Journal of Machine Learning Research 27 (2026) 1-182 Submitted 2/26; Published 6/26

High-Dimensional Analysis of Gradient Flow for Extensive-Width Quadratic Neural Networks

Simon Martin SMARTIN@DI.ENS.FR
INRIA
Laboratoire de Physique de l'École Normale Supérieure
ENS, Université PSL
Paris, France

Giulio Biroli GIULIO.BIROLI@PHYS.ENS.FR
Laboratoire de Physique de l'École Normale Supérieure
ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris
Paris, France

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA
École Normale Supérieure
PSL Research University
Paris, France

Editor: Lorenzo Rosasco

Abstract

We study the high-dimensional training dynamics of a shallow neural network with quadratic activation in a teacher–student setup. We focus on the extensive-width regime, where the teacher and student network widths scale proportionally with the input dimension, and the sample size grows quadratically. This scaling aims to describe overparameterized neural networks in which feature learning still plays a central role. In the high-dimensional limit, we derive a dynamical characterization of the gradient flow, in the spirit of dynamical mean-field theory (DMFT). Under ℓ_2-regularization, we analyze these equations at long times and characterize the performance and spectral properties of the resulting estimator. This result provides a quantitative understanding of the effect of overparameterization on learning and generalization, and reveals a double descent phenomenon in the presence of label noise, where generalization improves beyond interpolation. In the small regularization limit, we obtain an exact expression for the perfect recovery threshold as a function of the network widths, providing a precise characterization of how overparameterization influences recovery.

Keywords: shallow neural networks, high-dimensional learning, training dynamics, overparameterized neural networks, generalization

1. Introduction

Deep neural networks have achieved remarkable success across many domains such as image and speech recognition (Krizhevsky et al., 2012; Hinton et al., 2012), protein structure prediction (Jumper et al., 2021), natural language processing (Vaswani et al., 2017; Brown et al., 2020) and autonomous systems (Bojarski et al., 2016). On the theoretical side, while

the expressive power of neural architectures has been clarified (Cybenko, 1989; Hornik et al., 1989), many fundamental challenges related to their loss landscapes and training dynamics remain open. The main obstacles to developing a thorough and general analysis are the high-dimensionality and the non-convexity of the loss landscape.

Overparameterized neural networks. This gap has motivated a line of work focused on gradient-based training and its success, despite the non-convexity of neural networks objectives. In theory, gradient descent could get stuck in poor local minima, even for relatively simple architectures. Indeed, several works have shown that loss landscapes exhibit spurious local minima in underparameterized networks (Christof and Kowalczyk, 2023) and shallow ReLU models (Safran and Shamir, 2018; Yun et al., 2018). However, recent works suggest that in some regimes, optimization algorithms often find global minimizers of their training loss. For instance, in the overparameterized regime, Chizat and Bach (2018) have shown that two-layer networks converge globally, while Jacot et al. (2018) and Chizat et al. (2019) have described the lazy and neural tangent kernel (NTK) regimes, in which the network behaves nearly linearly during training. In addition, in heavily overparameterized regimes, several works have shown that deep architectures enjoy polynomial-time convergence guarantees (Allen-Zhu et al., 2019; Kawaguchi and Huang, 2019; Du et al., 2019). Most of these investigations focus on the case in which the number of hidden nodes diverges but the dimension of data is kept fixed.

High-dimensional proportional regime. In contrast to overparameterized networks, the proportional regime corresponds to a fixed number of hidden units while the dimension and the number of samples diverge proportionally. In this regime, a long line of work has analyzed the high-dimensional learning abilities of simple models such as multi-index and shallow neural networks. More precisely, tools from statistical physics have been applied to predict Bayes-optimal errors, phase transitions, and information-theoretic limits (Loureiro et al., 2022; Aubin et al., 2019; Maillard et al., 2020b,a). Many of these predictions have since been confirmed through rigorous approaches such as adaptive interpolation (Barbier and Macris, 2019; Barbier et al., 2019, 2020) and approximate message passing (Donoho et al., 2009; Gerbelot and Berthier, 2023; Cornacchia et al., 2023; Gerbelot et al., 2022). Beyond these static analyses, dynamical mean-field methods have studied gradient-based trajectories (Mignacco et al., 2020; Sarao Mannelli et al., 2019; Agoritsas et al., 2018), and more recent rigorous works have clarified the high-dimensional behavior of stochastic gradient descent (Ben Arous et al., 2022; Gerbelot et al., 2024) and Langevin dynamics (Fan et al., 2025). This line of work has built a precise understanding of learning in the regime of high-dimensional data and finite number of hidden nodes.

Quadratic networks and matrix sensing. Shallow neural networks with quadratic activation functions provide a simple model to study nonconvex optimization, as they reduce learning to the estimation of a positive semidefinite matrix. In this setting, several works studied the impact of overparameterization on the loss landscape (Soltanolkotabi et al., 2018; Du and Lee, 2018; Gamarnik et al., 2019; Venturi et al., 2019) and consequences of the learning dynamics with gradient-based methods, with an emphasis on population dynamics (Sarao Mannelli et al., 2020; Martin et al., 2024), and stochastic gradient descent (Ben Arous et al., 2025). In addition, this setting is equivalent to a low-rank matrix sensing problem, with rank-one measurements: in this more general framework, convex relaxation

guarantees recovery (Recht et al., 2010; Candes and Plan, 2011; Gross, 2011). In the case of nonconvex factorizations, several results established the absence of spurious local minima and global convergence of gradient descent under restricted isometry properties (Zheng and Lafferty, 2015; Ge et al., 2017; Park et al., 2017; Li et al., 2018), with an implicit bias toward minimal nuclear norm solutions (Gunasekar et al., 2017; Arora et al., 2019). As we shall show, quadratic networks offer a complex but tractable framework to investigate the challenging regime in which the number of hidden nodes, the dimension of the data and the number of data points diverge (in fixed ratios to be precised below).

1.1 Related Works

Quadratic networks have been the object of recent attention: they provide a framework allowing to study the role of overparameterization in a nonconvex setting. Several works have already clarified convergence properties of gradient-based methods in the highly overparameterized regime, where the problem enjoys a convex relaxation. In this regime, it has been shown that gradient flow converges globally and several works studied the generalization properties of the minimizers found by the dynamics (Sarao Mannelli et al. (2020) for a rank-one teacher, Gamarnik et al. (2019) for a full-rank teacher and Soltanolkotabi et al. (2018) when the output weights are also optimized). However, in the case where the network has fewer neurons than the dimension, existing dynamical studies only include population dynamics, i.e., the infinite sample limit, with recent works by Martin et al. (2024) for gradient flow and Ben Arous et al. (2025) for stochastic gradient descent (SGD). This paper aims to go beyond these results by providing an exact description of gradient flow dynamics on the empirical loss for an arbitrary, yet extensive, number of neurons.

The extensive-width regime, in which the number of hidden nodes, the input dimension and the number of data points diverge, has recently been studied with the goal of describing overparameterized neural networks in a setting where feature learning is still present. In this setting, several works already clarified Bayes-optimal learning and empirical risk minimization: Maillard et al. (2024); Erba et al. (2025b); Defilippis et al. (2025) in the case of quadratic networks, Cui et al. (2023); Barbier et al. (2025b) for deep networks, Erba et al. (2025a) for bilinear regression, and Boncoraglio et al. (2025a,b) for attention networks. Additionally, note that similar scalings were originally studied in the context of matrix denoising (Maillard et al., 2022; Semerjian, 2024; Barbier et al., 2025a). However, in the dynamical setting, it is still an open question to know if gradient-based algorithms match these static predictions. In recent works, Montanari and Urbani (2025) leverage dynamical mean-field theory (DMFT) to identify a separation of timescales between learning and overfitting in the gradient flow setting, while Ren et al. (2025) characterize SGD dynamics through precise scaling laws. Our work provides a complementary perspective to these studies by deriving sharp asymptotic results in the extensive-width regime.

One goal of the present work is also to connect the dynamical perspective with the static predictions obtained in the analyses of Maillard et al. (2024) and Erba et al. (2025b), whose settings are essentially identical to ours. More precisely, our dynamical approach is inspired by the replica calculation introduced by Maillard et al. (2024), while the regularized formulation we adopt follows the framework proposed by Erba et al. (2025b). Within this setting, we are able to recover several conclusions reached in these studies. Together, these

observations clarify how gradient flow and Langevin dynamics connect to the underlying landscape structure and optimal learning predictions.

1.2 Contributions

In this work, we analyze the learning dynamics of a shallow quadratic neural network trained with gradient flow and Langevin dynamics on a regularized empirical loss. To study the role of overparameterization on learning and generalization, we work in the *extensive-width* regime, where the widths of the teacher and student networks scale proportionally with the dimension. For analytical tractability, we introduce a Gaussian surrogate model that preserves the covariance structure of the rank-one measurements associated with the quadratic network model. We have verified by numerical experiments that this simplification does not affect our predictions. Our main results are:

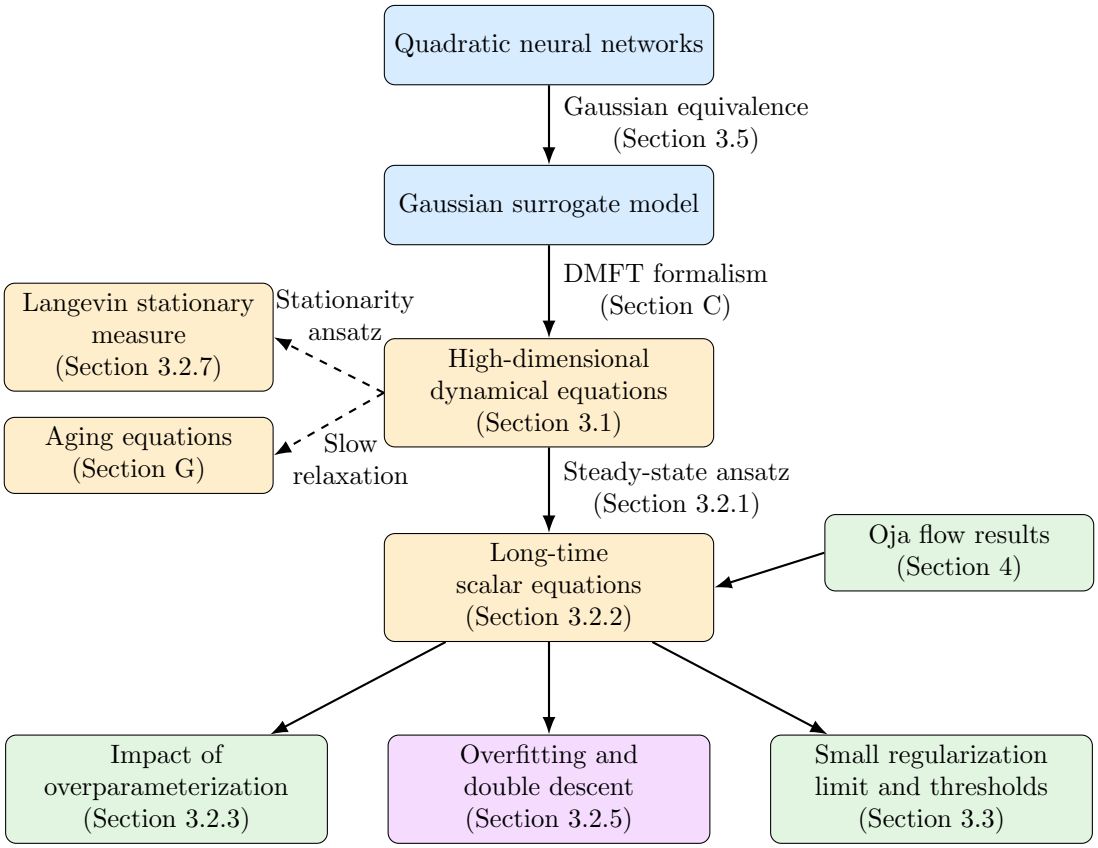

Figure 1: Logical structure of the main contributions. Solid arrows indicate the main chain of derivations and consequences, while dashed arrows represent complementary analyses. Node colors indicate the nature of the corresponding contribution: model formulation (blue), DMFT predictions and ansatz-based reductions (orange), rigorous analyses (green), and phenomena validated numerically from the theory (purple).

(*i*) *Description of the high-dimensional dynamics.* In the high-dimensional limit, we derive a set of dynamical equations that is distributionally equivalent to gradient flow and Langevin dynamics (Sections 3.1.1, 3.1.3). This result provides a self-consistent description of the learning dynamics in this asymptotic regime. Our equations are formulated and derived in the spirit of DMFT equations, but still involve high-dimensional objects. We believe that a reduction to a set of low-dimensional equations is impossible while keeping track of the relevant averaged quantities of the dynamics.

(*ii*) *Reduction to a denoising dynamics.* Under ℓ_2-regularization and an asymptotic simplification of the dynamics, we argue that the gradient flow dynamics can be reduced to a denoising gradient flow, known as the Oja flow (Section 3.2.1). We interpret this dynamics as a noisy version of the population dynamics (the infinite-sample gradient flow), where the effective noise variance arises from finite sample complexity and label noise. We analyze the Oja flow and obtain new results that are essential for our theoretical analysis (Section 4).

(*iii*) *Long-time analysis of the dynamics.* Building on this reduction, we study the long-time behavior of the gradient flow dynamics in high dimensions. This leads to a set of scalar equations describing the performance and spectral properties of the gradient flow estimator (Section 3.2.2). We support these results with theoretical insights (Section 3.2.4) and numerical confirmations, while noting that we leave a rigorous treatment of the simplifying assumptions for further work.

(*iv*) *Langevin dynamics.* We study the Langevin dynamics on the regularized empirical loss and derive its stationary measure under similar simplifications to those used for gradient flow (Section 3.2.7). In the zero-temperature limit, the stationary measure both recovers the long-time equations of the deterministic dynamics and concentrates on global minimizers, providing evidence for convergence of gradient flow to a global minimum.

(*v*) *Impact of overparameterization.* The long-time equations reveal that the overparameterized regime, in which the network has more neurons than the input dimension, persists down to a smaller width that we identify (Section 3.2.3). Above this threshold, gradient flow converges to the global minimizer of the regularized empirical loss over the set of positive semidefinite (PSD) matrices. Below it, the performance and behavior of the estimator depend on the effective number of neurons per dimension. These equations illustrate how overparameterization enables global convergence, whereas insufficient width traps the dynamics at higher loss solutions.

(*vi*) *Overfitting and double descent.* We provide numerical evidence of overfitting along the training dynamics. In the presence of label noise, our long-time equations capture a double descent phenomenon in the test error as a function of the sample complexity (Section 3.2.5). In the small regularization limit, we identify the associated interpolation threshold (Section 3.3.1). This result constitutes a theoretical characterization of double descent based on the dynamical study of a nonlinear model.

(*vii*) *Perfect recovery thresholds.* In the small regularization limit, we derive an explicit expression of the perfect recovery threshold as a function of the teacher and stu-

dent network widths, quantifying how overparameterization affects recovery (Section 3.3.3). For the unregularized case, we conjecture an expression of this threshold (Section 3.4.2) and provide numerical evidence supporting this conjecture.

(*viii*) *Insights into Gaussian equivalence.* We investigate the equivalence between the Gaussian surrogate model and the rank-one formulation induced by the quadratic network model (Section 3.5). We conjecture their asymptotic equivalence and provide numerical and theoretical evidence.

2. Setting

In this section, we introduce the notation and conventions used throughout the paper and describe the model and main assumptions of our work.

2.1 Notation and Conventions

For a matrix $A \in \mathbb{R}^{d\times m}$, we denote $A^\top$ its transpose, $\mathrm{Tr}(A)$ its trace and $\|A\|_F = \big[\mathrm{Tr}(AA^\top)\big]^{1/2}$ its Frobenius norm. We denote $\mathcal{S}_d(\mathbb{R}), \mathcal{S}_d^+(\mathbb{R})$ and $\mathcal{S}_d^{++}(\mathbb{R})$ the sets of $d \times d$ symmetric, positive semidefinite (PSD) and positive definite matrices. For $A \in \mathbb{R}^{d\times d}$, we denote $\mathrm{Sym}(A) = (A + A^\top)/2 \in \mathcal{S}_d(\mathbb{R})$ its symmetrization. Given E a Euclidean space and $L\colon E \to \mathbb{R}$ continuously differentiable, we let $\nabla L(x) \in E$ be its gradient at $x \in E$.

2.2 Model

Consider a data set composed of standard Gaussian inputs $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ and labels $z_1, \dots, z_n \in \mathbb{R}$ generated as:

$$z_k \sim P\big(\,\cdot\,\big|\, y_k^*\big), \qquad y_k^* = \frac{1}{m^*}\sum_{i=1}^{m^*} \frac{(x_k^\top w_i^*)^2 - \|w_i^*\|^2}{\sqrt{d}}, \tag{1}$$

where the vectors $w_1^*, \dots, w_{m^*}^* \in \mathbb{R}^d$ will be referred to as teacher vectors, and the distribution P encodes possible nonlinearity and noise applied to the true labels $(y_k^*)_{1\le k\le n}$. This structure corresponds to a two-layer neural network with a centered quadratic activation function and uniform output weights. The centering subtracts the expectation of the quadratic activation under the input distribution, since $\mathbb{E}\big[(x^\top w)^2\big] = \|w\|^2$ for $x \sim \mathcal{N}(0, I_d)$, therefore removing an uninformative contribution. Equivalently, the centered quadratic corresponds to the second-order Hermite polynomial, leading to an activation function with information exponent 2.

Unlike in generalized linear models (McCullagh, 2019; Barbier et al., 2019), our objective is not to recover the individual teacher vectors w_i^*. Indeed, the quadratic network is invariant under any orthogonal transformation of these vectors, leading to a degeneracy: many different sets of teacher vectors yield the same model output. As a result, only their second-order structure can be identified. Therefore, we instead aim to recover the *teacher matrix*:

$$Z^* = \frac{1}{m^*}\sum_{i=1}^{m^*} w_i^* w_i^{*\top} \in \mathcal{S}_d^+(\mathbb{R}), \tag{2}$$

which captures all the information about the teacher that we can recover from the observations. This matrix plays a central role in the quadratic model, as it allows one to express the true labels as:

$$y_k^* = \mathrm{Tr}\big(X_k Z^*\big), \qquad X_k = \frac{x_k x_k^\top - I_d}{\sqrt{d}}. \tag{3}$$

This formulation reveals an underlying linear structure: each true label y_k^* is a linear measurement of the teacher matrix Z^* through the sensing matrices X_k. The problem can therefore be viewed as a generalized *matrix sensing* task, where one aims to recover Z^* from the observations of $z_1, \dots, z_n$ that may depend nonlinearly on the true labels. In the particular case where P is Gaussian, this reduces to the standard matrix sensing model with additive Gaussian noise on the labels.

Under this representation, the effective parameter of the model becomes the matrix Z^*, that gathers all second-order information about the teacher. In the following, we will no longer refer to the individual teacher vectors, as the model's invariance makes them non-identifiable. Instead, we directly work with the matrix Z^*, whose structure is determined by the model: it is positive semidefinite and of rank m^*, with typically $m^* < d$.

2.3 Optimization

Given some cost $\ell\colon \mathbb{R}^2 \to \mathbb{R}^+$, we leverage the low-rank structure of Z^* and define the empirical loss function:

$$\mathcal{L}(W) = \frac{1}{2n} \sum_{k=1}^{n} \ell\Big(\mathrm{Tr}(X_k W W^\top), z_k\Big), \tag{4}$$

for $W \in \mathbb{R}^{d\times m}$. We assume that m^* is not known, therefore we potentially have $m \neq m^*$. While optimizing with respect to W directly exploits this low-rank structure, it is often simpler to analyze optimization with respect to the matrix $Z = WW^\top$, which can lead to convex formulations. Relevant works on this problem include the results of Burer and Monteiro (2003) and Gunasekar et al. (2017) on factorized approaches, and in a more general setting (Edelman et al., 1998; Journée et al., 2010; Massart and Absil, 2020) on the geometry and optimization of functions of $WW^\top$.

To optimize the student matrix W, we add a regularization $\Omega\colon \mathbb{R}^{d\times m} \to \mathbb{R}^+$ to the empirical loss and perform Langevin dynamics:

$$\mathrm{d}W(t) = -d\,\nabla\mathcal{L}\big(W(t)\big)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{5}$$

where B is a standard Brownian motion over $\mathbb{R}^{d\times m}$, and the inverse temperature $\beta \geq 0$ controls the intensity of the noise. When setting $\beta = \infty$, we recover the plain gradient flow dynamics, which is the setting for most of our results. The prefactor $1/\sqrt{d}$ in front of B ensures that the effect of the noise remains of order one in the high-dimensional limit, consistently with the scaling we introduce in Assumption 1.

In addition, although our analysis is formulated in continuous time, the results presented in Section 3.1 remain valid for gradient descent, corresponding to the discretized version of the dynamics (5), with $\beta = \infty$:

$$W_{k+1} = W_k - \eta d\,\nabla\mathcal{L}(W_k) - \eta\nabla\Omega(W_k), \tag{6}$$

where $\eta > 0$ corresponds to the stepsize of the algorithm.

Gradient descent simulations. The gradient descent algorithm (with small stepsize) will also be used as a proxy for the gradient flow dynamics in our numerical experiments. As some of our results will be derived using non-rigorous methods, numerical simulations are essential to confirm the validity of our claims. Throughout this paper, several figures are presented in order to compare the numerical integration of equation (6) to our theoretical claims. Unless specified in the corresponding figure, all of our simulations were performed at dimension $d = 100$, with stepsize $\eta = 5 \times 10^{-3}$, and error bars indicate the standard deviation under several realizations of the random initialization, teacher matrix and sensing matrices. However our results hold for a general class of initialization and teacher distributions (see Assumption 1 for more details), we have chosen the Gaussian distribution for our simulations:

$$W_{ij}(t=0) \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1}{m}\right), \qquad (w_i^*)_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{7}$$

In this case, the teacher matrix Z^*, expressed as the average of the rank-one matrices $w_i^* w_i^{*\top}$ in equation (2), is therefore distributed as a Wishart matrix (see Section B.1).

We give more details on our simulation setup in Section J.1.1 and we provide a code to reproduce the numerical experiments of the paper in the freely accessible GitHub repository: `https://github.com/simonmartin15/QuadraticNets`.

2.4 High-Dimensional Limit

In this work, we analyze the gradient flow dynamics (5) in the limit where the dimension d diverges. One key feature of our analysis is to study the *extensive-width regime* where the widths m, m^* of the student and teacher networks diverge proportionally to the dimension. This scaling is designed to reflect architectures with both large input dimension and hidden layers size, representing settings in which the feature space increases with the scale of the problem.

Moreover, one key feature of our analysis consists in replacing the centered rank-one matrices $(X_k)_{1 \le k \le n}$ in equation (3) by symmetric Gaussian matrices with the same mean and covariance. This distribution corresponds to the Gaussian orthogonal ensemble (denoted GOE(d) in the following), that is, the random symmetric matrices $X \in \mathcal{S}_d(\mathbb{R})$ whose coefficients are distributed as:

$$X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1+\delta_{ij}}{d}\right), \qquad \big(1 \le i \le j \le d\big). \tag{8}$$

Replacing the centered rank-one sensing matrices by Gaussian matrices is justified by recent universality and exact-asymptotics results in static problems (Maillard et al., 2024; Erba et al., 2025b; Xu et al., 2025). These works show that, for purposes of sharp risk and Bayes-optimal asymptotics, the quadratic network model can be traded for a sensing matrix problem with Gaussian observations. We claim that the same equivalence holds for the gradient flow trajectory in the high-dimensional limit, and that all our results apply to the observations introduced in equation (3). We refer to this property as *Gaussian equivalence*, which is supported by theoretical considerations and numerical experiments, presented in Section 3.5.

Assumption 1 *We make the following assumptions throughout this paper:*

(*i*) High-dimensional scaling. *The number of observations n and the number of neurons m, m^* scale with the dimension d as:*

$$n \underset{d\to\infty}{\sim} \alpha d^2, \qquad m \underset{d\to\infty}{\sim} \kappa d, \qquad m^* \underset{d\to\infty}{\sim} \kappa^* d, \tag{9}$$

where α, κ, κ^* are fixed positive constants, referred to respectively as the sample complexity and the effective width of the student and the teacher.

(*ii*) Data distribution. *The measurement matrices $X_1, \dots, X_n \in \mathcal{S}_d(\mathbb{R})$ are independent and drawn from the* GOE(d) *distribution.*

(*iii*) Teacher distribution. *The spectral distribution of the teacher matrix $Z^* \in \mathcal{S}_d^+(\mathbb{R})$ almost surely converges as $d \to \infty$ to some measure μ^* with compact support. Moreover, $\|Z^*\|_{\rm op} \le C$ almost surely for some C independent of d. When necessary, we will decompose $\mu^* = (1 - \min(\kappa^*, 1))\delta + \min(\kappa^*, 1)\nu^*$, where ν^* has a bounded support on $\mathbb{R}^+$ with no mass at zero, and δ is the Dirac point mass at zero.*

(*iv*) Initialization. *The dynamics* (5) *is initialized with a random matrix $W_0 \in \mathbb{R}^{d\times m}$ such that the empirical spectral distribution of $W_0 W_0^\top \in \mathcal{S}_d^+(\mathbb{R})$ almost surely converges as $d \to \infty$ to some measure μ_0 with compact support. Moreover, the distribution of W_0 admits a density with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$.*

Therefore, we let the width of the teacher and student to vary proportionally with the dimension, leading to what we refer to as the extensive-width regime. In this work, we quantify overparameterization through the width of the student network, as measured by κ, which reflects how the number of parameters grows with the dimension. Additionally, since the teacher has $m^* d \sim \kappa^* d^2$ coefficients, it is natural to require $\Theta(d^2)$ measurements in order to recover it. Together with the choice of κ, κ^*, this setting allows us to study the effect of overparameterization on the performance of gradient flow dynamics and signal recovery.

From the observations (1), we aim to recover the teacher matrix Z^*. We will be interested in several scalar quantities in the high-dimensional limit, including the mean-square error (MSE) between the teacher and the student, and the empirical loss value (or training loss):

$$\mathrm{MSE} = \frac{1}{d}\mathbb{E}\left\|Z - Z^*\right\|_F^2, \qquad \mathrm{Loss}_{\rm train} = \frac{1}{2n}\mathbb{E}\sum_{k=1}^{n} \ell\big(\mathrm{Tr}(X_k Z), z_k\big), \tag{10}$$

with $Z = WW^\top$, and the expectation is taken with respect to the distribution of the data and the teacher. Note that, when the cost ℓ in equation (4) is quadratic, the MSE is directly proportional to the generalization error, since the matrices $X_1, \dots, X_n$ are Gaussian with isotropic covariance. When writing $\mathrm{MSE}(t)$ and $\mathrm{Loss}_{\rm train}(t)$ we will refer to the MSE and the loss evaluated with $W(t)$ being the solution of the gradient flow dynamics (5).

3. Main Results

This section is dedicated to our main results. We briefly outline its organization:

- In Section 3.1, we derive a set of effective dynamical equations governing gradient flow and Langevin dynamics in the high-dimensional limit. These equations provide a self-consistent description of the learning dynamics and serve as the basis for the analyses that follow.
- In Section 3.2, under a long-time simplification, we reduce the high-dimensional dynamics to an effective denoising gradient flow and analyze its asymptotic behavior. We characterize the spectral properties and performance of the estimator, and discuss the impact of overparameterization through the network width, as well as the emergence of a double descent phenomenon.
- In Section 3.3, we investigate the behavior of the long-time equations as the regularization vanishes. This leads to an exact characterization of the perfect recovery and interpolation thresholds.
- In Section 3.4, we study the unregularized setting. While a complete analytical characterization remains challenging, we formulate a conjecture for the perfect recovery threshold and support it with numerical evidence.
- In Section 3.5, we study the relationship between the Gaussian surrogate model and the original quadratic neural network formulation. We provide theoretical and numerical evidence supporting their asymptotic equivalence.

3.1 High-Dimensional Dynamics

In this section, we study the high-dimensional limit associated with the gradient flow dynamics (5). More precisely, we present a system of stochastic differential equations that is equivalent in distribution to the gradient flow dynamics.

3.1.1 General High-Dimensional Equations

We now present a high-dimensional characterization of the gradient flow dynamics. In the high-dimensional limit, the training dynamics admits an equivalent description in terms of a self-consistent stochastic system. In this description, the dynamics is driven by Gaussian processes and deterministic scalar functions, both determined self-consistently from expectations along the training trajectory.

The resulting system involves two coupled random processes. The first is a matrix-valued process $\overline{W}(t) \in \mathbb{R}^{d\times m}$, which can be interpreted as the student matrix and remains high-dimensional. The second is a scalar process $\overline{y}(t) \in \mathbb{R}$ corresponding to a typical label, defined as one of the training labels sampled uniformly at random. Together, these processes describe the evolution of the network weights and the associated labels.

For conciseness, we do not report the full system of equations here and refer to Section C.2 for a complete statement.

Claim 2 *Let $W(t) \in \mathbb{R}^{d\times m}$ be the solution of the dynamics (5), with observed labels $z_1, \dots, z_n$ and initial condition $W_0 \in \mathbb{R}^{d\times m}$. Consider an index K uniformly drawn in $\{1, \dots, n\}$, and the typical label defined as:*

$$y(t) = \mathrm{Tr}\big(X_K W(t) W(t)^\top\big). \tag{11}$$

Then, in the $d \to \infty$ limit, under Assumption 1, $\big(W, y\big) \overset{\text{distrib}}{=} \big(\overline{W}, \overline{y}\big)$, solution of the stochastic equations:

$$\mathrm{d}\overline{W}(t) = \left(\mathcal{H}(t) + r(t)Z^* - \int_0^t \Gamma(t,t')\overline{Z}(t')\mathrm{d}t'\right)\overline{W}(t)\mathrm{d}t - \nabla\Omega\big(\overline{W}(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{12}$$

$$0 = \int_0^t R(t,t')\overline{y}(t')\mathrm{d}t' + \eta(t) - m(t)y^* + \frac{2}{\alpha}\ell'\big(\overline{y}(t), z\big), \tag{13}$$

where $\overline{Z}(t) = \overline{W}(t)\overline{W}(t)^\top$, $z \sim P(\,\cdot\,|\,y^*)$, and $\overline{W}(t=0) = W_0$. The functions $\mathcal{H}$ and η are independent Gaussian processes respectively belonging to $\mathcal{S}_d(\mathbb{R})$ and $\mathbb{R}$, with covariances:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{2d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\,\mathcal{K}_Z(t,t'), \qquad \mathbb{E}\,\eta(t)\eta(t') = \mathcal{K}_y(t,t'), \tag{14}$$

and $r, \Gamma, R, m, \mathcal{K}_Z, \mathcal{K}_y$ are deterministic scalar functions expressed as expectations over the stochastic processes $\overline{W}, \overline{y}$.

In these equations, the random variables y^* and z respectively correspond to the true label and its noisy version, associated with the index K picked at random in Claim 2. In the high-dimensional limit, as a consequence of the central-limit theorem, y^* is a Gaussian variable with zero mean and variance:

$$\mathbb{E}\,y^{*2} = \lim_{d\to\infty}\frac{2}{d}\mathbb{E}\,\mathrm{Tr}\big(Z^{*2}\big), \tag{15}$$

and $z \sim P\big(\,\cdot\,\big|\,y^*\big)$.

This result suggests that, in the high-dimensional limit, the dynamics of $\big(W(t), y(t)\big)$ can be accurately described by $\big(\overline{W}(t), \overline{y}(t)\big)$. More precisely, for any fixed time horizon, averaged quantities of the original system can be computed from the previous equations with asymptotically vanishing error. To illustrate this fact, consider for instance the MSE and the empirical loss defined in (10). Then, for any fixed $T > 0$, both the differences:

$$\sup_{t\in[0,T]}\left|\mathrm{MSE}(t) - \frac{1}{d}\mathbb{E}\left\|\overline{Z}(t) - Z^*\right\|_F^2\right|, \qquad \sup_{t\in[0,T]}\left|\mathrm{Loss}_{\mathrm{train}}(t) - \frac{1}{2}\mathbb{E}\,\ell\big(\overline{y}(t), z\big)\right|, \tag{16}$$

converge to zero as $d \to \infty$.

In this set of equations, the dynamics for the student matrix $\overline{W}$ and the typical label $\overline{y}$ are driven by scalar functions. As mentioned, these functions can be themselves computed from the law of the random processes $\overline{W}, \overline{y}$, leading to a highly nonlinear and self-consistent dynamics, in the spirit of McKean–Vlasov processes (Chaintron and Diez, 2022).

We emphasize the generality of this result: the derivation of Claim 2 only makes use of the high-dimensional scaling chosen for the number of samples n and the Gaussian distribution of the data. It remains valid for any choice of regularization Ω, cost function ℓ,

and noise distribution P. Furthermore, since the equations we obtain still explicitly depend on the teacher matrix Z^* and the initialization of the dynamics W_0, our result remains independent of the choice of the distributions of these matrices: the requirements made in Assumption 1 are only necessary to guarantee that the dynamical equations in Claim 2 are well-posed.

Derivation of the equations. We derive the equations of Claim 2 in Section C.3. Our calculation is based on the rewriting of the dynamical partition function associated with the dynamics (5). One key point of our analysis is the possibility to average the partition function with respect to the sensing matrices, thanks to the Gaussian distribution hypothesis in Assumption 1. Finally, a saddle-point calculation in our high-dimensional setting allows to obtain the self-consistent set of equations in Claim 2. Therefore, as $d \to \infty$, we obtain the equality between the dynamical partition function of the gradient flow dynamics and that of the stochastic process $\big(\overline{W}, \overline{y}\big)$ of Claim 2. This result turns out to be equivalent to equality in distribution in the high-dimensional limit. More details can be found in Section C.1.

We insist on the fact that several steps in our calculation are non-rigorous, but build on objects and methods that have been applied in similar problems, see for instance Agoritsas et al. (2018) for the perceptron model, Mignacco et al. (2020) for Gaussian mixture classification and Bordelon and Pehlevan (2022) for wide neural networks. More generally, our calculation falls inside the class of dynamical mean field theory (DMFT), which gathers several methods originally used to derive dynamical equations for spin glass models (Sompolinsky and Zippelius, 1982; Cugliandolo and Kurchan, 1993). Since then, several works have succeeded in showing rigorously that these asymptotic equations are exact (Ben Arous and Guionnet, 1995; Ben Arous et al., 2006; Celentano et al., 2021; Gerbelot et al., 2024).

Dimensionality reduction. In similar studies of the finite-rank case $m = O(1)$, a common procedure is to write a low-dimensional set of equations on the correlations and the overlaps between the neurons, leading to a finite number of summary statistics (see for instance Celentano et al., 2021; Gerbelot et al., 2024; Montanari and Urbani, 2025). In our case, where m grows with the dimension, we believe that our system of equations cannot be reduced to a finite-dimensional one while still capturing the relevant dynamical quantities such as the MSE.

3.1.2 Learning the Second Layer

The method used to derive the high-dimensional dynamics in Claim 2 extends to the more general setting where a second layer of weights is also learned. To be more precise, consider the following predictor:

$$\mathrm{Tr}\big(X_k W D_a W^\top\big) = \frac{1}{m}\sum_{i=1}^{m} a_i \frac{(x_k^\top w_i)^2 - \|w_i\|^2}{\sqrt{d}}, \qquad W = \frac{1}{\sqrt{m}}\big(w_1, \dots, w_m\big) \in \mathbb{R}^{d\times m}, \tag{17}$$

where $a \in \mathbb{R}^m$ and $D_a \in \mathbb{R}^{m\times m}$ is the diagonal matrix with the same diagonal coefficients as a. As it is common in machine learning settings where all layers of the network possess trainable parameters, this vector can also be optimized. We then consider the joint

dynamics:

$$\mathrm{d}a(t) = -\vartheta d\,\nabla_a \mathcal{L}\big(a(t), W(t)\big)\mathrm{d}t - \nabla_a \Omega\big(a(t), W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B_a(t), \tag{18}$$

$$\mathrm{d}W(t) = -d\,\nabla_W \mathcal{L}\big(a(t), W(t)\big)\mathrm{d}t - \nabla_W \Omega\big(a(t), W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B_W(t), \tag{19}$$

where the loss function $\mathcal{L}$ from equation (4) is considered with the predictor given in equation (17). The regularization Ω is now a function of a, W, the constant $\vartheta > 0$ controls the learning rate of the dynamics of $a(t)$ with respect to $W(t)$, and B_a, B_W are independent standard Brownian motions over $\mathbb{R}^m$ and $\mathbb{R}^{d\times m}$, respectively.

Using the same method as in Claim 2, we derive an equivalent description of the dynamics in the high-dimensional limit, under Assumption 1. We recover similar self-consistent equations (given in Section C.2), but the effective dynamics in equation (12) is replaced by a joint stochastic evolution for two processes $\overline{a}, \overline{W}$:

$$\begin{aligned}\mathrm{d}\overline{a}(t) = \frac{\vartheta}{2}\mathrm{diag}\left(\overline{W}(t)^\top\left[\mathcal{H}(t) + r(t)Z^* - \int_0^t \Gamma(t,t')\overline{Z}(t')\mathrm{d}t'\right]\overline{W}(t)\right)\mathrm{d}t \\ - \nabla_a\Omega\big(\overline{a}(t), \overline{W}(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B_a(t)\end{aligned} \tag{20}$$

$$\begin{aligned}\mathrm{d}\overline{W}(t) = \left(\mathcal{H}(t) + r(t)Z^* - \int_0^t \Gamma(t,t')\overline{Z}(t')\mathrm{d}t'\right)\overline{W}(t)D_{\overline{a}(t)}\mathrm{d}t \\ - \nabla_W\Omega\big(\overline{a}(t), \overline{W}(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B_W(t),\end{aligned} \tag{21}$$

where $\overline{Z}(t) = \overline{W}(t)D_{\overline{a}(t)}\overline{W}(t)^\top$, and for $A \in \mathbb{R}^{m\times m}$, $\mathrm{diag}(A) \in \mathbb{R}^m$ denotes the vector composed of the diagonal elements of A. We refer to Section C.5 for a derivation of these equations. In the following, we do not analyze this particular setting and leave this study for further work.

Another option is to train the output layer only, while keeping the inner weights fixed. This corresponds to the random feature model, originally introduced by Rahimi and Recht (2007). This setting has been extensively analyzed in high dimensions (Gerace et al., 2020; Mei and Montanari, 2022; Hastie et al., 2022). In this regime, the optimization is convex and therefore simpler than when training the inner weights.

3.1.3 Simplified Setting: Quadratic Cost and Gaussian Label Noise

We now give an illustration of the system of equations of Claim 2 in a less general setting where we specify the cost function and the noisy channel that generates the labels. In this case, the full set of equations and order parameters is less cumbersome and will be presented. This setup will be studied in more detail when analyzing the gradient flow dynamics at long times, in Section 3.2.

Assumption 3 *In the following, we consider the setting:*

(*i*) Cost function. *Optimization is performed using the quadratic cost:* $\ell(y,z) = \frac{1}{2}(y-z)^2$.

(*ii*) Noisy channel. *The labels are generated using an additive Gaussian channel with variance* Δ*:*

$$P(z \mid y) = \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{1}{2\Delta}(y-z)^2\right). \tag{22}$$

Under this assumption, Claim 2 can be simplified and leads to the following set of self-consistent equations:

Claim 4 *Under Assumptions 1, 3, the system of equations* (12), (13) *can be written in the form:*

$$\mathrm{d}\overline{W}(t) = 2\left(\int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - \overline{Z}(t')\Big)\mathrm{d}t'\right)\overline{W}(t)\mathrm{d}t - \nabla\Omega\big(\overline{W}(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{23}$$

$$\overline{y}(t) = y^* + \sqrt{\Delta}\zeta + \int_0^t \mathrm{d}t'\, R(t,t')\left(\xi(t') - \sqrt{\Delta}\zeta + \left(\frac{m_Z(t')}{Q_*} - 1\right)y^*\right), \tag{24}$$

where $\overline{Z}(t) = \overline{W}(t)\overline{W}(t)^\top$*, and* $y^* \sim \mathcal{N}(0, 2Q_*), \zeta \sim \mathcal{N}(0,1)$*,* $\mathcal{G}(t) \in \mathcal{S}_d(\mathbb{R})$ *and* $\xi(t) \in \mathbb{R}$ *are independent centered Gaussian variables and processes with covariances:*

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{2\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\left(C_Z(t,t') - m_Z(t) - m_Z(t') + Q_* + \frac{\Delta}{2}\right), \tag{25}$$

$$\mathbb{E}\,\xi(t)\xi(t') = 2C_Z(t,t') - \frac{2}{Q_*}m_Z(t)m_Z(t'). \tag{26}$$

Finally, the following relationships close the system of equations:

$$\begin{aligned} Q_* &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}), & C_Z(t,t') &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\overline{Z}(t)\overline{Z}(t')\big), \\ m_Z(t) &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\overline{Z}(t)Z^*\big), & R(t,t') &= \delta(t-t') - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,\overline{Z}(t)}{\partial H(t')}\bigg|_{H=0}\right). \end{aligned} \tag{27}$$

This last set of equations defines averaged quantities with respect to $\overline{W}(t)$: covariances, overlaps and responses. δ is the Dirac delta distribution supported at 0 and the equation defining R is to be interpreted in the sense of distributions (see Section C.1 for more details). The response $R(t,t')$ quantifies the average change of $\overline{Z}(t)$ in response to an infinitesimal perturbation $H(t') \in \mathcal{S}_d(\mathbb{R})$ added to the Gaussian noise $\mathcal{G}(t') \to \mathcal{G}(t') + H(t')$ into equation (23). More precisely, the partial derivative in equation (27) can be seen as the differential of the function mapping the perturbation $H(t') \in \mathcal{S}_d(\mathbb{R})$ to the corresponding perturbed solution $\overline{Z}(t)$. Then R is computed by taking the trace over linear maps acting on $\mathcal{S}_d(\mathbb{R})$.

Finally, we remark that in equation (24) the label $\overline{y}(t)$ is explicitly expressed as a linear combination of the independent centered Gaussian variables $y^*, \zeta, \xi(t)$. Therefore, $(\overline{y}(t))_{t\geq 0}$ is itself a Gaussian process with zero mean. This shows that in the high-dimensional limit, a label drawn at random remains Gaussian along the gradient flow dynamics. Figure 2 presents a numerical confirmation of this result.

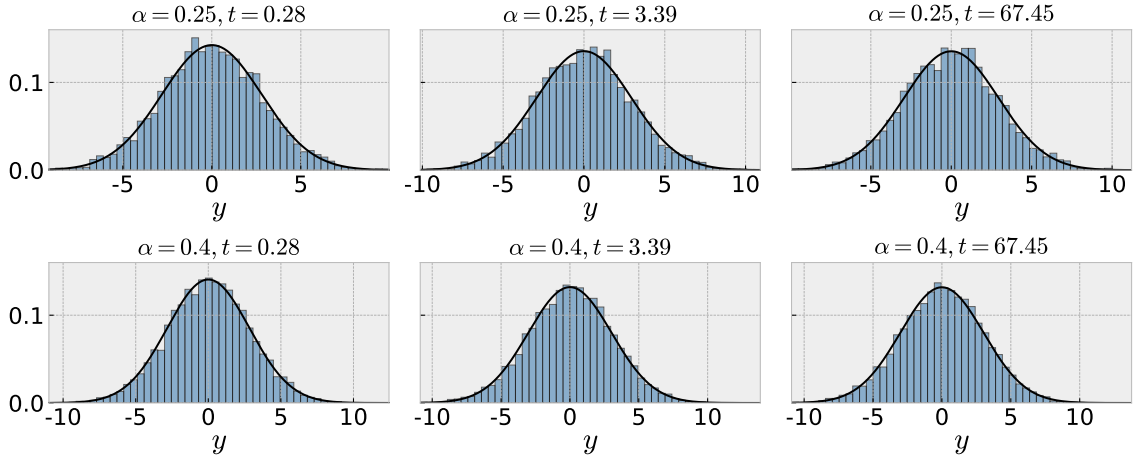

Figure 2: Empirical distribution of the student labels $y_k(t) = \mathrm{Tr}\big(W(t)W(t)^\top X_k\big)$ during optimization with gradient descent, defined in equation (6), with parameters $\kappa = 0.4, \kappa^* = 0.3$, $d = 150$, quadratic cost and no regularization, for two values of α and three values of time t. The black curve corresponds to the Gaussian density with zero mean and variance equal to the empirical variance of the labels.

3.1.4 Recovering Population Dynamics

A natural setting to simplify the equations corresponds to the population case, where the student has access to an infinite number of observations. In the expression of the loss in equation (4), this corresponds to replacing the average over the n data points with the expectation over the data distribution. Regarding our high-dimensional dynamics, these equations can be recovered by taking the $\alpha \to \infty$ limit in those of Claim 4. In this case we obtain that the Gaussian noise $\mathcal{G}(t)$ vanishes and the response R simplifies to $R(t, t') = \delta(t - t')$, i.e., the memory effect disappears and equations (23), (24) simply write:

$$\mathrm{d}\overline{W}(t) = 2\Big(Z^* - \overline{W}(t)\overline{W}(t)^\top\Big)\overline{W}(t)\mathrm{d}t - \nabla\Omega\big(\overline{W}(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{28}$$

$$\overline{y}(t) = \frac{m_Z(t)}{Q_*}y^* + \xi(t). \tag{29}$$

This first equation on $\overline{W}$ is precisely the one we would obtain by writing the Langevin dynamics (5) on the population loss. This evolution is very similar to those found in the population limit for shallow quadratic neural networks (Gamarnik et al., 2019; Sarao Mannelli et al., 2020; Martin et al., 2024). In addition, it can be shown that the equation on the typical label $\overline{y}(t)$ corresponds to the evolution of the student matrix $\overline{W}(t)\overline{W}(t)^\top$ projected on a random direction: indeed, when it has access to an infinite number of observations, the student does not correlate with any particular example.

In the gradient flow setting ($\beta = \infty$), and with the choice of ℓ_2-regularization, the dynamics (28) on $W(t)$ can be interpreted as an Oja flow (see Section 4 for a definition and properties). In the following section, we study the equations given in Claim 4 in the

long-time limit. One special case of our analysis will cover this population limit $\alpha \to \infty$. Further analysis can be found in Section D.6.

3.2 Long-Time Analysis of the Regularized Dynamics

This section is dedicated to the study of the long-time asymptotics of the set of equations from Claim 4, with the choice of ℓ_2-regularization:

$$\Omega(W) = \lambda \mathrm{Tr}(WW^\top). \tag{30}$$

The following results focus on the gradient flow setting ($\beta = \infty$), and we refer to Section 3.2.7 for an analysis of the Langevin dynamics at finite temperature.

The goal is to understand the $t \to \infty$ limit of the gradient flow dynamics (5), and to describe it in the high-dimensional limit. To do so, we start from the dynamical equations of Claim 4. Such equations, often referred to as DMFT equations, are known to be difficult to analyze and typically require a guess on the structure of the dynamics. In what follows, we introduce such a guess, which we refer to as the *steady-state assumption*.

3.2.1 Steady-State Assumption

The way to simplify the equations is to assume that the memory effect in equation (23) disappears at long times, i.e., the weight of the integral over $[0, t]$ concentrates on times t' that are close to t. In this case we say that the dynamics only possesses a short-term memory. We formalize this idea through the following assumption.

Assumption 5 *The self-consistent dynamics given in Claim 4 satisfy:*

(i) Response decay. *The response function $R(t, t')$ decays fast enough to zero as $t - t' \to \infty$.*

(ii) Constant noise. *The Gaussian process $\mathcal{G}(t)$ converges fast enough as $t \to \infty$ so that it can be considered constant in the evolution equation for $W(t)$.*

Although we do not exactly quantify what fast enough means, we assume that these conditions are strong enough to justify the following long-time approximation of the memory term in (23):

$$\int_0^t R(t, t')\Big(\mathcal{G}(t') + Z^* - Z(t')\Big)\mathrm{d}t' \underset{t\to\infty}{\approx} r_\infty\Big(\sqrt{\xi}\mathcal{G} + Z^* - Z(t)\Big), \tag{31}$$

where the constants ξ, r_∞ are such that:

$$r_\infty = \lim_{t\to\infty} \int_0^t R(t, t')\mathrm{d}t', \qquad \lim_{t\to\infty} \mathbb{E}\, \mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t) = \frac{\xi}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{i'j}\delta_{ij'}\big), \tag{32}$$

and $\mathcal{G} \sim \mathrm{GOE}(d)$. Despite being a strong assumption on the structure of the dynamics, it is directly motivated by a line of work on DMFT and generalized Langevin equations similar to (23). More precisely, our approach is closely related to the *time-translational invariance* (TTI) approximation, which has been used to analyze the asymptotic behavior of DMFT equations (Sompolinsky and Zippelius, 1982; Sompolinsky et al., 1988; Bordelon

et al., 2024). In these works, such assumptions are introduced as physically motivated ansätze and later verified through consistency checks and numerical simulations. Also note that related approaches have been studied rigorously in simpler settings (Celentano et al., 2021; Fan et al., 2025; Chen and Shen, 2025).

Our assumption should be understood in the same spirit: it represents a conjectured structural property of the long-time dynamics, and is natural when describing systems that rapidly reach a steady-state regime. Importantly, the consequences of this simplification will be systematically compared to high-dimensional numerical simulations of the gradient flow dynamics (5). We refer to Section D.1 for a more detailed discussion of this assumption.

This assumption leads to an effective dynamics, which we adopt as the starting point for the long-time analysis:

$$\dot{W}(t) = 2r_\infty\Big(\sqrt{\xi}\mathcal{G} + Z^* - W(t)W(t)^\top\Big)W(t) - 2\lambda W(t). \tag{33}$$

In addition, the self-consistent expressions of the covariance of $\mathcal{G}$ and the function R in equations (25) and (27) lead to the equations on ξ, r_∞:

$$\xi = \frac{1}{2\alpha}\left(\mathrm{MSE} + \frac{\Delta}{2}\right), \tag{34}$$

$$r_\infty = 1 - \frac{1}{\alpha d^2}\lim_{t\to\infty}\int_0^t \mathrm{Tr}\left(\left.\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\right|_{H=0}\right)\mathrm{d}t', \tag{35}$$

where H is a perturbation entering additively in the drift term multiplying $W(t)$ in equation (33). Then, one can express the solution of the dynamics (33) as a function of the variables ξ, r_∞. Using the expression of ξ and the definition of r_∞, one can deduce self-consistent equations on these two variables.

The dynamics in equation (33) is known as an Oja flow, a nonlinear matrix flow that has been studied in prior works (Yan et al., 1994; Bodin and Macris, 2023; Martin et al., 2024). Despite its nonlinearity, this equation admits a closed-form solution, and its convergence properties are well understood. We devote Section 4 to the study of this flow and provide new results that allow us to derive a closed system of finite-dimensional equations for r_∞ and ξ. We present these equations in Claim 6 and derive them in Section D.

A denoising formulation. Interestingly, the dynamics (33) can be viewed as a denoising problem of the matrix Z^* corrupted by the Gaussian noise $\mathcal{G}$, solved through regularized gradient flow. This interpretation is similar to the one in the replica calculation done by Maillard et al. (2024) for the same model, where the inference problem was mapped onto a matrix denoising formulation in the Bayes-optimal setting.

Remark that this denoising formulation is very similar to the population dynamics given in equation (28) when considering Ω to be the ℓ_2-regularization and the gradient flow setting ($\beta = \infty$). Equation (33) introduces two additional parameters: an effective noise variance ξ and a time reparameterization r_∞. Interestingly, the expression for ξ in equation (34) suggests that the noise arises from the finite number of training samples (through the parameter α), and the label noise Δ, which prevents a clean observation of the teacher labels.

At long times, the simplified dynamics reveal that, near the point of convergence, the landscape of the regularized empirical loss (4) exhibits the same landscape structure as

a regularized matrix denoising problem. In this regime, the quantity r_∞ plays a central role: it represents the local curvature of the landscape around the limiting point and thus quantifies the sharpness of the minimum and the associated convergence rate.

3.2.2 Set of Equations at Long Times

In the following, we give the set of equations resulting from the steady-state assumption (see Section D for their derivation). Before doing so, we define the following operator: for a symmetric matrix $A \in \mathcal{S}_d(\mathbb{R})$ with spectral decomposition $A = U\mathrm{diag}(\lambda_1, \dots, \lambda_d)U^\top$ and $\lambda_1 \geq \cdots \geq \lambda_d$, we define for $m \leq d$:

$$A^+_{(m)} = U\mathrm{diag}\big(\lambda_1^+, \dots, \lambda_m^+, 0, \dots, 0\big)U^\top, \qquad \lambda^+ = \max(\lambda, 0). \tag{36}$$

The matrix $A^+_{(m)}$ selects the m largest positive eigenvalues of A, and is known to be the best rank-m positive semidefinite approximation of A for the Frobenius norm.

As a consequence of the simplifications introduced in Section 3.2.1, the long-time behavior of the dynamics can be characterized as follows.

Claim 6 *Consider the variables ξ, r_∞ defined in equation (34), (35), and set $q = \lambda/r_\infty$. Define μ_ξ to be the asymptotic spectral distribution of the random matrix $Z^* + \sqrt{\xi}\mathcal{G}$, where $\mathcal{G} \sim \mathrm{GOE}(d)$. Then, under Assumptions 1, 3 and 5 in the $d \to \infty$ limit, ξ, q solve the equations:*

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x), \tag{37a}$$

$$1 = \frac{\lambda}{q} + \frac{1}{\alpha}\int_{\max(q,\omega)} (x - q)h_\xi(x)\mathrm{d}\mu_\xi(x), \tag{37b}$$

$$2\alpha\xi - \frac{\Delta}{2} = Q_* + \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x) + 4\xi\int_{\max(q,\omega)} (x - q)h_\xi(x)\mathrm{d}\mu_\xi(x), \tag{37c}$$

where h_ξ is the Hilbert transform of μ_ξ (see Definition 29) and:

$$Q_* = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}) = \int x^2 \mathrm{d}\mu^*(x). \tag{38}$$

Moreover, for almost all initializations, the limit of the dynamics (23) is given by:

$$Z_\infty = \Big(Z^* + \sqrt{\xi}\mathcal{G} - qI_d\Big)^+_{(m)}, \tag{39}$$

with $\mathcal{G} \sim \mathrm{GOE}(d)$. The MSE and the training loss are given by:

$$\mathrm{MSE} = 2\alpha\xi - \frac{\Delta}{2}, \qquad \mathrm{Loss}_{\mathrm{train}} = \frac{\lambda^2\alpha\xi}{q^2}. \tag{40}$$

More precisely, the measure μ_ξ corresponds to the free additive convolution between μ^*, the asymptotic spectral measure of the teacher Z^* and a semicircular density of variance ξ (see Biane, 1997, and Section B.1.3 for more details).

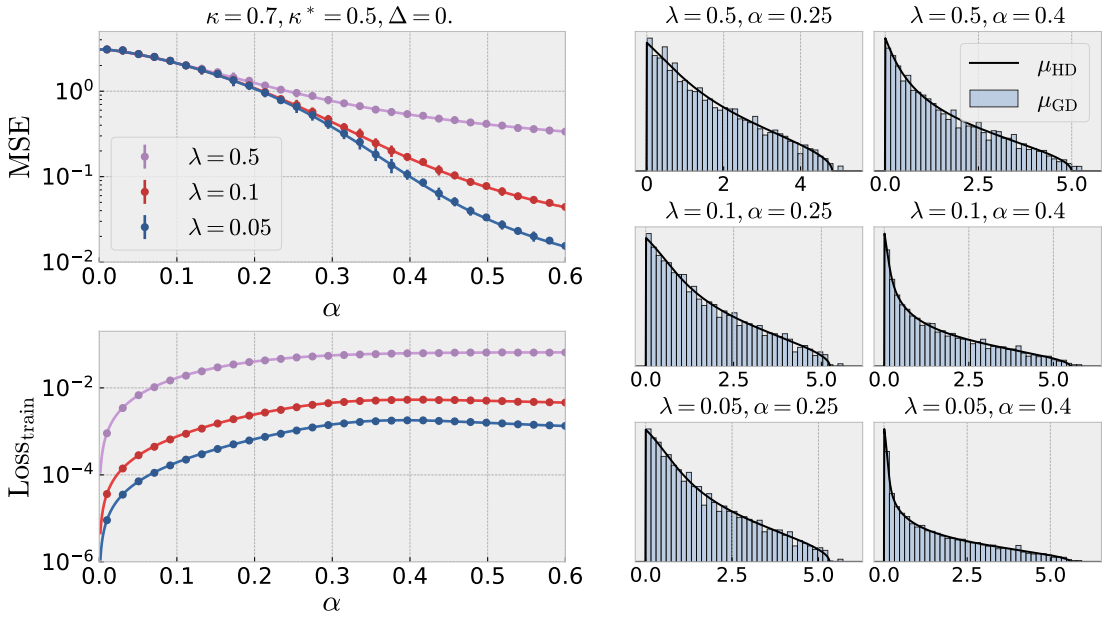

Figure 3: Comparison between simulations of gradient descent, defined in equation (6), and numerical integration of the system of equations (37), for $\kappa = 0.7, \kappa^* = 0.5$ and zero label noise. Gradient descent results are averaged over 10 (left) and 50 (right) realizations of the initialization, teacher and data. Left: MSE and empirical loss value as a function of α, for different values of λ. Dots correspond to GD simulations and full lines to the solution of equations (37). Right: eigenvalue distribution of $Z = WW^\top$ reached by GD (blue), restricted to its nonzero eigenvalues, and associated density computed with the asymptotic spectral distribution of the matrix in equation (39) (black line) for three values of λ and two values of α.

In contrast with the high-dimensional formulation of Claim 4, the system (37) is now finite-dimensional and involves only scalar quantities. The ambient dimension no longer appears explicitly, and the long-time limit of the dynamics is described by a small number of order parameters.

One of the key results is the limit of the flow found in equation (39). It shows that, asymptotically, the student matrix selects the m largest positive eigenvalues of a noisy version of the teacher matrix, with an eigenvalue shift that is characteristic of the regularized dynamics. In a similar fashion, in the system of equations (37), the variable ω selects a mass κ of the measure μ_ξ.

Claim 6 is based on assumptions regarding the long-time behavior of the gradient flow dynamics: it is still an open question to prove rigorously that these assumptions are verified. However, as a result of a large number of numerical simulations, we believe that this result holds for any value in our set of parameters $\kappa, \kappa^*, \alpha, \lambda, \Delta$, as soon as the regularization strength remains positive. In addition to Figures 3, 5 that compare the equations of Claim 6 with the results of gradient descent simulations and show excellent agreement, we provide

additional numerical evidence in Section J.2.1. Further details on how to simulate the system of equations (37) are given in Section J.1.2.

Additionally, let us remark that the assumption we made on the dynamics may not be specific to the choice of our setting: ℓ_2-regularization, Gaussian label noise and quadratic cost. In the general gradient flow case, the dynamics obtained in Claim 2 can also be approximated by a similar dynamics. This should lead to a comparable, although more complicated, set of equations as the one we present here. A sizable challenge would then be to validate numerically or theoretically these approximations in the general case.

For completeness, and in a spirit of coherence with the results of Section 3.1, in which the high-dimensional limit is first taken before the long-time limit, we also show that as soon as the dynamics is approximated by equation (33), the set of equations we derive is the same no matter in which order the limits are taken. This result also ensures robustness regarding the behavior of the dynamics (33): the only relevant timescale for the dynamics is the one that we study in this section. More details on this can be found in Section D.3.

Universality over the teacher's distribution. Interestingly, it appears that the system of equations (37) holds no matter the choice of the teacher distribution, provided that its spectral density converges as $d \to \infty$. However, there are several settings of interest that this result does not include but could be potentially generalized to:

- *Power-law teacher.* In the case where the spectrum of Z^* exhibits a power-law behavior, one has to take into account finite-dimensional corrections to obtain a contribution from the large eigenvalues of the teacher. This generalization has successfully been applied by Defilippis et al. (2025) for the empirical risk minimization problem in the same setting as ours.

- *Finite-width teacher.* Although we consider an extensive-width teacher ($m^* \sim \kappa^* d$), Claim 6 should remain valid when m^* remains of order one, in which case the teacher's spectral distribution μ^* would collapse onto a Dirac mass at zero. As it has been shown by Sarao Mannelli et al. (2020); Bonnaire et al. (2025) in the case $m^* = 1$, this setting only requires a number of observations proportional to d (and not d^2 as in our case). Therefore, we conjecture that our setting is unable to capture the finite-width case.

3.2.3 Overparameterization and Global Optimality

We shall now give some remarks on the impact of the overparameterization of the student network (which is controlled by the parameter κ) on the performance of the gradient flow estimator.

First of all, remark that our system of equations (37) only depends on κ through the threshold ω that selects the m largest positive eigenvalues of the noisy teacher $Z^* + \sqrt{\xi}\mathcal{G}$. Due to the dependence of equations (37b), (37c) on ω, it can be shown that for a given set of parameters $\kappa^*, \lambda, \alpha, \Delta$, there exists a value $\kappa_{\min}$ such that, as soon as $\kappa \geq \kappa_{\min}$, the solution of the system (37) does not depend on κ and is the same in the case $\kappa \geq 1$. In this region, gradient flow is able to converge to the global minimizer of the training loss over the set of PSD matrices. This solution has rank $\sim \kappa_{\min} d$, which can be expressed:

$$\kappa_{\min} = \int_q \mathrm{d}\mu_\xi(x). \tag{41}$$

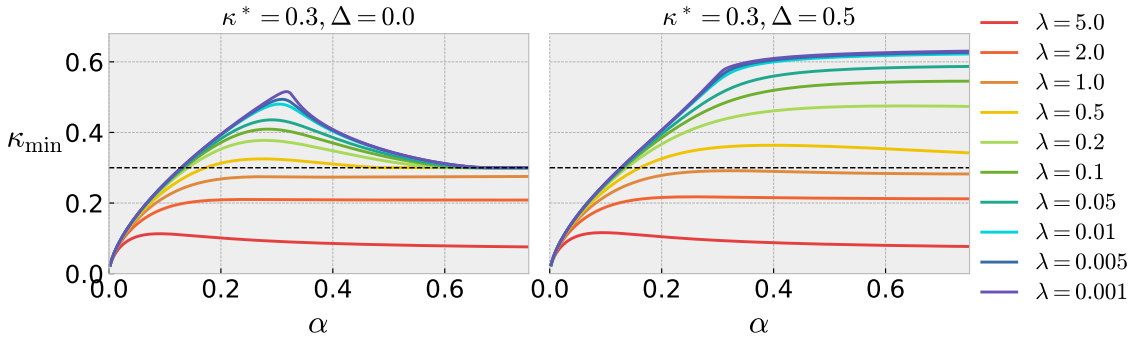

Figure 4: Rank of the solution $\kappa_{\min}$ obtained with $\kappa = 1$, as a function of α and for different values of λ. $\kappa^* = 0.3$ (horizontal dashed line), $\Delta = 0$ (left) and $\Delta = 0.5$ (right). Curves obtained by simulating the system (37) and using equation (41) to compute $\kappa_{\min}$. Above these curves, the solution found by gradient flow does not depend on κ.

In this equation, the parameters q and ξ are obtained by solving equations (37b), (37c) at $\kappa = 1$, for fixed values of the parameters $\alpha, \kappa^*, \lambda, \Delta$. Therefore, only a mild overparameterization ($\kappa \geq \kappa_{\min}$, compared to $\kappa \geq 1$) is necessary to reach the global minimizer of the regularized loss. This conclusion is non-trivial: for general functions of the quadratic form $WW^\top$, it is not guaranteed that the student matrix W finds the global minimizer of the loss over all PSD matrices, even if it has a high-enough rank to recover it. We give more details on this result in Section D.5.

On the other hand, for $\kappa \leq \kappa_{\min}$, the rank of the student is too small to recover the global minimizer (in terms of $Z = WW^\top$). Then, gradient flow seems to converge to a solution that depends on κ (and has maximal rank). Although we cannot directly conclude that this solution corresponds to a global minimizer of the regularized loss (now expressed in terms of W), the analysis of the Langevin dynamics in Section 3.2.7 in the low temperature regime suggests that this is indeed the case. We refer to this section for more details.

In Figure 4 we plot the threshold $\kappa_{\min}$ as a function of α, for a wide range of λ, and with $\kappa^* = 0.3$. The first observation is that this function decreases when increasing λ: a stronger regularization leads to a lower-rank global minimizer. This behavior is not a surprise, since this minimizer is obtained by optimizing the regularized loss over all positive semidefinite matrices Z. In this case, the ℓ_2-regularization on W translates into a nuclear norm penalty for $Z = WW^\top$, which is known to favor low-rank solutions (Fazel et al., 2001; Recht et al., 2010). In addition, in the presence of label noise (right panel), the rank of the global minimizer tends to increase, especially for larger values of α: in this region, the model needs more degrees of freedom to compensate for the variability induced by the noise.

Link with empirical risk minimization. Erba et al. (2025b) studied the global minimizer of the regularized empirical loss, in the case $\kappa \geq 1$. In this regime, it is known that the gradient flow always converges to the global minimizer of the loss (over all PSD matrices $Z = WW^\top$). In Section D.4, we show that the system of equations we derive in Claim 6 matches theirs.

This agreement, although expected, is of interest as the two sets of equations were derived using different approaches. In their work, the result follows from an exact analysis of the fixed point equations associated with an approximate message passing (AMP) iteration. In addition, the authors provide a study of the stability of the AMP fixed point and derive a condition that coincides with ours for the steady-state assumption. We show that whenever $\kappa \geq 1$, the stability condition is met.

3.2.4 Stability of the Steady-State Solution

In Section E, we provide theoretical insights to assess whether the steady-state assumption holds. To do so, a common approach is to study the response operator (also known as the susceptibility) associated with a perturbation of the steady-state solution (see for instance Mézard et al., 1987). This operator characterizes the robustness and convergence of the dynamics under a perturbation, which is particularly relevant since the steady-state dynamics (33) was derived in a perturbative way from the high-dimensional system in Claim 4. However, this approach only covers the steady-state dynamics, not its stability with respect to the high-dimensional system of equations of Claim 4.

In our case, the susceptibility operator is defined as:

$$\mathcal{X} = \frac{\partial Z_\infty}{\partial H}\bigg|_{H=0}, \qquad Z_\infty = \Big(Z^* + \sqrt{\xi}\mathcal{G} - qI_d + H\Big)^+_{(m)}. \tag{42}$$

Z_∞ is the limit of the steady-state dynamics obtained under a perturbation H, and the susceptibility can simply be interpreted as the differential of the map $H \in \mathcal{S}_d(\mathbb{R}) \mapsto Z_\infty \in \mathcal{S}_d(\mathbb{R})$. In Section E.1, we analyze both the spectrum and the normalized Frobenius norm of $\mathcal{X}$, that allows to investigate stability with respect to average and worst-case perturbations. Overall stability is guaranteed as soon as the spectrum of $\mathcal{X}$ (or its Frobenius norm) remains bounded. We derive the following results:

- In the region where $\kappa \geq \kappa_{\min}$, the spectrum of the susceptibility remains in $[0, 1]$, and it can be shown that the steady-state dynamics remains stable both at finite and infinite dimension.
- When $\kappa \leq \kappa_{\min}$, there exists a small proportion of unstable modes, with susceptibility eigenvalues diverging with the dimension. Overall, this leads to an average susceptibility of order $\log d$. This suggests that at finite d, the steady-state dynamics remains stable, but with a potential instability occurring in the high-dimensional limit.

However, this does not imply that the earlier approximation fails. Stability should instead be evaluated with respect to the original high-dimensional dynamical equations, rather than the reduced steady-state equations alone. A more accurate analysis would require treating perturbatively the system of Claim 4, and linearize the dynamics around the steady-state solution. In Section E.3, we explain how such a calculation can be carried out. In this setting, the specific structure of the high-dimensional perturbation may regularize the unstable modes observed in the steady-state solution.

Based on extensive numerical simulations (see Section J.2), we believe that this is indeed what happens in practice. This conclusion is supported by systematic comparisons between the theoretical predictions and empirical learning curves for averaged quantities, such as

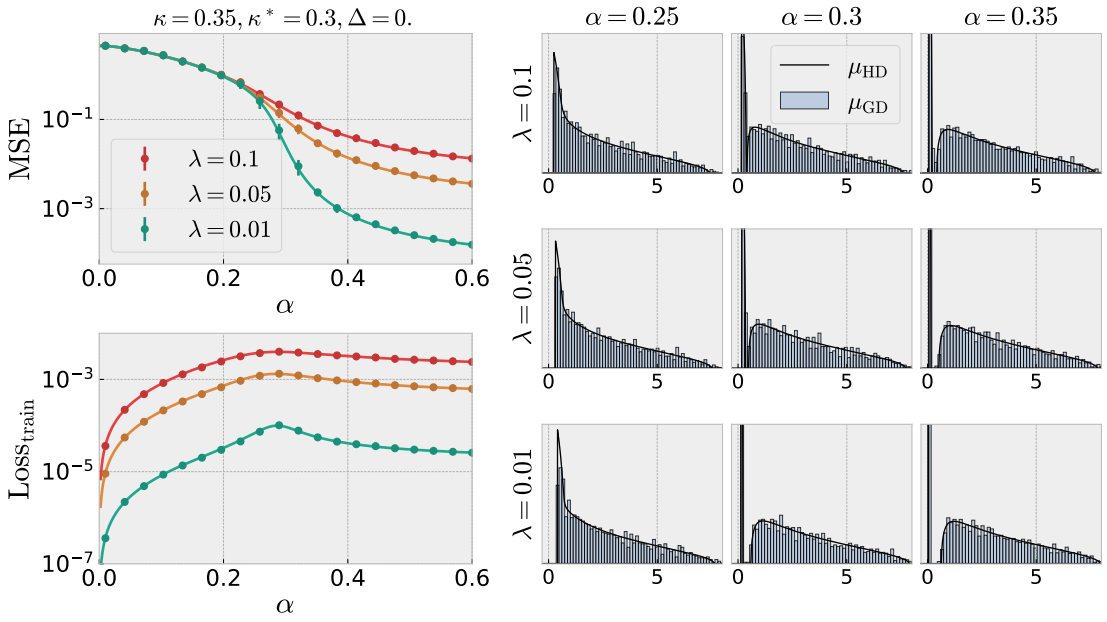

Figure 5: Comparison between simulations of gradient descent, defined in equation (6), and numerical integration of the system of equations (37), for $\kappa = 0.35, \kappa^* = 0.3$ and zero label noise. Gradient descent results are averaged over 10 (left) and 50 (right) realizations of the initialization, teacher and data. Left: MSE and empirical loss value as a function of α for different values of λ. Dots correspond to GD simulations and full lines to the solution of the system (37). Right: eigenvalue distribution of $Z = WW^\top$ reached by GD (blue), restricted to its nonzero eigenvalues, and associated density computed with the asymptotic spectral distribution of the matrix in equation (39) (black line) for three values of λ and α.

