Variational inference, spin glasses, and TAP free energy

Song Mei

Stanford University

September 19, 2018

Joint work with Zhou Fan and Andrea Montanari

Song Mei (Stanford University)

TAP free energy

< □ > < ≥ > < ≥ > ≥
 September 19, 2018

1 / 29

< 口 > < 同 >

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....
 ... within which "MF free energy" is known to be not optimal.
- ► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

イロト 不得 とうせい きゅうしょう

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....
 ... within which "MF free energy" is known to be not optimal.
- ► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

イロト 不得 とうせい きゅうしょう

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....
 ... within which "MF free energy" is known to be not optimal.
- ► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....

... within which "MF free energy" is known to be not optimal.

► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

・ロット (雪) (き) (き) (き)

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....
 - ... within which "MF free energy" is known to be not optimal.
- ► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

◆□ ▶ ◆□ ▶ ▲ ■ ▶ ▲ ■ ▶ ● ● ●

- Bayesian inference: high dimensional integration is hard!
- ► Variational inference: integration/summation → optimization. A popular objective function: "mean field free energy".
- Applications: topic modeling, stochastic block model, low rank matrix estimation, compressed sensing....
 - ... within which "MF free energy" is known to be not optimal.
- ► Today: introduce the optimal objective "TAP free energy", and provide rigorous results.

► Signal:

 $oldsymbol{x} = [oldsymbol{x}_1, \dots, oldsymbol{x}_n]^\mathsf{T} \in \mathbb{Z}_2^n, \quad oldsymbol{x}_i \stackrel{i.i.d.}{\sim} ext{Unif}(\mathbb{Z}_2), \quad \mathbb{Z}_2 = \{+1, -1\}.$

$$Y_{ij} = rac{\lambda}{n} x_i x_j + W_{ij}.$$

- Noise $W_{ij} \sim \mathcal{N}(0, 1/n)$.
- ▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.
- In matrix notation:

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

▶ Task: given $\mathbf{Y} = (Y_{ij})$, estimate \mathbf{x} (or say $\mathbf{X} = \mathbf{x}\mathbf{x}^{\mathsf{T}}$).

► Signal:

 $oldsymbol{x} = [x_1,\ldots,x_n]^{\mathsf{T}} \in \mathbb{Z}_2^n, \quad oldsymbol{x}_i \stackrel{i.i.d.}{\sim} \operatorname{Unif}(\mathbb{Z}_2), \quad \mathbb{Z}_2 = \{+1,-1\}.$

• Observation: for $1 \leq i < j \leq n$

$$Y_{ij} = rac{\lambda}{n} x_i x_j + W_{ij}.$$

• Noise
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$
.

▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.

In matrix notation:

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

▶ Task: given $Y = (Y_{ij})$, estimate x (or say $X = xx^{\top}$).

► Signal:

 $oldsymbol{x} = [x_1,\ldots,x_n]^\mathsf{T} \in \mathbb{Z}_2^n, \quad oldsymbol{x}_i \stackrel{i.i.d.}{\sim} \operatorname{Unif}(\mathbb{Z}_2), \quad \mathbb{Z}_2 = \{+1,-1\}.$

• Observation: for $1 \leq i < j \leq n$

$$Y_{ij} = rac{\lambda}{n} x_i x_j + W_{ij}.$$

- Noise $W_{ij} \sim \mathcal{N}(0, 1/n)$.
- ▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.
- ▶ In matrix notation:

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

▶ Task: given $Y = (Y_{ij})$, estimate x (or say $X = xx^{\top}$).

◆□ ▶ ◆□ ▶ ▲ ■ ▶ ▲ ■ ▶ ● ● ●

► Signal:

 $oldsymbol{x} = [x_1,\ldots,x_n]^\mathsf{T} \in \mathbb{Z}_2^n, \quad oldsymbol{x}_i \stackrel{i.i.d.}{\sim} \operatorname{Unif}(\mathbb{Z}_2), \quad \mathbb{Z}_2 = \{+1,-1\}.$

• Observation: for $1 \leq i < j \leq n$

$$Y_{ij} = rac{\lambda}{n} x_i x_j + W_{ij}.$$

• Noise
$$W_{ij} \sim \mathcal{N}(0, 1/n)$$
.

▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.

