ECE 598 - Fall 2025

Lecture 15: Sequential Inference Problems

21st October, 2025

Instructor: Shubhanshu Shekhar

With this lecture, we begin the discussion of the topic of *sequential anytime-valid inference*. This field of study was initiated by Robbins and collaborators in the 1960s-70s, with key contributions from Lai, Seigmund, Darling, and others. Over the last decade there has been a resurgence of interest in this area, partially driven by the importance of efficient large-scale A/B testing done in the industry (**?**), and partially by some theoretical and conceptual advances in the foundations in statistics and probability (**?**). A summary of the recent advances in this area can be found in the survey paper by **?**.

In this lecture, we will introduce the three main inference tasks within this framework: power-one testing, estimation via confidence sequences, and sequential change detection. We will see a duality between the first two tasks, and then discuss a result by **?** that establishes a reduction from the problem of change detection to power-one testing.

1 Motivation for Designing Sequential Procedures

In a typical fixed sample size inference problem, we assume that we have access to a dataset $X^n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ drawn i.i.d. (for simplicity) from a distribution P_θ with $\theta \in \Theta$, and our goal is to make a decision about the unknown parameter θ. Here the sample-size n is a quantity that is decided before the experiment or data-collection. In contrast, for sequential procedures, the sample size itself is a random variable, and the process of data-collection and inference are tightly coupled. The motivation of designing sequential procedures are usually for two reasons: (i) sequential methods can lead to some benefits over their fixed sample size (FSS) counterparts, and (ii) there exist problems which can only be solved sequentially. We illustrate both situations through examples next.

Example 1. Suppose $\mathcal{X} = \{0,1\}$ and $P_\theta = \mathrm{Bernoulli}(\theta)$ with $\theta \in \Theta = [0,1]$. Consider the hypothesis testing problem

$$H_0 : \theta \in \Theta_0 = [0, \theta_0], \quad \text{versus} \quad H_1 : \theta \in \Theta_1 = (\theta_0, 1].$$

In the FSS setting, we have $X^n \overset{i.i.d.}{\sim} P_\theta$, and we wish to design a deterministic test $\Psi : \mathcal{X}^n \to \{0,1\}$ which satisfies $\mathbb{P}_{H_0}(\Psi(X^n) = 1) \leq \alpha$ for a prespecified α. A natural test in this setting is the likelihood ratio test that takes the form

$$\Psi(x^n) = \begin{cases} 1, & \text{if } S_n = \sum_{i=1}^n X_i \geq t_\alpha, \\ 0, & \text{otherwise,} \end{cases}$$

where $t_\alpha = \min\{m \in \{0, \ldots, n\} : \mathbb{P}_{\theta_0}(S_n \geq m) \leq \alpha\}$. Now, suppose that instead of working with the entire dataset X^n at once, we observed the samples one at a time, and we define a stopping time τ and a hypothesis test Ψ_τ as

$$\tau = \min\{i \leq n : S_i \geq t_\alpha\}, \quad \Psi_\tau = \begin{cases} 1, & \text{if } \tau \leq n, \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to observe that

$$\mathbb{P}_{H_0}(\Psi_\tau = 1) = \mathbb{P}_{H_0}(\Psi(X^n) = 1), \quad \text{and} \quad \tau \overset{a.s.}{\leq} n,$$

by construction. Hence, Ψ_τ has the same significance level as the FSS test Ψ, but it requires a random number $\tau \leq n$ of observations; that is, it allows "early stopping". Thus, the sequential procedure Ψ_τ is a strict improvement over its FSS analog Ψ. In general, sequential methods do not result in such "pathwise" improvement; this is a special case due to the monotonicity of S_n and the one-sided nature of the testing problem.

