Stochastic Gradient Flow Dynamics of Test Risk and its Exact Solution for Weak Features

Rodrigo Veiga [1] **Anastasia Remizova** [1] **Nicolas Macris** [1]

Abstract

We investigate the test risk of a continuous time stochastic gradient flow dynamics in learning theory. Using a path integral formulation we provide, in the regime of small learning rate, a general formula for computing the difference between test risk curves of pure gradient and stochastic gradient flows. We apply the general theory to a simple model of weak features, which displays the double descent phenomenon, and explicitly compute the corrections brought about by the added stochastic term in the dynamics, as a function of time and model parameters. The analytical results are compared to simulations of discrete time stochastic gradient descent and show good agreement.

1. Introduction

In supervised learning of neural networks and regression models, understanding the dynamics of optimization algorithms, and in particular stochastic gradient descent (SGD), is of utmost importance. However, despite much progress in a number of directions, this still remains a highly challenging theoretical problem. A fruitful approach that allows making analytical progress consists of suitably approximating SGD by a continuous time approximation, henceforth called stochastic gradient flow (SGF). In this contribution, we build up on this approach, to develop a general formalism characterizing the dynamics of the stochastic process, and apply it to the investigation of the test risk (or generalization error) as a function of time.

As is well known, the classical bias-variance trade-off has been challenged in a number of models displaying the double descent phenomenon (Belkin et al., 2019; 2020; Belkin, 2021). Analytical derivations of double descent curves have been achieved for relatively simple models, but are limited to the use of least squares estimators (no dynamics) and pure gradient flow (GF) approximations of gradient descent (GD). The present work goes one step further by investigating the effects of stochasticity on the double descent curve.

Main contributions – Our main contributions are summarized as follows:

C1 We consider a general Itô stochastic differential equation (SDE) and represent the Markovian transition probability as a path integral, Eq. (11). A general 'explicit' formula for the transition probability, Eq. (17), is derived in the limit of a small learning rate by using a Laplace approximation. This constitutes one of the main results of this paper: suppose that $\boldsymbol{w}^{\mathrm{ode}}(t)$ is the deterministic solution of pure GF (with no stochastic term), then by representing the large deviation of the Itô diffusion of SGF as $\boldsymbol{w}^{\mathrm{ode}}(t) + \sqrt{\gamma}\boldsymbol{z}(t)$ we provide a general formula for the covariance $\boldsymbol{\mathscr{C}}(t) \equiv \mathbb{E}[\boldsymbol{z}(t)\boldsymbol{z}(t)^\top]$. The formula is confronted against a simple rigorously solvable SDE and we find exact agreement.

C2 We use the simplest relevant SGF approximation of SGD in a learning theory setting, which is known to come with a first-order guarantee with respect to small learning rates (Li et al., 2019). We use our general theory to express the covariance $\boldsymbol{\mathscr{C}}(t)$ in terms of the data matrix, training cost and its Hessian, and the deterministic GF trajectory.

C3 This general theoretical framework is applied to a particular random weak features model, displaying the double descent phenomenon, first discussed by (Breiman & Freedman, 1983) and revisited in (Belkin et al., 2020). We explicitly compute the corrections brought about by the added stochastic term in the *whole* learning dynamics. For i.i.d. Gaussian data, and in a suitable large-size asymptotic regime, the formula for the test risk of SGF is given explicitly by an expression involving only single and double integrals over the Marchenko-Pastur distribution, Eqs. (29) and (30). We check our analytical predictions against numerical simulations of SGD and find good agreement.

[1]École Polytechnique Fédérale de Lausanne (EPFL), Lab for Statistical Mechanics of Inference in Large Systems (SMILS), CH-1015 Lausanne, Switzerland. Correspondence to: R. Veiga <rodrigo.veiga@epfl.ch>, A. Remizova <anastasia.remizova@epfl.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Relation to previous work – Approximating discrete time stochastic processes by continuous time dynamics has been studied for more than three decades (see Kushner & Yin 2003 for a comprehensive monograph). In particular, many works have attempted to approximate SGD by considering continuous time Ornstein-Uhlenbeck processes (Mandt et al., 2016; 2017; Jastrzębski et al., 2018; Fan et al., 2018; Wang & Wu, 2020) assuming constant covariance in the continuous time limit; however this is not necessarily a good approximation (Ali et al., 2020). In (Mandt & Blei, 2015; Li et al., 2015) the authors heuristically proposed 'better' SDEs with a non-trivial covariance. This framework is a stochastic version of the method of modified equations in the analysis of classical finite difference methods (Warming & Hyett, 1974). Further developments are found in (Li et al., 2017; 2019) where rigorous approximation results, in the sense of convergence in probability, are proved.

The case of regularized least squares in high-dimensions is discussed in (Paquette et al., 2022b;a) where the dynamics of the so-called *homogenized* SGD (Paquette & Paquette, 2021) is characterized with rigorous guarantees in terms of Volterra integral equations. We expect the solution of these equations to agree with ours when applied to the weak features model in the limit of a small learning rate. Nevertheless, the saddle point calculation of the path integral employed here circumvents the need to solve Volterra integral equations, which is itself a non-trivial task.

Other interesting formulations of SGF were also analyzed in (Mignacco et al., 2020; 2021) through Dynamical Mean-Field Theory (DMFT) in the contexts of high-dimensional Gaussian mixture classification and phase retrieval. Rigorous advances along this line of work were recently obtained (Gerbelot et al., 2024). Furthermore, the role of stochasticity in escaping flat directions near initialization is discussed in (Arnaboldi et al., 2023) in the context of two-layer neural networks.

The path integral formulation used here remains at a heuristic level. We point out that for more restricted SDEs (with constant covariance and small noise) there exist rigorous formulations belonging to the Freidlin-Wentzell large deviation theory (see Freidlin & Wentzell 2012 and Dembo & Zeitouni 2010, Chapter 5.6). Extending the Freidlin-Wentzell large deviation theory with state-dependent covariances is, to the best of our knowledge, an open mathematical problem.

The complete time dependence of training and generalization errors of GF has been calculated for the random features and Gaussian covariate models in (Bodin & Macris, 2021; 2022) using advanced random matrix machinery. However, the corresponding curves for SGF have not been investigated within these models. Here, as our main focus is on SGF dynamics, we limit ourselves to the simplest possible weak features model which requires minimal random matrix theory, and postpone the harder calculation for other models to later work.

An interesting study of SGF, and notably its implicit bias, has been initiated for diagonal networks in a recent series of papers (Pesme et al., 2021; Pesme & Flammarion, 2023), where the benefit of the stochasticity for large training times has been uncovered, provided that the weights are properly initialized. Additionally, we mention the *decoupling approximation* near local minima proposed in (Mori et al., 2022), making the noise contribution additive near these minima and considerably simplifying the stochastic dynamics. These perspectives could also constitute an interesting test lab for the present approach, with which one could analyze the effects of stochasticity over the whole dynamics.

Outline – The simplest relevant SGF approximation of SGD is briefly described in Section 2. The general path integral is formulated in Section 3, where we also derive the Laplace approximation, test the theory against a simple solvable model, and apply it to a general learning theory setting. In Section 3.1 we compute the whole time evolution of the test risk under SGF for the weak features model, and compare the theoretical predictions to numerical simulations of SGD.

Reproducibility – Github repository.

2. Stochastic modeling of SGD

Consider a data set composed of n pairs $(\boldsymbol{x}^k, y^k)_{k\in[n]} \in \mathbb{R}^{d+1} \sim \mathbb{P}(\boldsymbol{x}, y)$ i.i.d. and a model parametrized by $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$. Starting with $\hat{\boldsymbol{\beta}}^0$, the usual full-batch GD update is $\hat{\boldsymbol{\beta}}^{\nu+1} = \hat{\boldsymbol{\beta}}^\nu - \gamma \boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}^\nu; \boldsymbol{X}, \boldsymbol{y}\big)$, where $\mathscr{L}\big(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}\big) = \frac{1}{n}\sum_{k=1}^n l\big(\boldsymbol{\beta}; \boldsymbol{x}^k, y^k\big)$ is the total training loss, with $\boldsymbol{X} = [\boldsymbol{x}_1|\dots|\boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times d}$ being the data matrix, $\boldsymbol{y} = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$ the target vector and $l\big(\boldsymbol{\beta}; \boldsymbol{x}^k, y^k\big)$ a penalization function. The parameter $\gamma > 0$ is the *learning rate* and $\nu = 0, 1, 2, \dots$. The GD rule can be thought of as a discretization of the continuous ordinary differential equation (ODE):

$$\frac{\mathrm{d}\hat{\boldsymbol{\beta}}(t)}{\mathrm{d}t} = -\boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big) \ , \tag{1}$$

with $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}}^0$, known as *gradient flow* (GF).

Stochastic gradient descent (SGD), on the other hand, does not make use of the whole training set at each optimization step. In its *single batch* form, it can be written as $\hat{\boldsymbol{\beta}}^{\nu+1} = \hat{\boldsymbol{\beta}}^\nu - \gamma \boldsymbol{\nabla} l\big(\hat{\boldsymbol{\beta}}^\nu; \boldsymbol{x}^{k_\nu}, y^{k_\nu}\big)$ with k_ν sampled uniformly in $\{1, n\}$. In general, one could perform *batch* SGD by sampling a set $\mathcal{B}^\nu$ of pairs, and replace the gradient above by $(1/|\mathcal{B}^\nu|)\sum_{k\in\mathcal{B}^\nu} l\big(\hat{\boldsymbol{\beta}}^\nu; \boldsymbol{x}^k, y^k\big)$. Conceptually, the

general path-integral approach we introduce in Section 3 does not change with the batch size. Therefore, for the sake of simplicity in the exposition, we adopt single batch SGD in the mathematical formulation. The expressions can be straightforwardly extended to the batch case by replacing the single sample gradient by the average gradient over the batch.

Summing and subtracting the total training loss gradient $\gamma \boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}^{\nu}; \boldsymbol{X}, \boldsymbol{y}\big)$ on the SGD update makes the effective noise explicit:

$$\hat{\boldsymbol{\beta}}^{\nu+1} = \underbrace{\hat{\boldsymbol{\beta}}^{\nu} - \gamma \boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}^{\nu}; \boldsymbol{X}, \boldsymbol{y}\big)}_{\text{Full-bath GD}} - \gamma \underbrace{\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)}_{\text{Stochastic perturbation}} , \quad (2)$$

where the stochastic perturbation is by definition:

$$\boldsymbol{\nabla} l\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{x}^{k_\nu}, y^{k_\nu}\big) - \boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big) . \quad (3)$$

Since k_ν is sampled uniformly in $\{1, n\}$, the stochastic perturbation is a zero mean vector, $\mathbb{E}_{k_\nu \sim \mathcal{U}\{1,n\}}\left[\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)\right] = \boldsymbol{0}$. In the spirit of *gradient flow*, Eq. (2) can be thought of as a discretization of a stochastic differential equation (SDE), which we will call *stochastic gradient flow* (SGF), in the continuous-time limit. Here we choose the simplest representation proposed in (Mandt & Blei, 2015) and (Li et al., 2015):

$$\mathrm{d}\hat{\boldsymbol{\beta}}(t) = -\boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)\mathrm{d}t + \sqrt{\gamma\, \boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)}\, \mathrm{d}\boldsymbol{b}(t) , \quad (4)$$

where $\mathrm{d}\boldsymbol{b}(t) = \boldsymbol{b}(t+\mathrm{d}t) - \boldsymbol{b}(t) \in \mathbb{R}^n$ is the forward Itô increment associated to a standard n-dimensional Wiener process with $\boldsymbol{b}(0) = \boldsymbol{0}, \mathbb{E}[\boldsymbol{b}(\tau)] = \boldsymbol{0}, \mathbb{E}[\boldsymbol{b}(t)\boldsymbol{b}(t')^\top] = \mathbb{I}_n \min(t, t')$, with $\mathbb{I}_n$ as the $n \times n$ identity matrix. The matrix $\boldsymbol{\Sigma}$ is matched to the covariance matrix of the stochastic perturbation in (3):

$$\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big) \equiv \mathbb{E}_{k_\nu}\left[\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)^\top\right] . \quad (5)$$

This is a $d \times d$ positive semi-definite matrix and its "square-root" is defined as the $d \times n$ matrix such that $\sqrt{\boldsymbol{\Sigma}}\sqrt{\boldsymbol{\Sigma}}^\top = \boldsymbol{\Sigma}$.

Solving the SGF (4) can be far from trivial, since a priori both the *drift vector* $\boldsymbol{\nabla} \mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)$ and the *diffusion matrix* $\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)$ depend on the stochastic process $\hat{\boldsymbol{\beta}}(t)$ and on the data $(\boldsymbol{X}, \boldsymbol{y})$. We provide in the next section a general method to compute the fluctuations around the deterministic trajectory when the learning rate is small.

3. General path integral formulation

Inspired by Eq. (4), we consider the following general SDE for a process $\boldsymbol{w}(\tau) \in \mathbb{R}^d$, sampled between t_0 and $t > t_0$:

$$\mathrm{d}\boldsymbol{w}(\tau) = \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\mathrm{d}\tau + \sqrt{\gamma}\, \boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right)\mathrm{d}\boldsymbol{\eta}(\tau) , \quad (6)$$

where $\mathrm{d}\boldsymbol{\eta}(\tau) = \boldsymbol{\eta}(\tau + \mathrm{d}\tau) - \boldsymbol{\eta}(\tau) \in \mathbb{R}^n$ is the forward Itô increment of a standard n-dimensional Wiener process (as defined previously). We assume this SDE has a unique solution. Standard conditions ensuring existence and uniqueness are Lipshitzness and linear growth of $\boldsymbol{f}$ and $\boldsymbol{G}$ w.r.t $\boldsymbol{w}$ uniformly in τ (see for example Evans 2012, Section 5.B.3). The explicit expressions of the drift vector $\boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right) \in \mathbb{R}^d$ and the diffusion matrix $\boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right) \in \mathbb{R}^{d \times n}$ in the learning theory context will depend on the model (data distribution, architecture and loss function). However, the method discussed here is not restricted to machine learning models: it can be viewed as a general approximation scheme to solve SDEs for small γ. To stress that fact, we generically named the stochastic process as $\boldsymbol{w}(\tau)$ in this section, which in the learning theory context is the estimator $\hat{\boldsymbol{\beta}}(\tau)$ itself or some parametrization of it. The general path integral formulation is outlined in this section; the complete detailed derivation is contained in Appendix A.

Let $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0)$ be the transition probability of the process associated to Eq. (6) to go from $\boldsymbol{w}_0$ at time t_0 to $\boldsymbol{w}$ at time t. If $\gamma = 0$ the SDE (6) becomes a first-order ODE:

$$\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} = \boldsymbol{f}(\tau, \boldsymbol{w}(\tau)) . \quad (7)$$

The solution of this ODE with initial condition $\boldsymbol{w}^{\text{ode}}(t_0) = \boldsymbol{w}_0$ (we assume a unique global solution here) is denoted by $\boldsymbol{w}^{\text{ode}}(\tau)$. In such case: $\mathbb{P}_{\gamma=0}(\boldsymbol{w}, t|\boldsymbol{w}_0, t) = \delta\left(\boldsymbol{w} - \boldsymbol{w}^{\text{ode}}(t)\right)$. Our goal is to study the fluctuations around the deterministic trajectories $\boldsymbol{w}^{\text{ode}}(\tau)$ when γ is small.

The continuous-time Itô SDE (6) must be understood according to its discrete-time companion process $\boldsymbol{w}_k = \boldsymbol{w}(k\Delta\tau)$:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \boldsymbol{f}(k, \boldsymbol{w}_k)\Delta\tau + \sqrt{\gamma}\, \boldsymbol{G}(k, \boldsymbol{w}_k)\Delta\boldsymbol{\eta}_k , \quad (8)$$

where the time interval $[t_0, t]$ has been discretized into N slices of length $\Delta\tau \equiv \frac{t-t_0}{N}$ with $\tau_k = t_0 + k\Delta\tau$ for $k = 0, \ldots, N$ and $\tau_N = t$. The discrete quantities $\Delta\boldsymbol{\eta}_k$ are sampled independently at each step k from a Gaussian with $\mathbb{E}[\Delta\boldsymbol{\eta}_k] = \boldsymbol{0}$, $\mathbb{E}[\Delta\boldsymbol{\eta}_k(\Delta\boldsymbol{\eta}_k)^\top] = \Delta\tau\ \mathbb{I}_n$. In general the covariance matrix $\boldsymbol{G}(k, \boldsymbol{w}_k)\boldsymbol{G}(k, \boldsymbol{w}_k)^\top$ is only positive semi-definite, and in particular when $d > n$ it certainly has zero eigenvalues. Thus the process might become singular in sub-manifolds of $\mathbb{R}^d$. This is cured by introducing a regularization parameter $\epsilon > 0$ and replacing Eq. (8) by

$$\Delta\boldsymbol{w}_k \equiv \boldsymbol{w}_{k+1} - \boldsymbol{w}_k \sim \mathcal{N}\left(\boldsymbol{f}_k\Delta\tau, \gamma\Delta\tau\left(\boldsymbol{G}_k\boldsymbol{G}_k^\top + \epsilon\mathbb{I}_d\right)\right) , \quad (9)$$

with $\boldsymbol{f}_k \equiv \boldsymbol{f}(k, \boldsymbol{w}_k)$ and $\boldsymbol{G}_k \equiv \boldsymbol{G}(k, \boldsymbol{w}_k)$.

In order to compute the continuous time limiting transition probability and associated expected values, the correct prescription is to take the limit $\epsilon \to 0$ *after* $N \to +\infty$.

The one-step propagator $\mathbb{P}_{\Delta\tau}(\boldsymbol{w}_{k+1}, \tau_{k+1}|\boldsymbol{w}_k, \tau_k)$ associated with Eq. (8) is known from Eq. (9). Then from the Chapman-Kolmogorov equation, $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0)$ equals:

$$\lim_{\epsilon\to 0}\lim_{N\to\infty}\int\prod_{k'=1}^{N-1}\mathrm{d}\boldsymbol{w}_{k'}\prod_{k=0}^{N-1}\mathbb{P}_{\Delta\tau}(\boldsymbol{w}_{k+1}, \tau_{k+1}|\boldsymbol{w}_k, \tau_k)$$

which yields

$$\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \lim_{\epsilon\to 0}\lim_{N\to\infty}\int\prod_{k'=1}^{N-1}\frac{\mathrm{d}\boldsymbol{w}_{k'}}{(2\pi\gamma\Delta\tau)^{d/2}\left(\det\boldsymbol{\Omega}_k^\epsilon\right)^{1/2}}$$
$$\exp\left[-\frac{\Delta\tau}{2\gamma}\sum_{k=0}^{N-1}\left(\frac{\Delta\boldsymbol{w}_k}{\Delta\tau}-\boldsymbol{f}_k\right)^\top(\boldsymbol{\Omega}_k^\epsilon)^{-1}\left(\frac{\Delta\boldsymbol{w}_k}{\Delta\tau}-\boldsymbol{f}_k\right)\right], \tag{10}$$

where $\boldsymbol{\Omega}_k^\epsilon \equiv \boldsymbol{G}_k\boldsymbol{G}_k^\top + \epsilon\mathbb{I}_d$ is the regularized covariance.

In the continuous-time limit, the expression above becomes a *path integral* over all possible trajectories $\boldsymbol{w}(\tau) \in \mathbb{R}^d$ connecting $\boldsymbol{w}_0$ and $\boldsymbol{w}$ from time t_0 to t:

$$\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \lim_{\epsilon\to 0}\int_{\boldsymbol{w}(t_0)=\boldsymbol{w}_0}^{\boldsymbol{w}(t)=\boldsymbol{w}}\mathcal{D}\left[\boldsymbol{w}(\cdot)\right]\exp\left(-\frac{1}{\gamma}\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right]\right). \tag{11}$$

The *action* functional is defined as

$$\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right] \equiv \int_{t_0}^{t}\mathrm{d}\tau\ \mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right), \tag{12}$$

with the *Lagrangian*:

$$\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = \frac{1}{2}\left(\dot{\boldsymbol{w}}(\tau) - \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\right)^\top$$
$$\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\left(\dot{\boldsymbol{w}}(\tau) - \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\right), \tag{13}$$

with $\dot{\boldsymbol{w}}(\tau) = \frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau}$, $\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau)) \equiv \boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right)\boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right)^\top + \epsilon\mathbb{I}_d$ and $\mathcal{D}\left[\boldsymbol{w}(\cdot)\right]$ as defined in Eq. (42) of the Appendix A.

The path measure and action functional are to be understood under the Itô discretization (10). If the SDE had been formulated with Stratonovich or other intermediate Brownian increments, the path integral would have to be understood accordingly (de Pirey et al., 2023).

Laplace approximation – For small enough γ, the path integral (11) is dominated by trajectories minimizing the action $\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right]$. This is the path given by Hamilton's action principle (e.g., like in classical analytical mechanics) $\delta\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right] = 0$ for variations over paths satisfying the *two boundary conditions* $\boldsymbol{w}(t_0) = \boldsymbol{w}_0$, $\boldsymbol{w}(t) = \boldsymbol{w}$. A path is a stationary point of the action if and only if it satisfies the Euler-Lagrange equations:

$$\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\partial}{\partial\dot{w}_j} - \frac{\partial}{\partial w_j}\right)\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = 0 \tag{14}$$

for $j = 1, \ldots, d$, supplemented by the two boundary conditions. Replacing (13) into the Euler-Lagrange equations (14) renders a complicated set of second-order differential equations. However, it is not very difficult to prove that (see Appendix A.2):

- If the terminal point $\boldsymbol{w}$ at time t is *on* the trajectory $\boldsymbol{w}^{\mathrm{ode}}(\cdot)$ i.e., it is on the unique solution of the first-order initial value problem $\dot{\boldsymbol{w}}(\tau) = \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))$, with $\boldsymbol{w}(0) = \boldsymbol{w}_0$, then the solution of the Euler-Lagrange equation (14) is precisely $\boldsymbol{w}^{\mathrm{ode}}(\tau)$ and $\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = 0$.
- If the terminal point $\boldsymbol{w}$ at time t is *not on* the trajectory $\boldsymbol{w}^{\mathrm{ode}}(\cdot)$, then the solution of the Euler-Lagrange equation cannot satisfy exactly the first order ODE i.e., $\dot{\boldsymbol{w}}(\tau) \neq \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))$ and $\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) > 0$.

In particular, we conclude that for small γ's:

- When the terminal condition is away from the trajectory $\boldsymbol{w}^{\mathrm{ode}}(\cdot)$ by an amount $|\boldsymbol{w}(t) - \boldsymbol{w}^{\mathrm{ode}}(t)| = \Omega_\gamma(1)$, then $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \mathcal{O}(e^{-C/\gamma})$ is an exponentially small probability.
- When the terminal condition is close enough to the trajectory $\boldsymbol{w}(\cdot)$ in the sense $|\boldsymbol{w}(t) - \boldsymbol{w}^{\mathrm{ode}}(t)| = \mathcal{O}(\sqrt{\gamma})$, then the $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0)$ is given by the fluctuations around the stationary point of the action functional. This picture is schematically illustrated in Fig. 1.

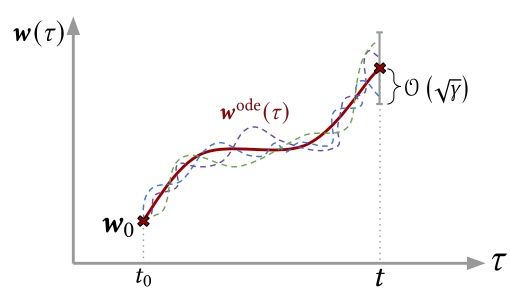


Figure 1. The solution $\boldsymbol{w}^{\mathrm{ode}}(\tau)$ of the first order ODE $\dot{\boldsymbol{w}}(\tau) = \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))$, $\boldsymbol{w}(0) = \boldsymbol{w}_0$ yields a minimum vanishing action. The dominant contributions to the path integral for small γ are fluctuations of order $\mathcal{O}(\sqrt{\gamma})$ around this trajectory.

