Free Dynamics of Feature Learning Processes

Cyril Furtlehner

Free Dynamics of Feature Learning Processes

Cyril Furtlehner [*]

Abstract

Regression models usually tend to recover a noisy signal in the form of a combination of regressors, also called features in machine learning, themselves being the result of a learning process. The alignment of the prior covariance feature matrix with the signal is known to play a key role in the generalization properties of the model, i.e. its ability to make predictions on unseen data during training. We present a statistical physics picture of the learning process. First we revisit the ridge regression to obtain compact asymptotic expressions for train and test errors, rendering manifest the conditions under which efficient generalization occurs. It is established thanks to an exact test-train sample error ratio combined with random matrix properties. Along the way in the form of a self-energy emerges an effective ridge penalty — precisely the train to test error ratio — which offer a very simple parameterization of the problem. This formulation appears convenient to tackle the learning process of the feature matrix itself. We derive an autonomous dynamical system in terms of elementary degrees of freedom of the problem determining the evolution of the relative alignment between the population matrix and the signal. A macroscopic counterpart of these equations is also obtained and various dynamical mechanisms are unveiled, allowing one to interpret the dynamics of simulated learning processes and reproduce trajectories of single experimental run with high precision.

1 Introduction

The rapid developments of the last decades in machine learning (ML), in particular around the deep neural networks (DNN), have led statisticians to revisit the mathematical foundations of the field [1, 2], where at odds with intuition, highly over-parameterized models (typically DNN) perform well in generalization [3, 4]. In the context of supervised learning that we are interested in here, data are given in the form of input-output pairs $(\mathbf{x}, y)$ and the goal is to learn a function f associating $\mathbf{x}$ to y. The input $\mathbf{x} \in \mathbb{R}^d$ belong to some d-dimensional embedding space, d being possibly large and the $y \in \mathbb{R}$ is taken as a scalar for simplicity. The function is often chosen in some parameterized class of function f_θ, like neural networks, after a loss

[*]Inria Saclay - Tau team, LISN - AO team, Bât 660 Université Paris-Saclay, Orsay Cedex 91405
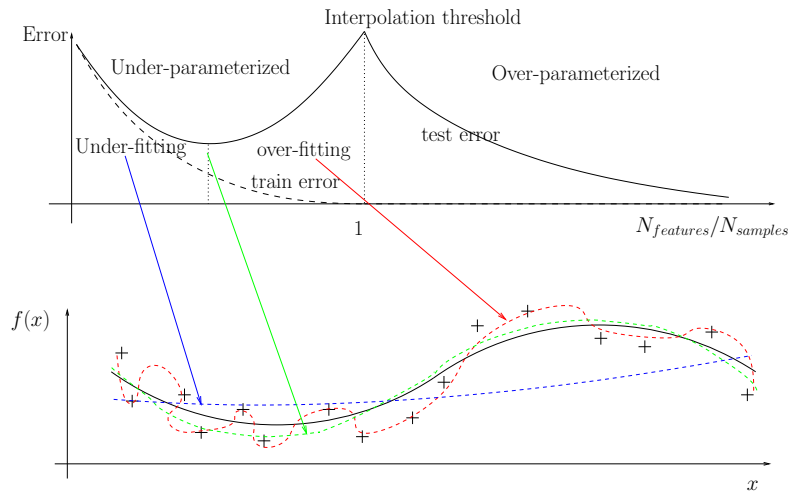
Fig. 1.1: Double descent "paradox". On the left is the traditional picture of a U-shape curve of the test error showing first under-fitting corresponding to high bias, followed by over-fitting characterized by high variance as illustrated on the 1-d regression problem. The right part corresponds often to the deep learning setting with high complexity/expressivity of the model along with low test error.

function $\mathcal{L}(\theta|\mathcal{D})$ has been designed and minimized conditionally to some training set $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots N\}$. Then its ability to generalize well on unseen data is assessed from the test error, estimated on another set of data called the test set, statistically independent from the training set. The essence of the dilemma has been summarized as the double descent picture in [5], which is reproduced here in Figure 1.1. The traditional picture concerns the transition from under-fitting to over-fitting as the complexity of the model is increased corresponding to the best trade-off between bias and variance as illustrated on the 1-d regression model. At some (sweet) point, while the train error continues to decrease, the test error instead starts to increase again to reach a maximum at the interpolation threshold where the number of parameters equals the number of train examples. The novelty with the advent of deep learning is that when going far into the over-parameterized region, better generalization performances can sometime be obtained than those corresponding to the sweet spot separating the underfitting from the overfitting region.

This observation has revived the interest in the generalization mechanism in ML and triggered a vast number of works during the last few years, which we only give here a very brief and partial account. Most of them point to the effect of some form of implicit regularization, meaning that either the architecture itself or the learning process itself provide some inductive bias penalizing automatically "complex" features schematically corresponding to the UV spectrum, and favoring "simpler" one, i.e. typically in the IR part of the spectrum. In a series of paper the linear ridge regression was revisited [6, 7, 8, 9], with various averaging hypothesis regarding the signal and the input data, and showing to exhibit in some cases this double descent

behaviour. In [7, 10], is remarked in particular that some implicit regularization is provided by the learning dynamics itself, where the lower part of the spectrum of the feature matrix which is potentially harmful is also learned after the stronger hence justifying early stopping strategies, while some spectral gap expected from random matrix theory may protect against overfitting in the overparameterized regime. In [8] another mechanism of implicit regularization called benign over-fitting is unveiled. It may occur when the population matrix has a large number of low eigenvalues decaying slowly. This is interpreted as self induced regularization [11], where the lower part of the prior spectrum, the "tail" features, acts as an L_2 regularization if the corresponding eigenvalues span a bounded interval. Extending asymptotic results of [6] based on random matrix theory (RMT), another mechanism explaining good/bad generalization properties is identified in [12, 13]. This mechanism involves both the signal and the population matrix, by considering the alignment between the two and was already discussed in [14] in the context of NN with random projections. Good generalization properties are observed whenever the signal decomposes preferentially on the strong modes of the population matrix. This in our point of view is a key point to characterize feature quality and generalization, and we elaborate on this in this work. These considerations may have also implications to define pre-conditioning strategies [15] in order to favor the modes mostly correlated with the signal. Extensions of double-descent analysis to non-linear models have been also explored in various works [16, 17]. Concerning the conditions of use of RMT, these are often met in the context of ML to allow precise analysis [18], because ML models typically involve large weights matrices of weekly correlated entries. This also offer the possibility to study the learning dynamics as already exemplified in [7, 10].

On the statistical physics side many works have also been discussing the double descent and generalization mechanisms. In [19] it is argued that large neural network perform as model ensembles through collective effects of roughly independent part of the NN, automatically reducing the variance which is otherwise expected to dominate in the overparameterized regime. Methods inherited from the spin-glass theory [20] allow one in principle to study a great variety of learning scenario [21] but need some efforts to go beyond oversimplified toy model architecture and data distributions to be of practical interest to ML. For quite some time now, many results have been obtained in this direction by various groups, concerning prediction for learning curves in various non-trivial regimes [22, 23, 24, 25, 26, 27, 28] and the learning dynamics [29, 30, 31] with various settings around the teacher-student one [32] for instance. Another convenient setting to analyze the learning process is to consider the kernel regime or equivalently the random feature model [33]. In these cases the data are sent respectively into an infinite dimensional space or at least to a sufficiently large dimensional space where regression or classification tasks can be done efficiently with linear methods. These regimes are found to be relevant when analyzing large NN when the width of the layers are very large. It is referred to as the lazy training regime [34], because there the weights defining the features barely change after initialization. In this case the neural network behave as a Gaussian process [35, 36] as can be proved in the limit of infinite size layers [37]. It then performs as kernel regression with a deterministic kernel called neural tangent kernel [38] resulting from central limit theorem applied layer-wise.

In this work we provide new insights to generalization mechanisms and feature

learning, by exploiting asymptotic results obtained in the standard RMT regime, namely the planar diagram approximation, and setting up a statistical physics description of the dynamics of the learning process. In order to make the paper self-contained, we first recall in Section 2 to which extent linear ridge regression is relevant to understand NN and discuss in Section 2.2 the bias-variance dilemma in this context. Then Section 3 is devoted to obtain general asymptotic expressions, in the context of the ridge regression, for the train, test error and loss functions. First we show in Section 3.1 the existence of a simple relation between train and test error for a given sample and how this is converted asymptotically to a deterministic train-test error ratio in Section 3.2, while in Section 3.3, the Feynman diagram formalism used to justify these formulas leads us also to define the effective ridge penalty of the problem, which then considerably simplifies the analysis of the generalization error given at the end of this section. In Section 4 we turn to the analysis of the learning dynamics intended to understand the learning process of the features themselves. We consider a semi-lazy regime, the parameters of the features being in the lazy regime, while the weights of the last layer solve a ridge regression conditionally on the features themselves. In Section 4.1, based on the asymptotic expressions obtained in Section 3, we derive an autonomous set of dynamical equations corresponding to the microscopic level description, i.e. concerning the dynamics of the basic degrees of freedom of the problem. Macroscopic counterparts of these equations are given in Section 4.3 and special cases are studied in Section 4.5.

2 From neural networks to ridge regression

2.1 Linear regime of learning

In some regimes neural networks can behave approximately as linear regression models [38]. The starting point is to regress some signal observed in the form

$$y = f(\mathbf{x}) + \epsilon \tag{2.1}$$

where the signal itself corresponds to some unknown function f on some input $\mathbf{x} \in \mathbb{R}^d$, with d the supposedly large embedding dimension of the input, sumperimposed with some decorrelated noise $\epsilon = \mathcal{N}(0, \sigma_\epsilon^2)$. Given some training data $\mathcal{D} = \{(\mathbf{x}^{(s)}, y^{(s)}), s = 1, \dots N\}$, the regression problem amounts then to find the best minimizer f_θ among a parametric family indexed by θ, of the loss function chosen by convenience to be MSE:

$$\mathcal{L}(\theta) = \frac{1}{2N} \sum_{s=1}^{N} |y^{(s)} - f_\theta(\mathbf{x}^{(s)})|^2. \tag{2.2}$$

Assume first that we are given a supposedly large set of N_f random features, $\{f_k(\mathbf{x}), k = 1, \dots N_f\}$, then our model can be chosen in the linear family

$$f_\theta(\mathbf{x}) = \sum_{k=1}^{N_f} w_k f_k(\mathbf{x}), \tag{2.3}$$

with $\theta = \{w_1, \ldots, w_{N_f}\}$. In order this to have a chance to work we need at least that the projection $f^{\parallel}$ of f on the subspace spanned by these features represents a significant fraction of the whole signal. If for instance we look for a smooth function with no privileged direction in space, requiring a resolution a on a domain of linear size L, we typically need a number of features

$$D = \left(\frac{L}{a}\right)^d,$$

which becomes prohibitive as soon as $d \gg 1$. The way to overcome the so-called "curse of dimensionality" is to consider non-random features obtained by some selection procedure. Multilayer neural network for instance constitute an efficient solution to do that, where typically the last layer corresponds to (2.3) with $f_k = f_k^{(N_l)}$ and the recursive definition

$$f_k^{(n)}(\mathbf{x}) = \phi\Big[\sum_\ell w_{k\ell}^{(n-1)} f_\ell^{(n-1)}(\mathbf{x}) - b_k^{(n)}\Big]$$

of the feature of the nth layer, ϕ being the so-called activation function. The parameters of this family now correspond to $\theta = \{w_{k,\ell}^{(n)}, b_k^{(n)}\}$ and the retro-propagation of the gradient of the loss (2.2) through the layer will lead in principle to find better than random features.

Still, as noticed in [38] a linear regime of learning can be recovered on this complex models. To see this first we may consider the continuous limit of the learning process indexed by "time" t

$$\partial_t \theta_t = -\gamma \nabla_\theta \mathcal{L}(\theta_t), \tag{2.4}$$

γ representing the learning rate in this limit. In [38] is introduced the following quantity

$$G_\theta(\mathbf{x}, \mathbf{x}') \stackrel{\text{def}}{=} \nabla_\theta^T f_\theta(\mathbf{x}) \nabla_\theta f_\theta(\mathbf{x}'), \tag{2.5}$$

called "neural tangent kernel" (NTK) which as we shall see is the equivalent of a Green function in the context of field theories, and can serve for instance as a measure of similarity between points as seen from a NN [39]. This kernel allows one to study the learning dynamics directly in the linear space of functions $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ according to the following equation

$$\partial_t f_{\theta_t} = -\gamma G_{\theta_t} \hat{P}\big[f_{\theta_t} - f^\star\big], \tag{2.6}$$

where

$$\hat{P}(\mathbf{x}, \mathbf{x}') = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)\delta(\mathbf{x}' - \mathbf{x}_i),$$

is the projection operator on the data and $f^\star$ is a function verifying

$$f^\star(\mathbf{x}_i) = y_i, \qquad i = 1, \ldots N.$$

In the limit of infinitely large layers and with weights scaling like $w_{k\ell} = \mathcal{O}(1/\sqrt{N_l})$ if N_l is the size of the layers the NTK becomes a deterministic constant kernel G_{θ_0} yielding a linear dynamics in (2.6). In this regime each component of $\big[f_{\theta_t} - f^\star\big]$ on the eigenmodes of G_{θ_0} evolves independently, and converges with some exponential

decay at a speed determined by the corresponding eigenvalue. As remarked in [7] the strategy called "early stopping" yields a form of regularization by avoiding the learning of the components of f_{θ_t} corresponding to low eigenvalues, i.e. direction of the feature space which are poorly explored by the data as will be made more clear later on. This regime, where the parameters of the NN are displaced by the learning process from their initial values by quantities scaling like $\mathcal{O}\left(1/\sqrt{N_l}\right)$, is also called "lazy training" regime. The interpretation of this is the following. Let θ_0 be the parameter point at which f_θ is initialized at the beginning of the learning. Assuming small variations around θ_0 we have at first order

$$f_\theta(\mathbf{x}) = f_{\theta_0}(\mathbf{x}) + \sum_{k=1}^{N_f} \delta\theta_k \frac{\partial}{\partial\theta_k} f_\theta(\mathbf{x})\Big|_{\theta=\theta_0}$$

with $\delta\theta = \theta - \theta_0$ representing now the weights of the linear regression model defined in (2.3), after identifying the features

$$f_k(\mathbf{x}) = \frac{\partial}{\partial\theta_k} f_\theta(\mathbf{x})\Big|_{\theta=\theta_0},$$

where the input vector is now $(f_k(\mathbf{x}_i), k = 1, \ldots N_f)$ of dimension N_f, for a given input data $\mathbf{x}_i$. As discussed above, for large dimensional problem it is not clear that the features defined at initialization in this way have a good chance to be adapted to the problem, hence this regime, even if interesting, is likely not able to explain alone the success of neural network on difficult tasks [34].

2.2   The Bias-Variance trade-off for the ridge regression

A fundamental aspect of learning with a limited number of data is the problem of choosing properly the complexity of the model given the size of the training set of examples. This problem is formalized in terms of the bias-variance dilemma [1] when the loss is a mean-squared error as in (2.2). Indeed in that case the test error can be decomposed into bias and variance as follows. The optimal solution is formally given by

$$f(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}].$$

Then given a predictor $f(\mathbf{x}|\mathcal{D})$ obtained from a specific training set, we have when averaging over $\mathcal{D}$ the bias and variance decomposition of the test error:

$$\mathbb{E}_\mathcal{D}\left[\left(f(\mathbf{x}|\mathcal{D}) - \mathbb{E}[y|\mathbf{x}]\right)^2\right] = \left(\mathbb{E}_\mathcal{D}[f(\mathbf{x}|\mathcal{D})] - \mathbb{E}[y|\mathbf{x}]\right)^2 + \mathbb{E}_\mathcal{D}\left[\left(f(\mathbf{x}|\mathcal{D}) - \mathbb{E}_\mathcal{D}[f(\mathbf{x}|\mathcal{D})]\right)^2\right].$$

Consistency of the predictor is realized when both bias and variance go to zero along with the size of the dataset $\mathcal{D}$ going to infinity. Choosing the right complexity of the model at finite N is therefore equivalent, at least in the traditional ML view, as illustrated on Figure 1.1, to obtain the best trade-off between bias and variance. For model with low complexity we expect a high bias due to the lack of expressivity of the model, but a low variance as there are few parameters to estimate compared to the number of training samples. Instead when the model has high complexity, it will be able to fit precisely a noisy observation of the signal (2.1), so that when

averaging over $\mathcal{D}$ the bias get small, but the variance is expected to be larger since it cumulates the noise of the signal with the noise of the model.

