Graph-based approximate message passing iterations

Cedric Gerbelot

Courant Institute of Mathematical Sciences, NYU

Joint work with Raphaël Berthier (EPFL), arXiv 2109.11905
To appear in *Information and Inference : A Journal of the IMA*

Cargese 2023
Statistical physics and machine learning back together again

The starting point : probabilistic inference

A hidden process generates $\mathbf{w} \in \mathbb{R}^d$ with large d

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w})$$

The starting point : probabilistic inference

A hidden process generates $\mathbf{w} \in \mathbb{R}^d$ with large d

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w})$$

We observe $\mathbf{y} \in \mathbb{R}^n$ s.t.

$$\mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}|\mathbf{w})$$

Estimate $\mathbf{w}$?

The starting point : probabilistic inference

A hidden process generates $\mathbf{w} \in \mathbb{R}^d$ with large d

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w})$$

We observe $\mathbf{y} \in \mathbb{R}^n$ s.t.

$$\mathbf{y} \sim p_{\mathbf{y}}\left(\mathbf{y}|\mathbf{w}\right)$$

Estimate $\mathbf{w}$?

MMSE estimator : $\hat{\mathbf{w}} = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right]$, i.e.

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{y}}\left(\mathbf{y}|\mathbf{w}\right) d\mu(\mathbf{w})$$

The starting point : probabilistic inference

A hidden process generates $\mathbf{w} \in \mathbb{R}^d$ with large d

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w})$$

We observe $\mathbf{y} \in \mathbb{R}^n$ s.t.

$$\mathbf{y} \sim p_{\mathbf{y}}\left(\mathbf{y}|\mathbf{w}\right)$$

Estimate $\mathbf{w}$?

MMSE estimator : $\hat{\mathbf{w}} = \mathbb{E}\left[\mathbf{w}|\mathbf{y}\right]$, i.e.

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{y}}\left(\mathbf{y}|\mathbf{w}\right) d\mu(\mathbf{w})$$

Problem : This is a high-dimensional integral

Link with statistical physics

Typically $p_{\mathbf{w}} \propto \exp(-\beta f(\mathbf{w}))$ and $p_{\mathbf{y}} \propto \exp(-\beta g(\mathbf{w}, \mathbf{y}))$

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w})$$

Link with statistical physics

Typically $p_{\mathbf{w}} \propto \exp(-\beta f(\mathbf{w}))$ and $p_{\mathbf{y}} \propto \exp(-\beta g(\mathbf{w}, \mathbf{y}))$

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w})$$

- Equilibrium Gibbs measure
- Hamiltonian $\mathcal{H}(\mathbf{w}, \mathbf{y}) = g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})$
- β is the inverse temperature

Link with statistical physics : disordered systems

- distributions involve a dense, random interaction matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, here i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements
- $\mathbf{y}$ can come from another random model, i.e.

$$\mathbf{y} \sim \mathbf{p}_{0,\mathbf{y}}(\mathbf{y}|\mathbf{X}, \mathbf{w}_0, \epsilon), \mathbf{w}_0 \sim p_{\mathbf{w}_0}(\mathbf{w}_0)$$

- estimate the generative model with postulated densities g, f

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta\,(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w}))) d\mu(\mathbf{w})$$

Link with statistical physics : disordered systems

- distributions involve a dense, random interaction matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, here i.i.d. $\mathcal{N}(0, \frac{1}{d})$ elements
- $\mathbf{y}$ can come from another random model, i.e.

$$\mathbf{y} \sim \mathbf{p}_{0,\mathbf{y}}(\mathbf{y}|\mathbf{X}, \mathbf{w}_0, \epsilon), \mathbf{w}_0 \sim p_{\mathbf{w}_0}(\mathbf{w}_0)$$

- estimate the generative model with postulated densities g, f

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w})$$

Typical setup

$$P(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \frac{1}{\mathcal{Z}} \prod_{\mu=1}^{n} p_z(y_\mu|z_\mu) \prod_{i=1}^{d} p_w(w_i) \quad \text{where} \quad z_\mu = \sum_{i=1}^{d} X_{\mu i} w_i$$

Back to machine learning and inference

Recall

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{x})$$

For $\beta \to +\infty$, Laplace's method gives

$$\hat{\mathbf{w}} \quad \xrightarrow[\beta \to +\infty]{} \quad \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^d} \quad g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})$$

Empirical risk minimization with n samples in $\mathbb{R}^d$

Examples : LASSO, logistic regression, etc ...

Back to machine learning and inference

Recall
$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta\,(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})))d\mu(\mathbf{x})$$

For $\beta \to +\infty$, Laplace's method gives

$$\hat{\mathbf{w}} \quad \xrightarrow[\beta \to +\infty]{} \quad \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^d} \quad g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})$$

Empirical risk minimization with n samples in $\mathbb{R}^d$

Examples : LASSO, logistic regression, etc ...

Goal : single letter formulas for the properties of $\hat{\mathbf{w}}$ when $n, d \to \infty$ with aspect ratio $\alpha \in (0, \infty)$

Belief propagation (BP) and AMP iterations

Consider the LASSO problem : $\mathbf{X} \in \mathbb{R}^{n\times d}$ i.i.d. $\mathcal{N}(0, 1/d)$.

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 ||\mathbf{w}||_1 \right\}$$

where $\quad \mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon}_0$

Belief propagation (BP) and AMP iterations

Consider the LASSO problem : $\mathbf{X} \in \mathbb{R}^{n \times d}$ i.i.d. $\mathcal{N}(0, 1/d)$.

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 ||\mathbf{w}||_1 \right\}$$

where $\quad \mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon}_0$

- **Can we solve this optimization problem ?**
  (existing methods : subgradient, proximal point)
- **Theoretical guarantees of $\mathbf{w}^*$?**

Belief propagation (BP) and AMP iterations

Consider the LASSO problem : $\mathbf{X} \in \mathbb{R}^{n\times d}$ i.i.d. $\mathcal{N}(0, 1/d)$.

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^d} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 ||\mathbf{w}||_1 \right\}$$

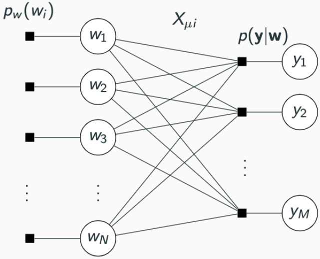
where $\quad \mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon}_0$

- **Can we solve this optimization problem ?**
  (existing methods : subgradient, proximal point)
- **Theoretical guarantees of $\mathbf{w}^*$?**

Can do both at the same time

Belief propagation (BP) and AMP iterations

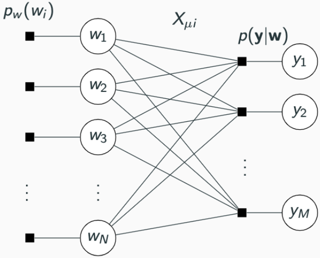
Factor graph representation of LASSO Gibbs measure



Factor graph for $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$

Belief propagation (BP) and AMP iterations

Factor graph representation of LASSO Gibbs measure



Factor graph for $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$

Relaxation of BP equations + concentration for $n, d \to +\infty$

$\Downarrow$

TAP equations [Mézard, Parisi & Virasoro '87]

AMP iteration [Donoho et al. '09]

AMP for the LASSO

LASSO problem. $\mathbf{X} \in \mathbb{R}^{n\times d}$ i.i.d. $\mathcal{N}(0, 1/d)$.

