Lecture-17: Minimax theorem

1 Minimax theorem

Consider the statistical decision theory simple setting, where the estimator $\hat{\theta}$ takes values in the action space $\hat{\Theta}$ with a loss function $\ell : \Theta \times \hat{\Theta} \to \mathbb{R}$. A very general result asserts that $R^* = R^*_B$, provided that the following condition hold.

1. The experiment is dominated, i.e., $P_\theta \ll \nu$ holds for all $\theta \in \Theta$ and for for some $\nu \in \mathcal{M}(\mathcal{X})$.
2. The action space $\hat{\Theta}$ is a locally compact topological space with a countable base e.g. the Euclidean space.
3. The loss function is level-compact i.e., for each $\theta \in \Theta, \ell(\theta, \cdot)$ is bounded from below and the sub-level set $\left\{\hat{\theta} \in \hat{\Theta} : \ell(\theta, \hat{\theta}) \leqslant a\right\}$ is compact for each $a \in \mathbb{R}$.

This result shows that for virtually all problems encountered in practice, the minimax risk coincides with the least favorable Bayes risk. At the heart of any minimax theorem, there is an application of the separating hyperplane theorem. Below we give a proof of a special case illustrating this type of argument.

Definition 1.1. Let parameter space Θ be a finite set, and $\mathbb{R}^\Theta$ denote the Euclidean space of real-valued vectors. Given an estimator $\hat{\theta}$, denote its risk vector $R(\hat{\theta}) \triangleq (R_\theta(\hat{\theta}) : \theta \in \Theta)$. We define

$$S \triangleq \left\{R(\hat{\theta}) \in \mathbb{R}^\Theta : \hat{\theta} \text{ is a randomized estimator}\right\}, \qquad T \triangleq \left\{t \in \mathbb{R}^\Theta : t_\theta < R^*, \theta \in \Theta\right\}.$$

The average risk $R_\pi(\hat{\theta})$ with respect to a prior $\pi \in \mathcal{M}(\Theta)$ is given by the inner product $R_\pi(\hat{\theta}) \triangleq \left\langle \pi, R(\hat{\theta}) \right\rangle$.

Remark 1. Recall that Bayes risk $R^*_\pi \triangleq \inf_{\hat{\theta}} R_\pi(\theta) = \inf_{\hat{\theta}} \left\langle \pi, R(\hat{\theta}) \right\rangle$ for a prior $\pi \in \mathcal{M}(\Theta)$. From the definition of S, we get $R^*_\pi(\hat{\theta}) = \inf_{s \in S} \langle \pi, s \rangle$. Further, from the definition of T, we obtain $R^* > \langle \pi, t \rangle$ for any $t \in T$. It follows that $\sup_{t \in T} \langle \pi, t \rangle = R^*$.

Lemma 1.2. *The sets S, T defined in Definition 1.1 are convex and disjoint.*

Proof. We will show the convexity and disjointness separately.

(a) **Convexity.** Let $\lambda \in [0,1]$, and $\hat{\theta}_1 \triangleq \hat{T}_1(X), \hat{\theta}_2 \triangleq \hat{T}_2(X)$ be two randomized estimators, then we can define another randomized estimator $\hat{\theta}(X)$ for an independent external randomness $U : \Omega \to [0,1]$, as

$$\hat{\theta}(X) \triangleq \hat{\theta}_1 \mathbb{1}_{\{U \leqslant \lambda\}} + \hat{\theta}_2 \mathbb{1}_{\{U > \lambda\}}.$$

It follows that $R_\theta(\hat{\theta}) \in S$, and the convexity of S follows from the following observation,

$$R_\theta(\hat{\theta}) = \mathbb{E}[\ell(\theta, \hat{\theta}) \mid \theta] = \lambda R_\theta(\hat{\theta}_1) + \bar{\lambda} R_\theta(\hat{\theta}_2).$$

Similarly, we take $t^1, t^2 \in T$ and hence $t^i_\theta < R^*$ for all $\theta \in \Theta$. It follows that $\lambda t^1_\theta + \bar{\lambda} t^2_\theta < R^*$ for all $\theta \in \Theta$. Hence $\lambda t^1 + \bar{\lambda} t^2 \in T$, showing the convexity of T.

