Optimization Theory and Algorithm II **October 13, 2022**

Lecture 10

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

1 SVRG

1.1 Covariate of Stochastic Gradient

We still consider the finite-sum optimization problem. Let $\{(\mathbf{a}_i, b_i)\}_{i=1}^m$ be a dataset, $\mathcal{F} = \{h_{\mathbf{x}} | h_{\mathbf{x}} : \mathcal{A} \to \mathcal{B}, \mathbf{x} \in \mathbb{R}^n\}$ be a class of predictor function and ℓ be a loss function. Then the corresponding finite-sum optimization problem is

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{m} \sum_{i}^{m} f_i(\mathbf{x}), \tag{1}$$

where $f_i(\mathbf{x}) := \ell(b_i, h_{\mathbf{x}}(\mathbf{a}_i))$.

Though the stochastic gradient is an unbiased estimator of the gradient, it may have high variance. Indeed, to analysis SGD we had to start by imposing the assumption about its variance

$$\mathbb{E}_{i_t}[\|\nabla f_{i_t}(\mathbf{x}^t)\|^2] \leq \sigma^2 + \|\nabla f(\mathbf{x}^t)\|^2.$$

Even with the above assumption, we required decreasing step sizes to gradually kill off the variance. Yet another glaring issue with SGD is that even if we start the SGD algorithm on the optimal point $\mathbf{x}^0 = \mathbf{x}^*$, then method will not stop. This is because the stochastic gradients are not necessarily zero on the solution, that is $\nabla f_{i_t}(\mathbf{x}^*) \neq 0$ is entirely possible (see the example in SGD section).

The aim of SVRG is to construct a new "gradient" $\mathbf{g}^t$ such that

- Unbiased: $\mathbb{E}[\mathbf{g}^t] = \nabla f(\mathbf{x}^t)$
- Reducing Variance:
$$\mathbb{E}[\|\mathbf{g}^t\|_2^2] \to 0, \text{ as } \mathbf{x}^t \to \mathbf{x}^*.$$

How to reduce the variance of stochastic gradient? The basic idea is to consider an important method in MCMC. That is to construct a covariate variable of $\nabla f_{i_t}(\mathbf{x}^t)$.

Let $\mathbf{x}$ be a random variable. We asy that a random variable $\mathbf{z}$ is a covariate of $\mathbf{x}$ if $cov(\mathbf{x}, \mathbf{z}) > 0$. Then, utilizing $\mathbf{z}$ can build an unbiased estimator of $\mathbf{x}$ that has a small variance. Let

$$\mathbf{x}_{\mathbf{z}} = \mathbf{x} - \mathbf{z} + \mathbb{E}[\mathbf{z}]. \tag{2}$$

And note that $\mathbb{E}[\mathbf{x}_{\mathbf{z}}] = \mathbb{E}[\mathbf{x}]$, and

$$Var[\mathbf{x}_{\mathbf{z}}] = Var[\mathbf{x}] + Var[\mathbf{z}] - 2Cov(\mathbf{x}, \mathbf{z}). \tag{3}$$

Consequently, if $Cov(\mathbf{x}, \mathbf{z})$ is sufficiently large, then $Var[\mathbf{x}_{\mathbf{z}}]$ is small.

1.2 SVRG Algorithm

We can build an estimate of the gradient with reduced variance by finding covariates for the stochastic gradient.

Let $\mathbf{x}^t \in \mathbb{R}^n$ be our current iterate and $\tilde{\mathbf{x}}^k$ be a *reference point*. If $\mathbf{x}^t$ is sufficiently close to $\tilde{\mathbf{x}}^k$ it is reasonable to expect that $\nabla f_{i_t}(\mathbf{x}^t)$ is the covariate of $\nabla f_{i_t}(\tilde{\mathbf{x}}^k)$. Then, i_t is uniformly sampled from $[m]$ and

$$\mathbf{g}^t = \nabla f_{i_t}(\mathbf{x}^t) - \nabla f_{i_t}(\tilde{\mathbf{x}}^k) + \nabla f(\tilde{\mathbf{x}}^k). \tag{4}$$

Let us first refer to the SVRG [1] Algorithm 1.

