UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Characterizing Dynamical Stability of Stochastic Gradient Descent in Overparameterized Learning

Chemnitz, Dennis; Engel, Maximilian

Publication date
2025
Document Version
Final published version
Published in
Journal of Machine Learning Research
License
CC BY

Link to publication

Citation for published version (APA):
Chemnitz, D., & Engel, M. (2025). Characterizing Dynamical Stability of Stochastic Gradient Descent in Overparameterized Learning. *Journal of Machine Learning Research*, *26*. https://jmlr.org/papers/v26/24-1547.html

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:15 Oct 2026

Journal of Machine Learning Research 26 (2025) 1-46 Submitted 9/24; Revised 7/25; Published 7/25

Characterizing Dynamical Stability of Stochastic Gradient Descent in Overparameterized Learning

Dennis Chemnitz DENNIS.CHEMNITZ2@FU-BERLIN.DE
Fachbereich Mathematik und Informatik
Freie Universität Berlin
14195 Berlin, Germany

Maximilian Engel M.R.ENGEL@UVA.NL
KdV Institute for Mathematics
University of Amsterdam
1090 GE Amsterdam, Netherlands
and
Fachbereich Mathematik und Informatik
Freie Universität Berlin
14195 Berlin, Germany

Editor: Gergely Neu

Abstract

For overparameterized optimization tasks, such as those found in modern machine learning, global minima are generally not unique. In order to understand generalization in these settings, it is vital to study to which minimum an optimization algorithm converges. The possibility of having minima that are unstable under the dynamics imposed by the optimization algorithm limits the potential minima that the algorithm can find. In this paper, we characterize the global minima that are dynamically stable/unstable for both deterministic and stochastic gradient descent (SGD). In particular, we introduce a characteristic Lyapunov exponent that depends on the local dynamics around a global minimum and rigorously prove that the sign of this Lyapunov exponent determines whether SGD can accumulate at the respective global minimum.

Keywords: machine learning, stochastic gradient descent, overparameterization, linear stability, Lyapunov exponents

1. Introduction

Since the success of "AlexNet" (Krizhevsky et al., 2012), the overparameterization paradigm has become ubiquitous in modern machine learning. Although it is well known that artificial neural networks with sufficiently many parameters can approximate arbitrary goal functions (see, e.g., Hornik et al., 1989), the training process, especially for overparameterized networks, is not well understood and leaves many open questions. For one, the fact that the loss landscapes are usually non-convex makes it difficult to rigorously guarantee that the optimization algorithms converge to global minima.

Another key open problem in understanding the training of overparameterized networks is that of generalization (see, e.g., Zhang et al., 2021). Classical wisdom suggests that overparameterized networks should suffer from overfitting problems due to their high expressivity. However, in practice, this is not the case. In other words, sufficiently large (that is, deep, wide, or both) artificial neural networks are capable of fitting arbitrary data, due to their large number of pa-

©2025 Dennis Chemnitz and Maximilian Engel.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/24-1547.html.

rameters (typically weights and biases). Thus, finding a set of parameters for which the network interpolates the training data, in itself, does not indicate that the network makes reasonable predictions for inputs outside the training data set. However, the parameters found by common optimization algorithms tend to perform well on unseen data. This is essential for the success of modern machine learning.

One possible explanation for this phenomenon is that of dynamical stability (see, e.g., Wu et al., 2018). Since the learning rate is in practice not infinitesimally small, some global minima can become dynamically unstable, in the sense that the optimization algorithm will not converge to these solutions, even if it is initialized arbitrarily close to them. This excludes unstable optimal solutions from being found by the optimization algorithm, thereby reducing the effective hypothesis class to stable global minima.

For gradient descent, a second-order Taylor expansion shows that a global minimum is only stable if the largest eigenvalue of the Hessian of the empirical loss is less than $2/\eta$, where η is the learning rate. Thus, gradient descent can only converge to minima that are sufficiently flat, which has been associated with good generalization (see, e.g., Hochreiter and Schmidhuber, 1997). Moreover, numerical experiments by Cohen et al. (2021) have demonstrated that (deterministic) gradient descent operates at the *edge of stability*, which means that the largest eigenvalue of the Hessian oscillates around the critical value $2/\eta$ for the later stages of training.

Transferring these ideas to the stochastic case requires an appropriate notion of dynamical stability under stochastic gradient descent (see Wu et al., 2018 and Ma and Ying, 2021). The goal of this work is to establish a coherent mathematical framework in which the dynamical stability of deterministic and stochastic optimization algorithms can be investigated and the convergence properties of global minima can be characterized.

1.1 Contributions

We investigate which global minima of the empirical loss function can be obtained as limits of the two most fundamental optimization algorithms, namely gradient descent (GD) and stochastic gradient descent (SGD). Specifically, we consider the limits of GD and SGD as random variables, denoted by $X_{\lim}^{\mathrm{GD}}$ and $X_{\lim}^{\mathrm{SGD}}$, respectively. These random variables depend on the randomness during initialization in the case of GD and on the randomness during initialization and training in the case of SGD. For a given global minimum x^* of the empirical loss $\mathcal{L}$, we introduce the quantity

$$\mu(x^*) := \log \|\mathbb{1} - \eta \operatorname{Hess} \mathcal{L}(x^*)\|.$$

As is readily seen in Section 2.3, the above condition $\|\operatorname{Hess} \mathcal{L}(x^*)\| < \frac{2}{\eta}$ is equivalent to $\mu(x^*) < 0$. We choose this reformulation because it has a direct analogue for SGD.

Our main contributions are the following.

- We introduce a set of mild conditions (see (H1), (H2), (H3) below) under which we can rigorously prove that the sign of μ characterizes the support of $X_{\lim}^{\mathrm{GD}}$ (see Theorem A).
- For global minima x^*, we introduce a new quantity, denoted by $\lambda(x^*)$ (see (2.12) below), which can be seen as a characteristic Lyapunov exponent (Oseledec, 1968) and corresponds to a new notion of dynamic stability/instability for SGD. It should be noted that this notion differs[1] from the ones introduced in Wu et al. (2018) and Ma and Ying (2021) (see Appendix A).

1. In particular our notion of linear stability is strictly weaker than the one introduced in (Wu et al., 2018, Defintion 1) (see Appendix A).

- Under an additional mild assumption on the global minimum in question, we rigorously prove that the sign of λ characterizes the support of $X_{\text{lim}}^{\text{SGD}}$ (see Theorem B).

This clarifies the concept of dynamical stability for stochastic gradient descent and establishes a connection to a substantial body of mathematical and physical literature on Lyapunov exponents (see, e.g., Oseledec, 1968; Bougerol and Lacroix, 1985; Arnold, 1998). Moreover, it provides a foundation for new theoretic and numerical investigations of the learning process under stochastic gradient descent. Possible questions of interest include how the value of λ evolves during training with stochastic gradient descent and what implications a small, respectively large, value of λ has on generalization.

From a mathematical perspective, our analysis of SGD amounts to a study of the asymptotic behavior of discrete-time random dynamical systems (see, e.g., Arnold, 1998 and also Section 3.5) possessing a manifold of equilibria. In the unstable case, we use a method based on moment Lyapunov exponents (see, e.g., Arnold et al., 1986). A more detailed discussion of this approach is given in Section 3.1 and throughout the proofs.

For simplicity, we restrict our study to scalar regression problems and, in the case of SGD, to mini-batch size 1, but we point out that, with minor modifications, our main result applies to a broader class of tasks and algorithms (see Section 2.7 below).

1.2 Related Literature

Closest to our work is a series of two papers by Wu et al. (2018) and Ma and Ying (2021), which study the dynamical stability of stochastic gradient descent. We explain the relation to this work in detail in Appendix A.

