Lecture 3: Elements of Decision Theory

Thibault Randrianarisoa

UTSC

January 29, 2026

Computer & Mathematical Sciences
UNIVERSITY OF TORONTO
SCARBOROUGH

Outline

- Decision Theory Framework
- Loss Functions and Risk
  - Frequentist Risk vs. Bayesian Risk
  - Posterior Risk
- Bayes Estimators
- Comparison of Estimators
  - Admissibility
  - Minimaxity

Introduction and Motivation

In a statistical experiment, a given prior distribution corresponds to a posterior distribution. From this posterior, we can deduce several estimators (mean, median, mode, etc.).

Questions:

- Which one should we choose in practice?
- What criteria can we state for this choice?
- More generally, are there **"optimal" estimators** among all possible estimators?

To answer this, we must define notions of **Risk** and **Loss Function**. We will study three classic criteria: Admissibility, Bayes Risk, and Minimax Risk.

The Loss Function

Consider an experiment $(\mathbf{X}, \mathcal{P})$ with $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$.

Definition:

A loss function ℓ is a measurable function $\ell : \Theta \times \Theta \to \mathbb{R}_+$ such that:

$$\forall \theta, \theta' \in \Theta, \quad \ell(\theta, \theta') = 0 \iff \theta = \theta'.$$

Relaxed Definition: Sometimes we only require $\forall \theta, \theta' \in \Theta, \theta = \theta' \Rightarrow \ell(\theta, \theta') = 0$.

- This relaxed version allows including the classification loss.

Examples of Loss Functions

- **Quadratic Loss:** If $\Theta \subset \mathbb{R}$:

$$\ell(\theta, \theta') = (\theta - \theta')^2$$

  More generally, in $\Theta \subset \mathbb{R}^d$:

$$\ell(\theta, \theta') = \|\theta - \theta'\|^2 = \sum_{i=1}^{d} (\theta_i - \theta_i')^2$$

- **Absolute Loss:** If $\Theta \subset \mathbb{R}$:

$$\ell(\theta, \theta') = |\theta - \theta'|$$

Distance-based Loss Functions

For arbitrary Θ, we can define loss based on distances between probability distributions P_θ and $P_{\theta'}$.

- **Total Variation Loss:** $\ell(\theta, \theta') = d_{\mathrm{TV}}(P_\theta, P_{\theta'})$ where

$$d_{\mathrm{TV}}(P, Q) = \sup_{A \in \mathcal{E}} |P(A) - Q(A)|$$

- **Hellinger Loss:** $\ell(\theta, \theta') = h(P_\theta, P_{\theta'})$, where

$$h(P, Q)^2 = \int_E \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 d\mu(x), \quad p, q \text{ densities of } P, Q$$

⚠ *Note: These define valid loss functions (where $\ell(\theta, \theta') = 0 \iff \theta = \theta'$) only if the model is* ***identifiable****.*

Example: Classification Loss

Suppose $\Theta = \Theta_0 \cup \Theta_1$ with $\Theta_0 \cap \Theta_1 = \varnothing$.

We define the **classification loss function** by:

$$L_C(\theta, \theta') = \mathbb{1}_{\theta \in \Theta_0, \theta' \in \Theta_1} + \mathbb{1}_{\theta \in \Theta_1, \theta' \in \Theta_0}$$

- $L_C(\theta, \theta') = 0$ if and only if θ and θ' are in the same region (Θ_0 or Θ_1).
- This is the natural loss used when constructing a test to answer a binary question about θ (cf. Lecture 4).

Definition

The risk function (or simply the risk) of an estimator T for the loss function ℓ is the map $\mathbf{R}(\cdot, T) : \Theta \to \mathbb{R}_+$ defined by:

$$\theta \longmapsto \mathbf{R}(\theta, T) = \mathbf{E}_\theta[\ell(\theta, T(\mathbf{X}))] = \int_E \ell(\theta, T(x))dP_\theta(x).$$

The risk at point θ is the **average loss** of T at θ (also called *pointwise risk*), under distribution P_θ.

