

Revisiting statistical principles for risk minimization

Reese Pathak

UC Berkeley, Department of Statistics

Cornell University, School of Operations Research and Industrial Engineering (ORIE)

October 21, 2026

Joint work with...



Nikita Zhivotovskiy
(UC Berkeley)



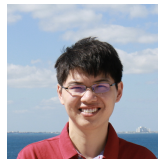
Annie Ulichney
(UC Berkeley)



Mike Jordan
(UC Berkeley & INRIA)



Liviu Aolaritei
(UC Berkeley)



Cong Ma
(University of Chicago)

Outline

1. M -estimation and some motivations
2. Gaussian sequence model and basic results
3. Failure mode: over-smoothing
4. Failure mode: heavy tails

1. M -estimation and some motivations

What is M -estimation?

M -ESTIMATION

given class of models Θ , observed data $Z \in \mathcal{Z}$, loss function $\Psi: \Theta \times \mathcal{Z} \rightarrow \mathbf{R}$, pick:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \Psi(\theta, Z)$$

- encompasses maximum likelihood estimation (MLE), minimum contrast estimation, (regularized) empirical risk minimization (ERM), *etc.*
- supported by classical *large-sample theory*:
 - Le Cam-Hájek: local asymptotic minimaxity for regular parametric models
 - Asymptotic nonparametric/semiparametric efficiency
 - Minimax analyses for estimators of interest (*e.g.*, Lasso, penalized methods, *etc.*)
- however, such results ignore several important issues:
 - finite n
 - complexity (or lack of complexity) from irregular (or very regular) covariate laws
 - problems where the SNR is (sufficiently) high
 - effects at or near the boundary of the constraint set

Issues with M -estimation

L. Le Cam, 1990:

Maximum Likelihood: An Introduction

L. Le Cam

Department of Statistics, University of California, Berkeley, California 94720, USA

Summary

Maximum likelihood estimates are reported to be best under all circumstances. Yet there are numerous simple examples where they plainly misbehave. One gives some examples for problems that had not been invented for the purpose of annoying maximum likelihood fans. Another example, imitated from Bahadur, has been specially created with just such a purpose in mind. Next, we present a list of principles leading to the construction of good estimates. The main principle says that one should not believe in principles but study each problem for its own sake.

other examples:

- *Birgé-Massart '93, Devroye-Lugosi '01*: suboptimality of MLE for density estimation
- *Kur-Rakhlin-Guntuboyina '20*: suboptimality of LSE in support function estimation
- *Kur-Gao-Guntuboyina-Sen '24*: suboptimality of LSE in convex regression ($d \geq 5$)
- *P.-Ma '24*: suboptimality of Lasso in ill-conditioned sparse regression

Reconciling with empirical success of ML/DL/LLMs?

falls *completely out of scope* of classical “fixed sample, fixed loss, fixed model class” theory!

3.1.2 Determining the Data Mix

To obtain a high-quality language model, it is essential to carefully determine the proportion of different data sources in the pre-training data mix. Our main tools in determining this data mix are knowledge classification and scaling law experiments.

Knowledge classification. We develop a classifier to categorize the types of information contained in our web data to more effectively determine a data mix. We use this classifier to downsample data categories that are over-represented on the web, for example, arts and entertainment.

Scaling laws for data mix. To determine the best data mix, we perform scaling law experiments in which we train several small models on a data mix and use that to predict the performance of a large model on that mix (see Section 3.2.1). We repeat this process multiple times for different data mixes to select a new data mix

(Llama 3, '24)

2.3.3 Rejection Sampling and Supervised Fine-Tuning

When reasoning-oriented RL converges, we utilize the resulting checkpoint to collect SFT (Supervised Fine-Tuning) data for the subsequent round. Unlike the initial cold-start data, which primarily focuses on reasoning, this stage incorporates data from other domains to enhance the model's capabilities in writing, role-playing, and other general-purpose tasks. Specifically, we generate the data and fine-tune the model as described below.