In matrix notation:

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

▶ Task: given $Y = (Y_{ij})$, estimate x (or say $X = xx^{\top}$).

► Signal:

 $oldsymbol{x} = [x_1, \ldots, x_n]^\mathsf{T} \in \mathbb{Z}_2^n, \quad oldsymbol{x}_i \stackrel{i.i.d.}{\sim} \operatorname{Unif}(\mathbb{Z}_2), \quad \mathbb{Z}_2 = \{+1, -1\}.$

• Observation: for $1 \leq i < j \leq n$

$$Y_{ij} = rac{\lambda}{n} x_i x_j + W_{ij}.$$

- Noise $W_{ij} \sim \mathcal{N}(0, 1/n)$.
- ▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.
- In matrix notation:

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

▶ Task: given $Y = (Y_{ij})$, estimate x (or say $X = xx^{\top}$).

Signal:

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

• Observation: for 1 < i < j < n

$$Y_{ij} = \frac{\lambda}{n} x_i x_j + W_{ij}.$$

- ▶ Noise $W_{ii} \sim \mathcal{N}(0, 1/n)$.
- ▶ SNR $\lambda \in [0, \infty)$ fixed, dimension $n \to \infty$.
- In matrix notation:

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

• Task: given $Y = (Y_{ij})$, estimate x (or say $X = xx^{\top}$).

Settings:

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

• Estimate $X = xx^{\mathsf{T}}$ with loss:

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

• For $\lambda < 1$, estimation is impossible.

- ▶ For λ > 1, estimation is possible and efficient, e.g., spectral estimator (Baik, Ben Arous, Peche phase transition).
- The optimal estimator is the Bayes estimator (also minimax estimator):

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

イロト 不得 とうせい きゅうしょう

Settings:

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

• Estimate $X = xx^{\mathsf{T}}$ with loss:

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

• For $\lambda < 1$, estimation is impossible.

- ▶ For λ > 1, estimation is possible and efficient, e.g., spectral estimator (Baik, Ben Arous, Peche phase transition).
- The optimal estimator is the Bayes estimator (also minimax estimator):

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

- 3

・ロッ ・ 一 ・ ・ ・ ・ ・ ・ ・ ・

Settings:

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

• Estimate $X = xx^{\mathsf{T}}$ with loss:

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

• For $\lambda < 1$, estimation is impossible.

- For λ > 1, estimation is possible and efficient, e.g., spectral estimator (Baik, Ben Arous, Peche phase transition).
- The optimal estimator is the Bayes estimator (also minimax estimator):

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

イロト 不得 とうせい きゅうしょう

Settings:

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

• Estimate $X = xx^{\top}$ with loss:

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

- For $\lambda < 1$, estimation is impossible.
- For λ > 1, estimation is possible and efficient, e.g., spectral estimator (Baik, Ben Arous, Peche phase transition).
- The optimal estimator is the Bayes estimator (also minimax estimator):

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

・ロト ・ 一日 ・ ・ 日 ・ ・ 日 ・ ・ 日

Settings:

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

• Estimate $X = xx^{\top}$ with loss:

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

- For $\lambda < 1$, estimation is impossible.
- For λ > 1, estimation is possible and efficient, e.g., spectral estimator (Baik, Ben Arous, Peche phase transition).
- The optimal estimator is the Bayes estimator (also minimax estimator):

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

► Settings:

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

Risk:

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



5 / 29

< □ > < / → >

Compute the Bayesian estimator

▶ The Bayesian estimator:

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

$$p(oldsymbol{\sigma}|oldsymbol{Y}) = rac{1}{Z} \exp\{\lambda \langle oldsymbol{\sigma},oldsymbol{Y}oldsymbol{\sigma}
angle / 2\}.$$

Compute the Bayesian estimator

▶ The Bayesian estimator:

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

▶ The posterior distribution:

$$p(\sigma|\mathbf{Y}) = rac{1}{Z} \exp\{ \lambda \langle \sigma, \mathbf{Y} \sigma
angle / 2 \}.$$

▶ The posterior distribution:

$$p(oldsymbol{\sigma}|oldsymbol{Y}) = rac{1}{Z} \exp\{oldsymbol{\lambda} \langle oldsymbol{\sigma}, oldsymbol{Y} oldsymbol{\sigma}
angle / 2\}.$$