Risk functions allow us to compare estimators. However, defining a "best possible estimator" is delicate.

Is there a "Best" Estimator?

Consider the Gaussian model $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \theta \in \mathbb{R}\}$ and the quadratic loss.

- **Estimator 1:** The constant estimator $T = \theta_0$.
- **Estimator 2:** The sample mean $T = \bar{X}_n$.

Comparison:

- At $\theta = \theta_0$, the constant estimator has **zero risk**, making it better than any other estimator at that specific point.
- However, for all θ such that $(\theta - \theta_0)^2 > 1/n$, we prefer $\bar{X}_n$ (which has constant risk $1/n$).

$\rightarrow$ Usually, no estimator is uniformly better than all others for all θ.

Critique of the Frequentist Risk

The definition of the risk function $\mathbf{R}(\theta, \delta) = \mathbf{E}_\theta[\ell(\theta, \delta(\mathbf{X}))]$ is not without issues:

- **Frequentist Assumption:** It tacitly assumes that the problem will be encountered many times so that a frequency-based evaluation makes sense:

$$\mathbf{R}(\theta, \delta) \simeq \text{average cost over repetitions.}$$

- **Lack of Total Ordering:** As we saw earlier (e.g., constant estimator vs. sample mean), this criterion typically does not lead to a *total order* on the set of estimators.

Admissibility

Definition:

An estimator T is inadmissible if there exists an estimator T_1 such that:

$$\forall \theta \in \Theta, \quad \mathbf{R}(\theta, T_1) \leqslant \mathbf{R}(\theta, T)$$
$$\text{and } \exists \theta_1 \in \Theta, \quad \mathbf{R}(\theta_1, T_1) < \mathbf{R}(\theta_1, T).$$

An estimator T is admissible if it is not inadmissible. In other words, for any other estimator T_1, if T_1 beats T somewhere, T must beat T_1 somewhere else.

Strategies for Optimality

Since minimizing risk pointwise everywhere is impossible, we need global criteria:

1. **Bayesian Risk:**
   - Depends on a chosen prior.
   - Gives a possible answer to "optimal estimator".
   - *Drawback:* The answer is not "universal" (depends on the prior).

2. **Minimax Risk:**
   - More universal (independent of a prior).
   - *Drawback:* Pessimistic approach, we seek an estimator T that minimizes the worst possible risk

$$\inf_T \sup_{\theta \in \Theta} \mathbf{R}(\theta, T)$$

Definition

For an estimator T and a prior distribution π, the Bayes Risk is defined as:

$$\begin{aligned}\mathbf{R}_B(\pi, T) &= \mathbf{E}[\ell(\theta, T(\mathbf{X}))] \qquad \text{(in the Bayesian model)} \\ &= \int_\Theta \int_E \ell(\theta, T(x)) dP_\theta(x) d\pi(\theta) \\ &= \int_\Theta \mathbf{R}(\theta, T) d\pi(\theta) = \mathbf{E}[\mathbf{R}(\theta, T)] \qquad \text{(expectation over the prior)}\end{aligned}$$

Alternatively, by conditioning (*law of total expectation*):

$$\mathbf{E}[\ell(\theta, T(\mathbf{X}))] = \mathbf{E}\big[\mathbf{E}[\ell(\theta, T(\mathbf{X})) \mid \theta]\big] = \mathbf{E}[\mathbf{R}(\theta, T)].$$

Definition

An estimator T^* is called a Bayes estimator for the prior π if:

$$\mathbf{R}_B(\pi, T^*) = \inf_T \mathbf{R}_B(\pi, T),$$

where the infimum is taken over all possible estimators T.

We denote the minimum value as $\mathbf{R}_B(\pi) = \inf_T \mathbf{R}_B(\pi, T)$, which is called the Bayes risk for the prior π.

