Lecture-12: Statistical decision theory

1 Setting

Definition 1.1. Consider a measurable space $(\Omega,\mathcal{F})$, a parameter space Θ, and a set of probability distributions $\mathcal{M}(\Theta) \triangleq \left\{P \in [0,1]^{\mathcal{B}(\Theta)} : P \text{ satisfies probability axioms}\right\}$. Let $P_\theta \in \mathcal{M}(\Theta)$ be a probability distribution for each $\theta \in \Theta$. A statistical model refers to a collection of probability distributions denoted as

$$\mathcal{P}(\Theta) \triangleq \{P_\theta \in \mathcal{M}(\Theta) : \theta \in \Theta\}. \tag{1}$$

Remark 1. Without loss of generality, all statistical models can be expressed in the parametric form of (1).

Definition 1.2. A statistical model is called *parametric* if Θ is a finite-dimensional Euclidean space so that each distribution is specified by finitely many parameters, and *nonparametric* if Θ is an infinite-dimensional space e.g. density estimation or sequence model.

Assumption 1.3. Let $\mathcal{X},\mathcal{Y},\Theta$ be the input, output, and parameter spaces respectively. Let estimand map be $T : \Theta \to \mathcal{Y}$ and $\mathcal{P}(\Theta)$ be a statistical model parametrized over parameter space Θ. The observation random variable $X : \Omega \to \mathcal{X}$ is assumed to be $\mathcal{F}$ measurable and generated by distribution $P_\theta \in \mathcal{P}(\Theta)$ and the goal is to estimate $T(\theta)$ based on the observation X.

Example 1.4. Some examples of *estimand* $T(\theta)$ are $\theta, \mathbb{1}_{\{\theta>0\}}, \text{sign}(\theta)$, or $\|\theta\|_p$ for some $p \geqslant 1$. If $\Theta \subseteq \mathbb{R}^d$, then an interesting estimand is $T(\theta) \triangleq \max\{\theta_i : i \in [d]\}$. If $\Theta \subseteq \mathbb{R}^{d\times d}$, then an interesting estimand is $T(\theta) \triangleq \max\{\lambda_i : i \in [d]\}$ where $(\lambda_1,\dots,\lambda_d)$ are eigenvalues of θ.

Example 1.5 (Binary classification). Let $\mathcal{X} \triangleq \mathbb{R}^d, \Theta = \mathcal{Y} \triangleq \{-1,1\}$, estimand $T(\theta) = \theta$, and an independent labeled training sample $z \in (\mathcal{X}\times\mathcal{Y})^m$. Defining $I_\theta \triangleq \{i \in [m] : y_i = \theta\}$, we observe that $(I_\theta : \theta \in \Theta)$ partitions $[m]$. For any parameter $\theta \in \Theta$, we define random vector $x_{I_\theta} \triangleq (x_i : i \in I_\theta)$, which is *i.i.d.* with a common distribution P_θ. We note that this assumption is different than assuming $x \in \mathcal{X}^m$ is *i.i.d.* . However, assuming a prior distribution $\pi \in [0,1]^{\mathcal{B}(\Theta)}$ such that

$$D(x) \triangleq \mathbb{E}_{\theta\sim\pi} P_\theta(x) = \int_{\theta\in\Theta} d\pi(\theta) P_\theta(x), \text{ for each } x \in \mathcal{X},$$

we can assume that $x \in \mathcal{X}^m$ is *i.i.d.* with common distribution $D \in \mathcal{M}(\mathcal{X})$.

Definition 1.6. An estimator is a function $\hat{T} : \mathcal{X} \times [0,1] \to \mathcal{Y}'$, where prediction space $\mathcal{Y}'$ need not be same as output space $\mathcal{Y}$.

Remark 2. The estimator $\hat{T}$ is a map such that $(X,U) \mapsto \mathcal{Y}'$ where $U : \Omega \to [0,1]$ is a random variable independent of observation X and models external randomness. In general, the estimator is random. A deterministic estimator doesn't depend on the uniform random variable U.

Example 1.7. $\hat{T}$ may be a confidence interval that aims to contain the scalar $T(\theta)$.

