Problem 1

Given a set $T^d(s)$ which is defined as follows:

$$T^d(s) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \le s, \|\theta\|_2 \le 1\} \tag{1}$$

For a subset $V \subseteq \mathbb{R}^d$, we have $\mathcal{G}(V)$ defined as:

$$\mathcal{G}(V) = \mathbb{E}[\max_{\boldsymbol{v}\in V} \boldsymbol{v}^\top \boldsymbol{w}] \tag{2}$$

where $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \forall i \in [d]$.

(a) Let $\mathcal{S}$ be defined as $\mathcal{S} = \{S \subseteq [d] : |S| = s\}$. Let $\boldsymbol{v}_S$ and $\boldsymbol{w}_S$ be the subvectors of $\boldsymbol{v}$ and $\boldsymbol{w}$ respectively, indexed by $S \in \mathcal{S}$. $\forall \boldsymbol{v} \in T^d(s)$, define $\mathcal{P}_{\boldsymbol{v}} = \{\boldsymbol{v}_S : \forall S \in \mathcal{S}\}$ and $\mathcal{P} = \{\boldsymbol{v}_S : \boldsymbol{v}_S \in \mathcal{P}_v, \forall \boldsymbol{v} \in T^d(s)\}$. By definition, if $\boldsymbol{v} \in T^d(s), \exists S \in \mathcal{S}$, such that $\boldsymbol{v}^\top \boldsymbol{w} = \sum_{i=1}^d v_i w_i = \sum_{i\in S} v_i w_i = \boldsymbol{v}_S^\top \boldsymbol{w}_S$. So,

$$\begin{aligned}
\mathcal{G}(T^d(s)) &= \mathbb{E}\left[\max_{\boldsymbol{v}\in T^d(s)} \boldsymbol{v}^\top \boldsymbol{w}\right] \\
&= \mathbb{E}\left[\max_{\boldsymbol{v}_S\in \mathcal{P}} \boldsymbol{v}_S^\top \boldsymbol{w}_S\right] \\
&= \mathbb{E}\left[\max_{S\in \mathcal{S}} \boldsymbol{v}_S^\top \boldsymbol{w}_S\right] \\
&\le \mathbb{E}\left[\max_{S\in \mathcal{S}} \|\boldsymbol{v}_S\|_2 \|\boldsymbol{w}_S\|_2\right] && \text{Cauchy-Schwarz inequality} \\
&\le \mathbb{E}\left[\max_{S\in \mathcal{S}} \|\boldsymbol{w}_S\|_2\right] && \|\boldsymbol{v}_S\| \le 1
\end{aligned}$$

(b) We use the following result presented in the class.

Theorem 1. *Let $(X_1, \cdots, X_n)$ be a vector of i.i.d. Gaussian random variables such that $X_i \sim \mathcal{N}(0,1)$ and $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz w.r.t. ℓ_2-norm, then $f(X) - \mathbb{E}[f(X)]$ is sub-Gaussian with parameter at most L. In particular,*

$$P(f(X) - \mathbb{E}[f(X)] > \delta) \le \exp\left(\frac{-\delta^2}{2L^2}\right) \tag{3}$$

We derive an upper bound on $\mathbb{E}[\|w_S\|_2]$ as follows:

$$\mathbb{E}[\|w_S\|_2] = \mathbb{E}\left[\sqrt{\sum_i w_i^2}\right] \overset{(a)}{\le} \sqrt{\mathbb{E}\left[\sum_i w_i^2\right]} = \sqrt{\sum_i \mathbb{E}[w_i^2]} = \sqrt{s} \tag{4}$$

where (a) follows from Jensen's inequality and the last statement follows because $\mathbb{E}[w_i^2] = 1$ for $w_i \sim \mathcal{N}(0,1)$. Note that $f(x) = \|x\|_2$ is 1-Lipschitz, i.e, $|\,\|x\|_2 - \|y\|_2\,| \le 1 \cdot \|x - y\|_2$ using triangle inequality. So, for Gaussian vector w_S, we apply Theorem 1 and obtain:

$$\begin{aligned} P(\|w_S\|_2 - \sqrt{s} > \delta) &\overset{(4)}{\leq} P(\|w_S\| - \mathbb{E}[\|w_S\|] > \delta) \\ &\leq \exp(-\delta^2/2) \\ \implies P(\|w_S\|_2 > \sqrt{s} + \delta) &\leq \exp(-\delta^2/2) \end{aligned}$$

