CHAPTER 9. VARIANCE REDUCTION

Kun Yuan

November 16, 2023

1 Finite-sum minimization

An important practical scenario for stochastic optimization is finite-sum minimization. Suppose the random variable ξ takes samples **uniformly** from a finite dataset $\{\xi_1, \xi_2, \cdots, \xi_m\}$, i.e., $\Pr(\xi = \xi_i) = 1/m$, where m is the size of the finite dataset, it then follows that

$$\mathbb{E}[F(x;\xi)] = \frac{1}{m}\sum_{i=1}^{m} F(x;\xi_i) = \frac{1}{m}\sum_{i=1}^{m} F_i(x) \tag{1}$$

where $F_i(x) := F(x;\xi_i)$ is defined as the loss function associated with data sample ξ_i. With relation (1), the stochastic optimization problem reduces to

$$\min_{x\in\mathbb{R}^d} \quad f(x) \quad \text{where} \quad f(x) = \frac{1}{m}\sum_{i=1}^{m} F_i(x), \tag{2}$$

which is referred to as finite-sum minimization or empirical risk minimization.

2 SGD for finite-sum minimization

Since problem (2) is a special example of stochastic optimization, any algorithm that can solve stochastic optimization can also apply to problem (2). Now we examine the SGD algorithm for finite-sum minimization. Since the random variable ξ is sampled from the finite dataset $\{\xi_1, \xi_2, \cdots, \xi_m\}$, SGD for finite-sum minimization is

$$\begin{aligned} &\text{Sample } i_t \in [m] \text{ uniformly randomly} \\ &x^{(t+1)} = x^{(t)} - \gamma \nabla F_{i_t}(x^{(t)}), \quad \forall\ t = 0, 1, 2, \cdots \end{aligned} \tag{3}$$

where $F_i(x) = F(x;\xi_i)$ and i_t is a random index variable at iteration t. In convergence analysis, we typically assume the stochastic gradient $\nabla F_{i_t}(x^{(t)})$ is unbiased and has bounded variance, i.e.

$$\mathbb{E}[\|\nabla F_{i_t}(x^{(t)}) - \nabla f(x^{(t)})\|^2 | \mathcal{F}^t] \leq \sigma^2, \tag{4}$$

where $\mathcal{F}^t = \{x^{(t)}, x^{(t-1)}, \cdots, x^{(0)}\}$. The non-vanishing gradient variance σ^2 is the root reason resuling in the slow convergence in SGD.

3 Stochastic variance-reduced gradient

We introduce a sequence of auxiliary variables $\{\alpha_j\}_{j=1}^m$ with each $\alpha_j \in \mathbb{R}^d$ share the same dimension as x. Consider the following new stochastic gradient:

$$g_t(x) = \nabla F_{i_t}(x) - \nabla F_{i_t}(\alpha_{i_t}) + u \quad \text{where} \quad u = \frac{1}{m}\sum_{j=1}^m \nabla F_j(\alpha_j) \tag{5}$$

where the random index variable $i_t \in [m]$ is sampled uniformly. It is easy to verify that

- $g_t(x)$ is unbiased. To see it, we can derive that

$$\mathbb{E}[g_t(x)] = \frac{1}{m}\sum_{j=1}^m \left[\nabla F_j(x) - \nabla F_j(\alpha_j)\right] + \frac{1}{m}\sum_{j=1}^m \nabla F_j(\alpha_j) = \frac{1}{m}\sum_{j=1}^m \nabla F_j(x) = f(x) \tag{6}$$

- $g_t(x)$ has bounded variance. To see it, we can derive that

$$\begin{aligned}
\mathbb{E}\|g_t(x) - \nabla f(x)\|^2 &= \mathbb{E}\|\nabla F_{i_t}(x) - \nabla F_{i_t}(\alpha_{i_t}) - (\nabla f(x) - u)\|^2 \\
&\leq \mathbb{E}\|\nabla F_{i_t}(x) - \nabla F_{i_t}(\alpha_{i_t})\|^2 \\
&\leq \frac{1}{m}\sum_{j=1}^m \|\nabla F_j(x) - \nabla F_j(\alpha_j)\|^2 \\
&\leq \frac{L^2}{m}\sum_{j=1}^m \|x - \alpha_j\|^2
\end{aligned} \tag{7}$$

  where we assume each $F_i(x)$ is L-smooth.

It is worth noting that the variance in (7) is not a constant σ^2 anymore. If we develop strategies that allow $\alpha_j \to x$, the gradient variance can vanish to 0, which will accelerate the convergence rate.

