Selected Topics in Statistical Decision Theory

Scribe: Jingxuan Xu
Lecturer: Prof. Harrison Zhou

May 1, 2025

Contents

1 Nonparametric Estimation 5

1.1 Gaussian Sequence Model . . . 5

1.1.1 Discrete Fourier Transform . . . 5

1.1.2 Sobolev Ellipsoid . . . 6

1.1.3 Minimax Risk Upper and Lower Bound . . . 6

1.2 Best Linear Procedure . . . 9

1.2.1 Pinsker Upper Bound . . . 10

1.2.2 Pinsker Lower Bound . . . 11

1.3 Adaptive Estimation . . . 14

1.3.1 Review: James-Stein Estimator . . . 14

1.3.2 Blockwise James Stein Estimator . . . 15

2 Frequentist Bayesian Investigation 17

2.1 Posterior Contraction . . . 17

2.1.1 Example: Gaussian Sequence Model . . . 18

2.1.2 Prior Mass and Testing . . . 19

3 High Dimensional Estimation 23

3.1 Lasso Estimator . . . 23

3.1.1 Gaussian Sequence Model . . . 23

3.1.2 High-dimensional Linear Regression . . . 25

3.2 Structured Vector/Matrix Estimation . . . 26

3.2.1 Sparse Vector Estimation with All Subset Selection . . . 26

3.2.2 Stochastic Block Model and Graphon Estimation . . . 27

3.3 Some Negative Results for Bayesian Estimation . . . 29

3.3.1 Natural Prior may not Work (LASSO prior Example) . . . 29

3.3.2 Le-Cam Schwarz Argument may Fail . . . 31

3.4 High-Dimensional Inference . . . 31

4 Topics on Diffusion Model 35

4.1 Kernel Smoothing . . . 35

4.1.1 Upper Bound . . . 35

4.1.2 Lower Bound . . . 37

4.2 Diffusion and Score Estimation . . . 40

5 More on Lower Bounds 43

5.1 Le Cam's Two-Points Argument . . . 43

5.2 Assouad's Lemma . . . 44

5.3 Fano's Lemma . . . 44

This note is a scribe of S&DS611 course: Selected Topics in Statistical Decision Theory taught by Prof. Harrison Zhou at Yale University.

Chapter 1

Nonparametric Estimation

1.1 Gaussian Sequence Model

1.1.1 Discrete Fourier Transform

Consider the nonparametric regression model

$$Y_i = f(X_i) + Z_i, \quad i = 1, 2, \cdots, n,\ Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

where X_i are deterministic and $f \in L_2[0,1] : [0,1] \to \mathbb{R}$ is a periodic function. We want to estimate f. Consider the case $X_i = i/n$. Let θ_i be the Fourier coefficients of f w.r.t. orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ of $L_2[0,1]$:

$$\theta_i = \int_0^1 f(x)\varphi_i(x)\,\mathrm{d}x, \quad f(x) = \sum_{i=1}^{\infty} \theta_i \varphi_i(x)$$

A natural estimator for θ_i is

$$\hat{\theta}_i = \frac{1}{n}\sum_{j=1}^{n} Y_j \varphi_i(X_j)$$

Here $X_i = i/n$ are nicely spread in $[0,1]$. Then, for large n,

$$\frac{1}{n}\sum_{i=1}^{n} \phi_j(X_i)\varphi_k(X_i) \approx \int_0^1 \varphi_j(x)\varphi_k(x)\,\mathrm{d}x = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \tag{1.1}$$

Hence $\{\varphi_i(X_j)\}_{j=1}^{\infty}$ approximately behaves like an orthonormal system in the discrete sense. The approximate expectation and variance is

$$\mathbb{E}\left[\hat{\theta}_i\right] = \frac{1}{n}\sum_{j=1}^{n} \varphi_i(X_j)\mathbb{E}[Y_j] = \frac{1}{n}\sum_{j=1}^{n} \varphi_i(X_j) f(X_j) \approx \int_0^1 \varphi_i(x) f(x) = \theta_i$$

$$\mathrm{Var}\left[\hat{\theta}_i\right] = \frac{1}{n^2}\sum_{j=1}^{n} \varphi_i(X_j)^2 \mathrm{Var}[Y_j] = \frac{1}{n^2}\sum_{i=1}^{n} \phi_i(X_j)^2 \approx \frac{1}{n}$$

where the last step uses Equation (1.4). Put it together, by central limit theorem, we have

$$\hat{\theta}_i = \theta_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \cdots, \; Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

which is the form of **Gaussian sequence model**.

1.1.2 Sobolev Ellipsoid

We assume the regression function f is sufficiently smooth. We will assume that it belongs to the *periodic Sobolev class*

$$W^{\text{per}}(\alpha, L) = \left\{ f : [0,1] \to \mathbb{R} \text{ periodic } : f^{(\alpha-1)} \text{ is absolutely continuous and } \int_0^1 (f^{(\alpha)}(x))^2 \, \mathrm{d}x \leqslant L^2 \right\}$$

It can be proved that, this function space, after mapping to the sequence space using discrete Fourier transform with trigonometric basis $\phi_1(x) = 1, \phi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \phi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx),\ k = 1, 2, \cdots,$ is isomorphic to the ellipsoid

$$\Theta' = \left\{ \theta : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \leqslant M \right\}$$

where $a_1 = 0, a_{2i} = a_{2i+1} = (2i)^\alpha$. This is called a **Sobolev ellipsoid**. For proof of this result, see Tsybakov Lemma A.3. Often time, we set $a_i = i^\alpha$ so that

$$\Theta = \left\{ \theta : \sum_{i=1}^{\infty} i^{2\alpha} \theta_i^2 \leqslant M \right\}$$

One can show that these two sets' minimax rates are very close, and $\Theta \subseteq \Theta'$. See Johnstone Section 3.1.

1.1.3 Minimax Risk Upper and Lower Bound

Consider the Gaussian sequence model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, 2, \cdots, \; Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

with parameter space being the Sobolev ellipsoid

$$\Theta = \left\{ \theta : \sum_{i=1}^{\infty} i^{2\alpha} \theta_i^2 \leqslant M \right\}$$

Our goal is to find the lower bound and upper bound of minimax risk.

1. Upper Bound

Consider the estimator

$$\hat{\theta}_i = \begin{cases} Y_i, & i \leqslant k \\ 0, & i > k \end{cases}$$

The risk function w.r.t. the squared loss is then

$$\begin{aligned}\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\right] &= \mathbb{E}\left[\sum_{i=1}^{k}(Y_i-\theta_i)^2\right]+\mathbb{E}\left[\sum_{i=k+1}^{\infty}\theta_i^2\right] \\ &= \sum_{i=1}^{k}\mathbb{E}\left[\frac{1}{n}Z_i^2\right]+\sum_{i=k+1}^{\infty}\theta_i^2 && (Y_i-\theta_i=\tfrac{1}{\sqrt{n}}Z_i) \\ &\leqslant \frac{k}{n}+\frac{1}{(k+1)^{2\alpha}}\sum_{i=k+1}^{\infty}i^{2\alpha}\theta_i^2 \\ &\leqslant \frac{k}{n}+\frac{M}{k^{2\alpha}} && (\theta\in\Theta)\end{aligned}$$

Now we want to get the tightest upper bound by choosing appropriate k. Set the variance and bias^2 equal, we have

$$\frac{k}{n}=\frac{M}{k^{2\alpha}} \quad\Longrightarrow\quad k=(Mn)^{\frac{1}{2\alpha+1}}$$

We then get the upper bound:

$$\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\right]\leqslant\frac{(Mn)^{\frac{1}{2\alpha+1}}}{n}+\frac{M}{(Mn)^{\frac{2\alpha}{2\alpha+1}}}=2M^{\frac{1}{2\alpha+1}}n^{-\frac{2\alpha}{2\alpha+1}}=Cn^{-\frac{2\alpha}{2\alpha+1}}$$

where C is a constant not depending on n. This is lower than typical parametric models, which have convergence rate n^{-1}. Since it holds for all $\theta\in\Theta$, we have the minimax upper bound

$$\inf_{\hat{\theta}}\sup_{\theta\in\Theta}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right]\leqslant Cn^{-\frac{2\alpha}{2\alpha+1}}$$

2. Lower Bound

To find the lower bound, we seek a sub-parameter space Θ_0. We choose

$$\Theta_0=\left\{\theta:\theta_i=\frac{1}{\sqrt{n}},\ i\leqslant k,\quad \theta_i=0,\ i\geqslant k+1\right\}$$

where $k=(Mn)^{\frac{1}{1+2\alpha}}$. To show that it is indeed a subspace of Θ, note that

$$\sum_{i=1}^{\infty}i^{2\alpha}\theta^2=\sum_{i=1}^{k}\frac{1}{n}i^{2\alpha}\leqslant\frac{k}{n}k^{2\alpha}=M$$

In this smaller parameter space, we can always truncate the tail from $k+1$ term, since we can always precisely estimate these 0's. For any estimator $\hat{\theta}$, we have

$$\sup_{\theta\in\Theta}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right]\geqslant\sup_{\theta\in\Theta_0}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right]$$

since $\Theta_0 \subseteq \Theta$. Take infimum on both sides, we have that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \geqslant \inf_{\hat{\theta}} \sup_{\theta \in \Theta_0} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \tag{1.2}$$

Therefore, we only need to find the lower bound in this sub-paremeter space. Take prior distribution π of θ_i: $\mathbb{P}(\theta_i = 0) = \mathbb{P}(\theta_i = \frac{1}{\sqrt{n}}) = 1/2$. Let $\hat{\theta}_{i,Bayes}$ denotes the Bayes estimator of θ_i under this prior. Let $\varphi_{\alpha,\beta}$ be the normal density with mean α and variance β. Since minimax risk is lower bounded by the average risk, we have

$$\begin{aligned}
\sup_{\theta \in \Theta_0} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] &\geqslant \mathbb{E}_{\theta}\mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] && \text{(worst-case risk greater than average risk)}\\
&\geqslant \mathbb{E}_{\theta}\mathbb{E}_{Y|\theta}\left[\sum_{i=1}^{k}(\hat{\theta}_{i,Bayes}-\theta_i)^2\right] && \text{(Truncate to } k\text{)}\\
&= \sum_{i=1}^{k}\left(\frac{1}{2}\mathbb{E}_{Y_i|\theta_i=0}\left[\left(\hat{\theta}_{i,Bayes}-0\right)^2\right] + \frac{1}{2}\mathbb{E}_{Y_i|\theta_i=\frac{1}{\sqrt{n}}}\left[\left(\hat{\theta}_{i,Bayes}-\frac{1}{\sqrt{n}}\right)^2\right]\right)\\
&= \sum_{i=1}^{k}\left(\frac{1}{2}\int\left(\hat{\theta}_{i,Bayes}\right)^2\varphi_{0,\frac{1}{n}}(x_i)\,\mathrm{d}x_i + \frac{1}{2}\left(\hat{\theta}_{i,Bayes}-\frac{1}{\sqrt{n}}\right)^2\varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x_i)\,\mathrm{d}x_i\right)\\
&\geqslant \sum_{i=1}^{k}\frac{1}{2}\int\left(\hat{\theta}_{i,Bayes}^2 + \left(\hat{\theta}_{i,Bayes}-\frac{1}{\sqrt{n}}\right)^2\right)\min\left\{\varphi_{0,\frac{1}{n}}(x_i), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x_i)\right\}\,\mathrm{d}x_i\\
&\geqslant \sum_{i=1}^{k}\frac{1}{2}\int\left(\frac{1}{2}\left(\frac{1}{\sqrt{n}}\right)^2\right)\min\left\{\varphi_{0,\frac{1}{n}}(x_i), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x_i)\right\}\,\mathrm{d}x_i && (a^2+b^2 \geqslant \tfrac{1}{2}(a-b)^2)\\
&= \frac{k}{4n}\int\min\left\{\varphi_{0,\frac{1}{n}}(x), \varphi_{\frac{1}{\sqrt{n}},\frac{1}{n}}(x)\right\}\,\mathrm{d}x
\end{aligned}$$

(HW 1: Show that this integral is bounded away by a positive constant c')

where $c' > 0$ is a positive constant that is not dependent on n. Therefore, we can get the lower bound

$$\sup_{\theta \in \Theta_0} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \geqslant \frac{c'k}{4n} = \frac{c'}{4n}(Mn)^{\frac{1}{1+2\alpha}} = \frac{c'}{4}M^{\frac{1}{1+2\alpha}}n^{-\frac{2\alpha}{2\alpha+1}} = cn^{-\frac{2\alpha}{2\alpha+1}}$$

Since this is true for all $\hat{\theta}$, we take infimum on both sides, and get

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_0} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \geqslant cn^{-\frac{2\alpha}{2\alpha+1}}$$

Then, combine this with Equation 1.2, we have the final lower bound

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \geqslant \inf_{\hat{\theta}} \sup_{\theta \in \Theta_0} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \geqslant cn^{-\frac{2\alpha}{2\alpha+1}}$$

With the upper bound and the lower bound, we have

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}-\theta\|_2^2\right] \asymp n^{-\frac{2\alpha}{2\alpha+1}}$$

where $f(n) \asymp g(n)$ denotes the case when there exists two positive constants c, C independent of n such that $cg(n) \leqslant f(n) \leqslant Cg(n)$.

1.2 Best Linear Procedure

To get a better bound, instead of just cutting the estimator discretely at a specific k (as what we did in upper bound calculation, where we set $\hat{\theta}_i = Y_i$ when $i \leqslant k$ and $\hat{\theta}_i = 0$ otherwise), we can make it gradually decreases to 0, where we introduce a linear procedure $c_i Y_i$ here. We consider a more general Gaussian sequence model

$$Y_i = \theta_i + \sigma Z_i, \quad i = 1, 2, \cdots, \ Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

with variance σ^2 instead of $1/n$. Consider $c = (c_1, c_2, \cdots)$, where each $c_i \in [0,1]$ and the estimator $\hat{\theta}_i = c_i Y_i$. Then, the minimax risk

$$\begin{aligned}
\inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \mathbb{E}\left[\sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2\right] &= \inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \sum_{i=1}^{\infty} \mathbb{E}\left[(c_i Y_i - \theta_i)^2\right] \\
&= \inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \sum_{i=1}^{\infty} \left(c_i^2 \mathbb{E}[Y_i^2] - 2c_i\theta_i \mathbb{E}[Y_i] + \theta_i^2\right) \\
&= \inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \sum_{i=1}^{\infty} \left(c_i^2(\sigma^2 + \theta_i^2) - 2c_i\theta_i^2 + \theta_i^2\right) \\
&= \inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \sum_{i=1}^{\infty} \left[\underbrace{(c_i - 1)^2\theta_i^2}_{\text{bias}^2} + \underbrace{c_i^2\sigma^2}_{\text{variance}}\right] \qquad (1.3)
\end{aligned}$$

The goal here is to find the optimal c (find the *best linear procedure*) that attains this minimax risk. To make the calculation simpler, we use the concave-convex function version of minimax theorem: Let X be a convex set and Y be a convex, compact set. If $f : X \times Y \to \mathbb{R}$ such that

- $f(\cdot, y)$ is continuous and convex on X, for any fixed $y \in Y$, and
- $f(x, \cdot)$ is continuous and concave on Y, for any fixed $x \in X$

Then,

$$\inf_{x\in X} \sup_{y\in Y} f(x,y) = \sup_{y\in Y} \inf_{x\in X} f(x,y)$$

In our case, if we set $\theta_i^2 = \omega_i$, then

$$X = [0,1]^\infty, \quad Y = \left\{\omega : \sum_{i=1}^{\infty} a_i^2 \omega_i \leqslant M\right\}$$

Our $f(c,\omega) = \sum_{i=1}^{\infty} \left[(c_i - 1)^2\omega_i + c_i^2\sigma^2\right]$. Then, it is convex on c (quadratic), and concave on ω (linear). Therefore, we can continue 1.3 and get

$$\inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \mathbb{E}\left[\sum_{i=1}^{\infty} (c_i Y_i - \theta_i)^2\right] = \inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \sum_{i=1}^{\infty} \left[(c_i - 1)^2\theta_i^2 + c_i^2\sigma^2\right] = \sup_{\theta\in\Theta} \inf_{c\in[0,1]^\infty} \sum_{i=1}^{\infty} \left[(c_i - 1)^2\theta_i^2 + c_i^2\sigma^2\right]$$

To find the optimal c, we set derivative w.r.t. each c_i to zero:

$$2(c_i - 1)\theta_i^2 + 2c_i\sigma^2 = 0 \quad \Longrightarrow \quad c_i = \frac{\theta_i^2}{\theta_i^2 + \sigma^2}$$

