# Mean Field Asymptotics in High-dimensional Statistics

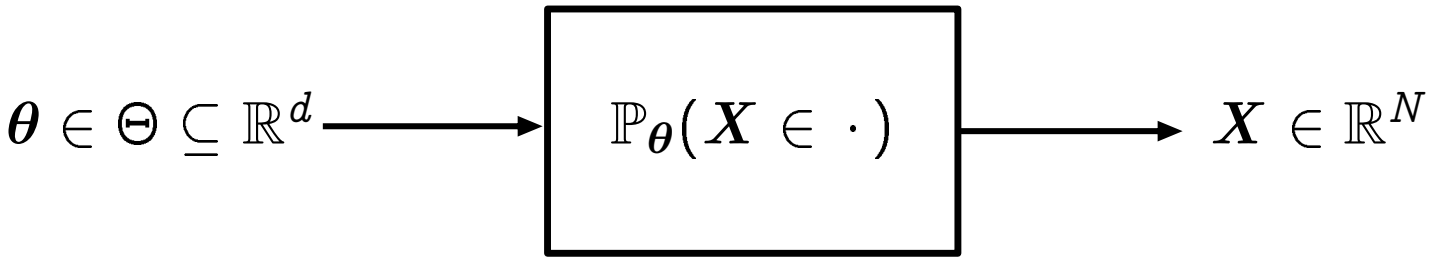
Andrea Montanari

Stanford University

August 8, 2018

# The big picture

# Statistical estimation

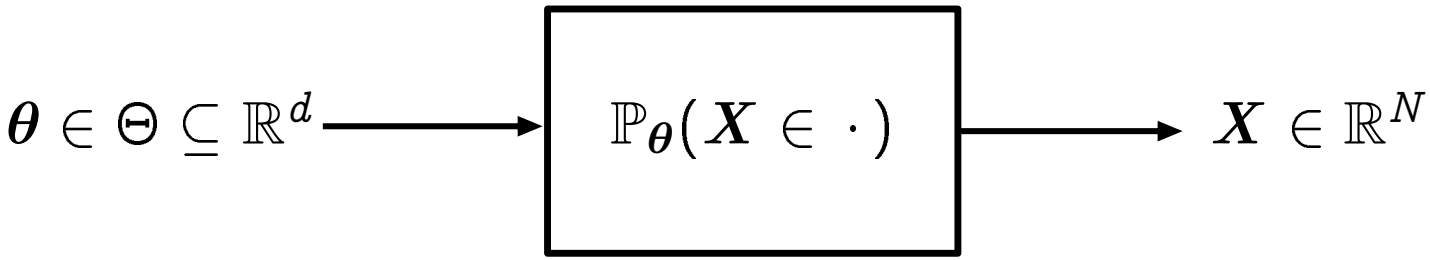



**Estimator:** $\hat{\boldsymbol{\theta}} : \mathbb{R}^N \to \mathbb{R}^d \qquad \boldsymbol{X} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{X})$

Amount of data : $N$

Model complexity : $d$

# Statistical estimation




**Estimator:** $\hat{\boldsymbol{\theta}} : \mathbb{R}^N \to \mathbb{R}^d \qquad \boldsymbol{X} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{X})$

Amount of data : $N$

Model complexity : $d$

# Regimes: $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^N$

**Classical – Data rich**

$$N \gg d \quad \Rightarrow \quad \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0\| \xrightarrow{\text{p}} 0$$

[Fisher, Le Cam, Huber, ... 1920-...]

**Modern – Data poor – High-dimensional**

$$d \gg N \gg \text{eff.dim}(\Theta) \quad \Rightarrow \quad \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0\| \xrightarrow{\text{p}} 0$$

[Donoho, Candés, Tao, ... 2006-...]

**Noisy high-dimensional**

$$N \asymp d \asymp \text{eff.dim}(\Theta) \quad \Rightarrow \quad (\hat{\theta}(\boldsymbol{X}), \theta_0) \xRightarrow{\text{d}} \text{Non-trivial limit}$$

# Regimes: $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, $\boldsymbol{X} \in \mathbb{R}^N$

**Classical – Data rich**

$$N \gg d \quad \Rightarrow \quad \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0\| \xrightarrow{\text{p}} 0$$

[Fisher, Le Cam, Huber, ... 1920-...]

**Modern – Data poor – High-dimensional**

$$d \gg N \gg \text{eff.dim}(\Theta) \quad \Rightarrow \quad \|\hat{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{\theta}_0\| \xrightarrow{\text{p}} 0$$

[Donoho, Candés, Tao, ... 2006-...]

## Noisy high-dimensional

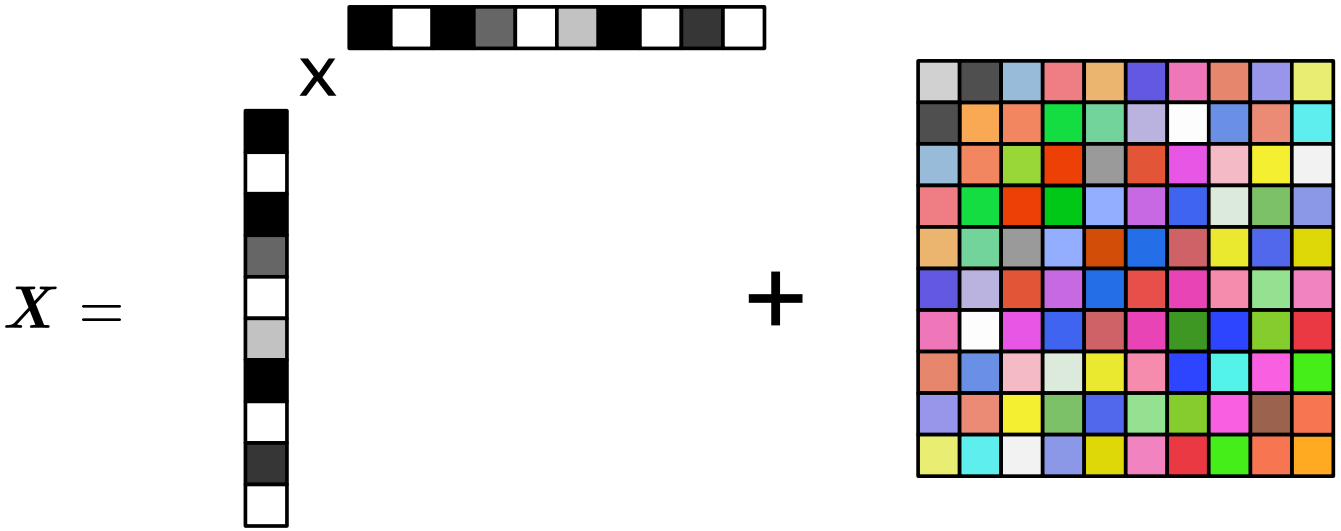
$$N \asymp d \asymp \text{eff.dim}(\Theta) \quad \Rightarrow \quad (\hat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}_0) \xRightarrow{\text{d}} \text{Non-trivial limit}$$

# Outline


1. A simple model
2. Bayes optimal estimation
3. Efficient algorithms
4. Generic/robust algorithms
5. Conclusion

# A simple model

# Rank-one matrix estimation




$$X = \frac{\lambda}{n}\theta_0\theta_0^{\mathsf{T}} + W$$

# Rank-one matrix estimation

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$$

- **Parameter:** $\boldsymbol{\theta}_0 \in \mathbb{R}^n$, $\|\boldsymbol{\theta}_0\|_2^2 = n + o(n)$
- **Data:** $\boldsymbol{X} \in \mathbb{R}^{n\times n}$
- **Noise:** $\boldsymbol{W} \sim \mathsf{GOE}(n)$

$$(\boldsymbol{W} = \boldsymbol{W}^{\mathsf{T}},\ W_{ii} \sim \mathsf{N}(0, 2/n),\ W_{ij} \sim \mathsf{N}(0, 1/n))$$

# Structured parameter

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$$

$$\widehat{p}_{\boldsymbol{\theta}} \equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}\,,$$

$$\Theta(\varepsilon) \equiv \left\{\boldsymbol{\theta} \in \mathbb{R}^n : \quad \widehat{p}_{\boldsymbol{\theta}} = p_\varepsilon \equiv \varepsilon\delta_{a_+} + (1-\varepsilon)\delta_{-a_-}\right\},$$

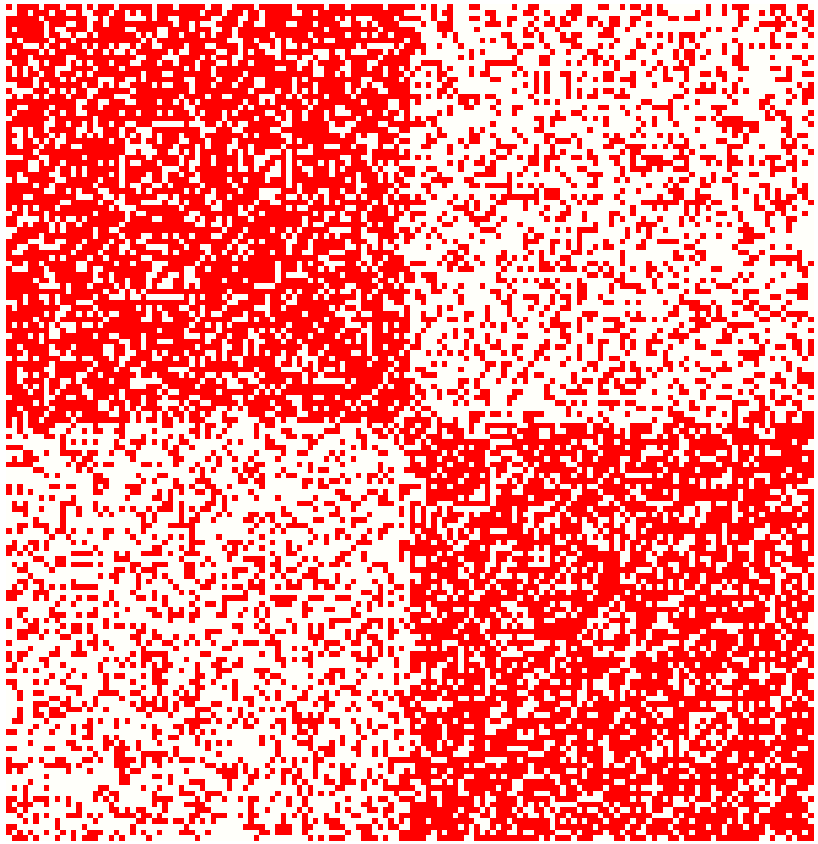
$$\int x\, p_\varepsilon(\mathrm{d}x) = 0 \qquad \int x^2\, p_\varepsilon(\mathrm{d}x) = 1\,.$$

