Lecture 6: SGD Continued, Minibatching, and Learning Rates

CS4787 — Principles of Large-Scale Machine Learning Systems

Where we left off: we looked at how stochastic gradient descent performs on non-convex objectives. But what happens in the "nice" case when we assume convexity?

Stochastic gradient descent for strongly convex objectives. Recall that the update step for SGD is

$$w_{t+1} = w_t - \alpha_t \nabla f_{\tilde{i}_t}(w_t).$$

As before, we start with Taylor's theorem. From Taylor's theorem, there exists a ξ_t such that

$$\begin{aligned}
f(w_{t+1}) &= f(w_t - \alpha_t \nabla f_{\tilde{i}_t}(w_t)) \\
&= f(w_t) - \left(\alpha_t \nabla f_{\tilde{i}_t}(w_t)\right)^T \nabla f(w_t) + \frac{1}{2}\left(\alpha_t \nabla f_{\tilde{i}_t}(w_t)\right)^T \nabla^2 f(\xi_t) \left(\alpha_t \nabla f_{\tilde{i}_t}(w_t)\right) \\
&\leq f(w_t) - \alpha_t \nabla f_{\tilde{i}_t}(w_t)^T \nabla f(w_t) + \frac{\alpha_t^2 L}{2} \left\|\nabla f_{\tilde{i}_t}(w_t)\right\|^2.
\end{aligned}$$

If we take the expected value, by our analysis last time we'll get

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1})\right] &\leq \mathbf{E}\left[f(w_t)\right] - \alpha_t \mathbf{E}\left[\nabla f_{\tilde{i}_t}(w_t)^T \nabla f(w_t)\right] + \frac{\alpha_t^2 L}{2} \mathbf{E}\left[\left\|\nabla f_{\tilde{i}_t}(w_t)\right\|^2\right] \\
&\leq \mathbf{E}\left[f(w_t)\right] - \alpha_t \mathbf{E}\left[\left\|\nabla f(w_t)\right\|^2\right] + \frac{\alpha_t^2 L}{2} \mathbf{E}\left[\left\|\nabla f_{\tilde{i}_t}(w_t)\right\|^2\right].
\end{aligned}$$

Last time, we assumed that the magnitude of the gradient samples $\left\|\nabla f_{\tilde{i}_t}(w)\right\|$ had some global upper bound. But it turns out that this is inconsistent with strong convexity, so we can't use it here. Instead of a global upper bound on the gradient samples, let's assume a bound on their variance instead: for some $\sigma > 0$, we require that for all $w \in \mathbb{R}^d$,

$$\mathbf{E}\left[\left\|\nabla f_{\tilde{i}}(w) - \nabla f(w)\right\|^2\right] = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w) - \nabla f(w)\right\|^2 \leq \sigma^2.$$

This is equivalent to writing

$$\begin{aligned}
\sigma^2 &\geq \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w) - \nabla f(w)\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(\left\|\nabla f_i(w)\right\|^2 - 2\nabla f(w)^T \nabla f_i(w) + \left\|\nabla f(w)\right\|^2\right) \\
&= \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w)\right\|^2 - 2\nabla f(w)^T \left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(w)\right) + \frac{1}{n}\sum_{i=1}^n \left\|\nabla f(w)\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w)\right\|^2 - 2\nabla f(w)^T \left(\nabla f(w)\right) + \left\|\nabla f(w)\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(w)\right\|^2 - \left\|\nabla f(w)\right\|^2 \\
&= \mathbf{E}\left[\left\|\nabla f_{\tilde{i}_t}(w)\right\|^2\right] - \left\|\nabla f(w)\right\|^2.
\end{aligned}$$

So this gives us a way to bound $\left\|\nabla f_{\tilde{i}_t}(w_t)\right\|^2$ as well. (You may notice that this is the vector analogue of the classic statistical formula for the variance in terms of the expected value and the second moment.) Substituting this bound into our expression for SGD above gives us

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1})\right] &\leq \mathbf{E}\left[f(w_t)\right] - \alpha_t \mathbf{E}\left[\nabla f_{\tilde{i}_t}(w_t)^T \nabla f(w_t)\right] + \frac{\alpha_t^2 L}{2} \mathbf{E}\left[\left\|\nabla f_{\tilde{i}_t}(w_t)\right\|^2\right] \\
&\leq \mathbf{E}\left[f(w_t)\right] - \alpha_t \mathbf{E}\left[\left\|\nabla f(w_t)\right\|^2\right] + \frac{\alpha_t^2 L}{2}\left(\sigma^2 + \mathbf{E}\left[\left\|\nabla f(w_t)\right\|^2\right]\right) \\
&\leq \mathbf{E}\left[f(w_t)\right] - \alpha_t \left(1 - \frac{\alpha_t L}{2}\right) \mathbf{E}\left[\left\|\nabla f(w_t)\right\|^2\right] + \frac{\alpha_t^2 \sigma^2 L}{2}.
\end{aligned}$$