Substitute this c_i into the equation, we have

$$\begin{aligned}
\inf_{c\in[0,1]^\infty} \sup_{\theta\in\Theta} \mathbb{E}\left[\sum_{i=1}^{\infty}(c_iY_i-\theta_i)^2\right] &= \sup_{\theta\in\Theta}\sum_{i=1}^{\infty}\left[\left(\frac{\theta_i^2}{\theta_i^2+\sigma^2}-1\right)^2\theta_i^2+\left(\frac{\theta_i^2}{\theta_i^2+\sigma^2}\right)^2\sigma^2\right] \\
&= \sup_{\theta\in\Theta}\sum_{i=1}^{\infty}\left[\left(\frac{\sigma^2}{\theta_i^2+\sigma^2}\right)^2\theta_i^2+\left(\frac{\theta_i^2}{\theta_i^2+\sigma^2}\right)^2\sigma^2\right] \\
&= \sup_{\theta\in\Theta}\sum_{i=1}^{\infty}\frac{\sigma^2\theta_i^2(\sigma^2+\theta_i^2)}{(\theta_i^2+\sigma^2)^2} = \sup_{\theta\in\Theta}\sum_{i=1}^{\infty}\frac{\sigma^2\theta_i^2}{\theta_i^2+\sigma^2}
\end{aligned}$$

Now, this becomes a constrained (on the set Θ) optimization problem. We introduce *Lagrangian multiplier*

$$L(\theta)=\sum_{i=1}^{\infty}\frac{\sigma^2\theta_i^2}{\theta_i^2+\sigma^2}-\frac{1}{\lambda^2}\left(\sum_{i=1}^{\infty}a_i^2\theta_i^2-M\right)$$

Here we use $1/\lambda^2$ instead of λ just for some simplicity of representation later. Reparametrize by $\omega_i=\theta_i^2$ and set the derivative to zero:

$$\frac{\partial L(\omega)}{\partial\omega_i}=\frac{\sigma^2(\sigma^2+\omega_i)-\sigma^2\omega_i}{(\sigma^2+\omega_i)^2}-\frac{1}{\lambda^2}a_i^2=\frac{\sigma^4}{(\sigma^2+\omega_i)^2}-\frac{1}{\lambda^2}a_i^2=0$$

$$\Longleftrightarrow \frac{\sigma^2}{\sigma^2+\omega_i}=\frac{a_i}{\lambda} \quad \Longleftrightarrow \quad \lambda\sigma^2=a_i\sigma^2+a_i\omega_i \quad \Longleftrightarrow \quad \omega_i=\frac{\lambda\sigma^2-a_i\sigma^2}{a_i}=\left(\frac{\lambda}{a_i}-1\right)\sigma^2$$

Since $\omega_i\geqslant 0$, typically we choose

$$\theta_i^{*^2}=\left(\frac{\lambda}{a_i}-1\right)_+\sigma^2 \tag{1}$$

Note that we directly take derivative of $L(\theta)$ since $\frac{\sigma^2\theta_i^2}{\theta_i^2+\sigma^2}$ is an increasing function w.r.t. each ω_i, thus the constraint (sobolev ball) is in active. Therefore, we choose λ at the boundary:

$$\lambda^*:\sum_{i=1}^{\infty}a_i^2\left(\frac{\lambda^*}{a_i}-1\right)_+\sigma^2=\sum_{i=1}^{\infty}a_i\left(\lambda^*-a_i\right)_+\sigma^2=M \tag{2}$$

The corresponding procedure is then:

$$c_i^*=\frac{\theta_i^{*^2}}{\theta_i^{*^2}+\sigma^2}=\frac{\left(\frac{\lambda^*}{a_i}-1\right)_+\sigma^2}{\left(\frac{\lambda^*}{a_i}-1\right)_+\sigma^2+\sigma^2}=\begin{cases}1-\dfrac{a_i}{\lambda^*}, & a_i\leqslant\lambda^*\\ 0, & a_i>\lambda^*\end{cases} \tag{3}$$

The cooresponding best linear risk is then:

$$R_L^*(\Theta)=\sum_{i=1}^{\infty}\frac{\theta_i^{*^2}\sigma^2}{\sigma^2+\theta_i^{*^2}}=\sum_{i=1}^{\infty}c_i^*\sigma^2 \tag{4}$$

where the subscription 'L' means 'linear'.

1.2.1 Pinsker Upper Bound

Solving λ^*

Since in sobolev ellipsoid, we make $a_i\to\infty$ for compactness, there must be some point such that $a_i>\lambda^*$, then $c_i^*=0$

by formula, and the best linear procedure terminates. Therefore, we don't actually need to calculate the infinite sum when solving λ^*.

For example, if we recover $a_i = i^\alpha$ and $\sigma^2 = 1/n$, then

$$\begin{aligned}
&\sum_{i=1}^{\infty} i^\alpha \left(\lambda^* - i^\alpha\right)_+ \frac{1}{n} = \sum_{i=1}^{I^*} i^\alpha (\lambda^* - i^\alpha)_+ \frac{1}{n} = M, \quad (I^*)^\alpha \leqslant \lambda^* < (I^*+1)^\alpha \\
\Longrightarrow &\lambda^* \left(\sum_{i=1}^{I^*} i^\alpha\right) - \sum_{i=1}^{I^*} i^{2\alpha} = Mn
\end{aligned}$$

Use the integration to approximate the sum. Note that when I^* is sufficiently large,

$$\sum_{i=1}^{I^*} i^\alpha = \frac{1}{\alpha+1}(I^*)^{\alpha+1}(1+o(1))$$

by Euler-Maclaurin. Choose $\lambda^* = (I^*)^\alpha(1+o(1))$ (since they have at most difference $(I^*+1)^\alpha - (I^*)^\alpha$, and this difference cancelled out the highest order term), we have

$$\begin{aligned}
&\left[\lambda^* \frac{1}{\alpha+1}(I^*)^{\alpha+1} - \frac{1}{2\alpha+1}(I^*)^{2\alpha+1}\right](1+o(1)) = Mn \\
\Longrightarrow \quad &(I^*)^{2\alpha+1}\left(\frac{1}{\alpha+1} - \frac{1}{2\alpha+1}\right) = Mn(1+o(1)) \\
\Longrightarrow \quad &I^* = \left(\frac{Mn}{\frac{1}{\alpha+1} - \frac{1}{2\alpha+1}}\right)^{\frac{1}{2\alpha+1}}(1+o(1)) = \left(\frac{Mn(\alpha+1)(2\alpha+1)}{\alpha}\right)^{\frac{1}{2\alpha+1}}(1+o(1))
\end{aligned}$$

(HW2: Find the Pinsker constant)

Therefore, we get the minimax upper bound: For any $\hat{\theta}$, linear or nonlinear, we have

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|_2^2\right] \leqslant R_L^*(\Theta) = c_p(1+o(1))n^{-\frac{2\alpha}{2\alpha+1}}$$

1.2.2 Pinsker Lower Bound

A lower bound is much more difficult. One typically constructs a prior, and use the average risk to bound. What prior should we choose? Naively, we can just choose the Gaussian prior as our first trial.

Naive attempt: Gaussian prior π

Let $\theta_i \sim N(0, \tau_i^2)$, where θ_i's are independent. Then, by the knowledge of Bayesian estimator with Gaussian prior, we have

$$\mathbb{E}\left[\theta_i | Y_i\right] = \frac{\tau_i^2}{\tau_i^2 + \frac{1}{n}} Y_i$$

Optimally, one choose

$$\tau_i^2 = \left(\frac{\lambda^\star}{a_i} - 1\right)_+ \frac{1}{n}$$

to get the least favorable prior, and we have

$$\mathbb{E}\left[\theta_i|Y_i\right] = \frac{\left(\frac{\lambda^\star}{a_i}-1\right)_+}{\left(\frac{\lambda^\star}{a_i}-1\right)_+ + 1} Y_i = \begin{cases} \left(1-\frac{a_i}{\lambda^\star}\right)Y_i, & \text{if } \lambda^\star \geqslant a_i \\ 0, & \text{if } \lambda^\star < a_i \end{cases} = \left(1-\frac{a_i}{\lambda^\star}\right)_+ Y_i$$

However, this prior is not supported on Θ (indeed, on $\mathbb{R}^\infty$). Indeed,

$$\mathbb{E}\left[\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2\right] = \sum_{i=1}^{\infty} i^{2\alpha}\tau_i^2 = \sum_{i=1}^{\infty} i^{2\alpha}\left(\frac{\lambda^\star}{i^\alpha}-1\right)_+ \frac{1}{n} = \sum_{i=1}^{\infty} i^{\alpha}\left(\lambda^\star - i^\alpha\right)_+ \frac{1}{n} = M$$

It is exactly on the boundary. Thus, about half time, the θ is outside the ball. This is not expected. However, notice that

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2\right) &= \sum_{i=1}^{\infty} i^{4\alpha}\mathrm{Var}\left(\theta_i^2\right) = \sum_{i=1}^{\infty} i^{4\alpha}\cdot 2\tau_i^4 \qquad \text{(Last equality follows from variance of chi-squared)} \\ &= 2\sum_{i=1}^{\infty} i^{4\alpha}\left(\frac{\lambda}{i^\alpha}-1\right)_+^2 \frac{1}{n^2} \\ &= 2\sum_{i=1}^{\infty} i^{2\alpha}(\lambda^\star - i^\alpha)_+^2 \frac{1}{n^2} \\ &\leqslant 2\sum_{i=1}^{I^\star} i^{2\alpha}(\lambda^\star)^2 \frac{1}{n^2} \asymp n^{-\frac{1}{1+2\alpha}} \end{aligned}$$

where the last line follows $I^\star \asymp n^{\frac{1}{1+2\alpha}}$, and $\lambda^\star = (I^\star)^\alpha$, and then $\sum_{i=1}^{I^\star} i^{2\alpha} \asymp I^\star \cdot (I^\star)^{2\alpha} \asymp n$, and $(\lambda^\star)^2 \asymp n^{\frac{2\alpha}{2\alpha+1}}$. The variance is extremely small.

Second Attempt: Move into interior

Therefore, if we just move the mean a little bit into the interior, most of the time the θ would be inside the ball. Thus, instead, we pick

$$\pi_1 : \tau_{\star.i}^2 = \left(\frac{\lambda^*}{a_i}-1\right)_+ \frac{1}{n}\left(1-\frac{1}{\log n}\right), \quad \theta_i \sim N(0, \tau_{\star,i}^2)$$

The $1/\log n$ can be replaced by any factor that have $o(1)$ rate. In this case,

$$\mathbb{E}\left[\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2\right] = \left(1-\frac{1}{\log n}\right)M$$

the mean is inside the ball. Moreover,

$$\mathrm{Var}\left(\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2\right) \lesssim \left(1-\frac{1}{\log n}\right)^2 n^{-\frac{1}{1+2\alpha}}$$

The variance is in a small scale. Using the Gaussian tail bound that for $X \sim N(0, \sigma^2)$, $\mathbb{P}[X \geqslant t] \lesssim e^{-t^2/2\sigma^2}$ for all $t \geqslant 0$, we get the consequence

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 \geqslant M\right] &= \mathbb{P}\left[\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 - \left(1 - \frac{1}{\log n}\right) M \geqslant M - \left(1 - \frac{1}{\log n}\right) M\right] \\
&= \mathbb{P}\left[\sum_{i=1}^{\infty} i^{2\alpha}\theta_i^2 - \left(1 - \frac{1}{\log n}\right) M \geqslant \frac{1}{\log n} M\right] \\
&\lesssim e^{-c\left(\frac{1}{\log n}M\right)^2/n^{-\frac{1}{1+2\alpha}}} \leqslant o(n^{-k}), \quad \forall k \geqslant 0
\end{aligned}$$

It decreses with a faster speed than any power rate, e.g., $o(n^{-100})$. Thus, 'almost all the time' the θ would be in the parameter space. Therefore, we can finally choose the proper prior such that:

$$\pi^\star : \text{Generate } \theta \sim \pi_1 \text{ and throw away } \theta \text{ if it is outside } \Theta$$

Therefore,

$$\mathrm{d}\pi^\star(\theta) = \frac{\mathrm{d}\pi_1(\theta)\mathbb{1}_{\{\theta \in \Theta\}}}{\pi_1(\Theta)}$$

We can bound below using the average risk such that

$$\begin{aligned}
\sup_\Theta \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] &\geqslant \mathbb{E}_{\theta \sim \pi^\star} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] \\
&= \frac{1}{\pi_1(\Theta)} \int \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] \mathbb{1}_{\{\theta \in \Theta\}} \, \mathrm{d}\pi_1(\theta) \\
&= \frac{1}{\pi_1(\Theta)} \mathbb{E}_{\theta \sim \pi_1}\left[\mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] \mathbb{1}_{\{\theta \in \Theta\}}\right] \\
&= \frac{1}{\pi_1(\Theta)} \mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] \\
&= (1 - o(1)) \mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right]
\end{aligned}$$

where the last line follows since $\mathbb{P}[\theta \notin \Theta] \leqslant o(n^{-k})$, and thus $\pi_1(\Theta) = 1 - o(1)$. Substitute using the Bayes estimator

$$\sup_\Theta \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] \geqslant \mathbb{E}_{\theta \sim \pi^\star} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] = (1 - o(1)) \mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta} - \theta\|^2\right] \geqslant (1 - o(1)) \mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}_{Bayes} - \theta\|^2\right]$$

Now we directly input our chosen $\hat{\theta}_{Bayes,i} = \frac{\tau_i^2}{\tau_i^2 + \sigma^2} Y_i$, and we can get

$$\begin{aligned}
\mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta}\left[\|\hat{\theta}_{Bayes} - \theta\|^2\right] &= \sum_{i=1}^{\infty} \mathbb{E}_{\theta_i \sim \pi_1} \mathbb{E}_{Y_i|\theta_i}\left[\left(\hat{\theta}_{Bayes,i} - \theta_i\right)^2\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}_{\theta_i \sim \pi_1} \mathbb{E}_{Y_i|\theta_i}\left[\left(\frac{\tau_i^2}{\tau_i^2 + \sigma^2} Y_i - \theta_i\right)^2\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}_{\theta_i \sim \pi_1} \mathbb{E}_{Y_i|\theta_i}\left[\left(\frac{\tau_i^2}{\tau_i^2 + \sigma^2}\right)^2 Y_i^2 - 2\theta_i \frac{\tau_i^2}{\tau_i^2 + \sigma^2} Y_i + \theta_i^2\right] \qquad (1.4)
\end{aligned}$$

Under θ_i, $Y_i \sim N(\theta_i, \sigma^2)$, where $\sigma^2 = 1/n$. Therefore, $\mathbb{E}[Y_i^2] = \mathrm{Var}(Y_i) + \mathbb{E}[Y_i]^2 = \sigma^2 + \theta_i^2$. Then,

$$\begin{aligned}
(1.4) &= \sum_{i=1}^{\infty} \mathbb{E}_{\theta_i \sim \pi_1} \left[\left(\frac{\tau_i^2}{\tau_i^2 + \sigma^2} \right)^2 (\sigma^2 + \theta_i^2) - 2\theta_i^2 \frac{\tau_i^2}{\tau_i^2 + \sigma^2} + \theta_i^2 \right] \\
&= \sum_{i=1}^{\infty} \left[\left(\frac{\tau_i^2}{\tau_i^2 + \sigma^2} \right)^2 (\sigma^2 + \tau_i^2) - 2\tau_i^2 \frac{\tau_i^2}{\tau_i^2 + \sigma^2} + \tau_i^2 \right] && (\theta_i \sim N(0, \tau_i^2)) \\
&= \sum_{i=1}^{\infty} \frac{\tau_i^4 \sigma^2 + \tau_i^6 - 2\tau_i^4(\tau_i^2 + \sigma^2) + \tau_i^2(\tau_i^2 + \sigma^2)^2}{(\tau_i^2 + \sigma^2)^2} \\
&= \sum_{i=1}^{\infty} \frac{\tau_i^2 \sigma^2}{\tau_i^2 + \sigma^2} = \sum_{i=1}^{\infty} c_i^* \frac{1}{n} = R_L^\star(\Theta)
\end{aligned}$$

Therefore, we have the conclusion:

$$\sup_{\Theta} \mathbb{E}\left[\|\hat{\theta} - \theta\|^2 \right] \geqslant (1 - o(1)) \mathbb{E}_{\theta \sim \pi_1} \mathbb{E}_{Y|\theta} \left[\|\hat{\theta} - \theta\|^2 \right] \geqslant (1 - o(1)) R_L^\star(\Theta)$$

Since this is true for all estimator $\hat{\theta}$, take infimum on both sides,

$$R_N(\Theta) = \inf_{\hat{\theta}} \sup_{\Theta} \mathbb{E}\left[\|\hat{\theta} - \theta\|^2 \right] \geqslant (1 - o(1)) R_L^\star(\Theta)$$

Since we have proven that the upper bound is $R_L^*(\Theta) = c_p(1 + o(1)) n^{-\frac{2\alpha}{2\alpha+1}}$, combining upper and lower bound, we have

$$c_p(1 - o(1)) n^{-\frac{2\alpha}{2\alpha+1}} \leqslant \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\left[\|\hat{\theta} - \theta\|^2 \right] \leqslant c_p(1 + o(1)) n^{-\frac{2\alpha}{2\alpha+1}}$$

This is a sharper bound than the naive minimax upper and lower bounds. We can also see that the linear procedure is asymptotically the best procedure, since it only differ with R_N by a factor $(1 + o(1))$.

