Mathematical Toolkit **Spring 2021**

Homework 5

Due: May 27, 2021

Note: *You may discuss these problems in groups. However, you must write up your own solutions and mention the names of the people in your group. Also, please do mention any books, papers or other sources you refer to. It is recommended that you typeset your solutions in LaTeX. Homeworks are due by the start of class on the due date.*

1. **Uniform Convergence.**

   In machine learning, we are typically given a *training set* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of labeled examples that are assumed to be drawn independently from some underlying probability distribution $\mathcal{D}$. Here, x_i is an example and y_i is its associated label. E.g., x_i could be an image taken from the web or from the ImageNet database, and y_i could be a labeling of that image according to what is in it. A learning algorithm uses this training set S in order to produce a classifier h (a function over the x's) that it hopes will have low error on new examples drawn from $\mathcal{D}$. This is typically done by fixing a family $\mathcal{H}$ of classifiers, such as a particular deep-network architecture, and then using one of various methods to find some $h \in \mathcal{H}$ with low error on S (e.g., for deep networks, this might be done using a greedy procedure called stochastic gradient descent). The hope is that by achieving low error on S, this will translate to low error with respect to $\mathcal{D}$ (i.e., the classifier will "generalize well").

   For a classifier h, define its *true error* as $err_{\mathcal{D}}(h) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$ and its *empirical error* as $err_S(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h(x_i)\neq y_i}$. In other words, true error is the probability of making a mistake on a new random example whereas empirical error is the fraction of mistakes on S.

   Use Chernoff-Hoeffding bounds to prove the following. There exists constants c_1, c_2 such that for any family of classifiers $\mathcal{H}$, and any $\varepsilon, \delta > 0$, if $S \sim \mathcal{D}^n$ for

   $$n \geq \frac{c_1}{\varepsilon^2}\left[\ln|\mathcal{H}| + \ln\left(\frac{c_2}{\delta}\right)\right],$$

   then with probability at least $1-\delta$, all $h \in \mathcal{H}$ satisfy $|err_S(h) - err_{\mathcal{D}}(h)| \leq \varepsilon$.

   For example, if $\mathcal{H}$ is a deep-network architecture with s tunable weights that are 32-bit floating point numbers, then $\log|\mathcal{H}| = O(s)$. Interestingly, deep networks tend to generalize even when given much less data than in the above bound, and trying to give mathematical guarantees for this is a major direction of current research.

2. **Randomization and Non-Uniformity.**

   Prove that **BPP** $\subseteq$ **P/poly**. Hint: Use Chernoff-Hoeffding bounds.

3. **Gaussian Random Variables.**

   Prove the following useful facts about Gaussian random variables:

   (a) Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ be two vectors. Let $\mathbf{g} \in \mathbb{R}^n$ be a random vector such that each coordinate g_i of $\mathbf{g}$ is distributed as a Gaussian random variable with mean 0 and variance 1, and any two coordinates g_i, g_j (for $i \neq j$) are independent. Then show that

   $$\mathop{\mathbb{E}}_{\mathbf{g}} \left[\langle \mathbf{u}, \mathbf{g} \rangle \cdot \langle \mathbf{v}, \mathbf{g} \rangle \right] \;=\; \langle \mathbf{u}, \mathbf{v} \rangle \; .$$

   (b) Let g be a Gaussian random variable with mean 0 and variance 1. Show that for any $t \in \mathbb{R}$, we have

   $$\mathbb{E}\left[e^{tg} \right] \;=\; e^{t^2/2} \, .$$

4. **Supremum of Gaussians.**

   (a) Let $g \sim N(0,1)$ be a Gaussian random variable with mean 0 and variance 1. Show that for $t > 0$,

   $$\mathbb{P}\left[g \geq t \right] \;\leq\; e^{-t^2/2} \, .$$

   Hint: Use 3(b).

   (b) Let $g_1, \ldots, g_n \sim N(0,1)$ be independent Gaussian random variables. Show that for some constants c_1, c_2 we have

   $$\mathbb{E}\left[\max_{i \in [n]} |g_i| \right] \;\leq\; c_1 \sqrt{\ln n} + c_2 \, .$$

   You may use the fact that for a non-negative random variable Z, the expectation can be computed as $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}[Z \geq t] \, dt$.