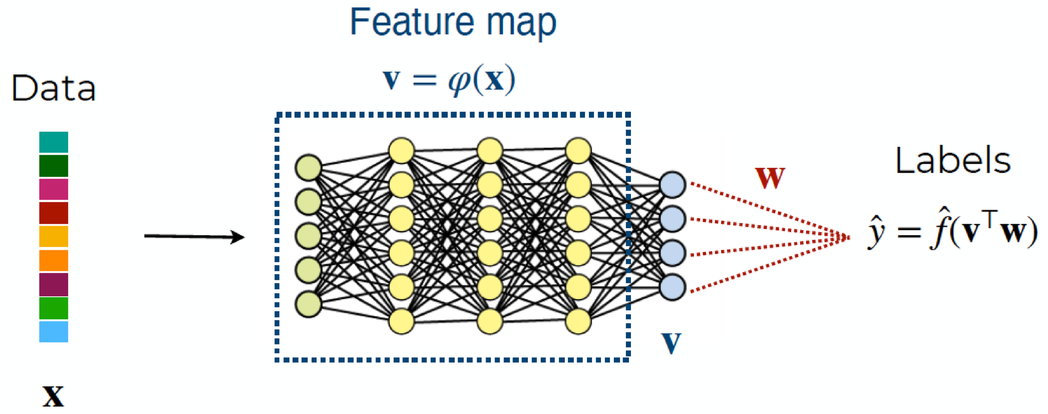
Learning curves of generic features maps for realistic datasets with a teacher-student model

Bruno Loureiro[1], **Cédric Gerbelot**[2], Hugo Cui[3], Sebastian Goldt[4], Florent Krzakala[1], Marc Mézard[2], Lenka Zdeborová[3]

[1] IdePHICS, EPFL; [2] **ENS Paris**; [3] SPOC, EPFL; [4] SISSA

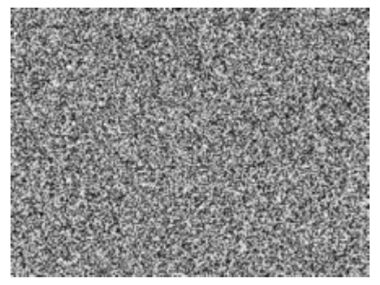
Motivation

- Choosing the right **features** in a Machine Learning task is an important step to achieve good generalisation performance. Indeed, neural networks can be seen as good feature learners, with the last layer performing a generalised linear task.



- Yet, the traditional statistical analysis based on uniform convergence bounds is agnostic to **data structure**.

Worst-case	Typical-case
Agnostic to structure	Model for data
Often too broad	Often too simplistic



Aim: *building better models for structured features*

Gaussian covariate model

We introduce a teacher-student **Gaussian covariate model** (GCM) for studying structured features. Consider a set of jointly Gaussian feature vectors:

$$\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix} \in \mathbb{R}^{p+d} \sim \mathcal{N}\left(0, \begin{bmatrix}\Psi & \Phi\\ \Phi^\top & \Omega\end{bmatrix}\right). \quad (1)$$

Labels are generated from the teacher features $\mathbf{u} \in \mathbb{R}^p$:

$$y^\mu = f_0\left(\frac{1}{\sqrt{p}}\boldsymbol{\theta}_0^\top \mathbf{u}^\mu\right), \quad (2)$$

Where $f_0 : \mathbb{R} \to \mathbb{R}$ is a (potentially random) scalar function and $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ are fixed weights. The goal is to characterise the learning performance of a student model:

$$\hat{y}(\mathbf{v}) = \hat{f}\left(\frac{1}{\sqrt{d}}\mathbf{v}^\top \hat{\mathbf{w}}\right) \quad (3)$$

obtained through empirical risk minimisation:

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^d}{\text{argmin}}\left[\sum_{\mu=1}^n g\left(\frac{\mathbf{w}^\top\mathbf{v}^\mu}{\sqrt{d}}, y^\mu\right) + \frac{\lambda}{2}||\mathbf{w}||_2^2\right], \quad (4)$$

where g is a convex loss function, $\lambda > 0$ the regularisation strength.