1.3 Adaptive Estimation

Can we have a procedure without knowledge of α and M, but still achieve the same upper bound? The answer is yes. This is called the *adaptive estimation.*

1.3.1 Review: James-Stein Estimator

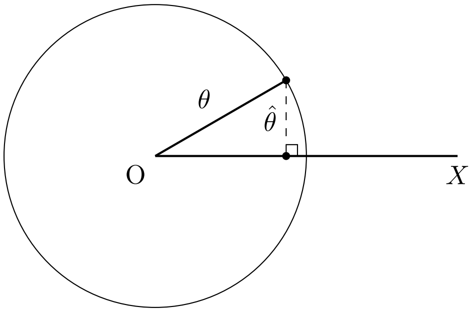
Intuition:
For multivariate standard normal distributed variable $Z \sim N(0, I_d)$, consider $X = \theta + Z$, we have

$$\mathbb{E}[\|X\|^2] = \mathbb{E}[\|\theta\|^2 + 2\theta^T Z + \|Z\|^2] = \mathbb{E}[\|\theta\|^2 + \|Z\|^2] = \|\theta\|^2 + d$$

Therefore, the sample mean may over-estimate the true value of θ. Instead, we project θ on to X as shown in figure below. Then, we can get a better result, that is,

$$\hat{\theta} = \frac{\langle \theta, X \rangle}{\|X\|^2} X \doteq \left(\frac{\|X\|^2 - d}{\|X\|^2} \right) X = \left(1 - \frac{d}{\|X\|^2} \right) X$$

where the second dot equality holds because $\mathbb{E}\langle\theta, X\rangle = \|\theta\|^2 \doteq \|X\|^2 - d$.



With some minor modification, the *James-Stein estimator* is of the form

$$\hat{\theta}_{JS} = \left(1 - \frac{d-2}{\|X\|^2}\right)_+ X, \quad \text{(more generally,}\hat{\theta}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{\|X\|^2}\right)_+ X \text{ for } Z \sim N(0, \sigma^2 I_d))$$

and it is proved in S&DS 610 that it is uniformly better than sample mean estimation.

Adaptivity:
Consider a d-dimensional Gaussian sequence model, and consider a linear procedure. We want to find a best linear procedure with the parameter being a ball

$$\Theta_b = \left\{\sum_{i=1}^{d} \theta_i^2 \leqslant M\right\}$$

We do the same thing as before, finding the best linear procedure c,

$$\inf_{c\in[0,1]} \mathbb{E}\left[\|cX - \theta\|^2\right] = \inf_{c\in[0,1]} \left\{\underbrace{(c-1)^2\|\theta\|^2}_{\text{bias}^2} + \underbrace{c^2 d}_{\text{variance}}\right\} = \frac{\|\theta\|^2 d}{\|\theta\|^2 + d} > \frac{\|\theta\|^2(d-2)}{\|\theta\|^2 + (d-2)}$$

where the last term is just the risk of JS estimator. It does a better job even than best linear procedure! JS is mimicking Best linear procedure when Θ is a ball, and even doing a greater job.

No matter what M is, JS is adaptive to the ball (dependent on data X), but not parameter M.

1.3.2 Blockwise James Stein Estimator

What are we going to do is to separate parameters into blocks. For example, we can separate into blocks of sizes $2, 4, 8, 16, \cdots$.

$$\overbrace{\theta_1, \theta_2}^{2}, \overbrace{\underset{\underset{3^{2\alpha}}{\uparrow}}{\theta_3}, \theta_4, \theta_5, \underset{\underset{6^{2\alpha}}{\uparrow}}{\theta_6}}^{4}, \overbrace{\underset{\underset{7^{2\alpha}}{\uparrow}}{\theta_7}, \cdots, \underset{\underset{14^{2\alpha}}{\uparrow}}{\theta_{14}}}^{8}, \cdots$$

Note that in this case, the ratio between the largest and the smallest semiaxes of the ellipsoid is $2^{2\alpha}$, which is reasonably big. Instead, we want the size of the block has a smaller increasing speed. Thus, we instead divide blocks by power α.

$$\lfloor a^1 \rfloor, \lfloor a^2 \rfloor, \lfloor a^3 \rfloor, \lfloor a^4 \rfloor, \cdots, \quad a = 1 + \frac{1}{\log n}$$

Then, the semiaxes ratio

$$\text{Ratio} \leqslant a^{2\alpha} = \left(1 + \frac{1}{\log n}\right)^{2\alpha} = 1 + o(1)$$

Then, each block is like a ball! Therefore, we can then apply JS to each block. It is constructed as follow:

- Choose integer K_0 so that the initial block length $I_0 = \lfloor a^{K_0} \rfloor \geqslant 3$ and $\lfloor a^k \rfloor - \lfloor a^{k-1} \rfloor \geqslant 3$ for all $k \geqslant K_0 + 1$. In this case, all blocks can be applied with James Stein estimator (since the dimension is larger or equal to 3).

- Choose $K_1 = \lfloor \log_a n \rfloor - 1$ as the stopping point, and denote $I_1 = \lfloor a^{K_1} \rfloor$ the whole length of blocks. Note that $I_1 \geqslant n/a$.

- Let $G_{K_0} = \{i : 1 \leqslant i \leqslant I_0\}$, and let $G_k = \{i : \lfloor a^{k-1} \rfloor < i \leqslant \lfloor a^k \rfloor\}$ for $k \geqslant K_0 + 1$. Let $m_k = \lfloor a^k \rfloor - \lfloor a^{k-1} \rfloor$ denote the number of indices in G_k and $m_{K_0} = I_0$. Denote $\|Y\|^2_{(k)} = \sum_{i \in G_k} y_i^2$, and $\|\theta\|^2_{(k)} = \sum_{i \in G_k} \theta_i^2$.

With these, the corresponding blockwise JS estimator is

$$\hat{\theta}_{BJS,i} = \begin{cases} \left(1 - \dfrac{m_k - 2}{n\|Y\|_k^2}\right)_+ y_i, & \text{if } i \in G_k \text{ for some } k = K_0, \cdots, K_1 \\ 0, & \text{if } i > I_1 \end{cases}$$

(HW3: Prove that Blockwise JS achieves the Pinsker's bound)

We have

$$\sup_{\Theta} \mathbb{E}\left[\|\hat{\theta}_{BJS} - \theta\|^2\right] \leqslant c_p(1 + o(1))n^{-\frac{2\alpha}{2\alpha+1}}$$

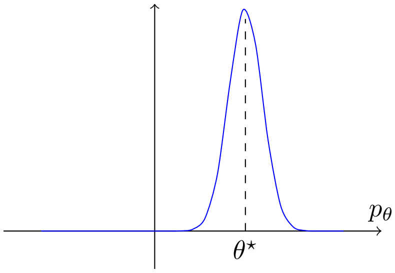
which shows that the blockwise James-Stein estimator attains the best rate of convergence.

Chapter 2

Frequentist Bayesian Investigation

2.1 Posterior Contraction

For Bayesian, we observe $Y \sim P_{Y|\theta^\star}$, and we have a prior $\theta \sim \pi$, and we calculate the posterior $\theta|Y$. We use the notation $\pi(\theta|Y) = \mathbb{E}_{\theta|Y}[\theta]$. The question is: does $\theta \approx \theta^*$? i.e., does θ 'contracts' to θ^*?



To measure this contraction, there are to criteria:

1. $\mathbb{E}_{Y|\theta^\star}\left[\|\pi(\theta|Y) - \theta^\star\|^2\right] \lesssim \varepsilon_n^2$

2. $\mathbb{E}_{Y|\theta^\star}\left[\pi\left(\|\theta - \theta^\star\|^2 \geqslant \varepsilon_n^2 | Y\right)\right] \lesssim e^{-cn\varepsilon_n^2}$

Usually, (2) implies (1), but we still need some mild assumption. Indeed, we want

$$(2) \overset{?}{\Longrightarrow} \mathbb{E}_{Y|\theta^\star}\left[\pi\left(\|\theta - \theta^\star\|^2|Y\right)\right] \lesssim \varepsilon_n^2 \Longrightarrow \mathbb{E}_{Y|\theta^\star}\left[\|\pi(\theta|Y) - \theta^\star\|^2\right] \lesssim \varepsilon_n^2 = (1)$$

where the second implication simply follows Jensen's inequality. Therefore, (2) implies (1) iff the first implication holds. To specify this, note that

$$\mathbb{E}_{Y|\theta^\star}\left[\pi\left(\|\theta - \theta^\star\|^2|Y\right)\right] = \mathbb{E}_{Y|\theta^\star}\mathbb{E}_{\theta|Y}\left[\|\theta - \theta^\star\|^2\right]$$

and the fact that

$$\mathbb{E}\left[X^2\right] = \mathbb{E}\left[X^2\{X^2 \leqslant \varepsilon_n^2\}\right] + \mathbb{E}\left[X^2\{X^2 > \varepsilon_n^2\}\right] \leqslant \varepsilon_n^2 + \sqrt{\mathbb{E}[X^4]} \cdot \underbrace{\sqrt{\mathbb{E}\left[\{X^2 > \varepsilon_n^2\}\right]}}_{\text{exponentially small}}$$

where the inequality follows from Cauchy Schwartz: $\mathbb{E}[XY] \leqslant \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$. Therefore, to make this converge, we only need to require that $\sqrt{\mathbb{E}[X^4]}$ increases by a speed no more than a power of n. Therefore, for (2) implies (1), we need the mild assumption that

$$\mathbb{E}_{Y|\theta^\star}\mathbb{E}_{\theta|Y}\left[\|\theta-\theta^\star\|^4\right] \leqslant n^c \quad \text{for some } c$$

Having this implication, we only need to evaluate (2) for posterior contraction.

2.1.1 Example: Gaussian Sequence Model

Again, consider Gaussian sequence model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}Z_i, \quad Z_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

We consider the prior

$$\text{prior: } \begin{cases} \theta_i \overset{\text{i.i.d.}}{\sim} N(0,1), & i \leqslant k \\ \theta_i = 0, & i > k \end{cases}, k = n^{\frac{1}{1+2\alpha}}$$

and we get the posterior

$$\text{posterior: } \begin{cases} \theta_i|Y \sim N\left(\frac{1}{1+\frac{1}{n}}Y_i, \frac{1\cdot\frac{1}{n}}{1+\frac{1}{n}}\right), & i \leqslant k \\ \theta_i|Y \sim 0, & i > k \end{cases}, k = n^{\frac{1}{1+2\alpha}}$$

Criteria (1):
Then, to verify criteria (1),

$$\mathbb{E}_{Y|\theta^\star}\left[\|\mathbb{E}[\theta|Y]-\theta^\star\|^2\right] = \mathbb{E}_{Y|\theta^\star}\left[\sum_{i=1}^{k}\left(\frac{n}{1+n}Y_i-\theta_i^\star\right)^2+\sum_{i>k}(\theta_i^\star)^2\right]$$

The second term can be bounded by

$$\sum_{i>k}(\theta_i^\star)^2 \leqslant \frac{1}{k^{2\alpha}}\sum_{i>k}i^{2\alpha}(\theta_i^\star)^2 \leqslant \frac{M}{n^{\frac{2\alpha}{1+2\alpha}}} \lesssim n^{-\frac{2\alpha}{2\alpha+1}} = \varepsilon_n^2$$

The second term can be bounded by

$$\begin{aligned}\sum_{i=1}^{k}\mathbb{E}_{Y|\theta^\star}\left[\left(\frac{n}{1+n}\right)^2Y_i^2-\frac{2n\theta_i^\star}{1+n}Y_i+(\theta_i^\star)^2\right] &= \sum_{i=1}^{k}\left[\left(\frac{n}{1+n}\right)^2\left((\theta_i^\star)^2+\frac{1}{n}\right)-\frac{2n}{1+n}(\theta_i^\star)^2+(\theta_i^\star)^2\right] \\ &= \underbrace{\sum_{i=1}^{k}\left(\frac{n}{n+1}\theta_i^\star-\theta_i^\star\right)^2}_{\text{bias}^2}+\underbrace{\sum_{i=1}^{k}\left(\frac{n}{n+1}\right)^2\frac{1}{n}}_{\text{variance}} \\ &\leqslant \frac{1}{(n+1)^2}\sum_{i=1}^{k}(\theta_i^\star)^2+\frac{k}{n} \lesssim n^{-\frac{2\alpha}{2\alpha+1}} = \varepsilon_n^2\end{aligned}$$

This shows that the convergence of (1) is achieved.

Criteria (2):

To verify criteria (2), we write

$$\theta_i = \frac{n}{n+1}Y_i + \sqrt{\frac{1}{n+1}}W_i, \quad W_i \overset{\text{i.i.d.}}{\sim} N(0,1)$$

for $i \leqslant k$. We then denote $A = \{\theta : \|\theta - \theta^\star\|^2 \geqslant \varepsilon_n^2\}$. In our case, this set is

$$\begin{aligned} A &= \left\{ \sum_{i=1}^k \left(\frac{n}{n+1}Y_i + \sqrt{\frac{1}{n+1}}W_i - \theta_i^\star \right)^2 + \sum_{i>k} (\theta_i^\star)^2 \geqslant Cn^{-\frac{2\alpha}{2\alpha+1}} \right\} \\ &= \left\{ \sum_{i=1}^k \left(\frac{n}{n+1}\theta_i^\star + \frac{n}{n+1}\frac{1}{\sqrt{n}}Z_i + \sqrt{\frac{1}{n+1}}W_i - \theta_i^\star \right)^2 + \sum_{i>k} (\theta_i^\star)^2 \geqslant Cn^{-\frac{2\alpha}{2\alpha+1}} \right\} \end{aligned}$$

To verify criterion (2), we need to verify $\mathbb{E}_{Y|\theta^\star}[A|Y] \lesssim e^{-cn\varepsilon_n^2}$. We manipulate the LHS in the set,

$$\begin{aligned} \sum_{i=1}^k \left(\frac{n}{n+1}\theta_i^\star + \frac{n}{n+1}\frac{1}{\sqrt{n}}Z_i + \sqrt{\frac{1}{n+1}}W_i - \theta_i^\star \right)^2 &= \sum_{i=1}^k \left(-\frac{1}{n+1}\theta_i^\star + \sqrt{\left(\frac{n}{n+1}\right)^2 \frac{1}{n} + \frac{1}{n+1}} Z_i' \right)^2 \\ &\leqslant 2\sum_{i=1}^k \left(\frac{1}{n+1}\theta_i^\star \right)^2 + 2\sum_{i=1}^k \frac{2n+1}{(n+1)^2} Z_i'^2 \qquad ((a+b)^2 \leqslant 2a^2 + 2b^2) \\ &\leqslant 2\frac{k}{n^2}M + 4\frac{1}{n+1}\sum_{i=1}^k Z_i'^2 \end{aligned}$$

where $Z_i' \overset{\text{i.i.d.}}{\sim} N(0,1)$ and in the last inequality we make bound $2n+1 \leqslant 2n+2$ to make the form clearer. The first term vanishes with rate $o(n^{-\frac{2\alpha}{2\alpha+1}})$, so we only need to get the tail bound of $Z_i'^2$. We use the fact:

$$\mathbb{P}\left[\chi_k^2 \geqslant k + 2\sqrt{kt} + 2t\right] \leqslant e^{-t}, \quad t > 0$$

With $k = t$, we get

$$\mathbb{P}\left[\chi_k^2 \geqslant 5k\right] \leqslant e^{-k}$$

Therefore,

$$\mathbb{P}\left[4\frac{1}{n+1}\sum_{i=1}^k Z_i'^2\right] = \mathbb{P}\left[\frac{4}{n+1}\chi^2_{n^{\frac{1}{1+2\alpha}}} \geqslant Cn^{-\frac{2\alpha}{2\alpha+1}}\right] \leqslant e^{-cn\varepsilon_n^2}$$

2.1.2 Prior Mass and Testing

Now we consider general prior and general model. We still want to derive the same convergence, i.e.,

$$\mathbb{E}_{Y|\theta^\star}\left[\pi\left(\|\theta - \theta^\star\|^2 \geqslant Ln^{-\frac{2\alpha}{2\alpha+1}} \,\middle|\, Y\right)\right] \leqslant e^{-cn^{-\frac{1}{2\alpha+1}}} \tag{2.1}$$

By Shwartz (1960, 1965), Le Cam (1973) and Barron (1988) together, they showed that three **prior mass and testing** conditions will lead to this conclusion:

(i) $\pi\left(\|\theta - \theta^\star\| \leqslant \varepsilon_n^2\right) \gtrsim e^{-cn\varepsilon_n^2}$. (There must be at least some mass around $\theta^\star$).

