Understanding Phase Transitions via Mutual Information and MMSE

Galen Reeves

Duke University, Durham NC, USA

ISIT Tutorial — July 7, 2019

Table of Contents

Bayes Optimal Inference

Distribution $p(\boldsymbol{x}, \boldsymbol{y})$ on random vectors (matrices, tensors)

$$\begin{aligned} \text{unknown:} \qquad & \boldsymbol{X} = (X_1, \dots, X_n) \\ \text{observed data:} \qquad & \boldsymbol{Y} = (Y_1, \dots, Y_m) \end{aligned}$$

Posterior distribution of unknowns:

$$p(\boldsymbol{x} \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, \boldsymbol{x}) p(\boldsymbol{x})}{p(\boldsymbol{y})}.$$

Basic questions:

1. What are the fundamental limits of inference?
2. Can we achieve these limits using efficient methods?

Marginals and Moments of Posterior Distribution

- Marginal of (random) posterior on set of indices $S \subset [n]$,

$$p(\boldsymbol{x}_S \mid \boldsymbol{Y}) = \int p(\boldsymbol{x} \mid \boldsymbol{Y}) \, \mathrm{d}\boldsymbol{x}_{S^c}$$

- Mean and covariance of (random) posterior distribution

$$\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}] = \int \boldsymbol{x} \, p(\boldsymbol{x} \mid \boldsymbol{Y}) \, \mathrm{d}\boldsymbol{x}$$

$$\mathsf{Cov}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathbb{E}\big[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}])^T \mid \boldsymbol{Y}\big]$$

Marginals and Moments of Posterior Distribution

- Marginal of (random) posterior on set of indices $S \subset [n]$,

$$p(\boldsymbol{x}_S \mid \boldsymbol{Y}) = \int p(\boldsymbol{x} \mid \boldsymbol{Y}) \, \mathrm{d}\boldsymbol{x}_{S^c}$$

- Mean and covariance of (random) posterior distribution

$$\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}] = \int \boldsymbol{x} \, p(\boldsymbol{x} \mid \boldsymbol{Y}) \, \mathrm{d}\boldsymbol{x}$$

$$\mathsf{Cov}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathbb{E}\big[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}])^T \mid \boldsymbol{Y}\big]$$

- Minimum mean-squared error (MMSE)

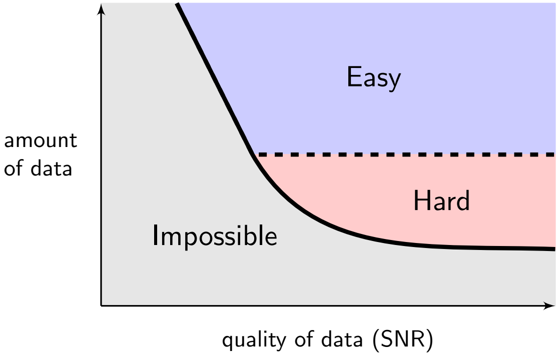
$$\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathbb{E}\Big[\|\boldsymbol{X} - \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}]\|^2\Big] = \mathbb{E}[\mathrm{tr}(\mathsf{Cov}(\boldsymbol{X} \mid \boldsymbol{Y}))]$$

  Also known as Bayes risk under squared-error loss

Approaches to Analysis

- **Information Theory**
  - Single-letter formulas describe fundamental limits
  - Models with structured randomness: e.g., Markov sources (compression) and finite-state channels (communication)
- **Replica Method from Statistical Physics**
  - Conjectured single-letter formulas via heuristics
  - Models with disordered randomness: e.g., spin glasses and modern high-dimensional inference problems
- **Approximate Inference**
  - Fixed-points of message passing algorithms related stationary points of potential functions.

Phase Diagram for High-Dimensional Inference



Easy – Computationally efficient methods succeed.

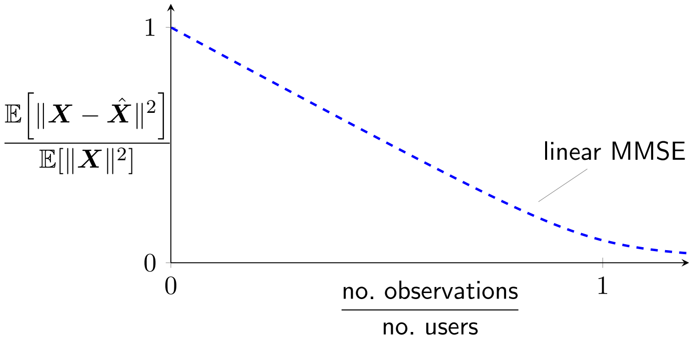
Hard – All known efficient methods fail but intractable methods succeed.

Impossible – All methods fail, regardless of complexity.

Example: Multiuser Detection in CDMA

$$\boldsymbol{Y} = \sum_j \boldsymbol{A}_j X_j + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 I)$$

- $\{\boldsymbol{A}_j\}$ are known signature vectors with IID entries
- $\{X_j\}$ are IID $\{+1, -1\}$ messages

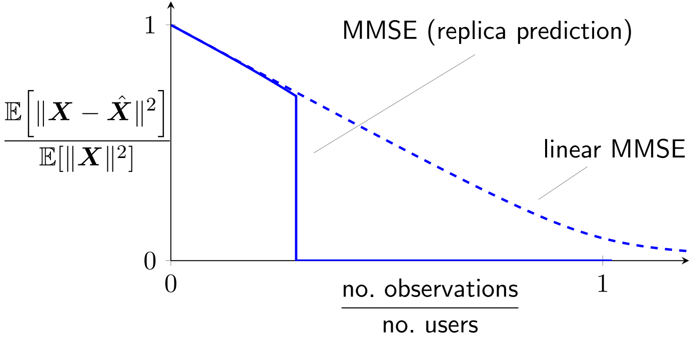


Tse,Hanly'99, Verdu,Shamai'99,

Example: Multiuser Detection in CDMA

$$\boldsymbol{Y} = \sum_j \boldsymbol{A}_j X_j + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 I)$$

- $\{\boldsymbol{A}_j\}$ are known signature vectors with IID entries
- $\{X_j\}$ are IID $\{+1, -1\}$ messages

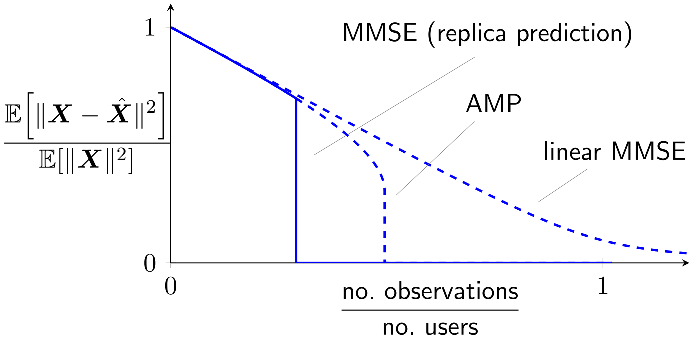


Tse,Hanly'99, Verdu,Shamai'99, Tanaka'02, Guo,Verdu'05,

Example: Multiuser Detection in CDMA

$$\boldsymbol{Y} = \sum_j \boldsymbol{A}_j X_j + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 I)$$

- $\{\boldsymbol{A}_j\}$ are known signature vectors with IID entries
- $\{X_j\}$ are IID $\{+1, -1\}$ messages

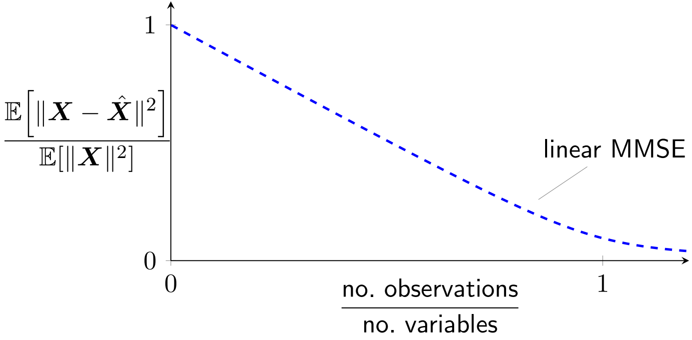


Tse,Hanly'99, Verdu,Shamai'99, Tanaka'02, Guo,Verdu'05, Bayati,Montanari'11

Example: Compressed Sensing

$$\boldsymbol{Y} = \sum_j \boldsymbol{A}_j X_j + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 I)$$

- $\{\boldsymbol{A}_j\}$ are columns of known sensing matrix with IID entries
- $\{X_j\}$ are IID Bernoulli-Gaussian: $p(x) = (1-\epsilon)\delta_0(x) + \epsilon\mathcal{N}(x \mid 0, \epsilon^{-1})$

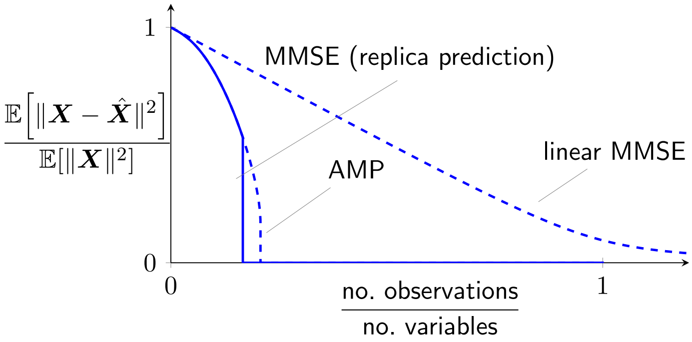


Donoho,Maleki,Montanari'09, Bayati,Montanari'11, R,Gastpar'12

Example: Compressed Sensing

$$\boldsymbol{Y} = \sum_j \boldsymbol{A}_j X_j + \boldsymbol{W}, \qquad \boldsymbol{W} \sim \mathcal{N}(0, \sigma^2 I)$$

- $\{\boldsymbol{A}_j\}$ are columns of known sensing matrix with IID entries
- $\{X_j\}$ are IID Bernoulli-Gaussian: $p(x) = (1-\epsilon)\delta_0(x) + \epsilon\mathcal{N}(x \mid 0, \epsilon^{-1})$



Donoho,Maleki,Montanari'09, Bayati,Montanari'11, R,Gastpar'12

Motivating Questions

- Where do these formulas come from?
- When is AMP optimal?
- Are the replica predictions correct?
- What about other models?