Reasoning data

We curate reasoning prompts and generate reasoning trajectories by performing rejection sampling from the checkpoint from the above RL training. In the previous stage, we only included data that could be evaluated using rule-

(DeepSeek-R1, '25)

more on source-level features, avoiding unnecessary semantic complexity.

In Qwen2.5-Coder, we applied this process iteratively. As shown in Figure 1, each iteration resulted in improvement for Qwen2.5-Coder-1.5B. Through 4-stage filtering, the average scores on HumanEval and MBPP increased from 41.6% to 46.8% compared to the baseline, demonstrating the value of high-quality Text-Code Grounding Data for code generation.

Synthetic Data Synthetic data offers a promising way to address the anticipated scarcity of training data. We used CodeQwen1.5, the predecessor of Qwen2.5-Coder, to generate large-scale synthetic datasets. To mitigate the risk of hallucinations during this process, we introduced an executor for validation, ensuring that only executable code was retained.

(Qwen2.5-Coder, '24)

other examples:

optimizer tuning, curriculum training, test-time training/compute, RLHF, distillation, ...

2. Gaussian sequence model and basic results

Gaussian sequence model and minimax risk

observational model: for a nonempty constraint set $\Theta \subset \mathbf{R}^d$, observe

$$Y = \theta^* + \xi, \quad \xi \sim \mathbf{N}(0, I_d), \quad \theta^* \in \Theta$$

minimax L^q risk: for $1 \leq q < \infty$, define

$$\mathcal{M}_q(\Theta) := \inf_{\hat{\theta}} \sup_{\theta^* \in \Theta} \left(\mathbf{E}_{\theta^*} \|\hat{\theta}(Y) - \theta^*\|_2^q \right)^{1/q}$$

main focus of this talk: for $q = 2$, for which we write

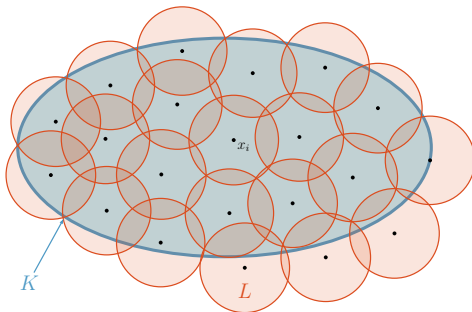
$$\mathcal{M}(\Theta) := \mathcal{M}_2(\Theta)$$

(thus $\mathcal{M}(\Theta)$, $\mathcal{M}(\Theta)^2$ are the minimax RMSE and MSE)

Metric entropy

covering number: for bounded $K \subset \mathbf{R}^d$ and compact $L \subset \mathbf{R}^d$ with nonempty interior,

$$N(K, L) := \min \left\{ m \geq 1 : \exists x_1, \dots, x_m \in \mathbf{R}^d, \quad K \subset \bigcup_{i=1}^m (x_i + L) \right\}$$



metric entropy: the logarithm of the covering number, $\log N(K, L)$

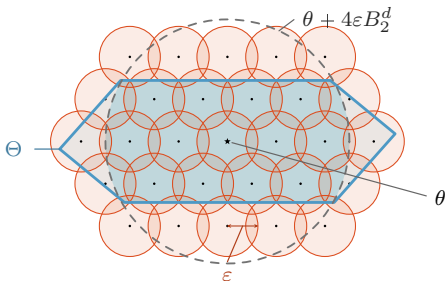
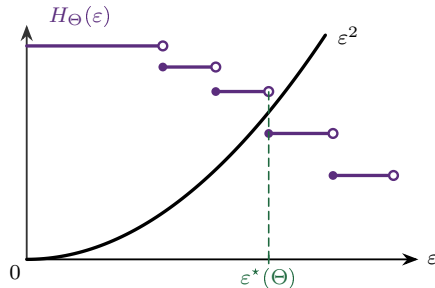
The critical radius