Gurbuzbalaban et al. (2021) consider Lyapunov exponents of SGD in a setting where SGD does not converge to a global minimum, but the transition densities approach a heavy-tailed stationary distribution (see also, e.g., Hodgkinson and Mahoney, 2021). The convergence rates for SGD with decaying learning rate have been studied in Fehrman et al. (2020).

The edge of stability phenomenon for gradient descent mentioned in the introduction has been numerically observed by Wu et al. (2018), as well as by Cohen et al. (2021) and theoretically studied by Arora et al. (2022). Moreover, the difference between the global minima found by GD and SGD (see Keskar et al., 2017; Wu et al., 2018) and between SGD and Adam (see Keskar and Socher, 2017) has been numerically investigated.

The dynamics of stochastic gradient descent have also been studied using approximations by stochastic differential equations, called stochastic modified equations (see, e.g., Li et al. (2017, 2019)). We note that such an approach is not compatible with our analysis as it assumes an infinitesimally small learning rate. Moreover, the modified stochastic equations have been extended to a stochastic flow in Gess et al. (2024), building a framework analogous to the one we introduce in Section 3.5.

The most critical part of our proof is based on techniques originally developed to study synchronization in stochastic differential equations by Baxendale and Stroock (1988) (see also Baxendale, 1991), which has received recent attention motivated by fluid dynamics (Coti Zelati and Hairer, 2021; Bedrossian et al., 2022, 2025; Blumenthal et al., 2023).

1.3 Organization of the Paper

We start by introducing the fundamental setting of the optimization task (Section 2.1), as well as the learning algorithms that we consider (2.2). In Section 2.3, we heuristically derive the notions of linear stability for GD and SGD. These derivations are made rigorous in our main

results, which we present in Section 2.4. In Section 2.5, the main quantities of interest, μ and λ, are related to the neural tangent kernel (see Jacot et al., 2018). Finally, we discuss possible generalizations of our main results, such as extensions to other optimizers, in Section 2.7. Section 3 provides the proofs of the main theorems. An overview of the structure of the proofs is given in Section 3.1.

2. Setting and Main Result

2.1 Network Model

In the following, we consider a scalar regression problem. Let $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ be a ground truth function, which is supposed to be reconstructed from N given data pairs $(y_i, z_i = \hat{f}(y_i))_{i\in[N]}$, where $[N] = \{1, \dots, N\}$. To do so, we consider a parameterized network model given by a smooth function[2] $\mathfrak{F} : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}$ and try, using the given data, to find an $x \in \mathbb{R}^D$ such that $\mathfrak{F}(x, \cdot) \approx \hat{f}$. In the absence of other information, it is reasonable to prescribe

$$x \in \operatorname{argmin} \frac{1}{N} \sum_{i\in[N]} \ell(\mathfrak{F}(x, y_i), z_i), \tag{2.1}$$

for some loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, with $\ell(\bar{z}, z) = 0$ if and only if $\bar{z} = z$, $\partial_1\ell(z, z) = 0$ and $\partial_1^2\ell(z, z) = 1$. Although other choices for ℓ are possible, it is recommended to think of the square loss function $\ell(\bar{z}, z) := \frac{1}{2}(\bar{z} - z)^2$. For each $i \in [N]$, we define the *individual loss function* $\mathcal{L}_i : \mathbb{R}^D \to [0, \infty)$ by

$$\mathcal{L}_i(x) := \ell(\mathfrak{F}(x, y_i), z_i) \tag{2.2}$$

and the *empirical loss function* $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}$ as the average of the individual loss functions, that is,

$$\mathcal{L}(x) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i(x). \tag{2.3}$$

Equation (2.1) can thus be rewritten as

$$x \in \operatorname{argmin} \mathcal{L}(x). \tag{2.4}$$

If the number of given training examples N exceeds the number of learnable parameters D, the problem is overdetermined and the global minimum of $\mathcal{L}$ is usually unique, so that x is fully determined by (2.4). However, in general, it will not be possible to find an $x \in \mathbb{R}^D$ such that $\mathfrak{F}(x, \cdot)$ interpolates the given data, that is, the set

$$\mathcal{M} := \big\{x \in \mathbb{R}^D : \mathcal{L}(x) = 0\big\} = \big\{x \in \mathbb{R}^D : \mathfrak{F}(x, y_i) = z_i,\ \forall i \in [N]\big\}$$

will be empty.

On the other hand, if the number of learnable parameters exceeds the number of training examples, that is, $D > N$, the optimization problem is overparameterized. As described in the introduction, this will be the case of interest here. Given a sufficiently expressive network model $\mathfrak{F}$ (meaning that $\mathfrak{F}(x, \cdot)$ can express a sufficiently rich family of functions), the set of interpolation solutions $\mathcal{M}$ will usually be infinite. In fact, if the set of gradients

$$\{\nabla_x\mathfrak{F}(x, y_i) : i \in [N]\} \subset R^D$$

2. For the purpose of being general, we will not specify how the network function $\mathfrak{F}$ looks like. Typical network architectures used in practice include fully connected networks, convolutional neural networks (image classification), and transformers (large language models).

is linearly independent for every $x \in \mathcal{M}$, the set $\mathcal{M}$ is an embedded smooth $(D-N)$-dimensional submanifold of $\mathbb{R}^D$ with normal space

$$\mathcal{N}(x) = \operatorname{span}\{\nabla_x \mathfrak{F}(x, y_i) : i \in [N]\} \tag{2.5}$$

and tangent space

$$\mathcal{T}(x) = \mathcal{N}(x)^\perp = \{v \in \mathbb{R}^D : w^t v = 0,\ \forall\, w \in \mathcal{N}(x)\}. \tag{2.6}$$

As a consequence of Sard's theorem (see, e.g., Milnor, 1997), this condition will be satisfied for generic training data[3] and in the following we will assume that $\mathcal{M}$ is a smooth manifold as described (see Hypothesis (H1) below).

While $x \in \mathcal{M}$ ensures $\mathfrak{F}(x, y_i) = \hat{f}(y_i)$ for all $i \in [N]$, it does not guarantee $\mathfrak{F}(x, y) \approx \hat{f}(y)$ for any y outside the training data set. This is the generalization gap mentioned in the introduction. If $\mathcal{M}$ is infinite, (2.4) does not determine x uniquely and the question of generalization is highly dependent on which $x \in \mathcal{M}$ is chosen.

2.2 Learning Algorithms

In practice, a wide range of optimization algorithms (see, e.g., Schmidt et al., 2021) are used to find a global minimum $x^* \in \mathcal{M}$. In this work, we will restrict our study to the two most basic optimization algorithms: gradient descent (GD) and stochastic gradient descent (SGD). Both algorithms are initialized at a random point $X_0^{\mathrm{GD/SGD}} \in \mathbb{R}^D$, which is distributed according to some probability measure $\nu : \mathfrak{B}(\mathbb{R}^D) \to [0, 1]$, called the *initial distribution*.[4] We will assume that ν is equivalent to Lebesgue measure[5] (see Hypothesis (H2) below).