MSE and training loss, as well as eigenvalue distributions of the gradient flow predictors. Across a wide range of parameter values, we observe an excellent agreement.

To support this claim, we compare in Figure 5 the behavior of gradient descent at convergence with the numerical integration of the system (37). The choice of $\kappa = 0.35$ and $\kappa^* = 0.3$ ensures that most of the values of α fall in the regime $\kappa \leq \kappa_{\text{min}}$ (this can be checked in Figure 4). As shown in Figure 5, the agreement between theory and simulations is excellent, both for averaged quantities (MSE and empirical loss, left panel) and spectral distribution (right panel). In addition, for large values of α, the spectral distribution of the student matrix develops a spike away from zero. This effect can be understood from equation (39): when ξ is small (corresponding to large α and near-zero MSE), the spectrum of $Z^* + \sqrt{\xi}\mathcal{G}$ splits into two bulks. Since κ is slightly larger than κ^*, the student matrix recovers the bulk associated with Z^*, along with a small fraction of the second, corresponding to a Gaussian matrix with small variance, producing the observed spike.

Finally, a complementary approach to assess the validity of the steady-state approximation is to check *a posteriori* whether the assumptions made in Section 3.2.1 are satisfied. In Section D.1, we qualitatively relate Assumption 5 to the fast convergence of the matrix

$Z(t)$. This motivates the study of the convergence rates associated with the steady-state solution. In finite dimension, we show in Section E.2.1 that the convergence is exponentially fast, but that there exists a few directions with a relaxation time diverging with the dimension. As a consequence, in Section E.2.2, we study these convergence rates after taking the high-dimensional limit and show that in this case they are degraded into a power-law decay. More precisely, we show that:

$$\lim_{d\to\infty} \frac{1}{d}\big\|Z(t) - Z_\infty\big\|_F^2 \underset{t\to\infty}{=} \begin{cases} \Theta(t^{-3}), & \text{if } \kappa > \kappa_{\min}, \\ \Theta(t^{-1}), & \text{if } \kappa < \kappa_{\min}. \end{cases} \tag{43}$$

In the underparameterized region, the convergence is much slower, hence unveiling the existence of a new dynamical regime. To the best of our knowledge, these asymptotics were not known before and come as new instances of scaling laws for optimization dynamics.

Beyond the steady-state ansatz. In a complementary approach to investigate the validity of our assumptions, we propose in Section G a more general approximation of the dynamics, which is often referred to as *aging* in the statistical physics literature (see for instance Cugliandolo and Kurchan, 1993; Ben Arous et al., 2001; Sarao Mannelli et al., 2019; Altieri et al., 2020). The idea is to decompose the dynamics between a steady-state part and another regime which is very slow. Assuming a separation of timescales as well as a quasi-static equilibrium for the slow dynamics, we derive a more general set of self-consistent equations than the one in Claim 6. However, these equations involve a matrix-valued distribution whose analysis in the high-dimensional limit is not tractable in general. While we do not analyze these equations further, the excellent agreement between our numerical simulations and the steady-state solution suggests that this more general solution coincides with the steady-state one (physically, one would say that aging is absent). Making this identification explicit from the aging equations is left for future work.

3.2.5 Overfitting and Double Descent

Several numerical simulations suggest the presence of overfitting during the dynamics, that is, a positive gap:

$$\delta_{\text{MSE}} = \text{MSE}_\infty - \inf_{t\geq 0} \text{MSE}(t). \tag{44}$$

Our simulations of the gradient descent algorithm (6) show that this phenomenon already appears in the noiseless setting but is amplified in the presence of label noise ($\Delta > 0$). Figure 6 features the MSE as a function of time for two different label noises $\Delta > 0$, and for different values of α, revealing the overfitting phenomenon. It shows that for small values of α, when the student learns with few data, the model exhibits mild overfitting. As the sample complexity increases, the gap between the final MSE and its time-optimal value grows, revealing a progressively stronger overfitting.

In addition, for large values of Δ (right panel in Figure 6), the monotonicity of the MSE with α breaks. This is characteristic of the double descent phenomenon (Belkin et al., 2019; Nakkiran et al., 2021): as the number of observations increases (up to a certain point), the estimator fits all the data points, leading to poor generalization. For larger values of α, fitting is not possible anymore, and the student starts to learn the latent structure of the labels. In this double descent regime, Figure 7 features the dependence of the MSE

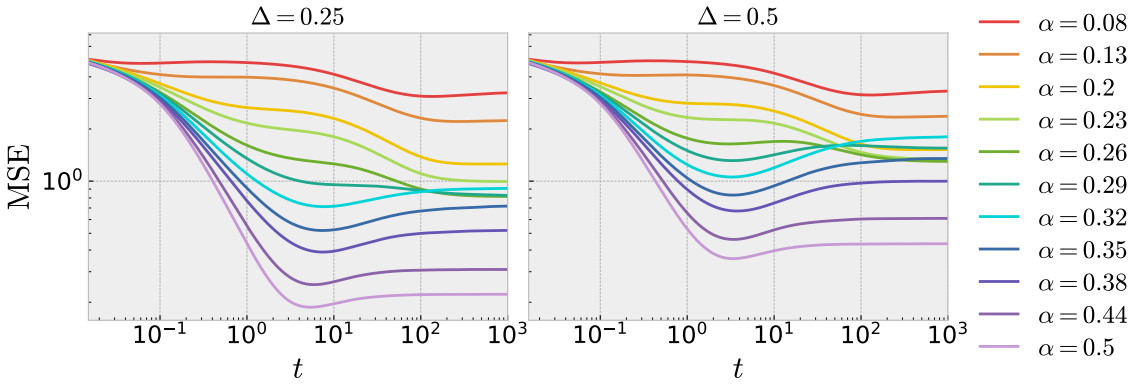

Figure 6: MSE as a function of time for gradient descent trajectories (see equation 6) averaged over 10 realizations of the initialization, teacher and data. Parameters $\kappa = 0.4, \kappa^* = 0.3, \lambda = 0.01$, $\Delta = 0.25$ (left) and 0.5 (right). Overfitting is characterized by portions where the MSE increases with time.

on α for different values of label noise Δ (left panel) and regularization strength λ (right panel). While such double descent curves are often plotted as a function of the number of parameters, we use here the sample complexity α to remain consistent with the rest of the paper. The figure also highlights the interpolation peak in the limit $\lambda \to 0^+$ (vertical dashed line), that we derive in Section 3.3.

Interestingly, overfitting is present even when regularizing the dynamics, but is reduced when increasing the regularization strength λ, within the range of values considered in Figure 7, right panel. This observation is consistent with known results in linear problems (Krogh and Hertz, 1991; Nakkiran et al., 2020; Mei and Montanari, 2022) and modern neural networks (Nakkiran et al., 2021; Zhang et al., 2019; D'Angelo et al., 2024).

Unfortunately, we are not able to precisely characterize the parameter regime in which double descent occurs. For instance, the left panel of Figure 7 shows that small values of Δ do not lead to this phenomenon. This suggests that the emergence of double descent may depend on a Δ-dependent scale in the regularization strength λ. However, the system of equations given in Claim 6 matches the empirical curves almost exactly, indicating that double descent is already encoded in the simplified dynamics of equation (33). Therefore, a more quantitative understanding of overfitting and double descent could, in principle, be obtained by analyzing the generalization properties of these denoising dynamics.

Overall, our results provide a characterization of double descent in a genuinely nonlinear model. Previous theoretical studies have analyzed this phenomenon in linear regression and least-squares settings (Nakkiran et al., 2020; Belkin et al., 2020; Wu and Xu, 2020; Derezinski et al., 2020; Hastie et al., 2022; Bach, 2024a) and in random feature models (d'Ascoli et al., 2020a,b; Gerace et al., 2020; Adlam and Pennington, 2020a,b; Mei and Montanari, 2022). Closest to our setting is the recent work of Erba et al. (2025b), who study overparameterized quadratic networks and also observe a double descent phenomenon. However, their analysis builds on a convex relaxation, which allows the problem to be treated within a linear setting. Our work instead goes beyond linearized models and provides, to the best of our

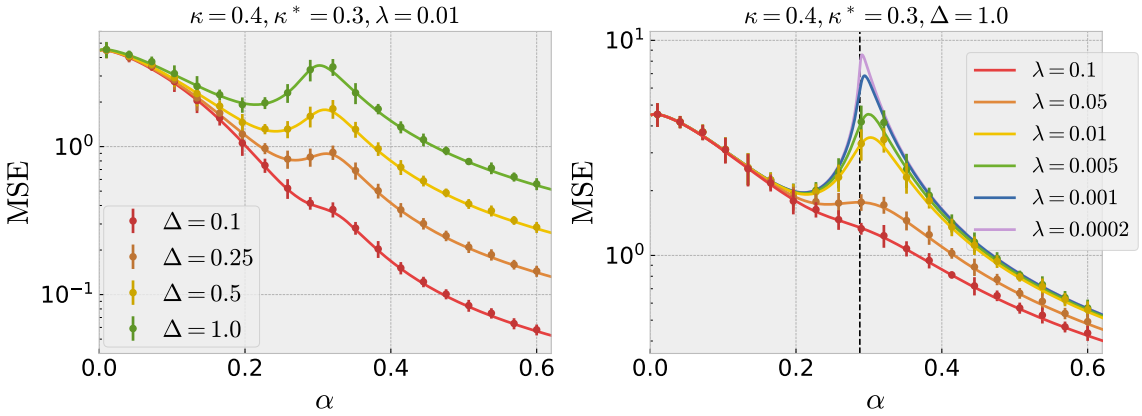

Figure 7: MSE as a function of α for different values of label noise Δ (left) and regularization strength λ (right). Dots: simulations of gradient descent, defined in equation (6), averaged over 10 realizations of the initialization, teacher and data. Full lines: numerical integration of the system of equations (37). Vertical dashed line: interpolation threshold (see Section 3.3.1).

knowledge, the first explicit dynamical characterization of double descent in a nonlinear high-dimensional model.

To conclude this part, let us compare our results with the recent and closely related work of Montanari and Urbani (2025), who report that learning and overfitting take place on two different timescales. In our case, both phenomena arise at the same timescale. A key distinction lies in the choice of activation function. We consider a centered quadratic activation with information exponent two, whereas their analysis focuses on activations with information exponent one. As a consequence, this difference leads the two models to exhibit different learning dynamics. Additional differences include the teacher structure (rank one in their case versus extensive in our setting) and the use of a sequential extensive-width limit, where the number of hidden units is taken to infinity after the dimension and sample size.

3.2.6 Population Limit

Let us now consider the population limit, i.e., the regime $\alpha \to \infty$, corresponding to a number of observations $n \gg d^2$. In this case, we can derive the associated limit of the system of equations in Claim 6.

Proposition 7 *Consider the case $\kappa \geq \min(\kappa^*, 1)$ and the variables q, ξ, solutions of the system of equations (37). Then, as $\alpha \to \infty$, we have $q \to \lambda$ and $\xi = \Theta(\alpha^{-1})$. The limit of the gradient flow is then given by:*

$$Z_\infty = \left(Z^* - \lambda I_d\right)^+, \tag{45}$$

and the MSE and the training loss write:

$$\text{MSE} = \int \min(x, \lambda)^2 \mathrm{d}\mu^*(x), \qquad \text{Loss}_{\text{train}} = \frac{1}{2}\text{MSE} + \frac{\Delta}{4}. \tag{46}$$

We prove this result in Section D.6.4. Interestingly, in the regime where the student has enough parameters to recover the teacher $\kappa \geq \min(\kappa^*, 1)$, these population equations do not depend on κ anymore. In addition, since μ^* is by assumption supported on $\mathbb{R}^+$ (due to the fact that Z^* is positive semidefinite), the MSE vanishes as $\lambda \to 0$: provided that the student has enough rank, it always recovers the teacher in this limit. This result is coherent with prior works on the population loss (Sarao Mannelli et al., 2020; Martin et al., 2024).

Interestingly, one recovers the same result when taking the limit $n \to \infty$ in the finite-dimensional expression of the loss (4), with our choice of quadratic cost, and Gaussian label noise. Indeed, we have:

$$\lim_{n\to\infty} \mathcal{L}_n(W) = \frac{1}{2d}\left\|WW^\top - Z^*\right\|_F^2 + \frac{\Delta}{4}. \tag{47}$$

Then, one can study the gradient flow associated with this loss, leading to a study of the Oja flow (see Section 4 for relevant results on this dynamics). Finally, we show that taking the high-dimensional limit in this setting recovers Proposition 7, leading to the equivalence between the $n, d \to \infty$ limits taken jointly before taking $\alpha = n/d^2 \to \infty$, and the sequential limit $n \to \infty$ then $d \to \infty$. More details can be found in Section D.6.

3.2.7 Langevin Dynamics

While the previous results focused on the gradient flow setting, it is natural to ask whether the same analysis can be performed for Langevin dynamics (in the case $\beta < \infty$). In this setting, we are interested in the stationary measure of the stochastic dynamics (23): due to the Brownian motion, the matrix $W(t)$ does not settle at long times but keeps fluctuating.

As a consequence, Assumption 5 is not suited for the study of Langevin dynamics, since it assumes that the Gaussian noise $\mathcal{G}(t)$ converges as $t \to \infty$. Instead, we rely on standard assumptions inspired by statistical physics and generalized Langevin equations: time-translational invariance and fluctuation–dissipation.

- *Time-translational invariance* (TTI) ensures that the correlation of the Gaussian process $\mathcal{G}$ and the memory kernel R (in equation 23) depend only on time differences. It reflects that the dynamics has reached a stationary regime in which statistical properties no longer drift with time.

- *Fluctuation–dissipation relation* links the covariance of the Gaussian process $\mathcal{G}$ to the memory kernel R in a physically consistent manner, enforcing a balance between random forcing and dissipation that enables relaxation toward equilibrium.

Together, these conditions ensure the existence of a stationary structure for the dynamics and allow one to show that it is driven toward equilibrium. Such assumptions have been commonly used, and confirmed, in spin-glass models (Sompolinsky and Zippelius, 1982; Altieri et al., 2020) and in high-dimensional learning problems (Chen and Shen, 2025; Fan et al., 2025). Our precise assumption can be found in Assumption 43.

Claim 8 *Consider the stochastic differential equation* (23)*, along with the self-consistent equations on the covariance of* $\mathcal{G}$ *in* (25) *and the memory kernel* R *in* (27)*. Under Assumption 43, the stationary measure of* (23) *is given by:*

$$\mathbb{P}_\beta(W) \propto \exp\left(-r\beta d\left\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\right\|_F^2 - 2\beta d\,\Omega(W)\right), \tag{48}$$

where $\mathcal{G} \sim \mathrm{GOE}(d)$ *and the variables* r, ξ *are self-consistently computed from* $\mathbb{P}_\beta$*:*

$$\xi = \frac{1}{2\alpha}\left(\frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\left\|\mathbb{E}_\beta\left[WW^\top\right] - Z^*\right\|_F^2 + \frac{\Delta}{2}\right), \tag{49}$$

$$r = \frac{\alpha}{\alpha + 2\beta V_\beta}, \tag{50}$$

$$V_\beta = \frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\mathbb{E}_\beta\left\|\mathbb{E}_\beta\left[WW^\top\right] - WW^\top\right\|_F^2, \tag{51}$$

where $\mathbb{E}_\beta$ *denotes the expectation with respect to* $\mathbb{P}_\beta$*.*

This stationary measure is computed in Section F.1. The main technical tool involved is the mapping of the dynamics (23) onto a coupling with a Gaussian auxiliary matrix process that allows to use standard results for the stationary measure of Langevin dynamics. The self-consistent equations for the variables ξ, r are a direct consequence of the self-consistency of the covariance of $\mathcal{G}(t)$ and $R(t,t')$ in Claim 4. These equations are derived in Section F.2.2.

It is interesting to note that this result exhibits strong similarities with the one in the gradient flow setting. First of all, the expression of the noise variance ξ in equation (49) is almost the same as in equation (34). The only difference is that at positive temperature, the Langevin predictor is stochastic and in this case, the MSE is computed as the one of the mean of $WW^\top$ under $\mathbb{P}_\beta$. In addition, the variable r is defined the same way as r_∞ in equation (35), and it precisely plays the same role as in the approximate dynamics (33).

Additionally, we derive in Section F.3 the stationary measure of the typical label, whose dynamics is given in equation (24). As remarked, the typical label remains Gaussian at all times. In the long-time limit, we obtain self-consistent expressions of its mean and covariance, summarized in Section F.3.2. Together with the result of Claim 8, this provides a set of self-consistent equations that allows to compute averaged quantities of the Langevin dynamics in the long-time limit.

The reasoning carried out in Section 3.2.1 allowed to obtain the set of low-dimensional equations of Claim 6, through the use of random matrix theory. Even though the set of equations of Claim 8 is still high-dimensional, we believe that it could be reduced to a system of scalar equations, as it was done by Maillard et al. (2024) in the Bayes-optimal setting. This would require to understand the distribution $\mathbb{P}_\beta$, and imply the use of results on HCIZ integrals (Harish-Chandra, 1957; Itzykson and Zuber, 1980; Guionnet and Zeitouni, 2002). Two cases can be directly analyzed already.

Zero-temperature limit. When considering ℓ_2-regularization $\Omega(W) = \lambda \mathrm{Tr}\left(WW^\top\right)$ and taking the $\beta \to \infty$ limit in Claim 8, we show in Section F.4 that we recover the same result as in Section 3.2.2. Interestingly, as it is known that the stationary measure of Langevin dynamics concentrates on the set of the global minimizers of its associated potential, this

implies that the gradient flow dynamics converges to a global minimizer of the regularized empirical loss. While we already claimed this was the case for large values of κ in Section 3.2.3, this general conclusion was still an open question.

This claim is valid under the assumptions of the section. Indeed, in both the gradient flow and Langevin settings, these assumptions allowed to interpret the high-dimensional dynamics as an effective gradient system with the potential:

$$U_{\text{eff}}(W) = \frac{r}{2d}\left\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\right\|_F^2 + \frac{\lambda}{d}\mathrm{Tr}\big(WW^\top\big). \tag{52}$$

As we show in Section 4.1, all local minimizers of this potential are global. Therefore, in this setting, it is no surprise to observe the match between gradient flow dynamics at long times and the zero-temperature limit of the stationary measure of Langevin dynamics.

Bayes-optimal learning. In addition, following an appropriate choice of the inverse temperature β in the Langevin dynamics, we show that the stationary measure of Claim 8 matches with the Bayes-optimal posterior distribution derived by Maillard et al. (2024) for the same problem. This establishes a direct connection between the dynamical formulation considered here and the Bayes-optimal analysis. In addition, it reveals the equivalence between the replica computation performed by Maillard et al. (2024) under the replica-symmetric ansatz, and the dynamical mean-field analysis presented here in the TTI regime.

3.3 Small Regularization Limit

In this section we study the small regularization limit ($\lambda \to 0^+$) of the system of equations given in Claim 6. Note that we do not expect the associated estimator to coincide with the one obtained when running gradient flow on the unregularized empirical loss. Indeed, in the following, we first consider the long-time ($t \to \infty$) limit before studying the small regularization limit.

The results derived in this section share several similarities with those of Erba et al. (2025b). As mentioned earlier, when $\kappa \geq 1$, the system of equations presented in Claim 6 coincides exactly with theirs. However, our equations (and the conclusions we draw from them) remain valid for all values of κ, including the underparameterized regime $\kappa < 1$. It is precisely in this regime that the system (37) depends explicitly on κ. This dependence allows us to study how this parameter, and thus the amount of overparameterization, affects the performance of the gradient flow estimator.

In the following, we will focus on the case $\kappa \geq \min(\kappa^*, 1)$, when the student possesses enough parameters to be able to recover the teacher. In particular, this setting will allow us to study perfect recovery in the small regularization limit (when $\Delta = 0$).

To capture the dependence on the teacher's effective rank κ^*, we recall Assumption 1 and decompose the teacher's asymptotic spectral measure μ^* as:

$$\mu^* = (1 - \min(\kappa^*, 1))\delta + \min(\kappa^*, 1)\nu^*, \tag{53}$$

where ν^* is a probability measure with support on $\mathbb{R}^+$ that we assume in addition to be bounded away from zero. Some of our proofs will also require ν^* to admit a smooth density, but we believe that our results can be easily extended to more general settings.

3.3.1 Interpolation Threshold

In a similar fashion as in the work of Erba et al. (2025b), we are able to derive an interpolation threshold $\alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$ such that:

- For $\alpha > \alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$, the performance of the gradient flow estimator in the limit $\lambda \to 0^+$ is found by taking $q = 0$ in the system (37) while λ/q remains of order one. In this case the gradient flow estimator behaves as if it was minimizing the unregularized loss. We will detail this correspondence in Section 3.4.1.
- For $\alpha < \alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$, the system of equations at $\lambda = 0^+$ is obtained by plugging $\lambda = 0$ into the system (37). In this case the effect of the regularization remains in the limit $\lambda \to 0^+$ and the performance differs from the one of the unregularized dynamics.

In Section 3.3.2, we show that this threshold coincides with the largest sample complexity for which the training labels can still be fitted exactly, in the presence of label noise. This interpretation is consistent with the standard notion of the interpolation threshold in noisy learning settings.

As a consequence of the previous characterization, the interpolation threshold is given by the smallest value of α for which the system (37) admits a solution in the limit where q vanishes proportionally to λ. For the following, we define the function:

$$I_\omega(\xi) = \int_{\max(0,\omega)} x h_\xi(x) \mathrm{d}\mu_\xi(x). \tag{54}$$

Recall that h_ξ is the Hilbert transform of the measure μ_ξ. In the end, we have the following characterization of the interpolation threshold:

Proposition 9 *For $\kappa \geq \min(\kappa^*, 1)$, consider ω, ξ to be the solution of the system:*

$$\begin{aligned} \min(\kappa, 1) &= \int_\omega \mathrm{d}\mu_\xi(x), \\ 2\xi I_\omega(\xi) &= \int_{\max(0,\omega)} x^2 \mathrm{d}\mu_\xi(x) - \frac{\Delta}{2} - Q_*. \end{aligned} \tag{55}$$

The interpolation threshold is then given by $\alpha_{\text{inter}}(\kappa, \kappa^*, \Delta) = I_\omega(\xi)$.

We emphasize that the dependence of these equations on κ^* enters only through the teacher's asymptotic spectral distribution μ^*, which appears in μ_ξ, the free additive convolution of μ^* with a semicircular distribution of variance ξ.

When $\kappa \geq 1$, our system of equations coincides with the one of Erba et al. (2025b) (Result 1). In their work, the interpolation threshold is defined as the smallest value of α for which the empirical loss admits a unique global minimizer. We conjecture that a similar interpretation holds for general values of κ. Indeed, the region $\alpha > \alpha_{\text{inter}}$ is characterized by the fact that the ratio λ/q remains positive in the limit of vanishing regularization. As discussed in Section 3.2.1, this quantity coincides with the long-time integrated response r_∞, defined in equation (35). Within the dynamical approximation introduced in equation (33), this scalar variable is directly related to the curvature of the loss landscape near the point of convergence. This leads to two possibilities:

- When $r_\infty = 0$, the landscape becomes flat in a neighborhood of the limit point, suggesting that the set of global minimizers of the empirical loss (in terms of W) forms a manifold of positive dimension.
- When $r_\infty > 0$, the convergence shares the same properties as in the case $\lambda > 0$, and we expect the global minimizer reached by gradient flow to be isolated (up to the representation $W \mapsto WW^\top$).

This dynamical interpretation only provides local information about the landscape close to the point reached by gradient flow. When $\kappa \geq 1$, the optimization problem is convex and so is the set of global minimizers. In this regime, the interpolation threshold α_{inter} describes the threshold at which the loss admits a single global minimizer. This observation was already made by Erba et al. (2025b). However, when $\kappa < 1$, convexity breaks, and the set of global minimizers may be more complicated, potentially splitting into multiple components.

Interestingly, as illustrated in Figure 7, the interpolation threshold appears to coincide with the location of the peak of the MSE in the double descent regime. This correspondence is not specific to our setting and has previously been observed in linear models (Belkin et al., 2019; Hastie et al., 2022) as well as in random feature models (d'Ascoli et al., 2020a; Mei and Montanari, 2022). More generally, this behavior is consistent with the idea that, in the presence of label noise ($\Delta > 0$), the interpolation threshold marks the largest sample complexity for which the model can exactly fit the training data, leading to vanishing training error but potentially poor generalization. In Section 3.4, we further investigate this regime by analyzing the unregularized dynamics and providing additional numerical evidence.

The following corollary gives a more explicit characterization of the interpolation threshold in the noiseless case $\Delta = 0$ and for $\kappa \geq \min(\kappa^*, 1)$, that is, when the student can recover the teacher at a finite value of α.

Corollary 10 *Consider the case where $\kappa \geq \min(\kappa^*, 1)$ and $\Delta = 0$, and let σ denote the semicircular distribution (defined in Section B.1.1). Then the solution ω, ξ of the system of equations of Proposition 9 are reached at $\xi = 0, \omega = 0$. The value of the interpolation threshold is then, for $\kappa^* \leq 1$:*

$$\alpha_{\text{inter}}(\kappa, \kappa^*) = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(0,\tilde{\omega})} x^2 \mathrm{d}\sigma(x), \tag{56}$$

where $\tilde{\omega}$ solves the equation:

$$\frac{\min(\kappa, 1) - \kappa^*}{1 - \kappa^*} = \int_{\tilde{\omega}} \mathrm{d}\sigma(x). \tag{57}$$

In addition, if $\kappa, \kappa^* \geq 1$, then $\alpha_{\text{inter}}(\kappa, \kappa^*) = 1/2$.

The quantity $\tilde{\omega}$ is a rescaled version of the threshold ω. Recall that ω selects a fraction κ of the mass of the measure μ_ξ. When $\kappa^* \leq 1$ and $\kappa > \kappa^*$, this selection involves a part of the semicircular component of μ_ξ, whose support has width of order $\sqrt{\xi}$. As a result, ω must

scale with ξ and vanish as $\xi \to 0$. The rescaled variable $\tilde{\omega}$ captures how the semicircular density should be cut in order to select a mass κ. The corresponding expression for the interpolation threshold is derived in Section H.1.

As a first remark, this threshold does not depend on the teacher's distribution μ^*, unlike for positive Δ (see Proposition 9). The semicircular distribution σ appears as a universal object in these equations: it is directly linked to the Gaussian noise added to the teacher to form the matrix $Z^* + \sqrt{\xi}\mathcal{G}$ studied earlier. Indeed, it is well known (Wigner, 1955) that the semicircular distribution corresponds to the high-dimensional limit of the spectral density of a GOE matrix.

3.3.2 Set of Equations in the Small Regularization Limit

Considering the previous observations, we shall now formulate the set of equations that describes the performance and spectral properties of the gradient flow predictor in the small regularization limit.

Proposition 11 *Recall the definition of the function I_ω in equation (54). Assume that given the values $\alpha, \kappa, \kappa^*, \Delta$ and $\lambda > 0$, the system of equations (37) has a unique solution (ξ, q). Then, as $\lambda \to 0^+$, we have:*

1. *If $\alpha \leq \alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$, the gradient flow predictor remains:*

$$Z_\infty = \left(Z^* + \sqrt{\xi}\mathcal{G} - qI_d\right)_{(m)}^+, \tag{58}$$

with values of MSE and training loss:

$$\text{MSE} = 2\alpha\xi - \frac{\Delta}{2}, \qquad \text{Loss}_{\text{train}} = 0, \tag{59}$$

where ξ, q are solutions of the system of equations (37) with $\lambda = 0$.

2. *If $\alpha \geq \alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$, the system of equations reduces to the variables ω, ξ, solution of:*

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x), \tag{60a}$$

$$2\alpha\xi - \frac{\Delta}{2} = Q_* - \int_{\max(0,\omega)} x^2 \mathrm{d}\mu_\xi(x) + 4\xi I_\omega(\xi). \tag{60b}$$

In this case the gradient flow predictor is given by:

$$Z_\infty = \left(Z^* + \sqrt{\xi}\mathcal{G}\right)_{(m)}^+, \tag{61}$$

and the MSE and the loss write:

$$\text{MSE} = 2\alpha\xi - \frac{\Delta}{2}, \qquad \text{Loss}_{\text{train}} = \alpha\xi\left(1 - \frac{I_\omega(\xi)}{\alpha}\right)^2. \tag{62}$$

Even in the presence of label noise, for $\alpha < \alpha_{\text{inter}}$ the empirical loss vanishes. In the small regularization limit, this implies that gradient flow converges to a predictor that exactly fits all the training labels. When $\alpha \geq \alpha_{\text{inter}}$, we consider two cases:

- If $\Delta = 0$, interpolation occurs after perfect recovery. In this case, the loss remains zero, and so does the MSE.

- In the presence of label noise, beyond the interpolation threshold, the student can no longer fit all the observed labels, and the training loss becomes positive. In this regime, we have the characterization of the interpolation threshold:

$$\alpha \leq \alpha_{\text{inter}} \iff \text{Loss}_{\text{train}} = 0, \tag{63}$$

  meaning that α_{inter} is the largest sample complexity for which exact fitting of the training data is still possible. In noisy settings, this criterion is often taken as the definition of the interpolation threshold.

In addition, one can remark that the system of equations obtained for $\alpha > \alpha_{\text{inter}}$ is precisely the same as we would get when performing the same dynamical assumptions as in Section 3.2, but for the unregularized dynamics. This observation motivates the conjecture formulated in Section 3.4: in this regime, the limit $\lambda \to 0^+$ coincides with the unregularized dynamics.

Figure 8 illustrates Proposition 11 in the noiseless case. It features both gradient descent simulations at $\lambda > 0$, and the solution of the system of equations of Proposition 11 at $\lambda = 0^+$. In this limit, the training loss is always zero and the MSE vanishes at a finite value of α. In the next part, we derive the perfect recovery threshold, corresponding to this critical value.

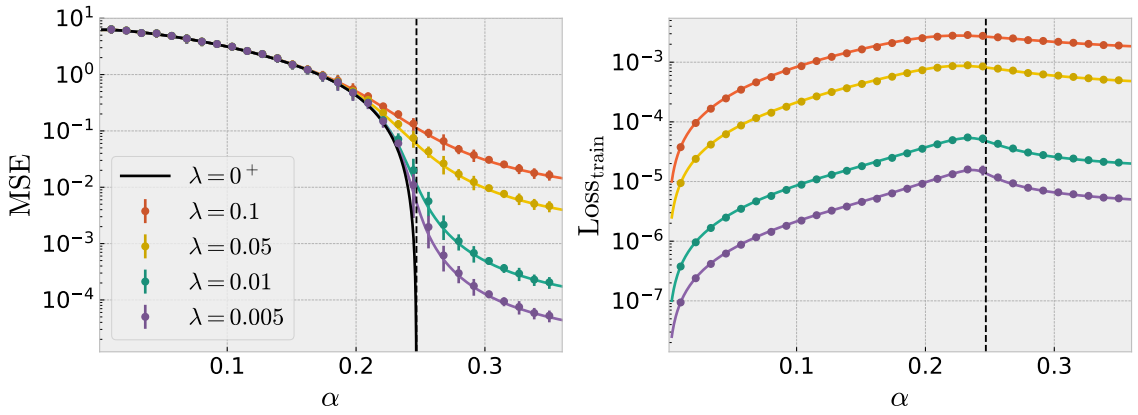

Figure 8: MSE (left) and training loss (right) as a function of α, for $\kappa = 0.3, \kappa^* = 0.2, \Delta = 0$ and several values of regularization strength λ. Dots: simulations of gradient descent, defined in equation (6), for $\lambda > 0$ and averaged over 10 realizations of the initialization, teacher and data. Full lines: numerical integration of the system of equations (37). The black line corresponds to simulations of the system of equations in Proposition 11. Vertical dashed line: perfect recovery threshold, defined in Proposition 12.

3.3.3 Perfect Recovery Threshold

From these results, we are now able to access the perfect recovery (PR) threshold associated with the system of equations in the small regularization limit, that is the value of α (that we denote $\alpha_{\rm PR}^+$) for which the MSE is zero as soon as $\alpha \geq \alpha_{\rm PR}^+$. Note that for $\lambda > 0$ the MSE always remains positive, so that the PR transition can only happen in the small regularization limit.

Proposition 12 *Consider the setting $\Delta = 0$ and $\kappa \geq \min(\kappa^*, 1)$. Let σ denote the semicircular distribution (defined in Section B.1.1). Then, if $\kappa^* < 1$, the perfect recovery threshold is given by:*

$$\alpha_{\rm PR}^+(\kappa, \kappa^*) = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(h,\tilde{\omega})} x(x-h) \mathrm{d}\sigma(x), \tag{64}$$

where $h, \tilde{\omega}$ solve the equations:

$$\begin{aligned} \frac{\min(\kappa, 1) - \kappa^*}{1 - \kappa^*} &= \int_{\tilde{\omega}} \mathrm{d}\sigma(x), \\ \frac{\kappa^*}{1-\kappa^*} &= \frac{1}{h} \int_{\max(h,\tilde{\omega})} (x-h) \mathrm{d}\sigma(x). \end{aligned} \tag{65}$$

In addition, if $\kappa^* \geq 1$, $\alpha_{\rm PR}^+(\kappa, \kappa^*) = \dfrac{1}{2}$ for any $\kappa \geq 1$.

For $\kappa \geq 1$, this quantity is exactly the same as the one identified by Erba et al. (2025b). This threshold can be derived by sending the MSE to zero, which amounts to taking the $\xi \to 0$ limit. Then, the result is obtained by choosing the scaling $q \propto \sqrt{\xi}$. As previously, the variable $\tilde{\omega}$ plays the role of the rank constraint. We refer to Section H.2 for the derivation of this threshold.

3.3.4 Comparison of the Thresholds

We shall now compare the interpolation and the PR thresholds derived in Corollary 10 and Proposition 12.

A first observation follows directly from equations (56) and (64): for almost all values of κ and κ^*, the perfect recovery threshold is strictly smaller than the interpolation threshold. This implies that, in the small regularization limit, perfect recovery typically occurs in a regime where the empirical loss still admits multiple global minimizers. In this sense, the solution selected in the small regularization limit has better generalization properties than a generic interpolator.

The two thresholds only coincide in specific situations: either when the student and teacher ranks match, $\kappa = \kappa^*$, or when both κ and κ^* are larger than one. In these cases, we have the expressions:

$$\alpha_{\rm inter}(\kappa, \kappa^*) = \alpha_{\rm PR}^+(\kappa, \kappa^*) = \begin{cases} \kappa^* - \dfrac{\kappa^{*2}}{2}, & \text{if } \kappa = \kappa^* \leq 1, \\ \dfrac{1}{2}, & \text{if } \kappa, \kappa^* \geq 1. \end{cases} \tag{66}$$

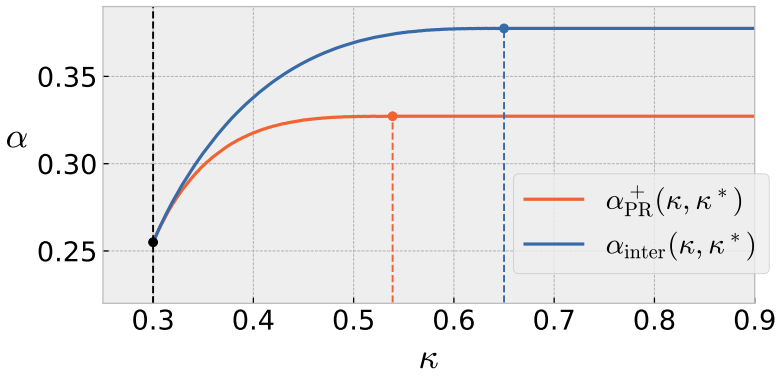

Figure 9: Interpolation and PR thresholds as a function of κ for $\kappa^* = 0.3$, computed from Corollary 10 and Proposition 12. The black dot indicates the common value of the thresholds $\kappa^* - \kappa^{*2}/2$ at $\kappa = \kappa^*$. Vertical colored lines indicate the value of κ at which each threshold becomes constant.

In the matched-rank case, $\kappa = \kappa^*$, this expression matches with the one previously obtained by Maillard et al. (2024). The regime $\kappa, \kappa^* \geq 1$ also recovers known results for this problem, consistent with earlier works such as Donoho et al. (2013); Gamarnik et al. (2019).

Moreover, it is clear that these thresholds only depend on the effective rank κ close to κ^*. More precisely, one can show that for each threshold there exists a critical value of κ beyond which the threshold no longer depends on κ. For $\kappa^* < 1$, a direct calculation gives the expressions:

$$\alpha_{\mathrm{inter}}(\kappa, \kappa^*) = \frac{1}{2}\left(\frac{1}{2} + \kappa^* - \frac{\kappa^{*2}}{2}\right) \qquad \Longleftrightarrow \kappa \geq \frac{1+\kappa^*}{2}, \tag{67}$$

$$\alpha^+_{\mathrm{PR}}(\kappa, \kappa^*) = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2}\int_h x(x-h)\mathrm{d}\sigma(x) \quad \Longleftrightarrow \kappa \geq \kappa^* + (1-\kappa^*)\int_h \mathrm{d}\sigma(x), \tag{68}$$

where $h > 0$ is solution of:

$$\frac{\kappa^*}{1-\kappa^*} = \frac{1}{h}\int_h (x-h)\mathrm{d}\sigma(x). \tag{69}$$

These two values of the interpolation and perfect recovery thresholds match with the ones obtained by Erba et al. (2025b), for $\kappa \geq 1$. Moreover, the critical value of κ at which α^+_{PR} becomes constant in equation (68) exactly matches the quantity κ_{min} defined in (41), in the small regularization limit. Recall that κ_{min} corresponds to the (normalized) rank of the global minimizer of the regularized loss over the set of PSD matrices. As discussed in Section 3.2.3, this quantity separates two regimes at positive regularization, depending on whether the student has enough parameters to reach this global minimizer. This distinction persists in the small regularization limit. For small values of κ, the student converges to a suboptimal predictor in terms of the empirical loss (when viewed as a function of $Z = WW^\top$). Nevertheless, the low-rank structure of the solution allows the student to recover the teacher using fewer samples.

This analysis reveals an interesting conclusion about the role of overparameterization, quantified by the network normalized width κ. As illustrated in Figure 9, increasing κ requires more observations to reach perfect recovery, but only up to a point. For larger values of κ (while still $\kappa < 1$), the perfect recovery threshold becomes independent of the number of parameters. This behavior is driven by the low-rank structure of the teacher, that influences $\kappa_{\min}$, the rank of the global minimizer of the empirical loss over all PSD matrices. Once κ exceeds this value, further overparameterization does not increase the sample complexity required for perfect recovery.

To conclude this part, we insist on the generality of the previous results: the expressions of $\alpha_{\rm PR}^+$ and $\alpha_{\rm inter}$ only depend on the variables κ, κ^*, but not on the distribution of the teacher or the student at initialization. Although the derivation of equations (56) and (64) requires some mild assumptions (that we detail in Section H), these remain very general.

Small-width teacher. One can wonder about the behavior of the thresholds $\alpha_{\rm inter}$ and $\alpha_{\rm PR}^+$ in the limit $\kappa^* \to 0$, i.e., when the teacher matrix possesses a sub-extensive rank $m^* \ll d$. In this case, as a consequence of Corollary 10 and Proposition 12, we have the equations:

$$\begin{aligned} \alpha_{\rm inter}(\kappa, 0^+) &= \frac{1}{2}\int_{\max(0,\tilde{\omega})} x^2 \mathrm{d}\sigma(x), \qquad \min(\kappa, 1) = \int_{\tilde{\omega}} \mathrm{d}\sigma(x), \\ \alpha_{\rm PR}^+(\kappa, 0^+) &= 0. \end{aligned} \tag{70}$$

Therefore, even an extensive-width student (with $\kappa > 0$) is able to recover this small teacher with a number of observations $n \ll d^2$, but interpolation only arises at $n = \Theta(d^2)$. In addition, we have the values:

$$\alpha_{\rm inter}(\kappa, 0^+) \xrightarrow[\kappa \to 0^+]{} 0, \qquad \alpha_{\rm inter}(\kappa, 0^+) = \frac{1}{4} \iff \kappa \geq \frac{1}{2}. \tag{71}$$

This value of $1/4$ is already present in present in the work of Erba et al. (2025b). In comparison, Sarao Mannelli et al. (2020) derived an interpolation threshold $n = 2d$ in the case of a rank-one teacher. As we do not recover this result when taking the $\kappa^* \to 0$ limit, this suggests the presence of different regimes depending on the scaling between m^* and the dimension.

3.3.5 Minimal Regularization Estimator

The previous observations reveal that in the small regularization limit, the gradient flow dynamics is selecting a specific global minimizer of the empirical loss that exhibits favorable generalization properties. In this part we relate this observation to the notion of minimal regularization interpolator.

We start by mentioning a standard proposition that gives a general characterization of the behavior of the global minimizer of a regularized loss in the small regularization limit.

Proposition 13 *Consider a loss of the form $\mathcal{L}_\lambda = \mathcal{L} + \lambda\Omega$, where $\mathcal{L}, \Omega$ are both continuous and take positive values. Assume moreover that the regularization Ω is coercive:*

$$\big\|\Omega(W)\big\| \xrightarrow[\|W\| \to \infty]{} \infty.$$

Define $S^* = \operatorname{argmin}(\mathcal{L})$ and assume S^* is non-empty. Then, given a family $(W_\lambda)_{\lambda>0}$ such that W_λ is a global minimizer of $\mathcal{L}_\lambda$ for all $\lambda > 0$, every cluster point of (W_λ) as $\lambda \to 0$ belongs to S^* and minimizes Ω over S^*. In addition, if the minimizer of Ω over S^* is unique, then W_λ converges toward it.

Linking this proposition back to our setting, a natural candidate for the predictor selected in the small regularization limit is the minimal ℓ_2-norm interpolator:

$$\operatorname*{argmin}_{W \in S^*} \|W\|_F^2, \qquad S^* = \operatorname*{argmin}_{W \in \mathbb{R}^{d\times m}} \left[\frac{1}{n} \sum_{k=1}^{n} \Big(\operatorname{Tr}(X_k W W^\top) - z_k \Big)^2 \right], \tag{72}$$

corresponding to the minimal ℓ_2-norm interpolator. However, convergence toward this predictor is only guaranteed when gradient flow reaches a global minimizer of the regularized loss. As emphasized earlier, our results allow us to establish this property only in the regime $\kappa \geq \kappa_{\min}$, where gradient flow converges to the global minimizer over all PSD matrices. In this case, the link with the minimal-norm interpolator has already been made precise by Erba et al. (2025b). Moreover, rewriting the optimization problem in terms of $Z = WW^\top$ shows that minimizing $\|W\|_F^2$ is equivalent to minimizing $\operatorname{Tr}(Z)$, yielding the minimal nuclear-norm interpolator.

In the underparameterized regime $\kappa < \kappa_{\min}$, we cannot directly guarantee that the solution reached by gradient flow at positive λ corresponds to a global minimizer of the regularized loss. Nevertheless, as discussed in Section 3.2.7, the analysis of low-temperature Langevin dynamics suggests that this may still hold, provided that our asymptotic simplifications on the dynamics (namely time-translational invariance and fluctuation–dissipation) are valid. If confirmed, this would imply that, for all values of κ, Proposition 11 characterizes the performance of the minimal ℓ_2-norm interpolator.

3.4 Unregularized Dynamics

In this section, we investigate the gradient flow dynamics (5) in the absence of regularization. First of all, note that the dynamical result of Claim 4 still holds in the absence of regularization, but the dynamics exhibit quite different behaviors than in the regularized case. Indeed, numerical simulations suggest that the dynamics at long times strongly depends on the initial condition. For instance, we show in Figure 10 the MSE reached by the unregularized dynamics initialized with Gaussian weights with different variances γ and remark that the curves we obtain as a function of α strongly depend on this parameter. In contrast, it is known that the presence of regularization tends to make the landscape more convex (see for instance Hoerl and Kennard, 1970; Du and Lee, 2018; Kobayashi et al., 2024). It also reduces the dependence on initialization by shrinking the weight components orthogonal to the data, which sometimes leads to the independence of the long-time dynamics with respect to the initial condition.

This observation reveals that the assumption used in Section 3.2 is not relevant to describe the unregularized dynamics, since it strongly relies on the hypothesis that the early stages of the dynamics are forgotten at long times.

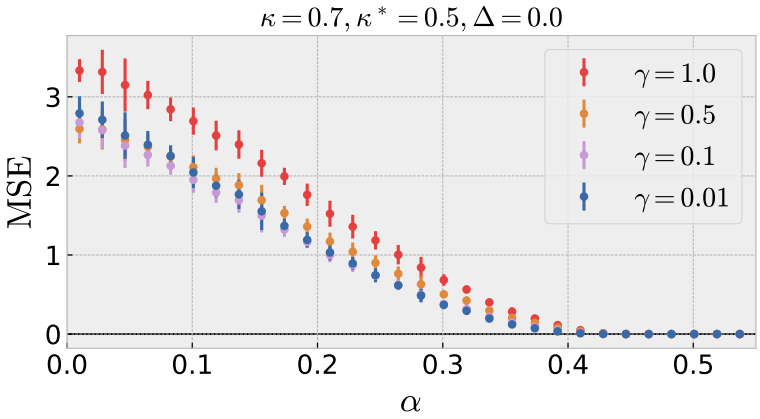

Figure 10: MSE reached by gradient descent, defined in equation (6), as a function of α and for different values of initialization variance γ. Simulations are averaged over 10 realizations of the initialization, teacher and data. $\kappa = 0.7, \kappa^* = 0.5$ and zero label noise.

3.4.1 Beyond the Interpolation Threshold

Despite these observations, recall that in Section 3.3.1 we introduced the interpolation threshold α_{inter}, and remarked that for $\alpha > \alpha_{\text{inter}}$, the statistics of the gradient flow estimator in the small regularization limit are the same as if there were no regularization. In this regime, we therefore conjecture that the equations derived in Proposition 11 also apply to the unregularized setting.

Conjecture 14 *Consider the gradient flow dynamics* (5) *under Assumption 3, in the absence of regularization. Then, for* $\alpha \geq \alpha_{\text{inter}}(\kappa, \kappa^*, \Delta)$*, the statistics of the gradient flow estimator are given by* ***2.*** *in Proposition 11.*

In Figure 11, we present a numerical illustration of this conjecture: the dependence on the initialization strength γ appears to be present only for $\alpha < \alpha_{\text{inter}}$. Beyond the interpolation threshold, this dependence vanishes and the performance of the gradient flow estimator matches the one in the small regularization limit. In this figure, we plot the MSE, the empirical loss and the in-sample error, corresponding to the error on the true labels:

$$\text{Err}_{\text{in}} = \frac{1}{4n} \sum_{k=1}^{n} \Big(\text{Tr}\big(X_k W W^\top\big) - \text{Tr}\big(X_k Z^*\big) \Big)^2, \tag{73}$$

where W is the gradient flow estimator reached at convergence. Similarly to the expressions of the MSE and training loss derived in Proposition 11, we derive an expression for the in-sample error in Section H.1.

Figure 11 reveals that for $\alpha < \alpha_{\text{inter}}$, the student is able to perfectly fit all the noisy labels, leading to a zero loss, and an in-sample error equal to $\Delta/4$: in this case there are too few observations for the student to capture the teacher structure behind the label noise. On

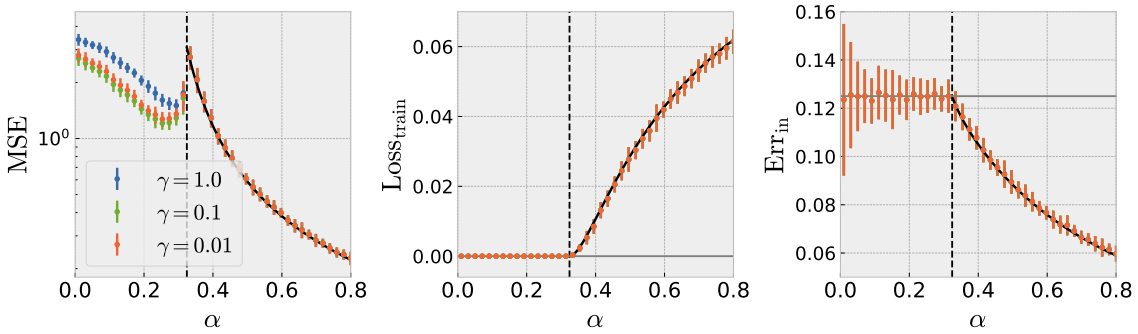

Figure 11: Comparison between simulations of gradient descent, defined in equation (6), and numerical integration of the system of equations (60) for $\alpha > \alpha_{\text{inter}}$ (vertical dashed line), for $\kappa = 0.7, \kappa^* = 0.5$ and $\Delta = 0.5$. MSE (left), empirical loss (middle) and error on the true labels, also known as in-sample error (right), as a function of α for three values of initialization variance γ, and with no regularization. The horizontal gray line on the right panel corresponds to the value $\Delta/4$. Gradient descent simulations are averaged over 20 realizations of the initialization, teacher, and data.

the contrary, for $\alpha > \alpha_{\text{inter}}$, the noisy labels cannot all be fitted, the loss becomes positive (and increases with α), but the representation learned by the student leads to a smaller error on the labels as well as a smaller MSE. In this region, despite the label noise, the high number of observations leads to an improved generalization.

3.4.2 Perfect Recovery Threshold

We shall now investigate the perfect recovery threshold achieved by the gradient flow estimator in the unregularized case. We therefore consider the case $\kappa \geq \min(\kappa^*, 1)$ and $\Delta = 0$, allowing for perfect recovery. As shown previously in Section 3.3.1, the MSE is already zero in the region $\alpha > \alpha_{\text{inter}}$. This leads to the question: does the PR threshold match the interpolation threshold, i.e., does gradient flow require the loss to have a single minimizer (at least locally near the point of convergence) in order to recover the teacher, or does the minimizer that is chosen have some particular properties that trigger perfect recovery before interpolation?

As a result of a large number of simulations, we formulate the following conjecture:

Conjecture 15 *Let σ denote the semicircular distribution (defined in Section B.1.1). In the setting where $\Delta = 0$ and $\kappa \geq \min(\kappa^*, 1)$, the value of the perfect recovery threshold is given by:*

$$\alpha_{\text{PR}}(\kappa, \kappa^*) = \min\left(\kappa - \frac{\kappa^2}{2}, \alpha_{\text{inter}}(\kappa, \kappa^*)\right), \tag{74}$$

where α_{inter} is given in Corollary 10 for $\kappa^* \leq 1$:

$$\alpha_{\text{inter}}(\kappa, \kappa^*) = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(0,\tilde{\omega})} x^2 \mathrm{d}\sigma(x), \tag{75}$$

with $\tilde{\omega}$ being solution of:

$$\frac{\min(\kappa, 1) - \kappa^*}{1 - \kappa^*} = \int_{\tilde{\omega}} \mathrm{d}\sigma(x). \tag{76}$$

This conjecture gives the expression of the perfect recovery threshold as the minimum of the interpolation threshold α_{inter} and a quantity $\alpha_{\text{dof}}(\kappa) = \kappa - \kappa^2/2$ that corresponds to the (normalized) number of degrees of freedom of the set of PSD matrices with rank κd:

$$\kappa - \frac{\kappa^2}{2} = \lim_{d\to\infty} \frac{1}{d^2} \dim\big\{ Z \in \mathcal{S}_d^+(\mathbb{R}),\, \mathrm{rank}(Z) = \lfloor \kappa d \rfloor \big\}. \tag{77}$$

Therefore, for $n \geq \alpha_{\text{dof}}(\kappa)\, d^2$, the number of observations exceeds the number of free parameters of the student matrix. Interestingly, this value does not coincide with the interpolation threshold derived in Section 3.3.1. Indeed, the quantity $\alpha_{\text{dof}}(\kappa)$ is a geometric quantity determined by the symmetries of the predictor, whereas the interpolation threshold is problem-specific and accounts for the structure of the teacher (note that α_{inter} depends on κ^* but not on α_{dof}).

More precisely, for $\kappa^* \leq 1$, an analysis of the function $\alpha_{\text{PR}}(\kappa, \kappa^*)$ in Conjecture 15 reveals the existence of a critical value $\kappa_c \in [\kappa^*, 1]$ such that

$$\alpha_{\text{PR}}(\kappa, \kappa^*) = \begin{cases} \kappa - \dfrac{\kappa^2}{2}, & \text{if } \kappa \leq \kappa_c, \\ \alpha_{\text{inter}}(\kappa, \kappa^*), & \text{if } \kappa \geq \kappa_c. \end{cases} \tag{78}$$

Therefore, for $\kappa < \kappa_c$, the teacher is perfectly recovered, despite the presence of multiple global minimizers in its vicinity. This conclusion is similar to the case of the small regularization limit Section 3.3: this phenomenon is an instance of the implicit bias induced by gradient flow dynamics.

Figure 12 compares the three thresholds derived in this section and in Section 3.3. Interestingly, the figure features a region in which the PR threshold in the $\lambda \to 0^+$ limit is larger than its unregularized counterpart. While this observation is plausible and does not contradict our previous results, the behavior differs from that observed at larger values of κ. In particular, this suggests that the implicit bias of the unregularized dynamics and the explicit bias induced by vanishing regularization select different solutions. As a result, each mechanism may be advantageous in different regions of the parameter space.

In the unregularized case, the perfect recovery threshold exhibits the same qualitative behavior as in the small regularization limit. At fixed κ^*, it increases with κ up to a certain value and then becomes constant. This occurs before $\kappa = 1$ and indicates that, in this regime, increasing the number of parameters does not require more data to achieve perfect recovery. From the expression (74), one can show that α_{PR} becomes constant at $\kappa = (1 + \kappa^*)/2$, provided that $\kappa^* \leq 1$. This value coincides with the interpolation threshold given in (67).

Finally, we emphasize that Conjecture 15 is confirmed by a large number of numerical simulations. More details on these simulations, the results and the measure of the PR threshold can be found in Section J.2.2.

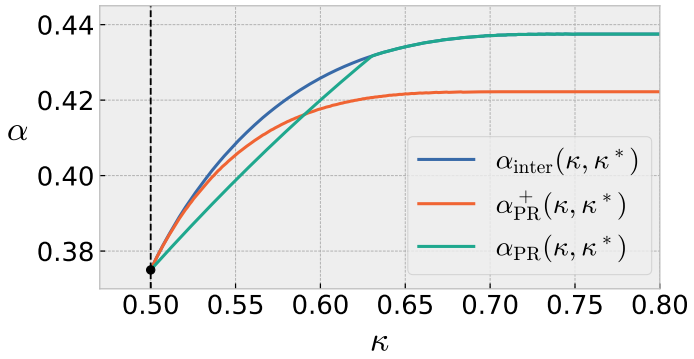

Figure 12: Conjectured PR threshold α_{PR} (Conjecture 15), along with the interpolation threshold α_{inter} (Corollary 10) and the PR threshold α_{PR}^{+} (Proposition 12) in the small regularization limit as a function of κ for $\kappa^* = 0.5$ (vertical dashed line).

3.5 Gaussian Equivalence and Quadratic Neural Networks

As emphasized earlier, the Gaussian structure of the data is a key assumption in our results. It enables to rewrite the dynamical partition function associated with the gradient flow dynamics (5) by averaging with respect to this distribution and leads to the dynamics given in Claim 2.

However, we believe that our conclusions should extend beyond the Gaussian setting to a broader class of distributions. The idea that key phenomena or asymptotic behaviors in complex systems remain unchanged regardless of the specific underlying distribution is often referred to as Gaussian universality. This principle has been supported by a long line of work establishing such universality results through asymptotic analyses of high-dimensional inference problems (Hu and Lu, 2022; Montanari and Saeed, 2022; Dandi et al., 2023; Gerace et al., 2024; Bandeira and Maillard, 2025; Xu et al., 2025; Wen et al., 2025). Unlike these works that derive universality results for static problems (either empirical risk minimization or Bayes-optimal learning), our conjecture bears on the behavior of gradient flow trajectories. Such dynamical universality results, although less studied, have been investigated in a few prior works (Celentano et al., 2021; Goldt et al., 2022; Han, 2025).

In the following, we study the case where the sensing matrices are (centered) rank-one measurements:

$$X_k = \frac{x_k x_k^\top - I_d}{\sqrt{d}}, \qquad x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d). \tag{79}$$

As underlined in Section 2, this setting corresponds to a shallow neural network with quadratic activation function, and fixed output weights. Our conjecture bears on the equivalence between this model and the Gaussian matrix sensing setting for which we derived our results.

Conjecture 16 *The conclusion of Claim 2 still holds if the sensing matrices $X_1, \dots, X_n$ are i.i.d. distributed as in* (79).

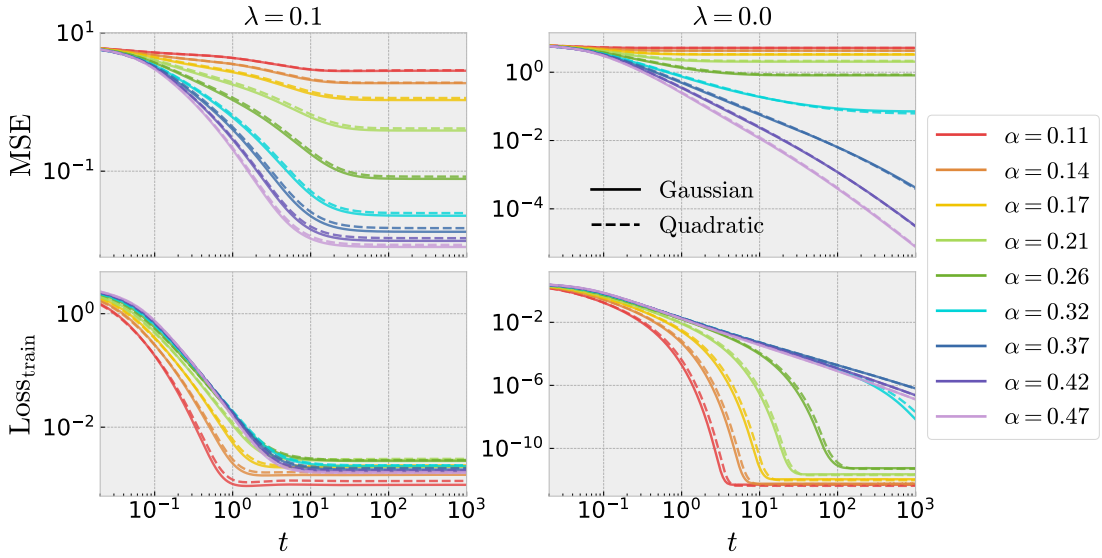

Figure 13: MSE and loss achieved by gradient descent, defined in equation (6), as a function of time for $d = 150, \kappa = 0.5, \kappa^* = 0.2$, $\Delta = 0$, $\lambda = 0.1$ (left) and $\lambda = 0$ (right), for several values of α. Full lines: Gaussian data. Dashed lines: data generated as in equation (79). Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data.

This conjecture is inspired by some recent works that derived universality results for quadratic neural networks, in the Bayes-optimal setting (Maillard et al., 2024) and for empirical risk minimization (Erba et al., 2025b). Although we formulate it as a conjecture, we build in Section C.6 an argument that would lead to a proof, but requires a more rigorous treatment.

In Figure 13, we numerically justify this claim by comparing time-dependent averaged quantities (MSE, loss) along both gradient descent dynamics (for the Gaussian and quadratic cases). Despite some slight discrepancies that should originate from the finite dimension (here $d = 150$), the trajectories look very similar and provide a first confirmation of this equivalence result.

To conclude this part, one question that would require a deeper clarification is to which extent this Gaussian universality result holds. Indeed, as it is often the case in such results, we believe that Conjecture 16 can be extended to a larger class of distributions. Therefore, it would be of interest to understand what would be the necessary requirements on this distribution to guarantee universality in our setting.

4. Additional Results on the Oja Flow

In this section, we consider the Oja flow dynamics, corresponding to the approximate dynamics we use in Section 3.2 to derive our long-time equations for the gradient flow trajec-

tory. In the following, we derive several results that are then used to prove the main claims of Section 3.2.

In the following, we let $A \in \mathcal{S}_d(\mathbb{R})$ be a symmetric matrix, and consider the nonlinear matrix differential equation:

$$\dot{W}(t) = \big(A - W(t)W(t)^\top\big)W(t), \tag{80}$$

where $W(t) \in \mathbb{R}^{d\times m}$. Moreover, note that $W(t)$ is solution of the gradient flow associated with the loss:

$$U(W) = \frac{1}{4}\big\|WW^\top - A\big\|_F^2. \tag{81}$$

Before presenting some results, we mention several key references on this topic. The Oja flow was introduced in the seminal paper by Oja (1982), and was designed as a continuous-time model for principal component analysis. Subsequent works studied its convergence properties (Yan et al., 1994; Tsuzuki and Ohki, 2025). This flow is naturally connected to high-dimensional inference problems, such as low-rank matrix factorization and matrix sensing. Such formulations are central to the Burer–Monteiro approach to semidefinite programming (Burer and Monteiro, 2003; Boumal et al., 2016) and can also be viewed as part of the broader framework of optimization on matrix manifolds (Absil et al., 2008). Finally, related ideas appear in matrix denoising: for instance, Bodin and Macris (2023) use the structure of the Oja flow to derive a high-dimensional limit for extensive-rank positive semidefinite denoising dynamics.

In the following, we study this dynamics and present new results, that we directly use to study the high-dimensional dynamics for our matrix sensing problem. More precisely:

- In Section 4.1, we start by mentioning several existing results on the Oja flow.
- In Section 4.2 and Section 4.3, we then derive tight convergence rates in both finite and infinite dimension. We compare those with existing literature on the subject.
- In Section 4.4, we finally derive the linear response of the solution with respect to a time-dependent perturbation.

4.1 Known Properties of the Oja Flow

We start with some known results on the Oja flow dynamics. The first one we present is a closed-form solution, expressed in terms of the matrix $Z(t) = W(t)W(t)^\top$.

Proposition 17 *Let $\big(W(t)\big)_{t\geq 0}$ be the solution of (80) with initial condition W_0. Then, at all times:*

$$W(t)W(t)^\top = e^{tA}W_0\left(I_m + 2W_0^\top\int_0^t e^{2sA}\mathrm{d}s\, W_0\right)^{-1}W_0^\top e^{tA}. \tag{82}$$

This result, already present in several previous works (Yan et al., 1994; Bodin and Macris, 2023; Martin et al., 2024), can easily be derived by writing the differential equation solved by $Z(t)$ and remarking that (82) solves this equation with the right initial condition. This explicit solution of the Oja flow is very powerful and will be a key result when studying the high-dimensional behavior of the dynamics.

We shall now move on to the convergence properties of the Oja flow. Indeed, this dynamics is very well understood at long times. Originally designed as a dynamical model for principal component recovery, it is no surprise that the dynamics converges to a point that selects the largest eigenvalues of the target matrix A.

Proposition 18 *Let $W(t)$ be solution of the Oja flow* (80)*. Assume that:*

- *The dynamics is initialized at $t = 0$ with a random matrix whose distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$.*
- *The eigenvalues of A are simple.*

Then, with probability one, $W(t)$ converges as $t \to \infty$ toward some W_∞ satisfying:

$$W_\infty W_\infty^\top = A^+_{(m)}, \tag{83}$$

where $A^+_{(m)}$ is defined in equation (36) *and is obtained spectrally by selecting the m largest positive eigenvalues of A.*

The proof of this result is standard and is deferred to Section I.1. It uses the analysis of the local minimizers associated with the loss (81), along with the fact that, under a random initialization, gradient flow almost always converges to a local minimizer.

4.2 Finite-Dimensional Convergence Rates

In this part, we mention the convergence rates associated with the Oja flow toward a solution of Proposition 18. In the interest of our main results in Section 3.2, we restrict ourselves to the case where the target matrix A is invertible and has simple eigenvalues. Indeed, this assumption allows us to obtain some exponentially fast convergence rates. Regarding prior works on this problem, several works have already quantified such convergence in the presence of a positive-definite (Yan et al., 1994) or rank-deficient (Martin et al., 2024) target, and Tsuzuki and Ohki (2025) analyzed the convergence on the Stiefel manifold. Interestingly, several works (Sarao Mannelli et al., 2020; Martin et al., 2024) also identified a power-law convergence in the overparameterized setting.

In the following, we denote p the number of positive eigenvalues of A. Our method is the following:

- For the case $m \leq p$, the gradient flow dynamics converges toward a matrix W which is of full rank. We then prove that in the general case, for functions of $WW^\top$, exponentially fast convergence is guaranteed provided that the Hessian is positive definite when restricted to an appropriate subspace.
- When $m > p$, the gradient flow converges toward a matrix of rank $p < \min(m, d)$, and the previous reasoning does not apply anymore. Using a similar technique to Martin et al. (2024), we derive the rates via Grönwall-type bounds.

4.2.1 Convergence to a Full-Rank Matrix

We first consider the case $m \leq p$. Then, A has more positive eigenvalues than m and the limit derived in Proposition 18 is of rank m (obviously in this case one has $m \leq d$). The first result uses the fact that the loss function (81) exhibits symmetries: it is a function of $WW^\top$ and associated with this symmetry is a set of what we call *uninformative directions*. These directions are flat for the loss U and should not compromise exponentially fast convergence. Indeed, we show that these directions are not seen through the gradient flow trajectory and that we can restrict the Hessian on the informative directions to study convergence.

Proposition 19 *Let $F(W) = G(WW^\top)$ for $W \in \mathbb{R}^{d\times m}$ where $G\colon \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}$ is $\mathcal{C}^2$. Define:*

$$\mathcal{H}_W = \Big\{K \in \mathbb{R}^{d\times m},\, W^\top K = K^\top W\Big\}, \tag{84}$$

and consider the gradient flow dynamics $\dot{W}(t) = -\nabla F(W(t))$. Assume that:

- *$W(t) \xrightarrow[t\to\infty]{} W_\infty$ with $\operatorname{rank}(W_\infty) = m \leq d$.*
- *The restriction of the Hessian of F on $\mathcal{H}_{W_\infty}$ is positive definite with smallest eigenvalue $\varrho > 0$.*

Then, for all $0 < c < \varrho$, we have:

$$F\big(W(t)\big) - F(W_\infty) \underset{t\to\infty}{=} o\big(e^{-2ct}\big), \qquad \big\|Z(t) - Z_\infty\big\| \underset{t\to\infty}{=} o\big(e^{-ct}\big), \tag{85}$$

with $Z(t) = W(t)W(t)^\top$ and $Z_\infty = W_\infty W_\infty^\top$.

We prove this proposition in Section I.2. The proof only involves elementary linear algebra and ordinary differential equations tools, but one could interpret the gradient flow dynamics as a projected gradient flow on an appropriate quotient space. Indeed, the subspace $\mathcal{H}_W$ precisely corresponds to the horizontal space associated with the quotient manifold $\big\{W \in \mathbb{R}^{d\times m}, \operatorname{rank}(W) = m\big\}$ with respect to the equivalence relation $V \sim W \iff VV^\top = WW^\top$. Then, the orthogonal complement of $\mathcal{H}_W$ corresponds to the directions such that the map $W \mapsto WW^\top$ remains constant at first order. Regarding the gradient flow dynamics, these correspond to the uninformative directions, that are inherently flat but do not influence the convergence. More details on the study of this manifold can be found in (Massart and Absil, 2020).

