Information-theoretically Optimal Sparse PCA

Yash Deshpande and Andrea Montanari

Stanford University

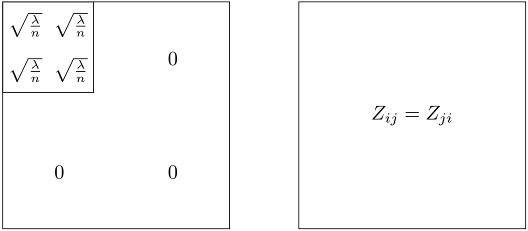
July 3rd, 2014

Problem Definition

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z}.$$

Problem Definition

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z}.$$

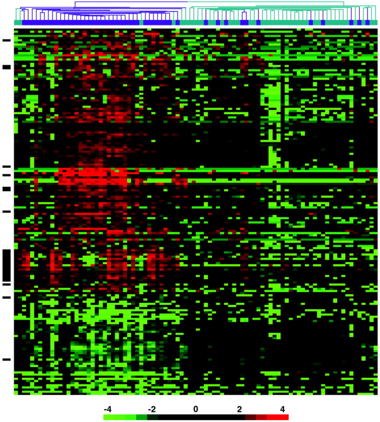


$x_i \sim \text{Bernoulli}(\varepsilon)$, $Z_{ij} \sim \text{Normal}(0, 1)$ independent.

Estimate $\mathbf{X} = \mathbf{x}\mathbf{x}^\mathsf{T}$ from $\mathbf{Y}_\lambda$

An example: gene expression data

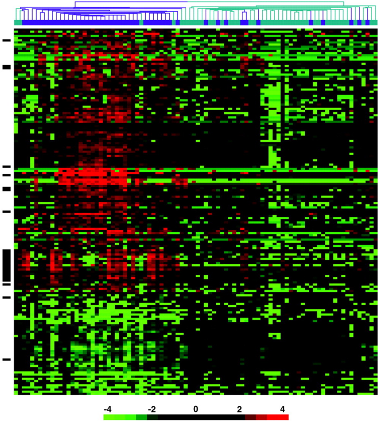
[Baechler et al, 2003 PNAS]



- Genes $\times$ patients matrix
- Blue - lupus patients, Aqua - healthy controls
- Black - a subset of immune system specific genes

An example: gene expression data

[Baechler et al, 2003 PNAS]



- Genes $\times$ patients matrix
- Blue - lupus patients,
  Aqua - healthy controls
- Black - a subset of immune system specific genes

A simple probabilistic model

Related work

Detection and estimation:

$$Y = X + \text{ noise } .$$

- $X \in \mathcal{S} \subset \{0,1\}^n$, a known set
- Goal: hypothesis testing, support recovery
- [Donoho, Jin 2004], [Addario-Berry et al. 2010], [Arias-Castro et al. 2011] ...

Related work

Machine learning:

$$\begin{aligned} \text{maximize} \quad & \langle \mathbf{v}, \mathbf{Y}_\lambda \mathbf{v} \rangle \\ \text{subject to:} \quad & \|\mathbf{v}\|^2 \le 1, \mathbf{v} \text{ is sparse.} \end{aligned}$$

- Goal: maximize “variance”, support recovery
- [d’Aspremont et al. 2004], [Moghaddam et al. 2005], [Zou et al. 2006], [Amini, Wainwright 2009] , [Papailiopoulos et al. 2013]. . .

Related work

Information theory:

$$\text{minimize } \|\mathbf{Y}_\lambda - \mathbf{v}\mathbf{v}^\mathsf{T}\|_F^2 + f(\mathbf{v}).$$

- Probabilistic model for $\mathbf{x}, \mathbf{Y}_\lambda$
- Propose approximate message passing algorithm
- [Rangan, Fletcher 2012], [Kabashima et al. 2014]

A first try: simple PCA

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z}.$$

A first try: simple PCA

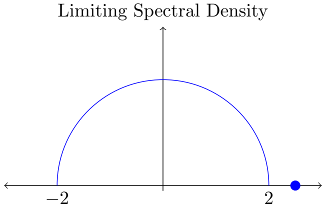
$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z}.$$

Estimate $\mathbf{x}$ using scaled principal eigenvector $\mathbf{x}_1(\mathbf{Y}_\lambda)$.