Example 1.8 (Binary classification). We take $\mathcal{Y}' = \mathcal{Y}$ and define linear estimator $\hat{T} : \mathcal{X} \times [0,1] \to \mathcal{Y}$ as $\hat{T}(x,u) \triangleq \text{sign}\langle w, x\rangle$ for some $w \in \mathbb{R}^d$. This is a deterministic estimator and not depending on the external randomness.

Definition 1.9. To measure the quality of an estimator $\hat{T}$, we introduce a loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ such that $(T(\theta), \hat{T}(X,U)) \mapsto L(T(\theta), \hat{T}(X,U))$ is the risk of $\hat{T}$ estimating T at parameter θ.

Remark 3. Since we are dealing with loss, all the negative or converse results are lower bounds and all the positive or achievable results are upper bounds. Note that X is a random variable, so is the estimate $\hat{T}(X,U)$ and the loss $L(T(\theta), \hat{T}(X,U))$.

Definition 1.10. The risk of estimator $\hat{T}$ at a parameter θ under loss L is defined as

$$R_\theta(T, \hat{T}) \triangleq \mathbb{E}[L(T(\theta), \hat{T}(X,U)) \mid \theta] = \int dP_\theta(x) \int_{u \in [0,1]} L(T(\theta), \hat{T}(x,u)) du.$$

Example 1.11 (Binary classification for general parameter space). We take the following parameter space, estimation goal, estimator, and loss functions for $\Theta_0 \cap \Theta_1 = \emptyset$,

$$\Theta \triangleq \Theta_0 \cup \Theta_1, \qquad T(\theta) \triangleq \mathbb{1}_{\Theta_1}(\theta), \qquad \mathcal{Y}' \triangleq \{0,1\}, \qquad L(T(\theta), \hat{T}(X,U)) \triangleq \mathbb{1}_{\{T(\theta) \neq \hat{T}(X,U)\}}.$$

Denoting the random set $\Theta_{\hat{T}} \triangleq \left\{\theta \in \Theta : \hat{T}(X,U) = T(\theta)\right\}$, we can write the expected risk as $R_\theta(T, \hat{T}) = P_\theta \left\{\theta \notin \Theta_{\hat{T}}\right\}$ the probability of error.

Example 1.12 (Confidence interval estimation). Consider the problem of inference where the goal is to output a confidence interval or region which covers the true parameter with high probability. In this case, $\mathcal{Y} = \Theta = \mathcal{X} \subseteq \mathbb{R}^d$, estimand $T(\theta) = \theta$, and estimate $\hat{T} : \mathcal{X} \times [0,1] \to \mathcal{Y}' \triangleq \mathcal{P}(\Theta)$. Estimate $\hat{\theta} \triangleq \hat{T}(X,U) \subseteq \Theta$ is a subset of parameter Θ for observation X and external randomness $U : \Omega \to [0,1]$. For loss function $L : \mathcal{Y} \times \mathcal{Y}'$ defined as $L(T(\theta), \hat{T}(X,U)) = L(\theta, \hat{\theta}) \triangleq \mathbb{1}_{\{\theta \in \hat{\theta}\}} + \lambda \left|\hat{\theta}\right|$ where $\left|\hat{\theta}\right|$ is the volume of region $\hat{\theta}$ and $\lambda > 0$ is some regularization parameter.

Remark 4 (Randomized versus deterministic estimators). Although most of the estimators used in practice are deterministic, there are a number of reasons to consider randomized estimators.

(a) For certain formulations, such as the minimizing worst-case risk (minimax approach), deterministic estimators are suboptimal and it is necessary to randomize. On the other hand, if the objective is to minimize the average risk (Bayes approach), then it does not lose generality to restrict to deterministic estimators.

(b) The space of randomized estimators (viewed as Markov kernels) is convex which is the convex hull of deterministic estimators. This convexification is needed for example for the treatment of minimax theorems.

Lemma 1.13. *When loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ is convex in the second argument, the best estimator is deterministic.*

Proof. We denote $y \triangleq T(\theta)$ and $\hat{Y} \triangleq \hat{T}(X,U)$. From conditional Jensen's inequality applied to the second argument of function L, it follows that $R_\theta(T, \hat{T}) = \mathbb{E}[L(y, \hat{Y}) \mid \theta] \geqslant \mathbb{E}[L(y, \mathbb{E}[\hat{Y} \mid X, \theta]) \mid \theta]$. □

Remark 5. For any randomized estimator $\hat{T}(X,U)$, we can derandomize it by considering its conditional expectation $\mathbb{E}[\hat{T}(X,U) \mid X]$, which is a deterministic estimator. For convex loss functions, the risk for deterministic estimator dominates that of the random estimator at every parameter θ.