Therefore, in order to derive the convergence rates, we only need to study the Hessian of the loss (81). It can be easily computed:

$$\mathrm{d}^2U_W(H) = \big(WW^\top - A\big)H + HW^\top W + WH^\top W. \tag{86}$$

We will study this linear map at the point of convergence of the dynamics, i.e., a $W \in \mathbb{R}^{d\times m}$ such that $WW^\top = A^+_{(m)}$. We gather the results obtained in the following proposition:

Proposition 20 *Consider $W \in \mathbb{R}^{d\times m}$ such that $WW^\top = A^+_{(m)}$, and the restriction of d^2U_W on the space $\mathcal{H}_W$ (defined in Proposition 19). Denote $\lambda_1 > \cdots > \lambda_d$ the ordered eigenvalues of A and p the number of its positive eigenvalues. Then, for $m \leq p$, on $\mathcal{H}_W$, the Hessian d^2F_W has eigenvalues:*

Eigenvalue	*Multiplicity*	*Indices*
$\lambda_i + \lambda_j$	1	$1 \leq i \leq j \leq m$
$\lambda_i - \lambda_j$	1	$1 \leq i \leq m,\ m+1 \leq j \leq d$

Then, if the flow (80) *converges toward a solution given in Proposition 18, the convergence is exponentially fast, in the sense of equation* (85)*, with rate:*

$$\varrho_{\mathrm{CV}} = \min\big(2\lambda_m, \lambda_m - \lambda_{m+1}\big). \tag{87}$$

This proposition is proved in Section I.3. As a remark, we obtain exponentially fast convergence, but in the usual random matrix theory setting, we expect these rates to be of order d^{-1}. Of course, given the spectrum of the Hessian, this only concerns a very small number of directions.

4.2.2 Convergence to a Rank-Deficient Matrix

In the case $m > p$, the dynamics converges to a matrix of rank p. In this case the result of Proposition 19 does not apply. The convergence rate of the dynamics can be derived by decomposing the dynamics onto the subspaces associated with the positive and negative eigenvalues of A. Some standard bounds and Grönwall-type arguments lead to the following result:

Proposition 21 *Consider the case $m > p$ and denote $W(t)$ the solution of* (80)*. Assume that $W(t)$ converges to the limit given in Proposition 18. Denote $\lambda_1 > \cdots > \lambda_d$ the ordered eigenvalues of A and p the number of its positive eigenvalues. Then, for all $0 < c < \min(2\lambda_p, |\lambda_{p+1}|)$, we have:*

$$U(W(t)) - U(W_\infty) = o(e^{-2ct}), \qquad \big\|Z(t) - Z_\infty\big\| = o(e^{-ct}). \tag{88}$$

We prove this proposition in Section I.4. As suggested, the rate depends on the eigenvalues of A near zero, since λ_p is the smallest positive eigenvalue of A and λ_{p+1} is the largest negative. For instance, in high dimension, if A admits a converging spectral density with mass near zero, this rate is typically of order d^{-1}, which is the same scaling as the previous case. This invites us to investigate the convergence rates in the high-dimensional limit, which we do in the next part.

4.3 Infinite-Dimensional Convergence Rates

Building on the previous convergence rates at finite dimension, we derive similar results but by first taking the high-dimensional limit at fixed $t \geq 0$, and then the long-time limit. In the following, we assume that the empirical spectral distribution of A converges to some μ_A in the high-dimensional limit. Moreover, consistently with the rest of the paper, we consider the limit where $m \sim \kappa d$ as $d \to \infty$. We then define the rescaled number of positive eigenvalues of A in the limit:

$$\kappa_A = \int \mathbf{1}_{x>0} \mathrm{d}\mu_A(x). \tag{89}$$

As in Proposition 20 and 21, the following results will depend on the value of κ with respect to κ_A.

For the sake of simplicity, and in order to use random matrix theory results, we will assume that the initialization W_0 is a Gaussian matrix. However, since we are only interested in the long-time limit, we believe that our results will hold for more general distributions.

We also insist on the fact that the distribution of A remains arbitrary in our case. Indeed, several similar results have been derived, for instance by Bodin and Macris (2023), but for more specific distributions. The result we prove is obtained in two steps:

- We exploit the explicit solution of the Oja flow in Proposition 17 in order to derive an exact high-dimensional limit for the function:

$$\frac{1}{d}\big\|Z(t)-\phi(A)\big\|_F^2, \qquad\qquad Z(t)=W(t)W(t)^\top, \tag{90}$$

  for all $t \geq 0$, where ϕ is a spectral function of A.

- By choosing ϕ associated with the limit of the dynamics, we then compute the $t \to \infty$ asymptotic of the previous quantity, depending on the value of κ, κ_A.

4.3.1 High-Dimensional Limit of the Dynamics

We first derive the high-dimensional limit associated with the Oja flow.

Proposition 22 *Let $W(t) \in \mathbb{R}^{d\times m}$ be the solution of the Oja flow* (80) *with initial condition $W(t=0)=W_0$. Assume that:*

- *W_0 is a centered Gaussian matrix with i.i.d. coefficients of variance $1/m$.*
- *A is almost surely invertible, and its empirical spectral distribution converges as $d \to \infty$ to some probability measure μ_A.*

Let ϕ be a spectral function on $\mathcal{S}_d(\mathbb{R})$. Then, with $Z(t)=W(t)W(t)^\top$, for all $t \geq 0$, as $d \to \infty$ with $m \sim \kappa d$:

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{d}\Big\|Z(t)-\phi(A)\Big\|_F^2 = \mathfrak{g}(t)\left(\int\left(\frac{xe^{xt}}{q_t(x)}\right)^2 \mathrm{d}\mu_A(x)\right)^2\left(\kappa+\int\frac{z(e^{2zt}-1)}{q_t(z)^2}\mathrm{d}\mu_A(z)\right)^{-1}\\ +\int\left(\mathfrak{g}(t)\frac{xe^{2xt}}{q_t(x)}-\phi(x)\right)^2\mathrm{d}\mu_A(x),\end{aligned} \tag{91}$$

where $q_t(x)=(e^{2xt}-1)\mathfrak{g}(t)+x$, and $\mathfrak{g}(t)$ solves the self-consistent equation:

$$\kappa\mathfrak{g}(t)+1-\kappa=\int\frac{x}{(e^{2xt}-1)\mathfrak{g}(t)+x}\mathrm{d}\mu_A(x). \tag{92}$$

This proposition is proved in Section I.5. The proof uses the explicit solution derived in Proposition 17 and interprets this solution using the Stieltjes transform of a Gaussian matrix with time-dependent covariance. Then, the application of some standard random matrix results (that can be found in Bun et al., 2017), as well as some lemmas given in Section I.5, allow us to access (91).

To the best of our knowledge, such a result is new and can be more widely applied to understand the Oja flow dynamics in high dimension.

4.3.2 Convergence Rates in the High-Dimensional Limit

We now explain how to derive convergence rates starting from Proposition 22. To obtain the distance to the limit using (91), one has to pick the spectral function ϕ corresponding to the m largest positive eigenvalues of A. This is done by choosing:

$$\phi(x) = x\mathbf{1}_{x \geq \max(0,\omega)}, \qquad \kappa = \int \mathbf{1}_{x \geq \omega} \mathrm{d}\mu_A(x). \tag{93}$$

Here ω selects a fraction κ of the measure μ_A corresponding to the largest eigenvalues and ϕ applies a threshold using ω and 0 to ensure that the selected eigenvalues are positive. Then, one can study the long-time limit of equation (91) to get the following result:

Proposition 23 *Assume that μ_A admits a density ρ_A. Then, there exists absolute constants $C_1, C_2 > 0$ such that:*

$$\lim_{d\to\infty} \frac{1}{d}\left\|Z(t) - Z_\infty\right\|_F^2 \underset{t\to\infty}{\sim} \begin{cases} C_1 \dfrac{\rho_A(0)}{t^3}, & \text{if } \kappa > \min(\kappa_A, 1) \\ C_2 \dfrac{\omega^2 \rho_A(\omega)}{t}, & \text{if } \kappa < \min(\kappa_A, 1). \end{cases} \tag{94}$$

This result is proved in Section I.6. In both cases, the convergence speed is proportional to the value of the density ρ_A near the smallest eigenvalue of Z_∞. Interestingly, if the density is zero in a neighborhood of this value, the convergence remains exponential, which is consistent with the previous observations at finite dimension. On the other hand, when the density is positive at the threshold, because of the directions with vanishing exponential rates identified in Section 4.2, the convergence becomes a power-law in the high-dimensional limit.

The critical case $\kappa = \min(\kappa_A, 1)$ is not covered by Proposition 23. We believe this regime to be more technical to analyze, and we leave its study for future work.

Figure 14 confirms the previous asymptotics. In this figure, we plot the distance to convergence (left-hand side of equation 94) as a function of time, and with the choice of a target matrix $A \sim \mathrm{GOE}(d)$ (see Definition 26 for details), so that in this case $\kappa_A = 1/2$. This function of the dynamics is then compared with the asymptotic prediction of (94) (black dashed line): the left panel shows the case $\kappa > \kappa_A$ and the right panel $\kappa < \kappa_A$, leading to the distinct power-law behaviors. The comparison reveals an excellent match, down to the constant factor. For very large values of t, the asymptotics seem to break, but we believe that this is due to numerical precision limits in the left panel, and to finite-dimensional effects on the right panel. Indeed, recall that Section 4.2 derived exponentially fast convergence rates: these still occur when the dimension is large, but only on a timescale of order d.

4.4 Linear Response

We shall now study the linear response associated with the Oja flow. This object is central in the high-dimensional equations derived in Section 3.1 and will be necessary to close the system of equations in Section 3.2. We shall now consider the perturbed Oja flow dynamics:

$$\dot{W}(t) = \big(A - W(t)W(t)^\top\big)W(t) + H(t)W(t), \tag{95}$$

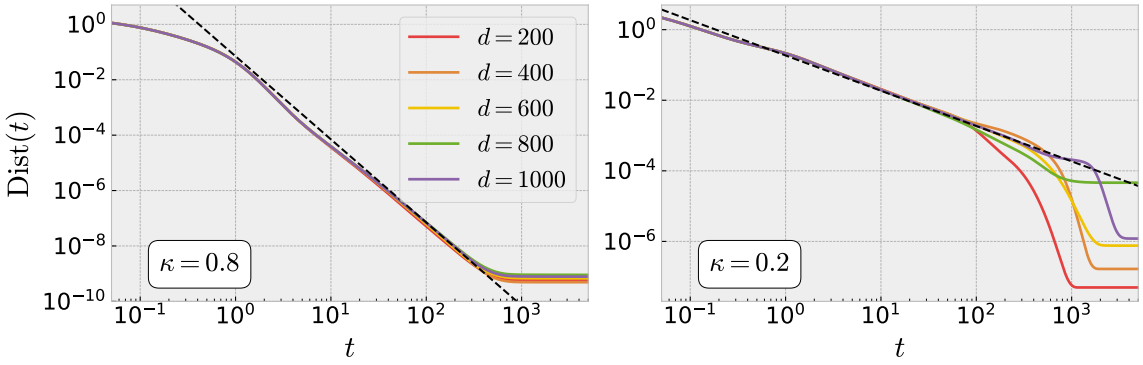

Figure 14: Comparison between the numerical integration of the Oja flow (80) (with step-size $\eta = 5 \times 10^{-3}$) and the power-law asymptotic of equation (94). Colored lines: the quantity $\mathrm{Dist}(t) = \frac{1}{d}\|Z(t) - Z_\infty\|_F^2$, averaged over 20 repetitions of the initialization and target matrix (here distributed as a GOE), and plotted for several values of the dimension d. Black dashed lines: exact power-law asymptotics given by equation (94) (the value of the constants can be found in Section I.6).

for some $H(t) \in \mathcal{S}_d(\mathbb{R})$, and we compute the linear response operator:

$$R(t,t') = \left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0}, \qquad Z(t) = W(t)W(t)^\top. \tag{96}$$

Note that, as emphasized earlier, $R(t,t')$ is a linear map $\mathcal{S}_d(\mathbb{R}) \to \mathcal{S}_d(\mathbb{R})$. The following proposition gives its expression:

Proposition 24 *The linear response of the Oja flow is given by, for $H \in \mathcal{S}_d(\mathbb{R})$:*

$$R(t,t')(H) = \Big(U(t,t')HV(t,t')^\top + V(t,t')HU(t,t')^\top\Big)\mathbf{1}_{t'\le t}, \tag{97}$$

where:

$$U(t,t') = P(t)^{-1}P(t'), \qquad V(t,t') = P(t)^{-1}Z_0 e^{t'A}, \qquad P(t) = e^{-tA} + Z_0 s_t(A), \tag{98}$$

and the function s_t, defined by $s_t(\lambda) = \lambda^{-1}\big(e^{t\lambda} - e^{-t\lambda}\big)$ if $\lambda \neq 0$ and $s_t(0) = 2t$ is applied spectrally to A. In the basis of $\mathcal{S}_d(\mathbb{R})$:

$$E_{ij} = \frac{e_i e_j^\top + e_j e_i^\top}{2},$$

for $i \le j$:

$$\begin{aligned} R_{ijkl}(t,t') &\equiv \mathrm{Tr}\Big(R(t,t')\big(E_{kl}\big)E_{ij}\Big) \\ &= \frac{1}{2}\Big(U_{ik}(t,t')V_{jl}(t,t') + U_{il}(t,t')V_{jk}(t,t') \\ &\qquad + U_{jk}(t,t')V_{il}(t,t') + U_{jl}(t,t')V_{ik}(t,t')\Big)\mathbf{1}_{t'\le t}. \end{aligned} \tag{99}$$

This result is exact and does not require any long-time or high-dimensional assumption. The key point is to decompose $Z(t)$ between the solution at $H = 0$ and a solution of order H, and obtain a differential equation on this solution. Finally, the differential equation can be solved by exploiting the explicit solution at $H = 0$ derived in Proposition 17. The proof details are given in Section I.7.

We now compute the high-dimensional limit of the response operator. To do so, we fix some times t, t' and consider the mean diagonal response:

$$R_{\text{diag}}(t,t') = \lim_{d\to\infty} \frac{1}{d^2}\text{Tr}\left(\frac{\partial Z(t)}{\partial H(t')}\bigg|_{H=0}\right), \qquad r_{\text{diag}}(t) = \int_0^t R_{\text{diag}}(t,t')\mathrm{d}t'. \tag{100}$$

These scalar quantities play an essential role in our system of equations, and in the case of the Oja flow they can be exactly computed. In the same fashion as in Proposition 22, we exploit some well-known random matrix results that require the initialization W_0 to be a Gaussian matrix. However, we believe that all the results that will be considered in the $t \to \infty$ limit will not depend on this choice of distribution.

Proposition 25 *Under the same assumptions as in Proposition 22, with $m \sim \kappa d$ in the $d \to \infty$ limit, we have:*

$$R_{\text{diag}}(t,t') = \frac{\mathfrak{g}(t)}{2}\iint \frac{1}{q_t(x)q_t(y)} e^{(x+y)t}\Big(y(x-\mathfrak{g}(t))e^{(y-x)t'} + x(y-\mathfrak{g}(t))e^{(x-y)t'} \\ + (x+y)\mathfrak{g}(t)e^{(x+y)t'}\Big)\mathrm{d}\mu_A(x)\mathrm{d}\mu_A(y), \tag{101}$$

$$r_{\text{diag}}(t) = \frac{\mathfrak{g}(t)}{2}\iint \frac{1}{y-x}\left[\frac{ye^{2yt}}{q_t(y)} - \frac{xe^{2xt}}{q_t(x)}\right]\mathrm{d}\mu_A(x)\mathrm{d}\mu_A(y), \tag{102}$$

where $q_t(x) = \mathfrak{g}(t)(e^{2xt}-1) + x$ and $\mathfrak{g}(t)$ solves the self-consistent equation:

$$\kappa\mathfrak{g}(t) + 1 - \kappa = \int \frac{x}{(e^{2xt}-1)\mathfrak{g}(t) + x}\mathrm{d}\mu_A(x). \tag{103}$$

We prove this proposition in Section I.8.

Finally, we emphasize that it is technically possible to derive a closed-form expression for higher-order responses for the Oja flow dynamics. In Section E.3, we explain how this can be achieved and how these responses can be used to investigate the stability of the long-time approximations made on the high-dimensional equations in Section 3.2.

5. Conclusion and Perspectives

In this work, we studied the high-dimensional training dynamics of a shallow neural network with quadratic activation function in a teacher–student setting. In the extensive-width regime, we derived a high-dimensional description of the gradient flow dynamics on the empirical loss and obtained asymptotic equations governing its long-time behavior. These results clarify how overparameterization affects learning and generalization, and help understand overfitting and double descent phenomena. Overall, our findings contribute to a clearer theoretical picture of learning dynamics in high-dimensional, overparameterized neural networks.

Toward a more rigorous analysis. Despite the theoretical insights and numerical evidence presented in this paper, several aspects of our analysis call for a more rigorous treatment.

- The derivation of the equivalent dynamics in the high-dimensional limit (Section 3.1) relies on statistical physics techniques that have been applied to several learning problems. To obtain fully rigorous results, one would need to generalize the mathematical analysis developed for finite width to the extensive width case, in order to characterize the limiting high-dimensional object and to formalize convergence in distribution.
- The simplification of the dynamics at long times (Section 3.2) is based on structural assumptions on the dynamics. Confirming this ansatz mathematically is challenging, and comparable rigorous results exist only for simpler models (Celentano et al., 2021; Fan et al., 2025; Chen and Shen, 2025).
- Our analysis relies on a Gaussian surrogate for the rank-one sensing matrices associated with the quadratic model. While we provide theoretical arguments and numerical evidence supporting the equivalence between these two models (Section 3.5), we believe that this equivalence could be established rigorously and extended to a broader class of distributions. Similar universality results are well understood in the static setting (Maillard et al., 2024; Erba et al., 2025b; Xu et al., 2025), but extending them to high-dimensional trajectories remains an open problem.

Generalizations. We expect that the dynamical framework developed in this work can be extended to other settings of interest. As noted by Maillard et al. (2024), studying polynomial activation functions leads to a lifting of the model to a linear one in the space of tensors. This extension introduces new mathematical challenges, in particular the analysis of the resulting nonlinear tensor dynamics under similar assumptions as those introduced in Section 3.2.

Another promising direction concerns the study of learning dynamics in transformer architectures. Simplified attention mechanisms have already been analyzed in the Bayes-optimal and empirical risk minimization settings, including in deep architectures (Troiani et al., 2025; Erba et al., 2025a; Boncoraglio et al., 2025a,b). Extending these approaches to a dynamical setting remains challenging and calls for further attention.

Acknowledgments

We thank Lenka Zdeborová, Florent Krzakala, Emanuele Troiani, Vittorio Erba, Antoine Maillard, Bruno Loureiro, Guilhem Semerjian, Gérard Ben Arous, Lénaïc Chizat, Louis-Pierre Chaintron, Nicolas Boumal and Andreea-Alexandra Muşat for useful remarks and discussions. We also thank the anonymous reviewers for their helpful comments and suggestions.

This work has received support from the French government, managed by the National Research Agency, under the France 2030 program with the reference "PR[AI]RIE-PSAI" (ANR-23-IACL-0008).

Appendix A. Guide to the Appendices

The appendices contain several types of technical material. This short guide summarizes the role of each appendix and clarifies its connections to the main text as well as its level of rigor. The latter is indicated in the following table with the color code of the first column: a green cell refers to a section containing rigorous results, orange stands for a physics-based derivation, and purple is used for numerical confirmation.

Section	**Content**	**Used in**
Section B	Preliminary technical results.	
Section C	Derivation of the high-dimensional DMFT equations, study of the simplified setting (Assumption 3), and discussion on Gaussian equivalence.	Section 3.1, Section 3.5
Section D	Derivation of the long-time scalar equations from the steady-state ansatz (Assumption 5), correspondence with Erba et al. (2025b), impact of overparameterization, population limit.	Section 3.2
Section E	Stability analysis of the steady-state solution.	Section 3.2.4
Section F	Analysis of Langevin dynamics from time-translational invariance and fluctuation-dissipation (Assumption 43).	Section 3.2.7
Section G	Derivation of the aging equations under the assumption of slow relaxation to equilibrium.	Section 3.2.4
Section H	Study of the small regularization limit and derivation of the interpolation and perfect recovery thresholds.	Section 3.3
Section I	Analysis of the Oja flow: finite and infinite-dimensional convergence rates, linear response.	Section 4
Section J	Numerical details and additional simulations.	

Additionally, we refer to Figure 1 for the dependencies between the main results and appendices of the paper.

Appendix B. Technical Background

This section introduces additional definitions, background material, and technical lemmas that will be used throughout the appendix. More precisely:

- In Section B.1, we recall some standard results on random matrices, introduce Stieltjes and Hilbert transforms of probability measures, and study the case of the free convolution with a semicircular density.

- In Section B.2, we state useful properties of the best low-rank positive semidefinite approximation of a symmetric matrix.

- Finally, in Section B.3, we prove technical lemmas used in the derivation of the high-dimensional equations in Section C.3.

B.1 Asymptotic Spectral Properties of Random Matrices

We start by recalling standard random matrix ensembles and their spectral properties in the high-dimensional limit.

B.1.1 Standard Matrix Ensembles

We now proceed to define the GOE and Wishart distributions, and state the well-known convergence results regarding their respective spectral distributions in the large dimension limit.

Definition 26 *We say that $X \in \mathcal{S}_d(\mathbb{R})$ is distributed according to the GOE(d) distribution if:*

$$X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{1+\delta_{ij}}{d}\right), \qquad \big(1 \le i \le j \le d\big).$$

As a consequence, if $X \sim \mathrm{GOE}(d)$, the mapping $A \in \mathcal{S}_d(\mathbb{R}) \mapsto \mathrm{Tr}(AX) \in \mathbb{R}$ defines a centered Gaussian process on $\mathcal{S}_d(\mathbb{R})$ with covariance function:

$$\mathbb{E}\,\mathrm{Tr}(AX)\mathrm{Tr}(BX) = \frac{2}{d}\mathrm{Tr}(AB).$$

More properties of this distribution can be found, for instance, in Anderson et al. (2010, Chapter 2).

Definition 27 *We say that $Z \in \mathcal{S}_d^+(\mathbb{R})$ is distributed according to the Wishart distribution $\mathcal{W}(m,d)$ if:*

$$Z = \frac{1}{m} W W^\top,$$

where $W \in \mathbb{R}^{d\times m}$ has i.i.d. standard Gaussian entries.

These two matrix ensembles are fundamental in random matrix theory and are used as key models for understanding the behavior of large complex systems. They appear in a wide range of fields, from statistical physics to statistics and machine learning, where they model random data and covariance structures. A key aspect of both ensembles is the behavior of their eigenvalue distributions in the high-dimensional limit.

Theorem 28 *(Wigner, 1955; Marchenko and Pastur, 1967). Let $\mathcal{G} \sim \mathrm{GOE}(d)$ and $Z \sim \mathcal{W}(m,d)$, with $m \sim \kappa d$ as $d \to \infty$. Then, almost surely as $d \to \infty$, the empirical spectral distributions of $\mathcal{G}$ and Z respectively weakly converge to the probability distributions:*

- *The semicircular distribution σ:*

$$\mathrm{d}\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{|x|\le 2}\mathrm{d}x,$$

- *The Marchenko–Pastur distribution:*

$$\mathrm{d}\nu_\kappa(x) = (1-\kappa)^+\delta(x) + \frac{\kappa}{2\pi x}\sqrt{(\lambda_+ - x)(x-\lambda_-)}\mathrm{d}x, \qquad \lambda_\pm = \left(1 \pm \frac{1}{\sqrt{\kappa}}\right)^2.$$

It will be useful to consider the rescaled semicircular distribution with variance ξ:

$$\mathrm{d}\sigma_\xi(x) = \frac{1}{2\pi\xi}\sqrt{4\xi - x^2}\mathbf{1}_{|x|\le 2\sqrt{\xi}}\mathrm{d}x. \tag{104}$$

B.1.2 Stieltjes and Hilbert Transforms

We shall now define two important transforms of probability measures. More details can be found in Bun et al. (2017).

Definition 29 *For a probability measure μ on $\mathbb{R}$, we define its Stieltjes transform m_μ and Hilbert transform h_μ as:*

$$m_\mu(z) = \int \frac{\mathrm{d}\mu(y)}{z-y}, \qquad\qquad h_\mu(x) = \text{P.V.} \int \frac{\mathrm{d}\mu(y)}{x-y},$$

for $z \in \mathbb{C}$ and $x \in \mathbb{R}$. The notation P.V. denotes Cauchy's principal value. More precisely:

$$h_\mu(x) = \lim_{\epsilon \to 0} h_\mu^\epsilon(x), \qquad\qquad h_\mu^\epsilon(x) = \int \mathbf{1}_{|x-y|>\epsilon} \frac{\mathrm{d}\mu(y)}{x-y}. \tag{105}$$

Moreover, if μ admits a density ρ, then, for $x \in \mathrm{Supp}(\mu)$:

$$\lim_{\eta \to 0^+} m_\mu(x + i\eta) = h_\mu(x) - i\pi\rho(x).$$

We now state a technical lemma on the limit in equation (105).

Lemma 30 *Let μ be a compactly supported probability measure with no atoms, and h_μ its Hilbert transform. For $\epsilon > 0$, consider h_μ^ϵ as defined in equation* (105)*.*

1. *Let $f : \mathbb{R} \to \mathbb{R}$ be either Lipschitz or non-decreasing on the support of μ. Then:*

$$\lim_{\epsilon \to 0^+} \int f(x) h_\mu^\epsilon(x) \mathrm{d}\mu(x) = \frac{1}{2} \iint \frac{f(x) - f(y)}{x-y} \mathrm{d}\mu(x) \mathrm{d}\mu(y).$$

2. *Moreover, if μ admits a bounded square-integrable density with respect to the Lebesgue measure, then for any $f \in L^2(\mu)$:*

$$\lim_{\epsilon \to 0^+} \int f(x) h_\mu^\epsilon(x) \mathrm{d}\mu(x) = \int f(x) h_\mu(x) \mathrm{d}\mu(x).$$

Proof For $\epsilon > 0$, define:

$$I_\epsilon \equiv \frac{1}{2} \iint \frac{f(x) - f(y)}{x-y} \mathbf{1}_{|x-y|>\epsilon} \mathrm{d}\mu(x) \mathrm{d}\mu(y).$$

Then, it is clear, by symmetry, that:

$$I_\epsilon = \int f(x) h_\mu^\epsilon(x) \mathrm{d}\mu(x). \tag{106}$$

Now, if f is non-decreasing, for any $x, y \in \mathrm{Supp}(\mu)$ such that $x \neq y$, the family:

$$\left(\frac{f(x) - f(y)}{x-y} \mathbf{1}_{|x-y|>\epsilon} \right)_{\epsilon > 0},$$

is non-decreasing as $\epsilon \to 0$, and I_ϵ converges by the monotone convergence theorem. On the other hand, if f is Lipschitz, the previous family is bounded in absolute value by the Lipschitz constant of f and the integral converges toward a finite value due to the dominated convergence theorem. In both cases the limit is given by:

$$\lim_{\epsilon \to 0} I_\epsilon = \frac{1}{2} \iint \frac{f(x) - f(y)}{x - y} \mathbf{1}_{x \neq y} \mathrm{d}\mu(x) \mathrm{d}\mu(y).$$

Since μ has not atoms, the diagonal has zero mass and the identity (106) concludes the first part.

For the second point, denote ρ the density of μ, and assume that ρ is bounded and in $L^2(\mathrm{d}x)$. Now, it is standard (see for instance Grafakos et al., 2008, Theorem 5.1.12) that h_μ^ϵ converges to h_μ in $L^2(\mathrm{d}x)$. Therefore:

$$\begin{aligned}\left| \int f(x) \big(h_\mu(x) - h_\mu^\epsilon(x)\big) \mathrm{d}\mu(x) \right| &\leq \big\| f(h_\mu - h_\mu^\epsilon) \big\|_{L^1(\mu)} \\ &\leq \|f\|_{L^2(\mu)} \sqrt{\|\rho\|_\infty} \|h_\mu - h_\mu^\epsilon\|_{L^2(\mathrm{d}x)}.\end{aligned}$$

By assumptions, this goes to zero as $\epsilon \to 0$. This gives the result. ■

We now formulate a direct corollary of this result.

Lemma 31 *Let μ be a compactly supported probability measure, with no atoms, with a bounded and square-integrable density, and h_μ its Hilbert transform. Then:*

$$\int x h_\mu(x) \mathrm{d}\mu(x) = \frac{1}{2}.$$

Proof As a consequence of Lemma 30, we have with $f(x) = x$:

$$\int x h_\mu(x) \mathrm{d}\mu(x) = \lim_{\epsilon \to 0^+} \int x h_\mu^\epsilon(x) \mathrm{d}\mu(x) = \frac{1}{2} \iint \frac{x - y}{x - y} \mathrm{d}\mu(x) \mathrm{d}\mu(y) = \frac{1}{2}.$$

■

B.1.3 Free Convolution with a Semicircular Density

In this section, we will be interested in the free additive convolution between some given probability distribution μ^* and σ_ξ, defined in equation (104). We will denote this resulting measure as $\mu_\xi = \mu^* \boxplus \sigma_\xi$. This object is of great interest for this work as it describes the asymptotic spectral distribution of a random matrix of the form $Z^* + \sqrt{\xi}\mathcal{G}$, for $\mathcal{G} \sim \mathrm{GOE}(d)$, and where Z^* is a random matrix independent from $\mathcal{G}$, with asymptotic spectral distribution μ^*. We refer to Biane (1997) for several results on this object.

The first relevant result for us is the so-called subordination identity between the Stieltjes transforms of μ_ξ and μ^*.

Lemma 32 *Let μ_ξ be the free additive convolution between some probability measure μ^* and a semicircular density with variance $\xi > 0$. Consider m_ξ and m_* the Stieltjes transforms of μ_ξ and μ^*. We have the equation, for all $z \in \mathbb{C} \setminus \mathrm{Supp}(\mu_\xi)$:*

$$m_\xi(z) = m_*\big(z - \xi m_\xi(z)\big). \tag{107}$$

A proof of this can be found in Biane (1997). This relationship between Stieltjes transforms is known to characterize the distribution μ_ξ.

Lemma 33 *Let m_ξ be the Stieltjes transform of μ_ξ. Then the map $(\xi, z) \mapsto m_\xi(z)$ is $\mathcal{C}^1$ on $\mathbb{R}^+ \times \mathbb{C} \setminus \mathbb{R}$ and analytic in z for fixed ξ. Moreover, m_ξ satisfies the complex Burgers' equation:*

$$\partial_\xi m_\xi(z) + m_\xi(z)\partial_z m_\xi(z) = 0, \tag{108}$$

whenever $\xi \geq 0$ and $z \in \mathbb{C} \setminus \mathbb{R}$, with initial condition $m_0(z) = m_*(z)$.

A reference for this result can be found in Voiculescu (1997).

B.2 Best PSD Low-Rank Approximation

Throughout the main text and the appendix, we often deal with the notation $A^+_{(m)}$ for a matrix $A \in \mathcal{S}_d(\mathbb{R})$. Although it is already defined in equation (36), we now provide a formal definition as well as some relevant properties.

Definition 34 *Let $A \in \mathcal{S}_d(\mathbb{R})$ with eigenvalue decomposition $A = U\mathrm{diag}(\lambda_1, \dots, \lambda_d)U^\top$ and $\lambda_1 \geq \cdots \geq \lambda_d$, we define for $m \leq d$:*

$$A^+_{(m)} = U\mathrm{diag}\big(\lambda_1^+, \dots, \lambda_m^+, 0, \dots, 0\big)U^\top, \qquad \lambda^+ = \max(\lambda, 0).$$

By convention, if $m > d$, we define $A^+_{(m)} = A^+$ to be the positive part of A.

This matrix is known to be the best PSD approximation of A by a matrix of rank at most m, for the Frobenius norm.

Lemma 35 *Assume that A has simple eigenvalues. Then:*

$$\Big\{A^+_{(m)}\Big\} = \underset{\substack{S \in \mathcal{S}_d^+(\mathbb{R}) \\ \mathrm{rank}(S) \leq m}}{\mathrm{argmin}} \big\|A - S\big\|_F^2.$$

Proof This property is standard. To prove it, one can realize the argmin is the image through the map $W \mapsto WW^\top$ of the set:

$$\underset{W \in \mathbb{R}^{d \times m}}{\mathrm{argmin}} \big\|A - WW^\top\big\|_F^2.$$

Now, we precisely show in Section I.1 that whenever A has simple eigenvalues, any local minimizer of the map $W \mapsto \big\|A - WW^\top\big\|_F^2$ verifies $WW^\top = A^+_{(m)}$. Since any element in the previous argmin is a global minimizer, this concludes the proof. ■

The following lemma computes the differential of the map $A \mapsto A^+_{(m)}$ on $\mathcal{S}_d(\mathbb{R})$.

Lemma 36 *Let $A \in \mathcal{S}_d(\mathbb{R})$ with simple and non-zero eigenvalues $\lambda_1 > \cdots > \lambda_d$. Write $A = U\mathrm{diag}(\lambda_1, \dots, \lambda_d)U^\top$. Then, the map $\Phi\colon X \in \mathcal{S}_d(\mathbb{R}) \mapsto X^+_{(m)} \in \mathcal{S}_d(\mathbb{R})$ is differentiable at A and has differential:*

$$\mathrm{d}\Phi_A(H) = U\Big(D \circ \big(U^\top H U\big)\Big)U^\top,$$

where $\circ$ denotes the Hadamard product (pointwise product between matrices), and $D \in \mathcal{S}_d(\mathbb{R})$ is defined as:

$$D_{ij} = \begin{cases} \dfrac{\lambda_i \mathbf{1}_{\lambda_i>0}\mathbf{1}_{i\le m} - \lambda_j \mathbf{1}_{\lambda_j>0}\mathbf{1}_{j\le m}}{\lambda_i - \lambda_j}, & \text{if } i \neq j, \\ \mathbf{1}_{\lambda_i>0}\mathbf{1}_{i\le m}, & \text{if } i = j. \end{cases}$$

As a consequence, the spectrum of $\mathrm{d}\Phi_A$ is given by:

$$\mathrm{Sp}\big(\mathrm{d}\Phi_A\big) = \big\{D_{ij},\, 1 \le i \le j \le d\big\},$$

and we have:

$$\mathrm{Tr}\,\mathrm{d}\Phi_A = \sum_{1\le i\le j\le d} D_{ij}, \qquad \big\|\mathrm{d}\Phi_A\big\|_F^2 = \sum_{1\le i\le j\le d} D_{ij}^2.$$

Proof As a consequence of Daleckiĭ–Kreĭn theorem (see for instance Noferini, 2017, Theorem 2.11), provided that $F\colon \mathbb{R} \to \mathbb{R}$ is continuously differentiable on the spectrum of A, and if F^{sp} is the associated spectral function on $\mathcal{S}_d(\mathbb{R})$:

$$F^{\mathrm{sp}}\Big(U\mathrm{diag}(\alpha_1, \dots, \alpha_d)U^\top\Big) = U\mathrm{diag}\big(F(\alpha_1), \dots, F(\alpha_d)\big)U^\top, \tag{109}$$

for any orthogonal matrix $U \in O_d(\mathbb{R})$ and $\alpha_1, \dots, \alpha_d \in \mathbb{R}^d$, then F^{sp} is differentiable (in the Fréchet sense) at A and has differential:

$$\mathrm{d}F^{\mathrm{sp}}_A(H) = U\Big(D \circ \big(U^\top H U\big)\Big)U^\top, \tag{110}$$

where:

$$D_{ij} = \begin{cases} \dfrac{F(\alpha_i) - F(\alpha_j)}{\alpha_i - \alpha_j}, & \text{if } i \neq j, \\ F'(\alpha_i), & \text{if } i = j. \end{cases} \tag{111}$$

Now recall that in our case, we are interested in $\Phi(X) = X^+_{(m)}$, i.e., the truncation of X onto the subspace associated with the m largest positive eigenvalues of X. Note that such a map is not properly a spectral function in the sense of equation (109). To solve this problem, we introduce some threshold ω such that $\lambda_{m+1} < \omega < \lambda_m$. Due to the continuity of the spectrum at A, that has simple eigenvalues by assumption, there exists a neighborhood $\mathcal{V} \in \mathcal{S}_d(\mathbb{R})$ of A such that for every $X \in \mathcal{V}$, we also have $\lambda_{m+1}(X) < \omega < \lambda_m(X)$. Therefore, on this neighborhood, Φ coincides with the spectral function associated with the real function $F(x) = x\mathbf{1}_{x\ge\max(0,\omega)}$. By assumption F is differentiable on the spectrum of A, and combining the identities (110), (111) this leads to the desired result.

Regarding the spectrum of $\mathrm{d}\Phi_A$, note that $H \in \mathcal{S}_d(\mathbb{R})$ solves $\mathrm{d}\Phi_A(H) = \gamma H$ for some $\gamma \in \mathbb{R}$ if and only if $K = U^\top H U$ solves $D \circ K = \gamma K$. Taking K to be, for $1 \le i \le j \le d$:

$$K = \frac{e_i e_j^\top + e_j e_i^\top}{2},$$

leads to the desired identity with $\gamma = D_{ij}$.

The trace of $\mathrm{d}\Phi_A$ can now be computed as the sum of its eigenvalues. For the Frobenius norm, let us remark that, for any $H, K \in \mathcal{S}_d(\mathbb{R})$:

$$\mathrm{Tr}\big(\mathrm{d}\Phi_A(H)K\big) = \sum_{i,j=1}^{d} D_{ij}\big(U^\top H U\big)_{ij}\big(U^\top K U\big)_{ij},$$

so that $\mathrm{d}\Phi_A$ is self-adjoint (as a linear map on $\mathcal{S}_d(\mathbb{R})$) and $\big\|\mathrm{d}\Phi_A\big\|_F^2$ can be computed as the sum of the squared eigenvalues. ■

B.3 Prerequisite for the High-Dimensional Equations

We now state a few lemmas that will be used in Section C.3 for the derivation of high-dimensional dynamics. We start with a lemma giving the derivative of a Gaussian expectation with respect to its covariance matrix.

Lemma 37 *Let $X \in \mathbb{R}^N$ be a centered Gaussian vector with covariance $K \in \mathcal{S}_N^{++}(\mathbb{R})$. Let $F\colon \mathbb{R}^N \to \mathbb{C}$ such that:*

$$\mathbb{E}\big[|F(X)|\big] < \infty, \qquad \mathbb{E}\Big[|F(X)|\big\|XX^\top\big\|_F\Big] < \infty.$$

Then, viewing $\mathbb{E}\big[F(X)\big]$ as a function of K on $\mathcal{S}_N^{++}(\mathbb{R})$:

$$\nabla_K \mathbb{E}\big[F(X)\big] = -\frac{1}{2}\mathbb{E}\big[F(X)\big]K^{-1} + \frac{1}{2}K^{-1}\,\mathbb{E}\big[F(X)XX^\top\big]K^{-1}.$$

Proof This result is a simple consequence of the formula:

$$\mathbb{E}\big[F(X)\big] = \frac{1}{(2\pi)^{N/2}}\big(\det K\big)^{-1/2}\int \exp\left(-\frac{1}{2}x^\top K^{-1}x\right)F(x)\mathrm{d}x.$$

Now, it is well-known that on the space of positive-definite symmetric matrices:

$$\nabla_K \det K = (\det K)K^{-1}, \qquad \nabla_K(x^\top K^{-1}x) = -K^{-1}xx^\top K^{-1}.$$

Now, the assumption on F guarantees that we can apply dominated convergence and differentiate under the integral. This leads to the desired. ■

In Section C.3, we will use a generalization of this result for a multi-dimensional Gaussian process $\big(X(t)\big)_{t\in[0,T]}$ on $\mathbb{R}^q$, with covariance $K_{ij}(s,t) = \mathbb{E}\,X_i(s)X_j(t)$ for $i,j \in \{1,\dots,q\}$ and $s,t \in [0,T]$. Informally, the result of Lemma 37 still applies provided that we add contractions over time variables in addition to those over indices:

$$\begin{aligned}\frac{\partial\,\mathbb{E}\big[F(X)\big]}{\partial K_{ij}(s,t)} &= -\frac{1}{2}\mathbb{E}\big[F(X)\big]\big(K^{-1}\big)_{ij}(s,t) \\ &\quad + \frac{1}{2}\sum_{i',j'=1}^{q}\int_0^T\!\!\int_0^T \big(K^{-1}\big)_{ii'}(s,s')\big(K^{-1}\big)_{jj'}(t,t')\mathbb{E}\Big[F(X)X_{i'}(s')X_{j'}(t')\Big]\mathrm{d}s'\mathrm{d}t'.\end{aligned} \tag{112}$$

K^{-1} denotes the functional inverse of K with respect to both time and indices:

$$\sum_{k=1}^{q} \int_0^T K_{ik}(t,u)\big(K^{-1}\big)_{kj}(u,s)\mathrm{d}u = \delta(t-s)\delta_{ij}. \tag{113}$$

δ denotes the Dirac delta distribution supported at zero. See Section C.1 for more details.

The following lemma gives the inverse of a certain type of 3×3 block matrix, that will be useful in Section C.3.

Lemma 38 *Let $A \in \mathcal{S}_N(\mathbb{R})$, $B \in \mathbb{R}^{N\times N}$ invertible, $c \in \mathbb{R}^N$ and $\lambda \in \mathbb{R}^*$. Then:*

$$\begin{pmatrix} A & B & c \\ B^\top & 0 & 0 \\ c^\top & 0 & \lambda \end{pmatrix}^{-1} = \begin{pmatrix} 0 & (B^\top)^{-1} & 0 \\ B^{-1} & M & -\lambda^{-1}B^{-1}c \\ 0 & -\lambda^{-1}c^\top(B^\top)^{-1} & \lambda^{-1} \end{pmatrix},$$

with:

$$M = B^{-1}\left(\frac{1}{\lambda}cc^\top - A\right)(B^\top)^{-1}.$$

Again, this result can be generalized in order to compute the inverse of the covariance function K as in equation (113), with $q = 3$. The specific structure of the matrix in Lemma 38 would correspond, in continuous-time, to a Gaussian process in $\mathbb{R}^3$ with a time-independent third coordinate.

Appendix C. Derivation of the High-Dimensional Dynamics

In this section, we derive the results presented in Section 3.1. The plan goes as follows:

- In Section C.1, we give some background on the methods we use and introduce the relevant objects for the derivation of the high-dimensional equations.
- In Section C.2, we give the full set of self-consistent equations mentioned in Claim 2, associated with the gradient flow dynamics (5) for general cost function, noise channel, and regularization.
- In Section C.3, we derive this set of equations.
- In Section C.4, we simplify the set of equations when dealing with the quadratic cost and a Gaussian noise channel. This leads to the system of equations of Claim 4.
- In Section C.5, we show that our method can be generalized to learning the output weights of the network. We explain how this can be achieved and give the associated set of equations.
- In Section C.6, we discuss Conjecture 16 and the relationship with quadratic networks, and give the steps that would lead to a rigorous proof of this conjecture.

C.1 Path Integral Formalism

Before giving and deriving the set of equations of Claim 2, we give some definitions and prior results on the objects involved in our calculation.

The technique we use to derive the high-dimensional equations is often referred to as the *path integral* method. This formalism provides a powerful way to study the dynamics of high-dimensional disordered systems. It was first introduced in spin-glass dynamics by De Dominicis (1978), who formulated the dynamics using a functional-integral and generating-functional approach. This idea was further developed by Sompolinsky and Zippelius (1982), who used it to derive the dynamical mean-field equations for spin-glass models. Since then, the same approach has been applied in many contexts, from classical spin models to modern learning and inference problems (Crisanti et al., 1993; Agoritsas et al., 2018; Mignacco et al., 2020; Bordelon and Pehlevan, 2022), to obtain self-consistent dynamical equations describing the collective behavior of complex random systems. While these derivations are non-rigorous, such results have been rigorously obtained on similar problems, for instance by Ben Arous et al. (2001), who derives a large deviation principle for spherical spin glasses. From this perspective, the path integral formulation is closely related, as it relies on a saddle-point asymptotic that is similar to the large deviation approach, though in a non-rigorous setting.

The dynamical partition function. For simplicity, consider a gradient flow dynamics similar to ours in equation (5):

$$\dot{W}(t) = -\nabla_W F\big(W(t), X\big),$$

where F depends on some random parameters X. In our case this would correspond to the sensing matrices and random labels. To study this dynamics, one can construct a functional integral representation, leading to an object encoding the full trajectory of the system. Formally, one can integrate over the set of all possible trajectories $\big\{W(t)\big\}_{t\in[0,T]}$ using a measure that we write $\mathcal{D}W$. This measure can be interpreted as the continuous limit of the product measure at discrete time instants:

$$\mathcal{D}W = \lim_{N\to\infty} \prod_{p=0}^{N} \mathrm{d}W_{pT/N}.$$

The object we then consider is the dynamical partition function:

$$\mathcal{Z}_{\mathrm{dyn}} = \int \mathcal{D}W\, \delta\Big(\dot{W}(t) + \nabla_W F\big(W(t), X\big)\Big).$$

δ denotes a functional Dirac delta that imposes the equation of motion at every instant. Integrating over all possible trajectories simply counts the unique trajectory consistent with a given initial condition. The resulting integral is therefore a constant, independent of the particular realization of the disorder or the function F.

Path integrals have been widely used to describe stochastic and dynamical systems, and several works have studied how to define them more rigorously. For readers interested in more systematic or rigorous discussions of these constructions, we refer to Chow and Buice (2015), Cugliandolo et al. (2019), De Pirey et al. (2022), and also Dupuis (2023, Chapter 1) for a general introduction and physical applications.

Representation of the functional Dirac. To make the expression of the partition function easy to work with, it is convenient to rewrite the functional Dirac delta in its exponential form. Similarly to the well-known expression in $\mathbb{R}^p$:

$$\delta(y) = \frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} \mathrm{d}\hat{y}\, e^{iy^\top \hat{y}},$$

the same identity extends to functional integrals by introducing a time-dependent *conjugate field* $\hat{W}(t)$. The partition function then rewrites, up to a constant depending only on the dimension:

$$\mathcal{Z}_{\mathrm{dyn}} \propto \int \mathcal{D}W \mathcal{D}\hat{W}\, \exp\left(i \int_0^T \mathrm{Tr}\Big(\Big[\dot{W}(t) + \nabla_W F\big(W(t), X\big) \Big] \hat{W}(t)^\top \Big) \mathrm{d}t \right). \tag{114}$$

$\hat{W}$ plays the role of a Lagrange multiplier that enforces the equation of motion. This exponential form is particularly useful since it allows to manipulate the dynamics using field-theoretic tools. For instance, this formulation will later allow us to average the dynamical partition function with respect to the Gaussian randomness of X.

Link with the generating functional. The dynamical partition function can be viewed as a central object characterizing the full distribution of a trajectory. In practice, one can introduce external source fields $J(t), \hat{J}(t)$ that couple linearly to $W(t)$ and $\hat{W}(t)$:

$$\mathcal{Z}_{\mathrm{dyn}}[J, \hat{J}] = \int \mathcal{D}W \mathcal{D}\hat{W} \exp\left(-S(W, \hat{W}) + \int_0^T \Big[\mathrm{Tr}\big(J(t) W(t)^\top\big) + \mathrm{Tr}\big(\hat{J}(t) \hat{W}(t)^\top\big) \Big] \mathrm{d}t \right),$$

where $S(W, \hat{W})$ is the quantity already appearing inside the exponential in equation (114). The function $\mathcal{Z}_{\mathrm{dyn}}[J, \hat{J}]$, often called the generating functional of the dynamics (the analogue of a generating function for finite-dimensional random variables), contains all statistical information about the process. Correlations of any observable can be obtained by taking functional derivatives of $\mathcal{Z}_{\mathrm{dyn}}[J, \hat{J}]$ with respect to $J, \hat{J}$, and then setting these variables to zero. In other words, the partition function (and its extension to the generating functional) fully characterizes the trajectory distribution of the random process W.

The temporal Dirac delta distribution. Above, we introduced the Dirac delta functional on trajectories. In what follows, we also use the Dirac delta distribution on a real scalar variable, denoted again by δ, with the usual meaning of a distribution supported at zero. As in Claim 4, equation (27), we sometimes use the abusive notation:

$$R(t, t') = \delta(t - t') + G(t, t'),$$

for some function $G\colon \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$. The previous expression is to be understood in the sense of distributions: R implicitly represents a linear operator $\overline{R}$ acting on test functions $\phi\colon \mathbb{R}^+ \to \mathbb{R}$:

$$\big(\overline{R}\phi\big)(t) = \phi(t) + \int_0^t G(t, t')\phi(t')\mathrm{d}t'.$$

C.2 Full Set of Equations

In this section, we give the general set of high-dimensional equations associated with the dynamics of the matrix $W(t)$ and the typical label $y(t)$. The general structure of the equations was given in Claim 2, but for completeness we give them in detail.

In the following, we show that in the high-dimensional limit, starting from the dynamics (5), the evolution of the student matrix and the typical label is equivalent, in distribution, to the following set of equations:

$$\mathrm{d}W(t) = \left(\mathcal{H}(t) + r(t)Z^* - \int_0^t \Gamma(t,t')Z(t')\mathrm{d}t'\right)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{115}$$

$$0 = \int_0^t R(t,t')y(t')\mathrm{d}t' + \eta(t) - m(t)y^* + \frac{2}{\alpha}\ell'\big(y(t),z\big), \tag{116}$$

with $Z(t) = W(t)W(t)^\top$, $y^* \sim \mathcal{N}(0, 2Q_*)$ and $z \sim P\big(\,\cdot\,|\,y^*\big)$. The functions $\mathcal{H}$ and η are independent centered Gaussian processes respectively belonging to $\mathcal{S}_d(\mathbb{R})$ and $\mathbb{R}$, with covariances:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{2d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\mathcal{K}_Z(t,t'), \qquad \mathbb{E}\,\eta(t)\eta(t') = \mathcal{K}_y(t,t'). \tag{117}$$

We then consider the averaged quantities with respect to the dynamics of $Z(t), y(t)$:

$$C_y(s,t) = \mathbb{E}\,y(s)y(t), \qquad m_y(t) = \mathbb{E}\,y(t)y^*, \tag{118}$$

$$C_Z(s,t) = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(s)Z(t)\big), \qquad m_Z(t) = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z^*\big), \tag{119}$$

$$Q_* = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}), \tag{120}$$

as well as the first-order derivatives:

$$R_y(s,t) = \left.\frac{\partial\,\mathbb{E}\,y(s)}{\partial h(t)}\right|_{h=0}, \qquad R_Z(s,t) = \frac{1}{d^2}\mathrm{Tr}\left(\left.\frac{\partial\,\mathbb{E}\,Z(s)}{\partial H(t)}\right|_{H=0}\right), \tag{121}$$

in response to a perturbation $\eta(t) \mapsto \eta(t) - h(t)$ and $\mathcal{H}(t) \mapsto \mathcal{H}(t) + H(t)$ in equations (115), (116). Then, the scalar deterministic functions $r, \Gamma, \mathcal{K}_Z, \mathcal{K}_y, m, R$ can be self-consistently computed as:

$$r(t) = \alpha\left(\frac{1}{2Q_*}\int_0^t R(t,t')m_y(t')\mathrm{d}t' - \int_0^t\int_0^{t'} R(t,t')R_y(t',s)m(s)\mathrm{d}s\mathrm{d}t'\right), \tag{122}$$

$$\Gamma(s,t) = \alpha\left(R(s,t) - \int_t^s\int_t^{s'} R(s,s')R_y(s',t')R(t',t)\mathrm{d}t'\mathrm{d}s'\right), \tag{123}$$

$$\begin{aligned}\mathcal{K}_Z(s,t) = \frac{\alpha}{2}\Bigg(&\frac{1}{2}\int_0^s\int_0^t R(s,s')C_y(s',t')R(t,t')\mathrm{d}t'\mathrm{d}s' \\ &- \int_0^s\int_0^{s'} R(s,s')R_y(s',t')\mathcal{K}_y(t,t')\mathrm{d}t'\mathrm{d}s' \\ &- m(t)\int_0^s R(s,s')m_y(s')\mathrm{d}s' + \frac{1}{2}\mathcal{K}_y(s,t) + Q_*m(s)m(t)\Bigg) + \mathrm{Sym},\end{aligned} \tag{124}$$

$$\mathcal{K}_y(t,t') = 2\int_0^t \int_0^s R(s,s')R(t,t')\Big(C_Z(s',t') - Q_*^{-1} m_Z(s')m_Z(t')\Big)\mathrm{d}s'\mathrm{d}t', \tag{125}$$

$$m(s) = \frac{1}{Q_*}\int_0^s R(s,s')m_Z(s')\mathrm{d}s', \tag{126}$$

$$\delta(t-t') = \int_{t'}^t R(t,s)R_Z(s,t')\mathrm{d}s. \tag{127}$$

The notation Sym in equation (124) indicates the symmetrization with respect to the variables s, t. The last equation defines R as the functional inverse of the response function R_Z. Although we do not prove that only one function R satisfies this relationship for all instants t, t', we believe that due to the causal structure of the responses, such a result should hold.

This set of equations is truly self-consistent, in the usual spirit of DMFT equations: the stochastic processes $W(t), y(t)$ are driven by scalar and deterministic functions that are themselves computed as expectations over these processes.

C.3 Derivation of the Equations

In this section we derive the set of equations presented in the previous section. We start from the dynamics in equation (5). Recalling the expression of the empirical loss in equation (4) and computing its gradient, the dynamics then writes:

$$\mathrm{d}W(t) = -\frac{d}{n}\sum_{k=1}^n \ell'\Big(\mathrm{Tr}\big(X_k W(t) W(t)^\top\big), z_k\Big) X_k W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t). \tag{128}$$

For simplicity we denote ℓ' the derivative of ℓ with respect to the first variable. We recall that z_k is generated by the teacher matrix Z^* through the noisy channel P:

$$z_k \sim P\Big(\,\cdot\,\Big|\,\mathrm{Tr}(X_k Z^*)\Big).$$

To keep the notation light, we derive the dynamical equations with $\Omega = 0$ and $\beta = \infty$, only keeping the first term in (128). In fact, since our calculation only transforms the empirical term, any additive term independent of the sensing matrices $(X_k)_{1\le k\le n}$ could be integrated in equation (128) and would appear in the same manner in the resulting dynamics.

Plan of the derivation. We derive our system of equations in several steps:

- We start by computing the dynamical partition function and average it with respect to the observations (Section C.3.1). The $d \to \infty$ limit allows to write a saddle-point equation on the covariance function of the typical label (Section C.3.2).
- We then study the response operator associated with the dynamics to simplify the structure of this covariance function (Section C.3.3).
- This simplification allows to write stochastic integro-differential equations for the evolution of the typical label (Section C.3.4) and student matrix (Section C.3.6).
- We link several averaged quantities to the coefficients driving the dynamics in order to close the equations (Section C.3.5).

The first step of this derivation is inspired by the replica calculation by Maillard et al. (2024). In our case, the integration over multiple replicas is replaced by one over the time instants of the dynamics. However, note several differences: in the Bayes-optimal setting, replicas are introduced in order to compute the expectation of the logarithm of the partition function. Since our derivation only requires averaging the partition function (rather than its logarithm) with respect to the observations, we avoid the non-rigorous step of sending the number of replicas to zero. In addition, while the work of Maillard et al. (2024) relies on a replica-symmetric ansatz (justified in the Bayes-optimal setting) to simplify the structure of the overlaps between replicas, our approach does not require such a simplification for the time-dependent overlaps that appear in the dynamics.

C.3.1 The Dynamical Partition Function

Following our introduction on path integrals in Section C.1, we write the dynamical partition function associated with the gradient flow dynamics on $[0, T]$ for some finite-time horizon T:

$$\mathcal{Z}_{\mathrm{dyn}} = \int \mathcal{D}W \, \delta\left(\dot{W}(t) + \frac{d}{n}\sum_{k=1}^{n} \ell'\big(y_k(t), z_k\big) X_k W(t)\right),$$

where $y_k(t) = \mathrm{Tr}\big(X_k W(t) W(t)^\top\big)$. The Dirac delta notation indicates here that the constraint must be verified for all $t \in [0, T]$. Then, using the Fourier representation of the Dirac delta function (as explained in Section C.1), we can rewrite:

$$\mathcal{Z}_{\mathrm{dyn}} \propto \int \mathcal{D}W \mathcal{D}\hat{W} \, \exp\bigg(- id \int_0^T \mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big)\mathrm{d}t - \frac{id^2}{n}\sum_{k=1}^{n}\int_0^T \ell'\big(y_k(t), z_k\big)\hat{y}_k(t)\mathrm{d}t\bigg),$$

with $\hat{y}_k(t) = \mathrm{Tr}\big(X_k W(t)\hat{W}(t)^\top\big)$ and $\propto$ indicates proportionality, up to a constant that may depend only on the dimension. The goal is now to compute the average of $\mathcal{Z}_{\mathrm{dyn}}$ with respect to the i.i.d. observations $X_1, \dots, X_n$. To do so, we write, when taking the expectation with respect to $X_1, \dots, X_n$ and $z_1, \dots, z_n$:

$$\mathbb{E}\,\exp\left(-\frac{id^2}{n}\sum_{k=1}^{n}\int_0^T \ell'\big(y_k(t), z_k\big)\hat{y}_k(t)\mathrm{d}t\right) = \left[\mathbb{E}_{X,z}\exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(t), z\big)\hat{y}(t)\mathrm{d}t\right)\right]^n.$$

Here $z \sim P\big(\,\cdot\,\big|\, y^*\big)$, and the variable $y(t)$ corresponds to the typical label introduced in Claim 2. We also used that $n \sim \alpha d^2$ in the large d limit. Since the variables $W(t), \hat{W}(t), Z^*$ are fixed, and $X \sim \mathrm{GOE}(d)$ (see Definition 26), the process $\mathbf{y}(t) = \big(y(t), \hat{y}(t), y^*\big)$ is Gaussian with zero mean and covariance:

$$Q(s,t) = \mathbb{E}\,\mathbf{y}(s)\mathbf{y}(t)^\top = \frac{2}{d}\begin{pmatrix} \mathrm{Tr}\, Z(s)Z(t) & \mathrm{Tr}\, Z(s)\hat{Z}(t) & \mathrm{Tr}\, Z(s)Z^* \\ \mathrm{Tr}\,\hat{Z}(s)Z(t) & \mathrm{Tr}\,\hat{Z}(s)\hat{Z}(t) & \mathrm{Tr}\,\hat{Z}(s)Z^* \\ \mathrm{Tr}\, Z^*Z(t) & \mathrm{Tr}\, Z^*\hat{Z}(t) & \mathrm{Tr}\, Z^{*2} \end{pmatrix}, \tag{129}$$

with $\hat{Z}(t) = \mathrm{Sym}\big(W(t)\hat{W}(t)^\top\big)$. Therefore, we can finally write the averaged partition function (with respect to the observations):

$$\overline{\mathcal{Z}}_{\mathrm{dyn}} \propto \int \mathcal{D}W\mathcal{D}\hat{W}\, \exp\left(-id\int_0^T \mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big)\mathrm{d}t\right) \times \left[\mathbb{E}_{\mathbf{y},z}\exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(t),z\big)\hat{y}(t)\mathrm{d}t\right)\right]^n.$$

Finally, we introduce the covariance matrix as an integration variable using a delta function, and obtain the identity:

$$\overline{\mathcal{Z}}_{\mathrm{dyn}} \propto \int \mathcal{D}W\mathcal{D}\hat{W}\mathcal{D}Q\mathcal{D}\hat{Q}\exp\left(-id^2\sum_{a,b=1}^3 \iint \hat{Q}_{ab}(s,t)\left(Q_{ab}(s,t) - \frac{2}{d}\mathrm{Tr}\, Z_a(s)Z_b(t)\right)\mathrm{d}s\mathrm{d}t\right) \times \exp\left(-id\int_0^T \mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big)\mathrm{d}t\right)\left[\mathbb{E}_{\mathbf{y},z}\exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(t),z\big)\hat{y}(t)\mathrm{d}t\right)\right]^n,$$

where $Z_1(t) = Z(t)$, $Z_2(t) = \hat{Z}(t)$ and $Z_3(t) = Z^*$. Rearranging the terms, we finally obtain the expression:

$$\overline{\mathcal{Z}}_{\mathrm{dyn}} = \int \mathcal{D}Q\mathcal{D}\hat{Q}\exp\left(-id^2\mathcal{T}(Q,\hat{Q}) + d^2\mathcal{F}(\hat{Q}) + \alpha d^2\mathcal{F}_{\mathrm{out}}(Q)\right), \tag{130}$$

with:

$$\mathcal{T}(Q,\hat{Q}) = \sum_{a,b=1}^3 \int_0^T\int_0^T \hat{Q}_{ab}(s,t)Q_{ab}(s,t)\mathrm{d}s\mathrm{d}t, \tag{131}$$

$$\mathcal{F}(\hat{Q}) = \frac{1}{d^2}\log\int \mathcal{D}W\mathcal{D}\hat{W}\, \exp\left(-id\int_0^T \mathrm{Tr}\,\dot{W}(t)\hat{W}(t)^\top\mathrm{d}t + 2id\sum_{a,b=1}^3\int_0^T\int_0^T \hat{Q}_{ab}(s,t)\mathrm{Tr}\big(Z_a(s)Z_b(t)\big)\mathrm{d}s\mathrm{d}t\right), \tag{132}$$

$$\mathcal{F}_{\mathrm{out}}(Q) = \log\mathbb{E}_{\mathbf{y}}\int \mathrm{d}z\, P\big(z\,\big|\,y^*\big)\exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(t),z\big)\hat{y}(t)\mathrm{d}t\right). \tag{133}$$

C.3.2 Saddle-Point Equations

Now, we can take the $d\to\infty$ limit in equation (130). Although W is still a high-dimensional object, the covariance functions $Q, \hat{Q}$ are finite-dimensional functions, when considered on timescales of order one. Then, the saddle-point asymptotics can be performed, and we obtain the equations:

$$i\hat{Q}(s,t) = \alpha\frac{\partial\mathcal{F}_{\mathrm{out}}(Q)}{\partial Q(s,t)}, \qquad iQ(s,t) = \frac{\partial\mathcal{F}(\hat{Q})}{\partial\hat{Q}(s,t)}. \tag{134}$$

Note that the second identity leads back to the definition of $Q(s,t)$ in equation (129), but averaged with respect to the dynamics.

The goal is now to compute the derivative of $\mathcal{F}_{\text{out}}$. To do so, recall that $\mathcal{F}_{\text{out}}$ depends on Q through the expectation with respect to $\mathbf{y}$ which is a Gaussian process with covariance Q. Therefore, we can write:

$$\mathcal{F}_{\text{out}}(Q) = \log \mathbb{E}_{\mathbf{y}\sim\mathcal{N}(0,Q)} \mathbb{E}_{z\sim P(\cdot\,|\,y^*)} \exp\Big(- i\Phi(y,\hat{y},z)\Big),$$
$$\Phi(y,\hat{y},z) = \frac{1}{\alpha}\int_0^T \ell'\big(y(t),z\big)\hat{y}(t)\mathrm{d}t.$$

To compute the derivative of $\mathcal{F}_{\text{out}}$, we shall now use Lemma 37 and the discussion that follows. As an application of the identity (112):

$$\frac{\partial \mathcal{F}_{\text{out}}(Q)}{\partial Q_{ij}(s,t)} = -\frac{1}{2}(Q^{-1})_{ij}(s,t) + \frac{1}{2}\sum_{a,b=1}^{3}\int_0^T\int_0^T (Q^{-1})_{ia}(s,s')(Q^{-1})_{jb}(t,t')\overline{\mathbb{E}}\big[y_a(s')y_b(t')\big]\mathrm{d}s'\mathrm{d}t', \tag{135}$$

where $y_1(t) = y(t), y_2(t) = \hat{y}(t), y_3(t) = y^*$, Q^{-1} is the functional inverse of Q and the expectation $\overline{\mathbb{E}}$ on $\mathbf{y}$ is computed as:

$$\overline{\mathbb{E}}\, f(\mathbf{y}) = \frac{\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(0,Q)}\mathbb{E}_{z\sim P(\cdot\,|\,y^*)} f(\mathbf{y}) e^{-i\Phi(y,\hat{y},z)}}{\mathbb{E}_{\mathbf{y}\sim\mathcal{N}(0,Q)}\mathbb{E}_{z\sim P(\cdot\,|\,y^*)} e^{-i\Phi(y,\hat{y},z)}}. \tag{136}$$

This distribution will correspond to the trajectory distribution of the typical label $y(t)$. Now, since this distribution involves the expectation over a Gaussian process with covariance function $Q(s,t)$, one needs to compute Q^{-1}. In the following, using the structure of the dynamics, we simplify Q and compute its inverse.

C.3.3 Form of the Covariance Function

We now give some arguments in order to simplify the form of the covariance $Q(s,t)$. Recall that the saddle-point equations led us to the same expression of Q as in equation (129), but averaged with respect to the distribution of the dynamics. To simplify this covariance, we study the general perturbed gradient flow:

$$\dot{W}(t) = -\nabla F(W(t)) + H(t)W(t),$$

with $H(t) \in \mathcal{S}_d(\mathbb{R})$, and $F\colon \mathbb{R}^{d\times m} \to \mathbb{R}$ is some continuously differentiable function. The dynamical partition function for this equation writes:

$$\mathcal{Z}_{\text{GF}} \propto \int \mathcal{D}W\mathcal{D}\hat{W} \exp\left(-id\int_0^T \mathrm{Tr}\Big[\Big(\dot{W}(t) + \nabla F(W(t)) - H(t)W(t)\Big)\hat{W}(t)^\top\Big]\mathrm{d}t\right). \tag{137}$$

Therefore, for any scalar function f of the dynamics, we have the identities, when averaging with respect to the distribution associated with $\mathcal{Z}_{\text{GF}}$:

$$\left.\frac{\partial\, \mathbb{E}\, f(W)}{\partial H(t)}\right|_{H=0} = id\,\mathbb{E}\big[f(W)\hat{Z}(t)\big],$$
$$\left.\frac{\partial^2\, \mathbb{E}\, f(W)}{\partial H(t)\partial H(t')}\right|_{H=0} = -d^2\,\mathbb{E}\Big[f(W)\big(\hat{Z}(t)\otimes\hat{Z}(t')\big)\Big].$$

Here we view $\mathbb{E}\, f(W)$ as a function of the perturbation $\big(H(t)\big)_{t\geq 0}$ on $\mathcal{S}_d(\mathbb{R})$, and we compute the first and second derivatives of this function in the space $\mathcal{S}_d(\mathbb{R})$ at $H = 0$. To prove these identities, one can write $\mathbb{E}\, f(W)$ as a function of H with the partition function (137) and use that the gradient of the map $H \mapsto \mathrm{Tr}(HM)$ on $\mathcal{S}_d(\mathbb{R})$ is $\mathrm{Sym}(M)$ (see Srinivasan and Panda, 2023, for the notion of symmetric gradient). We recall that $\hat{Z}(t) = \mathrm{Sym}\big(W(t)\hat{W}(t)^\top\big)$. In the second derivative, the notation $\otimes$ refers to the tensor product.

