ECE 598 - Fall 2025

Lecture 22: Designing Sequential Tests: Part VII

November 18th, 2025

Instructor: Shubhanshu Shekhar

This is our final lecture on the topic of designing and analyzing sequential power-one tests. So far, we have discussed how to construct sequential tests for finite alphabets, for testing bounded means, and two-sample tests. We discussed some common design principles, such as using test martingales, Ville's inequality, and Wald's inequality, used in all these problems. In today's lecture, we will build upon all these ideas to show how we can study a larger class of nonparametric testing problems in a unified manner.

1 An Abstract Testing Problem

In this section, we will introduce an abstract class of hypothesis testing problems, that can be characterized in terms of invariance to certain operators. To describe the problem, we assume that $\{Z_n : n \geq 1\}$ denote a stream of i.i.d. observations from some distribution P_Z, and let $\mathcal{T}_i : \mathcal{Z} \to \mathcal{W}$ for $i = 1, 2$ denote two operators from $\mathcal{Z}$ to some space $\mathcal{W}$ (possibly different from $\mathcal{Z}$). Then, we are interested in deciding between

$$H_0 : \mathcal{T}_1(Z) \overset{d}{=} \mathcal{T}_2(Z) \quad \text{versus} \quad H_1 : \mathcal{T}_1(Z) \overset{d}{\neq} \mathcal{T}_2(Z). \tag{1}$$

This abstract formulation enables a unified treatment of several classical and modern testing problem; from two-sample testing, independence testing, symmetry testing, to testing for fairness and adversarial robustness of machine learning models.

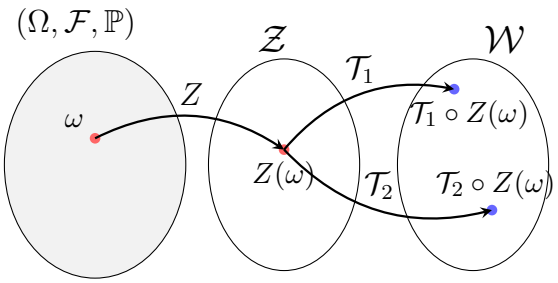

Figure 1: Let Z denote a $\mathcal{Z}$-valued random variable on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By definition, the distribution of the Z is equal to $P_Z = \mathbb{P} \circ Z^{-1}$. The two black curved lines from $\mathcal{Z}$ to $\mathcal{W}$ denote the operators $\mathcal{T}_i$ for $i \in \{1, 2\}$, used to characterize the class of null distributions. In particular, the distribution of the resulting $\mathcal{W}$-valued random variables is $\mathbb{P} \circ (\mathcal{T}_i \circ Z)^{-1} = \mathbb{P} \circ Z^{-1} \circ \mathcal{T}_i^{-1} = P_Z \circ \mathcal{T}_i^{-1}$, and the null hypothesis of our abstract testing problem states that the two distributions, $P_Z \circ \mathcal{T}_1^{-1}$ and $P_Z \circ \mathcal{T}_2^{-1}$, are the same.

1.1 Examples

Before going further, we show how the formulation above models various important problems. We borrow the exposition from Pandeva et al. (2024).

Example 1.1 (Paired Two-Sample Testing)**.** *Given a stream of paired observations:* $\{(X_t, Y_t) : t \geq 1\}$ *drawn i.i.d. from a distribution* $P_X \times P_Y$ *on a product space* $\mathcal{X} \times \mathcal{X}$*, our goal is to decide between the null,* $H_0 : P_X = P_Y$*, against the alternative* $H_1 : P_X \neq P_Y$*. This is a nonparametric testing problem with a composite null and a composite alternative. The null hypothesis class, however, has an interesting symmetry: the joint distribution of* (X, Y) *is the same as the joint distribution of* (Y, X)*. We can formally state this as* $H_0 : (X, Y) \overset{d}{=} \mathcal{T}_{swap}((X, Y))$*, where* $\mathcal{T}_{swap} : \mathcal{X} \times \mathcal{X} \to \mathcal{X} \times \mathcal{X}$*, and* $\mathcal{T}_{swap}((x, y)) = (y, x)$*.*

