1. Uniform Convergence. [3+7]

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 \( D \). 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 \( D \). 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 optimization 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 \( D \) (i.e., the classifier will “generalize well”).

For a classifier \( h \), define its true error as \( \text{err}_D(h) = \mathbb{P}_{(x,y) \sim D}[h(x) \neq y] \) and its empirical error as \( \text{err}_S(h) = \frac{1}{n} \sum_{i=1}^{n} 1_{h(x_i) \neq y_i} \). 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 \( S \). We will use Chernoff-Hoeffding bounds to show that if we have a sufficiently large data sample, then finding a hypothesis with low empirical error, also finds a hypothesis with low true error (with high probability over the choice of the data sample).

(a) Fix a hypothesis \( h \in \mathcal{H} \), and let the probability space be defined by choosing a data set \( S \) of \( n \) independent samples, each drawn according to the distribution \( D \) i.e., \( S \sim D^n \). Prove that we can write \( \text{err}_S(h) \) as

\[
\text{err}_S(h) = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i,
\]

where \( X_1, \ldots, X_n \) are independent Bernoulli variables with parameter \( p = \text{err}_D(h) \).
(b) Use Chernoff-Hoeffding bounds to prove that there exists constants $c_1, c_2$ such that for any family of classifiers $\mathcal{H}$, and any $\varepsilon, \delta > 0$, if $S \sim D^n$ 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 $|\text{err}_S(h) - \text{err}_D(h)| \leq \varepsilon$.

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. **Gaussian Random Variables.** [5+5+5]

Prove the following very useful facts about Gaussian random variables:

(a) Let $u, v \in \mathbb{R}^n$ be two vectors. Let $g \in \mathbb{R}^n$ be a random vector such that each coordinate $g_i$ of $g$ 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

$$\mathbb{E}_g [(u, g) \cdot (v, g)] = \langle u, v \rangle.$$

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

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

Comparing coefficients of $t^{2k}$ on both sides, use this to show that for any $k \in \mathbb{N}$,

$$\mathbb{E} [g^{2k}] = \frac{(2k)!}{2^k \cdot k!}.$$

(c) Let $g_1, g_2, g_3$ and $g_4$ be (not necessarily independent) Gaussian random variables with mean 0. Additionally, assume that for all coefficients $a_1, \ldots, a_4 \in \mathbb{R}$, the linear combination $a_1 g_1 + \cdots + a_4 g_4$ is also a Gaussian random variable (note that you were asked to prove this in class for independent Gaussian random variables, and this property is not always true if $g_1, \ldots, g_4$ are not independent. But here we are restricting ourselves to $g_1, \ldots, g_4$ which satisfy this assumption).

Consider the function $\mathbb{E}_{g_1, g_2, g_3, g_4} [e^{a_1 g_1 + a_2 g_2 + a_3 g_3 + a_4 g_4}]$ in the variables $t_1, t_2, t_3, t_4$ and use it to show that

$$\mathbb{E} [g_1 g_2 g_3 g_4] = \mathbb{E} [g_1 g_2] \cdot \mathbb{E} [g_3 g_4] + \mathbb{E} [g_1 g_3] \cdot \mathbb{E} [g_2 g_4] + \mathbb{E} [g_1 g_4] \cdot \mathbb{E} [g_2 g_3].$$

This shows that for any four Gaussian random variables, the expectation of their product can be expressed in terms of their pairwise correlations! This is a special case of what is known as Wick’s theorem, which can also be proved by the above method.

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

$$
P[g \geq t] = \int_t^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \, dx \leq e^{-t^2/2}.
$$

(b) Let $g_1, \ldots, g_n \sim N(0,1)$ be independent Gaussian random variables. Show that

$$
\mathbb{E} \left[ \max_{i \in [n]} |g_i| \right] \leq 4\sqrt{n \ln n}.
$$

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