Approximate Message Passing for High-Dimensional Inference, I

Ramji Venkataramanan

University of Cambridge

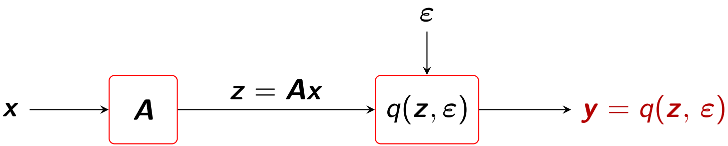
ISIT 2023

Focus of tutorial

Approximate Message Passing (AMP) for

1. Estimation in linear and generalized linear models
2. Low-rank matrix estimation

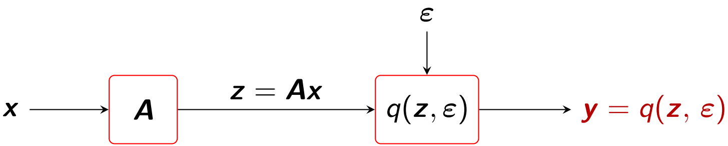
Generalized Linear Models (GLMs)



GOAL:

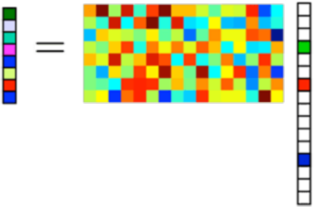
- Estimate signal $\boldsymbol{x} \in \mathbb{R}^d$ from observations $\boldsymbol{y} \equiv (y_1, \ldots, y_n)$
- Known sensing matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ and output function q

Example: Linear model

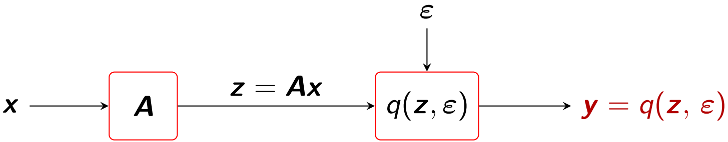


Linear model: $\boldsymbol{y} = \boldsymbol{Ax} + \varepsilon$

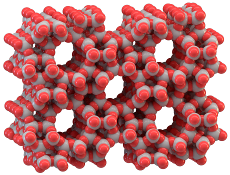
- Widely used model in signal processing and communications: CDMA, MIMO, sparse regression codes ...
- Compressed sensing: Signal $\boldsymbol{x}$ assumed to be sparse



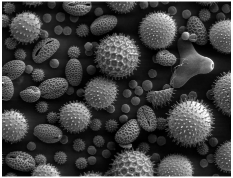
Example: Phase retrieval



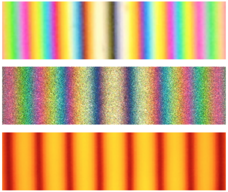
Phase retrieval: $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}|^2 + \varepsilon$



X-ray crystallography

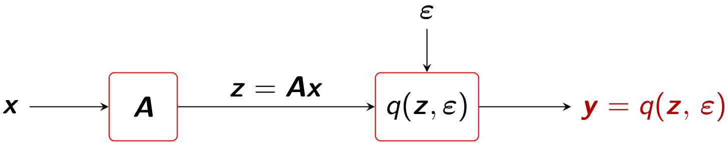


Microscopy



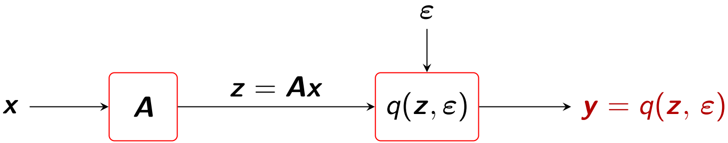
Interferometry

Example: 1-bit compressed sensing



1-bit compressed sensing [Boufounos '08]: $\boldsymbol{y} = \mathrm{sign}(\boldsymbol{A}\boldsymbol{x} + \varepsilon)$

Example: 1-bit compressed sensing



1-bit compressed sensing [Boufounos '08]: $\boldsymbol{y} = \mathrm{sign}(\boldsymbol{Ax} + \varepsilon)$

Many other popular GLMs, e.g.,

- Logistic, probit regression (Binary classification)
- Poisson regression (count data)

Low-rank models

$$\underset{n \times d}{\begin{bmatrix} \mathbf{A} \end{bmatrix}} \approx \underset{n \times k}{\begin{bmatrix} \mathbf{U} \end{bmatrix}} \overset{k \times d}{\begin{bmatrix} \mathbf{V}^{\mathsf{T}} \end{bmatrix}}$$

Topic Modelling

- Each row of $\boldsymbol{A}$ is a document
- Each row of $\boldsymbol{V}^{\mathsf{T}}$ is a topic
- Each document convex combination of k topics

[Blei, Ng, Jordan '03]

Low-rank models

$$\underset{n \times d}{\begin{bmatrix} \\ \quad \mathbf{A} \quad \\ \\ \end{bmatrix}} \approx \underset{n \times k}{\begin{bmatrix} \\ \mathbf{U} \\ \\ \end{bmatrix}} \overset{k \times d}{\begin{bmatrix} \quad \mathbf{V}^{\mathsf{T}} \quad \end{bmatrix}}$$

Collaborative filtering

- $\boldsymbol{A}$ contains ratings of users for items (e.g, films or books)
- Rows represent users, columns represent items
- Each rating is a combination of weights corresponding to a small number of factors

Hidden clique

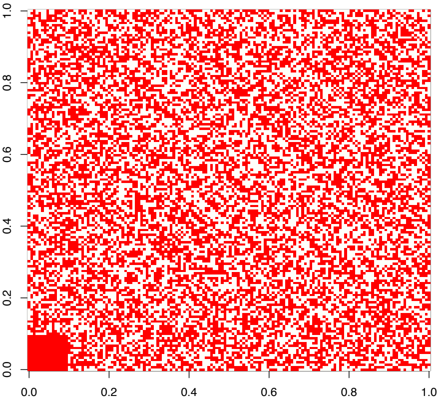


Image from *Statistical Estimation: From Denoising to Sparse Regression and Hidden Cliques* by A. Montanari

