

# Stochastic Optimization

Yajie Bao\*

February 13, 2020

For random sample  $X_i, i = 1, 2, \dots, n$  and lost function  $f(X, \beta)$ , denote the real estimate of  $\beta$  by  $\beta^*$ , which can be obtained by

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}f(X, \beta).$$

And denote the estimate of  $\beta$  based on sample by  $\hat{\beta}$ , which can be obtained by

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f(X_i, \beta).$$

**Lemma 0.1** *For convex function  $f(x)$ ,  $x^*$  is the global minimizer of  $f(x)$ . If for any  $x \in \{x : |x - \tilde{x}|^2 = a\}$ , s.t.,  $f(x) \geq f(\tilde{x})$ , then*

$$|x^* - \tilde{x}| \leq a.$$

**Proof:** If there exists  $x'$  such that  $|x' - \tilde{x}|^2 > a$  and  $f(x') \leq f(x^*)$ . By the convexity of  $f$ , we have

$$f(\alpha x' + (1 - \alpha)\tilde{x}) \leq \alpha f(x') + (1 - \alpha)f(\tilde{x}) < f(\tilde{x}),$$

where  $0 < \alpha < 1$ . Note that

$$|\alpha x' + (1 - \alpha)\tilde{x} - \tilde{x}| = \alpha |x' - \tilde{x}|,$$

let  $\alpha = |x' - \tilde{x}|/|x^* - \tilde{x}|$ , then  $|\alpha x' + (1 - \alpha)\tilde{x} - \tilde{x}| = a$ . But

$$f(\alpha x' + (1 - \alpha)\tilde{x}) < f(\tilde{x}),$$

which is a contradiction. ■

**Theorem 0.2** *If  $|\beta - \beta^*| = O_p(\frac{1}{\sqrt{n}})$  and  $f(x)$  is  $\mu$ -strongly convex, then  $|\hat{\beta} - \beta^*| = O_p(\frac{1}{\sqrt{n}})$ .*

---

\*Department of Mathematics, Shanghai Jiao Tong University, Email: baoyajie2019stat@sjtu.edu.cn

**Proof:** Perform Taylor expansion on  $f(X, \beta)$  with respect to  $\beta$ ,

$$f(X, \beta) = f(X, \beta^*) + \partial f(X, \beta^*)^T (\beta - \beta^*) + \frac{1}{2} (\beta - \beta^*)^T \partial^2 f(X, \tilde{\beta}) (\beta - \beta^*),$$

where  $|\tilde{\beta} - \beta^*| \leq |\beta - \beta^*|$ . Use the fact that  $E \partial f(X, \beta^*) = 0$ , we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n f(X_i, \beta) &= \frac{1}{n} \sum_{i=1}^n f(X_i, \beta^*) + \frac{1}{n} \sum_{i=1}^n [\partial f(X_i, \beta^*) - E \partial f(X_i, \beta^*)]^T (\beta - \beta^*) \\ &\quad + \frac{1}{2n} (\beta - \beta^*)^T \left( \sum_{i=1}^n \partial^2 f(X_i, \tilde{\beta}) \right) (\beta - \beta^*) \\ &\geq \frac{1}{n} \sum_{i=1}^n f(X_i, \beta^*) + O_p\left(\frac{1}{\sqrt{n}}\right) \frac{C}{\sqrt{n}} + \frac{\mu C^2}{n}, \end{aligned}$$

where  $C$  is a constant (positive or negative). If we set  $|C|$  sufficient large, then

$$\frac{1}{n} \sum_{i=1}^n f(X_i, \beta) \geq \frac{1}{n} \sum_{i=1}^n f(X_i, \beta^*).$$

Then the result follows from the Lemma [0.1](#). ■

## 1 Stochastic gradient descent

The optimization problem is

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}; \boldsymbol{\xi})],$$

where  $\boldsymbol{\xi}$  is random and  $f(\mathbf{x}; \boldsymbol{\xi})$  is convex for every  $\boldsymbol{\xi}$ . The SGD iteration step is

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t),$$

where  $\mathbf{g}(\mathbf{x}^t)$  is an unbiased estimate of  $\nabla F(\mathbf{x}^t)$ , i.e.,  $\mathbb{E}[\mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t)] = \nabla F(\mathbf{x}^t)$ . The empirical risk minimization is

$$\min_{\mathbf{x}} F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}; \boldsymbol{\xi}_i),$$

for  $t = 0, 1, \dots$  choose  $i_t$  uniformly at random then update  $\mathbf{x}^t$  by

