Approximate Message Passing for Matrix Regression

Thesis Defense

Nelvin Tan

Signal Processing and Communications Laboratory,
University of Cambridge

October 2024

Acknowledgements

The thesis consists of joint work with:

- Ramji Venkataramanan (University of Cambridge)
- Pablo Pascual Cobo (University of Cambridge)
- Jonathan Scarlett (National University of Singapore)

Standard Generalized Linear Model (GLM)

- **Model.** $y_i = q(\beta^\top X_{i,:}, \psi_i)$ for $i \in [n] := \{1, \dots, n\}$.
  - $X_{i,:}$ is the ith row of the design matrix $X \in \mathbb{R}^{n \times p}$.
  - y_i is the ith entry of the observation $y \in \mathbb{R}^n$.
  - ψ_i is the ith entry of the noise $\psi \in \mathbb{R}^n$.
  - $\beta \in \mathbb{R}^p$ is the target signal.
  - $q : \mathbb{R}^2 \to \mathbb{R}$ is a known function.
- **Goal.** Given X and y, estimate β.
- **Applications:**
  - Linear model: Compressed sensing and sparse regression codes.
  - Phase retrieval: Optics and X-ray crystallography.
  - Logistic regression: Fraud detection and disease prediction.

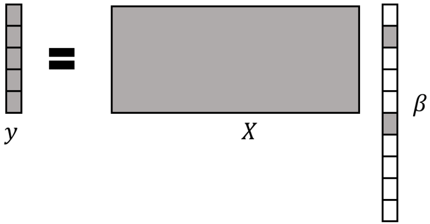
Standard Generalized Linear Model (GLM)

- **Model.** $y_i = q(\beta^\top X_{i,:}, \psi_i)$ for $i \in [n] := \{1, \dots, n\}$.
  - $X_{i,:}$ is the ith row of the design matrix $X \in \mathbb{R}^{n \times p}$.
  - y_i is the ith entry of the observation $y \in \mathbb{R}^n$.
  - ψ_i is the ith entry of the noise $\psi \in \mathbb{R}^n$.
  - $\beta \in \mathbb{R}^p$ is the target signal.
  - $q : \mathbb{R}^2 \to \mathbb{R}$ is a known function.
- **Goal.** Given X and y, estimate β.
- **Applications:**
  - Linear model: Compressed sensing and sparse regression codes.
  - Phase retrieval: Optics and X-ray crystallography.
  - Logistic regression: Fraud detection and disease prediction.

Standard Generalized Linear Model (GLM)

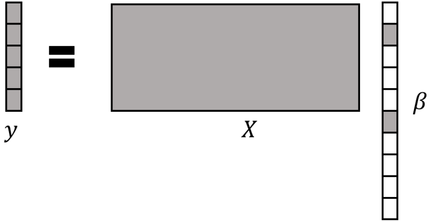
- **Model.** $y_i = q(\beta^\top X_{i,:}, \psi_i)$ for $i \in [n] := \{1, \dots, n\}$.
  - $X_{i,:}$ is the ith row of the design matrix $X \in \mathbb{R}^{n \times p}$.
  - y_i is the ith entry of the observation $y \in \mathbb{R}^n$.
  - ψ_i is the ith entry of the noise $\psi \in \mathbb{R}^n$.
  - $\beta \in \mathbb{R}^p$ is the target signal.
  - $q : \mathbb{R}^2 \to \mathbb{R}$ is a known function.
- **Goal.** Given X and y, estimate β.
- **Applications:**
  - Linear model: Compressed sensing and sparse regression codes.
  - Phase retrieval: Optics and X-ray crystallography.
  - Logistic regression: Fraud detection and disease prediction.

High-Dimensional Regime



- Motivated by massive data sets in recent times.
- #features is of comparable size, or larger, than #observations.
- Specifically, $n/p \to \delta \in (0, \infty)$ as $n, p \to \infty$.
- Common approaches when data has some form of structure:
  - Feature selection (e.g., forward/backward selection) then estimate.
  - Feature reduction (e.g., principal component analysis) then estimate.
  - Incorporate signal structure into estimation (this thesis).

High-Dimensional Regime



- Motivated by massive data sets in recent times.
- #features is of comparable size, or larger, than #observations.
- Specifically, $n/p \to \delta \in (0, \infty)$ as $n, p \to \infty$.
- Common approaches when data has some form of structure:
  - Feature selection (e.g., forward/backward selection) then estimate.
  - Feature reduction (e.g., principal component analysis) then estimate.
  - Incorporate signal structure into estimation (this thesis).

Matrix Generalized Linear Model (GLM)

$$Y = \begin{bmatrix} -Y_{1,:}- \\ \vdots \\ -Y_{n,:}- \end{bmatrix}, \quad X = \begin{bmatrix} -X_{1,:}- \\ \vdots \\ -X_{n,:}- \end{bmatrix}, \quad \Psi = \begin{bmatrix} -\Psi_{1,:}- \\ \vdots \\ -\Psi_{n,:}- \end{bmatrix}.$$

- **Model.** $Y_{i,:} = q(B^\top X_{i,:}, \Psi_{i,:})$ for $i \in [n]$.
  - Observation $Y \in \mathbb{R}^{n \times L_{\text{out}}}$.
  - Auxiliary matrix $\Psi \in \mathbb{R}^{n \times L_\Psi}$.
  - Matrix signal $B \in \mathbb{R}^{p \times L}$.
  - $q : \mathbb{R}^L \times \mathbb{R}^{L_\Psi} \to \mathbb{R}^{L_{\text{out}}}$ is a known function.
- **Goal.** Given X and Y, estimate B.
- **Applications.** Mixed regression, quantitative group testing, pooled data.

Matrix Generalized Linear Model (GLM)

$$Y = \begin{bmatrix} -Y_{1,:}- \\ \vdots \\ -Y_{n,:}- \end{bmatrix}, \quad X = \begin{bmatrix} -X_{1,:}- \\ \vdots \\ -X_{n,:}- \end{bmatrix}, \quad \Psi = \begin{bmatrix} -\Psi_{1,:}- \\ \vdots \\ -\Psi_{n,:}- \end{bmatrix}.$$

- **Model.** $Y_{i,:} = q(B^\top X_{i,:}, \Psi_{i,:})$ for $i \in [n]$.
  - Observation $Y \in \mathbb{R}^{n \times L_{\text{out}}}$.
  - Auxiliary matrix $\Psi \in \mathbb{R}^{n \times L_\Psi}$.
  - Matrix signal $B \in \mathbb{R}^{p \times L}$.
  - $q : \mathbb{R}^L \times \mathbb{R}^{L_\Psi} \to \mathbb{R}^{L_{\text{out}}}$ is a known function.
- **Goal.** Given X and Y, estimate B.
- **Applications.** Mixed regression, quantitative group testing, pooled data.

