Mean-Field fluctuations in Quadratic interaction models in low SNR regimes

Nabarun Deb
Statistics and Probability Seminar
Department of Mathematics and Statistics
Boston University

Joint work with Seunghyun (Sky) Li and Sumit Mukherjee

Outline

Ising Model

Consider

$$\frac{d\mathbb{P}}{d\prod_{i=1}^{n}\mu_i}(\boldsymbol{\beta}) := \frac{1}{Z_n}\exp\left(\frac{1}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + \mathbf{c}^\top \boldsymbol{\beta}\right),$$

μ_i supported on $[-1,1]$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$, and field vector $\mathbf{c} = (c_1, \ldots, c_n)$.

Ising Model

Consider

$$\frac{d\mathbb{P}}{d\prod_{i=1}^{n}\mu_i}(\boldsymbol{\beta}) := \frac{1}{Z_n}\exp\left(\frac{1}{2}\boldsymbol{\beta}^\top \mathbf{A}_n\boldsymbol{\beta} + \mathbf{c}^\top\boldsymbol{\beta}\right),$$

μ_i supported on $[-1,1]$, $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_n)$, and field vector $\mathbf{c} = (c_1,\ldots,c_n)$.

1. $\mathbf{A}_n = 0$ implies β_is are independent
2. $\mathbf{A}_n(i,j) > 0$ implies that sites i and j are inclined to align in the same direction.
3. Large c_i implies site i is more likely to take larger values
4. Interaction matrix - $\mathbf{A}_n$, Partition function - Z_n

Our goal

To study the asymptotic distribution of

$$T_n = \mathbf{q}^\top(\boldsymbol{\beta} - ??), \quad \|\mathbf{q}\| = 1$$

for certain linear combinations.

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2} \boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B \sum_{i=1}^{n} \beta_i \right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. θ — temperature parameter.

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B\sum_{i=1}^{n} \beta_i\right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. θ — temperature parameter.

- In social networks, to study trends in opinions (voting choices), where G_N could be determined by "friendships" within the network.

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B\sum_{i=1}^{n} \beta_i\right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. θ — temperature parameter.

- In social networks, to study trends in opinions (voting choices), where G_N could be determined by "friendships" within the network.
- (Phase transition) Modeling ferromagnetic properties (i.e., sharp change in magnetic properties of magnetic materials when heated beyond a certain (Curie) temperature — coded into $\mathbf{A}_n$).

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2} \boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B \sum_{i=1}^{n} \beta_i \right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector **c** is constant at B. θ — temperature parameter.

- In social networks, to study trends in opinions (voting choices), where G_N could be determined by "friendships" within the network.
- (Phase transition) Modeling ferromagnetic properties (i.e., sharp change in magnetic properties of magnetic materials when heated beyond a certain (Curie) temperature — coded into $\mathbf{A}_n$).

Natural question

Behavior of the "sufficient statistic"

$$n^{-??} \sum_{i=1}^{n} (\beta_i - ??) \xrightarrow{d} \ldots$$

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2} \boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B \sum_{i=1}^n \beta_i \right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. θ — temperature parameter.

- In social networks, to study trends in opinions (voting choices), where G_N could be determined by "friendships" within the network.
- (Phase transition) Modeling ferromagnetic properties (i.e., sharp change in magnetic properties of magnetic materials when heated beyond a certain (Curie) temperature — coded into $\mathbf{A}_n$).

Natural question

Behavior of the "sufficient statistic"

$$n^{-??} \sum_{i=1}^n (\beta_i - ??) \xrightarrow{d} \ldots$$

CLT for Sufficient statistic $\leftrightarrow$ CLT for MLE of B.

First motivating example

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B\sum_{i=1}^n \beta_i\right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. θ — temperature parameter.

- In social networks, to study trends in opinions (voting choices), where G_N could be determined by "friendships" within the network.
- (Phase transition) Modeling ferromagnetic properties (i.e., sharp change in magnetic properties of magnetic materials when heated beyond a certain (Curie) temperature — coded into $\mathbf{A}_n$).

Natural question

Behavior of the "sufficient statistic"

$$n^{-??}\sum_{i=1}^n (\beta_i - ??) \xrightarrow{d} \ldots$$

CLT for Sufficient statistic $\leftrightarrow$ CLT for MLE of B. What is the centering and scaling in terms of $\mathbf{A}_n$ and B?

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Here
  - $\mathbf{y} \in \mathbb{R}^n$ is the observed data,

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Here
  - $\mathbf{y} \in \mathbb{R}^n$ is the observed data,
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the (non-random) design matrix;

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Here
  - $\mathbf{y} \in \mathbb{R}^n$ is the observed data,
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the (non-random) design matrix;
  - $\boldsymbol{\beta}$ is the coefficient vector;

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Here
  - $\mathbf{y} \in \mathbb{R}^n$ is the observed data,
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the (non-random) design matrix;
  - $\boldsymbol{\beta}$ is the coefficient vector;
  - $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of i.i.d. $N(0, \sigma^2)$ random variables.

Second motivating example: Standard linear regression

- Suppose we have the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Here
  - $\mathbf{y} \in \mathbb{R}^n$ is the observed data,
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the (non-random) design matrix;
  - $\boldsymbol{\beta}$ is the coefficient vector;
  - $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of i.i.d. $N(0, \sigma^2)$ random variables.
- Assume that σ is known (and equals 1), but the parameter $\boldsymbol{\beta} \in [-1, 1]^p$ is unknown.

Bayesian perspective: put a prior

- Suppose π is a probability distribution on $[-1, 1]$.

Bayesian perspective: put a prior

- Suppose π is a probability distribution on $[-1, 1]$.
- Let $\boldsymbol{\beta}$ have a product prior, under which $\{\beta_i\}_{1\le i\le p} \overset{i.i.d.}{\sim} \pi$.

Bayesian perspective: put a prior

- Suppose π is a probability distribution on $[-1, 1]$.
- Let $\boldsymbol{\beta}$ have a product prior, under which $\{\beta_i\}_{1 \le i \le p} \overset{i.i.d.}{\sim} \pi$.
- Then the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ of $\boldsymbol{\beta}$ given $\mathbf{y}$ is given by

$$\frac{d\mu}{d\pi^{\otimes p}}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right)$$

Bayesian perspective: put a prior

- Suppose π is a probability distribution on $[-1, 1]$.

- Let $\boldsymbol{\beta}$ have a product prior, under which $\{\beta_i\}_{1 \le i \le p} \overset{i.i.d.}{\sim} \pi$.

- Then the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ of $\boldsymbol{\beta}$ given $\mathbf{y}$ is given by

$$\frac{d\mu}{d\pi^{\otimes p}}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right)$$

- The posterior is a quadratic interaction model where interaction matrix comes from $\sigma^{-2}\mathbf{X}^\top\mathbf{X}$ and the field vector $\sigma^{-2}\mathbf{X}^\top\mathbf{y}$ is non-constant.

Bayesian perspective: put a prior

- Suppose π is a probability distribution on $[-1, 1]$.
- Let $\boldsymbol{\beta}$ have a product prior, under which $\{\beta_i\}_{1 \le i \le p} \overset{i.i.d.}{\sim} \pi$.
- Then the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ of $\boldsymbol{\beta}$ given $\mathbf{y}$ is given by

$$\frac{d\mu}{d\pi^{\otimes p}}(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right)$$

- The posterior is a quadratic interaction model where interaction matrix comes from $\sigma^{-2}\mathbf{X}^\top\mathbf{X}$ and the field vector $\sigma^{-2}\mathbf{X}^\top\mathbf{y}$ is non-constant.
- We want to understand the behavior of this posterior distribution.

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime — no posterior contraction + no Laplace method

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime — no posterior contraction + no Laplace method

- To simplify, let's focus on the very special case $p = 1$:

$$Y_1, \cdots, Y_n \overset{iid}{\sim} N(\theta, 1) \text{ vs. } Y_1, \cdots, Y_n \overset{iid}{\sim} N\Big(\frac{h}{\sqrt{n}}, 1\Big).$$

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime — no posterior contraction + no Laplace method

- To simplify, let's focus on the very special case $p = 1$:

$$Y_1, \cdots, Y_n \overset{iid}{\sim} N(\theta, 1) \text{ vs. } Y_1, \cdots, Y_n \overset{iid}{\sim} N\Big(\frac{h}{\sqrt{n}}, 1\Big).$$

- The first example has design vector $\mathbf{1}_n$, which has norm $\sqrt{n}$.

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime — no posterior contraction + no Laplace method

- To simplify, let's focus on the very special case $p = 1$:

$$Y_1, \cdots, Y_n \overset{iid}{\sim} N(\theta, 1) \text{ vs. } Y_1, \cdots, Y_n \overset{iid}{\sim} N\Big(\frac{h}{\sqrt{n}}, 1\Big).$$

- The first example has design vector $\mathbf{1}_n$, which has norm $\sqrt{n}$.

- Hence this is high SNR.

