Variational inference, spin glasses, and TAP free energy

Song Mei

April 20, 2021

Joint work with Zhou Fan and Andrea Montanari

General motivation

- Bayesian inference: topic modeling, Bayesian GLM, Bayesian NN... High dim. integration is hard!
- Variational inference: integration $\to$ optimization.
- Popular objective function: "mean field free energy".
- ... but not optimal for moderate SNR.
- Today: optimal objective "TAP free energy".

$\mathbb{Z}_2$ synchronization

- Signal:

$$\boldsymbol{x} = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{Z}_2^n, \qquad x_i \overset{i.i.d.}{\sim} \mathrm{Unif}(\mathbb{Z}_2), \qquad \mathbb{Z}_2 = \{+1, -1\}.$$

- Observation $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$:

$$\boldsymbol{Y} = \frac{\lambda}{n} \boldsymbol{x}\boldsymbol{x}^\mathsf{T} + \boldsymbol{W}.$$

- Noise: $\boldsymbol{W} \sim \mathrm{GOE}(n)$.
- SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.
- Task: given $\boldsymbol{Y} = (Y_{ij})$, estimate $\boldsymbol{x}$ (or say $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\mathsf{T}$).

Bayes estimation in $\mathbb{Z}_2$ synchronization

- Estimate $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\mathsf{T}$ with loss:

$$\ell(\boldsymbol{X}, \widehat{\boldsymbol{X}}) = (1/n^2)\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F^2.$$

- $\text{MSE}(\widehat{\boldsymbol{X}}) = \mathbb{E}[(1/n^2)\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F^2]$.
- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient. BBP.
- The optimal estimator is the Bayes estimator:

$$\widehat{\boldsymbol{X}}_{\text{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}].$$

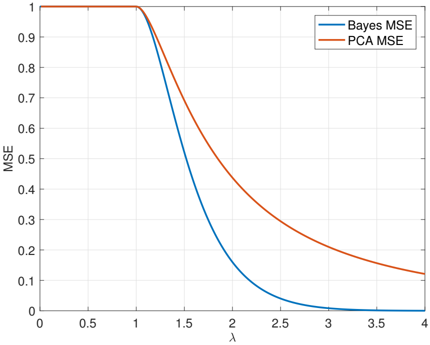
Bayes estimation in $\mathbb{Z}_2$ synchronization

- Settings:

$$\boldsymbol{x} \sim \mathrm{Unif}(\mathbb{Z}_2^n), \qquad \boldsymbol{Y} = (\lambda/n)\boldsymbol{x}\boldsymbol{x}^\mathsf{T} + \boldsymbol{W}.$$

- Risk:

$$\mathrm{MSE}_\lambda(\widehat{\boldsymbol{X}}) = (1/n^2)\mathbb{E}[\|\boldsymbol{x}\boldsymbol{x}^\mathsf{T} - \widehat{\boldsymbol{X}}\|_F^2].$$



Compute the Bayesian estimator

- The Bayesian estimator:

$$\widehat{\boldsymbol{X}}_{\text{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}] = \sum_{\boldsymbol{\sigma} \in \mathbb{Z}_2^n} \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} p(\boldsymbol{\sigma} | \boldsymbol{Y}).$$

- The posterior distribution:

$$p(\boldsymbol{\sigma} | \boldsymbol{Y}) = \frac{1}{Z} \exp\{\lambda \langle \boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma} \rangle / 2\}.$$

Mean field variational inference

- The posterior distribution:

$$p(\boldsymbol{\sigma}|\boldsymbol{Y}) = \frac{1}{Z}\exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\}.$$

- Approximate $p(\boldsymbol{\sigma}|\boldsymbol{Y})$ by $q \in \mathcal{P}_{\rm MF}$:

$$\mathcal{P}_{\rm MF} = \Big\{q(\boldsymbol{\sigma}) = \prod_{i=1}^n q_i(\sigma_i) : q_i \in \mathcal{P}(\mathbb{Z}_2)\Big\} \cong [-1,1]^n.$$

- Minimize the relative entropy between q and $p(\boldsymbol{\sigma}|\boldsymbol{Y})$:

$$\min_{q\in\mathcal{P}_{\rm MF}} {\rm D_{kl}}(q\|p(\boldsymbol{\sigma}|\boldsymbol{Y})).$$

- Equivalently minimizing $\min_{\boldsymbol{m}\in[-1,1]^n} \mathcal{F}_{\rm MF}(\boldsymbol{m})$

$$\mathcal{F}_{\rm MF}(\boldsymbol{m}) \equiv -\sum_{i=1}^n {\sf h}(m_i) - \lambda\langle\boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2 \geq -\log Z,$$

where ${\sf h}(m) = -\frac{1-m}{2}\log(\frac{1-m}{2}) - \frac{1+m}{2}\log(\frac{1+m}{2})$.

