

## Lecture 18: Designing Sequential Tests: Part III

November 4th, 2025

*Instructor: Shubhangshu Shekhar*

So far our discussion on designing sequential tests has focused on the special case of discrete distributions over finite alphabets. This simplified setting allowed us to focus on the key ideas (constructing test martingales, or  $e$ -processes) without being encumbered by technical details. One key benefit of working with distributions over finite alphabets is that the distributions can be completely specified by their probability mass functions. In the next few lectures, we consider scenarios where this is not the case. In particular, we will work with nonparametric distribution classes which do not admit a common dominating measure to define likelihood functions. As a result, we have to develop different approaches for quantifying the amount of evidence against the null.

In this lecture, we consider one of the simplest nonparametric testing problem: given a stream of bounded observations decide whether their mean is equal to a specified value or not. Despite its simplicity, this problem allows us to illustrate some general ideas to design optimum power one tests that will be applicable in many other problems.

### 1 Problem Definition

Consider the following problem: Given a stream of observations  $\{X_n : n \geq 1\}$  drawn i.i.d. from a distribution  $P_X$  supported on  $[0, 1]$  with an unknown mean  $\mu_X$ , consider the hypothesis

$$H_{0,m} : \mu_X = m, \quad \text{versus} \quad H_{1,m} : \mu_X \neq m.$$

As before, we are interested in constructing a level- $\alpha$  power-one test  $\tau$  such that

$$\mathbb{P}_{H_{0,m}}(\tau < \infty) \leq \alpha, \quad \text{and} \quad \mathbb{P}_{H_{1,m}}(\tau < \infty) = 1.$$

Note that this problem can be equivalently stated as follows

$$H_{0,m} : P_X \in \mathcal{P}_{0,m}, \quad \text{versus} \quad H_{1,m} : P_X \in \mathcal{P}_{1,m},$$

where the two distribution classes  $\mathcal{P}_{j,m}$  for  $j = 0, 1$  are defined as

$$\mathcal{P}_{0,m} = \{P \in \mathcal{P}([0, 1]) : \mathbb{E}_P[X] = m\}, \quad \mathcal{P}_{1,m} = \{P \in \mathcal{P}([0, 1]) : \mathbb{E}_P[X] \neq m\}.$$

As mentioned earlier, the distribution classes involved are both infinite-dimensional and we cannot define densities w.r.t. some common dominating measure for these two classes.

The general idea as before is to define the stopping time as the first  $1/\alpha$  crossing of an  $e$ -process; that is,

$$\tau = \inf\{n \geq 1 : W_n \geq 1/\alpha\}.$$

The main challenge lies in figuring out how to construct such a process  $\{W_n : n \geq 1\}$  that is unlikely to take large values when the null is true, and grows at an exponential rate under the alternative.

## 2 Proposed Power-One Test

In this section, we will propose and analyze a test for the mean testing problem introduced in the previous section. We first discuss a test for a simpler setting in Section 2.1, and then proceed to the general case in Section 2.2, where we draw an interesting connection to a two-stock portfolio.

### 2.1 One-sided Hypotheses

Let us first consider the following simpler problem.

$$H_0 : \mu_X = m, \quad \text{versus} \quad H_1 : \mu_X \geq m + \Delta, \quad (1)$$

for some  $\Delta \in (0, 1 - m)$ . To design a test for this problem, we start with the observation that the random variable  $U = X - m$  satisfies  $\mathbb{E}_{H_0}[U] = 0$ , and  $\mathbb{E}_{H_1}[U] \geq \Delta$ . Thus, for any  $\lambda \in (-1/(1-m), 1/m)$ , we know that

$$\mathbb{E}_{H_0}[1 + \lambda U] = 1, \quad \text{and} \quad \mathbb{E}_{H_1}[1 + \lambda U] \geq 1 + \Delta.$$

This naturally suggests the following construction with  $W_0(\lambda) = 1$ , and

$$W_n(\lambda) = W_{n-1}(\lambda)(1 + \lambda(X_n - m)),$$

which satisfies two crucial properties. The first is that it is a nonnegative martingale under the null:

$$\mathbb{E}_{H_0}[W_n(\lambda) \mid \mathcal{F}_{n-1}] = \mathbb{E}_{H_0}[W_{n-1}(\lambda)(1 + \lambda(X_n - m)) \mid \mathcal{F}_{n-1}] \stackrel{(a.s.)}{=} W_{n-1}(\lambda).$$

