Multi-scale Feature Learning Dynamics: Insights for Double Descent

Mohammad Pezeshki [1 2] **Amartya Mitra** [3] **Yoshua Bengio** [1 2 4] **Guillaume Lajoie** [1 5 4]

Abstract

An intriguing phenomenon that arises from the high-dimensional learning dynamics of neural networks is the phenomenon of "double descent". The more commonly studied aspect of this phenomenon corresponds to *model-wise* double descent where the test error exhibits a second descent with increasing model complexity, beyond the classical U-shaped error curve. In this work, we investigate the origins of the less studied *epoch-wise* double descent in which the test error undergoes two non-monotonous transitions, or descents as the training time increases. We study a linear teacher-student setup exhibiting epoch-wise double descent similar to that in deep neural networks. In this setting, we derive closed-form analytical expressions describing the generalization error in terms of low-dimensional scalar macroscopic variables. We find that double descent can be attributed to distinct features being learned at different scales: as fast-learning features overfit, slower-learning features start to fit, resulting in a second descent in test error. We validate our findings through numerical simulations where our theory accurately predicts empirical findings and remains consistent with observations in deep neural networks.

1. Introduction

Classical wisdom in statistical learning theory predicts a trade-off between the generalization ability of a machine learning model and its complexity, with highly complex models less likely to generalize well (Friedman et al., 2001). If the number of parameters measures complexity, deep learning models sometimes go against this prediction (Zhang et al., 2016): deep neural networks trained by stochastic gradient descent exhibit a so-called *double descent* behavior (Spigler et al., 2019; Belkin et al., 2019b) with increasing model parameters. Specifically, with increasing complexity, the generalization error first obeys the classical U-shaped curve consistent with statistical learning theory. However, a second regime emerges as the number of parameters is further increased past a transition threshold where generalization error drops again, hence the "double descent" or more accurately *model-wise double descent*.

Nakkiran et al. (2019) showed that the phenomenon of double descent is not limited to varying model size and is also observed as a function of training time or epochs. In this case as well, the so-called *epoch-wise double descent* is in apparent contradiction with the classical understanding of overfitting (Vapnik, 1998), where one expects that longer training of a sufficiently large model beyond a certain threshold should result in overfitting. This has important implications for practitioners and raises questions about one of the most widely used regularization method in deep learning (Goodfellow et al., 2016): early stopping. Indeed, while one might expect early stopping to prevent overfitting, it might in fact prevent models from being trained at their fullest potential.

While there has been significant interest, starting from 1990s, to understand the origins of the non-trivial generalization behaviors of neural networks (Opper, 1995; Opper & Kinzel, 1996; Ba et al., 2019; Mei & Montanari, 2019; d'Ascoli et al., 2020; Gerace et al., 2020), the majority of this previous work has been to understand the *asymptotic* or end-of-training model performance. In recent years though, there has been an interest in studying the *non-asymptotic* (finite training) performance (e.g. Saxe et al., 2013; Advani & Saxe, 2017; Kalimeris et al., 2019; Pezeshki et al., 2020; Stephenson & Lee, 2021). Among the limited work studying the particular epoch-wise double descent, Nakkiran et al. (2019) introduces the notion of *effective model complexity* and hypothesizes that it increases with training time and hence unifies both model-wise and epoch-wise double descent phenomena. Heckel & Yilmaz (2020) also study the dynamics of evolution of single and two layer networks and show that the superposition of two bias/variance trade-off curves with different minima leads to a double descent.

[1]Mila, Québec AI Institute [2]Dept. of Computer Science and Operational Research, Université de Montréal [3]University of California, Riverside [4]Canada CIFAR AI Chair [5]Dept. of Mathematics and Statistics, Université de Montréal. Correspondence to: Mohammad Pezeshki <pezeshki@mila.quebec>, Guillaume Lajoie <g.lajoie@umontreal.ca>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

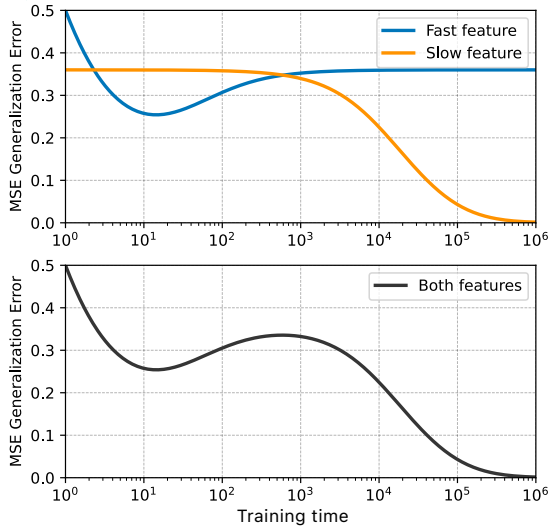

Figure 1. The generalization error as the training time proceeds. (top): The case where only the fast-learning feature or slow-learning feature are trained. (bottom): The case with both features. Features that are learned on a faster time-scale are responsible for the classical U-shaped generalization curve, while the second descent can be attributed to the features that are learned at a slower rate.

In this work, we build on Bös et al. (1993); Bös (1998); Advani & Saxe (2017); Mei & Montanari (2019) which analyze *model-wise* double descent through the lens of linear models, **to probe the origins of *epoch-wise* double descent**. In particular,

- We introduce a linear teacher-student model with features of different strengths. Despite its simplicity, such a model exhibits the epoch-wise double descent of the generalization error under gradient-based training. (Section 2.1)
- In the high-dimensional limit (of number of parameters and sample size), we derive the dynamics of a pair of low-dimensional macroscopic variables, R and Q, describing the generalization behavior of the model. (Eqs. 6, 7)
- Consistent with recent findings, we provide an explanation for the existence of epoch-wise double descent, suggesting that epoch-wise double descent can be attributed to different features being learned at different time-scales. (Figure 1 and Eqs. 12-14)
- We perform simulation experiments to validate our analytical predictions. Furthermore, we conduct experiments with a ResNet-18 model, to demonstrate qualitative similarity between the generalization behavior of our teacher-student setup and that of the former. (Figures 5, 6)

2. Analytical Framework

In this work, we focus on studying the generalization behavior of neural networks under the quintessential gradient-based training scenario, namely (stochastic) gradient descent (SGD/GD). SGD — the de facto optimization algorithm for neural networks — exhibits complex dynamics arising from a large number of parameters (Kunin et al., 2020). While an exact analysis of such dynamics is intractable due to the large number of *microscopic* parameters, it is though possible to capture various aspects of this high-dimensional dynamics in terms of certain low-dimensional comprehensible *macroscopic* entities. This was first demonstrated in a series of seminal papers by Gardner (Gardner, 1988; Gardner & Derrida, 1988; 1989), where the *replica method* of statistical physics was adopted to derive expressions describing the generalization behavior of linear models. In this paper, we employ Gardner's analysis to build upon an established line of work studying linear and generalized linear models (Seung et al., 1992; Kabashima et al., 2009; Krzakala et al., 2012). While most of previous work study the asymptotic ($t \to \infty$) generalization behavior, we adapt these methods to study transient learning dynamics of generalization for finite training time. In the following, we introduce a particular linear teacher-student model and study its generalization performance as a function of training time and regularization strength.

Notation. Scalar variables are denoted in lower case (y), while vectorial entities are represented in boldface ($\boldsymbol{x}$). Lastly, matrices are shown capitalized (F).

2.1. A Teacher-Student Setup

Teacher. We study a supervised linear regression problem in which the training labels y, are generated by a noisy linear model (Figure 2),

$$y := y^* + \epsilon, \quad y^* := \boldsymbol{z}^T \boldsymbol{w}, \quad z_i \sim \mathcal{N}(0, \frac{1}{\sqrt{d}}), \qquad (1)$$

where $\boldsymbol{z} \in \mathbb{R}^d$ is the teacher's input and $y^*, y \in \mathbb{R}$ are the teacher's noiseless and noisy outputs, respectively. $\boldsymbol{w} \in \mathbb{R}^d$ represents the (fixed) weights of the teacher and $\epsilon \in \mathbb{R}$ is the label noise. Here, both w_i and ϵ are drawn i.i.d. from Gaussian distributions with zero mean and variances of 1 and σ_ϵ^2, respectively. Additionally, we choose to set $||\boldsymbol{w}|| = 1$, without loss of generality.

Student. A student model is correspondingly chosen to be a similar shallow network with trainable weights $\hat{\boldsymbol{w}} \in \mathbb{R}^d$. The student model is trained on n training pairs $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^n$, with the labels y^μ being generated by the above teacher network and where student's inputs $\boldsymbol{x}^\mu$ correspond to teacher inputs $\boldsymbol{z}^\mu$ multiplied a predefined and fixed

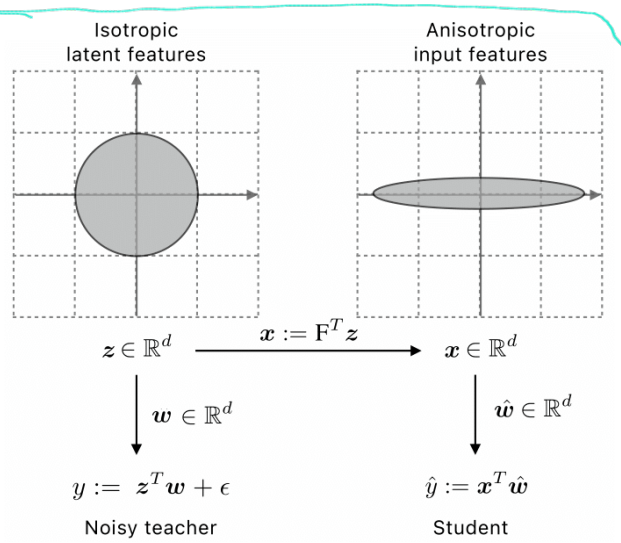

Figure 2. The teacher/student setup: The teacher is the data generating process that given the latent features in $\boldsymbol{z}$, generates student's input, $\boldsymbol{x}$ and its target, y. Student is trained on pairs of $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ where $\boldsymbol{x} := \mathrm{F}^T\boldsymbol{z}$ follow an anisotropic Gaussian distribution such that the directions with larger/smaller variance are learned faster/slower. The condition number of F determines how much faster some features are learned than the others. One can think of $\boldsymbol{z}$ as the latent factors of variation on which the teacher operates, while $\boldsymbol{x}$ can be thought as the pixels that the student learns from.

modulation matrix $\mathrm{F} \in \mathbb{R}^{d\times d}$ that regulates input features' strengths:

$$\hat{y} := \boldsymbol{x}^T\hat{\boldsymbol{w}}, \qquad s.t. \qquad \boldsymbol{x} := \mathrm{F}^T\boldsymbol{z}. \tag{2}$$

One can perceive $\boldsymbol{z}$ to be the latent factors of variation on which the teacher operates, while $\boldsymbol{x}$ corresponds to the pixels that the student learns from. (See Figure 2)

Learning paradigm. To train our student network, we use stochastic gradient descent (SGD) on the regularized mean-squared loss, evaluated on the n training examples as,

$$\mathcal{L}_T := \frac{1}{2n}\sum_{\mu=1}^{n}(y^\mu - \hat{y}^\mu)^2 + \frac{\lambda}{2}||\hat{\boldsymbol{w}}||_2^2, \tag{3}$$

where $\lambda \in [0, \infty)$ is the regularization coefficient. Optimizing Eq. 3 with stochastic gradient descent (SGD) yields the typical update rule,

$$\hat{\boldsymbol{w}}_t \leftarrow \hat{\boldsymbol{w}}_{t-1} - \eta\nabla_{\hat{\boldsymbol{w}}}\mathcal{L}_T + \xi, \tag{4}$$

in which t denotes the training step and η is the learning rate. Following the setup of Kuhn & Bos (1993), $\xi \sim \mathcal{N}(0, \frac{2}{\beta})$ approximates the stochasticity noise of the optimization algorithm, with β corresponding to an inverse *temperature parameter*. The shape of the noise is assumed to be Gaussian by virtue of the central limit theorem. See Bottou et al. (1991); Mandt et al. (2017); Wu et al. (2020) for more details on modeling the stochasticity of SGD with Gaussian noise.

Macroscopic variables. The quantity of interest in this work is the average generalization error of the student determined by averaging the student's error over all possible input-target pairs of a **noiseless** teacher, as

$$\mathcal{L}_G := \frac{1}{2}\mathbb{E}_z\big[(y^* - \hat{y})^2\big]. \tag{5}$$

As shown in Bös et al. (1993), if $n, d \to \infty$ with a constant ratio $\frac{n}{d} < \infty$, Eq. 5 can be written as a function of two macroscopic scalar variables $R, Q \in \mathbb{R}$,

$$\mathcal{L}_G = \frac{1}{2}(1 + Q - 2R), \tag{6}$$

where,

$$R := \frac{1}{d}\boldsymbol{w}^T\mathrm{F}\hat{\boldsymbol{w}}, \qquad Q := \frac{1}{d}\hat{\boldsymbol{w}}^T\mathrm{F}^T\mathrm{F}\hat{\boldsymbol{w}}, \tag{7}$$

(See App. B.1 for Proof.)