Therefore, the matrix $\hat{Z}(t)$ acts as a derivative operator when averaging quantities with respect to the dynamics. Thus, from the previous identities, we deduce that:

$$\mathbb{E}\,\mathrm{Tr}\big(Z^*\hat{Z}(t)\big) = 0, \qquad\qquad \mathbb{E}\,\mathrm{Tr}\big(\hat{Z}(s)\hat{Z}(t)\big) = 0.$$

Indeed, the first quantity can be rewritten as the derivative of Z^* in response to a perturbation of the dynamics, and the second one corresponds to the second-order derivative of a constant function. In addition, we obtain:

$$\mathbb{E}\,\mathrm{Tr}\big(Z(s)\hat{Z}(t)\big) = -\frac{i}{d}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(s)}{\partial H(t)}\bigg|_{H=0}\right).$$

Here the response operator can be interpreted as the differential of the function $\mathcal{S}_d(\mathbb{R}) \to \mathcal{S}_d(\mathbb{R})$ that maps the perturbation H introduced at time t to the perturbed solution $\mathbb{E}\, Z(s)$ at time s. This term is zero when $t > s$. Indeed, the dynamics cannot be influenced by a perturbation added at a later time. Finally, we can conclude that the covariance function $Q(s,t)$ is of the form:

$$Q(s,t) = 2\begin{pmatrix} C_Z(s,t) & -iR_Z(s,t) & m_Z(s) \\ -iR_Z(t,s) & 0 & 0 \\ m_Z(t) & 0 & Q_* \end{pmatrix},$$

with:

$$\begin{aligned} C_Z(s,t) &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(s)Z(t)\big), & R_Z(s,t) &= \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(s)}{\partial H(t)}\bigg|_{H=0}\right), \\ m_Z(t) &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z^*\big), & Q_* &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}). \end{aligned} \tag{138}$$

From this expression of Q, we can deduce the structure of its inverse by applying Lemma 38. Although this lemma is formulated with discrete variables, it is easily seen that we can adapt the result when dealing with two-time functions. In this case, the matrix inverse is replaced by the functional inverse, and the standard matrix product now corresponds to integration with respect to time variables. Then, we can conclude that the function Q^{-1} appearing in equation (135) has the form:

$$Q^{-1}(s,t) = \frac{1}{2}\begin{pmatrix} 0 & iR(t,s) & 0 \\ iR(s,t) & K(s,t) & -im(s) \\ 0 & -im(t) & Q_*^{-1} \end{pmatrix},$$

and we have the relationships, from Lemma 38:

$$\begin{aligned} R(s,t) &= R_Z^{-1}(s,t), \\ K(s,t) &= \int_0^t \int_0^s R(s,s')R(t,t')\Big(C_Z(s',t') - Q_*^{-1} m_Z(s')m_Z(t')\Big)\mathrm{d}s'\mathrm{d}t', \\ m(s) &= \frac{1}{Q_*}\int_0^s R(s,s')m_Z(s')\mathrm{d}s'. \end{aligned} \tag{139}$$

R is defined as the inverse of R_Z, therefore also has a causal structure, i.e., $R(t,t') = 0$ for $t' > t$, and we can deduce the relationship for all $0 \leq t' \leq t \leq T$:

$$\int_{t'}^t R(t,s)R_Z(s,t')\mathrm{d}s = \delta(t-t').$$

C.3.4 Evolution of the Labels

Before deriving the self-consistent set of equations, recall that equation (135) involves averages of the labels with respect to a reweighted distribution, defined in equation (136). We investigate this distribution and derive the evolution equation for the label $y(t)$. To do so, we use the fact that $\mathbf{y}$ is a Gaussian process with covariance Q and we compute:

$$\begin{aligned} \overline{\mathbb{E}}\, f(y,y^*) \propto \int \mathcal{D}y\mathcal{D}\hat{y}\mathcal{D}y^* f(y,y^*) \exp\left(-\frac{1}{2}\int_0^T\int_0^T \mathbf{y}(s)^\top Q^{-1}(s,t)\mathbf{y}(t)\mathrm{d}t\mathrm{d}s\right) \\ \times \mathbb{E}_z \exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(s),z\big)\hat{y}(s)\mathrm{d}s\right). \end{aligned}$$

Using the expression of Q^{-1}, we have:

$$\begin{aligned} \int_0^T\int_0^T \mathbf{y}(s)^\top Q^{-1}(s,t)\mathbf{y}(t)\mathrm{d}t\mathrm{d}s = \frac{1}{2Q_*}y^{*2} + i\int_0^T\int_0^s R(s,t)y(t)\hat{y}(s)\mathrm{d}t\mathrm{d}s \\ - iy^*\int_0^T m(s)\hat{y}(s)\mathrm{d}s + \frac{1}{2}\int_0^T\int_0^T K(s,t)\hat{y}(s)\hat{y}(t)\mathrm{d}t\mathrm{d}s. \end{aligned}$$

We now use the identity:

$$\exp\left(-\frac{1}{4}\int_0^T\int_0^T K(s,t)\hat{y}(s)\hat{y}(t)\mathrm{d}t\mathrm{d}s\right) = \mathbb{E}_\eta \exp\left(-\frac{i}{2}\int_0^T \eta(s)\hat{y}(s)\mathrm{d}s\right),$$

where $\eta(t)$ is a Gaussian process with zero mean and covariance $2K(s,t)$. Finally, we end up with:

$$\begin{aligned} \overline{\mathbb{E}}\, f(y,y^*) \propto \int \mathcal{D}y\mathcal{D}\hat{y}\mathcal{D}y^* f(y,y^*)\mathbb{E}_{z,\eta}\exp\left(-\frac{1}{4Q_*}y^{*2}\right) \\ \exp\left(\frac{i}{2}\int_0^T\left[-\int_0^s R(s,t)y(t)\mathrm{d}t + m(s)y^* - \eta(s) - \frac{2}{\alpha}\ell'\big(y(s),z\big)\right]\hat{y}(s)\mathrm{d}s\right). \end{aligned} \tag{140}$$

Therefore, integrating with respect to the variable $\hat{y}$, and using the exponential representation of the delta Dirac function (see Section C.1), we end up with the equation for $y(t)$:

$$\int_0^t R(t,t')y(t')\mathrm{d}t' + \eta(t) - m(t)y^* + \frac{2}{\alpha}\ell'\big(y(t),z\big) = 0, \qquad \mathbb{E}\,\eta(s)\eta(t) = 2K(s,t), \tag{141}$$

where $z \sim P(\cdot \,|\, y^*)$. Finally, we determine the form of the covariances for the y variables that appear in equation (135). Following the same arguments as for the covariance Q, we can show that:

$$\overline{\mathbb{E}}\, \hat{y}(s) y^* = 0, \qquad \overline{\mathbb{E}}\, \hat{y}(s)\hat{y}(t) = 0, \qquad \overline{\mathbb{E}}\, y(s)\hat{y}(t) = -2iR_y(s,t),$$

where R_y is a response function and is non-zero only for $t \in [0, s]$. Similarly to the response identities we derived for $W(t)$, we can prove, using the distribution in equation (140), that:

$$R_y(s,t) = \left.\frac{\partial \overline{\mathbb{E}}\, y(s)}{\partial h(t)}\right|_{h=0},$$

in response to a perturbation of the noise $\eta(t) \mapsto \eta(t) - h(t)$ in equation (141). In addition, equation (140) shows that $\overline{\mathbb{E}}\, y^{*2} = 2Q_*$. Finally, we denote the covariances:

$$C_y(s,t) = \overline{\mathbb{E}}\, y(s)y(t), \qquad m_y(t) = \overline{\mathbb{E}}\, y(t)y^*.$$

With all of these quantities, we are now ready to go back to equation (135) and derive the set of self-consistent equations.

C.3.5 Equations between Averaged Quantities

Back to equation (135), we can compute the derivative of $\mathcal{F}_{\text{out}}$ with respect to Q, and with the saddle-point equation (134), show that $\hat{Q}$ is of the form:

$$\hat{Q}(s,t) = \frac{1}{4}\begin{pmatrix} 0 & -\Gamma(t,s) & 0 \\ -\Gamma(s,t) & i\mathcal{K}_Z(s,t) & r(t) \\ 0 & r(s) & 0 \end{pmatrix}, \tag{142}$$

and that we have the relationships:

$$\Gamma(s,t) = \alpha\left(R(s,t) - \int_t^s \int_t^{s'} R(s,s')R_y(s',t')R(t',t)\mathrm{d}t'\mathrm{d}s'\right), \tag{143}$$

$$\begin{aligned}\mathcal{K}_Z(s,t) = \alpha\Bigg(&\frac{1}{4}\int_0^s \int_0^t R(s,s')C_y(s',t')R(t,t')\mathrm{d}t'\mathrm{d}s' \\ &- \int_0^s \int_0^{s'} R(s,s')R_y(s',t')K(t,t')\mathrm{d}t'\mathrm{d}s' \\ &- \frac{1}{2}m(t)\int_0^s R(s,s')m_y(s')\mathrm{d}s' + \frac{1}{2}K(s,t) + \frac{1}{2}Q_* m(s)m(t)\Bigg) + \text{Sym},\end{aligned}$$

$$r(s) = \alpha\left(\frac{1}{2Q_*}\int_0^s R(s,s')m_y(s')\mathrm{d}s' - \int_0^s \int_0^{s'} R(s,s')R_y(s',t')m(t')\mathrm{d}t'\mathrm{d}s'\right).$$

Sym indicates the symmetrization of the previous expression with respect to the time variables s, t. These equations relate the coefficients of $\hat{Q}$ to the quantities derived from Q (namely the functions R, K, m that are the coefficients of Q^{-1}), and the averaged quantities of the dynamics of $y(t)$, themselves driven by the functions R, K, m.

C.3.6 Evolution of the Weights

To close the system of self-consistent equations, we finally need another relationship between $\hat{Q}$ and Q. To do so, recall that the Q variables are related to averages with respect to the dynamics of W. Then, from the expression of the dynamical partition function in the high-dimensional limit, we can derive equivalent dynamical equations for the weights $W(t)$. To do so, we go back to the expression of $\mathcal{F}$ in equation (132), along with the expression of $\hat{Q}$ in equation (142):

$$\begin{aligned}\mathcal{F}(\hat{Q}) = \frac{1}{d^2}\log\int \mathcal{D}W\mathcal{D}\hat{W}\exp\Bigg(&-id\int_0^T \mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big)\mathrm{d}t\\ &-id\int_0^T\int_0^t \Gamma(t,t')\mathrm{Tr}\big(\hat{Z}(t)Z(t')\big)\mathrm{d}t'\mathrm{d}t\\ &+id\int_0^T r(t)\mathrm{Tr}\big(\hat{Z}(t)Z^*\big)\mathrm{d}t\\ &-\frac{d}{2}\int_0^T\int_0^T \mathcal{K}_Z(s,t)\mathrm{Tr}\big(\hat{Z}(s)\hat{Z}(t)\big)\mathrm{d}s\mathrm{d}t\Bigg).\end{aligned}$$

We now use the identity:

$$\exp\left(-\frac{d}{2}\int_0^T\int_0^T \mathcal{K}_Z(s,t)\mathrm{Tr}\big(\hat{Z}(s)\hat{Z}(t)\big)\mathrm{d}s\mathrm{d}t\right) = \mathbb{E}_V\exp\left(id\int_0^T \mathrm{Tr}\big(\hat{Z}(s)V(s)\big)\mathrm{d}s\right),$$

where $V(t)\in\mathbb{R}^{d\times d}$ is a centered Gaussian matrix with covariance:

$$\mathbb{E}\,V_{ij}(s)V_{i'j'}(t) = \frac{1}{d}\delta_{ii'}\delta_{jj'}\mathcal{K}_Z(s,t).$$

Therefore, using that $\hat{Z}(t) = \mathrm{Sym}\big(W(t)\hat{W}(t)^\top\big)$, we have:

$$\begin{aligned}\mathcal{F}(\hat{Q}) = \frac{1}{d^2}\log\int \mathcal{D}W\mathcal{D}\hat{W}\exp\Bigg(&-id\int \mathrm{d}t\,\mathrm{Tr}\Bigg[\dot{W}(t)+\int_0^t\Gamma(t,t')Z(t')\mathrm{d}t'\,W(t)\\ &-r(t)Z^*W(t)-\mathrm{Sym}\big(V(t)\big)W(t)\Bigg]\hat{W}(t)^\top\Bigg).\end{aligned}$$

We can integrate with respect to $\hat{W}(t)$ and use the exponential formulation of the delta functional in Section C.1. This leads to the following equation for $W(t)$:

$$\dot{W}(t) = \left(\mathrm{Sym}\big(V(t)\big) + r(t)Z^* - \int_0^t \Gamma(t,t')Z(t')\mathrm{d}t'\right)W(t). \tag{144}$$

Now setting $\mathcal{H}(t) = \mathrm{Sym}(V(t))$, we get that $\mathcal{H}$ is still a centered Gaussian process with covariance:

$$\mathbb{E}\,\mathcal{H}_{ij}(s)\mathcal{H}_{i'j'}(t) = \frac{1}{2d}\Big(\delta_{ii'}\delta_{jj'}+\delta_{ij'}\delta_{i'j}\Big)\mathcal{K}_Z(s,t).$$

As mentioned earlier, it is easily seen that we can add back the regularization and the Brownian motion from (128) into equation (144). Then, gathering equations (138), (139), (141), (143), with the above dynamics on $W(t)$, and simply writing $\mathcal{K}_y = 2K$, we obtain the equations of Section C.2.

C.4 Simplified Set of Equations

We now prove the simplified set of equations in Claim 4 under Assumption 3. More precisely, we take ℓ to be the quadratic cost and assume that the labels are generated using a Gaussian noisy channel. With these assumptions, we simplify the previous set of equations. The plan is the following:

- We identify the label evolution as a Gaussian dynamics and transform the equation so that it is driven by quantities associated with $W(t)$. This allows to write a system of equations on the variables R_y, m_y (Section C.4.1).
- We then use these simplifications to eliminate the variables $r, \Gamma, \mathcal{K}_Z$ (Section C.4.2).
- We finally rewrite the response function R_Z and obtain the desired set of equations (Section C.4.3).

C.4.1 Label Equation

We proceed to simplify the label equation (141). Using Assumption 3 on the noisy channel, we write $z = y^* + \sqrt{\Delta}\zeta$ where $\zeta \sim \mathcal{N}(0,1)$, we get the equation:

$$\int_0^t R(t,t')y(t')\mathrm{d}t' + \eta(t) - m(t)y^* + \frac{2}{\alpha}\big(y(t) - y^* - \sqrt{\Delta}\zeta\big) = 0. \tag{145}$$

Remark now that $y(t)$ is a centered Gaussian process. Before deriving its covariance, we apply the linear time transformation R_Z (equal to R^{-1} from equation 139) to equation (145), to get:

$$y(t) + \underbrace{\int_0^t R_Z(t,t')\eta(t')\mathrm{d}t'}_{\equiv \xi(t)} - \int_0^t R_Z(t,t')m(t')\mathrm{d}t'\, y^* + \frac{2}{\alpha}\int_0^t R_Z(t,t')\big(y(t') - y^* - \sqrt{\Delta}\zeta\big)\mathrm{d}t' = 0.$$

Now, from the expression of m in equation (139), and the fact that the covariance of η is K, written in equation (139), we have the identities:

$$\int_0^t R_Z(t,t')m(t')\mathrm{d}t' = \frac{1}{Q_*}m_Z(t),$$
$$\mathbb{E}\,\xi(t)\xi(t') = 2C_Z(t,t') - \frac{2}{Q_*}m_Z(t)m_Z(t').$$

This transformation does not change the expression of $y(t)$, nor those of m_y, C_y. However, we should recompute the response with respect to a perturbation of the new dynamics. We then consider R_y^{old} to be the response associated with equation (145) and R_y^{new} the response after the transformation by R_Z. In addition, since the responses are defined as additive perturbations of the noise (with a minus sign in this case), we simply have:

$$R_y^{\mathrm{old}}(t,t') = -\frac{\partial\,\mathbb{E}\,y(t)}{\partial\eta(t')}, \qquad R_y^{\mathrm{new}}(t,t') = -\frac{\partial\,\mathbb{E}\,y(t)}{\partial\xi(t')}, \tag{146}$$

and due to the relationship between $\xi(t), \eta(t)$, we get:

$$R_y^{\text{old}}(t,t') = \int_{t'}^{t} R_y^{\text{new}}(t,t'')R_Z(t'',t')\mathrm{d}t''. \tag{147}$$

From now on, we denote $R_y = R_y^{\text{new}}$. This leads to the equation on y:

$$y(t) + \xi(t) - \frac{1}{Q_*}m_Z(t)y^* + \frac{2}{\alpha}\int_0^t R_Z(t,t')\big(y(t') - y^* - \sqrt{\Delta}\zeta\big)\mathrm{d}t' = 0. \tag{148}$$

From this stochastic evolution for $y(t)$, we can obtain equations for the deterministic functions R_y, m_y. This is possible since y is explicitly written as a sum of independent Gaussian processes. Starting with the response, we use the previous identity and differentiate the new equation on $y(t)$ with respect to $\xi(t)$. This leads to:

$$\delta(t-t') = R_y(t,t') + \frac{2}{\alpha}\int_{t'}^{t} R_Z(t,t'')R_y(t'',t')\mathrm{d}t''. \tag{149}$$

This implies that we have the identity, from equation (148):

$$y(t) = y^* + \sqrt{\Delta}\zeta + \int_0^t R_y(t,t')\Big(\xi(t') + \big(\chi_Z(t') - 1\big)y^* - \sqrt{\Delta}\zeta\Big)\mathrm{d}t', \tag{150}$$

with $\chi_Z(t) = m_Z(t)/Q_*$. Here we have transformed ξ to $-\xi$, which leaves unchanged the statistics of this process. However, similarly to equation (146), R_y is now computed as the derivative of $\mathbb{E}\, y$ with respect to ξ. In order to arrive at equation (24), the only step remaining will be to reexpress R_y.

Finally, the expression of $y(t)$ in equation (150) leads to the identity on m_y:

$$\frac{1}{2Q_*}m_y(t) = 1 - \int_0^t R_y(t,t')\big(1 - \chi_Z(t')\big)\mathrm{d}t'. \tag{151}$$

C.4.2 Hat Variables

We now compute the variables $\Gamma, \mathcal{K}_Z, r$, which are the coefficients of the conjugate matrix $\hat{Q}$. To simplify notation, we consider one-time and two-time functions as vectors and matrices, and replace integration by matrix and vector products. We start from equation (143), and recall that in these equations the response corresponds to R_y^{old}, that we should replace using equation (147). Therefore, the system of equations (143) writes, in compact notations:

$$\begin{aligned}
\Gamma &= \alpha(R - RR_y), \\
\mathcal{K}_Z &= \alpha\bigg(\frac{1}{2}RC_yR^\top - RR_yR_ZK - KR_Z^\top R_y^\top R^\top \\
&\qquad\qquad - \frac{1}{2}Rm_ym^\top - \frac{1}{2}mm_y^\top R^\top + K + Q_*mm^\top\bigg), \\
r &= \alpha\left(\frac{1}{2Q_*}Rm_y - RR_yR_Zm\right).
\end{aligned} \tag{152}$$

Now, from equation (149) on R_y, we simply have $\Gamma = 2R_y$. Now using the expression of m_y in equation (151), as well as the expression of m in equation (139), we easily obtain that $r = \Gamma \mathbf{1}$, that is:

$$r(t) = \int_0^t \Gamma(t, t') \mathrm{d}t'.$$

Remains to compute the covariance $\mathcal{K}_Z$. To do so, let us now write the covariance C_y, that we can compute from equation (150). In this equation, y is written as a sum of the three independent Gaussians y^*, ζ, ξ. Therefore, we have:

$$\begin{aligned} C_y = 2Q_* \big(\mathbf{1} + R_y(\chi_Z - \mathbf{1})\big)\big(\mathbf{1} + R_y(\chi_Z - \mathbf{1})\big)^\top + \Delta \big(\mathbf{1} - R_y \mathbf{1}\big)\big(\mathbf{1} - R_y \mathbf{1}\big)^\top \\ + 2R_y \big(C_Z - Q_*^{-1} m_Z m_Z^\top\big) R_y^\top. \end{aligned}$$

Since we have the identities, from equations (151), (149) and the definition of K in equation (139):

$$\begin{aligned} \mathbf{1} + R_y(\chi_Z - \mathbf{1}) &= \frac{1}{2Q_*} m_y, \\ \mathbf{1} - R_y \mathbf{1} &= \frac{2}{\alpha} R_Z R_y \mathbf{1}, \\ C_Z - Q_*^{-1} m_Z m_Z^\top &= R_Z K R_Z^\top, \end{aligned}$$

we get the expression:

$$R C_y R^\top = \frac{1}{2Q_*} R m_y m_y^\top R^\top + \frac{4\Delta}{\alpha^2} R_y \mathbf{1}\mathbf{1}^\top R_y^\top + 2R R_y R_Z K R_Z^\top R_y^\top R^\top.$$

Then, using the expression of m in equation (139) and rearranging the expression of $\mathcal{K}_Z$ in equation (152):

$$\begin{aligned} \frac{1}{\alpha} \mathcal{K}_Z = \big(I - R R_y R_Z\big) K \big(I - R_Z^\top R_y^\top R^\top\big) \\ + \frac{1}{4Q_*} R(2m_Z - m_y)(2m_Z - m_y)^\top R^\top + \frac{2\Delta}{\alpha^2} R_y \mathbf{1}\mathbf{1}^\top R_y^\top. \end{aligned} \tag{153}$$

Now, with the fact that $R R_Z = I$, we get by multiplying equation (149) by R on the left and R_Z on the right:

$$I - R R_y R_Z = \frac{2}{\alpha} R_y R_Z.$$

Therefore, back to the definition of K in equation (139):

$$\big(I - R R_y R_Z\big) K \big(I - R_Z^\top R_y^\top R^\top\big) = \frac{4}{\alpha^2} R_y \big(C_Z - Q_*^{-1} m_Z m_Z^\top\big) R_y^\top.$$

This gives the expression of the first term of equation (153). For the second, we use the expression of m_y in equation (151) to get:

$$2m_Z - m_y = 2(I - R_y)(m_Z - Q_* \mathbf{1}).$$

Therefore, since $\alpha(R - R R_y) = 2R_y$, we get:

$$\frac{1}{4Q_*} R(2m_Z - m_y)(2m_Z - m_y)^\top R^\top = \frac{4}{\alpha^2 Q_*} R_y (m_Z - Q_* \mathbf{1})(m_Z - Q_* \mathbf{1})^\top R_y^\top.$$

Putting everything together, we arrive at the expression for $\mathcal{K}_Z$:

$$\mathcal{K}_Z = \frac{4}{\alpha} R_y \left(C_Z - \mathbf{1} m_Z^\top - m_Z \mathbf{1}^\top + Q_* \mathbf{1}\mathbf{1}^\top + \frac{\Delta}{2} \mathbf{1}\mathbf{1}^\top \right) R_y^\top .$$

Due to the expression of C_Z, m_Z, Q_*, we get the expression:

$$\mathcal{K}_Z(s,t) = \frac{4}{\alpha} \int_0^s \int_0^t R_y(s,s') R_y(t,t') \left(\frac{1}{d} \mathbb{E} \operatorname{Tr} \Big[\big(Z(s') - Z^*\big) \big(Z(t') - Z^*\big) \Big] + \frac{\Delta}{2} \right) \mathrm{d}t' \mathrm{d}s'.$$

Now recall that $\mathcal{K}_Z$ is linked to the covariance of the Gaussian process $\mathcal{H}$ in equation (117). Due to the integral structure of $\mathcal{K}_Z$, we can write:

$$\mathcal{H}(t) = 2 \int_0^t R_y(t,t') \mathcal{G}(t') \mathrm{d}t',$$

where $\mathcal{G}$ is also a centered Gaussian process taking values in the space of symmetric matrices, and whose covariance is given by:

$$\mathbb{E}\, \mathcal{G}_{ij}(t) \mathcal{G}_{i'j'}(t') = \frac{1}{2\alpha d} \big(\delta_{ii'} \delta_{jj'} + \delta_{ij'} \delta_{i'j} \big) \left(\frac{1}{d} \mathbb{E} \operatorname{Tr} \Big[\big(Z(t) - Z^*\big) \big(Z(t') - Z^*\big) \Big] + \frac{\Delta}{2} \right).$$

Now, this is precisely the same covariance as in equation (25). Putting everything together, equation (115) can be rewritten as:

$$\mathrm{d}W(t) = 2 \left(\int_0^t R_y(t,t') \Big(\mathcal{G}(t') + Z^* - Z(t') \Big) \mathrm{d}t' \right) W(t) \mathrm{d}t - \nabla \Omega\big(W(t)\big) \mathrm{d}t + \frac{1}{\sqrt{\beta d}} \mathrm{d}B(t). \tag{154}$$

C.4.3 Response Function

The last step of the simplification is to rewrite the responses R_y, R_Z. At the moment we have the relationships:

$$R_Z(t,t') = \frac{1}{d^2} \operatorname{Tr} \left(\frac{\partial\, \mathbb{E}\, Z(t)}{\partial H(t')} \bigg|_{H=0} \right), \tag{155}$$

$$\delta(t-t') = R_y(t,t') + \frac{2}{\alpha} \int_{t'}^t \mathrm{d}t''\, R_Z(t,t'') R_y(t'',t'), \tag{156}$$

and $H(t) \in \mathcal{S}_d(\mathbb{R})$ perturbs the equation for $W(t)$ as:

$$\mathrm{d}W(t) = 2 \left(\int_0^t R_y(t,t') \Big(\mathcal{G}(t') + Z^* - Z(t') \Big) \mathrm{d}t' \right) W(t) \mathrm{d}t + H(t) W(t) \mathrm{d}t + \dots,$$

where the dots include the other terms of equation (154). We now consider the following perturbation:

$$H(t) = 2 \int_0^t R_y(t,t') \tilde{H}(t') \mathrm{d}t',$$

so that $\tilde{H}(t')$ appears as an additive perturbation of the noise $\mathcal{G}(t)$ into the equation for W. The response associated with this perturbation writes:

$$\frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial Z(t)}{\partial \tilde{H}(t')}\right) = \frac{2}{d^2}\int_0^t \mathrm{Tr}\left(\frac{\partial Z(t)}{\partial H(t'')}\right) R_y(t'',t')\mathbf{1}_{t''\geq t'}\mathrm{d}t''.$$

Therefore, when averaging, we obtain the response associated with $\tilde{H}$:

$$\begin{aligned}\tilde{R}_Z(t,t') &\equiv \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\, \mathbb{E}\, Z(t)}{\partial \tilde{H}(t')}\Big|_{H=0}\right)\\ &= 2\int_{t'}^t \mathrm{d}t''\, R_Z(t,t'')R_y(t'',t')\\ &= \alpha\big(\delta(t-t') - R_y(t,t')\big),\end{aligned}$$

where we used the relationship between R_Z, R_y in equation (156). This means that we can replace R_y using the previous equation, which decouples the equation for $W(t)$ from the one on $y(t)$. We finally arrive at our system of equations by setting $R(t,t') \equiv R_y(t,t') = \delta(t-t') - \frac{1}{\alpha}\tilde{R}_Z(t,t')$ into equations (150) and (154).

C.5 Learning the Second Layer

In this section, following from Section 3.1.2, we derive a similar set of equations when also optimizing the output weights of the neural network.

C.5.1 Gradient Flow Dynamics

With the notations of Section 2, we are now interested in the predictor:

$$X \mapsto \mathrm{Tr}\big(XWD_aW^\top\big) = \frac{1}{m}\sum_{i=1}^m a_i \mathrm{Tr}\big(Xw_iw_i^\top\big), \qquad W = \frac{1}{\sqrt{m}}\Big(w_1\,\Big|\,\dots\,\Big|\,w_m\Big),$$

with $a = \big(a_1,\dots,a_m\big)^\top \in \mathbb{R}^m$ and D_a is the diagonal matrix with the same coefficients as a. We then consider the loss function:

$$\mathcal{L}(a,W) = \frac{1}{2n}\sum_{k=1}^n \ell\big(\mathrm{Tr}\big(X_kWD_aW^\top\big), z_k\big),$$

where z_k is still drawn from the distribution $z_k \sim P\big(\,\cdot\,\big|\,\mathrm{Tr}(X_kZ^*)\big)$. For this derivation, the teacher matrix can remain as is, but we can think of it as $Z^* = W^*D_{a^*}W^{*\top}$ for $W^* \in \mathbb{R}^{d\times m^*}$ and $a^* \in \mathbb{R}^{m^*}$. We are then interested in the gradient flow dynamics:

$$\begin{aligned}\dot{a}(t) = -\vartheta d\,\nabla_a\mathcal{L}\big(a(t),W(t)\big) \quad &= -\frac{\vartheta d}{2n}\sum_{k=1}^n \ell'\big(y_k(t),z_k\big)\mathrm{diag}\big(W(t)^\top X_kW(t)\big),\\ \dot{W}(t) = -d\,\nabla_W\mathcal{L}\big(a(t),W(t)\big) \quad &= -\frac{d}{n}\sum_{k=1}^n \ell'\big(y_k(t),z_k\big)X_kW(t)D_{a(t)},\end{aligned}$$

where $y_k(t) = \mathrm{Tr}\big(X_k W(t) D_{a(t)} W(t)^\top\big)$, $\mathrm{diag}(A)$ is the vector composed of the diagonal elements of A, and $\vartheta > 0$ is a parameter that allows for different learning speed for $a(t)$ and $W(t)$. Note that the dynamics considered in Section 3.1.2 includes regularization and thermal noise terms, but those remain unchanged in the resulting dynamics as it is still formulated in terms of the variables $a(t), W(t)$.

C.5.2 Derivation of the Equations

We shall now derive the set of equations given in Section 3.1.2.

The dynamical partition function. Writing the dynamical partition function similarly to what was done in Section C.3, we get by introducing conjugate variables $\hat{W}(t), \hat{a}(t)$:

$$\begin{aligned}\mathcal{Z}_{\mathrm{dyn}} \propto \int \mathcal{D}W\mathcal{D}\hat{W}\mathcal{D}a\mathcal{D}\hat{a}\, \exp\left(-id\int_0^T \Big(\mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big) + \dot{a}(t)^\top \hat{a}(t)\Big)\mathrm{d}t\right) \\ \times \exp\left(-\frac{id^2}{n}\sum_{k=1}^n \int_0^T \ell'\big(y_k(t), z_k\big)\Big[\mathrm{Tr}\big(X_k W(t) D_{a(t)} \hat{W}(t)^\top\big) \\ + \frac{\vartheta}{2}\mathrm{Tr}\big(X_k W(t) D_{\hat{a}(t)} W(t)^\top\big)\Big]\mathrm{d}t\right).\end{aligned} \tag{157}$$

We then define:

$$\hat{Z}(t) = \mathrm{Sym}\left(W(t)D_{a(t)}\hat{W}(t)^\top + \frac{\vartheta}{2}W(t)D_{\hat{a}(t)}W(t)^\top\right), \qquad \hat{y}_k(t) = \mathrm{Tr}\big(X_k\hat{Z}(t)\big), \tag{158}$$

and find ourselves in the same setup as the is Section C.3 when considering this $\hat{Z}(t)$ and $Z(t) = W(t)D_{a(t)}W(t)^\top$. The only difference is the extra term involving $\dot{a}(t)$ in the dynamical partition function.

Response structure. In Section C.3, we studied the structure of the overlap matrix when considering perturbed dynamics solely on $W(t)$ (see equation 137). Here, we study a similar generic dynamics taking the form of a perturbed gradient flow associated with a function of $Z = WD_aW^\top$:

$$\begin{aligned}\dot{a}(t) &= -\frac{\vartheta}{2}\mathrm{diag}\Big(W(t)^\top\Big(\nabla F\big(Z(t)\big) + H(t)\Big)W(t)\Big), \\ \dot{W}(t) &= -\Big(\nabla F\big(Z(t)\big) + H(t)\Big)W(t)D_{a(t)},\end{aligned} \tag{159}$$

where $H(t) \in \mathcal{S}_d(\mathbb{R})$ is to be considered as a perturbation of the dynamics. Then, we can write the dynamical partition function in a similar fashion to equation (137). Only showing the terms associated with the perturbation, we have:

$$\begin{aligned}\mathcal{Z}_{\mathrm{GF}} &\propto \int \mathcal{D}W\mathcal{D}\hat{W}\mathcal{D}a\mathcal{D}\hat{a}\, \exp\left(id\int_0^T \Big[\mathrm{Tr}\big(H(t)W(t)D_{a(t)}\hat{W}(t)^\top\big) \right. \\ &\qquad \left. + \frac{\vartheta}{2}\hat{a}(t)^\top \mathrm{diag}\big(W(t)^\top H(t) W(t)\big)\Big]\mathrm{d}t\right) \\ &= \int \mathcal{D}W\mathcal{D}\hat{W}\mathcal{D}a\mathcal{D}\hat{a}\, \exp\left(id\int_0^T \mathrm{Tr}\big(H(t)\hat{Z}(t)\big)\mathrm{d}t\right),\end{aligned}$$

where $\hat{Z}(t)$ is defined in equation (158). Then, as previously done, we can show that differentiating averages of a function of the dynamics (159) with respect to $H(t')$ is akin to multiplying by $\hat{Z}(t')$ and take the average. In particular, we find the same structure of the overlap matrix as in Section C.3 and we have the response identity:

$$\mathbb{E}\,\mathrm{Tr}\big(Z(s)\hat{Z}(t)\big) = -\frac{i}{d}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(s)}{\partial H(t)}\Big|_{H=0}\right),$$

where $Z(t) = W(t)D_{a(t)}W(t)^\top$ and $H(t)$ is introduced as in the dynamics (159).
End of the calculation. Once the response is computed, the exact same calculation can be carried out as in Section C.3. In the end, to derive the dynamical equations on $a(t), W(t)$, we obtain a function that is similar to the one in equation (132):

$$\begin{aligned}\mathcal{F} = \frac{1}{d^2}\log\int \mathcal{D}W\mathcal{D}\hat{W}\mathcal{D}a\mathcal{D}\hat{a}\exp\Bigg(-id\int_0^T\Big[&\mathrm{Tr}\big(\dot{W}(t)\hat{W}(t)^\top\big) + \dot{a}(t)^\top\hat{a}(t)\\ &+\int_0^t\Gamma(t,t')\mathrm{Tr}\big(Z(t')\hat{Z}(t)\big)\mathrm{d}t'\\ &-r(t)\mathrm{Tr}\big(Z^*\hat{Z}(t)\big) - \mathrm{Tr}\Big(\mathrm{Sym}\big(V(t)\big)\hat{Z}(t)\Big)\Big]\mathrm{d}t\Bigg).\end{aligned}$$

Expanding $\hat{Z}(t)$ using equation (158) and integrating with respect to $\hat{W}, \hat{a}$, we get the dynamical equations:

$$\begin{aligned}\dot{W}(t) &= \left(\mathcal{H}(t) + r(t)Z^* - \int_0^t\Gamma(t,t')Z(t')\mathrm{d}t'\right)W(t)D_{a(t)},\\ \dot{a}(t) &= \frac{\vartheta}{2}\mathrm{diag}\left(W(t)^\top\left[\mathcal{H}(t) + r(t)Z^* - \int_0^t\Gamma(t,t')Z(t')\mathrm{d}t'\right]W(t)\right),\end{aligned} \tag{160}$$

where $Z(t) = W(t)D_{a(t)}W(t)^\top$. Then, in order to get the set of self-consistent equations, one can simply consider the ones given in Section C.2 while replacing the evolution of $W(t)$ in (115) by the joint dynamics for $W(t), a(t)$ in equation (160). As underlined previously, the response function R_Z in equation (121) is defined under a perturbation of the noise $\mathcal{H}(t) \mapsto \mathcal{H}(t) + H(t)$, in both the dynamics for $a(t)$ and $W(t)$. In the end, the regularization and thermal noise introduced in Section 3.1.2 can be added back into the dynamics (160) and lead to the dynamical equations (20), (21).

To conclude, note the similarities of the dynamical structure between equations (159) (when $H = 0$) and (160). In these dynamical equations, the structure of the optimization problem (gradient flow associated with a function of $Z = WD_aW^\top$) remains, and the gradient of the optimized function is replaced by a sum of nonlinear terms involving a high-dimensional Gaussian process, the teacher matrix and a non-local memory contribution. Interestingly, as it is the case in equation (159), these terms only depend on $Z(t)$, and not directly on $W(t), a(t)$.

C.6 Gaussian Equivalence

In this section we give some intuition regarding Conjecture 16 on the equivalence between the Gaussian matrix sensing model and the shallow quadratic networks setting. The dynam-

ical equations of Section C.2 were derived under the first model, with the sensing matrices $X_1, \dots, X_n$ being i.i.d. drawn from the GOE distribution. From now on we consider the second one, i.e., we assume the $(X_k)_{1 \le k \le n}$ to be distributed as:

$$X_k = \frac{x_k x_k^\top - I_d}{\sqrt{d}}, \qquad x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d). \tag{161}$$

As a first remark, one can easily see that $\mathbb{E}\, X_k = 0$ and that the covariance of X_k is given by:

$$\mathbb{E}\,(X_k)_{ij}(X_k)_{i'j'} = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big),$$

which is the same as for the GOE. This is the first requirement for equivalence to hold: one needs to match the first and second moments.

In Section C.3, we exploited the Gaussian structure of the observations to derive self-consistent equations for the process $\mathbf{y}$ (see equation 129). For more general distributions of $X_1, \dots, X_n$, this process is *a priori* non-Gaussian. However, when the matrices are drawn as in equation (161), we argue that $\mathbf{y}$ becomes Gaussian in the high-dimensional limit. In what follows, without giving a rigorous proof, we analyze the cumulants of this process and explain how we can show that they match the ones of a Gaussian process in the limit.

C.6.1 Cumulants of the Quadratic Networks Distribution

Let us start by defining the notion of cumulants. Consider some real random variables $Y_1, \dots, Y_r$ with finite moments. Their joint cumulant is defined as:

$$K_r\big(Y_1, \dots, Y_r\big) = \frac{\partial^r}{\partial s_1 \dots \partial s_r} \log \mathbb{E}\left[\exp\left(\sum_{i=1}^r s_i Y_i\right)\right]\Bigg|_{s_1, \dots, s_r = 0}. \tag{162}$$

It is easily seen that $K_1(X) = \mathbb{E}[X]$ and $K_2(X, Y) = \mathrm{Cov}(X, Y)$. In addition, we have the following characterization of the multivariate Gaussian distribution in terms of cumulants: a random vector $Y \in \mathbb{R}^p$ is Gaussian if and only if:

$$K_r\big(Y_{i_1}, \dots, Y_{i_r}\big) = 0,$$

for any $r > 2$ and indices $i_1, \dots, i_r \in \{1, \dots, p\}$. In addition, this property still holds for Gaussian processes, since such processes are characterized by the Gaussianity of their finite-dimensional marginals.

In our case, we replace the Gaussian observations by the quadratic Gaussian matrices in equation (161). As suggested by the following lemma, the structure of the cumulants is then more complex, but still can be understood:

Lemma 39 *Let $x \sim \mathcal{N}(0, I_d)$ and $A_1, \dots A_r \in \mathcal{S}_d(\mathbb{R})$. Then, for $r \ge 1$:*

$$K_r\big(x^\top A_1 x, \dots, x^\top A_r x\big) = \frac{2^{r-1}}{r} \sum_{\sigma \in \mathfrak{S}_r} \mathrm{Tr}\big(A_{\sigma(1)} \dots A_{\sigma(r)}\big),$$

where $\mathfrak{S}_r$ denotes the set of permutations of r elements.

Proof Start from the definition of the multidimensional cumulants in equation (162), and plug in $Y_i = x^\top A_i x$. Then:

$$K_r\big(Y_1,\dots,Y_r\big) = \frac{\partial^r}{\partial s_1 \dots \partial s_r} \log \mathbb{E}\Big[e^{x^\top M_s x}\Big]\Big|_{s_1,\dots,s_r=0},$$

with:

$$M_s = \sum_{i=1}^r s_i A_i.$$

Then, choosing $s_1,\dots,s_r$ close enough to zero, M_s has a small enough spectral radius so that the previous expectation is finite and reads:

$$\log \mathbb{E}\Big[e^{x^\top M_s x}\Big] = -\frac{1}{2}\log\det\big(I_d - 2M_s\big) = \frac{1}{2}\sum_{p=1}^\infty \frac{1}{p} 2^p \operatorname{Tr}\big(M_s^p\big).$$

Then developing $\operatorname{Tr}(M_s^p)$, we obtain:

$$\log \mathbb{E}\Big[e^{x^\top M_s x}\Big] = \sum_{p=1}^\infty \frac{2^{p-1}}{p} \sum_{1\le i_1,\dots,i_p\le r} s_{i_1}\dots s_{i_p} \operatorname{Tr}\big(A_{i_1}\dots A_{i_p}\big).$$

Since we are taking the partial derivative with respect to $s_1,\dots,s_r$ and then setting these variables to zero, all the terms $p \neq r$ vanish. In addition, it is clear that $i_1,\dots,i_r$ has to correspond to a permutation of $\{1,\dots,r\}$, otherwise one of the s_i would not be represented and lead to a zero derivative. Therefore we get the desired. ■

As a consequence of this lemma, if we let:

$$y_i = \operatorname{Tr}\left(A_i \frac{xx^\top - I_d}{\sqrt{d}}\right) = \frac{x^\top A_i x - \operatorname{Tr}(A_i)}{\sqrt{d}},$$

due to the homogeneity of the cumulants and the invariance with respect to a constant shift (for $r \ge 2$), we have the expression:

$$K_r(y_1,\dots,y_r) = \frac{1}{d^{r/2}} \frac{2^{r-1}}{r} \sum_{\sigma\in\mathfrak{S}_r} \operatorname{Tr}\big(A_{\sigma(1)}\dots A_{\sigma(r)}\big).$$

Then, if all traces of products involving the matrices $A_1,\dots,A_r$ are of order d (as it will be the case for us), we have that in the $d\to\infty$ limit, $K_r(y_1,\dots,y_r) \longrightarrow 0$ as soon as $r > 2$, leading to the asymptotic Gaussianity of the random vector $(y_1,\dots,y_r)$ in the high-dimensional limit.

C.6.2 Equivalence for the Dynamical Partition Function

Let us apply this result in our dynamical setting. Recall the expression of the partition function in Section C.3:

$$\overline{\mathcal{Z}}_{\text{dyn}} \propto \int \mathcal{D}W\mathcal{D}\hat{W}\, \exp\left(-id\int_0^T \operatorname{Tr}\dot{W}(t)\hat{W}(t)^\top \mathrm{d}t\right) \times \left[\mathbb{E}_{\mathbf{y},z} \exp\left(-\frac{i}{\alpha}\int_0^T \ell'\big(y(t),z\big)\hat{y}(t)\mathrm{d}t\right)\right]^n,$$

where:

$$\mathbf{y}(t) = \Big(\mathrm{Tr}(XZ(t)), \mathrm{Tr}(X\hat{Z}(t)), \mathrm{Tr}(XZ^*)\Big).$$

This expression of the partition function holds no matter the distribution of X. When X was drawn from the GOE, the next steps were to introduce the covariance function Q of the Gaussian process $\mathbf{y}$, and to write saddle-point equations on this covariance. Similarly here, we could perform the saddle-point with respect to the higher-order cumulants of the process $\mathbf{y}(t)$:

$$Q_{i_1,\dots,i_r}(t_1,\dots,t_r) = K_r\big(\mathbf{y}_{i_1}(t_1),\dots,\mathbf{y}_{i_r}(t_r)\big), \tag{163}$$

with $i_1,\dots,i_r \in \{1,2,3\}$ (corresponding to $Z, \hat{Z}$ or Z^*) and $t_1,\dots,t_r \in [0,T]$. As it was already shown for the Gaussian case, in the high-dimensional limit, these cumulants would concentrate around the ones when averaging with respect to the true dynamics, and the order-r cumulant would then be expressed as a finite sum of terms of the form:

$$\frac{1}{d^{r/2}}\mathbb{E}\,\mathrm{Tr}\Big(Z_{i_1}(t_1)\dots Z_{i_r}(t_r)\Big), \tag{164}$$

with $Z_1 = Z, Z_2 = \hat{Z}, Z_3 = Z^*$. We now claim the following: as a consequence of the scaling chosen for the initialization, teacher, and the dynamics, all of these traces (at least in expectation) will remain of order d. To be more precise, as we showed in Section C.3.3 when considering a general gradient flow dynamics (137), the contraction with the matrix $\hat{Z}$ acts as a derivative when perturbing the dynamics for $Z(t)$. Therefore, any trace as in equation (164) can in principle be expressed as a derivative of a contraction between the matrix $Z(t)$ (at different instants) and the teacher Z^*. Although we do not prove rigorously that these traces all remain of order d, we believe that such a property holds.

In the end, this would allow to treat $\mathbf{y}$ as a Gaussian process, and therefore would lead to the same result derived for the Gaussian case.

Some remarks. We give some remarks regarding our previous arguments:

- The previous argument, if made rigorous, would only guarantee the convergence in distribution of the finite-dimensional marginals of $\mathbf{y}$. To prove the convergence in distribution (in the space of continuous functions on $[0,T]$) of $\mathbf{y}$ toward a Gaussian process, one would require a tightness argument (see for instance Billingsley, 2013, Section 7).
- Regarding the generality of our result, our calculation heavily rests on the quadratic Gaussian structure of the sensing matrices (161): it enables us to compute exactly the cumulants. As a consequence, it is not clear whether this cumulant method could be extended to more general distributions.

Appendix D. Derivation of the Long-Time Equations

This section is devoted to the long-time analysis of the system of equations stated in Claim 4. In particular, we derive the system of equations of Claim 6 and provide several proofs and insights supporting the results presented in Section 3.2. The plan of the section is as follows:

- In Section D.1, we discuss the steady-state assumption formulated in Assumption 5 and relate it to the fast convergence of the dynamics.

- In Section D.2, we build on this assumption to derive the set of equations at long times given in Claim 6.
- In Section D.3, we show that the same set of equations can be obtained by taking the high-dimensional limit before the long-time limit, reinforcing our claim.
- In Section D.4, we show that in the overparameterized case $\kappa \geq 1$, our system of equations exactly matches the one derived by Erba et al. (2025b) in the empirical risk minimization setting.
- In Section D.5, we analyze the system of equations at long times and derive some additional claims made in Section 3.2.
- In Section D.6, we study the population limit and prove the claims made in Section 3.1.4 and Proposition 7.

D.1 Steady-State Assumption

In this section we discuss the steady-state assumption formulated in Assumption 5. We start by recalling the set of equations from Claim 4. With the choice of the ℓ_2-regularization, the student matrix $W(t)$ solves the dynamics:

$$\dot{W}(t) = 2\int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - W(t')W(t')^\top\Big)\mathrm{d}t'\, W(t) - 2\lambda W(t). \tag{165}$$

From the equations in Claim 4, the covariance of the centered Gaussian process $\mathcal{G}(t)$ and the kernel $R(t,t')$ can be computed as averages with respect to equation (165):

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{2\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\left(\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big((Z(t)-Z^*)(Z(t')-Z^*)\Big) + \frac{\Delta}{2}\right), \tag{166}$$

$$R(t,t') = \delta(t-t') - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\bigg|_{H=0}\right), \tag{167}$$

with $Z(t) = W(t)W(t)^\top$, and the response is defined in terms of a perturbation $\mathcal{G}(t) \to \mathcal{G}(t) + H(t)$ in equation (165).

The steady-state assumption bears on the fast convergence of the memory kernel $R(t,t')$ as $t - t' \to \infty$, and the one of the Gaussian process $\mathcal{G}(t)$ toward its final value, that we denote $\mathcal{G}_\infty$. As mentioned in Section 3.2.1, we formulate this assumption in such a way that we can approximate the dynamics (165) by:

$$\dot{W}(t) = 2r_\infty\Big(\mathcal{G}_\infty + Z^* - W(t)W(t)^\top\Big)W(t) - 2\lambda W(t), \tag{168}$$

where:

$$r_\infty = \lim_{t\to\infty}\int_0^t R(t,t')\mathrm{d}t'.$$

We now explain how the steady-state assumption can be interpreted as one on the fast convergence of the matrix $Z(t) = W(t)W(t)^\top$.

Gaussian noise. We first examine the Gaussian process $\mathcal{G}(t)$. From the expression of its covariance in equation (166), we obtain:

$$\frac{1}{d}\mathbb{E}\Big[\big\|\mathcal{G}(t)-\mathcal{G}(t')\big\|_F^2\Big] = \frac{1}{2\alpha d}\left(1+\frac{1}{d}\right)\mathbb{E}\Big[\big\|Z(t)-Z(t')\big\|_F^2\Big].$$

Therefore, the convergence to $Z(t)$ at long times implies the one of $\mathcal{G}(t)$ toward a Gaussian matrix $\mathcal{G}_\infty$. In addition, the convergence rate of $\mathcal{G}(t)$ as $t\to\infty$ is directly given by the one of $Z(t)$. Then, Assumption 5 on the noise is not an independent dynamical property, but simply reflects the fast convergence of $Z(t)$.

Response. We now consider the response function defined in equation (167). By definition, $R(t,t')$ measures the effect at time t of a perturbation applied at an earlier time t' to the trajectory $Z(t)$. As the gradient flow converges at long times, the influence of the initial stages of the dynamics progressively vanishes. As a consequence, perturbations applied far in the past have little impact on the state of the system at long times, and the response $R(t,t')$ decays as $t-t'$ increases. Therefore, a faster relaxation of the dynamics leads to a faster decay of R.

Once the dynamics is close to convergence, the response is therefore localized near $t'=t$. This motivates approximations in which nonlocal memory terms vanish. In particular, one expects:

$$\int_0^t R(t,t')\,\phi(t')\,\mathrm{d}t' \underset{t\to\infty}{\approx} \left(\int_0^t R(t,t')\,\mathrm{d}t'\right)\phi(t),$$

with an error controlled by the relaxation rate of the dynamics. As for the noise, the decay of the response is not an independent assumption, but follows from the loss of memory along the gradient flow.

D.2 Derivation of the Long-Time Equations

The steady-state assumption allows to simplify the dynamics at long times. In addition to the dynamics (168), the variables r_∞ and the covariance of $\mathcal{G}_\infty$ are self-consistently computed from $W(t)$ at long times. These self-consistent relations are given in equations (34), (35).

In this section we start from these equations and derive the set of self-consistent scalar equations of Claim 6. The section is organized as follows:

- In Section D.2.1, we derive the long-time limit of the dynamics given in equation (39).
- In Section D.2.2, we use the simplified dynamics to compute the associated response function, which allows to close the set of equations as $t\to\infty$, and obtain equation (37b).
- In Section D.2.3, we compute the high-dimensional expression of the MSE using the expression of Z_∞ as a function of the variables ξ, q, ω. This leads to equation (37c).

D.2.1 Limit of the Dynamics

Let us now derive the limit of the dynamics under the steady-state assumption.

Limit of the student matrix. Given the simplified equation (168), one can interpret the dynamics as an Oja flow (see Section 4), whose limit is derived in Proposition 18. More precisely, for almost all initializations, the dynamics (168) converges to a point $W_\infty \in \mathbb{R}^{d\times m}$ such that:

$$W_\infty W_\infty^\top = \left(Z^* + \mathcal{G}_\infty - \frac{\lambda}{r_\infty} I_d\right)^+_{(m)}. \tag{169}$$

We recall that $X \in \mathcal{S}_d(\mathbb{R}) \mapsto X^+_{(m)} \in \mathcal{S}_d(\mathbb{R})$ is the spectral map selecting the m largest positive eigenvalues (see Definition 34). Then, setting $q = \lambda/r_\infty$ and $\mathcal{G}_\infty = \sqrt{\xi}\mathcal{G}$ with $\mathcal{G} \sim \mathrm{GOE}(d)$, we precisely get the limit of the dynamics given in equation (39).

Now, taking the limit $t, t' \to \infty$ in the expression of the covariance of $\mathcal{G}$ in equation (166) leads to the self-consistent equation between ξ and Z_∞:

$$\xi = \frac{1}{2\alpha}\left(\frac{1}{d}\mathbb{E}\left\|Z_\infty - Z^*\right\|_F^2 + \frac{\Delta}{2}\right), \tag{170}$$

where Z_∞ is the matrix in equation (169). This leads to the same expression of ξ as in (34), and therefore the expression of the MSE in equation (40) by rearranging.

Limit of the labels. Let us now apply the steady-state assumption to derive the limit of the label $y(t)$. To do so, recall the expression of $y(t)$ in equation (24).

Similarly to Assumption 5, we assume that the Gaussian process $\xi(t)$ appearing in the dynamics (24) converges fast enough as $t \to \infty$. Indeed, we have the covariance:

$$\mathbb{E}\,\xi(t)\xi(t') = 2C_Z(t,t') - \frac{2}{Q_*} m_Z(t) m_Z(t'),$$

where C_Z, m_Z, Q_* are defined in equation (27). Therefore, as it was discussed in Section D.1 for the case of the Gaussian process $\mathcal{G}(t)$, the fast convergence of the process $\xi(t)$ is a direct consequence of the one of $Z(t)$.

Then, using the steady-state assumption in equation (24), we obtain the following expression of the label as $t \to \infty$:

$$y_\infty = \left(1 - r_\infty + r_\infty \frac{m_Z^\infty}{Q_*}\right) y^* + r_\infty \xi_\infty + \sqrt{\Delta}(1 - r_\infty)\zeta, \tag{171}$$

where $y^* \sim \mathcal{N}(0, 2Q_*), \zeta \sim \mathcal{N}(0,1)$ and ξ_∞ are three independent centered Gaussian variables, and ξ_∞ has variance:

$$\mathbb{E}\,\xi_\infty^2 = 2C_Z^\infty - \frac{2}{Q_*}\left(m_Z^\infty\right)^2,$$

and we have:

$$C_Z^\infty = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\left(Z_\infty^2\right), \qquad m_Z^\infty = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\left(Z_\infty Z^*\right).$$

This gives access to the expression of the training loss. Indeed, recall that we have:

$$\mathrm{Loss}_{\mathrm{train}} = \frac{1}{4}\mathbb{E}\,(y_\infty - z)^2.$$

In our case, since the noisy label writes $z = y^* + \sqrt{\Delta}\zeta$, we obtain:

$$\mathrm{Loss}_{\mathrm{train}} = \frac{r_\infty^2}{2}\left(\mathrm{MSE} + \frac{\Delta}{2}\right).$$

Now using the relationship between the MSE and ξ in equation (34) and the fact that $r_\infty = \lambda/q$, we get the expression of the training loss (40).

D.2.2 Response Equation

The goal of this section is to derive equation (37b). To do so, we use the fact that the variable r_∞ is itself expressed as a function of the dynamics, through the response function:

$$R_Z(t,t') = \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z(t)}{\partial H(t')}\bigg|_{H=0}\right), \qquad r_\infty = 1 - \frac{1}{\alpha}\lim_{t\to\infty}\int_0^t R_Z(t,t')\mathrm{d}t'. \tag{172}$$

$H(t)$ is a perturbation of the high-dimensional dynamics (165):

$$\dot{W}(t) = 2\int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - Z(t') + H(t')\Big)\mathrm{d}t'\, W(t) - 2\lambda W(t). \tag{173}$$

We start by showing that we can compute the quantity r_∞ only using the long-time limit of the dynamics. This enables to use the result of the previous section. Recall that the response operator at times t, t' quantifies the change of $Z(t)$ in response to a perturbation introduced at t'. Therefore, the integrated response operator:

$$\int_0^t \frac{\partial \mathbb{E}\, Z(t)}{\partial H(t')}\bigg|_{H=0}\mathrm{d}t',$$

quantifies the response to a constant perturbation H (present since $t = 0$). Therefore, we have the identity:

$$\lim_{t\to\infty}\int_0^t R_Z(t,t')\mathrm{d}t' = \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z_\infty}{\partial H}\bigg|_{H=0}\right), \tag{174}$$

where Z_∞ is the limit of the perturbed dynamics (173). Now, similarly to the previous section, with the additional constant matrix H, we can deduce the limit from Proposition 18:

$$Z_\infty = \left(Z^* + \sqrt{\xi}\mathcal{G} - \frac{\lambda}{r_\infty}I_d + H\right)^+_{(m)}, \tag{175}$$

and we only need to compute the derivative of this matrix with respect to H. Here, we wrote the limit $\mathcal{G}(t) \xrightarrow[t\to\infty]{} \sqrt{\xi}\mathcal{G}$, with $\mathcal{G}$ being a GOE matrix.

Now the averaged response in equation (174) can be computed using Lemma 36. Indeed, equation (175) shows that we can write $Z_\infty = A^+_{(m)}$ at $H = 0$, with A having simple and non-zero eigenvalues, with probability one with respect to the Gaussian matrix $\mathcal{G}$. Therefore:

$$\frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z_\infty}{\partial H}\bigg|_{H=0}\right) = \frac{1}{d^2}\mathbb{E}\sum_{1\le i<j\le d}\frac{\lambda_i \mathbf{1}_{\lambda_i>0}\mathbf{1}_{i\le m} - \lambda_j\mathbf{1}_{\lambda_j>0}\mathbf{1}_{j\le m}}{\lambda_i - \lambda_j} + \frac{1}{d^2}\mathbb{E}\sum_{i=1}^d \mathbf{1}_{\lambda_i>0}\mathbf{1}_{i\le m},$$

where $\lambda_1 > \cdots > \lambda_d$ are the ordered eigenvalues of A. In the high-dimensional limit, the second term vanishes, and using the expression of r_∞ in equation (172), we are left with:

$$r_\infty = 1 - \frac{1}{2\alpha}\iint \frac{x\mathbf{1}_{x\ge\max(0,\omega)} - y\mathbf{1}_{y\ge\max(0,\omega)}}{x-y}\mathrm{d}\mu_A(x)\mathrm{d}\mu_A(y),$$

where μ_A is the limiting spectral density of A, and ω selects a fraction κ of this distribution, and solves the equation:

$$\min(\kappa, 1) = \int \mathbf{1}_{x \geq \omega} \mathrm{d}\mu_A(x).$$

Going back to the expression of Z_∞ in equation (175) with $H = 0$, we define $q = \lambda / r_\infty$. We then set μ_ξ to be the asymptotic spectral distribution of $Z^* + \sqrt{\xi}\mathcal{G}$. Therefore, ω now selects the m largest eigenvalues of μ_ξ above $q > 0$. This means that in the integral in the expression of r_∞ we should replace the lower bound of the integral by $\max(\omega, q)$. In addition, since the eigenvalues of A are simply those of $Z^* + \sqrt{\xi}\mathcal{G}$ shifted by q, we get the equation:

$$r_\infty = 1 - \frac{1}{2\alpha} \iint \frac{(x-q)\mathbf{1}_{x \geq \max(q,\omega)} - (y-q)\mathbf{1}_{y \geq \max(q,\omega)}}{x - y} \mathrm{d}\mu_\xi(x) \mathrm{d}\mu_\xi(y), \tag{176}$$

$$\min(\kappa, 1) = \int \mathbf{1}_{x \geq \omega} \mathrm{d}\mu_\xi(x). \tag{177}$$

Then using Lemma 30 with $\mu = \mu_\xi$ which admits a bounded square-integrable density as soon as $\xi > 0$, we have the identity by choosing $f(x) = (x - q)\mathbf{1}_{x \geq \max(q,\omega)}$:

$$\frac{1}{2} \iint \frac{(x-q)\mathbf{1}_{x \geq \max(q,\omega)} - (y-q)\mathbf{1}_{y \geq \max(q,\omega)}}{x - y} \mathrm{d}\mu_\xi(x) \mathrm{d}\mu_\xi(y) = \int_{\max(q,\omega)} (x - q) h_\xi(x) \mathrm{d}\mu_\xi(x),$$

and we obtain equation (37b) by replacing $r_\infty = \lambda / q$ in equation (176).

D.2.3 MSE in the High-Dimensional Limit

In the following, we compute the high-dimensional limit of the MSE associated with the predictor in equation (39):

$$Z_\infty = \left(Z^* + \sqrt{\xi}\mathcal{G} - q I_d \right)^+_{(m)},$$

where $\mathcal{G} \sim \mathrm{GOE}(d)$.

Proposition 40 *For $\xi > 0$ and $q \in \mathbb{R}$, consider:*

$$M_d(\xi, q) = \frac{1}{d} \left\| \left(Z^* + \sqrt{\xi}\mathcal{G} - q I_d \right)^+_{(m)} - Z^* \right\|_F^2,$$

where $\mathcal{G} \sim \mathrm{GOE}(d)$ and $m \sim \kappa d$ as $d \to \infty$. If the empirical spectral distribution of Z^* converges to some μ^* as $d \to \infty$, then:

$$\lim_{d \to \infty} M_d(\xi, q) = \int x^2 \mathrm{d}\mu^*(x) + \int_{\max(q,\omega)} (q^2 - x^2) \mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x - q) h_\xi(x) \mathrm{d}\mu_\xi(x).$$

where μ_ξ is the free additive convolution between μ^* and a semicircular distribution of variance ξ, and h_ξ is the Hilbert transform of μ_ξ (see Definition 29). Finally, ω verifies:

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x).$$

Proof We start by decomposing $M_d(\xi, q)$ at finite dimension, with $Z_\xi^* = Z^* + \sqrt{\xi}\mathcal{G}$:

$$M_d(\xi,q) = \frac{1}{d}\|Z^*\|_F^2 - \frac{2}{d}\mathrm{Tr}\Big(Z^*\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big) + \frac{1}{d}\Big\|\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big\|_F^2.$$

Then, by assumption, the first term concentrates in the $d \to \infty$ limit:

$$\frac{1}{d}\|Z^*\|_F^2 \xrightarrow[d\to\infty]{} \int x^2 \mathrm{d}\mu^*(x).$$

For the two other terms, we denote $\lambda_1, \dots, \lambda_d$ the eigenvalues of Z^* and $u_1, \dots, u_d$ a family of associated eigenvectors. We do the same for Z_ξ^* and write $\lambda_1^\xi, \dots, \lambda_d^\xi$ and $u_1^\xi, \dots, u_d^\xi$ its eigenvalues and eigenvectors. Now, we have:

$$\frac{1}{d}\mathrm{Tr}\Big(Z^*\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big) = \frac{1}{d}\sum_{i=1}^m\sum_{j=1}^d \lambda_j\big(\lambda_i^\xi - q\big)^+\big(u_j^\top u_i^\xi\big)^2,$$
$$\frac{1}{d}\Big\|\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big\|_F^2 = \frac{1}{d}\sum_{i=1}^m \big(\lambda_i^\xi - q\big)^{+2}.$$

We can derive the limit of the second term by remarking that the eigenvalues of Z_ξ^* which are selected are the m largest ones that are larger than q. Therefore, the second term concentrates around:

$$\frac{1}{d}\Big\|\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big\|_F^2 \xrightarrow[d\to\infty]{} \int_{\max(q,\omega)} (x-q)^2 \mathrm{d}\mu_\xi(x), \qquad \min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x).$$

For the first term, we use the result of Bun et al. (2017) (see Section 4 and Appendix D), and obtain:

$$\mathbb{E}\,\big(u_j^\top u_i^\xi\big)^2 \underset{d\to\infty}{\sim} \frac{1}{d}\frac{\xi}{(\lambda_j - \lambda_i^\xi + \xi h_\xi(\lambda_i^\xi))^2 + \pi^2\xi^2\rho_\xi(\lambda_i^\xi)^2},$$

where h_ξ is the Hilbert transform of μ_ξ, and ρ_ξ its density. Therefore:

$$\frac{1}{d}\mathrm{Tr}\Big(Z^*\big(Z_\xi^* - qI_d\big)_{(m)}^+\Big) \xrightarrow[d\to\infty]{} \int_{\max(q,\omega)}\int \frac{\xi y(x-q)}{(y - x + \xi h_\xi(x))^2 + \pi^2\xi^2\rho_\xi(x)^2}\mathrm{d}\mu^*(y)\mathrm{d}\mu_\xi(x).$$

Note that to be fully rigorous, one should prove that the variance of this last term vanishes. Putting everything together, we get:

$$\begin{aligned}\lim_{d\to\infty} M_d(\xi,q) = &\int x^2\mathrm{d}\mu^*(x) + \int_{\max(q,\omega)}(x-q)^2\mathrm{d}\mu_\xi(x) \\ &- 2\xi\int_{\max(q,\omega)}\int \frac{y(x-q)}{(y - x + \xi h_\xi(x))^2 + \pi^2\xi^2\rho_\xi(x)^2}\mathrm{d}\mu^*(y)\mathrm{d}\mu_\xi(x).\end{aligned} \tag{178}$$

Now that we have obtained an expression in the high-dimensional limit, we proceed to simplify it. Let us rewrite this equation as:

$$\lim_{d\to\infty} M_d(\xi,q) = Q_* + \int_{\max(q,\omega)}(x-q)^2\mathrm{d}\mu_\xi(x) - 2\xi\int_{\max(q,\omega)}(x-q)I_{\mu^*}\big(z_\xi(x)\big)\mathrm{d}\mu_\xi(x), \tag{179}$$

with:

$$I_\nu(z) = \int \frac{y}{|y-z|^2} \mathrm{d}\nu(y), \qquad z_\xi(x) = x - \xi h_\xi(x) + i\pi\xi\rho_\xi(x).$$

Then, using the identity:

$$\frac{y}{|y-z|^2} = \frac{1}{z-\bar{z}}\left(\frac{z}{y-z} - \frac{\bar{z}}{y-\bar{z}}\right),$$

we obtain that:

$$I_\nu(z) = -\frac{\operatorname{Im} z m_\nu(z)}{\operatorname{Im} z}, \qquad m_\nu(z) = \int \frac{\mathrm{d}\nu(y)}{z-y}.$$

m_ν is known as the Stieltjes transform of ν (see Definition 29). Now, since μ_ξ is the free additive convolution between μ^* and a semicircular density with variance ξ, as a consequence of Lemma 32, we have for all $z \in \mathbb{C} \setminus \operatorname{Supp}(\mu_\xi)$:

$$m_\xi(z) = m_*\big(z - \xi m_\xi(z)\big),$$

where m_ξ and m_* are the Stieltjes transforms of μ_ξ and μ^*. Taking $z = x + i\eta$ and letting $\eta \to 0^+$, we then obtain, using Definition 29:

$$h_\xi(x) - i\pi\rho_\xi(x) = m_*\big(z_\xi(x)\big).$$

Therefore:

$$\xi I_{\mu^*}\big(z_\xi(x)\big) = -\xi\frac{\operatorname{Im}\big[z_\xi(x) m_*(z_\xi(x))\big]}{\operatorname{Im} z_\xi(x)} = x - 2\xi h_\xi(x).$$

Plugging this into equation (179) leads to:

$$\lim_{d\to\infty} M_d(\xi, q) = Q_* + \int_{\max(q,\omega)} \big(q^2 - x^2\big)\mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x),$$

which is the desired. ■

To conclude, using the expression $Z_\infty = \big(Z^* + \sqrt{\xi}\mathcal{G} - qI_d\big)^+_{(m)}$, and the fact that $M_d(\xi, q)$ is precisely the MSE associated with Z_∞, the relationship between the MSE and ξ in equation (34) can be rewritten as:

$$2\alpha\xi - \frac{\Delta}{2} = Q_* + \int_{\max(q,\omega)} \big(q^2 - x^2\big)\mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x),$$

which is precisely equation (37c).

D.3 Equations in the High-Dimensional Limit

As mentioned earlier, for the sake of consistency with our earlier results, we derive our set of high-dimensional equations (37) in a slightly different fashion. Indeed, the method proposed in Section D.2 first uses the long-time limit of the Oja flow (and its response),

before taking the high-dimensional limit. However, we recall that the results derived in Section 3.1 already rely on a large dimension and are only valid for a fixed time horizon.

Therefore, in order to strengthen our results, we show that we recover the same equations when first taking the high-dimensional limit (associated with the simplified dynamics) before the limit $t \to \infty$. To do so, we use the results of Section 4 on the Oja flow dynamics.

Limit of the dynamics. In Section 4.3, we derive convergence rates for the Oja flow in the high-dimensional limit. In particular, we show that if $W(t)$ solves the dynamics:

$$\dot{W}(t) = \big(A - W(t)W(t)^\top\big)W(t),$$

then:

$$\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\big\|Z(t) - Z_\infty\big\|_F^2 = 0,$$

with $Z(t) = W(t)W(t)^\top$ and $Z_\infty = A^+_{(m)}$. Crucially in the previous equation, the high-dimensional limit is taken before the long-time limit. Applying this result to the target matrix $A = Z^* + \sqrt{\xi}\mathcal{G} - qI_d$, we end up with the same limit as the one found in Section D.2.1. Using again the result of Section D.2.3 as well as the relationship (34), this precisely leads to equation (37c).

Response equation. In order to derive the response equation, recall the expression of R_Z and its link to r_∞ in equation (172). Using Proposition 25, we have the high-dimensional limit:

$$\int_0^t R_Z(t,t')\mathrm{d}t' \xrightarrow[d\to\infty]{} \frac{\mathfrak{g}(t)}{2}\iint\frac{1}{y-x}\left[\frac{ye^{2yt}}{q_t(y)} - \frac{xe^{2xt}}{q_t(x)}\right]\mathrm{d}\mu_A(x)\mathrm{d}\mu_A(y), \tag{180}$$

where $q_t(x) = \mathfrak{g}(t)(e^{2xt}-1) + x$ and $\mathfrak{g}(t)$ solves the self-consistent equation:

$$\kappa\mathfrak{g}(t) + 1 - \kappa = \int\frac{x}{(e^{2xt}-1)\mathfrak{g}(t) + x}\mathrm{d}\mu_A(x).$$

We have proven this proposition in the case where the initialization of the flow is a Gaussian matrix, but we believe that when taking the $t \to \infty$ limit, the result becomes independent of the initialization (as it is the case for our derivation in Section D.2.2).

Now, in order to take the long-time limit in equation (180), one can simply use Lemma 52 that derives the long-time asymptotics of the function $\mathfrak{g}(t)$. As a consequence, we have the asymptotic:

$$\mathfrak{g}(t)\frac{xe^{2xt}}{q_t(x)} \xrightarrow[t\to\infty]{} x\mathbf{1}_{x\geq\max(0,\omega)}, \qquad \kappa = \int\mathbf{1}_{x\geq\omega}\mathrm{d}\mu_A(x).$$

Now plugging the expression of $A = Z^* + \sqrt{\xi}\mathcal{G} - qI_d$, we arrive at the identity:

$$\lim_{t\to\infty}\lim_{d\to\infty}\int_0^t R_Z(t,t')\mathrm{d}t' = \frac{1}{2}\iint\frac{(x-q)\mathbf{1}_{x\geq\max(q,\omega)} - (y-q)\mathbf{1}_{y\geq\max(q,\omega)}}{x-y}\mathrm{d}\mu_\xi(x)\mathrm{d}\mu_\xi(y).$$

Using Lemma 30, we end up with equation (37b).

Conclusion. In conclusion, the equations we obtain are unchanged when first taking the high-dimensional limit. This suggests a robustness of the Oja flow dynamics and reveals that the only relevant timescale is $t = O_d(1)$.

Indeed, first taking the limit $t \to \infty$ allows the dynamics to explore all timescales, including those that may depend on the dimension. On the other hand, taking the limit $d \to \infty$ restricts the analysis to a timescale of order one. The fact that both limits lead to the same system of equations indicates that no additional dynamical behavior emerges beyond the order-one timescale.

D.4 Link with Empirical Risk Minimization

In this part we link the system of equations in Claim 6 to the recent results of Erba et al. (2025b). In this work, the authors derive the statistics of the global minimizer of the same regularized loss as ours, in the case $\kappa \geq 1$.

In the following, we show that our system of equations is the same as theirs after matching our conventions. This result is no surprise for the following reason: it is known that when optimizing a function of $WW^\top$ (as in our case) in the setting where $m \geq d$, the gradient flow always converges to a global minimizer of the loss over the PSD matrices (see for instance Bach, 2024b, Section 12.3.3).