Limitations of PCA

Limitations of PCA

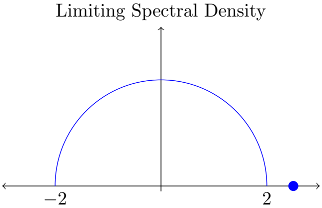
If $\lambda\varepsilon^2 > 1$



$\lim_{n\to\infty} \frac{\langle \mathbf{x}_1(\mathbf{Y}_\lambda),\mathbf{x}\rangle}{\sqrt{n\varepsilon}} > 0$ a. s.

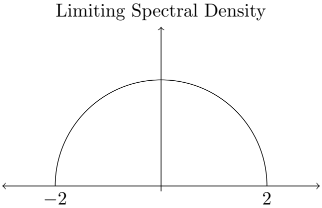
Limitations of PCA

If $\lambda\varepsilon^2 > 1$



$\lim_{n\to\infty} \frac{\langle \mathbf{x}_1(\mathbf{Y}_\lambda), \mathbf{x}\rangle}{\sqrt{n\varepsilon}} > 0$ a. s.

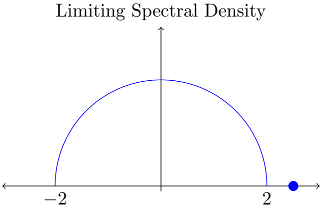
If $\lambda\varepsilon^2 < 1$



$\lim_{n\to\infty} \frac{\langle \mathbf{x}_1(\mathbf{Y}_\lambda), \mathbf{x}\rangle}{\sqrt{n\varepsilon}} = 0$ a. s.

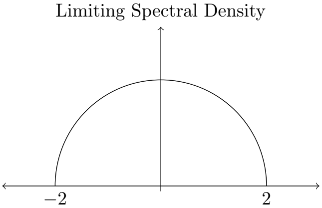
Limitations of PCA

If $\lambda\varepsilon^2 > 1$



$\lim_{n\to\infty} \frac{\langle \mathbf{x}_1(\mathbf{Y}_\lambda), \mathbf{x} \rangle}{\sqrt{n\varepsilon}} > 0$ a. s.

If $\lambda\varepsilon^2 < 1$



$\lim_{n\to\infty} \frac{\langle \mathbf{x}_1(\mathbf{Y}_\lambda), \mathbf{x} \rangle}{\sqrt{n\varepsilon}} = 0$ a. s.

[Knowles, Yin, 2011]

Our contributions

- Poly-time algorithm that exploits sparsity

Our contributions

- Poly-time algorithm that exploits sparsity
- Provably optimal in terms of MSE when $\varepsilon > \varepsilon_c$

Our contributions

- Poly-time algorithm that exploits sparsity
- Provably optimal in terms of MSE when $\varepsilon > \varepsilon_c$
- "Single-letter" characterization of MMSE

Single letter characterization

Original high-dimensional problem

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z},$$

Single letter characterization

Original high-dimensional problem

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^{\mathsf{T}} + \mathbf{Z},$$

$$\text{M-mmse}(\lambda, n) \equiv \frac{1}{n^2}\mathbb{E}\left\{\|\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}_\lambda\}\|_F^2\right\}.$$

Single letter characterization

Original high-dimensional problem

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z},$$

$$\text{M-mmse}(\lambda, n) \equiv \frac{1}{n^2}\mathbb{E}\left\{\|\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}_\lambda\}\|_F^2\right\}.$$

Scalar problem

$$Y_\lambda = \sqrt{\lambda}X_0 + Z,$$

Single letter characterization

Original high-dimensional problem

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^{\mathsf{T}} + \mathbf{Z},$$

$$\text{M-mmse}(\lambda, n) \equiv \frac{1}{n^2}\mathbb{E}\left\{\|\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}_\lambda\}\|_F^2\right\}.$$

Scalar problem

$$Y_\lambda = \sqrt{\lambda}X_0 + Z,$$

$$\text{S-mmse}(\lambda) \equiv \mathbb{E}\left\{(X_0 - \mathbb{E}\{X_0|Y_\lambda\})^2\right\}.$$

