ECE 598 - Fall 2025

Lecture 21: Designing Sequential Tests: Part VI

November 13th, 2025

Instructor: Shubhanshu Shekhar

We continue with our discussion of sequential two-sample tests, but in this lecture we go beyond real-valued observations to observations lying in some general alphabet $\mathcal{X}$. Surprisingly, the exact same design principle underlying our sequential Kolmogorov-Smirnov (KS) test is also applicable in this more general setting. We introduce a class of distance metrics, called integral probability metrics or IPMs, and use them to design and analyze sequential two-sample tests. In the process, we will establish some interesting connections between regret of an prediction strategy and performance of our sequential tests.

1 Designing Two-Sample Tests Using IPMs

As before, we consider the sequential two-sample testing problem, where we have a stream of paired observations $\{(X_i, Y_i) : i \geq 1\} \overset{i.i.d.}{\sim} P_X \times P_Y$ lying in an alphabet $\mathcal{X}$, and our goal is to decide between

$$H_0 : P_X = P_Y \quad \text{versus} \quad H_1 : P_X \neq P_Y.$$

Formally, we want to construct a level-α test of power-one, which is a stopping time τ, such that $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$, and $\mathbb{P}_{H_1}(\tau < \infty) = 1$.

In the previous lecture, we constructed a sequential test based on the Kolmogorov-Smirnov metric, defined as

$$D_{KS}(P_X, P_Y) = \sup_{x \in \mathbb{R}} |F_X(x) - F_Y(y)| = \sup_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)],$$

where $\mathcal{G} = \{\pm \mathbf{1}_{(-\infty, x]} : x \in \mathbb{R}\}$. To generalize this idea to arbitrary alphabets, we can choose a class of "test functions" $\mathcal{G} \subset \{g : \mathcal{X} \to \mathbb{R}\}$ and use it to define a notion of distance as follows (assuming $\mathcal{G}$ is closed under negation):

$$D_{\mathcal{G}}(P_X, P_Y) = \sup_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)].$$

If $\mathcal{G}$ is "rich enough", it should contain enough test functions to distinguish between any two probability measures on $\mathcal{X}$, and thus define an actual metric (Müller, 1997). We give some examples next:

- For distributions on $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$, let $\mathcal{G} = \{\mathbf{1}_E : E \in \mathcal{F}_{\mathcal{X}}\}$ denote the set of all indicator functions of measurable sets. Then, $D_{\mathcal{G}}$ is equal to the total variation metric.
- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote a positive definite kernel on the alphabet $\mathcal{X}$, and let $\mathcal{H}_k$ denote the reproducing kernel Hilbert space associated with this kernel. Then, if $\mathcal{G}$ is the unit-norm ball in $\mathcal{H}_k$; that is, $\mathcal{G} = \{h \in \mathcal{H} : \|h\|_k \leq 1\}$, then the associated $D_{\mathcal{G}}$ is the well known kernel maximum mean discrepancy (or kernel MMD) distance.

- Let $\mathcal{X}$ be some compact subset of $\mathbb{R}^d$, and let $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R} : g \text{ is 1-Lipschitz}\}$ denote the set of all 1-Lipshitz functions on $\mathcal{X}$. Then, by Kantorovich duality, we know that the Wasserstein-1 metric (W_1-metric) between two distributions P_X and P_Y on $\mathcal{X}$ is equal to $D_{\mathcal{G}}(P_X, P_Y)$.

In the next two subsections, we show how these distance metrics can be used to design sequential tests, starting with an oracle test that depends on some population terms, and then discussing practical tests that replace the population terms with their data-driven counterparts.