LOCAL METRIC ENTROPY

for a nonempty convex set $\Theta \subset \mathbb{R}^d$, define

$$H_{\Theta}(\varepsilon) := \sup_{\theta \in \Theta} \log N\left(\Theta \cap (\theta + 4\varepsilon B_2^d), \varepsilon B_2^d\right), \quad \text{and} \quad \varepsilon^*(\Theta) := \sup\{\varepsilon > 0 \mid H_{\Theta}(\varepsilon) \geq \varepsilon^2\}$$

illustration:



Metric characterization of the minimax risk for convex sets

critical radius: for nonempty convex $\Theta \subset \mathbf{R}^d$,

$$\varepsilon^*(\Theta) := \sup \left\{ \varepsilon > 0 \mid \sup_{\theta \in \Theta} \log N(\Theta \cap (\theta + 4\varepsilon B_2^d), \varepsilon B_2^d) \geq \varepsilon^2 \right\},$$

THEOREM (P.-ZHIVOTOVSKIY '26)

for every nonempty convex set $\Theta \subset \mathbf{R}^d$,

$$\mathcal{M}(\Theta) \simeq \varepsilon^*(\Theta)$$

more generally, for every $1 \leq q < \infty$,

$$\mathcal{M}_q(\Theta) \simeq \varepsilon^*(\Theta) + (\text{diam}(\Theta) \wedge \sqrt{q})$$

comments:

- based on ideas in *Le Cam '73, Birgé '83, Yang-Barron '99*
- paper of *Neykov '23* establishes a similar result for $q = 2$

Reduction from n samples and general noise levels

n -sample model:

- observe independent

$$Y_i \sim \mathcal{N}(\theta^*, \sigma^2 I_d), \quad \theta^* \in \Theta, \quad i = 1, \dots, n, \quad \sigma > 0$$

- for $1 \leq q < \infty$, define

$$\mathcal{M}_q^{(n)}(\Theta, \sigma) := \inf_{\hat{\theta}} \sup_{\theta^* \in \Theta} \left(\mathbf{E}_{\theta^*} \|\hat{\theta}(Y_1, \dots, Y_n) - \theta^*\|_2^q \right)^{1/q}$$

rescaling relation via sufficient statistics:

- with $\tau_n = \sigma/\sqrt{n}$, the sample mean satisfies $\bar{Y} \sim \mathcal{N}(\theta^*, \tau_n^2 I_d)$
- rescaling gives $\bar{Y}/\tau_n \sim \mathcal{N}(\theta^*/\tau_n, I_d)$
- the minimax risks satisfy

$$\mathcal{M}_q^{(n)}(\Theta, \sigma) = \tau_n \mathcal{M}_q^{(1)}(\tau_n^{-1}\Theta, 1) = \tau_n \mathcal{M}_q(\tau_n^{-1}\Theta)$$

3. Failure mode: over-smoothing

Gaussian sequence model under ℓ_p constraints

observational model: suppose we are given IID observations

$$Y_i \sim \mathcal{N}(\theta^*, \sigma^2 I_d), \quad \text{where} \quad \|\theta^*\|_p \leq r, \quad \text{for } i = 1, \dots, n$$

minimax (root) mean squared error: seek to estimate θ^* to optimal MSE

$$\mathcal{M}^{(n)}(rB_p^d, \sigma) := \inf_{\hat{\theta}} \sup_{\|\theta^*\|_p \leq r} \left(\mathbf{E} [\|\hat{\theta}(Y_1, \dots, Y_n) - \theta^*\|_2^2] \right)^{1/2}$$

(the order of the minimax rate above is known, Donoho-Johnstone '94)

QUESTION:

does the **maximum likelihood estimator (MLE)**, given below, achieve this rate?

$$\Pi(Y_1, \dots, Y_n) := \arg \min_{\|\vartheta\|_p \leq r} \left\{ \frac{1}{n} \sum_{i=1}^n \|Y_i - \vartheta\|_2^2 \right\}$$

(without loss, we take $n = 1, r = 1$; only interesting parameters are σ, p, d)