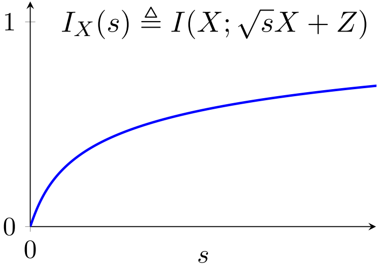
Signal-Plus Noise Model (Scalar Case)

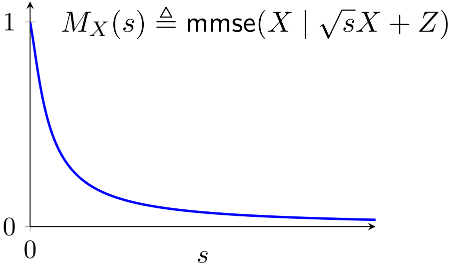
$$Y = \sqrt{s}X + Z, \qquad X \sim p(x), \qquad Z \sim \mathcal{N}(0,1)$$

Guo,Shamai,Verdu'05, R,Pfister,Dytso'18

Signal-Plus Noise Model (Scalar Case)

$$Y = \sqrt{s}X + Z, \qquad X \sim p(x), \qquad Z \sim \mathcal{N}(0,1)$$

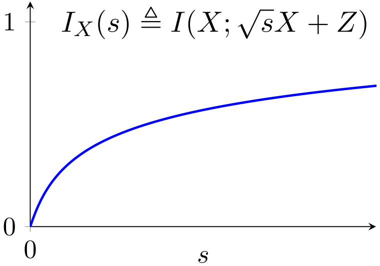


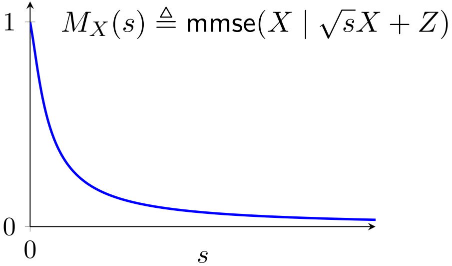


Guo,Shamai,Verdu’05, R,Pfister,Dytso’18

Signal-Plus Noise Model (Scalar Case)

$$Y = \sqrt{s}X + Z, \qquad X \sim p(x), \qquad Z \sim \mathcal{N}(0,1)$$

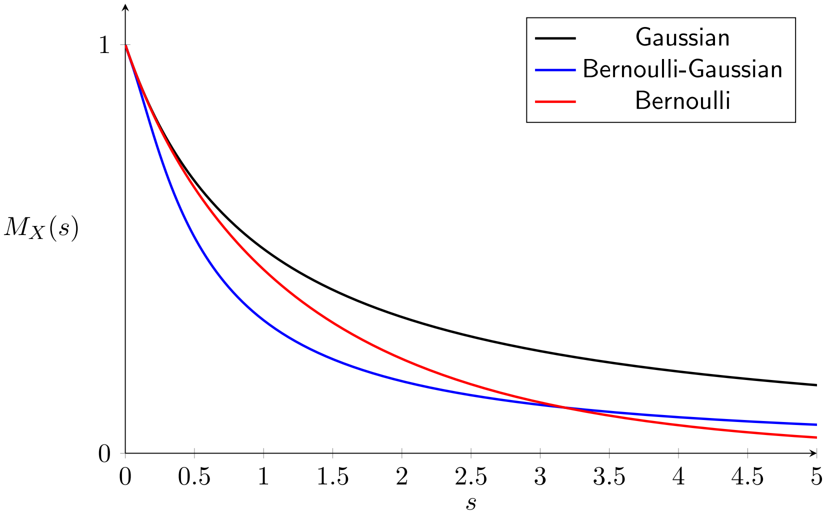




I-MMSE Relation: $\frac{\mathrm{d}}{\mathrm{d}s} I_X(s) = \frac{1}{2} M_X(s)$

Guo,Shamai,Verdu'05, R,Pfister,Dytso'18

Effect of Signal Prior on MMSE Function



Effect of Signal Prior on MMSE Function (Log Scale)

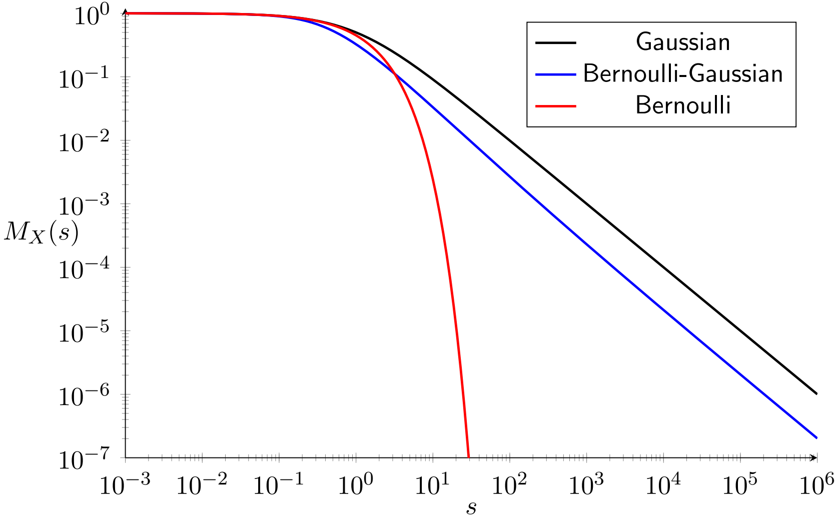


Table of Contents

Experiment: Linear Regression with IID Gaussian matrix

- Length $n = 10^4$ vector $\boldsymbol{X}$ with IID Bernoulli-Gaussian entries

$$X_i \sim p(x) = 0.8\,\delta_0(x) + 0.2\,\mathcal{N}(x \mid 0, 5)$$

- Observations $\boldsymbol{Y} \in \mathbb{R}^m, \boldsymbol{A} \in \mathbb{R}^{m \times n}$ with SNR $=$ 60 dB

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{X}, 10^{-6} I_m), \qquad A_{ij} \sim \mathcal{N}(0, 1/n)$$

Donoho,Maleki,Montanari'09, Matlab implementation by Schniter et. al.

Experiment: Linear Regression with IID Gaussian matrix

- Length $n = 10^4$ vector $\boldsymbol{X}$ with IID Bernoulli-Gaussian entries

$$X_i \sim p(x) = 0.8\,\delta_0(x) + 0.2\,\mathcal{N}(x \mid 0, 5)$$

- Observations $\boldsymbol{Y} \in \mathbb{R}^m, \boldsymbol{A} \in \mathbb{R}^{m \times n}$ with SNR $=$ 60 dB

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{X}, 10^{-6} I_m), \qquad A_{ij} \sim \mathcal{N}(0, 1/n)$$

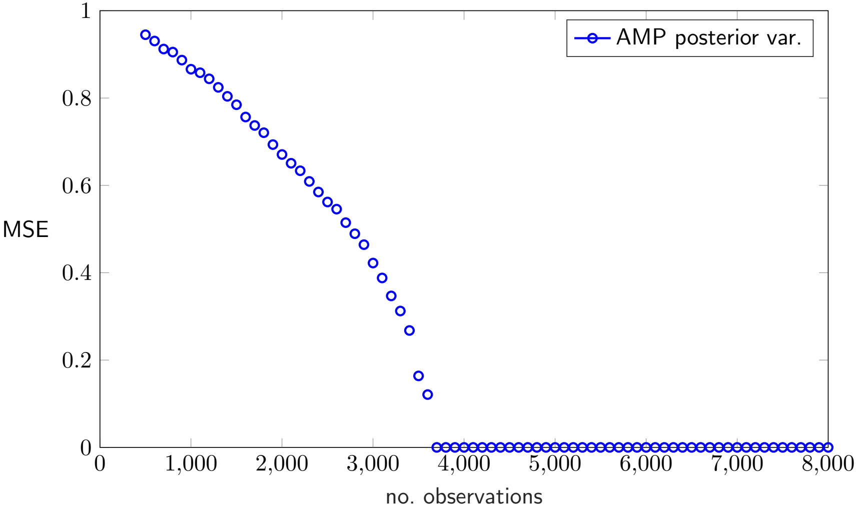
- Use AMP to approximate marginals of posterior: $\hat{p}(x_i \mid \boldsymbol{Y}, \boldsymbol{A})$

AMP squared error: $$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}_{\hat{p}}[X_i \mid \boldsymbol{Y}, \boldsymbol{A}])^2$$

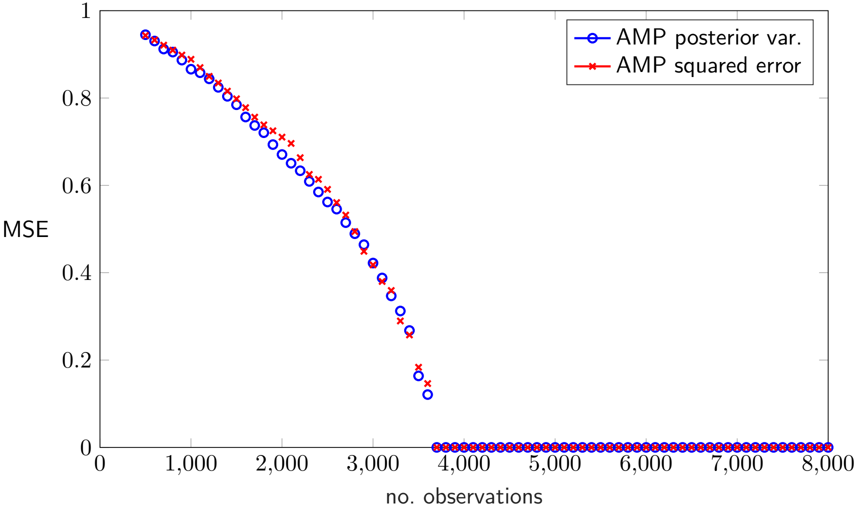
AMP posterior variance: $$\frac{1}{n}\sum_{i=1}^{n}\mathsf{Var}_{\hat{p}}(X_i i \mid \boldsymbol{Y}, \boldsymbol{A}).$$

Donoho,Maleki,Montanari'09, Matlab implementation by Schniter et. al.

