6.S951 Modern Mathematical Statistics Fall 2024

Lecture 7 — September 26, 2024

Prof. Stephen Bates

Scribe: Otavio Tecchio

1 Outline

So far:

1. Statistical decision theory (notions of optimality).
2. Estimation, prediction, and testing (with examples).

Agenda:

1. Total Variation Distance.
2. Composite hypothesis testing.
3. First steps on minimax bounds and information theory.

2 Total Variation (TV) Distance

Consider two probability distributions P_0 and P_1 over a common sample space $\mathcal{X}$. The definition of TV distance is as follows.

Definition 1. *The TV distance between P_0 and P_1, denoted by $\|P_0 - P_1\|_{TV}$, is given by*

$$\|P_0 - P_1\|_{TV} = \sup_{B \subseteq \mathcal{X}} |P_0(B) - P_1(B)|.$$

The following lemma, stated and proved in Lecture 6, relates the TV distance to the likelihood ratio.

Lemma 1. *Assume that P_0 and P_1 admit density functions denoted by f_0 and f_1 respectively. The set $B^{opt} \subseteq \mathcal{X}$ that maximizes $|P_0(B) - P_1(B)|$ is given by*

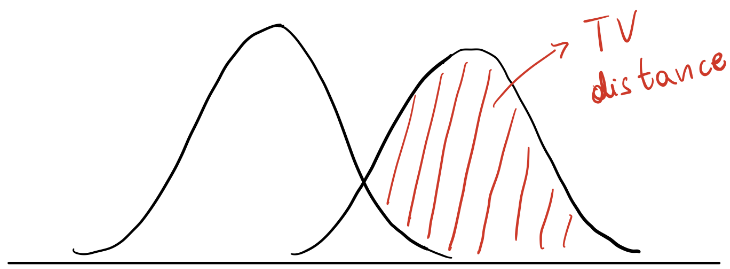
$$B^{opt} = \{x : f_1(x) > f_0(x)\}.$$

Note: equivalently, we could consider the complement of B^{opt}.

From the above lemma, we get the following corollary.

Corollary 1. *Assume that P_0 and P_1 admit density functions denoted by f_0 and f_1 respectively, then*

$$\|P_0 - P_1\|_{TV} = \int_{B^{opt}} f_1(x) - f_0(x)dx.$$



The following theorem motivates/justifies the introduction of TV distance in the context of simple-simple hypothesis testing. It shows that the TV distance encodes the difficulty of simple-simple hypothesis testing.

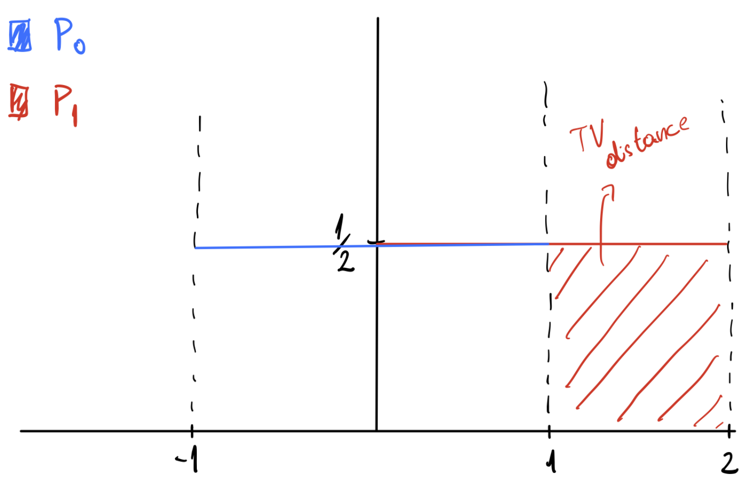
Theorem 1 (TV distance to testing link)**.** *Consider a simple-simple hypothesis test with $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, then*

$$\inf_{A:\mathcal{X}\to\{0,1\}} P_{\theta_0}(A(X) = 1) + P_{\theta_1}(A(X) = 0) = 1 - \|P_{\theta_0} - P_{\theta_1}\|_{TV}.$$

A proof of this theorem for the case where both P_{θ_0} and P_{θ_1} admit density functions is given below.