Behavior of the minimax risk

1994 result of Donoho & Johnstone: established two main results,

- (asymptotic) minimax RMSE for all $p \in (0, \infty)$
- dichotomy in the behavior of *linear* estimates

behavior of the minimax RMSE when $p \geq 2$:

$$\mathcal{M}(B_p^d, \sigma) \simeq \min\{\sigma\sqrt{d}, d^{1/2-1/p}\}$$

behavior of the minimax RMSE when $p < 2$:

$$\mathcal{M}(B_p^d, \sigma) \simeq \begin{cases} 1 & \text{when } \sigma \geq \frac{1}{\sqrt{1+\log d}} \\ [\sigma^2 \log(ed\sigma^p)]^{1/2-p/4} & \text{when } \sigma \in [\frac{1}{d^{1/p}}, \frac{1}{\sqrt{1+\log d}}] \\ \sigma\sqrt{d} & \text{when } \sigma \leq \frac{1}{d^{1/p}} \end{cases}$$

- “*moderate*” *noise regime* requires non-trivial procedure for rate-optimality...
- result above can be directly established via metric entropy of B_p^d

Donoho-Johnstone “nonlinearity phenomenon”

(terminology taken from Johnstone’s monograph)

Minimax risk over l_p -balls for l_q -error

David L. Donoho, Iain M. Johnstone

Department of Statistics, Stanford University, Stanford, CA 94305, USA

Received: 23 June 1992/In revised form: 13 November 1993

Summary. Consider estimating the mean vector θ from data $N_n(\theta, \sigma^2 I)$ with l_q norm loss, $q \geq 1$, when θ is known to lie in an n -dimensional l_p ball, $p \in (0, \infty)$. For large n , the ratio of minimax linear risk to minimax risk can be arbitrarily large if $p < q$. Obvious exceptions aside, the limiting ratio equals 1 only if $p = q = 2$. Our arguments are mostly indirect, involving a reduction to a univariate Bayes minimax problem. When $p < q$, simple non-linear co-ordinatewise threshold rules are asymptotically minimax at small signal-to-noise ratios, and within a bounded factor of asymptotic minimaxity in general. We also give asymptotic evaluations of the minimax linear risk. Our results are basic to a theory of estimation in Besov spaces using wavelet bases (to appear elsewhere).

Mathematics Subject Classification (1985): 62C20, 62F12, 62G20

(appeared in PTRF in 1994)

Donoho-Johnstone “nonlinearity phenomenon”

linear minimax RMSE: usual definition, but now restricted to linear estimators $Y \mapsto \hat{\theta}(Y)$,

$$\mathcal{M}^{\text{lin}}(B_p^d, \sigma) := \inf_{\hat{\theta} \text{ linear}} \sup_{\|\theta^*\|_p \leq 1} \left(\mathbf{E} [\|\hat{\theta}(Y) - \theta^*\|_2^2] \right)^{1/2}$$

failure of linear estimates:

THEOREM (DONOHO & JOHNSTONE '94)

fix $p \in (0, 2)$ and take $\sigma \equiv \sigma(d)$ such that

$$\frac{1}{d^{1/p}} \ll \sigma \ll \frac{1}{\sqrt{\log d}}$$

then, linear estimates are arbitrarily suboptimal as $d \rightarrow \infty$:

$$\lim_{d \rightarrow \infty} \frac{\mathcal{M}^{\text{lin}}(B_p^d, \sigma)}{\mathcal{M}(B_p^d, \sigma)} = \infty$$

Main result: good news

worst-case RMSE of the MLE:

$$\mathcal{R}^{\text{MLE}}(B_p^d, \sigma) := \sup_{\|\theta^*\|_p \leq 1} \left(\mathbf{E} \left[\|\Pi_{B_p^d}(Y) - \theta^*\|_2^2 \right] \right)^{1/2}$$

THEOREM (AOLARITEI, JORDAN, P., ULICHNEY '25)

the MLE is minimax rate-optimal in the following two settings:

- (i) if $p \in [0, 1 + \frac{1}{1+\log d}] \cup [2, \infty]$, for any $\sigma > 0$; or
- (ii) if $p \in [1 + \frac{1}{1+\log d}, 2]$, and for $\sigma \in [0, \frac{1}{d^{1/p}}] \cup [\frac{1}{\sqrt{1+\log d}}, \infty]$.

indeed, in the above two settings, we have

$$\mathcal{R}^{\text{MLE}}(B_p^d, \sigma) \simeq \mathcal{M}(B_p^d, \sigma)$$

Main result: bad news

THEOREM (AOLARITEI, JORDAN, P., ULICHNEY '25)

fix $p \in (1, 2)$, and suppose $\sigma = \sigma(d)$ satisfies

$$\frac{1}{d^{1/p}} \ll \sigma \ll \frac{1}{\sqrt{\log d}}$$

then, the MLE is minimax rate-suboptimal by an arbitrarily large factor:

$$\lim_{d \rightarrow \infty} \frac{\mathcal{R}^{\text{MLE}}(B_p^d, \sigma)}{\mathcal{M}(B_p^d, \sigma)} = \infty$$

(simplified) summary:

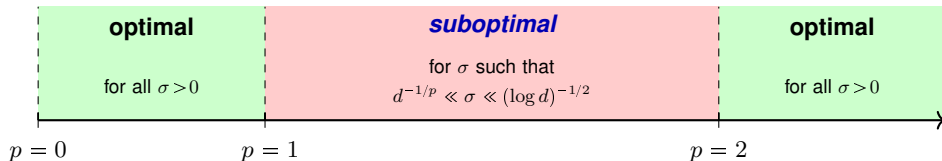
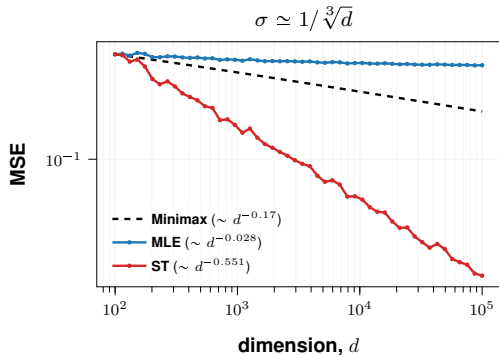
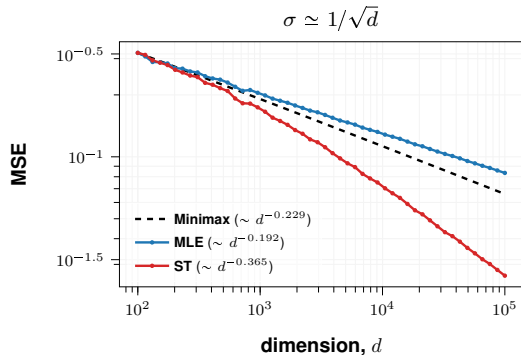


Illustration of suboptimality, $p = 1.5$



comments:

- comparison above with rate-optimal estimator based on soft-thresholding (ST):

$$\hat{\theta}^{\text{ST}}(Y) := \arg \min_{\vartheta \in \mathbf{R}^d} \left\{ \|\vartheta - Y\|_2^2 + 2\lambda \|\vartheta\|_1 \right\} \quad \text{where} \quad \lambda = \sqrt{2\sigma^2 \log(ed\sigma^p)}$$

- in paper, we establish two phases in asymptotics of sup-risk of the MLE
 - non-vanishing error: $\frac{1}{d^{1-1/p}} \ll \sigma \ll \frac{1}{\sqrt{\log d}}$
 - vanishing error: $\frac{1}{d^{1/p}} \ll \sigma \ll \frac{1}{d^{1-1/p}}$

4. Failure mode: heavy tails

Reduction to the L^q norm of projections

least squares estimator (LSE)/maximum likelihood estimator (MLE):

- for $\Theta \subset \mathbf{R}^d$ closed and convex, observe:

$$Y = \theta + \sigma\xi, \quad \xi \sim \mathbf{N}(0, I_d), \quad \theta \in \Theta \subset \mathbf{R}^d$$