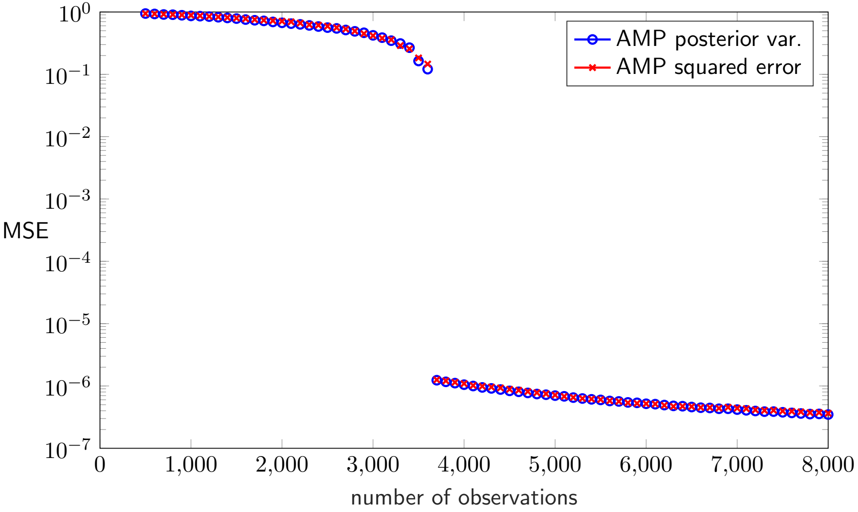
Experiment: Linear Regression with IID Gaussian matrix



Experiment: Linear Regression with IID Gaussian matrix



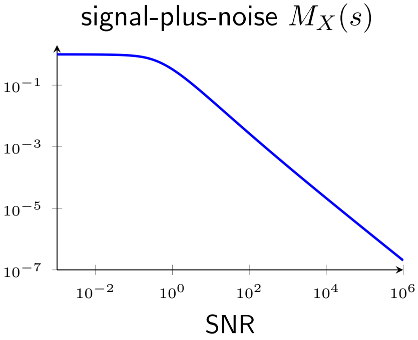
Experiment: Linear Regression with IID Gaussian matrix



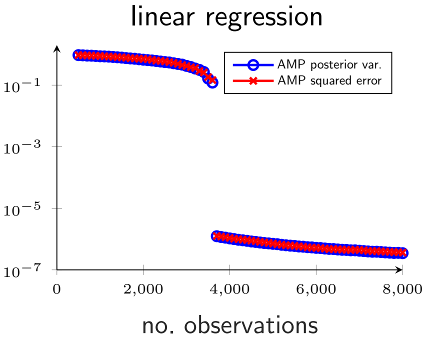
Motivating Questions

- Why is there a big change in behavior around 3,500 observations?
- Do AMP estimates correspond to true posterior distribution?

MSE of AMP

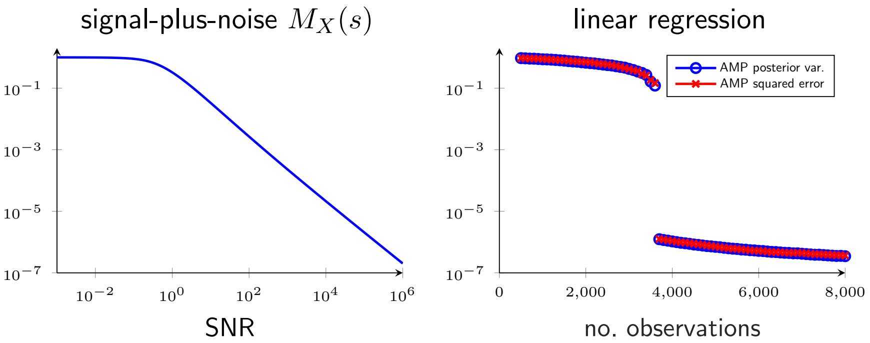


signal-plus-noise $M_X(s)$
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}
SNR



linear regression
AMP posterior var.
AMP squared error
10^{-1}
10^{-3}
10^{-5}
10^{-7}
0
2,000
4,000
6,000
8,000
no. observations

MSE of AMP



Define asymptotic MSE

$$\mathcal{E}_{\mathsf{AMP}}(\delta) \triangleq \lim_{m,n\to\infty} \frac{1}{n}\mathbb{E}\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{\mathsf{AMP}}\right\|^2\right], \qquad m/n \to \delta$$

State Evolution for AMP

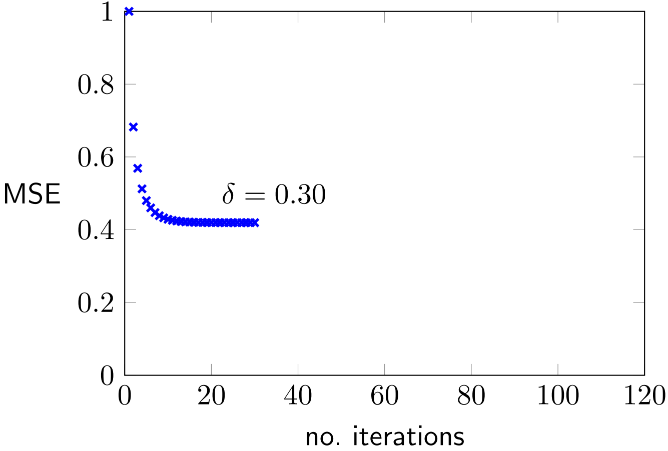
Recall rom Phil's talk, the AMP per-iteration MSE $\mathcal{E}^t$ satisfies

$$\mathcal{E}^{t+1} = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}^t}\right), \qquad \delta = m/n$$

State Evolution for AMP

Recall rom Phil's talk, the AMP per-iteration MSE $\mathcal{E}^t$ satisfies

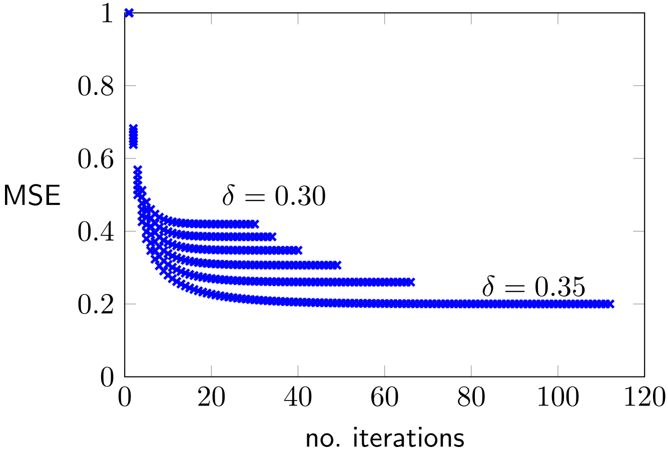
$$\mathcal{E}^{t+1} = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}^t}\right), \qquad \delta = m/n$$



State Evolution for AMP

Recall rom Phil's talk, the AMP per-iteration MSE $\mathcal{E}^t$ satisfies

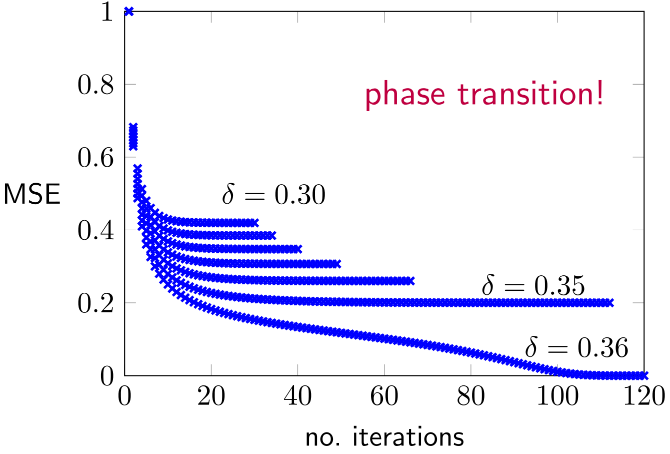
$$\mathcal{E}^{t+1} = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}^t}\right), \qquad \delta = m/n$$



State Evolution for AMP

Recall rom Phil's talk, the AMP per-iteration MSE $\mathcal{E}^t$ satisfies

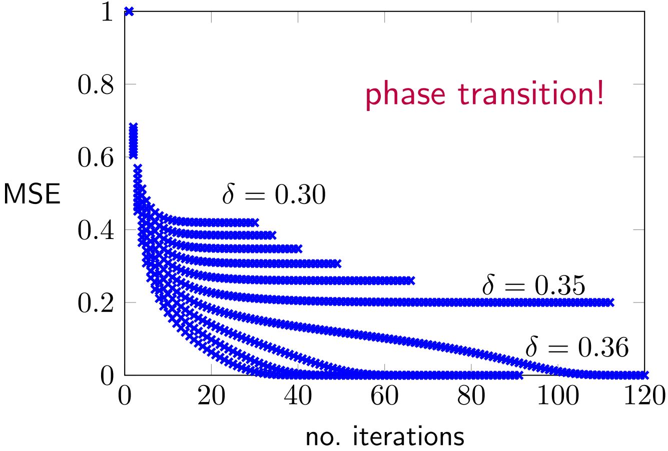
$$\mathcal{E}^{t+1} = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}^t}\right), \qquad \delta = m/n$$



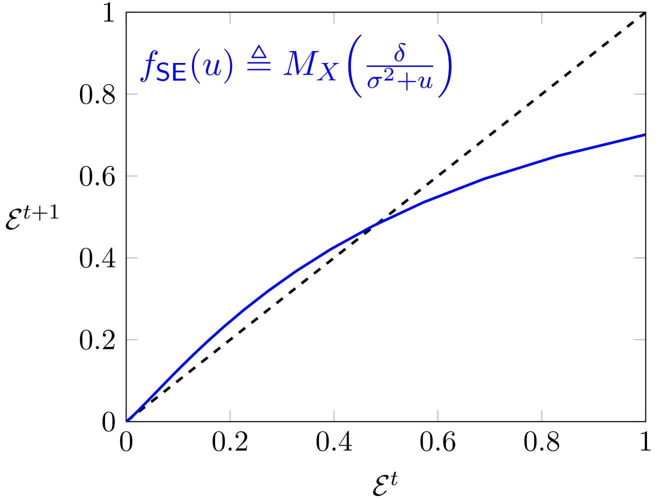
State Evolution for AMP

Recall rom Phil's talk, the AMP per-iteration MSE $\mathcal{E}^t$ satisfies

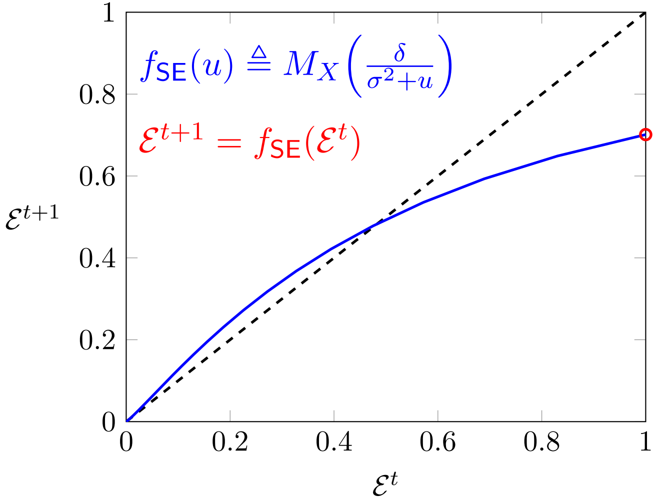
$$\mathcal{E}^{t+1} = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}^t}\right), \qquad \delta = m/n$$