Mean field variational inference

- The posterior distribution:

$$p(\boldsymbol{\sigma}|\boldsymbol{Y}) = \frac{1}{Z}\exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\}.$$

- Approximate $p(\boldsymbol{\sigma}|\boldsymbol{Y})$ by $q \in \mathcal{P}_{\rm MF}$:

$$\mathcal{P}_{\rm MF} = \Big\{q(\boldsymbol{\sigma}) = \prod_{i=1}^n q_i(\sigma_i) : q_i \in \mathcal{P}(\mathbb{Z}_2)\Big\} \cong [-1,1]^n.$$

- Minimize the relative entropy between q and $p(\boldsymbol{\sigma}|\boldsymbol{Y})$:

$$\min_{q\in\mathcal{P}_{\rm MF}} \mathrm{D}_{\rm kl}(q\|p(\boldsymbol{\sigma}|\boldsymbol{Y})).$$

- Equivalently minimizing $\min_{\boldsymbol{m}\in[-1,1]^n}\mathcal{F}_{\rm MF}(\boldsymbol{m})$

$$\mathcal{F}_{\rm MF}(\boldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(m_i) - \lambda\langle\boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2 \geq -\log Z,$$

where $\mathsf{h}(m) = -\frac{1-m}{2}\log(\frac{1-m}{2}) - \frac{1+m}{2}\log(\frac{1+m}{2})$.

Mean field variational inference

- The posterior distribution:

$$p(\boldsymbol{\sigma}|\boldsymbol{Y}) = \frac{1}{Z}\exp\{\lambda\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\}.$$

- Approximate $p(\boldsymbol{\sigma}|\boldsymbol{Y})$ by $q \in \mathcal{P}_{\rm MF}$:

$$\mathcal{P}_{\rm MF} = \Big\{ q(\boldsymbol{\sigma}) = \prod_{i=1}^n q_i(\sigma_i) : q_i \in \mathcal{P}(\mathbb{Z}_2) \Big\} \cong [-1,1]^n.$$

- Minimize the relative entropy between q and $p(\boldsymbol{\sigma}|\boldsymbol{Y})$:

$$\min_{q\in\mathcal{P}_{\rm MF}} {\rm D}_{\rm kl}(q\|p(\boldsymbol{\sigma}|\boldsymbol{Y})).$$

- Equivalently minimizing $\min_{\boldsymbol{m}\in[-1,1]^n} \mathcal{F}_{\rm MF}(\boldsymbol{m})$

$$\mathcal{F}_{\rm MF}(\boldsymbol{m}) \equiv -\sum_{i=1}^n {\sf h}(m_i) - \lambda\langle\boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2 \ge -\log Z,$$

where ${\sf h}(m) = -\frac{1-m}{2}\log(\frac{1-m}{2}) - \frac{1+m}{2}\log(\frac{1+m}{2})$.

Mean field variational inference

- The posterior distribution:

$$p(\boldsymbol{\sigma}|\boldsymbol{Y}) = \frac{1}{Z}\exp\{\lambda\langle\boldsymbol{\sigma},\boldsymbol{Y}\boldsymbol{\sigma}\rangle/2\}.$$

- Approximate $p(\boldsymbol{\sigma}|\boldsymbol{Y})$ by $q \in \mathcal{P}_{\rm MF}$:

$$\mathcal{P}_{\rm MF} = \Big\{q(\boldsymbol{\sigma}) = \prod_{i=1}^{n} q_i(\sigma_i) : q_i \in \mathcal{P}(\mathbb{Z}_2)\Big\} \cong [-1,1]^n.$$

- Minimize the relative entropy between q and $p(\boldsymbol{\sigma}|\boldsymbol{Y})$:

$$\min_{q\in\mathcal{P}_{\rm MF}} \mathrm{D}_{\rm kl}(q\|p(\boldsymbol{\sigma}|\boldsymbol{Y})).$$

- Equivalently minimizing $\min_{\boldsymbol{m}\in[-1,1]^n}\mathcal{F}_{\rm MF}(\boldsymbol{m})$

$$\mathcal{F}_{\rm MF}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n}\mathsf{h}(m_i) - \lambda\langle\boldsymbol{m},\boldsymbol{Y}\boldsymbol{m}\rangle/2 \geq -\log Z,$$

where $\mathsf{h}(m) = -\frac{1-m}{2}\log(\frac{1-m}{2}) - \frac{1+m}{2}\log(\frac{1+m}{2})$.

