Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

www.stat.berkeley.edu/∼stark/index.html

March 7, 1998

Sixth Set of Notes

1 Using the Bounded Normal Mean to Study Higher Dimensional Problems

Donoho (1994, *Ann. Stat., 22*, 238–270) shows how to connect certain infinitely parametric problems with Gaussian data errors to the one-dimensional problem of estimating a bounded normal mean. The connection is through the notion of the "hardest one-dimensional subproblem."

We shall change notation in this section. Let $\mathcal{X}$ be a convex subset of ℓ_2, the space of square-summable sequences. Suppose we observe

$$y = Kx + z, \tag{1}$$

where $x \in \mathcal{X}$, K is a linear operator from $\mathcal{X}$ into $\mathbf{R}^n$ (with n possibly infinite), z is a noise vector, and we want to estimate an affine functional $L(x)$, so as to do as well as possible for the worst case $x \in \mathcal{X}$ (minimax). Donoho relates the difficulty of this problem when z is chosen maliciously by a clever opponent, subject only to the constraint $\|z\|^2 \le \epsilon^2$ (the optimal recovery problem) to the difficulty of the problem when z is chosen randomly as a Gaussian noise vector with covariance matrix $\sigma^2 I$. By "difficulty" is meant a measure of the best, worst-case error of reconstruction, either as a maximum (for optimal recovery) or as risk (for the minimax problem). Examples of problems that can be cast in this form include nonparametric regression and density estimation; think of x as the coefficients of a function f in some orthonormal basis. The functional $L(x)$ could be the value of the regression function or density at a point, a derivative at a point, or a weighted average over some interval.

Donoho considers squared-error loss, absolute error loss, and length of a fixed-length confidence interval. We'll just go over squared-error loss. The other results are analogous, but involve the other measures of the difficulty of estimating a bounded normal mean.

Define the minimax affine risk for squared-error loss

$$R_A^*(\sigma; \mathcal{X}) = \inf_{\hat{L} \text{ affine}} \sup_{x \in \mathcal{X}} E(\hat{L}(y) - L(x))^2 \tag{2}$$

and the minimax risk for squared-error loss

$$R_N^*(\sigma; \mathcal{X}) = \inf_{\hat{L}} \sup_{x \in \mathcal{X}} E(\hat{L}(y) - L(x))^2. \tag{3}$$

Let $\rho_A(\tau, \sigma)$ be the affine minimax MSE for estimating θ from data $X \sim N(\theta, \sigma^2)$ subject to $\theta \in \Theta = [-\tau, \tau]$, and let $\rho_N(\tau, \sigma)$ be the minimax MSE for the same problem.

Suppose x were known to lie not just in $\mathcal{X}$, but in a one-dimensional subfamily of $\mathcal{X}$:

$$[x_{-1}, x_1] = \{\lambda x_{-1} + (1 - \lambda) x_1 : \lambda \in [0, 1]\}, \tag{4}$$

with x_{-1} and x_1 in $\mathcal{X}$. By convexity, $[x_{-1}, x_1] \subset \mathcal{X}$, so clearly

$$R_A^*(\sigma; \mathcal{X}) \geq R_A^*(\sigma; [x_{-1}, x_1]), \tag{5}$$

and similarly for R_N^*. Furthermore,

$$R_A^*(\sigma; \mathcal{X}) \geq \sup_{x_{-1}, x_1 \in \mathcal{X}} R_A^*(\sigma; [x_{-1}, x_1]), \tag{6}$$

and similarly for R_N^*.

Consider how difficult the problem would be if it were known that x was in the family $[x_{-1}, x_1]$. Let $x_0 = (x_{-1} + x_1)/2$, $w_0 = K(x_1 - x_{-1})/\|K(x_1 - x_{-1})\|$. Let $\theta = \langle w_0, Kx - Kx_0 \rangle$.

Let $\tau = \|K(x_1 - x_{-1})\|/2$. If $x \in [x_{-1}, x_1]$, the parameter $\theta \in [-\tau, \tau]$, and the distribution of $Y = \langle w_0, y - Kx_0 \rangle$ is $N(\theta, \sigma^2)$, so estimating θ is just a bounded normal mean problem. We already know something about the minimax MSE difficulty of such problems. If $\delta(\cdot)$ is minimax for estimating ν from $X \sim N(\nu, \sigma^2)$ with $\nu \in [-\tau, \tau]$, then $\delta(Y)$ is minimax for estimating θ.

