Statistics 210B, Spring 1998
Class Notes

P.B. Stark

`stark@stat.berkeley.edu`

`www.stat.berkeley.edu/~stark/index.html`

March 7, 1998

Fourth Set of Notes

1 Some remarks on Bayes and Minimax estimators

We observe $X \sim \mathbf{P}_\theta$, $\theta \in \Theta$. Let π be the prior distribution of θ; we assume that the support of π is a subset of Θ. Suppose the conditional distribution of X given θ is $\mathbf{P}_\theta$, with corresponding expectation operator E_θ. The action space is $\mathcal{A}$, and we seek a decision rule $\delta : \mathcal{X} \to \mathcal{A}$. The risk of a decision δ is $R(\theta, \delta) = E_\theta \ell(\theta, \delta(X))$. Define the *average risk* of an estimator δ to be

$$\begin{aligned} r_\pi(\delta) &\equiv E_{X,\theta} \ell(\theta, \delta(X)) \\ &= E[E(\ell(\theta, \delta(X)))|X)], \end{aligned} \tag{1}$$

where the expectation is with respect to the product measure of X and θ. The Bayes estimator minimizes the average risk.

The *posterior risk* of an action a is $r_\pi(a, x) = E_\pi(\ell(\theta, a)|X = x)$, where the subscript π is to remind us of the prior, but the expectation is with respect to the conditional distribution of θ given X, which is derived from the product measure on X and θ. Ideally, we would like to find the decision rule $\delta_\pi : \mathbf{R} \to \mathcal{A}$ that minimizes $r(\delta|x)$ for each x; such a rule would also minimize the Bayes risk. In general, such a rule need not exist; if one exists, it need not be unique (*vide infra*).

An estimator δ is unbiased for $\tau(\theta)$ if $E_\theta \delta(X) = \theta$. Recall that an estimator δ is inadmissible if there exists another estimator that does at least as well for all values of θ, and better for some value of θ. That is, if there is a δ_0 and θ_0 such that

$$R(\theta, \delta_0) = E_\theta(\ell(\theta, \delta(X)) \leq R(\theta, \delta) \;\; \forall \theta \in \Theta, \tag{2}$$

and

$$R(\theta_0, \delta_0) <\leq R(\theta_0, \delta). \tag{3}$$

One of the nice properties of Bayes estimators is that if they are unique, they are admissible.

Lehmann, TPE, §4.1 Theorem 1.1 states (in slightly different notation)

Theorem 1 *Let θ have distribution π, and, given $\theta = \gamma$, let X have distribution $\mathbf{P}_\gamma$. Suppose $\ell(\theta, a)$ is nonnegative for all θ, and that there exists an estimator δ_0 with finite risk for estimating estimating $\tau(\theta)$. If for almost all x there exists a rule $\delta_\pi(x)$ minimizing $E_\pi\{\ell(\theta, \delta(x))|X = x\}$, then δ_π is a Bayes estimator.*

Corollary 1 *If $\ell(\theta, a) = |a - \tau(\theta)|^2$, then $\delta_\pi = E_\pi\{\tau(\theta)|X = x\}$.*

Corollary 2 *If $\ell(\theta, a)$ is strictly convex in a, a Bayes estimator δ_π is unique a.e. $\mathcal{P} = \{\mathbf{P}_\theta\}$, provided the average risk of δ_π is finite, and provided the marginal distribution Q of X*

$$Q(A) = \int \mathbf{P}_\theta\{X \in A\} d\pi(\theta) \tag{4}$$

is such that a.e. Q implies a.e. $\mathcal{P}$.

The condition on Q ensures that measures $\mathbf{P}_\theta$ that are the only ones to assign mass to some points $x \in \mathcal{X}$ are not themselves given zero measure by π.

Note that we typically give up unbiasedness in moving to Bayes decisions:

Theorem 2 *(Lehmann, TPE, 4.4 Theorem 1.2) Let $\theta \sim \pi$ and let $\mathbf{P}_\theta$ be the conditional distribution of X given θ. Consider estimating $\tau(\theta)$ for squared-error loss. If $\delta(X)$ is unbiased, it cannot be Bayes unless*

$$E_{X,\theta}[\delta(X) - \tau(\theta)]^2 = 0. \tag{5}$$

Proof. Suppose δ is unbiased and is Bayes for $\tau(\theta)$. Then $\delta(X) = E_\pi[\tau(\theta)|X]$ a.e. Unbiasedness implies $E[\delta(X)|\theta = \gamma] = \tau(\gamma)$ for all $\gamma \in \Theta$. Conditioning on X gives

$$\begin{aligned} E[\tau(\theta)\delta(X)] &= E\{\delta(X)E[\tau(\theta)|X]\} \\ &= E\delta^2(X). \end{aligned} \tag{6}$$

