Homework 4

Akshay Krishnamurthy
Due: Thursday 11/2

October 17, 2017

Instructions: Turn in your homework in class on Tuesday 11/2/2017

1. **Perceptron Mistake Bound.** Perceptron can also be analyzed as an online learning algorithm. In this setting, we assume that \(\{(x_t, y_t)\}_{t=1}^T\) are a sequence of examples in \(\mathbb{R}^d\) and labels chosen adversarially. We assume the margin-style realizability condition that there exists \(w^*\) such that \(\gamma = \min_t \frac{(w^* x_t) y_t}{\|w^*\|^2}, \gamma > 0\). Further assume that \(\max_t \|x_t\| \leq R\).

The learning process proceeds in rounds, on round \(t\) the example \(x_t\) is presented to the learner, who makes a prediction \(\hat{y}_t\). The learner incurs loss \(1\{\hat{y}_t \neq y_t\}\) and label \(y_t\) is revealed. Ultimately we would like to bound the number of mistakes

\[
M = \sum_{t=1}^T 1\{\hat{y}_t \neq y_t\}.
\]

(a) Prove that for perceptron, we get \(M \leq R^2/\gamma^2\).

(b) There is also a multiclass generalization of perceptron. Assume there are \(K\) classes, so \(y_t \in \{1, \ldots, K\}\), and as usual assume \(\max_t \|x_t\| \leq R\). The parameter is a weight matrix \(W \in \mathbb{R}^{K \times d}\) and the prediction is \(h_W(x) = \arg\max_k (Wx)_k\). In this setting the perceptron algorithm can be expressed as, with \(W^{(0)} = 0\)

\[
W^{(t+1)} \leftarrow W^{(t)} + U^{(t)}, \quad U^{(t)}_k = x_t (1\{y_t = k\} - 1\{\hat{y}_t = k\}), \quad \hat{y}_t = \arg\max_k (W^{(t)} x_t)_k.
\]

Here \(U^{(t)} \in \mathbb{R}^{K \times d}\) and \(U^{(t)}_k\) is the \(k\)th row of the matrix. Ties are broken arbitrarily. Here the new notion of margin is as follows. Assume there exists \(W^*\) such that for all \(t\) and all \(k \neq y_t\)

\[
\frac{(W^* x_t)_{y_t} - (W^* x_t)_k}{\|W^*\|_F} \geq \gamma
\]

Prove that the number of mistakes for the multiclass perceptron is at most

\[
M \triangleq \sum_{t=1}^T 1\{\hat{y}_t \neq y_t\} \leq \frac{2R^2}{\gamma^2}
\]

2. **Calibration.** In this problem we’ll prove a different calibration statement for the multiclass square loss. Let \(\mathcal{D}\) be a distribution on \(\mathcal{X} \times \mathcal{Y}\) where \(\mathcal{X}\) is an abstract feature space and \(\mathcal{Y} = \{1, \ldots, K\}\), so we are doing multiclass classification. Let \(\mathcal{F} : \mathcal{X} \times \mathcal{Y} \rightarrow [0, 1]\) be a set of regression functions and associate with each \(\mathcal{F}\) a hypothesis \(h_f(x) = \arg\max_y f(x, y)\). The multiclass square loss is

\[
R_{\text{msq}}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \sum_k (f(x, k) - 1\{y = k\})^2
\]

Let \(f^*(x, y) = \mathbb{P}[Y = y|X = x]\) be the Bayes regression function, and let \(y^*(x) = \arg\max_y f^*(x, y)\) be the best label. Assume the realizability condition that \(f^* \in \mathcal{F}\). Define, for any \(\zeta > 0\)

\[
P_\zeta = \mathbb{P}_{x \sim \mathcal{D}}[f^*(x, y^*(x)) \leq \max_{y \neq y^*(x)} f^*(x, y) + \zeta]
\]
which is in some sense the noise level in the problem. Prove that for any \( f \in \mathcal{F} \) and any \( \zeta \)

\[
\mathbb{P}_{(x,y) \sim \mathcal{D}}[h_f(x) \neq y] - \mathbb{P}_{(x,y) \sim \mathcal{D}}[h_{f^*}(x) \neq y] \leq \zeta P_{\zeta} + \frac{2K}{\zeta}(R_{\text{msq}}(f) - R_{\text{msq}}(f^*))
\]

Note that this can lead to fast rates for multiclass classification when there is low noise. For example if there
is some \( \zeta \) for which \( P_{\zeta} = 0 \), which is called the Massart noise condition, this will produce a \( O(d/n\zeta) \) rate, since square loss admits \( O(d/n) \) generalization bounds, where \( d \) is the rademacher complexity or an analog of the VC-dimension.

3. **Convex Optimization.** In this problem you’ll derive a convergence rate for gradient descent on a strongly convex and smooth function. Consider the unconstrained optimization problem

\[
\min_{x \in \mathbb{R}^d} f(x)
\]

where \( f \) is differentiable, \( \lambda \)-strongly convex and \( \mu \)-smooth, which means that

\[
\begin{align*}
    f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{\lambda}{2}||y-x||^2_2 \\
    f(y) &\leq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||y-x||^2_2,
\end{align*}
\]

applies for all \( x, y \).

The following lemma about smooth and strongly convex functions will be helpful. If \( \phi \) is an \( \alpha \)-strongly convex, \( \beta \)-smooth function, then for all \( x, y \)

\[
(\nabla \phi(x) - \nabla \phi(y))^T(x-y) \geq \frac{\alpha\beta}{\alpha+\beta}||x-y||^2_2 + \frac{1}{\alpha+\beta}||\nabla \phi(x) - \nabla \phi(y)||^2_2
\]

Observe that if \( \phi(x) \) is a quadratic, then \( \alpha = \beta \) and the inequality is tight.

(a) Prove that if we run gradient descent with step size \( \eta_t = \frac{2}{\lambda+\mu} \), then

\[
f(x^{(t)}) - f^* \leq c^t||x^{(0)} - x^*||^2_2,
\]

where \( c = \frac{(\lambda-\mu)^2}{(\lambda+\mu)^2} < 1 \) and \( f^* = \min_x f(x) \).

(b) Prove that this implies a bound on \( ||x^{(t)} - x^*||^2_2 \).

4. **Hedge.** Prove that the regret bound for hedge is tight. That is, prove that for any learner, there exists an adversary, producing losses in \([0, 1]\), such that

\[
\mathbb{E} \sum_{t=1}^T \ell_t(a_t) - \min_a \sum_{t=1}^T \ell_t(a) \geq \Omega(\sqrt{T \log(K)}).
\]

You may use the fact that \( Z_j = \frac{1}{n} \sum_{i=1}^n \epsilon_{j,i} \) for \( j = 1, \ldots, d \) where \( \{\epsilon_{i,j}\} \) are iid rademacher variables, then

\[
\mathbb{E} \max_{j=1,\ldots,d} Z_j = \Omega(\sqrt{n \log d}).
\]