Mean field variational inference

- Mean field free energy:

$$\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n} \mathsf{h}(m_i) - \lambda\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2.$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{MF}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} \approx \widehat{\boldsymbol{X}}_{\mathrm{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}].$$

- It was shown that $\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} \not\approx \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

Mean field variational inference

- Mean field free energy:

$$\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n} \mathsf{h}(m_i) - \lambda\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2.$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{MF}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} \approx \widehat{\boldsymbol{X}}_{\mathsf{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}].$$

- It was shown that $\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} \not\approx \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

Mean field variational inference

- Mean field free energy:

$$\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n} \mathsf{h}(m_i) - \lambda \langle \boldsymbol{m}, \boldsymbol{Y} \boldsymbol{m} \rangle / 2.$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{MF}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} \approx \widehat{\boldsymbol{X}}_{\mathrm{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}].$$

- It was shown that $\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} \not\approx \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

Mean field variational inference

- Mean field free energy:

$$\mathcal{F}_{\mathrm{MF}}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n} \mathsf{h}(m_i) - \lambda\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle/2.$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{MF}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} \approx \widehat{\boldsymbol{X}}_{\mathsf{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}].$$

- It was shown that $\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} \not\approx \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

The TAP free energy

- Thouless, Anderson, and Palmer (1977) proposed the TAP free energy when they study the Sherrington-Kirkpatrick model, whose Gibbs measure gives

$$G_{\beta,\lambda}(\boldsymbol{\sigma}) = \frac{1}{Z_{\beta,\lambda}} \exp\{\beta\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle\}.$$

where $Y_{ij} \sim \mathcal{N}(\lambda/n, 1/n)$.

- When $\beta = \lambda$, the Gibbs measure of SK model is the same as the posterior of $\mathbb{Z}_2$ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle\boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle}_{\mathcal{F}_{\mathrm{MF}}} \underbrace{-\frac{n\lambda^2}{4}\left[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\right]^2}_{\text{Onsager's correction term}}.$$

The TAP free energy

- Thouless, Anderson, and Palmer (1977) proposed the TAP free energy when they study the Sherrington-Kirkpatrick model, whose Gibbs measure gives

$$G_{\beta,\lambda}(\boldsymbol{\sigma}) = \frac{1}{Z_{\beta,\lambda}} \exp\{\beta\langle\boldsymbol{\sigma}, \boldsymbol{Y}\boldsymbol{\sigma}\rangle\}.$$

  where $Y_{ij} \sim \mathcal{N}(\lambda/n, 1/n)$.
- When $\beta = \lambda$, the Gibbs measure of SK model is the same as the posterior of $\mathbb{Z}_2$ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle\boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle}_{\mathcal{F}_{\mathrm{MF}}} \underbrace{-\frac{n\lambda^2}{4}\left[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\right]^2}_{\text{Onsager's correction term}}.$$

The TAP free energy

- Thouless, Anderson, and Palmer (1977) proposed the TAP free energy when they study the Sherrington-Kirkpatrick model, whose Gibbs measure gives

$$G_{\beta,\lambda}(\boldsymbol{\sigma}) = \frac{1}{Z_{\beta,\lambda}} \exp\{\beta \langle \boldsymbol{\sigma}, \boldsymbol{Y} \boldsymbol{\sigma} \rangle\}.$$

  where $Y_{ij} \sim \mathcal{N}(\lambda/n, 1/n)$.
- When $\beta = \lambda$, the Gibbs measure of SK model is the same as the posterior of $\mathbb{Z}_2$ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2} \langle \boldsymbol{m}, \boldsymbol{Y} \boldsymbol{m} \rangle}_{\mathcal{F}_{\mathrm{MF}}} \underbrace{- \frac{n\lambda^2}{4} \left[1 - \frac{\|\boldsymbol{m}\|_2^2}{n} \right]^2}_{\text{Onsager's correction term}}.$$

The TAP free energy

- The TAP free energy

$$\mathcal{F}_{\rm TAP}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle}_{\mathcal{F}_{\rm MF}} \underbrace{-\frac{n\lambda^2}{4}\Big[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\Big]^2}_{\text{Onsager's correction term}} .$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\rm TAP}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^\top \approx \widehat{\boldsymbol{X}}_{\rm Bayes} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top | \boldsymbol{Y}].$$

- Our main theorem shows that this is correct.

