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

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
\[
w_{t+1} = w_t - \alpha_t \nabla f_{\tau_t}(w_t).
\]

Using the same analysis as before, but with \( \alpha_t \) in place of \( \alpha \), and assuming that \( \alpha_t L < 1 \), we can get
\[
\mathbb{E} [f(w_{t+1})] \leq \mathbb{E} [f(w_t)] - \frac{\alpha_t}{2} \mathbb{E} \left[ \| \nabla f(w_t) \|^2 \right] + \frac{\alpha_t^2 \sigma^2 L}{2}.
\]

Rearranging the terms, summing up over \( T \) iterations, and telescoping the sum,
\[
\mathbb{E} [f(w_T)] \leq \mathbb{E} [f(w_0)] - \sum_{t=0}^{T-1} \frac{\alpha_t}{2} \mathbb{E} \left[ \| \nabla f(w_t) \|^2 \right] + \sum_{t=0}^{T-1} \frac{\alpha_t^2 \sigma^2 L}{2}.
\]

If we define \( \tau \) 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 \( t \in \{0, \ldots, T-1\} \)
\[
P(\tau = t) = \frac{\alpha_t}{\sum_{t=0}^{T-1} \alpha_t},
\]
then
\[
\mathbb{E} \left[ \| \nabla f(w_\tau) \|^2 \right] = \sum_{t=0}^{T-1} \frac{\alpha_t}{\sum_{s=0}^{T-1} \alpha_s} \cdot \mathbb{E} \left[ \| \nabla f(w_t) \|^2 \right] = 2 \left( \sum_{s=0}^{T-1} \alpha_s \right)^{-1} \cdot \sum_{t=0}^{T-1} \frac{\alpha_t}{2} \mathbb{E} \left[ \| \nabla f(w_t) \|^2 \right] = 2 \left( \sum_{s=0}^{T-1} \alpha_s \right)^{-1} \cdot \left( \mathbb{E} [f(w_0)] - \mathbb{E} [f(w_T)] + \sum_{t=0}^{T-1} \frac{\alpha_t^2 \sigma^2 L}{2} \right).
\]

The norm of the gradient of the output \( w_\tau \) 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 \( \alpha_t = \frac{1}{L \sqrt{t+1}} \). Then we have
\[
\sum_{t=0}^{T-1} \alpha_t = \sum_{t=0}^{T-1} \frac{1}{L \sqrt{t+1}} \geq \int_1^{T+1} \frac{1}{L \sqrt{x}} \, dx = \frac{2 (\sqrt{T+1} - 1)}{L} \quad \text{and} \quad \sum_{t=0}^{T-1} \alpha_t^2 = \sum_{t=0}^{T-1} \frac{1}{L^2 (t+1)} \leq 1 + \int_1^{T} \frac{1}{L^2 x} \, dx = \frac{\log(T) + 1}{L^2}.
\]
With this, we get

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

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

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

and apply the Polyak–Lojasiewicz condition,

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

this gives us

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

Subtracting $f^*$ from both sides, we get

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

Now subtracting the fixed point from both sides gives us

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

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

Now applying this recursively,

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

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

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

What can we learn from this expression?