Lecture notes on Stochastic Variance Reduced Methods.

Robert M. Gower

February 5, 2019

Abstract

Lecture notes on variance reduction techniques.

Contents

1 Introduction

Consider the following optimziation problem

$$w^* = \arg \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n f_i(w) =: f(w), \tag{1}$$

where f is L–smooth, λ–strongly convex and f_i is convex and L_i–smooth for $i = 1, \ldots, n$. In other words

$$f(y) + \langle \nabla f(y), w - y \rangle + \frac{\lambda}{2} ||w - y||_2^2 \leq f(w) \leq f(y) + \langle \nabla f(y), w - y \rangle + \frac{L}{2} ||w - y||_2^2, \tag{2}$$

and

$$f_i(w) \leq f_i(y) + \langle \nabla f_i(y), w - y \rangle + \frac{L_i}{2} ||w - y||_2^2, \quad \text{for } i = 1, \ldots, n. \tag{3}$$

In last weeks lecture we saw that using Stochastic gradient descent (SGD) to solve (1) can converge much faster in the early iterations as compared to the gradient descent algorithm. But in later iterations the SGD algorithm slows down and thus struggles to reach an accurate solution. Can we get the best of both the fast initial convergence of SGD and the steady linear convergence of gradient descent? Yes we can! The trick is to solve SGD's issues with variance.

What are SGD's issues with variance? Though the stochastic gradient is an unbiased estimator of the gradient, it may have high variance. Indeed, to analyse SGD we had to start by imposing the rather awkward following assumption: That there exits $B > 0$ such that

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD.}$$

Even with the above assumption, we required decreasing stepsizes to gradually kill off the variance. Yet another glaring issue with SGD is that even if we start the SGD algorithm on the solution $w^* = w^0$, the method will not stop. This is because the stochastic gradients are not necessarily zero on the solution, that is $\nabla f_i(w^*) \neq 0$ is entirly possible. While $\nabla f(w^*) = 0$, thus gradient descent will stop once it has reached the solution.

In these notes we set out to describe methods that fix the above issues. Our aim is to have an iterative algorithm of the form

$$w^{t+1} = w^t - \alpha g^t, \tag{4}$$

where $\alpha > 0$ is a stepsize and g^t is an estimate of the gradient that satisfies

$$\begin{aligned}
&\text{Unbiased:} & \mathbf{E}\left[g^t\right] &= \nabla f(w^t) & (5)\\
&\text{Reducing Variance:} & \mathbf{E}\left[\|g^t\|_2^2\right] &\to_{w^t \to w^*} 0. & (6)
\end{aligned}$$

Note that the

$$\mathbf{VAR}\left[g^t\right] = \mathbf{E}\left[\|g^t\|\right] - \|\nabla f(w^t)\|_2^2.$$

Consequently if (6) holds, then the variance of g^t also tends to zero as w^t tends to w^*.

Our main tool for building an estimate of the gradient that satisfies the above will be covariates.

2 Covariates

Let x be a random variable. We say that a random variable z is a covariate of x if $\text{cov}[x, z] > 0$. We can use the covariate z to build an unbiased estimator of x that has a small variance. Indeed let

$$x_z = x - z + \mathbf{E}[z],$$

and note that $\mathbf{E}[x_z] = \mathbf{E}[x]$. Furthermore

$$\mathbf{VAR}[x_z] = \mathbf{VAR}[x] + \mathbf{VAR}[z] - 2\,\text{cov}[x, z].$$

Consequently if $\text{cov}[x, z]$ is sufficiently large, then $\mathbf{VAR}[x_z]$ is small. We can build an estimate of the gradient with reduced variance by finding covariates for the stochastic gradient.

3 The Stochastic Variance Reduced (SVRG) method

How do we choose this reference point $\tilde{w}$? One strategy is to choose $\tilde{w}$ as a past iterate

Let $w^k \in \mathbb{R}^d$ be our current iterate and let $\tilde{w}^t \in \mathbb{R}^d$ be a *reference point.* If w^k is sufficiently close to $\tilde{w}^t$ it is reasonable to expect that $\nabla f_i(w^k)$ and $\nabla f_i(\tilde{w}^t)$ are covariates for every $i = 1, \ldots n$. Consequently, if $i \in \{1, \ldots, n\}$ is sampled uniformly then

$$g^k = \nabla f_i(w^k) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t), \tag{7}$$

is an unbiased estimate of the gradient with reduced variance.

