证明

kaiyun li

2021 年 6 月 27 日

1 问题设定

2 主要的证明思路

Lemma 2.1 有很高的概率使得下面的式子成立

$$\|g(w) - \nabla F(w)\|_2 \leq \triangle \tag{1}$$

其中，$w \in \mathcal{W}$

Theorem 2.1 假设

1. 对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot; z)$ 的第一个参数的第 k 个坐标 $(k \in [d])$ 的偏导数 $\partial_k f(\cdot; z)$ 是 L_k-*Lipschitz* 连续函数。并且，$f(\cdot; z)$ 是 L-*smooth* 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^{d} L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-*smooth*。

   Remark: 易知$L_F \leq L \leq \hat{L}$

2. *(Bounded variance of gradient)*。对于任意的 $w \in \mathcal{W}, \mathrm{Var}(\nabla f(w; z)) \leq V^2$

3. *(Bounded skewness of gradient)*。对于任意的 $w \in \mathcal{W}, \|\gamma(\nabla f(w; z))\|_\infty \leq S$

只要算法使得引理 *2.1*成立，选择步长 $\eta = 1/L_F$。那么，有很高的概率使得 T 次迭代后，有

$$\|w^{t+1} - w^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F})\|w^t - w^*\|_2 + \frac{1}{L_F}\triangle \tag{2}$$

Proof 2.1 定义

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

因此，我们有 $w^{t+1} = \Pi_{\mathcal{W}}(\hat{w}^{t+1})$。通过 *Euclidean projection* 的性质，

$$\|w^{t+1} - w^*\|_2 \leq \|\hat{w}^{t+1} - w^*\|_2 \tag{4}$$

进一步

$$\begin{aligned} \|\underline{w^{t+1}} - w^*\|_2 &\leq \|\underline{w^t - \eta g(w^t)} - w^*\|_2 \\ &= \|w^t - \eta\nabla F(w^t) + \eta\nabla F(w^t) - \eta g(w^t) - w^*\|_2 \\ &\leq \|w^t - \eta\nabla F(w^t) - w^*\|_2 + \eta \underbrace{\|g(w^t) - \nabla F(w^t)\|_2}_{\triangle} \end{aligned} \tag{5}$$

同时，我们有

$$\|w^t - \eta\nabla F(w^t) - w^*\|_2^2 = \|w^t - w^*\|_2^2 - 2\eta\langle w^t - w^*, \nabla F(w^t)\rangle + \eta^2\|\nabla F(w^t)\|_2^2 \tag{6}$$

因为 $F(w)$ 是 $\lambda_F - strongly\ convex$，通过 *Bubeck et al.(p278,Lemma 3.11)* 知

$$\langle w^t - w^*, \nabla F(w^t) - \underbrace{\nabla F(w^*)}_{=0} \rangle \geq \frac{L_F \lambda_F}{L_F + \lambda_F} \|w^t - w^*\|_2^2 + \frac{1}{L_F + \lambda_F} \|\nabla F(w^t)\|_2^2 \tag{7}$$

令 $\eta = \frac{1}{L_F}$，将 7式带入 6式中。那么，

$$\begin{aligned} \|w^t - \eta \nabla F(w^t) - w^*\|_2^2 &\leq (1 - \frac{\overbrace{2\lambda_F}^{\lambda_F \leq L_F}}{L_F + \lambda_F}) \|w^t - w^*\|_2^2 - \underbrace{(\frac{2}{L_F + \lambda_F} + \frac{1}{L_F^2}) \|\nabla F(w^t)\|_2^2}_{\geq 0} \\ &\leq (1 - \frac{2\lambda_F}{L_F + \lambda_F}) \|w^t - w^*\|_2^2 \end{aligned} \tag{8}$$

使用不等式 $\sqrt{1-x} \leq 1 - \frac{x}{2}$，有

$$\|w^t - \eta \nabla F(w^t) - w^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F}) \|w^t - w^*\|_2 \tag{9}$$

合并 9与 5，得到

$$\|w^{t+1} - w^*\|_2 \leq (1 - \frac{\lambda_F}{L_F + \lambda_F}) \|w^t - w^*\|_2 + \frac{1}{L_F} \triangle \tag{10}$$

■

3 median of mean

首先考虑简单的一维随机变量的鲁棒性估计问题。

假设有 m 个 worker machines，其中有 q 个 Byzantine machines，(令 $\alpha := \frac{q}{m}$) 每个 machines 有 n 个 i.i.d. 样本，服从 $x \sim \mathcal{D}$。第 i 个 machine 的第 j 个样本表示为 $x^{i,j}$，令 $\mu := \mathbb{E}[x], \sigma^2 := Var(x)$ 和 $\gamma(x)$ 表示 x 的绝对偏度。此外，$\bar{x}^i$ 表示第 i 个 machine 的样本均值，即 $\bar{x}^i = \frac{1}{n} \sum_{j=1}^n x^{i,j}$。对于任意的 $z \in \mathbb{R}$，定义正常 worker machines 上的经验分布函数为 $\tilde{p}(z) := \frac{1}{m(1-\alpha)} \sum_{i \in [m] \backslash \mathcal{B}} \mathbb{1}(\bar{x}^i \leq z)$

Lemma 3.1 给定一个 $t > 0$，对于任意的 $\epsilon > 0$。我们有

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma(x)}{\sqrt{n}} \leq \frac{1}{2} - \epsilon \tag{11}$$

那么，至少有 $1 - 4e^{-2t}$ 的概率，使得

$$\tilde{p}\left(\mu + C_\epsilon \frac{\sigma}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma(x)}{\sqrt{n}}) \right) \geq \frac{1}{2} + \alpha \tag{12}$$

和

$$\tilde{p}\left(\mu - C_\epsilon \frac{\sigma}{\sqrt{n}} (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748 \frac{\gamma(x)}{\sqrt{n}}) \right) \leq \frac{1}{2} + \alpha \tag{13}$$

其中

$$C_\epsilon := \sqrt{2\pi} \exp\left(\frac{1}{2} (\Phi^{-1}(1-\epsilon))^2 \right) \tag{14}$$

Proof 3.1 令 $\sigma_n := \frac{\sigma}{\sqrt{n}}$ 和 $c_n := 0.4748 \frac{\mathbb{E}[|x-\mu|^3]}{\sigma^3 \sqrt{n}} = 0.4748 \frac{\gamma(x)}{\sqrt{n}}$。对于所有的 $i \in [m]$ 定义 $W_i := \frac{\bar{x}^i - \mu}{\sigma_n}$。对于 $i \in [m] \backslash \mathcal{B}$, $\Phi_n(\cdot)$ 是 W_i 的分布函数。同样也定义 $\{W_i : i \in [m] \backslash \mathcal{B}\}$ 的经验分布函数 $\tilde{\Phi}_n(\cdot), i.e., \tilde{\Phi}_n(z) = \frac{1}{m(1-\alpha)} \sum_{i \in [m] \backslash \mathcal{B}} \mathbb{1}(W_i \leq z)$。因此，由中心化，简单整理有

$$\tilde{\Phi}_n(z) = \frac{1}{m(1-\alpha)} \sum_{i \in [m] \backslash \mathcal{B}} \mathbb{1}(W_i = \frac{\bar{x}^i - \mu}{\sigma_n} \leq z) = \tilde{p}(\sigma_n z + \mu) \tag{15}$$

$\tilde{\Phi}_n(z)$ 是关于随机变量是 W_i 的函数。根据*Bounded Difference Inequality*，我们需要说明 $\tilde{\Phi}_n(z)$ 满足有界差分的假设，即

$$|\tilde{\Phi}_n(z) - \tilde{\Phi}_n^{'}(z)| \leq c_j \tag{16}$$

代入并整理

$$\left|\frac{1}{m(1-\alpha)}\left[\mathbb{1}(W_i \leq z) - \mathbb{1}(W_i^{'} \leq z)\right]\right| \leq \frac{1}{m(1-\alpha)} \tag{17}$$

现在可以使用 Bounded Difference Inequality 定理。其目的是：说明有很大的概率经验分布集中在均值附近，具体地说，后面会看到 $|\hat{p}(z) - \tilde{p}(z)| \leq \alpha$，所以我们要得到 $\tilde{p}(z) = \frac{1}{2} \pm \alpha$ 的概率以及此时 z 的值。

其想法是: Bounded Difference Inequality 定理建立了经验分布与期望分布之间的联系，期望分布与正态分布又可能通过 Berry-Esseen 定理建立联系，那么寻找 z 的问题可以通过正态分布求解出来。

Theorem 3.1 *Berry-Esseen 中心极限定理:* 假设 $Y_1, \ldots, Y_n$ 是从随机变量 Y 中采样的 *iid* 的样本点。均值为 μ，方差 σ^2，并且满足 $\mathbb{E}[|Y-\mu|^3] < \infty$。那么

$$\sup_{s \in \mathbb{R}} \left|\mathbb{P}\left\{\sqrt{n}\frac{\hat{Y}-\mu}{\sigma} \leq s\right\} - \Phi(s)\right| \leq 0.4748 \frac{\mathbb{E}[|Y-\mu|^3]}{\sigma^3\sqrt{n}} \tag{18}$$

这里 $\hat{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$，并且 $\Phi(s)$ 是标准正态分布的累积分布函数。

Theorem 3.2 *Bounded Difference Inequality* 定理：令 $X_1, \ldots, X_n$ 是独立同分布的随机变量，假设 $Z = g(X_1, \ldots, X_n)$，这里对于所有的 $j \in [n]$ 和所有的 $x_1, x_2, \ldots, x_j, x_j^{'}, \ldots, x_n$，

$$|g(x_1, x_2, \ldots, x_j, \ldots, x_n) - g(x_1, x_2, \ldots, x_j^{'}, \ldots, x_n)| \leq c_j \tag{19}$$

那么对于任意的 $t > 0$，

$$\mathbb{P}\{|Z - \mathbb{E}[Z]| \geq t\} \leq 2\exp\left\{-\frac{2t^2}{\sum_{j=1}^{n} c_j^2}\right\} \tag{20}$$

已知对于任意的 $z \in \mathbb{R}, \mathbb{E}[\tilde{\Phi}_n(z)] = \Phi_n(z)$。因此, 对于任意的 $t > 0$。

$$\mathbb{P}\left\{|\tilde{\Phi}_n(z) - \Phi_n(z)| \geq \sqrt{\frac{t}{m(1-\alpha)}}\right\} \leq 2\exp\left(-\frac{2\frac{t}{m(1-\alpha)}}{\sum_{j=1}^{n} c_j^2}\right) = 2\exp(-2t) \tag{21}$$

Tips: 为了结果的简洁!

也就是说，有至少 $1 - 2\exp(-2t)$ 的概率，使得

$$\left|\tilde{\Phi}_n(z) - \Phi_n(z)\right| \leq \sqrt{\frac{t}{m(1-\alpha)}} \tag{22}$$

令 $z_1 \geq z_2$，满足 $\Phi_n(z_1) \geq \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}}$，并且 $\Phi_n(z_2) \leq \frac{1}{2} - \alpha - \sqrt{\frac{t}{m(1-\alpha)}}$。因为满足半边的概率就是 $2\exp(-2t)$ 所以根据 union bound，我们知道至少有 $1 - 4\exp(-2t)$ 的概率使得 $\tilde{\Phi}_n(z) \geq \frac{1}{2} + \alpha$ 和 $\tilde{\Phi}_n(z) \leq \frac{1}{2} - \alpha$

下一步就是通过 Berry-Esseen 中心极限定理（该定理说明了与正态分布之间的近似误差）将 z 表示出来。

根据 Berry-Esseen 中心极限定理，知

$$\Phi_n(z_1) \geq \Phi(z_1) - \underbrace{c_n}_{\text{近似误差项}} \tag{23}$$

因此，我们要找到的 z_1 需满足

$$\Phi(z_1) = \frac{1}{2} + \alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \tag{24}$$

可以看到，通过转化此时可以通过正态分布这个已知的分布将 z_1 表示出来。

因为 $\Phi_n(z_1) \in [0,1]$，所以存在 $\epsilon \in (0, 1/2)$

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \leq \frac{1}{2} - \epsilon \tag{25}$$

那么，$z_1 \leq \Phi^{-1}(1-\epsilon)$。

下面要构造 z_1 与一个已知的值之间的关系，才方便表示出 z_1。

根据中值定理，存在 $\xi \in [0, z_1]$ 满足

$$\begin{aligned} \Phi(z_1) - \underbrace{\Phi(0)}_{=\frac{1}{2}} &= z_1 \Phi^{'}(\xi) = \alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \\ &= \frac{z_1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2}\right) \\ &\geq \frac{z_1}{\sqrt{2\pi}} \exp\left(-\frac{z_1^2}{2}\right) \end{aligned} \tag{26}$$

因此，

$$\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n \geq \frac{z_1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2\right) \tag{27}$$

整理一下，

$$z_1 \leq \underbrace{\sqrt{2\pi} \exp\left(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2\right)}_{C_\epsilon} \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right) \tag{28}$$

同理

$$z_2 \geq \underbrace{-\sqrt{2\pi} \exp\left(\frac{1}{2}(\Phi^{-1}(1-\epsilon))^2\right)}_{C_\epsilon} \left(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n\right) \tag{29}$$

综上，有

$$\tilde{p}(\sigma_n z_1 + \mu) = \tilde{p}\left(\mu + C_\epsilon \sigma_n (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)\right) \geq \frac{1}{2} + \alpha \tag{30}$$

和

$$\tilde{p}(\sigma_n z_2 + \mu) = \tilde{p}\left(\mu - C_\epsilon \sigma_n (\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + c_n)\right) \leq \frac{1}{2} - \alpha \tag{31}$$

■

前面考虑的是正常的 worker machines 的情况，进一步，我们要说明 median-mean 算法计算出来的梯度与实际的梯度以一个很高的概率仅相差一个有界的误差。定义

$$\hat{p}(z) = \frac{1}{m} \sum_{i \in [m]} \mathbb{1}(\bar{x}^i \leq z) \tag{32}$$

那么，有以下推论

Corollary 3.1 根据 *3.1*，至少有 $1-4\exp(-2t)$ 的概率，使得

$$\begin{aligned}\hat{p}\left(\mu + C_\epsilon\sigma_n(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+c_n)\right) &\geq \frac{1}{2}\\ \hat{p}\left(\mu - C_\epsilon\sigma_n(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+c_n)\right) &\leq \frac{1}{2}\end{aligned} \tag{33}$$

那么，至少有 $1-4\exp(-2t)$ 的概率，使得

$$|\text{med}\{\bar{x}^i : i\in[m]\}-\mu| \leq C_\epsilon\frac{\sigma}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+c_n) \tag{34}$$

前面定义了 $\sigma_n = \frac{\sigma}{\sqrt{n}}$

Proof 3.2 对于任意的 $z\in\mathbb{R}$，有

$$\begin{aligned}|\hat{p}(z)-\tilde{p}(z)| &= \underbrace{\frac{1}{m}\sum_{i\in\mathcal{B}}\mathbb{1}(W_i\leq z)}_{=\frac{q}{m}=\alpha} + \underbrace{(\frac{1}{m}-\frac{1}{m(1-\alpha)})\sum_{i\in[m]\backslash\mathcal{B}}\mathbb{1}(W_i\leq z)}_{\leq 0}\\ &\leq \alpha\end{aligned} \tag{35}$$

