Beyond i.i.d. Gaussian Models : Exact Asymptotics with Realistic Data

Cedric Gerbelot

ENS, Paris, France

June 18, 2021

Stat. Phys./High-Dimensional Approach

- typical case
- benchmark, random design problems
- exact solutions
- strong assumptions

Stat. Phys./High-Dimensional Approach

- typical case
- benchmark, random design problems
- exact solutions
- strong assumptions

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*"

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

- 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*"

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

- 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]

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

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

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

Solution to Block Covariance model

Theorem (informal)[B. Loureiro, CG, 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 :

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

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

Proof uses **convex Gaussian comparison inequalities**

[M. Stojnic, '13][C. Thrampoulidis, E. Abbasi, B. Hassibi '18]

How well does is work ?

Ridge regression works well ...

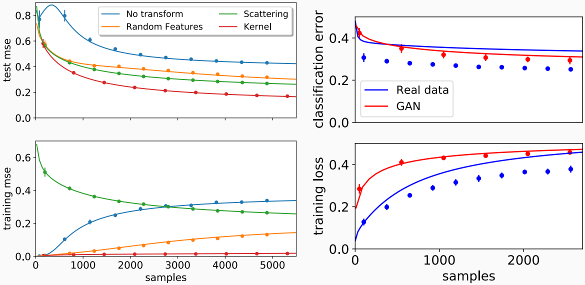

Figure 1: (Left) Ridge regression on real data. (Right) Logistic regression with real and synthetic (GAN) data

... but classification is more problematic

Ridge regression works well ...

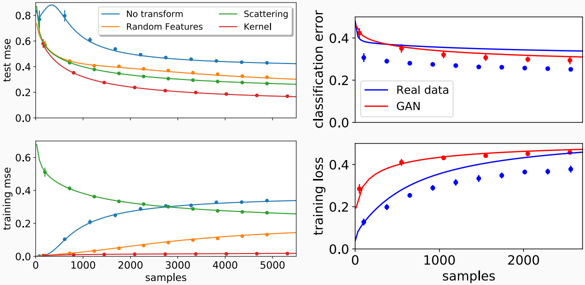

Figure 1: (Left) Ridge regression on real data. (Right) Logistic regression with real and synthetic (GAN) data

... 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

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

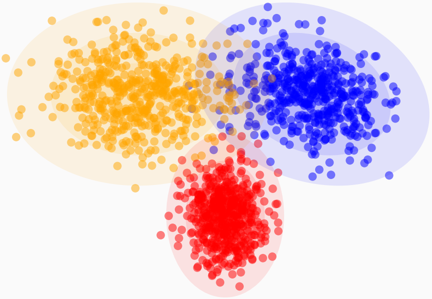

Figure 2: K=3, d=2

Classifying Gaussian Mixtures with Convex GLM

Data and teacher

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

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

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

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

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

Main result (informal)

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

Fixed-point of self-consistent equations

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

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

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

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 : 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

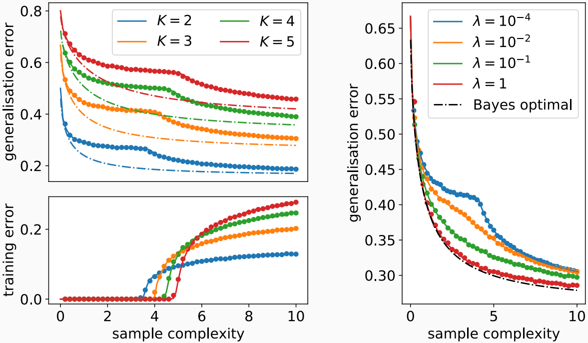

Figure 3: Ridge penalized logistic regression on K Gaussian clusters, $\Sigma_k = \Delta Id$. (Left) Sample complexity (Right) Regularization

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

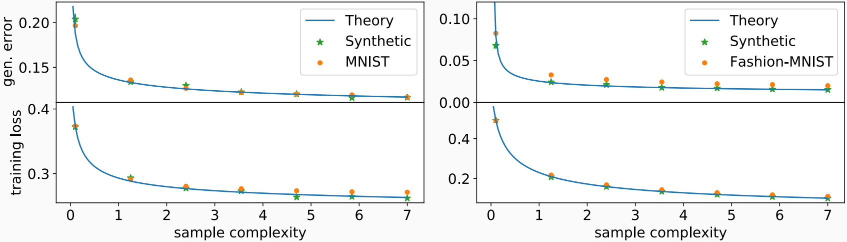

Figure 4: Binary classification on Mnist/Fashion-Mnist, odd vs even, Gaussian approximation and real data

