Statistical physics of learning : a mathematical perspective

Cedric Gerbelot

ENS & INRIA, Paris, France

April 7, 2022

An introductory example : binary classification on MNIST

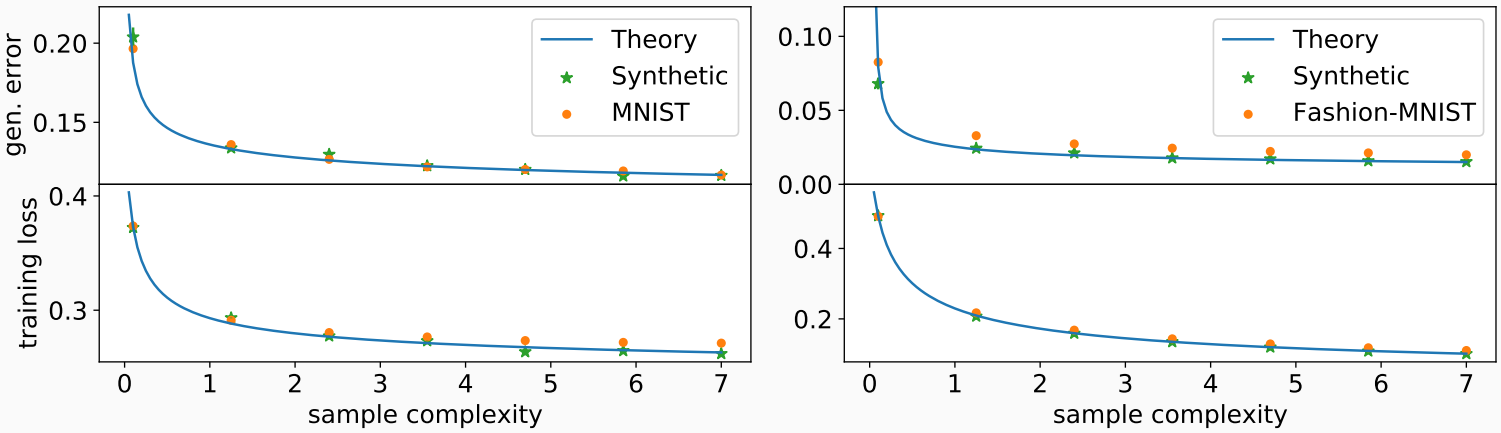


Figure 1: Binary classification on Mnist/Fashion-Mnist, odd vs even, ℓ_2 regularized logistic regression

An introductory example : binary classification on MNIST

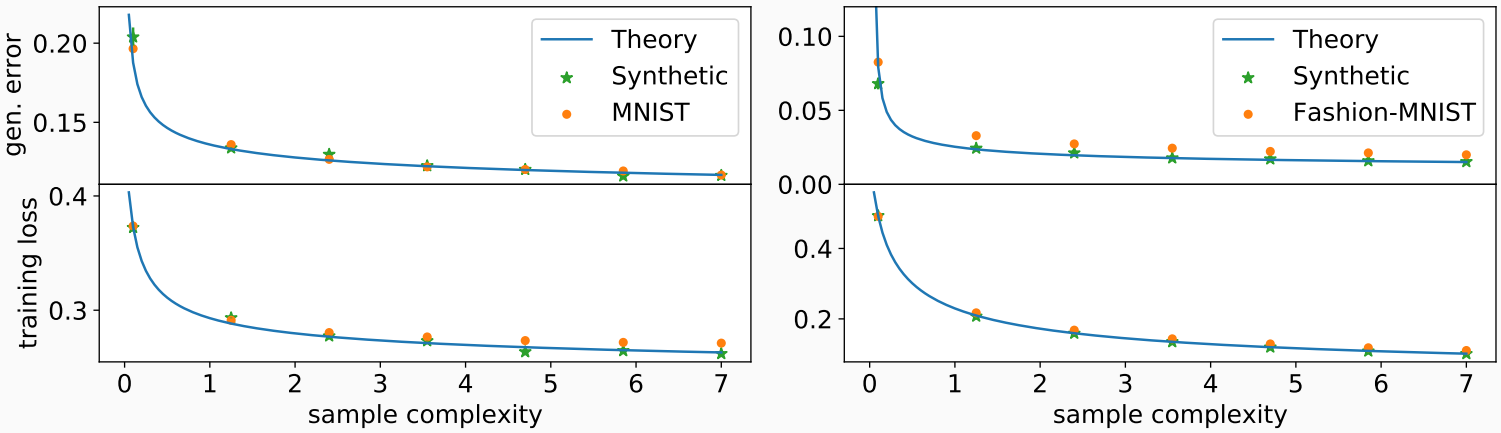


Figure 1: Binary classification on Mnist/Fashion-Mnist, odd vs even, ℓ_2 regularized logistic regression

Exact theory ? So what's the answer ?

[B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, L. Zdeborova '21]

An introductory example : binary classification on MNIST

Theorem 1 (Concentration properties of the estimator). *Let $\boldsymbol{\xi}_{k\in[K]} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$ be collection of K-dimensional standard normal vectors independent of other quantities. Let also be $\{\boldsymbol{\Xi}_k\}$ a set of K matrices, $\boldsymbol{\Xi}_k \in \mathbb{R}^{K\times d}$, with i.i.d. standard normal entries, independent of other quantities. Under the set of assumptions (A1–A5), for any pseudo-Lispchitz functions of finite order $\phi_1 : \mathbb{R}^{K\times d} \to \mathbb{R}, \phi_2 : \mathbb{R}^{K\times n} \to \mathbb{R}$, the estimator $\boldsymbol{W}^\star$ and the matrix $\boldsymbol{Z}^\star = \frac{1}{\sqrt{d}}\boldsymbol{W}^\star\boldsymbol{X}$ verify:*

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\boldsymbol{G})\right], \qquad \phi_2(\boldsymbol{Z}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\xi}}\left[\phi_2(\boldsymbol{H})\right], \tag{6}$$

where we have introduced the proximal for the loss:

$$\boldsymbol{h}_k = \boldsymbol{V}_k^{1/2}\,\mathrm{Prox}_{\ell(\boldsymbol{e}_k, \boldsymbol{V}_k^{1/2}\bullet)}(\boldsymbol{V}_k^{-1/2}\boldsymbol{\omega}_k) \in \mathbb{R}^K, \qquad \boldsymbol{\omega}_k \equiv \boldsymbol{m}_k + \boldsymbol{b} + \boldsymbol{Q}_k^{1/2}\boldsymbol{\xi}_k, \tag{7}$$

and $\boldsymbol{H} \in \mathbb{R}^{K\times n}$ is obtained by concatenating each $\boldsymbol{h}_k$, $\rho_k n$ times. We have also introduced the matrix proximal $\boldsymbol{G} \in \mathbb{R}^{K\times d}$:

