Lecture Notes 14: Variance Reduction

Instructor: Ashok Cutkosky

For most of this course we have considered fairly generic stochastic optimization problems. Now, we will consider a particular kind of problem that is of great interest in machine learning: the so-called *finite-sum* objectives:

$$\mathcal{L}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \ell(\mathbf{w}, z_i)$$

For example, the *training loss* for most standard machine learning problems takes this form, where $z_1, \dots, z_N$ represent the training set and $\ell(\mathbf{w}, z)$ is the loss on a training example. Let's consider the case that $\mathcal{L}$ is convex and H-smooth. Without invoking complicated procedures like acceleration, there are roughly two ways one might consider to apply gradient descent in this setting:

1. Use *full gradient* descent: that is, compute a gradient $\nabla\mathcal{L}(\mathbf{w}_t)$ and take a gradient descent step $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\mathcal{L}(\mathbf{w}_t)$ for $\eta = \frac{1}{H}$. This guarantees:

$$\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star) \leq O(1/T)$$

2. Use *stochastic* gradient descent: sample a training point z_t at random and take a step $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla\ell(\mathbf{w}_t, z_t)$. Since $\mathbb{E}[\nabla\ell(\mathbf{w}_t, z_t)] = \nabla\mathcal{L}(\mathbf{w}_t)$, by setting $\eta = O(1/\sqrt{T})$, we can guarantee:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq O(1/\sqrt{T})$$

As previously discussed, although the dependence on T is worse for the SGD approach, the total computation scales much better with N: every single iterate of the full gradient approach requires $O(N)$ computation, but every iterate of the SGD approach requires $O(1)$ computation. So in terms of total compute C, we are comparing $O(N/C)$ to $O(1/sqrt(C))$. As long as $N^2 \geq C$, we would prefer the stochastic approach. Even if we allow for accelerated gradient descent, it turns out that regular SGD is still better for sufficiently large N.

Now, the natural question is: can we improve at all on stochastic gradient descent? Surprisingly, by leveraging the finite-sum structure of the problem, the answer is *yes*! The key idea is to create a better gradient estimator than simply sampling one element at random. The method we discuss here has been found in various forms at around the same time, but arguable the simplest presentation is [1]. Other essentially concurrent methods include [2, 3].

We will create a better gradient estimate by *occasionally* computing a full gradient. Intuitively, if we compute one single full-gradient N iterations, we will still on average use only $O(1)$ computation per iteration. By some clever trickery, we will be able to leverage this perfectly-accurate gradient to improve subsequent gradients.

Let's suppose that we know the true gradient $\nabla\mathcal{L}(\mathbf{v})$ at some point $\mathbf{v}$ which we will call the *checkpoint* value. We wish now to compute an estimate of the gradient at a *nearby* point $\mathbf{w}$. To form the estimate, we randomly sample some z, and compute:

$$\mathbf{g} = \nabla\ell(\mathbf{w}, z) - \nabla\ell(\mathbf{v}z) + \nabla\mathcal{L}(\mathbf{v})$$

The first important property of this estimate is that it is *unbiased*:

$$\begin{aligned}\mathbb{E}[\mathbf{g}] &= \mathbb{E}[\nabla\ell(\mathbf{w}, z) - \nabla\ell(\mathbf{v}z) + \nabla\mathcal{L}(\mathbf{v})] \\ &= \nabla\mathcal{L}(\mathbf{w}) - \nabla\mathcal{L}(\mathbf{v}) + \nabla\mathcal{L}(\mathbf{v}) \\ &= \nabla\mathcal{L}(\mathbf{w})\end{aligned}$$

However, we can hope that the *variance* of $\mathbf{g}$ is rather small. Why might this be? Notice that the only random part of $\mathbf{g}$ is the difference $\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z)$. Now, the function ℓ is a smooth function of $\mathbf{w}$ for all z, then we have:

$$\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z)\| \leq H\|\mathbf{w} - \mathbf{v}\|$$

so that if $\mathbf{w} \approx \mathbf{v}$, the randomness in $\mathbf{g}$ is necessarily quite small. Note however that we have subtly changed assumptions here: *assuming that ℓ is smooth is a stronger assumption than assuming that $\mathcal{L}$ is smooth.*

In fact, by leveraging convexity, we can provide a somewhat more subtle bound on the variance of $\mathbf{g}$. To this end, we have the following Lemma:

Lemma 1. *Suppose $\ell(\mathbf{w}, z)$ is an H-smooth convex function of $\mathbf{w}$ for all z. Then:*

$$\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v})\|^2] \leq 8H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star)) + 4H(\mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{w}_\star))$$