Gradient descent tries to successively reduce the empirical loss $\mathcal{L}$ by taking small steps in the opposite direction of its gradient $\nabla\mathcal{L}$. Formally, the update rule is given by

$$X_{n+1}^{\mathrm{GD}} := X_n^{\mathrm{GD}} - \eta \nabla \mathcal{L}(X_n^{\mathrm{GD}}), \tag{2.7}$$

where $\eta > 0$ is a small real number called *learning rate*. Ideally, the hope is that the gradient descent converges to some point $X_{\lim}^{\mathrm{GD}} \in \mathcal{M}$ as n tends to infinity. Although gradient descent is remarkably reliable in finding global minima in practice, convergence to a global minimum (or even convergence in the first place) cannot be guaranteed in our general setting, since $\mathcal{L}$ is generally non-convex. Therefore we define $X_{\lim}^{\mathrm{GD}}$ as an $\mathcal{M} \cup \{\emptyset\}$-valued random variable by

$$X_{\lim}^{\mathrm{GD}} := \begin{cases} \lim_{n\to\infty} X_n^{\mathrm{GD}} & , \text{ if } (X_n^{\mathrm{GD}}) \text{ converges to some point in } \mathcal{M}, \\ \emptyset & , \text{ otherwise,} \end{cases}$$

accounting for a possible failure of convergence. Note that the randomness in $X_{\lim}^{\mathrm{GD}}$ stems purely from the random initialization X_0^{GD}. Once X_0^{GD} is drawn, all subsequent steps, and thus $X_{\lim}^{\mathrm{GD}}$ are deterministic.

In contrast, stochastic gradient descent is a stochastic optimization algorithm. After initializing X_0^{SGD} randomly according to the measure ν, SGD updates are performed according to the

3. Put more precisely, for every smooth network model $\mathfrak{F}$ and every choice of $(y_1, \dots, y_N)$, there exists a set $\mathcal{Z} \subseteq \mathbb{R}^N$ of full Lebesgue measure, such that, if $(z_1, \dots, z_N) \in \mathcal{Z}$, then $\mathcal{M}$ is an embedded smooth $(D-N)$-dimensional manifold (see Cooper, 2021).
4. Here and in the following, $\mathfrak{B}(\mathbb{R}^D)$ denotes the family of Borel sets.
5. This is for example consistent with ν being a normal distribution, which is the most common choice in practice.

gradients of the individual loss functions $\mathcal{L}_i$ (2.2), corresponding to randomly chosen training data. Concretely, we set

$$X_{n+1}^{\mathrm{SGD}} := X_n^{\mathrm{SGD}} - \eta \nabla \mathcal{L}_{\xi_{n+1}}(X_n^{\mathrm{SGD}}), \tag{2.8}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables, which are uniformly distributed over $[N]$. Note that $\nabla \mathcal{L}_{\xi_n}$ can be interpreted as an unbiased estimator for the gradient $\nabla \mathcal{L}$ of the empirical loss function $\mathcal{L}$ (2.3) since

$$\mathbb{E}[\nabla \mathcal{L}_{\xi_n}(x)] = \nabla \mathcal{L}(x),$$

for each $x \in \mathbb{R}^D$ and each $n \in \mathbb{N}$. Unfortunately, the problems with guaranteeing convergence for SGD persist such that, again, we define $X_{\lim}^{\mathrm{SGD}}$ as a $\mathcal{M} \cup \{\emptyset\}$-valued random variable by

$$X_{\lim}^{\mathrm{SGD}} := \begin{cases} \lim_{n\to\infty} X_n^{\mathrm{SGD}} & , \text{ if } (X_n^{\mathrm{SGD}}) \text{ converges to some point in } \mathcal{M}, \\ \emptyset & , \text{ otherwise.} \end{cases}$$

In the following, we are interested in studying which points GD and SGD can find. Put more precisely, we aim to characterize the support of the random variable $X_{\lim}^{\mathrm{GD/SGD}}$, which is defined as

$$\operatorname{supp}\left(X_{\lim}^{\mathrm{GD/SGD}}\right) := \left\{ x \in \mathcal{M} : \mathbb{P}\left(X_{\lim}^{\mathrm{GD/SGD}} \in U\right) > 0, \ \forall\, U \subseteq \mathcal{M} \text{ open nbhd. of } x \right\}.$$

2.3 Linear Stability

If the (stochastic) gradient descent is randomly initialized at some $x \in \mathcal{M}$ or reaches $\mathcal{M}$ after some finite number of iterations, it remains stationary from then on, that is,

$$X_n^{\mathrm{GD/SGD}} \in \mathcal{M} \Rightarrow X_n^{\mathrm{GD/SGD}} = X_m^{\mathrm{GD/SGD}} = X_{\lim}^{\mathrm{GD/SGD}}, \ \forall\, m > n.$$

This can be seen easily by computing the individual and empirical loss functions as

$$\nabla \mathcal{L}_i(x) = \partial_1 \ell(\mathfrak{F}(x, y_i), z_i) \nabla_x \mathfrak{F}(x, y_i),$$
$$\nabla \mathcal{L}(x) = \frac{1}{N} \sum_{i=1}^{N} \partial_1 \ell(\mathfrak{F}(x, y_i), z_i) \nabla_x \mathfrak{F}(x, y_i)$$

and noting that the right-hand sides vanish if $x \in \mathcal{M}$. Since the initial distribution ν is assumed to have full support, it is possible that $X_{\lim}^{\mathrm{GD/SGD}}$ takes any value in $\mathcal{M}$. However, under mild assumptions (see (H2) and (H3) below), the optimization algorithm can only reach a global minimum in a finite number of steps with probability zero. In other words, (S)GD will usually converge to some global minimum, say $x^* \in \mathcal{M}$, without reaching it in a finite number of steps. This is only possible if x^* is *dynamically stable*, which means that, once the optimization algorithm gets sufficiently close to x^*, it will stay close.

In order to study the dynamical stability at some $x^* \in \mathcal{M}$, we *linearize* the optimization step (2.7) around x^*. We introduce the function $\mathcal{G}_\eta : \mathbb{R}^D \to \mathbb{R}^D$ by

$$\mathcal{G}_\eta(x) := x - \eta \nabla \mathcal{L}(x), \tag{2.9}$$

such that $X_{n+1}^{\mathrm{GD}} = \mathcal{G}(X_n^{\mathrm{GD}})$. If x is close to x^*, by differentiability, $\mathcal{G}_\eta(x)$ is well approximated by

$$\mathcal{G}_\eta(x) = x^* + \mathcal{G}_\eta'(x^*)(x - x^*) + o(\|x - x^*\|).$$

Here, $\mathcal{G}'_\eta(x^*) \in \mathbb{R}^{D\times D}$ denotes the Jacobian of $\mathcal{G}_\eta$ at x^*. Consequently, for several iterations, we have

$$\mathcal{G}^n_\eta(x) = x^* + \mathcal{G}'_\eta(x^*)^n(x - x^*) + o(\|x - x^*\|).$$

The Jacobian $\mathcal{G}'_\eta(x^*)$ can be computed as

$$\mathcal{G}'_\eta(x^*) = \mathbb{1}_D - \eta \operatorname{Hess} \mathcal{L}(x^*) = \mathbb{1}_D - \frac{\eta}{N}\sum_{i=1}^{N} \nabla_x \mathfrak{F}(x^*, y_i)\nabla_x \mathfrak{F}(x^*, y_i)^t. \tag{2.10}$$

Note that, by definition of the normal space $\mathcal{N}$ (2.5) and the tangent space $\mathcal{T}$ (2.6), the matrix $\mathcal{G}'_\eta(x^*)$ satisfies $\mathcal{G}'_\eta(x^*)v = v \in \mathcal{T}(x^*)$ for all $v \in \mathcal{T}(x^*)$ and $\mathcal{G}'_\eta(x^*)w \in \mathcal{N}(x^*)$ for all $w \in \mathcal{N}(x^*)$. Thus, $\mathcal{G}'_\eta(x^*)$ respects the splitting $\mathbb{R}^D = \mathcal{T}(x^*) \oplus \mathcal{N}(x^*)$ and the restriction $\mathcal{G}'_\eta(x^*)|_{\mathcal{T}(x^*)} : \mathcal{T}(x^*) \to \mathcal{T}(x^*)$ equals the identity. We say that x^* is *linearly stable* under GD with learning rate η if

$$\mu(x^*) := \lim_{n\to\infty} \frac{1}{n} \log(\|\mathcal{G}'_\eta(x^*)^n|_{\mathcal{N}(x^*)}\|) = \log(\rho_{\mathrm{Spec}}(\mathcal{G}'_\eta(x^*)|_{\mathcal{N}(x^*)})) < 0.$$

