Proof for Byzantine Papers

1. Krum

1.1 Resource

Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent (NeurIPS 2017)

1.2 Model

1.2.1 Setup

Suppose there are n workers, and f of them are Byzantine workers. Each worker i sent gradient vector V_i to the parameter server.

1.2.2 Prerequisites

(i) (Unbiased expectation) Let G be the gradient distribution where $V_i \sim G$, we have $\mathbb{E}G = g$.

(ii) (Bounded variance) $\mathbb{E}||G-g||^2 = d\sigma^2$ where the gradient vectors are d-dimensional.

(iii) (Convex cost function) The cost function $Q(x)$ needs to be convex.

(iv) (Extra conditions) See **Proposition 2** for details.

1.3 Defense Method

Krum is to preclude the vectors that are too far away. For any $i \neq j$, we denote by $i \rightarrow j$ the fact that V_j belongs to the $n-f-2$ closest vectors to V_i. Then we define a score $s(i) = \sum_{i \rightarrow j} ||V_i - V_j||^2$ and Krum defense aggregation rule $Krum(V_1, \dots, V_n) = V_{i_*}$ where i_* refers to the worker which has the lowest score.

Definition 1. ($(\alpha, f) - \text{Byzantine Resilience.}$) We say an an aggregation rule F is $(\alpha, f) - \text{Byzantine Resilient}$ if F satisfies:

(i) $\langle \mathbb{E}F, g \rangle \geq (1-\sin\alpha) \cdot ||g||^2 > 0$

(ii) for $r = 2, 3, 4, \mathbb{E}||F||^r$ is bounded above by a linear combination of terms $\mathbb{E}||G||^{r_1} \dots E||G||^{r_{n-1}}$ with $r_1 + \dots + r_{n-1} = r$.

Proposition 1. Let $V_1, \dots, V_n$ be any independent and i.i.d random d-dimensional vectors s.t $V_i \sim G$ with $\mathbb{E}G = g$ and $\mathbb{E}||G-g||^2 = d\sigma^2$. If $2f+2 < n$ and $\eta(n,f)\sqrt{d} \cdot \sigma < ||g||$ where

$$\eta(n,f) := \sqrt{2(n-f+\frac{f \cdot (n-f-2) + f^2 \cdot (n-f-1)}{n-2f-2})} = \begin{cases} O(n) & \text{if } f = O(n) \\ O(\sqrt{n}) & \text{if} f = O(1) \end{cases}$$

then the Krum function is $(\alpha, f) - \text{Byzantine Resilient}$ where

$$\sin\alpha = \frac{\eta(n,f) \cdot \sqrt{d} \cdot \sigma}{||g||}$$

Proposition 2. We assume that :

(i) the cost function Q is three times differentiable with continuous derivatives, and is non-negative

(ii) the learning rates satisfy $\sum_t \gamma_t = \infty$ and $\sum_t \gamma_t^2 < \infty$

(iii) $\mathbb{E}G(x,\xi)=\nabla Q(x)$ and $\forall r\in\{2,3,4\},\mathbb{E}||G(x,\xi)||^r\le A_r+B_r||x||^r$ for constants A_r, B_r where G is a gradient estimator

(iv) $\exists\, 0\le\alpha\le\pi/2, \eta(n,f)\cdot\sqrt{d}\cdot\sigma(x)\le||\nabla Q(x)||\cdot\sin\alpha$

(v) beyond a certain horizon, $||x||^2\ge D$, there exists $\epsilon>0$ and $0\le\beta\le\pi/2-\alpha$ such that $||\nabla Q(x)||\ge\epsilon>0$ and $\frac{\langle x,\nabla Q(x)\rangle}{||x||\cdot||\nabla Q(x)||}\ge\cos\beta$

Then the sequence of gradients $\nabla Q(x_t)$ converges almost surely to zero.

1.4 Proof

1.4.1 For Byzantine Resilient

Consider $Krum=Krum(V_1,\dots,V_{n-f},B_1,\dots,B_f)$ and i_* is the index chosen by $Krum$, we have:

$$\begin{gathered}\delta_c(i)+\delta_b(i)=n-f-2\\ n-2f-2\le\delta_c(i)\le n-f-2\\ \delta_b(i)\le f\end{gathered}\tag{1-1}$$

where $\delta_c(i)/\delta_b(i)$ is the number of correct/Byzantine neighbors worker i has.

At first, we focus on the condition (i) of **Definition 1**:

$$\begin{aligned}||\mathbb{E}Krum-g||^2&\le||\mathbb{E}(Krum-\frac{1}{\delta_c(i_*)}\sum_{i_*\to correct\ j}V_j)||^2\\ &\le\mathbb{E}||Krum-\frac{1}{\delta_c(i_*)}\sum_{i_*\to correct\ j}V_j||^2\quad\text{(Jensen inequality)}\\ &\le\sum_{correct\ j}\mathbb{E}||V_i-\frac{1}{\delta_c(i)}\sum_{i\to correct\ j}V_j||^2\mathbb{I}(i_*=i)\\ &+\sum_{byz\ k}\mathbb{E}||B_k-\frac{1}{\delta_c(k)}\sum_{k\to correct\ j}V_j||^2\mathbb{I}(i_*=k)\end{aligned}\tag{1-2}$$

where $\mathbb{I}$ denotes the indicator function. $\mathbb{I}(P)=1$ if predicate P is true, and 0 otherwise.

> **Lemma 1.** (Jensen inequality) If $f(x)$ is convex:
>
> (i) $\mathbb{E}(f(x))\ge f(\mathbb{E}(x))$
>
> (ii) $f(\sum_i\lambda_i x_i)\le\sum_i\lambda_i f(x_i)$, if $\sum_i\lambda_i=1$

When we consider $f(x)=||x||^2$ and $x=Krum-\frac{1}{\delta_c(i_*)}\sum_{i_*\to correct\ j}V_j$ in **Lemma 1**-(i), the second inequality sign in equation 1-2 is true.

