

# CSE 6740 Lecture 9

## *What Loss Function Should I Use? (Estimation Theory)*

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

# Quiz Answers

1. For a Bayesian, parameters are drawn from probability distributions. T.
2. Even a flat prior has some effect on an estimator. T.
3. The effect of the prior on the estimator diminishes as  $N \rightarrow \infty$ . T.

# Today

1. Robustness (“How safe/stable is my loss function?”)
2. Comparing Estimators (“How can I say one loss function is superior to another?”)

# Robustness

We often choose according to mathematical/computational convenience. Otherwise, mostly robustness decides.

# Robustness

In the (approximate) words of [Huber, 1981]: Any statistical procedure should possess the following desirable features:

- It has reasonably good efficiency under the assumed model.
- It is robust in the sense that small deviations from the assumed model assumptions should impair the performance only slightly.
- Somewhat larger deviations from the model should not cause a catastrophe.

# MLE vs. L2E

Let's revisit  $L_2$  estimation (L2E), which we used for KDE. If  $f$  is the true density and  $\hat{f}_\theta$  is an estimate with parameters  $\theta$ , the  $L_2$  error or  $L_2$  distance is

$$\begin{aligned} L_2(\theta) &= \int (\hat{f}_\theta(x) - f(x))^2 dx & (1) \\ &= \int \hat{f}_\theta^2(x) dx - 2 \int \hat{f}_\theta(x) f(x) dx + \int f^2(x) dx. \end{aligned}$$

Note that the third term can be ignored for the purpose of comparing different estimators.

# MLE vs. L2E

Given a dataset, we wish to find the parameters which minimize the  $L_2$  risk

$$\mathbb{E} [L_2(\theta)] = \int \hat{f}_\theta^2(x) dx - \frac{2}{N} \sum_{i=1}^N \hat{f}_\theta(x_i). \quad (2)$$

The term  $\int \hat{f}_\theta^2(x) dx$  can be thought of as a kind of built-in regularization term, which acts to penalize spikes or overly large densities (due to, say, overlapped components in a mixture), and the second term as a goodness-of-fit term.

# MLE vs. L2E

Let's do L2E for a mixture of Gaussians

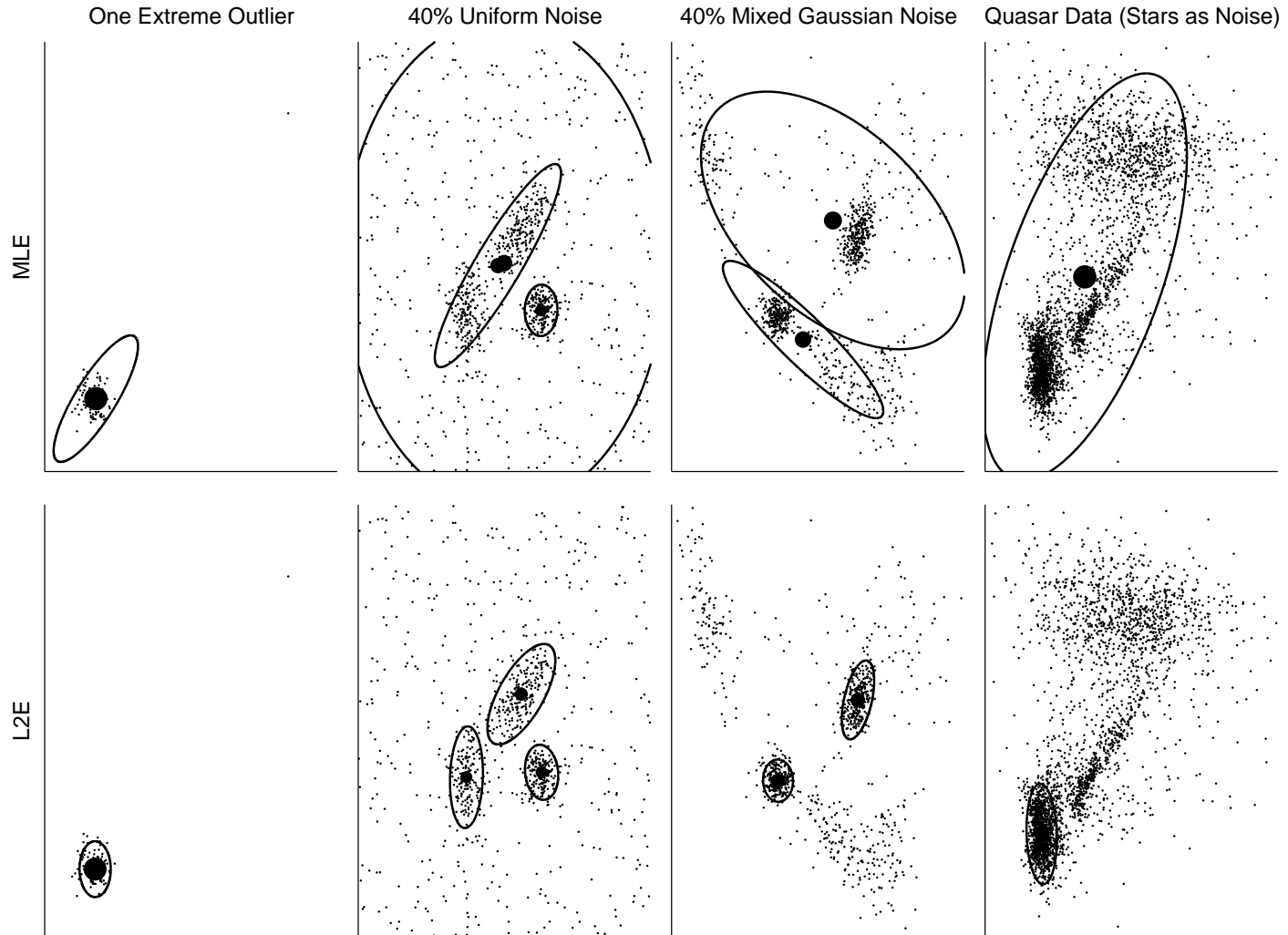
$$\hat{f}_\theta(x) = \sum_{k=1}^K \omega_k \phi(x|\mu_k, \Sigma_k). \quad (3)$$