Example 1.2 (Conditional Independence Testing)**.** *Given observations* $\{(U_t, V_t, W_t) : t \geq 1\}$ *drawn i.i.d. from* P_{UVW}*, we want to test whether* $U \perp\!\!\!\perp V \mid W$ *or not. This problem is fundamentally impossible without further assumptions (Shah and Peters, 2020), and a common structural assumption is that the conditional* $P_{U|W}$ *is known (the* model-X *assumption (Candes et al., 2018)). We can now reframe this problem as follows:*

- *Given* (U, V, W)*, generate a new* $\widetilde{U} \sim P_{U|W}(\cdot|W)$*, and let* Z *denote* $\big((U, V, W), (\widetilde{U}, V, W)\big)$*.*
- *Let* $\mathcal{T}_1$ *and* $\mathcal{T}_2$ *denote the coordinate projections:* $\mathcal{T}_1(Z) = (U, V, W)$ *and* $\mathcal{T}_2(Z) = (\widetilde{U}, V, W)$*.*

With these definitions, conditional independence (CI) testing falls under the abstract framework defined in (1)*.*

Testing for invariance under group actions, such as rotations, is another instantiation of (1).

Example 1.3 (Rotation invariance testing)**.** *Given a stream of observations* $\{(X_t, Y_t) : t \geq 1\}$*, where the* X_t*'s denote images of the (handwritten) digit "*6*", while dataset* Y_t*'s are images that, at a glance, represent the digit "*9*". However, these may essentially be the digit "*6*" but rotated. We aim to determine the statistical relationship between* X_t *and* $\mathcal{T}_{180}(Y_t)$*: the* 180 *degree rotations of* Y_t*. Essentially, we want to decide whether* $(Y_t)_{t \geq 1}$ *are merely rotated versions of "*6*", or truly represent the digit "*9*". Using the swap operator from Theorem 1.1, we define two distinct operators:* $\mathcal{T}_1 = (\mathcal{T}_{180}, \mathcal{T}_{id}) \circ \mathcal{T}_{swap}$*, and* $\mathcal{T}_2 = (\mathcal{T}_{id}, \mathcal{T}_{180})$*, with* $\mathcal{T}_{id}$ *being the identity mapping. Then, the above-defined test is equivalent to testing* $H_0 : \mathcal{T}_1(Z) \overset{d}{=} \mathcal{T}_2(Z)$*, where* $Z = (X, Y)$*.*

Our general framework is not restricted to simple operators with closed-form expressions, as in the examples above. In fact, the operators involved can even be general function approximators, such as deep neural networks.

Example 1.4 (Adversarial Examples)**.** *We now consider the problem of certifying the robustness of a trained machine learning model* h *to adversarial perturbations (Szegedy et al., 2013). In particular, let* $\mathcal{T}_{\text{adv}}$ *denote the adversarial attack that maps an input* X *to its adversarially perturbed version* $\widetilde{X}$*. Furthermore, let* $\mathcal{T}_h$ *denote the output of a specific layer (for example, a bottleneck layer) of the model. Then, our goal is to decide if the distributions of* $Y = \mathcal{T}_h(X)$ *and* $\widetilde{Y} = \mathcal{T}_h(\widetilde{X})$ *are equal or not. In other words, the null states that the distribution of* X *after applying* $\mathcal{T}_h$ *and* $\mathcal{T}_h \circ \mathcal{T}_{\text{adv}}$ *is the same.*