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

The collection of parameters $(\boldsymbol{Q}_k, \boldsymbol{m}_k, \boldsymbol{V}_k, \hat{\boldsymbol{Q}}_k, \hat{\boldsymbol{m}}_k, \hat{\boldsymbol{V}}_k)_{k\in[K]}$ is given by the fixed point of the following self-consistent equations:

$$\begin{cases} \boldsymbol{Q}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\Sigma}_k\boldsymbol{G}^\top] \\ \boldsymbol{m}_k = \frac{1}{\sqrt{d}}\mathbb{E}_{\boldsymbol{\Xi}}[\boldsymbol{G}\boldsymbol{\mu}_k] \\ \boldsymbol{V}_k = \frac{1}{d}\mathbb{E}_{\boldsymbol{\Xi}}\left[\left(\boldsymbol{G} \odot \left(\hat{\boldsymbol{Q}}_k \otimes \boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}} \odot (\mathbf{I}_K \otimes \boldsymbol{\Sigma}_k)\right)\boldsymbol{\Xi}_k^\top\right] \end{cases} \quad \begin{cases} \hat{\boldsymbol{Q}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{f}_k^\top\right] \\ \hat{\boldsymbol{V}}_k = -\alpha\rho_k\boldsymbol{Q}_k^{-\frac{1}{2}}\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\boldsymbol{\xi}^\top\right] \\ \hat{\boldsymbol{m}}_k = \alpha\rho_k\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{f}_k\right] \end{cases} \tag{8}$$

where $\boldsymbol{f}_k \equiv \boldsymbol{V}_k^{-1}(\boldsymbol{h}_k - \boldsymbol{\omega}_k)$, and the vector $\boldsymbol{b}^\star$ is such that $\sum_k \rho_k \mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{V}_k\boldsymbol{f}_k\right] = \mathbf{0}$ holds.

Figure 2: Main theorem from [B. Loureiro, G. Sicuro, C. Gerbelot, A. Pacco, F. Krzakala, L. Zdeborova '21]

An introductory example : binary classification on MNIST

What's the motivation for these formulas ? Why are they useful ? How are they obtained ?

The starting point : probabilistic inference

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

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w}) \tag{1}$$

The starting point : probabilistic inference

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

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w}) \tag{1}$$

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

$$\mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) \tag{2}$$

Estimate $\mathbf{w}$?

The starting point : probabilistic inference

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

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w}) \tag{1}$$

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

$$\mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) \tag{2}$$

Estimate $\mathbf{w}$?

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

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) \, d\mu(\mathbf{w}) \tag{3}$$

The starting point : probabilistic inference

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

$$\mathbf{w} \sim p_{\mathbf{w}}(\mathbf{w}) \tag{1}$$

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

$$\mathbf{y} \sim p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) \tag{2}$$

Estimate $\mathbf{w}$?

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

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} p_{\mathbf{w}}(\mathbf{w}) p_{\mathbf{y}}(\mathbf{y}|\mathbf{w}) \, d\mu(\mathbf{w}) \tag{3}$$

Problem : This is a high-dimensional integral !

Link with statistical physics

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

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w}) \tag{4}$$

Link with statistical physics

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

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w}) \tag{4}$$

- Equilibrium Boltzmann 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

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

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w}) \tag{4}$$

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

Stat. phys. toolbox deals with this problem

Link with statistical physics : disordered systems

- distributions involve a dense, random interaction matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$
- $\mathbf{y}$ can come from another stochastic model, i.e.

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

- estimate the generative model with postulated densities g, f

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w}) \tag{5}$$

Link with statistical physics : disordered systems

- distributions involve a dense, random interaction matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$
- $\mathbf{y}$ can come from another stochastic model, i.e.

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

- estimate the generative model with postulated densities g, f

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{w}) \tag{5}$$

- *Disordered* equilibrium Boltzmann measure
- toolbox : replica/cavity method, belief-propagation, etc...
- focus on $\mathbf{X}$ with i.i.d. $\mathcal{N}(0, 1)$ elements

What is meant by "solving" ?

Reducing the dimension and exact computation

Central limit theorem:
If $\{w_i\}_{i=1}^{d}$ i.i.d. in L^2, then

$$\frac{1}{\sqrt{d}}\left(\sum_{i=1}^{d} w_i - d\mathbb{E}\left[w\right]\right) \xrightarrow[d\to+\infty]{\text{in law}} \mathcal{N}(0,\sigma^2) \tag{6}$$

What is meant by "solving" ?

Reducing the dimension and exact computation

Central limit theorem:
If $\{w_i\}_{i=1}^{d}$ i.i.d. in L^2, then

$$\frac{1}{\sqrt{d}}\left(\sum_{i=1}^{d} w_i - d\mathbb{E}\left[w\right]\right) \xrightarrow[d\to+\infty]{\text{in law}} \mathcal{N}(0,\sigma^2) \tag{6}$$

Concentration of measure :

$$\Phi(\mathbf{w}) \xrightarrow[d\to\infty]{P} \mathbb{E}\left[\Phi(\mathbf{w})\right] = \mathbb{E}\left[\phi(w)\right] \tag{7}$$

for sufficiently regular Φ and fast decaying $p_{\mathbf{w}}$.

What is meant by "solving" ?

Reducing the dimension and exact computation

Central limit theorem:
If $\{w_i\}_{i=1}^d$ i.i.d. in L^2, then

$$\frac{1}{\sqrt{d}}\left(\sum_{i=1}^{d} w_i - d\mathbb{E}[w]\right) \xrightarrow[d\to+\infty]{\text{in law}} \mathcal{N}(0,\sigma^2) \tag{6}$$

Concentration of measure :

$$\Phi(\mathbf{w}) \xrightarrow[d\to\infty]{P} \mathbb{E}[\Phi(\mathbf{w})] = \mathbb{E}[\phi(w)] \tag{7}$$

for sufficiently regular Φ and fast decaying $p_{\mathbf{w}}$.

Want the same thing for $\hat{\mathbf{w}}$

What is meant by "solving" ?