Example 2: Fixed-width estimation with unknown variance. Next, we consider a simple problem that clearly illustrates a situation in which non-sequential solutions are impossible. Suppose $X^n = (X_1, \ldots, X_n)$ is drawn i.i.d. from $N(\theta, \sigma^2)$ with both (θ, σ^2) unknown, and fix an $\epsilon > 0$ and $\alpha \in (0, 1)$. Then, does their exist an $n \geq 1$ and an estimator $\widehat{\mu}_n \equiv \widehat{\mu}_n(X^n)$, such that we have

$$\inf_{\mu \in \mathbb{R}, \sigma^2 > 0} \mathbb{P}_{\mu, \sigma^2}\left(\mu \in [\widehat{\mu}_n - \epsilon, \widehat{\mu}_n + \epsilon]\right) \geq 1 - \alpha? \tag{1}$$

In other words, can we construct a "fixed-width" estimate of Gaussian mean with unknown variance that is uniformly valid at level $1 - \alpha$? It turns out that it is not too difficult to show that such an estimator cannot exist (**?**), and we can establish this with a simple application of LeCam's two-point method due to the fact that σ^2 can be arbitrarily large.

Proposition 1.1. *For a fixed $\alpha \in (0, 1/2)$ and $\epsilon > 0$, there exists no n and estimator $\widehat{\mu}_n$ for which* (1) *holds.*

Proof. Suppose there exists an $n \geq 1$ and an estimator $\widehat{\mu}_n$ for which (1) holds. Now, consider two mean values $\mu_1 = -2\epsilon$ and $\mu_2 = 2\epsilon$, and define the sets

$$E_j = \{x^n \in \mathcal{X}^n : |\widehat{\mu}_n(x^n) - \mu_j| \leq \epsilon\}, \quad \text{for} \quad j = 1, 2.$$

Then, by (1), we know that

$$\inf_{\sigma^2 > 0} \mathbb{P}_{\mu_j, \sigma^2}(X^n \in E_j) \geq 1 - \alpha, \quad \text{for} \quad j = 1, 2.$$

Note that since $|\mu_1 - \mu_2| = 4\epsilon$, E_1 and E_2 are disjoint subsets of $\mathcal{X}^n$. Hence, $E_2 \subset E_1^c$, which leads to the following

$$\begin{aligned} \mathbb{P}_{\mu_1, \sigma^2}(X^n \in E_1) - \mathbb{P}_{\mu_2, \sigma^2}(X^n \in E_1) &= \mathbb{P}_{\mu_1, \sigma^2}(X^n \in E_1) - 1 + \mathbb{P}_{\mu_2, \sigma^2}(X^n \in E_1^c) \\ &\geq \mathbb{P}_{\mu_1, \sigma^2}(X^n \in E_1) - 1 + \mathbb{P}_{\mu_2, \sigma^2}(X^n \in E_2). \end{aligned}$$

On taking 1 to the other side, and by using the definition of total variation and (1), we get

$$TV(\mathbb{P}_{\mu_1,\sigma^2}, \mathbb{P}_{\mu_2,\sigma^2}) + 1 \geq \mathbb{P}_{\mu_1,\sigma^2}(X^n \in E_1) + \mathbb{P}_{\mu_2,\sigma^2}(X^n \in E_2) \geq 2 - 2\alpha.$$

Next, by using Pinsker's inequality, we get that $TV(\mathbb{P}_{\mu_1,\sigma^2}, \mathbb{P}_{\mu_2,\sigma^2}) \leq \sqrt{n}\epsilon/\sigma$, which implies

$$\frac{\epsilon}{\sigma}\sqrt{n} \geq 1 - 2\alpha, \quad \text{which is violoated for all} \quad \sigma > \frac{\epsilon\sqrt{n}}{1-2\alpha}.$$

This completes the proof. □

Thus, we cannot construct an estimator which can achieve a fixed width ϵ uniformly over all σ^2 for any finite n: there always will be some problems with large enough σ^2 to make this impossible. However, sequential methods are naturally suited to this task, since we can hope to construct procedures that automatically adapt the (random) sample-size τ to the unknown variance, with τ taking larger values for larger σ^2 to attain the same ϵ width. This was first proved by **?**, who used a generalized likelihood ratio test using improper priors in order to design the test. We state a version of this result described by **?**, Theorem 4.1.

