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

Bruno Loureiro
IdePHICS, EPFL, Lausanne

Cédric Gerbelot
Lab. de Physique de l’École Normale Supérieure, Paris

Hugo Cui
SPOC, EPFL, Lausanne

Sebastian Goldt
SISSA, Trieste

Florent Krzakala
IdePHICS, EPFL, Lausanne

Marc Mézard
École Normale Supérieure, Paris

Lenka Zdeborová
SPOC, EPFL, Lausanne

Abstract

Teacher-student models provide a framework in which the typical-case performance of high-dimensional supervised learning can be described in closed form. The assumptions of Gaussian i.i.d. input data underlying the canonical teacher-student model may, however, be perceived as too restrictive to capture the behaviour of realistic data sets. In this paper, we introduce a Gaussian covariate generalisation of the model where the teacher and student can act on different spaces, generated with fixed, but generic feature maps. While still solvable in a closed form, this generalization is able to capture the learning curves for a broad range of realistic data sets, thus redeeming the potential of the teacher-student framework. Our contribution is then two-fold: first, we prove a rigorous formula for the asymptotic training loss and generalisation error. Second, we present a number of situations where the learning curve of the model captures the one of a realistic data set learned with kernel regression and classification, with out-of-the-box feature maps such as random projections or scattering transforms, or with pre-learned ones - such as the features learned by training multi-layer neural networks. We discuss both the power and the limitations of the framework.

1 Introduction

Teacher-student models are a popular framework to study the high-dimensional asymptotic performance of learning problems with synthetic data, and have been the subject of intense investigations spanning three decades [1–7]. In the wake of understanding the limitations of classical statistical learning approaches [8–10], this direction is witnessing a renewal of interest [10–15]. However, this framework is often assuming the input data to be Gaussian i.i.d., which is arguably too simplistic to be able to capture properties of realistic data. In this paper, we redeem this line of work by defining a Gaussian covariate model where the teacher and student act on different Gaussian correlated spaces with arbitrary covariance. We derive a rigorous asymptotic solution of this model generalizing the formulas found in the above mentioned classical works.

We then put forward a theory, supported by universality arguments and numerical experiments, that this model captures learning curves, i.e. the dependence of the training and test errors on the number of samples, for a generic class of feature maps applied to realistic datasets. These maps can be deterministic, random, or even learnt from the data. This analysis thus gives a unified framework to describe the learning curves of, for example, kernel regression and classification, the analysis of
feature maps – random projections [16], neural tangent kernels [17], scattering transforms [18] – as well as the analysis of transfer learning performance on data generated by generative adversarial networks [19]. We also discuss limits of applicability of our results, by showing concrete situations where the learning curves of the Gaussian covariate model differ from the actual ones.

Model definition — The Gaussian covariate teacher-student model is defined via two vectors \( u \in \mathbb{R}^p \) and \( v \in \mathbb{R}^d \), with correlation matrices \( \Psi \in \mathbb{R}^{p \times p} \), \( \Omega \in \mathbb{R}^{d \times d} \) and \( \Phi \in \mathbb{R}^{p \times d} \), from which we draw \( n \) independent samples:

\[
\begin{bmatrix}
    u^\mu \\
    v^\mu
\end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix}
    \Psi & \Phi \\
    \Phi^\top & \Omega
\end{bmatrix}\right), \quad \mu = 1, \ldots, n.
\]  

The labels \( y^\mu \) are generated by a teacher function that is only using the vectors \( u^\mu \):

\[
y^\mu = f_0 \left( \frac{1}{\sqrt{p}} \theta_0^\top u^\mu \right),
\]

where \( f_0 : \mathbb{R} \to \mathbb{R} \) is a function that may include randomness such as, for instance, an additive Gaussian noise, and \( \theta_0 \in \mathbb{R}^p \) is a vector of teacher-weights with finite norm which can be either random or deterministic. Learning is performed by the student with weights \( w \) via empirical risk minimization that has access only to the features \( v^\mu \):

\[
\hat{w} = \arg \min_{w \in \mathbb{R}^d} \left[ \frac{1}{n} \sum_{\mu=1}^n g \left( \frac{w^\top v^\mu}{\sqrt{d}}, y^\mu \right) + r(w) \right],
\]

where \( r \) and \( g \) are proper, convex, lower-semicontinuous functions of \( w \in \mathbb{R}^d \) (e.g. \( g \) can be a logistic or a square loss and \( r \) is \( \ell_p \) \((p=1, 2)\) regularization). The key quantities we want to compute in this model are the averaged training and generalisation errors for the estimator \( \hat{w} \),

\[
\mathcal{E}_{\text{train}}(w) = \frac{1}{n} \sum_{\mu=1}^n g \left( \frac{w^\top v^\mu}{\sqrt{d}}, y^\mu \right) \quad \text{and} \quad \mathcal{E}_{\text{gen}}(w) = \mathbb{E} \left[ \hat{g} \left( \frac{\hat{v}_{\text{new}}^\top w}{\sqrt{d}} \right), f_0 \left( \frac{u_{\text{new}}^\top \theta_0}{\sqrt{p}} \right) \right],
\]

where \( g \) is the loss function in eq. (3), \( \hat{f} \) is a prediction function (e.g. \( \hat{f} = \text{sign} \) for a classification task), \( \hat{g} \) is a performance measure (e.g. \( \hat{g}(\hat{y}, y) = (\hat{y} - y)^2 \) for regression or \( \hat{g}(\hat{y}, y) = \mathbb{P}(\hat{y} \neq y) \) for classification) and \((u_{\text{new}}, v_{\text{new}})\) is a fresh sample from the joint distribution of \( u \) and \( v \).