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

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

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

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

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

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - \lambda \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2 \geq -\log Z,$$

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

• The posterior distribution:

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

• Approximate $p(\sigma|Y)$ by $q \in \mathcal{P}_{\mathrm{MF}}$:

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

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

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

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

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - \lambda \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2 \geq -\log Z,$$

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

• The posterior distribution:

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

• Approximate $p(\sigma|Y)$ by $q \in \mathcal{P}_{\mathrm{MF}}$:

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

• Minimize the relative entropy between q and $p(\sigma|Y)$:

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

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

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(m_i) - \lambda \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle 2 - \log Z,$$

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

• The posterior distribution:

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

• Approximate $p(\sigma|Y)$ by $q \in \mathcal{P}_{\mathrm{MF}}$:

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

• Minimize the relative entropy between q and $p(\sigma|Y)$:

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

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

$$\mathcal{F}_{ ext{MF}}(m)\equiv -\sum_{i=1}^n \mathsf{h}(m_i)-oldsymbol{\lambda}\langle m, Ym
angle/2\geq -\log Z,$$

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

▶ Mean field free energy:

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - oldsymbol{\lambda} \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2.$$

For $m_{\star} = \operatorname{arg\,min}_{m} \mathcal{F}_{\mathrm{MF}}(m)$, we hope

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

- ▶ It was shown that $m_{\star}m_{\star}^{\top} \approx \mathbb{E}[xx^{\top}|Y]$ [Ghorbani, Javadi, and Montanari, 2017].
- ▶ The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

Mean field free energy:

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - oldsymbol{\lambda} \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2.$$

▶ For
$$m_{\star} = \arg\min_{m} \mathcal{F}_{\mathrm{MF}}(m)$$
, we hope $m_{\star}m_{\star}^{\mathsf{T}} pprox \widehat{oldsymbol{X}}_{\mathrm{Bayes}} = \mathbb{E}[xx^{\mathsf{T}}|oldsymbol{Y}].$

- ▶ It was shown that $m_{\star}m_{\star}^{\top} \approx \mathbb{E}[xx^{\top}|Y]$ [Ghorbani, Javadi, and Montanari, 2017].
- ► The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

▶ Mean field free energy:

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - oldsymbol{\lambda} \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2.$$

▶ For
$$m_{\star} = rgmin_{m} \mathcal{F}_{MF}(m)$$
, we hope

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

- ▶ It was shown that $m_{\star}m_{\star}^{\top} \approx \mathbb{E}[xx^{\top}|Y]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

Mean field free energy:

$$\mathcal{F}_{ ext{MF}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(oldsymbol{m}_i) - oldsymbol{\lambda} \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle / 2.$$

▶ For
$$m_{\star} = rgmin_m \mathcal{F}_{\mathsf{MF}}(m)$$
, we hope

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

- ▶ It was shown that $m_{\star}m_{\star}^{\top} \approx \mathbb{E}[xx^{\top}|Y]$ [Ghorbani, Javadi, and Montanari, 2017].
- The assumption that posterior distribution can be approximately factorized into the product of marginals is wrong!

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

$$G_{eta,oldsymbol{\lambda}}(oldsymbol{\sigma}) = rac{1}{Z_{eta,oldsymbol{\lambda}}} \exp\{eta\langle oldsymbol{\sigma},oldsymbol{Y}oldsymbol{\sigma}
angle\}.$$

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

- When β = λ, the Gibbs measure of SK model is the same as the posterior of Z₂ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

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

イロト 不得 とうせい きゅうしょう

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

$$G_{eta,oldsymbol{\lambda}}(oldsymbol{\sigma}) = rac{1}{Z_{eta,oldsymbol{\lambda}}} \exp\{eta\langle oldsymbol{\sigma},oldsymbol{Y}oldsymbol{\sigma}
angle\}.$$

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

- When β = λ, the Gibbs measure of SK model is the same as the posterior of Z₂ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

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

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

$$G_{eta,oldsymbol{\lambda}}(oldsymbol{\sigma}) = rac{1}{Z_{eta,oldsymbol{\lambda}}} \exp\{eta\langle oldsymbol{\sigma},oldsymbol{Y}oldsymbol{\sigma}
angle\}.$$