# $\varepsilon = 0.5$: '$\mathbb{Z}_2$-synchronization'

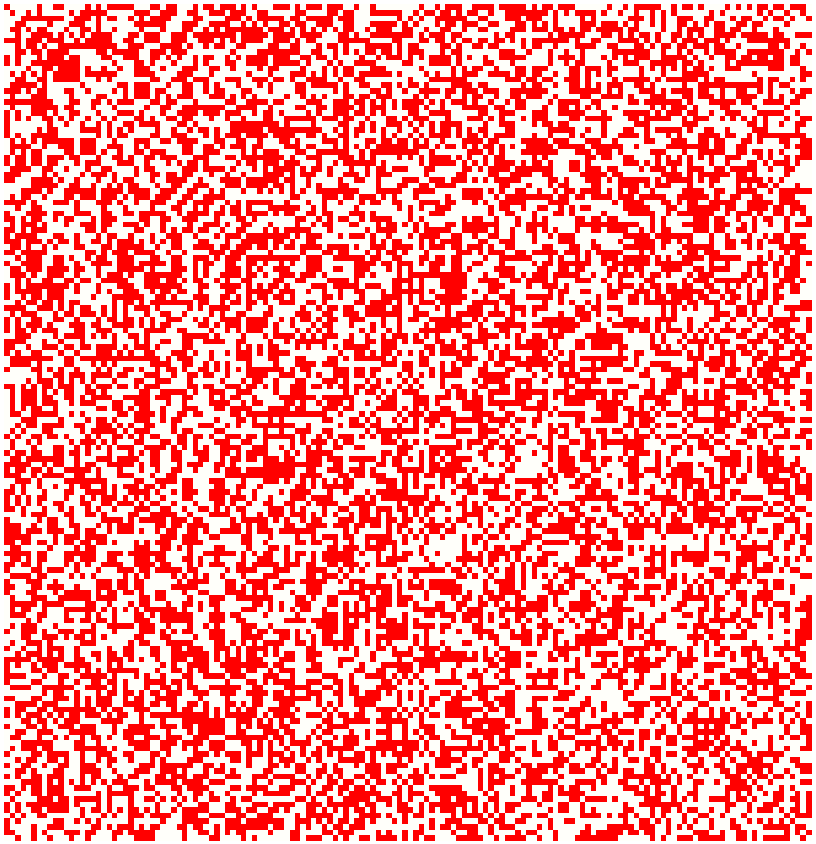
$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$$

$$\boldsymbol{\theta}_0 \in \{+1, -1\}^n\,; \quad \sum_{i=1}^{n} \theta_{0,i} = 0\,.$$

# Example



# The same example



# Goal

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$$

**Estimator:**

$$\hat{\boldsymbol{\theta}} : \boldsymbol{X} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{X}) \in \mathbb{R}^n$$

**Accuracy:**

$$\text{Minimax:} \qquad \min_{\boldsymbol{\theta}_0 \in \Theta(\varepsilon)} \mathbb{E}_{\boldsymbol{X}} \left\{ \frac{|\langle \hat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\},$$

$$\text{Bayesian:} \qquad \mathbb{E}_{\boldsymbol{\theta}_0 \sim \text{Unif}(\Theta(\varepsilon))} \mathbb{E}_{\boldsymbol{X}} \left\{ \frac{|\langle \hat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\}$$

Minimax $\approx$ Bayesian (by symmetry)

# Goal

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}$$

**Estimator:**

$$\hat{\boldsymbol{\theta}} : \boldsymbol{X} \mapsto \hat{\boldsymbol{\theta}}(\boldsymbol{X}) \in \mathbb{R}^n$$

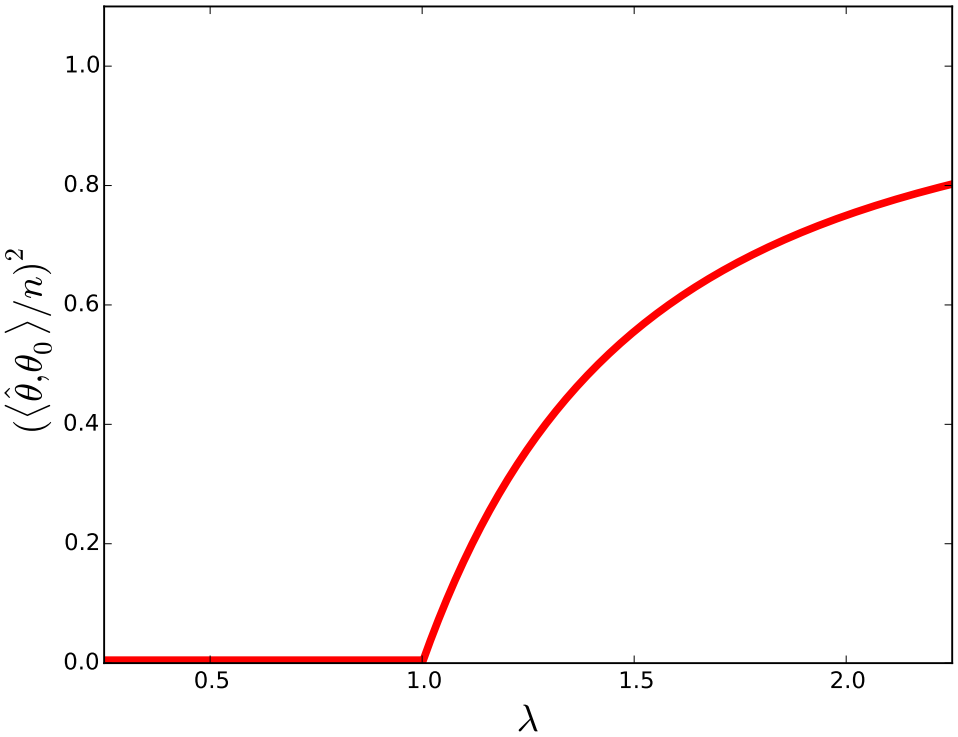
**Accuracy:**

$$\text{Minimax:} \qquad \min_{\boldsymbol{\theta}_0 \in \Theta(\varepsilon)} \mathbb{E}_{\boldsymbol{X}} \left\{ \frac{|\langle \hat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\},$$

$$\text{Bayesian:} \qquad \mathbb{E}_{\boldsymbol{\theta}_0 \sim \text{Unif}(\Theta(\varepsilon))} \mathbb{E}_{\boldsymbol{X}} \left\{ \frac{|\langle \hat{\boldsymbol{\theta}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\}$$

$$\text{Minimax} \approx \text{Bayesian} \qquad \text{(by symmetry)}$$

# Example: Principal Component Analysis




$$\hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}) = \sqrt{n}\, \boldsymbol{v}_1(\boldsymbol{X}) \qquad \text{(principal eigenvector)}$$

# Example: Principal Component Analysis

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}) = \boldsymbol{v}_1(\boldsymbol{X})\,.$$

**Theorem** (Baik, Ben Arous, Péché, 2005)

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \sqrt{\left(1 - \frac{1}{\lambda^2}\right)_+}$$

Can we do better?

# Example: Principal Component Analysis

$$\boldsymbol{X} = \frac{\lambda}{n}\boldsymbol{\theta}_0\boldsymbol{\theta}_0^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}) = \boldsymbol{v}_1(\boldsymbol{X})\,.$$

**Theorem** (Baik, Ben Arous, Péché, 2005)

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{PCA}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \sqrt{\left(1 - \frac{1}{\lambda^2}\right)_+}$$

**Can we do better?**

# Agenda

- Bayes optimal estimation
- Efficient algorithms
- Generic/robust algorithms

# Bayes optimal estimation

# Bayesian estimation ($\approx$ minimax)

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_\varepsilon(\mathrm{d}\theta_i)\,.$$

**Estimator**

$$\hat{\boldsymbol{\theta}}^{\text{Bayes}}(\boldsymbol{X}) = \mathbb{E}\{\boldsymbol{\theta}|\boldsymbol{X}\}\,.$$

[For $\varepsilon = 0.5$ need to break symmetry...]

# Bayesian estimation: Analysis

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_{\varepsilon}(\mathrm{d}\theta_i)\,.$$

- Generalized spin-glass model
- Replica symmetric [Korada, Macris 2007; M. 2008]

$$\boldsymbol{X} \sim p_{\text{Bayes}}(\,\cdot\,)\,, \qquad (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X}) \otimes p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$$

$$\frac{1}{n}\langle\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\rangle \xrightarrow{\text{p}} q_* \qquad \text{(nonrandom)}$$

Intuition: for $\boldsymbol{\theta} \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$, $\theta_i, \theta_j$, $i \neq j$ approx independent

# Bayesian estimation: Analysis

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_{\varepsilon}(\mathrm{d}\theta_i)\,.$$

- **Generalized spin-glass model**
- **Replica symmetric** [Korada, Macris 2007; M. 2008]

$$\boldsymbol{X} \sim p_{\text{Bayes}}(\,\cdot\,)\,, \qquad (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X}) \otimes p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$$

$$\frac{1}{n}\langle\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\rangle \xrightarrow{\text{p}} q_* \qquad \text{(nonrandom)}$$

Intuition: for $\boldsymbol{\theta} \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$, $\theta_i, \theta_j$, $i \neq j$ approx independent

# Bayesian estimation: Analysis

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_{\varepsilon}(\mathrm{d}\theta_i)\,.$$

- **Generalized spin-glass model**
- **Replica symmetric** [Korada, Macris 2007; M. 2008]

$$\boldsymbol{X} \sim p_{\text{Bayes}}(\,\cdot\,)\,, \qquad (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X}) \otimes p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$$

$$\frac{1}{n}\langle\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\rangle \xrightarrow{\text{p}} q_* \qquad \text{(nonrandom)}$$

Intuition: for $\boldsymbol{\theta} \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$, $\theta_i, \theta_j$, $i \neq j$ approx independent

# Bayesian estimation: Analysis

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_{\varepsilon}(\mathrm{d}\theta_i)\,.$$

- **Generalized spin-glass model**
- **Replica symmetric** [Korada, Macris 2007; M. 2008]

$$\boldsymbol{X} \sim p_{\text{Bayes}}(\,\cdot\,)\,, \qquad (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X}) \otimes p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$$

$$\frac{1}{n}\langle\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\rangle \xrightarrow{\text{p}} q_* \qquad \text{(nonrandom)}$$

Intuition: for $\boldsymbol{\theta} \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$, $\theta_i, \theta_j$, $i \neq j$ approx independent

# Bayesian estimation: Analysis

**Posterior**

$$p_{\text{Bayes}}(\mathrm{d}\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \exp\left\{\frac{\lambda}{2}\langle\boldsymbol{\theta}, \boldsymbol{X}\boldsymbol{\theta}\rangle - \frac{\lambda^2}{4n}\|\boldsymbol{\theta}\|_2^4\right\} \prod_{i=1}^{n} p_{\varepsilon}(\mathrm{d}\theta_i)\,.$$

- **Generalized spin-glass model**
- **Replica symmetric** [Korada, Macris 2007; M. 2008]

$$\boldsymbol{X} \sim p_{\text{Bayes}}(\,\cdot\,)\,, \qquad (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2) \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X}) \otimes p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$$

$$\frac{1}{n}\langle\boldsymbol{\theta}^1, \boldsymbol{\theta}^2\rangle \xrightarrow{\text{p}} q_* \qquad \text{(nonrandom)}$$

Intuition: for $\boldsymbol{\theta} \sim p_{\text{Bayes}}(\,\cdot\,|\boldsymbol{X})$, $\theta_i, \theta_j$, $i \neq j$ approx independent

# Bayesian estimation: Analysis

**Theorem** (Lelarge, Miolane, 2017; Barbier et al. 2016)

*Consider the rank-one model with $\theta_{0,i} \sim p$ (+ conditions). Let* $\mathrm{I}(\gamma) = \mathbb{E} \log \frac{dp_{Y|X_0}}{dp_Y}(Y, X_0)$ *for* $(X_0, G) \sim p \otimes \mathsf{N}(0,1)$, $Y = \sqrt{\gamma} X_0 + G$, *and define*

$$\Psi(\gamma, \lambda) \equiv \frac{\lambda^2}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathrm{I}(\gamma)\,,$$
$$\gamma_{\mathrm{Bayes}}(\lambda) \equiv \arg\max_{\gamma \geq 0} \Psi(\gamma, \lambda)\,.$$

*Then*

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \frac{\sqrt{\gamma_{\mathrm{Bayes}}(\lambda)}}{\lambda}\,.$$

[see also: Deshpande, M., 2014; Deshpande, Abbe, M., 2016; Barbier, Dia, Macris, Krzakala, Lesieur, Zdeborová, 2016; Krzakala, Xu, Zdeborová, 2016; El Alaoui, Jordan, 2018; ...]

