ECE 598 - Fall 2025

Lecture 20: Designing Sequential Tests: Part V

November 11th, 2025

Instructor: Shubhanshu Shekhar

In this and the next lecture, we will study an important nonparametric testing problem called two-sample testing. We will begin with a discussion of the two-sample testing problem in the non-sequential setting and discuss a general approach for designing such tests based on statistical distances. We will instantiate this idea for the case of real-valued observations using the Kolmogorov-Smirnov (KS) metric, and end the lecture by presenting a scheme for designing a sequential analog of the KS test.

1 Two Sample Testing Problem

Let $\mathcal{X}$ denote any general alphabet, and let $\{(X_i, Y_i) : 1 \leq i \leq n\}$ denote a n i.i.d. $\mathcal{X}$-valued pairs of random variables drawn from a product distribution $P_X \times P_Y$. Then, the goal is to test

$$H_0 : P_X = P_Y, \quad \text{versus} \quad H_1 : P_X \neq P_Y. \tag{1}$$

Equivalently, we can restate this problem as

$$\begin{aligned} &H_0 : P_{XY} \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P_{XY} \in \mathcal{P}_1, \quad \text{with} \\ &\mathcal{P}_0 = \{P \times P : P \in \mathcal{P}(\mathcal{X})\}, \quad \text{and} \quad \mathcal{P}_1 = \{P \times Q : P \neq Q \in \mathcal{P}(\mathcal{X})\}. \end{aligned}$$

As the above formulation illustrates, both the null and hypothesis classes are composite and nonparametric. This problem arises in various applications, such as:

- **A/B Testing.** Used for comparing two-versions of a website, and identifying whether they lead to the same amount of traffic or not.
- **Astrophysics:** Two-sample tests are often used to compare the radiations from different regions of a galaxy.
- **Machine Learning:** Two-sample tests can be used for detecting shifts in distributions between test and training datasets.

Remark 1.1. In general, we may have different number of X and Y observations. We assume the case of paired observations for simplicity.

1.1 General Idea

A general approach for designing two-sample tests proceeds as follows: We choose a distance measure $D : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to [0, \infty]$, and then given the samples $\{X_i : i \in [n]\}$ and $\{Y_i : i \in [n]\}$, we construct a statistic

$$T_n \approx D(\widehat{P}_n, \widehat{Q}_n),$$

where we use $\approx$ to indicate that often the statistic we construct is not exactly the one obtained by plugging-in $\widehat{P}_n, \widehat{Q}_n$ into D. For instance, we may need debiasing in some cases. Now, assuming D is a metric, it follows that $D(P_X, P_Y) = 0$ under H_0 and > 0 under H_1. As a result, a suitably scaled version of the statistic T_n (e.g., by $\sqrt{n}$ or n depending on the situation) will be expected to take small values under H_0, and larger values under H_1. Thus, a two-sample test can be defined as

$$\Psi(X^n, Y^n) = \mathbf{1}_{T_n > t^*_\alpha}, \quad \text{where} \quad t^*_\alpha = \inf\{t \in \mathbb{R} : \mathbb{P}_{H_0}(T_n > t) \leq \alpha\}.$$

Now, the crucial part in designing a (non-sequential) two-sample test is finding the right critical value to reject the null. The optimal value t^*_α depends on the unknown distribution $P_X = P_Y$, and thus cannot be directly employed. In practice, the following methods are often used:

- **Permutation Test.** If the null is true, then the dataset $\{Z_i : 1 \leq i \leq 2n\}$ is i.i.d. from P_X, where $Z_i = X_i \mathbf{1}_{i \leq n} + Y_{i-n} \mathbf{1}_{i > n}$. This means that for any permutation σ over the set $[2n]$, $\{Z_{\sigma(i)} : i \in [2n]\} \stackrel{d}{=} \{Z_i : i \in [2n]\}$. Let T^σ_n denote a statistic computed using a permuted dataset (with the first n observations considered X's). Now, suppose we draw M i.i.d. permutations, and re-compute $\{T^{\sigma_j}_n : 1 \leq j \leq M\}$, then by the invariance to permutations, the rank of the original statistic $R = 1 + |\{j : T^{\sigma_j}_n > T_n\}|$ is uniformly distributed in the set $\{1, \ldots, M+1\}$. We can use this fact to calibrate the test.

