Beyond Blackbox Models: Stochastic Variance Reduced Gradient Methods

Yudong Chen

In this lecture, we discuss how to speed up stochastic optimization by leveraging the structure of the problem. In particular, we consider the Stochastic Variance Reduced Gradient (SVRG) method for minimizing a finite sum of smooth and strongly convex functions.

Readings:

- Original SVRG paper: Johnson and Zhang 2013
- Section 6.3 of Bubeck's monograph
- See Defazio 2014 for related methods and their relationship.

1 Finite-sum minimization

Throughout this lecture, we use $\|\cdot\|$ to denote the Euclidean ℓ_2 norm.

Consider unconstrained minimization of a function $f : \mathbb{R}^d \to \mathbb{R}$ given by

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x),$$

where each *individual* f_i is L-smooth and convex, and f is m-strongly convex w.r.t. $\|\cdot\|$. Let $\kappa := \frac{L}{m}$ be the condition number. Let $x^* = \operatorname{argmin}_{x \in \mathbb{R}^d} f(x)$ be the unique minimizer of f. Note that $\nabla f(x^*) = 0$.

The gradient descent (GD) update is given by

$$x_{t+1} = x_t - \alpha \nabla f(x_t) = x_t - \alpha \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_t), \tag{1}$$

which uses the full gradient ∇f. The stochastic gradient descent (SGD) update is given by

$$x_{t+1} = x_t - \alpha \cdot \nabla f_{i_t}(x_t), \tag{2}$$

where i_t is chosen uniformly at random from $\{1, \ldots, n\}$.

From previous lecture, we know that GD converges to an ϵ-optimal solution in $O\left(\kappa \log(1/\epsilon)\right)$ iterations, and each iteration takes $O(n)$ computation since we need to compute the sum of the gradients of n functions (ignoring dependence on d). On the other hand, SGD can be shown to converge in $O\left(\frac{1}{m\epsilon}\right)$ iterations (see Lecture 18, Section 3.3.3), and each iteration takes $O(1)$ computation. The total computation is $O\left(n\kappa \log(1/\epsilon)\right)$ for GD and $O\left(\frac{1}{m\epsilon}\right)$ for SGD.

Can we do better by leveraging the finite-sum structure? Below we show that this is possible and one can achieve an $O\left((n+\kappa)\log(1/\epsilon)\right)$ complexity.

2 Stochastic Variance Reduced Gradient (SVRG)

We want to reduce the variance of the stochastic gradient $\nabla f_i(x)$. One idea is to subtract from $\nabla f_i(x)$ a mean-zero random variable Z that positively correlates with $\nabla f_i(x)$. In this case, we have $\mathbb{E}\left[\nabla f_i(x) - Z\right] = \mathbb{E}\left[\nabla f_i(x)\right] - 0 = \nabla f(x)$, which is still unbiased. Moreover, the new variance $\mathrm{Var}(\nabla f_i(x) - Z) = \mathrm{Var}(\nabla f_i(x)) + \mathrm{Var}(Z) - 2\,\mathrm{cov}(\nabla f_i(x), Z)$ may be smaller than $\mathrm{Var}(\nabla f_i(x))$.

Ideally one may want to subtract $Z = \nabla f_i(x^*) - \nabla f(x^*)$, but we do not know x^*. But we can approximate x^* using the average y of the past iterates. Doing so requires computing the full gradient $\nabla f(y)$, an expensive operation that we will perform only once in a while.

This leads to the SVRG method, given in Algorithm 1.

Algorithm 1 SVRG

input: initial $y^{(1)}$, strong convexity parameter m, smoothness parameter L, stepsize α, number of inner iterations K

for $s = 0, 1, 2, \ldots$

  $x_1^{(s)} = y^{(s)}$

  for $k = 1, \ldots K$

    $x_{k+1}^{(s)} = x_k^{(s)} - \alpha \left(\nabla f_{i_{s,k}}(x_k^{(s)}) - \nabla f_{i_{s,k}}(y^{(s)}) + \nabla f(y^{(s)}) \right),$

    where $i_{s,k} \sim \mathrm{uniform}\{1, \ldots, n\}$

  $y^{(s+1)} = \frac{1}{K} \sum_{k=1}^{K} x_k^{(s)}$

The following lemma quantifies the variance (the second moment to be precise) of the "ideal" stochastic gradient $\nabla f_i(x) - (\nabla f_i(x^*) - \nabla f(x^*))$. In particular, the closer $f(x)$ is to $f(x^*)$, the smaller the variance is.[1] The proof of the lemma exploits the property that smoothness is satisfied by each individual f_i, not just the the overall objective f.

Lemma 1. *Let $i \sim$ uniform $\{1, \ldots, n\}$. We have*

$$\mathbb{E}_i \left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|_2^2 \le 2L \left(f(x) - f(x^*) \right).$$

Proof. By convexity and L-smoothness of f_i, we have

$$\left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|_2^2 \le 2L \left[f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle \right]$$

(we proved this in HW2 Q1.2 as an intermediate step for proving co-coercivity). Taking the expectation of both sides gives

$$\begin{aligned} \mathbb{E}_i \left\| \nabla f_i(x) - \nabla f_i(x^*) \right\|_2^2 &\le 2L \left[\mathbb{E}_i \left[f_i(x) \right] - \mathbb{E}_i \left[f_i(x^*) \right] - \langle \mathbb{E}_i \left[\nabla f_i(x^*) \right], x - x^* \rangle \right] \\ &= 2L \left[f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle \right] \\ &= 2L \left(f(x) - f(x^*) \right), \end{aligned}$$

where the last step holds because $\nabla f(x^*) = 0$. We have proved the lemma. □

We can show that the outer iteration of SVRG achieves geometric convergence.