Conditioning on θ gives

$$\begin{aligned} E[\tau(\theta)\delta(X)] &= E\{\tau(\theta)E[\delta(X)|\theta]\} \\ &= Eg^2(\theta). \end{aligned} \tag{7}$$

Thus

$$E[\delta(X) - \tau(\theta)] = E\delta^2(X) + Eg^2(\theta) - 2E[\tau(\theta)\delta(X)] = 0. \tag{8}$$

$\square$.

The Bayes estimator minimizes a weighted average of the risks for different possible values of the parameter $\theta \in \Theta$, where the weight is the prior distribution on those values. In contrast, the minimax decision rule minimizes the largest risk for any $\theta \in \Theta$:

$$\sup_{\theta \in \Theta} R(\theta, \delta). \tag{9}$$

There is a truly wonderful duality between the risks. A prior π for θ is *least favorable* if the Bayes risk is no larger for any other prior than for it; *i.e.*, if δ_π denotes the Bayes estimator for prior π on θ, then π^* is *least favorable* if

$$r_{\pi^*}(\delta_{\pi^*}) \geq r_\pi(\delta_\pi) \tag{10}$$

for all priors π on Θ.

Theorem 3 *(Lehmann, TPE, 4.2 Theorem 2.1) Suppose that π is a prior distribution on Θ such that*

$$E_\pi R(\theta, \delta_\pi) = \sup_{\theta \in \Theta} R(\theta, \delta_\pi), \tag{11}$$

where δ_π is the Bayes decision for prior π, as before. Then

1. *δ_π is minimax over Θ.*

2. *If δ_π is the unique Bayes decision for prior π, it is the unique minimax decision.*

3. *π is least favorable.*

Proof.

1. Let δ be a different decision rule. Then

$$\begin{aligned} \sup_{\theta \in \Theta} R(\theta, \delta) &\geq E_\pi R(\theta, \delta) \\ &\geq E_\pi R(\theta, \delta_\pi) \\ &= \sup_{\theta \in \Theta} R(\theta, \delta_\pi). \end{aligned} \tag{12}$$

2. same proof as (1), using $>$.

3. Let π_1 be another prior distribution on Θ. Then

$$\begin{aligned} r_{\pi_1}(\delta_{\pi_1}) &= E_{\pi_1} R(\theta, \delta_{\pi_1}) \\ &\leq E_{\pi_1} R(\theta, \delta_\pi) \\ &\leq \sup_{\theta \in \Theta} R(\theta, \delta_\pi) \\ &= r_\pi. \end{aligned} \tag{13}$$

For the Bayes risk of the Bayes estimator to equal the maximum risk of the Bayes estimator implies that

$$\mathbf{P}_\pi \{ R(\theta, \delta_\pi) = \sup_{\nu \in \Theta} R(\nu, \delta_\pi) \} = 1. \tag{14}$$

This, together with the theorem, implies that if a Bayes estimator has constant risk (over Θ), it is minimax. Moreover, if there is a set $\omega \subset \Theta$ with $\pi(\omega) = 1$ such that $R(\theta, \delta_\pi)$ attains its maximum at all $\theta \in \omega$, then δ_π is minimax.

The preceeding development has tacitly assumed that we are restricting attention to non-randomized estimators. When the loss function is strictly convex, the every randomized estimator is dominated by a non-randomized estimator. When the loss function is merely convex, for each randomized estimator, there is a non-randomized estimator whose risk is no larger than that of the randomized estimator. Thus in many situations (squared-error loss, in particular) it suffices to consider non-randomized estimators.

The following material is drawn primarily from TPE.

Lemma 1 *Jensen's inequality. Let $f : \mathcal{X} \rightarrow \mathbf{R}$ be a convex function, and let X be a random variable taking values in $\mathcal{X}$. Then*

$$f(EX) \leq Ef(X). \tag{15}$$

If f is strictly convex, the inequality is strict unless X is almost surely constant.

