

STA3000F: Advanced Theory of Statistics

Fall 2026

Lecture 5: October 9, 2026

Lecturer: Wenlong Mou

1 Testing continued

Motivation: in most of practical problems, we may not be able to find UMP tests. So we need to consider more general notions of optimality.

1.1 Two-sided testing and UMPU

Two-sided testing problem:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$

Problem: we can always enhance the power of the test on one side by sacrificing the power on the other side. So we cannot find a UMP test for two-sided testing problem.

This motivates the concept of unbiased test:

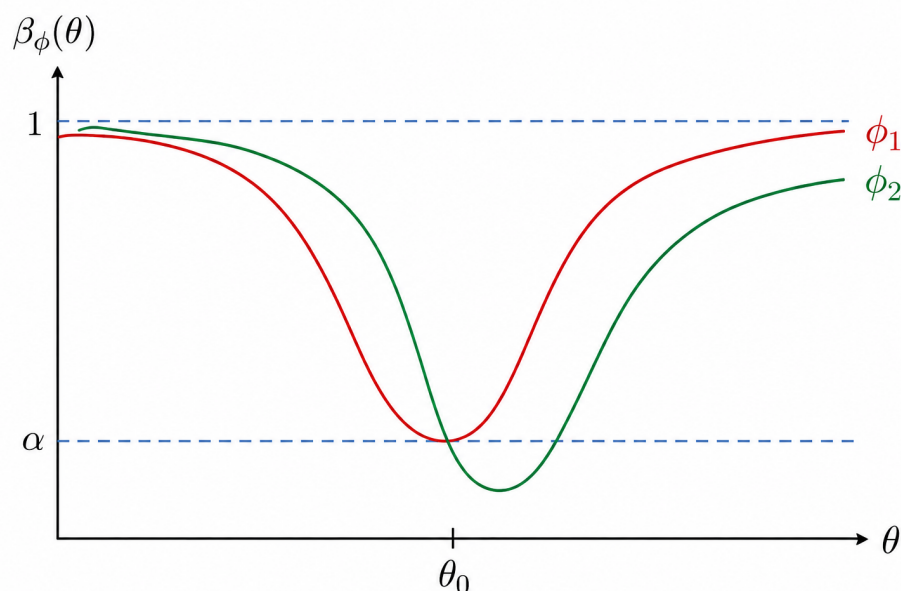
Definition 5.1 (unbiased test and UMPU test). *Within the two-sided testing framework, a test ϕ is called unbiased if*

$$\mathbb{E}_\theta[\phi(X)] \geq \mathbb{E}_{\theta_0}[\phi(X)], \quad \forall \theta \in \Theta_1.$$

A test ϕ is called uniformly most powerful unbiased (UMPU) if it is unbiased and has the highest power among all unbiased tests, uniformly for all $\theta \in \Theta_1$.

This is analogous to UMVU in estimation theory.

The following illustration compares an unbiased test with a biased test.



Both tests have rejection probability α at θ_0 . The red curve ϕ_1 stays above α under the alternative, whereas the green curve ϕ_2 falls below α for some alternatives and is therefore biased.

Finding UMPU tests: In this case, we restrict our attention to exponential families.

$$p_\eta(x) = h(x) \exp(\eta \cdot T(x) - A(\eta))$$

Testing problem: $\eta = \eta_0$ vs. $\eta \neq \eta_0$, where η_0 is an interior point of the interval $\eta(\Theta)$.