Let us illustrate this on the ridge regression, in order also to introduce some notations to be used later on. Consider the noisy observation $y \in \mathbb{R}$

$$y = \mathbf{f}^t \mathbf{x} + \epsilon, \tag{2.7}$$

of some signal defined by the inner product of some fixed unknown vector $\mathbf{f} \in \mathbb{R}^D$ with some known input vector $\mathbf{x} \in \mathbb{R}^D$ of basic features and the noise ϵ of variance σ^2. The statistics by which $\mathbf{x}$ is represented in the data is summarized by the so-called "population matrix" in the statistics literature, i.e. the prior cross-product of the basic features, that we denote here by C,

$$C \stackrel{\text{def}}{=} \mathbb{E}_{\mathbf{x}}[\mathbf{x}\mathbf{x}^t]. \tag{2.8}$$

The ridge regression problem amounts to find $\hat{\mathbf{w}} \in \mathbb{R}^D$ minimizing the loss

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^t\mathbf{w} + \frac{\alpha}{2N}\sum_{s=1}^{N}\left|y^{(s)} - \mathbf{w}^t\mathbf{x}^{(s)}\right|^2 \tag{2.9}$$

where $\mathcal{D} = \{(\mathbf{x}^{(s)}, y^{(s)}), s = 1, \dots N\}$ is a training set generated by the teacher model (2.7) and α corresponds here to the inverse of the ridge penalty, introduced this way by convenience and to be later interpreted as a bare coupling constant. $\rho = N/N_f$ is the aspect ratio of the data matrix. Given the empirical averages

$$C^{(N)} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N}\mathbf{x}^{(s)}\mathbf{x}^{(s)t},$$

$$Z^{(N)} \stackrel{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N}\mathbf{x}^{(s)}y^{(s)},$$

and the resolvent

$$G_\alpha^{(N)} = \left(\mathbb{I} + \alpha C^{(N)}\right)^{-1}, \tag{2.10}$$

with $\mathbb{I}$ being the identity matrix, we obtain the optimal solution

$$\hat{\mathbf{w}}_\alpha^{(N)} = \alpha G_\alpha^{(N)} Z^{(N)}. \tag{2.11}$$

From this we can deduce the following expressions from the train and test errors for a given dataset

$$E_{\text{train}}^{(N)}(\alpha) = \frac{1}{\alpha}\text{Tr}\left[G_\alpha^{(N)}(\mathbb{I} - G_\alpha^{(N)})\mathbf{f}\mathbf{f}^t\right] + \sigma^2\left(1 - \rho^{-1} + \frac{1}{N}\text{Tr}\left[G_\alpha^{(N)2}\right]\right), \tag{2.12}$$

$$E_{\text{test}}^{(N)}(\alpha) = \text{Tr}\left[G_\alpha^{(N)}CG_\alpha^{(N)}\mathbf{f}\mathbf{f}^t\right] + \sigma^2\left(1 + \frac{\alpha}{N}\text{Tr}\left[G_\alpha^{(N)}(\mathbb{I} - G_\alpha^{(N)})C\right]\right). \tag{2.13}$$

after averaging over the noise and the test set following (2.8). This last equation exhibits the bias-variance decomposition of the generalization error, the first term on the rhs corresponding to the bias and the second one proportional to σ^2 corresponding to the variance. The effect of the coupling α is to sharpen the model,

by reducing the influence of the regularization term in (2.2) and the bias-variance dilemma becomes manifest by looking at the role of α. As we see, the variance term is composed of the contribution σ^2 due to the noise ϵ of the signal itself and a second term coming from the fact that this noise is fitted by the model. This latter contribution can be harmful when $C^{(N)}$ has a high density of low eigenvalues especially when α is large and the way to contain it is to reduce α. Instead, the bias term which is also problematic when $C^{(N)}$ has a high density of low eigenvalues can be reduced only by increasing α.

For now we assumed that the model has all the necessary features at disposal to reconstruct the signal which is not the standard case in practice. To take this into account we will in the following consider as in [7] a slightly more general setting: the dimension D of the embedding space of the input $\mathbf{x} \in \mathbb{R}^D$ will be assume to be large, while the model is given a set of $N_f < D$ features in the form of an $N_f \times D$ matrix F, and we look now for a student model of the form

$$y = \mathbf{w}^t F \mathbf{x} \tag{2.14}$$

with now $\mathbf{w} \in \mathbb{R}^{N_f}$. The population matrix has now some structure and reads

$$C = \mathbb{E}_{\mathbf{x}}\Big[F\mathbf{x}\mathbf{x}^t F^t \Big] = FF^t, \tag{2.15}$$

assuming without loss of generality, i.e. up to a linear change of input vector, that $\mathbf{x}$ follows an isotropic and normalized distribution, with $\mathbb{E}(\|\mathbf{x}\|^2) = D$. The optimal solution is still given by (2.11) but with now

$$C^{(N)} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{s=1}^{N} F\mathbf{x}^{(s)}\mathbf{x}^{(s)^t} F^t, \tag{2.16}$$

$$Z^{(N)} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{s=1}^{N} F\mathbf{x}^{(s)} y^{(s)}. \tag{2.17}$$

In addition $\mathbf{f}\mathbf{f}^t$ is replaced by $F^{+t}\mathbf{f}\mathbf{f}^t F^+$ in the expression of (2.12,2.13) of the train and test errors, where F^+ denotes the pseudo-inverse of F. Since F has a rank smaller than D we decompose the input in $\mathbf{x} = \mathbf{x}^{\|} + \mathbf{x}^{\perp}$, where $\mathbf{x}^{\perp}$ is the component of $\mathbf{x}$ in the kernel of F s.t. $F\mathbf{x} = F\mathbf{x}^{\|}$. The part $\mathbf{f}^t\mathbf{x}^{\perp}$ of the signal cannot be recovered by the model which is now misspecified. As discussed in [7], this part can actually be considered as an additional contribution to the noise ϵ as soon as $\mathbf{x}^{\perp}$ is decorrelated from $\mathbf{x}^{\|}$, the noise variance being now changed to

$$\sigma_{\text{eff}}^2 \stackrel{\text{def}}{=} \sigma^2 + \mathbb{E}\big[\|\mathbf{f}^t\mathbf{x}^{\perp}\|^2\big]. \tag{2.18}$$

This may change the traditional picture of a U-shape curve when increasing the number of parameters in the model and be responsible in some cases, by decreasing the effective noise level as some particular rate, for the double descent scenario observed in [5]. In summary the train and test error of the misspecified ridge

regression problem are given by

$$E_{\text{train}}^{(N)}(\alpha) = \frac{1}{\alpha}\text{Tr}\Big[G_\alpha^{(N)}(\mathbb{I} - G_\alpha^{(N)})F^{+t}\mathbf{f}\mathbf{f}^t F^+\Big] + \sigma_{\text{eff}}^2\Big(1 - \rho^{-1} + \frac{1}{N}\text{Tr}\Big[G_\alpha^{(N)\,2}\Big]\Big),$$
$$(2.19)$$

$$E_{\text{test}}^{(N)}(\alpha) = \text{Tr}\Big[G_\alpha^{(N)}C G_\alpha^{(N)}F^{+t}\mathbf{f}\mathbf{f}^t F^+\Big] + \sigma_{\text{eff}}^2\Big(1 + \frac{\alpha}{N}\text{Tr}\Big[G_\alpha^{(N)}(\mathbb{I} - G_\alpha^{(N)})C\Big]\Big). \quad (2.20)$$

2.3 Kernel learning, Gaussian processes and Kernel regime

In order to avoid the reader to get lost with the ML jargon already used at some places in particular in the introduction, we need to introduce closely related learning schemes and models which lead in the end to very similar analysis as the one developed for the ridge regression. The first one is kernel learning (see e.g. [40]). Based on a set of observations $\{\mathbf{x}^{(s)}, y^{(s)}, s = 1, \ldots N\}$ it consists to define a prediction model for unseen data $y = f(\mathbf{x})$ as a combination of training set observations:

$$f(\mathbf{x}) = \sum_{s=1}^{N} K\big(\mathbf{x}, \mathbf{x}^{(s)}\big)y^{(s)}, \qquad (2.21)$$

where K is a well choosen kernel. Then to make connection with the linear regression model considered previously, consider the case where f is sought in the form of a linear combination of N_f basic functions $\{\phi_k, k = 1, \ldots N_f\}$, namely the features functions:

$$f(\mathbf{x}) = \sum_{k=1}^{N_f} w_k \phi_k(\mathbf{x}),$$

where $\mathbf{w}$ is as before the vector of regression coefficients. The solution (2.11) can be made in correspondence with the kernel regression by simply identifying K with

$$K(\mathbf{x}, \mathbf{x}') = \frac{\alpha}{N}\Phi^t(\mathbf{x})\Big[\mathbb{I} + \frac{\alpha}{N}\sum_{s=1}^{N}\Phi\big(\mathbf{x}^{(s)}\big)\Phi^t\big(\mathbf{x}^{(s)}\big)\Big]^{-1}\Phi(\mathbf{x}') \qquad (2.22)$$

where $\Phi(\mathbf{x}) = (\phi_k(\mathbf{x}), k = 1, \ldots N_f)$ is the feature vector. This solution can be obtained as the mean of a Gaussian process which corresponds to a special case of a probabilistic, Bayesian approach to learning. In this setting f is a random function which distribution called posterior, is given by

$$P[f] = \frac{1}{Z}P_{\text{prior}}[f]e^{-N\mathcal{L}[f]}$$

where P_{prior} is the prior distribution, which happens to be Gaussian for Gaussian processes, $\mathcal{L}[f]$ is the loss term like e.g. (2.2) attached to the training data, while Z is the normalization constant. Hence if f is assumed to decompose on the set of feature $\phi_k(\mathbf{x})$ previously considered, a Gaussian prior distribution on the weights $\mathbf{w}$ would therefore be equivalent as to solving the ridge regression. Gaussian processes when considered as a Gaussian field theory can be analyzed with the ordinary tools of field theory [41, 22, 27]. For instance, the NTK mentioned in Section 2.1 appears

to be the 2-points function of the prior distribution as can be verified by identifying features $\phi_k(\mathbf{x})$ as

$$\phi_k(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\partial}{\partial \theta_k} f_\theta(\mathbf{x})$$

and the adequate Gaussian field theory having f specified by (2.21,2.22) as the ground state.

3   Spectral decomposition of ridge regression errors

3.1   Test-train error ratio for the ridge regression

At least for simple machine learning models it is often possible to obtain relation between the train and test error, which for instance leads to the notion of stability explored in [42], by bounding the difference between the test and train error typically. Here on the specific case of the ridge regression, we obtain a different type of relation, namely a ratio, between test and train error, first when evaluated on a given sample. The idea is to compare the solution (2.11) when the considered sample is incorporated or not in the train set.

We will consider from now on the misspecified model (2.14). The solution obtained with N training samples is given by (2.11). When one training data point is added we have the following recurrence:

$$Z^{(N+1)} = \frac{N}{N+1} Z^{(N)} + \frac{1}{N+1} F\mathbf{x}^{(N+1)} y^{(N+1)}$$

$$G_\alpha^{(N+1)} = G_{\alpha'}^{(N)} - \gamma^{(N)}(\alpha') G_{\alpha'}^{(N)} F\mathbf{x}^{(N+1)} \mathbf{x}^{(N+1)t} F^t G_{\alpha'}^{(N)},$$

thanks to the Sherman-Morrison formula, with $\alpha' = \frac{N}{N+1}\alpha$ and

$$\gamma^{(N)}(\alpha) \stackrel{\text{def}}{=} \frac{\alpha}{N + \alpha \mathbf{x}^{(N+1)t} F^t G_\alpha^{(N)} F\mathbf{x}^{(N+1)}}.$$

As a result we can compare $\hat{\mathbf{w}}_\alpha^{(N)}$ with $\hat{\mathbf{w}}_\alpha^{(N+1)}$. Taking the transpose, we have

$$\hat{\mathbf{w}}_\alpha^{(N+1)t} = \hat{\mathbf{w}}_{\alpha'}^{(N)t} + \frac{\alpha' \mathbf{x}^{(N+1)t} F^t G_{\alpha'}^{(N)}}{N + \alpha' \mathbf{x}^{(N+1)t} F^t G_{\alpha'}^{(N)} F\mathbf{x}^{(N+1)}} \left[y^{(N+1)} - \hat{\mathbf{w}}_{\alpha'}^{(N)t} F\mathbf{x}^{(N+1)} \right].$$

Multiplying right by $F\mathbf{x}^{(N+1)}$ and substracting $\mathbf{f}^t \mathbf{x}^{(N+1)}$ on both side of the equation we finally end up with the relation

$$\hat{\mathbf{w}}_{\alpha'}^{(N)t} F\mathbf{x}^{(N+1)} - y^{(N+1)} = \mathcal{R}^{(N)}(\alpha', \mathbf{x}^{(N+1)}) \left[\hat{\mathbf{w}}_\alpha^{(N+1)t} F\mathbf{x}^{(N+1)} - y^{(N+1)} \right] \qquad (3.1)$$

with

$$\mathcal{R}^{(N)}(\alpha, \mathbf{x}) \stackrel{\text{def}}{=} 1 + \frac{\alpha}{N} \mathbf{x}^t F^t G_\alpha^{(N)} F\mathbf{x}.$$

This quantity, greater than one, represents a test to train absolute error ratio, since on the l.h.s of (3.1) the difference of the model with the observation is evaluated on $\mathbf{x}^{(N+1)}$ not present in the train set while on the r.h.s $\mathbf{x}^{(N+1)}$ is used to train the

model $\hat{\mathbf{w}}^{(N+1)}$. In term of squared error the sample wise relationship between the test and train reads

$$E_{\text{test}}^{(N)}(\alpha', \mathbf{x}) = \mathcal{R}^{(N)}(\alpha', \mathbf{x})^2 E_{\text{train}}^{(N)}(\alpha, \mathbf{x}). \tag{3.2}$$

which has to be average w.r.t. $\mathbf{x}$ to get the generalization error

$$E_{\text{test}}^{(N)} = \mathbb{E}_{\mathbf{x}}\Big[E_{\text{test}}^{(N)}(\alpha, \mathbf{x}) \Big],$$

Incidentally we found a similar statement made already some time ago in [23] in the context of Gaussian processes, resulting however from a quite remote approach based on variational approximation to Bayesian learning. As we shall see, $\mathcal{R}^{(N)}(\alpha, \mathbf{x})$ for large N is actually a self-averaging quantity, and since α and α' differ only by a $\mathcal{O}(1/N)$ quantity, it constitutes the actual test to train error ratio.

3.2   Asymptotic limits

Equations (2.19,2.20) have been analyzed in the asymptotic limits by various authors recently [6, 13, 12, 9], using RMT when $N, N_f \to \infty$ with fixed $\rho = N/N_f$. Let us first discuss informally these results. The main difficulty resides in the first term of the l.h.s of equation (2.20) which involves three matrices, the resolvent $G^{(N)}$ (from now on we drop in the notation the explicit dependency on α), the population matrix C and the representation of the signal matrix $\tilde{\mathbf{f}}\tilde{\mathbf{f}}^t \overset{\text{def}}{=} F^+ \tilde{\mathbf{f}}\tilde{\mathbf{f}}^t F^{+t}$ in feature space, with non mutually aligned sets of eigenvectors. For instance, given the spectral density ν_∞ associated to C and h an arbitrary function, results of Ledoit-Péchet [43] allows one to get the following asymptotic limits:

$$\lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \frac{1}{N_f} \text{Tr}\Big[G^{(N)} h(C) \Big] = \int dx \frac{\nu_\infty(x) h(x)}{1 + \alpha x \big[1 - \rho^{-1} + \rho^{-1} g(\rho) \big]}$$

with

$$g(\rho) \overset{\text{def}}{=} \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \frac{1}{N_f} \text{Tr}\Big[G^{(N)} \Big],$$

solution to the basic self-consistent equation of Marchenko-Pastur [44]

$$g(\rho) = \int dx \frac{\nu_\infty(x)}{1 + \alpha x \big[1 - \rho^{-1} + \rho^{-1} g(\rho) \big]},$$

assuming no specific form of the population matrix. Instead, the combined term of interest in equation (2.20) cannot be obtained in general by the Ledoit-Péchet formula, except for some special cases as in [6] where $\tilde{\mathbf{f}}\tilde{\mathbf{f}}^t$ is assumed isotropic or as in [13, 12] where $\tilde{\mathbf{f}}\tilde{\mathbf{f}}^t$ is assumed to be diagonal on the bases of C eigenvectors, while in [9] non-asymptotic expression of the terms in (2.19) are obtained with weaker assumptions. In all these works, the expressions which are given are quite involved and difficult to interpret from the physics point of view. Here we propose to take advantage of the relation (3.2) and argue that it can be transposed at the asymptotic level in the form

$$E_{\text{test}}(\rho, \alpha) = \mathcal{R}^2(\rho, \alpha) E_{\text{train}}(\rho, \alpha) \tag{3.3}$$

with

$$E_{\text{test,train}}(\rho, \alpha) \stackrel{\text{def}}{=} \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \mathbb{E}_{\mathbf{x}} \left[E_{\text{test,train}}^{(N)}(\alpha, \mathbf{x}) \right]$$

$$\mathcal{R}(\rho, \alpha) \stackrel{\text{def}}{=} \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \mathbb{E}_{\mathbf{x}} \left[\mathcal{R}^{(N)}(\alpha, \mathbf{x}) \right] = 1 + \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \left[\frac{\alpha}{N} \text{Tr} \left(G_\alpha^{(N)} C \right) \right].$$

