Midterm Solutions

Problem 1

(a) For \(i = 1, \ldots, n\), let \(\mathbb{E}X_i = \mu_i\). Without loss of generality, suppose that \(\mu_1 \leq \ldots \leq \mu_n\). Suppose that \(\mu_H \leq n\), otherwise trivially we have that \(P(X \geq (1 + \delta)\mu_H) = 0\), which is upper bounded by any positive quantity.

Consider the sequence \(\mu_i^H\) defined as follow. Let \(j\) be the integer such that \(\sum_{i=1}^{j-1}(1 - \mu_i) < \mu_H - \mu\) and \(\sum_{i=1}^{j}(1 - \mu_i) \geq \mu_H - \mu\). For \(i = 1, \ldots, j-1\), let \(\mu_i^H = 1\). Let \(\mu_j^H = \mu_j + \left(\mu_H - \mu - \sum_{i=1}^{j-1}(1 - \mu_i)\right)\). For \(i = j+1, \ldots, n\), let \(\mu_i^H = \mu_i\).

We want to prove that \(\sum_{i=1}^{n}\mu_i^H = \mu_H\). We have that:

\[
\sum_{i=1}^{n}\mu_i^H = \mu + \left(\sum_{i=1}^{j-1}(1 - \mu_i)\right) + \left(\mu_H - \mu - \sum_{i=1}^{j-1}(1 - \mu_i)\right) = \mu_H
\]

Note that the construction above it’s always feasible if \(\mu \leq \mu_H \leq n\). The idea is that \(X_1, \ldots, X_n\) are Bernoulli random variable, and we want to construct a sequence \(Z_1, \ldots, Z_n\), such that \(Z_i \geq X_i\) and \(Z_i \sim B(\mu_i^H)\). In order to do so, we increase the values \(\mu_1, \ldots, \mu_j\) to \(\mu_1^H, \ldots, \mu_j^H\) so that \(\sum_{i=1}^{n}\mu_i^H = \mu_H\). That’s why we are considering the quantities \((1 - \mu_i)\), which is how much we can make \(Z_i\) more probable to be 1 compared to \(X_i\). Note that if \(\mu_H \leq n\), it’s always to construct such \(\mu_i^H\) sequence.

Note that \(Z_1, \ldots, Z_{j-1}\) are always equal to 1, so \(Z_i \geq X_i\) for \(i = 1, \ldots, j-1\). We define \(Z_i = X_i\) for \(i = j+1, \ldots, n\). If \(X_j = 1\), we define \(Z_j = X_j\), if \(X_j = 0\), then we set \(Z_j = 1\) w.p. \(\frac{\mu_H^j - \mu_j}{1 - \mu_j}\), and \(Z_j = 0\) w.p. \(1 - \frac{\mu_H^j - \mu_j}{1 - \mu_j}\).

It’s clear that for \(i \neq j\), \(Z_i \sim B(\mu_i^H)\). Also, by using law of total probability, and the fact that \(Y_i\) is equal to 1 iff \(X_i = 1\), we have that:

\[
\mathbb{E}Z_j = P(Z_j = 1|X_j = 1)\mu_j + P(Z_j = 1|Z_j = 0)(1 - \mu_j) = \mu_j + (1 - \mu_j)\frac{\mu_H^j - \mu_j}{1 - \mu_j} = \mu_H^j
\]
Hence \( Z_j \sim B(\mu_H^j) \). The variables \( Z_1, \ldots, Z_n \) are clearly 0–1 independent random variables (\( Z_i \) depends on \( X_i \), but \( Z_i \) is independent to \( Z_k \) for \( i \neq k \), as \( X_i \) is independent to \( X_k \) by assumption). Let \( Z = \sum_{i=1}^n Z_i \), by linearity of expectation we have that \( \mathbb{E}Z = \sum_{i=1}^n \mu_H^i = \mu_H \). Also, by construction we have that \( Z \geq X \).

