ECE 598 - Fall 2025

Lecture 5: Statistical Decision Theory

September 9, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we will start our formal discussion of statistical inference problems. Since the eventual goal of statistical inference is often to aid decision-making, we will take a decision-theoretic approach to develop the formalism. In particular, we will introduce the notion of Loss functions (or negative utility functions) to quantify the quality of our inference, which will then help us define some notions of optimal inference procedures.

This somewhat abstract formulation is useful as it allows us to analyze a large class of problems, such as estimation, hypothesis testing, and stochastic optimization. In later lectures, we will expand this framework to allow for interactive and sequential decision-making (bandits, RL, etc.).

1 The general framework

Throughout this section, we will assume that there is an underlying probability space, $(\Omega, \mathcal{F}_\Omega, \mathbb{P})$, large enough to contain all the random variables that arise in our discussion. To study the optimality of statistical inference procedures, we require the following components:

- **Statistical model or experiment.** Let $\{P_\theta : \theta \in \Theta\}$ denote a family of probability distributions on some observation space $(\mathcal{X}, \mathcal{F}_\mathcal{X})$ indexed by a "parameter space" $(\Theta, \mathcal{F}_\Theta)$ endowed with a "pseudo-metric" $\rho : \Theta \times \Theta \to [0, \infty)$. By pseudo-metric, we mean that ρ is symmetric, satisfies triangle inequality, and $\rho(\theta, \theta) = 0$ for all $\theta \in \Theta$. That is, $\rho(\theta, \theta') = 0$ does not necessarily imply that $\theta' = \theta$ (as it would if ρ were a metric).

- **Observations.** Let $\boldsymbol{X} \sim P_\theta$ denote an $\mathcal{X}$-valued observation drawn from the distribution corresponding to the unknown parameter θ. In many important cases, we have $\mathcal{X} = \otimes_{i=1}^n \mathcal{X}_i$ for some spaces $\{\mathcal{X}_i : 1 \le i \le n\}$, and $\boldsymbol{X} = (X_1, \ldots, X_n)$ denotes n observations sampled according to the distribution P_θ on $\mathcal{X}$. This can be further specialized to the case of independent observations, when $P_\theta = \otimes_{i=1}^n p_{\theta,i}$. Finally, when $\mathcal{X}_i = \mathcal{X}$ and $p_{\theta,i} = p_\theta$ for all $i \in [n]$, then we reduce to the case of i.i.d. observations.

- **Decision.** Let $(\mathcal{W}, \mathcal{F}_\mathcal{W})$ denote the "decision space", and let $\boldsymbol{W} \sim P_{\boldsymbol{W}|\boldsymbol{X}}$ denote the $\mathcal{W}$-valued "decision" made by the statistician or data-analyst based on the observations $\boldsymbol{X}$. The stochastic kernel $P_{\boldsymbol{W}|\boldsymbol{X}}$ is referred to as the (randomized) "decision rule", and can be equivalently represented (in most problems) by a function $h(\boldsymbol{X}, U)$, where $h : \mathcal{X} \times [0,1] \to \mathcal{W}$, and $U \sim \mathrm{Uniform}([0,1])$ random variable independent of $\mathcal{X}$.

- **Loss Function.** For some loss function $L : \mathcal{W} \times \Theta \to \mathbb{R}$, the term $L(\boldsymbol{W}, \theta)$ assigns a numerical value to the quality of the decision W when the true underlying parameter is θ (i.e., when $\boldsymbol{X} \sim P_\theta$). Throughout this course, we will assume that a loss function (or a utility function) has already been specified for us, and not concern ourselves with the

design of appropriate loss or utility functions for a given task. An axiomatic approach for the existence and design of utility functions lies under the field of *Utility theory*; see for example, Bernardo and Smith (2009).

Thus, there are three major *spaces* under consideration: the *parameter space* $(\Theta, \mathcal{F}_\Theta)$, the *decision space* (also sometimes known as the action space) $(\mathcal{W}, \mathcal{F}_\mathcal{W})$, and the *observation space* $(\mathcal{X}, \mathcal{F}_\mathcal{X})$. Our goal is to construct a possibly randomized decision rule; $P_{W|\boldsymbol{X}}$, which results in a small expected loss, also called the *risk*:

$$R(P_{W|\boldsymbol{X}}, \theta) = \mathbb{E}_{(\boldsymbol{X},W)}\left[L(W, \theta)\right].$$

2 Examples

In this section, we will illustrate how the general framework introduced in the previous section models various tasks in statistics and machine learning.