The L2E regularization term for a mixture of Gaussians is

$$\int \hat{f}_\theta^2(x) dx = \sum_{k=1}^K \sum_{j=1}^K \omega_k \omega_j \phi(\mu_j|\mu_k, \Sigma_k + \Sigma_j). \quad (4)$$

The expression  $\phi(\mu_j|\mu_k, \Sigma_k + \Sigma_j)$  comes from the identity  $\phi(x|\mu_k, \Sigma_k)\phi(x|\mu_j, \Sigma_j) = \phi(\mu_j|\mu_k, \Sigma_k + \Sigma_j)\phi(x|\mu'_{k,j}, \Sigma'_{k,j})$ ; this and other properties of Gaussians make the integral tractable.

# MLE vs. L2E



# Robustness

Let  $X \sim N(\mu, \sigma^2)$ . The value which minimizes squared-error, or  $L_2$  loss,  $\arg \min_{\theta} \mathbb{E}(X - \theta)^2$ , is the mean of  $X$ :

$$\frac{d}{d\theta} \mathbb{E}(X - \theta)^2 = 0 \quad \Leftrightarrow \quad \theta = \mathbb{E}X \quad (5)$$

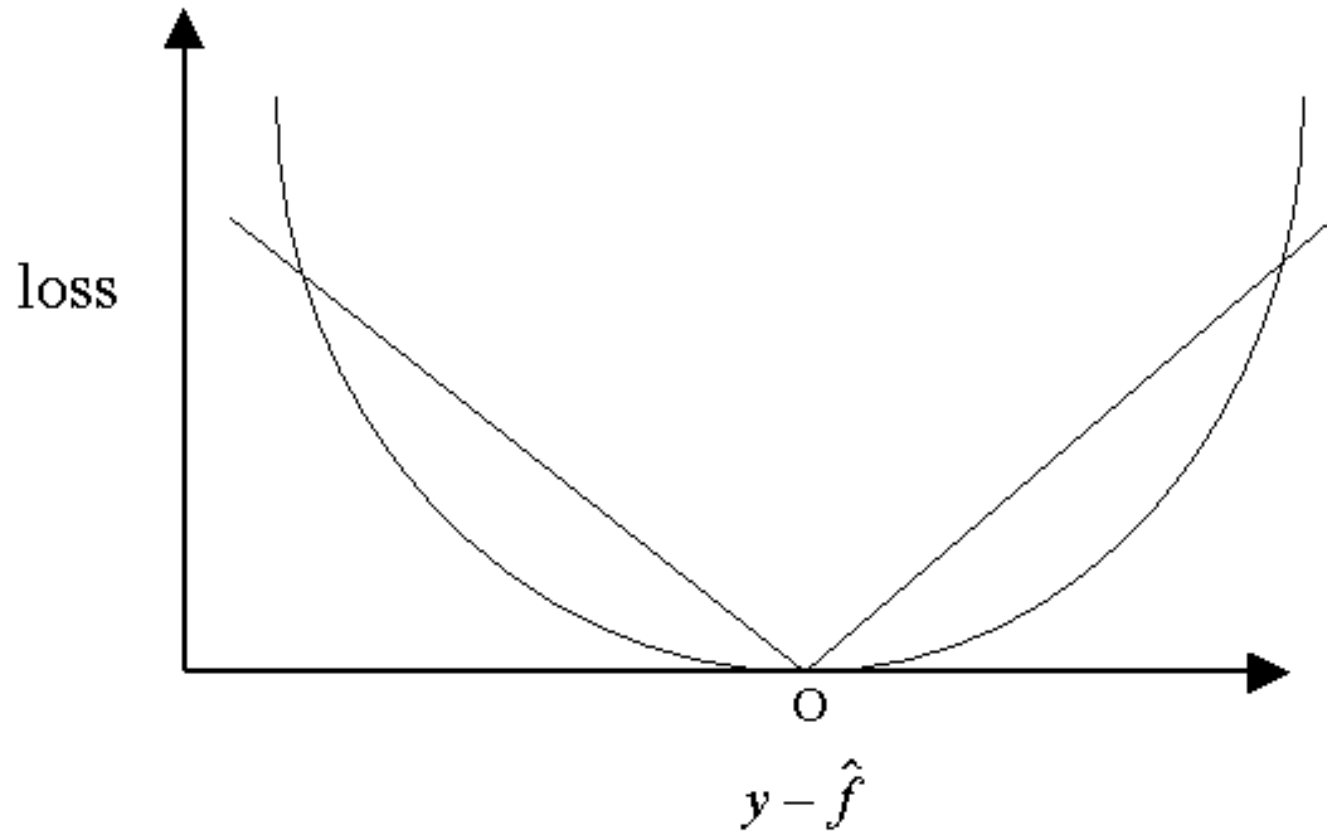
The value which minimizes absolute, or  $L_1$  loss,  $\arg \min_{\theta} \mathbb{E}|X - \theta|$ , is the median:

$$\frac{d}{d\theta} \mathbb{E}|X - \theta| = 0 \quad \Leftrightarrow \quad \theta = m \quad (6)$$

where  $m$  is the median of  $X$ , *i.e.*  $\mathbb{P}(X \leq m) \geq 1/2$  and  $\mathbb{P}(X > m) \geq 1/2$ . (If  $X$  is continuous,

$$\int_{-\infty}^m f(x) dx = \int_m^{\infty} f(x) dx = 1/2.)$$

# Loss Functions



# Robustness: Efficiency

Let's consider the performance of the sample mean  $\bar{X}$ .