D.4.1 Correspondence of the Equations

The first step when comparing both setups is to match the constants used. By matching the expressions of our respective loss functions, we arrive at the expression of their regularization parameter:

$$\lambda_{\text{ERM}} = \frac{4\alpha}{\sqrt{\kappa}}\lambda.$$

In addition, their results involve the free additive convolution between the teacher spectral distribution and a semicircular density with radius 2δ, i.e., with variance δ^2. Then, with our notations, their set of self-consistent equations (Theorem 1) is given by:

$$4\alpha\delta - \frac{\delta}{\epsilon} = 2\delta\, \partial_1 J(\delta^2, 4\alpha\lambda\epsilon), \tag{181}$$

$$Q_* + \frac{\Delta}{2} + 2\alpha\delta^2 - \frac{\delta^2}{\epsilon} = \big(1 - 4\alpha\lambda\epsilon\partial_2\big) J(\delta^2, 4\alpha\lambda\epsilon), \tag{182}$$

where the unknowns are δ, ϵ, and:

$$J(a, b) = \int_b (x-b)^2 \mathrm{d}\mu_a(x),$$

and μ_a corresponds to the asymptotic spectral density of the matrix $Z^* + \sqrt{a}\mathcal{G}$ with $\mathcal{G} \sim \text{GOE}(d)$. We refer to Section B.1.3 for more details.

Derivatives of J. We shall now compute the partial derivatives of the function J. Let us start with the derivative with respect to b. Since the map $b \mapsto (x-b)^2 \mathbf{1}_{x \geq b}$ is $\mathcal{C}^1$, we have, interchanging integration and differentiation:

$$\partial_2 J(a, b) = -2 \int_b (x-b) \mathrm{d}\mu_a(x). \tag{183}$$

Regarding the derivative with respect to a, we start from the complex Burgers' equation satisfied by m_a, the Stieltjes transform of μ_a (see Lemma 33):

$$\partial_a m_a(z) + m_a(z)\partial_z m_a(z) = 0,$$

Evaluating at $z = x + i\eta$ and taking imaginary parts while $\eta \to 0$, one gets the continuity equation using Definition 29:

$$\partial_a \rho_a(x) + \partial_x\big(h_a(x)\rho_a(x)\big) = 0.$$

Note that this equation is only verified in the sense of distributions. In the following, we use this equation non-rigorously and differentiate under the integral:

$$\begin{aligned}\partial_1 J(a,b) &= \int_b (x-b)^2 \partial_a \rho_a(x) \mathrm{d}x \\ &= -\int_b (x-b)^2 \partial_x\big(h_a(x)\rho_a(x)\big)\mathrm{d}x.\end{aligned}$$

Then integrating by parts, we finally get:

$$\partial_1 J(a,b) = 2\int_b (x-b) h_a(x) \mathrm{d}\mu_a(x). \tag{184}$$

Equivalence of the systems of equations. We now go back to the system of equations (181) and (182). We set $\xi = \delta^2$ and $q = 4\alpha\lambda\epsilon$, and using the derivatives in equations (183) and (184), we obtain that q, ξ solve the equations:

$$1 - \frac{\lambda}{q} = \frac{1}{\alpha}\int_q (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x), \tag{185}$$

$$Q_* + \frac{\Delta}{2} + 2\alpha\xi - \frac{4\alpha\lambda\xi}{q} = \int_q \big(x^2 - q^2\big)\mathrm{d}\mu_\xi(x). \tag{186}$$

Already the first equation is the same as (37b) in Claim 6, in the case where $\kappa \geq 1$. In addition, equation (37c) is directly obtained by replacing λ/q in equation (186) using equation (185).

Let us now show that this also leads to the same expression of the MSE and loss. In our case, recall the expressions:

$$\mathrm{MSE} = \frac{1}{d}\big\|WW^\top - Z^*\big\|_F^2, \qquad \mathrm{Loss}_{\mathrm{train}} = \frac{1}{4n}\sum_{k=1}^n \big(\mathrm{Tr}(X_k Z) - z_k\big)^2.$$

Now, Erba et al. (2025b) studied the test error and the loss value, that we respectively denote e_{test}, L. Taking into account their conventions, we reach the relationships with ours quantities:

$$e_{\mathrm{test}} = \mathrm{MSE}, \qquad L = 4\alpha\,\mathrm{Loss}_{\mathrm{train}} + 4\alpha\lambda \lim_{d\to\infty}\frac{1}{d}\|W_\infty\|_F^2. \tag{187}$$

Back to the correspondence between our variables ξ, q and their variables δ, ϵ, their expression of e_{test} directly leads to the MSE equation (40). Regarding their loss, we use the expression of the derivative of J in equation (183) and reach the expression of their loss:

$$L = \frac{4\alpha^2\xi\lambda^2}{q^2} + 4\alpha\lambda\int_q (x-q)\mathrm{d}\mu_\xi(x).$$

Now, since we have:

$$\int_q (x-q)\mathrm{d}\mu_\xi(x) = \lim_{d\to\infty}\frac{1}{d}\|W_\infty\|_F^2,$$

we indeed recover the expression of our loss in equation (40) using equation (187).

D.4.2 Stability Condition

In the same work, the authors derive a stability condition for the previous set of equations. This criterion is derived from an approximate message passing (AMP) iteration and reads, with our notations:

$$\iint \left(\frac{(x-q)^+ - (y-q)^+}{x-y} \mathrm{d}\mu_\xi(x) \mathrm{d}\mu_\xi(y) \right)^2 < 2\alpha. \tag{188}$$

Interestingly, this criterion matches the one we later derive in Section E.1.1, up to the fact that we only require the above quantity to be finite. We now show, using our system of equations, that the above criterion is verified. Recall that when $\kappa \geq 1$, as a consequence of Lemma 30, q, ξ are linked through the equation:

$$1 - \frac{\lambda}{q} = \frac{1}{2\alpha} \iint \frac{(x-q)^+ - (y-q)^+}{x-y} \mathrm{d}\mu_\xi(x) \mathrm{d}\mu_\xi(y).$$

Now, since the map $x \mapsto (x-q)^+$ is continuous and 1-Lipschitz, we have the bound, for all $x \neq y$:

$$\left(\frac{(x-q)^+ - (y-q)^+}{x-y} \right)^2 \leq \frac{(x-q)^+ - (y-q)^+}{x-y}.$$

Putting everything together, and since $\lambda > 0$, the criterion (188) is satisfied. Through the work of Erba et al. (2025b) on the AMP iteration associated with the same problem as ours, the direct verification of the stability criterion allows to reinforce the validity of the simplifications regarding the dynamics introduced in Section 3.2 (at least in the case $\kappa \geq 1$).

D.5 Analysis of the Long-Time Equations

In this part we analyze the system of equations given in Claim 6. In the following, we give the results:

- We derive the existence of the two regions with respect to the variable κ claimed in Section 3.2.3.
- We confirm that in the overparameterized region $\kappa \geq \kappa_{\min}$, the minimum reached by gradient flow is a global minimizer of the loss.

D.5.1 Overparameterized Region

We now prove the claim made in Section 3.2.3 regarding the two regimes depending on the value of κ. Indeed, as we claimed, there exists a value $\kappa_{\min}$, depending on $\alpha, \lambda, \kappa^*, \Delta$, such that the set of equations is independent of κ for $\kappa > \kappa_{\min}$. We recall the system of equations (37):

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x), \tag{189a}$$

$$1 = \frac{\lambda}{q} + \frac{1}{\alpha} \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x), \tag{189b}$$

$$2\alpha\xi - \frac{\Delta}{2} = Q_* + \int_{\max(q,\omega)} (q^2 - x^2) \mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x). \tag{189c}$$

Remark that these equations only depend on κ through the variable ω that selects a mass κ of the measure μ_ξ. Also remark that for a triplet (q, ξ, ω) solution of the system, having $q \geq \omega$ makes the last two equations independent of ω, and therefore κ. In this case only the pair (q, ξ) matters to solve the system of equations.

Lemma 41 *For a fixed set of parameters $\alpha, \lambda, \kappa^* > 0$ and $\Delta \geq 0$, consider a pair (q^*, ξ^*) solution of the system of equations* (189) *for $\kappa = 1$ (in this case one can choose $\omega = -\infty$). Let:*

$$\kappa_{\min} = \int_{q^*} \mathrm{d}\mu_{\xi^*}(x),$$

then, for $\kappa \geq \kappa_{\min}$, one can choose $\omega \leq q^*$ so that the triplet (q^*, ξ^*, ω) is solution of the system of equations (189)*.*

This lemma shows that for κ larger than $\kappa_{\min}$ we can pick the same solution q^*, ξ^* as when solving with $\kappa = 1$. Provided that for given values of our parameters $\alpha, \lambda, \kappa, \kappa^*, \Delta$, there is a unique solution in terms of the variables (q, ξ), this guarantees that for $\kappa \geq \kappa_{\min}$, the solution of the system (37) does not depend on κ.

Proof The key point is that the system of equations (189) only depends on κ through the variable ω. Let us consider these equations for some $\kappa \geq \kappa_{\min}$. We consider ω to be solution of:

$$\min(\kappa, 1) = \int_{\omega} \mathrm{d}\mu_{\xi^*}(x).$$

Remark that we used ξ^* to define ω. Since $\min(\kappa, 1) \geq \kappa_{\min}$ and due to the definition of $\kappa_{\min}$ we immediately get that $\omega \leq q^*$. We now plug the triplet (q^*, ξ^*, ω) into the system of equations (189), and remark that:

- Equation (189a) is verified due to the definition of ω.
- Since $\omega \leq q^*$, equations (189b), (189c) are exactly the same as in the $\kappa = 1$ case. Therefore they are solved by (ξ^*, q^*).

As a conclusion, as soon as $\kappa \geq \kappa_{\min}$, we can find some value of ω such that (q^*, ξ^*, ω) indeed solves the system (189). ■

D.5.2 Global Minimizer in the Overparameterized Region

We now briefly explain why in the region $\kappa \geq \kappa_{\min}$, the gradient flow estimator corresponds to a global minimizer of the loss over all PSD matrices. The first evidence is the calculation of Section D.4.1 where it was proved earlier that the system of equations in this region matches the one of Erba et al. (2025b), who worked in the empirical risk minimization setting.

The deeper reason behind this correspondence can be clarified by the following fact, which we do not prove in detail:

Proposition 42 *Let $L \colon \mathbb{R}^{d \times m} \to \mathbb{R}$ be such that $L(W) = G(WW^\top)$ for $G \colon \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}$ convex and real analytic. Let $\big(W(t)\big)_{t \geq 0}$ to be solution of the gradient flow:*

$$\dot{W}(t) = -\nabla L\big(W(t)\big).$$

Then, if $m \geq d$, for almost all initializations, $W(t)$ converges to a point W_∞ such that $W_\infty W_\infty^\top$ is a global minimizer of G over $\mathcal{S}_d^+(\mathbb{R})$.

The proof of this can be carried out as follows:

- Standard results on the convergence of gradient flow to local minimizers (see for instance Lee et al., 2016; Panageas and Piliouras, 2016) guarantee that for almost all initializations, $W(t)$ converges to a local minimizer of L.
- The convexity of G allows to conclude that any local minimizer of L translates into a global minimizer of G over the set of PSD matrices. For instance, this result can be found in Bach (2024b, Exercise 12.8).

Finally, in our case of interest in Section 3.2.3, the loss is indeed expressed as a convex function of $WW^\top$:

$$\mathcal{L}(W) = \frac{1}{4n}\sum_{k=1}^{n}\Big(\mathrm{Tr}\big(X_k WW^\top\big) - z_k\Big)^2 + \frac{\lambda}{d}\mathrm{Tr}\big(WW^\top\big).$$

All of these ingredients allow us to identify an overparameterized region, where the set of equations does not depend on κ anymore, and where the gradient flow converges to a point that minimizes the loss over all PSD matrices. As mentioned earlier, this minimizer has rank $\sim \kappa_{\min} d$, and for $\kappa < \kappa_{\min}$, the flow is unable to converge to such a point, due to the rank constraint.

D.6 Population Limit

In this section, we derive the population limit of the dynamics (5), corresponding to the regime where the student has access to an infinite number of observations. In line with our main results in this paper, we study this limit under Assumption 3, when using the quadratic cost for the loss and generating the labels with a noisy Gaussian channel. More precisely, we derive the following results:

- We start by taking the $n \to \infty$ limit (at fixed dimension) in the expression of the loss (4) and study the associated dynamics.
- Then, we take the limit $\alpha \to \infty$ (corresponding to the regime $n \gg d^2$) in the system of equations of Claim 4 and show that we obtain the same dynamics as in the previous step.
- Finally, we take the $\alpha \to \infty$ limit in the system of equations at long times derived in Claim 6, and show that the equations we obtained are coherent with the previous calculations. This provides a proof of Proposition 7.

Studying this limit from different angles leads to the following conclusions: first of all, taking the sequential limit $n \to \infty$ and then $d \to \infty$ leads to the same result as taking the joint limit $n \sim \alpha d^2$ and then sending $\alpha \to \infty$. This means that there is no intermediate scaling of n with the dimension that produces a different dynamics. Secondly, the same conclusion holds between the $\alpha \to \infty$ limit and the long times, meaning that the population limit preserves the relevant timescale for the dynamics.

D.6.1 Sequential Limit

Let us start with the expression of the loss (4) under Assumption 3:

$$\mathcal{L}(W) = \frac{1}{4n}\sum_{k=1}^{n}\Big(\mathrm{Tr}\big(X_k W W^\top\big) - \mathrm{Tr}\big(X_k Z^*\big) - \sqrt{\Delta}\zeta_k\Big)^2, \tag{190}$$

where $\zeta_1, \dots, \zeta_n$ are i.i.d. standard Gaussian variables, independent from all the other random variables of the problem. Taking the $n \to \infty$ limit allows to replace the empirical average over the n samples by the expectation over their distribution. Since $X_k \sim \mathrm{GOE}(d)$, we obtain that:

$$\mathcal{L}_{\mathrm{pop}}(W) = \frac{1}{2d}\big\|WW^\top - Z^*\big\|_F^2 + \frac{\Delta}{4}. \tag{191}$$

In this case the loss is directly related to the distance between the teacher and the student: in the population limit, we are simply optimizing the MSE. In this case, we can write the Langevin dynamics (5) with regularization Ω and inverse temperature β as:

$$\mathrm{d}W(t) = 2\Big(Z^* - W(t)W(t)^\top\Big)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t). \tag{192}$$

Already remark that the first term is very similar to the Oja flow dynamics studied in Section 4.

D.6.2 Limit of the Dynamical Equations

We now consider the $\alpha \to \infty$ limit in the system of equations in Claim 4, and show that it corresponds to the dynamics (192). To see this, first note that the evolution of $W(t)$ and the typical label in equations (23), (24) do not explicitly depend on α, and this dependence enters only through the covariance of the Gaussian process $\mathcal{G}$ and the response R that drive their evolution. We recall that:

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{2\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\left(\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big[\big(Z(t) - Z^*\big)\big(Z(t') - Z^*\big)\Big] + \frac{\Delta}{2}\right), \tag{193}$$

$$R(t,t') = \delta(t-t') - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\bigg|_{H=0}\right). \tag{194}$$

Now, as $\alpha \to \infty$, the covariance function of $Z(t)$ and its average response should remain of order one, so that from equations (193), (194) we get $\mathcal{G}(t) = 0$ almost surely, and $R(t,t') = \delta(t-t')$. Therefore, from equation (23), we get that W exactly solves equation (192). This means we have recovered the sequential limit by taking the $\alpha \to \infty$ limit. Regarding the evolution of the label, we simply get the expression from equation (24):

$$y(t) = \frac{m_Z(t)}{Q_*}y^* + \xi(t), \qquad\qquad y^* \sim \mathcal{N}(0, 2Q_*), \tag{195}$$

where ξ is a centered Gaussian process with covariance given in equation (26). We will now show that this evolution of the labels corresponds to a random Gaussian projection of the student. Indeed, consider $X \sim \mathrm{GOE}(d)$ independent of all other random variables. Then,

conditionally on $Z(t), Z^*$, the random projections $\tilde{y}(t) = \mathrm{Tr}\big(XZ(t)\big)$ and $\tilde{y}^* = \mathrm{Tr}\big(XZ^*\big)$ are Gaussian with zero mean and statistics:

$$\mathbb{E}\,\tilde{y}(t)\tilde{y}(t') = \frac{2}{d}\mathrm{Tr}\big(Z(t)Z(t')\big), \qquad \mathbb{E}\,\tilde{y}(t)\tilde{y}^* = \frac{2}{d}\mathrm{Tr}\big(Z(t)Z^*\big), \qquad \mathbb{E}\,\tilde{y}^{*2} = \frac{2}{d}\mathrm{Tr}\big(Z^{*2}\big).$$

Now, due to the covariance of $\xi(t)$ in equation (26), the couple $y(t), y^*$ in equation (195) has precisely the same statistics as $\tilde{y}(t), \tilde{y}^*$. This conclusion is intuitive: with a finite number of observations, the student remains correlated with the training examples, leading to the evolution of the typical label in Claim 4. In the population limit, the student becomes independent of the examples, and the evolution of the typical label is the same as for a label associated with a previously unseen sample.

D.6.3 Long-Time Analysis of the Population Equations

Let us now go back to equation (192), and similarly to what was done in Section 3.2, consider the gradient flow setting ($\beta = \infty$), and the ℓ_2-regularization $\Omega(W) = \lambda\mathrm{Tr}(WW^\top)$. Then, the population dynamics in equation (192) writes:

$$\dot{W}(t) = 2\Big(Z^* - W(t)W(t)^\top\Big)W(t) - 2\lambda W(t).$$

This now precisely corresponds to an Oja flow (see Section 4) with the target matrix $Z^* - \lambda I_d$. Under a random initialization, we can apply Proposition 18 to get:

$$W(t)W(t)^\top \xrightarrow[t\to\infty]{} \big(Z^* - \lambda I_d\big)^+_{(m)}.$$

Recall that the operator $A \mapsto A^+_{(m)}$ selects the m largest positive eigenvalues (see Definition 34). In the case $\kappa \geq \min(\kappa^*, 1)$, this simply writes $\big(Z^* - \lambda I_d\big)^+$, since $Z^* - \lambda I_d$ cannot have more than m positive eigenvalues. Then, denoting $\mu_1, \dots, \mu_d$ the eigenvalues of Z^*, the MSE in the high-dimensional limit writes:

$$\mathrm{MSE} = \lim_{d\to\infty} \frac{1}{d}\sum_{k=1}^{d}\Big((\mu_k - \lambda)^+ - \mu_k\Big)^2.$$

Using that $u^+ = \max(u, 0)$ and the convergence of the empirical spectral distribution of Z^*, we get:

$$\mathrm{MSE} = \lim_{d\to\infty} \frac{1}{d}\sum_{k=1}^{d}\min(\lambda, \mu_k)^2 = \int \min(\lambda, x)^2 \mathrm{d}\mu^*(x).$$

This is precisely the expression of the MSE in Proposition 7. The expression of the loss is a simple consequence of equation (191).

D.6.4 Population Limit in the Long-Time Equations

In the previous steps, we proved Proposition 7 starting from the dynamical equations in the population limit, and then studied the long times. For completeness, we show that the same

result holds when starting from the long-time result Claim 6 and then taking the $\alpha \to \infty$ limit. We recall the system:

$$\begin{aligned}
\min(\kappa, 1) &= \int_\omega \mathrm{d}\mu_\xi(x), \\
1 &= \frac{\lambda}{q} + \frac{1}{\alpha}\int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x), \\
2\alpha\xi - \frac{\Delta}{2} &= Q_* + \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x)\mathrm{d}\mu_\xi(x),
\end{aligned}$$

where q, ξ are the unknowns. Moreover, we have:

$$\mathrm{MSE} = 2\alpha\xi - \frac{\Delta}{2}, \qquad \mathrm{Loss}_{\mathrm{train}} = \frac{\lambda^2 \alpha \xi}{q^2}, \qquad Z_\infty = \Big(Z^* + \sqrt{\xi}\mathcal{G} - qI_d\Big)^+_{(m)}. \tag{196}$$

Now, as $\alpha \to \infty$, the finiteness of the MSE and the loss require that $\xi = \Theta(\alpha^{-1})$ and $q = \Theta(1)$. Therefore, the quantity:

$$\int_{\max(q,\omega)} (x-q) h_\xi(x)\mathrm{d}\mu_\xi(x),$$

remains of order one. Since $\alpha \to \infty$ and $\xi \to 0$, one gets the equations:

$$q = \lambda, \qquad 2 \lim_{\substack{\alpha\to\infty \\ \xi\to 0}} \alpha\xi = \frac{\Delta}{2} + Q_* + \lim_{\xi\to 0} \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x).$$

This means that from equation (196), we again obtain $Z_\infty = \big(Z^* - \lambda I_d\big)^+_{(m)}$, and we can drop the m largest eigenvalues selection when $\kappa \geq \min(\kappa^*, 1)$. As a consequence, we can pick $\omega = 0$ in the previous equations. Let us now compute the $\xi \to 0$ limit. In Section H.4, we compute the small ξ asymptotics of several integrals involving μ_ξ. As a consequence of Lemma 46, one can show that with $\omega = 0$ and $q = \lambda$, we have at leading order:

$$\begin{aligned}
\lim_{\xi\to 0} \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x) &= \lim_{\xi\to 0} \int_\lambda \big(\lambda^2 - x^2\big)\mathrm{d}\mu_\xi(x) \\
&= \int_\lambda \big(\lambda^2 - x^2\big)\mathrm{d}\mu^*(x).
\end{aligned}$$

Now recalling that:

$$Q_* = \int x^2 \mathrm{d}\mu^*(x),$$

one finally has the expression, using equation (196):

$$\begin{aligned}
\mathrm{MSE} &= \int^\lambda x^2 \mathrm{d}\mu^*(x) + \lambda^2 \int_\lambda \mathrm{d}\mu^*(x) \\
&= \int \min(x, \lambda)^2 \mathrm{d}\mu^*(x).
\end{aligned}$$

Finally, the expression of the loss can be also deduced from equation (196) since $\text{Loss}_{\text{train}} = \alpha\xi$ and $\text{MSE} = 2\alpha\xi - \Delta/2$. Again, this leads to the result of Proposition 7. As already mentioned, the fact that we recover the same equations starting from the dynamical equations and the long-time one reveal that both the population ($\alpha \to \infty$) and long-time ($t \to \infty$) limits commute and that the population limit does not introduce a new timescale for the dynamics.

Appendix E. Stability of the Steady-State Solution

We devote this section to the analysis of the stability of the steady-state equations presented in Section 3.2.4. More precisely, we consider a perturbation of the simplified dynamics introduced in Section 3.2.1, and investigate if this perturbation may grow and lead to instability. The plan goes as follows:

- In Section E.1, we first introduce the susceptibility operator associated with the long-time steady-state solution and compute its Frobenius norm. We identify a regime where this quantity slowly diverges as the dimension grows, leading to a potential instability, but only in the high-dimensional limit.
- In Section E.2, we then corroborate our previous observations with a study of the convergence rates associated with the steady-state approximation, in both finite and infinite dimension. We identify some weak directions associated with a vanishing curvature of the Hessian in high dimension.
- In Section E.3, we give the steps necessary for a thorough analysis of the dynamical stability. This could be, in principle, achieved by linearizing the true dynamical equations of Claim 4 around the steady-state solution. This would lead to an exact stability analysis of the high-dimensional dynamical system.

E.1 Susceptibility Operator

We now consider the susceptibility operator associated with the long-time limit of the dynamics:

$$\mathcal{X} = \left.\frac{\partial Z_\infty}{\partial H}\right|_{H=0}.$$

This operator can be viewed as the differential of the map that returns, for a given $H \in \mathcal{S}_d(\mathbb{R})$, the limit Z_∞ of the approximated dynamics with a perturbation H. Following the results of Section D.2.1:

$$Z_\infty = \left(Z^* + \sqrt{\xi}\mathcal{G} - qI_d + H\right)^+_{(m)}.$$

The susceptibility operator encodes how the system responds to perturbations. We can either analyze its spectrum to investigate the stability of each direction of the system, or its normalized Frobenius norm that gives access to a measure of the averaged susceptibility over all directions. Then, the stability criterion we consider is the finiteness of the normalized Frobenius norm.

To compute the spectrum and the Frobenius norm of the susceptibility, one can apply the result of Lemma 36. To do so, denote $\lambda_1, \dots, \lambda_d$ the eigenvalues of $Z^* + \sqrt{\xi}\mathcal{G}$. Then, the spectrum of $\mathcal{X}$, viewed as a linear map $\mathcal{S}_d(\mathbb{R}) \to \mathcal{S}_d(\mathbb{R})$ is given by:

$$\mathrm{Sp}(\mathcal{X}) = \left\{ \frac{\phi(\lambda_i) - \phi(\lambda_j)}{\lambda_i - \lambda_j}, 1 \le i \le j \le d \right\}, \tag{197}$$

with:

$$\phi(x) = (x - q)\mathbf{1}_{x \ge \max(q,\omega)}, \tag{198}$$

and ω is a threshold selecting the m largest eigenvalues of $Z^* + \sqrt{\xi}\mathcal{G}$:

$$\min(m, d) = \sum_{k=1}^{d} \mathbf{1}_{\lambda_k \ge \omega}.$$

For $i = j$, the corresponding eigenvalue of $\mathcal{X}$ is understood as $\phi'(\lambda_i) = \mathbf{1}_{\lambda_i \ge \max(q,\omega)}$. Note that Lemma 36 only applies if the matrix of interest has simple and non-zero eigenvalues. In our case this applies with probability one under the randomness of the Gaussian matrix $\mathcal{G}$, whenever $\xi > 0$.

E.1.1 Norm of the Susceptibility

We start by investigating the Frobenius norm of this susceptibility operator. As a consequence of Lemma 36, we have:

$$\mathcal{R}_d = \frac{1}{d^2}\left\|\mathcal{X}\right\|_F^2 = \frac{1}{d^2} \sum_{1 \le i \le j \le d} \left(\frac{\phi(\lambda_i) - \phi(\lambda_j)}{\lambda_i - \lambda_j} \right)^2. \tag{199}$$

In finite dimension, this quantity is almost surely finite. Let us study it in the high-dimensional limit. We have the deterministic limit:

$$\mathcal{R}_d \xrightarrow[d\to\infty]{} \frac{1}{2} \iint \left(\frac{(x - q)\mathbf{1}_{x \ge \max(q,\omega)} - (y - q)\mathbf{1}_{y \ge \max(q,\omega)}}{x - y} \right)^2 \mathrm{d}\mu_\xi(x)\mathrm{d}\mu_\xi(y) \equiv \mathcal{R}_\infty.$$

Again, ω selects the m largest eigenvalues:

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x).$$

Let us now consider two cases:

- When $q \ge \omega$, the matrix $Z^* + \sqrt{\xi}\mathcal{G}$ has less than m eigenvalues larger than q. In this case, the function ϕ in equation (198) is continuous and 1-Lipschitz, which guarantees the finiteness of $\mathcal{R}_\infty$.
- On the other hand, if $q < \omega$, ϕ is discontinuous at $x = \omega$. Near ω, it jumps from 0 to $\omega - q > 0$. To investigate the behavior of $\mathcal{R}_\infty$, we focus near $x = y = \omega$, where the jump happens. We have, changing variables $x = \omega - u$ and $y = \omega + v$:

$$\mathcal{R}_\infty \ge \frac{1}{2} \iint \rho_\xi(\omega - u)\rho_\xi(\omega + v)\mathbf{1}_{u \ge 0}\mathbf{1}_{v \ge 0} \left(\frac{\omega + v - q}{u + v} \right)^2 \mathrm{d}u\mathrm{d}v.$$

ρ_ξ denotes the density of μ_ξ. Near the points $u, v = 0$, the integrand is proportional to $(u+v)^{-2}$ which is not integrable with respect to u, v. Therefore, as soon as μ_ξ has positive mass on both sides of ω, the quantity $\mathcal{R}_\infty$ is infinite. Finally, we claim that this situation is generic: unless the density ρ_ξ splits into two parts of mass κ and $1-\kappa$, ρ_ξ will have positive mass on both sides of ω, leading to the divergence of $\mathcal{R}_\infty$.

Therefore, it is natural to understand the typical value of $\mathcal{R}_d$ as $d \to \infty$. To do so, we start from equation (199) and focus on the eigenvalues pairs (λ_i, λ_j) where the divergence happens. This corresponds to eigenvalues close to the threshold ω on each side of it. At this point, the spacing between two eigenvalues is of order $\epsilon_d \propto d^{-1}$. Taking the continuous approximation with the cutoff ϵ_d, we get that:

$$\begin{aligned}\mathcal{R}_d &\underset{d\to\infty}{\sim} \frac{1}{2}\rho_\xi(\omega)^2(\omega-q)^2 \iint \mathbf{1}_{u\geq 0}\mathbf{1}_{v\geq 0}\mathbf{1}_{|u-v|\geq\epsilon_d}\frac{\mathrm{d}u\mathrm{d}v}{(u+v)^2} \\ &\underset{d\to\infty}{\sim} \frac{1}{2}\rho_\xi(\omega)^2(\omega-q)^2\log d.\end{aligned}$$

Therefore, in the case $q < \omega$, the typical value of $\mathcal{R}_d$ is of order $\log d$ as $d \to \infty$, corresponding to a mild divergence. This invites us to study the spectrum of the susceptibility in order to characterize the directions of potential instability.

Finally, to relate this conclusion to Sections 3.2.3, 3.2.4, recall the definition of $\kappa_{\min}$ in equation (41). Then, the region $q \geq \omega$ corresponds to the case where the gradient flow dynamics converges to a rank-deficient matrix. In Section 3.2.3, we have precisely identified this region to be the overparameterized one, that is $\kappa \geq \kappa_{\min}$.

E.1.2 Spectrum of the Susceptibility

Recall that the spectrum of the susceptibility $\mathcal{X}$ is given in equation (197). As discussed earlier, the unstable modes correspond to pairs of eigenvalues (λ_i, λ_j) close and on each side of the cutoff ω. For such pairs such that $\lambda_i < \omega < \lambda_j$ and $\lambda_j - \lambda_i \propto \epsilon$, the associated eigenvalue for the susceptibility operator is:

$$\frac{\lambda_j - q}{\lambda_j - \lambda_i} \propto \frac{1}{\epsilon}.$$

As the dimension increases, there are $\Theta(\epsilon^2 d^2)$ pairs with amplitude larger than ϵ^{-1}. The most unstable directions correspond to the scaling $\epsilon \propto d^{-1}$ with a susceptibility eigenvalue proportional to d. However, these directions are only in a finite number. Combining all the scales from d^{-1} to order 1 leads to the divergence proportional to $\log d$.

Eigenvalue jump. Recall that the gradient flow selects the m largest positive eigenvalues of the target matrix $A = Z^* + \sqrt{\xi}\mathcal{G} - qI_d$. Letting p be the number of positive eigenvalues of this matrix, we have the dichotomy:

- In the stable case, when $m \geq p$, the student has enough rank to select all the positive eigenvalues. In this case any infinitesimal perturbation leads to a response of the same order, indicating stability.
- For $m < p$, the rank constraint imposes that several order one eigenvalues are sent to zero. Therefore, for very close eigenvalues, a perturbation can reorder the eigenvalues

and change those that are selected by the dynamics: the potential instability originates from the fact that some previously selected eigenvalues (of order one) become zero, and *vice versa*.

As it was made clear in Section E.1.1, these two cases respectively correspond to the overparameterized ($\kappa \geq \kappa_{\min}$) and underparameterized ($\kappa < \kappa_{\min}$) regions identified in Section 3.2.3.

E.1.3 A Refined Analysis

Despite the previous observations, simulations suggest that the gradient descent algorithm (with small stepsize) always converges toward the solution given by Claim 6, even in what we called the unstable region. As a first attempt to explain this, consider Z_∞ to be perturbed by some matrix H, and the associated response:

$$\frac{1}{d}\big\|\delta Z_\infty\big\|_F^2 = \frac{1}{d}\sum_{1\leq i\leq j\leq d}\left(\frac{\phi(\lambda_i)-\phi(\lambda_j)}{\lambda_i-\lambda_j}\right)^2 H_{ij}^2.$$

The key point here is that we do not ask the system to be stable under any perturbation, but specific perturbations originating from the dynamics of Claim 4. For instance, the previous equation is written in the basis that diagonalizes the susceptibility $\mathcal{X}$, which is directly related to the one that diagonalizes the matrix $Z^* + \sqrt{\xi}\mathcal{G}$. Now, it is easily shown that if the perturbation H is generic with respect to this basis, the previous quantity is related to the Frobenius norm of the susceptibility (199) and may diverge in the unstable region.

However, a perturbation originating from the dynamics of Claim 4 should necessarily be correlated with the simplified dynamics. Then, the stability of the system would be guaranteed provided that this correlated perturbation puts slightly less weight on the unstable modes. For instance, one could imagine a high-dimensional behavior:

$$\mathbb{E}\, H_{ij}^2 \approx \frac{\epsilon^2}{d} V(\lambda_i, \lambda_j),$$

where ϵ is the magnitude of the perturbation. We would then get:

$$\frac{1}{d}\big\|\delta Z_\infty\big\|_F^2 \xrightarrow[d\to\infty]{} \frac{\epsilon^2}{2}\iint V(x,y)\left(\frac{(x-q)\mathbf{1}_{x\geq\max(q,\omega)} - (y-q)\mathbf{1}_{y\geq\max(q,\omega)}}{x-y}\right)^2 \mathrm{d}\mu_\xi(x)\mathrm{d}\mu_\xi(y).$$

Now, a mild decay of V around the point (ω,ω) would guarantee the finiteness of the integral, therefore the stability of the system. On the other side, it is also possible that H puts more weight on the unstable modes and amplifies the perturbation.

In line with our numerical observations, we conjecture that this specific perturbation tends to regularize the dynamical system, i.e., it attenuates the weak modes associated with a potential instability of the approximate dynamics.

E.2 Convergence Rates for the Approximate Dynamics

In this part, we analyze the convergence rates of the approximate dynamics of Section 3.2.1:

$$\dot{W}(t) = 2r_\infty\Big(Z^* + \sqrt{\xi}\mathcal{G} - qI_d - W(t)W(t)^\top\Big)W(t). \tag{200}$$

The aim is twofold:

- Interpret the weak or unstable directions of the susceptibility operator in terms of convergence rates and landscape curvature.
- Examine Assumption 5 in light of the discussion in Section D.1, that relates the steady-state assumption with the fast convergence of the dynamics.

We build on the results for the Oja flow dynamics derived in Section 4.2 and Section 4.3 that derive its convergence rates in both finite and infinite dimension. To match with these results, we define:

$$A = Z^* + \sqrt{\xi}\mathcal{G} - qI_d, \tag{201}$$

the target matrix of the approximate dynamics (200). As soon as $\xi > 0$, the presence of the GOE matrix $\mathcal{G}$ guarantees that A is invertible and has simple eigenvalues with probability one. Moreover, we denote by p the number of positive eigenvalues of A. Now, the previous section has shown that the stability of the approximate dynamics is guaranteed whenever $m \geq p$ (overparameterized regime), but it is still to be clarified in the small-rank regime $m < p$.

Finally, note that the presence of the factor $2r_\infty$ in the dynamics acts as a time renormalization and only alters the convergence rates by a constant factor (independent of the dimension). For simplicity, we set this quantity to 1 in the following.

E.2.1 Finite-Dimensional Rates

As shown in Proposition 20 and Proposition 21, when the dimension remains finite, the Oja flow converges exponentially fast. Denoting $\lambda_1 > \cdots > \lambda_d$ the ordered eigenvalues of A in equation (201), the convergence rates of the dynamics (200) are given by:

$$\varrho_{\mathrm{CV}} = \begin{cases} \min(2\lambda_m, \lambda_m - \lambda_{m+1}), & \text{if } m \leq p, \\ \min(\lambda_p, |\lambda_{p+1}|), & \text{if } m > p, \end{cases}$$

in the sense that for all $c < \varrho_{\mathrm{CV}}$:

$$\big\|Z(t) - Z_\infty\big\|_F \underset{t\to\infty}{=} o\big(e^{-ct}\big).$$

Here $Z(t) = W(t)W(t)^\top$ and Z_∞ is the limit of $Z(t)$ as $t \to \infty$.

We believe that an exponentially fast convergence should validate Assumption 5. This result also matches with the conclusion of Section E.1 that the steady-state approximation remains stable in finite dimension, no matter the value of κ.

It is interesting to notice that in both cases, these rates vanish with the dimension as soon as the asymptotic spectral distribution of A has mass either near zero (when $m > p$) or near its m^{th} eigenvalue λ_m (when $m \leq p$). In both cases the convergence rates are of order d^{-1} when d grows large. This observation was verified for generic values of our parameters (for instance $\kappa > \kappa^*$) with the values of q, ξ obtained from the numerical integration of the system (37). In most cases, the spectral density of A possesses non-zero mass around λ_m for $m \leq p$ and 0 for $m > p$.

Therefore, there exists a few directions in the space $\mathbb{R}^{d\times m}$ associated with a slow relaxation time. In terms of landscape, these directions become flatter as the dimension

increases. Interestingly, this phenomenon happens in both the stable and unstable cases derived in Section E.1. As already mentioned, these two regions precisely correspond to the ones identified in Sections 3.2.3, 3.2.4.

E.2.2 Infinite-Dimensional Rates

Following the previous observations in finite dimension, we also study the convergence rates of the Oja flow by first letting the dimension go to infinity. As a consequence of the presence of directions with vanishing curvature, our analysis leads to the conclusion of non-exponential rates. More precisely, in Proposition 23, we show that, with the same notations as in the previous section:

$$\lim_{t\to\infty}\lim_{d\to\infty}\frac{1}{d}\big\|Z(t)-Z_\infty\big\|_F^2 \underset{t\to\infty}{=} \begin{cases} \Theta\big(t^{-3}\big), & \text{if } \kappa>\kappa_A,\\ \Theta\big(t^{-1}\big), & \text{if } \kappa<\kappa_A, \end{cases} \tag{202}$$

where:

$$\kappa_A = \int \mathbf{1}_{x>0}\mathrm{d}\mu_A(x),$$

is the fraction of positive eigenvalues of A. With our expression for A, the threshold κ_A coincides with $\kappa_{\min}$ identified in Section 3.2. This value separates two regimes with distinct convergence properties. In comparison with our stability result, the region we identified as stable corresponds to the fastest convergence rates, namely $\kappa > \kappa_A$. The potentially unstable region, $\kappa < \kappa_A$ is characterized by the slowest convergence rates.

When convergence is slow, perturbations decay over longer timescales and therefore may interact more strongly with unstable modes, leading to instability. On the other hand, faster convergence limits this effect by damping perturbations more rapidly.

Interestingly, these rates can be compared to those obtained from the gradient descent dynamics (6). In Figure 15, we plot the function:

$$\mathrm{Loss}_\lambda = \frac{1}{4n}\sum_{k=1}^{n}\Big(\mathrm{Tr}\big(X_kWW^\top\big)-z_k\Big)^2+\frac{\lambda}{d}\mathrm{Tr}\big(WW^\top\big). \tag{203}$$

Since the dynamics (5) is exactly the gradient flow associated with the loss function Loss_λ, this quantity is known to decrease over time along the flow. Figure 15 is split into two panels, separating values of α corresponding to the stable region $\kappa > \kappa_{\min}$ and the potentially unstable region $\kappa < \kappa_{\min}$. The observed convergence rates are then compared with the power-law asymptotic $t \mapsto t^{-3}$. Interestingly, in both regions the convergence rates are very close to this asymptotic behavior. Therefore, in terms of convergence rates alone, there is no clear difference between the stable and unstable regions. Since we conjectured in Section D.1 that the stability of the steady-state solution is directly related to the convergence rates of the dynamics, this numerical observation provides evidence supporting the stability of the underparameterized region $\kappa < \kappa_{\min}$.

As a technical remark, we emphasize that the cubic power-law behavior identified in Figure 15 is not directly related to the one in equation (202). Indeed, equation (202) concerns the distance to convergence, whereas Figure 15 displays the loss optimized during gradient flow. A more accurate comparison would require extending the convergence rate result of Proposition 23 to the high-dimensional limit of the loss function. We expect this analysis to be significantly more involved and leave it for future work.

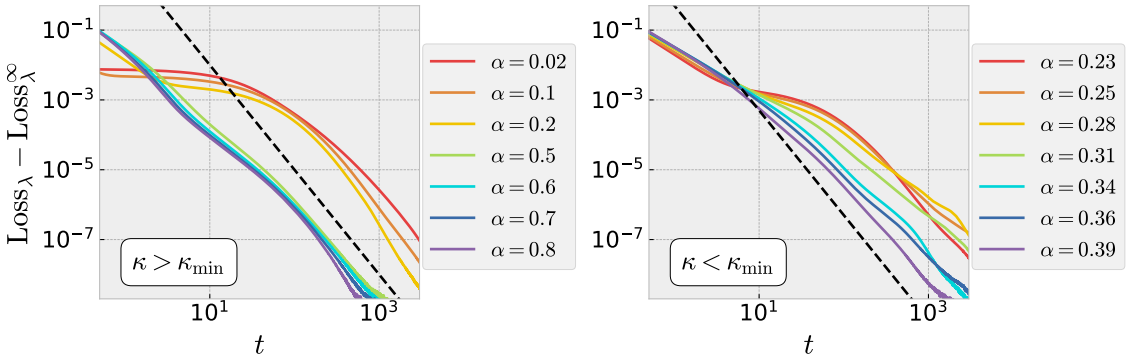

Figure 15: The function Loss_λ (defined in equation 203), with limiting value subtracted, as a function of time t for gradient descent trajectories (see equation 6). Parameters $\kappa = 0.4, \kappa^* = 0.3, \lambda = 0.01$, $\Delta = 0$. Values of α are chosen so that $\kappa > \kappa_{\min}$ on the left panel and $\kappa < \kappa_{\min}$ on the right panel. Black dashed line: the function $t \mapsto Ct^{-3}$, where C is chosen for visual comparison. Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data.

E.3 Dynamical Stability

To conclude this section on the stability of the dynamics, we outline an approach that would be worth investigating in order to characterize potential instabilities of the steady-state dynamics. Earlier in this section, we analyzed the stability of the steady-state solution under generic and worst-case perturbations. Here, we aim to go beyond this analysis by taking into account the original set of equations in Claim 4. More precisely, we decompose these equations into a steady-state solution and the exact associated perturbation, and then linearize the dynamics around the steady state for this specific perturbation. This type of approach is not new and has already been applied to simpler settings, such as spin-glass models (Sompolinsky and Zippelius, 1982; Crisanti et al., 1993).

We only sketch the main steps that would lead to such a result. While the calculation can, in principle, be carried out, it quickly becomes technically involved. For this reason, we leave a detailed implementation of this approach for future work.

Recall the high-dimensional equations of Claim 4. With the ℓ_2-regularization, $W(t)$ is solution of the dynamics:

$$\dot{W}(t) = 2\left(\int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - Z(t')\Big)\mathrm{d}t'\right)W(t) - 2\lambda W(t). \tag{204}$$

The covariance of the Gaussian process $\mathcal{G}$ and the response kernel R are directly related to averages with respect to the dynamics, in equations (25) and (27). From now on, we denote:

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C(t,t'), \tag{205}$$

E.3.1 Linearization of the Dynamics

We start by linearizing the dynamical equation (204) around the steady-state solution. To do so, we introduce the functions $\delta R, \delta C$ such that:

$$R(t,t') = r_\infty \delta(t-t') + \delta R(t,t'), \qquad C(t,t') = \xi + \delta C(t,t').$$

The first terms $r_\infty\delta(t-t')$ and $\xi = \lim_{t,t'\to\infty} C(t,t')$ correspond to the approximation of the response kernel and the noise covariance in the steady-state approximation. Then, the dynamics on $W(t)$ can be rewritten as:

$$\begin{aligned}
\dot W(t) &= 2r_\infty \left(Z^* + \sqrt{\xi}\mathcal{G} - \frac{\lambda}{r_\infty} I_d - W(t)W(t)^\top \right) W(t) + 2r_\infty H(t) W(t), \\
H(t) &= V(t) - \frac{1}{r_\infty}\int_0^t \delta R(t,t') \Big(\sqrt{\xi}\mathcal{G} + Z^* - Z(t') \Big) \mathrm{d}t',
\end{aligned}$$

where V is a centered Gaussian process taking values in the space of symmetric matrices, with covariance:

$$\mathbb{E}\, V_{ij}(t) V_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\delta C(t,t').$$

The key point of our analysis is to consider $H(t)$ as a perturbation and linearize the solution $Z(t) = W(t)W(t)^\top$ around $H(t)$:

$$\begin{aligned}
Z(t) = Z_0(t) &+ 2r_\infty \int_0^t \left.\frac{\partial Z_0(t)}{\partial H(t')}\right|_{H=0} \Big(H(t') \Big) \mathrm{d}t' \\
&+ 2r_\infty^2 \int_0^t \int_0^t \left.\frac{\partial^2 Z_0(t)}{\partial H(t')\partial H(t'')}\right|_{H=0} \Big(H(t'), H(t'') \Big) \mathrm{d}t' \mathrm{d}t'' + O\big(\|H\|^3\big),
\end{aligned} \tag{206}$$

where $Z_0(t) = W_0(t)W_0(t)^\top$ and W_0 is solution of the dynamics:

$$\dot W_0(t) = 2r_\infty \left(Z^* + \sqrt{\xi}\mathcal{G} - \frac{\lambda}{r_\infty} I_d - W_0(t)W_0(t)^\top \right) W_0(t). \tag{207}$$

Using equations (25) and (205), we can rewrite the covariance $C(t,t')$:

$$C(t,t') = \frac{1}{2\alpha}\left(\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big((Z(t) - Z^*)(Z(t') - Z^*) \Big) + \frac{\Delta}{2} \right).$$

Plugging the solution (206) into the expression of $C(t,t')$ will allow to obtain a self-consistent expression of δC. Since $H(t)$ involves the Gaussian process $V(t)$ whose covariance is proportional to δC, it is required that we study this perturbation up to second order.

The same can be done for the response $R(t,t')$. Equation (27) shows that this function is directly related to the response of Z, therefore one can directly differentiate equation (206) with respect to an external field perturbing the dynamics.

Ultimately, it is possible to identify a 2×2 matrix $K(t,t',s,s')$ such that, after linearizing the dynamics, we have:

$$\begin{pmatrix} \delta C(t,t') \\ \delta R(t,t') \end{pmatrix} = \begin{pmatrix} C_0(t,t') - \xi \\ (1-r_\infty)\delta(t-t') - \alpha^{-1} R_Z^0(t,t') \end{pmatrix} + \int \mathrm{d}s\mathrm{d}s'\, K(t,t',s,s') \begin{pmatrix} \delta C(s,s') \\ \delta R(s,s') \end{pmatrix}, \tag{208}$$

where C_0, R_Z^0 are the correlation and response computed from the steady-state solution. In addition, the kernel K only involves functions computed from this approximate solution. This is a consequence of the linearization of the dynamics around this solution. More precisely, we can show that K depends on the response kernels of the dynamics (207) up to order three:

$$\mathcal{R}_1(t,t') = \left.\frac{\partial Z_0(t)}{\partial H(t')}\right|_{H=0}, \qquad \mathcal{R}_2(t,t',t'') = \left.\frac{\partial^2 Z_0(t)}{\partial H(t')\partial H(t'')}\right|_{H=0},$$

and likewise for $\mathcal{R}_3$.

E.3.2 Response Kernels of the Oja Flow

The dynamics that describes our steady-state approximation is nonlinear, but it is possible to compute its response kernels of any order. We briefly explain how this can be done, and refer to Section 4.4 for a complete derivation of the first-order response. We consider the dynamics:

$$\dot{W}(t) = 2r_\infty\Big(A - W(t)W(t)^\top + \epsilon H(t)\Big)W(t).$$

For simplicity we drop the factor $2r_\infty$, that can be simply recovered by a time reparameterization. We decompose the solution $Z(t) = W(t)W(t)^\top$ in powers of ϵ:

$$Z(t) = Z_0(t) + \epsilon Z_1(t) + \cdots + \epsilon^n Z_n(t) + \dots.$$

Plugging this into the equation for $W(t)$, we get the equations for each order:

$$\dot{Z}_0(t) = AZ_0(t) + Z_0(t)A - 2Z_0(t)^2,$$

$$\dot{Z}_n(t) = B(t)Z_n(t) + Z_n(t)B(t) + \underbrace{H(t)Z_{n-1}(t) + Z_{n-1}(t)H(t) - 2\sum_{k=1}^{n-1} Z_k(t)Z_{n-k}(t)}_{\equiv M_n(t)},$$

with $B(t) = A - 2Z_0(t)$, and with initial conditions $Z_n(0) = 0$ for $n \geq 1$. The dynamics on Z_n is linear but is driven by the time-dependent matrix $B(t)$. To solve this, let us introduce $P(t)$ as the solution of $\dot{P}(t) = -B(t)P(t)$ with $P(0) = I_d$. Then, for $n \geq 1$, it is easily seen that:

$$Z_n(t) = P(t)^{-1}\int_0^t P(s)M_n(s)P(s)^\top \mathrm{d}s\, P(t)^{-\top}. \tag{209}$$

Now, as already been done in Section I.7, the matrix $P(t)$ can be computed explicitly and writes:

$$P(t) = e^{-tA} + Z_0A^{-1}\big(e^{tA} - e^{-tA}\big).$$

From equation (209), this leads to an explicit recursive expression of $Z_n(t)$. It is easily shown by recursion that $Z_n(t)$ is a homogeneous polynomial of degree n in H, and therefore the n^{th}-order response kernel of the Oja flow can be directly computed from Z_n. In Proposition 24, we give an explicit computation for the case $n = 1$. Then, it is possible to iterate: plug in the solution $Z_1(t)$ into the expression of $Z_2(t)$ and compute the second-order response, and so on. However, the calculations rapidly become heavy and it is not clear if one can practically obtain an expression of the third-order response, which is necessary to compute the kernel K.

E.3.3 Stability Criterion

Once the response kernels are computed, it is possible to go back to equation (208) and either let the dimension grow to infinity to test the stability in the high-dimensional limit, or leave it fixed. In any case, it is practical to look for a solution of the form:

$$X(t,t') \equiv \begin{pmatrix} \delta C(t,t') \\ \delta R(t,t') \end{pmatrix} = e^{\sigma T} \begin{pmatrix} \delta C_0(\tau) \\ \delta R_0(\tau) \end{pmatrix}, \qquad T = \frac{t+t'}{2}, \;\; \tau = t - t'.$$

This separates variables between the time-translational invariant part and a potential exponential growth as $T \to \infty$. Using that the kernel all originates from a steady-state solution (the approximate dynamics), we can integrate with respect to T, which should lead to an equation of the form:

$$\sigma \begin{pmatrix} \delta C_0(\tau) \\ \delta R_0(\tau) \end{pmatrix} = \sigma \epsilon(\tau) + \int_0^\infty \mathcal{K}(\tau, \tau') \begin{pmatrix} \delta C_0(\tau') \\ \delta R_0(\tau') \end{pmatrix} \mathrm{d}\tau'.$$

Then, a sizable challenge would be to compute the spectrum of the operator $\mathcal{K}$ (that is computed from K), and stability would be guaranteed if its eigenvalues are all with negative real part. However, it is not clear if such a calculation is possible.

Appendix F. Analysis of Langevin Dynamics

In this section we go beyond the gradient flow setting and analyze the Langevin dynamics, with $\beta < \infty$. We start from the dynamical equations in Claim 4 and study the stochastic evolution of the matrix $W(t)$. More precisely, under simplifying assumptions on the dynamics at long times, we derive the stationary measure of this process. Here, by stationary measure we mean the probability distribution reached by the process in the long-time limit. The plan of the section goes as follows:

- In Section F.1, we consider a stochastic evolution similar to equation (23) and derive its stationary measure.
- In Section F.2, we build on this result to derive the stationary measure of the dynamics on $W(t)$, leading to the equations in Claim 8. This result requires several assumptions on the dynamics (time-translational invariance and fluctuation–dissipation) that we detail in Assumption 43.
- In Section F.3, under similar assumptions, we derive the stationary measure for the typical label, whose expression is given in equation (24).
- In Section F.4, we study the zero-temperature ($\beta \to \infty$) limit of the stationary measure previously derived. At positive regularization, we show that we recover the results obtained in the gradient flow setting.
- In Section F.5, we show that an appropriate choice of the inverse temperature β leads to the Bayes-optimal equations of Maillard et al. (2024).

F.1 Derivation of a Stationary Measure

In this section, we introduce an auxiliary stochastic differential equation and derive its stationary measure using a Gaussian coupling.

F.1.1 Summary of the Result

In the following, we consider the stochastic differential equation on $W(t) \in \mathbb{R}^{d\times m}$:

$$\begin{aligned} \mathrm{d}W(t) = 2\left[\int_0^t R(t-t')\Big(\mathcal{H}(t') + A - W(t')W(t')^\top\Big)\mathrm{d}t'\right]W(t)\mathrm{d}t \\ - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \end{aligned} \tag{210}$$

where $A \in \mathcal{S}_d(\mathbb{R})$ is a fixed matrix, $\Omega\colon \mathbb{R}^{d\times m} \to \mathbb{R}^+$ is a continuously differentiable regularization, B is a standard Brownian motion over $\mathbb{R}^{d\times m}$, and $\mathcal{H}$ is a Gaussian process on $\mathcal{S}_d(\mathbb{R})$, independent of B, with zero mean and stationary covariance:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C(t-t').$$

Moreover, we assume that:

- The stationary covariance C can be expressed as:

  $$C(t) = \int_0^\infty \phi(\theta)e^{-\theta|t|}\mathrm{d}\theta,$$

  for some $\phi\colon \mathbb{R}^+ \to \mathbb{R}^+$ and one has $C(t) \xrightarrow[t\to\infty]{} 0$.

- The functions R, C satisfy the fluctuation–dissipation relation:

  $$R(t) - 4\beta\int_0^t C'(t-t')R(t')\mathrm{d}t' = \delta(t), \tag{211}$$

  where δ is the Dirac delta distribution supported at zero.

When these assumptions are satisfied, we show that the stochastic equation given in equation (210) has stationary measure:

$$\mathbb{P}_\beta(W) \propto \exp\left(-\frac{\beta d}{1+4\beta C(0)}\left\|WW^\top - A\right\|_F^2 - 2\beta d\,\Omega(W)\right). \tag{212}$$

For the following, it will be interesting to note that the fluctuation–dissipation relation in equation (211) implies the equality:

$$\int_0^\infty R(t)\mathrm{d}t = \frac{1}{1+4\beta C(0)}.$$

The following sections are dedicated to the derivation of this result. More precisely, we introduce in the following section a Gaussian coupling that allows to derive the stationary measure in equation (212) from standard results on Langevin dynamics.

F.1.2 Gaussian Coupling

In this part, we consider the following potential on $W \in \mathbb{R}^{d\times m}$ and the symmetric matrices $\mathcal{Q}_1, \dots, \mathcal{Q}_K \in \mathcal{S}_d(\mathbb{R})$:

$$U_{\text{coupling}}\Big(W, \big\{\mathcal{Q}_k\big\}\Big) = \left\| WW^\top - A - \sum_{k=1}^K g_k \mathcal{Q}_k \right\|_F^2 + \sum_{k=1}^K \theta_k \big\|\mathcal{Q}_k\big\|_F^2 + 2\Omega(W), \tag{213}$$

where $A \in \mathcal{S}_d(\mathbb{R})$ is a fixed matrix and Ω is a smooth and coercive regularization. While we are mainly interested in W, we introduce the matrices $\mathcal{Q}_1, \dots, \mathcal{Q}_K$ as auxiliary variables that couple to W.

The goal is to study a Langevin dynamics on the matrices $W, \big\{\mathcal{Q}_k\big\}$ such that:

- The stationary distribution of the dynamics is given by the Boltzmann–Gibbs distribution associated with the potential U_{coupling}, namely:

$$\mathbb{P}_{\text{coupling}}\big(W, \{\mathcal{Q}_k\}\big) \propto \exp\left(-\beta d\, U_{\text{coupling}}\Big(W, \big\{\mathcal{Q}_k\big\}\Big)\right), \tag{214}$$

  for some inverse temperature $\beta > 0$. Remark that the potential U_{coupling} is quadratic in the $\{\mathcal{Q}_k\}$, so that these variables are Gaussian when drawn from $\mathbb{P}_{\text{coupling}}$.

- The dynamics for the auxiliary variables $\{\mathcal{Q}_k\}$ can be solved explicitly, and leads to an effective stochastic dynamics for W that is equivalent to the dynamics (210).

To achieve this, we consider the stochastic differential equations:

$$\begin{aligned} \mathrm{d}W(t) &= -\frac{1}{2}\nabla_W U_{\text{coupling}}\Big(W(t), \big\{\mathcal{Q}_k(t)\big\}\Big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \\ \mathrm{d}\mathcal{Q}_k(t) &= -\frac{1}{2}\nabla_{\mathcal{Q}_k} U_{\text{coupling}}\Big(W(t), \big\{\mathcal{Q}_k(t)\big\}\Big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}\Xi_k(t), \end{aligned} \tag{215}$$

where B is a standard Brownian motion on $\mathbb{R}^{d\times m}$, and $\Xi_1, \dots, \Xi_K$ are independent (and independent of B) standard Brownian motions over $\mathcal{S}_d(\mathbb{R})$. As U_{coupling} is confining, it is well known (see for instance Pavliotis, 2014, Proposition 4.2) that the dynamics (215) admits a unique stationary distribution, given by (214).

We now analyze the Langevin dynamics (215). Computing the gradients, we arrive at:

$$\mathrm{d}W(t) = -2\Big(W(t)W(t)^\top - A - \mathcal{Q}(t)\Big)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{216}$$

$$\mathrm{d}\mathcal{Q}_k(t) = -\Big(\theta_k \mathcal{Q}_k(t) + g_k \mathcal{Q}(t) - g_k\Big(W(t)W(t)^\top - Z^*\Big)\Big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}\Xi_k(t), \tag{217}$$

$$\mathcal{Q}(t) = \sum_{k=1}^K g_k \mathcal{Q}_k(t). \tag{218}$$

Equation (217) can be integrated, leading to the expression of $\mathcal{Q}_k(t)$ depending on $\mathcal{Q}(t)$ and $W(t)$:

$$\begin{aligned} \mathcal{Q}_k(t) = e^{-\theta_k t}\mathcal{Q}_k(0) - g_k\int_0^t e^{-\theta_k(t-t')}\mathcal{Q}(t')\mathrm{d}t' + g_k\int_0^t e^{-\theta_k(t-t')}\Big(W(t')W(t')^\top - A\Big)\mathrm{d}t' \\ + \frac{1}{\sqrt{\beta d}}\int_0^t e^{-\theta_k(t-t')}\mathrm{d}\Xi_k(t'). \end{aligned}$$

Plugging this expression into equation (218), we obtain the self-consistent expression of $\mathcal{Q}$:

$$\mathcal{Q}(t) = \sum_{k=1}^{K} g_k e^{-\theta_k t}\mathcal{Q}_k(0) - \int_0^t \left(\sum_{k=1}^{K} g_k^2 e^{-\theta_k(t-t')}\right)\Big(\mathcal{Q}(t') + A - W(t')W(t')^\top\Big)\mathrm{d}t' + \frac{1}{\sqrt{\beta d}}\sum_{k=1}^{K} g_k \int_0^t e^{-\theta_k(t-t')}\mathrm{d}\Xi_k(t').$$

To solve for $\mathcal{Q}$, we introduce the following scalar function and random process:

$$\Gamma(t) = \sum_{k=1}^{K} g_k^2 e^{-\theta_k t}, \tag{219}$$

$$\mathcal{H}(t) = \sum_{k=1}^{K} g_k e^{-\theta_k t}\mathcal{Q}_k(0) + \frac{1}{\sqrt{\beta d}}\sum_{k=1}^{K} g_k \int_0^t e^{-\theta_k(t-t')}\mathrm{d}\Xi_k(t'), \tag{220}$$

so that we have the self-consistent expression for $\mathcal{Q}$:

$$\mathcal{Q}(t) = -\int_0^t \Gamma(t-t')\Big(\mathcal{Q}(t') + A - W(t')W(t')^\top\Big)\mathrm{d}t' + \mathcal{H}(t).$$

We can write $\mathcal{Q}$ explicitly by introducing R such that:

$$R(t) + \int_0^t \Gamma(t-t')R(t')\mathrm{d}t' = \delta(t). \tag{221}$$

As a consequence, $\mathcal{Q}$ can be expressed as:

$$\mathcal{Q}(t) = W(t)W(t)^\top - A - \int_0^t R(t-t')\Big(W(t')W(t')^\top - A - \mathcal{H}(t')\Big)\mathrm{d}t'.$$

Then, plugging this expression into the dynamics on W in equation (216), we precisely obtain the stochastic equation (210).

F.1.3 Covariance of the Noise

We now study the random process $\mathcal{H}(t)$ defined in equation (220). Provided that the initializations $\{\mathcal{Q}_k^0\}$ are Gaussian, $\mathcal{H}(t)$ is a matrix-valued Gaussian process. More precisely, we assume that the $\{\mathcal{Q}_k^0\}$ are Gaussian matrices with zero mean and a covariance of the form:

$$\mathbb{E}\left(\mathcal{Q}_k^0\right)_{ij}\left(\mathcal{Q}_{k'}^0\right)_{i'j'} = \frac{\xi_k^0}{d}\delta_{kk'}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big).$$

Then, since $\Xi_1, \dots, \Xi_K$ are independent Brownian motions over $\mathcal{S}_d(\mathbb{R})$, one has the expression for the covariance of $\mathcal{H}$:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\sum_{k=1}^{K} g_k^2 e^{-\theta_k(t+t')}\left(\xi_k^0 + \frac{1}{4\theta_k\beta}\big(e^{2\theta_k \min(t,t')} - 1\big)\right).$$

This covariance has a transient part that vanishes as $t, t' \to \infty$, and a time-translational invariant part which is only a function of $t - t'$. Since we are only interested in the dynamics of $W(t)$ at long times, and the initialization of the $\{\mathcal{Q}_k\}$ is not relevant, we choose the variance:

$$\xi_k^0 = \frac{1}{4\theta_k \beta}.$$

In this case the covariance of $\mathcal{H}$ is time-translational invariant:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{4\beta d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\sum_{k=1}^{K}\frac{g_k^2}{\theta_k}e^{-\theta_k|t-t'|}.$$

We then consider the function:

$$C(\tau) = \frac{1}{4\beta}\sum_{k=1}^{K}\frac{g_k^2}{\theta_k}e^{-\theta_k|\tau|}, \tag{222}$$

so that we have the fluctuation–dissipation relation between C and the function Γ, defined in equation (219):

$$-C'(\tau) = \frac{1}{4\beta}\Gamma(\tau),$$

for $\tau > 0$. Thanks to the relationship between R and Γ in equation (221), we arrive at the fluctuation–dissipation relation:

$$R(t) - 4\beta\int_0^t C'(t-t')R(t')\mathrm{d}t' = \delta(t). \tag{223}$$

This relationship fully characterizes R. In addition, equation (222) can be interpreted as the discretization (at finite K) of the continuous representation:

$$C(\tau) = \int_0^\infty \phi(\theta)e^{-\theta|\tau|}\mathrm{d}\theta,$$

for $\phi\colon \mathbb{R}^+ \to \mathbb{R}^+$. Combining this remark with the fluctuation–dissipation relation (223), this leads to the result given in Section F.1.1.

F.1.4 Stationary Measure

We shall now derive the stationary measure given in equation (212). Recall that we integrated the variables $\{\mathcal{Q}_k\}$ in the Langevin dynamics (215) to obtain an equation solely on $W(t)$. Now, to access the stationary measure on W itself, we apply the same procedure and average the joint distribution (214) with respect to the $\{\mathcal{Q}_k\}$:

$$\mathbb{P}_\beta(W) \propto \int \exp\Big(-\beta d\, U_{\text{coupling}}\Big(W, \big\{\mathcal{Q}_k\big\}\Big)\Big)\,\mathrm{d}\mathcal{Q}_1 \ldots \mathrm{d}\mathcal{Q}_K.$$

Due to the expression of the potential U_{coupling}, we define:

$$\mathcal{M} = \begin{pmatrix} g_1\big(WW^\top - A\big) \\ \vdots \\ g_K\big(WW^\top - A\big) \end{pmatrix} \in \mathcal{S}_d(\mathbb{R})^K, \qquad \mathcal{T} = \mathrm{diag}\Big(\theta_1 I_d, \ldots, \theta_K I_d\Big) + \Big(g_k g_l I_d\Big)_{1\le k,l\le K},$$

where $\mathcal{T}$ is viewed as a linear map on $\mathcal{S}_d(\mathbb{R})^K$. Then, denoting $\langle\,,\,\rangle$ the Euclidean inner product on $\mathcal{S}_d(\mathbb{R})^K$, we have:

$$\begin{aligned}\mathbb{P}_\beta(W) &\propto \exp\Big(-\beta d\big\|WW^\top - A\big\|_F^2 - 2\beta d\,\Omega(W)\Big)\\ &\qquad\qquad \int_{\mathcal{S}_d(\mathbb{R})^K} \mathrm{d}\mathcal{Q}\,\exp\Big(2\beta d\,\langle \mathcal{M},\mathcal{Q}\rangle - \beta d\,\langle \mathcal{Q},\mathcal{T}(\mathcal{Q})\rangle\Big) \qquad (224)\\ &\propto \exp\Big(-\beta d\big\|WW^\top - A\big\|_F^2 - 2\beta d\,\Omega(W)\Big)\exp\Big(\beta d\,\langle \mathcal{M},\mathcal{T}^{-1}(\mathcal{M})\rangle\Big).\end{aligned}$$

We computed the Gaussian integral with respect to $\mathcal{Q}$ and ignored the term proportional to $\det(\mathcal{T})$ that is independent of W. In the following, we view $\mathcal{S}_d(\mathbb{R})^K$ as the tensor product $\mathbb{R}^K \otimes \mathcal{S}_d(\mathbb{R})$. This induces a tensor structure for the space of linear maps on $\mathcal{S}_d(\mathbb{R})^K$. With this identification, we write:

$$\mathcal{T} = \big(\Theta + gg^\top\big)\otimes \mathrm{Id}, \qquad\qquad \mathcal{M} = g\otimes\big(WW^\top - A\big).$$

where Id denotes the identity map on the space of symmetric matrices, $g = \big(g_1,\dots,g_K\big)^\top \in \mathbb{R}^K$ and $\Theta = \mathrm{diag}(\theta_1,\dots,\theta_K)\in\mathbb{R}^{K\times K}$. Therefore, we have:

$$\langle \mathcal{M},\mathcal{T}^{-1}(\mathcal{M})\rangle = g^\top\big(\Theta + gg^\top\big)^{-1}g\,\big\|WW^\top - A\big\|_F^2.$$

As a consequence of the Sherman–Morrison formula (see Hager, 1989):

$$g^\top\big(\Theta + gg^\top\big)^{-1}g = \frac{g^\top\Theta^{-1}g}{1+g^\top\Theta^{-1}g}.$$

Computing this last quantity, along with the expression of $\mathbb{P}_\beta(W)$ in equation (224), we finally obtain:

$$\mathbb{P}_\beta(W) \propto \exp\left(-\frac{\beta d}{1+c}\big\|WW^\top - A\big\|_F^2 - 2\beta d\,\Omega(W)\right),$$

with:

$$c = \sum_{k=1}^K \frac{g_k^2}{\theta_k}.$$

From the expression of the covariance of the noise in equation (222), we have that $c = 4\beta C(0)$. This leads to equation (212).

F.2 Mapping onto the Dynamics

In this section, we apply the result of Section F.1 to compute the stationary measure of the dynamics given in Claim 4. We recall that $W(t)$ is solution of the equation:

$$\mathrm{d}W(t) = 2\left(\int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - Z(t')\Big)\mathrm{d}t'\right)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \qquad (225)$$

where the kernel R and the covariance of $\mathcal{G}$ are self-consistently computed from equation (225):

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{2\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\left(\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big((Z(t)-Z^*)(Z(t')-Z^*)\Big) + \frac{\Delta}{2}\right), \quad (226)$$

$$R(t,t') = \delta(t-t') - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\bigg|_{H=0}\right), \quad (227)$$

and the derivative in equation (227) is defined under a perturbation $\mathcal{G}(t') \mapsto \mathcal{G}(t') + H(t')$. The stochastic differential equation on W is very similar to the one in equation (210), with some differences. In the dynamics (210), it is required that the noise $\mathcal{H}$ decouples at large times, therefore we decompose $\mathcal{G}(t)$ as a static part and an independent Gaussian noise $\mathcal{H}(t)$:

$$\mathcal{G}(t) = \sqrt{\xi}\mathcal{G} + \mathcal{H}(t), \quad (228)$$

where $\mathcal{G} \sim \mathrm{GOE}(d)$ and $\mathcal{H}(t)$ is a centered Gaussian process such that:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C(t,t'), \quad (229)$$

with $C(t,t')$ going to zero as $t-t' \to \infty$. The variable ξ then corresponds to the covariance of the static part of the noise. Therefore, in order to match the setting of Section F.1, we set $A = Z^* + \sqrt{\xi}\mathcal{G}$, so that equation (225) matches (210).

Additionally, we introduce an assumption regarding the system of equations (225), (226), (227) allowing to apply the results obtained in Section F.1.1.