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \nabla_{\mathbf{x}} f(\mathbf{x}^t; \boldsymbol{\xi}_{i_t}).$$

**Assumption 1.1** Given  $\{\boldsymbol{\xi}^0, \dots, \boldsymbol{\xi}^{t-1}\}$ ,  $\mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t)$  is an unbiased estimate of  $\nabla F(\mathbf{x}^t)$ ; And for all  $\mathbf{x}$ , we have  $\mathbb{E}[\|\mathbf{g}(\mathbf{x}; \boldsymbol{\xi})\|_2^2] \leq \sigma_g^2$ .

**Theorem 1.2** Under Assumption 1.1, suppose  $F$  is  $\mu$ -strongly convex and the above assumptions are satisfied. If  $\eta_t = \frac{\theta}{t+1}$  for some  $\theta > \frac{1}{2\mu}$ , then SGD achieves

$$\mathbb{E} \left[ \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 \right] \leq \frac{c_\theta}{t+1}$$

where  $c_\theta = \max \left\{ \frac{2\theta^2\sigma_g^2}{2\mu\theta-1}, \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \right\}$ .

**Proof:** Using SGD update rule, we have

$$\begin{aligned} \|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}^t - \eta_t \mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t) - \mathbf{x}^*\|_2^2 \\ &= \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - 2\eta_t (\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t) + \eta_t^2 \|\mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t)\|_2^2 \end{aligned}$$

Since  $\mathbf{x}^t$  depend on  $\xi_1, \dots, \xi_{t-1}$ ,

$$\begin{aligned} \mathbb{E} \left[ (\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t) \right] &= \mathbb{E} \left[ \mathbb{E} \left[ (\mathbf{x}^t - \mathbf{x}^*)^\top \mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t) \mid \xi_1, \dots, \xi_{t-1} \right] \right] \\ &= \mathbb{E} \left[ (\mathbf{x}^t - \mathbf{x}^*)^\top \mathbb{E} [\mathbf{g}(\mathbf{x}^t; \boldsymbol{\xi}^t) \mid \xi_1, \dots, \xi_{t-1}] \right] \\ &= \mathbb{E} \left[ (\mathbf{x}^t - \mathbf{x}^*)^\top \nabla F(\mathbf{x}^t) \right] \end{aligned}$$

Furthermore, strong convexity gives

$$\langle \nabla F(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle = \langle \nabla F(\mathbf{x}^t) - \nabla F(\mathbf{x}^*), \mathbf{x}^t - \mathbf{x}^* \rangle \geq \mu \|\mathbf{x}^t - \mathbf{x}^*\|_2^2,$$

which implies

$$\mathbb{E} [\langle \nabla F(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle] \geq \mu \mathbb{E} [\|\mathbf{x}^t - \mathbf{x}^*\|_2^2].$$

Combine the above results to obtain,

$$\mathbb{E} [\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2] \leq (1 - 2\mu\eta_t) \mathbb{E} [\|\mathbf{x}^t - \mathbf{x}^*\|_2^2] + \eta_t^2 \sigma_g^2.$$

First for  $t = 0$ ,

$$\mathbb{E} [\|\mathbf{x}^1 - \mathbf{x}^*\|_2^2] \leq (1 - 2\mu\theta) \mathbb{E} [\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2] + \theta^2 \sigma_g^2 \leq c_\theta.$$

Assume the result holds for some  $k \leq 1$ , it follows that

$$\begin{aligned} \mathbb{E} [\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2] &\leq \frac{t+1-2\mu\theta}{(t+1)^2} c_\theta + \frac{\theta^2 \sigma_g^2}{(t+1)^2} \\ &= \frac{t}{(t+1)^2} c_\theta - \frac{2\mu\theta-1}{(t+1)^2} c_\theta + \frac{\theta^2 \sigma_g^2}{(t+1)^2} \\ &\leq \frac{t}{(t+1)^2} c_\theta - \frac{\theta^2 \sigma_g^2}{2(t+1)^2} \\ &\leq \frac{1}{(t+2)^2} c_\theta. \end{aligned}$$