1.1 An Oracle Sequential Test

The key idea is to reduce two-sample testing to the problem of testing for means of bounded random variables, that we already know how to solve from Lectures 18-19. To do this, fix a function class $\mathcal{G}$, and assume it is large enough to ensure that $P \neq Q \implies D_{\mathcal{G}}(P, Q) > 0$. Under this condition, for any pair P_X, P_Y under H_1, we can define the *witness function*

$$g^* \equiv g^*(P_X, P_Y, \mathcal{G}) \in \operatorname*{argmax}_{g \in \mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)].$$

Then, for $(X, Y) \sim P_X \times P_Y$, the two-sample testing problem reduces to

$$H_0 : \mathbb{E}[V^*] = 0, \quad \text{versus} \quad H_1 : \mathbb{E}[V^*] > 0, \quad \text{where} \quad V^* = g^*(X) - g^*(Y).$$

Assuming that each $g \in \mathcal{G}$ satisfies $\sup_{x \in \mathcal{X}} |g(x)| \leq 1/2$, we can define a test for this problem as

$$\tau^* = \inf\{n \geq 1 : W_n^* \geq 1/\alpha\}, \quad \text{with} \quad W_0^* = 1,\ W_n^* = W_{n-1}^* \times (1 + \lambda^* V_n^*),$$

and λ^* is the log-optimal or Kelly bet

$$\lambda^* = \operatorname*{argmax}_{\lambda \in \left[-\frac{1}{2}, \frac{1}{2}\right]} \mathbb{E}_{P_X \times P_Y}[\log(1 + \lambda\,(g^*(X) - g^*(Y))].$$

The reason for restricting λ to $[-1/2, 1/2]$ and $\mathcal{G}$ to be uniformly bounded in absolute value by $1/2$ is to ensure the nonnegativity of the process $\{W_n^* : n \geq 1\}$. It is not too difficult to show that τ^* is a level-α test with $\mathbb{E}[\tau^*] = \mathcal{O}(\log(1/\alpha)/D_{\mathcal{G}}^2(P_X, P_Y))$ under the alternative.

1.2 A Practical Sequential Test

Clearly the test τ^* discussed in the previous subsection is not practical. Nevertheless it serves as a good template for designing practical tests, by replacing the two population terms g^* and λ^* (both of which depend on the unknown distributions P_X, P_Y) with their data-driven counterparts:

- We refer to any policy for selecting $\{\lambda_n : n \geq 1\}$ in a predictable manner (i.e., each λ_n is $\mathcal{F}_{n-1}$-measurable) as a *betting strategy*, and denote it by $\mathcal{A}_b$. A natural choice from the theoretical perspective is the *mixture method* or the Online Newton Step (ONS) method, both of which achieve a $c \log n$ regret, with some universal constant $c > 0$.

- We refer to any policy for selecting a predictable sequence $\{g_n : n \geq 1\}$ that approximates the unknown witness function g^* as the *prediction strategy*, and denote it by $\mathcal{A}_p$. To measure the quality of a prediction strategy, we introduce the notion of regret, which is defined as

$$\mathrm{Reg}_n(\mathcal{A}_p, X^n, Y^n) = \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} g(X_i) - g(Y_i) - \sum_{i=1}^{n} g_i(X_i) - g_i(Y_i),$$

  where $\{(X_i, Y_i) : i \geq 1\}$ denote the sequence of paired observations, and $\{g_i : i \geq 1\}$ is the sequence of data-driven functions selected by the prediction strategy. As we will see, the regret of $\mathcal{A}_p$ will play a crucial role in determining the performance of our practical sequential test.

With these two components, we can now construct a purely data-driven practical sequential two-sample test:

Definition 1.1 (Practical Two-Sample Test)**.** Fix $W_0 = 1$, $\lambda_1 = 0$, and set g_1 to be an arbitrary element of $\mathcal{G}$.
For $t = 1, 2, \ldots$:

- Observe the next pair $(X_t, Y_t) \sim P_X \times P_Y$.
- Update $W_t \leftarrow W_{t-1} \times (1 + \lambda_t V_t)$, where $V_t = g_t(X_t) - g_t(Y_t)$.
- Stop if $W_t \geq 1/\alpha$.
- Update $\lambda_{t+1} = \mathcal{A}_b(X^t, Y^t)$.
- Update $\boldsymbol{g}_{t+1} = \mathcal{A}_p(X^t, Y^t)$.

In other words, we define the stopping time $\tau = \inf\{n \geq 1 : W_n \geq 1/\alpha\}$.

We can establish the following performance guarantees for this test.

