ECE 598 - Fall 2025

Lecture 17: Designing Sequential Tests: Part II

28th October, 2025

Instructor: Shubhanshu Shekhar

We continue our discussion of designing sequential anytime-valid tests for observations lying in finite alphabets. In the previous lecture, we designed and analyzed tests for three increasingly complex settings: point null and alternatives, point-vs-composite, and the most general case of composite null versus composite alternatives. For the first two settings, our tests achieve nonasymptotic type-I error control and near-optimal expected stopping time under the alternative (as we will establish in later lectures). For the case of composite nulls, we constructed a simple but suboptimal test. Additionally, all our results made a strong assumption of bounded log-likelihood ratios, which can often be too restrictive. In this lecture, we will address these two issues.

First, we will propose and analyze an improved test for the composite-null version of the problem using ideas from Wasserman et al. (2020, Section 8). Then, we will look at the abstract problem of analyzing the expected boundary crossing time of a random walk under weaker assumptions (than bounded increments), and show that we can essentially obtain the same bound.

1 Optimal Test for Composite Nulls

As in the previous lecture, we assume that $\mathcal{X} = \{x_1, \dots, x_m\}$ is a finite alphabet, and $\{\mathbb{P}_\theta = P_\theta^\infty : \theta \in \Delta_m\}$ denotes the set of all i.i.d. distributions on $\mathcal{X}^\infty$. In our notation, P_θ denotes a probability distribution on $\mathcal{X}$ with probability mass function (pmf) $\theta \in \Delta_m$. With this notation, consider the following problem: Given $\{X_n : n \geq 1\} \sim \mathbb{P}_\theta$ and a confidence parameter $\alpha \in (0, 1]$, construct a level-α power-one test to decide between

$$H_0 : \theta \in \Theta_0, \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \tag{1}$$

for two disjoint subsets Θ_0, Θ_1 of the $(m-1)$-dimensional simplex Δ_m. Throughout this lecture, we will assume that Θ_0 is compact and convex.

Notation. For any pmf $\theta \in \Delta_m$, and a sequence $x^n = (x_1, \dots, x_n)$, we will use the shorthand $\theta^n[x^n]$ to represent $\prod_{i=1}^n \theta[x_i]$. For any $\theta \in \Theta_1$, we define $\gamma_\theta(\Theta_0)$ (we will omit the Θ_0 dependence when it is clear from context) as the divergence to the null set; that is,

$$\gamma_\theta = \inf_{\theta_0 \in \Theta_0} D_{\mathrm{KL}}(P_\theta \parallel P_{\theta_0}), \quad \text{and} \quad \theta_0^* \equiv \theta_0^*(\theta) \in \operatorname*{argmin}_{\theta_0 \in \Theta_0} D_{\mathrm{KL}}(P_\theta \parallel P_{\theta_0}).$$

If Θ_0 is a closed subset of Δ_m, then γ_θ is well defined, and for the moment, we will not worry about the existence and properties of $\theta_0^*(\theta)$.

Now, our goal is to construct a level-α stopping time whose expectation under the alternative depends on the term γ_θ. We constructed a suboptimal test in the previous lecture for which we showed that

$$\mathbb{E}_\theta[\tau] = \mathcal{O}\left(\frac{\log(1/\alpha)}{\bar{\gamma}_\theta}\right), \quad \text{where} \quad \bar{\gamma}_\theta = \sum_{x \in \mathcal{X}} \theta[x] \log\left(\frac{\theta[x]}{\bar{\theta}_0[x]}\right), \text{ for } \bar{\theta}_0[x] = \sup_{\theta_0 \in \Theta_0} \theta_0[x].$$

Construction of the improved test. The idea behind this test is simple: instead of using $\bar{\theta}_0$ that assigns the largest probability among all pmfs in Θ_0 on a per-symbol basis, we replace it with the running maximum likelihood (ML) parameter from the null, which assigns the largest probability on a per-sequence basis. This simple idea, introduced by Wasserman et al. (2020, Section 8), ends up giving us an (as we will show later) optimal test for this problem. Formally, given the stream of observations $\{X_n : n \geq 1\}$, we define our stopping time as

$$\tau = \inf\{n \geq 1 : L_n^{\mathrm{UI}} \geq 1/\alpha\}, \quad \text{with} \quad L_0^{\mathrm{UI}} = 1, \text{ and } L_n^{\mathrm{UI}} = \frac{q_J^n(X^n)}{\sup_{\theta_0 \in \Theta_0} \theta_0[X^n]}. \tag{2}$$