Remark: Both R and Q have clear interpretations; R is the dot-product between the teacher's weights $\boldsymbol{w}$ and the student's *modulated* weights $\mathrm{F}\hat{\boldsymbol{w}}$, hence can be interpreted as the **alignment between the teacher and the student**. Similarly, Q can be interpreted as the **student's modulated norm**. The negative sign of R in Eq. 6 suggests that the larger R is, the smaller the generalization error gets. At the same time, Q appears with a positive sign suggesting the students with smaller (modulated) norm generalize better.

Note that both R and Q are functions of $\hat{\boldsymbol{w}}$, which itself is a function of training iteration t and the regularization coefficient λ. Therefore, from hereon, we denote the above quantities as $\mathcal{L}_G(t, \lambda)$, $R(t, \lambda)$, and $Q(t, \lambda)$.

2.2. Main Results

In this Section, we present our main analytical results, with Section 2.3 containing a sketch of our derivations. For brevity, here, we only present the results for $\sigma_\epsilon^2 = \lambda = 0$. See App. B for the general case and the detailed proofs.

General matrix F. Let $\mathrm{Z} := [\boldsymbol{z}^\mu]_{\mu=1}^n \in \mathbb{R}^{n\times d}$ and $\mathrm{X} := [\boldsymbol{x}^\mu]_{\mu=1}^n \in \mathbb{R}^{n\times d}$ denote the input matrices for the teacher and student such that $\mathrm{X} := \mathrm{ZF}$. For a general modulation matrix F, the input covariance matrix has the following singular value decomposition (SVD),

$$\mathrm{X}^T\mathrm{X} = \mathrm{F}^T\mathrm{Z}^T\mathrm{ZF} = \mathrm{V}\Lambda\mathrm{V}^T, \tag{8}$$

with Λ containing the singular values of the student's input covariance matrix. Solving the dynamics of exact gradient descent as in Eq. 4, we arrive at the following exact

analytical expressions for $R(t)$ and $Q(t)$,

$$R(t) = \frac{1}{d}\mathbf{Tr}\left(\mathrm{D}\right), \quad \text{where,} \quad \mathrm{D} := \mathrm{I} - \left[\mathrm{I} - \eta\Lambda\right]^t, \tag{9}$$

$$Q(t) = \frac{1}{d}\mathbf{Tr}\left(\mathrm{A}^T\mathrm{A}\right), \text{ where, } \mathrm{A} := \mathrm{F}\mathrm{V}\mathrm{D}\mathrm{V}^T\mathrm{F}^{-1}, \tag{10}$$

in which $\mathbf{Tr}(.)$ is the trace operator. (See App. B.2 for Proof.)

By plugging Eqs. 9 and 10 into Eq. 6, one obtains an exact expression for $\mathcal{L}_G(t)$. Unfortunately, Eqs. 9 and 10 are not straightforward to treat generally, and require the numerical evaluation of the singular values in Λ. Nevertheless, with some simple but informative assumptions on the modulation matrix F's structure, one can derive approximate solutions, as we now demonstrate.

Bipartite matrix F. We now study a case where F obeys the following Assumption.

Assumption 2.1. The modulation matrix, F, under a SVD, $\mathrm{F} := \mathrm{U}\Sigma\mathrm{V}^T$ has two sets of singular values such that the first p singular values are equal to σ_1 and the remaining $d-p$ singular values are equal to σ_2. We let the condition number of F to be denoted by $\kappa := \frac{\sigma_1}{\sigma_2} \geq 1$.

By employing the replica method of statistical physics (Gardner, 1988; Gardner & Derrida, 1988) and approximation of gradient descent dynamics with ridge regression, we derive closed-form expressions for $R(t)$ and $Q(t)$. To present the results, we first define the following auxiliary variables,

$$\alpha_1 := \frac{n}{p}, \qquad \alpha_2 := \frac{n}{d-p}, \tag{11}$$

$$\tilde{\lambda}_1 := \frac{d}{p}\underbrace{\frac{1}{\eta\sigma_1^2 t}}_{\text{time scaled by } \sigma_1^2}, \qquad \tilde{\lambda}_2 := \frac{d}{d-p}\underbrace{\frac{1}{\eta\sigma_2^2 t}}_{\text{time scaled by } \sigma_2^2}, \tag{12}$$

and also let, for $i \in \{1, 2\}$,

$$a_i = 1 + \frac{2\tilde{\lambda}_i}{(1-\alpha_i-\tilde{\lambda}_i) + \sqrt{(1-\alpha_i-\tilde{\lambda}_i)^2 + 4\tilde{\lambda}_i}}. \tag{13}$$

The scalar expression for $R(t)$ is then given by,

$$R(t) = R_1 + R_2, \quad \text{where,}$$
$$R_1 := \frac{n}{a_1 d}, \quad \text{and,} \quad R_2 := \frac{n}{a_2 d}. \tag{14}$$

Similarly, for $Q(t)$, we have, $Q(t) = Q_1 + Q_2$, where

$$Q_1 := \frac{b_1 b_2 c_2 + b_1 c_1}{1 - b_1 b_2}, \quad \text{and,} \quad Q_2 := \frac{b_1 b_2 c_1 + b_2 c_2}{1 - b_1 b_2}. \tag{15}$$

with ($i \in \{1, 2\}$),

$$b_i = \frac{\alpha_i}{a_i^2 - \alpha_i}, \quad c_i = 1 - 2R_i - \frac{n}{d}\frac{2 - a_i}{a_i}, \tag{16}$$

Plugging Eqs. 14 and 15 into Eq. 6, one obtains an (approximate) expression for $\mathcal{L}_G(t)$ as a function of the training time. (See App. B.3 for Proof.)

Remark: Eq. 12 indicates that the singular values of F, are directly multiplied by t. That implies that the learning speed of each feature is scaled by the magnitude of its corresponding singular value.

2.3. Sketch of derivations

In this Section, we sketch the key steps in the derivation of our main results. For the sake of simplicity, here again we only treat the case where $\sigma_\epsilon = \lambda = 0$. (See App. B for the general case and detailed Proofs.)

General matrix F: Exact dynamics. Recall the gradient descent update rule in Eq. 4. For the linear model defined in Eqs. 1-2, learning is governed by the following discrete-time dynamics,

$$\hat{\boldsymbol{w}}_t = \hat{\boldsymbol{w}}_{t-1} - \eta\nabla_{\hat{\boldsymbol{w}}_{t-1}}\mathcal{L}_T, \tag{17}$$

$$= \hat{\boldsymbol{w}}_{t-1} - \eta\left[-\mathrm{X}^T(y - \mathrm{X}\hat{\boldsymbol{w}}_{t-1})\right]. \tag{18}$$

With the assumption that $\hat{\boldsymbol{w}}_{t=0} = \mathbf{0}$, the dynamics admit the following exact closed-form solution,

$$\hat{\boldsymbol{w}}_t = \left(I - \left[I - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T y := \tilde{\boldsymbol{w}}(t). \tag{19}$$

With a SVD on $\mathrm{X}^T\mathrm{X}$, Eqs. 9-10 can then be obtained by substituting $\hat{\boldsymbol{w}}_t$ in Eq. 7. As a remark, note that one can recover the results of Advani & Saxe (2017) by setting $\mathrm{F} = \mathbb{I}$. In that case, the eigenvalues of $\mathrm{X}^T\mathrm{X}$ follow a Marchenko–Pastur distribution (Marchenko & Pastur, 1967).

Bipartite matrix F: Approximate dynamics. To employ the replica method, we first invoke the results in Eq. 9 of Solla (1995) and Kuhn & Bos (1993) which state that the equilibrium distribution of weights $\hat{\boldsymbol{w}}$ trained via SGD on a loss $\mathcal{L}(\hat{\boldsymbol{w}})$, follow the Gibbs-Boltzmann distribution, such that,

$$P(\hat{\boldsymbol{w}}) = \frac{1}{Z_\beta}e^{-\beta\mathcal{L}(\hat{\boldsymbol{w}})}, \tag{20}$$

in which $Z_\beta = \int \mathrm{d}\hat{\boldsymbol{w}}\exp(-\beta\mathcal{L}(\hat{\boldsymbol{w}}))$ is the partition function and β is called the *inverse temperature* and is inversely proportional to the stochasticity of SGD (see Eq. 4). Such distribution is a standard choice in statistical mechanics (see

page 53 of Engel & Van den Broeck (2001)). Intuitively, for small β, the distribution of $P(\hat{\boldsymbol{w}})$ is almost uniform, while as $\beta \to \infty$, $P(\hat{\boldsymbol{w}})$ becomes more concentrated around the minimum of the loss $\mathcal{L}(\hat{\boldsymbol{w}})$.

It is important to highlight that Eq. 20 describes the *equilibrium* distribution of the student network's weights, i.e., at the end of training ($t \to \infty$). However, we are interested in studying the trajectory of student's weights *during* the course of training, i.e., for finite t. To this end, we employ the connection between (continuous-time) SGD and L_2 regularization, as first quantified in Ali et al. (2019; 2020). Specifically, it states that the MSE loss of a linear regression model under stochastic gradient flow at time t is bounded from above by the end-of-training loss in the presence of ridge regression with an L_2 regularization coefficient $\lambda = 1/\eta t$. We note that while there is no guarantee that this bound is tight in general, we do observe that it matches the behavior of a wide range of numerical experiments extremely well (see Section 3).

Accordingly, we study the equilibrium distribution of the modified loss $\tilde{\mathcal{L}}(\hat{\boldsymbol{w}}, t)$, such that,

$$P(\hat{\boldsymbol{w}}) = \frac{1}{Z_{\beta,t}} e^{-\beta \tilde{\mathcal{L}}(\hat{\boldsymbol{w}},t)}, \quad \text{and,} \tag{21}$$

$$\tilde{\mathcal{L}}(\hat{\boldsymbol{w}}, t) := \frac{1}{2n} \sum_{\mu=1}^{n} (y^\mu - \hat{y}^\mu)^2 + \frac{1}{2}\left(\lambda + \frac{1}{\eta t}\right) ||\hat{\boldsymbol{w}}||_2^2. \tag{22}$$

See App. B.4 for proof.

To determine the *typical* generalization performance of students distributed according to $P(\hat{\boldsymbol{w}})$, one proceeds by computing the free-energy of the system as,

$$f := -\frac{1}{\beta d} \mathbb{E}_{w,\boldsymbol{z}} \left[\ln Z_{\beta,t} \right]. \tag{23}$$

Free-energy is a self-averaging property where its *typical/most probable* value coincides with its *average* over proper probability distributions (Engel & Van den Broeck, 2001). Therefore, to determine the typical values of R and Q, we extremize the free-energy w.r.t. those variables.

Due to the logarithm inside the expectation, analytical computation of Eq. 23 is intractable. However, the replica method (Mézard et al., 1987) allows us to tackle this through the following identity,

$$\mathbb{E}_{w,\boldsymbol{z}}[\ln Z_{\beta,t}] = \lim_{r \to 0} \frac{\mathbb{E}_{w,\boldsymbol{z}}[Z_{\beta,t}^r] - 1}{r}. \tag{24}$$

Computation of the free-energy via replica method and its subsequent extremization w.r.t R and Q, we arrive at Eqs. 14 and 15. See App. B.3 for more details.

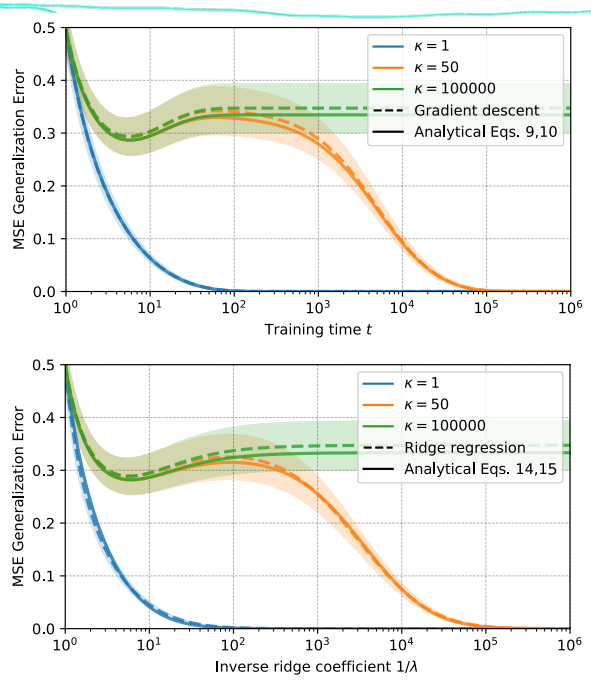

Figure 3. **Top:** Analytical results of Eqs. 9, 10 compared to gradient descent dynamics. The x-axis denotes the training time t. **Bottom:** Analytical results of scalar Eqs. 14, 15 compared to ridge regression dynamics. The x-axis denotes the inverse ridge (L2) coefficient $1/\lambda$. Analytical results closely match with empirical simulations. Consistent with Ali et al. (2019), ridge regression appears to reasonably approximate gradient descent dynamics. **Analysis:** With $\kappa = 1$, all the features are learned at the same rate (no double descent). $\kappa = 50$ corresponds to the case where a subset of features are learned 50 times faster than the rest and hence epoch-wise double descent is observed. Finally, $\kappa = 100000$ implies that a subset of of features are extremely slow to learn that practically do not get learned (typical overfitting).