Approximate Message Passing for Mixed Regression

Approximate Message Passing (AMP)

- **AMP algorithm:**

$$\Theta^k = X\widehat{B}^k - \widehat{R}^{k-1}(F^k)^\top, \quad \widehat{R}^k = g_k(\Theta^k, Y),$$

$$B^{k+1} = X^\top \widehat{R}^k - \widehat{B}^k (C^k)^\top, \quad \widehat{B}^{k+1} = f_{k+1}(B^{k+1}).$$

  - $C^k = \frac{1}{n}\sum_{i=1}^n g_k'(\Theta_i^k, Y_i)$ and $F^{k+1} = \frac{1}{n}\sum_{j=1}^p f_{k+1}'(B_j^{k+1})$.
  - Iteratively produces estimates $\widehat{B}^k$ of B.
  - Denoisers g_k and f_{k+1} are Lipschitz and applied component wise.

- **Main assumptions:**
  - As $n, p \to \infty$, we have $n/p = \delta > 0$;
  - $X_{i,:} \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$;
  - Empirical distribution of the rows of B converge to $\bar{B}$.

Approximate Message Passing (AMP)

- **AMP algorithm:**

$$\Theta^k = X\widehat{B}^k - \widehat{R}^{k-1}(F^k)^\top, \quad \widehat{R}^k = g_k(\Theta^k, Y),$$

$$B^{k+1} = X^\top \widehat{R}^k - \widehat{B}^k(C^k)^\top, \quad \widehat{B}^{k+1} = f_{k+1}(B^{k+1}).$$

  - $C^k = \frac{1}{n}\sum_{i=1}^n g_k'(\Theta_i^k, Y_i)$ and $F^{k+1} = \frac{1}{n}\sum_{j=1}^p f_{k+1}'(B_j^{k+1})$.
  - Iteratively produces estimates $\widehat{B}^k$ of B.
  - Denoisers g_k and f_{k+1} are Lipschitz and applied component wise.

- **Main assumptions:**
  - As $n, p \to \infty$, we have $n/p = \delta > 0$;
  - $X_{i,:} \sim_{\text{i.i.d.}} \mathcal{N}(0, I_p/n)$;
  - Empirical distribution of the rows of B converge to $\bar{B}$.

State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$\big(B, B^{k+1}\big) \to \big(\bar{B}, \bar{B}^{k+1}\big), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters $\mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L \times L}$ are defined as $(g_{k-1} := g_{k-1}(Z^{k-1}, q(Z.\bar{\Psi})))$
$$\mathrm{M}_B^k = \mathbb{E}\big[\partial_Z g_{k-1}\big] \text{ and } \mathrm{T}_B^k = \mathbb{E}\big[g_{k-1} g_{k-1}^\top\big],$$
where Z and Z^k are the limiting distributions of $\Theta = XB$ and Θ^k.
- **Choice of denoisers.** We propose:
$$f_k(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k \bar{B} + G_B^k = s]$$
$$g_k(u, y) = \mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u],$$
which minimizes the effective noise in each iteration.
- **Proof idea.** Via reduction of our AMP to an abstract AMP [Feng et al. 2022].

State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of

$$\big(B, B^{k+1}\big) \to \big(\bar{B}, \bar{B}^{k+1}\big), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$

where $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters $\mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L\times L}$ are defined as $(g_{k-1} := g_{k-1}(Z^{k-1}, q(Z, \bar{\Psi})))$

$$\mathrm{M}_B^k = \mathbb{E}\big[\partial_Z g_{k-1}\big] \text{ and } \mathrm{T}_B^k = \mathbb{E}\big[g_{k-1} g_{k-1}^\top\big],$$

where Z and Z^k are the limiting distributions of $\Theta = XB$ and Θ^k.

- **Choice of denoisers.** We propose:

$$f_k(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k \bar{B} + G_B^k = s]$$

$$g_k(u, y) = \mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u],$$

which minimizes the effective noise in each iteration.

- **Proof idea.** Via reduction of our AMP to an abstract AMP [Feng et al. 2022].

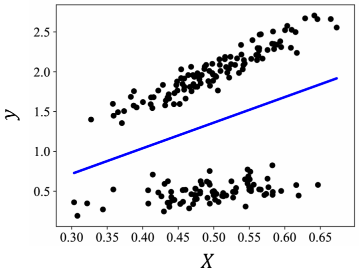
State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$\big(B, B^{k+1}\big) \to \big(\bar{B}, \bar{B}^{k+1}\big), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters $\mathrm{M}_B^k, \mathrm{T}_B^k \in \mathbb{R}^{L\times L}$ are defined as $\big(g_{k-1} := g_{k-1}(Z^{k-1}, q(Z.\bar{\Psi}))\big)$
$$\mathrm{M}_B^k = \mathbb{E}\big[\partial_Z g_{k-1}\big] \text{ and } \mathrm{T}_B^k = \mathbb{E}\big[g_{k-1}g_{k-1}^\top\big],$$
where Z and Z^k are the limiting distributions of $\Theta = XB$ and Θ^k.

- **Choice of denoisers.** We propose:
$$f_k(s) = \mathbb{E}[\bar{B} \mid \mathrm{M}_B^k\bar{B} + G_B^k = s]$$
$$g_k(u, y) = \mathbb{E}[Z \mid Z^k = u, \bar{Y} = y] - \mathbb{E}[Z \mid Z^k = u],$$
which minimizes the effective noise in each iteration.

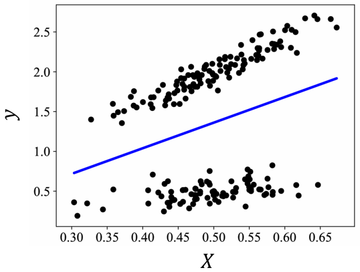
- **Proof idea.** Via reduction of our AMP to an abstract AMP [Feng et al. 2022].