The second is that its expectation grows exponentially under the alternative; that is, if  $\mu_X \geq m + \Delta$ , then we have the following due to the independence of the observations:

$$\mathbb{E}_{H_1}[W_n(\lambda)] = \prod_{i=1}^n \mathbb{E}_{H_1}[1 + \lambda(X_i - m)] \geq (1 + \lambda\Delta)^n.$$

Hence, this hints that the process might also grow (a.s.) at an exponential rate under the alternative for any  $\lambda > 0$  (that ensures non-negativity of the process).

*Remark 2.1.* Note that the martingale property of the process constructed above also holds when instead of a fixed  $\lambda$ , we choose a *predictable sequence*  $\{\lambda_n : n \geq 1\}$  where each  $\lambda_n$  is  $\mathcal{F}_{n-1}$  measurable. We will use this fundamental fact to learn the “best”  $\lambda$  in a data-driven manner in the next subsection.

We now analyze the performance of the test.

**Proposition 2.2.** Fix a  $\lambda \in [0, 1/m]$ , and consider the test  $\tau_\lambda = \inf\{n \geq 1 : W_n(\lambda) \geq 1/\alpha\}$  for the one-sided problem stated in (1). Then, this test satisfies the following:

$$\mathbb{P}_{H_0}(\tau_\lambda < \infty) \leq \alpha, \quad \text{and} \quad \mathbb{E}_{H_1}[\tau_\lambda] \leq \frac{\log(1/\alpha) + \log(1/m)}{\gamma(\lambda, P_X)},$$

where the term  $\gamma(\lambda, P_X, m)$  is equal to the following, when  $P_X$  is the true distribution under  $H_1$ :

$$\gamma(\lambda, P_X, m) = \mathbb{E}_{P_X}[\log(1 + \lambda(X - m))].$$

*Remark 2.3.* In the construction of our test, so far we have only placed minimal restrictions on  $\lambda$  (that it should be nonnegative and less than  $1/m$  to ensure the non-negativity of  $W_n(\lambda)$ ). The result above highlights its important role in the performance of the test, and suggests that an optimal choice of  $\lambda$  should be

$$\lambda^* \equiv \lambda^*(P_X, m) = \operatorname{argmax}_{\lambda \in [0, 1/m]} \mathbb{E}_{P_X} [\log(1 + \lambda(X - m))].$$

In the next lecture we will see that the maximization problem above is simply an information projection operation (mean-constrained relative entropy problem) in its dual form!

*Remark 2.4.* Also, note that whenever  $\Delta > 0$ , the term  $\mathbb{E}_{P_X} [\log(1 + \lambda(X - m))]$  is strictly positive for  $\lambda > 0$  under  $H_1$ . This is simply an application of Jensen's inequality which says that

$$\gamma(\lambda, P_X, m) = \mathbb{E}_{P_X} [\log(1 + \lambda(X - m))] \leq \log(1 + \lambda(\mu_X - m)) \geq \log(1 + \lambda\Delta) > 0.$$

### 2.1.1 Proof of Theorem 2.2

**Level  $\alpha$  property.** This is immediate from the fact that  $\{W_n(\lambda) : n \geq 1\}$  is a nonnegative martingale with an initial value of 1, and an application of Ville's inequality.

**Expected stopping time.** We can rewrite the test as  $\tau_\lambda = \inf\{n \geq 1 : \log W_n(\lambda) \geq \log(1/\alpha)\}$ , and let  $L_n = \log(W_n(\lambda))$ . Observe, that for every  $\lambda$ , the one-step increment in  $L_n$ , denoted by  $B_n = \log(1 + \lambda(X_n - m))$  is upper bounded by  $\log(1 + \lambda(1 - m))$ , which can be further upper bounded by  $C_m = \log(1 + (1 - m)/m) = -\log m$ .

Now,  $\tau$  is the first  $\log(1/\alpha)$  crossing of a random walk  $L_n = \sum_{i=1}^n B_i$ , with a positive drift  $\mathbb{E}[B_i] = \gamma \equiv \gamma(\lambda, P_X, m)$ . A simple application of Wald's identity implies that

$$\mathbb{E}_{L_{\tau_\lambda}} = \mathbb{E}[\tau_\lambda]\gamma \leq \log(1/\alpha) + C_m.$$

Dividing both sides by  $\gamma$  gives us the required upper bound.