(b) **Disjointness.** Recall the definition of minimax risk $R^* \triangleq \inf_{\hat{\theta}} \sup_\theta R_\theta(\hat{\theta})$. Fix $\epsilon > 0$. Then, for any estimator $\hat{\theta}$, there exists $\theta \in \Theta$ such that $R_\theta(\hat{\theta}) > R^* - \epsilon$. Since the choice of $\epsilon > 0$ is arbitrary, it follows that $R(\hat{\theta}) \notin T$ for any estimator $\hat{\theta}$, and hence $S \cap T = \emptyset$.

□

Theorem 1.3 (Minimax theorem). *Let Θ be a finite set, then $R^* = R^*_B$ in either of the following cases.*
(a) Input space $\mathcal{X}$ is finite.
(b) The loss function ℓ is bounded from below, i.e., $\inf_{\theta, \hat{\theta}} \ell(\theta, \hat{\theta}) > -\infty$.

Proof. We will show for both of the conditions.

(a) When the input space $\mathcal{X}$ is finite, the equality follows directly from the duality interpretation of minimax and Bayes risk and the fact that strong duality holds for finite-dimensional linear programming.

(b) We start by showing that if $R^* = \infty$, then $R^*_B = \infty$. To see this, consider the uniform prior $\pi \in \mathcal{M}(\Theta)$ and $M \in \mathbb{N}$. Then for any estimator $\hat{\theta}$, there exists $\theta \in \Theta$ such that $R_\theta(\hat{\theta}) \geqslant M$. It follows that $R_\pi(\hat{\theta}) \geqslant \frac{1}{|\Theta|} R_\theta(\hat{\theta}) \geqslant \frac{M}{|\Theta|}$. Since the choice of M was arbitrary, the result follows. Therefore, we can assume that $R^* < \infty$ without any loss of generality. From theorem hypothesis ℓ is bounded from below, and hence $R^* \in \mathbb{R}$. From Lemma 1.2, we observe that the sets S, T of Definition 1.1 are convex and disjoint. Applying the separating hyperplane theorem to S and T, there exists a separating hyperplane (π, b) where non-zero $\pi \in \mathbb{R}^\Theta$ and $b \in \mathbb{R}$ such that $\inf_{s \in S} \langle \pi, s \rangle + b \geqslant \sup_{t \in T} \langle \pi, t \rangle + b$. That is, there exists a $c \in \mathbb{R}$, such that $\inf_{s \in S} \langle \pi, s \rangle \geqslant c \geqslant \sup_{t \in T} \langle \pi, t \rangle$. We observe that π must be componentwise positive, otherwise $\sup_{t \in T} \langle \pi, t \rangle = \infty$ contradicting the finite upper bound c. Normalizing π, we can assume that $\pi \in \mathcal{M}(\Theta)$, a prior on Θ. The result follows from the observation that

$$R^*_B \geqslant R^*_\pi = \inf_{s \in S} \langle \pi, s \rangle \geqslant \sup_{t \in T} \langle \pi, t \rangle = R^*.$$

□

1.1 Multiple observations and sample complexity

Definition 1.4 (Independent sampling model). Given $m \in \mathbb{N}$ and an experiment or statistical model $\mathcal{P}(\Theta) \triangleq \{P_\theta \in \mathcal{M}(\mathcal{X}) : \theta \in \Theta\}$, the *independent sampling model* is the experiment or statistical model $\mathcal{P}_m(\Theta) \triangleq \left\{P_\theta^{\otimes m} \in \mathcal{M}(\mathcal{X}^m) : \theta \in \Theta\right\}$. In this experiment, observation sample $X : \Omega \to \mathcal{X}^m$ is an *i.i.d.* random vector drawn from $P_\theta \in \mathcal{M}(\mathcal{X})$ for some $\theta \in \Theta$.