(c) Note that $\mathcal{S} = \binom{d}{s}$. Using the result from part (a), we get:

$$\begin{aligned} \mathcal{G}(T^d(s)) &\leq \mathbb{E}\left[\max_{S\in\mathcal{S}} \|\boldsymbol{w}_S\|_2\right] \\ &= \mathbb{E}[\|\boldsymbol{w}_S\|_2] + \mathbb{E}\left[\max_{S\in\mathcal{S}} (\|\boldsymbol{w}_S\|_2 - \mathbb{E}[\|\boldsymbol{w}_S\|_2])\right] \\ &\overset{(4)}{\leq} \sqrt{s} + \mathbb{E}\left[\max_{S\in\mathcal{S}} (\|\boldsymbol{w}_S\|_2 - \mathbb{E}[\|\boldsymbol{w}_S\|_2])\right] \\ &\overset{(a)}{\leq} \sqrt{s} + \sqrt{2\ln\binom{d}{s}} \\ &\leq \sqrt{s} + \sqrt{2s\ln\left(\frac{ed}{s}\right)} && \binom{d}{s} \leq \left(\frac{ed}{s}\right)^s \end{aligned}$$

The statement (a) holds because from Theorem 1, we get $\|\boldsymbol{w}_S\|_2 - \mathbb{E}[\|\boldsymbol{w}_S\|_2]$ is sub-Gaussian with variance 1. So, by invoking Massart's Lemma over set $\mathcal{S}$, we get the result.

■

Problem 2

Choose the potential function ϕ_t as defined in (5):

$$\phi_t = \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} \exp(-\eta L_t(i)) \right) \tag{5}$$

where $L_t(i)$ is the cumulative loss incurred by ith expert till time step t. i.e, $L_t(i) = \sum_{t=1}^{T} l_t(i)$.

$$\begin{aligned}
\phi_t - \phi_{t-1} &= \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} \exp(-\eta L_t(i)) \right) - \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} \exp(-\eta L_{t-1}(i)) \right) \\
&= \frac{1}{\eta} \ln \left(\frac{\sum_{i=1}^{N} \exp(-\eta l_t(i)) \cdot \exp(-\eta L_{t-1}(i))}{\sum_{i=1}^{N} \exp(-\eta L_{t-1}(i))} \right) \\
&= \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} p_t(i) \cdot e^{-\eta l_t(i)} \right) && \text{where } p_t(i) = \frac{e^{-\eta L_{t-1}(i)}}{\sum_{j=1}^{N} e^{-\eta L_{t-1}(j)}} \\
&\leq \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} p_t(i) \left(1 - \eta l_t(i) + \eta l_t(i)^2\right) \right) && e^{-x} \leq 1 - x + x^2, x > 0 \\
&= \frac{1}{\eta} \ln \left(1 - \eta \langle p_t, l_t \rangle + \eta^2 \langle p_t, l_t^2 \rangle\right) \\
&\leq -\langle p_t, l_t \rangle + \eta \langle p_t, l_t^2 \rangle && e^x \geq 1 + x, x \in \mathbb{R}
\end{aligned}$$

Summing from $t = 1$ to $t = T$,

$$\phi_T - \phi_0 \leq -\sum_{t=1}^{T} \langle p_t, l_t \rangle + \sum_{t=1}^{T} \eta \langle p_t, l_t^2 \rangle \leq -\widehat{L}_T + \eta T \tag{6}$$

where the last inequality holds due to the assumption that $l_t(i) \in [0, 1] \implies \langle p_t, l_t^2 \rangle \leq \langle p_t, \mathbf{1} \rangle = 1$ and by definition $\widehat{L}_T = \sum_{t=1}^{T} \langle p_t, l_t \rangle$. As the initial losses are 0, clearly $\phi_0 = (\ln N)/\eta$. Let $\mathcal{S} = \{i : L_T(i) \leq L, i \in [N]\}$ and $|\mathcal{S}| = N_L$. So, we get:

$$\begin{aligned}
\phi_T &= \frac{1}{\eta} \ln \left(\sum_{i=1}^{N} \exp(-\eta L_T(i)) \right) \\
&= \frac{1}{\eta} \ln \left(\sum_{i \in \mathcal{S}} \exp(-\eta L_T(i)) + \sum_{i \notin \mathcal{S}} \exp(-\eta L_T(i)) \right) \\
&\geq \frac{1}{\eta} \ln \left(\sum_{i \in \mathcal{S}} \exp(-\eta L_T(i)) \right) \\
&\geq \frac{1}{\eta} \ln \left(\sum_{i \in \mathcal{S}} \exp(-\eta L) \right) \\
&= -L + \frac{1}{\eta} \ln N_L
\end{aligned} \tag{7}$$

Using (6) and (7), we get,

$$
\begin{aligned}
-L + \frac{1}{\eta}\ln N_L - \frac{1}{\eta}\ln N &\leq \phi_T - \phi_0 \\
&\leq -\widehat{L}_T + \eta T \\
\implies \widehat{L}_T &\leq L + \frac{1}{\eta}\ln\frac{N}{N_L} + \eta T \qquad (8)
\end{aligned}
$$

■

Problem 3

We reduce the boosting problem to an online learning setting. The N training examples $\{(x_i, f(x_i))\}_{i=1}^{N}$ are considered as the N experts. We have access to a hypothesis class $\mathcal{H}$ where $\exists h \in \mathcal{H}$, such that $P(h(x) \neq f(x)) \leq \frac{1}{2} - \gamma \implies P(h(x) = f(x)) \geq \frac{1}{2} + \gamma$, where $f(\cdot)$ is the target function and $f(x) \in \{\pm 1\}$. Counter-intuitively, we consider our loss function as $l_t(i) = \mathbf{1}(h_t(x_i) = f(x_i))$ for time step t and expert i. Let p_t be the probability distribution obtained from Hedge algorithm at time step t. So, the total loss upto time T obtained is given as:

$$\widehat{L}_T = \sum_{t=1}^{T}\sum_{i=1}^{N} p_t(i)\mathbf{1}(h_t(x_i) = f(x_i)) = \sum_{t=1}^{T} P(h_t(x) = f(x)) \geq T\left(\frac{1}{2} + \gamma\right) \tag{9}$$

Let $\mathcal{S} = \{i : H(x_i) \neq f(x_i)\}$, where $H(x) = \text{sign}\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right) \forall t, \alpha_t > 0, \sum_{t=1}^{T}\alpha_t = 1$. Choose a distribution such that $\alpha_t = \frac{1}{T}, \forall t \in [T] \implies H(x) = \text{sign}\left(\sum_{t=1}^{T} h_t(x)\right)$. Observe that $\forall i \in \mathcal{S}, f(x_i)\sum_{t=1}^{T} h(x_i) \leq 0$. Let L_i be the loss attained by any expert in $\mathcal{S}$. So,

$$\begin{aligned}
\frac{L_i}{T} &= \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(h_t(x_i) = f(x_i)) \\
&= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}(1 + h_t(x_i)f(x_i)) && f(x_i), h(x_i) \in \{\pm 1\} \\
&= \frac{1}{2} + \frac{1}{T}\sum_{t=1}^{T} h_t(x_i)f(x_i) \\
&= \frac{1}{2} + \frac{1}{T}\cdot f(x_i)\sum_{t=1}^{T} h_t(x_i) \leq \frac{1}{2} \qquad (10)
\end{aligned}$$

So, $L_T(i)/T \leq 1/2, \forall i \in \mathcal{S}$. We want to show that there is a hypothesis in the class of weighted majority vote functions that misclassifies at most an ϵ fraction $\implies |\mathcal{S}| \leq \epsilon N$. Applying the regret bound for Hedge with many good experts from (8)(here, i is a good expert if $i \in \mathcal{S}$):

$$\begin{aligned}
\frac{\widehat{L}_T}{T} &\leq \frac{L}{T} + \frac{1}{\eta T}\ln\frac{N}{|\mathcal{S}|} + \eta \\
\implies \frac{1}{2} + \gamma &\overset{(9)}{\leq} \frac{L}{T} + \frac{1}{\eta T}\ln\frac{N}{|\mathcal{S}|} + \eta \\
&\overset{(a)}{\leq} \frac{1}{2} + 2\sqrt{\frac{1}{T}\ln\frac{N}{|\mathcal{S}|}} \\
\implies \gamma &\leq 2\sqrt{\frac{1}{T}\ln\frac{N}{|\mathcal{S}|}}
\end{aligned}$$