To find the probability distribution of these fluctuations, we set $\boldsymbol{w}(\tau) = \boldsymbol{w}^{\mathrm{ode}}(\tau) + \sqrt{\gamma}\ \boldsymbol{z}(\tau)$ with $\boldsymbol{z}(0) = \boldsymbol{0}$[1] and

[1] At time t we have $\boldsymbol{w} = \boldsymbol{w}^{\mathrm{ode}}(t) + \sqrt{\gamma}\boldsymbol{z}(t)$ and because $\boldsymbol{w}$ is 'random' we should not constrain the terminal condition $\boldsymbol{z}(t)$.

expand the action functional (12) to second order in $\sqrt{\gamma}\boldsymbol{z}(\cdot)$. The zero-th order term vanishes because $\boldsymbol{w}^{\text{ode}}(\tau)$ satisfies the first order ODE, and the first order term also vanishes because it satisfies the Euler-Lagrange equation. Thus we find:

$$\mathcal{S}^{\epsilon}\left[\boldsymbol{w}(\cdot)\right] \approx \frac{\gamma}{2}\int_{t_0}^{t} \mathrm{d}\tau \left(\frac{\mathrm{d}\boldsymbol{z}(\tau)}{\mathrm{d}\tau} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\boldsymbol{z}(\tau)\right)^{\top} \boldsymbol{\Omega}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) \left(\frac{\mathrm{d}\boldsymbol{z}(\tau)}{\mathrm{d}\tau} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\boldsymbol{z}(\tau)\right) , \quad (15)$$

where $\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}$ is the Jacobian matrix of $\boldsymbol{f}$ along the trajectory $\boldsymbol{w}^{\text{ode}}(\cdot)$ (see Eq. (49) for an explicit definition). Eq. (15) gives us the Gaussian probability distributions of fluctuations $\boldsymbol{z}(\cdot)$ around the trajectory $\boldsymbol{w}^{\text{ode}}(\cdot)$ (see Fig. 1). This has zero mean $\mathbb{E}[\boldsymbol{z}(\cdot)] = \boldsymbol{0}$ and covariance $\boldsymbol{\mathscr{C}}(\cdot) \equiv \mathbb{E}[\boldsymbol{z}(\cdot)\boldsymbol{z}(\cdot)^{\top}]$ given by

$$\boldsymbol{\mathscr{C}}(t) = \int_{t_0}^{t} \mathrm{d}\tau\, \boldsymbol{U}^{\text{ode}}(t,\tau)\, \boldsymbol{\Omega}^{0}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\, \boldsymbol{U}^{\text{ode}}(t,\tau)^{\top} , \quad (16a)$$

where

$$\boldsymbol{U}^{\text{ode}}(t,\tau) \equiv \mathcal{T}\left\{e^{-\int_t^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} , \quad (16b)$$

with $\mathcal{T}\{\cdots\}$ denoting the time-ordered exponential (see Appendix B). The complete derivation of the covariance is found in the Appendix A.

Using Wick's theorem, this covariance allows to compute any expectation that reduces to multinomials of $\boldsymbol{z}(t_1), \boldsymbol{z}(t_2), \ldots, \boldsymbol{z}(t_m)$. In particular, we will see below that the generalization error involves the simplest such polynomial $\boldsymbol{z}(t)^{\top}\boldsymbol{z}(t)$ at the terminal time.

Finally, within our approximate theory, we have:

$$\mathbb{P}_{\gamma}(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) \approx \frac{e^{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}^{\text{ode}}(t))^{\top}\boldsymbol{\mathscr{C}}(t)^{-1}(\boldsymbol{w}-\boldsymbol{w}^{\text{ode}}(t))}}{(2\pi)^{d/2}(\det \boldsymbol{\mathscr{C}}(t))^{1/2}} . \quad (17)$$

We note that for $d > n$ the $d \times d$ covariance matrix has zero eigenvalues and this expression is somewhat formal. This means that the probability distribution has its support in an n-dimensional subspace of $\mathbb{R}^d$. In the context of machine learning, $d > n$ represents the over-parameterized regime.

3.1. An exactly solvable example

In this short section, we verify the above result for a simple SDE which is exactly solvable. Consider the one-dimensional stochastic process $\{w(\tau), \tau \geq t_0\}$ defined by

$$\mathrm{d}w(\tau) = h(\tau)(y - w(\tau))\mathrm{d}\tau + \sqrt{\gamma}\sigma(\tau)\mathrm{d}\eta(\tau) \quad (18)$$

with initial condition $w(t_0) = w_0$. The functions h and σ are real valued continuous for $\tau \in [0, 1)$, $y \in \mathbb{R}$, and η is a standard one dimensional Brownian motion. One checks explicitly that the solution is:

$$w(\tau) = \phi(\tau)w_0 + (1 - \phi(\tau))y + \sqrt{\gamma}\phi(\tau)\int_{t_0}^{\tau} \mathrm{d}\eta(s)\frac{\sigma(s)}{\phi(s)} \quad (19)$$

where $\phi(\tau) = \exp\Big(-\int_{t_0}^{\tau} \mathrm{d}s\, h(s)\Big)$. This equation implies that the process is Gaussian with transition probability $\mathbb{P}_{\gamma}(w, t|w_0, t_0)$ with mean and variance:

$$\mathbb{E}[w(t)] = y + (w_0 - y)e^{-\int_{t_0}^{t} \mathrm{d}s\, h(s)} , \quad (20a)$$

$$\mathrm{Var}[w(t)] = \gamma\int_{t_0}^{t} \mathrm{d}u\, \sigma(u)^2 e^{-2\int_u^t \mathrm{d}s\, h(s)} . \quad (20b)$$

It is an exercise to check that (20a) coincides with $w^{\text{ode}}(t)$, the solution of the first order ODE $\frac{\mathrm{d}w(t)}{\mathrm{d}t} = h(t)(y - w(t))$ and (20b) coincides with (16). We conclude that our general theory is exact for the simple linear SDE (18) .

In (Hildebrandt & Rœlly, 2020) the authors discuss in detail a *pinning phenomenon* which is of relevance to us. Suppose h and σ satisfy (i) $h \geq 0$; (ii) $\lim_{t\to+\infty}\int_{t_0}^{\infty} \mathrm{d}s\ h(s) = +\infty$; (iii) $\int_{t_0}^{\infty} \mathrm{d}s\sigma(s)^2$ is finite. Then we have for $t \to +\infty$, $\mathbb{E}[w(t)] \to y$ and $\mathrm{Var}[w(t)] \to 0$ (the later point can be shown by noticing that $\mathrm{Var}[w(t)] \leq \gamma\int_{t_0}^{+\infty} \mathrm{d}u\, \sigma(u)^2 e^{-2\int_u^t \mathrm{d}s\, h(s)}$ and applying the dominated convergence theorem to the right hand side; for more details we refer to Hildebrandt & Rœlly 2020). These diffusion processes become 'pinned' onto y as $t \to +\infty$. For a simple picture, one may keep in mind two continuous functions, such that for $r \to +\infty$, on one hand h does not tend to zero fast enough and on the other hand $\sigma(r)$ tends to zero fast enough. Roughly speaking for large times the process fluctuates around the gradient flow path for a 'loss function' $\frac{1}{2}h(\tau)(y - w)^2$ and for large times, as the fluctuations become smaller, it tends to the minimum.

Note that if $\sigma(s)^2$ is not integrable, pinning does not necessarily occur. For example, suppose $\sigma(s) = b$ and $h(s) = a > 0$ two constants. Then an explicit computation shows $\mathbb{E}[w(t)] \to y$ and $\mathrm{Var}[w(t)] \to \gamma b^2/2a$.

3.2. Test risk of SGF

We now come back to the setting of learning theory. Comparing equations (4) and (5), we have the correspondence $\boldsymbol{w}(t) \to \hat{\boldsymbol{\beta}}(t)$, $\boldsymbol{f}(t, \boldsymbol{w}(t)) \to -\boldsymbol{\nabla}\mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)$, $G(t, \boldsymbol{w}(t)) \to \sqrt{\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)}$ and $\mathrm{d}\boldsymbol{\eta}(t) \to \mathrm{d}\boldsymbol{b}(t)$. Therefore, we approximate the solution of (4) by $\hat{\boldsymbol{\beta}}(t) = \hat{\boldsymbol{\beta}}^{\text{GF}} + \sqrt{\gamma}\boldsymbol{z}(t)$ with $\hat{\boldsymbol{\beta}}^{\text{GF}}$ the solution of (1) and the Gaussian

process $\boldsymbol{z}(t)$ has covariance (setting now $t_0 = 0$):

$$\boldsymbol{\mathscr{C}}(t) = \int_0^t \mathrm{d}u\, \boldsymbol{U}^{\mathrm{GF}}(t,u)\, \boldsymbol{\Sigma}(u, \hat{\boldsymbol{\beta}}^{\mathrm{GF}}(u))\, \boldsymbol{U}^{\mathrm{GF}}(t,u)^\top \,, \tag{21a}$$

where now

$$\boldsymbol{U}^{\mathrm{GF}}(t,u) = \mathcal{T}\left\{ e^{-\int_u^t \mathrm{d}r \boldsymbol{\mathcal{H}}_{\mathscr{L}}(\hat{\boldsymbol{\beta}}^{\mathrm{GF}}(r))} \right\} \,, \tag{21b}$$

with $\boldsymbol{\mathcal{H}}_{\mathscr{L}} = \boldsymbol{\nabla} \otimes \boldsymbol{\nabla} \mathscr{L}$ the $d \times d$ Hessian matrix of the loss function (see also A.4 for a slightly more general setting).

From these formulas, we may expect a set of qualitative behaviors for the generalization error, which we now discuss. We stress, however, that these are by no means necessarily universal and their validity should be checked for each model. For $d < n$, below the interpolation threshold, $\boldsymbol{\Sigma}$ is generically full rank, so the covariance is also expected to be full rank. In other words, we expect the stochastic process to fluctuate in parts of $\mathbb{R}^d$ of 'full measure' even for large times. The difference between the generalization errors of GF and SGF should be non-trivial. On the other hand, above the interpolation threshold, $\boldsymbol{\Sigma}(\tau, \hat{\boldsymbol{\beta}}^{\mathrm{GF}}(\tau))$ certainly has $d - n$ zero eigenvalues (and possibly more). Therefore, we expect the process to fluctuate in a 'submanifold' of $\mathbb{R}^d$. Such submanifold can be quite complicated as it will in general depend on time and on the GF trajectory. For large times we furthermore expect that $\boldsymbol{\Sigma}$ tends to zero altogether. Indeed, for linear networks this matrix is proportional to the loss (see next section), which tends to zero above the interpolation threshold; and more generally if the number of data points is much less than the number of parameters, the data points are seen many times and there should not be a major difference between a full batch and a mini batch. Consequently, for large times we expect that the diffusion is pinned on the minima of the loss, and that there should be no essential difference between generalization error of GF and SGF. These considerations are corroborated in the next section by an explicit analysis of a simple model.

4. Application to a weak features model

4.1. Setting

We consider a special regression model with random projections, first studied by (Breiman & Freedman, 1983) in the underparametrized regime, and later extended by (Belkin et al., 2020) to the overparametrized regime. Despite its simplicity, the test risk displays the double descent shape, calculated explicitly in (Belkin et al., 2020) for the least-squares estimator.[2] The model is a suitable laboratory to apply our formalism and compute the whole time-dependence of the test risk for SGF, and compare with the GF solution.

[2]Here the least-squares estimator equals the GF estimator for infinite times.

Again, the data set is composed of n pairs $(\boldsymbol{x}^k, y^k) \in \mathbb{R}^{d+1}$, $k \in [n] \equiv \{1, \dots, n\}$ sampled i.i.d. from $\mathbb{P}(\boldsymbol{x}, y)$. The distribution generating the data features is Gaussian, $\mathbb{P}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{0}, \mathbb{I}_d)$, while $\mathbb{P}(y|\boldsymbol{x})$ is modeled by a linear function of $\boldsymbol{\beta} \in \mathbb{R}^d$, $\|\boldsymbol{\beta}\| = 1$, $y^k = \boldsymbol{\beta}^\top \boldsymbol{x}^k + \mu \epsilon^k$ with $k \in [n]$. The constant $\mu > 0$ controls the strength of the additive Gaussian noise $\epsilon^k \sim \mathcal{N}(0,1)$. The feature matrix is denoted $\boldsymbol{X} = [\boldsymbol{x}^1 | \dots | \boldsymbol{x}^n]^\top \in \mathbb{R}^{n \times d}$.

The weights $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ are learned using only a subset $\mathcal{A} \subseteq [d] \equiv \{1, \dots, d\}$ of $p \equiv |\mathcal{A}|$ components. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, we denote $\boldsymbol{v}_{\mathcal{A}} \equiv (v_j : j \in \mathcal{A}) \in \mathbb{R}^p$ its subvector of entries from $\mathcal{A}$; besides, we denote $\mathcal{A}^c \equiv [d] \backslash \mathcal{A}$ to be the complement of $\mathcal{A}$. Then, the training loss is defined as $\mathscr{L}(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}) = \frac{1}{n} \sum_{k=1}^n l(\hat{\boldsymbol{\beta}}; \boldsymbol{x}^k, y^k; \mathcal{A}) = \frac{1}{n} \sum_{k=1}^n \frac{1}{2} \left(y^k - \hat{\boldsymbol{\beta}}_{\mathcal{A}}^\top \boldsymbol{x}_{\mathcal{A}}^k \right)^2$. Only components $\hat{\boldsymbol{\beta}}_{\mathcal{A}} \in \mathbb{R}^p$ are learned, while $\hat{\boldsymbol{\beta}}_{\mathcal{A}^c} \in \mathbb{R}^{d-p}$ are set to zero. Besides, we denote $\boldsymbol{X}_{\mathcal{A}} = \left[\boldsymbol{x}_{\mathcal{A}}^1 | \dots | \boldsymbol{x}_{\mathcal{A}}^n \right]^\top \in \mathbb{R}^{n \times p}$ as the matrix of features used for regression.

The SGD process associated with training of the regression leads to the SDE modeling of Eq. (4). From (5) we approximate the diffusion matrix with (see Appendix C.2):

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}) \approx \frac{2}{n} \mathscr{L}(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}) \boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{A}}, \tag{22}$$

which results in the SGF:

$$\mathrm{d}\hat{\boldsymbol{\beta}}_{\mathcal{A}} = \frac{1}{n} \boldsymbol{X}_{\mathcal{A}}^\top (\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}} \hat{\boldsymbol{\beta}}_{\mathcal{A}}) \mathrm{d}t + \frac{\sqrt{\gamma}}{n} \|\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}} \hat{\boldsymbol{\beta}}_{\mathcal{A}}\| \boldsymbol{X}_{\mathcal{A}}^\top \mathrm{d}\boldsymbol{b}. \tag{23}$$

For $\gamma = 0$ this becomes the respective pure GF equation. Setting $t_0 = 0$ and assuming the initial condition $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) = \hat{\boldsymbol{\beta}}_{\mathcal{A}}^0$, we obtain in Appendix C.3 the exact GF trajectory:

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t) = e^{-\frac{1}{n} \boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{A}} t} \hat{\boldsymbol{\beta}}_{\mathcal{A}}^0 + \boldsymbol{X}_{\mathcal{A}}^\dagger \left(\mathbb{I}_n - e^{-\frac{1}{n} \boldsymbol{X}_{\mathcal{A}} \boldsymbol{X}_{\mathcal{A}}^\top t} \right) \boldsymbol{y} \,, \tag{24}$$

where $\dagger$ denotes Moore-Penrose inverse. Note that this tends to the usual least-squares estimator for $t \to +\infty$. Translating to the general setting of Section 3, the solution of SGF will have the form $\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{SGF}}(t) = \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t) + \sqrt{\gamma} \boldsymbol{z}(t)$, with $\boldsymbol{z}(0) = \boldsymbol{0}$ and $\boldsymbol{z}(t)$ being a zero mean Gaussian process with the covariance matrix $\boldsymbol{\mathscr{C}}(t)$.

4.2. Test risk of GF and SGF

We are particularly interested in the test risk provided by GF and SGF and how they compare. Let $\mathcal{A}$ be a uniformly random subset of $[d]$ of cardinality p. If $(\boldsymbol{x}, y)$ is a new sample from distribution $\mathbb{P}(\boldsymbol{x}, y)$, the test risk on it is defined

as:

$$\mathcal{E}_{\text{test}}^{\text{SGF}}(t) = \frac{1}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{b};\boldsymbol{x},y}\left[\left(y - \boldsymbol{x}_{\mathcal{A}}^{\top}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{SGF}}(t)\right)^2\right]$$
$$= \frac{1}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{b}}\left[\left\|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{SGF}}(t)\right\|^2 + \|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2\right] + \mu^2 \ . \tag{25}$$

As $\mathcal{E}_{\text{test}}^{\text{SGF}}(t)\big|_{\gamma=0} = \mathcal{E}_{\text{test}}^{\text{GF}}(t)$ and, for any fixed training data and choice of features, $\mathbb{E}_{\boldsymbol{b}}\left[\boldsymbol{z}(t)\right] = \mathbf{0}$, we get

$$\mathcal{E}_{\text{test}}^{\text{SGF}}(t) = \mathcal{E}_{\text{test}}^{\text{GF}}(t) + \frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{b}}\left[\boldsymbol{z}^{\top}(t)\boldsymbol{z}(t)\right]$$
$$= \mathcal{E}_{\text{test}}^{\text{GF}}(t) + \frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\text{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] \ . \tag{26}$$

Therefore, inclusion of the perturbation has an additive effect on the test risk proportional to the learning rate γ.

Let $\boldsymbol{\Lambda}_r \in \mathbb{R}^{r\times r}$ with $r = \min(n,p)$ be a diagonal matrix consisting of the r highest singular values of $\boldsymbol{X}_{\mathcal{A}}$; with probability 1 they are all strictly positive. Besides, all other singular values of $\boldsymbol{X}_{\mathcal{A}}$ are zero. From the GF solution, the test risk can be expressed as (see Appendix C.4 for a detailed derivation and Appendix E for plots of the GF test risk):

$$2\,\mathcal{E}_{\text{test}}^{\text{GF}}(t) = \frac{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2}{d}\left[\max(0, p-n) + \mathbb{E}_{\boldsymbol{\Lambda}_r}\text{Tr}\, e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}t}\right]$$
$$+ \left(\left(1 - \frac{p}{d}\right)\|\boldsymbol{\beta}\|^2 + \mu^2\right)\left[1 + \mathbb{E}_{\boldsymbol{\Lambda}_r}\text{Tr}\,\boldsymbol{\Lambda}_r^{-2}\left(\mathbb{I}_r - e^{-\frac{\boldsymbol{\Lambda}_r^2}{n}t}\right)^2\right] \tag{27}$$

When $t \to \infty$, this expression becomes exactly the one from (Belkin et al., 2020) (deduced directly form least-squares and can also be computed for finite n, p, d).

For the second term in (26), giving the correction to the test error coming from the stochastic term, we obtain:

$$\frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\text{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] = \frac{\gamma}{2}\cdot\frac{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2}{d}\cdot$$
$$\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{\Lambda}_r}\left[\text{Tr}\left\{\frac{\boldsymbol{\Lambda}_r^2}{n}e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}(t-\tau)}\right\}\text{Tr}\left\{\frac{\boldsymbol{\Lambda}_r^2}{n}e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}\tau}\right\}\right]$$
$$+ \frac{\gamma}{2}\left(\left(1 - \frac{p}{d}\right)\|\boldsymbol{\beta}\|^2 + \mu^2\right)\cdot$$
$$\cdot\frac{1}{n}\left(\max\left(0, \frac{n-p}{2}\right)\cdot\mathbb{E}_{\boldsymbol{\Lambda}_r}\left[\text{Tr}\left\{\mathbb{I}_r - e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}t}\right\}\right]\right.$$
$$\left. + \int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{\Lambda}_r}\left[\text{Tr}\left\{\frac{\boldsymbol{\Lambda}_r^2}{n}e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}(t-\tau)}\right\}\text{Tr}\left\{e^{-2\frac{\boldsymbol{\Lambda}_r^2}{n}\tau}\right\}\right]\right) \ . \tag{28}$$

In the remaining part of this paragraph, we discuss the *asymptotic regime* $n, d, p \to \infty$ with $p/n \to \alpha$, $d/n \to \psi$ with α and ψ fixed positive numerical constants. Note that $p \le d$ hence $\alpha \le \psi$. At the same time, in order to obtain a well defined limit for (28), one should rescale the learning rate in SGF as follows: $\gamma = \gamma'/d$ with γ' fixed. Indeed, in (6) the order of magnitude of the diffusion term is $\mathbb{E}\left\|\sqrt{\gamma}\,\mathrm{d}\boldsymbol{\eta}(\tau)\right\|^2 = \gamma n \mathrm{d}\tau$ and the scaling ensures that it is of the same order $\mathrm{d}\tau$ as the drift. This sort of scaling has been discussed in (Veiga et al., 2022; Ben Arous et al., 2022). We observe that the diffusion contribution vanishes if γ goes to zero faster than $1/d$.

With our normalizations the empirical distribution $n^{-1}\#\{\text{eigenvalues of } \boldsymbol{\Lambda}_r^2/n \in [a,b]\}$ tends weakly to the (a.c part of) Marchenko-Pastur law

$$\rho_\alpha([a,b]) = \int_a^b \mathrm{d}\sigma\,\frac{\sqrt{(\alpha_+ - \sigma)(\sigma - \alpha_-)}}{2\pi\alpha\sigma}\mathbb{1}(\sigma \in [\alpha_-, \alpha_+])$$

with $\alpha_\pm = (1 \pm \sqrt{\alpha})^2$; from which we can express the expectations of traces in (27) in the limit $n, d, p \to \infty$ (noting that $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2 \to 2$ for two random vectors on the high dimensional unit sphere):

$$\lim \mathcal{E}_{\text{test}}^{\text{GF}}(t) = \max\left(0, \frac{\alpha - 1}{\psi}\right) + \frac{\alpha}{2\psi}\int \rho_\alpha(\mathrm{d}\sigma)e^{-2\sigma t}$$
$$+ \frac{1}{2}\left(\left(1 - \frac{\alpha}{\psi}\right) + \mu^2\right)\left[1 + \alpha\int\rho_\alpha(\mathrm{d}\sigma)\frac{(1 - e^{-\sigma t})^2}{\sigma}\right] . \tag{29}$$

Using the concentration of expressions $n^{-1}\,\text{Tr}\{\varphi(\boldsymbol{\Lambda}_r^2/n)\}$ for any 'reasonable' function φ, we can compute the limit for $n, p, d \to \infty$ using again the Marchenko-Pastur distribution. With the scaling $\gamma = \gamma'/d$, we get the well-defined limit (again note $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2 \to 2$ for two random vectors on the high dimensional unit sphere):

$$\lim \frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\text{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] =$$
$$= \frac{\gamma'\alpha}{2\psi}\left[\frac{2\alpha}{\psi}F_1(\alpha,t) + \left(1 - \frac{\alpha}{\psi} + \mu^2\right)\cdot\right.$$
$$\left.\cdot\left(\alpha\, F_2(\alpha,t) + \max(1-\alpha, 0)\int\rho_\alpha(\mathrm{d}\sigma)\,\frac{1 - e^{-2\sigma t}}{2}\right)\right] , \tag{30}$$

where

$$F_1(\alpha,t) = \iint \rho_\alpha(\mathrm{d}\sigma_1)\rho_\alpha(\mathrm{d}\sigma_2)\sigma_1\sigma_2 K(t,\sigma_1,\sigma_2) \ ,$$
$$F_2(\alpha,t) = \frac{1}{2}\iint \rho_\alpha(\mathrm{d}\sigma_1)\rho_\alpha(\mathrm{d}\sigma_2)(\sigma_1 + \sigma_2)K(t,\sigma_1,\sigma_2) \ ,$$

and

$$K(t,\sigma_1,\sigma_2) = \frac{e^{-2\sigma_1 t} - e^{-2\sigma_2 t}}{2(\sigma_2 - \sigma_1)} \ .$$

4.3. Discussion and comparison with simulations

Eqs. (29) and (30) give the complete time dependence of the test risk (or average generalization error) under SGF

dynamics for small learning rate. To interpret these formulas correctly, one should recall that $\alpha \in [0, \psi]$ and $\psi \geq 0$.