Proposition 1.2 (Theorem 4.1 of **?**). *Fix an $m \geq 2$, and for a given $\alpha \in (0,1)$, let a denote a solution of the equation*

$$2(1 - F_{m-1}(a) + af_{m-1}(a)) = \alpha,$$

where F_{m-1} and f_{m-1} denote the CDF and PDF (resp.) of the t-distribution with $m-1$ degrees of freedom. Then, we have the following:

$$\mathbb{P}_{\mu,\sigma^2}\left(\forall n \geq m : \mu \in [\widehat{\mu}_n - A_n s_n,\, \widehat{\mu}_n + A_n s_n]\right) \geq 1 - \alpha,$$

$$\textit{where} \quad A_n = \sqrt{\left(\frac{1}{m}\left(1 + \frac{a^2}{m-1}\right)^m n\right)^{1/n} - 1}, \ \textit{and} \ s_n = \frac{1}{n}\sum_{i=1}^n \left(X_i - \widehat{\mu}_n\right)^2.$$

This result guarantees the existence of a so-called *confidence sequence* that we will describe later in Section3, and it can be used to define the random stopping time τ as

$$\tau = \inf\{n \geq 1 : A_n s_n \leq \epsilon\}, \quad \text{with} \quad \inf \emptyset = +\infty.$$

It is not too difficult to show that for every (μ, σ^2) pair, we have $\mathbb{P}_{\mu,\sigma^2}(\tau < \infty) = 1$. Hence, the sequential procedure $I_\tau = [\widehat{\mu}_\tau \pm A_\tau s_\tau]$ satisfies the required condition (1).

2 Sequential "Power-one" Tests

For the remainder of this lecture, we will work with some measurable space $(\mathcal{X}^\infty, \mathcal{F})$, containing an indexed class of probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$.Let Θ_0, Θ_1 denote two disjoint subsets of the parameter set Θ, and given a stream of $\mathcal{X}$-valued observations $\{X_i : i \geq 1\} \sim \mathbb{P}_\theta$, consider the hypothesis testing problem

$$H_0 : \theta \in \Theta_0, \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{2}$$

Our goal is to construct a level-α power-one sequential test for this problem, that we formally define next.

Definition 2.1. Let $\{\mathcal{F}_n \subset \mathcal{F} : n \geq 0\}$ denote a filtration such that $\sigma(X_1, \ldots, X_n) = \mathcal{F}_n$ for all $n \geq 1$. Then, a level-α power-one test for the problem (2) is a pair $(\tau, (\phi_k)_{k\in\mathbb{N}})$, such that τ is a $\mathbb{N} \cup \{\infty\}$ valued stopping time, and $\phi_k : \mathcal{X}^k \to \{0, 1\}$ is a $\mathcal{F}_k$-measurable "decision rule" for all $k \in \mathbb{N}$, such that the following conditions hold:

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\phi_\tau(X^\tau) = 1\right) \leq \alpha, \qquad \text{(Level-}\alpha\text{ Property)}$$
$$\inf_{\theta\in\Theta_1} \mathbb{P}_\theta\left(\phi_\tau(X^\tau) = 1\right) = 1. \qquad \text{(Power-one Property)}$$

In many instances we simply have $\phi_\tau = \mathbf{1}_{\tau<\infty}$; the decision is to always reject the null, and we will often follow this convention and simply refer to the stopping time τ as the level-α power-one test.

Example 2.2. *Consider a simple null $\Theta_0 = \{\theta_0\}$ and simple alternative $\Theta_1 = \{\theta_1\}$, and suppose each $\mathbb{P}_\theta$ is an infinite product of P_θ with density p_θ (w.r.t. some common dominating measure). Then, given a stream $\{X_i : i \geq 1\}$ drawn from an unknown $\mathbb{P}_\theta = \otimes_{i=1}^{\infty} p_\theta$, we can define a level-$\alpha$ power-one stopping time τ as*

$$\tau = \inf\{n \geq 1 : W_n \geq 1/\alpha\}, \quad \textit{with} \quad W_0 = 1,\ W_n = W_{n-1} \times \frac{p_{\theta_1}(X_n)}{p_{\theta_0}(X_n)}.$$

The process $\{W_n : n \geq 1\}$ is the likelihood ratio process, and as we will see that is a nonnegative martingale *with an initial value 1 under the null. That is $\mathbb{E}[W_n \mid \mathcal{F}_{n-1}] \overset{a.s.}{=} W_{n-1}$, and hence such a process satisfies a time-uniform variant of Markov's inequality, called Ville's inequality, that says that for any $a > 0$, we have*

$$\mathbb{P}_{\theta_0}\left(\exists n \geq 1 : W_n \geq a\right) \leq \frac{1}{a}.$$

This immediately implies the "level-α property" of τ by selecting $a \leftarrow 1/\alpha$. Furthermore, when the alternative is true, then we can show that

$$\mathbb{E}[\tau] = \mathcal{O}\left(\frac{\log(1/\alpha)}{D_{KL}(P_{\theta_1} \parallel P_{\theta_0})}\right),$$

which is finite whenever $D_{KL}(P_{\theta_1} \parallel P_{\theta_0}) > 0$, and hence also satisfies the power-one property.

