Lecture Notes 8

1 Minimax Theory

Suppose we want to estimate a parameter θ using data $X^n = (X_1, \ldots, X_n)$. What is the best possible estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ? Minimax theory provides a framework for answering this question.

1.1 Introduction

Let $\widehat{\theta} = \widehat{\theta}(X^n)$ be an estimator for the parameter $\theta \in \Theta$. We start with a **loss function** $L(\theta, \widehat{\theta})$ that measures how good the estimator is. For example:

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$	L_p loss,
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$	zero–one loss,
$L(\theta, \widehat{\theta}) = I(\lvert\widehat{\theta} - \theta\rvert > c)$	large deviation loss,
$L(\theta, \widehat{\theta}) = \int \log\left(\frac{p(x;\theta)}{p(x;\widehat{\theta})}\right) p(x;\, \theta)dx$	Kullback–Leibler loss.

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector then some common loss functions are

$$L(\theta, \widehat{\theta}) = ||\theta - \widehat{\theta}||^2 = \sum_{j=1}^{k} (\widehat{\theta}_j - \theta_j)^2,$$

$$L(\theta, \widehat{\theta}) = ||\theta - \widehat{\theta}||_p = \left(\sum_{j=1}^{k} |\widehat{\theta}_j - \theta_j|^p \right)^{1/p}.$$

When the problem is to predict a $Y \in \{0, 1\}$ based on some classifier $h(x)$ a commonly used loss is

$$L(Y, h(X)) = I(Y \neq h(X)).$$

For real valued prediction a common loss function is

$$L(Y, \widehat{Y}) = (Y - \widehat{Y})^2.$$

The **risk** of an estimator $\widehat{\theta}$ is

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta\Big(L(\theta, \widehat{\theta}) \Big) = \int L(\theta, \widehat{\theta}(x_1, \ldots, x_n)) p(x_1, \ldots, x_n; \theta) dx. \tag{1}$$

When the loss function is squared error, the risk is just the MSE (mean squared error):

$$R(\theta, \widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2 = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2. \tag{2}$$

If we do not state what loss function we are using, assume the loss function is squared error.

The **minimax risk** is

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta})$$

where the infimum is over all estimators. An estimator $\widehat{\theta}$ is **a minimax estimator** if

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta}).$$

Example 1 *Let $X_1, \ldots, X_n \sim N(\theta, 1)$. We will see that $\overline{X}_n$ is minimax with respect to many different loss functions. The risk is $1/n$.*

Example 2 *Let $X_1, \ldots, X_n$ be a sample from a density p. Let $\mathcal{P}$ be the class of smooth densities (defined more precisely later). We will see (later in the course) that the minimax risk for estimating f is $Cn^{-4/5}$ for some constant $C > 0$.*

1.2 Comparing Risk Functions

To compare two estimators, we compare their risk functions. However, this does not provide a clear answer as to which estimator is better. Consider the following examples.

Example 3 *Let $X \sim N(\theta, 1)$ and assume we are using squared error loss. Consider two estimators: $\widehat{\theta}_1 = X$ and $\widehat{\theta}_2 = 3$. The risk functions are $R(\theta, \widehat{\theta}_1) = \mathbb{E}_\theta(X - \theta)^2 = 1$ and $R(\theta, \widehat{\theta}_2) = \mathbb{E}_\theta(3 - \theta)^2 = (3 - \theta)^2$. If $2 < \theta < 4$ then $R(\theta, \widehat{\theta}_2) < R(\theta, \widehat{\theta}_1)$, otherwise, $R(\theta, \widehat{\theta}_1) < R(\theta, \widehat{\theta}_2)$. Neither estimator uniformly dominates the other; see Figure 1.*

Example 4 *Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Consider squared error loss and let $\widehat{p}_1 = \overline{X}$. Since this has zero bias, we have that*

$$R(p, \widehat{p}_1) = \mathsf{Var}(\overline{X}) = \frac{p(1-p)}{n}.$$

Another estimator is

$$\widehat{p}_2 = \frac{Y + \alpha}{\alpha + \beta + n}$$

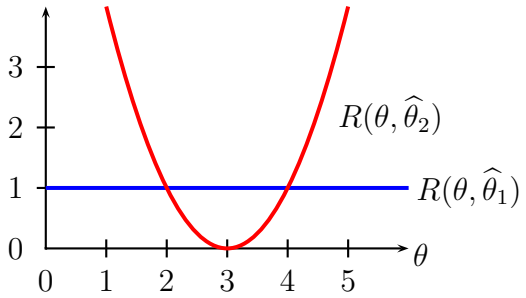

Figure 1: Comparing two risk functions. Neither risk function dominates the other at all values of θ.