Clearly, there are functions of y other than Y that we might consider in trying to estimate θ, but Y is sufficient for θ, so we can do at least as well just using it. Note that the minimax risk in the problem of estimating $s\theta + t$ from Y has s^2 times the minimax risk of estimating θ from Y. Restricting L to the subfamily $[x_{-1}, x_1]$ reduces L to an affine function of θ: $L(x) = L(x_0) + s\theta$, where

$$s = \frac{L(x_1) - L(x_{-1})}{\|Kx_1 - Kx_{-1}\|}. \tag{7}$$

Thus the minimax MSE risk for estimating Lx from y, if we know that $x \in [[x_{-1}, x_1]$, is

$$R_A^*(\sigma; [x_{-1}, x_1]) = \left[\frac{L(x_1) - L(x_{-1})}{\|Kx_1 - Kx_{-1}\|}\right]^2 \rho_A(\|Kx_1 - Kx_{-1}\|/2, \sigma), \tag{8}$$

and similarly for R_N^* in terms of $\rho_n(\cdot, \cdot)$.

For $v \in \mathcal{X}$, define the seminorm $\|v\|_K \equiv \|Kv\|$, where $\|Kv\|$ is the ordinary Euclidean norm. The *modulus of continuity of L with respect to the seminorm $\|\cdot\|_K$ over $\mathcal{X}$* is

$$\omega(\epsilon; L, K, \mathcal{X}) \equiv \sup_{x_{-1}, x_1 \in \mathcal{X}} \{|L(x_1) - L(x_{-1})| : \|x_1 - x_{-1}\|_K \leq \epsilon\}. \tag{9}$$

Donoho makes the following definitions: L is *well-defined* if the modulus of continuity of L with respect to the norm $\|\cdot\|$ over $\mathcal{X}$ is continuous at zero; *i.e.*, if

$$\lim_{\epsilon \downarrow 0} \omega(\epsilon; L, I, \mathcal{X}) = 0. \tag{10}$$

The operator K is *well-defined* if the modulus of continuity of K over $\mathcal{X}$ for the ℓ_2 norm is continuous at 0:

$$\lim_{\epsilon \downarrow 0} \sup_{x_{-1}, x_1 \in \mathcal{X},\, \|x_1 - x_{-1}\| \leq \epsilon} \|Kx_1 - Kx_{-1}\| = 0, \tag{11}$$

where the norm is the usual Euclidean norm.

Consider now the difficulty of the hardest one-dimensional subproblem:

$$\sup_{[x_{-1}, x_1] \in \mathcal{X}} R_A^*(\sigma, [x_{-1}, x_1]) \;=\; \sup_{\epsilon \geq 0} \sup_{x_{-1}, x_1 \in \mathcal{X},\, \|x_1 - x_{-1}\|_K = \epsilon} \left[\frac{L(x_1) - L(x_{-1})}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma)$$

$$= \sup_{\epsilon \geq 0} \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma). \tag{12}$$

Lemma 1 *(Donoho, 1994, Lemma 2) If $\mathcal{X}$ is closed, convex, and bounded, if L and K are well-defined, and if $\omega(\epsilon; L, K, \mathcal{X})$ is finite for each $\epsilon \geq 0$, then for each $\epsilon \geq 0$, there exists $x_{-1}, x_1 \in \mathcal{X}$ s.t. $\|x_1 - x_{-1}\|_K \leq \epsilon$ and $|L(x_1) - L(x_{-1})| = \omega(\epsilon)$. Also, there exists a hardest one-dimensional subfamily for affine estimates, that is, a pair $x_{-1}, x_1 \in \mathcal{X}$ s.t.*

$$R_A^*(\sigma, [x_{-1}, x_1]) = \sup_{\epsilon \geq 0} \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma). \tag{13}$$

This lemma follows from results on convexity and weak convergence.