[Alon, Krivelivich, Sudakov '98], [Deshpande, Montanari '15]

Hidden clique

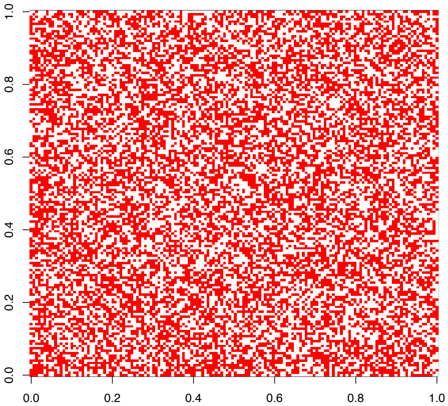


Image from *Statistical Estimation: From Denoising to Sparse Regression and Hidden Cliques* by A. Montanari

[Alon, Krivelivich, Sudakov '98], [Deshpande, Montanari '15]

Hidden clique

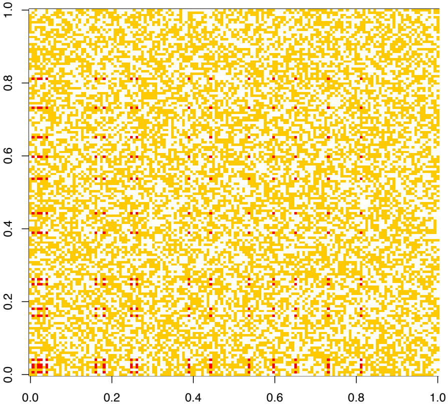
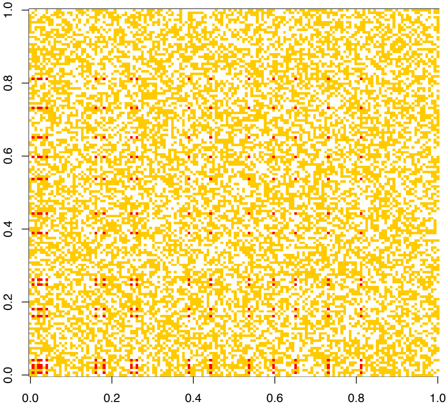


Image from *Statistical Estimation: From Denoising to Sparse Regression and Hidden Cliques* by A. Montanari

[Alon, Krivelivich, Sudakov '98], [Deshpande, Montanari '15]

Hidden clique



For hidden clique S, adjacency matrix has the form

$$\boldsymbol{A} = \mathbf{1}_S \mathbf{1}_S^{\mathsf{T}} + \boldsymbol{W}$$

[Alon, Krivelivich, Sudakov '98], [Deshpande, Montanari '15]

Structure of Tutorial

1. Introduction to AMP, application to low-rank matrix estimation

2. AMP to derive exact asymptotics in generalized linear models (Cynthia Rush)

3. AMP as a flexible tool in high-dimensional statistics (Marco Mondelli)

Origins of AMP

- Relaxation of belief propagation for CDMA multiuser detection:
  [Kabashima '03], [Caire, Muller, Tanaka '04], [Tanaka, Okada '05]

- Via systematic approximation of BP iterations:
  1. Compressed sensing (linear models): [Donoho, Maleki, Montanari '09], [Krzakala et. al '11]
  2. Generalized linear models: [Rangan '11]
  3. Low-rank matrix estimation: [Parker, Schniter, Cevher '14], [Fletcher, Rangan '18], [Lesieur et al., '17]

Origins of AMP

- Relaxation of belief propagation for CDMA multiuser detection:
  [Kabashima '03], [Caire, Muller, Tanaka '04], [Tanaka, Okada '05]

- Via systematic approximation of BP iterations:
  1. Compressed sensing (linear models): [Donoho, Maleki, Montanari '09], [Krzakala et. al '11]
  2. Generalized linear models: [Rangan '11]
  3. Low-rank matrix estimation: [Parker, Schniter, Cevher '14], [Fletcher, Rangan '18], [Lesieur et al., '17]

We'll take a different approach to understanding AMP:
Study it as an iteration defined via a random matrix

Gaussian Orthogonal Ensemble (GOE)

Consider a symmetric Gaussian matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$

$$W_{ij} \ \text{ independent for } \ 1 \leq i \leq j \leq n$$

$$W_{ij} \sim \mathrm{N}\left(0, \frac{1}{n}\right) \ \text{ for } \ i \neq j, \qquad W_{ij} \sim \mathrm{N}\left(0, \frac{2}{n}\right) \ \text{ for } \ i = j.$$

We write $\boldsymbol{W} \sim \mathrm{GOE}(n)$

Gaussian Orthogonal Ensemble (GOE)

Consider a symmetric Gaussian matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$

$$W_{ij} \ \text{ independent for } \ 1 \leq i \leq j \leq n$$

$$W_{ij} \sim \mathrm{N}\left(0, \frac{1}{n}\right) \ \text{ for } \ i \neq j, \qquad W_{ij} \sim \mathrm{N}\left(0, \frac{2}{n}\right) \ \text{ for } \ i = j.$$

We write $\boldsymbol{W} \sim \mathrm{GOE}(n)$

Property

If $\boldsymbol{W} \sim \mathrm{GOE}(n)$ and $\boldsymbol{Q}$ is any $n \times n$ orthogonal matrix, then:

$$\boldsymbol{Q}^{\mathsf{T}} \boldsymbol{W} \boldsymbol{Q} \sim \mathrm{GOE}(n)$$

An iteration with a GOE matrix

Let $\boldsymbol{W}$ be a GOE matrix

Starting with an initialization $\boldsymbol{h}^0 \in \mathbb{R}^n$, define for $t \geq 0$:

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t - \mathrm{b}_t\,\boldsymbol{m}^{t-1}$$

