Local convexity of the TAP free energy and AMP convergence for Z2-synchronization

Song Mei

UC Berkeley

September 15, 2021

Joint work with Michael Celentano and Zhou Fan

Motivation

- Variational Bayesian inference:
  - Optimizing variational free energy (VB, Bethe, TAP).
  - Iterative algorithms (BP, EP, MP, AMP).
- These methods have been established for more than 20 years. Have been written into softwares and works well in practice.
- Theoretical challenges:
  - Minimizer of free energy approximates Bayesian posterior mean?
  - Landscape of the free energy?
  - Convergence of iterative algorithms?

Today

- High dimensional statistical model:
  $\mathbb{Z}_2$ synchronization in the weak signal regime.
- Landscape of the TAP free energy.
- Convergence of iterative algorithms (NGD, AMP).

$\mathbb{Z}_2$ 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: for $1 \le i < j \le n$

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

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

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

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

$\mathbb{Z}_2$ 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: for $1 \le i < j \le n$

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

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

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

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

$\mathbb{Z}_2$ 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: for $1 \le i < j \le n$

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

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

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

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

$\mathbb{Z}_2$ synchronization

- Signal:

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

- Observation: for $1 \le i < j \le n$

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

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

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

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

$\mathbb{Z}_2$ 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: for $1 \le i < j \le n$

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

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

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

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

$\mathbb{Z}_2$ 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: for $1 \le i < j \le n$

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

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

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

- 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

- Settings:

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

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

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

- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient, e.g., spectral estimator (BBAP phase transition).
- Optimal estimator:

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

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

- Settings:

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

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

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

- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient, e.g., spectral estimator (BBAP phase transition).
- Optimal estimator:

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

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

- Settings:

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

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

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

- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient, e.g., spectral estimator (BBAP phase transition).
- Optimal estimator:

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

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

- Settings:

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

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

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

- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient, e.g., spectral estimator (BBAP phase transition).
- Optimal estimator:

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

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

- Settings:
$$x \sim \mathrm{Unif}(\mathbb{Z}_2^n), \qquad \boldsymbol{Y} = (\lambda/n)\boldsymbol{x}\boldsymbol{x}^\mathsf{T} + \boldsymbol{W}.$$
- Estimate $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\mathsf{T}$ with loss:
$$\ell(\boldsymbol{X}, \widehat{\boldsymbol{X}}) = (1/n^2)\|\boldsymbol{X} - \widehat{\boldsymbol{X}}\|_F^2.$$
- For $\lambda < 1$, impossible.
- For $\lambda > 1$, possible and efficient, e.g., spectral estimator (BBAP phase transition).
- Optimal estimator:
$$\widehat{\boldsymbol{X}}_{\text{Bayes}} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\mathsf{T}|\boldsymbol{Y}].$$

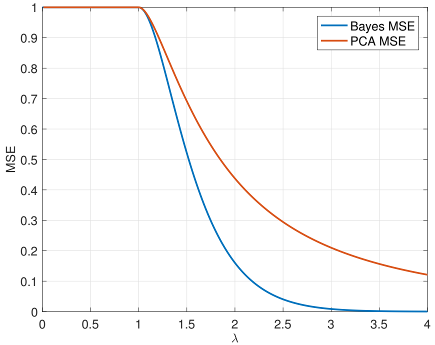
Bayes estimation in $\mathbb{Z}_2$ synchronization

- 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:

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

- The posterior distribution: (relationship with SK measure)

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

- Two viewpoints in variational inference:
  - Variational free energies (TAP free energy).
  - Iterative algorithms (NGD, AMP).

Compute the Bayesian estimator

- The Bayesian estimator:

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

- The posterior distribution: (relationship with SK measure)

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

- Two viewpoints in variational inference:
  - Variational free energies (TAP free energy).
  - Iterative algorithms (NGD, AMP).

Compute the Bayesian estimator

- The Bayesian estimator:

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

- The posterior distribution: (relationship with SK measure)

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

- Two viewpoints in variational inference:
  - Variational free energies (TAP free energy).
  - Iterative algorithms (NGD, AMP).