To summarize, using the replica method, we are able to cast the high-dimensional dynamics of SGD into simple scalar equations governing R and Q and, consequently, the generalization error $\mathcal{L}_G$. While our analysis is limited to the specific teacher and student setup, this simple model already exhibits dynamics qualitatively similar to those observed in more complex networks, as we now illustrate.

3. Experimental Results

In this Section, we conduct numerical simulations to validate our analytical results and provide clear insights on the macroscopic dynamics of generalization. We also conduct experiments on real-world neural networks showing a close qualitative match between the generalization behavior of neural networks and our teacher-student setup. To ensure reproducibility, we include the complete source code in a `GitHub repository` as well as a `Colab notebook`.

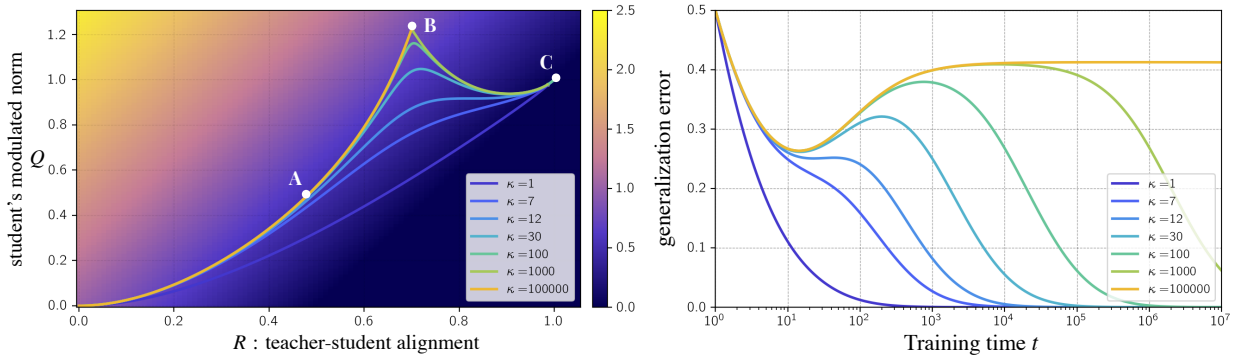

Figure 4. **Left**: Phase diagram of the generalization error as a function of $R(t)$ and $Q(t)$ (Eqs. 14 and 15). The generalization error for all pairs of $(R, Q) \in [0.0, 1.0] \times [0.0, 1.2]$ is contour-plotted in the background, with the best generalization performance being attained on the lower right part of the plot. The trajectories describe the evolution of $R(t)$ and $Q(t)$ as training proceeds. Each trajectory correspond to a different κ, the condition number of the modulation matrix F in Eq. 2. κ describes the ratio of the rates at which two sets of features are learned. **Right**: The corresponding generalization curves. **Analysis**: The trajectory with $\kappa = 1e5$ starts at the origin and advances towards point A (a descent in generalization error). Then by over-training, it converges to point B (an ascent). For the other trajectories with smaller κ, a first descent occurs up to the point A, then an ascent happens, but they no longer converge to point B. Instead, by further training, these trajectories converge to point C implying a second descent.

3.1. Analytical results compared with simulations

Through numerical simulations, we validate our analytical results presented in Section 2.2. Figure 3 depicts the comparisons for a teacher-student setup with $d = 100$, $p = 50$, and $n = 150$. Several similar experiments for different configurations are available in our provided notebook. It is observed that with $\kappa = 1$, the generalization error does not follow a double descent curve. Recall that $\kappa = 1$ implies that all the features are learned at the same rate. However, by increasing the value of κ, double descent curves are observed. Very large values of κ imply that some features are practically non-learnable and hence a typical overfitting curve is observed.

3.2. The Phase diagram

To further investigate the transition between the two phases of *classical single descent* and *double descent*, we explore the phase diagram. Recall that with Eq. 6, one can fully characterize the evolution of the generalization dynamics in terms of two scalar variables instead of the d-dimensional parameter space. R and Q presented in Eq. 7 are macroscopic variables where R represents **the alignment between the teacher and the student** and Q is the **student's (modulated) norm**. Hence, a better generalization performance is achieved with larger R and smaller Q.

The quantities R and Q are not free parameters and both depend on the training dynamics through Eqs. 14 and 15. Nevertheless, it is instructive to visualize the generalization error for all pairs of (R, Q). In Figure 4, we visualize the RQ-plane for $(R, Q) \in [0.0, 1.0] \times Q \in [0.0, 1.2]$. At the time of initialization, $(R, Q) = (0, 0)$ as the models are initialized at the origin. As training time proceeds, values of R and Q follow the depicted trajectories. In Figure 4, different trajectories correspond to different values of κ, the condition number of the modulation matrix F in Eq. 2. It is important to note that *the closer a trajectory is to the lower-right, the better the generalization error gets.*

The yellow curve corresponds to the case with large $\kappa = 1e5$, meaning that a subset of features are extremely slower than the others that practically do not get learned. In that case, generalization error exhibits traditional overfitting due to over-training. On the phase diagram, the yellow trajectory starts at $(0, 0)$ and moves towards Point A which has the lowest generalization error of this curve. Then as the training continues, Q increases and as $t \to \infty$ the trajectory lands at Point B which has the worse generalization error (highly-overfitted). Other curves follow the case of $\kappa = 1e5$ up to the vicinity of Point B, but then the trajectories slowly incline towards another fixed point, Point C signalling a second descent in the generalization error.

The phase diagram along with the corresponding generalization curves in Figure 4 illustrate that features that are learned on a faster time-scale are responsible for the initial conventional U-shaped generalization curve, while the second descent can be attributed to the features that are learned at a slower time-scale.

3.3. Qualitative comparison with ResNet on Cifar-10

We train a ResNet-18 (He et al., 2016) with layer widths $[64, 2 \times 64, 4 \times 64, 8 \times 64]$. We follow the training setup of Nakkiran et al. (2019); label noise with a probability 0.15

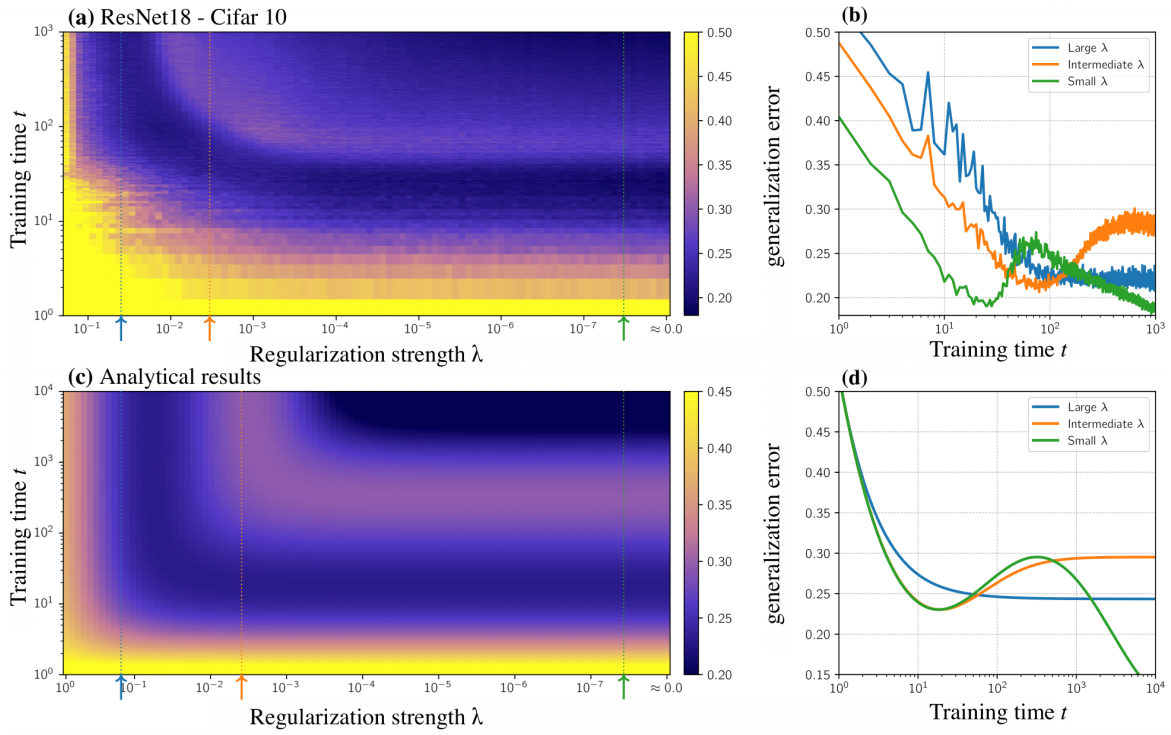

Figure 5. **A qualitative comparison between a ResNet-18 and our analytical results. (a)**: Heat-map of empirical generalization error (0-1 classification error) for the ResNet-18 trained on Cifar-10 with 15% label noise. X-axis denotes the inverse of weight-decay regularization strength and Y-axis represents the training time. **(c)**: Heat-map of the analytical generalization error (mean squared error) for the linear teacher-student setup with $\kappa = 100$, the condition number of the modulation matrix. **(b, d)**: Three slices of the heat-maps for large, intermediate, and small amounts of regularization. **Analysis**: As predicted by Eqs. 14 and 15, $\kappa = 100$ implies that a subset of features are learned 100 times faster that the rest. Intuitively, large amounts of regularization (↑) allow for the fast-learning features to be learned but cause overfitting. Intermediate levels of regularization (↑) result in a classical U-shaped generalization curve but prevent learning of slow features. Small amounts of regularization (↑) allow for both fast and slow features to be learned, leading to a double descent curve.

randomly assign an incorrect label to training examples. Noise is sampled only once before the training starts. We train using Adam (Kingma & Ba, 2014) with learning rate of $1e-4$ for 1K epochs. Experiments are averaged over 50 random seeds.

We conduct an experiment on the classification task of Cifar-10 (Krizhevsky et al., 2009) with varying amount of weight decay regularization strength λ. We monitor the generalization error (0-1 test error) during the course of training and visualize a heat-map of the generalization error for different λ's in Figure 5 (a).

We also conduct a similar experiment with the teacher-student setup presented in Section 2.1. We visualize a heat-map of the generalization error which is the mean squared error (MSE) over test distribution in Figure 5 (c). Particularly, we plot Eqs. 14 and 15 with a $\kappa = 100$. It is observed that in both experiments, a model with intermediate levels of regularization displays a typical overfitting behavior where the generalization error decreases first and then overfits. This is consistent with Eq. 61 of the appendix: The amount of regularization λ, is inversely proportional to the training time t implying that larger amounts of regularization act as early stopping.

3.4. Diminishing the temporary overfitting

The phase diagram in Figure 4 along with Eq. 6 suggest that an inflation in the value Q is responsible for the temporary overfitting observed in epoch-wise double descent. As an illustrative experiment, if we could diminish this temporary overfitting, we could expect to observe a *single descent* rather than a double descent curve. To that end, a natural solution is to penalize Q during training. To do that, we introduce the following lemma.

Lemma 3.1. *For a linear/linearized model, penalizing Q amounts to adding the following regularizer to the loss,*

$$\mathcal{L}_T \leftarrow \mathcal{L}_T + \alpha ||\hat{y}||^2, \tag{25}$$

previously introduced in Pezeshki et al. (2020). (See App. B.5 for Proof).