An iteration with a GOE matrix

Let $\boldsymbol{W}$ be a GOE matrix

Starting with an initialization $\boldsymbol{h}^0 \in \mathbb{R}^n$, define for $t \geq 0$:

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t - \mathrm{b}_t\,\boldsymbol{m}^{t-1}$$

- Function f_t is Lipschitz and acts component-wise, for $t \geq 1$
- Coefficient $\mathrm{b}_t = \frac{1}{n}\sum_{i=1}^n f_t^{\,'}(h_i^t)$
- First step: $\boldsymbol{h}^1 = \boldsymbol{W}\,f_0(\boldsymbol{h}^0)$

We call this the **abstract AMP** recursion

State Evolution

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \;-\; \mathsf{b}_t\,\boldsymbol{m}^{t-1}$$

Key result (informal): If initialization $\boldsymbol{h}^0$ is independent of $\boldsymbol{W}$, then for $t \geq 1$, as $n \to \infty$, the empirical distribution of $\boldsymbol{h}^t$ converges to $\mathsf{N}(0, \tau_t^2)$, where

$$\tau_{t+1}^2 = \mathbb{E}\{(f_t(G_t))^2\}, \qquad G_t \sim \mathsf{N}(0, \tau_t^2)$$

- The $\tau_t \to \tau_{t+1}$ recursion is called **state evolution**
- Initialized with $\tau_1^2 = \lim_{n\to\infty} \frac{\|f_0(\boldsymbol{h}^0)\|^2}{n}$

State Evolution

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \;-\; \mathsf{b}_t\,\boldsymbol{m}^{t-1}$$

Key result (informal): If initialization $\boldsymbol{h}^0$ is independent of $\boldsymbol{W}$, then for $t \geq 1$, as $n \to \infty$, the empirical distribution of $\boldsymbol{h}^t$ converges to $\mathsf{N}(0, \tau_t^2)$, where

$$\tau_{t+1}^2 = \mathbb{E}\{(f_t(G_t))^2\}, \qquad G_t \sim \mathsf{N}(0, \tau_t^2)$$

- The $\tau_t \to \tau_{t+1}$ recursion is called **state evolution**
- Initialized with $\tau_1^2 = \lim_{n\to\infty} \frac{\|f_0(\boldsymbol{h}^0)\|^2}{n}$

Why is this true? Why is it interesting?

Heuristic for state evolution

First step: $\boldsymbol{h}^1 = \boldsymbol{W}\boldsymbol{m}^0$

- Let $\nu_n(\boldsymbol{h}^1)$ denote empirical distribution of $\boldsymbol{h}^1$
- Since $\boldsymbol{m}^0$ is independent of $\boldsymbol{W}$, we have $\boldsymbol{h}^1$ Gaussian with $\nu_n(\boldsymbol{h}^1) \to \mathrm{N}(0, \tau_1^2)$ where

$$\tau_1^2 = \lim_{n\to\infty} \frac{\|\boldsymbol{m}^0\|^2}{n} = \lim_{n\to\infty} \frac{\|f_0(\boldsymbol{h}^0)\|^2}{n}$$

Heuristic for state evolution

First step: $\boldsymbol{h}^1 = \boldsymbol{W}\boldsymbol{m}^0$

- Let $\nu_n(\boldsymbol{h}^1)$ denote empirical distribution of $\boldsymbol{h}^1$
- Since $\boldsymbol{m}^0$ is independent of $\boldsymbol{W}$, we have $\boldsymbol{h}^1$ Gaussian with $\nu_n(\boldsymbol{h}^1) \to \mathrm{N}(0, \tau_1^2)$ where

$$\tau_1^2 = \lim_{n\to\infty} \frac{\|\boldsymbol{m}^0\|^2}{n} = \lim_{n\to\infty} \frac{\|f_0(\boldsymbol{h}^0)\|^2}{n}$$

Second step: $\boldsymbol{h}^2 = \boldsymbol{W}\,\boldsymbol{m}^1 - \mathrm{b}_1\,\boldsymbol{m}^0$, with $\boldsymbol{m}^1 = f_1(\boldsymbol{h}^1)$

- $\boldsymbol{W}$ and $\boldsymbol{m}^1$ are *dependent*, so $\boldsymbol{W}\boldsymbol{m}^1$ is **not** Gaussian

Heuristic for state evolution

First step: $\boldsymbol{h}^1 = \boldsymbol{W}\boldsymbol{m}^0$

- Let $\nu_n(\boldsymbol{h}^1)$ denote empirical distribution of $\boldsymbol{h}^1$
- Since $\boldsymbol{m}^0$ is independent of $\boldsymbol{W}$, we have $\boldsymbol{h}^1$ Gaussian with $\nu_n(\boldsymbol{h}^1) \to \mathrm{N}(0, \tau_1^2)$ where

$$\tau_1^2 = \lim_{n\to\infty} \frac{\|\boldsymbol{m}^0\|^2}{n} = \lim_{n\to\infty} \frac{\|f_0(\boldsymbol{h}^0)\|^2}{n}$$

Second step: $\boldsymbol{h}^2 = \boldsymbol{W}\,\boldsymbol{m}^1 - \mathrm{b}_1\,\boldsymbol{m}^0$, with $\boldsymbol{m}^1 = f_1(\boldsymbol{h}^1)$

- $\boldsymbol{W}$ and $\boldsymbol{m}^1$ are *dependent*, so $\boldsymbol{W}\boldsymbol{m}^1$ is **not** Gaussian
- For $\tilde{\boldsymbol{W}} \sim \mathrm{GOE}(n)$ independent of $\boldsymbol{m}^1$, we have $\tilde{\boldsymbol{W}}\boldsymbol{m}^1$ Gaussian with $\nu_n(\tilde{\boldsymbol{W}}\boldsymbol{m}^1) \to \mathrm{N}(0, \tau_2^2)$, where

$$\tau_2^2 = \lim_{n\to\infty} \frac{\|\boldsymbol{m}^1\|^2}{n} = \lim_{n\to\infty} \frac{\|f_1(\boldsymbol{h}^1)\|^2}{n} = \mathbb{E}\{f_1(G_1)^2\}, \quad G_1 \sim \mathrm{N}(0, \tau_1^2)$$

