Information Theoretic Limits of Robust Sub-Gaussian Mean Estimation Under Star-Shaped Constraints

Matey Neykov

Department of Statistics and Data Science
Northwestern

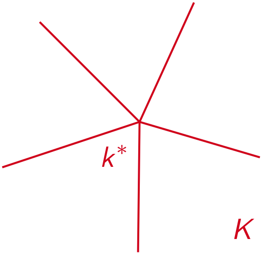
Joint Work With

Akshay Prasadan (CMU)

What We Wish We Observed

$$\tilde{X}_i = \mu + \xi_i,\ i \in [N], \quad \xi \sim \mathrm{SG}(\sigma^2), \quad \mathbb{E}\xi = 0$$
$$\mu \in K \subseteq \mathbb{R}^n, K \text{ is known and star-shaped.}$$

$$\mathbb{E}\xi = 0, \xi \sim \mathrm{SG}(\sigma^2):$$
$$\sup_{v \in S^{n-1}} \mathbb{E}e^{\lambda v^{\mathrm{T}}\xi} \le e^{\lambda^2\sigma^2/2}$$



$$\forall x \in K,\ \forall \alpha \in [0,1] \Rightarrow$$
$$\alpha k^* + (1-\alpha)x \in K$$

All Powerful Adversary

- Corrupts $\leq$ to $\epsilon < 1/2$ fraction of the N observations

What We Actually Observe

We observe $X_i = \mathcal{C}(\tilde{X}_i), i \in [N]$,
$\mathcal{C}(\tilde{X}_i) = \tilde{X}_i$ for $\geq (1-\epsilon)N$ observations,
but can be arbitrary on the rest!

- Compare and contrast to Huber contamination model $X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon)P_\mu + \epsilon Q$
- In the adversarial model X_i even the "good" samples are non i.i.d!
- We have guaranteed bounded number of outliers
- In Huber model risk is infinite on unbounded sets [Bateni and Dalalyan, 2020]

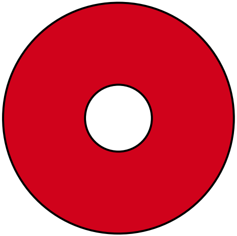
Relevant Literature

[Chen et al., 2018]
[Lugosi and Mendelson, 2021], [Neykov, 2022],
[Diakonikolas et al., 2022]
[Diakonikolas et al., 2019, Diakonikolas et al., 2017]

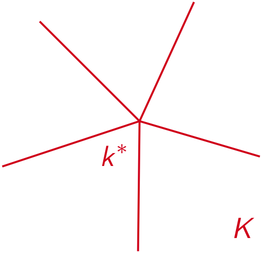
- There are (too) many relevant papers to fit on one slide
- Unconstrained setting
- Error bounds with high probability rather than expectation,
- Sample sizes required sufficiently large;
- Non-matching lower and upper bounds,
- A non-adversarial Huber contamination model,
- Distinct distributional assumptions on the noise term.

Examples of Star-Shaped Sets

- K — all $\leq s$-sparse vectors in $\mathbb{R}^n$ for some $s \leq n$.
- Any convex set!



Non star-shaped set!



$\forall x \in K, \forall \alpha \in [0,1] \Rightarrow$
$\alpha k^* + (1-\alpha)x \in K$

Outline of the Talk

1. Entropic Characterization of the Minimax Rate
2. Upper Bound Details
3. Examples

Minimax Rate

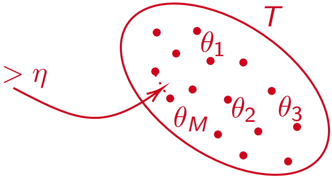
- Known (or symmetric) Noise:

$$\inf_{\hat{\mu}} \sup_{\mu \in K} \sup_{\mathcal{C}} \mathbb{E}\|\hat{\mu}(X) - \mu\|^2$$

- Unknown Noise:

$$\inf_{\hat{\mu}} \sup_{\mu \in K} \sup_{\xi \sim \mathsf{SG}(\sigma^2)} \sup_{\mathcal{C}} \mathbb{E}\|\hat{\mu}(X) - \mu\|^2$$

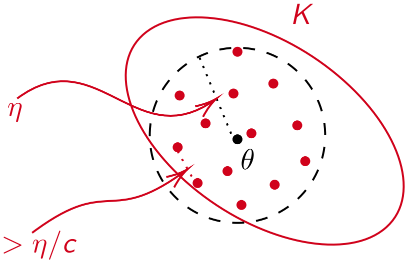
Global Entropy



Definition (Global Entropy)

For a set $T \subset \mathbb{R}^n$, a set $\theta_1, \theta_2, \ldots, \theta_M \in T$ is called a packing set if $\|\theta_i - \theta_j\| > \eta$ for all $i \neq j$. The η packing number is the cardinality of the maximal packing set. The log of that packing number is called (global) entropy.