Our two main technical contributions are:

(C1) In Theorems 1 & 2, we give a rigorous closed-form characterisation of the properties of the estimator \( \hat{w} \) for the Gaussian covariate model (1), and the corresponding training and generalisation errors in the high-dimensional limit. We prove our result using Gaussian comparison inequalities [20];

(C2) We show how the same expression can be obtained using the replica method from statistical physics [21]. This is of additional interest given the wide range of applications of the replica approach in machine learning and computer science [22]. In particular, this allows to put on a rigorous basis many results previously derived with the replica method.

Towards realistic data — In the second part of our paper, we argue that the above Gaussian covariate model (1) is generic enough to capture the learning behaviour of a broad range of realistic data. Let \( \{x^\mu\}_{\mu=1}^n \) denote a data set with \( n \) independent samples on \( \mathcal{X} \subset \mathbb{R}^p \). Based on this input, the features \( u, v \) are given by (potentially) elaborated transformations of \( x \), i.e.

\[
u = \varphi_\ell(x) \in \mathbb{R}^p \quad \text{and} \quad v = \varphi_s(x) \in \mathbb{R}^d
\]

for given centred feature maps \( \varphi_\ell : \mathcal{X} \to \mathbb{R}^p \) and \( \varphi_s : \mathcal{X} \to \mathbb{R}^d \), see Fig. 1. Uncentered features can be taken into account by shifting the covariances, but we focus on the centred case to lighten notation.

The Gaussian covariate model (1) is exact in the case where \( x \) are Gaussian variables and the feature maps \( (\varphi_\ell, \varphi_s) \) preserve the Gaussianity, for example linear features. In particular, this is the case for \( u = v = x \), which is the widely-studied vanilla teacher-student model [24]. The interest of the model (1) is that it also captures a range of cases in which the feature maps \( \varphi_\ell \) and \( \varphi_s \) are deterministic, or even learnt from the data. The covariance matrices \( \Psi, \Phi, \Omega \) then represent different aspects of the data-generative process and learning model. The student (5) then corresponds to the last layer of the learning model. These observation can be distilled into the following conjecture:
Figure 1: **Left:** Given a data set \( \{ x_i^{(n)} \}_{i=1}^{n} \), teacher \( u = \varphi_t(x) \) and student maps \( v = \varphi_s(x) \), we assume \( [u, v] \) to be jointly Gaussian random variables and apply the results of the Gaussian covariate model \([1]\). **Right:** Illustration on real data, here ridge regression on even vs odd MNIST digits, with regularisation \( \lambda = 10^{-2} \). Full line is theory, points are simulations. We show the performance with no feature map (blue), random feature map with \( \sigma = \text{erf} \) & Gaussian projection (orange), the scattering transform with parameters \( J = 3, L = 8 \) \([18]\) (green), and of the limiting kernel of the random map \([23]\) (red). The covariance \( \Omega \) is empirically estimated from the full data set, while the other quantities appearing in the Theorem \([1]\) are expressed directly as a function of the labels, see Section \([3,4]\). Simulations are averaged over 10 independent runs.

**Conjecture 1.** *(Gaussian equivalent model)* For a wide class of data distributions \( \{ x_i^{(n)} \}_{i=1}^{n} \), and features maps \( u = \varphi_t(x), v = \varphi_s(x) \), the generalisation and training errors of estimator \([3]\), are asymptotically captured by the equivalent Gaussian model \([1]\), where \( [u, v] \) are jointly Gaussian variables, and thus by the closed-form expressions of Theorem \([1]\).

The second part of our main contributions are:

(C3) In Sec. \([3,3]\) we show that the theoretical predictions from (C1) captures the learning curves in non-trivial cases, e.g. when input data are generated using a trained generative adversarial network, while extracting both the feature maps from a neural network trained on real data.

(C4) In Sec. \([3,4]\) we show empirically that for ridge regression the asymptotic formula of Theorem \([1]\) can be applied directly to real data sets, even though the Gaussian hypothesis is not satisfied. This universality-like property is a consequence of Theorem \([3]\) and is illustrated in Fig. 1 (right) where the real learning curve of several features maps learning the odd-versus-even digit task on MNIST is compared to the theoretical prediction.

**Related work —** Rigorous results for teacher-student models: The Gaussian covariate model \([1]\) contains the vanilla teacher-student model as a special case where one takes \( u \) and \( v \) identical, with unique covariance matrix \( \Omega \). This special case has been extensively studied in the statistical physics community using the heuristic replica method \([1,3,24,25]\). Many recent rigorous results for such models can be rederived as a special case of our formula, e.g. refs. \([10,15,26,29]\). Numerous of these results are based on the same proof technique as we employed here: the Gordon’s Gaussian min-max inequalities \([20,30,31]\). The asymptotic analysis of kernel ridge regression \([32]\), of margin-based classification \([33]\) also follow from our theorem. See also Appendix \([A,6]\) for the details on these connections. Other examples include models of the double descent phenomenon \([34]\). Closer to our work is the recent work of \([35]\) on the random feature model. For ridge regression, there are also precise predictions thanks to random matrix theory \([12,36,41]\). A related set of results was obtained in \([42]\) for orthogonal random matrix models. The main technical novelty of our proof is the handling of a generic loss and regularisation, not only ridge, representing convex empirical risk minimization, for both classification and regression, with the generic correlation structure of the model \([1]\).

