Lecture 11

Last time

- Missing lemmas
- SGD
- Expected descent

Today

- Reaching a critical point
- PL is back
- Variance Reduction

SGD reaches criticality

Let's continue with the proof of Lemma (ED).

Lemma (ED). Suppose Assumption (SGD) holds. Then, at each k we have

(a) $$\mathbb{E}[f(w_{k+1}) \mid \mathcal{F}_k] \leq f(w_k) - \eta_k\left(1 - \frac{L\eta_k}{2}\right)\|\nabla f(w_k)\|^2 + \frac{L\eta_k^2\sigma^2}{2}.$$

Proof: Using Taylors Theorem

(c) $$f(w_{k+1}) \leq f(w_k) - \eta_k\langle \nabla f(w_k), \underbrace{g_k}_{g(w_k,z_k)}\rangle + \frac{L\eta_k^2}{2}\|g_k\|^2.$$

Let's bound the norm $\|g_k\|^2$ in expectation

(b) $$\mathbb{E}[\|g_k\|^2 \mid \mathcal{F}_k]$$

$$= \mathbb{E}[\|\underbrace{g_k - \nabla f(w_k)}_{\xi_k} + \nabla f(w_k)\|^2 \mid \mathcal{F}_k]$$

(Linearity) (pull-out)

$$= \|\nabla f(w_k)\|^2 + 2\langle \nabla f(w_k), \mathbb{E}[\xi_k | \mathcal{F}_k]\rangle + \mathbb{E}[\|\xi_k\|^2 | \mathcal{F}_k]$$

(unbiased)

$$= \|\nabla f(w_k)\|^2 + \mathbb{E}[\|\xi_k\|^2 | \mathcal{F}_k] \leq \|\nabla f(w_k)\|^2 + \sigma^2.$$

Combining (i) and (ii) gives the result. □

We are now ready to establish a guarantee.

Theorem (SGD to criticality). Suppose that Assumption (SGD) holds. Then, the iterates of (SGD) with $\eta_k = \eta \in (0, 1/L]$ satisfy that for any $T \geq 1$,

$$(o) \quad \frac{1}{T}\sum_{k=0}^{T-1} \mathbb{E}\|\nabla f(w_k)\|^2 \leq \frac{2(f(w_0) - f^*)}{\eta T} + L\eta\sigma^2$$

Tension between controlling noise and convergence.

Proof: Taking total expectation in (N), and applying the tower law gives

$$\mathbb{E}[f(w_{k+1})] \le \mathbb{E}[f(w_k)] - \frac{\eta}{2}\mathbb{E}\|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2}.$$

Rearranging and summing up to $T-1$,

$$\frac{\eta}{2}\sum \mathbb{E}\|\nabla f(w_k)\|^2 \le \sum_{k=0}^{T-1}(\mathbb{E}f(w_k) - \mathbb{E}f(w_{k+1})) + \frac{TL\eta^2\sigma^2}{2}$$

$$= f(w_0) - \mathbb{E}f(w_T) + \frac{TL\eta^2\sigma^2}{2}$$

Dividing by $\eta T/2$ gives the result. □

Corollary: Suppose Assumption (SGD) holds. Then, for a given time horizon $T \ge 1$, the iterates of (SGD) with

$$\eta_k = \frac{1}{L\sqrt{T}}$$

satisfy

$$\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\|\nabla f(w_k)\| \le \frac{\sqrt{2L(f(w_0)-f^*) + \sigma^2}}{T^{1/4}}.$$

The proof follows from (0) and Jensen's.

If we take $R \sim \text{Unif}[T-1]$, then the same upper bound applies for $\mathbb{E}\|\nabla f(w_R)\|$.

This guarantee is much worse than what we got when $\sigma^2 = 0$, namely

$$\frac{\sqrt{2L(f(w_0) - f^*)}}{T}.$$

Also it requires knowing T. If we want an any-time guarantee, we might take $\eta_k = \frac{1}{L\sqrt{k+1}}$, (why?).

PL is back

Recall that we proved much better guarantee under the stronger, PL inequality:

(PL) $$\frac{1}{2}\|\nabla f(w)\|^2 \geq \mu(f(w) - f^*).$$

The next result shows that although convergence can be faster, the noise

floor doesn't change much.