Compute the Bayesian estimator

- The Bayesian estimator:

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

- The posterior distribution: (relationship with SK measure)

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

- Two viewpoints in variational inference:
  - Variational free energies (TAP free energy).
  - Iterative algorithms (NGD, AMP).

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: (relationship with SK measure)

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

- Two viewpoints in variational inference:
  - Variational free energies (TAP free energy).
  - Iterative algorithms (NGD, AMP).

Variational free energies

- Find a function $\mathcal{F} : [-1, 1]^n \to \mathbb{R}$, such that

$$\hat{\boldsymbol{m}} \equiv \arg\min_{\boldsymbol{m}} \mathcal{F}(\boldsymbol{m}) \longleftrightarrow \widehat{\boldsymbol{X}}_{\text{Bayes}}.$$

- The mean field variational Bayes (the KL minimization)

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

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

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

Variational free energies

- Find a function $\mathcal{F} : [-1,1]^n \to \mathbb{R}$, such that

$$\hat{\boldsymbol{m}} \equiv \arg\min_{\boldsymbol{m}} \mathcal{F}(\boldsymbol{m}) \longleftrightarrow \widehat{\boldsymbol{X}}_{\text{Bayes}}.$$

- The mean field variational Bayes (the KL minimization)

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

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

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

Variational free energies

- Find a function $\mathcal{F} : [-1, 1]^n \to \mathbb{R}$, such that

$$\hat{\boldsymbol{m}} \equiv \arg\min_{\boldsymbol{m}} \mathcal{F}(\boldsymbol{m}) \longleftrightarrow \widehat{\boldsymbol{X}}_{\text{Bayes}}.$$

- The mean field variational Bayes (the KL minimization)

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

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

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

Variational free energies

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

- This is non-convex.

Q: Nice landscape? Global minimizer corresponds to $\widehat{X}_{\mathrm{Bayes}}$?

Variational free energies

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

- This is non-convex.

Q: Nice landscape? Global minimizer corresponds to $\widehat{\boldsymbol{X}}_{\mathrm{Bayes}}$?

Variational free energies

- The TAP free energy (Thouless, Anderson, and Palmer (1977))

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

- This is non-convex.

Q: Nice landscape? Global minimizer corresponds to $\widehat{\boldsymbol{X}}_{\rm Bayes}$?

Approximate message passing

- AMP iteration [Bolthausen, 2012] [Donoho, Maleki, Montanari, 2009]

$$\boldsymbol{m}^{k+1} = \mathtt{tanh}(\lambda \boldsymbol{Y} \boldsymbol{m}^{k} - \lambda^2 (1 - Q(\boldsymbol{m}^{k})) \boldsymbol{m}^{k-1}).$$

- Fixed point is a stationery point of the TAP free energy

$$\boldsymbol{m}_\star = \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) \qquad \longleftrightarrow \qquad \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = \boldsymbol{0}.$$

- Not a descent algorithm on $\mathcal{F}_{\mathrm{TAP}}$.

Approximate message passing

- AMP iteration [Bolthausen, 2012] [Donoho, Maleki, Montanari, 2009]
$$\boldsymbol{m}^{k+1} = \tanh(\lambda \boldsymbol{Y} \boldsymbol{m}^{k} - \lambda^2 (1 - Q(\boldsymbol{m}^{k})) \boldsymbol{m}^{k-1}).$$
- Fixed point is a stationery point of the TAP free energy
$$\boldsymbol{m}_\star = \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) \qquad \longleftrightarrow \qquad \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = \boldsymbol{0}.$$
- Not a descent algorithm on $\mathcal{F}_{\mathrm{TAP}}$.

Approximate message passing

- AMP iteration [Bolthausen, 2012] [Donoho, Maleki, Montanari, 2009]

$$\boldsymbol{m}^{k+1} = \tanh(\lambda \boldsymbol{Y} \boldsymbol{m}^{k} - \lambda^2 (1 - Q(\boldsymbol{m}^{k})) \boldsymbol{m}^{k-1}).$$

- Fixed point is a stationery point of the TAP free energy

$$\boldsymbol{m}_\star = \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) \qquad \longleftrightarrow \qquad \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = \mathbf{0}.$$

- Not a descent algorithm on $\mathcal{F}_{\mathrm{TAP}}$.

