ECE 598 - Fall 2025

Lecture 7: Minimax Lower Bounds Part II

18th September, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we will study a generalization of the two-point method in which the second point is replaced with a mixture. This minor change makes this method significantly more potent in testing problems involving a "k-subset" structure and for functional estimation tasks.

1 Generalized Two-Point Method

As before, we are working in a decision-theoretic setting with model $\{P_\theta : \theta \in \Theta\}$, decision space $\mathcal{W}$, and a loss function $L : \Theta \times \mathcal{W} \to \mathbb{R}$.

Theorem 1.1. *Suppose there exist $\theta_0 \in \Theta$, and $\Theta_1 \subset \Theta$, satisfying the following* uniform separation *condition with some $\omega > 0$:*

$$\inf_{w \in \mathcal{W}} \inf_{\theta_1 \in \Theta_1} \frac{L(\theta_0, w) + L(\theta_1, w)}{2} \geq \omega.$$

Let μ denote any probability measure supported on Θ_1, and let P_μ denote the mixture distribution satisfying

$$P_\mu(E) = \int_{\Theta_1} P_\theta(E) d\mu(\theta), \quad \text{for all} \quad E \in \mathcal{F}_{\mathcal{X}}.$$

Then, we have the following lower bound:

$$R^*(\Theta, \mathcal{W}) = \inf_{P_{W|\boldsymbol{X}}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[L(\theta, W) \right] \ \geq \ \omega \left(1 - TV(P_{\theta_0}, P_\mu) \right).$$

Remark 1.2. Observe that the only difference from the two-point lower bound we saw in the previous lecture is that $TV(P_{\theta_0}, P_{\theta_1})$ is replaced with $TV(P_{\theta_0}, P_\mu)$. Since $P_\mu = \mathbb{E}_{\theta \sim \mu}[P_\theta]$, the convexity property of total variation implies that $TV(P_{\theta_0}, P_\mu) \leq \mathbb{E}_{\theta_1 \sim \mu}[TV(P_{\theta_0}, P_{\theta_1})]$. This fact hints at the use cases of Theorem 1.1 to be problems where the total variation (and other divergences) between the mixture P_μ and P_{θ_0} can be much smaller than the total variation between P_{θ_0} and any individual P_θ for $\theta \in \Theta_1$.

Proof of Theorem 1.1. Let $\pi = \frac{1}{2}\left(\delta_{\theta_0} + \mu\right)$ denote a prior distribution over the parameter space. Then, we know that the minimax risk is always lower bounded by any Bayes risk, which implies

$$\sup_{\theta \in \Theta} R(\theta, P_{W|\boldsymbol{X}}) \geq R(\pi, P_{W|\boldsymbol{X}}) = \frac{1}{2}\left(\mathbb{E}_{\theta_0}[L(\theta_0, W)] + \mathbb{E}_{\underline{\theta} \sim \mu}[L(\underline{\theta}, W)] \right)$$

Now, we can expand the two terms in the RHS as follows (assuming densities p_θ, p_μ):

$$\mathbb{E}_{\theta_0}[L(\theta_0, W)] = \int_{\mathcal{X}} \left(\int_{\mathcal{W}} L(\theta_0, w) p_{W|\boldsymbol{X}}(w|x) dw \right) p_{\theta_0}(x) =: \int f_0(x) p_{\theta_0}(x) dx, \quad \text{and}$$

$$\begin{aligned}
\mathbb{E}_{\underline{\theta}\sim\mu}[L(\underline{\theta}, W)] &= \int_{\mathcal{X}} p_\mu(x) \left(\int_{\Theta_1} \frac{p_\theta(x)}{p_\mu(x)} \mu(\theta) \left(\int_{\mathcal{W}} L(\theta, w) p_{W|\boldsymbol{X}}(w|x) dw \right) d\theta \right) dx \\
&= \int_{\mathcal{X}} p_\mu(x) \left(\int_{\mathcal{W}} p_{W|\boldsymbol{X}}(w) \left(\int_{\Theta_1} L(\theta, w) \frac{p_\theta(x)}{p_\mu(x)} \mu(\theta) d\theta \right) dw \right) dx \\
&:= \int_{\mathcal{X}} g(x) p_\mu(x) dx
\end{aligned}$$