Algorithm 1 SVRG

Parameters update frequency T and learning rate s
Initialize $\widetilde{\mathbf{x}}^0$
for $k = 1, 2, \ldots$ **do**
  $\mathbf{x}^0 = \widetilde{\mathbf{x}}^{k-1}$
  for $t = 1, 2, \ldots, T$ **do**
    Randomly pick $i_t \in \{1, \ldots, n\}$ and update weight

$$\mathbf{g}^{t-1} = \nabla f_{i_t}(\mathbf{x}^{t-1}) - \nabla f_{i_t}(\widetilde{\mathbf{x}}^{k-1}) + \nabla f(\widetilde{\mathbf{x}}^{k-1}), \tag{5}$$

$$\mathbf{x}^t = \mathbf{x}^{t-1} - s\mathbf{g}^{t-1}. \tag{6}$$

  end for
  Last Option: $\widetilde{\mathbf{x}}^k = \mathbf{x}^T$;
  Average Option: $\widetilde{\mathbf{x}}^k = 1/T \sum_t \mathbf{x}^t$;
  Random Option : $\widetilde{\mathbf{x}}^k = \mathbf{x}^t$ for randomly chosen $t \in \{1, \ldots, T\}$.
end for

1.3 Convergence Analysis

We suppose that f is β smooth and α-strongly convex, and f_i is convex and β_i smooth. Before the convergence analysis, we present the following lemmas.

Lemma 1 *Let f is a β smooth function then*

$$f(\mathbf{x} - \frac{1}{\beta}\nabla f(\mathbf{x})) - f(\mathbf{x}) \leq -\frac{1}{2\beta}\|\nabla f(\mathbf{x})\|^2. \tag{7}$$

Lemma 2 *If each f_{i_t} is β_{i_t} smooth and convex, then*

$$\mathbb{E}[\|\nabla f_{i_t}(\mathbf{x}) - \nabla f_{i_t}(\mathbf{x}^*)\|^2] \leq 2\beta_{\max}(f(\mathbf{x}) - f(\mathbf{x}^*)). \tag{8}$$

Proof 1 *Let $h_{i_t}(\mathbf{x}) = f_{i_t}(\mathbf{x}) - f_{i_t}(\mathbf{x}^*) - \langle \nabla f_{i_t}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0$ due to the convexity. By Lemma 2, we have that*

$$-h_{i_t}(\mathbf{x}) \leq h_{i_t}(\mathbf{x} - \frac{1}{\beta_{i_t}}\nabla h_{i_t}(\mathbf{x})) - h_{i_t}(\mathbf{x}) \leq -\frac{1}{2\beta_{i_t}}\|\nabla h_{i_t}(\mathbf{x})\|^2 \leq -\frac{1}{2\beta_{\max}}\|\nabla h_{i_t}(\mathbf{x})\|^2. \tag{9}$$

By substituting h_{i_t}, then

$$\|\nabla f_{i_t}(\mathbf{x}) - \nabla f_{i_t}(\mathbf{x}^*)\|^2 \leq 2\beta_{\max}(f_{i_t}(\mathbf{x}) - f_{i_t}(\mathbf{x}^*) - \langle \nabla f_{i_t}(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle). \tag{10}$$

Then taking expectation with respect to i_t and using that $\mathbb{E}[\nabla f_{i_t}(\mathbf{x}^*)] = \nabla f(\mathbf{x}^*) = 0$, we can finish the prove.

Now we bound the variance of stochastic gradient.

Lemma 3 *The second moment of SVRG gradient is bounded as*

$$\mathbb{E}_{i_t}[\|\mathbf{g}^t\|^2] \leq 4\beta_{\max}(f(\mathbf{x}^t) - f(\mathbf{x}^*)) + 4\beta_{\max}(f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x}^*)). \tag{11}$$

Proof 2

$$\begin{aligned}
\mathbb{E}_{i_t}\|\mathbf{g}^t\|^2 =& \mathbb{E}_{i_t}[\|\nabla f_{i_t}(\mathbf{x}^t) - \nabla f_{i_t}(\mathbf{x}^*) + \nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}}^k) + \nabla f(\tilde{\mathbf{x}}^k)\|^2] && (12)\\
\leq& 2\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}^t) - \nabla f_{i_t}(\mathbf{x}^*)\|^2 + 2\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}}^k) + \nabla f(\tilde{\mathbf{x}}^k)]\|^2 && (13)\\
\leq& 2\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}^t) - \nabla f_{i_t}(\mathbf{x}^*)\|^2 + 2\mathbb{E}_{i_t}\|\nabla f_{i_t}(\mathbf{x}^*) - \nabla f_{i_t}(\tilde{\mathbf{x}}^k)\|^2 && (14)\\
\leq& 4\beta_{\max}(f(\mathbf{x}^t) - f(\mathbf{x}^*)) + 4\beta_{\max}(f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x}^*)). && (15)
\end{aligned}$$

Based on these lemmas, we can prove the convergence of SVRG.

