1. [20 points] TV distance: In studying the NP classifier we introduced the TV distance. For the entire question to simplify things we will assume everything is defined on a discrete domain Ω. Suppose we have two distributions P and Q with pmfs p and q. We defined the total variation distance between these distributions as:

\[ TV(P, Q) = \sum_{\{x \in \Omega : p(x) \geq q(x)\}} [p(x) - q(x)]. \]

This distance is closely related to the \( \ell_1 \) distance between the two distributions. Show that,

\[ TV(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |p(x) - q(x)|. \]

The TV distance is a very strong notion of distance. In particular, if the TV is small then the probability of any event under the two distributions must be close. Show that,

\[ TV(P, Q) = \sup_{A \subseteq \Omega} |P(A) - Q(A)|. \]

The Total Variation distance also belongs to a popular class of distances between probability distributions. It is an integral probability metric (other popular examples of IPMs include the Wasserstein distance). Show that, the TV distance can also be written as:

\[ TV(P, Q) = \frac{1}{2} \sup_{f : \|f\|_\infty \leq 1} |E_{X \sim P} f(X) - E_{Y \sim Q} f(Y)| \]

Solution:

Part A. Let \( E = \{x \in \Omega : p(x) \geq q(x)\} \) and \( E^c = \{x \in \Omega : p(x) < q(x)\} \). Under the event \( E \), it is clear to see that \( p(x) - q(x) = |p(x) - q(x)| \). Hence

\[ \sum_{x \in E} [p(x) - q(x)] = \sum_{x \in E} |p(x) - q(x)|. \] (1)

Now we claim that

\[ \sum_{x \in E} |p(x) - q(x)| = \sum_{x \in E^c} |p(x) - q(x)|. \] (2)
If this is the case, then the result follows by
\[
\text{TV}(P, Q) \overset{\text{(1)}}{=} \sum_{x \in E} |p(x) - q(x)| \overset{\text{(2)}}{=} \frac{1}{2} \sum_{x \in E} |p(x) - q(x)| + \frac{1}{2} \sum_{x \in E^c} |p(x) - q(x)|
\]
\[
= \frac{1}{2} \sum_{x \in \Omega} |p(x) - q(x)|.
\]

To prove (2), we have similarly to (1)
\[
- \sum_{x \in E^c} |p(x) - q(x)| = \sum_{x \in E^c} [p(x) - q(x)].
\]

Therefore
\[
\sum_{x \in E} |p(x) - q(x)| - \sum_{x \in E^c} |p(x) - q(x)| = \sum_{x \in E} [p(x) - q(x)] + \sum_{x \in E^c} [p(x) - q(x)]
\]
\[
= \sum_{x \in \Omega} p(x) - \sum_{x \in \Omega} q(x) = 0,
\]
where the last equality uses the fact that \(p(x)\) and \(q(x)\) are pmfs. Hence (2) follows and the proof is complete.

**Part B.** Let \(A\) be any arbitrary measurable set and \(E = \{x \in \Omega : p(x) \geq q(x)\}\) as before. Then
\[
P(A) - Q(A) \leq \sum_{x \in A \cap E} [p(x) - q(x)] \leq \sum_{x \in E} [p(x) - q(x)]
\]
since \(p(x) - q(x)\) is always nonnegative under the event \(E\). Similarly we have
\[
P(A) - Q(A) \geq \sum_{x \in A \cap E^c} [p(x) - q(x)] \geq \sum_{x \in E^c} [p(x) - q(x)].
\]
Combining these two inequalities and taking the supremum over \(A\),
\[
\sup_{A \subseteq \Omega} |P(A) - Q(A)| \leq \max \left\{ \sum_{x \in E} [p(x) - q(x)], \sum_{x \in E^c} [q(x) - p(x)] \right\}
\]
\[
= \text{TV}(P, Q).
\]
(3)

Set \(A = \{x \in \Omega : p(x) \geq q(x)\}\). Then it is clear to see that
\[
\sup_{A \subseteq \Omega} |P(A) - Q(A)| \geq \sum_{\{x \in \Omega : p(x) \geq q(x)\}} [p(x) - q(x)] = \text{TV}(P, Q).
\]
(4)
From (3) and (4), we conclude that TV($P, Q$) = sup$_{A \subseteq \Omega} |P(A) - Q(A)|$.

**Part C.** Based on the result from Part A, we have

$$|\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y)| = \left| \sum_{x \in \Omega} f(x) [p(x) - q(x)] \right| \leq \sum_{x \in \Omega} |f(x)||p(x) - q(x)|$$

$$\leq 2||f||_{\infty} TV(P, Q) \leq 2TV(P, Q).$$

The right-hand side does not depend on $f$. Therefore

$$\sup_{f: ||f||_{\infty} \leq 1} \mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y) \leq 2TV(P, Q).$$

By choosing $g(x) = I\{p(x) \geq q(x)\} - I\{p(x) < q(x)\}$ such that $||g||_{\infty} \leq 1$, we have

$$\sup_{f: ||f||_{\infty} \leq 1} \mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y) \geq \mathbb{E}_{X \sim P} g(X) - \mathbb{E}_{Y \sim Q} g(Y)$$

$$= \sum_{\{x \in \Omega: p(x) \geq q(x)\}} [p(x) - q(x)] + \sum_{\{x \in \Omega: p(x) < q(x)\}} [q(x) - p(x)]$$

$$= \sum_{x \in \Omega} |p(x) - q(x)| = 2TV(P, Q).$$

Therefore we conclude that TV($P, Q$) = $\frac{1}{2} \sup_{f: ||f||_{\infty} \leq 1} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y)|$.