2.1 Estimation

The most important instance of the general decision problem discussed in Section 1 is that of estimation. In this case, we have the following:

- For simplicity, let us assume the case of i.i.d. observations $\boldsymbol{X} = (X_1, \ldots, X_n)$. This corresponds to the case of $\mathcal{X} = \mathcal{X}_i^n$ and $\boldsymbol{P}_\theta = P_\theta^{\otimes n}$.
- The decision space in this case is the parameter space Θ itself, and often we refer to the decision rule $P_{W|\boldsymbol{X}}$ with $\widehat{\theta} : \mathcal{X} \times [0,1] \to \Theta$ (the extra argument in the definition of $\widehat{\theta}$ is to allow for randomization if necessary). This mapping $\widehat{\theta}$ is referred to as the estimator.
- The loss function often has the form $L(\widehat{\theta}, \theta) = \Phi \circ \rho(\widehat{\theta}, \theta)$, where recall that ρ is the pseudo-metric on the parameter space Θ, and $\Phi : [0, \infty) \to [0, \infty)$ is a non-decreasing function (for example $\Phi(t) = t^2$).

The risk associated with any estimator is defined as

$$R_n(\widehat{\theta}, \theta) = \mathbb{E}_{(\boldsymbol{X},\widehat{\theta})}[\Phi \circ \rho(\widehat{\theta}(\boldsymbol{X}), \theta)].$$

Example 2.1. *Consider an estimation problem with $\Theta = \mathbb{R}$, $P_\theta = N(\theta, \sigma^2)$, data $\boldsymbol{X} = (X_1, \ldots, X_n) \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$. The decision space is also Θ, and let $P_{W|\boldsymbol{X}}$ denote any estimator. The usual loss function is $L(\theta, w) = |\theta - w|^2$, which corresponds to $\rho(\theta, w) = |\theta - w|$ and $\Phi(t) = t^2$.*

A powerful estimator in this case is the maximum likelihood estimator (MLE), defined as

$$\widehat{\theta}(\boldsymbol{X}) \in \underset{w \in \Theta}{\operatorname{argmax}} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n |X_i - w|^2\right)$$

In this case, the MLE turns out to have the familiar form of the sample mean

$$\widehat{\theta}(\boldsymbol{X}) = \bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^n X_i, \quad \textit{and} \quad \mathbb{E}\left[|\widehat{\theta}(\boldsymbol{X}) - \theta|^2\right] = \frac{\sigma^2}{n}.$$

Example 2.2. *Consider a parameter space for some $M > 0$ and $\gamma \in (0, 1]$:*

$$\Theta \equiv \Theta(M, \gamma) = \{(M, \gamma) \textit{ Hölder continuous densities on } [0, 1]\}.$$

Recall that a function is (M, γ) Hölder continuous if $|f(x) - f(x')| \leq M|x - x'|^\gamma$. Now, for every $\theta \in \Theta(M, \gamma)$, let P_θ denote the distribution on $[0, 1]$ with density θ. Given $\boldsymbol{X} = (X_1, \ldots, X_n)$ i.i.d. from some distribution P_θ, our goal is to construct an estimate $\widehat{\theta} \in \mathcal{W}$, where $\mathcal{W} = \{\textit{all densities on } [0, 1]\}$. Let $\rho(\theta, \theta') = \int(\theta(x) - \theta'(x))^2 dx$, and $\Phi(t) = t$ be the identity function.

In this case, a natural idea is to use the histogram estimator: partition the domain $[0, 1]$ into m equal intervals $I_1, \ldots, I_m$, with $I_j = [(j - 1)/m, j/m]$, and define

$$\widehat{\theta}_n(x) = \sum_{j=1}^{m} m\hat{p}_j \mathbf{1}_{x \in I_j}, \quad \textit{where} \quad \hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in I_j}.$$

We can verify that the risk of this histogram estimator is equal to

$$R_n(\theta, \widehat{\theta}) \leq C_1 m^{-2\gamma} + C_2 \frac{m}{n},$$

which on balancing with $m \asymp n^{1/(2\gamma+1)}$ gives us the rate $n^{-2\gamma/(2\gamma+1)}$. In the next week's lectures we will see that this rate cannot be improved.

