35th Conference on Neural Information Processing Systems (NeurIPS 2021)

Ising Model Selection Using ℓ_1-Regularized Linear Regression:

A Statistical Mechanics Analysis

Xiangming Meng
The University of Tokyo
Institute for Physics of Intelligence

Tomoyuki Obuchi
Kyoto University

Yoshiyuki Kabashima
The University of Tokyo
Institute for Physics of Intelligence

Oct 18th, 2021

東京大学 THE UNIVERSITY OF TOKYO

知の物理学研究センター Institute for Physics of Intelligence

京都大学 KYOTO UNIVERSITY

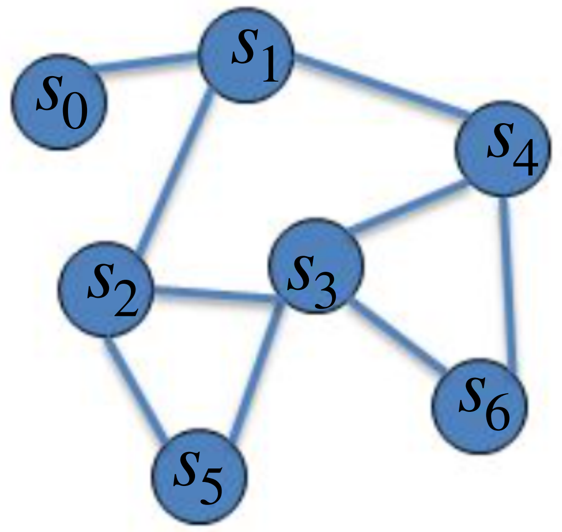
Ising Model Selection

- **Ising Model**

Binary Markov random field (**MRF**) with **pairwise** potentials [Wainwright & Jordan, 2008]

Binary spins $\boldsymbol{s} = \left(s_i\right)_{i=0}^{N-1} \in \{-1, +1\}^N$

Pairwise couplings: $\boldsymbol{J}^* = \left(J_{ij}^*\right)_{i,j} \in \mathbf{R}^{N\times N}$



$G = (\mathtt{V}, \mathtt{E})$

node set $\mathtt{V} = \{0,1,\ldots,N-1\}$

edge set $\mathtt{E} = \left\{(i,j) \mid J_{ij}^* \neq 0\right\}$

The Joint Distribution

$$P_{\text{Ising}}\left(\boldsymbol{s}|\boldsymbol{J}^*\right) = \frac{1}{Z_{\text{Ising}}\left(\boldsymbol{J}^*\right)} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Partition function

$$Z_{\text{Ising}}\left(\boldsymbol{J}^*\right) = \sum_{\boldsymbol{s}} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Wide Applications: statistical physics, image analysis, social networking, biology, etc. [Nguyen et al., 2017; Aurell & Ekeberg, 2012; BachschmidRomano & Opper, 2015; Berg, 2017; Bachschmid-Romano & Opper, 2017; Abbara et al., 2020].

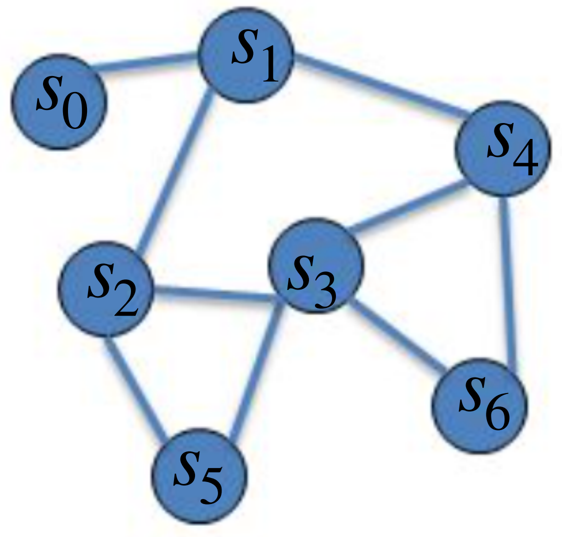
Ising Model Selection

Ising Model

Binary Markov random field (**MRF**) with **pairwise** potentials [Wainwright & Jordan, 2008]

Binary spins $\boldsymbol{s} = (s_i)_{i=0}^{N-1} \in \{-1, +1\}^N$

Pairwise couplings: $\boldsymbol{J}^* = \left(J_{ij}^*\right)_{i,j} \in \mathbf{R}^{N \times N}$



$G = (\mathbb{V}, \mathbb{E})$

node set $\mathbb{V} = \{0, 1, \ldots, N-1\}$

edge set $\mathbb{E} = \left\{(i,j) \mid J_{ij}^* \neq 0\right\}$

The Joint Distribution

$$P_{\text{Ising}}(\boldsymbol{s}|\boldsymbol{J}^*) = \frac{1}{Z_{\text{Ising}}(\boldsymbol{J}^*)} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Partition function

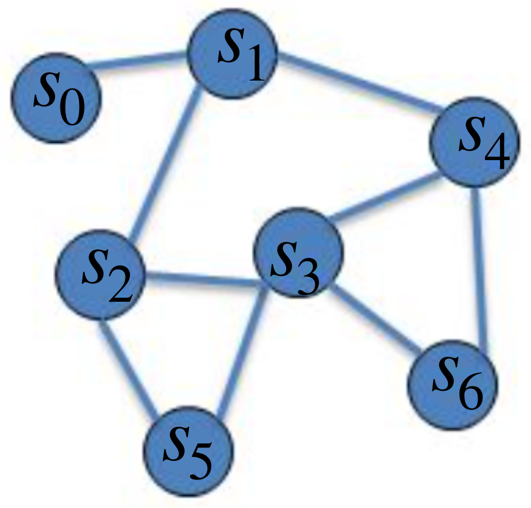
$$Z_{\text{Ising}}(\boldsymbol{J}^*) = \sum_{\boldsymbol{s}} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Wide Applications: statistical physics, image analysis, social networking, biology, etc. [Nguyen et al., 2017; Aurell & Ekeberg, 2012; BachschmidRomano & Opper, 2015; Berg, 2017; Bachschmid-Romano & Opper, 2017; Abbara et al., 2020].

Ising Model Selection

The edge set $\mathbb{E} = ?$

$G = (\mathbb{V}, \mathbb{E})$



$\boldsymbol{J}^* = \left(J_{ij}^*\right)_{i,j} \in \mathbf{R}^{N \times N}$

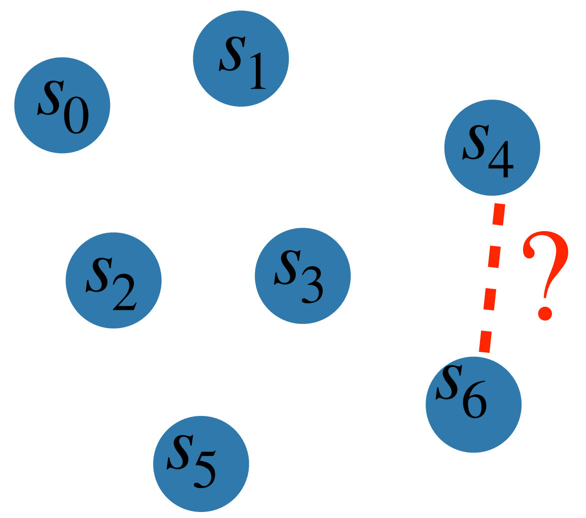
Generate i.i.d. data →

Collected Data

$$\mathcal{D}^M = \left\{\boldsymbol{s}^{(\mu)}\right\}_{\mu=1}^{M}$$

M samples

Inference →



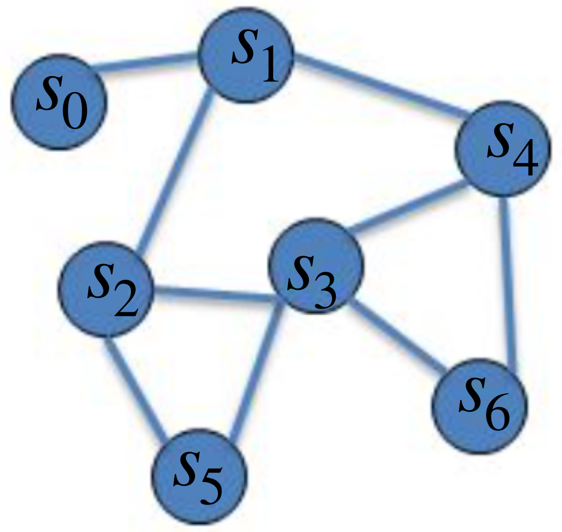
Ising Model Selection

Ising Model

Binary Markov random field (**MRF**) with **pairwise** potentials [Wainwright & Jordan, 2008]

Binary spins $\boldsymbol{s} = \left(s_i\right)_{i=0}^{N-1} \in \{-1, +1\}^N$

Pairwise couplings: $\boldsymbol{J}^* = \left(J_{ij}^*\right)_{i,j} \in \mathbf{R}^{N \times N}$



$G = (\mathbb{V}, \mathbb{E})$

node set $\mathbb{V} = \{0, 1, \ldots, N-1\}$

edge set $\mathbb{E} = \left\{(i, j) \mid J_{ij}^* \neq 0\right\}$

The Joint Distribution

$$P_{\text{Ising}}\left(\boldsymbol{s} \mid \boldsymbol{J}^*\right) = \frac{1}{Z_{\text{Ising}}\left(\boldsymbol{J}^*\right)} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Partition function

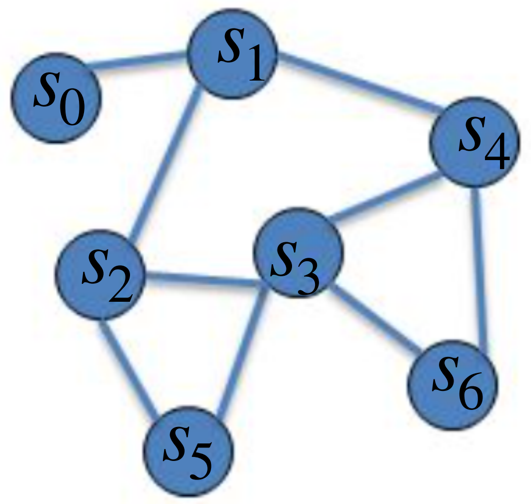
$$Z_{\text{Ising}}\left(\boldsymbol{J}^*\right) = \sum_{\boldsymbol{s}} \exp\left\{\sum_{i<j} J_{ij}^* s_i s_j\right\}$$

Wide Applications: statistical physics, image analysis, social networking, biology, etc. [Nguyen et al., 2017; Aurell & Ekeberg, 2012; BachschmidRomano & Opper, 2015; Berg, 2017; Bachschmid-Romano & Opper, 2017; Abbara et al., 2020].