Interpretation: A Bayes estimator minimizes the "average risk" weighted by the prior belief π on Θ.

Example: Classification Loss

Associated Frequentist Risk:

$$\mathbf{R}(\theta, T) = \mathbf{E}_\theta[\ell(\theta, T(\mathbf{X}))] = \begin{cases} P_\theta(T(\mathbf{X}) \in \Theta_1) & \text{if } \theta \in \Theta_0 \quad \text{(Type I Error)} \\ P_\theta(T(\mathbf{X}) \in \Theta_0) & \text{otherwise} \quad \text{(Type II Error)} \end{cases}$$

The Bayes risk associated with any prior π and the classification loss is:

$$\int_{\Theta_0} P_\theta(T(\mathbf{X}) \in \Theta_1)\pi(\theta)d\theta + \int_{\Theta_1} P_\theta(T(\mathbf{X}) \in \Theta_0)\pi(\theta)d\theta$$

Example: The Gaussian Model

Setting:

- *Model:* $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}, \theta \in \mathbb{R}\}$.
- *Prior:* $\pi = \mathcal{N}(0, 1)$.
- *Loss:* Quadratic loss $\ell(\theta, \theta') = (\theta - \theta')^2$.

We calculate the Bayes risk for π for the following three estimators:

$$T_1(\mathbf{X}) = 0, \quad T_2(\mathbf{X}) = \bar{X}_n, \quad T_3(\mathbf{X}) = \frac{n}{n+1}\bar{X}_n.$$

Bayes Risk for T_1 and T_2

1. *The Constant Estimator* $T_1 = 0$:

$$\mathbf{R}_B(\pi, T_1) = \int_\Theta \mathbf{R}(\theta, T_1) d\pi(\theta)$$
$$= \int_\Theta \mathbf{E}_\theta[(\theta - 0)^2] d\pi(\theta) = \int_\Theta \theta^2 d\pi(\theta) = 1. \quad \text{(Variance of prior)}$$

2. *The Sample Mean* $T_2 = \bar{X}_n$:
Recall that under P_θ, $\bar{X}_n \sim \mathcal{N}(\theta, 1/n)$. Thus, $\mathbf{R}(\theta, T_2) = 1/n$.

$$\mathbf{R}_B(\pi, T_2) = \int_\Theta \frac{1}{n} d\pi(\theta) = \frac{1}{n}.$$

Bayes Risk for T_3

3. *The Shrinkage Estimator* $T_3 = \frac{n}{n+1}\bar{X}_n$:
First, compute the pointwise risk $\mathbf{R}(\theta, T_3)$ (Bias-Variance decomposition):

$$\mathbf{R}(\theta, T_3) = \mathbf{E}_\theta\left[\left(\frac{n}{n+1}\bar{X}_n - \theta\right)^2\right] = \mathbf{E}_\theta\left[\left(\frac{n}{n+1}(\bar{X}_n - \theta) - \frac{\theta}{n+1}\right)^2\right]$$

$$= \left(\frac{n}{n+1}\right)^2 \underbrace{\mathbf{E}_\theta[(\bar{X}_n - \theta)^2]}_{1/n} + \left(\frac{\theta}{n+1}\right)^2 \quad \text{(Cross term is 0)}$$

$$= \frac{n}{(n+1)^2} + \frac{\theta^2}{(n+1)^2}.$$

Now, integrate over the prior π (recall $\int \theta^2 d\pi = 1$):

$$\mathbf{R}_B(\pi, T_3) = \frac{n}{(n+1)^2} + \frac{1}{(n+1)^2}\int_\Theta \theta^2 d\pi(\theta) = \frac{n+1}{(n+1)^2} = \frac{1}{n+1}.$$

Comparison of Estimators

We have the following Bayes risks for prior $\pi = \mathcal{N}(0,1)$:

- $T_1 = 0 \implies \mathbf{R}_B = 1$
- $T_2 = \bar{X}_n \implies \mathbf{R}_B = \frac{1}{n}$
- $T_3 = \frac{n}{n+1}\bar{X}_n \implies \mathbf{R}_B = \frac{1}{n+1}$

For all $n \geqslant 2$:

$$\mathbf{R}_B(\pi, T_3) < \mathbf{R}_B(\pi, T_2) < \mathbf{R}_B(\pi, T_1).$$

Maximal and Minimax Risk

Before constructing Bayes estimators, let's briefly define the alternative criterion.