2. **[10 points] Exponential LRT:** Suppose that, $X_1, \ldots, X_n$ is drawn from an Exponential distribution with density:

$$p(x|\theta) = \begin{cases} \exp(-(x - \theta)) & x \geq \theta, \\ 0 & \text{otherwise}. \end{cases}$$

Construct the (generalized) LRT statistic for distinguishing the hypotheses:

$$H_0 : \theta \leq \theta_0,$$

$$H_1 : \theta > \theta_0,$$

for some fixed $\theta_0$.

**Solution:**

Denote the minimum of $x_1, \ldots, x_n$ by $x_{(1)}$ and let $\mathbf{x} = \{x_1, \ldots, x_n\}$. Using these notations, the likelihood function is

$$L(\theta|x_1, \ldots, x_n) = \begin{cases} \exp (-\sum_{i=1}^n x_i + n\theta), & \text{if } \theta \leq x_{(1)}, \\ 0, & \text{if } \theta > x_{(1)}. \end{cases}$$
Note that $L(\theta|\mathbf{x})$ is an increasing function of $\theta$ on $\theta \leq x_{(1)}$. Hence

$$\sup_{\theta \in \Theta} L(\theta|\mathbf{x}) = L(x_{(1)}|\mathbf{x}) = \exp \left( -\sum_{i=1}^{n} x_i + nx_{(1)} \right).$$

On the other hand, the maximum of $L(\theta|\mathbf{x})$ is given as $L(\theta_0|\mathbf{x})$ under the null $H_0 : \theta \leq \theta_0$ and $x_{(1)} > \theta_0$. Hence the (generalized) LRT statistic is

$$\lambda(x_1, \ldots, x_n) = \begin{cases} 1, & \text{if } x_{(1)} \leq \theta_0, \\ \exp\{-n(x_{(1)} - \theta_0)\}, & \text{if } x_{(1)} > \theta_0. \end{cases}$$

We reject the null for a small value of $\lambda(x_1, \ldots, x_n)$ i.e. $\lambda(x_1, \ldots, x_n) \leq c$. By taking the log function, the critical region can be described as

$$\{ \mathbf{x} : x_{(1)} \geq \theta_0 - \frac{\log c}{n} \}.$$

We can choose $c = \alpha$ to have a valid test. To see how it works, note that

$$\mathbb{P}_\theta \left( X_{(1)} \geq \theta_0 - \frac{\log \alpha}{n} \right) \leq \mathbb{P}_{\theta_0} \left( X_{(1)} \geq \theta_0 - \frac{\log \alpha}{n} \right), \text{ for any } \theta \geq \theta_0.$$

Therefore

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left( X_{(1)} \geq \theta_0 - \frac{\log \alpha}{n} \right) = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta \left( X_{(1)} \geq \theta_0 - \frac{\log \alpha}{n} \right) = \mathbb{P}_{\theta_0} \left( X_{(1)} \geq \theta_0 - \frac{\log \alpha}{n} \right) = \prod_{i=1}^{n} \mathbb{P}_{\theta_0} \left( X_i \geq \theta_0 - \frac{\log \alpha}{n} \right) = \alpha,$$

where the last inequality uses $X_i - \theta_0 \sim \text{Exp}(1)$ and its survival function.

3. **[20 points]**

   (a) Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Construct the likelihood ratio test for

   \[ H_0 : \sigma = \sigma_0, \mu \text{ unknown} \]
   \[ H_1 : \sigma \neq \sigma_0, \mu \text{ unknown}. \]

   Compare to the Wald test.
Solution:

Likelihood Ratio Test. Let \( x = \{x_1, \ldots, x_n\} \) and \( \bar{x} = n^{-1} \sum_{i=1}^{n} x_i \). For the normal distribution, the log-likelihood function is

\[
\ell(\mu, \sigma) = \log L(\mu, \sigma|\mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.
\]

We first find the maximum likelihood estimators for \((\mu, \sigma)\) without restrictions. Notice that

\[
\frac{\partial}{\partial \mu} \ell(\mu, \sigma) = \frac{1}{\sigma^2} \left( \sum_{i=1}^{n} x_i - n\mu \right) = 0,
\]

\[
\frac{\partial}{\partial \sigma} \ell(\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - \mu)^2 = 0.
\]

Solving these two equations simultaneously, we obtain the solution as

\[
\hat{\mu}_{\text{mle}} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2_{\text{mle}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]