Theorem 1.2. *The test τ constructed in Theorem 1.1 using a betting strategy $\mathcal{A}_b$ and a prediction strategy $\mathcal{A}_p$ satisfies the following:*

- *For any betting and prediction strategies, τ satisfies the level-α property under the null: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
- *Suppose the regret of the betting strategy is $c \log n$; that is, for any sequence $n \geq 1$ and $v_1, v_2, \ldots, v_n \in [-1, 1]$, we have $\sum_{i=1}^{n} \log(1 + \lambda_i v_i) \geq \sup_{\lambda \in [-1/2, 1/2]} \log(1 + \lambda v_i) - c \log n$. Then, we have the following under H_1, with $\Delta = D_{\mathcal{G}}(P_X, P_Y)$:*

$$\limsup_{n \to \infty} \frac{\mathrm{Reg}_n(\mathcal{A}_p, X^n, Y^n)}{n} \overset{a.s.}{<} \Delta \quad \Longrightarrow \quad \mathbb{P}_{H_1}(\tau < \infty) = 1.$$

- *Under a stronger high probability control of the regret of $\mathcal{A}_p$, we can also bound the expected stopping time; that is,*

$$\mathbb{P}\left(\frac{\mathrm{Reg}_n(\mathcal{A}_p, X^n, Y^n)}{n} \geq \frac{c}{\sqrt{n}}\right) \leq \frac{1}{2n^2} \quad \Longrightarrow \quad \mathbb{E}[\tau] = \mathcal{O}\left(\frac{\log(1/\alpha\Delta)}{\Delta^2}\right).$$

The proof of the level-α property simply follows from the fact that for any feasible betting and prediction strategies, the process $\{W_n : n \geq 1\}$ is a non-negative martingale with an initial value of 1, and hence Ville's inequality applies. The proof of the remaining two-steps is in the next section.

2 Analysis of the Sequential Test

Power-one property. We first show that if

$$\limsup_{n\to\infty} \frac{\mathrm{Reg}_n}{n} < \Delta(P_X, P_Y) := \sup_{g\in\mathcal{G}} \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)] \quad \implies \quad \mathbb{P}_{H_1}(\tau < \infty) = 1.$$

To do this observe the following:

$$\mathbb{P}_{H_1}(\tau < \infty) = \lim_{n\to\infty} \mathbb{P}_{H_1}(\tau \leq n) = 1 - \lim_{n\to\infty} \mathbb{P}_{H_1}(\tau > n).$$

Thus, it suffices to analyze the term $\mathbb{P}_{H_1}(\tau > n)$. Observe that

$$\mathbb{P}_{H_1}(\tau > n) \leq \mathbb{P}_{H_1}\left(\log W_n < \log(1/\alpha)\right).$$

By the regret guarantee on the betting strategy, we know that

$$\log w_n \geq \sup_{|\lambda|\leq 1/2} \sum_{i=1}^{n} \log(1 + \lambda v_i) - c \log n,$$

where we use v_i to denote $g_i(X_i) - g_i(Y_i)$. On taking the second order Taylor approximation, and optimizing this lower bound over λ, we get

$$\sup_{|\lambda|\leq 1/2} \sum_{i=1}^{n} \log(1 + \lambda v_i) \geq \sup_{|\lambda|\leq 1/2} \lambda S_n - \lambda^2 M_n \geq \sup_{|\lambda|\leq 1/2} \lambda S_n - \lambda^2 n = \frac{S_n^2}{4n},$$

where we use $S_n = \sum_{i=1}^n v_i$ and $M_n = \sum_{i=1}^n v_i^2$, and c is some universal constant ≥ 1. Thus, we have

$$\mathbb{P}_{H_1}(\tau > n) \leq \mathbb{P}_{H_1}\left(\frac{S_n^2}{4n} < \log(1/\alpha) + c \log n\right) \leq \mathbb{P}_{H_1}\left(\frac{|S_n|}{n} < 2\sqrt{\frac{c\log(n/\alpha)}{n}}\right).$$

Now, we know that $|S_n| \geq S_n$, and hence

$$\mathbb{P}_{H_1}\left(\frac{|S_n|}{n} < 2\sqrt{\frac{c\log(n/\alpha)}{n}}\right) \leq \mathbb{P}_{H_1}\left(\frac{S_n}{n} < 2\sqrt{\frac{c\log(n/\alpha)}{n}}\right).$$