Goal

Characterise the generalisation and training performances of the ERM predictor $\hat{\mathbf{w}} \in \mathbb{R}^d$:

$$\mathcal{E}_{\text{gen.}}(\hat{\mathbf{w}}) = \mathbb{E}\left[\hat{g}\left(\hat{f}(\mathbf{v}_{\text{new}}^\top\hat{\mathbf{w}}), f_0(\mathbf{u}_{\text{new}}^\top\boldsymbol{\theta}_0)\right)\right] \quad (5)$$

$$\mathcal{E}_{\text{train.}}(\hat{\mathbf{w}}) = \frac{1}{n}\sum_{\mu=1}^n g\left(\hat{\mathbf{w}}^\top\mathbf{v}^\mu, y^\mu\right) \quad (6)$$

in the high-dimensional limit $n, p, d \to \infty$ with $\alpha \equiv n/d$ and $\gamma \equiv p/d$ fixed, where g is the loss and $\hat{g}$ is a performance measure.

Main technical result

Let $\Omega = \mathrm{S}^\top \mathrm{diag}(\omega_i)\mathrm{S}$ be the spectral decomposition of Ω. Let:

$$\rho \equiv \frac{1}{d}\boldsymbol{\theta}_0^\top\Psi\boldsymbol{\theta}_0 \in \mathbb{R}, \qquad \bar{\boldsymbol{\theta}} \equiv \frac{\mathrm{S}\Phi^\top\boldsymbol{\theta}_0}{\sqrt{\rho}} \in \mathbb{R}^d \quad (7)$$

and define the joint empirical density $\hat{\mu}_d$ between $(\omega_i, \bar{\theta}_i)$:

$$\hat{\mu}_d(\omega, \bar{\theta}) \equiv \frac{1}{d}\sum_{i=1}^d \delta(\omega - \omega_i)\delta(\bar{\theta} - \bar{\theta}_i). \quad (8)$$

We assume that in the high-dimensional limit the spectral distributions of the matrices Φ, Ψ and Ω converge to distributions such that the limiting joint distribution μ is well-defined, and their maximum singular values are bounded with high probability.

Closed-form asymptotics

Theorem 1. (informal) In the asymptotic limit, the training and generalisation errors (5) of the estimator $\hat{\mathbf{w}} \in \mathbb{R}^d$ solving the empirical risk minimisation problem in eq. (4) verify:

$$\mathcal{E}_{\text{train.}}(\hat{\mathbf{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[g\left(z(V^\star, m^\star, q^\star), f_0(\sqrt{\rho}s)\right)\right]$$
$$\mathcal{E}_{\text{gen.}}(\hat{\mathbf{w}}) \xrightarrow[d\to\infty]{P} \mathbb{E}_{(\nu,\lambda)}\left[\hat{g}\left(\hat{f}(\lambda), f_0(\nu)\right)\right] \quad (9)$$

where we have defined the scalar random function $z(V, m, q) = \mathrm{prox}_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q - \rho^{-1}m^2}h)$, with:

$$\mathrm{prox}_{Vg(.,y)}(x) = \underset{z\in\mathbb{R}}{\text{argmin}}\left\{g(z, y) + \frac{1}{2V}(x - z)^2\right\} \quad (10)$$

and where (ν, λ) are jointly Gaussian scalar variables:

$$(\nu, \lambda) \sim \mathcal{N}\left(0, \begin{bmatrix}\rho & m^\star\\ m^\star & q^\star\end{bmatrix}\right). \quad (11)$$

The overlap parameters $(V^\star, q^\star, m^\star)$ are prescribed by the unique fixed point of the following set of self-consistent equations:

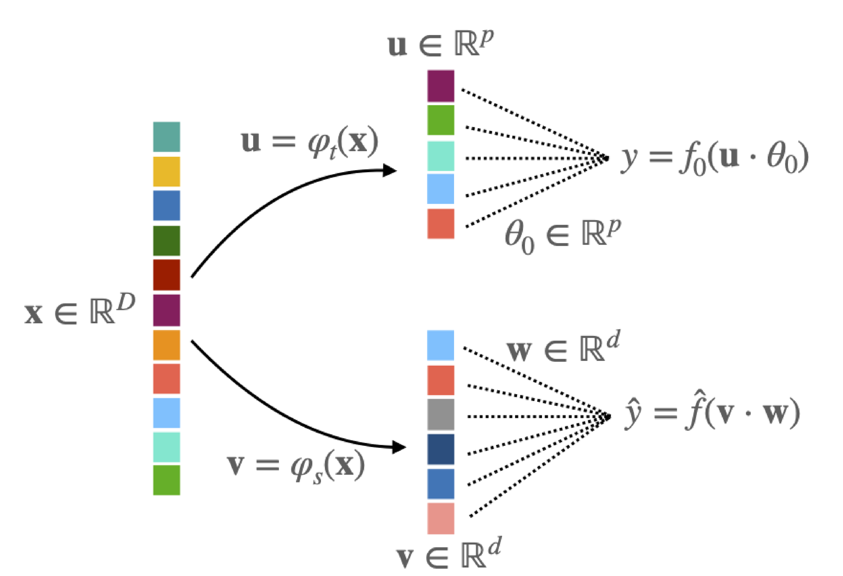
$$\begin{cases} V = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\omega}{\lambda+\hat{V}\omega}\right]\\ m = \frac{\hat{m}}{\sqrt{\gamma}}\mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\bar{\theta}^2}{\lambda+\hat{V}\omega}\right]\\ q = \mathbb{E}_{(\omega,\bar{\theta})\sim\mu}\left[\frac{\hat{m}^2\bar{\theta}^2\omega + \hat{q}\omega^2}{(\lambda+\hat{V}\omega)^2}\right]\end{cases}$$
$$\begin{cases} \hat{V} = \frac{\alpha}{V}(1 - \mathbb{E}_{s,h\sim\mathcal{N}(0,1)}[z'(V, m, q)])\\ \hat{m} = \frac{1}{\sqrt{\rho\gamma}}\frac{\alpha}{V}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[sz(V, m, q) - \frac{m}{\sqrt{\rho}}z'(V, m, q)\right]\\ \hat{q} = \frac{\alpha}{V^2}\mathbb{E}_{s,h\sim\mathcal{N}(0,1)}\left[\left(\frac{m}{\sqrt{\rho}}s + \sqrt{q - \frac{m^2}{\rho}}h - z(V, m, q)\right)^2\right]\end{cases} \quad (12)$$

and $z'(V, m, h) = \mathrm{prox}'_{Vg(.,f_0(\sqrt{\rho}s))}(\rho^{-1/2}ms + \sqrt{q - \rho^{-1}m^2}h)$ is the first derivative of the proximal operator.

Modelling realistic data

Let $\{\mathbf{x}^\mu\}_{\mu=1}^n$ denote n independent samples from a data set on $\mathcal{X}$ which we would like to learn. The idea is to use the GCM to capture the learning performance with the following non-linear features:

$$\mathbf{x} \mapsto \mathbf{u} = \boldsymbol{\varphi}_t(\mathbf{x}) \in \mathbb{R}^p, \qquad \mathbf{x} \mapsto \mathbf{v} = \boldsymbol{\varphi}_s(\mathbf{x}) \in \mathbb{R}^d \quad (13)$$



In general $[\mathbf{u}, \mathbf{v}]$ stemming from non-linear feature maps are not jointly Gaussian, but in the high-dimensional limit we observe the generalisation and training error often depend only on the second order statistics.

Conjecture: Gaussian equivalence [2, 3]

For a wide class of data distributions $\{\mathbf{x}^\mu\}_{\mu=1}^n$, and features maps $\boldsymbol{u} = \boldsymbol{\varphi}_t(\boldsymbol{x})$, $\boldsymbol{v} = \boldsymbol{\varphi}_s(\boldsymbol{x})$, the generalisation and training errors of estimator (4) are asymptotically captured by the equivalent Gaussian model (1), where $[\mathbf{u}, \mathbf{v}]$ are jointly Gaussian variables, and thus by the closed-form expressions of Theorem 1:

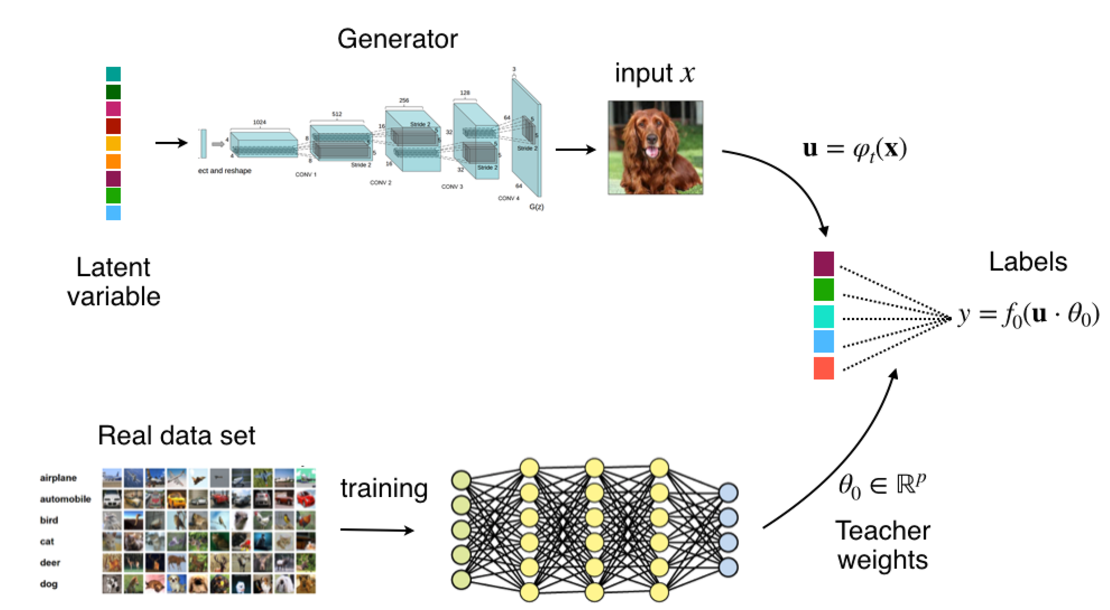
$$\mathcal{E}_{\text{gen.}\backslash\text{train.}}(\nu, \beta) \underset{n,p,d\to\infty}{\asymp} \mathcal{E}_{\text{gen.}\backslash\text{train.}}(\nu_2, \beta_2) \quad (14)$$

where $\nu = \boldsymbol{\theta}_0^\top\boldsymbol{u}$, $\beta = \hat{\boldsymbol{w}}^\top\boldsymbol{v}$ and (ν_2, β_2) are their Gaussian equivalent obtained by matching the first moments.

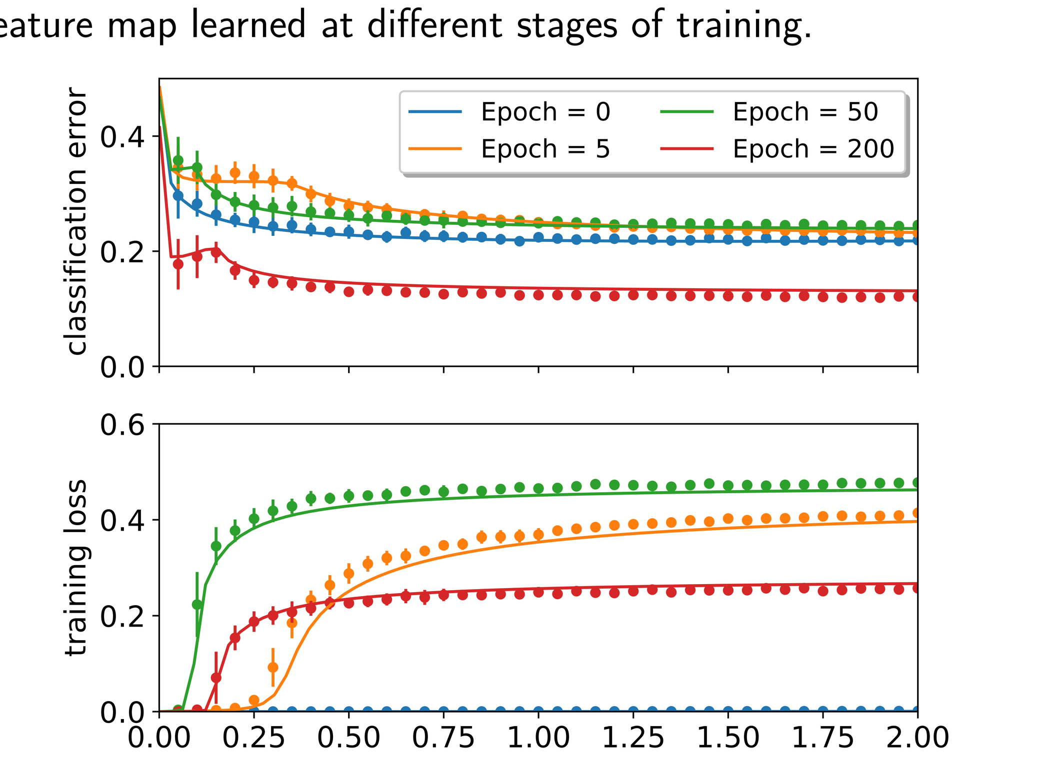
Note: the Gaussian equivalence has been rigorously proven in the random features case $\mathbf{u} \sim \mathcal{N}(0, \mathrm{I}_p)$ $\mathbf{v} = \sigma(\mathbf{F}\mathbf{u})$ in [3, 4].