Now, observe that $(p_\theta(x)\mu(\theta)/p_\mu(x)) = \mu(\theta|x)$ is the posterior distribution of $\underline{\theta}$, which implies

$$\begin{aligned}
f(x) + g(x) &= \int_{\mathcal{W}} p_{W|\boldsymbol{X}}(w|x) \left(\int_{\Theta_1} \left(L(\theta, w) + L(\theta_0, w) \right) \mu(\theta|x) d\theta \right) dw \\
&\geq 2\omega \int_{\mathcal{W}} p_{W|\boldsymbol{X}}(w|x) \left(\int_{\Theta_1} \mu(\theta|x) d\theta \right) dw = 2\omega.
\end{aligned}$$

The inequality above relies on the separation assumption. Finally, this implies that

$$\begin{aligned}
\sup_{\theta\in\Theta} R(\theta, P_{W|\boldsymbol{X}}) \;&\geq\; \frac{1}{2} \left(\int f(x) p_{\theta_0}(x) dx + \int g(x) p_\mu(x) dx \right) \\
&\geq\; \frac{1}{2} \int \left(f(x) + g(x) \right) \min\{p_{\theta_0}(x),\, p_\mu(x)\} dx \\
&\geq \omega \int \min\{p_{\theta_0}(x),\, p_\mu(x)\} dx = \omega(1 - TV(P_{\theta_0},\, P_\mu)),
\end{aligned}$$

where the last equality uses the fact that $TV(P, Q) = 1 - \int \min\{p, q\}$. □

Remark 1.3. A close look at the proof suggests that the same argument would have also worked for the case of two mixtures (instead of point-vs-mixture).

1.1 Divergence Bound

To apply Theorem 1.1 in practice, we first need to fix (θ_0, Θ_1), then select an appropriate mixture distribution μ, and finally get an upper bound on $TV(P_{\theta_0}, P_\mu)$. In most cases, it suffices to choose (θ_0, Θ_1) to maximize ω (the separation) while controlling $TV(P_{\theta_0}, P_\mu)$ to a value less than $1/2$. It turns out that for mixtures, working with chi-squared distance is most convenient, as we explain next.

For simplicity, throughout we will assume that P_θ has a density p_θ with respect to some common dominating measure (which we simply denote by dx). Recall that the chi-squared divergence between any pair (P, Q) with densities (p, q) is defined as

$$\chi^2(P \parallel Q) = \int \left(\frac{p(x)}{q(x)} - 1 \right)^2 q(x) dx = \int \frac{\left(p(x) - q(x)\right)^2}{q(x)} dx = \int \frac{p(x)^2}{q(x)} dx - 1.$$

Our next result shows why this formulation of chi-squared is useful in handling mixtures.

Lemma 1.4. *Assume throughout that $\int p_\theta(x)^2/p_{\theta_0}(x)dx < \infty$ for all $\theta \in \Theta_1$. Then, we have the following:*

$$1 + \chi^2(P_\mu, P_{\theta_0}) = \mathbb{E}_{\theta,\theta'\sim\mu}\left[\int \frac{p_\theta(x)p_{\theta'}(x)}{p_{\theta_0}(x)}dx\right] = \mathbb{E}_{\theta,\theta'\sim\mu}\left[\int \ell_{\theta,\theta_0}(x)\ell_{\theta',\theta_0}(x)p_{\theta_0}(x)dx\right]$$
$$= \mathbb{E}_{\theta,\theta'\sim\mu}\left[\langle \ell_{\theta,\theta_0}, \ell_{\theta',\theta_0}\rangle_{L^2(P_{\theta_0})}\right],$$

where $\ell_{\theta,\theta_0}(x) = p_\theta(x)/p_{\theta_0}(x)$, and we use $\langle\cdot,\cdot\rangle_{L^2(P_{\theta_0})}$ to denote the inner product in the space of square integrable functions (w.r.t. P_{θ_0}).