Example 1.5 (Independence Testing)**.** *Independence testing is another well-studied problem in statistics, where given observations $\{(X_i, Y_i) : 1 \leq i \leq n\}$ drawn i.i.d. from a distribution P_{XY} on a product space $\mathcal{X} \times \mathcal{Y}$, we want to test whether $P_{XY} = P_X \times P_Y$ or not. By working with two pairs of observations at a time, we can again describe the null as being invariant to an operator. In particular, let given $Z_1 = (X_1, Y_1)$ and $Z_2 = (X_2, Y_2)$, let $\mathcal{T}$ denote the operator that maps (Z_1, Z_2) to (Z_1', Z_2'), with $Z_1' = (X_1, Y_2)$ and $Z_2' = (X_2, Y_1)$. Clearly, the distribution of (Z_1, Z_2) is the same as that of (Z_1', Z_2') under the null, while this invariance to $\mathcal{T}$ is broken under the alternative.*

Example 1.6 (Symmetry Testing)**.** *In the simplest version of this problem, we consider real-valued observations (that is, $\mathcal{Z} = \mathbb{R}$), and the operators $\mathcal{T}_2 = \mathcal{T}_{id}$, and $\mathcal{T}_1 = \mathcal{T}_{flip}$, where the operator $\mathcal{T}_{flip}$ simply flips the observations about the origin; that is, $\mathcal{T}_{flip}(x) = -x$. The resulting null hypothesis asserts that P_Z is symmetric about the origin. The same formulation also covers other kinds of symmetry, such as rotational invariance or invariance to horizontal or vertical flips in the case of images.*

Example 1.7 (Group Invariance Testing)**.** *Suppose we have a collection of images Z containing equilateral triangles. Each edge of these triangles is colored either blue or green. A practitioner would like to find out if the edges of the triangles are colored without any particular pattern, or if some hidden rule controls their coloring. To do this, we will examine whether the triangles remain the same when rotated 120 or 240 degrees. We will therefore introduce a set of operators that represent the aforementioned rotations: $\mathcal{T}_{120}$ and $\mathcal{T}_{240}$. Next, we formulate the following composite null hypothesis:*

$$H_0 : \mathcal{T}_{120}(Z) = Z \text{ and } \mathcal{T}_{240}(Z) = Z$$

Thus, we want to test whether the distribution of Z remains invariant w.r.t. the two operators $\mathcal{T}_{120}$ and $\mathcal{T}_{240}$.

2 Proposed Test

We will use the exact same design idea that we used for two-sample testing for this more general problem. For simplicity, let us consider a simplified version of (1) in which $\mathcal{T}_2 = \mathcal{T}_{\text{id}}$, and we will drop the subscript from $\mathcal{T}_1$ (and assume $\mathcal{W} = \mathcal{Z}$):

$$H_0 : \mathcal{T}(Z) \stackrel{d}{=} Z \quad \text{versus} \quad H_1 : \mathcal{T}(Z) \stackrel{d}{\neq} Z. \tag{2}$$

Since this is essentially a two-sample testing problem in disguise, our general template from the previous lecture carries over to this case almost directly. In particular, we proceed as follows:

- Choose a function class $\mathcal{G} \subset \{g : \mathcal{X} \to [-1/2, 1/2]\}$, and define a notion of distance

$$D_{\mathcal{G}}(P_Z, \mathcal{T}) = \sup_{g \in \mathcal{G}} \mathbb{E}_{P_Z}[g(Z) - g(\mathcal{T}(Z))].$$

  As before, we assume that $\mathcal{G}$ is rich enough to ensure that whenever $\mathcal{T}(Z) \stackrel{d}{\neq} Z$, the distance defined above is strictly positive.

- For a given distribution P_Z, and operator $\mathcal{T}$, we can define the "witness function" $g^* \equiv g^*(P_Z, \mathcal{T})$ as

$$g^* \equiv g^*(P_Z, \mathcal{T}) \in \operatorname*{argmax}_{g \in \mathcal{G}} \mathbb{E}_{P_Z}[g(Z) - g(\mathcal{T}(Z))].$$

- A natural definition of "oracle test" follows

$$\tau^* = \inf\{n \geq 1 : W_n^* \geq 1/\alpha\}, \quad \text{where}$$
$$W_0^* = 1,\ W_n^* = W_{n-1}^* \times (1 + \lambda^*(g^*(Z) - g^*(\mathcal{T}(Z)))),$$

  and $\lambda^* = \operatorname{argmax}_{\lambda \in [-0.5, 0.5]} \mathbb{E}_{P_Z}\left[\log\left(1 + \lambda(g(Z) - g(\mathcal{T}(Z)))\right)\right]$.