- the LSE is $\Pi_{\Theta}(Y) = \arg \min_{\vartheta \in \Theta} \|Y - \vartheta\|_2$
- *translation and scaling*: it holds that: $\hat{\theta}(\theta + \sigma\xi) - \theta = \sigma \Pi_{(\Theta - \theta)/\sigma}(\xi)$
- hence, risk satisfies

$$\left(\mathbf{E} \|\hat{\theta}(Y) - \theta\|_2^q \right)^{1/q} = \sigma \left(\mathbf{E} \|\Pi_{(\Theta - \theta)/\sigma}(\xi)\|_2^q \right)^{1/q}$$

UPSHOT:

tails of risk of LSE for closed, convex sets



L^q norms of projections onto closed, convex sets containing the origin

Formulation as an extremal problem

optimal $L^{q_1} - L^{q_0}$ inequality:

- consider class of convex sets

$$\mathcal{C}_d = \{C \subset \mathbf{R}^d : C \text{ closed, convex, and } 0 \in C\}$$

- for $q_0, q_1 \in [1, \infty)$ with $q_0 < q_1$, define

$$A_d^*(q_0, q_1) := \sup_{\substack{C \in \mathcal{C}_d \\ C \neq \{0\}}} \frac{(\mathbf{E} \|\Pi_C(\xi)\|_2^{q_1})^{1/q_1}}{(\mathbf{E} \|\Pi_C(\xi)\|_2^{q_0})^{1/q_0}}, \quad \text{where } \xi \sim \mathbf{N}(0, I_d)$$

QUESTION:

is $A_d^*(q_0, q_1)$ small if q_1 is comparable to q_0 ?

Main result: No!

THEOREM (P. '26)

for $d \geq 1$ and $q_0, q_1 \in [1, \infty)$ with $q_0 < q_1$, it holds that

$$A_d^*(q_0, q_1) \simeq \begin{cases} \left(\frac{d}{q_0 \log(ed/q_0)} \right)^\tau, & q_0 \leq d, \ q_0 < q_1 \leq q_0 \log \frac{ed}{q_0}, \\ \sqrt{\frac{d}{q_0 \log(ed/q_1)}}, & q_0 \leq d, \ q_0 \log \frac{ed}{q_0} < q_1 \leq d, \\ \sqrt{\frac{q_1}{q_0}}, & q_1 > d \end{cases}$$

$$\text{where } \tau = \frac{1}{2} \left(1 - \frac{q_0}{q_1} \right).$$

- *e.g.*, for $q_0 = 1, q_1 = 2$,

$$A_d^*(1, 2) \simeq \left(\frac{d}{1 + \log d} \right)^{1/4}$$

Qualitative behavior of projections and supporting vectors

behavior on dilates:

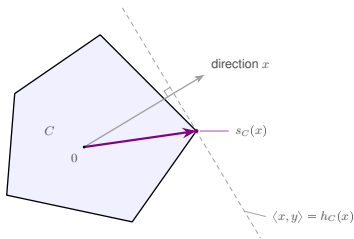
- for compact $C \in \mathcal{C}_d$,

$$\|\Pi_{\lambda C}\|_{L^p} = (1 + o(1)) \lambda \|s_C\|_{L^p} \quad \text{as } \lambda \rightarrow 0^+$$

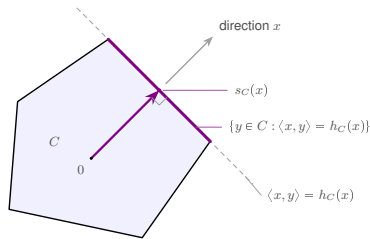
- above, s_C is the supporting vector, defined by

$$s_C(x) := \arg \min \left\{ \|y\|_2 : y \in C, \langle x, y \rangle = h_C(x) \right\}, \quad h_C(x) := \sup_{z \in C} \langle z, x \rangle$$

two examples:



unique maximizer



multiple maximizers

Proof idea: extremal sets

bad sets:

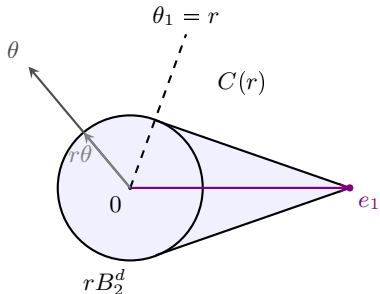
$$C(r) = \text{conv}(rB_2^d, e_1), \quad r \in (0, 1]$$

The supporting vector satisfies

$$\|s_{C(r)}(\theta)\|_2 = \begin{cases} 1, & \theta_1 > r, \\ r, & \theta_1 \leq r, \end{cases} \quad \theta \in \mathbf{S}^{d-1}$$

With previous limit relation:

$$\frac{\|\Pi_{\lambda C(r)}\|_{L^{q_1}}}{\|\Pi_{\lambda C(r)}\|_{L^{q_0}}} \longrightarrow \frac{\|s_{C(r)}\|_{L^{q_1}}}{\|s_{C(r)}\|_{L^{q_0}}} \quad (\lambda \rightarrow 0^+)$$



intuition:

- “most” directions have supporting vector $r\theta$, except on cap $\{\theta_1 > r\}$, we get spike e_1
- pick mass of cap to dominate the L^{q_1} norm but be negligible for the L^{q_0} norm
- up to a logarithm in d , this gives the lower bound

Corollary: suboptimality of the LSE

- for closed convex $\Theta \subset \mathbf{R}^d$, define

$$\mathcal{M}(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \left(\mathbf{E}_{\xi \sim N(0, I_d)} \|\hat{\theta}(\theta + \xi) - \theta\|_2^2 \right)^{1/2}$$

- for sufficiently large d , take

$$r_d = \left(\frac{1 + \log d}{d} \right)^{1/4}, \quad \lambda_d = \sqrt{d} r_d$$

$$\Theta_d = \lambda_d C(r_d) = \sqrt{d} \mathbf{conv}(r_d^2 B_2^d, r_d e_1)$$

- by similar argument to previous theorem, have:

$$\mathcal{M}(\Theta_d) \leq \sqrt{1 + dr_d^4}, \quad \left(\mathbf{E} \|\Pi_{\Theta_d}(\sqrt{d} r_d e_1 + \xi) - \sqrt{d} r_d e_1\|_2^2 \right)^{1/2} \gtrsim \sqrt{d} r_d$$

- therefore, the LSE/MLE satisfies

$$\frac{\sup_{\theta \in \Theta_d} \left(\mathbf{E} \|\Pi_{\Theta_d}(\theta + \xi) - \theta\|_2^2 \right)^{1/2}}{\mathcal{M}(\Theta_d)} \gtrsim \sqrt{\frac{dr_d^2}{1 + dr_d^4}} \simeq \sqrt[4]{\frac{d}{1 + \log d}}$$

Summary

1. several issues can arise in risk minimization, here we focused on *Gaussian sequence model*
2. over ℓ_p balls, minimizing square loss (i.e., the MLE) can be *rate-suboptimal due to over-smoothing*, when $p \in (1, 2)$, at moderate noise
3. even under Gaussian noise and convex constraints, *adjacent moments of the risk* of the MLE can blow up!
4. yields a convex constraint set for which the RMSE of the MLE is larger than the optimal (minimax) RMSE by a factor $\gtrsim \sqrt[4]{d/(1 + \log d)}$
5. in contrast, there's an (inefficient) rate-optimal method (simultaneously for all L^q)!

OPEN QUESTION

for every convex constraint set, is there an estimator which is both *statistically optimal* and *computationally efficient*?

Thank you!

based on the following papers:

arxiv.org/abs/2402.00382 (Biometrika, 2026+)

arxiv.org/abs/2506.10354 (Annals of Statistics, 2026+)

arxiv.org/abs/2603.02714 (2026)

arxiv.org/abs/2607.05117 (2026)