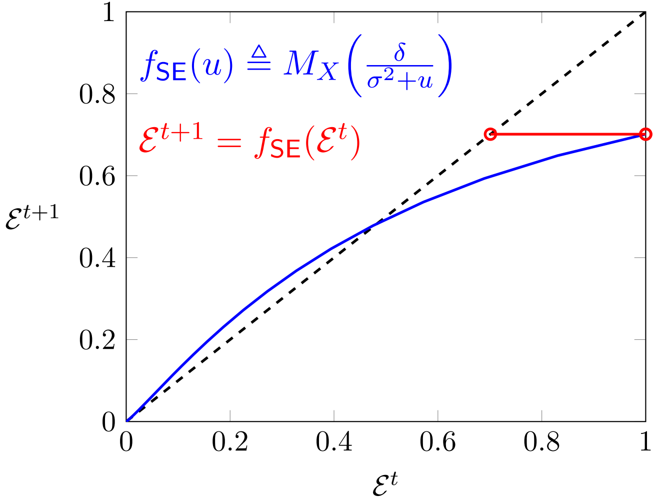
Fixed-Point Analysis of State Evolution



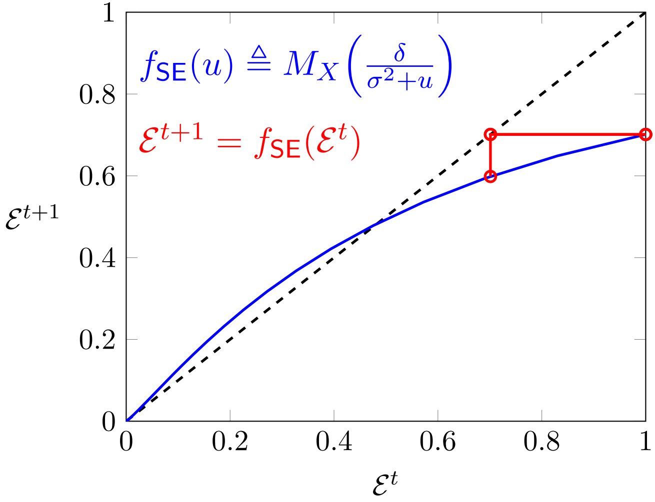
Fixed-Point Analysis of State Evolution



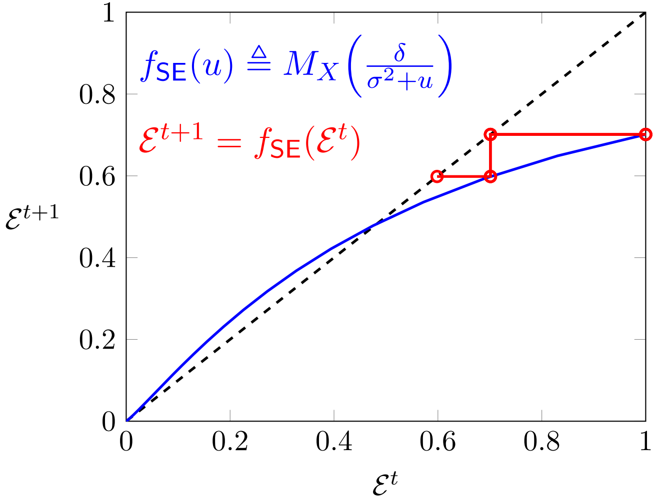
Fixed-Point Analysis of State Evolution



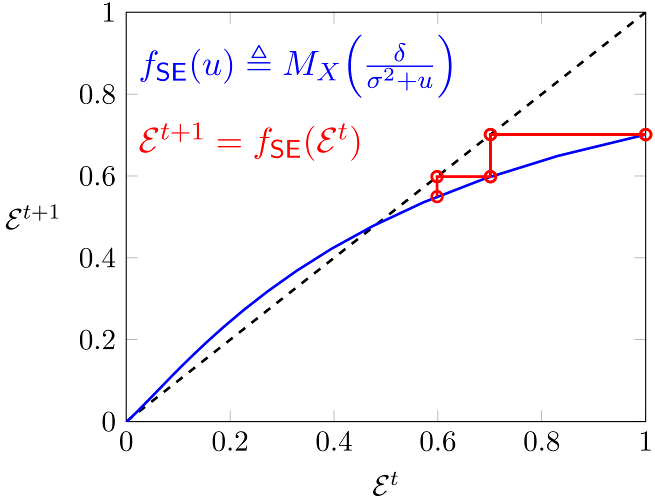
Fixed-Point Analysis of State Evolution



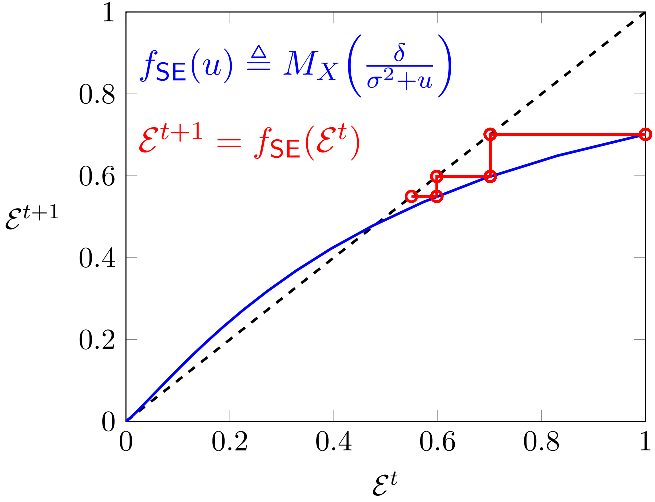
Fixed-Point Analysis of State Evolution



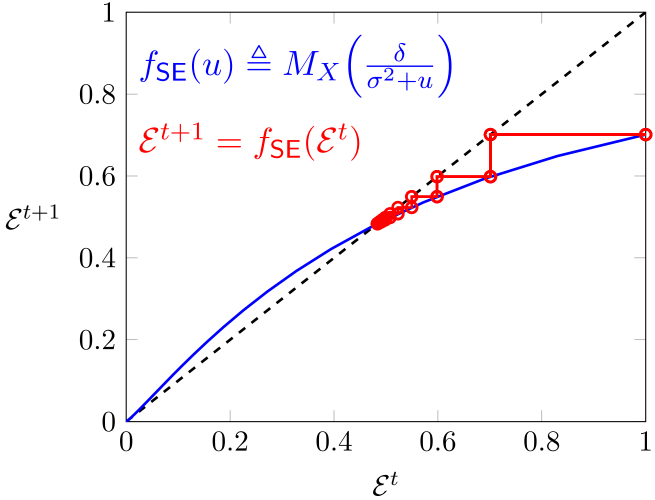
Fixed-Point Analysis of State Evolution



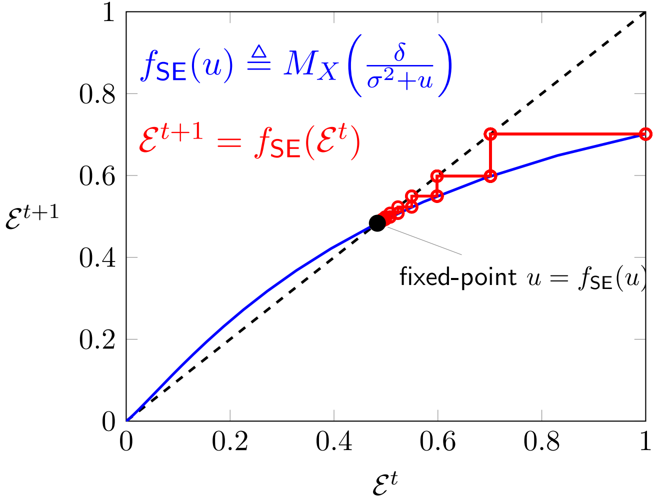
Fixed-Point Analysis of State Evolution



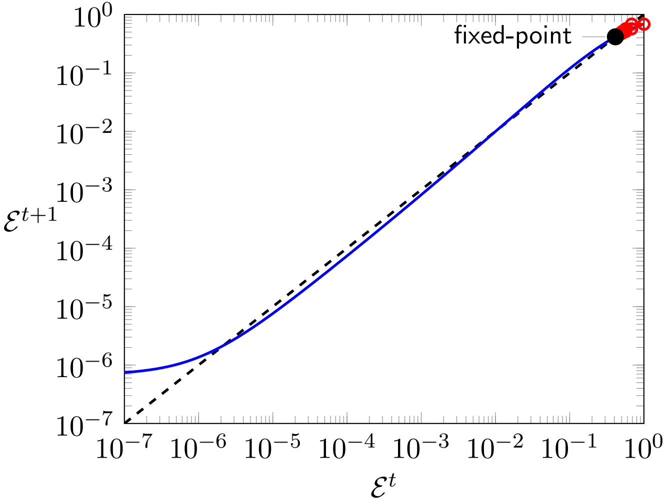
Fixed-Point Analysis of State Evolution



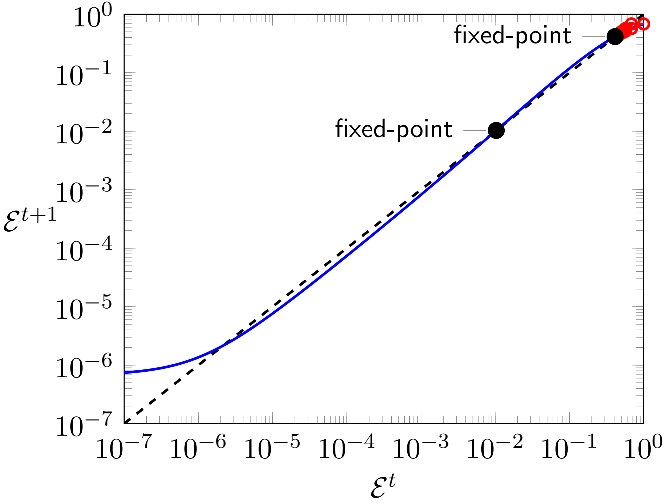
Fixed-Point Analysis of State Evolution



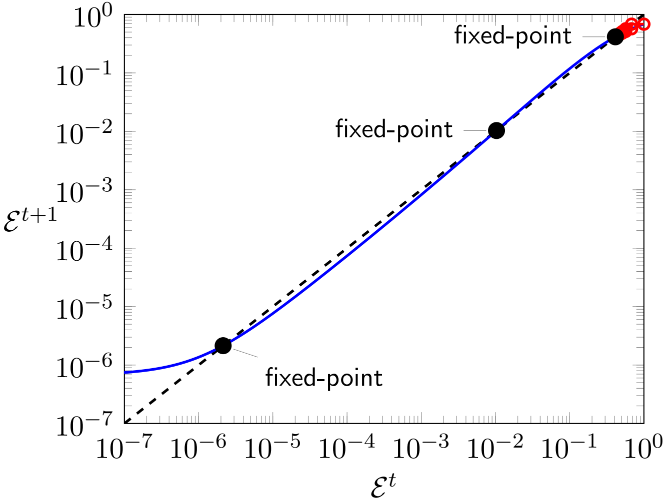