where $Y = \sum_{i=1}^n X_i$ *and* α *and* β *are positive constants.*[1] *Now,*

$$\begin{aligned} R(p, \widehat{p}_2) &= \mathsf{Var}_p(\widehat{p}_2) + (\mathrm{bias}_p(\widehat{p}_2))^2 \\ &= \mathsf{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathbb{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2. \end{aligned}$$

Let $\alpha = \beta = \sqrt{n/4}$. *The resulting estimator is*

$$\widehat{p}_2 = \frac{Y + \sqrt{n/4}}{n + \sqrt{n}}$$

and the risk function is

$$R(p, \widehat{p}_2) = \frac{n}{4(n+\sqrt{n})^2}.$$

The risk functions are plotted in Figure 2. As we can see, neither estimator uniformly dominates the other.

These examples highlight the need to be able to compare risk functions. To do so, we need a one-number summary of the risk function. Two such summaries are the maximum risk and the Bayes risk.

The **maximum risk** is

$$\overline{R}(\widehat{\theta}) = \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \tag{3}$$

[1] This is the posterior mean using a Beta (α, β) prior.

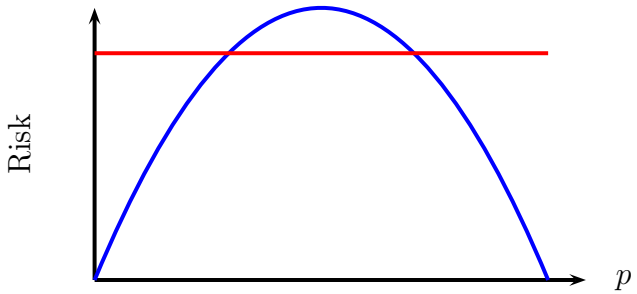

Figure 2: Risk functions for $\widehat{p}_1$ and $\widehat{p}_2$ in Example 4. The solid curve is $R(\widehat{p}_1)$. The dotted line is $R(\widehat{p}_2)$.

and the **Bayes risk** under prior π is

$$B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta. \tag{4}$$

Example 5 *Consider again the two estimators in Example 4. We have*

$$\overline{R}(\widehat{p}_1) = \max_{0\leq p\leq 1} \frac{p(1-p)}{n} = \frac{1}{4n}$$

and

$$\overline{R}(\widehat{p}_2) = \max_p \frac{n}{4(n+\sqrt{n})^2} = \frac{n}{4(n+\sqrt{n})^2}.$$

Based on maximum risk, $\widehat{p}_2$ is a better estimator since $\overline{R}(\widehat{p}_2) < \overline{R}(\widehat{p}_1)$. However, when n is large, $\overline{R}(\widehat{p}_1)$ has smaller risk except for a small region in the parameter space near $p = 1/2$. Thus, many people prefer $\widehat{p}_1$ to $\widehat{p}_2$. This illustrates that one-number summaries like maximum risk are imperfect.

These two summaries of the risk function suggest two different methods for devising estimators: choosing $\widehat{\theta}$ to minimize the maximum risk leads to minimax estimators; choosing $\widehat{\theta}$ to minimize the Bayes risk leads to Bayes estimators.

An estimator $\widehat{\theta}$ that minimizes the Bayes risk is called a **Bayes estimator**. That is,

$$B_\pi(\widehat{\theta}) = \inf_{\widetilde{\theta}} B_\pi(\widetilde{\theta}) \tag{5}$$

where the infimum is over all estimators $\widetilde{\theta}$. An estimator that minimizes the maximum risk is called a **minimax estimator**. That is,

$$\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widetilde{\theta}} \sup_{\theta} R(\theta, \widetilde{\theta}) \tag{6}$$

where the infimum is over all estimators $\widetilde{\theta}$. We call the right hand side of (6), namely,

$$R_n \equiv R_n(\Theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}), \tag{7}$$

the **minimax risk**. Statistical decision theory has two goals: determine the minimax risk R_n and find an estimator that achieves this risk.