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 ||\mathbf{w}||_1 \right\}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon}_0$$

AMP for the LASSO

LASSO problem. $\mathbf{X} \in \mathbb{R}^{n \times d}$ i.i.d. $\mathcal{N}(0, 1/d)$.

$$\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w} \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda_1 ||\mathbf{w}||_1 \right\}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\epsilon}_0$$

Approximate message-passing for the LASSO [Donoho et al. '09, Bayati & Montanari '11]

$$\mathbf{z}^t = \mathbf{y} - \mathbf{X}\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{z}^{t-1} \langle \eta'(\hat{\mathbf{w}}^{t-1} + \frac{1}{\alpha}\mathbf{X}^T \mathbf{z}^{t-1}, \theta^{t-1}) \rangle$$

$$\hat{\mathbf{w}}^{t+1} = \eta(\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{X}^T \mathbf{z}^t, \theta^t)$$

- η is the soft-thresholding operator (proximal of ℓ_1)
- θ^t is a tunable parameter

AMP and state evolution equations

AMP for LASSO

$$
\begin{aligned}
\mathbf{z}^t &= y - \mathbf{X}\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{z}^{t-1}\langle \eta'(\hat{\mathbf{w}}^{t-1} + \frac{1}{\alpha}\mathbf{X}^T\mathbf{z}^{t-1}, \theta^{t-1})\rangle \\
\hat{\mathbf{w}}^{t+1} &= \eta(\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{X}^T\mathbf{z}^t, \theta^t)
\end{aligned}
$$

AMP and state evolution equations

AMP for LASSO

$$\mathbf{z}^t = y - \mathbf{X}\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{z}^{t-1}\langle \eta'(\hat{\mathbf{w}}^{t-1} + \frac{1}{\alpha}\mathbf{X}^T\mathbf{z}^{t-1}, \theta^{t-1})\rangle$$
$$\hat{\mathbf{w}}^{t+1} = \eta(\hat{\mathbf{w}}^t + \frac{1}{\alpha}\mathbf{X}^T\mathbf{z}^t, \theta^t)$$

State evolution for $n, d \to \infty$, $\mathrm{Z} \sim \mathcal{N}(0,1)$ [Donoho et al. '09,Bayati & Montanari '11]

$$V = \mathbb{E}_{z,w_0}\{[\eta'(W_0 + \sqrt{\frac{\Delta_0 + E}{\alpha}}Z; \theta(V))]^2\}$$
$$E = \mathbb{E}_{z,w_0}\{[\eta(W_0 + \sqrt{\frac{\Delta_0 + E}{\alpha}}Z; \theta(V)) - W_0]^2\}$$

Same result as the replica computation [Krzakala et al. '12]

AMP and state evolution equations

Generic AMP iteration

- $\mathbf{A} \in \mathbb{R}^{n \times d}$ Gaussian random matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, \frac{1}{d})$
- $e_t : \mathbb{R}^d \to \mathbb{R}^d$, $g_t : \mathbb{R}^n \to \mathbb{R}^n$, pseudo-Lipschitz functions

$$\begin{aligned} \mathbf{u}^{t+1} &= \mathbf{A}^\top g_t(\mathbf{v}^t) - d_t e_t(\mathbf{u}^t) \\ \mathbf{v}^t &= \mathbf{A} e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1}) \end{aligned}$$

where

$$\begin{aligned} d_t &= \frac{1}{m} \mathrm{div}(g_t(\mathbf{v}^t)) \\ b_t &= \frac{1}{m} \mathrm{div}(e_t(\mathbf{u}^t)) \end{aligned}$$

AMP and state evolution equations

$$\mathbf{u}^{t+1} = \mathbf{A}^\top g_t(\mathbf{v}^t) - d_t e_t(\mathbf{u}^t)$$
$$\mathbf{v}^t = \mathbf{A} e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1})$$

Define the recursion

$$\tau_{t+1} = \mathbb{E}\left[g_t^2(Z_\sigma^t)\right] \qquad Z_\sigma^t \sim \mathcal{N}(0, \sigma_t)$$
$$\sigma_t = \mathbb{E}\left[e_t^2(Z_\tau^t)\right] \qquad Z_\tau^t \sim \mathcal{N}(0, \tau_t)$$

AMP and state evolution equations

$$\mathbf{u}^{t+1} = \mathbf{A}^\top g_t(\mathbf{v}^t) - d_t e_t(\mathbf{u}^t)$$
$$\mathbf{v}^t = \mathbf{A} e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1})$$

Define the recursion

$$\tau_{t+1} = \mathbb{E}\left[g_t^2(Z_\sigma^t)\right] \qquad Z_\sigma^t \sim \mathcal{N}(0, \sigma_t)$$
$$\sigma_t = \mathbb{E}\left[e_t^2(Z_\tau^t)\right] \qquad Z_\tau^t \sim \mathcal{N}(0, \tau_t)$$

Theorem (Bayati & Montanari 2011)

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

$$\frac{1}{d}\sum_{i=1}^{d} \phi(u_i^t) \xrightarrow[n,d\to\infty]{a.s.} \mathbb{E}\left[\phi(Z_\tau^t)\right]$$
$$\frac{1}{n}\sum_{i=1}^{n} \phi(v_i^t) \xrightarrow[n,d\to\infty]{a.s.} \mathbb{E}\left[\phi(Z_\sigma^t)\right]$$

AMP and state evolution equations

Recall inference problem :

- recover $\mathbf{w}_0$ from observations $\mathbf{y} = \phi(\mathbf{A}\mathbf{w}_0)$
- prior $p_{\mathbf{w}_0}$

$$
\begin{aligned}
\mathbf{u}^{t+1} &= \mathbf{A}^\top g_t(\mathbf{v}^t) - d_t e_t(\mathbf{u}^t) \\
\mathbf{v}^t &= \mathbf{A} e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1})
\end{aligned}
$$

AMP and state evolution equations

Recall inference problem :

- recover $\mathbf{w}_0$ from observations $\mathbf{y} = \phi(\mathbf{A}\mathbf{w}_0)$
- prior $p_{\mathbf{w}_0}$

$$\begin{aligned} \mathbf{u}^{t+1} &= \mathbf{A}^\top g_t(\mathbf{v}^t) - d_t e_t(\mathbf{u}^t) \\ \mathbf{v}^t &= \mathbf{A} e_t(\mathbf{u}^t) - b_t g_{t-1}(\mathbf{v}^{t-1}) \end{aligned}$$

- function e_t estimates $\mathbf{w}_0$: denoising $p_{\mathbf{w}_0}$ blurred with additive Gaussian noise
- function g_t estimates $\mathbf{A}\mathbf{w}_0$: same for $p(\mathbf{y}|\mathbf{A}\mathbf{w}_0)$

Composing AMP iterations

Multilayer AMP [Mézard '17, Manoel et al. '17] :

Random neural networks, generative models, structured matrices ...
Used in other works [Gabrié et al. '19, Aubin et al. '20]

Recover $\mathbf{w}_0 \in \mathbb{R}^{N_1}$, prior $p_{\mathbf{w}_0}$, $\mathbf{A}_l \in \mathbb{R}^{N_l \times N_{l-1}}$

$$\mathbf{y} = \phi_L(\boldsymbol{A}_L \phi_{L-1}(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1 \mathbf{w}_0))))$$

Standard AMP iteration at each layer l

Composing AMP iterations

Multilayer AMP [Mézard '17, Manoel et al. '17] :

