New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 6763: Homework 3.

Collaboration is allowed on this problem set, but solutions must be written-up individually. Please list collaborators for each problem separately, or write "No Collaborators" if you worked alone.

Problem 1: Compressed classification.

(10 pts) In machine learning, the goal of many classification methods (like support vector machines) is to separate data into classes using a separating hyperplane.

Recall that a hyperplane in $\mathbb{R}^d$ is defined by a unit vector $a \in \mathbb{R}^d$ ($\|a\|_2 = 1$) and scalar $c \in \mathbb{R}$. It contains all $h \in \mathbb{R}^d$ such that $\langle a, h\rangle = c$.

Suppose our dataset consists of n unit vectors in $\mathbb{R}^d$ (i.e., each data point is normalized to have norm 1). These points can be separated into two sets X, Y, with the guarantee that there exists a hyperplane such that every point in X is on one side and every point in Y is on the other. In other words, for all $x \in X, \langle a, x\rangle > c$ and for all $y \in Y, \langle a, y\rangle < c$.

Furthermore, suppose that the ℓ_2 distance of each point in X and Y to this separating hyperplane is at least ϵ. When this is the case, the hyperplane is said to have "margin" ϵ.

1. Show that this margin assumption equivalently implies that for all $x \in X, \langle a, x\rangle \geq c + \epsilon$ and for all $y \in Y, \langle a, y\rangle \leq c - \epsilon$.

2. Show that if we use a Johnson-Lindenstrauss map Π to reduce our data points to $O(\log n/\epsilon^2)$ dimensions, then the dimension reduced data can still be separated by a hyperplane with margin $\epsilon/4$, with high probability (say $> 9/10$).

Problem 2: Concentration of Random Vectors

(12 pts) In Stochastic Gradient Descent, we replace the true gradient vector with a stochastic gradient that is equal to the true gradient in expectation. Our analysis in class only used equality in expectation, although more refined analysis of SGD often requires understanding how well the stochastic gradient concentrates around its expectation. Previously, all concentration results we studied apply to random numbers. For this problem, you will prove a basic concentration inequality for random vectors.

In particular, let $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^d$ be i.i.d. random vectors in d dimensions (independent, drawn from the same distribution) with mean $\boldsymbol{\mu}$. I.e., $\mathbb{E}[\mathbf{x}_i] = \boldsymbol{\mu}$. Further suppose that $\mathbb{E}\left[\|\mathbf{x}_i - \boldsymbol{\mu}\|_2^2\right] = \sigma^2$. σ^2 is a natural generalization of "variance" to a random vector. Let $\mathbf{s} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{x}_i$. Prove that if $k \geq O\left(\frac{1/\delta}{\epsilon^2}\right)$ then

$$\Pr\left[\|\mathbf{s} - \boldsymbol{\mu}\|_2 \geq \epsilon\sigma\right] \leq \delta.$$

Problem 3: Gradient Descent with Decaying Step-size

(10 pts) In class we showed that gradient descent with step size $\eta = R/G\sqrt{T}$ converges to an ϵ approximate minimizer of a convex G-Lipschitz function in $T = R^2G^2/\epsilon^2$ steps if our starting point $\mathbf{x}^{(0)}$ satisfies $\|\mathbf{x}^{(0)} - \mathbf{x}^*\|_2 \leq R$. Choosing this step size requires knowing G, R and moreover T in advance, which might not be reasonable in a lot of settings. For example, when training machine learning models, we might not be able to estimate how long it will take to reach a point where test accuracy levels off. Instead, we want to be able to keep running the algorithm, achieving better and better accuracy as we do.

Here, we analyze a variant of gradient descent with a variable step size that avoids this limitation. In particular, consider running gradient descent with the update $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \eta\nabla f(\mathbf{x}^{(i)})$, where

$$\eta = \frac{f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)}{\|\nabla f(\mathbf{x}^{(i)})\|_2^2}.$$

This step size requires knowledge of $f(\mathbf{x}^*)$, but not $\mathbf{x}^*$, which may be reasonable in some settings. Moreover, since it's just one parameter, grid search can be more easily used to "guess" $f(\mathbf{x}^*)$ than the three parameters

G, R, T. More complex approaches can remove the need to know this value entirely. Prove that, if we run gradient descent for $T = O(R^2G^2/\epsilon^2)$ steps using the step size above then $\hat{\mathbf{x}} = \min_{i \in 0,\ldots,T} f(\mathbf{x}^{(i)})$ satisfies $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \epsilon$.

Hint: Prove that our distance from the optimum $\|\mathbf{x}^{(i)} - \mathbf{x}^*\|_2$ always decreases with this choice of step size, and the decrease is larger if our gap from the objective value $f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)$ is larger.

Problem 4: Separation Oracles

(12 pts) Describe efficient separation oracles for each of the following families of convex sets. Here, "efficient" means linear time plus $O(1)$ calls to any additional oracles provided to you.

(a) The set $A \cap B$, given separation oracles for A and B.

(b) The ℓ_1 ball: $\|\mathbf{x}\|_1 \leq 1$.

(c) Any convex set A, given a projection oracle for A. Recall that a projection oracle, given a point $\mathbf{x}$, returns

$$\text{Proj}_A(\mathbf{x}) = \arg\min_{y \in A} \|\mathbf{x} - \mathbf{y}\|_2.$$

Above you may wish to use the following fact that was stated but not formally proven in class: for any point $\mathbf{x}$, convex set A, and $\mathbf{z} \in A$, $\|\mathbf{z} - \text{Proj}_A(\mathbf{x})\|_2 \leq \|\mathbf{z} - \mathbf{x}\|_2$.