Lecture-16: Bayes and minimax risk

1 Bayes and minimax risk

Definition 1.1 (Simple setting). For notational simplicity, we consider the task of estimating $T(\theta) \triangleq \theta$, such that label, prediction, and parameter spaces are identical, i.e. $\mathcal{Y} = \mathcal{Y}' = \Theta$. The observation $X : \Omega \to \mathcal{X}$ is sampled with distribution $P_\theta \in \mathcal{M}(\mathcal{X})$, and the random estimate $\hat{\theta} \triangleq \hat{T}(X)$ is a function of observation X and external randomness independent of everything else.

Remark 1. 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 1.2 (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*.

Remark 2. 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.

1.1 Bayes risk

Definition 1.3 (Bayes risk). Let $\pi \in \mathcal{M}(\Theta)$ be a *prior* probability distribution on parameter space Θ. Then the average risk with respect to prior π of an estimator $\hat{\theta}$ is defined as $R_\pi(\hat{\theta}) \triangleq \mathbb{E}_{\theta \sim \pi} R_\theta(\hat{\theta}) = \mathbb{E}[\mathbb{E}[\ell(\theta, \hat{\theta}) \mid \theta]]$. 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 $R^*_\pi = \mathbb{E}_{\theta \sim \pi}[R_\theta(\hat{\theta}_B)]$. We define the *worst case Bayes risk* as $R^*_B \triangleq \sup_{\pi \in \mathcal{P}(\Theta)} R^*_\pi$. If the supremum is attained for some prior π, the prior is called *least favorable*.

Lemma 1.4. *Bayes estimator is always deterministic for any loss function.*

Proof. Any randomized estimator $\hat{\theta} \triangleq \hat{T}(X)$ is a random variable conditioned on X and independent of everything else. We observe that for each $x \in \mathcal{X}$,

$$\mathbb{E}[\ell(\theta, \hat{T}(X)) \mid X = x, \theta] = \int_{\hat{\theta} \in \Theta} \ell(\theta, \hat{\theta}) dP_{\hat{T}(X) \mid X = x}(\hat{\theta}) \geqslant \inf\left\{\ell(\theta, \hat{\theta}) : dP_{\hat{T}(X) \mid X = x} > 0\right\}.$$

We denote this pointwise lower bound on the conditional expectation as $\inf \hat{T}(X)$. Hence, the risk of any randomized estimator is lower bounded by

$$R_\pi(\hat{\theta}) = \mathbb{E}\ell(\theta, \hat{T}(X)) = \mathbb{E}[\mathbb{E}[\ell(\theta, \hat{T}(X)) \mid X, \theta]] \geqslant \mathbb{E} \inf \ell(\theta, \hat{T}(X)).$$

□

Exercise 1.5 (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 $\ell : (\theta, \hat{\theta}) \mapsto \left\|\theta - \hat{\theta}\right\|^2$.

(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 deterministic 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 1.6 (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 $\ell : (\theta,\hat{\theta}) \mapsto \left\|\theta - \hat{\theta}\right\|^2$.

1.2 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 1.7 (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 3. 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 $R_\theta(\hat{\theta}) \geqslant R^* - \epsilon$.

Definition 1.8. We say that two nets $a,b \in \mathbb{R}^I$ are *asymptotically equal* if $ca_i \leqslant b_i \leqslant Ca_i$ for each $i \in I$ and some universal constants $c,C \geqslant 0$. We denote $a \asymp b$ and call a a *constant factor approximation* of b.

Remark 4. 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 $R^* \asymp \Psi$. Establishing Ψ is the minimax rate still entails proving the minimax upper and lower bounds, albeit within multiplicative constant factors.

Remark 5. 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 1.9. *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 $R_\pi(\hat{\theta}) = \mathbb{E}_{\theta\sim\pi} R_\theta(\hat{\theta}) \leqslant \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$. 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 1.10 (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 $\ell(\theta,\hat{\theta}) = \mathbb{1}_{\{\theta\neq\hat{\theta}\}}$. Consider a Bernoulli estimator $\hat{\theta} : \Omega \to \Theta$ 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 $\hat{\theta}$.

Example 1.11 (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 $\ell : (\theta, \hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|^2$. Recall that the minimax risk is defined as $R^* \triangleq \inf_{\hat{\theta}} \sup_{\theta\in\Theta} R_\theta(\hat{\theta})$, where $R_\theta(\hat{\theta}) \triangleq \mathbb{E}[\ell(\theta,\hat{\theta}) \mid \theta]$. 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,

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

(a) For the upper bound, we consider a maximum likelihood estimator and 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 $\theta \in \Theta$.

(b) For the lower bound, we consider prior distribution $\pi \triangleq \mathcal{N}(0, sI_d)$ parametrized by variance s. The Bayes estimator for quadratic loss is $\hat{\theta} = \mathbb{E}[\theta \mid X]$, and the Bayesian risk under this prior is $R^*_\pi = \frac{s\sigma^2}{s+\sigma^2} d$ which is increasing in s. The least favorable prior is the one with the worst variance, and it follows that $R^*_B = \lim_{s\to\infty} R^*_\pi = d\sigma^2$. It follows that $R^* = d\sigma^2$.

Remark 6 (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 1.12 (Minimax risk greater than Bayes risk). Consider the statistical decision theory simple setting with $\Theta \triangleq \mathbb{N}$ and loss function $\ell : (\theta, \hat{\theta}) \mapsto \mathbb{1}_{\{\hat{\theta} < \theta\}}$. For no observation case, estimate $\hat{\theta} \triangleq \hat{T}$ is a uniformly distributed random variable $\hat{T} : \Omega \to [0,1]$ independent of everything else, and risk $R_\theta(\hat{\theta}) = \mathbb{E}[\ell(\theta,\hat{\theta}) \mid \theta] = P(\{\hat{\theta} < \theta\} \mid \theta)$ 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 lower bound, we consider a prior $\pi \in \mathcal{M}(\mathbb{N})$, which results in Bayes risk $R_\pi(\hat{\theta}) = \sum_{\theta\in\Theta} \pi_\theta P(\{\hat{\theta} < \theta\} \mid \theta)$, a non-increasing function in estimate $\hat{\theta}$. Taking $\hat{\theta}_n \triangleq n \in \Theta$ for any $n \in \mathbb{N}$, we observe that $R^*_\pi \triangleq \inf_{\hat{\theta}_n} R_\pi(\hat{\theta}_n) = 0$ 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 1.13. 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 $\ell : (\theta, \hat{\theta}) \mapsto \left\|\hat{\theta} - \theta\right\|^2$.

(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 $\theta \in \mathbb{R}_+$.

1.3 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 1.14. *Minimax risk exceeds worst case Bayes risk, i.e. $R^* \geqslant R^*_B$.*

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

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

Since $P_{\hat{\theta}|X} \mapsto R_\theta(\hat{\theta}) = \sum_{v \in \Theta} \ell(\theta, v) \int_{\mathcal{X}} P_{\hat{\theta}|X}(v \mid x) dP_\theta(x)$ 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 $\min_{P_{\hat{\theta}|X}, t} \mathcal{L}(P_{\hat{\theta}|X}, t, \pi) = -\infty$. Why is that a problem? Thus π must be a probability measure and the dual problem is

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

□

Remark 7. 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.