 x_\ell^2 (w_i - w_i^*)^2 (w_j - w_j^*)^2 (w_\ell - w_\ell^*)^2]\\ =& W_1 + 15W_2 + 15W_3.\end{aligned} \tag{10}$$

Combining (9) and (10), we get

$$\mathbb{E}[\Delta_k^6(\mathbf{w})] \le 32(W_1 + 15W_2 + 15W_3 + 15\sigma^6). \tag{11}$$

Combining (8) and (11), we get

$$\gamma(\Delta_k(\mathbf{w})) \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \le \sqrt{\frac{32(W_1 + 15W_2 + 15W_3 + 15\sigma^6)}{W_1 + 3W_2 + W_3 + \sigma^6}} \le 480.$$

A.2 Example of Regression with Gaussian Features

Claim 1. *Suppose that each data point consists of a feature $\mathbf{x} \in \mathbb{R}^d$ and a label $y \in \mathbb{R}$, and the label is generated by*

$$y = \mathbf{x}^{\mathrm{T}}\mathbf{w}^* + \xi$$

with some $\mathbf{w}^* \in \mathcal{W}$. Assume that the elements of $\mathbf{x}$ are i.i.d. samples of standard Gaussian distribution, and that the noise ξ is independent of $\mathbf{x}$ and drawn from Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Define the quadratic loss function $f(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}(y - \mathbf{x}^{\mathrm{T}}\mathbf{w})^2$. Then, we have

$$\mathrm{Var}(\nabla f(\mathbf{w}; \mathbf{x}, y)) = (d+1)\|\mathbf{w} - \mathbf{w}^*\|_2^2 + d\sigma^2,$$

and

$$\|\gamma(\nabla f(\mathbf{w}; \mathbf{x}, y))\|_\infty \le 429.$$

Proof. We use the same simplified notation as in Appendix A.1. One can also see that (1) still holds for in the Gaussian setting. Thus,

$$\begin{aligned}\mathbb{E}[\Delta_k^2(\mathbf{w})] &= \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_k^2 x_i^2 (w_i - w_i^*)^2 + (x_k^2-1)^2(w_k-w_k^*)^2 + \xi^2 x_k^2] \\ &= \sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2 + 2(w_k - w_k^*)^2 + \sigma^2 \qquad (12)\\ &= \|\mathbf{w}-\mathbf{w}^*\|_2^2 + (w_k - w_k^*)^2 + \sigma^2, \qquad (13)\end{aligned}$$

which yields

$$\mathrm{Var}(\nabla f(\mathbf{w})) = \mathbb{E}[\|\nabla f(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2] = (d+1)\|\mathbf{w}-\mathbf{w}^*\|_2^2 + d\sigma^2.$$

Then we proceed to bound $\gamma(\Delta_k(\mathbf{w}))$. By Jensen's inequality, we know that

$$\gamma(\Delta_k(\mathbf{w})) = \frac{\mathbb{E}[|\Delta_k(\mathbf{w})|^3]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^{3/2}} \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \qquad (14)$$

We first find a lower bound for $\mathrm{Var}(\Delta_k(\mathbf{w}))^3$. According to (12), we know that

$$\begin{aligned}\mathrm{Var}(\Delta_k(\mathbf{w}))^3 &= \Big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2 + 2(w_k-w_k^*)^2 + \sigma^2\Big)^3 \\ &\ge \Big(\sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2\Big)^3 + 8(w_k-w_k^*)^6 + \sigma^6.\end{aligned}$$

Define the W_1, W_2, and W_3 as in (4), (5), and (6). We can also see that (7) still holds, and thus

$$\mathrm{Var}(\Delta_k(\mathbf{w}))^3 \ge W_1 + 3W_2 + W_3 + 8(w_k - w_k^*)^6 + \sigma^6. \qquad (15)$$

Then, we find an upper bound on $\mathbb{E}[\Delta_k^6(\mathbf{w})]$. According to (1), and Hölder's inequality, we know that

$$\begin{aligned}\mathbb{E}[\Delta_k^6(\mathbf{w})] &= \mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*) + (x_k^2-1)(w_k - w_k^*) - \xi x_k)^6] \\ &\le 243\big(\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i (w_i - w_i^*))^6] + \mathbb{E}[(x_k^2-1)^6(w_k-w_k^*)^6] + \mathbb{E}[\xi^6 x_k^6]\big) \\ &= 243\big(15\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i(w_i - w_i^*))^6] + 6040(w_k-w_k^*)^6 + 225\sigma^6\big), \qquad (16)\end{aligned}$$

where in the last inequality we use the moments of Gaussian random variables. Then, we compute the first term in (16). By algebra, one can obtain

$$\begin{aligned}\mathbb{E}[(\sum_{\substack{1\le i\le d\\ i\neq k}} x_i(w_i - w_i^*))^6] =& \mathbb{E}[\sum_{\substack{1\le i\le d\\ i\neq k}} x_i^6 (w_i - w_i^*)^6] + 15\mathbb{E}[\sum_{\substack{1\le i,j\le d\\ i,j\neq k\\ i\neq j}} x_i^4 x_j^2 (w_i-w_i^*)^4 (w_j - w_j^*)^2] \\ &+ 15\mathbb{E}[\sum_{\substack{1\le i,j,\ell\le d\\ i,j,\ell\neq k\\ i\neq j, i\neq \ell, j\neq \ell}} x_i^2 x_j^2 x_\ell^2 (w_i-w_i^*)^2 (w_j-w_j^*)^2 (w_\ell - w_\ell^*)^2] \\ =& 15W_1 + 45W_2 + 15W_3. \qquad (17)\end{aligned}$$

Combining (16) and (17), we get

$$\mathbb{E}[\Delta_k^6(\mathbf{w})] \le 243(225W_1 + 675W_2 + 225W_3 + 6040(w_k - w_k^*)^6 + 225\sigma^6). \qquad (18)$$

Combining (15) and (18), we get

$$\gamma(\Delta_k(\mathbf{w})) \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(\mathbf{w})]}{\mathrm{Var}(\Delta_k(\mathbf{w}))^3}} \le \sqrt{\frac{243(225W_1 + 675W_2 + 225W_3 + 6040(w_k-w_k^*)^6 + 225\sigma^6)}{W_1 + 3W_2 + W_3 + 8(w_k-w_k^*)^6 + \sigma^6}} \le 429.$$

□

A.3 Proof of Proposition 2

We use the same notation as in Appendix A.1. We have

$$\partial_k f(\mathbf{w};\mathbf{z}) - F(\mathbf{w}) = \Delta_k(\mathbf{w}) = \sum_{\substack{1\le i\le d\\ i\neq k}} x_k x_i(w_i - w_i^*) + (x_k^2 - 1)(w_k - w_k^*) - \xi x_k$$
$$= x_k(-\xi + \sum_{\substack{1\le i\le d\\ i\neq k}} x_i(w_i - w_i^*)) := x_k \Delta_k'(\mathbf{w})$$

Since $\Delta_k'(\mathbf{w})$ has symmetric distribution and x_k is uniformly distributed in $\{-1, 1\}$, we know that the distributions of $\Delta_k(\mathbf{w})$ and $\Delta_k'(\mathbf{w})$. We then prove a stronger result on $\Delta_k'(\mathbf{w})$. We first recall the definition of v-sub-Gaussian random variables. A random variable X with mean $\mu = \mathbb{E}[X]$ is v-sub-Gaussian if for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{v^2\lambda^2/2}$. We can see that v-sub-Gaussian random variables are also v-sub-exponential. One can also check that x_i's are i.i.d. 1-sub-Gaussian random variables, and then $\Delta_k'(\mathbf{w})$ is v-sub-exponential with

$$v = \big(\sigma^2 + \sum_{\substack{1\le i\le d\\ i\neq k}} (w_i - w_i^*)^2\big)^{1/2} \le \sqrt{\sigma^2 + \|\mathbf{w} - \mathbf{w}^*\|_2^2}.$$

B Proof of Theorem 1

The proof of Theorem 1 consists of two parts: 1) the analysis of coordinate-wise median estimator of the population gradients, and 2) the convergence analysis of the robustified gradient descent algorithm.

Recall that at iteration t, the master machine sends $\mathbf{w}^t$ to all the worker machines. For any normal worker machine, say machine $i \in [m] \setminus \mathcal{B}$, the gradient of the local empirical loss function $\mathbf{g}^i(\mathbf{w}^t) = \nabla F_i(\mathbf{w}^t)$ is computed and returned to the center machine, while the Byzantine machines, say machine $i \in \mathcal{B}$, the returned message $\mathbf{g}^i(\mathbf{w}^t)$ can be arbitrary or even adversarial. The master machine then compute the coordinate-wise median, i.e.,

$$\mathbf{g}(\mathbf{w}^t) = \mathsf{med}\{\mathbf{g}^i(\mathbf{w}^t) : i \in [m]\}.$$

The following theorem provides a uniform bound on the distance between $\mathbf{g}(\mathbf{w}^t)$ and $\nabla F(\mathbf{w}^t)$.