Numerical Simulations – Mixed Linear Regression



- **Model.** $y_i = \langle X_{i,:}, \beta^{(1)} \rangle c_{i1} + \cdots + \langle X_{i,:}, \beta^{(L)} \rangle c_{iL} + \epsilon_i$ for $i \in [n]$
  - Observation $y \in \mathbb{R}^n$, signal $\beta^{(l)} \in \mathbb{R}^p$, noise $\epsilon \in \mathbb{R}^n$.
  - Latent variables $c_{i1}, \dots, c_{iL} \in \{0, 1\}$ such that $\sum_{l=1}^{L} c_{il} = 1$.
  - Used when data comes from unknown sub-populations.
- **Reduction to matrix GLM.** $B = (\beta^{(1)}, \dots, \beta^{(L)})$, $\Psi_{i,:} = (c_{i1}, \dots, c_{iL}, \epsilon_i)$
- **Other models.** Max-affine regression, mixed GLM, mixture-of-experts

Numerical Simulations – Mixed Linear Regression



- **Model.** $y_i = \langle X_{i,:}, \beta^{(1)} \rangle c_{i1} + \cdots + \langle X_{i,:}, \beta^{(L)} \rangle c_{iL} + \epsilon_i$ for $i \in [n]$
  - Observation $y \in \mathbb{R}^n$, signal $\beta^{(l)} \in \mathbb{R}^p$, noise $\epsilon \in \mathbb{R}^n$.
  - Latent variables $c_{i1}, \ldots, c_{iL} \in \{0, 1\}$ such that $\sum_{l=1}^{L} c_{il} = 1$.
  - Used when data comes from unknown sub-populations.
- **Reduction to matrix GLM.** $B = (\beta^{(1)}, \ldots, \beta^{(L)})$, $\Psi_{i,:} = (c_{i1}, \ldots, c_{iL}, \epsilon_i)$
- **Other models.** Max-affine regression, mixed GLM, mixture-of-experts

Numerical Simulations

- **Two-component case for simulation:**

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i,$$

where $c_i \sim_{\text{i.i.d.}}$ Bernoulli(α), with $\alpha \in (0,1)$, and $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

- **Normalized squared correlation.** #Categories $= L$

$$\underbrace{\frac{\langle \hat{\beta}^{(l),k}, \beta^{(l)} \rangle^2}{\|\hat{\beta}^{(l),k}\|_2^2 \cdot \|\beta^{(1)}\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\mathbb{E}[f_{k,l}(\bar{B}^k)\bar{B}_l])^2}{\mathbb{E}[f_{k,l}(\bar{B}^k)^2] \cdot \mathbb{E}[\bar{B}_l^2]}}_{\text{theoretical}}, \quad \text{for } l \in [L]$$

- **Gaussian prior.** The prior distribution of the two signals follows

$$(\beta_j^{(1)}, \beta_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, \rho \\ \rho, 1 \end{bmatrix} \right), \quad j \in [p]$$

- **Plots.** For each setting in the plots, we plot
  - Empirical normalized squared correlation (labeled as 'AMP');
  - Theoretical normalized squared correlation (labeled as 'SE').

Numerical Simulations

- **Two-component case for simulation:**

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i,$$

where $c_i \sim_{\text{i.i.d.}}$ Bernoulli(α), with $\alpha \in (0,1)$, and $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

- **Normalized squared correlation.** #Categories $= L$

$$\underbrace{\frac{\langle \hat{\beta}^{(l),k}, \beta^{(l)} \rangle^2}{\|\hat{\beta}^{(l),k}\|_2^2 \cdot \|\beta^{(1)}\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\mathbb{E}[f_{k,l}(\bar{B}^k)\bar{B}_l])^2}{\mathbb{E}[f_{k,l}(\bar{B}^k)^2] \cdot \mathbb{E}[\bar{B}_l^2]}}_{\text{theoretical}}, \quad \text{for } l \in [L]$$

- **Gaussian prior.** The prior distribution of the two signals follows

$$(\beta_j^{(1)}, \beta_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, \rho \\ \rho, 1 \end{bmatrix} \right), \quad j \in [p]$$

- **Plots.** For each setting in the plots, we plot
  - Empirical normalized squared correlation (labeled as 'AMP');
  - Theoretical normalized squared correlation (labeled as 'SE').

Numerical Simulations

- **Two-component case for simulation:**

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i,$$

where $c_i \sim_{\text{i.i.d.}}$ Bernoulli(α), with $\alpha \in (0,1)$, and $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

- **Normalized squared correlation.** $\#$Categories $= L$

$$\underbrace{\frac{\langle \hat{\beta}^{(l),k}, \beta^{(l)} \rangle^2}{\|\hat{\beta}^{(l),k}\|_2^2 \cdot \|\beta^{(1)}\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\mathbb{E}[f_{k,l}(\bar{B}^k)\bar{B}_l])^2}{\mathbb{E}[f_{k,l}(\bar{B}^k)^2] \cdot \mathbb{E}[\bar{B}_l^2]}}_{\text{theoretical}}, \quad \text{for } l \in [L]$$

- **Gaussian prior.** The prior distribution of the two signals follows

$$(\beta_j^{(1)}, \beta_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, \rho \\ \rho, 1 \end{bmatrix} \right), \quad j \in [p]$$

- **Plots.** For each setting in the plots, we plot
  - Empirical normalized squared correlation (labeled as 'AMP');
  - Theoretical normalized squared correlation (labeled as 'SE').

Numerical Simulations

- **Two-component case for simulation:**

$$Y_i = \langle X_i, \beta^{(1)} \rangle c_i + \langle X_i, \beta^{(2)} \rangle (1 - c_i) + \epsilon_i,$$

where $c_i \sim_{\text{i.i.d.}}$ Bernoulli(α), with $\alpha \in (0, 1)$, and $\epsilon_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2)$.