**1. Efficiency.** We know it has variance  $\mathbb{V}\bar{X} = \sigma^2/N$ , which is the Cramer-Rao lower bound.

# Robustness: Small Deviations

**2. Small deviations.** Consider a  $\epsilon$ -contamination model. Suppose that we observe

$$X_i \sim \begin{cases} N(\mu, \sigma^2) & \text{with probability } 1 - \epsilon \\ f(x) & \text{with probability } \epsilon \end{cases} \quad (7)$$

where  $f$  is some other distribution, with mean  $\mu_f$  and variance  $\sigma_f^2$ ; then

$$\mathbb{V}\bar{X} = (1 - \epsilon) \frac{\sigma^2}{N} + \epsilon \frac{\sigma_f^2}{N} + \frac{\epsilon(1 - \epsilon)(\mu_f - \mu)^2}{N}. \quad (8)$$

If  $|\mu_f - \mu|$  and  $|\sigma_f - \sigma|$  are small, this is near optimal.

# Robustness: Large Deviations

**3. Large deviations.** Let  $X_{(1)}, \dots, X_{(N)}$  be the data sorted, and let  $T_N$  be a statistic based on this sample. We say  $T_N$  has *breakdown value*  $0 \leq b \leq 1$  if for every  $\epsilon > 0$ ,

$$\lim_{X_{((1-b)N)} \rightarrow \infty} T_N < \infty \quad \text{and} \quad \lim_{X_{((1-b+\epsilon)N)} \rightarrow \infty} T_N = \infty, \quad (9)$$

*i.e.* the percentage of outliers (which go to infinity) in the dataset needed to make  $T_N$  go to infinity.