Local Entropy



Definition (Local Entropy)

Let $\theta \in K$ be a point. Consider the set $B(\theta, \eta) \cap K$. Let $M(\eta/c, B(\theta, \eta) \cap K)$ denote the largest cardinality of an η/c packing set in $B(\theta, \eta) \cap K$. Let

$$\log M_K^{\text{loc}}(\eta, c) := \sup_{\theta \in K} \log M(\eta/c, B(\theta, \eta) \cap K).$$

A Fact for Local Entropy

For star-shaped sets the map $\eta \mapsto \log M_K^{\text{loc}}(\eta, c)$ is non-increasing!

Known or Symmetric Noise Minimax Rate

Theorem (A. Prasadan and N. (2024))

We have (for sufficiently large c) and any $\epsilon < c_0 < 1/2$

$$\inf_{\hat{\mu}} \sup_{\mu \in K} \sup_{\mathcal{C}} \mathbb{E}\|\hat{\mu}(X) - \mu\|^2 \asymp \max(\eta^{*2}, \sigma^2\epsilon^2) \wedge d^2$$

where $d = \operatorname{diam}(K)$ ($d = \infty$ if K is unbounded), and η^* solves the entropic equation

$$\eta^* = \sup\left\{\eta : \frac{N\eta^2}{\sigma^2} \leq \log M^{\mathrm{loc}}(\eta, c)\right\},$$

- $\eta^* \wedge d \gtrsim \sigma/\sqrt{N} \wedge d$ so that when $\epsilon < 1/\sqrt{N}$ outliers do not affect the rate!

Unknown Noise Minimax Rate

Theorem (A. Prasadan and N. (2024))

We have (for sufficiently large c) and any $\epsilon < 1/16$

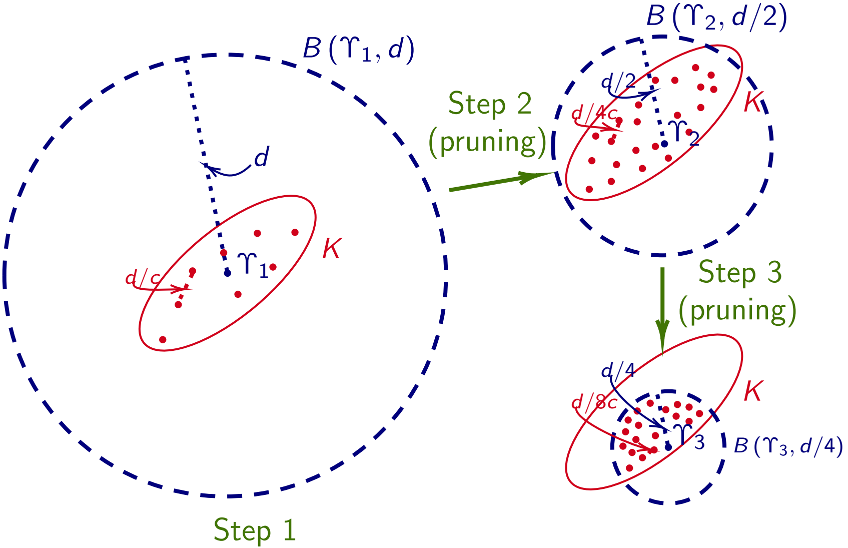
$$\inf_{\hat{\mu}} \sup_{\mu \in K} \sup_{\xi \sim \mathrm{SG}(\sigma^2)} \sup_{\mathcal{C}} \mathbb{E}\|\hat{\mu}(X) - \mu\|^2 \asymp \max(\eta^{*2}, \sigma^2\epsilon^2 \log(1/\epsilon)) \wedge d^2$$

where $d = \mathrm{diam}(K)$ ($d = \infty$ if K is unbounded), and η^* solves the entropic equation

$$\eta^* = \sup\left\{\eta : \frac{N\eta^2}{\sigma^2} \le \log M^{\mathrm{loc}}(\eta, c)\right\},$$

Outline of the Talk

Algorithm (Directed Tree Construction)



Algorithm (Comparison Between Two Points)

Definition

Given an ordered pair (ν_1, ν_2) of points $\nu_1, \nu_2 \in \mathbb{R}^n$, define the test ψ_{ν_1,ν_2} by

$$\psi_{\nu_1,\nu_2}(\{X_i\}_{i\in[N]}) = \mathbb{1}(|\{i \in [N] : \|X_i - \nu_1\| \geq \|X_i - \nu_2\|\}| \geq N/2).$$

We drop the subscripts and write ψ when the context is clear.

Definition

Assume points ν_1 and ν_2 are in lexicographic order.
If $\psi_{\nu_1,\nu_2}(\{X_i\}_{i\in[N]}) = 0$, then $\nu_1 \succ \nu_2$ (or $\nu_2 \prec \nu_1$).
If $\psi_{\nu_1,\nu_2}(\{X_i\}_{i\in[N]}) = 1$ then $\nu_2 \succ \nu_1$ (or $\nu_1 \prec \nu_2$).