Debiasing term

$$\boldsymbol{h}^2 = \boldsymbol{W}\,\boldsymbol{m}^1 \;-\; \mathrm{b}_1\,\boldsymbol{m}^0, \qquad\qquad \mathrm{b}_1 = \frac{1}{n}\sum_{i=1}^{n} f_1^{\,\prime}(h_i^1)$$

- The 'Onsager' correction $-\mathrm{b}_1\boldsymbol{m}^0$ is as a **debiasing** term
- Ensures that $\boldsymbol{h}^2$ asymptotically has the same empirical distribution as $\tilde{\boldsymbol{W}}\boldsymbol{m}^1$. That is, $\nu_n(\boldsymbol{h}^2) \to \mathrm{N}(0, \tau_2^2)$

$$\boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \;-\; \mathrm{b}_t\,\boldsymbol{m}^{t-1}, \qquad\qquad \mathrm{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t^{\,\prime}(h_i^t)$$

Debiasing term

$$\boldsymbol{h}^2 = \boldsymbol{W}\,\boldsymbol{m}^1 \,-\, \mathrm{b}_1\,\boldsymbol{m}^0, \qquad \mathrm{b}_1 = \frac{1}{n}\sum_{i=1}^{n} f_1^{'}(h_i^1)$$

- The 'Onsager' correction $-\mathrm{b}_1\boldsymbol{m}^0$ is as a **debiasing** term
- Ensures that $\boldsymbol{h}^2$ asymptotically has the same empirical distribution as $\tilde{\boldsymbol{W}}\boldsymbol{m}^1$. That is, $\nu_n(\boldsymbol{h}^2) \to \mathrm{N}(0, \tau_2^2)$

$$\boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \,-\, \mathrm{b}_t\,\boldsymbol{m}^{t-1}, \qquad \mathrm{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t^{'}(h_i^t)$$

- Conditional distribution of $\boldsymbol{W}\,\boldsymbol{m}^t$ given $(\boldsymbol{m}^0, \ldots, \boldsymbol{m}^t)$ can be decomposed into Gaussian component and a non-Gaussian one
- Non-Gaussian part asymptotically cancelled out by $-\mathrm{b}_t\boldsymbol{m}^{t-1}$

Pseudo-Lipschitz test functions

A function $\phi : \mathbb{R}^m \to \mathbb{R}$ is called pseudo-Lipschitz if for all inputs $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$,

$$|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})| \leq C\|\boldsymbol{x} - \boldsymbol{y}\|(1 + \|\boldsymbol{x}\| + \|\boldsymbol{y}\|)$$

for some constant $C > 0$

- Roughly: Functions with at most quadratic growth
- Examples: $\phi(x) = x^2$, $\phi(x, y) = xy$

State evolution results for AMP often stated in terms of pseudo-Lipschitz test functions

E.g., mean-squared error (MSE) of estimate $\phi(x, y) = (x - y)^2$

Main result for abstract AMP

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \,-\, \mathsf{b}_t\,\boldsymbol{m}^{t-1}$$

Assumptions:

- Functions f_t Lipschitz, for $t \geq 1$
- Initialization $\boldsymbol{h}^0$ is independent of $\boldsymbol{W}$

Theorem [Bolthausen '10, Bayati-Montanari '11]

For $t \geq 1$, and any pseudo-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi\big(h_i^t\big) \;=\; \mathbb{E}\{\phi(G_t)\} \;\text{ almost surely}$$

where $G_t \sim \mathsf{N}(0, \tau_t^2)$, with $\tau_{t+1}^2 = \mathbb{E}\{f_t(G_t)^2\}$.

Main result for abstract AMP

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \; - \; \mathrm{b}_t\,\boldsymbol{m}^{t-1}$$

Assumptions:

- Functions f_t Lipschitz, for $t \ge 1$
- Initialization $\boldsymbol{h}^0$ is independent of $\boldsymbol{W}$

Theorem [Bolthausen '10, Bayati-Montanari '11]

For $t \ge 1$, and any pseudo-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi\big(h_i^t\big) \; = \; \mathbb{E}\{\phi(G_t)\} \;\; \text{almost surely}$$

where $G_t \sim \mathrm{N}(0, \tau_t^2)$, with $\tau_{t+1}^2 = \mathbb{E}\{f_t(G_t)^2\}$.

Equivalent to: empirical distribution $\nu_n(\boldsymbol{h}^t)$ converges to $\mathrm{N}(0, \tau_t^2)$ almost surely (in Wasserstein-2 distance)

Stronger statement

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \;-\; \mathsf{b}_t\,\boldsymbol{m}^{t-1}$$

Theorem [Javanmard-Montanari '13]

For $t \geq 1$, and any pseudo-Lipschitz function $\phi : \mathbb{R}^t \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \phi\big(h_i^1, h_i^2, \ldots, h_i^t\big) \;=\; \mathbb{E}\{\phi(G_1, G_2, \ldots, G_t)\} \text{ almost surely}$$

where $(G_1, \ldots, G_t) \sim \mathrm{N}(0, \Sigma_t)$, where $\Sigma_t \in \mathbb{R}^{t\times t}$ can be recursively computed via state evolution, for $t \geq 1$.

Empirical distribution of rows of $\nu_n(\boldsymbol{h}^1, \ldots, \boldsymbol{h}^t)$ converges (in Wasserstein-2 distance) to $\mathrm{N}(0, \Sigma_t)$ almost surely

---

Unifying tutorial on approximate message passing, [Feng et al. '22]

Rank-1 matrix estimation

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v}\, \boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

- Signal $\boldsymbol{v} \in \mathbb{R}^n$, entries $v_i \sim_{\text{iid}} P_V$
- Noise matrix $\boldsymbol{W} \sim \text{GOE}(n)$

[Baik, Ben Arous, Péché '05], [Baik, Silverstein '06], [Capitaine, Donati-Martin, Féral '09], [Benaych-Georges and Nadakuditi '11], ...

