Lecture 4: Stochastic Gradient Descent Part 2

CS4787 — Principles of Large-Scale Machine Learning Systems

So SGD with constant step size converges to a noise ball!

Even if we run for a very large number of iterations,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{E}\left[\|\nabla f(w_t)\|^2\right] \le \lim_{T\to\infty} \frac{2\left(f(w_0) - f^*\right)}{\alpha T} + \frac{\alpha\sigma^2 L}{2} = \frac{\alpha\sigma^2 L}{2} \ne 0.$$

For many applications this is fine...but it seems somehow lacking.

What if we want an algorithm that actually converges to the optimum? Intuition: for the constant step size approach, we converge down to a gradient magnitude that is proportional to the step size. So if we use a *decreasing step size scheme*, can we get arbitrarily small gradients? That is, we can run the update

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

Using the same analysis as before, but with α_t in place of α, and assuming that $\alpha_t L < 1$, we can get

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

Rearranging the terms, summing up over T iterations, and telescoping the sum,

$$\mathbf{E}\left[f(w_T)\right] \le \mathbf{E}\left[f(w_0)\right] - \sum_{t=0}^{T-1} \frac{\alpha_t}{2}\mathbf{E}\left[\|\nabla f(w_t)\|^2\right] + \sum_{t=0}^{T-1} \frac{\alpha_t^2\sigma^2 L}{2}.$$

If we define τ as being the index of a random output model that is selected at random from a weighted distribution over the iterates of SGD, such that for $t \in \{0, \ldots, T-1\}$

$$\mathbf{P}\left(\tau = t\right) = \frac{\alpha_t}{\sum_{t=0}^{T-1} \alpha_t},$$

then

$$\begin{aligned}\mathbf{E}\left[\|\nabla f(w_\tau)\|^2\right] &= \sum_{t=0}^{T-1} \frac{\alpha_t}{\sum_{s=0}^{T-1}\alpha_s} \cdot \mathbf{E}\left[\|\nabla f(w_t)\|^2\right] = 2\left(\sum_{s=0}^{T-1}\alpha_s\right)^{-1} \cdot \sum_{t=0}^{T-1} \frac{\alpha_t}{2}\mathbf{E}\left[\|\nabla f(w_t)\|^2\right] \\ &\le 2\left(\sum_{s=0}^{T-1}\alpha_s\right)^{-1} \cdot \left(\mathbf{E}\left[f(w_0)\right] - \mathbf{E}\left[f(w_T)\right] + \sum_{t=0}^{T-1}\frac{\alpha_t^2\sigma^2 L}{2}\right).\end{aligned}$$

The norm of the gradient of the output w_τ will be guaranteed to go to zero if

$$\sum_{t=0}^{T-1} \alpha_t \text{ grows much faster than } \sum_{t=0}^{T-1} \alpha_t^2.$$

One example of such a step size rule is $\alpha_t = \frac{1}{L \cdot \sqrt{t+1}}$. Then we have

$$\sum_{t=0}^{T-1} \alpha_t = \sum_{t=0}^{T-1} \frac{1}{L\sqrt{t+1}} \ge \int_1^{T+1} \frac{1}{L\sqrt{x}}\,dx = \frac{2\left(\sqrt{T+1}-1\right)}{L} \quad \text{and} \quad \sum_{t=0}^{T-1}\alpha_t^2 = \sum_{t=0}^{T-1}\frac{1}{L^2(t+1)} \le 1 + \int_1^T \frac{1}{L^2 x}\,dx = \frac{\log(T)+1}{L^2}.$$

With this, we get

$$\mathbf{E}\left[\|\nabla f(w_\tau)\|^2\right] \le 2\left(\frac{2\left(\sqrt{T+1}-1\right)}{L}\right)^{-1} \cdot \left(\mathbf{E}\left[f(w_0)\right] - \mathbf{E}\left[f(w_T)\right] + \frac{\log(T)+1}{L^2}\right) = \mathcal{O}\left(\frac{\log(T)}{L\sqrt{T}}\right).$$

This is indeed going to go to zero as $T \to \infty$.

How does this compare to the expression that we got for gradient descent?

Gradient descent for strongly convex objectives. This was without assuming strong convexity. But how does SGD perform on strongly convex problems? As before, we start from this sort of expression

$$\mathbf{E}\left[f(w_{t+1})\right] \le \mathbf{E}\left[f(w_t)\right] - \frac{\alpha}{2}\mathbf{E}\left[\|\nabla f(w_t)\|^2\right] + \frac{\alpha^2\sigma^2 L}{2}$$

and apply the Polyak–Lojasiewicz condition,

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

this gives us

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

Subtracting f^* from both sides, we get

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

Now subtracting the fixed point from both sides gives us

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

Now applying this recursively,

$$\mathbf{E}\left[f(w_T) - f^*\right] - \frac{\alpha^2\sigma^2 L}{4\mu\alpha} \le (1-\mu\alpha)^K\left(f(w_0) - f^* - \frac{\alpha^2\sigma^2 L}{2\mu\alpha}\right),$$

and so since $(1-\mu\alpha) \le \exp(-\mu\alpha)$,

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

What can we learn from this expression?