- Clearly the oracle test cannot be implemented in practice. So we replace the population terms with their data driven versions, using a *betting strategy* $\mathcal{A}_b$ (for choosing the bets $\{\lambda_n : n \geq 1\}$) and a *prediction strategy* $\mathcal{A}_p$ (for selecting $\{g_n : n \geq 1\}$), which gives us

$$\tau = \inf\{n \geq 1 : W_n \geq 1/\alpha\}, \quad \text{with} \quad W_0 = 1,\ W_n = W_{n-1} \times (1 + \lambda_n V_n), \tag{3}$$

  where $V_n = g_n(Z_n) - g_n(\mathcal{T}(Z_n))$.

It is easy to show, as in our previous lecture, that the test τ defined above is level-α for all betting and prediction strategies. Furthermore, as in the previous lecture, the power of this sequential test can again be characterized in terms of the regret of the prediction strategy $\mathcal{A}_p$, defined as

$$\mathcal{R}_n \equiv \mathcal{R}_n\left(\mathcal{A}_p, \mathcal{G}, \mathcal{T}, Z_1^n\right) \coloneqq \sup_{g \in \mathcal{G}} \sum_{t=1}^{n} \left(\big(g(Z_t) - g(\mathcal{T}Z_t)\big) - \big(g_t(Z_t) - g_t(\mathcal{T}Z_t)\big) \right).$$

Corollary 2.1. *For the hypothesis testing problem defined in* (2)*, let $\tau \equiv \tau(\mathcal{A}_b, \mathcal{A}_p)$ denote the sequential test introduced in* (3) *with function class $\mathcal{G}$. Then, we have the following:*

- *For any prediction strategy $\mathcal{A}_p$, the test $\tau(\mathcal{A}_b, \mathcal{A}_p)$ controls the type-I error uniformly over the null; that is, $\sup_{P \in \mathcal{P}_{\text{null}}} \mathbb{P}_P(\tau < \infty) \leq \alpha$.*
- *If $\mathcal{A}_p$ ensures that $\limsup_{n \to \infty} \frac{\mathcal{R}_n\left(\mathcal{A}_p, \mathcal{G}, \mathcal{T}, Z_1^n\right)}{n} < D_{\mathcal{G}}\left(P, P \circ \mathcal{T}^{-1}\right)$ almost surely under the alternative, then the test τ is consistent. That is, $\mathbb{P}_P\left(\tau < \infty\right) = 1$ for all $P \in \mathcal{P}_{\text{alt}}$.*
- *If $\mathcal{A}_p$ ensures that there exists a sequence $\{r_n : n \geq 1\}$ with $r_n \to 0$, and events $E_n = \{\mathcal{R}_n / n \leq r_n\}$ with $\sum_{n=1}^{\infty} \mathbb{P}(E_n^c) < \infty$, then the expected stopping time satisfies the upper bound $\mathbb{E}_{H_1}[\tau] = \mathcal{O}\left(\frac{\log(1/\alpha\Delta)}{\Delta^2}\right)$, where $\Delta = D_{\mathcal{G}}(P_Z, \mathcal{T})$.*

The proof follows the exact same steps as in the analysis of the two-sample test in previous lecture, and we omit the details.

While Corollary 2.1 identifies sufficient conditions for the consistency of the test τ, it is non-constructive in nature. We now analyze the properties of our test τ initialized with a natural prediction strategy, called the empirical risk minimization (ERM) strategy.