Proof of Lemma 1.4. We know from the definition of chi-squared divergence that

$$1 + \chi^2(P_\mu \parallel P_{\theta_0}) = \int \frac{p_\mu(x)^2}{p_{\theta_0}(x)}dx = \int \frac{\left(\int p_\theta(x)d\mu(\theta)\right)\left(\int p_{\theta'}(x)d\mu(\theta')\right)}{p_{\theta_0(x)}}dx$$
$$= \int d\mu(\theta) \int d\mu(\theta') \int \frac{p_\theta(x)p_{\theta'}(x)}{p_{\theta_0}(x)}dx.$$

This completes the proof. □

This result indicates that the chi-squared divergence between a point and a mixture distribution depends on the average "similarity" (as measured by the inner product) between two randomly drawn independent distributions according to μ.

In many applications, we work with i.i.d. observations, and our next result shows the crucial property of tensorization which makes chi-squared divergence the appropriate choice when working with mixtures.

Lemma 1.5. *For any $\theta \in \Theta$, let P_θ^n denote the n-fold product measure, and for some probability measure μ supported on Θ_1, let P_μ^n denote $\mathbb{E}_{\underline{\theta}\sim\mu}[P_{\underline{\theta}}^n]$. Then, with $\kappa(\theta,\theta') := \langle \ell_{\theta,\theta_0}, \ell_{\theta',\theta_0}\rangle_{L^2(P_{\theta_0})}$, we have the following:*

$$1 + \chi^2\left(P_\mu^n \parallel P_{\theta_0}^n\right) = \mathbb{E}_{\underline{\theta},\underline{\theta}'\sim\mu}\left[\kappa(\underline{\theta},\underline{\theta}')^n\right].$$

Proof of Lemma 1.5. The proof is a simple consequence of the previous derivation. In particular, from Lemma 1.4, we know that

$$1 + \chi^2(P_\mu^n \parallel P_{\theta_0}^n) = \mathbb{E}_{\underline{\theta},\underline{\theta}'\sim\mu}\left[\left\langle \frac{p_{\underline{\theta}}^n(x^n)}{p_{\theta_0}^n(x^n)}, \frac{p_{\underline{\theta}'}^n(x^n)}{p_{\theta_0}^n(x^n)}\right\rangle_{L^2(P_{\theta_0^n})}\right].$$

Now, on expanding the inner product term, we get

$$\left\langle \frac{p_{\underline{\theta}}^n(x^n)}{p_{\theta_0}^n(x^n)}, \frac{p_{\underline{\theta}'}^n(x^n)}{p_{\theta_0}^n(x^n)}\right\rangle_{L^2(P_{\theta_0^n})} = \int \frac{p_{\underline{\theta}}^n(x^n)p_{\underline{\theta}'}^n(x^n)}{p_{\theta_0}^n(x^n)}dx^n$$
$$= \int \frac{p_{\underline{\theta}}(x_1)p_{\underline{\theta}'}(x_1)}{p_{\theta_0}(x_1)}dx_1 \ldots \int \frac{p_{\underline{\theta}}(x_1)p_{\underline{\theta}'}(x_1)}{p_{\theta_0}(x_1)}dx_n = \kappa(\underline{\theta},\underline{\theta}')^n.$$

This completes the proof. $\square$

Remark 1.6. The previous two lemmas tell us that the chi-squared divergence between $P_\mu^n = \mathbb{E}_{\underline{\theta}\sim\mu}[P_{\underline{\theta}}^n]$ and P_{θ_0} is controlled by the average value of $\kappa(\underline{\theta}, \underline{\theta}')$; a measure of how similar two randomly drawn product distributions in Θ_1 are. This gives us an indication of the type of problems in which the point-vs-mixture approach is useful: if each θ_1 is such that P_{θ_0} and P_{θ_1} are quite distinct from each other, but any two P_{θ_1} and $P_{\theta_1'}$ are almost "orthogonal". In such cases, the two-point method would lead to a suboptimal lower bound (owing to the large distinctness between P_{θ_0} and P_{θ_1}), but a mixture method may be more useful (owing to the almost orthogonality between two randomly drawn elements from μ).