Fix a \( \delta > 0 \). As \( Z \geq X \), the event \( X \geq (1+\delta)\mu_H \) implies that \( Z \geq (1+\delta)\mu_H \).

Hence, we have that:

\[
\mathbb{P}(X \geq (1+\delta)\mu_H) \leq \mathbb{P}(Z \geq (1+\delta)\mu_H) \leq \left( \frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu_H}
\]

First inequality is due to the fact that the first event implies the other, the second inequality is the standard Chernoff bound.

(b) The strategy is similar to the one discussed in point (a), but in this case we want to construct a sequence of independent 0–1 random variables \( Y_1, \ldots, Y_n \) such that \( Y_i \leq X_i \) for \( i = 1, \ldots, n \) and the sequence is less likely to have values equal to 1, so that \( \mathbb{E} \sum_{i=1}^n Y_i = \mu_L \).

Assume that \( \mu_L \geq 0 \), otherwise we trivially have that \( \mathbb{P}(X \leq (1-\delta)\mu_L) = 0 \) for any \( 0 < \delta < 1 \). Consider the sequence \( \mu_L^i \) defined as follow. Let \( j \) be the integer such that \( \sum_{i=1}^{j-1} \mu_i < \mu - \mu_L \) and \( \sum_{i=1}^{j-1} \mu_i \geq \mu - \mu_L \). For \( i = 1, \ldots, j-1 \), we define \( \mu_L^i = 0 \). For \( i = j+1, \ldots, n \), we define \( \mu_L^i = \mu_i \).

We let \( \mu_L^j = \mu_j - \left( \mu - \mu_L - \sum_{i=1}^{j-1} \mu_i \right) \). We have that:

\[
\sum_{i=1}^n \mu_L^i = \mu - \left( \sum_{i=1}^{j-1} \mu_i \right) - \left( \mu - \mu_L - \sum_{i=1}^{j-1} \mu_i \right) = \mu_L
\]

We construct a sequence of 0–1 random variables \( Y_1, \ldots, Y_n \) as follows. For \( i = 1, \ldots, j-1 \), let \( Y_i = 0 \) and note that \( Y_i \sim B(\mu_L^i) \) and \( Y_i \leq X_i \). For \( i = j+1, \ldots, n \), let \( Y_i = X_i \) and note that \( Y_i \leq X_i \) and \( Y_i \sim B(\mu_L^i) \). Let \( Y_j = 0 \) if \( X_j = 0 \), else if \( X_j = 1 \), then \( Y_j = 1 \) w.p. \( \frac{\mu_L^j}{\mu} \) and \( Y_j = 0 \) w.p. \( 1 - \frac{\mu_L^j}{\mu} \). Clearly, \( Y_j \leq X_j \). Also, by law of total probability:

\[
\mathbb{E}Y_j = \mathbb{P}(Y_j = 1|X_j = 1)\mathbb{P}(X_j = 1) = \frac{\mu_L^j}{\mu_j} \cdot \mu_j = \mu_L^j
\]

thus \( Y_j \sim B(\mu_L^j) \). As argued for the variables \( Z_i \), variables \( Y_1, \ldots, Y_n \) are independent (but \( Y_i \) depends on \( X_i \)), and we have that \( \sum_{i=1}^n Y_i := Y \leq X \)
and by linearity of expectation $\mathbb{E}Y = \mu_L$. To conclude, for any positive $\delta < 1$, we have that:

$$
\mathbb{P}(X \leq (1 - \delta)\mu_L) \leq \mathbb{P}(Y \leq (1 - \delta)\mu_L) \leq \left(\frac{e^{-\delta}}{(1 - \delta)(1 - \delta)}\right)^{\mu_L}
$$

The first inequality is due to the fact that the first event implies the other, as $Y \leq X$, the second inequality is standard Chernoff’s bound.

