SGD vs GD: Rank Deficiency in Linear Networks

Aditya Varre
EPFL
aditya.varre@epfl.ch

Margarita Sagitova
EPFL
margarita.sagitova@epfl.ch

Nicolas Flammarion
EPFL
nicolas.flammarion@epfl.ch

Abstract

In this article, we study the behaviour of continuous-time gradient methods on a two-layer linear network with square loss. A dichotomy between SGD and GD is revealed: GD preserves the rank at initialization while (label noise) SGD diminishes the rank regardless of the initialization. We demonstrate this rank deficiency by studying the time evolution of the *determinant* of a matrix of parameters. To further understand this phenomenon, we derive the stochastic differential equation (SDE) governing the eigenvalues of the parameter matrix. This SDE unveils a *replusive force* between the eigenvalues: a key regularization mechanism which induces rank deficiency. Our results are well supported by experiments illustrating the phenomenon beyond linear networks and regression tasks.

1 Introduction

Deep neural networks have significantly advanced machine learning in recent decades. A key attribute of these models is their ability, despite being heavily overparameterized, to learn effective representations which generalizes well across different tasks. This capability has sparked substantial interest in understanding how neural networks learn internal representations for specific tasks [Bengio et al., 2013]. Gaining deeper insights into these mechanisms is crucial for enhancing model interpretability and refining training and application methodologies in real-world scenarios.

The success in learning these representations is often attributed to the gradient methods used in training. These methods navigate complex non-convex landscapes, finding solutions that not only minimize the training objective but also yield effective representations. They achieve this generalization while avoiding the spurious features that could potentially arise from the models' large number of parameters. Empirical studies have shown that the stochastic noise in gradient algorithms enhances generalization [Keskar et al., 2017] by favoring solutions with simpler structures that mitigate spurious features [Andriushchenko et al., 2022]. This paper address the overarching question:

How does stochasticity facilitate the discovery of solutions with simplified structures?

We explore this question using a simplified model: a single hidden-layer linear network. Despite lacking non-linearity, such networks capture some intricate phenomena of real-world deep networks and have been extensively studied to understand convergence [Arora et al., 2019a, Min et al., 2021], learning dynamics [Saxe et al., 2014], and the implicit bias of optimization algorithms [Gunasekar et al., 2017, Soudry et al., 2018]. Our work builds on this foundation by comparing stochastic algorithms with their deterministic counterparts, focusing on how these differences influence the learning of simpler structures.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Specifically, we analyze vector regression on two-layer linear networks trained with both gradient flow and stochastic gradient flow methods. Our contributions include:

- In Section 4, we track the evolution of the determinant of the parameter matrix under gradient flow and stochastic gradient flow. We show that stochastic gradient flow drives the determinant towards zero, effectively removing irrelevant direction(s).
- In Section 5, we derive a stochastic differential equation that describes the behavior of the eigenvalues of the parameter matrix. This analysis reveals a repulsive force between eigenvalues that pushes them apart and a geometric Brownian motion that pulls them toward zero.
- In Section 6, we discuss the generalizability of our approach beyond square loss and various noise models, including discrete step sizes. Finally, we present experimental results in Section 7 that support our theoretical findings.

2 Related Work

Our work lies at the convergence of distinct research topics:

Effect of SGD on generalization. The relationship between the stochasticity of SGD and its generalization capabilities has been extensively examined [Mandt et al., 2016, Jastrzebski et al., 2018, He et al., 2019, Hoffer et al., 2017, Kleinberg et al., 2018]. Notably, SGD tends to yield models with superior generalization compared to gradient descent [Keskar et al., 2017, Jastrzebski et al., 2018, He et al., 2019]. Various explorations into this phenomenon have been conducted through various approaches: hypothesizing that SGD favors flatter minima linked to better generalization, as opposed to sharp minima associated with poor generalization [Hochreiter and Schmidhuber, 1997, Keskar et al., 2017, Andriushchenko et al., 2023], using a random walk on a random landscape model to understand the impact of stochasticity [Hoffer et al., 2017], proposing that the inherent noise in SGD smooths the loss landscape [Kleinberg et al., 2018], and exploring the implications of dynamical stability [Wu et al., 2018].

Stochastic dynamics and Label Noise. Recent literature has explored label noise-driven Gradient Descent as an effective method to probe the beneficial impact of stochasticity on generalization, with two distinct perspectives emerging. Firstly, an asymptotic view on general model parametrization is considered, where Blanc et al. [2020], Damian et al. [2021] suggest that stochastic dynamics preferentially optimize a hidden objective linked to the curvature of the loss. In a related vein, Li et al. [2021] demonstrates appropriate limiting dynamics on the manifold of interpolators through time rescaling. Secondly, specifically for diagonal linear networks, HaoChen et al. [2021], Pillaud-Vivien et al. [2022] observe a similar collapsing effect due to label noise but with a finer characterization of the limiting process. Finally, in the absence of label noise, Pesme et al. [2021], Even et al. [2023] have characterized the outcomes of stochastic GF and GD for diagonal linear networks as the solutions to an implicit regularization problem that results in sparser solutions than without stochasticity. Recently, Ghosh et al. [2023] further exhibit a similar sparser features effect for single-neuron autoencoder. Chen et al. [2023] provides a condition under which an invariant set is attractive for SGD — characterizing the local behavior around these sets. The paper also studies linear networks in a teacher-student setup, however due to structured label-noise [Chen et al., 2023, A2 in p.30], the analysis falls short of capturing the repulsive force in the singular values.

Linear Networks. The study of two-layer linear networks has been explored extensively, particularly when optimized using gradient flow on the square loss, across various settings including zero-balance initialization and whitened data Fukumizu [1998], Saxe et al. [2014, 2019], Braun et al. [2022]. Early work by Saxe et al. [2014, 2019] elucidates the temporal changes in the singular values of the predictor, assuming decoupled dynamics and a specific data-dependent weight initialization. This condition is broadened by the analyses of Fukumizu [1998] and Braun et al. [2022], Tarmoun et al. [2021], who apply solutions from a matrix Riccati equation to characterize the weights dynamics under full-rank network initialization. Furthermore, Gidel et al. [2019] extends the existing framework by relaxing the whitened data assumption, conducting a perturbation analysis, and discussing the temporal evolution of the weight matrices' singular values. Additionally, Varre et al. [2024] eliminates the need for zero-balanced and full-rank initializations. Their study provides detailed formulas for weight evolution as a function of the initial scale , also studies a simple version of a stochastic flow without the drift. Wang and Jacot [2023] studied the implicit bias of SGD with ℓ_2-regularization.

Matrix valued stochastic process and their eigenvalues. Stochastic process on the space of symmetric (or Hermitian) matrices and the evolution of their eigenvalues are well studied since Dyson [1962]. These techniques were further developed by Bru [1989, 1991] to study perturbations of principal component analysis and the eigenvalues of Wishart processes. Norris et al. [1986], Graczyk and Małecki [2013] applied SDE-based techniques to study the eigenvalues and eigenvectors of Brownian motion on ellipsoids.

3 Linear networks and continuous-time gradient method

Notation We use $\langle .,. \rangle$ to denote the inner product, i.e., $\langle u, v\rangle = u^\top v$ for vectors, and $\langle A, B\rangle = \mathrm{Tr}\left(AB^\top\right)$ for matrices. I_d denotes the identity matrix of dimension d and $0_{p\times k}$ denote the matrix with all zero entries of dimension $p \times k$.

Vector regression. We study the vector regression problems with inputs $x_1, \dots, x_n$ in $(\mathbb{R}^p)^n$ and outputs $y_1, \dots, y_n$ in $(\mathbb{R}^k)^n$. We consider the minimization of the square loss over a class of parametric models $\mathcal{H} = \{f_\theta(\cdot) : \mathbb{R}^p \to \mathbb{R}^k \mid \theta \in \mathbb{R}^d\}$ specified in the next paragraph. The train loss therefore can be written as $\mathcal{L}\left(\theta\right) = \frac{1}{2n}\sum_{i=1}^n \|y_i - f_\theta(x_i)\|^2$.

Parameterization with a linear network. We focus on two-layer linear neural networks of width $l \in \mathbb{N}^*$. The model is described by the parameterization $\theta = (\mathbf{W}_1, \mathbf{W}_2)$, where $\mathbf{W}_1 \in \mathbb{R}^{p\times l}$ and $\mathbf{W}_2 \in \mathbb{R}^{l\times k}$, and the function $f_\theta(x) = \mathbf{W}_2^\top \mathbf{W}_1^\top x$. This model is linear with respect to the input x. In terms of expressivity, it is comparable to the linear class of predictors, represented as $f_{\boldsymbol{\beta}}(x) = \boldsymbol{\beta}^\top x$, where $\boldsymbol{\beta}$ equals $\mathbf{W}_1\mathbf{W}_2$. Throughout our analysis, we denote the equivalent linear predictor of the network as $\boldsymbol{\beta}$. A key aspect of this parametrization is that the prediction function f_θ is positive homogeneous of degree 2 with respect to θ: specifically, for any $\lambda \in \mathbb{R}$, $f_{\lambda\theta} = \lambda^2 f_\theta$. This property mirrors that of two-layer ReLU networks and significantly influences the loss landscape navigated by the parameters θ. It is important to note that this parameterization introduces some redundancy, a single linear predictor $\boldsymbol{\beta}$ can have multiple representations $\mathbf{W}_1, \mathbf{W}_2$ such that $\mathbf{W}_1\mathbf{W}_2 = \boldsymbol{\beta}$. Some representations have a rich structure whereas other resemble random features. For example, consider the case of scalar regression ($k = 1$), for a vector $\boldsymbol{\beta}$ there exists rich parameterizations where all the neurons, i.e., columns of $\mathbf{W}_1$ align with $\boldsymbol{\beta}$ and also some lazy structures where $\mathbf{W}_1$ resembles a random matrix [Chizat et al., 2019, Varre et al., 2023].

Train loss. By defining $X^\top = [x_1, \dots, x_n]$ and $Y^\top = [y_1, \dots, y_n]$, the loss function is given by:

$$\mathcal{L}\left(\mathbf{W}_1, \mathbf{W}_2\right) = \frac{1}{2n}\|X\mathbf{W}_1\mathbf{W}_2 - Y\|^2. \tag{3.1}$$

For simplicity, we adjust for the normalization factor n by rescaling the data to $(X, Y) \leftarrow (X/\sqrt{n}, Y/\sqrt{n})$, thereby implicitly considering it in the loss function without directly mentioning n in the formula. Note that the loss is non-convex in $\mathbf{W}_1, \mathbf{W}_2$.

Gradient flow. The dynamics induced in parameter space by running GF on Equation (3.1) is given by

$$\mathrm{d}\mathbf{W}_1 = -\nabla_{\mathbf{W}_1}\mathcal{L}\left(\mathbf{W}_1, \mathbf{W}_2\right)\mathrm{d}t = X^\top(Y - X\mathbf{W}_1\mathbf{W}_2)\mathbf{W}_2^\top \mathrm{d}t, \tag{3.2}$$

$$\mathrm{d}\mathbf{W}_2 = -\nabla_{\mathbf{W}_2}\mathcal{L}\left(\mathbf{W}_1, \mathbf{W}_2\right)\mathrm{d}t = \mathbf{W}_1^\top X^\top(Y - X\mathbf{W}_1\mathbf{W}_2)\mathrm{d}t. \tag{3.3}$$

Introducing the block matrix, $\boldsymbol{\Theta} = \left[\mathbf{W}_1^\top \mid \mathbf{W}_2\right] \in \mathbb{R}^{l\times(p+k)}$ and denoting the residual matrix by $\mathbf{R} = X^\top(Y - X\mathbf{W}_1\mathbf{W}_2)$, the evolution of $\boldsymbol{\Theta}$ can be written as

$$\mathrm{d}\boldsymbol{\Theta} = \left[\mathrm{d}\mathbf{W}_1{}^\top \mid \mathrm{d}\mathbf{W}_2\right] = \left[\mathbf{W}_2\mathbf{R}^\top\mathrm{d}t \mid \mathbf{W}_1^\top\mathbf{R}\mathrm{d}t\right] = \left[\mathbf{W}_1^\top \mid \mathbf{W}_2\right]\begin{bmatrix}0_{p\times p} & \mathbf{R}\\ \mathbf{R}^\top & 0_{k\times k}\end{bmatrix}\mathrm{d}t.$$

The gradient flow can therefore be compactly written as

$$\mathrm{d}\boldsymbol{\Theta} = \boldsymbol{\Theta}\mathbf{J}\mathrm{d}t, \quad \text{where } \mathbf{J} = \begin{bmatrix}0_{p\times p} & \mathbf{R}\\ \mathbf{R}^\top & 0_{k\times k}\end{bmatrix}. \tag{3.4}$$

The gradient flow (GF), when expressed in this form, reveals an inherent multiplicative structure with respect to $\boldsymbol{\Theta}$ in the gradient of the loss. As we see in subsequent sections, this representation of the gradient flow with block matrices proves to be very convenient.

Label noise gradient descent. Label noise gradient descent (LNGD) is a theoretically studied alternative to SGD that mirrors its practical behavior by sharing the geometric properties of the noise Blanc et al. [2020], Damian et al. [2021]. Let $\varepsilon_t \in \mathbb{R}^{n\times k}$, where each entry of ε_t is an independent Gaussian random variable. At iteration t, the labels are perturbed with this Gaussian noise at an intensity δ, i.e., $\widetilde{Y} = Y + \sqrt{\delta}\varepsilon_t$. The LNGD algorithm updates the iterates with a step size η in the direction of the gradient computed after the labels have been perturbed, as follows:

$$\mathbf{W}_1^{t+1} = \mathbf{W}_1^t - \eta \nabla_{\mathbf{W}_1} \mathcal{L}\left(\widetilde{Y}, \mathrm{X}, \mathbf{W}_1^t, \mathbf{W}_2^t\right); \quad \mathbf{W}_2^{t+1} = \mathbf{W}_2^t - \eta \nabla_{\mathbf{W}_2} \mathcal{L}\left(\widetilde{Y}, \mathrm{X}, \mathbf{W}_1^t, \mathbf{W}_2^t\right),$$

where, by an abuse of notation, $\mathcal{L}\left(Y, \mathrm{X}, \mathbf{W}_1, \mathbf{W}_2\right) = 1/2\|X\mathbf{W}_1\mathbf{W}_2 - Y\|^2$. The iterates can then be restructured into a block matrix:

$$\mathbf{\Theta}^{t+1} = \mathbf{\Theta}^t - \eta\mathbf{\Theta}^t\mathbf{J}_t - \eta\sqrt{\delta}\mathbf{\Theta}^t\xi_t, \quad \text{where } \xi_t = \begin{bmatrix} 0_{p\times p} & X^\top\varepsilon_t \\ \varepsilon_t^\top X & 0_{k\times k} \end{bmatrix}, \tag{3.5}$$

and J_t is defined as in Equation (3.4).