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

- When β = λ, the Gibbs measure of SK model is the same as the posterior of Z₂ synchronization
- The TAP free energy (when $\beta = \lambda$) gives

$${\cal F}_{
m TAP}({m m})\equiv \underbrace{-\sum\limits_{i=1}^n {\sf h}({m m}_i)-rac{\lambda}{2}\langle{m m},{m Y}{m m}
angle}_{{\cal F}_{
m MF}}\underbrace{-rac{n\lambda^2}{4}\Big[1-rac{\|{m m}\|_2^2}{n}\Big]^2}_{
m Onsager's\ correction\ term}}$$

・ロト ・ 一日 ・ ・ 日 ・ ・ 日 ・ ・ 日 ・

► The TAP free energy

$${\mathcal F}_{\mathrm{TAP}}(m) \equiv \underbrace{-\sum\limits_{i=1}^n \mathsf{h}(m_i) - rac{\lambda}{2} \langle m, Ym
angle}_{{\mathcal F}_{\mathrm{MF}}} \underbrace{-rac{n\lambda^2}{4} \Big[1 - rac{\|m\|_2^2}{n} \Big]^2}_{\mathrm{Onsager's \ correction \ term}}.$$

▶ For $m_{\star} = rgmin_{m} \mathcal{F}_{\mathrm{TAP}}(m)$, we hope

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

Our main theorem shows that this is correct.

э

10 / 29

< ロ > < 同 > < 回 > < 回 > < 回 > <

▶ The TAP free energy

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

• For $m_{\star} = \arg\min_{m} \mathcal{F}_{\text{TAP}}(m)$, we hope

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

< (□)

э.

▶ The TAP free energy

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

▶ For $m_{\star} = rgmin_{m} \mathcal{F}_{\texttt{TAP}}(m)$, we hope

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

• Our main theorem shows that this is correct.

B 🖌 🖌 B 🛌 👘

э.

Proof of the main theorem

Theorem (Fan, M., Montanari, 2018)

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

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

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

Proof of the main theorem

Theorem (Fan, M., Montanari, 2018)

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

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

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

Another way to construct the Bayes estimator is approximate message passing [Donoho, Maleki, and Montanari, 2009], [Bolthausen, 2014]:

$$oldsymbol{m}^{k+1} = anh({oldsymbol{\lambda}}oldsymbol{Y}oldsymbol{m}^k - \lambda^2 [1-\|oldsymbol{m}^k\|_2^2/n]oldsymbol{m}^{k-1})$$
 .

- ▶ Fixed point of AMP is a critical point of the TAP free energy.
- The risk of AMP iterations converge to the Bayes risk [Deshpande, Abbes, and Montanari, 2016], [Montanari and Venkataramanan, 2017]:

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{1}{n^2}\|\boldsymbol{m}^k(\boldsymbol{m}^k)^{\mathsf{T}}-\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2=\lim_{n\to\infty} {\tt MSE}_n(\widehat{\boldsymbol{X}}_{\tt Bayes}).$$

But it is not known if AMP will converge to a fixed point (It is still an open problem).

12 / 29

< ロ > < 同 > < 回 > < 回 > < 回 > <

Another way to construct the Bayes estimator is approximate message passing [Donoho, Maleki, and Montanari, 2009], [Bolthausen, 2014]:

$$m{m}^{k+1} = anh(m{\lambda} m{Y} m{m}^k - m{\lambda}^2 [1 - \|m{m}^k\|_2^2 / n] m{m}^{k-1})$$
 .

- Fixed point of AMP is a critical point of the TAP free energy.
- The risk of AMP iterations converge to the Bayes risk [Deshpande, Abbes, and Montanari, 2016], [Montanari and Venkataramanan, 2017]:

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{1}{n^2}\|\boldsymbol{m}^k(\boldsymbol{m}^k)^{\mathsf{T}}-\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2=\lim_{n\to\infty} {\tt MSE}_n(\widehat{\boldsymbol{X}}_{{\tt Bayes}}).$$

But it is not known if AMP will converge to a fixed point (It is still an open problem).

12 / 29

イロト イボト イヨト イヨト

Another way to construct the Bayes estimator is approximate message passing [Donoho, Maleki, and Montanari, 2009], [Bolthausen, 2014]:

$$oldsymbol{m}^{k+1} = anh(oldsymbol{\lambda}oldsymbol{Y}oldsymbol{m}^k - \lambda^2 [1 - \|oldsymbol{m}^k\|_2^2/n]oldsymbol{m}^{k-1})$$
 .