Definition 2.2 (ERM strategy)**.** For a stream of observations $\{Z_t : t \geq 1\}$, the ERM prediction strategy, $\mathcal{A}_{\mathrm{ERM}}$, selects $\{g_t \equiv g_t(Z_1^{t-1}) : t \geq 1\}$ as follows:

$$g_t \in \operatorname*{argmax}_{g \in \mathcal{G}} \frac{1}{t-1} \sum_{i=1}^{t-1} g(Z_i) - g(\mathcal{T} Z_i), \quad \text{for all } t \geq 2,$$

and at $t = 1$, $\mathcal{A}_{\mathrm{ERM}}$ sets g_1 to be an arbitrary element of $\mathcal{G}$.

We will analyze the performance of our test $\tau(\mathcal{A}_b, \mathcal{A}_{\mathrm{ERM}})$ under certain assumptions on the richness of the function class $\mathcal{G}$. A suitable measure of complexity is the Rademacher complexity, whose definition we recall next.

Definition 2.3. Consider a function class $\mathcal{H}$ containing mappings from some observations space $\mathcal{Z}$ to $\mathbb{R}$, and let $P \in \mathcal{P}(\mathcal{Z})$ denote a probability distribution on $\mathcal{Z}$. For a natural number $n \geq 1$, let $\boldsymbol{\sigma_n} = (\sigma_1, \ldots, \sigma_n)$ denote a random vector distributed uniformly over $\{-1, +1\}^n$. Then, given $Z_1, Z_2, \ldots, Z_n$ drawn i.i.d. from P, introduce the the following complexity terms:

$$C_n(\mathcal{H}, P) \coloneqq \frac{1}{n} \mathbb{E}\left[\sup_{h \in \mathcal{H}} \sum_{t=1}^{n} h(Z_t)\sigma_t\right], \quad \text{and} \quad C_n(\mathcal{H}) \coloneqq \sup_{P \in \mathcal{P}(\mathcal{Z})} C_n(\mathcal{H}, P).$$

Before stating Theorem 2.4, we need to introduce two more terms: the function class $\widetilde{\mathcal{G}}$, and the notion of Δ-separated alternatives, $\mathcal{P}_{\mathrm{alt}}(\Delta)$.

$$\widetilde{\mathcal{G}} \coloneqq \{\widetilde{g}(\cdot) = g(\cdot) - g(\mathcal{T}\cdot),\ g \in \mathcal{G}\}, \quad \text{and} \quad \mathcal{P}_{\mathrm{alt}}(\Delta) = \{P \in \mathcal{P}_{\mathrm{alt}} : D_{\mathcal{G}}(P, P \circ \mathcal{T}^{-1}) > \Delta)\}.$$

We now present the main result of this section, that relates the consistency and detection boundary of $\tau(\mathcal{A}_b, \mathcal{A}_{\mathrm{ERM}})$ to the complexity of the function class $\widetilde{\mathcal{G}}$.

Theorem 2.4. *For the sequential test $\tau \equiv \tau(\mathcal{A}_b, \mathcal{A}_{\mathrm{ERM}})$ for the testing problem of* (2)*, with prediction strategy $\mathcal{A}_{\mathrm{ERM}}$ introduced in Theorem 2.2, we have the following:*

- *τ is consistent against any $P \in \mathcal{P}_{\mathrm{alt}}$, for which $C_n(\widetilde{\mathcal{G}}, P)$ converges to 0, that is,*

$$\lim_{n \to \infty} C_n(\widetilde{\mathcal{G}}, P) \to 0 \quad \Longrightarrow \quad \mathbb{P}_P(\tau < \infty) = 1.$$

- *Suppose $C_n(\widetilde{\mathcal{G}})$ converges to 0 with n, and for a small $\gamma \in (0, 1)$, introduce the term*

$$\Delta_n^* = \sqrt{\frac{8 \log n/\alpha}{n}} + \frac{2}{n}\left(2 + \sum_{t=1}^{n-1}\left(C_t(\widetilde{\mathcal{G}}) + 5\sqrt{\frac{\log(16n/\gamma)}{2t}}\right)\right) + \sqrt{\frac{8\log(4/\gamma)}{n}}.$$

Then, for any $n \geq 1$, and $\Delta_n > \Delta_n^*$,

$$\sup_{P \in \mathcal{P}_{\mathrm{alt}}(\Delta_n)} \mathbb{P}_P(\tau > n) \leq \gamma.$$

In other words, Δ_n^* denotes the minimum separation that can be detected with power greater than $1 - \gamma$ by our sequential test within the first n observations.