Ising Model Selection

The edge set $\mathbb{E} = ?$

$G = (\mathbb{V}, \mathbb{E})$



$\boldsymbol{J}^* = \left(J_{ij}^*\right)_{i,j} \in \mathbf{R}^{N \times N}$

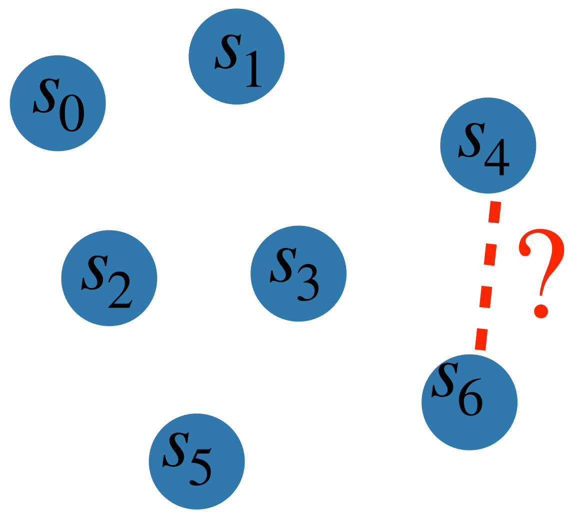
Generate i.i.d. data

Collected Data

$$\mathcal{D}^M = \left\{\boldsymbol{s}^{(\mu)}\right\}_{\mu=1}^{M}$$

M samples

Inference



Structure Learning Problem
(Inverse Ising problem)

Overview and Motivations

- **Popular Algorithms**
  - **Mean field methods** [Nguyen &Berg, 2012,Nguyen et al., 2017] **; Boltzmann learning** [Ackley et al. 1985]**, etc**
  - **Neighborhood based Methods** [Ravikumar et al., 2010;Aurell, Erik&Ekeberg 2012;Lokhov et al., 2018;Wu et al., 2019]

Recovering full edge set

$$\mathtt{E} = \left\{ (i,j) \mid J^*_{ij} \neq 0 \right\}$$

Equivalent

Recovering neighborhood of each node

$$\hat{\mathcal{N}}(i) = \left\{ j \mid \hat{J}_{ij} \neq 0, j \in \mathbb{V} \backslash i \right\}, \forall i \in \mathbb{V}$$

$$\boldsymbol{J}_{\backslash i} \equiv (J_{ij})_{j(\neq i)}$$

Overview and Motivations

- **Popular Algorithms**
  - **Mean field methods** [Nguyen &Berg, 2012,Nguyen et al., 2017] **; Boltzmann learning** [Ackley et al. 1985]**, etc**
  - **Neighborhood based Methods** [Ravikumar et al., 2010;Aurell, Erik&Ekeberg 2012;Lokhov et al., 2018;Wu et al., 2019]

Recovering full edge set

$$\mathrm{E} = \left\{ (i,j) \,|\, J^*_{ij} \neq 0 \right\}$$

Equivalent

Recovering neighborhood of each node

$$\hat{\mathcal{N}}(i) = \left\{ j \,|\, \hat{J}_{ij} \neq 0, j \in \mathbb{V} \backslash i \right\}, \forall i \in \mathbb{V}$$

$$\boldsymbol{J}_{\backslash i} \equiv (J_{ij})_{j(\neq i)}$$

ℓ_1-LogR Estimator

[Ravikumar et al., 2010]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} -\log P\left(s_i^{(\mu)} \,|\, \boldsymbol{s}_{\backslash i}^{(\mu)}, \boldsymbol{J}_i \right) + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

pseudo-likelihood (PL)
[Besag, 1975]

$$P\left(s_i \,|\, \boldsymbol{s}_{\backslash i}, \boldsymbol{J}_i \right) = \frac{1}{Z_i} e^{s_i \sum_{j(\neq i)} J_{ij} s_j}$$

Interaction Screening (IS)

[Lokhov et al., 2018]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}} + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

IS objective (ISO)
[Lokhov et al., 2018]

$$e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}}$$

Overview and Motivations

Popular Algorithms

- **Mean field methods** [Nguyen &Berg, 2012,Nguyen et al., 2017] **; Boltzmann learning** [Ackley et al. 1985]**, etc**
- **Neighborhood based Methods** [Ravikumar et al., 2010;Aurell, Erik&Ekeberg 2012;Lokhov et al., 2018;Wu et al., 2019]

Recovering full edge set

$$\mathbb{E} = \left\{ (i,j) \mid J^*_{ij} \neq 0 \right\}$$

Equivalent

Recovering neighborhood of each node

$$\hat{\mathcal{N}}(i) = \left\{ j \mid \hat{J}_{ij} \neq 0, j \in \mathbb{V} \backslash i \right\}, \forall i \in \mathbb{V}$$

$$\boldsymbol{J}_{\backslash i} \equiv (J_{ij})_{j(\neq i)}$$

ℓ_1-LogR Estimator
[Ravikumar et al., 2010]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} -\log P\left(s_i^{(\mu)} \mid \boldsymbol{s}_{\backslash i}^{(\mu)}, \boldsymbol{J}_i \right) + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

pseudo-likelihood (PL)
[Besag, 1975]

$$P\left(s_i \mid \boldsymbol{s}_{\backslash i}, \boldsymbol{J}_i \right) = \frac{1}{Z_i} e^{s_i \sum_{j(\neq i)} J_{ij} s_j}$$

Interaction Screening (IS)
[Lokhov et al., 2018]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}} + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

IS objective (ISO)
[Lokhov et al., 2018]

$$e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}}$$

A Unified View as M-estimator

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} \ell\left(s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)} \right) + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

$$\ell(x) = \begin{cases} \log\left(1 + e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

Overview and Motivations

- **Popular Algorithms**
  - **Mean field methods** [Nguyen &Berg, 2012,Nguyen et al., 2017] **; Boltzmann learning** [Ackley et al. 1985]**, etc**
  - **Neighborhood based Methods** [Ravikumar et al., 2010;Aurell, Erik&Ekeberg 2012;Lokhov et al., 2018;Wu et al., 2019]

Recovering full edge set $\mathrm{E} = \left\{(i,j) \mid J^*_{ij} \neq 0\right\}$ ⟺ **Equivalent** ⟺ Recovering neighborhood of each node $\hat{\mathcal{N}}(i) = \left\{j \mid \hat{J}_{ij} \neq 0, j \in \mathrm{V}\backslash i\right\}, \forall i \in \mathrm{V}$

$$\boldsymbol{J}_{\backslash i} \equiv (J_{ij})_{j(\neq i)}$$

ℓ_1-LogR Estimator [Ravikumar et al., 2010]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M}\sum_{\mu=1}^{M} -\log P\left(s_i^{(\mu)} \mid \boldsymbol{s}_{\backslash i}^{(\mu)}, \boldsymbol{J}_i\right) + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

pseudo-likelihood (PL) [Besag, 1975] $P\left(s_i \mid \boldsymbol{s}_{\backslash i}, \boldsymbol{J}_i\right) = \frac{1}{Z_i} e^{s_i \sum_{j(\neq i)} J_{ij} s_j}$

Interaction Screening (IS) [Lokhov et al., 2018]

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M}\sum_{\mu=1}^{M} e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}} + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

IS objective (ISO) [Lokhov et al., 2018] $e^{-s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}}$

A Unified View as M-estimator

$$\hat{\boldsymbol{J}}_{\backslash i} = \arg\min_{\boldsymbol{J}_i} \left\{ \frac{1}{M}\sum_{\mu=1}^{M} \ell\left(s_i^{(\mu)} \sum_{j(\neq i)} J_{ij} s_j^{(\mu)}\right) + \lambda \left\| \boldsymbol{J}_{\backslash i} \right\|_1 \right\}$$

$$\ell(x) = \begin{cases} \log\left(1 + e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

One Natural Question: How about other loss functions, e.g., quadratic loss?

Main Contributions

Does it work for binary data?