Stochastic gradient flow (SGF). We aim to model the aforementioned LNGD in continuous time using an appropriate SDE. Stochastic continuous-time counterparts of discrete stochastic gradient algorithms are favored for their enhanced amenability to theoretical analysis. We propose the following stochastic differential equation (SDE) to model LNGD in continuous time:

$$\mathrm{d}\mathbf{\Theta} = \mathbf{\Theta}\left[\mathbf{J}\mathrm{d}t + \sqrt{\eta\delta}\mathrm{d}\xi\right], \text{where } \mathrm{d}\xi = \begin{bmatrix} 0_{p\times p} & X^\top\mathrm{d}\mathbf{B}_t \\ \mathrm{d}\mathbf{B}_t{}^\top X & 0_{k\times k}, \end{bmatrix} \tag{3.6}$$

where $\mathbf{B}_t$ denotes a matrix Brownian motion in $\mathbb{R}^{n\times k}$. LNGD as defined in Equation (3.5), can be interpreted as the the Euler-Maryama discretization of the above SGF with a stepsize η. Although the inclusion of step size in the continuous-time modeling of an SDE may seem counter-intuitive, it is a necessary component [Li et al., 2019b]. As all the terms of the SDE in Equation (3.6) are polynomial in $\mathbf{\Theta}$, both the drift and diffusion terms are locally Lipschitz continuous. Hence, the solution of the SDE is uniquely defined up to the explosion time τ_∞ [see, e.g., Khasminskii, 2012]. Furthermore, the explosion time can be proven to be infinite ($\tau_\infty = \infty$ almost surely), by using that the GF does not diverge and applying the techniques outlined by Pillaud-Vivien et al. [2022, Proposition 10].

Initialization. The dynamics of gradient methods on homogeneous models are significantly influenced by initialization, which determines the regime they operate in—specifically, the lazy regime for large initializations and the rich regime for small ones [Chizat et al., 2019, Woodworth et al., 2020]. Thus, the scale of initialization has garnered significant interest, particularly its impact on the training of linear and non-linear networks with GD [Woodworth et al., 2020, Boursier et al., 2022]. It is observed that stochastic methods eliminate the dependence on initialization [Pesme et al., 2021].

Conserved quantities and balanceness. Gradient flows follow specific conservation laws along their trajectory [Marcotte et al., 2023], maintaining characteristics of the initial conditions. For linear networks, this conservation manifests as the *balanceness property* [Du et al., 2018], described by:

$$\mathbf{\Delta} = \mathbf{W}_1^\top\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_2^\top = \mathbf{W}_1^\top(0)\mathbf{W}_1(0) - \mathbf{W}_2(0)\mathbf{W}_2^\top(0).$$

As a result, Saxe et al. [2014], Arora et al. [2018, 2019b] have adopted *balanced initialization*, where $\mathbf{\Delta}(0) = 0$, to ensure that weight matrices remain low rank throughout the trajectory. However, unbalanced initialization do not preserve these simple low-rank structures, as aspects of the initial conditions persist.

In contrast, stochastic methods do not adhere to these conservation laws [Ziyin et al., 2023] and the evolution of the imbalance $\mathbf{\Delta}$ for SGF is

$$\mathrm{d}\mathbf{\Delta} = \mathrm{d}\big(\mathbf{W}_1^\top\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_2^\top\big) = \mathrm{tr}\left(XX^\top\right)\mathbf{W}_2\mathbf{W}_2^\top\mathrm{d}t - k\ \mathbf{W}_1^\top X^\top X\mathbf{W}_1\mathrm{d}t.$$

While there is no diffusion term in the derivative, the matrices remain stochastic and no definitive conclusions can be drawn from this. However, in the case where $k = p$ and $X^\top X = \mathrm{I}_p$, it can be shown that $\mathbf{W}_1^\top\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_2^\top \to 0$, indicating that the stochastic noise eliminates initial imbalance.

Conclusion. Understanding how stochastic methods mitigate dependency on initialization requires exploring beyond the evolution of the imbalance $\mathbf{\Delta}$. To this end, we identify and discuss other conserved quantities, such as the determinant of the block matrix $\mathbf{\Theta}^\top\mathbf{\Theta}$ in the following sections.

4 Separation between Gradient Flow through determinant

Here, we present our first separation result between GF and SGF. While the determinant of the parameters is preserved in GF, it is driven to zero by the stochasticity of SGF, leading to a simplistic low-rank structure.

4.1 Determinant evolution of the gradient flow

The theorem below demonstrates that the determinant of the parameters is preserved in gradient flow.

Theorem 4.1. *For the gradient flow defined in Equation (3.4), the following property holds,*

$$\mathrm{d}\big(\det\left(\mathbf{\Theta}^\top\mathbf{\Theta}\right)\big) = 0.$$

Hence, $\det\left(\mathbf{\Theta}(t)^\top\mathbf{\Theta}(t)\right) = \det\left(\mathbf{\Theta}_0^\top\mathbf{\Theta}_0\right)$*, where* $\mathbf{\Theta}_0 = \mathbf{\Theta}(0)$ *is the initialisation at time* $t = 0$*.*

The proof presented in the App. B.1, is based on straightforward computations of the derivative of the determinant and the fact that the matrix $\mathbf{J}$ has zero trace. We note that the simplicity of the proof arises from the strategically chosen block structure of $\mathbf{\Theta}$. This result would have been less straightforward with different parametrizations, which likely explains why such a simple finding appears to be novel. The theorem implies that the determinant of $\mathbf{M}$ along the trajectory remains equal to the determinant at initialization. If $\mathbf{\Theta}_0^\top\mathbf{\Theta}_0$ is full-rank initially, meaning the determinant is non-zero, the theorem ensures that the determinant of $\mathbf{M}$ remains non-zero. Consequently, the rank of $\mathbf{\Theta}$ does not diminish along the trajectory. When $l \geq p + k$, i.e., the hidden layer has a large width and $\mathbf{W}_1, \mathbf{W}_2$ are initialized randomly from a Gaussian distribution, $\mathbf{\Theta}_0^\top\mathbf{\Theta}_0$ has full rank almost surely. The theorem also reveals some implications regarding the impact of initialization scale. Note that $\lambda_{min}(A) \leq \sqrt[n]{\det A}$, indicating that when the scale of initialization is very small, at least one singular value of $\mathbf{\Theta}$ is small.

4.2 Determinant evolution of the stochastic gradient flow

In contrast, the theorem presented below demonstrates that the determinant of the parameters converges to zero in stochastic gradient flow.

Theorem 4.2. *For the SDE, defined in the Equation (3.6), for* $t \leq \tau_\infty$*, the following property holds for the evolution of determinant*

$$\mathrm{d}\big(\det\left(\mathbf{\Theta}^\top\mathbf{\Theta}\right)\big) = -2\eta\delta k \operatorname{tr}\left(X^\top X\right)\det\left(\mathbf{\Theta}^\top\mathbf{\Theta}\right)\mathrm{d}t.$$

Hence, $\det\left(\mathbf{\Theta}(t)^\top\mathbf{\Theta}(t)\right) = \det\left(\mathbf{\Theta}_0^\top\mathbf{\Theta}_0\right)\exp\big\{-2\eta\delta k \operatorname{tr}\left(X^\top X\right)t\big\}$*, where* $\mathbf{\Theta}_0$ *is the initialization.*

Although the evolution of the parameters in SGF is random, the evolution of the determinant is deterministic. The theorem highlights a striking phenomenon: the noise in SGF diminishes the determinant along the trajectory, leading to a simplification of the network over time. The larger the noise and the stepsize, the faster the determinant vanishes. The vanishing of the determinant suggests that the rank of the parameters decreases by at least one, effectively eliminating some components. It holds for any initialization of $\mathbf{\Theta}_0$ and indicates how the SGF overrides some aspects of initialization. The proof uses the fact that stochastic Brownian term in the SDE, through Itô's calculus, introduces a negative drift, ultimately driving the determinant to zero (refer to B.3 for the proof).

Limitations. Given the large width of the hidden layer, the determinant converging to zero does not fully reveal the complexity of the situation. It merely indicates that at least one singular value is approaching zero. Furthermore, the theorem provides limited insights when the determinant is already 0 at initialization, $\det \mathbf{\Theta}_0 = 0$ which happens whenever $l < p + k$. Next, we explore the mechanisms behind this low-rank phenomenon, suggesting that the repulsive forces induced by stochasticity drive the spurious singular values to zero as seen in the right plot of Figure 1.

5 Mechanism behind the low-rank phenomenon

In this section, we investigate the evolution of singular values under stochastic training to gain deeper insights into the low-rank phenomenon. To simplify the discussion, throughout the section we consider the case where $k = 1$ and for notational convenience, we let $\mathbf{W}_1 = \mathbf{W}, \mathbf{W}_2 = \mathbf{a}$. Additionally, we assume that $l \leq p$, however the results can be extended to any l.

Warm-up: Comparison with diagonal networks. Let $\mathbf{W} = \mathbf{U}\Sigma\mathbf{V}^\top$ be the singular value decomposition (assuming $l \leq p$). The predictor $\boldsymbol{\beta}$ can be expressed as

$$\mathbf{W}\mathbf{a} = \mathbf{U}\Sigma\mathbf{V}^\top\mathbf{a} = \mathbf{U}\left[\boldsymbol{\sigma} \odot \mathbf{c}\right], \text{where } \mathbf{c} = \mathbf{V}^\top\mathbf{a}.$$

This expression reveals a Hadamard product between $\boldsymbol{\sigma}$ and $\mathbf{c}$, reminiscent of diagonal networks which are widely studied to understand the nonconvex dynamics of gradient algorithms [Woodworth et al., 2020, Pesme et al., 2021, Pillaud-Vivien et al., 2022]. In the context of diagonal networks, SGD is known to provably induce sparsity in predictions. Similarly, for linear networks, SGF may induce sparsity in terms of the singular value σ. We next derive the SDE governing the evolution of the singular values Σ of the weight matrix to gain a clearer understanding of the low-rank phenomenon.

Scalar Regression. We assume that the data is isotropic, i.e., $X = \mathrm{I}_p$. Under these conditions, the loss function for scalar regression can be written as

$$\mathcal{L}\left(\mathbf{W}, \mathbf{a}\right) = \frac{1}{2}\|y - \mathbf{W}\mathbf{a}\|^2. \tag{5.1}$$

We train the above objective with SGF, formulated as follows,

$$\mathrm{d}\mathbf{W} = (y - \mathbf{W}\mathbf{a})\mathbf{a}^\top \mathrm{d}t + \sqrt{\eta\delta}\ \mathrm{d}\mathbf{B}_t\mathbf{a}^\top; \qquad \mathrm{d}\mathbf{a} = \mathbf{W}^\top(y - \mathbf{W}\mathbf{a})\mathrm{d}t + \sqrt{\eta\delta}\ \mathbf{W}^\top\mathrm{d}\mathbf{B}_t. \tag{5.2}$$

where $\mathbf{B}_t$ is the standard Brownian motion in $\mathbb{R}^p$. For analytical convenience, we rescale the time $t \to {}^{t}/_{\eta\delta}$ and use the process $\mathrm{d}\mathbf{X} = {}^{1}/_{\eta\delta}(y - \mathbf{W}\mathbf{a})\mathrm{d}t + \mathrm{d}\mathbf{B}_t$. The SGF can then be rewritten as,

$$\mathrm{d}\mathbf{W} = \ \mathrm{d}\mathbf{X}\mathbf{a}^\top; \qquad \mathrm{d}\mathbf{a} = \mathbf{W}^\top\mathrm{d}\mathbf{X}. \tag{5.3}$$

Our focus is on understanding the evolution of the singular values of the matrix $\mathbf{W}$. This aim is facilitated by considering the symmetric matrix $\mathbf{M} = \mathbf{W}^\top\mathbf{W}$, whose eigenvalues are the squares of the singular values of $\mathbf{W}$. Taking the derivative of $\mathbf{M}$, we find

$$\mathrm{d}\mathbf{M} = \mathrm{d}\mathbf{W}^\top\mathbf{W} + \mathbf{W}^\top\mathrm{d}\mathbf{W} + \mathrm{d}\mathbf{W}^\top\mathrm{d}\mathbf{W} = \mathbf{a}\mathrm{d}\mathbf{X}^\top\mathbf{W} + \mathbf{W}^\top\mathrm{d}\mathbf{X}\mathbf{a}^\top + p\mathbf{a}\mathbf{a}^\top\mathrm{d}t. \tag{5.4}$$

Note that $\mathrm{d}x\mathrm{d}y$ represents $d[x, y]$ for any continuous semi-martingales x, y [see, e.g., Ikeda and Watanabe, 1981, chapter 3 for reference].

Eigenvalues of a matrix-valued stochastic process. We leverage tools from the study of eigenvalues of matrix-valued stochastic processes [Bru, 1989, Graczyk and Małecki, 2013] to derive the evolution of the eigenvalues of $\mathbf{M}$ in the theorem that follows.