Thus, the main result holds for every  $k \geq 1$ . ■

## 2 Stochastic variance reduced gradient

Consider the following optimization problem

$$\min P(w), \quad P(w) := \frac{1}{n} \sum_{i=1}^n \psi_i(w) \quad (2.1)$$

Because of the variance of SGD, we need a small learning rate, which leads a slower convergence. To fix this, SVRG proposed by [Johnson and Zhang \[2013\]](#) keep a snapshot of the estimator  $\tilde{w}$  after every  $m$  SGD iterations. Moreover, we maintain the average gradient

$$\tilde{\mu} = \nabla P(\tilde{w}) = \frac{1}{n} \sum_{i=1}^n \nabla \psi_i(\tilde{w}),$$

and the expectation of  $\nabla \psi_i(\tilde{w}) - \tilde{\mu}$  over  $i$  is 0. And thus the following update rule is generalized SGD: randomly draw  $i_t$  from  $\{1, 2, \dots, n\}$ :

$$w^{(t)} = w^{(t-1)} - \eta_t \left( \nabla \psi_{i_t}(w^{(t-1)}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu} \right). \quad (2.2)$$

And we have

$$\mathbb{E} [w^{(t)} | w^{(t-1)}] = w^{(t-1)} - \eta_t \nabla P(w^{(t-1)}).$$

The SVRG is presented in Algorithm 1.

---

### Algorithm 1: Stochastic Variance Reduced Gradient

---

**Input** : Data  $\{X_i, i = 1, 2, \dots, n\}$ , the number of iterations  $L$ , update frequency  $m$  and learning rate  $\eta$

**Output**: The estimator  $\hat{w}_L$

```

1 Compute the initial estimator  $\tilde{w}_0$ 
2 for  $s = 1, 2, \dots, L$  do
3   Set  $\tilde{w} = \tilde{w}_{s-1}$ 
4   Compute the average gradient  $\tilde{\mu} = \frac{1}{n} \sum_{i=1}^n \nabla \psi_i(\tilde{w})$ 
5   Set  $w_0 = \tilde{w}$ 
6   for  $t = 1, 2, \dots, m$  do
7     Randomly pick  $i_t$  from  $\{1, 2, \dots, n\}$  and update estiamte

```

$$w_t = w_{t-1} - \eta \left( \nabla \psi_{i_t}(w_{t-1}) - \nabla \psi_{i_t}(\tilde{w}) + \tilde{\mu} \right)$$

```

8   end
9   Option 1: Set  $\tilde{w}_s = w_m$ 
10  Option 2: Set  $\tilde{w}_s = w_t$  for randomly chosen  $t \in \{1, 2, \dots, n\}$ 
11 end

```

---

**Theorem 2.1** *For the optimization problem (2.1), consider SVRG in Algorithm 1 with option 2. Assume that all  $\psi_i(x)$  are convex and  $L$ -smooth and  $P(w)$  is  $\lambda$ -strongly convex. Let  $w_* = \arg \min_w P(w)$ . Assume that  $m$  is sufficiently large so that*

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$

*then we have geometric convergence in expectation for SVRG:*

$$\mathbb{E}P(\tilde{w}_s) \leq \mathbb{E}P(w_*) + \alpha^s [P(\tilde{w}_0) - P(w_*)].$$

### 3 Distributed SGD

Let  $L_n$  be the global loss function, and expand it to an infinite series:

$$\mathcal{L}_N(\theta) = \mathcal{L}_N(\bar{\theta}) + \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_N(\bar{\theta}) (\theta - \bar{\theta})^{\otimes j}, \quad (3.1)$$

where  $\bar{\theta}$  is the initial estimator of  $\theta$ . Because the data is split across machines, evaluating the derivatives  $\nabla^j \mathcal{L}_N(\bar{\theta})$ ,  $j \geq 1$  requires one communication round. However the higher-order derivatives ( $j \geq 2$ ) require communicating more than  $O(d^2)$  bits from each machine. This reasoning motivates [Jordan, Lee, and Yang \[2019\]](#) to replace the global higher-order derivatives  $\nabla^j \mathcal{L}_N(\bar{\theta})$ ,  $j \geq 2$  with the local derivatives, leading to the following approximation of  $\mathcal{L}_N(\theta)$

$$\tilde{\mathcal{L}}(\theta) = \mathcal{L}_N(\bar{\theta}) + \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_1(\bar{\theta}) (\theta - \bar{\theta})^{\otimes j}. \quad (3.2)$$