The proof of this statement is given in Shekhar and Ramdas (2023). The above result implies that the detection boundary for our test in terms of the $D_{\mathcal{G}}$ distance measure is given by $\Delta_n^* = \Omega\left(\frac{1}{n}\sum_{t=1}^{n-1} C_t(\widetilde{\mathcal{G}}) + \sqrt{\frac{\log(n/\alpha)}{n}} + \sqrt{\frac{\log(n/\gamma)}{n}}\right)$, where α and γ correspond to the type-I and type-II errors. For the bounded-mean test we considered in Lectures 18-19, and the kernel-MMD test from the previous lecture, it is known that $C_t(\widetilde{\mathcal{G}})$ decays to zero at a $1/\sqrt{t}$ rate. Hence, for both these tests, we have $\Delta_n^* = \Omega\left(\sqrt{\frac{\log(n/\alpha)}{n}} + \sqrt{\frac{\log(n/\gamma)}{n}}\right)$.

Testing for independence. In this case, we have two observation spaces $\mathcal{X}$ and $\mathcal{Y}$, not necessarily the same, and define $\mathcal{Z} = (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})$. Let P_{XY} denote a distribution in $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and let $P_X \in \mathcal{P}(\mathcal{X})$ and $P_Y \in \mathcal{P}(\mathcal{Y})$ denote its marginals. Under the null hypothesis, we have $P_{XY} = P_X \times P_Y$, which can be encoded via the operator $\mathcal{T} : \mathcal{Z} \to \mathcal{Z}$, with $\mathcal{T}((x, y), (x', y')) = ((x, y'), (x', y))$.

When $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, we can again select $\mathcal{G} = \{g(x) = \mathbf{1}_{x \le u} : u \in \mathbb{R}\}$, which leads to $D_{\mathcal{G}}$ being the KS distance between P_{XY} and the product of its marginals $P_X \times P_Y$. For general $\mathcal{X} \neq \mathcal{Y}$, a suitable choice of $\mathcal{G}$ is a norm ball in the RKHS of the product kernel $K((x, y), (x', y')) := K_X(x, x')K_Y(y, y')$ for positive definite kernels $K_X : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. In this case, the distance $D_{\mathcal{G}}$ is the kernel-MMD distance between P_{XY} and $P_X \times P_Y$; also called the HSIC criterion (Gretton et al., 2005). In both cases, Theorem 2.4 implies that the test is ERM strategy is consistent and, furthermore, has a detection boundary of the order $\mathcal{O}(\sqrt{\log n/n})$ in their respective distance metrics.

3 Conclusion

This concludes our module on designing sequential power-one tests by using ideas from universal portfolios and online learning. All the tests we designed had the general form $\tau = \inf\{n \ge 1 : W_n \ge 1/\alpha\}$, and hence the task reduces to constructing appropriate stochastic processes $\{W_n : n \ge 1\}$. In the simplest case (finite alphabet), we can use likelihood functions, but in more general nonparametric problems, we developed alternative techniques, relying on the dual representations of mean-constrained divergence (bounded mean), and using integral probability metrics (two-sample testing and beyond).

References

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

T. Pandeva, P. Forré, A. Ramdas, and S. Shekhar. Deep anytime-valid hypothesis testing. In *International Conference on Artificial Intelligence and Statistics*, pages 622–630. PMLR, 2024.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020.

S. Shekhar and A. Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2023.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.