Reducing the dimension and exact computation

Central limit theorem:
If $\{w_i\}_{i=1}^d$ i.i.d. in L^2, then

$$\frac{1}{\sqrt{d}}\left(\sum_{i=1}^{d} w_i - d\mathbb{E}[w]\right) \xrightarrow[d\to+\infty]{\text{in law}} \mathcal{N}(0,\sigma^2) \tag{6}$$

Concentration of measure :

$$\Phi(\mathbf{w}) \xrightarrow[d\to\infty]{P} \mathbb{E}[\Phi(\mathbf{w})] = \mathbb{E}[\phi(w)] \tag{7}$$

for sufficiently regular Φ and fast decaying $p_{\mathbf{w}}$.

Want the same thing for $\hat{\mathbf{w}}$

Problem : $\hat{\mathbf{w}}$ is not i.i.d.! Strong coupling notably due to X

Finding a decoupled measure

Stat.phys. toolbox (and ensuing math !!)

$$\Downarrow$$

Product measure of simple components describing $\hat{\mathbf{w}}$ for large n, d

Finding a decoupled measure

Stat.phys. toolbox (and ensuing math !!)

$\Downarrow$

Product measure of simple components describing $\hat{\mathbf{w}}$ for large n, d

Theorems then take the form

$$\Phi(\hat{\mathbf{w}}) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}\left[\Phi\left(\Omega(\mathbf{Z}), \{q_s\}_{s\in\mathcal{S}}\right)\right]$$

$$\Phi(\hat{\mathbf{w}}) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}\left[\phi\left(\omega(z), \{q_s\}_{s\in\mathcal{S}}\right)\right] \quad \text{if separable problem, 1D integral}$$

Finding a decoupled measure

Stat.phys. toolbox (and ensuing math !!)

$\Downarrow$

Product measure of simple components describing $\hat{\mathbf{w}}$ for large n, d

Theorems then take the form

$$\Phi(\hat{\mathbf{w}}) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}\left[\Phi\left(\Omega(\mathbf{Z}), \{q_s\}_{s\in\mathcal{S}}\right)\right]$$

$$\Phi(\hat{\mathbf{w}}) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}\left[\phi\left(\omega(z), \{q_s\}_{s\in\mathcal{S}}\right)\right] \quad \text{if separable problem, 1D integral}$$

- $\mathbf{Z} \in \mathbb{R}^d$ i.i.d. $\mathcal{N}(0,1)$
- Ω explicit, bounded variation function
- q_s are low-dimensional paramaters
- given by explicit, self-consistent equations

Back to machine learning and inference

Recall

$$\hat{\mathbf{w}} = \frac{1}{\mathcal{Z}(\mathbf{y})} \int \mathbf{w} \exp(-\beta \left(g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w})\right)) d\mu(\mathbf{x}) \tag{8}$$

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

$$\hat{\mathbf{w}} \quad \xrightarrow[\beta \to +\infty]{} \quad \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \quad g(\mathbf{X}\mathbf{w}, \mathbf{y}) + f(\mathbf{w}) \tag{9}$$

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

Examples : LASSO, logistic regression, etc ...

From the decoupled measure : training, generalization error, phenomenology, etc ...

Stat. Phys./High-Dimensional Approach

- typical case
- benchmark, random design problems
- exact solutions
- strong assumptions (i.i.d. Gaussian !)

Stat. Phys./High-Dimensional Approach

- typical case
- benchmark, random design problems
- exact solutions
- strong assumptions (i.i.d. Gaussian !)

How realistic are the stat. phys. benchmarks ?
What can we do to make them more realistic ?

Teacher Student Generalized Linear Model

Observe "*teacher*" generative model

$$\boldsymbol{y} = f_0(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n, \quad \boldsymbol{w}_0 \in \mathbb{R}^d \quad \boldsymbol{X} \in \mathbb{R}^{n \times d} \text{ i.i.d. } \mathcal{N}(0,1)$$

Teacher Student Generalized Linear Model

Observe "*teacher*" generative model

$$\boldsymbol{y} = f_0(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n, \quad \boldsymbol{w}_0 \in \mathbb{R}^d \quad \boldsymbol{X} \in \mathbb{R}^{n \times d} \text{ i.i.d. } \mathcal{N}(0,1)$$

Learn with "*student*"

$$\boldsymbol{w}^\star \in \underset{\boldsymbol{w} \in \mathbb{R}^d}{\operatorname{argmin}} \, L\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}\right) + r(\boldsymbol{w})$$

- L, r are a convex loss and penalty
- $n, d \to +\infty$ with fixed ratio

Teacher Student Generalized Linear Model

Observe "*teacher*" generative model

$$\boldsymbol{y} = f_0(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n, \quad \boldsymbol{w}_0 \in \mathbb{R}^d \quad \boldsymbol{X} \in \mathbb{R}^{n \times d} \text{ i.i.d. } \mathcal{N}(0,1)$$

Learn with "*student*"

$$\boldsymbol{w}^\star \in \underset{\boldsymbol{w} \in \mathbb{R}^d}{\operatorname{argmin}} \, L\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}\right) + r(\boldsymbol{w})$$

- L, r are a convex loss and penalty
- $n, d \to +\infty$ with fixed ratio

Goal : statistical properties of $\boldsymbol{w}^\star$

Teacher Student Generalized Linear Model

Observe "*teacher*" generative model

$$\boldsymbol{y} = f_0(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n, \quad \boldsymbol{w}_0 \in \mathbb{R}^d \quad \boldsymbol{X} \in \mathbb{R}^{n \times d} \text{ i.i.d. } \mathcal{N}(0,1)$$

Learn with "*student*"

$$\boldsymbol{w}^\star \in \underset{\boldsymbol{w} \in \mathbb{R}^d}{\operatorname{argmin}} \, L\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}\right) + r(\boldsymbol{w})$$

- L, r are a convex loss and penalty
- $n, d \to +\infty$ with fixed ratio

Goal : statistical properties of $\boldsymbol{w}^\star$

Beyond i.i.d. assumption : introduce correlation

Introducing Correlation : a Block Covariance Model

Teacher and student with different feature spaces

Block covariate model proposed in

[B. Loureiro, **CG**, H. Cui, S. Goldt, M. Mézard, F. Krzakala, L. Zdeborova '21]

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{bmatrix} \in \mathbb{R}^{p+d} \sim \mathcal{N}\left(0, \begin{bmatrix} \Psi & \Phi \\ \Phi^\top & \Omega \end{bmatrix}\right) \quad y^\mu = f_0\left(\frac{1}{\sqrt{p}} \boldsymbol{w}_0^\top \boldsymbol{u}^\mu\right) ,$$

$$\boldsymbol{w}^\star = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left[\sum_{\mu=1}^{n} l\left(\frac{\boldsymbol{w}^\top \boldsymbol{v}^\mu}{\sqrt{d}}, y^\mu\right) + r(\boldsymbol{w})\right]$$