The TAP free energy

- The TAP free energy

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle}_{\mathcal{F}_{\mathrm{MF}}} \underbrace{-\frac{n\lambda^2}{4}\Big[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\Big]^2}_{\text{Onsager's correction term}}.$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} \approx \widehat{\boldsymbol{X}}_{\mathsf{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}} | \boldsymbol{Y}].$$

- Our main theorem shows that this is correct.

The TAP free energy

- The TAP free energy

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv \underbrace{-\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle}_{\mathcal{F}_{\mathrm{MF}}} \underbrace{-\frac{n\lambda^2}{4}\left[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\right]^2}_{\text{Onsager's correction term}} .$$

- For $\boldsymbol{m}_\star = \arg\min_{\boldsymbol{m}} \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})$, we hope

$$\boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T} \approx \widehat{\boldsymbol{X}}_{\mathrm{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T} | \boldsymbol{Y}].$$

- Our main theorem shows that this is correct.

Remarks

- Derivation of TAP can be obtained by expectation consistency/Plefka's expansion.
- The stationery equation of TAP $\nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = 0$ is the fixed point equation for AMP algorithm
$$\boldsymbol{m}^{k+1} = \mathrm{AMP}(\boldsymbol{m}^k, \boldsymbol{m}^{k-1}), \qquad \boldsymbol{m}^{k+1} = \boldsymbol{m}^k = \boldsymbol{m}^{k-1}.$$
- AMP is not a descent algorithm on TAP free energy.

Main theorem

Theorem (Fan, M., Montanari, 2018)

Denote $\mathcal{C}_{\lambda,n} = \{\boldsymbol{m} \in [-1,1]^n : \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = \boldsymbol{0}, \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \le -\lambda^2/3\}$. *There exists* $\lambda_0 > 0$, *such that for any* $\lambda > \lambda_0$, *we have*

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{\boldsymbol{m}\in\mathcal{C}_{\lambda,n}} \frac{1}{n^2}\|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 \wedge 1\Big] = 0. \tag{1}$$

All the critical points (below a threshold) are close to the Bayesian estimator.

Main theorem

Theorem (Fan, M., Montanari, 2018)

Denote $\mathcal{C}_{\lambda,n} = \{\boldsymbol{m} \in [-1,1]^n : \nabla\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = \boldsymbol{0}, \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \le -\lambda^2/3\}$. *There exists* $\lambda_0 > 0$, *such that for any* $\lambda > \lambda_0$, *we have*

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{\boldsymbol{m}\in\mathcal{C}_{\lambda,n}} \frac{1}{n^2}\|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\mathrm{Bayes}}\|_F^2 \wedge 1\Big] = 0. \tag{1}$$

All the critical points (below a threshold) are close to the Bayesian estimator.

Related literatures in spin glass theory

TAP free energy in unbiased SK.

- TAP equations: [Talagrand, 2004], [Chatterjee, 2009], [Chen, 2011], [Auffinger and Jagannath, 2016], Posterior means/Pure states satisfy TAP equations.
- TAP free energy: [Chen and Panchenko, 2017], constrained TAP minimum are exact.

Calculating the complexity.

- [Auffinger, Ben Arous, and Cerny, 2010], [Subag, 2016].

And a few more after 2018

Main theorem

Theorem (Fan, M., Montanari, 2018)

Denote $\mathcal{C}_{\lambda,n} = \{\boldsymbol{m} \in [-1,1]^n : \nabla \mathcal{F}_{\rm TAP}(\boldsymbol{m}) = \boldsymbol{0}, \mathcal{F}_{\rm TAP}(\boldsymbol{m}) \le -\lambda^2/3\}$. *There exists* $\lambda_0 > 0$, *such that for any* $\lambda > \lambda_0$, *we have*

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{\boldsymbol{m}\in\mathcal{C}_{\lambda,n}} \frac{1}{n^2}\|\boldsymbol{m}\boldsymbol{m}^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\rm Bayes}\|_F^2 \wedge 1\Big] = 0. \quad (1)$$

All the critical points (below a threshold) are close to the Bayesian estimator.

