1 [Indicator Functions] An indicator function $I_A$ of a subset $A$ of the universe $\Omega$ is defined as $I_A : \Omega \rightarrow \{0, 1\}$, $I_A(x) = 1$ iff $x \in A$.

(a) If the universe represents the set of outcomes, and $A$ a class defined by a subset of the outcomes, show that $E(I_A) = P(A)$ (you may assume a discrete universe).

(b) We can think of $A$ as our ‘true’ model, i.e. the assumption on the data. We decide to model $A$ by another set $B$. If our loss function is 1 whenever in error, 0 otherwise, show that the risk, as a function of $B$, is equal to the probability of the symmetric difference of $A$ and $B$.

2 [Perceptrons] We have a sequence $(x_i, y_i), 1 \leq i \leq l$, generated by the true model:

- $x$ is sampled in $\mathbb{R}^2$ according to the probability density $f(x)$.
- $y$ is a label, categorizing $x$ into one of two linearly separable classes. If $y$ is positive [negative], we say $x$ is a “positive” [“negative”] sample.
- $\theta$ is an unknown $\mathbb{R}^2$ vector, which characterizes the two classes:
  $$y = \text{sgn}(\theta'x).$$

Our task is to estimate the vector $\theta$. We always assume that $|x_k| \leq R$, and that all data points lie a margin away from the separating line, i.e. we have $y_k \cdot \theta'x_k \geq \gamma > 0$.

Consider the algorithm that starts with some $\hat{\theta}_0$ and proceeds:

$$\hat{\theta}_k = \begin{cases} 
\hat{\theta}_{k-1} + y_kx_k, & \text{if } y_k \cdot \hat{\theta}_{k-1}'x_k < 0 \\
\hat{\theta}_{k-1}, & \text{otherwise.}
\end{cases}$$

(a) Assume that $f(x)$ is uniform over a disk of radius $R$, excluding the band due to the margin. Show that $\hat{\theta}_k$ in the above algorithm converges to the true $\theta$, in the sense that, for all $\epsilon \geq \epsilon(\gamma, R) > 0$:

$$P\{\cos(\theta, \hat{\theta}_k) > 1 - \epsilon\} \rightarrow 0 \quad \text{as } k \rightarrow \infty,$$

$\epsilon(\gamma, R)$ is the accuracy beneath which no updates will occur, due to the margin.

(b) Show the stronger statement that $\hat{\theta}_k$ converges to $\theta$ after a finite number of updates.

(c) Construct a distribution $f(x)$ where convergence as in (a) is not guaranteed.

(d) Show that, for any distribution, we have:

$$P\{x \mid \hat{\theta}_k'x \leq 0 \text{ and } \theta'x \geq 0\} \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$
(e) [optional] If the data is generated as in part (a), but each sample is independently corrupted: \( y \) is flipped with a small probability \( \alpha \), then would the algorithm still converge, in a finite number of updates or otherwise?

3 [Maximum Entropy] Given a discrete random variable \( X \) taking values in the finite set \( \{1, \ldots, k\} \), find the probability mass function \( p(x) \) that maximizes the entropy \( H(X) \), subject to the constraint that:

\[
E[X] = \sum_{i=1}^{k} i \cdot p(i) = \mu.
\]

4 [Chernoff Bounds] Let \( Z \) be an arbitrary random variable admitting a moment generating function \( M_Z(s) = E[e^{sZ}] \).

(a) Use Markov’s inequality to show that, for all \( a \), we have:

\[
P(Z \geq a) \leq e^{-sa} M_Z(s), \quad \text{for} \ s \geq 0,
\]

\[
P(Z \leq a) \leq e^{-sa} M_Z(s), \quad \text{for} \ s \leq 0.
\]

(b) Define

\[
\phi_Z^+(a) = \max_{s \geq 0} [sa - \log M_Z(s)],
\]

\[
\phi_Z^-(a) = \max_{s \leq 0} [sa - \log M_Z(s)],
\]

and show that

\[
P(Z \geq a) \leq e^{-\phi_Z^+(a)},
\]

\[
P(Z \leq a) \leq e^{-\phi_Z^-(a)}.
\]

(c) Let \( Z_1, Z_2, \ldots, Z_n \) be independent random variables with the same distribution as \( Z \). Let \( S_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \). Show that \( \phi_{S_n}^+(a) = n\phi_Z^+(a) \) and \( \phi_{S_n}^-(a) = n\phi_Z^-(a) \).

