1. This exam is open notes and open laptop, but you must turn off the wireless.

2. Please write clearly and give rigorous justification for each non-trivial step.

3. You may cite results from the course notes without proof.

4. When we ask for an upper bound, to get full credit, in addition to being mathematically correct, you must have the right dependencies on all the problem-dependent quantities (e.g., if it is possible to get $d$, then $d \log d$ is not acceptable, but $2d$ is acceptable).
1. **Regularized exponential families (20 points)**

Consider a standard exponential family:

\[ p_\theta(x) = \exp \{ \phi(x) \cdot \theta - A(\theta) \}, \tag{1} \]

where \( \theta \in \mathbb{R}^d \) are the parameters, \( \phi: \mathcal{X} \to [0, 1]^d \) is a bounded feature map, and \( A(\theta) = \log \sum_{x \in \mathcal{X}} \exp \{ \phi(x) \cdot \theta \} \) is the log-partition function.

Suppose we have \( n \) i.i.d. data points \( x_1, \ldots, x_n \sim p_{\theta^*} \), where \( \theta^* \) are the true parameters. Define the expected and empirical risks, respectively, in terms of the negative log-likelihood:

\[ L(\theta) = \mathbb{E}_{x \sim p^*} [-\log p_\theta(x)], \tag{2} \]
\[ \hat{L}(\theta) = \frac{1}{n} \sum_{i=1}^n -\log p_\theta(x_i). \tag{3} \]

In class, we obtained generalization bounds by uniform convergence over all \( \theta \in \mathbb{R}^d \), but in this problem we will take a different route.

**a. (5 points)**

Define \( \mu^* = \mathbb{E}_{x \sim p^*}[\phi(x)] \) and \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^n \phi(x_i) \) to be the expected and empirical sufficient statistics.

Show that the expected and empirical risks differ by an amount depending on how much the sufficient statistics deviate:

\[ |\hat{L}(\theta) - L(\theta)| \leq \|\mu^* - \hat{\mu}\|_\infty \|\theta\|_1. \tag{4} \]

**Solution:**

A simple calculation:

\[ \hat{L}(\theta) - L(\theta) = [A(\theta) - \hat{\mu} \cdot \theta] - [A(\theta) - \mu^* \cdot \theta] \] [definition of \( L, \hat{L} \)]
\[ = (\mu^* - \hat{\mu}) \cdot \theta \] [cancel]
\[ = \|\mu^* - \hat{\mu}\|_\infty \|\theta\|_1 \] [Hölder’s inequality]. \tag{5}

The same calculation applies to the negated version: \( L(\theta) - \hat{L}(\theta) \). Putting the two together yields the result.
b. (5 points)
Show that with probability at least $1 - \delta$, the empirical sufficient statistics concentrates:

$$\|\hat{\mu} - \mu^*\|_{\infty} \leq \sqrt{\frac{\log(2d) + \log(1/\delta)}{2n}} \overset{\text{def}}{=} \lambda_n.$$  \hspace{1cm} (8)

Here, we have defined $\lambda_n$ to be the upper bound.

Solution:
For each component $j = 1, \ldots, d$, we have that $\hat{\mu}_j$ is just an average of bounded i.i.d. random variables with mean $\mu_j^*$ (recall that $\phi_j(x) \in [0, 1]$), so applying Hoeffding’s inequality yields:

$$\mathbb{P}[\hat{\mu}_j \geq \mu_j^* + \epsilon] \leq \exp\{-2n\epsilon^2\}. \hspace{1cm} (9)$$

Applying Hoeffding’s bound in the other direction, and taking a union bound over all $j = 1, \ldots, d$ yields:

$$\mathbb{P}[\|\hat{\mu} - \mu^*\|_{\infty} \geq \epsilon] \leq 2d \exp\{-2n\epsilon^2\}. \hspace{1cm} (10)$$

Setting the RHS to $\delta$ and solving for $\epsilon$ yields the result.
c. (5 points)
Define the following $L_1$-regularized estimator:

$$\hat{\theta} \overset{\text{def}}{=} \arg \min_{\theta} \hat{L}(\theta) + \lambda_n \|\theta\|_1. \quad (11)$$

Show that this estimator enjoys the following bound on excess risk: with probability at least $1 - \delta$,

$$L(\hat{\theta}) \leq L(\theta^*) + 2\lambda_n \|\theta^*\|_1. \quad (12)$$

(Recall that $\theta^*$ is the parameter of the true distribution; i.e., $x_i \sim p_{\theta^*}$).