The dominant contribution is given by (29). The $t \to +\infty$ limit of this expression agrees with the well known double descent curve obtained by the least-squares estimator. Such finite time dependencies have been obtained and discussed in detail in (Bodin & Macris, 2021; 2022) for the random features and Gaussian covariate models.

Here we concentrate on the correction term which is our main interest, i.e., $\mathcal{E}_{\text{test}}^{\text{SGF}}(t) - \mathcal{E}_{\text{test}}^{\text{GF}}(t)$ or Eq. (30). It can be computed for $t \to +\infty$ after the large size limit $n, p, d \to \infty$, and we find:

$$\frac{\gamma'}{4}\frac{\alpha}{\psi}\left(\left(1-\frac{\alpha}{\psi}\right)+\mu^2\right)\max(1-\alpha, 0), \quad \alpha \leq \psi\,. \quad (32)$$

For $\alpha \leq 1$ this is a cubic polynomial (in α) with roots at $\alpha_1 = 0, \alpha_2 = 1$ and $\alpha_3 = (1+\mu^2)\psi$. The last root is always 'unphysical' since $\alpha_3 > \psi$. The root $\alpha_2 = 1$ (the interpolation threshold) is present only for $\psi \geq 1$. For infinitely large times the correction to GF brought about by the stochastic term in SGF is present only in the underparametrized regime $\alpha < 1$. In the overparametrized regime $\alpha > 1$ (hence also $\psi > 1$) there is no such correction for infinite time. This result is compared with numerical simulations of the discrete time stochastic gradient descent in Fig. 2. The agreement is satisfying given that (i) the continuous time dynamics is a heuristic approximation of stochastic gradient descent and, (ii) the analysis of the stochastic process uses a Laplace approximation of the path integral.

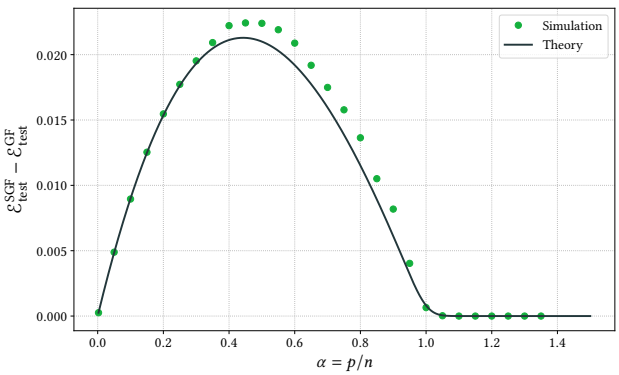


Figure 2. Continuous curve: difference between SGF and GF test risks for $t \to +\infty$ according to Eq. (32). Dots: difference of SGD and GD test risks obtained from numerical simulations (averaged over 1000 different random subsets $\mathcal{A}$). Simulation parameters: $d = 1000$, $n = 400$, such that $\psi = 2.5$, and $\gamma = 10^{-3}$ (hence $\gamma' = 1$). Vectors $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ are taken at random on the unit 1000-dimensional sphere and here $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 \approx 2.11$.

Figs. 3 and 4 show various aspects of the finite time behavior of $\mathcal{E}_{\text{test}}^{\text{SGF}}(t) - \mathcal{E}_{\text{test}}^{\text{GF}}(t)$. In Fig. 3 we observe that (for $\psi = 2.5$) the maximum difference between the test risk curves occurs above the interpolation threshold for 'early' times and this maximum moves below the interpolation threshold at 'later' times. For intermediate times, however, our analytical theory underestimates the difference in test risks (see Appendix E for more details).

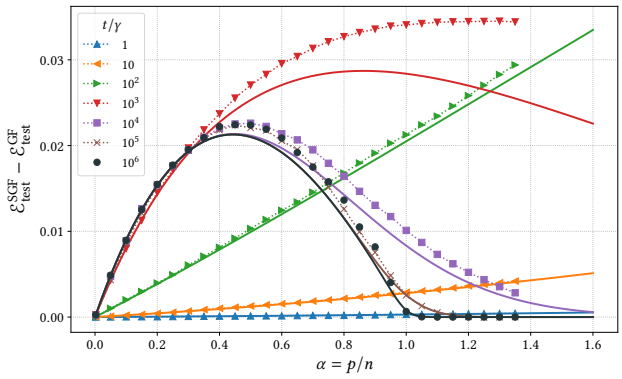


Figure 3. Continuous curves: difference between test risk of SGF and GF for various times $t \in [10^{-3}, 10^3]$ according to (30). Dots: difference of test risk of SGD and GD obtained from numerical simulations with same parameters as in Fig 2. Asymptotic analytical theory underestimates SGD quantities for intermediate times $1 \lessapprox t \lessapprox 10$ but works well for small and large times.

In Fig. 4 the difference in test risks of GF and SGF shows a knee structure. This bending occurs roughly at times corresponding to the onset of double descent.

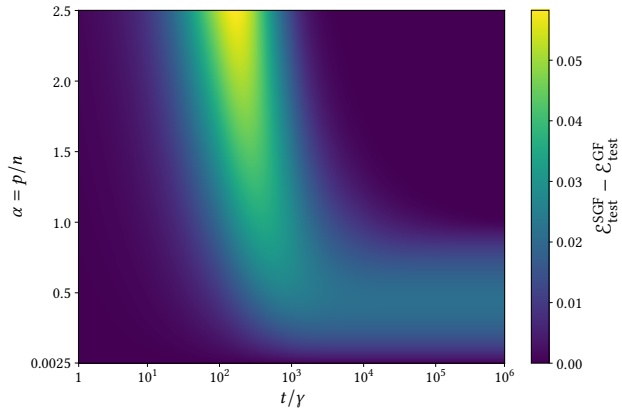


Figure 4. Theoretical prediction for the difference between test risk of SGF and GF in the (t, α) plane. Here we fix $\psi = 2.5$.

5. Conclusion

The general framework developed here gives concrete means to compute the whole-time evolution of the generalization error under stochastic gradient flow dynamics.

In particular, we provide general formulas, Eqs. (21), for the covariance of the fluctuations around the deterministic pure gradient flow.

In this work the formalism is applied only to a situation where the dynamics is very simple. It will be interesting to investigate to what extent one can apply it beyond the weak features model for more complicated landscapes with many minima and saddles. Random features with non-linear activation, diagonal linear networks, generalized linear or multi-index models might still be tractable. More broadly, it is of interest to further develop the theory extracting general properties from the covariance formula by leveraging statistical analysis of the data matrix and of the Hessian. Recent activity in the literature connects the eigenvalues of the Hessian to the optimization landscape, see e.g., (Gur-Ari et al., 2018; Sagun et al., 2018; Xie et al., 2022; Sabanayagam et al., 2023).

There exist more sophisticated SGFs yielding better approximations of SGD (Li et al., 2019). The analysis is then more complicated as there are new correction terms in the path integral. In the same vein, it would be desirable to investigate the use of alternative processes, other than Brownian motion, to model the SGD fluctuations.

We plan to come back to these issues in forthcoming works.

Acknowledgements

We acknowledge funding from EPFL and from the Swiss National Science Foundation grant number 200021-204119.

Impact Statement

Broadly speaking, theoretical results on continuous time models of stochastic gradient flows may lead to better guidelines for using practical gradient descent methods in Machine Learning. This may in particular help design better and less wasteful learning systems.

References

Ali, A., Dobriban, E., and Tibshirani, R. The implicit regularization of stochastic gradient flow for least squares. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 233–244. PMLR, 13–18 Jul 2020.

Arnaboldi, L., Krzakala, F., Loureiro, B., and Stephan, L. Escaping mediocrity: how two-layer networks learn hard generalized linear models. In *OPT 2023: Optimization for Machine Learning*, 2023.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021. doi: 10.1017/S0962492921000039.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. doi: 10.1137/20M1336072.

Ben Arous, G., Gheissari, R., and Jagannath, A. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25349–25362, 2022.

Bodin, A. and Macris, N. Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21605–21617, 2021.

Bodin, A. and Macris, N. Gradient flow in the Gaussian covariate model: exact solution of learning curves and multiple descent structures, 2022.

Breiman, L. and Freedman, D. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983. ISSN 01621459.

de Pirey, T. A., Cugliandolo, L. F., Lecomte, V., and van Wijland, F. Path integrals and stochastic calculus. *Advances in Physics*, 0(0):1–85, 2023. doi: 10.1080/00018732.2023.2199229.

Dembo, A. and Zeitouni, O. *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, 2010. ISBN 9783642033117. doi: 10.1007/978-3-642-03311-7.

Dyson, F. J. The s matrix in quantum electrodynamics. *Phys. Rev.*, 75:1736–1755, Jun 1949. doi: 10.1103/PhysRev.75.1736.

Evans, L. C. *An introduction to stochastic differential equations*, volume 82. American Mathematical Soc., 2012.

Fan, J., Gong, W., Li, C. J., and Sun, Q. Statistical sparse online regression: A diffusion approximation perspective. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1017–1026. PMLR, 09–11 Apr 2018.

Fetter, A. L. and Walecka, J. D. *Theoretical Mechanics of Particles and Continua*. Dover Books on Physics. Dover Publications, Mineola, NY, dec 2003. ISBN 978-0-486-43261-8.

Freidlin, M. I. and Wentzell, A. D. *Random Perturbations of Dynamical Systems*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-25847-3.

Gerbelot, C., Troiani, E., Mignacco, F., Krzakala, F., and Zdeborová, L. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024. doi: 10.1137/23M1594388.

Gur-Ari, G., Roberts, D. A., and Dyer, E. Gradient descent happens in a tiny subspace, 2018.

Hildebrandt, F. and Rœlly, S. Pinned diffusions and Markov bridges. *Journal of Theoretical Probability*, 33(2):906–917, Jun 2020. ISSN 1572-9230. doi: 10.1007/s10959-019-00954-5.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pp. 392–402, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01424-7.

Kushner, H. J. and Yin, G. G. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer-Verlag, 2 edition, 2003. doi: 10.1007/b97441.

Li, Q., Tai, C., and E, W. Dynamics of stochastic gradient algorithms, 2015.

Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110. PMLR, 06–11 Aug 2017.

Li, Q., Tai, C., and E, W. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

Mandt, S. and Blei, D. M. Continuous-time limit of stochastic gradient descent revisited, 2015.

Mandt, S., Hoffman, M., and Blei, D. A variational analysis of stochastic gradient algorithms. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 354–363, New York, New York, USA, 20–22 Jun 2016. PMLR.

Mandt, S., t, Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Mignacco, F., Krzakala, F., Urbani, P., and Zdeborová, L. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9540–9550. Curran Associates, Inc., 2020.

Mignacco, F., Urbani, P., and Zdeborová, L. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*, 2(3):035029, jul 2021. doi: 10.1088/2632-2153/ac0615.

Mori, T., Ziyin, L., Liu, K., and Ueda, M. Power-law escape rate of SGD. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15959–15975. PMLR, 17–23 Jul 2022.

Paquette, C. and Paquette, E. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9229–9240. Curran Associates, Inc., 2021.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Implicit regularization or implicit conditioning? Exact risk trajectories of SGD in high dimensions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35984–35999, 2022a.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties, 2022b.

Pesme, S. and Flammarion, N. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7475–7505. Curran Associates, Inc., 2023.

Pesme, S., Pillaud-Vivien, L., and Flammarion, N. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230, 2021.

Sabanayagam, M., Behrens, F., Adomaityte, U., and Dawid, A. Unveiling the Hessian's connection to the decision boundary. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Sagun, L., Evci, U., Güney, V. U., Dauphin, Y. N., and Bottou, L. Empirical analysis of the Hessian of over-parametrized neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

Sakurai, J. J. and Napolitano, J. *Modern Quantum Mechanics*. Cambridge University Press, 3 edition, 2020. doi: 10.1017/9781108587280.

Veiga, R., Stephan, L., Loureiro, B., Krzakala, F., and Zdeborová, L. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23244–23255, 2022.

Wang, Y. and Wu, S. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *Journal of Machine Learning Research*, 21(199):1–103, 2020.

Warming, R. and Hyett, B. The modified equation approach to the stability and accuracy analysis of finite-difference methods. *Journal of Computational Physics*, 14(2):159–179, 1974. ISSN 0021-9991. doi: https://doi.org/10.1016/0021-9991(74)90011-4.

Weinberg, S. *The Quantum Theory of Fields*, volume 1. Cambridge University Press, 1995. doi: 10.1017/CBO9781139644167.

Xie, Z., Tang, Q.-Y., Cai, Y., Sun, M., and Li, P. On the power-law Hessian spectrums in deep learning, 2022.

A. General Path Integral Formulation

In this appendix, we provide the details on the general path integral framework and on the derivation of the covariance matrix. For easy readability and to make this section as self-contained as possible, some equations from the main part are rewritten here.

Consider the following general stochastic differential equation (SDE) for a process $\boldsymbol{w}(\tau) \in \mathbb{R}^d$, sampled between t_0 and $t > t_0$:

$$\mathrm{d}\boldsymbol{w}(\tau) = \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right) \mathrm{d}\tau + \sqrt{\gamma}\, \boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right) \mathrm{d}\boldsymbol{\eta}(\tau) \ , \tag{33}$$

where $\boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right) \in \mathbb{R}^d$ and $\boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right) \in \mathbb{R}^{d\times n}$ are the drift vector and the diffusion matrix, respectively; while $\boldsymbol{\eta}(\tau)$ is a standard n-dimensional Wiener process, with $\mathrm{d}\boldsymbol{\eta}(\tau) = \boldsymbol{\eta}(\tau + \mathrm{d}\tau) - \boldsymbol{\eta}(t) \in \mathbb{R}^n$ as its forward increment. The initial and terminal conditions are $\boldsymbol{w}(t_0) = \boldsymbol{w}_0$ and $\boldsymbol{w}(t) = \boldsymbol{w}$, respectively. We assume that this SDE has a unique solution. Standard conditions ensuring existence and uniqueness are Lipshitzness and linear growth of $\boldsymbol{f}$ and $\boldsymbol{G}$ w.r.t $\boldsymbol{w}$ uniformly in τ (see for example Evans 2012, Section 5.B.3).

The continuous-time equation Eq. (33) must be understood according to its discrete-time companion process $\boldsymbol{w}_k = \boldsymbol{w}(k\Delta\tau)$:

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \boldsymbol{f}(k, \tilde{\boldsymbol{w}}_k)\Delta\tau + \sqrt{\gamma}\, \boldsymbol{G}(k, \tilde{\boldsymbol{w}}_k)\Delta\boldsymbol{\eta}_k \ , \tag{34}$$

where the time interval $[t_0, t]$ has been discretized into N slices of length $\Delta\tau \equiv \frac{t-t_0}{N}$ with $\tau_k = t_0 + k\Delta\tau$ for $k = 0, \dots, N$ and $\tau_N = t$. The drift and the diffusion functions are evaluated at the point

$$\tilde{\boldsymbol{w}}_k = \boldsymbol{w}_k + \alpha\Delta\boldsymbol{w}_k \ , \tag{35}$$

with $0 \leq \alpha \leq 1$ and $\Delta\boldsymbol{w}_k \equiv \boldsymbol{w}_{k+1} - \boldsymbol{w}_k$. The choice of α fully determines the discretization scheme and $\alpha \neq 0$ introduces additional contributions to the drift on the path-integral (de Pirey et al., 2023). Such contributions are also related to different SDEs as studied in (Li et al., 2019). In the present work, we adopt the simplest Itô discretization, $\alpha = 0$, $\tilde{\boldsymbol{w}}_k = \boldsymbol{w}_k$. The discrete quantities $\Delta\boldsymbol{\eta}_k$ are sampled independently at each step k from a Gaussian with $\mathbb{E}[\Delta\boldsymbol{\eta}_k] = \boldsymbol{0}$, $\mathbb{E}[\Delta\boldsymbol{\eta}_k(\Delta\boldsymbol{\eta}_k)^\top] = \Delta\tau\, \mathbb{I}_n$.

The matrix $\boldsymbol{G}(k, \boldsymbol{w}_k)\boldsymbol{G}(k, \boldsymbol{w}_k)^\top$ is positive semi-definite, and in particular when $d > n$ it certainly has zero eigenvalues[3], which can make the process singular in sub-manifolds of $\mathbb{R}^d$. A regularization parameter $\epsilon > 0$ is then introduced, allowing us to write:

$$\Delta\boldsymbol{w}_k \equiv \boldsymbol{w}_{k+1} - \boldsymbol{w}_k \sim \mathcal{N}\left(\boldsymbol{f}_k\Delta\tau, \gamma\Delta\tau\left(\boldsymbol{G}_k\boldsymbol{G}_k^\top + \epsilon\mathbb{I}_d\right)\right) \ , \tag{36}$$

with $\boldsymbol{f}_k \equiv \boldsymbol{f}(k, \boldsymbol{w}_k)$ and $\boldsymbol{G}_k \equiv \boldsymbol{G}(k, \boldsymbol{w}_k)$.

Let $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0)$ be the probability of the process to be at $\boldsymbol{w}$ at time t given it starts from $\boldsymbol{w}_0$ at time t_0. And $\mathbb{P}_{\Delta\tau}(\boldsymbol{w}_{k+1}, \tau_{k+1}|\boldsymbol{w}_k, \tau_k)$ the one-step propagator associated with Eq. (33). Since the process is Markovian, the Chapman–Kolmogorov equation holds over the intermediate time windows $[t_k, t_{k+1}]$:

$$\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \lim_{\epsilon\to 0}\lim_{N\to\infty}\int\prod_{k'=1}^{N-1}\mathrm{d}\boldsymbol{w}_{k'}\prod_{k=0}^{N-1}\mathbb{P}_{\Delta\tau}(\boldsymbol{w}_{k+1}, \tau_{k+1}|\boldsymbol{w}_k, \tau_k) \ . \tag{37}$$

The transition probability over each intermediate time window $[\tau_k, \tau_{k+1}]$ is known from Eq.(36). Replacing in Eq.(37), we have:

$$\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \lim_{\epsilon\to 0}\lim_{N\to\infty}\int\prod_{k'=1}^{N-1}\frac{\mathrm{d}\boldsymbol{w}_{k'}}{\left(2\pi\gamma\Delta\tau\right)^{d/2}\left(\det\boldsymbol{\Omega}_k^\epsilon\right)^{1/2}}\exp\left[-\frac{\Delta\tau}{2\gamma}\sum_{k=0}^{N-1}\left(\frac{\Delta\boldsymbol{w}_k}{\Delta\tau} - \boldsymbol{f}_k\right)^\top(\boldsymbol{\Omega}_k^\epsilon)^{-1}\left(\frac{\Delta\boldsymbol{w}_k}{\Delta\tau} - \boldsymbol{f}_k\right)\right] \ , \tag{38}$$

where $\boldsymbol{\Omega}_k^\epsilon \equiv \boldsymbol{G}_k\boldsymbol{G}_k^\top + \epsilon\mathbb{I}_d$ is the regularized covariance.

In the continuous-time limit, the expression above becomes a *path integral* over all possible trajectories $\boldsymbol{w}(\tau) \in \mathbb{R}^d$ for the process to be in $\boldsymbol{w}$ at time t given the process starts in $\boldsymbol{w}_0$ at time t_0:

$$\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) = \lim_{\epsilon\to 0}\int_{\boldsymbol{w}(t_0)=\boldsymbol{w}_0}^{\boldsymbol{w}(t)=\boldsymbol{w}}\mathcal{D}\left[\boldsymbol{w}(\cdot)\right]\exp\left(-\frac{1}{\gamma}\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right]\right) \ . \tag{39}$$

[3] In the learning theory context, $d > n$ means overpametrization.

The *action* functional is defined as

$$\mathcal{S}^{\epsilon}\left[\boldsymbol{w}(\cdot)\right] \equiv \int_{t_0}^{t} \mathrm{d}\tau\ \mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right)\ , \tag{40}$$

where the continuous-time Lagrangian function is given by

$$\mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = \frac{1}{2}\left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\right)^{\top} \boldsymbol{\Omega}^{\epsilon}(\tau, \boldsymbol{w}(\tau))^{-1}\left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\right) \tag{41}$$

with $\boldsymbol{\Omega}^{\epsilon}(\tau, \boldsymbol{w}(\tau)) \equiv \boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right)\boldsymbol{G}\left(\tau, \boldsymbol{w}(\tau)\right)^{\top} + \epsilon\, \mathbb{I}_d$. The path measure is expressed as

$$\mathcal{D}\left[\boldsymbol{w}(\cdot)\right] \equiv \lim_{N\to\infty} \prod_{k'=1}^{N-1} \frac{\mathrm{d}\boldsymbol{w}_{k'}}{\left(2\pi\gamma\Delta\tau\right)^{d/2}\left(\det \boldsymbol{\Omega}_k^{\epsilon}\right)^{1/2}}\ , \tag{42}$$

and the integrals must be understood under the Itô discretization. For example

$$\int_{t_0}^{t} \mathrm{d}\tau \left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)\right) = \lim_{N\to\infty} \sum_{k=0}^{N-1} \Delta\tau \left(\frac{\Delta \boldsymbol{w}_k}{\Delta\tau} - \boldsymbol{f}\left(k, \boldsymbol{w}_k\right)\right)\ . \tag{43}$$

We observe that, by construction, $\boldsymbol{G}(\tau, \boldsymbol{w}(\tau))\boldsymbol{G}(\tau, \boldsymbol{w}(\tau))^{\top}$ is positive semi-definite, then the pseudo-inverse $\boldsymbol{\Omega}^{\epsilon}(\tau, \boldsymbol{w}(\tau))^{-1}$ is well-defined and also positive semi-definite. Therefore, for any point τ in the time-interval $[t_0, t]$:

$$\mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) \geq 0\ , \tag{44}$$

which allows us to write:

$$\mathcal{S}^{\epsilon}\left[\boldsymbol{w}(\cdot)\right] \geq 0\ . \tag{45}$$

Realizing that the action is non-negative will play an important role on identifying the high-probability paths in the limit $\gamma \to 0$.

A.1. Laplace approximation

If γ is small enough, the path integral in Eq. (39) is dominated by the paths minimizing the action $\mathcal{S}^{\epsilon}\left[\boldsymbol{w}(\cdot)\right]$. This is equivalent to the variational statement of classical mechanics known as the Hamilton's action principle (Fetter & Walecka, 2003), which requires the vanishing of the first order variation of the action functional,

$$\delta\mathcal{S}^{\epsilon}\left[\boldsymbol{w}(\cdot)\right] = \delta \int_{t_0}^{t} \mathrm{d}\tau\ \mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right)\ , \tag{46}$$

satisfying the *two boundary conditions* $\boldsymbol{w}(t_0) = \boldsymbol{w}_0$, $\boldsymbol{w}(t) = \boldsymbol{w}$. The path is a stationary point of the action if and only if it satisfies the Euler-Lagrange equations

$$\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\partial}{\partial \dot{w}_j} - \frac{\partial}{\partial w_j}\right)\mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = 0\ , \tag{47}$$

for $j = 1, \ldots, d$.