因此可以得到 *33*。又因为 *median* 算法总是取最中间的那个数，故

$$\hat{p}(\text{med}\{\bar{x}^i : i\in[m]\}) = \frac{1}{2} \tag{36}$$

用这个式子代替上面的 $\frac{1}{2}$，得到

$$|\text{med}\{\bar{x}^i : i\in[m]\}-\mu| \leq C_\epsilon\frac{\sigma}{\sqrt{n}}(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+c_n) \tag{37}$$

Remark：证明过程中最关键的地方在于为 median 算法总是取最中间的那个数，其累积分布函数值 $=\frac{1}{2}$。 ■

有了引理 3.1和推论 3.1。我们可以通过 union bound 推广到高维，并用 $\epsilon-net$ 理论将参数推广到 $w\in\mathcal{W}$。

定义

$$g^i(w) = \begin{cases}\nabla F_i(w) & i\in[m]\backslash\mathcal{B}\\ * & i\in\mathcal{B}\end{cases} \tag{38}$$

和

$$g(w) = \text{med}\{g^i(w) : i\in[m]\} \tag{39}$$

用 $g_k^i(w)$ 和 $g_k(w)$ 表示 $g^i(w)$ 和 $g(w)$ 的第 k 个坐标位置的值。此外，定义 nomal machies 上和所以的 machines 上的梯度的第 k 维处的经验分布函数，分别为

$$\tilde{p}(z;w,k) = \frac{1}{m(1-\alpha)}\sum_{i\in[m]\backslash\mathcal{B}}\mathbb{1}(g_k^i(w)\leq z), \tag{40}$$

和

$$\hat{p}(z;w,k) = \frac{1}{m}\sum_{i\in[m]}\mathbb{1}(g_k^i(w)\leq z), \tag{41}$$

使用 ∂_k 表示任意函数的偏导数的第 k 个坐标位置的值。也使用 $\sigma_k^2(w) := \mathrm{Var}(\partial_k f(w;z))$，以及 $\gamma_k(w) := \gamma(\partial_k f(w;z))$。

代入引理 3.1中得

$$\begin{aligned}\tilde{p}\left(\partial_k F(w) + C_\epsilon \frac{\sigma_k(w)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}); w, k\right) \geq \frac{1}{2} + \alpha \\ \tilde{p}\left(\partial_k F(w) - C_\epsilon \frac{\sigma_k(w)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}); w, k\right) \leq \frac{1}{2} - \alpha\end{aligned} \tag{42}$$

进一步，根据推论 3.1知，至少 $1 - 4\exp(-2t)$ 的概率，使得

$$|g_k(w) - \partial_k F(w)| \leq C_\epsilon \frac{\sigma_k(w)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}) \tag{43}$$

上述结果是在固定的 w 和固定的 k 下得到的。现在需要通过 union bound 和 $\epsilon - net$ 的结论作进一步的推广。

令 $\mathcal{W}_\delta = \{w^1, w^2, \ldots, w^{N_\delta}\}$ 是 $\mathcal{W}$ 的有限子集，满足对于任意的 $w \in \mathcal{W}$，存在 $w^l \in \mathcal{W}_\delta$ 满足 $\|w^l - w\|_2 \leq \delta$。根据 Vershynin2010(p64) 中的引理 5.2，引理 5.3 知 $N_\delta \leq (1 + \frac{D}{\delta})^d$。通过 union bound 知，至少有 $1 - 4(dN_\delta)\exp(-2t)$ 的概率 (有 d 维，每维对应 N_δ 个向量，全部加起来)。

此时 42、43满足在$w = w^l \in \mathcal{W}_\delta$，以及 $k \in [d]$。

根据假设 2，假设 3，有

$$|g_(w^l) - \nabla F(w^l)| \leq \frac{C_\epsilon}{\sqrt{n}} V(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{S}{\sqrt{n}}) \tag{44}$$

然后，考虑任意的 $w \in \mathcal{W}$。假设 $\|w^l - w\|_2 \leq \delta$。根据假设 1，对于 $k \in [d]$ 和任意的 $z \in \mathbb{R}$，$\partial_k f(w;z)$ 是 L_k-Lipschitz 函数，因此对于每个 nomal machine i $in[m]\backslash\mathcal{B}$，我们有

$$|g_k^i(w) - g_k^i(w^l)| \leq L_k\delta \tag{45}$$

Tips:这里是一个关键点，通过 Lipschitz 连续的定义将 $w \in w^l$ 推广到 $w \in \mathcal{W}$。

根据 $\tilde{p}(z;w,k)$ 的定义知

$$\begin{aligned}\tilde{p}(z + L_k\delta; w, k) \geq \tilde{p}(z\delta; w, k) \\ \tilde{p}(z - L_k\delta; w, k) \leq \tilde{p}(z\delta; w, k)\end{aligned} \tag{46}$$

那么可以稍微放缩一下

$$\begin{aligned}\tilde{p}\left(\partial_k F(w) + L_k\delta + C_\epsilon \frac{\sigma_k(w)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}); w, k\right) \geq \frac{1}{2} + \alpha \\ \tilde{p}\left(\partial_k F(w) - L_k\delta - C_\epsilon \frac{\sigma_k(w)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}); w, k\right) \leq \frac{1}{2} - \alpha\end{aligned} \tag{47}$$

根据推论 3.1，有

$$|g_k(w) - \partial_k F(w)| \leq 2L_k\delta + C_\epsilon \frac{\sigma_k(w^l)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}}) \tag{48}$$

然后合并所有的 k

$$\begin{aligned}\|g(w) - \nabla F(w)\|_2^2 &= \sum_{k=1}^d \left(2L_k\delta + C_\epsilon \frac{\sigma_k(w^l)}{\sqrt{n}}(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}})\right)^2 \\ &\leq 8\delta^2 \sum_{k=1}^d L_k^2 + 2\frac{C_\epsilon^2}{n}\sum_{k=1}^d \sigma_k^2(w^l)(\alpha + \sqrt{\frac{t}{m(1-\alpha)}} + 0.4748\frac{\gamma_k(w)}{\sqrt{n}})^2\end{aligned} \tag{49}$$

这里使用不等式 $(a+b)^2 \le 2(a^2+b^2)$。再通过假设 2、假设 3，有

$$\|g(w)-\nabla F(w)\|_2 \le 2\sqrt{2}\delta \underbrace{\hat{L}}_{\text{假设 1}} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{S}{\sqrt{n}}\right) \tag{50}$$

这里使用了不等式 $\sqrt{a+b} \le \sqrt{a}+\sqrt{b}$。

综上，对于任意的 $\delta>0$，我们至少有 $1-4dN_\delta \exp(-2t)$ 的概率满足对于任意的 $w \in \mathcal{W}$。为了简单，选取 $\delta = \frac{1}{nm\hat{L}}$ 和 $t = d\log(1+nm\hat{L}D)$，那么，我们至少有 $1-\frac{4d}{(1+nm\hat{L}D)^d}$ 的概率，使得

$$\|g(w)-\nabla F(w)\|_2 \le 2\sqrt{2}\frac{1}{nm} + \sqrt{2}\frac{C_\epsilon}{\sqrt{n}}V\left(\alpha+\sqrt{\frac{d\log(1+nm\hat{L}D)}{m(1-\alpha)}}+0.4748\frac{S}{\sqrt{n}}\right) \tag{51}$$

成立。 ■

3.1 non strongly convex losses

假设

1. 对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot;z)$ 的第一个参数的第 k 个坐标（$k \in [d]$）的偏导数 $\partial_k f(\cdot;z)$ 是 L_k-Lipschitz 连续函数。并且，$f(\cdot;z)$ 是 L-smooth 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^d L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-smooth。

   Remark: 易知$L_F \le L \le \hat{L}$

2. (Bounded variance of gradient)。对于任意的 $w \in \mathcal{W}, \mathrm{Var}(\nabla f(w;z)) \le V^2$

3. (Bounded skewness of gradient)。对于任意的 $w \in \mathcal{W}, \|\gamma(\nabla f(w;z))\|_\infty \le S$

4. (Size of $\mathcal{W}$)。参数空间 $\mathcal{W}$ 包含在下面的以 w^* : $\{w \in \mathbb{R}^d : \|w-w^*\|_2 \le 2\|w^0-w^*\|_2\}$ 为中心的 l_2 球里面。

Theorem 3.3 如果假设 *1-4* 成立，损失函数是凸函数且对于任意的 $\epsilon>0$，α 满足

$$\alpha+\sqrt{\frac{t}{m(1-\alpha)}}+0.4748\frac{\gamma(x)}{\sqrt{n}} \le \frac{1}{2}-\epsilon \tag{52}$$

选取 $\eta = \frac{1}{L_F}$，那么，至少有 $1-\frac{4d}{(1+nm\hat{L}D)^d}$ 的概率，在 $T = \frac{L_F}{\triangle}\|w^0-w^*\|_2$ 后，有

$$F(w^T)-F(w^*) \le 16\|w^0-w^*\|_2\triangle\left(1+\frac{1}{2L_F}\triangle\right) \tag{53}$$

Proof 3.3 证明分两步，首先说明当假设 *4* 满足的时候，没有做映射的时，w^t 也在 $\mathcal{W}$ 中。然后证明结论。

对于 $T=0,1,\ldots,T-1$，那么 $w^t \in \mathcal{W}$ 以及对于所有的 $t=0,1,\ldots,T$。定义

$$w^{t+1} = w^t - \eta g(w^t) \tag{54}$$

有

$$\|w^{t+1}-w^*\|_2 \le \|w^t-\eta\nabla F(w^t)-w^*\|_2 + \eta\|g(w^t)-\nabla F(w^t)\|_2 \tag{55}$$

其中

$$\begin{aligned}\|w^t-\eta\nabla F(w^t)-w^*\|_2^2 &= \|w^t-w^*\|_2^2 - 2\eta\langle\nabla F(w^t), \underbrace{w^t-w^*}_{\geq\frac{1}{L_F}\|\nabla F(w^t)-\nabla F(w^*)\|_2^2}\rangle + \eta^2\|\nabla F(w^t)\|_2\\ &\leq \|w^t-w^*\|_2^2 - 2\eta\frac{1}{L_F}\|\nabla F(w^t)\|_2^2+\eta^2\|\nabla F(w^t)\|_2^2\\ &\leq \|w^t-w^*\|_2^2-\frac{1}{L_F^2}\|\nabla F(w^t)\|_2^2\\ &\leq \|w^t-w^*\|_2^2\end{aligned} \tag{56}$$

结合式 *55*和式 *56*有

$$\|w^{t+1}-w^*\|_2\leq\|w^t-w^*\|_2+\frac{\triangle}{L_F} \tag{57}$$

因为 $T=\frac{L_FD_0}{\triangle}$，根据假设 *4* 我们知 $w^t\in\mathcal{W}$ 对于所以的 $t=1,2,\ldots,T$ 成立。

对于 $t=1,2,\ldots,T$，定义 $D_t:=\|w^0-w^*\|_2+\frac{t\triangle}{L_F}$。使用 $F(w)$ 的平滑性质

$$\begin{aligned}F(w^{t+1}) &\leq F(w^t)+\langle\nabla F(w^t),w^{t+1}-w^t\rangle+\frac{L_F}{2}\|w^{t+1}-w^t\|_2^2\\ &= F(w^t)+\eta\langle\nabla F(w^t),\underline{-g(w^t)+\nabla F(w^t)}-\nabla F(w^t)\rangle+\eta^2\frac{L_F}{2}\|\underline{g(w^t)-\nabla F(w^t)}+\nabla F(w^t)\|_2^2\\ &= F(w^t)-\frac{1}{2L_F}\|\nabla F(w^t)\|_2^2+\frac{1}{2L_F}\triangle^2\end{aligned} \tag{58}$$

上面已经找到了递推关系，下面要找 $F(w^t)$ 与 $F(w^*)$ 的关系。

注意到$D_t\leq 2D_0$ 对于所以的 $t=0,1,\ldots,T$ 都成立。对于任意的 w 有

$$F(w)-F(w^*)\leq\langle\nabla F(w),w-w^*\rangle\leq\|\nabla F(w)\|_2\|w-w^*\|_2 \tag{59}$$

因此，

$$\|\nabla F(w)\|_2\geq\frac{F(w)-F(w^*)}{\|w^{t_1-1}-w^*\|_2} \tag{60}$$

1. 假设**存在**$t\in\{0,1,\ldots,T-1\}$ 满足 $\|\nabla F(w^t)\|_2<\sqrt{2}\triangle$。

$$F(w)-F(w^*)\leq\|\nabla F(w)\|_2\|w-w^*\|_2\leq 2\sqrt{2}D_0\triangle \tag{61}$$

2. 假设对于**所有的** $t\in\{0,1,\ldots,T-1\}$ 满足 $\|\nabla F(w^t)\|_2\geq\sqrt{2}\triangle$。那么根据式 *58*和式 *60*有

$$\begin{aligned}F(w^{t+1})\underline{-F(w^*)} &\leq F(w^t)\underline{-F(w^*)}-\frac{1}{2L_F}\underbrace{\|\nabla F(w^t)\|_2^2}_{\geq 2\triangle^2}+\frac{1}{2L_F}\triangle^2\\ &\leq F(w^t)-F(w^*)-\frac{1}{4L_F}\|\nabla F(w^t)\|_2^2\\ &\leq F(w^t)-F(w^*)-\frac{1}{4L_FD_t^2}(F(w^t)-F(w^*))^2\end{aligned} \tag{62}$$

用 $[(F(w^{t+1})-F(w^*))(F(w^t)-F(w^*))]^{-1}$ 同乘上式左右两边有

$$\frac{1}{F(w^{t+1})-F(w^*)}\geq\frac{1}{F(w^t)-F(w^*)}+\frac{1}{4L_FD_t^2}\frac{F(w^t)-F(w^*)}{F(w^{t+1})-F(w^*)}\geq\frac{1}{F(w^t)-F(w^*)}+\frac{1}{16L_FD_0^2} \tag{63}$$

因此根据这个递推式有（T 个）

$$\frac{1}{F(w^{t+1})-F(w^*)}\geq\frac{1}{F(w^0)-F(w^*)}+\frac{T}{16L_FD_0^2} \tag{64}$$

把 $T=\frac{L_F D_0}{\triangle}$ 代入有 $F(w^{t+1})-F(w^*)\leq 16D_0\triangle$

接下来，我们证明 $F(w^{t+1})-F(w^*)\leq 16D_0\triangle+\frac{1}{2L_F}\triangle^2$ 成立。因为我们要证明的结论是 T 次迭代后的情境，所以仅仅是说明 T 次前的情况不够。采用反证法。