- **ℓ_1-Regularized Linear Regression (ℓ_1-LinR)** [Tibshirani, 1996]

Our main focus

$$\hat{\boldsymbol{J}}_{\setminus i} = \arg\min_{\boldsymbol{J}_{\setminus i}} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} \frac{1}{2} \left(s_i^{(\mu)} - \sum_{j(\neq i)} J_{ij} s_j^{(\mu)} \right)^2 + \lambda \left\| \boldsymbol{J}_{\setminus i} \right\|_1 \right\}$$

quadratic loss $\ell(x) = \frac{1}{2}(1-x)^2$

- **One representative example of *model misspecification***
- **ℓ_1-LinR (LASSO), as one most popular linear estimator, is *more efficient than nonlinear ones***

Main Contributions

- ℓ_1**-Regularized Linear Regression (**ℓ_1**-LinR)** [Tibshirani, 1996]

Does it work for binary data?

Our main focus

$$\hat{J}_{\setminus i} = \arg\min_{J_{\setminus i}} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} \frac{1}{2} \left(s_i^{(\mu)} - \sum_{j(\neq i)} J_{ij} s_j^{(\mu)} \right)^2 + \lambda \left\| J_{\setminus i} \right\|_1 \right\}$$

quadratic loss $\ell(x) = \frac{1}{2}(1-x)^2$

  - **One representative example of** ***model misspecification***
  - ℓ_1**-LinR (LASSO), as one most popular linear estimator, is** ***more efficient than nonlinear ones***

- **Main Contributions**
  - **A statistical mechanics analysis of the** ***typical*** **learning performances of** ℓ_1**-LinR for** ***typical*** **paramagnetic random regular (RR) graphs**
    - — **An accurate estimate of the** ***typical*** **sample complexity of** ℓ_1**-LinR: same order** $M = \mathcal{O}(\log N)$ **as** ℓ_1**-LogR!**
    - — **A sharp** ***quantitative*** **prediction of** ***non-asymptotic*** **(moderate** M, N**) performances of** ℓ_1**-LinR, e.g., precision, recall, RSS**
  - **Our analysis method applies to** ***any*** ℓ_1**-regularized M-estimator including** ℓ_1**-LogR and IS**

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

general loss function

(s_0 is considered)

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \arg\min_{\boldsymbol{J}} \left[\frac{1}{M} \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda \left\| \boldsymbol{J} \right\|_1 \right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

general loss function

(s_0 is considered)

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\arg\min}\left[\frac{1}{M}\sum_{\mu=1}^{M}\ell\left(s_0^{(\mu)}h^{(\mu)}\right) + \lambda\,\|\boldsymbol{J}\|_1\right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

A Statistical Mechanics System

Hamiltonian $$\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M}\ell\left(s_0^{(\mu)}h^{(\mu)}\right) + \lambda M\|\boldsymbol{J}\|_1$$

Boltzmann distribution $$P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z}e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)} \qquad Z = \int d\boldsymbol{J}e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$$

$\mathcal{D}^M$
plays the role of
quenched disorder

[Opper & Saad, 2001; Nishimori, 2001; Mezard& Montanari, 2009]

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

general loss function

(s_0 is considered)

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \arg\min_{\boldsymbol{J}} \left[\frac{1}{M} \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda \left\| \boldsymbol{J} \right\|_1 \right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

A Statistical Mechanics System

Hamiltonian $\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda M \|\boldsymbol{J}\|_1$

Boltzmann distributior $P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z} e^{-\beta \mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$ $\quad Z = \int d\boldsymbol{J} e^{-\beta \mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$

$\beta \to +\infty$

$\delta\left(\boldsymbol{J} - \hat{\boldsymbol{J}}\left(\mathcal{D}^M\right)\right)$

The Boltzmann distribution freezes onto the solution $\hat{\boldsymbol{J}}$ as $\beta \to +\infty$!

$\mathcal{D}^M$ plays the role of quenched disorder

[Opper & Saad, 2001; Nishimori, 2001; Mezard& Montanari, 2009]

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

general loss function

(s_0 is considered)

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \underset{\boldsymbol{J}}{\arg\min}\left[\frac{1}{M}\sum_{\mu=1}^{M}\ell\left(s_0^{(\mu)}h^{(\mu)}\right) + \lambda \left\|\boldsymbol{J}\right\|_1\right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

A Statistical Mechanics System

Hamiltonian $$\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M}\ell\left(s_0^{(\mu)}h^{(\mu)}\right) + \lambda M\left\|\boldsymbol{J}\right\|_1$$

Boltzmann distribution $$P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z}e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)} \qquad Z = \int d\boldsymbol{J}e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$$

$\beta \to +\infty$

$$\delta\left(\boldsymbol{J} - \hat{\boldsymbol{J}}\left(\mathcal{D}^M\right)\right)$$

The Boltzmann distribution freezes onto the solution $\hat{\boldsymbol{J}}$ as $\beta \to +\infty$!

$\mathcal{D}^M$ plays the role of quenched disorder

[Opper & Saad, 2001; Nishimori, 2001; Mezard& Montanari, 2009]

Statistical mechanics analysis

The key quantity $$f(\mathcal{D}^M) = -\frac{1}{N\beta}\log Z$$

free energy density

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

general loss function

(s_0 is considered)

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \arg\min_{\boldsymbol{J}} \left[\frac{1}{M}\sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda \left\| \boldsymbol{J} \right\|_1\right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

A Statistical Mechanics System

Hamiltonian $\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda M \|\boldsymbol{J}\|_1$

Boltzmann distribution $P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z} e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$ $\quad Z = \int d\boldsymbol{J} e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$

$\beta \to +\infty$

$\delta\left(\boldsymbol{J} - \hat{\boldsymbol{J}}\left(\mathcal{D}^M\right)\right)$

The Boltzmann distribution freezes onto the solution $\hat{\boldsymbol{J}}$ as $\beta \to +\infty$!

$\mathcal{D}^M$ plays the role of quenched disorder

[Opper & Saad, 2001; Nishimori, 2001; Mezard& Montanari, 2009]

Statistical mechanics analysis

The key quantity $f(\mathcal{D}^M) = -\frac{1}{N\beta}\log Z$

free energy density

[Nishimori, 2001]

Self-Averaging

for large N, M

$$f = -\frac{1}{N\beta}\left[\log Z\right]_{\mathcal{D}^M}$$

***average* free energy density**

averaged over the disorder, i.e. dataset

Problem Formulation

- **Statistical Mechanics Perspective**

The ℓ_1-regularized M-estimator

(s_0 is considered)

general loss function

$$\hat{\boldsymbol{J}}\left(\mathcal{D}^M\right) \equiv \hat{\boldsymbol{J}} = \arg\min_{\boldsymbol{J}} \left[\frac{1}{M}\sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda \left\| \boldsymbol{J} \right\|_1\right]$$

$$\ell(x) = \begin{cases} \frac{1}{2}(1-x)^2 & \ell_1\text{-LinR} \\ \log\left(1+e^{-2x}\right) & \ell_1\text{-LogR} \\ e^{-x} & \text{IS} \end{cases}$$

A Statistical Mechanics System

Hamiltonian $$\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right) = \sum_{\mu=1}^{M} \ell\left(s_0^{(\mu)} h^{(\mu)}\right) + \lambda M \left\|\boldsymbol{J}\right\|_1$$

Boltzmann distribution $$P\left(\boldsymbol{J}|\mathcal{D}^M\right) = \frac{1}{Z} e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)} \qquad Z = \int d\boldsymbol{J} e^{-\beta\mathcal{H}\left(\boldsymbol{J}|\mathcal{D}^M\right)}$$

$\beta \to +\infty$

$$\delta\left(\boldsymbol{J} - \hat{\boldsymbol{J}}\left(\mathcal{D}^M\right)\right)$$

The Boltzmann distribution freezes onto the solution $\hat{\boldsymbol{J}}$ as $\beta \to +\infty$!

$\mathcal{D}^M$ plays the role of quenched disorder

[Opper & Saad, 2001; Nishimori, 2001; Mezard& Montanari, 2009]

Statistical mechanics analysis

The key quantity $$f(\mathcal{D}^M) = -\frac{1}{N\beta}\log Z$$

free energy density

[Nishimori, 2001]

Self-Averaging

for large N, M

$$f = -\frac{1}{N\beta}\left[\log Z\right]_{\mathcal{D}^M}$$

averaged over the disorder, i.e. dataset

average free energy density

Difficult to calculate and we resort to the **replica method!**

Replica Method

- **Basic Idea**

$$f = -\frac{1}{N\beta}[\log Z]_{\mathcal{D}^M} = -\lim_{n\to 0}\frac{1}{N\beta}\frac{\partial \log [Z^n]_{\mathcal{D}^M}}{\partial n}$$

[Mézard et al 1987; Opper & Saad, 2001; Nishimori, 2001; Mézard & Montanari, 2009]

- **Procedure**

1. Compute $[Z^n]_{\mathcal{D}^M}$ for $n \in \mathbb{N}$
2. Take $N \to \infty$ limit using Laplace/Saddle-point method
3. Obtain an analytically continuable form w.r.t. n under appropriate ansatz
   - replica symmetry (RS) is used here (*due to convexity of estimator*)
4. Take $n \to 0$ limit using the obtained analytically continuable form

Replica Method

- **Basic Idea**

$$f = -\frac{1}{N\beta}[\log Z]_{\mathcal{D}^M} = -\lim_{n\to 0}\frac{1}{N\beta}\frac{\partial \log [Z^n]_{\mathcal{D}^M}}{\partial n}$$

[Mézard et al 1987; Opper & Saad, 2001; Nishimori, 2001; Mézard & Montanari, 2009]

- **Procedure**

1. Compute $[Z^n]_{\mathcal{D}^M}$ for $n \in \mathbb{N}$
2. Take $N \to \infty$ limit using Laplace/Saddle-point method
3. Obtain an analytically continuable form w.r.t. n under appropriate ansatz
   - replica symmetry (RS) is used here (*due to convexity of estimator*)
4. Take $n \to 0$ limit using the obtained analytically continuable form

- **Comments**

1. In present case for Ising model selection, the detailed replica computation is still far from trivial
   - We use an approach based on *cavity method* [Bachschmid-Romano & Opper 2017, Abbara et al., 2020; Meng et al., 2021]
   - We propose two ansatzs to enable the calculation, which can be (numerically) verified.