Next, as we did in the analysis of gradient descent, we can apply the Polyak–Lojasiewicz condition,

$$\left\|\nabla f(x)\right\|^2 \geq 2\mu\left(f(x) - f^*\right);$$

this gives us

$$\mathbf{E}\left[f(w_{t+1})\right] \leq \mathbf{E}\left[f(w_t)\right] - 2\alpha_t \mu \left(1 - \frac{\alpha_t L}{2}\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha_t^2 \sigma^2 L}{2}.$$

Subtracting f^* from both sides, we get

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1}) - f^*\right] &\leq \mathbf{E}\left[f(w_t) - f^*\right] - 2\alpha_t \mu \left(1 - \frac{\alpha_t L}{2}\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha_t^2 \sigma^2 L}{2} \\
&= \left(1 - 2\alpha_t \mu \left(1 - \frac{\alpha_t L}{2}\right)\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha_t^2 \sigma^2 L}{2}.
\end{aligned}$$

To simplify this a bit, let's add the requirement that for all time,

$$\alpha_t L \leq 1 \quad \text{which implies that} \quad 1 - \frac{\alpha_t L}{2} \geq \frac{1}{2}.$$

This simplifies our bound to

$$\mathbf{E}\left[f(w_{t+1}) - f^*\right] \leq \left(1 - \alpha_t \mu\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha_t^2 \sigma^2 L}{2}.$$

Now we will analyze this in two different settings: first for constant learning rate, and then for a decreasing step size.

SGD with a constant step size. With a constant learning rate $\alpha_t = \alpha$, we get

$$\mathbf{E}\left[f(w_{t+1}) - f^*\right] \leq \left(1 - \alpha\mu\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha^2 \sigma^2 L}{2}.$$

This is a simple linear recurrence relation. To solve it, we first find the fixed point. This fixed point occurs at ρ, where

$$\rho = \left(1 - \alpha\mu\right)\rho + \frac{\alpha^2 \sigma^2 L}{2} \quad \Rightarrow \quad \rho = \frac{\alpha \sigma^2 L}{2\mu}.$$

Subtracting the fixed point from both sides of the inequality above, we get

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1}) - f^*\right] - \frac{\alpha\sigma^2 L}{2\mu} &\leq \left(1 - \alpha\mu\right) \mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha^2 \sigma^2 L}{2} - \frac{\alpha\sigma^2 L}{2\mu} \\
&= \left(1 - \alpha\mu\right)\left(\mathbf{E}\left[f(w_t) - f^*\right] - \frac{\alpha\sigma^2 L}{2\mu}\right).
\end{aligned}$$

Now applying this recursively gives us

$$\begin{aligned}\mathbf{E}\left[f(w_T) - f^*\right] - \frac{\alpha\sigma^2 L}{2\mu} &\leq (1-\alpha\mu)^T \left((f(w_0) - f^*) - \frac{\alpha\sigma^2 L}{2\mu}\right) \\ &\leq (1-\alpha\mu)^T \cdot (f(w_0) - f^*).\end{aligned}$$

And so

$$\mathbf{E}\left[f(w_T) - f^*\right] \leq (1-\alpha\mu)^T \cdot (f(w_0) - f^*) + \frac{\alpha\sigma^2 L}{2\mu}.$$

Interpreting the result. What this says is that SGD with constant step size converges at a linear rate to a noise ball of size proportional to α. How fast that linear rate is is also a function of α: namely, it converges *faster* the larger α is. This exposes some tradeoffs in the parameters used in the expression.

Activity: the tradeoffs of SGD, as guided by our theoretical formula.

What tradeoffs happen when we change α?	How could we change σ^2? What tradeoffs happen?	How could we change μ? What tradeoffs happen?