设 $t=t_0$ 是首次使得 $F(w^{t+1})-F(w^*)\leq 16D_0\triangle$ 成立，按照前面的推导我们有对于任意的 $t>t_0$ 时有 $F(w^{t+1})-F(w^*)\leq 16D_0\triangle+\frac{1}{2L_F}\triangle^2$ 成立。现在我们假设前式不成立，即我们令 $t_1>t_0$ 是首次使得假设成立的时刻，有 $F(w^{t+1})-F(w^*)>16D_0\triangle+\frac{1}{2L_F}\triangle^2$ 成立。那么，有$F(w^{t_1-1})<F(w^{t_1})$。

根据式 *58*有

$$F(w^{t_1-1})-F(w^*)>F(w^{t_1})-F(w^*)+\frac{1}{2L_F}\|\nabla F(w^t)\|_2^2-\frac{1}{2L_F}\triangle^2>16D_0\triangle \tag{65}$$

同样根据式 *60*有

$$\|\nabla F(w^{t_1-1})\|_2\geq\frac{F(w^{t_1-1})-F(w^*)}{\|w^{t_1-1}-w^*\|_2}>8\triangle \tag{66}$$

再将式 *66*代入到式 *65*中，有有$F(w^{t_1-1})\geq F(w^{t_1})$，因此得到矛盾，故结论成立

■

4 Trim-Mean

Lemma 4.1 假设 *nomal machines* 的样本是一维的，且独立同分布，满足 $v-sub-exponential$ 分布且均值为 μ。那么，对于任意的 $t\geq 0$

$$\mathbb{P}\left\{\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}\bar{x}^i-\mu\right|\geq t\right\}\leq 2\exp\left(-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right) \tag{67}$$

并且，对于任意的 $s\geq 0$

$$\mathbb{P}\left\{\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\}\geq s\right\}\leq 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{68}$$

当 $\beta>\alpha,\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}\bar{x}^i-\mu\right|\leq t$，和 $\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\}\leq s$ 时，有

$$|\text{trmean}_\beta\{\bar{x}^i:i\in[m]\}-\mu|\leq\frac{t+3\beta s}{1-2\beta} \tag{69}$$

Proof 4.1 ***Theorem 4.1*** *Bernstein's* 不等式：假设 $X_1,X_2,\ldots,X_n$ 是独立同分布的 $v-sub-exponential$ 随机变量，且均值为 μ，那么，对于任意的 $t\geq 0$

$$\mathbb{P}\left\{|\frac{1}{n}\sum_{i=1}^n X_i-\mu|\geq t\right\}\leq 2\exp\left\{-n\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right\} \tag{70}$$

因此，对于任意的 $t\geq 0$

$$\mathbb{P}\left\{\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}\bar{x}^i-\mu\right|\geq t\right\}\leq 2\exp\left(-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right) \tag{71}$$

其中的常数因子 n 是由于 $\bar{x}^i$ 的缘故。同样地，对于任意的 $i\in[m]\backslash\mathcal{B},s\geq 0,$

$$\mathbb{P}\left\{|\bar{x}^i-\mu|\geq s\right\}\leq 2\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{72}$$

那么，通过 *union bound* 知

$$\mathbb{P}\left\{\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\}\geq s\right\}\leq 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{73}$$

下面分析 *trimmed mean of means*。

为了表示简单，定义 $\mathcal{M}=[m]\backslash\mathcal{B}$ 为 *nomal worker machines*。$\mathcal{U}\subseteq[m]$ 表示 *untrimmed machines* 的集合。$\mathcal{T}\subseteq[m]$ 表示所有 *trimmed machines* 的集合。

trimmed mean of means 的计算为

$$\text{trmean}_\beta\{\bar{x}^i:i\in[m]\}=\frac{1}{(1-2\beta)m}\sum_{i\in\mathcal{U}}\bar{x}^i \tag{74}$$

进一步有

$$\begin{aligned}|\text{trmean}_\beta\{\bar{x}^i:i\in[m]\}-\mu|&=\left|\frac{1}{(1-2\beta)m}\sum_{i\in\mathcal{U}}\bar{x}^i-\mu\right|\\&=\frac{1}{(1-2\beta)m}\left|\sum_{i\in\mathcal{M}}(\bar{x}^i-\mu)+\sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i-\mu)+\sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i-\mu)\right|\\&\le\frac{1}{(1-2\beta)m}\left(\underbrace{|\sum_{i\in\mathcal{M}}(\bar{x}^i-\mu)|}_{①}+\underbrace{|\sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i-\mu)|}_{②}+\underbrace{|\sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i-\mu)|}_{③}\right)\end{aligned} \tag{75}$$

已知

$$\begin{aligned}&|\sum_{i\in\mathcal{M}}(\bar{x}^i-\mu)|\le\underbrace{t(1-\alpha)m}_{\le tm}\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\\&|\sum_{i\in\mathcal{M}\cap\mathcal{T}}(\bar{x}^i-\mu)|\le 2\beta m\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\\&|\sum_{i\in\mathcal{B}\cap\mathcal{U}}(\bar{x}^i-\mu)|\le\underbrace{\alpha m}_{\le\beta m}\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i-\mu|\end{aligned} \tag{76}$$

所以代入上式，有

$$|\text{trmean}_\beta\{\bar{x}^i:i\in[m]\}-\mu|\le\frac{t+3\beta s}{1-2\beta} \tag{77}$$

■

Remark: 证明中首先用 *Bernstein* 集中不等式给在 *nomal worker* 的梯度聚合给了一个 *bound*。然后利用独立性把式子拆开，得到单项的 *bound*。然后利用 *union bound* 得到了 max 项的 *bound*。

接着分析 trmean 算法的向均值集中的概率及 *bound*。分为三部分，一部分是 *nomal worker machines* 它的概率好分析。第二部分是 *nomal worker machines* 中被 *Trimmed* 的部分它的概率也好分析（因为还是关于 *nomal worker machines* 的）。关键是第三部分，*Byzantine* 中的没有 *Trimmed* 的部分。因为既然没有 *Trimmed* 说明其值的绝对值小于等于 max，所以有上述结果。

我们假设对于任意的 $k\in[d]$ 和 $w\in\mathcal{W},\partial_k f(w;z)$ 是亚指数分布，因此引理 4.1可以直接应用在损失函数的偏导数上。有

$$\mathbb{P}\left\{\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}g_k^i(w)-\partial_k f(w;z)\right|\ge t\right\}\le 2\exp\left(-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right) \tag{78}$$

和

$$\mathbb{P}\left\{\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w)-\partial_k f(w;z)|\}\ge s\right\}\le 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{79}$$

因此，至少有

$$1-2\exp\left(-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right)-2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{80}$$

的概率。有

$$|g_k^i(w)-\partial_k f(w;z)|=|\text{trmean}_\beta\{g_k^i(w):i\in[m]\}-\partial_k f(w;z)|\le\frac{t+3\beta s}{1-2\beta} \tag{81}$$

同样地，我们使用 union bound 和 $\epsilon-net$ 理论将上述结果扩展到 $w\in\mathcal{W}$。

令 $\mathcal{W}_\delta=\{w^1,w^2,\dots,w^{N_\delta}\}$ 是 $\mathcal{W}$ 的有限子集，满足对于任意的 $w\in\mathcal{W}$，存在 $w^l\in\mathcal{W}_\delta$ 满足 $\|w^l-w\|_2\le\delta$。根据 Vershynin2010(p64) 中的引理 5.2，引理 5.3 知 $N_\delta\le(1+\frac{D}{\delta})^d$。通过 union bound 知，至少有

$$1-2dN_\delta\exp\left(-(1-\alpha)mn\min\{\frac{t}{2v},\frac{t^2}{2v^2}\}\right) \tag{82}$$

的概率，使得 $w=w^l\in\mathcal{W}_\delta$ 和 $k\in[d]$ 有下式成立

$$\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}g_k^i(w)-\partial_k f(w;z)\right|\le t \tag{83}$$

同理，至少有

$$1-2(1-\alpha)mdN_\delta\exp\left(-n\min\{\frac{s}{2v},\frac{s^2}{2v^2}\}\right) \tag{84}$$

的概率，使得 $w=w^l\in\mathcal{W}_\delta$ 和 $k\in[d]$ 有下式成立

$$\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w)-\partial_k f(w;z)|\}\le s \tag{85}$$

然后，合并所有的 k 知

$$\|g(w^l)-\nabla F(w^l)\|_2\le\sqrt{d}\frac{t+3\beta s}{1-2\beta} \tag{86}$$

上面考虑的是 $w=w^l\in\mathcal{W}_\delta$ 下面考虑任意的 $w\in\mathcal{W}$

同样是利用 Lipschitz 连续性条件。假设 $\|w^l-w\|_2\le\delta$，因为假设 1（对于任意的 $z\in\mathbb{R}$ 和每一个 $k\in[d]$，$\partial_k f(w;z)$）是 L_k-Lipschitzl 连续函数。那么，对于 nomal machine $i\in[m]\backslash\mathcal{B}$

$$|g_k^i(w)-g_k^i(w^l)|\le L_k\delta,\quad |\partial_k f(w)-\partial_k f(w^l)|\le L_k\delta \tag{87}$$

这意味着：如果 $w^l\in\mathcal{W}_\delta$ 和 $k\in[d]|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}g_k^i(w^l)-\partial_k f(w^l)|\le t$和$\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w^l)-\partial_k f(w^l)|\}\le s$。那么对于任意的 $w\in\mathcal{W}$ 和 $k\in[d]$，有

$$\begin{aligned}\left|\frac{1}{(1-\alpha)m}\sum_{i\in[m]\backslash\mathcal{B}}g_k^i(w)-\partial_k f(w;z)\right|&\le t+2L_k\delta\\ \max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w)-\partial_k f(w;z)|\}&\le s+2L_k\delta\end{aligned} \tag{88}$$

因此

$$|g_k(w)-\partial_k f(w)|=|\text{trmean}_\beta\{g_k(w):i\in[m]\}-\partial_k f(w)|\le\frac{t+3\beta s}{1-2\beta}+\frac{2(1+3\beta)}{1-2\beta}L_k\delta \tag{89}$$

推出

$$\|g(w)-\nabla F(w)\|_2\le\sqrt{2d}\frac{t+3\beta s}{1-2\beta}+\sqrt{2}\frac{2(1+3\beta)}{1-2\beta}\hat{L}\delta \tag{90}$$

■

5 geometric median

Lemma 5.1 令 $z_1,\dots,z_n$ 是 *Hilbert* 空间中的点。令 z_* 表示 *geometric median* 的 $(1+\gamma)$-*approximation*。即，对于任意的 $\gamma>0$

$$\sum_{i=1}^{n}\|z_*-z_i\|\le(1+\gamma)\min_z\sum_{i=1}^{n}\|z-z_i\| \tag{91}$$

对于任意的 $\alpha\in(0,\frac{1}{2})$ 和给定 $r\in\mathbb{R}$，如果 $\sum_{i=1}^{n}\mathbb{1}\{\|z_i\|\le r\}\ge(1-\alpha)n$，那么

$$\|z_*\|\le C_\alpha r+\gamma\frac{\min_z\sum_{i=1}^{n}\|z-z_i\|}{(1-2\alpha)n}\le C_\alpha r+\gamma\frac{\max_{1\le i\le n}\|z_i\|}{(1-2\alpha)} \tag{92}$$

这里

$$C_\alpha=\frac{2(1-\alpha)}{1-2\alpha} \tag{93}$$

Proof 5.1 令 $S=\{i:\|z_i\|\le r\}$，对于任意的 $i\in S$，我们有

$$\|z_*-z_i\|\ge\|z_*\|-\|z_i\|\ge\|z_*\|-\underbrace{2r}_{2\|z_i\|\le 2r}+\|z_i\| \tag{94}$$

此外，对于任意的 $i\notin S$ 根据三角不等式

$$\|z_*-z_i\|\ge\|z_i\|-\|z_*\| \tag{95}$$

合并上面两个不等式（把 $i\in[0,n]$ 分为 $i\in S$ 和 $i\notin S$）来计算，有

$$\sum_{i=1}^{n}\|z_*-z_i\|\ge\sum_{i=1}^{n}\|z_i\|+(2|S|-n)\|z_*\|-2|S|r \tag{96}$$

因为 z_* 是 *geometric median* 的 $(1+\gamma)$-*approximation*。因此，

$$\sum_{i=1}^{n}\|z_i\|+(2|S|-n)\|z_*\|-2|S|r\le(1+\gamma)\min_z\sum_{i=1}^{n}\|z-z_i\| \tag{97}$$

注意到 $\sum_{i=1}^{n}\|z_i\|=\sum_{i=1}^{n}\|0-z_i\|\ge\min_z\|z-z_i\|$。因此（代入到第一项中）有

$$(2|S|-n)\|z_*\|-2|S|r\le\gamma\min_z\sum_{i=1}^{n}\|z-z_i\| \tag{98}$$

因此

$$\|z_*\|\le\frac{2|S|r}{2|S|-n}+\gamma\frac{\min_z\sum_{i=1}^{n}\|z-z_i\|}{2|S|-n}\le\frac{2(1-\alpha)r}{1-2\alpha}r+\gamma\frac{\min_z\sum_{i=1}^{n}\|z-z_i\|}{(1-2\alpha)n} \tag{99}$$

这里使用了 $|S|\ge(1-\alpha)n$ ■

Remark：引理 5.1说明只要有超过半数的 z_i 的模长在 r 范围以内，那么通过 geometric median 方法计算出来的 z_* 在半径为 $C_\alpha r$ 的球内。可以从证明的最后一步中看到，为满足不等式，分母为正数，所以这个半数也可以这么得来。

5.1 ALGORITHM

5.2 CONVERGENCE RESULTS AND ANALYSIS

假设对于任意的 $\epsilon>0$ 有 $2(1+\epsilon)q\le k\le m$。固定一个常数 $\alpha\in(\frac{1}{2+2\epsilon},\frac{1}{2})$ 和任意的 $\delta>0$ 满足 $\delta\le\alpha-q/k$。

算法将 m 个梯度分为 k 批，定义一个函数 $Z_l : \theta \to \mathbb{R}^d$，它刻画梯度的均值与期望之间的差距

$$\begin{aligned} Z_l(\theta) &:= \frac{1}{b} \sum_{j=(l-1)b+1}^{lb} \nabla \bar{f}^j(\theta) - \nabla F(\theta) \\ &= \frac{k}{N} \sum_{j=(l-1)b+1}^{lb} \sum_{i \in S_j} \nabla f(X_i, \theta) - \nabla F(\theta) \end{aligned} \tag{100}$$

这里 $b = m/k$，本地数据集的大小为 $|S_j| = N/m$。

定义一个 good event

$$\varepsilon_{\alpha,\xi_1,\xi_2} := \left\{ \sum_{l=1}^{k} \mathbb{1}\{\forall \theta : C_\alpha \|z_l(\theta)\| \le \xi_2 \|\theta - \theta^*\| + \xi_1\} \ge k(1-\alpha) + q \right\} \tag{101}$$

第一步：说明事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生的条件下，算法收敛；第二步：说明该事件发生的概率很大。

定义向量函数 $g_t(\cdot)$

$$g_t(\theta) = (g_t^1(\theta), \ldots, g_t^m(\theta)), \quad \forall \theta \tag{102}$$