2. Although the replica method is non-rigorous, our results are supported by experimental results.

Free Energy Result

- **Result of replica method**

In the case of ℓ_1-LinR estimator

$$f(\beta\to\infty) = -\underset{\Theta}{\mathrm{Extr}}\left\{\begin{array}{c} -\frac{\alpha}{2(1+\chi)}\mathbb{E}_{s,z}\left(\left(s_0-\sum_{j\in\Psi}\bar{J}_j s_j-\sqrt{Q}z\right)^2\right)-\lambda\alpha\sum_{j\in\Psi}\left|\bar{J}_j\right| \\ +(-ER+F\eta)G'(-E\eta)+\frac{1}{2}EQ-\frac{1}{2}F\chi+\frac{1}{2}KR-\frac{1}{2}H\eta \\ -\mathbb{E}_z\min_w\left\{\frac{K}{2}w^2-\sqrt{H}zw+\frac{\lambda M}{\sqrt{N}}|w|\right\}\end{array}\right\}$$

Notations Definition

$$G(x) = -\frac{1}{2}\log x-\frac{1}{2}+\underset{\Lambda}{\mathrm{Extr}}\left\{-\frac{1}{2}\int\log(\Lambda-\gamma)\rho(\gamma)\,d\gamma+\frac{\Lambda}{2}x\right\}$$

$\rho(\lambda)$ eigenvalue distribution (EVD) of covariance matrix $C^{\backslash 0}$ of Ising model without s_0 (available for RR graph)

$$\Theta = \left\{\chi, Q, E, R, F, \eta, K, H, \{\bar{J}_j\}_{j\in\Psi}\right\}$$

$\underset{\Theta}{\mathrm{Extr}}\{\cdot\}$ denotes extremization operation over parameters

$\mathbb{E}_{s,z}$ denotes joint expectation with s, z, where $z\sim\mathcal{N}(0,1)$ and $s\propto e^{s_0\sum_{j\in\Psi}J_j^* s_j}$

How to solve f? **Equations of state (EOS)**

$$\begin{cases} E = \frac{\alpha}{(1+\chi)}, \\ F = \frac{\alpha}{(1+\chi)^2}\left[\mathbb{E}_s\left(s_0-\sum_{j\in\Psi}s_j\bar{J}_j\right)^2+Q\right], \\ R = \frac{1}{K^2}\left[\left(H+\frac{\lambda^2M^2}{N}\right)\mathrm{erfc}\left(\frac{\lambda M}{\sqrt{2HN}}\right)-2\lambda M\sqrt{\frac{H}{N}}\frac{1}{\sqrt{2\pi}}e^{-\frac{\lambda^2M^2}{2HN}}\right], \\ E\eta = -\int\frac{\rho(\gamma)}{\tilde{\Lambda}-\gamma}d\gamma, \\ Q = \frac{F}{E^2}+R\tilde{\Lambda}-\frac{(-ER+F\eta)\eta}{\int\frac{\rho(\gamma)}{(\tilde{\Lambda}-\gamma)^2}d\gamma}, \\ K = E\tilde{\Lambda}+\frac{1}{\eta}, \\ \chi = \frac{1}{E}+\eta\tilde{\Lambda}, \\ H = \frac{R}{\eta^2}+F\tilde{\Lambda}+\frac{(-ER+F\eta)E}{\int\frac{\rho(\gamma)}{(\tilde{\Lambda}-\gamma)^2}d\gamma}, \\ \eta = \frac{1}{K}\mathrm{erfc}\left(\frac{\lambda M}{\sqrt{2HN}}\right), \\ \bar{J}_j = \frac{\mathrm{soft}(\tanh(K_0),\lambda(1+\chi))}{1+(d-1)\tanh^2(K_0)}, j\in\Psi, \end{cases}$$

The EOS can be efficiently solved numerically!

Equivalent Probabilistic Model of ℓ_1-LinR

- **The estimates of ℓ_1-LinR are decoupled**

$$\hat{\boldsymbol{J}} = \arg\min_{\boldsymbol{J}} \left\{ \frac{1}{M} \sum_{\mu=1}^{M} \frac{1}{2} \left(s_i^{(\mu)} - \sum_{j(\neq i)} J_{ij} s_j^{(\mu)} \right)^2 + \lambda \left\| \boldsymbol{J} \right\|_1 \right\}$$

Highly coupled
&
Difficult to analyze

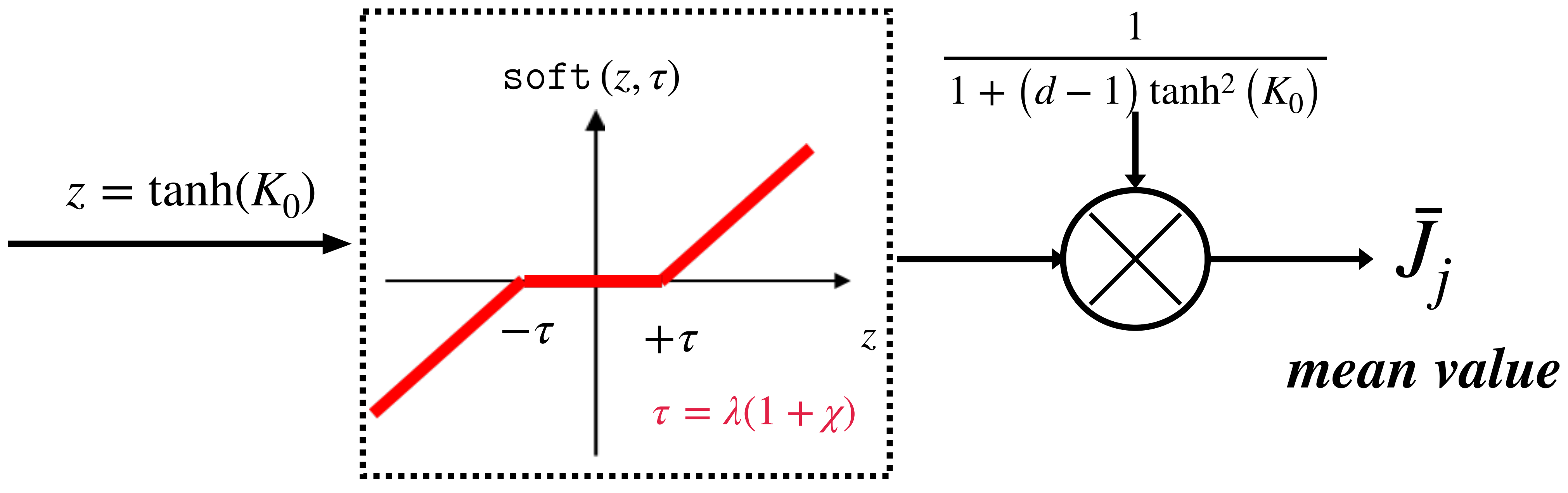
Decoupled (Replica method)

Probabilistic Model of ℓ_1-LinR

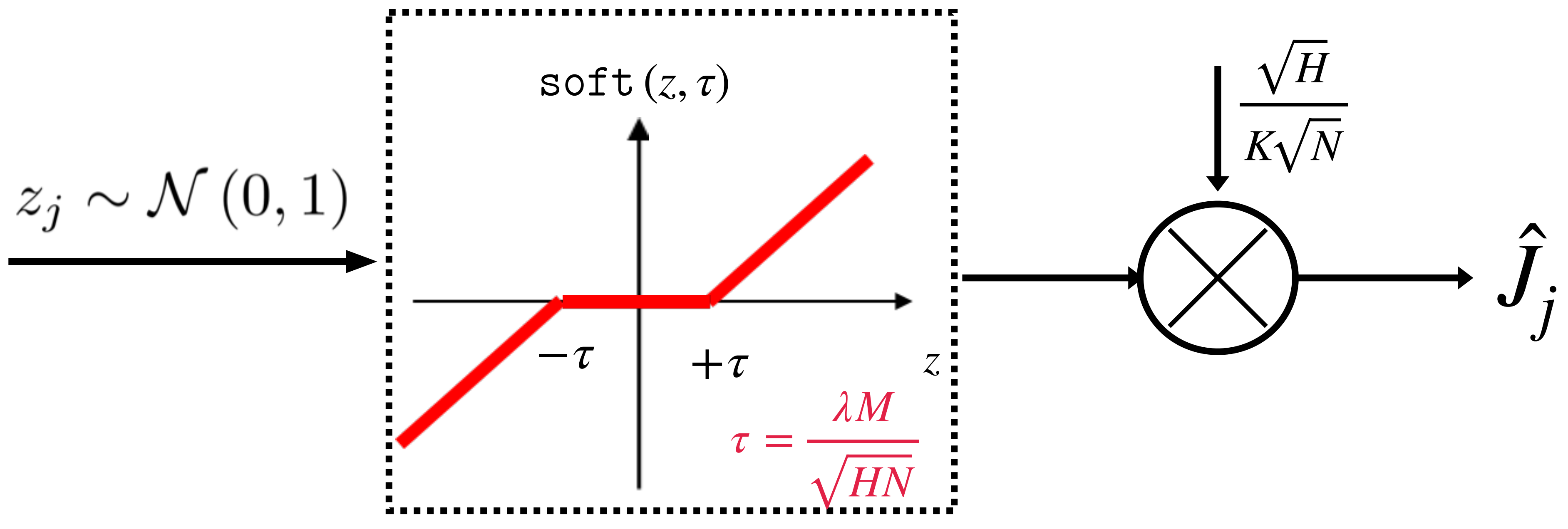
Statistically equivalent to two scalar estimators !