- Fixed point of AMP is a critical point of the TAP free energy.
- The risk of AMP iterations converge to the Bayes risk [Deshpande, Abbes, and Montanari, 2016], [Montanari and Venkataramanan, 2017]:

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{1}{n^2}\|\boldsymbol{m}^k(\boldsymbol{m}^k)^{\mathsf{T}}-\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2=\lim_{n\to\infty} {\tt MSE}_n(\widehat{\boldsymbol{X}}_{\tt Bayes}).$$

But it is not known if AMP will converge to a fixed point (It is still an open problem).

12 / 29

Another way to construct the Bayes estimator is approximate message passing [Donoho, Maleki, and Montanari, 2009], [Bolthausen, 2014]:

$$oldsymbol{m}^{k+1} = anh(oldsymbol{\lambda}oldsymbol{Y}oldsymbol{m}^k - \lambda^2 [1 - \|oldsymbol{m}^k\|_2^2/n]oldsymbol{m}^{k-1})$$
 .

- Fixed point of AMP is a critical point of the TAP free energy.
- The risk of AMP iterations converge to the Bayes risk [Deshpande, Abbes, and Montanari, 2016], [Montanari and Venkataramanan, 2017]:

$$\lim_{k\to\infty}\lim_{n\to\infty}\frac{1}{n^2}\|\boldsymbol{m}^k(\boldsymbol{m}^k)^{\mathsf{T}}-\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\|_F^2=\lim_{n\to\infty}\mathtt{MSE}_n(\widehat{\boldsymbol{X}}_{\mathtt{Bayes}}).$$

 But it is not known if AMP will converge to a fixed point (It is still an open problem).

・ロト ・ 一下・ ・ ヨト ・ ・ ヨト

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].

Proof of the main theorem

Theorem (Fan, M., Montanari, 2018)

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

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

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

Recall

$$\mathcal{F}_{ ext{TAP}}(oldsymbol{m}) \equiv -\sum_{i=1}^n \mathsf{h}(m_i) - rac{\lambda}{2} \langle oldsymbol{m}, oldsymbol{Y}oldsymbol{m}
angle - rac{n\lambda^2}{4} \Big[1 - rac{\|oldsymbol{m}\|_2^2}{n} \Big]^2.$$

Define some important statistics of m:

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

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

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

Proposition

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

Song Mei (Stanford University)

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

where

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

and

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

Song Mei (Stanford University)

September 19, 2018

16 / 29

イロト 不得 とうせい きゅうしょう

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

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

- For any U such that T(U) > 0, there could potentially be critical points of \mathcal{F}_{TAP} in U.
- ▶ For any U such that T(U) < 0, there is no critical points of F_{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.

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

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

- ▶ For any U such that T(U) > 0, there could potentially be critical points of F_{TAP} in U.
- ▶ For any U such that T(U) < 0, there is no critical points of F_{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.

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

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

- ▶ For any U such that T(U) > 0, there could potentially be critical points of F_{TAP} in U.
- ▶ For any U such that T(U) < 0, there is no critical points of F_{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.

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

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

- ▶ For any U such that T(U) > 0, there could potentially be critical points of F_{TAP} in U.
- ▶ For any U such that T(U) < 0, there is no critical points of F_{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.

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



• At e_{\star} , $S_{\star}(e_{\star}) = 0$.

< 口 > < 同 >

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



• At φ_{\star} , $S_{\star}(\varphi_{\star}) = 0$.

< 口 > < 同 >

(大臣) (大臣) (臣)

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



• At q_{\star} , $S_{\star}(q_{\star}) = 0$.

< 口 > < 同 >

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

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

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.

$$\mathbb{E}[\operatorname{Crit}_n(U)] \leq \expig\{n\sup_{(q,arphi,e)\in U}S_\star(q,arphi,e)+o(n)ig\}.$$

Now suffice to show the key inequality.

・ロト ・ 一日 ・ ・ 日 ・ ・ 日 ・ ・ 日

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

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

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.

$$\mathbb{E}[\operatorname{Crit}_n(U)] \leq \expig\{n\sup_{(q,arphi,e)\in U}S_\star(q,arphi,e)+o(n)ig\}.$$

Now suffice to show the key inequality.