# Bayesian estimation: Analysis

**Theorem** (Lelarge, Miolane, 2017; Barbier et al. 2016)

*Consider the rank-one model with $\theta_{0,i} \sim p$ (+ conditions). Let* $\mathrm{I}(\gamma) = \mathbb{E} \log \frac{dp_{Y|X_0}}{dp_Y}(Y, X_0)$ *for* $(X_0, G) \sim p \otimes \mathsf{N}(0,1)$, $Y = \sqrt{\gamma} X_0 + G$, *and define*

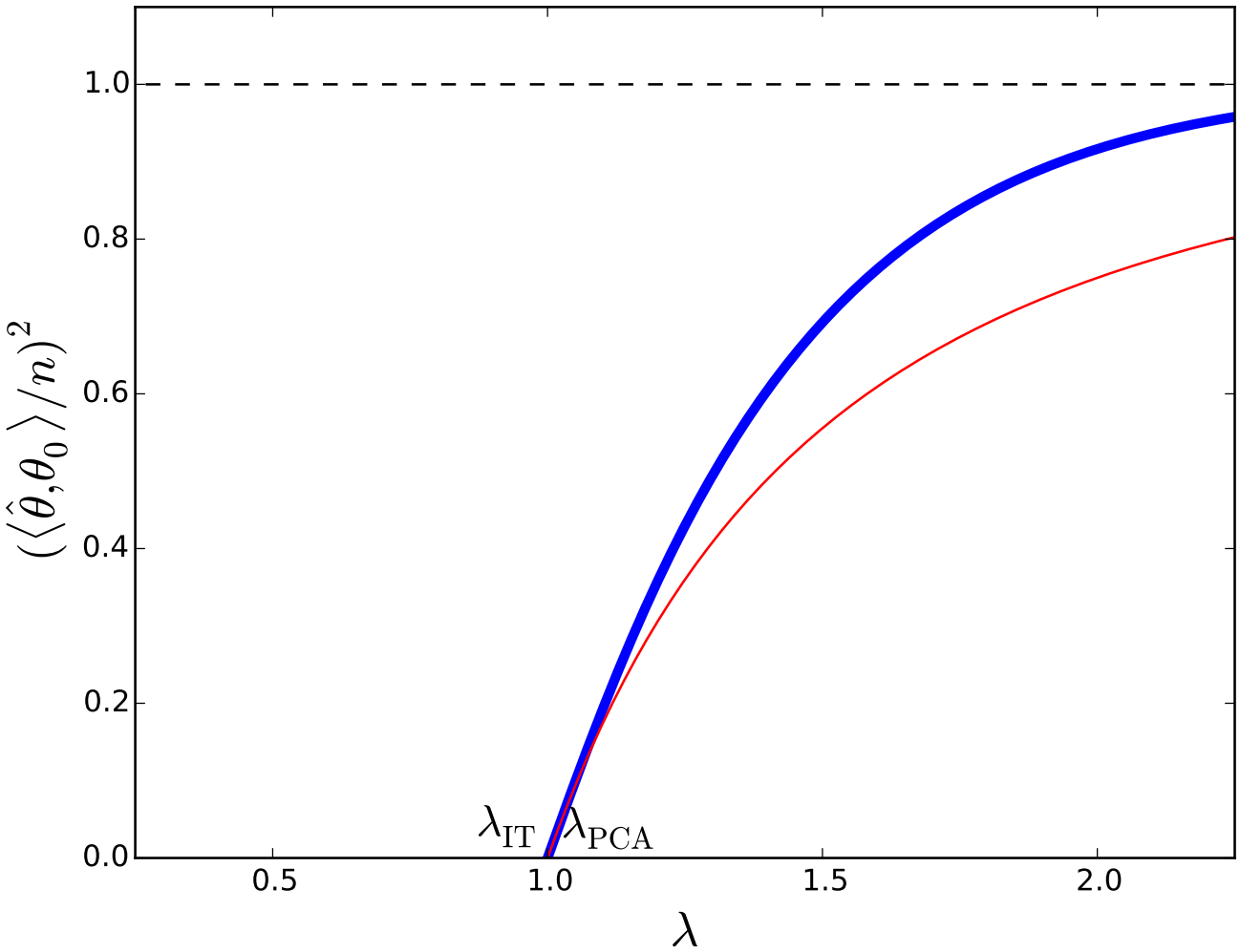
$$\Psi(\gamma, \lambda) \equiv \frac{\lambda^2}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathrm{I}(\gamma)\,,$$
$$\gamma_{\mathrm{Bayes}}(\lambda) \equiv \arg\max_{\gamma \geq 0} \Psi(\gamma, \lambda)\,.$$

*Then*

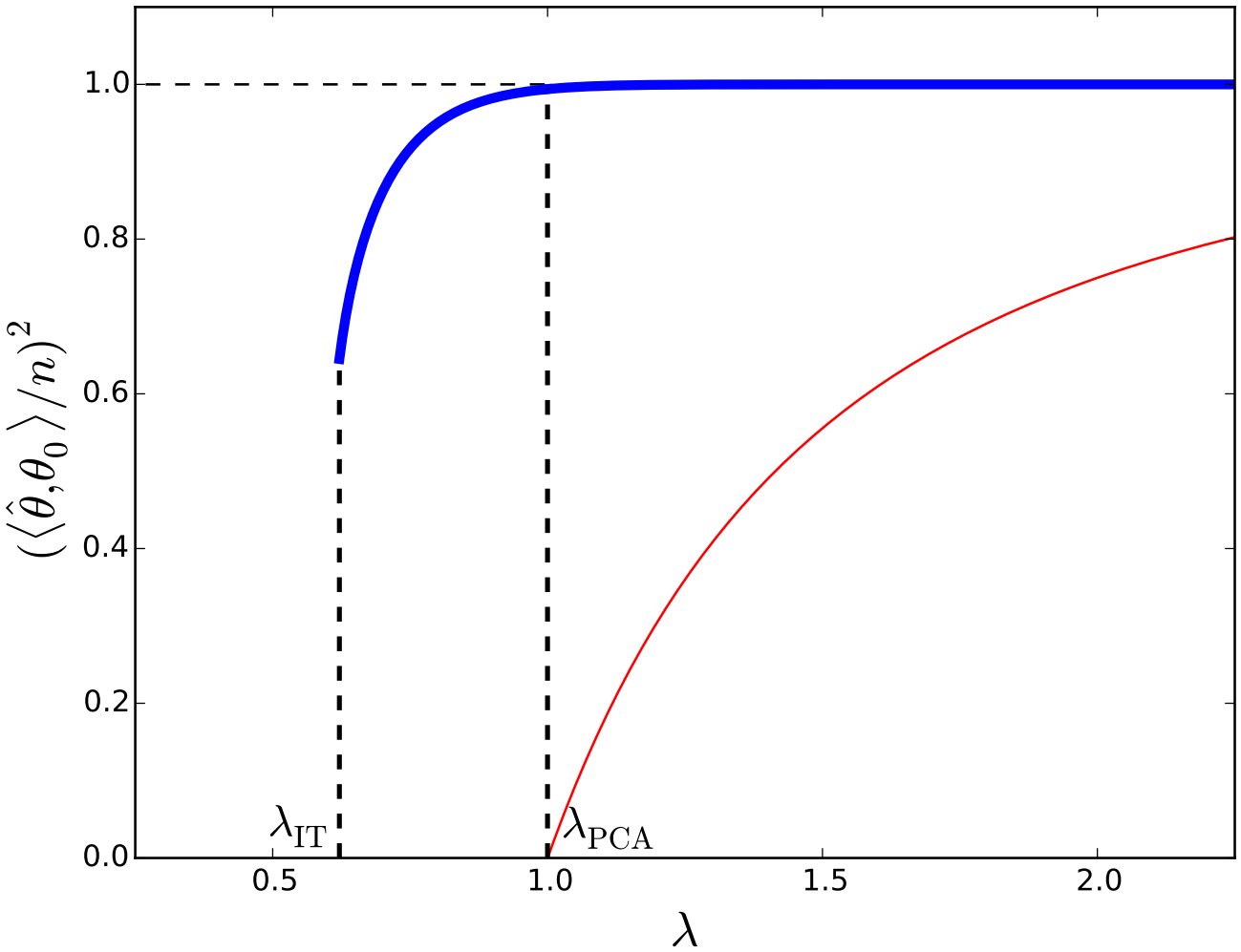
$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \frac{\sqrt{\gamma_{\mathrm{Bayes}}(\lambda)}}{\lambda}\,.$$

[see also: Deshpande, M., 2014; Deshpande, Abbe, M., 2016; Barbier, Dia, Macris, Krzakala, Lesieur, Zdeborová, 2016; Krzakala, Xu, Zdeborová, 2016; El Alaoui, Jordan, 2018; ...]

