Low-rank Matrix Estimation via Approximate Message Passing

Ramji Venkataramanan
Department of Engineering
(Joint work with Andrea Montanari)

October 31, 2018

CCIMI Seminar
Symmetric Low-rank Model

\[
A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \in \mathbb{R}^{n \times n}
\]

- \(\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k\) are deterministic scalars
- \(v_1, \ldots, v_k \in \mathbb{R}^n\) are orthonormal vectors ("spikes")
- \(W\) is a symmetric noise matrix

GOAL: To estimate the vectors \(v_1, \ldots, v_k\) from \(A\)
Rectangular Low-rank model

\[ A = \sum_{i=1}^{k} \lambda_i u_i v_i^T + W \in \mathbb{R}^{m \times n} \]

- \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \) are deterministic scalars
- \( u_1, \ldots, u_k \in \mathbb{R}^m \) are left singular vectors
  \( v_1, \ldots, v_k \in \mathbb{R}^n \) are right singular vectors
- \( W \) is a noise matrix

GOAL: Estimate the singular vectors \( u_1, \ldots, u_k \) and \( v_1, \ldots, v_k \)
Applications

\[ A \approx U V^T \]

\[ n \times d \approx n \times k \]

**Topic Modelling**

- Each row of \( A \) is a document
- Each row of \( V^T \) is a topic
- Each document convex combination of \( k \) topics

[Blei, Ng, Jordan '03]
Applications

$A \approx U V^T$

$n \times d \approx n \times k \times d$

Collaborative filtering

- $A$ 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
On the other hand, if \( i \notin S \), then \( D_i \sim k + \text{Binom}(n, 1/2) \). Hence, by a similar argument, we have
\[
\min_{i \in S} D_i \geq n^k r (1 + \epsilon) n \log_2 k^2 .
\]
(143)
The claim follows by using together the above, and selecting a suitable value \( \epsilon \).

Viceversa, for \( k < 2(1 - \epsilon) \log_2 n \) there exists an estimator \( b_S \) such that \( b_S = S \) with probability converging to one as \( n \to \infty \).

Proof. We will not present a complete proof but only sketch the fundamental reason for a threshold \( k \sim 2 \log_2 n \) and leave to the reader the task of filling the details.

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

<table>
<thead>
<tr>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(a) A random graph with a planted clique.

(b) The same graph, but with the vertices shuffled.

(c) Retrieving the clique in the shuffled graph.

On the other hand, if $i \not\in S$, then $D_i \sim \text{Binom}(n, 1/2)$. Hence, by a similar union bound

$$\min_{i \in S} D_i \geq n + k - 2\sqrt{n \log |S|}.$$  \hfill (143)

The claim follows by using together the above, and selecting a suitable value $\epsilon > 0$.

For $k$ too small, the problem becomes statistically intractable because the planted clique is not the unique clique of size $k$. Hence no estimator can distinguish between the set $S$ and another set $S_0$ that supports a different (purely random) clique. The next theorem characterizes this statistical threshold.

**Proposition 2.** Let $\epsilon > 0$ be fixed. Then, for $k < 2(1 - \epsilon) \log_2 n$ any estimator $\hat{b}_S$ is such that $\hat{b}_S = S$ with probability converging to one as $n \to \infty$.

Vice versa, for $k < 2(1 - \epsilon) \log_2 n$ there exists an estimator $\hat{b}_S$ such that $\hat{b}_S = S$ with probability converging to one as $n \to \infty$.

**Proof.** We will not present a complete proof but only sketch the fundamental reason for a threshold $k \sim 2 \log_2 n$ and leave to the reader the task of filling the details.
Hidden clique

(a) A random graph with a planted clique.

(b) The same graph, but with the vertices shuffled.

(c) Retrieving the clique in the shuffled graph.

On the other hand, if $i \in S$, then $D_i \sim \text{Binom}(n, \frac{1}{2})$. Hence, by a similar union bound, $\min_i 2 \leq S D_i \leq n + k^2 r (1 + \varepsilon \frac{\log_2 k}{2})$.

The claim follows by using together the above, and selecting a suitable value $\varepsilon \subset 0$.

For $k$ too small, the problem becomes statistically intractable because the planted clique is not the unique clique of size $k$. Hence no estimator can distinguish between the set $S$ and another set $S'$ that supports a different (purely random) clique. The next theorem characterizes this statistical threshold.

**Proposition 2.** Let $\varepsilon > 0$ be fixed. Then, for $k < 2(1 - \varepsilon) \log_2 n$, any estimator $b_S$ is such that $b_S = S$ with probability converging to one as $n \to \infty$.

Vice versa, for $k < 2(1 - \varepsilon) \log_2 n$, there exists an estimator $b_S$ such that $b_S = S$ with probability converging to one as $n \to \infty$.