Claim 2. *Define*

$$\mathbf{g}^i(\mathbf{w}) = \begin{cases} \nabla F_i(\mathbf{w}) & i \in [m] \setminus \mathcal{B}, \\ * & i \in \mathcal{B}. \end{cases} \tag{19}$$

and the coordinate-wise median of $\mathbf{g}^i(\mathbf{w})$*:*

$$\mathbf{g}(\mathbf{w}) = \mathsf{med}\{\mathbf{g}^i(\mathbf{w}) : i \in [m]\}. \tag{20}$$

Suppose that Assumption 2, 3, and 1 hold, and inequality (2) is satisfied with some $\epsilon > 0$. Then, we have with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, we have

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \le 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha + \sqrt{\frac{d\log(1+nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right), \tag{21}$$

for all $\mathbf{w} \in \mathcal{W}$, where C_ϵ is defined as in (4) in the main paper.

Proof. See Appendix B.1. □

Then, we proceed to analyze the convergence of the robust distributed gradient descent algorithm. We condition on the event that the bound in (21) is satisfied for all $\mathbf{w} \in \mathcal{W}$. Then, in the t-th iteration, we define

$$\widehat{\mathbf{w}}^{t+1} = \mathbf{w}^t - \eta \mathbf{g}(\mathbf{w}^t).$$

Thus, we have $\mathbf{w}^{t+1} = \Pi_{\mathcal{W}}(\widehat{\mathbf{w}}^{t+1})$. By the property of Euclidean projection, we know that

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\widehat{\mathbf{w}}^{t+1} - \mathbf{w}^*\|_2.$$

We further have

$$\begin{aligned} \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 &\leq \|\mathbf{w}^t - \eta \mathbf{g}(\mathbf{w}^t) - \mathbf{w}^*\|_2 \\ &\leq \|\mathbf{w}^t - \eta \nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 + \eta \|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2. \end{aligned} \tag{22}$$

Meanwhile, we have

$$\|\mathbf{w}^t - \eta \nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 = \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta \langle \mathbf{w}^t - \mathbf{w}^*, \nabla F(\mathbf{w}^t) \rangle + \eta^2 \|\nabla F(\mathbf{w}^t)\|_2^2. \tag{23}$$

Since $F(\mathbf{w})$ is λ_F-strongly convex, by the co-coercivity of strongly convex functions (see Lemma 3.11 in [2] for more details), we obtain

$$\langle \mathbf{w}^t - \mathbf{w}^*, \nabla F(\mathbf{w}^t) \rangle \geq \frac{L_F \lambda_F}{L_F + \lambda_F} \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \frac{1}{L_F + \lambda_F} \|\nabla F(\mathbf{w}^t)\|_2^2.$$

Let $\eta = \frac{1}{L_F}$. Then we get

$$\begin{aligned} \|\mathbf{w}^t - \eta \nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 &\leq (1 - \frac{2\lambda_F}{L_F + \lambda_F}) \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \frac{2}{L_F(L_F + \lambda_F)} \|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{L_F^2} \|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq (1 - \frac{2\lambda_F}{L_F + \lambda_F}) \|\mathbf{w}^t - \mathbf{w}^*\|_2^2, \end{aligned}$$

where in the second inequality we use the fact that $\lambda_F \leq L_F$. Using the fact $\sqrt{1-x} \leq 1 - \frac{x}{2}$, we get

$$\|\mathbf{w}^t - \eta \nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F}) \|\mathbf{w}^t - \mathbf{w}^*\|_2. \tag{24}$$

Combining (22) and (24), we get

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F}) \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{1}{L_F} \Delta, \tag{25}$$

where

$$\Delta = 2\sqrt{2} \frac{1}{nm} + \sqrt{2} \frac{C_\epsilon}{\sqrt{n}} V (\alpha + \sqrt{\frac{d \log(1 + nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748 \frac{S}{\sqrt{n}}).$$

Then we can complete the proof by iterating (25).

B.1 Proof of Claim 2

The proof of Claim 2 relies on careful analysis of the median of means estimator in the presence of adversarial data and a covering net argument.

We first consider a general problem of robust estimation of a one dimensional random variable. Suppose that there are m worker machines, and q of them are Byzantine machines, which store n adversarial data (recall that $\alpha := q/m$). Each of the other $m(1-\alpha)$ normal worker machines stores n i.i.d. samples of some one dimensional random variable $x \sim \mathcal{D}$. Denote the j-th sample in the i-th worker machine by $x^{i,j}$. Let $\mu := \mathbb{E}[x]$, $\sigma^2 := \mathrm{Var}(x)$, and $\gamma(x)$ be the absolute skewness of x. In addition, define $\bar{x}^i$ as the average of samples in the i-th machine, i.e., $\bar{x}^i = \frac{1}{n} \sum_{j=1}^n x^{i,j}$. For any $z \in \mathbb{R}$, define $\widetilde{p}(z) := \frac{1}{m(1-\alpha)} \sum_{i \in [m] \setminus \mathcal{B}} \mathbb{1}(\bar{x}^i \leq z)$ as the empirical distribution function of the sample averages on the *normal* worker machines. We have the following result on $\widetilde{p}(z)$.

Lemma 1. *Suppose that for a fixed $t > 0$, we have*

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma(x)}{\sqrt{n}} \leq \frac{1}{2} - \epsilon, \tag{26}$$

for some $\epsilon > 0$. Then, with probability at least $1 - 4e^{-2t}$, we have

$$\widetilde{p}\left(\mu + C_\epsilon \frac{\sigma}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma(x)}{\sqrt{n}}) \right) \geq \frac{1}{2} + \alpha, \tag{27}$$

and

$$\widetilde{p}\left(\mu - C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \leq \frac{1}{2} - \alpha, \tag{28}$$

where C_ϵ is defined as in (4) in the main paper.

Proof. See Appendix B.2. □

We further define the distribution function of all the m machines as $\widehat{p}(z) := \frac{1}{m}\sum_{i\in[m]} \mathbb{1}(\bar{x}^i \leq z)$. We have the following direct corollary on $\widehat{p}(z)$ and the median of means estimator $\mathsf{med}\{\bar{x}^i : i \in [m]\}$.

Corollary 1. *Suppose that condition (26) is satisfied. Then, with probability at least $1 - 4e^{-2t}$, we have,*

$$\widehat{p}\left(\mu + C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \geq \frac{1}{2}, \tag{29}$$

and

$$\widehat{p}\left(\mu - C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})\right) \leq \frac{1}{2}. \tag{30}$$

Thus, we have with probability at least $1 - 4e^{-2t}$,

$$|\mathsf{med}\{\bar{x}^i : i \in [m]\} - \mu| \leq C_\epsilon \frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}}). \tag{31}$$

Proof. One can easily check that for any $z \in \mathbb{R}$, we have $|\widehat{p}(z) - \widetilde{p}(z)| \leq \alpha$, which yields the results (29) and (30). The result (31) can be derived using the fact that $\widehat{p}(\mathsf{med}\{\bar{x}^i : i \in [m]\}) = 1/2$. □

Lemma 1 and Corollary 1 can be translated to the estimators of the gradients. Define $\mathbf{g}^i(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$ as in (19) and (20), and let $g_k^i(\mathbf{w})$ and $g_k(\mathbf{w})$ be the k-th coordinate of $\mathbf{g}^i(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$, respectively. In addition, for any $\mathbf{w} \in \mathcal{W}$, $k \in [d]$, and $z \in \mathbb{R}$, we define the empirical distribution function of the k-th coordinate of the gradients on the normal machines:

$$\widetilde{p}(z; \mathbf{w}, k) = \frac{1}{m(1-\alpha)} \sum_{i\in[m]\setminus\mathcal{B}} \mathbb{1}(g_k^i(\mathbf{w}) \leq z), \tag{32}$$

and on all the m machines

$$\widehat{p}(z; \mathbf{w}, k) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(g_k^i(\mathbf{w}) \leq z). \tag{33}$$

We use the symbol ∂_k to denote the partial derivative of any function with respect to its k-th argument. We also use the simplified notation $\sigma_k^2(\mathbf{w}) := \mathrm{Var}(\partial_k f(\mathbf{w}; \mathbf{z}))$, and $\gamma_k(\mathbf{w}) := \gamma(\partial_k f(\mathbf{w}; \mathbf{z}))$. Then, according to Lemma 1, when (26) is satisfied, for any fixed $\mathbf{w} \in \mathcal{W}$ and $k \in [d]$, we have with probability at least $1 - 4e^{-2t}$,

$$\widetilde{p}\left(\partial_k F(\mathbf{w}) + C_\epsilon \frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}); \mathbf{w}, k\right) \geq \frac{1}{2} + \alpha, \tag{34}$$

and

$$\widetilde{p}\left(\partial_k F(\mathbf{w}) - C_\epsilon \frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}); \mathbf{w}, k\right) \leq \frac{1}{2} - \alpha. \tag{35}$$

Further, according to Corollary 1, we know that with probability $1 - 4e^{-2t}$,

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq C_\epsilon \frac{\sigma_k(\mathbf{w})}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}). \tag{36}$$

Here, the inequality (36) gives a bound on the accuracy of the median of means estimator for the gradient at any fixed $\mathbf{w}$ and any coordinate $k \in [d]$. To extend this result to all $\mathbf{w} \in \mathcal{W}$ and all the d coordinates, we need to use union bound and a covering net argument.