Here, q_J^n denotes the mixture distribution over $\mathcal{X}^n$ w.r.t. Jeffreys prior (i.e., the Krichevsky-Trofimov estimator/predictor), and observe that the denominator is simply the maximum likelihood value based on the first n observations:

$$q_J^n(x^n) = \int_{\Delta_m} \theta^n[x^n] \pi_J(\theta) d\theta, \quad \text{and} \quad \widehat{\theta}_0^n[x^n] = \sup_{\theta_0 \in \Theta_0} \theta_0^n[x^n].$$

Remark 1.1. From our discussion of universal prediction, we know that the mixture q_J^n can be written as

$$q_J^n[x^n] = \prod_{i=1}^n q_J(x_i \mid x^{i-1}), \quad \text{with} \quad q_J(x \mid x^{i-1}) = \frac{1/2 + \sum_{j=1}^{i-1} \mathbf{1}_{x_j = x}}{m/2 + i - 1}.$$

That is, each conditional $q_J(\cdot \mid x^{i-1})$ is the so-called "add-$1/2$" estimator, where we simply add a fictitious count of $1/2$ to each $x \in \mathcal{X}$ to smooth out our estimate. Thus, the numerator in (2) can be updated in an incremental and computationally efficient manner.

However, the denominator will require recomputing the maximum likelihood value from scratch in each round for most problems. Thus, compared to our previous tests, this particular scheme can end up being significantly more computationally demanding.

We now proceed to the analysis of the test defined in (2).

Theorem 1.2. *For the testing problem with composite null and composite alternative* (1)*, the test defined in* (2) *satisfies the following:*

$$\sup_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}(\tau < \infty) \leq \alpha, \quad \textit{and} \quad \mathbb{E}_\theta[\tau] = \frac{\log(1/\alpha) + \frac{m-1}{2} \log\left(\frac{\log(1/\alpha)}{\gamma_{\theta_1}}\right)}{\gamma_\theta} (1 + o(1)),$$

for all $\theta_1 \in \Theta_1$ such that $\gamma_{\theta_1} > 0$. In particular, the finiteness of the expected stopping times also implies the weaker power-one property.

An important fact about the process $\{L_n^{\mathrm{UI}} : n \geq 0\}$ is that it is neither a martingale nor a supermartingale. Yet, as we show in the proof in the next subsection, we can still employ Ville's inequality on this process to control the type-I error. This is because, for every null $\theta_0 \in \Theta_0$, we can show that there exists a nonnegative supermartingale (or actually a nonnegative martingale) that dominates the process $\{L_n^{\mathrm{UI}} : n \geq 0\}$ pointwise. This is one of the defining properties of *e-processes* (Ramdas et al., 2020, Lemma 6), which are fundamental tools in sequential anytime-valid inference.

1.1 Proof of Theorem 1.2

Level-α property. The proof of the level-α property of the test τ follows exactly the same argument as in our previous lecture. In particular, observe that under the null, if $\theta_0 \in \Theta_0$ is the true parameter, then

$$L_n^{\mathrm{UI}} = \frac{q_J^n(X^n)}{\widehat{\theta}_0^n[X^n]} = \frac{q_J^n(X^n)}{\sup_{\theta \in \Theta_0} \theta^n[X^n]} \overset{a.s.}{\leq} \frac{q_J^n(X^n)}{\theta_0^n[X^n]} =: L_n^{\theta_0}.$$

It is easy to verify that $\{L_n^{\theta_0} : n \geq 0\}$ is a nonnegative supermartingale with an initial value of $L_0^{\theta_0} = 1$, which implies that

$$\mathbb{P}_{\theta_0}(\tau < \infty) = \mathbb{P}_{\theta_0}\left(\exists n \geq 1 : L_n^{\mathrm{UI}} \geq 1/\alpha\right) \leq \mathbb{P}_{\theta_0}\left(\exists n \geq 1 : L_n^{\theta_0} \geq 1/\alpha\right) \leq \alpha,$$

where the last inequality follows from an application of Ville's inequality. Taking a supremum over all $\theta_0 \in \Theta_0$ gives us the first part of Theorem 1.2.