Proof idea - Count the number of critical points

- Recall

$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \equiv -\sum_{i=1}^{n} \mathsf{h}(m_i) - \frac{\lambda}{2}\langle \boldsymbol{m}, \boldsymbol{Y}\boldsymbol{m}\rangle - \frac{n\lambda^2}{4}\left[1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\right]^2.$$

- Define some important statistics of $\boldsymbol{m}$:

$$E(\boldsymbol{m}) = \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})/n, \quad Q(\boldsymbol{m}) = \|\boldsymbol{m}\|_2^2/n, \quad M(\boldsymbol{m}) = \langle \boldsymbol{m}, \boldsymbol{x}\rangle/n.$$

- For any $U \subseteq \mathbb{R}^3$, define

$$\mathrm{Crit}_n(U) \equiv \#\{\boldsymbol{m} : \nabla E(\boldsymbol{m}) = \boldsymbol{0}, (Q(\boldsymbol{m}), M(\boldsymbol{m}), E(\boldsymbol{m})) \in U\}. \quad (2)$$

Calculating the Crit: Kac-Rice formula

Lemma (Kac-Rice formula, c.f. [Adler and Taylor, 2007)

] Let $f : \mathbb{R}^d \to \mathbb{R}$ be a "sufficiently regular" random morse function. Let $p_{\boldsymbol{m}}(\boldsymbol{z})$ be the density of $\nabla f(\boldsymbol{m})$ at $\boldsymbol{z}$. For any Borel measurable set $T \subseteq \mathbb{R}^d$, denote

$$\mathrm{Crit}(T) = \#\{\boldsymbol{m} \in T : \nabla f(\boldsymbol{m}) = \boldsymbol{0}\}.$$

Then

$$\begin{aligned}\mathbb{E}[\mathrm{Crit}(T)] =& \mathbb{E}\Big[\int_T \big|\det \nabla^2 f(\boldsymbol{m})\big| \cdot \delta(\nabla f(\boldsymbol{m})) \cdot \mathrm{d}\boldsymbol{m}\Big] \\ =& \int_T \mathbb{E}\Big[\big|\det \nabla^2 f(\boldsymbol{m})\big| \Big| \nabla f(\boldsymbol{m}) = \boldsymbol{0}\Big] p_{\boldsymbol{m}}(\boldsymbol{0})\mathrm{d}\boldsymbol{m}.\end{aligned}$$

- $|\det \nabla^2 f(\boldsymbol{m})|$ is the correct weight function so that each critical point count exactly once.

Proof idea - Count the number of critical points

Proposition

$$\mathbb{E}[\mathrm{Crit}_n(U)] \le \exp\Big\{n \sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e) + o(n)\Big\}.$$

$$S_\star(q,\varphi,e) = \sup_{a\in\mathbb{R}} \inf_{(\mu,\nu,\tau,\gamma)\in\mathbb{R}^4} S(q,\varphi,a,e;\mu,\nu,\tau,\gamma),$$

where

$$\begin{aligned} S(q,\varphi,a,e;\mu,\nu,\tau,\gamma) =& \frac{1}{4\beta^2}\Big[\frac{a}{q} - \frac{\beta\lambda\varphi^2}{q} - \beta^2(1-q)\Big]^2 \\ & - q\mu - \varphi\nu - a\tau - \Big[-\frac{\beta^2}{4}(1-q^2) + \frac{a}{2} - e\Big]\gamma + \log I, \end{aligned}$$

and

$$\begin{aligned} I = \int_{-\infty}^{\infty} \frac{1}{(2\pi\beta^2 q)^{1/2}} \exp\Big\{ & -\frac{(x-\beta\lambda\varphi)^2}{2\beta^2 q} \\ & + \mu\tanh^2(x) + \nu\tanh(x) + \tau x\tanh(x) + \gamma\log[2\cosh(x)]\Big\}\mathrm{d}x. \end{aligned}$$

Proof idea - Count the number of critical points

- Key proposition: for $U \subseteq \mathbb{R}^3$,

$$\mathbb{E}[\mathrm{Crit}_n(U)] \leq \exp\Big\{n \overbrace{\sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e)}^{T(U)} + o(n)\Big\},$$

- For any U such that $T(U) > 0$, there could potentially be critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U.
- For any U such that $T(U) < 0$, there is no critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U, with high probability.
- If we admit the key proposition, suffice to show that $T(U) < 0$ unless U contains a neighborhood of the Bayes estimator.