2.2 Hypothesis Testing

Another important instance of the general decision problem is that of (binary) hypothesis testing. In this case, we have the following:

- i.i.d. observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ corresponding to $\mathcal{X} = \mathbb{X}^n$ and $\boldsymbol{P}_\theta = P_\theta^{\otimes n}$.
- For two disjoint subsets Θ_0, Θ_1 of the parameter space Θ, the goal is to decide whether the true parameter lies in Θ_0 or Θ_1. Hence, the decision space in this case is $\mathcal{W} = \{0, 1\}$.
- A possibly randomized hypothesis test corresponds to a kernel $P_{W|\boldsymbol{X}}$, which can be represented by a map $\Psi : \mathcal{X} \to [0, 1]$, with $\Psi(\boldsymbol{x}) = \mathbb{P}(W = 1|\boldsymbol{X} = \boldsymbol{x})$. A deterministic hypothesis test maps $\mathcal{X}$ to $\mathcal{W} = \{0, 1\}$.
- One possible way of designing a loss function is of the form

$$L(\theta, w) = c_{ij} \quad \text{if} \quad \theta \in \Theta_i,\ w \in \Theta_j,\ \text{for } i, j \in \{0, 1\}.$$

Example 2.3. *With the same setting as Example 3.3, let us consider the case of $\Theta_0 = \{0\}$ and $\Theta_1 = \Theta \setminus \{0\}$ (a two-sided test).*

A natural (likelihood based) test for this problem is

$$\Psi(\boldsymbol{X}) = \begin{cases} 1 & \textit{if } \frac{\sqrt{n}|\bar{\boldsymbol{X}}_n|}{\sigma} > t, \\ 0 & \textit{if } \frac{\sqrt{n}|\bar{\boldsymbol{X}}_n|}{\sigma} \leq t. \end{cases}$$

for an appropriate threshold t.

We now go to a simple nonparametric testing problem.

Example 2.4. *Suppose we have a data $\{(U_i, V_i) : 1 \le i \le n\}$ drawn i.i.d. from a joint distribution $P_U \times P_V$, with both P_U, P_V supported on the real line $\mathbb{R}$. Consider the hypotheses*

$$H_0 : P_U = P_V, \quad \textit{versus} \quad H_1 : P_U \neq P_V.$$

Here $\Theta = \mathcal{P} \times \mathcal{P}$, where $\mathcal{P} \equiv \mathcal{P}(\mathbb{R})$ is the space of all distributions on $\mathbb{R}$; $\Theta_0 = \{(P, P) : P \in \mathcal{P}\}$ and $\Theta_1 = \Theta \setminus \Theta_0$.

In this case, we cannot use likelihood based tests (there is no common dominating measure for the distribution classes involved to define the family of likelihood ratios). A natural idea is to use the data to estimate the distance between the two distributions, and reject the null if the empirical distance is large. In this case,

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_{n,U}(x) - \hat{F}_{n,V}(x)|,$$

where $\hat{F}_{n,U}, \hat{F}_{n,V}$ denote the empirical CDFs. Then, we can define a hypothesis test

$$\Psi(\boldsymbol{X}) = \begin{cases} 1, \text{ if } D_n > t, \\ 0, \text{ if } D_n \le t. \end{cases}$$

The threshold can be selected either by permutations or via the limiting Kolmogorov distribution.

2.3 Binary Classification

We now discuss how the framework can be used to study the classical machine learning task of binary classification.

- In this case, our observation space is $\mathcal{X} = \otimes_{i=1}^n (\mathbb{Z} \times \{0, 1\})$, and we have n i.i.d. feature-label pairs denoted by $\boldsymbol{X} = \{(Z_i, Y_i) : 1 \le i \le n\}$. Each distribution $\boldsymbol{P}_\theta = (P_{\theta,ZY})^{\otimes n} = (P_{\theta,Z} P_{\theta,Y|Z})^{\otimes n}$.
- The decision space is some family of binary classifiers $\mathcal{W} = \{w : \mathbb{Z} \to \{0, 1\}\}$, and the decision rule is represented by some, potentially randomized, binary classifier $\boldsymbol{P}_{W|\boldsymbol{X}}$.
- The loss function $L : \mathcal{W} \times \Theta \to \mathbb{R}$ is the expected value of some classification error $\phi : \mathcal{W} \times \mathbb{Z} \times \{0, 1\} \to \mathbb{R}$:

  $$L(w, \theta) = \mathbb{E}_{P_{\theta,ZY}}[\phi(w, Z, Y)], \quad \text{where} \quad (Z, Y) \perp \boldsymbol{X}.$$

  For example, the 0-1 loss corresponds to $\phi(w, z, y) = \mathbf{1}_{w(z) \neq y}$.