**Proof.** We will not present a complete proof but only sketch the fundamental reason for a threshold $k \sim 2 \log_2 n$ and leave to the reader the task of filling the details.
For hidden clique $S$, adjacency matrix has the form
\[ A = 1_S 1_S^T + W \]

[Alon, Krivelivich, Sudakov ’98], . . .
Symmetric Spiked Model

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \quad \in \mathbb{R}^{n \times n} \]

- \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \) are deterministic scalars
- \( v_1, \ldots, v_k \in \mathbb{R}^n \) are orthonormal vectors ("spikes")
- \( W \sim \text{GOE}(n) \implies W \) symmetric with
  \( (W_{ii})_{i \leq n} \sim \text{i.i.d. } \mathcal{N}(0, \frac{2}{n}) \) and \( (W_{ij})_{i < j \leq n} \sim \text{i.i.d. } \mathcal{N}(0, \frac{1}{n}) \)
Spectrum of spiked matrix

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \]

Random matrix theory and the ‘BBAP’ phase transition:

- Bulk of eigenvalues of \( A \) in \([-2, 2]\) distributed according to Wigner’s semicircle
- Outlier eigenvalues corresponding to \(|\lambda_i|’s greater than 1: \n  \[ z_i \rightarrow \lambda_i + \frac{1}{\lambda_i} > 2 \]
- Eigenvectors \( \varphi_i \) corresponding to outliers \( z_i \) satisfy
  \[ |\langle \varphi_i, v_i \rangle| \rightarrow \sqrt{1 - \frac{1}{\lambda_i^2}} \]

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

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + \mathcal{W} \]

When \( v_i \)'s are unstructured, e.g., drawn uniformly at random from the unit sphere,

- Best estimator of \( v_i \) is the \( i \)th eigenvector \( \varphi_i \)
- If \( |\lambda_i| \geq 1 \), then \( |\langle v_i, \varphi_i \rangle| \to \sqrt{1 - \frac{1}{\lambda_i^2}} \)
Structural information

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \]

When \( v_i \)'s are unstructured, e.g., drawn uniformly at random from the unit sphere,

- Best estimator of \( v_i \) is the \( i \)th eigenvector \( \varphi_i \)
- If \( |\lambda_i| \geq 1 \), then \( |\langle v_i, \varphi_i \rangle| \to \sqrt{1 - \frac{1}{\lambda_i^2}} \)

But we often have structural information about \( v_i \)'s

- For example, \( v_i \)'s may be sparse, bounded, non-negative etc.
- Relevant in sparse PCA, non-negative PCA, hidden clique, community detection under stochastic block model, . . .
- Can improve on spectral methods
Prior on eigenvectors

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + \mathcal{W} \equiv V \Lambda V^T + \mathcal{W} \]

\[ V = [v_1 \ v_2 \ldots v_k] \quad \mathbb{R}^{n \times k} \]

If each row of \( V \) is \( \sim_{i.i.d} P_V \), then Bayes-optimal estimator (for squared error loss) is

\[ \hat{V}_{\text{Bayes}} = \mathbb{E} [V | A] \]

- Generally not computable
- Closed-form expressions for asymptotic Bayes risk

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

\[
A = \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mathbf{v}_i^T + \mathbf{W} \equiv \mathbf{V} \Lambda \mathbf{V}^T + \mathbf{W}
\]

- Convex relaxations generally do not achieve Bayes-optimal performance [Javanmard, Montanari, Ricci-Tersinghi '16]

- MCMC can approximate Bayes estimator, but can have large mixing time and hard to analyze
Computable estimators

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + \mathcal{W} \equiv V \Lambda V^T + \mathcal{W} \]

- Convex relaxations generally do not achieve Bayes-optimal performance [Javanmard, Montanari, Ricci-Tersinghi '16]

- MCMC can approximate Bayes estimator, but can have large mixing time and hard to analyze

In this talk
Approximate Message Passing (AMP) algorithm to estimate \( V \)
Rank one spiked model

\[ A = \frac{\lambda}{n} v v^T + W, \quad v \sim_{i.i.d.} P_V, \quad \mathbb{E} V^2 = 1 \]

Power iteration for principal eigenvector:

\[ x^{t+1} = A x^t, \text{ with } x^0 \text{ chosen at random} \]
Rank one spiked model

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W}, \quad \mathbf{v} \sim_{i.i.d.} P_{\mathbf{V}}, \quad \mathbb{E} V^2 = 1 \]

Power iteration for principal eigenvector:
\[ x^{t+1} = A x^t, \text{ with } x^0 \text{ chosen at random} \]

**AMP:**

\[ x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}), \quad b_t = \frac{1}{n} \sum_{i=1}^{n} f'_t(x^t_i) \]

- Non-linear function \( f_t: \mathbb{R} \rightarrow \mathbb{R} \) chosen based on structural info on \( \mathbf{v} \)
- **Memory term** ensures a nice distributional property for the iterates in high dimensions
- Can be derived via approximation of belief propagation equations
State evolution

\[ x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}), \quad \text{with } b_t = \frac{1}{n} \sum_{i=1}^{n} f'_t(x_i^t) \]