Writing $\mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right)$ explicitly as sum over the components,

$$\mathcal{L}^{\epsilon}\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = \frac{1}{2}\sum_{l,m=1}^{d}\left(\frac{\mathrm{d}w_l(\tau)}{\mathrm{d}\tau} - f_l(\tau, \boldsymbol{w}(\tau))\right)\left(\boldsymbol{\Omega}^{\epsilon}(\tau, \boldsymbol{w}(\tau))^{-1}\right)_{lm}\left(\frac{\mathrm{d}w_m(\tau)}{\mathrm{d}\tau} - f_m(\tau, \boldsymbol{w}(\tau))\right)\ . \tag{48}$$

and the Jacobian matrix of $\boldsymbol{f}$ along the trajectory $\boldsymbol{w}(\tau)$:

$$\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}(\tau)) \equiv \begin{bmatrix} \boldsymbol{\nabla} f_1(\tau, \boldsymbol{w}(\tau))^{\top} \\ \vdots \\ \boldsymbol{\nabla} f_d(\tau, \boldsymbol{x}(\tau))^{\top} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\tau, \boldsymbol{w}(\tau))}{\partial w_1} & \frac{\partial f_1(\tau, \boldsymbol{w}(\tau))}{\partial w_2} & \cdots & \frac{\partial f_1(\tau, \boldsymbol{w}(\tau))}{\partial w_d} \\ \frac{\partial f_2(\tau, \boldsymbol{w}(\tau))}{\partial w_1} & \frac{\partial f_2(\tau, \boldsymbol{w}(\tau))}{\partial w_2} & \cdots & \frac{\partial f_2(\tau, \boldsymbol{w}(\tau))}{\partial w_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_d(\tau, \boldsymbol{w}(\tau))}{\partial w_1} & \frac{\partial f_d(t, \boldsymbol{w}(\tau))}{\partial w_2} & \cdots & \frac{\partial f_d(\tau, \boldsymbol{w}(\tau))}{\partial w_d} \end{bmatrix} \in \mathbb{R}^{d\times d}\ , \tag{49}$$

the derivatives of the Lagrangian with respect to the components of $\boldsymbol{w}$ and $\dot{\boldsymbol{w}}$ are given by

$$\frac{\partial}{\partial \dot{w}_j}\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = \sum_{l=1}^{d}\left(\frac{\mathrm{d}w_l(\tau)}{\mathrm{d}\tau} - f_l(\tau, \boldsymbol{w}(\tau))\right)\left(\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\right)_{lj} \ , \tag{50a}$$

$$\begin{aligned}\frac{\partial}{\partial w_j}\mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}(\tau), \dot{\boldsymbol{w}}(\tau)\right) = &- \sum_{l,m=1}^{d}\left(\frac{\mathrm{d}w_l(\tau)}{\mathrm{d}\tau} - f_l(\tau, \boldsymbol{w}(\tau))\right)\left(\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\right)_{lm}\left(\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}(\tau))\right)_{mj} \\ &+ \frac{1}{2}\sum_{l,m=1}^{d}\left(\frac{\mathrm{d}w_l(\tau)}{\mathrm{d}\tau} - f_l(\tau, \boldsymbol{w}(\tau))\right)\left(\frac{\partial}{\partial w_j}\left(\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\right)_{lm}\right)\left(\frac{\mathrm{d}w_m(\tau)}{\mathrm{d}\tau} - f_m(\tau, \boldsymbol{w}(\tau))\right) \ ,\end{aligned} \tag{50b}$$

where we have used the fact that $\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))$ is symmetric.

Reverting to matrix notation, the set of Euler-Lagrange equations (47) can then be represented by:

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\tau}\left[\left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))\right)^\top \boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\right] \\ &+ \left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - f(\tau, \boldsymbol{w}(\tau))\right)^\top \boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}(\tau)) \\ &- \frac{1}{2}\left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))\right)^\top \frac{\partial \boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}}{\partial \boldsymbol{w}}\left(\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} - \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))\right)^\top = \boldsymbol{0} \ ,\end{aligned} \tag{51}$$

where $\frac{\partial \boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}}{\partial \boldsymbol{w}}$ is defined as the tensor with components

$$\left(\frac{\partial \boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}}{\partial \boldsymbol{w}}\right)_j^{lm} \equiv \frac{\partial}{\partial w_j}\left(\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}\right)_{lm} \ . \tag{52}$$

Similarly to the covariant formulation of classical electromagnetism and special relativity, we assumed that upper indices are the ones to be contracted on vector-tensor multiplication.

A.2. Dominant paths

Eqs. (51) constitute an intricate set of second order differential equations. Nevertheless, we recall that $\boldsymbol{w}^{\mathrm{ode}}(\tau)$ is assumed to be the unique global solution of the initial value problem:

$$\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} = \boldsymbol{f}(\tau, \boldsymbol{w}(\tau)) \quad , \quad \boldsymbol{w}^{\mathrm{ode}}(t_0) = \boldsymbol{w}_0 \ . \tag{53}$$

If the terminal condition $\boldsymbol{w}^{\mathrm{ode}}(t) = \boldsymbol{w}$ is also satisfied, then $\boldsymbol{w}^{\mathrm{ode}}(\cdot)$ is also a solution of the Euler-Lagrange equations for all τ in the interval $[t_0, t]$. Moreover, the action would vanish for this solution:

$$\mathcal{S}^\epsilon\left[\boldsymbol{w}^{\mathrm{ode}}(\cdot)\right] = \int_{t_0}^{t}\mathrm{d}\tau\ \mathcal{L}^\epsilon\left(\tau, \boldsymbol{w}^{\mathrm{ode}}(\tau), \dot{\boldsymbol{w}}^{\mathrm{ode}}(\tau)\right) = 0 \ . \tag{54}$$

From Eq. (45), we know that $\mathcal{S}^\epsilon\left[\boldsymbol{w}(\cdot)\right]$ is non-negative. Thus, for any solution of the initial value problem (53) with terminal condition $\boldsymbol{w}^{\mathrm{ode}}(t) \neq \boldsymbol{w}$, the action is greater than zero. Since the $\mathbb{P}_\gamma(\boldsymbol{w}, t|\boldsymbol{w}_0, t_0) \propto e^{-\frac{1}{\gamma}\mathcal{S}^\epsilon[\boldsymbol{w}(\cdot)]}$, paths $\boldsymbol{w}^{\mathrm{ode}}(\tau)$ not satisfying both initial $\boldsymbol{w}^{\mathrm{ode}}(t_0) = \boldsymbol{w}_0$ and terminal conditions $\boldsymbol{w}^{\mathrm{ode}}(t) = \boldsymbol{w}$ are exponentially less probable in the limit $\gamma \to 0$. Henceforth, we will refer to the paths $\boldsymbol{w}^{\mathrm{ode}}(\tau)$ satisfying both initial and terminal conditions,

$$\begin{cases}\boldsymbol{w}^{\mathrm{ode}}(t_0) &= \boldsymbol{w}_0 \ , \\ \boldsymbol{w}^{\mathrm{ode}}(t) &= \boldsymbol{w} \ ,\end{cases} \tag{55}$$

as the *dominant paths*.

A.3. Fluctuations around dominant paths

Consider now a small time-dependent stochastic perturbation $\boldsymbol{z}(\tau)$ around a dominant path such that the trajectory $\bar{\boldsymbol{w}}(\tau)$ satisfying Eq. (33) can be written as

$$\bar{\boldsymbol{w}}(\tau) = \boldsymbol{w}^{\text{ode}}(\tau) + \sqrt{\gamma}\boldsymbol{z}(\tau) \ , \tag{56}$$

where $\boldsymbol{w}^{\text{ode}}(\tau)$ is the high-probability path satisfying the initial and terminal conditions $\boldsymbol{w}^{\text{ode}}(t_0) = \boldsymbol{w}_0$ and $\boldsymbol{w}^{\text{ode}}(t) = \boldsymbol{w}$, respectively. The random perturbation $\boldsymbol{z}(\tau)$ is constrained to satisfy the boundary conditions $\boldsymbol{z}(t_0) = \boldsymbol{z}(t \to \infty) = \boldsymbol{0}$. For intermediate times, there is no constrain on the fluctuations.

In order to study the fluctuations effects on the probability measure, the key quantity is the covariance matrix:

$$\boldsymbol{\mathscr{C}}(\tau) \equiv \mathbb{E}\left[\boldsymbol{w}(\tau) - \mathbb{E}\left[\boldsymbol{w}(\tau)\right]\right] \mathbb{E}\left[\boldsymbol{w}(\tau) - \mathbb{E}\left[\boldsymbol{w}(\tau)\right]\right]^\top \ , \tag{57}$$

where $\mathbb{E}[\cdot]$ stands for the expectation over all sources of randomness. For $\boldsymbol{w}(\tau) = \bar{\boldsymbol{w}}(\tau)$, the only stochasticity is the perturbation and the covariance reduces to

$$\boldsymbol{\mathscr{C}}(\tau) = \gamma \ \left(\mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}(\tau)\boldsymbol{z}(\tau)^\top\right] - \mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}(\tau)\right] \mathbb{E}_{\boldsymbol{z}}\left[\boldsymbol{z}(\tau)\right]^\top\right) \ , \tag{58}$$

where $\mathbb{E}_{\boldsymbol{z}}[\cdot]$ stands for the expectation over the perturbation $\boldsymbol{z}(\cdot)$. Up to now, the probability distribution for $\boldsymbol{z}$ is unknown. Assuming $\gamma \to 0$, we will construct a measure for $\boldsymbol{z}$ in order to compute the covariance. We first Taylor expand the elements of $\boldsymbol{f}\left(\tau, \boldsymbol{w}(\tau)\right)$ and $\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}(\tau))^{-1}$ around $\boldsymbol{w}^{\text{ode}}(\tau)$:

$$f_l\left(\tau, \bar{\boldsymbol{w}}(\tau)\right) = f_l(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) + \sqrt{\gamma}\ \boldsymbol{z}(\tau)^\top \boldsymbol{\nabla} f_l(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) + \mathcal{O}(\gamma) \ , \tag{59a}$$

$$\left(\boldsymbol{\Omega}^\epsilon\left(\tau, \bar{\boldsymbol{w}}(\tau)\right)^{-1}\right)_{lm} = \left(\boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1}\right)_{lm} + \sqrt{\gamma}\ \boldsymbol{z}(\tau)^\top \boldsymbol{\nabla} \left(\boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1}\right)_{lm} + \mathcal{O}(\gamma) \ . \tag{59b}$$

for $l, m = 1, \ldots, d$. Additionally, since by assumption $\boldsymbol{w}^{\text{ode}}(\tau)$ is the solution of $\frac{\mathrm{d}\boldsymbol{w}(\tau)}{\mathrm{d}\tau} = \boldsymbol{f}(\tau, \boldsymbol{w}(\tau))$ satisfying the boundary conditions (55):

$$\frac{\mathrm{d}\bar{w}_l(\tau)}{\mathrm{d}\tau} - f_l(\tau, \boldsymbol{w}(\tau)) \approx \frac{\mathrm{d}z_l(\tau)}{\mathrm{d}\tau} - \boldsymbol{z}(\tau)^\top \boldsymbol{\nabla} f_l(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) \ . \tag{60}$$

Plugging it into the Lagrangian together with Eqs. (59) and retaining contributions up to order two on $\boldsymbol{z}(\cdot)$, we obtain:

$$\begin{aligned}
&\mathcal{L}^\epsilon\left(\tau, \bar{\boldsymbol{w}}(\tau), \dot{\bar{\boldsymbol{w}}}(\tau)\right) = \\
&= \frac{\gamma}{2} \sum_{l,m=1}^{d} \left(\frac{\mathrm{d}z_l(\tau)}{\mathrm{d}\tau} - \boldsymbol{z}(\tau)^\top \boldsymbol{\nabla} f_l(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\right) \left(\boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1}\right)_{lm} \left(\frac{\mathrm{d}z_m(\tau)}{\mathrm{d}\tau} - \boldsymbol{z}(\tau)^\top \boldsymbol{\nabla} f_m(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\right) \\
&= \frac{\gamma}{2} \left(\frac{\mathrm{d}\boldsymbol{z}(\tau)}{\mathrm{d}\tau} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\boldsymbol{z}(\tau)\right)^\top \boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1} \left(\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(t, \boldsymbol{w}^{\text{ode}}(t))\boldsymbol{z}(t)\right) \ ,
\end{aligned} \tag{61}$$

where in the second line we have used the Jacobian matrix defined in (49). Therefore, the measure governing the fluctuations $\boldsymbol{z}(t)$ in the limit $\gamma \to 0$ is given by

$$\bar{\mathbb{P}}\left(\boldsymbol{z}, t|\boldsymbol{0}, t_0\right) \propto \lim_{\epsilon \to 0} \int_{\boldsymbol{z}(t_0)=\boldsymbol{0}}^{\boldsymbol{z}(t)=\boldsymbol{z}} \bar{\mathcal{D}}\left[\boldsymbol{z}(\cdot)\right] \exp\left(-\int_{t_0}^{t} \mathrm{d}\tau\ \bar{\mathcal{L}}^\epsilon\left(\tau, \boldsymbol{z}(\tau), \dot{\boldsymbol{z}}(\tau)\right)\right) \ , \tag{62}$$

where

$$\bar{\mathcal{L}}^\epsilon\left(\tau, \boldsymbol{z}(\tau), \dot{\boldsymbol{z}}(\tau)\right) \equiv \frac{1}{2} \left(\frac{\mathrm{d}\boldsymbol{z}(\tau)}{\mathrm{d}\tau} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\boldsymbol{z}(\tau)\right)^\top \boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1} \left(\frac{\mathrm{d}\boldsymbol{z}(t)}{\mathrm{d}t} - \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(t, \boldsymbol{w}^{\text{ode}}(t))\boldsymbol{z}(t)\right) \ . \tag{63}$$

The diffusion matrix appearing on the path-measure,

$$\bar{\mathcal{D}}\left[\boldsymbol{z}(\cdot)\right] \equiv \lim_{N \to \infty} \prod_{k'=1}^{N-1} \frac{\Delta \boldsymbol{z}_{k'}}{\left(2\pi\Delta\tau\right)^{d/2} \left(\det \boldsymbol{\Omega}_k^{\epsilon,\text{ode}}\right)^{1/2}} \ , \tag{64}$$

is now calculated on the discretized dominant trajectory: $\boldsymbol{\Omega}_k^{\epsilon,\text{ode}} = \boldsymbol{G}(k, \boldsymbol{w}_k^{\text{ode}})\boldsymbol{G}(k, \boldsymbol{w}_k^{\text{ode}})^\top + \epsilon\, \mathbb{I}_d$.

For the purpose of simplifying the Lagrangian (63), we perform the change of variables:

$$\boldsymbol{z}(\tau) = \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \tilde{\boldsymbol{z}}(\tau)\ , \tag{65}$$

where the object inside the brackets is a matrix in $\mathbb{R}^{d\times d}$. The exponential of a matrix is, as usual, defined via Taylor series. The symbol $\mathcal{T}$ indicates time-ordering in the sense that the exponential is understood as a Dyson series (Weinberg, 1995). For the convenience of the reader, we review the time-ordered exponential for time dependent non-commutative matrices in Appendix B.

Under the change of variables, the time derivative of $\boldsymbol{z}(t)$ is written as

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{z}(\tau)}{\mathrm{d}\tau} &= \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\ \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \tilde{\boldsymbol{z}}(\tau) + \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau} \\ &= \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\boldsymbol{z}(\tau) + \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}\ ,\end{aligned} \tag{66}$$

so that the Lagrangian (63), in terms of $\tilde{\boldsymbol{z}}(\tau)$, is written as

$$\bar{\mathcal{L}}^{\epsilon}\left(\tau, \tilde{\boldsymbol{z}}(\tau), \dot{\tilde{\boldsymbol{z}}}(\tau)\right) \equiv \frac{1}{2}\frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}^{\top}\ \boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}\ . \tag{67a}$$

with

$$\boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) \equiv \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}^{\top} \boldsymbol{\Omega}^{\epsilon}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1}\ \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s\, \boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}\ . \tag{67b}$$

Let us mention in passing an analogy that can be made with techniques from quantum mechanics at this point. One could interpret (67b) as a kind of transformation from a mix of Schrödinger and Heisenberg pictures to the interaction picture, with $\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))$ and $\boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))$ playing the role of time-dependent Hamiltonian and potential, respectively (Sakurai & Napolitano, 2020).

With the intention of computing the expectations in (58), which are over the original coordinates $\boldsymbol{z}(\cdot)$, we consider the quantities $\mathbb{E}_{\tilde{\boldsymbol{z}}}\left[\tilde{\boldsymbol{z}}(\cdot)\right]$ and $\mathbb{E}_{\tilde{\boldsymbol{z}}}\left[\tilde{\boldsymbol{z}}(\cdot)\tilde{\boldsymbol{z}}(\cdot)^{\top}\right]$, where $\mathbb{E}_{\tilde{\boldsymbol{z}}}$ indicates expectation over the measure $\propto e^{-\frac{1}{2}\int_{t_0}^{t} \mathrm{d}\tau \frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}^{\top} \boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) \frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}}$. At this point, it is convenient to return to the discretized form and to define a new discrete variable:

$$\tilde{\boldsymbol{z}}_k = \sum_{k'=0}^{k-1}\left(\tilde{\boldsymbol{z}}_{k'+1} - \tilde{\boldsymbol{z}}_{k'}\right) \equiv \sum_{k'=0}^{k-1} \boldsymbol{u}_{k'}\ , \tag{68}$$

as the assumption $\tilde{\boldsymbol{z}}(t_0) = \boldsymbol{0}$ implies that $\tilde{\boldsymbol{z}}(t_0) = \tilde{\boldsymbol{z}}_0 = \boldsymbol{0}$. Similarly, the discretized derivative in terms of the variable $\boldsymbol{u}_k$ is given by:

$$\frac{\Delta\tilde{\boldsymbol{z}}_k}{\Delta\tau} = \frac{\tilde{\boldsymbol{z}}_{k+1} - \tilde{\boldsymbol{z}}_k}{\Delta\tau} = \frac{\boldsymbol{u}_k}{\Delta\tau}\ . \tag{69}$$

The discretized measure then reads:

$$\begin{aligned}\exp\left[-\frac{1}{2}\int_{t_0}^{t} \mathrm{d}\tau \frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}^{\top}\ \boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))\frac{\mathrm{d}\tilde{\boldsymbol{z}}(\tau)}{\mathrm{d}\tau}\right] &\to \exp\left[-\frac{1}{2}\sum_{k=0}^{N-1}\Delta\tau\ \frac{\Delta\tilde{\boldsymbol{z}}_k}{\Delta\tau}^{\top}\ \boldsymbol{\Psi}^{\epsilon}(k, \boldsymbol{w}_k^{\text{ode}})\frac{\Delta\tilde{\boldsymbol{z}}_k}{\Delta\tau}\right] \\ &= \exp\left[-\frac{1}{2}\sum_{k=0}^{N-1}\frac{\Delta\tau}{(\Delta\tau)^2}\boldsymbol{u}_k^{\top}\boldsymbol{\Psi}^{\epsilon}(k, \boldsymbol{w}_k^{\text{ode}})\boldsymbol{u}_k\right] \\ &= \prod_{k=0}^{N-1}\exp\left[-\frac{1}{2\Delta\tau}\boldsymbol{u}_k^{\top}\boldsymbol{\Psi}^{\epsilon}(k, \boldsymbol{w}_k^{\text{ode}})\boldsymbol{u}_k\right]\ ,\end{aligned} \tag{70}$$

where $\boldsymbol{\Psi}^{\epsilon}(k, \boldsymbol{w}_k^{\text{ode}})$ means the discretization of $\boldsymbol{\Psi}^{\epsilon}(\tau, \boldsymbol{w}^{\text{ode}}(\tau))$ on the discretized dominant path. We then conclude that the first and second moments of $\boldsymbol{u}_k$ are given by:

$$\mathbb{E}\left[\boldsymbol{u}_k\right] = \boldsymbol{0}\ , \tag{71a}$$

$$\mathbb{E}\left[\boldsymbol{u}_k\boldsymbol{u}_k^{\top}\right] = \Delta\tau\ \boldsymbol{\Psi}^{\epsilon}(k, \boldsymbol{w}_k^{\text{ode}})^{-1}\ , \tag{71b}$$

respectively. From Eq. (71a), we immediately conclude that $\mathbb{E}\left[\tilde{\boldsymbol{z}}(\tau)\right] = \mathbb{E}\left[\boldsymbol{z}(\tau)\right] = \mathbf{0}$. With respect to the second moment, we observe that:

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{z}}_k \tilde{\boldsymbol{z}}_k^\top\right] &= \sum_{k',k''=0}^{k-1} \mathbb{E}\left[\boldsymbol{u}_{k'} \boldsymbol{u}_{k''}^\top\right] = \sum_{k'=0}^{k-1} \mathbb{E}\left[\boldsymbol{u}_{k'} \boldsymbol{u}_{k'}^\top\right] \\ &= \sum_{k'=0}^{k-1} \Delta\tau \; \boldsymbol{\Psi}^\epsilon(k, \boldsymbol{w}_k^{\text{ode}})^{-1} \; ,\end{aligned} \tag{72}$$

where we have used the Markovian property. In the continuous-time limit:

$$\mathbb{E}\left[\tilde{\boldsymbol{z}}(t)\tilde{\boldsymbol{z}}(t)^\top\right] = \int_{t_0}^{t} \mathrm{d}\tau \; \boldsymbol{\Psi}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau))^{-1} \; . \tag{73}$$

Returning to the original coordinates:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{z}(t)\boldsymbol{z}(t)^\top\right] &= \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \mathbb{E}\left[\tilde{\boldsymbol{z}}(t)\tilde{\boldsymbol{z}}(t)^\top\right] \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}^\top = \\ &= \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \int_{t_0}^{t} \mathrm{d}\tau \; \boldsymbol{\Psi}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau))^{-1} \; \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}^\top \; .\end{aligned} \tag{74}$$

Hitherto, the covariance matrix can then be expressed as

$$\boldsymbol{\mathscr{C}}(t) = \gamma \lim_{\epsilon \to 0} \; \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \int_{t_0}^{t} \mathrm{d}\tau \; \boldsymbol{\Psi}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau))^{-1} \; \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\} \; , \tag{75}$$

with

$$\boldsymbol{\Psi}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) = \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\} \boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^{-1} \; \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \; , \tag{76}$$

and

$$\boldsymbol{\Omega}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) = \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^\top + \epsilon \, \mathbb{I}_d \; . \tag{77}$$

We continue by explicitly inverting $\boldsymbol{\Psi}^\epsilon$,

$$\begin{aligned}&\boldsymbol{\Psi}^\epsilon(\tau, \boldsymbol{w}^{\text{ode}}(\tau))^{-1} = \\ &= \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}\right)^{-1} \boldsymbol{\Omega}^\epsilon\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \; \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\}\right)^{-1} \\ &= \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}\right)^{-1} \left(\boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^\top + \epsilon \, \mathbb{I}_d\right) \; \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\}\right)^{-1}\end{aligned} \tag{78}$$

and replacing it in the expression (75) for the covariance matrix, which leads to:

$$\begin{aligned}\boldsymbol{\mathscr{C}}(t) = \gamma \, \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \int_{t_0}^{t} \mathrm{d}\tau \; \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}\right)^{-1} \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^\top \cdot \\ \cdot \left(\mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\}\right)^{-1} \; \mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\}\end{aligned} \tag{79}$$

where we have already computed the limit $\epsilon \to 0$. Since, for any τ such that $t_0 < \tau < t$:

$$\mathcal{T}\left\{e^{\int_{t_0}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} = \mathcal{T}\left\{e^{\int_{\tau}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \cdot \mathcal{T}\left\{e^{\int_{t_0}^{\tau} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \; , \tag{80}$$

we have the final expression for the covariance matrix:

$$\boldsymbol{\mathscr{C}}(t) = \gamma \int_{t_0}^{t} \mathrm{d}\tau \; \mathcal{T}\left\{e^{\int_{\tau}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\} \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^\top \mathcal{T}\left\{e^{\int_{\tau}^{t} \mathrm{d}s \; \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))^\top}\right\} \; . \tag{81}$$

Setting $\boldsymbol{U}^{\text{ode}}(t, \tau) \equiv \mathcal{T}\left\{e^{\int_{\tau}^{t} \mathrm{d}s \boldsymbol{\mathcal{J}}_{\boldsymbol{f}}(s, \boldsymbol{w}^{\text{ode}}(s))}\right\}$ and observing that $\boldsymbol{\Omega}^0(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) = \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right) \boldsymbol{G}\left(\tau, \boldsymbol{w}^{\text{ode}}(\tau)\right)^\top$, we obtain:

$$\boldsymbol{\mathscr{C}}(t) = \gamma \int_{t_0}^{t} \mathrm{d}\tau \; \boldsymbol{U}^{\text{ode}}(t, \tau) \; \boldsymbol{\Omega}^0(\tau, \boldsymbol{w}^{\text{ode}}(\tau)) \; \boldsymbol{U}^{\text{ode}}(t, \tau)^\top \; , \tag{82}$$

which is exactly the same expression as in Eqs. (16) presented in the main text.