Here $\|\cdot\|$ denotes the operator norm and ρ_{Spec} the spectral radius of a matrix which is given by

$$\rho_{\mathrm{Spec}}(A) := \lim_{n\to\infty} \|A^n\|^{1/n} = \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}.$$

Conversely, if $\mu(x^*) > 0$, we say that x^* is *linearly unstable* under GD. By (2.10), we have

$$\operatorname{Spec}(\mathcal{G}'_\eta(x^*)|_{\mathcal{N}(x^*)}) = 1 - \eta \operatorname{Spec}(\operatorname{Hess} \mathcal{L}(x^*)|_{\mathcal{N}(x^*)}).$$

While the Hessian $\operatorname{Hess} \mathcal{L}(x^*)$ is only positive semi-definite, its restriction to $\mathcal{N}(x^*)$ is even positive definite and thus

$$\operatorname{Spec}(\operatorname{Hess} \mathcal{L}(x^*)|_{\mathcal{N}(x^*)}) \subset \mathbb{R}_{>0}.$$

In particular, the Hessian $\operatorname{Hess} \mathcal{L}(x^*)|_{\mathcal{N}(x^*)}$ is a symmetric positive definite matrix. Therefore, the condition for linear stability $\mu(x^*) < 0$ can be equivalently expressed as

$$\begin{aligned}\mu(x^*) < 0 &\Leftrightarrow \operatorname{Spec}(\mathcal{G}'_\eta(x^*)|_{\mathcal{N}(x^*)}) \subset (-1, 1) \Leftrightarrow \operatorname{Spec}(\operatorname{Hess} \mathcal{L}(x^*)|_{\mathcal{N}(x^*)}) \subset (0, 2/\eta) \\ &\Leftrightarrow \|\operatorname{Hess} \mathcal{L}(x^*)|_{\mathcal{N}(x^*)}\| < \frac{2}{\eta} \Leftrightarrow \|\operatorname{Hess} \mathcal{L}(x^*)\| < \frac{2}{\eta},\end{aligned}$$

where the last equivalence holds if $\eta < 2$. Although this formulation is more common in the literature (see, e.g., Wu et al., 2018; Cohen et al., 2021; Arora et al., 2022), we will stick to the expression $\mu(x^*) < 0$ here, as it can be more easily extended to stochastic gradient descent.

Our stability analysis for stochastic gradient descent is analogous to that for gradient descent. For notational convenience, we introduce the functions $\mathcal{G}_{\eta,i} : \mathbb{R}^D \to \mathbb{R}^D$ for each $i \in [N]$ by

$$\mathcal{G}_{\eta,i}(x) := x - \eta \nabla \mathcal{L}_i(x). \tag{2.11}$$

Their Jacobians at some global minimum $x^* \in \mathcal{M}$ can be computed as

$$\mathcal{G}'_{\eta,i}(x^*) = \mathbb{1}_D - \eta \operatorname{Hess} \mathcal{L}_i(x^*) = \mathbb{1}_D - \eta \nabla_x \mathfrak{F}(x^*, y_i)\nabla_x \mathfrak{F}(x^*, y_i)^t.$$

For points $x \in \mathbb{R}^D$ close to x^*, we have

$$\big[\mathcal{G}_{\eta,\xi_n} \circ \cdots \circ \mathcal{G}_{\eta,\xi_1}\big](x) = x^* + \big[\mathcal{G}'_{\eta,\xi_n}(x^*) \dots \mathcal{G}'_{\eta,\xi_1}(x^*)\big](x - x^*) + o(\|x - x^*\|).$$

Note that the matrices $\mathcal{G}'_{\eta,i}(x^*)$ also all respect the splitting $\mathbb{R}^D = \mathcal{T}(x^*) \oplus \mathcal{N}(x^*)$ with $\mathcal{G}'_{\eta,i}(x^*)|_{\mathcal{T}(x^*)} = \mathrm{Id}_{\mathcal{T}(x^*)}$. Analogously to gradient descent, we would like to call x^* linearly stable if

$$\lambda(x^*) = \lim_{n\to\infty} \frac{1}{n} \log\left(\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right) < 0.$$

At first glance, this definition raises two problems. First, it is not clear whether the limit exists. Secondly, the value of $\lambda(x^*)$ seems to depend on the random sequence $(\xi_1, \xi_2, \dots)$. However, Kingman's sub-additive ergodic theorem (Kingman, 1968, see also Steele, 1989)[6] states that for almost every realization $(\xi_1, \xi_2, \dots)$ we have

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right) = \inf_{n\in\mathbb{N}} \frac{1}{n}\mathbb{E}\left[\log\left(\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right)\right].$$

Hence, using the right-hand side as a definition for $\lambda(x^*)$, that is,

$$\lambda(x^*) := \inf_{n\in\mathbb{N}} \frac{1}{n}\mathbb{E}\left[\log\left(\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right)\right] \in [-\infty, \infty), \tag{2.12}$$

solves both issues. It should be noted that the sequence

$$n \mapsto \frac{1}{n}\mathbb{E}\left[\log\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right]$$

is monotonically decreasing and one may equivalently define $\lambda(x^*)$ by

$$\lambda(x^*) := \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\log\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right] \in [-\infty, \infty). \tag{2.13}$$

In fact, the following theorem due to Oseledec (1968) (see also, e.g., chapter 3.4 in Arnold, 1998) shows the existence of so-called *Oseledec's subspaces*, which are the equivalent notion to generalized eigenspaces for random-matrix products.

Theorem 2.1 (Multiplicative Ergodic Theorem) *There exists a natural number $1 \le k \le N$ and a tuple of extended real numbers*

$$-\infty \le \lambda_k < \dots < \lambda_2 < \lambda_1 = \lambda(x^*)$$

and a tuple of multiplicities $m_1, m_2, \dots, m_k$ with $m_1 + \dots + m_k = N$ such that the following holds: For almost every[7] *sequence $(\xi_1, \xi_2, \dots)$ there exists a tuple of subspaces*

$$\{0\} = V_{k+1}(\xi_1, \xi_2, \dots) \subsetneq V_k(\xi_1, \xi_2, \dots) \subsetneq \dots \subsetneq V_2(\xi_1, \xi_2, \dots) \subsetneq V_1(\xi_1, \xi_2, \dots) = \mathcal{N}(x^*)$$

with $\dim(V_i) = m_k + \dots + m_i$ such that for every $1 \le i \le k$

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)y\right\|\right] = \lambda_i, \ \forall\, y \in V_i(\xi_1, \xi_2, \dots) \setminus V_{i+1}(\xi_1, \xi_2, \dots).$$

In particular, for every $y \in \mathcal{N} \setminus V_2(\xi_1, \xi_2, \dots)$, we have

$$\lim_{n\to\infty} \frac{1}{n} \log\left[\left\|\mathcal{G}'_{\eta,\xi_n}(x^*)\dots\mathcal{G}'_{\eta,\xi_1}(x^*)y\right\|\right] = \lambda(x^*).$$

We will call x^* *linearly stable* (or *linearly unstable*) under stochastic gradient descent if $\lambda(x^*) < 0$ (or $\lambda(x^*) > 0$, respectively).[8]

6. The integrability condition for Kingman's ergodic theorem holds trivially here, as we only consider finitely many different matrices.
7. with respect to the uniform i.i.d. measure,
8. As mentioned in the introduction, this definition differs from the ones introduced by Wu et al. (2018) and Ma and Ying (2021). See Appendix A for a detailed comparison.