Comparing (3.1) and (3.2), using the fact  $\|\nabla^2 \mathcal{L}_N(\bar{\theta}) - \nabla^2 \mathcal{L}_1(\bar{\theta})\|_2 = O(n^{-1/2})$ , we can obtain the approximation error

$$\begin{aligned} \tilde{\mathcal{L}}(\theta) - \mathcal{L}_N(\theta) &= \mathcal{L}_N(\bar{\theta}) + \langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \bar{\theta} \rangle + \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_1(\bar{\theta}) (\theta - \bar{\theta})^{\otimes j} \\ &\quad - \left( \mathcal{L}_N(\bar{\theta}) + \left\langle \nabla \mathcal{L}_N(\bar{\theta}), \theta - \sum_{j=2}^{\infty} \frac{1}{j!} \nabla^j \mathcal{L}_N(\bar{\theta}) (\theta - \bar{\theta})^{\otimes j} \right\rangle \right) \\ &= \frac{1}{2} \langle \theta - \bar{\theta}, (\nabla^2 \mathcal{L}_1(\bar{\theta}) - \nabla^2 \mathcal{L}_N(\bar{\theta})) (\theta - \bar{\theta}) \rangle + O(\|\theta - \bar{\theta}\|_2^3) \\ &= O\left(\frac{1}{\sqrt{n}} \|\theta - \bar{\theta}\|_2^2 + \|\theta - \bar{\theta}\|_2^3\right). \end{aligned}$$

Now we replace the infinite sum of high-order derivatives in expression (3.1) with  $\mathcal{L}_1(\theta) - \mathcal{L}_1(\bar{\theta}) - \langle \nabla \mathcal{L}_1(\bar{\theta}), \theta - \bar{\theta} \rangle$  and omit the additive constants, which yields

$$\tilde{\mathcal{L}}(\theta) := \mathcal{L}_1(\theta) - \langle \theta, \nabla \mathcal{L}_1(\bar{\theta}) - \nabla \mathcal{L}_N(\bar{\theta}) \rangle. \quad (3.3)$$

We define the distributed SGD method as Algorithm 2, and the error of first iteration is given by Theorem 3.1.

---

**Algorithm 2:** Distributed Stochastic Gradient Descent

---

**Input** : Data on local machines  $\{X_i, i \in H_k\}$  for  $k = 1, 2, \dots, N$ , the number of iterations  $L$

**Output**: The final median estimate  $\hat{\beta}_t$

1 Compute the initial estimator  $\hat{\beta}_0$  based on  $\{X_i, i \in H_1\}$  such that

$$|\hat{\beta}_0 - \beta_*| = O_p\left(\frac{1}{\sqrt{m}}\right)$$

2 **for**  $t = 1, 2, \dots, L$  **do**

3     Transmit  $\hat{\beta}_{t-1}$  to all local machines.

4     **for**  $i = 1, 2, \dots, N$  **do**

5         In each machine, compute the gradient of empirical lost function

$$\hat{g}_i(\hat{\beta}_{t-1}) = \frac{1}{m} \sum_{k \in H_i} \partial f(X_k, \hat{\beta}_{t-1})$$

6     **end**

7     Compute the pooled gradient:

$$\hat{g}(\hat{\beta}_{t-1}) = \frac{1}{N} \sum_{i=1}^N \hat{g}_i(\hat{\beta}_{t-1})$$

8     Compute update of  $\beta$  based on  $\{X_i, i \in H_1\}$ :

$$\hat{\beta}_t = \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{m} \sum_{k \in H_1} f(X_k, \beta) - \beta^T [\hat{g}_1(\hat{\beta}_{t-1}) - \hat{g}(\hat{\beta}_{t-1})] \right\}$$

9 **end**

---

**Theorem 3.1** Assume that for all  $\beta, \beta' \in \mathbb{R}^p$

$$\|\partial f(X, \beta) - \partial f(X, \beta')\|_2 \leq M(X) \|\beta - \beta'\|_2$$

and  $E(M^2(X)) \leq M$ . The error of first update  $\hat{\beta}_1$  is

$$|\hat{\beta}_1 - \beta_*| = O_p\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{\log n}}{m}\right). \quad (3.4)$$