Theorem 5.1. *Let $\mathrm{s}_1 > \ldots > \mathrm{s}_l$ be the order of the eigenvalues of the matrix $\mathbf{M}$ defined by Equation (5.4). Let the collision time for the eigenvalues be defined as*

$$\tau = \{\inf t : \mathbf{s}_i(t) = \mathbf{s}_j(t)\, \mathit{for}\ 1 \leq i \neq j \leq l\}. \tag{5.5}$$

For $t \leq \tau$, the eigenvalues are semi-martingales given by the solution of the following SDE

$$\mathrm{d}(\mathbf{s}_i) = p\mathbf{c}_i^2\ \mathrm{d}t + \sum_{\substack{j=1,\\ j\neq i}}^{l} \frac{\mathbf{s}_i\mathbf{c}_j^2 + \mathbf{s}_j\mathbf{c}_i^2}{\mathbf{s}_i - \mathbf{s}_j}\mathrm{d}t + 2\sqrt{\mathbf{s}_i\mathbf{c}_i^2}\left(\mathrm{d}\tilde{\mathbf{X}}\right)_i \tag{5.6}$$

where $\mathbf{c} = \mathbf{V}^\top\mathbf{a}$ and $\left(\mathrm{d}\tilde{\mathbf{X}}\right)_i = {}^{1}/_{\eta\delta}\left(\langle\mathbf{u}_i, y\rangle - \sqrt{\mathbf{s}_i\mathbf{c}_i^2}\right)\mathrm{d}t + \mathrm{d}\varepsilon_i$ with $\mathbf{u}_i$ being the i^{th} column of $\mathbf{U}$ and $(\varepsilon_0, \ldots, \varepsilon_{l-1})$ is the standard Brownian motion in $\mathbb{R}^l$. The evolution of $\mathbf{c}_i$ and $\mathbf{U}$ are presented in the appendix B.5.

This theorem can be interpreted as the stochastic counterpart to the evolution of eigenvalues previously described for linear networks by Arora et al. [2019c], Varre et al. [2023]. The derivation of the eigenvalues is inspired by the work of Bru [1989].

The evolution of the eigenvalues features a key term highlighted in Equation (5.6) consisting of the sum of skew-symmetric elements ${}^{\mathbf{s}_i\mathbf{c}_j^2 + \mathbf{s}_j\mathbf{c}_i^2}/_{\mathbf{s}_i - \mathbf{s}_j}$. For a pair of indices (i_0, j_0) with $i_0 < j_0$ and thus $\mathbf{s}_{i_0} > \mathbf{s}_{j_0}$, the term ${}^{\mathbf{s}_{i_0}\mathbf{c}_{j_0}^2 + \mathbf{s}_{j_0}\mathbf{c}_{i_0}^2}/_{\mathbf{s}_{i_0} - \mathbf{s}_{j_0}}$ positively influences the evolution of the larger eigenvalue $\mathrm{d}\mathbf{s}_{i_0}$ and negatively affects the smaller eigenvalue $\mathrm{d}\mathbf{s}_{j_0}$. Therefore, this force is repulsive, driving the eigenvalues apart and increasing their gap. Another factor influencing the dynamics is the presence of Geometric Brownian motion, where the singular value σ_i multiplicatively influences the Brownian motion as $\sqrt{\mathbf{s}_i\mathbf{c}_i^2}\left(\mathrm{d}\tilde{\mathbf{X}}\right)_i$, similar to what is observed in diagonal linear networks (refer to the previous discussion for similarities). This effect tends to pull the singular values toward zero. Together with the fact that $(\mathbf{s}_i, \mathbf{c}_i) = (0, 0)$ represents a fixed point of the dynamics, these two forces collectively push redundant singular values toward zero.

To further understand the interplay of repulsive forces and geometric Brownian motion, we consider the evolution of the smaller singular value $\mathbf{s}_p$ for $l = p$. Using the Ito chain rule, we analyze the evolution of $\log \mathbf{s}_p$, expressed as,

$$\mathrm{d}(\log \mathbf{s}_p) = p\frac{\mathbf{c}_p^2}{\mathbf{s}_p}\,\mathrm{d}t + \frac{1}{\mathbf{s}_p}\sum_{\substack{j=1,\\ j\neq p}}^{p} \frac{\mathbf{s}_p\mathbf{c}_j^2 + \mathbf{s}_j\mathbf{c}_p^2}{\mathbf{s}_p - \mathbf{s}_j}\mathrm{d}t - 2\frac{\mathbf{c}_p^2}{\mathbf{s}_p} + 2\sqrt{\frac{\mathbf{c}_p^2}{\mathbf{s}_p}}\left(\mathrm{d}\tilde{\mathbf{X}}\right)_p .$$

Using that ${}^{\mathbf{s}_p\mathbf{c}_j^2+\mathbf{s}_j\mathbf{c}_p^2}/_{\mathbf{s}_p-\mathbf{s}_j} < -\mathbf{c}_p^2$, for all indices j, the repulsive force accumulates to $-(p-1)(\mathbf{c}_p^2/\mathbf{s}_p)$ and the Ito correction term from the logarithm contributes an additional $-2(\mathbf{c}_p^2/\mathbf{s}_p)$ (the GBM component) thus offsetting the positive drift of $p(\mathbf{c}_p^2/\mathbf{s}_p)$. In the case of $l \neq p$, considering a polynomial x^α with an appropriate α would demonstrate similar behaviour. This discussion outlines the forces at play, yet a complete characterization of the solution of the SDE Equation (5.6) remains missing. Moreover, we have not established that the eigenvalues avoid a.s. collision, i.e., the explosion time $\tau_\infty = \infty$ which is in itself a significant challenge [Bru, 1989, Graczyk and Małecki, 2014].

A simplified two-vector problem. To enhance our understanding of the SDE governing the evolution of the eigenvalues detailed in Equation (5.6), we consider the large noise limit. In this scenario, the process described in Equation (5.3) simplifies to a purely noise-driven process without drift:

$$\mathrm{d}\mathbf{W} = \mathrm{d}\mathbf{B}_t\mathbf{a}^\top; \qquad \mathrm{d}\mathbf{a} = \mathbf{W}^\top\mathrm{d}\mathbf{B}_t.$$

This SDE exhibits notable symmetry; allowing for an analysis using a matrix with sub-sampled columns. Let S be any subset of $1, \ldots, l$, with $(\mathbf{w}_i)_{i=1}^l$ representing the columns of $\mathbf{W}$. We define $\mathbf{W}_S \in \mathbb{R}^{p\times|S|}$ as the subsampled matrix obtained by selecting columns $\mathbf{w}_i$ where $i \in S$, and similarly, we define a subsampled vector $\mathbf{a}_S$ by selecting the corresponding coordinates. The SDE restricted to the set S is structured as follows:

$$\mathrm{d}\mathbf{W}_S = \mathrm{d}\mathbf{B}_t\mathbf{a}_S^\top; \qquad \mathrm{d}\mathbf{a}_S = \mathbf{W}_S^\top\mathrm{d}\mathbf{B}_t.$$

To demonstrate that the columns of $\mathbf{W}$ align, we leverage the symmetry of the SDE by examining the restricted problem on every pair of rows $S = \{i, j\}$, and proving alignment within this subset. This approach leads us to consider the two vector problem ($l = 2$), where $\mathbf{W} = [\mathbf{w}_1|\mathbf{w}_2]$ and $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^p$, $\mathbf{a} \in \mathbb{R}^2$. We describe the behavior of the eigenvalues for this two-vector problem in the theorem below.

Theorem 5.2. *In the large noise limit, let $\mathbf{s}_0 > \mathbf{s}_1$ be the eigenvalues of $\mathbf{W}$, the following properties hold, for $t \leq \tau$ defined by $\tau = \{\inf t : \mathbf{s}_0(t) = \mathbf{s}_1(t)\}$,*

- *(a) $\mathbf{s}_0, \mathbf{s}_1$ are greater than zero almost surely,*
- *(b) for $\alpha = (p-3)/2$, $\mathbf{s}_0^{-\alpha}$ is a super-martingale while $\mathbf{s}_1^{-\alpha}$ is a sub-martingale.*

This model for $l = 2$ mirrors the dynamics of the Wishart process studied by Bru [1991], motivating the exploration of the evolution of an appropriately chosen exponent of $\mathbf{s}_0, \mathbf{s}_1$. The first part of the theorem arises from the fact that $\mathbf{s}_1^{-\alpha}\mathbf{s}_2^{-\alpha}$ is a local continuous martingale that cannot explode to infinity in finite time. The second part highlights a clear separation between the eigenvalues: one is a sub-martingale that consistently increases in expectation, while the other is a super-martingale that diminishes (note that the eigenvalues are raised to a negative power). This dynamic, coupled with the symmetry argument, suggests that for every pair of columns, there is a component that strengthens the alignment through its increases in expectation. Refer to App. B.6 for the proof.

Conclusion. In this section, we derive the SDE of eigenvalues for the matrix of parameters evolving under SGF. This derivation provides deeper insights into the mechanisms contributing to low-rank behavior. Specifically, repulsive forces drive the eigenvalues apart, while the geometric Brownian motion pulls them towards zero. These forces, unique to training with SGF, highlight the regularization effects of stochastic methods compared to gradient flow. However, fully characterizing the solution of this SDE remains a challenging open problem we let as future work.

6 Generalization to other settings

In this section, we generalize our results beyond the square loss and the label noise gradient flow. We consider the general framework of a loss function over the weight product $\mathbf{W}_1\mathbf{W}_2$ defined as

$$\mathcal{L}\left(\mathbf{W}_1, \mathbf{W}_2\right) = \widehat{\mathcal{L}}(\mathbf{W}_1\mathbf{W}_2) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(\mathbf{W}_1\mathbf{W}_2; x, y)\right],$$

In this framework, the loss function ℓ combines the prediction loss directly with the parametrized model f_θ. This approach applies, for example, to classification problems using linear networks where ℓ might represent any classification loss and $f_\theta = \mathbf{W}_1\mathbf{W}_2$. It also directly extends to more complex architectures where $f_\theta = \sigma(\mathbf{W}_1\mathbf{W}_2)$ for an activation function σ, including settings like a self-attention layer with frozen value vectors. We denote the product by $\boldsymbol{\beta} = \mathbf{W}_1\mathbf{W}_2$ noting it solely controls the loss. We investigate the evolution of the weight matrix determinant for a general loss across various algorithms, from gradient flow to gradient descent, and demonstrate that a similar separation occurs due to stochasticity.

Warm-up: Gradient flow. The gradient flow on the loss $\mathcal{L}$ can be written as the following,

$$\mathrm{d}\boldsymbol{\Theta} = \boldsymbol{\Theta}\mathbf{J}\mathrm{d}t, \qquad \text{where } \mathbf{J} = \begin{bmatrix} 0_{p\times p} & -\nabla\widehat{\mathcal{L}}(\boldsymbol{\beta}) \\ -\nabla\widehat{\mathcal{L}}(\boldsymbol{\beta})^\top & 0_{k\times k} \end{bmatrix}. \tag{6.1}$$

Following a similar proof as in Theorem 4.1, we obtain that $\mathrm{d}\big(\det\left(\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\right)\big) = 0$. For separable classification problem, the gradient flow converges to infinity [Soudry et al., 2018, Ji and Telgarsky, 2019], hence, after appropriate rescaling, the layers are aligned, as shown by Ji and Telgarsky [2019]. Next, we contrast this result with the outcomes observed in stochastic and discrete algorithms.

Continuous modelling of SGD. We consider the SGD algorithm with a batch size B. We denote the mini-batch version of the loss functions $\mathcal{L}$ and $\widehat{\mathcal{L}}$ as $\mathcal{L}_B$ and $\widehat{\mathcal{L}}_B$, respectively. The SGD update with stepsize η can be represented with the following block structure,

$$\boldsymbol{\Theta}^{t+1} = \boldsymbol{\Theta}^t - \eta\boldsymbol{\Theta}^t\mathbf{J}^t - \eta\boldsymbol{\Theta}^t\xi^t, \quad \text{where } \xi^t = \begin{bmatrix} 0_{p\times p} & -\left(\nabla\widehat{\mathcal{L}}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{L}}_B(\boldsymbol{\beta})\right) \\ -\left(\nabla\widehat{\mathcal{L}}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{L}}_B(\boldsymbol{\beta})\right)^\top & 0_{k\times k} \end{bmatrix}.$$

We denote the SGD noise as $g_t = \left(\nabla\widehat{\mathcal{L}}(\boldsymbol{\beta}) - \nabla\widehat{\mathcal{L}}_B(\boldsymbol{\beta})\right)$ and the noise covariance as $\Sigma_t = \mathbb{E}\left[g^t\left(g^t\right)^\top\right]$ where the expectation is over all the minibatches. Following Li et al. [2019a], the SGD update can be modelled with the following SDE,

$$\mathrm{d}\boldsymbol{\Theta} = -\boldsymbol{\Theta}\mathbf{J}\mathrm{d}t - \sqrt{\eta}\mathrm{d}\xi, \text{where } \mathrm{d}\xi = \begin{bmatrix} 0_{p\times p} & -\Sigma_t^{1/2}\mathrm{d}\mathbf{B}_t \\ -\left(\Sigma_t^{1/2}\mathrm{d}\mathbf{B}_t\right)^\top & 0_{k\times k} \end{bmatrix}. \tag{6.2}$$

The main difference with SGF is that, in overparameterized problems, the noise covariance is time-varying and decreases to zero upon convergence. Using Theorem B.3, the evolution of the determinant of $\mathbf{M} = \boldsymbol{\Theta}^\top\boldsymbol{\Theta}$ is given by $\mathrm{d}(\det\left(\mathbf{M}\right)) = -\eta\det\left(\mathbf{M}\right)\mathrm{Tr}\left(\Sigma(t)\right)\mathrm{d}t$ and can be explicitly solved as

$$\mathrm{d}(\det\left(\mathbf{M}\right)(t)) = \det\left(\mathbf{M}(0)\right)\exp\{-\eta\int_0^t \mathrm{Tr}\left(\Sigma(s)\right)\mathrm{d}s\}.$$

Hence, the decay in the determinant is governed by the integral $\int_0^\infty \mathrm{Tr}\left(\Sigma(t)\right)\mathrm{d}t$ which is a stochastic quantity. $\mathrm{Tr}\left(\Sigma(t)\right)$ represents the strength of the stochastic noise, which, in over-parameterized regression, is proportional to the loss, i.e., $\mathrm{Tr}\left(\Sigma(t)\right) \propto \mathcal{L}\left(\boldsymbol{\Theta}\right)$ [Pesme et al., 2021]. Therefore, the rate of decay in the determinant depends on $\int_0^\infty \mathcal{L}\left(\left(\boldsymbol{\Theta}(t)\right)\right)\mathrm{d}t$, with slower convergence leading to a simpler model at convergence, as observed in the case of diagonal networks by Pesme et al. [2021]. The result above also holds for *non-separable* classification tasks where the noise of SGD drives the determinant to 0, a scenario not covered by the previous analysis of Ji and Telgarsky [2019].