High vs Low Signal to Noise

- Most of the existing literature on this problem focuses on the setting where the operator norm $\|\mathbf{X}\| \to \infty$, which we will refer to as high signal-to-noise (SNR) regime — posterior contraction + Laplace method

- A different regime is the case when $\|\mathbf{X}\|_2 = O(1)$, which results in the low SNR regime — no posterior contraction + no Laplace method

- To simplify, let's focus on the very special case $p = 1$:

$$Y_1, \cdots, Y_n \overset{iid}{\sim} N(\theta, 1) \text{ vs. } Y_1, \cdots, Y_n \overset{iid}{\sim} N\Big(\frac{h}{\sqrt{n}}, 1\Big).$$

- The first example has design vector $\mathbf{1}_n$, which has norm $\sqrt{n}$.

- Hence this is high SNR. In this case consistent estimation of θ is possible, and the posterior contracts at rate $\sqrt{n}$ (Bernstein-von-Mises).

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.

- Hence this is low SNR.

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.

- Hence this is low SNR. In this case consistent estimation of h is impossible, and the posterior does not contract.

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.
- Hence this is low SNR. In this case consistent estimation of h is impossible, and the posterior does not contract.
- In a very similar manner, it is expected that more generally when $\|\mathbf{X}\|_2 = O(1)$, consistent estimation of $\boldsymbol{\beta}$ is impossible.

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.
- Hence this is low SNR. In this case consistent estimation of h is impossible, and the posterior does not contract.
- In a very similar manner, it is expected that more generally when $\|\mathbf{X}\|_2 = O(1)$, consistent estimation of $\boldsymbol{\beta}$ is impossible.
- This is known in the special case when the design matrix $\mathbf{X}$ has IID Gaussian entries, given by $X_{ij} \sim N\left(0, \frac{1}{n}\right)$ (Barbier et al., IEEE-20).

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.
- Hence this is low SNR. In this case consistent estimation of h is impossible, and the posterior does not contract.
- In a very similar manner, it is expected that more generally when $\|\mathbf{X}\|_2 = O(1)$, consistent estimation of $\boldsymbol{\beta}$ is impossible.
- This is known in the special case when the design matrix $\mathbf{X}$ has IID Gaussian entries, given by $X_{ij} \sim N\left(0, \frac{1}{n}\right)$ (Barbier et al., IEEE-20).
- Also true in the Gaussian sequence model $Y_i = \beta_i + \epsilon_i$, where the design $\mathbf{X}$ is the identity matrix

High vs Low Signal to Noise

- The second example has design vector $n^{-1/2}\mathbf{1}_n$, which has norm 1.

- Hence this is low SNR. In this case consistent estimation of h is impossible, and the posterior does not contract.

- In a very similar manner, it is expected that more generally when $\|\mathbf{X}\|_2 = O(1)$, consistent estimation of $\boldsymbol{\beta}$ is impossible.

- This is known in the special case when the design matrix $\mathbf{X}$ has IID Gaussian entries, given by $X_{ij} \sim N\left(0, \frac{1}{n}\right)$ (Barbier et al., IEEE-20).

- Also true in the Gaussian sequence model $Y_i = \beta_i + \epsilon_i$, where the design $\mathbf{X}$ is the identity matrix

- In this talk we will focus on this low SNR regime, which is a (non-trivial) extension of the LAN regime of classical statistics.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.
- At a high level, because of the scaling choice, the prior never gets washed away, but competes with the likelihood.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.
- At a high level, because of the scaling choice, the prior never gets washed away, but competes with the likelihood.
- Thus one expects the prior to show up in non-trivial ways in the asymptotics, affecting LLNs and CLTs.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.
- At a high level, because of the scaling choice, the prior never gets washed away, but competes with the likelihood.
- Thus one expects the prior to show up in non-trivial ways in the asymptotics, affecting LLNs and CLTs.

Natural question

A high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.
- At a high level, because of the scaling choice, the prior never gets washed away, but competes with the likelihood.
- Thus one expects the prior to show up in non-trivial ways in the asymptotics, affecting LLNs and CLTs.

Natural question

A high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior. Note $\mathbf{q}^\top \boldsymbol{\beta}$ is the posterior prediction when the covariate is $\mathbf{q}$.

What can we expect?

- As suggested in the last slide, we expect the posterior to not contract, and do not expect Bernstein von Mises type CLT to hold.
- At a high level, because of the scaling choice, the prior never gets washed away, but competes with the likelihood.
- Thus one expects the prior to show up in non-trivial ways in the asymptotics, affecting LLNs and CLTs.

Natural question

A high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \ldots \text{ conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior. Note $\mathbf{q}^\top \boldsymbol{\beta}$ is the posterior prediction when the covariate is $\mathbf{q}$.

Ideally the analysis will apply to both deterministic and random $\mathbf{X}$, and allows for possibly dependent entries.

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —
  - Phase transitions in the fluctuations of $\sum_{i=1}^n \beta_i$ and universal limit laws (free of $\mathbf{A}_n$).

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —
  - Phase transitions in the fluctuations of $\sum_{i=1}^{n} \beta_i$ and universal limit laws (free of $\mathbf{A}_n$).
  - (Potential) non-universality and lack of phase transitions for $\sum_{i=1}^{n} q_i \beta_i$ for some other $\mathbf{q}$.

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —
  - Phase transitions in the fluctuations of $\sum_{i=1}^n \beta_i$ and universal limit laws (free of $\mathbf{A}_n$).
  - (Potential) non-universality and lack of phase transitions for $\sum_{i=1}^n q_i\beta_i$ for some other $\mathbf{q}$.
- In Bayesian linear regression (only in high temperature)
  - Effect of prior in the centering (and limiting variance).

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —
  - Phase transitions in the fluctuations of $\sum_{i=1}^{n} \beta_i$ and universal limit laws (free of $\mathbf{A}_n$).
  - (Potential) non-universality and lack of phase transitions for $\sum_{i=1}^{n} q_i\beta_i$ for some other $\mathbf{q}$.
- In Bayesian linear regression (only in high temperature)
  - Effect of prior in the centering (and limiting variance).
  - Bayesian low SNR credible sets

What to expect?

- For two-spin Ising model with constant magnetization on approximately regular graphs —
  - Phase transitions in the fluctuations of $\sum_{i=1}^{n} \beta_i$ and universal limit laws (free of $\mathbf{A}_n$).
  - (Potential) non-universality and lack of phase transitions for $\sum_{i=1}^{n} q_i\beta_i$ for some other $\mathbf{q}$.
- In Bayesian linear regression (only in high temperature)
  - Effect of prior in the centering (and limiting variance).
  - Bayesian low SNR credible sets
- (Not in the talk) General Berry-Esseen bounds — `https://arxiv.org/abs/2005.00710` and `https://arxiv.org/abs/2503.21152`.

Outline

Proposed approach: Naive Mean Field

- One frequently used method to “understand” the posterior μ is the naive mean field (NFM) variational approximation.

Proposed approach: Naive Mean Field

- One frequently used method to "understand" the posterior μ is the naive mean field (NFM) variational approximation.

- This is obtained by projecting μ to the space of product measures on $[-1, 1]^p$, using the Kullback-Leibler divergence.

Proposed approach: Naive Mean Field

- One frequently used method to "understand" the posterior μ is the naive mean field (NFM) variational approximation.
- This is obtained by projecting μ to the space of product measures on $[-1, 1]^p$, using the Kullback-Leibler divergence.

Typical result (Basak and Mukherjee, 2015)

Let P_{prod} denote the space of product measures on $[-1, 1]^p$, then under the Mean-Field assumption $\|\mathbf{A}\|_F^2 = o(p)$ (Frobenius norm of interaction matrix), the following holds:

$$\inf_{\nu \in P_{\text{prod}}} \text{KL}(\nu | \mu_{\mathbf{y},\mathbf{X},\pi}) = o(p).$$

Proposed approach: Naive Mean Field

- One frequently used method to “understand” the posterior μ is the naive mean field (NFM) variational approximation.
- This is obtained by projecting μ to the space of product measures on $[-1, 1]^p$, using the Kullback-Leibler divergence.

Typical result (Basak and Mukherjee, 2015)

Let P_{prod} denote the space of product measures on $[-1, 1]^p$, then under the Mean-Field assumption $\|\mathbf{A}\|_F^2 = o(p)$ (Frobenius norm of interaction matrix), the following holds:

$$\inf_{\nu \in P_{\text{prod}}} \text{KL}(\nu | \mu_{\mathbf{y},\mathbf{X},\pi}) = o(p).$$

- The optimization resulting from the above projection is usually easy to compute.

Proposed approach: Naive Mean Field

- One frequently used method to "understand" the posterior μ is the naive mean field (NFM) variational approximation.

- This is obtained by projecting μ to the space of product measures on $[-1,1]^p$, using the Kullback-Leibler divergence.

Typical result (Basak and Mukherjee, 2015)

Let P_{prod} denote the space of product measures on $[-1,1]^p$, then under the Mean-Field assumption $\|\mathbf{A}\|_F^2 = o(p)$ (Frobenius norm of interaction matrix), the following holds:

$$\inf_{\nu \in P_{\text{prod}}} \mathrm{KL}(\nu | \mu_{\mathbf{y},\mathbf{X},\pi}) = o(p).$$

- The optimization resulting from the above projection is usually easy to compute.

- NMF is computationally efficient (see Jain et al. 2018) as opposed to MCMC based methods, particularly in high dimensions.

How does the projection look like?

- Write

$$\sigma^2 \mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2 \mathbf{c} = \mathbf{X}'\mathbf{y}.$$

How does the projection look like?