Rank-1 matrix estimation

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v} \, \boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

- Signal $\boldsymbol{v} \in \mathbb{R}^n$, entries $v_i \sim_{\text{iid}} P_V$
- Noise matrix $\boldsymbol{W} \sim \text{GOE}(n)$

Natural estimator: $\hat{\varphi}$ the principal eigenvector of $\boldsymbol{A}$

Random matrix theory shows phase transition:

$$\text{Principal eigenvalue} \quad \lambda_1(\boldsymbol{A}) \to \begin{cases} \lambda + \lambda^{-1}, & \text{if } \lambda > 1, \\ 2, & \text{if } \lambda \in (0, 1] \end{cases}$$

$$\text{Correlation} \quad \frac{|\langle \hat{\varphi}, \boldsymbol{v} \rangle|}{\|\hat{\varphi}\| \|\boldsymbol{v}\|} \to \begin{cases} \sqrt{1 - \lambda^{-2}}, & \text{if } \lambda > 1, \\ 0, & \text{if } \lambda \in (0, 1] \end{cases}$$

[Baik, Ben Arous, Péché '05], [Baik, Silverstein '06], [Capitaine, Donati-Martin, Féral '09], [Benaych-Georges and Nadakuditi '11], ...

Structural information

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v}\, \boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

Spectral estimator $\hat{\varphi}$ doesn't use *structural* information about $\boldsymbol{v}$

- For example, $\boldsymbol{v}$ may be sparse, bounded, non-negative etc.
- Relevant in sparse PCA, non-negative PCA, hidden clique, community detection under stochastic block model, . . .

[Deshpande, Montanari '14], [Barbier *et al.* '16], [Lesieur *et al.* '17], [Miolane, Lelarge '16] . . .

Structural information

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

Spectral estimator $\hat{\varphi}$ doesn't use *structural* information about $\boldsymbol{v}$

- For example, $\boldsymbol{v}$ may be sparse, bounded, non-negative etc.
- Relevant in sparse PCA, non-negative PCA, hidden clique, community detection under stochastic block model, ...

If we know prior P_V on entries of $\boldsymbol{v}$, MMSE estimator is

$$\widehat{\boldsymbol{M}}_{\text{Bayes}} = \mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} \mid \boldsymbol{A}\right]$$

$\widehat{\boldsymbol{M}}_{\text{Bayes}}$ is generally not computable, but computable formula for asymptotic Bayes risk available

[Deshpande, Montanari '14], [Barbier *et al.* '16], [Lesieur *et al.* '17], [Miolane, Lelarge '16] ...

AMP for rank-1 estimation

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \boldsymbol{W} \sim \mathrm{GOE}(n)$$

Let's try same AMP iteration as before, but defined via $\boldsymbol{A}$

$$\hat{\boldsymbol{v}}^{t} = f_t(\boldsymbol{v}^{t}), \qquad \boldsymbol{v}^{t+1} = \boldsymbol{A}\,\hat{\boldsymbol{v}}^{t} - \mathrm{b}_t\,\hat{\boldsymbol{v}}^{t-1}, \qquad \mathrm{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t{}'(v_i^t)$$

AMP for rank-1 estimation

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \boldsymbol{W} \sim \mathrm{GOE}(n)$$

Let's try same AMP iteration as before, but defined via $\boldsymbol{A}$

$$\hat{\boldsymbol{v}}^t = f_t(\boldsymbol{v}^t), \qquad \boldsymbol{v}^{t+1} = \boldsymbol{A}\,\hat{\boldsymbol{v}}^t - \mathrm{b}_t\,\hat{\boldsymbol{v}}^{t-1}, \qquad \mathrm{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t^{\,\prime}(v_i^t)$$

Using the expression for $\boldsymbol{A}$:

$$\boldsymbol{v}^{t+1} = \lambda \frac{\langle \boldsymbol{v},\, \hat{\boldsymbol{v}}^t \rangle}{n}\,\boldsymbol{v} + \boldsymbol{W}\,\hat{\boldsymbol{v}}^t - \mathrm{b}_t\,\hat{\boldsymbol{v}}^{t-1}$$

Shift + abstract AMP iterate

First iteration

Suppose

$$\boldsymbol{v}^0 = \mu_0\, \boldsymbol{v} \,+\, \boldsymbol{g}^0\,, \qquad \text{with } \boldsymbol{g}_0 \sim \mathrm{N}(0, \sigma_0^2 \boldsymbol{I}_n)$$

for some constants μ_0, σ_0. Then

$$\boldsymbol{v}^1 = \lambda \frac{\langle \boldsymbol{v},\ \hat{\boldsymbol{v}}^0 \rangle}{n}\, \boldsymbol{v} \,+\, \boldsymbol{W}\, \hat{\boldsymbol{v}}^0, \qquad \hat{\boldsymbol{v}}^0 = f_0(\boldsymbol{v}_0)$$

First iteration

Suppose

$$\boldsymbol{v}^0 = \mu_0\,\boldsymbol{v} + \boldsymbol{g}^0\,, \qquad \text{with } \boldsymbol{g}_0 \sim \mathrm{N}(0, \sigma_0^2 \boldsymbol{I}_n)$$

for some constants μ_0, σ_0. Then

$$\boldsymbol{v}^1 = \lambda\frac{\langle \boldsymbol{v},\ \hat{\boldsymbol{v}}^0\rangle}{n}\,\boldsymbol{v} + \boldsymbol{W}\,\hat{\boldsymbol{v}}^0, \qquad \hat{\boldsymbol{v}}^0 = f_0(\boldsymbol{v}_0)$$