A.4. Special cases

A.4.1. Conservative drifts

If there exists a scalar function $\Phi(\tau, \boldsymbol{w}(\tau))$ such that the drift $\boldsymbol{f}(\tau, \boldsymbol{w}(\tau)) \in \mathbb{R}^d$ is given by:

$$\boldsymbol{f}(\tau, \boldsymbol{w}(\tau)) = -\boldsymbol{\nabla}\Phi(\tau, \boldsymbol{w}(\tau)) \ , \tag{83}$$

we say that the drift is *conservative*, in analogy with the conservative forces in classical mechanics (Fetter & Walecka, 2003). In this case, the Jacobian matrix of $\boldsymbol{f}$ is nothing but the negative of the Hessian of Φ:

$$\boldsymbol{\mathscr{J}}_{\boldsymbol{f}}(\tau, \boldsymbol{w}(\tau)) = -\boldsymbol{\mathcal{H}}_{\Phi}(\tau, \boldsymbol{w}(\tau)) \ , \tag{84}$$

where

$$\boldsymbol{\mathcal{H}}_{\Phi}(\tau, \boldsymbol{w}(\tau)) \equiv \boldsymbol{\nabla} \otimes \boldsymbol{\nabla}\Phi(\tau, \boldsymbol{w}(t)) = \begin{bmatrix} \frac{\partial\Phi(\tau,\boldsymbol{w}(\tau))}{\partial w_1^2} & \frac{\partial\Phi(\tau,\boldsymbol{w}(t))}{\partial w_1 \partial w_2} & \cdots & \frac{\partial\Phi(\tau,\boldsymbol{w}(\tau))}{\partial w_1 \partial w_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\Phi(\tau,\boldsymbol{w}(\tau))}{\partial w_d \partial w_1} & \frac{\partial\Phi(\tau,\boldsymbol{w}(\tau))}{\partial w_d \partial w_2} & \cdots & \frac{\partial\Phi(\tau,\boldsymbol{w}(\tau))}{\partial w_d^2} \end{bmatrix} \in \mathbb{R}^{d\times d} \ . \tag{85}$$

A.4.2. Cholesky-like diffusions

If the diffusion matrix $\boldsymbol{G}(\tau, \boldsymbol{w}(\tau)) \in \mathbb{R}^{d\times n}$ is such that there exists a matrix $\boldsymbol{\Xi}(\tau, \boldsymbol{w}(\tau)) \in \mathbb{R}^{d\times d}$ given by

$$\boldsymbol{\Xi}(\tau, \boldsymbol{w}(\tau)) = \boldsymbol{G}(\tau, \boldsymbol{w}(\tau))\boldsymbol{G}(\tau, \boldsymbol{w}(\tau))^{\top} \ , \tag{86}$$

we say that $\boldsymbol{G}(\tau, \boldsymbol{w}(\tau))$ is a Cholesky-like diffusion matrix. In this case, we use the convention $\boldsymbol{G}(\tau, \boldsymbol{w}(\tau)) \equiv \sqrt{\boldsymbol{\Xi}(\tau, \boldsymbol{w}(\tau))}$.

A.5. Machine learning realm

We consider now the particular case of the stochastic gradient flow modelling of SGD:

$$\mathrm{d}\hat{\boldsymbol{\beta}}(t) = -\boldsymbol{\nabla}\mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)\mathrm{d}t + \sqrt{\gamma\, \boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}\big)}\ \mathrm{d}\boldsymbol{b}(t) \ , \tag{87}$$

which is Eq. (4) in the main text (where all the quantities are defined). Clearly, the drift is conservative with the Jacobian equal to minus the Hessian of the loss. The diffusion matrix is also Cholesky-like. Additionally, if $\gamma = 0$, the SDE is reduced to the gradient flow equation, and this is why we term the solution of the ODE in this context as $\hat{\boldsymbol{\beta}}^{\mathrm{GF}}(\tau)$.

In summary, for SGF the covariance (82) reduces to

$$\boldsymbol{\mathscr{C}}^{\mathrm{SGF}}(t) = \gamma \int_{t_0}^{t} \mathrm{d}\tau\ \boldsymbol{U}^{\mathrm{GF}}(t,\tau)\ \boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}^{\mathrm{GF}}(\tau); \boldsymbol{X}, \boldsymbol{y}\big)\ \boldsymbol{U}^{\mathrm{GF}}(t,\tau)^{\top} \ , \tag{88a}$$

with

$$\boldsymbol{U}^{\mathrm{GF}}(t,\tau) \equiv \mathcal{T}\left\{ e^{-\int_{\tau}^{t} \mathrm{d}s\, \boldsymbol{\mathcal{H}}_{\mathscr{L}}(\hat{\boldsymbol{\beta}}^{\mathrm{GF}}(s); \boldsymbol{X}, \boldsymbol{y})} \right\} \ , \tag{88b}$$

which are equivalent to Eqs. (21) in the main text.

A.6. Discrete to continuous-time approximation guarantees

If the drift is conservative and the diffusion Cholesky-like, we have:

$$\mathrm{d}\boldsymbol{w}(\tau) = -\boldsymbol{\nabla}\Phi(\tau, \boldsymbol{w}(\tau))\mathrm{d}\tau + \sqrt{\gamma\ \boldsymbol{\Xi}(\tau, \boldsymbol{w}(\tau))}\ \mathrm{d}\boldsymbol{\eta}(\tau) \ . \tag{89}$$

From Corollary 10 in (Li et al., 2019):

$$\max_{k=0,\ldots,N} \left|\mathbb{E}\left[h\left(\boldsymbol{w}_k\right)\right] - \mathbb{E}\left[h\left(\boldsymbol{w}(k\Delta\tau)\right)\right]\right| \leq C\gamma \ , \tag{90}$$

for a suitable set of 'smooth' functions h with 'polynomial growth'.

Let us also mention that one may look at an 'improved' SGF modelling:

$$\mathrm{d}\boldsymbol{w}(\tau) = -\boldsymbol{\nabla}\left(\Phi(\tau,\boldsymbol{w}(\tau)) + \frac{1}{4}\gamma\|\boldsymbol{\nabla}\Phi(\tau,\boldsymbol{w}(\tau))\|^2\right)\mathrm{d}\tau + \sqrt{\gamma\,\boldsymbol{\Xi}(\tau,\boldsymbol{w}(\tau))}\,\mathrm{d}\boldsymbol{\eta}(\tau)\ , \tag{91}$$

where the convergence is improved to (Theorem 9 in (Li et al., 2019)):

$$\max_{k=0,\dots,N}|\mathbb{E}\left[h\left(\boldsymbol{w}_k\right)\right] - \mathbb{E}\left[h\left(\boldsymbol{w}(k\Delta\tau)\right)\right]| \leq C\gamma^2\ , \tag{92}$$

again for a suitable set of 'smooth' functions h with 'polynomial growth'.

B. Time-ordered exponential

The concept of time-ordering integrals was first introduced by Dyson (Dyson, 1949) in the context of time-dependent perturbation theory applied to scattering problems in quantum mechanics. Let $\boldsymbol{B}(t)$ and $\boldsymbol{A}(t)$ be two time-dependent square real matrices of the same dimension. Consider the following first order differential equation:

$$\frac{\mathrm{d}\boldsymbol{B}(t)}{\mathrm{d}t} = \boldsymbol{A}(t)\boldsymbol{B}(t) \tag{93}$$

with initial condition $\boldsymbol{B}(t_0) = \mathbb{I}$. In quantum-mechanics, both matrices could be complex with $\boldsymbol{B}(t)$ representing the time-evolution operator and $\boldsymbol{A}(t)$ a potential in the interaction picture.

Observe that Eq. (93) (with initial condition) can be reformulated as an integral equation:

$$\boldsymbol{B}(t) = \mathbb{I} - \int_{t_0}^{t}\mathrm{d}\tau\ \boldsymbol{A}(\tau)\boldsymbol{B}(\tau)\ . \tag{94}$$

Iterating the expression above:

$$\boldsymbol{B}(t) = \mathbb{I} - \int_{t_0}^{t}\mathrm{d}\tau_1\ \boldsymbol{A}(\tau_1) + \int_{t_0}^{t}\mathrm{d}\tau_1\int_{t_0}^{\tau_1}\mathrm{d}\tau_2\ \boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_2) - \int_{t_0}^{t}\mathrm{d}\tau_1\int_{t_0}^{\tau_1}\mathrm{d}\tau_2\int_{t_0}^{\tau_2}\mathrm{d}\tau_3\ \boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_3)\boldsymbol{A}(\tau_3) + \dots\ . \tag{95}$$

The *time-ordered product* $\mathcal{T}\{\cdot\}$ of any time-dependent matrix is defined as the product with factors arranged in such a way that the factor latest in time takes the leftmost position, the next-latest next to the leftmost, and so on. For example:

$$\mathcal{T}\{\boldsymbol{A}(\tau)\} = \boldsymbol{A}(\tau)\ , \tag{96a}$$

$$\mathcal{T}\{\boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_2)\} = \Theta(\tau_1-\tau_2)\boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_2) + \Theta(\tau_2-\tau_1)\boldsymbol{A}(\tau_2)\boldsymbol{A}(\tau_1)\ , \tag{96b}$$

where $\Theta(\tau)$ is the Heaviside step function equal to 1 for $\tau > 0$ and to zero for $\tau < 0$. This definition generalizes to the time ordered product of $\boldsymbol{A}$'s for 'n time instants' as a sum over all $n!$ permutations. Each permutation gives the same integral over $\tau_1, \tau_2, \dots, \tau_n$. Eq. (95) can then be written as

$$\boldsymbol{B}(t) = \mathbb{I}_d + \frac{(-1)^n}{n!}\sum_{n=1}^{\infty}\int_{t_0}^{t}\mathrm{d}\tau_1\int_{t_0}^{\tau_1}\mathrm{d}\tau_2\dots\int_{t_0}^{\tau_{n-1}}\mathrm{d}\tau_n\ \mathcal{T}\left\{\boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_2)\dots\boldsymbol{A}(\tau_n)\right\}\ , \tag{97}$$

which in the context of quantum mechanics is referred to as the *Dyson series*.

If the $\boldsymbol{A}(\tau)$'s would be commuting matrices for all τ, one could simply represent the series (97) by $e^{\int_{t_0}^{t}\mathrm{d}\tau\boldsymbol{A}(\tau)}$. Nevertheless, since commutation in time is not guaranteed in general, we introduce the notation for the *time-ordered exponential*:

$$\mathcal{T}\left\{e^{\int_{t_0}^{t}\mathrm{d}\tau\boldsymbol{A}(\tau)}\right\} \equiv \mathbb{I} + \frac{(-1)^n}{n!}\sum_{n=1}^{\infty}\int_{t_0}^{t}\mathrm{d}\tau_1\int_{t_0}^{\tau_1}\mathrm{d}\tau_2\dots\int_{t_0}^{\tau_{n-1}}\mathrm{d}\tau_n\ \mathcal{T}\left\{\boldsymbol{A}(\tau_1)\boldsymbol{A}(\tau_2)\dots\boldsymbol{A}(\tau_n)\right\}\ , \tag{98}$$

for any square time-dependent real matrix $\boldsymbol{A}(\tau)$.

C. Application to the weak features model: details of derivations

C.1. Preliminaries

For arbitrary $n_1, n_2 \in \mathbb{N}$, we use notation $\mathbf{0}_{n_1\times n_2}$ for all-zero $n_1 \times n_2$ matrix, $\mathbb{I}_{n_1\times n_2}$ for $n_1 \times n_2$ matrix with elements $(\mathbb{I}_{n_1\times n_2})_{ij} = \delta_{ij}$. $(\boldsymbol{A})^\dagger \in \mathbb{R}^{n_2\times n_1}$ denotes Moore-Penrose inverse of a matrix $\boldsymbol{A} \in \mathbb{R}^{n_1\times n_2}$. We use Unif to denote uniform (Haar) measure over a group.

We will find SVD's useful for our derivations. Suppose $r = \operatorname{rank} \boldsymbol{X}_\mathcal{A} \leq \min(n, p)$. Now, let

$$\boldsymbol{X}_\mathcal{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top,\ \boldsymbol{U} \in \mathbb{R}^{n\times n}, \boldsymbol{\Lambda} \in \mathbb{R}^{n\times p}, \boldsymbol{V} \in \mathbb{R}^{p\times p} \tag{99a}$$

be a 'full' singular value decomposition, and

$$\boldsymbol{X}_\mathcal{A} = \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^\top,\ \boldsymbol{U}_r \in \mathbb{R}^{n\times r}, \boldsymbol{\Lambda}_r \in \mathbb{R}^{r\times r}, \boldsymbol{V}_r \in \mathbb{R}^{p\times r} \tag{99b}$$

be 'truncated' the one. We also denote columns of $\boldsymbol{U}$ as $\{\boldsymbol{u}_i\}_{i=1}^n$, of $\boldsymbol{V}$ as $\{\boldsymbol{v}_i\}_{i=1}^p$ and positive singular values as $\{\lambda_i\}_{i=1}^r$. For convenience, we also utilize thenotation $\lambda_i = 0$ for $i > r$. Besides, we make use of $\boldsymbol{U}_{r^\perp} := [\boldsymbol{u}_{r+1} \mid \cdots \mid \boldsymbol{u}_n] \in \mathbb{R}^{n\times(n-r)}$ and $\boldsymbol{V}_{r^\perp} := [\boldsymbol{v}_{r+1} \mid \cdots \mid \boldsymbol{v}_p] \in \mathbb{R}^{p\times(p-r)}$ – i.e., $\boldsymbol{U}_{r^\perp}$ and $\boldsymbol{V}_{r^\perp}$ consist of columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively which are not included in $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$.

Let us state some useful properties. We keep in mind that $\boldsymbol{U}_r^\top\boldsymbol{U}_r = \boldsymbol{V}_r^\top\boldsymbol{V}_r = \mathbb{I}_r$, $\boldsymbol{U}_{r^\perp}^\top\boldsymbol{U}_{r^\perp} = \mathbb{I}_{n-r}$ and $\boldsymbol{V}_{r^\perp}^\top\boldsymbol{V}_{r^\perp} = \mathbb{I}_{p-r}$, and $\boldsymbol{U}_r^\top\boldsymbol{U}_{r^\perp} = \mathbf{0}_{r\times(n-r)}$ and $\boldsymbol{V}_r^\top\boldsymbol{V}_{r^\perp} = \mathbf{0}_{r\times(p-r)}$. Besides, the Moore-Penrose inverse of $\boldsymbol{X}_\mathcal{A}$ can be written as

$$(\boldsymbol{X}_\mathcal{A})^\dagger = \boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top\ ,. \tag{100}$$

Using properties of matrix exponentials,

$$e^{-\frac{1}{n}\boldsymbol{X}_\mathcal{A}\boldsymbol{X}_\mathcal{A}^\top t} = \boldsymbol{U}e^{-\frac{1}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top t}\boldsymbol{U}^\top = \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top + \boldsymbol{U}_{r^\perp}\boldsymbol{U}_{r^\perp}^\top \overset{t\to\infty}{\longrightarrow} \boldsymbol{U}_{r^\perp}\boldsymbol{U}_{r^\perp}^\top\ , \tag{101a}$$

$$e^{-\frac{1}{n}\boldsymbol{X}_\mathcal{A}^\top\boldsymbol{X}_\mathcal{A} t} = \boldsymbol{V}e^{-\frac{1}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda} t}\boldsymbol{V}^\top = \boldsymbol{V}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{V}_r^\top + \boldsymbol{V}_{r^\perp}\boldsymbol{V}_{r^\perp}^\top \overset{t\to\infty}{\longrightarrow} \boldsymbol{V}_{r^\perp}\boldsymbol{V}_{r^\perp}^\top\ . \tag{101b}$$

In the model we look at, elements of $\boldsymbol{X}_\mathcal{A} \in \mathbb{R}^{n\times p}$ are sampled independently from a standard normal distribution. This has several consequences. First, with probability 1, $\boldsymbol{X}_\mathcal{A}$ is a full rank matrix, i.e. $r = \operatorname{rank} \boldsymbol{X}_\mathcal{A} = \min(n, p)$. Second, this allows us to characterize some properties of distribution of SVD components from Eq. (99) – we can consider $\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}$ to be mutually independent; in addition, $\boldsymbol{U}$ and $\boldsymbol{V}$ are distributed uniformly on orthogonal groups of unitary matrices $O(n)$ and $O(p)$ of size n and p and respectively.

C.2. SDE modeling

We apply SDE modeling of Eq. (4) to the simple regression model. To do this, we first find the drift vector

$$\boldsymbol{\nabla}\mathscr{L}\big(\hat{\boldsymbol{\beta}}(t); \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}\big) = -\frac{1}{n}\boldsymbol{X}_\mathcal{A}^\top\Big(\boldsymbol{y} - \boldsymbol{X}_\mathcal{A}\hat{\boldsymbol{\beta}}_\mathcal{A}(t)\Big) \tag{102}$$

and then the diffusion matrix. For the latter, we define, according to Eq. (3), stochastic perturbation

$$\begin{aligned}\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}\big) &= \boldsymbol{\nabla} l\big(\hat{\boldsymbol{\beta}}; \boldsymbol{x}^{k_\nu}, y^{k_\nu}; \mathcal{A}\big) - \boldsymbol{\nabla}\mathscr{L}\big(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}\big) \\ &= -\Big(y^{k_\nu} - \hat{\boldsymbol{\beta}}_\mathcal{A}^\top\boldsymbol{x}_\mathcal{A}^{k_\nu}\Big)\boldsymbol{x}_\mathcal{A}^{k_\nu} + \frac{1}{n}\boldsymbol{X}_\mathcal{A}^\top(\boldsymbol{y} - \boldsymbol{X}_\mathcal{A}\hat{\boldsymbol{\beta}}_\mathcal{A}) \\ &= -\left[\Big(y^{k_\nu} - \hat{\boldsymbol{\beta}}_\mathcal{A}^\top\boldsymbol{x}_\mathcal{A}^{k_\nu}\Big)\boldsymbol{x}_\mathcal{A}^{k_\nu} - \frac{1}{n}\sum_{k=1}^n\Big(y^k - \hat{\boldsymbol{\beta}}_\mathcal{A}^\top\boldsymbol{x}_\mathcal{A}^k\Big)\boldsymbol{x}_\mathcal{A}^k\right]\ .\end{aligned} \tag{103}$$

Let $\{\mathbf{e}_j\}_{j=1}^n$ be a set of vectors composing the standard orthonormal basis in $\mathbb{R}^n$: $\mathbf{e}_j^\top\mathbf{e}_l = \delta_{jl}$. The random variable $\boldsymbol{\xi}$ can then be rewritten as

$$\boldsymbol{\xi}_{k_\nu}\big(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}\big) = -\boldsymbol{X}_\mathcal{A}^\top\left[\Big(y^{k_\nu} - \hat{\boldsymbol{\beta}}_\mathcal{A}^\top\boldsymbol{x}_\mathcal{A}^{k_\nu}\Big)\mathbf{e}_{k_\nu} - \frac{1}{n}\sum_{k=1}^n\Big(y^k - \hat{\boldsymbol{\beta}}_\mathcal{A}^\top\boldsymbol{x}_\mathcal{A}^k\Big)\mathbf{e}_k\right]\ . \tag{104}$$

Additionally, since the samples are picked uniformly at random, we can replace the sum in the second term with the expectation:

$$\boldsymbol{\xi}_{k_\nu}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)=-\boldsymbol{X}_{\mathcal{A}}^{\top}\left[\left(y^{k_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{k_\nu}\right)\mathbf{e}_{k_\nu}-\mathbb{E}_{q_\nu\sim\mathcal{U}\{1,n\}}\left[\left(y^{q_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{q_\nu}\right)\mathbf{e}_{q_\nu}\right]\right] . \tag{105}$$

The first moment is zero:

$$\mathbb{E}_{k_\nu\sim\mathcal{U}\{1,n\}}\left[\boldsymbol{\xi}_{k_\nu}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\right]=\mathbf{0} . \tag{106}$$

We now aim to compute the covariance matrix

$$\begin{aligned}
\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right) &= \mathbb{E}_{k_\nu\sim\mathcal{U}\{1,n\}}\left[\boldsymbol{\xi}_{k_\nu}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\boldsymbol{\xi}_{k_\nu}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)^{\top}\right]= \\
&= \boldsymbol{X}_{\mathcal{A}}^{\top}\left[\frac{1}{n}\sum_{k_\nu=1}^{n}\left(y^{k_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{k_\nu}\right)^2\mathbf{e}_{k_\nu}\mathbf{e}_{k_\nu}^{\top}-\frac{1}{n^2}\sum_{k_\nu,q_\nu=1}^{n}\left(y^{k_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{k_\nu}\right)\left(y^{q_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{q_\nu}\right)\mathbf{e}_{k_\nu}\mathbf{e}_{q_\nu}^{\top}\right]\boldsymbol{X}_{\mathcal{A}} \\
&= \boldsymbol{X}_{\mathcal{A}}^{\top}\left[\frac{2}{n}\sum_{k_\nu=1}^{n}l\left(\hat{\boldsymbol{\beta}};\boldsymbol{x}^{k_\nu},y^{k_\nu};\mathcal{A}\right)\mathbf{e}_{k_\nu}\mathbf{e}_{k_\nu}^{\top}-\frac{1}{n^2}\sum_{k_\nu,q_\nu=1}^{n}\left(y^{k_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{k_\nu}\right)\left(y^{q_\nu}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\top}\boldsymbol{x}_{\mathcal{A}}^{q_\nu}\right)\mathbf{e}_{k_\nu}\mathbf{e}_{q_\nu}^{\top}\right]\boldsymbol{X}_{\mathcal{A}} \\
&= \frac{2}{n}\boldsymbol{X}_{\mathcal{A}}^{\top}\begin{pmatrix} l\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}};\boldsymbol{x}^1,y^1;\mathcal{A}\right) & 0 & \dots & 0 \\ 0 & l\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}};\boldsymbol{x}^2,y^2;\mathcal{A}\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & l\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}};\boldsymbol{x}^n,y^n;\mathcal{A}\right)\end{pmatrix}\boldsymbol{X}_{\mathcal{A}}+\mathcal{O}\left(\frac{1}{n^2}\right) .
\end{aligned} \tag{107}$$

We neglect $\mathcal{O}\left(n^{-2}\right)$ contributions and assume that the individual loss function $l\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}};\boldsymbol{x}^k,y^k;\mathcal{A}\right)$ is equal to its average $\mathscr{L}\left(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)$. Indeed, the result is then

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}(t);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\approx\frac{2}{n}\mathscr{L}\left(\hat{\boldsymbol{\beta}}(t);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\boldsymbol{X}_{\mathcal{A}}^{\top}\boldsymbol{X}_{\mathcal{A}} , \tag{108}$$

which matches Eq. (22). Along with Eq. (102), we obtain Eq. (23).