- Write

$$\sigma^2 \mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2 \mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$.

How does the projection look like?

- Write
$$\sigma^2 \mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2 \mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$. This assumption is not needed for most of our results, but simplifies the argument+notations.

How does the projection look like?

- Write
$$\sigma^2 \mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2 \mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$. This assumption is not needed for most of our results, but simplifies the argument+notations.

- Define a linear+quadratic of the prior π by setting
$$\frac{d\pi_{\theta,d}}{d\pi}(w) = e^{\theta x - \frac{d}{2}w^2 - \alpha(\theta)},$$
where
$$\alpha(\theta) := \alpha_{\pi,d}(\theta) = \log \int_{[-1,1]} e^{\theta w - \frac{d}{2}w^2} d\pi(w).$$

How does the projection look like?

- Write
$$\sigma^2 \mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2 \mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$. This assumption is not needed for most of our results, but simplifies the argument+notations.

- Define a linear+quadratic of the prior π by setting
$$\frac{d\pi_{\theta,d}}{d\pi}(w) = e^{\theta x - \frac{d}{2}w^2 - \alpha(\theta)},$$
where
$$\alpha(\theta) := \alpha_{\pi,d}(\theta) = \log \int_{[-1,1]} e^{\theta w - \frac{d}{2}w^2} d\pi(w).$$

- Then $\alpha''(\theta) = Var_{\pi_{\theta,d}}(W) > 0$, and so $\alpha'(\cdot) : \mathbb{R} \mapsto (-1,1)$ is strictly monotone.

How does the projection look like?

- Write
$$\sigma^2\mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2\mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$. This assumption is not needed for most of our results, but simplifies the argument+notations.

- Define a linear+quadratic of the prior π by setting
$$\frac{d\pi_{\theta,d}}{d\pi}(w) = e^{\theta x - \frac{d}{2}w^2 - \alpha(\theta)},$$
where
$$\alpha(\theta) := \alpha_{\pi,d}(\theta) = \log \int_{[-1,1]} e^{\theta w - \frac{d}{2}w^2} d\pi(w).$$

- Then $\alpha''(\theta) = Var_{\pi_{\theta,d}}(W) > 0$, and so $\alpha'(\cdot) : \mathbb{R} \mapsto (-1,1)$ is strictly monotone. Note that $\alpha(\cdot)$ depends on the prior π.

How does the projection look like?

- Write
$$\sigma^2\mathbf{X}'\mathbf{X} = A + D, \text{ and } \sigma^2\mathbf{c} = \mathbf{X}'\mathbf{y}.$$

- Suppose for simplicity that all diagonal entries of $\mathbf{X}'\mathbf{X}$ are the same, i.e. $D = d\mathbf{I}$. This assumption is not needed for most of our results, but simplifies the argument+notations.

- Define a linear+quadratic of the prior π by setting
$$\frac{d\pi_{\theta,d}}{d\pi}(w) = e^{\theta x - \frac{d}{2}w^2 - \alpha(\theta)},$$
where
$$\alpha(\theta) := \alpha_{\pi,d}(\theta) = \log \int_{[-1,1]} e^{\theta w - \frac{d}{2}w^2} d\pi(w).$$

- Then $\alpha''(\theta) = Var_{\pi_{\theta,d}}(W) > 0$, and so $\alpha'(\cdot) : \mathbb{R} \mapsto (-1,1)$ is strictly monotone. Note that $\alpha(\cdot)$ depends on the prior π.

- Let $h(\cdot) : (-1,1) \mapsto \mathbb{R}$ be the inverse of the function $\alpha'(\cdot)$,

How does the projection look like?

- Consider the following auxiliary optimization problem

$$\sup_{\mathbf{u}\in[-1,1]^p}\left\{-\frac{1}{2}\mathbf{u}'A\mathbf{u}+\mathbf{c}'\mathbf{u}-\sum_{i=1}^{p}\mathrm{KL}(\pi_{h(u_i),d}|\pi_{0,d})\right\}. \qquad (1)$$

How does the projection look like?

- Consider the following auxiliary optimization problem

$$\sup_{\mathbf{u}\in[-1,1]^p}\left\{-\frac{1}{2}\mathbf{u}'A\mathbf{u}+\mathbf{c}'\mathbf{u}-\sum_{i=1}^{p}\mathrm{KL}(\pi_{h(u_i),d}|\pi_{0,d})\right\}. \qquad (1)$$

- Then any optimizer over the space of product measures is of the form

$$\prod_{i=1}^{p}\pi_{\theta_i^{\mathrm{opt}},d},$$

where $\boldsymbol{\theta}^{\mathrm{opt}}$ is chosen such that $\mathbb{E}_{\pi_{\boldsymbol{\theta}^{\mathrm{opt}},d}}(W)=\mathbf{u}^{\mathrm{opt}}$, and $\mathbf{u}^{\mathrm{opt}}$ is a global optimizer to (1). $\mathbf{u}^{\mathrm{opt}}$ will serve as the centering term.

How does the projection look like?

- Consider the following auxiliary optimization problem

$$\sup_{\mathbf{u}\in[-1,1]^p} \left\{ -\frac{1}{2}\mathbf{u}' A\mathbf{u} + \mathbf{c}'\mathbf{u} - \sum_{i=1}^{p} \mathrm{KL}(\pi_{h(u_i),d} | \pi_{0,d}) \right\}. \qquad (1)$$

- Then any optimizer over the space of product measures is of the form

$$\prod_{i=1}^{p} \pi_{\theta_i^{\text{opt}},d},$$

where $\boldsymbol{\theta}^{\text{opt}}$ is chosen such that $\mathbb{E}_{\pi_{\boldsymbol{\theta}^{\text{opt}},d}}(W) = \mathbf{u}^{\text{opt}}$, and $\mathbf{u}^{\text{opt}}$ is a global optimizer to (1). $\mathbf{u}^{\text{opt}}$ will serve as the centering term.

- Thus we have reduced the optimization over the space of measures to an optimization over the space of mean vectors $\mathbf{u}$.

How does the projection look like?

- Consider the following auxiliary optimization problem

$$\sup_{\mathbf{u}\in[-1,1]^p} \left\{ -\frac{1}{2}\mathbf{u}' A\mathbf{u} + \mathbf{c}'\mathbf{u} - \sum_{i=1}^{p} \mathrm{KL}(\pi_{h(u_i),d} | \pi_{0,d}) \right\}. \quad (1)$$

- Then any optimizer over the space of product measures is of the form

$$\prod_{i=1}^{p} \pi_{\theta_i^{\mathrm{opt}},d},$$

where $\boldsymbol{\theta}^{\mathrm{opt}}$ is chosen such that $\mathbb{E}_{\pi_{\boldsymbol{\theta}^{\mathrm{opt}},d}}(W) = \mathbf{u}^{\mathrm{opt}}$, and $\mathbf{u}^{\mathrm{opt}}$ is a global optimizer to (1). $\mathbf{u}^{\mathrm{opt}}$ will serve as the centering term.

- Thus we have reduced the optimization over the space of measures to an optimization over the space of mean vectors $\mathbf{u}$. We will refer to (1) as the mean-field prediction formula.

How does the projection look like?

- Consider the following auxiliary optimization problem

$$\sup_{\mathbf{u}\in[-1,1]^p} \left\{ -\frac{1}{2}\mathbf{u}' A\mathbf{u} + \mathbf{c}'\mathbf{u} - \sum_{i=1}^{p} \mathrm{KL}(\pi_{h(u_i),d} | \pi_{0,d}) \right\}. \quad (1)$$

- Then any optimizer over the space of product measures is of the form

$$\prod_{i=1}^{p} \pi_{\theta_i^{\text{opt}},d},$$

where $\boldsymbol{\theta}^{\text{opt}}$ is chosen such that $\mathbb{E}_{\pi_{\boldsymbol{\theta}^{\text{opt}},d}}(W) = \mathbf{u}^{\text{opt}}$, and $\mathbf{u}^{\text{opt}}$ is a global optimizer to (1). $\mathbf{u}^{\text{opt}}$ will serve as the centering term.

- Thus we have reduced the optimization over the space of measures to an optimization over the space of mean vectors $\mathbf{u}$. We will refer to (1) as the mean-field prediction formula.

- Note that $\mathbf{u}^{\text{opt}}$ depends on $(\mathbf{y}, \mathbf{X}, \pi)$, which we omit in the notation.

Uniqueness of optimizer

- In general the mean-field prediction formula does not have a unique solution.

Uniqueness of optimizer

- In general the mean-field prediction formula does not have a unique solution.

- In M.-Sen, JMLR-2022, Lacker-M-Yeung, IMRN-2024 we provide sufficient conditions on $(\mathbf{A}, \pi)$ under which there is a unique well-separated global optimizer.

Uniqueness of optimizer

- In general the mean-field prediction formula does not have a unique solution.

- In M.-Sen, JMLR-2022, Lacker-M-Yeung, IMRN-2024 we provide sufficient conditions on $(\mathbf{A}, \pi)$ under which there is a unique well-separated global optimizer.

- One sufficient condition is that $\|\mathbf{A}\|_2 \leq 1 - \rho$, for some ρ (free of n, p).

Uniqueness of optimizer

- In general the mean-field prediction formula does not have a unique solution.