$$\hat{J}_j = \begin{cases} \dfrac{\texttt{soft}(\tanh(K_0), \lambda(1+\chi))}{1+(d-1)\tanh^2(K_0)} \equiv \bar{J}_j, & j \in \Psi \quad \text{Active set} \\ \dfrac{\sqrt{H}}{K\sqrt{N}} \texttt{soft}\left(z_j, \dfrac{\lambda M}{\sqrt{HN}}\right), \; z_j \sim \mathcal{N}(0,1), & j \in \bar{\Psi} \quad \text{Inactive set} \end{cases}$$

Easy to analyze



(a) Equivalent scalar estimator for the active set



(b) Equivalent scalar estimator for the inactive set

High-dimensional Asymptotic Result

- **Sample complexity of ℓ_1-LinR**

Definition 1: An estimator is called *model selection consistent* if both the associated precision and recall satisfy $Precision \to 1$ and $Recall \to 1$ as $N \to \infty$.

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}$$

Estimated Results

True Results	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

High-dimensional Asymptotic Result

- **Sample complexity of ℓ_1-LinR**

Definition 1: An estimator is called *model selection consistent* if both the associated precision and recall satisfy *Precision* $\to 1$ and *Recall* $\to 1$ as $N \to \infty$.

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}$$

Estimated Results

True Results	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Results from the two scalar estimators:

$$\triangle = \mathbb{E}_{s_0}\left(s - \sum_{j\in\Psi} s_j \bar{J}_j\right)^2$$

$$FP < \frac{1}{\sqrt{\pi}} e^{-\frac{\lambda^2 M}{2\triangle} + \log N} \to 0 \text{ as } N \to \infty \quad \text{if } M > \frac{2\triangle \log N}{\lambda^2}$$

$$FN \to 0 \text{ as } N \to \infty \quad \text{if } 0 < \lambda < \tanh(K_0)$$

To achieve model selection consistency

Sample complexity $$M > \frac{c(\lambda, K_0)\log N}{\lambda^2}, \lambda \in \left(0, \tanh(K_0)\right)$$

Lower bound $$M > \frac{2\log N}{\tanh^2(K_0)} \quad \lambda \to \tanh(K_0)$$

High-dimensional Asymptotic Result

- **Sample complexity of ℓ_1-LinR**

Definition 1: An estimator is called *model selection consistent* if both the associated precision and recall satisfy $Precision \to 1$ and $Recall \to 1$ as $N \to \infty$.

$$Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}$$

Estimated Results

True Results	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Results from the two scalar estimators:

$$\triangle = \mathbb{E}_{s_0}\Big(s - \sum_{j\in\Psi} s_j \bar{J}_j\Big)^2$$

$$FP < \frac{1}{\sqrt{\pi}} e^{-\frac{\lambda^2 M}{2\triangle} + \log N} \to 0 \text{ as } N \to \infty \qquad \text{if } M > \frac{2\triangle \log N}{\lambda^2}$$