The proof of this lemma relies on characterizing the superdifferential

Theorem 1 *(Donoho, 1994, Theorem 1.) If $\mathcal{X}$ is closed, bounded, and convex, if L and K are well-defined, and if $\omega(\epsilon)$ is finite for every $\epsilon \geq 0$, then the affine difficulty of the full problem is equal to the affine difficulty of a hardest one-dimensional subproblem:*

$$R_A^*(\sigma) = \max_{x_1, x_{-1} \in \mathcal{X}} R_A^*(\sigma; [x_{-1}, x_1]). \tag{14}$$

The estimator that is minimax for a hardest one-dimensional subproblem is minimax for the full problem.

Donoho's proof of Theorem 1 relies on showing that there exists an affine estimator of the form

$$L_0(y) = L(x_0) + d\langle w_0, y - Kx_0\rangle \tag{15}$$

that (1) is minimax for the subproblem $[x_{-1}, x_1]$, and (2) attains its worst-case risk over all of $\mathcal{X}$ in the subproblem $[x_{-1}, x_1]$. That is done using results from convex analysis to find a particular value of d, then showing that for that d, (1) and (2) hold. The construction of d Dohono uses gives (1) directly. Because the variance of L_0 does not depend on x, the dependence of the risk on x is through the bias of L_0. If the bias is at least as large within the subproblem as anywhere else, the maximum risk is attained in the subproblem. Donoho establishes the second property by showing that the absolute value of the bias of $L_0(y)$ is larger at x_1 than for any other $x \in \mathcal{X}$.

Theorem 2 *(Donoho, 1994, Theorem 2.) If L is affine, if $\mathcal{X}$ is convex, and if L and K are well-defined, then*

$$R_A^*(\sigma) = \sup_{\epsilon \geq 0} \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma). \tag{16}$$

Corollary 1 *(Donoho, 1994, Corollary 1.) Under the assumptions of Theorem 2,*

$$R_A^*(\sigma) \leq 1.25 R_N^*(\sigma). \tag{17}$$

Corollary 2 *(Donoho, 1994, Corollary 2.) Under the assumptions of Theorem 2,*

$$\rho_N(1/2, 1)\omega^2(\sigma) \leq R_N^*(\sigma) \leq R_A^*(\sigma) \leq \omega^2(\sigma), \tag{18}$$

so the modulus of continuity determines the minimax risk, up to a constant factor.

Proof. The modulus of continuity of a linear functional over a convex set is subadditive:

$$\omega(\sigma_1 + \sigma_2) \leq \omega(\sigma_1) + \omega(\sigma_2), \tag{19}$$

as we may readily show.

$$\begin{aligned}
\omega(\sigma_1 + \sigma_2) &= \sup_{x_1, x_{-1} \in \mathcal{X} : \|x_1 - x_{-1}\|_K \leq \sigma_1 + \sigma_2} |L(x_1) - L(x_{-1})| \\
&\leq \sup_{x_1, x_{-1}, x_2, x_{-2} \in \mathcal{X} : \|x_1 - x_{-1}\|_K \leq \sigma_1, \|x_2 - x_{-2}\| \leq \sigma_2} |L(x_2) - L(x_{-2}) + L(x_1) - L(x_{-1})| \\
&\leq \sup_{x_1, x_{-1}, x_2, x_{-2} \in \mathcal{X} : \|x_1 - x_{-1}\|_K \leq \sigma_1, \|x_2 - x_{-2}\| \leq \sigma_2} |L(x_2) - L(x_{-2})| + |L(x_1) - L(x_{-1})| \\
&\leq \sup_{x_1, x_{-1} \in \mathcal{X} : \|x_1 - x_{-1}\|_K \leq \sigma_1} |L(x_1) - L(x_{-1})| + \sup_{x_2, x_{-2} \in \mathcal{X} : \|x_2 - x_{-2}\| \leq \sigma_2} |L(x_2) - L(x_{-2})| \\
&= \omega(\sigma_1) + \omega(\sigma_2).
\end{aligned} \tag{20}$$