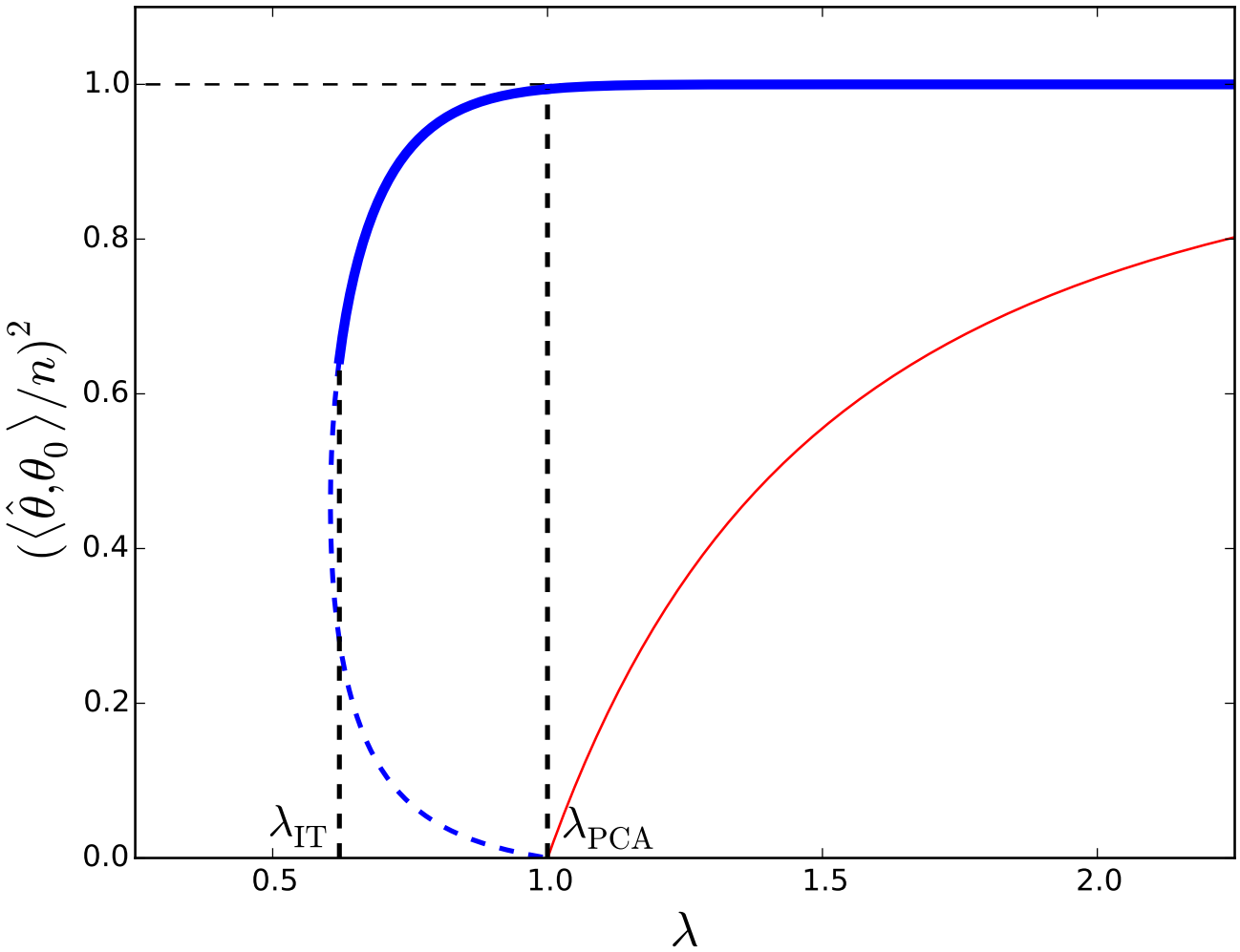
# $\varepsilon = 0.5$



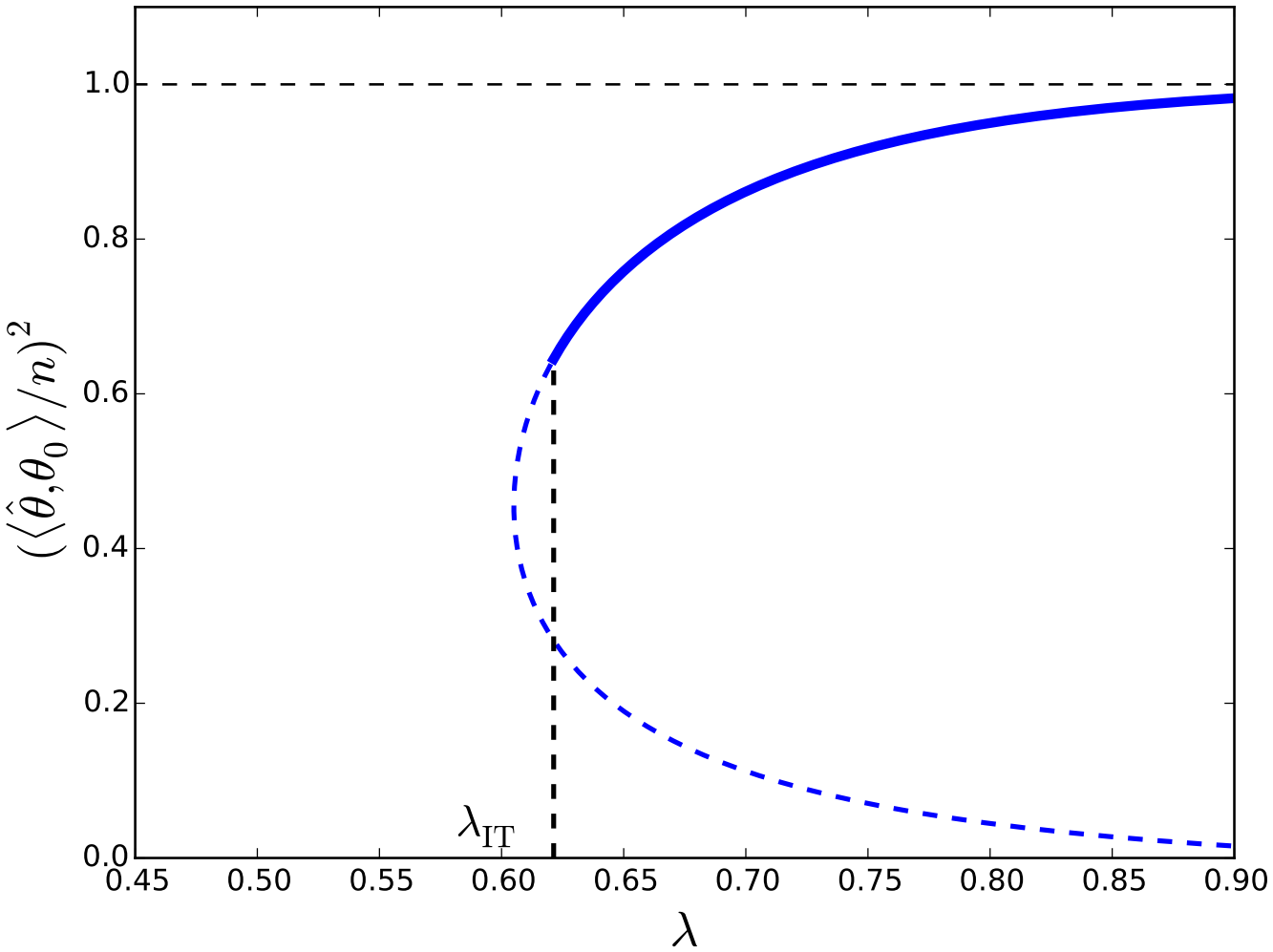
# $\varepsilon = 0.025$



# $\varepsilon = 0.025$

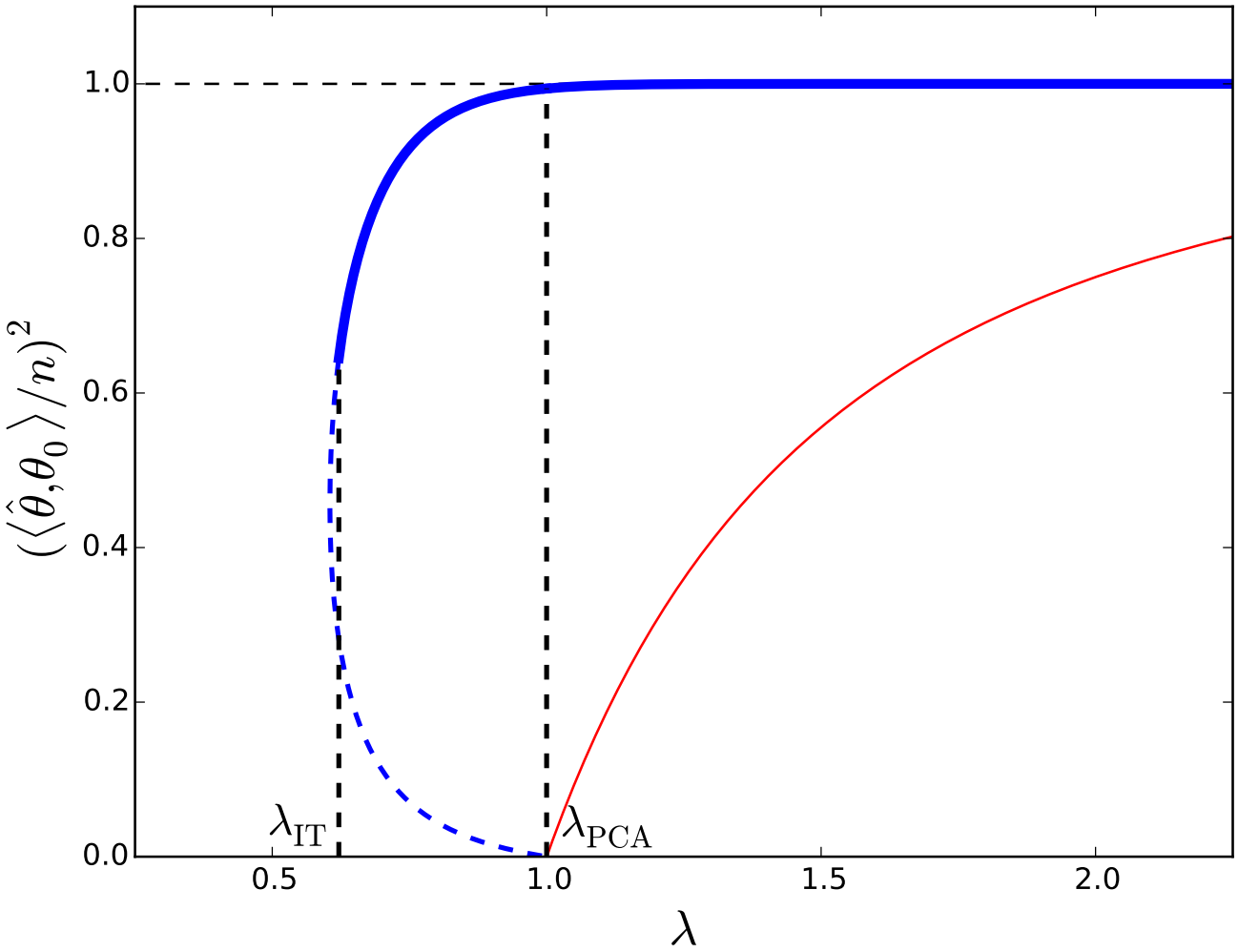


$\varepsilon = 0.025$



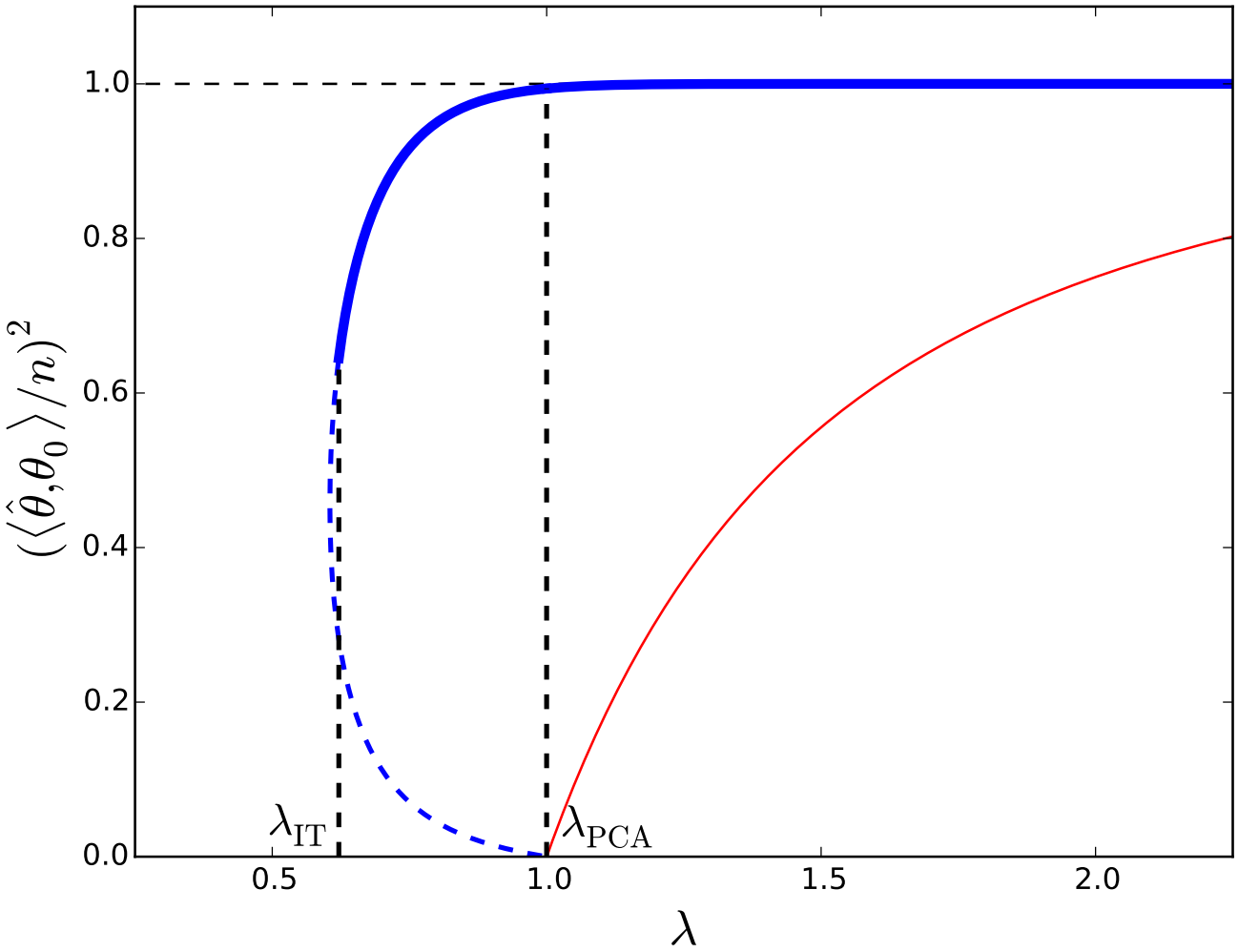
# Efficient algorithms

# Are we happy about this?




Not really...

# Are we happy about this?




**Not really...**

# Approximate Message Passing (AMP)

**Data:** $\boldsymbol{X} \in \mathbb{R}^{n\times n}$

**Initialization:** $\hat{\boldsymbol{\theta}}^0 \in \mathbb{R}^n$

**Iteration:**

$$\hat{\boldsymbol{\theta}}^{t+1} = \boldsymbol{X}\, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t) - \mathsf{b}_t\, \boldsymbol{f}_{t-1}(\hat{\boldsymbol{\theta}}^{t-1})\,,$$

$$\mathsf{b}_t \equiv \frac{1}{n}\mathrm{div}\boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)\,.$$

# Approximate Message Passing (AMP)

$$\hat{\boldsymbol{\theta}}^{t+1} = \boldsymbol{X}\, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t) - \mathsf{b}_t\, \boldsymbol{f}_{t-1}(\hat{\boldsymbol{\theta}}^{t-1})\,,$$
$$\mathsf{b}_t \equiv \frac{1}{n}\mathrm{div}\boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)\,.$$

- Generalization of iterative scheme to solve TAP equations for mean field spin glasses
  [Thouless, Anderson, Palmer, 1977; Kabashima 2003]
- Closely related to loopy belief propagation
  [Gallager, 1962; Pearl, 1986; ...]

# Analysis (heuristic!!!)

$$\hat{\boldsymbol{\theta}}^t \approx \mu_t \boldsymbol{\theta}_0 + \sigma_t \boldsymbol{g}\,, \quad \boldsymbol{g} = \mathsf{N}(\mathbf{0}, \mathrm{I}_{n\times n})$$

$$\boldsymbol{X} f_t(\hat{\boldsymbol{\theta}}^t) = \frac{\lambda}{n}\langle \boldsymbol{\theta}_0, f_t(\hat{\boldsymbol{\theta}}^t)\rangle \boldsymbol{\theta}_0 + \boldsymbol{W} f_t(\hat{\boldsymbol{\theta}}^t)$$

$$\approx \mu_{t+1}\boldsymbol{\theta}_0 + \sigma_{t+1}\boldsymbol{g}'$$

Separable $f_t(x) = (f_t(x_1), \dots, f_t(x_n))$, $\qquad ((X_0, G) \sim \widehat{p}_{\theta_0}(\,\cdot\,) \otimes \mathsf{N}(0,1))$

$$\mu_{t+1} = \lambda \mathbb{E}\{X_0 f_t(\mu_t X_0 + \sigma_t G)\}\,,$$
$$\sigma^2_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sigma_t G)^2\}\,.$$

# Analysis (heuristic!!!)

$$\hat{\boldsymbol{\theta}}^t \approx \mu_t \boldsymbol{\theta}_0 + \sigma_t \boldsymbol{g}\,, \quad \boldsymbol{g} = \mathsf{N}(\mathbf{0}, \mathrm{I}_{n\times n})$$

$$\boldsymbol{X} \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t) = \frac{\lambda}{n}\langle \boldsymbol{\theta}_0, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)\rangle \boldsymbol{\theta}_0 + \boldsymbol{W}\, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)$$

$$\approx \mu_{t+1} \boldsymbol{\theta}_0 + \sigma_{t+1} \boldsymbol{g}'$$