2.4 Main Result

The main contribution of our work is to rigorously show that the definitions for linear stability/instability, heuristically derived in Section 2.3, characterize the supports of $X_{\lim}^{\mathrm{GD/SGD}}$ and thus characterize the qualitative implicit bias of gradient descent/stochastic gradient descent. In our derivation, we made some assumptions which can be formalized as follows.

Hypothesis (H1) *For each $x \in \mathcal{M}$ the set of vectors*

$$\{\nabla_x \mathfrak{F}(x, y_i) : i \in [N]\} \subset R^D$$

is linearly independent. In particular, $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^D$.

Hypothesis (H2) *The initial distribution ν is equivalent to the Lebesgue measure, that is,*

$$\nu(B) = 0 \Leftrightarrow \mathrm{Leb}(B) = 0, \ \forall\, B \in \mathfrak{B}(\mathbb{R}^D).$$

Recall that, by an argument by Cooper (2021), Hypothesis (H1) is fulfilled for generic training data, while (H2) holds for the most common choice of initial distribution.

Furthermore, the heuristic argument presented in the previous section relied on the assumption that with probability one the optimization algorithms do not reach $\mathcal{M}$ in a finite number of steps. Hypothesis (H2) ensures that, with probability one, the algorithms are not initialized on $\mathcal{M}$. However, the setting presented so far is too general to exclude that the optimization algorithms reach $\mathcal{M}$ after any positive finite number of iterations with positive probability. Thus, we require an additional assumption. From now on, consider a fixed learning rate $\eta > 0$.

Hypothesis (H3) *The map $\mathcal{G}_\eta$ and the maps $\mathcal{G}_{\eta,1}, \ldots, \mathcal{G}_{\eta,N}$, defined in* (2.9) *and* (2.11)*, are non-singular, that is, the pre-image of every Lebesgue-null set is a Lebesgue-null set.*

In particular, a continuously differentiable map $\mathcal{G} : \mathbb{R}^D \to \mathbb{R}^D$ is non-singular if its Jacobian is invertible Lebesgue almost everywhere. Whether this is true for the maps $\mathcal{G}_\eta, \mathcal{G}_{\eta,1}, \ldots, \mathcal{G}_{\eta,N}$ depends on the network function $\mathfrak{F}$. Yet, it is reasonable to assume that this should be satisfied for the common neural network architectures, at least for almost every learning rate η.

Recall that the support of $X_{\lim}^{\mathrm{GD/SGD}}$ is defined as

$$\mathrm{supp}\left(X_{\lim}^{\mathrm{GD/SGD}}\right) := \left\{x \in \mathcal{M} : U \subseteq \mathcal{M} \text{ open nbhd. of } x \Rightarrow \mathbb{P}\left(X_{\lim}^{\mathrm{GD/SGD}} \in U\right) > 0\right\}.$$

For gradient descent, we will show the following main result.

Theorem A *Suppose (H1), (H2), and (H3) are satisfied. Let $x^* \in \mathcal{M}$.*

(i) If $\mu(x^*) < 0$, then $x^* \in \mathrm{supp}\left(X_{\lim}^{\mathrm{GD}}\right)$.

(ii) If $\mu(x^*) > 0$, then $x^* \notin \mathrm{supp}\left(X_{\lim}^{\mathrm{GD}}\right)$.

The analogous result for stochastic gradient descent requires some additional assumptions on the global minimum x^* in question.

Definition 2.2 *A global minimum $x^* \in \mathcal{M}$ is said to be regular if*

(i) for every $i \in [N]$, we have

$$\|\nabla_x \mathfrak{F}(x^*, y_i)\|^2 \notin \left\{\frac{1}{\eta}, \frac{2}{\eta}\right\}, \ \textit{and} \tag{2.14}$$

(ii) there exists no proper sub-set $\emptyset \subsetneq \mathcal{A} \subsetneq [N]$*, such that*

$$\nabla_x \mathfrak{F}(x^*, y_i) \cdot \nabla_x \mathfrak{F}(x^*, y_j) = 0, \ \forall\, i \in \mathcal{A},\, j \in [N] \setminus \mathcal{A}.$$

Note that almost every family of vectors $(\nabla_x \mathfrak{F}(x^*, y_i))_{i\in[N]}$ will satisfy these conditions. Thus, it is reasonable to assume that, for most network functions $\mathfrak{F}$, almost every global minimum $x^* \in \mathcal{M}$ is regular. Together, the conditions (i) and (ii) in Definition 2.2 imply that the semi-group generated by the operators $\mathcal{G}'_{\eta,1}(x^*), \dots, \mathcal{G}'_{\eta,N}(x^*)$ has nice algebraic properties, namely being contracting and strongly irreducible (see Definition 3.8 below). These properties are well established in the literature on random-matrix products (see, e.g., Bougerol and Lacroix, 1985) and allow us to use tools from that theory which are crucial for our proof. With this definition in hand, we get the following result for SGD.

Theorem B *Suppose (H1), (H2), and (H3) are satisfied. Let* $x^* \in \mathcal{M}$ *be regular.*

(i) If $\lambda(x^*) < 0$*, then* $x^* \in \operatorname{supp}\left(X_{\lim}^{\mathrm{SGD}}\right)$*.*

(ii) If $\lambda(x^*) > 0$*, then* $x^* \notin \operatorname{supp}\left(X_{\lim}^{\mathrm{SGD}}\right)$*.*

Remark 2.3 *Both Theorem A and Theorem B do not address the case* $\mu(x^*) = 0$*, respectively* $\lambda(x^*) = 0$*. Since the support of* $X_{\lim}^{GD/SGD}$ *is closed by definition and* $\mu : \mathcal{M} \to \mathbb{R}$ *is continuous, most points with* $\mu(x^*) = 0$ *should have global minima* $x' \in \mathcal{M}$ *with* $\mu(x') < 0$ *near them and therefore be in the support of* $X_{\lim}^{GD}$*. Making this argument rigorous would require a more precise knowledge of the network function* $\mathfrak{F}$ *and is beyond the scope of this paper. The continuity of* λ *is a more subtle issue. In general,* $\lambda : \mathcal{M} \to \mathbb{R}$ *is only upper semi-continuous*[9]*. However, a recent result of Avila et al. (2023) shows that* λ *is continuous at all points* x^**, where the matrices* $\mathcal{G}'_{\eta,1}(x^*), \dots, \mathcal{G}'_{\eta,N}(x^*)$ *are all invertible. This is, in particular, the case for all regular* $x^* \in \mathcal{M}$*.*

2.5 Relation to the Neural Tangent Kernel

We point out that, for any $x^* \in \mathcal{M}$, the values of $\mu(x^*)$ and $\lambda(x^*)$ can be deduced entirely from the *neural tangent kernel* $K_{x^*} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, first introduced by Jacot et al. (2018), which is defined by

$$K_{x^*}(y, y') = \nabla_x \mathfrak{F}(x^*, y)^t \nabla_x \mathfrak{F}(x^*, y'),$$

or more specifically, its *Gram matrix* $G_{x^*} \in \mathbb{R}^{N\times N}$ given by

$$[G_{x^*}]_{i,j} := K_{x^*}(y_i, y_j).$$

In more detail, let $S_{x^*} \in \mathbb{R}^{D\times N}$ denote the matrix whose i-th column is given by $\nabla_x \mathfrak{F}(x^*, y_i)$, that is,

$$S_{x^*} := \begin{pmatrix} | & & | \\ \nabla_x \mathfrak{F}(x^*, y_1) & \dots & \nabla_x \mathfrak{F}(x^*, y_N) \\ | & & | \end{pmatrix}. \tag{2.15}$$