Gaussian equivalence: A similar Gaussian conjecture has been discussed in a series of recent works, and some authors proved partial results in this direction \([11,12,28,35,43,46]\). Ref. \([45]\) analyses a special case of the Gaussian model (corresponding to \( \varphi_t = \text{id} \) here), and proves a Gaussian equivalence theorem (GET) for feature maps \( \varphi_s \), given by single-layer neural networks with fixed weights. They also show that for Gaussian data \( x \sim N(0, I_D) \), feature maps of the form \( v = \sigma(Wx) \) (with some technical restriction on the weights) led to the jointly-Gaussian property for the two scalars \( \langle v, w, u, \theta_0 \rangle \) for almost any vector \( w \). However, their stringent assumptions on random teacher weights limited the scope of applications to unrealistic label models. A related line of work
discussed similar universality through the lens of random matrix theory \cite{47,49}. In particular, Seddik et al. \cite{50} showed that, in our notations, vectors \([u, v]\) obtained from Gaussian inputs \(x \sim N(0, 1_D)\) with Lipschitz feature maps satisfy a concentration property. In this case, again, one can expect the two scalars \((v \cdot w, u \cdot \theta_0)\) to be jointly Gaussian with high-probability on \(w\). Remarkably, in the case of random feature maps, \cite{46} could go beyond this central-limit-like behavior and established the universality of the Gaussian covariate model \cite{1} for the actual learned weights \(\hat{w}\).

## 2 Main technical results

Our main technical result is a closed-form expression for the asymptotic training and generalisation errors \cite{4} of the Gaussian covariate model introduced above. We start by presenting our result in the most relevant setting for the applications of interest in Section \ref{sec:applications} which is the case of the \(\ell_2\) regularization. Next, we briefly present our result in larger generality, which includes non-asymptotic results for non-separable losses and regularizations.

We start by defining key quantities that we will use to characterize the estimator \(\hat{w}\). Let \(\Omega = S^\top \text{diag}(\omega_\ell) S\) be the spectral decomposition of \(\Omega\). Let:

\[
\rho \equiv \frac{1}{d} \theta_0^\top \Psi \theta_0 \in \mathbb{R}, \quad \vartheta \equiv \frac{S \Phi \Psi \theta_0}{\sqrt{\rho}} \in \mathbb{R}^d
\]

and define the joint empirical density \(\hat{\mu}_d\) between \((\omega_i, \theta_i)\):

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

Note that \(\Phi \Psi \theta_0\) is the projection of the teacher weights on the student space, and therefore \(\vartheta\) is the rotated projection on the basis of the student covariance, rescaled by the teacher variance. Together with the student eigenvalues \(\omega_i\), these are relevant statistics of the model, encoded here in the joint distribution \(\hat{\mu}_d\).

**Assumptions** — Consider the high-dimensional limit in which the number of samples \(n\) and the dimensions \(p, d\) go to infinity with fixed ratios:

\[
\alpha \equiv \frac{n}{d}, \quad \text{and} \quad \gamma \equiv \frac{p}{d}.
\]

Assume that the covariance matrices \(\Psi, \Omega\) are positive-definite and that the Schur complement of the block covariance in equation \ref{eq:block} is positive semi-definite. Additionally, the spectral distributions of the matrices \(\Phi, \Psi\) and \(\Omega\) converge to distributions such that the limiting joint distribution \(\hat{\mu}\) is well-defined, and their maximum singular values are bounded with high probability as \(n, p, d \to \infty\). Finally, regularity assumptions are made on the loss and regularization functions mainly to ensure feasibility of the minimization problem. We assume that the cost function \(r + g\) is coercive, i.e. \(\lim_{\|w\| \to +\infty} (r + g)(w) = +\infty\) and that the following scaling condition holds: for all \(n, d \in \mathbb{N}, z \in \mathbb{R}^n\) and any constant \(c > 0\), there exist a finite, positive constant \(C\), such that, for any standard normal random vectors \(h \in \mathbb{R}^d\) and \(g \in \mathbb{R}^n\):

\[
\|z\|_2 \leq c\sqrt{n} \implies \sup_{x \in \partial g(x)} \|x\|_2 \leq C\sqrt{n}, \quad \frac{1}{d} \mathbb{E} [r(h)] < +\infty, \quad \frac{1}{n} \mathbb{E} [g(g)] < +\infty.
\]

The relevance of these assumptions in a supervised machine learning context is discussed in Appendix \ref{app:assumptions}. We are now in a position to state our result.