(ii) There exists $\mathcal{F}_n$, such that $\pi(\mathcal{F}_n^c) \lesssim e^{-(c+4)n\varepsilon_n^2}$.

(iii) There exists a test ϕ_n, such that for $H_0 : \theta = \theta^\star$, $H_1 : \|\theta - \theta^\star\|^2 \geqslant L\varepsilon_n^2$, we have

$$\underbrace{\mathbb{E}_{Y|\theta^\star}[\phi_n]}_{\text{Type I Error}} + \underbrace{\sup_{\theta \in A \cap \mathcal{F}_n \cap \text{supp}(\pi)} \mathbb{E}_{Y|\theta}[1-\phi_n]}_{\text{Type II Error}} \lesssim e^{-(c+4)n\varepsilon_n^2}, \quad A = \{\theta : \|\theta - \theta^\star\|^2 \geqslant L\varepsilon_n^2\}$$

With conditon (i) - (iii), we can get the conclusion 2.1 for general prior and general model. The condition (ii) provides that the test in (iii) can be found, and the condition (iii) may not hold since the alternative hypothesis is on an infinite-dimensional space.

We first note that (i) implies the following condition:
(i)':

$$\mathbb{P}_{Y|\theta^\star}\left(\int \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi(\theta) \geqslant e^{-(c+2)n\varepsilon_n^2}\right) \geqslant 1 - e^{-c' n\varepsilon_n^2}$$

Proof. Let

$$H = \left\{\int \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi(\theta) \geqslant e^{-(c+2)n\varepsilon_n^2}\right\}, \quad B = \left\{\theta : \|\theta - \theta^\star\|^2 \leqslant \varepsilon_n^2\right\}$$

Consider the event

$$\begin{aligned}
H &\supseteq \left\{\int_B \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi(\theta) \geqslant e^{-(c+2)n\varepsilon_n^2}\right\} = \left\{\int_B \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \frac{\mathrm{d}\pi(\theta)}{\pi(B)} \geqslant \frac{e^{-(c+2)n\varepsilon_n^2}}{\pi(B)}\right\} \\
&\supseteq \left\{\int_B \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \frac{\mathrm{d}\pi(\theta)}{\pi(B)} \geqslant e^{-2n\varepsilon_n^2}\right\} = \left\{\int_B \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi_B(\theta) \geqslant e^{-2n\varepsilon_n^2}\right\} && \text{(By condition (i))} \\
&= \left\{-2\log \int_B \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi_B(\theta) \leqslant 4n\varepsilon_n^2\right\} \supseteq \left\{\int_B -2\log \frac{\mathbb{P}_{Y|\theta}}{\mathbb{P}_{Y|\theta^\star}} \,\mathrm{d}\pi_B(\theta) \leqslant 4n\varepsilon_n^2\right\} && \text{(Jensen's inequality)} \\
&= \left\{\int n\left(\|Y-\theta\|^2 - \|Y-\theta^\star\|^2\right) \mathrm{d}\pi_B(\theta) \leqslant 4n\varepsilon_n^2\right\} && (Y \text{ is Guassian, substitute in Guassian density})
\end{aligned}$$

Since we are considering $\mathbb{P}_{Y|\theta^\star}[H]$ in condition $(i)'$, we substitute $Y = \theta^\star + \frac{1}{\sqrt{n}}Z$, hence

$$\begin{aligned}
H &\supseteq \left\{\int_B n\left(\|Y-\theta\|^2 - \|Y-\theta^\star\|^2\right) \mathrm{d}\pi_B(\theta) \leqslant 4n\varepsilon_n^2\right\} = \left\{\int_B 2n\left\langle \theta^\star - \theta, \frac{1}{\sqrt{n}}Z\right\rangle + n\|\theta^\star - \theta\|^2 \,\mathrm{d}\pi_B(\theta) \leqslant 4n\varepsilon_n^2\right\} \\
&\supseteq \left\{\underbrace{\int_B 2n\left\langle \theta^\star - \theta, \frac{1}{\sqrt{n}}Z\right\rangle \mathrm{d}\pi_B(\theta)}_{N(0, 4n\|\int \theta \,\mathrm{d}\pi_B(\theta) - \theta^\star\|^2)} \leqslant 3n\varepsilon_n^2\right\} && (\|\theta^\star - \theta\|^2 \leqslant \varepsilon_n^2 \text{ on B}) \\
&= \left\{W \leqslant \frac{3n\varepsilon_n^2}{\sqrt{4n\|\int \theta \,\mathrm{d}\pi_B(\theta) - \theta^\star\|^2}}\right\}, \quad W \sim N(0,1) \\
&\supseteq \left\{W \leqslant \sqrt{\frac{9}{4}n\varepsilon_n^2}\right\} && (\|\textstyle\int \theta \,\mathrm{d}\pi_B(\theta) - \theta^\star\|^2 \leqslant \varepsilon_n^2)
\end{aligned}$$

This can be tackled by a gauss tail bound $\mathbb{P}[W > t] \leqslant e^{-t^2/2}$, where we have

$$\mathbb{P}_{Y|\theta^\star}[H] \geqslant \mathbb{P}_{Y|\theta^\star}\left\{W \leqslant \sqrt{\frac{9}{4}n\varepsilon_n^2}\right\} \geqslant 1 - e^{-\frac{9}{8}n\varepsilon_n^2}$$

which completes the proof. □

Now we are going to use (i)’, (ii) and (iii) to derive the posterior contraction rate 2.1.

Proof.

$$\begin{aligned}
\mathbb{E}_{Y|\theta^\star}\left[\pi(A|Y)\right] &= \mathbb{E}_{Y|\theta^\star}\left[\pi(A|Y)\left(\phi_n + 1 - \phi_n\right)\right] \\
&\leqslant \mathbb{E}_{Y|\theta^\star}\left[\phi_n\right] + \mathbb{E}_{Y|\theta^\star}\left[\pi(A|Y)(1-\phi_n)\right] && (\mathbb{E}_{Y|\theta^\star}[\pi(A|Y)] \in [0,1]) \\
&\lesssim e^{-c'n\varepsilon_n^2} + \mathbb{E}_{Y|\theta^\star}\left[\frac{\int_A P_{Y|\theta}\,\mathrm{d}\pi(\theta)}{\int P_{Y|\theta}\,\mathrm{d}\pi(\theta)}(1-\phi_n)\right] && \text{(Condition (iii) Type I error bound)} \\
&= e^{-c'n\varepsilon_n^2} + \mathbb{E}_{Y|\theta^\star}\left[\frac{\int_A \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}{\int \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}(1-\phi_n)\right] \\
&= e^{-c'n\varepsilon_n^2} + \mathbb{E}_{Y|\theta^\star}\left[\frac{\int_A \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}{\int \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}(1-\phi_n)\right](\mathbb{1}_H + \mathbb{1}_{H^c}) \\
&\lesssim e^{-c'n\varepsilon_n^2} + \mathbb{E}_{Y|\theta^\star}\left[\frac{\int_A \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}{\int \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)}(1-\phi_n)\right]\mathbb{1}_H + e^{-c'n\varepsilon_n^2} && \text{(By (i)’, on } H^c \text{ bounded above)} \\
&\lesssim e^{-c'n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_A \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] && \text{(By (i)’, on } H \text{ denominator lower bound)} \\
&= e^{-c'n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n^c} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] \\
&= e^{-c'n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] + e^{(c+2)n\varepsilon_n^2}\int P_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n^c} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right]\mathrm{d}y \\
&= e^{-c'n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] + e^{(c+2)n\varepsilon_n^2}\int_{A\cap\mathcal{F}_n^c}\int P_{Y|\theta}\,\mathrm{d}y\,\mathrm{d}\pi(\theta)(1-\phi_n) \\
&\lesssim e^{-c'n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\mathbb{E}_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right] + e^{(c+2)n\varepsilon_n^2}\pi(\mathcal{F}_n^c) \\
&= e^{-c'n\varepsilon_n^2} + e^{-2n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\int P_{Y|\theta^\star}\left[\int_{A\cap\mathcal{F}_n} \frac{P_{Y|\theta}}{P_{Y|\theta^\star}}\,\mathrm{d}\pi(\theta)(1-\phi_n)\right]\mathrm{d}y && \text{(Condition (ii))} \\
&= e^{-c'n\varepsilon_n^2} + e^{-2n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\int_{A\cap\mathcal{F}_n}\int P_{Y|\theta}(1-\phi_n)\,\mathrm{d}y\,\mathrm{d}\pi(\theta) \\
&\leqslant e^{-c'n\varepsilon_n^2} + e^{-2n\varepsilon_n^2} + e^{(c+2)n\varepsilon_n^2}\sup_{\theta\in A\cap\mathcal{F}_n\cap\mathrm{supp}(\pi)}\int P_{Y|\theta}(1-\phi_n)\,\mathrm{d}y \\
&\lesssim e^{-c'n\varepsilon_n^2} + e^{-2n\varepsilon_n^2} + e^{-2n\varepsilon_n^2} = e^{-c'n\varepsilon_n^2}
\end{aligned}$$

which completes the proof. □

(HW4: Check the three conditions for Gaussian sequence model)

Chapter 3

High Dimensional Estimation

3.1 Lasso Estimator

Lasso estimator is the form

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \left(\|Y-\theta\|_2^2 + \lambda\|\theta\|_1\right)$$

3.1.1 Gaussian Sequence Model

Consider the model

$$Y_i = \theta_i + \frac{1}{\sqrt{n}}Z_i, \quad Z_i \overset{\text{i.i.d}}{\sim} N(0,1), \quad i = 1,2,\cdots,p$$

with parameter space

$$\Theta = \{\theta : \|\theta\|_0 \leqslant s\}, \text{ where } \|\theta\|_0 = \#\{i : \theta_i \neq 0\}$$

1. Upper Bound

We use the **basic inequality** here: For $Y = \theta^\star + \frac{1}{\sqrt{n}}Z$, we have

$$\|Y-\hat{\theta}\|_2^2 + \lambda\|\hat{\theta}\|_1 \leqslant \|Y-\theta^\star\|_2^2 + \lambda\|\theta^\star\|_1$$

Actually, it holds for all θ substuting $\theta^\star$ on the RHS. Substitute $Y = \theta^\star + \frac{1}{\sqrt{n}}Z$ in, we have

$$\begin{aligned}
& \left\|\theta^\star - \hat{\theta} + \frac{1}{\sqrt{n}}Z\right\|_2^2 + \lambda\left\|\hat{\theta}\right\|_1 \leqslant \left\|\frac{1}{\sqrt{n}}Z\right\|_2^2 + \lambda\left\|\theta^\star\right\|_1 \\
\Longrightarrow \quad & \left\|\theta^\star - \hat{\theta}\right\|_2^2 + \left\|\frac{1}{\sqrt{n}}Z\right\|_2^2 + 2\left\langle \theta^\star - \hat{\theta}, \frac{1}{\sqrt{n}}Z\right\rangle + \lambda\left\|\hat{\theta}\right\|_1 \leqslant \left\|\frac{1}{\sqrt{n}}Z\right\|_2^2 + \lambda\left\|\theta^\star\right\|_1 \\
\Longrightarrow \quad & \left\|\theta^\star - \hat{\theta}\right\|_2^2 \leqslant 2\left\langle \hat{\theta} - \theta^\star, \frac{1}{\sqrt{n}}Z\right\rangle + \lambda\left\|\theta^\star\right\|_1 - \lambda\left\|\hat{\theta}\right\|_1
\end{aligned}$$

Use the notation

$$S^\star = \{i : \theta_i^\star \neq 0\}, \quad |S^\star| = s, \quad |(S^\star)^c| = p - s, \quad \theta_{A,i} = \begin{cases}\theta_i, & i \in A \\ 0, & i \notin A\end{cases}$$

We can rewrite the inequality in the form:

$$\left\|\theta^\star - \hat{\theta}\right\|_2^2 \leqslant 2\left\langle \hat{\theta} - \theta^\star, \frac{1}{\sqrt{n}}Z \right\rangle + \lambda \left\|\theta^\star_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{(S^\star)^c}\right\|_1$$

We use two facts here:

$$\textbf{Fact I: } |\langle a, b\rangle| \leqslant \|a\|_1 \|b\|_\infty$$

$$\textbf{Fact II: } W \sim N(0,1) \Longrightarrow \mathbb{P}(W \geqslant t) \leqslant \frac{1}{\sqrt{2\pi}t}e^{-t^2/2}$$

If we let $t = \sqrt{2\log p}$ in Fact II, we have

$$\mathbb{P}(W \geqslant \sqrt{2\log p}) \leqslant \frac{1}{\sqrt{2\pi}\sqrt{2\log p}}e^{-\log p} = o(1/p)$$

Summing over all coordinate in our case, we will have $o(1)$ convergence. Therefore, we have

$$\left\|\frac{1}{\sqrt{n}}Z\right\|_\infty \leqslant \frac{1}{\sqrt{n}}\sqrt{2\log p} \quad \text{with high probability } (1 - o(1))$$

For simplicity, we will write *w.h.p.* for 'with high probability'. Substitute this in the inequality, we have

$$\begin{aligned}
\left\|\theta^\star - \hat{\theta}\right\|_2^2 &\leqslant 2\left\|\hat{\theta} - \theta^\star\right\|_1 \left\|\frac{1}{\sqrt{n}}Z\right\|_\infty + \lambda \left\|\theta^\star_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{(S^\star)^c}\right\|_1 \\
&\leqslant 2\sqrt{\frac{2\log p}{n}}\left(\left\|(\hat{\theta} - \theta^\star)_{S^\star}\right\|_1 + \left\|(\hat{\theta} - \theta^\star)_{(S^\star)^c}\right\|_1\right) + \lambda \left\|\theta^\star_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{(S^\star)^c}\right\|_1 \quad \text{w.h.p.} \\
&\leqslant 2\sqrt{\frac{2\log p}{n}}\left(\left\|(\hat{\theta} - \theta^\star)_{S^\star}\right\|_1 + \left\|\hat{\theta}_{(S^\star)^c}\right\|_1\right) + \lambda \left\|(\theta^\star - \hat{\theta})_{S^\star}\right\|_1 - \lambda \left\|\hat{\theta}_{(S^\star)^c}\right\|_1 \quad \text{w.h.p.}
\end{aligned}$$

where the last step uses the fact that $\theta^\star = 0$ on $(S^\star)^c$ (on the first term) and triangular inequality (on the second term). Set $\lambda = 2\sqrt{\frac{2\log p}{n}}$, we have

$$\left\|\theta^\star - \hat{\theta}\right\|_2^2 \leqslant 4\sqrt{\frac{2\log p}{n}}\left\|(\hat{\theta} - \theta^\star)_{S^\star}\right\|_1 \text{ w.h.p.} \leqslant 4\sqrt{\frac{2\log p}{n}}\sqrt{s}\left\|\hat{\theta} - \theta^\star\right\|_2 \text{ w.h.p.}$$

where we use the fact $\|a\|_1 \leqslant \sqrt{d}\|a\|_2$ for $a \in \mathbb{R}^d$. Thus,

$$\left\|\theta^\star - \hat{\theta}\right\|_2^2 \leqslant \frac{32 s\log p}{n} \text{ w.h.p.}$$

This is not a sharp bound, but with the correct rate.