Approximate message passing

- Convergence starting from spectral initialization [Montanari, Venkataramanan, 2017]

$$\lim_{n\to\infty} \|\boldsymbol{m}^k\boldsymbol{m}^k - \widehat{\boldsymbol{X}}_{\text{Bayes}}\|_F^2/n^2 = \varepsilon(k) \to 0, \qquad \text{as } k \to \infty.$$

- Analysis based on state evolution. Best known is $k = o(\log(n)/\log\log n)$ [Rush, Venkataramanan, 2016].

Q: Convergence of AMP for fixed n as $k \to \infty$?

Approximate message passing

- Convergence starting from spectral initialization [Montanari, Venkataramanan, 2017]

$$\lim_{n\to\infty} \|\boldsymbol{m}^k \boldsymbol{m}^k - \widehat{\boldsymbol{X}}_{\rm Bayes}\|_F^2 / n^2 = \varepsilon(k) \to 0, \qquad \text{as } k \to \infty.$$

- Analysis based on state evolution. Best known is $k = o(\log(n)/\log\log n)$ [Rush, Venkataramanan, 2016].

Q: Convergence of AMP for fixed n as $k \to \infty$?

Approximate message passing

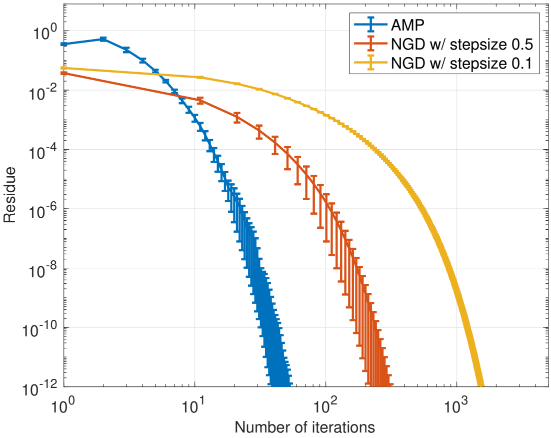
- Convergence starting from spectral initialization [Montanari, Venkataramanan, 2017]

$$\lim_{n\to\infty} \|\boldsymbol{m}^k \boldsymbol{m}^k - \widehat{\boldsymbol{X}}_{\rm Bayes}\|_F^2 / n^2 = \varepsilon(k) \to 0, \qquad \text{as } k \to \infty.$$

- Analysis based on state evolution. Best known is $k = o(\log(n)/\log\log n)$ [Rush, Venkataramanan, 2016].

Q: Convergence of AMP for fixed n as $k \to \infty$?

Numerical simulations



Main results

Landscape of the TAP free energy

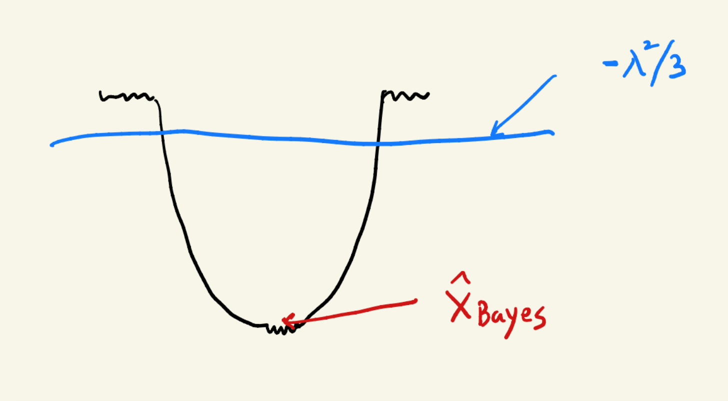
Theorem (Fan, Mei, Montanari, 2018)