Many works: [E. Dobriban, S. Wager '15][PL. Bartlett, PM. Long, G. Lugosi, A. Tsigler '19][T. Hastie, A. Montanari, S. Rosset, RJ. Tibshirani '19][M. Celentano, A. Montanari, Y. Wei '20]

Solution to Block Covariance model

Theorem (informal)[B. Loureiro, C.Gerbelot, H. Cui, S. Goldt, M. Mézard, F. Krzakala, L. Zdeborova '21]

Unique fixed point of self-consistent equations

$$\begin{cases} V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right] \\ m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right] \\ q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{(\lambda+\hat{V}\omega)^2}\right] \end{cases}, \quad \begin{cases} \hat{V} = \frac{\alpha}{V}(1-\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[z'(V,m,q)]) \\ \hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[sz(V,m,q)-\frac{m}{\sqrt{\rho}}z'(V,m,q)\right] \\ \hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s+\sqrt{q-\frac{m^2}{\rho}}h-z(V,m,q)\right)^2\right] \end{cases}$$

where $z(V,m,q) = \text{prox}_{Vl(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q-\rho^{-1}m^2}h)$

Solution to Block Covariance model

Theorem (informal)[B. Loureiro, C.Gerbelot, H. Cui, S. Goldt, M. Mézard, F. Krzakala, L. Zdeborova '21]

Unique fixed point of self-consistent equations

$$\begin{cases} V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right] \\ m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right] \\ q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega+\hat{q}\omega^2}{(\lambda+\hat{V}\omega)^2}\right] \end{cases}, \quad \begin{cases} \hat{V} = \frac{\alpha}{V}(1-\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[z'(V,m,q)]) \\ \hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[sz(V,m,q)-\frac{m}{\sqrt{\rho}}z'(V,m,q)\right] \\ \hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s+\sqrt{q-\frac{m^2}{\rho}}h-z(V,m,q)\right)^2\right] \end{cases}$$

where $z(V,m,q) = \text{prox}_{Vl(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q-\rho^{-1}m^2}h)$

$n, p, d \to \infty$, training and generalization error :

$$\mathcal{E}_{\text{train.}}(\hat{\mathbf{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[l\left(\text{prox}_{V^\star l(.,f_0(\sqrt{\rho}s))}\left(\frac{m^\star}{\sqrt{\rho}}s+\sqrt{q^\star-\frac{m^{\star\,2}}{\rho}}h\right), f_0(\sqrt{\rho}s)\right)\right]$$

$$\mathcal{E}_{\text{gen.}}(\hat{\mathbf{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{(\nu,\lambda)}\left[\hat{g}\left(\hat{f}(\lambda), f_0(\nu)\right)\right]$$

How well does it work ?

Ridge regression works well ...

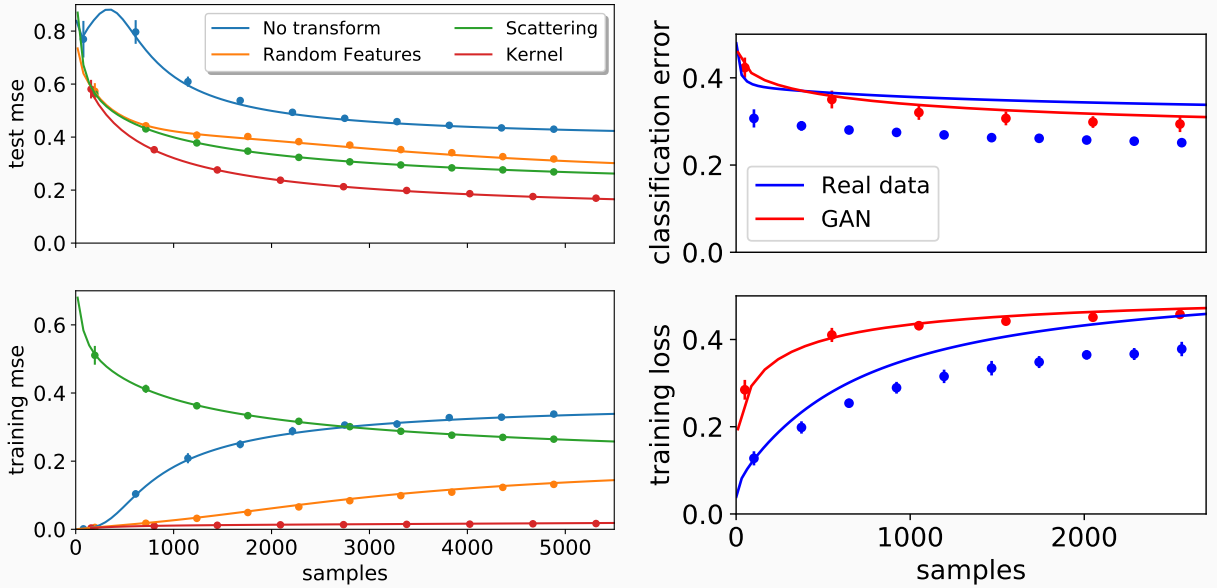


Figure 3: (Left) Ridge regression on real data. (Right) Logistic regression with real and synthetic (GAN) data

... but classification is more problematic

How well does it work ?

Ridge regression works well ...