In addition, the Hessian matrix is

\[
H = \begin{pmatrix}
\frac{\partial^2 \ell(\mu, \sigma)}{\partial \mu^2} & \frac{\partial^2 \ell(\mu, \sigma)}{\partial \mu \partial \sigma} \\
\frac{\partial^2 \ell(\mu, \sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 \ell(\mu, \sigma)}{\partial \sigma^2}
\end{pmatrix}_{(\mu, \sigma) = (\hat{\mu}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}})} = \begin{pmatrix}
-\frac{n}{\hat{\sigma}^2_{\text{mle}}} & 0 \\
0 & -\frac{2n}{\hat{\sigma}^2_{\text{mle}}}
\end{pmatrix},
\]

which is negative definite. Hence \((\hat{\mu}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}})\) are the maximum likelihood estimators for \((\mu, \sigma)\). Under the null where \(\sigma = \sigma_0\), we see that \((\hat{\mu}_{\text{mle}}, \sigma_0)\) maximizes the likelihood function. Therefore, the LRT statistic is

\[
\lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta} L(\mu, \sigma|\mathbf{x})}{\sup_{\theta \in \Theta_0} L(\mu, \sigma|\mathbf{x})} = \frac{L(\hat{\mu}_{\text{mle}}, \sigma_0)}{L(\hat{\mu}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}})}
\]

where \(\theta = (\mu, \sigma), \Theta = \{\theta : -\infty < \mu < \infty, \sigma > 0\}\) and \(\Theta_0 = \{\theta : -\infty < \mu < \infty, \sigma = \sigma_0\}\). We reject the null for a small value of \(\lambda(\mathbf{x})\). Specifically we have

\[
L(\hat{\mu}_{\text{mle}}, \sigma_0) = (2\pi \sigma_0^2)^{-n/2} \exp \left( -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right),
\]

\[
L(\hat{\mu}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}}) = (2\pi)^{-n/2} \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-n/2} \exp \left( -\frac{n}{2} \right).
\]
Hence the null is rejected when
\[
\lambda(x) = \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n\sigma_0^2} \right)^{n/2} \exp \left( -\frac{n}{2} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n\sigma_0^2} + \frac{n}{2} \right)
\]
= \left( \frac{\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} \right)^{n/2} \exp \left( -\frac{n}{2} \frac{\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} \right) \exp \left( \frac{n}{2} \right) < k
\]
for some \( k < 1 \). Viewing \( \lambda(x) \) as a function of \( \hat{\sigma}_{\text{mle}}^2 / \sigma_0^2 \), the above inequality holds when
\[
\frac{\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} < a \quad \text{or} \quad \frac{\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} > b
\]
for some \( a < b \). Since under \( H_0 \)
\[
\frac{n\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma_0^2} \sim \chi^2_{n-1}, \quad (5)
\]
we can simply decide to reject the null when
\[
\frac{n\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2} \quad \text{or} \quad \frac{n\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2}. \quad (6)
\]
We can also use the asymptotic property of the likelihood ratio statistic and reject the null when
\[-2 \log \lambda(x) > \chi^2_{1,\alpha},\]
where we use the fact that \( \dim(\Theta) = 2 \) and \( \dim(\Theta_0) = 1 \).

Wald Test. Due to the relationship in (5), we have
\[
\operatorname{Var}_{H_0} \left( \frac{n\hat{\sigma}_{\text{mle}}^2}{\sigma_0^2} \right) = 2(n-1) \quad \text{and thus} \quad \operatorname{Var}_{H_0} \left( \hat{\sigma}_{\text{mle}}^2 \right) = \frac{2(n-1)}{n^2} \sigma_0^4.
\]
Then a Wald statistic can be written as
\[
T_{\text{Wald}} = \frac{\hat{\sigma}_{\text{mle}}^2 - \sigma_0^2}{\sqrt{\frac{2(n-1)}{n^2} \sigma_0^4}}
\]
Alternatively the scaling part can be replaced by the square root of the inverse Fisher information \( \{nI(\sigma_0^2)\}^{-1} = 2\sigma_0^4 / n \). By the asymptotic normality of the MLE, we have
\[
T_{\text{Wald}} \sim N(0, 1).
\]
Using this asymptotic result, we can reject the null when
\[
T_{\text{Wald}}^2 > \chi^2_{1,\alpha}.
\]
By a Taylor expansion as in the lecture note, it can be seen that \( T_{\text{Wald}}^2 \) is asymptotically equivalent to \(-2 \log \lambda(X_1, \ldots, X_n)\).
(b) Let $X \sim \text{Bin}(n,p)$. Construct the likelihood ratio test for

\[ H_0 : p = p_0 \]
\[ H_1 : p \neq p_0. \]

Compare to the Wald test.

**Solution:**

**Likelihood Ratio Test.** The log-likelihood function for Binomial distribution is

\[ \ell(p) = \log L(p|x) = \log \left( \binom{n}{x} \right) + x \log p + (n-x) \log(1-p). \]

As before, calculate the first and second derivatives as

\[ \frac{d}{dp} \ell(p) = \frac{x}{p} - \frac{n-x}{1-p} = 0 \iff p = \frac{x}{n}, \]

\[ \frac{d^2}{dp^2} \ell(p) = - \frac{x}{p^2} - \frac{n-x}{(1-p)^2} < 0, \quad \text{for } 0 < p < 1. \]

Hence $\hat{p}_{\text{mle}} = \frac{x}{n}$. Using this result, the LRT statistic is

\[ \lambda(x) = \frac{L(p_0|x)}{L(\hat{p}_{\text{mle}}|x)} = \frac{p_0^x(1-p_0)^{n-x}}{\hat{p}_{\text{mle}}^x(1-\hat{p}_{\text{mle}})^{n-x}}. \]

We reject the null when

\[ \frac{p_0^x(1-p_0)^{n-x}}{\hat{p}_{\text{mle}}^x(1-\hat{p}_{\text{mle}})^{n-x}} < k \]

for some $k < 1$. For a large sample size, we use the $\chi^2$ approximation

\[ -2 \log \lambda(X) = -2X \log \left( \frac{p_0}{X/n} \right) - 2(n-X) \log \left( \frac{1-p_0}{1-X/n} \right) \sim \chi^2_1 \]

and reject the null when

\[ -2 \log \lambda(x) > \chi^2_{1,\alpha}. \]