Theorem 1 *The sequence $\{\widetilde{\mathbf{x}}_k\}$ in Algorithm 1 has the following property*

$$\mathbb{E}[f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x}^*)] \leq \left[\frac{1}{\alpha s(1-2s\beta)T} + \frac{2s\beta}{1-2s\beta}\right]\mathbb{E}[f(\widetilde{\mathbf{x}}^{k-1}) - f(\mathbf{x}^*)], \tag{16}$$

where $\beta = \beta_{\max}$.

Proof 3 *By conditioning on $\mathbf{x}^{t-1}$, we have $\mathbb{E}\mathbf{g}^{t-1} = \nabla f(\mathbf{x}^{t-1})$ and this leads to*

$$\begin{aligned}
\mathbb{E}\|\mathbf{x}^t - \mathbf{x}^*\|^2 =& \|\mathbf{x}^{t-1} - \mathbf{x}^*\|^2 - 2s(\mathbf{x}^{t-1} - \mathbf{x}^*)^\top \mathbb{E}\mathbf{g}^{t-1} + s^2\mathbb{E}\|\mathbf{g}^{t-1}\|^2 && (17)\\
\leq& \|\mathbf{x}^{t-1} - \mathbf{x}^*\|^2 - 2s(\mathbf{x}^{t-1} - \mathbf{x}^*)^\top \nabla f(\mathbf{x}^{t-1}) + 4\beta s^2[f(\mathbf{x}^{t-1}) - f(\mathbf{x}^*) + f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)] && (18)\\
\leq& \|\mathbf{x}^{t-1} - \mathbf{x}^*\|^2 - 2s(f(\mathbf{x}^{t-1}) - f(\mathbf{x}^*)) + 4\beta s^2[f(\mathbf{x}^{t-1}) - f(\mathbf{x}^*) + f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)] && (19)\\
=& \|\mathbf{x}^{t-1} - \mathbf{x}^*\|^2 - 2s(1-2s\beta)[f(\mathbf{x}^{t-1} - f(\mathbf{x}^*)] + 4\beta s^2[f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)]. && (20)
\end{aligned}$$

Take total expectation, summing up over $t = 1, \ldots, T$, then

$$\mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\mathbf{x}^0 - \mathbf{x}^*\|^2 - 2s(1-2s\beta)\mathbb{E}[\sum_{t=0}^{T-1}(f(\mathbf{x}^t) - f(\mathbf{x}^*))] + 4\beta s^2 T[f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)]. \tag{21}$$

Using that $\mathbf{x}^0 = \widetilde{\mathbf{x}}^k$, strong convexity says $f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x}^*) \geq \frac{\alpha}{2}\|\widetilde{\mathbf{x}}^k - \mathbf{x}^*\|^2$ and rearranging the formulation, we have

$$\begin{aligned}
2s(1-2s\beta)\mathbb{E}[\sum_{t=0}^{T-1}(f(\mathbf{x}^t) - f(\mathbf{x}^*))] \leq& \mathbb{E}\|\mathbf{x}^0 - \mathbf{x}^*\|^2 - \mathbb{E}\|\mathbf{x}^T - \mathbf{x}^*\|^2 + 4\beta s^2 T[f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)] && (22)\\
\leq& (4T\beta s^2 + 2/\alpha)[f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)]. && (23)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[f(\widetilde{\mathbf{x}}^{k+1})] - f^* =& \mathbb{E}[f(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{x}^t)] - f^* && (24)\\
\leq& \frac{1}{T}\mathbb{E}[\sum_{t=0}^{T-1} f(\mathbf{x}^t)] - f^* && (25)\\
\leq& \frac{4T\beta s^2 + 2\alpha^{-1}}{T2s(1-2s\beta)}\mathbb{E}[f(\widetilde{\mathbf{x}}_k) - f(\mathbf{x}^*)] && (26)\\
=& \left[\frac{1}{\alpha s(1-2s\beta)T} + \frac{2s\beta}{1-2s\beta}\right]\mathbb{E}[f(\widetilde{\mathbf{x}}^k) - f(\mathbf{x}^*)]. && (27)
\end{aligned}$$

Mover, we can substitute $s = 1/10\beta$ *and* $T = 20\beta/\alpha$*, then*

$$\left[\frac{1}{\alpha s(1-2s\beta)T} + \frac{2s\beta}{1-2s\beta}\right] = 7/8 < 0.9$$

.

Finally, we can obtain

$$\mathbb{E}[f(\widetilde{\mathbf{x}}^k)] - f^* \leq (0.9)^k \mathbb{E}[f(\widetilde{\mathbf{x}}^0) - f(\mathbf{x}^*)]. \tag{28}$$

References

[1] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.