Definition 1 *A randomized decision rule δ is a mapping from the sample space $\mathcal{X}$ to a random variable $Y(x)$ that takes values in the action space $\mathcal{A}$ (which is assumed to be a measurable space). To each $x \in \mathcal{X}$, δ assigns a random variable $Y(x)$ with* known *distribution $\mathbf{P}_x$. The decision rule assigns to an observed value x an observation from the random variable $Y(x) \sim \mathbf{P}_x$. The risk of a randomized decision rule is $E_\theta E_X \ell(\theta, Y(X))$.*

Theorem 4 *(Lehmann, TPE, §1.5, Theorem 5.1) Suppose $X \sim \mathbf{P}_\theta$, $\theta \in \Theta$, and let T be sufficient for $\mathbf{P}_\Theta$. For any estimator $\delta(X)$ of $\tau(\theta)$ there exists a (possibly randomized) estimator based on T that has the same risk function as $\delta(X)$.*

Sketch of proof. Given T, the conditional distribution of X does not depend on θ. Let $\mathbf{P}(\cdot|T = t)$ denote this distribution. Given $T = t$, one can construct a random variable X'_t that has distribution $\mathbf{P}_(\cdot |T = t)$ The unconditional distributions of X'_t and X are the same: $\mathbf{P}_\theta\{X'_t \in A\} = \mathbf{P}_\theta\{X \in A\}$ for all measurable subsets $A \subset \mathcal{X}$. Thus if one knows the value of T, performing a subsequent randomization by drawing from $\mathbf{P}_(\cdot |T = t)$, allows one to generate data with the same distribution as the original experiment gave. One can therefore construct an estimator $\delta'(t)$ that depends on the data only through T and that is risk-equivalent to $\delta(x)$ by taking $\delta(t)$ to be $\delta(X'_t)$, whose value depends on the data only through T.

Remark. Any randomized estimator from data X is equivalent to a non-randomized estimator from data $X' = (X, U)$, where $U \sim U[0,1]$ is independent of X.

Theorem 5 *The Rao-Blackwell Theorem (see Lehmann, TPE, §1.6, Theorem 6.4). Let X have distribution $\mathbf{P}_\theta \in \mathbf{P}_\Theta = \{\mathbf{P}_\nu : \nu \in \Theta\}$, and let T be sufficient for $\mathbf{P}_\Theta$. Let $\delta : \mathcal{X} \to \mathcal{A}$ be an estimator of $\tau(\theta)$, and let the loss $\ell(\theta, a)$ be strictly convex in a. Suppose $E_\theta \delta(X) < \infty$ and $E_\theta \ell(\tau(\theta), \delta(X)) < \infty$, $\theta \in \Theta$. Let the estimator $\eta(t) \equiv E[\delta(X)|T = t]$. Then*

$$R(\theta, \eta) < R(\theta, \delta) \tag{16}$$

unless $\delta(X) = \eta(T)$ with probability 1.

Proof. If ℓ is strictly convex in a, then applying Jensen's inequality to the conditional expectation given $T = t$,

$$\ell(\theta, \eta(t)) < E\{\ell(\theta, \delta(X))|T = t\}, \tag{17}$$

unless $\delta(X) = \eta(t)$ a.s. Thus

$$E_\theta \ell(\theta, \eta(t)) < E_\theta E\{\ell(\theta, \delta(X))|T = t\}, \tag{18}$$

which was to be shown.

Corollary 3 *(Lehmann, TPE, \$1.6, Corollary 6.2) If the loss function ℓ is strictly convex, every randomized estimator of $\tau(\theta)$ is dominated by a non-randomized estimator. If ℓ is convex, there is a non-randomized estimator whose risk function is pointwise no larger than that of any randomized estimator.*

Proof. Any randomized estimator is equivalent to a nonrandomized estimator based on (X, U), and X is sufficient for X.

Note that the "zero-one" loss associated with confidence intervals is not convex. If the loss is

$$\ell(\theta, a) = \begin{cases} 0, & |\theta - a| \le \chi \\ 1, & |\theta - a| > \chi, \end{cases} \tag{19}$$

then the risk of δ is the non-coverage probability of the fixed-length interval $[\delta - \chi, \delta + \chi]$, which one would like to minimize for a given χ. This loss is not convex: take $a_0 = \theta$ and $a_1 = \theta + 3\chi$. Then $\ell(\theta, a_0) = 0$, $\ell(\theta, a_1) = 1$, and

$$\ell(\theta, (a_0 + a_1)/2) = 1 > (\ell(\theta, a_0) + \ell(\theta, a_1))/2 = 1/2. \tag{20}$$

(This loss is, however, *quasiconvex*. A quasi-convex function f is one for which

$$f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}, \tag{21}$$

for all x, y, and for all $\lambda \in [0, 1]$. If the inequality is strict whenever $\lambda \in (0, 1)$ and $x \neq y$, f is strictly quasiconvex. For any two actions a_0 and a_1, we have

$$\ell(\theta, \lambda a_0 + (1 - \lambda)a_1) \leq \max(\ell(\theta, a_0), \ell(\theta, a_1)), \;\; \forall \lambda \in [0, 1], \tag{22}$$

so ℓ is quasiconvex (but not strictly) in a. A different characterization of quasiconvex functions is that f is quasiconvex iff its level sets $\{x : f(x) \leq b\}$ are convex for every b. A local minimum of a strictly quasiconvex function is a global minimum.)