The above example illustrates the general steps (in the simplest setting) we will follow in constructing power-one tests, and we will explore this in more details in the next lecture.

3 Estimation via Confidence Sequences

Definition 3.1. Let $\{X_n : n \geq 1\} \sim \mathbb{P}_\theta$ denote an infinite sequence of observations in $\mathcal{X}^\infty$ for some unknown $\theta \in \Theta$. Then, a sequence of possibly random subsets $\{C_n : n \geq 0\}$ is said to form a level-$(1-\alpha)$ confidence sequence for θ if each $C_n \equiv C_n(X^n)$ is constructed based on $(X_1, \ldots, X_n)$, and

$$\mathbb{P}_\theta \left(\forall n \geq 1 : \theta \in C_n \right) \geq 1 - \alpha, \qquad \Longleftrightarrow \qquad \mathbb{P}_\theta \left(\exists n \geq 1 : \theta \not\in C_n \right) \leq \alpha.$$

That is, a confidence sequences (CSs) satisfy a *uniform validity* property (the quantifier $\forall n \geq 1$ is inside the probability statement).

Remark 3.2. A natural consequence of the above definition is that if N is a random stopping time w.r.t. to the natural filtration $\{\mathcal{F}_n : n \geq 1\}$, then we also have

$$\mathbb{P}_\theta \left(\theta \in C_N \right) \geq 1 - \alpha.$$

In other words, confidence sequences retain their validity at data-driven stopping rules. This makes confidence sequences valid under "peeking".

Remark 3.3. Recall that a level-$(1-\alpha)$ confidence interval (CI) for θ constructed using X^n is any set $C_n \subset \Theta$ such that

$$\forall n \geq 1 : \quad \mathbb{P}(\theta \in C_n) \geq 1 - \alpha.$$

This illustrates the crucial difference between CIs and CSs: the quantifier "$\forall n \geq 1$" is outside the probability statement in CIs, and hence their validity guarantee is not time-uniform (it holds only for a fixed n).

Remark 3.4. Due to the uniform validity property of CSs, it is without loss of generality to assume that $C_n \subset C_m$ for all $n \geq m$ (if not, we can simply take the running intersection without violating the validity properties). In most cases, reasonable CSs satisfy $|C_n| \to 0$ as $n \to \infty$ for appropriate notions of size of the sets. If $\mathcal{X} = \mathbb{R}$, then this suggests that CSs are an appropriate tool for "fixed width estimation" problem of Section 1. More specifically, if we wish to construct an ϵ-width estimate of a parameter θ, we can do that by using C_τ with $\tau = \inf\{n \geq 1 : |C_n| \leq \epsilon\}$.

Duality between CSs and power-one tests. As in the fixed sample size case, there exists a natural duality between power-one tests and confidence sequences:

- Consider a testing problem in which given $\{X_i : i \geq 1\} \sim \mathbb{P}$s, we wish the test the null $H_\theta : \mathbb{P} = \mathbb{P}_\theta$ against the alternative $H_{\theta,1} : \mathbb{P} \neq \mathbb{P}_\theta$. Now, suppose we know how to construct a level-α test τ_θ for every such null; that is, $\mathbb{P}_\theta(\tau_\theta < \infty) \leq \alpha$. Then, we can use these to design a confidence sequence $\{C_n : n \geq 0\}$ with $C_0 = \Theta$, and

  $$C_n = \{\theta \in \Theta : \tau_\theta > n\}, \quad \text{for} \quad n \geq 1.$$