For the sample mean it is 0. For the median it is 50%.

# M-Estimators

Consider any loss function of the form

$$\sum_{i=1}^N \Phi(x_i - \theta). \quad (10)$$

An estimator which minimizes such a loss function, an *M-estimator*, is the solution to

$$\sum_{i=1}^N \phi(x_i - \theta) = 0 \quad (11)$$

where  $\phi = \Phi'$ . An example is the MLE.

# M-Estimators

Recall that the MLE is asymptotically normal around the true parameter. So is any M-estimator  $\hat{\theta}$ , as well as being consistent, and we can compute its asymptotic variance.

Thus we can compute its asymptotic relative efficiency (ARE) with respect to the optimal variance (which the MLE achieves):

$$\text{ARE}(\hat{\theta}, \theta^*) = \frac{\left( \mathbb{E}_{\theta^*} \phi(X - \hat{\theta}) l'(\hat{\theta} | X) \right)^2}{\mathbb{E}_{\theta^*} \phi(X - \theta^*)^2 \mathbb{E}_{\theta^*} l'(\theta^* | X)^2} \leq 1 \quad (12)$$

where  $\theta^*$  is the true parameter. For example the ARE of the sample median with respect to the sample mean is 0.64. This is the price of robustness.

# Huber Estimator

Consider an M-estimator called the *Huber estimator*, which minimizes

$$\sum_{i=1}^N \Phi(x_i - \theta) \quad (13)$$

where

$$\Phi(t) = \begin{cases} \frac{1}{2}t^2 & \text{if } |t| \leq c \\ c|t| - \frac{1}{2}c^2 & \text{if } |t| \geq c \end{cases} \quad (14)$$

where the constant  $c$  must be chosen.  $\Phi(t)$  is a function which acts like  $t^2$  for  $|t| \leq c$  and like  $|t|$  for  $|t| > c$ , and is continuous and differentiable.

Indeed, its behavior is a compromise between the mean and the median.

# Influence Function

Define the *influence function* of a statistic  $T = T(F)$  at a point  $x$  drawn from distribution  $F$  as

$$U_T(x) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} \quad (15)$$

where  $X \sim F_\epsilon$  if

$$X \sim \begin{cases} F & \text{with probability } 1 - \epsilon \\ x & \text{with probability } \epsilon, \end{cases} \quad (16)$$

*i.e.*  $F_\epsilon$  is a mixture of  $F$  and a point  $x$ .

# Influence Function: Mean

Let's compare the influence functions of the mean and median. Let  $T()$  be the functional that computes the mean of a population. Thus  $T(\hat{F}_N) = \bar{X}$ ,  $T(F) = \mu$ , and

$$T(F_\epsilon) = (1 - \epsilon)T(F) + \epsilon T(x) = (1 - \epsilon)\mu + \epsilon x \quad (17)$$

$$U_T(x) = \lim_{\epsilon \rightarrow 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \lim_{\epsilon \rightarrow 0} \frac{(1 - \epsilon)\mu + \epsilon x - \mu}{\epsilon} = x - \mu. \quad (18)$$

So as  $x$  gets larger, its influence on  $\bar{X}$  becomes larger.

# Influence Function: Median

If  $T()$  computes the median  $m$ , we have

$$U_T(x) = \begin{cases} \frac{1}{2}f(m) & \text{if } x > m \\ -\frac{1}{2}f(m) & \text{otherwise.} \end{cases} \quad (19)$$

So unlike the mean, the median has a bounded influence function.

# Influence Function: Any M-Estimator

For an M-estimator  $\hat{\theta}$  that is the solution to  $\sum_i \phi(x_i - \theta)$  where  $X \sim f$ , the influence function of  $\hat{\theta}$  is

$$U_{\hat{\theta}}(x) = \frac{\phi(x - \theta^*)}{-\int \phi'(t - \theta^*) f(t) dt} = \frac{\phi(x - \theta^*)}{-\mathbb{E}_{\theta^*}(\phi'(X - \theta^*))}. \quad (20)$$

The expected square of the influence function gives the asymptotic variance of  $\hat{\theta}$ , *i.e.*

$$\sqrt{N}(\hat{\theta} - \theta^*) \rightsquigarrow N(0, \mathbb{E}_{\theta^*} U_{\hat{\theta}}^2(X)). \quad (21)$$

# Comparing Estimators

Given a loss function, which estimator? *Decision theory*.