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*

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

Landscape of the TAP free energy

All the critical points below a threshold are close to the Bayesian estimator. [Fan, Mei, Montanari, 2018]



Landscape of the TAP free energy

Theorem (Celentano, Fan, Mei, 2021)

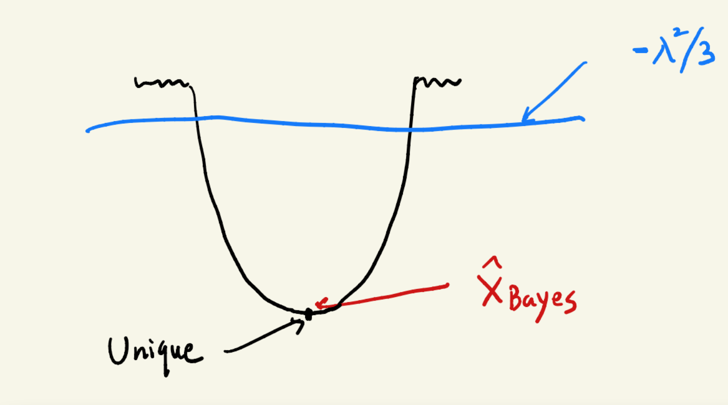
Fix any $\lambda > 1$ (1 is the IT thresholds). For any sufficiently small $\iota > 0$, with probability approaching 1 as $n \to \infty$, there exists a unique critical point and unique local minimizer $\boldsymbol{m}_\star$ of $\mathcal{F}_{\rm TAP}(\boldsymbol{m})$ (up to $\pm$ sign) such that

$$\frac{1}{n^2}\|\boldsymbol{m}_\star \boldsymbol{m}_\star^{\mathsf{T}} - \widehat{\boldsymbol{X}}_{\rm Bayes}\|_{\rm F}^2 < \iota. \tag{2}$$

Moreover, $\mathcal{F}_{\rm TAP}$ is strongly convex over $(-1,1)^n \cap {\sf B}_{\sqrt{\varepsilon n}}(\boldsymbol{m}_\star)$ for some n independent constant ε.

Landscape of the TAP free energy

There is a local strongly convex region near the Bayes estimator, which contains a unique local minimizer. [Celentano, Fan, Mei, 2021]



Iterative algorithms

- Approximate message passing (AMP).

$$\begin{aligned} \boldsymbol{m}^k &= \tanh(\boldsymbol{h}^k), \\ \boldsymbol{h}^{k+1} &= \Big(\lambda \boldsymbol{Y} \boldsymbol{m}^k - \lambda^2 [1 - Q(\boldsymbol{m}^k)] \boldsymbol{m}^{k-1}\Big). \end{aligned}$$

- Natural gradient descent (NGD) or Bregman gradient descent.

$$\begin{aligned} \boldsymbol{m}^k &= \tanh(\boldsymbol{h}^k) \\ \boldsymbol{h}^{k+1} &= \boldsymbol{h}^k - \eta n \cdot \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^k) \\ &= (1-\eta)\boldsymbol{h}^k + \eta \Big(\lambda \boldsymbol{Y} \boldsymbol{m}^k - \lambda^2 [1 - Q(\boldsymbol{m}^k)] \boldsymbol{m}^k\Big). \end{aligned}$$

A hybrid algorithm converges

A hybrid algorithm

Spectral initialization. Run AMP for T steps, and then continue to run NGD with stepsize η.

Theorem (Celentano, Fan, Mei, 2021)

Fix any $\lambda > 1$*. There exist* λ*-dependent constants* $C, \mu, \eta_0 > 0$ *and* $T \geq 1$ *such that with probability approaching 1 as* $n \to \infty$*, we have*