Expected Stopping Time. To study the expected stopping time property, consider the alternative with the true parameter $\theta_1 \in \Theta_1$. By the regret of the universal compression scheme, we know that

$$S_n := \log L_n^{\mathrm{UI}} \geq \sup_{\theta \in \Delta_m} \log \theta^n[X^n] - \log \widehat{\theta}_0^n[X^n] - r_n, \quad \text{with} \quad r_n = \frac{m-1}{2} \log n + C_m.$$

Let $\widehat{p}_n$ denote the empirical distribution based on X^n. Observe that for any $\theta \in \Delta_m$, we have

$$\log \theta^n[X^n] = n \sum_{x \in \mathcal{X}} \widehat{p}_n[x] \log \theta[x] = n \left(-H(\widehat{p}_n) - D_{\mathrm{KL}}(\widehat{p}_n \parallel \theta)\right).$$

By the nonnegativity of relative entropy, this implies that $\widehat{p}_n$ is also the MLE (over all Δ_m) based on X^n, which means that

$$S_n \geq -nH(\widehat{p}_n) - \log \widehat{\theta}_0^n[X^n] - r_n. \tag{3}$$

Now, we analyze the second term above to see that

$$\begin{aligned} \log \sup_{\theta_0 \in \Theta_0} \theta_0^n[X^n] = n \sum_{x \in \mathcal{X}} \widehat{p}_n \log \theta_0[x] &= \sup_{\theta_0 \in \Theta_0} -n \left(H(\widehat{p}_n) + D_{\mathrm{KL}}(\widehat{p}_n \parallel \theta_0)\right) \\ &= -nH(\widehat{p}_n) - nD_{\mathrm{KL}}(\widehat{p}_n \parallel \Theta_0), \end{aligned} \tag{4}$$

where we define $D_{\mathrm{KL}}(\widehat{p}_n \parallel \Theta_0)$ as $\inf_{\theta_0 \in \Theta_0} D_{\mathrm{KL}}(\widehat{p}_n \parallel \theta_0)$. On combining (3) and (4), we get

$$S_n \geq nD_{\mathrm{KL}}(\widehat{p}_n \parallel \Theta_0) - r_n. \tag{5}$$

Now, we recall the Donsker-Varadhan (DV) variational representation (Section 4 of Lecture 3) of relative entropy, to observe that

$$D_{\mathrm{KL}}(p \parallel \Theta_0) = \inf_{\theta_0 \in \Theta_0} \sup_{f : \mathcal{X} \to \mathbb{R}} \left(\mathbb{E}_p[f] - \log E_{\theta_0}[e^f]\right) =: \inf_{\theta_0 \in \Theta_0} \sup_{f : \mathcal{X} \to \mathbb{R}} \Phi(f, \theta_0).$$

Now, observe that the function Φ is concave in f (due to the convexity of log-sum-exp function), and convex in θ (due to the concavity of $\log$ function). We have assumed that the null class Θ_0 is closed and convex and that $\mathbb{R}^{\mathcal{X}}$ is closed. Hence, the conditions for applying the minimax theorem are satisfied, and we have

$$D_{\mathrm{KL}}(p \parallel \Theta_0) = \sup_{f:\mathcal{X}\to\mathbb{R}} \inf_{\theta_0\in\Theta_0} \mathbb{E}_p[f] - \log \mathbb{E}_{\theta_0}[e^f] = \sup_{f:\mathcal{X}\to\mathbb{R}} \mathbb{E}_p[f] - \psi_0(f),$$

where $\psi_0(f) \coloneqq \sup_{\theta_0\in\Theta_0} \log \mathbb{E}_{\theta_0}[e^f]$. Now, let us fix an arbitrary (for now) $f : \mathcal{X} \to \mathbb{R}$, which gives a lower bound in the equality above, and using this with (5), we get

$$S_n \geq nD_{\mathrm{KL}}(\widehat{p}_n \parallel \Theta_0) \geq \mathbb{E}_{\widehat{p}_n}[f] - \psi_0(f) - r_n = \underbrace{\sum_{i=1}^n f(X_i) - n\psi_0(f)}_{\coloneqq S_n(f)} - r_n.$$