- **Asymptotic Null.** In some instances, we can show that the a properly scaled statistic, e.g., $\sqrt{n} T_n \Rightarrow P_0$, for some tractable limiting null distribution P_0 (such as Gaussian, chi-squared, etc.). Hence, we can set t_α to be the $(1-\alpha)$ quantile of the limiting null distribution, and obtain an (asymptotically valid) level-α test.

- **Concentration Inequalities.** Finally, in some situations, we can establish non-asymptotic deviation results of the form: $\mathbb{P}(T_n > \epsilon) \leq f(\epsilon)$ for some closed-form function f. We can then invert this to set $t_\alpha = f^{-1}(\alpha)$, and get a valid test. In practice, this approach often leads to excessively conservative tests.

1.2 Non-sequential KS Test

To illustrate the general discussion in the last subsection, we consider the case of $\mathcal{X} = \mathbb{R}$-valued observations $\{(X_i, Y_i) : 1 \leq i \leq n\}$. In this case, the distributions P_X and P_Y are completely specified by their CDFs, denoted by F_X and F_Y. Based on this fact, we consider a test associated with the KS-distance, defined as

$$D_{KS}(F_X, F_Y) = \|F_X - F_Y\|_\infty = \sup_{u \in \mathbb{R}} |F_X(u) - F_Y(u)|.$$

In other words, the KS distance between two real-valued distributions is the pointwise supremum of the absolute deviation between their CDFs. Thus, given the dataset, we can define the statistic

$$T_n = \sup_{u \in \mathcal{X}} |\hat{F}_{X,n}(u) - \hat{F}_{Y,n}(u)|.$$

Note that this can be done in $O(n \log n)$ cost: we simply need to sort the observations ($O(n \log n)$ operation), and then check the difference at the jumps.

To calibrate the test, we can use the permutation test, but that can often be computationally too expensive. An alternative is to use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, which says

$$\mathbb{P}(D_{KS}(\hat{F}_n, F) > \epsilon) \leq 2e^{-2n\epsilon^2} \quad \Longrightarrow \quad \mathbb{P}_{H_0}(T_n > \epsilon) \leq 4e^{-2n\epsilon^2},$$

where the second statement follows by the triangle inequality (and the fact that $F_X = F_Y$ under H_0). This inequality suggests a critical value of

$$t_\alpha = \sqrt{\frac{\log(4/\alpha)}{2n}}.$$

This approach gives us a nonasymptotic control over the type-I error, but often leads to a loss of power.

An alternative is to look at the asymptotic null distribution of the statistic T_n. It turns out that

$$\sqrt{\frac{n}{2}} T_n \Rightarrow \sup_{0 \leq u \leq 1} |\mathbb{B}(u)| \quad \Longrightarrow \quad t_\alpha^{\text{asy}} = c_\alpha \sqrt{\frac{2}{n}}, \ \mathbb{P}(\sup_{0 \leq u \leq 1} |\mathbb{B}(u)| \leq c_\alpha) = 1 - \alpha.$$

where $\mathbb{B}(u)$ is the Brownian bridge (it is the usual Brownian motion, with the constraint that the end points are equal to 0). The supremum of the Brownian bridge follows the so-called Kolmogorov distribution, and it has a closed form expression of its CDF which can be used to calculate c_α.

2 Sequential Kolmogorov-Smirnov (KS) Test

One key drawback of the non-sequential approach is that the sample-size n must be decided before implementing the test, and hence there is always a chance of allocating too many observations on a simple problem (thus wasting resources), or allocating too few observations on a difficult problem (leading to inconclusive tests). These issues can be addressed in a sequential setting.

We assume that we have a stream of paired observations $\{(X_i, Y_i) : i \geq 1\}$ drawn i.i.d. from a product distribution $P_X \times P_Y$, and as before, we want to construct a level-α power-one test for the problem described in (1). Formally, we want to construct a stopping time τ such that

$$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha, \quad \text{and} \quad \mathbb{P}_{H_1}(\tau < \infty) = 1.$$

We discuss a simple approach customized for the case of KS tests in the next subsection.