Random neural networks, generative models, structured matrices ...
Used in other works [Gabrié et al. '19, Aubin et al. '20]

Recover $\mathbf{w}_0 \in \mathbb{R}^{N_1}$, prior $p_{\mathbf{w}_0}$, $\mathbf{A}_l \in \mathbb{R}^{N_l \times N_{l-1}}$

$$\mathbf{y} = \phi_L(\boldsymbol{A}_L \phi_{L-1}\left(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1 \mathbf{w}_0))\right))$$

Standard AMP iteration at each layer l

$$\begin{aligned} \mathbf{u}_l^{t+1} &= \mathbf{A}_l^\top g_{l,t}(\mathbf{v}_l^t) - d_{l,t} e_{l,t}(\mathbf{u}_l^t) \\ \mathbf{v}_l^t &= \mathbf{A}_l e_{l,t}(\mathbf{u}_l^t) - b_{l,t} g_{l,t-1}(\mathbf{v}_l^{t-1}) \end{aligned}$$

Layerwise state evolution equations

Composing AMP iterations

Spiked matrix with generative prior [Aubin et al. '19]

Recover $\mathbf{v}_0$ from

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{W} \quad \mathbf{W} \sim \mathrm{GOE(N)}$$

with $\mathbf{v}_0$ generated from

$$\mathbf{v}_0 = \phi_L(\boldsymbol{A}_L\phi_{L-1}\left(\boldsymbol{A}_{L-1}(\ldots\phi_1(\boldsymbol{A}_1\mathbf{w}_0)\right)))$$

Composing AMP iterations

Spiked matrix with generative prior [Aubin et al. '19]

Recover $\mathbf{v}_0$ from

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{W} \quad \mathbf{W} \sim \mathrm{GOE(N)}$$

with $\mathbf{v}_0$ generated from

$$\mathbf{v}_0 = \phi_L(\boldsymbol{A}_L\phi_{L-1}\left(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1\mathbf{w}_0)\right)))$$

AMP for the spike part

$$\mathbf{x}^{t+1} = \mathbf{W}f^t(\mathbf{x}^t) - b^t f^{t-1}(\mathbf{x}^{t-1})$$

Combined with MLAMP, also admits layerwise state evolution equations.

Composition preserves state evolution property

State evolution property is preserved when composing AMP iterations

$\Downarrow$

Common structure in AMP iterations ? Can we use this structure to give a modular proof of SE equations ?

Composition preserves state evolution property

State evolution property is preserved when composing AMP iterations

$\Downarrow$

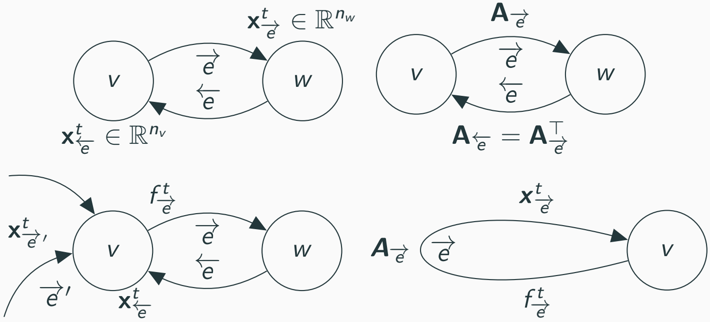
Common structure in AMP iterations ? Can we use this structure to give a modular proof of SE equations ?

Proposed Solution

- **indexation of AMP iterations on an oriented graph**
- **modular proof of SE equations based on this graph**

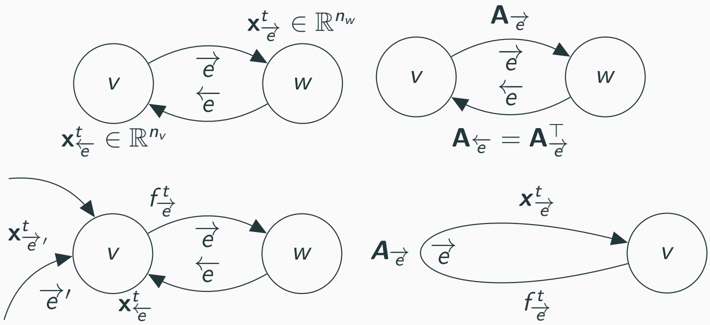
Graph-based AMP iterations : the oriented graph

Variables $\mathbf{x}^t_{\vec{e}}$, random matrices $\mathbf{A}_{\vec{e}}$, non-linearities $f^t_{\vec{e}}$



Graph-based AMP iterations : the oriented graph

Variables $\mathbf{x}^t_{\overrightarrow{e}}$, random matrices $\mathbf{A}_{\overrightarrow{e}}$, non-linearities $f^t_{\overrightarrow{e}}$



Arbitrary composition of this structure

Graph-based AMP : the iteration

The graph-based AMP iteration reads

$$\begin{aligned}
\mathbf{x}^{t+1}_{\vec{e}} &= \boldsymbol{A}_{\vec{e}}\mathbf{m}^{t}_{\vec{e}} - b^{t}_{\vec{e}}\mathbf{m}^{t-1}_{\overleftarrow{e}}\,,\\
\mathbf{m}^{t}_{\vec{e}} &= f^{t}_{\vec{e}}\left(\left(\mathbf{x}^{t}_{\vec{e}\,'}\right)_{\vec{e}\,':\,\vec{e}\,'\to\vec{e}}\right)\,,
\end{aligned}$$

where $b^{t}_{\vec{e}}$ is the *Onsager term*

$$b^{t}_{\vec{e}} = \frac{1}{N}\,\mathrm{Tr}\,\frac{\partial f^{t}_{\vec{e}}}{\partial \mathbf{x}_{\overleftarrow{e}}}\left(\left(\mathbf{x}^{t}_{\vec{e}\,'}\right)_{\vec{e}\,':\,\vec{e}\,'\to\vec{e}}\right) \qquad \in \mathbb{R}\,.$$

The graph-based AMP iteration reads

$$\mathbf{x}^{t+1}_{\overrightarrow{e}} = \boldsymbol{A}_{\overrightarrow{e}} \mathbf{m}^{t}_{\overrightarrow{e}} - b^{t}_{\overrightarrow{e}} \mathbf{m}^{t-1}_{\overleftarrow{e}} \,,$$
$$\mathbf{m}^{t}_{\overrightarrow{e}} = f^{t}_{\overrightarrow{e}}\left(\left(\mathbf{x}^{t}_{\overrightarrow{e}'}\right)_{\overrightarrow{e}': \overrightarrow{e}' \to \overrightarrow{e}} \right) \,,$$

where $b^{t}_{\overrightarrow{e}}$ is the *Onsager term*

$$b^{t}_{\overrightarrow{e}} = \frac{1}{N} \operatorname{Tr} \frac{\partial f^{t}_{\overrightarrow{e}}}{\partial \mathbf{x}_{\overleftarrow{e}}} \left(\left(\mathbf{x}^{t}_{\overrightarrow{e}'}\right)_{\overrightarrow{e}': \overrightarrow{e}' \to \overrightarrow{e}} \right) \qquad \in \mathbb{R} \,.$$

Theorem (informal):

- **any Graph-based AMP iterations admits rigorous SE**
- **equations can be deduced from the graph**

Graph-based AMP : the state evolution equations

Definition (State evolution iterates)