**Wald Test.** Under $H_0$, the variance of $\hat{p}_{\text{mle}}$ is

\[ \text{Var}_{H_0}(\hat{p}_{\text{mle}}) = \frac{1}{nI(p_0)} = \frac{p_0(1-p_0)}{n}. \]
where $I(p_0)$ is the Fisher information. Based on this, we have a Wald test statistic converging to the normal distribution

$$T_{\text{Wald}} = \frac{X/n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1).$$

Hence the Wald test rejects the null when

$$T_{\text{Wald}}^2 > \chi^2_{1,\alpha}.$$

As before, it is seen that $T_{\text{Wald}}^2$ is asymptotically equivalent to $-2 \log \lambda(X)$.

4. [20 points] p-values: Throughout this question we fix an arbitrary test statistic. Most natural hypothesis tests have (strictly) nested rejection regions. What this means is that if we denote the rejection region as $R_\alpha$ then:

$$R_\alpha \subset R_{\alpha'} \quad \text{if} \quad \alpha < \alpha'.$$

It is in this setting that p-values make the most sense. We will assume that our test statistic has nested rejection regions. Now, consider the general hypothesis testing problem:

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \notin \Theta_0.$$

(a) Show that if the test is valid, i.e.,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\{X_1, \ldots, X_n\} \in R_\alpha) \leq \alpha \quad \text{for all} \quad 0 \leq \alpha \leq 1,$$

then for any $\theta \in \Theta_0$ we have that the distribution of the p-value satisfies:

$$\mathbb{P}_\theta(p \leq u) \leq u \quad \text{for all} \quad 0 \leq u \leq 1.$$

This means that the p-value is “stochastically dominated” by the uniform distribution.

Solution:

The p-value is defined as the smallest $\alpha$ at which we would reject $H_0$. That is, the p-value is given by

$$p = \inf \{\alpha : \{X_1, \ldots, X_n\} \in R_\alpha\}.$$

Therefore if $\theta \in \Theta_0$, then $\{p \leq u\}$ implies that $\{X_1, \ldots, X_n\} \in R_\theta$ for any $u < v$. Hence

$$\mathbb{P}_\theta(p \leq u) \leq \mathbb{P}_\theta(\{X_1, \ldots, X_n\} \in R_\theta) \leq v.$$

Now by taking $v \to u$, we obtain the result.
(b) Show that if there is a \( \theta \in \Theta_0 \) such that,
\[
P_{\theta}(\{X_1, \ldots, X_n\} \in R_\alpha) = \alpha \quad \text{for all } 0 \leq \alpha \leq 1,
\]
then
\[
P_{\theta}(p \leq u) = u \quad \text{for all } 0 \leq u \leq 1.
\]
This shows that in cases when the null is simple (or if there is a "worst-case" null distribution) then the p-value has a uniform distribution.

Solution:

Since \( \{X_1, \ldots, X_n\} \in R_u \) implies that \( \{p \leq u\} \) by the definition of the p-value, we have
\[
P_{\theta}(p \leq u) \geq P_{\theta}(\{X_1, \ldots, X_n\} \in R_u) = u.
\]

Combining this with the previous problem, we see that
\[
P_{\theta}(p \leq u) = u \quad \text{for all } 0 \leq u \leq 1.
\]

5. [10 points] More LRT: Suppose that, \( X_1, \ldots, X_n \) is drawn from a Pareto distribution with density:
\[
p(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}}I_{[\nu, \infty)}(x), \quad \theta > 0, \nu > 0.
\]
Construct the (generalized) LRT statistic for distinguishing the hypotheses:
\[
H_0 : \theta = 1, \nu \text{ unknown},
H_1 : \theta \neq 1, \nu \text{ unknown}.
\]

Suppose Wilk’s approximation was accurate. Derive the critical region for the GLRT.

Solution:

Let \( x = \{x_1, \ldots, x_n\} \) and \( x_{(1)} = \min_i x_i \). The log-likelihood function is
\[
\log L(\theta, \nu|x) = \left[ n \log \theta + n \theta \log \nu - (\theta + 1) \log \left( \prod_{i=1}^n x_i \right) \right] \times I\{\nu \leq x_{(1)}\}.
\]
For any value of \( \theta \), the likelihood function is increasing in \( \nu \) on \( \nu \leq x_{(1)} \). Therefore, we have \( \hat{\nu}_{\text{MLE}} = x_{(1)} \) for both the restricted and unrestricted cases. Next we find the MLE of \( \theta \) by taking the derivative of the log likelihood function and setting equal to zero
\[
\frac{\partial}{\partial \theta} \log L(\theta, x_{(1)}|x) = \frac{n}{\theta} + n \log x_{(1)} - \log \left( \prod_{i=1}^n x_i \right) = 0,
\]

which gives

\[ \hat{\theta}_{\text{mle}} = \frac{n}{\log \left( \prod_{i=1}^{n} x_i / x_{(1)} \right)} . \]

Also check that \( \partial^2 / \partial \theta^2 \log L(\theta, x|\mathbf{x}) = -n/\theta^2 < 0 \), which ensures that \( \hat{\theta}_{\text{mle}} \) is a maximum.