Let $\mathcal{W}_\delta = \{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{N_\delta}\}$ be a finite subset of $\mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$, there exists $\mathbf{w}^\ell \in \mathcal{W}_\delta$ such that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \leq \delta$. According to the standard covering net results [8], we know that $N_\delta \leq (1 + \frac{D}{\delta})^d$. By union bound, we know that with probability at least $1 - 4dN_\delta e^{-2t}$, the bounds in (34) and (35) hold for all $\mathbf{w} = \mathbf{w}^\ell \in \mathcal{W}_\delta$, and $k \in [d]$. By gathering all the k coordinates and using Assumption 3, we know that this implies for all $\mathbf{w}^\ell \in \mathcal{W}_\delta$,

$$\|\mathbf{g}(\mathbf{w}^\ell) - \nabla F(\mathbf{w}^\ell)\|_2 \leq \frac{C_\epsilon}{\sqrt{n}} V \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{S}{\sqrt{n}} \right). \tag{37}$$

Then, consider an arbitrary $\mathbf{w} \in \mathcal{W}$. Suppose that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \leq \delta$. Since by Assumption 1, we assume that for each $k \in [d]$, the partial derivative $\partial_k f(\mathbf{w}; \mathbf{z})$ is L_k-Lipschitz for all $\mathbf{z}$, we know that for every normal machine $i \in [m] \setminus \mathcal{B}$,

$$\left| g_k^i(\mathbf{w}) - g_k^i(\mathbf{w}^\ell) \right| \leq L_k \delta.$$

Then, according to the definition of $\widetilde{p}(z; \mathbf{w}, k)$ in (33), we know that for any $z \in \mathbb{R}$, $\widetilde{p}(z + L_k\delta; \mathbf{w}, k) \geq \widetilde{p}(z; \mathbf{w}^\ell, k)$ and $\widetilde{p}(z - L_k\delta; \mathbf{w}, k) \leq \widetilde{p}(z; \mathbf{w}^\ell, k)$. Then, the bounds in (34) and (35) yield

$$\widetilde{p}\left(\partial_k F(\mathbf{w}^\ell) + L_k\delta + C_\epsilon \frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}}); \mathbf{w}, k \right) \geq \frac{1}{2} + \alpha, \tag{38}$$

and

$$\widetilde{p}\left(\partial_k F(\mathbf{w}^\ell) - L_k\delta - C_\epsilon \frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}}); \mathbf{w}, k \right) \leq \frac{1}{2} - \alpha. \tag{39}$$

Using the fact that $|\partial_k F(\mathbf{w}^\ell) - \partial_k F(\mathbf{w})| \leq L_k\delta$, and Corollary 1, we have

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq 2L_k\delta + C_\epsilon \frac{\sigma_k(\mathbf{w}^\ell)}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}}).$$

Again, by gathering all the k coordinates we get

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2 \leq 8\delta^2 \sum_{k=1}^d L_k^2 + 2\frac{C_\epsilon^2}{n} \sum_{k=1}^d \sigma_k^2(\mathbf{w}^\ell) (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma_k(\mathbf{w}^\ell)}{\sqrt{n}})^2,$$

where we use the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Then, by Assumption 2 and 3, we further obtain

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\delta\widehat{L} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}} V \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{S}{\sqrt{n}} \right), \tag{40}$$

where we use the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Combining (37) and (40), we conclude that for any $\delta > 0$, with probability at least $1 - 4dN_\delta e^{-2t}$, (40) holds for all $\mathbf{w} \in \mathcal{W}$. We simply choose $\delta = \frac{1}{nm\widehat{L}}$, and $t = d\log(1 + nm\widehat{L}D)$. Then, we know that with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$, we have

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}} V \left(\alpha + \sqrt{\frac{d\log(1 + nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748 \frac{S}{\sqrt{n}} \right)$$

for all $\mathbf{w} \in \mathcal{W}$.

B.2 Proof of Lemma 1

We recall the Berry-Esseen Theorem [1, 4, 7] and the bounded difference inequality, which are useful in this proof.

Claim 3 (Berry-Esseen Theorem)**.** *Assume that $Y_1, \ldots, Y_n$ are i.i.d. copies of a random variable Y with mean μ, variance σ^2, and such that $\mathbb{E}[|Y-\mu|^3] < \infty$. Then,*

$$\sup_{s\in\mathbb{R}} \left| \mathbb{P}\left\{ \sqrt{n}\frac{\bar{Y}-\mu}{\sigma} \le s \right\} - \Phi(s) \right| \le 0.4748 \frac{\mathbb{E}[|Y-\mu|^3]}{\sigma^3\sqrt{n}},$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\Phi(s)$ is the cumulative distribution function of the standard normal random variable.

Claim 4 (Bounded Difference Inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables, and assume that $Z = g(X_1, \ldots, X_n)$, where g satisfies that for all $j \in [n]$ and all $x_1, x_2, \ldots, x_j, x_j', \ldots, x_n$,*

$$|g(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_n) - g(x_1, \ldots, x_{j-1}, x_j', x_{j+1}, \ldots, x_n)| \le c_j.$$

Then for any $t \ge 0$,

$$\mathbb{P}\left\{ Z - \mathbb{E}[Z] \ge t \right\} \le \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2} \right)$$

and

$$\mathbb{P}\left\{ Z - \mathbb{E}[Z] \le -t \right\} \le \exp\left(-\frac{2t^2}{\sum_{j=1}^n c_j^2} \right).$$

Let $\sigma_n := \frac{\sigma}{\sqrt{n}}$ and $c_n := 0.4748\frac{\mathbb{E}[|x-\mu|^3]}{\sigma^3\sqrt{n}} = 0.4748\frac{\gamma(x)}{\sqrt{n}}$. Define $W_i := \frac{\bar{x}^i - \mu}{\sigma_n}$ for all $i \in [m]$, and $\Phi_n(\cdot)$ be the distribution function of W_i for any $i \in [m] \setminus \mathcal{B}$. We also define the empirical distribution function of $\{W_i : i \in [m] \setminus \mathcal{B}\}$ as $\widetilde{\Phi}_n(\cdot)$, i.e., $\widetilde{\Phi}_n(z) = \frac{1}{m(1-\alpha)}\sum_{i\in[m]\setminus\mathcal{B}} \mathbb{1}(W_i \le z)$. Thus, we have

$$\widetilde{\Phi}_n(z) = \widetilde{p}(\sigma_n z + \mu). \tag{41}$$

We then focus on $\widetilde{\Phi}_n(z)$. We know that for any $z \in \mathbb{R}$, $\mathbb{E}[\widetilde{\Phi}_n(z)] = \Phi_n(z)$. Then, since the bounded difference inequality is satisfied with $c_j = \frac{1}{m(1-\alpha)}$, we have for any $t > 0$,

$$\left| \widetilde{\Phi}_n(z) - \Phi_n(z) \right| \le \sqrt{\frac{t}{m(1-\alpha)}}, \tag{42}$$

on the draw of W_i, $i \in [m] \setminus \mathcal{B}$ with probability at least $1 - 2e^{-2t}$. Let $z_1 \ge z_2$ be such that $\Phi_n(z_1) \ge \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}}$, and $\Phi_n(z_2) \le \frac{1}{2} - \alpha - \sqrt{\frac{t}{m(1-\alpha)}}$. Then, by union bound, we know that with probability at least $1 - 4e^{-2t}$, $\widetilde{\Phi}_n(z_1) \ge 1/2 + \alpha$ and $\widetilde{\Phi}_n(z_2) \le 1/2 - \alpha$. The next step is to choose z_1 and z_2. According to Claim 3, we know that

$$\Phi_n(z_1) \ge \Phi(z_1) - c_n,$$

and thus, it suffices to find z_1 such that

$$\Phi(z_1) = \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n.$$

By mean value theorem, we know that there exists $\xi \in [0, z_1]$ such that

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n = z_1\Phi'(\xi) = \frac{z_1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} \ge \frac{z_1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}}$$

Suppose that for some fix constant $\epsilon \in (0, 1/2)$, we have

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \le \frac{1}{2} - \epsilon.$$

Then, we know that $z_1 \le \Phi^{-1}(1-\epsilon)$, and thus we have

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \ge \frac{z_1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2),$$

which yields

$$z_1 \leq \sqrt{2\pi}\exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right).$$

Similarly

$$z_2 \geq -\sqrt{2\pi}\exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right).$$

For simplicity, let $C_\epsilon := \sqrt{2\pi}\exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)$. We conclude that with probability $1-4e^{-2t}$, we have

$$\widetilde{p}(\mu + C_\epsilon\sigma_n(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)) \geq \frac{1}{2} + \alpha,$$

and

$$\widetilde{p}(\mu - C_\epsilon\sigma_n(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)) \leq \frac{1}{2} - \alpha.$$