Discrete gradient algorithms. We can extend the previous results to discrete (possibly stochastic) gradient algorithm. Both algorithms can be written as

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t\left(\mathrm{I}_{p+k} + \eta\mathbf{J}_t\right),$$

for stepsize η and $\mathbf{J}_t$ the possibly stochastic block gradient matrix defined in Equation (6.1). In the context of discrete algorithms, the determinant is controlled by the following lemma (refer to B.4 for the proof).

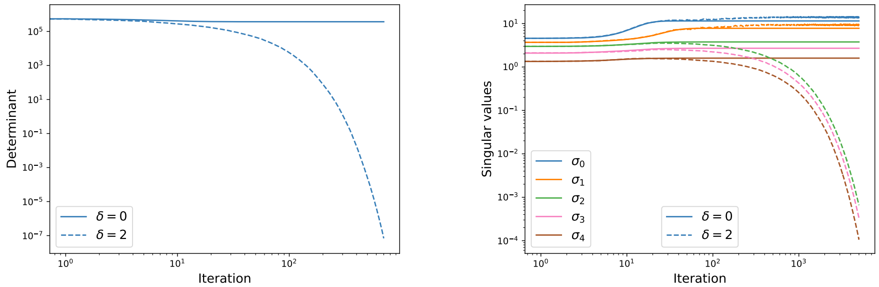

Figure 1: Evolution of the model characteristics for gradient flow ($\delta = 0$) and stochastic gradient flow ($\delta = 2$). Left: Determinant of $\mathbf{M}$. Right: Top-5 singular values of $\mathbf{W}_1$.

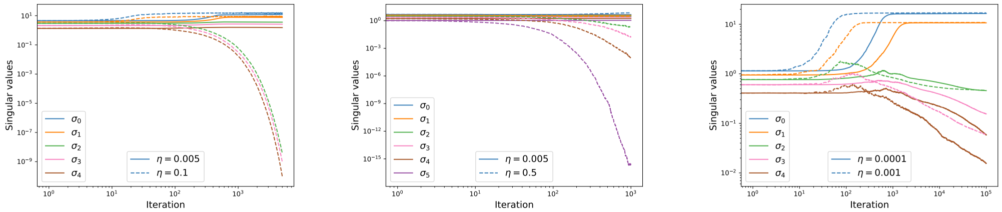

Figure 2: Evolution of the top-5 singular values of $\mathbf{W}_1$ for SGD with small and large stepsizes η. Left: Regression with MSE loss, linear network. Middle: Classification with logistic loss, linear network. Right: Regression with MSE loss, 2-layer ReLU network.

Lemma 6.1. *When $l = p + k$ and $\eta^2 \left\|\mathbf{J}_t\right\|_F^2 \le 1$, the following property holds for the determinant,*

$$|\det \mathbf{\Theta}_{t+1}| \le \exp\left(-\frac{\eta^2}{2}\left\|\mathbf{J}_t\right\|_F^2\right)|\det \mathbf{\Theta}_t|.$$

If the factor $\eta^2 \left\|\mathbf{J}_t\right\|_F^2 \le 1$ at every iteration t, the determinant is reduced by the discrete step size. However, there is a tradeoff: the sum $S := \sum_{t=0}^{\infty} \eta^2 \left\|\mathbf{J}_t\right\|_F^2$ can be finite, indicating that it does not completely drive the determinant to zero. Increasing η to increase S might lead to instability and divergence. Furthermore, since $\left\|\mathbf{J}_t\right\|_F^2 \propto \mathcal{L}\left(\mathbf{\Theta}_t\right)$, there is an additional tradeoff between convergence and the simplicity of the parameters. This illustrates how step sizes that produce non-convergent training loss patterns, such as the catapult effect [Lewkowycz et al., 2020] or the edge of stability mechanisms [Cohen et al., 2020], can simplify the network's parameters.

7 Experimental evidence

We consider a regression problem on synthetic data with $n = 1000$ samples of Gaussian data in $\mathbb{R}^5$ ($p = 5$) with labels in $\mathbb{R}^2$ ($k = 2$) generated by some ground truth $\boldsymbol{\beta} \in \mathbb{R}^{5\times 2}$, the width of the network is $l = 10$. We use Gaussian initialization of the network parameters with entries from $\mathcal{N}(0, 1)$. Experiments details can be found in the appendix C. In the left plot of Figure 1, we show the time evolution of the determinant of matrix $\mathbf{M}$. As suggested by theorems 4.1 and 4.2, in the case without label noise, $\det\left(\mathbf{\Theta}^\top \mathbf{\Theta}\right)$ stays constant, while with the Label Noise of intensity $\delta = 2$ it goes to zero with time. In the right plot of Figure 1, we demonstrate the time evolution of the top-5 singular values of the matrix $\mathbf{W}_1$. Note that in the case of Gradient Flow all except the first k singular values (σ_0 and σ_1) stay at the same scale, while adding Label Noise forces smallest $d + l - k$ singular values ($\sigma_2, \sigma_3,$ and σ_4) to tend toward zero. Further experiments illustrate in Figure 2 the evolution of singular values of parameter matrix $\mathbf{W}_1$ when optimized with SGD, for classification tasks and with ReLU network. These results also confirm that the beneficial effects of stochasticity hold in these contexts.

8 Conclusion

In this paper, we demonstrate a distinct separation between GF and SGF when trained on linear networks. This separation is obtained by tracking the evolution of the determinant of the parameter matrix. However, while the determinant is a significant factor, it does not fully capture the implicit regularization effects. Notably, the determinant mirrors the imbalance $\mathbf{u}^2 - \mathbf{v}^2$ in diagonal networks represented by $\mathbf{u} \odot \mathbf{v}$, whose dynamics play a crucial role in attuning the implicit regularization across various algorithms [Woodworth et al., 2020, Pesme et al., 2021, Papazov et al., 2024]. Our analysis presents the initial step in deciphering implicit regularization for stochastic methods in linear networks, yet achieving a complete characterization remains a promising direction for future research.

Acknowledgments and Disclosure of Funding

AV is supported by Swiss data science fellowship. This work was supported by the Swiss National Science Foundation (grant number 212111).

References

M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion. Sgd with large step sizes learns sparse features. *arXiv preprint arXiv:2210.05337*, 2022.

M. Andriushchenko, F. Croce, M. Müller, M. Hein, and N. Flammarion. A modern look at the relationship between sharpness and generalization. In *International Conference on Machine Learning*, pages 840–902. PMLR, 2023.

S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=SkMQg3C5K7.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019b.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019c. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

G. Blanc, N. Gupta, G. Valiant, and P. Valiant. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on Learning Theory, COLT 2020*, Proceedings of Machine Learning Research. PMLR, 2020.

E. Boursier, L. Pillaud-Vivien, and N. Flammarion. Gradient flow dynamics of shallow reLU networks for square loss and orthogonal inputs. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=L74c-iUxQ1I.

L. Braun, C. C. J. Dominé, J. E. Fitzgerald, and A. M. Saxe. Exact learning dynamics of deep linear networks with prior knowledge. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=lJx2vng-KiC.

M.-F. Bru. Diffusions of perturbed principal component analysis. *Journal of Multivariate Analysis*, 29(1):127–136, 1989. ISSN 0047-259X. doi: https://doi.org/10.1016/0047-259X(89)90080-8. URL https://www.sciencedirect.com/science/article/pii/0047259X89900808.

M.-F. Bru. Wishart processes. *Journal of Theoretical Probability*, 4:725–751, 1991.

F. Chen, D. Kunin, A. Yamamura, and S. Ganguli. Stochastic collapse: How gradient noise attracts SGD dynamics towards simpler subnetworks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=iFxWrxDekd`.

L. Chizat, E. Oyallon, and F. Bach. *On Lazy Training in Differentiable Programming*. Curran Associates Inc., Red Hook, NY, USA, 2019.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2020.

A. Damian, T. Ma, and J. D. Lee. Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

S. S. Du, W. Hu, and J. D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, 2018.

F. J. Dyson. A Brownian-Motion Model for the Eigenvalues of a Random Matrix. *Journal of Mathematical Physics*, 3(6):1191–1198, 11 1962. ISSN 0022-2488. doi: 10.1063/1.1703862. URL `https://doi.org/10.1063/1.1703862`.

M. Even, S. Pesme, S. Gunasekar, and N. Flammarion. (s) gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems*, 2023.

K. Fukumizu. Effect of batch learning in multilayer neural networks. *Gen*, 1(04):1E–03, 1998.

N. Ghosh, S. Frei, W. Ha, and B. Yu. The effect of sgd batch size on autoencoder learning: Sparsity, sharpness, and feature learning. *arXiv preprint arXiv:2308.03215*, 2023.

G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

P. Graczyk and J. Małecki. Multidimensional yamada-watanabe theorem and its applications to particle systems. *Journal of Mathematical Physics*, 54(2), 2013.

P. Graczyk and J. Małecki. Strong solutions of non-colliding particle systems. 2014.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

J. Z. HaoChen, C. Wei, J. D. Lee, and T. Ma. Shape matters: Understanding the implicit bias of the noise covariance. In M. Belkin and S. Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, 2021.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

F. He, T. Liu, and D. Tao. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, jan 1997.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1729–1739, 2017.

N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 1981. ISBN 0-444-86172-6.

S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, A. Storkey, and Y. Bengio. Three factors influencing minima in SGD. In *International Conference on Learning Representations*, 2018.

Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJflg30qKX`.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

R. Khasminskii. *Stochastic Stability of Differential Equations*, volume 66 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, second edition, 2012.

B. Kleinberg, Y. Li, and Y. Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

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

Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019b. URL `http://jmlr.org/papers/v20/17-526.html`.

Z. Li, T. Wang, and S. Arora. What happens after sgd reaches zero loss?–a mathematical framework. In *International Conference on Learning Representations*, 2021.

S. Mandt, M. D. Hoffman, and D. M. Blei. A variational analysis of stochastic gradient algorithms. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 354–363, 2016.

S. Marcotte, R. Gribonval, and G. Peyré. Abide by the law and follow the flow: conservation laws for gradient flows. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=kMueEV8Eyy`.

E. Mayerhofer, O. Pfaffel, and R. Stelzer. On strong solutions for positive definite jump diffusions. *Stochastic processes and their applications*, 121(9):2072–2086, 2011.

H. Min, S. Tarmoun, R. Vidal, and E. Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

J. R. Norris, L. C. G. Rogers, and D. Williams. Brownian motions of ellipsoids. *Transactions of the American Mathematical Society*, 294(2):757–765, 1986. ISSN 00029947. URL `http://www.jstor.org/stable/2000214`.

H. Papazov, S. Pesme, and N. Flammarion. Leveraging continuous time to understand momentum when training diagonal linear networks. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2024.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

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

L. Pillaud-Vivien, J. Reygner, and N. Flammarion. Label noise (stochastic) gradient descent implicitly solves the lasso for quadratic parametrisation. In *Conference on Learning Theory*, pages 2127–2159. PMLR, 2022.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

S. Tarmoun, G. Franca, B. D. Haeffele, and R. Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 18–24 Jul 2021.

J. Townsend. Differentiating the singular value decomposition. Technical report, Technical Report 2016, https://j-towns. github. io/papers/svd-derivative . . . , 2016.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

A. V. Varre, M.-L. Vladarean, L. Pillaud-Vivien, and N. Flammarion. On the spectral bias of two-layer linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=FFdrXkm3Cz`.

A. V. Varre, M.-L. Vladarean, L. Pillaud-Vivien, and N. Flammarion. On the spectral bias of two-layer linear networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Z. Wang and A. Jacot. Implicit bias of sgd in l_2-regularized linear dnns: One-way jumps from high to low rank, 2023.

B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3635–3673. PMLR, 09–12 Jul 2020.

L. Wu, C. Ma, and W. E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

L. Ziyin, H. Li, and M. Ueda. Law of balance and stationary distribution of stochastic gradient descent. *arXiv preprint arXiv:2308.06671*, 2023.

A Notations

Notation S_d, S_d^+, S_d^{++} denote the set of symmetric, positive semi-definite and positive definite matrices in $R^{d\times d}$. We use $\odot$ to denote the Hadamard product.

B Proofs

Theorem B.1. *For the gradient flow defined in Equation* (3.4)*, the following property holds,*

$$\mathrm{d}\big(\det\big(\mathbf{\Theta}^\top\mathbf{\Theta}\big)\big) = 0.$$

Hence, $\det\big(\mathbf{\Theta}(t)^\top\mathbf{\Theta}(t)\big) = \det\big(\mathbf{\Theta}_0^\top\mathbf{\Theta}_0\big)$*, where* $\mathbf{\Theta}_0 = \mathbf{\Theta}(0)$ *is the initialisation at time* $t = 0$*.*

First, we present a proof of this theorem, based on straightforward computations of the derivative of the determinant and the fact that the matrix $\mathbf{J}$ has zero trace.