If we initialize with \( x^0 \) independent of \( A \), then as \( n \to \infty \):

\[ x^t \xrightarrow{\to} \mu_t v + \sigma_t g \]

\[ g \sim \text{i.i.d. } \mathcal{N}(0, 1), \text{ independent of } v \sim \text{i.i.d. } P_V \]

[Bayati, Montanari '11], [Rangan, Fletcher '12], [Deshpande, Montanari '14]
State evolution

\[ \mathbf{x}^{t+1} = \mathbf{A} f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}), \quad \text{with } b_t = \frac{1}{n} \sum_{i=1}^{n} f'_t(\mathbf{x}_i^t) \]

If we initialize with \( \mathbf{x}^0 \) independent of \( \mathbf{A} \), then as \( n \to \infty \):

\[ \mathbf{x}^t \longrightarrow \mu_t \mathbf{v} + \sigma_t \mathbf{g} \]

- \( \mathbf{g} \sim i.i.d. \ N(0, 1) \), independent of \( \mathbf{v} \sim i.i.d. \ P_V \)

- Scalars \( \mu_t, \sigma_t^2 \) recursively determined as

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

- Initialize with \( \mu_0 = \frac{1}{n} |\mathbb{E}[\mathbf{x}^0, \mathbf{v}]|, \quad \sigma_0^2 = \mathbb{E} V^2 - \mu_0^2 \)

[Bayati, Montanari '11], [Rangan, Fletcher '12], [Deshpande, Montanari '14]
Bayes-optimal AMP

Assuming \( x^t = \mu_t v + \sigma_t g \), choose \( f_t(y) = \lambda \mathbb{E}[V \mid \mu_t V + \sigma_t G = y] \)
Bayes-optimal AMP

Assuming \( x^t = \mu_t v + \sigma_t g \), choose \( f_t(y) = \lambda \mathbb{E}[V \mid \mu_t V + \sigma_t G = y] \)

State evolution becomes \( \gamma_{t+1} = \lambda^2 \{1 - \text{mmse}(\gamma_t)\} \) with \( \mu_t = \sigma_t^2 = \gamma_t \)

\[ P_V \sim \text{uniform}\{1, -1\}, \quad \lambda = \sqrt{2} \]

Initial value \( \gamma_0 \propto \frac{1}{n} |\mathbb{E}\langle x^0, v \rangle| \), what is \( \lim_{t \to \infty} \gamma_t? \)
If $\mathbb{E}\langle x^0, \nu \rangle = 0$, then $\gamma_t = 0$ is an (unstable) fixed point.

This is the case when $\nu$ has zero mean, as $x^0$ is independent of $\nu$. 
Spectral Initialization

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathcal{W}, \quad \lambda > 1 \]

- Compute \( \varphi_1 \), the principal eigenvector of \( A \)
- Run AMP with initialization \( x^0 = \sqrt{n} \varphi_1 \)
- \( \gamma_0 > 0 \) as \( \frac{1}{n} |\mathbb{E}\langle x^0, \mathbf{v} \rangle| \rightarrow \sqrt{1 - \lambda^{-2}} \)
AMP with spectral initialization

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W} \]

Existing AMP analysis does not apply for initialization \( \mathbf{x}^0 \) correlated with \( \mathbf{v} \)
Standard AMP analysis

With $\mathbf{W} \sim \text{GOE}(n)$, consider

$$h^{t+1} = \mathbf{W} f_t(h^t) - b_t f_{t-1}(h^{t-1})$$

Initialised with $h^0$ independent of $\mathbf{W}$. Let $\mathcal{V}_t := \{h^0, \ldots, h^t\}$

[Bolthausen '10], [Bayati-Montanari '11], [Rush-Venkataramanan '16]
Standard AMP analysis

With $\mathbf{W} \sim \text{GOE}(n)$, consider

$$h^{t+1} = \mathbf{W} f_t(h^t) - b_t f_{t-1}(h^{t-1})$$

Initialised with $h^0$ independent of $\mathbf{W}$. Let $\vartheta_t := \{h^0, \ldots, h^t\}$

- Conditional distribution

$$\mathbf{W} | \vartheta_t \stackrel{d}{=} \mathbb{E}[\mathbf{W} | \vartheta_t] + P_{\vartheta_t} \tilde{\mathbf{W}} P_{\vartheta_t}$$

---

[Bolthausen '10], [Bayati-Montanari '11], [Rush-Venkataramanan '16]
Standard AMP analysis

With $W \sim \text{GOE}(n)$, consider

$$h^{t+1} = W f_t(h^t) - b_t f_{t-1}(h^{t-1})$$

Initialised with $h^0$ independent of $W$. Let $\vartheta_t := \{h^0, \ldots, h^t\}$

- Conditional distribution

$$W|_{\vartheta_t} \overset{d}{=} \mathbb{E}[W | \vartheta_t] + P_{\vartheta_t} W \tilde{P}_{\vartheta_t}$$