The argument here is already used in the derivation of the Marchenko-Pastur distribution (see Appendix A), namely that some traces are self-averaging when $N \to \infty$. Consider the $D \times D$ matrix $F^t G_\alpha^{(N)} F$, denote by $\{\lambda_k, k = 1, \ldots N_f\}$ its non-zero eigenvalues assumed to be bounded and $\{x_k, k = 1, \ldots D\}$ the components of $\mathbf{x}$ on these modes. We have

$$\frac{1}{N_f} \mathbf{x}^t F^t G_\alpha^{(N)} F \mathbf{x} = \frac{1}{N_f} \sum_{k=1}^{D} x_k^2 \lambda_k = \frac{1}{N_f} \text{Tr} \left[G_\alpha^{(N)} C \right] + \mathcal{O} \left(\frac{1}{N_f} \right),$$

by law of large numbers, given that the x_k are iid with $x_k = \mathcal{N}(0, 1)$. At this point we can directly use the Ledoit-Péchet formula in a similar fashion to what is done in [6] to express the asymptotic limits of $\mathcal{R}(\rho, \alpha)$ and $E_{\text{train}}(\rho, \alpha)$ thereby giving $E_{\text{test}}(\rho, \alpha)$ in (3.3). As explained in Appendix A this can also be obtained thanks to a diagrammatic expansion restricted to a summation over non-crossing diagrams, which corresponds to the combinatorial interpretation of free probabilities [45, 46], which somehow backs all these asymptotic expressions of random matrix theory. The usefulness of the diagrammatic expansion formalism will be made more obvious in the next Section when interpreting the results. Eventually we arrive at the following spectral decomposition of the train error and the test-train error ratio (see Appendix B):

$$\mathcal{R}(\rho, \alpha) = 1 + \frac{\alpha}{\rho} \int dx \nu_\infty(x) x g(x, \rho, \alpha), \tag{3.4}$$

$$E_{\text{train}}(\rho, \alpha) = - \int \frac{dx}{x} \mu^{\|}(x) \frac{\partial}{\partial \alpha} g(x, \rho, \alpha) + \frac{\sigma^2 + \mu^\perp}{\rho} \left(\rho - 1 + \frac{\partial}{\partial \alpha} \left[\alpha g(\rho, \alpha) \right] \right). \tag{3.5}$$

with

$$g(\rho, \alpha) = \int dx \nu_\infty(x) g(x, \rho, \alpha),$$

$$g(x, \rho, \alpha) = \left(1 + \frac{\alpha x}{\mathcal{R}(\rho, \alpha)} \right)^{-1}.$$

Here, in addition to ν_∞, the spectral density of the population matrix C hence associated to the features, appear also spectral quantities associated to the signal $\mu^{\|}(x)$ and $\mu^\perp$. Let us define them properly in thermodynamic limits. Consider first the singular value decomposition (SVD) of the feature matrix:

$$F = \sum_{a=1}^{N_f} \sqrt{c_a} \mathbf{u}_a \mathbf{v}_a^t \tag{3.6}$$

where the right singular vectors $\{\mathbf{v}_a, a = 1, \dots D\}$ are arbitrarily completed for $a = N_f, \dots D$ to span the entire embedding space. The population matrix C reads then

$$C = \sum_{a=1}^{N_f} c_a \mathbf{u}_a \mathbf{u}_a^t,$$

where $\{\mathbf{u}_a, a = 1, \dots N_f\}$ and $\{c_a, a = 1, \dots N_f\}$ are the eigenvectors and eigenvalues of C at finite N_f. In thermodynamic limit we have

$$\nu_\infty(x) \overset{\text{def}}{=} \lim_{N_f \to \infty} \frac{1}{N_f} \sum_{a=1}^{N_f} \delta(x - c_a).$$

The signal itself is represented by $\mathbf{f}$ defined on the embedding space $\mathbb{R}^D$. On this bases we have the decomposition

$$\mathbf{f} = \mathbf{f}^\parallel + \mathbf{f}^\perp = \sum_{a=1}^{N_f} f_a \mathbf{v}_a + \sum_{a=N_f+1}^{D} f_a \mathbf{v}_a.$$

Then

$$\mu^\parallel(x) \overset{\text{def}}{=} \lim_{N_f \to \infty} \sum_{a=1}^{N_f} f_a^2 \delta(x - c_a). \tag{3.7}$$

It is normalized to

$$\mu^\parallel \overset{\text{def}}{=} \sum_{a=1}^{N_f} f_a^2,$$

while

$$\mu^\perp \overset{\text{def}}{=} \sum_{a=N_f+1}^{D} f_a^2$$

represents the squared norm of the orthogonal part of the signal to the features. The effective variance of the noise defined previously in (2.18) reads in these notations $\sigma_{\text{eff}}^2 = \sigma^2 + \mu^\perp$.

3.3 Effective coupling constant and spectral parameters

In the curse of derivation of these equations via the diagrammatic expansion, we obtain a closed form expression of the self energy involved in the Dyson equation thanks to planar diagram approximation (see Appendix A). In itself the self-energy, also called $\mathcal{R}$-transform in free probabilities, represents a kind of dressing of the coupling constant α in the present expansion which suggests to define the following effective coupling constant (see Appendix B)

$$\Lambda(\rho, \alpha) \overset{\text{def}}{=} \frac{\alpha}{\mathcal{R}(\rho, \alpha)} \leq \alpha, \tag{3.8}$$

as the inverse of the test-train ratio. Incidentally we found out in [47] that this quantity has a counterpart called the resolvent bias factor in the probabilistic/statistic literature. We found also a similar definition in [48] of an effective ridge penalty in

the context of kernel regression, but resulting from different considerations. Let us introduce two sets of statistical parameters:

$$g_k[\Lambda] \overset{\text{def}}{=} \int dx \, \frac{\nu_\infty(x)}{[1 + \Lambda x]^k}, \qquad (3.9)$$

$$\mu_k[\Lambda] \overset{\text{def}}{=} \frac{1}{\mu^\|} \int dx \frac{\mu^\|(x)}{\left[1 + \Lambda x\right]^k}, \qquad (3.10)$$

the former being associated to the feature matrix while the latter is associated to the signal. These will be of use later on. They represent the moments of the variable $y = 1/(1 + \Lambda x)$ when x is respectively drawn from ν_∞ and from $\mu^\|(x)/\mu^\|$. For the moment let us present some properties concerning $\Lambda(\rho, \alpha) = \Lambda$ (the explicit dependency of Λ w.r.t. α and ρ will be omitted from now on). From (3.4) we see that Λ is solution of the self-consistent equation

$$\Lambda = \alpha\left(1 - \frac{1}{\rho}\right) + \frac{\alpha}{\rho} \int dx \frac{\nu_\infty(x)}{1 + \Lambda x}$$

$$= \alpha\left(1 - \frac{1}{\rho}\right) + \frac{\alpha}{\rho} g[\Lambda] \overset{\text{def}}{=} h_{\rho,\alpha}(\Lambda), \qquad (3.11)$$

where $g \, (= g_1)$ now reads in terms of this effective coupling,

$$g[\Lambda] = \int dx \, \frac{\nu_\infty(x)}{1 + \Lambda x},$$

therefore $h_{\rho,\alpha}$ is a strictly decreasing convex function of Λ with $h_{\rho,\alpha}(0) = \alpha > 0$. As a result equation (3.11) admits a single solution $\Lambda(\rho, \alpha) \in [0, \Lambda_{\max}(\rho, \alpha)[$ with

$$\Lambda_{\max}(\rho, \alpha) = \begin{cases} \alpha & \rho > 1, \\ \min(\alpha, g^{-1}[1 - \rho]) & \rho \le 1. \end{cases} \qquad (3.12)$$

In practice this solution can be rapidly obtained by iteration of $h_{\rho,\alpha}(\Lambda)$ or by dichotomic search in absence of convergence of the fixed point equation. Notice also that $g[\Lambda]$ is strictly decreasing, hence $g[\Lambda] \in]g_{\min}(\rho), 1]$ with

$$g_{\min}(\rho) \overset{\text{def}}{=} \max(0, 1 - \rho)$$

Using the fixed point equation (3.11) we get the following expressions for the derivative of Λ w.r.t. α:

$$\frac{\partial}{\partial \alpha} \Lambda(\rho, \alpha) = \frac{\Lambda^2}{\alpha^2} \frac{\rho}{\rho - \mathcal{Q}[\Lambda]}, \qquad (3.13)$$

with

$$\mathcal{Q}[\Lambda] \overset{\text{def}}{=} \int dx \, \nu_\infty(x) \frac{\Lambda^2 x^2}{\left[1 + \Lambda x\right]^2}, \qquad (3.14)$$

$$= 1 - 2g[\Lambda] + g_2[\Lambda], \qquad (3.15)$$

$$= \rho + \Lambda\left(g'[\Lambda] - \frac{\rho}{\alpha}\right).$$

This last equality combined to (3.11) allows one to rewrite (3.13) in the form

$$\frac{\partial}{\partial \alpha} \Lambda(\rho, \alpha) = \frac{\Lambda}{\alpha} \frac{\rho}{\rho - \alpha g'[\Lambda]} > 0, \tag{3.16}$$

the inequality following the fact that $g'[\Lambda] < 0$. Hence we get here the important property that the effective coupling Λ is a strictly increasing function of the bare coupling α, meaning that Λ can replace α to parameterize the ridge regression. We are now in position to simplify the asymptotic expression of the test error given in the preceding section. First we introduce the fraction r of the signal that can be possibly recovered

$$r = \frac{\mu^{\parallel}}{\mu^{\parallel} + \mu^{\perp} + \sigma^2} = \frac{SNR}{1 + SNR} \tag{3.17}$$

where

$$SNR \overset{\text{def}}{=} \frac{\mu^{\parallel}}{\mu^{\perp} + \sigma^2}$$

is the effective signal to noise ratio, from the definition (2.18) of the effective noise variance. In order to have a meaningful comparison between various situation we normalize the errors by $\mu^{\parallel} + \mu^{\perp} + \sigma^2$, or equivalently assume that the observation y in (2.7) is normalized s.t.

$$\mu^{\parallel} + \mu^{\perp} + \sigma^2 = 1. \tag{3.18}$$

We have

$$\frac{\partial}{\partial \alpha} g(x, \rho, \alpha) = -\frac{1}{\rho} \frac{(\rho + g[\Lambda] - 1)^2}{\rho - \mathcal{Q}[\Lambda]} \frac{x}{(1 + \Lambda x)^2}$$

$$\frac{\partial}{\partial \alpha} \big[\alpha g(\rho, \alpha)\big] = g[\Lambda] - \frac{\rho - 1 + g[\Lambda]}{\rho - \mathcal{Q}[\Lambda]} \big[1 - g[\Lambda] - \mathcal{Q}[\Lambda]\big],$$

with the shorthand notation $\Lambda = \Lambda(\rho, \alpha)$. Combining this with (3.11) in (3.4,3.5) and arranging the various terms leads us to our final expression for the (normalized) test error

$$E_{\text{test}}[\rho, r, \Lambda] = \frac{\rho}{\rho - \mathcal{Q}[\Lambda]} \big(1 - r + r\mu_2[\Lambda]\big) \qquad \in [1 - r, 1] \tag{3.19}$$

expressed now as a function of ρ, r and Λ. The contributions corresponding respectively to the bias and variance are then identified as

$$\mathcal{B}[\rho, r, \Lambda] = \frac{\rho}{\rho - \mathcal{Q}[\Lambda]} r\mu_2[\Lambda], \tag{3.20}$$

$$\mathcal{V}[\rho, r, \Lambda] = \frac{\rho}{\rho - \mathcal{Q}[\Lambda]} (1 - r). \tag{3.21}$$

The expressions which appeared recently in statistics [6, 9] or ML literature [13, 12], are either less general or/and maybe more obscure to us. At least we could check that ours coincide with that of [6] when restricting the signal's distribution to an isotropic one. Then looking at (3.19) we immediately see two counter effects from the regularization to be at work. One comes from the features and the other
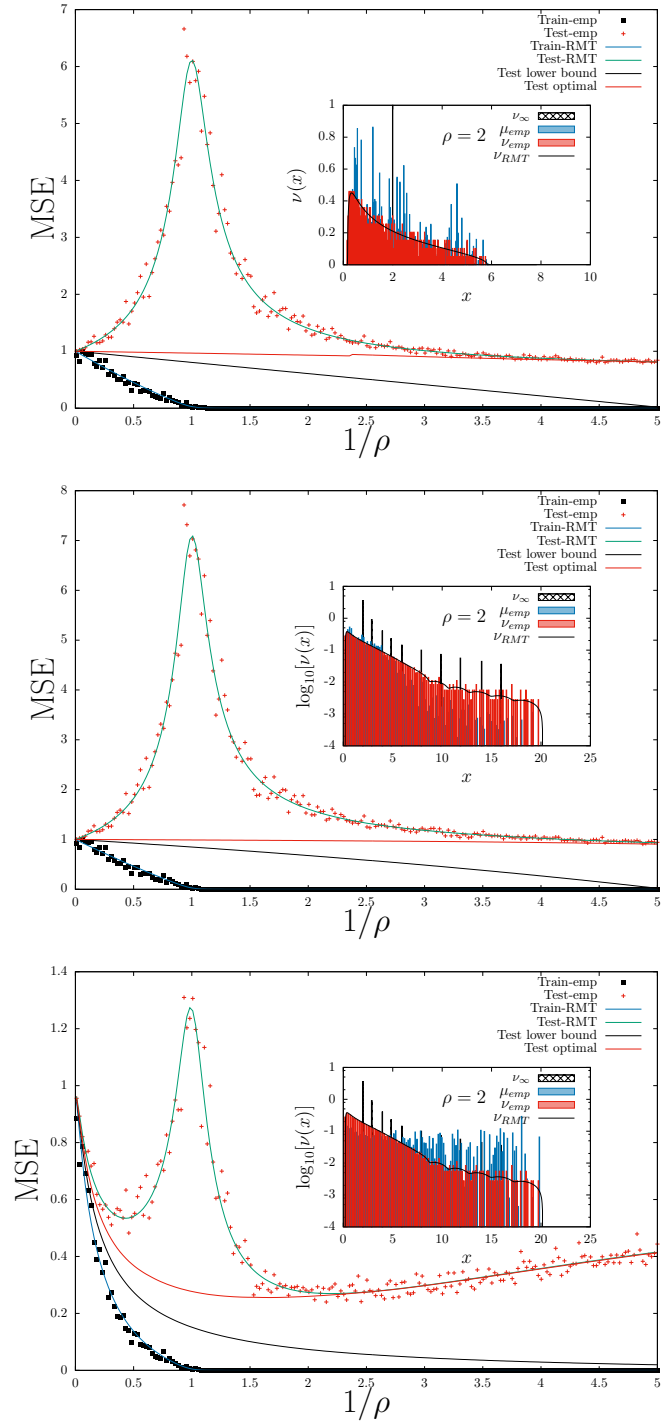
Fig. 3.2: Main plots: Test and train error as a function of ρ^{-1} given by RMT (plain lines) and a single experiments with $N = 200$, when the population matrix has one degenerate level (top panel) and ten degenerate levels, with wrong alignment (middle panel) and good alignment (bottom panel) of the signal with the population matrix. Inset: spectral densities of the population (in gray), empirical feature matrices (in red) and spectral power of the signal on the corresponding levels (blue).

one from the signal and are simply expressed by the spectral coefficients (3.14) and (3.10) with $k = 2$. When the coupling is increased,

$$\mathcal{Q}'[\Lambda] = 2\big(g[\Lambda] + g_3[\Lambda] - 2g_2[\Lambda]\big) = 2 \int dx \nu_\infty(x) \frac{\Lambda x^2}{(1 + \Lambda x)^3} > 0,$$

so $\mathcal{Q}$ increases, while μ_2 obviously decreases, inducing the two counter effects on E_{test}; the first contributes to some increase, the latter to a decrease, and the optimal tuning of the penalty corresponds to the best trade-off between these two. We also immediately see, as expected, that E_{test} is lower bounded by $1 - r$ ($\mu_2 = \mathcal{Q} = 0$). In order to get as close as possible from this lower bound, as stated in [13], a favorable situation corresponds to have a good alignment of the population matrix with the signal, meaning that the signal decomposes preferentially on the stronger modes of the population matrix. Equation (3.19) allows us to have more precise indications on the necessary ingredients for the ideal situation concerning generalization. Ideally we want the signal to decompose mainly on a small fraction ν_s of strong modes of the population matrix, with $\nu_s \ll \rho$ separated by a large gap from the other modes. This situation offers then a good choice for the ridge penalty which is to select Λ^{-1} somewhere in the middle of the gap. Doing that insures on one side that $\Lambda x \ll 1$ on the weak modes and on the other side that $\Lambda x \gg 1$ on the strong modes; in turn this insures that $\mathcal{Q} \sim \nu_s \ll \rho$ and $\mu_2 \ll 1$. This mechanism is illustrated on Figure 3.2 where we also see that up to fluctuations due mainly to the finite number $N_s = 200$ of train samples (the number of test samples $> 10^4$ is chosen sufficiently large to have no influence on the variance of the results), the test and train error are already properly estimated from RMT results given above. An example of a double descent is seen on the bottom panel, but as shown from the curve corresponding to optimal penalty this is an artifact of a wrongly tuned regularization, which is kept constant while varying ρ. In the next sections of the paper we will exploit this formulation which will appear very convenient when discussing the feature learning process.

