Lecture-13: Minimax theorem

1 Minimax Risk

Example 1.1 (Minimax quadratic risk of GLM). Consider the statistical decision theory simple setting for Gaussian location model with unconstrained parameter space $\Theta \triangleq \mathbb{R}^d$, input space $\mathcal{X} = \Theta$, observation $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$, and quadratic loss function $L : (\theta, \hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|^2$. Recall that the minimax risk is defined as $R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R_\theta(\hat{\theta})$, where $R_\theta(\hat{\theta}) \triangleq \mathbb{E}[L(\theta, \hat{\theta}) \mid \theta]$. The upper bound is achieved by any estimate, and the lower bound is achieved by Bayes risk under any prior $\pi \in \mathcal{P}(\Theta)$. That is,

$$R_\pi^* \leqslant R^* \leqslant \sup_{\theta \in \Theta} R_\theta(\hat{\theta}).$$

Recall that the maximum likelihood estimate for GLM and quadratic cost is $\hat{\theta}_{\mathrm{ML}} \triangleq X$. Since $Z = X - \theta$ is zero mean Gaussian with distribution $\mathcal{N}(0, \sigma^2 I_d)$, the risk for ML estimate and quadratic loss is $R_\theta(\hat{\theta}_{\mathrm{ML}}) = \mathbb{E}[\|Z\|^2 \mid \theta] = d\sigma^2$ for all $\theta \in \Theta$. For prior distribution $\pi \triangleq \mathcal{N}(0, sI_d)$ parametrized by variance s, we have $R_\pi = \frac{s\sigma^2}{s+\sigma^2} d$ increasing in s. The least favorable prior is the one with the worst variance, and it follows that $R_\pi^* = \lim_{s \to \infty} R_\pi = d\sigma^2$. It follows that $R^* = d\sigma^2$.

Remark 1 (Non-uniqueness of minimax estimators). In general, estimators that achieve the minimax risk need not be unique. For instance, as shown in Example 1.1, the MLE $\hat{\theta}_{\mathrm{ML}} = X$ is minimax for the unconstrained GLM in any dimension. On the other hand, it is known that whenever $d \geqslant 3$, the risk of the James-Stein estimator is smaller that of the MLE everywhere and thus is also minimax. In fact, there exist a continuum of estimators that are minimax for this problem.

Example 1.2 (Minimax risk greater than Bayes risk). Consider the statistical decision theory simple setting with $\Theta \triangleq \mathbb{N}$ and loss function $L : (\theta, \hat{\theta}) \mapsto \mathbb{1}_{\left\{\hat{\theta} < \theta\right\}}$. For no observation case, estimate $\hat{\theta} \triangleq \hat{T}(U)$ for external randomness $U : \Omega \to [0,1]$, and risk $R_\theta(\hat{\theta}) = \mathbb{E}[L(\theta, \hat{\theta}) \mid \theta] = P\left\{\hat{\theta} < \theta\right\}$ is a non-decreasing function of θ. It follows that $\sup_\theta R_\theta(\hat{\theta}) = 1$ for any estimator $\hat{\theta}$. From the definition of minimax risk $R^* \triangleq \inf_{\hat{\theta}} \sup_\theta R_\theta(\hat{\theta}) = 1$.

For any prior $\pi \in \mathcal{M}(\mathbb{N})$, we have $R_\pi(\hat{\theta}) = \sum_{\theta \in \Theta} \pi_\theta P\left\{\hat{\theta} < \theta\right\}$, a non-increasing function in $\hat{\theta}$. Therefore, $R_\pi^* \triangleq \inf_{\hat{\theta}} R_\pi(\hat{\theta}) = 0$ for any prior $\pi \in \mathcal{M}(\mathbb{N})$. It follows that $R_B^* = \sup_{\pi \in \mathcal{M}(\mathbb{N})} R_\pi^* = 0$. Therefore, in this case $R^* = 1 > R_B^* = 0$.

Exercise 1.3. Consider the statistical decision theory simple setting for Gaussian location model with constrained parameter space $\Theta \triangleq \mathbb{R}_+$, input space $\mathcal{X} = \mathbb{R}$, observation $X \sim \mathcal{N}(\theta, \sigma^2)$, and quadratic loss function $L : (\theta, \hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|^2$.

(a) Show that the minimax quadratic risk of the GLM $X \sim \mathcal{N}(\theta, \sigma^2)$ with constrained parameter space $\Theta = \mathbb{R}_+$ is the same as the unconstrained case $\Theta = \mathbb{R}$.

(b) Show that the thresholded estimator $X_+ = X \vee 0$ achieves a better risk compared to maximum likelihood estimator, pointwise at every $\theta \in \mathbb{R}_+$.

1.1 Duality of minimax and Bayes risk

Recall the inequality $R^* \geqslant R_B^*$. This result can be interpreted from an optimization perspective. More precisely, R^* is the value of a primal convex optimization problem and R_B^* is precisely the value of its dual program. Thus the inequality that minimax risk exceeds Bayes risk is simply *weak duality*. If *strong duality* holds, then this is in fact an equality, in which case the minimax theorem holds.

