Stat450 2025 Winter - Final Report

On Minimax Rate of ℓ_1 norm in Gaussian Location Model

Tuorui Peng[1]

Abstract

In this project, we followed the idea of [2] to present the minimax rate for estimating non-smooth functions of mean in Gaussian location model. Specifically, we focused on the estimation of $\|\theta\|_1 := \sum_{i=1}^p |\theta_i|$ from an observation $X \sim \mathcal{N}(\theta, I_p)$. The main result presented is that: the minimax rate under ℓ_2 loss is $M^2 p^2 \left(\frac{\log\log p}{\log p}\right)^2$ for M-bounded parameter, and $\frac{p^2}{\log p}$ for unbounded case.

1 Motivation and Introduction

1.1 Introduction

Gaussian location model is a basic and important model in statistics. Here we focus on the ℓ_2^2 loss Minimax rate for location parameter estimation. To be specific, we have $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, I_p)$ where $\theta \in \Theta_p \subseteq \mathbb{R}^p$ is the parameter of interest. Some estimation problem includes $T(\theta) = \theta$, $T(\theta) = \theta_{\max}$, $T(\theta) = \|\theta\|_2$, $T(\theta) = \|\theta\|_1$, etc. The minimax rate under ℓ_2^2 loss for these problems are well studied.

$$\mathfrak{M}^*(T, \Theta_p) := \inf_{\hat{T}} \sup_{\theta \in \Theta_p} \mathbb{E}\left[(T(\theta) - \hat{T}(Y))^2\right]$$

For example, we have some well-known results as follows:

Estimation Problem	Minimax Rate
$T(\theta) = \theta$	$\frac{p}{n}$
$T(\theta) = \theta_{\max}$	$\frac{p^2}{n}$
$T(\theta) = \|\theta\|_2$	$\frac{\sqrt{p}}{n}$

In the lecture, we already proved the case of $T(\theta) = \theta$ (c.f. Lecture 5) and $T(\theta) = \|\theta\|_2$ (c.f. Lecture 7[2]). Here in this project, we will focus on the case of $T(\theta) = \|\theta\|_1$.

Due to the property of ℓ_2^2 loss, it suffices to consider one sample case and deduce that $\mathfrak{M}_{n=n} = \mathfrak{M}_{n=1}/n$. i.e. our setting is that $X \sim \mathcal{N}(\theta, I_p)$ and the task is to estimate $\|\theta\|_1 := \sum_{i=1}^p |\theta_i|$.

1.2 Challenge

$\|\theta\|_1$ is a typical example of a non-smooth functional of parameter. While the distribution of normal distribution is 'much too nice' by being analytic. e.g. our first trial of estimation might be $\hat{T}(X) = \|X\|_1$, however we

[1] TuoruiPeng2028@u.northwestern.edu

[2] We got that $\varepsilon^* \asymp d^{1/4}/n^{1/2}$. Applying Le Cam's two-point method we have $\mathfrak{M}^* \gtrsim (\varepsilon^*)^2 \asymp d/n$.

would immediately see that it's biased by noticing that (take 1-dim as example):

$$\int_{\mathbb{R}} |X| \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-\theta)^2}{2}\right] \mathrm{d}x \overset{?}{=} |\theta|$$

in which left hand side is a smooth function of θ and right hand side is not. Actually no estimator $\hat{T}(X)$ can satisfy the analytic property, so we have no access to an unbiased estimator to $|\theta|$.

However, we do have access to some other analytic estimators. For example, we can safely estimate moments (i.e. polynomial functions) of θ. Thus here the key recipe is to approximate $|\theta|$ by a polynomial function (of proper degree) of θ.

1.3 Outline of the Proof

We would first do the proof for bounded case:

1. We can approximate $|x|$ by a degree-d polynomial up to $\asymp \frac{1}{d}$ precision;
2. The variance of the polynomial estimator is $\asymp p \cdot d!$;
3. A bias-variance trade-off gives us a desired upper bound side;
4. Control χ^2 divergence to get the lower bound side.

Then a extension to unbounded case is easy by noticing that: when $\|\theta\|_1$ is large enough, $\|X\|$ would be similar to $\|\theta\|_1$, so we only need to consider the case that $\|\theta\|_1$ is small enough, and study the rate of the 'boundary', then go back to bounded case.