C Proof of Theorem 2

Since Claim 2 holds without assuming the convexity of $F(\mathbf{w})$, when $F(\mathbf{w})$ is non-strongly convex, the event that (21) holds for all $\mathbf{w} \in \mathcal{W}$ still happens with probability at least $1 - \frac{4d}{(1+nm\widehat{L}D)^d}$. We condition on this event. We first show that when Assumption 4 is satisfied and we choose $\eta = \frac{1}{L_F}$, the iterates $\mathbf{w}^t$ stays in $\mathcal{W}$ without using projection. Namely, define

$$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\mathbf{g}(\mathbf{w}^t),$$

for $T = 0, 1, \ldots, T-1$, then $\mathbf{w}^t \in \mathcal{W}$ for all $t = 0, 1, \ldots, T$. To see this, we have

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2 + \eta\|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2,$$

and

$$\begin{aligned}
\|\mathbf{w}^t - \eta\nabla F(\mathbf{w}^t) - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta\langle\nabla F(\mathbf{w}^t), \mathbf{w}^t - \mathbf{w}^*\rangle + \eta^2\|\nabla F(\mathbf{w}^t)\|_2^2 \\
&\leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - 2\eta\frac{1}{L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \eta^2\|\nabla F(\mathbf{w}^t)\|_2^2 \\
&= \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 - \frac{1}{L_F^2}\|\nabla F(\mathbf{w}^t)\|_2^2 \\
&\leq \|\mathbf{w}^t - \mathbf{w}^*\|_2^2
\end{aligned}$$

where the inequality is due to the co-coercivity of convex functions. Thus, we get

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{\Delta}{L_F},$$

and since $T = \frac{L_F D_0}{\Delta}$, according to Assumption 4 we know that $\mathbf{w}^t \in \mathcal{W}$ for all $t = 0, 1, \ldots, T$. Then, we proceed to study the algorithm without projection. Here, we define $D_t := \|\mathbf{w}^0 - \mathbf{w}^*\|_2 + \frac{t\Delta}{L_F}$ for $t = 0, 1, \ldots, T$.

Using the smoothness of $F(\mathbf{w})$, we have

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

Since $\eta = \frac{1}{L_F}$ and $\|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2 \leq \Delta$, by simple algebra, we obtain

$$F(\mathbf{w}^{t+1}) \leq F(\mathbf{w}^t) - \frac{1}{2L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{2L_F}\Delta^2. \tag{43}$$

We now prove the following lemma.

Lemma 2. *Condition on the event that (21) holds for all $\mathbf{w} \in \mathcal{W}$. When $F(\mathbf{w})$ is convex, by running $T = \frac{L_F D_0}{\Delta}$ parallel iterations, there exists $t \in \{0, 1, 2, \ldots, T\}$ such that*

$$F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16 D_0 \Delta.$$

Proof. We first notice that since $T = \frac{L_F D_0}{\Delta}$, we have $D_t \leq 2D_0$ for all $t = 0, 1, \ldots, T$. According to the first order optimality of convex functions, for any $\mathbf{w}$,

$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^* \rangle \leq \|\nabla F(\mathbf{w})\|_2 \|\mathbf{w} - \mathbf{w}^*\|_2,$$

and thus

$$\|\nabla F(\mathbf{w})\|_2 \geq \frac{F(\mathbf{w}) - F(\mathbf{w}^*)}{\|\mathbf{w} - \mathbf{w}^*\|_2}. \tag{44}$$

Suppose that there exists $t \in \{0, 1, \ldots, T-1\}$ such that $\|\nabla F(\mathbf{w}^t)\|_2 < \sqrt{2}\Delta$. Then we have

$$F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq \|\nabla F(\mathbf{w}^t)\|_2 \|\mathbf{w}^t - \mathbf{w}^*\|_2 \leq 2\sqrt{2} D_0 \Delta.$$

Otherwise, for all $t \in \{0, 1, \ldots, T-1\}$, $\|\nabla F(\mathbf{w}^t)\|_2 \geq \sqrt{2}\Delta$. Then, according to (43) and (44), we have for all $t < T$,

$$\begin{aligned} F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*) &\leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{4L_F} \|\nabla F(\mathbf{w}^t)\|_2^2 \\ &\leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{4L_F D_t^2} (F(\mathbf{w}^t) - F(\mathbf{w}^*))^2. \end{aligned}$$

Multiplying both sides by $[(F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*))(F(\mathbf{w}^t) - F(\mathbf{w}^*)]^{-1}$ and rearranging the terms, we obtain

$$\frac{1}{F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^t) - F(\mathbf{w}^*)} + \frac{1}{4L_F D_t^2} \frac{F(\mathbf{w}^t) - F(\mathbf{w}^*)}{F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^t) - F(\mathbf{w}^*)} + \frac{1}{16 L_F D_0^2},$$

which implies

$$\frac{1}{F(\mathbf{w}^T) - F(\mathbf{w}^*)} \geq \frac{1}{F(\mathbf{w}^0) - F(\mathbf{w}^*)} + \frac{T}{16 L_F D_0^2} \geq \frac{T}{16 L_F D_0^2}.$$

Then, we obtain $F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq 16 D_0 \Delta$ using the fact that $T = \frac{L_F D_0}{\Delta}$. □

Next, we show that $F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq 16 D_0 \Delta + \frac{1}{2L_F}\Delta^2$. More specifically, let $t = t_0$ be the first time that $F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16 D_0 \Delta$, and we show that for any $t > t_0$, $F(\mathbf{w}^t) - F(\mathbf{w}^*) \leq 16 D_0 \Delta + \frac{1}{2L_F}\Delta^2$. If this statement is not true, then we let $t_1 > t_0$ be the first time that $F(\mathbf{w}^t) - F(\mathbf{w}^*) > 16 D_0 \Delta + \frac{1}{2L_F}\Delta^2$. Then there must be $F(\mathbf{w}^{t_1 - 1}) < F(\mathbf{w}^{t_1})$. According to (43), there should also be

$$F(\mathbf{w}^{t_1 - 1}) - F(\mathbf{w}^*) \geq F(\mathbf{w}^{t_1}) - F(\mathbf{w}^*) - \frac{1}{2L_F}\Delta^2 > 16 D_0 \Delta.$$

Then, according to (44), we have

$$\|\nabla F(\mathbf{w}^{t_1 - 1})\|_2 \geq \frac{F(\mathbf{w}^{t_1 - 1}) - F(\mathbf{w}^*)}{\|\mathbf{w}^{t_1 - 1} - \mathbf{w}^*\|_2} > 8\Delta.$$

Then according to (43), this implies $F(\mathbf{w}^{t_1}) \leq F(\mathbf{w}^{t_1 - 1})$, which contradicts with the fact that $F(\mathbf{w}^{t_1 - 1}) < F(\mathbf{w}^{t_1})$.

D Proof of Theorem 3

Since Claim 2 holds without assuming the convexity of $F(\mathbf{w})$, when $F(\mathbf{w})$ is non-convex, the event that (21) holds for all $\mathbf{w} \in \mathcal{W}$ still happens with probability at least $1 - \frac{4d}{(1 + nm\widehat{L}D)^d}$. We condition on this event. We first show that when Assumption 5 is satisfied and we choose $\eta = \frac{1}{L_F}$, the iterates $\mathbf{w}^t$ stays in $\mathcal{W}$ without using projection. Since we have

$$\|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2 \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \eta(\|\nabla F(\mathbf{w}^t)\|_2 + \|\mathbf{g}(\mathbf{w}^t) - \nabla F(\mathbf{w}^t)\|_2) \leq \|\mathbf{w}^t - \mathbf{w}^*\|_2 + \frac{1}{L_F}(M + \Delta).$$

Then, we know that by running $T = \frac{2L_F}{\Delta^2}(F(\mathbf{w}^0) - F(\mathbf{w}^*))$ parallel iterations, using Assumption 5, we know that $\mathbf{w}^t \in \mathcal{W}$ for $t = 0, 1, \ldots, T$ without projection.

We proceed to study the convergence rate of the algorithm. By the smoothness of $F(\mathbf{w})$, we know that when choosing $\eta = \frac{1}{L_F}$, the inequality (43) still holds. More specifically, for all $t = 0, 1, \ldots, T-1$,

$$F(\mathbf{w}^{t+1}) - F(\mathbf{w}^*) \leq F(\mathbf{w}^t) - F(\mathbf{w}^*) - \frac{1}{2L_F}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{1}{2L_F}\Delta^2. \tag{45}$$

Sum up (45) for $t = 0, 1, \ldots, T-1$. Then, we get

$$0 \leq F(\mathbf{w}^T) - F(\mathbf{w}^*) \leq F(\mathbf{w}^0) - F(\mathbf{w}^*) - \frac{1}{2L_F}\sum_{t=0}^{T-1}\|\nabla F(\mathbf{w}^t)\|_2^2 + \frac{T}{2L_F}\Delta^2.$$

This implies that

$$\min_{t=0,1,\ldots,T}\|\nabla F(\mathbf{w}^t)\|_2^2 \leq 2\frac{L_F}{T}(F(\mathbf{w}^0) - F(\mathbf{w}^*)) + \Delta^2,$$

which completes the proof.