Single letter characterization

Original high-dimensional problem

$$\mathbf{Y}_\lambda = \sqrt{\frac{\lambda}{n}}\mathbf{x}\mathbf{x}^\mathsf{T} + \mathbf{Z},$$

$$\text{M-mmse}(\lambda, n) \equiv \frac{1}{n^2}\mathbb{E}\left\{\|\mathbf{X} - \mathbb{E}\{\mathbf{X}|\mathbf{Y}_\lambda\}\|_F^2\right\}.$$

Scalar problem

$$Y_\lambda = \sqrt{\lambda}X_0 + Z,$$

$$\text{S-mmse}(\lambda) \equiv \mathbb{E}\left\{(X_0 - \mathbb{E}\{X_0|Y_\lambda\})^2\right\}.$$

Here $X_0 \sim \text{Bernoulli}(\varepsilon)$, $Z \sim \text{Normal}(0, 1)$.

Main result

Theorem (Deshpande, Montanari 2014)

There exists an $\varepsilon_c < 1$ *such that the following happens. For every* $\varepsilon > \varepsilon_c$

$$\lim_{n\to\infty} \text{M-mmse}(\lambda, n) = \varepsilon^2 - \tau_*^2$$

$$\text{where } \tau_* = \varepsilon - \text{S-mmse}(\lambda\tau_*).$$

Further there exists a polynomial time algorithm that achieves this MSE.

Main result

Theorem (Deshpande, Montanari 2014)

There exists an $\varepsilon_c < 1$ *such that the following happens. For every* $\varepsilon > \varepsilon_c$

$$\lim_{n\to\infty} \text{M-mmse}(\lambda, n) = \varepsilon^2 - \tau_*^2$$

$$\textit{where } \tau_* = \varepsilon - \text{S-mmse}(\lambda\tau_*).$$

Further there exists a polynomial time algorithm that achieves this MSE.

$\varepsilon_c \approx 0.05$ (solution to scalar non-linear equation)

Making use of sparsity

Making use of sparsity

The power iteration with $\mathbf{A} = \mathbf{Y}_\lambda/\sqrt{n}$:

$$\mathbf{x}^{t+1} = \mathbf{A}\ \mathbf{x}^{t}.$$

Making use of sparsity

The power iteration with $\mathbf{A} = \mathbf{Y}_\lambda/\sqrt{n}$:

$$\mathbf{x}^{t+1} = \mathbf{A}\ \mathbf{x}^t.$$

Improvement:

$$\mathbf{x}^{t+1} = \mathbf{A}F_t(\mathbf{x}^t),$$

where $F_t(\mathbf{x}^t) = (f_t(x_1^t), \ldots f_t(x_n^t))^{\mathsf{T}}$.

Choose f_t to exploit sparsity.

A heuristic analysis

Expanding the i^{th} entry of $\mathbf{x}^{t+1}$:

$$x_i^{t+1} = \underbrace{\left(\sqrt{\lambda}\frac{\langle \mathbf{x}, F_t(\mathbf{x}^t)\rangle}{n}\right)}_{\approx \mu_t} x_i + \underbrace{\frac{1}{\sqrt{n}}\sum_j Z_{ij} f_t(x_j^t)}_{\approx \text{Normal}(0,\tau_t)}$$

A heuristic analysis

Expanding the i^{th} entry of $\mathbf{x}^{t+1}$:

$$x_i^{t+1} = \underbrace{\left(\sqrt{\lambda}\frac{\langle \mathbf{x}, F_t(\mathbf{x}^t)\rangle}{n}\right)}_{\approx \mu_t} x_i + \underbrace{\frac{1}{\sqrt{n}}\sum_j Z_{ij} f_t(x_j^t)}_{\approx \text{Normal}(0,\tau_t)}$$

Thus:

$$\mathbf{x}^{t+1} \overset{d}{\approx} \mu_t \mathbf{x} + \sqrt{\tau_t}\mathbf{z}, \text{ where } \mathbf{z} \sim \text{Normal}(0, \mathrm{I}_n)$$

Approximate Message Passing (AMP)

This analysis is obviously wrong, but. . .

Approximate Message Passing (AMP)

This analysis is obviously wrong, but...