**Theorem 1.** *(Closed-form asymptotics for \(\ell_2\) regularization)* In the asymptotic limit defined above, the training and generalisation errors \cite{4} of the estimator \(\hat{w} \in \mathbb{R}^d\) solving the empirical risk minimisation problem in eq. \ref{eq:empirical-risk} with \(\ell_2\) regularization \(r(w) = \frac{1}{2} \|w\|_2^2\) verify:

\[
\mathcal{E}_{\text{train}}(\hat{w}) \xrightarrow{d \to \infty} \mathbb{E}_{\omega, h \sim N(0, 1)} \left[ g \left( \text{prox}_{\nu^*(g(\cdot), \nu_0(\sqrt{\rho}))} \left( \frac{m^*}{\sqrt{\rho}} h + \sqrt{q^* - \frac{m^*^2}{\rho}} \tilde{f}_0(\sqrt{\rho}s) \right) , \nu_0(\sqrt{\rho}s) \right) \right]
\]

\[
\mathcal{E}_{\text{gen}}(\hat{w}) \xrightarrow{d \to \infty} \mathbb{E}_{(\nu, \lambda)} \left[ g \left( \hat{f}(\lambda), \nu_0(\nu) \right) \right]
\]

\[
\mathcal{E} \equiv \mathcal{E}_{\text{train}}(\hat{w}) - \mathcal{E}_{\text{gen}}(\hat{w})
\]

\[
\mathcal{E} \xrightarrow{d \to \infty} \mathbb{E}_{(\nu, \lambda)} \left[ g \left( \hat{f}(\lambda), \nu_0(\nu) \right) \right] - \mathbb{E}_{\omega, h \sim N(0, 1)} \left[ g \left( \text{prox}_{\nu^*(g(\cdot), \nu_0(\sqrt{\rho}))} \left( \frac{m^*}{\sqrt{\rho}} h + \sqrt{q^* - \frac{m^*^2}{\rho}} \tilde{f}_0(\sqrt{\rho}s) \right) , \nu_0(\sqrt{\rho}s) \right) \right]
\]

\[
\mathcal{E} \xrightarrow{d \to \infty} \mathbb{E}_{(\nu, \lambda)} \left[ g \left( \hat{f}(\lambda), \nu_0(\nu) \right) \right] - \mathbb{E}_{\omega, h \sim N(0, 1)} \left[ g \left( \text{prox}_{\nu^*(g(\cdot), \nu_0(\sqrt{\rho}))} \left( \frac{m^*}{\sqrt{\rho}} h + \sqrt{q^* - \frac{m^*^2}{\rho}} \tilde{f}_0(\sqrt{\rho}s) \right) , \nu_0(\sqrt{\rho}s) \right) \right]
\]
where prox stands for the proximal operator defined as
\[ \text{prox}_{\nu g(\cdot,y)}(x) = \arg\min_z \{g(z,y) + \frac{1}{2\nu} (x-z)^2\} \]  
and where \((\nu, \lambda)\) are jointly Gaussian scalar variables:
\[ (\nu, \lambda) \sim \mathcal{N}\left(0, \begin{bmatrix} \rho & m^* \\ m^* & q^* \end{bmatrix}\right), \]
and the overlap parameters \((V^*, q^*, m^*)\) are prescribed by the unique fixed point of the following set of self-consistent equations:
\[
\begin{align*}
V &= \mathbb{E}_{(\omega, \vartheta) \sim \mu} \frac{\omega}{\lambda + V \omega} \\
m &= \frac{m}{\sqrt{q}} \mathbb{E}_{(\omega, \vartheta) \sim \mu} g^2 \frac{\lambda + V \omega}{\sqrt{q + \rho^{-1} m^2 h}} \\
q &= \mathbb{E}_{(\omega, \vartheta) \sim \mu} \frac{m^2 \omega^2 + q \vartheta^2}{\sqrt{q + \rho^{-1} m^2 h}} \\
\hat{V} &= \frac{\mathbb{E}_{s, h \sim \mathcal{N}(0,1)} f_g(V, m, q)}{\mathbb{E}_{s, h \sim \mathcal{N}(0,1)} \left[ s f_g(V, m, q) - m \rho q f_g(V, m, q) \right]} \\
\hat{m} &= \frac{m}{\sqrt{q}} \mathbb{E}_{s, h \sim \mathcal{N}(0,1)} \left[ \frac{m^2}{\sqrt{q + \rho^{-1} m^2 h}} - f_g(V, m, q) \right]
\end{align*}
\]
where we defined the scalar random functions \(f_g(V, m, q) = \text{prox}_{\nu g(\cdot,y)}(\rho^{-1/2} m s + \sqrt{q - \rho^{-1} m^2 h})\) and \(f_g^s(V, m, h) = \text{prox}_{\nu g(\cdot,y)}(\rho^{-1/2} m s + \sqrt{q - \rho^{-1} m^2 h})\) as the first derivative of the proximal operator.