E Proof of Theorem 4

The proof of Theorem 4 consists of two parts: 1) the analysis of coordinate-wise trimmed mean of means estimator of the population gradients, and 2) the convergence analysis of the robustified gradient descent algorithm. Since the second part is essentially the same as the proof of Theorem 1, we mainly focus on the first part here.

Claim 5. *Define*

$$\mathbf{g}^i(\mathbf{w}) = \begin{cases} \nabla F_i(\mathbf{w}) & i \in [m] \setminus \mathcal{B}, \\ * & i \in \mathcal{B}. \end{cases} \tag{46}$$

and the coordinate-wise trimmed mean of $\mathbf{g}^i(\mathbf{w})$*:*

$$\mathbf{g}(\mathbf{w}) = \mathsf{trmean}_\beta\{\mathbf{g}^i(\mathbf{w}) : i \in [m]\}. \tag{47}$$

Suppose that Assumptions 1 and 6 are satisfied, and that $\alpha \leq \beta \leq \frac{1}{2} - \epsilon$*. Then, with probability at least* $1 - \frac{2d(m+1)}{(1+nm\widehat{L}D)^d}$*,*

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq \frac{v}{\epsilon}\left(\frac{3\sqrt{2}\beta d}{\sqrt{n}} + \frac{2d}{\sqrt{nm}}\right)\sqrt{\log(1 + nm\widehat{L}D) + \frac{1}{d}\log m} + \widetilde{\mathcal{O}}(\frac{\beta}{n} + \frac{1}{nm})$$

for all $\mathbf{w} \in \mathcal{W}$*.*

Proof. See Appendix E.1 □

The rest of the proof is essentially the same as the proof of Theorem 1. In fact, we essentially analyze a gradient descent algorithm with bounded noise in the gradients. In the proof of Theorem 1 in Appendix B. The bound on the noise in the gradients is

$$\Delta = \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V(\alpha + \sqrt{\frac{d\log(1 + nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}) + 2\sqrt{2}\frac{1}{nm},$$

while here we replace Δ with Δ':

$$\Delta' := \frac{v}{\epsilon}\left(\frac{3\sqrt{2}\beta d}{\sqrt{n}} + \frac{2d}{\sqrt{nm}}\right)\sqrt{\log(1 + nm\widehat{L}D) + \frac{1}{d}\log m} + \widetilde{\mathcal{O}}(\frac{\beta}{n} + \frac{1}{nm}),$$

and the same analysis can still go through. Therefore, we omit the details of the analysis here.

Remark 1. *The same arguments still go through when the population risk function* $F(\mathbf{w})$ *is non-strongly convex or non-convex. One can simply replace the bound on the noise in the gradients* Δ *in Theorems 2 and 3 with* Δ' *here. Thus we omit the details here.*

E.1 Proof of Claim 5

The proof of Claim 5 relies on the analysis of the trimmed mean of means estimator in the presence of adversarial data and a covering net argument. We first consider a general problem of robust estimation of a one dimensional random variable. Suppose that there are m worker machines, and q of them are Byzantine machines, which store n adversarial data (recall that $\alpha := q/m$). Each of the other $m(1-\alpha)$ normal worker machines stores n i.i.d. samples of some one dimensional random variable $x \sim \mathcal{D}$. Suppose that x is v-sub-exponential and let $\mu := \mathbb{E}[x]$. Denote the j-th sample in the i-th worker machine by $x^{i,j}$. In addition, define $\bar{x}^i$ as the average of samples in the i-th machine, i.e., $\bar{x}^i = \frac{1}{n}\sum_{j=1}^{n} x^{i,j}$. We have the following result on the trimmed mean of $\bar{x}^i$, $i \in [m]$.

Lemma 3. *Suppose that the one dimensional samples on all the normal machines are i.i.d. v-sub-exponential with mean μ. Then, we have for any $t \geq 0$,*

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}} \bar{x}^i - \mu| \geq t\} \leq 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\},$$

and for any $s \geq 0$,

$$\mathbb{P}\{\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \geq s\} \leq 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\},$$

and when $\beta \geq \alpha$, $|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}} \bar{x}^i - \mu| \leq t$, and $\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \leq s$, we have

$$|\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} - \mu| \leq \frac{t + 3\beta s}{1-2\beta}.$$

Proof. See Appendix E.2. □

Lemma 3 can be directly applied to the k-th partial derivative of the loss functions. Since we assume that for any $k \in [d]$ and $\mathbf{w} \in \mathcal{W}$, $\partial_k f(\mathbf{w}; \mathbf{z})$ is v-sub-exponential, we have for any $t \geq 0$, $s \geq 0$,

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}} g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})| \geq t\} \leq 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\}, \tag{48}$$

$$\mathbb{P}\{\max_{i\in[m]\setminus\mathcal{B}}\{|g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})|\} \geq s\} \leq 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}, \tag{49}$$

and consequently with probability at least

$$1 - 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\} - 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\},$$

we have

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| = |\mathsf{trmean}_\beta\{g_k^i(\mathbf{w}) : i \in [m]\} - \partial_k F(\mathbf{w})| \leq \frac{t + 3\beta s}{1-2\beta}. \tag{50}$$

To extend this result to all $\mathbf{w} \in \mathcal{W}$ and all the d coordinates, we need to use union bound and a covering net argument. Let $\mathcal{W}_\delta = \{\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^{N_\delta}\}$ be a finite subset of $\mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$, there exists $\mathbf{w}^\ell \in \mathcal{W}_\delta$ such that $\|\mathbf{w}^\ell - \mathbf{w}\|_2 \leq \delta$. According to the standard covering net results [8], we know that $N_\delta \leq (1 + \frac{D}{\delta})^d$. By union bound, we know that with probability at least

$$1 - 2dN_\delta\exp\{-(1-\alpha)mn\min\{\frac{t}{2v}, \frac{t^2}{2v^2}\}\},$$

the bound $|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}} g_k^i(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq t$ holds for all $\mathbf{w} = \mathbf{w}^\ell \in \mathcal{W}_\delta$, and $k \in [d]$, and with probability at least

$$1 - 2(1-\alpha)dmN_\delta\exp\{-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\}$$

the bound $\max_{i\in[m]\setminus\mathcal{B}}\{|g_k^i(\mathbf{w})-\partial_k F(\mathbf{w})|\}\le s$ holds for all $\mathbf{w}=\mathbf{w}^\ell\in\mathcal{W}_\delta$, and $k\in[d]$. By gathering all the k coordinates, we know that this implies for all $\mathbf{w}^\ell\in\mathcal{W}_\delta$,

$$\|\mathbf{g}(\mathbf{w}^\ell)-\nabla F(\mathbf{w}^\ell)\|_2\le\sqrt{d}\frac{t+3\beta s}{1-2\beta}. \tag{51}$$

Then, consider an arbitrary $\mathbf{w}\in\mathcal{W}$. Suppose that $\|\mathbf{w}^\ell-\mathbf{w}\|_2\le\delta$. Since by Assumption 1, we assume that for each $k\in[d]$, the partial derivative $\partial_k f(\mathbf{w};\mathbf{z})$ is L_k-Lipschitz for all $\mathbf{z}$, we know that for every normal machine $i\in[m]\setminus\mathcal{B}$,

$$\left|g_k^i(\mathbf{w})-g_k^i(\mathbf{w}^\ell)\right|\le L_k\delta,\quad \left|\partial_k F(\mathbf{w})-\partial_k F(\mathbf{w}^\ell)\right|\le L_k\delta.$$

This means that if $|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}}g_k^i(\mathbf{w}^\ell)-\partial_k F(\mathbf{w}^\ell)|\le t$ and $\max_{i\in[m]\setminus\mathcal{B}}\{|g_k^i(\mathbf{w}^\ell)-\partial_k F(\mathbf{w}^\ell)|\}\le s$ hold for all $\mathbf{w}^\ell\in\mathcal{W}_\delta$, and $k\in[d]$, then

$$|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}}g_k^i(\mathbf{w})-\partial_k F(\mathbf{w})|\le t+2L_k\delta,$$

and

$$\max_{i\in[m]\setminus\mathcal{B}}\{|g_k^i(\mathbf{w})-\partial_k F(\mathbf{w})|\}\le s+2L_k\delta$$

hold for all $\mathbf{w}\in\mathcal{W}$. This implies that for all $\mathbf{w}\in\mathcal{W}$ and $k\in[d]$,

$$|g_k(\mathbf{w})-\partial_k F(\mathbf{w})|=|\mathsf{trmean}_\beta\{g_k^i(\mathbf{w}):i\in[m]\}-\partial_k F(\mathbf{w})|\le\frac{t+3\beta s}{1-2\beta}+\frac{2(1+3\beta)}{1-2\beta}\delta L_k,$$

which yields

$$\|\mathbf{g}(\mathbf{w})-\nabla F(\mathbf{w})\|_2\le\sqrt{2d}\frac{t+3\beta s}{1-2\beta}+\sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\delta\widehat{L}.$$

The proof is completed by choosing $\delta=\frac{1}{nm\widehat{L}}$,

$$t=v\max\{\frac{8d}{nm}\log(1+nm\widehat{L}D),\sqrt{\frac{8d}{nm}\log(1+nm\widehat{L}D)}\},$$
$$s=v\max\{\frac{4}{n}(d\log(1+nm\widehat{L}D)+\log m),\sqrt{\frac{4}{n}(d\log(1+nm\widehat{L}D)+\log m)}\},$$

and using the fact that $\beta\le\frac{1}{2}-\epsilon$.