SGD with a decreasing step size: motivation. This analysis is a bit less straightforward than the fixed-step-size analysis. We're going to look for an *optimal* step size rule. We start with our inequality from before:

$$\mathbf{E}\left[f(w_{t+1}) - f^*\right] \leq (1-\alpha_t\mu)\,\mathbf{E}\left[f(w_t) - f^*\right] + \frac{\alpha_t^2\sigma^2 L}{2}.$$

If we minimize the right side of this over α_t by differentiating, we get

$$0 = -\mu\mathbf{E}\left[f(w_t) - f^*\right] + \alpha_t\sigma^2 L \quad \Rightarrow \quad \mathbf{E}\left[f(w_t) - f^*\right] = \frac{\alpha_t\sigma^2 L}{\mu}$$

If we suppose that we used the optimal step size at each iteration, we would get

$$\begin{aligned}\frac{\alpha_{t+1}\sigma^2 L}{\mu} &\leq (1-\alpha_t\mu)\,\frac{\alpha_t\sigma^2 L}{\mu} + \frac{\alpha_t^2\sigma^2 L}{2} \\ &= \frac{\alpha_t\sigma^2 L}{\mu} - \frac{\alpha_t^2\sigma^2 L}{2},\end{aligned}$$

which simplifies to

$$\alpha_{t+1} \leq \alpha_t - \frac{\alpha_t^2\mu}{2}.$$

Finally, if we invert this, we get

$$\frac{1}{\alpha_{t+1}} \geq \frac{1}{\alpha_t - \frac{\alpha_t^2\mu}{2}}.$$

Since the function $1/x$ is convex, it follows that

$$\frac{1}{x+y} \geq \frac{1}{x} - \frac{y}{x^2},$$

and so

$$\frac{1}{\alpha_{t+1}} \geq \frac{1}{\alpha_t} + \frac{\mu}{2}.$$

That is, α_t^{-1} is increasing by about a constant amount each iteration. This motivates us to propose the $1/t$ *step size scheme* in which the learning rate decreases proportional to the inverse of the iteration number.

SGD with a $1/t$ step size: analysis. (This proof is adapted from *Optimization Methods for Large-Scale Machine Learning*.) Suppose that we pick, for some constant $c > 0$,

$$\alpha_t = \frac{c\alpha_0}{c+t}$$

where as before we require that $\alpha_0 L \leq 1$. Then we'd like to prove that

$$\mathbf{E}\left[f(w_t) - f^*\right] \leq \frac{c}{c+t} \max\left(\frac{c\alpha_0^2\sigma^2 L}{2(c\alpha_0\mu - 1)}, f(w_0) - f^*\right) = \frac{c\nu}{c+t}$$

where ν is defined to be equal to the max such that this will hold. Clearly, this holds when $t = 0$, so all we need to do to prove this is to validate the inductive case. From our analysis above, we have

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1}) - f^*\right] &\leq \left(1 - \frac{c\alpha_0}{c+t}\mu\right)\mathbf{E}\left[f(w_t) - f^*\right] + \frac{\sigma^2 L}{2}\left(\frac{c\alpha_0}{c+t}\right)^2 \\
&\leq \left(1 - \frac{c\alpha_0}{c+t}\mu\right)\frac{c\nu}{c+t} + \frac{\sigma^2 L}{2}\left(\frac{c\alpha_0}{c+t}\right)^2 \\
&= \left(\frac{c+t-c\alpha_0\mu}{(c+t)^2}\right)c\nu + \frac{c^2\alpha_0^2\sigma^2 L}{2(c+t)^2} \\
&= \left(\frac{c+t-1}{(c+t)^2}\right)c\nu - \left(\frac{c\alpha_0\mu - 1}{(c+t)^2}\right)c\nu + \frac{c^2\alpha_0^2\sigma^2 L}{2(c+t)^2} \\
&\leq \left(\frac{c+t-1}{(c+t)^2}\right)c\nu - \left(\frac{c\alpha_0\mu - 1}{(c+t)^2}\right)c \cdot \frac{c\alpha_0^2\sigma^2 L}{2(c\alpha_0\mu - 1)} + \frac{c^2\alpha_0^2\sigma^2 L}{2(c+t)^2} \\
&= \left(\frac{c+t-1}{(c+t)^2}\right)c\nu.
\end{aligned}$$

Finally, since $(c+t)^2 \geq (c+t-1)(c+t+1)$,

$$\begin{aligned}
\mathbf{E}\left[f(w_{t+1}) - f^*\right] &\leq \left(\frac{c+t-1}{(c+t-1)(c+t+1)}\right)c\nu \\
&= \frac{c\nu}{c+t+1}.
\end{aligned}$$

This is what we wanted to prove.

Minibatching. One way to make all these rates smaller is by decreasing the value of σ^2. A simple way to do this is by using *minibatching*. With minibatching, we use a sample of the gradient examples of size larger than 1. If the batch size is B, this results in an estimator with variance B times smaller. **How does this trade off work for faster convergence?**