Clearly, S_{x^*} maps $\mathbb{R}^N$ isomorphically onto $\mathcal{N}(x^*)$ and straightforward calculations, using linear independence of the gradients, show that

$$[\operatorname{Hess} \mathcal{L}(x^*)] S_{x^*} = \frac{1}{N} S_{x^*} G_{x^*} \qquad \text{and} \qquad [\operatorname{Hess} \mathcal{L}_i(x^*)] S_{x^*} = S_{x^*} G_{x^*,[i]},$$

9. This can readily be seen from the fact that it is defined as an infimum of continuous function.

where $G_{x^*,[i]}$ denotes the matrix G_{x^*} with every row but the i-th set to zero. It should be noted that S_{x^*} is not a square matrix and is thus not invertible as a matrix. However, as an isomorphism from $\mathbb{R}^N$ to $\mathcal{N}(x^*)$ it can be inverted, and we let $S_{x^*}^{-1} \in \mathbb{R}^{N\times D}$ denote the matrix associated with this inverse isomorphism. It satisfies $S_{x^*}^{-1} S_{x^*} = \mathbb{1}_N$ and $(S_{x^*} S_{x^*}^{-1})|_{\mathcal{N}(x^*)} = \mathbb{1}_D|_{\mathcal{N}(x^*)}$. Thus, the restrictions of the Jacobians of $\mathcal{G}_\eta$ and $\mathcal{G}_{\eta,i}$ to $\mathcal{N}(x^*)$ can be expressed as

$$\mathcal{G}'_\eta(x^*)|_{\mathcal{N}(x^*)} = S_{x^*} \left(\mathbb{1}_N - \frac{\eta}{N} G_{x^*}\right) {S_{x^*}}^{-1} \tag{2.16}$$

and

$$\mathcal{G}'_{\eta,i}(x^*)|_{\mathcal{N}(x^*)} = S_{x^*} \left(\mathbb{1}_N - \eta G_{x^*,[i]}\right) {S_{x^*}}^{-1}. \tag{2.17}$$

From this we get $\rho_{\text{Spec}}(\mathcal{G}'_\eta(x^*)|_{\mathcal{N}(x^*)}) = \rho_{\text{Spec}}(\mathbb{1}_N - \frac{\eta}{N} G_{x^*})$ and, using the symmetry of $(\mathbb{1}_N - \frac{\eta}{N} G_{x^*})$, we also get

$$\mu(x^*) = \log\left[\rho_{\text{Spec}}\left(\mathbb{1}_N - \frac{\eta}{N} G_{x^*}\right)\right] = \log\left\|\mathbb{1}_N - \frac{\eta}{N} G_{x^*}\right\|. \tag{2.18}$$

An analogous statement can be obtained for $\lambda(x^*)$. Indeed, we have the estimates

$$\begin{aligned}
&\log\left\| \left(\mathcal{G}'_{\eta,\xi_n}(x^*) \dots \mathcal{G}'_{\eta,\xi_1}(x^*)\right)|_{\mathcal{N}(x^*)}\right\| \\
&= \log\left\|S_{x^*}\left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right) {S_{x^*}}^{-1}\right\| \\
&\leq \log\|S_{x^*}\| + \log\left\|\left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right)\right\| + \log\left\|{S_{x^*}}^{-1}\right\|
\end{aligned}$$

and

$$\begin{aligned}
&\log\left\|\left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right)\right\| \\
&= \log\left\|{S_{x^*}}^{-1}\left(\mathcal{G}'_{\eta,\xi_n}(x^*) \dots \mathcal{G}'_{\eta,\xi_1}(x^*)\right) S_{x^*}\right\| \\
&\leq \log\left\|{S_{x^*}}^{-1}\right\| + \log\left\|\left(\mathcal{G}'_{\eta,\xi_n}(x^*) \dots \mathcal{G}'_{\eta,\xi_1}(x^*)\right)|_{\mathcal{N}(x^*)}\right\| + \log\|S_{x^*}\|.
\end{aligned}$$

Together, they imply

$$\begin{aligned}
\lambda(x^*) &= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\log\left\|\mathcal{G}'_{\eta,\xi_n}(x^*) \dots \mathcal{G}'_{\eta,\xi_1}(x^*)|_{\mathcal{N}(x^*)}\right\|\right] \\
&= \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\log\left\|\left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right)\right\|\right] \\
&= \inf_{n\in\mathbb{N}} \frac{1}{n} \mathbb{E}\left[\log\left\|\left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right)\right\|\right].
\end{aligned} \tag{2.19}$$

This observation is useful for two reasons. On the one hand, it connects the central quantities in the present work to an ongoing line of research of neural networks in the infinite-width limit (see Jacot et al., 2018). Additionally, it turns out that the expressions (2.18) and (2.19) are nicer to work with than the corresponding expressions in the previous section and will be used in the proofs of the main theorems. Finally, these expressions are well defined for any $x \in \mathbb{R}^D$ and not only for global minima $x^* \in \mathcal{M}$. This is especially useful for the potential empirical studies suggested in the following section.

2.6 Edge of Stability and Empirical Studies

While the present work is purely theoretical, the Lyapunov exponent $\lambda(x^*)$ of a global minimum x^* is also an interesting object for empirical studies. As mentioned in the introduction, a highly interesting study by Cohen et al. (2021) observed that training with gradient descent undergoes two distinct phases. At initialization, the Hessian of the loss landscape satisfies $\|\operatorname{Hess}\mathcal{L}(X_0^{\mathrm{GD}})\| < \frac{2}{\eta}$. During the early stages of training with gradient descent, the parameters X_n^{GD} move to progressively sharper regions of the loss landscape, that is, the operator norm of the Hessian of the loss function increases. During this phase, called *progressive sharpening*, the loss $\mathcal{L}(X_n^{\mathrm{GD}})$ decreases monotonically. Once $\|\operatorname{Hess}\mathcal{L}(X_n^{\mathrm{GD}})\|$ reaches the threshold $\frac{2}{\eta}$, progressive sharpening stops and training continues in the so-called *edge-of-stability* phase. During this phase, the Hessian of the loss function remains approximately constant close to the stability threshold $\frac{2}{\eta}$. Meanwhile, the loss is still decreasing in the long run, but no longer monotonically. Cohen et al. (2021) also studied whether stochastic gradient descent exhibits the same behavior. In Appendix G, they observe that for stochastic gradient descent progressive sharpening stops before the threshold $\|\operatorname{Hess}\mathcal{L}(X_n^{\mathrm{GD}})\| \approx \frac{2}{\eta}$, is reached, suggesting that SGD does not enter the edge of stability. In Appendix H, however, they observe that the expected change of the loss does behave similarly to the non-monotonous decrease observed for gradient descent at the edge of stability. Based on this, they conjecture (see Section 6 in Cohen et al., 2021) that SGD does enter an edge-of-stability regime, except that $\|\operatorname{Hess}\mathcal{L}(x)\| \leq \frac{2}{\eta}$ is no longer the correct notion of stability for gradient descent (see also the discussions in Wu et al., 2018 and Andreyev and Beneventano, 2025).