Proof. Recall from previous lectures the fact that for all x, and all H-smooth functions f with $x_\star = \operatorname{argmin} f(x)$:

$$\|\nabla f(x)\|^2 \leq 2H(f(x) - f(x_\star)) \tag{1}$$

Let's consider the function $f(x) = \ell(x, z) - \langle \nabla \ell(\mathbf{w}_\star, z), x - \mathbf{w}_\star \rangle$. Notice that since ℓ is H-smooth and convex, so is f (because we only added a linear function to ℓ). However, also we have that $\nabla f(\mathbf{w}_\star) = \nabla \ell(\mathbf{w}_\star) - \nabla \ell(\mathbf{w}_\star) = 0$, so that f is actually minimized at $\mathbf{w}_\star$ (which is the argmin of $\mathcal{L}$, not ℓ). Therefore:

$$\|\nabla f(x)\|^2 \leq 2H(\nabla f(x) - f(\mathbf{w}_\star))$$
$$\|\nabla \ell(x, z) - \nabla \ell(\mathbf{w}_\star, z)\|^2 \leq 2H(\ell(x, z) - \ell(\mathbf{w}_\star, z) - \langle \nabla \ell(\mathbf{w}_\star, z), x - \mathbf{w}_\star \rangle)$$

Further, by a more direct application of (1):

$$\|\nabla \mathcal{L}(\mathbf{w})\|^2 \leq 2H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star))$$

Now, we can compute:

$$\begin{aligned}
\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v})\|^2] &= \mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v}) - \nabla \mathcal{L}(\mathbf{w}) + \nabla \mathcal{L}(\mathbf{w})\|^2] \\
&\leq 2\,\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v}) - \nabla \mathcal{L}(\mathbf{w})\|^2 + \|\nabla \mathcal{L}(\mathbf{w})\|^2] \\
&\leq 2\,\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v}) - \nabla \mathcal{L}(\mathbf{w})\|^2] + 4H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star)
\end{aligned}$$

Now, notice that $\mathbb{E}[\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z)] = \nabla \mathcal{L}(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{v})$. Further, for any random variable X, $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$. Therefore the first term in the above expression can be bounded:

$$\begin{aligned}
\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z) + \nabla \mathcal{L}(\mathbf{v}) - \nabla \mathcal{L}(\mathbf{w})\|^2] &\leq \mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{v}, z)\|^2] \\
&= \mathbb{E}[\|(\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{w}_\star, z)) - (\nabla \ell(\mathbf{v}, z) - \nabla \ell(\mathbf{w}_\star, z))\|^2] \\
&\leq 2\,\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \ell(\mathbf{w}_\star, z)\|^2 + \|\nabla \ell(\mathbf{v}, z) - \nabla \ell(\mathbf{w}_\star, z)\|^2] \\
&\leq 4H\,\mathbb{E}\left[\ell(\mathbf{w}, z) - \ell(\mathbf{w}_\star, z) + \ell(\mathbf{v}, z) - \ell(\mathbf{w}_\star, z) - \langle \nabla \ell(\mathbf{w}_\star, z), \mathbf{w} + \mathbf{v} - 2\mathbf{w}_\star \rangle\right] \\
&= 4H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) + \mathcal{L}(\mathbf{v}) - \mathcal{L}(\mathbf{w}_\star))
\end{aligned}$$

□

This lemma tells us that the estimate $\mathbf{g}$ has low variance if both $\mathbf{w}$ and $\mathbf{v}$ have small function value. We can use this fact to set up a kind of "virtuous cycle": using estimates $\mathbf{g}$, we run N steps of SGD to obtain iterates with smaller function value. Then, we choose a new checkpoint $\mathbf{v}$ from these iterates and recompute the new $\nabla \mathcal{L}(\mathbf{v})$, and then run N more iterates of SGD. Now, as SGD converges, $\mathbf{w}$ and $\mathbf{v}$ will obtain smaller and smaller objective values, which will make the value of $\mathbb{E}[\|\mathbf{g}\|^2]$ small, which will make SGD convergence even faster, making $\mathbb{E}[\|\mathbf{g}\|^2]$ even smaller, and so on.

The algorithm is presented in Algorithm 1:

Algorithm 1 SVRG

Input: Initial Point $\mathbf{w}_1$, learning rate η.
Set initial checkpoint $\mathbf{v}_1 = \mathbf{w}_1$.
Compute $\nabla\mathcal{L}(\mathbf{v}_1)$ (takes $O(N)$ time).
Set $e = 1$ //e is an "epoch" counter.
for $t = 1 \dots T$ **do**
  Sample z_t at random from $\{z_1, \dots, z_N\}$.
  Form $\mathbf{g}_t = \nabla\ell(\mathbf{w}_t, z_t) - \nabla\ell(\mathbf{v}_e, z_t) + \nabla\mathcal{L}(\mathbf{v}_e)$
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{g}_t$.
  if $t \equiv 0 \mod N$ **then**
    //If we've finished one epoch, pick a new $\mathbf{v}$.
    $e = e + 1$.
    $\mathbf{v}_e = \frac{1}{N}\sum_{i=t-N+1}^{t} \mathbf{w}_i$
    Compute $\nabla\mathcal{L}(\mathbf{v}_e)$ (takes $O(N)$ time).
  end if
end for

Theorem 2. *Suppose $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\ell(\mathbf{w}, z_i)$ is such that each $\ell(\mathbf{w}, z)$ is H-smooth and convex in $\mathbf{w}$ for all z_i. Suppose T is a multiple of N. Suppose $\eta = \frac{c}{\sqrt{N}}$ and η satisfies $\eta \le \frac{1}{8H}$. Then Algorithm 1 guarantees:*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \le \frac{\sqrt{N}[8cH(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)) + \frac{2}{c}\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{T}$$

Let's take a minute to interpret this result. First, since Algorithm 1 only performs one $O(N)$ full-gradient computation every N iterations, and normally each iteration takes 2 gradient computations, the total computation used by Algorithm 1 is still $O(T)$, same as it would be for SGD. Therefore, if we use C for the total computation, we need to compare the following numbers:

1. $O\left(\frac{N}{C}\right)$ for full gradient descent.
2. $O\left(\frac{1}{\sqrt{C}}\right)$ for SGD.
3. $O\left(\frac{\sqrt{N}}{C}\right)$ for SVRG.