Define independently for each $\overrightarrow{e} \in \overrightarrow{E}$, $\boldsymbol{Z}^0_{\overrightarrow{e}} = \boldsymbol{x}^0_{\overrightarrow{e}}$ and $(\boldsymbol{Z}^1_{\overrightarrow{e}}, \ldots, \boldsymbol{Z}^t_{\overrightarrow{e}})$ a centered Gaussian random vector of covariance $(\boldsymbol{\kappa}^{r,s}_{\overrightarrow{e}})_{r,s \leq t} \otimes I_{n_w}$. We then define new state evolution iterates

$$\boldsymbol{\kappa}^{t+1,s+1}_{\overrightarrow{e}} = \lim_{n \to \infty} \frac{1}{N} \mathbb{E}\left[\left\langle f^s_{\overrightarrow{e}}\left((\boldsymbol{Z}^s_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}' \to \overrightarrow{e}}\right), f^t_{\overrightarrow{e}}\left((\boldsymbol{Z}^t_{\overrightarrow{e}'})_{\overrightarrow{e}':\overrightarrow{e}' \to \overrightarrow{e}}\right)\right\rangle\right]$$

for all $\quad s \in \{1, \ldots, t\}, \overrightarrow{e} \in \overrightarrow{E}$.

Theorem (Gerbelot & Berthier '21)

Under regularity assumptions, for any sequence of uniformly (in n) pseudo-Lipschitz function $\Phi : \mathbb{R}^{(t+1)N} \to \mathbb{R}$,

$$\Phi\left((\boldsymbol{x}^s_{\overrightarrow{e}})_{0 \leq s \leq t, \overrightarrow{e} \in \overrightarrow{E}}\right) \overset{P}{\simeq} \mathbb{E}\left[\Phi\left((\boldsymbol{Z}^s_{\overrightarrow{e}})_{0 \leq s \leq t, \overrightarrow{e} \in \overrightarrow{E}}\right)\right]$$

Graph-based AMP : recovering known examples

Symmetric AMP : spiked matrix recovery, SK model [Rangan et al. '12,Javanmard & Montanari '12,Deshpande & Montanari '14,Bolthausen '14]

Recover $\mathbf{v}_0 \in \mathbb{R}^N$ from:

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}} \mathbf{v}_0 \mathbf{v}_0^\top + \mathbf{W}$$

where the noise matrix $\mathbf{W} \in GOE(N)$.

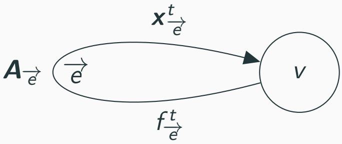
Graph-based AMP : recovering known examples

Symmetric AMP : spiked matrix recovery, SK model [Rangan et al. '12,Javanmard & Montanari '12,Deshpande & Montanari '14,Bolthausen '14]

Recover $\mathbf{v}_0 \in \mathbb{R}^N$ from:

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{W}$$

where the noise matrix $\mathbf{W} \in GOE(N)$.



$$\mathbf{x}^{t+1}_{\vec{e}} = \mathbf{A}_{\vec{e}}\mathbf{m}^t_{\vec{e}} - b^t_{\vec{e}}\mathbf{m}^{t-1}_{\vec{e}} \,,$$
$$\mathbf{m}^t_{\vec{e}} = f^t_{\vec{e}}\left(\mathbf{x}^t_{\vec{e}}\right) \,,$$

Graph-based AMP : recovering known examples

Asymmetric AMP : LASSO, GLM, ... [Donoho et al. '09, Bayati & Montanari '11, Rangan '11]

Recover $\mathbf{x}_0 \in \mathbb{R}^N$ from

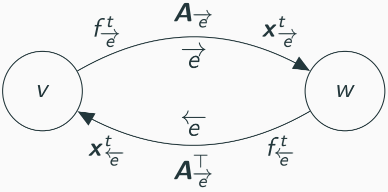
$$\mathbf{y} = \phi(\mathbf{A}\mathbf{x}_0, \boldsymbol{\epsilon}_\mathbf{0})$$

Graph-based AMP : recovering known examples

Asymmetric AMP : LASSO, GLM, ... [Donoho et al. '09, Bayati & Montanari '11, Rangan '11]

Recover $\mathbf{x}_0 \in \mathbb{R}^N$ from

$$\mathbf{y} = \phi(\mathbf{A}\mathbf{x}_0, \boldsymbol{\epsilon}_\mathbf{0})$$



$$\begin{aligned}
\mathbf{x}^{t+1}_{\overrightarrow{e}} &= \mathbf{A}_{\overrightarrow{e}}\mathbf{m}^t_{\overrightarrow{e}} - b^t_{\overrightarrow{e}}\mathbf{m}^{t-1}_{\overleftarrow{e}}\,, \\
\mathbf{m}^t_{\overrightarrow{e}} &= f^t_{\overrightarrow{e}}\left(\mathbf{x}^t_{\overleftarrow{e}}\right)\,, \\
\mathbf{x}^{t+1}_{\overleftarrow{e}} &= \mathbf{A}^\top_{\overrightarrow{e}}\mathbf{m}^t_{\overleftarrow{e}} - b^t_{\overleftarrow{e}}\mathbf{m}^{t-1}_{\overrightarrow{e}}\,, \\
\mathbf{m}^t_{\overleftarrow{e}} &= f^t_{\overleftarrow{e}}\left(\mathbf{x}^t_{\overrightarrow{e}}\right)\,.
\end{aligned}$$

Multilayer AMP [Manoel et al. '17] :

Random neural networks, generative models, structured matrices ...
Used in other works [Gabrie et al. '19, Aubin et al. '20]

Recover $\mathbf{x}_0 \in \mathbb{R}^{N_1}$

$$\mathbf{y} = \phi_L(\boldsymbol{A}_L \phi_{L-1} (\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1 \mathbf{x}_0)))$$

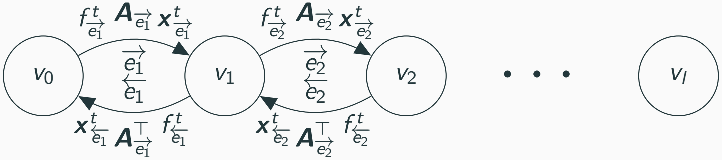
Graph-based AMP iterations : proving heuristic SE equations

Multilayer AMP [Manoel et al. '17] :

Random neural networks, generative models, structured matrices ...
Used in other works [Gabrie et al. '19, Aubin et al. '20]

Recover $\mathbf{x}_0 \in \mathbb{R}^{N_1}$

$$\mathbf{y} = \phi_L(\boldsymbol{A}_L \phi_{L-1}(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1 \mathbf{x}_0)))$$



Spiked matrix with generative prior [Aubin et al. '19]

Recover $\mathbf{v}_0$ from

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{A}_0$$

with $\mathbf{v}_0$ generated from

$$\mathbf{v}_0 = \phi_L(\boldsymbol{A}_L\phi_{L-1}\,(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1\mathbf{w}_0))))$$

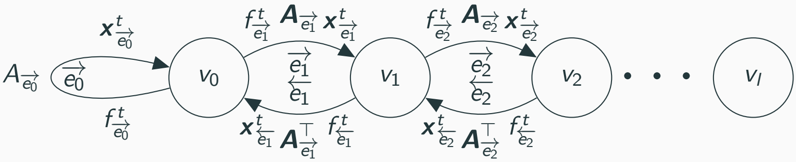
Spiked matrix with generative prior [Aubin et al. '19]

