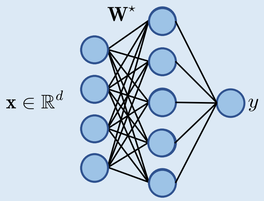
Bayes-optimal learning of an extensive-width neural network from quadratically many samples

Antoine Maillard[1], **Emanuele Troiani**[2], **Simon Martin**[3], **Florent Krzakala**[2], **Lenka Zdeborová**[2]
[1] ETH Zürich [2] EPFL [3] ENS Paris

Learning in large neural networks



with m hidden units

$$y_i = f_{\mathbf{W}^\star}(\mathbf{x}_i) \coloneqq \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^\star)^T \cdot \mathbf{x}_i\right]^2$$

$\sim \mathcal{N}(0, \mathrm{I}_d)$ $\quad$ $\mathbf{w}_k^\star \sim \mathcal{N}(0, \mathrm{I}_d)$

Learning from data

$\{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\} \Rightarrow \mathbf{W}^\star$?

High-dimensional limit
$d \to \infty\,;\ m = \Theta(d)$

Optimal generalization error

$$\mathcal{E}_{\text{gen.}} \coloneqq \mathbb{E}_{\mathbf{W}^\star,\{\mathbf{x}_i\}} \min_{\hat{y}(\{y_i,\mathbf{x}_i\})} \mathbb{E}_{\mathbf{x}_{\text{test}}}[(\hat{y}(\mathbf{x}_{\text{test}}) - f_{\mathbf{W}^\star}(\mathbf{x}_{\text{test}}))^2]$$

- The scaling $m = \Theta(d)$ is an interesting one to study wide networks:
  Scaling $m = \mathcal{O}(1)$: multi-index phase retrieval [5]
  Scaling $m \gg d$: linear regression reaches the optimal error
- With $n = \mathcal{O}(d)$ samples, no information on $\mathbf{W}^\star$ can be retrieved [1]
- But there are $\Theta(d^2)$ weights to learn
  Information theory hints at the scaling $n = \Theta(d^2)$

Optimal generalization error

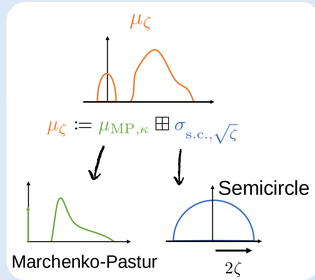
$d \to \infty \qquad m = \kappa d \qquad n = \alpha d^2$

$$\left\{y_i = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T \cdot \mathbf{x}_i + \sqrt{\Delta}\xi_k\right]^2\right\}_{i=1}^{n} \Rightarrow \lim_{d\to\infty} \mathcal{E}_{\text{gen.}} = 2\kappa\alpha\zeta - \Delta(2+\Delta)$$

ζ solves the self-consistent equation

$$(1-2\alpha) + \frac{\Delta(2+\Delta)}{\kappa\zeta} = \frac{4\pi^2\zeta}{3}\int \mu_\zeta(y)^3\,\mathrm{d}y$$

➢ **Easy-to-evaluate formula** for the Optimal generalization error



Open questions

❖ Analyze **other activations** (beyond quadratic)
Related to extensive rank tensor denoising

❖ Account for the **structure** of the data

❖ **Theoretical** analysis of GD properties

Sketch of the analysis

1 A planted linear matrix model

$$y = \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k^*)^T \cdot \mathbf{x}\right]^2 = \frac{1}{d}\mathbf{x}^\intercal\mathbf{S}^\star\mathbf{x} = \mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}]$$

$$\mathbf{S}^\star \coloneqq \frac{1}{m}\sum_{k=1}^{m}\mathbf{w}_k^\star(\mathbf{w}_k^\star)^\intercal \sim \mathcal{W}_{m,d} \qquad \mathbf{\Phi} \coloneqq \frac{1}{d}\mathbf{x}\mathbf{x}^\intercal$$