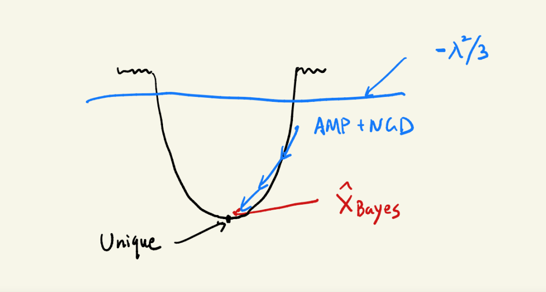
$$\mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}^{T+k}) - \mathcal{F}_{\mathrm{TAP}}(\pm\boldsymbol{m}_\star) < C(1-\mu\eta)^k,$$

$$\|\boldsymbol{m}^{T+k} - (\pm\boldsymbol{m}_\star)\|_2 < C(1-\mu\eta)^k\sqrt{n}.$$

In particular, $\lim_{k\to\infty} \boldsymbol{m}^{T+k} \in \{+\boldsymbol{m}_\star, -\boldsymbol{m}_\star\}$.

A hybrid algorithm converges

A hybrid algorithm converges. [Celentano, Fan, Mei, 2021]



- AMP drives the iterates into the local region.
- Local convergence of NGD is due to the relative smoothness and relative strong convexity with respect to the entropy function.

How about AMP or NGD individually?

Theorem (Celentano, Fan, Mei, 2021)

There exists λ_0, such that for any $\lambda > \lambda_0$, with probability converging to 1 as $n \to \infty$, either spectral initialized AMP or spectral initialized NGD converges to the global minimizer $\boldsymbol{m}_\star$.

The proof for the case $\lambda \geq \lambda_0$ is much more easier than $\lambda > 1$.

How about AMP or NGD individually?

Theorem (Celentano, Fan, Mei, 2021)

There exists λ_0, such that for any $\lambda > \lambda_0$, with probability converging to 1 as $n \to \infty$, either spectral initialized AMP or spectral initialized NGD converges to the global minimizer $\boldsymbol{m}_\star$.

The proof for the case $\lambda \geq \lambda_0$ is much more easier than $\lambda > 1$.

What do we know about AMP when $\lambda > 1$?

Theorem (Celentano, Fan, Mei, 2021)

For any $\lambda > 1$, with probability converging to 1 as $n \to \infty$, $\boldsymbol{m}_\star$ is an asymptotically stable fixed point of AMP. That means, there exists a neighborhood $\mathsf{B}(\boldsymbol{m}_\star, \delta)$ (the size of δ may depend on n), such that if AMP is initialized in the neighborhood, it converges to $\boldsymbol{m}_\star$.

Open problem

Show that spectral initialized AMP converges as long as $\lambda > 1$.

What do we know about AMP when $\lambda > 1$?

Theorem (Celentano, Fan, Mei, 2021)

For any $\lambda > 1$, with probability converging to 1 as $n \to \infty$, $\boldsymbol{m}_\star$ is an asymptotically stable fixed point of AMP. That means, there exists a neighborhood $\mathsf{B}(\boldsymbol{m}_\star, \delta)$ (the size of δ may depend on n), such that if AMP is initialized in the neighborhood, it converges to $\boldsymbol{m}_\star$.

Open problem

Show that spectral initialized AMP converges as long as $\lambda > 1$.

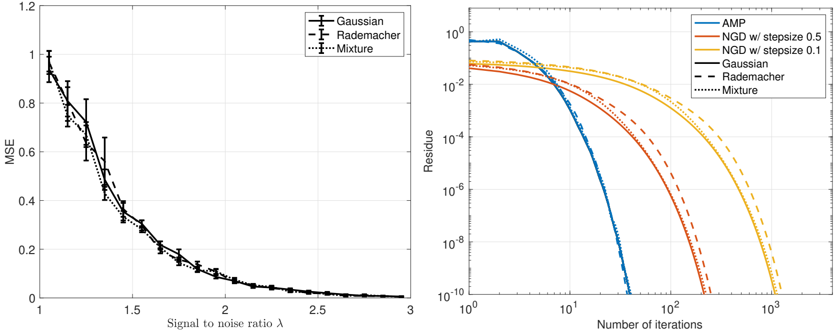
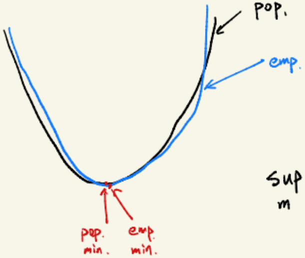


Figure: Universality with respect to the noise distribution.