Definition 1.5. Given a loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$, the minimax risk for simple setting is denoted by

$$R^*_m(\Theta) \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\theta, \hat{\theta}) \mid \theta].$$

Remark 2. It follows that $m \mapsto R^*_m(\Theta)$ is a non-increasing map. Typically, $\lim_{m\to\infty} R^*_m(\Theta) = 0$ for a fixed $\Theta \subseteq \mathbb{R}^d$. A natural question to ask is the rate of convergence of minimax risk as a function of sample size m.

Definition 1.6 (Parametric rate). In the classical large-sample asymptotics, the rate of convergence for the quadratic risk is usually of order $\Theta(\frac{1}{m})$, which is commonly referred to as the *parametric rate*.

Definition 1.7 (Sample complexity). The minimum sample size to attain a minimax risk of $\epsilon > 0$ is called *sample complexity* and denoted by $m^*(\epsilon) \triangleq \min\left\{m \in \mathbb{N} : R^*_m(\Theta) \leqslant \epsilon\right\}$.

Example 1.8 (GLM). Consider GLM statistical model under simpler setting with unconstrained parameter space $\Theta \triangleq \mathbb{R}^d$, observation space $\mathcal{X} = \Theta$, identity matrix I_d in d dimensions, and *i.i.d.* sample $X : \Omega \to \mathcal{X}^m$ with common Gaussian distribution $\mathcal{N}(\theta, \sigma^2 I_d)$. We note that $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i$ is a sufficient statistic of X for θ, and therefore the model reduces to a single observation $\bar{X}$ that has a Gaussian distribution $\mathcal{N}(\theta, \frac{\sigma^2}{m} I_d)$. The minimax quadratic risk for this single Gaussian observation is $\frac{d\sigma^2}{m}$. We conclude that the sample complexity is $m^*(\epsilon) = \left\lceil \frac{d\sigma^2}{\epsilon} \right\rceil$, which grows linearly with the dimension d.

Exercise 1.9 (Sample complexity as a function of dimensions). Consider the matrix case $\Theta \triangleq \mathbb{R}^{d\times d}$ with m independent observations in zero mean unit variance Gaussian noise, and let ϵ be a small constant. Then we have

(a) For quadratic loss, namely, $\left\|\theta - \hat{\theta}\right\|_F^2$, we have $R^*_m = \frac{d^2}{m}$ and hence $m^*(\epsilon) = \Theta(d^2)$.

(b) If the loss function is $\left\|\theta - \hat{\theta}\right\|_{\text{op}}^2$ then $R^*_m \asymp \frac{d}{m}$ and hence $m^*(\epsilon) = \Theta(d)$.

(c) If $T(\theta) \triangleq \max_{i \in [d]} \theta_i$, then $m^*(\epsilon) = \Theta(\sqrt{\ln d})$.

1.2 Tensor product of experiments

Tensor product is a way to define a high-dimensional model from low-dimensional models.

Definition 1.10. For each $i \in [d]$, consider parameter space Θ_i, input space $\mathcal{X}_i$ generated by statistical experiment $\mathcal{P}_i \triangleq \{\mathcal{P}_{\theta_i} : \theta_i \in \Theta_i\}$, label space $\mathcal{Y}_i$ generated by estimand map $T_i : \Theta_i \to \mathcal{Y}_i$, prediction space $\mathcal{Y}'_i$ generated by estimator $\hat{T}_i : \Omega \to (\mathcal{Y}'_i)^{\mathcal{X}_i}$, and the corresponding loss function $\ell_i : \mathcal{Y}_i \times \mathcal{Y}'_i \to \mathbb{R}$. We respectively define the tensor product of parameter, input, label, prediction spaces, and statistical experiments, as

$$\Theta \triangleq \prod_{i\in[d]} \Theta_i, \qquad \mathcal{X} \triangleq \prod_{i\in[d]} \mathcal{X}_i, \qquad \mathcal{Y} \triangleq \prod_{i\in[d]} \mathcal{Y}_i, \qquad \mathcal{Y}' \triangleq \prod_{i\in[d]} \mathcal{Y}'_i, \qquad \mathcal{P} \triangleq \left\{ P_\theta \triangleq \prod_{i=1}^d P_{\theta_i} : \theta \in \Theta \right\}.$$