2. Lower Bound

(HW5:)

The goal is: For $s = p^\gamma$, where $0 < \gamma < 1$, we achieve minimax lower bound:

$$\sup_{\theta\in\Theta} \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] \gtrsim \frac{s\log p}{n}$$

3.1.2 High-dimensional Linear Regression

Consider the high-dimensional linear model

$$Y_{n\times p} = X_{n\times p}\beta_{p\times 1} + Z, \quad Z \sim N(0, I_n), \quad p \gg n, \quad \|\beta\|_0 \leqslant s$$

The Lasso estimator is

$$\hat{\theta} = \underset{\beta:\|\beta\|_0\leqslant s}{\operatorname{argmin}} \left(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\right)$$

Apply basic inequality:

$$\begin{aligned}
& \|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leqslant \|Y - X\beta^\star\|_2^2 + \lambda\|\beta^\star\|_1 \\
\Longrightarrow\quad & \|X\beta^\star + Z - X\hat{\beta}\|_2^2 \leqslant \|Z\|_2^2 + \lambda\|\beta^\star\|_1 - \lambda\|\hat{\beta}\|_1 && (Y = X\beta^\star + Z) \\
\Longrightarrow\quad & \|X\beta^\star - X\hat{\beta}\|_2^2 \leqslant 2\left\langle X(\hat{\beta} - \beta^\star), Z\right\rangle + \lambda\|\beta^\star\|_1 - \lambda\|\hat{\beta}\|_1 \\
\Longrightarrow\quad & \|X\beta^\star - X\hat{\beta}\|_2^2 \leqslant 2\left\langle \hat{\beta} - \beta^\star, X^TZ\right\rangle + \lambda\|\beta^\star\|_1 - \lambda\|\hat{\beta}\|_1 \\
& \qquad\qquad\qquad \leqslant 2\|\hat{\beta} - \beta^\star\|_1\|X^TZ\|_\infty + \lambda\|\beta^\star\|_1 - \lambda\|\hat{\beta}\|_1
\end{aligned}$$

Note that $X^TZ \sim N(0, X^TX)$, we assume that $\|X_{\cdot j}\|^2 = n$, i.e., X is normalized, then $\|X^TZ\|_\infty \leqslant \sqrt{2n\log p}$ with high probability. Therefore,

$$\|X\beta^\star - X\hat{\beta}\|_2^2 \leqslant 2\sqrt{2n\log p}\|\hat{\beta} - \beta^\star\|_1 + \lambda\|\beta^\star\|_1 - \lambda\|\hat{\beta}\|_1 \text{ w.h.p.}$$

Our first attempt i, to choose $\lambda = 2\sqrt{2n\log p}$, just like what we did in Gaussian sequence model, to eliminate the terms containing $(S^*)^c$, that is

$$\begin{aligned}
\|X\beta^\star - X\hat{\beta}\|_2^2 &\leqslant 2\sqrt{2n\log p}\left(\|(\hat{\beta} - \beta^\star)_{S^\star}\|_1 + \|(\hat{\beta} - \beta^\star)_{(S^\star)^c}\|_1\right) + \lambda\|(\beta^\star)_{S^\star}\|_1 - \lambda\|\hat{\beta}_{S^\star}\|_1 - \lambda\|\hat{\beta}_{(S^\star)^c}\|_1 \text{ w.h.p.} \\
&\leqslant 2\sqrt{2n\log p}\left(\|(\hat{\beta} - \beta^\star)_{S^\star}\|_1 + \|\hat{\beta}_{(S^\star)^c}\|_1\right) + \lambda\|(\beta^\star - \hat{\beta})_{S^\star}\|_1 - \lambda\|\hat{\beta}_{(S^\star)^c}\|_1 \text{ w.h.p.} \\
&= 4\sqrt{2n\log p}\|(\hat{\beta} - \beta^\star)_{S^\star}\|_1 \text{ w.h.p.} && (\lambda = 2\sqrt{2n\log p}) \\
&\leqslant 4\sqrt{s}\sqrt{2n\log p}\|(\hat{\beta} - \beta^\star)_{S^\star}\|_2
\end{aligned}$$

However, the challenge is that, $\|X\beta^\star - X\hat{\beta}\|_2^2 = (\beta^\star - \hat{\beta})^TX^TX(\beta^\star - \hat{\beta})$, the smallest eigenvalue of X^TX may be zero, and we may never have remove X from LHS to get a bound for $\|\beta^\star - \hat{\beta}\|_2^2$. Therefore, we instead choose

$$\lambda = 4\sqrt{2n\log p}$$

This will lead to

$$\begin{aligned}
& 0 \leqslant \|X\hat{\beta} - X\beta^\star\|_2^2 \leqslant 6\sqrt{2n\log p}\|(\hat{\beta} - \beta^\star)_{S^\star}\|_1 - 2\sqrt{2n\log p}\|(\hat{\beta} - \beta^\star)_{(S^\star)^c}\|_1 \\
\Longrightarrow\quad & 3\|(\hat{\beta} - \beta^\star)_{S^\star}\|_1 \geqslant \|(\hat{\beta} - \beta^\star)_{(S^\star)^c}\|_1
\end{aligned}$$

This leads to the **restricted eigenvalue condition**:

$$\|X\theta\|^2 \geqslant \gamma n \|\theta\|^2 \quad \text{for } 3\|\theta_S\|_1 \geqslant \|\theta_{S^c}\|_1 \text{ and } |S| = s$$

Under this condition, we finally have

$$\begin{aligned} & \gamma n \|\beta^\star - \hat{\beta}\|_2^2 \leqslant 6\sqrt{s}\sqrt{2n\log p}\|(\beta^\star - \hat{\beta})_{S^\star}\|_2 \\ \Longrightarrow \quad & \|\beta^\star - \hat{\beta}\|_2^2 \lesssim \frac{s\log p}{n} \end{aligned}$$

3.2 Structured Vector/Matrix Estimation

3.2.1 Sparse Vector Estimation with All Subset Selection

We continue considering the high-dimensional linear regression

$$Y_{n\times p} = X_{n\times p}\beta_{p\times 1} + Z, \quad Z \sim N(0, I_n), \quad p \gg n, \quad \|\beta\|_0 \leqslant s$$

This time, we consider the MLE, or all subset selection estimator

$$\hat{\beta} = \underset{\beta:\|\beta\|_0 = s}{\operatorname{argmin}} \|Y - X\beta\|^2$$

It is called all subset selection since totally $\binom{p}{s}$ ways of picking a support is valid. Use basic inequality

$$\begin{aligned} & \|Y - X\hat{\beta}\|^2 \leqslant \|Y - X\beta^\star\|^2 \\ \Longrightarrow \quad & \|X\beta^\star + Z - X\hat{\beta}\|^2 \leqslant \|Z\|^2 && (Y = X\beta^\star + Z) \\ \Longrightarrow \quad & \|X\beta^\star - X\hat{\beta}\|^2 + 2\left\langle X\beta^\star - X\hat{\beta}, Z \right\rangle \leqslant 0 \\ \Longrightarrow \quad & \|X\beta^\star - X\hat{\beta}\|^2 \leqslant 2\left\langle Z, X\hat{\beta} - X\beta^\star \right\rangle \\ \Longrightarrow \quad & \|X\beta^\star - X\hat{\beta}\| \leqslant 2\left\langle Z, \frac{X\hat{\beta} - X\beta^\star}{\|X\hat{\beta} - X\beta^\star\|} \right\rangle \\ & \qquad \leqslant 2 \sup_{\beta:\|\beta\|_0 = s} \left\langle Z, \frac{X\beta - X\beta^\star}{\|X\beta - X\beta^\star\|} \right\rangle \\ & \qquad = 2 \sup_{S:|S|=s} \sup_{\beta_S:\text{supp}(\beta_S) = S} \left\langle Z, \frac{X\beta_S - X\beta^\star}{\|X\beta_S - X\beta^\star\|} \right\rangle, \quad S \subseteq \{1, 2, \cdots, p\} \end{aligned}$$

To deal with this, we use the so-called **Hanson-Wright inequality**:

For sub-Gaussian random vector $\xi \in \mathbb{R}^d$, i.e., for $\|v\|_2 = 1$, $v \in \mathbb{R}^d$ we have

$$\mathbb{E}[\xi] = 0, \quad \mathbb{E}\left[e^{tv^T\xi}\right] \leqslant e^{\frac{1}{2}t^2\sigma^2}$$

Let W be a S-dimensional subspace in $\mathbb{R}^d$, then

$$\mathbb{P}\left(\|P_W\xi\|^2 \geqslant (s + 2\sqrt{st} + 2t)\sigma^2\right) \leqslant e^{-t}, \quad t > 0$$

In our case, let $W_S = \{\beta_S : \text{supp}(\beta_S) = S\}$, and write $Z = P_{W_S}Z + P_{W_S^\perp}S$, then we have

$$\sup_{\|w\|_2=1} \langle P_{W_S}Z, w\rangle = \left\langle P_{W_S}Z, \frac{P_{W_S}Z}{\|P_{W_S}Z\|_2}\right\rangle = \|P_{W_S}Z\|_2$$

This has implication of our equation such that

$$\mathbb{P}\left(\sup_{\beta:\|\beta\|_0=s}\left\langle Z, \frac{X\beta - X\beta^\star}{\|X\beta - X\beta^\star\|}\right\rangle^2 \geqslant s + 2\sqrt{st} + 2t\right) \leqslant \binom{p}{s}e^{-t}$$

where $\binom{p}{s}$ appears since we have this number of choices of W_S. Then, we can choose, for example, $t = \log\binom{p}{s}^{1+\epsilon}$, $\epsilon > 0$, so that $\binom{p}{s}e^{-t} = \binom{p}{s}^{-\epsilon} \to 0$. Thus,

$$\sup_{\beta:\|\beta\|_0=s}\left\langle Z, \frac{X\beta - X\beta^\star}{\|X\beta - X\beta^\star\|}\right\rangle \leqslant \gamma \text{ w.h.p.,} \qquad \text{where } \gamma^2 = s + 2\sqrt{s(1+\epsilon)\log\binom{p}{s}} + 2(1+\epsilon)\log\binom{p}{s}$$

Note that $\log\binom{p}{s} \leqslant s\log\frac{ep}{s}$, and the dominating term in γ^2 is the last term, so we finally get

$$\|X\beta^\star - X\hat{\beta}\|^2 \leqslant 4\gamma^2 \lesssim s\log\frac{ep}{s}$$

Remark: More general form of Hanson-Wright inequality:

$$\mathbb{P}\left(\|A\xi\|^2 \geqslant \left(\text{tr}(\Sigma) + 2\sqrt{\text{tr}(\Sigma^2)t} + 2\lambda_{\max}(\Sigma)t\right)\sigma^2\right) \leqslant e^{-t}, \quad \Sigma = AA^T$$

3.2.2 Stochastic Block Model and Graphon Estimation

Suppose we have a graph with nodes $1, 2, \cdots, n$, and edges A_{ij}, with $A_{ij} = 1$ if i, j is connected, and $A_{ij} = 0$ otherwise. Assume $A_{ij} \sim \text{Ber}(p_{ij})$ independent. Moreover, $A_{ii} = 0$ for all i. It is impossible to directly estimate p_{ij}, so we assume certain structure of p_{ij}.

Consider the two blocks case $g : \{1, 2, \cdots, n\} \to \{1, 2\}$, and $p_{ij} = B_{g(i)g(j)}$, where B is a constant only depends on group information. Then, B is a 2×2 matrix where we assume $B_{12} = B_{21}$. More generally, we may have k blocks

$$g : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, k\}, \text{ where } P_{ij} = B_{g(i)g(j)}, \quad B \text{ is } k \times k \text{ symmetric matrix}$$

Note that the cardinality of g is k^n in this case. The goal is to have the *graphon estimation* such that $\|\hat{p} - p\|_{\mathcal{F}}^2 \leqslant \gamma$ w.h.p.. We consider the estimator

$$\hat{p} = \underset{p_{g,B}}{\text{argmin}}\|A - p\|_{\mathcal{F}}^2$$

Use basic inequality:

$$\begin{aligned}
& \|A-\hat{p}\|_{\mathcal{F}}^2 \leqslant \|A-p^\star\|_{\mathcal{F}}^2 \\
\Longrightarrow\quad & \|p^\star+Z-\hat{p}\|_{\mathcal{F}}^2 \leqslant \|Z\|_{\mathcal{F}}^2 && (A=p^\star+Z,\, \mathbb{E}[Z]=0) \\
\Longrightarrow\quad & \|p^\star-\hat{p}\|_{\mathcal{F}}^2 \leqslant 2\langle \hat{p}-p^\star, Z\rangle = 2\|p^\star-\hat{p}\|_{\mathcal{F}} \left\langle Z, \frac{\hat{p}-p^\star}{\|p^\star-\hat{p}\|_{\mathcal{F}}} \right\rangle \\
\Longrightarrow\quad & \|p^\star-\hat{p}\|_{\mathcal{F}} \leqslant 2\left\langle Z, \frac{\hat{p}-p^\star}{\|p^\star-\hat{p}\|_{\mathcal{F}}} \right\rangle \leqslant 2\sup_{p_{g,B}} \left\langle Z, \frac{p_{g,B}-p^\star}{\|p_{g,B}-p^\star\|_{\mathcal{F}}} \right\rangle \\
& \qquad\qquad = 2\sup_{g}\sup_{p_B^g} \left\langle Z, \frac{p_B^g-p^\star}{\|p_B^g-p^\star\|_{\mathcal{F}}} \right\rangle
\end{aligned}$$

where the supscripts g, B denotes that p depends on these quantities, and p_B^g denotes that g is fixed. Use Hanson-Wright inequality,

$$\mathbb{P}\left(\sup_{g}\sup_{p_B^g} \left\langle Z, \frac{p_B^g-p^\star}{\|p_B^g-p^\star\|_{\mathcal{F}}} \right\rangle^2 \geqslant (s+2\sqrt{st}+2t)\sigma^2 \right) \leqslant k^n e^{-t}$$

Note that $s = k + \binom{k}{2}$ (dimension of B), and we can choose t large enough such that $e^t = (k^n)^{1+\epsilon}$, $\epsilon > 0$ so that $k^n e^{-t} = (k^n)^{-\epsilon} \to 0$ $(t = (1+\epsilon)n\log k)$. Thereofore,

$$\|p^\star-\hat{p}\|_{\mathcal{F}}^2 \leqslant s + 2\sqrt{st} + 2t \text{ w.h.p.}$$

Since $s \asymp k^2$, $t \asymp n\log k$, we have

$$\|p^\star-\hat{p}\|_{\mathcal{F}}^2 \lesssim k^2 + n\log k$$

There are two phase transitions in this bound: $k = 1$, and $k \asymp \sqrt{n\log n}$. when $k \ll \sqrt{n\log n}$, $n\log k$ is the leading term, and when $k \gg \sqrt{n\log n}$, k^2 is the leading term.

(HW6:)

The goal is to get the bound of bi-clustering model. In this model, we assume that the observed matrix

$$A \in \{0,1\}^{n\times m} \quad \text{saitisfies} \quad A_{ij} \sim \text{Ber}(p_{ij}), \quad \text{with } p_{ij} = B_{g(i),h(j)}$$

where

- $g : \{1,\cdots,n\} \to \{1,\cdots,k_1\}$ assigns each row to one of k_1 clusters.
- $h : \{1,\cdots,m\} \to \{1,\cdots,k_2\}$ assigns each column to one of k_2 clusters.
- $B \in \mathbb{R}^{k_1\times k_2}$ is the matrix of parameters.

Consider the estimator

$$\hat{p} = \underset{p_{g,h,B}}{\text{argmin}} \|A-p\|_{\mathcal{F}}^2$$

and our goal is to bound the error $\|p^\star-\hat{p}\|_{\mathcal{F}}^2$ with high probability.

3.3 Some Negative Results for Bayesian Estimation

3.3.1 Natural Prior may not Work (LASSO prior Example)

Consider the model

$$Y_i = \theta_i + \sigma Z_i, \quad i = 1, 2, \cdots, p, \quad p \to \infty, \|\theta\|_0 \leqslant s \ll p$$

and the LASSO estimator

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \frac{\|Y - \theta\|^2}{2\sigma^2} + \frac{\lambda}{\sigma}\|\theta\|_1 \right\}, \quad \lambda = \sqrt{2\log p}$$

It is equivalent to the maximum a posteriori (MAP) estimator

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left\{ \exp\left(-\frac{\|Y - \theta\|^2}{2\sigma^2} - \frac{\lambda}{\sigma}\|\theta\|_1 \right) \right\}, \quad \lambda = \sqrt{2\log p}$$

Recall that $\mathbb{P}\left\{ \|\hat{\theta} - \theta^\star\|^2 \leqslant c\frac{s\log p}{n} \right\} \to 1$, and when $s = 0$, $\mathbb{P}\{\hat{\theta} = \theta^\star\} \to 1$, from our previous analysis.

