Exercises for Sample complexity and uniform convergence for learning and data analysis

1 Large Deviation Bounds

1. Suppose that we can obtain independent samples $X_1, X_2, \ldots,$ of a random variable X, and we want to use these samples to estimate $\mathbf{E}[X]$. Using t samples, we use $\sum_{i=1}^{t} X_i/t$ for our estimate of $\mathbf{E}[X]$. We want the estimate to be within $\epsilon\mathbf{E}[X]$ from the true value of $\mathbf{E}[X]$ with probability at least $1-\delta$. We may not be able to use Chernoff's bound directly to bound how good our estimate is if X is not a $0-1$ random variable, and we do not know its moment generating function. We develop an alternative approach that requires only having a bound on the variance of X. Let $r = \frac{\sqrt{\mathbf{Var}[X]}}{\mathbf{E}[X]}$.

   (a) Show using Chebyshev's inequality that $O(\frac{r^2}{\epsilon^2\delta})$ samples are sufficient to solve the above problem.

   (b) Suppose that we only need a weak estimate that is within $\epsilon\mathbf{E}[X]$ of $\mathbf{E}[X]$ with probability at least $3/4$. Argue that only $O(\frac{r^2}{\epsilon^2})$ samples are enough for this weak estimate.

   (c) Show that by taking the median of $O(\log \frac{1}{\delta})$ weak estimates, we can obtain an estimate within $\epsilon\mathbf{E}[X]$ of $\mathbf{E}[X]$ with probability at least $1-\delta$. Conclude that we only need $O(\frac{r^2 \log 1/\delta}{\epsilon^2})$ samples.

2. A casino is testing a new class of simple slot machines. Each game, the player puts in one dollar, and the slot machine is supposed to return either three dollars to the player with probability $4/25$, one hundred dollars with proability $1/200$, and nothing with all remaining probability. Each game is supposed to be independent of other games.

   The casino has been surprised to find in testing that the machines have lost ten thousand dollars over the first millions games. Derive a Chernoff bound for the probability of this event. You may want to use a calculator or program to help you choose appropriate values as you derive your bound.

3. In many wireless communication systems, each receiver listens on a specific frequency. The bit $b(t)$ sent at time t is represented by a 1 or -1. Unfortunately, noise from other nearby communications can affect the receiver's signal. A simplified model of this noise is the following: there are n other senders, and the ith has strength p_i. At any time t the ith sender is also trying to send a bit $b_i(t)$, represented by 1 or -1. The receiver obtains the signal $s(t)$ given by

$$s(t) = b(t) + \sum_{i=1}^{n} p_i b_i(t).$$

If $s(t)$ is closer to 1 than -1, the receiver assumes that the bit sent at time t was a 1; otherwise, the receiver assumes that is was a -1.

Assume that all the bits $b_i(t)$ can be considered independent, uniform random variables. Give a Chernoff bound to estimate the probability that the receiver makes an error in determining $b(t)$.

4. Recall that a function f is said to be convex if for any x_1, x_2, and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

(a) Let Z be a random variable that takes on a (finite) set of values in the interval $[0, 1]$, and let $p = \mathbf{E}[Z]$. Define the Bernoulli random variable X by $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$. Show that $\mathbf{E}[f(Z)] \leq \mathbf{E}[f(X)]$ for any convex function f.

(b) Use the fact that $f(x) = \mathrm{e}^{tx}$ is convex for any $t \geq 0$ to obtain a Chernoff-like bound for Z based on a Chernoff bound for X.

5. We prove that the randomized Quicksort algorithm sorts a set of n numbers in time $O(n \log n)$ with high probability. Consider the following view of Quicksort. Every point in the algorithm where it decides on a pivot element is called a *node*. Suppose the size of the set to be sorted at a particular node is S. The node is called *good* if the pivot element divides the set into two parts, each of size not exceeding $2S/3$. Otherwise the node is called *bad*. The nodes can be thought of as forming a tree in which the root node has the whole set to be sorted and its children have the two sets formed after the first pivot step and so on.

(a) Show that the number of good nodes in any path from the root to a leaf in the above tree is not greater than $c \log_2 n$, where c is some positive constant.

(b) Show that with high probability (greater than $1 - 1/n^2$), the number of nodes in a given root to leaf path of the above tree is not greater than $c' \log_2 n$ where c' is another constant.