Assumption 43 *There exists $C_{\mathrm{TTI}}, R_{\mathrm{TTI}} \colon \mathbb{R}^+ \to \mathbb{R}$ such that:*

(*i*) Asymptotic time-translational invariance. *The functions C, R are asymptotically close to $C_{\mathrm{TTI}}, R_{\mathrm{TTI}}$ in the sense that:*

$$\begin{aligned} \sup_{t \geq t'} \big|C(t,t') - C_{\mathrm{TTI}}(t-t')\big| &\xrightarrow[t'\to\infty]{} 0, \\ \int_0^t \big|R(t,t') - R_{\mathrm{TTI}}(t-t')\big|\mathrm{d}t' &\xrightarrow[t\to\infty]{} 0. \end{aligned} \quad (230)$$

(*ii*) Fluctuation–dissipation. *$C_{\mathrm{TTI}}, R_{\mathrm{TTI}}$ are linked by the fluctuation–dissipation relation:*

$$\delta(t) = R_{\mathrm{TTI}}(t) - 4\beta\int_0^t C'_{\mathrm{TTI}}(t-t')R_{\mathrm{TTI}}(t')\mathrm{d}t'. \quad (231)$$

Indeed, the setting of Section F.1.1 requires the covariance of $\mathcal{H}$ and the function $R(t,t')$ to be time-translational invariant, i.e., to only depend on the time difference $t-t'$. This assumption is not satisfied by the dynamics (225), for which non-stationary effects may persist at early times.

However, it is standard in dynamical mean-field theory and generalized Langevin equations to assume that deviations of time-translational invariance (TTI) become negligible in the long-time regime. Following prior works such as those of Fan et al. (2025), Chen

and Shen (2025), we assume that the correlation $C(t,t')$ and response $R(t,t')$ converge to TTI limits in the sense of equation (230). In addition, the conditions in equation (230) ensure that all the non-TTI contributions vanish in the long-time limit. These subleading corrections do not contribute to the stationary measure, which is solely determined by the asymptotic of the TTI components C_{TTI} and R_{TTI}.

F.2.1 Fluctuation–Dissipation Relation

In this part, we explain how we can recover the fluctuation–dissipation relation given in Assumption 43 (and obtained in equation 211) from the system of equations (225), (226), (227). To do so, we consider the perturbed dynamics:

$$\mathrm{d}W(t) = 2M(t)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + 2\tilde{H}(t)W(t)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{232}$$

$$M(t) = \int_0^t R(t,t')\Big(\mathcal{G}(t') + Z^* - Z(t')\Big)\mathrm{d}t'. \tag{233}$$

Formally, the additional drift term originating from $\tilde{H}$ can be interpreted as the gradient of the quadratic function:

$$2\tilde{H}W = \nabla_W \mathrm{Tr}\big(\tilde{H}WW^\top\big),$$

except that $\tilde{H}$ is time-dependent. In this case we say that $\tilde{H}$ is conjugate to the observable $Z(t) = W(t)W(t)^\top$. In the setting of Langevin dynamics driven to equilibrium, it is well known that the linear response to a perturbation conjugate to an observable is related to the correlation function of that observable through the fluctuation–dissipation theorem (Kubo, 1966).

Assuming that the unperturbed dynamics converges to a stationary state and satisfies the asymptotic time-translational invariance in Assumption 43, we therefore assume that the standard fluctuation–dissipation relation holds for the observable $Z(t)$. More precisely:

$$\left.\frac{\partial\,\mathbb{E}\,Z_{ij}(t)}{\partial\tilde{H}_{ij}(t')}\right|_{\tilde{H}=0} = 2\beta d\,\partial_{t'}\mathbb{E}\big[Z_{ij}(t)Z_{ij}(t')\big].$$

This relation expresses that, at equilibrium, the response of Z_{ij} to a small perturbation is governed by its spontaneous fluctuations. The factor $2\beta d$ is the inverse of the diffusion constant in the dynamics (232). Then, summing over the indices i,j, we get the relationship:

$$\frac{1}{d^2}\mathrm{Tr}\left(\left.\frac{\partial\,\mathbb{E}\,Z(t)}{\partial\tilde{H}(t')}\right|_{\tilde{H}=0}\right) = 2\beta\partial_{t'}C_Z(t,t'), \tag{234}$$

with:

$$C_Z(t,t') = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z(t')\big).$$

Now, recall that the fluctuation–dissipation relation in equation (231) relates the response in equation (227) and the covariance of the noise $\mathcal{H}(t)$. Considering the asymptotic TTI regime, we may write $C_Z(t,t') \approx C_Z^{\mathrm{TTI}}(t-t')$ for large values of t,t' and obtain that:

$$C_{\mathrm{TTI}}(t) = \frac{1}{2\alpha}\Big(C_Z^{\mathrm{TTI}}(t) - C_Z^{\mathrm{TTI}}(\infty)\Big).$$

This relationship is a consequence of the expression of the covariance of $\mathcal{G}$ in equation (226), as well as the decomposition of $\mathcal{G}$ in equation (228). Therefore:

$$\partial_{t'} C_Z(t,t') = -2\alpha C'_{\mathrm{TTI}}(t-t'). \tag{235}$$

We now consider the left side of equation (234). Recall equations (232) and (227) that respectively define how the perturbations $\tilde{H}(t)$ and $H(t)$ appear in the dynamics. Then, one can relate the associated responses by setting:

$$\tilde{H}(t) = \int_0^t R(t,t')H(t')\mathrm{d}t'.$$

This leads to the relationship:

$$\frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z(t)}{\partial H(t')}\Big|_{H=0}\right) = \int_{t'}^t \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z(t)}{\partial \tilde{H}(t'')}\Big|_{\tilde{H}=0}\right) R(t'',t')\mathrm{d}t''. \tag{236}$$

Now, the left side of equation (236) can be reexpressed using equation (227) to obtain:

$$\delta(t-t') = R(t,t') + \frac{1}{\alpha}\int_{t'}^t \tilde{R}_Z(t,t'')R(t'',t')\mathrm{d}t'', \tag{237}$$

$$\tilde{R}_Z(t,t') = \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial \mathbb{E}\, Z(t)}{\partial \tilde{H}(t')}\Big|_{\tilde{H}=0}\right). \tag{238}$$

In the asymptotic TTI regime, we precisely get equation (211) when combining this previous equation, the fluctuation–dissipation relation (234) and the identity (235).

F.2.2 Self-Consistent Equations

Recall that in the dynamics (225), the covariance of $\mathcal{G}$ and the response R are computed self-consistently with respect to averaged quantities of W. This does not change the stationary measure of equation (212), with $A = Z^* + \sqrt{\xi}\mathcal{G}$:

$$\mathbb{P}_\beta(W) \propto \exp\left(-\frac{\beta d}{1+4\beta C_{\mathrm{TTI}}(0)}\left\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\right\|_F^2 - 2\beta d\,\Omega(W)\right). \tag{239}$$

However, the scalars $\xi, C_{\mathrm{TTI}}(0)$ need to be computed self-consistently from the distribution $\mathbb{P}_\beta$.

Self-consistency of the covariances. From the decomposition of the Gaussian processes in equation (228) and the expression for the covariance of $\mathcal{G}(t)$ in equation (226), we have the expressions:

$$\xi = \frac{1}{2\alpha}\left(\mathrm{MSE} + \frac{\Delta}{2}\right), \qquad C_{\mathrm{TTI}}(0) = \frac{1}{2\alpha}\Big(\overline{\mathrm{MSE}} - \mathrm{MSE}\Big), \tag{240}$$

where the quantities $\mathrm{MSE}, \overline{\mathrm{MSE}}$ are computed as expectations over the distribution of W in equation (239), the GOE matrix $\mathcal{G}$ and the teacher matrix Z^*:

$$\mathrm{MSE} = \frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\left\|\mathbb{E}_\beta\big[WW^\top\big] - Z^*\right\|_F^2, \qquad \overline{\mathrm{MSE}} = \frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\mathbb{E}_\beta\left\|WW^\top - Z^*\right\|_F^2, \tag{241}$$

where $\mathbb{E}_\beta$ designates the average solely with respect to $\mathbb{P}_\beta$ in equation (239).

Self-consistency of the response. In addition, one can impose the self-consistency of the response in equation (227) by integrating with respect to t':

$$\lim_{t\to\infty}\int_0^t R(t,t')\mathrm{d}t' = 1 - \frac{1}{\alpha}\lim_{t\to\infty}\int_0^t \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\bigg|_{H=0}\right)\mathrm{d}t'.$$

Now, it is clear that integrating the response over $\mathbb{R}^+$ is the same as having a time-independent perturbation H (see Section D.2.2). This leads to:

$$\lim_{t\to\infty}\int_0^t R(t,t')\mathrm{d}t' = 1 - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}_H\big[WW^\top\big]}{\partial H}\bigg|_{H=0}\right), \tag{242}$$

when considering the perturbed distribution:

$$\mathbb{P}_{\beta,H}(W) \propto \exp\left(-\frac{\beta d}{1+4\beta C_{\mathrm{TTI}}(0)}\Big\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G} - H\Big\|_F^2 - 2\beta d\,\Omega(W)\right).$$

Now, it is possible to compute the derivative in equation (242). By taking into account the fact that the normalization constant in $\mathbb{P}_{\beta,H}$ also depends on H, we obtain:

$$\frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}_H\big[WW^\top\big]}{\partial H}\bigg|_{H=0}\right) = \frac{2\beta}{1+4\beta C_{\mathrm{TTI}}(0)}\Big[\mathbb{E}_W\,\mathrm{Tr}\Big(\big(WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\big)WW^\top\Big) - \mathrm{Tr}\Big(\mathbb{E}_W\big[WW^\top\big]\mathbb{E}_W\big[WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\big]\Big)\Big].$$

Since the matrices $Z^*, \mathcal{G}$ are fixed when averaging with respect to W, we get that:

$$\frac{1}{d^2}\mathbb{E}_{\mathcal{G},Z^*}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}_H\big[WW^\top\big]}{\partial H}\bigg|_{H=0}\right) = \frac{2\beta}{1+4\beta C_{\mathrm{TTI}}(0)}\Big(\overline{\mathrm{MSE}} - \mathrm{MSE}\Big).$$

Therefore, with the relationship (240) between $C_{\mathrm{TTI}}(0)$ and $\overline{\mathrm{MSE}}-\mathrm{MSE}$, we get the identity:

$$\lim_{t\to\infty}\int_0^t R(t,t')\mathrm{d}t' = \frac{1}{1+4\beta C_{\mathrm{TTI}}(0)}. \tag{243}$$

It is interesting to note that this identity is consistent with the fluctuation–dissipation relation (231). To see this, remark that our assumptions for the asymptotic TTI in equation (230) imply that:

$$\lim_{t\to\infty}\int_0^t R(t,t')\mathrm{d}t' = \int_0^\infty R_{\mathrm{TTI}}(t')\mathrm{d}t'.$$

Then, integrating the fluctuation–dissipation relation (231) with respect to t and using that $C_{\mathrm{TTI}}(\infty) = 0$ exactly yields the relation (243). This correspondence confirms the fluctuation–dissipation relation.

Conclusion. To conclude, it is easily seen that we have:

$$\overline{\mathrm{MSE}} - \mathrm{MSE} = \frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\mathbb{E}_\beta\Big\|\mathbb{E}_\beta\big[WW^\top\big] - WW^\top\Big\|_F^2 \equiv V_\beta. \tag{244}$$

Then, given the expression of the integrated response in equation (243) (that we call r from now on), and replacing the expression of $C_{\mathrm{TTI}}(0)$ by its expression in equation (240), we obtain the self-consistent equations of Claim 8, along with the stationary measure in equation (239).

F.3 Label Equation

In this section, we perform a similar calculation to derive the stationary measure for the dynamics of the typical label, given in equation (24). To study this dynamics, we rather use the equivalent formulation given in equation (148):

$$y(t) + \xi(t) - \chi_Z(t)y^* + \frac{2}{\alpha}\int_0^t \tilde{R}_Z(t,t')\Big(y(t') - y^* - \sqrt{\Delta}\zeta\Big)\mathrm{d}t' = 0, \tag{245}$$

where $\tilde{R}_Z$ is the response function defined in equation (238), $\zeta \sim \mathcal{N}(0,1)$ and ξ is a Gaussian process with mean and covariance:

$$\mathbb{E}\,\xi(t)\xi(t') = 2C_Z(t,t') - 2Q_*\chi_Z(t)\chi_Z(t') \equiv K_Z(t,t'),$$

and we have the statistics given in Claim 4:

$$\begin{aligned} C_Z(t,t') &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z(t')\big), \qquad & \chi_Z(t) &= \frac{1}{Q_*}\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z^*\big), \\ Q_* &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}). \end{aligned} \tag{246}$$

Moreover, at long times, under time-translational invariance, equation (234) guarantees the fluctuation–dissipation relation:

$$\tilde{R}_Z(\tau) = -\beta K_Z'(\tau), \tag{247}$$

for $\tau > 0$. In the following, we shall introduce a coupling similar to the one of Section F.1 that can be mapped onto the dynamics for y.

F.3.1 Gaussian Coupling

We introduce auxiliary variables $\mathfrak{q}_1, \dots, \mathfrak{q}_k$ and the potential:

$$U_y\big(\{\mathfrak{q}_k\}\big) = \sum_{k=1}^K \theta_k \mathfrak{q}_k^2 + \frac{1}{\alpha}\left(\sum_{k=1}^K g_k\mathfrak{q}_k\right)^2,$$

for some constants $\theta_1, \dots, \theta_K > 0$ and $g_1, \dots, g_K \in \mathbb{R}$. We then consider the Langevin dynamics:

$$\mathrm{d}\mathfrak{q}_k(t) = -\frac{1}{2}\frac{\partial U_y}{\partial \mathfrak{q}_k}\Big(\big\{\mathfrak{q}_k(t)\big\}\Big)\mathrm{d}t + \frac{1}{\sqrt{\beta}}\mathrm{d}B_k(t), \tag{248}$$

where $B_1, \dots, B_K$ are independent Brownian motions. Then, it is clear that the stationary measure for the process $\{\mathfrak{q}_k(t)\}$ is given by the Boltzmann–Gibbs distribution associated with U_y. Then, computing the partial derivative of U_y, we can integrate the $\mathfrak{q}_k$ to get:

$$\begin{aligned} \mathfrak{q}_k(t) &= e^{-\theta_k t}\mathfrak{q}_k(0) - \frac{1}{\alpha}g_k\int_0^t e^{-\theta_k(t-t')}\mathfrak{q}(t')\mathrm{d}t' + \frac{1}{\sqrt{\beta}}\int_0^t e^{-\theta_k(t-t')}\mathrm{d}B_k(t'), \\ \mathfrak{q}(t) &= \sum_{k=1}^K g_k\mathfrak{q}_k(t). \end{aligned}$$

Then, we have the expression for $\mathfrak{q}(t)$:

$$\mathfrak{q}(t) = -\frac{1}{\alpha}\sum_{k=1}^{K} g_k^2 \int_0^t e^{-\theta_k(t-t')}\mathfrak{q}(t')\mathrm{d}t' + \sum_{k=1}^{K} g_k e^{-\theta_k t}\mathfrak{q}_k(0) + \frac{1}{\sqrt{\beta}}\sum_{k=1}^{K} g_k \int_0^t e^{-\theta_k(t-t')}\mathrm{d}B_k(t'). \tag{249}$$

We then choose the constants $g_1, \dots, g_K$ and $\theta_1, \dots, \theta_K$ so that:

$$\tilde{R}_Z(t) = \frac{1}{2}\sum_{k=1}^{K} g_k^2 e^{-\theta_k t}. \tag{250}$$

Of course it is not obvious that we can decompose this function as a finite sum of exponentials, but as argued in Section F.1.3 one can in principle take the $K \to \infty$ limit once the stationary measure is obtained. We then set:

$$\xi_0(t) = \sum_{k=1}^{K} g_k e^{-\theta_k t}\mathfrak{q}_k(0) + \frac{1}{\sqrt{\beta}}\sum_{k=1}^{K} g_k \int_0^t e^{-\theta_k(t-t')}\mathrm{d}B_k(t'). \tag{251}$$

Assuming that the $\mathfrak{q}_k$ are initialized as independent Gaussians with zero mean and variance:

$$\mathbb{E}\big[\mathfrak{q}_k(0)^2\big] = \frac{1}{2\theta_k\beta},$$

we obtain that the covariance function of ξ_0 is time-translational invariant:

$$\mathbb{E}\,\xi_0(t)\xi_0(t') = \frac{1}{2\beta}\sum_{k=1}^{K}\frac{g_k^2}{\theta_k} e^{-\theta_k|t-t'|}.$$

Now, as a consequence of the fluctuation–dissipation relation in equation (247), we have for $t > 0$:

$$\partial_t \frac{1}{2\beta}\sum_{k=1}^{K}\frac{g_k^2}{\theta_k} e^{-\theta_k|t|} = -\frac{1}{\beta}\tilde{R}_Z(t) = K_Z'(t).$$

Therefore, the covariances of ξ_0 and ξ appearing in equation (245) only differ by a constant. We now define $y_{\text{eff}}(t) = \mathfrak{q}(t) + m$ for some constant m. Then, using the expressions of $\mathfrak{q}(t), \tilde{R}_Z^{\text{TTI}}(t)$ and $\xi(t)$ in equations (249), (250), (251), we have:

$$y_{\text{eff}}(t) + \frac{2}{\alpha}\int_0^t \tilde{R}_Z(t-t')y_{\text{eff}}(t')\mathrm{d}t' - \xi_0(t) - \left(1 + \frac{2}{\alpha}r_Z(t)\right)m = 0,$$

where:

$$r_Z(t) = \int_0^t \tilde{R}_Z(s)\mathrm{d}s.$$

In order to match the noises ξ and ξ_0 remark that the covariance of ξ_0 between t, t' vanishes when $t - t' \to \infty$. Then, in distribution, we can decompose $\xi(t) = \xi_\infty + \xi_0(t)$, where ξ_∞ is a Gaussian variable with variance:

$$\mathbb{E}\,\xi_\infty^2 = \lim_{t\to\infty} K_Z(t).$$

Then, at long times, we shall replace the one-time functions by their final values, and the dynamics on y_{eff} matches equation (245) if we choose:

$$m = \frac{1}{\alpha + 2r_Z}\Big(\big(\alpha\chi_Z + 2r_Z\big)y^* + \alpha\xi_\infty + 2r_Z\sqrt{\Delta}\zeta \Big). \tag{252}$$

with r_Z, χ_Z being respectively the long-time limits of $r_Z(t), \chi_Z(t)$. Therefore, the previous calculation shows that the dynamics (245) can be mapped onto the Langevin equation (248). Now, the stationary measure for the variables $\{\mathfrak{q}_k\}$ is given by:

$$\begin{aligned}\mathbb{P}\big(\{\mathfrak{q}_k\}\big) &\propto \exp\left(-\beta \sum_{k=1}^K \theta_k \mathfrak{q}_k^2 - \frac{\beta}{\alpha}\left(\sum_{k=1}^K g_k \mathfrak{q}_k \right)^2 \right) \\ &= \exp\left(-\beta \mathfrak{q}^\top \Theta \mathfrak{q} - \frac{\beta}{\alpha}(g^\top \mathfrak{q})^2 \right),\end{aligned}$$

with $g = \big(g_1, \dots, g_K\big)^\top \in \mathbb{R}^K$ and $\Theta = \operatorname{diag}(\theta_1, \dots, \theta_K) \in \mathbb{R}^{K\times K}$. Since $y_{\text{eff}} = g^\top \mathfrak{q} + m$, it can be shown that y_{eff} is Gaussian with mean m and variance:

$$\operatorname{Var} y_{\text{eff}} = \frac{1}{2\beta}\frac{\alpha c}{\alpha + c}, \qquad\qquad c = \sum_{k=1}^K \frac{g_k^2}{\theta_k}.$$

Now, from equations (247), (250), we have the identity:

$$c = 2r_Z = 2\beta\big(K_Z(0) - K_Z^\infty\big). \tag{253}$$

F.3.2 Self-Consistent Equations

The previous calculation shows that we cannot write a self-consistent set of equations on the variable y only. Indeed, the distribution of y itself depends on averaged quantities with respect to W. Let us now express the quantities involved in the distribution of y as averages over the distribution of W in equation (48):

$$\chi_Z = \frac{1}{Q_*}\frac{1}{d}\mathbb{E}_{\mathcal{G},Z^*}\mathbb{E}_\beta \operatorname{Tr}\big(WW^\top Z^*\big), \tag{254}$$

$$K_Z^\infty = \frac{2}{d}\mathbb{E}_{Z^*,\mathcal{G}} \operatorname{Tr}\Big(\mathbb{E}_\beta\big[WW^\top\big]^2\Big) - 2Q_*\chi_Z^2, \tag{255}$$

$$K_Z(0) - K_Z^\infty = 2V_\beta, \tag{256}$$

where V_β is defined in equation (244). With these notations, the stationary measure for the label dynamics is Gaussian with mean and variance (conditionally on y^*, ξ, ζ):

$$\mathbb{E}\big[y \,\big|\, y^*, \xi, \zeta\big] = \frac{1}{\alpha + 4\beta V_\beta}\Big(\big(\alpha\chi_Z + 4\beta V_\beta\big)y^* + \alpha\sqrt{K_Z^\infty}\xi + 4\beta V_\beta\sqrt{\Delta}\zeta \Big), \tag{257}$$

$$\operatorname{Var}\big(y \,\big|\, y^*, \xi, \zeta\big) = \frac{2\alpha V_\beta}{\alpha + 4\beta V_\beta}, \tag{258}$$

where $y^* \sim \mathcal{N}(0, 2Q_*)$ and $\xi, \zeta \sim \mathcal{N}(0,1)$ are independent.

F.4 Zero-Temperature Limit

In this part, we focus on the stationary distribution for the matrix W, and take the $\beta \to \infty$ limit with the choice of the ℓ_2-regularization $\Omega(W) = \lambda \mathrm{Tr}(WW^\top)$. This choice allows to compare these results with the ones obtained in the gradient flow setting, mentioned in Section 3.2.

We can rewrite the stationary distribution for W in equation (239):

$$\mathbb{P}_\beta(W) \propto \exp\left(-\beta r_\beta d \left\| WW^\top - Z^* - \sqrt{\xi}\mathcal{G} + qI_d \right\|_F^2\right),$$

where $q = \lambda / r_\beta$ and r_β is the integrated response given in equation (243):

$$r_\beta = \int_0^\infty R_{\mathrm{TTI}}(t')\mathrm{d}t' = \frac{\alpha}{\alpha + 2\beta V_\beta}. \tag{259}$$

F.4.1 Positive Regularization

When $\lambda > 0$, the expression of q indicates that the integrated response r_β remains of order one. Therefore, the distribution $\mathbb{P}_\beta$ collapses onto the set:

$$\operatorname*{argmin}_{W \in \mathbb{R}^{d\times m}} \left\| WW^\top - Z^* - \sqrt{\xi}\mathcal{G} + qI_d \right\|_F^2.$$

As a consequence of Lemma 35, this is precisely the set of the $W \in \mathbb{R}^{d\times m}$ such that:

$$WW^\top = \left(Z^* + \sqrt{\xi}\mathcal{G} - qI_d\right)^+_{(m)}. \tag{260}$$

The spectral operator $X \mapsto X^+_{(m)}$ selects the m largest positive eigenvalues of X. We refer to Definition 34 for more details. Since $WW^\top$ is now a deterministic function of the matrices $Z^*, \mathcal{G}$, the two quantities MSE and $\overline{\mathrm{MSE}}$ (defined in equation 241) collapse onto the same value. Combining the definition of ξ, the expression of response and the limit of the dynamics, respectively in equations (240), (242) and (260), we arrive at the same system of equations as in Sections D.2.1, D.2.2, that eventually yields Claim 6.

This suggests that in the regularized case, the zero temperature limit of the long-time equations for the Langevin dynamics matches the gradient flow setting, which is a non-trivial result. Indeed, in the zero-temperature limit, the stationary measure of the Langevin dynamics collapses onto the set of global minimizers of the regularized empirical loss. This agreement implies that, under our dynamical assumptions, the gradient flow also converges to a global minimizer of the loss.

F.4.2 Unregularized Dynamics

We now consider the unregularized case $\lambda = 0$. In this setting, it is now possible for r_β to vanish as $\beta \to \infty$.

Let us start by considering the case where r_β remains of order one as $\beta \to \infty$. Then, the same argument as earlier applies: the distribution $\mathbb{P}_\beta$ collapses, and we recover the same equations as in Claim 6 with $\lambda, q = 0$. Interestingly, this leads to the same result as in the second part of Proposition 11 (corresponding to α larger than the interpolation threshold). As the calculation remains identical as in the gradient flow setting, we can conclude that:

- Combining equation (242) with the calculation carried out in Section D.2.2 leads to the same set of equations characterizing the interpolation thresholds as in Proposition 9.
- Since r_β is precisely the integrated response of equation (242), we obtain the same characterization of the interpolation threshold as in Section 3.3.1: it is the smallest value of α for which the integrated response remains positive in the small-regularization limit.

In this regime, it is interesting to note the agreement between the unregularized Langevin dynamics in the zero-temperature limit and the small regularization limit of the gradient flow dynamics. Moreover, as discussed in Conjecture 14, we expect these results to extend to the unregularized gradient flow itself. The regime of large sample complexity α therefore appears to be associated with a simple landscape, in which the different dynamics we studied converge to the same solution.

In the case where r_β vanishes as $\beta \to \infty$, equation (259) yields:

$$\beta r_\beta \xrightarrow[\beta\to\infty]{} \frac{\alpha}{2V},$$

where V is the positive limit of the variance V_β. Note that the intermediate scalings of r_β with β would also lead to the collapse of $\mathbb{P}_\beta$. When $r_\beta = \Theta(\beta^{-1})$, the distribution $\mathbb{P}_\beta$ does not concentrate:

$$\lim_{\beta\to\infty} \mathbb{P}_\beta(W) \propto \exp\left(-\frac{\alpha d}{2V}\left\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\right\|_F^2\right).$$

The scalars ξ, V are self-consistently computed from W thanks to equations (240), (244). This distribution describes the zero-temperature limit for $\alpha \le \alpha_{\text{inter}}$. In this regime we know that the empirical loss (4) exhibits many global minimizers, so the random nature of the Langevin predictor is no surprise.

This observation is confirmed by the statistics of the typical label in equations (257), (258). Taking the $\beta \to \infty$ limit in these equations, with $V_\beta = \Theta(1)$, leads to:

$$\mathbb{E}\, y = y^* + \sqrt{\Delta}\zeta, \qquad \text{Var}\, y = 0.$$

In this region the noisy labels are perfectly fitted.

However, this result relies on the asymptotic TTI assumption (see Assumption 43), whose validity in this regime remains unclear. In particular, as shown numerically in Figure 10, the gradient flow dynamics displays a strong dependence on initialization, even at long times. Such a dependence is incompatible with a TTI structure of the dynamics. Therefore, it remains an open question whether the Langevin dynamics satisfies the asymptotic TTI assumption in the small α regime. Overall, these observations suggest that, in this regime, the zero-temperature limit of the stationary measure of the Langevin dynamics does not coincide with the long-time asymptotics of the gradient flow dynamics.

F.5 Link with Bayes-Optimal Learning

In this part, we make the connection with the Bayes-optimal analysis of Maillard et al. (2024). To do so, we consider the stationary measure associated with our Langevin dynam-

ics, given in equation (48):

$$\mathbb{P}_\beta(W) \propto \exp\left(-r\beta d\left\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\right\|_F^2 - 2\beta d\,\Omega(W)\right).$$

From a Bayesian perspective, this distribution can be interpreted as a posterior measure over W, that includes a prior distribution:

$$\mathbb{P}_{\text{prior}}(W) \propto \exp\Big(-2\beta d\,\Omega(W)\Big).$$

In Maillard et al. (2024), the prior on the weights W is chosen to be Gaussian, inducing a Wishart distribution for $WW^\top$. In our framework, this corresponds to a ℓ_2-regularization:

$$\Omega(W) = \frac{\kappa}{4\beta}\mathrm{Tr}\big(WW^\top\big).$$

We recall that κ is the width ratio m/d as $d \to \infty$. The Bayes-optimal setting corresponds to the matched case in which the inference model coincides with the true model. In particular, the width of the student must match the one of the teacher, i.e., $\kappa = \kappa^*$.

In the following, we show two properties:

- We prove that a specific choice of the inverse temperature β allows to match our setting with the one of Maillard et al. (2024). This leads to the relationship $\beta = \alpha/\Delta$. We recall that Δ corresponds to the label noise variance when generating the teacher's labels.

- Using some known results in the Bayes-optimal setting, we show that the following relationship should hold:
$$\mathrm{MSE} = \frac{1}{2}\overline{\mathrm{MSE}}. \tag{261}$$

We recall that these two quantities are defined in equation (241). Combining these two identities with the ones obtained in Section F.2.2:

$$\xi = \frac{1}{2\alpha}\left(\mathrm{MSE} + \frac{\Delta}{2}\right), \qquad r = \frac{\alpha}{\alpha + 2\beta\big(\overline{\mathrm{MSE}} - \mathrm{MSE}\big)},$$

we obtain the simple identity:

$$r = \frac{1}{4\beta\xi}.$$

This leads to the posterior distribution:

$$\mathbb{P}_\beta(W) \propto \exp\left(-\frac{d}{4\xi}\big\|WW^\top - Z^* - \sqrt{\xi}\mathcal{G}\big\|_F^2\right)\mathbb{P}_{\text{prior}}(W).$$

Therefore, with the choice $\hat{q} = \xi^{-1}$, we obtain the same posterior distribution for W as in Maillard et al. (2024, equation 46). In addition, up to a normalization convention, our self-consistent relation between ξ and the MSE coincides with their equation (7).

F.5.1 The Bayes-Optimal Temperature

We first fix the value of the inverse temperature β to match our stationary measure with the Bayes-optimal posterior distribution, at the level of the empirical model. In Maillard et al. (2024), the Bayes posterior distribution over W is given by:

$$\mathbb{P}_{\text{Bayes}}(W) \propto \mathbb{P}_{\text{prior}}(W) \prod_{k=1}^{n} P\Big(z_k \,\Big|\, \operatorname{Tr}(X_k W W^\top)\Big). \tag{262}$$

As we have chosen the square loss in our analysis, we assume this noisy channel to be Gaussian with zero mean and variance Δ:

$$P\big(z \,|\, y\big) = \frac{1}{\sqrt{2\pi\Delta}} \exp\left(-\frac{(y-z)^2}{2\Delta}\right).$$

Then, the noisy labels can be written as $z_k = \operatorname{Tr}(X_k Z^*) + \sqrt{\Delta}\xi_k$, where $\xi_1, \dots, \xi_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. On our side, the Langevin dynamics in equation (5) can be rewritten as:

$$\mathrm{d}W(t) = -d\,\nabla_W \left(\frac{1}{4n}\sum_{k=1}^{n}\Big(\operatorname{Tr}(X_k W W^\top) - z_k\Big)^2\right)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t). \tag{263}$$

Then, the stationary measure for this dynamics writes:

$$\mathbb{P}_\beta(W) \propto \exp\left(-\frac{\beta d^2}{2n}\sum_{k=1}^{n}\Big(\operatorname{Tr}(X_k W W^\top) - z_k\Big)^2 - 2\beta d\,\Omega(W)\right).$$

The regularization term Ω has already been chosen to match the prior distribution. Therefore, to identify $\mathbb{P}_\beta$ with the Bayes posterior $\mathbb{P}_{\text{Bayes}}$ in equation (262), it suffices to match the empirical terms. This leads to the expression of the inverse temperature:

$$\beta = \frac{\alpha}{\Delta}.$$

With this choice of β, the stationary measure of the Langevin dynamics coincides with the Bayes-optimal posterior distribution. This provides an *a priori* guarantee that the Langevin dynamics at long times samples from the Bayes-optimal posterior measure.

F.5.2 The Nishimori Condition

In the Bayes-optimal setting, with a planted teacher Z^*, the following relationship holds:

$$\mathbb{E}_{W_1,W_2}\, f\big(W_1 W_1^\top, W_2 W_2^\top\big) = \mathbb{E}_{W,Z^*} f\big(W W^\top, Z^*\big), \tag{264}$$

where W_1, W_2, W are drawn from the Bayes-optimal posterior distribution, with W_1 independent of W_2. This identity is known as the Nishimori condition, and a general proof of this fact can be found in Zdeborová and Krzakala (2016, Section I.B.3).

Let us apply this identity to relate the quantities MSE and $\overline{\text{MSE}}$. For simplicity, we denote by $\mathbb{E}_W$ the expectation with respect to $\mathbb{P}_\beta$, and $\mathbb{E}$ the joint expectation over the

random variables $W, \mathcal{G}, Z^*$. Denoting by $\hat{Z}$ the posterior mean $\mathbb{E}_W[WW^\top]$, we have:

$$\begin{aligned}\overline{\mathrm{MSE}} &= \frac{1}{d}\mathbb{E}\|WW^\top - Z^*\|_F^2 \\ &= \frac{1}{d}\mathbb{E}\|WW^\top - \hat{Z}\|_F^2 + \frac{1}{d}\mathbb{E}\|\hat{Z} - Z^*\|_F^2 - \frac{2}{d}\mathbb{E}_W \operatorname{Tr}\Big((WW^\top - \hat{Z})(Z^* - \hat{Z})\Big) \\ &= \frac{1}{d}\mathbb{E}\|WW^\top - \hat{Z}\|_F^2 + \mathrm{MSE}. \end{aligned} \tag{265}$$

It is clear that the crossed term vanishes. Now, if W_1, W_2 are independently drawn from the posterior distribution, one has, inserting again $\hat{Z}$:

$$\begin{aligned}\frac{1}{d}\mathbb{E}_{W_1,W_2}\|W_1W_1^\top - W_2W_2^\top\|_F^2 = \frac{1}{d}\mathbb{E}_{W_1}\|W_1W_1^\top - \hat{Z}\|_F^2 + \frac{1}{d}\mathbb{E}_{W_2}\|W_2W_2^\top - \hat{Z}\|_F^2 \\ - \frac{2}{d}\mathbb{E}_{W_1,W_2}\operatorname{Tr}\Big((W_1W_1^\top - \hat{Z})(W_2W_2^\top - \hat{Z})\Big) \\ = \frac{2}{d}\mathbb{E}_W\|WW^\top - \hat{Z}\|_F^2.\end{aligned}$$

The last term vanishes since W_1, W_2 are independent and $\mathbb{E}_{W_1} W_1W_1^\top = \mathbb{E}_{W_2} W_2W_2^\top = \hat{Z}$. Then using the Nishimori condition in equation (264) with $f(Z_1, Z_2) = \|Z_1 - Z_2\|_F^2$, this yields:

$$\frac{1}{d}\mathbb{E}\|WW^\top - \hat{Z}\|_F^2 = \frac{1}{2}\overline{\mathrm{MSE}}.$$

Plugging this identity into equation (265) leads to the relationship (261).

Appendix G. Aging Ansatz

In this section, we go beyond the assumptions made in Section 3.2 and introduce a more general ansatz on the dynamics, that is usually referred to as aging. This framework is designed to capture situations in which the system does not reach equilibrium on the timescales of interest, and where relaxation becomes increasingly slow as time evolves. In this regime, memory effects remain and the system keeps evolving without settling into a stationary state. Additionally, correlations decay on timescales comparable to the age of the system. This behavior is characteristic of glassy and mean-field systems, and has been extensively studied in the context of dynamical mean-field theory and spin-glass dynamics (Bouchaud, 1992; Cugliandolo and Kurchan, 1993, 1994; Bouchaud et al., 1998). Additionally, our calculation follows the same steps than of Altieri et al. (2020) and uses results that can be found in Cugliandolo and Kurchan (2000).

In what follows, we decompose the stochastic equations of Claim 4 into two timescales: a fast converging part that verifies time-translational invariance (for which results have been derived in Section F), and an aging component that captures the slow relaxation of the dynamics. Our main assumptions for the section are:

- In the TTI regime, the slow variables evolving on aging timescales can be considered as effectively frozen parameters. In this regime, the system is locally at equilibrium at inverse temperature β, and the corresponding correlations and response functions satisfy a fluctuation–dissipation relation with inverse temperature β.

- For time separations comparable to the age of the system, correlations and responses enter an aging regime characterized by a violation of time-translational invariance. In this regime, we assume that a generalized fluctuation–dissipation relation holds, with an effective temperature $\beta_e < \beta$. We give more details on this parameter in Section G.4.

In order to derive the aging equations (stated in Section G.1), we perform a finite-temperature analysis in Section G.2 under the two-timescale ansatz. The self-consistent equations are then obtained by taking the zero-temperature limit (Section G.3). We conclude with a discussion of the effective temperature β_e in Section G.4 and whether it can be imposed at a finite value.

G.1 Summary of the Result

In the following, we derive a set of self-consistent equations describing the long-time limit, under the aging ansatz, of the system given in Claim 4, in the case of ℓ_2-regularization and zero label noise ($\Delta = 0$). While we derive general equations at finite temperature ($\beta < \infty$) in Section G.2, we simplify them in the zero-temperature limit ($\beta \to \infty$) in Section G.3. In this limit, we show that the self-consistent equations involve a random matrix $X \in \mathcal{S}_d(\mathbb{R})$ whose distribution is given by:

$$\mathbb{P}(X) \propto \exp\left(-d\left(2r\beta_e + \frac{\alpha}{V_Z}\right)\|X\|_F^2 - \alpha r\beta_e d\left\|X_{(m)}^+\right\|_F^2 + 2r\beta_e d\,\mathrm{Tr}\big(M^*X\big)\right), \tag{266}$$

$$M^* = \alpha Z^* + \left(2 + \frac{\alpha}{r\beta_e V_Z}\right)\Big(Z^* + \sigma\mathcal{G} - qI_d.\Big), \tag{267}$$

where $\mathcal{G} \sim \mathrm{GOE}(d)$ and $q = \lambda/r$. The variables r, V_Z, σ can be self-consistently computed as:

$$r = 1 - \frac{1}{\alpha d^2}\mathbb{E}_{Z^*,\mathcal{G}}\mathbb{E}_X \mathrm{Tr}\left(\frac{\partial}{\partial H}\big(X+H\big)_{(m)}^+\Big|_{H=0}\right), \tag{268}$$

$$V_Z = \frac{2}{d}\mathbb{E}_{Z^*,\mathcal{G}}\mathbb{E}_X\left\|\mathbb{E}_X\big[X_{(m)}^+\big] - X_{(m)}^+\right\|_F^2, \tag{269}$$

$$\sigma^2 = \frac{\alpha}{2}\frac{1}{(2r\beta_e V_Z + \alpha)^2}\frac{1}{d}\mathbb{E}_{Z^*,\mathcal{G}}\left\|\mathbb{E}_X\big[X_{(m)}^+\big] - Z^*\right\|_F^2. \tag{270}$$

The notation $X_{(m)}^+$ is defined in Definition 34, and the partial derivative corresponds to the differential on the space of symmetric matrices. Note that, in the high-dimensional limit, the limit of r can be computed by using similar arguments as in Section D.2.2. However, these equations still involve high-dimensional objects, and it seems challenging to analyze the distribution (266) in the high-dimensional limit. We leave this investigation for future work.

The only quantity that is not self-consistently determined is the effective inverse temperature β_e. This parameter is fixed by imposing a marginality condition on the TTI part of the dynamics. We give more details in Section G.4.

Finally, the gradient flow predictor is simply given by $Z_\infty = X_{(m)}^+$. This allows, in addition to a set of equations on the typical label that is derived in the following, to access all the relevant averaged quantities of the dynamics in the long-time limit.

G.2 Derivation of the Aging Equations

This section is dedicated to the derivation of the system of equations given in Section G.1.

G.2.1 Dynamical Equations

We start by giving the set of self-consistent equations that we start from in order to derive the aging equations. Instead of directly using those of Claim 4, we go back to Section C.4. In this section, it has been shown that the evolution of the student matrix $W(t)$ and typical label $y(t)$ can be written as:

$$\mathrm{d}W(t) = 2\left(\mathcal{H}(t) + \int_0^t R_y(t,t')\big(Z^* - Z(t')\big)\mathrm{d}t'\right)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t), \tag{271}$$

$$0 = y(t) - \chi_Z(t)y^* + \xi(t) + \frac{2}{\alpha}\int_0^t \tilde{R}_Z(t,t')\big(y(t') - y^*\big)\mathrm{d}t', \tag{272}$$

where $Z(t) = W(t)W(t)^\top$ and $\mathcal{H}(t)$ and $\xi(t)$ are independent centered Gaussian processes with covariances:

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{2\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\int_0^t\int_0^{t'} R_y(t,s)R_y(t',s')\mathcal{M}_Z(s,s')\mathrm{d}s'\mathrm{d}s, \tag{273}$$

$$\mathcal{M}_Z(s,s') = \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\Big[\big(Z(s) - Z^*\big)\big(Z(s') - Z^*\big)\Big], \tag{274}$$

$$\mathbb{E}\,\xi(t)\xi(t') = 2C_Z(t,t') - 2Q_*\chi_Z(t)\chi_Z(t'), \tag{275}$$

and:

$$\begin{aligned} C_Z(t,t') &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z(t')\big), \qquad & \xi_Z(t) &= \frac{1}{Q_*}\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(Z(t)Z^*\big), \\ Q_* &= \frac{1}{d}\mathbb{E}\,\mathrm{Tr}(Z^{*2}). \end{aligned} \tag{276}$$

The responses $R_Z, \tilde{R}_Z$ are defined as:

$$R_Z(t,t') = \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial H(t')}\bigg|_{H=0}\right), \qquad \tilde{R}_Z(t,t') = \frac{1}{d^2}\mathrm{Tr}\left(\frac{\partial\,\mathbb{E}\,Z(t)}{\partial \tilde{H}(t')}\bigg|_{H=0}\right), \tag{277}$$

in response to respective perturbations $H, \tilde{H}$ in the dynamics (271):

$$\mathrm{d}W(t) = \cdots + 2\int_0^t R_y(t,t')H(t')\,\mathrm{d}t'W(t)\mathrm{d}t, \qquad \mathrm{d}W(t) = \cdots + \tilde{H}(t)W(t)\mathrm{d}t.$$

Likewise, R_y is the response associated with y, that can be written:

$$R_y(t,t') = -\frac{\partial\,\mathbb{E}\,y(t)}{\partial\xi(t')}.$$

This leads to the relationships between $R_y, R_Z, \tilde{R}_Z$:

$$\begin{aligned} \delta(t-t') &= R_y(t,t') + \frac{1}{\alpha}R_Z(t,t'), \\ \delta(t-t') &= R_y(t,t') + \frac{2}{\alpha}\int_{t'}^t \tilde{R}_Z(t,t'')R_y(t'',t')\mathrm{d}t''. \end{aligned} \tag{278}$$

Finally, one can derive the following relationship, starting from equation (272) and using equations (273), (275), (276):

$$\mathbb{E}\,\mathcal{H}_{ij}(t)\mathcal{H}_{i'j'}(t') = \frac{1}{4\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)\mathbb{E}\Big[\big(y(t)-y^*\big)\big(y(t')-y^*\big)\Big]. \tag{279}$$

In the following we shall denote C_y the covariance function of $y - y^*$.

G.2.2 Timescale Decomposition and Fluctuation–Dissipation

We now start from the previous set of equations and decompose the responses and correlations depending on two timescales, that we call time-translational invariant (TTI) and aging. For the responses, we write:

$$\begin{aligned} R_y(t,t') &= R_y^{\mathrm{TTI}}(t-t') + R_y^A(t,t'),\\ \tilde{R}_Z(t,t') &= \tilde{R}_Z^{\mathrm{TTI}}(t-t') + \tilde{R}_Z^A(t,t'), \end{aligned} \tag{280}$$

and likewise for the correlations:

$$\begin{aligned} C_y(t,t') &= C_y^{\mathrm{TTI}}(t-t') + C_y^A(t,t'),\\ K_Z(t,t') &= K_Z^{\mathrm{TTI}}(t-t') + K_Z^A(t,t'), \end{aligned} \tag{281}$$

where K_Z is the covariance of ξ given in equation (275). As we will be interested in the $t, t' \to \infty$ limit, one-time functions can be replaced by their limit, which yields in particular $K_Z(t,t') = 2C_Z(t,t') - 2Q_*(\chi_Z^\infty)^2$. In the previous equations, the TTI contributions vary when $t-t'$ is of order one, whereas the aging regime corresponds to $t-t'$ diverging. In this regime we choose the TTI covariances to be zero, that is:

$$C_y^{\mathrm{TTI}}(\tau) \xrightarrow[\tau\to\infty]{} 0, \qquad\qquad K_Z^{\mathrm{TTI}}(\tau) \xrightarrow[\tau\to\infty]{} 0.$$

In addition, equation (281) on the covariances induces the decomposition of ξ and $\mathcal{H}$ as sums of independent Gaussian processes:

$$\xi(t) = \xi_{\mathrm{TTI}}(t) + \xi_A(t), \qquad\qquad \mathcal{H}(t) = \mathcal{H}_{\mathrm{TTI}}(t) + \mathcal{H}_A(t),$$

where ξ_{TTI} and ξ_A have respective covariances K_Z^{TTI} and K_Z^A. The same holds for $\mathcal{H}_{\mathrm{TTI}}$ and $\mathcal{H}_A$, with the additional normalization:

$$\mathbb{E}\,\big(\mathcal{H}_{\mathrm{TTI}}(t)\big)_{ij}\big(\mathcal{H}_{\mathrm{TTI}}(t')\big)_{i'j'} = \frac{1}{4\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C_y^{\mathrm{TTI}}(t-t'), \tag{282}$$

$$\mathbb{E}\,\big(\mathcal{H}_A(t)\big)_{ij}\big(\mathcal{H}_A(t')\big)_{i'j'} = \frac{1}{4\alpha d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C_y^{A}(t-t'), \tag{283}$$

where we used the covariance structure of $\mathcal{H}$ in equation (279).

In addition, we assume the fluctuation–dissipation relations in the TTI regime:

$$R_y^{\mathrm{TTI}}(\tau) = -\beta\big(C_y^{\mathrm{TTI}}\big)'(\tau), \qquad\qquad \tilde{R}_Z^{\mathrm{TTI}}(\tau) = -\beta\big(K_Z^{\mathrm{TTI}}\big)'(\tau). \tag{284}$$

While the second expression is a consequence of the calculation carried out in Section F.2.1 (see in particular equation 234), along with the definition of $\tilde{R}_Z, C_Z$ and K_Z, the first one can be derived in a similar fashion.

In the aging regime, fluctuation–dissipation does not hold at inverse temperature β, but with an effective inverse temperature, that we denote β_e:

$$R_y^A(t,t') = \beta_e\partial_{t'}C_y^A(t,t'), \qquad\qquad \tilde{R}_Z^A(t,t') = \beta_e\partial_{t'}K_Z^A(t,t'). \tag{285}$$

G.2.3 TTI Regime

Using the previous decompositions, we can write the joint dynamics on $W(t)$ and $y(t)$ given in equations (271), (272):

$$\mathrm{d}W(t) = 2\left(\mathcal{H}_{\mathrm{TTI}}(t) + \int_0^t R_y^{\mathrm{TTI}}(t-t')\big(Z^* - Z(t')\big)\mathrm{d}t'\right)W(t)\mathrm{d}t - \nabla\Omega\big(W(t)\big)\mathrm{d}t \qquad (286)$$
$$+ 2\Psi(t)W(t)\mathrm{d}t + \frac{1}{\sqrt{\beta d}}\mathrm{d}B(t),$$

$$0 = y(t) - \chi_Z(t)y^* + \xi_{\mathrm{TTI}}(t) + \frac{2}{\alpha}\int_0^t \tilde{R}_Z^{\mathrm{TTI}}(t-t')\big(y(t') - y^*\big)\mathrm{d}t' + h(t), \qquad (287)$$

where the slow fields $\Psi(t), h(t)$ are given by:

$$\Psi(t) = \mathcal{H}_A(t) + \int_0^t R_y^A(t,t')\big(Z^* - Z(t')\big)\mathrm{d}t', \qquad (288)$$

$$h(t) = \xi_A(t) + \frac{2}{\alpha}\int_0^t \tilde{R}_Z^A(t,t')\big(y(t') - y^*\big)\mathrm{d}t'. \qquad (289)$$

The key point is to consider the slow fields $\Psi(t), h(t)$ as frozen in the TTI regime, that corresponds to timescales of order one. On the other hand, we assume that the slow aging variables typically vary on a much larger timescale, comparable to the age of the system.

Student equation. Let us start by analyzing the dynamics of $W(t)$ in equation (286). In the TTI regime, this equation can be written exactly as the one studied in Section F.1, with an additional gradient term involving Ψ. Since the covariance of $\mathcal{H}_{\mathrm{TTI}}$ between instants t, t' vanishes as $t - t' \to \infty$, this implies the stationary distribution, conditionally on $\Psi(t)$:

$$\mathbb{P}_W^{\mathrm{TTI}}\big(W \mid \Psi(t)\big) \propto \exp\left(-r\beta d\left\|WW^\top - Z^*\right\|_F^2 - 2\beta d\,\Omega(W) + 2\beta d\,\mathrm{Tr}\big(\Psi(t)\big(WW^\top - Z^*\big)\big)\right). \qquad (290)$$

In addition, r is given by:

$$r = \int_0^\infty R_y^{\mathrm{TTI}}(\tau)\mathrm{d}\tau = \beta C_y^{\mathrm{TTI}}(0), \qquad (291)$$

where we used the fluctuation–dissipation relation for y in equation (284).

Label equation. To compute the stationary distribution of the label in the TTI regime, we introduce the same coupling as the one studied in Section F.3. We obtain that conditionally on the slow field $h(t)$, the typical label y is Gaussian, with statistics:

$$\mathbb{E}\, y = \frac{1}{\alpha + 2r_Z}\Big(\big(\alpha\chi_Z + 2r_Z\big)y^* - \alpha h(t)\Big),$$
$$\mathrm{Var}\, y = \frac{1}{\beta}\frac{\alpha r_Z}{\alpha + 2r_Z},$$

where χ_Z is the limit of $\chi_Z(t)$ as $t \to \infty$ and we have from equation (284):

$$r_Z = \int_0^\infty \tilde{R}_Z^{\mathrm{TTI}}(t)\mathrm{d}t = \beta K_Z^{\mathrm{TTI}}(0).$$

Therefore, we can write the distribution of y in the TTI regime as:

$$\mathbb{P}_y^{\mathrm{TTI}}\big(y \,|\, h(t)\big) \propto \exp\left(-\frac{\beta}{2\alpha r_Z}\Big((\alpha + 2r_Z)y^2 - 2(\alpha\chi_Z + 2r_Z)yy^* + 2\alpha h(t)(y - y^*)\Big)\right). \tag{292}$$

G.2.4 An Auxiliary Coupling

In order to study the aging regime and derive the stationary distribution of the slow variables $\Psi(t), h(t)$, defined in equations (288), (289), we introduce an auxiliary coupling. This coupling is inspired by the one introduced by Cugliandolo and Kurchan (2000) in the context of aging in spin-glass models.

We consider a smooth function $F\colon \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}$, a fixed matrix $A \in \mathcal{S}_d(\mathbb{R})$, and define the potential on the variables $\mathcal{Q}_1, \dots, \mathcal{Q}_K \in \mathcal{S}_d(\mathbb{R})$:

$$U_{\mathrm{aging}}\big(\{\mathcal{Q}_k\}\big) = \rho F\left(A + \sum_{k=1}^K g_k \mathcal{Q}_k\right) + \sum_{k=1}^K \theta_k \|\mathcal{Q}_k\|_F^2.$$

We then study the associated Langevin dynamics:

$$\mathrm{d}\mathcal{Q}_k(t) = -\frac{\rho g_k}{2}\nabla F\left(A + \sum_{j=1}^K g_j \mathcal{Q}_j(t)\right)\mathrm{d}t - \theta_k \mathcal{Q}_k(t)\mathrm{d}t + \frac{1}{\sqrt{\beta_e d}}\mathrm{d}\Xi_k(t), \tag{293}$$

where $\Xi_1, \dots, \Xi_k$ are independent standard Brownian motions over $\mathcal{S}_d(\mathbb{R})$. The presence of β_e is motivated by the assumptions of the section: in the aging regime, correlations and responses obey a generalized fluctuation–dissipation relation with effective temperature β_e. The stationary measure associated with the dynamics (293) is given by:

$$\mathbb{P}_{\mathrm{aging}}\big(\{\mathcal{Q}_k\}\big) \propto \exp\Big(-\beta_e d\, U_{\mathrm{aging}}\big(\{\mathcal{Q}_k\}\big)\Big). \tag{294}$$

Moreover, we can integrate the dynamics (293) to get:

$$\mathcal{Q}_k(t) = e^{-\theta_k t}\mathcal{Q}_k(0) - \frac{\rho g_k}{2}\int_0^t e^{-\theta_k(t-t')}\nabla F\left(A + \sum_{j=1}^K g_j \mathcal{Q}_j(t')\right)\mathrm{d}t' + \frac{1}{\sqrt{\beta_e d}}\int_0^t e^{-\theta_k(t-t')}\mathrm{d}\Xi_k(t').$$

We then define:

$$\Psi(t) = A + \sum_{k=1}^K g_k \mathcal{Q}_k(t), \qquad R_0(t) = \sum_{k=1}^K g_k^2 e^{-\theta_k t},$$
$$\mathcal{G}(t) = \sum_{k=1}^K g_k e^{-\theta_k t}\mathcal{Q}_k(0) + \frac{1}{\sqrt{\beta_e d}}\sum_{k=1}^K g_k \int_0^t e^{-\theta_k(t-t')}\mathrm{d}\Xi_k(t'),$$

so that Ψ is solution of the equation:

$$\Psi(t) = A + \mathcal{G}(t) - \frac{\rho}{2}\int_0^t R_0(t-t')\nabla F\big(\Psi(t')\big)\mathrm{d}t'. \tag{295}$$

Similarly to what was done in Section F.1.3, we pick the initializations of the $\mathcal{Q}_k$ as independent centered Gaussian matrices, with covariance:

$$\mathbb{E}\left(\mathcal{Q}_k^0\right)_{ij}\left(\mathcal{Q}_{k'}^0\right)_{i'j'} = \frac{1}{4\theta_k\beta_e d}\delta_{kk'}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big).$$

This choice guarantees that the covariance of the Gaussian process $\mathcal{G}$ is time-translational invariant:

$$\mathbb{E}\,\mathcal{G}_{ij}(t)\mathcal{G}_{i'j'}(t') = \frac{1}{d}\big(\delta_{ii'}\delta_{jj'} + \delta_{ij'}\delta_{i'j}\big)C(t-t'), \qquad C(t) = \frac{1}{4\beta_e}\sum_{k=1}^K \frac{g_k^2}{\theta_k}e^{-\theta_k|t|}. \tag{296}$$

From this it is clear that we have the fluctuation–dissipation relation, for $t > 0$:

$$R_0(t) = -4\beta_e C'(t). \tag{297}$$

Then, following a similar reasoning as in Section F.1.4, we can integrate the stationary distribution in equation (294) to get the one on Ψ:

$$\mathbb{P}_{\text{aging}}(\Psi) \propto \exp\left(-\rho\beta_e d\,F(\Psi) - \frac{\beta_e d}{c}\big\|\Psi - A\big\|_F^2\right), \qquad c = \sum_{k=1}^K \frac{g_k^2}{\theta_k}. \tag{298}$$

With the previous equations, we have:

$$c = \int_0^\infty R_0(t)\mathrm{d}t = 4\beta_e C(0).$$

G.2.5 Aging Regime

We shall now apply the previous result to the aging regime. More precisely, we now derive the stationary measures for the slow fields $\Psi(t), h(t)$. First recall the TTI stationary distributions given in equations (290), (292), and define the free energies:

$$F_W(\Psi) = \frac{1}{\beta d}\log\int \mathrm{d}W \exp\Bigg(-r\beta d\left\|WW^\top - Z^*\right\|_F^2 - 2\beta d\,\Omega(W) + 2\beta d\,\mathrm{Tr}\Big(\Psi\big(WW^\top - Z^*\big)\Big)\Bigg), \tag{299}$$

$$F_y(h) = \frac{r_Z}{\beta}\log\int \mathrm{d}y \exp\left(-\frac{\beta}{2\alpha r_Z}\Big((\alpha + 2r_Z)y^2 - 2(\alpha\chi_Z + 2r_Z)yy^* + 2\alpha h(y-y^*)\Big)\right). \tag{300}$$

Then, we have:

$$\nabla F_W(\Psi) = 2\,\mathbb{E}\big[WW^\top - Z^*\big], \qquad F_y'(h) = -\mathbb{E}\big[y - y^*\big],$$

where the expectations are computed with respect to the TTI distributions in (290), (292). Now, in the slow aging regime, the variables $W(t)W(t)^\top, y(t)$ can be replaced by their averages with respect to fluctuations on short timescales, that is with respect to the TTI distributions. This means we can rewrite:

$$\Psi(t) = \mathcal{H}_A(t) - \frac{1}{2}\int_0^t R_y^A(t,t')\nabla F_W\big(\Psi(t')\big)\mathrm{d}t', \tag{301}$$

$$h(t) = \xi_A(t) - \frac{2}{\alpha}\int_0^t \tilde{R}_Z^A(t,t')F_y'\big(h(t')\big)\mathrm{d}t'. \tag{302}$$

In order to match with the setting of the previous section, we decompose the Gaussian noises:

$$\mathcal{H}_A(t) = \tilde{\mathcal{H}}_A(t) + \mathcal{H}_A^\infty, \qquad \xi_A(t) = \tilde{\xi}_A(t) + \xi_A^\infty,$$

so that $\tilde{\mathcal{H}}_A, \tilde{\xi}_A$ have vanishing covariances between instants t, t' whose difference diverges beyond the aging regime.

Let us now identify the dynamics (301) with the one of the auxiliary coupling in equation (295). First of all, we identify the constant Gaussian noise $\mathcal{H}_A^\infty$ with the matrix A, as well as $\mathcal{G}(t) = \tilde{\mathcal{H}}_A(t)$ in equation (295), which implies the identity between covariances, using equations (283) and (296):

$$C = \frac{1}{4\alpha}\big(C_y^A - C_y^A(\infty)\big).$$

Now, in order to match the expressions of Ψ in equations (295) and (301), we choose $F = F_W$ and $\rho R_0 = R_y^A$. Then, we can identify the fluctuation-dissipation relations in equations (285) and (297), that leads to the choice $\rho = \alpha$. These identities show that we can apply the result of the previous section to the dynamics (301), and obtain, thanks to equation (298), that Ψ has stationary measure:

$$\mathbb{P}_W^{\text{aging}}(\Psi) \propto \exp\left(-\alpha\beta_e d\, F_W(\Psi) - \frac{\alpha d}{C_y^A(0) - C_y^A(\infty)}\big\|\Psi - \tilde{\mathcal{H}}_A^\infty\big\|_F^2\right). \tag{303}$$

The same identification can be performed for the dynamics of $h(t)$ in equation (302), in one dimension. In this case, we choose:

$$F = F_y, \qquad A = \xi_A^\infty, \qquad C = \frac{K_Z - K_Z(\infty)}{2}, \qquad R_0 = \frac{4}{\alpha\rho}\tilde{R}_Z^A.$$

The third equation originates from matching the covariance of $\tilde{\xi}_A^\infty$ with the one of $\mathcal{G}$ in equation (296), equal to $2C$ in one dimension. Finally, imposing the equivalence of the fluctuation-dissipation relations in equations (285), (297) leads to the relation $\alpha\rho = 2$. Using equation (298), this shows that h has stationary distribution:

$$\mathbb{P}_y^{\text{aging}}(h) \propto \exp\left(-\frac{2\beta_e}{\alpha}F_y(h) - \frac{1}{2}\frac{1}{K_Z^A(0) - K_Z^A(\infty)}\big(h - \tilde{\xi}_A^\infty\big)^2\right). \tag{304}$$

Already it is interesting to remark that the distribution for y conditionally to h in equation (292) is Gaussian, and so is the distribution of h itself (this can be shown by computing F_y). Therefore, the typical label y is itself Gaussian, which is consistent with the observation made in Section 3.1.3.

G.2.6 Closing the System of Equations

The previous calculations in the TTI and aging regimes allow to access the distribution of W and y under the aging ansatz. These are given by:

$$\mathbb{P}(W) \propto \int \mathbb{P}_W^{\mathrm{TTI}}\big(W \,|\, \Psi\big)\mathbb{P}_W^{\mathrm{aging}}(\Psi)\mathrm{d}\Psi,$$
$$\mathbb{P}(y) \propto \int \mathbb{P}_y^{\mathrm{TTI}}\big(y \,|\, h\big)\mathbb{P}_y^{\mathrm{aging}}(h)\mathrm{d}h.$$

Then, all the quantities appearing in these four distributions can be expressed as expectations over these distributions. We have:

$$r = \beta C_y^{\mathrm{TTI}}(0) = \beta\, \mathbb{E}_h\mathbb{E}_y\big[(\overline{y} - y)^2\big], \tag{305}$$
$$r_Z = \beta K_Z^{\mathrm{TTI}}(0) = \frac{2\beta}{d}\mathbb{E}_\Psi\mathbb{E}_W\Big[\big\|\overline{Z} - WW^\top\big\|_F^2\Big], \tag{306}$$
$$\chi_Z = \frac{1}{Q_*}\frac{1}{d}\mathbb{E}_\Psi\mathbb{E}_W\Big[\mathrm{Tr}\big(WW^\top Z^*\big)\Big], \tag{307}$$
$$C_y^A(0) - C_y^A(\infty) = \mathbb{E}_h\Big[\big(\mathbb{E}_h[\overline{y}] - \overline{y}\big)^2\Big] \equiv V_y, \tag{308}$$
$$K_Z^A(0) - K_Z^A(\infty) = \frac{2}{d}\mathbb{E}_\Psi\Big[\big\|\mathbb{E}_\Psi[\overline{Z}] - \overline{Z}\big\|_F^2\Big] \equiv V_Z, \tag{309}$$

where $\mathbb{E}_W, \mathbb{E}_y$ denote the expectations with respect to the TTI distributions, conditionally on Ψ, h, as well as $\overline{y} = \mathbb{E}_y[y]$ and $\overline{Z} = \mathbb{E}_W[WW^\top]$. Moreover, in the following, we denote the Gaussian noises in equations (303), (304) by:

$$\tilde{\mathcal{H}}_A^\infty = \sigma_y \mathcal{G}, \qquad \tilde{\xi}_A^\infty = \sigma_W \phi,$$

where $\mathcal{G} \sim \mathrm{GOE}(d), \phi \sim \mathcal{N}(0,1)$, and:

$$\sigma_y^2 = \frac{1}{4\alpha}\big(\mathbb{E}_h[\overline{y}] - y^*\big)^2, \qquad \sigma_W^2 = \frac{2}{d}\mathrm{Tr}\big(\mathbb{E}_\Psi[\overline{Z}]^2\big) - 2Q_*\chi_Z^2. \tag{310}$$

This provides a set of self-consistent equations describing the long-time behavior of the Langevin dynamics under the aging ansatz. In the following, we take the zero-temperature limit ($\beta \to \infty$) in order to simplify these equations.

G.3 Zero-Temperature Limit

In the $\beta \to \infty$ limit, we analyze the previous equations with the choice of the regularization $\Omega(W) = \lambda \mathrm{Tr}(WW^\top)$. In this limit, the equations can be simplified and lead to the result given in Section G.1. However, it is not guaranteed that taking the zero-temperature limit of the stationary equations (i.e., after taking $t \to \infty$) leads to the long-time behavior of gradient flow dynamics. Nevertheless, as shown in Section F.4 in the TTI analysis (with positive regularization), both lead to the same limiting behavior, which suggests that they may remain closely related under the aging ansatz.

G.3.1 Student Equations

Let us start by computing the free energy F_W in equation (299). The first thing to notice is that the density:

$$\exp\left(-r\beta d\left\|WW^\top - Z^*\right\|_F^2 - 2\lambda\beta d\,\mathrm{Tr}(WW^\top) + 2\beta d\,\mathrm{Tr}\Big(\Psi\big(WW^\top - Z^*\big)\Big)\right),$$

concentrates as $\beta \to \infty$ on the set:

$$\left\{W \in \mathbb{R}^{d\times m}, WW^\top = \left(Z^* + \frac{1}{r}\Psi - \frac{\lambda}{r}I_d\right)^+_{(m)}\right\}, \tag{311}$$

where the operator $X \mapsto X^+_{(m)}$ is defined in Definition 34 and selects the m largest positive eigenvalues of X. The concentration of the previous density is a consequence of Lemma 35. Therefore, under the TTI distribution (290), $WW^\top$ is a deterministic function of Ψ.

Now, the free energy defined in equation (299) has the asymptotic:

$$F_W(\Psi) \xrightarrow[\beta\to\infty]{} r\left\|\big(Z^*_\lambda + r^{-1}\Psi\big)^+_{(m)}\right\|_F^2 - 2\mathrm{Tr}\big(\Psi Z^*\big) - r\|Z^*\|_F^2,$$

with $Z^*_\lambda = Z^* - \lambda r^{-1} I_d$. We can conclude the expression of the distribution of Ψ given in equation (303):

$$\mathbb{P}_W^{\text{aging}}(\Psi) \propto \exp\left(-\alpha r\beta_e d\left\|\big(Z^*_\lambda + r^{-1}\Psi\big)^+_{(m)}\right\|_F^2 + 2\alpha\beta_e d\,\mathrm{Tr}\big(\Psi Z^*\big) - \frac{\alpha d}{V_y}\big\|\Psi - \sigma_y\mathcal{G}\big\|_F^2\right). \tag{312}$$

As a consequence of this concentration, the variance term in equation (306) is of order β^{-1}, leading to r_Z of order one. To compute r_Z, we can go back to equation (278) that links the responses $\tilde{R}_Z, R_y$. In the TTI regime, this shows the relation:

$$\alpha + 2r_Z = \frac{\alpha}{r}.$$

Moreover, we can apply the same argument as in Section D.2.2 to compute the integrated response as a derivative over the stationary distribution with respect to a constant perturbation:

$$r = 1 - \frac{1}{\alpha d^2}\mathrm{Tr}\left(\frac{\partial}{\partial H}\left(Z^* + \frac{1}{r}\Psi - \frac{\lambda}{r}I_d + H\right)^+_{(m)}\Bigg|_{H=0}\right), \tag{313}$$

where $H \in \mathcal{S}_d(\mathbb{R})$ and the trace is to be interpreted as taken over the linear maps on the space of symmetric matrices.

G.3.2 Label Equations

Let us now move on to the label equation. In the $\beta \to \infty$ limit, the distribution of y conditionally on h in equation (292) concentrates around:

$$y = \frac{\alpha\chi_Z + 2r_Z}{\alpha + 2r_Z}y^* - \frac{\alpha}{\alpha + 2r_Z}h.$$

Therefore, y is a deterministic function of h. In addition, we have the limit for the free energy in equation (300):

$$F_y(h) \xrightarrow[\beta\to\infty]{} hy^* + \frac{1}{2\alpha(\alpha+2r_Z)}\Big((\alpha\chi_Z + 2r_Z)y^* - \alpha h\Big)^2.$$

Plugging this expression into the distribution of h in equation (304), we obtain that:

$$\mathbb{P}_y^{\text{aging}}(h) \propto \exp\left(-\frac{2\beta_e}{\alpha}hy^* - \frac{r\beta_e}{\alpha}\left(\left(\chi_Z + \frac{2r_Z}{\alpha}\right)y^* - h\right)^2 - \frac{1}{2V_Z}\big(h - \sigma_W\phi\big)^2\right).$$

Therefore h is Gaussian, with mean and variance:

$$\begin{aligned}\mathbb{E}\,h &= \frac{1}{2r\beta_e V_Z + \alpha}\Big(-2r\beta_e V_Z(1-\chi_Z)y^* + \alpha\sigma_W\phi\Big),\\ \text{Var}\,h &= \frac{\alpha V_Z}{2r\beta_e V_Z + \alpha}.\end{aligned}$$

This leads to the equation on V_y in equation (308):

$$V_y = r^2\text{Var}\,h = \frac{\alpha r^2 V_Z}{2r\beta_e V_Z + \alpha}. \tag{314}$$

In addition, the variance of the GOE noise given in equation (310) writes:

$$\sigma_y^2 = \frac{\alpha r^2}{2}\frac{1}{(2r\beta_e V_Z + \alpha)^2}\frac{1}{d}\Big\|\mathbb{E}_\Psi\big[WW^\top\big] - Z^*\Big\|_F^2. \tag{315}$$

These expressions guarantee that we can write a closed set of equations on the variable Ψ.

G.3.3 Set of Equations

We now explain how to arrive at the system of equations given in Section G.1. Let us start by defining:

$$X = Z^* - \frac{\lambda}{r}I_d + \frac{1}{r}\Psi,$$

so that X is a linear transform of Ψ. Then:

- Equation (311) guarantees that the limit of the dynamics is simply $WW^\top = X^+_{(m)}$.
- Equation (268) is easily derived from equation (313).
- Equation (269) is a consequence of equation (309).
- Letting $\sigma = \sigma_y/r$ in equation (315) yields equation (270).
- Plugging equation (314) and the expression of X into the distribution (312) leads to equation (266).

G.4 The Effective Temperature

The previous calculation leads to a set of self-consistent equations relating the distribution of X in (266) to the parameters appearing in this distribution. The only quantity that remains undetermined is the effective inverse temperature β_e.

The value of β_e cannot be determined from the aging equations alone and must be fixed by an additional condition. To do so, we impose that the time-translational invariant (TTI) solution should be *marginally stable*, meaning that small perturbations around this solution relax on arbitrarily long timescales. This marginality condition selects a unique value of β_e and ensures a consistent matching between the TTI and aging regimes.

To impose the effective temperature, recall that we studied in Section E.1 the susceptibility operator associated with the TTI solution at zero temperature. Since the aging solution is also given by a matrix of the form $Z = X^+_{(m)}$, a similar computation leads to an expression for the Frobenius norm of the susceptibility in the high-dimensional limit:

$$\mathcal{R} = \frac{1}{2} \iint \left(\frac{x \mathbf{1}_{x \geq \max(0,\omega)} - y \mathbf{1}_{y \geq \max(0,\omega)}}{x - y} \right)^2 \mathrm{d}\mu_X(x) \mathrm{d}\mu_X(y),$$

where μ_X is the asymptotic spectral distribution of X, and ω is a threshold selecting a fraction κ of the largest eigenvalues of X. In Section E.1, we have seen that this quantity may remain finite because of two mechanisms: if $\omega \leq 0$ or if the measure μ_X does not have mass in the vicinity of ω.