Thus, the stopping time $T_f = \inf\{n \geq 1 : S_n(f) \geq \log(1/\alpha) + r_n\}$ is an upper bound on $\tau = \inf\{n \geq 1 : S_n \geq \log(1/\alpha)\}$. Now, suppose we choose f to be the element

$$f \in \operatorname*{argmax}_{f':\mathcal{X}\to\mathbb{R}} \mathbb{E}_{\theta_1}[f'] - \psi_0(f') \quad \Longrightarrow \quad \mathbb{E}_{\theta_1}[f] - \psi_0(f) = D_{\mathrm{KL}}(\theta_1 \parallel \Theta_0) = \gamma_{\theta_1}.$$

We now note the following with the above choice of f:

- The stopping time T_f is finite almost surely. This is because

  $$\lim_{n\to\infty} \frac{1}{n} S_n(f) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n f(X_i) - \psi_0(f) \overset{\text{a.s.}}{=} \mathbb{E}_{\theta_1}[f(X)] - \psi_0(f) = \gamma_{\theta_1} > 0,$$

  where we used the assumption on f in the last inequality. On the other hand, since $r_n \asymp \log n$, we have $(\log(1/\alpha) + r_n)/n \to 0$, which means that $S_n(f)$ eventually crosses the boundary $\log(1/\alpha) + r_n$ with probability 1; that is, $T_f < \infty$ almost surely.

- Since $f(x) = \log \theta_1[x]/\theta^\dagger[x]$ and $\theta^\dagger[x] > 0$ for all x such that $\theta_1[x] > 0$, we can observe that

  $$f(x) - \psi_0(f) \leq f(x) - \sup_{\theta_0\in\Theta_0} \log \sum_{y\in\mathcal{X}} \theta_0[y] e^{f(y)} \leq f(x) - \log \sum_{y\in\mathcal{X}} \theta^\dagger[y] \times \frac{\theta_1[y]}{\theta^\dagger[y]} = f(x).$$

  Hence, the one-step increment of $S_n(f)$ is upper bounded by $C_f \coloneqq \max_{x\in\mathcal{X}} \log \theta_1[x]/\theta^\dagger[x]$.

Together, these two facts imply that we can use Wald's identity to conclude that

$$\mathbb{E}_{\theta_1}[S_{T_f}(f)] = \mathbb{E}_{\theta_1}[T_f]\mathbb{E}_{\theta_1}[f(X) - \psi_0(f)] \leq \log(1/\alpha) + \frac{m-1}{2} \log \mathbb{E}_{\theta_1}[T_f] + C_m + C_f.$$

Let γ_{θ_1} denote $\mathbb{E}_{\theta_1}[f(X) - \psi_0(f)]$, and observe that with $y = \mathbb{E}_{\theta_1}[T_f]$, we have

$$y\gamma_{\theta_1} \leq \log(1/\alpha) + \frac{m-1}{2} \log y + C_m + C_f.$$

By some standard calculations (see Appendix A) we can show that

$$\begin{aligned}\mathbb{E}_{\theta_1}[T_f] &\leq \frac{\log(1/\alpha) + \frac{m-1}{2}\log(\log(1/\alpha)/\gamma_{\theta_1}) + C_1 + C_m}{\gamma_{\theta_1}} + \frac{m-1}{2\gamma_{\theta_1}}\log^+\left(\frac{m-1}{2\gamma_{\theta_1}}\right) \\ &= \frac{\log(1/\alpha) + \frac{m-1}{2}\log\left(\frac{\log(1/\alpha)}{\gamma_{\theta_1}}\right)}{\gamma_{\theta_1}}(1+o(1)),\end{aligned}$$

where $o(1)$ indicates terms that go to zero as $\alpha \to 0$ (with fixed γ_{θ_1}) or with $\gamma_{\theta_1}) \to 0$ (with fixed α). This completes our proof.

2 Boundary crossing with unbounded increments

So far, we have mostly considered the case of random walks with bounded increments. In many important cases, this simplifying condition is not satisfied. Nevertheless, we can still employ the general outlines discussed above to derive bounds on expected stopping times. We illustrate an elementary, although suboptimal (by a factor of 2) argument in this section, and briefly discuss how this can be improved.