Separable $f_t(x) = (f_t(x_1), \dots, f_t(x_n))$, $((X_0, G) \sim \widehat{p}_{\theta_0}(\,\cdot\,) \otimes \mathsf{N}(0,1))$

$$\mu_{t+1} = \lambda \mathbb{E}\{X_0 f_t(\mu_t X_0 + \sigma_t G)\}\,,$$
$$\sigma^2_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sigma_t G)^2\}\,.$$

# Analysis (heuristic!!!)

$$\hat{\boldsymbol{\theta}}^t \approx \mu_t \boldsymbol{\theta}_0 + \sigma_t \boldsymbol{g}\,, \quad \boldsymbol{g} = \mathsf{N}(\mathbf{0}, \mathrm{I}_{n\times n})$$

$$\boldsymbol{X} \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t) = \frac{\lambda}{n}\langle \boldsymbol{\theta}_0, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)\rangle \boldsymbol{\theta}_0 + \boldsymbol{W}\, \boldsymbol{f}_t(\hat{\boldsymbol{\theta}}^t)$$

$$\approx \mu_{t+1} \boldsymbol{\theta}_0 + \sigma_{t+1} \boldsymbol{g}'$$

Separable $\boldsymbol{f}_t(\boldsymbol{x}) = (f_t(x_1), \dots, f_t(x_n))$, $\qquad ((X_0, G) \sim \widehat{p}_{\boldsymbol{\theta}_0}(\,\cdot\,) \otimes \mathsf{N}(0,1))$

$$\mu_{t+1} = \lambda \mathbb{E}\{X_0 f_t(\mu_t X_0 + \sigma_t G)\}\,,$$
$$\sigma^2_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sigma_t G)^2\}\,.$$

# Consequence

$$\mu_{t+1} = \lambda \mathbb{E}\{X_0 f_t(\mu_t X_0 + \sigma_t G)\}\,,$$
$$\sigma^2_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sigma_t G)^2\}\,.$$

**Bayes-AMP algorithm**

- Optimal denoiser: $f_t(y) = \mathbb{E}\{X_0 | \mu_t X_0 + \sigma_t G = y\}.$
- Initialization $\hat{\boldsymbol{\theta}}^0 = c \cdot \hat{\boldsymbol{\theta}}^{\rm PCA}(\boldsymbol{X})$

# AMP estimation

**Theorem** (M., Venkatarmanan, 2018)

*Consider the rank-one model with $\theta_{0,i} \sim p$ (+ conditions), $\lambda > 1$ and define*

$$\Psi(\gamma,\lambda) \equiv \frac{\lambda^2}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathrm{I}(\gamma)\,,$$
$$\gamma_{\mathrm{ALG}}(\lambda) \equiv \min\left\{\gamma > 0 : \ \ \partial_\gamma \Psi(\gamma,\lambda) = 0\right\}.$$

*(Recall that $\gamma_{\mathrm{Bayes}}(\lambda) \equiv \arg\max_{\gamma\geq 0} \Psi(\gamma,\lambda)$) .*
*Then*

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^t(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^t(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \frac{\sqrt{\gamma_{\mathrm{ALG}}(\lambda)}}{\lambda}\,.$$

[see also: **Bolthausen, 2014**; Bayati, M., 2011; Bayati, Lelarge, M., 2015; Berthier, M., Nguyen 2018; . . . ]

# AMP estimation

**Theorem** (M., Venkatarmanan, 2018)

*Consider the rank-one model with $\theta_{0,i} \sim p$ (+ conditions), $\lambda > 1$ and define*

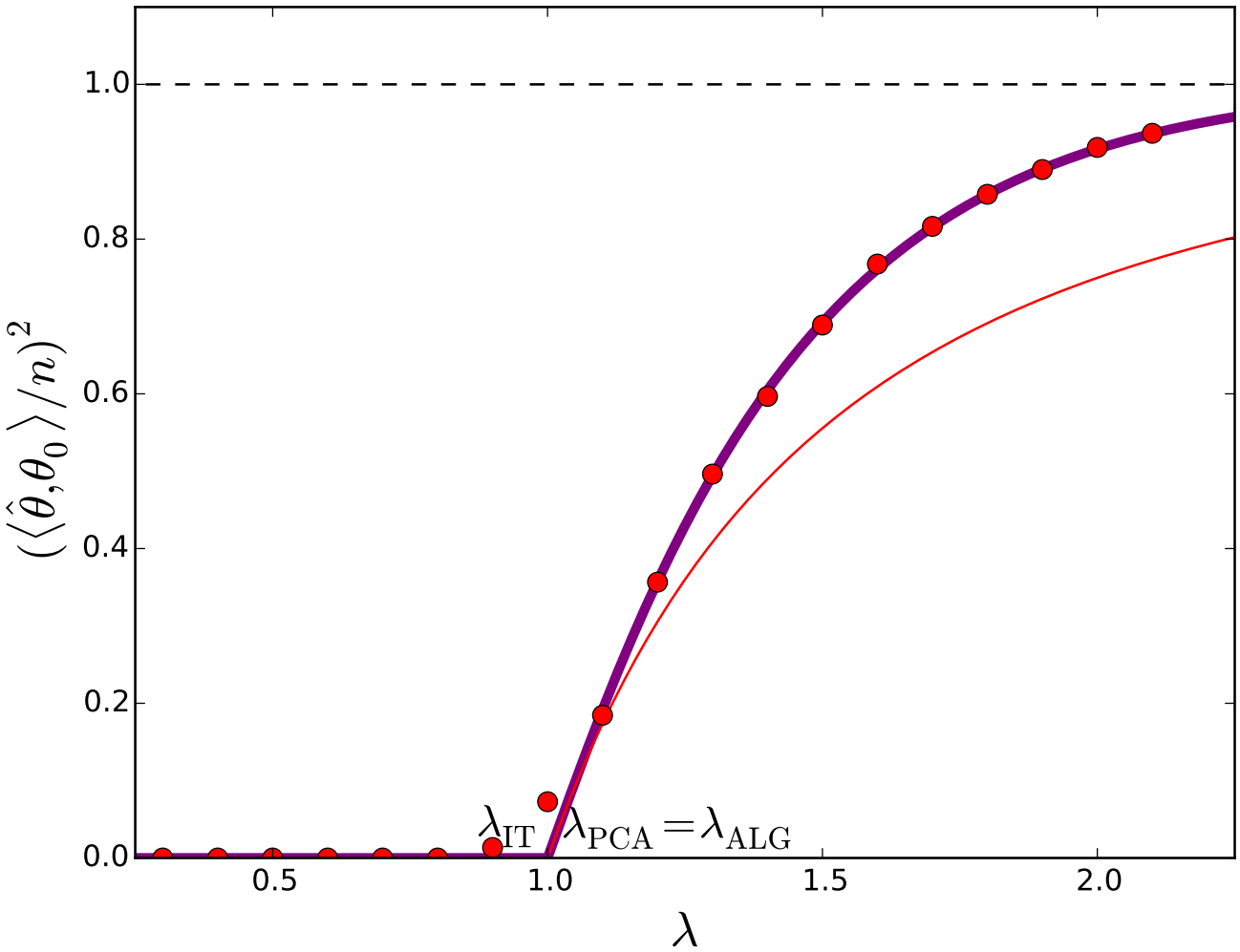
$$\Psi(\gamma,\lambda) \equiv \frac{\lambda^2}{4} + \frac{\gamma^2}{4\lambda} - \frac{\gamma}{2} + \mathrm{I}(\gamma)\,,$$
$$\gamma_{\mathrm{ALG}}(\lambda) \equiv \min\left\{\gamma > 0 : \ \ \partial_\gamma \Psi(\gamma,\lambda) = 0\right\}.$$