Proof. Let $\mathbf{M} = \mathbf{\Theta}^\top\mathbf{\Theta}$. The dynamics of $\mathbf{M}$ are governed by the ODE,

$$\mathrm{d}\mathbf{M} = \mathrm{d}\mathbf{\Theta}^\top\mathbf{\Theta} + \mathbf{\Theta}^\top\mathrm{d}\mathbf{\Theta} = \mathbf{\Theta}^\top\mathbf{\Theta}\mathbf{J}\mathrm{d}t + \mathbf{J}\mathbf{\Theta}^\top\mathbf{\Theta}\mathrm{d}t = (\mathbf{MJ} + \mathbf{JM})\mathrm{d}t.$$

Using the gradient of the determinant given in Proposition B.2, the determinant of $\mathbf{M}$ evolves as follows,

$$\begin{aligned}\mathrm{d}(\det(\mathbf{M})) &= \langle\nabla\det(\mathbf{M}), \mathrm{d}\mathbf{M}\rangle = \det(\mathbf{M})\left\langle\mathbf{M}^{-1}, \mathbf{MJ} + \mathbf{JM}\right\rangle \mathrm{d}t,\\ &= \det(\mathbf{M})\left\langle\mathbf{M}^{-1}, \mathbf{MJ}\right\rangle + \left\langle\mathbf{M}^{-1}, \mathbf{JM}\right\rangle = 2\det(\mathbf{M})\left\langle \mathrm{I}_{p+k}, \mathbf{J}\right\rangle = 2\det(\mathbf{M})\mathrm{Tr}(\mathbf{J}).\end{aligned}$$

Given that $\mathrm{Tr}(\mathbf{J}) = 0$, it follows that $\mathrm{d}(\det(\mathbf{M})) = 0$. □

Proposition B.2. *For any matrix M in* S_d^{++}*, the first two derivatives of the determinant of M, denoted by* $\det(M)$ *are the following*

(i) $\nabla\det(M) = \det(M)M^{-1}$

(ii) For $1 \le a, b, k, l \le d$*, the second order partial derivative is given by*

$$\frac{\partial^2\det(M)}{\partial M_{ab}\partial M_{kl}} = \det(M)\left[(M^{-1})_{ba}(M^{-1})_{lk} - (M^{-1})_{bk}(M^{-1})_{la}\right] \tag{B.1}$$

Theorem B.3. *For a stochastic process given by the SDE,*

$$\mathrm{d}\mathbf{\Theta} = \mathbf{\Theta}\left[\mathbf{J}\mathrm{d}t + \mathrm{d}\xi\right] \tag{B.2}$$

with $\mathrm{Tr}\,\mathbf{J} = \mathrm{Tr}\,\xi = 0$*, the determinant of the* $\mathbf{M} = \mathbf{\Theta}^\top\mathbf{\Theta}$ *evolves as*

$$\mathrm{d}(\det(\mathbf{M})) = -\det(\mathbf{M})\mathrm{Tr}\left[\mathrm{d}\xi\mathrm{d}\xi\right]. \tag{B.3}$$

Proof. First, we compute the evolution of $\mathbf{M} = \mathbf{\Theta}^\top\mathbf{\Theta}$ using the Ito's product rule,

$$\mathrm{d}\mathbf{M} = \mathrm{d}\big(\mathbf{\Theta}^\top\mathbf{\Theta}\big) = \mathrm{d}\mathbf{\Theta}^\top\mathbf{\Theta} + \mathbf{\Theta}^\top\mathrm{d}\mathbf{\Theta} + \mathrm{d}\mathbf{\Theta}^\top\mathrm{d}\mathbf{\Theta}$$

The last term is interpreted as a derivative of the finite variation and it should be computed using $\mathrm{d}t.\,(\mathrm{d}\mathbf{B}_t)_{ij} = 0$ and $(\mathrm{d}\mathbf{B}_t)_{ij}\,.\,(\mathrm{d}\mathbf{B}_t)_{kl} = \delta_{i=k\wedge j=l}\mathrm{d}t$. Using Eq. (3.6),

$$\begin{aligned}\mathrm{d}\mathbf{M} &= [\mathbf{J}\mathrm{d}t + \mathrm{d}\xi]\mathbf{\Theta}^\top\mathbf{\Theta} + \mathbf{\Theta}^\top\mathbf{\Theta}[\mathbf{J}\mathrm{d}t + \mathrm{d}\xi] + \mathrm{d}\xi\mathbf{\Theta}^\top\mathbf{\Theta}\mathrm{d}\xi,\\ &= \mathbf{JM}\mathrm{d}t + \mathbf{MJ}\mathrm{d}t + \mathrm{d}\xi\mathbf{M}\mathrm{d}\xi + \mathrm{d}\xi\mathbf{M} + \mathbf{M}\mathrm{d}\xi.\end{aligned}$$

Using the Ito chain rule, we can compute the evolution of determinant as following,

$$\mathrm{d}(\det(\mathbf{M})) = \langle\nabla\det(\mathbf{M}), \mathrm{d}\mathbf{M}\rangle + \frac{1}{2}\sum_{a,b,k,l}\frac{\partial^2\det(\mathbf{M})}{\partial\mathbf{M}_{ab}\partial\mathbf{M}_{kl}}\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl},$$

The first term is

$$\begin{aligned}\langle \nabla \det(\mathbf{M}), \mathrm{d}\mathbf{M}\rangle &= \det(\mathbf{M})\left\langle \mathbf{M}^{-1}, \mathbf{J}\mathbf{M}\mathrm{d}t + \mathbf{M}\mathbf{J}\mathrm{d}t + \mathrm{d}\xi\mathbf{M}\mathrm{d}\xi + \mathrm{d}\xi\mathbf{M} + \mathbf{M}\mathrm{d}\xi\right\rangle, \\ &= 2det(\mathbf{M})\left\langle \mathrm{I}_{p+k}, \mathbf{J}\right\rangle \mathrm{d}t + 2\det(\mathbf{M})\left\langle \mathrm{I}_{p+k}, \mathrm{d}\xi\right\rangle + \det(\mathbf{M})\left\langle \mathbf{M}^{-1}, \mathrm{d}\xi\mathbf{M}\xi\right\rangle\end{aligned}$$

Using the property that $\mathrm{Tr}\,(\mathbf{J}) = \mathrm{Tr}\,(\mathrm{d}\xi) = 0$. We get that $\langle \nabla \det(\mathbf{M}), \mathrm{d}\mathbf{M}\rangle = \left\langle \mathbf{M}^{-1}, \mathrm{d}\xi\mathbf{M}\xi\right\rangle$. For the second term

$$\begin{aligned}\frac{1}{2}\sum_{a,b,k,l}\frac{\partial^2 \det(\mathbf{M})}{\partial \mathbf{M}_{ab}\partial \mathbf{M}_{kl}}\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl} &= \frac{1}{2}\sum_{a,b,k,l}\det\mathbf{M}\left[(\mathbf{M}^{-1})_{ba}(\mathbf{M}^{-1})_{lk} - (\mathbf{M}^{-1})_{bk}(\mathbf{M}^{-1})_{la}\right]\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl}, \\ &= \frac{det(\mathbf{M})}{2}\sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{ba}(\mathbf{M}^{-1})_{lk}\right]\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl} \\ &\qquad - \sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{bk}(\mathbf{M}^{-1})_{la}\right]\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl},\end{aligned}$$

Rearranging the terms in the summation, we get,

$$\begin{aligned}\sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{ba}(\mathbf{M}^{-1})_{lk}\right]\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl} &= \sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{ba}\mathrm{d}\mathbf{M}_{ab}\right]\left[(\mathbf{M}^{-1})_{lk}\mathrm{d}\mathbf{M}_{kl}\right], \\ &= \sum_{b,l}\left[\sum_{a}(\mathbf{M}^{-1})_{ba}\mathrm{d}\mathbf{M}_{ab}\right]\left[\sum_{k}(\mathbf{M}^{-1})_{lk}\mathrm{d}\mathbf{M}_{kl}\right], \\ &= \sum_{b,l}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{bb}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{ll} = \sum_{b}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{bb}\sum_{l}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{ll}, \\ &= \mathrm{Tr}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\mathrm{Tr}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right).\end{aligned}$$

Similarly for the other term, we get,

$$\begin{aligned}\sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{bk}(\mathbf{M}^{-1})_{la}\right]\mathrm{d}\mathbf{M}_{ab}\mathrm{d}\mathbf{M}_{kl} &= \sum_{a,b,k,l}\left[(\mathbf{M}^{-1})_{bk}\mathrm{d}\mathbf{M}_{kl}\right]\left[(\mathbf{M}^{-1})_{la}\mathrm{d}\mathbf{M}_{ab}\right], \\ &= \sum_{b,l}\left[\sum_{a}(\mathbf{M}^{-1})_{ba}\mathrm{d}\mathbf{M}_{al}\right]\left[\sum_{k}(\mathbf{M}^{-1})_{bk}\mathrm{d}\mathbf{M}_{kl}\right], \\ &= \sum_{b}\left[\sum_{l}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{bl}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{lb}\right] = \sum_{b}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)_{bb}, \\ &= \mathrm{Tr}\left[\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\right].\end{aligned}$$

Note that the diffusion part of $\mathbf{M}^{-1}\mathrm{d}\mathbf{M}$ is $\mathrm{d}\xi + \mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}$. Using this

$$\mathrm{Tr}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\mathrm{Tr}\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right) = \mathrm{Tr}\left[\mathrm{d}\xi + \mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\right]\mathrm{Tr}\left[\mathrm{d}\xi + \mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\right] = 0,$$

as $\mathrm{Tr}\,\mathrm{d}\xi = 0$. For the other term,

$$\begin{aligned}\mathrm{Tr}\left[\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\left(\mathbf{M}^{-1}\mathrm{d}\mathbf{M}\right)\right] &= \mathrm{Tr}\left[\left(\mathrm{d}\xi + \mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\right)\left(\mathrm{d}\xi + \mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\right)\right], \\ &= 2\mathrm{Tr}\left[\mathrm{d}\xi\mathrm{d}\xi\right] + 2\mathrm{Tr}\left[\mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\mathrm{d}\xi\right].\end{aligned}$$

Putting everything together, we get,

$$\frac{1}{2}\sum_{a,b,k,l}\frac{\partial^2 \det(\mathbf{M})}{\partial \mathbf{M}_{ab}\partial \mathbf{M}_{kl}} = -\det\mathbf{M}\left(\mathrm{Tr}\left[\mathrm{d}\xi\mathrm{d}\xi\right] + \mathrm{Tr}\left[\mathbf{M}^{-1}\mathrm{d}\xi\mathbf{M}\mathrm{d}\xi\right]\right)$$

which gives us

$$\mathrm{d}(\det(\mathbf{M})) = -\det(\mathbf{M})\ \mathrm{Tr}\left[\mathrm{d}\xi\mathrm{d}\xi\right].$$

□

Lemma B.4. *When $l = p + k$ and $\eta^2 \left\|\mathbf{J}_t\right\|_F^2 \leq 1$, the following property holds for the determinant,*

$$|\det \mathbf{\Theta}_{t+1}| \leq \exp\left(-\frac{\eta^2}{2}\left\|\mathbf{J}_t\right\|_F^2\right)|\det \mathbf{\Theta}_t|.$$

Proof. Note that because of the block structure of the matrix $\mathbf{J}_t$, its nonzero eigenvalues come in $\pm$-pairs: $\pm\sigma_1, \ldots, \pm\sigma_m$, moreover, since $\mathbf{J}_t$ is symmetric, singular values of $\mathbf{J}_t$ are the absolute values of eigenvalues, i.e. $\sigma_1, \ldots, \sigma_m$. Then, the determinant of $\mathbf{\Theta}_{t+1}$ can be written as the following,

$$\det \mathbf{\Theta}_{t+1} = \det \mathbf{\Theta}_t \det\left(\mathrm{I}_{p+k} + \eta \mathbf{J}_t\right) = \det \mathbf{\Theta}_t \prod_{i=1}^{m}(1 - \eta^2\sigma_i^2).$$

Using that $1 - x^2 \leq e^{-x^2}$ for all x, we can estimate

$$\prod_{i=1}^{m}(1 - \eta^2\sigma_i^2) \leq \exp\left(-\eta^2\sum_{i=1}^{m}\sigma_i^2\right) = \exp\left(-\frac{\eta^2}{2}\left\|\mathbf{J}_t\right\|_F^2\right).$$

We obtain the required inequality by observing that $\prod_{i=1}^{m}(1 - \eta^2\sigma_i^2) = \left|\prod_{i=1}^{m}(1 - \eta^2\sigma_i^2)\right|$ since each term $1 - \eta^2\sigma_i^2 \geq 0$ when $\eta^2\left\|\mathbf{J}_t\right\|_F^2 < 1$. □

Theorem B.5. *Let $\mathbf{s}_1 > \ldots \mathbf{s}_l$ be the order of the eigenvalues of the matrix $\mathbf{M}$ defined by Equation (5.4). Let the collision time for the eigenvalues be defined as*

$$\tau = \{\inf t : \mathbf{s}_i(t) = \mathbf{s}_j(t) \mathit{\ for\ } 1 \leq i \neq j \leq l\}. \tag{B.4}$$

For $t \leq \tau$, the eigenvalues are semi-martingales given by the solution of the following SDE

$$\mathrm{d}(\mathbf{s}_i) = p\mathbf{c}_i^2\,\mathrm{d}t + \sum_{\substack{j=1,\\ j\neq i}}^{l}\frac{\mathbf{s}_i\mathbf{c}_j^2 + \mathbf{s}_j\mathbf{c}_i^2}{\mathbf{s}_i - \mathbf{s}_j}\mathrm{d}t + 2\sqrt{\mathbf{s}_i\mathbf{c}_i^2}\left(\mathrm{d}\tilde{\mathbf{X}}\right)_i \tag{B.5}$$

where $\left(\mathrm{d}\tilde{\mathbf{X}}\right)_i = {}^1\!/\!_{\eta\delta}\left(\langle\mathbf{u}_i, y\rangle - \sqrt{\mathbf{s}_i\mathbf{c}_i^2}\right)\mathrm{d}t + \mathrm{d}\varepsilon_i$ with $\mathbf{u}_i$ being the i^{th} column of $\mathbf{U}$ and $(\varepsilon_0, \ldots, \varepsilon_{l-1})$ is the standard Brownian motion in $\mathbb{R}^l$. The evolution of $\mathbf{c}_i$ and $\mathbf{U}$ are presented in the appendix.

Proof. The proof follows the approach of Bru [1989]. Let $\mathbf{W} = \mathbf{U\Sigma V}^\top$ be the singularvalue decomposition (see Def.D.1 involved with $r = l$ and $l < p$ and it will be the rank). Our focus is on understanding the evolution of the singular values and singular vectors of the matrix $\mathbf{W}$. To derive the evolution of $\mathbf{\Sigma}, \mathbf{V}$ we can consider the eigenvalues and eigenvectors of the PSD matrix process $\mathbf{M}$. Note that $\mathbf{M} = \mathbf{V\Sigma}^2\mathbf{V}^\top$, let $\mathbf{D} = \mathbf{\Sigma}^2$.