2 Gaussian location model (GLM)

Definition 2.1 (Gaussian location model (GLM) or normal mean model). Consider parameter space $\Theta \subseteq \mathbb{R}^d$ where I_d denotes the d-dimensional identity matrix. *Gaussian location model (GLM)* is the collection of d-dimensional Gaussian distributions parameterized by mean θ and variance σ^2, and denoted

$$\mathcal{P}(\Theta) \triangleq \left\{\mathcal{N}(\theta, \sigma^2 I_d) : \theta \in \Theta\right\}.$$

Remark 6. For an observation $X : \Omega \to \mathbb{R}^d$ generated by GLM on parameter space $\Theta \subseteq \mathbb{R}^d$, we can write the observation as $X = \theta + Z$ where Z is a zero-mean Gaussian random variable $\mathcal{N}(0, \sigma^2 I_d)$.

Example 2.2 (Parametric spaces $\Theta \subseteq \mathbb{R}^d$ for GLM). Following are some of the examples.

(a) **Unconstrained**. $\Theta = \mathbb{R}^d$.

(b) **ℓ_p-norm balls.** $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_p \leqslant 1\right\}$.

(c) **k-sparse vectors.** $\Theta = \left\{\theta \in \mathbb{R}^d : \|\theta\|_0 \leqslant k\right\}$ where $\|\theta\|_0 \triangleq |\{i \in [d] : \theta_i \neq 0\}|$ is the size of the support of θ.

(d) **r-rank matrices.** $\Theta = \left\{\theta \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}\theta \leqslant r\right\}$. A matrix $\theta \in \mathbb{R}^{d_1 \times d_2}$ can be vectorized into a $d = d_1 \times d_2$ dimensional vector.

Example 2.3 (Loss functions and estimators for GLM). Let $\mathcal{Y} = \mathcal{Y}' = \Theta$ where $T(\theta) = \theta$ and denote $\hat{\theta} = \hat{T}(X,U)$. We consider following loss functions $L : \Theta \times \Theta \to \mathbb{R}_+$ defined for all $(\theta,\hat{\theta}) \in \Theta \times \Theta$.

(a) A p-norm loss function defined as $L(\theta,\hat{\theta}) \triangleq \left\|\theta - \hat{\theta}\right\|_p^\alpha$ for $p \geqslant 1$ and $\alpha > 0$.

(b) Log-likelihood loss function defined as $L(\theta,\hat{\theta}) \triangleq -\ln P_{\hat{\theta}}$, and the resulting estimator that minimizes the loss is called the maximum likelihood estimator (MLE) and denoted by $\hat{\theta}_{\mathrm{ML}}$. From the definition of the log-likelihood loss function, minimizing loss is equivalent to maximizing log likelihood of sample X, which for GLM is given by

$$\ln P_{\hat{\theta}} = -\frac{d}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\|X - \hat{\theta}\right\|^2.$$

We observe that $\hat{\theta}_{\mathrm{ML}} = X$ maximizes the log-likelihood for GLM.

(c) The resulting estimator based on shrinkage is called the James-Stein estimator $\hat{\theta}_{\mathrm{JS}} \triangleq \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right)X$.

3 Bayes and minimax risk

Definition 3.1 (Simple setting). For notational simplicity, we consider the task of estimating $T(\theta) = \theta$, such that $\mathcal{Y} = \mathcal{Y}' = \Theta$ and $\hat{\theta} \triangleq \hat{T}(X,U)$ for observation $X : \Omega \to \mathcal{X}$ with distribution P_θ and external randomness $U : \Omega \to [0,1]$ independent of everything else.

The risk $R_\theta(\hat{\theta})$ of an estimator $\hat{\theta}$ depends on the ground truth θ. To choose an estimator, we need to compare the risk profiles of different estimators meaningfully.