Then, the risk associated with a learning algorithm $\boldsymbol{P}_{W|\boldsymbol{X}}$ is equal to

$$R_n(\boldsymbol{P}_{W|\boldsymbol{X}}, \theta) = \mathbb{E}_{\boldsymbol{X} \sim (P_{\theta,ZY})^{\otimes n}}[\mathbb{E}_{\boldsymbol{P}_{W|\boldsymbol{X}}}[L(W, \theta) \mid \boldsymbol{X}]].$$

In practice, a popular method of constructing classifiers is via empirical risk minimization or ERM procedures, that set $\boldsymbol{P}_{W|\boldsymbol{X}} = \delta_{\hat{W}(\boldsymbol{X})}$, where

$$\hat{W}(\boldsymbol{X}) \in \operatorname*{argmin}_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \phi(w, Z_i, Y_i).$$

3 Notions of Optimality

If our goal is to identify optimal decision rules, the above definition of risk presents a problem. In most cases, the risk curves of any two decision rules, $P_{\boldsymbol{W}|\boldsymbol{X}}$ and $Q_{\boldsymbol{W}|\boldsymbol{X}}$ are not comparable; that is, there may be regions in Θ where $R_n(P_{W|\boldsymbol{X}}, \theta) > R_n(Q_{W|\boldsymbol{X}}, \theta)$, and there may be other regions where the strict inequality is reversed. The best we can hope to do is to discard inadmissible decision rules: $Q_{W|\boldsymbol{X}}$ is said to be inadmissible if there exists a $P_{W|\boldsymbol{X}}$ such that $R_n(Q_{W|\boldsymbol{X}}, \theta) \geq R_n(P_{W|\boldsymbol{X}}, \theta)$ for all $\theta \in \Theta$, and the inequality is strict for at least one θ.

Thus, in order to compare different decision rules and find the optimal one among them, we need to develop methods of assigning numerical values that represent the quality of the decision rules. There are two common ways of doing this: the average-case or *Bayesian* approach, and the worst-case or *minimax* approach.

3.1 Bayesian Framework

In the Bayesian setting, we consider the parameter θ itself as a random variable, which we denote by the bold symbol $\boldsymbol{\theta}$ drawn according to a *prior distribution* π defined on the measurable space $(\Theta, \mathcal{F}_\Theta)$. Then, we have the Markov chain:

$$\boldsymbol{\theta} \sim \pi \xrightarrow{P_{\boldsymbol{X}|\boldsymbol{\theta}}} \boldsymbol{X} \sim \xrightarrow{P_{W|\boldsymbol{X}}} W.$$

Thus, for any decision rule $P_{W|\boldsymbol{X}}$ and prior π, we have the following Bayesian risk:

$$R_n\left(P_{W|\boldsymbol{X}}, \pi\right) = \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{X},W)}\left[L(W, \boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[R_n(P_{W|\boldsymbol{X}}, \boldsymbol{\theta})\right].$$

Thus, we now have a numerical value, called the Bayes Risk, associated with each decision rule, $P_{W|\boldsymbol{X}}$, and we can define a natural notion of the Bayes optimal decision rule as

$$P^*_{W|\boldsymbol{X}}(\pi) \in \operatorname*{argmin}_{P_{W|\boldsymbol{X}}} R_n\left(P_{W|\boldsymbol{X}}, \pi\right).$$

The decision rule $P^*_{W|X}(\pi)$ is referred to as the Bayes optimal decision rule, and our next two results characterize it.