满足

$$g_t^j(\theta) = \begin{cases} \nabla \bar{f}^j(\theta) & \text{if} j \notin \mathcal{B}_t \\ \bigstar & o.w. \end{cases} \tag{103}$$

如果第 t 次迭代时，machine j 不是 Byzantine 那么定义

$$\tilde{Z}_l(\theta) := \frac{1}{b} \sum_{j=(l-1)b+1}^{lb} g_t^j(\theta) - \nabla F(\theta) \tag{104}$$

Lemma 5.2 如果事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，对于第 $t > 1$ 次迭代

$$\|\mathcal{A}_k(g_t(\theta)) - \nabla F(\theta)\| \le \xi_2 \|\theta - \theta^*\| + \xi_1, \quad \forall \theta \in \Theta \tag{105}$$

Proof 5.2

$$\|\mathcal{A}_k(g_t(\theta)) - \nabla F(\theta)\| = \|\tilde{Z}_1(\theta), \ldots, \tilde{Z}_m(\theta)\| \tag{106}$$

事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，则是说 k 个中至少有 $k(1-\alpha)+q$ 个 *batches* $(\{Z_l : 1 \le l \le k\})$ 满足 $C_\alpha \|z_l(\theta)\| \le \xi_2 \|\theta - \theta^*\| + \xi_1\}$。正常的 *machines* 占了 $k(1-\alpha)$，所以说明事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生则有超过半数的 *machines* 梯度是正常的。使用引理 *5.1* (引理说明，只要有过半数的模长满足条件，则使用 $(1+\gamma)$ 近似得到的模长也是满足相应关系的)，即证！

■

引理 5.2说明 geometic median 算法得到的梯度满足一个模长的关系式（也说明事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，则意味着这个结论)，即是说明计算结果在真实的梯度附近（有界)。下面说明事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，则算法收敛

假设 1：期望风险函数 $F : \Theta \to \mathbb{R}$ 是 $L - strongly$ 凸函数。并且梯度是 $M - Lipschitz$ 平滑的。也就是说，对于所有的 $\theta, \theta \in \Theta$ 有

$$F(\theta') \ge F(\theta) + \langle \nabla F(\theta), \theta' - \theta \rangle + \frac{L}{2} \|\theta' - \theta\|^2 \tag{107}$$

和

$$\|\nabla F(\theta) - \nabla F(\theta')\| \le M \|\theta - \theta'\| \tag{108}$$

Lemma 5.3 如果假设 *1* 满足，令 $\eta = L/(2M^2)$，定义

$$\theta' = \theta - \eta \times \nabla F(\theta) \tag{109}$$

那么

$$\|\theta' - \theta^*\| \le \sqrt{1 - L^2/(4M^2)} \|\theta - \theta^*\| \tag{110}$$

Proof 5.3 由于 $\nabla F(\theta^*)=0$，所以有

$$\begin{aligned}\|\theta^{'}-\theta^*\|^2 &= \|\theta-\theta^*-\eta\nabla F(\theta)\|^2 \\ &= \|\theta-\theta^*-\eta(\nabla F(\theta)-\nabla F(\theta^*))\|^2 \\ &= \|\theta-\theta^*\|^2+\eta^2\underbrace{\|\nabla F(\theta)-\nabla F(\theta^*)\|^2}_{\leq M^2\|\theta-\theta^*\|^2}-2\eta\langle\theta-\theta^*,\nabla F(\theta)-\nabla F(\theta^*)\rangle\end{aligned} \tag{111}$$

最后一项，由凸函数的性质有

$$\begin{aligned}F(\theta) &\geq F(\theta^*)+\langle\nabla F(\theta^*),\theta-\theta^*\rangle+\frac{L}{2}\|\theta-\theta^*\|^2 \\ F(\theta^*) &\geq F(\theta)+\langle\nabla F(\theta),\theta^*-\theta\rangle\end{aligned} \tag{112}$$

两式合并整理后有

$$0\geq\langle\theta-\theta^*,\nabla F(\theta)-\nabla F(\theta^*)\rangle+\frac{L}{2}\|\theta-\theta^*\|^2 \tag{113}$$

因此，

$$\|\theta^{'}-\theta^*\|\leq\sqrt{1-L^2/(4M^2)}\|\theta-\theta^*\| \tag{114}$$

■

下面将算法求出来的梯度，代到引理 5.3中去，目的是得到当梯度被 bound 的时候的一个收敛性结果，这是我们想要的。

Lemma 5.4 如果，令 $G_t(\theta)$ 是关于梯度的函数，满足下式

$$\|G_t(\theta)-\nabla F(\theta)\|\leq\xi_2\|\theta-\theta^*\|+\xi_1,\quad\forall\theta\in\Theta \tag{115}$$

选取 $\eta=L/(2M^2)$，对于每个 $t\geq 1$ 和

$$\rho:=1-\sqrt{1-L^2/(4M^2)}-\xi_2L/(2M^2)>0 \tag{116}$$

那么，第 t 次迭代 $\{\theta_t\}$ 时，满足

$$\|\theta_t-\theta^*\|\leq(1-\rho)^t\|\theta_0-\theta^*\|+\eta\xi_1/\rho \tag{117}$$

Proof 5.4 选定任意的 $t\leq 1$，有

$$\begin{aligned}\|\theta_t-\theta^*\| &= \|\theta_{t-1}-\eta G_t(\theta_{t-1})-\theta^*\| \\ &= \|\theta_{t-1}-\eta\nabla F(\theta_{t-1})-\theta^*+\eta(\nabla F(\theta_{t-1})-G_t(\theta_{t-1}))\| \\ &\leq \|\theta_{t-1}-\eta\nabla F(\theta_{t-1})-\theta^*\|+\eta\|(\nabla F(\theta_{t-1})-G_t(\theta_{t-1}))\|\end{aligned} \tag{118}$$

根据引理 5.3第一项

$$\|\underbrace{\theta_{t-1}-\eta\nabla F(\theta_{t-1})}_{\theta'}-\theta^*\|\leq\sqrt{1-L^2/(4M^2)}\|\theta-\theta^*\| \tag{119}$$

另外，第二项是我们的假设

$$\|G_t(\theta)-\nabla F(\theta)\|\leq\xi_2\|\theta-\theta^*\|+\xi_1,\quad\forall\theta\in\Theta \tag{120}$$

因此，

$$\|\theta_t-\theta^*\|\leq\left(\sqrt{1-L^2/(4M^2)}+\eta\xi_2\right)\|\theta_{t-1}-\theta^*\|+\eta xi_1 \tag{121}$$

再通过递推关系有

$$\|\theta_t-\theta^*\|\leq(1-\rho)^t\|\theta_0-\theta^*\|+\eta\xi_1\underbrace{\sum_{\tau=0}^{t-1}(1-\rho)^\tau}_{\leq\frac{1}{\rho}} \tag{122}$$

■

Theorem 5.1 假设事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，则算法满足

$$\|\theta_t - \theta^*\| \leq (1-\rho)^t \|\theta_0 - \theta^*\| + \eta\xi_1/\rho \tag{123}$$

Proof 5.5 引理 5.2说明事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生，则算法计算出来的梯度在真实的梯度周围一个常数 *bound* 中。再用引理 5.4即证！ ■

上面已经证完了第一步，下面是说明第二步。 我们要说明的是事件 $\varepsilon_{\alpha,\xi_1,\xi_2}$ 发生的概率，因为可以使用 Chernoff 不等式，所以下面我们只需要考察

$$\mathbb{P}\{\forall\theta : C_\alpha\|Z_l(\theta)\| \leq \xi_2\|\theta - \theta^*\| + \xi_1\} \geq 1 - \delta \tag{124}$$

这里，$\alpha \geq (q/k, 1/2)$ 和 $0 < \delta \leq \alpha - q/k$。要想说清楚这个事件发生的概率，需要使用到 $\epsilon - net$ 的理论

分析：给定一个 l，令 $\theta_1, \ldots, \theta_{N_{\epsilon_l}}$ 是 Θ_l 的 ϵ_l-cover。固定任意的 $\theta \in \Theta_l$。存在一个 $1 \leq j_l \leq N_{\epsilon_l}$ 满足 $\|\theta - \theta_{jl}\|_2 \leq \epsilon_l$ 因为

$$\begin{aligned}\|Z_l(\theta)\| &= \|\nabla\bar{f}_n(\theta) - \nabla F(\theta)\| \\ &\leq \underbrace{\|\nabla F(\theta) - \nabla F(\theta_{jl})\|}_{①} + \underbrace{\|\nabla\bar{f}_n(\theta) - \nabla\bar{f}_n(\theta_{jl})\|}_{②} + \underbrace{\|\nabla\bar{f}_n(\theta_{jl}) - \nabla F(\theta_{jl})\|}_{③}\end{aligned} \tag{125}$$

为了求出概率，要做两件事，第一是利用温和的假设来得到界，第二是根据这个界算出相应的概率。

针对式子 ①，由假设 1 有

$$\|\nabla F(\theta) - \nabla F(\theta_{jl})\| \leq M\|\theta - \theta_{jl}\| \leq M\epsilon_l \tag{126}$$

针对式子 ②，

假设 2：对于任意的 $\delta \in (0,1)$，存在关于 n 是非增的 $M' = M'(n,\delta)$，满足

$$\mathbb{P}\left\{\sup_{\theta,\theta'\in\Theta;\theta\neq\theta'} \frac{\|\frac{1}{n}\sum_{i=1}^n(\nabla f(X_i,\theta) - \nabla f(X_i,\theta'))\|}{\|\theta - \theta_{'}\|} \leq M'\right\} \geq 1 - \frac{\delta}{3} \tag{127}$$

那么根据假设 2

$$\sup_{\theta\in\Theta}\|\nabla\bar{f}_n(\theta) - \nabla\bar{f}_n(\theta_{jl})\| \leq M'\epsilon_l \tag{128}$$

为方便后面描述，定义事件

$$\varepsilon_1 = \left\{\sup_{\theta,\theta'\in\Theta;\theta\neq\theta'} \frac{\|\nabla f(X_i,\theta) - \nabla f(X_i,\theta')\|}{\|\theta - \theta_{'}\|} \leq M'\right\} \tag{129}$$

它发生的概率为 $1 - \frac{\delta}{3}$。

针对式子 ③。使用三角不等式

$$\begin{aligned}\|\nabla\bar{f}_n(\theta_{jl}) - \nabla F(\theta_{jl})\| &\leq \|\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*)\| + \|\nabla\bar{f}_n(\theta_{jl}) - \nabla\bar{f}_n(\theta^*) - (\nabla F(\theta_{jl}) - \nabla F(\theta^*))\| \\ &\leq \underbrace{\|\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*)\|}_{④} + \underbrace{\|\frac{1}{n}\sum_{i=1}^n h(X_i,\theta_{jl}) - \mathbb{E}\left[h(X_i,\theta_{jl})\right]\|}_{⑤}\end{aligned} \tag{130}$$

观察到式④和⑤都是要求样本均值与期望的差距，自然会想到大数定律。另外我们要刻画分布在均值附近的概率，我们可以假设满足次指数分布（直觉上，次指数分布说明的是样本均值有很大的概率向其期望集中）。现在将③分解为④和⑤两个子式。

针对式子 ④。

假设 3：存在正常数 σ_1 和 α_1 满足对于单位向量 $v \in B, \langle\nabla f(X,\theta^*), v\rangle$ 是放缩因子为 σ_1 和 α_1 的次指数分布。即

$$\sup_{v\in B}\mathbb{E}[\exp(\lambda\langle\nabla f(X,\theta^*), v\rangle)] \leq \exp(\sigma_1^2\lambda^2/2), \quad \forall|\lambda| \leq \frac{1}{\alpha_1}, \tag{131}$$

这里 B 表示单位球 $\{\theta : \|\theta\|_2 = 1\}$

Lemma 5.5 如果假设 *3* 成立，对于 $\delta \in (0,1)$ 和任意正整数 n，令

$$\triangle_1(n,d,\delta,\sigma_1) = \sqrt{2}\sigma_1\sqrt{\frac{d\log 6 + \log(3/\delta)}{n}} \tag{132}$$

如果 $\triangle_1(n,d,\delta,\sigma_1) \le \sigma_1^2/\alpha_1$，那么

$$\mathbb{P}\left\{\|\frac{1}{n}\sum_{i=1}^n \nabla f(X_i,\theta^*) - \nabla F(\theta^*)\| \ge 2\triangle_1(n,d,\delta,\sigma_1)\right\} \le \frac{\delta}{3} \tag{133}$$

Proof 5.6 为表示简单，记 $\triangle_1(n,d,\delta,\sigma_1)$ 为 $\triangle_1$，记 $\frac{1}{n}\sum_{i=1}^n \nabla f(X_i,\theta^*) = \nabla \bar{f}_n(\theta)$。

参考资料：（ϵ 网覆盖数）对于任何 $\epsilon > 0$，单位欧几里德球 B_2^n 的覆盖数满足结论

$$\left(\frac{1}{\epsilon}\right)^n \le \mathcal{N}(B_2^n,\epsilon) \le \left(1+\frac{2}{\epsilon}\right)^n \tag{134}$$

当 $\epsilon \in (0,1]$ 内，我们有

$$\left(\frac{1}{\epsilon}\right)^n \le \mathcal{N}(B_2^n,\epsilon) \le \left(\frac{3}{\epsilon}\right)^n \tag{135}$$

参考资料：（在 ϵ-网上计算子范数）设 A 是一个 $m\times n$ 矩阵，$\epsilon \in [0,1)$，那么，对于 S^{n-1} 的 ϵ-网，有

$$\sup_{x\in\mathcal{N}}\|Ax\| \le \|A\| \le (1-\epsilon)^{-1}\sup_{x\in\mathcal{N}}\|Ax\| \tag{136}$$

参考资料：（次指数分布）均值为 μ 的随机变量 X 是次指数分布，如果 $\exists v>0$，和 $\alpha>0$ 满足

$$\mathbb{E}[\exp(\lambda(X-\mu))] \le \exp\left(\frac{v^2\lambda^2}{2}\right), \quad \forall|\lambda| \le \frac{1}{\alpha} \tag{137}$$

参考资料：（次指数分布有关定理）如果 $X_1,\dots,X_n$ 是独立随机变量，这里 X_i 是放缩因子为 (v_i,α_i)，均值为 μ_i 的次指数分布。那么 $\sum_{i=1}^n X_i$ 是 (v_*,α_*) 的次指数分布，这里 $v_* = \sum_{i=1}^n v_i^2, \alpha_* = \max_{1\le i\le n}\alpha_i$。那么

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i-\mu_i) \ge t\right\} \le \begin{cases}\exp(-t^2/(2v_*^2)) & if\ 0\le t\le v_*^2/\alpha_* \\ \exp(-t/(2\alpha_*)) & o.w.\end{cases} \tag{138}$$