(c) Show that with high probability (greater than $1 - 1/n$), the number of nodes in the longest root to leaf path is not greater than $c' \log_2 n$. (Hint: How many nodes are there in the tree?)

(d) Use the above to show that the running time of Quicksort is $O(n \log n)$ with probability $1 - 1/n$.

6. In this problem, we design a randomized algorithm for the following packet routing problem: we are given a network which is an undirected connected graph G where nodes represent processors and the edges between the nodes represent wires. We are also given a set of N packets to route. For each packet we are given a source node, a destination node, and the exact route (path in the graph) that the packet should take from the source to its destination. (We may assume that there are no loops in the path.) In each time step, only one packet can traverse an edge. A packet can wait at any node during any time step and we assume unbounded queue sizes at each node.

   A schedule for a set of packets specifies the timing for the movement of packets along their respective routes. That is, it specifies which packet should move and which should wait at each time step. Our goal is to produce a schedule for the packets that tries to minimize the total time and the maximum queue size needed to route all the packets to their destinations.

   (a) The dilation d is the maximum distance traveled by any packet. The congestion c is the maximum number of packets that must traverse a single edge during the entire course of the routing. Argue that the time required for any schedule should be at least $\Omega(c + d)$.

   (b) Consider the following unconstrained schedule, where many packets may traverse an edge during a single time step. Assign each packet an integral delay chosen randomly, independently and uniformly from the interval $[1, \frac{\alpha c}{\log(Nd)}]$, where α is a constant. A packet that is assigned a delay of x waits in its source node for x time steps, and then moves on to its final destination through its specified route without ever stopping. Give an upper bound on the probability that more than $O(\log(Nd))$ packets use a particular edge e at a particular time step t.

   (c) Again using the unconstrained schedule above, show that the probability that more than $O(\log(Nd))$ packets pass through any edge at any time step is at most $1/(Nd)$ for a sufficiently large α.

   (d) Use the unconstrained schedule to devise a simple randomized algorithm that with high probability produces a schedule following the constraint of only one packet crossing an edge per time step of length $O(c + d \log(Nd))$ using queues of size $O(\log(Nd))$.

2 Martingales

1. Show that if $Z_0, Z_1, \ldots Z_n$ is a martingale with respect to $X_0, X_1, \ldots X_n$, then it is a martingale with respect to itself.

2. Let $X_0 = 0$ and for $j \geq 0$ let X_{j+1} be chosen uniformly over the real interval $[X_j, 1]$. Show that for $k \geq 0$ the sequence

$$Y_k = 2^k(1 - X_k)$$

is a martingale.

3. Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with expectation 0 and variance $\sigma^2 < \infty$. Let

$$Z_n = \left(\sum_{i=1}^{n} X_i\right)^2 - n\sigma^2.$$

Show that $Z_1, Z_2, \ldots$ is a martingale.

4. Consider an n-cube with $N = 2^n$ nodes. Let S be a non-empty set of vertices on the cube, and let x be a random vertex chosen uniformly at random among all vertices of the cube. Let $D(x, S)$ be the minimum number of coordinates that x and y differ in over all points $y \in S$. Give a bound on

$$\Pr(|D(x, S) - \mathbf{E}[D(x, S)]| > \lambda).$$

5. A *subsequence* of a string s is any string that can that can be obtained by deleting characters from s. Consider two strings x and y of length n, where each character in each string is independently a 0 with probability 1/2 and a 1 with probability 1/2. We consider the *longest common subsequence* of the two strings.

   (a) Show that the expected length of the longest common subsequence is greater than $c_1 n$ and less than $c_2 n$ for constants $c_1 > 1/2$ and $c_2 < 1$ when n is sufficiently large. (Any constants c_1 and c_2 are sufficient; as a challenge, you may attempt to find the best constants c_1 and c_2 that you can.)

   (b) Use a martingale inequality to show that the length of the longest common subsequence is highly concentrated around its mean.

6. Given a bag with r red balls and g green balls, suppose that we uniformly sample n balls from the bin without replacement. Set up an appropriate martingale and use it to show that the number of red balls in the sample is tightly concentrated around $\frac{nr}{r+g}$.

7. Consider a random graph from $G_{n,N}$, where $N = cn$ for some constant $c > 0$. Let X be the expected number of isolated vertices, that is, vertices of degree 0.

   (a) Determine $\mathbf{E}[X]$.