Lemma 3.1. *For every $P_{W|\boldsymbol{X}}$ there exists a deterministic decision rule $h : \mathcal{X} \to \mathcal{W}$ such that $R_n(P_{W|\boldsymbol{X}}, \pi) \geq R_n(h, \pi)$. In other words, we can restrict our attention to deterministic decision rules without loss of optimality.*

Proof. [We will not focus on any measurability issues in this proof.] Recall that the Bayes risk of a procedure $P_{W|\boldsymbol{X}}$ is defined as

$$R_n(P_{W|\boldsymbol{X}}, \pi) = \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{X},W)}\left[L(W, \boldsymbol{\theta})\right].$$

Since $\boldsymbol{\theta} \longrightarrow \boldsymbol{X} \longrightarrow W$ form a Markov chain, we know that $\mathbb{P}_{(\theta,\boldsymbol{X},W)} = \mathbb{P}_X \mathbb{P}_{W|\boldsymbol{X}} \mathbb{P}_{\boldsymbol{\theta}|\boldsymbol{X}}$, hence we have the following:

$$R_n(P_{W|\boldsymbol{X}}, \pi) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{W|\boldsymbol{X}}\left[\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[L(W, \boldsymbol{\theta}) \mid \boldsymbol{X}\right] \mid \boldsymbol{X}\right]\right].$$

Let $\ell(X, W)$ denote the conditional expectation $\mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[L(W, \boldsymbol{\theta}) \mid \boldsymbol{X}\right]$, and define for any $x \in \mathcal{X}$,

$$h(x) \in \operatorname*{argmin}_{w \in \mathcal{W}} \ell(w, x), \quad \Longrightarrow \quad \ell(X, W) \overset{a.s.}{\geq} \ell(\boldsymbol{X}, h(\boldsymbol{X})).$$

Thus, we have

$$R_n(P_{W|\boldsymbol{X}}, \pi) = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}_{W|\boldsymbol{X}}\left[\ell(\boldsymbol{X}, W) \mid \boldsymbol{X}\right]\right] \geq \mathbb{E}_{\boldsymbol{X}}\left[\ell(\boldsymbol{X}, h(\boldsymbol{X}))\right] = R_n(h, \pi).$$

For any fixed $(\boldsymbol{\theta}, \boldsymbol{X}) = (\theta, \boldsymbol{x})$, we have

$$\mathbb{E}_W\left[L(W, \boldsymbol{\theta}) \mid \boldsymbol{X}, \boldsymbol{\theta}\right] \geq \inf_{w \in \mathcal{W}} L(w, \boldsymbol{\theta}).$$

This completes the proof. □

As an immediate corollary of the proof, we can obtain a complete characterization of a deterministic Bayes optimal decision rule.

Corollary 3.2. *A Bayes optimal decision rule is the one that minizes the average loss conditioned on the data:*

$$h_B^*(X) \in \operatorname*{argmin}_{w \in \mathcal{W}} \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[L(\boldsymbol{\theta}, w) \mid \boldsymbol{X}\right].$$

This optimal procedure has a nice interpretation: π denotes our "prior" belief over the parameter space, and the kernel $\mathbb{P}_{\boldsymbol{\theta}|\boldsymbol{X}}$ denotes our "posterior" belief over the parameter space after observing the data. Then the Bayes optimal decision rule is to observe the data, update our posterior belief, and take the action that minimizes the expected loss according to our updated posterior.

Example 3.3. *Consider the mean estimation problem of $\Theta = \mathbb{R}$, $P_\theta = N(\theta, \sigma^2)$. Let π denote a prior over the parameter space. Then, given $\boldsymbol{X}|\boldsymbol{\theta} \sim P_\theta^{\otimes n}$, the Bayes optimal estimator of the mean under quadratic loss $L(\theta, w) = |\theta - w|^2$ is*

$$\begin{aligned} h_B^*(\boldsymbol{X}) &\in \operatorname*{argmin}_{w \in \Theta} \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[|\boldsymbol{\theta} - w|^2 \mid \boldsymbol{X}\right] \\ &= \operatorname*{argmin}_{w \in \Theta} \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{X}}\left[\left(|\boldsymbol{\theta} - \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]|^2 + |\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}] - w|^2\right) \mid \boldsymbol{X}\right] \\ &= \operatorname*{argmin}_{w \in \Theta} \left(\operatorname{Var}\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right) + |\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X}] - w|^2\right) \end{aligned}$$

Since only the second term depends on w, it follows that the Bayes optimal estimator is $\mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{X}]$; the posterior mean of $\boldsymbol{\theta}$ given the data $\boldsymbol{X}$.

Furthermore, the Bayes optimal risk is equal to the expected value of the conditional posterior variance:

$$R_n(h_B^*, \pi) = \mathbb{E}_{\boldsymbol{X}}\left[\operatorname{Var}\left(\boldsymbol{\theta} \mid \boldsymbol{X}\right)\right].$$

The main criticism of the above approach is the reliance on the prior distribution π: two decision-makers may prefer different priors, π_1 and π_2, which in turn would lead them to two different "optimal" decision rules. Apriori, there is no way of distinguishing between them.