How come we did not have to employ uniform convergence arguments over the entire hypothesis space of $\theta \in \mathbb{R}^d$ in order to control $\hat{\theta}$? Explain briefly in one or two sentences.

Solution:
Condition on the event $A$ that $d_n \overset{\text{def}}{=} \|\mu^* - \hat{\mu}\|_\infty \leq \lambda_n$.

$$L(\hat{\theta}) \leq \hat{L}(\hat{\theta}) + d_n \|\hat{\theta}\|_1 \quad \text{[apply (a) to \hat{\theta}]} \quad (13)$$

$$\leq \hat{L}(\hat{\theta}) + \lambda_n \|\hat{\theta}\|_1 \quad \text{[event A holds]} \quad (14)$$

$$\leq \hat{L}(\theta^*) + \lambda_n \|\theta^*\|_1 \quad \text{[\hat{\theta} is the minimizer]} \quad (15)$$

$$\leq L(\theta^*) + 2\lambda_n \|\theta^*\|_1 \quad \text{[apply (a) again to \theta^*]} \quad (16)$$

$$\leq L(\theta^*) + 2\lambda_n \|\theta^*\|_1 \quad \text{[event A holds]} \quad (17)$$

By (b), event $A$ (which by the previous calculation implies the result) holds with probability at least $1 - \delta$.

Note that (a) already holds deterministically and uniformly for all $\theta$. Therefore, we only need to bound the difference $\|\mu^* - \hat{\mu}\|_\infty$ once to get a result for all $\theta$, including $\hat{\theta}$ and $\theta^*$. 

4
d. (5 points)
Now let us consider the unregularized estimator ($\lambda_n = 0$). Compute the asymptotic distribution of $n(L(\hat{\theta}) - L(\theta^*))$. Explain briefly in what regimes the $L_1$-regularized estimator would be superior to the unregularized estimator. Your explanation should describe the regimes in terms of $n, d, ||\theta^*||_1, ||\hat{\theta}||_1$.

Solution:
By the general asymptotics result from class,
\[
n(L(\hat{\theta}) - L(\theta^*)) \xrightarrow{d} \frac{1}{2} \text{tr} W(I, d).
\]  
(18)
Therefore,
\[
L(\hat{\theta}) - L(\theta^*) \sim \frac{d}{2n}.
\]
From part (b), in the case $\lambda_n > 0$,
\[
L(\hat{\theta}) - L(\theta^*) \leq 2\|\theta^*\|_1 \sqrt{\frac{\log(2d) + \log(1/\delta)}{2n}}.
\]
Regularization performs better when
\[
4\|\theta^*\|_1^2 \cdot n \cdot \frac{\log d}{d^2} \lesssim 1.
\]
For example, this holds when $n$ is small compared to $d$, in particular when
\[
n \lesssim \frac{d^2}{\log d},
\]
or when $\|\theta^*\|_1$ is small compared to $n$, in particular when
\[
\|\theta^*\|_1 \lesssim \frac{1}{\sqrt{n}}.
\]
2. Convergence with kernels (20 points)

In this problem, we will show convergence of functions in an RKHS. Tip: if kernels make your head hurt, first think about what the answer should be for finite-dimensional vectors.

a. (7 points)

Let’s start with a basic result. Let \( k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \) be a kernel and \( \mathcal{H} \) be its associated RKHS. Assume \( k(x, x) \leq C^2 \) for all \( x \in \mathcal{X} \). Let \( p^* \) be a distribution over \( \mathcal{X} \). Our goal is to estimate the mean function \( f^* = \mathbb{E}_{x \sim p^*}[k(x, \cdot)] \in \mathcal{H} \).