Definition

The maximal risk of an estimator T is:

$$\mathbf{R}_{\max}(T) = \sup_{\theta \in \Theta} \mathbf{R}(\theta, T).$$

The minimax risk $\mathbf{R}_M$ is:

$$\mathbf{R}_M = \inf_T \mathbf{R}_{\max}(T) = \inf_T \sup_{\theta \in \Theta} \mathbf{R}(\theta, T) \qquad \text{(infimum over all estimators } T\text{)}$$

An estimator T^* is minimax if $\mathbf{R}_{\max}(T^*) = \mathbf{R}_M$.

Interpretation: Minimax seeks the "least worst" estimator (*pessimistic*), while Bayes seeks the best "average" estimator.

Posterior Risk

Instead of minimizing the global Bayes risk directly, we can minimize a conditional quantity.

Definition

Let ℓ be a loss function and π a prior. The posterior risk $\rho(\pi, T \mid \mathbf{X})$ is defined as:

$$\rho(\pi, T \mid \mathbf{X}) = \mathbf{E}[\ell(\theta, T(\mathbf{X})) \mid \mathbf{X}] = \int_{\Theta} \ell(\theta, T(\mathbf{X})) d\pi(\theta \mid \mathbf{X}).$$

- Unlike the Bayes risk (which is a scalar), the posterior risk is a random variable depending on $\mathbf{X}$.
- It represents the expected loss **after observing the data**.

Theorem

Given a loss function ℓ and a prior π, if an element $T^*(\mathbf{X})$ satisfies:

$$T^*(\mathbf{X}) \in \arg\min_{T} \rho(\pi, T \mid \mathbf{X})$$

(if it exists), then T^* is a **Bayes estimator** for π.

Why is this useful?

- It simplifies the problem: instead of minimizing an integral over both $\mathcal{X}$ and Θ, we minimize the integral over Θ for each fixed $\mathbf{X}$.

$$\int_{\Theta}\int_{E} \ell(\theta, T(x))dP_{\theta}(x)d\pi(\theta) \text{ vs. } \int_{\Theta} \ell(\theta, T(\mathbf{X}))d\pi(\theta \mid \mathbf{X})$$

- We simply find the estimator that minimizes the loss pointwise for every x.

Bayes Estimator: Quadratic Loss

Consider the quadratic loss $\ell(\theta, \theta') = (\theta - \theta')^2$ with $\Theta \subset \mathbb{R}$.

Proposition

If $\int_\Theta \theta^2 d\pi(\theta) < \infty$, the Bayes estimator for quadratic loss is the **Posterior Mean**:

$$T^*(\mathbf{X}) = \mathbf{E}[\theta \mid \mathbf{X}] = \int_\Theta \theta d\pi(\theta \mid \mathbf{X}).$$

Proof Sketch: The problem reduces to finding a constant a that minimizes $\mathbf{E}[(\theta - a)^2 \mid \mathbf{X}]$. This is a classic result: the minimum of $f(a) = \mathbf{E}[(Z - a)^2]$ is achieved at $a = \mathbf{E}[Z]$. Here, Z is θ distributed according to the posterior.

Calculating Bayes Risk (Quadratic Case)

Remark

For quadratic loss, the Bayes risk $\mathbf{R}_B(\pi)$ is the **expected posterior variance**:

$$\mathbf{R}_B(\pi) = \mathbf{E}[\mathrm{Var}(\theta \mid \mathbf{X})].$$

Two ways to compute Bayes Risk:

1. Compute the risk function $\theta \mapsto \mathbf{R}(\theta, T^*)$ and integrate against the prior π.
2. (Often simpler) Compute the posterior variance $v_{\mathbf{X}} = \mathrm{Var}(\theta \mid \mathbf{X})$ and take its expectation.