Adversarial Generative Network (GAN) data

In this section, the input $\mathbf{x} = \mathcal{G}(\mathbf{z})$ is drawn from a GAN trained on a data set of interest. Labels are generated from a teacher model trained on the real data set. This generative process allow us to sample and estimate the covariances required in Theorem 1.



As an example, we have trained a dcGAN to generate CIFAR10-like images, and have trained a fully-connected two-layer teacher network to assign labels for a binary animal vs. not animal classification task on CIFAR10. The student features were obtained by training a fully connected three-layer neural network on 30k samples from the generative data set with the square loss. Logistic regression is then performed on the features with vanishing $\lambda \to 0^+$, and the performance is shown for the feature map learned at different stages of training.



Real data

Next, we investigate whether our model can capture the learning curves of real data sets $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^{n_{\text{tot}}}$. Indeed, if the teacher weights $\boldsymbol{\theta}_0$ and features $\mathbf{u}^\mu = \boldsymbol{\varphi}_t(\mathbf{x}^\mu)$ generating the labels were known, we could estimate the covariances for the student features of interest $\mathbf{v} = \boldsymbol{\varphi}_s(\mathbf{x})$ empirically:

$$\Psi = \sum_{\mu=1}^{n_{\text{tot}}}\frac{\mathbf{u}^\mu\mathbf{u}^{\mu\top}}{n_{\text{tot}}}, \qquad \Phi = \sum_{\mu=1}^{n_{\text{tot}}}\frac{\mathbf{u}^\mu\mathbf{v}^{\mu\top}}{n_{\text{tot}}}, \qquad \Omega = \sum_{\mu=1}^{n_{\text{tot}}}\frac{\mathbf{v}^\mu\mathbf{v}^{\mu\top}}{n_{\text{tot}}}. \quad (15)$$

In principle, many teachers $(\boldsymbol{\theta}_0, \mathbf{u})$ interpolate the data, and it is not clear how to choose one. With one exception: linear teachers $y^\mu = \boldsymbol{\theta}_0^\top\mathbf{u}^\mu$.

Theorem 2. For any teacher feature map $\boldsymbol{\varphi}_t$, and for any $\boldsymbol{\theta}_0$ that interpolates the data so that $y^\mu = \boldsymbol{\theta}_0^\top\mathbf{u}^\mu$ $\forall\mu$, the asymptotic predictions of model (1) are equivalent. Indeed, the teacher only appear through quantities which can be directly expressed in terms of the labels:

$$\rho = \frac{1}{n_{\text{rot}}}\sum_{\mu=1}^{n_{\text{rot}}}(y^\mu)^2, \qquad \Phi^\top\boldsymbol{\theta}_0 = \frac{1}{n_{\text{rot}}}\sum_{\mu=1}^{n_{\text{rot}}}y^\mu\mathbf{v}^\mu. \quad (16)$$