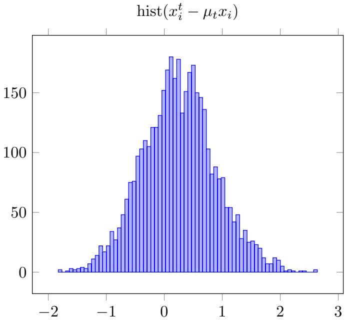
is asymptotically exact for the modified iteration:

$$\begin{aligned}\mathbf{x}^{t+1} &= \mathbf{A}\widehat{\mathbf{x}}^t - b_t \widehat{\mathbf{x}}^{t-1},\\ \widehat{\mathbf{x}}^t &= F_t(\mathbf{x}^t).\end{aligned}$$

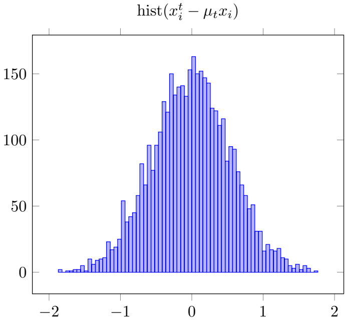
[Donoho, Maleki, Montanari 2009], [Bayati, Montanari 2011], [Rangan, Fletcher 2012].

Asymptotic behavior

$t = 2$



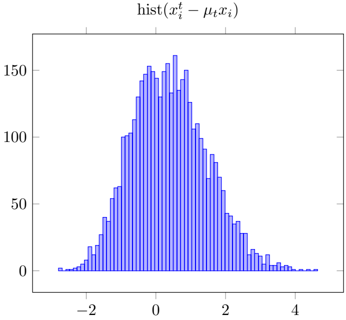
Power method



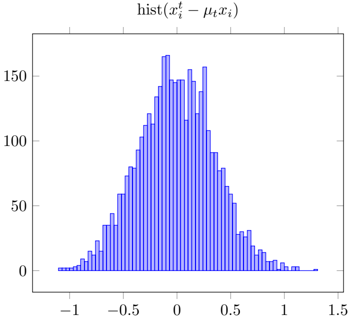
AMP

Asymptotic behavior

$t = 4$



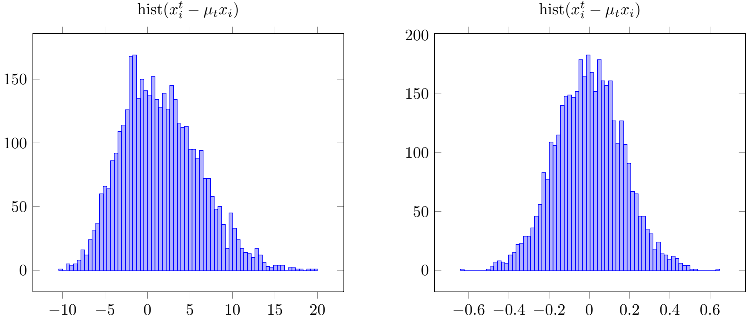
Power method



AMP

Asymptotic behavior

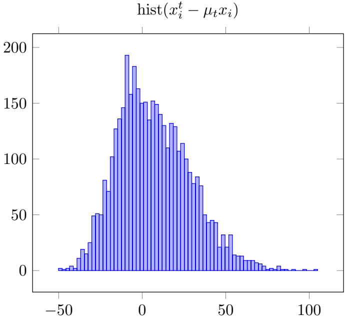
$t = 8$

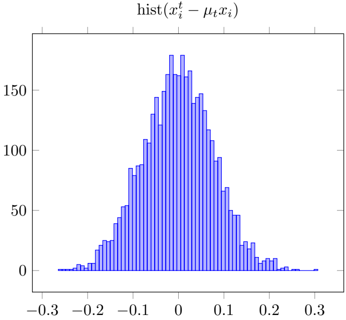


Power method AMP

Asymptotic behavior

$t = 12$



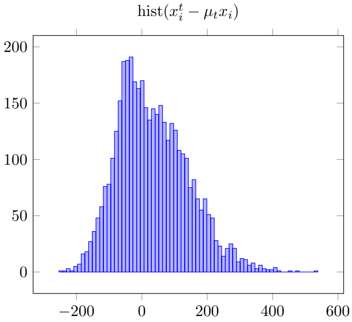


Power method

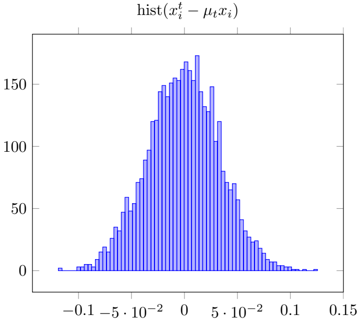
AMP