4 SAGA

To achieve $\alpha_j \to x$, a natural strategy is to update auxiliary variables $\{\alpha_j\}_{j=1}^m$ with the most recent $x^{(t)}$:

$$\alpha_i^{(t+1)} = \begin{cases} x^{(t)} & \text{if index } i \text{ is sampled;} \\ \alpha_i^{(t)} & \text{otherwise.} \end{cases} \tag{8}$$

Algorithm 1 SAGA for Finite-sum Minimization

Input: an arbitrary x^0; let $\alpha_j^{(0)} = x^0$ for any j, and $u^{(0)} = \frac{1}{m}\sum_{j=1}^m \nabla F_j(\alpha_j^{(0)})$.
for $t = 0, 1, 2, ..., T-1$ **do**
  sample $i_t \in \{1, 2, \cdots, m\}$ uniformly randomly
  update $g_{i_t} = \nabla F_{i_t}(x^{(t)}) - \nabla F_{i_t}(\alpha_{i_t}^{(t)}) + u^{(t)}$
  update $x^{(t+1)} = x^{(t)} - \gamma g_{i_t}$
  update $\alpha_{i_t}^{(t+1)} = x^{(t)}$ and $\alpha_j^{(t+1)} = \alpha_j^{(t)}$ if $j \neq i_t$
  update $u^{(t+1)} = u^{(t)} + \frac{1}{m}\big(\nabla F_i(x^{(t)}) - \nabla F_i(\alpha_i^{(t)})\big)$
end for
Output: variable $x^{(T)}$

With the above update, we can compute $u^{(t+1)}$ in a cheap recursive manner

$$\begin{aligned} u^{(t+1)} &= \frac{1}{m}\sum_{j=1}^m \nabla F_j(\alpha_j^{(t+1)}) \\ &= \frac{1}{m}\sum_{j\neq i} \nabla F_j(\alpha_j^{(t)}) + \frac{1}{m}\nabla F_i(x^{(t)}) \\ &= u^{(t)} + \frac{1}{m}\big(\nabla F_i(x^{(t)}) - \nabla F_i(\alpha_i^{(t)})\big). \end{aligned} \tag{9}$$

With the above strategy, we can develop the SAGA algorithm [1] in Algorithm 1. It is observed that SAGA achieves an one-loop algorithm by maintaining $\{\alpha_j\}_{j=1}^m$ with $O(md)$ memory, which can be very expensive when d or m is large. In real implementations, we can store $\{\nabla F_j(\alpha_j)\}_{j=1}^m$ instead of $\{\alpha_j\}_{j=1}^m$ which can save the cost to evaluate $\nabla F_{i_t}(\alpha_{i_t}^{(t)})$ in (5).

5 Convergence analysis

5.1 Assumption and supporting lemma

We introduce the following assumption throughout this chapter

Assumption 5.1. We assume each $F_i(x)$ is L-smooth, and each random index variable i_t is sampled uniformly and randomly from set $\{1, \cdots, m\}$.

We let $\mathcal{F}^t = \{x^{(t)}, \{\alpha_j^{(t)}\}_{j=1}^m, \cdots, x^{(0)}, \{\alpha_j^{(0)}\}_{j=1}^m\}$. Following the arguments in Sec. 3, we can prove the following lemma:

Lemma 5.2. Under Assumption 5.1, it holds that

$$\mathbb{E}[g_{i_t}|\mathcal{F}^t] = \nabla f(x^{(t)}) \tag{10}$$

$$\mathbb{E}[\|g_{i_t} - \nabla f(x^{(t)})\|^2|\mathcal{F}^t] \leq \frac{L^2}{m}\sum_{j=1}^m \|x^{(t)} - \alpha_j^t\|^2 \tag{11}$$

5.2 Descent lemma

Lemma 5.3. Under Assumption 5.1, if the learning rate $\gamma \leq \frac{1}{L}$, the iterate $x^{(t)}$ generated by SAGA satisfies

$$\mathbb{E}[f(x^{(t+1)})] \leq \mathbb{E}[f(x^{(t)})] - \frac{\gamma}{2}\mathbb{E}\|\nabla f(x^{(t)})\|^2 + \frac{L^3\gamma^2}{2m}\sum_{j=1}^{m}\mathbb{E}\|x^{(t)} - \alpha_j^{(t)}\|^2. \tag{12}$$