$$FN \to 0 \text{ as } N \to \infty \qquad \text{if } 0 < \lambda < \tanh(K_0)$$

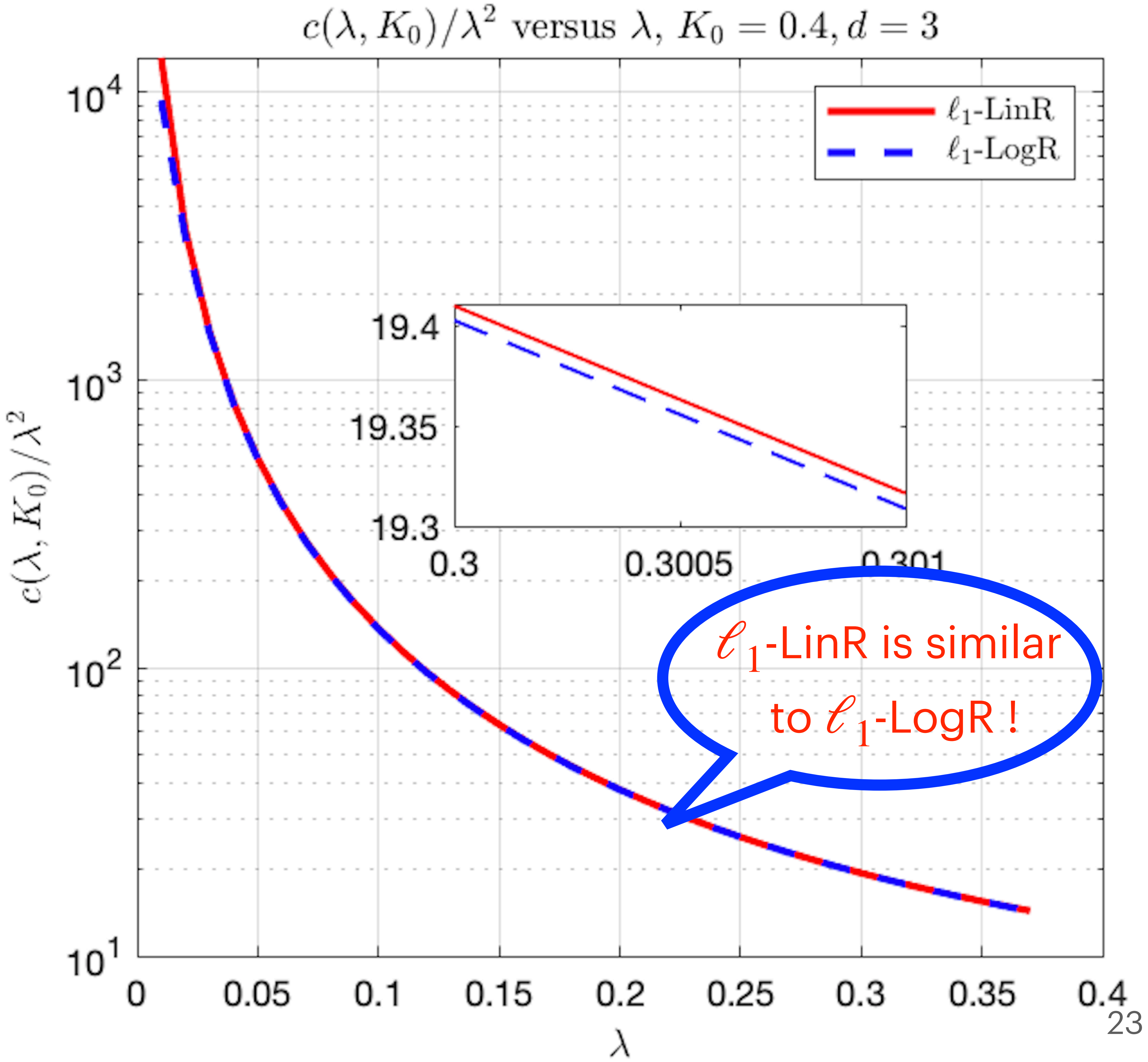
To achieve model selection consistency

Sample complexity

$$M > \frac{c(\lambda, K_0)\log N}{\lambda^2}, \lambda \in (0, \tanh(K_0))$$

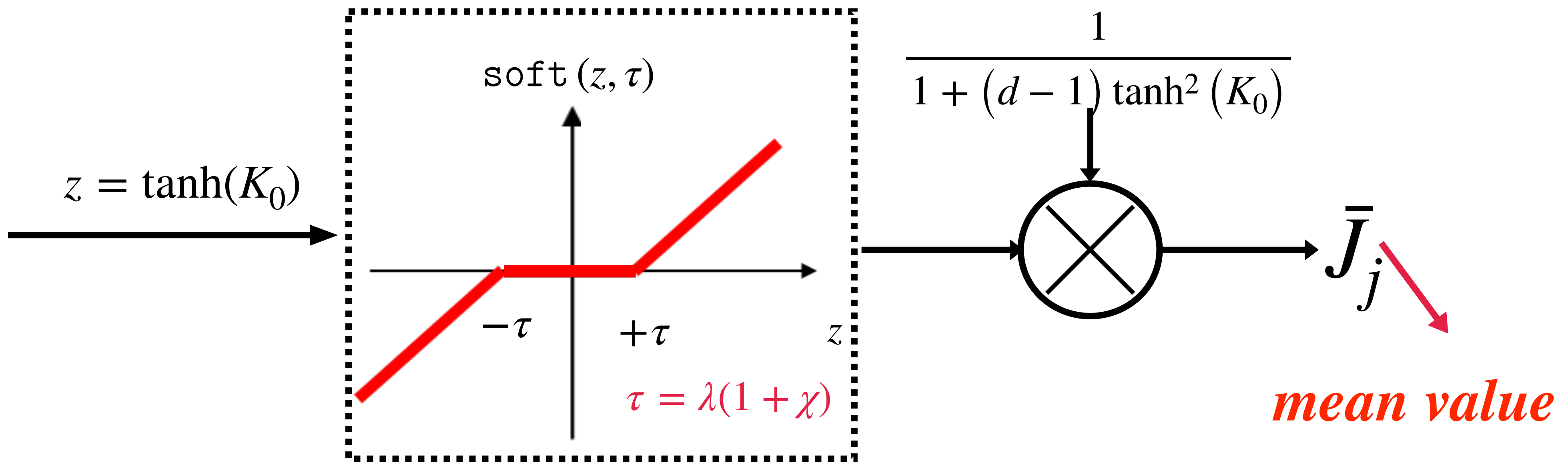
Lower bound

$$M > \frac{2\log N}{\tanh^2(K_0)} \qquad \lambda \to \tanh(K_0)$$



Non-Asymptotic Predictions

- **To account for the finite-size effect**

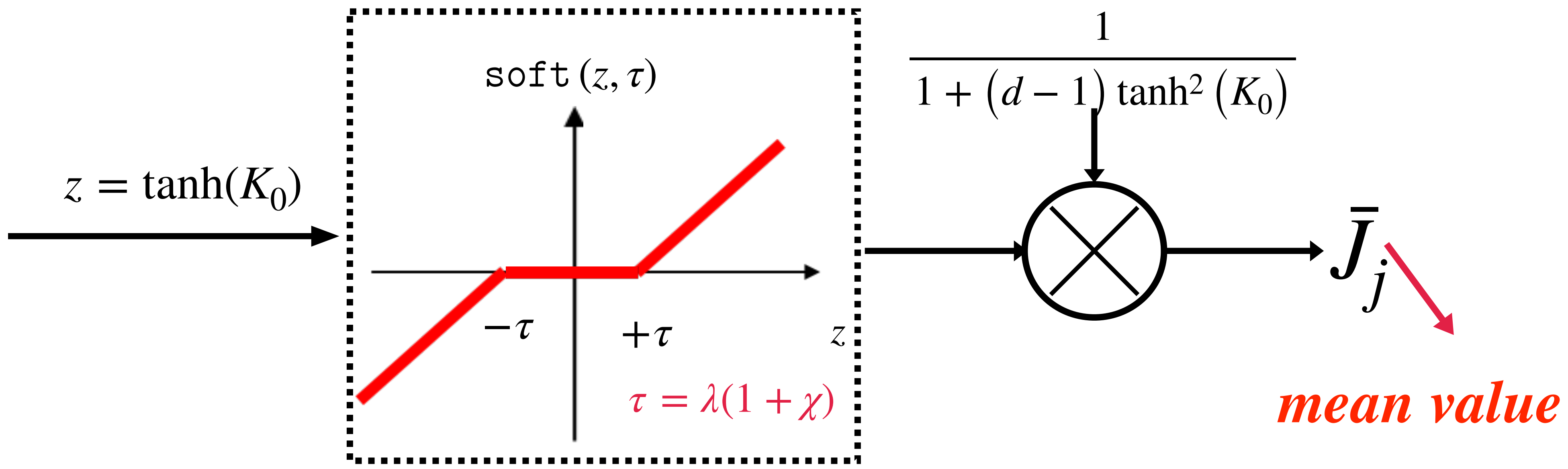


(a) Equivalent scalar estimator for the active set

- **Current scalar estimator (a) only produces the mean-value result**
  - The fluctuations of estimates in the active set Ψ are *averaged out*

Non-Asymptotic Predictions

■ **To account for the finite-size effect**



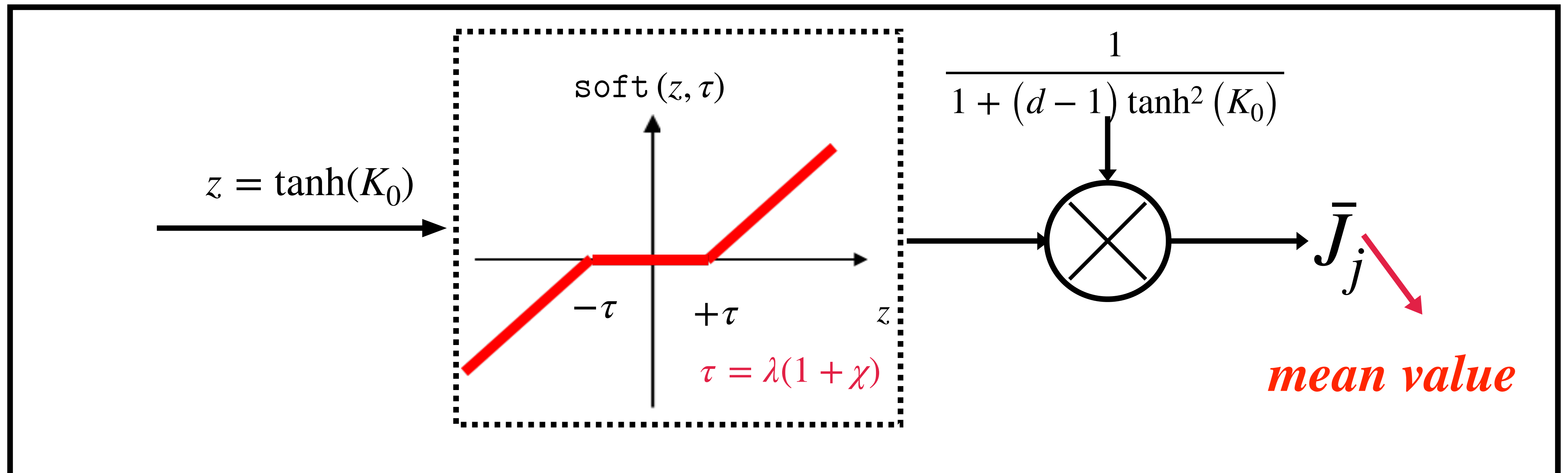
(a) Equivalent scalar estimator for the active set

- **Current scalar estimator (a) only produces the mean-value result**
  - The fluctuations of estimates in the active set Ψ are *averaged out*
- **New idea**: **Replacing expectation in free energy with sample average**
  - The modified free energy can be solved iteratively (Algorithm 1)

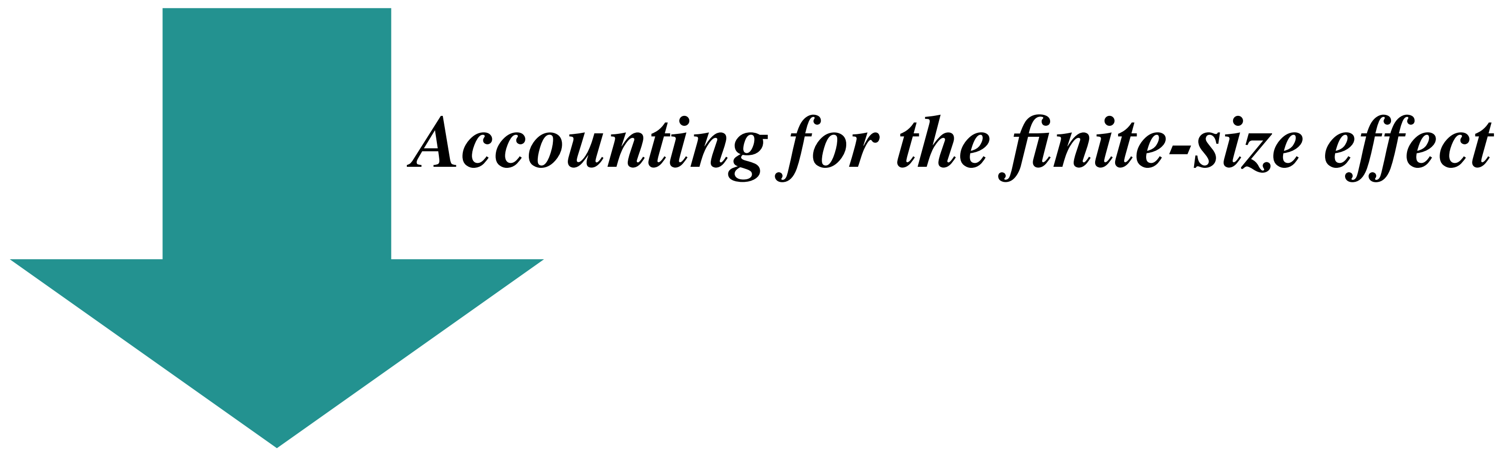
$$f(\beta \to \infty) = -\underset{\Theta}{\mathrm{Extr}} \left\{ \begin{array}{c} -\frac{\alpha}{2(1+\chi)} \frac{1}{T_{MC} M} \sum_{t=1}^{T_{MC}} \sum_{\mu=1}^{M} \left(\left(s_0^{\mu,t} - \sum_{j \in \Psi} J_j s_j^{\mu,t} - \sqrt{Q} z^{\mu,t} \right)^2 \right) \\ -\lambda\alpha \sum_{j \in \Psi} \left| \bar{J}_j \right| + (-ER + F\eta) G^{'}(-E\eta) + \frac{1}{2} EQ - \frac{1}{2} F\chi + \frac{1}{2} KR - \frac{1}{2} H\eta \\ -\mathbb{E}_z \underset{w}{\min} \left\{ \frac{K}{2} w^2 - \sqrt{H} z w + \frac{\lambda M}{\sqrt{N}} |w| \right\} \end{array} \right\}$$