- **Normalized squared correlation.** $\#$Categories $= L$

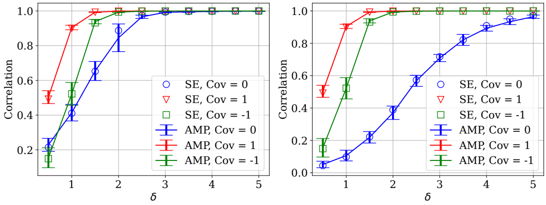
$$\underbrace{\frac{\langle \hat{\beta}^{(l),k}, \beta^{(l)} \rangle^2}{\|\hat{\beta}^{(l),k}\|_2^2 \cdot \|\beta^{(1)}\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\mathbb{E}[f_{k,l}(\bar{B}^k)\bar{B}_l])^2}{\mathbb{E}[f_{k,l}(\bar{B}^k)^2] \cdot \mathbb{E}[\bar{B}_l^2]}}_{\text{theoretical}}, \quad \text{for } l \in [L]$$

- **Gaussian prior.** The prior distribution of the two signals follows

$$(\beta_j^{(1)}, \beta_j^{(2)}) \sim_{\text{i.i.d.}} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1, \rho \\ \rho, 1 \end{bmatrix} \right), \quad j \in [p]$$

- **Plots.** For each setting in the plots, we plot
  - Empirical normalized squared correlation (labeled as 'AMP');
  - Theoretical normalized squared correlation (labeled as 'SE').

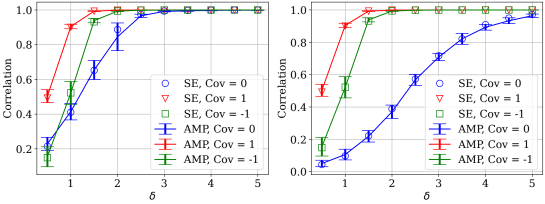
Numerical Simulations



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

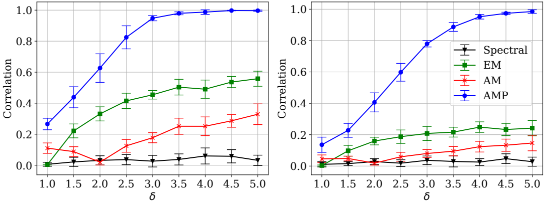
Figure: Normalized squared correlation vs. δ, with different values of signal covariance ρ, $\alpha = 0.7$, $\sigma = 0$.

Numerical Simulations



(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure: Normalized squared correlation vs. δ, with different values of signal covariance ρ, $\alpha = 0.7$, $\sigma = 0$.

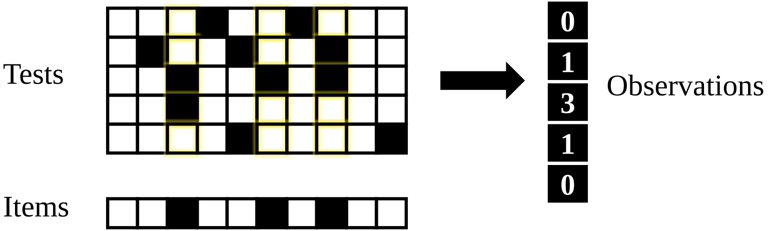


(a) $\beta^{(1)}$ (b) $\beta^{(2)}$

Figure: Comparison of different estimators; Normalized squared correlation vs. δ, with $\rho = 0$, $\alpha = 0.6$, and $\sigma = 0$.

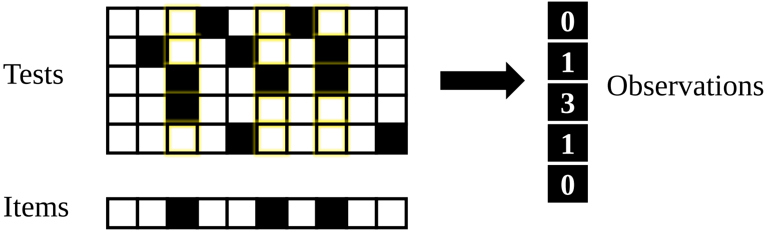
AMP for Pooled Data and QGT with i.i.d. Bernoulli Design

Application: Quantitative Group Testing (QGT)



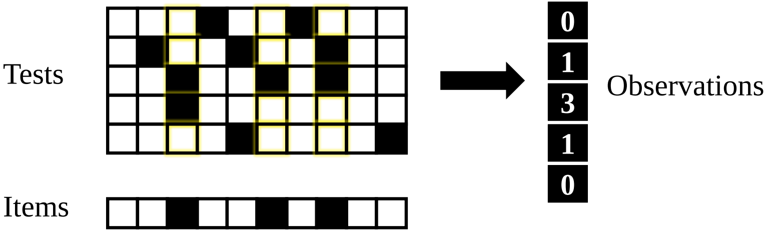
- **Setup:**
  - Given p items, where d are defective, recover the defective set.
  - Each test returns the number of defective items in the test.
  - Goal. Minimize number of pooled tests n required.
- **Medical use.** Aim to find the infected people and each test outputs the number of infected people (e.g., viral load) in the test.
- **Model.** $y = X\beta$ where $y \in \mathbb{R}^n$, $X \in \{0,1\}^{n \times p}$, and $\beta \in \{0,1\}^p$.

Application: Quantitative Group Testing (QGT)



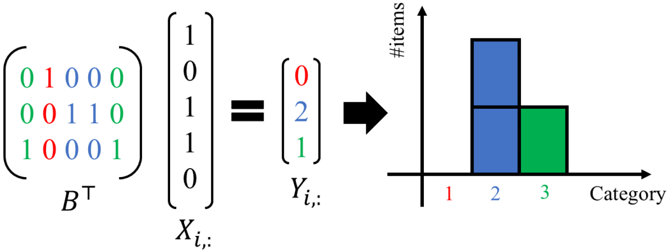
- **Setup:**
  - Given p items, where d are defective, recover the defective set.
  - Each test returns the number of defective items in the test.
  - Goal. Minimize number of pooled tests n required.
- **Medical use.** Aim to find the infected people and each test outputs the number of infected people (e.g., viral load) in the test.
- **Model.** $y = X\beta$ where $y \in \mathbb{R}^n$, $X \in \{0,1\}^{n \times p}$, and $\beta \in \{0,1\}^p$.