From unbiasedness, we know that the power function $\beta_\phi(\eta)$ is minimized at $\eta = \eta_0$. Under regularity conditions, this implies

$$0 = \frac{d}{d\eta} \beta_\phi(\eta) \Big|_{\eta=\eta_0} = \frac{d}{d\eta} \mathbb{E}_\eta[\phi(X)] \Big|_{\eta=\eta_0} = \mathbb{E}_{\eta_0}[\phi(X)T(X)] - \mathbb{E}_{\eta_0}[\phi(X)]\mathbb{E}_{\eta_0}[T(X)].$$

Since we want to utilize the full significance level α , we want to have

$$\mathbb{E}_{\eta_0}[\phi(X)] = \alpha, \tag{5.1}$$

so we have

$$\mathbb{E}_{\eta_0}[T(X) \cdot (\phi(X) - \alpha)] = 0. \tag{5.2}$$

Intuition from exponential family tells us that we should only reject when $T(X)$ is either too small or too large. So we can consider the following form of the test:

$$\phi(x) = \begin{cases} 1, & T(x) < c_1 \text{ or } T(x) > c_2, \\ \pi_i, & T(x) = c_i, \ i = 1, 2, \\ 0, & c_1 < T(x) < c_2. \end{cases} \tag{5.3}$$

(canonical example: binomial testing, where we put π_1 and π_2 to fill in the gap to achieve level α).

This test is determined by the parameters (c_1, c_2, π_1, π_2) . Similar to the one-sided case, π_i 's are only used to fill in the gaps when we change c_i discontinuously. Intuitively, this is closer to a two-parameter family, and we have two equations.

Theorem 5.1. *Under above assumptions, for $\alpha \in (0, 1)$, there exists a test in the form of Equation (5.3) that satisfies Equations (5.1) and (5.2). This test is UMPU at level α .*

Proof (see textbook): generalized Neyman–Pearson lemma.

1.2 Multivariate testing

Canonical example: $X \sim \mathcal{N}(\theta, I_d)$, where $\theta \in \mathbb{R}^d$. We want to test $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$.

- Even in $d = 1$, UMP does not exist (two-sided testing).
- When $d \geq 2$, UMPU does not exist either, e.g.

$$\phi_1(x) = \mathbf{1}_{|x_1| \geq c}, \quad \phi_2(x) = \mathbf{1}_{|x_2| \geq c}, \quad \phi_3(x) = \mathbf{1}_{\|x\|_2 \geq c'}, \quad \phi_4(x) = \mathbf{1}_{\|x\|_\infty \geq c''}.$$

Similar to estimation theory, this makes us turn to minimax optimality.

Framework of minimax testing radius: for a given $\varepsilon > 0$, we want to test

$$\begin{aligned} H_0 : \theta &\in \Theta_0, \\ H_1 : \theta &\in \Theta_1(\varepsilon) = \Theta_1 \cap \{\theta : d(\theta, \Theta_0) \geq \varepsilon\}, \end{aligned}$$

We need to put some margin here to make it possible to distinguish the two hypotheses.

Risk function for testing:

$$R_\varepsilon(\phi) = \sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\phi(X)] + \sup_{\theta \in \Theta_1(\varepsilon)} \mathbb{E}_\theta[1 - \phi(X)].$$

Minimax testing radius: the smallest ε such that there exists a test ϕ with $R_\varepsilon(\phi) \leq \alpha$ for some $\alpha \in (0, 1)$. Usually, we only care about the order of the minimax testing radius, and the constant α is not important. (e.g. we can take $\alpha = 1/4$).

Revisiting the one-dimensional Gaussian example: $X \sim \mathcal{N}(\theta, 1)$, we want to test $H_0 : \theta = 0$ vs. $H_1 : |\theta| \geq \varepsilon$. We have seen that the minimax testing radius is of order $\varepsilon \asymp 1/\sqrt{n}$.

Now we consider $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, I_d)$, and we want to test $H_0 : \theta = 0$ vs. $H_1 : \|\theta\|_2 \geq \varepsilon$. First attempt: we use the same idea in one-dimensional case, we know that

$$\mathbb{E}_\theta[\|\bar{X}_n - \theta\|_2^2] = \frac{d}{n}.$$

We may naturally consider the test that rejects when $\|\bar{X}_n\|_2^2$ is large. There is no problem with this test per se, but how about the minimax testing radius?