*(Recall that $\gamma_{\mathrm{Bayes}}(\lambda) \equiv \arg\max_{\gamma\geq 0} \Psi(\gamma,\lambda)$) .*
*Then*

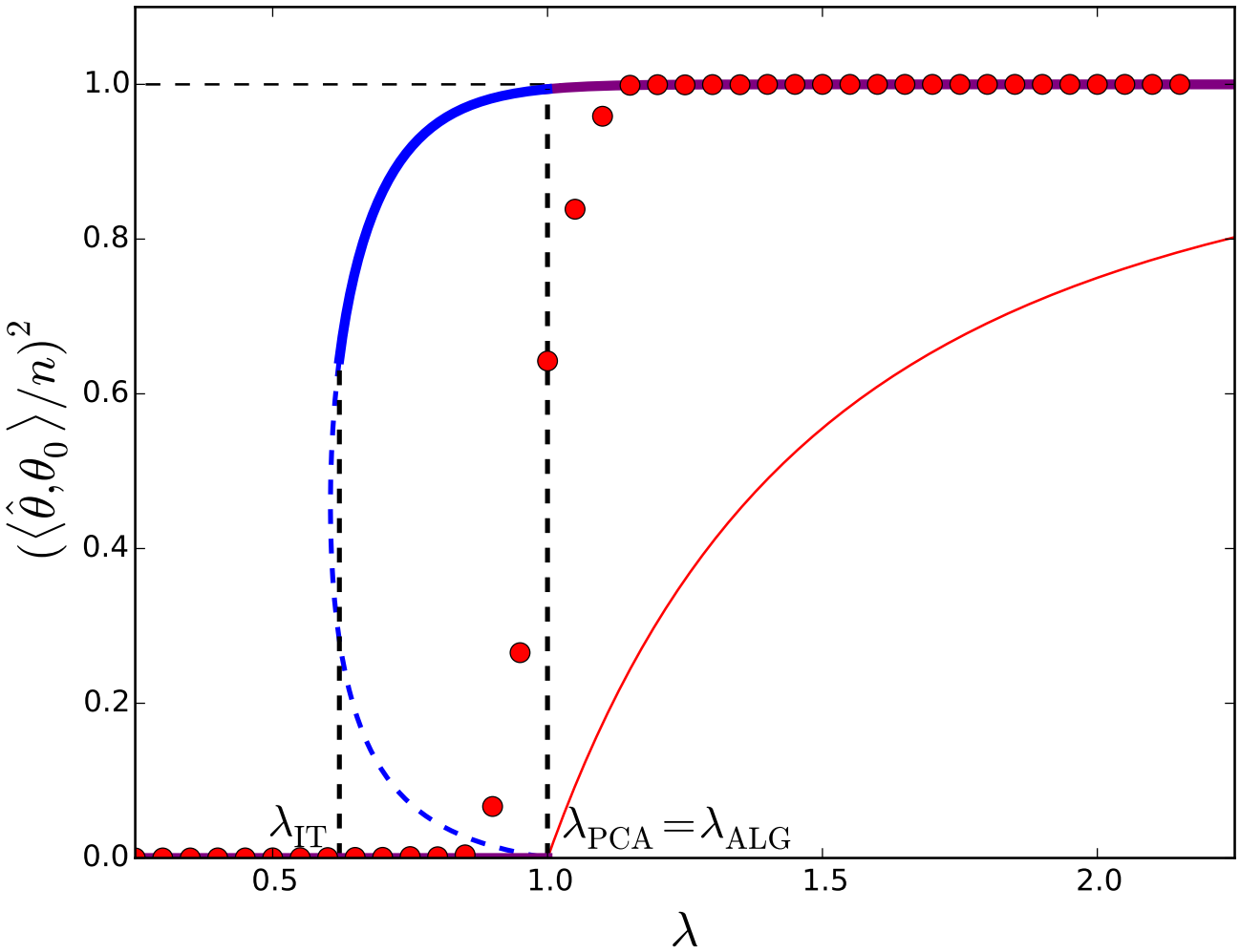
$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^t(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^t(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \frac{\sqrt{\gamma_{\mathrm{ALG}}(\lambda)}}{\lambda}\,.$$

[see also: **Bolthausen, 2014**; Bayati, M., 2011; Bayati, Lelarge, M., 2015; Berthier, M., Nguyen 2018; . . . ]