Application: Quantitative Group Testing (QGT)



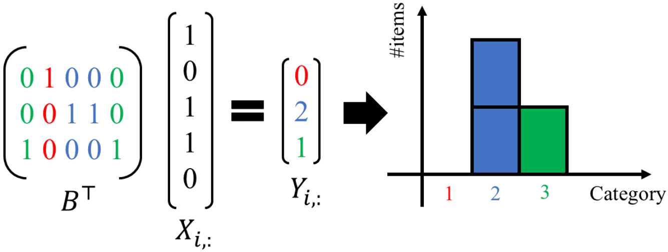
- **Setup:**
  - Given p items, where d are defective, recover the defective set.
  - Each test returns the number of defective items in the test.
  - Goal. Minimize number of pooled tests n required.
- **Medical use.** Aim to find the infected people and each test outputs the number of infected people (e.g., viral load) in the test.
- **Model.** $y = X\beta$ where $y \in \mathbb{R}^n$, $X \in \{0,1\}^{n \times p}$, and $\beta \in \{0,1\}^p$.

Application: Pooled Data



- **Setup.** $Y = XB$, generalization of QGT
  - Items have more than 2 categories; QGT $=$ pooled data with 2 categories.
  - Each test returns the total number of items corresponding to each category that are present in the pool.
  - Goal. Minimize the number of pooled test required.
- **Medical use:**
  - Each person has one out of several diseases.
  - We want to recover the disease each person is infected with.

Application: Pooled Data



- **Setup.** $Y = XB$, generalization of QGT
  - Items have more than 2 categories; QGT = pooled data with 2 categories.
  - Each test returns the total number of items corresponding to each category that are present in the pool.
  - Goal. Minimize the number of pooled test required.
- **Medical use:**
  - Each person has one out of several diseases.
  - We want to recover the disease each person is infected with.

AMP with Generalized White Noise Design

- **Issue.** Matrix-AMP algorithm for mixed regression requires Gaussian entries while QGT and pooled data require binary entries.
- **Algorithm.** Same as the one for matrix GLM but with a different design matrix $\widetilde{X}$.
- **Generalized white noise matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$:**
  - All entries $\widetilde{X}_{ij}$ are independent, have zero mean, and bounded moments.
  - Example: Sub-Gaussian entries.
- QGT $y = X\beta$ and pooled data $Y = XB$ with a Bernoulli design X can be recentered and rescaled to give $\tilde{y} = \widetilde{X}\beta$ and $\widetilde{Y} = \widetilde{X}B$ resp.

AMP with Generalized White Noise Design

- **Issue.** Matrix-AMP algorithm for mixed regression requires Gaussian entries while QGT and pooled data require binary entries.
- **Algorithm.** Same as the one for matrix GLM but with a different design matrix $\widetilde{X}$.
- **Generalized white noise matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$:**
  - All entries $\widetilde{X}_{ij}$ are independent, have zero mean, and bounded moments.
  - Example: Sub-Gaussian entries.
- QGT $y = X\beta$ and pooled data $Y = XB$ with a Bernoulli design X can be recentered and rescaled to give $\tilde{y} = \widetilde{X}\beta$ and $\widetilde{Y} = \widetilde{X}B$ resp.

AMP with Generalized White Noise Design

- **Issue.** Matrix-AMP algorithm for mixed regression requires Gaussian entries while QGT and pooled data require binary entries.
- **Algorithm.** Same as the one for matrix GLM but with a different design matrix $\widetilde{X}$.
- **Generalized white noise matrix $\widetilde{X} \in \mathbb{R}^{n \times p}$:**
  - All entries $\widetilde{X}_{ij}$ are independent, have zero mean, and bounded moments.
  - Example: Sub-Gaussian entries.
- QGT $y = X\beta$ and pooled data $Y = XB$ with a Bernoulli design X can be recentered and rescaled to give $\tilde{y} = \widetilde{X}\beta$ and $\widetilde{Y} = \widetilde{X}B$ resp.

State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$(B, B^{k+1}) \to (\bar{B}, \bar{B}^{k+1}), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $\bar{B}$ is independent of $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters M_B^k, T_B^k are defined previously.
- **Choice of denoiser.** Defined previously.
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].
- Our result makes the previous AMP performance guarantees in [Alaoui et al. 2018] rigorous.
- **Performance measure.** Normalized correlation
$$\frac{1}{p}\sum_{j=1}^{p}\langle \widehat{B}^k_{j,:}, B_{j,:}\rangle \to \mathbb{E}\Big[\big\langle f_k(\bar{B}^k), \bar{B}\big\rangle\Big].$$

State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$(B, B^{k+1}) \to (\bar{B}, \bar{B}^{k+1}), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $\bar{B}$ is independent of $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters M_B^k, T_B^k are defined previously.
- **Choice of denoiser.** Defined previously.
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].
- Our result makes the previous AMP performance guarantees in [Alaoui et al. 2018] rigorous.
- **Performance measure.** Normalized correlation
$$\frac{1}{p}\sum_{j=1}^{p}\langle \widehat{B}_{j,:}^k, B_{j,:}\rangle \to \mathbb{E}\Big[\big\langle f_k(\bar{B}^k), \bar{B}\big\rangle\Big].$$

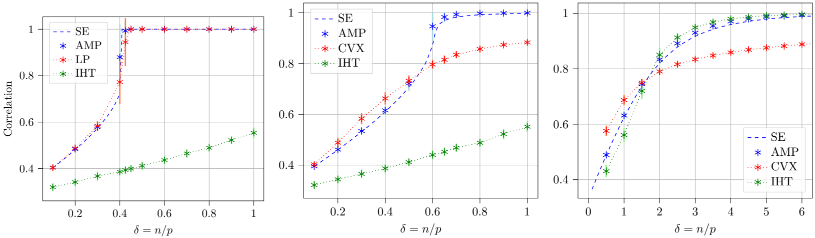
State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$(B, B^{k+1}) \to (\bar{B}, \bar{B}^{k+1}), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $\bar{B}$ is independent of $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters M_B^k, T_B^k are defined previously.
- **Choice of denoiser.** Defined previously.
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].
- Our result makes the previous AMP performance guarantees in [Alaoui et al. 2018] rigorous.
- **Performance measure.** Normalized correlation
$$\frac{1}{p}\sum_{j=1}^{p}\langle \widehat{B}_{j,:}^k, B_{j,:}\rangle \to \mathbb{E}\Big[\big\langle f_k(\bar{B}^k), \bar{B}\big\rangle\Big].$$