Definition 3.2 (Inadmissible estimator). Consider two estimators $\hat{\theta}_1, \hat{\theta}_2$ such that $R_\theta(\hat{\theta}_1) \leqslant R_\theta(\hat{\theta}_2)$ pointwise for all θ, then $\hat{\theta}_2$ is *inadmissible*.

However, if two estimators $\hat{\theta}_1, \hat{\theta}_2$ do not dominate each other point wise, then the comparison is not clear. For example, consider the case when peak of risk $\hat{\theta}_2$ is bigger than the peak of risk $\hat{\theta}_1$, however the average risk of $\hat{\theta}_2$ is smaller than the average risk of $\hat{\theta}_1$. From worst-case (minimax) view, $\hat{\theta}_1$ is a better estimator, whereas from average-case (Bayesian) view, $\hat{\theta}_2$ is a better estimator.

3.1 Bayes risk

Definition 3.3 (Bayes risk). Let π be a *prior* probability distribution on Θ. Then the average risk with respect to π of an estimator $\hat{\theta}$ is defined as $R_\pi(\hat{\theta}) \triangleq \mathbb{E}_{\theta\sim\pi} R_\theta(\hat{\theta}) = \mathbb{E}[\mathbb{E}[L(\theta,\hat{\theta}) \mid \theta]]$. Given a prior π, Bayes risk of estimator $\hat{\theta}$ is the minimal average risk $R_\pi^* \triangleq \inf_{\hat{\theta}} R_\pi(\hat{\theta})$. An estimator $\hat{\theta}_B$ is called a Bayes estimator if it attains the Bayes risk $R_\pi^* = \mathbb{E}_{\theta\sim\pi}[R_\theta(\hat{\theta}_B)]$.

Lemma 3.4. *Bayes estimator is always deterministic for any loss function.*

Proof. Consider any randomized estimator $\hat{\theta}(U) \triangleq \hat{T}(X,U)$ where U is external randomness independent of observation X and parameter θ. Then, its risk is lower bounded by

$$R_\pi(\hat{\theta}) = \mathbb{E}L(\theta,\hat{T}(X,U)) = \mathbb{E}[\mathbb{E}[L(\theta,\hat{\theta}(U)) \mid U]] = \mathbb{E}R_\pi(\hat{\theta}(U)) \geqslant \inf_u R_\pi(\hat{\theta}(u)).$$

□

Exercise 3.5 (Bayes risk for square loss function). Consider the statistical decision theory simple setting with unconstrained parameter set $\Theta \triangleq \mathbb{R}^d$, input space $\mathcal{X} = \Theta$, a prior $\pi \in \mathcal{M}(\Theta)$, and the quadratic loss $L : (\theta,\hat{\theta}) \mapsto \left\|\theta - \hat{\theta}\right\|^2$.

(a) Show that the best Bayes estimator is deterministic for any loss function. Consequently, it suffices to focus on deterministic estimators $\hat{T}(X)$.

(b) Show that for any estimator $\hat{T}(X)$, we have $\mathbb{E}[(\theta - \mathbb{E}[\theta \mid X])\hat{T}(X)] = 0$.

(c) Show that the Bayes estimator for quadratic loss is $\hat{T}_B(X) \triangleq \mathbb{E}[\theta \mid X]$.

(d) Show that the Bayes risk is $\mathbb{E}[\mathrm{tr}(\mathrm{cov}(\theta \mid X))]$.

Exercise 3.6 (Bayes risk for GLM). Consider the statistical decision theory simple setting with unconstrained parameter space $\Theta \triangleq \mathbb{R}^d$ and input space $\mathcal{X} = \Theta$. For GLM, the observation $X \triangleq \theta + Z$, where Z is independent of θ and has a zero-mean Gaussian distribution $\mathcal{N}(0,\sigma^2 I_d)$. Consider a Gaussian prior $\pi \in \mathcal{M}(\mathcal{X})$ with zero mean and covariance matrix sI_d.

(a) Given the observation X, derive the posterior distribution $P_{\theta|X}$.

(b) Find the Bayes estimator and Bayes risk for quadratic loss function $L : \theta \times \hat{\theta} \mapsto \left\|\theta - \hat{\theta}\right\|^2$.

3.2 Minimax risk

A common criticism of the Bayesian approach is the arbitrariness of the selected prior. Instead, we take a frequentist viewpoint by considering the worst-case situation.