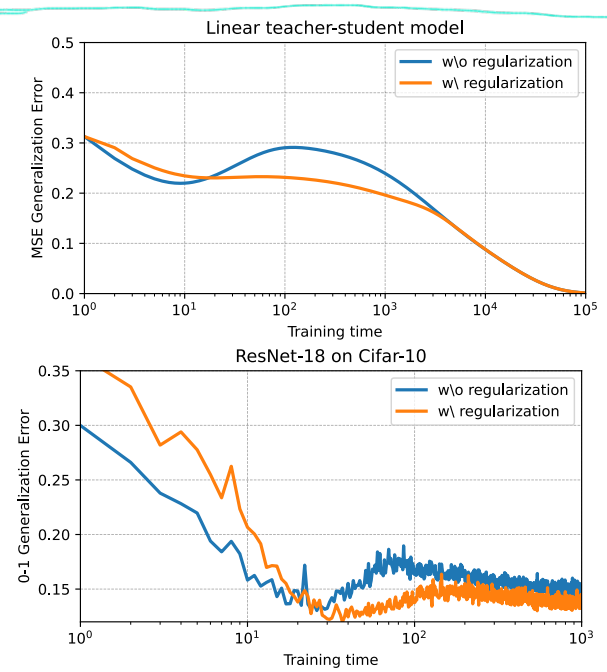

Figure 6. The effect of regularizing the quantity Q on the generalization curve. Two setups with (w/) and without (w/o) regularization are compared. Both the linear teacher-student model and a ResNet-18 on a binary Cifar-10 benefit from such regularization as the temporary overfitting is diminished. In accordance with Lemma 3.1, Q regularization is implemented by simply penalizing the norm of the model's output.

Figure 6 depicts the effect of this regularizer on the generalization curve. Both linear teacher-student model and ResNet-18 show curves in which the overfitting cusps are diminished. The ResNet experiment is on a binary classification version of the Cifar dataset.

We note that, for any linear model $\hat{y} = Xw$, the regularization $||\hat{y}||^2$ translates to an L2 regularization on the weights that is scaled by the input covariance matrix, as $||\hat{y}||^2 = w^T X^T X w$. Therefore, such regularization slows down the learning along the direction of faster features and hence attempts to equalize the learning scale of different features. We should highlight that mitigating double descent is not the purpose of our work and this experiment is presented to support that the findings from a linear model can still carry over to non-linear networks.

4. Related Work and Discussion

Although the term *double descent* has been introduced rather recently (Belkin et al., 2019a), similar behaviors had already been observed and studied in several decades-old works form a statistical physics perspective (Krogh & Hertz, 1992; Opper, 1995; Opper & Kinzel, 1996; Bös, 1998). More recently, these behaviors have been investigated in the context of modern machine learning, both from an empirical (Amari et al., 2020; Yang et al., 2020) and theoretical perspectives (Geiger et al., 2019; d'Ascoli et al., 2021; Geiger et al., 2020).

Hastie et al. (2019); Advani & Saxe (2017); Belkin et al. (2020) use random matrix theory (RMT) tools to characterize the asymptotic generalization behavior of over-parameterized linear and random feature models. Mei & Montanari (2019) extend the same analysis to a random feature model and theoretically derive the model-wise double descent curve for a model with Tikhonov regularization. Jacot et al. (2020) also study double descent in ridge estimators and show an equivalence to kernel ridge regression.

While most of the related work study the non-monotonicity of the generalization error as a function of the model size or sample size, Nakkiran et al. (2019) introduced the epoch-wise double descent, where the double descent occurs as the training time increases. There has been limited work on studying of epoch-wise double descent. Very recently, Heckel & Yilmaz (2020) and Stephenson & Lee (2021) have focused on finding the roots of this phenomenon.

Heckel & Yilmaz (2020) provides *upper bounds* on the risk of single and two layer models in a regression setting where the input data has distinct feature variances. Heckel & Yilmaz (2020) demonstrate that a superposition of two or more bias-variance tradeoff curves leads to epoch-wise double descent. The authors also show that different layers of the network are learned at different epochs. For that reason, epoch-wise double descent can be eliminated by appropriate selection of learning rates for individual weights. Stephenson & Lee (2021) arrive at similar conclusions. A data model is constructed so that the noise is explicitly added *only* to the fast-learning features while slow-learning features remain noise-free. Consequently, the noisy features form a U-shaped generalization curve while noiseless but slow features are responsible for the second descent.

Our findings and those of Heckel & Yilmaz (2020) and Stephenson & Lee (2021) reinforce one another with a common central finding that the epoch-wise double descent results from different features/layers being learned at different time-scales. However, we also highlight that both Heckel & Yilmaz (2020) and Stephenson & Lee (2021) use tools from random matrix theory to study distinct data models from our teacher-student setup. We study a similar phenomenon by leveraging the replica method from statistical physics to characterize the generalization behavior using a set of informative macroscopic parameters. The key novel contribution from our approach is the derivation of the macroscopic quantities R and Q (see Eq. 7) which track teacher-student alignment, and the student's modulated norm, respectively. Crucially, these quantities can be used to study other generalization phenomena and/or to modify the learning dynamics

via their explicit regularization as illustrated in Section 3.4.

We believe our framework sets the stage for further understanding of generalization dynamics beyond the double descent. A future direction to study is a case in which the first descent is strong enough to bring down the training loss to zero such that learning slower features is practically impossible (Pezeshki et al., 2020) or happens after a very large number of epochs (Power et al., 2021). *Grokking* is an instance of such behavior reported by Power et al. (2021) in which the model abruptly learns to perfectly generalize but long after the training loss has reached very small values.

Finally, we note that while our simple teacher-student setup exhibits the epoch-wise double descent, its simplicity introduces several limitations. Studying finer details of the dynamics of neural networks requires more precise, nonlinear, and multi-layered models, which introduce novel challenges that remain to be studied in future work.

Acknowledgments and Disclosure of Funding

The authors are grateful to Samsung Electronics Co., Ltd., CIFAR, IVADO, and the Canada Research Chair program for funding support, and Calcul Québec and Compute Canada for providing us with the computing resources. We would further like to acknowledge the significance of discussions and supports from Reyhane Askari Hemmat, Faruk Ahmed, David Yu-Tung Hui, and Aristide Baratin.

References

Advani, M. S. and Saxe, A. M. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1370–1378. PMLR, 2019.

Ali, A., Dobriban, E., and Tibshirani, R. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pp. 233–244. PMLR, 2020.

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

Ba, J., Erdogdu, M., Suzuki, T., Wu, D., and Zhang, T. Generalization of two-layer neural networks: An asymptotic viewpoint. In *International conference on learning representations*, 2019.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks, 2020.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019a.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019b.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Bender, C. M. and Orszag, S. A. *Advanced mathematical methods for scientists and engineers I: Asymptotic methods and perturbation theory*. Springer Science & Business Media, 2013.

Bös, S. Statistical mechanics approach to early stopping and weight decay. *Physical Review E*, 58(1):833, 1998.

Bös, S., Kinzel, W., and Opper, M. Generalization ability of perceptrons with continuous outputs. *Physical Review E*, 47(2):1384, 1993.

Bottou, L. et al. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

Chen, L., Min, Y., Belkin, M., and Karbasi, A. Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*, 2020.

Chizat, L. and Bach, F. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pp. 1305–1338. PMLR, 2020.

d'Ascoli, S., Sagun, L., and Biroli, G. Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*, 2020.

d'Ascoli, S., Gabrié, M., Sagun, L., and Biroli, G. On the interplay between data structure and loss function in classification problems. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

Edwards, S. F. and Anderson, P. W. Theory of spin glasses. *Journal of Physics F: Metal Physics*, 5(5):965, 1975.

Engel, A. and Van den Broeck, C. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Gardner, E. The space of interactions in neural network models. *Journal of physics A: Mathematical and general*, 21(1):257, 1988.

Gardner, E. and Derrida, B. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

Gardner, E. and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

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

Gerace, F., Loureiro, B., Krzakala, F., Mézard, M., and Zdeborová, L. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Goldt, S., Reeves, G., Mézard, M., Krzakala, F., and Zdeborová, L. The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv e-prints*, pp. arXiv–2006, 2020.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*. MIT press Cambridge, 2016.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heckel, R. and Yilmaz, F. F. Early stopping in deep networks: Double descent and how to eliminate it. *arXiv preprint arXiv:2007.10099*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Jacot, A., Simsek, B., Spadaro, F., Hongler, C., and Gabriel, F. Implicit regularization of random feature models. In *International Conference on Machine Learning*, pp. 4631–4640. PMLR, 2020.

Kabashima, Y., Wadayama, T., and Tanaka, T. A typical reconstruction limit for compressed sensing based on lp-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Krogh, A. and Hertz, J. A. Generalization in a linear perceptron in the presence of noise. *Journal of Physics A: Mathematical and General*, 25(5):1135, 1992.

Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

Kuhn, R. and Bos, S. Statistical mechanics for neural networks with continuous-time dynamics. *Journal of Physics A: Mathematical and General*, 26(4):831, 1993.

Kunin, D., Sagastuy-Brena, J., Ganguli, S., Yamins, D. L., and Tanaka, H. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. *arXiv preprint arXiv:2012.04728*, 2020.

Le Cun, Y., Kanter, I., and Solla, S. A. Eigenvalues of covariance matrices: Application to neural-network learning. *Physical Review Letters*, 66(18):2396, 1991.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *arXiv preprint arXiv:1704.04289*, 2017.

Marchenko, V. A. and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Mei, S. and Montanari, A. The generalization error of random features regression: precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Mézard, M., Parisi, G., and Virasoro, M. *Spin glass theory and beyond: an introduction to the Replica Method and its applications*, volume 9. World Scientific Publishing Company, 1987.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: where bigger models and more data hurt. *ICLR 2020, arXiv preprint arXiv:1912.02292*, 2019.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Neyshabur, B., Tomioka, R., and Srebro, N. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5947–5956, 2017.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. Towards understanding the role of over-parametrization in generalization of neural networks, 2018.

Opper, M. Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pp. 922–925, 1995.

Opper, M. and Kinzel, W. Statistical mechanics of generalization. In *Models of neural networks III*, pp. 151–209. Springer, 1996.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*, 2020.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets. In *ICLR MATH-AI Workshop*, 2021.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

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

Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

Solla, S. A. A bayesian approach to learning in neural networks. *International Journal of Neural Systems*, 6: 161–170, 1995.

Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., and Wyart, M. A jamming transition from under-to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, 52 (47):474001, 2019.

Stephenson, C. and Lee, T. When and how epochwise double descent happens. *arXiv preprint arXiv:2108.12006*, 2021.

Vapnik, V. N. *The nature of statistical learning theory*. Wiley, New York, 1st edition, September 1998. ISBN 978-0-471-03003-4.

Wu, J., Hu, W., Xiong, H., Huan, J., Braverman, V., and Zhu, Z. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pp. 10367–10376. PMLR, 2020.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Zhang, X., Xiong, H., and Wu, D. Rethink the connections among generalization, memorization and the spectral bias of dnns. *arXiv preprint arXiv:2004.13954*, 2020.

A. Further Related Work and Discussion

If we consider plots where the generalization error on the y-axis is plotted against other quantities on the x-axis, we find earlier works that have identified double descent behavior for quantities such as the number of parameters, the dimensionality of the data, the number of training samples, or the training time on the x-axis. In this paper, we studied epoch-wise double descent, *i.e.* we plot the training time t, or the number of training epochs, on the x-axis. Literature displaying double descent phenomena in generalization behavior w.r.t. other quantities do so in the limit of $t \to \infty$.

From a random matrix theory perspective, Le Cun et al. (1991); Hastie et al. (2019); Advani & Saxe (2017), and Belkin et al. (2020) are among works which have analytically studied the spectral density of the Hessian matrix. According to their analyses, at intermediate levels of complexity, the presence of small but non-zero eigenvalues in the Hessian matrix results in high generalization error as the inverse of the Hessian is calculated for the pseudo-inverse solution.

Neyshabur et al. (2014) demonstrated that over-parameterized networks does not necessarily overfit thus suggesting the need of a new form of measure of model complexity other than network size. Subsequently, Neyshabur et al. (2018) suggest a novel complexity measure based on unit-wise capacities which correlates better with the behavior of test error with increasing network size. Chizat & Bach (2020) study the global convergence and superior generalization behavior of infinitely wide two-layer neural networks with logistic loss. Goldt et al. (2020) make use of the Gaussian Equivalence Theorem to study the generalization performance of two-layer neural networks and kernel models trained on data drawn from pre-trained generative models. Bai & Lee (2020) investigated the gap between the empirical performance of over-parameterized networks and their NTK counterparts, first proposed by Jacot et al. (2018).