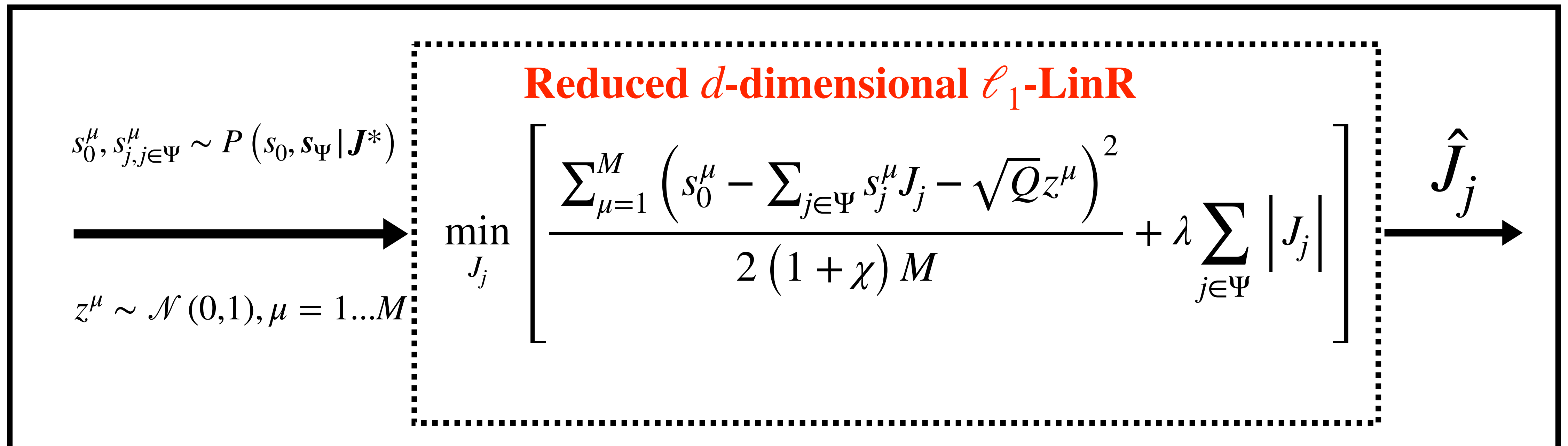
Non-Asymptotic Predictions

■ **To account for the finite-size effect**



(a) Equivalent scalar estimator for the active set





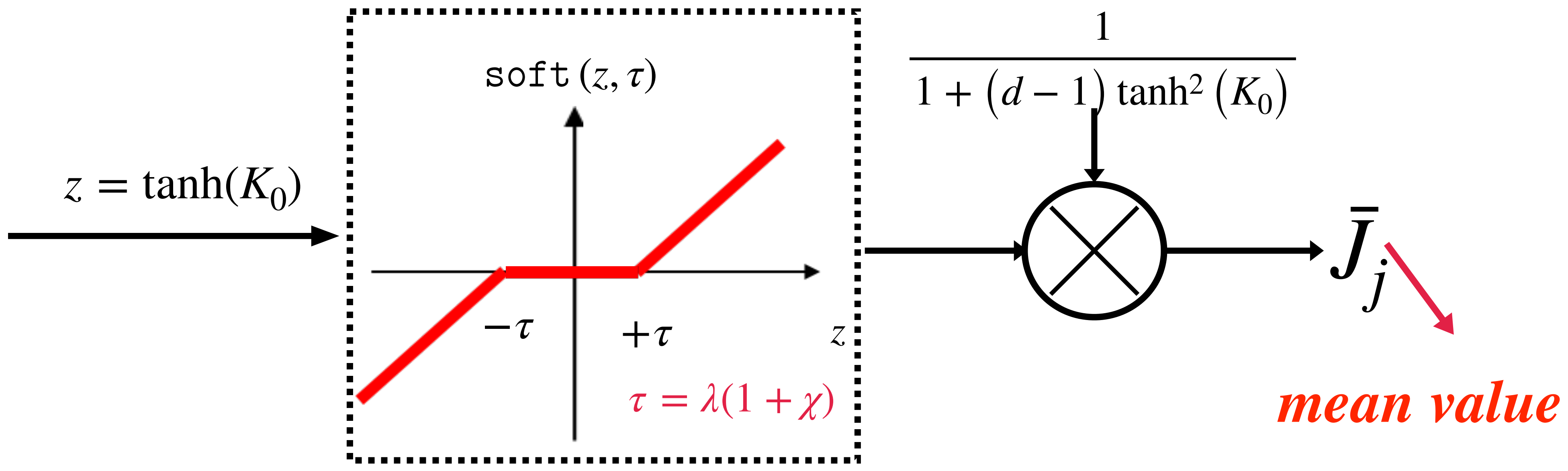
(c) Equivalent d-dimensional estimator for active set

- **Current scalar estimator (a) only produces the mean-value result**
  - The fluctuations of estimates in the active set Ψ are *averaged out*
- **New idea**: **Replacing expectation in free energy with sample average**
  - The modified free energy can be solved iteratively (Algorithm 1)

$$f(\beta \to \infty) = -\underset{\Theta}{\mathrm{Extr}} \left\{ \begin{array}{c} -\frac{\alpha}{2(1+\chi)} \frac{1}{T_{MC} M} \sum_{t=1}^{T_{MC}} \sum_{\mu=1}^{M} \left(\left(s_0^{\mu,t} - \sum_{j\in\Psi} J_j s_j^{\mu,t} - \sqrt{Q} z^{\mu,t} \right)^2 \right) \\ -\lambda\alpha \sum_{j\in\Psi} \left| \bar{J}_j \right| + (-ER + F\eta) G'(-E\eta) + \frac{1}{2}EQ - \frac{1}{2}F\chi + \frac{1}{2}KR - \frac{1}{2}H\eta \\ -\mathbb{E}_z \min_{w} \left\{ \frac{K}{2} w^2 - \sqrt{H} z w + \frac{\lambda M}{\sqrt{N}} |w| \right\} \end{array} \right\}$$

Non-Asymptotic Predictions

■ To account for the finite-size effect



(a) Equivalent scalar estimator for the active set

Accounting for the finite-size effect

Reduced d-dimensional ℓ_1-LinR

$s_0^\mu, s_{j,j\in\Psi}^\mu \sim P\left(s_0, s_\Psi \mid \boldsymbol{J}^*\right)$

$z^\mu \sim \mathcal{N}(0,1), \mu = 1 \ldots M$

$$\min_{J_j}\left[\frac{\sum_{\mu=1}^{M}\left(s_0^\mu - \sum_{j\in\Psi} s_j^\mu J_j - \sqrt{Q} z^\mu\right)^2}{2\left(1+\chi\right)M} + \lambda \sum_{j\in\Psi}\left|J_j\right|\right] \rightarrow \hat{J}_j$$

(c) Equivalent d-dimensional estimator for active set

- **Current scalar estimator (a) only produces the mean-value result**
  - The fluctuations of estimates in the active set Ψ are *averaged out*
- **New idea**: **Replacing expectation in free energy with sample averages**
  - The modified free energy can be solved iteratively (Algorithm 1)

$$f(\beta \to \infty) = -\underset{\Theta}{\mathrm{Extr}}\left\{\begin{array}{c} -\frac{\alpha}{2(1+\chi)}\frac{1}{T_{MC}M}\sum_{t=1}^{T_{MC}}\sum_{\mu=1}^{M}\left(\left(s_0^{\mu,t} - \sum_{j\in\Psi} J_j s_j^{\mu,t} - \sqrt{Q} z^{\mu,t}\right)^2\right) \\ -\lambda\alpha\sum_{j\in\Psi}\left|\bar{J}_j\right| + \left(-ER + F\eta\right)G'\left(-E\eta\right) + \frac{1}{2}EQ - \frac{1}{2}F\chi + \frac{1}{2}KR - \frac{1}{2}H\eta \\ -\mathbb{E}_z \min_w\left\{\frac{K}{2}w^2 - \sqrt{H}zw + \frac{\lambda M}{\sqrt{N}}\left|w\right|\right\} \end{array}\right\}$$

■ Predicting Non-Asymptotic performances

Given modified estimator (c) and scalar estimator (b), one can then easily obtain the non-asymptotic performances of ℓ_1-LinR, e.g., Precision, Recall, RSS, with a number of T_{MC} MC simulations

$$\begin{cases} Precision = \frac{1}{T_{\mathrm{MC}}}\sum_{t=1}^{T_{\mathrm{MC}}}\frac{\left\|\hat{J}_{j,j\in\Psi}^t\right\|_0}{\left\|\hat{J}_{j,j\in\Psi}^t\right\|_0 + \left\|\hat{J}_{j,j\in\bar{\Psi}}^t\right\|_0} \\ Recall = \frac{1}{T_{\mathrm{MC}}}\sum_{t=1}^{T_{\mathrm{MC}}}\frac{\left\|\hat{J}_{j,j\in\Psi}^t\right\|_0}{d} \\ RSS = \frac{1}{T_{\mathrm{MC}}}\sum_{t=1}^{T_{\mathrm{MC}}}\sum_{j\in\Psi}\left|\hat{J}_j^t - K_0\right|^2 + R \end{cases}$$