Due to the form of the distribution $\mathbb{P}(X)$ in equation (266), we expect μ_X to have mass near the threshold ω, at least when it is positive. Indeed, the second term in equation (266) can be interpreted as a penalization of the eigenvalues exceeding ω, which should favor an accumulation of eigenvalues close to this threshold.

This observation suggests that the marginal stability of the TTI solution cannot be achieved by requiring μ_X to vanish near the threshold ω. We therefore propose that the marginality condition should be imposed by setting $\omega = 0$, i.e., by constraining the matrix X to have exactly m positive eigenvalues. This condition implicitly defines the effective inverse temperature β_e.

It is, however, not obvious if this marginality condition can be imposed. Addressing this question requires understanding the effect of the parameter β_e on the distribution in equation (266), and in particular how it controls the number of positive eigenvalues of the matrix X. If such a control is not possible, the marginality condition cannot be imposed, and the time-translational invariant solution would remain stable, leading to the absence of aging in the dynamics. We leave this challenging question for further work.

Perfect recovery thresholds. A more reasonable analysis could help derive the perfect recovery thresholds from the system of equations in Section G.1. In this limit, we expect that the distribution of X can be understood using a Laplace method and a perturbative analysis, leading to a simplification of the self-consistent equations. However, it remains open whether the thresholds obtained from this aging system coincide with those of Proposition 12 and Conjecture 15.

Appendix H. Small Regularization Limit

In this section, we derive the results presented in Section 3.3 on the small regularization limit. The section is organized as follows:

- In Section H.1 and Section H.2, we compute the interpolation and perfect recovery thresholds respectively introduced in Section 3.3.1 and Section 3.3.3.
- In Section H.3, for the sake of completeness, we give a proof of Proposition 13 regarding the notion of minimal regularization interpolator.
- In Section H.4, we derive the asymptotics associated with the functions introduced in Claim 6 in the limit of the small noise ξ. These technical results allow to compute the interpolation and perfect recovery thresholds in the noiseless case.

In the following, and in agreement with the factorization of the teacher $Z^* = W^* W^{*\top}$, where $W^* \in \mathbb{R}^{d\times m^*}$, we adopt the following decomposition for the limiting distribution of the teacher μ^*:

$$\mu^* = (1 - \min(\kappa^*, 1))\delta + \min(\kappa^*, 1)\nu^*,$$

where ν^* has support in $\mathbb{R}^+$ bounded away from zero. In the following we will also assume that ν^* admits a smooth density with respect to the Lebesgue measure. However, since the results we give do not depend on this property, we believe they can be extended in the general case.

This representation will be useful to study our system of equations in the $\xi \to 0$ limit. In this limit, provided that the support of ν^* is bounded away from zero, the support of the measure μ_ξ (the free additive convolution between μ^* and a semicircular density of variance ξ) splits into two parts when $\kappa^* < 1$: a semicircular density with radius $\propto \sqrt{\xi}$ and mass $1-\kappa^*$, and a measure of mass κ^* that resembles ν^*. More quantitative details can be found in Section H.4, and we also refer to Maillard et al. (2024, Section F).

H.1 Interpolation Threshold

In this section we prove the expressions of the interpolation threshold given in Proposition 9 and Corollary 10.

H.1.1 Derivation of the Interpolation Threshold

We shall now derive the system of equations that characterizes the interpolation threshold in Proposition 9. As a consequence, we also access the system of equations in the small regularization limit given in Proposition 11.

Let us recall the system of equations (37). The equations on the variables (q, ξ) can be written:

$$1 = \frac{\lambda}{q} + \frac{1}{\alpha} I_\omega(q, \xi), \tag{316}$$

$$2\alpha\xi - \frac{\Delta}{2} = Q_* + \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x) + 4\xi I_\omega(q, \xi), \tag{317}$$

with:

$$I_\omega(q,\xi) = \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x), \tag{318}$$

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x). \tag{319}$$

From the expression of the MSE in equation (40), we have MSE $\leq 2\alpha\xi$. This implies that when α is smaller than the perfect recovery threshold, ξ remains positive in the small regularization limit. Regarding the behavior of q in this limit, we consider two cases in the following.

We work with the scaling $q = r\lambda$, with fixed $r > 0$. Here r corresponds to the variable r_∞ introduced in Section 3.2.1. Then, equation (317) rewrites:

$$2\alpha\xi - \frac{\Delta}{2} = Q_* - \int_{\max(0,\omega)} x^2 \mathrm{d}\mu_\xi(x) + 4\xi I_\omega(\xi), \tag{320}$$

with $I_\omega(\xi) = I_\omega(0,\xi)$. This corresponds to the second part of Proposition 11. In this case the first equation of (316) simply gives the expression of r:

$$r = 1 - \frac{1}{\alpha} I_\omega(\xi).$$

This is only valid if $r > 0$, which leads to the constraint $\alpha > I_\omega(\xi)$. This leads to the definition of the interpolation threshold as the limiting value for which these equations are verified. Combining equation (320) with the equality $\alpha_{\text{inter}} = I_\omega(\xi)$ leads to the result of Proposition 9.

Now considering the regime where q remains of order one as $\lambda \to 0^+$, we obtain the first part of Proposition 11 for $\alpha < \alpha_{\text{inter}}$. Note that one could also consider intermediate regimes for the dependence between q and λ, but all of these collapse into the equations at $\alpha = \alpha_{\text{inter}}$.

We also derive the expression of the in-sample error that is defined in equation (73) and plotted in Figure 11. To do so, we go back to the expression of the label y_∞ at long times in equation (171). In the same way, we computed the empirical loss on the noisy labels, the in-sample error writes:

$$\begin{aligned} \mathrm{Err}_{\text{in}} &= \frac{1}{4}\mathbb{E}\left(y_\infty - y^*\right)^2 \\ &= \frac{r_\infty^2}{2}\mathrm{MSE} + \frac{\Delta}{4}(1-r_\infty)^2. \end{aligned}$$

For $\alpha < \alpha_{\text{inter}}$, in the small regularization limit, we have $r_\infty = 0$ and therefore $\mathrm{Err}_{\text{in}} = \Delta/4$. Above the interpolation threshold, we have $r_\infty = 1 - I_\omega(\xi)/\alpha$, and replacing the value of the MSE, we get:

$$\mathrm{Err}_{\text{in}} = \alpha\xi\left(1 - \frac{I_\omega(\xi)}{\alpha}\right)^2 + \frac{\Delta}{4}\left(\frac{2I_\omega(\xi)}{\alpha} - 1\right).$$

H.1.2 The Noiseless Case

We now compute the interpolation threshold in the case where $\Delta = 0$, proving Corollary 10. In this case we show that the interpolation threshold is reached at $\xi = 0$, meaning that in the small regularization limit, the interpolation threshold is larger than the perfect recovery one.

To prove the result, we can combine equation (320) at $\Delta = 0$ with the identity $\alpha_{\text{inter}} = I_\omega(\xi)$ to obtain a relationship solely on ξ:

$$I_\omega(\xi) = \frac{1}{2\xi}\left(\int_{\max(0,\omega)} x^2 \mathrm{d}\mu_\xi(x) - Q_*\right). \tag{321}$$

Therefore, to prove the result, it is enough to show that this relationship is verified in the limit $\xi \to 0$. As shown in Lemma 45, for $\kappa^* < 1$, we have $\omega \sim \sqrt{(1-\kappa^*)\xi}\tilde{\omega}$ as $\xi \to 0$, with:

$$\frac{\min(\kappa, 1) - \kappa^*}{1 - \kappa^*} = \int_{\tilde{\omega}} \mathrm{d}\sigma(x).$$

Now using Lemma 46 with $q, h = 0$, and using the definition of $I_\omega(\xi)$ in equation (318), we have:

$$\int_{\max(0,\omega)} x^2 \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} Q_* + 2\xi\left(\kappa^* - \frac{\kappa^{*2}}{2}\right) + \xi(1-\kappa^*)^2 \int_{\max(0,\tilde{\omega})} x^2 \mathrm{d}\sigma(x) + o(\xi), \tag{322}$$

$$I_\omega(\xi) \underset{\xi\to 0}{=} \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(0,\tilde{\omega})} x^2 \mathrm{d}\sigma(x) + o(1). \tag{323}$$

Rearranging, this shows that equation (321) is verified in the $\xi \to 0$ limit. We then obtain the result of Corollary 10 since α_{inter} is now given by $\lim_{\xi\to 0} I_\omega(\xi)$.

For the case $\kappa, \kappa^* \geq 1$, the result is simply an application of Lemma 47. Indeed, since $\kappa \geq 1$, there is no need for the threshold ω. We then find the desired value $\alpha_{\text{inter}}(\kappa, \kappa^*) = 1/2$.

H.2 Perfect Recovery Threshold

In this section we show Proposition 12 that gives the expression of the perfect recovery threshold in the small regularization limit. To do so, we consider the first part of Proposition 11 with $\Delta = 0$.

To derive the value of the perfect recovery threshold, remark that as $\alpha \to \alpha_{\text{PR}}^+$, we have $Z_\infty \to Z^*$ and from equation (58) we can conclude that the variables q, ξ vanish. Moreover, when $\kappa^* < 1$, we show in Lemma 45 that the cutoff ω needs to scale with ξ as $\omega \sim \sqrt{(1-\kappa^*)\xi}\tilde{\omega}$. Finally, we choose the scaling $q \sim \sqrt{(1-\kappa^*)\xi}h$, with $h = \Theta(1)$. Using the system of equations in the small regularization limit, we then have the equations describing the perfect recovery threshold:

$$\alpha_{\text{PR}}^+ = \lim_{\xi\to 0} \int_{\max(q,\omega)} (x - q) h_\xi(x) \mathrm{d}\mu_\xi(x), \tag{324}$$

$$\alpha_{\text{PR}}^+ = \lim_{\xi\to 0} \frac{1}{2\xi}\left(Q_* + \int_{\max(q,\omega)} (q^2 - x^2)\mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x)\right), \tag{325}$$

and one has to solve this for the variables $\alpha_{\rm PR}^+, q$. Then, taking into account the scalings of ω, q as $\xi \to 0$, we have the asymptotics from Lemma 46, for $\kappa^* < 1$:

$$\lim_{\xi\to 0} \frac{1}{\xi}\left(Q_* + \int_{\max(q,\omega)} (q^2 - x^2) \mathrm{d}\mu_\xi(x)\right) = (1-\kappa^*)h^2\left(\kappa^* + (1-\kappa^*)\int_{\max(h,\tilde{\omega})} \mathrm{d}\sigma(x)\right) - 2\left(\kappa^* - \frac{\kappa^{*2}}{2}\right) - (1-\kappa^*)^2 \int_{\max(h,\tilde{\omega})} x^2 \mathrm{d}\sigma(x),$$

$$\lim_{\xi\to 0} \int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x) = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(h,\tilde{\omega})} x(x-h) \mathrm{d}\sigma(x).$$

Rearranging, we get the system:

$$\alpha_{\rm PR}^+ = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2} \int_{\max(h,\tilde{\omega})} x(x-h) \mathrm{d}\sigma(x),$$

$$\alpha_{\rm PR}^+ = \kappa^* - \frac{\kappa^{*2}}{2} + \frac{\kappa^*(1-\kappa^*)}{2} h^2 + \frac{(1-\kappa^*)^2}{2} \int_{\max(h,\tilde{\omega})} (x-h)^2 \mathrm{d}\sigma(x).$$

This leads to the following equation on h:

$$\kappa^* h^2 = (1-\kappa^*) h \int_{\max(h,\tilde{\omega})} (x-h) \mathrm{d}\sigma(x).$$

Now, there are two possible solutions to this equation: the one given in Proposition 12, and $h = 0$, which leads back to the interpolation threshold. Now, one needs to take the solution that leads to the smallest value of $\alpha_{\rm PR}^+$, since the perfect recovery threshold is defined as the minimal value of α such that the MSE vanishes. Therefore we need to choose the solution $h > 0$ and obtain the result of Proposition 12.

Whenever $\kappa, \kappa^* \geq 1$, we simply apply Lemma 47, in which case there is no need for the threshold ω. Then, equation (324) leads to $\alpha_{\rm PR}^+ = 1/2$.

H.3 Minimal Regularization Interpolator

In this section, we prove Proposition 13, which concerns the minimal regularization interpolator. This result is already known, but we give a proof for completeness. We recall that the loss we consider here writes $\mathcal{L}_\lambda = \mathcal{L} + \lambda\Omega$. We denote $\mathcal{S}^* = \mathrm{argmin}(\mathcal{L})$ and assume without loss of generality that $\mathcal{S}^* = \mathcal{L}^{-1}(\{0\})$. We let $(W_\lambda)_{\lambda>0}$ to be such that W_λ is a global minimizer of $\mathcal{L}_\lambda$ for all $\lambda > 0$.

Due to the optimality of W_λ for $\mathcal{L}_\lambda$, we have for any $V \in \mathcal{S}^*$:

$$\mathcal{L}(W_\lambda) + \lambda\Omega(W_\lambda) \leq \lambda\Omega(V), \tag{326}$$

where we used that $\mathcal{L}(V) = 0$. Therefore $\Omega(W_\lambda) \leq \Omega(V)$, i.e., W_λ has a smaller regularization than any element of $\mathcal{S}^*$. In addition, since Ω is coercive, the sublevel set $\{W, \Omega(W) \leq \Omega(V)\}$ is compact, and so is the family $(W_\lambda)_{\lambda>0}$. Finally, taking the $\lambda \to 0$ limit in equation (326) leads to the fact that $\mathcal{L}_\lambda(W_\lambda) \xrightarrow[\lambda\to 0]{} 0$.

Now, we let $\overline{W}$ be a cluster point of W_λ, and a sequence $\lambda_k \xrightarrow[k\to\infty]{} 0$ such that $W_{\lambda_k} \xrightarrow[k\to\infty]{} \overline{W}$. By continuity of $\mathcal{L}$ and since $\mathcal{L}(W) \leq \mathcal{L}_\lambda(W)$ for all W, we have:

$$\mathcal{L}\big(\overline{W}\big) = \lim_{k\to\infty} \mathcal{L}(W_{\lambda_k}) \leq \lim_{k\to\infty} \mathcal{L}_{\lambda_k}\big(W_{\lambda_k}\big) = 0.$$

Therefore $\overline{W} \in \mathcal{S}^*$, and $\overline{W}$ minimizes Ω on $\mathcal{S}^*$ since $\Omega(W_\lambda) \leq \Omega(V)$ for all $V \in \mathcal{S}^*$ and $\lambda > 0$. Now, if such an element is unique, the family $(W_\lambda)_{\lambda>0}$ is compact and has a single cluster point, therefore it converges to $\overline{W}$.

H.4 Small Noise Asymptotics

In the following, we derive the small ξ asymptotics used in the proofs of Corollary 10 and Proposition 12. These computations rely on the understanding of the measure μ_ξ at small ξ. We recall that we define the measure μ_ξ and give some basic properties in Section B.1.3. The following lemma gives an expansion of the density ρ_ξ and the Hilbert transform of this measure up to the needed order.

Lemma 44 *Consider $\kappa^* \in (0,1)$. Let ρ_ξ be the density of μ_ξ and h_ξ its Hilbert transform (see Definition 29). Decompose:*

$$\mu^* = (1-\kappa^*)\delta + \kappa^*\nu^*,$$

where ν^* has a smooth density with a compact support away from zero (that we also denote ν^* for the sake of simplicity). Then, for all $x \neq 0$, as $\xi \to 0$:

$$\rho_\xi(x) = \kappa^*\nu^*(x) + \xi\kappa^*(1-\kappa^*)\left(\frac{\nu^*(x)}{x^2} - \frac{\partial_x\nu^*(x)}{x}\right) + \frac{\xi\kappa^{*2}}{\pi}\mathrm{Im}\Big(m_{\nu^*}(x)\partial_x m_{\nu^*}(x)\Big) + O(\xi^2),$$
$$h_\xi(x) = \kappa^* h_{\nu^*}(x) + \frac{1-\kappa^*}{x} + O(\xi),$$

where h_{ν^*} and m_{ν^*} are respectively the Hilbert and Stieltjes transforms of ν^* (see Definition 29), and $m_{\nu^*}(x)$ is understood as the limit $m_{\nu^*}(x+i\eta)$ as $\eta \to 0^+$.

In addition, for any $|y| < 2$, with $s = 1-\kappa^*$:

$$\rho_\xi\big(\sqrt{s\xi}y\big) = \sqrt{\frac{s}{\xi}}\frac{1}{2\pi}\sqrt{4-y^2} + O(1),$$
$$h_\xi\big(\sqrt{s\xi}y\big) = \sqrt{\frac{s}{\xi}}\frac{y}{2} + O(1).$$

Proof Let us start with the subordination equation between the Stieltjes transforms of μ^* and μ_ξ (see Lemma 32):

$$m_\xi(z) = m_*\big(z - \xi m_\xi(z)\big).$$

Due to the decomposition of μ^*, this also writes:

$$m_\xi(z) = \frac{1-\kappa^*}{z-\xi m_\xi(z)} + \kappa^* m_{\nu^*}\big(z - \xi m_\xi(z)\big), \tag{327}$$

Then decomposing $m_\xi(z) = m_0(z) + \xi m_1(z)$ at leading order, we arrive at:

$$m_0(z) = \frac{1-\kappa^*}{z} + \kappa^* m_{\nu^*}(z),$$
$$m_1(z) = m_0(z)\left(\frac{1-\kappa^*}{z^2} - \kappa^* \partial_z m_{\nu^*}(z)\right).$$

Taking $z = x + i\eta$ with η going to 0^+, we use the identity (see Definition 29):

$$m_\xi(x+i\eta) \to h_\xi(x) - i\pi\rho_\xi(x).$$

This immediately gives the relationship at first order for h_ξ, ρ_ξ (by considering m_0). Now at second order for ρ_ξ, we have:

$$\begin{aligned}\rho_1(x) &= -\frac{1}{\pi}\lim_{\eta\to 0^+} \operatorname{Im} m_1(x+i\eta) \\ &= \frac{\kappa^*(1-\kappa^*)}{x^2}\nu^*(x) - \frac{\kappa^*(1-\kappa^*)}{x}\partial_x\nu^*(x) + \frac{\kappa^{*2}}{\pi}\operatorname{Im}\Big(m_{\nu^*}(x)\partial_x m_{\nu^*}(x)\Big).\end{aligned}$$

This gives the result knowing that $\rho_\xi(x) = \kappa^*\nu^*(x) + \xi\rho_1(x) + O(\xi^2)$.

Let us now consider the microscopic part, i.e., when $x = \sqrt{s\xi}y$ for some $y = O(1)$. To do so, we start from equation (327) with $z = \sqrt{s\xi}y + i\eta$ and take the limit $\eta \to 0$, to obtain:

$$\begin{aligned}h_\xi\big(\sqrt{s\xi}y\big) - i\pi\rho_\xi\big(\sqrt{s\xi}y\big) = \frac{1-\kappa^*}{\sqrt{s\xi}y - \xi h_\xi\big(\sqrt{s\xi}y\big) + i\pi\xi\rho_\xi\big(\sqrt{s\xi}y\big)} \\ + \kappa^* m_{\nu^*}\Big(\sqrt{s\xi}y - \xi h_\xi\big(\sqrt{s\xi}y\big) + i\pi\xi\rho_\xi\big(\sqrt{s\xi}y\big)\Big).\end{aligned}$$

Looking at the first term, we should have $h_\xi\big(\sqrt{s\xi}y\big)$ and $\rho_\xi\big(\sqrt{s\xi}y\big)$ of order $\xi^{-1/2}$. In this case the second term remains of order one and we can ignore it at leading order. Let us define:

$$\tilde{h}(y) = \lim_{\xi\to 0}\sqrt{\frac{\xi}{s}}h_\xi\big(\sqrt{s\xi}y\big), \qquad \tilde{\rho}(y) = \lim_{\xi\to 0}\sqrt{\frac{\xi}{s}}\rho_\xi\big(\sqrt{s\xi}y\big). \tag{328}$$

Then, we get the equation:

$$\tilde{h}(y) - i\pi\tilde{\rho}(y) = \frac{1}{y - \tilde{h}(y) + i\pi\tilde{\rho}(y)}.$$

Solving for $\tilde{m}(y) = \tilde{h}(y) - i\pi\tilde{\rho}(y)$, we get that $\tilde{m}(y)^2 - y\tilde{m}(y) + 1 = 0$, and solving for $|y| \le 2$, we obtain:

$$\tilde{h}(y) = \frac{y}{2}, \qquad \tilde{\rho}(y) = \frac{1}{2\pi}\sqrt{4-y^2}.$$

This leads to the result with the definition of $\tilde{h}, \tilde{\rho}$ in equation (328). Note that we have recovered the density and Hilbert transform of the semicircular density. ■

Interestingly, such a result holds for arbitrary measure ν^*, provided that it admits a smooth density. Before computing the asymptotics of the relevant quantities in our high-dimensional equations, we recall the definition of the semicircular distribution:

$$\mathrm{d}\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{|x|\leq 2}\mathrm{d}x.$$

As underlined by the previous lemma, this corresponds to the limiting measure of μ_ξ (with an appropriate rescaling) in the microscopic part $x = \Theta\big(\sqrt{\xi}\big)$.

Lemma 45 *Assume $0 < \kappa^* < \kappa < 1$ and consider ω to be solution of the equation:*

$$\kappa = \int_\omega \mathrm{d}\mu_\xi(x).$$

Then, if μ^* is decomposed as $\mu^* = (1-\kappa^*)\delta + \kappa^*\nu^*$, we have as $\xi \to 0$, $\omega \sim \sqrt{(1-\kappa^*)\xi}\tilde{\omega}$, where $\tilde{\omega}$ is solution of the equation:

$$\frac{\kappa - \kappa^*}{1-\kappa^*} = \int_{\tilde{\omega}} \mathrm{d}\sigma(x).$$

Proof For $\kappa > \kappa^*$, it is clear that ω goes to zero with ξ, since ω selects a proportion κ of the mass of μ_ξ. If ω were to remain of order 1, it would select either a proportion κ^* corresponding to the mass of ν^* (if $\omega > 0$) or a mass of one if $\omega < 0$. Therefore, we let $\omega \sim \sqrt{s\xi}\tilde{\omega}$, with $s = 1-\kappa^*$.

Then, using Lemma 44, we decompose the integral between $x = O(1)$ and $x = O(\sqrt{\xi})$ and obtain the asymptotics:

$$\begin{aligned}\int \mathbf{1}_{x\geq\omega}\mathrm{d}\mu_\xi(x) &\underset{\xi\to 0}{=} \kappa^* \int \mathrm{d}\nu^*(x) + \sqrt{s\xi}\int_{|y|\leq 2}\mathbf{1}_{y\geq\tilde{\omega}}\rho_\xi(\sqrt{s\xi}y)\mathrm{d}y + o(1)\\ &\xrightarrow[\xi\to 0]{} \kappa^* + (1-\kappa^*)\int \mathbf{1}_{x\geq\tilde{\omega}}\mathrm{d}\sigma(x).\end{aligned}$$

Since the integral on the left side is equal to κ, we get the derived self-consistent equation on $\tilde{\omega}$. When $\tilde{\omega}$ ranges from -2 to 2 (the edges of σ), κ goes from 1, in which case it selects the whole spectrum, to κ^* where it only retains the positive mass of ν^*. ■

Although we assumed $\kappa^* < \kappa < 1$, the previous result still applies to the limit cases:

- $\kappa = \kappa^* < 1$. In this case there is no need to cut through the semicircular density, and we can pick any scaling $\omega \sim \sqrt{s\xi}\tilde{\omega}$ with $\tilde{\omega} > 2$.
- $\kappa^* < 1 \leq \kappa$. Now ω needs to select the whole spectrum of μ_ξ, so we can take the same scaling $\omega \sim \sqrt{s\xi}\tilde{\omega}$ with $\tilde{\omega} < -2$.

With this understanding of the threshold ω in the small ξ limit, we can now compute the following asymptotics:

Lemma 46 *Consider $\kappa^* \in (0,1)$ and $\kappa \geq \kappa^*$. Then, we have, with $\omega \sim \sqrt{s\xi}\tilde{\omega}$ and $q \sim \sqrt{s\xi}h$:*

$$\int_{\max(q,\omega)} \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} \kappa^* + (1-\kappa^*)\int_{\max(h,\tilde{\omega})} \mathrm{d}\sigma(x) + o(1), \tag{329}$$

$$\int_{\max(q,\omega)} x^2 \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} Q_* + 2\xi\left(\kappa^* - \frac{\kappa^{*2}}{2}\right) + \xi(1-\kappa^*)^2 \int_{\max(h,\tilde{\omega})} x^2 \mathrm{d}\sigma(x) + o(\xi), \tag{330}$$

$$\int_{\max(q,\omega)} (x-q)h_\xi(x)\mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} \kappa^* - \frac{\kappa^{*2}}{2} + \frac{(1-\kappa^*)^2}{2}\int_{\max(h,\tilde{\omega})} x(x-h)\mathrm{d}\sigma(x) + o(1). \tag{331}$$

Proof The derivation of equation (329) is precisely the same as in Lemma 45 with a slightly different threshold. Then, let us show equation (330). We have, as a consequence of Lemma 44:

$$\begin{aligned}\int_{\max(q,\omega)} x^2 \mathrm{d}\mu_\xi(x) = \kappa^* \int x^2 \mathrm{d}\nu^*(x) - \xi\kappa^*(1-\kappa^*)\int x^2 \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\nu^*(x)}{x}\right)\mathrm{d}x \\ + \frac{\xi\kappa^{*2}}{\pi}\int x^2 \operatorname{Im}\Big(m_{\nu^*}(x)\partial_x m_{\nu^*}(x)\Big)\mathrm{d}x \\ + (1-\kappa^*)^2\xi\int_{\max(h,\tilde{\omega})} x^2 \mathrm{d}\sigma(x) + o(\xi).\end{aligned} \tag{332}$$

In the last expression, the three first term come from the expansion of $\rho_\xi(x)$ for x fixed and non-zero, and the second comes from the microscopic behavior of $\rho_\xi(x)$ with $x = O(\sqrt{\xi})$. Also $m_{\nu^*}(x)$ is understood as the limit of $m_*(x+i\eta)$ as $\eta \to 0^+$. Now, integrating by parts:

$$\int x^2 \operatorname{Im}\Big(m_{\nu^*}(x)\partial_x m_{\nu^*}(x)\Big)\mathrm{d}x = \left[\frac{x^2}{2}\operatorname{Im}\big(m_{\nu^*}(x)^2\big)\right] - \int x \operatorname{Im}\big(m_{\nu^*}(x)^2\big)\mathrm{d}x.$$

Now, since $m_{\nu^*}(x+i\eta) = h_{\nu^*}(x) - i\pi\nu^*(x)$ as $\eta \to 0^+$, we have:

$$\operatorname{Im}\big(m_{\nu^*}(x)^2\big) = -2\pi\nu^*(x)h_{\nu^*}(x),$$

so the bracket vanishes. Therefore:

$$\int x^2 \operatorname{Im}\Big(m_{\nu^*}(x)\partial_x m_{\nu^*}(x)\Big)\mathrm{d}x = 2\pi\int x h_{\nu^*}(x)\mathrm{d}\nu^*(x).$$

Now, using Lemma 31, we have:

$$\int x h_{\nu^*}(x)\mathrm{d}\nu^*(x) = \frac{1}{2}.$$

Now going back to equation (332), we have, integrating by parts:

$$\int x^2 \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\nu^*(x)}{x}\right)\mathrm{d}x = -2.$$

Therefore, we have:

$$\int_{\max(q,\omega)} x^2 \mathrm{d}\mu_\xi(x) = Q_* + 2\xi\left(\kappa^* - \frac{\kappa^{*2}}{2}\right) + \xi(1-\kappa^*)^2 \int_{\max(h,\tilde{\omega})} x^2 \mathrm{d}\sigma(x) + o(\xi),$$

which is precisely equation (330).

Let us show equation (331). We know from Lemma 44 that:

$$h_\xi(x) \underset{\xi\to 0}{\longrightarrow} \kappa^* h_{\nu^*}(x) + \frac{1-\kappa^*}{x}, \qquad h_\xi\big(x\sqrt{s\xi}\big) \underset{\xi\to 0}{\sim} \sqrt{\frac{s}{\xi}}\frac{x}{2}.$$

Therefore at first order:

$$\int_{\max(q,\omega)} (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x) = \kappa^{*2} \int x h_{\nu^*}(x) \mathrm{d}\nu^*(x) + \kappa^*(1-\kappa^*) + \frac{(1-\kappa^*)^2}{2} \int_{\max(h,\tilde{\omega})} x(x-h)\mathrm{d}\sigma(x) + o(1).$$

As shown before, the first integral is simply $1/2$. Combining the first and second term, we get the expected result. ■

To conclude these computations, we consider the case where $\kappa^* \geq 1$, and derive the same asymptotics as previously.

Lemma 47 *Consider the case $\kappa, \kappa^* \geq 1$. Then, as $\xi \to 0$, with $q \to 0$:*

$$\int_q \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{\longrightarrow} 1,$$
$$\int_q x^2 \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} Q_* + \xi + o(\xi),$$
$$\int_q (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{\longrightarrow} \frac{1}{2}.$$

Proof In this case, the measure μ_ξ converges smoothly to μ^* and there is no semicircular density near zero. Therefore, we can adapt the result of Lemma 44 and obtain for all $x \geq 0$:

$$\rho_\xi(x) = \mu^*(x) + \frac{\xi}{\pi}\mathrm{Im}\Big(m_*(x)\partial_x m_*(x)\Big) + O(\xi^2)$$
$$h_\xi(x) = h^*(x) + o(1),$$

with h^*, m_* being the Hilbert and Stieltjes transform of μ^* (see Definition 29). Therefore, at the necessary orders:

$$\int_q \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{\longrightarrow} \int \mathrm{d}\mu^*(x) = 1,$$
$$\int_q x^2 \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{=} \int x^2 \mathrm{d}\mu^*(x) + \frac{\xi}{\pi}\int x^2\, \mathrm{Im}\Big(m_*(x)\partial_x m_*(x)\Big)\mathrm{d}x + o(\xi),$$
$$\int_q (x-q) h_\xi(x) \mathrm{d}\mu_\xi(x) \underset{\xi\to 0}{\longrightarrow} \int x h^*(x) \mathrm{d}\mu^*(x).$$

Now we have shown in the proof of Lemma 46 that:

$$\frac{1}{\pi}\int x^2 \operatorname{Im}\Big(m_*(x)\partial_x m_*(x)\Big)\mathrm{d}x = 2\int x h^*(x)\mathrm{d}\mu^*(x).$$

Finally, as a consequence of Lemma 31, we have:

$$\int x h^*(x)\mathrm{d}\mu^*(x) = \frac{1}{2},$$

and putting everything together, this concludes the proof. ■

Appendix I. Proofs of the Results on the Oja Flow

In this section, we prove the results on the Oja flow presented in Section 4. More precisely:

- In Section I.1, we characterize the limit of the Oja flow by studying the local minimizers of the potential in equation (81).
- We then derive finite-dimensional convergence rates, both in the full-rank setting (Sections I.2, I.3) and in the rank-deficient case (Section I.4).
- Next, we study the high-dimensional regime by proving the limit of the distance to convergence (Section I.5) and the corresponding long-time convergence rates (Section I.6).
- Finally, we analyze the linear response of the Oja flow, first in finite dimension (Section I.7) and then in the high-dimensional limit (Section I.8).

I.1 Proof of Proposition 18

In the following, we prove Proposition 18, giving the limit of the Oja flow dynamics (80) at long times, under a random initialization. To do so, we study the local minimizers of the associated loss and invoke the stable manifold theorem (see Smale, 1963). We recall that the dynamics we study is the gradient flow associated with the loss defined in equation (81):

$$U(W) = \frac{1}{4}\big\|WW^\top - A\big\|_F^2.$$

The gradient and Hessian of U are given by:

$$\nabla U(W) = \big(WW^\top - A\big)W, \tag{333}$$

$$\operatorname{Tr}\big(\mathrm{d}^2 U_W(K)K^\top\big) = \operatorname{Tr}\Big(\big(WW^\top - A\big)KK^\top\Big) + \operatorname{Tr}\big(KW^\top KW^\top\big) + \operatorname{Tr}\big(KW^\top WK^\top\big), \tag{334}$$

for $K \in \mathbb{R}^{d\times m}$. Now, if $\nabla U(W) = 0$, then with $Z = WW^\top$, we have $Z^2 = AZ$, so that Z and A commute. Thus, we can diagonalize Z in the same basis as A. Then, we write the eigenvalues and singular values decompositions for A and W:

$$A = \sum_{k=1}^{d} \lambda_k u_k u_k^\top, \qquad W = \sum_{k=1}^{\min(m,d)} \sqrt{\mu_k} v_k u_k^\top.$$

Plugging these expressions into the equation $\nabla U(W) = 0$ leads to $\mu_k \in \{0, \lambda_k\}$ for all $k \in \{1, \dots, \min(m, d)\}$. Of course if $\lambda_k < 0$ then $\mu_k = 0$ since Z is positive semidefinite. We now assume that $W \in \mathbb{R}^{d\times m}$ is a local minimizer and evaluate equation (334) with $K = v_j u_i^\top$. Under the constraint that $\mathrm{Tr}\big(\mathrm{d}^2 U_W(K) K^\top\big) \geq 0$, we get:

$$\mu_j - \lambda_j + \mu_i \delta_{ij} + \mu_i \geq 0.$$

We then choose the indices $i \neq j$ such that $\mu_i = \lambda_i > 0$ and $\lambda_j > \lambda_i$. Then, we obtain that $\mu_j \geq \lambda_j - \lambda_i > 0$ so that necessarily $\mu_j = \lambda_j$. Therefore, as soon as Z matches a positive eigenvalue of A, it needs to match the larger eigenvalues of A.

We now show that Z recovers as many eigenvalues as it is possible under the rank constraint. To do so, we assume that $\mathrm{rank}(W) < m$ and introduce $u \in \mathbb{R}^m$ non-zero such that $Wu = 0$. Then evaluating equation (334) with $K = v_j u^\top$ we get $\mu_j \geq \lambda_j$. Therefore if $\lambda_j > 0$, necessarily $\mu_j = \lambda_j$ and Z matches all the positive eigenvalues of A. Otherwise $\mathrm{rank}(W) = m$ and the previous reasoning shows that Z recovers the m largest positive eigenvalues of A. Therefore, we can conclude the equivalence:

$$W \text{ local minimizer of } U \iff WW^\top = A^+_{(m)}.$$

We now use the stable manifold theorem (Smale, 1963): since U is analytic in the coefficients of W, and the gradient flow $W(t)$ associated with U is initialized with a random matrix W_0 whose distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{d\times m}$, then with probability one, $W(t)$ converges toward some local minimizer of U. This leads to the result.

I.2 Proof of Proposition 19

We now prove Proposition 19 that gives the exponentially fast convergence of the gradient flow dynamics $\dot{W}(t) = -\nabla F(W(t))$ when F is of the form $F(W) = G(WW^\top)$, for some twice continuously differentiable $G\colon \mathcal{S}_d(\mathbb{R}) \to \mathbb{R}$. Note that the result we show in the following only applies whenever $W(t)$ converges to a full-rank matrix. We start with a few useful lemmas.

Lemma 48 *Consider the family of subspaces indexed by $W \in \mathbb{R}^{d\times m}$:*

$$\mathcal{H}_W = \Big\{ K \in \mathbb{R}^{d\times m},\, W^\top K = K^\top W \Big\}, \tag{335}$$

and π_W to be the orthogonal projection onto $\mathcal{H}_W$. Then, for $m \leq d$, the map $W \mapsto \pi_W$ is continuous on the open set $\big\{W \in \mathbb{R}^{d\times m},\, \mathrm{rank}(W) = m\big\}$.

Proof It is known (see Massart and Absil, 2020) that the orthogonal complement $\mathcal{V}_W$ of $\mathcal{H}_W$ is given by:

$$\mathcal{V}_W = \big\{ W\Omega,\, \Omega \in \mathcal{A}_m(\mathbb{R}) \big\}, \tag{336}$$

where $\mathcal{A}_m(\mathbb{R})$ is the set of $m \times m$ skew-symmetric matrices. Now decomposing some $K \in \mathbb{R}^{d\times m}$ as $K = H + W\Omega$, where $H \in \mathcal{H}_W$ and $\Omega \in \mathcal{A}_m(\mathbb{R})$, we obtain:

$$W^\top K - K^\top W = \phi_{W^\top W}(\Omega), \qquad \phi_M(X) = MX + XM.$$

Now, provided that M is invertible, ϕ_M is also invertible, and we get the expression:

$$\pi_W(K) = K - W\phi^{-1}_{W^\top W}\big(W^\top K - K^\top W\big).$$

This leads to the result thanks to the continuity of the inverse. ■

Before proving Proposition 19, we give the structure of the Hessian of the function F at a critical point.

Lemma 49 *Let $W \in \mathbb{R}^{d\times m}$ such that $\nabla F(W) = 0$ and $\operatorname{rank}(W) = m$. Then the Hessian $\mathrm{d}^2 L_W$ has the following representation:*

$$\begin{array}{c} \begin{array}{cc} \mathcal{H}_W & \mathcal{V}_W \end{array} \\ \begin{pmatrix} * & 0 \\ 0 & 0 \end{pmatrix} \begin{array}{c} \mathcal{H}_W \\ \mathcal{V}_W \end{array} \end{array},$$

where the orthogonal subspaces $\mathcal{H}_W$ and $\mathcal{V}_W$ are defined in equations (335), (336).

Proof Let us show first of all that for all $H \in \mathbb{R}^{d\times m}$, $\mathrm{d}^2 L_W(H) \in \mathcal{H}_W$, whenever W has full rank and $\nabla F(W) = 0$. First of all, remark that, due to the structure of $F(W) = G(WW^\top)$:

$$\nabla F(W) = 2\nabla G(WW^\top)W \in \mathcal{H}_W.$$

Now, as $\epsilon \to 0$, since $\nabla F(W) = 0$ and F is $\mathcal{C}^2$:

$$\nabla F(W + \epsilon H) = \epsilon \mathrm{d}^2 F_W(H) + o(\epsilon). \tag{337}$$

Now, let $K \in \mathcal{V}_W$ and denote $W_\epsilon = W + \epsilon H$. We have:

$$K = \pi^\perp_W(K) = \pi^\perp_{W_\epsilon}(K) + \big(\pi^\perp_W - \pi^\perp_{W_\epsilon}\big)(K),$$

where $\pi^\perp_W$ denotes the projection onto $\mathcal{V}_W$. Then, since $\nabla F(W_\epsilon) \in \mathcal{H}_{W_\epsilon}$ is orthogonal to $\mathcal{V}_{W_\epsilon}$:

$$\Big|\operatorname{Tr}\big(\nabla F(W_\epsilon)K^\top\big)\Big| \leq \big\|\pi^\perp_W - \pi^\perp_{W_\epsilon}\big\|_{op}\big\|\nabla F(W_\epsilon)\big\|\big\|K\big\|.$$

Now by Lemma 48, $V \mapsto \pi^\perp_V$ is continuous at $V = W$, and using equation (337), we have that the previous quantity is $o(\epsilon)$. Taking the scalar product with K in equation (337), we obtain that:

$$\operatorname{Tr}\big(\mathrm{d}^2 F_W(H)K^\top\big) = 0.$$

Therefore, for all $H \in \mathbb{R}^{d\times m}$, $\mathrm{d}^2 L_W(H) \in \mathcal{H}_W$. Now, since $\mathrm{d}^2 L_W$ is self-adjoint, for any $H \in \mathbb{R}^{d\times m}$ and $K \in \mathcal{V}_W$:

$$\operatorname{Tr}\big(\mathrm{d}^2 L_W(K)H^\top\big) = \operatorname{Tr}\big(\mathrm{d}^2 L_W(H)K^\top\big) = 0,$$

since $\mathrm{d}^2 L_W(H) \in \mathcal{H}_W$. This shows that $\mathrm{d}^2 L_W$ is zero on $\mathcal{V}_W$ and concludes the proof. ■

With these tools, we are now ready to prove Proposition 19. In the following, we will denote $\pi_t, \pi_t^\perp, \pi_\infty, \pi_\infty^\perp$ the orthogonal projections onto $\mathcal{H}_{W(t)}, \mathcal{V}_{W(t)}, \mathcal{H}_{W_\infty}$ and $\mathcal{V}_{W_\infty}$. First of all, we start with the standard identity for gradient flow:

$$F(W(t)) - F(W_\infty) = \int_t^\infty \big\|\nabla F(W(s))\big\|_F^2 \mathrm{d}s. \tag{338}$$

This can be proven by differentiating the function $t \mapsto F(W(t))$ and using the gradient flow structure. Now, since $\nabla F(W(s)) \in \mathcal{H}_{W(s)}$, and by continuity of the projection derived in Lemma 48, we have:

$$\big\|\nabla F(W(s))\big\|_F^2 \underset{s\to\infty}{\sim} \big\|\pi_\infty\big(\nabla F(W(s))\big)\big\|_F^2. \tag{339}$$

We now define:

$$\phi(t) = \frac{1}{2}\big\|\pi_\infty\big(\nabla F(W(t))\big)\big\|_F^2.$$

It is easily shown that:

$$\dot\phi(t) = -\mathrm{Tr}\Big(\mathrm{d}^2 F_{W(t)}\big(\nabla F(W(t))\big)\pi_\infty\big(\nabla F(W(t))\big)^\top\Big). \tag{340}$$

We decompose this term by introducing $\mathrm{d}^2 F_{W_\infty}$. We have, using the Cauchy–Schwarz inequality:

$$\begin{aligned}\Big|\mathrm{Tr}\Big(\big(\mathrm{d}^2 F_{W(t)} - \mathrm{d}^2 F_{W_\infty}\big)\big(\nabla F(W(t))\big)\pi_\infty\big(\nabla F(W(t))\big)^\top\Big)\Big| \\ \leq \big\|\mathrm{d}^2 F_{W(t)} - \mathrm{d}^2 F_{W_\infty}\big\|_{op}\big\|\nabla F(W(t))\big\|_F\big\|\pi_\infty\big(\nabla F(W(t))\big)\big\|_F.\end{aligned} \tag{341}$$

Since F is $\mathcal{C}^2$ and using equation (339), this quantity is of the form $\epsilon(t)\phi(t)$ with $\epsilon(t) \xrightarrow[t\to\infty]{} 0$. Moreover, with the representation of the Hessian in Lemma 49 and the fact that it is positive definite on $\mathcal{H}_{W_\infty}$ with smallest eigenvalue ϱ, we have the lower bound:

$$\mathrm{Tr}\Big(\mathrm{d}^2 F_{W_\infty}\big(\nabla F(W(t))\big)\pi_\infty\big(\nabla F(W(t))\big)^\top\Big) \geq 2\varrho\phi(t). \tag{342}$$

Therefore, gathering equations (340), (341), (342), we have:

$$\dot\phi(t) \leq -2\big(\varrho - \epsilon(t)\big)\phi(t).$$

Now, given some $c < \varrho$, integrating the previous inequality leads to:

$$\phi(t) = o(e^{-2ct}),$$

as $t \to \infty$. Combining this with equations (338), (339), we get:

$$F(W(t)) - F(W_\infty) \underset{t\to\infty}{\sim} 2\int_t^\infty \phi(s)\mathrm{d}s = o(e^{-2ct}).$$

For the distance to convergence, we have due to the gradient flow structure:

$$\big\|W(t) - W_\infty\big\|_F \leq \int_t^\infty \big\|\nabla F(W(s))\big\|_F \mathrm{d}s = o(e^{-ct}).$$

Therefore, with $Z(t) = W(t)W(t)^\top$ and $Z_\infty = W_\infty W_\infty^\top$:

$$\begin{aligned}\big\|Z(t) - Z_\infty\big\|_F &\leq \big\|(W(t) - W_\infty)W(t)^\top\big\|_F + \big\|W_\infty(W(t) - W_\infty)^\top\big\|_F \\ &\leq \big(\|W(t)\|_F + \|W_\infty\|_F\big)\big\|W(t) - W_\infty\|_F \\ &= o(e^{-ct}).\end{aligned}$$

This concludes the proof.

I.3 Proof of Proposition 20

Let us now conclude the proof of the convergence rates in the case where the Oja flow dynamics converges to a full-rank local minimizer. To do so, we study the eigenvalues of the Hessian:

$$\mathrm{d}^2 U_W(H) = \big(WW^\top - A\big)H + HW^\top W + WH^\top W.$$

Given Proposition 18, we write $A = Q\Lambda Q^\top$ and $W = QDU^\top$, where:

$$\Lambda = \begin{pmatrix}\Lambda_1 & 0 \\ 0 & \Lambda_2\end{pmatrix}, \qquad D = \begin{pmatrix}\Lambda_1^{1/2} \\ 0\end{pmatrix}, \tag{343}$$

where $\Lambda_1 \in \mathbb{R}^{m\times m}$ is diagonal and gathers the m largest positive eigenvalues of A, and $\Lambda_2 \in \mathbb{R}^{(d-m)\times(d-m)}$ contains all the remaining eigenvalues. Moreover, $Q \in O_d(\mathbb{R})$ and $U \in O_m(\mathbb{R})$. Then, changing variables $H = QKU^\top$, we get:

$$\mathrm{d}^2 U_W(QKU^\top) = Q\Big(\big(DD^\top - \Lambda\big)K + KD^\top D + DK^\top D\Big)U^\top.$$

Since the mapping $K \in \mathbb{R}^{d\times m} \mapsto QKU^\top \in \mathbb{R}^{d\times m}$ is invertible, we can diagonalize the Hessian by considering the representation (343). We therefore consider the eigenvalue problem:

$$\big(DD^\top - \Lambda\big)K + KD^\top D + DK^\top D = \mu K. \tag{344}$$

To solve this, we decompose:

$$K = \begin{pmatrix}K_1 \\ K_2\end{pmatrix}, \qquad K_1 \in \mathbb{R}^{m\times m}, \quad K_2 \in \mathbb{R}^{(d-m)\times m}.$$

Since we restrict $K \in \mathcal{H}_W$, we have the relationship $\Lambda_1^{1/2}K_1 = K_1^\top \Lambda_1^{1/2}$. Then, equation (344) writes:

$$K_1\Lambda_1 + \Lambda_1^{1/2}K_1^\top\Lambda_1^{1/2} = \mu K_1, \tag{345}$$

$$K_2\Lambda_1 - \Lambda_2 K_2 = \mu K_2. \tag{346}$$

This corresponds to two separate eigenvalue equations that we can solve independently. Starting with equation (346), it is easily seen that the solutions are of the form $\lambda_i - \lambda_j$ for $i \in \{1,\dots,m\}$ and $j \in \{m+1,\dots,d\}$, with $\lambda_1 \geq \cdots \geq \lambda_d$ being the ordered eigenvalues of A. Now considering equation (345), we choose, for $1 \leq i \leq j \leq d$:

$$K_1 = e_i e_j^\top + \sqrt{\frac{\lambda_i}{\lambda_j}}e_j e_i^\top,$$

so that:

$$\Lambda_1^{1/2} K_1 = \sqrt{\lambda_i}\big(e_i e_j^\top + e_j e_i^\top\big) = K_1^\top \Lambda^{1/2},$$

and:

$$\begin{aligned} K_1\Lambda_1 + \Lambda_1^{1/2} K_1^\top \Lambda_1^{1/2} &= (\lambda_i + \lambda_j) e_i e_j^\top + \left(\lambda_i \sqrt{\frac{\lambda_i}{\lambda_j}} + \sqrt{\lambda_i \lambda_j}\right) e_j e_i^\top \\ &= (\lambda_i + \lambda_j) K_1. \end{aligned}$$

This corresponds to $\dfrac{1}{2}m(m+1)$ eigenvalues, which is the dimension of the admissible K_1 under the symmetry constraint $\Lambda_1^{1/2} K_1 = K_1^\top \Lambda_1^{1/2}$. This concludes the proof.

I.4 Proof of Proposition 21

Let us now prove the convergence rates of the Oja flow in the case where it converges toward a rank-deficient matrix. To do so, we decompose $A = Q\Lambda Q^\top$, with:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & -\Lambda_2 \end{pmatrix}. \tag{347}$$

In this case $\Lambda_1 \in \mathbb{R}^{p\times p}, \Lambda_2 \in \mathbb{R}^{(d-p)\times(d-p)}$ are diagonal and are both positive definite. Indeed, we assumed here that A was invertible. If this is not the case, it is known (see for instance Martin et al., 2024; Sarao Mannelli et al., 2020) that the convergence is not exponentially fast anymore. From the gradient flow dynamics:

$$\dot{W}(t) = \big(A - W(t)W(t)^\top\big)W(t),$$

we get the equation on $Z(t) = W(t)W(t)^\top$:

$$\dot{Z}(t) = \big(A - Z(t)\big)Z(t) + Z(t)\big(A - Z(t)\big).$$

Now, we consider $Y(t) = Q^\top Z(t) Q$ and obtain:

$$\dot{Y}(t) = \big(\Lambda - Y(t)\big)Y(t) + Y(t)\big(\Lambda - Y(t)\big). \tag{348}$$

In addition, we have $\|Z(t) - Z_\infty\|_F^2 = \|Y(t) - Y_\infty\|_F^2$, so we restrict ourselves to the study of $Y(t)$. In this case, we have from Proposition 18:

$$Y_\infty = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We then decompose:

$$Y(t) = \begin{pmatrix} Y_1(t) & Y_2(t) \\ Y_2(t)^\top & Y_3(t) \end{pmatrix},$$

where $Y_1(t) \in \mathcal{S}_p^+(\mathbb{R}), Y_3(t) \in \mathcal{S}_{d-p}^+(\mathbb{R})$ and $Y_2(t) \in \mathbb{R}^{p\times(d-p)}$. Then, equation (348) leads to the dynamics:

$$\begin{aligned} \dot{Y}_1(t) &= (\Lambda_1 - Y_1(t))Y_1(t) + Y_1(t)(\Lambda_1 - Y_1(t)) - 2Y_2(t)Y_2(t)^\top, \\ \dot{Y}_2(t) &= (\Lambda_1 - 2Y_1(t))Y_2(t) - Y_2(t)(\Lambda_2 + 2Y_3(t)), \\ \dot{Y}_3(t) &= -(\Lambda_2 + Y_3(t))Y_3(t) - Y_3(t)(\Lambda_2 + Y_3(t)) - 2Y_2(t)^\top Y_2(t). \end{aligned}$$

In order to study the distance to Y_∞, we define the functions:

$$\phi_1(t) = \big\|Y_1(t) - \Lambda_1\big\|_F^2, \qquad \phi_2(t) = \big\|Y_2(t)\big\|_F^2, \qquad \phi_3(t) = \big\|Y_3(t)\big\|_F^2.$$

We obtain the dynamics:

$$\begin{aligned}
\frac{1}{2}\dot{\phi}_1(t) &= -2\mathrm{Tr}\Big(Y_1(t)(Y_1(t)-\Lambda_1)^2\Big) - 2\mathrm{Tr}\Big(Y_2(t)Y_2(t)^\top(Y_1(t)-\Lambda_1)\Big),\\
\frac{1}{2}\dot{\phi}_2(t) &= \mathrm{Tr}\Big(Y_2(t)Y_2(t)^\top(\Lambda_1 - 2Y_1(t))\Big) - \mathrm{Tr}\Big(Y_2(t)^\top Y_2(t)(\Lambda_2 + 2Y_3(t))\Big),\\
\frac{1}{2}\dot{\phi}_3(t) &= -2\mathrm{Tr}\Big(Y_3(t)^2(Y_3(t)+\Lambda_2)\Big) - 2\mathrm{Tr}\Big(Y_2(t)^\top Y_2(t)Y_3(t)\Big).
\end{aligned}$$

Since $Y_2^\top Y_2$ and Y_3 are both PSD, we have $\mathrm{Tr}\big(Y_2^\top Y_2 Y_3\big) \geq 0$ and $\mathrm{Tr}(Y_3^3) \geq 0$, so that:

$$\begin{aligned}
\dot{\phi}_1(t) &\leq -4\lambda_{\min}\big(Y_1(t)\big)\phi_1(t) - 4\lambda_{\min}\big(Y_1(t)-\Lambda_1\big)\phi_2(t),\\
\dot{\phi}_2(t) &\leq -2\Big(\lambda_{\min}\big(Y_1(t)-\Lambda_1\big) + \lambda_{\min}\big(Y_1(t)\big) + \lambda_{\min}(\Lambda_2)\Big)\phi_2(t),\\
\dot{\phi}_3(t) &\leq -4\lambda_{\min}(\Lambda_2)\phi_3(t),
\end{aligned}$$

where we used the inequality:

$$\lambda_{\min}(B)\mathrm{Tr}(A) \leq \mathrm{Tr}(AB),$$

for $B \in \mathcal{S}_d(\mathbb{R})$ and $A \in \mathcal{S}_d^+(\mathbb{R})$. Therefore, defining:

$$\begin{aligned}
\alpha_1(t) &= \int_0^t \lambda_{\min}\big(Y_1(s)\big)\mathrm{d}s,\\
\alpha_2(t) &= \alpha_1(t) + \lambda_{\min}(\Lambda_2)t + \int_0^t \lambda_{\min}\big(Y_1(s)-\Lambda_1\big)\mathrm{d}s,
\end{aligned}$$

we can integrate the previous bounds and get:

$$\phi_1(t) \leq e^{-4\alpha_1(t)}\left(\phi_1(0) - 4\int_0^t \lambda_{\min}\big(Y_1(s)-\Lambda_1\big)e^{4\alpha_1(s)}\phi_2(s)\mathrm{d}s\right), \tag{349}$$

$$\phi_2(t) \leq e^{-2\alpha_2(t)}\phi_2(0), \tag{350}$$

$$\phi_3(t) \leq e^{-4\lambda_{\min}(\Lambda_2)t}\phi_3(0). \tag{351}$$

Since $Y_1(t)$ converges to Λ_1, we have the asymptotics:

$$\begin{aligned}
\alpha_1(t) &= \lambda_{\min}(\Lambda_1)t + o(t),\\
\alpha_2(t) &= \big(\lambda_{\min}(\Lambda_1) + \lambda_{\min}(\Lambda_2)\big)t + o(t).
\end{aligned}$$

For simplicity, let us now write $\rho_1 = \lambda_{\min}(\Lambda_1)$ and $\rho_2 = \lambda_{\min}(\Lambda_2)$. Then, from equations (350), (351), as $t \to \infty$:

$$\phi_2(t) \leq e^{-2(\rho_1+\rho_2)t}e^{o(t)}, \qquad \phi_3(t) \leq e^{-4\rho_2 t}e^{o(t)}. \tag{352}$$

In addition, the integrand in equation (349) verifies:

$$\Big|\lambda_{\min}\big(Y_1(t)-\Lambda_1\big)e^{4\alpha_1(t)}\phi_2(t)\Big| \le e^{2(\rho_1-\rho_2)t}e^{o(t)}.$$

Therefore, we have as $t\to\infty$:

$$\left|\int_0^t \lambda_{\min}\big(Y_1(s)-\Lambda_1\big)e^{4\alpha_1(s)}\phi_2(s)\mathrm{d}s\right| \le \begin{cases} e^{2(\rho_1-\rho_2)t}e^{o(t)}, & \text{if } \rho_1>\rho_2,\\ e^{o(t)}, & \text{otherwise.}\end{cases}$$

Plugging this into equation (349), we obtain:

$$\phi_1(t) \le e^{-2\rho_1 t}e^{-2\min(\rho_1,\rho_2)t}e^{o(t)}. \tag{353}$$

Combining the bounds (352) and (353), we obtain that the three functions $\phi_1(t),\phi_2(t)$ and $\phi_3(t)$ are all bounded by $e^{-4\min(\rho_1,\rho_2)t}e^{o(t)}$ as $t\to\infty$. To conclude the proof, we use the expressions:

$$\begin{aligned} U(W(t))-U(W_\infty) &= \frac{1}{4}\Big(\phi_1(t)+2\phi_2(t)+\phi_3(t)+2\mathrm{Tr}\big(Y_3(t)\Lambda_2\big)\Big),\\ \big\|Z(t)-Z_\infty\big\|_F^2 &= \frac{1}{4}\Big(\phi_1(t)+2\phi_2(t)+\phi_3(t)\Big).\end{aligned}$$

Using the Cauchy–Schwarz inequality guarantees that:

$$\mathrm{Tr}\big(Y_3(t)\Lambda_2\big) \le \|\Lambda_2\|_F\sqrt{\phi_3(t)} \le e^{-2\rho_2 t}e^{o(t)}.$$

Now introducing $c<\min(2\rho_1,\rho_2)$, we obtain that as $t\to\infty$:

$$U(W(t))-U(W_\infty)=o(e^{-2ct}), \qquad \big\|Z(t)-Z_\infty\big\|_F=o(e^{-ct}).$$

To conclude the proof, it is clear that we have $\rho_1=\lambda_p$ and $\rho_2=|\lambda_{p+1}|$, where $\lambda_1>\cdots>\lambda_d$ are the ordered eigenvalues of A.

I.5 Proof of Proposition 22

Before proving Proposition 22, we give two essential lemmas for our proof. The first lemma establishes high-dimensional limits for traces of functions of a covariance matrix composed with a Gaussian transformation.

Lemma 50 *Let $W\in\mathbb{R}^{d\times m}$ be a Gaussian matrix with i.i.d. coefficients of variance $1/m$, and $K\in\mathcal{S}_d^+(\mathbb{R})$ with converging empirical spectral distribution as $d\to\infty$ to some probability measure μ with compact support. Let $V=K^{1/2}W\in\mathbb{R}^{d\times m}$ and:*

$$G=V\big(I_m+V^\top V\big)^{-1}V^\top.$$

Then, given ϕ a spectral function on $\mathcal{S}_d(\mathbb{R})$, we have in the $d\to\infty$ limit, with $m\sim\kappa d$:

$$\begin{aligned} \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\phi(K)G\big) &\xrightarrow[d\to\infty]{} \mathfrak{g}\int\frac{x\phi(x)}{1+\mathfrak{g}x}\mathrm{d}\mu(x),\\ \frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\phi(K)G\phi(K)G\big) &\xrightarrow[d\to\infty]{} \mathfrak{g}^2\int\phi(x)^2\frac{x^2}{(1+x\mathfrak{g})^2}\mathrm{d}\mu(x)\\ &\quad+\mathfrak{g}\left(\int\frac{x\phi(x)}{1+x\mathfrak{g}}\mathrm{d}\mu(x)\right)^2\left(\kappa+\int\frac{x}{(1+x\mathfrak{g})^2}\mathrm{d}\mu(x)\right)^{-1},\end{aligned}$$

where $\mathfrak{g}$ *is the unique solution of the equation:*

$$\kappa\mathfrak{g} + 1 - \kappa = \int \frac{\mathrm{d}\mu(x)}{1 + x\mathfrak{g}}.$$

Proof In a basis where K is diagonal, we have:

$$\frac{1}{d}\mathrm{Tr}\big(\phi(K)G\big) = \frac{1}{d}\sum_{i=1}\phi(\lambda_i)G_{ii}, \tag{354}$$

$$\frac{1}{d}\mathrm{Tr}\big(\phi(K)G\phi(K)G\big) = \frac{1}{d}\sum_{i,j=1}^{d}\phi(\lambda_i)\phi(\lambda_j)G_{ij}^2. \tag{355}$$

We now compute the statistics of the random variables G_{ij}. Let us denote $w_1, \dots, w_d \in \mathbb{R}^m$ the rows of W, so that:

$$G_{ij} = \sqrt{\lambda_i\lambda_j}w_i^\top \left(I_m + \sum_{k=1}^{d}\lambda_k w_k w_k^\top\right)^{-1} w_j.$$

Let us start by taking $i = j$, and define:

$$M = \left(I_m + \sum_{k=1}^{d}\lambda_k w_k w_k^\top\right)^{-1}, \qquad M_i = \left(I_m + \sum_{k\neq i}\lambda_k w_k w_k^\top\right)^{-1}.$$

The key point is that, due to the covariance structure of the Gaussian matrix W, M_i is independent from w_i. Now, as a consequence of the Sherman–Morison formula (see for instance Hager, 1989), we have:

$$G_{ii} = \frac{\lambda_i w_i^\top M_i w_i}{1 + \lambda_i w_i^\top M_i w_i}.$$

Since M_i and w_i are independent, we have:

$$\begin{aligned}\mathbb{E}\, w_i^\top M_i w_i &= \frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M_i),\\ \mathrm{Var}\, w_i^\top M_i w_i &= \frac{2}{m^2}\mathbb{E}\,\mathrm{Tr}(M_i^2) + \frac{1}{m^2}\mathrm{Var}\,\mathrm{Tr}(M_i).\end{aligned}$$

Now, as $d \to \infty$, one can replace M_i by M in the previous equations since we only removed one index. Now, it is known that (see for instance Bun et al., 2017):

$$\frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M) \xrightarrow[d\to\infty]{} \mathfrak{g}, \qquad \lim_{m\to\infty}\frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M^2) < \infty,$$

where $\mathfrak{g}$ solves the self-consistent equation:

$$\kappa\mathfrak{g} + 1 - \kappa = \int \frac{\mathrm{d}\mu(x)}{1 + x\mathfrak{g}}. \tag{356}$$

Therefore, $w_i^\top M_i w_i$ concentrates as $d \to \infty$ toward $\mathfrak{g}$, and:

$$\mathbb{E}\, G_{ii} \xrightarrow[d\to\infty]{} \frac{\lambda_i \mathfrak{g}}{1+\lambda_i \mathfrak{g}}.$$

This shows that, using equation (354):

$$\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\phi(K)G\big) \xrightarrow[d\to\infty]{} \mathfrak{g}\int \frac{x\phi(x)}{1+\mathfrak{g}x}\mathrm{d}\mu(x).$$

For the off-diagonal case, the same can be done by removing the rows i, j. Then using again the Sherman–Morrison formula with the rank-two matrix update $\lambda_i w_i w_i^\top + \lambda_j w_j w_j^\top$, we get:

$$G_{ij} = \sqrt{\lambda_i \lambda_j}\frac{w_i^\top M_{ij} w_j}{(1+\lambda_i w_i^\top M_{ij} w_i)(1+\lambda_j w_j^\top M_{ij} w_j) - \lambda_i\lambda_j (w_i^\top M_{ij} w_j)^2},$$

with:

$$M_{ij} = \left(I_m + \sum_{k\neq i,j} \lambda_k w_k w_k^\top\right)^{-1}.$$

Again, $w_i^\top M_{ij} w_i$ and $w_j^\top M_{ij} w_j$ concentrate around $\mathfrak{g}$. Now, we have:

$$\begin{aligned}
\mathbb{E}\,(w_i^\top M_{ij} w_j)^2 &= \frac{1}{m^2}\mathbb{E}\,\mathrm{Tr}(M_{ij}^2),\\
\mathrm{Var}\,(w_i^\top M_{ij} w_j)^2 &= \frac{3}{m^4}\mathbb{E}\big[\mathrm{Tr}(M_{ij}^2)^2\big] - \frac{1}{m^4}\mathbb{E}\big[\mathrm{Tr}(M_{ij}^2)\big]^2 + \frac{6}{m^4}\mathbb{E}\,\mathrm{Tr}(M_{ij}^4).
\end{aligned}$$

This shows that:

$$d\,\mathbb{E}\, G_{ij}^2 \xrightarrow[d\to\infty]{} \frac{1}{\kappa}\frac{\lambda_i\lambda_j}{(1+\lambda_i\mathfrak{g})^2(1+\lambda_j\mathfrak{g})^2}\lim_{d\to\infty}\frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M^2). \tag{357}$$

However, note that G_{ij}^2 has a variance which is of the same order than its expectation. This means that this quantity does not concentrate in the high-dimensional limit.

Let us compute the expectation of $\mathrm{Tr}(M^2)$ in the limit. From Bun et al. (2017, Chapter 3), if we consider $M(z) = \big(zI_m - V^\top V\big)^{-1}$, then we have the asymptotic:

$$\frac{1}{m}\mathbb{E}\,\mathrm{Tr}\, M(z) \xrightarrow[d\to\infty]{} \mathfrak{g}(z), \qquad\qquad \kappa z\mathfrak{g}(z) + 1 - \kappa = \int \frac{\mathrm{d}\mu(x)}{1-x\mathfrak{g}(z)}.$$

Then:

$$\frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M(z)^2) \xrightarrow[d\to\infty]{} -\mathfrak{g}'(z),$$

and we can compute $\mathfrak{g}'(z)$ by differentiating the self-consistent equation for $\mathfrak{g}(z)$:

$$\mathfrak{g}'(z) = -\kappa\mathfrak{g}(z)\left(\kappa z - \int \frac{x\mathrm{d}\mu(x)}{(1-x\mathfrak{g}(z))^2}\right)^{-1}.$$

Then, taking $z = -1$, we obtain:

$$\lim_{d\to\infty}\frac{1}{m}\mathbb{E}\,\mathrm{Tr}(M^2) = \kappa\mathfrak{g}\left(\kappa + \int \frac{x\mathrm{d}\mu(x)}{(1+x\mathfrak{g})^2}\right)^{-1}, \tag{358}$$

where $\mathfrak{g}$ is the solution of the self-consistent equation (356). Then, back to equation (355), using that G_{ii} concentrates around its mean as well as equations (357), (358):

$$\begin{aligned}\frac{1}{d}\mathbb{E}\,\mathrm{Tr}\big(\phi(K)G\phi(K)G\big) &= \frac{1}{d}\sum_{i=1}^{d}\phi(\lambda_i)^2\mathbb{E}\,G_{ii}^2 + \frac{1}{d}\sum_{i\neq j}\phi(\lambda_i)\phi(\lambda_j)\mathbb{E}\,G_{ij}^2 \\ &\xrightarrow[d\to\infty]{} \mathfrak{g}^2\int \phi(x)^2\frac{x^2}{(1+x\mathfrak{g})^2}\mathrm{d}\mu(x) \\ &\quad + \mathfrak{g}\left(\int\frac{x\phi(x)}{(1+x\mathfrak{g})^2}\mathrm{d}\mu(x)\right)^2\left(\kappa+\int\frac{x\mathrm{d}\mu(x)}{(1+x\mathfrak{g})^2}\right)^{-1},\end{aligned}$$

which is the desired. ■

The previous lemma established the convergence of the expectation of our quantities of interest. The following guarantees concentration in the high-dimensional limit.

Lemma 51 *Consider the same setup as in Lemma 50. Then the two quantities:*

$$\frac{1}{d}\mathrm{Tr}\big(\phi(K)G\big), \qquad\qquad \frac{1}{d}\mathrm{Tr}\big(\phi(K)G\phi(K)G\big),$$

have vanishing variance as $d\to\infty$.

Proof We see both of these quantities as functions of W. Following from the Poincaré inequality, we have for a $\mathcal{C}^1$ function $F\colon\mathbb{R}^{d\times m}\to\mathbb{R}$:

$$\mathrm{Var}(F(W)) \le \frac{1}{m}\mathbb{E}\left\|\nabla F(W)\right\|_F^2.$$

Writing $G = K^{1/2}E(W)K^{1/2}$ with $E(W) = W(I_m + W^\top KW)^{-1}W^\top$ we have the identities:

$$\begin{aligned}\nabla_W\frac{1}{d}\mathrm{Tr}\big(\phi(K)G\big) &= \frac{1}{d}\mathrm{d}E_W^*\big(K\phi(K)\big),\\ \nabla_W\frac{1}{d}\mathrm{Tr}\big(\phi(K)G\phi(K)G\big) &= \frac{2}{d}\mathrm{d}E_W^*\big(K\phi(K)E(W)K\phi(K)\big),\end{aligned}$$

where $\mathrm{d}E_W^*$ is the adjoint of the differential of E at W. Now, with $M = I_m + W^\top KW$, one has for $U\in\mathcal{S}_d(\mathbb{R})$:

$$\mathrm{d}E_W^*(U) = 2UWM^{-1} - 2KWM^{-1}W^\top UWM^{-1}.$$

Now computing the squared norms of the gradients, one can write in both cases:

$$\mathrm{Var}\big(F(W)\big) \le \frac{1}{md^2}\mathbb{E}\,\mathrm{Tr}\Big(\mathcal{H}\big(K,\phi(K),E,\tilde{E}\big)\Big), \qquad\qquad \tilde{E} = WM^{-2}W^\top,$$

and $\mathcal{H}$ is a sum of products of its arguments. Since $E,\tilde{E}$ only depend on K and the Gaussian matrix W whose covariance scales as d^{-1}, in the high-dimensional limit, these traces are of order d as $d\to\infty$. This leads to a variance of order d^{-2}. ■

We are now ready to prove Proposition 22, that derives the high-dimensional limit associated with a scalar quantity of the Oja flow:

$$\frac{1}{d}\big\|Z(t) - \phi(A)\big\|_F^2,$$

where $Z(t)$ is solution of the Oja flow (80) with target matrix A, and ϕ is a spectral function on $\mathcal{S}_d(\mathbb{R})$.