Notice that we have \( \hat{\theta}_0 = 1 \) and \( \hat{\nu}_{\text{mle}} = x_{(1)} \) under the null. Therefore the LRT statistic is

\[ \lambda(x) = \frac{\prod_{i=1}^{n} \left( x_{(1)}/x_i^2 \right)}{\prod_{i=1}^{n} \left( \hat{\theta}_{\text{mle}}^{\hat{\theta}} / x_i^{\hat{\theta}+1} \right)} . \]

Suppose Wilk’s approximation was accurate. Then we have

\[-2 \log \lambda(X_1, \ldots, X_n) \sim \chi^2_{\omega}, \]

where \( \omega = \dim(\Theta) - \dim(\Theta_0) = 2 - 1 = 1 \). Hence the critical region for the GLRT is

\[ R_\alpha = \{ x : -2 \log \lambda(x) > \chi^2_{1, \alpha} \} . \]

6. **[20 points]** Minimax Rate Lower Bounds: We will derive something that is popularly known as Le Cam’s (two-point) lower bound. Minimax rate lower bounds for point estimation are usually proved via a reduction to testing. Suppose we have a collection of distributions \( P_\theta = \{ p_\theta : \theta \in \Theta \} \), and we are interested in proving a lower bound on the minimax rate for estimating \( \theta \). Concretely, for some loss function \( \ell \), which we will assume is symmetric and satisfies the triangle inequality, we want to show something of the form:

\[ \inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \ldots . \]

(a) Fix an estimator \( \hat{\theta}(X_1, \ldots, X_n) \). First show that for any \( \delta > 0 \),

\[ \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \delta \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta) . \]

**Solution:**

We assume \( \ell(\hat{\theta}, \theta) \) is nonnegative. Hence by Markov’s inequality,

\[ \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta) \leq \delta^{-1} \mathbb{E}[\ell(\hat{\theta}, \theta)], \]

which implies the result.
Now, select $\theta_1, \theta_2 \in \Theta$, such that $\ell(\theta_1, \theta_2) > 2\delta$. From our estimator $\hat{\theta}$ we are going to create a test $\phi(\hat{\theta})$ for distinguishing:

$$H_0 : \theta = \theta_1$$
$$H_1 : \theta = \theta_2.$$ 

In particular,

$$\phi(\hat{\theta}) = \begin{cases} \theta_1 & \text{if } \ell(\hat{\theta}, \theta_1) < \ell(\hat{\theta}, \theta_2), \\ \theta_2 & \text{otherwise}. \end{cases}$$ 

Show that,

$$\sup_{\theta \in \Theta} \delta \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta) \geq \frac{\delta}{2} \left[ \mathbb{P}_{\theta_1}(\phi(\hat{\theta}) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(\hat{\theta}) \neq \theta_2) \right].$$

**Solution:**

Suppose that $\phi(\hat{\theta}) \neq \theta_1$. Then

$$2\delta \overset{(i)}{<} \ell(\theta_1, \theta_2) \overset{(ii)}{\leq} \ell(\theta_1, \hat{\theta}) + \ell(\theta_2, \hat{\theta}) \overset{(iii)}{\leq} 2\ell(\theta_1, \hat{\theta}),$$

where $(i)$ is from the construction of $\theta_1$ and $\theta_2$, $(ii)$ uses the triangle inequality and $(iii)$ uses the fact that $\phi(\hat{\theta}) \neq \theta_1$ implies $\ell(\hat{\theta}, \theta_1) \geq \ell(\hat{\theta}, \theta_2)$ by the definition of $\phi(\hat{\theta})$. Therefore

$$\mathbb{P}_{\theta_1}(\ell(\hat{\theta}, \theta_1) \geq \delta) \geq \mathbb{P}_{\theta_1}(\phi(\hat{\theta}) \neq \theta_1).$$

Similarly we have

$$\mathbb{P}_{\theta_2}(\ell(\hat{\theta}, \theta_2) \geq \delta) \geq \mathbb{P}_{\theta_2}(\phi(\hat{\theta}) \neq \theta_2).$$

From these,

$$\sup_{\theta \in \Theta} \delta \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta) \geq \max_{j=1,2} \delta \mathbb{P}_{\theta_j}(\ell(\hat{\theta}, \theta_j) \geq \delta) \geq \max_{j=1,2} \delta \mathbb{P}_{\theta_j}(\phi(\hat{\theta}) \neq \theta_j) \geq \frac{\delta}{2} \left[ \mathbb{P}_{\theta_1}(\phi(\hat{\theta}) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(\hat{\theta}) \neq \theta_2) \right],$$

where the last inequality holds because a maximum is larger than an average.
(c) Now, show that:

\[
\inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \frac{\delta}{2} \inf_{\phi} \left[ \mathbb{P}_{\theta_1}(\phi(X_1, \ldots, X_n) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(X_1, \ldots, X_n) \neq \theta_2) \right],
\]

where \( \phi \) is a test of the form above for distinguishing the hypotheses \( H_0 \) and \( H_1 \).

**Solution:**

Combining the results in (a) and (b),

\[
\sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \frac{\delta}{2} \left[ \mathbb{P}_{\theta_1}(\phi(\hat{\theta}) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(\hat{\theta}) \neq \theta_2) \right]
\]

\[
\geq \frac{\delta}{2} \inf_{\phi} \left[ \mathbb{P}_{\theta_1}(\phi(X_1, \ldots, X_n) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(X_1, \ldots, X_n) \neq \theta_2) \right].
\]

Now the result follows by taking the infimum over any estimators on both sides.