From the perspective of bias/variance trade-off, Geman et al. (1992), and more recently, Neal et al. (2018) empirically observe that while bias is monotonically decreasing, variance could be decreasing too or unimodal as the number of parameters increases, thus manifesting a double descent generalization curve. Hastie et al. (2019) analytically study the variance. More recently, Yang et al. (2020) provides a new bias/variance decomposition of bias exhibiting double descent in which the variance follows a bell-shaped curve. However, the decrease in variance as the model size increases remains unexplained. For high dimensional regression with random features, d'Ascoli et al. (2020) provides an asymptotic expression for the bias/variance decomposition and identifies three sources of variance with non-monotonous behavior as the model size or dataset size varies. d'Ascoli et al. (2020) also employs the analysis of random feature models and identifies two forms of overfitting which leads to the so-called sample-wise triple descent. More recently, Chen et al. (2020) show that as a result of the interaction between the data and the model, one may design generalization curves with multiple descents.

From a statistical physics perspective, Opper (1995); Bös et al. (1993); Bös (1998); Opper & Kinzel (1996) are among the first studies which theoretically observe sample-wise double-descent in a ridge regression setup where the solution is obtained by the pseudo-inverse method. Most of these studies employ the "Gardner analysis" (Gardner, 1988; Gardner & Derrida, 1988; 1989) for models where the number of parameters and the dimensionality of data are coupled and hence the observed form of double descent is different from that observed in deep neural networks. A beautiful extended review of this line of work is provided in Engel & Van den Broeck (2001). Among recent works, Gerace et al. (2020) also apply the Gardner analysis but to a novel generalized data generating process called the hidden manifold model and derive the model-wise double-descent equations analytically.

Finally, recall that towards providing an explanation for the epoch-wise double descent, we argue that *the epoch-wise double descent can be attributed to different features being learned at different time-scales*, resulting in a non-monotonous generalization curve. In relation to the aspect of different feature learning scales, Rahaman et al. (2019) had observed that DNNs have a tendency towards learning simple target functions first that can allow for good generalization behavior of various data samples. Pezeshki et al. (2020) also identify and provide explanation for a feature learning imbalance exhibited by over-parameterized networks trained via gradient descent on cross-entropy loss, with the networks learning only a subset of the full feature spectrum over training. More recently though, Zhang et al. (2020), show that certain DNNs models prioritize learning high-frequency components first followed by the learning of slow but informative features, leading to the second descent of the test error as observed in epoch-wise double descent.

On the difference between model-wise and epoch-wise double descent curves. In accordance with its name, model-wise double descent (in the test error) occurs due to an increase in model-size (number of its parameters), i.e., as the model transitions from an under-parameterized to an over-parameterized regime. A variety of works have tried to understand this phenomenon from the lens of implicit regularization (Neyshabur et al., 2014) or defining novel complexity measures (Neyshabur et al., 2017). On the other hand, epoch-wise double descent (in the test error) as treated in our work, is observed

to occur for both over-parameterized (Nakkiran et al., 2019) and under-parameterized (Heckel & Yilmaz, 2020) setups. As found in our work along with the latter reference, this phenomenon seems to be a result of different feature learning speeds rather than the extent of model parameterization. The overlap of the test-error contributions from the different weights with varying scales of learning henceforth leads to a non-monotonous evolution of the model test error as exemplified by epoch-wise double descent.

We also note that the peak in model-wise double descent is associated with the model's capacity to perfectly interpolate the data, we do not think an analogous notion exists for the case of epoch-wise double descent. Our understanding of the peak in the latter is that it corresponds to a training time configuration whereby a subclass of features are already learnt (due to a larger associated signal-to-noise-ratio) and are being overfitted upon to fit the target. As training proceeds further, the remaining set of features are eventually learnt thus allowing for a lowering of the test error.

On the link to complex networks. Generally, exact study of complex neural networks is often intractable. A common practice is to study a simpler system that conserves key attributes and then validate the findings on the original complex system. In this work, we build on the same established practices: we propose a simple linear model with two key advantages, a) it can be solved analytically, b) exhibits double descent, the property of interest. Subsequently, our experiments support the extension of our findings and intuitions to complex neural networks.

B. Technical Proofs

B.1. The generalization error as a function of R and Q (Eq. 6)

Recall that the teacher is the data generator and is defined as,

$$y := y^* + \epsilon, \qquad y^* := \boldsymbol{z}^T \boldsymbol{W}, \qquad z_i \sim \mathcal{N}(0, \frac{1}{\sqrt{d}}), \tag{26}$$

where $\boldsymbol{z} \in \mathbb{R}^d$ is the teacher's input and $y^*, y \in \mathbb{R}$ are the teacher's noiseless and noisy outputs, respectively. $\boldsymbol{W} \in \mathbb{R}^d$ represents the (fixed) weights of the teacher and $\epsilon \in \mathbb{R}$ is the label noise.

While the student network is defined as,

$$\hat{y} := \boldsymbol{x}^T \hat{\boldsymbol{W}}, \qquad s.t. \qquad \boldsymbol{x} := \mathrm{F}^T \boldsymbol{z}, \tag{27}$$

where the matrix $\mathrm{F} \in \mathbb{R}^{d \times d}$ is a predefined and fixed modulation matrix regulating the student's access to the true input $\boldsymbol{z}$.

The average generalization error of the student, determined by averaging the student's error over all possible input configurations and label noise realizations is given by,

$$\mathcal{L}_G := \frac{1}{2} \mathbb{E}_{\boldsymbol{z},\epsilon} \big[(y^* - \hat{y} + \epsilon)^2 \big], \tag{28}$$

in which the variables $(y^*, \hat{y})$ form a bi-variate Gaussian distribution with zero mean and a covariance of,

$$\Sigma = \begin{bmatrix} < y^*, y^* >_{\boldsymbol{z}} & < y^*, \hat{y} >_{\boldsymbol{z}} \\ < y^*, \hat{y} >_{\boldsymbol{z}} & < \hat{y}, \hat{y} >_{\boldsymbol{z}} \end{bmatrix} = \begin{bmatrix} 1 & R \\ R & Q \end{bmatrix}, \tag{29}$$

Here,

$$R := \mathbb{E}_{\boldsymbol{z}}[y^* \hat{y}] = \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{W}^T \boldsymbol{z} \boldsymbol{z}^T \mathrm{F} \hat{\boldsymbol{W}}] = \frac{1}{d} \boldsymbol{W}^T \mathrm{F} \hat{\boldsymbol{W}}, \quad \text{and,} \tag{30}$$

$$Q := \mathbb{E}_{\boldsymbol{z}}[\hat{y}^T \hat{y}] = \mathbb{E}_{\boldsymbol{z}}[\hat{\boldsymbol{W}}^T \mathrm{F}^T \boldsymbol{z} \boldsymbol{z}^T \mathrm{F} \hat{\boldsymbol{W}}] = \frac{1}{d} \hat{\boldsymbol{W}}^T \mathrm{F}^T \mathrm{F} \hat{\boldsymbol{W}}. \tag{31}$$

Utilizing this, Eq. 28 can be expressed as,

$$\mathcal{L}_G := \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[(y^* - \hat{y} + \epsilon)^2\right], \tag{32}$$

$$= \frac{1}{2}\mathbb{E}_{\tilde{y}^*,\tilde{\hat{y}}}\left[(\tilde{y}^* - (R\tilde{y}^* + \sqrt{Q-R^2}\tilde{\hat{y}}) + \epsilon)^2\right], \tag{33}$$

$$= \frac{1}{2}(1 + \epsilon^2 + Q - 2R). \tag{34}$$

Additionally, we note that expectation w.r.t. a Gaussian variable x is defined as,

$$\mathbb{E}_x[f(x)] := \int_{-\infty}^{+\infty} \frac{dx}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) f(x). \tag{35}$$

B.2. The general case exact dynamics (Eqs. 9-10)

Recall that to train our student network, we use gradient descent (GD) on the regularized mean-squared loss, evaluated on the n training examples as,

$$\mathcal{L}_T := \frac{1}{2n}\sum_{\mu=1}^{n}(y^\mu - \hat{y}^\mu)^2 + \frac{\lambda}{2}||\hat{\boldsymbol{W}}||_2^2, \tag{36}$$

where $\lambda \in [0, \infty)$ is the regularization coefficient.

The minimum of the loss function, denoted by $\hat{\boldsymbol{W}}_{\text{gd}}$, is achieved at,

$$\nabla_{\hat{\boldsymbol{W}}}\mathcal{L}_T = 0 \Rightarrow \nabla_{\hat{\boldsymbol{W}}}\left[\frac{1}{2}||\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{W}}||_2^2 + \frac{\lambda}{2}||\hat{\boldsymbol{W}}||_2^2\right] = 0 \tag{37}$$

$$\Rightarrow -\mathrm{X}^T(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{W}}_{\text{gd}}) + \lambda\hat{\boldsymbol{W}}_{\text{gd}} = 0 \tag{38}$$

$$\Rightarrow \hat{\boldsymbol{W}}_{\text{gd}} := (\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I})^{-1}\mathrm{X}^T y. \tag{39}$$

Additionally, the exact dynamics under gradient-descent, correspond to,

$$\begin{aligned}
\hat{\boldsymbol{W}}_t &= \hat{\boldsymbol{W}}_{t-1} - \eta\nabla_{\hat{\boldsymbol{W}}_{t-1}}\mathcal{L}_T, \\
&= \hat{\boldsymbol{W}}_{t-1} - \eta\big[-\mathrm{X}^T(\boldsymbol{y} - \mathrm{X}\hat{\boldsymbol{W}}_{t-1}) + \lambda\hat{\boldsymbol{W}}_{t-1}\big] \\
&= (1-\eta\lambda)\hat{\boldsymbol{W}}_{t-1} - \eta\mathrm{X}^T\mathrm{X}\hat{\boldsymbol{W}}_{t-1} + \eta\mathrm{X}^T\boldsymbol{y}, \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]\hat{\boldsymbol{W}}_{t-1} + \eta\mathrm{X}^T\boldsymbol{y}, \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]\hat{\boldsymbol{W}}_{t-1} + \eta(\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I})(\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I})^{-1}\mathrm{X}^T\boldsymbol{y}, \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]\hat{\boldsymbol{W}}_{t-1} + \eta(\mathrm{X}^T\mathrm{X} + \lambda\mathrm{I})\hat{\boldsymbol{W}}_{\text{gd}}, \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]\hat{\boldsymbol{W}}_{t-1} + (\eta\mathrm{X}^T\mathrm{X} + \eta\lambda\mathrm{I})\hat{\boldsymbol{W}}_{\text{gd}}, \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]\hat{\boldsymbol{W}}_{t-1} + (\eta\mathrm{X}^T\mathrm{X} + (\eta\lambda - 1)\mathrm{I})\hat{\boldsymbol{W}}_{\text{gd}} + \hat{\boldsymbol{W}}_{\text{gd}},
\end{aligned} \tag{40}$$

which leads to,

$$\begin{aligned}
\hat{\boldsymbol{W}}_t - \hat{\boldsymbol{W}}_{\text{gd}} &= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}](\hat{\boldsymbol{W}}_{t-1} - \hat{\boldsymbol{W}}_{\text{gd}}), \\
&= [(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}]^t(\hat{\boldsymbol{W}}_0 - \hat{\boldsymbol{W}}_{\text{gd}}).
\end{aligned} \tag{41}$$

Assuming $\hat{\boldsymbol{W}}_0 = 0$, we arrive at the following closed-form equation,

$$\hat{\boldsymbol{W}}_t = \left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)\hat{\boldsymbol{W}}_{\text{gd}}, \tag{42}$$

where $\hat{\boldsymbol{W}}_{\text{gd}}$ is defined in Eq 39.