Recover $\mathbf{v}_0$ from

$$\mathbf{Y} = \sqrt{\frac{\lambda}{N}}\mathbf{v}_0\mathbf{v}_0^\top + \mathbf{A}_0$$

with $\mathbf{v}_0$ generated from

$$\mathbf{v}_0 = \phi_L(\boldsymbol{A}_L\phi_{L-1}(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1\mathbf{w}_0))))$$



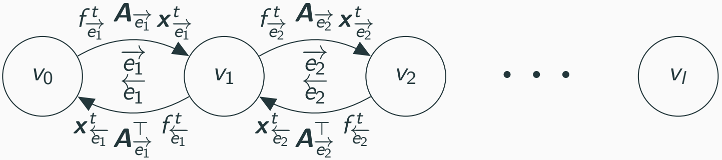
Graph-based AMP iterations : proving new SE equations

Convolutional multilayer generalized linear estimation [Gerbelot et al. 22]

Recover $\mathbf{x}_0 \in \mathbb{R}^{N_1}$ from

$$\mathbf{y} = \phi_L(\boldsymbol{A}_L \phi_{L-1}(\boldsymbol{A}_{L-1}(...\phi_1(\boldsymbol{A}_1 \mathbf{x}_0))))$$

Random sparse circulant $\boldsymbol{A}_l \in \mathbb{R}^{N_{l+1} \times N_l}$



Proof of Graph-AMP theorem

$\mathbf{A} \sim \mathrm{GOE}(N)$, then

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

with initialization at $\mathbf{x}^0$ and Onsager correction

$$b_t = \mathrm{div}\left[f_t(\mathbf{x}^t)\right]$$

Proof of Graph-AMP theorem

$\mathbf{A} \sim \mathrm{GOE}(N)$, then

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

with initialization at $\mathbf{x}^0$ and Onsager correction

$$b_t = \mathrm{div}\left[f_t(\mathbf{x}^t)\right]$$

State evolution : for any t, $\mathbf{x}^t$ behaves as $\mathbf{Z}^t \sim \mathcal{N}(0, \kappa_{t,t}\mathbf{I}_N)$

$$\text{where} \quad \kappa_{t+1} = \mathbb{E}\left[(f^t(z^t))^2\right], \qquad z^t \sim \mathcal{N}(0, \kappa_t)$$

Proof idea due to E. Bolthausen '09, '14

Sketch of proof : Bolthausen conditioning

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

Define the σ-algebra $\mathfrak{S}_t = \sigma(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^t)$. We then have :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}|_{\mathfrak{S}_t}\mathbf{m}^t - b_t \mathbf{m}^{t-1}$$

Sketch of proof : Bolthausen conditioning

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

Define the σ-algebra $\mathfrak{S}_t = \sigma(\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^t)$. We then have :

$$\mathbf{x}^{t+1}|_{\mathfrak{S}_t} = \mathbf{A}|_{\mathfrak{S}_t}\mathbf{m}^t - b_t\mathbf{m}^{t-1}$$

Gaussian conditioning lemma

$$\begin{aligned}\mathbf{A}|_{\mathfrak{S}_t} &= \mathbb{E}\left[\mathbf{A}|\mathfrak{S}_t\right] + \mathcal{P}_t(\mathbf{A}) \\ &= \mathbf{A} - \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}} + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\end{aligned}$$

where $\mathbf{M}_{t-1} = \left[m^0|...|m^{t-1}\right]$ and $\tilde{\mathbf{A}}$ is an independent copy of $\mathbf{A}$.

Sketch of proof : Bolthausen conditioning

A bit of algebra leads to

$$\mathbf{x}^{t+1}_{|\mathfrak{S}_t} = \underbrace{\mathbf{X}_{t-1}\alpha_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t}_{\textit{Part } 1} + \underbrace{[0|\mathbf{M}_{t-2}]\,\mathbf{B}_t\alpha_t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}}_{\textit{Part } 2}$$

Sketch of proof : Bolthausen conditioning

A bit of algebra leads to

$$\mathbf{x}^{t+1}_{|\mathfrak{S}_t} = \underbrace{\mathbf{X}_{t-1}\alpha_t + \mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\tilde{\mathbf{A}}\mathbf{P}^{\perp}_{\mathbf{M}_{t-1}}\mathbf{m}^t}_{Part\ 1} + \underbrace{[0|\mathbf{M}_{t-2}]\,\mathbf{B}_t\alpha_t + \mathbf{P}_{\mathbf{M}_{t-1}}\mathbf{A}\mathbf{m}^t_{\perp} - b_t\mathbf{m}^{t-1}}_{Part\ 2}$$

- Part 1 concentrates (induction+Gaussian concentration)
- Part 2 goes to zero w.h.p. as $N \to \infty$
- Onsager correction b_t cancels the bothersome part

Sketch of proof : embedding of the graph

Embed the graph into a large, matrix valued, non-separable iteration of the form

$$\boldsymbol{X}^{t+1} = \boldsymbol{A}\boldsymbol{M}^{t} - \boldsymbol{M}^{t-1}\mathbf{b}_t^\top \,.$$

Sketch of proof : embedding of the graph

Embed the graph into a large, matrix valued, non-separable iteration of the form

$$\boldsymbol{X}^{t+1} = \boldsymbol{A}\boldsymbol{M}^{t} - \boldsymbol{M}^{t-1}\mathbf{b}_t^{\top} .$$

where

$$\boldsymbol{A}\boldsymbol{M}^{t} = \begin{pmatrix} \boldsymbol{A}_{\overrightarrow{e}_1} & & & & & \\ & \ddots & & & & \\ & & \boldsymbol{A}_{\overrightarrow{e}_l} & & & \\ & & & * & \boldsymbol{A}_{\overrightarrow{e}_{l+1}} & \\ & & & \boldsymbol{A}_{\overleftarrow{e}_{l+1}} & * & \\ & & & & & \ddots \end{pmatrix} \begin{pmatrix} f^{t}_{\overrightarrow{e}_1}(.) & & & & & \\ & \ddots & & & & \\ & & f^{t}_{\overrightarrow{e}_l}(.) & & & \\ & & & 0 & f^{t}_{\overleftarrow{e}_{l+1}}(.) & \\ & & & f^{t}_{\overrightarrow{e}_{l+1}}(.) & 0 & \\ & & & & & \ddots \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{A}_{\overrightarrow{e}_1} f^{t}_{\overrightarrow{e}_1}\left(\left(\boldsymbol{x}^{t}_{\overrightarrow{e}}\right)_{\overrightarrow{e} : \overrightarrow{e} \to \overrightarrow{e_1}}\right) & & & & & \\ & \ddots & & & * & \\ & & \boldsymbol{A}_{\overrightarrow{e}_l} f^{t}_{\overrightarrow{e}_l}(.) & & & \\ & & & \boldsymbol{A}_{\overrightarrow{e}_{l+1}} f^{t}_{\overrightarrow{e}_{l+1}}(.) & & \\ & * & & & \boldsymbol{A}_{\overleftarrow{e}_{l+1}} f^{t}_{\overleftarrow{e}_{l+1}}(.) & \\ & & & & & \ddots \end{pmatrix} .$$