- By induction, show that for $t \geq 0$:

$$h^{t+1} = \sum_{i=0}^{t} \alpha_i h^i + g_t + \Delta_t$$

[Bolthausen '10], [Bayati-Montanari '11], [Rush-Venkataramanan '16]
Standard AMP analysis

With $W \sim \text{GOE}(n)$, consider

$$h^{t+1} = W f_t(h^t) - b_t f_{t-1}(h^{t-1})$$

Initialised with $h^0$ independent of $W$. Let $\mathcal{V}_t := \{h^0, \ldots, h^t\}$

- Conditional distribution

$$W|\mathcal{V}_t \overset{d}{=} \mathbb{E}[W|\mathcal{V}_t] + P_{\mathcal{V}_t}^\perp \tilde{W} P_{\mathcal{V}_t}^\perp$$

- By induction, show that for $t \geq 0$:

$$h^{t+1} = \sum_{i=0}^{t} \alpha_i h^i + g_t + \Delta_t$$

$$h^{t+1} \overset{d}{\approx} \tau_t g$$

$$\tau_t^2 = \mathbb{E}[f_t(\tau_{t-1} G)^2], \quad \tau_0^2 = \|f(h^0)\|^2 / n$$

[Bolthausen '10], [Bayati-Montanari '11], [Rush-Venkataramanan '16]
AMP with spectral initialization

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathcal{W} \]

Let \((\varphi_1, z_1)\) be principal eigenvector and eigenvalue of \(A\)

\[ x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}) \]

initialised with \(x^0 = \sqrt{n} \varphi_1\)
AMP with spectral initialization

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W} \]

Let \((\varphi_1, z_1)\) be principal eigenvector and eigenvalue of \(A\)

\[ \mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}) \]

initialised with \(\mathbf{x}^0 = \sqrt{n} \varphi_1\)

We write

\[ A = z_1 \varphi_1 \varphi_1^\top + P_\perp \left( \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W} \right) P_\perp \]

\( \rightarrow P_\perp = I - \varphi_1 \varphi_1^\top \)
AMP with spectral initialization

\[ A = \frac{\lambda}{n} v v^\top + W \]

Let \((\varphi_1, z_1)\) be principal eigenvector and eigenvalue of \(A\)

\[ x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}) \]

initialised with \(x^0 = \sqrt{n} \varphi_1\)

Instead of \(A\), we will analyze AMP on

\[ \tilde{A} = z_1 \varphi_1 \varphi_1^\top + P^\perp \left( \frac{\lambda}{n} v v^\top + \tilde{W} \right) P^\perp \]

\(P^\perp = I - \varphi_1 \varphi_1^\top\)

\(\tilde{W} \sim \text{GOE}(n)\) is independent of \(W\)

1. Conditioned on \(z_1\) and \((\varphi_1^\top v)^2\) being close to limiting values, total variation distance between \(A\) and \(\tilde{A}\) is small
AMP with spectral initialization

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W} \]

Let \((\varphi_1, z_1)\) be principal eigenvector and eigenvalue of \(A\)

\[ x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}) \]

initialised with \(x^0 = \sqrt{n} \varphi_1\)

Instead of \(A\), we will analyze AMP on

\[ \tilde{A} = z_1 \varphi_1 \varphi_1^\top + \mathbf{P}^\perp \left( \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \tilde{\mathbf{W}} \right) \mathbf{P}^\perp \]

- \(\mathbf{P}^\perp = \mathbf{I} - \varphi_1 \varphi_1^\top\)
- \(\tilde{\mathbf{W}} \sim \text{GOE}(n)\) is independent of \(\mathbf{W}\)

1. Conditioned on \(z_1\) and \((\varphi_1^\top \mathbf{v})^2\) being close to limiting values, total variation distance between \(A\) and \(\tilde{A}\) is small
2. Analyze AMP on \(\tilde{A}\) by extending standard AMP analysis
Model assumptions

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W} \]

Let \( \mathbf{v} = \mathbf{v}(n) \in \mathbb{R}^n \) be a sequence such that the empirical distribution of entries of \( \mathbf{v}(n) \) converges weakly to \( P_V \),

\[ \psi(\mathbf{v}, \hat{\mathbf{v}}) = \frac{1}{n} \sum_{i=1}^{n} \psi(v_i, \hat{v}_i) \]

\( \psi \) assumed to be pseudo-Lipschitz:

\[ |\psi(x) - \psi(y)| \leq C \|x - y\|_2 (1 + \|x\|_2 + \|y\|_2) \]

\( \forall x, y \in \mathbb{R}^2 \)

\( L_2 \) loss, \( L_1 \) loss are both pseudo-Lipschitz.
Model assumptions

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W} \]

Let \( \mathbf{v} = \mathbf{v}(n) \in \mathbb{R}^n \) be a sequence such that the empirical distribution of entries of \( \mathbf{v}(n) \) converges weakly to \( P_V \),