Asymptotic behavior

$t = 16$



Power method



AMP

Asymptotic behavior: a lemma

Lemma

Let f_t be a sequence of Lipschitz functions. For every fixed t and uniformly random i:

$$(x_i, x_i^t) \quad \stackrel{\mathrm{d}}{\to} \quad (X_0, \mu_t X_0 + \sqrt{\tau_t} Z) \text{ almost surely.}$$

State evolution

Deterministic recursions:

$$\mu_{t+1} = \mathbb{E}\{\sqrt{\lambda} f_t(\mu_t X_0 + \sqrt{\tau_t} Z)\}$$
$$\tau_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sqrt{\tau_t} Z)^2\}.$$

State evolution

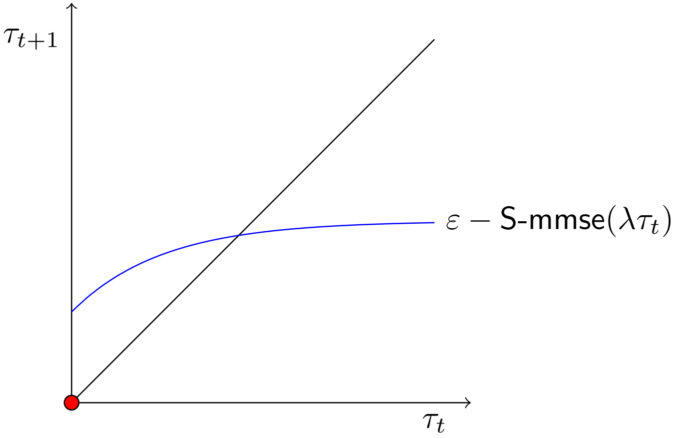
Deterministic recursions:

$$\mu_{t+1} = \mathbb{E}\{\sqrt{\lambda} f_t(\mu_t X_0 + \sqrt{\tau_t} Z)\}$$
$$\tau_{t+1} = \mathbb{E}\{f_t(\mu_t X_0 + \sqrt{\tau_t} Z)^2\}.$$

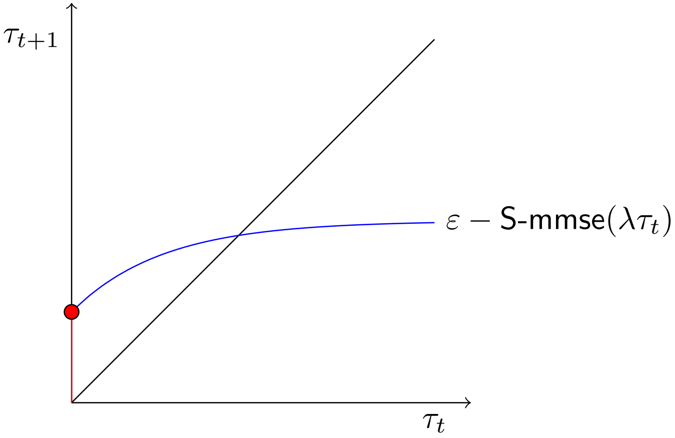
With optimal f_t:

$$\mu_{t+1} = \sqrt{\lambda}\tau_{t+1}$$
$$\tau_{t+1} = \varepsilon - \text{S-mmse}(\lambda\tau_t).$$