Lehmann (TPE, 4.2 Example 2.2) gives an example for this zero-one loss where a randomized decision does better than a non-randomized one. Suppose we are estimating the probability p of success in n i.i.d. Bernoulli(p) trials from the total number X of successes in the trials, which is a binomially-distributed sufficient statistic. Suppose the interval half-width is $\chi < 1/(2(n + 1))$. There are only $n + 1$ possible data, so a non-randomized rule can take only $n + 1$ possible values. Because the interval is so short, the union of the intervals centered at those values cannot include all of $\Theta = [0, 1]$, and thus the maximum risk for the minimax non-randomized rule is 1. (Hence, just picking $\delta(X) = 0$ is minimax among non-randomized decisions.) On the other hand, suppose we use the randomized rule $\delta_r(X) \sim U(0, 1)$, independent of the data and ignoring the data completely. Then

$$\sup_{\theta \in [0,1]} \mathbf{P}\{|U - \theta| > \chi\} = 1 - \chi < 1. \tag{23}$$

In this case, a randomized rule does uniformly better (as measured by maximum risk over Θ) than the best non-randomized rule.

2 Some Math

Before we begin, some math.

Definition 2 *A set $\mathcal{X}$ is* partially ordered *by a relation $\leq$ if for $x, y, z \in \mathcal{X}$,*

1. *$x \leq y$ and $y \leq z \Rightarrow x \leq z$ (transitivity)*
2. *$x \leq x$ for all $x \in \mathcal{X}$ (reflexivity)*
3. *$x \leq y$ and $y \leq x \Rightarrow x = y$.*

A subset $\mathcal{X}_0$ of $\mathcal{X}$ is totally ordered by $\leq$ if for every $x, y \in \mathcal{X}$, either $x \leq y$ or $y \leq x$. If $\mathcal{X}_0$ is totally ordered, $x, y \in \mathcal{X}_0$, $x \leq y$, and $x \neq y$, we write $x < y$.

That every nonempty partially ordered set contains a maximal totally ordered subset is Hausdorff's maximality theorem.

Definition 3 *Suppose the sets $\mathcal{X}$ and $\mathcal{Y}$ are totally ordered. Let $K(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$. We say $K(x, y)$ is* sign regular of order r *(SR_r) if for every $1 \leq m \leq r$ there is a constant $\epsilon_m = \pm 1$ such that for every pair of increasing sets of elements $(x_1 < x_2 < \ldots < x_m)$ and $(y_1 < y_2 < \ldots < y_m)$,*

$$\epsilon_m K \begin{pmatrix} x_1, x_2, \ldots, x_m \\ y_1, y_2, \ldots, y_m \end{pmatrix} \equiv \begin{vmatrix} K(x_1, y_1) & K(x_1, y_2) & \cdots & K(x_1, y_m) \\ K(x_2, y_1) & K(x_2, y_2) & \cdots & K(x_2, y_m) \\ \cdots & \cdots & \cdots & \cdots \\ K(x_m, y_1) K(x_m, y_2) \cdots K(x_m, y_m) & & & \end{vmatrix} \geq 0, \tag{24}$$

where the vertical bars denote the determinant of the matrix. If the inequality 24 is strict, K is said to be strictly sign regular of order r *(SSR_r). If all ϵ_j equal $+1$, $1 \leq j \leq r$, K is said to be* totally positive of order r *(TP_r). If all ϵ_j equal $+1$, $1 \leq j \leq r$, and the inequality 24 is strict, we say K is* strictly totally positive of order r *(STP_r). If the inequality 24 holds for all finite r, r is omitted from the notation, and K is said to be sign regular (SR), strictly sign regular (SSR), totally positive (TP), or strictly totally positive (STP), respectively.*

For statistical applications, a very useful fact is that the "kernel" $K(x,y)$ associated with the "density" of a one-parameter exponential family is totally positive. That is, if $\mathcal{X}$ and $\mathcal{Y}$ are totally ordered subsets of $\mathbf{R}$, the kernel $K(x,y) = \beta(x)e^{xy}$ is totally positive. This follows from the fact that an exponential polynomial $\sum_{j=1}^{n} p_j(y)e^{c_j y}$, where $c_j \neq c_j$ for $i \neq j$, and p_j is a real polynomial of degree d_j, either vanishes identically, or has at most $n - 1 + \sum_{j=1}^{n} d_j$ zeros (counting multiplicities).