Before convergence, we need the following three Lemmas.

Lemma 1 *If f is an L–smooth function then*

$$f(y - \frac{1}{L}\nabla f(y)) - f(y) \leq -\frac{1}{2L}\nabla f(y). \tag{8}$$

Proof: Setting $w = y - \frac{1}{L}\nabla f(y)$ in the right hand of (2) gives

$$f(y - \frac{1}{L}\nabla f(y)) - f(y) \leq \langle \nabla f(y), -\frac{1}{L}\nabla f(y)\rangle + \frac{L}{2}|| - \frac{1}{L}\nabla f(y)||_2^2 = -\frac{1}{2L}\nabla f(y). \quad \blacksquare$$

Lemma 2 *If each f_i is L_i–smooth then*

$$\mathbf{E}\left[\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2\right] \leq 2L_{\max}(f(w) - f(w^*)). \tag{9}$$

Proof: Let $g_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle$ which is L_i–smooth. By the convexity of f_i we have that $g_i(w) \geq 0$ for all w. From (8) we have that

$$-g_i(w) \overset{g_i(w - \frac{1}{L_i}\nabla g_i(w)) \geq 0}{\leq} g_i(w - \frac{1}{L_i}\nabla g_i(w)) - g_i(w) \leq -\frac{1}{2L_i}\|\nabla g_i(w)\|_2^2 \leq -\frac{1}{2L_{\max}}\|\nabla g_i(w)\|_2^2.$$

By substituting $g_i(w) = f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle$ the above can be re-written as

$$\frac{1}{2L_{\max}}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \leq f_i(w) - f_i(w^*) - \langle \nabla f_i(w^*), w - w^*\rangle.$$

Taking expectation with respect to i and using that $\frac{1}{n}\sum_{i=1}^n \nabla f_i(w^*) = \nabla f(w^*) = 0$ gives the result. $\blacksquare$

Lemma 3 *The second moment of the SVRG gradient estimate is bounded*

$$\mathbf{E}\left[\|g^t\|_2^2\right] \leq 4L_{\max}(f(w^k) - f(w^*)) + 2L_{\max}(f(\tilde{w}^t) - f(w^*)). \tag{10}$$

Proof:

$$\begin{aligned}
\mathbf{E}\left[\|g^k\|_2^2\right] &\leq \mathbf{E}\left[\|\nabla f_i(w^k) - \nabla f_i(w^*) + \nabla f_i(w^*) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)\|_2^2\right] \\
&= 2\mathbf{E}\left[\|\nabla f_i(w^k) - \nabla f_i(w^*)\|_2^2\right] + 2\|\nabla f_i(w^*) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w}^t)\|_2^2 \\
&\leq 2\mathbf{E}\left[\|\nabla f_i(w^k) - \nabla f_i(w^*)\|_2^2\right] + 2\|\nabla f_i(w^*) - \nabla f_i(\tilde{w}^t)\|_2^2 \\
&\overset{(9)}{\leq} 4L_{\max}(f(w^k) - f(w^*)) + 4L_{\max}(f(\tilde{w}^t) - f(w^*)).
\end{aligned}$$

Where we used in the first inequality that $\mathbf{E}\left[\|X-\mathbf{E}[X]\|_2^2\right] \leq \mathbf{E}\left[\|X\|_2^2\right]$ with $X = \nabla f_i(w^*) - \nabla f_i(\tilde{w}^t) + \nabla f(\tilde{w})$.

Finally we prove the convergence of the SVRG method in the following theorem.

