CIS3990-002: Mathematics of Machine Learning Fall 2023

Lecture: SGD Convergence

Date: November 27th, 2023 *Author: Eric Wong*

In stochastic gradient descent (SGD), we assumed that the objective function was decomposable into a sum of N objectives, i.e.

$$\min_x f(x) = \min_x \frac{1}{N} \sum_i f_i(x)$$

where $f(x) = \sum_i f_i(x)$. Then, the SGD update randomly samples one of these sub-objectives and performs a gradient descent update as follows:

$$x_{t+1} = x_t - \gamma_t \nabla f_n(x_t)^\top, \quad n \sim \text{Uniform}[1, \dots, N]$$

Recall that the classic gradient descent step can ensure improvement by taking a small enough of a step size. In contrast, SGD is not guaranteed to always improve the objective because the gradient with respect to a single f_n may not decrease the total sum $f(x) = \frac{1}{N} \sum_i f_i(x)$. So why does SGD even work in the first place? In this final example, we will prove why it works. Even better, we will prove a convergence rate—a measure of how fast it takes for SGD to converge.

These notes are based upon the follwoing notes:

https://www.cs.cornell.edu/courses/cs4787/2019sp/notes/lecture5.pdf

For an in-depth plunge into many convergence proofs, see

https://arxiv.org/abs/2301.11235

1 Taylor's Theorem

Before we prove the convergence of SGD, we first need one additional tool. In particular, we need a variation of the Taylor series approximation, which is a more exact formalization of the Taylor series.

- Recall the multivariate Taylor series around x_0:

$$f(x) = \sum_{k=0}^{\infty} \frac{D^k f(x_0)}{k!} \delta^k$$

  where $\delta = x - x_0$ and $D^k f(x_0)$ is the kth total derivative tensor.

- Also recall that the first three terms are the following:

  1. $k = 0$ we have $D^0 f(x_0) \delta^0 = f(x_0) \in \mathbb{R}$

2. $k = 1$ we have $D^1 f(x_0)\delta^1 = \nabla f(x_0)\delta \in \mathbb{R}$
3. $k = 2$ we have $D^2 f(x_0)\delta^2 = \delta^\top \nabla^2 f(x_0)\delta \in \mathbb{R}$

- For a second order approximation we have

$$f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(x_0)(x - x_0)$$

- The above approximation does not say explicitly how close or far the approximation is from the true function f. Taylor's theorem formalizes the exact approximation of f around x_0. For example, for a first order approximation of

$$f_{\text{approx}}(x) = f(x_0) + \nabla f(x_0)(x - x_0)$$

, we want to quantify the remainder $R(x)$:

$$R(x) = f(x) - f_{\text{approx}}(x) = f(x) - (f(x_0) + \nabla f(x_0)(x - x_0))$$

- Taylor's Theorem: Let f be $k+1$ times differentiable and $D^k f$ be continuous on the interval $[x_0, x]$. Then, the remainder is

$$R_k(x) = \frac{D^{k+1} f(\xi)}{(k+1)!}(x - x_0)^{k+1}$$

for some $\xi \in [x_0, x]$

- For a first order approximation, this means that

$$f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)(x - x_0)$$

for some $\xi \in [x_0, x]$

2 SGD Analysis

Proof sketch: The proof will consist of three main steps. The end goal of this proof is to drive the gradient to zero—if the gradient is zero, then the algorithm has converged.

1. First, we'll use a Taylor approximation to calculate the error of the SGD update (e.g. how far off the update is from the true value).
2. Then, we'll use an expectation to handle the randomness of the SGD update.
3. Third, we'll use a telescoping sum to combine the progress from all T steps, and bound the norm of the gradient.
4. Finally, we'll pick a smart step size to drive the norm of the gradient to zero.

- To analyze SGD, we need to make a few assumptions on the function. In particular, these assumptions will ensure that the function is not too "crazy" and behaves with some regularity.

- Bounded gradient: we first assume that the norm of the gradient is globally bounded, i.e. there exists some $G > 0$ such that
$$\|\nabla f_i(x)\| \leq G$$
for all functions f_i and all inputs x.