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

- $\blacktriangleright S_{\star}(q_{\star},\varphi_{\star},e_{\star})=0, \text{ where } (q_{\star},\varphi_{\star},e_{\star})\approx (Q(m_{\star}),M(m_{\star}),E(m_{\star}))$ for $\widehat{X}_{ ext{Baves}} pprox m_{\star} m_{\star}^{\mathsf{T}}$.
- ► $S_{\star}(q, \varphi, e) < 0$ for any $e < -\lambda^2/3$ 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.

$$\mathbb{E}[\operatorname{Crit}_n(U)] \leq \expig\{n\sup_{(q,arphi,e)\in U}S_\star(q,arphi,e)+o(n)ig\}.$$

Now suffice to show the key inequality.

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 $p_m(z)$ be the density of $\nabla f(m)$ at z. For any Borel measurable set $T \subseteq \mathbb{R}^d$, denote

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

Then

$$egin{aligned} \mathbb{E}[\operatorname{Crit}(T)] = & \mathbb{E}\Big[\int_{T} \big|\det
abla^2 f(oldsymbol{m})ig| \cdot \delta(
abla f(oldsymbol{m}))\cdot \mathrm{d}oldsymbol{m}\Big] \ &= \int_{T} \mathbb{E}\Big[ig|\det
abla^2 f(oldsymbol{m})ig|ig|
abla f(oldsymbol{m}) = oldsymbol{0}\Big] p_{oldsymbol{m}}(oldsymbol{0})\mathrm{d}oldsymbol{m}. \end{aligned}$$

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

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

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

where $D = \text{diag}(d_i)$, and $W \sim \text{GOE}(n)$.

Song Mei (Stanford University)

・ロト ・ 日 ・ ・ ヨ ・ ・ 日 ・ ・ う へ つ ・ September 19, 2018

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

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

where $D = \text{diag}(d_i)$, and $W \sim \text{GOE}(n)$.

▶ The low rank perturbation has vanishing effects. Therefore, we just need to calculate $\mathbb{E}[|\det(H)|]$, with

$$H = D + W.$$

$$m{H} = m{D} + m{W} = ext{diagonal} + ext{GOE}.$$

 $rac{1}{n} \log \mathbb{E}[|\det(m{H})|] = rac{1}{n} \log \mathbb{E}\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big] pprox rac{1}{n} \log\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big]$
 $= rac{1}{n} \sum_{i=1}^{n} \log |\lambda_i(m{H})| = \int_{\mathbb{R}} \log |x| \cdot \mu_H(ext{d}x) pprox \mathbb{E}\Big[\int_{\mathbb{R}} \log |x| \cdot \mu_H(ext{d}x)\Big].$

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

- Approximate equalities are due to concentration.
- ▶ The Stieltjes transform of μ_H can be approximately calculated using free probability theory.
- Once the Stieltjes transform of μ_H is known, the quantity $\mathbb{E}\left[\int_{\mathbb{R}} (\log |x|) \mu_H(\mathrm{d}x)\right]$ can be computed.

$$oldsymbol{H} = oldsymbol{D} + oldsymbol{W} = ext{diagonal} + ext{GOE}.$$

 $rac{1}{n} \log \mathbb{E}[|\det(oldsymbol{H})|] = rac{1}{n} \log \mathbb{E}\Big[\prod_{i=1}^n |\lambda_i(oldsymbol{H})|\Big] pprox rac{1}{n} \log\Big[\prod_{i=1}^n |\lambda_i(oldsymbol{H})|\Big]$
 $= rac{1}{n} \sum_{i=1}^n \log |\lambda_i(oldsymbol{H})| = \int_{\mathbb{R}} \log |x| \cdot \mu_H(ext{d}x) pprox \mathbb{E}\Big[\int_{\mathbb{R}} \log |x| \cdot \mu_H(ext{d}x)\Big].$

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

- Approximate equalities are due to concentration.
- The Stieltjes transform of µ_H can be approximately calculated using free probability theory.
- Once the Stieltjes transform of μ_H is known, the quantity $\mathbb{E}\left[\int_{\mathbb{R}} (\log |x|) \mu_H(\mathrm{d}x)\right]$ can be computed.