Continue, we use **Lemma 1**-(ii) and we consider $f(x)=||x||^2$ and $x=V_i-V_j$ in it.

$$\begin{gathered}||V_i-\frac{1}{\delta_c(i)}\sum_{i_*\to correct\ j}V_j||^2=||\frac{1}{\delta_c(i)}\sum_{i_*\to correct\ j}(V_i-V_j)||^2\le\frac{1}{\delta_c(i)}\sum_{i_*\to correct\ j}||V_i-V_j||^2\qquad\text{(Jensen inequality)}\\ \mathbb{E}||V_i-\frac{1}{\delta_c(i)}\sum_{i\to correct\ j}V_j||^2\le\frac{1}{\delta_c(i)}\sum_{i\to correct\ j}\mathbb{E}||V_i-V_j||^2\le 2d\sigma^2\\ \sum_{correct\ j}\mathbb{E}||V_i-\frac{1}{\delta_c(i)}\sum_{i\to correct\ j}V_j||^2\le(n-f)\cdot 2d\sigma^2\end{gathered}\tag{1-3}$$

Now we consider the case where $V_{i_*}=B_k$ is proposed by a Byzantine worker. This represents that k minimizes the score for all indexes i even it is proposed by correct worker:

$$\sum_{k\to correct\ j}||B_k-V_j||^2+\sum_{k\to byz\ l}||B_k-B_l||^2\le\sum_{i\to correct\ j}||V_i-V_j||^2+\sum_{i\to byz\ l}||V_i-B_l||^2\tag{1-4}$$

Consider the last term of the equation 1-2, for all indexes i of vectors proposed by correct workers:

$$\begin{aligned} ||B_k - \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j||^2 &\leq \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} ||B_k - V_j||^2 \qquad \text{(Jensen inequality)} \\ &\leq \frac{1}{\delta_c(k)} \sum_{i \to correct\ j} ||V_i - V_j||^2 + \frac{1}{\delta_c(k)} \sum_{i \to byz\ l} ||V_i - B_l||^2 \qquad \text{(Krum definition)} \end{aligned} \tag{1-5}$$

We denote $\sum_{i \to byz\ l} ||V_i - B_l||^2$ as $D^2(i)$ and focus on it:

There exists a correct worker $\zeta(i)$ which is farther from i than every neighbor j of i. In particular, for all l such that $i \to l$, $||V_i - B_l|| \leq ||V_i - V_{\zeta(i)}||^2$. Based on the last inequality of equation 1-5, we have:

$$||B_k - \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j||^2 \leq \frac{1}{\delta_c(k)} \sum_{i \to correct\ j} ||V_i - V_j||^2 + \frac{\delta_b(i)}{\delta_c(k)} ||V_i - V_{\zeta(i)}||^2 \tag{1-6}$$

For the changing last term, we just replace $D^2(i)$. And we also have:

$$\begin{aligned} \mathbb{E}||B_k - \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j||^2 &\leq \frac{\delta_c(i)}{\delta_c(k)} \cdot 2d\sigma^2 + \frac{\delta_b(i)}{\delta_c(k)} \sum_{correct\ j \neq i} \mathbb{E}||V_i - V_j||^2 \mathbb{I}(\zeta(i) = j) \\ &\leq (\frac{\delta_c(i)}{\delta_c(k)} + \frac{\delta_b(i)}{\delta_c(k)} (n - f - 1)) \cdot 2d\sigma^2 \\ &\leq (\frac{n-f-2}{n-2f-2} + \frac{f}{n-2f-2} \cdot (n-f-1)) \cdot 2d\sigma^2 \end{aligned} \tag{1-7}$$

For the last inequality, we choose the max of the numerator and the min of the denominator with equation 1-1.

!!! I have a question for the $(n - f - 1)$. Although it is true, it can be reduce to 1. !!!

Combining equation 1-2, 1-3, 1-7, we obtain:

$$||\mathbb{E}Krum - g||^2 \leq [(n-f) + f \cdot (\frac{n-f-2}{n-2f-2} + \frac{f(n-f-1)}{n-2f-2})] \cdot 2d\sigma^2 \leq \eta^2(n,f) \cdot d\sigma^2 \tag{1-8}$$

By the assumption of the **Proposition 1**, we have $\eta\sqrt{d}\sigma < ||g||$, so $\mathbb{E}Krum$ belongs to a ball centered at g with radius $\eta(n,f)\sqrt{d}\sigma$. This implies $\langle \mathbb{E}Krum, g \rangle \geq (1 - \sin\alpha)||g||^2$.

We now focus on condition (ii) of the **Definition 1**:

$$\mathbb{E}||Krum||^r = \sum_{correct\ i} \mathbb{E}||V_i||^r \mathbb{I}(i_* = i) + \sum_{byz\ k} \mathbb{E}||B_k||^r \mathbb{I}(i_* = k) \leq (n-f)\mathbb{E}||G||^r + \sum_{byz\ k} \mathbb{E}||B_k||^r \mathbb{I}(i_* = k) \tag{1-9}$$

When $i_* = k$, for all correct indexes i, based on equation 1-6, we have:

$$\begin{aligned} ||B_k - \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j|| &\leq \sqrt{\frac{1}{\delta_c(k)} \sum_{i \to correct\ j} ||V_i - V_j||^2 + \frac{\delta_b(i)}{\delta_c(k)} ||V_i - V_{\zeta(i)}||^2} \\ \leq C \cdot (\sqrt{\frac{1}{\delta_c(k)}} \cdot \sum_{i \to correct\ j} ||V_i - V_j|| &+ \sqrt{\frac{\delta_b(i)}{\delta_c(k)}} ||V_i - V_{\zeta(i)}||) \leq C \cdot \sum_{correct\ j} ||V_j|| \quad \text{(triangular inequality)} \end{aligned} \tag{1-10}$$

The second inequality comes from the equivalence of norms in finite dimension. Denoting by C a generic constant, we have:

$$\begin{aligned} ||B_k|| &\leq ||B_k - \frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j|| + ||\frac{1}{\delta_c(k)} \sum_{k \to correct\ j} V_j|| \leq C \cdot \sum_{correct\ j} ||V_j|| \\ ||B_k||^r &\leq C \cdot \sum_{r_1 + \ldots + r_{n-f} = r} ||V_1||^{r_1} \cdots ||V_{n-f}||^{r_{n-f}} \end{aligned} \tag{1-11}$$

Since V_i are independent, we finally obtain that $\mathbb{E}||Krum||^r$ is bounded above by a linear combination of the terms because:

$$\mathbb{E}||V_1||^{r_1} \cdots \mathbb{E}||V_{n-f}||^{r_{n-f}} = \mathbb{E}||G||^{r_1} \cdots \mathbb{E}||G||^{r_{n-f}}$$

where $r_1 + \cdots + r_{n-f} = r$. And this completes the proof of **Definition 1**-(ii).