Now back to definition of R in Eq. 30 and by substitution of Eq. 42, we have,

$$\begin{aligned}
R(t) :&= \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\hat{\boldsymbol{W}}_t, \\
&= \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)\hat{\boldsymbol{W}}_{\text{gd}}, \\
&= \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)(\mathrm{X}^T\mathrm{X}+\lambda\mathrm{I})^{-1}\mathrm{X}^T\boldsymbol{y}, \\
&= \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\mathrm{V}\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\Lambda\right]^t\right)(\Lambda+\lambda\mathrm{I})^{-1}\mathrm{V}^T\mathrm{X}^T\boldsymbol{y}, \quad (\mathrm{X}^T\mathrm{X} = \mathrm{V}\Lambda\mathrm{V}^T) \\
&= \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\mathrm{V}\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\Lambda\right]^t\right)(\Lambda+\lambda\mathrm{I})^{-1}(\Lambda\mathrm{V}^T\mathrm{F}^{-1}\boldsymbol{W} + \Lambda^{\frac{1}{2}}\boldsymbol{\epsilon}), \\
&= \boxed{\frac{1}{d}\mathbf{Tr}\left[\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\Lambda\right]^t\right)\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\right]}.
\end{aligned} \tag{43}$$

Similarly for Q, let $D := \left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\Lambda\right]^t\right)$, then we have,

$$\begin{aligned}
Q(t) :&= \frac{1}{d}\hat{\boldsymbol{W}}^T\mathrm{F}^T\mathrm{F}\hat{\boldsymbol{W}}, \\
&= \frac{1}{d}\hat{\boldsymbol{W}}_{\text{gd}}^T\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)\mathrm{F}^T\mathrm{F}\left(\mathrm{I} - \left[(1-\eta\lambda)\mathrm{I} - \eta\mathrm{X}^T\mathrm{X}\right]^t\right)\hat{\boldsymbol{W}}_{\text{gd}}, \\
&= \frac{1}{d}\hat{\boldsymbol{W}}_{\text{gd}}^T\mathrm{V}\mathrm{D}\mathrm{V}^T\mathrm{F}^T\mathrm{F}\mathrm{V}\mathrm{D}\mathrm{V}^T\hat{\boldsymbol{W}}_{\text{gd}}, \\
&= \frac{1}{d}\hat{\boldsymbol{W}}_{\text{gd}}^T\mathrm{V}\mathrm{D}\tilde{\mathrm{F}}^T\tilde{\mathrm{F}}\mathrm{D}\mathrm{V}^T\hat{\boldsymbol{W}}_{\text{gd}}, \qquad (\tilde{\mathrm{F}} := \mathrm{F}\mathrm{V}, \mathrm{X} = \mathrm{U}\Lambda^{1/2}\mathrm{V}^T, \tilde{\boldsymbol{\epsilon}} := \mathrm{U}^T\boldsymbol{\epsilon}) \\
&= \frac{1}{d}(\boldsymbol{W}^T\mathrm{F}^{-1^T}\mathrm{V} + \Lambda^{-1/2}\tilde{\epsilon})\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\mathrm{D}\tilde{\mathrm{F}}^T\tilde{\mathrm{F}}\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}(\mathrm{V}^T\mathrm{F}^{-1}\boldsymbol{W} + \Lambda^{-1/2}\tilde{\boldsymbol{\epsilon}}), \\
&= \frac{1}{d}(\boldsymbol{W}^T\tilde{\mathrm{F}}^{-1^T} + \Lambda^{-1/2}\tilde{\boldsymbol{\epsilon}})\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\mathrm{D}\tilde{\mathrm{F}}^T\tilde{\mathrm{F}}\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}(\tilde{\mathrm{F}}^{-1}\boldsymbol{W} + \Lambda^{-1/2}\tilde{\boldsymbol{\epsilon}}), \\
&= \frac{1}{d}\boldsymbol{W}^T\tilde{\mathrm{F}}^{-1^T}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\mathrm{D}\tilde{\mathrm{F}}^T\tilde{\mathrm{F}}\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\tilde{\mathrm{F}}^{-1}\boldsymbol{W}, \\
&\quad + \frac{1}{d}\Lambda^{-1/2}\tilde{\boldsymbol{\epsilon}}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\mathrm{D}\tilde{\mathrm{F}}^T\tilde{\mathrm{F}}\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\Lambda^{-1/2}\tilde{\boldsymbol{\epsilon}}, \\
&= \boxed{\frac{1}{d}\mathbf{Tr}\left[\mathrm{A}^T\mathrm{A}\right] + \frac{\sigma_\epsilon^2}{d}\mathbf{Tr}\left[\mathrm{B}^T\mathrm{B}\right]}
\end{aligned} \tag{44}$$

where,

$$\mathrm{A} := \tilde{\mathrm{F}}\,\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\tilde{\mathrm{F}}^{-1} \quad \text{and,} \quad \mathrm{B} := \tilde{\mathrm{F}}\,\mathrm{D}\frac{\Lambda}{\Lambda+\lambda\mathrm{I}}\Lambda^{-\frac{1}{2}}. \tag{45}$$

B.3. Special case of approximate dynamics (Eqs. 14 and 15)

Recall that the teacher and student are defined as,

$$y := y^* + \epsilon, \qquad y^* := \boldsymbol{z}^T\boldsymbol{W}, \qquad \hat{y} := \boldsymbol{x}^T\hat{\boldsymbol{W}}, \qquad \boldsymbol{x} := \mathrm{F}^T\boldsymbol{z}, \tag{46}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is the label noise, F is the modulation matrix, and $||\boldsymbol{z}||_2^2 = ||\boldsymbol{W}||_2^2 = 1$.

The training and generalization losses are defined as,

$$\mathcal{L}_T := \frac{1}{2n}\sum(\hat{y}-y)^2 + \frac{\lambda}{2}||\hat{\boldsymbol{W}}||_2^2, \qquad \mathcal{L}_G := \frac{1}{2}\mathbb{E}_{\boldsymbol{z},\epsilon}[(\hat{y}-y)^2]. \tag{47}$$

According to Eq. 6, the generalization loss can be written in terms of two scalar variables R and Q,

$$\mathcal{L}_G = \frac{1}{2}(1+\sigma_\epsilon^2+Q-2R), \quad \text{where,} \tag{48}$$

$$R := \mathbb{E}_{\boldsymbol{z}}[y^*\hat{y}] = \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{W}^T\boldsymbol{z}\boldsymbol{z}^T\mathrm{F}\hat{\boldsymbol{W}}] = \frac{1}{d}\boldsymbol{W}^T\mathrm{F}\hat{\boldsymbol{W}}, \quad \text{and,} \tag{49}$$

$$Q := \mathbb{E}_{\boldsymbol{z}}[\hat{y}\hat{y}] = \mathbb{E}_{\boldsymbol{z}}[\hat{\boldsymbol{W}}^T\mathrm{F}^T\boldsymbol{z}\boldsymbol{z}^T\mathrm{F}\hat{\boldsymbol{W}}] = \frac{1}{d}\hat{\boldsymbol{W}}^T\mathrm{F}^T\mathrm{F}\hat{\boldsymbol{W}}. \tag{50}$$

In the following, we next determine the most probable values of the above scalar entities, from statistical perspective.

Application of t steps of GD on $\mathcal{L}_T$ results in the following distribution for the student's weights:

$$P(\hat{\boldsymbol{W}},t) = \frac{1}{Z_{\beta,t}}e^{-\beta\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}},t)}, \tag{51}$$

in which $\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}},t)$ is a modified loss that dictates the distribution of student weights $\hat{\boldsymbol{W}}$ upon t^{th} iterations of GD on the original loss $\mathcal{L}_T(\hat{\boldsymbol{W}})$, while β corresponds to an (inverse) temperature parameter of our student weight distribution.

In Eq. 51, $Z_{\beta,t}$ is the partition function which is defined as,

$$Z_{\beta,t} = \frac{\int_{-\infty}^{\infty}\prod_{i=1}^{d}\mathrm{d}(\hat{\boldsymbol{W}}_i)\delta\Big(\frac{1}{d}\hat{\boldsymbol{W}}_i^T\mathrm{F}^T\mathrm{F}\hat{\boldsymbol{W}}_i-Q_0\Big)e^{-\beta\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}},t)}}{\int_{-\infty}^{\infty}\prod_{i=1}^{d}\mathrm{d}(\hat{\boldsymbol{W}}_i)\delta\Big(\frac{1}{d}\hat{\boldsymbol{W}}_i^T\mathrm{F}^T\mathrm{F}\hat{\boldsymbol{W}}_i-Q_0\Big)}, \tag{52}$$

in which, Q_0 can be perceived to be a target norm the student weights $\hat{\boldsymbol{W}}$ are being constrained to and d is the dimensionality of the data.

We are now interested in finding R and Q of the typical (most probable) students. Therefore, it suffices to find the students that dominate the partition function (or more precisely the free-energy). The free-energy is defined as,

$$f := -\frac{1}{\beta d}\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}\big[\ln Z_{\beta,t}\big], \tag{53}$$

where $\boldsymbol{W}$ and $\boldsymbol{z}$ are the teacher's weight and input, respectively.

Due to the logarithm inside the expectation, analytical computation of Eq. 53 is intractable. However, the replica method (Mézard et al., 1987) allows us to tackle this through the following identity,

$$\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[\ln Z_{\beta,t}] = \lim_{r\to 0}\frac{\mathbb{E}_{\boldsymbol{W},\boldsymbol{z}}[Z_{\beta,t}^r]-1}{r}. \tag{54}$$

Case 1: F = I. As a first step, we first study a case where $\mathrm{F}=\mathrm{I}$. In that case, as derived in Bös (1998), Eq. 53 can be simplified to,

$$-\beta f = \frac{1}{2}\frac{Q-R^2}{Q_0-Q}+\frac{1}{2}\ln(Q_0-Q)-\frac{n}{2d}\ln[1+\beta(Q_0-Q)]-\frac{n\beta}{2d}\frac{G-2HR+Q}{1+\beta(Q_0-Q)}, \tag{55}$$

in which the scalar variables G and H are defined as,

$$H := \mathbb{E}_{y^*,\epsilon}[y^*y] = \mathbb{E}_{y^*}[y^*(y^*+\epsilon)] = 1, \tag{56}$$

$$G := \mathbb{E}_{y^*,\epsilon}[yy] = \mathbb{E}_{y^*}[(y^*+\epsilon)(y^*+\epsilon)] = 1+\sigma_\epsilon^2. \tag{57}$$

At this point, in order to find the most probable students, one can extremize the free-energy $f(R,Q,Q_0)$ in Eq. 55. The solution to this extremisation is derived in Bös et al. (1993) and reads,

$$\nabla_R f = 0 \quad \Rightarrow \quad R = \frac{n}{d}\frac{1}{a}, \tag{58}$$

$$\nabla_Q f = 0 \quad \Rightarrow \quad Q = \frac{n}{d}\frac{1}{a^2-n/d}\left(G-\frac{n}{d}\frac{2-a}{a}\right), \tag{59}$$

$$\nabla_{Q_0} f = 0 \quad \Rightarrow \quad a = 1+\frac{2\tilde{\lambda}}{1-n/d-\tilde{\lambda}+\sqrt{(1-n/d-\tilde{\lambda})^2+4\tilde{\lambda}}}, \tag{60}$$

in which,

$$a := 1 + \frac{1}{\beta(Q_0 - Q)}, \qquad \text{and,} \qquad \tilde{\lambda} := \lambda + \frac{1}{\eta t}. \tag{61}$$

Case 2: F follows Assumption 2.1. The modulation matrix, F, under a SVD, $\mathrm{F} := \mathrm{U}\Sigma\mathrm{V}^T$ has two sets of singular values such that the first p singular values are equal to σ_1 and the remaining $d-p$ singular values are equal to σ_2. We let the condition number of F to be denoted by $\kappa := \frac{\sigma_1}{\sigma_2} > 1$.

Without loss of generality, we hereby assume that $\mathrm{U} = \mathrm{V} = \mathrm{I}$. Consequently, the (noiseless) teacher and the student can be written as the composition of two sub-models as following,

$$y^* = y_1^* + y_2^* = \boldsymbol{z}_1^T \boldsymbol{W}_1 + \boldsymbol{z}_2^T \boldsymbol{W}_2, \qquad \text{(teacher decomposition)} \tag{62}$$

$$\hat{y} = \hat{y}_1 + \hat{y}_2 = \sigma_1 \boldsymbol{z}_1^T \hat{\boldsymbol{W}}_1 + \sigma_2 \boldsymbol{z}_2^T \hat{\boldsymbol{W}}_2, \qquad \text{(student decomposition)} \tag{63}$$

in which $\boldsymbol{z}_1 \in \mathbb{R}^p$ and $\boldsymbol{z}_2 \in \mathbb{R}^{d-p}$.

Let $\hat{y}_i$ denote the output of the i^{th} component of the student. Also let y_i^* and y_i denote the noiseless and noisy targets, respectively. Therefore, for the student components $i \in 1, 2$, we have,

$$\begin{aligned} \hat{y}_1 &= \sigma_1 z_1^T \hat{\boldsymbol{W}}_1, \\ y_1^* &= z_1^T W_1, \\ y_1 &= y_1^* + \underbrace{z_2^T W_2 - \sigma_2 z_2^T \hat{\boldsymbol{W}}_2}_{y_2^* - \hat{y}_2 = \epsilon_2(t)} + \epsilon, \end{aligned} \qquad \begin{aligned} \hat{y}_2 &= \sigma_2 z_2^T \hat{\boldsymbol{W}}_2, \\ y_2^* &= z_2^T W_2, \\ y_2 &= y_2^* + \underbrace{z_1^T W_1 - \sigma_1 z_1^T \hat{\boldsymbol{W}}_1}_{y_1^* - \hat{y}_1 = \epsilon_1(t)} + \epsilon, \end{aligned}$$

in which ϵ is the *explicit noise*, added to the teacher's output while $\epsilon_j(t)$ is an *implicit variable noise* which decreases as the component $j \neq i$ learns to match $\hat{y}_j$ and y_j.