C.3. Solving gradient flow equation

We assume initial condition $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\equiv\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{0}$. When $\gamma=0$, SDE (23) becomes the following ODE:

$$\frac{\mathrm{d}\hat{\boldsymbol{\beta}}_{\mathcal{A}}}{\mathrm{d}t}=\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^{\top}(\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}) . \tag{109}$$

This is a first-order matrix differential equation. Applying (99), we can write it as:

$$\frac{\mathrm{d}\hat{\boldsymbol{\beta}}_{\mathcal{A}}}{\mathrm{d}t}=-\frac{1}{n}\boldsymbol{V}\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}\boldsymbol{V}^{\top}\hat{\boldsymbol{\beta}}_{\mathcal{A}}+\frac{1}{n}\boldsymbol{V}\boldsymbol{\Lambda}^{\top}\boldsymbol{U}^{\top}\boldsymbol{y} . \tag{110}$$

Now, we can transform it to a system of independent linear ODEs by multiplying both sides by $\boldsymbol{V}^{\top}$, which is full-rank:

$$\frac{\mathrm{d}(\boldsymbol{V}^{\top}\hat{\boldsymbol{\beta}}_{\mathcal{A}})}{\mathrm{d}t}=-\frac{1}{n}\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}(\boldsymbol{V}^{\top}\hat{\boldsymbol{\beta}}_{\mathcal{A}})+\frac{1}{n}\boldsymbol{\Lambda}^{\top}\boldsymbol{U}^{\top}\boldsymbol{y} , \tag{111}$$

i.e., we solve:

$$\frac{\mathrm{d}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}}{\mathrm{d}t}=-\frac{1}{n}\boldsymbol{\Lambda}^{\top}\boldsymbol{\Lambda}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}+\frac{1}{n}\boldsymbol{\Lambda}^{\top}\boldsymbol{U}^{\top}\boldsymbol{y} , \tag{112}$$

where $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}\equiv\boldsymbol{V}^{\top}\hat{\boldsymbol{\beta}}_{\mathcal{A}}$. Let $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(t)=\left(\tilde{\beta}_1(t),\dots,\tilde{\beta}_p(t)\right)$ be its component representation. Then, the system can be written as:

$$\frac{\mathrm{d}\tilde{\beta}_i}{\mathrm{d}t}=\begin{cases}-\frac{1}{n}\lambda_i^2\tilde{\beta}_i+\frac{1}{n}\lambda_i\boldsymbol{u}_i^{\top}\boldsymbol{y} & \text{for } i\le r ; \\ 0 & \text{for } i>r .\end{cases} \tag{113}$$

with initial condition $\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(0) = \boldsymbol{V}^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) = \boldsymbol{V}^\top \hat{\boldsymbol{\beta}}^0_{\mathcal{A}}$. The solution of the system is then:

$$\tilde{\beta}_i(t) = \begin{cases} e^{-\frac{1}{n}\lambda_i^2 t}\left(\tilde{\beta}_i(0) - \frac{\boldsymbol{u}_i^\top \boldsymbol{y}}{\lambda_i}\right) + \frac{\boldsymbol{u}_i^\top \boldsymbol{y}}{\lambda_i} & \text{for } i \le r\ ; \\ \tilde{\beta}_i(0) & \text{for } i > r\ . \end{cases} \tag{114}$$

In vector form it can be written as:

$$\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(t) = e^{-\frac{1}{n}\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} t}\tilde{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \mathbb{I}_{p\times r}\boldsymbol{\Lambda}_r^{-1}\left(\mathbb{I}_r - e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\right)\boldsymbol{U}_r^\top \boldsymbol{y}\ . \tag{115}$$

Multiplying by $\boldsymbol{V}$, we get back to the original basis

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}(t) = \boldsymbol{V} e^{-\frac{1}{n}\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} t}\boldsymbol{V}^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \boldsymbol{V}_r \boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top\left(\mathbb{I}_n - \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top\right)\boldsymbol{y}\ , \tag{116}$$

or

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}}(t) = e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{A}} t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top\right)\boldsymbol{y}\ . \tag{117}$$

Note that, using Eq. (101):

$$\begin{aligned} \boldsymbol{X}_{\mathcal{A}}^\dagger e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t} &= \boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top \cdot \boldsymbol{U} e^{-\frac{1}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top t}\boldsymbol{U}^\top = \boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top\left(\boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top + \boldsymbol{U}_{r^\perp}\boldsymbol{U}_{r^\perp}^\top\right) = \\ &= \boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top\boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top = \boldsymbol{X}_{\mathcal{A}}^\dagger \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top\ , \end{aligned} \tag{118}$$

which leads us to the concluding form of the GF trajectory solution shown in Eq. (24):

$$\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t) = e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top \boldsymbol{X}_{\mathcal{A}} t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{y}\ . \tag{119}$$

When $t \to \infty$, we can observe using Eq. (101):

$$\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t) \overset{t\to\infty}{\longrightarrow} \boldsymbol{V}_{r^\perp}\boldsymbol{V}_{r^\perp}^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - \boldsymbol{U}_{r^\perp}\boldsymbol{U}_{r^\perp}^\top\right)\boldsymbol{y} = \boldsymbol{V}_{r^\perp}\boldsymbol{V}_{r^\perp}^\top \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + \boldsymbol{X}_{\mathcal{A}}^\dagger \boldsymbol{y}\ . \tag{120}$$

Therefore, GF converges to the minimum norm solution of the least-squares problem $\min_{\hat{\boldsymbol{\beta}}_{\mathcal{A}}\in\mathbb{R}^p}\mathscr{L}\big(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{y}; \mathcal{A}\big)$ with one caveat – when $r < p$ the problem is overparametrized and therefore the solution includes the projection of $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)$ onto the subspace orthogonal to row space of $\boldsymbol{X}_{\mathcal{A}}$. In the Gaussian setting, aside of cases with probability 0, this happens iff $n < p$.

C.4. Evaluating the generalization error for gradient flow

Our goal is to compute:

$$\mathcal{E}^{\text{GF}}_{\text{test}}(t) = \mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left[l\big(\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t); \boldsymbol{x}^{\text{new}}, y^{\text{new}}\big)\right]\right] = \frac{1}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left[\left(y^{\text{new}} - (\boldsymbol{x}^{\text{new}}_{\mathcal{A}})^\top \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\right)^2\right]\ , \tag{121}$$

where $\boldsymbol{x}^{\text{new}}, y^{\text{new}} \sim \mathbb{P}(\boldsymbol{x}^{\text{new}}, y^{\text{new}})$ is a new sample from the distribution described in Section 4.

First, we look at the expectation over a new sample; as $y^{\text{new}} = \boldsymbol{\beta}^\top \boldsymbol{x}^{\text{new}} + \mu\epsilon^{\text{new}} = \boldsymbol{\beta}_{\mathcal{A}}^\top \boldsymbol{x}^{\text{new}}_{\mathcal{A}} + \boldsymbol{\beta}_{\mathcal{A}^c}^\top \boldsymbol{x}^{\text{new}}_{\mathcal{A}^c} + \mu\epsilon^{\text{new}}$ and $\boldsymbol{x}^{\text{new}}_{\mathcal{A}} \sim \mathcal{N}(\boldsymbol{x}^{\text{new}}_{\mathcal{A}}|\boldsymbol{0}, \mathbb{I}_p), \boldsymbol{x}^{\text{new}}_{\mathcal{A}^c} \sim \mathcal{N}(\boldsymbol{x}^{\text{new}}_{\mathcal{A}^c}|\boldsymbol{0}, \mathbb{I}_{d-p}), \epsilon^{\text{new}} \sim \mathcal{N}(\epsilon^{\text{new}}|0, 1)$ are mutually independent,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left[\left(y^{\text{new}} - (\boldsymbol{x}^{\text{new}}_{\mathcal{A}})^\top \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\right)^2\right] = \\ &= \mathbb{E}_{\boldsymbol{x}^{\text{new}}_{\mathcal{A}}}\left[\left(\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\right)^\top \boldsymbol{x}^{\text{new}}_{\mathcal{A}}\right)^2\right] + \mathbb{E}_{\boldsymbol{x}^{\text{new}}_{\mathcal{A}^c}}\left[\left(\boldsymbol{\beta}_{\mathcal{A}^c}^\top \boldsymbol{x}^{\text{new}}_{\mathcal{A}^c}\right)^2\right] + \mu^2\mathbb{E}_{\epsilon^{\text{new}}}\left[(\epsilon^{\text{new}})^2\right] \\ &= \left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\right)^\top \mathbb{E}_{\boldsymbol{x}^{\text{new}}_{\mathcal{A}}}\left[\boldsymbol{x}^{\text{new}}_{\mathcal{A}}(\boldsymbol{x}^{\text{new}}_{\mathcal{A}})^\top\right]\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\right) + \boldsymbol{\beta}_{\mathcal{A}^c}^\top \mathbb{E}_{\boldsymbol{x}^{\text{new}}_{\mathcal{A}^c}}\left[\boldsymbol{x}^{\text{new}}_{\mathcal{A}^c}(\boldsymbol{x}^{\text{new}}_{\mathcal{A}^c})^\top\right]\boldsymbol{\beta}_{\mathcal{A}^c} + \mu^2 \\ &= \|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2 + \|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2 + \mu^2\ . \end{aligned} \tag{122}$$

We then continue deriving expression for Eq. (121) by taking the expectation over $(\boldsymbol{X}, \boldsymbol{y})$. Only the first term of the resulting sum of Eq. (122) depends on it, therefore our goal now is to compute $\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[\|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right]$.

We denote $\boldsymbol{\epsilon} := \left(\epsilon^1, \ldots, \epsilon^n\right) \in \mathbb{R}^n$ as a vector of additive Gaussian noises. Then, we can write target vector as:

$$\boldsymbol{y} = \boldsymbol{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} + \boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon} \ . \tag{123}$$

The vector of differences between true value of parameter and GF estimate from Eq. (119) is:

$$\begin{aligned}
&\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\mathrm{GF}}_{\mathcal{A}}(t) = \boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{y} = \\
&= \left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right) \ .
\end{aligned} \tag{124}$$

As $\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}$ is zero-mean vector independent from $\boldsymbol{X}_{\mathcal{A}}$, we can write:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[\|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}^{\mathrm{GF}}_{\mathcal{A}}(t)\|_2^2\right] &= \mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\|\left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\right] \\
&\quad + \mathbb{E}_{\boldsymbol{X}_{\mathcal{A}},\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)\|^2\right] \ .
\end{aligned} \tag{125}$$

Using Eqs. (99), (100), (101), we can simplify the first term, because:

$$\begin{aligned}
&\left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} = \\
&= \left(\mathbb{I}_p - \boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^\top\left(\mathbb{I}_n - \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{U}_r^\top - \boldsymbol{U}_{r\perp}\boldsymbol{U}_{r\perp}^\top\right)\boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^\top\right)\boldsymbol{\beta}_{\mathcal{A}} = \\
&= \left(\mathbb{I}_p - \boldsymbol{V}_r\boldsymbol{V}_r^\top + \boldsymbol{V}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{V}_r^\top\right)\boldsymbol{\beta}_{\mathcal{A}} = \left(\boldsymbol{V}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2 t}\boldsymbol{V}_r^\top + \boldsymbol{V}_{r\perp}\boldsymbol{V}_{r\perp}^\top\right)\boldsymbol{\beta}_{\mathcal{A}} = \\
&= e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\boldsymbol{\beta}_{\mathcal{A}} \ ,
\end{aligned} \tag{126}$$

and, therefore:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\|\left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\right] = \\
&\quad = \mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\|e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0))\|^2\right] = \left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)^\top\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\right]\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right) \ .
\end{aligned} \tag{127}$$

Going back to representation of Eq. (101), we can rewrite it in terms of the singular values of the matrix $\boldsymbol{X}_{\mathcal{A}}$. Indeed, applying trace trick (again, $\boldsymbol{\Lambda}$ and $\boldsymbol{V}$ are independent):

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\|\left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\right] = \\
&\quad = \mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda},\boldsymbol{V}}\left[\left(\boldsymbol{V}^\top\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\right)^\top e^{-\frac{2}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}t}\boldsymbol{V}^\top\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\right]\right\} = \\
&\quad = \mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}t}\right]\mathbb{E}_{\boldsymbol{V}}\left[\boldsymbol{V}^\top\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)^\top\boldsymbol{V}\right]\right\} \ .
\end{aligned} \tag{128}$$

As $\boldsymbol{V} \sim \mathrm{Unif}\left(O\left(p\right)\right)$, if $\boldsymbol{a} \in \mathbb{R}^p$ is arbitrary vector of dimensionality p, then

$$\boldsymbol{V}^\top\boldsymbol{a} \sim \mathrm{Unif}\left(\mathcal{S}^{p-1}\left(\|\boldsymbol{a}\|\right)\right) \ , \tag{129a}$$

where $\mathcal{S}^{d-1}\left(r\right)$ denotes d-dimensional sphere with radius r, and thus

$$\mathbb{E}_{\boldsymbol{V}}\left[\boldsymbol{V}^\top\boldsymbol{a}\right] = \boldsymbol{0}_p, \ \mathbb{E}_{\boldsymbol{V}}\left[\boldsymbol{V}^\top\boldsymbol{a}\left(\boldsymbol{V}^\top\boldsymbol{a}\right)^\top\right] = \frac{1}{p}\|\boldsymbol{a}\|^2 \cdot \mathbb{I}_p \ . \tag{129b}$$

This results in:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\|\left(\mathbb{I}_p - \boldsymbol{X}_{\mathcal{A}}^\dagger\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top t}\right)\boldsymbol{X}_{\mathcal{A}}\right)\boldsymbol{\beta}_{\mathcal{A}} - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\right] = \\
&\quad = \frac{1}{p}\|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\,\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}t}\right]\right\} \\
&\quad = \|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2 \cdot \frac{1}{p}\,\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 t} + (p-r)\right]\right\} = \\
&\quad = \|\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2 \cdot \left(\frac{1}{p}\,\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 t}\right]\right\} + \max\left(0, 1 - \frac{n}{p}\right)\right) \ .
\end{aligned} \tag{130}$$

For the second term of Eq. (125), we can apply trace trick again:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}},\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)\|^2\right]= \\
&=\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}},\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\mathrm{Tr}\left\{\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)^{\top}\left(\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\right)^{\top}\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)\right\}\right]= \\
&=\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}}}\left[\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\left(\boldsymbol{X}_{\mathcal{A}}^{\dagger}\right)^{\top}\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\right]\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)^{\top}\right]\right\}.
\end{aligned} \tag{131}$$

We note that the vector $\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}$ is distributed as:

$$\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\sim\mathcal{N}\left(\boldsymbol{0},\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\mathbb{I}_n\right)\ , \tag{132}$$

and that on the other hand,

$$\begin{aligned}
\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)&=\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\boldsymbol{U}_r^{\top}\left(\mathbb{I}_n-\boldsymbol{U}_re^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\boldsymbol{U}_r^{\top}-\boldsymbol{U}_{r\perp}\boldsymbol{U}_{r\perp}^{\top}\right)= \\
&=\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\left(\boldsymbol{U}_r^{\top}-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\boldsymbol{U}_r^{\top}\right)=\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)\boldsymbol{U}_r^{\top}\ ;
\end{aligned} \tag{133}$$

thus it follows from Eq. (131):

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}},\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{X}_{\mathcal{A}}^{\dagger}\left(\mathbb{I}_n-e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^{\top}t}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c}+\mu\boldsymbol{\epsilon}\right)\|^2\right]= \\
&=\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{U},\boldsymbol{\Lambda},\boldsymbol{V}}\left[\boldsymbol{U}_r\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)\boldsymbol{\Lambda}_r^{-1}\boldsymbol{V}_r^{\top}\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)\boldsymbol{U}_r^{\top}\right]\right\}= \\
&=\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\Lambda}_r^{-2}\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)^2\right]\right\}.
\end{aligned} \tag{134}$$

Gathering expressions from (130) and (134) back to (125), we obtain:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\|_2^2\right]&=\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\cdot\left(\max\left(0,1-\frac{n}{p}\right)+\frac{1}{p}\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}\right)+ \\
&+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\Lambda}_r^{-2}\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)^2\right]\right\},
\end{aligned} \tag{135}$$

which can be used to obtain average test error over $\boldsymbol{X},\boldsymbol{y}$ by taking the expectation over both first and last equal expressions of Eq. (122) and substituting one of the terms with this one.

We return to test error Eq. (121) and average over $\mathcal{A}$. The only terms depending on $\mathcal{A}$ are $\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2$ and $\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2$. As $\mathcal{A}$ in our model is a uniformly random subset of $[d]$ of cardinality p,

$$\mathbb{E}_{\mathcal{A}}\left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\right]=\frac{p}{d}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2,\ \mathbb{E}_{\mathcal{A}}\left[\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2\right]=\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|^2\ . \tag{136}$$

Finally, we obtain:

$$\begin{aligned}
\mathcal{E}_{\mathrm{test}}^{\mathrm{GF}}(t)&=\frac{1}{2}\mathbb{E}_{\mathcal{A}}\left[\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\|_2^2\right]+\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right]= \\
&=\frac{1}{2}\left(\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot\left(\max\left(0,\ \frac{p-n}{d}\right)+\frac{1}{d}\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}\right)+\right. \\
&\left.+\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\left[1+\mathrm{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\Lambda}_r^{-2}\left(\mathbb{I}_r-e^{-\frac{1}{n}\boldsymbol{\Lambda}_r^2t}\right)^2\right]\right\}\right]\right),
\end{aligned} \tag{137}$$

which matches expression (27).

C.5. Evaluating train error for gradient flow

In this section, we compute the average train error for GF solution:

$$\mathcal{E}^{\text{GF}}_{\text{train}}(t) = \mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathscr{L}\big(\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)\right] = \frac{1}{2n}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right] . \tag{138}$$

While in this paper we do not insist on the properties of the train error, the derivation itself will prove useful in C.6. This is not surprising, as the train error is involved in modeling of diffusion matrix.

First, we look at the 'vector of train errors'; we substitute the gradient flow estimator with the solution from Eq. (119):

$$\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t) = \left[\mathbb{I}_n - \boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\dagger}_{\mathcal{A}}\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\right)\right]\boldsymbol{y} - \boldsymbol{X}_{\mathcal{A}}e^{-\frac{1}{n}\boldsymbol{X}^{\top}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) . \tag{139}$$

The multiplier in the first term can be simplified utilizing Eqs. (133), (101):

$$\begin{aligned}\mathbb{I}_n - \boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\dagger}_{\mathcal{A}}\left(\mathbb{I}_n - e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\right) &= \mathbb{I}_n - \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\top}\boldsymbol{V}_r\boldsymbol{\Lambda}_r^{-1}\left(\mathbb{I}_r - e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\right)\boldsymbol{U}_r^{\top} = \\ &= \boldsymbol{U}_r\boldsymbol{U}_r^{\top} + \boldsymbol{U}_{r\perp}\boldsymbol{U}_{r\perp}^{\top} - \boldsymbol{U}_r\boldsymbol{U}_r^{\top} + \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{U}_r^{\top} = \boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{U}_r^{\top} + \boldsymbol{U}_{r\perp}\boldsymbol{U}_{r\perp}^{\top} = e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t} ;\end{aligned} \tag{140}$$

with this and the explicit form of $\boldsymbol{y}$ from Eq. (123), we obtain:

$$\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t) = e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \boldsymbol{X}_{\mathcal{A}}e^{-\frac{1}{n}\boldsymbol{X}^{\top}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}t}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0) + e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right) . \tag{141}$$

Now, using properties of SVD of $\boldsymbol{X}_{\mathcal{A}}$, we can see that:

$$\begin{aligned} e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\boldsymbol{X}_{\mathcal{A}} &= \left(\boldsymbol{U}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{U}_r^{\top} + \boldsymbol{U}_{r\perp}\boldsymbol{U}_{r\perp}^{\top}\right)\boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\top} = \boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top} ; \\ \boldsymbol{X}_{\mathcal{A}}e^{-\frac{1}{n}\boldsymbol{X}^{\top}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}t} &= \boldsymbol{U}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^{\top}\left(\boldsymbol{V}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top} + \boldsymbol{V}_{r\perp}\boldsymbol{V}_{r\perp}^{\top}\right) = \boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top} .\end{aligned} \tag{142}$$

Then, the vector of train errors can be split into two terms – one depending on $\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)$, another on $\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}$:

$$\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t) = \boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right) + e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right) . \tag{143}$$

Now, we would like to take expectation of $\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2$ over $\boldsymbol{X},\boldsymbol{y}$, or, equivalently, over $\boldsymbol{X}_{\mathcal{A}},\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}$. We start with the latter two. Again $\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}$ is zero-mean vector independent from $\boldsymbol{X}_{\mathcal{A}}$, so we can write:

$$\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right] = \|\boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\|^2 + \mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)\|^2\right] . \tag{144}$$

The first term can be reduced to:

$$\begin{aligned}\|\boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\|^2 &= \left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)^{\top}\boldsymbol{V}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{\Lambda}_r\boldsymbol{U}_r^{\top}\boldsymbol{U}_r\boldsymbol{\Lambda}_r e^{-\frac{1}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right) \\ &= \left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)^{\top}\boldsymbol{V}_r\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}^2_r t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right) ,\end{aligned} \tag{145}$$

To take the expectation in the second one, along the lines of Eq. (131), we use again use trace trick and the covariance of vector $\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}$ (see Eq. (132)):

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)\|^2\right] &= \mathrm{Tr}\Big\{\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)^{\top}e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)\right]\Big\} = \\ &= \mathrm{Tr}\Big\{e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\cdot\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)\left(\boldsymbol{X}_{\mathcal{A}^c}\boldsymbol{\beta}_{\mathcal{A}^c} + \mu\boldsymbol{\epsilon}\right)^{\top}\right]\Big\} = \left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2 + \mu^2\right)\mathrm{Tr}\Big\{e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\Big\} .\end{aligned} \tag{146}$$

We write this result in terms of SVD's of matrices (noting that in fact the expectation over $\boldsymbol{U}$ is not necessary here):

$$\mathrm{Tr}\Big\{e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}^{\top}_{\mathcal{A}}t}\Big\} = \mathrm{Tr}\Big\{\boldsymbol{U}e^{-\frac{2}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top}t}\boldsymbol{U}^{\top}\Big\} = \mathrm{Tr}\Big\{e^{-\frac{2}{n}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top}t}\Big\} = \mathrm{Tr}\Big\{e^{-\frac{2}{n}\boldsymbol{\Lambda}^2_r t}\Big\} + n - r . \tag{147}$$

All in all, we collect everything back into Eq. (144):

$$\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right]= \\ =\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)^{\top}\boldsymbol{V}_r\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\left(\text{Tr}\left\{e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right\}+n-r\right)\ . \quad (148)$$

Now, to average over $\boldsymbol{X}_{\mathcal{A}}$, we start with the expectation over $\boldsymbol{U}$ and $\boldsymbol{V}$ first, because the result would be convenient to employ in section C.6. $\boldsymbol{U}$ is not present at all, and $\boldsymbol{V}$ is found only in the first term. With an application of trace trick,

$$\mathbb{E}_{\boldsymbol{V}}\left[\left(\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\right)^{\top}\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\right]= \\ =\text{Tr}\left\{\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\mathbb{E}_{\boldsymbol{V}}\left[\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\left(\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)\right)^{\top}\right]\right\}\ . \quad (149)$$

We discussed distribution of the vector $\boldsymbol{V}^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)$ in Eq. (129); then, we can see that $\boldsymbol{V}_r^{\top}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\right)$ are simply first r components of it, and, therefore, the new vector is also zero-mean with covariance matrix is $\frac{1}{p}\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\mathbb{I}_r$. This leads us to:

$$\mathbb{E}_{\boldsymbol{U},\boldsymbol{V};\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right]=\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\frac{1}{p}\,\text{Tr}\left\{\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right\}+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\left(\text{Tr}\left\{e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right\}+n-r\right)\ . \quad (150)$$

At last, the whole expectation of square of train error over $\boldsymbol{X},\boldsymbol{y}$ is found to be

$$\mathbb{E}_{\boldsymbol{X},y}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right]=\mathbb{E}_{\boldsymbol{X}_{\mathcal{A}};\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)\|^2\right]= \\ =\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\frac{1}{p}\,\text{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\left(\text{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}+\max(0,n-p)\right)\ , \quad (151)$$

and, by taking average over feature set $\mathcal{A}$ (sampled as in Eq. (136)) we obtain the final expression for Eq. (138):

$$\mathcal{E}^{\text{GF}}_{\text{train}}(t)=\frac{1}{2}\Bigg(\frac{1}{d}\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot\frac{1}{n}\,\text{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}+ \\ +\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\left(\frac{1}{n}\,\text{Tr}\left\{\mathbb{E}_{\boldsymbol{\Lambda}}\left[e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2t}\right]\right\}+\max\left(0,1-\frac{p}{n}\right)\right)\Bigg)\ . \quad (152)$$

C.6. Evaluating generalization error for stochastic gradient flow

Now, we would like to evaluate stochastic gradient flow test error, i.e. error for estimator $\hat{\boldsymbol{\beta}}^{\text{SGF}}_{\mathcal{A}}(t)=\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}(t)+\sqrt{\gamma}\boldsymbol{z}(t)$, with $\boldsymbol{z}(0)=\boldsymbol{0}$ and $\boldsymbol{z}(t)$ being a mean-zero Gaussian process with the covariance matrix $\boldsymbol{\mathscr{C}}(t)$.