- Signal term:

$$\frac{\lambda\langle \boldsymbol{v},\ \hat{\boldsymbol{v}}^0\rangle}{n} = \frac{\lambda}{n}\sum_{i=1}^{n} v_i\, f_0(v_i^0) \to \mathbb{E}\{V f_0(\mu_0 V + G_0)\} =: \mu_1$$

where $V \sim P_V$ and $G_0 \sim \mathrm{N}(0, \sigma_0^2)$ are independent

First iteration

Suppose

$$\boldsymbol{v}^0 = \mu_0\, \boldsymbol{v} \;+\; \boldsymbol{g}^0\,, \qquad \text{with } \boldsymbol{g}_0 \sim \mathrm{N}(0, \sigma_0^2 \boldsymbol{I}_n)$$

for some constants μ_0, σ_0. Then

$$\boldsymbol{v}^1 = \lambda \frac{\langle \boldsymbol{v},\; \hat{\boldsymbol{v}}^0 \rangle}{n}\, \boldsymbol{v} \;+\; \boldsymbol{W}\, \hat{\boldsymbol{v}}^0, \qquad \hat{\boldsymbol{v}}^0 = f_0(\boldsymbol{v}_0)$$

- Signal term:

$$\frac{\lambda \langle \boldsymbol{v},\; \hat{\boldsymbol{v}}^0 \rangle}{n} = \frac{\lambda}{n} \sum_{i=1}^{n} v_i\, f_0(v_i^0) \;\to\; \mathbb{E}\{V f_0(\mu_0 V + G_0)\} =: \mu_1$$

  where $V \sim P_V$ and $G_0 \sim \mathrm{N}(0, \sigma_0^2)$ are independent

- Empirical distribution of $\boldsymbol{W} f_0(\boldsymbol{v}^0) \;\to\; \mathrm{N}(0, \sigma_1^2)$ where

$$\sigma_1^2 := \lim_{n\to\infty} \frac{\|\boldsymbol{v}^0\|^2}{n} = \mathbb{E}\{f_0(\mu_0 V + G_0)^2\}$$

$\Rightarrow$ Empirical dist. $\nu_n(\boldsymbol{v}^1) \;\to\; \mu_1 V \;+\; G_1$, with $G_1 \sim \mathrm{N}(0, \sigma_1^2)$

Subsequent iterations

Recall the AMP iteration:

$$\hat{\boldsymbol{v}}^t = f_t(\boldsymbol{v}^t), \qquad \boldsymbol{v}^{t+1} = \boldsymbol{A}\,\hat{\boldsymbol{v}}^t - \mathrm{b}_t\,\hat{\boldsymbol{v}}^{t-1}, \qquad \mathrm{b}_t = \frac{1}{n}\sum_{i=1}^{n} f_t^{\,\prime}(v_i^t)$$

Suppose $\nu_n(\boldsymbol{v}^t) \to \mu_t V + G_t$, with $G_t \sim \mathrm{N}(0, \sigma_t^2)$

$$\boldsymbol{v}^{t+1} = \underbrace{\lambda\frac{\langle \boldsymbol{v},\, \hat{\boldsymbol{v}}^t\rangle}{n}\,\boldsymbol{v}}_{\approx\,\mu_{t+1}\boldsymbol{v}} + \underbrace{\boldsymbol{W}\,\hat{\boldsymbol{v}}^t - \mathrm{b}_t\,\hat{\boldsymbol{v}}^{t-1}}_{\approx\,\mathrm{N}(0,\sigma_{t+1}^2\boldsymbol{I}_n)}$$

State evolution recursion

$$\mu_{t+1} = \lambda\,\mathbb{E}[V\,f_t(\mu_t V + G_t)], \quad \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t V + G_t)^2]$$

where $G_t \sim \mathrm{N}(0, \sigma_t^2)$ indep. of $V \sim P_V$. Initialize with μ_0, σ_0

Main result for rank-one AMP

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v}\, \boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \boldsymbol{W} \sim \mathrm{GOE}(n)$$

Assumptions:

- Functions f_t Lipschitz, for $t \geq 1$
- Initialization $\boldsymbol{v}^0$ is independent of $\boldsymbol{W}$

Theorem

For $t \geq 1$, and any pseudo-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi\big(v_i,\, v_i^t\big) = \mathbb{E}\{\phi(V,\, \mu_t V + G_t)\} \;\; \text{almost surely}$$

where $G_t \sim \mathrm{N}(0, \sigma_t^2)$ independent of $V \sim P_V$

Main result for rank-one AMP

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}\,, \qquad \boldsymbol{W} \sim \mathrm{GOE}(n)$$

Assumptions:

- Functions f_t Lipschitz, for $t \geq 1$
- Initialization $\boldsymbol{v}^0$ is independent of $\boldsymbol{W}$

Theorem

For $t \geq 1$, and any pseudo-Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \phi\big(v_i,\, v_i^t\big) = \mathbb{E}\{\phi(V,\, \mu_t V + G_t)\} \text{ almost surely}$$

where $G_t \sim \mathrm{N}(0, \sigma_t^2)$ independent of $V \sim P_V$

$$\boxed{\text{Implies } \lim_{n\to\infty} \frac{\langle \boldsymbol{v}, \hat{\boldsymbol{v}}^t\rangle}{n} = \mathbb{E}\{V\, f_t(\mu_t V + G_t)\}, \text{ for each } t \geq 1}$$

Choosing f_t

AMP result says $\boldsymbol{v}^t \overset{\mathrm{d}}{\approx} \mu_t V + G_t$, with $G_t \sim \mathrm{N}(0, \sigma_t^2)$

$$\mu_{t+1} = \lambda\, \mathbb{E}[V\, f_t(\mu_t V + G_t)], \quad \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t V + G_t)^2]$$

