CSE 547 - Assignment 2

Philip Pham

April 23, 2018

Problem 0

List of collaborators: I have not collaborated with anyone.

List of acknowledgements: None.

Certify that you have read the instructions: I have read and understood these policies.

Problem 1: Generalization, Streaming, and SGD

In class, we examined using Stochastic Gradient Descent (SGD) for empirical loss minimization, where we have an N sized training set $\mathcal{T}$. The empirical loss considered was:

$$F(w) = \frac{1}{N} \sum_{(x,y)\in\mathcal{T}} l\left(w, (x, y)\right). \tag{1}$$

Here, gradient descent for the function F is the algorithm:

1. Initialize at some point $w^{(0)}$.
2. Sample (x, y) uniformly at random from the set $\mathcal{T}$.
3. Update the parameters:

$$w^{(k+1)} = w^{(k)} - \eta_k \cdot \nabla l\left(w^{(k)}, (x, y)\right), \tag{2}$$

   and go back to 2.

We provided guarantees assuming that F was smooth and the gradients in our training set were uniformly bounded, $\|\nabla l\left(w, (x, y)\right)\| \leq B$.

However, in practice, we care about generalization, that is, statements on how well we do on the underlying distribution. Define:

$$\mathcal{L}(w) = \mathbb{E}_{(x,y)\in\mathcal{D}} l\left(w, (x, y)\right), \tag{3}$$

where $\mathcal{D}$ is the underlying distribution.