The corresponding tensor product of estimand $T : \Theta \to \mathcal{Y}$, estimator $\hat{T} : \Omega \to (\mathcal{Y}')^{\mathcal{X}}$, and loss function $\ell : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$ are defined respectively for all $\theta \in \Theta$, observation $X \in \mathcal{X}$, and pair $(y, y') \in \mathcal{Y} \times \mathcal{Y}'$, as

$$T(\theta) \triangleq (T_i(\theta_i) : i \in [d]), \qquad \hat{T}(X) \triangleq (\hat{T}_i(X) : i \in [d]), \qquad \ell(y, y') \triangleq \sum_{i=1}^d \ell_i(y_i, y'_i).$$

Remark 3. The observation X consists of independent and not identically distributed $X_i \sim P_{\theta_i}$ and the loss function takes a *separable* form. This should be contrasted with the multiple-observation model, in which m *i.i.d.* observations drawn from the same distribution P_θ are given.

Theorem 1.11 (Minimax risk of tensor product). *For the minimax risk of the tensorized experiment* $\sum_{i=1}^d R_B^*(\mathcal{P}_i) \leqslant R^*(\mathcal{P}) \leqslant \sum_{i=1}^d R^*(\mathcal{P}_i)$. *Consequently, if minimax theorem holds for each experiment, i.e.,* $R^*(\mathcal{P}_i) = R_B^*(\mathcal{P}_i)$ *for each* $i \in [d]$*, then it also holds for the tensorized experiment, i.e.* $R^*(\mathcal{P}) = \sum_{i=1}^d R^*(\mathcal{P}_i)$.

Proof. We will show the upper and lower bound separately.

(a) **Upper bound.** The upper bound follows by taking a sub-class of estimators where $\hat{T}_i(X) \triangleq \hat{T}_i(X_i)$. We can rewrite the minimax risk for the tensorized experiment as

$$\begin{aligned} R^*(\mathcal{P}) &= \inf_{\hat{T}} \sup_{\theta\in\Theta} \mathbb{E}[\ell(T(\theta), \hat{T}(X)) \mid \theta] \leqslant \inf_{\hat{T}} \sup_{\theta\in\Theta} \sum_{i=1}^d \mathbb{E}[\ell_i(T_i(\theta_i), \hat{T}_i(X)) \mid \theta] \\ &\leqslant \sum_{i=1}^d \inf_{\hat{T}_i(X_i)} \sup_{\theta_i\in\Theta_i} \mathbb{E}[\ell_i(T_i(\theta_i), \hat{T}_i(X_i)) \mid \theta] = \sum_{i=1}^d R^*(\mathcal{P}_i). \end{aligned}$$

(b) **Lower bound.** For the lower bound, we take a product prior $\pi \triangleq \prod_{i=1}^d \pi_i$ under which $\theta : \Omega \to \Theta$ is an independent vector, and consequently $X : \Omega \to \mathcal{X}$ is an independent vector. This implies that X_j has no information regarding θ_i for $j \neq i$, and hence for any random estimator $\hat{T}(X)$, we have $P_{\hat{T}(X)|X} = \prod_{i=1}^d P_{\hat{T}_i(X_i)|X_i}$, and

$$R_{\pi_i}(\hat{T}_i(X_i)) = \mathbb{E}\ell_i(T_i(\theta_i), \hat{T}_i(X_i)) \geqslant \inf_{\hat{T}_i} R_{\pi_i}(\hat{T}_i(X_i)) = R^*_{\pi_i}.$$

From the fact that sup is greater than mean, and the linearity of expectation, we get

$$\sup_{\theta\in\Theta} \mathbb{E}[\ell(T(\theta), \hat{T}(X)) \mid \theta] \geqslant \mathbb{E}\ell(T(\theta), \hat{T}(X)) = \sum_{i=1}^d \mathbb{E}\ell_i(T_i(\theta_i), \hat{T}_i(X)) = \sum_{i=1}^d R_{\pi_i}(\hat{T}_i(X_i)).$$

Since the choices of prior π_i and estimator $\hat{T}$ were arbitrary, the lower bound follows.