Note: see Theorem 15.1.1 in Lehman and Romano (4th edition) for a version of this result that allows for randomized tests.

Example 1 (Uniform Location Models)**.** *Consider $P_0 = uniform[-1, 1]$ and $P_1 = uniform[0, 2]$, illustrated in the figure below.*



Let A be a decision rule given by

$$A(x) = \begin{cases} 1 & \text{if } x > 1 \\ 0 & \text{otherwise} \end{cases}.$$

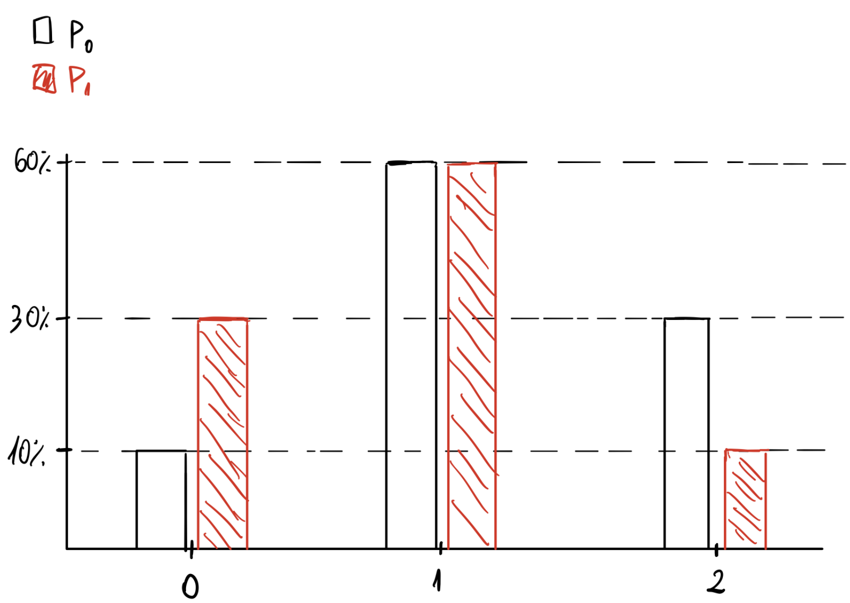
Note that $\|P_0 - P_1\|_{TV} = P_1([1,2]) - P_0([1,2]) = 1/2$ *since* $B^{opt} = [1,2]$ *in this case. Moreover,*

$$\begin{aligned} P_0\left(A(X)=1\right) + P_1\left(A(X)=0\right) =& P_0\left(X > 1\right) + P_1\left(X \le 1\right) \\ =& 0 + \frac{1}{2} = 1 - \|P_0 - P_1\|_{TV}. \end{aligned}$$

Thus, A hits the lower bound and is optimal in the sense that minimizes the sum of the probabilities of Type I and Type II errors. Also, A is Bayes optimal for a specific prior and loss function (see the proof of Theorem 1).

Note: *we could have chosen the cutoff value to be any number between 0 and 1 and the associated decision rule would still be optimal in the sense of minimizing the sum of the probabilities of Type I and Type II errors.*

Example 2 (Discrete Distribution)**.** *Consider the following discrete distribution.*



Note that $B^{opt} = \{0\}$ *so* $\|P_0 - P_1\|_{TV} = P_1(\{0\}) - P_0(\{0\}) = 20\%$. *Consider a decision rule A given by*

$$A(x) = \begin{cases} 1 & \text{if } x = 0 \\ 1 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}.$$