The inequality (a) is a direct consequence of (10) and using $\eta^* = \sqrt{\frac{1}{T}\ln\frac{1}{\epsilon}}$ to get the strictest upper bound. To ensure that $|\mathcal{S}| \leq \epsilon N \implies \frac{N}{|\mathcal{S}|} \geq \epsilon \implies RHS \geq \sqrt{(1/T)\ln(1/\epsilon)}$, we need the following condition to hold true:

$$\gamma \geq 2\sqrt{\frac{1}{T}\ln\frac{1}{\epsilon}} \implies T \geq \frac{4}{\gamma^2}\ln\frac{1}{\epsilon} = \mathcal{O}\left(\frac{1}{\gamma^2}\ln\frac{1}{\epsilon}\right)$$

So, for $T = \mathcal{O}\left(\frac{1}{\gamma^2}\ln\frac{1}{\epsilon}\right)$, there exists a hypothesis in the class of weighted majority functions such that the misclassfication error is atmost ϵ.

■

Problem 4

We have to determine k locations of ones in a d-dimensional binary vector b. We query a binary vector ϕ which gives an output $y = \vee_i \phi_i b_i$. Suppose we make a total of T queries to achieve probability of error atmost ϵ. We need to find an lower bound on T for two different cases. Consider $\boldsymbol{X} = [b_1, \ldots, b_T]^\top$ to be a $T \times d$ matrix containing the ground truth value of b for the T queries, $\boldsymbol{Y} = [y_1, \ldots, y_T]^\top$ to be a T-dimensional binary vector containing the query results and $\widehat{\boldsymbol{X}} = [\hat{b}_1, \ldots, \hat{b}_T]^\top$ to be a $T \times d$ matrix which is the estimated value of the vector after each of the T queries[1].

(a) We first consider the noiseless case. Notice that $\boldsymbol{X} \to \boldsymbol{Y} \to \widehat{\boldsymbol{X}}$ form a Markov Chain. So, by data processing inequality we have:

$$I(\boldsymbol{X}, \boldsymbol{Y}) \geq I(\boldsymbol{X}, \widehat{\boldsymbol{X}}) \tag{11}$$

Furthermore, using Fano's inequality, we get:

$$1 + P_e \log |\mathcal{X}| \geq H(\boldsymbol{X}|\widehat{\boldsymbol{X}}) \geq H(\boldsymbol{X}|\boldsymbol{Y}) \tag{12}$$

where H is the Shannon entropy, $P_e = P(\boldsymbol{X} \neq \widehat{\boldsymbol{X}})$. We need to find conditions for $P_e \leq \epsilon$. Observe that $|\mathcal{X}| = \binom{d}{k} \geq (d/k)^k$. So, from (12),

$$\begin{aligned}
1 + \epsilon \log \binom{d}{k} &\geq 1 + P_e \log |\mathcal{X}| \\
&\geq H(\boldsymbol{X}|\boldsymbol{Y}) \\
&= H(\boldsymbol{X}) - I(\boldsymbol{X}, \boldsymbol{Y}) && \text{By definition of mutual information} \\
&= H(\boldsymbol{X}) + H(\boldsymbol{Y}|\boldsymbol{X}) - H(\boldsymbol{Y}) && \text{By definition of mutual information} \\
&\geq H(\boldsymbol{X}) - H(\boldsymbol{Y}) && H(\boldsymbol{X}|\boldsymbol{Y}) \geq 0 \\
&= \log \binom{d}{k} - \log 2^T
\end{aligned}$$