Performance of any estimator \( \hat{\mathbf{v}} \) measured via loss function
\[ \psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R} : \]
\[ \psi(\mathbf{v}, \hat{\mathbf{v}}) = \frac{1}{n} \sum_{i=1}^{n} \psi(v_i, \hat{v}_i) \]

\( \psi \) assumed to be pseudo-Lipschitz:
\[ |\psi(\mathbf{x}) - \psi(\mathbf{y})| \leq C \|\mathbf{x} - \mathbf{y}\|_2 (1 + \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^2 \]

\( L_2 \) loss, \( L_1 \) loss are both pseudo-Lipschitz
Result for rank one case

\[ A = \frac{\lambda}{n} \mathbf{vv}^T + \mathbf{W} \]

**Theorem:** Let \( \lambda > 1 \). Consider the AMP

\[ \mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}) \]

- Assume \( f_t : \mathbb{R} \to \mathbb{R} \) is Lipschitz continuous
- Initialize with \( \mathbf{x}^0 = \sqrt{n} \varphi_1 \)

Then for any pseudo-Lipschitz loss function \( \psi \) and \( t \geq 0 \),

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(v_i, x_i^t) = \mathbb{E} \{ \psi(V, \mu_t V + \sigma_t G) \} \quad \text{a.s.} \]
Result for rank one case

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W} \]

**Theorem:** Let \( \lambda > 1 \). Consider the AMP

\[ \mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}) \]

- Assume \( f_t : \mathbb{R} \to \mathbb{R} \) is Lipschitz continuous
- Initialize with \( \mathbf{x}^0 = \sqrt{n} \varphi_1 \)

Then for any pseudo-Lipschitz loss function \( \psi \) and \( t \geq 0 \),

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(\mathbf{v}_i, \mathbf{x}_i^t) = \mathbb{E}\{\psi(\mathbf{V}, \mu_t \mathbf{V} + \sigma_t \mathbf{G})\} \quad \text{a.s.} \]

**State evolution parameters:** \( \mu_0 = \sqrt{1 - \lambda^{-2}}, \quad \sigma_0 = 1/\lambda, \)

\[ \mu_{t+1} = \lambda \mathbb{E}[V f_t(\mu_t V + \sigma_t G)], \quad \sigma_{t+1}^2 = \mathbb{E}[f_t(\mu_t V + \sigma_t G)^2], \]
Proof Sketch

True vs conditional model

\[ A = \frac{\lambda}{n} \nu \nu^T + W \]

\[ \tilde{A} = z_1 \varphi_1 \varphi_1^T + P^\perp \left( \frac{\lambda}{n} \nu \nu^T + \tilde{W} \right) P^\perp \]

Lemma

For \((z_1, \varphi_1) \in \left\{ |z_1 - (\lambda + \lambda^{-1})| \leq \epsilon, \quad (\varphi_1^T \nu)^2 \geq 1 - \lambda^{-2} - \epsilon \right\}\),

we have

\[ \sup_{(z, \Phi) \in E_\epsilon} \left\| \mathbb{P}(A \in \cdot | z_1, \varphi_1) - \mathbb{P}(\tilde{A} \in \cdot | z_1, \varphi_1) \right\|_{TV} \leq \frac{1}{c(\epsilon)} e^{-nc(\epsilon)} \]
AMP on conditional model

\[ \tilde{A} = z_1 \varphi_1 \varphi_1^T + P^\perp \left( \frac{\lambda}{n} \nu \nu^T + \tilde{W} \right) P^\perp \]

AMP with \( \tilde{A} \) instead of \( A \):

\[ \tilde{x}^{t+1} = \tilde{A} f(\tilde{x}^t; t) - b_t f(\tilde{x}^{t-1}; t-1), \quad \tilde{x}^0 = \sqrt{n} \varphi_1 \]

Analyze using existing AMP analysis + results from random matrix theory
Bayes-optimal AMP

$$A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W}$$

$$\mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1})$$

▸ Bayes-optimal choice
  $$f_t(y) = \lambda \mathbb{E}(V | \gamma_t V + \sqrt{\gamma_t} G = y)$$

▸ State evolution:
  $$\gamma_{t+1} = \lambda^2 \{1 - \text{mmse}(\gamma_t)\}, \quad \gamma_0 = \lambda^2 - 1$$
  where
  $$\text{mmse}(\gamma) = \mathbb{E}\{ [V - \mathbb{E}(V | \sqrt{\gamma} V + G)]^2 \}$$

▸ $$\mu_t = \sigma_t^2 = \gamma_t$$
Bayes-optimal AMP

\[ A = \frac{\lambda}{n} \mathbf{vv}^T + \mathbf{W} \]

Let \( \gamma_{\text{AMP}}(\lambda) \) be the \textit{smallest} strictly positive solution of

\[ \gamma = \lambda^2 [1 - \text{mmse}(\gamma)]. \quad (1) \]