令 $\mathcal{V} = \{v_1,\dots,v_{N_{1/2}}\}$ 表示单位球的 $\frac{1}{2}$-*cover*。由 *vershynin* 的引理 *5.2*、引理 *5.3* 知 $\log N_{1/2} \le d\log 6$，以及

$$\|\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*)\| \le 2\sup_{v\in\mathcal{V}}\left\{\langle\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*), v\rangle\right\} \tag{139}$$

因为有 $\nabla F(\theta^*) = 0$，所以 $\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*)$ 满足假设 *3* 的前提条件。如果假设 *3* 满足且 $\triangle_1 \le \sigma_*^2/\alpha_*$，那么对于 $v\in\mathcal{V}$

$$\mathbb{P}\left\{\langle\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*), v\rangle \ge \triangle_1\right\} \le \exp(-n\triangle_1^2/(2\sigma_1^2)) \tag{140}$$

Tips: 这里的 n 是求均值的 $1/n$ 乘过去的，注意我们的假设的条件！

我们知 $\mathcal{V}$ 包含 6^d 个向量。使用 *union bound*，得到

$$\begin{aligned}\mathbb{P}\underbrace{\left\{2\sup_{v\in\mathcal{V}}\{\langle\nabla\bar{f}_n(\theta^*) - \nabla F(\theta^*), v\rangle\} \ge 2\triangle_1\right\}}_{\times 2\&\sup} &\le 6^d\exp(-n\triangle_1^2/(2\sigma_1^2)) \\ &= \exp(-n\triangle_1^2/(2\sigma_1^2) + d\log 6)\end{aligned} \tag{141}$$

进一步，有

$$\mathbb{P}\left\{\|\frac{1}{n}\sum_{i=1}^n \nabla f(X_i,\theta^*) - \nabla F(\theta^*)\| \ge 2\triangle_1(n,d,\delta,\sigma_1)\right\} \le \frac{\delta}{3} \tag{142}$$

为得到结论，首先是将公式 *139*代入公式 *141*中，然后把 $\triangle_1$ 代入，即证！

■

为描述方便，记事件

$$\varepsilon_2 = \left\{ \|\frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i,\theta^*) - \nabla F(\theta^*)\| \le 2\triangle_1(n,d,\delta,\sigma_1) \right\} \tag{143}$$

那么事件发生的概率为 $1-\frac{\delta}{3}$
针对式子 ⑤。定义

$$\begin{aligned} h(x,\theta) &:= \nabla f(x,\theta) - \nabla f(x,\theta^*) \\ \mathbb{E}[h(x,\theta)] &:= \nabla F(x,\theta) - \nabla F(x,\theta^*) \end{aligned} \tag{144}$$

Remark：这里想要说明样本均值与期望的差值，前面的假设 1 是说明函数 F 的性质，而这里是关于向均值集中的讨论，注意区分。
假设 4：存在正常数 σ_2 和 α_2 满足对于任意的 $\theta \in \Theta$ 且 $\theta \neq \theta^*$ 和单位向量 $v \in B$，$\langle h(X,\theta) - \mathbb{E}[h(X,\theta)], v\rangle / \|\theta-\theta^*\|$ 是放缩因子为 (σ_2,α_2) 的次指数分布，对于所有的 $|\lambda| \le \frac{1}{\alpha_2}$

$$\sup_{\theta\in\Theta, v\in B} \mathbb{E}\left[\exp\left(\frac{\lambda\langle h(X,\theta)-\mathbb{E}[h(X,\theta)], v\rangle}{\|\theta-\theta^*\|}\right)\right] \le \exp(\sigma_2^2\lambda^2/2) \tag{145}$$

Lemma 5.6 如果假设 *4* 成立，那么对于任意固定的 $\theta\in\Theta$，令

$$\triangle_1^{'}(n,d,\delta,\sigma_2) = \sqrt{2}\sigma_2\sqrt{\frac{d\log 6 + \log(3/\delta)}{n}} \tag{146}$$

如果 $\triangle_1^{'}(n,d,\delta,\sigma_2) \le \sigma_2^2/\alpha_2$，那么

$$\mathbb{P}\left\{ \|\frac{1}{n}\sum_{i=1}^{n} h(X,\theta) - \mathbb{E}[h(X,\theta)]\| \ge 2\triangle_1^{'}(n,d,\delta,\sigma_2)\|\theta-\theta^*\| \right\} \le \frac{\delta}{3} \tag{147}$$

其证明同引理 5.5。
Remark：子式⑤是 $h(X,\theta_{jl})$ 是 ϵ 网里的元素，为了扩展到任意的 $\theta\in\Theta$。我们要用到假设 2。
综上，可以提出下面命题

Theorem 5.2 如果假设 *2-4* 满足，并且对于正数 r 有 $\Theta \subset \{\theta : \|\theta-\theta^*\| \le r\sqrt{d}\}$，任意的 $\delta\in(0,1)$ 和正数 n。$\triangle_1$ 如上述定义，还定义

$$\triangle_2(n) = \sigma_2\sqrt{\frac{2}{n}}\sqrt{d\log\left(\frac{18M\vee M'}{\sigma_2}\right) + \frac{1}{2}d\log\frac{n}{d} + \log\left(\frac{6\sigma_2^2 r\sqrt{n}}{\alpha_2\sigma_1\delta}\right)} \tag{148}$$

如果 $\triangle_1(n) \le \sigma_1^2/\alpha_2$ 和 $\triangle_2(n) \le \sigma_2^2/\alpha_2$，那么

$$\mathbb{P}\left\{ \forall\theta\in\Theta : \left\| \frac{1}{n}\sum_{i=1}^{n}\nabla f(X_i,\theta) - \nabla F(\theta) \right\| \le 8\triangle_2\|\theta-\theta^*\| + 4\triangle_1 \right\} \ge 1-\delta \tag{149}$$

Proof 5.7 记

$$\tau = \frac{\alpha_2\sigma_1}{2\sigma_2^2},\ and\ l^* = \lceil r\sqrt{d}/\tau \rceil \tag{150}$$

假设 l^* 是一个整数，对于整数 $1\le l\le l^*$ 定义

$$\Theta_l := \{\theta : \|\theta-\theta^*\| \le \underbrace{\tau l}_{\le r\sqrt{d}}\} \tag{151}$$

给定 l，令 $\theta_1,\dots,\theta_{N_{\epsilon_l}}$ 是 Θ_l 的 ϵ_l 网覆盖，这里的 ϵ_l 定义为

$$\epsilon_l = \frac{\sigma_2 \tau l}{M \vee M'} \sqrt{\frac{d}{n}} \tag{152}$$

其中 $M \vee M' = \max\{M, M'\}$。根据 *Verchynin* 的引理知，

$$\log N_{\epsilon_l} \leq d \log(3\tau l/\epsilon_l) \tag{153}$$

根据前述分析，

$$\begin{aligned}
\|\nabla \bar{f}_n(\theta) - \nabla F(\theta)\| &\leq \underbrace{\|\nabla F(\theta) - \nabla F(\theta_{jl})\|}_{①} + \underbrace{\|\nabla \bar{f}_n(\theta) - \nabla \bar{f}_n(\theta_{jl})\|}_{②} + \underbrace{\|\nabla \bar{f}_n(\theta_{jl}) - \nabla F(\theta_{jl})\|}_{③} \\
&\leq ① + ② + \underbrace{\|\nabla \bar{f}_n(\theta^*) - \nabla F(\theta^*)\|}_{④} + \underbrace{\|\frac{1}{n}\sum_{i=1}^{n} h(X_i, \theta_{jl}) - \mathbb{E}\left[h(X_i, \theta_{jl})\right]\|}_{⑤} \\
&= ① + ② + ④ + ⑤
\end{aligned} \tag{154}$$

由假设 *1*，对于①一定有 *126*成立；由假设 *2*，对于②，事件 ε_1 发生的概率为 $\mathbb{P}\{\varepsilon_1\} \geq 1 - \delta/3$；由假设 *3*，对于④，事件 ε_2 发生的概率为 $\mathbb{P}\{\varepsilon_2\} \geq 1 - \delta/3$；最后同样为表述方便，记事件

$$\mathcal{F}_l = \left\{ \sup_{1 \leq j \leq N_{\epsilon_l}} \left\| \frac{1}{n}\sum_{i=1}^{n} h(X, \theta) - \mathbb{E}[h(X, \theta)] \right\| \leq 2\tau l \triangle_2 \right\} \tag{155}$$

由引理 *5.6*，把 $\triangle_2$ 这个界代进去有

$$\begin{aligned}
\mathbb{P}\{\mathcal{F}_l^c\} &= \mathbb{P}\left\{ \sup_{1 \leq j \leq N_{\epsilon_l}} \left\| \frac{1}{n}\sum_{i=1}^{n} h(X, \theta) - \mathbb{E}[h(X, \theta)] \right\| \leq 2 \underbrace{\tau l}_{\geq \|\theta - \theta^*\|} \triangle_2 \right\} \\
&\leq \sum_{j=1}^{N_{\epsilon l}} \mathbb{P}\left\{ \left\| \frac{1}{n}\sum_{i=1}^{n} h(X, \theta) - \mathbb{E}[h(X, \theta)] \right\| \leq 2\tau l \triangle_2 \right\} \quad \text{用到} N_{\epsilon_l} \leq (3\tau l/\epsilon_l)^d \\
&\leq \frac{\delta}{3l^*} \frac{1}{(3\tau l/\epsilon_l)^d} \left((3\tau l/\epsilon_l)^d\right) = \frac{\delta}{3l^*}
\end{aligned} \tag{156}$$

因此，有 $\mathbb{P}\{\mathcal{F}_l\} \geq 1 - \frac{\delta}{3l^*}$。

整理一下，$\varepsilon_1 \cap \varepsilon_2 \cap \mathcal{F}_l$

$$\begin{aligned}
\sup_{\theta \in \Theta_l} \left\|\nabla \bar{f}_n(\theta) - \nabla F(\theta)\right\| &\leq (M + M')\epsilon_l + 2\triangle_1 + 2\triangle_2 \tau l \\
&\leq 4\triangle_2 \tau l + 2\triangle_1
\end{aligned} \tag{157}$$

最后使用了不等式 $(M \vee M')\epsilon_l \leq \triangle_2 \tau l$。令

$$\varepsilon = \varepsilon_1 \cap \varepsilon_2 \cap (\cap_{l=1}^{l^*} \mathcal{F}_l) \tag{158}$$

那么，使用 *union bound* 有 $\mathbb{P}\{\varepsilon\} \geq 1 - \delta$。前面考虑的是在 $\theta \in \Theta_l$ 上的结论，对于 $\theta \in \Theta_{l^*}$。存在 $1 \leq l \leq l^*$ 满足 $(l-1)\tau < \|\theta - \theta^*\| \leq l\tau$。如果 $l > 2$，那么 $l \leq 2(l-1)$（把 $l > 2$ 代一下即可）有

$$\begin{aligned}
\left\|\nabla \bar{f}_n(\theta) - \nabla F(\theta)\right\| &\leq 4\triangle_2 \tau l + 2\triangle_1 \\
&\leq 4\triangle_2 \tau 2(l-1) + 2\triangle_1 \\
&\leq 8\triangle_2 \|\theta - \theta^*\| + 2\triangle_1
\end{aligned} \tag{159}$$

总结一下，事件 ε 发生的概率为 $1 - \delta$，有

$$\sup_{\theta \in \Theta_{l^*}} \left\|\nabla \bar{f}_n(\theta) - \nabla F(\theta)\right\| \leq 8\triangle_2 \|\theta - \theta^*\| + 2\triangle_1 \tag{160}$$

这里基于了假设 $\Theta \in \Theta_{l^*}$

■

6 draft

Algorithm 1: mean-based robust SGD

Input: 上传的梯度

Output: 聚合梯度

1 按列计算均值并减去均值;

2 按列统计正、负数的个数然后去掉个数少的部分;

3 按列将剩下数据重复一遍上述步骤;

4 执行完毕后，剩下数据按列求均值;

先记作:

$$mbrSGD\{x_i; i = 1, \dots, m\} \tag{161}$$

Lemma 6.1 令点集 $\mathcal{I} \in [a, b]$，均值为 $\mu_{\mathcal{I}}$；点集 $\mathcal{O} \in [c, d]$ 且 $\mathcal{O} \cap \mathcal{I} = \emptyset$，其中 $a, b, c, d \in \mathcal{R}$ 满足 $c \le a < b \le d$。令 $\mathcal{T} = \mathcal{I} \cup \mathcal{O}$ 且均值为 *0*，如果 $\frac{3|\mathcal{T}|}{4} < |\mathcal{I}| \le |\mathcal{T}|$（$|\cdot|$ 表示集合 $\cdot$ 中元素的个数），那么有

$$|\tilde{\mu}| = |mbrSGD\{x; x \in \mathcal{T}\}| < \left(\frac{2k}{m+4} + 1\right) r \quad r = max\{|a|, |b|\} \tag{162}$$

Proof 6.1 若 $\mathcal{O} = \emptyset$，结论显然成立！若 $\mathcal{O} \neq \emptyset$，分情况讨论。为描述方便记负（非负）数点集分别为 $\mathcal{T}^-, \mathcal{T}^+$，其均值分别为 μ^-, μ^+；执行算法第 *2* 步后的集合分别为 $\mathcal{T}^{'}, \mathcal{I}^{'}, \mathcal{O}^{'}$，记 $\mathcal{T}^{'}$ 的均值为 $\mu^{'}$，大于均值 $\mu^{'}$ 的点集合为 $\mathcal{T}^{'+}$ 和 $\mathcal{T}^{'-}$。

为讨论出 $\tilde{\mu}$ 的边界，则考虑最坏的情况：**集合 $\mathcal{O}$ 将集合 $\mathcal{I}$ 的均值拉得尽可能的偏离** $\mu_{\mathcal{I}}$。所以先对集合 $\mathcal{O}$ 在集合 $\mathcal{I}$ 的一侧时的情况进行讨论。然后讨论集合 $\mathcal{O}$ 在集合 $\mathcal{I}$ 的两侧的情况。令 $r = \max\{|a|, |b|\}, k = \max\{|c|, |d|\}/r$

算法第 *2* 步中，均值为 *0*，所以有

$$\sum_{i \in \mathcal{T}^-} |x_i| = \sum_{i \in \mathcal{T}^+} |x_i| \tag{163}$$

可以推出 $k \le (m-1)r$ 算法第 *3* 步中，均值为 $\mu^{'}$

$$\sum_{i \in \mathcal{T}^{'-}} |\mu^{'} - x_i| = \sum_{i \in \mathcal{T}^{'+}} |x_i - \mu^{'}| \tag{164}$$

1. 集合 $\mathcal{O}$ 在集合 $\mathcal{I}$ 的一侧时。不妨令集合 $\mathcal{O}$ 在集合 $\mathcal{I}$ 的右侧。均考虑极端情况

(a) 如果算法第 *2* 步去掉 $\mathcal{T}^-$，算法第 *3* 步去掉 $\mathcal{T}^{'-}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^-$，那么 $0 \in [a, b)$，算法第 *3* 步为去掉 $\mathcal{T}^{'-}$，那么 $\mu^{'} \in (0, b]$。因此，极端情况为：a 处放置 $\frac{m}{2} - 1$ 个点，0 处放置 $\frac{m}{4}$ 个点，b 处放置集合 $\mathcal{I}$ 剩下的点（最少有 *2* 个）。

$$\left(\frac{m}{2} - 1\right)|a| = 2r + \sum_{i \in \mathcal{T}^{'+}} x_i \tag{165}$$