Algorithm (Tournament)

At any point ν, given a radius $\delta > 0$ and finite set $S \subset K$, define

$$T(\delta, \nu, S) = \begin{cases} \max_{\nu' \in S} \|\nu - \nu'\| \text{ if } \nu \prec \nu' \text{ and } \|\nu - \nu'\| \geq (c/2 - 1)\delta \\ 0 \text{ otherwise.} \end{cases}$$

Algorithm 1: Robust Upper Bound Algorithm

Input: A point $\Upsilon_1 \in K$

$k \leftarrow 1$;

$\Upsilon \leftarrow [\Upsilon_1]$;

while *TRUE* **do**

  $\Upsilon_{k+1} \leftarrow \operatorname{argmin}_{\nu \in \mathcal{O}(\Upsilon_k)} T\left(\frac{d}{2^{k-1}c}, \nu, \mathcal{O}(\Upsilon_k)\right)$

  Υ.append(Υ_{k+1});

  $k \leftarrow k + 1$;

return $\Upsilon = [\Upsilon_1, \Upsilon_2, \ldots]$

Algorithm (Some Omitted Details)

- We actually add a $R_i \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ variable to each observation
- When noise is unknown we change the def of $\prec$ and $\succ$:

$$\psi_{\nu_1,\nu_2}(\{X_i\}_{i=1}^{2N}) = \begin{cases} \mathbb{1}(\mathrm{TM}(\{V_i\}_{i=1}^{2N}) > 0) & \text{if } \frac{\delta^2}{\sigma^2} \leq C \\ \mathbb{1}(|\{i \in [2N] : \|X_i + R_i - \nu_1\| \geq \|X_i + R_i - \nu_2\|\}| \geq N) & \text{if } \frac{\delta^2}{\sigma^2} > C \end{cases}$$

$$\text{where } V_i = \|X_i + R_i - \nu_1\|^2 - \|X_i + R_i - \nu_2\|^2$$

- In the unbounded K case we first trap μ in a bounded set with high probability
- Then kind of reuse previous results for bounded sets

Outline of the Talk

Example 1

- $K = \mathbb{R}^n$
- $\log M^{\mathrm{loc}}(\eta, c) \asymp n$
- Hence $\eta^{*2} \asymp n\sigma^2/N$ and the rate is
- $\max(n\sigma^2/N, \sigma^2\epsilon^2)$ or $\max(n\sigma^2/N, \sigma^2\epsilon^2 \log(1/\epsilon))$

Example 2

- $K = s$-sparse vectors
- Lemma: $\log M^{\text{loc}}(\eta, c) \asymp s \log(1 + n/s)$
- Hence $\eta^{*2} \asymp s \log(1 + n/s)\sigma^2/N$ and the rate is
- $\max(s \log(1 + n/s)\sigma^2/N, \sigma^2\epsilon^2)$ or
  $\max(s \log(1 + n/s)\sigma^2/N, \sigma^2\epsilon^2 \log(1/\epsilon))$

Food for thought

- Clearly, there are many more examples like ℓ_p bodies for $p \in [1, \infty]$ and even $p < 1$
- Even the case with $N = 1, \epsilon = 0$ is interesting!
- E.g. K is a d-dimensional subspace – linear regression
- Or $K = \{(f(x_1), f(x_2), \dots, f(x_n)) | f \in \mathcal{F}\}$ with x_i being fixed design points – nonparametric regression with fixed design

Thanks!

Thank You!

Bateni, A.-H. and Dalalyan, A. S. (2020).
Confidence regions and minimax rates in outlier-robust estimation on the probability simplex.
Electron. J. Statist., 14(2):2653–2677.

Chen, M., Gao, C., and Ren, Z. (2018).
Robust covariance and scatter matrix estimation under huber?s contamination model.
The Annals of Statistics, 46(5):1932–1960.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Moitra, A., and Stewart, A. (2019).
Robust estimators in high-dimensions without the computational intractability.
SIAM Journal on Computing, 48(2):742–864.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Moitra, A., and Stewart, A. (2017).
Being robust (in high dimensions) can be practical.

In *International Conference on Machine Learning*, pages 999–1008. PMLR.

Diakonikolas, I., Kane, D. M., Karmalkar, S., Pensia, A., and Pittas, T. (2022).
Robust sparse mean estimation via sum of squares.
In *Conference on Learning Theory*, pages 4703–4763. PMLR.

Lugosi, G. and Mendelson, S. (2021).
Robust multivariate mean estimation: The optimality of trimmed mean.
The Annals of Statistics, 49(1):393 – 410.

Neykov, M. (2022).
On the Minimax Rate of the Gaussian Sequence Model Under Bounded Convex Constraints.
IEEE Transactions on Information Theory, 69(2):1244–1260.