Then the AMP estimate \( \hat{x}^t = f_t(x^t) \) achieves

\[ \lim_{t \to \infty} \lim_{n \to \infty} \min_{s \in \{+1, -1\}} \frac{1}{n} \| \hat{x}^t - sv \|_2^2 = \frac{1 - \frac{\gamma_{\text{AMP}}(\lambda)}{\lambda^2}}{2} \]
Bayes-optimal AMP

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathbf{W} \]

Let \( \gamma_{\text{AMP}}(\lambda) \) be the \textit{smallest} strictly positive solution of

\begin{equation}
\gamma = \lambda^2 [1 - \text{mmse}(\gamma)].
\end{equation}

Then the AMP estimate \( \hat{x}^t = f_t(x^t) \) achieves

\[
\text{Overlap} : \lim_{t \to \infty} \lim_{n \to \infty} \frac{|\langle \hat{x}^t, \mathbf{v} \rangle|}{\| \hat{x}^t \|_2 \| \mathbf{v} \|_2} = \frac{\sqrt{\gamma_{\text{AMP}}(\lambda)}}{\lambda}
\]
Bayes-optimal AMP

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathcal{W} \]

Let \( \gamma_{\text{AMP}}(\lambda) \) be the *smallest* strictly positive solution of

\[ \gamma = \lambda^2 [1 - \text{mmse}(\gamma)]. \tag{1} \]

Then the AMP estimate \( \hat{\mathbf{x}}^t = f_t(\mathbf{x}^t) \) achieves

\[
\text{Overlap : } \lim_{t \to \infty} \lim_{n \to \infty} \frac{|\langle \hat{\mathbf{x}}^t, \mathbf{v} \rangle|}{\|\hat{\mathbf{x}}^t\|_2 \|\mathbf{v}\|_2} = \frac{\sqrt{\gamma_{\text{AMP}}(\lambda)}}{\lambda}
\]

Bayes-optimal overlap [Miolane-Lelarge '16]

For (almost) all \( \lambda > 0 \)

\[
\lim_{n \to \infty} \sup_{\hat{x}(\cdot)} \frac{|\langle \hat{\mathbf{x}}^t, \mathbf{v} \rangle|}{\|\hat{\mathbf{x}}^t\|_2 \|\mathbf{v}\|_2} = \frac{\sqrt{\gamma_{\text{Bayes}}(\lambda)}}{\lambda}
\]

\( \gamma_{\text{Bayes}}(\lambda) \): fixed point of (1) that maximizes a free-energy functional
Example: Two-point mixture

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W} \]

\[ P_V = \varepsilon \delta_{a_+} + (1 - \varepsilon) \delta_{a_-} \]

\[ a_+ = \sqrt{\frac{1 - \varepsilon}{\varepsilon}} \quad a_- = -\sqrt{\frac{\varepsilon}{1 - \varepsilon}}. \]
Example: Two-point mixture

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^T + \mathbf{W} \]

\[ P_V = \varepsilon \delta_{a_+} + (1 - \varepsilon) \delta_{a_-} \quad a_+ = \sqrt{\frac{1 - \varepsilon}{\varepsilon}} \quad a_- = -\sqrt{\frac{\varepsilon}{1 - \varepsilon}}. \]

\[ \varepsilon = 0.05 \]
Confidence intervals

\[
A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathcal{W}
\]

AMP: \( x^{t+1} = A f_t(x^t) - b_t f_{t-1}(x^{t-1}) \)

- Convergence result tells us that \( x^t \approx \mu_t \mathbf{v} + \sigma_t \mathbf{g} \)
Confidence intervals

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathcal{W} \]

AMP: \[ \mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}) \]

- Convergence result tells us that \[ \mathbf{x}^t \approx \mu_t \mathbf{v} + \sigma_t \mathbf{g} \]
- State evolution parameters can be estimated:

\[ \hat{\sigma}^2_t \equiv \frac{1}{n} \| f_{t-1}(\mathbf{x}^{t-1}) \|_2^2, \]

\[ \hat{\mu}^2_t \equiv \frac{1}{n} \| \mathbf{x}^t \|_2^2 - \frac{1}{n} \| f_{t-1}(\mathbf{x}^{t-1}) \|_2^2. \]
Confidence intervals

\[ A = \frac{\lambda}{n} \mathbf{v} \mathbf{v}^\top + \mathcal{W} \]

AMP: \[ \mathbf{x}^{t+1} = A f_t(\mathbf{x}^t) - b_t f_{t-1}(\mathbf{x}^{t-1}) \]

- Convergence result tells us that \( \mathbf{x}^t \approx \mu_t \mathbf{v} + \sigma_t \mathbf{g} \)
- State evolution parameters can be estimated:
  \[ \hat{\sigma}^2_t \equiv \frac{1}{n} \| f_{t-1}(\mathbf{x}^{t-1}) \|_2^2, \]
  \[ \hat{\mu}^2_t \equiv \frac{1}{n} \| \mathbf{x}^t \|_2^2 - \frac{1}{n} \| f_{t-1}(\mathbf{x}^{t-1}) \|_2^2. \]