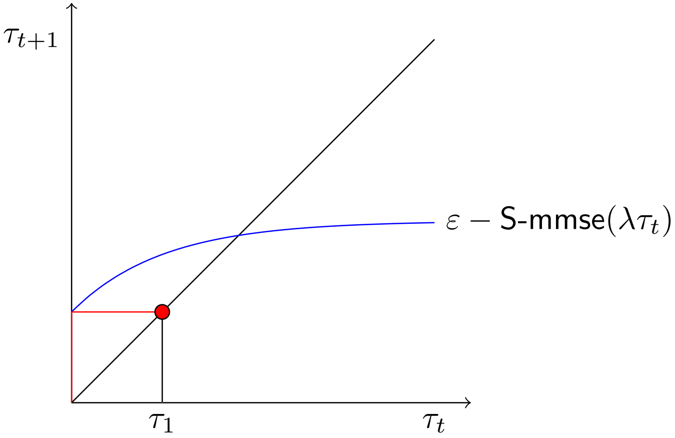
State evolution: an illustration



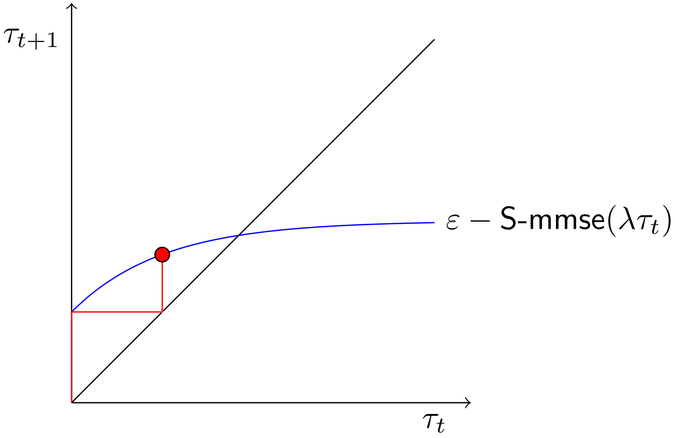
State evolution: an illustration



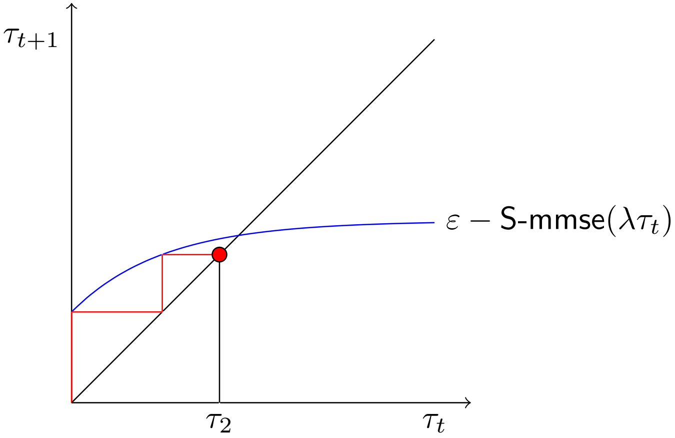
State evolution: an illustration



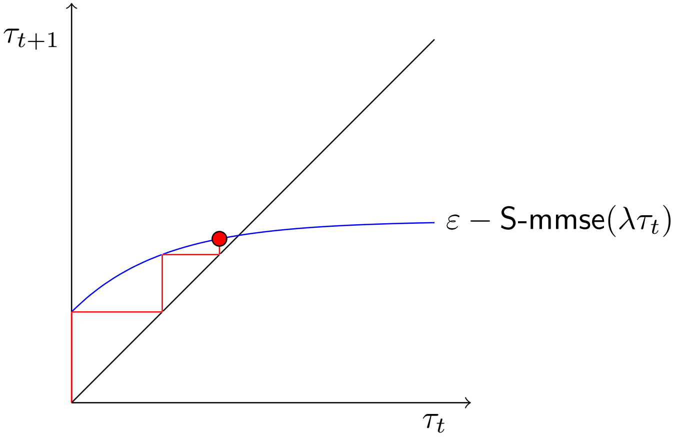
State evolution: an illustration



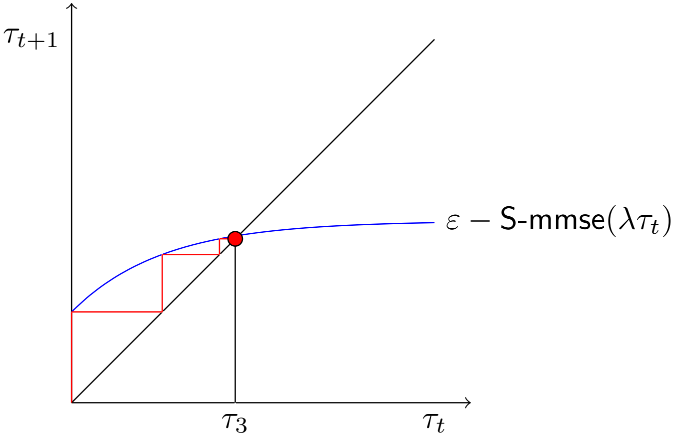
State evolution: an illustration



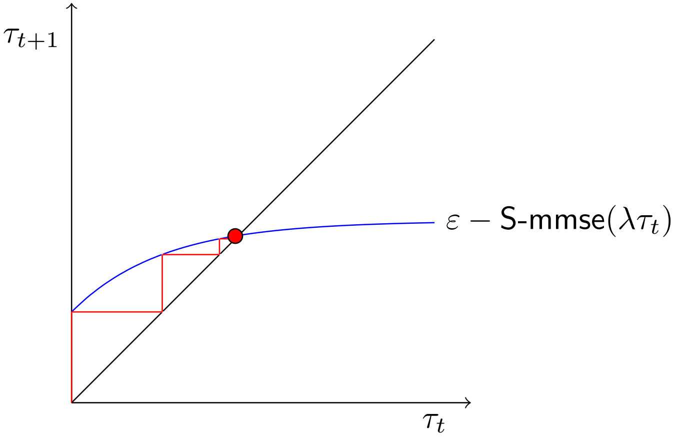
State evolution: an illustration



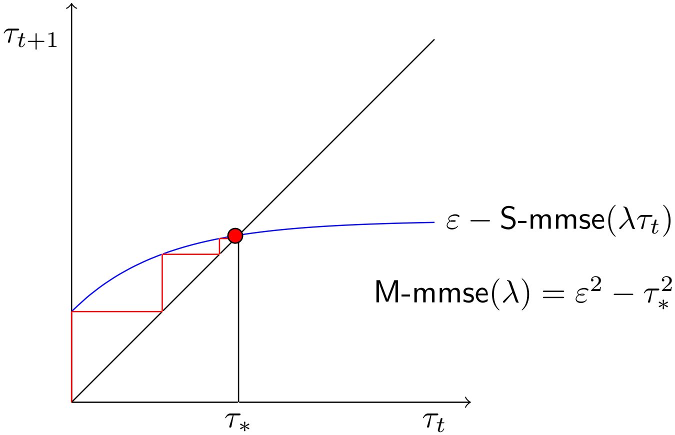
State evolution: an illustration



State evolution: an illustration



State evolution: an illustration



Proof sketch: MSE expression

Using estimator $\widehat{\mathbf{X}}^t = \widehat{\mathbf{x}}^t(\widehat{\mathbf{x}}^t)^{\mathsf{T}}$:

$$\begin{aligned}
\mathrm{mse}(\widehat{\mathbf{X}}^t, \lambda) &= \frac{1}{n^2}\mathbb{E}\{\|\widehat{\mathbf{x}}(\widehat{\mathbf{x}}^t)^{\mathsf{T}} - \mathbf{x}\mathbf{x}^{\mathsf{T}}\|_F^2\} \\
&= \frac{1}{n^2}\mathbb{E}\{\|\mathbf{x}\|^4\} + \frac{1}{n^2}\mathbb{E}\{\|\widehat{\mathbf{x}}\|^4\} - 2\mathbb{E}\left\{\frac{\langle \widehat{\mathbf{x}}^t, \mathbf{x}\rangle^2}{n^2}\right\} \\
&\to \varepsilon^2 - \tau_{t+1}^2.
\end{aligned}$$