1.4.2 For Convergence

The SGD equation is expressed as follows

$$x_{t+1} = x_t - \gamma_t \cdot Krum(V_1^t, \ldots, V_n^t) = x_t - \gamma_t \cdot Krum_t \tag{2-1}$$

We first show that x_t is almost surely globally confined within the region $||x||^2 \leq D$. And we let $u_t = \phi(||x_t||^2)$ where

$$\phi(a) = \begin{cases} 0 & \text{if } a < D \\ (a-D)^2 & \text{otherwise} \end{cases}$$

And we note that:

$$\phi(a) - \phi(b) \leq (b-a)\phi'(a) + (b-a)^2 \tag{2-2}$$

this becomes a equality when $a, b \geq D$. Applying this inequality to $u_{t+1} - u_t$ yields:

$$\begin{aligned} u_{t+1} - u_t \leq & (-2\gamma_t \langle x_t, Krum_t \rangle + \gamma_t^2 ||Krum_t||^2) \cdot \phi'(||x_t||^2) + 4\gamma_t^2 \langle x_t, Krum_t \rangle^2 \\ & -4\gamma_t^3 \langle x_t, Krum_t \rangle ||Krum_t||^2 + \gamma_t^4 ||Krum||^4 \\ \leq & -2\gamma_t \langle x_t, Krum_t \rangle \phi'(||x_t||^2) + \gamma_t^2 ||Krum_t||^2 \phi'(||x_t||^2) \\ & +4\gamma_t^2 ||x_t||^2 ||Krum_t||^2 + 4\gamma_t^3 ||x_t|| ||Krum_t||^3 + \gamma_t^4 ||Krum_t||^4 \end{aligned} \tag{2-3}$$

where we use $\langle a, b \rangle \leq ||a|| \cdot ||b||$.

Let P_t denote the σ-algebra encoding all the information up to t. We have:

$$\begin{aligned} \mathbb{E}(u_{t+1} - u_t | P_t) \leq & -2\gamma_t \langle x_t, \mathbb{E}(Krum_t) \rangle + \gamma_t^2 \mathbb{E}(||Krum_t||^2) \phi'(||x_t||^2) + 4\gamma_t^2 ||x_t||^2 \mathbb{E}(||Krum_t||^2) \\ & +4\gamma_t^3 ||x_t|| \mathbb{E}(||Krum_t||^2) + \gamma_t^4 \mathbb{E}(||Krum_t||^4) \end{aligned} \tag{2-4}$$

Applying **Definition 1**-(ii) and **Proposition 2**-(iii), we have:

$$\mathbb{E}(u_{t+1} - u_t | P_t) \leq -2\gamma_t \langle x_t, \mathbb{E}(Krum_t) \rangle \phi'(||x_t||^2) + \gamma_t^2 (A_0 + B_0 ||x_t||^4) \tag{2-5}$$

Thus, there exists positive constant A, B such that

$$\mathbb{E}(u_{t+1} - u_t | P_t) \leq -2\gamma_t \langle x_t, \mathbb{E}(Krum_t) \rangle \phi'(||x_t||^2) + \gamma_t^2 (A + B \cdot u_t) \tag{2-6}$$

And because $\langle x_t, \mathbb{E}(Krum_t) \rangle \geq ||x_t|| \cdot ||\mathbb{E}Krum_t|| \cdot \cos(\alpha + \beta) > 0$

$$\mathbb{E}(u_{t+1} - u_t | P_t) \leq \gamma_t^2 (A + B \cdot u_t) \tag{2-7}$$

!!! I don't why equation 2-5 is true. Why $\phi'(\cdot)$ could be added to the tail of the right-hand first term based on equation 2-5? !!!

!!!----- I can't understand the following proof process. -----!!!

Then we define two auxiliary sequences:

$$\begin{gathered} \mu_t = \prod_{i=1}^{t} \frac{1}{1 - \gamma_i^2 B} \xrightarrow{t \to \infty} \mu_\infty \\ u_t' = \mu_t u_t \end{gathered} \tag{2-8}$$

Note that μ_t converges because $\sum_t \gamma_t^2 < \infty$. Then we have:

$$\mathbb{E}(u_{t+1}' - u_t' | P_t) \leq \gamma^2 \mu_t A \tag{2-9}$$

And we define an indicator of the right hand of equation 2-9:

$$\chi_t = \begin{cases} 1 & \text{if } \mathbb{E}(u_{t+1}' - u_t' | P_t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2-10}$$

Then we have:

$$\mathbb{E}(\chi_t \cdot (u'_{t+1} - u'_t)) \leq \mathbb{E}(\chi_t \cdot \mathbb{E}(u'_{t+1} - u'_t|P_t)) \leq \gamma_t^2 \mu_t A \tag{2-11}$$

The right-hand side of the previous inequality is the summand of a convergent series. By the quasi-martingale convergence theorem, this shows that the sequence u'_t converges almost surely, which in turn shows that the sequence u_t converges almost surely, $u_t \to u_\infty \geq 0$.

Let us assume $u_\infty > 0$. When t is large enough, this implies that $||x_t||^2, ||x_{t+1}||^2 > D$ and equation 2-2 becomes an equality which implies that the following infinite sum converges almost surely:

$$\sum_{t=1}^{\infty} \gamma_t \langle x_t, \mathbb{E}Krum_t \rangle \phi'(||x_t||^2) < \infty \tag{2-12}$$

Note that the sequence $\phi'(||x_t||^2)$ converges to a positive value. In the region $||x_t||^2 > D$, we have:

$$\begin{aligned} \langle x_t, \mathbb{E}Krum_t \rangle &\geq \sqrt{D} \cdot ||\mathbb{E}Krum_t|| \cdot \cos(\alpha + \beta) \\ &\geq \sqrt{D} \cdot (||\nabla Q(x_t) - \eta(n, f) \cdot \sqrt{d} \cdot \sigma(x_t)||) \cdot \cos(\alpha + \beta) \\ &\geq \sqrt{D} \cdot \epsilon \cdot (1 - \sin \alpha) \cdot \cos(\alpha + \beta) > 0 \end{aligned} \tag{2-13}$$

This contradicts the fact that $\sum_{t=1}^{\infty} \gamma_t = \infty$. Therefore, the sequence u_t converges to zero. This convergence implies that the sequence $||x_t||^2$ is bounded.