2 Application: Uniformity Testing

Let us consider the hypothesis testing problem within the minimax framework. For some $\{P_\theta : \theta \in \Theta\}$, we are given n i.i.d. observations $\boldsymbol{X} = X^n = (X_1, \ldots, X_n)$ drawn from an unknown P_θ. Our goal is to test between

$$H_0 : \theta \in \Theta_0, \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \quad \text{for disjoint } \Theta_0, \Theta_1 \subset \Theta.$$

A randomized hypothesis test can be represented by a mapping $\Psi : \mathcal{X}_n := \mathcal{X}^n \to [0,1]$, with $\Psi(\boldsymbol{x})$ denoting the probability of deciding that H_1 is true. In other words, the decision space is $\mathcal{W} = \{0,1\}$, and our decision is $W \sim \mathrm{Bernoulli}(\Psi(X^n))$. Then, the minimax risk with the 0-1 loss is defined as

$$R_n^*(\Theta_0, \Theta_1) = \inf_{\Psi} \sup_{\theta \in \Theta_0 \cup \Theta_1} \mathbb{E}_\theta[\mathbf{1}_{W \neq h_\theta}] = \inf_{\Psi} \sup_{\theta \in \Theta_0 \cup \Theta_1} \mathbb{P}_\theta(W \neq h_\theta), \quad \text{where} \quad h_\theta = \mathbf{1}_{\theta \in \Theta_1}.$$

A simple instance of this problem is for the identity testing for discrete distributions. Assume that $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_X$ for some distribution supported on a finite alphabet $\mathcal{X}$ with $|\mathcal{X}| = k$, and let U_k denote the uniform distribution over $\mathcal{X}$. Then, for some $\epsilon > 0$, consider the problem:

$$H_0 : P_X = U_k, \quad \text{versus} \quad H_1 : \|P_X - U_k\|_1 \geq \epsilon.$$

Here, the parameter space is $\Theta = \Delta_k$, with $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta : \|\theta_0 - \theta\|_1 \geq \epsilon\}$, where $\theta_0 = (1/k, \ldots, 1/k)$. This task is called uniformity testing in the theoretical computer science literature.

Lower Bound via Theorem 1.1. The first step is note that the "separation condition" is satisfied with $\omega = 1/2$: for any $\theta \in \Theta_1$, we have

$$L(\theta_0, w) + L(\theta, w) = \mathbf{1}_{w=1} + \mathbf{1}_{w=0} = 1.$$

Hence, Theorem 1.1 implies that the minimax risk in this case is lower bounded by

$$R_n^*(\theta_0, \epsilon) \geq \sup_\mu \frac{1}{2}\left(1 - TV(P_\mu, P_{\theta_0})\right),$$

where μ is any probability measure of the alternative set. We will now describe Paninski's construction of this mixture.

Assume that k is even, and pair off the coordinates into $\{(2j-1, 2j) : 1 \le j \le k/2\}$. let $\boldsymbol{v} = (v_1, \ldots, v_{k/2}) \in \{-1, \}^{k/2}$ be drawn i.i.d. from a Rademacher distribution (i.e., ± 1 w.p. $1/2$ each), and define

$$q_{\boldsymbol{v}} \in \Delta_k, \quad \text{with} \quad q_{\boldsymbol{v}}[2j-1] = \frac{1+\epsilon v_j}{k}, \quad \text{and} \quad q_{\boldsymbol{v}}[2j] = \frac{1-\epsilon v_j}{k}, \quad \text{for } j \in [k/2].$$

It is easy to verify that each $q_{\boldsymbol{v}}$ lies in Θ_1, and the mixture distribution μ is uniformly distributed over the subset $\{q_{\boldsymbol{v}} : \boldsymbol{v} \in \{-1,1\}^{k/2}\} \subset \Theta_1$. Interestingly, we have $\mathbb{E}_{\boldsymbol{v}}[q_{\boldsymbol{v}}] = p_{\theta_0}$. Based on this, we can compute the chi-squared divergence as follows:

$$\begin{aligned}
\kappa(\boldsymbol{v}, \boldsymbol{v}') &= \sum_{i=1}^{k} \frac{q_{\boldsymbol{v}}[i] q_{\boldsymbol{v}'}[i]}{1/k} = k \sum_{j=1}^{k/2} \left(\frac{1+\epsilon v_j}{k} \frac{1+\epsilon v_j'}{k} + \frac{1-\epsilon v_j}{k} \frac{1-\epsilon v_j'}{k} \right) \\
&= \frac{1}{k} \sum_{j=1}^{k/2} \left(1 + \epsilon(v_j + v_j') + \epsilon^2 v_j v_j' + 1 - \epsilon(v_j + v_j') + \epsilon^2 v_j v_j' \right) = 1 + \frac{2\epsilon^2}{k} \sum_{j=1}^{k/2} v_j v_j'.
\end{aligned}$$