- Confidence intervals for coverage level \( (1 - \alpha) \):
  \[ \hat{l}_i(\alpha; t) = \left[ \frac{1}{\hat{\mu}_t} x_i^t - \frac{\hat{\sigma}_t}{\hat{\mu}_t} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right), \frac{1}{\hat{\mu}_t} x_i^t + \frac{\hat{\sigma}_t}{\hat{\mu}_t} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right] \]

- Bayes-optimal choice minimizes length of confidence intervals, but requires knowledge of the empirical distribution of \( \mathbf{v} \)
For $1 \leq i \leq n$,

$$
\hat{l}_i(\alpha; t) = \left[ \frac{1}{\hat{\mu}_t} x_i^t - \frac{\hat{\sigma}_t}{\hat{\mu}_t} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right), \quad \frac{1}{\hat{\mu}_t} x_i^t + \frac{\hat{\sigma}_t}{\hat{\mu}_t} \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) \right]
$$

**Corollary:**

$$
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(v_i \in \hat{l}_i(\alpha; t)) = 1 - \alpha \quad \text{almost surely.}
$$
General case

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \equiv V \Lambda V^T + W. \]

- Assume \( k_* \) eigenvectors corresponding to outliers \( |\lambda_i| > 1 \)
- Outliers can be estimated from \( A \), as \( z_i \rightarrow \lambda_i + \lambda_i^{-1} \)
- Assume empirical distribution of rows of \( V \sim P_V \)
General case

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \equiv V \Lambda V^T + W. \]

- Assume \( k_* \) eigenvectors corresponding to outliers \( |\lambda_i| > 1 \)
- Outliers can be estimated from \( A \), as \( z_i \rightarrow \lambda_i + \lambda_i^{-1} \)
- Assume empirical distribution of rows of \( V \sim P_V \)
General case

\[ A = \sum_{i=1}^{k} \lambda_i v_i v_i^T + W \equiv V \Lambda V^T + W. \]

- Assume \( k_* \) eigenvectors corresponding to outliers \( |\lambda_i| > 1 \)
- Outliers can be estimated from \( A \), as \( z_i \to \lambda_i + \lambda_i^{-1} \)
- Assume empirical distribution of rows of \( V \sim P_V \)

**AMP:**

\[ x^{t+1} = A f_t(x^t) - f_{t-1}(x^{t-1}) B_t^T \]

- \( x^t \in \mathbb{R}^{n \times k_*} \) are estimates of the outlier eigenvectors
- \( f(\cdot; t) : \mathbb{R}^{k_*} \to \mathbb{R}^{k_*} \) applied row by row
- \( B_t = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial f_t}{\partial x} (x^t_i) \), where \( \frac{\partial f_t}{\partial x} \) is Jacobian of \( f(\cdot; t) \)

Spectral initialization: \( x^0 = [\sqrt{n} \varphi_1 | \ldots | \sqrt{n} \varphi_{k_*}] \)
The goal of community detection in this case is to obtain the right graph (with five communities) from the left graph (scrambled) up to some level of accuracy. In such a context, community detection may be called graph clustering. In general, communities may not only refer to denser clusters but more generally to groups of vertices that behave similarly.

The goal of this monograph is to describe recent developments aiming at answering these questions in the context of block models. Block models are a family of random graphs with planted clusters. The “mother model” is the stochastic block model (SBM), which has been widely employed as a canonical model for community detection. It is arguably the simplest model of a graph with communities (see definitions in the next section). Since the SBM is a generative model, it benefits from a ground truth for the communities, which allows to consider the previous questions in a formal context. As any model, it is not necessarily realistic, but it is insightful - judging for example from the powerful algorithms that have emerged from its study.

In a sense, the SBM plays a similar role to the discrete memoryless channel (DMC) in information theory. While the task of modelling external noise may be more amenable to simplifications that real data sets, the SBM captures some of the key bottleneck phenomena for community detection and admits many possible refinements that improve the fit to real data. Our focus will be here on the fundamental understanding of the core SBM, without diving too much into the refined extensions.

The core SBM is defined as follows. For positive integers $k, n$, a probability $p_{ij}$ is assigned to each pair of vertices, where $p_{ij} = \theta$ if $i$ and $j$ are in the same community and $p_{ij} = \phi$ otherwise. The graph is then generated by connecting each pair of vertices with probability $p_{ij}$.

Image from Community detection and stochastic block models by E. Abbe
Block model with multiple communities

Figure 1: The above two graphs are the same graph re-organized and drawn from the SBM model with 1000 vertices, 5 balanced communities, within-cluster probability of $1/50$ and across-cluster probability of $1/1000$. The goal of community detection in this case is to obtain the right graph (with five communities) from the left graph (scrambled) up to some level of accuracy. In such a context, community detection may be called graph clustering. In general, communities may not only refer to denser clusters but more generally to groups of vertices that behave similarly.