Free convolution of two distribution

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

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

Lemma (Due to free probability theory)

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

$$\mathbb{E}g_{H}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{d_{i} - z - \mathbb{E}g_{H}(z)} + o_{n}(1).$$
(3)

Free convolution of two distribution

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

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

Lemma (Due to free probability theory) Let $D = \text{diag}(d_i)$ be a diagonal matrix, and let H = D + W. Then

$$\mathbb{E}g_{H}(z) = rac{1}{n}\sum_{i=1}^{n}rac{1}{d_{i}-z-\mathbb{E}g_{H}(z)} + o_{n}(1).$$
 (3)

Song Mei (Stanford University)

・ロト ・日下 ・日下 ・日下 ・ うへつ September 19, 2018

$$m{H} = m{D} + m{W} = ext{diagonal} + ext{GOE}.$$

 $rac{1}{n} \log \mathbb{E}[|\det(m{H})|] = rac{1}{n} \log \mathbb{E}\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big] pprox rac{1}{n} \log\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big]$
 $= rac{1}{n} \sum_{i=1}^{n} \log |\lambda_i(m{H})| = \int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x) pprox \mathbb{E}\Big[\int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x)\Big].$

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

- Approximate equalities are due to concentration.
- The Stieltjes transform of µ_H can be approximately calculated using free probability theory.
- Once the Stieltjes transform of μ_H is known, the quantity $\mathbb{E}\left[\int_{\mathbb{R}} (\log |x|) \mu_H(\mathrm{d}x)\right]$ can be computed.

$\text{Calculate } \mathbb{E}[\text{f}_{\mathbb{R}} \log |x| \cdot \mu_{\boldsymbol{H}}(\mathrm{d} x)]$

Define

$$B(t) = \mathbb{E}\int_{\mathbb{R}} \log(x-it) \mu_{H}(\mathrm{d}x).$$

We have

$$egin{aligned} &\mathcal{R}B(0+) = &\mathbb{E}\int_{\mathbb{R}} \log |x| \cdot \mu_{oldsymbol{H}}(\mathrm{d}x), \ &B'(t) = -\,i\mathbb{E}\int_{\mathbb{R}} [1/(x-it)]\mu_{oldsymbol{H}}(\mathrm{d}x) = -i\mathbb{E}[g_{oldsymbol{H}}(it)]. \end{aligned}$$

We guess a formula

$$ilde{B}(t) = rac{1}{n}\sum_{i=1}^n \log(d_i - it - \mathbb{E}g_{oldsymbol{H}}(it)) + rac{1}{2}[\mathbb{E}g_{oldsymbol{H}}(it)]^2.$$

Then $\tilde{B}(t)$ satisfy all the conditions that B(t) approximately satisfy, so that $\tilde{B}(t) = B(t) + o_n(1)$.

Hence

$$\frac{1}{n}\log\mathbb{E}[|\det(\boldsymbol{H})|] = \tilde{B}(0) + o_n(1).$$

Song Mei (Stanford University)

$ext{Calculate } \mathbb{E}[\int_{\mathbb{R}} \log |x| \cdot \mu_{H}(\mathrm{d}x)]$

Define

$$B(t) = \mathbb{E}\int_{\mathbb{R}} \log(x-it) \mu_{oldsymbol{H}}(\mathrm{d} x).$$

We have

$$egin{aligned} \Re B(0+) =& \mathbb{E}\int_{\mathbb{R}}\log|x|\cdot\mu_{H}(\mathrm{d}x),\ B'(t) =& -i\mathbb{E}\int_{\mathbb{R}}[1/(x-it)]\mu_{H}(\mathrm{d}x) =& -i\mathbb{E}[g_{H}(it)]. \end{aligned}$$

We guess a formula

$$ilde{B}(t) = rac{1}{n}\sum_{i=1}^n \log(d_i - it - \mathbb{E}g_{oldsymbol{H}}(it)) + rac{1}{2}[\mathbb{E}g_{oldsymbol{H}}(it)]^2.$$

Then $\tilde{B}(t)$ satisfy all the conditions that B(t) approximately satisfy, so that $\tilde{B}(t) = B(t) + o_n(1)$.