We obtain \( n \) i.i.d. data points \( x_1, \ldots, x_n \sim p^* \) and form the following empirical estimate:

\[
\hat{f} = \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot).
\]

Show that with probability at least \( 1 - \delta \), the estimated function is close to the true function in the sense that

\[
D_n \overset{\text{def}}{=} \| \hat{f} - f^* \|_{\mathcal{H}} \leq C \left( 1 + \sqrt{\frac{2 \log(1/\delta)}{n}} \right).\]

Solution:

Let us apply McDiarmid’s inequality. To show bounded differences of \( D_n \), consider changing \( x_i \) to \( x'_i \), and let \( D'_n \) be the resulting quantity defined on \( x_1, \ldots, x'_i, \ldots, x_n \). Let us write \( \| \cdot \| \) for \( \| \cdot \|_{\mathcal{H}} \). Then we have that:

\[
D_n - D'_n = \| \hat{f} - f^* \| - \| \hat{f}' - f^* \|
\]

\[
= \| \hat{f} - \hat{f}' \| \quad \text{[triangle inequality]}
\]

\[
= \frac{1}{n} \| k(x_i, \cdot) - k(x'_i, \cdot) \| \quad \text{[definition]}
\]

\[
= \frac{1}{n} (\| k(x_i, \cdot) \| + \| k(x'_i, \cdot) \|) \quad \text{[triangle inequality]}
\]

\[
\leq \frac{2C}{n}. \quad \| k(x, \cdot) \| = \sqrt{k(x, x)} \leq C
\]

We can also bound the expectation:

\[
\mathbb{E}[D_n] \leq \sqrt{\mathbb{E}[D_n^2]} \quad \text{[Jensen’s inequality]}
\]

\[
= \frac{1}{n} \sqrt{\sum_{i=1}^{n} \mathbb{E}[\| k(x_i, \cdot) - f^* \|^2]} \quad \text{[cross terms have mean zero]}
\]

\[
\leq \frac{C}{\sqrt{n}} \quad \text{[variance is upper bounded by second moment].}
\]

Applying McDiarmid’s inequality, we get that

\[
\mathbb{P}[D_n \geq \mathbb{E}[D_n] + \epsilon] \leq \exp \left( \frac{-2\epsilon^2}{\sum_{i=1}^{n} \left( \frac{2C}{n} \right)^2} \right)
\]

\[
= \exp \left( \frac{-n \epsilon^2}{2C^2} \right).
\]
b. (7 points) 

Now let’s apply (a) to something interesting. We have an unknown distribution \( p^* \) over \((x, y) \in \mathcal{X} \times \mathcal{Y}\). Our goal is to ascertain whether \( x \) and \( y \) are independent. Suppose we have kernels \( k_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) and \( k_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \) with \( k_x(x, x) \leq C^2 \) for all \( x \in \mathcal{X} \) and \( k_y(y, y) \leq C^2 \) for all \( y \in \mathcal{Y} \).

For notational convenience, we will extend \( k_x \) and \( k_y \) to operate on the product space, so that

\[
k_x((x, y), (x', y')) = k_x(x, x'), \quad \text{and} \quad k_y((x, y), (x', y')) = k_y(y, y').
\]

This way, we can write \( k_x k_y \) for the product kernel; formally, \((k_x k_y)((x, y), (x', y')) = k_x(x, x')k_y(y, y')\).

Now let \( H_x \) be the RKHS corresponding to kernel \( k_x \), \( H_y \) be the RKHS corresponding to kernel \( k_y \), \( H_{xy} \) be the RKHS corresponding to kernel \( k_xk_y \). Define \( H \) to be \( H_x + H_y + H_{xy} \), i.e.,

\[
H \overset{\text{def}}{=} \{ f = f_x + f_y + f_{xy} : f_x \in H_x, f_y \in H_y, f_{xy} \in H_{xy} \}
\]

We also define the inner product in \( H \) as follows: for any two functions \( f = f_x + f_y + f_{xy} \in H \) and \( f' = f'_x + f'_y + f'_{xy} \in H \), where \( f_x, f'_x \in H_x, f_y, f'_y \in H_y, f_{xy}, f'_{xy} \in H_{xy} \), define:

\[
\langle f, f' \rangle_H \overset{\text{def}}{=} \langle f_x, f'_x \rangle_{H_x} + \langle f_y, f'_y \rangle_{H_y} + \langle f_{xy}, f'_{xy} \rangle_{H_{xy}}.
\]

Define the norm in the usual way: \( \| \cdot \|_H \overset{\text{def}}{=} \sqrt{\langle \cdot, \cdot \rangle_H} \).