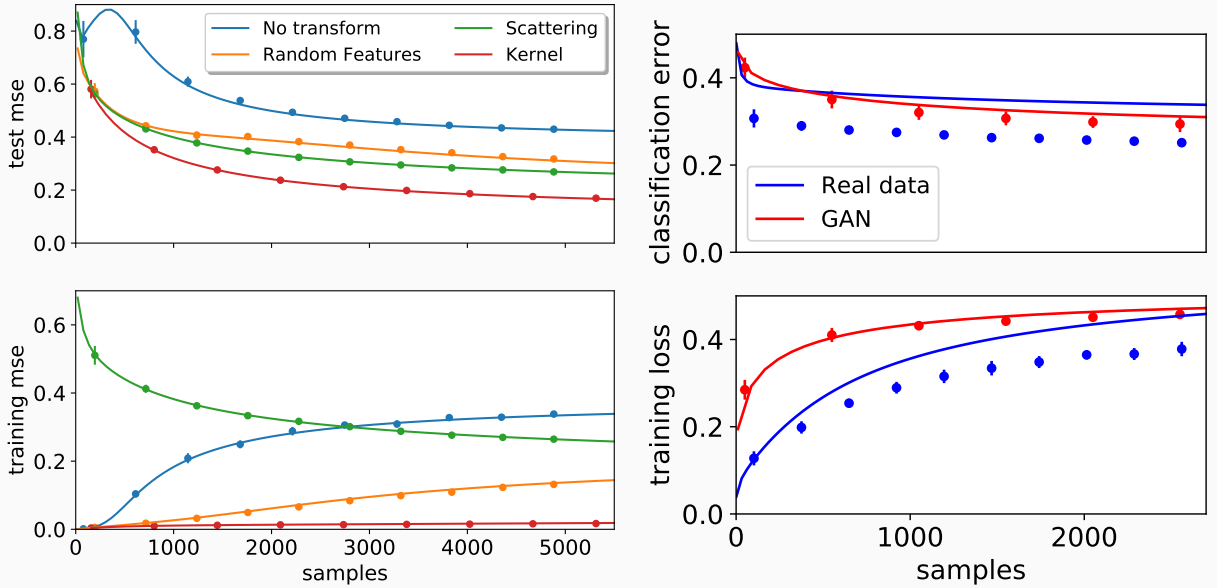


Figure 3: (Left) Ridge regression on real data. (Right) Logistic regression with real and synthetic (GAN) data

... but classification is more problematic

Need for another realistic benchmark problem

Gaussian mixtures are meaningful

Study classification of k-Gaussian mixture with convex GLM

Gaussian mixtures are meaningful

Study classification of k-Gaussian mixture with convex GLM

- Benchmark problem in ML, universal approximation, ...
- many scenarios described by Gaussian mixtures (GANs, 'Neural collapse', ...)

[M. Seddik, C. Louart, M. Tamaazousti, R. Couillet, '20][V. Papyan, X. Han, D.Donoho, '20]

Classifying Gaussian Mixtures with Convex GLM

Data and teacher

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

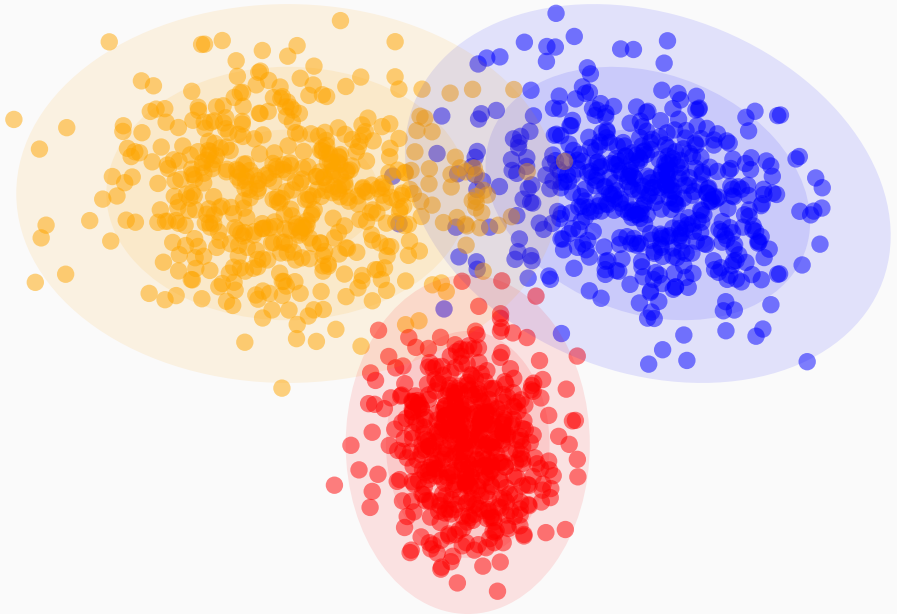


Figure 4: K=3, d=2

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

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

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

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

Main result (informal)

Theorem [B. Loureiro, G. Sicuro, CG, A. Pacco, F. Krzakala, L. Zdeborova '21]

Fixed-point of self-consistent equations

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

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

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

Main result (informal)

Theorem [B. Loureiro, G. Sicuro, CG, A. Pacco, F. Krzakala, L. Zdeborova '21]

Fixed-point of self-consistent equations

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

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 $n, d \to \infty$:

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

Main result : the intuition (forget dimensions)

For any convex GLM (separable, covariance, matrix variable, etc ...)

Same global form of the result

Main result : the intuition (forget dimensions)

For any convex GLM (separable, covariance, matrix variable, etc ...)

Same global form of the result

Problem defined by :

- estimator $\hat{\boldsymbol{w}}$
- "ground-truth" $\boldsymbol{w}_0/\boldsymbol{\mu}$, random data with covariance Σ
- cost function $L + r$.

Main result : the intuition (forget dimensions)

For any convex GLM (separable, covariance, matrix variable, etc ...)

Same global form of the result

Problem defined by :

- estimator $\hat{\boldsymbol{w}}$
- "ground-truth" $\boldsymbol{w}_0/\boldsymbol{\mu}$, random data with covariance Σ
- cost function $L + r$.

Statement

$$\hat{\boldsymbol{w}} \sim \text{nonlinearity}_{L,r}(\alpha * \text{ground-truth} + \text{gaussian})$$

With low-dim. closed form parameters.

Main result : a closer look

For any "well-behaved" observable ϕ_1:

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\boldsymbol{G})\right]$$

Main result : a closer look

For any "well-behaved" observable ϕ_1:

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\boldsymbol{G})\right]$$

$$\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}} \odot \mathrm{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}} \odot \boldsymbol{B}),$$

$$\mathbf{A}^{-1} \equiv \sum_k \hat{\boldsymbol{V}}_k \otimes \boldsymbol{\Sigma}_k, \qquad \boldsymbol{B} \equiv \sum_k \left(\boldsymbol{\mu}_k \hat{\boldsymbol{m}}_k^\top + \boldsymbol{\Xi}_k \odot \sqrt{\hat{\boldsymbol{Q}}_k \otimes \boldsymbol{\Sigma}_k}\right).$$