By the MSE bound above, we have

$$\|\bar{X}_n\|_2 \leq c\sqrt{d/n} \quad \text{with high probability.}$$

On the other hand, when $\varepsilon \geq 2c\sqrt{d/n}$, under $H_1(\varepsilon)$, we have

$$\|\bar{X}_n\|_2 \geq \|\theta\|_2 - \|\bar{X}_n - \theta\|_2 \geq \varepsilon - c\sqrt{d/n} \geq c\sqrt{d/n}.$$

This may suggest that the minimax testing radius is of order $\varepsilon \asymp \sqrt{d/n}$. But this is not correct. We will see that the minimax testing radius is actually of order $\varepsilon \asymp d^{1/4}/\sqrt{n}$.

The proof consists of two parts: (1) we need to show that there exists a test that can achieve this rate; (2) we need to show that no test can do better than this rate.

Construction of the test: Obviously, we still want to use the test that rejects when $\|\bar{X}_n\|_2^2$ is large. But we need to do more careful calculation about the threshold. Indeed, we have

$$\begin{aligned} \mathbb{E}_\theta[\|\bar{X}_n\|_2^2] &= \frac{d}{n} + \|\theta\|_2^2, \\ \text{var}_\theta(\|\bar{X}_n\|_2^2) &= \sum_{j=1}^d \text{var}_\theta((e_j^\top \bar{X}_n)^2) = \frac{4}{n} \|\theta\|_2^2 + \frac{2d}{n^2}. \end{aligned}$$

Intuitively, that is why a smaller testing radius is possible: the fluctuation of $\|\bar{X}_n\|_2^2$ is of order $\sqrt{d/n}$, which is smaller than the order of magnitude of $\|\bar{X}_n\|_2$, which is of order $\sqrt{d/n}$.

Based on above calculation, we can choose the rejection threshold as

$$C(a) = \mathbb{E}_0[\|\bar{X}_n\|_2^2] + a \cdot \sqrt{\text{var}_0(\|\bar{X}_n\|_2^2)}$$

Now we can calculate the type I error and type II error. For type I error, we have

$$\mathbb{E}_0[\phi(X)] = \mathbb{P}_0(\|\bar{X}_n\|_2^2 > C(a)) \leq \frac{1}{a^2},$$

which follows from Chebyshev's inequality.

For the type II error, we

$$\begin{aligned}\mathbb{E}_\theta[1 - \phi(X)] &= \mathbb{P}_\theta(\|\bar{X}_n\|_2^2 \leq C(a)) \\ &= \mathbb{P}_\theta\left(\|\bar{X}_n\|_2^2 - \mathbb{E}_\theta[\|\bar{X}_n\|_2^2] \leq \frac{d}{n} + \frac{2a\sqrt{d}}{n} - \mathbb{E}_\theta[\|\bar{X}_n\|_2^2]\right) \\ &= \mathbb{P}_\theta\left(\|\bar{X}_n\|_2^2 - \mathbb{E}_\theta[\|\bar{X}_n\|_2^2] \leq \frac{2a\sqrt{d}}{n} - \|\theta\|_2^2\right).\end{aligned}$$

By choosing $\varepsilon = 4ad^{1/4}/\sqrt{n}$, for any $\theta \in H_1(\varepsilon)$, we have

$$\mathbb{P}_\theta\left(\|\bar{X}_n\|_2^2 - \mathbb{E}_\theta[\|\bar{X}_n\|_2^2] \leq \frac{2a\sqrt{d}}{n} - \|\theta\|_2^2\right) \leq \frac{\frac{4}{n}\|\theta\|_2^2 + \frac{2d}{n^2}}{(\|\theta\|_2^2 - 2a\sqrt{d/n^2})^2}.$$

We want to make sure that this is smaller than $1/a^2$ as well, which is equivalent to

$$\|\theta\|_2^2 - 2a\sqrt{d/n^2} \geq 2a\sqrt{\frac{1}{n}\|\theta\|_2^2 + \frac{d}{2n^2}}.$$

We can solve for $\|\theta\|_2^2$, and this inequality holds true as long as $\|\theta\|_2^2 \geq c'(a)\frac{\sqrt{d}}{n}$, for some function $c'(a)$ that depends on a . This shows that the test is possible when $\varepsilon \asymp d^{1/4}/\sqrt{n}$.