State Evolution Result

- **Theorem.** The empirical joint distribution of the rows of
$$(B, B^{k+1}) \to (\bar{B}, \bar{B}^{k+1}), \quad \bar{B}^{k+1} := \mathrm{M}_B^{k+1}\bar{B} + G_B^{k+1}$$
where $\bar{B}$ is independent of $G_B^{k+1} \sim \mathcal{N}(0, \mathrm{T}_B^{k+1})$ and the state evolution parameters M_B^k, T_B^k are defined previously.
- **Choice of denoiser.** Defined previously.
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].
- Our result makes the previous AMP performance guarantees in [Alaoui et al. 2018] rigorous.
- **Performance measure.** Normalized correlation
$$\frac{1}{p}\sum_{j=1}^{p}\langle \widehat{B}_{j,:}^k, B_{j,:}\rangle \to \mathbb{E}\Big[\big\langle f_k(\bar{B}^k), \bar{B}\big\rangle\Big].$$

Numerical Simulations

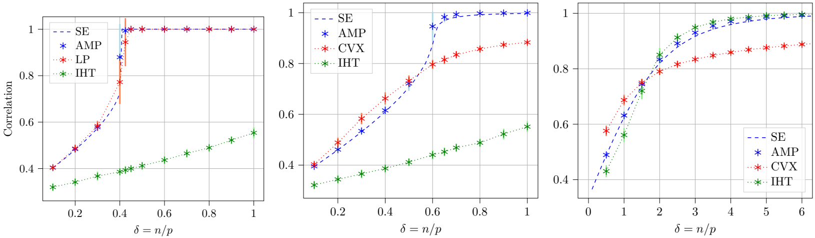


(a) $\sigma = 0$ (b) $\sigma = 0.1$ (c) $\sigma = 0.3$

Figure: AMP vs. other algorithms for pooled data: normalized correlation vs. δ, with $L = 3$ and $\pi = [1/3, 1/3, 1/3]$. The plots are similar for the case of non-uniform priors.

- **Algorithms.** linear programming (LP), convex optimization (CVX), iterative hard thresholding (IHT)
- AMP generally outperforms other algorithms.
- CVX is better than AMP for low δ when there is noise (why?)

Numerical Simulations



(a) $\sigma = 0$ (b) $\sigma = 0.1$ (c) $\sigma = 0.3$

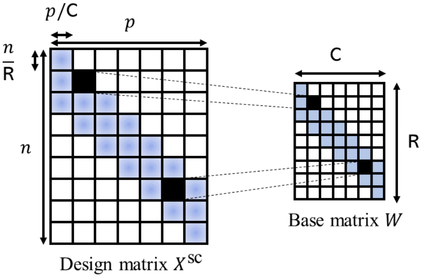
Figure: AMP vs. other algorithms for pooled data: normalized correlation vs. δ, with $L = 3$ and $\pi = [1/3, 1/3, 1/3]$. The plots are similar for the case of non-uniform priors.

- **Algorithms.** linear programming (LP), convex optimization (CVX), iterative hard thresholding (IHT)
- AMP generally outperforms other algorithms.
- CVX is better than AMP for low δ when there is noise (why?)

AMP for QGT and Pooled Data with Spatially Coupled Design

Improvement: Spatial Coupling Design

- Enforce a blockwise band-diagonal structure in X.
- All entries in X are either 0 or Bernoulli.

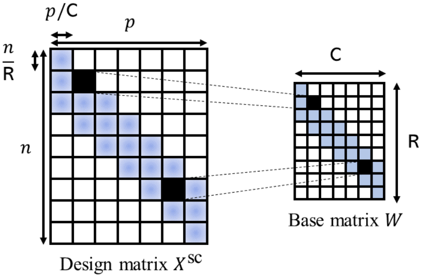


- R: Number of row blocks.
- C: Number of column blocks.
- ω: Coupling width.
- Relation: $\mathsf{R} = \mathsf{C} + \omega - 1$

- Visual representation of a $(\omega = 3, \mathsf{C} = 7)$ spatially coupled matrix.
- Additional tests associated to first and last entries of β. Edge entries recovered first, neighboring entries progressively reconstructed.

Improvement: Spatial Coupling Design

- Enforce a blockwise band-diagonal structure in X.
- All entries in X are either 0 or Bernoulli.



- R: Number of row blocks.
- C: Number of column blocks.
- ω: Coupling width.
- Relation: R = C + ω − 1

- Visual representation of a $(\omega = 3, \mathsf{C} = 7)$ spatially coupled matrix.
- Additional tests associated to first and last entries of β. Edge entries recovered first, neighboring entries progressively reconstructed.

Spatially Coupled (SC) AMP

- **AMP algorithm.** Iteratively produces estimates $\hat{\beta}^k$ of β.

  $\widetilde{\Theta}^k = \widetilde{y} - \widetilde{X}^{\rm sc}\hat{\beta}^k + b^k \odot Q^k \odot \widetilde{\Theta}^{k-1},\ \beta^{k+1} = (\widetilde{X}^{\rm sc})^\top (Q^k \odot \widetilde{\Theta}^k) - c^k \odot \hat{\beta}^k$
- **Signal estimate.** $\hat{\beta}^{k+1} = f_{k+1}(\beta^{k+1}, \mathsf{c})$.
  - b^k – determined by derivative of $f_k(\beta^k, \mathsf{c})$.
  - $\widetilde{y}$, $\widetilde{X}^{\rm sc}$ – recentered and rescaled versions of y and X.
  - Q^k, c^k – defined via state evolution parameters.
  - f_k – Lipschitz denoiser, applied component wise.
- State evolution parameters: $\chi_1^k, \ldots, \chi_{\mathsf{C}}^k$
- **Theorem.** For each block $\mathsf{c} \in [\mathsf{C}]$, the empirical joint distribution of

  $$(\beta_{\mathsf{c}}, \beta_{\mathsf{c}}^k) \to (\bar{\beta}, \bar{\beta}_{\mathsf{c}}^k), \quad \bar{\beta}_{\mathsf{c}}^k := (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G,$$

  where $\bar{\beta} \sim$ Bernoulli(π) is independent of $G \sim \mathcal{N}(0,1)$.
- **Choice of denoiser.** $f_k(s, \mathsf{c}) = \mathbb{E}[\bar{\beta} | (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G = s]$
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].