Suppose we have i.i.d. points \( z_1, \ldots, z_n \sim p^* \) where each \( z_i = (x_i, y_i) \). Define the statistic \( D_n \) as follows, which measures the covariance of \( k_x(x, \cdot) \) and \( k_y(y, \cdot) \):

\[
A_n = \frac{1}{n} \sum_{i=1}^{n} k_x(z_i, \cdot),
\]

\[
B_n = \frac{1}{n} \sum_{i=1}^{n} k_y(z_i, \cdot),
\]

\[
T_n = \frac{1}{n} \sum_{i=1}^{n} k_x(z_i, \cdot)k_y(z_i, \cdot),
\]

\[
D_n = \| T_n - A_nB_n \|_H.
\]

(The idea is that for universal kernels such as the Gaussian kernel, \( D_n \overset{p}{\to} 0 \) exactly when \( x \) and \( y \) are independent...but don’t worry about this.)

Show that if \( x \) and \( y \) are independent, then with probability at least \( 1 - \delta \),

\[
D_n \leq \frac{3C^2(1 + \sqrt{2 \log(3/\delta)})}{\sqrt{n}} + \frac{C^2(1 + \sqrt{2 \log(3/\delta)})^2}{n}.
\]

For this part, you may use the result of (a) without proof.

Solution:

Let us apply (a) to bound the errors of \( A_n, B_n, T_n \) from their mean. Note that \( k_x(z, z), k_y(z, z) \leq C^2 \) and \( k_{xy}(z, z) \leq C^4 \).

Define the mean functions: \( f_x \overset{\text{def}}{=} \mathbb{E}[A_n] \), \( f_y \overset{\text{def}}{=} \mathbb{E}[B_n] \). Since \( x \) and \( y \) are independent, we also have: \( \mathbb{E}[T_n] = f_x f_y \).
With probability at least $1 - \delta$ (applying the union bound over the following three quantities):

\[
\|A_n - f_x\| \leq \epsilon, \tag{40}
\]
\[
\|B_n - f_y\| \leq \epsilon, \tag{41}
\]
\[
\|T_n - f_x f_y\| \leq C\epsilon, \tag{42}
\]
\[
\epsilon \overset{\text{def}}{=} \frac{C(1 + \sqrt{2 \log(3/\delta)})}{\sqrt{n}}. \tag{43}
\]

Therefore,

\[
\|T_n - A_n B_n\| = \|(T_n - f_x f_y) - (A_n B_n - f_x f_y)\| \tag{44}
\]
\[
= \|(T_n - f_x f_y) - [(f_x + (A_n - f_x))(f_y + (B_n - f_y)) - f_x f_y]\| \tag{45}
\]
\[
= \|(T_n - f_x f_y) - f_x(B_n - f_y) + f_y(A_n - f_x) + (A_n - f_x)(B_n - f_y)\| \tag{46}
\]
\[
\leq C\epsilon + \epsilon\|f_x\| + \epsilon\|f_y\| + \epsilon^2 \tag{47}
\]
\[
\leq 3C\epsilon + \epsilon^2. \tag{48}
\]

The result follows.
c. (6 points)

Now we actually need to be able to compute $D_n$. Write $D_n$ in terms of only kernel evaluations $k_x, k_y$ on the data points $z_1, \ldots, z_n$. You might find it useful to write $D_n$ as a sum over terms which you define separately to avoid having a long, hard-to-read equation.

How fast can you compute $D_n$? More specifically, find the smallest integer $r$ for which computing $D_n$ takes $O(n^r)$ time, assuming kernel evaluations take $O(1)$ time. Briefly justify your answer.

Solution:

$$D_n^2 \overset{\text{def}}{=} \| T_n - A_n B_n \|^2$$

$$= \| T_n \|^2 - 2 \langle T_n, A_n B_n \rangle + \| A_n B_n \|^2,$$  \hspace{1cm} (49)

$$\| T_n \|^2 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k_x(z_i, z_j) k_y(z_i, z_j),$$ \hspace{1cm} (50)

$$\langle T_n, A_n B_n \rangle = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{j'=1}^{n} k_x(z_i, z_j) k_y(z_i, z_j'),$$ \hspace{1cm} (51)

$$\| A_n B_n \|^2 = \frac{1}{n^4} \sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{j=1}^{n} \sum_{j'=1}^{n} k_x(z_i, z_j) k_y(z_i', z_j'),$$ \hspace{1cm} (52)

$$D_n \text{ is gotten by taking the square root. The factorization makes it clear that we can compute everything in } O(n^2) \text{ time.}$$
3. **Distinguishing distributions with classification (20 points)**

Recall that the total variation distance between two distributions \( p \) and \( q \) over \( \mathcal{X} \) is defined as the maximum deviation over measurable subsets \( A \subseteq \mathcal{X} \):

\[
D(p, q) \overset{\text{def}}{=} \sup_A |p(A) - q(A)|.
\] (56)

In practice, we have i.i.d. points \( x_1, \ldots, x_n \sim p \) and \( x'_1, \ldots, x'_n \sim q \). Let \( \hat{p} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \) and \( \hat{q} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x'_i} \) be the corresponding empirical distributions.