  In other words, using the observations $\{X_i : i \geq 1\}$, we run in parallel a continuum of tests $\{\tau_\theta : \theta \in \Theta\}$, and our confidence set at time n is the collection of all parameters θ for which we do not have enough evidence yet to reject the (corresponding) null H_θ. It is easy to verify that if θ^* is the true parameter, then

  $$\mathbb{P}_{\theta^*} \left(\exists n \geq 1 : \theta^* \not\in C_n \right) = \mathbb{P}_{\theta^*} \left(\exists n \geq 1 : \tau_{\theta^*} \leq n \right) = \mathbb{P}_{\theta^*} \left(\tau_{\theta^*} < \infty \right) \leq \alpha.$$

- Now, suppose $\{X_i : i \geq 1\} \sim \mathbb{P}_\theta$, and we want to test the hypothesis $H_{\theta^*} : \theta = \theta^*$. If we know how to construct a level-$(1-\alpha)$ CS $\{C_n : n \geq 1\}$, we can define a stopping time

  $$\tau = \inf\{n \geq 1 : \theta^* \not\in C_n\} \quad \implies \quad \mathbb{P}_{\theta^*}\left(\tau < \infty\right) = \mathbb{P}_{\theta^*}\left(\exists n \geq 1 : \theta^* \not\in C_n\right) \leq \alpha.$$

  The power-one property will require that the width of the set $|C_n|$ converges to 0 almost surely.

4 Sequential Change Detection

This is the third main inference task we will study in this course. In this problem, we again assume we have a stream of observations $\{X_i : i \geq 1\}$, and there is an unknown time $T \in \mathbb{N} \cup \{\infty\}$ at which the distribution changes from $\mathbb{P}_{\theta_0}$ to $\mathbb{P}_{\theta_1}$. In general, we assume that $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ for some disjoint classes $\Theta_0, \Theta_1 \subset \Theta$, and our goal is to design a stopping time τ that satisfies the following:

$$\begin{aligned} & \mathrm{minimize}\ D(\tau, \Theta_0, \Theta_1), \ \text{subject to } \mathrm{ARL}(\tau, \Theta_0) \geq 1/\alpha, \\ \text{where} \quad & \mathrm{ARL}(\tau, \Theta_0) = \inf_{\theta_0 \in \Theta_1} \mathbb{E}_{\theta_0}[\tau], \quad \text{and} \quad D(\tau, \Theta_0, \Theta_1) = \sup_{T \in \mathbb{N}} \sup_{\theta_j \in \Theta_j} \mathbb{E}_{\theta_0, T, \theta_1}[(\tau - T + 1)^+ \mid \mathcal{F}_{T-1}]. \end{aligned}$$

Here we use the notation $\mathbb{E}_{\theta_0,T,\theta_1}$ to indicate the probability measure over $\mathcal{X}^\infty$ in which the first $T-1$ observations are from P_{θ_0} and the ones starting at T are from P_{θ_1}.

The term $D(\tau, \Theta_0, \Theta_1)$ is the worst-case (over the choices of the pre- and post-change parameters and the changepoint) detection delay, and $\mathrm{ARL}(\tau, \Theta_0)$ is the worst-case *average run length* when there is no change.

As in the case of confidence sequences, we can reduce the task of sequential change detection to that of constructing level-α tests for the pre-change class. In particular, consider a testing problem with the null $H_0 : \theta \in \Theta_0$, and for any $k \geq 1$, let N_k denote a level-α test for the null H_0 constructed using $\{X_i : i \geq k\}$. Then, we can show that

$$\tau = \inf_{k \geq 1}\{N_k + k - 1\}$$

is a valid change detection scheme; that is, for this τ, we have $\inf_{\theta_0 \in \Theta_0} \mathbb{E}_{\theta_0}[\tau] \geq 1/\alpha$. The detection delay of the stopping time τ will depend on the expected stopping time of the individual tests N_k, and we will explore this connection in a later lecture.

5 Conclusion

Our discussion reveals that, *in principle*, the problems of sequential estimation via confidence sequences and sequential change detection can be “reduced to” constructing power-one tests for appropriate hypotheses (*in principle*, because the reduction might end up being computationally infeasible). Hence, the key methodological issue for us is to develop general techniques for constructing powerful sequential power-one tests and establish their optimality. That will be the focus of the next few lectures.