The last equality holds because $\boldsymbol{X}$ and $\boldsymbol{Y}$ can take any value uniformly in their respective alphabet spaces. So, $H(\boldsymbol{X}) = \log |\mathcal{X}|$ and $H(\boldsymbol{Y}) = \log |\mathcal{Y}|$. Rearranging the terms, we get:

$$\begin{aligned}
\log 2^T &\geq (1-\epsilon) \log \binom{d}{k} - 1 \\
\implies T &\geq (1-\epsilon) \log \binom{d}{k} - 1 \\
\implies T &\geq (1-\epsilon) k \log(n/k) - 1
\end{aligned}$$

(b) In the noisy case, we introduce another intermediate random value $\widehat{\boldsymbol{Y}}$ which is a T-dimensional binary vector containing the values after elements of $\boldsymbol{Y}$ are randomly flipped. We further know

[1] Basic idea of the proof follows: Chan, C. L., Che, P. H., Jaggi, S., Saligrama, V. (2011, September). Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In 2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton) (pp. 1832-1839). IEEE.

that the bits are flipped with a probability $q \in [0, 1/2]$. So, in this case $\boldsymbol{X} \to \boldsymbol{Y} \to \widehat{\boldsymbol{Y}} \to \widehat{\boldsymbol{X}}$ forms a Markov Chain. Using data processing inequality, we obtain:

$$I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) \geq I(\boldsymbol{Y}, \widehat{\boldsymbol{X}}) \geq I(\boldsymbol{X}, \widehat{\boldsymbol{X}}) \tag{13}$$

Following the steps as done in the noiseless case, we obtain:

$$\begin{aligned}
1 + \epsilon \log \binom{d}{k} &\geq 1 + P_e \log |\mathcal{X}| \\
&\geq H(\boldsymbol{X} | \widehat{\boldsymbol{X}}) \\
&= H(\boldsymbol{X}) - I(\boldsymbol{X}, \widehat{\boldsymbol{X}}) && \text{By definition of mutual information} \\
&\overset{(13)}{\geq} H(\boldsymbol{X}) - I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) \\
\implies I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) &\geq (1 - \epsilon) \log \binom{d}{k} - 1 && H(\boldsymbol{X}) = \log |\mathcal{X}| = \log \binom{d}{k} \qquad (14)
\end{aligned}$$

Now, we need to upper bound $I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})$.

$$\begin{aligned}
I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) &= H(\widehat{\boldsymbol{Y}}) - H(\widehat{\boldsymbol{Y}} | \boldsymbol{Y}) \\
&= \sum_{t=1}^{T} H(\widehat{\boldsymbol{Y}}_t) - H(\widehat{\boldsymbol{Y}}_t | \boldsymbol{Y}_t) && \text{As } \boldsymbol{Y}_t\text{'s are independent} \\
&= T \left(H(\widehat{\boldsymbol{Y}}_1) - H(\widehat{\boldsymbol{Y}}_1 | \boldsymbol{Y}_1) \right) && \text{As } \boldsymbol{Y}_t\text{'s are identical} \\
&\leq T \left(1 - h(q)\right) && (15)
\end{aligned}$$

where $h(q)$ is the usual binary entropy function. The last inequality holds because $\widehat{\boldsymbol{Y}}_1$ being a binary r.v. has $H(\widehat{\boldsymbol{Y}}) \leq 1$ and $H(\widehat{\boldsymbol{Y}}_1 | \boldsymbol{Y}_1) = H(\widehat{\boldsymbol{Y}}_1 | \boldsymbol{Y}_1 = 0) P(\boldsymbol{Y}_1 = 0) + H(\widehat{\boldsymbol{Y}}_1 | \boldsymbol{Y}_1 = 1) P(\boldsymbol{Y}_1 = 1) = h(q)(P(\boldsymbol{Y}_1 = 0) + P(\boldsymbol{Y}_1 = 1)) = h(q)$. Using (14) and (15), we get:

$$\begin{aligned}
T(1 - h(q)) &\geq I(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) \\
&\geq (1 - \epsilon) \log \binom{d}{k} - 1 \\
&\geq (1 - \epsilon) k \log(n/k) - 1 \\
\implies T &\geq \frac{(1 - \epsilon) k \log(n/k) - 1}{1 - h(q)}
\end{aligned}$$

■