Theorem 1.4. *Minimax risk exceeds worst case Bayes risk, i.e. $R^* \geqslant R_B^*$.*

Proof. For simplicity, we consider the case where Θ is a finite set. Recalling that $R_\theta(\hat{\theta}) = \mathbb{E}[L(\theta,\hat{\theta}) \mid \theta]$, we write

$$R^* = \min_{P_{\hat{\theta}|X}} \max_{\theta \in \Theta} R_\theta(\hat{\theta}).$$

Since $P_{\hat{\theta}|X} \mapsto R_\theta(\hat{\theta}) = \sum_{v \in \Theta} L(\theta, v) \int_{\mathcal{X}} P_{\hat{\theta}|X}(v \mid x) dP_\theta(x)$ is an affine map and the pointwise supremum of affine functions is convex. Hence, minimax is a convex optimization problem. To write down its dual problem, we rewrite this in an augmented form

$$\begin{aligned} R^* = &\min_{P_{\hat{\theta}|X}, t} t \\ &\text{such that } R_\theta(\hat{\theta}) \leqslant t \text{ for all } \theta \in \Theta. \end{aligned}$$

Let $\pi_\theta \geqslant 0$ denote the Lagrange multiplier or the dual variable for each inequality constraint corresponding to $\theta \in \Theta$. We define $\pi \triangleq (\pi_\theta : \theta \in \Theta)$, and write the Lagrangian for the above primal problem as

$$\mathcal{L}(P_{\hat{\theta}|X}, t, \pi) \triangleq t + \sum_{\theta \in \Theta} \pi_\theta (R_\theta(\hat{\theta}) - t) = (1 - \sum_{\theta \in \Theta} \pi_\theta) t + \sum_{\theta \in \Theta} \pi_\theta R_\theta(\hat{\theta}).$$

By definition, we have $R^* \geqslant \min_{P_{\hat{\theta}|X}, t} \mathcal{L}(P_{\hat{\theta}|X}, t, \pi)$. We note that if $\sum_{\theta \in \Theta} \pi_\theta \neq 1$, then $\min_{P_{\hat{\theta}|X}, t} \mathcal{L}(P_{\hat{\theta}|X}, t, \pi) = -\infty$. Thus π must be a probability measure and the dual problem is

$$\max_\pi \min_{P_{\hat{\theta}|X}, t} \mathcal{L}(P_{\hat{\theta}|X}, t, \pi) = \max_{\pi \in \mathcal{M}(\Theta)} \min_{P_{\hat{\theta}|X}} R_\pi(\hat{\theta}) = \max_{\pi \in \mathcal{M}(\Theta)} R_\pi^*.$$

Hence, the result follows. □

Remark 2. In summary, the minimax risk and the worst-case Bayes risk are related by convex duality, where the primal variables are randomized estimators and the dual variables are priors. This view can in fact be operationalized.

1.2 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 $L : \Theta \times \hat{\Theta} \to \mathbb{R}$. 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 $\theta \in \Theta, L(\theta, \cdot)$ is bounded from below and the sub-level set $\left\{\hat{\theta} \in \hat{\Theta} : L(\theta, \hat{\theta}) \leqslant a\right\}$ 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 1.5. 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 $R_\pi(\hat{\theta}) \triangleq \langle \pi, R(\hat{\theta}) \rangle$.

Remark 3. 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 1.6. *The sets S, T defined in Definition 1.5 are convex and disjoint.*

Proof. Let $\lambda \in [0,1]$.

(a) Let $\hat\theta_1(X, U_1), \hat\theta_2(X, U_2)$ be two randomized estimators, then we can define another randomized estimator $\hat\theta(X, U)$ for *i.i.d.* external randomness $U : \Omega \to [0,1]^3$, as

$$\hat\theta(X,U) \triangleq \hat\theta_1(X,U_1)\mathbb{1}_{\{U_3 \leqslant \lambda\}} + \hat\theta_2(X,U_2)\mathbb{1}_{\{U_3 > \lambda\}}.$$

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

$$R_\theta(\hat\theta) = \mathbb{E}[L(\theta,\hat\theta) \mid \theta] = \lambda R_\theta(\hat\theta_1) + \bar\lambda R_\theta(\hat\theta_2).$$

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) 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 1.7 (Minimax theorem). *Let Θ be a finite set, then $R^* = R^*_B$ in either of the following cases.*

1. *Input space $\mathcal{X}$ is finite.*
2. *The loss function L is bounded from below, i.e., $\inf_{\theta,\hat\theta} L(\theta,\hat\theta) > -\infty$.*

Proof. The first case directly follows from the duality interpretation of minimax and Bayes risk and the fact that strong duality holds for finite-dimensional linear programming.

For the second case, 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 1.6, we observe that the sets S, T of Definition 1.5 are convex and disjoint. Applying the separating hyperplane theorem to S and T, there exists a non-zero $\pi \in \mathbb{R}^\Theta$ and $c \in \mathbb{R}$, 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

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

□

1.3 Multiple observations and sample complexity

Definition 1.8 (Independent sampling model). Given $m \in \mathbb{N}$ and an experiment or statistical model $\mathcal{P}(\Theta) \triangleq \{P_\theta : \theta \in \Theta\}$, the *independent sampling model* is the experiment or statistical model $\mathcal{P}_m(\Theta) \triangleq \left\{P_\theta^{\otimes m} : \theta \in \Theta\right\}$. 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 1.9. Given a loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$, the minimax risk for simple setting is denoted by

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

Remark 4. 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 1.10 (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 1.11 (Sample complexity). The minimum sample size to attain a minimax risk of $\epsilon > 0$ is called *sample complexity* and denoted by $m^*(\epsilon) \triangleq \min\{m \in \mathbb{N} : R^*_m(\Theta) \leqslant \epsilon\}$.

Example 1.12 (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 1.13 (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 $m^*(\epsilon) = \Theta(\sqrt{\ln d})$.