Theorem 4 *If we choose the stepsize $\alpha = 1/10L_{\max}$ and the number of inner iterations as $m = \lambda/L_{\max}$ then the SVRG method (7) converges according to*

$$\mathbb{E}[f(\tilde{w}^t)] - f(w^*) \quad \leq \quad 0.9^t(f(\tilde{w}^0) - f(w^*)). \tag{11}$$

Proof: First note that

$$\begin{aligned}
\mathbf{E}_j\left[\|w^{k+1}-w^*\|_2^2\right] &= \|w^k-w^*\|_2^2 - 2\alpha\left\langle \nabla f(w^k), w^k-w^*\right\rangle + \mathbf{E}_j\left[\|g^k\|_2^2\right] \\
&\leq \|w^k-w^*\|_2^2 - 2\alpha(f(w^k)-f(w^*)) + \mathbf{E}_j\left[\|g^k\|_2^2\right] \\
&\overset{(10)}{\leq} \|w^k-w^*\|_2^2 - 2\alpha(1-2\alpha L_{\max})(f(w^k)-f(w^*)) + 4\alpha L_{\max}(f(\tilde{w}^t)-f(w^*)).
\end{aligned}$$

Taking total expectation, summing up over $k = 0\ldots m-1$ and using telescopic cancellation we have that

$$\begin{aligned}
\mathbb{E}\left[||w^m-w^*||_2^2\right] \leq\ & \mathbb{E}\left[||w^0-w^*||_2^2\right] - 2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k)-f(w^*))] \\
& +4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right].
\end{aligned}$$

Using that $w^0 = \tilde{w}^t$, strong convexity $f(\tilde{w}^t) - f(w^*) \geq \frac{\lambda}{2}||\tilde{w}^t - w^*||_2^2$ and re-arranging we have that

$$\begin{aligned}
2\alpha(1-2\alpha L_{\max})\mathbb{E}[\textstyle\sum_{k=0}^{m-1}(f(w^k)-f(w^*))] &\leq \mathbb{E}\left[||\tilde{w}^t-w^*||_2^2\right] - \mathbb{E}\left[||w^m-w^*||_2^2\right] \\
&\quad +4m\alpha^2 L_{\max}\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right] \\
&\leq (4m\alpha^2 L_{\max} + \tfrac{2}{\lambda})\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]
\end{aligned}$$

Re-arranging again and using Jensen's inequality we have

$$\begin{aligned}
\mathbb{E}[f(\textstyle\sum_{k=0}^{m-1}\frac{w^k}{m})] - f(w^*) &\leq \tfrac{1}{m}\mathbb{E}[\textstyle\sum_{k=0}^{m-1} f(w^k)] - f(w^*) \\
&\leq \tfrac{4m\alpha^2 L_{\max}+2\lambda^{-1}}{2\alpha(1-2\alpha L_{\max})m}\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right] \\
&= \left(\tfrac{2\alpha L_{\max}}{1-2\alpha L_{\max}} + \tfrac{1}{\lambda\alpha(1-2\alpha L_{\max})m}\right)\mathbb{E}\left[f(\tilde{w}^t)-f(w^*)\right]
\end{aligned}$$

It now remains to substitute $\alpha = 1/10L_{\max}$ and $m = 20L_{\max}/\gamma$ to see that

$$\frac{2\alpha L_{\max}}{1-2\alpha L_{\max}} + \frac{1}{\lambda\alpha(1-2\alpha L_{\max})m} = \frac{2/10}{1-2/10} + \frac{1}{2(1-2/10)} = \frac{2}{8} + \frac{5}{8} = \frac{7}{8}. \quad \blacksquare$$

Algorithm 1 Stochastic Variance reduced gradient

```
**Parameters** number of inner iterations $m$ and learning rate $\alpha$.
Choose $\tilde{w}_0$.
**for** $t = 1, 2, \ldots$ **do**
    $\tilde{w} = \tilde{w}_{t-1}$.
    Compute the full gradient $\nabla f(\tilde{w})$.
    Set $w_0 = \tilde{w}$.
    **for** $k = 0, \ldots, m-1$ **do**
        Sample $i_k$ from $\{1, \ldots, n\}$.
        $g_k(w^k) = \nabla f_{i_k}(w^k) - \nabla f_{i_k}(\tilde{w}) + \nabla f(\tilde{w})$.
        Update $w^{k+1} = w^k - \alpha g_k(w^k)$.
    Choose the following reference point $\tilde{w}_s$, according to the options below.
    **Option Last:** $\tilde{w}_t = w_m$.
    **Option Average:** Choose $\tilde{w}_t$ such that $\tilde{w}_t = \frac{1}{m}\sum_{i=0}^{m-1} w^i$
```