Stochastic Optimization

Yajie Bao
Shanghai Jiao Tong University

January 25, 2020

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

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

And denote the estimate of β 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 μ-strongly convex, then $|\hat{\beta} - \beta^*| = O_p(\frac{1}{\sqrt{n}})$.*

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

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

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

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

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

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

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

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

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

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

for $t = 0, 1, ...$ choose i_t uniformly at random then update x^t by

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

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

Theorem 1.2 *Under Assumption 1.1, suppose F is μ-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*

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

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

Proof: Using SGD update rule, we have

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

Since $\boldsymbol{x}^t$ depend on $\xi_1, \cdots, \xi_{t-1}$,

$$\begin{aligned}\mathbb{E}\left[\left(\boldsymbol{x}^t - \boldsymbol{x}^*\right)^\top \boldsymbol{g}\left(\boldsymbol{x}^t; \boldsymbol{\xi}^t\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(\boldsymbol{x}^t - \boldsymbol{x}^*\right)^\top \boldsymbol{g}\left(\boldsymbol{x}^t; \boldsymbol{\xi}^t\right) | \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\left(\boldsymbol{x}^t - \boldsymbol{x}^*\right)^\top \mathbb{E}\left[\boldsymbol{g}\left(\boldsymbol{x}^t; \boldsymbol{\xi}^t\right) | \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_{t-1}\right]\right]. \\ &= \mathbb{E}\left[\left(\boldsymbol{x}^t - \boldsymbol{x}^*\right)^\top \nabla F\left(\boldsymbol{x}^t\right)\right]\end{aligned}$$

Furthermore, strong convexity gives

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

which implies

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

Combine the above results to obtain,

$$\mathbb{E}\left[\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\right\|_2^2\right] \leq \left(1 - 2\mu\eta_t\right)\mathbb{E}\left[\left\|\boldsymbol{x}^t - \boldsymbol{x}^*\right\|_2^2\right] + \eta_t^2\sigma_{\mathrm{g}}^2.$$

First for $t = 0$,

$$\mathbb{E}\left[\left\|\boldsymbol{x}^1 - \boldsymbol{x}^*\right\|_2^2\right] \leq (1 - 2\mu\theta)\mathbb{E}\left[\left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2^2\right] + \theta^2\sigma_{\mathrm{g}}^2 \leq c_\theta.$$

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

$$\begin{aligned}\mathbb{E}\left[\left\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^*\right\|_2^2\right] &\leq \frac{t + 1 - 2\mu\theta}{(t+1)^2}c_\theta + \frac{\theta^2\sigma_{\mathrm{g}}^2}{(t+1)^2} \\ &= \frac{t}{(t+1)^2}c_\theta - \frac{2\mu\theta - 1}{(t+1)^2}c_\theta + \frac{\theta^2\sigma_{\mathrm{g}}^2}{(t+1)^2} \\ &\leq \frac{t}{(t+1)^2}c_\theta - \frac{\theta^2\sigma_{\mathrm{g}}^2}{2(t+1)^2} \\ &\leq \frac{1}{(t+2)^2}c_\theta.\end{aligned}$$

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

2 Stochastic variance reduced gradient

Consider the following optimization problem

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

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 $\{1, 2, ..., n\}$:

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

And we have

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

The SVRG is presented in Algorithm 1.

Algorithm 1: Stochastic Variance Reduced Gradient

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

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

Compute the initial estimator $\tilde{w}_0$

for $s = 1, 2, ..., L$ **do**

  Set $\tilde{w} = \tilde{w}_{s-1}$

  Compute the average gradient $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_i(\tilde{w})$

  Set $w_0 = \tilde{w}$

  for $t = 1, 2, ..., m$ **do**

    Randomly pick i_t from $\{1, 2, ..., n\}$ and update estiamte

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

  end

  Option 1: Set $\tilde{w}_s = w_m$

  Option 2: Set $\tilde{w}_s = w_t$ for randomly chosen $t \in \{1, 2, ..., n\}$

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:

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

3 Distributed SGD

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

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

where $\bar{\theta}$ is the initial estimator of θ. 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 $\mathcal{L}_N(\theta)$

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

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

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

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

$$\widetilde{\mathcal{L}}(\theta) := \mathcal{L}_1(\theta) - \left\langle \theta, \nabla\mathcal{L}_1(\bar{\theta}) - \nabla\mathcal{L}_N(\bar{\theta}) \right\rangle. \tag{3.3}$$

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 $k = 1, 2, ..., N$, the number of iterations L

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

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

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

for $t = 1, 2, ..., L$ **do**

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

for $i = 1, 2, ..., N$ **do**

In each machine, compute the gradient of empirical lost function

$$\hat{g}_i(\hat{\beta}_{g-1}) = \frac{1}{m} \sum_{k \in H_i} \partial f(X_k, \hat{\beta}_0)$$

end

Compute the pooled gradient:

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

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

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

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*

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

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

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

where $h(\beta) = \frac{1}{m}\sum_{k\in H_i} f(X_k,\beta) - \beta^T(\hat{g}_1(\hat{\beta}_{t-1})) - \hat{g}(\hat{\beta}_{t-1})$. 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

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

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

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

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

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

where $f(\cdot,\boldsymbol{\xi}) : \mathbb{R}^p \to \mathbb{R}$ 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

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

where $\boldsymbol{\Sigma}$ is the population Hessian matrix and $\left(\frac{1}{n}\sum_{i=1}^{n} g\left(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i\right)\right)$ is the subgradient vector. The estimation of $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ for any $\boldsymbol{a} \in \mathcal{R}^p$ 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 $\boldsymbol{a} = \frac{1}{n}\sum_{i=1}^{n} g\left(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i\right)$.

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

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_0 - \widehat{\boldsymbol{\Sigma}^{-1}\boldsymbol{a}}, \quad \boldsymbol{a} = \left(\frac{1}{n}\sum_{i=1}^{n} g\left(\widehat{\boldsymbol{\theta}}_0, \boldsymbol{\xi}_i\right)\right). \tag{4.3}$$

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

$$\widetilde{\boldsymbol{z}}_t = \widetilde{\boldsymbol{z}}_{t-1} - \eta\left(\boldsymbol{\Sigma}\widetilde{\boldsymbol{z}}_{t-1} - \boldsymbol{a}\right), 1 \le t \le T. \tag{4.4}$$

When T is large enough, we have

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

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

$$-\boldsymbol{\Sigma}\tilde{\boldsymbol{z}}_{t-1} = \boldsymbol{\Sigma}\left(\boldsymbol{z}_{t-1} - \widehat{\boldsymbol{\theta}}_0\right) \approx g_{B_t}\left(\boldsymbol{z}_{t-1}\right) - g_{B_t}\left(\widehat{\boldsymbol{\theta}}_0\right), \tag{4.5}$$

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

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

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.