2. Let $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}X = nP \leq n$. We want to bound:

$$
\mathbb{P}(P \notin [X - \epsilon, X + \epsilon]) = \mathbb{P}(P < X - \epsilon) + \mathbb{P}(P > X + \epsilon) = \mathbb{P}(X > \mu + n\epsilon) + \mathbb{P}(X < \mu - n\epsilon)
$$

Observe that if $\epsilon \geq 1$, then the probability is trivially 0. Hence, we focus on the case $0 < \epsilon < 1$. The first equality is due to the fact that the events are disjoint, the second equality is obtained multiplying both sides of the argument of the probability by $n$.

Let $\mu_H = (\mu + n\epsilon)/(1 + \epsilon)$. Note that $\mu_H \geq \mu \iff \mu \leq n$ which is always true. Similarly, it’s straightforward to see that $\mu_H \leq n$. Hence, we have that:

$$
\mathbb{P}(X > \mu + n\epsilon) = \mathbb{P}(X > \mu_H(1 + \epsilon)) \leq \left(\frac{e^\epsilon}{(1 + \epsilon)(1 + \epsilon)}\right)^{\frac{\mu + n\epsilon}{1 + \epsilon}} \leq \left(\frac{e^\epsilon}{(1 + \epsilon)(1 + \epsilon)}\right)^{n\epsilon} \quad (*)
$$

The first inequality is thanks to the first part of the problem, the second inequality follows from the fact that the argument inside the brackets is less than 1 (if this wasn’t the case, the bound would be vacuous).

Similarly, let $\mu_L = \frac{\mu - n\epsilon}{1 - \epsilon}$. Again, it’s easy to see that $\mu_L \leq \mu$ iff $\mu \leq n$, which is true. We have that:

$$
\mathbb{P}(X < \mu - n\epsilon) = \mathbb{P}(X < \mu_L(1 - \epsilon)) \leq \left(\frac{e^{-\epsilon}}{(1 - \epsilon)(1 - \epsilon)}\right)^{\frac{\mu - n\epsilon}{1 - \epsilon}}
$$

Note that the bounds above is true only if $\mu - n\epsilon \geq 0 \iff \epsilon \leq P$, otherwise we have that $\mathbb{P}(X < \mu - n\epsilon) = 0$. Let:

$$
f(\epsilon) = \left(\frac{e^{+\epsilon}}{(1 + \epsilon)(1 + \epsilon)}\right)^{\frac{\mu + n\epsilon}{1 + \epsilon}}
$$

We have that if $\epsilon \geq P$, then $\mathbb{P}(P \notin [X - \epsilon, X + \epsilon]) \leq f(\epsilon)$. Note that in this case we can remove the dependency by $\mu$ by using the weaker bound $(*)$. If $\epsilon \leq P$, then $\mathbb{P}(P \notin [X - \epsilon, X + \epsilon]) \leq f(\epsilon) + f(-\epsilon)$. 


If $\epsilon < P/2$, we have that $n\epsilon \leq \mu - n\epsilon$, hence we can obtain a weaker bound that does not depend by $\mu$. 

$$P( P \notin [X - \epsilon, X + \epsilon]) \leq \left( \frac{e^\epsilon}{(1 + \epsilon)(1 + \epsilon)} \right)^{\frac{n\epsilon}{1+\epsilon}} + \left( \frac{e^{-\epsilon}}{(1 - \epsilon)(1 - \epsilon)} \right)^{\frac{n\epsilon}{1-\epsilon}}$$
Problem 2

a. If a packet must do $d$ hops to reach its destination, then it is clear that the time required for any scheduling is $\geq d$. Also, if an edge must be traversed by at least $c$ packets, any scheduling requires $\geq c$ time as there is the constraint that only one packet can traverse the same edge for any time step. By combining these two lower bounds, we obtain that the time required by any scheduling is at least $\geq \max\{c, d\}$. We conclude that any scheduling requires time $\Omega(c + d)$ by observing that $\max\{c, d\} \geq (c + d)/2$.