- In M.-Sen, JMLR-2022, Lacker-M-Yeung, IMRN-2024 we provide sufficient conditions on $(\mathbf{A}, \pi)$ under which there is a unique well-separated global optimizer.

- One sufficient condition is that $\|\mathbf{A}\|_2 \leq 1 - \rho$, for some ρ (free of n, p).

- This corresponds to the so called high temperature regime of statistical physics.

Uniqueness of optimizer

- In general the mean-field prediction formula does not have a unique solution.
- In M.-Sen, JMLR-2022, Lacker-M-Yeung, IMRN-2024 we provide sufficient conditions on $(\mathbf{A}, \pi)$ under which there is a unique well-separated global optimizer.
- One sufficient condition is that $\|\mathbf{A}\|_2 \leq 1 - \rho$, for some ρ (free of n, p).
- This corresponds to the so called high temperature regime of statistical physics.
- As a comment, typically they assume the somewhat stronger assumption $\|\mathbf{A}\|_\infty \leq 1 - \rho$ particularly for quantitative bounds (concentration inequalities).

Outline

Existing work

Recall

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B \sum_{i=1}^{n} \beta_i\right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. $\theta > 0$ — temperature parameter.

- Most of existing work analyzing $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$ focuses exclusively on the Curie-Weiss model (see Ellis-Newman (1978), Chatterjee-Shao (2011), Shao-Zhang (2017)), where $\mathbf{A}_n$ is the (scaled) complete graph.

Existing work

Recall

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B\sum_{i=1}^n \beta_i\right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. $\theta > 0$ — temperature parameter.

- Most of existing work analyzing $T_n = n^{-1}\sum_{i=1}^n \beta_i$ focuses exclusively on the Curie-Weiss model (see Ellis-Newman (1978), Chatterjee-Shao (2011), Shao-Zhang (2017)), where $\mathbf{A}_n$ is the (scaled) complete graph.

- It is thus perhaps not surprising that $(X_1, \ldots, X_n)$ can be expressed as an exact mixture of i.i.d. laws. This observation (implicitly) plays an important role in existing analysis

Existing work

Recall

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2} \boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B \sum_{i=1}^{n} \beta_i \right),$$

$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. $\theta > 0$ — temperature parameter.

- Most of existing work analyzing $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$ focuses exclusively on the Curie-Weiss model (see Ellis-Newman (1978), Chatterjee-Shao (2011), Shao-Zhang (2017)), where $\mathbf{A}_n$ is the (scaled) complete graph.

- It is thus perhaps not surprising that $(X_1, \ldots, X_n)$ can be expressed as an exact mixture of i.i.d. laws. This observation (implicitly) plays an important role in existing analysis

Existing work

Recall

$$\mathbb{P}(\boldsymbol{\beta}) := \frac{1}{Z_n} \exp\left(\frac{\theta}{2}\boldsymbol{\beta}^\top \mathbf{A}_n \boldsymbol{\beta} + B\sum_{i=1}^{n}\beta_i\right),$$

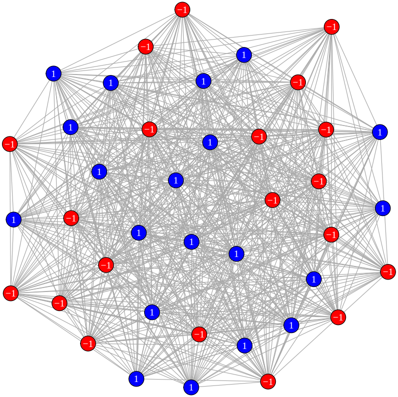
$\beta_i \in \{-1, 1\}$ binary. The field vector $\mathbf{c}$ is constant at B. $\theta > 0$ — temperature parameter.

- Most of existing work analyzing $T_n = n^{-1}\sum_{i=1}^{n}\beta_i$ focuses exclusively on the Curie-Weiss model (see Ellis-Newman (1978), Chatterjee-Shao (2011), Shao-Zhang (2017)), where $\mathbf{A}_n$ is the (scaled) complete graph.

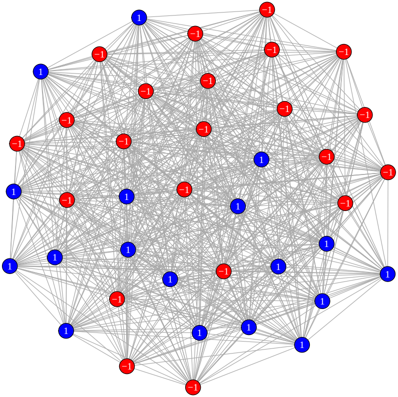
- It is thus perhaps not surprising that $(X_1, \ldots, X_n)$ can be expressed as an exact mixture of i.i.d. laws. This observation (implicitly) plays an important role in existing analysis

We focus on other approximately regular graphs with diverging degree satisfying a Mean-Field condition.

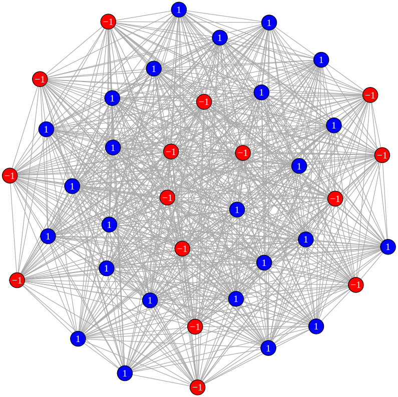
$N = 40,\ \kappa = 0.9,\ \theta = 0.5,\ B = 0$



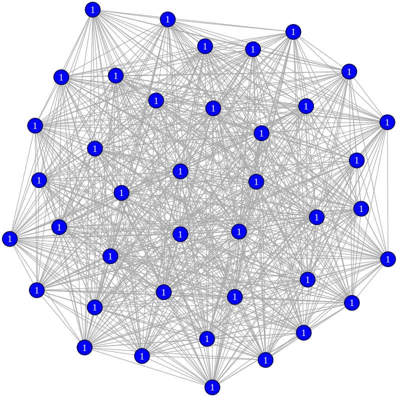
$N = 40,\ \kappa = 0.9,\ \theta = 0.5,\ B = 0$



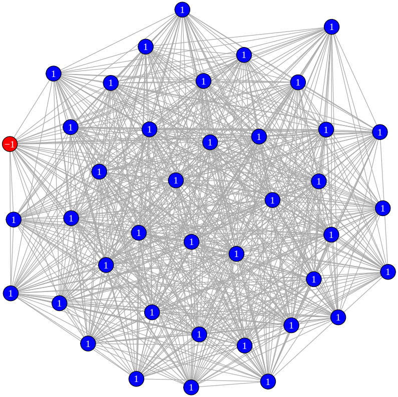
$N = 40,\ \kappa = 0.9,\ \theta = 0.5,\ B = 0$



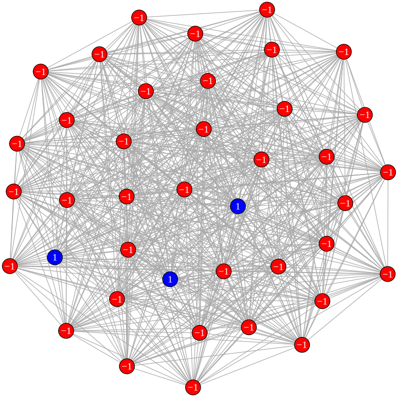
$N = 40,\ \kappa = 0.9,\ \theta = 1.5,\ B = 0$



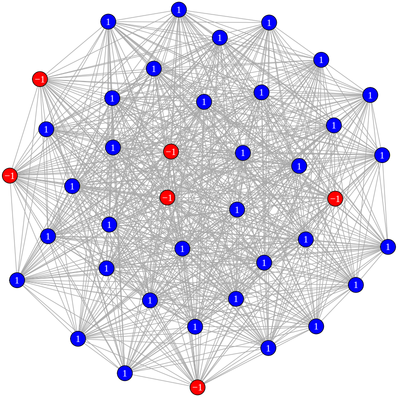
$N = 40,\ \kappa = 0.9,\ \theta = 1.5,\ B = 0$



$N = 40,\ \kappa = 0.9,\ \theta = 1.5,\ B = 0$



$N = 40,\ \kappa = 0.9,\ \theta = 1.5,\ B = 0$



Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

- Properties of above equation —
  - If $0 \leq \theta \leq 1$, $B = 0$, then unique solution $t_{\theta,B} = 0$.

Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

- Properties of above equation —
  - If $0 \leq \theta \leq 1$, $B = 0$, then unique solution $t_{\theta,B} = 0$.
  - If $\theta > 1$, $B = 0$, then unique positive solution $t_{\theta,B}$.

Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

- Properties of above equation —
  - If $0 \leq \theta \leq 1$, $B = 0$, then unique solution $t_{\theta,B} = 0$.
  - If $\theta > 1$, $B = 0$, then unique positive solution $t_{\theta,B}$.
  - If $\theta > 0$, $B > 0$, then unique positive solution $t_{\theta,B}$ (negative soln. for $B < 0$).

Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

- Properties of above equation —
  - If $0 \leq \theta \leq 1$, $B = 0$, then unique solution $t_{\theta,B} = 0$.
  - If $\theta > 1$, $B = 0$, then unique positive solution $t_{\theta,B}$.
  - If $\theta > 0$, $B > 0$, then unique positive solution $t_{\theta,B}$ (negative soln. for $B < 0$).
- Can be extended to the case beyond two-spins as long as $\mathbf{A}_n$ is "approximately" regular.