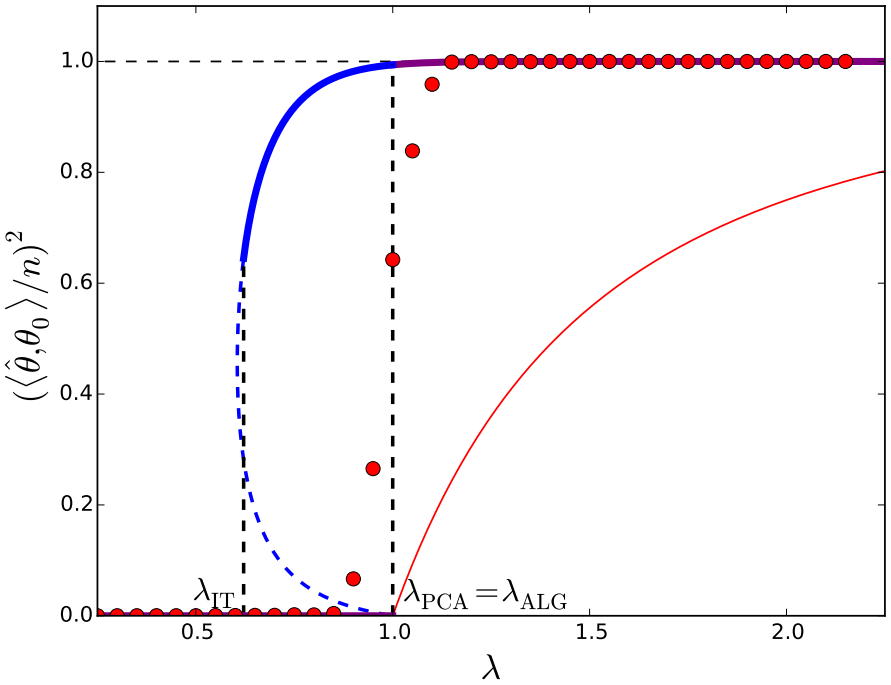
# $\varepsilon = 0.5$



# $\varepsilon = 0.025$



# A bold conjecture

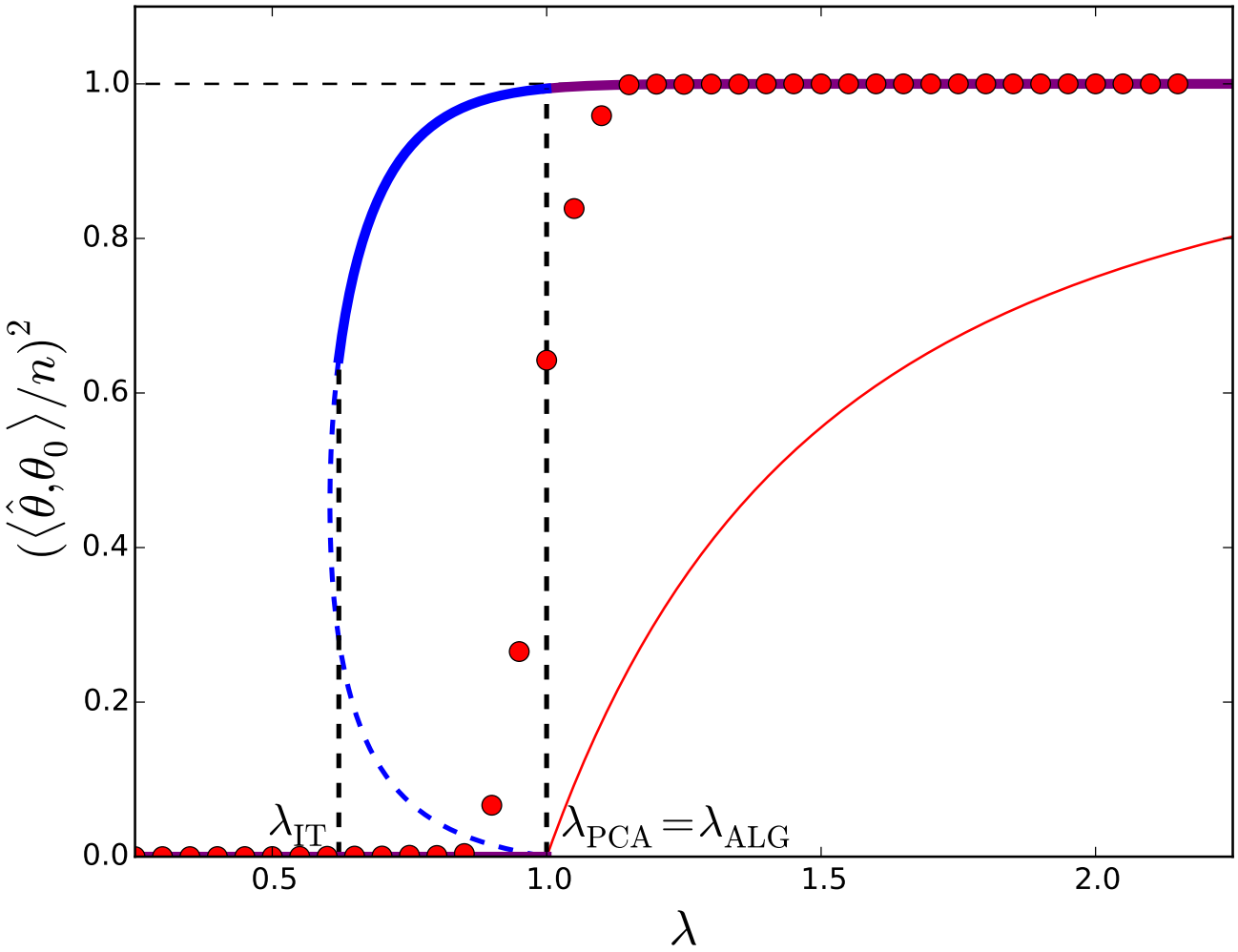



## Conjecture

*No polytime algorithm can beat AMP.*

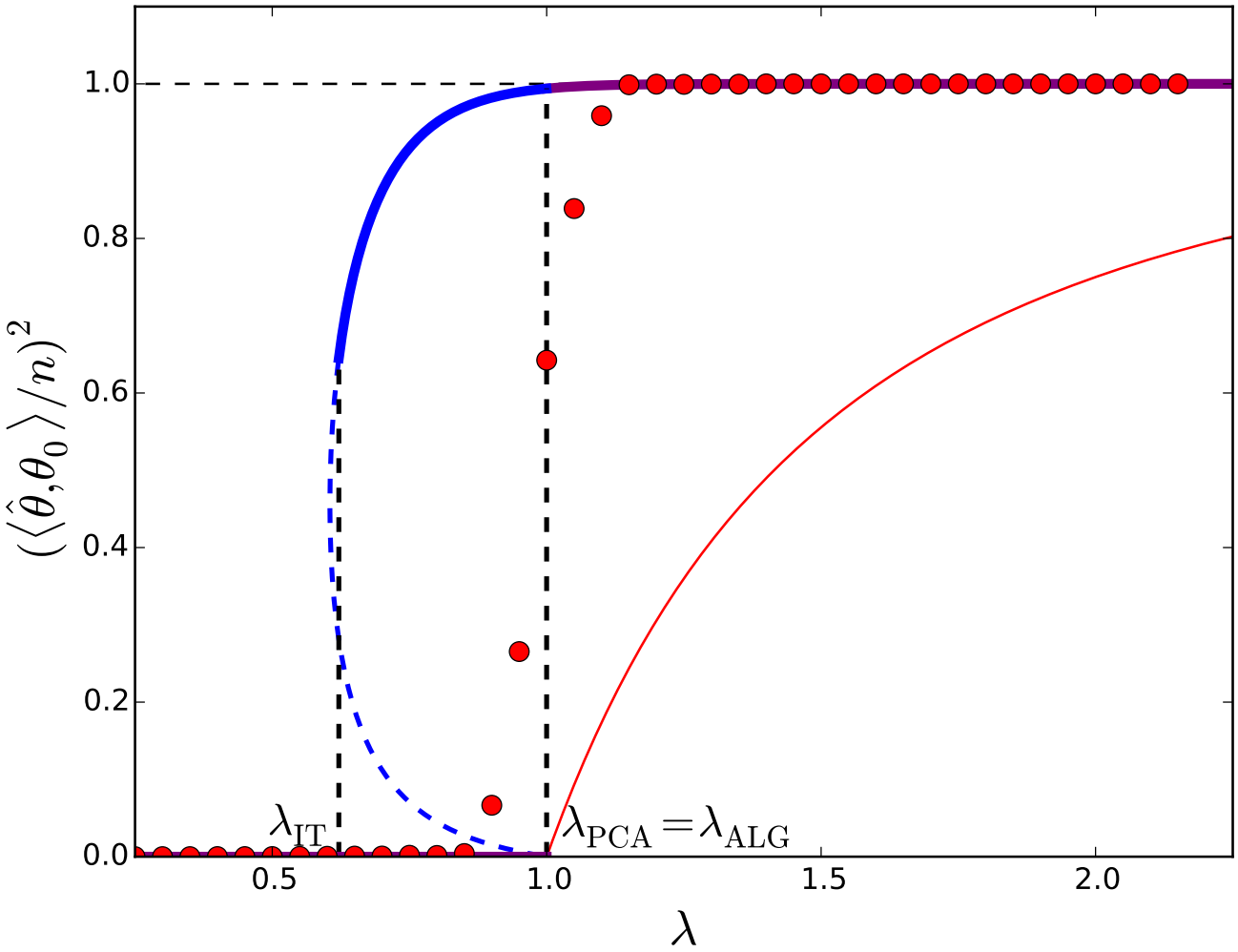
## Generic/robust algorithms

# Are we happy about this?




Not really...

# Are we happy about this?




**Not really...**

Hereafter $\varepsilon = 0.5$ (generalizable?)

# Starting point: Maximum likelihood

$\mathsf{OPT}(\boldsymbol{X})$

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{\sigma}, \boldsymbol{X}\boldsymbol{\sigma} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

- Meaningful irrespective of the model
- Robust to 'small' changes in $\boldsymbol{X}$
- NP hard

# Semi-Definite Programming relaxation

$\mathsf{OPT}(\boldsymbol{X})$

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

**Equivalent formulation**

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{Z} \in \mathbb{R}^{n \times n} \,, \quad \boldsymbol{Z} \succeq \boldsymbol{0} \,, \quad \mathrm{rank}(\boldsymbol{Z}) = 1 \\ & Z_{ii} = 1 \quad \forall i \in \{1, \dots, n\}. \end{aligned}$$

# Semi-Definite Programming relaxation

$\mathsf{OPT}(\boldsymbol{X})$

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

**Relaxation**

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{Z} \in \mathbb{R}^{n \times n}, \quad \boldsymbol{Z} \succeq \boldsymbol{0}, \quad \cancel{\mathrm{rank}(\boldsymbol{Z}) = 1} \\ & Z_{ii} = 1 \quad \forall i \in \{1, \dots, n\}. \end{aligned}$$

# Semi-Definite Programming relaxation

OPT($\boldsymbol{X}$)

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

SDP($\boldsymbol{X}$)

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{Z} \in \mathbb{R}^{n \times n} \,, \quad \boldsymbol{Z} \succeq \boldsymbol{0} \,, \\ & Z_{ii} = 1 \quad \forall i \in \{1, \dots, n\}. \end{aligned}$$

$$\hat{\theta}^{\mathsf{SDP}}(\boldsymbol{X}) = c \cdot v_1(Z_*) \,, \quad \text{(principal eigenvector)}$$

$$\hat{\theta}^{\mathsf{SDP,theor.}}(\boldsymbol{X}) \sim \mathsf{N}(0, c Z_*) \,.$$

# Semi-Definite Programming relaxation

$\mathsf{OPT}(\boldsymbol{X})$

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{\sigma}\boldsymbol{\sigma}^{\mathsf{T}} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{\sigma} \in \{+1, -1\}^n \,. \end{aligned}$$

$\mathsf{SDP}(\boldsymbol{X})$

$$\begin{aligned} \text{maximize} \quad & \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \,, \\ \text{subject to} \quad & \boldsymbol{Z} \in \mathbb{R}^{n \times n} \,, \quad \boldsymbol{Z} \succeq \boldsymbol{0} \,, \\ & Z_{ii} = 1 \quad \forall i \in \{1, \dots, n\}. \end{aligned}$$

$$\hat{\boldsymbol{\theta}}^{\mathsf{SDP}}(\boldsymbol{X}) = c \cdot \boldsymbol{v}_1(\boldsymbol{Z}_*) \,, \quad \text{(principal eigenvector)}$$

$$\hat{\boldsymbol{\theta}}^{\mathsf{SDP,theor.}}(\boldsymbol{X}) \sim \mathsf{N}(0, c\boldsymbol{Z}_*) \,.$$

# Optimal weak recovery threshold

**Theorem** (M., Sen, 2016)

*Consider the $\mathbb{Z}_2$-synchronization model with $\boldsymbol{\theta}_0 \in \{+1,-1\}^n$. For any $\lambda > 1$, there exist $\varepsilon(\lambda) > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{E}\left\{ \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathsf{SDP,theor.}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathsf{SDP,theor.}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\} \geq \varepsilon(\lambda)\,.$$

[see also: Guédon, Vershynin, 2016; Hajek, Wu, Xu, 2016; Perry, Wein, Moitra, 2016; Bandeira, 2018; ...]

# Optimal weak recovery threshold

**Theorem** (M., Sen, 2016)

*Consider the $\mathbb{Z}_2$-synchronization model with $\boldsymbol{\theta}_0 \in \{+1,-1\}^n$. For any $\lambda > 1$, there exist $\varepsilon(\lambda) > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{E}\left\{ \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathsf{SDP,theor.}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathsf{SDP,theor.}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} \right\} \geq \varepsilon(\lambda)\,.$$

[see also: Guédon, Vershynin, 2016; Hajek, Wu, Xu, 2016; Perry, Wein, Moitra, 2016; Bandeira, 2018; ...]

# Exact limit: Statistical physics prediction

**Conjecture** (Javanmard, M., Ricci-Tersenghi, 2016)

*In the $\mathbb{Z}_2$ synchronization model*

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathsf{SDP}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathsf{SDP}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \mathrm{ExplicitFormula}(\lambda)\,.$$