**Proof:** First we will prove for any  $\beta$  satisfying  $|\beta - \beta_*| = O_p\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{\log n}}{m}\right)$ ,

$$h(\beta) > h(\beta_*),$$

where  $h(\beta) = \frac{1}{m} \sum_{k \in H_1} f(X_k, \beta) - \beta^T (\hat{g}_1(\hat{\beta}_{t-1})) - \hat{g}(\hat{\beta}_{t-1})$ . Taking Taylor expansion on  $h(\beta)$  we have

$$\begin{aligned} h(\beta) &= h(\beta^*) + \left[ \hat{g}_1(\beta^*) - \hat{g}_1(\hat{\beta}_0) + \hat{g}(\hat{\beta}_0) \right]^T (\beta - \beta^*) \\ &\quad + \frac{1}{2} (\beta - \beta^*)^T \left( \frac{1}{m} \sum_{k \in H_1} \partial^2 f(X_k, \tilde{\beta}) \right) (\beta - \beta^*). \end{aligned}$$

Let  $G(\beta) = E(\partial f(X, \beta))$ , then we consider the case of 1-dimension. Note that

$$\begin{aligned} \hat{g}(\hat{\beta}_0) &= \frac{1}{n} \sum_{i=1}^n \partial f(X_i, \hat{\beta}_0) - G(\hat{\beta}_0) - \left[ \frac{1}{n} \sum_{i=1}^n \partial f(X_i, \beta^*) - G(\beta^*) \right] \\ &\quad + \left[ \frac{1}{n} \sum_{i=1}^n \partial f(X_i, \beta^*) - G(\beta^*) \right] + G(\hat{\beta}_0) \\ &\leq \sup_{|\beta - \beta^*| = \frac{1}{\sqrt{m}}} \frac{1}{n} \sum_{i=1}^n \partial f(X_i, \beta) - G(\beta) - \left[ \frac{1}{n} \sum_{i=1}^n \partial f(X_i, \beta^*) - G(\beta^*) \right] + O_p\left(\frac{1}{\sqrt{n}}\right) + G(\hat{\beta}_0) \\ &= O_p\left(\frac{1}{\sqrt{m}} \frac{\log n}{\sqrt{n}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right) + G(\hat{\beta}_0), \end{aligned}$$

thus we have  $\hat{g}(\hat{\beta}_0) - G(\hat{\beta}_0) = O_p\left(\frac{1}{\sqrt{n}}\right)$ . Then using the fact that  $G(\beta^*) = 0$  we can obtain

$$\begin{aligned} \hat{g}_1(\beta^*) - \hat{g}_1(\hat{\beta}_0) + \hat{g}(\hat{\beta}_0) &= \hat{g}_1(\beta^*) - \hat{g}_1(\hat{\beta}_0) + G(\hat{\beta}_0) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \hat{g}_1(\beta^*) - G(\beta^*) - \hat{g}_1(\hat{\beta}_0) + G(\hat{\beta}_0) + O_p\left(\frac{1}{\sqrt{n}}\right) \\ &= O_p\left(\frac{\sqrt{\log m}}{m} + \frac{1}{\sqrt{n}}\right). \end{aligned}$$

Therefore,  $h(\beta) > h(\beta^*)$ . According to Lemma 0.1, the conclusion holds. ■

## 4 First order Newton method

Consider a general statistical estimation problem in the following risk minimization form

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} F(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\xi} \sim \Pi} f(\boldsymbol{\theta}, \boldsymbol{\xi}) \quad (4.1)$$

where  $f(\cdot, \boldsymbol{\xi}) : \mathbb{R}^p \rightarrow \mathbb{R}$  is a convex loss function that can be non-differentiable and  $\boldsymbol{\xi}$  denotes the random sample from a probability distribution. The “one-step estimator” essentially performs the following Newton-type step based on

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 - \boldsymbol{\Sigma}^{-1} \left( \frac{1}{n} \sum_{i=1}^n g(\hat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i) \right), \quad (4.2)$$

where  $\Sigma$  is the population Hessian matrix and  $\left(\frac{1}{n} \sum_{i=1}^n g\left(\hat{\theta}_0, \xi_i\right)\right)$  is the subgradient vector. The estimation of  $\Sigma$  is not easy when  $f$  is non-differentiable or in high-dimensional situation, and the empirical Hessian matrix does not exist.

