Minimax lower bound for 1-bit Matrix Completion

Tri Nguyen

Reading group - Summer 2022
Oregon State University

October 6, 2022

Main References

- John Duchi. "Lecture notes for statistics 311/electrical engineering 377". In: *URL: https://stanford. edu/class/stats311/Lectures/full notes. pdf. Last visited on* 2 [2016], p. 23
- Jonathan Scarlett et al. "An introductory guide to Fano's inequality with applications in statistical estimation". In: *arXiv preprint arXiv:1901.00555* [2019]
- Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Dordrecht: Springer, 2009. DOI: 10.1007/b13794. URL: https://cds.cern.ch/record/1315296
- Mark A Davenport et al. "1-bit matrix completion". In: *Information and Inference: A Journal of the IMA* 3.3 [2014], pp. 189–223

Recap: General Setting

- From a distribution family $\mathcal{P} = \mathcal{N}_d = \left\{N(\theta, I_d) | \theta \in \mathbb{R}^d\right\}$, God chooses a distribution $P \in \mathcal{P}$.
- A set of N (i.i.d) samples X_1^N are drawn from P, denoted as $\boldsymbol{X}$.
- Task: estimating $\theta(P)$ from given samples.
- Quality of estimator $\widehat{\theta}$ is measured by $\Phi(\rho(\theta, \widehat{\theta})) = \left\|\theta - \widehat{\theta}\right\|^2$, where:
  - $\theta = \theta(P)$ is expectation of $P = N(\theta, I_d)$
  - $\widehat{\theta} = \widehat{\theta}(X_1^n)$ is the estimator of interest. Examples: $n^{-1}(\sum_{i=1}^n X_i)$, X_1.
  - $\Phi(t) = t^2$ is a non-decreasing function
  - $\rho(\theta, \widehat{\theta}) = \left\|\theta - \widehat{\theta}\right\|$ is a semimetric
- Question: What would be the best performance of an ideal estimator in the worse case?

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right]$$

Finding exact $\mathcal{M}()$ is difficult, instead our attempt is to find a lower bound of it.

Recap: General Approach to Find Lower Bound

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right]$$

- Translate to probability (Markov inequality)

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}\left[\Phi(\rho(\theta, \widehat{\theta}))\right] \geq \Phi(\delta) \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta)$$

- Reduce the whole space $\mathcal{P}$ to a finite set $\{\theta_v | v \in \mathcal{V}\}$

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\theta, \widehat{\theta}) \geq \delta) \geq \inf_{\widehat{\theta}} \max_{v} \mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta)$$

- Reduce to a hypothesis testing error by constructing 2δ-**packing set**.

$$\inf_{\widehat{\theta}} \max_{v} \mathbb{P}(\rho(\theta_v, \widehat{\theta}) \geq \delta) \geq \inf_{\Psi} \max_{v} \mathbb{P}(v \neq \Psi(\widetilde{X}_1^n)),$$

where $\Psi(\widetilde{X}_1^N) \triangleq \arg\min_v \rho(\theta_v, \widehat{\theta}(\widetilde{X}_1^N))$

Recap

- Fano's method is to switch to the average.

$$\inf_{\Psi} \max_{v} \mathbb{P}(v \neq \Psi(\widetilde{X}_1^n)) \geq \inf_{\Psi} \frac{1}{|\mathcal{V}|} \sum_{v} \mathbb{P}(v \neq \Psi(\widetilde{X}_1^N))$$
$$= \inf_{\Psi} \mathbb{P}(V \neq \Psi(\widetilde{X}_1^N)), \quad \text{where } V \text{ is a uniform RV.}$$

Lemma

For any discrete RVs V, V' on the same alphabet $\mathcal{V}$,

$$\mathbb{P}(V \neq V') \geq 1 - \frac{I(V;V') + \log 2}{\log |\mathcal{V}|}$$

where $\mathbb{P}$ is taken with respect to both V, V'.

- There are other alternatives which do not consider RV V [Tsybakov 2009].

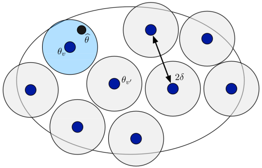
Recap: Fano's Method - The Recipe

We want to lower bound the RHS of

$$\mathbb{P}(V \neq V') \geq 1 - \frac{I(V;V') + \log 2}{\log |\mathcal{V}|}$$

by construct a **packing set** $\{\theta_v | v \in \mathcal{V}\}$, such that

- (required) $\rho(\theta_v, \theta_{v'}) \geq 2\delta \quad \forall v, v' \in \mathcal{V}$
- (desired) $|\mathcal{V}|$ is *large*
- (desired) $I(V;V')$ is *small*
- In the Gaussian mean estimation example, $|\mathcal{V}| \geq 2^d$, $I(V;V') \leq O(n\delta^2)$.