Proof idea - Count the number of critical points

- Key proposition: for $U \subseteq \mathbb{R}^3$,

$$\mathbb{E}[\mathrm{Crit}_n(U)] \le \exp\Big\{n \overbrace{\sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e)}^{T(U)} + o(n)\Big\},$$

- For any U such that $T(U) > 0$, there could potentially be critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U.
- For any U such that $T(U) < 0$, there is no critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U, with high probability.
- If we admit the key proposition, suffice to show that $T(U) < 0$ unless U contains a neighborhood of the Bayes estimator.

Proof idea - Count the number of critical points

- Key proposition: for $U \subseteq \mathbb{R}^3$,

$$\mathbb{E}[\mathrm{Crit}_n(U)] \le \exp\Big\{ n \overbrace{\sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e)}^{T(U)} + o(n) \Big\},$$

- For any U such that $T(U) > 0$, there could potentially be critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U.
- For any U such that $T(U) < 0$, there is no critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U, with high probability.
- If we admit the key proposition, suffice to show that $T(U) < 0$ unless U contains a neighborhood of the Bayes estimator.

Proof idea - Count the number of critical points

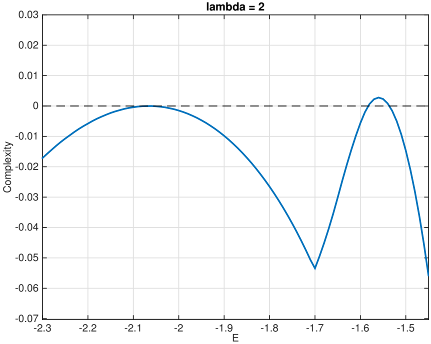
- Key proposition: for $U \subseteq \mathbb{R}^3$,

$$\mathbb{E}[\mathrm{Crit}_n(U)] \le \exp\Big\{ n \overbrace{\sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e)}^{T(U)} + o(n) \Big\},$$

- For any U such that $T(U) > 0$, there could potentially be critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U.
- For any U such that $T(U) < 0$, there is no critical points of $\mathcal{F}_{\mathrm{TAP}}$ in U, with high probability.
- If we admit the key proposition, suffice to show that $T(U) < 0$ unless U contains a neighborhood of the Bayes estimator.

Proof idea - the complexity function $S_\star$

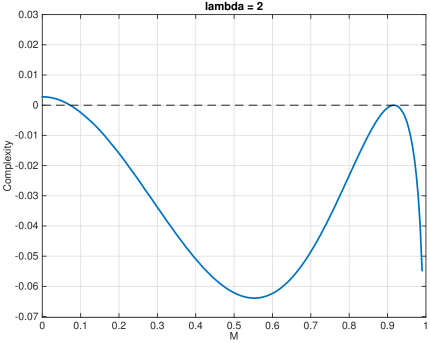
- $S_\star(e) = \sup_{q,\varphi} S_\star(q, \varphi, e)$.



- At $e_\star$, $S_\star(e_\star) = 0$.

Proof idea - the complexity function $S_\star$

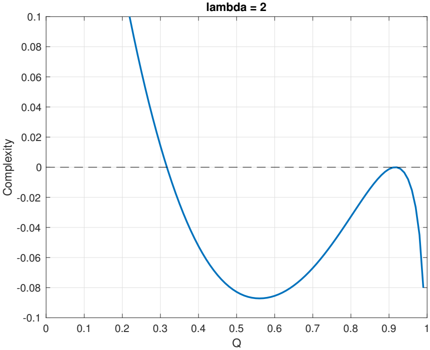
- $S_\star(\varphi) = \sup_{q,e} S_\star(q, \varphi, e)$.



- At $\varphi_\star$, $S_\star(\varphi_\star) = 0$.

Proof idea - the complexity function $S_\star$

- $S_\star(q) = \sup_{\varphi,e} S_\star(q, \varphi, e)$.



- At $q_\star$, $S_\star(q_\star) = 0$.

Proof idea - the complexity function $S_\star$

There exists λ_0, for $\lambda \geq \lambda_0$,

- $S_\star(q_\star, \varphi_\star, e_\star) = 0$, where $(q_\star, \varphi_\star, e_\star) \approx (Q(\boldsymbol{m}_\star), M(\boldsymbol{m}_\star), E(\boldsymbol{m}_\star))$ for $\widehat{\boldsymbol{X}}_{\mathrm{Bayes}} \approx \boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T}$.
- $S_\star(q, \varphi, e) < 0$ for any $e \leq -\lambda^2/3$ and $(q, \varphi, e) \neq (q_\star, \varphi_\star, e_\star)$.