3.4   Practical aspects from empirical estimations

Since (3.19) is written as function of ρ, r and Λ, finding the optimal ridge penalty for the test error is equivalent to optimize E_{test} w.r.t. Λ at fixed ρ and r. Doing this yields the following condition on Λ:

$$E_{\text{test}}[\rho, r, \Lambda] \int dx \nu_\infty(x) \frac{\Lambda^2 x^2}{\left[1 + \Lambda x\right]^3} - \frac{\rho r}{\mu^\|} \int dx \mu^\|(x) \frac{\Lambda x}{\left[1 + \Lambda x\right]^3} = 0. \qquad (3.22)$$

It is not clear yet how to solve this equation in practice. Just notice here that the previous condition gives us a way to parameterize the optimal $E_{\text{test}}^\star$ in terms of the optimal value $\Lambda^\star$ of the effective coupling as

$$E_{\text{test}}^\star[\rho, r] = \rho r \, \frac{\mu_2[\Lambda^\star] - \mu_3[\Lambda^\star]}{g[\Lambda^\star] + g_3[\Lambda^\star] - 2g_2[\Lambda^\star]}$$

written in terms of the spectral coefficients defined in (3.9,3.10) when $\Lambda^\star$ is finite. Yet there are considerations of more practical interest that can be deduced from

the asymptotic theory in complement of some others recently proposed in [47] for instance. First from (3.3,3.8) and (3.11) we see that the estimation

$$g[\Lambda] \approx \hat{g}[\alpha] \overset{\text{def}}{=} \frac{1}{N_f} \text{Tr}\Big[\frac{1}{\mathbb{I}_{\mathcal{F}} + \alpha C^{(N)}}\Big],$$

directly made on the training data gives us an estimation of the effective coupling

$$\Lambda \approx \hat{\Lambda} = \frac{\alpha}{\rho}\big(\rho - 1 + \hat{g}[\alpha]\big)$$

and hence an estimation of the train-test error ratio, so that finally from the train error we can get a prediction of the test error without resorting to cross-validation:

$$\hat{E}_{\text{test}} = \frac{\rho^2}{\big(\rho - 1 + \hat{g}[\alpha]\big)^2} E_{\text{train}}. \tag{3.23}$$

Again, a similar proposal is made in [23] to exploit asymptotic relations between test and train error in the context of Gaussian processes. As seen on the plot of Figure 3.3, a good correlation between this estimation and the actual test error is observed, with a variance increasing with errors.

Nevertheless it is interesting to realize that other statistics concerning the hidden prior, namely the population matrix, are accessible via other estimators in this asymptotic regime. For instance the coefficient g_2 can be estimated as well. Let

$$\hat{g}_2(\alpha) \overset{\text{def}}{=} \frac{1}{N_f} \text{Tr}\Big[\frac{1}{\big(\mathbb{I}_{\mathcal{F}} + \alpha C^{(N)}\big)^2}\Big].$$

On the one hand we have

$$\frac{d}{d\alpha}\hat{g}(\alpha) = \frac{1}{\alpha}\big[\hat{g}_2(\alpha) - \hat{g}(\alpha)\big].$$

On the other hand

$$\frac{d}{d\alpha}g[\Lambda] = \frac{d\Lambda}{d\alpha}g'[\Lambda] = \frac{1}{\Lambda}\frac{d\Lambda}{d\alpha}\big(g_2[\Lambda] - g[\Lambda]\big).$$

Identifying the two derivatives and making use of (3.11,3.13), we can rearrange things to obtain the following estimator for g_2:

$$g_2[\Lambda] \approx \hat{g} + \frac{\hat{g}_2 - \hat{g}}{1 + \frac{\hat{g}_2 - \hat{g}}{\rho - 1 + \hat{g}}}.$$

In principle we should be able to carry on this program for $g_k, k > 2$ which eventually is equivalent to trying to solve the inverse problem of the population matrix given the sample covariance matrix [49]. Typically, if C has K distinct levels, we have to go until $k = K$ to be able to resolve all its levels. For instance, the 2-level problem considered in the next Section 3.5 has two unknowns regarding the population matrix, namely ν the degeneracy of the level associated to the strong modes and Δ the gap between the two levels. Obviously the knowledge of g and g_2 is enough to find these two unknowns. While this is an interesting direction relevant to practical ML applications, it is out of the scope of the present paper and we leave it aside for future works.
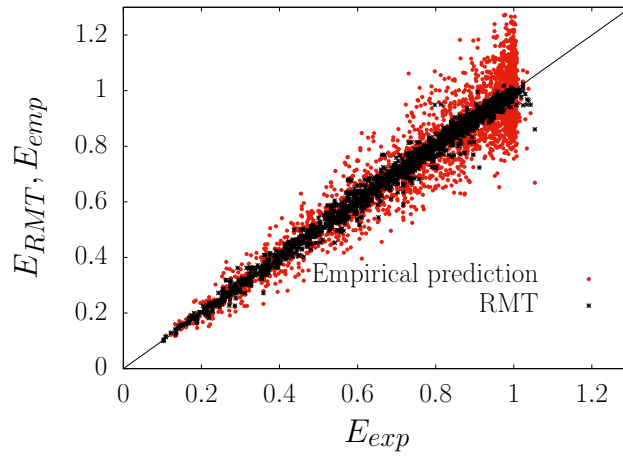
Fig. 3.3: Scatter plot of the test error E_{exp} measured on experiments done with a two 2-level population matrix model with eigenvalues $\{1, 50\}$, versus the expected results (3.19) from RMT and the empirical prediction (3.23) obtained from the training. The experimental point are obtained with a train set of size $N_s = 200$ and test set of size 10^4. By convenience, the default coupling is set to $\alpha = \alpha^\star$ delivered by the RMT, ρ is varied $\in \{0.5, 1.0, 2.0\}$, $\nu \in \{0.01, 0.99\}$ and $\mu \in [0, 1]$.

Concerning the signal itself, it is also possible to extract information. Similarly we would like to be able to estimate the coefficients μ_k, at least for small k, but also $\mu^{\|,\perp}$. First the mean-squared norm of the observed signal is given empirically by

$$\mu^{\|} + \sigma_{\text{eff}}^2 \approx \frac{1}{N} \sum_{s=1}^{N_s} y^{(s)\,2}. \tag{3.24}$$

Then the train error normalized by this quantity is already a source of information since its expression in terms of μ_k coefficients actually read

$$E_{\text{train}} = \frac{\rho \Lambda^2}{\alpha^2 (\rho - \mathcal{Q}[\Lambda])} \big[r\mu_2 + (1-r) \big] \tag{3.25}$$

where here Λ and $\mathcal{Q}[\Lambda] = 1 - 2g[\Lambda] + g_2[\Lambda]$ can be estimated empirically from the preceding remarks, hence the unknown are r and μ_2. Another interesting quantity is the squared norm $\|\mathbf{w}\|^2$ of the regression vector (2.11),

$$\mathbf{w}^t \mathbf{w} = \mathbf{f}^t F^+ \big(\mathbb{I} - G^{(N)} \big)^2 F^{+t} \mathbf{f} + \frac{\alpha \sigma_{\text{eff}}^2}{N} \text{Tr}\Big[G^{(N)} \big(\mathbb{I} - G^{(N)} \big) \Big],$$

which when normalized by (3.24) is given asymptotically by

$$\mathbf{w}^t \mathbf{w} = r\Lambda \Big[\mu_1 - \frac{\rho \Lambda}{\alpha (\rho - \mathcal{Q}[\Lambda])} \mu_2 \Big] + (1-r)\Lambda \Big[1 - \frac{\rho \Lambda}{\alpha (\rho - \mathcal{Q}[\Lambda])} \Big].$$

Combining these two expressions yields the asymptotic expression of the normalized loss:

$$\mathcal{L} = \frac{1}{2} \big(\mathbf{w}^t \mathbf{w} + \alpha E_{\text{train}} \big) = \frac{\Lambda}{2} \big[1 + r(\mu_1 - 1) \big], \tag{3.26}$$

which will be used to determine the dynamics in Section 4. Hence combining the empirical measurement of $\mathbf{w}^t \mathbf{w}$ and of E_{train} allows us to disentangle μ_1 from μ_2 when $r = 1$, i.e. in absence of effective noise. In this case, derivation w.r.t. α would let us obtain empirical estimation of higher order $\mu_k, k > 1$ coefficients. When noise is present or whenever the model is misspecified a possible way to proceed, but not explored here, would consist in probing the system by adding some noise and extract empirically the rate of dependence of $\mathcal{L}$ to that noise, namely $\Lambda(1 - \mu_1)/2$ as seen from (3.26).

3.5   Special cases

To illustrate the preceding statements let us consider some special cases.

The one-level case: it corresponds to the fully degenerate case, when the population matrix $C = \mathbb{I}$, i.e. when

$$\nu_\infty(x) = \delta(x - 1).$$

This case corresponds for instance to the functional case where the features are normalized Fourier modes along with a uniform spatial distribution of input sample points. The relative longitudinal spectral density of the signal is then necessary of the form
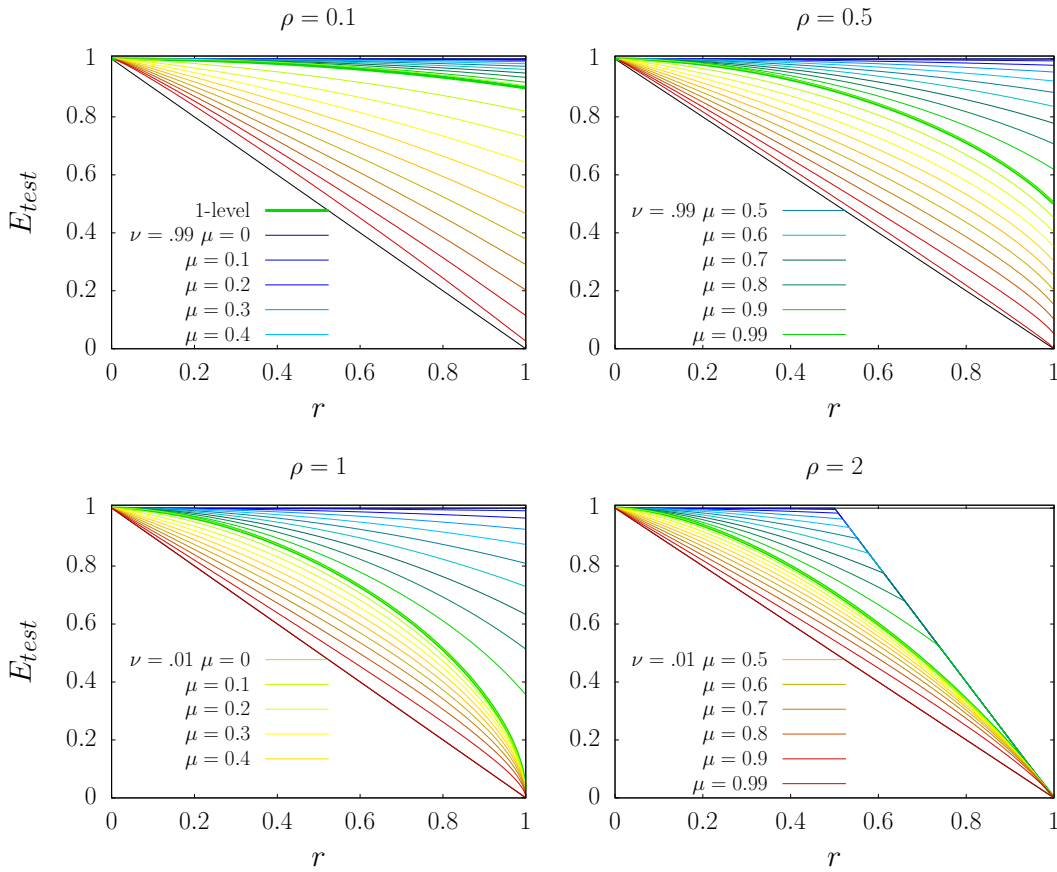
$$\mu^{\|}(x) = \mu^{\|} \delta(x - 1),$$

Fig. 3.4: The optimal performance in terms of E_{test} of the 2-level model as a function of r, depending on the fraction ν of strong modes and on the fraction μ of longitudinal spectral power of the signal on these modes. The performance of the 1-level model defines the frontier between the aligned and anti-aligned domains, represented here respectively by the $\nu = 0.01$ and $\nu = 0.99$ set of curves.

so that we have the form $g_k = \mu_k = y^k$ with $y \in [\max(0, 1 - \rho), 1]$ yielding

$$E_{\text{test}} = \frac{\rho}{\rho - (1 - y)^2} \left[1 + r(y^2 - 1) \right]. \tag{3.27}$$

The optimization can be done here directly on y (which is a monotonic function of Λ) and yields

$$y^\star = \frac{1}{2r} \left[2r + \sqrt{(1 + \rho r)^2 - 4\rho r^2} - 1 - \rho r \right], \tag{3.28}$$

giving the optimal test error as a function of r in the form

$$E_{\text{test}}^\star[\rho, r] = \rho r \frac{2r + \sqrt{(1 + \rho r)^2 - 4\rho r^2} - 1 - \rho r}{1 + \rho r - \sqrt{(1 + \rho r)^2 - 4\rho r^2}},$$

plotted on Figure 3.4.

Two-level case: it corresponds now to the case (also considered in [12]) were the spectrum of the population matrix has two levels representing weak and strong modes. Note that the problem is invariant when multiplying the ridge penalty and the population matrix by a common factor, so in great generality for this case we can consider a spectral density of the form

$$\nu_\infty(x) = (1 - \nu)\delta(x - 1) + \nu\delta(x - c),$$

with $c > 1$. $\nu \in [0, 1]$ represents the fraction of strong modes.

Then the relative spectral density of the signal takes the form

$$\mu^\parallel(x) = \mu^\parallel \left[(1 - \mu)\delta(x - 1) + \mu\delta(x - c) \right],$$

where $\mu \in [0, 1]$ represents the weight of the signal on the strong modes. Given this we obtain a test error given by (3.19) with now

$$\mu_2[\Lambda] = \frac{1 - \mu}{\left[1 + \Lambda \right]^2} + \frac{\mu}{\left[1 + \Lambda c \right]^2},$$

$$\mathcal{Q}[\Lambda] = (1 - \nu)\frac{\Lambda^2}{\left[1 + \Lambda \right]^2} + \nu\frac{\Lambda^2 c^2}{\left[1 + \Lambda c \right]^2}.$$

Looking for the optimal $\Lambda^\star$ amounts to solve (3.22) as a complicated polynomial equation. To simplify let us consider the case of a large gap, i.e. $c \gg 1$. In that case we can distinguish two distinct asymptotic regimes when $c \to \infty$. First from the discussion of the end of Section 3.3 we may expect Λ^{-1} to scale like c, i.e. $\Lambda c = \mathcal{O}(1)$. Letting

$$y \overset{\text{def}}{=} \frac{1}{1 + \Lambda c} \in [0, 1],$$

we obtain

$$\mu_2 = 1 - \mu + \mu y^2 + \mathcal{O}\left(\frac{1}{c}\right),$$

$$\mathcal{Q} = \nu(1 - y)^2 + \mathcal{O}\left(\frac{1}{c^2}\right),$$

and the test error then reads

$$E_{\text{test}} = \frac{\rho}{\rho - \nu(1-y)^2}\big[1 + r\mu(y^2 - 1)\big],$$

which derivative w.r.t. y vanishes at

$$y^\star = \frac{1}{2a}\Big[\sqrt{\big(1 + a\eta\big)^2 + 4a(1-a)} - 1 - a\eta\Big],$$

with $a \overset{\text{def}}{=} r\mu$ and $\eta \overset{\text{def}}{=} \rho/\nu - 2$. The second asymptotic regime corresponds to the situation where $\Lambda c \to \infty$, in which case Λ is not necessarily vanishing. Letting now

$$y \overset{\text{def}}{=} \frac{1}{1 + \Lambda} \in [0, 1],$$

we obtain

$$\mu_2 = (1 - \mu)y^2 + \mathcal{O}\Big(\frac{1}{c}\Big),$$

$$\mathcal{Q} = (1 - \nu)(1 - y)^2 + \nu + \mathcal{O}\Big(\frac{1}{c}\Big),$$

and the test error then reads

$$E_{\text{test}} = \frac{\rho}{\rho - \nu - (1-\nu)(1-y)^2}\big[1 - r + r(1-\mu)y^2\big].$$

This expression actually coincides with the previous one up to a constant factor when doing the change $r \to r/(1 - r\mu)$, $\rho \to \rho - \nu$, $\nu \to 1 - \nu$ and $\mu \to 1 - \mu$, leading to the same $y^\star$ in terms of the modified parameters a and η. Note that in this regime for $\rho < 1$, the constraint $g > 1 - \rho$ becomes $y > \frac{1-\rho}{1-\nu}$, hence for $\nu > \rho$, $y_{\min} > 1$ so this asymptotic regime is irrelevant and only the first one needs to be considered. Otherwise, given the two minimal values of E_{test} obtained, the lowest is selected. The result of this is shown on Figure 3.4, where we see that optimal performance is obtained with the perfect alignment when the signal is concentrated on very few strong modes, i.e. $\nu \to 0$ and $\mu \to 1$. The 1-level case appears to separate the good from bad alignment settings in these plots.