E.2 Proof of Lemma 3

We first recall Bernstein's inequality for sub-exponential random variables.

Claim 6 (Bernstein's inequality)**.** *Suppose that $X_1,X_2,\ldots,X_n$ are i.i.d. v-sub-exponential random variables with mean μ. Then for any $t\ge0$,*

$$\mathbb{P}\{|\frac{1}{n}\sum_{i=1}^n X_i-\mu|\ge t\}\le 2\exp\{-n\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\}.$$

Thus, for any $t\ge0$

$$\mathbb{P}\{|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\setminus\mathcal{B}}\bar{x}^i-\mu|\ge t\}\le 2\exp\{-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\}. \tag{52}$$

Similarly, for any $i\in[m]\setminus\mathcal{B}$, and any $s\ge0$

$$\mathbb{P}\{|\bar{x}^i-\mu|\ge s\}\le 2\exp\{-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\}.$$

Then, by union bound we know that

$$\mathbb{P}\{\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i-\mu|\}\ge s\}\le 2(1-\alpha)m\exp\{-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\}. \tag{53}$$

We proceed to analyze the trimmed mean of means estimator. To simplify notation, we define $\mathcal{M} = [m] \setminus \mathcal{B}$ as the set of all normal worker machines, $\mathcal{U} \subseteq [m]$ as the set of all untrimmed machines, and $\mathcal{T} \subseteq [m]$ as the set of all trimmed machines. The trimmed mean of means estimator simply computes

$$\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} = \frac{1}{(1-2\beta)m} \sum_{i \in \mathcal{U}} \bar{x}^i.$$

We further have

$$\begin{aligned}
|\mathsf{trmean}_\beta\{\bar{x}^i : i \in [m]\} - \mu| =& \left| \frac{1}{(1-2\beta)m} \sum_{i \in \mathcal{U}} \bar{x}^i - \mu \right| \\
=& \frac{1}{(1-2\beta)m} \left| \sum_{i \in \mathcal{M}} (\bar{x}^i - \mu) - \sum_{i \in \mathcal{M} \cap \mathcal{T}} (\bar{x}^i - \mu) + \sum_{i \in \mathcal{B} \cap \mathcal{U}} (\bar{x}^i - \mu) \right| \\
=& \frac{1}{(1-2\beta)m} \big(|\sum_{i \in \mathcal{M}} (\bar{x}^i - \mu)| + |\sum_{i \in \mathcal{M} \cap \mathcal{T}} (\bar{x}^i - \mu)| + |\sum_{i \in \mathcal{B} \cap \mathcal{U}} (\bar{x}^i - \mu)| \big)
\end{aligned} \tag{54}$$

We also know that $|\sum_{i \in \mathcal{M} \cap \mathcal{T}} (\bar{x}^i - \mu)| \leq 2\beta m \max_{i \in \mathcal{M}} \{|\bar{x}^i - \mu|\}$. In addition, since $\beta \geq \alpha$, without loss of generality, we assume that $\mathcal{M} \cap \mathcal{T} \neq \emptyset$, and then $|\sum_{i \in \mathcal{B} \cap \mathcal{U}} (\bar{x}^i - \mu)| \leq \alpha m \max_{i \in \mathcal{M}} \{|\bar{x}^i - \mu|\}$. Then we directly obtain the desired result.

F Proof of Theorem 7

Since the loss functions are quadratic, we denote the loss function $f(\mathbf{w}; \mathbf{z}^{i,j})$ by

$$f(\mathbf{w}; \mathbf{z}^{i,j}) = \frac{1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{H}_{i,j} \mathbf{w} + \mathbf{p}_{i,j}^{\mathrm{T}} \mathbf{w} + c_{i,j}.$$

We further define $\mathbf{H}_i := \frac{1}{n} \sum_{j=1}^n \mathbf{H}_{i,j}$, $\mathbf{p}_i := \frac{1}{n} \sum_{j=1}^n \mathbf{p}_{i,j}$, and $c_i := \frac{1}{n} \sum_{j=1}^n c_{i,j}$. Thus the empirical risk function on the i-th machine is

$$F_i(\mathbf{w}) = \frac{1}{2} \mathbf{w}^{\mathrm{T}} \mathbf{H}_i \mathbf{w} + \mathbf{p}_i^{\mathrm{T}} \mathbf{w} + c_i.$$

Then, for any worker machine $i \in [m] \setminus \mathcal{B}$, $\widehat{\mathbf{w}}^i = -\mathbf{H}_i^{-1} \mathbf{p}_i$. In addition, the population risk minimizer is $\mathbf{w}^* = -\mathbf{H}_F^{-1} \mathbf{p}_F$. We further define $\mathbf{U}_{i,j} := \mathbf{H}_{i,j} - \mathbf{H}_F$, $\mathbf{U}_i = \mathbf{H}_i - \mathbf{H}_F$, $\mathbf{v}_{i,j} = \mathbf{p}_{i,j} - \mathbf{p}_F$, and $\mathbf{v}_i = \mathbf{p}_i - \mathbf{p}_F$. Then

$$\widehat{\mathbf{w}}^i = -(\mathbf{U}_i + \mathbf{H}_F)^{-1} (\mathbf{v}_i + \mathbf{p}_F).$$

Let $\mathbf{e}_k$ be the k-th vector in the standard basis, i.e., the k-th column of the $d \times d$ identity matrix. We proceed to study the distribution of the k-th coordinate of $\widehat{\mathbf{w}}^i - \mathbf{w}^*$, $i \in [m] \setminus \mathcal{B}$, i.e.,

$$\widehat{w}_k^i - w_k^* = \mathbf{e}_k^{\mathrm{T}} \mathbf{H}_F^{-1} \mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}} (\mathbf{U}_i + \mathbf{H}_F)^{-1} (\mathbf{v}_i + \mathbf{p}_F).$$

Similar to the proof of Theorem 1, we need to obtain a Berry-Esseen type bound for $\widehat{w}_k^i - w_k^*$. However, here, $\widehat{w}_k^i$ is not a sample mean of n i.i.d. random variables, and thus we cannot directly apply the vanilla Berry-Esseen bound. Instead, we apply the following bound in [6] on functions of sample means.

Claim 7 (Theorem 2.11 in [6], simplified)**.** *Let $\mathcal{X}$ be a Hilbert space equipped with norm $\|\cdot\|$. Let $f : \mathcal{X} \to \mathbb{R}$ be a function on $\mathcal{X}$. Suppose that there exists linear functions $\ell : \mathcal{X} \to \mathbb{R}$, $\theta > 0$, $M_\theta > 0$ such that*

$$|f(X) - \ell(X)| \leq \frac{M_\theta}{2} \|X\|^2, \ \forall \ \|X\| \leq \theta. \tag{55}$$

Suppose that there is a probability distribution $\mathcal{D}_X$ over $\mathcal{X}$, and let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables drawn from $\mathcal{D}_X$. Assume that $\mathbb{E}[X] = 0$, and define

$$\widetilde{\sigma} := (\mathbb{E}[\ell(X)^2])^{1/2}, \quad \nu_p := (\mathbb{E}[\|X\|^p])^{1/p}, p = 2, 3, \quad \varsigma := \frac{(\mathbb{E}[|\ell(X)|^3])^{1/3}}{\widetilde{\sigma}}.$$

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. *Then for any* $z \in \mathbb{R}$, *we have*

$$\left|\mathbb{P}\left\{\frac{f(\bar{X})}{\widetilde{\sigma}/\sqrt{n}} \le z\right\} - \Phi(z)\right| \le \frac{C}{\sqrt{n}}, \tag{56}$$

where $C = C_0 + C_1\varsigma^3 + (C_{20} + C_{21}\varsigma)\nu_2^2 + (C_{30} + C_{31}\varsigma)\nu_3^2 + C_4$, *with*

$$\begin{aligned} &C_0 = 0.1393, \quad C_1 = 2.3356 \\ &(C_{20}, C_{21}, C_{30}, C_{31}) = \frac{M_\theta}{2\widetilde{\sigma}}\left(2(\frac{2}{\pi})^{1/6}, 2 + \frac{2^{2/3}}{n^{1/6}}, \frac{(8/\pi)^{1/6}}{n^{1/3}}, \frac{2}{n^{1/2}}\right) \\ &C_4 = \min\{\frac{\nu_2^2}{\theta^2 n^{1/2}}, \frac{2\nu_2^3 + \nu_3^3/n^{1/2}}{\theta^3 n}\}. \end{aligned} \tag{57}$$

Define the function $\psi_k(\mathbf{U}, \mathbf{v}) : \mathbb{R}^{d\times d} \times \mathbb{R} \to \mathbb{R}$:

$$\psi_k(\mathbf{U}, \mathbf{v}) := \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}(\mathbf{U} + \mathbf{H}_F)^{-1}(\mathbf{v} + \mathbf{p}_F),$$

and thus

$$\widehat{w}_k^i - w_k^* = \psi_k(\mathbf{U}_i, \mathbf{v}_i) = \psi_k(\frac{1}{n}\sum_{j=1}^n \mathbf{U}_{i,j}, \frac{1}{n}\sum_{j=1}^n \mathbf{v}_{i,j}).$$

On the product space $\mathbb{R}^{d\times d} \times \mathbb{R}$, define the element-wise inner product:

$$\langle(\mathbf{U}, \mathbf{v}), (\mathbf{X}, \mathbf{y})\rangle = \sum_{i,j=1}^d U_{i,j}X_{i,j} + \sum_{i=1}^d v_i y_i,$$

and thus $\mathbb{R}^{d\times d} \times \mathbb{R}$ is associated with the norm

$$\|(\mathbf{U}, \mathbf{v})\| = \sqrt{\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2},$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrices. We then provide the following lemma on $\psi_k(\mathbf{U}, \mathbf{v})$.