(b) Show that

$$\Pr(|X - \mathbf{E}[X]| \geq 2\lambda\sqrt{cn}) \leq 2e^{-\lambda^2/2}.$$

(Hint: use a martingale that reveals the locations of the edges that are in the graph, one at a time.)

8. We improve our bound from the Azuma-Hoeffding inequality for the problem where m balls are thrown into n bins. We let F be the number of empty bins after the m balls are thrown, and we let X_i be the bin in which the i-th ball lands. We define $Z_0 = \mathbf{E}[F]$, and $Z_i = \mathbf{E}[F \mid X_1, \ldots, X_i]$.

(a) Suppose that the number of bins that are empty after the i-th ball is thrown is A_i. Show that in this case

$$Z_{i-1} = A_{i-1}\left(1 - \frac{1}{n}\right)^{m-i+1}.$$

(b) Show that if the i-th ball lands in a bin that is empty when it is thrown, then

$$Z_i = (A_{i-1} - 1)\left(1 - \frac{1}{n}\right)^{m-i}.$$

(c) Show that if the i-th ball lands in a bin that is not empty when it is thrown, then

$$Z_i = A_{i-1}\left(1 - \frac{1}{n}\right)^{m-i}.$$

(d) Show that the Azuma-Hoeffding inequality in Theorem **??** applies with $d_i = \left(1 - \frac{1}{n}\right)^{m-i}$.

(e) Using the above, prove that

$$\Pr(|F - \mathbf{E}[F]| \geq \lambda) \leq 2\mathrm{e}^{-\lambda^2(2n-1)/(n^2-(\mathbf{E}[F])^2)}.$$

3 Exercises - Uniform Convergence

1. Consider a range space $(X, \mathcal{C})$ where $X = \{1, 2, \ldots, n\}$ and $\mathcal{C}$ is the set of all subsets of X of size k for some $k < n$. What is the VC dimenion of $\mathcal{C}$?

2. Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all axis-aligned rectangles in $\mathbb{R}^2$. That is, $c \in \mathcal{C}$ if for some $x_0 < x_1$ and $y_0 < y_1$, $c = \{(x, y) \in \mathbb{R}^2 \mid x_0 \leq x \leq x_1 \text{ and } y_0 \leq y \leq y_1\}$.

(a) Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 4. You should show both a set of four points that can be shattered, and show that no larger set can be shattered.

(b) Construct and analyze a PAC learning algorithm for the concept class of all axis-aligned rectangles in $\mathbb{R}^2$.

3. Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all axis-aligned squares in $\mathbb{R}^2$. Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 3.

4. Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all squares (that need not be axis-aligned) in $\mathbb{R}^2$. Show that the VC dimension of $(\mathbb{R}^2, \mathcal{C})$ is equal to 5.

5. Consider a range space $(\mathbb{R}^2, \mathcal{C})$ of all axis-aligned rectangular boxes in $\mathbb{R}^3$. Find the VC dimension of $(\mathbb{R}^2, \mathcal{C})$; you should show both the largest number of points that can be shattered, and show that no larger set can be shattered.

6. Prove that the VC dimension of the collection of all closed disks on the plane is 3.

7. Prove that the VC dimension of the range space $(\mathbb{R}^d, \mathcal{R})$, where $\mathcal{R}$ is the set of all half-spaces in $\mathbb{R}^d$, is at least $d+1$, by showing that the set consisting of the origin $(0, 0, \ldots, 0)$ and the d unit points $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0), \ldots, (0, 0, \ldots, 1)$ is shattered by $\mathcal{R}$.

8. Let $S = (X, R)$ and $S' = (X, R')$ be two range spaces. Prove that if $R' \subseteq R$ then the VC dimension of S' is no larger than the VC dimension of S.

9. Given a set of functions $\mathcal{F}$ and constants $a, b \in \mathbb{R}$, consider the set of functions

$$\mathcal{F}_{a,b} = \{af + b \mid f \in \mathcal{F}\}.$$

Let $R_m()$ and $\tilde{R}_m()$ denote the Rademacher complexity and the empirical Rademacher complexity, respectively. Prove that

(a) $\tilde{R}_m(\mathcal{F}_{a,b}) = |a|\tilde{R}_m(\mathcal{F})$,

(b) $R_m(\mathcal{F}_{a,b}) = |a|R_m(\mathcal{F})$.