This is in contrast with the error bound for estimation, where the MSE is of order d/n .

Detecting the existence of a signal is easier than estimating the signal.

This is a general phenomenon in high-dimensional statistics, e.g. in graph theory, detecting the existence of a community is easier (requiring weaker signal strength) than estimating the community structure.

Information theoretic lower bound For any test ϕ , we want to show that

$$\mathbb{E}_0[\phi(X)] + \sup_{\theta \in H_1(\varepsilon)} \mathbb{E}_\theta[1 - \phi(X)] \geq 1/2,$$

when $\varepsilon \leq c_1 d^{1/4}/\sqrt{n}$.

First attempt: Le Cam's two-point method with one vs. one testing. We can arbitrarily choose θ_1 such that $\|\theta_1\|_2 = \varepsilon$. If we consider the one-vs-one testing problem

$$H_0 : \theta = 0, \quad H_1 : \theta = \theta_1, \tag{5.4}$$

this is essentially a one-dimensional testing problem, and we could choose

$$\phi(X) = \mathbf{1}_{\{\bar{X}_n^\top \theta_1 \geq c\}}.$$

Using this test, we can show that the sum of type I and type II errors is at least $1/2$ when $\varepsilon \leq c_1/\sqrt{n}$. On the other hand, for the original problem, we know that we cannot use such a test, because we do not know the direction of θ_1 . So reducing to the test (5.4) is too weak. We need better tools.

From decision theory, we know that the minimax risk is lower bounded by the Bayes risk with respect to any prior. The one-vs-one attempt is equivalent to using a point mass prior on θ_1 . In general, we have

$$\begin{aligned} \mathbb{E}_0[\phi(X)] + \sup_{\theta \in H_1(\varepsilon)} \mathbb{E}_\theta[1 - \phi(X)] \\ \geq \mathbb{E}_0[\phi(X)] + \int_{\theta \in H_1(\varepsilon)} \mathbb{E}_\theta[1 - \phi(X)] d\pi(\theta), \end{aligned}$$

by constructing a prior distribution π supported on $H_1(\varepsilon)$. Operationally, we can also think of this as a one-vs-one testing problem, but the alternative hypothesis is now a mixture distribution, instead of a point mass.

$$H_0 : \theta = 0, \quad H_1 : \theta \sim \pi.$$

Under this H_1 , the joint distribution of $X_1, \dots, X_n$ is a mixture of product distributions (a.k.a. hierarchical Bayesian model).

Since this is formulated as a one-vs-one testing problem, we can still use Neyman–Pearson lemma and/or total variation distance calculation to lower bound the sum of type I and type II errors. Here we need to compute the TV distance between a product distribution and a mixture of product distributions. Before that, we need to construct the prior distribution π . We need π to be genuinely d -dimensional.

- Natural choice: uniform distribution on the sphere $\mathbb{S}^{d-1}(\varepsilon) = \{\theta : \|\theta\|_2 = \varepsilon\}$. This works, but the calculation is a bit complicated.
- Simpler choice: a discrete construction:

$$\pi = \text{uniform distribution on } \{\pm\varepsilon/\sqrt{d}\}^d.$$

This is a discrete prior over 2^d points. Each coordinate of θ is either $+\varepsilon/\sqrt{d}$ or $-\varepsilon/\sqrt{d}$, each with probability $1/2$, and the coordinates are independent.