As a sequence, any continuous function of x_t is also bounded, such as $||x_t||^2, \mathbb{E}||G(x, \xi)||^2$ and all the derivatives of the cost function $Q(x_t)$. And we will use K_i as a positive constant whenever such a bound is used.

!!!----- I can't understand the proof process above. -----!!!

We proceed to show that the gradient $\nabla Q(x_t) = \nabla h_t$ converges almost to zero.

Using Taylor expansion and bounding the second derivative with K_1, we obtain:

$$|h_{t+1} - h_t + 2\gamma_t \langle Krum_t, \nabla Q(x_t) \rangle| \leq \gamma_t^2 ||Krum_t||^2 K_1 \tag{2-14}$$

Therefore,

$$\mathbb{E}(h_{t+1} - h_t|P_t) \leq -2\gamma_t \langle Krum_t, \nabla Q(x_t) \rangle + \gamma_t^2 \mathbb{E}(||Krum_t||^2|P_t) K_1 \tag{2-15}$$

Under **Definition 1**, this implies:

$$\mathbb{E}(h_{t+1} - h_t|P_t) \leq \gamma_t^2 K_2 K_1 \tag{2-16}$$

which in turn implies:

$$\mathbb{E}(\chi_t \cdot (h_{t+1} - h_t)) \leq \gamma_t^2 K_2 K_1 \tag{2-17}$$

The right-hand side is the summand of a convergent infinite sum. By the quasi-martingale convergence theorem, the sequence h_t converges almost surely, $Q(x_t) \to Q_\infty$.

Taking the expectation of equation 2-15, and computing the sum from $t = 1$, the convergence of $Q(x_t)$ implies that

$$\sum_{t=1}^{\infty} \gamma_t \langle \mathbb{E}Krum_t, \nabla Q(x_t) \rangle < \infty \tag{2-18}$$

Using a Taylor expansion and defining $\rho_t = ||\nabla Q(x_t)||^2$, we obtain:

$$\rho_{t+1} - \rho_t \leq -2\gamma_t \langle Krum_t, (\nabla^2 Q(x_t)) \cdot \nabla Q(x_t) \rangle + \gamma_t^2 ||Krum_t||^2 K_3 \tag{2-19}$$

Taking the conditional expectations, and bounding the second derivatives by K_4

$$\mathbb{E}(\rho_{t+1} - \rho_t|P_t) \leq 2\gamma_t \langle \mathbb{E}Krum_t, \nabla Q(x_t) \rangle K_4 + \gamma_t^2 K_2 K_3 \tag{2-20}$$

The positive expected variations of ρ_t are bounded

$$\mathbb{E}(\chi_t \cdot (\rho_{t+1} - \rho_t)) \leq 2\gamma_t \mathbb{E}\langle \mathbb{E}Krum_t, \nabla Q(x_t)\rangle K_4 + \gamma_t^2 K_2 K_3 \tag{2-21}$$

The two terms on the right-hand side are the summands of convergent infinite series. By the quasi-martingale convergence theorem, this shows that ρ_t converges almost surely.

We have

$$\langle \mathbb{E}Krum_t, \nabla Q(x_t)\rangle \geq (||\nabla Q(x_t)|| - \eta(n,f) \cdot \sqrt{d} \cdot \sigma(x_t)) \cdot ||\nabla Q(x_t)|| \geq (1 - \sin\alpha) \cdot \rho_t \tag{2-22}$$

This implies that the following infinite series converge almost surely:

$$\sum_{t=1}^{\infty} \gamma_t \cdot \rho_t < \infty \tag{2-23}$$

Since ρ_t converges almost surely, and series $\sum_{t=1}^{\infty} \gamma_t = \infty$ diverges, we conclude that the sequence $||\nabla Q(x_t)||$ converges almost surely to zero.

1.5 Results

Result 1. A single Byzantine worker can prevent the convergence a linear aggregation rule.

Result 2. The expected time complexity of the Krum Function is $O(n^2 \cdot d)$ where gradient vectors are d-dimensional.

Result 3. Krum is Byzantine Resilient.

Result 4. By using Krum, $\nabla Q(x_t)$ converges almost surely to zero.

2. Trimmed Mean

2.1 Resource

Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates (ICML 2018)

2.2 Model

2.2.1 Setup

Suppose that training data points are sampled from unknown distribution D on the sample space Z. Let $f(x;z)$ be the loss function where $w \in W \subseteq \mathbb{R}^d$ is the parameter vector and z is the data point. And we define $F(w) := \mathbb{E}_{z\sim D}[f(w;z)]$. Our goal is to learn a model:

$$w^* = \arg\min_{w\in W} F(w) \tag{1}$$

Suppose there are m workers and each worker stores n data points. Denote by $z^{i,j}$ the j-th data on the i-th worker. And $F_i(w) := \frac{1}{n}\sum_{j=1}^{n} f(w; z^{i,j})$ is the empirical risk function for the i-th worker.

We assume that an α fraction of m workers are Byzantine and denote the set of them by $[B]$ where $|B| = \alpha m$.

2.2.2 Prerequisites

(i) W is convex and $||w_1 - w_2|| \leq D \quad \forall w_1, w_2 \in W$.

(ii) (Smoothness) For any z, the partial derivative $\partial_k f(\cdot; z)$ is L_k-Lipschitz. We also assume the function $f(\cdot; z)$ is L-smooth and $F(\cdot)$ is L_F-smooth. Let $\hat{L} := \sqrt{\sum_k L_k^2}$.

(iii)

2.3 Defense Method

Definition 1. (Coordinate-wise median) For vectors $x^i \in \mathbb{R}^d$ the coordinate-wise median $g_k := \text{med}\{x_k^i\}$ for each $k \in [d]$.