First, note that the estimator \( D(\hat{p}, \hat{q}) \) won’t work. If \( p \) and \( q \) have densities, then all the points are distinct with probability 1, and we can choose \( A = \{x_1, \ldots, x_n\} \) to make \( D(\hat{p}, \hat{q}) = 1 \) regardless of how close \( p \) and \( q \) are. In a sense, this is because all measurable subsets is too rich.

Let us instead consider a restricted class \( \mathcal{H} \) consisting of boolean functions \( h : \mathcal{X} \rightarrow \{0, 1\} \) which is closed under complement (if \( h \in \mathcal{H} \), then \( 1 - h \in \mathcal{H} \)). Then we can define:

\[
D_{\mathcal{H}}(p, q) \overset{\text{def}}{=} \sup_{h \in \mathcal{H}} |\mathbb{E}_{x \sim p}[h(x)] - \mathbb{E}_{x' \sim q}[h(x')]|.
\] (57)

In practice, we would compute \( D_{\mathcal{H}}(\hat{p}, \hat{q}) \).

**a. (5 points)**

Let’s reduce the problem of computing \( D_{\mathcal{H}} \) to something we’re familiar with. Define a loss function \( \ell \) and a scheme for constructing a set of examples given the original points \( x_{1:n} \sim p, x'_{1:n} \sim q \), so that minimizing the empirical risk yields \( D_{\mathcal{H}}(\hat{p}, \hat{q}) \).

**Solution:**

For each \( x_i \), create the example \((x_i, 1)\); for each \( x'_i \), create the example \((x'_i, 0)\). Let \( \ell \) be the zero-one loss. Then choosing \( h \) to maximize the distance is equivalent to maximizing the number of points \( x_i \) such that \( h(x_i) = 1 \) and minimizing the number of points \( x'_i \) such that \( h(x'_i) = 1 \) (without loss of generality, since \( 1 - h \in \mathcal{H} \)). Then the empirical risk of the ERM is \( 1 - D_{\mathcal{H}}(\hat{p}, \hat{q}) \). Intuitively, we are fitting a binary classifier to separate the points from \( p \) and \( q \). The distance is large exactly when the classification problem is easy.
b. (5 points)
Now that we are in familiar territory, let us consider possible hypothesis classes $\mathcal{H}$. Let $\mathcal{H}$ be the set of ellipsoids:

$$\mathcal{H} = \{ x \mapsto \mathbb{I}[x^T \Sigma x \geq 1] : \Sigma \in \mathbb{R}^{d \times d}, \Sigma \succeq 0 \}. \quad (58)$$

Compute an upper bound on the VC dimension of $\mathcal{H}$ (your answer need only contain the appropriate polynomial dependence on $d$—$O(1), O(d), O(d^2)$, etc.).

Solution:
The idea is that $\mathcal{H}$ is the set of functions obtained thresholding the class of linear functions:

$$\mathcal{F} = \{ x \mapsto \text{tr}(\Sigma x x^T) - 1 : \Sigma \succeq 0 \}. \quad (59)$$

So it suffices to compute the dimension of $\mathcal{F}$ since $\text{VC}(\mathcal{H}) \leq \dim(\mathcal{F})$ as showed in class. Here, $\mathcal{F}$ is a set of linear functions with a quadratic basis function $xx^T$, so its dimensionality is that of $\Sigma$. Clearly, it is bounded by $d^2$, which is good enough for this problem; a tighter bound would be $\frac{d(d+1)}{2}$, the actual dimension of PSD matrices.
c. (5 points)

The problem with ellipsoids is that they are not a very rich hypothesis class. To construct a richer hypothesis class, consider boolean functions constructed out of random features. Draw $r$ i.i.d. vectors $\omega_1, \ldots, \omega_r \sim \mathcal{N}(0, I_{d \times d})$. Define the feature map $\phi_r(x) = [h(\omega_1 \cdot x), \cdots, h(\omega_r \cdot x)]$, where $h(z) = \max\{0, z\}$ is the ReLU non-linearity. Then let the hypothesis class be:

$$
\mathcal{H}_r = \{x \mapsto \mathbb{I}[\theta \cdot \phi_r(x) \geq 0] : \theta \in \mathbb{R}^r\}.
$$

(60)

Assume $d \geq 2$. Show that with probability 1, the VC dimension of $\mathcal{H}_r$ is $\Theta(r)$ (i.e., provide both an upper bound $C_1r$ and a lower bound $C_2r$). Warning: this problem might be somewhat involved, so you might want to save it for last.