Two tasks:

- Construct packing set.
- Lower bound mutual information.

Minimax Bound in 1-bit Matrix Completion Problem

- Given matrix $\boldsymbol{M} \in K \triangleq \{\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2} \mid \|\boldsymbol{M}\|_* \leq \alpha\sqrt{r d_1 d_2}, \|\boldsymbol{M}\|_\infty \leq \alpha\}$.
- A RV $\Omega \subset [d_1] \times [d_2]$ with $\mathbb{E}[|\Omega|] = n$
- A differential function $f : \mathbb{R} \to [1, 0]$ (cdf)
- Matrix $\boldsymbol{Y}$ such that

$$Y_{ij} = \begin{cases} +1 & \text{with probability } f(M_{ij}) \\ -1 & \text{with probability } 1 - f(M_{ij}) \end{cases}$$

- Task: Estimate $\boldsymbol{M}$ given $\boldsymbol{Y}, \Omega$
- Quality measurement: $\Phi(\rho(\boldsymbol{M}, \widehat{\boldsymbol{M}})) = \frac{1}{d_1 d_2} \left\| \boldsymbol{M} - \widehat{\boldsymbol{M}} \right\|_{\mathrm{F}}^2$.

Theorem (Davenport et al. 2014)

Given a fixed algorithm, there exists $\boldsymbol{M} \in K$ such that with probability at least 0.75 (over RV $\boldsymbol{Y}$),

$$\frac{1}{d_1 d_2} \left\| \boldsymbol{M} - \widehat{\boldsymbol{M}} \right\|_{\mathrm{F}}^2 \geq \min\left(C_1, C_2 \alpha \sqrt{\beta_{0.75\alpha}} \sqrt{\frac{r \max(d_1, d_2)}{n}} \right) = O\left(\frac{1}{\sqrt{n}} \right)$$

Prove by construction!

Sketch of Proof: Step 1

- Construct a set of matrices $\mathcal{X} = \{\boldsymbol{X}_v\}$ indexed by $v \in \mathcal{V}$ such that
  - $\mathcal{X} \subseteq K$
  - $\|\boldsymbol{X}_v - \boldsymbol{X}_{v'}\|_{\mathrm{F}}^2 \geq \epsilon^2, \quad \forall v, v' \in \mathcal{V}$ for some $\epsilon > 0$.
- Uniformly choose a $V \in \mathcal{V}$, constructing $P(\cdot; \boldsymbol{X}_V)$, draw set X_1^N of N samples from that $P(\cdot; \boldsymbol{X}_V)$.
- Let ψ be the algorithm in the theorem: it uses data $\boldsymbol{Y}, \Omega$ and outputs $\widehat{\boldsymbol{M}}$.
- Let Ψ define as $\widehat{V} = \Psi((\boldsymbol{Y}, \Omega)) = \arg\min_{v \in \mathcal{V}} \rho(\widehat{\boldsymbol{M}}, \boldsymbol{X}_v)$.

By construction,

$$\mathbb{P}(V \neq \widehat{V}) = \mathbb{P}\left(\left\|\boldsymbol{X}_V - \boldsymbol{X}_{\widehat{V}}\right\|_{\mathrm{F}}^2 \geq \epsilon^2\right)$$

Hence, the remaining part is to find a bound on the best possible prediction accuracy? i.e,

$$\inf_{\Psi} \mathbb{P}(V \neq \widehat{V}) \geq \; q(\epsilon)$$

where $\mathbb{P}$ is respect to RVs $V, \boldsymbol{Y}_\Omega$.
Then we can claim that there exists $\boldsymbol{M}$, with probability at least $q(\epsilon)$,

$$\left\|\boldsymbol{M} - \widehat{\boldsymbol{M}}\right\|_{\mathrm{F}}^2 \geq \epsilon^2$$

Sketch of Proof: Step 2

- Find the lower bound of $\inf_\Psi \mathbb{P}(V \neq \widehat{V})$
  - Fano's inequality: For any discrete RVs V, V' on the same alphabet $\mathcal{V}$,

$$P(V \neq \widehat{V}) \geq 1 - \frac{I(V; \widehat{V}) + \log 2}{\log \mathcal{V}}$$

  - $I(V; \widehat{V}) \leq I(V; \boldsymbol{Y}_\Omega)$ since we have a Markov chain $V \to \boldsymbol{Y}_\Omega \to \widehat{V}$
  - Bound the $I(V; \boldsymbol{Y}_\Omega)$ [Scarlett et al. 2019]
    - Tensorization if all data points are i.i.d
    - Otherwise,

$$I(V; \boldsymbol{Y}_\Omega) \leq \max_{v,v'} D_{\mathrm{kl}}(P(\cdot|v) \,||\, P(\cdot|v'))$$

  - Upper bound that KL, which is application-dependent. Also, the ϵ should appear in this step.
- Integrating everything together, choosing ϵ to have a tight/meaningful bound.