Definition 2. (Coordinate-wise trimmed mean) For $\beta \in [0, \frac{1}{2})$, we remove the the largest and smallest β fraction of x_k^i and compute the average of the remaining elements. We define it as $g := \text{trmean}_\beta$.

Definition 3. (Variance) $Var(x) := \mathbb{E}[||x - \mathbb{E}[x]||_2^2]$.

Definition 4. (Absolute skewness) $\gamma(X) = \frac{\mathbb{E}[|X - \mathbb{E}[x]|^3]}{Var(X)^{\frac{3}{2}}}$ and $\gamma(x) := [\gamma(x_1), \ldots, \gamma(x_d)]^T$.

Definition 5. (Sub-exponential random variables) Random variable X with $\mathbb{E}[X] = \mu$ is called v-sub-exponential if $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{1}{2}v^2\lambda^2},\ \forall |\lambda| < \frac{1}{v}$.

For a differentiable function $h(\cdot) : \mathbb{R}^d \to \mathbb{R}$,

Definition 6. (Lipschitz) h is L-Lipschitz if $|h(w_1) - h(w_2)| \le L||w_1 - w_2||_2,\ \forall w_1, w_2$.

Definition 7. (Smoothness) h is L'-smooth if $||\nabla h(w_1) - \nabla h(w_2)||_2 \le L'||w_1 - w_2||_2,\ \forall w_1, w_2$.

Definition 8. (Strong convexity) h is λ-strongly convex is $h(w_1) \ge h(w_2) + \langle \nabla h(w_2), w_1 - w_2 \rangle + \frac{\lambda}{2}||w_1 - w_2||_2^2,\ \forall w_1, w_2$.

2.4 Proof

2.4.1 For Median-based Gradient Descent

Extra assumptions:

Assumption 1. (Bounded variance of gradient) For any $w \in W$, $\text{Var}(\nabla f(w; z)) \le V^2$.

Assumption 2. (Bounded skewness of gradient) For any $w \in W$, $||\gamma(\nabla f(w; z))||_\infty \le S$.

Proposition 1-1. Suppose each data point $z = (x, y)$ is generated by $y = x^T w^* + \xi,\ w^* \in W$, x is i.i.d. in $\{1, -1\}$ and $\xi \sim N(0, \sigma^2)$. With $f(w; x, y) = \frac{1}{2}(y - x^T w)^2$, we have $\text{Var}(\nabla f(w; x, y)) = (d-1)||w - w^*||_2^2 + d\sigma^2$ and $||\gamma(\nabla f(w; x, y))||_\infty \le 480$.

Proof:

Applying $y = x^T w^* + \xi,\ w^* \in W$, we have:

$$\nabla f(w) = x(x^T w - y) = xx^T(w - w^*) - \xi x \tag{1-1}$$

Applying x is i.i.d. in $\{1, -1\}$, we have:

$$\nabla F(w) = \mathbb{E}[\nabla f(w)] = w - w^* \tag{1-2}$$

Define $\Delta(w) := \nabla f(w) - \nabla F(w)$, we now compute the variance and absolute skewness of $\Delta_k(w)$:

$$\Delta_k(w) = \sum_{i \ne k} x_k x_i (w_i - w_i^*) + (x_k^2 - 1)(w_k - w_k^*) - \xi x_k \tag{1-3}$$

Thus,

$$\mathbb{E}[\Delta_k^2(w)] = \mathbb{E}[\sum_{i \ne k} x_k^2 x_i^2 (w_i - w_i^*)^2 + \xi^2 x_k^2] = ||w - w^*||_2^2 - (w_k - w_k^*)^2 + \sigma^2 \tag{1-4}$$

which yields

$$\text{Var}(\nabla f(w)) = \mathbb{E}[||\nabla f(w) - \nabla F(w)||_2^2] = (d-1)||w - w^*||_2^2 + d\sigma^2 \tag{1-5}$$

Then we proceed to bound $\gamma(\Delta_k(w))$:

$$\gamma(\Delta_k(w)) = \frac{\mathbb{E}[|\Delta_k(w)|^3]}{\text{Var}(\Delta_k(w))^{3/2}} \le \sqrt{\frac{\mathbb{E}[\Delta_k^6(w)]}{\text{Var}(\Delta_k(w))^3}} \tag{1-6}$$

We first find a lower bound for $\text{Var}(\Delta_k(w))^3$, based on equation 1-4, we have:

$$\text{Var}(\Delta_k(w))^3 = (\sum_{i\neq k}(w_i - w_i^*)^2 + \sigma^2)^3 \geq (\sum_{i\neq k}(w_i - w_i^*)^2)^3 + \sigma^6 \tag{1-7}$$

Then we define the following quantities:

$$\begin{aligned} W_1 &= \sum_{i\neq k}(w_i - w_i^*)^6 \\ W_2 &= \sum_{i,j\neq k, i\neq j}(w_i - w_i^*)^4(w_j - w_j^*)^2 \\ W_3 &= \sum_{i,j,l\neq k, i\neq j\neq l}(w_i - w_i^*)^2(w_j - w_j^*)^2(w_l - w_l^*)^2 \end{aligned} \tag{1-8}$$

And we can compute that

$$(\sum_{i\neq k}(w_i - w_i^*)^2)^3 = W_1 + 3W_2 + W_3 \tag{1-9}$$

Combining with equation 1-7,

$$\text{Var}(\Delta_k(w))^3 \geq W_1 + 3W_2 + W_3 + \sigma^6 \tag{1-10}$$

Then we find an upper bound on $\mathbb{E}[\Delta_k^6(w)]$, from equation 1-3 and Holder's inequality, we have:

$$\begin{aligned} \mathbb{E}[\Delta_k^6(w)] = \mathbb{E}[(\sum_{i\neq k} x_k x_i (w_i - w_i^*) - \xi x_k)^6] &\leq 32(\mathbb{E}[(\sum_{i\neq k} x_k x_i (w_i - w_i^*))^6] + \mathbb{E}[\xi^6 x_k^6]) \\ &= 32(\mathbb{E}[(\sum_{i\neq k} x_i (w_i - w_i^*))^6] + 15\sigma^6) \end{aligned} \tag{1-11}$$