Mean-Field optimizers in the two-spin case

- The coordinates of the Mean-Field optimizer $\mathbf{u}^{\text{opt}}$ splits into solutions of n independent stationarity conditions given by

$$t = \tanh(\theta t + B).$$

- Properties of above equation —
  - If $0 \leq \theta \leq 1$, $B = 0$, then unique solution $t_{\theta,B} = 0$.
  - If $\theta > 1$, $B = 0$, then unique positive solution $t_{\theta,B}$.
  - If $\theta > 0$, $B > 0$, then unique positive solution $t_{\theta,B}$ (negative soln. for $B < 0$).
- Can be extended to the case beyond two-spins as long as $\mathbf{A}_n$ is "approximately" regular.

Does $n^{-??} \sum_{i=1}^{n} (\beta_i - t_{\theta,B})$ converge?

Law of Large numbers for $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$

Basak-Mukherjee (2015), D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of (scaled) "approximately" d_n-regular graphs, with $d_n \overset{n\to\infty}{\longrightarrow} \infty$, then:

Law of Large numbers for $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$

Basak-Mukherjee (2015), D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of (scaled) “approximately” d_n-regular graphs, with $d_n \xrightarrow{n\to\infty} \infty$, then:

1. If $B > 0$, then $T_n \xrightarrow{d} t_{\theta,B}$

Law of Large numbers for $T_n = n^{-1}\sum_{i=1}^n \beta_i$

Basak-Mukherjee (2015), D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of (scaled) "approximately" d_n-regular graphs, with $d_n \stackrel{n\to\infty}{\longrightarrow} \infty$, then:

1. If $B > 0$, then $T_n \stackrel{d}{\longrightarrow} t_{\theta,B}$
2. If $\theta \le 1, B = 0$, then $T_n \stackrel{w}{\longrightarrow} 0$.

Law of Large numbers for $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$

Basak-Mukherjee (2015), D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of (scaled) "approximately" d_n-regular graphs, with $d_n \overset{n\to\infty}{\longrightarrow} \infty$, then:

1. If $B > 0$, then $T_n \xrightarrow{d} t_{\theta,B}$
2. If $\theta \le 1, B = 0$, then $T_n \xrightarrow{w} 0$.
3. If $\theta > 1, B = 0$ and G_n is "well-connected", then

$$T_n \xrightarrow{w} \begin{cases} t_{\theta,B} & \text{w.p. } 0.5 \\ -t_{\theta,B} & \text{w.p. } 0.5 \end{cases}.$$

Law of Large numbers for $T_n = n^{-1} \sum_{i=1}^{n} \beta_i$

Basak-Mukherjee (2015), D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of (scaled) "approximately" d_n-regular graphs, with $d_n \xrightarrow{n\to\infty} \infty$, then:

1. If $B > 0$, then $T_n \xrightarrow{d} t_{\theta,B}$
2. If $\theta \leq 1, B = 0$, then $T_n \xrightarrow{w} 0$.
3. If $\theta > 1, B = 0$ and G_n is "well-connected", then

$$T_n \xrightarrow{w} \begin{cases} t_{\theta,B} & \text{w.p. } 0.5 \\ -t_{\theta,B} & \text{w.p. } 0.5 \end{cases}.$$

- Shades of phase transition at $\theta = 1$ when $B = 0$.
- More profound effect of phase transition in rates of convergence and fluctuations.

Fluctuations

$B > 0$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs, then:

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(\sqrt{n}(T_n - t_{\theta,B}) \leq x) - \mathbb{P}\left(\sqrt{\frac{1 - t_{\theta,B}^2}{1 - \theta(1 - t_{\theta,B}^2)}} Z \leq x \right) \right| \lesssim \frac{\sqrt{n}}{d_n}.$$

Here $Z \sim \mathcal{N}(0, 1)$.

Fluctuations

$B > 0$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs, then:

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}(\sqrt{n}(T_n - t_{\theta,B}) \le x) - \mathbb{P}\left(\sqrt{\frac{1 - t_{\theta,B}^2}{1 - \theta(1 - t_{\theta,B}^2)}} Z \le x \right) \right| \lesssim \frac{\sqrt{n}}{d_n}.$$

Here $Z \sim \mathcal{N}(0,1)$.

- In particular if $d_n \gg n^{1/2}$, then the distributional limit holds.
- This threshold is tight as there exists sequence of graphs with $d_n \sim n^{1/2}$ for which the limit does not hold.

Fluctuations

$B = 0$, $\theta < 1$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs, then:

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}(\sqrt{n}T_n \leq x) - \mathbb{P}\left(\sqrt{\frac{1}{1-\theta}}Z \leq x\right) \right| \lesssim \frac{n^{1/3}\text{poly}(\log n)}{d_n}.$$

Here $Z \sim \mathcal{N}(0,1)$.

Fluctuations

$B = 0$, $\theta < 1$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs, then:

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(\sqrt{n} T_n \leq x) - \mathbb{P}\left(\sqrt{\frac{1}{1-\theta}} Z \leq x \right) \right| \lesssim \frac{n^{1/3} \mathrm{poly}(\log n)}{d_n}.$$

Here $Z \sim \mathcal{N}(0,1)$.

- In particular if $d_n \gg n^{1/3}$, then the distributional limit holds.
- We do expect the limit to hold whenever $d_n \to \infty$.

Fluctuations

$B = 0$, $\theta = 1$, D.-Mukherjee (2023)

If G_N is a sequence of “approximately” d_n-regular graphs which are well-connected, then:

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(n^{1/4} T_n \leq x) - \mathbb{P}(W \leq x) \right| \lesssim \frac{\sqrt{n}\text{poly}(\log n)}{d_n}$$

where W now has density proportional to $\exp(-x^4/12)$.

Fluctuations

$B = 0$, $\theta = 1$, D.-Mukherjee (2023)

If G_N is a sequence of "approximately" d_n-regular graphs which are well-connected, then:

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}(n^{1/4} T_n \le x) - \mathbb{P}(W \le x) \right| \lesssim \frac{\sqrt{n}\,\mathrm{poly}(\log n)}{d_n}$$

where W now has density proportional to $\exp(-x^4/12)$.

- In particular if $d_n \gg n^{1/2}$, then the distributional limit holds.
- This threshold is tight as there exists sequence of graphs with $d_n \sim n^{1/2}$ for which the limit does not hold.

Fluctuations

$B = 0$, $\theta = 1$, D.-Mukherjee (2023)

If G_N is a sequence of "approximately" d_n-regular graphs which are well-connected, then:

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}(n^{1/4} T_n \leq x) - \mathbb{P}(W \leq x) \right| \lesssim \frac{\sqrt{n}\,\text{poly}(\log n)}{d_n}$$

where W now has density proportional to $\exp(-x^4/12)$.

- In particular if $d_n \gg n^{1/2}$, then the distributional limit holds.
- This threshold is tight as there exists sequence of graphs with $d_n \sim n^{1/2}$ for which the limit does not hold.
- The well-connectedness assumption is also tight.

Fluctuations

$B = 0$, $\theta > 1$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs which are well-connected, then:

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}(\sqrt{n}(T_n - t_{\theta,B}) \le x | T_n > 0) - \mathbb{P}\left(\sqrt{\frac{1-t_{\theta,B}^2}{1-\theta(1-t_{\theta,B}^2)}}Z \le x\right)\right| \lesssim \frac{\sqrt{n}}{d_n}$$

Similar result holds by conditioning on $T_n < 0$ and replacing $t_{\theta,B}$ by $-t_{\theta,B}$.

Fluctuations

$B = 0$, $\theta > 1$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of “approximately” d_n-regular graphs which are well-connected, then:

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}(\sqrt{n}(T_n - t_{\theta,B}) \le x | T_n > 0) - \mathbb{P}\left(\sqrt{\frac{1-t_{\theta,B}^2}{1-\theta(1-t_{\theta,B}^2)}}Z \le x\right)\right| \lesssim \frac{\sqrt{n}}{d_n}$$

Similar result holds by conditioning on $T_n < 0$ and replacing $t_{\theta,B}$ by $-t_{\theta,B}$.

- In particular if $d_n \gg n^{1/2}$, then the distributional limit holds.
- This threshold is tight as there exists sequence of graphs with $d_n \sim n^{1/2}$ for which the limit does not hold.

Fluctuations

$B = 0$, $\theta > 1$, D.-Mukherjee (2023)

If $\mathbf{A}_n$ is a sequence of "approximately" d_n-regular graphs which are well-connected, then:

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}(\sqrt{n}(T_n - t_{\theta,B}) \le x | T_n > 0) - \mathbb{P}\left(\sqrt{\frac{1-t_{\theta,B}^2}{1-\theta(1-t_{\theta,B}^2)}}Z \le x\right)\right| \lesssim \frac{\sqrt{n}}{d_n}$$

Similar result holds by conditioning on $T_n < 0$ and replacing $t_{\theta,B}$ by $-t_{\theta,B}$.

- In particular if $d_n \gg n^{1/2}$, then the distributional limit holds.
- This threshold is tight as there exists sequence of graphs with $d_n \sim n^{1/2}$ for which the limit does not hold.
- The well-connectedness assumption is also tight.