Spatially Coupled (SC) AMP

- **AMP algorithm.** Iteratively produces estimates $\hat{\beta}^k$ of β.

$$\widetilde{\Theta}^k = \widetilde{y} - \widetilde{X}^{\mathrm{sc}}\hat{\beta}^k + b^k \odot Q^k \odot \widetilde{\Theta}^{k-1},\ \beta^{k+1} = (\widetilde{X}^{\mathrm{sc}})^\top (Q^k \odot \widetilde{\Theta}^k) - c^k \odot \hat{\beta}^k$$

- **Signal estimate.** $\hat{\beta}^{k+1} = f_{k+1}(\beta^{k+1}, \mathsf{c})$.
  - b^k – determined by derivative of $f_k(\beta^k, \mathsf{c})$.
  - $\widetilde{y}$, $\widetilde{X}^{\mathrm{sc}}$ – recentered and rescaled versions of y and X.
  - Q^k, c^k – defined via state evolution parameters.
  - f_k – Lipschitz denoiser, applied component wise.
- State evolution parameters: $\chi_1^k, \ldots, \chi_{\mathsf{C}}^k$
- **Theorem.** For each block $\mathsf{c} \in [\mathsf{C}]$, the empirical joint distribution of

$$(\beta_{\mathsf{c}}, \beta_{\mathsf{c}}^k) \to (\bar{\beta}, \bar{\beta}_{\mathsf{c}}^k), \quad \bar{\beta}_{\mathsf{c}}^k := (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G,$$

  where $\bar{\beta} \sim$ Bernoulli(π) is independent of $G \sim \mathcal{N}(0,1)$.
- **Choice of denoiser.** $f_k(s, \mathsf{c}) = \mathbb{E}[\bar{\beta} | (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G = s]$
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].

Spatially Coupled (SC) AMP

- **AMP algorithm.** Iteratively produces estimates $\hat{\beta}^k$ of β.

$$\widetilde{\Theta}^k = \widetilde{y} - \widetilde{X}^{\rm sc}\hat{\beta}^k + b^k \odot Q^k \odot \widetilde{\Theta}^{k-1},\ \beta^{k+1} = (\widetilde{X}^{\rm sc})^\top (Q^k \odot \widetilde{\Theta}^k) - c^k \odot \hat{\beta}^k$$

- **Signal estimate.** $\hat{\beta}^{k+1} = f_{k+1}(\beta^{k+1}, \mathsf{c})$.
  - b^k – determined by derivative of $f_k(\beta^k, \mathsf{c})$.
  - $\widetilde{y}$, $\widetilde{X}^{\rm sc}$ – recentered and rescaled versions of y and X.
  - Q^k, c^k – defined via state evolution parameters.
  - f_k – Lipschitz denoiser, applied component wise.
- State evolution parameters: $\chi_1^k, ..., \chi_{\mathsf{C}}^k$
- **Theorem.** For each block $\mathsf{c} \in [\mathsf{C}]$, the empirical joint distribution of

$$(\beta_{\mathsf{c}}, \beta_{\mathsf{c}}^k) \to (\bar{\beta}, \bar{\beta}_{\mathsf{c}}^k), \quad \bar{\beta}_{\mathsf{c}}^k := (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G,$$

where $\bar{\beta} \sim \text{Bernoulli}(\pi)$ is independent of $G \sim \mathcal{N}(0,1)$.

- **Choice of denoiser.** $f_k(s, \mathsf{c}) = \mathbb{E}[\bar{\beta} | (\chi_{\mathsf{c}}^k)^2 \bar{\beta} + \chi_{\mathsf{c}}^k G = s]$
- **Proof idea.** Via reduction to an abstract AMP iteration, and applying the universality result of [Wang et al. 2022].

Almost-Exact Recovery

- **Theorem.** For any sampling ratio $n/p \to \delta > 0$ and $\omega = o(\mathsf{C})$, the SC-AMP estimate $\hat{\beta}$ achieves almost-exact recovery

$$\lim_{k,\omega,p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \mathbb{1}\{\hat{\beta}_j^k \neq \beta_j\} = 0,$$

with $n = o(p)$ tests.

- **Proof idea.** Apply previous theorem and
  - Characterize fixed points $\lim_{k\to\infty}(\chi_1^k, \dots, \chi_{\mathsf{C}}^k)$ via minimum of a potential function [Yedla et al. 2014].
  - Use above point to show that the asymptotic MSE vanishes for noiseless QGT.
- **Pooled data.** Run SC-AMP column-wise on $\widetilde{Y}$.
- **Performance measure.** Normalized squared correlation:

$$\underbrace{\frac{\langle \hat{\beta}^k, \beta\rangle}{\|\hat{\beta}^k\|_2^2 \cdot \|\beta\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{\left(\frac{1}{\mathsf{C}}\sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c}) \cdot \bar{\beta}]\right)^2}{\left(\frac{1}{\mathsf{C}}\sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c})^2]\right) \cdot (\mathbb{E}[\bar{\beta}^2])}}_{\text{theoretical}}$$

Almost-Exact Recovery

- **Theorem.** For any sampling ratio $n/p \to \delta > 0$ and $\omega = o(\mathsf{C})$, the SC-AMP estimate $\hat{\beta}$ achieves almost-exact recovery

$$\lim_{k,\omega,p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathbb{1}\{\hat{\beta}_j^k \neq \beta_j\} = 0,$$

  with $n = o(p)$ tests.

- **Proof idea.** Apply previous theorem and
  - Characterize fixed points $\lim_{k\to\infty}(\chi_1^k, \dots, \chi_{\mathsf{C}}^k)$ via minimum of a potential function [Yedla et al. 2014].
  - Use above point to show that the asymptotic MSE vanishes for noiseless QGT.

- **Pooled data.** Run SC-AMP column-wise on $\widetilde{Y}$.