等式左右两边同除 $\frac{m}{4} + 1$，即是 $\tilde{\mu}$。所以

$$\tilde{\mu} \le \frac{\left(\frac{m}{2} - 1\right)|a|}{\frac{m}{4} + 1} < 2|a| \tag{166}$$

考虑到对称性，则可以得到

$$|\tilde{\mu}| < 2r \tag{167}$$

(b) 如果算法第 *2* 步去掉 $\mathcal{T}^-$，算法第 *3* 步去掉 $\mathcal{T}^{'+}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^-$，那么 $0 \in [a,b)$，算法第 *3* 步为去掉 $\mathcal{T}^{'+}$，那么 $|\mu^{'}| < |\mu^+|$。因此，极端情况为：a 处放置 $\frac{m}{2}-1$ 个点。

$$\left(\frac{m}{2}-1\right)|a| = \left(\frac{m}{2}+2\right)\mu^+ \tag{168}$$

即 $|\mu^+| < |a|$，所以如果考虑两侧，有

$$|\tilde{\mu}| \leq r \tag{169}$$

(c) 如果算法第 *2* 步去掉 $\mathcal{T}^+$，算法第 *3* 步去掉 $\mathcal{T}^{'-}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^+$，那么 $0 \in [b,d)$，算法第 *3* 步为去掉 $\mathcal{T}^{'-}$，那么 $\mu^{'} \in (a,b)$。因此，极端情况为：a 处放置 $\frac{m}{4}-1$ 个点，d 处放置 *1* 个点。

$$\mu^{'} \leq \tilde{\mu} \leq b \tag{170}$$

因此，考虑两端的情况，有

$$|\tilde{\mu}| < r \tag{171}$$

(d) 如果算法第 *2* 步去掉 $\mathcal{T}^+$，算法第 *3* 步去掉 $\mathcal{T}^{'+}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^+$，那么 $0 \in [b,d)$，算法第 *3* 步为去掉 $\mathcal{T}^{'+}$，那么 $\mu^{'} \in (b,0)$。因此，极端情况为：d 处放置 *1* 个点，0 处放置 *1* 个点。

$$|\mu^{'}| < |\tilde{\mu}| < |a| \leq r \tag{172}$$

2. 集合 $\mathcal{O}$ 在集合 $\mathcal{I}$ 的两侧。

(a) 如果算法第 *2* 步去掉 $\mathcal{T}^+$，算法第 *3* 步去掉 $\mathcal{T}^{'+}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^+$，那么 $0 \in [a,d)$，算法第 *3* 步为去掉 $\mathcal{T}^{'+}$，那么 $\mu^{'} \in (a,0)$。因此，极端情况为：d 处放置 *1* 个点，0 处放置 *1* 个点。

因为算法第 *3* 步去除的是右边，所以最后的均值不会超过 $2|a|$。因此

$$|\tilde{\mu}| < 2r \tag{173}$$

(b) 如果算法第 *2* 步去掉 $\mathcal{T}^+$，算法第 *3* 步去掉 $\mathcal{T}^{'-}$。

【分析】算法第 *2* 步为去掉 $\mathcal{T}^+$，那么 $0 \in [a,d)$，算法第 *3* 步为去掉 $\mathcal{T}^{'-}$，那么 $\mu^{'} \in (c,b)$。因此，分为 $0 \in [a,b]$ 和 $0 \notin [a,b]$ 两种极端情况。只需要讨论 $0 \in [a,b]$ 的情况，它的界更大。

算法第 *2* 步尽可能多地去掉集合 $\mathcal{I}$ 中的元素，算法第 *3* 步，尽可能少地去掉集合 $\mathcal{O}$ 中的元素。d 处放置 *1* 个点。

$$\left(\frac{m}{2}+2\right)|\mu^{'}| \leq \left(\frac{m}{2}-2\right)|b| + |d| \tag{174}$$

由于算法第 *3* 步去掉集合 $\mathcal{T}^{'-}$，所以有 $|\tilde{\mu}| < |\mu^{'}|$

$$|\tilde{\mu}| < r + \frac{2k}{m+4}r \tag{175}$$

综上所述，考虑到所有的情况，命题得证！

■

Lemma 6.2 假设 *nomal machines* 的样本是一维的，且独立同分布，满足 $v-sub-exponential$ 分布且均值为 μ。那么，对于任意的 $s \geq 0$

$$\mathbb{P}\left\{\max_{i\in[m]\setminus\mathcal{B}}\{|\bar{x}^i - \mu|\} \geq s\right\} \leq 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{176}$$

当 $0 \leq \alpha < \frac{1}{4}$，$k \geq 0$ 是个有界的实数。和 $\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i - \mu|\} \leq s$ 时，有

$$|msbSGD\{\bar{x}^i : i \in [m]\} - \mu| \leq \left(1 + \frac{2k}{m+4}\right) s \tag{177}$$

Proof 6.2 根据 *Bernstein's* 不等式，对于任意的 $i \in [m]\backslash\mathcal{B}, s \geq 0$,

$$\mathbb{P}\left\{|\bar{x}^i - \mu|\right\} \geq s\} \leq 2\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{178}$$

那么，通过 *union bound* 知

$$\mathbb{P}\left\{\max_{i\in[m]\backslash\mathcal{B}}\{|\bar{x}^i - \mu|\} \geq s\right\} \leq 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{179}$$

根据引理 *6.1*，有

$$|msbSGD\{\bar{x}^i : i \in [m]\} - \mu| \leq \left(1 + \frac{2k}{m+4}\right) s \tag{180}$$

■

我们假设对于任意的 $k \in [d]$ 和 $w \in \mathcal{W}, \partial_k f(w;z)$ 是次指数分布，因此引理 6.2可以直接应用在损失函数的偏导数上。有

$$\mathbb{P}\left\{\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w) - \partial_k f(w;z)|\} \geq s\right\} \leq 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{181}$$

因此，至少有

$$1 - 2(1-\alpha)m\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{182}$$

的概率。有

$$|g_k^i(w) - \partial_k f(w;z)| \leq \left(1 + \frac{2k}{m+4}\right) s \tag{183}$$

同样地，我们使用 union bound 和 $\epsilon - net$ 理论将上述结果扩展到 $w \in \mathcal{W}$。

令 $\mathcal{W}_\delta = \{w^1, w^2, \ldots, w^{N_\delta}\}$ 是 $\mathcal{W}$ 的有限子集，满足对于任意的 $w \in \mathcal{W}$，存在 $w^l \in \mathcal{W}_\delta$ 满足 $\|w^l - w\|_2 \leq \delta$。根据 Vershynin2010(p64) 中的引理 5.2，引理 5.3 知 $N_\delta \leq (1 + \frac{D}{\delta})^d$。通过 union bound 知，至少有

$$1 - 2(1-\alpha)mdN_\delta\exp\left(-n\min\{\frac{s}{2v}, \frac{s^2}{2v^2}\}\right) \tag{184}$$

的概率，使得 $w = w^l \in \mathcal{W}_\delta$ 和 $k \in [d]$ 有下式成立

$$\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w) - \partial_k f(w;z)|\} \leq s \tag{185}$$

然后，合并所有的 k 知

$$\|g(w^l) - \nabla F(w^l)\|_2 \leq \sqrt{d}\left(1 + \frac{2k}{m+4}\right) s \tag{186}$$

上面考虑的是 $w = w^l \in \mathcal{W}_\delta$ 下面考虑任意的 $w \in \mathcal{W}$

同样是利用 Lipschitz 连续性条件。假设 $\|w^l - w\|_2 \leq \delta$，因为假设 1（对于任意的 $z \in \mathbb{R}$ 和每一个 $k \in [d]$，$\partial_k f(w;z)$）是 L_k-Lipschitzl 连续函数。那么，对于 nomal machine $i \in [m]\backslash\mathcal{B}$

$$|g_k^i(w) - g_k^i(w^l)| \leq L_k\delta, \quad |\partial_k f(w) - \partial_k f(w^l)| \leq L_k\delta \tag{187}$$

这意味着：如果 $w^l \in \mathcal{W}_\delta$ 和 $\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w^l) - \partial_k f(w^l)|\} \leq s$。那么对于任意的 $w \in \mathcal{W}$ 和 $k \in [d]$，有

$$\max_{i\in[m]\backslash\mathcal{B}}\{|g_k^i(w) - \partial_k f(w;z)|\} \leq s + 2L_k\delta \tag{188}$$

因此

$$|g_k(w) - \partial_k f(w)| \leq \left(1 + \frac{2k}{m+4}\right)(s + 2L_k\delta) \tag{189}$$

推出

$$\|g(w) - \nabla F(w)\|_2 \leq \sqrt{2d}\left(1 + \frac{2k}{m+4}\right)s + \sqrt{2}\left(1 + \frac{2k}{m+4}\right)\hat{L}\delta \tag{190}$$

■

1. 对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot; z)$ 的第一个参数的第 k 个坐标（$k \in [d]$）的偏导数 $\partial_k f(\cdot; z)$ 是 L_k-Lipschitz 连续函数。并且，$f(\cdot; z)$ 是 L-smooth 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^d L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-smooth。

2. (sub-exponential gradients)。假设对于所有的 $k \in [d]$ 和 $w \in \mathcal{W}$，$f(w; z)$ 的关于 w 的坐标维度 k 的偏导 $\partial_k f(w; z)$ 是次指数分布。

Theorem 6.1 如果假设

1. *对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot; z)$ 的第一个参数的第 k 个坐标（$k \in [d]$）的偏导数 $\partial_k f(\cdot; z)$ 是 L_k-Lipschitz 连续函数。并且，$f(\cdot; z)$ 是 L-smooth 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^d L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-smooth。*

2. *(sub-exponential gradients)。假设对于所有的 $k \in [d]$ 和 $w \in \mathcal{W}$，$f(w; z)$ 的关于 w 的坐标维度 k 的偏导 $\partial_k f(w; z)$ 是次指数分布。*

成立，$F(\cdot)$ 是 $\lambda_F - strongly$ 凸函数。并且对于任意的 $\epsilon > 0$ 有 $\alpha \leq \beta \leq \frac{1}{4} - \epsilon$。选择步长 $\eta = 1/L_F$。那么至少有 $1 - \frac{4d}{(1+nm\hat{L}D)^d}$ 的概率，在 T 次迭代后，有

$$\|w^T - w^*\|_2 \leq \left(1 - \frac{\lambda_F}{L_F + \lambda_F}\right)^T \|w^0 - w^*\|_2 + \frac{2}{\lambda_F}\triangle \tag{191}$$

这里

$$\triangle = \sqrt{2d}\left(1 + \frac{2k}{m+4}\right)s + \sqrt{2}\left(1 + \frac{2k}{m+4}\right)\hat{L}\delta \tag{192}$$

Theorem 6.2 如果假设

1. *对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot; z)$ 的第一个参数的第 k 个坐标（$k \in [d]$）的偏导数 $\partial_k f(\cdot; z)$ 是 L_k-Lipschitz 连续函数。并且，$f(\cdot; z)$ 是 L-smooth 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^d L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-smooth。*

2. *(sub-exponential gradients)。假设对于所有的 $k \in [d]$ 和 $w \in \mathcal{W}$，$f(w; z)$ 的关于 w 的坐标维度 k 的偏导 $\partial_k f(w; z)$ 是次指数分布。*

3. *(Size of $\mathcal{W}$)。参数空间 $\mathcal{W}$ 包含在下面的以 w^* : $\{w \in \mathbb{R}^d : \|w - w^*\|_2 \leq 2\|w^0 - w^*\|_2\}$ 为中心的 l_2 球里面。*

成立 $F(\cdot)$ 是凸函数。并且对于任意的 $\epsilon > 0$ 有 $\alpha \leq \beta \leq \frac{1}{4} - \epsilon$。选择步长 $\eta = 1/L_F$。那么至少有 $1 - \frac{4d}{(1+nm\hat{L}D)^d}$ 的概率，在 $T = \frac{L_F}{\triangle}\|w^0 - w^*\|_2$ 次迭代后，有

$$F(w^T) - F(w^*) \leq 16\|w^0 - w^*\|_2\triangle\left(1 + \frac{1}{2L_F}\triangle\right) \tag{193}$$

这里 $\triangle$ 定义为式 *192*

Theorem 6.3 如果假设

1. 对于任意的 $z \in \mathcal{Z}$，关于 $f(\cdot; z)$ 的第一个参数的第 k 个坐标（$k \in [d]$）的偏导数 $\partial_k f(\cdot; z)$ 是 L_k-*Lipschitz* 连续函数。并且，$f(\cdot; z)$ 是 *L-smooth* 函数。令 $\hat{L} := \sqrt{\sum_{k=1}^{d} L_k^2}$。同时也假设其分布函数 $F(\cdot)$ 也是 L_F-*smooth*。

2. (*sub-exponential gradients*)。假设对于所有的 $k \in [d]$ 和 $w \in \mathcal{W}$，$f(w; z)$ 的关于 w 的坐标维度 k 的偏导 $\partial_k f(w; z)$ 是次指数分布。

3. *(Size of $\mathcal{W}$)*。假设任意的 $\forall w \in \mathcal{W}, \|\triangle F(w)\|_2 \leq M$。假设 $\mathcal{W}$ 包含在下面的以 w^* : $\{w \in \mathbb{R}^d : \|w - w^0\|_2 \leq \frac{2}{\triangle^2}(M + \triangle)(F(w^0) - F(w^*))$ 为中心的 l_2 球里面。

成立。并且对于任意的 $\epsilon > 0$ 有 $\alpha \leq \beta \leq \frac{1}{4} - \epsilon$。选择步长 $\eta = 1/L_F$。那么至少有 $1 - \frac{4d}{(1+nm\tilde{L}D)^d}$ 的概率，在 $T = \frac{2L_F}{\triangle^2}(F(w^0) - F(w^*))$ 次迭代后，有

$$\min_{t=0,\ldots,T} \|\nabla F(w^t)\|_2 \leq \sqrt{2}\triangle \tag{194}$$

这里 $\triangle$ 定义为式 *192*

Theorem 6.4 令 $v_1, \ldots, v_n$ 是任意独立同分布的随机变量，即 $v_i \sim G$，并且 $\mathbb{E}[G] = g$ 和 $\mathbb{E}\|G - g\|^2 = d\sigma^2$。对于任意的 $j \in [d]$，$\{(v_1)_j, \ldots, (v_n)_j\}$ 中的 q 个值被替换成任意值，这里 $(v_i)_j$ 表示的是第 i 个向量的第 j 维的值。如果 $q < \frac{n}{4}$ 和 $\eta(n,q)\sqrt{d}\sigma < \|g\|$，这里 $\eta(n,q) = \sqrt{n-q}$，那么该算法是 *dimensional(α,q)-Byzantine resilient*，其中 $\sin\alpha = \frac{\eta(n,q)\sqrt{d}\sigma}{\|g\|}$。