Accordingly, the variables H_i and G_i for each component i are re-defined as,

$$\begin{aligned} H_1 &= \mathbb{E}[y_1^* y_1] = \mathbb{E}_{y_1^*}[y_1^* y_1^*] = \frac{p}{d}, \\ G_1 &= \mathbb{E}[y_1 y_1], \\ &= \mathbb{E}[(y_1^* + y_2^* - \hat{y}_2)(y_1^* + y_2^* - \hat{y}_2)] + \sigma_\epsilon^2, \\ &= \mathbb{E}[y_1^* y_1^*] + \mathbb{E}[y_2^* y_2^*] + \mathbb{E}[\hat{y}_2 \hat{y}_2], \\ &\quad - 2\mathbb{E}[y_2^* \hat{y}_2] + \sigma_\epsilon^2, \\ &= \frac{p}{d} + \frac{d-p}{d} + Q_2 - 2R_2 + \sigma_\epsilon^2, \\ &= 1 + Q_2 - 2R_2 + \sigma_\epsilon^2, \end{aligned} \qquad \begin{aligned} H_2 &= \mathbb{E}[y_2^* y_2] = \mathbb{E}_{y_2^*}[y_2^* y_2^*] = \frac{d-p}{d}, \\ G_2 &= \mathbb{E}[y_2^T y_2], \\ &= \mathbb{E}[(y_2^* + y_1^* - \hat{y}_1)^T (y_2^* + y_1^* - \hat{y}_1)] + \sigma_\epsilon^2, \\ &= \mathbb{E}[y_2^* y_2^*] + \mathbb{E}[y_1^* y_1^*] + \mathbb{E}[\hat{y}_1 \hat{y}_1], \\ &\quad - 2\mathbb{E}[y_1^* \hat{y}_1] + \sigma_\epsilon^2, \\ &= \frac{d-p}{d} + \frac{p}{d} + Q_1 - 2R_1 + \sigma_\epsilon^2, \\ &= 1 + Q_1 - 2R_1 + \sigma_\epsilon^2, \end{aligned}$$

in which R_i and Q_i are defined as,

$$R_i := \mathbb{E}_z[y_i^* \hat{y}_i] = \frac{1}{d} W_i^T \sigma_i \hat{\boldsymbol{W}}_i, \quad \text{and,} \quad Q_i := \mathbb{E}_z[\hat{y}_i \hat{y}_i] = \frac{1}{d} \hat{\boldsymbol{W}}_i^T \sigma_i^2 \hat{\boldsymbol{W}}_i,$$

where σ_i denotes the singular values of the matrix F as defined in Assumption 2.1.

Rewriting Eqs. 58, 59, and 60 for each of the student's components, we arrive at,

$$
\begin{aligned}
R_1 &= \frac{n}{d}\frac{1}{a_1}, \\
Q_1 &= \frac{n}{pa_1^2 - n}\left(1 + Q_2 - 2R_2 + \sigma_\epsilon^2 - \frac{n}{d}\frac{2-a_1}{a_1}\right), \\
a_1 &= 1 + \frac{2\tilde{\lambda}_1}{1 - \frac{n}{p} - \tilde{\lambda}_1 + \sqrt{(1-\frac{n}{p}-\tilde{\lambda}_1)^2 + 4\tilde{\lambda}_1}}, \\
\tilde{\lambda}_1 &:= \frac{d}{p}\frac{1}{\sigma_1^2}(\lambda + \frac{1}{\eta t}),
\end{aligned}
\qquad
\begin{aligned}
R_2 &= \frac{n}{d}\frac{1}{a_2}, \\
Q_2 &= \frac{n}{(d-p)a_1^2 - n}\left(1 + Q_1 - 2R_1 + \sigma_\epsilon^2 - \frac{n}{d}\frac{2-a_2}{a_2}\right), \\
a_2 &= 1 + \frac{2\tilde{\lambda}}{1 - \frac{n}{d-p} - \tilde{\lambda} + \sqrt{(1-\frac{n}{d-p}-\tilde{\lambda})^2 + 4\tilde{\lambda}}}, \\
\tilde{\lambda}_2 &:= \frac{d}{d-p}\frac{1}{\sigma_2^2}(\lambda + \frac{1}{\eta t}),
\end{aligned}
$$

where Q_1 depends on Q_2 and vice versa. However, with simple calculations, we can arrive at the following standalone equation. Let,

$$\alpha_1 = \frac{n}{p},\ \alpha_2 = \frac{n}{d-p}, \tag{64}$$

and also let,

$$b_i = \frac{\alpha_i}{a_i^2 - \alpha_i}, \quad c_i = 1 - 2R_i - \frac{n}{d}\frac{2-a_i}{a_i} \qquad \text{for} \qquad i \in \{1,2\}, \tag{65}$$

with which the closed-from scalar expression for $Q(t,\lambda)$ reads,

$$Q(t,\lambda) = Q_1 + Q_2, \quad \text{where,} \quad Q_1 := \frac{b_1 b_2 c_2 + b_1 c_1}{1 - b_1 b_2}, \quad \text{and,} \quad Q_2 := \frac{b_1 b_2 c_1 + b_2 c_2}{1 - b_1 b_2}. \tag{66}$$

B.4. Derivation of $\tilde{\mathcal{L}}(\hat{\boldsymbol{W}}, t)$ in Eq. 22.

The t^{th} iterate of gradient descent on $\mathcal{L}_T(\hat{\boldsymbol{W}})$ matches the minimum of $\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}}, t)$ defined as,

$$\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}}, t) := \frac{1}{2n}\sum\left[\hat{y}^\mu - y^\mu\right]^2 + \frac{1}{\eta t}||\hat{\boldsymbol{W}}||_2^2. \tag{67}$$

Proof. The goal is to show,

$$\hat{\boldsymbol{W}}_t = \arg\min_{\hat{\boldsymbol{W}}} \tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}}, t), \qquad \text{where,} \qquad \hat{\boldsymbol{W}}_t := \hat{\boldsymbol{W}}_{t-1} - \eta \nabla_{\hat{\boldsymbol{W}}_{t-1}} \mathcal{L}(\hat{\boldsymbol{W}}_{t-1}). \tag{68}$$

For brevity of derivations, here we only consider the case where $\lambda = \sigma_\epsilon^2 = 0$. Recall the closed-form derivation of $\hat{\boldsymbol{W}}_t$ in

Eq. 19,

$$\hat{\boldsymbol{W}}_t = \Big(I - \big[I - \eta X^T X\big]^t\Big)(X^T X)^{-1} X^T y, \tag{69}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \Big[X\hat{\boldsymbol{W}} - X\Big(I - \big[I - \eta X^T X\big]^t\Big)(X^T X)^{-1} X^T y\Big]^2, \tag{70}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T}\Big(I - \big[I - \eta X^T X\big]^t\Big) \underbrace{(X^T X)^{-1} X^T y}_{=W, \text{ assuming } \sigma_\epsilon^2 = 0}\Big]^2, \tag{71}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - \underbrace{x^{\mu^T}\Big(I - \big[I - \eta X^T X\big]^t\Big) W}_{\text{a dynamic target (function of } t)}\Big]^2, \tag{72}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \big[I - \eta\Lambda\big]^t\Big) V^T W\Big]^2, \qquad (X^T X = V\Lambda V^T) \tag{73}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(t \log[I - \eta\Lambda]\big)\Big) V^T W\Big]^2, \tag{74}$$

$$\approx \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\eta\Lambda t\big)\Big) V^T W\Big]^2, \qquad (\log(1+x) \approx x) \tag{75}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\frac{\Lambda}{1/\eta t}\big)\Big) V^T W\Big]^2, \tag{76}$$

$$\approx \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \exp\big(-\log(\frac{\Lambda}{1/\eta t} + I)\big)\Big) V^T W\Big]^2, \quad (\log(1+x) \approx x) \tag{77}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big(I - \big[\Lambda + \frac{1}{\eta t} I\big]^{-1} \frac{1}{\eta t}\Big) V^T W\Big]^2, \tag{78}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} V\Big((\Lambda + \frac{1}{\eta t} I)^{-1}\Lambda\Big) V^T W\Big]^2, \tag{79}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} (X^T X + \frac{1}{\eta t} I)^{-1} X^T X W\Big]^2, \tag{80}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \frac{1}{2n} \sum \Big[\hat{y}^\mu - x^{\mu^T} \underbrace{(X^T X + \frac{1}{\eta t} I)^{-1} X^T y}_{\text{the normal equation}}\Big]^2, \tag{81}$$

$$= \arg\min_{\hat{\boldsymbol{W}}} \underbrace{\frac{1}{2n} \sum \Big[\hat{y}^\mu - y^\mu\Big]^2 + \frac{1}{\eta t} ||\hat{\boldsymbol{W}}||_2^2}_{\tilde{\mathcal{L}}_T(\hat{\boldsymbol{W}}, t)}, \tag{82}$$

which concludes the proof.

□

This proof have a core dependence on the findings of Ali et al. (2019; 2020). These works first formalize the connection between (continuous-time) GD or SGD-based training of an ordinary least squares (OLS) setup and that of ridge regression, providing bounds on the test error under these algorithms over training time t, in terms of a ridge setup with ridge parameter $\lambda = 1/t$. We utilize these results in the sense that by evaluating the generalization error $\mathcal{L}_G$ of our student-teacher setup with explicit ridge regularization, we invoke the connection between the ridge coefficient λ and training time t as described in these works, to obtain the behavior of (ridgeless) $\mathcal{L}_G$ over training.

B.5. Proof of Lemma 3.1

For a linear/linearized model, penalizing Q amounts to adding the following regularizer to the loss,

$$\mathcal{L}_T \leftarrow \mathcal{L}_T + \alpha ||\hat{y}||^2,$$

previously introduced in Pezeshki et al. (2020).

Proof: Recall that the variable Q is defined as,

$$Q := \frac{1}{d}\hat{W}^T \mathrm{F}^T \mathrm{F}\hat{W}.$$

Since Z is normally distributed with unit covariance, we can rewrite Q as,

$$Q := \frac{1}{d}\hat{W}^T \mathrm{F}^T Z^T Z \mathrm{F}\hat{W} = \frac{1}{d}\hat{W}^T X^T X \hat{W}.$$

We note that for a linear/linearized model of form $\hat{\boldsymbol{y}} := X^T\hat{W}$, the following identity holds,

$$||\hat{\boldsymbol{y}}||^2 = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} = \hat{W}^T X^T X \hat{W} = dQ.$$

B.6. Replica Trick

In the following, we detail the mathematical arguments leading to the *replica trick* expression (Edwards & Anderson, 1975). For some $r \to 0$, we can write for any scalar x:

$$\begin{aligned} x^r = \exp(r\ln x) &= \lim_{r\to 0} 1 + r\ln x \\ \Rightarrow \lim_{r\to 0} r\ln x &= \lim_{r\to 0} x^r - 1 \\ \Rightarrow \ln x &= \lim_{r\to 0} \frac{x^r - 1}{r} \\ \therefore \mathbb{E}[\ln x] &= \lim_{r\to 0} \frac{\mathbb{E}[x^r] - 1}{r}, \;\; \mathbb{E}: \text{averaging} \end{aligned} \tag{83}$$

B.7. Computation of the free-energy

The self-averaged free energy (per unit weight) of our student network, is given by (Engel & Van den Broeck, 2001),

$$-\beta f = \frac{1}{d}\langle\langle \ln Z\rangle\rangle_{z,W} \tag{84}$$

Here, $\beta = 1/T$ is the inverse temperature parameter corresponding to our statistical ensemble, d the (teacher) student network width, and Z the partition function of the system defined as (n: number of training examples).