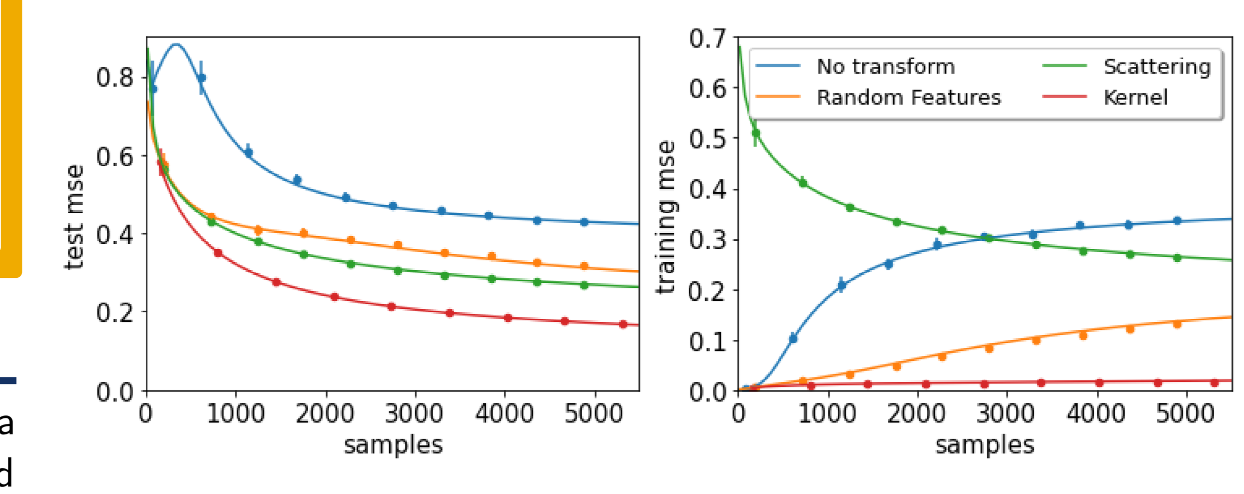


Figure 1: Test and training mean-squared errors as a function of the number of samples n for ridge regression on the MNIST data set, with regularisation $\lambda = 10^{-2}$. We show the performance with no feature map (blue), random feature map with $\sigma = \mathrm{erf}$ & Gaussian projection (orange), the scattering transform with parameters $J = 3, L = 8$ (green), and of the limiting kernel of the random map (red).

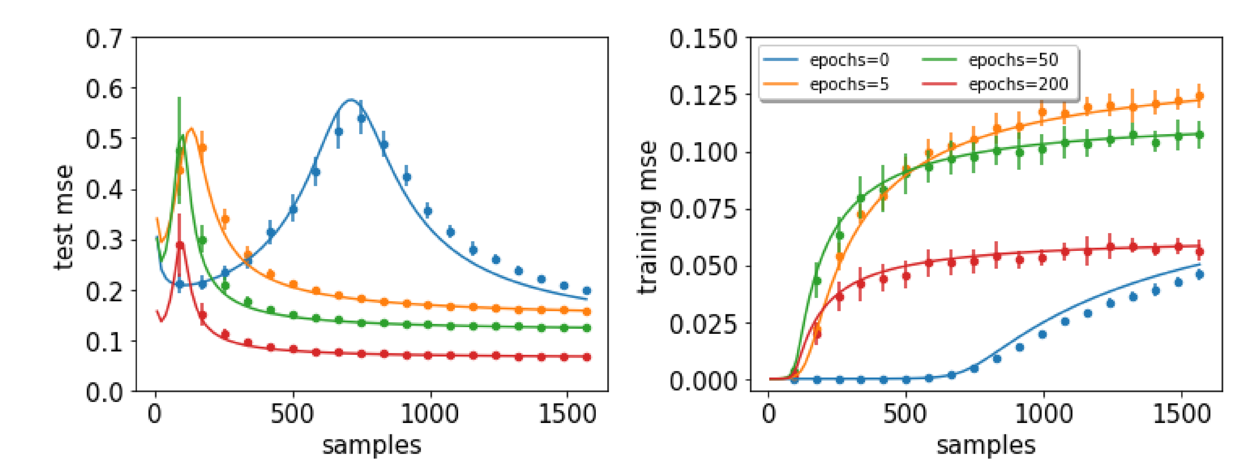


Figure 2: Test and training mean-squared errors as a function of the number of samples n for ridge regression on the Fashion-MNIST data set, with vanishing regularisation $\lambda = 10^{-5}$. In this plot, the student feature map $\boldsymbol{\varphi}_s$ is a 3-layer fully-connected neural network with $d = 2352$ hidden neurons trained on the full data set with the square loss. Different curves correspond to the feature map obtained at different stages of training.

Conclusion

We have shown that the training and generalisation performances of generalised linear models on a broad class of realistic features can be captured by a Gaussian covariate model, for which we provide rigorous and exact characerisation in the high-dimensional limit.

Perspectives:

- Applications to practical Machine Learning problems.
- Universality beyond linear teachers?

References

[1] ***Learning curves of generic features maps for realistic datasets with a teacher-student model***, B Loureiro, C Gerbelot, H Cui, S Goldt, F Krzakala, M Mézard, L Zdeborová, arXiv: 2102.08127 [stat.ML]

[2] *Modelling the influence of data structure on learning in neural networks: the hidden manifold model*, S Goldt, Mézard, F Krzakala, L Zdeborová, Physical Review X, Vol. 10, No. 4 (2020)

[3] *The Gaussian equivalence of generative models for learning with two-layer neural networks*, S Goldt, B Loureiro, G Reeves, M Mézard, F Krzakala, L Zdeborová, MSML 2021

[4] *Universality Laws for High-Dimensional Learning with Random Features*, H Hu, YM Lu, arXiv: 2009.07669 [cs.IT]

Contact: cedric.gerbelot@ens.fr