First of all, it is easy to see that the derivations in Eq. (122) are still valid if we replace $\hat{\boldsymbol{\beta}}^{\text{GF}}_{\mathcal{A}}$ with $\hat{\boldsymbol{\beta}}^{\text{SGF}}_{\mathcal{A}}$, and therefore:

$$\begin{aligned}\mathcal{E}^{\text{SGF}}_{\text{test}}(t)&=\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{b}}\left[\mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left[l\left(\hat{\boldsymbol{\beta}}^{\text{SGF}}_{\mathcal{A}}(t);\boldsymbol{x}^{\text{new}},y^{\text{new}}\right)\right]\right]\\&=\frac{1}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y};\boldsymbol{b}}\left[\mathbb{E}_{\boldsymbol{x}^{\text{new}},y^{\text{new}}}\left[\left(y^{\text{new}}-(\boldsymbol{x}^{\text{new}}_{\mathcal{A}})^{\top}\hat{\boldsymbol{\beta}}^{\text{SGF}}_{\mathcal{A}}(t)\right)^2\right]\right]\\&=\frac{1}{2}\mathbb{E}_{\mathcal{A}}\left[\mathbb{E}_{\boldsymbol{X},\boldsymbol{y};\boldsymbol{b}}\left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}^{\text{SGF}}_{\mathcal{A}}(t)\|^2\right]+\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right]\ ,\end{aligned} \quad (153)$$

so we can easily obtain equality from Eq. (25).

Now, as we can decompose:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{b}}\left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{SGF}}(t)\|^2\right] &= \mathbb{E}_{\boldsymbol{b}}\left[\left\|\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\right)-\sqrt{\gamma}\boldsymbol{z}(t)\right\|^2\right] \\ &= \left\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\right\|^2 - 2\sqrt{\gamma}\left(\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\right)\mathbb{E}_{\boldsymbol{b}}\left[\boldsymbol{z}(t)\right]+\gamma\,\mathbb{E}_{\boldsymbol{b}}\left[\boldsymbol{z}^\top(t)\boldsymbol{z}(t)\right] \\ &= \left\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t)\right\|^2 + \gamma\,\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\ ,\end{aligned} \tag{154}$$

we can see that $\mathcal{E}_{\mathrm{test}}^{\mathrm{SGF}}(t)$ indeed can be split into sum of test error $\mathcal{E}_{\mathrm{test}}^{\mathrm{GF}}(t)$ of deterministic gradient flow and a term coming from perturbation and proportional to learning rate γ as in Eq. (26):

$$\mathcal{E}_{\mathrm{test}}^{\mathrm{SGF}}(t) = \mathcal{E}_{\mathrm{test}}^{\mathrm{GF}}(t) + \frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right]\ . \tag{155}$$

Evaluation of the first term was done in Section C.4. To complete it for the second term, we refer to Eq. (21a). In our setting, $\mathscr{L}\big(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)=\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}\|^2$, and its Hessian is

$$\boldsymbol{\mathcal{H}}_{\mathscr{L}}\big(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)=\boldsymbol{\nabla}_{\hat{\boldsymbol{\beta}}_{\mathcal{A}}}\otimes\boldsymbol{\nabla}_{\hat{\boldsymbol{\beta}}_{\mathcal{A}}}\mathscr{L}\big(\hat{\boldsymbol{\beta}};\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)=\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}\in\mathbb{R}^{p\times p}\ . \tag{156}$$

Then, the term in Eq. (21b) in our model becomes:

$$\boldsymbol{U}^{\mathrm{GF}}\left(t,\tau;\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\equiv\mathcal{T}\left\{e^{-\int_\tau^t \mathrm{d}s\boldsymbol{\mathcal{H}}_{\mathscr{L}}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(t);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)}\right\}=\mathcal{T}\left\{e^{-\frac{1}{n}\int_\tau^t \mathrm{d}s\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}}\right\}=e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}(t-\tau)}\ . \tag{157}$$

On the other hand, the diffusion matrix is, as stated in Eq. (22):

$$\boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}(t);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)\approx\frac{1}{n^2}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}\cdot\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}(t)\|^2\ . \tag{158}$$

This results in the following covariance matrix of the process:

$$\begin{aligned}\boldsymbol{\mathscr{C}}(t) &= \int_0^t \mathrm{d}\tau\ \boldsymbol{U}^{\mathrm{GF}}\left(t,\tau;\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)\ \boldsymbol{\Sigma}\big(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(\tau);\boldsymbol{X},\boldsymbol{y};\mathcal{A}\big)\ \boldsymbol{U}^{\mathrm{GF}}\left(t,\tau;\boldsymbol{X},\boldsymbol{y};\mathcal{A}\right)^\top \\ &= \frac{1}{n^2}\int_0^t \mathrm{d}\tau\ e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}(t-\tau)}\cdot\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}\cdot\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(\tau)\|^2\cdot e^{-\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}(t-\tau)} \\ &= \frac{1}{n^2}\int_0^t \mathrm{d}\tau\ \|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(\tau)\|^2\cdot\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}\ e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}(t-\tau)}\ .\end{aligned} \tag{159}$$

As $\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(\tau)\|^2$ is scalar and with Eq. (101),

$$\begin{aligned}\mathrm{Tr}\Big\{\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}\ e^{-\frac{2}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}(t-\tau)}\Big\} &= \mathrm{Tr}\Big\{\boldsymbol{V}\boldsymbol{\Lambda}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top\boldsymbol{V}e^{-\frac{2}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}(t-\tau)}\boldsymbol{V}^\top\Big\}= \\ &= \mathrm{Tr}\Big\{\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}e^{-\frac{2}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}(t-\tau)}\Big\}=\mathrm{Tr}\Big\{\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\Big\}\ ,\end{aligned} \tag{160}$$

we obtain the following expression for trace of covariance matrix:

$$\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}=\frac{1}{n^2}\int_0^t \mathrm{d}\tau\ \|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{\mathrm{GF}}(\tau)\|^2\cdot\mathrm{Tr}\Big\{\boldsymbol{\Lambda}_r^2e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\Big\}\ . \tag{161}$$

Now, we would like to compute $\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right]$. Expectation over $\boldsymbol{X},\boldsymbol{y}$ is the same as expectation over

$\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{V}, \boldsymbol{X}_{\mathcal{A}^c}, \boldsymbol{\epsilon}$, and thus, following Eq. (150),

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] &= \frac{1}{n^2}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{U},\boldsymbol{V};\boldsymbol{X}_{\mathcal{A}^c},\boldsymbol{\epsilon}}\left[\|\boldsymbol{y}-\boldsymbol{X}_{\mathcal{A}}\hat{\boldsymbol{\beta}}^{\mathrm{GF}}_{\mathcal{A}}(\tau)\|^2\right]\cdot \mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right] \\
&= \frac{1}{n^2}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\int_0^t \mathrm{d}\tau\, \left(\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\frac{1}{p}\,\mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}+\right.\right. \\
&\left.\left.+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\left(\mathrm{Tr}\left\{e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}+n-r\right)\right)\cdot \mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right] \\
&= \frac{1}{n^2}\int_0^t \mathrm{d}\tau\, \left[\|\boldsymbol{\beta}_{\mathcal{A}}-\hat{\boldsymbol{\beta}}_{\mathcal{A}}(0)\|^2\frac{1}{p}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right]+\left(\|\boldsymbol{\beta}_{\mathcal{A}^c}\|^2+\mu^2\right)\cdot\right. \\
&\left.\cdot\left(\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right]+\max\left(0,n-p\right)\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right]\right)\right]\,.
\end{aligned} \tag{162}$$

We avoid taking integral of time over product of traces right now as it makes further approximation complicated; however, the term below is easy to simplify:

$$\begin{aligned}
\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\frac{1}{n}\boldsymbol{\Lambda}_r^2 e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right] &= \mathbb{E}_{\boldsymbol{\Lambda}}\left[\sum_{i=1}^r \frac{1}{n}\lambda_i^2\int_0^t \mathrm{d}\tau\, e^{\frac{2}{n}\lambda_i^2(\tau-t)}\right]= \\
&= \mathbb{E}_{\boldsymbol{\Lambda}}\left[\sum_{i=1}^r \frac{1}{n}\lambda_i^2\cdot\frac{n}{2\lambda_i^2}\left(1-e^{-\frac{2}{n}\lambda_i^2 t}\right)\right]=\frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\sum_{i=1}^r\left(1-e^{-\frac{2}{n}\lambda_i^2 t}\right)\right]=\frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\mathbb{I}_r-e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 t}\right\}\right]\,.
\end{aligned} \tag{163}$$

With averaging over feature set $\mathcal{A}$ sampling as in Eq. (136) and substituting as above, we get the expression:

$$\begin{aligned}
\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] &= \frac{1}{d}\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\frac{1}{n}\boldsymbol{\Lambda}_r^2\cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\frac{1}{n}\boldsymbol{\Lambda}_r^2\cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right]+ \\
&+\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\left(\frac{1}{n}\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\frac{1}{n}\boldsymbol{\Lambda}_r^2\cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2(t-\tau)}\right\}\right]+\right. \\
&\left.+\frac{1}{2}\max\left(0,1-\frac{p}{n}\right)\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathrm{Tr}\left\{\mathbb{I}_r-e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 t}\right\}\right]\right)\,,
\end{aligned} \tag{164}$$

which matches Eq. (28).

D. Asymptotic generalization error

In Appendix C, we derived precise formulas for generalization errors of GF and SGF solutions. However, to analyze these results, we need to find expectation for several functions of $\boldsymbol{\Lambda}$.

D.1. Reformulation with empirical distribution of eigenvalues

We consider $p\times p$ matrix $\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}$. As elements of $\boldsymbol{X}_{\mathcal{A}}$ are i.i.d. standard normal, the former matrix follows the Wishart distribution $W_p\left(n,\ \frac{1}{n}\mathbb{I}_p\right)$ with n degrees of freedom. Let $\nu_{(p\times n)}$ be empirical distribution of eigenvalues of $\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}$. As $\boldsymbol{X}_{\mathcal{A}}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, the spectral decomposition is

$$\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}=\boldsymbol{V}\left(\frac{1}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\right)\boldsymbol{V}^\top\,, \tag{165}$$

i.e., eigenvalues are values on the diagonal of $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$. Following that, the measure $\nu_{(p\times n)}$ can be defined as:

$$\nu_{(p\times n)}\left(A\right)=\frac{1}{p}\#\left\{\frac{1}{n}\lambda_i^2\in A\mid i\in\overline{1,p}\right\}\,,\ A\subset\mathbb{R}\,. \tag{166}$$

If $p \leq n$, with probability 1, all eigenvalues are positive; $\{0\}$ is a set of measure 0, and $\nu_{(p\times n)}$ is an absolutely continuous distribution. Otherwise, it has measure $1 - {}^{n}\!/_{p}$.

Let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be a pointwise function. For arbitrary diagonal $m \times m$ matrix $\boldsymbol{D} = \operatorname{diag}(d_{11}, \ldots, d_{mm})$, we define $\varphi(\boldsymbol{D}) := \operatorname{diag}\left(\varphi(d_{11}), \ldots, \varphi(d_{mm})\right) \in \mathbb{R}^{m\times m}$. Then,

$$\mathbb{E}_{\sigma\sim\nu_{(p\times n)}}\left[\varphi(\sigma)\right] = \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}_{\lambda_i}\left[\varphi\left(\frac{1}{n}\lambda_i^2\right)\right] = \frac{1}{p}\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{n}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}\right)\right\}\right] =$$
$$= \frac{1}{p}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right)\right\} + (p-r)\,\varphi(0)\right] = \max\left(0, 1-\frac{n}{p}\right)\varphi(0) + \frac{1}{p}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right)\right\}\right] . \quad (167)$$

Thus, when $p \leq n$, we can express $\frac{1}{p}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right)\right\}\right]$ in terms of expected value of absolutely continuous measure. When $p > n$, we can repeat our line of thought for $\nu_{(n\times p)}$, which is an empirical distribution of eigenvalues of $n \times n$ matrix $\frac{1}{p}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top$. $\frac{1}{p}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top$ also has Wishart distribution – $W_n\left(p,\ \frac{1}{p}\mathbb{I}_n\right)$, – and spectral decomposition

$$\frac{1}{p}\boldsymbol{X}_{\mathcal{A}}\boldsymbol{X}_{\mathcal{A}}^\top = \boldsymbol{U}\left(\frac{1}{p}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\boldsymbol{U}^\top , \quad (168)$$

and obtain:

$$\mathbb{E}_{\sigma\sim\nu_{(n\times p)}}\left[\varphi(\sigma)\right] = \max\left(0, 1-\frac{p}{n}\right)\varphi(0) + \frac{1}{n}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{p}\boldsymbol{\Lambda}_r^2\right)\right\}\right] . \quad (169)$$

Now, let $\varphi' : \mathbb{R} \mapsto \mathbb{R}$ be such function that $\forall x \in \mathbb{R} : \varphi'(x) = \varphi\left(\frac{n}{p}x\right)$. Then, $\varphi'\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) = \varphi\left(\frac{1}{p}\boldsymbol{\Lambda}_r^2\right)$, and we can write:

$$\begin{aligned}\frac{1}{p}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi'\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right)\right\}\right] &= \frac{n}{p}\cdot\frac{1}{n}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{p}\boldsymbol{\Lambda}_r^2\right)\right\}\right] \\ &= \frac{n}{p}\left(\mathbb{E}_{\sigma\sim\nu_{(n\times p)}}\left[\varphi(\sigma)\right] - \max\left(0, 1-\frac{p}{n}\right)\varphi(0)\right) \\ &= \frac{n}{p}\,\mathbb{E}_{\sigma\sim\nu_{(n\times p)}}\left[\varphi'\left(\frac{p}{n}\sigma\right)\right] - \max\left(0, \frac{n}{p}-1\right)\varphi(0) .\end{aligned} \quad (170)$$

To sum up, for $\varphi : \mathbb{R} \mapsto \mathbb{R}$, we can use the following representations:

$$\frac{1}{p}\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\operatorname{Tr}\left\{\varphi\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right)\right\}\right] = \begin{cases}\mathbb{E}_{\sigma\sim\nu_{(p\times n)}}\left[\varphi(\sigma)\right] & \text{for } p \leq n\ ; \\ \frac{n}{p}\,\mathbb{E}_{\sigma\sim\nu_{(n\times p)}}\left[\varphi\left(\frac{p}{n}\sigma\right)\right] & \text{for } p > n\ .\end{cases} \quad (171)$$

D.2. Asymptotic approximation with Marchenko-Pastur distribution

Unfortunately, to our knowledge, there is no explicit expression for probability density of $\nu_{(n\times p)}$ which would be convenient to use for analysis of generalization error of solutions for finite sizes. However, for high enough values of n, d, p we can approximate the results asymptotically.

More precisely, as mentioned in Section 4.2, we consider $n, d, p \to \infty$ such that ${}^{p}\!/_{n} \to \alpha$, ${}^{d}\!/_{n} \to \psi$, where α and ψ are fixed positive values. As p must be less or equal than d, $\alpha \leq \psi$. Under these conditions, as $\boldsymbol{X}_{\mathcal{A}}$ has i.i.d. standard normal elements, empirical distribution $\nu_{(p\times n)}$ of eigenvalues of $\frac{1}{n}\boldsymbol{X}_{\mathcal{A}}^\top\boldsymbol{X}_{\mathcal{A}}$ weakly converges to well-known Marchenko-Pastur law ν_α. Its measure can be written as

$$\nu_\alpha(A) = \begin{cases}\left(1-\frac{1}{\alpha}\right)\mathbb{1}\left(0 \in A\right) + \rho_\alpha(A)\ , & \alpha > 1 \\ \rho_\alpha(A)\ , & 0 \leq \alpha \leq 1\end{cases};\quad A \subset \mathbb{R}, \quad (172)$$

where ρ_α is a following measure:

$$\rho_\alpha(\mathrm{d}\sigma) = \frac{\sqrt{(\alpha_+ - \sigma)(\sigma - \alpha_-)}}{2\pi\alpha\sigma}\mathbb{1}(\sigma \in [\alpha_-, \alpha_+])\,\mathrm{d}\sigma\ ,\quad \text{where } \alpha_\pm = (1 \pm \sqrt{\alpha})^2\ . \quad (173)$$

We remark that ρ_α is not a probability measure when $\alpha > 1$ as $\rho_\alpha(\mathbb{R}) \neq 1$. Likewise to $\nu_{(p\times n)}$, $\nu_{(n\times p)}$ converges to $\nu_{1/\alpha}$. Then, we derive from Eq. (171) that if $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is a pointwise function,

$$\frac{1}{p}\,\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\varphi\left(\frac{1}{n}\mathbf{\Lambda}_r^2\right)\right\}\right] \xrightarrow{n,p\to\infty,\ p/n\to\alpha} \begin{cases} \mathbb{E}_{\sigma\sim\nu_\alpha}\left[\varphi(\sigma)\right]\ , & \alpha \leq 1; \\ \frac{1}{\alpha}\,\mathbb{E}_{\sigma\sim\nu_{1/\alpha}}\left[\varphi\left(\alpha\sigma\right)\right]\ , & \alpha > 1. \end{cases} \tag{174}$$

Now, when $\alpha \leq 1$,

$$\mathbb{E}_{\sigma\sim\nu_\alpha}\left[\varphi(\sigma)\right] = \int \varphi(\sigma)\,\rho_\alpha(\mathrm{d}\sigma), \tag{175}$$

and when $\alpha > 1$,

$$\frac{1}{\alpha}\,\mathbb{E}_{\sigma\sim\nu_{1/\alpha}}\left[\varphi\left(\alpha\sigma\right)\right] = \frac{1}{\alpha}\int \varphi(\alpha\sigma)\,\rho_{1/\alpha}(\mathrm{d}\sigma) = \frac{1}{\alpha}\int_{(1/\alpha)_-}^{(1/\alpha)_+} \varphi(\alpha\sigma)\frac{\sqrt{\left((1/\alpha)_+ - \sigma\right)\left(\sigma - (1/\alpha)_-\right)}}{2\pi\sigma/\alpha}\mathrm{d}\sigma\ . \tag{176}$$

We can transform the latter expression; as

$$\left(\frac{1}{\alpha}\right)_\pm = \left(1 \pm \frac{1}{\sqrt{\alpha}}\right)^2 = \frac{1}{\alpha}(1 \pm \sqrt{\alpha})^2 = \frac{1}{\alpha}\cdot\alpha_\pm\ , \tag{177}$$

we continue

$$\begin{aligned}\frac{1}{\alpha}\,\mathbb{E}_{\sigma\sim\nu_{1/\alpha}}\left[\varphi\left(\alpha\sigma\right)\right] &= \int_{\alpha_-/\alpha}^{\alpha_+/\alpha} \varphi(\alpha\sigma)\frac{\sqrt{\left(\alpha_+ - \alpha\sigma\right)\left(\alpha\sigma - \alpha_-\right)}}{2\pi\alpha(\alpha\sigma)}\,\alpha\,\mathrm{d}\sigma = \left|\begin{array}{l}\zeta := \alpha\sigma; \\ \mathrm{d}\zeta = \alpha\mathrm{d}\sigma\end{array}\right| = \\ &= \int_{\alpha_-}^{\alpha_+} \varphi(\zeta)\frac{\sqrt{\left(\alpha_+ - \zeta\right)\left(\zeta - \alpha_-\right)}}{2\pi\alpha\zeta}\,\mathrm{d}\zeta = \int \varphi(\zeta)\,\rho_\alpha(\mathrm{d}\zeta)\ .\end{aligned} \tag{178}$$

Combining results from Eqs. (175) and (178) with Eq. (174), we can state for any $\alpha \in (0, \psi]$ that

$$\frac{1}{p}\,\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\varphi\left(\frac{1}{n}\mathbf{\Lambda}_r^2\right)\right\}\right] \xrightarrow{n,p\to\infty,\ p/n\to\alpha} \int \varphi(\sigma)\,\rho_\alpha(\mathrm{d}\sigma)\ . \tag{179}$$

D.2.1. Gradient flow test error

With findings above, we can give approximate estimate for GF and SGF generalization error. Starting with the former, i.e., Eq. (137), we can provide asymptotic limits:

$$\frac{1}{d}\,\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{\Lambda}}\left[e^{-\frac{2}{n}\mathbf{\Lambda}_r^2 t}\right]\right\} = \left|\varphi(\sigma) := e^{-2\sigma t}\right| = \frac{p/n}{d/n}\cdot\frac{1}{p}\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\varphi\left(\frac{\mathbf{\Lambda}_r^2}{n}\right)\right\}\right] \xrightarrow[p/n\to\alpha]{n,p\to\infty} \frac{\alpha}{\psi}\int e^{-2\sigma t}\,\rho_\alpha(\mathrm{d}\sigma) \tag{180}$$

and

$$\begin{aligned}\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{\Lambda}}\left[\mathbf{\Lambda}_r^{-2}\left(\mathbb{I}_r - e^{-\frac{1}{n}\mathbf{\Lambda}_r^2 t}\right)^2\right]\right\} &= \left|\varphi(\sigma) := \frac{\left(1 - e^{-\sigma t}\right)^2}{\sigma}\right| = \\ &= \frac{p}{n}\cdot\frac{1}{p}\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\varphi\left(\frac{\mathbf{\Lambda}_r^2}{n}\right)\right\}\right] \xrightarrow[p/n\to\alpha]{n,p\to\infty} \alpha\int \frac{\left(1 - e^{-\sigma t}\right)^2}{\sigma}\,\rho_\alpha(\mathrm{d}\sigma)\ ,\end{aligned} \tag{181}$$

which results in

$$\begin{aligned}\lim_{\substack{n,d,p\to\infty \\ d/n\to\psi,\ p/n\to\alpha}} \mathcal{E}_{\text{test}}^{\text{GF}}(t) = \frac{1}{2}\Bigg(&\frac{1}{\psi}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2\left(\max\left(0,\ \alpha - 1\right) + \alpha\int e^{-2\sigma t}\,\rho_\alpha(\mathrm{d}\sigma)\right) + \\ &+ \left(\left(1 - \frac{\alpha}{\psi}\right)\|\boldsymbol{\beta}\|^2 + \mu^2\right)\left[1 + \alpha\int \frac{\left(1 - e^{-\sigma t}\right)^2}{\sigma}\,\rho_\alpha(\mathrm{d}\sigma)\right]\Bigg)\ ;\end{aligned} \tag{182}$$

this coincides with Eq. (29) when $\|\boldsymbol{\beta}\| = 1$ and $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2 \to 2$.