4   Learning dynamics of the population matrix

Now we are in position to analyze the learning dynamics in a more general setting corresponding for instance to neural networks (NN) regressions. Dynamics of neural networks learning has been studied for long time in statistical physics [29, 30]. In particular in [50] is characterized the (non-linear) dynamics of deep linear neural networks, while in [31] dynamical mean field equations are obtained to obtain the stochastic gradient dynamics of 2-layers NN (both linear of non linear) or more recently in [51] where non-linear pathways dynamics is studied. Here we study something close but different in spirit, by considering the architecture sketched on Figure 4.5, where at the first stage, features functions are provided by some arbitrary NN which don't need to be specified for the moment, while the last layer is linear and combine these features to solve a ridge regression. Our goal is to
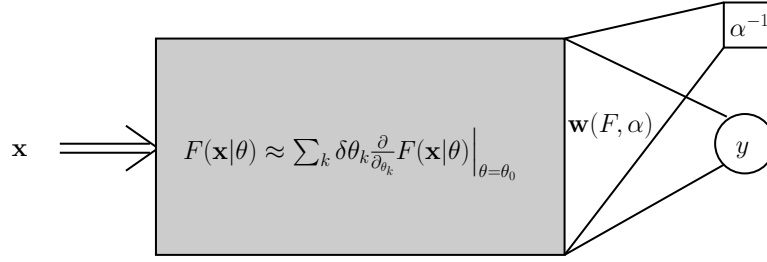
Fig. 4.5:　Schematic plot of the regression model considered here.　The weight vector $\mathbf{w}(F, \alpha)$ of the last layer solves a ridge regression with penalty α^{-1} depending on the vector of features functions $F(\mathbf{x}|\theta)$ defined by some NN with parameters θ represented by the grey box. In the "semi-lazy" regime $F(\mathbf{x}|\theta)$ is taken close to some reference parameter θ_0, with $\delta\theta = \theta - \theta_0$.

specifically investigate how the alignment of the population matrix with the signal, characterized in previous sections, evolves during learning. As a result, the weights of the last layer are given in closed form by (2.11) with explicit dependency on the feature matrix. We study whether the learning process leads to a better alignment of the features with the signal as the one chosen randomly at initialization.

4.1　Microsopic equations of learning

To this end let us reconsider the misspecified regression problem (2.14), with a feature matrix being now a function $F(\mathbf{x}|\theta)$ of the input data $\mathbf{x}$ and of some vector parameter θ of a NN, while the vector $\mathbf{w}$ constitutes the weight vector of the last layer

$$f(\mathbf{x}|\mathbf{w}, \theta) = \mathbf{w}^t F(\mathbf{x}|\theta).$$

We keep the L_2 penalty on the vector $\mathbf{w}$ in the loss function which is there to mimic some form of implicit regularization of the whole architecture:

$$\mathcal{L}(\mathbf{w}, \theta) = \frac{1}{2}\mathbf{w}^t\mathbf{w} + \frac{\alpha}{2N_s}\sum_{s=1}^{N_s}\left[y^{(s)} - \mathbf{w}^t F(\mathbf{x}^{(s)}|\theta)\right]^2. \tag{4.1}$$

In order to study the dynamics of the features under the gradient descent we consider a continuous limit θ_t of the learning process indexed by t and assume that $\mathbf{w} = \mathbf{w}(\theta_t)$ moves "adiabatically" along the gradient descent of θ, meaning that we use (2.11)

$$\mathbf{w}(\theta) = \alpha G^{(N)}(\theta)Z^{(N)}(\theta),$$

with now $G^{(N)}(\theta) = \left[\mathbb{I}_{\mathcal{F}} + \alpha C^{(N)}(\theta)\right]^{-1}$ and

$$C^{(N)}(\theta) \overset{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N} F(\mathbf{x}^{(s)}|\theta)F^t(\mathbf{x}^{(s)}|\theta), \qquad (4.2)$$

$$Z^{(N)}(\theta) \overset{\text{def}}{=} \frac{1}{N}\sum_{s=1}^{N} F(\mathbf{x}^{(s)}|\theta)y^{(s)}. \qquad (4.3)$$

The dynamics (2.4) is modified as ($\gamma = 1$)

$$\dot{\theta}_t = -\nabla_\theta \mathcal{L}\big[\mathbf{w}(\theta_t), \theta_t\big].$$

Next we write the singular value decomposition of the feature matrix as

$$F(\cdot|\theta) = \sum_{a=1}^{N_f} \sqrt{c_a(\theta)}\,\mathbf{u}_a(\theta)v_a(\cdot|\theta),$$

where the v_a which depends on θ, are elements of a functional basis on the input space $\mathbb{R}^d$. This is the counterpart of (3.6) written for a functional space, which dimension D grows potentially exponentially with the dimension d of the base space. We assume them to be orthonormal with respect to the scalar product relative to the prior density $\rho_\infty(\mathbf{x})$ of the data:

$$\int d^d\mathbf{x}\,\rho_\infty(\mathbf{x})v_a(\mathbf{x})v_b(\mathbf{x}) = \delta_{ab}.$$

The hidden function $f(\mathbf{x})$ may be considered indifferently through its vector representation $\mathbf{f} = \{f_a\}$ on this vector space, with

$$f(\mathbf{x}) = \sum_{a=1}^{D} f_a v_a(\mathbf{x}) \qquad \text{and} \qquad f_a = \int d^d\mathbf{x}\,\rho_\infty(\mathbf{x})v_a(\mathbf{x})f(\mathbf{x}).$$

The population matrix obtained by taking $N \to \infty$ in (2.16) reads as before

$$C(\theta) = \sum_{a=1}^{N_f} c_a(\theta)\mathbf{u}_a(\theta)\mathbf{u}_a^t(\theta),$$

thanks to the hypothesis that ρ_∞ is uniform, which is the counterpart in continuous space to the one considered in (2.15). This stated we can now express the evolution of the feature matrix by its derivative w.r.t. t on the basis $\{\mathbf{u}_a\}$ and $\{v_a(\mathbf{x})\}$. In order to lighten the notations we will use $\mathbf{v}_a$ instead of $v_a(\mathbf{x})$ to denote the right functional basis and hide any explicit dependency over $\mathbf{x}$. We follows the same procedure as what was done previously to model the learning dynamics of restricted Boltzmann machines, approximately for the binary-binary in [52, 53] or exactly for the Gaussian-spherical in [54]. Infinitesimal variations of the feature matrix are given by

$$dF = \sum_{a=1}^{N_f}\Big[\frac{dc_a}{2\sqrt{c_a}}\mathbf{u}_a\mathbf{v}_a^t + c_a\Big(\sum_{b=1}^{N_f} d\Omega_{ab}^{(u)}\mathbf{u}_b\mathbf{v}_a^t + \sum_{b=1}^{D} d\Omega_{ab}^{(v)}\mathbf{u}_a\mathbf{v}_b^t\Big)\Big], \qquad (4.4)$$

where $d\Omega^{(u,v)}$ are skew-symmetric operators representing infinitesimal rotation generators of the $\mathbf{u}_a$ and $\mathbf{v}_a$ vectors,

$$d\Omega_{ab}^{(u)} \stackrel{\text{def}}{=} (d\mathbf{u}_a^t)\mathbf{u}_b = -d\Omega_{ba}^{(u)},$$

$$d\Omega_{ab}^{(v)} \stackrel{\text{def}}{=} (d\mathbf{v}_a^t)\mathbf{v}_b = -d\Omega_{ba}^{(v)}.$$

At this point let us remark that the problem enjoys a gauge symmetry, namely the loss function is invariant under the combined transformation

$$(\mathbf{w}, F) \longrightarrow (R^t\mathbf{w}, RF),$$

where $R = e^{\Omega^{(u)}}$ is any left rotation, i.e. on feature space, so that we may focus on the effect of scalar transformations and right rotations. The scalar transformation corresponds to modifications of the $\{c_a\}$ i.e. modifying the relative strength between the features, while the right rotations corresponds to changing the features themselves. The relation (4.4) can be inverted, yielding in particular for right rotations (from now on we drop the (v) index)

$$d\Omega_{ab} = \frac{1}{\sqrt{c_a} + \sqrt{c_b}}\left(\mathbf{u}_a^t dF \mathbf{v}_b\right)^A + \frac{1}{\sqrt{c_a} - \sqrt{c_b}}\left(\mathbf{u}_a^t dF \mathbf{v}_b\right)^S, \tag{4.5}$$

where A, S denote respectively the anti-symmetric and symmetric parts w.r.t. a, b. In turn we can express infinitesimal transformations of the loss function under such deformations. Using its asymptotic form (3.26) obtained in Section 3.4 written at finite N_f, i.e. with

$$\Lambda = \frac{\alpha}{\rho}\left(\rho - 1 + \frac{1}{N_f}\sum_{a=1}^{N_f}\frac{1}{1 + \Lambda c_a}\right)$$

from (3.11), and assuming the normalization (3.18) of the signal we have

$$r = \sum_{a=1}^{N_f} f_a^2 \qquad \text{and} \qquad r\mu_1 = \sum_{a=1}^{N_f}\frac{f_a^2}{1 + \Lambda c_a}.$$

Since the signal $\mathbf{f}$ is fixed the effect of rotating the right (functional) bases induces variations of its components given by

$$df_a = \sum_{b=1}^{D} d\Omega_{ab} f_b.$$

As a result we obtain

$$d\Lambda = -\frac{1}{N_f}\sum_{a=1}^{N_f}\frac{\Lambda^2}{(1 + \Lambda c_a)^2}\frac{dc_a}{\rho - \mathcal{Q}}, \tag{4.6}$$

$$dr = \sum_{a,b=1}^{N_f,D} f_a d\Omega_{ab} f_b,$$

and

$$d(r\mu_1) = \sum_{a,b=1}^{N_f,D}\frac{f_a d\Omega_{ab} f_b}{1 + \Lambda c_a} - \sum_{a=1}^{N_f}\frac{\Lambda}{(1 + \Lambda c_a)^2}\left[f_a^2 - \frac{r}{N_f}(\mu_1 - \mu_2)\right]dc_a.$$

Using (4.5) we can now write the variations of the loss function in terms of dF as ($dF_{aa} = dc_a/(2\sqrt{c_a})$)

$$d\mathcal{L} = -\sum_{a=1}^{N_f} \frac{\Lambda\sqrt{c_a}}{(1+\Lambda c_a)^2}\Big(f_a^2 + \frac{\mathcal{B}}{N}\Big)dF_{aa} - \frac{\Lambda^2}{2}\sum_{a,b=1}^{N_f,D} \frac{f_a f_b \sqrt{c_a}}{(1+\Lambda c_a)(1+\Lambda c_b)}dF_{ab}, \quad (4.7)$$

letting $c_b = 0$ for $b > N_f$, $\mathcal{B}$ is the bias given by (3.20) and $N = \rho N_f$ the number of samples. From these we can now obtain the equations expressing the learning dynamics asymptotically. In absence of constraints from the NN structure, to be discussed in the next Section, the time variation of the feature matrix is simply given by the gradient descent of the loss function w.r.t. its coefficients

$$\frac{dF_{ab}}{dt} = -\frac{\partial\mathcal{L}}{\partial F_{ab}},$$

which can be read off from the previous equation. First the scalar deformation gives us the dynamics of the eigenvalues of the population matrix:

$$\frac{d\sqrt{c_a}}{dt} = \frac{dF_{aa}}{dt} = -\frac{\partial\mathcal{L}}{\partial F_{aa}}$$

while the rotation of the eigenvectors are determined by the variation of off-diagonal coefficients of the feature matrix

$$\frac{d\Omega_{ab}}{dt} = -\frac{1}{c_a - c_b}\Big(\sqrt{c_a}\,\frac{\partial\mathcal{L}}{\partial F_{ab}} + \sqrt{c_b}\,\frac{\partial\mathcal{L}}{\partial F_{ba}}\Big) \qquad \forall a \neq b.$$

Again, $c_b = 0$ by convention for transverse modes, i.e. $b > N_f$. Note that the rotation between vectors belonging to the same degenerate eigenspaces is not defined and can be set to zero by convention. Consequently we get

$$\frac{d\Omega_{ab}}{dt} = \Lambda^2 \frac{c_a + c_b}{c_a - c_b}\frac{f_a f_b}{\big[1+\Lambda c_a\big]\big[1+\Lambda c_b\big]}. \qquad (4.8)$$

The equations simplify by remarking that the transverse components are involved only through the macroscopic quantity (3.17). We have

$$r = \sum_{a=1}^{N_f} f_a^2 \qquad \text{and} \qquad \sum_{b=N_f+1}^{D} f_b^2 = 1 - r - \sigma^2, \qquad (4.9)$$

thanks to the normalization (3.18). As a result we end up with the following dynamical system for $a = 1,\dots N_f$:

$$\dot{c}_a = \frac{2\Lambda^2 c_a}{\big[1+\Lambda c_a\big]^2}\Big(f_a^2 + \frac{\mathcal{B}}{N}\Big), \qquad (4.10)$$

$$\dot{f}_a = \Lambda^2\Big[\frac{f_a}{1+\Lambda c_a}(1 - r - \sigma^2) + \sum_{b=1,b\neq a}^{N_f} \frac{c_a + c_b}{c_a - c_b}\frac{f_a f_b^2}{\big[1+\Lambda c_a\big]\big[1+\Lambda c_b\big]}\Big]. \qquad (4.11)$$

Given the definition (4.9,3.20) of r and the bias $\mathcal{B}$ and the fact that Λ is self-consistently determined through (3.11), all the macroscopic variables appearing in

these equations can be expressed in terms of spectral coefficients g_k and μ_k given by (3.9,3.10), which read

$$g_k[\Lambda] = \frac{1}{N_f} \sum_{a=1}^{N_f} \frac{1}{(1 + \Lambda c_a)^k}, \qquad \text{and} \qquad \mu_k[\Lambda] = \sum_{a=1}^{N_f} \frac{f_a^2}{(1 + \Lambda c_a)^k} \qquad (4.12)$$

when considered at finite size. Consequently the system (4.10,4.11) constitutes an autonomous dynamical system over the variables $\{c_a, f_a; a = 1, \ldots N_f\}$. It describes the evolution of the $2N_f$ microscopic degrees of freedom of the regression problem. As we see, the time derivatives of c_a or f_a^2 are of order $1/N_f$, while $c_a = \mathcal{O}(1)$ and $f_a^2 = \mathcal{O}(1/N_f)$, so c_a evolve at a different time scale than f_a^2. This means that the evolution of the c_a and the f_a decouple, the f_a being fast variables while the c_a (as well as Λ as we shall see) evolve adiabatically w.r.t. these fast variables.

Our next goal is to derive macroscopic equations associated to this dynamical system. Before that let us discuss qualitatively the behavior of the microscopic equations. First notice that equations (4.10) tend clearly to align the population matrix with the signal in the sense discussed in Section 3.3 for having good generalization. Indeed we see that the rate of increase of c_a is proportional to f_a^2, hence modes well aligned with the signal grow faster. The second equation is also driving the system into the good direction, regarding generalization. The first term on the rhs of equation (4.11) is a positive source term corresponding to increasing the specification of the feature w.r.t. the signal, i.e. to increase r as we shall see in Section 4.3. Because of the factor $1/(1 + \Lambda c_a)$ we see that this increase goes preferentially to the weak modes. The second term instead conserves the longitudinal spectral power of the signal wrt the features but it induces a net flow of spectral power from weak to strong modes. When combined these mechanisms contribute to improve the alignment of the population matrix with the signal.

4.2   Effect of the architecture

So far we did not take into account the limits and biases induced by the specific architecture of the model like for instance if the feature matrix $F(\theta)$ is obtained from the last layer of a deep neural network. Considering this effect would require to project equations (4.8) on the deformation axes favored by the architecture. The generators of rotation have indeed the following decomposition

$$d\Omega_{ab} = \sum_k \Omega_{ab}^k(\theta) d\theta_k$$

where Ω^k is the rotation induced by an infinitesimal variation of the kth component of the parameter vector. This does span only a subset of all possible rotations, and it is determined solely by the architecture of the model in a complicated way. In the end this would result in some damping factor on the evolution of the c_a in (4.10), and in some discount factors applied on the rotation generators especially on the one exchanging longitudinal $(a \leq N_f)$ with transverse vectors $(b > N_f)$. This latter effect is important as it corresponds to the inductive bias of the architecture which would for instance favor smooth functions and penalize rotation toward rugged ones. A minimal way to take this into account is to modify the ridge penalty

term in (4.1). By default is was choosen isotropic, but we may replace it by some structured one

$$\mathcal{L}(\mathbf{w}, \theta) = \frac{1}{2} \int d\mathbf{x} d\mathbf{x}' \mathbf{w}^t F(\mathbf{x}|\theta)^t S(\mathbf{x}, \mathbf{x}') F(\mathbf{x}'|\theta) \mathbf{w} + \frac{\alpha}{N_s} \sum_{s=1}^{N_s} \big[y^{(s)} - \mathbf{w}^t F(\mathbf{x}^{(s)}|\theta)\big]^2,$$
(4.13)

where S is a positive definite kernel function. For instance if the input space is a functional space spanned by a Fourier basis, S could be diagonal in this basis with some dependence in $|\mathbf{k}|^2$ with respect to wave numbers. This would add extra terms in the microscopic dynamical equations that we won't investigate in the present work. Keeping the ridge penalty term in isotropic form is already partially taking care of the implicit regularization of the model. Other arguments given in [14] provide some justification for such considerations. Additionally in order to justify that F can be considered as a matrix and derived components-wise requires to consider the system to be in a "semi-lazy" regime as explained on Figure 4.5, by reference to the lazy regime discussed in Section 2.1. We assume only the parameter of the vector of feature function to be in the lazy regime, the weights of the last layer remaining arbitrary. In such case we are send back to the problem of learning a $D \times N_f$ feature matrix, with D being however potentially prohibitively large as discussed in Section 2.1, depending on the chosen level of resolution when $d \gg 1$. There additional considerations have to come into play to make sense of an efficient learning dynamics in such cases, because $r(t = 0)$ would be otherwise very close to zero at initialization and the learning would have a hard time to start up. These considerations like intrinsic dimension of the input $\mathbf{x}$ much lower than d, the inductive bias of the NN or weight sharing due to the composite nature of the input $\mathbf{x}$ [55] have been discussed mostly qualitatively but could be formalized in present framework by considering a feature matrix with some structure and symmetries. This is however outside the scope of the present paper.