Technical challenge

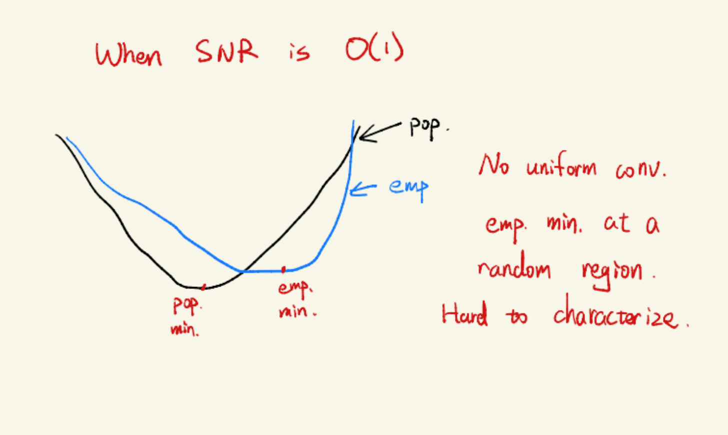
When SNR is super large ($\lambda = O(\log n)$)



$$\sup_m \left| \nabla^k \mathrm{emp}(m) - \nabla^k \mathrm{pop}(m) \right| \to 0$$

$k = 0, 1, 2.$

Technical challenge



Technical tool 1: Kac-Rice formula

- $f : \mathbb{R}^d \to \mathbb{R}$ a "sufficiently regular" random morse function.

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

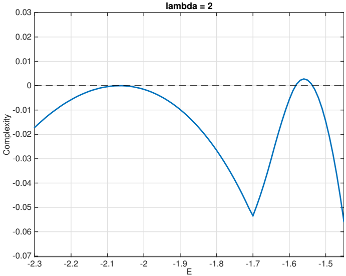
Kac-Rice formula:

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

- Technicality: determinant of Hessian $\det \nabla^2 f(\boldsymbol{m})$.
- Result: [Fan, Mei, Montanari, 2018]

$$\frac{1}{n} \log \mathbb{E}[\mathrm{Crit}_n(U)] \le \sup_{(q,\varphi,e) \in U} S_\star(q, \varphi, e) + o(1).$$

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



- All local min below some threshold are close to the global min.
- There could potentially be many local min near global.

Technical tool 2: Conditional Gaussian comparison

- Analyze the following Gaussian process ...

$$\min_{\boldsymbol{m} \in \mathsf{B}(\boldsymbol{m}_\star, \varepsilon\sqrt{n})} \min_{\boldsymbol{u} \in \mathbb{S}^{n-1}} \langle \boldsymbol{u}, \nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \boldsymbol{u} \rangle$$

  ... conditional on $\nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = 0$ (a linear constraint on Gaussian noise matrix).
- $[\nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) | \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) = 0]$ is a finite rank spiked GOE matrix.

Technical tool 2: Conditional Gaussian comparison

- Gaussian comparison lower bound (with many non-trivial idea) [Celentano, Fan, Mei, 2021]

$$\text{rescaled Hessian } \geq \inf_{u,p} \sup_{\alpha,\kappa,\gamma} H_\lambda(p,u;\alpha,\kappa,\gamma) + o(1) > 0.$$

- So, $\mathcal{F}_{\mathrm{TAP}}$ is locally strongly convex near $\boldsymbol{m}_\star$.

Some technical challenges showing this lower bound:

- Need control on the empirical distribution of coordinates of $\boldsymbol{m}_\star$. Use Slepian + Kac-Rice to give it a tight control.
- Need to analyze the variational formula $\inf \sup H_\lambda$. First handle the bulk, then show the spikes do not affect the bulk.