Step 1: Construct an Attentive Packing set

Lemma

Let $\gamma \leq 1$ be such that $r/\gamma^2 \in \mathbb{N}$, and suppose that $r/\gamma^2 \leq d_1$. There is a set $\mathcal{X} \subset K$ with

$$|\mathcal{X}| \geq \exp\left(\frac{rd_2}{16\gamma^2}\right)$$

with the following properties:

- *For all $\boldsymbol{X} \in \mathcal{X}$, each entry has $|X_{ij}| = \alpha\gamma$.*
- *For all $\boldsymbol{X} \neq \boldsymbol{X}' \in \mathcal{X}$,*

$$\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}}^2 > 0.5\alpha^2\gamma^2 d_1 d_2$$

Proof of the Existence of Packing Set

It is an interesting probabilistic argument.

Consider the following distribution over random matrix $\boldsymbol{X}$ with size of $d_1 \times d_2$:

- Let $d_1' \triangleq r/\gamma^2 (\leq d_1)$.
- Matrix $\boldsymbol{X}$ contains multiple blocks of size $d_1' \times d_2$.
- For the first block, all entries are i.i.d Bernoulli RVs, i.e., $X_{ij} \sim \text{Bernoulli}(0.5), X_{ij} \in \{\pm \alpha\gamma\}, \forall (i,j) \in [d_1'] \times [d_2]$.
- For other blocks are just copies of the first block (as much as possible).

We will draw from this distribution to construct set $\mathcal{X}$ of $\left\lceil \exp\left(\frac{rd_2}{16\gamma^2}\right)\right\rceil$ elements.

Then $\mathcal{X} \subset K \triangleq \left\{\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2} \mid \|\boldsymbol{M}\|_* \leq \alpha\sqrt{rd_1d_2}, \|\boldsymbol{M}\|_\infty \leq \alpha\right\}$ since

- $\|\boldsymbol{X}\|_\infty = \alpha\gamma \leq \alpha$
- $\|\boldsymbol{X}\|_* \leq \sqrt{\text{rank}(\boldsymbol{X})}\, \|\boldsymbol{X}\|_{\text{F}} \leq \sqrt{d_1'}\, \|\boldsymbol{X}\|_{\text{F}} = \sqrt{r/\gamma^2}\sqrt{d_1d_2}\alpha\gamma = \alpha\sqrt{rd_1d_2}$

For 2 RVs $\boldsymbol{X}, \boldsymbol{Y}$ followed the above distribution,

$$\begin{aligned}
\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2 &= \sum_{i,j}(X_{ij}-Y_{ij})^2 \\
&\geq \left\lfloor \frac{d_1}{d_1'} \right\rfloor \sum_{i\in[d_1'],j\in[d_2]} (X_{ij}-Y_{ij})^2 \\
&= 4\alpha^2\gamma^2 \left\lfloor \frac{d_1}{d_1'} \right\rfloor \sum_{i\in[d_1'],j\in[d_2]} \delta_{ij}, \qquad \delta_{ij} \sim_{\text{i.i.d}} \text{Bern}(0.5), \delta_{ij} \in \{0,1\}
\end{aligned}$$

Next, with union bound and Hoeffding's inequality, we obtain,

$$\begin{aligned}
P\left(\min_{\boldsymbol{X}\neq\boldsymbol{Y}} \sum_{i\in[d_1'],j\in[d_2]} \delta_{ij} \leq 0.25 d_1' d_2\right) &\leq \sum_{\boldsymbol{X}\neq\boldsymbol{Y}} P\left(\min_{\boldsymbol{X}\neq\boldsymbol{Y}} \sum_{i\in[d_1'],j\in[d_2]} \delta_{ij} \leq 0.25 d_1' d_2\right) \\
&\leq \binom{|\mathcal{X}|}{2} \exp\left(-d_1' d_2/8\right) < 1
\end{aligned}$$

That means that there is a non-zero probability that we obtain the set $\mathcal{X}$ such that

$$\|\boldsymbol{X}-\boldsymbol{Y}\|_{\mathrm{F}}^2 \geq \alpha^2\gamma^2 \left\lfloor \frac{d_1}{d_1'} \right\rfloor d_1' d_2 \geq 0.5\alpha^2\gamma^2 d_1 d_2$$