The goal of this monograph is to describe recent developments aiming at answering these questions in the context of block models. Block models are a family of random graphs with planted clusters. The "mother model" is the stochastic block model (SBM), which has been widely employed as a canonical model for community detection. It is arguably the simplest model of a graph with communities (see definitions in the next section). Since the SBM is a generative model, it benefits from a ground truth for the communities, which allows to consider the previous questions in a formal context. As any model, it is not necessarily realistic, but it is insightful - judging for example from the powerful algorithms that have emerged from its study.

In a sense, the SBM plays a similar role to the discrete memoryless channel (DMC) in information theory. While the task of modelling external noise may be more amenable to simplifications that real data sets, the SBM captures some of the key bottleneck phenomena for community detection and admits many possible refinements that improve the fit to real data. Our focus will be here on the fundamental understanding of the core SBM, without diving too much into the refined extensions.

The core SBM is defined as follows. For positive integers $k$, $n$, a probability

Wish to recover vertex labels (colours) from adjacency matrix

Image from *Community detection and stochastic block models* by E. Abbe
A closely related model . . .

- Let \( \sigma = (\sigma_1, \ldots, \sigma_n) \) be vector of vertex labels
- Labels \( \sigma_i \) uniformly distributed in \( \{1, 2, 3\} \)
- Consider the \( n \times n \) matrix \( \mathbf{A}_0 \) with entries

\[
\mathbf{A}_{0,ij} = \begin{cases} 
2/n & \text{if } \sigma_i = \sigma_j \\
-1/n & \text{otherwise.}
\end{cases}
\]

- \( \mathbf{A}_0 \) is an orthogonal projector onto a two-dimensional subspace \( \Rightarrow \mathbf{A}_0 \) is rank 2
A closely related model . . .

- Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be vector of vertex labels
- Labels $\sigma_i$ uniformly distributed in $\{1, 2, 3\}$
- Consider the $n \times n$ matrix $A_0$ with entries
  \[
  A_{0,ij} = \begin{cases} 
  2/n & \text{if } \sigma_i = \sigma_j \\
  -1/n & \text{otherwise}.
  \end{cases}
  \]

- $A_0$ is an orthogonal projector onto a two-dimensional subspace $\Rightarrow A_0$ is rank 2

Wish to estimate $A_0$ from noisy version:

\[ A = \lambda A_0 + W \]

- Degenerate eigenvalues: $\lambda_1 = \lambda_2 = \lambda > 1$
- $W \sim \text{GOE}(n)$
- $A$ similar to rescaled adjacency matrix in block model
AMP

\[ A = \frac{\lambda}{n} V V^T + W \]

Spectral initialization: \( x^0 = [\sqrt{n} \varphi_1 \sqrt{n} \varphi_2] \)

Main result

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(V_i, x^t_i) = \mathbb{E}\{\psi(V, M_t V + Q_t^{1/2} G)\} \quad \text{a.s.}
\]
AMP

\[ A = \frac{\lambda}{n} V V^T + W \]

Spectral initialization: \( x^0 = [\sqrt{n}\varphi_1 \quad \sqrt{n}\varphi_2] \)

**Main result**

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \psi(V_i, x_i^t) = \mathbb{E}\{\psi(V, M_t V + Q_t^{1/2} G)\} \quad \text{a.s.}
\]

**State evolution:**

\( M_0 = (x^0)^T V \) and \( Q_0 = \lambda^{-1} I \in \mathbb{R}^{2 \times 2} \)

\[
M_{t+1} = \lambda \mathbb{E}\left\{ f_t(M_t V + Q_t^{1/2} G) V^T \right\},
\]

\[
Q_{t+1} = \mathbb{E}\left\{ f_t(M_t V + Q_t^{1/2} G) f_t(M_t V + Q_t^{1/2} G)^T \right\}.
\]

Since \( V V^T = V R R^T V^T \) for any \( 2 \times 2 \) rotation matrix \( R \)

\( \Rightarrow \) state evolution starts from a *random* initial condition

\[
M_0 = (x^0)^T V \overset{d}{=} \sqrt{1 - \lambda^2} R
\]
\[ A = \frac{\lambda}{n} \mathbf{V} \mathbf{V}^T + \mathbf{W} \]

Gaussian block model with \( \lambda = 1.5, \quad n = 6000 \)
Summary

\[ A = \Lambda \Lambda V^T + W \]

AMP with spectral initialization

- Distributional property of the iterates gives succinct performance characterization via state evolution
- Can be used to construct confidence intervals
- AMP can achieve Bayes-optimal accuracy

Extensions and Future work

- Can be extended to rectangular low-rank matrix model:
  \[ A = U \Sigma V^T + W \]
- AMP with spectral initialization for generalized linear models, e.g., phase retrieval