Solution:
First, define the underlying linear function class

$$
\mathcal{F}_r = \{x \mapsto \theta \cdot \phi_r(x) : \theta \in \mathbb{R}^r\}.
$$

(61)

The dimensionality of $\mathcal{F}_r$ is clearly upper bounded by $r$, and thus the VC dimension of $\mathcal{H}_r$ is also upper bounded by $r$.

To show a lower bound, we need to exhibit $\frac{r}{2}$ points that can be shattered by $\mathcal{H}_r$. We will construct $k = \frac{r}{2}$ points $X_1, \ldots, X_k$ so that the set of points $Y_j = \phi_r(X_j) \in \mathbb{R}^r$ are linearly independent. Since we are considering linear classifiers over the points $Y_j$ (i.e., $Y_j \mapsto \theta \cdot Y_j$, where $\theta \in \mathbb{R}^r$), we know that the points $Y_j$ can be shattered (to obtain a labeling $Q \in \{-1, +1\}^k$, simply solve $[Y_1, \ldots, Y_k] \theta = Q$). This shows that the original points $X_j$ can be shattered by $\mathcal{H}_r$.

It suffices to perform this construction for $d = 2$. Suppose that we are given $r$ samples $\omega_1, ..., \omega_r \in \mathbb{R}^2$ from $\mathcal{N}(0, I_{2 \times 2})$.

First, note that with probability 1, each $\omega_i$ is non-zero. Then, let $\theta_i$ denote the angle of the vector $\omega_i$ measured from the positive $x$-axis. Again with probability 1, each angle $\theta_i$ is distinct. Therefore, assuming without loss of generality that the $\omega_i$ are ordered by angle, we have

$$
0 \leq \theta_1 < \theta_2 < ... < \theta_r < 2\pi.
$$

Note that either at least half of the $\omega_i$ lie in the upper half-plane (if $\theta_k < \pi$, where $k = \lceil \frac{r}{2} \rceil$) or at least half of the $\omega_i$ lie in the lower half-plane (if $\theta_k \geq \pi$); without loss of generality, we assume that the former holds, so

$$
0 \leq \theta_1 < \theta_2 < ... < \theta_k < \pi.
$$

For each $1 \leq i \leq k$, the point $X = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$ satisfies $\omega_i \cdot X \leq 0$, since by our assumption $\omega_i$ is in the upper half-plane. The idea is to rotate clockwise from $X$; as we do so, we can construct vectors $X_j$ so that

$$
\begin{cases}
\omega_i \cdot X_j > 0 & \text{if } i \leq j \\
\omega_i \cdot X_j \leq 0 & \text{if } i > j
\end{cases}
$$

for each $1 \leq j \leq k$. Then, letting

$$
\phi_k(x) = [h(\omega_1 \cdot x), \cdots, h(\omega_k \cdot x)]
$$

be the first $k$ components of $\phi_r$, the vectors $Y_j = \phi_k(X_j)$ (i.e., $Y_{ji} = h(\omega_i \cdot X_j)$) satisfy

$$
\begin{cases}
Y_{ji} > 0 & \text{if } i \leq j \\
Y_{ji} = 0 & \text{if } i > j.
\end{cases}
$$
In other words, the matrix with rows $Y^T_j$ form an lower triangular matrix with non-zero diagonal entries:

$$
\begin{bmatrix}
Y_1^T \\
\vdots \\
Y_k^T 
\end{bmatrix}
= 
\begin{bmatrix}
Y_{11} & 0 & 0 & \cdots & 0 \\
Y_{21} & Y_{22} & 0 & \cdots & 0 \\
Y_{31} & Y_{32} & Y_{33} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Y_{k1} & Y_{k2} & Y_{k3} & \cdots & Y_{kk} 
\end{bmatrix}
$$

Therefore, the set of vectors $\{Y_j : 1 \leq j \leq k\}$ are linearly independent.