Once we have found the minimax risk R_n we want to find the minimax estimator that achieves this risk:

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}). \tag{8}$$

Sometimes we settle for an asymptotically minimax estimator

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \sim \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \quad n \to \infty \tag{9}$$

where $a_n \sim b_n$ means that $a_n/b_n \to 1$. Even that can prove too difficult and we might settle for an estimator that achieves the minimax rate,

$$\sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \asymp \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}) \quad n \to \infty \tag{10}$$

where $a_n \asymp b_n$ means that both a_n/b_n and b_n/a_n are both bounded as $n \to \infty$.

1.3 Bayes Estimators

Let π be a prior distribution. After observing $X^n = (X_1, \ldots, X_n)$, the posterior distribution is, according to Bayes' theorem,

$$\mathbb{P}(\theta \in A | X^n) = \frac{\int_A p(X_1, \ldots, X_n|\theta)\pi(\theta)d\theta}{\int_\Theta p(X_1, \ldots, X_n|\theta)\pi(\theta)d\theta} = \frac{\int_A \mathcal{L}(\theta)\pi(\theta)d\theta}{\int_\Theta \mathcal{L}(\theta)\pi(\theta)d\theta} \tag{11}$$

where $\mathcal{L}(\theta) = p(x^n; \theta)$ is the likelihood function. The posterior has density

$$\pi(\theta|x^n) = \frac{p(x^n|\theta)\pi(\theta)}{m(x^n)} \tag{12}$$

where $m(x^n) = \int p(x^n|\theta)\pi(\theta)d\theta$ is the **marginal distribution** of X^n. Define the **posterior risk** of an estimator $\widehat{\theta}(x^n)$ by

$$r(\widehat{\theta}|x^n) = \int L(\theta, \widehat{\theta}(x^n))\pi(\theta|x^n)d\theta. \tag{13}$$

Theorem 6 *The Bayes risk $B_\pi(\widehat{\theta})$ satisfies*

$$B_\pi(\widehat{\theta}) = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \tag{14}$$

Let $\widehat{\theta}(x^n)$ be the value of θ that minimizes $r(\widehat{\theta}|x^n)$. Then $\widehat{\theta}$ is the Bayes estimator.

Proof. Let $p(x,\theta) = p(x|\theta)\pi(\theta)$ denote the joint density of X and θ. We can rewrite the Bayes risk as follows:

$$\begin{aligned} B_\pi(\widehat{\theta}) &= \int R(\theta,\widehat{\theta})\pi(\theta)d\theta = \int\left(\int L(\theta,\widehat{\theta}(x^n))p(x|\theta)dx^n\right)\pi(\theta)d\theta \\ &= \int\int L(\theta,\widehat{\theta}(x^n))p(x,\theta)dx^n d\theta = \int\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)m(x^n)dx^n d\theta \\ &= \int\left(\int L(\theta,\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta\right)m(x^n)\,dx^n = \int r(\widehat{\theta}|x^n)m(x^n)\,dx^n. \end{aligned}$$

If we choose $\widehat{\theta}(x^n)$ to be the value of θ that minimizes $r(\widehat{\theta}|x^n)$ then we will minimize the integrand at every x and thus minimize the integral $\int r(\widehat{\theta}|x^n)m(x^n)dx^n$.

Now we can find an explicit formula for the Bayes estimator for some specific loss functions.

Theorem 7 *If $L(\theta,\widehat{\theta}) = (\theta-\widehat{\theta})^2$ then the Bayes estimator is*

$$\widehat{\theta}(x^n) = \int \theta\pi(\theta|x^n)d\theta = \mathbb{E}(\theta|X=x^n). \tag{15}$$

If $L(\theta,\widehat{\theta}) = |\theta-\widehat{\theta}|$ then the Bayes estimator is the median of the posterior $\pi(\theta|x^n)$. If $L(\theta,\widehat{\theta})$ is zero–one loss, then the Bayes estimator is the mode of the posterior $\pi(\theta|x^n)$.

Proof. We will prove the theorem for squared error loss. The Bayes estimator $\widehat{\theta}(x^n)$ minimizes $r(\widehat{\theta}|x^n) = \int(\theta-\widehat{\theta}(x^n))^2\pi(\theta|x^n)d\theta$. Taking the derivative of $r(\widehat{\theta}|x^n)$ with respect to $\widehat{\theta}(x^n)$ and setting it equal to zero yields the equation $2\int(\theta-\widehat{\theta}(x^n))\pi(\theta|x^n)d\theta = 0$. Solving for $\widehat{\theta}(x^n)$ we get 15.