Experimental Results

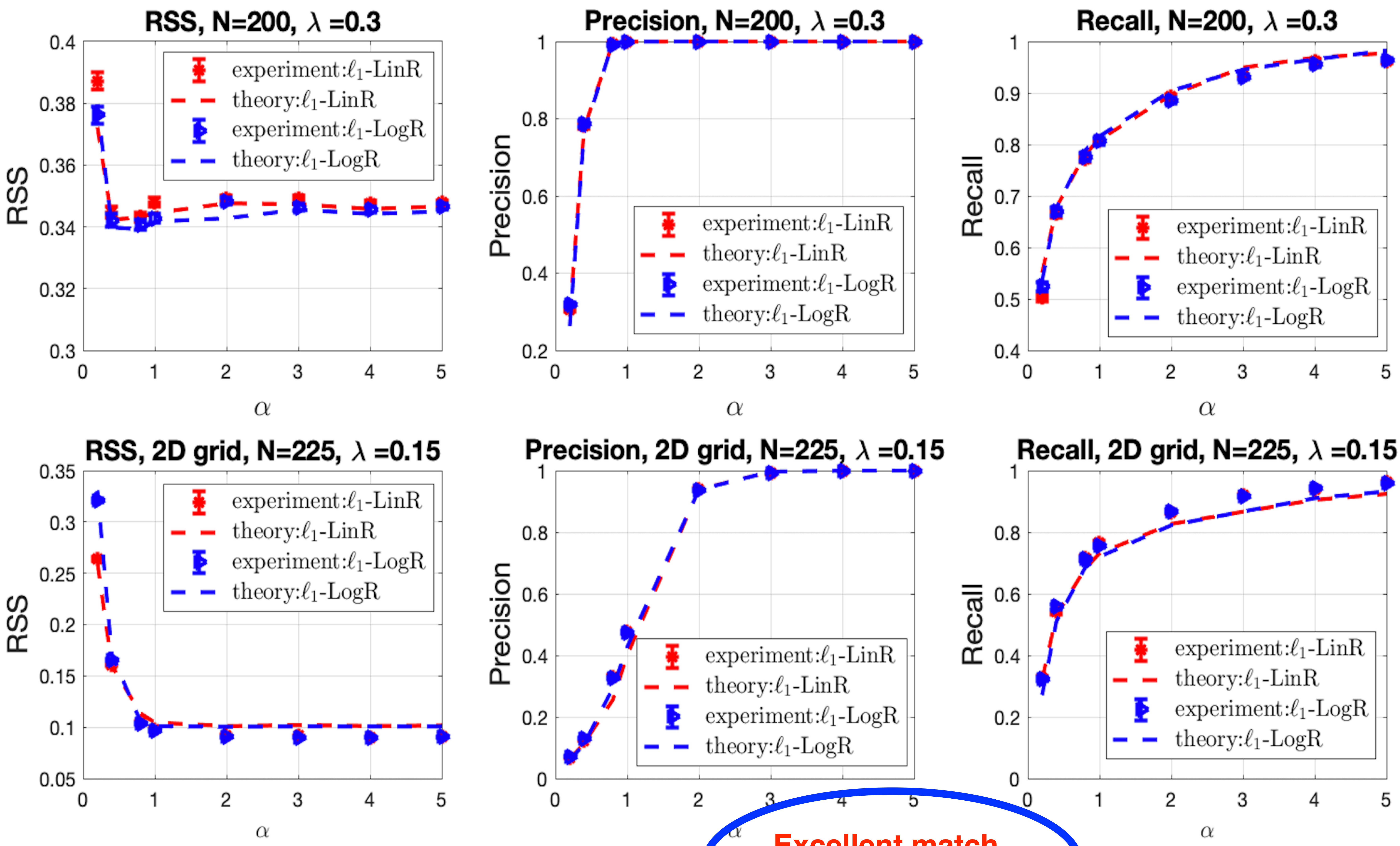
- **Accurate non-Asymptotic Predictions**

Ising model:

- RR graph, $K_0 = 0.4$, $d = 3$
- 2D grid (loopy), $K_0 = 0.2$, $d = 4$

Estimators:

ℓ_1-LinR and ℓ_1-LogR
$\lambda = 0.3$ for RR graph
$\lambda = 0.15$ for 2D grid graph



- Fairly good match between theory and experiments, even for 2D grid.
- ℓ_1-LinR behave similarly as ℓ_1-LogR for precision and recall.

Excellent match even for loopy graphs!

Experimental Results

■ **Accurate Sample Complexity Prediction**

Ising model: RR graph, $K_0 = 0.4$, $d = 3$

Estimators: ℓ_1-LinR and ℓ_1-LogR with $\lambda = 0.3$

 # samples scaling value

$$M = c \log N$$

Theoretical Prediction $c_0 \left(\lambda = 0.3, K_0\right) \approx 19.41$

- **Precision**

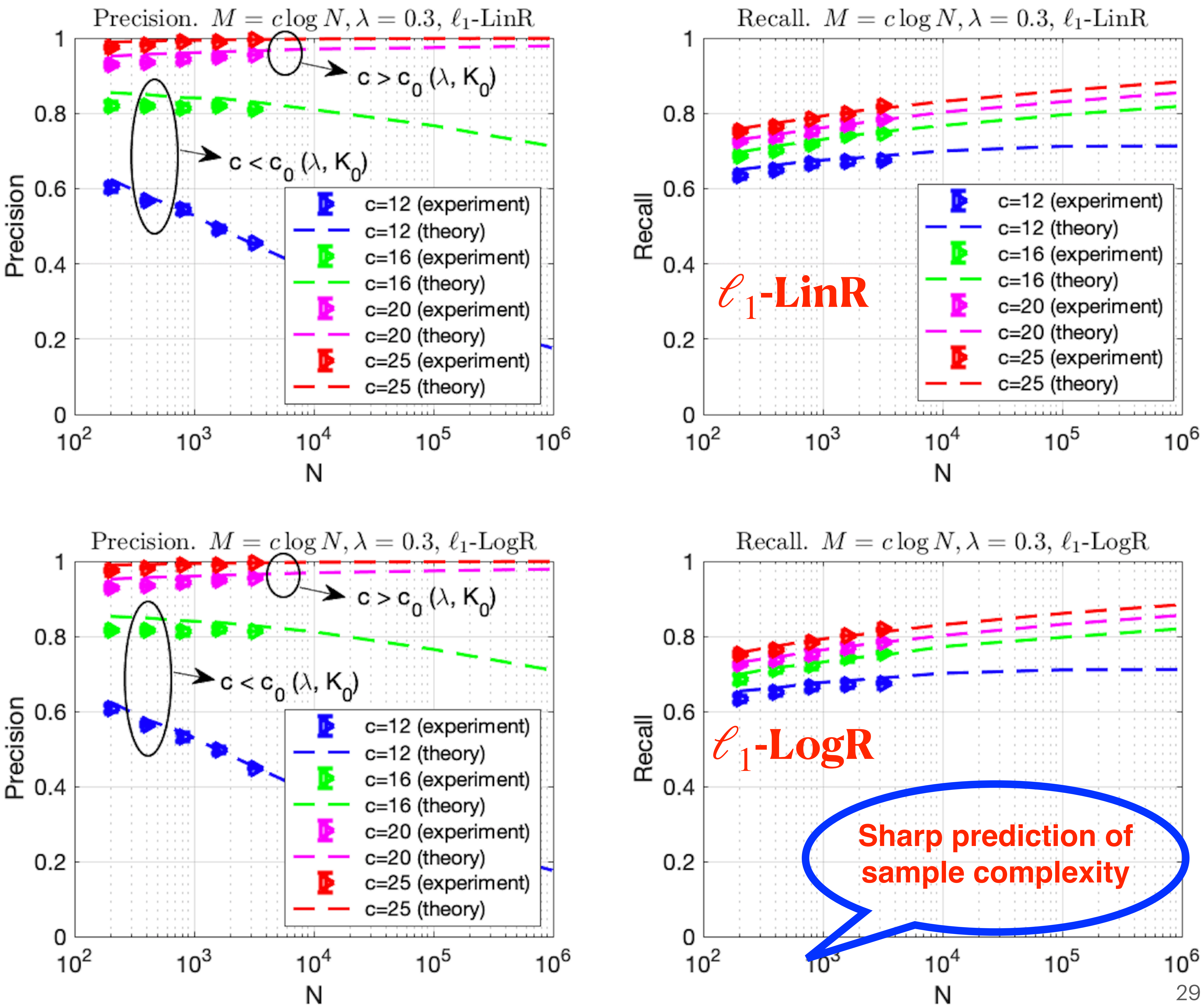
  $c > c_0 \left(\lambda, K_0\right)$: increasing to 1 as $N \to \infty$

  $c < c_0 \left(\lambda, K_0\right)$: decreasing to 0 as $N \to \infty$

- **Recall**

  Increasing to 1 as $N \to \infty$

The prediction of the sample complexity is accurate for ℓ_1-LinR (and ℓ_1-LinR) !



Summary

- **Our work**
  - **A unified statistical mechanics framework for precisely investigating the *typical* learning performances of ℓ_1-regularized M-estimators. In particular,**

— **Revealing that ℓ_1-LinR is model selection consistent with same order of sample complexity as ℓ_1-LogR**

— **Providing accurate predictions of both the sample complexity and *non-asymptotic* learning performances**

— **An excellent agreement between the theoretical predictions and experimental results, even for graphs with many loops, which supports our findings.**

Summary

- **Our work**
  - **A unified statistical mechanics framework for precisely investigating the *typical* learning performances of ℓ_1-regularized M-estimators. In particular,**

— **Revealing that ℓ_1-LinR is model selection consistent with same order of sample complexity as ℓ_1-LogR**

— **Providing accurate predictions of both the sample complexity and *non-asymptotic* learning performances**

— **An excellent agreement between the theoretical predictions and experimental results, even for graphs with many loops, which supports our findings.**

- **Main Limitations**
  - **Several Key assumptions are made in theoretical analysis, for example:**

— **Paramagnetic assumption of the Ising model**

— **Typical tree-like RR graph is considered**

  - **Overcoming such limitations is an important direction for future work**

Thank you!

Q&A