**Proof:** This result is a consequence of Theorem 2, whose proof can be found in appendix B.

The parameters of the model \((\theta_0, \Omega, \Phi, \Psi)\) only appear through \(\rho, q, \rho_q, \rho_{q^2}\), and the asymptotic limit \(\mu\) of the joint distribution eq. (7) and \((\hat{f}, g, \lambda)\). One can easily iterate the above equations to find their fixed point, and extract \((q^*, m^*)\) which appear in the expressions for the training and generalisation errors \((\mathcal{E}^\star_{\text{train}}, \mathcal{E}^\star_{\text{gen}})\), see eq. (4). Note that \((q^*, m^*)\) have an intuitive interpretation in terms of the estimator \(\hat{w} \in \mathbb{R}^d\):
\[
q^* = \frac{1}{d} \hat{w}^\top \Omega \hat{w}, \quad m^* = \frac{1}{\sqrt{d \rho}} \hat{\theta}_0^\top \Phi \hat{w}
\]
Or in words: \(m^*\) is the correlation between the estimator projected in the teacher space, while \(q^*\) is the reweighted norm of the estimator by the covariance \(\Omega\). The parameter \(V^*\) also has a concrete interpretation: it parametrizes the deformation that must be applied to a Gaussian field specified by the solution of the fixed point equations to obtain the asymptotic behaviour of \(\hat{z}\). It prescribes the degree of non-linearity given to the linear output by the chosen loss function. This is coherent with the robust regression viewpoint, where one introduces non-square losses to deal with the potential non-linearity of the generative model. \(V^*\) plays a similar role for the estimator \(\hat{w}\) through the proximal operator of the regularisation, see Theorem 4 and 5 in the Appendix. Two cases are of particular relevance for the experiments that follow. The first is the case of ridge regression, in which \(f_0(x) = \hat{f}(x)\) and both the loss \(g\) and the performance measure \(\hat{g}\) are taken to be the mean-squared error \(\text{mse}(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2\), and the asymptotic errors are given by the simple closed-form expression:
\[
\mathcal{E}^\star_{\text{gen}} = \rho + q^* - 2 m^*, \quad \mathcal{E}^\star_{\text{train}} = \frac{\mathcal{E}^\star_{\text{gen}}}{(1 + V^*)^2},
\]
The second case of interest is the one of a binary classification task, for which \(f_0(x) = \hat{f}(x) = \text{sign}(x)\), and we choose the performance measure to be the classification error \(\hat{g}(y, \hat{y}) = \mathbb{P}(y \neq \hat{y})\).
In the same notation as before, the asymptotic generalisation error in this case reads:
\[
\mathcal{E}^\star_{\text{gen}} = \frac{1}{\pi} \cos^{-1}\left(\frac{m^*}{\sqrt{\rho q^*}}\right),
\]
while the training error \(\mathcal{E}^\star_{\text{train}}\) depends on the choice of \(g\) - which we will take to be the logistic loss \(\hat{g}(y, x) = \log(1 + e^{-xy})\) in all of the binary classification experiments.

As mentioned above, this paper includes stronger technical results including finite size corrections and precise characterization of the distribution of the estimator \(\hat{w}\), for generic, non-separable loss and
regularization $g$ and $r$. This type of distributional statement is encountered for special cases of the model in related works such as [28, 29, 51]. Define $V \in \mathbb{R}^{n \times d}$ as the matrix of concatenated samples used by the student. Informally, in high-dimension, the estimator $\hat{w}$ and $\bar{z} = \frac{1}{\sqrt{d}} V \hat{w}$ roughly behave as non-linear transforms of Gaussian random variables centered around the teacher vector $\theta_0$ (or its projection on the covariance spaces) as follows:

$$\hat{w}^* = \Omega^{-1/2} \frac{\text{prox}_{\nu r(\Omega^{-1/2})} \left( \frac{1}{\sqrt{V_\sigma}} (\hat{m}_t t + \sqrt{q}^* g) \right)}{\|t\|_2}, \quad z^* = \frac{\text{prox}_{\nu^* g(\cdot, z)} \left( \frac{m^* s + \sqrt{q}^* - (m^*)^2}{\rho} h \right)}{\|s\|_2}.$$

where $s, h \sim \mathcal{N}(0, I_n)$ and $g \sim \mathcal{N}(0, I_d)$ are random vectors independent of the other quantities, $t = \Omega^{-1/2} \Phi^T \theta_0$, $y = f_0 (\sqrt{\nu} s)$, and $(V^*, \hat{V}^*, q^*, \hat{q}^*, m^*, \hat{m}^*)$ is the unique solution to the fixed point equations presented in Lemma 12 of appendix B. Those fixed point equations are the generalization of (13) to generic, non-separable loss function and regularization. The formal concentration of measure result can then be stated in the following way:

**Theorem 2.** *(Non-asymptotic version, generic loss and regularization)* Under Assumption (B.1), consider any optimal solution $\hat{w}$ to (3). Then, there exist constants $C, c, c' > 0$ such that, for any Lipschitz function $\phi_1 : \mathbb{R}^d \to \mathbb{R}$, and separable, pseudo-Lipschitz function $\phi_2 : \mathbb{R}^n \to \mathbb{R}$ and any $0 < \epsilon < c'$:

$$\mathbb{P} \left( \left| \phi_1 \left( \frac{\hat{w}}{\sqrt{d}} \right) - \mathbb{E} \phi_1 \left( \frac{w^*}{\sqrt{d}} \right) \right| \geq \epsilon \right) \leq \frac{C e^{-cn \epsilon^2}}{c^2}, \quad \mathbb{P} \left( \left| \phi_2 \left( \frac{\hat{z}}{\sqrt{n}} \right) - \mathbb{E} \phi_2 \left( \frac{z^*}{\sqrt{n}} \right) \right| \geq \epsilon \right) \leq \frac{C}{c^2} e^{-cn \epsilon^2}.$$

Note that in this form, the dimensions $n, p, d$ still appear explicitly, as we are characterizing the convergence of the estimator’s distribution for large but finite dimension. The clearer, one-dimensional statements are recovered by taking the $n, p, d \to \infty$ limit with separable functions and an $\ell_2$ regularization. Other simplified formulas can also be obtained from our general result in the case of an $\ell_1$ penalty, but since this breaks rotational invariance, they do look more involved than the $\ell_2$ case. From Theorem 2, one can deduce the expressions of a number of observables, represented by the test functions $\phi_1, \phi_2$, characterizing the performance of $\hat{w}$, for instance the training and generalization error. A more detailed statement, along with the proof, is given in appendix B.