Proof sketch: MSE expression

Using estimator $\widehat{\mathbf{X}}^t = \widehat{\mathbf{x}}^t(\widehat{\mathbf{x}}^t)^{\mathsf{T}}$:

$$\begin{aligned}
\mathrm{mse}(\widehat{\mathbf{X}}^t, \lambda) &= \frac{1}{n^2}\mathbb{E}\{\|\widehat{\mathbf{x}}(\widehat{\mathbf{x}}^t)^{\mathsf{T}} - \mathbf{x}\mathbf{x}^{\mathsf{T}}\|_F^2\} \\
&= \frac{1}{n^2}\mathbb{E}\{\|\mathbf{x}\|^4\} + \frac{1}{n^2}\mathbb{E}\{\|\widehat{\mathbf{x}}\|^4\} - 2\mathbb{E}\left\{\frac{\langle\widehat{\mathbf{x}}^t, \mathbf{x}\rangle^2}{n^2}\right\} \\
&\to \varepsilon^2 - \tau_{t+1}^2.
\end{aligned}$$

Thus

$$\mathrm{mse}_{\mathrm{AMP}}(\lambda) = \lim_{t\to\infty}\lim_{n\to\infty} \mathrm{mse}(\widehat{\mathbf{X}}^t, \lambda) = \varepsilon^2 - \tau_*^2.$$

Proof sketch: I-MMSE identity

$$\text{M-mmse}(\lambda) \quad \leq \quad \text{mse}_{\text{AMP}}(\lambda)$$

Proof sketch: I-MMSE identity

$$\frac{1}{4}\int_0^\infty \text{M-mmse}(\lambda)\mathrm{d}\lambda \qquad \leq \qquad \frac{1}{4}\int_0^\infty \mathsf{mse}_{\mathsf{AMP}}(\lambda)\mathrm{d}\lambda$$

Proof sketch: I-MMSE identity

$$\frac{1}{4}\int_0^\infty \text{M-mmse}(\lambda)\mathrm{d}\lambda \qquad \leq \qquad \frac{1}{4}\int_0^\infty \text{mse}_{\text{AMP}}(\lambda)\mathrm{d}\lambda$$

$$\big\downarrow$$

$$I(\mathbf{X};\mathbf{Y}_\infty) - I(\mathbf{X};\mathbf{Y}_0)$$

Proof sketch: I-MMSE identity

$$\frac{1}{4}\int_0^\infty \text{M-mmse}(\lambda)\mathrm{d}\lambda \qquad \leq \qquad \frac{1}{4}\int_0^\infty \mathsf{mse}_{\mathsf{AMP}}(\lambda)\mathrm{d}\lambda$$

$$\downarrow$$

$$I(\mathbf{X};\mathbf{Y}_\infty) - I(\mathbf{X};\mathbf{Y}_0)$$

$$= h(\varepsilon)$$

Proof sketch: I-MMSE identity

$$\begin{array}{ccc} \frac{1}{4}\int_0^\infty \text{M-mmse}(\lambda)\mathrm{d}\lambda & \leq & \frac{1}{4}\int_0^\infty \text{mse}_{\text{AMP}}(\lambda)\mathrm{d}\lambda \\ \downarrow & & \downarrow \\ I(\mathbf{X};\mathbf{Y}_\infty) - I(\mathbf{X};\mathbf{Y}_0) & & \frac{1}{4}\int_0^\infty (\varepsilon^2 - \tau_*(\varepsilon,\lambda)^2)\mathrm{d}\lambda \\ = h(\varepsilon) & & \end{array}$$

Proof sketch: I-MMSE identity

$$\frac{1}{4}\int_0^\infty \text{M-mmse}(\lambda)\mathrm{d}\lambda \quad \leq \quad \frac{1}{4}\int_0^\infty \text{mse}_{\text{AMP}}(\lambda)\mathrm{d}\lambda$$

$$\big\downarrow \qquad\qquad\qquad\qquad \big\downarrow$$

$$I(\mathbf{X};\mathbf{Y}_\infty) - I(\mathbf{X};\mathbf{Y}_0) \qquad\qquad \frac{1}{4}\int_0^\infty (\varepsilon^2 - \tau_*(\varepsilon,\lambda)^2)\mathrm{d}\lambda$$

$$= h(\varepsilon) \qquad\qquad = h(\varepsilon)$$

Conclusion

Some open problems. . .

- MMSE characterization with multiple fixed points
- General distributions for x

Conclusion

Some open problems. . .

- MMSE characterization with multiple fixed points
- General distributions for x

Thanks!