Fixed-Point Analysis of State Evolution (Log Scale)



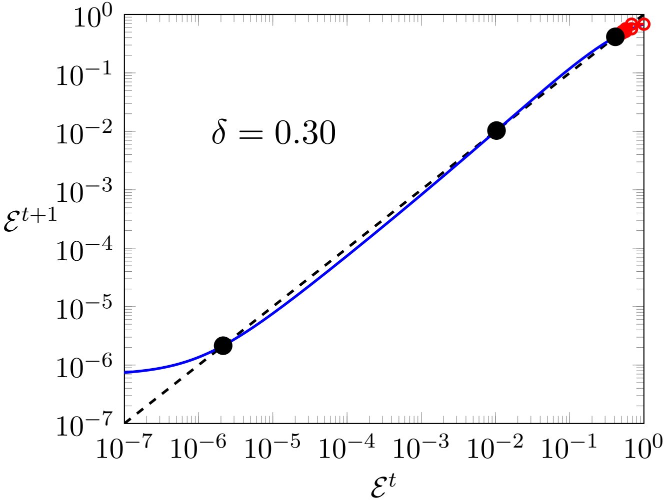
Fixed-Point Analysis of State Evolution (Log Scale)



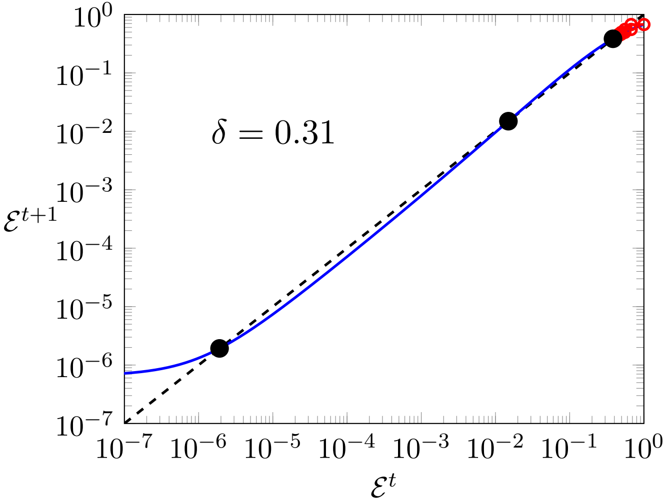
Fixed-Point Analysis of State Evolution (Log Scale)



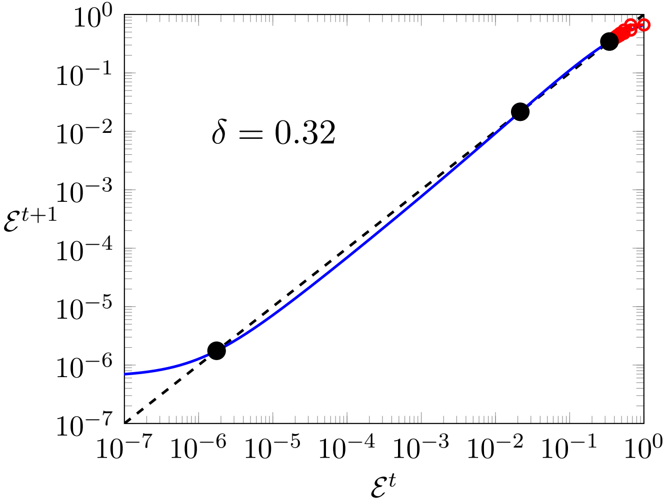
Fixed-Point Analysis of State Evolution (Log Scale)



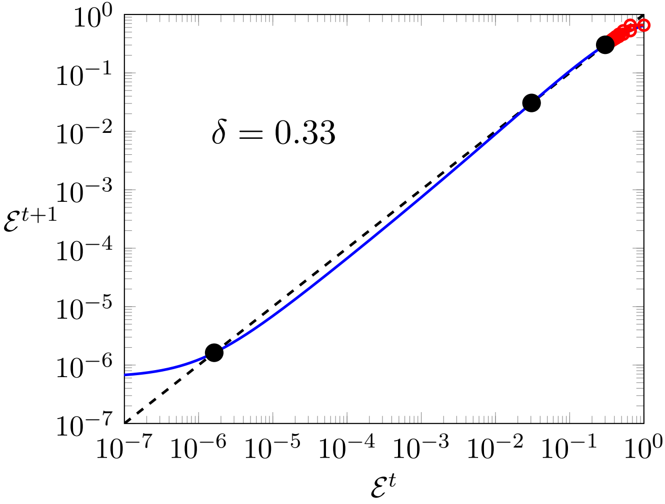
Fixed-Point Analysis of State Evolution (Log Scale)



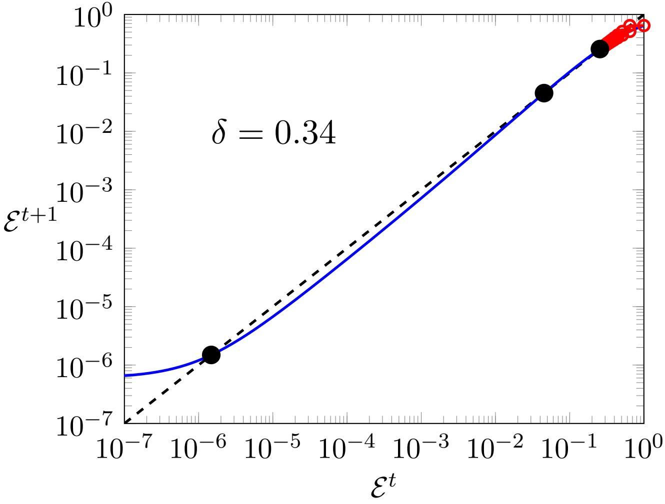
Fixed-Point Analysis of State Evolution (Log Scale)



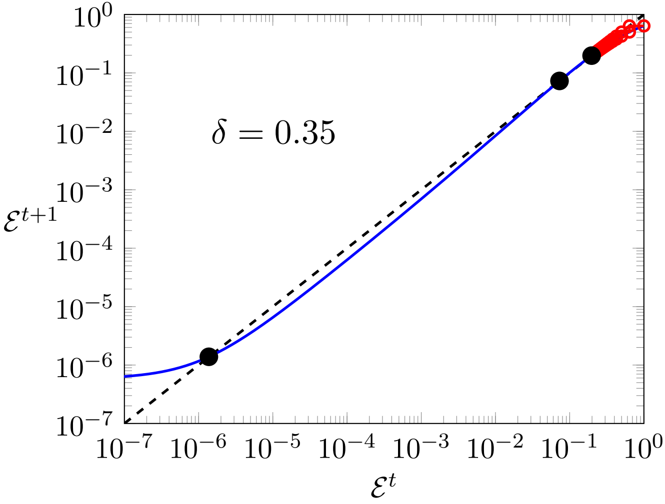
Fixed-Point Analysis of State Evolution (Log Scale)



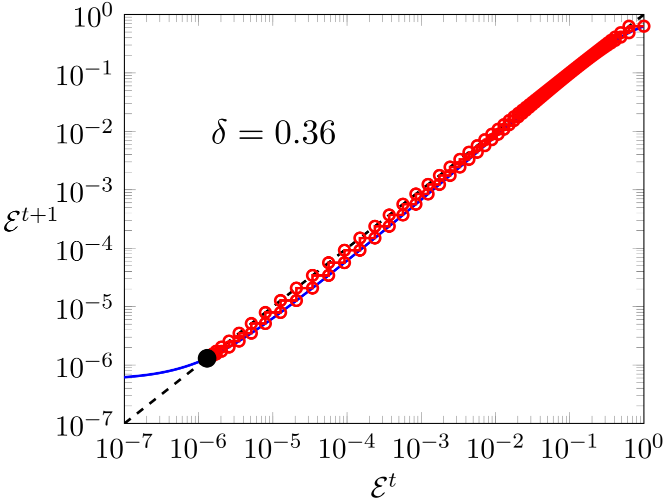
Fixed-Point Analysis of State Evolution (Log Scale)



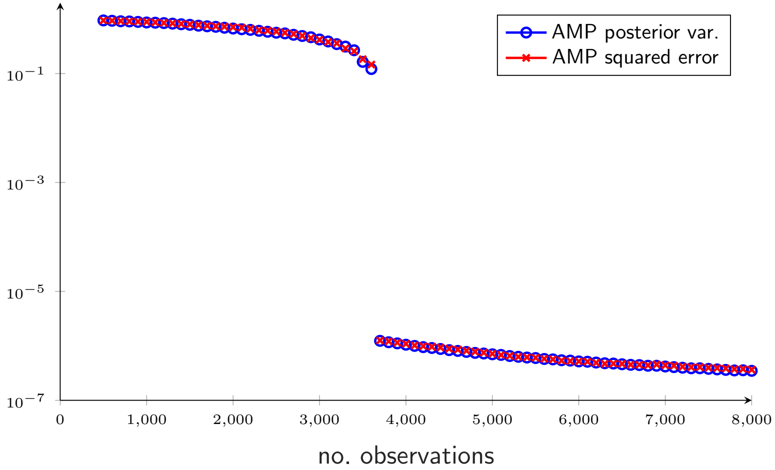
Fixed-Point Analysis of State Evolution (Log Scale)