- Without this assumption, the gradient could grow infinitely, which would result in an unbounded SGD step.

- Bounded Hessian: we next assume that the Hessian is similarly well-behaved, i.e. there exists an L such that
$$u^\top \nabla^2 f(x) u \leq L\|u\|^2$$
for all u, x. In other words, the inner product with the Hessian is at most L times the standard inner product (i.e. the Hessian does not explode the inner product by an infinite amount).

A typical way to approach convergence in optimization is to apply the Taylor series approximation and combine it with additional regularity assumptions. Putting aside the issue of stochasticity for now, this results in the following:

- The first step here is to apply the Taylor approximation to the SGD update around x_t. Plugging in the SGD update, we have:
$$f(x_{t+1}) = f(x_t - \gamma_t \nabla f_n(x_t)^\top)$$
and a direct application of Taylor's theorem results in
$$f(x_{t+1}) = f(x_t) - \nabla f(x_t)(\gamma_t \nabla f_n(x_t)^\top) + \frac{1}{2}(\gamma_t \nabla f_n(x_t))\nabla^2 f(\xi)(\gamma_t \nabla f_n(x_t)^\top)$$
Re-arranging a bit we get
$$f(x_{t+1}) = f(x_t) - \gamma_t \nabla f(x_t)\nabla f_n(x_t)^\top + \frac{\gamma_t^2}{2}\nabla f_n(x_t)\nabla^2 f(\xi)\nabla f_n(x_t)^\top$$

- Under the regularity assumption for the Hessian, we have
$$f(x_{t+1}) \leq f(x_t) - \gamma_t \nabla f(x_t)\nabla f_n(x_t)^\top + \frac{L\gamma_t^2}{2}\|\nabla f_n(x_t)^\top\|^2$$

- Under the bounded assumption for the gradient, we have
$$f(x_{t+1}) \leq f(x_t) - \gamma_t \nabla f(x_t)\nabla f_n(x_t)^\top + \frac{LG^2\gamma_t^2}{2}$$

Note here that the sign of the middle term, $\nabla f(x_t)\nabla f_n(x_t)^\top$ is unclear (since f_n is a randomly selected fnction), and so $f(x_{t+1})$ is not guaranteed to improve upon $f(x_t)$.

The key will be to show improvement in *expectation*.

- Let us take the expected value of both sides over the random index selection n:
$$\mathbb{E}[f(x_{t+1})] \leq \mathbb{E}[f(x_t)] - \gamma_t \mathbb{E}[\nabla f(x_t)\nabla f_n(x_t)^\top] + \frac{LG^2\gamma_t^2}{2}$$

- Let's tackle the middle term. Using the law of iterated expectation, we have that

$$\mathbb{E}[\nabla f(x_t)\nabla f_n(x_t)^\top] = \mathbb{E}[\mathbb{E}_n[\nabla f(x_t)\nabla f_n(x_t)^\top | x_t]]$$

  where the inner expectation is

$$\mathbb{E}_n[\nabla f(x_t)\nabla f_n(x_t)^\top | x_t] = \nabla f(x_t)\mathbb{E}_n[\nabla f_n(x_t)^\top | x_t]$$

  and the expectation with respect to n is

$$\mathbb{E}_n[\nabla f_n(x_t)^\top | x_t] = \sum_{i=1}^{N} \nabla f_i(x_t)^\top P(i = n|x_t) = \frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_t)^\top = \nabla f(x_t)^\top$$

  Plugging this back in, we get

$$\mathbb{E}[\nabla f(x_t)\nabla f_n(x_t)^\top] = \mathbb{E}[\nabla f(x_t)\nabla f(x_t)^\top] = \mathbb{E}[\|\nabla f(x_t)\|_2^2]$$

- Re-arranging this sum to put the norm of the gradient on the left, we have

$$\gamma_t \mathbb{E}[\|\nabla f(x_t)\|_2^2] \leq \mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})] + \frac{LG^2\gamma_t^2}{2}$$