$\tilde{O}$

7 safe guard SGD

7.1 ALGORITHM

7.2 CONVERGENCE RESULTS AND ANALYSIS

令 $\Xi = \sigma_t + \triangle_t$，这里

$$\begin{aligned} \sigma_t &= \frac{1}{|good_t|} \sum_{i \in good} (\nabla_{t,i} - \nabla f(w_t)) \\ \triangle_t &= \frac{1}{|good_t|} \sum_{i \in good \setminus good} (\nabla_{t,i} - \nabla f(w_t)) \end{aligned} \tag{195}$$

所以，扰动 SGD 可以写成期望 + 误差项（$good_t$ 中 *good* 的误差加非 *good* 中的误差）的形式：$w_{t+1} = w_t - \eta(\nabla f(w_t) + \xi_t + \Xi_t)$

Lemma 7.1 该算法中，假设选择 $\mathfrak{T} = 8\sqrt{T\log(16mT/p)}$。那么，至少有 $1 - p/4$ 的概率对于每一个 $t = 0, \ldots, T-1$，有

- $good_t \supseteq good$
- $\|\sigma_t\|^2 \leq O(\frac{\log(T/p)}{m})$ 和 $\|\sigma_0 + \cdots + \sigma_{t-1}\|^2 \leq O(\frac{T\log(T/p)}{m})$
- $\|\triangle_t\|^2 \leq \alpha^2$ 和 $\|\triangle_0 + \cdots + \triangle_{t-1}\|^2 \leq O(\alpha^2 T\log(mT/p))$
- $|\langle \nabla f(w_t), \xi_t \rangle| \leq \|\nabla f(w_t)\| \cdot O(v\sqrt{\log(T/p)})$
- $\|\xi_t\|^2 \leq O(v^2 d\log(T/p)), \|\xi_0 + \cdots + \xi_{t-1}\|^2 \leq O(v^2 dT\log(T/p))$

记 $\text{Event}_T^{\text{single}}(w_0)$

参考资料：（Pinelis' 不等式）令 $X_i, \dots, X_T \in \mathbb{R}^d$ 是一个随机过程，满足 $\mathbb{E}[X_t|X_1, \dots, X_{t-1}] = 0$ 以及 $\|X_t\| \le M$。那么

$$\Pr[\|X_1, \dots, X_T\|^2 > 2\log(2/\delta)M^2T] \le \delta \tag{196}$$

Proof 7.1 令

$$B_i^{(t)} = \frac{\nabla_{0,i}}{|good_0|} + \dots + \frac{\nabla_{t-1,i}}{|good_{t-1}|} \quad B_*^{(t)} = \frac{\nabla f(w_0)}{|good_0|} + \dots + \frac{\nabla f(w_{t-1})}{|good_{t-1}|} \tag{197}$$

令 $c = \log(16mT/p)$，下面的命题都至少有 $1 - \frac{p}{4}$ 概率发生

1. 对于所有的 $i \in good$ 和 $t \in [T]$，有 $\|B_i^{(t)} - B_*^{(t)}\| \le 4\sqrt{tC}/m$。

   证明：对于每个 $i \in good$，记 $\mathbb{E}[\nabla_{t,i}] = \nabla_t$ 并且 $\|\nabla_{t,i} - \nabla_i\| \le 1$（这里是一个常数，为便于分析所以设成 *1*）。令 $X_t = \frac{\nabla_{t,i} - \nabla_i}{|good_t|}$，它满足 $\|X_t\| \le \frac{1}{|good_t|} \le \frac{1}{(1-\alpha)m} \le \frac{2}{m}$。应用 *Pinelis'* 不等式，下式发生的概率至少为 $1 - \frac{p}{8mT}$

   $$\begin{aligned} \|X_0, \dots, X_{t-1}\|^2 &= \|\frac{\nabla_{0,i} - \nabla_0}{|good_0|} + \dots + \frac{\nabla_{t-1,i} - \nabla_{t-1}}{|good_{t-1}|}\|^2 \\ &= \|B_i^{(t)} - B_*^{(t)}\|^2 \le 2\log(2/p/8mT)\left(\frac{2}{m}\right)^2 t \end{aligned} \tag{198}$$

   再在 $i \in good$ 上使用 *Union bound*。有 $1 - \frac{p}{8Tm} \times m$ 的概率有

   $$\|B_i^{(t)} - B_*^{(t)}\| \le \sqrt{2\log(2/p/8mT)\left(\frac{2}{m}\right)^2 t} \le 4\sqrt{tC}/m \tag{199}$$

2. 对于所有的 $t \in [T]$，对于 $B_{med}^{(t)} = B_i^{(t)}$，每个 $i \in good$ 都是一个有效的选择。

   证明：由三角不等式再加上面的 *1* 可知

   $$\|B_i^{(t)} - B_j^{(t)}\| \le \|B_i^{(t)} - B_*^{(t)}\| + \|B_j^{(t)} - B_*^{(t)}\| \le 8\sqrt{tC}/m \tag{200}$$

   根据 $B_{med}^{(t)}$ 的定义，对于 $good_{t-1}$ 中的每一个 i 都满足至少有一半的 $j \in [m]$ 有 $\|B_j^{(t)} - B_i^{(t)}\| \le \mathfrak{T} = 8\sqrt{TC}$。可见 $i \in good$ 是满足的。

3. 对于所有的 $t \in [T]$ 和所有的 $i \in good$，有 $\|B_i^{(t)} - B_{med}^{(t)}\| \le 16\sqrt{tC}/m$ 以及 $\|B_*^{(t)} - B_{med}^{(t)}\| \le 12\sqrt{tC}/m$。

   证明：同样根据三角不等式

   $$\begin{aligned} \|B_i^{(t)} - B_{med}^{(t)}\| &\le \|B_i^{(t)} - B_j^{(t)}\| + \|B_j^{(t)} - B_{med}^{(t)}\| \le (8+8)\sqrt{tC}/m \\ \|B_*^{(t)} - B_{med}^{(t)}\| &\le \|B_i^{(t)} - B_*^{(t)}\| + \|B_i^{(t)} - B_{med}^{(t)}\| \le (4+8)\sqrt{tC}/m \end{aligned} \tag{201}$$

4. 显然，对于所有的 $i \in good$ 和所有的 $t \in [T]$，有 $i \in good_{t+1}$。

5. $\|\sum_{i \in good}(B_i^{(t)} - B_*^{(t)})\| \le O(\sqrt{t\log(T/p)}/\sqrt{m})$。

   证明：令 $\{X_1, X_2, \dots, X_{t\cdot|good|}\} = \{\frac{\nabla_{k,i} - \nabla f(w_k)}{|good_k|}\}_{k \in [t], i \in good}$，至少有 $1 - \frac{p}{8T}$ 的概率，使得下式成立。

   $$\|\sum_{i \in good}(B_i^{(t)} - B_*^{(t)})\| \le \sqrt{2\log(2/p/8T)\left(\frac{2}{m}\right)^2 |good|t} = O(\sqrt{t\log(T/p)}/\sqrt{m}) \tag{202}$$

下面证明引理 *7.1*,

- **(*good* 集合始终包含于 $good_t$) :** $good_t \supseteq good$

  证明: 由上面的 *4* 可知。

- **(每一轮聚合时由 *good* 引入的误差大小) :** $\|\sigma_t\|^2 \leq O(\frac{\log(T/p)}{m})$ 和 $\|\sigma_0 + \cdots + \sigma_{t-1}\|^2 \leq O(\frac{T\log(T/p)}{m})$

  证明:

$$\begin{aligned} \|\sigma_t\|^2 &= \| \sum_{i\in good} \frac{\nabla_{t,i} - \nabla f(w_t)}{|good_t|} \|^2 \\ &= \|X_1 + \cdots + X_{|good|}\|^2 \leq O(log(T/p) \left(\frac{2}{m}\right)^2 \cdot m) \\ &= O(log(T/p)/m) \end{aligned} \tag{203}$$

  另外由上面的 *5*

$$\begin{aligned} \|\sigma_0 + \cdots + \sigma_{t-1}\|^2 &= \| \sum_{t\in T} \sum_{i\in good} \frac{\nabla_{t,i} - \nabla f(w_t)}{|good_t|} \|^2 \\ &= \| \sum_{i\in good} (B_i^{(t)} - B_*^{(t)}) \|^2 \leq O(T\log(T/p)/m) \end{aligned} \tag{204}$$

- **(每一轮聚合时由 $good_t \backslash good$ 引入的误差大小):** $\|\triangle_t\|^2 \leq \alpha^2$ 和 $\|\triangle_0 + \cdots + \triangle_{t-1}\|^2 \leq O(\alpha^2 T\log(mT/p))$

  证明:

$$\|\triangle_t\|^2 = \| \sum_{i\in good_t\backslash good} \frac{\nabla_{t,i} - \nabla f(w_t)}{|good_t|} \|^2 \leq \| \frac{|good_t \backslash good|}{|good_t|} O(\mathfrak{T}) \|^2 \leq \alpha^2 \underbrace{O(\mathfrak{T}^2)}_{\text{不失一般性，把它看作 } 1 \text{ 了}} \tag{205}$$

  备注: 由于算法使用 $\mathfrak{T}$ 对数值进行了限制，且这个阈值是由 *good* 集合得出来的，因此通过这个过滤条件的梯度 $i \in good_t$ 应该都满足 $|\nabla f(w_t) - \nabla_{t,i}| < V = 1$。

$$\begin{aligned} \|\triangle_0 + \cdots + \triangle_{t-1}\|^2 &= \| \sum_{t\in[T]} \sum_{i\in good_t\backslash good} \frac{\nabla_{t-1,i} - \nabla f(w_{t-1})}{|good_{t-1}|} \|^2 \\ &= \| \sum_{i\in good_t\backslash good} (B_i^{(t)} - B_*^{(t)}) \|^2 \\ &= \alpha^2 m^2 \|B_i^{(t)} - B_*^{(t)})\|^2 \leq O(\frac{\alpha^2 m^2 TC}{m^2}) = O(\alpha^2 T\log(mT/p)) \end{aligned} \tag{206}$$

- $|\langle \nabla f(w_t), \xi_t \rangle| \leq \|\nabla f(w_t)\| \cdot O(v\sqrt{\log(T/p)})$

  证明:

$$\begin{aligned} |\langle \nabla f(w_t), \xi_t \rangle| &\leq \|\nabla f(w_t)\| \|\xi_t\| \\ &\leq \|\nabla f(w_t)\| \cdot O(v\sqrt{d\log(T/p)}) \end{aligned} \tag{207}$$

  备注: 不知在原文中是不是这里丢了一个 $\sqrt{d}$。

- **(每一轮聚合时由随机扰动引入的误差大小):** $\|\xi_t\|^2 \leq O(v^2 d\log(T/p)), \|\xi_0 + \cdots + \xi_{t-1}\|^2 \leq O(v^2 dT\log(T/p))$

  证明:

$$\|\xi_t\| \leq O(v^2 d\log(T/p)) \tag{208}$$

  后面一个

$$\|\xi_0 + \cdots + \xi_{t-1}\|^2 \leq O(v^2 dT\log(T/p)) \tag{209}$$

Lemma 7.2 假设选择引理 *7.1*中的 $\mathfrak{T}, C_1 = \log(T/p)$ 和 $C_2 = \alpha^2 \log(\frac{mT}{p}) + \frac{\log(T/p)}{m}$。假设 $\eta \le 0.01 \min\{1, \frac{1}{C_2}\}$, $T = \frac{1}{100\eta(1+\sqrt{C_2})}$，算法执行从 w_0 开始。如果事件 $\text{Event}_T^{\text{single}}(w_0)$ 成立，则有

$$f(w_0) - f(w_T) \ge 0.7\eta \sum_{t=0}^{T-1} \Big(\|\nabla f(w_t)\|^2 - \eta \cdot O(C_2 + (C_2)^{1.5}) - O(C_1 v^2 \eta (d + \sqrt{C_2})) \Big) \tag{210}$$

Proof 7.2 使用 *Lipschitz* 平滑假设，有

$$\begin{aligned} f(w_t) - f(w_{t+1}) &\ge \langle \nabla f(w_t), w_t - w_{t+1} \rangle - \frac{1}{2}\|w_t - w_{t+1}\|^2 \\ &= \langle \nabla f(w_t), w_t - (w_t - \eta(\nabla f(w_t) + \xi_t + \Xi_t)) \rangle - \frac{1}{2}\|w_t - w_{t+1}\|^2 \\ &= \eta \|f(w_t)\|^2 + \eta \underbrace{\langle \nabla f(w_t), \Xi \rangle}_{①} - \frac{1}{2} \underbrace{\|w_t - w_{t+1}\|^2}_{②} + \eta \underbrace{\langle \nabla f(w_t), \xi_t \rangle}_{③} \end{aligned} \tag{211}$$

对于 ②，使用 $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ 有

$$\|w_t - w_{t+1}\|^2 = \eta^2 \|\nabla f(w_t) + \Xi_t - \xi_t\|^2 \le 3\eta^2 (\|\nabla f(w_t)\|^2 + \|\Xi_t\|^2 - \|\xi_t\|^2) \tag{212}$$

对于 ③，下面会使用到它的和的形式，使用 $2ab \le a^2 + b^2$ 和 $a + b \ge \sqrt{a^2 + b^2}$ 有

$$\begin{aligned} |\sum_{t=0}^{T-1} \eta \langle \nabla f(w_t), \xi_t \rangle| &\le \sqrt{\Big| \sum_{t=0}^{T-1} \eta \|\nabla f(w_t)\| \cdot \|\xi_t\| \Big|^2} \le \sqrt{\eta} O(v\sqrt{C_1}) \cdot \sqrt{\eta} \sqrt{(\sum_{t=0}^{T-1} \|\nabla f(w_t)\|)^2} \\ &\le \frac{1}{2}\eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 + O(v^2 \eta C_1) \end{aligned} \tag{213}$$

备注： 这里的结果与答案不同，可通过 $\sqrt{0.1\eta} \cdot \sqrt{10\eta} \cdot O(v\sqrt{C_1})$ 来匹配，还不知是不是这个操作。进而，由迭代式有

$$\begin{aligned} f(w_0 - f(w^T)) \ge \eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 + \eta \sum_{t=0}^{T-1} \langle \nabla f(w_t), \Xi \rangle \\ - \frac{1}{2} \sum_{t=0}^{T-1} \|w_t - w_{t+1}\|^2 + \eta \sum_{t=0}^{T-1} \langle \nabla f(w_t), \xi_t \rangle \end{aligned} \tag{214}$$