The proof of these two properties is more than calculus. It requires bounds using concentration inequalities.

Combining with the key inequality it is easy to show the main theorem.

$$\mathbb{E}[\mathrm{Crit}_n(U)] \leq \exp\Big\{ n \sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e) + o(n) \Big\}.$$

Now suffice to show the key inequality.

Proof idea - the complexity function $S_\star$

There exists λ_0, for $\lambda \geq \lambda_0$,

- $S_\star(q_\star, \varphi_\star, e_\star) = 0$, where $(q_\star, \varphi_\star, e_\star) \approx (Q(\boldsymbol{m}_\star), M(\boldsymbol{m}_\star), E(\boldsymbol{m}_\star))$ for $\widehat{\boldsymbol{X}}_{\text{Bayes}} \approx \boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T}$.
- $S_\star(q, \varphi, e) < 0$ for any $e \leq -\lambda^2/3$ and $(q, \varphi, e) \neq (q_\star, \varphi_\star, e_\star)$.

The proof of these two properties is more than calculus. It requires bounds using concentration inequalities.

Combining with the key inequality it is easy to show the main theorem.

$$\mathbb{E}[\mathrm{Crit}_n(U)] \leq \exp\Big\{n \sup_{(q,\varphi,e)\in U} S_\star(q,\varphi,e) + o(n)\Big\}.$$

Now suffice to show the key inequality.

Proof idea - the complexity function $S_\star$

There exists λ_0, for $\lambda \geq \lambda_0$,

- $S_\star(q_\star, \varphi_\star, e_\star) = 0$, where $(q_\star, \varphi_\star, e_\star) \approx (Q(\boldsymbol{m}_\star), M(\boldsymbol{m}_\star), E(\boldsymbol{m}_\star))$ for $\widehat{\boldsymbol{X}}_{\text{Bayes}} \approx \boldsymbol{m}_\star \boldsymbol{m}_\star^\mathsf{T}$.
- $S_\star(q, \varphi, e) < 0$ for any $e \leq -\lambda^2/3$ and $(q, \varphi, e) \neq (q_\star, \varphi_\star, e_\star)$.

The proof of these two properties is more than calculus. It requires bounds using concentration inequalities.
Combining with the key inequality it is easy to show the main theorem.

$$\mathbb{E}[\mathrm{Crit}_n(U)] \leq \exp\Big\{ n \sup_{(q,\varphi,e)\in U} S_\star(q, \varphi, e) + o(n) \Big\}.$$

Now suffice to show the key inequality.

Dealing with determinant of Hessian

- The conditional Hessian is distributed as (up to some scaling)

  $$[\nabla^2 \mathcal{F}_{\rm TAP}(\boldsymbol{m}) | \nabla \mathcal{F}_{\rm TAP}(\boldsymbol{m}) = \boldsymbol{0}] \stackrel{d}{=} \boldsymbol{D} + \boldsymbol{W} + \text{low rank perturbation},$$

  where $\boldsymbol{D} = \text{diag}(d_i)$, and $\boldsymbol{W} \sim \text{GOE}(n)$.
- The low rank perturbation has vanishing effects. Therefore, we just need to calculate $\mathbb{E}[|\det(\boldsymbol{H})|]$, with

  $$\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W}.$$

Determinant of Hessian: "replica trick"

$$\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W} = \text{diagonal} + \text{GOE}.$$

We have

$$\begin{aligned}
\mathbb{E}[|\det(\boldsymbol{H})|] &\approx \mathbb{E}[|\det(\boldsymbol{H})|^{-1/2}]^2 \\
&= \mathbb{E}\Big[\int_{\mathbb{R}^d} \exp\{-\langle \boldsymbol{x}, \boldsymbol{H}\boldsymbol{x}\rangle/2\}\mathrm{d}\boldsymbol{x}\Big]^2 \\
&= \mathbb{E}\Big[\int_{\mathbb{R}^d} \exp\{-\langle \boldsymbol{x}, \boldsymbol{W}\boldsymbol{x}\rangle/2 - \langle \boldsymbol{x}, \boldsymbol{D}\boldsymbol{x}\rangle\}\mathrm{d}\boldsymbol{x}\Big]^2
\end{aligned}$$