b. Number the packets from 1 to $N$. Let $e \in E$ be an edge of the graph. Denote with $S(e) = \{i : \text{pkt } i \text{ traverses edge } e\}$. Note that for any edge $e$, $|S(e)| \leq c$. Let $C_{i,e}(t)$ be 1 if packet $i$ traverses edge $e$ at time $t$, 0 otherwise. For any $i \in S(e)$, $C_{i,e}(t)$ is a Bernoulli random variable and $\mathbb{E}C_{i,e}(t) \leq 1/(\lceil \alpha c \log(Nd) \rceil) \leq \frac{\log(Nd)}{\alpha c}$. This is due to the fact that the packet $i$ waits a random time between 1 and $\lceil \alpha c \log(Nd) \rceil$, so the probability that it traverses at time $t$ is at most the probability that it chooses the right value in this interval to start its routing (at most, as this may not be possible if $t$ is too large). Note that $C_{i,e}(t)$ and $C_{j,e}(t)$ are independent for $i, j \in S(e)$, with $i \neq j$. Let $B_1, \ldots, B_N$ be $N$ independent Bernoulli random variables with mean $\frac{\log(Nd)}{\alpha c}$.

Let $\mu > 0$ be a constant. We have that for any edge $e$ and time $t$:

$$\mathbb{P}\left(\sum_{i \in S(e)} C_{i,e}(t) \geq \mu \cdot \frac{\log(Nd)}{\alpha c}\right) \leq \mathbb{P}\left(\sum_{i=1}^{c} B_i \geq \mu \cdot \frac{\log(Nd)}{\alpha}\right)$$

where we used the fact that $|S(e)| \leq c$, i.e. we are taking the sum over more elements. We observe that $\mathbb{E}(\sum_{i=1}^{c} B_i) = \frac{\log(Nd)}{\alpha}$, hence for constant $\mu \geq 6$, we can apply Chernoff’s bound and obtain that:

$$\mathbb{P}\left(\sum_{i \in S(e)} C_{i,e}(t) \geq \mu \cdot \frac{\log(Nd)}{\alpha c}\right) \leq 2^{-\mu \cdot \frac{\log(Nd)}{\alpha}}$$

This yields an upper bound to the requested probability, as $\mu \log(Nd)/(\alpha c) = O(\log(Nd))$.

c. We observe that any packet arrives to its destination in time $\leq d + \lceil \frac{\alpha c}{\log(Nd)} \rceil$, hence for any packet $i$ and edge $e$, $C_{i,e}(t) = 0$ if $t > d + \lceil \frac{\alpha c}{\log(Nd)} \rceil$. Also, any packet can traverse at maximum $d$ edges, hence the number of edges that are traversed by at least one packet in the graph is at most $Nd$. 

Thus, we want to bound the sum \( \sum_{i \in S(e)} C_{i,e}(t) \) only for \( \leq Nd \) edges and \( \leq d + \left\lceil \frac{c\alpha}{\log(Nd)} \right\rceil \) possible time steps. By an union bound, we have that:

\[
\gamma = \mathbb{P}\left( \text{at any time, an edge in the network is traversed by } \geq \mu \cdot \frac{\log(Nd)}{\alpha c} \text{ packets} \right) \leq Nd \left( d + \left\lceil \frac{c\alpha}{\log(Nd)} \right\rceil \right) 2^{-\mu \frac{\log(Nd)}{\alpha}}
\]

We observe that \( c \leq N \), hence \( c\alpha \leq N\alpha \). Assuming that \( \alpha \geq 1 \), we have that \( \left( d + \left\lceil \frac{c\alpha}{\log(Nd)} \right\rceil \right) \leq N\alpha d \). Hence:

\[
\gamma \leq \alpha(Nd)^2 2^{-\mu \frac{\log(Nd)}{\alpha}}
\]

We choose \( \alpha = 1 \), \( \mu = 6 \), and we obtain that:

\[
\gamma \leq (Nd)^{2-6} \leq \frac{1}{Nd}
\]