What happens for $\mathbf{q}^\top \boldsymbol{\beta}$?

- Lack of phase transition or universality for "approximately" d_n-regular graphs.

What happens for $\mathbf{q}^\top\boldsymbol{\beta}$?

- Lack of phase transition or universality for "approximately" d_n-regular graphs.

Other directions in Erdős-Rényi/ random regular

Suppose $\mathbf{A}_n$ is the adjacency matrix of an Erdős-Rényi graph with edge probability κ_n, scaled by the edge density $n\kappa_n$. Assume the following

$$\kappa_n >> n^{-1/2}, \quad \|\mathbf{q}\|_2 = 1, \quad \sum_{i=1}^{n} q_i = o(n^{1/2}).$$

Then

$$\sum_{i=1}^{n} q_i(\beta_i - t_{\theta,B})|\mathbf{A}_n \xrightarrow{w} N(0, 1 - t_{\theta,B}^2).$$

What happens for $\mathbf{q}^\top \boldsymbol{\beta}$?

- Lack of phase transition or universality for "approximately" d_n-regular graphs.

Other directions in Erdős-Rényi/ random regular

Suppose $\mathbf{A}_n$ is the adjacency matrix of an Erdős-Rényi graph with edge probability κ_n, scaled by the edge density $n\kappa_n$. Assume the following

$$\kappa_n >> n^{-1/2}, \quad \|\mathbf{q}\|_2 = 1, \quad \sum_{i=1}^{n} q_i = o(n^{1/2}).$$

Then

$$\sum_{i=1}^{n} q_i(\beta_i - t_{\theta,B})|\mathbf{A}_n \xrightarrow{w} N(0, 1 - t_{\theta,B}^2).$$

Therefore when $\mathbf{q}$ is a contrast, there is no phase transition!

What happens for $\mathbf{q}^\top \boldsymbol{\beta}$?

- Lack of phase transition or universality for "approximately" d_n-regular graphs.

Other directions in Erdős-Rényi/ random regular

Suppose $\mathbf{A}_n$ is the adjacency matrix of an Erdős-Rényi graph with edge probability κ_n, scaled by the edge density $n\kappa_n$. Assume the following

$$\kappa_n >> n^{-1/2}, \quad \|\mathbf{q}\|_2 = 1, \quad \sum_{i=1}^n q_i = o(n^{1/2}).$$

Then

$$\sum_{i=1}^n q_i(\beta_i - t_{\theta,B})|\mathbf{A}_n \xrightarrow{w} N(0, 1 - t_{\theta,B}^2).$$

Therefore when $\mathbf{q}$ is a contrast, there is no phase transition!

- Generally it matters which eigenvalue of $\mathbf{A}_n$, the direction $\mathbf{q}$ is most aligned towards. The corresponding eigenvalue shows up in the limiting variance.

What happens for $\mathbf{q}^\top \boldsymbol{\beta}$?

- Lack of phase transition or universality for "approximately" d_n-regular graphs.

Other directions in Erdős-Rényi/ random regular

Suppose $\mathbf{A}_n$ is the adjacency matrix of an Erdős-Rényi graph with edge probability κ_n, scaled by the edge density $n\kappa_n$. Assume the following

$$\kappa_n >> n^{-1/2}, \quad \|\mathbf{q}\|_2 = 1, \quad \sum_{i=1}^{n} q_i = o(n^{1/2}).$$

Then

$$\sum_{i=1}^{n} q_i(\beta_i - t_{\theta,B})|\mathbf{A}_n \xrightarrow{w} N(0, 1 - t_{\theta,B}^2).$$

Therefore when $\mathbf{q}$ is a contrast, there is no phase transition!

- Generally it matters which eigenvalue of $\mathbf{A}_n$, the direction $\mathbf{q}$ is most aligned towards. The corresponding eigenvalue shows up in the limiting variance.
- In regular graphs, the leading eigenvector is always $n^{-1/2}\mathbf{1}$ with eigenvalue 1. Therefore choosing $\mathbf{q} = n^{-1/2}\mathbf{1}$ leads to universal behavior for "approximately" regular graphs.

Brief Summary

- We derive Berry-Esseen bounds between $n^{-1/2}\sum_{i=1}^{n}\beta_i$ and an appropriate limit (Gaussian or otherwise) through the full Ferromagnetic parameter regime $\theta > 0$ and $B \in \mathbb{R}$.

Brief Summary

- We derive Berry-Esseen bounds between $n^{-1/2} \sum_{i=1}^{n} \beta_i$ and an appropriate limit (Gaussian or otherwise) through the full Ferromagnetic parameter regime $\theta > 0$ and $B \in \mathbb{R}$.

- In particular, this condition holds both for deterministic and random interaction matrices.

Brief Summary

- We derive Berry-Esseen bounds between $n^{-1/2}\sum_{i=1}^{n}\beta_i$ and an appropriate limit (Gaussian or otherwise) through the full Ferromagnetic parameter regime $\theta > 0$ and $B \in \mathbb{R}$.

- In particular, this condition holds both for deterministic and random interaction matrices.

- Also works beyond binary $\{-1, 1\}$ spins. If instead the domination measure is some μ (symmetric around 0), one can still derive similar results with the phase transition point being at $(\alpha''(0))^{-1}$ where $\alpha(\cdot)$ is the cumulant generating function of μ.

Brief Summary

- We derive Berry-Esseen bounds between $n^{-1/2}\sum_{i=1}^{n}\beta_i$ and an appropriate limit (Gaussian or otherwise) through the full Ferromagnetic parameter regime $\theta > 0$ and $B \in \mathbb{R}$.
- In particular, this condition holds both for deterministic and random interaction matrices.
- Also works beyond binary $\{-1, 1\}$ spins. If instead the domination measure is some μ (symmetric around 0), one can still derive similar results with the phase transition point being at $(\alpha''(0))^{-1}$ where $\alpha(\cdot)$ is the cumulant generating function of μ.
- Non-universal behavior and lack of phase transitions for general linear combinations with $\mathbf{q}$.

Outline

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

- We make a high temperature assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$, $\mathbf{A} = \mathrm{Off}(\mathbf{X}^\top \mathbf{X})$.

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

- We make a high temperature assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$, $\mathbf{A} = \mathrm{Off}(\mathbf{X}^\top \mathbf{X})$.
    - In particular, this assumption guarantees uniqueness of the Mean-Field optimizer $\mathbf{u}^{\mathrm{opt}}$ (which will be our centering), for any prior π.

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

- We make a high temperature assumption $\|\mathbf{A}\|_4 \le 1 - \rho$, $\mathbf{A} = \mathrm{Off}(\mathbf{X}^\top \mathbf{X})$.
  - In particular, this assumption guarantees uniqueness of the Mean-Field optimizer $\mathbf{u}^{\mathrm{opt}}$ (which will be our centering), for any prior π.
  - Also, this assumption ensures that we are in low SNR regime $\|\mathbf{A}\|_2 = O(1)$.

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

- We make a high temperature assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$, $\mathbf{A} = \mathrm{Off}(\mathbf{X}^\top \mathbf{X})$.
  - In particular, this assumption guarantees uniqueness of the Mean-Field optimizer $\mathbf{u}^{\mathrm{opt}}$ (which will be our centering), for any prior π.
  - Also, this assumption ensures that we are in low SNR regime $\|\mathbf{A}\|_2 = O(1)$.
- Our second assumption is the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

Assumptions

Our goal is a high-dimensional limit $p, n \to \infty$

$$\sum_{i=1}^{p} q_i(\beta_i - ??) \xrightarrow{d} \text{... conditioned on } \mathbf{X}, \mathbf{y}$$

where $\boldsymbol{\beta}$ drawn from the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

- We make a high temperature assumption $\|\mathbf{A}\|_4 \le 1 - \rho$, $\mathbf{A} = \mathrm{Off}(\mathbf{X}^\top \mathbf{X})$.
    - In particular, this assumption guarantees uniqueness of the Mean-Field optimizer $\mathbf{u}^{\mathrm{opt}}$ (which will be our centering), for any prior π.
    - Also, this assumption ensures that we are in low SNR regime $\|\mathbf{A}\|_2 = O(1)$.
- Our second assumption is the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

- Finally, we assume that $\mathbf{q}$ is an arbitrary vector with $\|\mathbf{q}\|_2 = 1$.

Lee-D.-Mukherjee, 25+

(i) Under the above assumptions, there is a unique well separated optimizer $\mathbf{u}^{\text{opt}}$ for the mean-field prediction formula.

LLN

Lee-D.-Mukherjee, 25+

(i) Under the above assumptions, there is a unique well separated optimizer $\mathbf{u}^{\text{opt}}$ for the mean-field prediction formula.

(ii) Further we have

$$\mathbb{E}_\mu \Big[\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \Big]^2 = O(1).$$

Lee-D.-Mukherjee, 25+

(i) Under the above assumptions, there is a unique well separated optimizer $\mathbf{u}^{\text{opt}}$ for the mean-field prediction formula.

(ii) Further we have

$$\mathbb{E}_\mu\Big[\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}})\Big]^2 = O(1).$$

- Note that this bound is the best possible, as even if $\{\beta_i\}_{1\le i\le p}$ were independent with mean u_i^{opt}, the second moment would be $O(1)$.