Main result : a closer look

For any "well-behaved" observable ϕ_1:

$$\phi_1(\boldsymbol{W}^\star) \xrightarrow[n,d\to+\infty]{P} \mathbb{E}_{\boldsymbol{\Xi}}\left[\phi_1(\boldsymbol{G})\right]$$

$$\boldsymbol{G} = \mathbf{A}^{\frac{1}{2}} \odot \mathrm{Prox}_{r(\mathbf{A}^{\frac{1}{2}}\odot\bullet)}(\mathbf{A}^{\frac{1}{2}} \odot \boldsymbol{B}),$$

$$\mathbf{A}^{-1} \equiv \sum_k \hat{\boldsymbol{V}}_k \otimes \boldsymbol{\Sigma}_k, \qquad \boldsymbol{B} \equiv \sum_k \left(\boldsymbol{\mu}_k \hat{\boldsymbol{m}}_k^\top + \boldsymbol{\Xi}_k \odot \sqrt{\hat{\boldsymbol{Q}}_k \otimes \boldsymbol{\Sigma}_k}\right).$$

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

where

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

Main result : important points

- very generic statement
- greatly simplifies with assumptions on covariances, separability of functions, ...
- in most cases reduces to low dimensional statement

Examples : synthetic random design problems

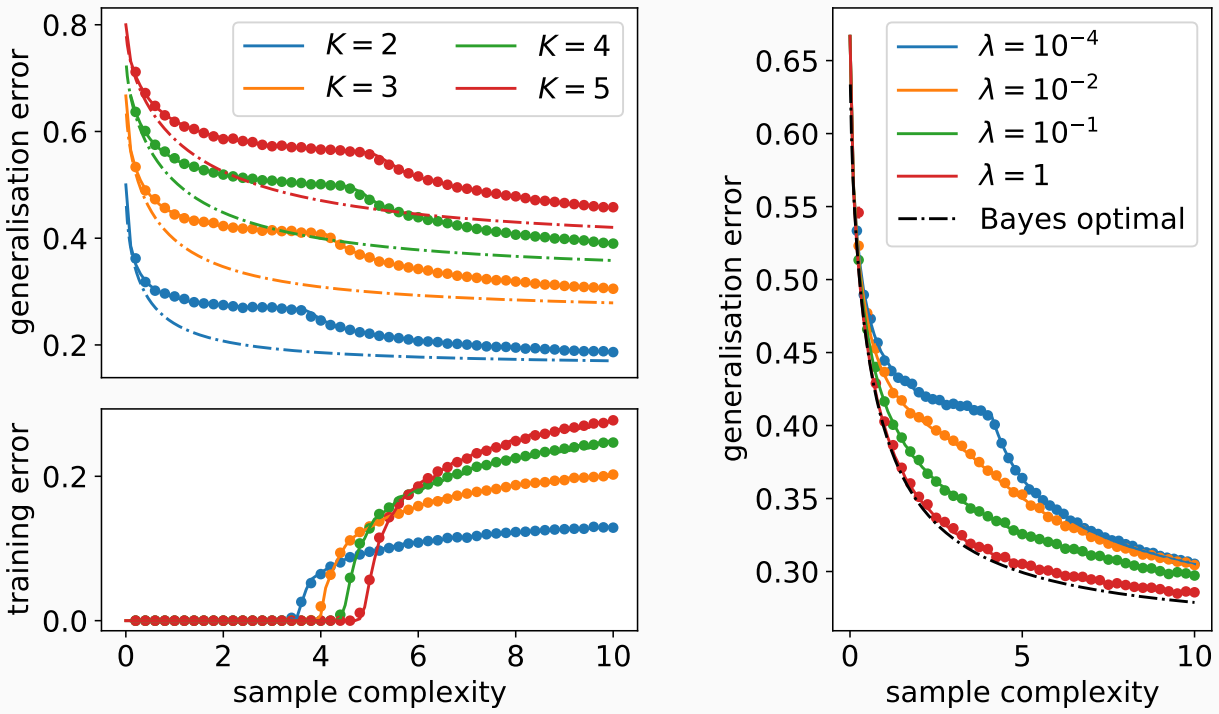


Figure 5: Ridge penalized logistic regression on K Gaussian clusters, $\Sigma_k = \Delta Id$. (Left) Sample complexity (Right) Regularization

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]

Examples : real data

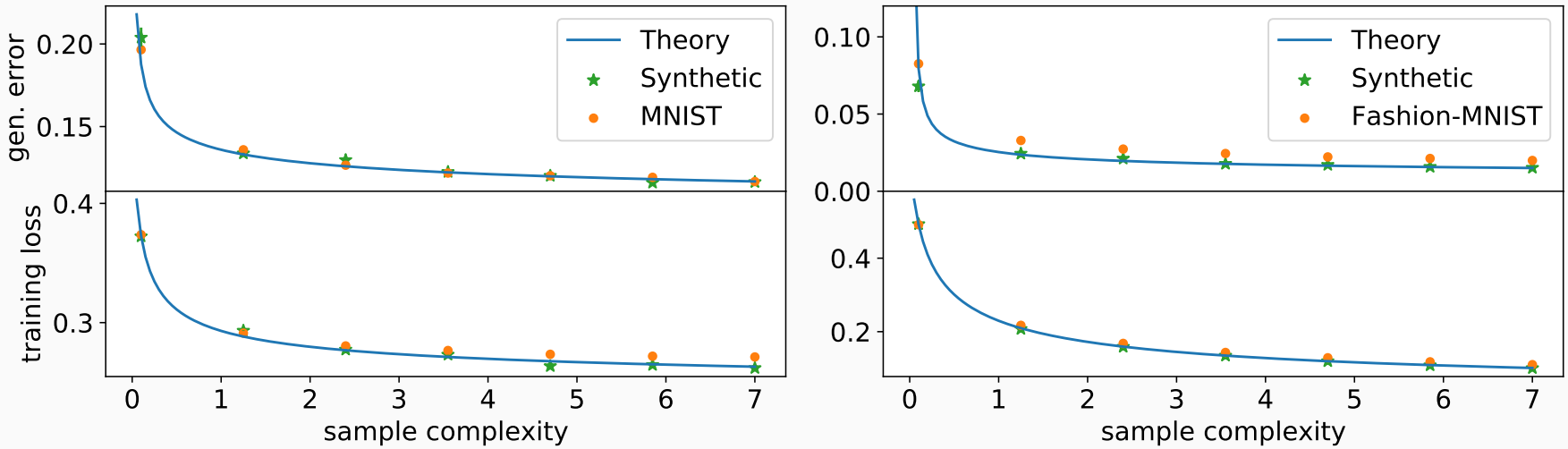


Figure 6: Binary classification on Mnist/Fashion-Mnist, odd vs even, Gaussian approximation and real data