Summing up, if we can construct $k$ such vectors $X_j \in \mathbb{R}^2$, then the corresponding vectors $Y_j \in \mathbb{R}^k$ are linearly independent. As a consequence, the vectors $Y_j' = \phi_r(X_j)$ will also be linearly independent, so we can conclude that $\mathcal{H}_r$ has VC dimension at least $\frac{r^2}{2}$.

In the remainder of the proof, we formally construct $X_j$ satisfying the desired properties. Our approach to constructing $X_j$ is to first take a vector with angle between $\theta_j$ and $\theta_{j+1}$, and then rotate it by $\frac{\pi}{2}$ radians. In particular, let

$$Z_j = \hat{\omega}_j + \hat{\omega}_{j+1}$$

for each $1 \leq j \leq k$, where we take $\omega_{k+1} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ with corresponding angle $\theta_{k+1} = \pi$. Here, $\hat{v} = \frac{v}{\|v\|}$ denotes the unit vector in the direction of $v$. Now, let

$$X_j = AZ_j = \begin{bmatrix} Z_{j2} \\ -Z_{j1} \end{bmatrix}$$

where

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

is the rotation by $-\frac{\pi}{2}$ radians. Note that

$$\omega_i \cdot (A\hat{\omega}_j) = \|\omega_i\| \cos \left(\theta_i - \theta_j + \frac{\pi}{2}\right) = \|\omega_i\| \sin(\theta_j - \theta_i),$$

which by assumption is (i) positive if $i < j$, (ii) zero if $i = j$, or (iii) negative if $i > j$. Therefore,

$$\omega_i \cdot X_j = \omega_i \cdot (A\hat{\omega}_j) + \omega_i \cdot (A\hat{\omega}_{j+1})$$

is (i) positive if $i \leq j$ (since the first term is either zero or positive and the second term is positive) or (ii) negative if $i > j$ (since the first term is negative and the second term is either zero or negative). This completes the proof.
d. (5 points)
Suppose the data generating distribution \( p \) places probability only on points \( x \in \mathbb{R}^d \) that lie in a low-dimensional subspace—that is, there exists \( A \in \mathbb{R}^{d \times k} \) such that every point \( x \) in the support of \( p \) can be written as \( x = Az \) for some \( z \in \mathbb{R}^k \). Consider the class of linear functions:

\[
F \overset{\text{def}}{=} \{ x \mapsto I[w \cdot x \geq 0] : w \in \mathbb{R}^d \}.
\]

Show that the Rademacher complexity is upper bounded by

\[
R_n(F) \leq \sqrt{\frac{2k(\log n + 1)}{n}}.
\]

Solution:
Intuitively, we are working in a dimension \( k \) subspace, so we should only be paying for \( k \)-dimensional linear classifiers. Fix a set of \( n \) points \( X_1, ..., X_n \in \mathbb{R}^d \) in the support of \( p \). By assumption, we can write \( X_i = Az_i \) for each \( i \), where \( Z_i \in \mathbb{R}^k \). Consider the hypothesis class

\[
F' \overset{\text{def}}{=} \{ z \mapsto I[v \cdot z \geq 0] : v \in \mathbb{R}^k \}.
\]

Note that

\[
\hat{R}_n(F) = \mathbb{E} \left[ \sup_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \sigma_i I[w \cdot X_i \geq 0] \mid X_{1:n} \right]
\]

\[
= \mathbb{E} \left[ \sup_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \sigma_i I[w \cdot AZ_i \geq 0] \mid Z_{1:n} \right] \quad \text{[substitution]}
\]

\[
= \mathbb{E} \left[ \sup_{v \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i I[A^\top w \cdot Z_i \geq 0] \mid Z_{1:n} \right] \quad \text{[rearranging terms]}
\]

\[
\leq \mathbb{E} \left[ \sup_{v \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} \sigma_i I[v \cdot Z_i \geq 0] \mid Z_{1:n} \right] \quad \text{[since } A^\top w \in \mathbb{R}^k]\]

\[
= \hat{R}_n(F') \quad \text{[definition]}
\]

\[
\leq \sqrt{\frac{2k(\log n + 1)}{n}} \quad \text{[VC dimension of linear classifiers]}
\]

as claimed.