Lemma 4. *There exits a linear function* $\ell_k(\mathbf{U}, \mathbf{v}) = \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}$ *such that for any* $\mathbf{U}$, $\mathbf{v}$ *with*

$$\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2 \le \frac{\lambda_F^2}{4},$$

we have

$$|\psi_k(\mathbf{U}, \mathbf{v}) - \ell_k(\mathbf{U}, \mathbf{v})| \le \frac{\lambda_F + 2\|\mathbf{p}_F\|_2}{\lambda_F^3}(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2).$$

Proof. See Appendix F.1. □

Lemma 4 tells us that the condition (55) is satisfied with $\theta = \frac{\lambda_F}{2}$ and $M_\theta = \frac{2\lambda_F + 4\|\mathbf{p}_F\|_2}{\lambda_F^3}$. For all normal worker machine $i \in [m] \setminus \mathcal{B}$, denote the distribution of $\mathbf{U}_{i,j}$ and $\mathbf{v}_{i,j}$ by $\mathcal{D}_U$ and $\mathcal{D}_v$, respectively. Since $\widehat{w}_k^i - w_k^* = \psi_k(\frac{1}{n}\sum_{j=1}^n \mathbf{U}_{i,j}, \frac{1}{n}\sum_{j=1}^n \mathbf{v}_{i,j})$, Claim 7 directly gives us the following lemma.

Lemma 5. *Let* $\mathbf{U} \sim \mathcal{D}_U$, $\mathbf{v} \sim \mathcal{D}_v$, *and* $\ell_k(\mathbf{U}, \mathbf{v}) = \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{v}$. *Define*

$$\widetilde{\sigma}_k := (\mathbb{E}[\ell_k(\mathbf{U}, \mathbf{v})^2])^{1/2}, \quad \nu_p := (\mathbb{E}[(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2)^{p/2}])^{1/p}, p = 2, 3, \quad \varsigma_k := \frac{(\mathbb{E}[|\ell_k(\mathbf{U}, \mathbf{v})|^3])^{1/3}}{\widetilde{\sigma}_k}.$$

Then for any $z \in \mathbb{R}$, $i \in [m] \setminus \mathcal{B}$, *we have*

$$\left|\mathbb{P}\left\{\frac{\widehat{w}_k^i - w_k^*}{\widetilde{\sigma}_k/\sqrt{n}} \le z\right\} - \Phi(z)\right| \le \frac{C_k}{\sqrt{n}}, \tag{58}$$

where

$$C_k = \widehat{C}_0 + \widehat{C}_1\varsigma_k^3 + \frac{1}{\widetilde{\sigma}_k}\left[(\widehat{C}_{20} + \widehat{C}_{21}\varsigma_k)\nu_2^2 + (\widehat{C}_{30} + \widehat{C}_{31}\varsigma_k)\nu_3^2\right] + \widehat{C}_4,$$

with

$$\widehat{C}_0 = 0.1393, \quad \widehat{C}_1 = 2.3356$$
$$(\widehat{C}_{20}, \widehat{C}_{21}, \widehat{C}_{30}, \widehat{C}_{31}) = \frac{\lambda_F + 2\|\mathbf{p}_F\|_2}{\lambda_F^3}\left(2(\frac{2}{\pi})^{1/6}, 2 + \frac{2^{2/3}}{n^{1/6}}, \frac{(8/\pi)^{1/6}}{n^{1/3}}, \frac{2}{n^{1/2}}\right) \tag{59}$$
$$\widehat{C}_4 = \min\{\frac{4\nu_2^2}{\lambda_F^2 n^{1/2}}, \frac{16\nu_2^3 + 8\nu_3^3/n^{1/2}}{\lambda_F^3 n}\}.$$

Then, we proceed to bound $\mathsf{med}\{\widehat{w}_k^i : i \in [m]\} - w_k^*$, the technique is similar to what we use in the proof of Claim 2. For every $z \in \mathbb{R}$, $k \in [d]$, define

$$\widetilde{p}(z;k) = \frac{1}{m(1-\alpha)} \sum_{i \in [m]\setminus\mathcal{B}} \mathbb{1}(\widehat{w}_k^i - w_k^* \le z).$$

We have the following lemma on $\widetilde{p}(z;k)$.

Lemma 6. *Suppose that for a fixed $t > 0$, we have*

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}} \le \frac{1}{2} - \epsilon, \tag{60}$$

for some $\epsilon > 0$. Then, with probability at least $1 - 4e^{-2t}$, we have

$$\widetilde{p}\left(C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \ge \frac{1}{2} + \alpha, \tag{61}$$

and

$$\widetilde{p}\left(-C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \le \frac{1}{2} - \alpha, \tag{62}$$

where C_ϵ is defined as in (4) in the main paper.

Proof. The proof is essentially the same as the proof of Lemma 1. One can simply replace σ in Lemma 1 with $\widetilde{\sigma}_k$ and $0.4748\gamma(x)$ in Lemma 1 with C_k. Then the same arguments still apply. Thus, we skip the details of this proof. □

Then, define $\widehat{p}(z;k) = \frac{1}{m}\sum_{i\in[m]} \mathbb{1}(\widehat{w}_k^i - w_k^* \le z)$. Using the same arguments as in Corollary 1, we know that

$$\widehat{p}\left(C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \ge \frac{1}{2},$$

and

$$\widehat{p}\left(-C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}}); k\right) \le \frac{1}{2},$$

which implies that $|\mathsf{med}\{\widehat{w}_k^i : i \in [m]\} - w_k^*| \le C_\epsilon \frac{\widetilde{\sigma}_k}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{C_k}{\sqrt{n}})$. Then, let

$$\widetilde{\sigma} := \sqrt{\sum_{k=1}^d \widetilde{\sigma}_k^2} = \sqrt{\mathbb{E}[\|\mathbf{H}_F^{-1}(\mathbf{U}\mathbf{H}_F^{-1}\mathbf{p}_F - \mathbf{v})\|_2^2]},$$

and $\widetilde{C} = \max_{k\in[d]} C_k$, we have with probability at least $1 - 4de^{-2t}$,

$$\|\mathsf{med}\{\widehat{\mathbf{w}}^i : i \in [m]\} - \mathbf{w}^*\|_2 \le \frac{C_\epsilon}{\sqrt{n}}\widetilde{\sigma}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + \frac{\widetilde{C}}{\sqrt{n}}).$$

We complete the proof by choosing $t = \frac{1}{2}\log(nmd)$.

Explicit expression of $\widetilde{C}$ To summarize, we provide an explicit expression of $\widetilde{C}$. Let $\mathbf{e}_k$ be the k-th vector in the standard basis, i.e., the k-th column of the $d \times d$ identity matrix, and define $\ell_k(\mathbf{U}, \mathbf{v}) : \mathbb{R}^{d\times d} \times \mathbb{R} \to \mathbb{R}$ as

$$\ell_k(\mathbf{U}, \mathbf{v}) = \mathbf{e}_k^{\mathrm{T}} \mathbf{H}_F^{-1} \mathbf{U} \mathbf{H}_F^{-1} \mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}} \mathbf{H}_F^{-1} \mathbf{v}.$$

Let $\mathbf{H} \sim \mathcal{D}_H$ and $\mathbf{p} \sim \mathcal{D}_p$ and define

$$\widetilde{\sigma}_k := (\mathbb{E}[\ell_k(\mathbf{H} - \mathbf{H}_F, \mathbf{p} - \mathbf{p}_F)^2])^{1/2}, \quad \varsigma_k := \frac{(\mathbb{E}[|\ell_k(\mathbf{H} - \mathbf{H}_F, \mathbf{p} - \mathbf{p}_F)|^3])^{1/3}}{\widetilde{\sigma}_k}.$$
$$\nu_p := (\mathbb{E}[(\|\mathbf{H} - \mathbf{H}_F\|_F^2 + \|\mathbf{p} - \mathbf{p}_F\|_2^2)^{p/2}])^{1/p}, p = 2, 3$$