[1] This is similar to the $B = 0$ case discussed in Section 3.3.2 of Lecture 18. Recall that in this case, geometric convergence can be achieved.

Theorem 1. *Let $f_1, \dots, f_n : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex, and $f = \frac{1}{n}\sum_{i=1}^n f_i$ be m-strongly convex. Then SVRG with stepsize $\alpha = \frac{1}{10L}$ and $K = 20\frac{L}{m}$ satisfies*

$$\mathbb{E}\left[f(y^{(s+1)})\right] - f(x^*) \le 0.9^s \left(f(y^{(1)}) - f(x^*)\right), \quad \forall s.$$

Under the above choice of parameters, each outer iteration of SVRG involves computing the full gradient once ($O(n)$ computation) and computing the stochastic gradient $K = O(\kappa)$ times. Thanks to geometric convergence, the number of outer iterations for achieving ϵ-optimality is $\log(1/\epsilon)$. Consequently, the overall computation is $O\left((n+\kappa)\log(1/\epsilon)\right)$.

Proof. It suffices to show that

$$\mathbb{E}\left[f(y^{(s+1)})\right] - f(x^*) \le 0.9 \left(f(y^{(s)}) - f(x^*)\right), \tag{3}$$

where $y^{(s+1)} = \frac{1}{K}\sum_{k=1}^K x_k^{(s)}$. Below we drop the dependence on the outer index s to simplify notation.

We denote the the variance-reduced stochastic gradient by the shorthand

$$v_k := \nabla f_{i_k}(x_k) - \nabla f_{i_k}(y) + \nabla f(y),$$

so $x_{k+1} = x_k - \alpha v_k$. It follows that

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\alpha\langle v_k, x_k - x^*\rangle + \alpha^2\|v_k\|^2. \tag{4}$$

For the second RHS term above, we have

$$\begin{aligned}
\mathbb{E}_{i_k}\langle v_k, x_k - x^*\rangle &= \left\langle \mathbb{E}_{i_k}\nabla_{i_k} f(x_k) - \mathbb{E}_{i_k}\nabla_{i_k} f(y) + \nabla f(y), x_k - x^*\right\rangle \\
&= \langle \nabla f(x_k), x_k - x^*\rangle && \text{stochastic gradient is unbiased} \\
&\ge f(x_k) - f(x^*). && \text{convexity of } f
\end{aligned}$$

For the last RHS term, we have

$$\begin{aligned}
&\mathbb{E}_{i_k}\|v_k\|^2 \\
\le& 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)\right\|^2 + 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(y) - \nabla f_{i_k}(x^*) - \nabla f(y)\right\|^2 && \because \|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2 \\
=& 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)\right\|^2 \\
&\quad + 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(y) - \nabla f_{i_k}(x^*) - \mathbb{E}_{i_k}\left[\nabla f_{i_k}(y) - \nabla f_{i_k}(x^*)\right]\right\|^2 && \text{unbiased stochastic gradient} \\
\le& 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(x^*)\right\|^2 + 2\mathbb{E}_{i_k}\left\|\nabla f_{i_k}(y) - \nabla f_{i_k}(x^*)\right\|^2 && \mathbb{E}\|X - \mathbb{E}X\|^2 \le \mathbb{E}\|X\|^2 \\
\le& 4L\left(f(x_k) - f(x^*)\right) + 4L\left(f(y) - f(x^*)\right). && \text{Lemma 1}
\end{aligned}$$

Plugging these bounds into (4), we get

$$\mathbb{E}_{i_k}\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 - 2\alpha(1 - 2\alpha L)\left(f(x_k) - f(x^*)\right) + 4\alpha^2 L\left(f(y) - f(x^*)\right).$$

Summing over $k = 1, \dots, K$ and taking expectation w.r.t. all $i_1, \dots i_K$, we obtain

$$0 \le \mathbb{E}\|x_{K+1} - x^*\|^2 \le \mathbb{E}\|x_1 - x^*\|^2 - 2\alpha(1 - 2\alpha L)\,\mathbb{E}\sum_{k=1}^K\left(f(x_k) - f(x^*)\right) + 4\alpha^2 LK\left(f(y) - f(x^*)\right). \tag{5}$$

Recall that $x_1 = y$. By m-strong convexity of f we have

$$\|x_1 - x^*\|^2 \le \frac{2}{m}\left(f(x_1) - f(x^*)\right) = \frac{2}{m}\left(f(y) - f(x^*)\right).$$

By convexity Jensen's we have

$$\frac{1}{K}\sum_{k=1}^{K}\left(f(x_k) - f(x^*)\right) \ge f\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) - f(x^*).$$

Combining the above two equations with (5), we obtain

$$0 \le \frac{2}{m}\left(f(y) - f(x^*)\right) - 2\alpha\left(1 - 2\alpha L\right)K\left[f\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) - f(x^*)\right] + 4\alpha^2 LK\left(f(y) - f(x^*)\right).$$

Rearranging gives

$$f\left(\frac{1}{K}\sum_{k=1}^{K} x_k\right) - f(x^*) \le \left[\frac{1}{m\alpha(1-2\alpha L)K} + \frac{2\alpha L}{1-2\alpha L}\right]\left(f(y) - f(x^*)\right).$$

With $\alpha = \frac{1}{10L}$ and $K = 20\frac{L}{m}$, the expression inside the square bracket becomes 0.9, which proves the desired inequality (3). □