Question:

- Does the posterior contract? i.e.,

$$\mathbb{E}_{Y|\theta^\star}\left[\pi\left\{ \|\theta - \theta^\star\|^2 \leqslant c\frac{s\log p}{n} \middle| Y \right\} \right] \to 1$$

- When $s = 0$, do we have

$$\mathbb{E}_{Y|\theta^\star}\left[\pi\left\{ \theta = \theta^\star | Y \right\} \right] \to 1$$

The claim of this section is that:

$$\exists\, c > 0, \ s.t.\ \mathbb{E}_{Y|\theta^\star = 0}\left[\pi\left\{ \|\theta - \theta^\star\|^2 \leqslant c\sigma^2 \frac{p}{\log p} \middle| Y \right\} \right] \to 0$$

This corresponds to the case $s = \frac{p}{\log^2 p}$. For the simplicity of analysis, we assume $\sigma = 1$, where with some minor effort it can be extended to general cases. Denote

$$A = \left\{ \theta : \|\theta - \theta^\star\|^2 \leqslant c\frac{p}{\log p} \right\}$$

Then,

$$\begin{aligned}
\mathbb{E}_{Y|\theta^\star = 0}\left[\pi\left\{ \|\theta - \theta^\star\|^2 \leqslant c\sigma^2 \frac{p}{\log p} \middle| Y \right\} \right] &= \mathbb{E}_{Y|\theta^\star = 0}\left[\frac{\int_A \exp\left(-\frac{\|Y-\theta\|^2}{2} - \lambda\|\theta\|_1 \right) \mathrm{d}\theta}{\int \exp\left(-\frac{\|Y-\theta\|^2}{2} - \lambda\|\theta\|_1 \right) \mathrm{d}\theta} \right] \\
&= \mathbb{E}_{Y|\theta^\star = 0}\left[\frac{\int_A \exp\left(\langle Y, \theta \rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1 \right) \mathrm{d}\theta}{\int \exp\left(\langle Y, \theta \rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1 \right) \mathrm{d}\theta} \right]
\end{aligned}$$

Define

$$T = \int \exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1 \right) \mathrm{d}\theta$$

Then,

$$\mathbb{E}_{Y|\theta^\star=0}\left[\pi\left\{\|\theta-\theta^\star\|^2 \leqslant c\sigma^2\frac{p}{\log p}\middle| Y\right\}\right] = \mathbb{E}_{Y|\theta^\star=0}\left[\frac{\int_A \exp\left(\langle Y,\theta\rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta}{T\int \exp\left(\langle Y,\theta\rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)/T\,\mathrm{d}\theta}\right]$$

where we multiply and divide by T on the denominator. Then the denominator can be operated by

$$\int \exp\left(\langle Y,\theta\rangle\right)\underbrace{\exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)/T}_{\text{a density of r.v. } W}\,\mathrm{d}\theta = \mathbb{E}_W\left[\exp\langle Y,\theta\rangle\right] \geqslant \exp\left(\mathbb{E}_W\langle Y,\theta\rangle\right) = 1$$

where the inequality follows from Jensen's inequality, and the final equality holds since Y is symmetry around 0. Denote

$$G = \mathbb{E}_{Y|\theta^\star=0}\left[\pi\left\{\|\theta-\theta^\star\|^2 \leqslant c\sigma^2\frac{p}{\log p}\middle| Y\right\}\right]$$

We have

$$G \leqslant \mathbb{E}_{Y|\theta^\star=0}\left[\frac{\int_A \exp\left(\langle Y,\theta\rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta}{T}\right], \quad Y \sim N(0, I_p)$$

Since the expectation is w.r.t. Y, and denominator does not depend on Y, we can put the expectation into integration, and get

$$\begin{aligned}
G &\leqslant \frac{\int_A \mathbb{E}_{Y|\theta^\star=0}\left[\exp\left(\langle Y,\theta\rangle - \frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\right]\,\mathrm{d}\theta}{\int \exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta} \\
&= \frac{\int_A \exp(-\lambda\|\theta\|_1)\,\mathrm{d}\theta}{\int \exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta} \qquad \left(\mathbb{E}_{Y|\theta^\star=0}\exp\left\{\langle Y,\theta\rangle - \frac{\|\theta\|^2}{2}\right\} = 1, \text{ integration of density}\right) \\
&= \frac{\int_{\{\|\theta\|^2 \leqslant c\frac{p}{\log p}\}} \exp(-\lambda\|\theta\|_1)\,\mathrm{d}\theta}{\int \exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta} \\
&\leqslant \frac{\int_{\{\|\theta\|^2 \leqslant c\frac{p}{\log p}\}} \exp(-\lambda\|\theta\|_1)\,\mathrm{d}\theta}{\int_{\{\|\theta\|^2 \leqslant \frac{p}{\log p}\}} \exp\left(-\frac{\|\theta\|^2}{2} - \lambda\|\theta\|_1\right)\,\mathrm{d}\theta}
\end{aligned}$$

We deal with terms separately. First, the integrand on numerator $\exp(-\lambda\|\theta\|_1) \leqslant 1$. Second, for the integrand on the denominator, we have

$$\frac{\|\theta\|^2}{2} \leqslant \frac{p}{2\log p} \quad\Longrightarrow\quad \exp\left(-\frac{\|\theta\|^2}{2}\right) \geqslant \exp\left(-\frac{p}{2\log p}\right)$$

$$\lambda\|\theta\|_1 = \sqrt{2\log p}\|\theta\|_1 \leqslant \sqrt{2\log p}\sqrt{p}\|\theta\|_2 \leqslant \sqrt{2\log p}\sqrt{p}\sqrt{\frac{p}{\log p}} = \sqrt{2}p \quad\Longrightarrow\quad \exp(-\lambda\|\theta\|_1) \geqslant e^{-\sqrt{2}p}$$

Finally, note that the region $\{\|\theta\|^2 \leqslant c\frac{p}{\log p}\}$ on numerator and region $\{\|\theta\|^2 \leqslant \frac{p}{\log p}\}$ has volume ratio $c^{p/2}$ in p-dimension. Therfore, we finally have

$$G \leqslant c^{p/2} \times e^{\frac{p}{2\log p}} \times e^{\sqrt{2}p} = \exp\left(\frac{p}{2}\log c + \frac{p}{2\log p} + \sqrt{2}p\right)$$

Set $\log c = -4$, we have

$$G \leqslant \exp\left(-2p + \frac{p}{2\log p} + \sqrt{2}p\right) \to 0$$

which is our desired result.

Why would this happen? An intuitive explanation is that, for $\theta \sim \pi(\theta) \propto e^{-\lambda\|\theta\|_1}$, we have a very heavy tail due to the L_1-norm, since

$$\mathbb{E}[\|\theta\|_1] = \frac{p}{\lambda} = \frac{p}{\sqrt{2\log p}} \geqslant \frac{p}{\log p}$$

3.3.2 Le-Cam Schwarz Argument may Fail

Consider the 1-dimension model

$$Y = \theta + \frac{1}{\sqrt{n}}Z, \quad Z \sim N(0,1)$$

For $\theta \sim \pi$, we want the Bayesian contraction:

$$\mathbb{E}_{Y|\theta^\star}\left[\pi\left\{|\theta - \theta^\star| \geqslant \epsilon_n | Y\right\}\right] \leqslant e^{-cn\epsilon_n^2}$$

For frequentist, we have the result $\mathbb{P}\{|Y - \theta^\star| \geqslant Ln^{-1/2}\} \leqslant 2e^{-L^2/2}$, which derives from the Gaussian tail bound $\mathbb{P}(|Z| \geqslant L) \leqslant 2e^{-L^2/2}$. For frequentist, we just have the Bayesian contraction result with $\epsilon_n = Ln^{-1/2}$. However, if we see the Le Cam-Schwarz argument:

$$\pi\left(|\theta - \theta^\star| \leqslant \epsilon_n\right) \geqslant e^{-c'n\epsilon_n^2}$$

For $\epsilon_n^2 = Ln^{-1/2}$, we have RHS $= e^{-c'L^2}$, which is a constant. The condition fails. A way to fix this is to let $L = c\sqrt{\log n}$.

3.4 High-Dimensional Inference

Consider the model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + Z_{n\times 1}, \quad Z \sim N(0, \sigma^2 I), \sigma^2 = 1$$

When $p < n$, we have $\hat{\beta} - \beta \sim N(0, (X^TX)^{-1})$. But when $p > n$, the inverse does not exist. How we should do inference?

CASE I: LASSO Consider

$$\hat{\beta} = \operatorname*{argmin}_{\beta}\left\{\frac{\|Y - X\beta\|^2}{2} + \lambda\|\beta\|_1\right\}$$

and the special case that $X_i \perp X_j$ when $i \neq j$, $\|X_j\| = \sqrt{n}$, $1 \leqslant j \leqslant p$. It can be shown that

$$\hat{\beta}_{1,OLS} = \frac{X_1^TY}{\|X_1\|^2} \sim N(\beta_1, \frac{1}{n})$$

$$\hat{\beta}_{1,LASSO} = \begin{cases} \dfrac{X_1^TY}{\|X_1\|^2} - \lambda, & \dfrac{X_1^TY}{\|X_1\|^2} \geqslant \lambda \\ \dfrac{X_1^TY}{\|X_1\|^2} + \lambda, & \dfrac{X_1^TY}{\|X_1\|^2} \leqslant -\lambda \\ 0, & \text{otherwise} \end{cases}$$

However, we cannot do inference on this estimator since we don't know its distribution.

CASE II: Write $X_1 = P_{X_1} + P_{X_1}^{\perp}$, where P is the projection to the space $\text{span}\{X_2, \cdots, X_p\}$. Denote $W_1 = P_{X_1}^{\perp}$. Then,

$$\begin{aligned} Y = \sum_{j=1}^{n} \beta_j X_j + Z \quad &\Longrightarrow \quad W_1^T Y = W_1^T \left(\sum_{j=1}^{n} \beta_j X_j \right) + W_1^T Z \\ &\Longrightarrow \quad W_1^T Y = W_1^T X_1 \beta_1 + W_1^T Z = \|W_1\|^2 \beta_1 + W_1^T Z \\ &\Longrightarrow \quad \hat{\beta}_1 = \frac{W_1^T Y}{\|W_1\|^2} \sim N\left(\beta_1, \frac{1}{\|W_1\|^2} \right) \end{aligned}$$

CASE III: Consider the case where p could be more than n, $X_i \overset{\text{i.i.d.}}{\sim} N(0, I)$. Then,

$$\mathbb{E}[X_i^T X_j] = \begin{cases} n, & i = j \\ 0, & i \neq j \end{cases}$$

The first attempt could be:

$$X_1^T Y = \sum_{j=1}^{p} \beta_j X_1^T X_j + X_1^T Z = \beta_1 \|X_1\|^2 + \sum_{j=2}^{p} \beta_j X_1^T X_j + X_1^T Z$$

The estimation is

$$\hat{\beta}_1 = \beta_1 + \sum_{j=2}^{p} \frac{X_1^T X_j}{\|X_1\|^2} \beta_j + \frac{X_1^T Z}{\|X_1\|^2}, \quad \beta_1 = \frac{X_1^T Y}{\|X_1\|^2}$$

The third term has distribution $N(0, 1/n)$. However, the second term has distribution $N(0, \frac{1}{n} \sum_{j=2}^{p} \beta_j^2)$, which could be very large when other β_j's are large. The estimator may fail. Therefore, we have our second attempt to try to estimate other $\hat{\beta}_j$ first, and

$$\hat{\beta}_1^{\star} = \frac{X_1^T Y}{\|X_1\|^2} - \sum_{j=2}^{p} \frac{X_1^T X_j}{\|X_1\|^2} \hat{\beta}_j = \beta_1 + \sum_{j=2}^{p} \frac{X_1^T X_j}{\|X_1\|^2} (\beta_j - \hat{\beta}_j) + \frac{X_1^T Z}{\|X_1\|^2}$$

when the second term vanishes at $o(1/\sqrt{n})$, we have $\hat{\beta}_1^{\star} - \beta_1 \approx N(0, 1/n)$. This is a debiased estimator.

Recall that when $\|\beta\|_0 \leqslant s \ll p$, we have restricted eigenvalue condition: $\|\hat{\beta} - \beta\|_2^2 \lesssim \frac{s \log p}{n}$ w.h.p.. If $\|\hat{\beta}\|_0 \leqslant cs$, where c is a constant, we have a stronger condition: $\|\hat{\beta} - \beta\|_1 \lesssim s\sqrt{\frac{\log p}{n}}$ w.h.p.. We claim that

$$\sup_{2 \leqslant j \leqslant p} \frac{|X_1^T X_j|}{\|X_j\|^2} \leqslant \sqrt{\frac{2 \log p}{n}} \quad w.h.p.$$

Thus,

$$\left| \sum_{j=2}^{p} \frac{X_1^T X_j}{\|X_1\|^2} (\beta_j - \hat{\beta}_j) \right| \lesssim \sqrt{\frac{\log p}{n}} \sum_{j=2}^{p} |\beta_j - \hat{\beta}_j| \lesssim \sqrt{\frac{\log p}{n}} s \sqrt{\frac{\log p}{n}} = \frac{s \log p}{n}$$

This has the rate $o(1/\sqrt{n})$ when $s = o(\sqrt{n}/\log p)$.

CASE IV: We now consider the general case.

$$Y = X_1\beta_1 + \sum_{j=2}^{p} X_j\beta_j + Z, \quad Z \sim N(0, I)$$

We need to find some W_1, e.g., regress X_1 w.r.t. X_2 to X_p, such that

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \left\{ \left\| X_1 - \sum_{j=2}^{p} X_j\gamma_j \right\|^2 + \lambda\|\gamma\|_1 \right\}, \quad \text{residual is } W_1 = X_1 - \sum_{j=2}^{p} X_j\hat{\gamma}_j$$

We have

$$W_1^T Y = W_1^T X_1\beta_1 + \sum_{j=2}^{p} W_1^T X_j\beta_j + W_1^T Z$$

Similarly, we have our debiased estimator.

$$\hat{\beta_1}^\star = \frac{W_1^T Y}{\|W_1\|^2} - \sum_{j=2}^{p} \frac{W_1^T X_j}{\|W_1\|^2}\hat{\beta}_j = \beta_1 + \sum_{j=2}^{p} \frac{W_1^T X_j}{\|W_1\|^2}(\beta_j - \hat{\beta}_j) + \frac{W_1^T Z}{\|W_1\|^2}$$

Example: Gaussian Graphical Model Let $Y_1, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} N(0, \Omega^{-1})$, where Ω is the $p \times p$ precision matrix. $\omega_{ij} = 0$ if and only if $Y_{1i} \perp Y_{1j}$ conditional on rest of the coordinates. If we have $\sup_j \|\Omega_{\cdot j}\|_0 \leqslant s = o(\sqrt{n}/\log p)$, we can do some inference following the procedure above.

(HW7: What we will get if we only have the L_2 restricted eigenvalue condition?)

Chapter 4

Topics on Diffusion Model

4.1 Kernel Smoothing

We observe $Y_1, Y_2, \cdots, Y_n \overset{\text{i.i.d.}}{\sim} f$, where f is a density. The goal is to estimate quantities such as f, f', and $\int f^2$. The assumption given here is

$$\int (f^m(x))^2 \leqslant M$$

A intuitive estimator is the *histogram*

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} I\{x - h/2 \leqslant Y_i < x + h/2\}$$

where we observe that

$$I\{x - h/2 \leqslant Y_i < x + h/2\} = \begin{cases} 1, & w.p. \displaystyle\int_{x-h/2}^{x+h/2} f(y)\,\mathrm{d}y \\ 0, & o.w. \end{cases}$$

Therefore, by LLN, we have $\hat{f}(x) \to \frac{\int_{x-h/2}^{x+h/2} f(y)\,\mathrm{d}y}{h} \approx f(x)$, an approximately unbiased estimator. Generally, we can have the form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{Y_i - x}{h}\right)$$

For example, if $k(y) = I\{-1/2 \leqslant y \leqslant 1/2\}$, then we recover the previous simple case of histogram estimator. We now compute the upper and lower bound of this general estimator.