(d) Consider the (optimal) NP classifier for distinguishing these hypotheses. Show that this classifier has error \( 1 - \text{TV}(p_{n}^{\theta_1}, p_{n}^{\theta_2}) \), where \( p_{n}^{\theta} \) denotes the \( n \)-fold product distribution (i.e. the distribution of \( n \) i.i.d. samples). Using this show the following minimax lower bound, for any \( \theta_1, \theta_2 \) such that \( \ell(\theta_1, \theta_2) > 2\delta \),

\[
\inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \frac{\delta}{2} \left[ 1 - \text{TV}(p_{n}^{\theta_1}, p_{n}^{\theta_2}) \right].
\]

**Solution:**

Let \( x = \{x_1, \ldots, x_n\} \) and \( q_{\theta} \) be the density of \( p_{\theta} \) with respect to a base measure \( \mu \). Consider the Neyman-Pearson (NP) test

\[
\phi^*(x) = \begin{cases} 
\theta_1, & \text{if } x \in A = \{x : q_{\theta_1}^{n}(x) \geq q_{\theta_2}^{n}(x)\}, \\
\theta_2, & \text{if } x \in A^c = \{x : q_{\theta_1}^{n}(x) < q_{\theta_2}^{n}(x)\}.
\end{cases}
\]

By the Neyman-Pearson lemma and Problem 1, we know

\[
\inf_{\phi} [\mathbb{P}_{\theta_1}(\phi(X) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(X) \neq \theta_2)] = \mathbb{P}_{\theta_1}(\phi^*(X) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi^*(X) \neq \theta_2)
\]

\[
= \int_{A} q_{\theta_2}^{n}(x)d\mu(x) + \int_{A^c} q_{\theta_1}^{n}(x)d\mu(x)
\]

\[
= 1 - \left( \int_{A^c} q_{\theta_2}^{n}(x)d\mu(x) - \int_{A^c} q_{\theta_1}^{n}(x)d\mu(x) \right)
\]

\[
= 1 - \text{TV}(p_{\theta_1}, p_{\theta_2}).
\]

Now the result follows from Part (c).
(e) Explain the bound intuitively, i.e. how should we try to select \( \theta_1, \theta_2 \) to obtain a strong lower bound.

Solution:
This lower bound tells us that we need to have a large \( \delta \) and a small \( \text{TV}(p^n_{\theta_1}, p^n_{\theta_2}) \) to have a tight lower bound. However, there is a trade-off in this procedure. For example, if we choose \( \theta_1 \) and \( \theta_2 \) far from each other, then we can choose \( \delta \) to be large but \( \text{TV}(p^n_{\theta_1}, p^n_{\theta_2}) \) becomes also large. Hence a strong lower bound can be obtained by choosing \( \theta_1 \) and \( \theta_2 \) strategically that balance \( \delta \) and \( \text{TV}(p^n_{\theta_1}, p^n_{\theta_2}) \).

7. [20 points] Extra Credit: The \( \chi^2 \) test: In lecture we claimed that the Pearson goodness-of-fit statistic for testing a multinomial \( p_0 \),

\[
T = \sum_{i=1}^{d} \frac{(X_i - np_{0i})^2}{np_{0i}},
\]

has a \( \chi^2 \) distribution with \( d - 1 \) degrees of freedom (we actually subtracted the mean of this statistic to center the test statistic to have mean 0 under the null but this is not important). We will assume throughout that \( d \) is fixed, and that \( p_{0i} > 0 \) for each category. Here \( X_1, \ldots, X_d \) are the counts for the \( d \) categories, so we have that:

\[
\sum_{i=1}^{d} X_i = n \quad \text{and} \quad \sum_{i=1}^{d} p_{0i} = 1.
\]

(a) As a warmup: Suppose that there are only two categories, i.e. the multinomial is \((p_{01}, p_{02})\). Show that the test statistic can be equivalently written as:

\[
T = \frac{(X_1 - np_{01})^2}{np_{01}(1 - p_{01})}.
\]

Show that \( \frac{(X_1 - np_{01})}{\sqrt{np_{01}(1 - p_{01})}} \) under the null has an asymptotically standard normal distribution, and conclude that in this case the Pearson statistic has a \( \chi^2_1 \) distribution.

Solution:
If there are just two categories, the test statistic can be written as:

\[
T = \frac{(X_1 - np_{01})^2}{np_{01}} + \frac{(X_2 - np_{02})^2}{np_{02}} = \frac{(X_1 - np_{01})^2}{np_{01}} + \frac{(n - X_1 - n + np_{01})^2}{n(1 - p_{01})}
\]
\[= \frac{(X_1 - np_{01})^2}{np_{01}} + \frac{(X_1 - np_{01})^2}{n(1 - p_{01})} = \frac{(X_1 - np_{01})^2}{n} \left[ \frac{1}{p_{01}} + \frac{1}{1 - p_{01}} \right] = \frac{(X_1 - np_{01})^2}{np_{01}(1 - p_{01})}.\]

Now by the central limit theorem under the null hypothesis,
\[\sqrt{n} \left( \frac{X_1}{n} - p_{01} \right) \sqrt{p_{01}(1 - p_{01})} \xrightarrow{d} N(0, 1).\]

Let \(Z = \frac{\sqrt{n}(X_1/n - p_{01})}{\sqrt{p_{01}(1 - p_{01})}}\), then \(Z \xrightarrow{d} N(0, 1)\) and
\[Z = \frac{\sqrt{n}(X_1/n - p_{01})}{\sqrt{p_{01}(1 - p_{01})}} = \frac{X_1 - np_{01}}{\sqrt{np_{01}(1 - p_{01})}}.\]

Therefore, we notice that \(T = Z^2 \xrightarrow{d} \chi_1^2\).