2 Useful Knowledge

2.1 Polynomial Approximation of $|x|$

A main recipe is the approximation of $|x|, x \in [-1,1]$ by a polynomial function. In this project, we use d for the degree of polynomial, and we seek a best degree-d polynomial estimator $G_d^*(x)$ in the sense that reaches:

$$\delta_d(f = |\cdot|) = \inf_{G_d \in \mathcal{P}_d} \max_{x \in [-1,1]} \big| |x| - G_d(x) \big|$$

i.e. optimal in $\|\cdot\|_\infty$, in which $\mathcal{P}_d$ is the set of degree-d polynomial functions. The classical Chebyshev's alternation theorem [2] states that the optimal polynomial $G_d^*(x)$ is s.t.: $|x| - G_d^*(x)$ take value consecutively $\delta_d, -\delta_d, \delta_d, \ldots$. And Berstein [1] proved that this optimal error is of order $\asymp \frac{1}{d}$:

$$\lim_{d\to\infty} d\delta_d(f = |\cdot|) = \beta_* \approx 0.2802 \qquad \text{(Berstein's constant)}$$

Upgrade to p dimensional, we have that the optimal polynomial approximation of $\|\theta\|_1$ with precision rate $\asymp \frac{p}{d}$.

2.2 Hermite Polynomial

Hermite polynomial(s) is a set of orthogonal polynomial with respect to the normal density weight function $\exp\left[-\frac{x^2}{2}\right]$, defined as:

$$H_r(x) := (-1)^r \exp\left[\frac{x^2}{2}\right] \frac{\mathrm{d}^r}{\mathrm{d}x^r} \exp\left[-\frac{x^2}{2}\right]$$

with orthogonality:

$$\langle H_r, H_s\rangle = \int_{\mathbb{R}} H_r(x)H_s(x)\exp\left[-\frac{x^2}{2}\right]\mathrm{d}x = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[H_r(Z)H_s(Z)\right] = r!\delta_{rs}$$

and it form a direct estimator to moments:

$$\mathbb{E}_{Z\sim\mathcal{N}(\theta,1)}\left[H_r(Z)\right] = \int_{\mathbb{R}} H_r(x)\exp\left[-\frac{(x-\theta)^2}{2}\right]\mathrm{d}x = \theta^r$$

When using Hermite polynomial up to degree d to approximate $|x|_1$ as

$$|x|_1 \approx G_d^*(x) = \sum_{r=0}^{d} \alpha_r H_r(x), \quad x\in[-1,1],$$

we have that $\alpha_r \lesssim \mathrm{const}^{\mathrm{const}\cdot d}$, i.e. $\log\alpha_r \lesssim d$ [2].[3]

3 Bounded Case

We first focus on the bounded case that $\theta\in\Theta_p(M) := \{\theta\in\mathbb{R}^p : |\theta|_\infty \le M\}$ for some $M>0$. WLOG we assume $M=1$.

3.1 Upper Bound Side

We use the above mentioned polynomial approximation (the degree to be determined later) to construct a (good enough) estimator, which would give us a upper bound side of the minimax rate.

$$\hat{T}_i(X) = G_d^*(X_i) = \sum_{r=0}^{d}\alpha_r H_r(X_i), \quad i\in[p],$$

for which we know that the ℓ_2^2 loss can be decomposed to the following bias-variance trade-off:

$$\text{Bias: } \left|\hat{T}(X) - |X|_1\right| = \delta_d(f=|\cdot|) \lesssim \frac{p}{d},$$

$$\text{Variance: } var(\hat{T}(X)) = \sum_{i=1}^{p} var(\hat{T}_i(X)) \lesssim p\cdot d!\cdot \mathrm{const}^{\mathrm{const}\cdot d} \lesssim p\cdot d!.$$

To obtain the optimal loss rate, we have:

$$\text{optimal } d: \quad (\frac{p}{d})^2 \asymp p\cdot d! \overset{(i)}{\Rightarrow} d_p \asymp \frac{\log p}{\log\log p}$$

$$\Rightarrow \text{optimal loss:} \quad \mathfrak{M}^* \lesssim \left(\frac{p}{d_p}\right)^2 + p\cdot d_p! \lesssim p^2\left(\frac{\log\log p}{\log p}\right)^2.$$

where (i) uses Stirling's formula $d!\sim(d/e)^d$.