- Next we will do what is called a telescoping sum. That is,

$$\sum_{i=1}^{N}(a_i - a_{i+1}) = a_N - a_0$$

- Sum all the gradients and take a sum over all iterations to get:

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[\|\nabla f(x_t)\|_2^2] \leq \sum_{t=0}^{T-1}(\mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})]) + \frac{LG^2}{2}\sum_{t=0}^{T-1}\gamma_t^2$$

  and applying the telescoping sum, we get

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[\|\nabla f(x_t)\|_2^2] \leq \mathbb{E}[f(x_0)] - \mathbb{E}[f(x_T)] + \frac{LG^2}{2}\sum_{t=0}^{T-1}\gamma_t^2$$

- Since $f^* = \min_x f(x) \leq \mathbb{E}[f(x_T)]$ and since $f(x_0)$ is not random, we have

$$\sum_{t=0}^{T-1} \gamma_t \mathbb{E}[\|\nabla f(x_t)\|_2^2] \leq f(x_0) - f^* + \frac{LG^2}{2}\sum_{t=0}^{T-1}\gamma_t^2$$

This is almost there, but we need to handle the term on the left. The simplest is to just take the minimum norm over all time steps:

$$\min_t \mathbb{E}[\|\nabla f(x_t)\|_2^2]\left(\sum_{t=0}^{T-1}\gamma_t\right) \leq f(x_0) - f^* + \frac{LG^2}{2}\sum_{t=0}^{T-1}\gamma_t^2$$

and then divide both sides by the sum of the step sizes to get the bound:

$$\min_t \mathbb{E}[\|\nabla f(x_t)\|_2^2] \leq \frac{f(x_0) - f^*}{\sum_{t=0}^{T-1}\gamma_t} + \frac{LG^2}{2}\frac{\sum_{t=0}^{T-1}\gamma_t^2}{\sum_{t=0}^{T-1}\gamma_t}$$

- A constant step size brings the first term to zero, but the second term stays as a non-zero constant!
- Constant step-size: $\gamma_t = \gamma$ for some $\gamma > 0$, then
  1. $\sum_t \gamma_t = T\gamma$
  2. $\sum_t \gamma_t^2 = T\gamma^2$.
  3. Then, the norm of the gradient is $O(1/t) + O(\gamma)$. This doesn't go to zero! In practice, if you use a constant step size, you'll notice that SGD oscillates around the minimum. This is why, and is sometimes called the *noise ball*.
- Instead, convergence is dictated by the fraction of step sizes. We need the ratio $\frac{\sum_{t=0}^{T-1} \gamma_t^2}{\sum_{t=0}^{T-1} \gamma_t} \to 0$, or in other words, we need $\sum_{t=0}^{T-1} \gamma_t$ to grow much faster than $\sum_{t=0}^{T-1} \gamma_t^2$.
- Classic decreasing step size: $\gamma_t = \gamma/t$ for some $\gamma > 0$, then
  1. $\sum_t \gamma_t = \gamma H_T = O(\log T)$ where H_T is the Tth Harmonic number
  2. $\sum_t \gamma_t^2 = \gamma\pi^2/6 = O(1)$.
  3. Then, the norm of the gradient is $O(1/\log T)$. This goes to zero, but is quite slow. In practice, you can use this but you'll find that SGD makes very, very slow progress.
- Bigger decreasing step sizes: $\gamma_t = \gamma/\sqrt{t}$ for some $\gamma > 0$, then
  1. $\sum_t \gamma_t = \gamma \sum_t \frac{1}{\sqrt{t}} \approx \gamma \cdot 2\sqrt{T} = O(\sqrt{T})$ (can check this by noting that $\sum_t \frac{1}{\sqrt{T}}$ is the lower Reimann sum of the integral $\int_0^T t^{-1/2} dt = 2t^{1/2}|_0^T = 2\sqrt{T}$)
  2. $\sum_t \gamma_t^2 = \gamma \sum_t \frac{1}{t} = \gamma H_T = O(\log T)$.
  3. Then, the norm of the gradient is $O(\log T/\sqrt{T}) = \tilde{O}(1/\sqrt{T})$. This is faster than $O(1/\log T)$.