Wishart prior

Can be generalized to **noisy pre-activations**

$$\mathbf{w}_k^\star \cdot \mathbf{x} \to \mathbf{w}_k^\star \cdot \mathbf{x} + \sqrt{\Delta}\xi_k \Rightarrow y \sim P_{\text{out}}(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}])$$

Universality of optimal generalization error [3]

$$\min \mathcal{E}_{\text{gen.}}(\hat{\mathbf{w}}_k) = \min \|\hat{\mathbf{S}} - \mathbf{S}^\star\|_F^2 \quad \text{from } \{y_i \sim P_{\text{out}}(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{\Phi}_i])\}_{i=1}^n \simeq \min \|\hat{\mathbf{S}} - \mathbf{S}^\star\|_F^2 \quad \text{from } \{\tilde{y}_i \sim P_{\text{out}}(\cdot|\mathrm{Tr}[\mathbf{S}^\star\mathbf{G}_i])\}_{i=1}^n$$

Gaussian matrix

Just a (generalized) **linear model on** $\mathbf{S}^\star$

2 Mapping to GLM: scalar estimation part

➢ Gaussian data $\mathbf{G} \coloneqq \begin{pmatrix}\mathrm{flatt}(\mathbf{G}_1)\\ \vdots \\ \mathrm{flatt}(\mathbf{G}_n)\end{pmatrix}$ ✚ ➢ Wishart prior $\mathbf{S}^\star \sim \mathcal{W}_{m,d}$

Formula for $\widetilde{\mathcal{E}}_{\text{gen.}}$ (generalization of [4]) It involves
- Scalar estimation problem involving P_{out}
- Matrix denoising on the Wishart prior

3 Mapping to GLM: matrix denoising part

Denoising problem $\quad \mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \mathbf{S}^\star$?

Gaussian (GOE) matrix

❑ The optimal estimator is **spectral** [3]: $\mathbf{Y} = \mathbf{O}\mathbf{D}\mathbf{O}^\intercal \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = \mathbf{O}f_{\text{opt.}}(\mathbf{D})\mathbf{O}^\intercal$

❑ Analytical expressions for $f_{\text{opt.}}$ and the **asymptotic MMSE** $\lim_{d\to\infty} \|\hat{\mathbf{S}}(\mathbf{Y}) - \mathbf{S}^\star\|_F^2$

Optimal algorithm: GAMP-RIE

A provably optimal for large d, easy-to-implement polynomial-time algorithm — **GAMP** / **RIE**

Generalized linear model w. Gaussian data with non-separable prior ⇒ **Generalized Approximate Message Passing (GAMP)** [2]

Each GAMP iteration solves $\mathbf{Y} = \sqrt{\lambda}\mathbf{S}^\star + \mathbf{Z} \longrightarrow \mathbf{S}^\star$? ⇒ **Rotationally-Invariant Estimator (RIE)** [3] $\mathbf{Y} = \mathbf{O}\mathbf{D}\mathbf{O}^\intercal \Rightarrow \hat{\mathbf{S}}(\mathbf{Y}) = f_{\text{RIE}}(\mathbf{Y}) = \mathbf{O}f_{\text{opt.}}(\mathbf{D})\mathbf{O}^\intercal$

$$\omega_i^t = \frac{\mathbf{x}_i^\intercal\hat{\mathbf{S}}\mathbf{x}_i}{d} - g_{\text{out}}(y_i, \omega_i^{t-1}, V^{t-1})V^t$$
$$A^t = \frac{2\alpha}{n}\sum_{i=1}^{n} g_{\text{out}}(y_i, \omega_i^t, V^t)^2$$
$$\mathbf{R}^t = \hat{\mathbf{S}} + \frac{1}{d^{3/2}A^t}\sum_{i=1}^{n} g_{\text{out}}(y_i, \omega_i^t, V^t)(\mathbf{x}_i\mathbf{x}_i^\intercal - \mathbf{1})$$
$$\hat{\mathbf{S}}^{t+1} = f_{\text{RIE}}\left(\mathbf{R}^t, \frac{1}{2A^t}\right) \qquad V^{t+1} = 2F_{\text{RIE}}\left(\frac{1}{2A^t}\right)$$