Theorem (SGD under PL). Suppose that Assumption (SGD) and (PL) hold. Then, the iterates of (SGD) with $\eta_k = \eta \in (0, 1/L)$ satisfy that for every $T \geq 0$,

$$\mathbb{E}[f(w_T) - f^*] \leq (1 - \mu\eta)^T (f(w_0) - f^*) + \frac{L\eta\sigma^2}{2\mu}.$$

Here is that tension again →

Proof: Taking total expectation in (A) and subtracting f^* yields

$$\underbrace{\mathbb{E}[f(w_{k+1}) - f^*]}_{\delta_{k+1}} \leq \mathbb{E}[f(w_k) - f^*] - \frac{\eta}{2}\mathbb{E}\|\nabla f(w_k)\|^2 + \frac{L\eta^2\sigma^2}{2}$$

PL → $\leq (1 - \eta\mu)\,\mathbb{E}[f(w_k) - f^*] + \frac{L\eta^2\sigma^2}{2}.$

Recursively applying this inequality

yields

$$\delta_T \le (1-\eta\mu)^T \delta_0 + \frac{L\eta^2\sigma^2}{2} \sum_{j=0}^{T-1} (1-\eta\mu)^j$$

$$= (1-\eta\mu)^T \delta_0 + \frac{L\eta^2\sigma^2 \left(1-(1-\eta\mu)^T\right)}{2\left(1-(1-\eta\mu)\right)}$$

$$\le (1-\eta\mu)^T \delta_0 + \frac{L\eta\sigma^2}{2\mu}.$$

□

Why can't we surpass the extra noise term with a constant step size?

To see exactly why, let's consider a finite sum example, i.e.,

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w). \tag{2}$$

Suppose we have a sequence

$$\nabla f(w_k) \to 0.$$

Even in that case, we don't need to see oracles that go to zero,

e.g., we might have that $\forall i \in [n]$

$$\nabla f_i(w_k) \not\to 0.$$

Thus, even though the average goes to zero, the individual components might not (they cancel each other, but aren't zero). Next we will see how to bypass this issue.

Variance reduction

This section only applies to finite sums as in (Σ). For this type of losses we could feasibly compute a gradient, yet it might be expensive. The key idea of variance reduction is to compute one full gradient every so often and note

$$\nabla f(w) = \nabla f(\tilde{w}) + \frac{1}{n} \sum_{i=1}^{n} (\nabla f_i(w) - \nabla f_i(\tilde{w}))$$

$\tilde{w}$: Nearby point

$\nabla f_i(w) - \nabla f_i(\tilde{w})$: Shouldn't grow too much if ∇f_i is Lipschitz

This naturally suggests a gradient estimate

$$g(w, i; \tilde{w}) = \nabla f(\tilde{w}) + \nabla f_i(w) - \nabla f_i(\tilde{w}).$$

Then, it is immediate that

$$\mathbb{E}\, g(w, i; \tilde{w}) = \nabla f(w).$$

So when would this estimator be less noisy? Suppose for instance that each ∇f_i is L-Lipschitz. Define

$$A_i = \nabla f_i(w) - \nabla f_i(\tilde{w})$$

then, $g(w, i; \tilde{w}) - \nabla f(w) = A_i - \mathbb{E} A_i$. So,

$$\begin{aligned}
\mathbb{E}\|g(w, i, \tilde{w} - \nabla f(w)\|^2 &= \mathbb{E}\, \|A_i - \mathbb{E} A_i\|^2 \\
&= \mathbb{E}\|A_i\|^2 - 2\langle \mathbb{E} A_i, \mathbb{E} A_i\rangle + \|\mathbb{E} A_i\|^2 \\
&= \mathbb{E}\,\|A_i\|^2 - \|\mathbb{E} A_i\|^2 \\
&\le \mathbb{E}\,\|A_i\|^2 \\
&= \mathbb{E}\,\|\nabla f_i(w) - \nabla f_i(\tilde{w})\|^2 \\
&\le L\,\|w - \tilde{w}\|^2.
\end{aligned}$$

Could be really small if we are close to $\tilde{w}$

This suggests a natural iterative method

Set $\tilde{w}_0 = w_k$

Compute $\nabla f(\tilde{w}_0)$.

For $t = 0, \ldots, T-1$

(SVRG) Draw a fresh $j \sim \text{Unif}[n]$

$\tilde{w}_{t+1} \leftarrow \tilde{w}_t - \eta\, g(\tilde{w}_t, j; \tilde{w}_0)$

Update $w_{k+1} \leftarrow \frac{1}{T} \sum_{t=0}^{T-1} \tilde{w}_t$.

Next class we will analyze this method.