4.1.1 Upper Bound

Consider $m = 2$ case, we impose four conditions on k:

$$\int k(y)\,\mathrm{d}y = 1, \quad \int y k(y)\,\mathrm{d}y = 0, \quad \int |k(y)| y^2 \leqslant M_1, \quad \int k^2(y)\,\mathrm{d}y \leqslant M_2$$

The first ensures that the window of k is always 1; the second makes k symmetric, the third is for bias analysis and the fourth is for variance calculation. The risk function is

$$R = \mathbb{E}\left[\int (\hat{f}(x) - f(x))^2 \,\mathrm{d}x\right] = \int \mathbb{E}\left[(\hat{f}(x) - f(x))^2\right] \mathrm{d}x = \int (\text{bias}^2 + \text{variance}) \,\mathrm{d}x$$

where

$$\text{bias}^2 = \left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2, \quad \text{variance} = \mathrm{Var}(\hat{f}(x))$$

bias^2 upper bound:

$$\mathbb{E}[\hat{f}(x)] - f(x) = \int \frac{1}{h} k\left(\frac{y-x}{h}\right) f(y) \,\mathrm{d}y - f(x) = \int k(z) f(x+hz) \,\mathrm{d}z - f(x) = \int k(z)[f(x+hz) - f(x)] \,\mathrm{d}z$$

where the first equality holds since Y_i's are i.i.d.; the second holds by change of variable $z = (y-x)/h$, and the third holds by condition 1. When h is small, we hope that the bias is small, with Taylor expansion:

$$\mathbb{E}[\hat{f}(x)] - f(x) = \int k(z) \left[f'(x)hz + (hz)^2 \int_0^1 f''(x+shz)(1-s) \,\mathrm{d}s\right] \mathrm{d}z$$

where we use the *integral remainder* of Taylor expansion. The first term vanishes when integrated w.r.t. z by condition 2. Therefore, we can write is as

$$\mathbb{E}[\hat{f}(x)] - f(x) = h^2 \iint_0^1 k(z) z^2 f''(x+shz)(1-s) \,\mathrm{d}s \,\mathrm{d}z$$

Using Cauchy-Schwarz, $(\int fg)^2 \leqslant \int f^2 \int g^2$, we have

$$\begin{aligned} &\int \left(\iint_0^1 k(z) z^2 f''(x+shz)(1-s) \,\mathrm{d}s \,\mathrm{d}z\right)^2 \mathrm{d}x \\ \leqslant &\int \left(\underbrace{\iint_0^1 |k(z)| z^2 \,\mathrm{d}s \,\mathrm{d}z}_{\leqslant M_1} \iint_0^1 \underbrace{|k(z)| z^2}_{\leqslant M_1} \underbrace{(f''(x+shz)^2)}_{\leqslant M} \underbrace{(1-s)}_{1/3} \,\mathrm{d}s \,\mathrm{d}z\right) \mathrm{d}x = \frac{1}{3} M_1^2 M \end{aligned}$$

Therefore, $\text{bias}^2 \leqslant h^4 \frac{1}{3} M_1^2 M$.

Variance upper bound:

$$\mathrm{Var}(\hat{f}(x)) = \frac{1}{n} \mathrm{Var}\left(\frac{1}{h} k\left(\frac{Y_1 - x}{h}\right)\right) = \frac{1}{n} \int \left(\frac{1}{h} k\left(\frac{Y_1 - x}{h}\right)\right)^2 f(y) \,\mathrm{d}y = \frac{1}{nh} \int k^2(z) f(x+hz) \,\mathrm{d}z$$

where the last equality is a change of variable $y = x + hz$. By condition 3, we have

$$\int \mathrm{Var}(\hat{f}(x)) \,\mathrm{d}x \leqslant \frac{1}{nh} M_2$$

To do the bias-varaince trade-off, since we have risk $\lesssim h^4 + \frac{1}{nh}$, we take $h = n^{-1/5}$, so risk $\lesssim n^{-4/5}$.

Remarks:

- In practice, we can choose h by considering

  $$R = \mathbb{E}\left[\int (\hat{f}(x) - f(x))^2 \,\mathrm{d}x\right] = \mathbb{E}\left[\int \hat{f}(x)^2 \,\mathrm{d}x - 2\int \hat{f}(x)f(x)\,\mathrm{d}x + const\right]$$

  So

  $$\hat{h} = \underset{h}{\operatorname{argmin}} \left\{\int \hat{f}(x)^2 \,\mathrm{d}x - \frac{2}{n}\sum_{i=1}^{n} \hat{f}(Y_i)\right\}$$

  where we replace the second term by the emprical mean: $\int g(x)f(x)\,\mathrm{d}x = \mathbb{E}[g(Y)] \approx \frac{1}{n}\sum_{i=1}^{n} g(Y_i)$.

- To choose k, a result is that $k(y) = \frac{3}{4}(1-y)_+^2$ is the k that minimize the asymptotic constant.

- When $m \neq 2$, we have

  $$R \leqslant c_{k,m}\left(h^{2m} + \frac{1}{nh}\right)$$

  with constraint 2 updated as

  $$\int y^a k(y)\,\mathrm{d}y = 0, \quad a = 1, 2, \cdots, m-1$$

  However, these constraints may make $k(y) \leqslant 0$ at some places. So, we typically let

  $$\hat{f}(x) = \frac{\hat{f}(x)_+}{\int \hat{f}(x)_+ \,\mathrm{d}x}$$

  and it can be shown that the estimation here still achieves the upper bound derived.

- When estimating $\int f^2$, the plug-in estimator using $\hat{f}(x)$ we derived above is not optimal.

(HW8: Estimated f' and get the bound $n^{-2/5}$)

4.1.2 Lower Bound

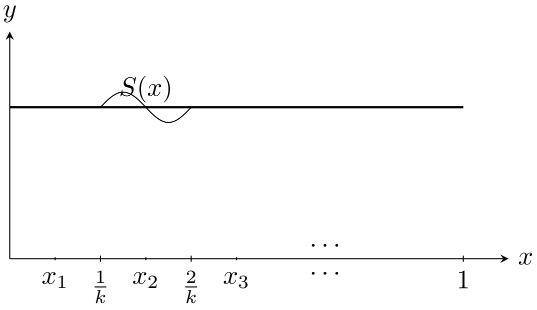
To calculate the lower bound, the first question is: how to construct a sub-parameter space? For simplicity, we consider the parameter space

$$\int_a^b \left(f^\alpha(x)\right)^2 \,\mathrm{d}x \leqslant M, \quad [a,b] = [0,1]$$

We can consider a constant function, separate into equal pieces at $0, 1/k, 2/k, \cdots, 1$, and put a Bayesian $S(x)$ (supported on $[-1/2, 1/2]$) on each part, such that

$$\int s(x)\,\mathrm{d}x = 0, \quad 0 < c_1 \leqslant \int (s^\alpha(x))^2 \,\mathrm{d}x \leqslant c_2 < \infty$$

This is shown in the figure below, where $x_1, x_2, \cdots$ denote the midpoint of each interval.



On each part, we can define

$$S\left(\frac{x-x_i}{1/k}\right), \quad \mathrm{supp} S\left(\frac{x-x_i}{1/k}\right) \subseteq \left[\frac{i-1}{k}, \frac{i}{k}\right], \quad S \text{ differentiable}$$

Note that, if S satisfies all the conditions, then so do aS, so we can always multiply S by a small constant. We consider a subparameter space

$$f_\theta(x) = f_0(x) + \sum_{j=1}^{k} \theta_j k^{-\alpha} S\left(\frac{x-x_j}{1/k}\right), \quad f_0(x) = 1, x \in [0,1], \theta_j \in \{0,1\}$$

Note that we have:

$$\int_0^1 f_\theta(x)\,\mathrm{d}x = 1, \quad f_0(x) \geqslant 0$$

and

$$\int_0^1 \left(f_\theta^{(\alpha)}(x)\right)^2 \mathrm{d}x = \sum_{j=1}^{k} \theta_j k^{-2\alpha} \int_0^1 \left(\left(S\left(\frac{x-x_j}{1/k}\right)\right)^{(\alpha)}\right)^2 \mathrm{d}x = \sum_{j=1}^{k} \theta_j \int_0^1 \left(S^\alpha\left(\frac{x-x_j}{1/k}\right)\right)^2 (k^\alpha)^2\,\mathrm{d}x \leqslant const. \leqslant M$$

where we can separate the integral into pieces because the supports of functions are disjoint. If for some specific S, the integration exceeds M, then we can just time S by a small constant to make it again smaller than M. The goal is to show that: For any $\hat{f}$,

$$\sup_\theta \mathbb{E}\left[\int (\hat{f} - f_\theta)^2\,\mathrm{d}x\right] \gtrsim n^{-\frac{2\alpha}{2\alpha+1}}$$

A natural estimator for the sub-parameter space is to estimate θ's, i.e., $f_{\hat{\theta}}$ where $\hat{\theta}_i \in \{0,1\}$.

$$\begin{aligned}
\mathbb{E}\left[\int (f_{\hat{\theta}}(x) - f_\theta(x))^2\,\mathrm{d}x\right] &= \int \left(\sum_{j=1}^{k} (\hat{\theta}_j - \theta_j) k^{-\alpha} S\left(\frac{x-x_j}{1/k}\right)\right)^2 \mathrm{d}x \\
&= \sum_{j=1}^{k} (\hat{\theta}_j - \theta_j)^2 k^{-2\alpha} \int S^2\left(\frac{x-x_j}{1/k}\right) \mathrm{d}x \\
&= \sum_{j=1}^{k} |\hat{\theta}_j - \theta_j| k^{-2\alpha-1} \int S^2(x)\,\mathrm{d}x \\
&= \sum_{j=1}^{k} |\hat{\theta}_j - \theta_j| \frac{1}{n} \int S^2(x)\,\mathrm{d}x \qquad (k = n^{\frac{1}{2\alpha+1}})
\end{aligned}$$

where in the last step we use change of variable $x' = \frac{x-x_j}{1/k}$. Then,

$$\sup_\theta \mathbb{E}\left[\int (f_{\hat\theta}(x) - f_\theta(x))^2 \,\mathrm{d}x\right] = \sup_\theta \mathbb{E}\left[\sum_{j=1}^k |\hat\theta_j - \theta_j| \cdot \underbrace{\frac{1}{n}\int S^2(x)\,\mathrm{d}x}_{\frac{1}{n}\cdot\text{constant}}\right]$$

If the part outside the underbrace $\geqslant c_3 k$, then we have the whole formula $\gtrsim \frac{1}{n}k = n^{-\frac{2\alpha}{2\alpha+1}}$. Indeed, we will show this below. For the justification that this sub-parameter space leads to the final lower bound, for any $\hat f$, define $f_{\hat\theta}(x)$ with $\hat\theta = \operatorname{argmin}_\theta \int (\hat f - f_\theta(x))^2 \,\mathrm{d}x$. Then,

$$2\int (\hat f(x) - f_\theta(x))^2 \,\mathrm{d}x \geqslant \int (\hat f(x) - f_\theta(x))^2 \,\mathrm{d}x + \int (\hat f(x) - f_{\hat\theta}(x))^2 \,\mathrm{d}x \geqslant \frac{1}{2}\int (f_{\hat\theta}(x) - f_\theta(x))^2 \,\mathrm{d}x$$

where in the first step we use definition of $\hat\theta$, and in the last step we use $a^2 + b^2 \geqslant \frac{1}{2}(a-b)^2$. Therefore, it is enough to consider estimation on this θ instead of the full function $\hat f$.

Now we prove the bound for $\sum_{j=1}^k |\hat\theta_j - \theta_j|$ part. Note that

$$\sup_\theta \mathbb{E}\left[\sum_{j=1}^k |\hat\theta_j - \theta_j|\right] \geqslant \frac{1}{2^k}\sum_{\theta\in\{0,1\}^k}\sum_{i=1}^k \mathbb{E}\left[|\hat\theta_i - \theta_i|\right] = \sum_{i=1}^k \frac{1}{2^k}\sum_{\theta\in\{0,1\}^k} \mathbb{E}\left[|\hat\theta_i - \theta_i|\right]$$

where the first inequality use the fact that supremum is larger than average. Therefore, the only thing we need to do is to show that $\frac{1}{2^k}\sum_\theta \mathbb{E}\left[|\hat\theta_i - \theta_i|\right] \geqslant c_3$. Without loss of generality, we use $i = 1$, and

$$\begin{aligned}
\frac{1}{2^k}\sum_{\theta\in\{0,1\}^k} \mathbb{E}\left[|\hat\theta_1 - \theta_1|\right] &= \frac{1}{2^k}\sum_{\theta_{-1}}\Big(\mathbb{E}[|\hat\theta_1 - 1|] + \mathbb{E}[|\hat\theta_1 - 0|]\Big) \\
&= \frac{1}{2^k}\left(\int |\hat\theta_1 - 1| f_{(1,\theta_{-1})} + \int |\hat\theta_1 - 0| f_{(0,\theta_{-1})}\right) \\
&\geqslant \frac{1}{2^k}\sum_{\theta_{-1}}\int \Big(|\hat\theta_1 - 1| + |\hat\theta_1 - 0|\Big)\min\{f_{(1,\theta_{-1})}(x), f_{(0,\theta_{-1})}\}\,\mathrm{d}x \\
&\geqslant \frac{1}{2^k}\sum_{\theta_{-1}}\int \min\{f_{(1,\theta_{-1})}(x), f_{(0,\theta_{-1})}\}\,\mathrm{d}x \\
&= \frac{1}{2^k}2^{k-1}\int \min\{f_{(1,\theta_{-1})}(x), f_{(0,\theta_{-1})}\}\,\mathrm{d}x \\
&\geqslant \frac{1}{2}\min_{H(\theta,\theta')=1}\int \min\{f_\theta(x), f_{\theta'}(x)\}\,\mathrm{d}x
\end{aligned}$$

where $H(\theta,\theta')$ denotes the Hamming distance between θ and θ'. This can be proved to be bounded away from zero just as what we did in Chapter 1.

(HW9: Estiamte f' and get the bout $n^{-\frac{2(\alpha-1)}{2\alpha+1}}$)

4.2 Diffusion and Score Estimation

Suppose $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} f$, where f is a density. The goal of diffusion model is to generate X_{n+1} with a density approximately f. We assume that f has smoothness α, and

$$\sup_x |f^{(\alpha)}(x)| \leqslant M, \quad x \in [0,1]$$

The diffusion model has two parts:
Forward Process: $X_t = X_0 + W_t, W_t \sim N(0,t)$, $X_T|X_0 \sim N(X_0, T)$ (Write $\mathrm{d}X_t = \mathrm{d}W_t$ in stochastic analysis context). If T is large enough, we will have $X_T|X_0 \approx N(0,T)$, a pure noise.
Backward Process: $\mathrm{d}Y_t = S(Y_t, T-t)\,\mathrm{d}t + \mathrm{d}W_t$, $0 \leqslant t \leqslant T$. $Y_0|X_0 \sim N(0,T)$. i.e., We try to recover X_0 from the pure noise. S is called the *score function.*

Properties of Diffusion:

1. $\mathcal{L}(Y_{T-t}) = \mathcal{L}(X_t)$ for all $0 \leqslant t \leqslant T$. where $\mathcal{L}$ denotes the distribution.
2. $d^2_{\mathrm{TV}}(\mathcal{L}(Y_0), N(0,T)) \lesssim \frac{1}{T}$, where d_{TV} is the total variation distance $d_{TV}(f,g) = \frac{1}{2}\int |f-g|$.
3. If

   (a) $\mathrm{d}Y_t = S(Y_t, T-t)\,\mathrm{d}t + \mathrm{d}W_t$, $0 \leqslant t \leqslant T$, $Y_0|X_0 \sim N(X_0, T)$.

   (b) $\mathrm{d}Y_t = \hat{S}(\tilde{Y_t}, T-t)\,\mathrm{d}t + \mathrm{d}W_t$, $0 \leqslant t \leqslant T$, $\tilde{Y_0} \sim N(0,T)$.

   Then,
   $$d^2_{\mathrm{TV}}(\mathcal{L}(\tilde{Y_T}), \mathcal{L}(X_0)) = d^2_{\mathrm{TV}}(\mathcal{L}(\tilde{Y_T}), \mathcal{L}(Y_t)) \lesssim \frac{1}{T} + \int_0^T \int \left(\hat{S}(x,t) - S(x,t)\right)^2 p(x,t)\,\mathrm{d}x\,\mathrm{d}t$$
   where $p(x,t) = f * \varphi_t$, the convolution of density of X_t (f_t) and density of W_t (φ_t).