Now, observe that each $s_j := v_j v_j'$ is also a Rademacher random variable. Hence, we have

$$\begin{aligned}
1 + \chi^2(P_\mu^n \parallel P_{\theta_0}) &= \mathbb{E}_{\boldsymbol{v},\boldsymbol{v}'} \left[\left(1 + \frac{2\epsilon^2}{k} \sum_{j=1}^{k/2} s_j \right)^n \right] \\
&\le \mathbb{E}_{\boldsymbol{v},\boldsymbol{v}'} \left[\exp\left(\frac{2n\epsilon^2}{k} \sum_{j=1}^{k/2} s_j \right) \right] && (\text{since } 1 + x \le e^x) \\
&= \prod_{j=1}^{k/2} \mathbb{E}\left[\exp\left(\frac{2n\epsilon^2}{k} s_j \right) \right] = \prod_{j=1}^{k/2} \frac{1}{2} \left(e^{2n\epsilon^2/k} + e^{-2n\epsilon^2/k} \right) \\
&\le \prod_{j=1}^{k/2} e^{2n^2\epsilon^4/k^2} && (\text{since } e^x + e^{-x} \le 2e^{x^2/2}) \\
&= e^{n^2\epsilon^4/k}.
\end{aligned}$$

Thus, using the fact that $TV(P,Q) \le \sqrt{\chi^2(P \parallel Q)/2}$, we get

$$R_n^*(\theta_0, \epsilon) \ge \frac{1}{2} \left(1 - \sqrt{\frac{e^{n^2\epsilon^4/k} - 1}{2}} \right). \tag{1}$$

Interpreting the lower bound. This lower bound can be used to characterize fundamental limits on either the detection boundary, or the sample complexity. In particular, suppose we wish to answer the question: For a fixed n, k, suppose we have a procedure that can achieve a minimax risk of $r \in (0,1)$. Then, what is the smallest possible value of $\epsilon \equiv \epsilon_{n,k,r}$? To answer this, note that (1) implies

$$\log(1+2(1+2r)^2) \le n^2\epsilon^4/k \quad \implies \quad \epsilon_{n,k,r} \ge \frac{c_r k^{1/4}}{\sqrt{n}} \quad \text{for} \quad c_r = \left(\log(1+2(1+2r)^2)\right)^{1/4}.$$

This characterizes the *detection boundary*, or a lower limit on the closest alternative that can be distinguished well enough by any test. Conversely, we can also characterize the sample complexity, which is the smallest n for which there exists a procedure with a minimax risk of r (with ϵ, k fixed). The above equation tells us that

$$n_{\epsilon,k,r} \ge \frac{c_r^2\sqrt{k}}{\epsilon^2}.$$

The key benefit of the two point method is that it can capture the k-dependence of the detection boundary / sample complexity.

Failure of the two-point method. If we were to use the two-point method, then we have for any Q in the alternative class

$$\chi^2(Q^n \parallel P_{\theta_0}^n) = \left(1+\chi^2(Q \parallel P_{\theta_0})^n\right) - 1.$$

Now, we can show that

$$\inf_{q:\|q-p_{\theta_0}\|_1 \ge \epsilon} \chi^2(q \parallel p_{\theta_0}) = \epsilon^2,$$

achieved at the pmf with equal $\pm\epsilon$ perturbation from the uniform pmf. This gives us

$$\chi^2(Q^n \parallel P_{\theta_0}^n) \le (1+\epsilon^2)^n - 1 \le e^{n\epsilon^2} - 1.$$

This will result in

$$\epsilon_{n,k,r} \gtrsim \frac{1}{\sqrt{n}}, \quad \text{and} \quad n_{\epsilon,k,r} \gtrsim \frac{1}{\epsilon^2},$$

thus not capturing the k dependence.

Achievability. One constructive approach for addressing this task is based on the so-called "collision statistic"

$$C_n = \frac{1}{\binom{n}{2}} \sum_{i \ne j} \mathbf{1}_{X_i = X_j}.$$

The idea is that in expectation the number of collisions will be the smallest under the uniform distribution, and hence we can reject the null if C_n is above an appropriately chosen threshold. We will work out the details in Homework 2.