Examples : real data

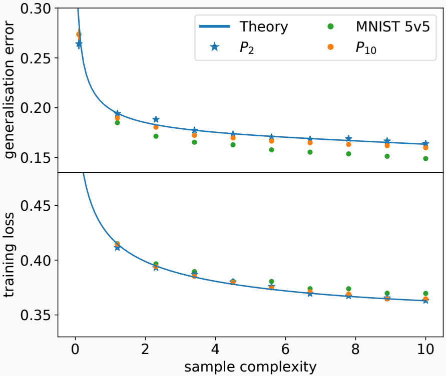

Figure 5: Adding more clusters to the Gaussian approximation

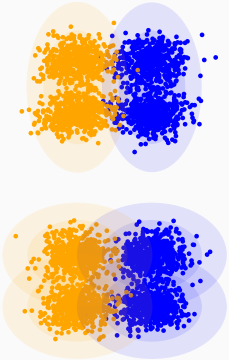

Figure 6: Idealized view

Sketch of proof

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

[C. Thrampoulidis, S. Oymak, M. Soltanolkotabi '20] (identity covariances)

Enter Approximate Message Passing (AMP)

Enter Approximate Message Passing (AMP)

Family of iterations with closed form exact asymptotics : **state evolution (SE) equations**

- enables matrix valued variables
- handles block correlation structures (spatial coupling)
- very adaptable !

Enter Approximate Message Passing (AMP)

Family of iterations with closed form exact asymptotics : **state evolution (SE) equations**

- enables matrix valued variables
- handles block correlation structures (spatial coupling)
- very adaptable !

First proof of SE equations due to E. Bolthausen (2009, math. phys.)
Then [M. Bayati, A. Montanari, '11]

Enter Approximate Message Passing (AMP)

What does an AMP look like ?

Enter Approximate Message Passing (AMP)

What does an AMP look like ?

Sequence of matrices $\boldsymbol{u}, \boldsymbol{v}$:

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

where $\boldsymbol{Z}$ (block-)Gaussian, $\boldsymbol{h}_t, \boldsymbol{e}_t$ are matrix valued functions.

Enter Approximate Message Passing (AMP)

What does an AMP look like ?

Sequence of matrices $\boldsymbol{u}, \boldsymbol{v}$:

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

where $\boldsymbol{Z}$ (block-)Gaussian, $\boldsymbol{h}_t, \boldsymbol{e}_t$ are matrix valued functions.

Brackets are Jacobian-like terms $\to$ **inherent to 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)

AMP for high-dim. stat : [M. Bayati, A. Montanari '11] [D. Donoho, A. Montanari '16]

Designing the AMP : a quick look

Often designed from a factor graph, see e.g. [L. Zdeborova, F. Krzakala '16]

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

Designing the AMP : a quick look

Often designed from a factor graph, see e.g. [L. Zdeborova, F. Krzakala '16]

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, see e.g. [L. Zdeborova, F. Krzakala '16]

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

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

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

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, matrix AMP : [A. Javanmard, A. Montanari '12]
Non-separable AMP : [R. Berthier, A. Montanari, P. Nguyen '18]

Validity of state evolution

Non-separable, block structure gradient

$$Z^\top \begin{bmatrix} \partial\tilde{L}_1(Z_1\tilde{W}_1) & & & \\ & \partial\tilde{L}_2(Z_2\tilde{W}_2) & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{L}_K(Z_K\tilde{W}_K) \end{bmatrix} + \begin{bmatrix} \partial\tilde{r}(\tilde{W})_1 & & & \\ & \partial\tilde{r}(\tilde{W})_2 & (0) & \\ & (0) & \ddots & \\ & & & \partial\tilde{r}(\tilde{W})_K \end{bmatrix}$$

Spatially-coupled, matrix AMP : [A. Javanmard, A. Montanari '12]
Non-separable AMP : [R. Berthier, A. Montanari, P. Nguyen '18]

Combination included in [**CG**, R. Berthier '21]

Future directions

Relevance to realistic scenarios

- Gaussian models are relevant to a certain degree
- Gaussian density estimators are universal ...
- ... becomes more complicated than original problem !
- middle ground/parametrization relevant for given tasks ?

Future directions

Relevance to realistic scenarios

- Gaussian models are relevant to a certain degree
- Gaussian density estimators are universal ...
- ... becomes more complicated than original problem !
- middle ground/parametrization relevant for given tasks ?

Technical improvements

- more possibilities using AMP methods
- finite size analysis [C. Rush, R. Venkataramanan '18]
- universality properties [M. Bayati, M. Lelarge, A. Montanari '15]

Thank you

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