(b) The general case roughly follows the same ideas but is more involved. I will outline the steps to help you but this is only one way to get to the answer (and ideally you should try to solve it without looking through these steps first).

- The first thing to note is that our representation is redundant in the sense that the sum of the counts and the sum of the probabilities is constrained, so we will eliminate \(p_0d\), and denote the vector \(\tilde{p} = (p_{01}, \ldots, p_{0(d-1)})\). Now, using the usual MLE asymptotics (you can look this up in the Wasserman book), you can conclude that under the null,
\[
\sqrt{n} \begin{pmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{pmatrix} - \tilde{p} \xrightarrow{d} N(0, \Sigma),
\]
where
\[
\Sigma = \begin{pmatrix} p_{01}(1 - p_{01}) & -p_{01}p_{02} & \cdots & -p_{01}p_{0(d-1)} \\ -p_{01}p_{02} & p_{02}(1 - p_{02}) & \cdots & -p_{02}p_{0(d-1)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{01}p_{0(d-1)} & -p_{02}p_{0(d-1)} & \cdots & p_{0(d-1)}(1 - p_{0(d-1)}) \end{pmatrix}.
\]

Convince yourself that the covariance matrix is non-degenerate (this is why we eliminated one of the categories).
• Show that the inverse of the covariance matrix is given by:

\[ I(p) = \begin{bmatrix}
\frac{1}{p_{01}} + \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\
\frac{1}{p_{0d}} & \frac{1}{p_{02}} + \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\
\vdots & \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\
\frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0(d-1)}} + \frac{1}{p_{0d}}
\end{bmatrix}.\]

Conclude that,

\[ n \left( \begin{bmatrix}
\frac{X_1}{n} \\
\frac{X_2}{n} \\
\vdots \\
\frac{X_{d-1}}{n}
\end{bmatrix} - \tilde{p} \right)^T I(p) \left( \begin{bmatrix}
\frac{X_1}{n} \\
\frac{X_2}{n} \\
\vdots \\
\frac{X_{d-1}}{n}
\end{bmatrix} - \tilde{p} \right) \xrightarrow{d} \chi^2_{d-1}.\]

• Show that the above statistic is equal to the Pearson $\chi^2$ statistic.

Solution:

Since $\sum_{i=1}^d p_{0i} = 1$, we eliminate $p_{0d}$ and write it in terms of the other probabilities. Hence we consider $\tilde{p} = (p_{01}, \ldots, p_{0(d-1)})$. We know by using MLE asymptotics that, under the null hypothesis,

\[ \sqrt{n} \left( \begin{bmatrix}
\frac{X_1}{n} \\
\frac{X_2}{n} \\
\vdots \\
\frac{X_{d-1}}{n}
\end{bmatrix} - \tilde{p} \right) \xrightarrow{d} N(0, \Sigma). \]

where

\[ \Sigma = \begin{bmatrix}
p_{01}(1 - p_{01}) & -p_{01}p_{02} & \cdots & -p_{01}p_{0(d-1)} \\
-p_{01}p_{02} & p_{02}(1 - p_{02}) & \cdots & -p_{02}p_{0(d-1)} \\
\vdots & \vdots & \ddots & \vdots \\
-p_{01}p_{0(d-1)} & -p_{02}p_{0(d-1)} & \cdots & p_{0(d-1)}(1 - p_{0(d-1)})
\end{bmatrix}. \]

Now we can show that the inverse of the covariance matrix, let us call it $\Sigma$ is given by

\[ I(p) = \Sigma^{-1} = \begin{bmatrix}
\frac{1}{p_{01}} + \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\
\frac{1}{p_{0d}} & \frac{1}{p_{02}} + \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\
\vdots & \frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0(d-1)}} + \frac{1}{p_{0d}}
\end{bmatrix}. \]

In order to show this, it is enough to verify that

\[ \Sigma \Sigma^{-1} = I_{d-1}. \]
For $i \neq j$

$$(\Sigma \Sigma^{-1})_{ij} = - \sum_{k \neq i,j} \frac{p_0i p_0k}{p_0d} + \frac{p_0i(1 - p_0i)}{p_0d} - p_0i p_0j \left( \frac{1}{p_0j} + \frac{1}{p_0d} \right)$$

$$= - \sum_{k=1}^{d-1} \frac{p_0i p_0k}{p_0d} + \frac{p_0i}{p_0d} - p_0i$$

$$= \frac{p_0i(1 - p_0d)}{p_0d} + \frac{p_0i}{p_0d} - p_0i = 0,$$

and for the diagonal elements

$$(\Sigma \Sigma^{-1})_{ii} = - \sum_{k \neq i} \frac{p_0i p_0k}{p_0d} + p_0i(1 - p_0i) \left( \frac{1}{p_0i} + \frac{1}{p_0d} \right)$$