To address this issue, [Chen, Liu, and Zhang \[2018\]](#) propose an estimator of  $\Sigma^{-1}\mathbf{a}$  for any  $\mathbf{a} \in \mathcal{R}^p$  only using the stochastic first-order information, which is called **First-Order Newton-type Estimator (FONE)**. Then it solves (4.2) as a special case  $\mathbf{a} = \frac{1}{n} \sum_{i=1}^n g\left(\hat{\theta}_0, \xi_i\right)$ .

For a given initial estimator  $\hat{\theta}_0$ , we can perform the Newton-type step in (4.2)

$$\tilde{\theta} = \hat{\theta}_0 - \widehat{\Sigma^{-1}\mathbf{a}}, \quad \mathbf{a} = \left(\frac{1}{n} \sum_{i=1}^n g\left(\hat{\theta}_0, \xi_i\right)\right). \quad (4.3)$$

Note that  $\Sigma^{-1}\mathbf{a} = \sum_{i=0}^{\infty} (1 - \eta\Sigma)^i \eta\mathbf{a}$ , for some small enough  $\eta$  such that  $\|\eta\Sigma\| < 1$ . Then we can use the following iterative procedure  $\{\tilde{\mathbf{z}}_t\}$  to approximate  $\Sigma^{-1}\mathbf{a}$ ,

$$\tilde{\mathbf{z}}_t = \tilde{\mathbf{z}}_{t-1} - \eta(\Sigma\tilde{\mathbf{z}}_{t-1} - \mathbf{a}), \quad 1 \leq t \leq T. \quad (4.4)$$

When  $T$  is large enough, we have

$$\begin{aligned} \tilde{\mathbf{z}}_T &= \tilde{\mathbf{z}}_{T-1} - \eta(\Sigma\tilde{\mathbf{z}}_{T-1} - \mathbf{a}) \\ &= (I - \eta\Sigma)\tilde{\mathbf{z}}_{T-1} + \eta\Sigma\mathbf{a} \\ &= (I - \eta\Sigma)^2\tilde{\mathbf{z}}_{T-2} + (I - \eta\Sigma)\eta\mathbf{a} + \eta\mathbf{a} \\ &= (I - \eta\Sigma)^{T-1}\tilde{\mathbf{z}}_1 + \sum_{i=0}^{T-2} (I - \eta\Sigma)^i \eta\mathbf{a} \approx \Sigma^{-1}\mathbf{a}, \end{aligned}$$

which implies that (4.4) leads to an approximation of  $\Sigma^{-1}\mathbf{a}$ . Let us define  $\mathbf{z}_t = \hat{\theta}_0 - \tilde{\mathbf{z}}_t$ , which is the LHS of (4.3). To avoid estimating  $\Sigma$  in (4.4), we adopt the following first-order approximation

$$-\Sigma\tilde{\mathbf{z}}_{t-1} = \Sigma(\mathbf{z}_{t-1} - \hat{\theta}_0) \approx g_{B_t}(\mathbf{z}_{t-1}) - g_{B_t}(\hat{\theta}_0), \quad (4.5)$$

where  $g_{B_t}(\theta) = \frac{1}{m} \sum_{i \in B_t} g(\theta, \xi_i)$  is the averaged stochastic subgradient over a subset of the data indexed by  $B_t \subseteq \{1, 2, \dots, n\}$ . Here  $B_t$  is randomly chosen from  $\{1, \dots, n\}$  with replacement for every iteration. Then we can construct FONE of  $\hat{\theta}_0 - \Sigma^{-1}\mathbf{a}$  by the following recursive update

$$\mathbf{z}_t = \mathbf{z}_{t-1} - \eta \left\{ g_{B_t}(\mathbf{z}_{t-1}) - g_{B_t}(\hat{\theta}_0) + \mathbf{a} \right\}, \quad \mathbf{z}_0 = \hat{\theta}_0. \quad (4.6)$$

Compare FONE (4.6) with the SVRG (2.2), then FONE can be reduces to a mini-batch version of SVRG.

## References

X. Chen, W. Liu, and Y. Zhang. First-order newton-type estimator for distributed estimation and inference. *arXiv preprint arXiv:1811.11368*, 2018.



- R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.
- M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.