[3] Intuitively, $\alpha_r x^r \lesssim 1$.

3.2 Lower Bound Side

We use the χ^2 divergence + Le Cam's two-point method we introduced in the lecture to get the lower bound side. We wanna construct

► (Upper) Bounding χ^2 divergence by selecting proper prior and degree of polynomial.

Denote the (posterior) distribution of Null and Alternative as $f_0(x) = \pi_0(\theta) \otimes \phi(x;\theta) = \int_{\mathbb{R}} \phi(x-t)\pi_0(\,\mathrm{d}t)$ and $f_1(x) = \pi_1(\theta) \otimes \phi(x;\theta) = \int_{\mathbb{R}} \phi(x-t)\pi_1(\,\mathrm{d}t)$ respectively. We have the following two useful lemmas:

- By the convexity of $\xi \mapsto e^{-\xi}$, we have

$$\begin{aligned} f_0(x) =& \int_{\mathbb{R}} \phi(x-t)\pi_0(\,\mathrm{d}t) \\ \geq& \frac{1}{\sqrt{2\pi}} \exp\left[-\int_{\mathbb{R}} \frac{(x-t)^2}{2}\pi_0(\,\mathrm{d}t)\right] \\ =& \phi(x)\exp\left[-\frac{1}{2}\int t^2\pi_0(\,\mathrm{d}t)\right] \\ \geq& \phi(x)\exp\left[-\frac{1}{2}\right] \end{aligned}$$

- Taylor expansion of normal distribution:

$$\phi(x-t) = \sum_{k=0}^{\infty} H_k(x)\phi(x)\frac{t^k}{k!}.$$

  and thus we have

$$\begin{aligned} f_1(x) =& \int_{\mathbb{R}} \phi(x-t)\pi_1(\,\mathrm{d}t) \\ =& \int_{\mathbb{R}} \sum_{k=0}^{\infty} H_k(x)\phi(x)\frac{t^k}{k!}\pi_1(\,\mathrm{d}t) \\ =& \sum_{k=0}^{\infty} \frac{1}{k!} H_k(x)\phi(x)\int_{\mathbb{R}} t^k\pi_1(\,\mathrm{d}t) \end{aligned}$$

  (and similarly for $f_0(x)$).

- We study the following χ^2 divergence:

$$\begin{aligned}
\chi^2(f_1\|f_0) &= \int_{\mathbb{R}} \frac{(f_1(x)-f_0(x))^2}{f_0(x)}\,\mathrm{d}x \\
&\overset{(i)}{=} \int_{\mathbb{R}} \frac{\left(\sum_{k=0}^{\infty}\frac{1}{k!}H_k(x)\phi(x)\int_{-1}^{1}t^k(\pi_1(\,\mathrm{d}t)-\pi_0(\,\mathrm{d}t))\right)^2}{f_0(x)}\,\mathrm{d}x \\
&\overset{(ii)}{=} \int_{\mathbb{R}} \frac{\left(\sum_{k=d+1}^{\infty}\frac{1}{k!}H_k(x)\phi(x)\int_{-1}^{1}t^k(\pi_1(\,\mathrm{d}t)-\pi_0(\,\mathrm{d}t))\right)^2}{f_0(x)}\,\mathrm{d}x \\
&\overset{(iii)}{\le} e^{1/2}\int_{\mathbb{R}}\Big(\sum_{k=d+1}^{\infty}\frac{1}{k!}H_k(x)\phi(x)\int_{\mathbb{R}}t^k(\pi_1(\,\mathrm{d}t)-\pi_0(\,\mathrm{d}t))\Big)^2\phi(x)^{-1}\,\mathrm{d}x \\
&\overset{(iv)}{=} e^{1/2}\sum_{k=d+1}^{\infty}\frac{1}{k!}\Big(\int_{-1}^{1}t^k\pi_1(\,\mathrm{d}t)-\int_{-1}^{1}t^k\pi_0(\,\mathrm{d}t)\Big)^2 \\
&\le e^{1/2}\sum_{k=d+1}^{\infty}\frac{1}{k!} \\
&\overset{(v)}{\lesssim} \frac{1}{d!}
\end{aligned}$$

  where (i) using the above mentioned Taylor expansion, (ii) since our prior is selected s.t. having matching first d moments, (iii) using the lemma mentioned above, and (iv) using the orthogonality of Hermite polynomial, and (v) since $1/k!$ decays faster than geometric series so is also dominated by the leading them $1/d!$.