Instead of directly tracking the sharpness of the loss, $\|\operatorname{Hess}\mathcal{L}(x)\|$, one can equivalently study the evolution of $\mu(x)$ along the trajectory. It should be mentioned that in Section 2.3 we have only defined $\mu(x)$ for the parameters x that lie in the manifold of interpolation solutions $\mathcal{M}$. However, the equivalent definition (2.18) of μ given in Section 2.5, is well defined for any $x \in \mathbb{R}^D$. Thus, it is possible to track $\mu(X_n^{\mathrm{GD}})$ during training. Since μ is related to the Hessian of the loss by $\mu(x) = \log\left|1 - \eta\|\operatorname{Hess}\mathcal{L}(x)\|\right|$, it contains the same information as $\|\operatorname{Hess}\mathcal{L}(x)\|$. In the experiments of Cohen et al. (2021), we would observe that $\mu(X_0^{\mathrm{GD}})$ is negative and that $\mu(X_n^{\mathrm{GD}})$ progressively increases until it enters the edge of stability $\mu(X_n^{\mathrm{GD}}) = 0$ from which point on $\mu(X_n^{\mathrm{GD}})$ remains close to zero. Based on our main results, we suggest that from a dynamical point of view, tracking $\lambda(X_n^{\mathrm{SGD}})$ during training with stochastic gradient descent is the appropriate analogue to the study of Cohen et al. (2021). We leave it as an open problem to the community to determine the behavior of λ by the means of empirical experiments. An interesting challenge in conducting such a study lies in the numerical computation of the Lyapunov exponent for high-dimensional random-matrix products. Numerical schemes for the computation of Lyapunov exponents can be found, for example, in Eckmann and Ruelle (1985) and Sandri (1996).

2.7 Possible Extensions and Outlook

While our main results are general in the sense that they require only very weak assumptions on the network function $\mathfrak{F}$, we want to point out that we only consider scalar regression problems and training with gradient descent or stochastic gradient descent with mini-batch size 1. As mentioned above, we have made these restrictions to simplify the already extensive proofs as much as possible. The main results can be extended, with minor modifications, to a more general setting.

In practice, SGD is usually implemented with so-called mini-batches to allow for parallel computations. For SGD with mini-batches of size $1 \leq B \leq N$, the iterative step (2.8) is replaced by

$$X_{n+1}^{\mathrm{SGD}} := X_n^{\mathrm{SGD}} - \eta \nabla \mathcal{L}_{\Xi_{n+1}}(X_n^{\mathrm{SGD}}),$$

where $(\Xi_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of size B subsets $\Xi_n \subseteq [N]$ chosen uniformly from the $\binom{N}{B}$ possible subsets and

$$\mathcal{L}_\Xi(x) := \frac{1}{B}\sum_{i\in\Xi} \mathcal{L}_i(x).$$

For $B = 1$ this coincides with the SGD algorithm presented in Section 2.2 and for $B = N$ the iterative step is deterministic and coincides with GD. Stochastic gradient descent with mini-batch size $1 < B < N$ can thus be seen as an interpolation between the two algorithms. The derivations from sections 2.3 and 2.5 can be carried out analogously for mini-batch SGD and one can express the Lyapunov exponent of mini-batch SGD with learning rate η and mini-batch size B as

$$\lambda(x^*) = \inf_{n\in\mathbb{N}} \frac{1}{n}\mathbb{E}\left[\log\left\|\left(\mathbb{1}_N - \frac{\eta}{B}G_{x^*,[\Xi_n]}\right)\dots\left(\mathbb{1}_N - \frac{\eta}{B}G_{x^*,[\Xi_1]}\right)\right\|\right],$$

where $G_{x^*,[\Xi]}$ denotes the Gram matrix G_{x^*} of the neural tangent kernel with all rows whose index is not in Ξ set to 0. In order to derive a version of Theorem B it is necessary to adapt the notion of regular global minima (see Definition 2.2) to ensure that the analog of Lemma 3.9 in the proof below still remains valid. We leave this as a problem for future work. The rest of the proof of Theorem B can be extended to mini-batch SGD with only minor changes. Such an extension of Theorem B could help to better understand the impact of the learning rate and the mini-batch size on generalization properties (see, e.g., Hoffer et al., 2017; Goyal et al., 2017; Keskar et al., 2017 for numerical studies).

Furthermore, it would be interesting to extend our analysis to other optimizers such as SGD with momentum (see Rumelhart et al., 1986) or Adam (see Kingma and Ba, 2014). The derivations in Section 2.3 can be extended to these algorithms by linearizing the iteration steps around fixed points where all moment terms are zero. In the case of Adam, the parameter ϵ appearing in the numerator of the final update step is usually chosen extremely small (Kingma and Ba, 2014 suggest $\epsilon = 10^{-8}$). Thus, the linearization only approximates the dynamics of the actual algorithm in a vanishingly small neighborhood, and it is questionable whether it is still meaningful for the dynamics of the algorithm.

Our analysis can be extended to multidimensional regression problems with ground truth function $\hat{f} : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ and network function $\mathfrak{F} : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$. In fact, one can interpret such a multidimensional regression problem as a scalar regression problem with a ground truth function $\hat{f} : \mathbb{R}^d \times [\tilde{d}] \to \mathbb{R}$ and a network function $\mathfrak{F} : \mathbb{R}^D \times (\mathbb{R}^d \times [\tilde{d}]) \to \mathbb{R}$. This has the effect of splitting one training example $(y, z) \in \mathbb{R}^d \times \mathbb{R}^{\tilde{d}}$ into $\tilde{d}$ training examples $((y, 1), z_1), \dots, ((y, \tilde{d}), z_{\tilde{d}}) \in (\mathbb{R}^d \times [\tilde{d}]) \times \mathbb{R}$. An SGD step with the training example (y, z) now corresponds to an SGD step with the mini-batch $((y, 1), z_1), \dots, ((y, \tilde{d}), z_{\tilde{d}})$. Similarly to mini-batch SGD, the main challenge in extending Theorem B to multidimensional regression problems is to adapt the notion of a regular global minimum (see Definition 2.2) in such a way that Lemma 3.9 still remains valid for multidimensional regression.

Unfortunately, our results cannot be easily extended to classification problems. For a classification task with k classes, it is common to encode the training data as pairs $(y, z) \in \mathbb{R}^d \times \mathbb{R}^k$, where $z = e_i$ is the i-th unit vector, where $i \in [k]$ is the class y belongs to. Then one considers

a network function $\mathfrak{F} : \mathbb{R}^D \times \mathbb{R}^d \to \mathbb{R}^k$, which is then postcomposed with the softmax function $\mathrm{softmax}_\beta : \mathbb{R}^k \to [0,1]^k$ given by

$$\mathrm{softmax}_\beta(z_1, \dots z_k)_i := \frac{e^{\beta z_i}}{\sum_{j=1}^k e^{\beta z_j}},$$

where $\beta > 0$ is some positive real number commonly referred to as inverse temperature. For training, the most common loss function is the cross-entropy loss given by

$$\ell(\bar{z}, z) := -\sum_{i=1}^n z_i \log(\bar{z}_i).$$

To have $\ell(\bar{z}, z) = 0$, one necessarily needs $\bar{z} = z$. However, the outputs of the softmax function are in $(0,1)^k$ for finite input values. Thus, it is not possible to reach training error zero and $\mathcal{M} = \emptyset$.

Finally, it should be mentioned that our analysis assumes that the model is trained to convergence. In practice, it has been observed that stopping the algorithm early can improve generalization. While the global minima with $\lambda(x^*) > 0$ are asymptotically unstable, the finite time Lyapunov exponents

$$\lambda_{w,n}(x^*) := \frac{1}{n} \log \left\| \left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right) w \right\|$$

can, even for large n, still be negative with small, but positive probability. This effect is captured by a large deviation principle (see Arnold and Kliemann, 1987), the rate function of which depends on the moment Lyapunov exponents (see Definition A.2).