The convexity of $\mathcal{X}$ was used in the second step. Obviously, $\omega(\epsilon)$ is monotone increasing for $\epsilon \geq 0$. Because ω is subadditive (and by assumption $\omega(0) = 0$), $\omega(\epsilon)/\epsilon$ is a decreasing function of ϵ for $\epsilon \geq 0$. Therefore,

$$\sup_{\epsilon \geq \sigma} \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma) \leq \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \sup_{\epsilon \geq \sigma} \rho_A(\epsilon/2, \sigma) = \omega^2(\sigma), \tag{21}$$

because $\rho_A(\epsilon/2, \sigma) \to \sigma^2$ as $\epsilon \to \infty$. Using the monotonicity of ω and the fact that $\rho_A(\epsilon/2, \sigma) \leq \sigma^2$,

$$\sup_{\epsilon \leq \sigma} \left[\frac{\omega(\epsilon)}{\epsilon}\right]^2 \rho_A(\epsilon/2, \sigma) \leq \omega^2(\sigma) \sup_{\epsilon \leq \sigma} \epsilon^{-2} \rho_A(\epsilon/2, \sigma) \leq \omega^2(\sigma). \tag{22}$$

Thus

$$R_A^*(\sigma) \le \omega^2; \tag{23}$$

the lower bound is part of Theorem 2.

Donoho uses this bound to calculate the minimax rates of convergence in certain problems in terms of the rate at which the modulus of continuity goes to zero as $\sigma \downarrow 0$.

Donoho also points out that the maximum risk of an affine estimator for convex $\mathcal{X}$ and well-defined L and K is the same for $\mathcal{X}$ and the closure of $\mathcal{X}$, so it is not necessary to assume that $\mathcal{X}$ is closed.

Donoho gives a variety of applications that can be cast in the form of the canonical problem; here are a few.

Approximately Linear Models. Observe

$$y_i = a + \beta t_i + \delta_i + z_i, \;\; i = 1, \ldots, n, \tag{24}$$

a, β unknown reals, δ_i unknown except for $|\delta_i| \le c_i$, $i = 1, \ldots, n$, and $\{z_i\}$ i.i.d. $N(0, \sigma^2)$. Seek to estimate β.

Set $x = (a, \beta, \delta_1, \ldots, \delta_n)$, $(Kx)_i = a + \beta t_i + \delta_i$, $Lx = \beta$, $\mathcal{X} = \mathbf{R}^2 \times \Delta$, where $\Delta \equiv \{(b_1, \ldots b_n) : |b_i| \le c_i, \;\; i = 1, \ldots, n\}$. This is an instance of the general problem (with $\mathcal{X}$ convex but unbounded).

Assignment. Let $c_i = 1$, $i = 1, \cdots, n$. Let the domain be $[0, 1]$, and let $t_i = (i-1)/(n-1)$. Find $\omega(\sigma)$. Does $\omega(\sigma) \to 0$ as $\sigma \to 0$? Find the minimax affine estimator of β.

Semiparametric Models. Observe

$$y_i = \beta t_i + f(u_i) + z_i, \;\; i = 1, \ldots, n, \tag{25}$$

with β an unknown real, $\{t_i\}$ and $\{u_i\}$ known reals, f unknown except $f \in \mathcal{F}$ (convex), and $\{z_i\}$ iid $N(0, \sigma^2)$.

Setting $\delta_i = f(u_i)$, $\Delta = \{(\delta_i) : \delta_i = f(u_i)$ for some $f \in \mathcal{F}\}$ transforms this to an instance of an approximately linear model; convexity of $\mathcal{F}$ yields convexity of Δ.

Nonparametric Regression. Observe

$$y_i = f(t_i) + z_i, \;\; i = 1, \ldots, n, \tag{26}$$

$f \in \mathcal{F}$, a convex class of L_2 functions on a domain $D \subset \mathbf{R}^d$, $\{t_i\} \subset D$, $\{z_i\}$ iid $N(0, \sigma^2)$. We want to estimate a linear functional $T(f)$, such as $T_0(f) = f(t_0)$ or $T_1(f) = f'(t_0)$.

Let $\{\phi_j(t)\}_{j=1}^{\infty}$ be an orthonormal basis for $L_2(D)$, $x_j = x_j(f) = \langle f, \phi_j \rangle$, $x = (x_j)$, $X = \{(x_j(f)) : f \in \mathcal{F}\}$, $(Kx)_i = \sum_j x_j \phi_j(t_i)$, $L(x) = T(f)$.

Linear Inverse Problems. Observe

$$y_i = (Pf)(t_i) + z_i, \quad i = 1, \ldots, n, \tag{27}$$

with P a linear operator (such as a convolution). This is the same as nonparametric regression, but with $(Kx)_i = \sum_j x_j (P\phi_j)(t_i)$.