## 2.2 General Case with a Two-Stock Portfolio Interpretation

As discussed in Theorem 2.3, we know that that optimal choice of  $\lambda^*$  for a given problem depends on the unknown distribution  $P_X$ . However, one option might be to select  $\{\lambda_n : n \geq 1\}$  predictably that learn to approximate the optimal  $\lambda^*$  in a data-driven manner. It turns out that the best way to do this is by interpreting our problem as a two-stock portfolio.

To see this, consider a stream of observations  $\{X_n : n \geq 1\}$ . To each  $X_n$  and a given null  $H_{0,m}$ , let us consider a two-stock portfolio with price relatives  $\{R_n \in [0, \infty)^2 : n \geq 1\}$ , where

$$R_n = [R_{n,1}, R_{n,2}], \quad \text{with} \quad R_{n,1} = 1 + \frac{X_n - m}{m} = \frac{X_n}{m}, \quad R_{n,2} = 1 - \frac{1 - X_n}{1 - m} = \frac{X_n - m}{1 - m}.$$

Now, observe the following:

- If the null is true, then both stocks have  $\mathbb{E}[R_{n,j}] = 1$  for  $j \in \{1, 2\}$ . Thus, no portfolio allocation strategy is expected to make large increases in the wealth of the investor.

- If the two-sided alternative is true, and  $\mu_X > m$ , then we can check that the first stock is profitable in expectation, and an allocation that places larger weight on this stock will lead to an exponential growth. Similarly, if  $\mu_X < m$ , then the optimal allocation places larger weight on the second stock.

Let  $\mathbf{b} = (b, 1 - b)$  denote any portfolio allocation. Then, observe that

$$\begin{aligned}\langle \mathbf{b}, R_n \rangle &= bR_{n,1} + (1 - b)R_{n,2} = b \left(1 + \frac{X_n - m}{m}\right) + (1 - b) \left(1 - \frac{X_n - m}{1 - m}\right) \\ &= b + (1 - b) + \frac{b}{m}(X_n - m) - \frac{1 - b}{1 - m}(X_n - m) \\ &= 1 + \left(\frac{b}{m} - \frac{1 - b}{1 - m}\right)(X_n - m) = 1 + \lambda(X_n - m).\end{aligned}$$

Thus, every allocation  $\mathbf{b}$  represented by  $b \in [0, 1]$  maps injectively to a “ $\lambda$ ” in our test statistic;  $b \rightarrow \lambda = \left(\frac{b}{m} - \frac{1-b}{1-m}\right) \in [-1/(1-m), 1/m]$ .

This simple observation now make all portfolio allocation algorithms available to us for constructing power-one tests. In particular, given a stream of observations  $\{X_n : n \geq 1\}$ , we can define a test as follows, with  $W_0 = 1$  and

$$\tau = \inf\{n \geq 1 : W_n \geq 1/\alpha\},$$

where  $W_n$  denotes the wealth of an investor who starts with \$1, and invests in the two-stock portfolio constructed above using Cover’s Universal portfolio algorithm (mixture method with Beta-prior):

$$W_n = \int_{[-1/(1-m), 1/m]} W_n(\lambda) \pi_J(\lambda) d\lambda.$$

From our discussion on mixture methods, we know that we can translate this into a predictable form using the fact that for  $n \geq 2$ , and with  $Z_n(\lambda) = 1 + \lambda(X_n - m)$ :

$$\begin{aligned}\frac{W_n}{W_{n-1}} &= \frac{\int W_{n-1}(\lambda) Z_n(\lambda) \pi_J(\lambda) d\lambda}{\int W_{n-1}(\lambda') \pi_J(\lambda') d\lambda'} = \int Z_n(\lambda) \left( \frac{W_{n-1}(\lambda) \pi_J(\lambda)}{\int W_{n-1}(\lambda') \pi_J(\lambda') d\lambda'} \right) d\lambda \\ &= \int (1 + \lambda(X_n - m)) \pi_J(\lambda | \mathcal{F}_{n-1}) d\lambda \\ &= 1 + \left( \int \lambda \pi_J(\lambda | \mathcal{F}_{n-1}) d\lambda \right) (X_n - m) = 1 + \lambda_n^J (X_n - m).\end{aligned}$$

Hence, the process  $\{W_n : n \geq 1\}$  consists of  $W_n = \prod_{i=1}^n (1 + \lambda_i(X_i - m))$ , where  $\{\lambda_i : i \geq 1\}$  is a predictable sequence of “investments” or “bets” associated with the mixture strategy with Jeffreys prior (i.e., Cover’s universal portfolio). The benefit of this approach can be understood from the individual sequence regret of mixture methods, restated here in our  $\lambda$  notation.