□

Example 1.12 (Unstructured GLM). Consider statistical decision theory simple setting with $\mathcal{Y} = \mathcal{Y}' = \Theta \triangleq \mathbb{R}^d$. An unstructured GLM statistical model $\mathcal{P} \triangleq (\mathcal{N}(\theta, \sigma^2 I_d) : \theta \in \Theta)$ with quadratic loss $\ell : (\theta, \hat{\theta}) \mapsto \|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^d (\theta_i - \hat{\theta}_i)^2$ is simply the d-fold tensor product of the one-dimensional GLM. Since minimax theorem holds for the GLM, Theorem 1.11 shows the minimax risks sum up to $d\sigma^2$.

Remark 4. In general, it is possible that the minimax risk of the tensorized experiment is strictly less than the sum of individual minimax risks. This may appear surprising since X_i only carries information about θ_i and it makes sense intuitively to estimate θ_i based solely on X_i. However, this is not always true.

Example 1.13 (Minimax risk of tensorized experiment strictly less than the sum of individual minimax risks). Consider statistical decision theory simple setting with label space $\mathcal{Y} = \mathcal{Y}' = \Theta \triangleq \mathbb{N}$, observation $X \triangleq \theta Z$ where $Z : \Omega \to \{0,1\}$ is an independent Bernoulli random variable with $\mathbb{E}Z = \frac{1}{2}$, and the loss function $\ell : (\theta,\hat{\theta}) \mapsto \mathbb{1}_{\{\hat{\theta}<\theta\}}$. If $Z = 0$, then all information about θ is erased. Therefore for any estimator $\hat{\theta} \triangleq \hat{T}(X,U)$, the risk is lower bounded by

$$R_\theta(\hat{\theta}) = P(\{\hat{\theta} < \theta\} \mid \theta) \geqslant P(\{\hat{\theta} < \theta, Z = 0\} \mid \theta) = \frac{1}{2} P(\{\hat{\theta} < \theta\} \mid \{Z = 0\}, \theta).$$

Taking supremum on both sides, we obtain $\sup_\theta R_\theta(\hat{\theta}) \geqslant \frac{1}{2}$. It follows that minimax risk $R^* \geqslant \frac{1}{2}$. For an estimator $\hat{T}(X,U) \triangleq X$, we obtain risk $R_\theta(\hat{\theta}) = \mathbb{E}[\mathbb{1}_{\{\theta Z<\theta\}}] = \mathbb{E}\mathbb{1}_{\{Z=0\}} = \frac{1}{2}$. It follows that minimax risk $R^* = \frac{1}{2}$. Recall that $R^*_\pi = 0$ in this case for any prior $\pi \in \mathcal{M}(\Theta)$.

Next consider the tensor product of two copies of this experiment with $\Theta \triangleq \mathbb{N}^2$ observation $X \triangleq \theta \circ Z$ where *i.i.d.* random vector $Z : \Omega \to \{0,1\}^2$ with $\mathbb{E}Z_1 = \frac{1}{2}$, and the loss function $L(\theta,\hat{\theta}) = \mathbb{1}_{\{\hat{\theta}_1<\theta_1\}} + \mathbb{1}_{\{\hat{\theta}_2<\theta_2\}}$. Consider the following estimator

$$\hat{\theta}_1 = \hat{\theta}_2 \triangleq X_1 \vee X_2 + \mathbb{1}_{\{X_1=X_2=0\}} = (\theta_1 \vee \theta_2) Z_1 Z_2 + \theta_1 Z_1 \bar{Z}_2 + \theta_2 \bar{Z}_1 Z_2 + \bar{Z}_1 \bar{Z}_2.$$