Thus we have:

$$\begin{aligned}
\chi^2\left(\mathbb{E}_{\theta\sim\pi_1^{\otimes p}}\left[p_\theta^{\otimes p}\right]\Big\|\mathbb{E}_{\theta\sim\pi_0^{\otimes p}}\left[p_\theta^{\otimes p}\right]\right)+1 &= \chi^2\left(f_1^{\otimes p}\|f_0^{\otimes p}\right) \\
&\overset{(i)}{=} \Big(\chi^2(f_1\|f_0)+1\Big)^p \\
&\le (\frac{c}{d!}+1)^p \\
&\le \exp\left[\frac{cp}{d!}\right]
\end{aligned}$$

Thus by selecting $d_p \gtrsim \dfrac{\log p}{\log\log p}$ we have $\chi^2\left(\mathbb{E}_{\theta\sim\pi_1^{\otimes p}}\left[p_\theta^{\otimes p}\right]\Big\|\mathbb{E}_{\theta\sim\pi_0^{\otimes p}}\left[p_\theta^{\otimes p}\right]\right)\lesssim 1$ and thus by Le Cam's two-point method we have:

$$\begin{aligned}
\mathfrak{M}^* &\gtrsim \varepsilon^2\Big(1-d_{\mathrm{TV}}\left(\mathbb{E}_{\theta\sim\pi_1^{\otimes p}}\left[p_\theta^{\otimes p}\right]\Big\|\mathbb{E}_{\theta\sim\pi_0^{\otimes p}}\left[p_\theta^{\otimes p}\right]\right)\Big) \\
&\ge \varepsilon^2\Big(1-\frac{1}{2}\sqrt{\chi^2\left(\mathbb{E}_{\theta\sim\pi_1^{\otimes p}}\left[p_\theta^{\otimes p}\right]\Big\|\mathbb{E}_{\theta\sim\pi_0^{\otimes p}}\left[p_\theta^{\otimes p}\right]\right)}\Big) \\
&\gtrsim \varepsilon^2
\end{aligned}$$

where for the rate of ε, we first describe its intuition, then formalize it.

▶ (Lower) Bounding the gap between Null and Alternative.

Intuition: As we mentioned in subsection 2.1, a polynomial of degree d can approximate $|x|$ up to precision $\delta_d(|\cdot|)=\frac{1}{d}$, so the smallest gap between Null and Alternative should also be of order $\asymp\frac{1}{d}$.

Formalization: We can formalize the above as a optimization problem:

$$\begin{aligned}&\arg\max_{\pi_0,\pi_1}\mathbb{E}_{\pi_0}\left[|\theta|\right]-\mathbb{E}_{\pi_1}\left[|\theta|\right]\\&w.r.t.\mathbb{E}_{\pi_0}\left[\theta^k\right]=\mathbb{E}_{\pi_1}\left[\theta^k\right],k=0,1,\ldots,d\end{aligned}$$

i.e. we just take π_0,π_1 as the optimizer and it would yield $\varepsilon=\mathbb{E}_{\pi_0}\left[|\theta|\right]-\mathbb{E}_{\pi_1}\left[|\theta|\right]$. In the preceeding part we have proved that with the constraint satisfied with proper d, we have $\mathfrak{M}^*\gtrsim\varepsilon^2$.

Now solve the optimization problem by writing down the Lagrangian:

$$\begin{aligned}\mathcal{L}(\pi_0,\pi_1,\lambda_1,\ldots,\lambda_d)=&\mathbb{E}_{\pi_0}\left[|\theta|\right]-\mathbb{E}_{\pi_1}\left[|\theta|\right]-\sum_{k=0}^{d}\lambda_k\big(\mathbb{E}_{\pi_0}\left[\theta^k\right]-\mathbb{E}_{\pi_1}\left[\theta^k\right]\big)\\=&\mathbb{E}_{\pi_0}\left[|\theta|-\sum_{k=0}^{d}\lambda_k\theta^k\right]-\mathbb{E}_{\pi_1}\left[|\theta|-\sum_{k=0}^{d}\lambda_k\theta^k\right]\end{aligned}$$