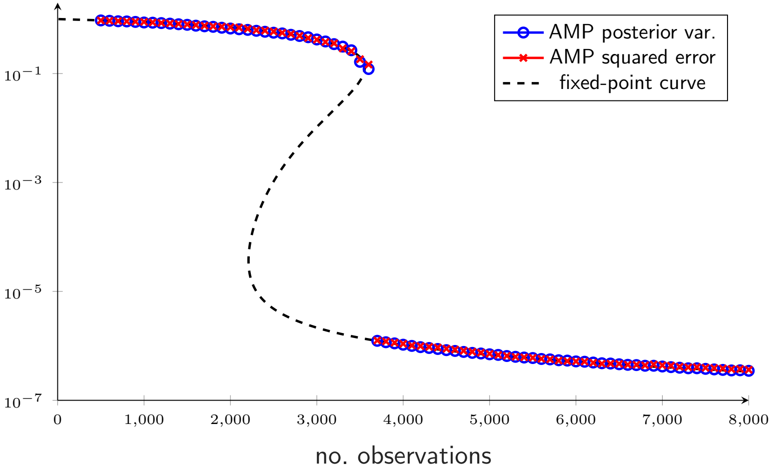
Fixed-Point Analysis of State Evolution (Log Scale)



AMP MSE is Upper Envelope of Fixed-Point Curve



AMP MSE is Upper Envelope of Fixed-Point Curve



$$\mathcal{E}_{\mathrm{AMP}}(\delta) = \max\left\{ u \,:\, u = M_X\left(\frac{\delta}{\sigma^2 + u}\right)\right\}$$

Asymptotic Mutual Information and MMSE

$$\mathcal{I}(\delta) \triangleq \lim_{\substack{m,n\to\infty \\ m/n\to\delta}} \frac{1}{n} I(\boldsymbol{X};\boldsymbol{Y}\,|\,\boldsymbol{A})$$

$$\mathcal{E}(\delta) \triangleq \lim_{\substack{m,n\to\infty \\ m/n\to\delta}} \frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X}\,|\,\boldsymbol{Y},\boldsymbol{A}).$$

Replica-Symmetric Formulas

Recall scalar mutual information function

$$I_X(s) = \frac{1}{2} \int_0^s M_X(s') \, \mathrm{d}s'$$

Replica-Symmetric Formulas

Recall scalar mutual information function

$$I_X(s) = \frac{1}{2}\int_0^s M_X(s')\,\mathrm{d}s'$$

Consider potential function:

$$\mathcal{F}_\delta(u) \triangleq I_X\left(\frac{\delta}{\sigma^2+u}\right) + \frac{\delta}{2}\left(\log\left(1+\frac{u}{\sigma^2}\right) - \frac{u}{\sigma^2+u}\right)$$

Replica-Symmetric Formulas

Recall scalar mutual information function

$$I_X(s) = \frac{1}{2}\int_0^s M_X(s')\,\mathrm{d}s'$$

Consider potential function:

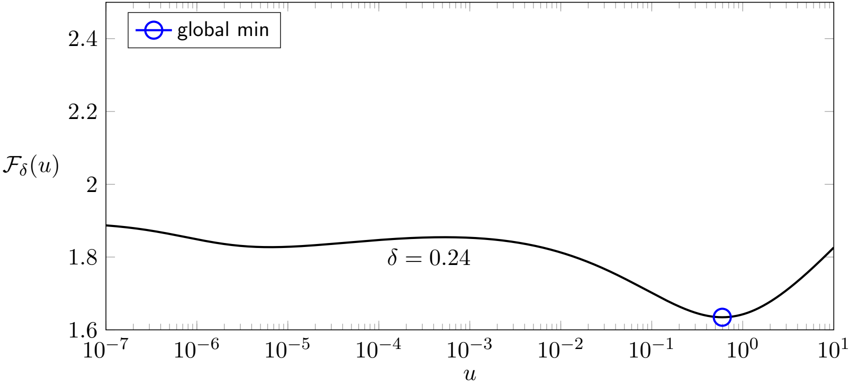
$$\mathcal{F}_\delta(u) \triangleq I_X\left(\frac{\delta}{\sigma^2+u}\right) + \frac{\delta}{2}\left(\log\left(1+\frac{u}{\sigma^2}\right) - \frac{u}{\sigma^2+u}\right)$$

RS formulas defined by global minimum

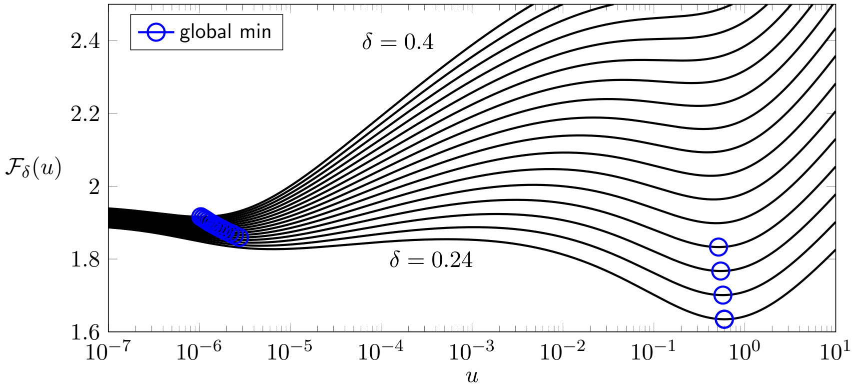
$$\mathcal{I}_{\mathsf{RS}}(\delta) \triangleq \min_u \mathcal{F}_\delta(u)$$

$$\mathcal{E}_{\mathsf{RS}}(\delta) \triangleq \arg\min_u \mathcal{F}_\delta(u)$$

RS MMSE is Global Minimizer of Potential Function



RS MMSE is Global Minimizer of Potential Function



Potential Function view of AMP MSE.

By I-MMSE relation, one finds that the stationary points of potential function satisfy fixed-point equation

$$\mathcal{F}'_\delta(u) = 0 \quad \iff \quad u = M_X\left(\frac{\delta}{\sigma^2 + u}\right)$$

Potential Function view of AMP MSE.

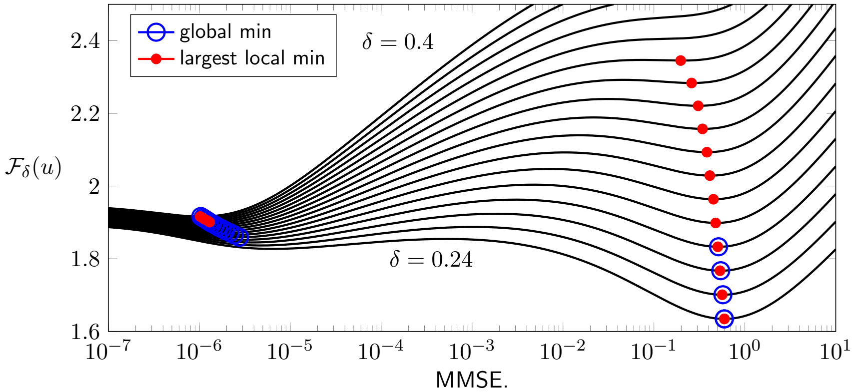
By I-MMSE relation, one finds that the stationary points of potential function satisfy fixed-point equation

$$\mathcal{F}_\delta'(u) = 0 \quad \iff \quad u = M_X\left(\frac{\delta}{\sigma^2 + u}\right)$$

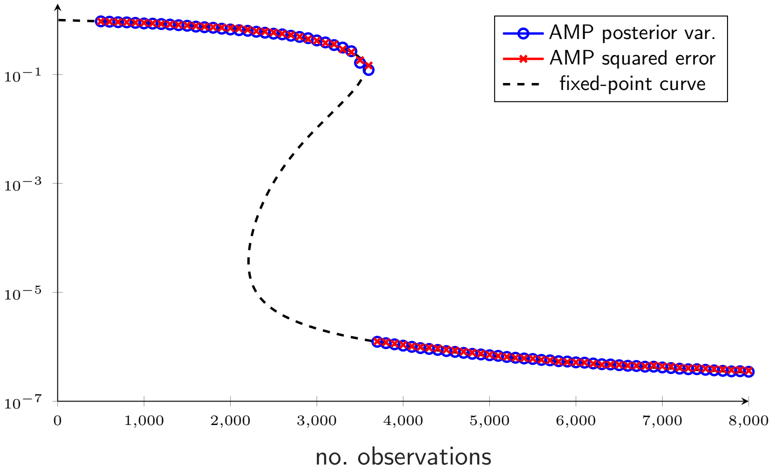
Hence, MSE of AMP is largest local minimizer:

$$\mathcal{E}_{\mathrm{AMP}}(\delta) = \max\left\{u \,:\, \mathcal{F}_\delta'(u) = 0\right\}$$

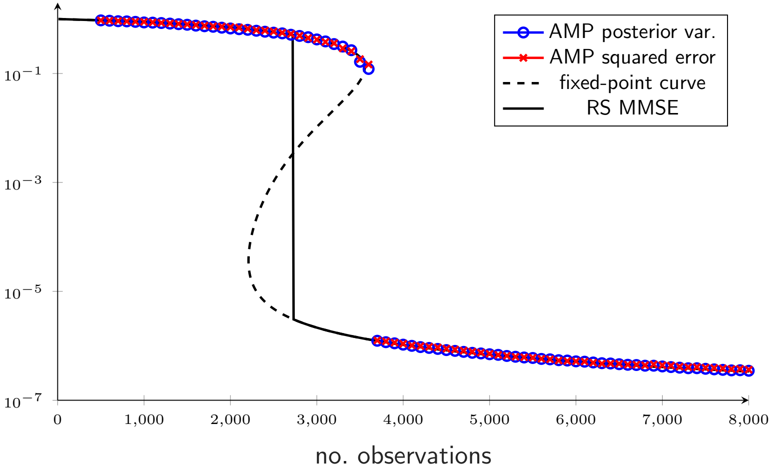
AMP MSE is Largest Local Minimizer



RS Formula for MMSE



RS Formula for MMSE



Rigorous Proofs of RS Formula

- Derived using heuristic replica method from statistical physics Tanaka 2001, Guo and Verdu 2005
- First proof by R and Pfister 2016
- Alternate proof using a different method Barbier, Dia, Macris, Krzakala, 2016

Recap of Main Points

- MSE of AMP given by largest solution to fixed-point equation

$$\mathcal{E}_{\mathrm{AMP}}(\delta) = \max\left\{u \,:\, u = M_X\left(\frac{\delta}{\sigma^2 + u}\right)\right\}$$

- MMSE is given by minimum energy solution to fixed-point equation

$$\mathcal{E}(\delta) = \arg\min_u \left\{ I_X\left(\frac{\delta}{\sigma^2 + u}\right) + \frac{\delta}{2}\left(\log\left(1 + \frac{u}{\sigma^2}\right) - \frac{u}{\sigma^2 + u}\right)\right\}$$

- Hence, AMP is optimal if and only if the largest solution is also the minimum energy solution.