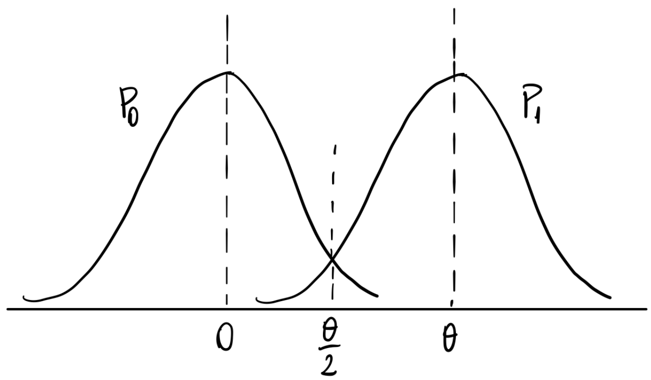
In this case, $P_0(A(X)=1) = P_0(X = 0 \text{ or } X = 1) = 70\%$ *and* $P_1(A(X)=0) = P_1(X=2) = 10\%$. *Therefore,*

$$P_0(A(X)=1) + P_1(A(X)=0) = 80\% = 1 - \|P_0 - P_1\|_{TV}.$$

Once again, the decision rule A hits the bound.

Note: *similar to the previous example, it does not matter what decision we make when $x = 1$ in the sense that if $A(1) = 0$ (or even if we randomized at $x = 1$), the sum of Type I and Type II errors probabilities would be the same.*

Example 3 (Gaussian Distributions)**.** *Let $P_0 = \mathcal{N}(0, 1)$ and $P_1 = \mathcal{N}(\theta, 1)$ for some given $\theta > 0$.*



In this case, $B^{\text{opt}} = (\theta/2, \infty)$ and

$$\begin{aligned}
\|P_0 - P_1\|_{TV} &= P_1\left(\left(\frac{\theta}{2}, \infty\right)\right) - P_0\left(\left(\frac{\theta}{2}, \infty\right)\right) \\
&= 1 - \Phi\left(\theta - \frac{\theta}{2}\right) - \left(1 - \Phi\left(\frac{\theta}{2}\right)\right) \\
&= \Phi\left(\frac{\theta}{2}\right) - \Phi\left(-\frac{\theta}{2}\right),
\end{aligned}$$

where Φ is the cdf of the standard normal.

A few numerical examples:

- *If $\theta = 5$, $\|P_0 - P_1\|_{TV} \approx 0.99$.*
- *If $\theta = 3$, $\|P_0 - P_1\|_{TV} \approx 0.87$.*
- *If $\theta = 1$, $\|P_0 - P_1\|_{TV} \approx 0.38$.*
- *If $\theta = 0.1$, $\|P_0 - P_1\|_{TV} \approx 0.04$.*

Proof of Theorem 1: Let f_0, f_1 be the density functions of P_{θ_0} and P_{θ_1} respectively.

Consider the notation introduced in Lecture 6 and set $\pi_0 = 0.5$ and $c_{FP} = c_{FN} = 1$. In Lecture 6, we showed that the optimal Bayes test A^{Bayes} in this setting is characterized by

$$A^{\text{Bayes}}(x) = \begin{cases} 1 & \text{if } f_1(x) > f_0(x) \\ 0 & \text{otherwise} \end{cases}.$$

Moreover, recall that for any $A : \mathcal{X} \to \{0,1\}$,

$$R_B(A, \pi_0 = 0.5) = \frac{1}{2}\left[P_{\theta_0}\left(A(X) = 1\right) + P_{\theta_1}\left(A(X) = 0\right)\right].$$

In particular,

$$\begin{aligned} R_B(A^{\text{Bayes}}, \pi_0 = 0.5) &= \frac{1}{2}\Big[P_{\theta_0}\left(f_1(X) > f_0(X)\right) + P_{\theta_1}\left(f_1(X) \le f_0(X)\right)\Big] \\ &= \frac{1}{2}\Big[P_{\theta_0}\left(f_1(X) > f_0(X)\right) + 1 - P_{\theta_1}\left(f_1(X) > f_0(X)\right)\Big] \\ &= \frac{1}{2}\Big[1 - \|P_0 - P_1\|_{TV}\Big] \end{aligned}$$

since $B^{\text{opt}} = \{x : f_1(x) > f_0(x)\}$ from Lemma 1. The result follows from the Bayes optimality of A^{Bayes}.