Evolution of D and V Taking the derivative of $\mathbf{M}$, we find

$$\mathrm{d}\mathbf{M} = \mathrm{d}\mathbf{W}^\top\mathbf{W} + \mathbf{W}^\top\mathrm{d}\mathbf{W} + \mathrm{d}\mathbf{W}^\top\mathrm{d}\mathbf{W} = \mathbf{a}\mathrm{d}\mathbf{X}^\top\mathbf{W} + \mathbf{W}^\top\mathrm{d}\mathbf{X}\mathbf{a}^\top + p\mathbf{a}\mathbf{a}^\top\mathrm{d}t. \tag{B.6}$$

We invoke the theorem D.2 we derived to give the eigenvalues of any matrix valued stochastic process. Note that $\mathbf{VV}^\top = \mathrm{I}_l$, so some terms of the computation are not required.

$$\mathrm{d}\mathbf{D} = \mathrm{I} \odot \tilde{\mathbf{N}}\,\mathrm{d}t + \mathrm{I} \odot \mathrm{d}\widetilde{\mathbf{M}}\,\mathrm{d}t + \mathrm{I} \odot \left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right).$$

and the evolution of the eigenvectors,

$$\mathrm{d}\mathbf{V} = \mathbf{V}\left(\mathbf{Q}_{\|}\,\mathrm{d}t + \mathbf{S} \odot (\tilde{\mathbf{N}}\mathrm{d}t + \mathrm{d}\widetilde{\mathbf{M}})\right)$$

where you define,

$$\mathbf{Q}_{\|} = \frac{\mathrm{I} \odot \left[\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right]}{2} - \mathbf{S} \odot \left[\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}} \odot \mathrm{I}\right]\right] + \mathbf{S} \odot \left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right)$$

where the matrix $\mathbf{S}$ is given by

$$\begin{aligned}
\mathbf{S}_{ij} &= \begin{cases} 0 & \text{if } i = j, \\ (\mathbf{s}_j - \mathbf{s}_i)^{-1} & \text{o.w.} \end{cases} \\
\widetilde{\mathbf{N}} &= \mathbf{V}^\top (p\mathbf{a}\mathbf{a}^\top)\mathbf{V} = p\mathbf{c}\mathbf{c}^\top. \\
\mathrm{d}\widetilde{\mathbf{M}} &= \mathbf{V}^\top \left[\mathbf{a}\mathrm{d}\mathbf{X}^\top\mathbf{W} + \mathbf{W}^\top\mathrm{d}\mathbf{X}\mathbf{a}^\top\right] \mathbf{V}, \\
&= \mathbf{c}\mathrm{d}\mathbf{X}^\top\mathbf{U}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{U}^\top\mathrm{d}\mathbf{X}\mathbf{c}^\top.
\end{aligned}$$

Note that $\boldsymbol{\Sigma} = \mathrm{diag}\left((\boldsymbol{\sigma}_0, \ldots, \boldsymbol{\sigma}_{l-1})\right)$ where $\boldsymbol{\sigma}_0 > \boldsymbol{\sigma}_1 \ldots > \boldsymbol{\sigma}_{l-1}$. Let $\mathbf{D} = \boldsymbol{\Sigma}^2$ and denote the entires of $\mathbf{D}$ as following, $\mathbf{D} = \mathrm{diag}\left((\mathbf{s}_0, \ldots, \mathbf{s}_{p-1})\right)$. Note that

$$\begin{aligned}
\mathbf{U}^\top\mathrm{d}\mathbf{X} &= \mathbf{U}^\top(\frac{1}{\eta\delta}(y - \mathbf{W}\mathbf{a})\mathrm{d}t + \mathrm{d}\mathbf{B}_t), \\
&= \frac{1}{\eta\delta}\left[\mathbf{U}^\top y - \boldsymbol{\Sigma}\mathbf{c}\right]\mathrm{d}t + \mathbf{U}^\top\mathrm{d}\mathbf{B}_t.
\end{aligned}$$

Using Levy's characterization $\mathbf{U}^\top\mathrm{d}\mathbf{B}_t$ is a Brownian motion in $\mathbb{R}^l$, lets call that $\mathrm{d}\tilde{\mathbf{B}}_t$. The diffusion part of $\mathrm{d}\widetilde{\mathbf{M}}$ (say $\mathrm{d}\mathbf{F}$)

$$\begin{aligned}
\mathrm{d}\mathbf{F} &= \boldsymbol{\Sigma}\mathbf{V}^\top\mathrm{d}\mathbf{B}_t\mathbf{c}^\top + \mathbf{c}\mathrm{d}\mathbf{B}_t{}^\top\mathbf{V}\boldsymbol{\Sigma}, \\
&= \left(\boldsymbol{\sigma} \odot \mathrm{d}\tilde{\mathbf{B}}_t\right)\mathbf{c}^\top + \mathbf{c}\left(\boldsymbol{\sigma} \odot \mathrm{d}\tilde{\mathbf{B}}_t\right)^\top \\
&= \mathrm{d}\mathbf{m}_t\mathbf{c}^\top + \mathbf{c}\mathrm{d}\mathbf{m}_t{}^\top
\end{aligned}$$

where $\mathrm{d}\mathbf{m}_t \overset{\text{def}}{=} (\boldsymbol{\sigma} \odot \mathrm{d}\tilde{\mathbf{B}}_t)$. We are required to compute $\mathrm{d}\mathbf{F}(\mathbf{S} \odot \mathrm{d}\mathbf{F})$ to compute the evolution of eigenvalues. Using the lemma D.4, we get

$$\begin{aligned}
\mathrm{d}\mathbf{F}(\mathbf{S} \odot \mathrm{d}\mathbf{F}) &= \mathbf{c}\mathbf{s}^\top\mathbf{S}\mathrm{diag}\left(\mathbf{c}\right)\mathrm{d}t - \mathbf{D}\mathrm{diag}\left(\mathbf{S}\mathrm{diag}\left(\mathbf{c}\right)\mathbf{c}\right)\mathrm{d}t + \mathbf{D}\mathrm{diag}\left(\mathbf{c}\right)\mathbf{S}\mathrm{diag}\left(\mathbf{c}\right)\mathrm{d}t, \\
\mathrm{I} \odot \left[\mathrm{d}\mathbf{F}(\mathbf{S} \odot \mathrm{d}\mathbf{F})\right] &= \mathrm{I} \odot \left[\mathbf{c}\mathbf{s}^\top\mathbf{S}\mathrm{diag}\left(\mathbf{c}\right)\mathrm{d}t - \mathbf{D}\mathrm{diag}\left(\mathbf{S}\mathrm{diag}\left(\mathbf{c}\right)\mathbf{c}\right)\mathrm{d}t\right]
\end{aligned}$$

The element wise computation of this term gives the required result for evolution of eigenvalues.

Evolution of c. Note that $c = \mathbf{V}^\top\mathbf{a}$. Computing the derivative using the Ito's product rule, we get,

$$\begin{aligned}
\mathrm{d}\mathbf{V}^\top\mathbf{a} &= \mathbf{V}^\top\mathrm{d}\mathbf{a} + \mathrm{d}\mathbf{V}^\top\mathbf{a} + \mathrm{d}\mathbf{V}^\top\mathrm{d}\mathbf{a}, \\
&= \mathbf{V}^\top\mathrm{d}\mathbf{a} + \mathrm{d}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathbf{a} + \mathrm{d}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathrm{d}\mathbf{a}, \\
\mathrm{d}\mathbf{V}^\top\mathbf{V} &= \left[\left(\mathbf{Q}_{\|}^\top\,\mathrm{d}t - \mathbf{S} \odot \mathrm{d}\mathbf{X}\right)\right], \\
\mathbf{V}^\top\mathrm{d}\mathbf{a} &= \mathbf{V}^\top\mathbf{W}^\top\mathrm{d}\mathbf{B}_t + \frac{1}{\eta\delta}\left[\mathbf{U}^\top y - \boldsymbol{\Sigma}\mathbf{c}\right]\mathrm{d}t = \boldsymbol{\Sigma}\mathrm{d}\tilde{\mathbf{B}}_t = \mathrm{d}\mathbf{m}_t + \frac{1}{\eta\delta}\left[\mathbf{U}^\top y - \boldsymbol{\Sigma}\mathbf{c}\right]\mathrm{d}t, \\
\mathrm{d}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathrm{d}\mathbf{a} &= -(\mathbf{S} \odot \mathrm{d}\mathbf{F})\mathrm{d}\mathbf{m}_t. \\
\mathrm{d}\mathbf{V}^\top\mathbf{V}\mathbf{V}^\top\mathrm{d}\mathbf{a} &= \left[\left(\mathbf{Q}_{\|}^\top\,\mathrm{d}t - \mathbf{S} \odot \left(\widetilde{\mathbf{N}}\mathrm{d}t + \mathrm{d}\widetilde{\mathbf{M}}\right)\right)\right]\mathbf{c}
\end{aligned}$$

Using the lemma D.6, D.5, D.4 and computing the element wise summation, we get the following evolution for $\mathrm{d}\mathbf{c}$

$$\begin{aligned}
\mathrm{d}\mathbf{c}_i = &-\frac{1}{2}\sum_{j=1}^{l}\mathbf{S}_{ij}(\mathbf{s}_i\mathbf{c}_j^2 + \mathbf{s}_j\mathbf{c}_i^2)\mathrm{d}t - \mathbf{c}_i\sum_{j=1}^{l}(\mathbf{S}_{ij}\mathbf{c}_j^2)\left(\sum_{k\neq i,j}\mathbf{s}_k\mathbf{S}_{ki}\right) \\
&- (p-2)\mathbf{c}_i\sum_{j=1}^{l}\mathbf{S}_{ij}\mathbf{c}_i^2\mathrm{d}t - \sum_{j=1}^{l}\mathbf{S}_{ij}\mathbf{s}_j\mathrm{d}t, \\
&+ \boldsymbol{\sigma}_i(\mathbf{U}^\top\mathrm{d}\mathbf{X})_i(1 - \sum_{j=1}^{l}\mathbf{S}_{ij}\mathbf{c}_j^2) - \mathbf{c}_i\sum_{j}\mathbf{S}_{ij}\boldsymbol{\sigma}_j\mathbf{c}_j(\mathbf{U}^\top\mathrm{d}\mathbf{X})_j
\end{aligned}$$

Evolution of U. To compute the evolution of $\mathbf{U}$, we invoke the theorem D.2 on the evolution of $\mathbf{W}\mathbf{W}^\top = \mathbf{U}\mathbf{D}\mathbf{U}^\top$. We ignore it here as it does not have much consequence on our results. □

Theorem B.6. *In the large noise limit, when $l = 2$, the following properties hold, for $t \leq \tau$,*

(a) $\mathbf{s}_0, \mathbf{s}_1$ *are greater than zero almost surely.*

(b) for $\alpha = (p-3)/2$, $\mathbf{s}_0^{-\alpha}$ *is a super-martingale while* $\mathbf{s}_1^{-\alpha}$ *is a sub-martingale.*

Proof. First, note that in the large noise limit with $l = 2$, the evolution of the eigenvalues is expressed as

$$\mathrm{d}(\mathbf{s}_0) = p\mathbf{c}_0^2 \mathrm{d}t + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_0, \tag{B.7}$$

$$\mathrm{d}(\mathbf{s}_1) = p\mathbf{c}_1^2 \mathrm{d}t - \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_1\mathbf{c}_1^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_1. \tag{B.8}$$

Using the Ito chain rule, for the evolution of $\mathbf{s}_0^{-\alpha}$ we can write

$$\begin{aligned}
\mathrm{d}\big(\mathbf{s}_0^{-\alpha}\big) = \frac{\partial(\mathbf{s}_0^{-\alpha})}{\partial \mathbf{s}_0}\left(p\mathbf{c}_0^2\mathrm{d}t + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_0\right) + \frac{1}{2}\frac{\partial^2(\mathbf{s}_0^{-\alpha})}{\partial^2\mathbf{s}_0}\left(2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\right)^2\mathrm{d}t \\
= -\alpha\mathbf{s}_0^{-\alpha-1}\left(p\mathbf{c}_0^2\mathrm{d}t + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t - 2(\alpha+1)\mathbf{c}_0^2\mathrm{d}t + 2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_0\right) \\
= -\alpha\mathbf{s}_0^{-\alpha-1}\left(\mathbf{c}_0^2\mathrm{d}t + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_0\right),
\end{aligned}$$

analogously

$$\mathrm{d}\big(\mathbf{s}_1^{-\alpha}\big) = -\alpha\mathbf{s}_1^{-\alpha-1}\left(\mathbf{c}_1^2\mathrm{d}t - \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_1\mathbf{c}_1^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_1\right),$$

and finally for $\mathbf{s}_0^{-\alpha}\mathbf{s}_1^{-\alpha}$

$$\begin{aligned}
\mathrm{d}\big(\mathbf{s}_0^{-\alpha}\mathbf{s}_1^{-\alpha}\big) = \mathrm{d}\big(\mathbf{s}_0^{-\alpha}\big)\mathbf{s}_1^{-\alpha} + \mathbf{s}_0^{-\alpha}\mathrm{d}\big(\mathbf{s}_1^{-\alpha}\big) + \mathrm{d}\big(\mathbf{s}_0^{-\alpha}\big)\mathrm{d}\big(\mathbf{s}_1^{-\alpha}\big) \\
= -\alpha\mathbf{s}_0^{-\alpha-1}\mathbf{s}_1^{-\alpha}\left(\mathbf{c}_0^2\mathrm{d}t + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_0\mathbf{c}_0^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_0\right) \\
-\alpha\mathbf{s}_0^{-\alpha}\mathbf{s}_1^{-\alpha-1}\left(\mathbf{c}_1^2\mathrm{d}t - \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\mathrm{d}t + 2\sqrt{\mathbf{s}_1\mathbf{c}_1^2}\left(\mathrm{d}\tilde{\mathbf{B}}_t\right)_1\right).
\end{aligned}$$

Now, we can show that the drift term in the SDE that describes the dynamics of $\mathbf{s}_0^{-\alpha}\mathbf{s}_1^{-\alpha}$ is zero, which gives us the first part of the result by Mckean's argument [Mayerhofer et al., 2011],

$$\begin{aligned}
-\alpha\mathbf{s}_0^{-\alpha-1}\mathbf{s}_1^{-\alpha-1}\left(\mathbf{s}_1\mathbf{c}_0^2 + \mathbf{s}_1\frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1} + \mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_0\frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\right) \\
= -\alpha\mathbf{s}_0^{-\alpha-1}\mathbf{s}_1^{-\alpha-1}\left(\mathbf{s}_1\mathbf{c}_0^2 + \mathbf{s}_0\mathbf{c}_1^2 + \frac{\mathbf{s}_0\mathbf{s}_1\mathbf{c}_1^2 + \mathbf{s}_1^2\mathbf{c}_0^2 - \mathbf{s}_0^2\mathbf{c}_1^2 + \mathbf{s}_0\mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1}\right) \\
= -\alpha\mathbf{s}_0^{-\alpha-1}\mathbf{s}_1^{-\alpha-1}\left(\mathbf{s}_1\mathbf{c}_0^2 + \mathbf{s}_0\mathbf{c}_1^2 + \frac{(\mathbf{s}_1 - \mathbf{s}_0)\left(\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2\right)}{\mathbf{s}_0 - \mathbf{s}_1}\right) = 0.
\end{aligned}$$

The second part is obtained by noticing that

$$\begin{aligned}
\mathbf{c}_0^2 + \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1} = \frac{\mathbf{s}_0\left(\mathbf{c}_1^2 + \mathbf{c}_0^2\right)}{\mathbf{s}_0 - \mathbf{s}_1} \geq 0, \\
\mathbf{c}_1^2 - \frac{\mathbf{s}_0\mathbf{c}_1^2 + \mathbf{s}_1\mathbf{c}_0^2}{\mathbf{s}_0 - \mathbf{s}_1} = -\frac{\mathbf{s}_1\left(\mathbf{c}_1^2 + \mathbf{c}_0^2\right)}{\mathbf{s}_0 - \mathbf{s}_1} \leq 0,
\end{aligned}$$

and hence the drift term of $\mathrm{d}\big(\mathbf{s}_0^{-\alpha}\big)$ is not positive, while the drift term of $\mathrm{d}\big(\mathbf{s}_1^{-\alpha}\big)$ is not negative.