In the Gaussian model, the posterior variance often *does not depend on* $\mathbf{X}$, making the calculation trivial.

Example: Gaussian Model Revisited

Model: $\mathcal{P} = \{\mathcal{N}(\theta, 1)^{\otimes n}\}$, Prior $\pi = \mathcal{N}(0, 1)$.

We saw in the previous lecture that the posterior distribution is:

$$\pi(\cdot \mid \mathbf{X}) = \mathcal{N}\left(\frac{n\bar{X}_n}{n+1}, \frac{1}{n+1}\right).$$

From the proposition:

- The Bayes estimator is the posterior mean:

$$\mathbf{E}[\theta \mid \mathbf{X}] = \frac{n}{n+1}\bar{X}_n.$$

- This confirms our earlier "guess" (T_3).
- The Bayes risk is the expectation of the posterior variance:

$$\mathbf{R}_B(\pi) = \mathbf{E}\left[\frac{1}{n+1}\right] = \frac{1}{n+1}.$$

Bayes Estimator: Absolute Loss

Consider the absolute loss $\ell(\theta, \theta') = |\theta - \theta'|$ with $\Theta \subset \mathbb{R}$.

Proposition

Let ℓ be the absolute value loss. The Bayes estimator is the **Posterior Median**:

$$T^*(\mathbf{X}) = \text{Median}(\pi(\cdot \mid \mathbf{X})).$$

Formally, $T^*(\mathbf{X}) = F_{\theta|\mathbf{X}}^{-1}(1/2)$ (generalized inverse of posterior CDF).

Intuition: Just as the mean minimizes mean squared error (L_2), the median minimizes mean absolute error (L_1).

Relationship between Bayes and Minimax Risk

We begin with a fundamental inequality relating the two major optimal risks.

Theorem

For any prior distribution π on Θ and any loss function, the Bayes risk always lower bounds the minimax risk:

$$\mathbf{R}_B(\pi) \leqslant \mathbf{R}_M.$$

Proof Idea: Recall that $\mathbf{R}_B(\pi) = \inf_T \int \mathbf{R}(\theta, T) d\pi(\theta)$. Since $\pi(\Theta) = 1$:

$$\int_\Theta \mathbf{R}(\theta, T) d\pi(\theta) \leqslant \sup_{\theta \in \Theta} \mathbf{R}(\theta, T) \int_\Theta d\pi(\theta) = \sup_{\theta \in \Theta} \mathbf{R}(\theta, T).$$

Taking the infimum over T on both sides yields the result.

Usage: This is often used to lower bound the minimax risk by finding a "least favorable" prior.

Admissibility: Sufficient Conditions

Definition

Two estimators T and T' are equivalent if their risk functions are identical:

$$\forall \theta \in \Theta, \quad \mathbf{R}(\theta, T) = \mathbf{R}(\theta, T').$$

Theorem (Unique and Bayes $\implies$ Admissible)

Let T^* be a Bayes estimator for prior π. If T^* is **unique** (up to equivalence), then T^* is **admissible**.

Proof Sketch: If T^* were inadmissible, there would exist a T with better or equal risk everywhere (and strictly better somewhere). Integrating this inequality against π would imply $\mathbf{R}_B(\pi, T) \leqslant \mathbf{R}_B(\pi, T^*)$. Since T^* is Bayes, equality must hold, and by uniqueness, T must be equivalent to T^*, contradicting strict inequality.