Example 8 *Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ where σ^2 is known. Suppose we use a $N(a,b^2)$ prior for μ. The Bayes estimator with respect to squared error loss is the posterior mean, which is*

$$\widehat{\theta}(X_1,\ldots,X_n) = \frac{b^2}{b^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{b^2+\frac{\sigma^2}{n}}a. \quad \blacksquare \tag{16}$$

1.4 Minimax Estimators

Finding minimax estimators is complicated and we cannot attempt a complete coverage of that theory here but we will mention a few key results. The main message to take away from this section is: **Bayes estimators with a constant risk function are minimax**.

Theorem 9 *Let $\widehat{\theta}$ be the Bayes estimator for some prior π. If*

$$R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \text{ for all } \theta \tag{17}$$

then $\widehat{\theta}$ is minimax and π is called a **least favorable prior**.

Proof. Suppose that $\widehat{\theta}$ is not minimax. Then there is another estimator $\widehat{\theta}_0$ such that $\sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta})$. Since the average of a function is always less than or equal to its maximum, we have that $B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0)$. Hence,

$$B_\pi(\widehat{\theta}_0) \leq \sup_\theta R(\theta, \widehat{\theta}_0) < \sup_\theta R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta}) \tag{18}$$

which is a contradiction.

Theorem 10 *Suppose that $\widehat{\theta}$ is the Bayes estimator with respect to some prior π. If the risk is constant then $\widehat{\theta}$ is minimax.*

Proof. The Bayes risk is $B_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta = c$ and hence $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ. Now apply the previous theorem.

Example 11 *Consider the Bernoulli model with squared error loss. In example 4 we showed that the estimator*

$$\widehat{p}(X^n) = \frac{\sum_{i=1}^n X_i + \sqrt{n/4}}{n + \sqrt{n}}$$

has a constant risk function. This estimator is the posterior mean, and hence the Bayes estimator, for the prior $\text{Beta}(\alpha, \beta)$ *with $\alpha = \beta = \sqrt{n/4}$. Hence, by the previous theorem, this estimator is minimax.*

Example 12 *Consider again the Bernoulli but with loss function*

$$L(p, \widehat{p}) = \frac{(p - \widehat{p})^2}{p(1-p)}.$$

Let $\widehat{p}(X^n) = \widehat{p} = \sum_{i=1}^n X_i/n$. The risk is

$$R(p, \widehat{p}) = E\left(\frac{(\widehat{p} - p)^2}{p(1-p)}\right) = \frac{1}{p(1-p)}\left(\frac{p(1-p)}{n}\right) = \frac{1}{n}$$

which, as a function of p, is constant. It can be shown that, for this loss function, $\widehat{p}(X^n)$ is the Bayes estimator under the prior $\pi(p) = 1$. Hence, $\widehat{p}$ is minimax.

What is the minimax estimator for a Normal model? To answer this question in generality we first need a definition. A function ℓ is **bowl-shaped** if the sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about the origin. A loss function L is bowl-shaped if $L(\theta, \widehat{\theta}) = \ell(\theta - \widehat{\theta})$ for some bowl-shaped function ℓ.

Theorem 13 *Suppose that the random vector X has a Normal distribution with mean vector θ and covariance matrix Σ. If the loss function is bowl-shaped then X is the unique (up to sets of measure zero) minimax estimator of θ.*

If the parameter space is restricted, then the theorem above does not apply as the next example shows.

Example 14 *Suppose that $X \sim N(\theta, 1)$ and that θ is known to lie in the interval $[-m, m]$ where $0 < m < 1$. The unique, minimax estimator under squared error loss is*

$$\widehat{\theta}(X) = m\left(\frac{e^{mX} - e^{-mX}}{e^{mX} + e^{-mX}}\right).$$

This is the Bayes estimator with respect to the prior that puts mass 1/2 at m and mass 1/2 at $-m$. The risk is not constant but it does satisfy $R(\theta, \widehat{\theta}) \leq B_\pi(\widehat{\theta})$ for all θ; see Figure 3. Hence, Theorem 9 implies that $\widehat{\theta}$ is minimax. This might seem like a toy example but it is not. The essence of modern minimax theory is that the minimax risk depends crucially on how the space is restricted. The bounded interval case is the tip of the iceberg.