From last lecture, we know that

$$\mathbb{E}_0[\phi(X)] + \int_{\theta \in H_1(\varepsilon)} \mathbb{E}_\theta[1 - \phi(X)] d\pi(\theta) \geq 1 - d_{\text{TV}}(P_0^{\otimes n}, \int_{\theta \in H_1(\varepsilon)} P_\theta^{\otimes n} d\pi(\theta)).$$

For notation simplicity, we denote $P_0^{\otimes n}$ as P_0 and $\int_{\theta \in H_1(\varepsilon)} P_\theta^{\otimes n} d\pi(\theta)$ as P_1 .

By Cauchy–Schwarz inequality, we have

$$d_{\text{TV}}(P_0, P_1) = \mathbb{E}_{P_0} \left[\left| \frac{dP_1}{dP_0} - 1 \right| \right] \leq \sqrt{\mathbb{E}_{P_0} \left[\left| \frac{dP_1}{dP_0} - 1 \right|^2 \right]} = \sqrt{\mathbb{E} \left[\left(\frac{dP_1}{dP_0} \right)^2 \right] - 1}.$$

Actually, the square of the right-hand-side is called the χ^2 divergence between P_1 and P_0 .

$$\chi^2(P_1 \| P_0) = \mathbb{E}_{P_0} \left[\left(\frac{dP_1}{dP_0} - 1 \right)^2 \right] = \mathbb{E}_{P_0} \left[\left(\frac{dP_1}{dP_0} \right)^2 \right] - 1.$$

Define the likelihood ratio

$$\begin{aligned} L(X_1, \dots, X_n) &= \frac{dP_1}{dP_0}(X_1, \dots, X_n) = \frac{1}{2^d} \sum_{z \in \{\pm 1\}^d} \frac{\exp \left(-\frac{1}{2} \sum_{i=1}^n \|X_i - \varepsilon z / \sqrt{d}\|_2^2 \right)}{\exp \left(-\frac{1}{2} \sum_{i=1}^n \|X_i\|_2^2 \right)} \\ &= \mathbb{E}_{z \sim \text{Unif}(\{\pm 1\}^d)} \left[\exp \left(\frac{\varepsilon}{\sqrt{d}} n \bar{X}_n^\top z - \frac{\varepsilon^2 n}{2} \right) \right]. \end{aligned}$$

Now we need to compute the second moment of L under P_0 :

$$\begin{aligned}
\mathbb{E}_{P_0}[L^2] &= \mathbb{E}_0 \left[\left(\mathbb{E}_{z \sim \text{Unif}(\{\pm 1\}^d)} \left[\exp \left(\frac{\varepsilon}{\sqrt{d}} n \bar{X}_n^\top z - \frac{\varepsilon^2 n}{2} \right) \right] \right)^2 \right] \\
&= \mathbb{E}_{z, z' \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\}^d)} \left[\mathbb{E}_0 \left[\exp \left(\frac{\varepsilon}{\sqrt{d}} n \bar{X}_n^\top (z + z') - \varepsilon^2 n \right) \right] \right] \\
&= \mathbb{E}_{z, z' \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\}^d)} \left[\exp \left(\frac{n \varepsilon^2}{2d} \|z + z'\|_2^2 - n \varepsilon^2 \right) \right] \\
&= \mathbb{E}_{z, z' \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(\{\pm 1\}^d)} \left[\exp \left(\frac{n \varepsilon^2}{d} z^\top z' \right) \right] \\
&= \left(\frac{1}{2} \exp(n \varepsilon^2 / d) + \frac{1}{2} \exp(-n \varepsilon^2 / d) \right)^d \\
&\leq \exp \left(\frac{1}{2} \left(\frac{n \varepsilon^2}{d} \right)^2 d \right) = \exp \left(\frac{n^2 \varepsilon^4}{2d} \right)
\end{aligned}$$

We want to make sure that $d_{\text{TV}}(P_0, P_1) \leq 1/2$, which is implied by $\chi^2(P_1 \| P_0) \leq 1/4$, which is implied by $\mathbb{E}_{P_0}[L^2] \leq 5/4$. Solve for ε . this holds true as long as $\varepsilon \leq c_1 d^{1/4} / \sqrt{n}$, for some universal constant c_1 . This shows that the minimax testing radius is of order $\varepsilon \asymp d^{1/4} / \sqrt{n}$.