Determinant of Hessian: Stieltjes transform

$$\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W} = \text{diagonal} + \text{GOE}.$$

$$\frac{1}{n} \log|\det(\boldsymbol{H})| = \frac{1}{n} \log \prod_{i=1}^{n} |\lambda_i(\boldsymbol{H})| = \int_{\mathbb{R}} \log|x| \cdot \mu_{\boldsymbol{H}}(\mathrm{d}x),$$

where $\mu_{\boldsymbol{H}} = (1/n) \sum_{i=1}^{n} \delta(\lambda_i(\boldsymbol{H}))$.

- The Stieltjes transform of $\mu_{\boldsymbol{H}}$ can be approximately calculated using free probability theory.
- Once the Stieltjes transform of $\mu_{\boldsymbol{H}}$ is known, the quantity $\mathbb{E}\Big[\int_{\mathbb{R}} (\log|x|) \mu_{\boldsymbol{H}}(\mathrm{d}x)\Big]$ can be computed.

Determinant of Hessian: Stieltjes transform

$$\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W} = \text{diagonal} + \text{GOE}.$$

$$\frac{1}{n} \log |\det(\boldsymbol{H})| = \frac{1}{n} \log \prod_{i=1}^{n} |\lambda_i(\boldsymbol{H})| = \int_{\mathbb{R}} \log |x| \cdot \mu_{\boldsymbol{H}}(\mathrm{d}x),$$

where $\mu_{\boldsymbol{H}} = (1/n) \sum_{i=1}^{n} \delta(\lambda_i(\boldsymbol{H}))$.

- The Stieltjes transform of $\mu_{\boldsymbol{H}}$ can be approximately calculated using free probability theory.
- Once the Stieltjes transform of $\mu_{\boldsymbol{H}}$ is known, the quantity $\mathbb{E}\Big[\int_{\mathbb{R}} (\log |x|) \mu_{\boldsymbol{H}}(\mathrm{d}x)\Big]$ can be computed.

Free convolution of two distribution

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, and $\mu_{\boldsymbol{A}} = (1/n)\sum_{i=1}^n \delta(\lambda_i(\boldsymbol{A}))$. For any $z \in \mathbb{C}_+$, the Stieltjes transform of $\mu_{\boldsymbol{A}}$ is defined as

$$g_{\boldsymbol{A}}(z) = \int_{\mathbb{R}} \frac{1}{x-z}\mu_{\boldsymbol{A}}(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i(\boldsymbol{A}) - z}.$$

Lemma (Free probability theory heuristics/Leave one out)

Let $\boldsymbol{D} = \mathrm{diag}(d_i)$ be a diagonal matrix, and let $\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W}$. Then

$$\mathbb{E}g_{\boldsymbol{H}}(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{d_i - z - \mathbb{E}g_{\boldsymbol{H}}(z)} + o_n(1). \tag{3}$$

Free convolution of two distribution

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, and $\mu_{\boldsymbol{A}} = (1/n) \sum_{i=1}^{n} \delta(\lambda_i(\boldsymbol{A}))$. For any $z \in \mathbb{C}_+$, the Stieltjes transform of $\mu_{\boldsymbol{A}}$ is defined as

$$g_{\boldsymbol{A}}(z) = \int_{\mathbb{R}} \frac{1}{x - z} \mu_{\boldsymbol{A}}(\mathrm{d}x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i(\boldsymbol{A}) - z}.$$

Lemma (Free probability theory heuristics/Leave one out)

Let $\boldsymbol{D} = \mathrm{diag}(d_i)$ *be a diagonal matrix, and let* $\boldsymbol{H} = \boldsymbol{D} + \boldsymbol{W}$. *Then*

$$\mathbb{E} g_{\boldsymbol{H}}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{d_i - z - \mathbb{E} g_{\boldsymbol{H}}(z)} + o_n(1). \tag{3}$$

Summary

- TAP free energy is accurate for $\mathbb{Z}_2$ synchronization.
- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- More sophiscated variational inference beyond naive mean field.

Summary

- TAP free energy is accurate for $\mathbb{Z}_2$ synchronization.
- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- More sophiscated variational inference beyond naive mean field.

Summary

- TAP free energy is accurate for $\mathbb{Z}_2$ synchronization.
- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- More sophiscated variational inference beyond naive mean field.