Technical tool 2: Conditional Gaussian comparison

- Gaussian comparison lower bound (with many non-trivial idea) [Celentano, Fan, Mei, 2021]

$$\text{rescaled Hessian } \geq \inf_{u,p} \sup_{\alpha,\kappa,\gamma} H_\lambda(p, u; \alpha, \kappa, \gamma) + o(1) > 0.$$

- So, $\mathcal{F}_{\text{TAP}}$ is locally strongly convex near $\boldsymbol{m}_\star$.

Some technical challenges showing this lower bound:

- Need control on the empirical distribution of coordinates of $\boldsymbol{m}_\star$. Use Slepian + Kac-Rice to give it a tight control.
- Need to analyze the variational formula $\inf \sup H_\lambda$.
  First handle the bulk, then show the spikes do not affect the bulk.

Technical tool 3: Union bound using Kac-Rice

Need to translate

$$\sup_{\boldsymbol{m}_\star \in \text{some region}} \mathbb{P}\Big(\min_{\boldsymbol{m} \in \mathsf{B}(\boldsymbol{m}_\star, \varepsilon\sqrt{n})} \lambda_{\min}(\nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})) < -\varepsilon \\ \Big| \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = 0 \Big) < e^{-n\delta}$$

to

$$\mathbb{P}\Big(\forall \boldsymbol{m}_\star \in \text{some region}, \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = 0, \\ \min_{\boldsymbol{m} \in \mathsf{B}(\boldsymbol{m}_\star, \varepsilon\sqrt{n})} \lambda_{\min}(\nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})) < -\varepsilon \Big) < e^{-n\delta}$$

Method: using Kac-Rice to upper bound

$$\mathbb{E}[\#\{\boldsymbol{m}_\star \in \text{some region}, \nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = 0, \\ \min_{\boldsymbol{m} \in \mathsf{B}(\boldsymbol{m}_\star, \varepsilon\sqrt{n})} \lambda_{\min}(\nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m})) < -\varepsilon\}].$$

Convergence of NGD

- Local geometry: relatively smooth and relatively strongly convex

$$c \cdot \nabla^2 \mathrm{Ent}(\boldsymbol{m}) \preceq \nabla^2 \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}) \preceq C \cdot \nabla^2 \mathrm{Ent}(\boldsymbol{m}).$$

- Local geometry implies convergence of NGD.

Convergence of AMP

- AMP iteration
$$\begin{bmatrix} \boldsymbol{m}^{k+1} \\ \boldsymbol{m}^{k} \end{bmatrix} = \mathrm{AMP}\left(\begin{bmatrix} \boldsymbol{m}^{k} \\ \boldsymbol{m}^{k-1} \end{bmatrix} \right)$$
- Analyze the linearized AMP operator $\nabla \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star)$, conditional on $\nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star, \boldsymbol{m}_\star) = \boldsymbol{0}$.

Convergence of AMP

- Analyze the linearized AMP operator $\nabla \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star)$, conditional on $\nabla \mathcal{F}_{\mathrm{TAP}}(\boldsymbol{m}_\star) = \mathbf{0}$.
- When $\lambda > 1$, w.h.p

$$\sup_{i \in [2n]} \left| \lambda_i(\nabla \mathrm{AMP}(\boldsymbol{m}_\star, \boldsymbol{m}_\star)) \right| < 1 - \varepsilon,$$

  implies asymptotic stability at $\boldsymbol{m}_\star$. (Quite non-trivial proof)
- When $\lambda \geq \lambda_0$, w.h.p

$$\sup_{\boldsymbol{m} \in \mathsf{B}(\boldsymbol{m}_\star, \varepsilon)} \|\nabla \mathrm{AMP}'\|_{\mathrm{op}} < 1 - \varepsilon,$$

  implies global convergence.

Summary

- TAP has no spurious local min below a threshold ($\lambda \geq \lambda_0$)
- TAP is locally strongly convex near Bayes estimator ($\lambda > 1$).
- A hybrid spec. + AMP + NGD algorithm converges ($\lambda > 1$).
- AMP, NGD converges ($\lambda \geq \lambda_0$).
- AMP is asymptotically stable at $\boldsymbol{m}_\star$ ($\lambda > 1$).
- The proof strategy can be extended to other problems.