Step 2: Apply Fano's Inequality

Let $\mathcal{X}'_{\alpha/2,\gamma}$ be the set constructed in the previous Lemma. Construct $\mathcal{X}$ as

$$\mathcal{X} \triangleq \left\{ \boldsymbol{X}' + \alpha \left(1 - \frac{\gamma}{2}\right) \mathbf{1} \mid \boldsymbol{X}' \in \mathcal{X}'_{\alpha/2,\gamma} \right\},$$

where γ is chosen as

$$4\sqrt{2}\epsilon/\alpha \leq \gamma \leq 8\epsilon/\alpha,$$

and ϵ is chosen such that

$$\|\boldsymbol{X} - \boldsymbol{X}'\|_{\mathrm{F}}^2 \leq 4 d_1 d_2 \epsilon^2$$

By construction, $\mathcal{X} \subset K$ (not obvious but easy to show), and $|\mathcal{X}| = \left|\mathcal{X}'_{\alpha/2,\gamma}\right|$.

Fano's Inequality

Now we show that if we choose $\boldsymbol{M} \in \mathcal{X}$ uniformly

$$P(V \neq \widehat{V}) \geq 1 - \frac{\max_{v,v'} \; D_{\mathrm{kl}}(P(\cdot|v) \;||\; P(\cdot|v')) + \log 2}{\log |\mathcal{V}|}$$

By property of KL divergence of product distributions,

$$\max_{v,v' \in \mathcal{V}} D_{\mathrm{kl}}(\boldsymbol{Y}_\Omega | v \;||\; \boldsymbol{Y}_\Omega | v') = \max_{v,v' \in \mathcal{V}} \sum_{(i,j) \in \Omega} D_{\mathrm{kl}}(Y_{ij} | v \;||\; Y_{ij} | v')$$

- All summands are D_{KL} between 2 Bernoulli RVs
- They are either $0, D_{\mathrm{kl}}(\alpha || \alpha'), D_{\mathrm{kl}}(\alpha' || \alpha)$ (because of our construction of the packing set).

Lemma

For $x, y \in (0,1), X \sim$ Bern$(x), Y \sim$ Bern(y). Then

$$D_{\mathrm{kl}}(x || y) \leq \frac{(x-y)^2}{y(1-y)}$$

Using the above Lemma,

$$
\begin{aligned}
D_{\mathrm{kl}}(Y_{ij}|v \,||\, Y_{ij}|v') &\leq \frac{(f(\alpha) - f(\alpha'))^2}{f(\alpha')(1 - f(\alpha'))} \\
&\leq \frac{(f'(\xi))^2(\alpha - \alpha')^2}{f(\alpha')(1 - f(\alpha'))} \quad \text{for some } \xi \in [\alpha', \alpha] \quad \text{(intermediate value theorem)} \\
&\leq \frac{(\gamma\alpha)^2}{\beta_{\alpha'}} \quad \text{(since } \alpha' = (1-\gamma)\alpha) \\
&\leq \frac{64\epsilon^2}{\beta_{\alpha'}} \quad \text{(by assumption)}
\end{aligned}
$$

$$
\Rightarrow I(V; \widehat{V}) \leq \frac{64n\epsilon^2}{\beta_{\alpha'}}
$$

Hence,

$$
\begin{aligned}
\inf_{\Psi} \mathbb{P}(\Psi(\boldsymbol{Y}_\Omega) \neq V) &\geq 1 - \frac{I(V; \widehat{V}) + \log 2}{\log |\mathcal{X}|} \\
&\geq 1 - 1024\epsilon^2 \left(\frac{64n\epsilon^2 / \beta_{\alpha'} + 1}{\alpha^2 r d_2} \right)
\end{aligned}
$$

$$\inf_{\Psi} \mathbb{P}(\Psi(\boldsymbol{Y}_\Omega) \neq V) \geq 1 - 1024\epsilon^2 \left(\frac{64n\epsilon^2/\beta_{\alpha'} + 1}{\alpha^2 r d_2} \right)$$

Recall that

$$\left\| \boldsymbol{M} - \widehat{\boldsymbol{M}} \right\|_{\mathrm{F}}^2 \geq 4 d_1 d_2 \epsilon^2$$

Lastly, choose ϵ so that we get a meaningful bound. Choose

$$\epsilon^2 = \dots$$

then they can conclude that

$$\left\| \boldsymbol{M} - \widehat{\boldsymbol{M}} \right\|_{\mathrm{F}}^2 \geq O(1/\sqrt{n})$$

with probability at least 0.75.

Some Comments

- The proof does not take into account RV Ω
- Proof of existence of packing set using probabilistic is a nice approach
- Data samples does not need to be independent
- Fano's inequality is a key step in the general minimax bound derivation