Definition 4 *The* lower number of sign changes *of a finite real-valued sequence* $(x_j)_{j=1}^{m}$, $S^-((x_j))$, *is the number of sign changes in the sequence, discarding zeros. The* upper number of sign changes *of* (x_j), $S^+((x_j))$, *is the maximum number of sign changes in the sequence when the terms that equal zero are counted as having arbitrary signs. Let f be a real-valued function defined on a totally ordered subset $\mathcal{I}$ of* $\mathbf{R}$. *The* lower number of sign changes of f, $S^-(f)$ *is*

$$S^-(f) = \sup_{m<\infty,\ \{x_j\}\subset\mathcal{I}: x_1<x_2<\ldots x_m} S^-((f(x_j))_{j=1}^{m}), \tag{25}$$

and the upper number of sign changes of f, $S^+(f)$, *is*

$$S^+(f) = \sup_{m<\infty,\ \{x_j\}\subset\mathcal{I}: x_1<x_2<\ldots x_m} S^+((f(x_j))_{j=1}^{m}). \tag{26}$$

A very important result (which we shall use presently) is that transformations induced by a sign-regular kernel are *variation diminishing*: they do not increase the number of zero-crossings of a function.

Theorem 6 *(Karlin, §3, Theorem 3.1) Let $K(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ be Borel measurable, where $\mathcal{X}$ and $\mathcal{Y}$ are totally ordered topological spaces. Let μ be a sigma-finite regular measure on $\mathcal{Y}$, such that $\mu(U) > 0$ for each open set U for which $U \cap \mathcal{Y} \neq \emptyset$. Let X be a totally ordered topological space, and let $K(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbf{R}$ be Borel-measureable, and assume that $\int_{\mathcal{Y}} K(x,y)d\mu(y)$ exists for every $x \in \mathcal{X}$. Let $f : \mathcal{Y} \to \mathbf{R}$ be a bounded, Borel-measurable function on $\mathcal{Y}$. Define the transformation $(Tf) : \mathcal{X} \to \mathbf{R}$ by*

$$(Tf)(x) = \int_{\mathcal{Y}} K(x,y)f(y)d\mu(y). \tag{27}$$

1. *If $K(x,y)$ is SR_r, then if $S^-(f) \leq r-1$,*

$$S^-(Tf) \leq S^-(f), \tag{28}$$

If K is TP_r and f is piecewise continuous, then if $S^-(f) = S^-(Tf) \leq r-1$, f and Tf have the same sequence of signs as their arguments increase.

2. *If K is SSR_r and $f \neq 0$ a.e.(μ),*

$$S^+(Tf) \leq S^-(f) \tag{29}$$

if $S^-(f) \leq r-1$.

A transformation that does not increase the number of zero crossings of a function is called *variation diminishing.* Because the kernel associated with a one-parameter exponential family is TP, the theorem implies that integration against the density of an exponential family is variation diminishing.

For example, we obtain the Normal distribution with unit variance by taking $\beta(x) = e^{-x^2/2}/\sqrt{2\pi}$ and $d\mu(y) = e^{-y^2/2}dy$. Suppose f is bounded and Borel-measurable. Let

$$\begin{aligned}(Tf)(x) &= \int_{\mathbf{R}} f(y)e^{-x^2/2}/\sqrt{2\pi}e^{xy}e^{-y^2/2}dy \\ &= \int_{\mathbf{R}} f(y)e^{-(x-y)^2/2}/\sqrt{2\pi}dy \\ &= \int_{\mathbf{R}} f(y)\phi(x-y)dy \\ &= f \star \phi, \end{aligned} \tag{30}$$

where ϕ is the density of the standard normal distribution and $\star$ denotes convolution. Then $S^-(f\star\phi) \leq S^-(f)$.

In this case, $K(x,y)d\mu(y)$ is a probability density for fixed x. Suppose Y is a random variable with that density. Then a different notation for the transformation T is $(Tf)(x) = E_x f(Y)$.

See Karlin, 1968, *Total Positivity*, Stanford Univ. Press, Stanford CA, for more on total positivity.