- Phase transition: when the solution jump from one location to another.

Table of Contents

Proof I: New Observations are Approx. Gaussian

Follow the approach of R. and Pfister 2016.

Proof I: New Observations are Approx. Gaussian

Follow the approach of R. and Pfister 2016.

- Let $Y^i = (Y_1, \ldots, Y_i)$ and apply chain rule

$$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{A}) = \sum_{i=0}^{m-1} \underbrace{I(\boldsymbol{X}; Y_{i+1} \mid Y^i, \boldsymbol{A})}_{\text{info. from new obs.}}$$

Proof I: New Observations are Approx. Gaussian

Follow the approach of R. and Pfister 2016.

- Let $Y^i = (Y_1, \dots, Y_i)$ and apply chain rule

$$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{A}) = \sum_{i=0}^{m-1} \underbrace{I(\boldsymbol{X}; Y_{i+1} \mid Y^i, \boldsymbol{A})}_{\text{info. from new obs.}}$$

- If $p(\boldsymbol{x} \mid Y^i, \boldsymbol{A})$ is weakly correlated, then $Y_{i+1} \mid Y^i, \boldsymbol{A}$ is approx. Gaussian:

$$\begin{aligned} Y_{i+1} - \mathbb{E}\big[Y_{i+1} \mid Y^i, \boldsymbol{A}\big] &= \sum_{j=1}^{n} A_{ij}(X_j - \mathbb{E}\big[X_j \mid Y^i, \boldsymbol{A}\big]) + W_i \\ &\approx \mathcal{N}\left(0, \frac{1}{n} \operatorname{mmse}(\boldsymbol{X} \mid Y^i, \boldsymbol{A}) + \sigma^2\right) \end{aligned}$$

Proof I: New Observations are Approx. Gaussian

Follow the approach of R. and Pfister 2016.

- Let $Y^i = (Y_1, \dots, Y_i)$ and apply chain rule

$$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{A}) = \sum_{i=0}^{m-1} \underbrace{I(\boldsymbol{X}; Y_{i+1} \mid Y^i, \boldsymbol{A})}_{\text{info. from new obs.}}$$

- If $p(\boldsymbol{x} \mid Y^i, \boldsymbol{A})$ is weakly correlated, then $Y_{i+1} \mid Y^i, \boldsymbol{A}$ is approx. Gaussian:

$$\begin{aligned} Y_{i+1} - \mathbb{E}\big[Y_{i+1} \mid Y^i, \boldsymbol{A}\big] &= \sum_{j=1}^{n} A_{ij}(X_j - \mathbb{E}\big[X_j \mid Y^i, \boldsymbol{A}\big]) + W_i \\ &\approx \mathcal{N}\left(0, \frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid Y^i, \boldsymbol{A}) + \sigma^2\right) \end{aligned}$$

- Hence,

$$I(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{A}) \approx \sum_{i=0}^{m-1} \frac{1}{2} \log\left(1 + \frac{\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid Y^i, \boldsymbol{A})}{\sigma^2}\right)$$

Proof II: MMSE Satisfies Fixed-Point Equation

- By symmetry, MMSE is the same for all entries:

$$\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A}) = \mathsf{mmse}(X_n \mid \boldsymbol{Y}, \boldsymbol{A})$$

Proof II: MMSE Satisfies Fixed-Point Equation

- By symmetry, MMSE is the same for all entries:

$$\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A}) = \mathsf{mmse}(X_n \mid \boldsymbol{Y}, \boldsymbol{A})$$

- Let Q be an orthogonal matrix such that the first $(m-1)$ transformed observations are independent of X_n

$$\begin{aligned}(Q\boldsymbol{Y})_i &= \tilde{\boldsymbol{A}}_i X^{n-1} + \mathcal{N}(0, \sigma^2), \qquad i = 1, \dots, m-1 \\ (Q\boldsymbol{Y})_m &\approx \sqrt{m/n}\,X_n + \underbrace{\tilde{\boldsymbol{A}}_m X^{n-1} + \mathcal{N}(0, \sigma^2)}_{\text{effective Gaussian noise}}\end{aligned}$$

Proof II: MMSE Satisfies Fixed-Point Equation

- By symmetry, MMSE is the same for all entries:

$$\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A}) = \mathsf{mmse}(X_n \mid \boldsymbol{Y}, \boldsymbol{A})$$

- Let Q be an orthogonal matrix such that the first $(m-1)$ transformed observations are independent of X_n

$$\begin{aligned}(Q\boldsymbol{Y})_i &= \tilde{\boldsymbol{A}}_i X^{n-1} + \mathcal{N}(0,\sigma^2), \qquad i = 1,\dots,m-1\\ (Q\boldsymbol{Y})_m &\approx \sqrt{m/n}\,X_n + \underbrace{\tilde{\boldsymbol{A}}_m X^{n-1} + \mathcal{N}(0,\sigma^2)}_{\text{effective Gaussian noise}}\end{aligned}$$

- If $p(\boldsymbol{x} \mid (Q\boldsymbol{Y})^{m-1}, \boldsymbol{A})$ is weakly correlated, then effective noise is approx. Gaussian and this leads to the fixed-point equation:

$$\mathsf{mmse}(X_n \mid \boldsymbol{Y}, \boldsymbol{A}) \approx M_X\left(\frac{m/n}{\sigma^2 + \frac{1}{n}\,\mathsf{mmse}(X^{n-1} \mid (Q\boldsymbol{Y})^m, \boldsymbol{A})}\right)$$

Proof III: Weak Correlation Almost Everywhere

- By chain rule and problem symmetry, we obtain:

$$I(\boldsymbol{X}; Y_{m+1} \mid Y^m, \boldsymbol{A}) = I(\boldsymbol{X}; Y_1 \mid \boldsymbol{A}) - \sum_{i=0}^{m-1} I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A})$$

Proof III: Weak Correlation Almost Everywhere

- By chain rule and problem symmetry, we obtain:

$$I(\boldsymbol{X}; Y_{m+1} \mid Y^m, \boldsymbol{A}) = I(\boldsymbol{X}; Y_1 \mid \boldsymbol{A}) - \sum_{i=0}^{m-1} I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A})$$

- Correlation in $p(\boldsymbol{x} \mid Y^i, \boldsymbol{A})$ is upper bounded by

$$\frac{1}{n^2} \sum_{k,\ell} \mathbb{E}\big[\mathsf{Cov}(X_k, X_\ell \mid Y^i, \boldsymbol{A})\big] \leq \Psi\big(I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A})\big)$$

where $\Psi(\cdot)$ is concave and non-decreasing.

Proof III: Weak Correlation Almost Everywhere

- By chain rule and problem symmetry, we obtain:

$$I(\boldsymbol{X}; Y_{m+1} \mid Y^m, \boldsymbol{A}) = I(\boldsymbol{X}; Y_1 \mid \boldsymbol{A}) - \sum_{i=0}^{m-1} I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A})$$

- Correlation in $p(\boldsymbol{x} \mid Y^i, \boldsymbol{A})$ is upper bounded by

$$\frac{1}{n^2} \sum_{k,\ell} \mathbb{E}\big[\mathsf{Cov}(X_k, X_\ell \mid Y^i, \boldsymbol{A})\big] \leq \Psi\big(I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A})\big)$$

  where $\Psi(\cdot)$ is concave and non-decreasing.

- But $\sum_{i=0}^{\infty} I(Y_{i+1}; Y_{i+2} \mid Y^i, \boldsymbol{A}) \leq I(Y_1; \boldsymbol{X} \mid \boldsymbol{A})$ is finite, and thus

$$\#\{i : p(\boldsymbol{x} \mid Y^i, \boldsymbol{A}) \text{ has significant correlation}\} = o(n)$$

Proof IV: Resolve Limits via Area Theorem

Show that

$$\frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{A}) \approx \frac{1}{n} \sum_{i=1}^{m} \frac{1}{2} \log\left(1 + \frac{\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid Y^i, \boldsymbol{A})}{\sigma^2}\right)$$

$$\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A}) \approx M_X\left(\frac{m/n}{\sigma^2 + \frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A})}\right)$$

Proof IV: Resolve Limits via Area Theorem

Show that

$$\frac{1}{n} I(\boldsymbol{X}; \boldsymbol{Y} \mid \boldsymbol{A}) \approx \frac{1}{n} \sum_{i=1}^{m} \frac{1}{2} \log\left(1 + \frac{\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid Y^i, \boldsymbol{A})}{\sigma^2}\right)$$

$$\frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A}) \approx M_X\left(\frac{m/n}{\sigma^2 + \frac{1}{n}\,\mathsf{mmse}(\boldsymbol{X} \mid \boldsymbol{Y}, \boldsymbol{A})}\right)$$

This implies asymptotic constraints

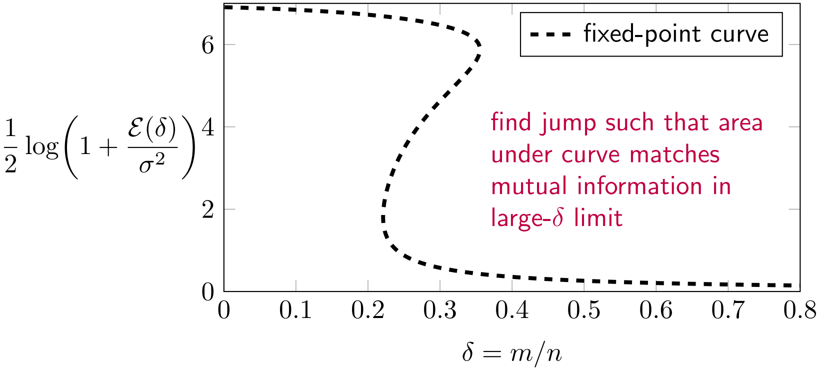
$$\mathcal{I}(\delta) = \int_0^\delta \frac{1}{2} \log\left(1 + \frac{\mathcal{E}(\delta')}{\sigma^2}\right) \mathrm{d}\delta' \qquad \text{integral constraint}$$

$$\mathcal{E}(\delta) = M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}(\delta)}\right), \quad \text{a.e.} \qquad \text{fixed-point constraint}$$

