1 Normal Means

Suppose that $Y \sim (\mu, 1)$ and define the loss function:

$$L_q(\mu) = 2^{q-2}(Y-\mu)^2 + \lambda|\mu|^q \tag{1}$$

We can also define our estimate of the mean as follows:

$$\hat{\mu}_q = \underset{\mu}{argmin} L_q(\mu) \tag{2}$$

Consider q=0, q=1, and q=2. We can find the estimate of μ by taking the derivative and setting to zero.

- q=0 $\implies \frac{d}{d\mu}L_q(\mu) = -\frac{1}{2}(Y-\mu) = 0 \implies \hat{\mu_0} = Y$
- q=2 $\implies \frac{d}{d\mu}L_q(\mu) = -2(Y-\mu) + 2\lambda\mu = 0 \implies \hat{\mu_2} = \frac{Y}{\lambda+1}$
- q=1 ????

For the case of q=1, we need to introduce the notion of a subderivative.

2 Subdifferential

Definition 2.1. *We call c a subderivative of f at x_0 provided $f(x) - f(x_0) \geq c(x - x_0)$*

A convex function can be optimized by setting the subderivative = 0.

Definition 2.2. *The subdifferential $\partial f|_{x_0}$ is the set of subderivatives.*

x_0 minimizes f if and only if $0 \in \partial f|_{x_0}$.

Example 2.3. *For $\rho(\mu) = |\mu|$,*

$$\partial\rho|_\mu = \begin{cases} \{-1\} & \textit{if } \mu < 0 \\ [-1, 1] & \textit{if } \mu = 0 \\ \{1\} & \textit{if } \mu > 0 \end{cases}$$

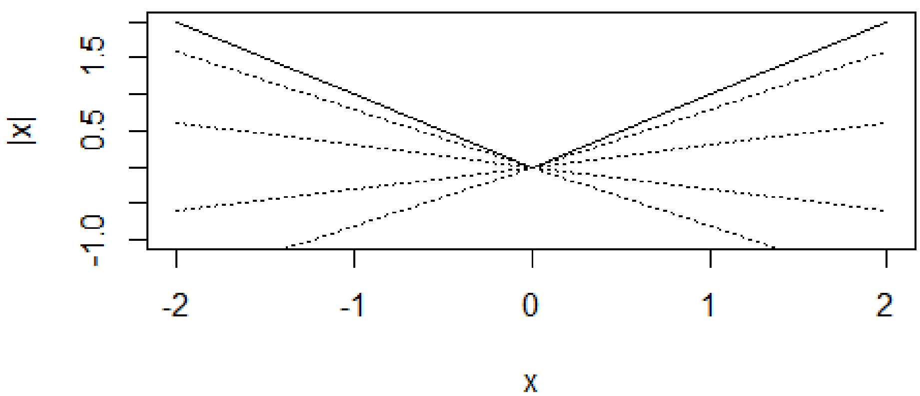

Figure 1: The subdifferential at 0 can vary between -1 and 1

While $|x|$ is not differentiable at 0, the subdifferential can be defined at 0.

Recall $L_1(\mu) = \frac{1}{2}(Y - \mu)^2 + \lambda|\mu|$. The subdifferential is as follows:

$$\partial L_1|_\mu = \begin{cases} \{\mu - Y - \lambda\} & \text{if } \mu < 0 \\ \{\mu - Y + \lambda z : -1 \leq z \leq 1\}[-1, 1] & \text{if } \mu = 0 \\ \{\mu - Y + \lambda\} & \text{if } \mu > 0 \end{cases}$$

$\hat{\mu}_1$ minimizes L_1 if and only if $0 \in \partial L_1$. Therefore, we obtain the following estimator.

$$\hat{\mu}_1 = \begin{cases} Y + \lambda & \text{if } Y < -\lambda \\ 0 & \text{if } -\lambda \leq Y \leq \lambda \\ Y - \lambda & \text{if } Y > \lambda \end{cases}$$

This can be written as

$$\hat{\mu}_1 = sgn(Y)(|Y| - \lambda)_+ \tag{3}$$

This is known as **soft thresholding**.

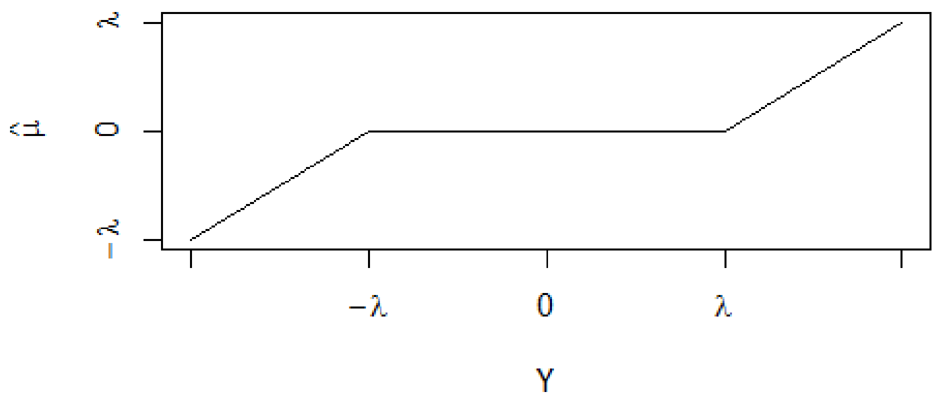

Figure 2: Soft thresholding sets the estimate to 0 when Y is near 0

3 Orthogonal Design

Suppose now that $(p \leq n)$, $Y = \mathbb{X}\beta + \epsilon$, where $\mathbb{X}^\top \mathbb{X}/n = I$

We want to solve:

$$\hat{\beta}_\lambda = argmin_\beta \frac{1}{2n} ||\mathbb{X}\beta - Y||_2^2 + \lambda||\beta||_1$$

We can minimize this component wise:

$$L(\beta) = \beta^2/2 - \beta\hat{\beta}_{LS} + \lambda|\beta|$$

This can be optimized using **subdifferentials** (see homework #2).

This results in **soft-thresholding** the least squares solution.

4 Non-orthogonal Design

An iterative algorithm for finding $\hat{\beta} \equiv \hat{\beta}_\lambda$ is:

Set $\hat{\beta} = (0, \ldots, 0)^\top$. Then for $j = 1, \ldots, p$:

1. Define $R_i = \sum_{k \neq j} \hat{\beta}_k X_{ik}$
2. Form $\hat{\beta}_j$ by simple least squares of $(R_i)_{i=1}^n$ on x_j
3. Soft-threshold these coefficients: $\hat{\beta}_j = \text{sgn}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda/||x_j||_2^2)_+$

5 Assumptions in High Dimension

Most theoretical papers on high-dimensional regression have several components:

- The linear model is correct.
- The variance is constant.
- The errors have a Normal distribution (or related distribution)
- The parameter vector is sparse.
- The design matrix has very weak collinearity.

In the next set of scribe notes (#6), low assumption prediction will be addressed.