> **Lemma 1-1.** (Holder's inequality) Suppose $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n \geq 0$ and $\frac{1}{p} + \frac{1}{q} = 1, \quad p, q > 1$. We have
>
> $$(a_1b_1 + a_2b_2 + \dots + a_nb_n)^{1/p} \leq (\sum a_i^p)^{1/p}(\sum b_i^q)^{1/q}$$

> **Lemma 1-2.** (k moment of origin of normal distribution) Suppose $x \sim N(0, \sigma^2)$. We have
>
> $$\mathbb{E}(x^{2n+1}) = (2n)!! \cdot \sigma^{2n}, \quad n = 1, 2, 3, \dots$$

Based on equation 1-11, we have

$$\mathbb{E}[(\sum_{i\neq k} x_i (w_i - w_i^*))^6] = W_1 + 15W_2 + 15W_3 \tag{1-12}$$

Combining equation 1-11 and 1-12, we have

$$\mathbb{E}[\Delta_k^6(w)] \leq 32(W_1 + 15W_2 + 15W_3 + 15\sigma^6) \tag{1-13}$$

Combining equation 1-6, 1-10, and 1-13, we have

$$\gamma(\Delta_k(w)) \leq \sqrt{\frac{\mathbb{E}[\Delta_k^6(w)]}{\text{Var}(\Delta_k(w))^3}} \leq \sqrt{\frac{32(W_1 + 15W_2 + 15W_3 + 15\sigma^6)}{W_1 + 3W_2 + W_3 + \sigma^6}} \leq 480 \tag{1-14}$$

Proposition 1-2. When the features x in **Proposition 1-1** are i.i.d. Gaussian distributed, the coordinate-wise skewness can be upper bounded by 429.

Theorem 1. Suppose **Prerequisites-(ii)** and **Assumption-1,2** are true, and $F(\cdot)$ is λ_F-strongly convex, we have

$$\alpha + \sqrt{\frac{d\log(1 + nm\hat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}} \leq \frac{1}{2} - \epsilon$$

for some $\epsilon > 0$. When we choose $\eta = 1/L_F$, with probability at least $1 - \frac{4d}{(1+nm\hat{L}D)^d}$, after T parallel iterations, we have

$$||w^T - w^*||_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F})^T ||w^0 - w^*||_2 + \frac{2}{\lambda_F}\Delta$$

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

where

$$\Delta := O(C_\epsilon V(\frac{\alpha}{\sqrt{n}} + \sqrt{\frac{d\log(1+nm\hat{L}D)}{nm}} + \frac{S}{n}))$$

and

$$C_\epsilon = \sqrt{2\pi}\exp(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)$$

with $\Phi^{-1}(\cdot)$ being the inverse of the cumulative distribution function of the standard Gaussian distribution.

Extra Definition 1:

Meanwhile, we define

$$g^i(w) = \nabla F_i(w) \qquad i \in [m]\backslash B$$

and the coordinate-wise median of $g^i(w)$:

$$g(w) = \text{med}\{g^i(w) : i \in [m]\}$$

Corollary 1. When $C_\epsilon \approx 4, \epsilon = 6$, after $T \geq \frac{L_F+\lambda_F}{\lambda_F}\log(\frac{\lambda_F}{2\Delta}||w^0 - w^*||_2)$ parallel iterations, with high probability we can obtain $\hat{w} = w^T$ with error $||\hat{w} - w^*||_2 \leq \frac{4}{\lambda_F}\Delta$.

Here we achieve an error rate of the form $O(\frac{\alpha}{\sqrt{n}} + \frac{1}{\sqrt{nm}} + \frac{1}{n})$.

Proof for Theorem 1:

Suppose that there are m workers and q of them are Byzantine workers where $q = m \cdot \alpha$. They store n adversarial data. For normal workers, each of them stores n one-dimensional data $x \sim D$ where $\mu = \mathbb{E}[x], \sigma^2 = Var(x)$. And $x^{i,j}$ represents the i-th worker's k-th data sample, $\bar{x}^i$ is the average of the i-th worker's data.

Suppose $\hat{p}(z) := \frac{1}{m(1-\alpha)}\sum_{i\in[m]\backslash B}\mathbb{1}(\bar{x}^i \leq z)$, we have the following result on it:

> **Lemma 2-1.** Suppose that for a fixed $t > 0$, we have
>
> $$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}} \leq 1/2 - \epsilon \tag{i}$$
>
> for some $\epsilon > 0$. Then with the probability at least $1 - 4e^{-2t}$, we have
>
> $$\hat{p}(\mu + C_\epsilon\frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})) \geq 1/2 + \alpha$$
>
> and
>
> $$\hat{p}(\mu - C_\epsilon\frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}})) \leq 1/2 - \alpha$$
>
> where C_ϵ is defined in **Theorem 1.**

> **Lemma 2-2.** (Berry-Essen Theorem). Assume that $Y_1, Y_2, \dots, Y_n$ 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,
>
> $$\sup_{s\in R}|\mathbb{P}\{\sqrt{n}\frac{\bar{Y}-\mu}{\sigma} \leq s\} - \Phi(s)| \leq 0.4748\frac{\mathbb{E}[|Y-\mu|^3]}{\sigma^3\sqrt{n}}$$
>
> where $\Phi(s)$ is the cumulative distribution function of the standard normal random variable.

> **Lemma 2-3.** (Bounded Difference Inequality). Let $X_1, \ldots, X_n$ be i.i.d. random variables, and assume that $Z = g(x_1, x_2, \ldots, x_n)$ where g satisfies that for all $j \in [n]$ and all $x_1, \ldots, x_j, x_j', \ldots, x_n$,
>
> $$|g(x_1, \ldots, x_{j-1}, x_j, x_{j+1} \ldots x_n) - g(x_1, \ldots, x_{j-1}, x_j', x_{j+1} \ldots x_n)| \le c_j$$
>
> Then for any $t \ge 0$,
>
> $$\mathbb{P}\{Z - \mathbb{E}[Z] \ge t\} \le \exp(-\frac{2t^2}{\sum_j c_j^2})$$
>
> and
>
> $$\mathbb{P}\{Z - \mathbb{E}[Z] \le t\} \le \exp(-\frac{2t^2}{\sum_j c_j^2})$$

Proof for Lemma 2-1:

Let $\sigma_n = \frac{\sigma}{\sqrt{n}}$ and $c_n = 0.4748\frac{\gamma(x)}{\sqrt{n}}$. 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 $i \in [m]$. We also define the empirical distribution function of $\{W_i : i \in [m]\}$ as $\hat{\Phi}_n(z) = \frac{1}{m(1-\alpha)} \sum_{i \in [m] \setminus B} \mathbb{I}(W_i \le z)$. Thus we have

$$\hat{\Phi}_n(z) = \hat{p}(\sigma_n z + \mu) \tag{2-1}$$

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

$$|\hat{\Phi}_n(z) - \Phi_n(z)| \le \sqrt{\frac{t}{m(1-\alpha)}} \tag{2-2}$$

with the probability a 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 $\Phi_n(z_2) \le \frac{1}{2} - \alpha - \sqrt{\frac{t}{m(1-\alpha)}}$. By union bound, we know that with probability at least $1 - 4e^{-2t}$, $\Phi_n(z_1) \ge \frac{1}{2} + \alpha$ and $\Phi_n(z_2) \le \frac{1}{2} - \alpha$.

According to **Lemma 2-2**, we know that

$$\Phi_n(z_1) \ge \Phi(z_1) - c_n \tag{2-3}$$

it suffices to find z_1 such that

$$\Phi(z_1) = \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \tag{2-4}$$

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

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n = z_1 \Phi'(\xi) = \frac{z_1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} \ge \frac{z_1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}} \tag{2-5}$$

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

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \le \frac{1}{2} - \epsilon \tag{2-6}$$

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

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \ge \frac{z_1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2) \tag{2-7}$$

which yields

$$z_1 \le \sqrt{2\pi} \exp(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n) \tag{2-8}$$

Similarly,

$$z_2 \geq -\sqrt{2\pi}\exp(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2)(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n) \tag{2-9}$$

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

$$\widetilde{p}\left(\mu + C_\epsilon\sigma_n\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right)\right) \geq \frac{1}{2} + \alpha \tag{2-10}$$

and

$$\widetilde{p}\left(\mu - C_\epsilon\sigma_n\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right)\right) \leq \frac{1}{2} - \alpha \tag{2-11}$$

Proof for the main part of Theorem 1:

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

Corollary 2. Suppose that **Lemma 2-1-(i)** is satisfied. Then, with the probability at least $1-4e^{-2t}$, we have equation 2-10 and 2-11. Thus, we have with probability at least $1-4e^{-2t}$,

$$|\operatorname{med}\{\bar{x}^i : i \in [m]\} - \mu| \leq C_\epsilon\frac{\sigma}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma(x)}{\sqrt{n}}) \tag{2-12}$$

Lemma 2-1 and **Corollary 2** can be translated to the estimators of the gradients. Define $g^i(w)$ and $g(w)$ as in **Extra Definition 1**. In addition, for any $w \in W, k \in [d], z \in \mathbb{R}$, we define the empirical distribution function of the k-th coordinate of the gradients on the normal machines:

$$\hat{p}(z; w, k) = \frac{1}{m(1-\alpha)}\sum_{i\in[m]\setminus B}\mathbb{I}(g_k^i(w) \leq z) \tag{2-13}$$

and on all the m machines

$$\hat{p}(z; w, k) = \frac{1}{m}\sum_{i=1}\mathbb{I}(g_k^i(w) \leq z) \tag{2-14}$$

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 notations $\sigma_k^2(w) = \operatorname{Var}(\partial_k f(w; z))$, and $\gamma_k(w) = \gamma(\partial_k f(w; z))$. Then, according to **Lemma 2-1-(i)**, for any fixed $w \in W$ and $k \in [d]$, we have with probability at least $1-4e^{-2t}$

$$\widetilde{p}\left(\partial_k F(\mathbf{w}) + C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}\right); \mathbf{w}, k\right) \geq \frac{1}{2} + \alpha \tag{2-15}$$

and

$$\widetilde{p}\left(\partial_k F(\mathbf{w}) - C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}\right); \mathbf{w}, k\right) \leq \frac{1}{2} - \alpha \tag{2-16}$$

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

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq C_\epsilon\frac{\sigma_k(\mathbf{w})}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(\mathbf{w})}{\sqrt{n}}\right) \tag{2-17}$$

Equation 2-17 gives a bound on the accuracy of the median of means estimator for the gradient at any fixed w and any coordinate $k \in [d]$. To extend this result to all $w \in W$ and all the d coordinates, we need to use union bound and a covering net argument.

Let $W_\delta = \{w^1, w^2, \ldots, w^{N_\delta}\}$ be a finite subset of W such that for any $w \in W$, there exists $w^l \in W_\delta$ such that $||w^l - w||_2 \leq \delta$. According to the standard covering net results, we know that $N_\delta \leq (1 + \frac{D}{\delta})^d$. By a union bound, we know that with probability at least $1 - 4dN_\delta e^{-2t}$, the bounds in equation 2-15 and 2-16 hold for all $w = w^l \in W_\delta, k \in [d]$. By gathering all the k coordinates and using **Assumption 2**, we know that for all $w^l \in W_\delta$

$$\left\|\mathbf{g}\left(\mathbf{w}^{\ell}\right) - \nabla F\left(\mathbf{w}^{\ell}\right)\right\|_2 \leq \frac{C_\epsilon}{\sqrt{n}} V\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right) \tag{2-18}$$

Then consider an arbitrary $w \in W$. Suppose that $||w^l - w||_2 \leq \delta$. Since by **Prerequisites (ii)**, we assume that for each $k \in [d]$, the partial derivative $\partial_k f(w; z)$ is L_k-Lipschitz for all z, we know that for every normal machine $i \in [m] \backslash B$

$$|g_k^i(w) - g_k^i(w^l)| \leq L_k \delta \tag{2-19}$$

Then according to the equation 2-14, we know that for any $z \in \mathbb{R}, \hat{p}(z + l_k\delta; w, k) \geq \hat{p}(z; w, k)$ and $z \in \mathbb{R}, \hat{p}(z - l_k\delta; w, k) \leq \hat{p}(z; w, k)$. Then the bounds in equation 2-15 and 2-16 yield

$$\widetilde{p}\left(\partial_k F\left(\mathbf{w}^{\ell}\right) + L_k\delta + C_\epsilon \frac{\sigma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\right); \mathbf{w}, k\right) \geq \frac{1}{2} + \alpha \tag{2-20}$$

and

$$\widetilde{p}\left(\partial_k F\left(\mathbf{w}^{\ell}\right) - L_k\delta + C_\epsilon \frac{\sigma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\right); \mathbf{w}, k\right) \leq \frac{1}{2} - \alpha \tag{2-21}$$