Then, $\widetilde{C} = \max_{k\in[d]} C_k$, with where

$$C_k = \widehat{C}_0 + \widehat{C}_1 \varsigma_k^3 + \frac{1}{\widetilde{\sigma}_k}\big[(\widehat{C}_{20} + \widehat{C}_{21}\varsigma_k)\nu_2^2 + (\widehat{C}_{30} + \widehat{C}_{31}\varsigma_k)\nu_3^2\big] + \widehat{C}_4,$$

with

$$\widehat{C}_0 = 0.1393, \quad \widehat{C}_1 = 2.3356$$
$$(\widehat{C}_{20}, \widehat{C}_{21}, \widehat{C}_{30}, \widehat{C}_{31}) = \frac{\lambda_F + 2\|\mathbf{p}_F\|_2}{\lambda_F^3}\left(2(\frac{2}{\pi})^{1/6}, 2 + \frac{2^{2/3}}{n^{1/6}}, \frac{(8/\pi)^{1/6}}{n^{1/3}}, \frac{2}{n^{1/2}}\right)$$
$$\widehat{C}_4 = \min\{\frac{4\nu_2^2}{\lambda_F^2 n^{1/2}}, \frac{16\nu_2^3 + 8\nu_3^3/n^{1/2}}{\lambda_F^3 n}\}.$$

F.1 Proof of Lemma 4

We use $\|\cdot\|_2$ and $\|\cdot\|_F$ to denote the operator norm and the Frobenius norm of matrices, respectively. We have the identity

$$(\mathbf{I} + \mathbf{A})^{-1} = \sum_{r=0}^{\infty} (-1)^r \mathbf{A}^r, \quad \forall \|\mathbf{A}\|_2 < 1.$$

Then, we have for all $\mathbf{U} \in \mathbb{R}^{d\times d}$ such that $\|\mathbf{H}_F^{-1}\mathbf{U}\|_2 < 1$,

$$(\mathbf{U} + \mathbf{H}_F)^{-1} = (\mathbf{I} + \mathbf{H}_F^{-1}\mathbf{U})^{-1}\mathbf{H}_F^{-1} = \mathbf{H}_F^{-1} - \mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1} + \sum_{r=2}^{\infty}(-1)^r(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}. \tag{63}$$

Let us consider the set of matrices such that $\|\mathbf{U}\|_F \le \frac{\lambda_F}{2}$. One can check that for any such matrix, we have $\|\mathbf{H}_F^{-1}\mathbf{U}\|_2 \le \frac{1}{2}$. Let

$$\ell_k(\mathbf{U}, \mathbf{v}) = \mathbf{e}_k^{\mathrm{T}} \mathbf{H}_F^{-1} \mathbf{U} \mathbf{H}_F^{-1} \mathbf{p}_F - \mathbf{e}_k^{\mathrm{T}} \mathbf{H}_F^{-1} \mathbf{v}.$$

Then, we know that

$$|\psi_k(\mathbf{U}, \mathbf{v}) - \ell_k(\mathbf{U}, \mathbf{v})| = \left|\mathbf{e}_k^{\mathrm{T}}\mathbf{H}_F^{-1}\mathbf{U}\mathbf{H}_F^{-1}\mathbf{v} - \sum_{r=2}^{\infty}(-1)^r\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}(\mathbf{v} + \mathbf{p}_F)\right|. \tag{64}$$

Denote the operator norm of matrices by $\|\cdot\|_2$. We further have for any $r \ge 1$,

$$|\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}\mathbf{v}| \le \frac{1}{2}\|\mathbf{H}_F^{-1}\mathbf{U}\|_2^{r-1}(\|\mathbf{H}_F^{-1}\mathbf{U}\|_2^2 + \|\mathbf{H}_F^{-1}\mathbf{v}\|_2^2) \le \frac{1}{2^r\lambda_F^2}(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2), \tag{65}$$

where we use the fact $\|\mathbf{U}\|_2 \le \|\mathbf{U}\|_F$. In addition, for any $r \ge 2$,

$$|\mathbf{e}_k^{\mathrm{T}}(\mathbf{H}_F^{-1}\mathbf{U})^r\mathbf{H}_F^{-1}\mathbf{p}_F| \le \|\mathbf{H}_F^{-1}\mathbf{U}\|_2^{r-2}\|\mathbf{H}_F^{-1}\|_2^3\|\mathbf{U}\|_2^2\|\mathbf{p}_F\|_2 \le \frac{\|\mathbf{p}_F\|_2}{2^{r-2}\lambda_F^3}\|\mathbf{U}\|_F^2. \tag{66}$$

Then, we plug (65) and (66) into (64), and obtain

$$|\psi_k(\mathbf{U}, \mathbf{v}) - \ell_k(\mathbf{U}, \mathbf{v})| \le \frac{1}{\lambda_F^2}(\|\mathbf{U}\|_F^2 + \|\mathbf{v}\|_2^2) + \frac{2\|\mathbf{p}_F\|_2}{\lambda_F^3}\|\mathbf{U}\|_F^2,$$

which completes the proof.

G Proof of Observation 1

This proof is essentially the same as the lower bound in the robust mean estimation literature [3] and [5]. We reproduce this result for the purpose of completeness. For a d dimensional Gaussian distribution $P = \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, we denote by P^n the joint distribution of n i.i.d. samples of P. Obviously P^n is equivalent to a dn dimensional Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^+, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu}^+ \in \mathbb{R}^{dn}$ is a vector generated by repeating $\boldsymbol{\mu}$ n times, i.e., $\boldsymbol{\mu}^+ = [\boldsymbol{\mu}^{\mathrm{T}}\ \boldsymbol{\mu}^{\mathrm{T}}\ \cdots\ \boldsymbol{\mu}^{\mathrm{T}}]^{\mathrm{T}}$.

We show that for two d dimensional distributions $P_1 = \mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I})$ and $P_2 = \mathcal{N}(\boldsymbol{\mu}_2, \sigma^2\mathbf{I})$, there exist two dn dimensional distributions Q_1 and Q_2 such that

$$(1-\alpha)P_1^n + \alpha Q_1 = (1-\alpha)P_2^n + \alpha Q_2. \tag{67}$$

If this happens, then no algorithm can distinguish between P_1 and P_2. Let ϕ_1 and ϕ_2 be the PDF of P_1^n and P_2^n, respectively. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be such that the total variation distance between P_1^n and P_2^n is

$$\frac{1}{2}\int \|\phi_1 - \phi_2\|_1 = \frac{\alpha}{1-\alpha}.$$

By the results of the total variation distance between Gaussian distributions, we know that

$$\|\boldsymbol{\mu}_1^+ - \boldsymbol{\mu}_2^+\|_2 \geq \frac{2\alpha\sigma}{1-\alpha}. \tag{68}$$

Let Q_1 be the distribution with PDF $\frac{1-\alpha}{\alpha}(\phi_2 - \phi_1)\mathbb{1}_{\phi_2 \geq \phi_1}$ and Q_2 be the distribution with PDF $\frac{1-\alpha}{\alpha}(\phi_1 - \phi_2)\mathbb{1}_{\phi_1 \geq \phi_2}$. One can verify that (67) is satisfied. In this case, by the lower bound in (68), we get

$$\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 \geq \frac{2\alpha\sigma}{\sqrt{n}(1-\alpha)} \geq \frac{2\alpha\sigma}{\sqrt{n}}.$$

This implies that for two Gaussian distributions such that $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 = \Omega(\frac{\alpha}{\sqrt{n}})$, in the worst case it can be impossible to distinguish these two distributions due to the existence of the adversary. Thus, to estimate the mean $\boldsymbol{\mu}$ of a Gaussian distribution in the distributed setting with α fraction of Byzantine machines, any algorithm that computes an estimation $\widehat{\boldsymbol{\mu}}$ of the mean has a constant probability of error $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = \Omega(\frac{\alpha}{\sqrt{n}})$.

Further, according to the standard results from minimax theory [9], we know that using $\mathcal{O}(nm)$ data, there is a constant probability that $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = \Omega(\sqrt{\frac{d}{nm}})$. Combining these two results, we know that $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 = \Omega(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d}{nm}})$.

Remark. In our paper, we consider the setting where a deterministic α fraction of worker machines are adversarial, whereas in this lower bound, we consider the probabilistic setting where the machines are adversarial with probability α. We note that Chernoff bound ensures that the number of Byzantine machines concentrates around αm. This fact then implies that the lower bound also holds for the case where $\Omega(\alpha m)$ Byzantine machines are selected uniformly at random without replacement from all machines. Since we can translate an average-case bound to a minimax lower bound, we can further show that the lower bound holds under the same setting of our main theorems, that is, an unknown set of αm Byzantine machines are selected without any assumption.

References

[1] A. C. Berry. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136, 1941.

[2] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[3] M. Chen, C. Gao, and Z. Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

[4] C.-G. Esseen. *On the Liapounoff limit of error in the theory of probability.* Almqvist & Wiksell, 1942.

[5] 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.

[6] I. Pinelis and R. Molzon. Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electronic Journal of Statistics*, 10(1):1001–1063, 2016.

[7] I. Shevtsova. On the absolute constants in the berry-esseen-type inequalities. In *Doklady Mathematics*, volume 89, pages 378–381. Springer, 2014.

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

[9] Y. Wu. Lecture notes for ece598yw: Information-theoretic methods for high-dimensional statistics. http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf, 2017.