Proof That $\overline{X}_n$ is Minimax Under Squared Error Loss. Now we will explain why $\overline{X}_n$ is justified by minimax theory. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2 I)$ be multivariate Normal with mean vector $\theta = (\theta_1, \ldots, \theta_d)$. We will prove that $\widehat{\theta} = X$ is minimax when $L(\theta, \widehat{\theta}) = ||\widehat{\theta} - \theta||^2$.

For simplicity, I will take $n = 1$ and $\sigma = 1$. You should do the more general case. The calculations are essentially the same.

Take the prior to be $\pi = N(0, c^2 I)$. Then the posterior is

$$\theta | X = x \ \sim N\left(\frac{c^2 x}{1 + c^2}, \frac{c^2}{1 + c^2} I\right). \tag{19}$$

The Bayes risk for an estimator $\widehat{\theta}$ is $R_\pi(\widehat{\theta}) = \int R(\theta, \widehat{\theta})\pi(\theta)d\theta$ which is minimized by the posterior mean $\widetilde{\theta} = c^2 X/(1 + c^2)$. Direct computation shows that $R_\pi(\widetilde{\theta}) = dc^2/(1 + c^2)$. Hence, if θ^* is any estimator, then

$$\begin{aligned} \frac{dc^2}{1 + c^2} &= R_\pi(\widetilde{\theta}) \leq R_\pi(\theta^*) && (20) \\ &= \int R(\theta^*, \theta)d\pi(\theta) \leq \sup_\theta R(\theta^*, \theta). && (21) \end{aligned}$$

We have now proved that $R_n \geq dc^2/(1 + c^2)$ for every $c > 0$ and hence

$$R_n \geq d. \tag{22}$$

But the risk of $\widehat{\theta} = X$ is d. So, $\widehat{\theta} = X$ is minimax.

Potential Test Question: fill in the details of the above proof for general n and σ^2.

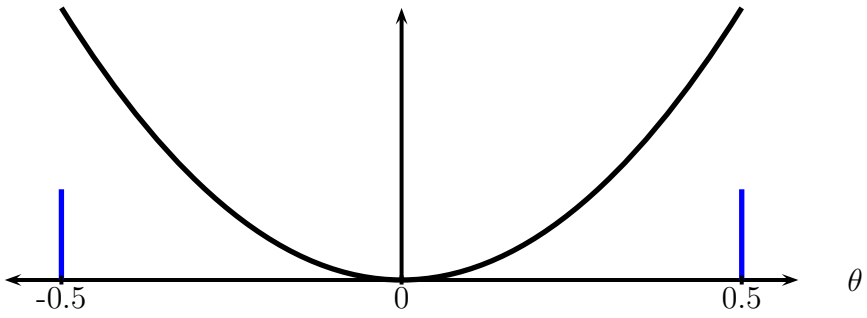

Figure 3: Risk function for constrained Normal with m=.5. The two short dashed lines show the least favorable prior which puts its mass at two points.

1.5 Maximum Likelihood

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the mle $\widehat{\theta}$ roughly equals the variance:[2]

$$R(\theta, \widehat{\theta}) = \mathsf{Var}_\theta(\widehat{\theta}) + \text{bias}^2 \approx \mathsf{Var}_\theta(\widehat{\theta}). \tag{23}$$

The variance of the mle is approximately $\mathsf{Var}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$ where $I(\theta)$ is the Fisher information. Hence,

$$nR(\theta, \widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{24}$$

For any other estimator θ', it can be shown that for large n, $R(\theta, \theta') \geq R(\theta, \widehat{\theta})$. So **the maximum likelihood estimator is approximately minimax. This assumes that the dimension of θ is fixed and n is increasing.**

1.6 The Hodges Example

Here is an interesting example about the subtleties of optimal estimators. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The mle is $\widehat{\theta}_n = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. But consider the following estimator due to

[2]Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

Hodges. Let

$$J_n = \left[-\frac{1}{n^{1/4}}, \ \frac{1}{n^{1/4}} \right] \tag{25}$$

and define

$$\widetilde{\theta}_n = \begin{cases} \overline{X}_n & \text{if } \overline{X}_n \notin J_n \\ 0 & \text{if } \overline{X}_n \in J_n. \end{cases} \tag{26}$$

Suppose that $\theta \neq 0$. Choose a small ϵ so that 0 is not contained in $I = (\theta - \epsilon, \theta + \epsilon)$. By the law of large numbers, $\mathbb{P}(\overline{X}_n \in I) \to 1$. In the meantime J_n is shrinking. See Figure 4. Thus, for n large, $\widetilde{\theta}_n = \overline{X}_n$ with high probability. We conclude that, for any $\theta \neq 0$, $\widetilde{\theta}_n$ behaves like $\overline{X}_n$.