- **Performance measure.** Normalized squared correlation:

$$\underbrace{\frac{\langle \hat{\beta}^k, \beta \rangle}{\|\hat{\beta}^k\|_2^2 \cdot \|\beta\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{\left(\frac{1}{\mathsf{C}} \sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c}) \cdot \bar{\beta}]\right)^2}{\left(\frac{1}{\mathsf{C}} \sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c})^2]\right) \cdot (\mathbb{E}[\bar{\beta}^2])}}_{\text{theoretical}}$$

Almost-Exact Recovery

- **Theorem.** For any sampling ratio $n/p \to \delta > 0$ and $\omega = o(\mathsf{C})$, the SC-AMP estimate $\hat{\beta}$ achieves almost-exact recovery
$$\lim_{k,\omega,p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \mathbb{1}\{\hat{\beta}_j^k \neq \beta_j\} = 0,$$
with $n = o(p)$ tests.
- **Proof idea.** Apply previous theorem and
  - Characterize fixed points $\lim_{k\to\infty}(\chi_1^k, \dots, \chi_{\mathsf{C}}^k)$ via minimum of a potential function [Yedla et al. 2014].
  - Use above point to show that the asymptotic MSE vanishes for noiseless QGT.
- **Pooled data.** Run SC-AMP column-wise on $\widetilde{Y}$.
- **Performance measure.** Normalized squared correlation:
$$\underbrace{\frac{\langle \hat{\beta}^k, \beta\rangle}{\|\hat{\beta}^k\|_2^2 \cdot \|\beta\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\frac{1}{\mathsf{C}}\sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c}) \cdot \bar{\beta}])^2}{(\frac{1}{\mathsf{C}}\sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c})^2]) \cdot (\mathbb{E}[\bar{\beta}^2])}}_{\text{theoretical}}$$

Almost-Exact Recovery

- **Theorem.** For any sampling ratio $n/p \to \delta > 0$ and $\omega = o(\mathsf{C})$, the SC-AMP estimate $\hat{\beta}$ achieves almost-exact recovery
$$\lim_{k,\omega,p\to\infty} \frac{1}{p} \sum_{j=1}^{p} \mathbb{1}\{\hat{\beta}_j^k \neq \beta_j\} = 0,$$
with $n = o(p)$ tests.
- **Proof idea.** Apply previous theorem and
  - Characterize fixed points $\lim_{k\to\infty}(\chi_1^k, \dots, \chi_{\mathsf{C}}^k)$ via minimum of a potential function [Yedla et al. 2014].
  - Use above point to show that the asymptotic MSE vanishes for noiseless QGT.
- **Pooled data.** Run SC-AMP column-wise on $\widetilde{Y}$.
- **Performance measure.** Normalized squared correlation:
$$\underbrace{\frac{\langle \hat{\beta}^k, \beta \rangle}{\|\hat{\beta}^k\|_2^2 \cdot \|\beta\|_2^2}}_{\text{empirical}} \to \underbrace{\frac{(\frac{1}{\mathsf{C}} \sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c}) \cdot \bar{\beta}])^2}{(\frac{1}{\mathsf{C}} \sum_{\mathsf{c}=1}^{\mathsf{C}} \mathbb{E}[f_k(\bar{\beta}_{\mathsf{c}}^k, \mathsf{c})^2]) \cdot (\mathbb{E}[\bar{\beta}^2])}}_{\text{theoretical}}$$

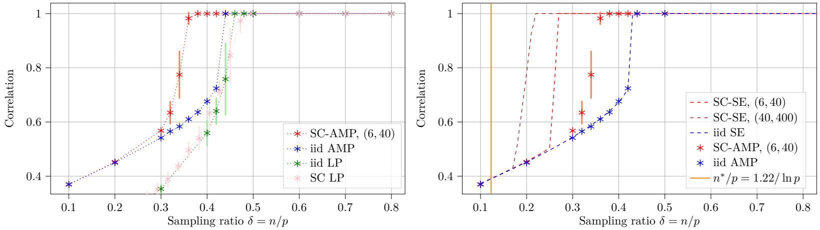
Numerical Simulations

- **Linear programming (LP).** Implemented LP using the i.i.d. matrix (iid LP) and spatially coupled matrix with $(\omega = 6, \mathsf{C} = 40)$ (SC LP).
- **Information theoretic lower bound.** $n^* = 2H(\pi)p/\ln p$.

- Spatially coupled AMP (SC-AMP) outperforms iid AMP and LP.
- As we increase coupling dimensions (ω, C), spatially coupled state evolution (SC-SE) approaches lower bound.

Numerical Simulations

- **Linear programming (LP).** Implemented LP using the i.i.d. matrix (iid LP) and spatially coupled matrix with $(\omega = 6, \mathsf{C} = 40)$ (SC LP).
- **Information theoretic lower bound.** $n^* = 2H(\pi)p/\ln p$.



(a) AMP vs. LP (b) SC-AMP vs. i.i.d. AMP

Figure: SC-AMP, iid AMP, and n^*/p used $p = 20000$, SC LP and iid LP used $p = 2000$. Defective probability π=0.3.

- Spatially coupled AMP (SC-AMP) outperforms iid AMP and LP.
- As we increase coupling dimensions (ω, C), spatially coupled state evolution (SC-SE) approaches lower bound.

Summary

- (Ch. 2) Extend AMP to account for matrix signals. To this end, we considered the matrix GLM model. Applied AMP to mixed regression.
- (Ch. 3) Extend AMP to account for generalized white noise design matrices under the matrix GLM model. Applied AMP to QGT and pooled data.
- (Ch. 4) Improve the performance of AMP for QGT and pooled data by considering a spatially coupled Bernoulli test matrix. We also extend the AMP algorithm to account for this matrix.

Summary

- (Ch. 2) Extend AMP to account for matrix signals. To this end, we considered the matrix GLM model. Applied AMP to mixed regression.
- (Ch. 3) Extend AMP to account for generalized white noise design matrices under the matrix GLM model. Applied AMP to QGT and pooled data.
- (Ch. 4) Improve the performance of AMP for QGT and pooled data by considering a spatially coupled Bernoulli test matrix. We also extend the AMP algorithm to account for this matrix.

Summary

- (Ch. 2) Extend AMP to account for matrix signals. To this end, we considered the matrix GLM model. Applied AMP to mixed regression.
- (Ch. 3) Extend AMP to account for generalized white noise design matrices under the matrix GLM model. Applied AMP to QGT and pooled data.
- (Ch. 4) Improve the performance of AMP for QGT and pooled data by considering a spatially coupled Bernoulli test matrix. We also extend the AMP algorithm to account for this matrix.