Generalization error curves

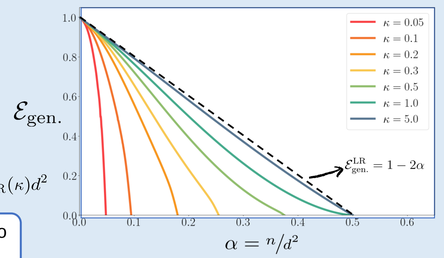
Noiseless setting : $\Delta = 0$

Perfect recovery threshold

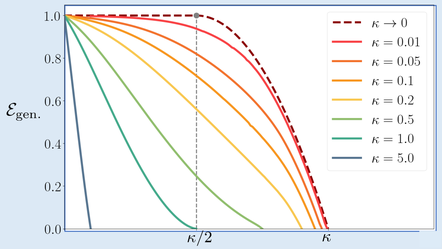
$$\alpha_{\text{PR}}(\kappa) = \min\left(\kappa - \frac{\kappa^2}{2}, \frac{1}{2}\right)$$

Matches a naïve "counting argument"

$$\mathrm{DOF}[\{\mathbf{S} : \mathbf{S} = \mathbf{S}^\intercal \text{ and } \mathrm{rk}(\mathbf{S}) \le \kappa d\}] \simeq \alpha_{\text{PR}}(\kappa)d^2$$



No computational to statistical gap



Significantly lower generalization error than linear regression ($\kappa \to \infty$):

$$\mathcal{E}_{\text{gen.}}^{\text{LR}} = 1 - 2\alpha$$

Smooth transition from hidden layer size $m = \mathcal{O}(1)$ to $m = \Theta(d)$ [5]

Gradient descent performance

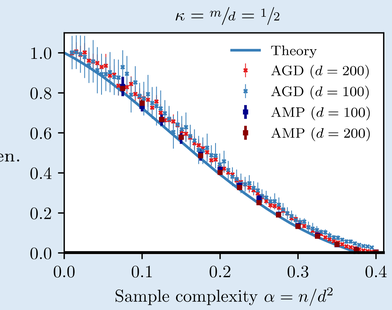
Noiseless setting : $\Delta = 0$

$$\mathcal{L}(\mathbf{W}) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{f}_{\mathbf{W}}(\mathbf{x}_i)\right)^2, \text{ where } \tilde{f}_{\mathbf{W}}(\mathbf{x}) \coloneqq \frac{1}{m}\sum_{k=1}^{m}\left[\frac{1}{\sqrt{d}}(\mathbf{w}_k)^T \cdot \mathbf{x}\right]^2$$

➢ For $\kappa \ge 1$ $(m \ge d)$, the problem is **convex** over $\mathbf{S} \coloneqq (1/m)\sum_{k=1}^{m}\mathbf{w}_k\mathbf{w}_k^\intercal$

➢ For $\kappa < 1$, **non-convex problem**. Still, naïve GD reaches optimal error !

For **any** κ, (averaged) GD seems to reach the optimal MMSE



References

[1] H. Cui, F. Krzakala, L. Zdeborová (2013). "Bayes-optimal learning of deep random networks of extensive-width" ICML 2023 [Arxiv 2302.00375]

[2] D. L. Donoho, A. Maleki, A. Montanari (2009). "Message-passing algorithms for compressed sensing" Proceeding of the National Academy of Sciences [Arxiv 0907.3574]

[3] J. Bun, R. Allez, J. P. Bouchaud, M. Potters (2016). "Rotational invariant estimators for general noisy matrices" IEEE Transactions on Information Theory [Arxiv 1502.06736]

[4] A. Maillard, A. Bandeira (2023). "Exact threshold for approximate ellipsoid fitting of random points" [Arxiv 2310.05787]

[5] E. Troiani, Y. Dandi, L. Defilippis, L. Zdeborová, B. Loureiro, F. Krzakala (2024). "Fundamental computational limits of weak learnability in high-dimensional multi-index models" [Arxiv 2405.15480]