$$= - \sum_{k=1}^{d-1} \frac{p_0i p_0k}{p_0d} + 1 - p_0i + \frac{p_0i}{p_0d}$$

$$= - \frac{p_0i(1 - p_0d)}{p_0d} + 1 - p_0i + \frac{p_0i}{p_0d} = 1.$$

Therefore,

$$\Sigma \Sigma^{-1} = I_{d-1}.$$

and we can conclude that

$$n \begin{pmatrix} X_1/n & X_2/n & \hdots & X_{d-1}/n \end{pmatrix}^T I(p) \begin{pmatrix} X_1/n & X_2/n & \hdots & X_{d-1}/n \end{pmatrix} \cdot (\Sigma^{-1})_{ij} \xrightarrow{d} \chi^2_{d-1}.$$

Next we show that the term on the left-hand side is equal to the Pearson $\chi^2$ statistic. After calculations, we see that

$$\text{LHS} = n \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} \left( \frac{X_i - np_0i}{n} \right) \left( \frac{X_j - np_0j}{n} \right) (\Sigma^{-1})_{ij}$$

$$= n \sum_{i=1}^{d-1} \sum_{j \neq i} \left( \frac{X_i - np_0i}{n} \right) \left( \frac{X_j - np_0j}{n} \right) \frac{1}{p_0d}$$

$$+ n \sum_{i=1}^{d-1} \left( \frac{X_i - np_0i}{n} \right) \left( \frac{X_j - np_0i}{n} \right) \left( \frac{1}{p_0i} + \frac{1}{p_0d} \right)$$

$$= \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} \frac{(X_i - np_0i)(X_j - np_0j)}{np_0d} + \sum_{i=1}^{d-1} \frac{(X_i - np_0i)^2}{np_0i}.$$
Furthermore

\[(X_d - np_{0d})^2 = \left(n - \sum_{i=1}^{d-1} X_i - n + n \sum_{i=1}^{d-1} p_{0i}\right)^2 \]

\[= \left(\sum_{i=1}^{d-1} (X_i - np_{0i})\right)^2 \]

\[= \sum_{i=1}^{d-1} \sum_{j=1}^{d-1} (X_i - np_{0i})(X_j - np_{0j}) . \]

By plugging this into the previous expression, we have

\[
\sum_{i=1}^{d} \frac{(X_i - np_{0i})^2}{np_{0i}} = n \left(\begin{bmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{bmatrix} - \tilde{p}\right)^T I(p) \left(\begin{bmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{bmatrix} - \tilde{p}\right),
\]

as desired.

**Alternate proof for finding** $I(p)$: Say if we wanted to derive the inverse matrix $(\Sigma^{-1})$, then we first notice that $\Sigma$ can be written as:

\[
\Sigma = A - vv^T, \quad A = \begin{bmatrix} p_{01} & 0 & \ldots & 0 \\ 0 & p_{02} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_{0(d-1)} \end{bmatrix}, \quad v = \begin{bmatrix} p_{01} \\ p_{02} \\ \vdots \\ p_{0(d-1)} \end{bmatrix}.
\]

Then by using the Sherman–Morrison formula, we get

\[
\Sigma^{-1} = A^{-1} + A^{-1} vv^T A^{-1} \frac{1}{1 - v^T A^{-1} v}, \quad A^{-1} = \begin{bmatrix} \frac{1}{p_{01}} & 0 & \ldots & 0 \\ 0 & \frac{1}{p_{02}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{p_{0(d-1)}} \end{bmatrix}.
\]

Now,

\[
A^{-1} v = \begin{bmatrix} \frac{1}{p_{01}} & 0 & \ldots & 0 \\ 0 & \frac{1}{p_{02}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{p_{0(d-1)}} \end{bmatrix} \begin{bmatrix} p_{01} \\ p_{02} \\ \vdots \\ p_{0(d-1)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = 1_{d-1}.
\]
Therefore,

\[ A^{-1} v v^T A^{-1} = 1_{d-1} 1_{d-1}^T, \]

and

\[ v^T A^{-1} v = v^T 1_{d-1} = \sum_{i=1}^{d-1} p_{0i} = 1 - p_{0d}. \]

Plugging everything in we get,

\[
I(p) = \Sigma^{-1} = \begin{bmatrix}
\frac{1}{p_{01}} & 0 & \ldots & 0 \\
0 & \frac{1}{p_{02}} & \ldots & 0 \\
\vdots & & & \\
0 & 0 & \ldots & \frac{1}{p_{0(d-1)}}
\end{bmatrix} + \frac{1}{p_{0d}} 1_{d-1} 1_{d-1}^T
\]

\[
= \begin{bmatrix}
\frac{1}{p_{01}} + \frac{1}{p_{0d}} & \frac{1}{p_{01}} + \frac{1}{p_{0d}} & \ldots & \frac{1}{p_{01}} + \frac{1}{p_{0d}} \\
\frac{1}{p_{01}} + \frac{1}{p_{0d}} & \frac{1}{p_{02}} + \frac{1}{p_{0d}} & \ldots & \frac{1}{p_{02}} + \frac{1}{p_{0d}} \\
\vdots & & & \\
\frac{1}{p_{0d}} & \frac{1}{p_{0d}} & \ldots & \frac{1}{p_{0(d-1)}} + \frac{1}{p_{0d}}
\end{bmatrix}.
\]