optimize w.r.t. π_0,π_1 gives us:

$$\max_{\pi_0,\pi_1}\mathcal{L}=\max_{\theta\in[-1,1]}[|\theta|-\mathrm{poly}_d(\theta)]-\min_{\theta\in[-1,1]}[|\theta|-\mathrm{poly}_d(\theta)]$$

then optimize w.r.t. λ_k suffices to choosing the optimal degree-d polynomial approximation of $|x|$, for which we have study in subsection 2.1:

$$\begin{aligned}\max_{\pi_0,\pi_1}\mathbb{E}_{\pi_0}\left[|\theta|\right]-\mathbb{E}_{\pi_1}\left[|\theta|\right]=&\min_{\lambda_{k=0}^d}\max_{\pi_0,\pi_1}\mathcal{L}\\=&\inf_{G_d\in\mathcal{P}_d}\max_{\theta\in[-1,1]}\big||\theta|-G_d(\theta)\big|\\=&\delta_d(|\cdot|)\\\asymp&\frac{1}{d}\end{aligned}$$

▶ Result

Together we have:

$$\mathfrak{M}^*\geq\left(p\max_{\pi_0,\pi_1}\mathbb{E}_{\pi_0}\left[|\theta|\right]-\mathbb{E}_{\pi_1}\left[|\theta|\right]\right)^2\left(1-d_{\mathrm{TV}}\left(\mathbb{E}_{\theta\sim\pi_1^{\otimes p}}\left[p_\theta^{\otimes p}\right]\big\|\mathbb{E}_{\theta\sim\pi_0^{\otimes p}}\left[p_\theta^{\otimes p}\right]\right)\right)\gtrsim(\frac{p}{d})^2\Theta(1)$$

with the optimal $d_p\asymp\dfrac{\log p}{\log\log p}$, we have:

$$\mathfrak{M}^*\gtrsim p^2\left(\frac{\log\log p}{\log p}\right)^2$$

3.3 Conclusion for Bounded Case

Now putting back the scaling M we have:

$$\mathfrak{M}^*\asymp M^2p^2\left(\frac{\log\log p}{\log p}\right)^2$$

In the original paper [2] they provided the sharp Minimax risk:

$$\mathfrak{M}^*(\Theta_p(M))=\beta_*^2M^2p^2\left(\frac{\log\log p}{\log p}\right)^2(1+o(1)),$$

where β_* is the Berstein's constant mentioned in subsection 2.1.

4 Upgrade to Unbounded Case

Here we focus more on the spirit of how to upgrade the bounded case to unbounded case, and we will not go through the details of the proof.

The idea is that: For small θ, we apply the preceeding polynomial approximation estimator, while for large θ, we can safely use $|X|$ as an estimator of $|\theta|$. The cut-off between the two case is estimated by the maximization of normal variable: For $X_1, \dots, X_p \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, w.h.p.

$$\max_{i=1}^{p} |X_i| \le 2\sqrt{\log p}.$$

4.1 Upper Bound Side

For the upper bound side, we modify the "good enough" estimator to something look like $\hat{T} = \text{poly}_{d_p}(X)\mathbf{1}_{|X|\le 2\sqrt{\log p}} + |X|\,\mathbf{1}_{|X|>2\sqrt{\log p}}$[4]. Detail computation of bias and variance is omitted here. The final result is:

$$\text{Bias: } \left|\hat{T}(X) - |X|_1\right| \lesssim \sqrt{\log p}\frac{p}{d}$$
$$\text{Variance: } var(\hat{T}(X)) \lesssim \sqrt{p}\log p$$

which gives optimal $d_p \sim p^{1/4}$ and thus:

$$\mathfrak{M}^* \lesssim$$

References

[1] Serge Bernstein. Sur la meilleure approximation de| x| par des polynomes de degrés donnés. Acta math, 37(1):1–57, 1914.

[2] T Tony Cai and Mark G Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. 2011.

[4] To be specific, we split the data into pairs $X = \{(X_{i1}, X_{i2})\}_{i=1}^{n/2}$. In each pair, accroding to $|X_{i1}| \gtrless 2\sqrt{\log p}$, we apply $\text{poly}(X_{i2})$ or $|X_{i2}|$ in the mean value $\hat{T} = \sum_{i=1}^{n/2} \text{func}(X_{i2})$.