WEEKLY PARTICIPATION 6: VERIFYING THE SGD ASSUMPTIONS

Recall that in our proof of the convergence of SGD, we assumed that our stochastic gradient at iteration t, $\boldsymbol{g}_t$, has the following properties:

(1) $\boldsymbol{g}_t$ is conditionally unbiased: $\mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_0] = \nabla f(\boldsymbol{x}_t)$, and

(2) $\boldsymbol{g}_t$ has bounded expected norm: there is a $C > 0$ satisfying $\mathbb{E}\|\boldsymbol{g}_t\|_2^2 \leq C$ for all t.

These properties respectively ensure that moving in the direction of the negative of the stochastic gradient tends to decrease the function, and that we can choose a stepsize small enough to prevent us from moving too far in the occasional bad direction.

Let's verify that a common stochastic gradient used in the case where f has the finite sum structure does indeed have these properties. Assume that f has the finite sum structure

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})$$

for some convex functions f_i; further assume that the gradients of these functions are bounded by $B > 0$ everywhere, so $\|\nabla f_i(\boldsymbol{x})\|_2^2 \leq B$ for all i and all points $\boldsymbol{x}$.

Consider the stochastic gradient $\boldsymbol{g}_t$ given by

$$\boldsymbol{g}_t = \frac{n}{k} \sum_{j \in J_t} \nabla f_j(\boldsymbol{x}_t),$$

where J_t is a subset of $[n]$ of size k that is sampled uniformly at random, i.e. J_t is a set of random indices that satisfies

$$\mathbb{P}[J_t = \{j_1, \ldots, j_k\}] = \frac{1}{\binom{n}{k}}$$

for any set of k unique indices $\{j_1, \ldots, j_k\} \subseteq [n]$.

Verify that $\boldsymbol{g}_t$ is a conditionally unbiased estimate of the gradient:

$$\mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_0] = \mathbb{E}[\boldsymbol{g}_t \,|\, \boldsymbol{x}_t] = \nabla f(\boldsymbol{x}_t).$$

Use Jensen's inequality with $f(z) = z^2$ and the triangle inequality for the ℓ_2 norm to conclude the generally useful fact[1] that for any set of vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$,

$$\left\| \sum_{j=1}^{k} \boldsymbol{z}_j \right\|_2^2 \leq k \sum_{j=1}^{k} \|\boldsymbol{z}_j\|_2^2.$$

Use this fact to find the smallest C you can that bounds the expected norm of $\boldsymbol{g}_t$.

[1] This is like a version of the triangle inequality, but for the square of the norm. Note that the inequality is sharp, because the equality holds if all the summands are the same vector.