□

The next lemma gives a coupling/optimal transport interpretation for the TV distance.

Lemma 2. *If there exists $\gamma \in [0,1]$ such that $\|P_0 - P_1\|_{TV} = \gamma$, then there exists a joint distribution for $(X_0, X_1) \in \mathcal{X} \times \mathcal{X}$ with $P(X_0 = X_1) = 1 - \gamma$ and $X_0 \sim P_0$ and $X_1 \sim P_1$ marginally.*

Note: we also have that $1 - \gamma$ is the maximum possible value for $P(X_0 = X_1)$ for X_0 and X_1 satisfying the conditions in the lemma.

The idea is that we can flip a coin that returns Heads with probability $1 - \gamma$ and returns Tails with probability γ. The distributions P_0 and P_1 are indistinguishable when the coin returns Heads, and the distributions P_0 and P_1 are different when the coin returns Tails.

Example 4 (Example 2 continued)**.** *We had that $\|P_0 - P_1\|_{TV} = 20\%$ and so we want to construct a joint distribution such that $P(X_0 = X_1) = 80\%$. Consider the following joint distribution:*

$$P(X_0 = x_0, X_1 = x_1) = \begin{cases} 0.6 & \text{if } (x_0, x_1) = (1,1) \\ 0.1 & \text{if } (x_0, x_1) = (0,0) \\ 0.1 & \text{if } (x_0, x_1) = (2,2) \\ 0.2 & \text{if } (x_0, x_1) = (2,0) \\ 0 & \text{if } (x_0, x_1) = (0,2) \end{cases}.$$

Clearly, $P(X_0 = X_1) = 80\%$. Also, $P(X_0 = 0) = 0.1$, $P(X_0 = 1) = 0.6$, $P(X_0 = 2) = 0.3$ so that $X_0 \sim P_0$ and $P(X_1 = 0) = 0.3$, $P(X_1 = 1) = 0.6$, $P(X_1 = 2) = 0.1$ so that $X_1 \sim P_1$.

3 Composite Hypothesis Tests

Up to now, we have considered simple-simple hypothesis tests and derived that the optimal test is based on a threshold for the likelihood ratio. Now, we consider hypotheses where Θ_0 and Θ_1 are not singletons and tests $A : \mathcal{X} \to [0,1]$. Note that we are now allowing for randomized tests, which can be interpreted as a non-randomized test $A' : \mathcal{X} \to \{0,1\}$ defined by $A'(X) = \mathbf{1}_{\{U \le A(X)\}}$, where U is distributed as a uniform between 0 and 1 and is independent of X. Thus, $A(x)$ is the probability of rejecting the null given that we observed $x \in \mathcal{X}$.

3.1 Uniformly Most Powerful Tests

Definition 2 (UMP)**.** *A test* $A_\alpha^{UMP} : \mathcal{X} \to [0,1]$ *is the uniformly most powerful test of size* α *if for all* $\theta_1 \in \Theta_1$*,*

$$E_{\theta_1}[A_\alpha^{UMP}(X)] = \sup_{A \in \mathcal{A}} E_{\theta_1}[A(X)],$$

where $\mathcal{A} = \{A : \mathcal{X} \to [0,1]$ *such that* $\sup_{\theta_0 \in \Theta_0} E_{\theta_0}[A(X)] \leq \alpha\}$*.*

Example 5. *Consider a Gaussian distribution with mean* $\theta \in \mathbb{R}$ *and unit variance and the hypothesis* $\Theta_0 = (-\infty, 0]$*,* $\Theta_1 = (0, \infty)$*.*