Admissibility in the Gaussian Model

Quadratic Loss: For the Gaussian model $\mathcal{P} = \{\mathcal{N}(\theta, \sigma^2)^{\otimes n}\}$, we obtain that Bayes estimators (for normal priors) are of the form (see Problem set 1 and Lecture 2):

$$T(\mathbf{X}) = \frac{n}{n+\lambda}\bar{X}_n + \frac{\lambda}{n+\lambda}\mu, \ \lambda > 0, \ \mu \in \mathbb{R} \quad \text{(Affine transformations)},$$

with Bayesian risk $\sigma^2/(n+\lambda)$.

Result: It can be shown that any estimator of the form:

$$\alpha \bar{X}_n + \beta, \quad \text{with } \alpha \in (0,1) \text{ and } \beta \in \mathbb{R}$$

is **admissible** for the quadratic loss.

Note: The sample mean $\bar{X}_n$ (where $\alpha = 1$ or $\lambda = 0$) is the limit of these estimators but requires a different tool (limit of priors) to analyze its minimaxity.

Finding Minimax Estimators: Constant Risk

A powerful method to identify minimax estimators is to look for those with **constant risk**.

Proposition

If an estimator T is **admissible** and has **constant risk** (i.e., $\theta \mapsto \mathbf{R}(\theta, T)$ is constant), then T is **minimax**.

Theorem

If T is a **Bayes estimator** for a prior π and has **constant risk**, then T is **minimax**.

Intuition: A Bayes estimator minimizes the *average* risk. If the risk is flat (constant), the average is equal to the maximum. Thus, it minimizes the maximum risk.

The Limiting Bayes Method

Sometimes the minimax estimator is not a Bayes estimator for any proper prior (e.g., $\bar{X}_n$ in the Gaussian model). We use sequences of priors.

Theorem

If there exists a sequence of priors $(\pi_k)_{k \geqslant 1}$ such that:

$$\mathbf{R}_{\max}(T) = \lim_{k \to \infty} \mathbf{R}_B(\pi_k),$$

then T is **minimax**.

Proof: We know $\mathbf{R}_B(\pi_k) \leqslant \mathbf{R}_M \leqslant \mathbf{R}_{\max}(T)$. Taking the limit:

$$\lim \mathbf{R}_B(\pi_k) \leqslant \mathbf{R}_M \leqslant \mathbf{R}_{\max}(T).$$

If the ends are equal, then $\mathbf{R}_{\max}(T) = \mathbf{R}_M$.

Application: Minimaxity of $\bar{X}_n$

Setting: Gaussian Model $\mathcal{N}(\theta, 1)^{\otimes n}$ with quadratic loss.

1. **Estimator:** Consider $T = \bar{X}_n$.
2. **Maximal Risk:** $\mathbf{R}(\theta, \bar{X}_n) = 1/n$ for all θ. Thus, $\mathbf{R}_{\max}(\bar{X}_n) = 1/n$.
3. **Sequence of Priors:** Let $\pi_{\sigma^2} = \mathcal{N}(0, \sigma^2)$.
4. **Bayes Risk:** We calculated previously that $\mathbf{R}_B(\pi_{\sigma^2}) = \frac{1}{n+\sigma^{-2}}$.

Conclusion:

$$\lim_{\sigma^2 \to \infty} \mathbf{R}_B(\pi_{\sigma^2}) = \lim_{\sigma^2 \to \infty} \frac{1}{n + \sigma^{-2}} = \frac{1}{n} = \mathbf{R}_{\max}(\bar{X}_n).$$

By the Theorem, $\bar{X}_n$ **is a minimax estimator**.

Bayes Estimators under Frequentist Criteria

Proposition

If a Bayes estimator, constructed from a prior $\pi(\theta)$, is associated with a strictly convex cost function, then it is **admissible**.

A Frequentist Perspective:

- Criteria such as **minimaxity** and **admissibility** are fundamentally *frequentist* (as they are built from the frequentist risk).
- According to these standards, Bayesian estimators perform better than, or at least as well as, standard frequentist estimators:
  - Their minimax risk is often equal or smaller.
  - They are often all **admissible** (provided the Bayes risk is well-defined).