Proof IV: Resolve Limits via Area Theorem

$$\mathcal{I}(\delta) \approx \int_0^\delta \frac{1}{2}\log\left(1 + \frac{\mathcal{E}(\delta')}{\sigma^2}\right) \mathrm{d}\delta' \qquad \text{integral constraint}$$

$$\mathcal{E}(\delta) \approx M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}(\delta)}\right), \quad \text{a.e.} \qquad \text{fixed-point constraint}$$



Proof IV: Resolve Limits via Area Theorem

$$\mathcal{I}(\delta) \approx \int_0^{\delta} \frac{1}{2}\log\left(1 + \frac{\mathcal{E}(\delta')}{\sigma^2}\right) \mathrm{d}\delta' \qquad \text{integral constraint}$$

$$\mathcal{E}(\delta) \approx M_X\left(\frac{\delta}{\sigma^2 + \mathcal{E}(\delta)}\right), \quad \text{a.e.} \qquad \text{fixed-point constraint}$$

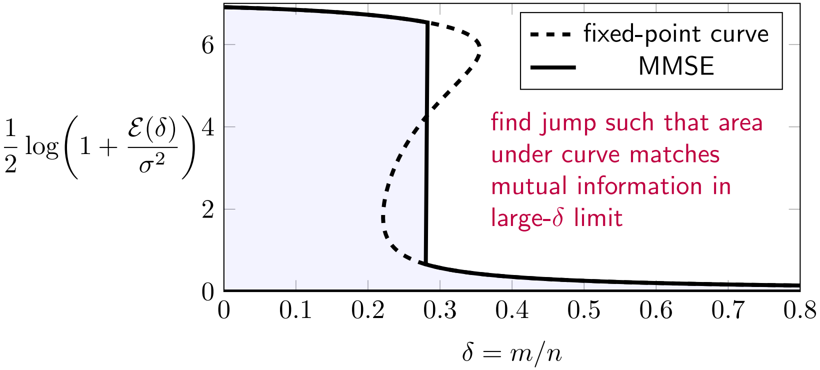


Table of Contents

Inference with Mismatched Prior

- Postulated distribution $q(\boldsymbol{x}, \boldsymbol{y})$ differs from true distribution $p(\boldsymbol{x}, \boldsymbol{y})$
- Example: MAP estimation via expectation with modified likelihood:

$$q_\beta(\boldsymbol{y}, \boldsymbol{x}) \propto [p(\boldsymbol{y} \mid \boldsymbol{x})]^\beta p(\boldsymbol{x})$$

$$\underbrace{\int \boldsymbol{x} q_\beta(\boldsymbol{x} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{x}}_{\text{postulated mean}} \quad \xrightarrow{\beta \to \infty} \quad \underbrace{\arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{MAP estimate}}$$

- Conjectured formulas for mismatched MMSE and M-estimators (e.g., MAP, MLE, LASSO)
- But replica symmetry does not hold in general!

Guo,Verdu'05, Rangan,Fletcher,Goyal'09, Bereyhi, Müller,Schulz-Baldes'17

Information in Multilayer Networks

Feed forward network with random weight matrices $\{W_\ell\}$

- Component-wise activation functions $\{\sigma_\ell\}$

$$\xrightarrow{\boldsymbol{x}} \boxed{\sigma_1(\boldsymbol{W}_1\boldsymbol{x})} \xrightarrow{\boldsymbol{h}_1} \cdots \xrightarrow{\boldsymbol{h}_{L-1}} \boxed{\sigma_L(\boldsymbol{W}_L\boldsymbol{h}_{L-1})} \xrightarrow{\boldsymbol{y}}$$

- Separable prior $p(\boldsymbol{x})$ and separable distributions $\{p_\ell(\cdot \mid \cdot)\}$

$$\boxed{p(\boldsymbol{x})} \xrightarrow{\boldsymbol{X}} \boxed{p_1(\cdot\,|\,\boldsymbol{W}_1\boldsymbol{x})} \xrightarrow{\boldsymbol{H}_1} \cdots \xrightarrow{\boldsymbol{H}_{L-1}} \boxed{p_L(\cdot\,|\,\boldsymbol{W}_L\boldsymbol{h}_{L-1})} \xrightarrow{\boldsymbol{Y}}$$

Manoel,Krzakala,Mézard,Zdeborová'17, Fletcher,Rangan,Schniter'18, R.'17, Gabrié,et.al'18

Information in Multilayer Networks

Feed forward network with random weight matrices $\{W_\ell\}$

- Component-wise activation functions $\{\sigma_\ell\}$

$$\xrightarrow{\boldsymbol{x}} \boxed{\sigma_1(\boldsymbol{W}_1\boldsymbol{x})} \xrightarrow{\boldsymbol{h}_1} \cdots \xrightarrow{\boldsymbol{h}_{L-1}} \boxed{\sigma_L(\boldsymbol{W}_L\boldsymbol{h}_{L-1})} \xrightarrow{\boldsymbol{y}}$$

- Separable prior $p(\boldsymbol{x})$ and separable distributions $\{p_\ell(\cdot \mid \cdot)\}$

$$\boxed{p(\boldsymbol{x})} \xrightarrow{\boldsymbol{X}} \boxed{p_1(\cdot\,|\,\boldsymbol{W}_1\boldsymbol{x})} \xrightarrow{\boldsymbol{H}_1} \cdots \xrightarrow{\boldsymbol{H}_{L-1}} \boxed{p_L(\cdot\,|\,\boldsymbol{W}_L\boldsymbol{h}_{L-1})} \xrightarrow{\boldsymbol{Y}}$$

What is mutual information between layers?

- State evolution for multilayer AMP / VAMP
- Conjectured formulas via heuristics
- Observed behavior under restricted learning of weights

Manoel,Krzakala,Mézard,Zdeborová'17, Fletcher,Rangan,Schniter'18, R.'17, Gabrié,et.al'18

Phase Diagrams in Other Problem

- Recent progress
  - Multilayer networks (random initialization)
  - Bilinear and multi-linear models (low rank) [See Jean's Talk]
  - Community detection (dense case)

Phase Diagrams in Other Problem

- Recent progress
  - Multilayer networks (random initialization)
  - Bilinear and multi-linear models (low rank) [See Jean's Talk]
  - Community detection (dense case)
- Future challenges
  - Multilayer networks (learned from data)
  - Bilinear and multi-linear models (extrinsic rank)
  - Community detection (sparse case)

References I

D. N. C. Tse and S. Hanly, “Linear multiuser receivers: Effective interference, effective bandwith and user capacity,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 641–657, Mar. 1999.

S. Verdú and S. Shamai, “Spectral efficiency of cdma with random spreading,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 622–640, Mar. 1999.

T. Tanaka, “A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors,” *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.

D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

G. Reeves and M. Gastpar, “The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 58, no. 5, pp. 3065–3092, May 2012.

——, “Compressed sensing phase transitions: Rigorous bounds versus replica predictions,” in *Proc. Conf. on Inform. Sciences and Systems*, Princeton, NJ, Mar. 2012.

References II

——, “Approximate sparsity pattern recovery: Information-theoretic lower bounds,” *IEEE Trans. Inform. Theory*, vol. 59, no. 6, pp. 3451–3465, Jun. 2013.

D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in Gaussian channels,” *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

G. Reeves, H. D. Pfister, and A. Dytso, “Mutual information as a function of matrix SNR for linear gaussian channels,” in *Proc. IEEE Int. Symp. Inform. Theory*, Vail, CO, Jun. 2018.

G. Reeves and H. D. Pfister, “The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact,” in *Proc. IEEE Int. Symp. Inform. Theory*, Barcelona, Spain, Jul. 2016, pp. 665 – 669.

——, “The replica-symmetric prediction for random linear estimation with Gaussian matrices is exact,” *IEEE Trans. Inform. Theory*, vol. 65, no. 4, pp. 2252–2283, Apr. 2019.

J. Barbier, M. Dia, N. Macris, and F. Krzakala, “The mutual information in random linear estimation,” in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, Monticello, IL, 2016.

S. Rangan, A. K. Fletcher, and V. K. Goyal, “Asymptotic analysis of map estimation via the replica method and applications to compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 58, no. 3, pp. 1902 – 1923, Mar. 2012.

References III

A. Bereyhi, R. R. Müller, and H. Schulz-Baldes, "Statistical mechanics of MAP estimation: General replica ansatz," Oct. 2017, [Online]. Available https://arxiv.org/pdf/1612.01980.pdf.

A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, "Multi-layer generalized linear estimation," in *Proc. IEEE Int. Symp. Inform. Theory*, Aachen, Germany, 2017, pp. 2098–2102.

G. Reeves, "Additivity of information in multilayer networks via additive Gaussian noise transforms," in *Proc. Annual Allerton Conf. on Commun., Control, and Comp.*, Monticello, IL, 2017, [Online]. Available https://arxiv.org/abs/1710.04580.

A. K. Fletcher, S. Rangan, and P. Schniter, "Inference in deep networks in high dimensions," in *Proc. IEEE Int. Symp. Inform. Theory*, Vail, CO, Jun. 2018.

M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," 2018, [Online]. Available https://arxiv.org/abs/1805.09785.

G. Reeves, J. Xu, and I. Zadik, "The all-or-nothing phenomenon in sparse linear regression," in *Conference On Learning Theory (COLT)*, 2019, [Online]. Available https://arxiv.org/abs/1903.05046.

References IV

G. Reeves and H. D. Pfister, "Understanding phase transitions via mutual information and mmse," 2019, [Online]. Available https://arxiv.org/abs/1907.02095.

G. Reeves, V. Mayya, and A. Volfovsky, "The geometry of community detection via the mmse matrix," in *Proc. IEEE Int. Symp. Inform. Theory*, Paris, France, Jul. 2019.