### 3 Applications of the Gaussian model

We now discuss how the theorems above are applied to characterise the learning curves for a range of concrete cases. We present a number of cases – some rather surprising – for which Conjecture 1 seems valid, and point out some where it is not. An out-of-the-box iterator for all the cases studied hereafter is provided in the GitHub repository for this manuscript at https://github.com/IdePHICS/GCMProject.

#### 3.1 Random kitchen sink with Gaussian data

If we choose random feature maps $\varphi_s(x) = \sigma(Fx)$ for a random matrix $F$ and a chosen scalar function $\sigma$ acting component-wise, we obtain the random kitchen sink model [16]. This model has seen a surge of interest recently, and a sharp asymptotic analysis was provided in the particular case of uncorrelated Gaussian data $x \sim \mathcal{N}(0, I_d)$ and $\varphi_l(x) = x$ in [11, 12] for ridge regression and generalised by [43, 46] for generic convex losses. Both results can be framed as a Gaussian covariate model with:

$$\Psi = I_p, \quad \Phi = \kappa_1 F^T, \quad \Omega = \kappa_0^2 I_d 1_d^T + \kappa_1^2 \frac{FF^T}{d} + \kappa_2^2 I_d,$$

where $1_d \in \mathbb{R}^d$ is the all-one vector and the constants $(\kappa_0, \kappa_1, \kappa_2)$ are related to the non-linearity $\sigma$:

$$\kappa_0 = \mathbb{E}_{z \sim \mathcal{N}(0,1)} [\sigma(z)], \quad \kappa_1 = \mathbb{E}_{z \sim \mathcal{N}(0,1)} [z \sigma(z)], \quad \kappa_2 = \sqrt{\mathbb{E}_{z \sim \mathcal{N}(0,1)} [\sigma(z)^2] - \kappa_0^2 - \kappa_1^2}.$$

In this case, the averages over $\mu$ in eq. (13) can be directly expressed in terms of the Stieltjes transform associated with the spectral density of $F F^T$. Note, however, that our present framework can accommodate more involved random sinks models, such as when the teacher features are also a random feature model or multi-layer random architectures.

#### 3.2 Kernel methods with Gaussian data
Another direct application of our formalism is to kernel methods. Kernel methods admit a dual representation in terms of optimization over feature space [53]. The connection is given by Mercer’s theorem, which provides an eigen-decomposition of the kernel and of the target function in the feature basis, effectively mapping kernel regression to a teacher-student problem on feature space. The classical way of studying the performance of kernel methods [54, 55] is then to directly analyze the performance of convex learning in this space. In our notation, the teacher and student feature maps are equal, and we thus set \( p = d, \Psi = \Phi = \Omega = \text{diag}(\omega_i) \) where \( \omega_i \) are the eigenvalues of the kernel and we take the teacher weights \( \theta_0 \) to be the decomposition of the target function in the kernel feature basis.

There are many results in classical learning theory on this problem for the case of ridge regression (where the teacher is usually called “the source” and the eigenvalues of the kernel matrix the “capacity”, see e.g. [54, 56]). However, these are worst case approaches, where no assumption is made on the true distribution of the data. In contrast, here we follow a typical case analysis, assuming Gaussianity in feature space. Through Theorem 1 this allows us to go beyond the restriction of the ridge loss. An example for logistic loss is in Fig. 2.

For the particular case of kernel ridge regression, Th. 1 provides a rigorous proof of the formula conjectured in [52]. App. A.6 presents an explicit mapping to their results. Hard-margin Support Vector Machines (SVMs) have also been studied using the heuristic replica method from statistical physics in [57, 58]. In our framework, this corresponds to the hinge loss \( g(x, y) = \max(0, 1 - xy) \) when \( \lambda \to 0^+ \). Our theorem thus puts also these works on rigorous grounds, and extends them to more general losses and regularization.

### 3.3 GAN-generated data and learned teachers

To approach more realistic data sets, we now consider the case in which the input data \( x \in \mathcal{X} \) is given by a generative neural network \( x = \mathcal{G}(z) \), where \( z \) is a Gaussian i.i.d. latent vector. Therefore, the covariates \( [u, v] \) are the result of the following Markov chain:

\[
\begin{align*}
  z &\rightarrow x \in \mathcal{X} \rightarrow u \in \mathbb{R}^p, \\
  z &\rightarrow x \in \mathcal{X} \rightarrow v \in \mathbb{R}^d.
\end{align*}
\]

(19)

With a model for the covariates, the missing ingredient is the teacher weights \( \theta_0 \in \mathbb{R}^p \), which determine the label assignment: \( y = f_0(u^\top \theta_0) \). In the experiments that follow, we fit the teacher weights from the original data set in which the generative model \( \mathcal{G} \) was trained. Different choices for the fitting yield different teacher weights, and the quality of label assignment can be assessed by the performance of the fit on the test set. The set \((\varphi_t, \varphi_s, \mathcal{G}, \theta_0)\) defines the data generative process. For predicting the learning curves from the iterative eqs. (13) we need to sample from the spectral measure \( \mu \), which amounts to estimating the population covariances \((\Psi, \Phi, \Omega)\). This is done from the generative process in eq. (19) with a Monte Carlo sampling algorithm. This pipeline is explained in detail in Appendix D. An open source implementation of the algorithms used in the experiments is available online at https://github.com/IdePHICS/GCMProject.