and,

$$
\begin{aligned}
\boldsymbol{M}^{t-1}\mathbf{b}_t &= \begin{pmatrix}
f^{t-1}_{\overrightarrow{e}_1}(.) & & & & & \\
& \ddots & & & & \\
& & f^{t-1}_{\overrightarrow{e}_l}(.) & & & \\
& & & 0 & f^{t-1}_{\overleftarrow{e}_{l+1}}(.) & \\
& & & f^{t-1}_{\overrightarrow{e}_{l+1}}(.) & 0 & \\
& & & & & \ddots
\end{pmatrix}
\begin{pmatrix}
\mathbf{b}^t_{\overrightarrow{e}_1} & & & & & \\
& \ddots & & & & \\
& & \mathbf{b}^t_{\overrightarrow{e}_l} & & & \\
& & & 0 & \mathbf{b}^t_{\overleftarrow{e}_{l+1}} & \\
& & & \mathbf{b}^t_{\overrightarrow{e}_{l+1}} & 0 & \\
& & & & & \ddots
\end{pmatrix} \\
&= \begin{pmatrix}
\mathbf{b}^t_{\overrightarrow{e}_1} f^{t-1}_{\overrightarrow{e}_1}\left(\left(\boldsymbol{x}^t_{\overrightarrow{e}}\right)_{\overrightarrow{e}:\overrightarrow{e}\to\overrightarrow{e_1}}\right) & & & & & \\
& \ddots & & & 0 & \\
& & \mathbf{b}^t_{\overrightarrow{e}_l} f^{t-1}_{\overrightarrow{e}_l}(.) & & & \\
& & & \mathbf{b}^t_{\overrightarrow{e}_{l+1}} f^{t-1}_{\overleftarrow{e}_{l+1}}(.) & & \\
& 0 & & & \mathbf{b}^t_{\overleftarrow{e}_{l+1}} f^{t-1}_{\overrightarrow{e}_{l+1}}(.) & \\
& & & & & \ddots
\end{pmatrix}.
\end{aligned}
$$

Low rank perturbations

Define the matrix

$$\hat{\mathbf{A}} = \mathbf{A} + \frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top \quad \in \mathbb{R}^{N\times N},$$

Low rank perturbations

Define the matrix

$$\hat{\mathbf{A}} = \mathbf{A} + \frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top \quad \in \mathbb{R}^{N\times N},$$

AMP iteration

$$\begin{aligned}
\boldsymbol{X}^{t+1} &= \hat{\mathbf{A}}\boldsymbol{M}^t - \boldsymbol{M}^{t-1}(\mathbf{b}^t)^\top && \in \mathbb{R}^{N\times q}\,, \\
\boldsymbol{M}^t &= f^t(\boldsymbol{X}^t) && \in \mathbb{R}^{N\times q}\,, \\
\mathbf{b}^t &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i^t}{\partial \boldsymbol{X}_i}(\boldsymbol{X}^t) && \in \mathbb{R}^{q\times q}\,.
\end{aligned}$$

Low rank perturbations

Define the matrix

$$\hat{\mathbf{A}} = \mathbf{A} + \frac{1}{N}\mathbf{V}_0\mathbf{V}_0^\top \quad \in \mathbb{R}^{N\times N},$$

AMP iteration

$$\begin{aligned}
\boldsymbol{X}^{t+1} &= \hat{\mathbf{A}}\boldsymbol{M}^t - \boldsymbol{M}^{t-1}(\mathbf{b}^t)^\top && \in \mathbb{R}^{N\times q}\,, \\
\boldsymbol{M}^t &= f^t(\boldsymbol{X}^t) && \in \mathbb{R}^{N\times q}\,, \\
\mathbf{b}^t &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i^t}{\partial \boldsymbol{X}_i}(\boldsymbol{X}^t) && \in \mathbb{R}^{q\times q}\,.
\end{aligned}$$

State evolution recursion, initialized with $\boldsymbol{\mu}_0 = 0_{q\times q}$,

$$\begin{aligned}
\boldsymbol{\kappa}^{1,1} &= \lim_{N\to\infty}\frac{1}{N}f^0(\mathbf{V}_0\boldsymbol{\mu}_0 + \boldsymbol{X}^0)^\top f^0(\mathbf{V}_0\boldsymbol{\mu}_0 + \boldsymbol{X}^0) \\
\boldsymbol{\mu}^{s+1} &= \lim_{N\to+\infty}\frac{1}{N}\mathbb{E}\left[(\mathbf{V}_0)^\top f^s\left(\mathbf{V}_0\boldsymbol{\mu}^s + \boldsymbol{Z}^s\right)\right] \\
\boldsymbol{\kappa}^{t+1,s+1} &= \boldsymbol{\kappa}^{s+1,t+1} = \lim_{N\to\infty}\frac{1}{N}\mathbb{E}\left[f^s(\mathbf{V}_0\boldsymbol{\mu}^s + \boldsymbol{Z}^s)^\top f^t(\mathbf{V}_0\boldsymbol{\mu}^t + \boldsymbol{Z}^t)\right]\,.
\end{aligned}$$

Sketch of proof

Similar proof with a single node in [Deshpande, Abbe & Montanari '17]. Define

$$\begin{aligned}
\boldsymbol{S}^{t+1} &= \mathbf{A}\tilde{\boldsymbol{M}}^{t} - \tilde{\boldsymbol{M}}^{t-1}(\tilde{\mathbf{b}}^{t})^{\top} && \in \mathbb{R}^{N\times q}\,,\\
\tilde{\boldsymbol{M}}^{t} &= f^{t}(\mathbf{V}_0\boldsymbol{\mu}^{t} + \boldsymbol{S}^{t}) && \in \mathbb{R}^{N\times q}\,,\\
\tilde{\mathbf{b}}^{t} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i^{t}}{\partial \boldsymbol{S}_i}(\mathbf{V}_0\boldsymbol{\mu}^{t} + \boldsymbol{S}^{t}) && \in \mathbb{R}^{q\times q}\,.
\end{aligned}$$

Has the structure required by main theorem. Then prove

$$\forall t \in \mathbb{N} \quad \frac{1}{\sqrt{N}}\|\mathbf{X}^{t} - \mathbf{S}^{t} - \mathbf{V}_0\boldsymbol{\mu}^{t}\|_F \xrightarrow[N\to\infty]{P} 0$$

by induction.

Dependence on linear observations

$$
\begin{aligned}
\boldsymbol{X}^{t+1} &= \mathbf{A}\boldsymbol{M}^t - \boldsymbol{M}^{t-1}(\mathbf{b}^t)^\top && \in \mathbb{R}^{N\times q}\,, \\
\boldsymbol{M}^t &= f^t(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \boldsymbol{X}^t) && \in \mathbb{R}^{N\times q}\,, \\
\mathbf{b}^t &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial f_i^t}{\partial \boldsymbol{X}_i}(\varphi\left(\mathbf{A}\mathbf{W}_0\right), \boldsymbol{X}^t) && \in \mathbb{R}^{q\times q}\,.
\end{aligned}
$$

Dependence on linear observations

$$
\begin{aligned}
\boldsymbol{X}^{t+1} &= \mathbf{A}\boldsymbol{M}^{t} - \boldsymbol{M}^{t-1}(\mathbf{b}^{t})^{\top} && \in \mathbb{R}^{N\times q}, \\
\boldsymbol{M}^{t} &= f^{t}(\varphi(\mathbf{A}\mathbf{W}_0), \boldsymbol{X}^{t}) && \in \mathbb{R}^{N\times q}, \\
\mathbf{b}^{t} &= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f_i^{t}}{\partial \boldsymbol{X}_i}(\varphi(\mathbf{A}\mathbf{W}_0), \boldsymbol{X}^{t}) && \in \mathbb{R}^{q\times q}.
\end{aligned}
$$