4.3 Macroscopic equations

The equations (4.10,4.11) determine completely the learning dynamics at the microscopic level in absence of architecture constraints. At the macroscopic level, the dynamics would be conveniently summarized by those of $g(t)$ and $g_2(t)$ (or equivalently $\Lambda(t)$ and $\mathcal{Q}(t)$), $\mu_1(t)$, $\mu_2(t)$ and $r(t)$. Unfortunately, as we shall see we cannot obtain in general a closed system of equations. Instead as usual in statistical physics, the macroscopic description is obtained in the form of a hierarchy of equations involving all the moments of some distribution. At least the dynamics of r is straightforward, it is directly obtained from (4.11) by summing the $f_a \dot{f}_a$. We get

$$\dot{r} = 2\Lambda^2 r(1 - r - \sigma^2)\mu_1, \tag{4.14}$$

with again $\mu_1 \in [0, 1]$ given by (3.10) ($\mu^{\parallel} = r$). This yields a sigmoid-like behaviour:

$$r(t) = \frac{1 - \sigma^2}{1 + C \exp\left(-2 \int_0^t d\tau \Lambda^2(\tau)\mu_1(\tau)\right)},$$

with $C > 0$ a constant determined by initial conditions. In order to get the dynamics of Λ, we simply combine (4.6) with(4.10) to obtain

$$\frac{d}{dt}\big[\Lambda(t)^{-1}\big] = \frac{2\Lambda r}{N_f}\frac{\tilde{\mu}_3 - \tilde{\mu}_4}{\rho - \mathcal{Q}}, \tag{4.15}$$

with $\mathcal{Q}$ given by (3.15) and

$$\tilde{\mu}_k \stackrel{\text{def}}{=} \mu_k + \frac{\mu_2}{\rho - \mathcal{Q}}g_k.$$

Since $\mu_3 > \mu_4$ and $g_3 > g_4$ and $\rho > \mathcal{Q}$, we see that the effective ridge penalty Λ^{-1}, or equivalently the test to train error ratio, is strictly increasing with time, leading eventually, as expected, to overfitting. But due to the $1/N_f$ factor, we see that its dynamics is slow compared to those of r. In order to close the equations we need also the derivatives of g_k and μ_k. For the g_k we get

$$\dot{g}_k = \frac{2kr\Lambda^2}{N_f}\Big[\tilde{\mu}_{k+3} - \tilde{\mu}_{k+2} + \frac{(\tilde{\mu}_3 - \tilde{\mu}_4)(g_k - g_{k+1})}{\rho - \mathcal{Q}}\Big], \tag{4.16}$$

the $1/N_f$ in front indicates that these are slow variables. Concerning the μ_k instead, their dynamics which is dominated by the variations of f_a, decouple from the previous ones. Neglecting the contributions from the variation of c_a and Λ of order $1/N_f$, and exploiting the anti-symmetry of the second r.h.s. term in (4.11) we obtain

$$\dot{\mu}_k = 2\Lambda^2\Big[(1 - r - \sigma^2)(\mu_{k+1} - \mu_1\mu_k) + r\sum_{q=0}^{k-1}(\mu_{q+2} - \mu_{q+1})\mu_{k+1-q}\Big]. \tag{4.17}$$

This last equation is valid as long as the spectral density of the signal do not condensate on a mode or on a subset of degenerate modes, otherwise additional terms need to be taken into account. The problem is that in this form this equation leads to a non-physical stationary state. The first term on the rhs is going to saturate at some point when r approaches its limit $1 - \sigma^2$ while the second one is strictly negative and vanishes when all the $\mu_k \to 0$. Coming back to the discussion of equation (4.11), it is actually already expected that the microscopic dynamics leads to concentrate the spectral power of the signal on the strong modes, eventually leading to the apparition of a condensate on the strongest one. To take this into account we need to modify slightly the equations which will be done in the next section. The system (4.14,4.17,4.15,4.16) constitute a hierarchical set of macroscopic equations describing the learning process in closed form. Actually it is decomposed into two subsystems (4.14,4.17) and (4.15,4.16) respectively of fast and slow variables, which means that in practice to solve it one should insert at each time step of the slow system, the stationary state of (4.14,4.17) conditionally on Λ and the g_k's.

4.4 Macroscopic dynamics with condensate

The equations derived so-far from RMT assume in particular that the signal decomposes along the modes of the population matrix with components of order $f_a = \mathcal{O}(1/\sqrt{N_f})$. Consequently we assume that the signal condensate on a degenerate level c_a with degeneracy $\mathcal{O}(N_f)$ to stay within valid hypothesis regarding

RMT. As a result equations (4.10,4.15) correspond still to a slow dynamics. Let $c_{\max}$ the level associated to the condensate and μ^c the longitudinal spectral power fraction of the signal on the condensate. Hence we have the decomposition

$$\mu^{\|}(x) = \mu^{\mathrm{bulk}}(x) + \mu^{\|}\mu_{\mathrm{cmax}}\delta(x - c_{\max}).$$

Then we can write accordingly the following decomposition

$$\mu_k = \mu_k^{\mathrm{bulk}} + \mu_{\mathrm{cmax}}y^k,$$

where

$$y \stackrel{\mathrm{def}}{=} \frac{1}{1 + \Lambda c_{\max}}.$$

Including the term which was neglected in (4.17) yields

$$\dot{\mu}_k = 2\Lambda^2\Big[(1-r-\sigma^2)(\mu_{k+1}-\mu_1\mu_k)+r\sum_{q=0}^{k-1}(\mu_{q+2}-\mu_{q+1})\mu_{k+1-q}+kr\big(\mu_{k+2}^{(2)}-\mu_{k+3}^{(2)}\big)\Big],$$

where

$$\mu_k^{(2)} \stackrel{\mathrm{def}}{=} \frac{1}{r^2}\sum_{a=1}^{N_f} \frac{f_a^4}{(1 + \Lambda c_a)^k}.$$

Here the index a may represent indifferently individual modes or more conveniently spectral level associated with c_a. In the latter case f_a^2 represents the spectral density on this level which can be of order $\mathcal{O}(1)$ instead of $\mathcal{O}(1/N_f)$ if the degeneracy of that level is extensive. This is the case for the condensate which therefore gives the contribution

$$\mu_k^{(2)} \stackrel{\mathrm{def}}{=} \mu_{\mathrm{cmax}}^2 y^k + \mathcal{O}(1/N_f).$$

Finally we get

$$\dot{\mu}_k = 2\Lambda^2\Big[(1-r-\sigma^2)(\mu_{k+1}-\mu_1\mu_k)+r\sum_{q=0}^{k-1}(\mu_{q+2}-\mu_{q+1})\mu_{k+1-q}+rk\mu_{\mathrm{cmax}}^2(1-y)y^{k+2}\Big].$$

(4.18)

Closing this system requires the introduction of additional macroscopic variables in order to express the dynamics of μ_{cmax} leading to cumbersome equations. Instead we will content ourselves with a simple bounding of the derivative of μ_{cmax}. From the microscopic equation (4.11) we have

$$\dot{\mu}_{\mathrm{cmax}} = \Lambda^2 y\mu_{\mathrm{cmax}}\Big[1 - r - \sigma^2 + \sum_{b=1,c_b\neq c_{\max}}^{N_f} \frac{c_{\max} + c_b}{c_{\max} - c_b}\frac{f_b^2}{1 + \Lambda c_b}\Big],$$

Both terms on the rhs are positive. Let $\Delta = c_{\max} - \max_{b,c_b\neq c_{\max}}(c_b)$ the spectral gap between the condensate and the bulk. We have

$$\Lambda^2 y\big[\mu_1^{\mathrm{bulk}} + (1 - r - \sigma^2)\big] \leq \frac{\dot{\mu}_{\mathrm{cmax}}}{2\mu_{\mathrm{cmax}}} \leq \Lambda^2 y\big[\frac{2c_{\max} - \Delta}{\Delta}\mu_1^{\mathrm{bulk}} + (1 - r - \sigma^2)\big].$$

From this it is now clear that μ_{cmax} increases as long as $r < 1 - \sigma^2$ and $\mu_1^{\mathrm{bulk}} > 0$, i.e. $\mu_{\mathrm{cmax}} < 1$. As a result, as can be checked, equation (4.18) has a stationary state

corresponding to $\mu_{\text{cmax}} = 1$ i.e. $\mu^{\text{bulk}} = 0$, and $r = 1 - \sigma^2$, meaning that the signal is completely condensed on the strong modes, which as discussed in Section 3.3 is a crucial point for obtaining good generalization properties. In practice it is not clear whether the dynamics can maintain a degenerated set of strong modes, or if instead isolated modes appear at the top of the spectrum. In that case the dynamics of previously slow variables become faster (because f_a^2 in equation (4.10) is $\mathcal{O}(1)$) and coupled to the dynamics of the μ_k, which should result in a more complex behavior. Since also the validity of RMT in this case is not granted we will leave this aside and focus instead on simple solvable special cases in Section 4.5.

On Figure 4.6 is shown a comparison between the actual gradient descent and the autonomous dynamical system (4.10,4.11), which actually confirms that condensation occurs. The dynamical equations are written with the hypothesis that the train data are un-correlated from one time step to the next one, which is not necessarily true in practice because in general the training set is finite, except for streaming data. To keep with a consistent definition of ρ that represents the ratio of training data to the number of parameters of the model the experimental setting is the following: to each time step is associated a batch of training data of size $N = \rho N_f$ out of a training set of overall size $N_{\text{train}} \stackrel{\text{def}}{=} \rho \times N_f(D+1)$, where $N_f(D+1)$ represent the total number parameters of the model, N_f being the number of weights of the last layer and DN_f being the number of entries in the feature matrix F. The dynamical system is initialized with the eigenvalues of the population matrix and the components of the signal on the corresponding eigenmodes given at the initialization of the experiment. On the examples of Figure 4.6 the population matrix is initialized with 5 equally spaced and degenerate levels and to compare we monitor the eigenvalues $c_a(t)$, the components $f_a(t)$ and various macroscopic variables of interest like train and test errors $E_{\text{train,test}}(t)$, the loss $\mathcal{L}(t)$, the longitudinal power $r(t)$ of the signal and the train-test error ratio $\mathcal{R}^2(t)$. For the actual learning experiment, at each time step, E_{train} as well as $\mathcal{L}$ are evaluated on the batch of training data, while E_{test} is evaluated with 10^4 independent test samples and $\mathcal{R}^2(t)$ is evaluated via the empirical estimation $\hat{\Lambda}$ of Λ given in Section 3.4. For the dynamical system, equations (3.25,3.19,3.26,3.8) are used along with the fixed point equation (3.11) for Λ and the finite size counterparts (4.12) of the spectral coefficients (3.9,3.10). As we see, for $\rho = 2$, i.e. in the under-parameterized regime the autonomous system reproduces exactly the dynamics of the experiments, remarkably in a situation outside of the domain of validity of the equations, because some condensate emerge from each levels, temporarily for all except the top level. In the over-parameterized regime, i.e. $\rho = 0.5$ the dynamics is also very-well reproduced, but we see a small deviation as time progresses due to the finite size of the training data set. This is to be expected from the finite size of the training set, the deviation becoming more pronounced for small values of ρ, but indeed disappear (not shown) when the experiment is feed with un-correlated streaming-data. In order to understand this consider the ratio

$$\rho_D \stackrel{\text{def}}{=} \rho \frac{N_f(D+1)}{D},$$

which represents the number of training data points per direction in the embedding space. Its value is respectively 50 and 200 for $\rho = 0.5$ and $\rho = 2$ in the experiments.
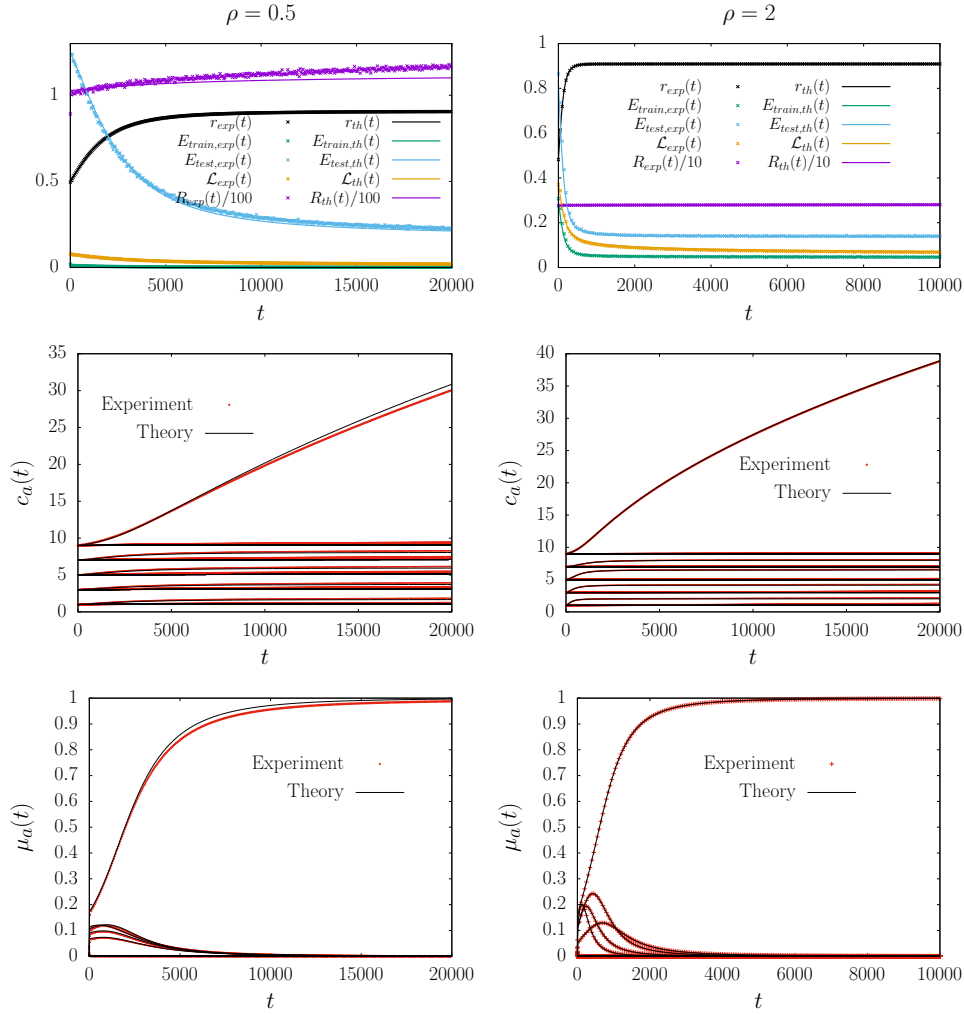
Fig. 4.6: Feature learning process dynamics: comparison between actual
          feature learning processes and the corresponding dynamical system
          (4.10,4.11). The left and right panels are respectively in the over
          and under-parameterized regime with same level of noise $\sigma^2 = 0.1$.
          The population matrix is initialized with 5 equally spaced and de-
          generate levels. Concerning the learning process we have in both
          cases $N_f = D/2 = 100$.

Each input vector $\mathbf{x}_s$ being normalized to $\sqrt{D}$, the empirical operator

$$\hat{\mathbb{I}} = \frac{1}{N_{\text{train}}} \sum_{s=1}^{N_{\text{train}}} \mathbf{x}_s \mathbf{x}_s^t \xrightarrow[N_{\text{train}} \to \infty]{} \mathbb{I}$$

is a finite N_{train} approximation to the identity operator in the embedding space and the vector

$$\tilde{f} \stackrel{\text{def}}{=} \hat{\mathbb{I}} f$$

is the signal as viewed from the training set. From RMT we have

$$\|\mathbf{f} - \tilde{\mathbf{f}}\|^2 \approx \frac{\|\mathbf{f}\|^2}{\rho_D},$$

and the training process is going to point to $\tilde{f}$ instead of $\mathbf{f}$ which qualitatively explain the observed deviation. Since $\rho_D \propto N_f$ this is however a finite size effect which vanishes in the thermodynamics limits.

4.5   Special cases

Let us consider some cases where the macroscopic equations can be closed. These correspond to a population matrix having a finite number of degenerate levels. According to the microscopic equations, all the modes of the same level evolve the same way in that case, provided that the spectral power of the signal is equally distributed at initialization among the modes belonging to the same levels (identical f_a^2). Here of course it is assumed that no spontaneous symmetry breaking is taking place. What is shown on Figure 4.6 proves actually the contrary, this instability is observed and well explained by equation (4.11) which tends to amplify differences between longitudinal modes. Anyway it is always possible to consider a scenario where degenerate levels of the population matrix are maintained artificially during the learning, by mixing the modes and averaging the eigenvalue within each level after each iteration. This gives consistent results (not shown) where the special cases considered here become relevant.