Leveraging the replica trick, we next obtain,

$$\begin{aligned} \langle\langle Z^r\rangle\rangle_{z,W} &= \prod_{a=1}^{r}\prod_{\mu=1}^{d}\int \mathrm{d}\mu\,(W^a)\,\mathrm{d}y_a^\mu \mathrm{d}(y^*)^\mu e^{-\beta N \mathcal{E}_T(y_a, y^*)} \\ &\quad\times \left\langle\left\langle \delta\left(y^{*\mu} - \frac{1}{\sqrt{d}}W^T x^{*\mu}\right)\delta\left(y_a^\mu - \frac{1}{\sqrt{d}}W_a^T x^\mu\right)\right\rangle\right\rangle_{z,W} \\ &= \prod_{a=1}^{r}\prod_{\mu=1}^{d}\int \mathrm{d}\mu\,(W^a)\,\frac{\mathrm{d}y_a^\mu \mathrm{d}\hat{y}_a^\mu}{2\pi}\frac{\mathrm{d}y^{*\mu}\mathrm{d}\hat{y}^{*\mu}}{2\pi} e^{-\beta N \mathcal{E}_T(y_a, y^*)} e^{iy^{*\mu}\hat{y}^{*\mu} + iy_a^\mu \hat{y}_a^\mu} \\ &\quad\times \left\langle\left\langle \exp\left(-\frac{i}{\sqrt{d}}\hat{y}^{*\mu}W^T x^{*\mu} - \frac{i}{\sqrt{d}}\hat{y}_a^\mu W_a^T x^\mu\right)\right\rangle\right\rangle_{z,W} \end{aligned} \tag{85}$$

where in the last line above, we have expressed the inserted δ functions using their integral representations. To make further progress, we introduce the auxiliary variables,

$$\sum_{ija} W_a^i \Delta_{ij} W^{*j} = dR_a, \tag{86}$$

$$\sum_{ij\langle a,b\rangle} W_a^i \Gamma_{ij} W_b^j = dQ_{ab} \tag{87}$$

via the respective δ functions, to arrive at,

$$\begin{aligned}\langle\langle Z^n\rangle\rangle_{z,W} = \prod_{\mu,a,b}\int \mathrm{d}\mu\left(\mathbf{W}^a\right)\frac{\mathrm{d}y_a^\mu \mathrm{d}\hat{y}_a^\mu}{2\pi}\frac{\mathrm{d}y^{*\mu}\mathrm{d}\hat{y}^{*\mu}}{2\pi}e^{-\beta N\mathcal{E}_T(y_a,y^*)}e^{iy^{*\mu}\hat{y}^{*\mu}+iy_a^\mu\hat{y}_a^\mu}\\ \times\int P\mathrm{d}Q^{ab}\int P\mathrm{d}R^a\;\delta\left(\sum_{i,j,a}W_a^i\Delta_{i,j}W^{*j}-PR_a\right)\delta\left(\sum_{ij\langle a,b\rangle}W_a^i\Gamma_{ij}W_b^j-PQ_{ab}\right)\\ \times\left\langle\left\langle\exp\Big(-\frac{Q_0}{2}\sum_{\mu,a}(\hat{y}_a^\mu)^2-\frac{1}{2}\sum_{\mu,\langle a,b\rangle}\hat{y}_a^\mu\hat{y}_b^\mu Q_{ab}-\sum_{\mu,a}\hat{y}^{*\mu}\hat{y}_a^\mu R_a-\frac{1}{2}\sum_\mu(\hat{y}^{*\mu})^2\Big)\right\rangle\right\rangle_W\end{aligned} \tag{88}$$

Repeating the procedure of expressing the above δ functions using their integral representations, we then get ($\alpha = n/d$),

$$\begin{aligned}\langle\langle Z^n\rangle\rangle_{x,x^*,W} = \int\prod_{a,b}\frac{\mathrm{d}Q_0}{\sqrt{2\pi}}\frac{\mathrm{d}\hat{Q}_{0a}}{4\pi}\frac{\mathrm{d}Q_{ab}\hat{Q}_{ab}}{2\pi/d}\frac{\mathrm{d}R_a\hat{R}_a}{2\pi/d}\exp\Big(\frac{iP}{2}\sum_a Q_0\hat{Q}_{0a}+iP\sum_{a<b}Q^{ab}\hat{Q}^{ab}\\ +iP\sum_a R^a\hat{R}^a\Big)\int\prod_{i,a}\frac{\mathrm{d}W_i^a}{\sqrt{2\pi}}\exp\Big(-\frac{i}{2}\sum_{i,j,a}\hat{Q}_{0a}W_a^i\Gamma_{ij}W_a^j\\ -i\sum_{i,j,a<b}\hat{Q}_{ab}W_a^i\Gamma_{ij}W_b^j-i\sum_{i,j,a}\hat{R}_a\Delta_{ij}W_a^j\Big)\times\\ \int\prod_{\mu,a}\frac{\mathrm{d}y_a^\mu\mathrm{d}\hat{y}_a^\mu}{2\pi}\frac{\mathrm{d}y^{*\mu}}{\sqrt{2\pi}}e^{-\beta N\mathcal{E}_T(y_a,y^*)}\exp\Big(-\frac{1}{2}\sum_\mu(y^{*\mu})^2+i\sum_{\mu,a}\hat{y}_a^\mu\hat{y}_a^\mu\\ -\frac{1}{2}\sum_{a,\mu}\left(1-R_a^2\right)(\hat{y}_a^\mu)^2-\frac{1}{2}\sum_{\mu,\langle a,b\rangle}\hat{y}_a^\mu\hat{y}_b^\mu\left(Q^{ab}-R^aR^b\right)-i\sum_{\mu,a}y^{*\mu}\hat{y}_a^\mu R^a\Big)\end{aligned} \tag{89}$$

If we now, perform a singular value decomposition of the covariance matrix Γ as, $\Gamma = \mathrm{U}^T\mathrm{S}\mathrm{U} = \mathrm{V}^T\mathrm{V}$, where S: matrix of singular values of Γ, and we have expressed, $\mathrm{V} = \mathrm{S}^{1/2}\mathrm{U}$, then one can proceed to write,

$$\begin{aligned}\langle\langle Z^n\rangle\rangle_{x,W} = \frac{1}{\det|V|}\int\prod_{a,b}\frac{\mathrm{d}Q_0}{\sqrt{2\pi}}\frac{\mathrm{d}\hat{Q}_{0a}}{4\pi}\frac{\mathrm{d}Q_{ab}\hat{Q}_{ab}}{2\pi/d}\frac{\mathrm{d}R_a\hat{R}_a}{2\pi/d}\exp\Big(\frac{iP}{2}\sum_a Q_0\hat{Q}_{0a}\\ +iP\sum_{a<b}Q^{ab}\hat{Q}^{ab}+iP\sum_a R^a\hat{R}^a\Big)\int\prod_{i,a}\frac{\mathrm{d}\tilde{W}_i^a}{\sqrt{2\pi}}\exp\Big(-\frac{i}{2}\sum_{i,a}\hat{Q}_{0a}\left(\tilde{W}_a^i\right)^2\\ -i\sum_{i,a<b}\hat{Q}_{ab}\tilde{W}_a^i\tilde{W}_b^i-i\sum_{i,j,a}\hat{R}_a\tilde{W}_a^j\Big)\times\int\prod_{\mu,a}\frac{\mathrm{d}y_a^\mu\mathrm{d}\hat{y}_a^\mu}{2\pi}\frac{\mathrm{d}y^{*\mu}}{\sqrt{2\pi}}e^{-\beta N\mathcal{E}_T(y_a,y^*)}\\ \exp\Big(-\frac{1}{2}\sum_\mu(y_\mu^*)^2+i\sum_{\mu,a}\hat{y}_a^\mu\hat{y}_a^\mu-\frac{1}{2}\sum_{a,\mu}\left(1-R_a^2\right)(\hat{y}_a^\mu)^2-i\sum_{\mu,a}y^{*\mu}\hat{y}_a^\mu R^a\\ -\frac{1}{2}\sum_{\mu,\langle a,b\rangle}\hat{y}_a^\mu\hat{y}_b^\mu\left(Q^{ab}-R^aR^b\right)\Big)\end{aligned} \tag{90}$$

having expressed, $\tilde{W}_a = \mathrm{V}W_a$, and identifying $\Delta = \mathrm{S}^{1/2}\mathrm{U}$ from our definitions. Now, since in the above, the W_i^a integrals factorize in i, and similarly the y_a^μ, $\hat{y}_a^\mu$ and $\mathrm{d}y^{*\mu}$ factorize in μ, one can proceed to write:

$$\begin{aligned}\langle\langle Z^n\rangle\rangle_{x,W} = \frac{1}{\det|V|}\int\prod_{a,b}\frac{\mathrm{d}Q_0\mathrm{d}\hat{Q}_{0a}}{\sqrt{2\pi}4\pi}\frac{\mathrm{d}Q_{ab}\hat{Q}_{ab}}{2\pi/d}\frac{\mathrm{d}R_a\hat{R}_a}{2\pi/d}\exp\Big(P\Big[\frac{i}{2}\sum_a Q_0\hat{Q}_{0a}\\ +i\sum_{a<b}Q^{ab}\hat{Q}^{ab}+i\sum_a R^a\hat{R}^a+G_S(\hat{Q}_{0a},\hat{Q}^{ab},\hat{R}^a)+\alpha G_E(Q^{ab},R^a)\Big]\Big)\end{aligned} \tag{91}$$

where,

$$
\begin{aligned}
G_S(\hat{Q}_{0a}, \hat{Q}^{ab}, \hat{R}^a) &= \ln \int \prod_a \frac{\mathrm{d}\tilde{W}^a}{\sqrt{2\pi}} \exp\Big(-\frac{i}{2}\sum_a \hat{Q}_{0a}\tilde{W}_a^i\tilde{W}_a^i - i\sum_{a<b}\hat{Q}_{ab}\tilde{W}_a\tilde{W}_b - i\sum_a \hat{R}_a\tilde{W}_a\Big) \\
G_E(Q^{ab}, R^a) &= \ln \int \prod_a \frac{\mathrm{d}y_a \mathrm{d}\hat{y}_a}{2\pi}\frac{\mathrm{d}y^*}{\sqrt{2\pi}} e^{-\beta N \mathcal{E}_T(y_a, y^*)} \exp\Big(-\frac{1}{2}(y^*)^2 + i\sum_a \hat{y}_a\hat{y}_a \\
&\quad - \frac{1}{2}\sum_a \left(1 - R_a^2\right)(\hat{y}_a)^2 - \frac{1}{2}\sum_{\langle a,b\rangle}\hat{y}_a\hat{y}_b\left(Q^{ab} - R^aR^b\right) - iy^{*\mu}\sum_a \hat{y}_a R^a\Big)
\end{aligned}
\tag{92}
$$

Now, in the limit $d \to \infty$, Eq. 91 can be approximated using the saddle-point approach (Bender & Orszag, 2013),

$$
\begin{aligned}
\langle\langle Z^n\rangle\rangle_{x,W} \approx \textbf{extr}_{Q_0,\hat{Q}_{0a},Q^{ab},\hat{Q}^{ab},R^a,\hat{R}^a} \exp\Big(P\Big[\frac{i}{2}\sum_a Q_0\hat{Q}_{0a} + i\sum_{a<b}Q^{ab}\hat{Q}^{ab} \\
+ i\sum_a R^a\hat{R}^a + G_S(\hat{Q}_{0a}, \hat{Q}^{ab}, \hat{R}^a) + \alpha G_E(Q^{ab}, R^a)\Big]\Big)
\end{aligned}
\tag{93}
$$

where, **extr** corresponds to extremization of $\langle\langle Z^n\rangle\rangle_{x,W}$ over the respective order parameters. Performing this extremization over $\hat{Q}_{0a}, \hat{Q}^{ab}$ and $\hat{R}^a$, then generates an expression of the form,

$$
\begin{aligned}
\langle\langle Z^n\rangle\rangle_{x,W} = \textbf{extr}_{Q_0,Q,R} \exp\Bigg\{nN\Bigg(\frac{1}{2}\frac{Q - R^2}{Q_0 - Q} + \frac{1}{2}\ln(Q_0 - Q) - \frac{\alpha}{2}\ln\left[1 + \beta(Q_0 - Q)\right] \\
- \frac{\alpha\beta}{2}\frac{1 - 2R + Q}{1 + \beta(Q_0 - Q)}\Bigg)\Bigg\}
\end{aligned}
\tag{94}
$$

where we have invoked *replica symmetry* in the form, $Q^{ab} = Q$ and $R^a = R$, and that $\mathcal{E}_T = (y^* - y)^2/2$. Plugging this back into Eq. 84, then finally yields,

$$
\begin{aligned}
\beta f = -\textbf{extr}_{Q_0,Q,R}\Bigg\{\frac{1}{2}\frac{Q - R^2}{Q_0 - Q} + \frac{1}{2}\ln(Q_0 - Q) - \frac{\alpha}{2}\ln\left[1 + \beta(Q_0 - Q)\right] \\
- \frac{\alpha\beta}{2}\frac{1 - 2R + Q}{1 + \beta(Q_0 - Q)}\Bigg\}
\end{aligned}
\tag{95}
$$

The remaining pair of order parameters generate the following set of transcendental equations on extremization (Bös, 1998):

$$
\begin{aligned}
R &= \frac{\alpha}{a} \\
Q &= \frac{\alpha}{a^2 - \alpha}\left(1 - \frac{2 - a}{a}\alpha\right) \\
Q_0 &= Q + \frac{1}{\beta\,(a - 1)}
\end{aligned}
\tag{96}
$$

where, $a = \max[1, \alpha]$ for $T \to 0$.

Now, the above determined values of R, Q and Q_0 can be perceived as the *maximally likely* values of R, Q and Q_0 of our teacher-student setup, for an inverse temperature β parameterizing the system.