State evolution

$$
\boldsymbol{\nu}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\mathbf{W}_0^{\top} f^{t}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t} + \mathbf{Z}^{t}\right)\right]
$$

$$
\hat{\boldsymbol{\nu}}^{t+1} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N} \frac{\partial f_i^{t}}{\partial \mathbf{Z}_{\mathbf{W}_0,i},\varphi}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t} + \mathbf{Z}^{t}\right)\right]
$$

$$
\begin{aligned}
&\boldsymbol{\kappa}^{t+1,s+1} = \\
&\lim_{N\to\infty} \frac{1}{N}\mathbb{E}\Bigg[\left(f^{s}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{s} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{s} + \mathbf{Z}^{s}\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{s+1}\right)^{\top} \\
&\qquad\qquad \left(f^{t}\left(\varphi(\mathbf{Z}_{\mathbf{W}_0}), \mathbf{Z}_{\mathbf{W}_0}\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t} + \mathbf{W}_0\hat{\boldsymbol{\nu}}^{t} + \mathbf{Z}^{t}\right) - \mathbf{W}_0\rho_{\mathbf{W}_0}^{-1}\boldsymbol{\nu}^{t+1}\right)\Bigg]
\end{aligned}
$$

Other directions

- finite size rates [Rush & Venkataramanan '18]
- subGaussian universality [Bayati, Lelarge & Montanari '15]
- rotationally invariant matrices [Rangan, Schniter & Fletcher '16], semi-random universality [Dudeja, Lu & Sen '22]

An application of non-separable non-linearities

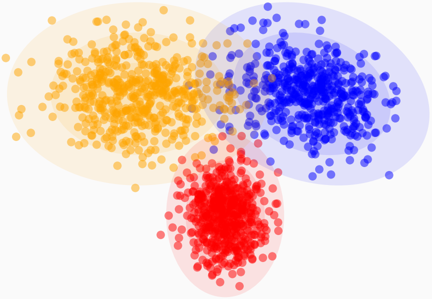
Classifying a high dimensional, non-isotropic Gaussian mixture

joint work with Loureiro. B., Sicuro. G., Pacco. A., Krzakala. F., Zdeborovà. L. *Neurips 2021*

Classifying Gaussian Mixtures with Convex GLM

Data and teacher

$$\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{y} \in \mathbb{R}^K \quad P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$



K=3, d=2

Classifying Gaussian Mixtures with Convex GLM

Data and teacher

$$\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{y} \in \mathbb{R}^K \quad P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}\left(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

Student

$$\boldsymbol{W}^\star \in \min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}} L\left(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}\right) + \frac{\lambda}{2} \|\boldsymbol{W}\|_2^2$$

Learn K separating hyperplanes, i.e. a matrix $\boldsymbol{W} \in \mathbb{R}^{d \times K}$

Examples : ridge regression, softmax with cross-entropy, ...

What are the difficulties ?

Learning a matrix : how are the different hyperplanes correlated/linked by the learning process ?

What are the difficulties ?

Learning a matrix : how are the different hyperplanes correlated/linked by the learning process ?

Different covariances : effect of each cluster cannot be characterized with the same quantities

What are the difficulties ?

Learning a matrix : how are the different hyperplanes correlated/linked by the learning process ?

Different covariances : effect of each cluster cannot be characterized with the same quantities

Convex Gaussian Comparison Inequalities break down beyond least-squares

[Thrampoulidis et al. 20] (identity covariances)

What are the difficulties ?

Learning a matrix : how are the different hyperplanes correlated/linked by the learning process ?

Different covariances : effect of each cluster cannot be characterized with the same quantities

Convex Gaussian Comparison Inequalities break down beyond least-squares

[Thrampoulidis et al. 20] (identity covariances)

Solve it with an AMP

Sketch of proof

Target :

$$\boldsymbol{W}^{\star} \in \min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}} L(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}) + r(\boldsymbol{W}) \tag{1}$$

Tool :

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \boldsymbol{Z}^{\top} \boldsymbol{h}_t(\boldsymbol{v}^t) - \boldsymbol{e}_t(\boldsymbol{u}^t)\langle \boldsymbol{h}_t' \rangle^{\top} \\ \boldsymbol{v}^t &= \boldsymbol{Z}\boldsymbol{e}_t(\boldsymbol{u}^t) - \boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle \boldsymbol{e}_t' \rangle^{\top} \end{aligned} \tag{2}$$

Sketch of proof

Target :

$$\boldsymbol{W}^\star \in \min_{\boldsymbol{W} \in \mathbb{R}^{d \times K}} L(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}) + r(\boldsymbol{W}) \tag{1}$$

Tool :

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \boldsymbol{Z}^\top \boldsymbol{h}_t(\boldsymbol{v}^t) - \boldsymbol{e}_t(\boldsymbol{u}^t)\langle \boldsymbol{h}_t' \rangle^\top \\ \boldsymbol{v}^t &= \boldsymbol{Z}\boldsymbol{e}_t(\boldsymbol{u}^t) - \boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle \boldsymbol{e}_t' \rangle^\top \end{aligned} \tag{2}$$

Instructions:

- design $\boldsymbol{h}_t, \boldsymbol{e}_t$ s.t. fixed point of (2) matches opt. cond. of (1)
- find a converging trajectory (convexity helps)
- use state evolution equations (fixed point)

Proof idea [Bayati and Montanari '11], see also [Feng et al. '22]

Designing the AMP : a quick look

Often designed from a factor graph

The factor graph for generic multiclass GLM is not obvious ...

Designing the AMP : a quick look

Often designed from a factor graph

The factor graph for generic multiclass GLM is not obvious ...

Reformulate the optimality condition

$$\boldsymbol{X}^\top \partial L(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}^\star) + \partial r(\boldsymbol{W}^\star) = 0$$

Designing the AMP : a quick look

Often designed from a factor graph

The factor graph for generic multiclass GLM is not obvious ...

Reformulate the optimality condition

$$\boldsymbol{X}^\top \partial L(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}^\star) + \partial r(\boldsymbol{W}^\star) = 0$$

Match it with the fixed point

$$(Id + \boldsymbol{e}(\bullet)\langle \boldsymbol{h}' \rangle)(\boldsymbol{u}) = Z^\top \boldsymbol{h}(\boldsymbol{v})$$
$$(Id + \boldsymbol{h}(\bullet)\langle \boldsymbol{e}' \rangle)(\boldsymbol{v}) = Z\boldsymbol{e}(\boldsymbol{u})$$

Validity of state evolution

Non-separable, block structure gradient

$$\boldsymbol{Z}^\top \begin{bmatrix} \partial\tilde{L}_1(\boldsymbol{Z}_1\tilde{\boldsymbol{W}}_1) & & & \\ & \partial\tilde{L}_2(\boldsymbol{Z}_2\tilde{\boldsymbol{W}}_2) & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{L}_K(\boldsymbol{Z}_K\tilde{\boldsymbol{W}}_K) \end{bmatrix} + \begin{bmatrix} \partial\tilde{r}(\tilde{\boldsymbol{W}})_1 & & & \\ & \partial\tilde{r}(\tilde{\boldsymbol{W}})_2 & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{r}(\tilde{\boldsymbol{W}})_K \end{bmatrix}$$