Note that we can rewrite the KS metric as

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

where we define $\mathcal{G} = \{\mathbf{1}_{(-\infty, x]}, -\mathbf{1}_{(-\infty, x]} : x \in \mathbb{R}\}$ as the set of all semi-infinite intervals with a $\pm$ sign. Since for every $P_X \neq P_Y$, we know that $D_{KS}(P_X, P_Y) > 0$, we can conclude that there exists a $g^* \equiv g^*(P_X, P_Y)$ such that

$$0 < D_{KS}(P_X, P_Y) = \mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)].$$

Such a g^* is referred to as the "witness function" of P_X, P_Y in $\mathcal{G}$. If on the other hand, the null distribution is true, then for all $g \in \mathcal{G}$ (and beyond), we must have $\mathbb{E}_{P_X}[g(X)] - \mathbb{E}_{P_Y}[g(Y)]$. So if

someone gave us the witness function g^*, then the two-sample problem reduces to that of (bounded) mean testing, and we can solve it by our general betting/portfolio based approach discussed in the previous lectures.

But since we don't know $g^*(P_X, P_Y)$, a natural idea is to learn it on the fly. That is, we can set

$$g_n \in \pm\mathbf{1}_{(-\infty,u_n]} \in \operatorname*{argmax}_{u\in\mathbb{R}} |\hat{F}_{n-1,X}(u) - \hat{F}_{n-1,X}(u)|.$$

This suggests the following definition of a sequential KS test.

Definition 2.1. Set $W_0 = 1$, $g_1 = \mathbf{1}_{\mathbb{R}}$, and for $t = 1, 2, \ldots$:

- Calculate g_n from the previous observations.
- Calculate $\lambda_n \in [-1/2, 1/2]$ based on the previous observations.
- Observe the new pair (X_n, Y_n).
- Update $W_n \leftarrow W_{n-1} \times (1 + \lambda_n(g_n(X_n) - g_n(Y_n)))$.
- Reject the null if $W_n \geq 1/\alpha$.

In other words, $\tau_{KS} = \inf\{n \geq 1 : W_n \geq 1/\alpha\}$.

If the functions g_n approximate g^* sufficiently well with high probability as n increases, we expect that the test defined above should be able to identify the alternatives in a near-optimal manner. We formalize this intuition in the next subsection.

2.1 Analysis of the Sequential KS Test

Theorem 2.2. *The sequential KS test introduced in Theorem 2.1 satisfies the following:*

$$\mathbb{P}_{H_0}(\tau_{KS} < \infty) \leq \alpha, \quad \textit{and} \quad \mathbb{P}_{H_1}(\tau_{KS} < \infty) = 1.$$

Furthermore, if P_X, P_Y are compactly supported, then we also have the following bound on the expected stopping time:

$$\mathbb{E}_{H_1}[\tau_{KS}] = \mathcal{O}\left(\frac{\log(1/\alpha\Delta)}{\Delta^2}\right),$$

where $\Delta = D_{KS}(P_X, P_Y)$.

Proof of the level-α property. This follows directly from the fact that the process $\{W_n : n \geq 0\}$ is a nonnegative martingale under the null.

Proof of the power-one property. We only provide an outline of the proof. First, note that if the regret of the "betting strategy" $\{\lambda_n : n \geq 1\}$ is less than $c \log n$ for some constant $c > 0$, then we have the following, with $v_n = g_n(X_n) - g_n(Y_n)$:

$$\log W_n \geq \left(\sup_{\lambda\in[-1/2,1/2]} \sum_{i=1}^{n} \log(1 + \lambda v_i)\right) - c \log n.$$

By using a logarithmic lower bound, $\log(1+\lambda v) \geq \lambda v - \lambda^2 v^2$, we can further lower bound this with

$$\log W_n \geq \sup_{\lambda \in [-1/2, 1/2]} \lambda S_n - \lambda^2 M_n - c \log n,$$

where we use $S_n = \sum_{i=1}^n v_i$, and $M_n = \sum_{i=1}^n v_i^2$. Now, using the fact that $M_n \leq n$, we get the further lower bound

$$\log W_n \geq \sup_{\lambda \in [-1/2, 1/2]} \lambda S_n - \lambda^2 n - c \log n, \tag{2}$$

which is optimized at $\lambda = S_n/2n \in [-1/2, 1/2]$. Hence, we can conclude that

$$\frac{1}{n} \log W_n \geq \frac{S_n^2}{4n^2} - \frac{c \log n}{n}.$$

To show the power-one property, we can use the fact that under some regularity conditions, $g_n \to g^*$ almost surely, and hence $S_n/n \to \Delta = D_{KS}(P_X, P_Y) > 0$ almost surely (by Cesaro means theorem). Hence, $\liminf_{n\to\infty} \log W_n/n > 0$ almost surely, which implies the power-one property.