**Fact 2.5.** *For any sequence  $\{X_n : n \geq 1\}$ , we have the following*

$$\log W_n \geq \sup_{\lambda \in [-1/(1-m), 1/m]} \sum_{i=1}^n \log(1 + \lambda(X_i - m)) - \underbrace{\left( \frac{1}{2} \log n + C \right)}_{:= r_n},$$

for some universal constant  $C > 0$ .

Perhaps more crucially, this fact implies that if for a given distribution  $P_X$  with mean different from  $m$ , we define  $\lambda^*(P_X, m) \in \operatorname{argmax}_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P_X} [\log(1 + \lambda(X - m))]$ , then, we also have

$$\log W_n \geq \sum_{i=1}^n \log(1 + \lambda^*(X_i - m)) - r_n := \sum_{i=1}^n B_i^* - r_n,$$

for all  $n \geq 1$ . This implies that

$$\tau = \inf\{n \geq 1 : \log W_n \geq \log(1/\alpha)\} \leq T := \inf\{n \geq 1 : \sum_{i=1}^n B_i^* \geq \log(1/\alpha) + r_n\}. \quad (2)$$

The stopping time  $T$  is a much easier to analyze using our usual techniques, and we can conclude the following.

**Theorem 2.6.** *Let  $\{X_n : n \geq 1\} \stackrel{i.i.d.}{\sim} P_X \in \mathcal{P}([0, 1])$ , and consider the two-sided testing problem with null  $H_{0,m} : \mu_X = m$ . Then, the test  $\tau = \inf\{W_n : n \geq 1\}$  constructed using the universal portfolio scheme satisfies the following:*

$$\mathbb{E}_{H_{0,m}}(\tau < \infty) \leq \alpha, \quad \text{and} \quad \mathbb{E}_{H_{1,m}}[\tau] \leq \frac{\log(1/\alpha) + \frac{1}{2} \log \left( \frac{\log(1/\alpha)}{\gamma^*(P_X, m)} + \mathcal{O}(1) \right)}{\gamma^*(P_X, m)},$$

where  $\gamma^*(P_X, m)$  is defined as  $\sup_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P_X} [\log(1 + \lambda(X - m))]$ .

*Proof.* For the power-one property, note that  $\{W_n : n \geq 1\}$  is a nonnegative martingale under the null; hence Ville's inequality applies to this case.

For the expected stopping time bound, we simply use  $\mathbb{E}[\tau] \leq \mathbb{E}[T]$ , where both are defined in (2). Since  $T$  is the first boundary crossing of the random walk  $\sum_{i=1}^n B_i^*$ , our analysis from the previous two lectures gives us the stated bound, noting that  $\gamma^* = \mathbb{E}[B_i^*]$ .  $\square$

### 3 Conclusion

For the case of finite alphabet testing, we saw that under the alternative the tests satisfied  $\mathbb{E}[\tau] \lesssim \log(1/\alpha)/\gamma^*$ , where  $\gamma^* = \inf_{Q \in \Theta_0} D_{\text{KL}}(P_X \parallel Q)$  represented the minimum divergence of the distribution  $P_X$  from the null class of distributions indexed by  $\Theta_0$ . Hence, it might be natural to expect a similar expression in the bounded mean testing problem, involving  $\inf_{Q: \mathbb{E}_Q[X] = m} D_{\text{KL}}(P_X \parallel Q)$ . However, Theorem 2.6 presents an upper bound on  $\mathbb{E}_{P_X}[\tau]$  in terms of  $\sup_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P_X} [\log(1 + \lambda(X - m))]$ , which looks different from the minimum divergence term. As we shall see in the next lecture, this is simply the minimum divergence term in disguise by establishing a convex duality relation between the two expressions.