The one-level case: is the first obvious case and corresponds to have a population matrix with one single degenerate level. So again in this case we have $g_k = \mu_k = y^k$ with $y \in [\max(0, 1 - \rho), 1]$, so that the dynamical equation (4.15) concerning Λ can be rewritten directly in term of y and r only and we end up with two coupled equations:

$$\dot{y} = -\frac{2\alpha^2 r}{\rho^2 N_f} \frac{(y + \rho - 1)^3}{\rho - (1 - y)^2} y^3 (1 - y) \left[1 + \frac{y^2}{\rho - (1 - y)^2}\right], \qquad (4.19)$$

$$\dot{r} = \frac{2\alpha^2}{\rho^2} r (1 - r - \sigma^2) y (\rho - 1 + y)^2. \qquad (4.20)$$

First r which is a fast variable tends to $1 - \sigma^2$, then y at a slower speed will eventually tends to 0. Note that equation (4.18) is irrelevant in this case, as can be checked we get $\dot{\mu}_k = 0$ which corresponds to absence of fast variations of μ_k. At

large time we therefore get the generalization error given by (3.27) with $g = g_{\min}$ which actually corresponds to the overfitting regime ($\Lambda = 0$). In order to obtain the best result, one should stop in that case the optimization when $g = g^\star$ given in (3.28) with $r = 1 - \sigma^2$ which maybe unknown actually.

The two-level case: is more interesting and will be useful to illustrate the reinforcement mechanism on the strong modes. As before we denote by $y_a \stackrel{\text{def}}{=} 1/(1 + \Lambda c_a)$ for each level $a \in 1, 2$. We have the decomposition

$$g_k = (1 - \nu)y_1^k + \nu y_2^k,$$

$$\mu_k = (1 - \mu)y_1^k + \mu y_2^k,$$

where ν is the fraction of strong longitudinal modes which remains fixed, while $\mu \in [0, 1]$ is the fraction of the spectral power of the signal aligned with the features which are contained in the strong level; Consequently the effective coupling reads

$$\Lambda(y_1, y_2) = \frac{\alpha}{\rho}\big[\rho - 1 + (1 - \nu)y_1 + \nu y_2\big]. \tag{4.21}$$

For $\rho \geq 1$ the valid domain $\mathbb{D}$ of these variable is $(y_1, y_2) \in [0, 1]^2$. Instead for $\rho < 1$, the constraint $g \in [1 - \rho, 1]$ (see Section 3.3) result in $\mathbb{D}$ being the intersection of $[0, 1]^2$ with

$$(1 - \nu)y_1 + \nu y_2 \geq 1 - \rho, \tag{4.22}$$

i.e. above the line

$$\nu y_2 = (\nu - 1)y_1 + 1 - \rho,$$

where $\Lambda = 0$ in the (y_1, y_2) plane. As already explained in the last section μ is a fast variable, hence if N_f is large enough its dynamics is decoupled from the slow one and we can directly assume $\mu = 1$. It remains to examine the slow dynamics, which is given by the evolution of Λ. We have

$$\dot{\Lambda} = -y_2^2 \frac{2r}{N_f} \frac{\Lambda^3}{\rho - \mathcal{Q}}\Big[y_2^3(1 - y_2) + \frac{g_3 - g_4}{\rho - \mathcal{Q}}\Big]$$

where the specification on the present case yields

$$\mathcal{Q} = (1 - \nu)(1 - y_1)^2 + \nu(1 - y_2)^2, \tag{4.23}$$

$$g_3 - g_4 = (1 - \nu)y_1^3(1 - y_1) + \nu y_2^3(1 - y_2)^2. \tag{4.24}$$

In addition we have

$$\dot{y}_a = -y_a(1 - y_a)\Big[\frac{\dot{\Lambda}}{\Lambda} + \frac{\dot{c}_a}{c_a}\Big]$$

and from (4.10) we have

$$\frac{\dot{c}_1}{c_1} = \frac{2\Lambda^2 r}{N_f}\frac{y_1^2 y_2^2}{\rho - \mathcal{Q}},$$

$$\frac{\dot{c}_2}{c_2} = \frac{2\Lambda^2 r}{N_f}y_2^2\Big(\frac{1}{\nu} + \frac{y_2^2}{\rho - \mathcal{Q}}\Big).$$
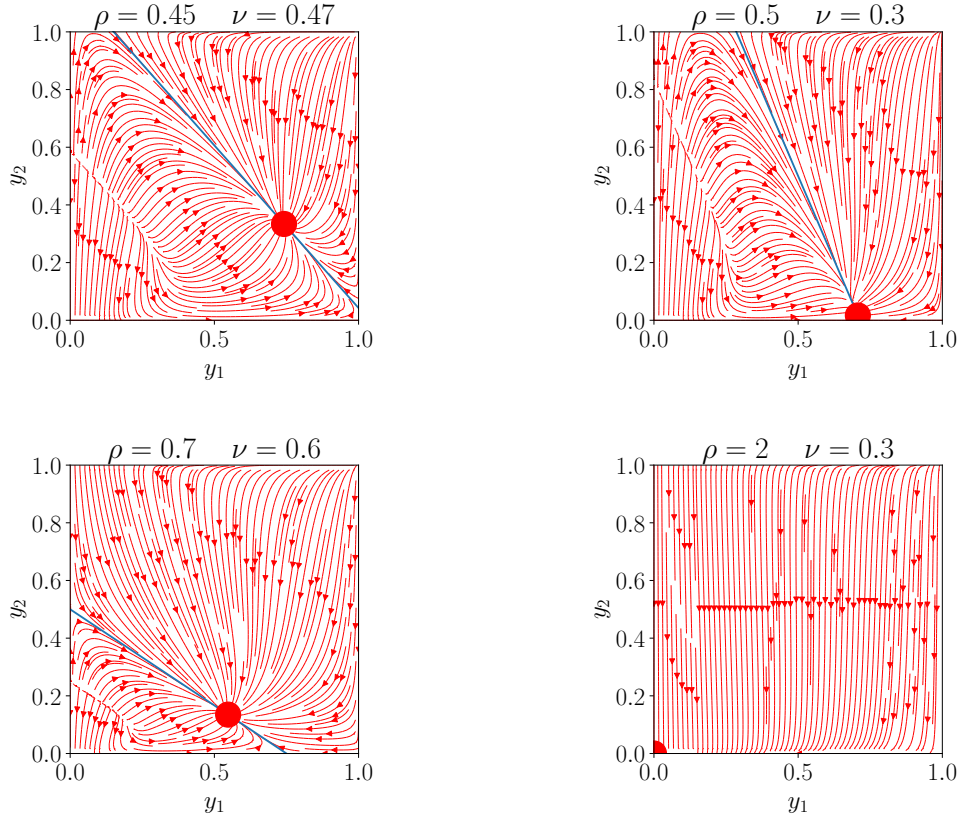
Fig. 4.7: Examples of phase portraits corresponding to the autonomous system (4.25,4.26). The red dot locate the stable fixed point while the blue line represents $\Lambda(y_1, y_2) = 0$ given by (4.21) above which is the valid domain $\mathbb{D}$.

with $f_a^2 = r\left[N_f \nu \min(1, \rho)\right]^{-1}$ for modes in the strong level ($f_a^2 = 0$ for the other ones). Finally we obtain

$$\dot{y}_1 = -\frac{2\Lambda^2 r}{N_f} \frac{y_1(1 - y_1)y_2^2}{\rho - \mathcal{Q}}\left(y_1^2 - y_2(1 - y_2) - \frac{g_3 - g_4}{\rho - \mathcal{Q}}\right), \qquad (4.25)$$

$$\dot{y}_2 = -\frac{2\Lambda^2 r}{N_f} y_2^3(1 - y_2)\left(\frac{1}{\nu} + \frac{y_2(2y_2 - 1)}{\rho - \mathcal{Q}} - \frac{g_3 - g_4}{(\rho - \mathcal{Q})^2}\right), \qquad (4.26)$$

which altogether with (4.23,4.24) and (4.21) defines an autonomous system for the variables (y_1, y_2). In terms of (y_1, y_2) the test error reads

$$E_{\text{test}}(y_1, y_2) = \frac{\rho\left[1 + r(y_2^2 - 1)\right]}{\rho - (1 - \nu)(1 - y_1)^2 - \nu(1 - y_2)^2}.$$

Some phase portraits of this simple dynamical system are shown on Figure 4.7 with different combinations of parameters ρ and ν. As can be easily checked, the

point $(y_1, y_2) = (0,0)$ is a stable fixed point of this system which is in the valid domain $\mathbb{D}$ defined above only for $\rho > 1$. Other stationary points like $(y_1, y_2) \in \{(1,0),(0,1),(1,1)\}$ are unstable. For $\rho > 1$ no other fixed points exit and the system always converges to $(0,0)$. For $\rho \leq 1$ this point is outside $\mathbb{D}$, but as can be shown, there are two additional possible fixed points: $(y_1, y_2) = \left(\frac{1-\rho}{1-\nu}, 0\right)$ and $(y_1, y_2) = \left(\sqrt{1-\rho}, \frac{1-\nu}{\nu}\left[\frac{1-\rho}{1-\nu} - \sqrt{1-\rho}\right]\right)$, both lying on the line $\Lambda(y_1, y_2) = 0$ but not necessarily in $\mathbb{D}$. When $\rho \in [\nu(2-\nu), 1]$ the first one is a stable fixed point, the second being outside $\mathbb{D}$, while for $\rho \in [0, \nu(2-\nu)]$ the second is always the only stable fixed point. These four situations are illustrated on Figure 4.7. The test error which is reached at these various fixed points is then given by

$$E_{\text{test}}(t \to \infty) = \begin{cases} \dfrac{\rho}{\rho-1}(1-r) & \rho \geq 1, \\[2ex] \dfrac{\rho}{1-\rho}\dfrac{1-\nu}{\rho-\nu}(1-r) & \nu(2-\nu) \leq \rho \leq 1, \\[2ex] E_{\text{test}}\left(\sqrt{1-\rho}, \dfrac{1-\nu}{\nu}\left[\dfrac{1-\rho}{1-\nu} - \sqrt{1-\rho}\right]\right) & \rho < \nu(2-\nu). \end{cases}$$

None of these limit points are optimal, in particular close to $\rho = 1$ the system is badly overfitting. A best scenario is obtained with ν small and (y_1, y_2) converging close to $(1,0)$ zero and y_1 close to one, which is obtained with $\rho \simeq \nu(2-\nu)$. Overall we see that the long term dynamics is not necessarily beneficial and may lead to overfitting, in contrary to the short term dynamics of the population matrix discussed in Section 4.1 and 4.4.

5 Discussion

The dynamical system (4.10,4.11) and associated macroscopic equations, obtained in the free probability regime describe very precisely the feature learning process of an ideal NN, sufficiently large and expressive to be considered in a free of constraint and "semi-lazy" regime defined in Section 4.2. Various dynamical features are revealed by this dynamical system. First the orientation of the population matrix with the signal appears to occur at a faster scale than the dynamics of it eigenvalues. Both scale are separated by a factor N_f, i.e. the number of features to be learned. In the first stage of the learning, information from the signal is mainly collected by the weak modes and then transferred to the strongest one which emerge spontaneously. This goes in the direction of having good generalization propertied of the solution, according the alignment criteria of the population matrix with the signal discussed around the general formula (3.19) of the test error derived in the free probability regime thanks to the train-test error ratio discussed in Sections 3.1,3.2 and 3.3. Instead the long term dynamics offers less guaranties as seen with the dynamical system (4.25,4.26) corresponding to a 2-level artificially maintained population matrix.

These observations call for further investigations in different avenues, both practical and theoretical. From the practical point of view, we may expect to be able to improve on feature learning algorithms, in order to insure that the result obtained is optimal for the generalization criteria and that the time to reach the result is

minimal. From the theoretical point of view, as suggested in Section 4.2, the unreasonable success of deep learning [56] could be investigated more quantitatively by rewriting the dynamical equations with a structured feature matrix and appropriate hypothesis on the signal, and adapting them to the classification context instead of the regression considered here. The possibility to transfer some observations of the present work to the unsupervised learning context seems also plausible thanks to the observation made in a previous work that a convex relaxation of a restricted Boltzmann machine, called "Coulomb machine" could be learned in principle by means of constraint linear regressions [57].

Acknowledgements

I wish to thank Guillaume Charpiat for insightful discussions especially on neural networks.

Data Availability

There are no data associated with this paper. Code used to produce the plots can be shared from the author on demand.

References

[1] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[2] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[4] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *proc. of ICML*, pages 541–549. PMLR, 2018.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS*, 116(32):15849–15854, 2019.

[6] E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

[7] M.S. Advani, A.M. Saxe, and H. Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[8] P.L. Bartlett, P.M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 117(48):30063–30070, 2020.

[9] T. Hastie, A. Montanari, S. Rosset, and R.J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

[10] Z. Liao and R. Couillet. The dynamics of learning: A random matrix approach. In *proceedings of ICML*, pages 3072–3081, 2018.

[11] P.L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

[12] D. Richards, J. Mourtada, and L. Rosasco. Asymptotics of ridge (less) regression under general source condition. In *proc. of AIStats*, pages 3889–3897. PMLR, 2021.

[13] D. Wu and J. Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *In proc. of NeurIPS*, 33:10112–10123, 2020.

[14] C. Louart, Z. Liao, and R. Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

[15] S.I. Amari, J. Ba, R. Grosse, X. Li, A. Nitanda, T. Suzuki, D. Wu, and J. Xu. When does preconditioning help or hurt generalization? *arXiv preprint arXiv:2006.10732*, 2020.

[16] J. Ba, M. Erdogdu, T. Suzuki, D. Wu, and T. Zhang. Generalization of two-layer neural networks: An asymptotic viewpoint. In *proc. of ICLR*, 2019.

[17] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[18] R. Couillet and Z. Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[19] M. Geiger, A. Jacot, S. Spigler, F. Gabriel, L. Sagun, S. dAscoli, G. Biroli, C. Hongler, and M. Wyart. Scaling description of generalization with number of parameters in deep learning. *J. Stat. Mech.: Theory and Experiment*, 2020(2):023401, 2020.

[20] M. Mézard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[21] H.S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A*, 45(8):6056, 1992.

[22] D. Malzahn and M. Opper. Statistical mechanics of learning: a variational approach for real data. *Phys. Rev. Lett.*, 89(10):108302, 2002.

[23] D. Malzahn and M. Opper. A statistical physics approach for the analysis of machine learning algorithms on real data. *J. Stat. Mech.: Theory and Experiment*, 2005(11):P11001, 2005.

[24] B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mezard, and L. Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *In proc. of NeurIPS*, 34:18137–18151, 2021.

[25] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *proc. of ICML*, pages 3452–3462. PMLR, 2020.

[26] S. Spigler, M. Geiger, and M. Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *J. Stat. Mech.: Theory and Experiment*, 2020(12):124001, 2020.

[27] O. Cohen, O. Malka, and Z. Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Phys. Rev. Research*, 3:023034, 2021.

[28] J.W. Rocks and P. Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Phys. Rev. Research*, 4(1):013201, 2022.

[29] D. Saad and S.A. Solla. On-line learning in soft committee machines. *Phys.Rev. E*, 52(4):4225, 1995.

[30] D. Saad and S.A. Solla. Exact solution for on-line learning in multilayer neural networks. *Phys.Rev.Lett.*, 74(21):4337, 1995.

[31] S. Goldt, M. Advani, A.M. Saxe, F. Krzakala, and L. Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

[32] L. Zdeborová and F. Krzakala. Statistical physics of inference: Thresholds and algorithms. *Adv. in Physics*, 65(5):453–552, 2016.

[33] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *In proc. of NeurIPS*, 20, 2007.

[34] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *In proc. of NeurIPS*, 32, 2019.

[35] R.M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[36] C. Williams and C. Rasmussen. Gaussian processes for regression. *In proc. of NeurIPS*, 8, 1995.

[37] J. Lee, Y. Bahri, R. Novak, S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as Gaussian processes. In *proc. of ICLR*, 2018.

[38] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *In proc. of NeurIPS*, volume 31, 2018.

[39] G. Charpiat, N. Girard, L. Felardos, and Y. Tarabalka. Input similarity from the neural network perspective. *Advances in Neural Information Processing Systems*, 32, 2019.

[40] C.M. Bishop and N.M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[41] W. Bialek, C.G. Callan, and S.P. Strong. Field theories for learning probability distributions. *Phys. Rev. Lett.*, 77(23):4693, 1996.

[42] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

[43] O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.

[44] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[45] D. Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104(1):201–220, 1991.

[46] J.A. Mingo and R. Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[47] S. Liu and E. Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *ICLR*, 2020.

[48] A. Jacot, B. Simsek, F. Spadaro, C. Hongler, and F. Gabriel. Implicit regularization of random feature models. In *proc. of ICML*, pages 4631–4640. PMLR, 2020.

[49] J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

[50] A.M. Saxe, J.L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

[51] A. Saxe, S. Sodhani, and S.J. Lewallen. The neural race reduction: Dynamics of abstraction in gated networks. In *International Conference on Machine Learning*, pages 19287–19309. PMLR, 2022.

[52] A. Decelle, G. Fissore, and C. Furtlehner. Spectral dynamics of learning in restricted Boltzmann machines. *EPL*, 119(6):60001, 2017.

[53] A. Decelle, G. Fissore, and C. Furtlehner. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J.Stat.Phys.*, 172(18):15761608, 2018.

[54] A. Decelle and C. Furtlehner. Gaussian-spherical restricted Boltzmann machines. *J.Phys. A: Mathematical and Theoretical*, 53(18):184002, 2020.

[55] H.W. Lin, M. Tegmark, and D. Rolnick. Why does deep and cheap learning work so well? *J.Stat.Phys.*, 168(6):1223–1247, 2017.