Examples : real data

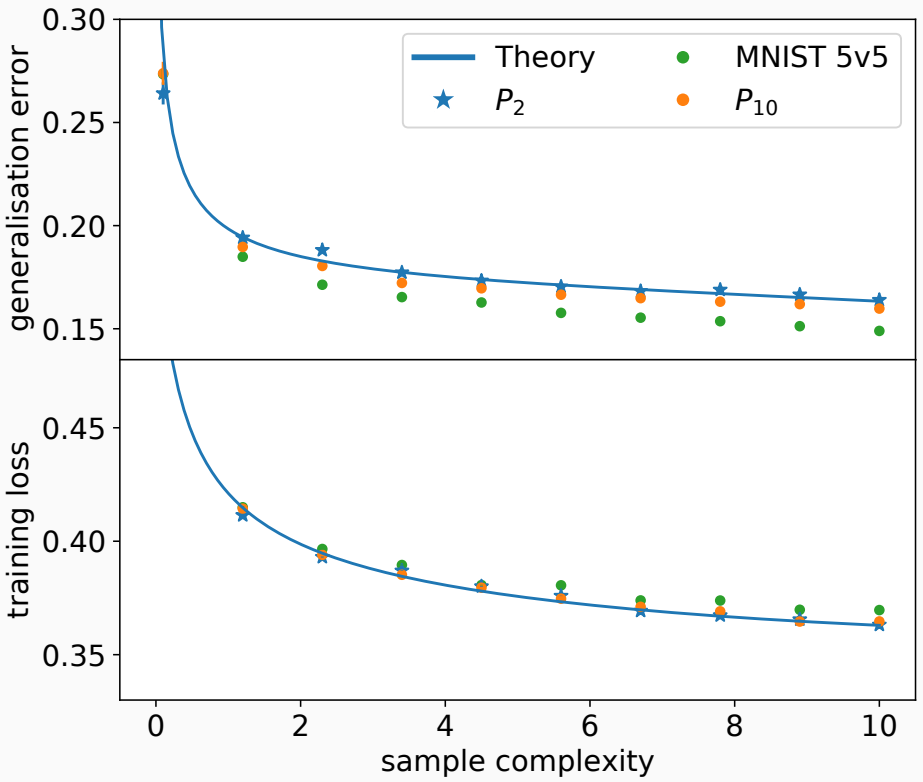


Figure 7: Adding more clusters to the Gaussian approximation

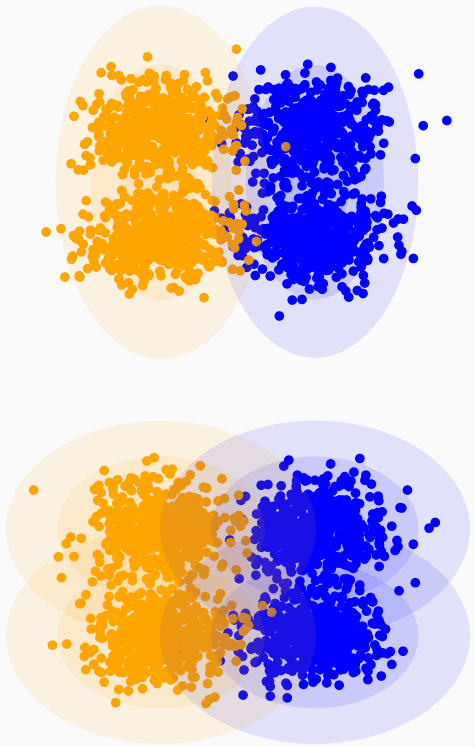


Figure 8: Idealized view

A few comments

- correlated Gaussian designs are meaningful
- study the phenomenology : sample complexity, regularization, etc...
- more models : ensembling, boosting
- dataset/feature map distribution AND predictor geometry
- both important for *Gaussian equivalence*

Ongoing research : what feature maps and predictors have Gaussian equivalence properties ?

The proof methods : a few words

Recall the problem : find a decoupled/product measure

- Guerra interpolation
- Gordon minmax
- cavity method
- approximate message passing iterations

Focus on AMP iterations

The proof methods : a few words

Recall the problem : find a decoupled/product measure

- Guerra interpolation - interpolate with a known replica solution
- Gordon minmax - Gaussian comparison inequalities
- cavity method - effect of a single variable among d
- approximate message passing iterations - more detail shortly

Focus on AMP iterations

Initially a relaxation of belief propagation equations

AMP iterations and SE equations

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

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t \mathbf{m}^{t-1} \tag{10}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{11}$$

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

$$b_t = \mathrm{div}\left[f_t(\mathbf{x}^t)\right] \tag{12}$$

AMP iterations and SE equations

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

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t \mathbf{m}^{t-1} \tag{10}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{11}$$

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

$$b_t = \mathrm{div}\left[f_t(\mathbf{x}^t)\right] \tag{12}$$

Theorem : for any t, $\mathbf{x}^t \xrightarrow[N\to\infty]{PLk} \mathbf{Z}^t \sim \mathcal{N}(0, \kappa_{t,t}\mathbf{I}_N)$

$$\text{where} \quad \kappa_{t,t} = \lim_{N\to\infty} \frac{1}{N}\left[f^{t-1}(\mathbf{Z}^{t-1})^\top f^{t-1}(\mathbf{Z}^{t-1})\right] \tag{13}$$

First proof due to E.Bolthausen '09, '14, M. Bayati & A. Montanari '11

Sketch of proof : Bolthausen conditioning

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{14}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{15}$$

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

$$\mathbf{x}^{t+1} = \mathbf{A}\mathbf{m}^t - b_t\mathbf{m}^{t-1} \tag{14}$$

$$\mathbf{m}^t = f_t(\mathbf{x}^t) \tag{15}$$

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

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

Sketch of proof : Bolthausen conditioning

A bit of algebra leads to

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

- 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

Moving forward on AMP iterations

Advantages:

- very generic and adaptable (non-convex, multilayer, committee,...)
- algorithm + fixed points
- control of trajectory understood in the convex case

Moving forward on AMP iterations

Advantages:

- very generic and adaptable (non-convex, multilayer, committee,...)
- algorithm + fixed points
- control of trajectory understood in the convex case

Disadvantage:

- **state evolution proofs are tedious, case by case basis**
- control of trajectory in non-convex/RS case ? (SK above AT line)

Moving forward on AMP iterations

Advantages:

- very generic and adaptable (non-convex, multilayer, committee,...)
- algorithm + fixed points
- control of trajectory understood in the convex case

Disadvantage:

- **state evolution proofs are tedious, case by case basis**
- control of trajectory in non-convex/RS case ? (SK above AT line)