Therefore, with probability \( \geq 1 - \frac{1}{Nd} \), any edge is traversed by at most \( 6 \log(Nd) / c \leq 6 \log(Nd) = O(\log(Nd)) \) at any time step. Equivalently, the probability that more than \( O(\log(Nd)) \) packets pass through any edge at any time step is at most \( 1/(Nd) \).

d. We use the strategy of point b., i.e. the random schedule where every node waits a random time between 1 and \( \left\lceil \frac{c\alpha}{\log(Nd)} \right\rceil \). The time required by this schedule in the unconstrained setting is at most \( O(d + \left\lceil \frac{c\alpha}{\log(Nd)} \right\rceil) \), where the second term is due to the maximum waiting time. To adapt this strategy in the constrained setting, where an edge can be traversed by at most one packet at any given time, we do as follows. We substitute the \( i \)-th time step of the schedule in the unconstrained setting with \( 6 \log(Nd) \) time steps in the constrained setting. The strategy is to simulate the schedule of the unconstrained setting: if in the \( i \)-th time step of the unconstrained schedule some packets traverse an edge, those same packets will traverse the edge in the time steps \( i \cdot 6 \log(Nd) + 1, \ldots, (i + 1)6 \log(Nd) \) in the constrained setting, and they will continue their routing only after the \( (i + 1)6 \log(Nd) \) time step. Similarly, if a packet must wait \( t \) time steps in the unconstrained setting, then it will wait \( t \cdot 6 \log(Nd) \) in the constrained setting.

Thanks to the analysis of points b and c, we know that with high probability no edge is traversed by more than \( 6 \log(Nd) \) with high probability, i.e. \( \geq 1 - 1/(Nd) \geq 1 - 1/N \). Hence, the simulation works as in \( 6 \log(Nd) \) time steps all the packets that must traverse an edge (following the unconstrained
strategy) can do so. This simulation requires that every edge has a buffer of size $6 \log(Nd) = O(6 \log(Nd))$, as in the simulation of a unconstrained time step at most $6 \log(Nd)$ packets can queue to an edge.

As the scheduling in the unconstrained setting requires at most $d + \lceil \frac{c\alpha}{\log(Nd)} \rceil$ time steps, the constrained scheduling strategy requires $(d + \lceil \frac{c\alpha}{\log(Nd)} \rceil) \cdot 6 \log(Nd) = O(c + d \log(Nd))$ time steps with probability $\geq 1 - 1/N$. 


Problem 3

a) We want to show that the VC-dimension \( d \) of \((C, X)\) is \( n + 1 \). Given a point \( x \in X \), denote with \(|x|\) the number of 1’s in it. Let \( S = \{x_0, \ldots, x_{n+1}\} \), where \(|x_i| = i\), i.e. all points in \( S \) have a different number of 1’s. For any set \( C \subseteq S \), there is a function \( f \in C \) such that \( f(x) = 1 \) iff \( x \in C \). This function \( f \) is the symmetric function that is positive only if the number of 1 is \(|x| \colon x \in C\). Hence \( C \) shatters \( S \), thus \( d \geq n + 1 \).

I want to show that it’s not possible to shatter a set of size \( n + 2 \), and let \( S \) be such a set. For the pigeonhole principle, there must be two points \( x, y \in S \) such that \(|x| = |y|\). From the definition of symmetric function, this implies that for any \( f \in C \) we have that \( f(x) = f(y) \), hence the points \( x \) and \( y \) will be always given the same label. Hence \( S \) cannot be shattered. This proves that \( d = n + 1 \).

b) Let \( m \) be the size of the training set. Let \( D \) be the uniform distribution over \( X = \{0, 1\}^n \). Fix a value \( 0 < \gamma < 1 \). Let \( p_n \) be the probability that a sample \( x \sim D \) is such that \((1 - \gamma)^\frac{n}{2} \leq |x| \leq (1 + \gamma)^\frac{n}{2}\). Observe that since we are considering an uniform distribution, we have that \( \mathbb{E}_{x \sim D} |x| = n/2 \).