Proof. Since $f(x)$ is L-smooth, it holds that

$$\begin{aligned}
\mathbb{E}[f(x^{(t+1)})|\mathcal{F}^{(t)}] &\leq f(x^{(t)}) - \gamma\mathbb{E}[\langle \nabla f(x^{(t)}), g_{i_t}\rangle|\mathcal{F}^{(t)}] + \frac{L\gamma^2}{2}\mathbb{E}[\|g_{i_t}\|^2|\mathcal{F}^{(t)}] \\
&\overset{(a)}{=} f(x^{(t)}) - \gamma\|\nabla f(x^{(t)})\|^2 + \frac{L\gamma^2}{2}\mathbb{E}[\|g_{i_t}\|^2|\mathcal{F}^{(t)}] \\
&\overset{(b)}{\leq} f(x^{(t)}) - (\gamma - \frac{L\gamma^2}{2})\|\nabla f(x^{(t)})\|^2 + \frac{L^3\gamma^2}{2m}\sum_{j=1}^{m}\|x^{(t)} - \alpha_j^{(t)}\|^2 \\
&\overset{(c)}{\leq} f(x^{(t)}) - \frac{\gamma}{2}\|\nabla f(x^{(t)})\|^2 + \frac{L^3\gamma^2}{2m}\sum_{j=1}^{m}\|x^{(t)} - \alpha_j^{(t)}\|^2
\end{aligned} \tag{13}$$

where (a) holds because of equality (10), and (b) holds because

$$\begin{aligned}
\mathbb{E}[\|g_{i_t}\|^2|\mathcal{F}^{(t)}] &= \mathbb{E}[\|g_{i_t} - \nabla f(x^{(t)}) + \nabla f(x^{(t)})\|^2|\mathcal{F}^{(t)}] \\
&\leq \mathbb{E}[\|g_{i_t} - \nabla f(x^{(t)})\|^2|\mathcal{F}^{(t)}] + \|\nabla f(x^{(t)})\|^2 \\
&\overset{(11)}{\leq} \frac{L^2}{m}\sum_{j=1}^{m}\|x^{(t)} - \alpha_j^{(t)}\|^2 + \|\nabla f(x^{(t)})\|^2.
\end{aligned} \tag{14}$$

Inequality (c) in (13) holds when $\gamma \leq 1/L$. By taking expectations over $\mathcal{F}^{(t)}$ on both sides of (13), we achieve the result in (12). □

5.3 Evolution of $\{\alpha_j\}_{j=1}^n$

Lemma 5.4. Under Assumption 5.1, it holds for any $t = 0, 1, 2, \cdots, T-1$ and $\beta > 0$ that

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|x^{(t+1)} - \alpha_j^{(t+1)}\|^2 \leq \Big(1 - \frac{1}{m} + \gamma\beta + \gamma^2L^2\Big)\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|x^{(t)} - \alpha_i^{(t)}\|^2 \\
+ \big(\gamma^2 + \frac{\gamma}{\beta}\big)\mathbb{E}\|\nabla f(x^{(t)})\|^2
\end{aligned} \tag{15}$$

Proof. Recalling the update of $\alpha_i^{(t+1)}$ in (8) for a given $i \in [m]$, we have

$$
\begin{aligned}
&\mathbb{E}[\|x^{(t+1)} - \alpha_i^{(t+1)}\|^2 | \mathcal{F}^{(t)}] \\
&= \frac{1}{m}\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2 | \mathcal{F}^{(t)}] + \frac{m-1}{m}\mathbb{E}[\|x^{(t+1)} - \alpha_i^{(t)}\|^2 | \mathcal{F}^{(t)}] \\
&= \frac{1}{m}\mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2 | \mathcal{F}^{(t)}] + \frac{m-1}{m}\mathbb{E}[\|x^{(t+1)} - x^{(t)} + x^{(t)} - \alpha_i^{(t)}\|^2 | \mathcal{F}^{(t)}] \\
&= \mathbb{E}[\|x^{(t+1)} - x^{(t)}\|^2 | \mathcal{F}^{(t)}] + \frac{m-1}{m}\|x^{(t)} - \alpha_i^{(t)}\|^2 \\
&\quad + \frac{2(m-1)}{m}\mathbb{E}[\langle x^{(t+1)} - x^{(t)}, x^{(t)} - \alpha_i^{(t)} \rangle | \mathcal{F}^{(t)}] \\
&= \frac{m-1}{m}\|x^{(t)} - \alpha_i^{(t)}\|^2 + \gamma^2\mathbb{E}[\|g_{i_t}\|^2 | \mathcal{F}^{(t)}] + \frac{2(m-1)\gamma}{m}\mathbb{E}[\langle g_{i_t}, x^{(t)} - \alpha_i^{(t)} \rangle | \mathcal{F}^{(t)}] \\
&= \frac{m-1}{m}\|x^{(t)} - \alpha_i^{(t)}\|^2 + \gamma^2\mathbb{E}[\|g_{i_t}\|^2 | \mathcal{F}^{(t)}] + \frac{2(m-1)\gamma}{m}\langle \nabla f(x^{(t)}), x^{(t)} - \alpha_i^{(t)} \rangle \\
&\overset{(a)}{\leq} \frac{m-1}{m}(1+\gamma\beta)\|x^{(t)} - \alpha_i^{(t)}\|^2 + \big(\gamma^2 + \frac{(m-1)\gamma}{m\beta}\big)\|\nabla f(x^{(t)})\|^2 \\
&\quad + \frac{\gamma^2 L^2}{m}\sum_{j=1}^{m}\|x^{(t)} - \alpha_j^{(t)}\|^2 \\
&\leq (1 - \frac{1}{m} + \gamma\beta)\|x^{(t)} - \alpha_i^{(t)}\|^2 + \big(\gamma^2 + \frac{\gamma}{\beta}\big)\|\nabla f(x^{(t)})\|^2 + \frac{\gamma^2 L^2}{m}\sum_{j=1}^{m}\|x^{(t)} - \alpha_j^{(t)}\|^2 \qquad (16)
\end{aligned}
$$