Proof of the expected stopping time. To analyze the expected stopping time, we start with the fact that

$$\mathbb{E}[\tau] = \sum_{n \geq 1} \mathbb{P}(\tau \geq n) = \sum_{n \geq 1} \mathbb{P}\left(\frac{1}{n} \log W_n < \frac{\log(1/\alpha)}{n} \right).$$

Now, from (2), we can further upper bound this with

$$\mathbb{E}[\tau] \leq \sum_{n \geq 1} \mathbb{P}\left(\frac{S_n^2}{4n^2} < \frac{\log(1/\alpha)}{n} + \frac{c \log n}{n} \right) = \sum_{n \geq 1} \mathbb{P}\left(\frac{|S_n|}{n} < \sqrt{\frac{\log(1/\alpha)}{n}} + \sqrt{\frac{c \log n}{n}} \right).$$

Finally, without going into details, we mention that for compactly supported distributions, there exists a strategy of selecting $\{g_n : n \geq 1\}$ in a predictable manner, for which we have

$$\frac{S_n}{n} \geq \sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n g_i(X_i) - g_i(Y_i) - \frac{c_2}{\sqrt{n}} \geq \frac{S_n^*}{n} - \frac{c_2}{\sqrt{n}},$$

where $S_n^* = \sum_{i=1}^n g^*(X_i) - g^*(Y_i)$, and $c_2 > 0$ is some universal constant. This allows us to show

$$\mathbb{P}\left(\frac{|S_n|}{n} < \sqrt{\frac{\log(1/\alpha)}{n}} + \sqrt{\frac{c \log n}{n}} \right) \leq \mathbb{P}\left(\frac{S_n^*}{n} < \sqrt{\frac{c_3 \log(n/\alpha)}{n}} \right).$$

Since $\mathbb{E}[S_n^*] = n\Delta$, we can show that the event $G_n = \{|S_n^*/n - \Delta| \leq \sqrt{c_4 \log n/n}\}$ occurs with probability at least $1 - 1/n^2$ (where $c_4 > 0$ is some universal constant) [We can show this, for example, by using Hoeffding's inequality]. Hence, we have

$$\mathbb{P}\left(\frac{S_n^*}{n} < \sqrt{\frac{c_3 \log(n/\alpha)}{n}} \right) \leq \mathbb{P}\left(\left\{ \frac{S_n^*}{n} < \sqrt{\frac{c_3 \log(n/\alpha)}{n}} \right\} \cap G_n \right) + \mathbb{P}(G_n^c).$$

Under the event G_n, we know that $S_n^* \geq \Delta - \sqrt{c_4 \log n/n}$, which means that we can further upper bound this

$$\mathbb{P}\left(\frac{S_n^*}{n} < \sqrt{\frac{c_3 \log(n/\alpha)}{n}}\right) \leq \mathbb{P}\left(\left\{\Delta < \sqrt{\frac{c_3 \log(n/\alpha)}{n}} + \sqrt{\frac{c_4 \log n}{n}}\right\} \cap G_n\right) + \frac{1}{n^2}.$$

Thus, we can conclude that

$$\mathbb{E}[\tau] \leq \sum_{n \geq 1} \mathbb{P}\left(\Delta < \sqrt{\frac{c_3 \log(n/\alpha)}{n}} + \sqrt{\frac{c_4 \log n}{n}}\right) + \frac{1}{n^2} \leq n_0 + \frac{\pi^2}{6},$$

where $n_0 = \inf\{n \geq 1 : \Delta \geq \sqrt{c_3 \log(n/\alpha)/n} + \sqrt{c_4 \log n/n}\} \asymp \log(1/\alpha\Delta)/\Delta^2$. This completes the proof. □

3 Conclusion

In this lecture, we proposed a sequential analog of Kolmogorov-Smirnov (KS) test and analyzed its performance. Under certain conditions, the expected number of observations needed by this test is $\asymp \mathcal{O}(\log(1/\alpha)/\Delta^2)$, which establishes a key property of sequential methods: they can adapt the expected sample size to the unknown hardness of the problem. In our next lecture, we will see how the same idea essentially can be employed to develop sequential two-sample tests more generally.