Learning Gaussian Mixtures with Generalised Linear Models: a brief look at the proof

BL, GS, **Cedric Gerbelot**, AP, FK and LZ

Laboratoire de Physique de l'Ecole normale supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, F-75005 Paris, France

Teacher Student Generalized linear model

Most supervised learning problems are formulated as

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

$$\text{where} \quad \boldsymbol{y} = \phi(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n$$

- L, r are a convex loss and penalty defining the *student*
- $\phi, \boldsymbol{w}_0 \in \mathbb{R}^d$ represent the *teacher*
- $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a random design matrix (e.g. Gaussian with covariance)

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

Teacher Student Generalized linear model

Most supervised learning problems are formulated as

$$\boldsymbol{w}^\star \in \min_{\boldsymbol{w}\in\mathbb{R}^d} L(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{w}) + r(\boldsymbol{w})$$
$$\text{where} \quad \boldsymbol{y} = \phi(\boldsymbol{X}\boldsymbol{w}_0) \in \mathbb{R}^n$$

- L, r are a convex loss and penalty defining the *student*
- $\phi, \boldsymbol{w}_0 \in \mathbb{R}^d$ represent the *teacher*
- $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is a random design matrix (e.g. Gaussian with covariance)

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

(And match the replica formula !)

What are the difficulties ?

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

Simplest case : ridge regression with i.i.d./correlated Gaussian data, closed-form solution $\longrightarrow$ **random matrix theory** [BLLT20, HMRT20]

Beyond ridge regression : no closed-form solutions. One popular method is **convex Gaussian comparison inequalities** (CGMT) [TAH18, LGC$^+$21]

What are the difficulties ?

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

Simplest case : ridge regression with i.i.d./correlated Gaussian data, closed-form solution $\longrightarrow$ **random matrix theory** [BLLT20, HMRT20]

Beyond ridge regression : no closed-form solutions. One popular method is **convex Gaussian comparison inequalities** (CGMT) [TAH18, LGC$^+$21]

Works well for vector estimator $\boldsymbol{w}^\star$, any convex GLM, various correlation structure in the data ...

So what's wrong ?

What are the difficulties ?

Here we are learning a matrix !

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

And the pair $(\boldsymbol{Y}, \boldsymbol{X})$ is taken from a Gaussian mixture

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \Sigma_k), \tag{1}$$

Harder to represent as a matrix (e.g. $\boldsymbol{X} = \boldsymbol{Z}\Sigma^{1/2}$ with i.i.d. $\boldsymbol{Z}$)

What are the difficulties ?

Here we are learning a matrix !

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

And the pair $(\boldsymbol{Y}, \boldsymbol{X})$ is taken from a Gaussian mixture

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{K} y_k \rho_k \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}_k, \Sigma_k), \qquad (2)$$

Harder to represent as a matrix (e.g. $\boldsymbol{X} = \boldsymbol{Z}\Sigma^{1/2}$ with i.i.d. $\boldsymbol{Z}$)

Convex Gaussian comparison inequalities break down [TOS20]

Enter Approximate Message Passing (AMP)

Family of iterations with closed form exact asymptotics : **state evolution equations** [BM11, JM13, BMN20, GB21]

- 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 equations** [BM11, JM13, BMN20, GB21]

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

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

where $\boldsymbol{Z}$ (block-)Gaussian, $\boldsymbol{h}_t, \boldsymbol{e}_t$ are matrix valued functions.
Brackets are Jacobian-like terms $\rightarrow$ **inherent to AMP**

Sketch of proof

Target :

$$\boldsymbol{W}^\star \in \min_{\boldsymbol{W}\in\mathbb{R}^{d\times K}} L(\boldsymbol{Y}, \boldsymbol{X}\boldsymbol{W}) + r(\boldsymbol{W}) \tag{3}$$

Tool :

$$\begin{aligned} \boldsymbol{u}^{t+1} &= \boldsymbol{Z}^\top \boldsymbol{h}_t(\boldsymbol{v}^t) - \boldsymbol{e}_t(\boldsymbol{u}^t)\langle \boldsymbol{h}_t' \rangle^\top \\ \boldsymbol{v}^t &= \boldsymbol{Z}\boldsymbol{e}_t(\boldsymbol{u}^t) - \boldsymbol{h}_{t-1}(\boldsymbol{v}^{t-1})\langle \boldsymbol{e}_t' \rangle^\top \end{aligned} \tag{4}$$

Instructions:

- design $\boldsymbol{h}_t, \boldsymbol{e}_t$ s.t. fixed point of (4) matches opt. cond. of (3)
- find a converging trajectory (convexity helps)
- use state evolution equations (fixed point)

All done !

Fixed point of SE equations match replica saddle-point
(Simulations as well)

Thank you !

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler.
Benign overfitting in linear regression.
Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

Mohsen Bayati and Andrea Montanari.
The dynamics of message passing on dense graphs, with applications to compressed sensing.
IEEE Transactions on Information Theory, 57(2):764–785, 2011.

Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen.
State evolution for approximate message passing with non-separable functions.
Information and Inference: A Journal of the IMA, 9(1):33–79, 2020.

Cédric Gerbelot and Raphaël Berthier.
Graph-based approximate message passing iterations.
To appear, 2021.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani.
Surprises in high-dimensional ridgeless least squares interpolation.
Preprint arXiv:1903.08560, 2020.

Adel Javanmard and Andrea Montanari.
State evolution for general approximate message passing algorithms, with applications to spatial coupling.
Information and Inference: A Journal of the IMA, 2(2):115–144, 2013.

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová.
Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model.
Preprint arXiv:2102.08127, 2021.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi.

Precise error analysis of regularized m-estimators in high dimensions.
IEEE Transactions on Information Theory, 64(8):5592–5628, 2018.

Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi.
Theoretical insights into multiclass classification: A high-dimensional asymptotic view.
In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8907–8920. Curran Associates, Inc., 2020.