It turns out that the correct score function is:

$$S(x,t) = \frac{p'(x,t)}{p(x,t)} = (\log p(x,t))', \quad p(x,t) = \int_0^1 \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} f(y)\,\mathrm{d}y$$

A justification: If we observe $Y = \theta + z$, where $Z \sim N(0,1)$, we want to estimate $\theta \sim \pi(\theta)$. One approach is to use the posterior mean:

$$\mathbb{E}[\theta|Y] = \frac{\int \theta\varphi(y-\theta)\pi(\theta)\,\mathrm{d}\theta}{\int \varphi(y-\theta)\pi(\theta)\,\mathrm{d}\theta} = Y + \frac{\int (\theta - Y)\varphi(y-\theta)\pi(\theta)\,\mathrm{d}\theta}{\int \varphi(y-\theta)\pi(\theta)\,\mathrm{d}\theta} = Y + \frac{p'(Y)}{p(Y)}$$

where

$$p(x) = \int \varphi(x-\theta)\pi(\theta)\,\mathrm{d}\theta$$

Our claim is that:

$$\inf_{\hat{S}} \sup_{S} \mathbb{E}\left[\int (\hat{S}(x,t) - S(x,t))^2 p(x,t)\,\mathrm{d}x\right] \asymp \begin{cases} n^{-\frac{2(\alpha-1)}{2\alpha+1}}, & 0 \leqslant t < n^{-\frac{2}{2\alpha+1}}, \\ \dfrac{1}{nt^{3/2}}, & n^{-\frac{2}{2\alpha+1}} \leqslant t \leqslant 1, \\ \dfrac{1}{nt^2}, & t > 1 \end{cases}$$

If this claim is true, $T = n$, then

$$\begin{aligned}\mathbb{E}\left[d_{\mathrm{TV}}^2(\mathcal{L}(\tilde{Y}_t), \mathcal{L}(X_0))\right] &\lesssim \frac{1}{T} + \int_0^T \left\{\mathbb{E}\left[\int (\hat{S}(x,t) - S(x,t))^2 p(x,t)\, \mathrm{d}x\right]\right\} \mathrm{d}t \\ &\lesssim \frac{1}{n} + \int_0^{n^{-\frac{2}{2\alpha+1}}} n^{-\frac{2(\alpha-1)}{2\alpha+1}}\, \mathrm{d}t + \int_{n^{-\frac{2}{2\alpha+1}}}^1 \frac{1}{nt^{3/2}}\, \mathrm{d}t + \int_1^T \frac{1}{nt^2}\, \mathrm{d}t \lesssim n^{-\frac{2\alpha}{2\alpha+1}}\end{aligned}$$

A proof sketch of the claim: Since p would be canceled out, we only consider the estimation of p'

(1) $0 < t < n^{-\frac{2}{1+2\alpha}}$:

$$p'(x.y) = \int f'(x-y) \frac{1}{\sqrt{2\pi t}} e^{-\frac{y^2}{2t}}\, \mathrm{d}y$$

From previous section, we have the kernel smoothing estimate f' has the rate $\frac{1}{nh^3} + h^{2(\alpha-1)}$, and when $h = n^{-\frac{1}{1+2\alpha}}$, we have the rate $n^{-\frac{2(\alpha-1)}{2\alpha+1}}$. When t is small, the estimation is like a kernel smoothing with bandwidth $\sqrt{t} = n^{-\frac{1}{1+2\alpha}}$, so have the same rate.

(2) $n^{-\frac{1}{1+2\alpha}} < t < 1$,

$$p'(x,t) = \int_0^1 -\frac{x-y}{t} \cdot \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} f(y)\, \mathrm{d}y = \mathbb{E}\left[-\frac{x-x_0}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_0)^2}{2t}}\right]$$

A natural estimator is the empirical estiamtion

$$\hat{p}'(x,t) = \frac{1}{n} \sum_{i=1}^n \left(-\frac{x-x_i}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_i)^2}{2t}}\right)$$

We have

$$\mathbb{E}\left[(\hat{p}'(x,t) - p'(x,t))^2\right] = \mathrm{Var}(\hat{p}'(x,t)) \lesssim \frac{1}{nt^2} \mathrm{Var}\left((x-x_0) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_0)^2}{2t}}\right)$$

Use change of variable $x = x_0 + \sqrt{t}z$,

$$\begin{aligned}\mathbb{E}\left[(\hat{p}'(x,t) - p'(x,t))^2\right] &\lesssim \frac{1}{nt^2} \mathrm{Var}\left((x-x_0) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_0)^2}{2t}}\right) \\ &= \frac{1}{nt^2} \int tz^2 \frac{1}{t} e^{-\frac{z^2}{2}} \sqrt{t}\, \mathrm{d}t = \frac{1}{nt^2} \sqrt{t} = \frac{1}{nt^{3/2}}\end{aligned}$$

(3) $t > 1$

$$p'(x,t) = \int_0^1 -\frac{x-y}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} f(y)\, \mathrm{d}y = -\frac{x}{t} p(x,t) + \int_0^1 \frac{y}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} f(y)\, \mathrm{d}y$$

For score estimation, the first term cancels p out, so we don't need the estimation of the first term. The only concern is the second term. Let $p_L'(x,t) = \int_0^1 \frac{y}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}} f(y)\, \mathrm{d}y$. Again, a natural estimator is the empirical estimation

$$\hat{p}_L'(x,t) = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_i)^2}{2t}}$$

Then,

$$\mathbb{E}\left[(\hat{p}'_L(x,t) - p_L(x,t))^2\right] \leqslant \frac{1}{nt^2}\underbrace{\mathbb{E}\left[\left(x_0\frac{1}{\sqrt{2\pi t}}e^{-\frac{(x-x_0)^2}{2t}}\right)^2\right]}_{\text{bounded}} \lesssim \frac{1}{nt^2}$$

(HW10: Get the lower bound)

Chapter 5

More on Lower Bounds

5.1 Le Cam's Two-Points Argument

$Y \sim P_\theta$, where $\theta \in \Theta$. The goal is to calculate

$$\inf_{\hat{\theta}} \sup_{\Theta} \mathbb{E}\left[d^p(\hat{\psi}(\theta), \psi(\theta))\right]$$

where d denotes the distance. We choose $\theta^{(0)}, \theta^{(1)} \in \Theta$, and for $p = 1$,

$$\begin{aligned}
\sup_{\theta \in \{\theta^{(0)}, \theta^{(1)}\}} \mathbb{E}\left[d\left(\hat{\psi}(\theta), \psi(\theta)\right)\right] &\geqslant \frac{1}{2}\int d\left(\hat{\psi}(\theta), \psi(\theta^{(0)})\right) P_{\theta^{(0)}}\,\mathrm{d}\mu + \frac{1}{2}\int d\left(\hat{\psi}(\theta), \psi(\theta^{(1)})\right) P_{\theta^{(1)}}\,\mathrm{d}\mu \\
&\geqslant \frac{1}{2}\int \min\{P_{\theta^{(0)}}, P_{\theta^{(1)}}\}\left[d\left(\hat{\psi}(\theta), \psi(\theta^{(0)})\right) + d\left(\hat{\psi}(\theta), \psi(\theta^{(1)})\right)\right]\mathrm{d}\mu \\
&\geqslant \frac{1}{2} d(\psi(\theta^{(0)}), \psi(\theta^{(1)})) \int \min\{P_{\theta^{(0)}}, P_{\theta^{(1)}}\}\,\mathrm{d}\mu \qquad \text{(Triangle Inequality)}
\end{aligned}$$

Examples:
(i) $Y \sim N(\theta, \frac{1}{n})$, $\theta \in \mathbb{R}$, choose $\Theta_{sub} = \{0, 1/\sqrt{n}\}$, rate $n^{-1/2}$.
(ii) $Y = (Y_1, \cdots, Y_n)$, $Y_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$, choose $\Theta_{sub} = \{1, 1 + 1/n\}$, rate n^{-1}.

Remark: for pth moment, we have, by Jensen's inequality

$$\mathbb{E}\left[d^p(\hat{\psi}(\theta), \psi(\theta))\right] \geqslant \left(\mathbb{E}\left[d(\hat{\psi}(\theta), \psi(\theta))\right]\right)^p, \quad p \geqslant 1$$

so if we solve the lower bound for $p = 1$, we solve the lower bound for higher order.

A slight Extension to the two-points argument: $\Theta_{sub} = \{\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(M)}\}$. Then,

$$\begin{aligned}
\sup_{\theta \in \{\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(M)}\}} \mathbb{E}\left[d\left(\hat{\psi}(\theta), \psi(\theta)\right)\right] &\geqslant \frac{1}{2}\int d(\hat{\psi}(\theta), \psi(\theta^{(0)})) P_{\theta^{(0)}}\,\mathrm{d}\mu + \frac{1}{2M}\sum_{i=1}^{M}\int d(\psi(\theta), \psi(\theta^{(i)})) P_{\theta^{(i)}}\,\mathrm{d}\mu \\
&\geqslant \frac{1}{2}\int d(\hat{\psi}(\theta), \psi(\theta^{(0)})) P_{\theta^{(0)}}\,\mathrm{d}\mu + \frac{1}{2M}\sum_{i=1}^{M}\int \min_i\{d(\psi(\theta), \psi(\theta^{(i)}))\} P_{\theta^{(i)}}\,\mathrm{d}\mu
\end{aligned}$$

$$= \frac{1}{2}\int d(\hat{\psi}(\theta), \psi(\theta^{(0)}))P_{\theta^{(0)}}\,\mathrm{d}\mu + \frac{1}{2}\int \min_i\{d(\psi(\theta), \psi(\theta^{(i)}))\} \qquad \underbrace{\frac{1}{M}\sum_{i=1}^{M} P_{\theta^{(i)}}}_{\text{mixture of distribution}} \qquad \mathrm{d}\mu$$

$$\geqslant \frac{1}{2}\min_i\left\{d(\psi(\theta^{(0)}), \psi(\theta^{(i)})\right\}\int \min\left\{P_{\theta^{(0)}}, \frac{1}{M}\sum_{i=1}^{M} P_{\theta^{(i)}}\right\}\mathrm{d}\mu$$

This is useful, for example, when you do kernel estimation, there must be more than two points that should be considered.

5.2 Assouad's Lemma

Let $\theta \in \{0,1\}^k$, $Y \sim P_\theta$, the Assouad's lemma said that

$$\mathbb{E}\left[2^p d^p(\hat{\psi}(\theta), \psi(\theta))\right] \geqslant \inf_{H(\theta,\theta')\geqslant 1}\frac{d^p(\psi(\theta), \psi(\theta'))}{H(\theta,\theta')}\cdot\frac{k}{2}\cdot\min_{H(\theta,\theta')=1}\int \min\{P_\theta, P_{\theta'}\}\,\mathrm{d}\mu$$

Example: For $p = 2$, $\int (f^{(\alpha)})^2 \leqslant M$, $x \in [0,1]$, we use the sub-parameter space

$$f_\theta(x) = 1 + \sum_{i=1}^{k}\theta_i h^\alpha k\left(\frac{x - x_i}{h}\right), \quad k \asymp n^{\frac{1}{2\alpha+1}}, \quad h \asymp \frac{1}{k}$$

(i) $\int (f_\theta(x) - f_{\theta'}(x))^2\,\mathrm{d}x \asymp \frac{1}{n}H(\theta,\theta')$, $\min_{H(\theta,\theta')=1}\int \min\{P_\theta, P_{\theta'}\}\,\mathrm{d}\mu \geqslant c$, which leads to the rate $n^{-\frac{2\alpha}{2\alpha+1}}$.
(ii) $\int (f'_\theta(x) - f'_{\theta'}(x))^2\,\mathrm{d}x \asymp n^{-\frac{2\alpha-1}{2\alpha+1}}H(\theta,\theta')$, $\min_{H(\theta,\theta')=1}\int \min\{P_\theta, P_{\theta'}\}\,\mathrm{d}\mu \geqslant c$, which leads to the rate $n^{-\frac{2(\alpha-1)}{2\alpha+1}}$.

5.3 Fano's Lemma

$Y \sim P_\theta$, $\theta \in \Theta$, construct $\Theta_{sub} = \{\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(M)}\} \subseteq \Theta$ such that $d(\psi(\theta^{(i)}), \psi(\theta^{(j)})) \geqslant \epsilon$ for $i \neq j$. Let

$$\max_{i,j}\mathrm{KL}(P_{\theta^{(i)}} \,\|\, P_{\theta^{(j)}}) \leqslant T$$

Then,

$$\sup_{\theta\in\Theta_{sub}}\mathbb{E}\left[d^p(\hat{\psi}(\theta), \psi(\theta))\right] \geqslant \left(\frac{\epsilon}{2}\right)^p\underbrace{\left(1 - \frac{T + \log 2}{\log M}\right)}_{\text{usually}\geqslant c}$$

Example: Let us still use the kernel smoothing example with $p = 2$. Then, $\psi(\theta) = f_\theta$. Choose

$$\sup_\theta \int (\hat{f}_\theta - f_\theta)^2 \gtrsim n^{-\frac{2\alpha}{2\alpha+1}} \quad \Longleftrightarrow \quad \epsilon \asymp n^{-\frac{\alpha}{2\alpha+1}}$$

If we still use the sub-parameter space as we did before, the problem is that, $\frac{1}{n}H(\theta,\theta')$ can be as small as $\frac{1}{n}$, which violates the assymption in the Fano's lemma. The solution is: choose $\Theta'_{sub} \subseteq \Theta_{sub} = \{0,1\}^k$ such that

$$(i) : \int (f_{\theta^{(i)}} - f_{\theta^{(j)}})^2 \gtrsim n^{-\frac{2\alpha}{2\alpha+1}}, \quad i \neq j$$

We claim that

$$(ii) : \mathrm{KL}\left(P_{\theta^{(i)}} \| P_{\theta^{(j)}}\right) \asymp n \int (f_{\theta^{(i)}} - f_{\theta^{(j)}})^2 \leqslant c_1 n \underbrace{\frac{1}{n} k}_{\max \frac{1}{n} H(\theta,\theta')} = c_1 k$$

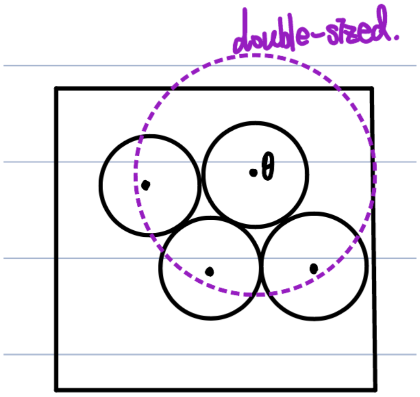
Finally, assume we can make

$$(iii) : \log M \geqslant c_2 k \quad \text{for } c_2 > 0$$

Then, the second term in Fano's bound becomes $1 - \frac{c_1 k + \log 2}{c_2 k}$. When k is large, $\log 2$ is negligible, and it becomes $1 - \frac{c_1}{c_2}$. To make it larger than zero, we can always times the kernel by a small constant to make c_1 smaller. With all $(i), (ii), (iii)$ holds, we can use Fano's lemma to get the same bound as before.

A justification on (iii): Use **Varshmov-Gilbert Lemma:**
There are M points from $\{0,1\}^k$ such that $H(\theta^{(i)}, \theta^{(j)}) \geqslant k/4$, and $M \geqslant c'^k$, where $c' > 1$.

proof sketch: If we want to make $H(\theta^{(i)}, \theta^{(j)}) \geqslant k/4$ and make M as much large as it can be, we are actually solving a *maximum packing problem*, where we try to pack as many 'Hamming ball' inside $\{0,1\}^k$ without overlapping. Each Hamming ball should have radius $H(\theta, \theta') \leqslant k/8$.



Assume that we achieve the maximum pack. Then, we cannot pack any ball with the same size in the space without overlapping. Equivalently, we cannot have any other ball with center having distance larger than $k/4$ with any other ball centers. Therefore, doubling the size of each ball would give a cover of this whole space. This gave us the ineuality

$$M \binom{k}{\frac{k}{4}} \geqslant 2^k \implies M \geqslant \frac{2^k}{\frac{k!}{(k/4)!(3k/4)!}}$$

Use the Stirling formula to approximate factorials

$$m! = (1 + o(1))\sqrt{2\pi m} \left(\frac{m}{e}\right)^m$$

We have

$$\begin{aligned}\frac{k!}{(k/4)!(3k/4)!} &\approx \frac{\sqrt{2\pi k}(k/e)^k}{\sqrt{2\pi k/4}(k/(4e))^{k/4}\sqrt{\frac{3}{4}2\pi k}(3k/(4e))^{3k/4}} \\ &= \frac{\sqrt{2\pi k}}{\sqrt{2\pi\frac{k}{4}}\sqrt{2\pi\frac{3}{4}k}}\left(4^{1/4}\right)^k\left[\left(\frac{4}{3}\right)^{\frac{3}{4}}\right]^k \\ &\asymp \left(4^{\frac{1}{4}}\cdot\left(\frac{4}{3}\right)^{\frac{3}{4}}\right)^k\end{aligned}$$

and it turns out that $4^{\frac{1}{4}}\cdot\left(\frac{4}{3}\right)^{\frac{3}{4}} < 2$, which finally shows that

$$M \geqslant c'^k \quad \text{with } c' > 1$$

The class ends! It is indeed my favorite class this semester!