By Chernoff’s inequality, as the number \(|x|\) can be seen as the sum of \( n \) independent fair coin flips, we have that:

\[
1 - p_n = \mathbb{P}_{x \sim D} \left[ |x| \leq (1 - \gamma)^\frac{n}{2} \right] + \mathbb{P}_{x \sim D} \left[ |x| \geq (1 + \gamma)^\frac{n}{2} \right] \leq 2e^{-\frac{n^2}{6}}
\]

The probability that for all samples we have that they have a number of 1’s within \((1 - \gamma)^\frac{n}{2}\) and \((1 + \gamma)^\frac{n}{2}\) is:

\[
(p_n)^m \geq (1 - 2e^{-\frac{n^2}{6}})^m \geq 1 - 2me^{-\frac{n^2}{6}}
\]

We observe that \( m = O \left( \frac{n+1}{\epsilon} \log \frac{n+1}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon} \right) \). Fixed the values of \( \epsilon \) and \( \delta \), we have that \( m = O(n \log n) \). Hence, we observe that \( m \cdot e^{-O(n)} \) converges to 0 as \( n \) grows. Hence, for any \( \gamma \), we can make arbitrarily unlikely the event that a sample falls outside of the interval of interest (for \( \gamma \geq 1 \) this event is impossible). Note that this implies that, given a training set, with very high probability (for large \( n \)), we decide on our hypothesis only by looking at the value of the functions for number of 1’s between \((1 - \gamma)^\frac{n}{2}\) and \((1 + \gamma)^\frac{n}{2}\). So our chosen hypothesis can be wrong in giving the correct answer for number of 1’s outside of this interval (as in the training set there are no examples of it). However, the error is weighted according to the probability to obtain
a certain element in the domain, and the probability of getting an element with a number of 1' between \((1 - \gamma)\frac{n}{2}\) and \((1 + \gamma)\frac{n}{2}\) is very unlikely. Hence, its contribution to the error will be very low. This is why, even if we don't get examples in our training set for a large interval of possible number of 1's, we can still have a small error. This discussion explains how there is no contradiction with the theorem.
Problem 4

Let $d$ be the VC-dimension of the rangespace $(V, R)$. We want to show that $d < 3$.

Consider three distinct vertices $u, v, w \in V$, such that $w$ is along the shortest path from $u$ to $v$. We want to show that $R$ cannot shatter $S = \{u, v, w\}$, hence proving that $d < 3$. Observe that such a $u, v$ and $w$ do not exist only if all the paths have length $\leq 2$, and in that case I can never find a path that contain three points, hence $d < 3$. Therefore, we assume that we can find such vertices.

Notation: we denote with $r_{st} \in R$ the shortest path between vertices $s$ and $t$.

Vertex $w$ is along the shortest path from $u$ to $v$. I want to show that there is no $r \in R$ such that $r \cap S = \{u, v\}$. Suppose by contradiction that such $r$ exists. This path $r$ represents the shortest path between two vertices, call them $x, y \in R$, and the path $r_{xy}$. As $u, v \in r_{xy}$, we can write the path as the concatenation of three paths $r_{xy} = r_1 \cup r_2 \cup r_3$, where $r_1$ is a path from $x$ to $u$, $r_2$ is a path from $u$ to $v$ and $r_3$ is a path from $v$ to $y$. By assumption, $r_{xy}$ does not contain $w$, therefore $r_2$ is not the shortest path from $u$ to $v$ (as this shortest path is unique, and the shortest path from $u$ to $v$ contains $w$).

Consider now the path $r_{xy}' = r_1 \cup r_{uv} \cup r_2$ from $x$ to $y$, where we replaced $r_2$ with $r_{uv}$. As $|r_{uv}| < |r_2|$, we have that $|r_{xy}'| < |r_{xy}|$ which is a contradiction.

To wrap it up, we can conclude that $R$ cannot shatter $S$. Therefore, we have that $d < 3 = |S|$.