When $\theta = 0$,

$$\begin{aligned} \mathbb{P}(\overline{X}_n \in J_n) &= \mathbb{P}(|\overline{X}_n| \leq n^{-1/4}) && (27) \\ &= \mathbb{P}(\sqrt{n}|\overline{X}_n| \leq n^{1/4}) = \mathbb{P}(|N(0,1)| \leq n^{1/4}) \to 1. && (28) \end{aligned}$$

Thus, for n large, $\widetilde{\theta}_n = 0 = \theta$ with high probability. This is a much better estimator of θ than $\overline{X}_n$.

We conclude that Hodges estimator is like $\overline{X}_n$ when $\theta \neq 0$ and is better than $\overline{X}_n$ when $\theta = 0$. So $\overline{X}_n$ is not the best estimator. $\widetilde{\theta}_n$ is better.

Or is it? Figure 5 shows the mean squared error, or **risk**, $R_n(\theta) = \mathbb{E}(\widetilde{\theta}_n - \theta)^2$ as a function of θ (for $n = 1000$). The horizontal line is the risk of $\overline{X}_n$. The risk of $\widetilde{\theta}_n$ is good at $\theta = 0$. At any θ, it will eventually behave like the risk of $\overline{X}_n$. But the maximum risk of $\widetilde{\theta}_n$ is terrible. We pay for the improvement at $\theta = 0$ by an increase in risk elsewhere.

There are two lessons here. First, we need to pay attention to the maximum risk. Second, it is better to look at uniform asymptotics $\lim_{n\to\infty} \sup_\theta R_n(\theta)$ rather than pointwise asymptotics $\sup_\theta \lim_{n\to\infty} R_n(\theta)$.

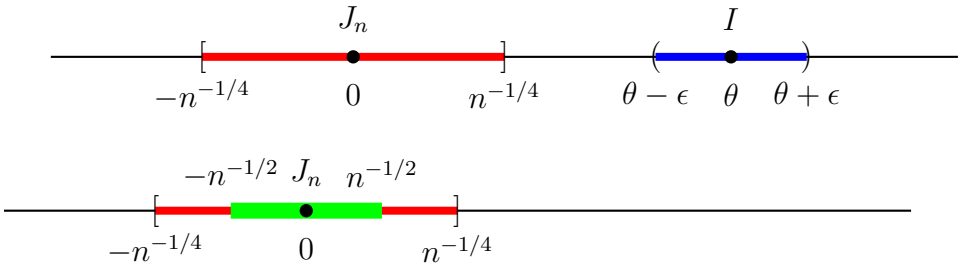

Figure 4: Top: when $\theta \neq 0$, $\overline{X}_n$ will eventually be in I and will miss the interval J_n. Bottom: when $\theta = 0$, $\overline{X}_n$ is about $n^{-1/2}$ away from 0 and so is eventually in J_n.

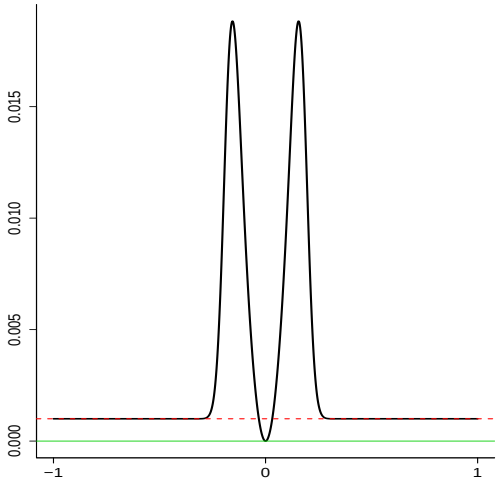

Figure 5: The risk of the Hodges estimator for $n = 1000$ as a function of θ. The horizontal line is the risk of the sample mean.