Now, it is clear here that SVRG always beats full gradient descent. But what about SGD? Let's consider the case that C is fairly large compared to N, say $O(N^{3/2})$. Then the bounds are $O(1/N^{3/4})$ for SGD compared with $O(1/N)$ for SVRG. Thus, SVRG clearly dominates in this case. However, it's also intuitive that SVRG may not do as well when N is large compared to C. By inspecting the equations, we can see that SVRG will always beat SGD (up to constants) so long as $N \le C$. This is extremely mild: essentially we are asking that we need to look at every point in the dataset at once. Anything less is essentially throwing out data, so the situation in which SGD is better is a bit of a degenerate case.

Now, let's give the proof of Theorem 2

Proof. The proof is similar to our bound for smooth convex losses (after reading the proof, you might want to take a minute to think about why we don't need projections here unlike some of our other convex analyses).

We will use e_t to indicate the current value of e at iteration t. So $e_t = \lceil \frac{t}{N} \rceil$.

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &= \|\mathbf{w}_t - \eta\mathbf{g}_t - \mathbf{w}_\star\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta\langle\mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle + \eta^2\|\mathbf{g}_t\|^2 \\
\langle\mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle &\le \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta\|\mathbf{g}_t\|^2}{2} \\
\mathbb{E}[\langle\mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_\star\rangle] &\le \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta\|\mathbf{g}_t\|^2}{2}\right]
\end{aligned}$$

Use the fact that $\mathbb{E}[\mathbf{g}_t] = \nabla\mathcal{L}(\mathbf{w}_t)$:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta\|\mathbf{g}_t\|^2}{2}\right]$$

$$\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta\|\mathbf{g}_t\|^2}{2}\right]$$

$$\leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{g}_t\|^2]$$

now we use Lemma 1:

$$\leq \frac{\eta\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}[8H(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)) + 4H(\mathcal{L}(\mathbf{v}_{e_t}) - \mathcal{L}(\mathbf{w}_\star))]$$

Using $\eta \leq \frac{1}{8H}$:

$$\leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] + \frac{N}{4}\sum_{e=2}^{T/N}\mathbb{E}[\mathcal{L}(\mathbf{v}_e) - \mathcal{L}(\mathbf{w}_\star)] + 2\eta HN\,\mathbb{E}[\mathcal{L}(\mathbf{v}_1) - \mathcal{L}(\mathbf{w}_\star)]$$

Now, let's take a minute to understand $\mathcal{L}(\mathbf{v}_e) - \mathcal{L}(\mathbf{w}_\star)$. By Jensen inequality, for all $e > 1$, we have:

$$\mathcal{L}(\mathbf{v}_e) \leq \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\mathbf{w}_{e(N-1)+1+i})$$

Therefore:

$$N\sum_{e=2}^{T/N-1}\mathbb{E}[\mathcal{L}(\mathbf{v}_e) - \mathcal{L}(\mathbf{w}_\star)] \leq \sum_{e=2}^{T/N-1}\sum_{i=1}^{N}\mathcal{L}(\mathbf{w}_{(e-1)N+i}) - \mathcal{L}(\mathbf{w}_\star)$$

$$= \sum_{t=N+1}^{T(N-1)/N}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)]$$

$$\leq \sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)]$$

Further, for $e = 1$, $\mathbf{v}_e = \mathbf{w}_1$. Therefore, we have:

$$\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} + \frac{3}{4}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] + 2\eta HN\,\mathbb{E}[\mathcal{L}(\mathbf{v}_1) - \mathcal{L}(\mathbf{w}_\star)]$$

$$\frac{1}{4}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta} + 2\eta HN\,\mathbb{E}[\mathcal{L}(\mathbf{v}_1) - \mathcal{L}(\mathbf{w}_\star)]$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{2\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{\eta T} + \frac{8\eta HN(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{T}$$

Now finally plugging in $\eta = \frac{c}{\sqrt{N}}$ concludes the result. □

References

[1] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in neural information processing systems* 26 (2013), pp. 315–323.

[2] Nicolas Le Roux, Mark Schmidt, and Francis Bach. "A stochastic gradient method with an exponential convergence rate for finite training sets". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 2*. 2012, pp. 2663–2671.

[3] Shai Shalev-Shwartz and Tong Zhang. "Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization." In: *Journal of Machine Learning Research* 14.2 (2013).