LLN

Lee-D.-Mukherjee, 25+

(i) Under the above assumptions, there is a unique well separated optimizer $\mathbf{u}^{\text{opt}}$ for the mean-field prediction formula.

(ii) Further we have

$$\mathbb{E}_\mu\Big[\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}})\Big]^2 = O(1).$$

- Note that this bound is the best possible, as even if $\{\beta_i\}_{1\le i\le p}$ were independent with mean u_i^{opt}, the second moment would be $O(1)$.
- Recall that Bayes optimal estimator for $\sum_{i=1}^{p} q_i\beta_i$ is $\sum_{i=1}^{p} q_i\mathbb{E}_\mu[\beta_i|\mathbf{y},\mathbf{X}]$.

Lee-D.-Mukherjee, 25+

(i) Under the above assumptions, there is a unique well separated optimizer $\mathbf{u}^{\text{opt}}$ for the mean-field prediction formula.

(ii) Further we have

$$\mathbb{E}_\mu\Big[\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}})\Big]^2 = O(1).$$

- Note that this bound is the best possible, as even if $\{\beta_i\}_{1\le i\le p}$ were independent with mean u_i^{opt}, the second moment would be $O(1)$.
- Recall that Bayes optimal estimator for $\sum_{i=1}^{p} q_i\beta_i$ is $\sum_{i=1}^{p} q_i\mathbb{E}_\mu[\beta_i|\mathbf{y},\mathbf{X}]$. Same computation shows that the Mean-Field estimator $\sum_{i=1}^{p} q_i u_i^{\text{opt}}$ is approximately Bayes optimal for $\sum_{i=1}^{p} q_i\beta_i$.

Some comments about LLN Theorem

- The above theorem applies for deterministic $\mathbf{y}$ and $\mathbf{X}$.

Some comments about LLN Theorem

- The above theorem applies for deterministic $\mathbf{y}$ and $\mathbf{X}$.
- If $\mathbf{y}$ (and $\mathbf{X}$) are random, it continues to hold, after conditioning on $\mathbf{y}$ (and $\mathbf{X}$).

Some comments about LLN Theorem

- The above theorem applies for deterministic $\mathbf{y}$ and $\mathbf{X}$.
- If $\mathbf{y}$ (and $\mathbf{X}$) are random, it continues to hold, after conditioning on $\mathbf{y}$ (and $\mathbf{X}$).
- The crucial assumption of the theorem is the strong mean-field assumption $\max_{i\in[p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2})$.

Some comments about LLN Theorem

- The above theorem applies for deterministic $\mathbf{y}$ and $\mathbf{X}$.
- If $\mathbf{y}$ (and $\mathbf{X}$) are random, it continues to hold, after conditioning on $\mathbf{y}$ (and $\mathbf{X}$).
- The crucial assumption of the theorem is the strong mean-field assumption $\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2})$.
- There are known examples which show that the mean-field centering $\mathbf{u}^{\text{opt}}$ is not the right one if $\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = O(p^{-1/2})$.

Assumptions1

- For CLT1 we still make the assumption that $X_{ij} = n^{-1/2} Z_{ij}$, where $(Z_{ij})_{1 \le i \le n, 1 \le j \le p}$ are IID observations from a sub-gaussian distribution.

Assumptions1

- For CLT1 we still make the assumption that $X_{ij} = n^{-1/2} Z_{ij}$, where $(Z_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ are IID observations from a sub-gaussian distribution.
- We also assume that the vector $\mathbf{q}$ is delocalized, in the sense that $\|\mathbf{q}\|_\infty \to 0$.

Assumptions1

- For CLT1 we still make the assumption that $X_{ij} = n^{-1/2} Z_{ij}$, where $(Z_{ij})_{1 \le i \le n, 1 \le j \le p}$ are IID observations from a sub-gaussian distribution.

- We also assume that the vector $\mathbf{q}$ is delocalized, in the sense that $\|\mathbf{q}\|_\infty \to 0$.

- Note that if $\mathbf{q}$ is not delocalized, CLT does not hold for $\mathbf{q}^\top \boldsymbol{\beta}$ even for IID random variables $\boldsymbol{\beta}$.

Assumptions1

- For CLT1 we still make the assumption that $X_{ij} = n^{-1/2} Z_{ij}$, where $(Z_{ij})_{1 \le i \le n, 1 \le j \le p}$ are IID observations from a sub-gaussian distribution.

- We also assume that the vector $\mathbf{q}$ is delocalized, in the sense that $\|\mathbf{q}\|_\infty \to 0$.

- Note that if $\mathbf{q}$ is not delocalized, CLT does not hold for $\mathbf{q}^\top \boldsymbol{\beta}$ even for IID random variables $\boldsymbol{\beta}$.

- Finally, we also assume that the true data $\mathbf{y}$ is generated from a frequentist linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \text{ where } \|\boldsymbol{\beta}^\star\|_\infty \le 1 \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}).$$

CLT1

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N(0, v),$$

provided $p \ll n^{-2/3}$.

CLT1

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N(0, \upsilon),$$

provided $p \ll n^{-2/3}$. Here $\upsilon := \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \stackrel{d}{\approx} N(0, \upsilon),$$

provided $p \ll n^{-2/3}$. Here $\upsilon := \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

- Note that all the quantities

$$\alpha(\cdot) = \alpha_{\pi,d}(\cdot), \quad \mathbf{c} = \mathbf{X}^\top \mathbf{y}, \quad \mathbf{u}^{\text{opt}} = \mathbf{u}^{\text{opt}}_{\mathbf{y},\mathbf{X},\pi}$$

can be computed once we know $(\mathbf{y}, \mathbf{X}, \pi)$. Here π-prior and d=Diagonal of $\mathbf{X}^\top \mathbf{X}$.

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N(0, \upsilon),$$

provided $p \ll n^{-2/3}$. Here $\upsilon := \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

- Note that all the quantities

  $$\alpha(\cdot) = \alpha_{\pi,d}(\cdot), \quad \mathbf{c} = \mathbf{X}^\top \mathbf{y}, \quad \mathbf{u}^{\text{opt}} = \mathbf{u}^{\text{opt}}_{\mathbf{y},\mathbf{X},\pi}$$

  can be computed once we know $(\mathbf{y}, \mathbf{X}, \pi)$. Here π-prior and d=Diagonal of $\mathbf{X}^\top \mathbf{X}$.

- In particular, they don't depend on the unknown true $\boldsymbol{\beta}^\star$.

- Note that the previous theorem does not directly make the design assumptions

$$\|\mathbf{A}\|_4 \leq 1 - \rho \text{ and } \max_{1\leq i\leq p} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}),$$

made for the LLN.

Comments about CLT1

- Note that the previous theorem does not directly make the design assumptions

$$\|\mathbf{A}\|_4 \leq 1 - \rho \text{ and } \max_{1 \leq i \leq p} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}),$$

  made for the LLN.

- We verify that both the assumptions hold if $p \ll n^{-2/3}$.

Comments about CLT1

- Note that the previous theorem does not directly make the design assumptions

$$\|\mathbf{A}\|_4 \leq 1-\rho \text{ and } \max_{1\leq i\leq p}\sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}),$$

  made for the LLN.

- We verify that both the assumptions hold if $p \ll n^{-2/3}$.

- This requires a new Chevet type inequality for bounding $(4,4)$ operator norm of random covariance matrices. A more naive $\|\mathbf{A}\|_\infty \leq 1-\rho$ condition would result in a weaker threshold $p \ll n^{-1/2}$.

Comments about CLT1

- Note that the previous theorem does not directly make the design assumptions

$$\|\mathbf{A}\|_4 \leq 1-\rho \text{ and } \max_{1\leq i\leq p} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}),$$

  made for the LLN.

- We verify that both the assumptions hold if $p \ll n^{-2/3}$.

- This requires a new Chevet type inequality for bounding $(4,4)$ operator norm of random covariance matrices. A more naive $\|\mathbf{A}\|_\infty \leq 1-\rho$ condition would result in a weaker threshold $p \ll n^{-1/2}$.

- We believe the dimension dependence is sharp, in that the mean-field centering has to be adjusted to go beyond this regime.

Comments about CLT1

- Note that the previous theorem does not directly make the design assumptions

$$\|\mathbf{A}\|_4 \leq 1 - \rho \text{ and } \max_{1 \leq i \leq p} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}),$$

  made for the LLN.

- We verify that both the assumptions hold if $p \ll n^{-2/3}$.

- This requires a new Chevet type inequality for bounding $(4, 4)$ operator norm of random covariance matrices. A more naive $\|\mathbf{A}\|_\infty \leq 1 - \rho$ condition would result in a weaker threshold $p \ll n^{-1/2}$.

- We believe the dimension dependence is sharp, in that the mean-field centering has to be adjusted to go beyond this regime.

- As a comment, similar dimension dependence also arises for Bernstein-von-Mises type CLT approximations in the high SNR regime (see Katsevich, Arxiv-2023), where they require $p \ll n^{-1/2}$.

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.
- We also explicitly assume the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.
- We also explicitly assume the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

- But now we assume that $\mathbf{q}$ is an (approximate) eigenvector of $-\mathbf{A}$, with eigenvalue λ.