3.2 Minimax Framework

One important alternative to the Bayesian approach is to take a more pessimistic or worst-case view of the world, and design procedures that minimize the following worst-case risk called the minimax-risk:

$$R_n\left(P_{W|\boldsymbol{X}}, \Theta\right) = \sup_{\theta \in \Theta} \mathbb{E}_{(\boldsymbol{X},W)}\left[L(W,\theta)\right].$$

Again by restricting our attention to the worst-case performance of a decision rule, we have obtained a way of assigning a numerical value of performance to each such decision rule. This naturally leads us to the definition of the *minimax optimal* decision rule:

$$P^{*,M}_{W|\boldsymbol{X}} \equiv P^{*,M}_{W|\boldsymbol{X}}(\Theta) \equiv \operatorname*{argmin}_{P_{W|\boldsymbol{X}}} R_n(P_{W|\boldsymbol{X}}, \Theta), \quad \text{with minimax value} \quad R^*_n(\Theta) = \inf_{P_{W|\boldsymbol{X}}} \sup_{\theta \in \Theta} R_n(P_{W|\boldsymbol{X}}, \theta).$$

In general, finding the exact minimax optimal decision procedure is not feasible, and we often aim to identify either asymptotically minimax optimal procedures. In most practical cases, even that is not possible, and we restrict our attention to minimax rate-optimal procedures.

- A procedure $P_{W|\boldsymbol{X}}$ is said to be asymptotically minimax optimal if

$$\lim_{n\to\infty} \frac{R_n(P_{W|\boldsymbol{X}}, \Theta)}{R^*_n(\Theta)} = 1.$$

- For a positive sequence $\{\psi_n : n \geq 1\}$ is called the minimax rate if

$$-\infty < \liminf_{n\to\infty} \frac{R^*_n(\Theta)}{\psi_n} \leq \limsup_{n\to\infty} \frac{R^*_n(\Theta)}{\psi_n} < \infty.$$

  In practice, we establish rate optimality of procedures by first showing that the minimax risk of a procedure satisfies $R_n(P_{W|\boldsymbol{X}}, \Theta) \leq C\psi_n$ for some rate ψ_n, and then show an impossibility result that any procedure $Q_{W|\boldsymbol{X}}$ must suffer a minimax risk of at least $c\psi_n$.

Unlike the Bayes optimal procedure, the minimax optimal decision rule requires randomization in general.

Duality between Bayesian and Minimax Frameworks. An immediate consequence of the definitions is that

$$R_n(P_{W|\boldsymbol{X}}, \pi) = \int_\Theta R_n(P_{W|\boldsymbol{X}}, \theta) d\pi(\theta) \leq \int_\Theta \left(\sup_{\theta\in\Theta} R_n(P_{W|\boldsymbol{X}}, \theta)\right) d\pi(\theta)$$
$$\int_\Theta R_n(P_{W|\boldsymbol{X}}, \Theta) d\pi(\theta) = R_n(P_{W|\boldsymbol{X}}, \Theta). \tag{1}$$

Note that here we have used the assumption that π is a probability measure on $(\Theta, \mathcal{F}_\Theta)$, and not an improper prior. Now, observe that (1) immediately implies that

$$\sup_{\pi \in \mathcal{P}(\Theta,\mathcal{F}_\Theta)} R_n(P_{W|\boldsymbol{X}}, \pi) \leq R_n(P_{W|\boldsymbol{X}}, \Theta).$$

If the supremum above is achieved at some prior π^*, then this prior is referred to as the *least favorable prior*. This inequality is essentially the same as the notion of *weak duality* in optimization, and in analogy with optimization, under certain regularity conditions, we have the following *strong duality* condition:

$$R_n^*(\Theta) = \inf_{P_{W|\boldsymbol{X}}} \sup_{\theta} R_n(P_{W|\boldsymbol{X}}, \theta) = \inf_{P_{W|\boldsymbol{X}}} \sup_{\pi \in \mathcal{P}(\theta)} R_n(P_{W|\boldsymbol{X}}, \pi) = \sup_{\pi} \inf_{P_{W|\boldsymbol{X}}} R_n(P_{W|\boldsymbol{X}}, \pi).$$

The last inequality can be justified via a minimax theorem under certain regularity conditions.

References

J. M. Bernardo and A. F. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.