- Given μ_t, σ_t, want to choose f_t to maximize

$$\gamma_{t+1} := \frac{\mu_{t+1}^2}{\sigma_{t+1}^2}$$

Choosing f_t

AMP result says $\boldsymbol{v}^t \stackrel{\mathrm{d}}{\approx} \mu_t V + G_t$, with $G_t \sim \mathrm{N}(0, \sigma_t^2)$

$$\mu_{t+1} = \lambda \, \mathbb{E}[V \, f_t(\mu_t V + G_t)], \quad \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t V + G_t)^2]$$

- Given μ_t, σ_t, want to choose f_t to maximize

$$\gamma_{t+1} := \frac{\mu_{t+1}^2}{\sigma_{t+1}^2}$$

- If we know the prior distribution $V \sim P_V$, optimal choice is

$$f_t^*(s) = \mathbb{E}\{V \mid \mu_t V + \sigma_t G_t = s\}$$

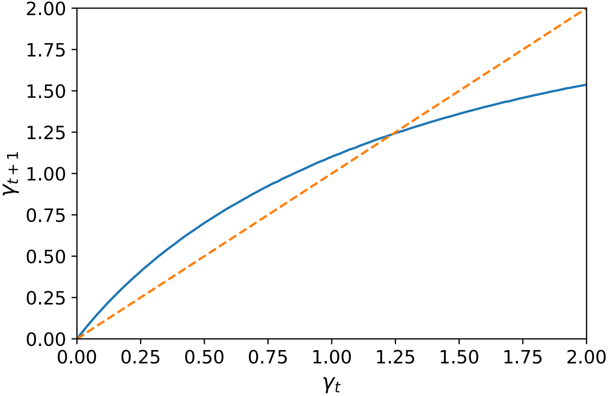
- State evolution with Bayes-optimal f_t^*

$$\gamma_{t+1} = \lambda^2 \{1 - \mathrm{mmse}(\gamma_t)\}$$

where $\mathrm{mmse}(\gamma) = \mathbb{E}\{(V - \mathbb{E}\{V \mid V + \sqrt{\gamma} G = s\})^2\}$

Fixed point of state evolution

$$\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v}\, \boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}, \quad P_V \sim \mathrm{Unif}\{1, -1\}, \quad \lambda = \sqrt{2}$$

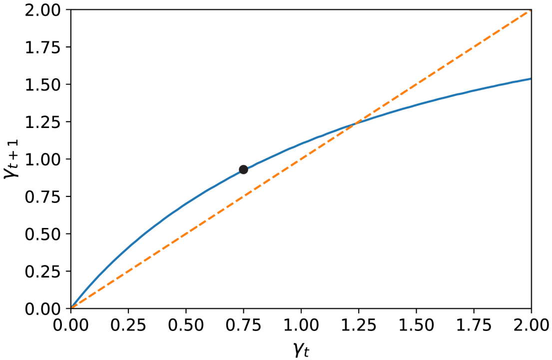


Recall $\boldsymbol{v}^0 \stackrel{\mathrm{d}}{=} \mu_0 V + \sigma_0 G$

- $\gamma_t = 0$ is an (unstable) fixed point: if $\gamma_0 = \frac{\mu_0^2}{\sigma_0^2} = 0$ then $\gamma_t = 0$ for all t!

Fixed point of state evolution

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}\,, \quad P_V \sim \mathrm{Unif}\{1,-1\}, \quad \lambda = \sqrt{2}$$

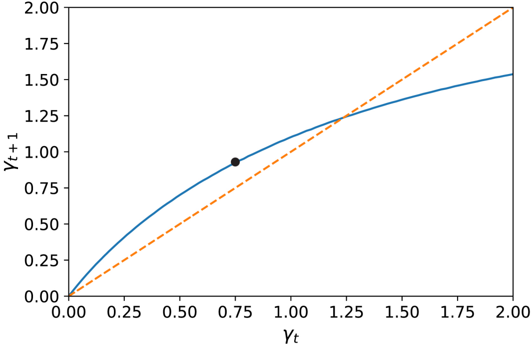


Recall $\boldsymbol{v}^0 \overset{\mathrm{d}}{=} \mu_0 V + \sigma_0 G$

- If $\gamma_0 \neq 0$, that is, $\boldsymbol{v}^0$ correlated with $\boldsymbol{v}$, AMP converges to the 'good' fixed point

Correlated initialization

Assuming correlated initialization often not realistic

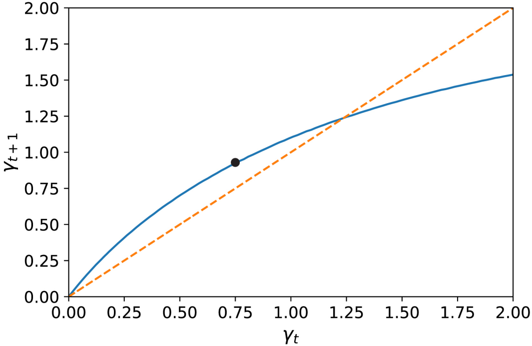


Natural initializer: $\hat{\varphi}$ the principal eigenvector of $\boldsymbol{A} = \frac{\lambda}{n} \boldsymbol{v} \, \boldsymbol{v}^\mathsf{T} + \boldsymbol{W}$

$$\text{Correlation} \quad : \qquad \frac{|\langle \hat{\varphi}, \, \boldsymbol{v} \rangle|}{\|\hat{\varphi}\| \|\boldsymbol{v}\|} \to \begin{cases} \sqrt{1 - \lambda^{-2}}, & \text{if } \lambda > 1, \\ 0, & \text{if } \lambda \in (0, 1] \end{cases}$$

Spectral initialization

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\,\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$



- Standard AMP theory assumes $\hat{\boldsymbol{v}}^0$ is independent of $\boldsymbol{A}$
- Spectral initialization requires special analysis [Montanari-Venkataramanan '21]
- With spectral initialization $\gamma_0 = 1 - \lambda^{-2}$ if $\lambda > 1$