# Comparing Risk

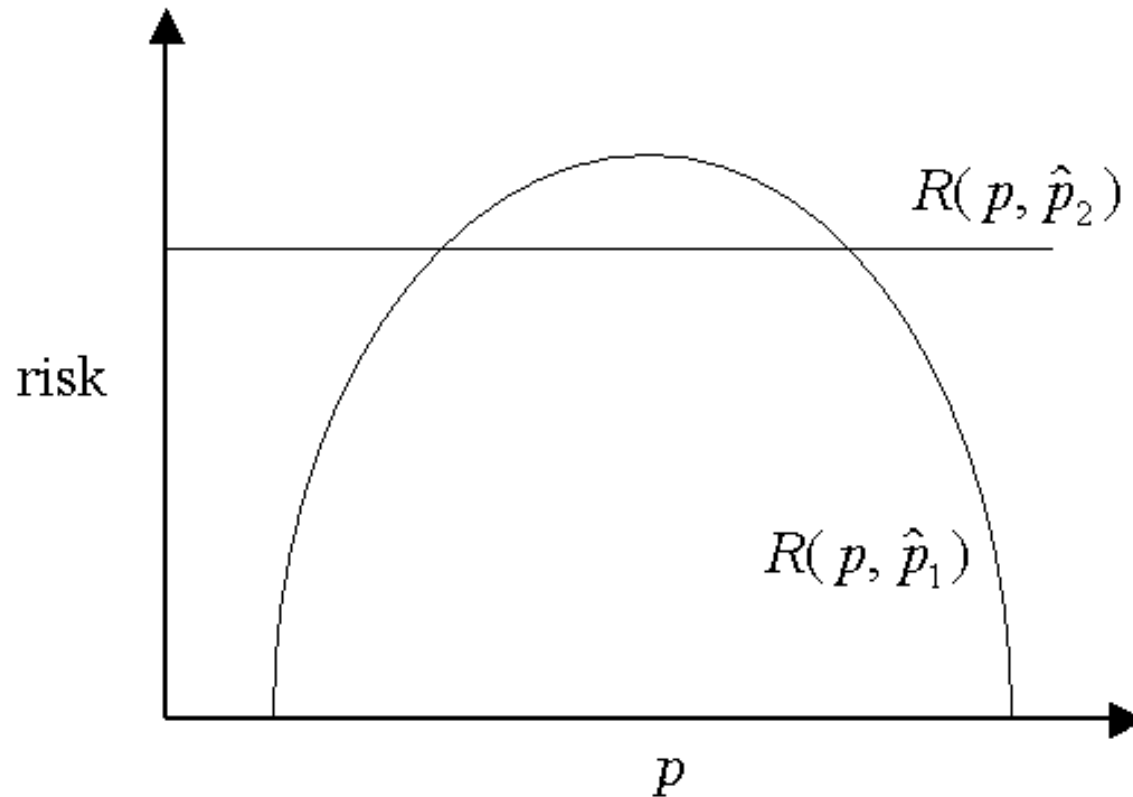
We can compare the risk of an estimator  $\hat{\theta}$

$$R(\theta, \hat{\theta}) = \mathbb{E}_{\theta} \left( L(\theta, \hat{\theta}) \right) = \int L(\theta, \hat{\theta}(x)) f(x; \theta) dx \quad (22)$$

with that of another estimator.

Note that the risk is a function of  $\theta$ .

# Comparing Risk



It may be that neither dominates at all values of  $\theta$ .

# Comparing Risk

We'll look at two one-number summaries of the risk function.

The *maximum risk* is

$$R_{\max}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}). \quad (23)$$

The *Bayes risk* is

$$r(f, \hat{\theta}) = \int R(\theta, \hat{\theta}) f(\theta) d\theta \quad (24)$$

where  $f(\theta)$  is a prior for  $\theta$ .

# Decision Rules

Recall that a *decision rule* is another name for an estimator, and that a decision rule which minimizes the Bayes risk is called a *Bayes rule* or *Bayes estimator*, i.e.  $\hat{\theta}_f$  is a Bayes rule with respect to the prior  $f$  if

$$r(f, \hat{\theta}_f) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}). \quad (25)$$

An estimator which minimizes the maximum risk is called a *minimax rule*, i.e.  $\hat{\theta}$  is minimax if

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta}) \quad (26)$$

where the infimum is over all estimators  $\tilde{\theta}$ .