[56] T.J. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *PNAS*, 117(48):30033–30038, 2020.

[57] A. Decelle and C. Furtlehner. Exact training of restricted Boltzmann machines on intrinsically low dimensional data. *Phys.Rev.Lett.*, 127(15):158303, 2021.

[58] W. Hachem, P. Loubaton, and J. Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

A   Random matrix theory and planar diagrams

Let us first summarize without much details the original setting and derivation of the Marchenko-Pastur distribution [44] adapted to our context. We assume that C is of the form

$$C^{(N)} = C_0 + \sum_{n=1}^{N} \omega_n \omega_n^t,$$

where $C^{(N)}$ and C_0 are $N_f \times N_f$ symmetric matrices, and ω_n are random vectors of the form

$$\omega_n = \|\omega_n\| \mathbf{x}_n$$

where $\mathbf{x}_n$ is a unit vector uniformly distributed on the N_f-dimensional sphere and the square norm $\|\omega_n\|^2$ of ω_n have the same independent distribution $\sigma(x)$. We assume this to have a second moment τ. As a result the population matrix is given by

$$C = C_0 + \mathbb{E}\Big[\sum_{n=1}^{N} \omega_n \omega_n^t\Big] = C_0 + \rho \tau \mathbb{I},$$

with $\rho = N/N_f$. Let us call $\nu_\rho^{(N)}$ the spectral measure of $C^{(N)}$ (normalized to one) and $g^{(N)}(\rho, z)$ its Stieltjes transform:

$$g^{(N)}(\rho, z) = \int_0^\infty \frac{\nu_\rho^{(N)}(x)}{z+x} = \mathbb{E}\Big[\frac{1}{N_f}\text{Tr}\Big(G^{(N)}(z)\Big)\Big],$$

where

$$G^{(N)}(z) \overset{\text{def}}{=} \big(z + C^{(N)}\big)^{-1}.$$

In order to obtain the limit spectral density of $C^{(N)}$, we have to find the limit $g(\rho, z)$ of $g^{(N)}(\rho, z)$ when N and N_f tend to infinity at fixed ρ. To this end Marchenko and Pastur proceed by recurrence. Thanks to the Sherman-Morrison formula they indeed have

$$g^{(N+1)}(\rho+d\rho, z) = g^{(N)}(\rho, z) - \frac{d\rho}{z + \omega_{N+1}^t G^{(N)}(z)\omega_{N+1}}\text{Tr}\big(G^{(N)}(z)\omega_{N+1}\omega_{N+1}^t G^{(N)}(z)\big).$$

with $d\rho = 1/N_f$. The limit $N \to \infty$ is taken at fixed ρ. Keeping $\|\omega_n\|^2 = \tau_n$ fixed, under simple hypothesis they show self-averaging properties of the resolvent

$$\lim_{N\to\infty}\Big|\frac{1}{N_f}\omega_{N+1}^t G^{(N)}(z)\omega_{N+1} - \tau_{N+1}g(\rho, z)\Big| = 0 \qquad (A.1)$$

and product of resolvents

$$\lim_{N\to\infty}\Big|\frac{1}{N_f}\text{Tr}\big[G^{(N)}(z)\omega_{N+1}\omega_{N+1}^t G^{(N)}(z)\big] - \tau_{N+1}\text{Tr}\big[\big(z+C^{(N)}\big)^{-2}\big]\Big| = 0. \qquad (A.2)$$

This then leads to get the following equation

$$\frac{\partial}{\partial\rho}g(\rho, z) + \frac{\tau(\rho)}{1+\tau(\rho)g(\rho, z)}\frac{\partial}{\partial z}g(\rho, z) = 0,$$

thanks to the identity

$$\frac{1}{N_f}\text{Tr}\Big(\big[z+C^{(N)}\big]^{-2}\Big) = \frac{\partial}{\partial z}g^{(N)}(\rho, z).$$

This is solved via to the method of characteristics as

$$g(\rho, z) = g\Big(0, z + \int_0^\rho dt \frac{\tau(t)}{1+\tau(t)g(\rho, z)}\Big),$$

with initial condition given by the Stieltjes transform of C_0 (assuming a spectral density independent of N_f). Upon ordering the ω_n along with increasing values of τ_n, the solution finally reads

$$g(\rho, z) = g\Big(0, z + \rho \int_0^\infty \frac{\tau d\sigma(\tau)}{1 + \tau g(\rho, z)}\Big). \tag{A.3}$$

Let us see how this solution can be recovered with the Dyson equation, based on Feynman diagrams expansion of the resolvent $G^{(N)}$. We detail this formalism as it can be used conveniently also to compute more complex quantities as train or test errors, i.e. product of matrices involving the resolvent multiple times. The starting point is the expansion

$$G^{(N)}(z) = G_0(z) + \sum_{p=1}^\infty (-1)^p G_0(z) \Big[\sum_{n=1}^N \omega_n \omega_n^t\Big]^p G_0(z), \tag{A.4}$$

with $G_0(z)$ the resolvent of C_0. When developing the terms $\Big[\sum_{n=1}^N \omega_n \omega_n^t\Big]^p$, averaging over the ω_n and taking the limit $N \to \infty$ at fixed $\rho = N/N_f$ we get a summation over terms, each one corresponding to a certain power p in τ and q in ρ, and organized topologically according to the Feynman diagram representation, each one coming with a factor $\rho^q \tau^p$. While p is the number of terms $\omega_n \omega_n^t$ counted with their multiplicity taken into account in the product, q is the number of distinct terms $\omega_n \omega_n^t$ w.r.t. n. These diagrams can be classified in four global topological classes, depending on whether at least two interaction lines cross, and if the diagram can be cut into 2 distinct diagrams by just cutting the propagator line (see Figure A.8). The connected diagrams are actually not directly present in the expansion (A.4) because they are amputated from their external propagator lines. They are building blocks of the complete expansion, similar in a way to what cumulants are for moments of a probability distribution. Defining the so-called self energy Σ_ρ, in the field theory formalism, as the sum over connected diagrams allows one to write the Dyson equation for the full propagator:

$$G_\rho(z) = G_0(z) - \rho G_0(z) \Sigma_\rho G_\rho(z),$$

i.e. the limit of $G^{(N)}$ in thermodynamic limit. In some cases the contribution of crossing diagrams becomes negligible in the asymptotic limit, and this corresponds precisely to the free probability setting [45, 46]. Then the expansion simplifies considerably. The first point is that all factors of the form $\omega^t G_0 \omega$ become deterministic and concentrated at its mean value at large N:

$$\lim_{N_f \to \infty} \omega^t G_0(z)\omega = \lim_{N_f \to \infty} \text{Tr}\Big[G_0(z)\omega\omega^t\Big] = \tau \lim_{N_f \to \infty} \frac{1}{N_f} \text{Tr}\Big[G_0(z)\Big] = \tau g_0(z),$$

and the same remains valid when $G_0(z)$ is replaced by $G_\rho(z)$, as already stated in A.1. Then a simple combinatorial argument shows that the self-energy can be expressed in closed form function of the full propagator $G_\rho(z)$ as a sum of connected diagrams indicated on Figure A.8. Then we have $\Sigma_\rho = \Sigma_\rho[G_\rho]$, with

$$\Sigma_\rho[G] = N_f \sum_{n=1}^\infty (-1)^{n-1} \mathbb{E}\Big[\omega\big(\omega^t G\omega\big)^{n-1}\omega^t\Big] = N_f \mathbb{E}\Big[\frac{\omega\omega^t}{1 + \omega^t G\omega}\Big], \tag{A.5}$$

Non crossing   Crossing
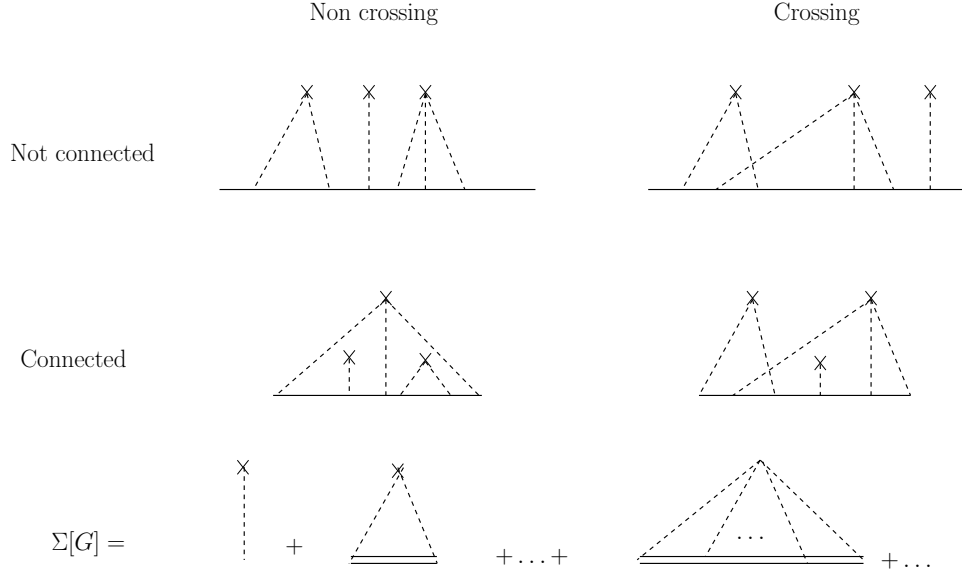

Not connected

Connected

$\Sigma[G] =$

Fig. A.8: Examples of Feynman Diagrams corresponding to a contribution proportional to $\rho^3 \tau^6$. The plain line corresponds to the propagator line and corresponds to insertion of the free propagator $G_0(z)$ between two successive interaction lines (dotted lines) or edges. The top left one for instance is $-\rho^3 \mathbb{E}\Big[G_0 \omega_1 \omega_1^t G_0 \omega_1 \omega_1^t G_0 \omega_2 \omega_2^t G_0 \omega_3 \omega_3^t G_0 \omega_3 \omega_3^t G_0 \omega_3 \omega_3^t G_0\Big]$, and is both non-crossing and not connected. The double plain line corresponds to use the full propagator instead in the expansion of the free energy.

(the N_f showing up comes from writing ρ instead of N in the previous equation). Hence we get G_ρ in a self-consistent form as

$$G_\rho(z) = G_0(z) - \rho G_0(z) \Sigma_\rho \big[G_\rho(z)\big] G_\rho(z), \qquad (A.6)$$

which leads to

$$G_\rho(z) = \Big(\mathbb{I} + \rho G_0(z) \Sigma_\rho\big[G_\rho(z)\big]\Big)^{-1} G_0(z). \qquad (A.7)$$

Note that $\Sigma(z)$ is noted $\mathcal{R}(z)$ and called the $\mathcal{R}$-transform in free probabilities [46]. Using (A.1) and the fact that the distribution of ω is isotropic we obtain

$$G_\rho(z) = G_0(z) - \rho G_0(z) \int_0^\infty d\sigma(\tau) \frac{\tau}{1 + \tau g(\rho, z)} G_\rho(z).$$

To recover (general) Marcenko-Pastur solution we re-arrange this as

$$g(\rho, z) = \text{Tr}\left[\Big(1 + \rho \int_0^\infty \frac{\tau d\sigma(\tau)}{1 + \tau g(\rho, z)} G_0(z)\Big)^{-1} G_0(z)\right]$$

after taking the trace. Then thanks to the identity

$$\text{Tr}\big[G_0(z)^{n+1}\big] = \frac{(-1)^n}{n!} \frac{\partial^n}{\partial z^n} g(0, z),$$

and formally identifying the Taylor expansion we arrive at (A.3).

B   Solution of the Dyson equation in planar approximation

We can now map the preceding equation to the specific problem under consideration to justify equations (3.4,3.5). This justification will remains at a non-rigorous level, even though existing tools of RMT, like deterministic equivalents [58], as those discussed and used in [47] on the same problem but assuming an isotropic signal, would possibly also work here to provide rigorous proofs.

At finite N and N_f our propagator is

$$G^{(N)} = \left(\mathbb{I} + \alpha C^{(N)}\right)^{-1}.$$

Therefore we have $z = 1$, $G_0(z) = \mathbb{I}$ and

$$\sum_{n=1}^{N} \omega_n^{(N)} \omega_n^{(N),t} = \alpha C^{(N)},$$

where the superscript N as been added to underline the fact that the ω_n have a $1/\sqrt{N}$ scaling with N at fixed N_f. We are interested to give a meaning to the limit of this operator in the thermodynamic limit

$$G_\rho = \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} G^{(N)}.$$

At finite N_f we assume also that there is a prior C to $C^{(N)}$, the population matrix, obtained by letting $N \to \infty$, so we have

$$\mathbb{E}\left(\omega^{(N)} \omega^{(N),t}\right) = \frac{\alpha C}{N}. \tag{B.1}$$

Let us denote by $\omega_a^{(N)}$ the components of $\omega^{(N)}$ on the eigenvectors of C:

$$\omega^{(N)} = \sum_{a=1}^{N_f} \omega_a^{(N)} \mathbf{u}_a.$$

We have

$$\mathbb{E}\left[\omega_a^{(N)} \omega_b^{(N)}\right] = \delta_{a,b} \frac{\alpha c_a}{N}. \tag{B.2}$$

The central point insuring the validity of the Marchenko-Pastur theory, hence justifying the planar approximation is that the coefficient $\omega^{(N),t} G_\alpha^{(N)} \omega^{(N)}$ becomes deterministic in thermodynamic limits. If the random vector $\omega^{(N)}$ is assumed to have the moment constraint (B.2) and assuming typically some higher moments to be finite, then there exists some function $h(\rho, \alpha)$ such that the following holds:

$$\lim_{N, N_f \to \infty} \omega^{(N)t} G^{(N)} \omega^{(N)} = h(\rho, \alpha).$$

As a result, combining this with (A.5,B.1) we get

$$\Sigma_\rho[G_\rho] = \frac{\alpha C}{\rho\big[1 + h(\rho, \alpha)\big]}, \tag{B.3}$$

therefore G_ρ, which expression in terms of the self-energy reads

$$G_\rho = \left(\mathbb{I} + \rho\Sigma_\rho[G_\rho]\right)^{-1},$$

commutes with C, being self-consistently defined as

$$G_\rho = \left(\mathbb{I} + \frac{\alpha C}{[1 + h(\rho, \alpha)]}\right)^{-1}, \tag{B.4}$$

Note that these expressions are obtained by averaging, and we used them by convenience, but in a rigorous perspective we should instead use deterministic equivalents [58] as is done in [47]. Nevertheless, this leads us to obtain a spectral decomposition of the Stieltjes transform of $C^{(N)}$ along the modes of C in thermodynamic limit. First define $g(x, \rho, \alpha)$ as

$$\nu_\infty(x)g(x, \rho, \alpha) = \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \mathbf{u}_a^t G^{(N)} \mathbf{u}_a \delta(x - c_a),$$

such that

$$g(\rho, \alpha) \stackrel{\text{def}}{=} \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \frac{1}{N_f} \text{Tr}\left[G_\alpha^{(N)}\right] = \int \nu_\infty(dx)g(x, \rho, \alpha).$$

Then we get

$$h(\rho, \alpha) = \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \text{Tr}\left[G^{(N)}\omega^{(N)}\omega^{(N)t}\right]$$

$$= \lim_{\substack{N, N_f \to \infty \\ N/N_f = \rho}} \frac{\alpha}{N} \text{Tr}\left[G^{(N)}C\right]$$

$$= \frac{\alpha}{\rho} \int \nu_\infty(dx)xg(x, \rho, \alpha)$$

which is a special case of the Ledoit-Péchet formula [43], along with the self-consistent equation

$$h(\rho, \alpha) = \lim_{N_f \to \infty} \frac{1}{N_f} \text{Tr}\left[G^{(N)}C\right] = \frac{\alpha}{\rho} \int \nu_\infty(dx)x\left(1 + \frac{\alpha x}{1 + h(\rho, \alpha)}\right)^{-1}.$$

In equation (B.4) we see that the full propagator differs form the resolvent of αC by all the contributions of correlations between data taken into the self energy. In this process the operator αC is replaced by $\rho\Sigma_\rho$ given in (B.3) or equivalently the bare coupling constant α is replaced by some effective coupling constant

$$\Lambda(\rho, \alpha) = \frac{\alpha}{1 + h(\rho, \alpha)},$$

corresponding to the dressing up of α with fluctuations in the traditional view-point of field theories. This justifies the equation (3.4) for the test-train error ratio, where we see that $\mathcal{R}(\rho, \alpha) = 1 + h(\rho, \alpha)$. Concerning the train error (2.19), to obtain its asymptotic counterpart (3.5) we observe first that power of $G^{(N)}$ are obtained by derivation w.r.t. α:

$$\frac{\partial}{\partial\alpha}G^{(N)} = -\frac{1}{\alpha}G^{(N)}\left(\mathbb{I} - G^{(N)}\right).$$

The fact that the expectation of G_ρ is aligned with C leads directly to (3.5). In particular the first term expressed as a summation over $\mu^\parallel(x)$ define by (3.7), results from the fact that it is a self-averaging quantity as long as the transverse components $\tilde{f}_a = (F^+ f)_a = \frac{f_a}{\sqrt{c_a}}$ on the modes of C are evenly distributed as $\tilde{f}_a^2 = \mathcal{O}(1/\sqrt{N_f})$, i.e. without condensing onto a specific mode.