(d) Show that if \( a > E[Z] \) then \( \phi_Z^+(a) > 0 \), and if \( a < E[Z] \) then \( \phi_Z^-(a) > 0 \). [Hint: explicitly compute the maximized expression and its derivative, at \( s = 0 \).]

5 [Maximum Likelihood Estimation] Let \( X_1, X_2, \ldots, X_n \) be i.i.d. Bernoulli random variables with \( P(X_i = 1) = p \). Our data is a set of observations \( x_1, x_2, \ldots, x_n \). If we correctly choose our class to be Bernoulli, parameterized by \( q \), then density estimation is equivalent to estimating \( p \). A natural choice for the estimator is one that maximizes the likelihood of an observation:

\[
\hat{p}_n = \arg\max_q P(X_1 = x_1, \ldots, X_n = x_n; q).
\]

(a) Show that this formulation is equivalent to empirical risk minimization, using the loss function \( L(x, p) = -\log P(X = x; p) \).

(b) Show that \( \hat{p}_n = \frac{1}{n} \sum_i x_i \), the empirical distribution.
6 [Types of Convergence] Consider the setting of problem 5. We are interested about whether the empirical distribution converges to the true distribution.

(a) Use Chebyshev’s inequality to show that \( \hat{p}_n \rightarrow p \) in probability, sometimes written \( \hat{p}_n \overset{p}{\rightarrow} p \), meaning:

\[
P(|\hat{p}_n - p| > \epsilon) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty \quad \text{for all} \quad \epsilon > 0.
\]

(b) Let \( X \) have the same distribution as the \( X_i \)'s. Using the definitions in problem 4, show that if \( p < a < 1 \) then \( \phi^+_X(a) = D(a\|p) \), where the latter expression denotes the KL distance between two Bernoulli distributions, parameterized by \( a \) and \( p \) respectively. Similarly, show that if \( 0 < a < p \) then \( \phi^-_X(a) = D(a\|p) \).

(c) Show that \( D(p + \epsilon\|p) \geq 2\epsilon^2 \), and similarly that \( D(p - \epsilon\|p) \geq 2\epsilon^2 \). Assume that \( \epsilon \ll \min\{p, 1-p\} \). [Hint: expand the logarithm around 1.]

(d) Use the results from parts (b) and (c), together with that of problem 4, to deduce that, for all \( \epsilon > 0 \) small enough, we have the additive Chernoff bound:

\[
P(|\hat{p}_n - p| > \epsilon) \leq 2e^{-2\epsilon^2 n}.
\]

(e) [optional] Unlike part (a), the result of part (d) gives a strong bound on the decay of the probability that the empirical distribution deviates from the true one. To see what this can buy us, consider the Borel-Cantelli lemma, stated as follows:

"Given a sequence \( A_n \) of events, if \( \sum_{n=1}^{\infty} P(A_n) < \infty \), then \( P(\limsup_{n \rightarrow \infty} A_n) = 0 \), i.e. the probability that infinitely many of the events occur is zero."

Use the lemma to show that \( \hat{p}_n \rightarrow p \) almost surely, sometimes written \( \hat{p}_n \overset{a.s.}{\rightarrow} p \), formally:

\[
P(\{\omega \mid p_n(\omega) \rightarrow p \quad \text{as} \quad n \rightarrow \infty\}) = 1.
\]

7 [Monotone Convergence of the Empirical Distribution] Consider once more the setting of problem 5. Let \( \hat{p}_n \) denote the empirical distribution, and let \( D(\hat{p}_n\|p) \) denote the KL distance between the empirical distribution and the true one. Note that since \( \hat{p}_n \) is a random variable, so is \( D(\hat{p}_n\|p) \).

(a) Show that \( E[D(\hat{p}_{2n}\|p)] \leq E[D(\hat{p}_n\|p)] \).

(b) Show that \( E[D(\hat{p}_{n+1}\|p)] \leq E[D(\hat{p}_n\|p)] \).