Fig. 3 shows an example of the learning curves resulting from the pipeline discussed above in a logistic regression task on data generated by a GAN trained on CIFAR10 images. More concretely,
we used a pre-trained five-layer deep convolutional GAN (dcGAN) from [59], which maps 100 dimensional i.i.d. Gaussian noise into $k = 32 \times 32 \times 3$ realistic looking CIFAR10-like images: $G: z \in \mathbb{R}^{100} \mapsto x \in \mathbb{R}^{32 \times 32 \times 3}$. To generate labels, we trained a simple fully-connected four-layer neural network on the real CIFAR10 data set, on a odd ($y = +1$) vs. even ($y = -1$) task, achieving $\sim 75\%$ classification accuracy on the test set. The teacher weights $\theta_0 \in \mathbb{R}^p$ were taken from the last layer of the network, and the teacher feature map $\varphi_t$ from the three previous layers. For the student model, we trained a completely independent fully connected 3-layer neural network on the dcGAN-generated CIFAR10-like images and took snapshots of the feature maps $\varphi_i^s$ induced by the 2-first layers during the first $i \in \{0, 5, 50, 200\}$ epochs of training. Finally, once $(G, \varphi_t, \varphi_i^t, \theta_0)$ have been fixed, we estimated the covariances $(\Psi, \Phi, \Omega)$ with a Monte Carlo algorithm. Details of the architectures used and of the training procedure can be found in Appendix [D.1].

Fig.3 depicts the resulting learning curves obtained by training the last layer of the student. Interestingly, the performance of the feature map at epoch 0 (random initialisation) beats the performance of the learned features during early phases of training in this experiment. Another interesting behaviour is given by the separability threshold of the learned features, i.e. the number of samples for which the training loss becomes larger than 0 in logistic regression. At epoch 50 the learned features are separable at lower sample complexity $\alpha = n/d$ than at epoch 200 - even though in the later the training and generalisation performances are better.

### 3.4 Learning from real data sets

**Applying teacher/students to a real data set** — Given that the learning curves of realistic-looking inputs can be captured by the Gaussian covariate model, it is fair to ask whether the same might be true for real data sets. To test this idea, we first need to cast the real data set into the teacher-student formalism, and then compute the covariance matrices $\Omega, \Psi, \Phi$ and teacher vector $\theta_0$ required by model [1].

---

[Figure 3: **Left:** generalisation classification error (top) and (unregularised) training loss (bottom) vs the sample complexity $\alpha = n/d$ for logistic regression on a learned feature map trained on dcGAN-generated CIFAR10-like images labelled by a teacher fully-connected neural network (see Appendix [D.1] for architecture details), with vanishing $\ell_2$ regularisation. The different curves compare featured maps at different epochs of training. The theoretical predictions based on the Gaussian covariate model (full lines) are in very good agreement with the actual performance (points). **Right:** Test classification error (top) and (unregularised) training loss, (bottom) for logistic regression as a function of the number of samples $n$ for an animal vs not-animal binary classification task with $\ell_2$ regularization $\lambda = 10^{-2}$, comparing real CIFAR10 grey-scale images (blue) with dcGAN-generated CIFAR10-like gray-scale images (red). The real-data learning curve was estimated, just as in Figs.4 from the population covariances on the full data set, and it is not in agreement with the theory in this case. On the very right we depict the histograms of the variable $\frac{1}{\sqrt{d}} v^\top \hat{w}$ for a fixed number of samples $n = 2d = 2048$ and the respective theoretical predictions (solid line). Simulations are averaged over 10 independent runs.](image-url)
Let \( \{ x^\mu, y^\mu \}_{\mu=1}^{n_{\text{tot}}} \) denote a real data set, e.g. MNIST or Fashion-MNIST for concreteness, where \( n_{\text{tot}} = 7 \times 10^4 \), \( x^\mu \in \mathbb{R}^D \) with \( D = 784 \). Without loss of generality, we can assume the data is centred. To generate the teacher, let \( u^\mu = \varphi_\theta(x^\mu) \in \mathbb{R}^p \) be a feature map such that data is invertible in feature space, i.e. that \( y^\mu = \theta_0^\top u^\mu \) for some teacher weights \( \theta_0 \in \mathbb{R}^p \), which should be computed from the samples. Similarly, let \( u^\mu = \varphi_\gamma(x^\mu) \in \mathbb{R}^d \) be a feature map we are interested in studying. Then, we can estimate the population covariances \( \Psi, \Phi, \Omega \) empirically from the entire data set as:

\[
\Psi = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} u^\mu u^\mu^\top, \quad \Phi = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} u^\mu v^\mu v^\mu^\top, \quad \Omega = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} v^\mu v^\mu^\top.
\]

At this point, we have all we need to run the self-consistent equations (13). The issue with this approach is that there is not a unique teacher map \( \varphi_\gamma \) and teacher vector \( \theta_0 \) that fit the true labels. However, we can show that all interpolating linear teachers are equivalent:

**Theorem 3.** (Universality of linear teachers) For any feature teacher map \( \varphi_\gamma \), and for any \( \theta_0 \) that interpolates the data so that \( y^\mu = \theta_0^\top u^\mu \forall \mu \), the asymptotic predictions of model (1) are equivalent.