To obtain an infinite-time limit for the test error, we can see that for $\alpha \neq 1$,

$$\begin{aligned} &\int e^{-2\sigma t}\, \rho_\alpha(\mathrm{d}\sigma) \leq e^{-2\alpha_- t} \left(\int \rho_\alpha(\mathrm{d}\sigma) \right) \xrightarrow{t\to+\infty} 0 \; ; \\ &\int \frac{e^{-k\sigma t}}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) \leq e^{-k\alpha_- t} \left(\int \frac{\rho_\alpha(\mathrm{d}\sigma)}{\sigma} \right) \xrightarrow{t\to+\infty} 0 \; , \; k \in \mathbb{N} \; . \end{aligned} \tag{183}$$

Therefore, the first integral term of Eq. (182) will go away, and in the integral second term, we can set aside a part independent of t:

$$\alpha \int \frac{(1-e^{-\sigma t})^2}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) = \alpha \int \frac{1}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) + \alpha \int \frac{e^{-2\sigma t} - 2e^{-\sigma t}}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) \; , \tag{184}$$

which will be the only one remaining when $t \to +\infty$ and can be simplified as

$$\alpha \int \frac{1}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) = \alpha \cdot \int_{\alpha_-}^{\alpha_+} \frac{\sqrt{(\alpha_+ - \sigma)(\sigma - \alpha_-)}}{2\pi\alpha\sigma^2} \mathrm{d}\sigma = \alpha \cdot \begin{cases} \frac{1}{1-\alpha} \; , & \alpha \leq 1 \; , \\ \frac{1}{\alpha-1} - \frac{1}{\alpha} \; , & \alpha > 1 \; . \end{cases} = \begin{cases} \frac{1}{1-\alpha} - 1 \; , & \alpha \leq 1 \; , \\ \frac{1}{\alpha-1} \; , & \alpha > 1 \; . \end{cases} \tag{185}$$

Now, as

$$1 + \alpha \int \frac{1}{\sigma}\, \rho_\alpha(\mathrm{d}\sigma) = \begin{cases} \frac{1}{1-\alpha} \; , & \alpha \leq 1 \; , \\ \frac{\alpha}{\alpha-1} \; , & \alpha > 1 \end{cases} = \frac{1}{1 - \min\left(\alpha, {}^1\!/\!_\alpha\right)} \; , \tag{186}$$

we retrieve a GF test error limit when $t \to +\infty$:

$$\lim_{t\to+\infty} \lim_{\substack{n,d,p\to\infty \\ d/n\to\psi,\, p/n\to\alpha}} \mathcal{E}_{\text{test}}^{\text{GF}}(t) = \frac{1}{2} \left(\frac{1}{\psi} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^0\|^2 \max\left(0,\; \alpha - 1\right) + \left(\left(1 - \frac{\alpha}{\psi} \right) \|\boldsymbol{\beta}\|^2 + \mu^2 \right) \frac{1}{1 - \min\left(\alpha, {}^1\!/\!_\alpha\right)} \right) . \tag{187}$$

D.2.2. Stochastic gradient flow test error

Now, we are going to SGF, i.e., the stochastic component additive to GF error from Eq. (164). There, however, we have to deal with expressions of form $\mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_1\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \mathrm{Tr}\left\{ \varphi_2\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right]$ for $\varphi_1, \varphi_2 : \mathbb{R} \mapsto \mathbb{R}$. However, for large n, p, we can assume that the (normalized) traces concentrate so that:

$$\mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_1\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \mathrm{Tr}\left\{ \varphi_2\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right] \approx \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_1\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right] \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_2\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right] \; . \tag{188}$$

Then

$$\begin{aligned} &\frac{1}{p^2} \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \frac{1}{n}\boldsymbol{\Lambda}_r^2 \cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 \tau} \right\} \mathrm{Tr}\left\{ \frac{1}{n}\boldsymbol{\Lambda}_r^2 \cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 (t-\tau)} \right\} \right] \\ &\quad = \left| \varphi_1(\sigma) := \sigma e^{-2\sigma\tau}, \varphi_2(\sigma) := \sigma e^{-2\sigma(t-\tau)} \right| \\ &\quad \approx \frac{1}{p} \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_1\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right] \cdot \frac{1}{p} \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ \varphi_2\left(\frac{1}{n}\boldsymbol{\Lambda}_r^2\right) \right\} \right] \\ &\quad \xrightarrow[p/n\to\alpha]{n,p\to\infty} \int \sigma_1 e^{-2\sigma_1\tau}\, \rho_\alpha(\mathrm{d}\sigma_1) \cdot \int \sigma_2 e^{-2\sigma_2(t-\tau)}\, \rho_\alpha(\mathrm{d}\sigma_2) \; . \end{aligned} \tag{189}$$

Similarly, by taking $\varphi_1(\sigma) := e^{-2\sigma\tau}, \varphi_2(\sigma) := \sigma e^{-2\sigma(t-\tau)}$, we get:

$$\frac{1}{p^2} \mathbb{E}_{\boldsymbol{\Lambda}} \left[\mathrm{Tr}\left\{ e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 \tau} \right\} \mathrm{Tr}\left\{ \frac{1}{n}\boldsymbol{\Lambda}_r^2 \cdot e^{-\frac{2}{n}\boldsymbol{\Lambda}_r^2 (t-\tau)} \right\} \right] \xrightarrow[p/n\to\alpha]{n,p\to\infty} \int e^{-2\sigma_1\tau} \rho_\alpha(\mathrm{d}\sigma_1) \cdot \int \sigma_2 e^{-2\sigma_2(t-\tau)} \rho_\alpha(\mathrm{d}\sigma_2) \; . \tag{190}$$

We can now integrate over τ by changing order of integration:

$$\begin{aligned}
&\frac{1}{p^2}\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{e^{-\frac{2}{n}\mathbf{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\frac{1}{n}\mathbf{\Lambda}_r^2\cdot e^{-\frac{2}{n}\mathbf{\Lambda}_r^2(t-\tau)}\right\}\right]\\
&\quad\xrightarrow[p/n\to\alpha]{n,p\to\infty}\int_0^t \mathrm{d}\tau\int\sigma_1 e^{-2\sigma_1\tau}\rho_\alpha(\mathrm{d}\sigma_1)\int\sigma_2 e^{-2\sigma_2(t-\tau)}\rho_\alpha(\mathrm{d}\sigma_2)\\
&\quad=\int\sigma_1\,\rho_\alpha(\mathrm{d}\sigma_1)\int e^{-2\sigma_2 t}\sigma_2\rho_\alpha(\mathrm{d}\sigma_2)\int_0^t \mathrm{d}\tau\, e^{-2(\sigma_1-\sigma_2)\tau}\\
&\quad=\int\sigma_1\,\rho_\alpha(\mathrm{d}\sigma_1)\int e^{-2\sigma_2 t}\sigma_2\rho_\alpha(\mathrm{d}\sigma_2)\,\frac{e^{-2(\sigma_1-\sigma_2)t}-1}{-2(\sigma_1-\sigma_2)}\\
&\quad=\int\int\rho_\alpha(\mathrm{d}\sigma_1)\,\rho_\alpha(\mathrm{d}\sigma_2)\,\sigma_1\sigma_2\frac{e^{-2\sigma_1 t}-e^{-2\sigma_2 t}}{2(\sigma_2-\sigma_1)}\ ,
\end{aligned}\tag{191}$$

In the same way:

$$\begin{aligned}
&\frac{1}{p^2}\int_0^t \mathrm{d}\tau\, \mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\frac{1}{n}\mathbf{\Lambda}_r^2\cdot e^{-\frac{2}{n}\mathbf{\Lambda}_r^2\tau}\right\}\mathrm{Tr}\left\{\frac{1}{n}\mathbf{\Lambda}_r^2\cdot e^{-\frac{2}{n}\mathbf{\Lambda}_r^2(t-\tau)}\right\}\right]\\
&\quad\xrightarrow[p/n\to\alpha]{n,p\to\infty}\int_0^t \mathrm{d}\tau\int e^{-2\sigma_1\tau}\rho_\alpha(\mathrm{d}\sigma_1)\cdot\int\sigma_2 e^{-2\sigma_2(t-\tau)}\rho_\alpha(\mathrm{d}\sigma_2)\\
&\quad=\int\,\rho_\alpha(\mathrm{d}\sigma_1)\int e^{-2\sigma_2 t}\sigma_2\rho_\alpha(\mathrm{d}\sigma_2)\int_0^t \mathrm{d}\tau\, e^{-2(\sigma_1-\sigma_2)\tau}\\
&\quad=\int\,\rho_\alpha(\mathrm{d}\sigma_1)\int e^{-2\sigma_2 t}\sigma_2\rho_\alpha(\mathrm{d}\sigma_2)\,\frac{e^{-2(\sigma_1-\sigma_2)t}-1}{-2(\sigma_1-\sigma_2)}\\
&\quad=\int\int\rho_\alpha(\mathrm{d}\sigma_1)\,\rho_\alpha(\mathrm{d}\sigma_2)\,\sigma_2\frac{e^{-2\sigma_1 t}-e^{-2\sigma_2 t}}{2(\sigma_2-\sigma_1)}\ .
\end{aligned}\tag{192}$$

Finally, $\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\mathbb{I}_r-e^{-\frac{2}{n}\mathbf{\Lambda}_r^2 t}\right\}\right]$ can be treated as seen before in Eq. (179):

$$\frac{1}{p}\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\mathbb{I}_r-e^{-\frac{2}{n}\mathbf{\Lambda}_r^2 t}\right\}\right]=\left|\varphi(\sigma):=1-e^{-2\sigma t}\right|=\frac{1}{p}\mathbb{E}_{\mathbf{\Lambda}}\left[\mathrm{Tr}\left\{\varphi\left(\frac{\mathbf{\Lambda}_r^2}{n}\right)\right\}\right]\xrightarrow[p/n\to\alpha]{n,p\to\infty}\int\left(1-e^{-2\sigma t}\right)\rho_\alpha(\mathrm{d}\sigma)\ .\tag{193}$$

Then, if we denote

$$F_1(\alpha,t)\equiv\int\int\rho_\alpha(\mathrm{d}\sigma_1)\rho_\alpha(\mathrm{d}\sigma_2)\sigma_1\sigma_2 K(t,\sigma_1,\sigma_2)\ ,\tag{194}$$

$$\begin{aligned}
F_2(\alpha,t)&\equiv\int\int\rho_\alpha(\mathrm{d}\sigma_1)\rho_\alpha(\mathrm{d}\sigma_2)\sigma_2 K(t,\sigma_1,\sigma_2)\\
&=\frac{1}{2}\int\int\rho_\alpha(\mathrm{d}\sigma_1)\rho_\alpha(\mathrm{d}\sigma_2)(\sigma_1+\sigma_2)K(t,\sigma_1,\sigma_2)\ ,
\end{aligned}\tag{195}$$

$$K(t,\sigma_1,\sigma_2)\equiv\frac{e^{-2\sigma_1 t}-e^{-2\sigma_2 t}}{2(\sigma_2-\sigma_1)}\ ,\tag{196}$$

we obtain an asymptotic result for stochastic component of SGF error:

$$\begin{aligned}\lim_{\substack{n,d,p\to\infty\\ d/n\to\psi,\, p/n\to\alpha}} \frac{1}{n}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] &= \lim_{\substack{n,d,p\to\infty\\ d/n\to\psi,\, p/n\to\alpha}} \left\{\frac{p^2}{dn}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot F_1(\alpha,t)+\right.\\ &\left.+\left(\left(1-\frac{p}{d}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\left(\frac{p^2}{n^2}F_2(\alpha,t)+\frac{1}{2}\max\left(0,1-\frac{p}{n}\right)\cdot\frac{p}{n}\int\left(1-e^{-2\sigma t}\right)\,\rho_\alpha(\mathrm{d}\sigma)\right)\right\}\\ &=\alpha\left[\frac{\alpha}{\psi}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot F_1(\alpha,t)+\left(\left(1-\frac{\alpha}{\psi}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\right.\\ &\left.\cdot\left(\alpha\cdot F_2(\alpha,t)+\max\left(0,1-\alpha\right)\cdot\int\frac{\left(1-e^{-2\sigma t}\right)}{2}\,\rho_\alpha(\mathrm{d}\sigma)\right)\right],\end{aligned} \tag{197}$$

or, with scaling

$$\frac{\gamma}{2}=\frac{\gamma'}{2d}=\frac{\gamma'}{2\psi n}\,, \tag{198}$$

the expression is

$$\begin{aligned}\lim_{\substack{n,d,p\to\infty\\ d/n\to\psi,\, p/n\to\alpha}} \frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right] &= \frac{\gamma'}{2}\cdot\frac{\alpha}{\psi}\left[\frac{\alpha}{\psi}\cdot\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\cdot F_1(\alpha,t)+\left(\left(1-\frac{\alpha}{\psi}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\right.\\ &\left.\cdot\left(\alpha\cdot F_2(\alpha,t)+\max\left(0,1-\alpha\right)\cdot\int\frac{\left(1-e^{-2\sigma t}\right)}{2}\,\rho_\alpha(\mathrm{d}\sigma)\right)\right],\end{aligned} \tag{199}$$

which matches Eq. (30) when $\|\boldsymbol{\beta}\|=1$ and $\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^0\|^2\to 2$.

Now, we obtain infinite-time limit. With argument similar to Eq. (183), we can write for $\alpha\neq 1$:

$$\lim_{t\to+\infty}F_1(\alpha,t)=0,\ \lim_{t\to+\infty}F_2(\alpha,t)=0\,; \tag{200}$$

besides, from Eq. (172) we can see:

$$\int\rho_\alpha(\mathrm{d}\sigma)=\min\left(1,\frac{1}{\alpha}\right)\,, \tag{201}$$

and as

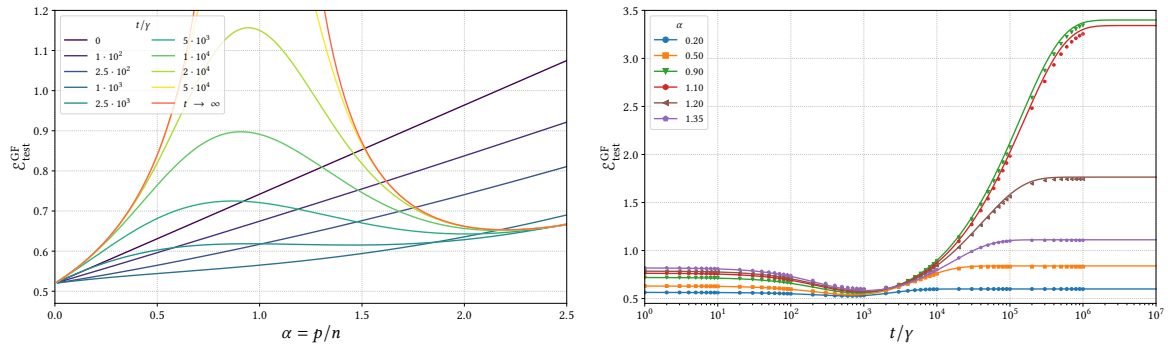
$$\max\left(0,1-\alpha\right)\cdot\min\left(1,\frac{1}{\alpha}\right)=\max\left(0,1-\alpha\right), \tag{202}$$

the infinite-time limit of difference between test errors of SGF and GF is:

$$\lim_{t\to+\infty}\lim_{\substack{n,d,p\to\infty\\ d/n\to\psi,\, p/n\to\alpha}}\frac{\gamma}{2}\mathbb{E}_{\mathcal{A};\boldsymbol{X},\boldsymbol{y}}\left[\mathrm{Tr}\{\boldsymbol{\mathscr{C}}(t)\}\right]=\frac{\gamma'}{4}\cdot\frac{\alpha}{\psi}\left(\left(1-\frac{\alpha}{\psi}\right)\|\boldsymbol{\beta}\|^2+\mu^2\right)\cdot\max\left(0,1-\alpha\right)\,. \tag{203}$$

E. Additional numerics for the weak features model

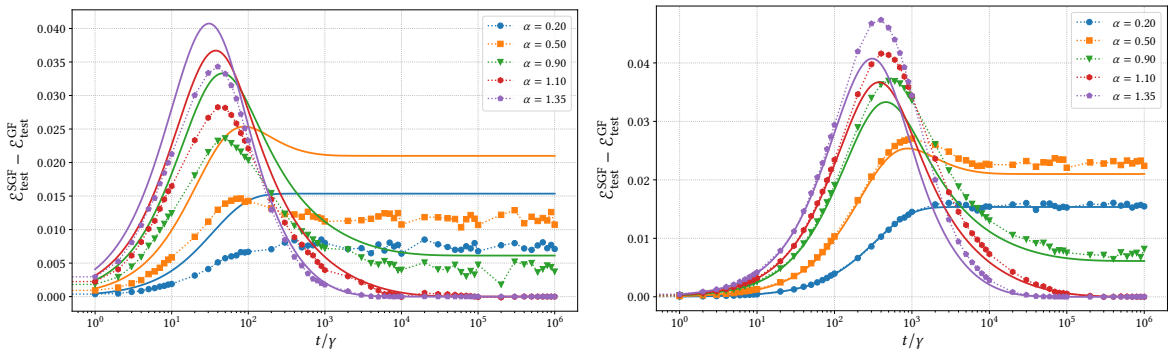
Gradient flow – Plots for the test risk of deterministic gradient flow estimator are found in Fig. 5. Fig. 5(a) focuses on the risk dependence on α. One can see a limiting double descent picture (with degenerate first descent) described in (Belkin et al., 2020). Besides, our derivations, particularly Eq. (29), also allows to see the emergence of the double descent with time. This is illustrated by Fig. 5(b) which depicts the time evolution of the test risk for GF estimator. We observe that under our choice of model parameters early stopping is beneficial for most values of α. The gradient descent simulations come very close to our prediction. The discrepancy between the theoretical and empirical predictions is slightly more pronounced for values of α close to 1, which we hypothesize to be caused by a lower rate of convergence of the empirical distribution of eigenvalues of $\boldsymbol{\Lambda}_r^2/n$ to Marchenko-Pastur distribution for such α's.



(a) (b)

Figure 5. Plots depicting $\mathcal{E}^{\mathrm{GF}}_{\mathrm{test}}(t)$ for $\psi = 2.5$. Solid lines represent theoretical asymptotic estimation. Markers on 5(b) represent the results of numerical simulations of GD (averaged over 1000 different random subsets $\mathcal{A}$). Simulation parameters: $d = 1000$, $\gamma = 10^{-3}$; vectors $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ are taken at random on the unit 1000-dimensional sphere and here $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 \approx 2.11$

Difference between SGF and GF – We illustrate the time-evolution of the difference $\mathcal{E}^{\mathrm{SGF}}_{\mathrm{test}} - \mathcal{E}^{\mathrm{GF}}_{\mathrm{test}}$ and its numerical simulation counterpart $\mathcal{E}^{\mathrm{SGD}}_{\mathrm{test}} - \mathcal{E}^{\mathrm{GD}}_{\mathrm{test}}$, as well as finite-size effects. These differences are presented in Fig. 6 as a function of t/γ for different values of α, below and above the interpolation threshold. Note that t/γ can be interpreted as teh number of iterations in gradient descent. Fig. 6(a) depicts the comparison with simulations for $d = 100$ and Fig. 6(b) with simulations for $d = 1000$.



(a) Simulations with $d = 100$ (b) Simulations with $d = 1000$

Figure 6. Time-evolution of $\mathcal{E}^{\mathrm{SGF}}_{\mathrm{test}} - \mathcal{E}^{\mathrm{GF}}_{\mathrm{test}}$ with $\psi = 2.5$ for different values of α. Markers represent simulations while lines represent the theoretical predictions. The learning rate is scaled as $\gamma = 1/d$. Vectors $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ are taken at random on the unit d-dimensional sphere. Here $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 \approx 1.78$ for $d = 100$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|^2 \approx 2.11$ for $d = 1000$.

We note pronounced finite-size effects when comparing Figs. 6(a) (with $d = 100$) and Figs. 6(b) (with $d = 1000$), particularly below the interpolation threshold and high values of t/γ.

Additionally, for finite-times, we observe that the theoretical results are less accurate for times of order $t \sim 1$ (corresponding to roughly $1/\gamma = d$ GD iterations). That is already apparent from the 'red curve' in Fig. 3 in the main text. To further investigate this point, we define $\Delta\mathcal{E}^{\mathrm{Theory}}_{\mathrm{test}} \equiv \mathcal{E}^{\mathrm{SGF}}_{\mathrm{test}} - \mathcal{E}^{\mathrm{GF}}_{\mathrm{test}}$ and $\Delta\mathcal{E}^{\mathrm{Simul}}_{\mathrm{test}} \equiv \mathcal{E}^{\mathrm{SGD}}_{\mathrm{test}} - \mathcal{E}^{\mathrm{GD}}_{\mathrm{test}}$, and compare the difference between $\Delta\mathcal{E}^{\mathrm{Theory}}_{\mathrm{test}} - \Delta\mathcal{E}^{\mathrm{Simul}}_{\mathrm{test}}$ (for the same setting as that of Fig. 6). These differences are presented in Fig. 7 for different values of α, and with $d = 100$, $d = 1000$.

For values of α well below the interpolation threshold, the finite size effects are clearly visible in the whole time-evolution and go away as d increases. Above the threshold, the discrepancy is more pronounced around $t \sim 1$ and there is no clear evidence that this is a finite size effect. It remains to be seen if better continuous time modelling (for example second order

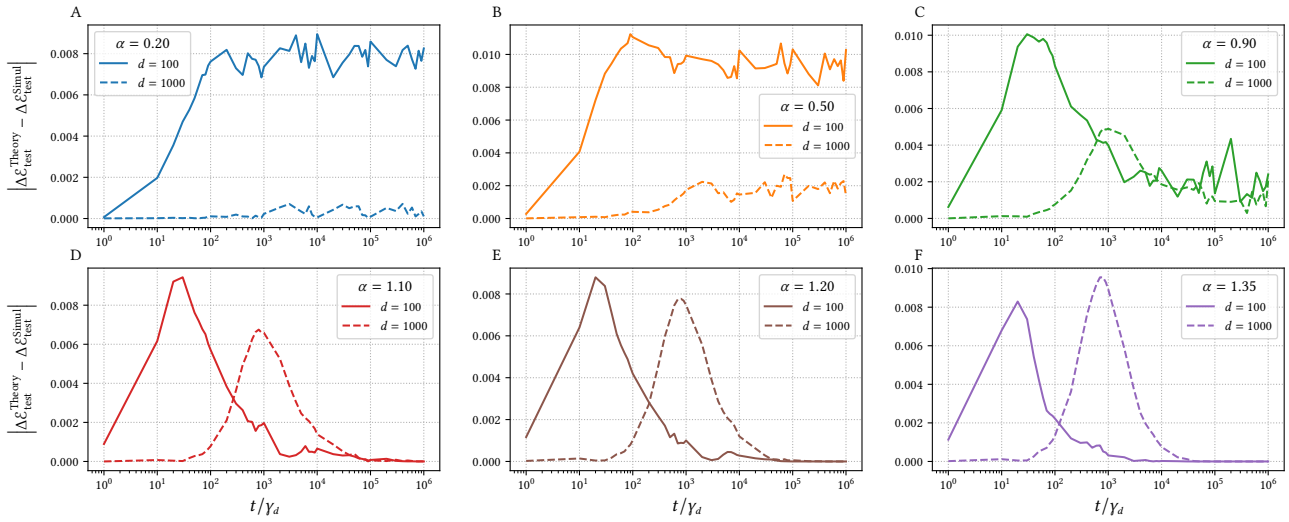


Figure 7. Time-evolution of the difference between the theory and simulation ($d = 100, 1000$ and $\psi = 2.5$) results for $\Delta\mathcal{E}_{\text{test}}^{\text{Theory}} - \Delta\mathcal{E}_{\text{test}}^{\text{Simul}}$ (defined in the text) for different values of α, below and above the interpolation threshold. The time scaling is adjusted according to the fine size simulation: $\gamma_d = {}^1\!/_d$.

SDE modelling) would be tractable and make the discrepancy disappear.