Using the fact that $\partial_k F(w^l) - \partial_k F(w) \leq L_k\delta$, and **Corollary 2**, we have

$$|g_k(\mathbf{w}) - \partial_k F(\mathbf{w})| \leq 2L_k\delta + C_\epsilon \frac{\sigma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\right) \tag{2-22}$$

Again, by gathering all the k coordinates we get

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2^2 \leq 8\delta^2 \sum_{k=1}^{d} L_k^2 + 2\frac{C_\epsilon^2}{n}\sum_{k=1}^{d}\sigma_k^2\left(\mathbf{w}^{\ell}\right)\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k\left(\mathbf{w}^{\ell}\right)}{\sqrt{n}}\right)^2 \tag{2-23}$$

where we use the fact that $(a+b)^2 \leq 2(a^2+b^2)$. Then by **Assumption 1,2**, we further obtain

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\delta\widehat{L} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}} V\left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right) \tag{2-24}$$

where we use the fact $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Combining equation 2-18 and 2-24, we conclude that for any $\delta > 0$, with probability at least $1 - 4dN_\delta e^{-2t}$, equation 2-24 hold for all $w \in W$. We simply choose $\delta = \frac{1}{nm\hat{L}}$ and $t = d\log(1 + nm\hat{L}D)$. Then, we know that with probability at least $1 - \frac{4d}{(1+nm\hat{L}D)^d}$, we have

$$\|\mathbf{g}(\mathbf{w}) - \nabla F(\mathbf{w})\|_2 \leq 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}} V\left(\alpha + \sqrt{\frac{d\log(1 + nm\widehat{L}D)}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}\right) \tag{2-25}$$

for all $w \in W$.

Proof for convergence:

We now proceed to show the convergence: in the t-th iteration, we define

$$\hat{w}^{t+1} = w^t - \eta g(w^t) \tag{3-1}$$

Thus we have $w^{t+1} = \prod_W(\hat{w}^{t+1})$ where $\prod_W(\cdot)$ is the Euclidean projection which ensures that the model parameter stays in the parameter space W. And by the property of it, we have:

$$||w^{t+1} - w^*||_2 \leq ||\hat{w}^{t+1} - w^*||_2 \tag{3-2}$$

We further have:

$$\begin{aligned}||w^{t+1}-w^*||_2 &\le ||w^t-\eta g(w^t)-w^*||_2 \\ &\le ||w^t-\eta\nabla F(w^t)-w^*||_2+\eta||g(w^t)-\nabla F(w^t)||_2\end{aligned} \tag{3-3}$$

and

$$||w^t-\eta\nabla F(w^t)-w^*||_2^2=||w^t-w^*||_2^2-2\eta\langle w^t-w^*,\nabla F(w^t)\rangle+\eta^2||\nabla F(w^t)||_2^2 \tag{3-4}$$

then we obtain

$$\langle w^t-w^*,\nabla F(w^t)\rangle \ge \frac{L_F\lambda_F}{L_F+\lambda_F}||w^t-w^*||_2^2+\frac{1}{L_F+\lambda_F}||\nabla F(w^t)||_2^2 \tag{3-5}$$

where we use $\nabla F(w^*)=0$ and **Lemma 3-1**.

> **Lemma 3-1** Suppose $f(x)$ is L-smooth m-strongly convex function, we have
>
> $$[\nabla f(x)-\nabla f(y)]^T(x-y)\ge\frac{mL}{m+L}||x-y||^2+\frac{1}{m+L}||\nabla f(x)-\nabla f(y)||^2$$

Let $\eta=1/L_F$, combining equation 3-4 and 3-5, we get

$$\begin{aligned}||w^t-\eta\nabla F(w^t)-w^*||_2^2 &\le (1-\frac{2\lambda_F}{L_F+\lambda_F})||w^t-w^*||_2^2-\frac{2}{L_F(L_F+\lambda_F)}||\nabla F(w^t)||_2^2+\frac{1}{L_F^2}||\nabla F(w^t)||_2^2 \\ &\le (1-\frac{2\lambda_F}{L_F+\lambda_F})||w^t-w^*||_2^2 \qquad (\lambda_F\le L_F)\end{aligned} \tag{3-6}$$

Using the fact $\sqrt{1-x}\le 1-x/2$, we get

$$||w^t-\eta\nabla F(w^t)-w^*||_2\le(1-\frac{\lambda_F}{L_F+\lambda_F})||w^t-w^*||_2 \tag{3-7}$$

Combining equation 3-3 and 3-7, we have

$$||w^{t+1}-w^*||_2\le(1-\frac{\lambda_F}{L_F+\lambda_F})||w^t-w^*||_2+\frac{1}{L_F}\Delta \tag{3-8}$$

where $\Delta=||g(w^t)-\nabla F(w^t)||_2=\frac{2\sqrt{2}}{nm}+\sqrt{\frac{2}{n}}C_\epsilon V(\alpha+\sqrt{\frac{d\log(1+nm\hat{L}D)}{m(1-\alpha)}}+0.4748\frac{S}{\sqrt{n}})$

2.5 Results

Result 1. For Median-based GD: error rate is $O(\frac{\alpha}{\sqrt{n}}+\frac{1}{\sqrt{nm}}+\frac{1}{n})$, order-optimal for strongly convex loss if $n\gtrsim m$.

Result 2. For Trimmed-mean-based GD: error rate is $O(\frac{\alpha}{\sqrt{n}}+\frac{1}{\sqrt{nm}})$, order-optimal for strongly convex loss.

Result 3. For Median-based one-round algorithm: error rate is $O(\frac{\alpha}{\sqrt{n}}+\frac{1}{\sqrt{nm}}+\frac{1}{n})$, order-optimal for strongly convex quadratic loss if $n\gtrsim m$.