□

C Experiment details

In all the graphs we plot the values averaged on 20 runs with different random seeds as well as the 95% confidence interval (lightly colored). To numerically emulate GF (Figure 1), we set a stepsize of $1e^{-6}$ in numerical simulation.

In the further experiments, we study the behaviour of the linear network for regression with the same synthetic data and same network initialization as in previous experiment. As seen in the left plot of the Figure 2, when the stepsize is large ($\eta = 0.1$), singular values exhibit behavior similar to the case of LNGF, while with the small stepsize ($\eta = 0.005$) the evolution of singular values is closer to GF case. Next, we examine the effect of SGD in the case of classification task with logistic loss, as illustrated in the middle plot of the Figure 2. We consider synthetic data with $n = 1000$ samples of Gaussian data in $\mathbb{R}^5$ ($d = 5$) constituting two clusters corresponding to two classes ($k = 1$). Note that larger stepsize ($\eta = 0.5$) in this case also forces the smallest singular value to tend to zero, however the effect is not so dramatic for the rest of singular values. Additionally, we study the 2-layer ReLu network optimized with SGD on the same regression task as before. As seen in the right plot of the Figure 2, the decrease of the last singular value σ_4 is much slower than in the case of the linear network, however, the larger stepsize still facilitates divergence of k largest (σ_0 and σ_1) and $p - k$ smallest (σ_2, σ_3 and σ_4) singular values.

All experiments are implemented with Python 3 [Van Rossum and Drake, 2009] under PSF license, NumPy [Harris et al., 2020] under BSD license, and PyTorch [Paszke et al., 2019] under BSD-3-Clause license.

The experiments were run on a Intel i5-8250U, 8-GB RAM, with OS Ubuntu 20.04.6.

D Supplementary material

D.1 Notations and preliminary definitions

Definition D.1 (Eigen decomposition and Singular Value decomposition)**.** *We discuss the eigen value decomposition for a symmetric square matrix, and the singular value decompostion for any matrix is defined as the following*

(a) ***Eigen decomposition**. For any rank r matrix $\mathbf{R} \in S_p$, $\mathbf{R} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ is the eigen decomposition, where $\mathbf{V} \in \mathbb{R}^{p\times r}$, $\mathbf{D} \in \mathbb{R}^{r\times r}$, $\mathbf{D}$ is a diagonal matrix and $\mathbf{V}^\top\mathbf{V} = \mathrm{I}_r$, however, $\mathbf{V}\mathbf{V}^\top$ is not necessarily an identity matrix unless $r = p$.*

(b) ***Singular Value Decomposition**. For any rank r matrix $\mathbf{W} \in \mathbb{R}^{p\times l}$, $\mathbf{W} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{p\times r}, \mathbf{V} \in \mathbb{R}^{l\times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, $\boldsymbol{\Sigma}$ is a diagonal matrix and $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathrm{I}_r$, however the $\mathbf{U}\mathbf{U}^\top$ and $\mathbf{V}\mathbf{V}^\top$ are not necessarily identity unless $r = p$ or $r = l$ respectively.*

D.2 Eigenvalues of matrix valued stochastic process

Theorem D.2. *For a matrix-valued stochastic process on S^{++}_{p+k},*

$$\mathrm{d}\mathbf{R} = \mathbf{N}\mathrm{d}t + \mathrm{d}\mathbf{M}$$

where $\mathrm{d}\mathbf{M}$ is a local martingale process. Let $R = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ is the eigenvalue decomposition of the process, the evolution of eigenvalues satisfy the SDE for time t less than the collision time,

$$\mathrm{d}\mathbf{D} = \mathrm{I} \odot \widetilde{\mathbf{N}}\,\mathrm{d}t + \mathrm{I} \odot \mathrm{d}\widetilde{\mathbf{M}}\,\mathrm{d}t + \mathrm{I} \odot \left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right) + \mathbf{D}^{-1} \odot \left(\mathbf{V}^\top \mathrm{d}\mathbf{R}\left(\mathrm{I} - \mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right).$$

where $\mathbf{S}$ is defined as per Eq. D.1 and $\mathrm{d}\widetilde{\mathbf{M}} = \mathbf{V}^\top\mathrm{d}\mathbf{M}\mathbf{V}, \widetilde{\mathbf{N}} = \mathbf{V}^\top\mathbf{N}\mathbf{V}$. The evolution of the eigenvectors,

$$\mathrm{d}\mathbf{V} = \mathbf{V}\left(\mathbf{Q}_{||}\,\mathrm{d}t + \mathbf{S} \odot \mathrm{d}\mathbf{F}\right) + (\mathrm{I} - \mathbf{V}\mathbf{V}^\top)\left(\mathbf{Q}_{\perp}\,\mathrm{d}t + \mathrm{d}\mathbf{R}\,\mathbf{V}\mathbf{D}^{-1}\right).$$

where you define,

$$\begin{aligned}\mathbf{Q}_{\|} &= \frac{\mathrm{I}\odot\left[\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right]}{2} - \frac{\mathrm{I}\odot\left[\mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]}{2} \\ &\quad - \mathbf{S}\odot\left[\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}}\odot\mathrm{I}\right]\right] + \mathbf{S}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) \\ &\quad + \mathbf{S}\odot\left(\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right), \\ \mathbf{Q}_{\perp} &= \left[\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\left[\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right]\mathbf{D}-\mathrm{d}\widetilde{\mathbf{M}}\right]\mathbf{D}^{-1}.\end{aligned}$$

Evolution of eigenvalues for general matrix SDE

Proof. Using the eigen decomposition, we have $\mathbf{R}=\mathbf{V}\mathbf{D}\mathbf{V}^{\top}$,

$$\begin{aligned}\mathbf{D} &= \mathbf{V}^{\top}\mathbf{R}\mathbf{V}, \\ \mathrm{d}\mathbf{D} &= \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathbf{V}^{\top}\mathbf{R}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathbf{R}\mathbf{V} + \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathbf{R}\mathrm{d}\mathbf{V}, \\ &= \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathbf{D}\mathbf{V}^{\top}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} + \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \left(\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\right)\mathbf{D}\left(\mathbf{V}^{\top}\mathrm{d}\mathbf{V}\right).\end{aligned}$$

The approach we follow is use the jacobian of the evolution of $\mathbf{V}$ (see [Townsend, 2016]) and solve the constrains equations to obtain the Ito correction term as done in Bru [1989]. Let $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_r)$ denote the diagonal entries of $\mathbf{D}$. Furthermore, we define the matrix $\mathbf{S}$, which plays a notable role in Jacobian w.r.t $\mathbf{V}$, as the following,

$$\mathbf{S}_{ij} = \begin{cases} 0 & \text{if } i=j, \\ (\mathbf{s}_j-\mathbf{s}_i)^{-1} & \text{o.w.} \end{cases} \tag{D.1}$$

For the sake of brevity, we denote the evolution

$$\begin{aligned}\mathrm{d}\mathbf{F} \overset{\text{def}}{=} \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} &= \mathbf{V}^{\top}\mathbf{N}\mathbf{V}\,\mathrm{d}t + \mathbf{V}^{\top}\mathrm{d}\mathbf{M}\mathbf{V}, \\ &\overset{\text{def}}{=} \widetilde{\mathbf{N}}\,\mathrm{d}t + \mathrm{d}\widetilde{\mathbf{M}}\end{aligned}$$

The evolution of the eigenvectors,

$$\mathrm{d}\mathbf{V} = \mathbf{V}\mathrm{d}\Omega_{\mathbf{V}} + (\mathrm{I}-\mathbf{V}\mathbf{V}^{\top})\mathrm{d}\Xi_{\mathbf{V}}.$$

Using the Jacobian of the eigen vectors, we write,

$$\begin{aligned}\mathrm{d}\Omega_{\mathbf{V}} &= \mathbf{Q}_{\|}\,\mathrm{d}t + \mathbf{S}\odot\mathrm{d}\mathbf{F}, \\ \mathrm{d}\Xi_{\mathbf{V}} &= \mathbf{Q}_{\perp}\,\mathrm{d}t + \mathrm{d}\mathbf{R}\,\mathbf{V}\mathbf{D}^{-1}.\end{aligned}$$

Note that $\mathbf{V}^{\top}\mathbf{V}=\mathrm{I}_r$, using this we have,

$$\begin{aligned}0 = \mathrm{d}\big(\mathbf{V}^{\top}\mathbf{V}\big) &= \mathrm{d}\mathbf{V}^{\top}\mathbf{V} + \mathbf{V}^{\top}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{V}, \\ &= \mathrm{d}\Omega_{\mathbf{V}}{}^{\top} + \mathrm{d}\Omega_{\mathbf{V}} + \mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{V}, \\ &= \mathrm{d}\Omega_{\mathbf{V}}{}^{\top} + \mathrm{d}\Omega_{\mathbf{V}} + \mathrm{d}\Omega_{\mathbf{V}}{}^{\top}\mathrm{d}\Omega_{\mathbf{V}} + \mathrm{d}\Xi_{\mathbf{V}}{}^{\top}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\Xi_{\mathbf{V}}, \\ &= \mathrm{d}\Omega_{\mathbf{V}}{}^{\top} + \mathrm{d}\Omega_{\mathbf{V}} - (\mathbf{S}\odot\mathrm{d}\mathbf{F})(\mathbf{S}\odot\mathrm{d}\mathbf{F}) + \mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}.\end{aligned}$$

Using $\mathrm{d}\Omega_{\mathbf{V}}^{\top} = \mathbf{Q}_{\|}^{\top}\mathrm{d}t - \mathbf{S}\odot\mathrm{d}\mathbf{F}$, we have $\mathrm{d}\Omega_{\mathbf{V}}{}^{\top} + \mathrm{d}\Omega_{\mathbf{V}} = \left(\mathbf{Q}_{\|}^{\top} + \mathbf{Q}_{\|}\right)\mathrm{d}t$.

$$\left(\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^{\top}\right)\mathrm{d}t = \left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) - \mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}. \tag{D.2}$$

Coming back to the evolution of singular values,

$$\begin{aligned}\mathrm{d}\mathbf{D} &= \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathbf{D}\mathbf{V}^{\top}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} + \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathrm{d}\mathbf{V} + \mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \left(\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\right)\mathbf{D}\left(\mathbf{V}^{\top}\mathrm{d}\mathbf{V}\right). \\ &= \mathrm{d}\mathbf{F} + \left(\mathbf{D}\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^{\top}\mathbf{D}\right)\mathrm{d}t + \mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\mathbf{F}\right) - \left(\mathbf{S}\odot\mathrm{d}\mathbf{F}\right)\mathbf{D} + \mathrm{d}\Omega_{\mathbf{V}}^{\top}\mathbf{D}\mathrm{d}\Omega_{\mathbf{V}} \\ &\qquad + \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left[\mathbf{V}\mathrm{d}\Omega_{\mathbf{V}} + \left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\Xi_{\mathbf{V}}\right] + \left[\mathrm{d}\Omega_{\mathbf{V}}{}^{\top}\mathbf{V}^{\top} + \mathrm{d}\Xi_{\mathbf{V}}{}^{\top}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^{\top}\right)\right]\mathrm{d}\mathbf{R}\mathbf{V},\end{aligned}$$

$$\begin{aligned}\mathrm{d}\mathbf{D} = \mathrm{I}\odot \mathrm{d}\mathbf{F} &+ \left(\mathbf{D}\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^\top \mathbf{D}\right)\mathrm{d}t - \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) + \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) \\ &- \left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathrm{d}\widetilde{\mathbf{M}} + \mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}.\end{aligned} \tag{D.3}$$

Note that $\mathrm{d}\mathbf{D}$ is diagonal, hence, $\mathrm{I}\odot\mathrm{d}\mathbf{D} = \mathrm{d}\mathbf{D}$.

$$\begin{aligned}\mathrm{I}\odot\mathrm{d}\mathbf{D} = \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathrm{I}\odot\left(\mathbf{D}\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^\top \mathbf{D}\right)\mathrm{d}t - \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right] \\ + 2\mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + 2\mathrm{I}\odot\left(\mathbf{D}^{-1}\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\right)\end{aligned}$$

Note that $\mathrm{I}\odot(DM) = \mathrm{I}\odot(MD) = D\odot M$ for any matrix M and diagonal matrix D, using this property, we can simplify the above expression as,

$$\begin{aligned}\mathrm{d}\mathbf{D} = \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathbf{D}\odot\left(\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^\top\right)\mathrm{d}t - \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right] \\ + 2\mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + 2\mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\right)\end{aligned}$$

Using Eq. D.2, we have,

$$\begin{aligned}\mathbf{D}\odot\left(\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^\top\right)\mathrm{d}t &= \mathbf{D}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) - \mathbf{D}^{-1}\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right], \\ &= \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\right] - \mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right).\end{aligned}$$

Using this,

$$\begin{aligned}\mathrm{d}\mathbf{D} &= \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\right] - \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right] \\ &\qquad + 2\mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + \mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right), \\ &= \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D} - \mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right]\right] \\ &\qquad + 2\mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + \mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right), \\ &= \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathrm{I}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathrm{d}\widetilde{\mathbf{M}}\right] \\ &\qquad + 2\mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + \mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right), \\ &= \mathrm{I}\odot \mathrm{d}\mathbf{F} + \mathrm{I}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) + \mathbf{D}^{-1}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\right).\end{aligned}$$