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.
- We also explicitly assume the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

- But now we assume that $\mathbf{q}$ is an (approximate) eigenvector of $-\mathbf{A}$, with eigenvalue λ.
- Finally, we assume that the true data $\mathbf{y}$ is generated from a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\beta}^\star \overset{IID}{\sim} \pi^\star, \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}).$$

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.
- We also explicitly assume the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

- But now we assume that $\mathbf{q}$ is an (approximate) eigenvector of $-\mathbf{A}$, with eigenvalue λ.
- Finally, we assume that the true data $\mathbf{y}$ is generated from a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\beta}^\star \overset{IID}{\sim} \pi^\star, \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}).$$

- Here $\pi^\star$ is a symmetric distribution with finite fourth moment.

Assumptions2 (True Bayesian model)

- For CLT2, we work with a non-random design matrix $\mathbf{X}$.
- We explicitly make the assumption $\|\mathbf{A}\|_4 \leq 1 - \rho$.
- We also explicitly assume the strong mean-field condition

$$\max_{i \in [p]} \sum_{j=1}^{p} A_{ij}^2 = o(p^{-1/2}).$$

- But now we assume that $\mathbf{q}$ is an (approximate) eigenvector of $-\mathbf{A}$, with eigenvalue λ.
- Finally, we assume that the true data $\mathbf{y}$ is generated from a linear model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\beta}^\star \overset{IID}{\sim} \pi^\star, \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}).$$

- Here $\pi^\star$ is a symmetric distribution with finite fourth moment. Thus we allow the prior to be misspecified as π, where the true data generating prior is $\pi^\star$.

CLT2

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N\Big(0, \frac{\upsilon}{1-\lambda\upsilon}\Big)$$

CLT2

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N\Big(0, \frac{\upsilon}{1-\lambda\upsilon}\Big)$$

Here $\upsilon = \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$ as before. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

CLT2

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N\Big(0, \frac{\upsilon}{1-\lambda\upsilon}\Big)$$

Here $\upsilon = \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$ as before. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

- Note that the centering or scaling quantities $\mathbf{u}^{\text{opt}}$ and υ do not depend on the (possibly unknown) true prior $\pi^\star$.

CLT2

Lee-D.-Mukherjee, 25+

Under the above assumptions, conditioning on $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}^\star$, under the posterior $\mu = \mu_{\mathbf{y},\mathbf{X},\pi}$ we have

$$\sum_{i=1}^{p} q_i(\beta_i - u_i^{\text{opt}}) \overset{d}{\approx} N\Big(0, \frac{\upsilon}{1 - \lambda\upsilon}\Big)$$

Here $\upsilon = \sum_{i=1}^{p} q_i^2 \alpha''(c_i)$ as before. Recall $\mathbf{c} = \mathbf{X}^\top \mathbf{y}$.

- Note that the centering or scaling quantities $\mathbf{u}^{\text{opt}}$ and υ do not depend on the (possibly unknown) true prior $\pi^\star$.
- For both CLTs we provide explicit convergence rates for the above theorem in Kolmogorov-Smirnov distance.

Application: Credible Intervals

- Based on CLT2, the set

$$\mathcal{I} := \mathcal{I}(\mathbf{y}, \mathbf{X}, \pi) := \Big[\sum_{i=1}^{p} q_i u_i^{\text{opt}} \pm z_{\alpha/2} \sqrt{\frac{\upsilon}{1 - \lambda \upsilon}} \Big]$$

  is asymptotically a $1 - \alpha$ credible interval under the posterior $\mu_{\mathbf{y}, \mathbf{X}, \pi}$.

Application: Credible Intervals

- Based on CLT2, the set

$$\mathcal{I} := \mathcal{I}(\mathbf{y}, \mathbf{X}, \pi) := \Big[\sum_{i=1}^{p} q_i u_i^{\text{opt}} \pm z_{\alpha/2} \sqrt{\frac{\upsilon}{1 - \lambda\upsilon}} \Big]$$

is asymptotically a $1 - \alpha$ credible interval under the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

Lee-D.-Mukherjee, 25+

- Suppose we are in the set up of CLT2.

Application: Credible Intervals

- Based on CLT2, the set

$$\mathcal{I} := \mathcal{I}(\mathbf{y}, \mathbf{X}, \pi) := \Big[\sum_{i=1}^{p} q_i u_i^{\text{opt}} \pm z_{\alpha/2} \sqrt{\frac{\upsilon}{1 - \lambda \upsilon}} \Big]$$

is asymptotically a $1 - \alpha$ credible interval under the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

Lee-D.-Mukherjee, 25+

- Suppose we are in the set up of CLT2.
- Then we have

$$\mathbb{P}_{\pi^*}\Big(\sum_{i=1}^{p} q_i \beta_i^\star \in \mathcal{I}(\mathbf{y}) \Big| \mathbf{X} \Big) \xrightarrow{P} F(d, \alpha, \pi, \pi^\star).$$

Application: Credible Intervals

- Based on CLT2, the set

$$\mathcal{I} := \mathcal{I}(\mathbf{y}, \mathbf{X}, \pi) := \Big[\sum_{i=1}^{p} q_i u_i^{\mathrm{opt}} \pm z_{\alpha/2}\sqrt{\frac{\upsilon}{1-\lambda\upsilon}}\Big]$$

is asymptotically a $1-\alpha$ credible interval under the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

Lee-D.-Mukherjee, 25+

- Suppose we are in the set up of CLT2.
- Then we have

$$\mathbb{P}_{\pi^*}\Big(\sum_{i=1}^{p} q_i \beta_i^\star \in \mathcal{I}(\mathbf{y})\Big|\mathbf{X}\Big) \xrightarrow{P} F(d, \alpha, \pi, \pi^\star).$$

- In particular $F(d, \alpha, \pi^\star, \pi^\star) = 1-\alpha$, so if we use the correct prior, we have asymptotically valid credible intervals.

Application: Credible Intervals

- Based on CLT2, the set

$$\mathcal{I} := \mathcal{I}(\mathbf{y}, \mathbf{X}, \pi) := \Big[\sum_{i=1}^{p} q_i u_i^{\text{opt}} \pm z_{\alpha/2} \sqrt{\frac{\upsilon}{1 - \lambda\upsilon}} \Big]$$

is asymptotically a $1 - \alpha$ credible interval under the posterior $\mu_{\mathbf{y},\mathbf{X},\pi}$.

Lee-D.-Mukherjee, 25+

- Suppose we are in the set up of CLT2.
- Then we have

$$\mathbb{P}_{\pi^*}\Big(\sum_{i=1}^{p} q_i \beta_i^\star \in \mathcal{I}(\mathbf{y}) \Big| \mathbf{X} \Big) \xrightarrow{P} F(d, \alpha, \pi, \pi^\star).$$

- In particular $F(d, \alpha, \pi^\star, \pi^\star) = 1 - \alpha$, so if we use the correct prior, we have asymptotically valid credible intervals.
- Also works if you sample split to estimate $\pi^\star$ by some $\hat{\pi}_n$ and use it as a plug-in prior.

Outline

Brief Summary

- We give general conditions on the design matrix $\mathbf{X}$ for LLN and CLT to hold, for Bayesian Linear Regression with a product prior.

Brief Summary

- We give general conditions on the design matrix $\mathbf{X}$ for LLN and CLT to hold, for Bayesian Linear Regression with a product prior.

- In particular, this condition holds both for deterministic and random matrices, and allows for dependence of entries.

Brief Summary

- We give general conditions on the design matrix $\mathbf{X}$ for LLN and CLT to hold, for Bayesian Linear Regression with a product prior.
- In particular, this condition holds both for deterministic and random matrices, and allows for dependence of entries.
- We give explicit error rates in terms of the Kolmogorov-Smirnov distance.

Brief Summary

- We give general conditions on the design matrix $\mathbf{X}$ for LLN and CLT to hold, for Bayesian Linear Regression with a product prior.
- In particular, this condition holds both for deterministic and random matrices, and allows for dependence of entries.
- We give explicit error rates in terms of the Kolmogorov-Smirnov distance.
- We apply our results to construct credible intervals, and compute their asymptotic coverage under possible prior misspecification.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$. In this case one can have more than one mean-field optimizer, and the right approximation should be a mixture of product measures.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$. In this case one can have more than one mean-field optimizer, and the right approximation should be a mixture of product measures.

- Another possible direction is to remove the compactness assumption, and allow for a broader class of priors.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$. In this case one can have more than one mean-field optimizer, and the right approximation should be a mixture of product measures.

- Another possible direction is to remove the compactness assumption, and allow for a broader class of priors. One can assume log-concavity of the prior, which also guarantees uniqueness of optimizer.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$. In this case one can have more than one mean-field optimizer, and the right approximation should be a mixture of product measures.
- Another possible direction is to remove the compactness assumption, and allow for a broader class of priors. One can assume log-concavity of the prior, which also guarantees uniqueness of optimizer.
- Finally, it remains to study similar questions for non-quadratic posteriors.

Future Scope

- A natural question is whether we can study the low temperature regime, where $\|\mathbf{A}\|_4 > 1$. In this case one can have more than one mean-field optimizer, and the right approximation should be a mixture of product measures.

- Another possible direction is to remove the compactness assumption, and allow for a broader class of priors. One can assume log-concavity of the prior, which also guarantees uniqueness of optimizer.

- Finally, it remains to study similar questions for non-quadratic posteriors. People apply NMF to a host of problems (GLMs, Topic Modeling).

Thank you. Questions?