where inequality (a) holds due to inequality (14) and the fact that $2\langle a, b\rangle \leq \beta\|a\|^2 + \frac{1}{\beta}\|b\|^2$ for any $\beta > 0$ and $a, b, \in \mathbb{R}^d$. By taking average over all indices on both sides of (16) and taking expectations over the filtration, we achieve the result in (15). □

5.4 Convergence theorem

Theorem 5.5. Under Assumption 5.1, if learning rate $\gamma = (3Lm^{2/3})^{-1}$, then $x^{(t)}$ generated by SAGA will converge as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^{(t)})\|^2 \leq \frac{10Lm^{2/3}\Delta_0}{T} \qquad (17)$$

where $\Delta_0 = f(x^{(0)}) - f^\star$.

Proof. We introduce the following Lyapunov function

$$
\begin{aligned}
L^{(t+1)} &= \mathbb{E}[f(x^{(t+1)})] + \frac{c_{t+1}}{m}\sum_{j=1}^{m}\mathbb{E}\|x^{(t+1)} - \alpha_i^{(t+1)}\|^2 \\
&\leq \mathbb{E}[f(x^{(t)})] - [\frac{\gamma}{2} - c_{t+1}(\gamma^2 + \frac{\gamma}{\beta})]\mathbb{E}\|\nabla f(x^{(t)})\|^2 \\
&\quad + [c_{t+1}\big(1 - \frac{1}{m} + \gamma\beta + \gamma^2 L^2\big) + \frac{L^3\gamma^2}{2}]\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\|x^{(t)} - \alpha_j^{(t)}\|^2 \qquad (18)
\end{aligned}
$$

If we define

$$c_t = c_{t+1}\big(1 - \frac{1}{m} + \gamma\beta + \gamma^2 L^2\big) + \frac{L^3\gamma^2}{2}, \quad c_T = 0 \tag{19}$$

$$r_t = \frac{\gamma}{2} - c_{t+1}(\gamma^2 + \frac{\gamma}{\beta}) \tag{20}$$

Inequality (16) is equivalent to

$$L^{(t+1)} \leq L^{(t)} - r_t \mathbb{E}\|\nabla f(x^{(t)})\|^2. \tag{21}$$

Next we evaluate c_t. Let $\theta = \frac{1}{m} - \gamma\beta - \gamma^2 L^2$. If we let $\beta = L/m^{\frac{1}{3}}$ and $\gamma = 1/(3Lm^{2/3})$, it holds that $\theta < 1$ and $\theta \geq \frac{5}{9m}$. Substituting θ into (19), we have

$$c_t = c_{t+1}\big(1-\theta\big) + \frac{L^3\gamma^2}{2}, c_T = 0 \quad \Longrightarrow \quad c_t \leq c_0 \leq \frac{L^3\gamma^2}{2\theta} \leq \frac{L}{10m^{1/3}}. \tag{22}$$

Substituting (22) to (20), we have $r_t \geq 1/(10Lm^{2/3})$ and hence (21) becomes

$$\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t)} - \frac{\mathbb{E}\|\nabla f(x^{(t)})\|^2}{10Lm^{2/3}}. \tag{23}$$

By taking average over iterations, and recalling that $\alpha_j^{(0)} = x^{(0)}$, we achieve

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla f(x^{(t)})\|^2 \leq \frac{10Lm^{2/3}(f(x^{(0)}) - f^\star)}{T}. \tag{24}$$

□

Remark. Recall that standard SGD converge at rate $O(1/\sqrt{T})$, we conclude that SAGA has a faster convergence rate.

References

[1] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," *Advances in neural information processing systems*, vol. 27, 2014.