**Proof.** It follows from the fact that the teacher weights and covariances only appear in eq. (13) through \( \tilde{\rho} = \frac{1}{p} \theta_0^\top \Psi \theta_0 \) and the projection \( \Phi^\top \theta_0 \). Using the estimation (20) and the assumption that it exists \( y^\mu = \theta_0^\top u^\mu \), one can write these quantities directly from the labels \( y^\mu \):

\[
\tilde{\rho} = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} (y^\mu)^2, \quad \Phi^\top \theta_0 = \frac{1}{n_{\text{tot}}} \sum_{\mu=1}^{n_{\text{tot}}} y^\mu u^\mu.
\]

For linear interpolating teachers, results are thus independent of the choice of the teacher. \( \square \)

Although this result might seen surprising at first sight, it is quite intuitive. Indeed, the information about the teacher model only enters the Gaussian covariate model (1) through the statistics of \( u^\top \theta_0 \).

For a linear teacher \( f_0(x) = x \), this is precisely given by the labels.

**Ridge Regression with linear teachers**

We now test the prediction of model (1) on real data sets, and show that it is surprisingly effective in predicting the learning curves, at least for the ridge regression task. We have trained a 3-layer fully connected neural network with ReLU activations on the full Fashion-MNIST data set to distinguish clothing used above vs. below the waist (60). The student feature map \( \varphi_\gamma : \mathbb{R}^{784} \rightarrow \mathbb{R}^d \) is obtained by removing the last layer, see Appendix D.1 for a detailed description. In Fig. 4, we show the test and training errors of the ridge estimator on a sub-sample of \( n < n_{\text{tot}} \) on the Fashion-MNIST images. We observe remarkable agreement between the learning curve obtained from simulations and the theoretical prediction by the matching Gaussian covariate model. Note that for the square loss and for \( \lambda \ll 1 \), the worst performance peak is located at the point in which the linear system becomes invertible. Curiously, Fig. 4 shows that the fully-connected network progressively learns a low-rank representation of the data as training proceeds. This can be directly verified by counting the number of zero eigenvalues of \( \Omega \), which go from a full-rank matrix to a matrix of rank 380 after 200 epochs of training.

Figure 4: Test and training mean-squared errors eqs. (15) as a function of the number of samples \( n \) for ridge regression. The Fashion-MNIST data set, with vanishing regularisation \( \lambda = 10^{-5} \). In this plot, the student feature map \( \varphi_\gamma \) 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. Simulations are averaged over 10 independent runs. Further details on the simulations are described in Appendix D.1.
characteristic double-descent behaviour [9, 25, 62], and the accurate prediction of the peak where the interpolation transition occurs. We note in Appendix D.1 that for both Figs. 4 and 1, for a number of samples \( n \) closer to \( n_{\text{tot}} \) we start to see deviations between the real learning curve and the theory. This is to be expected since in the teacher-student framework the student can, in principle, express the same function as the teacher if it recovers its weights exactly. Recovering the teacher weights becomes possible with a large training set. In that case, its test error will be zero. However, in our setup the test error on real data remains finite even if more training data is added, leading to the discrepancy between teacher-student learning curve and real data, see Appendix D.1 for further discussion.

Why is the Gaussian model so effective for describing learning with data that are not Gaussian? The point is that ridge regression is sensitive only to second order statistics, and not to the full distribution of the data. It is a classical property (see Appendix E) that the training and generalisation errors are only a function of the spectrum of the empirical and population covariances, and of their products. Random matrix theory teaches us that such quantities are very robust, and their asymptotic behaviour is universal for a broad class of distributions of \([u, v]\) [49, 63–65]. The asymptotic behavior of kernel matrices has indeed been the subject of intense scrutiny [11, 47, 48, 50, 66, 67]. Indeed, a universality result akin to Theorem [3] was noted in [41] in the specific case of kernel methods. We thus expect the validity of model (1) for ridge regression, with a linear teacher, to go way beyond the Gaussian assumption.

Beyond ridge regression — The same strategy fails beyond ridge regression and mean-squared test error. This suggests a limit in the application of model (1) to real (non-Gaussian) data to the universal linear teacher. To illustrate this, consider the setting of Figs. 4, and compare the model predictions for the binary classification error instead of the \( \ell_2 \) one. There is a clear mismatch between the simulated performance and prediction given by the theory (see Appendix D.1) due to the fact that the classification error does not depends only on the first two moments.

We present an additional experiment in Fig. 3. We compare the learning curves of logistic regression on a classification task on the real CIFAR10 images with the real labels versus the one on dCGAN-generated CIFAR10-like images and teacher generated labels from Sec. 3.3. While the Gaussian theory captures well the behaviour of the later, it fails on the former. A histogram of the distribution of the product \( u^\top \hat{w} \) for a fixed number of samples illustrates well the deviation from the prediction of the theory with the real case, in particular on the tails of the distribution. The difference between GAN generated data (that fits the Gaussian theory) and real data is clear. Given that for classification problems there exists a number of choices of “sign” teachers and feature maps that give the exact same labels as in the data set, an interesting open question is: is there a teacher that allows to reproduce the learning curves more accurately? This question is left for future works.

Acknowledgements

We thank Romain Couillet, Cosme Louart, Loucas Pillaud-Vivien, Matthieu Wyart, Federica Gerace, Luca Saglietti and Yue Lu for discussions. We are grateful to Kabir Aladin Chandrasekher, Ashwin Pananjady and Christos Thrampoulidis for pointing out discrepancies in the finite size rates and insightful related discussions. We acknowledge funding from the ERC under the European Union’s Horizon 2020 Research and Innovation Programme Grant Agreement 714608-SMiLe, and from the French National Research Agency grants ANR-17-CE23-0023-01 PAIL.
References