We start by using the explicit solution of the Oja flow in Proposition 17. Due to the invertibility of A, we have:

$$Z(t) = e^{tA}W_0\Big(I_m + W_0^\top A^{-1}\big(e^{2tA} - I_d\big)W_0\Big)^{-1}W_0^\top e^{tA}.$$

Let us now define:

$$\Sigma(t) = A^{-1}\big(e^{2tA} - I_d\big) \succeq 0, \qquad\qquad R(t) = e^{tA}\Sigma(t)^{-1/2},$$

which are both spectral functions of A. Now, with $V(t) = \Sigma(t)^{1/2}W_0$, we have:

$$Z(t) = R(t)\underbrace{V(t)\Big(I_m + V(t)^\top V(t)\Big)^{-1}V(t)^\top}_{\equiv G(t)}R(t).$$

Then, given ϕ a spectral function of A:

$$\frac{1}{d}\big\|Z(t) - \phi(A)\big\|_F^2 = \frac{1}{d}\mathrm{Tr}\Big(R(t)^2G(t)R(t)^2G(t)\Big) - \frac{2}{d}\mathrm{Tr}\Big(\phi(A)R(t)^2G(t)\Big) + \frac{1}{d}\mathrm{Tr}\big(\phi(A)^2\big).$$

Since $\Sigma(t), R(t)$ are both spectral functions of A, we can apply Lemma 50 and Lemma 51 to get that the first two terms concentrate in the $d \to \infty$ limit, for fixed $t \geq 0$:

$$\begin{aligned}
\frac{1}{d}\mathrm{Tr}\Big(\phi(A)R(t)^2G(t)\Big) &\xrightarrow[d\to\infty]{} \mathfrak{g}(t)\int \phi(x)\frac{xe^{2tx}}{(e^{2xt}-1)\mathfrak{g}(t)+x}\mathrm{d}\mu_A(x),\\
\frac{1}{d}\mathrm{Tr}\Big(R(t)^2G(t)R(t)^2G(t)\Big) &\xrightarrow[d\to\infty]{} \mathfrak{g}(t)^2\int\left(\frac{xe^{2tx}}{(e^{2xt}-1)\mathfrak{g}(t)+x}\right)^2\mathrm{d}\mu_A(x)\\
&\quad + \mathfrak{g}(t)\left(\int\left(\frac{xe^{tx}}{(e^{2xt}-1)\mathfrak{g}(t)+x}\right)^2\mathrm{d}\mu_A(x)\right)^2\\
&\quad\times\left(\kappa + \int\frac{x(e^{2xt}-1)}{((e^{2xt}-1)\mathfrak{g}(t)+x)^2}\mathrm{d}\mu_A(x)\right)^{-1},
\end{aligned}$$

where we used that the eigenvalues of $\Sigma(t)$ are of the form:

$$\frac{e^{2\lambda t}-1}{\lambda},$$

for $\lambda \in \mathrm{Sp}(A)$. Moreover, $\mathfrak{g}(t)$ is solution of the self-consistent equation:

$$\kappa\mathfrak{g}(t) + 1 - \kappa = \int\frac{x}{(e^{2xt}-1)\mathfrak{g}(t)+x}\mathrm{d}\mu_A(x). \tag{359}$$

Finally, since we have:

$$\frac{1}{d}\mathrm{Tr}(\phi(A)^2) \xrightarrow[d\to\infty]{} \int \phi(x)^2 \mathrm{d}\mu_A(x),$$

we get the desired by rearranging the terms.

I.6 Proof of Proposition 23

Let us now show Proposition 23, that gives access to the high-dimensional convergence rates for the Oja flow. To do so, we apply the result of Proposition 22 to the spectral function:

$$\phi(x) = x\mathbf{1}_{x\geq \max(0,\omega)}, \qquad \kappa = \int \mathbf{1}_{x\geq\omega}\mathrm{d}\mu_A(x).$$

The following lemma describes the behavior of the function $\mathfrak{g}(t)$, solution of the self-consistent equation (359).

Lemma 52 *Consider $\mathfrak{g}(t)$ to be the unique solution of equation (359). Define:*

$$\kappa_A = \int \mathbf{1}_{x>0}\mathrm{d}\mu_A(x).$$

- *If $\kappa > \kappa_A$, $\mathfrak{g}(t) \xrightarrow[t\to\infty]{} \mathfrak{g}_\infty \in (0,1)$, such that:*

$$\kappa\mathfrak{g}_\infty + 1 - \kappa = \int_{x<0} \frac{x}{x-\mathfrak{g}_\infty}\mathrm{d}\mu_A(x). \tag{360}$$

- *If $\kappa < \kappa_A$, $\mathfrak{g}(t) \underset{t\to\infty}{\sim} \omega e^{-2\omega t}$, where $\omega > 0$ is such that:*

$$\kappa = \int \mathbf{1}_{x\geq\omega}\mathrm{d}\mu_A(x). \tag{361}$$

Proof For the first case, we simply start to assume that $\mathfrak{g}(t)$ converges to a non-zero value, and take the pointwise limit in the integral of equation (359). We directly obtain equation (360).

Now, the positivity of $\mathfrak{g}_\infty$ imposes the inequalities:

$$1-\kappa < \kappa\mathfrak{g}_\infty + 1 - \kappa = \int_{x<0} \frac{x}{x-\mathfrak{g}_\infty}\mathrm{d}\mu_A(x) < 1-\kappa_A.$$

Therefore this asymptotic behavior is only suitable for $\kappa > \kappa_A$.

Now considering the case $\kappa < \kappa_A$, we assume an asymptotic of the form $\mathfrak{g}(t) = h(t)e^{-2\omega t}$ for some $\omega > 0$ and a sub-exponential function h, that is:

$$\frac{1}{t}\log h(t) \xrightarrow[t\to\infty]{} 0.$$

Then, going back to the self-consistent equation (359), we get:

$$1-\kappa = \int^{\omega} \mathrm{d}\mu_A(x).$$

We then obtain equation (361). Since $\kappa < \kappa_A$, we have that $\omega > 0$. Let us now compute the asymptotic of $h(t)$. We have, replacing the expression of $\mathfrak{g}(t)$ in equation (359) and using equation (361):

$$\kappa h(t) e^{-2\omega t} = \int \left(\frac{x}{(e^{2xt}-1)h(t)e^{-2\omega t} + x} - \mathbf{1}_{x<\omega} \right) \mathrm{d}\mu_A(x).$$

Changing variables $x = \omega + u/t$ and getting rid of the subleading terms, we get:

$$\kappa h(t) e^{-2wt} \underset{t\to\infty}{\sim} \frac{1}{t}\rho_A(\omega) \int_{-\infty}^{+\infty} \left(\frac{\omega}{e^{2u}h(t) + \omega} - \mathbf{1}_{u<0} \right) \mathrm{d}u.$$

Now, since $h(t)$ is sub-exponential, the integral converges to zero so that the LHS can indeed decay exponentially fast. Now, one can rearrange this integral to get:

$$0 = \lim_{t\to\infty} \int_{-\infty}^{+\infty} \left(\frac{\omega}{e^{2u}h(t) + \omega} - \mathbf{1}_{u<0} \right) \mathrm{d}u = \lim_{t\to\infty} \left[\theta\left(\frac{\omega}{h(t)} \right) - \theta\left(\frac{h(t)}{\omega} \right) \right],$$

with:

$$\theta(x) = \int_0^{+\infty} \frac{x}{e^{2u} + x} \mathrm{d}u = \frac{1}{2}\log(x+1).$$

Therefore, necessarily $\lim_{t\to\infty} h(t) = \omega$. This concludes the proof. ∎

Let us now prove the asymptotics given in Proposition 23. We start by the case $\kappa > \kappa_A$, in which case $\phi(x) = x\mathbf{1}_{x>0}$, since the Oja flow selects all the positive eigenvalues of A. We start with the first term of equation (91). We have, setting $u = tx$:

$$\begin{aligned}\int \left(\frac{xe^{xt}}{q_t(x)} \right)^2 \mathrm{d}\mu_A(x) &= \frac{1}{t^3} \int \rho_A\left(\frac{u}{t}\right) u^2 e^{2u} \left(\left(e^{2u}-1\right)\mathfrak{g}(t) + \frac{u}{t} \right)^{-2} \mathrm{d}u \\ &\underset{t\to\infty}{\sim} \frac{\rho_A(0)}{\mathfrak{g}_\infty^2 t^3} \int_{-\infty}^{+\infty} \frac{u^2 e^{2u}}{(e^{2u}-1)^2} \mathrm{d}u.\end{aligned}$$

Now, since the integral with respect to z in equation (91) remains positive, we obtain for the first term:

$$\mathfrak{g}(t) \left(\int \left(\frac{xe^{xt}}{q_t(x)} \right)^2 \mathrm{d}\mu_A(x) \right)^2 \left(\kappa + \int \frac{z(e^{2zt}-1)}{q_t(z)^2} \mathrm{d}\mu_A(z) \right)^{-1} = O\left(\frac{1}{t^6} \right).$$

Now, for the second term, we perform the same change of variables:

$$\begin{aligned}\int \left(\mathfrak{g}(t) \frac{xe^{2xt}}{q_t(x)} - \phi(x) \right)^2 \mathrm{d}\mu_A(x) &= \frac{1}{t^3} \int \rho_A\left(\frac{u}{t}\right) u^2 \left(\frac{\mathfrak{g}(t)e^{2u}}{\left(e^{2u}-1\right)\mathfrak{g}(t) + \frac{u}{t}} - \mathbf{1}_{u>0} \right)^2 \mathrm{d}u \\ &\underset{t\to\infty}{\sim} \frac{\rho_A(0)}{t^3} \int_{-\infty}^{+\infty} u^2 \left(\frac{e^{2u}}{e^{2u}-1} - \mathbf{1}_{u>0} \right)^2 \mathrm{d}u.\end{aligned}$$

This gives the asymptotic for $\kappa > \kappa_A$ since the integral is finite.

Regarding the underparameterized region $\kappa < \kappa_A$, we have $\phi(x) = x\mathbf{1}_{x\geq\omega}$, and we split the expression (91) as:

$$\lim_{d\to\infty} \frac{1}{d}\big\|Z(t) - Z_\infty\big\|_F^2 = \mathfrak{g}(t)\frac{I_1(t)^2}{\kappa + I_2(t)} + I_3(t), \tag{362}$$

with:

$$\begin{aligned}
I_1(t) &= \int \frac{x^2 e^{2xt}}{((e^{2xt}-1)\mathfrak{g}(t)+x)^2}\mathrm{d}\mu_A(x),\\
I_2(t) &= \int \frac{x(e^{2xt}-1)}{((e^{2xt}-1)\mathfrak{g}(t)+x)^2}\mathrm{d}\mu_A(x),\\
I_3(t) &= \int \left(\mathfrak{g}(t)\frac{xe^{2xt}}{(e^{2xt}-1)\mathfrak{g}(t)+x} - x\mathbf{1}_{x\geq\omega}\right)^2 \mathrm{d}\mu_A(x).
\end{aligned}$$

Now using the asymptotic of $\mathfrak{g}(t)$ and changing variables $x = \omega + u/t$, we get the asymptotics:

$$\begin{aligned}
I_1(t) &\underset{t\to\infty}{\sim} \frac{e^{2\omega t}}{t}\rho_A(\omega)\int_{-\infty}^{+\infty}\frac{e^{2u}}{(e^{2u}+1)^2}\mathrm{d}u,\\
I_2(t) &\underset{t\to\infty}{\sim} \frac{e^{2\omega t}}{t}\frac{\rho_A(\omega)}{\omega}\int_{-\infty}^{+\infty}\frac{e^{2u}}{(e^{2u}+1)^2}\mathrm{d}u,\\
I_3(t) &\underset{t\to\infty}{\sim} \frac{1}{t}\rho_A(\omega)\omega^2\int_{-\infty}^{+\infty}\left(\frac{e^{2u}}{e^{2u}+1} - \mathbf{1}_{u\geq 0}\right)^2\mathrm{d}u.
\end{aligned}$$

Therefore, combining these asymptotics with equation (362):

$$\lim_{d\to\infty}\frac{1}{d}\big\|Z(t)-Z_\infty\big\|_F^2 \underset{t\to\infty}{\sim} \frac{\omega^2\rho_A(\omega)}{t}\left(\int_{-\infty}^{+\infty}\frac{e^{2u}}{(e^{2u}+1)^2}\mathrm{d}u + \int_{-\infty}^{+\infty}\left(\frac{e^{2u}}{e^{2u}+1} - \mathbf{1}_{u\geq 0}\right)^2\mathrm{d}u\right).$$

The result follows from the finiteness of the integrals.

Since we plot in Figure 14 the convergence rates obtained in Proposition 22 with the exact constants, we give the values of the above integrals:

$$\begin{aligned}
\int_{-\infty}^{+\infty} u^2\left(\frac{e^{2u}}{e^{2u}-1} - \mathbf{1}_{u>0}\right)^2\mathrm{d}u &= \frac{\pi^2}{12} - \frac{1}{2}\zeta(3),\\
\int_{-\infty}^{+\infty}\frac{e^{2u}}{(e^{2u}+1)^2}\mathrm{d}u &= \frac{1}{2},\\
\int_{-\infty}^{+\infty}\left(\frac{e^{2u}}{e^{2u}+1} - \mathbf{1}_{u\geq 0}\right)^2\mathrm{d}u &= \log 2 - \frac{1}{2}.
\end{aligned}$$

I.7 Proof of Proposition 24

In this section we compute the linear response associated with the perturbed Oja flow dynamics:

$$\dot{W}(t) = \Big(A - W(t)W(t)^\top\Big)W(t) + H(t)W(t),$$

where $A \in \mathcal{S}_d(\mathbb{R})$ and $H(t) \in \mathcal{S}_d(\mathbb{R})$ is the perturbation. We are interested in computing:

$$R(t,t') = \left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0}, \qquad Z(t) = W(t)W(t)^\top.$$

$R(t,t')$ is an operator $\mathcal{S}_d(\mathbb{R}) \to \mathcal{S}_d(\mathbb{R})$ quantifying the change of $Z(t)$ in response to the introduction of H at time t'. To compute it, we replace H by ϵH and decompose the solution $Z(t) = Z_0(t) + \epsilon Y(t)$ at first order in ϵ. We then get the equations:

$$\dot{Z}_0(t) = AZ_0(t) + Z_0(t)A - 2Z_0(t)^2, \tag{363}$$

$$\dot{Y}(t) = Y(t)(A - 2Z_0(t)) + (A - 2Z_0(t))Y(t) + H(t)Z_0(t) + Z_0(t)H(t), \tag{364}$$

with $Y(0) = 0$. Remark that the dynamics on $Y(t)$ is linear, but driven by a time-dependent matrix. This type of equation cannot be solved in general, but remarkably it is possible in our case. First of all, the dynamics (363) is an Oja flow, and following Proposition 17, we have the expression for $Z_0(t)$:

$$Z_0(t) = e^{tA}W_0\left(I_m + 2W_0^\top \int_0^t e^{2sA}\mathrm{d}s\, W_0\right)^{-1} W_0^\top e^{tA}, \tag{365}$$

where $W_0W_0^\top = Z_0(t=0)$. To solve the equation on $Y(t)$, we set $B(t) = A - 2Z_0(t)$. Then, it is easily seen that $B(t)$ is solution of the dynamics:

$$\dot{B}(t) = B(t)^2 - 4A^2. \tag{366}$$

Now, from equation (364) on $Y(t)$ and since $Y(0) = 0$, we have the solution:

$$Y(t) = P(t)^{-1}\int_0^t P(s)\Big(H(s)Z_0(s) + Z_0(s)H(s)\Big)P(s)^\top\,\mathrm{d}s\,P(t)^{-\top}, \tag{367}$$

where $P(t)$ is such that $\dot{P}(t) = -P(t)B(t)$ with $P(0) = I_d$. Differentiating this equation together with the differential equation (366) on $B(t)$, we obtain:

$$\ddot{P}(t) = 4P(t)A^2.$$

We can integrate this equation with the initial conditions $P(0) = I_d$ and $\dot{P}(0) = 4Z_0 - 2A$. We obtain:

$$P(t) = e^{-tA} + Z_0A^{-1}\big(e^{tA} - e^{-tA}\big). \tag{368}$$

This expression remains well defined when A is not invertible if $s_t(\lambda) = \lambda^{-1}(e^{t\lambda} - e^{-t\lambda})$ is continuously extended to $\lambda = 0$ by setting $s_t(0) = 2t$. Now, in order to compute $R(t,t')$, we differentiate Y with respect to H using equation (367). We need to take into account that the derivative is taken on the space $\mathcal{S}_d(\mathbb{R})$. To do so, we define, for $1 \le i \le j \le d$:

$$E_{ij} = \frac{e_ie_j^\top + e_je_i^\top}{2}, \tag{369}$$

where $(e_1, \dots, e_d)$ is the standard basis of $\mathbb{R}^d$. Then, setting:

$$U(t,t') = P(t)^{-1}P(t'), \qquad V(t,t') = P(t)^{-1}P(t')Z_0(t'),$$

and evaluating equation (367) along E_{kl}, we get for $t' \leq t$:

$$\begin{aligned} R_{ijkl}(t,t') &\equiv \mathrm{Tr}\left(\left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0} (E_{kl}) E_{ij} \right) \\ &= \frac{1}{2}\Big(U_{ik}(t,t')V_{jl}(t,t') + U_{il}(t,t')V_{jk}(t,t') \\ &\qquad + U_{jk}(t,t')V_{il}(t,t') + U_{jl}(t,t')V_{ik}(t,t') \Big). \end{aligned}$$

To conclude, combining equations (365) and (368), we get:

$$P(t)Z_0(t) = W_0 W_0^\top e^{tA},$$

and the result of Proposition 24 follows.

I.8 Proof of Proposition 25

To prove Proposition 25, we compute the functions:

$$R_{\mathrm{diag}}(t,t') = \lim_{d\to\infty} \frac{1}{d^2}\mathrm{Tr}\left(\left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0} \right), \qquad r_{\mathrm{diag}}(t) = \int_0^t R_{\mathrm{diag}}(t,t')\mathrm{d}t'.$$

To compute the trace, we use the orthogonal basis of $\mathcal{S}_d(\mathbb{R})$ defined in equation (369). It is easily seen that $\|E_{ij}\|_F^2 = (1+\delta_{ij})/2$. Then, we have, with the notations of Proposition 24:

$$\begin{aligned} \frac{1}{d^2}\mathrm{Tr}\left(\left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0} \right) &= \sum_{1\leq i\leq j\leq d} \frac{1}{\|E_{ij}\|_F^2}\mathrm{Tr}\left(\left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0} (E_{ij}) E_{ij} \right) \\ &= \frac{1}{d^2}\sum_{i,j=1}^d \mathrm{Tr}\left(\left.\frac{\partial Z(t)}{\partial H(t')}\right|_{H=0} (E_{ij}) E_{ij} \right) \\ &= \frac{1}{d^2}\mathrm{Tr}\Big(U(t,t')\Big)\mathrm{Tr}\Big(V(t,t')\Big) + \frac{1}{d^2}\mathrm{Tr}\Big(V(t,t')U(t,t')^\top\Big). \end{aligned} \tag{370}$$

Now, $U(t,t'), V(t,t')$ are d-independent functions of the matrices Z_0, A. Since the scalings we chose guarantee that both Z_0 (as a Wishart matrix) and A have convergent empirical spectral distributions in the high-dimension limit, all traces of scalar (and independent of d) functions of these matrices should be of order d. This implies that, as $d \to \infty$:

$$\frac{1}{d^2}\mathrm{Tr}\Big(V(t,t')U(t,t')^\top\Big) = O\left(\frac{1}{d}\right). \tag{371}$$

Therefore, only the first term in equation (370) contributes in the high-dimensional limit. Now using the expressions of $U(t,t'), V(t,t')$, we have, denoting by $\lambda_1 \geq \cdots \geq \lambda_d$ the eigenvalues of A, and working in a basis where A is diagonal:

$$\begin{aligned} \frac{1}{d}\mathrm{Tr}\big(U(t,t')\big) &= \frac{1}{d}\sum_{i=1}^d \big(P(t)^{-1}Z_0\big)_{ii}\frac{e^{\lambda_i t'} - e^{-\lambda_i t'}}{\lambda_i} + \frac{1}{d}\sum_{i=1}^d \big(P(t)^{-1}\big)_{ii} e^{-\lambda_i t'}, \\ \frac{1}{d}\mathrm{Tr}\big(V(t,t')\big) &= \frac{1}{d}\sum_{i=1}^d \big(P(t)^{-1}Z_0\big)_{ii} e^{\lambda_i t'}. \end{aligned} \tag{372}$$

We now compute the limit of these traces as $d \to \infty$ by replacing the averages by integrals with respect to μ_A. The key point is to deal with the coefficients $\big(P(t)^{-1}Z_0\big)_{ii}$, $\big(P(t)^{-1}\big)_{ii}$, which need to be expressed as functions of the eigenvalues of A. To do so, we define:

$$\Sigma(t) = A^{-1}\big(e^{2tA} - I_d\big) \succeq 0, \qquad V(t) = \Sigma(t)^{1/2}W_0.$$

Then, we can express:

$$\begin{aligned} P(t)^{-1} &= e^{tA}\Big(I_d + W_0W_0^\top\Sigma(t)\Big)^{-1} \\ &= e^{tA}\Sigma(t)^{-1/2}\Big(I_d + V(t)V(t)^\top\Big)^{-1}\Sigma(t)^{1/2}. \end{aligned}$$

Therefore, since we are working in a basis where A (and therefore $\Sigma(t)$) is diagonal, we have the expressions:

$$\big(P(t)^{-1}Z_0\big)_{ii} = \frac{\lambda_i e^{\lambda_i t}}{e^{2\lambda_i t}-1} e_i^\top \Big(I_d + V(t)V(t)^\top\Big)^{-1} V(t)V(t)^\top e_i, \tag{373}$$

$$\big(P(t)^{-1}\big)_{ii} = e^{\lambda_i t} e_i^\top \Big(I_d + V(t)V(t)^\top\Big)^{-1} e_i. \tag{374}$$

Now, since:

$$\big(I_d + V(t)V(t)^\top\big)^{-1}V(t)V(t)^\top = I_d - \big(I_d + V(t)V(t)^\top\big)^{-1},$$

we can rewrite:

$$\big(P(t)^{-1}Z_0\big)_{ii} = \frac{\lambda_i e^{\lambda_i t}}{e^{2\lambda_i t}-1}\Big(1 - e_i^\top\big(I_d + V(t)V(t)^\top\big)^{-1}e_i\Big). \tag{375}$$

Therefore we only need to compute $e_i^\top\big(I_d + V(t)V(t)^\top\big)^{-1}e_i$. We now invoke the result of Bun et al. (2017, Section 4), that yields in our case:

$$e_i^\top\big(I_d + V(t)V(t)^\top\big)^{-1}e_i \longrightarrow \frac{1}{1+\mu_i(t)\mathfrak{g}(t)}, \qquad \mu_i(t) = \frac{e^{2\lambda_i t}-1}{\lambda_i}, \tag{376}$$

where:

$$\mathfrak{g}(t) = \lim_{d\to\infty}\frac{1}{m}\mathrm{Tr}\,\Big(I_m + V(t)^\top V(t)\Big)^{-1},$$

that solves the self-consistent equation:

$$\kappa\mathfrak{g}(t) + 1 - \kappa = \int \frac{x}{(e^{2xt}-1)\mathfrak{g}(t) + x}\mathrm{d}\mu_A(x).$$

Therefore, in the high-dimensional limit, combining equations (374), (375), we have the expression:

$$\begin{aligned} \big(P(t)^{-1}Z_0\big)_{ii} &\xrightarrow[d\to\infty]{} \mathfrak{g}(t)\frac{\lambda_i e^{\lambda_i t}}{\mathfrak{g}(t)(e^{2\lambda_i t}-1)+\lambda_i}, \\ \big(P(t)^{-1}\big)_{ii} &\xrightarrow[d\to\infty]{} \frac{\lambda_i e^{\lambda_i t}}{\mathfrak{g}(t)(e^{2\lambda_i t}-1)+\lambda_i}. \end{aligned} \tag{377}$$

Finally, using equations (370), (371), (372), (377), we get the expression of R_{diag}:

$$R_{\text{diag}}(t,t') = \mathfrak{g}(t) \iint \frac{1}{q_t(x)q_t(y)} y e^{y(t+t')} \Big(\big(x - \mathfrak{g}(t)\big) e^{x(t-t')} + \mathfrak{g}(t) e^{x(t+t')} \Big) \mathrm{d}\mu_A(x) \mathrm{d}\mu_A(y),$$

with $q_t(x) = \mathfrak{g}(t)(e^{2xt} - 1) + x$. Symmetrizing with respect to x, y we get:

$$R_{\text{diag}}(t,t') = \frac{\mathfrak{g}(t)}{2} \iint \frac{1}{q_t(x)q_t(y)} e^{(x+y)t} \Big(y(x - \mathfrak{g}(t)) e^{(y-x)t'} + x(y - \mathfrak{g}(t)) e^{(x-y)t'} + (x+y)\mathfrak{g}(t) e^{(x+y)t'} \Big) \mathrm{d}\mu_A(x) \mathrm{d}\mu_A(y).$$

This is precisely equation (101) in Proposition 25. Let us now compute r_{diag}:

$$\begin{aligned} r_{\text{diag}}(t) &= \int_0^t R_{\text{diag}}(t,t') \mathrm{d}t' \\ &= \frac{\mathfrak{g}(t)}{2} \iint \frac{1}{q_t(x)q_t(y)} e^{(x+y)t} \left[\frac{y(x - \mathfrak{g}(t))}{y - x} \Big(e^{(y-x)t} - 1 \Big) + \frac{x(y - \mathfrak{g}(t))}{x - y} \Big(e^{(x-y)t} - 1 \Big) + \mathfrak{g}(t) \Big(e^{(x+y)t} - 1 \Big) \right] \mathrm{d}\mu_A(x) \mathrm{d}\mu_A(y). \end{aligned}$$

First considering the terms with no exponential in the bracket, we have:

$$\frac{y(x - \mathfrak{g}(t))}{y - x} + \frac{x(y - \mathfrak{g}(t))}{x - y} + \mathfrak{g}(t) = 0.$$

Now splitting the last term into two, we can write the bracket as:

$$\begin{aligned} &\frac{y}{y - x} e^{yt} \Big(e^{xt} + (x - \mathfrak{g}(t)) e^{-xt} \Big) + \frac{x}{x - y} e^{xt} \Big(e^{yt} + (y - \mathfrak{g}(t)) e^{-yt} \Big) \\ &= \frac{y}{y - x} e^{(y-x)t} q_t(x) + \frac{x}{x - y} e^{(x-y)t} q_t(y). \end{aligned}$$

Putting everything together, this leads to equation (102) and concludes the proof.

Appendix J. Numerical Simulations

In this section, we give more details on our numerical simulations and provide additional figures that complement the results presented in the main text.

J.1 Details on the Numerics

We now detail our simulation setup, including the gradient descent algorithm and the numerical integration of the system of equations (37).

J.1.1 Gradient Descent Simulations

We refer to Section 2.3 for the details of the gradient descent simulations performed in this paper. All numerical simulations were implemented in `PyTorch` and executed on NVIDIA

RTX6000, RTX8000, and H100 GPUs provided by the CLEPS infrastructure at INRIA Paris.

The dimension d, stepsize η and total number of gradient descent steps were selected after preliminary experiments. We found that the chosen dimension captures the high-dimensional regime considered in our theoretical analysis, while the stepsize is sufficiently small for discrete gradient descent to approximate the corresponding gradient flow dynamics. Simulations were run for a time horizon long enough to observe convergence of the gradient descent trajectories.

Unless otherwise specified, all numerical results are averaged over several independent realizations of the random initialization, teacher matrix, and sensing matrices. Error bars in the figures represent two standard deviations across these realizations.

J.1.2 Simulating the System of Equations

Throughout this work, we compared gradient descent simulations with the numerical integration of the system of equations (37). We now explain how it is possible to simulate this set of equations.

In this system of equations, the unknowns are the variables ξ, q, ω, and the parameters are $\alpha, \kappa, \kappa^*, \lambda, \Delta$. To numerically integrate the equations, the key point is to consider ξ itself as a parameter. Therefore ω can be directly computed by solving:

$$\min(\kappa, 1) = \int_\omega \mathrm{d}\mu_\xi(x).$$

Then, we rather consider α as an unknown, and rewrite equation (37c):

$$\alpha = \frac{1}{2\xi}\left(\frac{\Delta}{2} + Q_* + \int_{\max(q,\omega)} \left(q^2 - x^2\right)\mathrm{d}\mu_\xi(x) + 4\xi \int_{\max(q,\omega)} (x-q) h_\xi(x)\mathrm{d}\mu_\xi(x)\right).$$

Then, this expression can be plugged in equation (37b), and leads to an equation solely on the variable q. This equation can then be solved using a numerical root solver (we used `scipy.brentq`). Then, the previous equation allows to deduce the value of α.

Still, the equation on q involves integrals (of simple functions) with respect to the measure μ_ξ, which is the free additive convolution between the teacher's spectral distribution and a semicircular density with variance ξ. For simplicity, we chose the teacher distribution to be the Marchenko–Pastur distribution with parameter κ^*. In this setting, the density of μ_ξ can be computed by solving a third-degree polynomial equation. We refer to Maillard et al. (2024, Appendix H.1) for details on this result. Then, the integration with respect to μ_ξ can be computed efficiently using `scipy.quad`.

Figure 16 displays q and the MSE as a function of ξ in the noisy setting $\Delta > 0$, for several choices of the regularization strength λ. We have observed that for most values of the parameters $\kappa, \kappa^*, \lambda, \Delta$, there was a unique solution q for a given ξ. However, for small λ and large Δ, as underlined by Figure 16, several solutions q may exist simultaneously. This makes it more difficult to integrate numerically the system of equations. To circumvent this problem, we used a technique known as arc-length continuation, that allows to compute the curve $q(\xi)$ in the 2D space and therefore capture the whole solution. Finally, we remark that this degeneracy in the solutions reveals the double descent phenomenon in our equations:

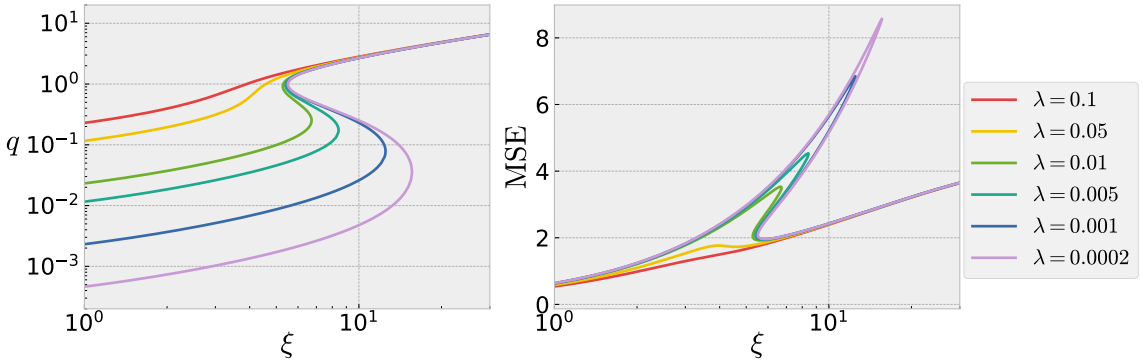

Figure 16: Left: value of q as a function of ξ when simulating the system of equations (37). Right: corresponding value of the MSE, for different values of λ and with $\kappa = 0.4, \kappa^* = 0.3$ and $\Delta = 1.0$. As λ decreases, a single value of ξ may lead to several solutions for q.

the non-monotonicity of the MSE as a function of α is linked to the fact that a single value of ξ may allow several solutions q.

J.2 Additional Experiments and Figures

In this section we present additional figures to support claims made in the main text.

J.2.1 Learning Curves

Claim 6 provides expressions for the MSE and the training loss $\text{Loss}_{\text{train}}$ as a function of the parameters $\alpha, \kappa, \kappa^*, \lambda, \Delta$. These expressions are obtained from the high-dimensional equations of Claim 4 using non-rigorous arguments. The figures below provide more numerical evidence supporting the validity of our results. In Figures 17, 18, 19, 20, we plot the MSE, loss, and in-sample error (the latter two coincide in the noiseless case $\Delta = 0$) as a function of α for a wide range of parameter values, and compare gradient descent simulations with the numerical integration of the system of equations of Claim 6. In addition, Figures 17, 20 feature results for values of κ close to $\kappa^* = 0.3$. As discussed in Section 3.2.4, this corresponds to a region where the steady-state solution may be unstable. Together with the numerical evidence of Figure 5, these figures support our theoretical predictions in this regime.

Moreover, Figure 17 highlights the dependence of the MSE and the training loss on κ. Increasing κ leads to a lower empirical loss, reflecting the increased representational capacity of the model. However, this improvement leads to a higher MSE and a poorer generalization. This effect becomes more pronounced as the regularization strength λ decreases.

Figures 18, 19 display the same learning curves for several values of κ, κ^*, respectively in the noiseless ($\Delta = 0$) and noisy ($\Delta = 0.5$) settings. These figures not only confirm the strong agreement between our theory and gradient descent simulations, but also highlight the role of regularization on the performance of gradient flow. In the noiseless case (Figure 18), reducing the regularization leads to simultaneous improvements in both training

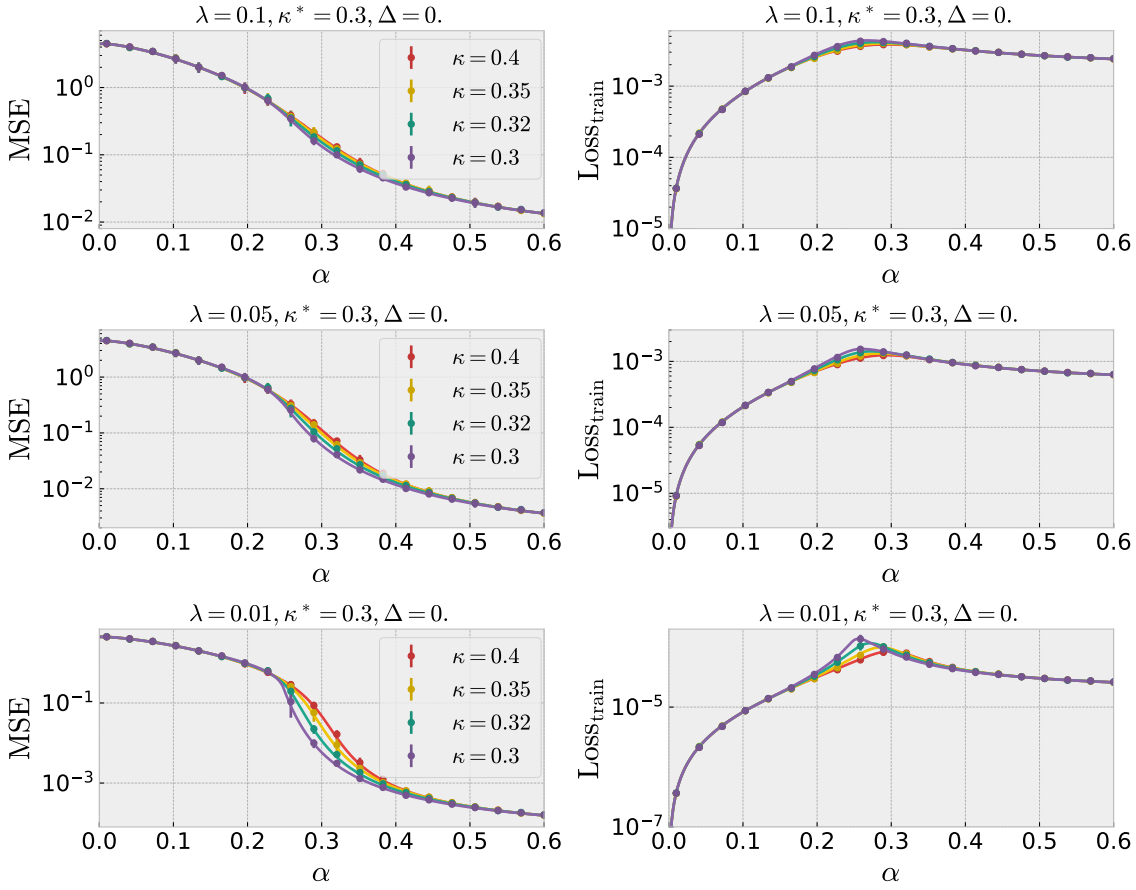

Figure 17: Comparison between simulations of gradient descent, defined in (6), and numerical integration of the system of equations (37), for $\kappa^* = 0.3, \Delta = 0$ and several values of the width κ and regularization strength λ. MSE (left) and training loss (right) as a function of the sample complexity α. Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data.

and generalization performance. This monotonic behavior holds in the range of regularization values considered here ($0.05 \leq \lambda \leq 0.5$) and suggests that in this regime, regularization mainly limits generalization. In the noisy setting (Figure 19), the dependence on the regularization strength is less clear, suggesting the existence of an optimal level of regularization. We leave this question for future work.

Finally, Figure 20 shows the MSE and in-sample error (see equation 73) in the double descent regime (introduced in Section 3.2.5), corresponding to large values of Δ and small values of λ. In contrast with the previous cases, decreasing the regularization strength leads to larger values of both the MSE and the in-sample error. This indicates that under weak regularization, the gradient flow predictor fits the noise present in the training labels. Beyond the interpolation threshold, the in-sample error and MSE starts to decrease again, suggesting that the structure of the teacher is progressively learned.

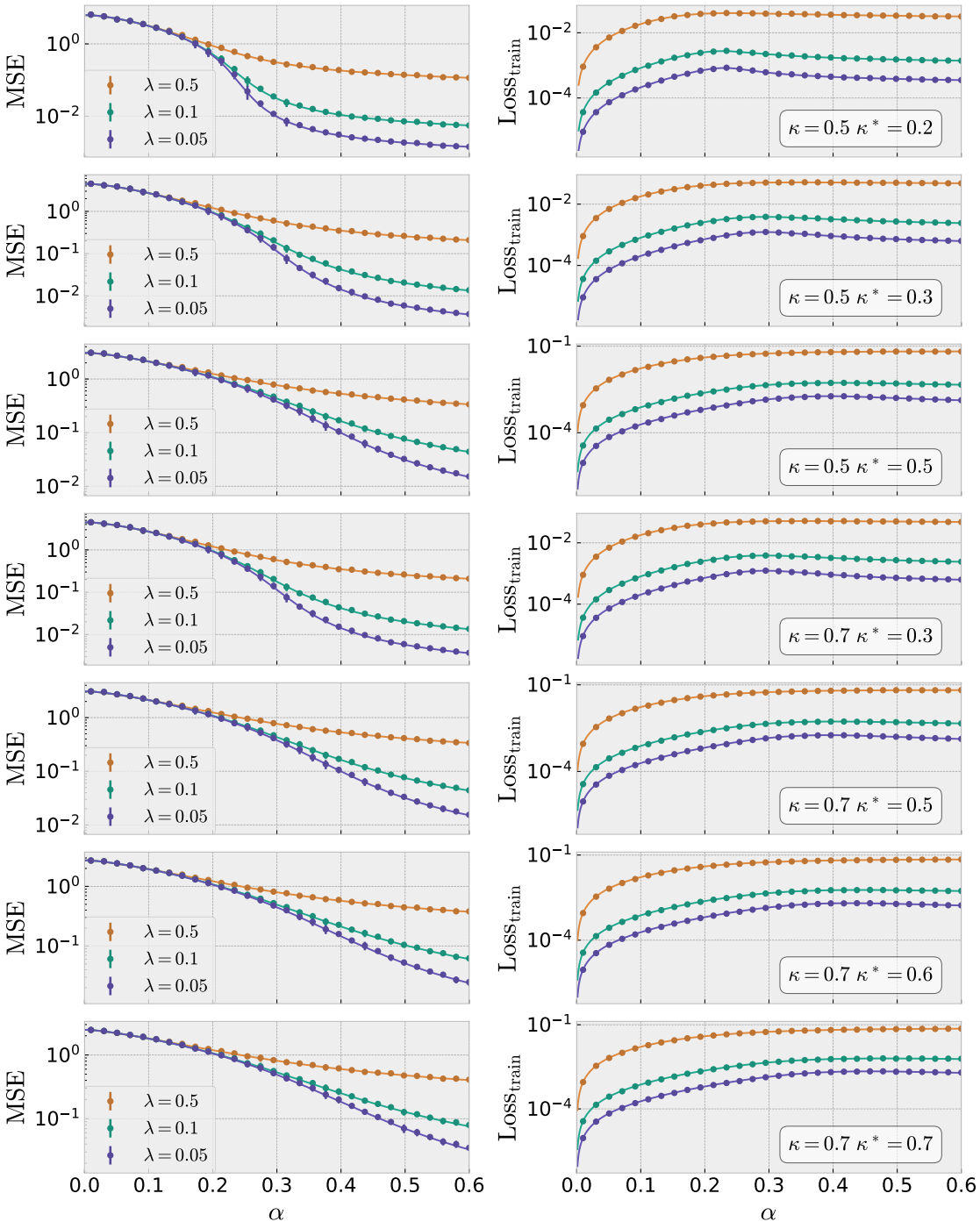

Figure 18: Comparison between simulations of gradient descent, defined in (6), and numerical integration of the system of equations (37), for $\Delta = 0$ and several values of the widths κ, κ^* and regularization strength λ. MSE (left) and training loss (right) as a function of the sample complexity α. Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data.

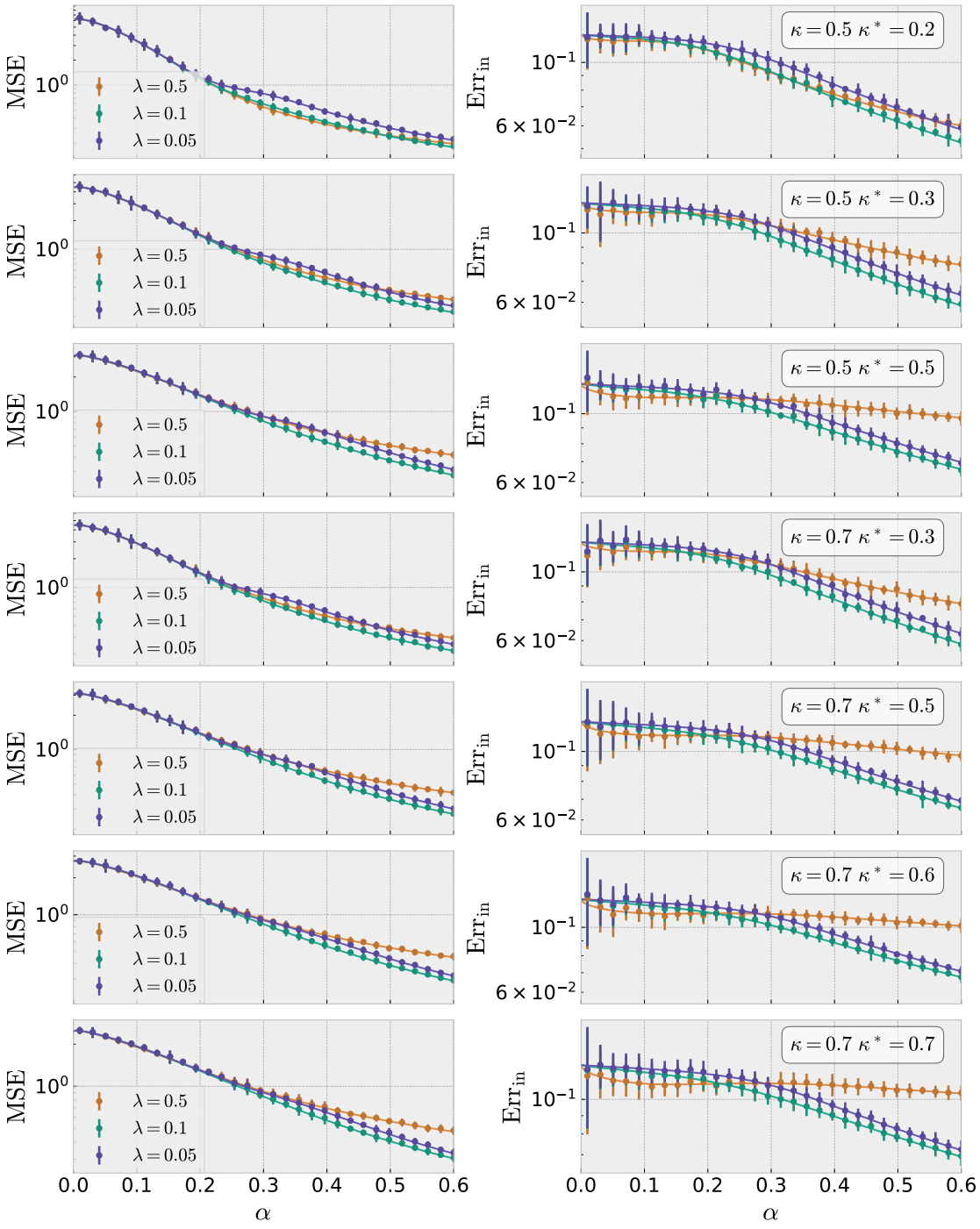

Figure 19: Comparison between simulations of gradient descent, defined in (6), and numerical integration of the system of equations (37), for $\Delta = 0.5$ and several values of the widths κ, κ^* and regularization strength λ. MSE (left) and in-sample error (right), as a function of the sample complexity α. Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data.

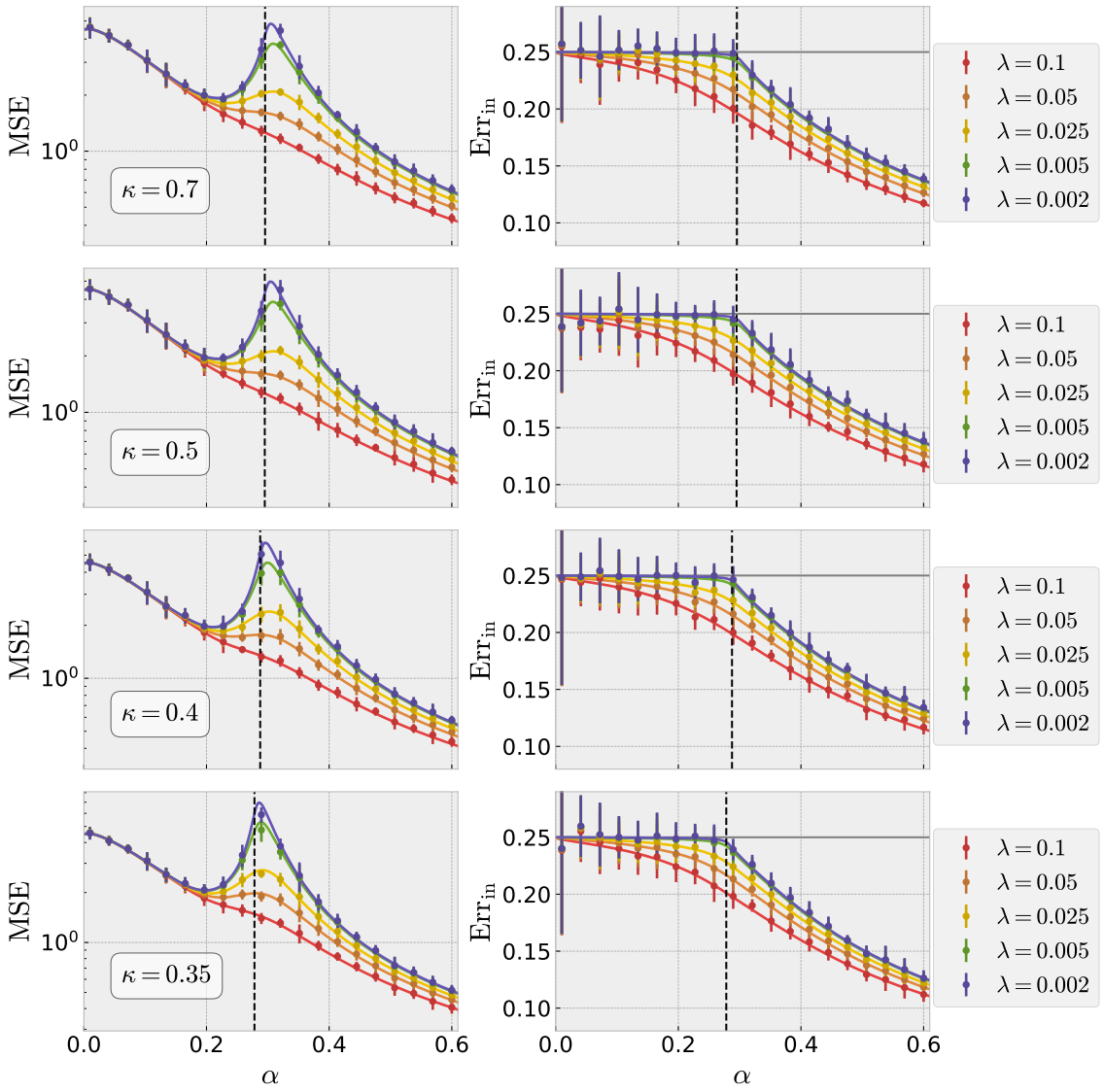

Figure 20: Comparison between simulations of gradient descent, defined in (6), and numerical integration of the system of equations (37), for $\kappa^* = 0.3, \Delta = 1.0$ and several values of the width κ and regularization strength λ. MSE (left) and in-sample error (right), defined in equation (73), as a function of the sample complexity α. Gradient descent simulations are averaged over 10 realizations of the initialization, teacher and data. The vertical dashed line indicates the interpolation threshold in the small regularization limit (see Proposition 9), and the horizontal gray line is the value $\Delta/4$.

J.2.2 Perfect Recovery Threshold

In this section, we give some numerical evidence regarding the perfect recovery threshold whose expression was conjectured in Conjecture 15. Because of the finite dimension and the finite time horizon at which our numerical simulations have been carried, the MSE always remains non-zero, but it is still possible to numerically investigate the perfect recovery threshold.

As $\alpha \to \alpha_{\mathrm{PR}}$, the MSE goes continuously to zero, but non-smoothly. In the statistical physics vocabulary, this corresponds to a second-order phase transition. Such transitions are often associated with a scaling of the form:

$$\mathrm{MSE}(\alpha) \underset{\alpha \to \alpha_{\mathrm{PR}}}{\sim} \big(\alpha_{\mathrm{PR}} - \alpha\big)^{\theta}, \tag{378}$$

for $\alpha < \alpha_{\mathrm{PR}}$ and $\theta > 0$. Then:

$$\frac{\mathrm{MSE}'(\alpha)}{\mathrm{MSE}(\alpha)} \underset{\alpha \to \alpha_{\mathrm{PR}}}{\sim} -\frac{\theta}{\alpha_{\mathrm{PR}} - \alpha} \xrightarrow[\alpha \to \alpha_{\mathrm{PR}}]{} -\infty.$$

Therefore, the perfect recovery threshold can be located by identifying a point where the above quantity exhibits a sharp minimum. This is illustrated in Figure 21. By computing the discrete derivative of log(MSE) with respect to α, we identify this minimum and use it as a measure of the perfect recovery threshold. In addition, we quantify the uncertainty of this estimation by measuring the width of the minimum at half depth.

This procedure, repeated over a large number of values for (κ, κ^*), allows to compare the numerical value of the perfect recovery threshold with the one given in Conjecture 15. In Figure 22, we compare the predictions of Conjecture 15 (continuous curves) with our numerical experiments (dots and error bars). This leads to a very convincing match and brings a numerical confirmation of our conjecture.

Additionally, the exponent θ in equation (378) can be numerically estimated using standard linear regression techniques. Based on the data obtained from simulations, we observe that θ depends on the parameters κ, κ^*, but further work is required to formulate a precise conjecture and investigate this dependence.

From a theoretical perspective, we expect that it is technically possible (although quite challenging) to derive an expression of this exponent. It would require working under the steady-state assumption, which under Conjecture 15 should hold for large values of κ and close enough to perfect recovery. In the end, such a calculation could be made possible by extending the results of Section H.4.

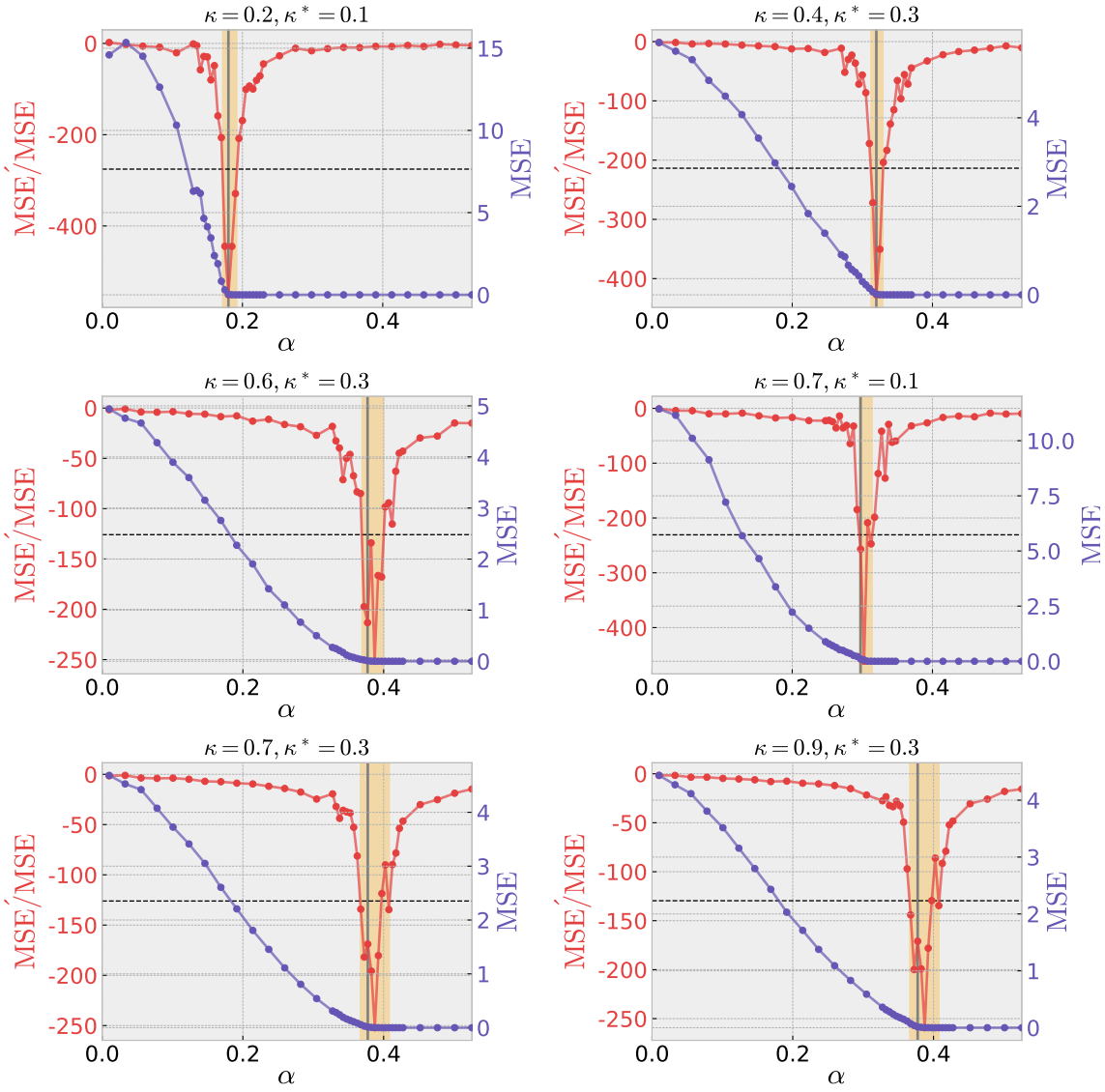

Figure 21: Measure of the perfect recovery threshold from gradient descent simulations. Red: the logarithmic derivative $\mathrm{MSE}'/\mathrm{MSE}$ as a function of α. This function exhibits a very sharp minimizer that coincides with the value of α where the MSE (purple curve) approaches zero. Horizontal dashed line: half depth of the minimizer, allowing to compute the uncertainty on the perfect recovery threshold (yellow region). Vertical gray line: conjectured value of the perfect recovery threshold in Conjecture 15.

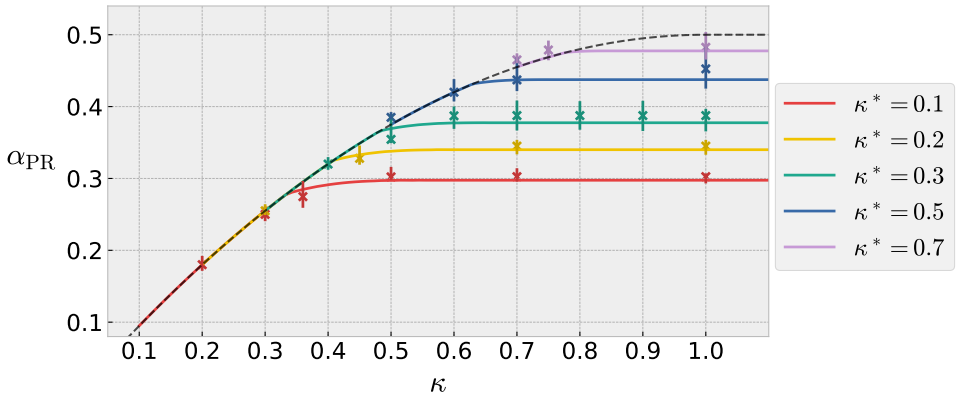

Figure 22: Perfect recovery threshold α_{PR} as a function of κ for different values of κ^*. Comparison between the prediction in equation (74) and the numerical estimation of α_{PR}. The vertical bars indicate the uncertainty on the measure computed from Figure 21.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84, 2020a.

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in Neural Information Processing Systems*, 33: 11022–11032, 2020b.

Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252, 2019.

Ada Altieri, Giulio Biroli, and Chiara Cammarota. Dynamical mean-field theory and aging dynamics. *Journal of Physics A: Mathematical and Theoretical*, 53(37):375006, 2020.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Cambridge university press, 2010.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: computational to statistical gaps in learning a two-layers neural network. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124023, jan 2019.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024a.

Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024b.

Afonso S Bandeira and Antoine Maillard. Exact threshold for approximate ellipsoid fitting of random points. *Electronic Journal of Probability*, 30:1–46, 2025.

Jean Barbier and Nicolas Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in Bayesian inference. *Probability Theory and Related Fields*, 174 (3):1133–1185, 2019.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *IEEE Transactions on Information Theory*, 66(7):4270–4303, 2020.

Jean Barbier, Francesco Camilli, Justin Ko, and Koki Okajima. Phase diagram of extensive-rank symmetric matrix denoising beyond rotational invariance. *Physical Review X*, 15 (2):021085, 2025a.

Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Statistical physics of deep learning: Optimal learning of a multi-layer perceptron near interpolation. *arXiv preprint arXiv:2510.24616*, 2025b.

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

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

Gérard Ben Arous and Alice Guionnet. Large deviations for Langevin spin glass dynamics. *Probability Theory and Related Fields*, 102(4):455–509, 1995.

Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Aging of spherical spin glasses. *Probability Theory and Related Fields*, 120(1):1–67, 2001.

Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-Kurchan equations for dynamics of spin-glasses. *Probability Theory and Related Fields*, 136(4):619–660, 2006.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

Gérard Ben Arous, Murat A Erdogdu, N Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. *arXiv preprint arXiv:2508.03688*, 2025.

Philippe Biane. On the free convolution with a semi-circular distribution. *Indiana University Mathematics Journal*, pages 705–718, 1997.

Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013.

Antoine Bodin and Nicolas Macris. Gradient flow on extensive-rank positive semi-definite matrix denoising. In *2023 IEEE Information Theory Workshop (ITW)*, pages 365–370, 2023.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Florent Krzakala, and Lenka Zdeborová. Inductive bias and spectral properties of single-head attention in high dimensions. *arXiv preprint arXiv:2509.24914*, 2025a.

Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborová. Bayes optimal learning of attention-indexed models. *arXiv preprint arXiv:2506.01582*, 2025b.

Tony Bonnaire, Giulio Biroli, and Chiara Cammarota. The role of the time-dependent Hessian in high-dimensional optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):083401, 2025.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35: 32240–32256, 2022.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pages 4345–4382, 2024.

Jean-Philippe Bouchaud. Weak ergodicity breaking and aging in disordered systems. *Journal de Physique I*, 2(9):1705–1713, 1992.

Jean-Philippe Bouchaud, Leticia F Cugliandolo, Jorge Kurchan, and Marc Mézard. Out of equilibrium dynamics in spin-glasses and other glassy systems. *Spin Glasses and Random Fields*, 12(161):9, 1998.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: a review of models, methods and applications. I. Models and methods. *arXiv preprint arXiv:2203.00446*, 2022.

Yuchen Chen and Yandi Shen. Learning single index model with gradient descent: spectral initialization and precise asymptotics. *arXiv preprint arXiv:2509.23527*, 2025.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Carson C Chow and Michael A Buice. Path integral methods for stochastic differential equations. *The Journal of Mathematical Neuroscience (JMN)*, 5(1):8, 2015.

Constantin Christof and Julia Kowalczyk. On the omnipresence of spurious local minima in certain neural network training problems. *Constructive Approximation*, pages 1–28, 2023.

Elisabetta Cornacchia, Francesca Mignacco, Rodrigo Veiga, Cédric Gerbelot, Bruno Loureiro, and Lenka Zdeborová. Learning curves for the multi-class teacher–student perceptron. *Machine Learning: Science and Technology*, 4(1):015019, 2023.

Andrea Crisanti, Heinz Horner, and H-J Sommers. The spherical p-spin interaction spin-glass model: the dynamics. *Zeitschrift für Physik B Condensed Matter*, 92(2):257–271, 1993.

Leticia Cugliandolo and Jorge Kurchan. A scenario for the dynamics in the small entropy production limit. *Journal of the Physical Society of Japan*, 69(Suppl. A):247–256, 2000.

Leticia F Cugliandolo and Jorge Kurchan. Analytical solution of the off-equilibrium dynamics of a long-range spin-glass model. *Physical Review Letters*, 71(1):173, 1993.

Leticia F Cugliandolo and Jorge Kurchan. On the out-of-equilibrium relaxation of the Sherrington-Kirkpatrick model. *Journal of Physics A: Mathematical and General*, 27 (17):5749, 1994.

Leticia F Cugliandolo, Vivien Lecomte, and Frédéric Van Wijland. Building a path-integral calculus: a covariant discretization approach. *Journal of Physics A: Mathematical and Theoretical*, 52(50):50LT01, 2019.

Hugo Cui, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pages 6468–6521, 2023.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for Gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36:54754–54768, 2023.

Francesco D'Angelo, Maksym Andriushchenko, Aditya Vardhan Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *Advances in Neural Information Processing Systems*, 37:23191–23223, 2024.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance(s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290, 2020a.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *Advances in Neural Information Processing Systems*, 33:3058–3069, 2020b.

C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

Thibaut Arnoulx De Pirey, Leticia F Cugliandolo, Vivien Lecomte, and Frédéric Van Wijland. Path integrals and stochastic calculus. *Advances in Physics*, 71(1-2):1–85, 2022.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Emanuele Troiani, Vittorio Erba, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. *arXiv preprint arXiv:2509.24882*, 2025.

Michal Derezinski, Feynman T Liang, and Michael W Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in Neural Information Processing Systems*, 33:5152–5164, 2020.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

David L Donoho, Matan Gavish, and Andrea Montanari. The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International Conference on Machine Learning*, pages 1329–1338, 2018.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685, 2019.

Nicolas Dupuis. *Field Theory of Condensed Matter and Ultracold Gases: Volume 1*. World Scientific, 2023.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2): 303–353, 1998.

Vittorio Erba, Emanuele Troiani, Luca Biggio, Antoine Maillard, and Lenka Zdeborová. Bilinear sequence regression: A model for learning from long sequences of high-dimensional tokens. *Physical Review X*, 15(2):021092, 2025a.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. *arXiv preprint arXiv:2505.17958*, 2025b.

Zhou Fan, Justin Ko, Bruno Loureiro, Yue M Lu, and Yandi Shen. Dynamical mean-field analysis of adaptive Langevin diffusions: Replica-symmetric fixed point and empirical Bayes. *arXiv preprint arXiv:2504.15558*, 2025.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference*, volume 6, pages 4734–4739, 2001.

David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Stationary points of shallow neural networks with quadratic activation function. *arXiv preprint arXiv:1912.01599*, 2019.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462, 2020.

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Physical Review E*, 109(3): 034305, 2024.

Cédric Gerbelot and Raphaël Berthier. Graph-based approximate message passing iterations. *Information and Inference: A Journal of the IMA*, 12(4):2562–2628, 2023.

Cedric Gerbelot, Alia Abbara, and Florent Krzakala. Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula). *IEEE Transactions on Information Theory*, 69(3):1824–1852, 2022.

Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471, 2022.

Loukas Grafakos et al. *Classical Fourier Analysis*, volume 2. Springer, 2008.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Alice Guionnet and Ofer Zeitouni. Large deviations asymptotics for spherical integrals. *Journal of Functional Analysis*, 188(2):461–515, 2002.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

William W Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, 1989.

Qiyang Han. Entrywise dynamics and universality of general first order methods. *The Annals of Statistics*, 53(4):1783–1807, 2025.

Harish-Chandra. Differential operators on a semisimple Lie algebra. *American Journal of Mathematics*, pages 87–120, 1957.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949, 2022.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Claude Itzykson and J-B Zuber. The planar approximation. ii. *Journal of Mathematical Physics*, 21(3):411–421, 1980.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Michel Journée, Francis Bach, P-A Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5): 2327–2351, 2010.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *Allerton Conference on Communication, Control, and Computing*, pages 92–99, 2019.

Seijin Kobayashi, Yassir Akram, and Johannes Von Oswald. Weight decay induces low-rank attention layers. *Advances in Neural Information Processing Systems*, 37:4481–4510, 2024.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in Neural Information Processing Systems*, 4, 1991.

Rep Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on Learning Theory*, pages 2–47, 2018.

Bruno Loureiro, Cedric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*, pages 14283–14314, 2022.

Antoine Maillard, Gérard Ben Arous, and Giulio Biroli. Landscape complexity for the empirical risk of generalized linear models. In *Mathematical and Scientific Machine Learning*, pages 287–327, 2020a.

Antoine Maillard, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase retrieval in high dimensions: Statistical and computational phase transitions. *Advances in Neural Information Processing Systems*, 33:11071–11082, 2020b.

Antoine Maillard, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Perturbative construction of mean-field equations in extensive-rank matrix factorization and denoising. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083301, 2022.

Antoine Maillard, Emanuele Troiani, Simon Martin, Florent Krzakala, and Lenka Zdeborová. Bayes-optimal learning of an extensive-width neural network from quadratically many samples. *Advances in Neural Information Processing Systems*, 37:82085–82132, 2024.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3655–3663, 2024.

Estelle Massart and P-A Absil. Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198, 2020.

Peter McCullagh. *Generalized Linear Models*. Routledge, 2019.

Song Mei and Andrea 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.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312, 2022.

Andrea Montanari and Pierfrancesco Urbani. Dynamical decoupling of generalization and overfitting in large two-layer networks. *arXiv preprint arXiv:2502.21269*, 2025.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Vanni Noferini. A formula for the Fréchet derivative of a generalized matrix function. *SIAM Journal on Matrix Analysis and Applications*, 38(2):434–457, 2017.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

Grigorios A Pavliotis. Stochastic processes and applications. *Texts in Applied Mathematics*, 60, 2014.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason D Lee. Emergence and scaling laws in SGD learning of shallow neural networks. *arXiv preprint arXiv:2504.19983*, 2025.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4433–4441, 2018.

Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is afraid of big bad minima? analysis of gradient-flow in spiked matrix-tensor models. *Advances in Neural Information Processing Systems*, 32, 2019.

Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. *Advances in Neural Information Processing Systems*, 33:13445–13455, 2020.

Guilhem Semerjian. Matrix denoising: Bayes-optimal estimators via low-degree polynomials. *Journal of Statistical Physics*, 191(10):139, 2024.

Stephen Smale. Stable manifolds for differential equations and diffeomorphisms. *Annali della Scuola Normale Superiore di Pisa-Scienze Fisiche e Matematiche*, 17(1-2):97–116, 1963.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Haim Sompolinsky and Annette Zippelius. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860, 1982.

Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259, 1988.

Shriram Srinivasan and Nishant Panda. What is the gradient of a scalar function of a symmetric matrix? *Indian Journal of Pure and Applied Mathematics*, 54(3):907–919, 2023.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds. *arXiv preprint arXiv:2502.00901*, 2025.

Daiki Tsuzuki and Kentaro Ohki. Global convergence of Oja's component flow for general square matrices and its applications. *arXiv preprint arXiv:2510.00801*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Luca Venturi, Afonso S Bandeira, and Joan Bruna. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133): 1–34, 2019.

Dan-Virgil Voiculescu. The derivative of order 1/2 of a free convolution by a semicircle distribution. *Indiana University Mathematics Journal*, pages 697–703, 1997.

Garrett G Wen, Hong Hu, Yue M Lu, Zhou Fan, and Theodor Misiakiewicz. When does Gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling. *arXiv preprint arXiv:2512.03325*, 2025.

Eugene P Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, pages 548–564, 1955.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Yizhou Xu, Antoine Maillard, Lenka Zdeborová, and Florent Krzakala. Fundamental limits of matrix sensing: Exact asymptotics, universality, and applications. *arXiv preprint arXiv:2503.14121*, 2025.

Wei-Yong Yan, Uwe Helmke, and John B Moore. Global analysis of Oja's flow for neural networks. *IEEE Transactions on Neural Networks*, 5(5):674–683, 1994.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.