把 ①, ②, ③ 代入上式有

$$\begin{aligned} f(w_0 - f(w^T)) &\ge \eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 + \eta \sum_{t=0}^{T-1} \langle \nabla f(w_t), \Xi_t \rangle - \frac{3}{2}\eta^2 \sum_{t=0}^{T-1} (\|\nabla f(w_t)\|^2 + \|\Xi_t\|^2 - \|\xi_t\|^2) \\ &\quad + \frac{1}{2}\eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 + O(v^2 \eta C_1) \\ &= (1 - \frac{3}{2}\eta + \frac{1}{2})\eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 + \eta \sum_{t=0}^{T-1} \langle \nabla f(w_t), \Xi_t \rangle \\ &\quad - \frac{3}{2}\eta^2 \sum_{t=0}^{T-1} \|\Xi_t\|^2 + \frac{3}{2}\eta^2 \sum_{t=0}^{T-1} \|\xi_t\|^2 + O(v^2 \eta C_1) \\ &\ge (1.5 - \frac{3}{2}\eta)\eta \sum_{t=0}^{T-1} (\|\nabla f(w_t)\|^2 - O(\eta C_2)) + \underbrace{\eta \sum_{t=0}^{T-1} \langle \nabla f(w_t), \Xi_t \rangle}_{④} - O(\eta T v^2 C_1 (\eta d + \frac{1}{T}) \end{aligned} \tag{215}$$

下面工作就是要求出 ④ 的界限。

$$\begin{aligned}\eta\sum_{t=0}^{T-1}\langle\nabla f(w_t),\Xi_t\rangle &= \frac{\eta}{T}\sum_{q=0}^{T-1}\sum_{t=0}^{T-1}\langle\nabla f(w_t)+\nabla f(w_q)-\nabla f(w_q),\Xi_t\rangle \\ &= \underbrace{\frac{\eta}{T}\sum_{q=0}^{T-1}\langle\nabla f(w_q),\sum_{t=0}^{T-1}\Xi_t\rangle}_{\clubsuit}+\underbrace{\frac{\eta}{T}\sum_{q=0}^{T-1}\sum_{t=0}^{T-1}\langle\nabla f(w_t)-\nabla f(w_q),\Xi_t\rangle}_{\spadesuit}\end{aligned} \tag{216}$$

对于 ♣，有

$$\begin{aligned}|\clubsuit| &\le \frac{\eta}{T}\sum_{q=0}^{T-1}\left|\langle\nabla f(w_q),\sum_{t=0}^{T-1}\Xi_t\rangle\right| \le \frac{\eta}{T}\sum_{q=0}^{T-1}\|\nabla f(w_q)\|\cdot\|\sum_{t=0}^{T-1}\Xi_t\| \\ &= \sum_{q=0}^{T-1}[\sqrt{\eta}\|\nabla f(w_q)\|]\cdot\left[\frac{\sqrt{\eta}}{T}\|\sum_{t=0}^{T-1}\Xi_t\|\right] \\ &\le \frac{\eta}{2}\sum_{q=0}^{T-1}\|\nabla f(w_q)\|^2+\frac{O(\eta)}{T^2}\sum_{q=0}^{T-1}\|\sum_{t=0}^{T-1}\Xi_t\|^2 \\ &\le \frac{\eta}{2}\sum_{q=0}^{T-1}\|\nabla f(w_q)\|^2+O(\eta C_2)\end{aligned} \tag{217}$$

对于 ♠，首先根据泰勒展开有

$$\begin{aligned}\nabla f(w_t) &\simeq \nabla f(w_0)+\nabla^2 f(w_0)(w_t-w_0)-\frac{1}{2}\|w_t-w_0\|^2 \\ \nabla f(w_q) &\simeq \nabla f(w_0)+\nabla^2 f(w_0)(w_q-w_0)-\frac{1}{2}\|w_q-w_0\|^2\end{aligned} \tag{218}$$

代入

$$\begin{aligned}|\spadesuit| &\le \frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle\nabla f(w_t)-\nabla f(w_q),\Xi_t\rangle\right| \\ &\le \frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle\nabla^2 f(w_0)(w_t-w_q)-\frac{1}{2}[(w_t-w_0)^2+(w_q-w_0)^2],\Xi_t\rangle\right| \\ &\le \frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle\nabla^2 f(w_0)(w_t-w_q),\Xi_t\rangle\right|+\frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle[(w_t-w_0)+(w_q-w_0)]^2,\Xi_t\rangle\right| \\ &\le \underbrace{\frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle\nabla^2 f(w_0)(w_t-w_q),\Xi_t\rangle\right|}_{\diamondsuit}+\underbrace{\frac{\eta}{T}\sum_{q=0}^{T-1}\sum_{t=0}^{T-1}(\|w_t-w_0\|+\|w_q-w_0\|)^2\|\Xi_t\|}_{\heartsuit}\end{aligned} \tag{219}$$

对于 $\heartsuit$，有

$$\begin{aligned}
\heartsuit &\leq \frac{\eta}{T}\sum_{q=0}^{T-1}\sum_{t=0}^{T-1}2(\|w_t-w_0\|^2+\|w_q-w_0\|^2)\|\Xi_t\| \\
&\leq \frac{\eta}{T}\sum_{q=0}^{T-1}\sum_{t=0}^{T-1}(\|w_t-w_0\|^2+\|w_q-w_0\|^2)\cdot O(\sqrt{C_2}) \\
&\leq \frac{\eta}{T}\sum_{t=0}^{T-1}T\|w_t-w_0\|^2\cdot O(\sqrt{C_2}) \\
&\leq \eta^3\sum_{t=0}^{T-1}\|\nabla f(w_0)+\cdots+\nabla f(w_{t-1})+\Xi_0+\cdots+\Xi_{t-1}+\xi_0+\cdots+\xi_{t-1}\|^2\cdot O(\sqrt{C_2}) \\
&\leq \eta^3\left[\sum_{t=0}^{T-1}\|\sum_{t=0}^{T-1}\nabla f(w_t)\|^2+\sum_{t=0}^{T-1}\|\sum_{t=0}^{T-1}\Xi_t\|^2+\sum_{t=0}^{T-1}\|\sum_{t=0}^{T-1}\xi_t\|^2\right]\cdot O(\sqrt{C_2}) \\
&\leq O(\sqrt{C_2}\eta^3)T\cdot(T\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2)+T\cdot O(\sqrt{C_2}C_2\eta^3T)+T\cdot O(\eta^3v^2TdC_1\sqrt{C_2}) \\
&\leq O(\sqrt{C_2}\eta^3T^2)\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2+O(\sqrt{C_2}C_2\eta^3T^2)+O(\eta^3v^2T^2dC_1\sqrt{C_2})
\end{aligned} \tag{220}$$

对于 $\diamondsuit$，从 $t=q$ 处断开为两部分有

$$\left|\sum_{t=0}^{T-1}\langle\nabla^2f(w_0)(w_t-w_q),\Xi_t\rangle\right|\leq\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(w_t-w_q),\Xi_t\rangle\right|+\left|\sum_{t=0}^{q-1}\langle\nabla^2f(w_0)(w_t-w_q),\Xi_t\rangle\right| \tag{221}$$

对于右式的第一部分，第二部分也同样处理

$$\begin{aligned}
&\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(w_t-w_q),\Xi_t\rangle\right| \\
&=\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(\nabla f(w_q)+\cdots+\nabla f(w_{t-1})+\Xi_q+\cdots+\Xi_{t-1}+\xi_q+\cdots+\xi_{t-1}),\Xi_t\rangle\right| \\
&\leq\underbrace{\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(\nabla f(w_q)+\cdots+\nabla f(w_{t-1}),\Xi_t\rangle\right|}_{⑤} \\
&+\underbrace{\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(\Xi_q+\cdots+\Xi_{t-1}),\Xi_t\rangle\right|}_{⑥}+\underbrace{\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(\xi_q+\cdots+\xi_{t-1}),\Xi_t\rangle\right|}_{⑦}
\end{aligned} \tag{222}$$

对于 ⑤,t 在变 q 没变，$\nabla^2f(w_0)$ 是一个常数 (把它看作 *1*)，拆开看是每一个 Ξ_t 在与 $\nabla f_(w_q)+\cdots+f(w_t)$ 做内积，等价于每个 $\nabla f(w_t)$ 与 $\Xi_t+\cdots+\Xi_{T-1}$ 做内积(仅仅是调整了顺序，这样做的好处是 $\Xi_t+\cdots+\Xi_{T-1}$ 我们可以计算出界限)。所以有

$$\begin{aligned}
\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2f(w_0)(\nabla f(w_q)+\cdots+\nabla f(w_{t-1}),\Xi_t\rangle\right| &= \eta\left|\sum_{t=q}^{T-1}\langle\nabla^2f(w_0)\nabla f(w_t),\Xi_{t+1}+\cdots+\Xi_{T-1}\rangle\right| \\
&\leq \eta\sum_{t=q}^{T-2}\|\nabla f(w_t)\|\cdot\|\Xi_{t+1}+\cdots+\Xi_{T-1}\| \\
&\leq O(\eta\sqrt{TC_2})\sum_{t=q}^{T-2}\|\nabla f(w_t)\|
\end{aligned} \tag{223}$$

对于 ⑥，有

$$\begin{aligned}\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2 f(w_0)(\Xi_q+\cdots+\Xi_{t-1}),\Xi_t\rangle\right| &\le \eta\langle\nabla^2 f(w_0)(\Xi_0+\cdots+\Xi_{T-1}),\Xi_0+\cdots+\Xi_{T-1}\rangle\\ &\le \eta\nabla^2 f(w_0)\|\Xi_0+\cdots+\Xi_{T-1}\|\cdot\|\Xi_0+\cdots+\Xi_{T-1}\|\\ &\le O(\eta TC_2)\end{aligned} \tag{224}$$

对于 ⑦，有

$$\begin{aligned}\eta\left|\sum_{t=q+1}^{T-1}\langle\nabla^2 f(w_0)(\xi_q+\cdots+\xi_{t-1}),\Xi_t\rangle\right| &\le \eta\langle\nabla^2 f(w_0)(\xi_0+\cdots+\xi_{T-1}),\Xi_0+\cdots+\Xi_{T-1}\rangle\\ &\le \eta\nabla^2 f(w_0)\|\xi_0+\cdots+\xi_{T-1}\|\cdot\|\Xi_0+\cdots+\Xi_{T-1}\|\\ &\le \cdot O(\eta\sqrt{dv^2TC_1}\cdot\sqrt{TC_2})\end{aligned} \tag{225}$$

将 ⑤,⑥,⑦ 的化简结果带入，有

$$\left|\sum_{t=q+1}^{T-1}\langle\nabla^2 f(w_0)(w_t-w_q),\Xi_t\rangle\right| \le O(\eta\sqrt{TC_2})\sum_{t=0}^{T-1}\|\nabla f(w_t)\|+O(\eta TC_2+\eta dv^2TC_1) \tag{226}$$

这里对 $O(\sqrt{\eta dv^2TC_1}\cdot\sqrt{\eta TC_2})$ 使用了 *Young's* 不等式，得到 $\le O(\eta dv^2TC_1+\eta TC_2)$。

进而得到 ♢

$$\begin{aligned}\frac{\eta}{T}\sum_{q=0}^{T-1}\left|\sum_{t=0}^{T-1}\langle\nabla^2 f(w_0)(w_t-w_q),\Xi_t\rangle\right| &\le \frac{\eta}{T}\sum_{q=0}^{T-1}\left\{O(\eta\sqrt{TC_2})\sum_{t=0}^{T-2}\|\nabla f(w_t)\|+O(\eta TC_2+\eta dv^2TC_1)\right\}\\ &= O(\eta^2\sqrt{TC_2})\sum_{t=0}^{T-1}\|\nabla f(w_t)\|+O(\eta^2TC_2+\eta^2dv^2TC_1)\\ &\le \eta\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2+O(\eta^3T^2C_2+\eta^2TC_2+\eta^2dv^2TC_1)\end{aligned} \tag{227}$$

这里最后一步把 $\eta^2\sqrt{TC_2}$ 分成两部分 $\frac{\eta^{\frac{1}{2}}}{\sqrt{T}}\cdot\sqrt{T}\eta^{\frac{3}{2}}\sqrt{TC_2}$，再使用了 $ab\le\frac{1}{2}(a^2+b^2)$ 和 $(a+b+c)^2\le 3(a^2+b^2+c^2)$。

最后，把 ♢,♡ 带回 ♠，然后联合 ♣ 带回 *216*中，最后得到整个结果

$$\begin{aligned}f(w_0-f(w^T)) \ge (1.5-\frac{3}{2}\eta+\frac{1}{2}+1)\eta\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2+O(\sqrt{C_2}\eta^3T^2)\sum_{t=0}^{T-1}\|\nabla f(w_t)\|^2\\ -\underbrace{C_2\cdot O(\eta+\eta^2T+\eta^3T^2+\sqrt{C_2}\eta^3T^2)}_{⑧}-\underbrace{C_1\cdot O(T\eta^2v^2d+T^2\eta^3v^2\sqrt{C_2}+\eta Tv^2(\eta d+\frac{1}{T}))}_{⑨}\end{aligned} \tag{228}$$

取 $T=\frac{1}{100\eta(1+\sqrt{C_2})}$ 和 $\eta\le 0.01\min\{1,\frac{1}{C_2}\}$，对 ⑧ 只保留主要的影响因子，有

$$\begin{aligned}C_2\cdot O(\eta+\eta^2T+\eta^3T^2+\sqrt{C_2}\eta^3T^2) &\le \eta T(\frac{1}{T}+\eta+\eta^2T+\sqrt{C_2}\eta^2T)\\ &= \eta T\left(100\eta(1+\sqrt{C_2})+\eta+\eta^2\frac{1}{100\eta(1+\sqrt{C_2})}+\sqrt{C_2}\eta^2\frac{1}{100\eta(1+\sqrt{C_2})}\right)\\ &\le \eta T\eta C_2\left(100(1+\sqrt{C_2})+\frac{101}{100}\right)=\eta T\eta\cdot O(C_2+(C_2)^{1.5})\end{aligned} \tag{229}$$

对 ⑨ 有

$$\begin{aligned}C_1O(T\eta^2v^2d+T^2\eta^3v^2\sqrt{C_2}+\eta Tv^2(\eta d+\frac{1}{T})) &\le C_1O(T\eta^2v^2d\\ &+\frac{1}{100^2\eta^2(1+\sqrt{C_2})^2}\eta^3v^2\sqrt{C_2}+\eta Tv^2(\eta d+(100\eta(1+\sqrt{C_2}))))\\ &= C_1\cdot O(\eta Tv^2\eta(d+\sqrt{C_2}))\end{aligned} \tag{230}$$

因此,

$$
\begin{aligned}
f(w_0 - f(w^T)) &\geq 0.7\eta \sum_{t=0}^{T-1} \|\nabla f(w_t)\|^2 - C_2 O(\eta + \eta^2 T + \eta^3 T^2 + \sqrt{C_2}\eta^3 T^2) \\
&\quad - C_1 O(T\eta^2 v^2 d + T^2 \eta^3 v^2 \sqrt{C_2} + \eta T v^2 (\eta d + \frac{1}{T})) \\
&\geq 0.7\eta \sum_{t=0}^{T-1} \Big(\|\nabla f(w_t)\|^2 - \eta \cdot O(C_2 + (C_2)^{1.5}) - O(C_1 v^2 \eta (d + \sqrt{C_2})) \Big)
\end{aligned} \tag{231}
$$