Claim 1. *The uniformly most powerful test of size* α *in this case is given by*

$$A_\alpha^{UMP}(x) = \begin{cases} 1 & \text{if } x > \Phi^{-1}(1-\alpha) \\ 0 & \text{otherwise} \end{cases}.$$

Proof. Consider a simple-simple hypothesis with $\Theta_0 = \{0\}$ and $\Theta_1 = \{\theta_1\}$, $\theta_1 > 0$. From the Neyman-Pearson Lemma, the optimal test for this simple-simple hypothesis is

$$A^{\text{NP}}(x) = \begin{cases} 1 & \text{if } \frac{f_1(x)}{f_0(x)} > \lambda \\ 0 & \text{otherwise} \end{cases},$$

where f_0 is the pdf of the standard normal, f_1 is the pdf of $\mathcal{N}(\theta_1, 1)$ and λ satisfies

$$P_0\left(\frac{f_1(X)}{f_0(X)} > \lambda\right) = \alpha.$$

Here,

$$\frac{f_1(x)}{f_0(x)} = \exp\left(\theta_1 x - \frac{\theta_1^2}{2}\right),$$

which is strictly increasing in x, so an above-the-cutoff rule based on the likelihood ratio is equivalent to an above-the-cutoff rule based on x. Moreover, from the equation that implicitly defines λ, the Neyman-Pearson test binds the constraint. Thus, we can rewrite it as

$$A^{\text{NP}}(x) = \begin{cases} 1 & \text{if } x > \Phi^{-1}(1-\alpha) \\ 0 & \text{otherwise} \end{cases},$$

which equals A_α^{UMP}.

Thus, because A_α^{UMP} satisfies the Type I Error constraint for $\theta_0 = 0$ and because

$$\begin{aligned} E_{\theta_0}[A_\alpha^{\text{UMP}}(X)] = P_{\theta_0}(X > \Phi^{-1}(1-\alpha)) &= 1 - \Phi(\Phi^{-1}(1-\alpha) - \theta_0) \\ &\leq 1 - \Phi(\Phi^{-1}(1-\alpha)) \\ &= P_0(X > \Phi^{-1}(1-\alpha)) = E_0[A_\alpha^{\text{UMP}}(X)] \end{aligned}$$

for any $\theta_0 \in (-\infty, 0]$, A_α^{UMP} satisfies the Type I Error constraint for all $\theta_0 \in (-\infty, 0]$. Therefore, because A_α^{UMP} is optimal when $\theta_0 = 0$, satisfies the Type I Error constraint, and does not depend on θ_1, it is the UMP of size α. □

A key step in the proof above is the monotonicity of the likelihood ratio. We can extend the existence of a UMP for a more general class of statistical models with that property. Consider the following class.

Definition 3. *A statistical model* $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}\}$ *has monotone likelihood ratio if there exists a function* $T : \mathbb{R} \to \mathbb{R}$ *such that for all* $\theta_0 < \theta_1$, P_{θ_0} *and* P_{θ_1} *are distinct and* $f_{\theta_1}(x)/f_{\theta_0}(x) = g_1(T(x))/g_0(T(x))$ *for some functions* g_0, g_1 *with* g_1/g_0 *nondecreasing.*

The following theorem characterizes the UMP in the class of models defined above for the hypothesis $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$.

Theorem 2 (UMP with Monotone Likelihood Ratio)**.** *Suppose the statistical model has monotone likelihood ratio, then there is a UMP for the hypothesis* $\Theta_0 = (-\infty, 0]$, $\Theta_1 = (0, \infty)$ *which is given by*

$$A^{UMP}(x) = \begin{cases} 1 & \textit{if } T(x) > c \\ \gamma & \textit{if } T(x) = c \\ 0 & \textit{if } T(x) < c \end{cases},$$

where $\gamma \in [0, 1]$ *and* c *and* γ *are uniquely determined by the Type I Error constraint.*

Section 3.4 of Lehman and Romano (4th edition) gives many examples of distributions satisfying the monotone likelihood ratio property.