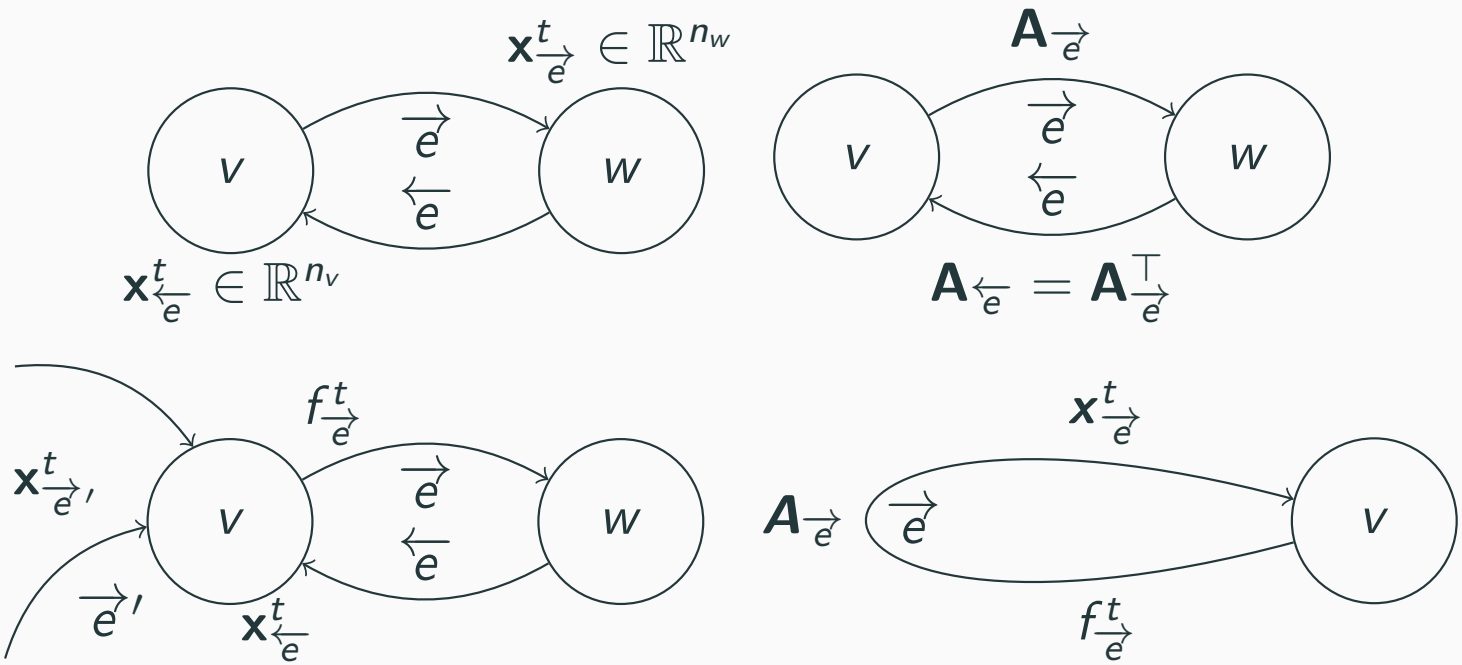
Solution

- unifying framework for AMP iterations
- generic, modular proof of SE equations

([C. Gerbelot, R. Berthier '21])

Graph-based AMP [C. Gerbelot, R. Berthier '21]

The oriented graph:



Arbitrary composition of this structure

Graph-based AMP [C. Gerbelot, R. Berthier '21]

The graph-based AMP iteration

$$\mathbf{x}_{\overrightarrow{e}}^{t+1} = \boldsymbol{A}_{\overrightarrow{e}} \mathbf{m}_{\overrightarrow{e}}^{t} - b_{\overrightarrow{e}}^{t} \mathbf{m}_{\overleftarrow{e}}^{t-1} , \tag{16}$$

$$\mathbf{m}_{\overrightarrow{e}}^{t} = f_{\overrightarrow{e}}^{t} \left(\left(\mathbf{x}_{\overrightarrow{e}'}^{t} \right)_{\overrightarrow{e}' : \overrightarrow{e}' \to \overrightarrow{e}} \right) , \tag{17}$$

where $b_{\overrightarrow{e}}^{t}$ is the so-called *Onsager term*

$$b_{\overrightarrow{e}}^{t} = \frac{1}{N} \operatorname{Tr} \frac{\partial f_{\overrightarrow{e}}^{t}}{\partial \mathbf{x}_{\overleftarrow{e}}} \left(\left(\mathbf{x}_{\overrightarrow{e}'}^{t} \right)_{\overrightarrow{e}' : \overrightarrow{e}' \to \overrightarrow{e}} \right) \qquad \in \mathbb{R} . \tag{18}$$

Graph-based AMP [C. Gerbelot, R. Berthier '21]

The graph-based AMP iteration

$$\mathbf{x}_{\vec{e}}^{t+1} = \boldsymbol{A}_{\vec{e}}\mathbf{m}_{\vec{e}}^{t} - b_{\vec{e}}^{t}\mathbf{m}_{\overleftarrow{e}}^{t-1}\,, \tag{16}$$

$$\mathbf{m}_{\vec{e}}^{t} = f_{\vec{e}}^{t}\left(\left(\mathbf{x}_{\vec{e}'}^{t},\right)_{\vec{e}':\vec{e}'\to\vec{e}}\right)\,, \tag{17}$$

where $b_{\vec{e}}^{t}$ is the so-called *Onsager term*

$$b_{\vec{e}}^{t} = \frac{1}{N}\,\mathrm{Tr}\,\frac{\partial f_{\vec{e}}^{t}}{\partial \mathbf{x}_{\overleftarrow{e}}}\left(\left(\mathbf{x}_{\vec{e}'}^{t},\right)_{\vec{e}':\vec{e}'\to\vec{e}}\right) \qquad \in \mathbb{R}\,. \tag{18}$$

Theorem (informal):

any Graph-based AMP iterations admits rigorous SE equations

Graph-based AMP: idea of the proof

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

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

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

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

and prove its SE equations.

Applications of AMP iterations

- SE+converging trajectory $\rightarrow$ properties of estimators
- design of the iteration is key (not discussed here)
- sample from non-convex priors/posteriors, multilayer, etc ...
- study the dynamics of learning algorithms, notably SGD [M.Celentano, C. Cheng, A. Montanari '21]

Thank you

Collaborators : Bruno Loureiro, Gabriele Sicuro, Raphaël Berthier, Lenka Zdeborova and Florent Krzakala