Since $\theta_1, \theta_2 \in \mathbb{N}$, we can write the indicators

$$\mathbb{1}_{\{\hat{\theta}_1<\theta_1\}} = \mathbb{1}_{\{\theta_2<\theta_1\}} \bar{Z}_1 Z_2 + \mathbb{1}_{\{1<\theta_1\}} \bar{Z}_1 \bar{Z}_2, \qquad \mathbb{1}_{\{\hat{\theta}_2<\theta_2\}} = \mathbb{1}_{\{\theta_1<\theta_2\}} Z_1 \bar{Z}_2 + \mathbb{1}_{\{1<\theta_2\}} \bar{Z}_1 \bar{Z}_2.$$

Since Z is *i.i.d.* Bernoulli random vector with $\mathbb{E}Z_1 = \frac{1}{2}$, we get $\mathbb{E}\bar{Z}_1 Z_2 = \mathbb{E}Z_1 \bar{Z}_2 = \mathbb{E}\bar{Z}_1 \bar{Z}_2 = \frac{1}{4}$. Therefore, for any $\theta_1, \theta_2 \in \mathbb{N}$, averaging over Z_1, Z_2, we can find the mean loss

$$\mathbb{E}L(\theta,\hat{\theta}) = \mathbb{E}\mathbb{1}_{\{\hat{\theta}_1<\theta_1\}} + \mathbb{E}\mathbb{1}_{\{\hat{\theta}_2<\theta_2\}} \leqslant \frac{1}{4}(\mathbb{1}_{\{\theta_1<\theta_2\}} + \mathbb{1}_{\{\theta_2<\theta_1\}} + 2) \leqslant \frac{3}{4}.$$

A Sufficient statistics

Definition A.1 (Sufficient statistic). Consider a parameter space Θ, input space $\mathcal{X}$, output space $\mathcal{Y}$, prediction space $\mathcal{Y}'$, a statistical model $\mathcal{P}(\Theta) \triangleq \{P_\theta \in \mathcal{M}(\mathcal{X}) : \theta \in \Theta\}$ for distribution of observation $X : \Omega \to \mathcal{X}$, an estimate $\hat{T} : \Omega \to (\mathcal{Y}')^{\mathcal{X}}$, and $P_{\hat{T}(X)|X}$ some Markov kernel such that $P^\theta_{\hat{T}(X)} \triangleq P_\theta P_{\hat{T}(X)|X}$ be the induced distribution on $\hat{T}(X)$ for each parameter θ. We say that $\hat{T}$ is a *sufficient statistic* of X for θ if there exists a transition probability kernel $P_{X|\hat{T}(X)}$ so that $P_\theta P_{\hat{T}(X)|X} = P^\theta_{\hat{T}(X)} P_{X|\hat{T}(X)}$, i.e., $P_{X|\hat{T}(X)}$ can be chosen to not depend on θ.

Remark 5. The intuitive interpretation of $\hat{T}$ being sufficient is that, with $\hat{T}$ at hand, one can ignore X. In other words, $\hat{T}$ contains all the relevant information to infer about θ. This is because X can be simulated on the sole basis of $\hat{T}$ without knowing θ. As such, X provides no extra information for identification of θ. Any one-to-one transformation of X is sufficient, however, this is not the interesting case. In the interesting cases dimensionality of T will be much smaller (typically equal to that of θ) than that of X.

Theorem A.2. *Let θ, X, T be as in the setting above. Then the following are equivalent.*
(a) T is a sufficient statistic of X for θ.
(b) For all P^θ, $\theta \to T \to X$.
(c) For all P^θ, $I(\theta; X|T) = 0$.
(d) For all P^θ, $I(\theta; X) = I(\theta; T)$, i.e., the data processing inequality for mutual information holds with equality.