Finally, we can use the regret guarantee to claim that $S_n > S_n^* - \mathrm{Reg}_n$, where $S_n^* = \sum_{i=1}^n v_i^* = \sum_{i=1}^n g^*(X_i) - g^*(Y_i)$. This implies that

$$\mathbb{P}_{H_1}(\tau > n) \leq \mathbb{P}\left(\frac{S_n^*}{n} - \frac{\mathrm{Reg}_n}{n} < 2\sqrt{\frac{c\log(n/\alpha)}{n}}\right).$$

Now, taking the limit to ∞, we can show that if $\limsup_{n\to\infty} \mathrm{Reg}_n/n < \lim_{n\to\infty} S_n^*/n = \Delta(P_X, P_Y)$, then we have

$$\lim_{n\to\infty} \mathbb{P}_{H_1}(\tau > n) = 0.$$

This in turn implies the power one property as required.

Analysis of Expected stopping time. To analyze the expected stopping time under the alternative, we need to introduce a "Good event":

$$G_n = \{\mathrm{Reg}_n < r_n\} \bigcap \left\{ \left| \frac{1}{n} S_n^* - \Delta \right| < c\sqrt{\frac{\log n}{n}} \right\},$$

and throughout we assume that $\mathbb{P}_{H_1}(G_n^c) \leq 1/n^2$ for all $n \geq 1$ and with $r_n = c\sqrt{n}$. Now, to analyze the expected stopping time, we begin with the observation that

$$\mathbb{E}[\tau] = \sum_{n\geq 0} \mathbb{P}_{H_1}(\tau > n).$$

Hence, the problem essentially reduces to that of analyzing the tail probability. To do that, we note the following.

$$\begin{aligned}
\mathbb{E}[\tau] &= \sum_{n\geq 0} \mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) + \mathbb{P}_{H_1}(\{\tau > n\} \cap G_n^c) \\
&\leq 1 + \sum_{n\geq 1} \mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) + \mathbb{P}_{H_1}(\cap G_n^c) && \text{(since } \mathbb{P}(\tau > 0) = 1\text{)} \\
&\leq 1 + \sum_{n\geq 1} \mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) + \sum_{n\geq 1} \frac{1}{n^2}, \\
&\leq \sum_{n\geq 1} \mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) + \mathcal{O}(1). && \left(\text{since } \sum_{n\geq 1} \frac{1}{n^2} = \frac{\pi^2}{6}\right)
\end{aligned}$$

So we will now analyze the event $\mathbb{P}_{H_1}(\{\tau > n\} \cap G_n)$. As before, we can upper bound this with

$$\mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) \leq \mathbb{P}\left(\left\{ \frac{S_n^*}{n} - \frac{\mathrm{Reg}_n}{n} < 2\sqrt{\frac{c\log(n/\alpha)}{n}} \right\} \cap G_n \right).$$

Under the good event G_n, we know that $\mathrm{Reg}_n/n \leq r_n/n$, and that $S_n^*/n > \Delta(P_X, P_Y) - c\sqrt{\log n/n}$. Combining these two facts, we see that

$$\mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) \leq \mathbb{P}\left(\left\{ \Delta < \frac{r_n}{n} + c\sqrt{\frac{\log n}{n}} + 2\sqrt{\frac{c\log(n/\alpha)}{n}} \right\} \cap G_n \right).$$

Define n_0 as

$$n_0 = \inf \left\{ n \geq 1 : \Delta \geq \frac{r_n}{n} + c\sqrt{\frac{\log n}{n}} + 2\sqrt{\frac{c\log(n/\alpha)}{n}} \right\},$$

and observe that

$$\mathbb{P}_{H_1}(\{\tau > n\} \cap G_n) = 0, \quad \text{for all} \quad n \geq n_0.$$

Hence, we have that

$$\mathbb{E}_{H_1}[\tau] \leq n_0 + \mathcal{O}(1).$$

Finally, if $r_n/n \asymp \sqrt{1/n}$, then it is easy to verify that

$$n_0 \asymp \frac{\log(1/\alpha\Delta)}{\Delta^2}.$$

This completes the proof.

References

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.