Evolution of eigenvectors for general matrix SDE. Here, we derive the evolution of eigenvectors,

Using Eq. D.2, we have,

$$\left(\mathbf{Q}_{\|}\mathbf{D} + \mathbf{Q}_{\|}^\top\mathbf{D}\right)\mathrm{d}t = \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D} - \mathbf{D}^{-1}\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}$$

Now further using the constrain that $\mathrm{d}\mathbf{D}$ needs to be diagonal we get,

$$\begin{aligned}\left(\mathbf{D}\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^\top\mathbf{D}\right)\mathrm{d}t &= \mathrm{d}\mathbf{D} - \mathrm{I}\odot\mathrm{d}\mathbf{F} + \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) - \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) + \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathrm{d}\widetilde{\mathbf{M}} \\ &\qquad - \mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}. \\ \left(\mathbf{D}\mathbf{Q}_{\|} - \mathbf{Q}_{\|}\mathbf{D}\right)\mathrm{d}t &= \mathrm{d}\mathbf{D} - \mathrm{I}\odot\mathrm{d}\mathbf{F} - \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\mathbf{D} - \mathbf{D}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right] - \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) \\ &\qquad + \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathrm{d}\widetilde{\mathbf{M}} - \mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}, \\ &= \mathrm{d}\mathbf{D} - \mathrm{I}\odot\mathrm{d}\mathbf{F} - \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}}\odot\bar{\mathrm{I}}\right] - \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) \\ &\qquad + \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\mathrm{d}\widetilde{\mathbf{M}} - \mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}, \\ &= \mathrm{d}\mathbf{D} - \mathrm{I}\odot\mathrm{d}\mathbf{F} + \left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}}\odot\mathrm{I}\right] - \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right) \\ &\qquad - \mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}. \\ \bar{\mathrm{I}}\odot\left(\mathbf{D}\mathbf{Q}_{\|} - \mathbf{Q}_{\|}\mathbf{D}\right)\mathrm{d}t &= \bar{\mathrm{I}}\odot\left(\mathrm{d}\mathbf{D} - \mathrm{I}\odot\mathrm{d}\mathbf{F}\right) + \bar{\mathrm{I}}\odot\left[\left(\mathbf{S}\odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}}\odot\mathrm{I}\right]\right] - \bar{\mathrm{I}}\odot\left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S}\odot\mathrm{d}\widetilde{\mathbf{M}}\right)\right) \\ &\qquad - \bar{\mathrm{I}}\odot\left(\mathbf{V}^\top\mathrm{d}\mathbf{R}\left(\mathrm{I}-\mathbf{V}\mathbf{V}^\top\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right).\end{aligned}$$

$$\left(\bar{\mathrm{I}} \odot \mathbf{Q}_{\|}\right) \mathrm{d}t = \mathbf{S} \odot \left[-\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}} \odot \mathrm{I}\right] + \mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right) + \mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right].$$

Combing these, we get the diagonal and off diagonal terms of $\mathbf{Q}_{\|}$

$$\begin{aligned}
\left(\mathrm{I} \odot \mathbf{Q}_{\|}\right) \mathrm{d}t &= \frac{1}{2}\,\mathrm{I} \odot \left(\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^{\top}\right)\mathrm{d}t, \\
&= \frac{\mathrm{I} \odot \left[\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right]}{2} - \frac{\mathrm{I} \odot \left[\mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]}{2}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{Q}_{\|} = {} & \frac{\mathrm{I} \odot \left[\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right]}{2} - \frac{\mathrm{I} \odot \left[\mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]}{2} \\
& - \mathbf{S} \odot \left[\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\left[\mathrm{d}\widetilde{\mathbf{M}} \odot \mathrm{I}\right]\right] + \mathbf{S} \odot \left(\mathrm{d}\widetilde{\mathbf{M}}\left(\mathbf{S} \odot \mathrm{d}\widetilde{\mathbf{M}}\right)\right) \\
& + \mathbf{S} \odot \left(\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right).
\end{aligned}$$

Computing of $\mathbf{Q}_{\perp}$. Recalling the evolution of the eigenvectors,

$$\mathrm{d}\mathbf{V} = \mathbf{V}\mathrm{d}\Omega_{\mathbf{V}} + (\mathrm{I} - \mathbf{V}\mathbf{V}^{\top})\mathrm{d}\Xi_{\mathbf{V}}.$$

Using the Jacobian of the eigen vectors, we write,

$$\begin{aligned}
\mathrm{d}\Omega_{\mathbf{V}} &= \mathbf{Q}_{\|}\ \mathrm{d}t + \mathbf{S} \odot \mathrm{d}\mathbf{F}, \\
\mathrm{d}\Xi_{\mathbf{V}} &= \mathbf{Q}_{\perp}\ \mathrm{d}t + \mathrm{d}\mathbf{R}\ \mathbf{V}\mathbf{D}^{-1}, \\
\mathrm{d}\mathbf{V} &= \mathbf{V}\left[\mathbf{Q}_{\|}\ \mathrm{d}t + \mathbf{S} \odot \mathrm{d}\mathbf{F}\right] + (\mathrm{I} - \mathbf{V}\mathbf{V}^{\top})\left[\mathbf{Q}_{\perp}\ \mathrm{d}t + \mathrm{d}\mathbf{R}\ \mathbf{V}\mathbf{D}^{-1}\right], \\
\mathrm{d}\mathbf{V}^{\top} &= \left[\mathbf{Q}_{\|}^{\top}\ \mathrm{d}t - \mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{V}^{\top} + \left[\mathbf{Q}_{\perp}^{\top}\mathrm{d}t + \mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\right](\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}).
\end{aligned}$$

Using the fact that $\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{R} = 0$ and deriving it,

$$\begin{aligned}
0 &= \left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{R}, \\
0 &= \mathrm{d}\left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{R}\right], \\
\mathrm{d}\mathbf{R} &= \mathrm{d}\left(\mathbf{V}\mathbf{V}^{\top}\mathbf{R}\right), \\
&= \mathrm{d}\mathbf{V}\mathbf{V}^{\top}\mathbf{R} + \mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{R} + \mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{R} + \mathrm{d}\mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{R} + \mathrm{d}\mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{R} + \mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}, \\
\mathrm{d}\mathbf{R}\mathbf{V} &= \mathrm{d}\mathbf{V}\mathbf{D} + \mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} + \mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathrm{d}\mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} + \mathrm{d}\mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} + \mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V},
\end{aligned}$$

$$\begin{aligned}
\mathrm{d}\mathbf{V}\mathbf{D} &= \mathbf{V}\left[\mathbf{Q}_{\|}\mathbf{D}\ \mathrm{d}t + (\mathbf{S} \odot \mathrm{d}\mathbf{F})\,\mathbf{D}\right] + (\mathrm{I} - \mathbf{V}\mathbf{V}^{\top})\left[\mathbf{Q}_{\perp}\mathbf{D}\ \mathrm{d}t + \mathrm{d}\mathbf{R}\ \mathbf{V}\right], \\
\mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} &= \mathbf{V}\left[\mathbf{Q}_{\|}^{\top}\ \mathrm{d}t - \mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D}, \\
\mathrm{d}\mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathbf{V}\mathbf{D} &= -\mathbf{V}\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} - \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D}, \\
\mathrm{d}\mathbf{V}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} &= \mathbf{V}\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathrm{d}\mathbf{F} + \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\mathrm{d}\mathbf{F}, \\
\mathbf{V}\mathrm{d}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}\mathbf{V} &= -\mathbf{V}\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathrm{d}\mathbf{F} + \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top})\mathrm{d}\mathbf{R}\mathbf{V}.
\end{aligned}$$

Adding the terms up we get,

$$\begin{aligned}
& \mathbf{V}\left[\mathbf{Q}_{\|} + \mathbf{Q}_{\|}^{\top}\right]\mathbf{D}\mathrm{d}t + \left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{Q}_{\perp}\mathbf{D}\mathrm{d}t \\
& - \mathbf{V}\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} - \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} \\
& + \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\mathrm{d}\mathbf{F} + \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{\top}\mathrm{d}\mathbf{R}(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top})\mathrm{d}\mathbf{R}\mathbf{V} = 0.
\end{aligned}$$

$$\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{Q}_{\perp}\mathbf{D}\mathrm{d}t - \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} + \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\mathrm{d}\mathbf{F} = 0.$$

$$\begin{aligned}
\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{Q}_{\perp}\mathbf{D}\mathrm{d}t &= \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} - \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\mathrm{d}\mathbf{F}, \\
\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathbf{Q}_{\perp} &= \left[\left(\mathrm{I} - \mathbf{V}\mathbf{V}^{\top}\right)\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} - \mathrm{d}\mathbf{F}\right]\mathbf{D}^{-1}
\end{aligned}$$

$$\mathbf{Q}_{\perp} = \left[\mathrm{d}\mathbf{R}\mathbf{V}\mathbf{D}^{-1}\right]\left[\left[\mathbf{S} \odot \mathrm{d}\mathbf{F}\right]\mathbf{D} - \mathrm{d}\mathbf{F}\right]\mathbf{D}^{-1}$$

This gives the expression for $\mathbf{Q}_{\perp}$ and this ends our computation. □

Lemma D.3. *For any matrix $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{n\times n}$, $m \times n$-dimensional Brownian motion $\mathrm{d}\mathbf{B}_t$, the following results hold on the covariance*

$$\mathrm{d}\mathbf{B}_t A \mathrm{d}\mathbf{B}_t = A^\top \mathrm{d}t, \tag{D.4}$$

$$\mathrm{d}\mathbf{B}_t B \mathrm{d}\mathbf{B}_t{}^\top = \mathrm{tr}\left(B\right) \mathrm{I}_m \mathrm{d}t. \tag{D.5}$$

Lemma D.4. *With $\mathbf{S}$ defined in Equation (D.1), $\mathbf{dF} = \mathbf{dF} = \mathbf{\Sigma}\mathbf{V}^\top \mathbf{dB}_t \mathbf{c}^\top + \mathbf{c}\mathbf{dB}_t{}^\top \mathbf{V}\mathbf{\Sigma}$ and $\mathbf{dm}_t \overset{\mathrm{def}}{=} (\boldsymbol{\sigma} \odot \mathrm{d}\tilde{\mathbf{B}}_t)$.*

$$\mathbf{dF}(\mathbf{S} \odot \mathbf{dF}) = \mathbf{c}\mathbf{s}^\top \mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t - \mathbf{D}\operatorname{diag}\left(\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathbf{c}\right)\mathrm{d}t + \mathbf{D}\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t. \tag{D.6}$$

Proof.

$$\begin{aligned}
\mathbf{S} \odot \mathbf{dF} &= \left[\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\right], \\
\mathbf{dF}(\mathbf{S} \odot \mathbf{dF}) &= \left(\mathbf{c}\mathbf{dm}_t{}^\top + \mathbf{dm}_t\mathbf{c}^\top\right)\left[\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\right], \\
&= \mathbf{c}\mathbf{s}^\top \mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t - \mathbf{D}\operatorname{diag}\left(\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathbf{c}\right)\mathrm{d}t + \mathbf{D}\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t.
\end{aligned}$$

□

Lemma D.5. *With $\mathbf{S}$ defined in Equation (D.1), $\mathbf{dF} = \mathbf{dF} = \mathbf{\Sigma}\mathbf{V}^\top \mathbf{dB}_t \mathbf{c}^\top + \mathbf{c}\mathbf{dB}_t{}^\top \mathbf{V}\mathbf{\Sigma}$ and $\mathbf{dm}_t \overset{\mathrm{def}}{=} (\boldsymbol{\sigma} \odot \mathrm{d}\tilde{\mathbf{B}}_t)$.*

$$(\mathbf{S} \odot \mathbf{dF})(\mathbf{S} \odot \mathbf{dF}) = \mathbf{D}\operatorname{diag}\left(\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)^2\mathbf{S}\right)\mathrm{d}t + \operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\mathbf{D}\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t. \tag{D.7}$$

Proof.

$$\begin{aligned}
(\mathbf{S} \odot \mathbf{dF}) = \mathbf{S} \odot \left(\mathbf{dm}_t\mathbf{c}^\top + \mathbf{c}\mathbf{dm}_t{}^\top\right), \\
&= \operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right).
\end{aligned}$$

Now, computing the product,

$$\begin{aligned}
(\mathbf{S} \odot \mathbf{dF})(\mathbf{S} \odot \mathbf{dF}) &= \left[\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\right]\left[\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\right], \\
&= \mathbf{D}\operatorname{diag}\left(\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)^2\mathbf{S}\right)\mathrm{d}t + \operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\mathbf{D}\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\mathrm{d}t.
\end{aligned}$$

□

Lemma D.6.

$$(S \odot \mathbf{dF})\mathbf{dm}_t\mathbf{c}^\top\mathbf{dF} =$$

Proof.

$$(S \odot \mathbf{dF})\mathbf{dm}_t = \left[\operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}\operatorname{diag}\left(\mathbf{dm}_t\right) + \operatorname{diag}\left(\mathbf{dm}_t\right)\mathbf{S}\operatorname{diag}\left(\mathbf{c}\right)\right]\mathbf{dm}_t = \operatorname{diag}\left(\mathbf{c}\right)\mathbf{S}(\boldsymbol{\sigma} \odot \boldsymbol{\sigma})$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The dichotomy between SGD and GD is revealed by theorems 4.1 and 4.2, the repulsive force between the eigenvalues of parameter matrix is discussed in theorems 5.1 and 5.2. Supporting experiments are discussed in section 7.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in section 4 in a designated paragraph as well as in section 5 in the discussion of theorem 5.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs of the theorems that are not presented in the main body are presented in appendix B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data as well as the models and optimization algorithms used are discussed in the section 7

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code in the form of Jupyter notebook is provided and all the random seeds are fixed for the reproducibility purposes.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The setup is discussed in the section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All the graphs of the parameters evolution are accompanied with the 95% confidence interval calculated on the 20 runs with different random seeds.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All relevant information is stated in the appendix section C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical and does not suggest any new model that can cause harm. All the data used is synthetic.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work investigates the effects of well known algorithms on the simple models and doesn't suggest any new applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work doesn't entail models or datasets releases.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All relevant information is stated in the appendix section C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.