# ExplicitFormula($\lambda$)

For $G \sim \mathsf{N}(0,1)$, define $\rho = \rho(G;\mu,q,r)$ via

$$1 = \frac{(\mu + \sqrt{q}\, G)^2}{(\rho + r)^2} + \frac{1-q}{\rho^2}\,.$$

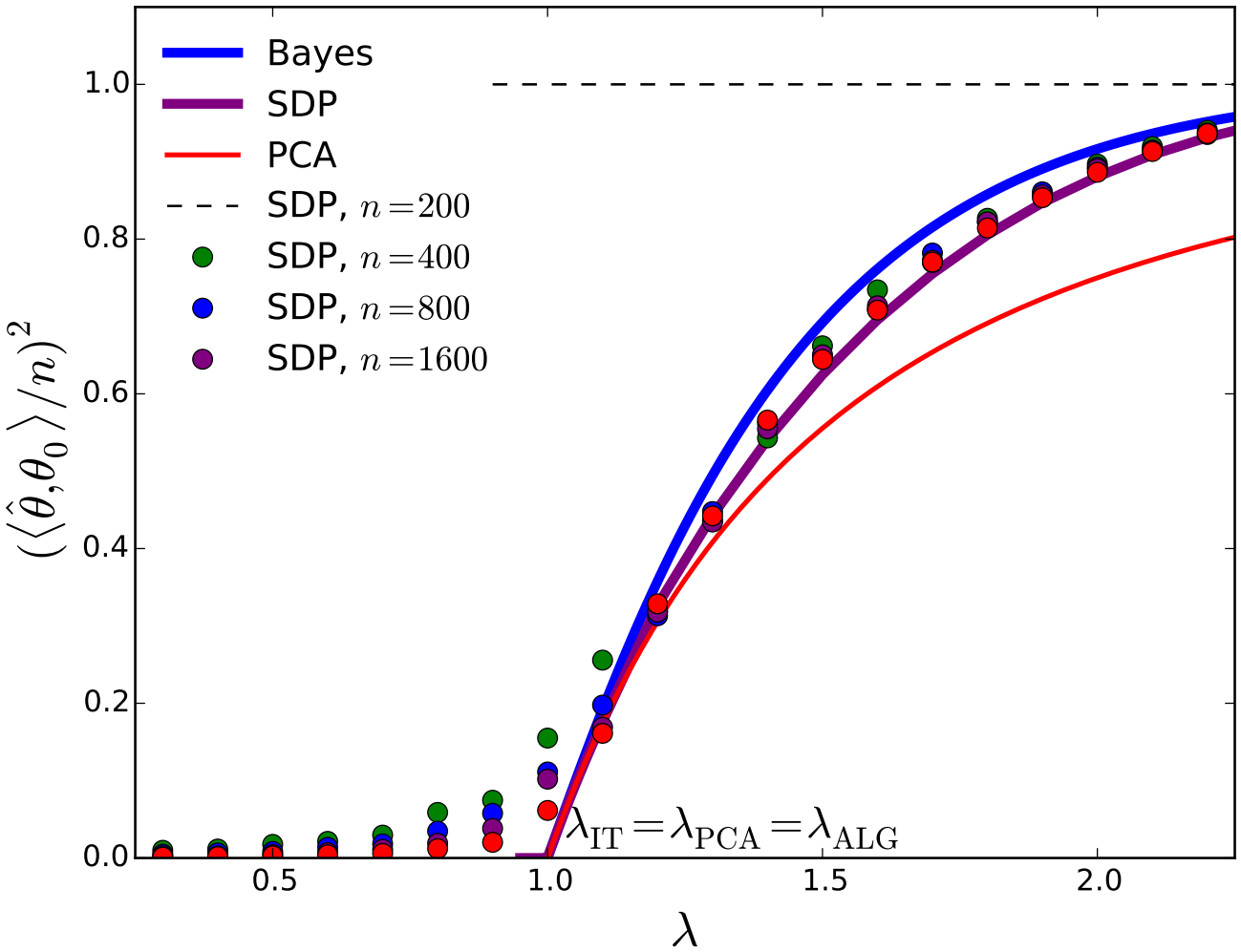
Let $\mu,\ q,\ r \in \mathbb{R}$ be the solution of

$$\mu = \lambda \mathbb{E}\left\{\frac{\mu + \sqrt{q}\, G}{\rho + r}\right\}, \qquad q = \mathbb{E}\left\{\frac{(\mu + \sqrt{q}\, G)^2}{(\rho + r)^2}\right\},$$

$$r = \mathbb{E}\Big\{\frac{1}{\rho} - \frac{\mu}{\sqrt{q}}\,\frac{G}{\rho + r} - \frac{G^2}{\rho + r}\Big\}\,.$$

$$\lim_{n\to\infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\mathsf{SDP}}(\boldsymbol{X}), \boldsymbol{\theta}_0\rangle|}{\|\hat{\boldsymbol{\theta}}^{\mathsf{SDP}}(\boldsymbol{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = 1 - 2\Phi\Big(-\frac{\mu(\lambda)}{\sqrt{q(\lambda)}}\Big)\,.$$

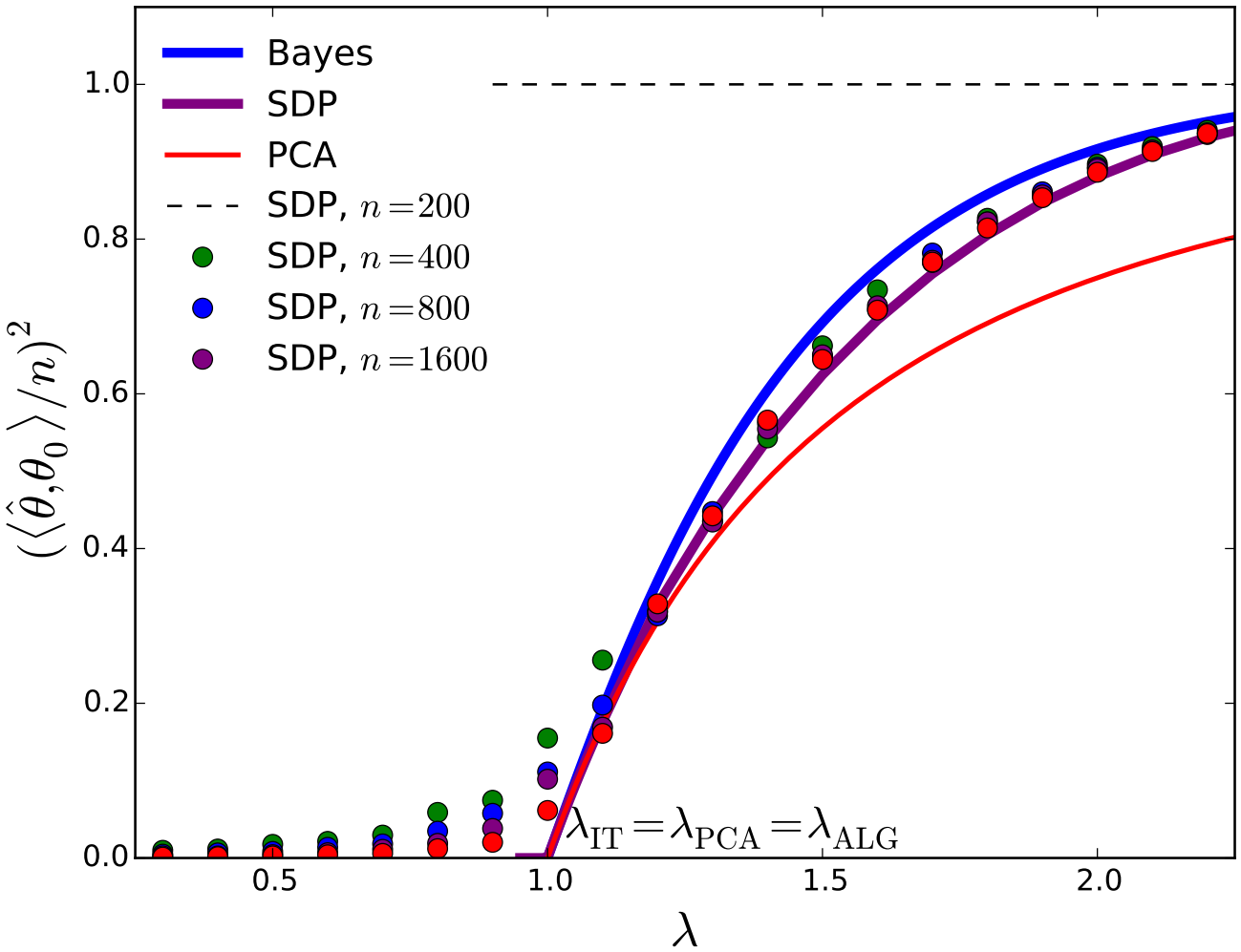
# Bayes vs SDP vs PCA



# SDP

- Nearly Bayes-optimal
- Computationally practical
- Generic/robust
- Analysis still incomplete

# Can we close the gap?



# TAP free energy

$$\mathscr{F}_{\lambda,\boldsymbol{X}} : [-1,1]^n \to \mathbb{R}$$

$$\mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m}) = -\frac{1}{n}\sum_{i=1}^{n} h(m_i) - \frac{\lambda}{2n}\langle \boldsymbol{m}, \boldsymbol{X}\,\boldsymbol{m}\rangle - \frac{\lambda^2}{4}\Big(1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\Big)^2,$$

[Thouless, Anderson, Palmer, 1977]

**Statistical physics prediction**

$$\hat{\boldsymbol{\theta}}^{\text{Bayes}}(\boldsymbol{X}) = \mathbb{E}\{\boldsymbol{\theta}|\boldsymbol{X}\} \approx \arg\min_{\boldsymbol{m}\in[-1,+1]^n} \mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m})$$

[Technical nuisance: need to break $\boldsymbol{m} \leftrightarrow -\boldsymbol{m}$ symmetry]

# TAP free energy

$$\mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m}) = -\frac{1}{n}\sum_{i=1}^{n} h(m_i) - \frac{\lambda}{2n}\langle \boldsymbol{m}, \boldsymbol{X}\,\boldsymbol{m}\rangle - \frac{\lambda^2}{4}\Big(1 - \frac{\|\boldsymbol{m}\|_2^2}{n}\Big)^2,$$

**Interpretation**

$$\mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m}) \approx \min\Big\{\mathscr{D}_{\mathrm{KL}}(p\|p_{\mathrm{Bayes}}(\,\cdot\,|\boldsymbol{X})) \quad \text{such that} \quad p \in \mathscr{P}(\{+1,-1\}^n),\quad \mathsf{E}_p(\boldsymbol{\sigma}) = \boldsymbol{m}\Big\}$$

[see also: Chatterjee 2010; Talagrand 2010; ...]

# TAP estimator

$$\hat{\boldsymbol{\theta}}^{\mathrm{TAP}}(\boldsymbol{X}) \equiv \arg\min_{\boldsymbol{m}\in[-1,+1]^n} \mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m})\,.$$

- Non-convex :-(
- Stationary points $\leftrightarrow$ AMP fixed points :-)
- More robust than AMP :-)
- Can we prove anything?

# TAP estimator

$$\hat{\boldsymbol{\theta}}^{\mathrm{TAP}}(\boldsymbol{X}) \equiv \arg \min_{\boldsymbol{m} \in [-1,+1]^n} \mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m})\,.$$

- Non-convex :-(
- Stationary points $\leftrightarrow$ AMP fixed points :-)
- More robust than AMP :-)
- Can we prove anything?

# TAP estimator

$$\hat{\boldsymbol{\theta}}^{\mathrm{TAP}}(\boldsymbol{X}) \equiv \arg\min_{\boldsymbol{m}\in[-1,+1]^n} \mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m})\,.$$

- Non-convex :-(
- Stationary points $\leftrightarrow$ AMP fixed points :-)
- More robust than AMP :-)
- Can we prove anything?

# What would we like to prove?

$$\hat{\boldsymbol{\theta}}^{\mathrm{TAP}}(\boldsymbol{X}) \equiv \arg \min_{\boldsymbol{m} \in [-1,+1]^n} \mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m})\,.$$

- Landscape is 'simple' (can be minimized efficiently)
- Global minimum close to $\hat{\boldsymbol{\theta}}^{\mathrm{Bayes}}(\boldsymbol{X})$

# A recent result

**Theorem** (Fan, Mei, M, 2018)

*Denote $\mathcal{C}_{\lambda,n}$ be the set of critical points of $\mathscr{F}_{\lambda,\boldsymbol{X}}$ with $\mathscr{F}_{\lambda,\boldsymbol{X}}(\boldsymbol{m}) \le -\lambda^2/3\}$. There exists $\lambda_0 > 0$, such that for any $\lambda > \lambda_0$, we have*

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

# Conclusion

# Conclusion: A general strategy

- High-dimensional Bayes posterior $\to$ Fundamental statistical limts
- AMP $\to$ Computational limits
- Convex relaxations $\to$ Robust/generic algorithms
- Free energy approximations $\to$ Better cost functions

Thanks!

# Conclusion: A general strategy

- High-dimensional Bayes posterior $\to$ Fundamental statistical limts
- AMP $\to$ Computational limits
- Convex relaxations $\to$ Robust/generic algorithms
- Free energy approximations $\to$ Better cost functions

Thanks!