Hence

$$\frac{1}{n}\log\mathbb{E}[|\det(\boldsymbol{H})|] = \tilde{B}(0) + o_n(1).$$

Song Mei (Stanford University)

Calculate $\mathbb{E}[\int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x)]$

Define

$$B(t) = \mathbb{E}\int_{\mathbb{R}} \log(x-it) \mu_{oldsymbol{H}}(\mathrm{d}x).$$

▶ We have

$$egin{aligned} \Re B(0+) =& \mathbb{E}\int_{\mathbb{R}}\log|x|\cdot\mu_{H}(\mathrm{d}x),\ B'(t) =& -i\mathbb{E}\int_{\mathbb{R}}[1/(x-it)]\mu_{H}(\mathrm{d}x) =& -i\mathbb{E}[g_{H}(it)]. \end{aligned}$$

We guess a formula

$$ilde{B}(t) = rac{1}{n}\sum_{i=1}^n \log(d_i - it - \mathbb{E}g_{oldsymbol{H}}(it)) + rac{1}{2}[\mathbb{E}g_{oldsymbol{H}}(it)]^2.$$

Then $\tilde{B}(t)$ satisfy all the conditions that B(t) approximately satisfy, so that $\tilde{B}(t) = B(t) + o_n(1)$.

Hence

$$\frac{1}{n}\log\mathbb{E}[|\det(\boldsymbol{H})|] = \tilde{B}(0) + o_n(1).$$

Song Mei (Stanford University)

TAP free energy

Calculate $\mathbb{E}[\int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x)]$

Define

$$B(t) = \mathbb{E}\int_{\mathbb{R}} \log(x-it) \mu_{oldsymbol{H}}(\mathrm{d} x).$$

We have

$$egin{aligned} \Re B(0+) =& \mathbb{E}\int_{\mathbb{R}}\log|x|\cdot\mu_{H}(\mathrm{d}x),\ B'(t) =& -i\mathbb{E}\int_{\mathbb{R}}[1/(x-it)]\mu_{H}(\mathrm{d}x) =& -i\mathbb{E}[g_{H}(it)]. \end{aligned}$$

We guess a formula

$$ilde{B}(t) = rac{1}{n}\sum_{i=1}^n \log(d_i - it - \mathbb{E}g_{oldsymbol{H}}(it)) + rac{1}{2}[\mathbb{E}g_{oldsymbol{H}}(it)]^2.$$

Then $\tilde{B}(t)$ satisfy all the conditions that B(t) approximately satisfy, so that $\tilde{B}(t) = B(t) + o_n(1)$.

Hence

$$\frac{1}{n}\log\mathbb{E}[|\det(\boldsymbol{H})|] = \tilde{B}(0) + o_n(1).$$

Song Mei (Stanford University)

▶ < E > E

$$m{H} = m{D} + m{W} = ext{diagonal} + ext{GOE}.$$

 $rac{1}{n} \log \mathbb{E}[|\det(m{H})|] = rac{1}{n} \log \mathbb{E}\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big] pprox rac{1}{n} \log\Big[\prod_{i=1}^{n} |\lambda_i(m{H})|\Big]$
 $= rac{1}{n} \sum_{i=1}^{n} \log |\lambda_i(m{H})| = \int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x) pprox \mathbb{E}\Big[\int_{\mathbb{R}} \log |x| \cdot \mu_H(\mathrm{d}x)\Big].$

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

- Approximate equalities are due to concentration.
- The Stieltjes transform of µ_H can be approximately calculated using free probability theory.
- Once the Stieltjes transform of μ_H is known, the quantity $\mathbb{E}\left[\int_{\mathbb{R}} (\log |x|) \mu_H(\mathrm{d}x)\right]$ can be computed.

Summary

▶ TAP free energy is accurate for \mathbb{Z}_2 synchronization.

- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- ▶ It is interesting to study and apply variational inference beyond mean field.

< ロ > < 同 > < 回 > < 回 > < 回 > <

- ▶ TAP free energy is accurate for \mathbb{Z}_2 synchronization.
- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- ▶ It is interesting to study and apply variational inference beyond mean field.

イロト 不得 とうせい きゅうしょう

- ▶ TAP free energy is accurate for \mathbb{Z}_2 synchronization.
- Can be generalized to topic modeling, low rank matrix estimation, compressed sensing, etc...
- It is interesting to study and apply variational inference beyond mean field.