# Minimax Rules

Finding minimax rules, or showing that something is minimax, is hairy. However, there is at least one easy way. Some Bayes rules are minimax.

Let  $\hat{\theta}_f$  be the Bayes rule for some prior  $f$ :  
 $r(f, \hat{\theta}_f) = \inf_{\tilde{\theta}} r(f, \tilde{\theta})$ . Suppose that

$$R(\theta, \hat{\theta}_f) \leq r(f, \theta_f) \quad \forall \theta. \quad (27)$$

Then  $\hat{\theta}_f$  is minimax and  $f$  is called a *least favorable prior*.

A simple consequence of this is that if a Bayes rule has constant risk  $R(\theta, \hat{\theta}_f) = c$  for some  $c$ , it is minimax.

# MLE is Approximately Minimax

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. For squared error loss, typically

$$R(\theta, \hat{\theta}) = \text{bias}^2 + \mathbb{V}_{\theta}(\hat{\theta}) \quad (28)$$

$$= O(1/N^2) + O(1/N) \quad (29)$$

$$\approx \mathbb{V}_{\theta}(\hat{\theta}) \quad (30)$$

$$\approx 1/N I(\theta). \quad (31)$$

The *Cramer-Rao Inequality* states that this is a lower bound on the variance for any unbiased estimator.

# MLE is Approximately Minimax

Then for any other estimator  $\tilde{\theta}$ ,

$$R(\theta, \tilde{\theta}) \geq R(\theta, \hat{\theta}) \quad (32)$$

for large  $N$ , *i.e.* the MLE is approximately minimax.

So, in most parametric models, with large samples, the MLE is approximately minimax and Bayes.

# Admissibility

Of course, any estimator which is dominated by another at all values of  $\theta$  is undesirable. We say an estimator  $\hat{\theta}$  is *inadmissible* if there exists another rule  $\tilde{\theta}$  such that

$$R(\theta, \tilde{\theta}) \leq R(\theta, \hat{\theta}) \quad \forall \theta \quad \text{and} \quad (33)$$

$$R(\theta, \tilde{\theta}) < R(\theta, \hat{\theta}) \quad \text{for at least one } \theta. \quad (34)$$

Otherwise,  $\hat{\theta}$  is *admissible*.

# Many Normal Means

The *many normal means* problem is a prototype problem which can be shown to be equivalent to general nonparametric regression or density estimation. For this problem, many of our positive results regarding maximum likelihood no longer hold.

Let  $Y_i \sim N(\theta_i, \sigma^2/N)$ ,  $i = 1, \dots, N$ . Let  $Y = (Y_1, \dots, Y_N)$  denote the data and  $\theta = (\theta_1, \dots, \theta_N)$  denote the unknown parameters. Our model class is

$$\mathcal{F} = \left\{ f : \int (f''(x))^2 dx < \infty \right\} \quad (35)$$

for some  $c > 0$ . Note that there are as many parameters as observations.

# MLE is Not Optimal Here

The MLE for this problem is  $\hat{\theta} = Y = (Y_1, \dots, Y_N)$ . Under the loss function  $L(\hat{\theta}, \theta) = \sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2$ , the risk of the MLE is  $R(\hat{\theta}, \theta) = \sigma^2$ .

It can be shown that the minimax risk is approximately  $\sigma^2 / (\sigma^2 + c^2)$  and that there is an estimator  $\tilde{\theta}$  which achieves this risk. In other words, there exists an estimator with smaller risk than that of the MLE, *i.e.* the MLE is inadmissible. In practice the difference in risk can be significant.

So in high-dimensional or nonparametric problems, the MLE is not an optimal estimator. There is also a robustness argument against the MLE in nonparametric settings.

# Bottom Line

So how to choose an estimator based on these tools?

- An inadmissible estimator is definitely bad.
- If you're Bayesian, you like Bayes rules.
- If minimaxity is satisfying to you, go with that.

# Main Things You Should Know

- What is robustness
- What the influence function is
- What it means to be minimax
- What it means to be admissible
- MLE is not best for nonparametric estimation

# Quiz

1. T/F: The Bayes risk is a function of a prior.
2. T/F: There is a tradeoff between statistical efficiency and robustness.
3. T/F: The likelihood is not as robust as  $L_2$ .
4. T/F: MLE is approximately minimax.