In general, this type of construction is called fuzzy hypotheses, where we construct a prior distribution over the null/alternative hypothesis. With increasing degree of generality, we have

- One vs. one testing (last lecture)
- One vs. mixture testing (this lecture)
- Mixture vs. mixture testing

2 M -estimation/Empirical risk minimization/Sample average approximation

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n f(X_i; \theta).$$

(sometimes you also see regularized variants).

Here we assume that $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$. Define the population loss function

$$F(\theta) = \mathbb{E}_{\mathbb{P}}[f(X; \theta)].$$

In our analysis, we want to show that $\hat{\theta}_n$ is “close” to the population minimizer

$$\theta^* \in \arg \min_{\theta \in \Theta} F(\theta),$$

in the sense of distance $\|\hat{\theta}_n - \theta^*\|$ or function value $F(\hat{\theta}_n) - F(\theta^*)$.

For notation convenience, we define the empirical loss function

$$F_n(\theta) = \frac{1}{n} \sum_{i=1}^n f(X_i; \theta).$$

2.1 Basic inequality

By law of large numbers or concentration inequality, we know that for each fixed θ , we have

$$|F_n(\theta) - F(\theta)| \begin{cases} \rightarrow 0, & \text{a.s., (assume finite first moment)} \\ = O(\sqrt{\frac{\log(1/\delta)}{n}}), & \text{with probability at least } 1 - \delta, \text{ (assume boundedness)} \end{cases}$$

Is this sufficient for our analysis?

$$F(\hat{\theta}_n) - F(\theta^*) = \underbrace{F(\hat{\theta}_n) - F_n(\hat{\theta}_n)}_{?} + \underbrace{F_n(\hat{\theta}_n) - F_n(\theta^*)}_{\leq 0} + \underbrace{F_n(\theta^*) - F(\theta^*)}_{\text{easy}}$$

The first term is evaluated at a random point $\hat{\theta}_n$, which is not fixed. So above arguments do not apply. Simple idea:

$$|F(\hat{\theta}_n) - F_n(\hat{\theta}_n)| \leq \sup_{\theta \in \Theta} |F(\theta) - F_n(\theta)|.$$

We need uniform law of large numbers, i.e., we need to show that

$$\sup_{\theta \in \Theta} |F(\theta) - F_n(\theta)| \xrightarrow{P} 0,$$

and also its quantitative version. This is the core theme of empirical process theory. Empirical process studies

- Uniform convergence.
- Quantitative bounds on the uniform convergence rate.
- Uniform central limit theorems.

2.2 When do we have uniform convergence?

Simple case: Θ is finite. Then we can use union bound to show that

$$\mathbb{P}\left(\sup_{\theta \in \Theta} |F(\theta) - F_n(\theta)| > \epsilon\right) \leq \sum_{\theta \in \Theta} \mathbb{P}(|F(\theta) - F_n(\theta)| > \epsilon).$$

Each terms go to zero (and have quantitative bounds). So the whole sum goes to zero.

How about infinite (especially continuous) Θ ? Idea: discretization.

Definition 5.2 (covering number). *Given a set K and a metric ρ , the covering number $N(K; \rho, \epsilon)$ is defined as*

$$N(K; \rho, \epsilon) = \min \left\{ n : \exists x_1, \dots, x_n \in K, K \subseteq \bigcup_{i=1}^n B_\rho(x_i, \epsilon) \right\},$$

Definition 5.3 (packing number). *Given a set K and a metric ρ , the packing number $M(K; \rho, \epsilon)$ is defined as*

$$M(K; \rho, \epsilon) = \max \left\{ n : \exists x_1, \dots, x_n \in K, \rho(x_i, x_j) \geq \epsilon, i \neq j \right\},$$