Definition 3.7 (Minimax risk). The *minimax risk* is defined as $R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$. If there exists $\hat{\theta}_m$ such that $\sup_{\theta\in\Theta} R_\theta(\hat{\theta}_m) = R^*$, then the estimator $\hat{\theta}_m$ is *minimax optimal*.

Remark 7. Let $\epsilon > 0$. Finding the value of the minimax risk R^* entails showing the following.

(a) **A minimax upper bound.** Find the minimax estimator $\hat{\theta}_m$ such that $\sup_{\theta\in\Theta} R_\theta(\hat{\theta}_m) \leqslant R^* + \epsilon$.

(b) **A minimax lower bound.** For any estimator $\hat{\theta}$, find a parameter $\theta \in \Theta$ such that $R_\theta(\hat{\theta}) \geqslant R^* - \epsilon$.

Remark 8. Often this task is difficult, especially in high dimensions. Instead of the exact minimax risk, it is often useful to find a constant-factor approximation Ψ, which we call *minimax rate*, such that $R^* \asymp \Psi$, i.e. $c\Psi \leqslant R^* \leqslant C\Psi$ for some universal constants $c, C \geqslant 0$. Establishing Ψ is the minimax rate still entails proving the minimax upper and lower bounds, albeit within multiplicative constant factors.

Remark 9. In practice, minimax lower bounds are rarely established according to the original definition. The next result shows that the Bayes risk is always lower than the minimax risk. All lower bound techniques essentially boil down to evaluating the Bayes risk with a sagaciously chosen prior.

Theorem 3.8. *Minimax risk is lower bounded by the worst Bayes risk, i.e.* $R^* \geqslant R_B^* \triangleq \sup_{\pi\in\mathcal{P}(\Theta)} R_\pi^*$. *If the supremum is attained for some prior, we say it is* least favorable.

Proof. Following are two equivalent ways to prove this fact.

(a) **max is greater than mean.** For any estimate $\hat{\theta}$ and prior π, we have average risk $R_\pi(\hat{\theta}) = \mathbb{E}R_\theta(\hat{\theta}) \leqslant \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$. Taking the infimum over $\hat{\theta}$ on both sides completes the proof.

(b) **min max greater than max min.** Recall that for any $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we have $\min_x \max_y f(x,y) \geqslant \max_y \min_x f(x,y)$. It follows that

$$R^* = \inf_{\hat{\theta}} \sup_{\theta} R_\theta(\hat{\theta}) = \inf_{\hat{\theta}} \sup_{\pi\in\mathcal{P}(\Theta)} R_\pi(\hat{\theta}) \geqslant \sup_{\pi\in\mathcal{P}(\Theta)} \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = \sup_{\pi\in\mathcal{P}(\Theta)} R_\pi^*.$$

□

Example 3.9 (Minimax risk is minimized by randomized estimators). Unlike Bayes estimators which are always deterministic, to minimize the worst-case risk it is sometimes necessary to randomize for example in the context of hypotheses testing. Specifically, consider a trivial experiment where parameter space $\Theta \triangleq \{0,1\}$ and there is no observation X, so that we are forced to guess the value of $\theta \in \Theta$ under the zero-one loss $L(\theta,\hat{\theta}) = \mathbb{1}_{\{\theta\neq\hat{\theta}\}}$. Consider an estimator $\hat{\theta}(U) \triangleq \mathbb{1}_{\{U>1-p\}}$

for external randomness $U : \Omega \to [0,1]$ uniformly distributed. We observe that $\hat{\theta}$ is a Bernoulli random variable with probability $P\left\{\hat{\theta} = 1\right\} = p$, such that $R_\theta(\hat{\theta}) = \bar{p}\theta + \bar{\theta}p$, and $\sup_\theta R_\theta(\hat{\theta}) = \bar{p} \vee p$. Infimum over all estimators is the infimum over all probabilities p, and we can find the minimax risk

$$R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta} R_\theta(\hat{\theta}) = \inf_{p} \sup_{\theta} \bar{p}\theta + \bar{\theta}p = \inf_{p} \bar{p} \vee p = \frac{1}{2}.$$

That is, the minimax risk $\frac{1}{2}$ is achieved by random guessing $\hat{\theta}$ with uniform Bernoulli distribution but not by any deterministic $\hat{\theta}$.