Validity of state evolution

Non-separable, block structure gradient

$$\boldsymbol{Z}^\top \begin{bmatrix} \partial\tilde{L}_1(\boldsymbol{Z}_1\tilde{\boldsymbol{W}}_1) & & & \\ & \partial\tilde{L}_2(\boldsymbol{Z}_2\tilde{\boldsymbol{W}}_2) & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{L}_K(\boldsymbol{Z}_K\tilde{\boldsymbol{W}}_K) \end{bmatrix} + \begin{bmatrix} \partial\tilde{r}(\tilde{\boldsymbol{W}})_1 & & & \\ & \partial\tilde{r}(\tilde{\boldsymbol{W}})_2 & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{r}(\tilde{\boldsymbol{W}})_K \end{bmatrix}$$

Spatially coupled, non-separable, matrix-valued, two-layer AMP

Converging trajectory

Prove $\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\|\mathbf{u}^t-\mathbf{u}^{t+1}\|_F^2=0$

Use SE equations and well chosen initialization

$$\lim_{d\to\infty}\frac{1}{d}\|\mathbf{u}^t-\mathbf{u}^{t+1}\|_F^2=\mathbf{C}-\mathbf{C}_{t,t+1}$$

$\mathbf{C}_{t,t+1}\overset{d\to+\infty}{=}\frac{1}{d}(\mathbf{u}^t)^\top\mathbf{u}^{t+1}=$ expectation over SE fields, $\mathbf{C}_{t,t}=\mathbf{C}$

Converging trajectory

Prove $\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\|\mathbf{u}^t-\mathbf{u}^{t+1}\|_F^2=0$

Use SE equations and well chosen initialization

$$\lim_{d\to\infty}\frac{1}{d}\|\mathbf{u}^t-\mathbf{u}^{t+1}\|_F^2=\mathbf{C}-\mathbf{C}_{t,t+1}$$

$\mathbf{C}_{t,t+1}\overset{d\to+\infty}{=}\frac{1}{d}(\mathbf{u}^t)^\top\mathbf{u}^{t+1}=$ expectation over SE fields, $\mathbf{C}_{t,t}=\mathbf{C}$

Prescribes an iteration $\mathbf{C}_{t,t+1}=\mathcal{O}\left(\mathbf{C}_{t,t-1}\right)$

In the strongly convex case, can prove $\mathcal{O}$ is a contraction

$\Downarrow$

Trajectory converges

Main result (informal)

Theorem [LSGPKZ21]

Fixed-point of self-consistent equations

$$\begin{cases} \boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\Sigma}_k\boldsymbol{G}^\top] \\ \boldsymbol{M}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\mu}_k] \\ \boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}\left[\left(\boldsymbol{G}\odot\left(\hat{\boldsymbol{Q}}_k\otimes\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}}\odot(\boldsymbol{I}_K\otimes\boldsymbol{\Sigma}_k)\right)\boldsymbol{\Xi}_k^\top\right] \end{cases} \qquad \begin{cases} \hat{\boldsymbol{Q}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{f}_k^\top\right] \\ \hat{\boldsymbol{V}}_k = -\alpha\rho_k\boldsymbol{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{\xi}^\top\right] \\ \hat{\boldsymbol{m}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right] \end{cases}$$

where $\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}}\odot\mathrm{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}}\odot\boldsymbol{B}),\ \mathbf{A}^{-1}\equiv\sum_k\hat{\boldsymbol{V}}_k\otimes\boldsymbol{\Sigma}_k,\ \boldsymbol{B}\equiv\sum_k\left(\boldsymbol{\mu}_k\hat{\boldsymbol{m}}_k^\top+\boldsymbol{\Xi}_k\odot\sqrt{\hat{\boldsymbol{Q}}_k\otimes\boldsymbol{\Sigma}_k}\right)$

$\boldsymbol{f}_k\equiv\boldsymbol{V}_k^{-1}(\boldsymbol{h}_k-\boldsymbol{\omega}_k),\ \boldsymbol{h}_k=\boldsymbol{V}_k^{1/2}\mathrm{Prox}_{\ell(\boldsymbol{e}_k,\boldsymbol{V}_k^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k),\ \boldsymbol{\omega}_k\equiv\boldsymbol{M}_k+\boldsymbol{b}+\boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k$

Main result (informal)

Theorem [LSGPKZ21]

Fixed-point of self-consistent equations

$$\begin{cases} \boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\Sigma}_k\boldsymbol{G}^\top] \\ \boldsymbol{M}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\mu}_k] \\ \boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}\left[\left(\boldsymbol{G}\odot\left(\hat{\boldsymbol{Q}}_k\otimes\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}}\odot(\boldsymbol{I}_K\otimes\boldsymbol{\Sigma}_k)\right)\boldsymbol{\Xi}_k^\top\right] \end{cases} \qquad \begin{cases} \hat{\boldsymbol{Q}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{f}_k^\top\right] \\ \hat{\boldsymbol{V}}_k = -\alpha\rho_k\boldsymbol{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{\xi}^\top\right] \\ \hat{\boldsymbol{m}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right] \end{cases}$$

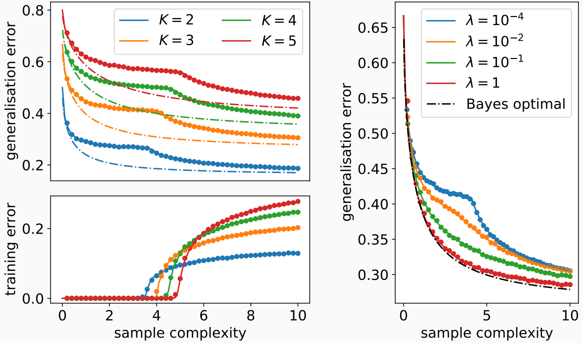
where $\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}}\odot\mathrm{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}}\odot\boldsymbol{B}),\ \mathbf{A}^{-1}\equiv\sum_k\hat{\boldsymbol{V}}_k\otimes\boldsymbol{\Sigma}_k,\ \boldsymbol{B}\equiv\sum_k\left(\boldsymbol{\mu}_k\hat{\boldsymbol{m}}_k^\top+\boldsymbol{\Xi}_k\odot\sqrt{\hat{\boldsymbol{Q}}_k\otimes\boldsymbol{\Sigma}_k}\right)$

$\boldsymbol{f}_k\equiv\boldsymbol{V}_k^{-1}(\boldsymbol{h}_k-\boldsymbol{\omega}_k),\ \boldsymbol{h}_k=\boldsymbol{V}_k^{1/2}\mathrm{Prox}_{\ell(\boldsymbol{e}_k,\boldsymbol{V}_k^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k),\ \boldsymbol{\omega}_k\equiv\boldsymbol{M}_k+\boldsymbol{b}+\boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k$

Training and generalization for $n,d\to\infty$:

$$\epsilon_t = 1-\sum_{k=1}^K\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{h}_k)\right], \quad \epsilon_g = 1-\sum_{k=1}^K\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\hat{y}_k(\boldsymbol{\omega}_k)\right].$$

Examples : synthetic random design problems



Ridge penalized logistic regression on K Gaussian clusters, $\Sigma_k = \Delta Id$. (Left) Sample complexity (Right) Regularization

Related works :[T. Cover '69][E. Gardner, B. Derrida '89] [EJ. Candès, P. Sur '20] [F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, L. Zdeborova '20][C. Thrampoulidis, S. Oymak, M. Soltanolkotabi '20]

Thank you