Consider an i.i.d. process $\{Y_i : i \geq 1\}$ and let $S_n = \sum_{i \leq n} Y_i$ denote a random walk. Suppose $\mathbb{E}[|Y_i|] < \infty$ and $\gamma = \mathbb{E}[Y_i] > 0$. Consider a boundary $b_n = b_0 + c\log n$ for $n \geq 1$, and define the stopping time

$$\tau = \inf\{n \geq 1 : S_n \geq b_n\}.$$

The idea is simple: We select a value $a \in [0, \infty)$ whose exact value will be decided later, and work with truncated increments $U_i = \min\{Y_i, a\}$ with mean $\gamma_a = \mathbb{E}[U_i]$. That is, we define a new process $V_n = \sum_{i=1}^n U_i$, and observe that $V_n \leq S_n$ almost surely. This leads to a new stopping time T_a which is almost surely larger than τ;

$$T_a = \inf\{n \geq 1 : V_n \geq b_n = b_0 + c\log n\}.$$

Now, T_a is the first boundary crossing of the process $\{V_n : n \geq 1\}$ with increments that are bounded from above. Hence, we can conclude that, with $y_a = \mathbb{E}[T_a]$, we have

$$y_a\gamma_a \leq b_0 + c\log y_a + a, \quad \implies y_a \leq \frac{b_0 + a}{\gamma_a} + \frac{c}{\gamma_a}\log(y_a).$$

Using the inequality from Appendix A, we can conclude that for all $a > 0$, we have

$$y_a \leq \frac{b_0 + a}{\gamma_a} + \frac{c}{\gamma_a}\left(\log\left(\frac{b_0 + a}{\gamma_a}\right) + 1\right) + \frac{c}{\gamma_a}\log^+\left(\frac{c}{\gamma_a}\right) \asymp \frac{b_0 + a + c\log\left(\frac{b_0+a}{\gamma_a}\right)}{\gamma_a}.$$

Now, it remains for us to select an appropriate a. To do this, note that $\gamma = \mathbb{E}[Y] = \mathbb{E}[U] + \mathbb{E}[(Y - a)^+] = \gamma_a + \mathbb{E}[(Y-a)^+]$. We know that $(Y-a)^+ \overset{a.s.}{\to} 0$ as $a \uparrow \infty$, and hence by monotone convergence theorem, we can conclude $\lim_{a\uparrow\infty}\mathbb{E}[(Y-a)^+] = 0$, which implies $\lim_{a\uparrow\infty}\gamma_a = \gamma$. Hence, we can select a value a_0 such that $\gamma_{a_0} \geq \gamma$, and that gives us

$$\mathbb{E}[\tau] \leq \mathbb{E}[T_{a_0}] \leq 2\frac{b_0 + a_0 + c\log\left(\frac{2(b_0+a_0)}{\gamma}\right)}{\gamma}.$$

Remark 2.1. Thus, we essentially get the same bound as in the case of bounded increments, but with a leading multiplicative factor 2 and an additive term a_0 (quantifying the size of truncated increments. We can make the multiplicative constant 2 go arbitrarily close to 1 from above, at the cost of increasing the additive constant $a_0 \to \infty$.

Remark 2.2. The key issue in deriving upper bounds on expected stopping times is to analyze the behavior of the last increment; that is, $\mathbb{E}[Y_\tau]$. Our derivation in this section relied only on the existence of finite first moments, by using a truncation device. If we further assume that increments have finite second moments, then we can get a cleaner bound on the "overshoot" of the form

$$\mathbb{E}[Y_\tau] \leq \frac{\mathbb{E}[(Y_1^+)^2]}{\mathbb{E}[(Y_1^+)]},$$

which is tighter, and does not lead to an extra multiplicative factor. We will explore this idea in a homework problem.

A A Useful Inequality

For some constants A, B, γ, suppose we have a quantity $y > 0$ such that

$$y \leq \frac{A}{\gamma} + \frac{B}{\gamma} \log(1 + y).$$

Then, we have the following bound on y:

$$y \leq \frac{A}{\gamma} + \frac{B}{\gamma}\left(\log\left(\frac{A}{\gamma}\right) + 1\right) + \frac{B}{\gamma}\log^+\left(\frac{B}{\gamma}\right).$$

References

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.