Example: Two-point mixture

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

$$P_V = \varepsilon\,\delta_{a_+} + (1-\varepsilon)\delta_{a_-} \qquad a_+ = \sqrt{\frac{1-\varepsilon}{\varepsilon}} \quad a_- = -\sqrt{\frac{\varepsilon}{1-\varepsilon}}\,.$$

Run AMP with spectral initialization

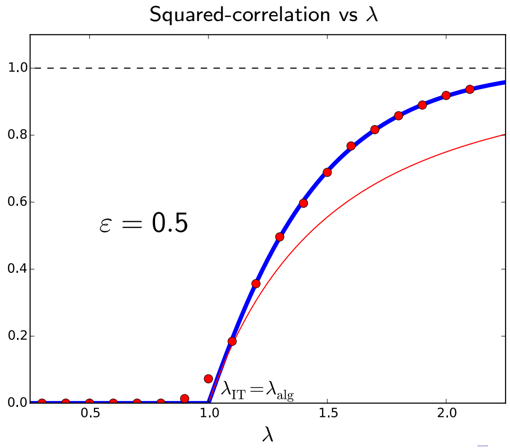
$$\gamma_{t+1} = \lambda^2\big\{1 - \mathrm{mmse}(\gamma_t)\big\}$$

- Can determine fixed point $\lim_{t\to\infty}\gamma_t$
- Initialization $\gamma_0 = 1 - \lambda^{-2}$

Example: Two-point mixture

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

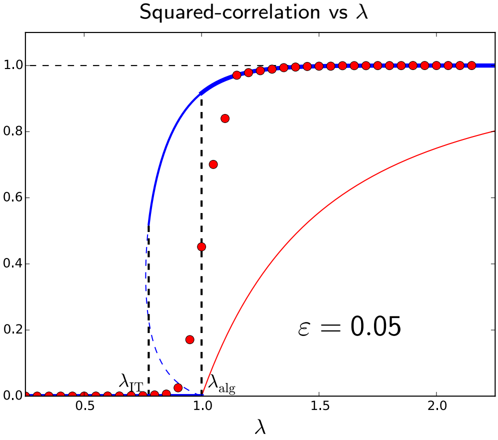
$$P_V = \varepsilon\,\delta_{a_+} + (1-\varepsilon)\delta_{a_-} \qquad a_+ = \sqrt{\frac{1-\varepsilon}{\varepsilon}} \quad a_- = -\sqrt{\frac{\varepsilon}{1-\varepsilon}}\,.$$



Example: Two-point mixture

$$\boldsymbol{A} = \frac{\lambda}{n}\boldsymbol{v}\boldsymbol{v}^{\mathsf{T}} + \boldsymbol{W}$$

$$P_V = \varepsilon\,\delta_{a_+} + (1-\varepsilon)\delta_{a_-} \qquad a_+ = \sqrt{\frac{1-\varepsilon}{\varepsilon}} \quad a_- = -\sqrt{\frac{\varepsilon}{1-\varepsilon}}\,.$$



Rank-k matrix estimation

Can generalize AMP to estimate rank-k signals

Symmetric:

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}} + \boldsymbol{W} \quad \in \mathbb{R}^{n \times n}$$

GOAL: To estimate the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ from $\boldsymbol{A}$

Non-symmetric:

$$\boldsymbol{A} = \sum_{i=1}^{k} \lambda_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}} + \boldsymbol{W} \quad \in \mathbb{R}^{m \times n}$$

GOAL: Estimate the singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$

Generalizations

Abstract AMP can be generalized to:

1. **Matrix-valued** iterates

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t - \mathrm{b}_t\,\boldsymbol{m}^{t-1}$$

   with $\boldsymbol{h}^t$, $\boldsymbol{m}^t$ being $n \times k$ matrices (k is fixed)

   Used for analyzing AMP for rank-k matrix estimation

Generalizations

Abstract AMP can be generalized to:

1. **Matrix-valued** iterates

$$\boldsymbol{m}^t = f_t(\boldsymbol{h}^t), \qquad \boldsymbol{h}^{t+1} = \boldsymbol{W}\,\boldsymbol{m}^t \,-\, \mathsf{b}_t\,\boldsymbol{m}^{t-1}$$

   with $\boldsymbol{h}^t$, $\boldsymbol{m}^t$ being $n \times k$ matrices (k is fixed)

   Used for analyzing AMP for rank-k matrix estimation

2. **Non-symmetric** i.i.d. Gaussian matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$.
   AMP defined via pairs of functions f_t, g_t for $t \geq 1$:

$$\boldsymbol{e}^t = \boldsymbol{A}\,f_t(\boldsymbol{h}^t) \,-\, \mathsf{b}_t\,g_{t-1}(\boldsymbol{e}^{t-1})$$
$$\boldsymbol{h}^{t+1} = \boldsymbol{A}^{\mathsf{T}}\,g_t(\boldsymbol{e}^t) \,-\, \mathsf{c}_t\,f_t(\boldsymbol{h}^t)$$

   - Empirical distributions of $\boldsymbol{e}^t \in \mathbb{R}^n$ and $\boldsymbol{h}^{t+1} \in \mathbb{R}^d$ converge to zero-mean Gaussians with variances given by SE
   - Used for analyzing AMP for linear models

Finite sample analysis of AMP

State evolution (SE) results in the large-but-finite n regime can be established under stronger assumptions

- [Rush, Venkataramanan '18]: Concentration inequality for AMP performance showing validity of SE for $\sim \frac{\log n}{\log \log n}$ iterations
- [Li, Wei '22], [Li, Fan, Wei '23]: Refined finite-sample SE for rank-1 AMP showing SE valid for $O(\frac{n}{\mathrm{polylog}(n)})$ iterations

Reference:
O. Feng, R. Venkataramanan, C. Rush, R. Samworth,
A unifying tutorial on Approximate Message Passing,
Foundations and Trends in Machine Learning, 2022
https://arxiv.org/abs/2105.02180

Free download during ISIT at
https://nowpublishers.com/conference/ISIT2023
Access code: ISIT-9502