3. Proofs of the Main Results

3.1 Overview

The remainder of this paper will consist of the proofs of Theorem A and Theorem B. For the main arguments, it will be convenient to work in a local coordinate system in which $\mathcal{M}$ corresponds to a linear subspace. Such a coordinate system will be introduced in Section 3.2. Theorem A (i) then follows from an elementary argument presented in Section 3.3. In order to prove Theorem A (ii), in Section 3.4, we employ a center-stable manifold theorem. In Section 3.5, we introduce a random-dynamical-system framework to treat stochastic gradient descent. Theorem B (i) is then proved in Section 3.6. The proof is similar to the proof of Theorem A (i) presented in Section 3.3, but the possibility of non-uniform hyperbolicity adds an additional challenge. Finally, Theorem B (ii) is proved in sections 3.7-3.9. The proof is inspired by previous work on the instability of invariant subspaces for stochastic differential equations (see, e.g., Baxendale and Stroock, 1988; Baxendale, 1991). While these works rely on Hörmander conditions (see Hörmander, 1967), we use a criterion by Le Page (1982) to establish a spectral gap for the projective semigroup. The assumptions for the criterion of Le Page are checked in Section 3.7. In Section 3.8, we construct a local Lyapunov function similar to the recent works of Bedrossian et al. (2022) and Blumenthal et al. (2023). The proof of Theorem B is then completed in Section 3.9.

Throughout the entire Section 3 we assume (H1), (H2) and (H3) as standing assumptions.

3.2 Local Coordinates

In the following, for some fixed $x^* \in \mathcal{M}$, we introduce a local coordinate system for a neighborhood of x^*, in which the generally nonlinear manifold $\mathcal{M}$ becomes a linear subspace aligned with the coordinate axes. Using these coordinates will be helpful in all further proofs.

Lemma 3.1 *There exist an open neighborhood $x^* \in \hat{U} \subset \mathbb{R}^D$ and an open neighborhood $(0,0) \in \hat{V} \subseteq \mathbb{R}^{D-N} \times \mathbb{R}^N$, as well as a smooth diffeomorphism $\chi : \hat{V} \to \hat{U}$, such that*

(i) $\chi(0,0) = x^*$,

(ii) $\chi(\hat{V} \cup (\mathbb{R}^{D-N} \times \{0\})) = \hat{U} \cap \mathcal{M}$,

(iii) the Jacobian at the origin is given by

$$\chi'(0,0) = \begin{pmatrix} A & S_{x^*} \end{pmatrix}, \tag{3.1}$$

where $A \in \mathbb{R}^{D\times(D-N)}$ is some matrix which induces an isomorphism from $\mathbb{R}^{D-N}$ onto $\mathcal{T}(x^*)$ and $S_{x^*} \in \mathbb{R}^{D\times N}$ is the matrix defined in (2.15)*,*

(iv) and χ is bi-Lipschitz with Lipschitz constant L_χ, that is, both χ and χ^{-1} are Lipschitz continuous with said Lipschitz constant.

Proof Let $A \in \mathbb{R}^{D\times(D-N)}$ be some matrix that induces an isomorphism from $\mathbb{R}^{D-N}$ to $\mathcal{T}(x^*)$. Since $\mathcal{M}$ is an embedded smooth manifold, we can find neighborhoods $0 \in U_\mathcal{T} \subseteq \mathcal{T}(x^*)$ and $x^* \in \tilde{U} \subseteq \mathbb{R}^D$ and a smooth map $\zeta : U_\mathcal{T} \to \mathcal{N}(x^*)$ with $\zeta(0) = 0$, $\zeta'(0) = 0$ and

$$\{x^* + x + \zeta(x) : x \in U_\mathcal{T}\} = \tilde{U} \cap \mathcal{M}.$$

This allows us to define a smooth map $\chi : A^{-1}U_\mathcal{T} \times \mathbb{R}^N \to \mathbb{R}^D$ by

$$\chi(v,w) = x^* + Av + S_{x^*}w + \zeta(Av).$$

It can be easily verified that χ is injective and

$$\chi(A^{-1}U_\mathcal{T} \times \{0\}) = \tilde{U} \cap \mathcal{M}.$$

Also, the Jacobian is given by

$$\chi'(v,w) = \begin{pmatrix} A + \zeta'(Av)A & S_{(x_0)} \end{pmatrix},$$

so in particular χ is a local diffeomorphism and (3.1) holds. Set $\hat{V} := \chi^{-1}(\tilde{U})$ and $\hat{U} := \chi(\hat{V})$. If we now let χ be its restriction $\chi : \hat{V} \to \hat{U}$, it is a smooth diffeomorphism, satisfying i) to iii). If χ is not bi-Lipschitz, we can reassign $\hat{V}$ to a smaller neighborhood, which is precompact in the original $\hat{V}$. After reassigning $\hat{U}$ and restricting χ accordingly, χ will be bi-Lipschitz. ■

The learning dynamics of (stochastic) gradient decent can, at least locally, be lifted in the new coordinates via the function χ. We define the functions $\hat{\mathcal{G}}_\eta, \hat{\mathcal{G}}_{\eta,1}, \dots \hat{\mathcal{G}}_{\eta,N} : V^* \to \hat{V}$ by

$$\hat{\mathcal{G}}_\eta(v,w) = \chi^{-1}(\mathcal{G}_\eta(\chi(v,w))) \qquad \text{and} \qquad \hat{\mathcal{G}}_{\eta,i}(v,w) = \chi^{-1}(\mathcal{G}_{\eta,i}(\chi(v,w))),$$

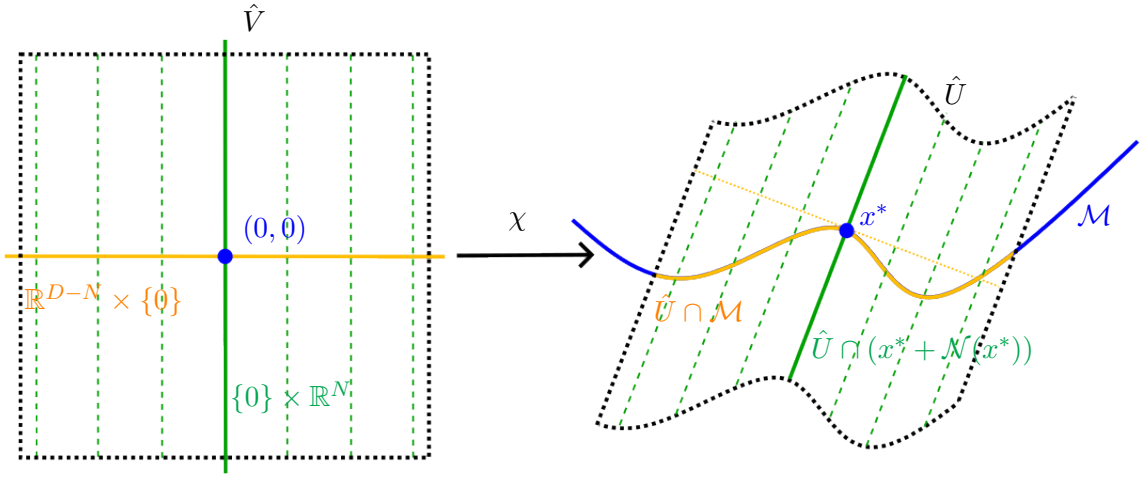

Figure 1: Schematic representation of the construction of χ. Objects with the same color are mapped onto each other.

where $V^* \subseteq \hat{V}$ is an open set given by

$$V^* := \chi^{-1}\left(\mathcal{G}_\eta^{-1}(\hat{U}) \cap \bigcap_{i\in[N]} \mathcal{G}_{\eta,i}^{-1}(\hat{U})\right). \tag{3.2}$$

Let $\tau : V^* \to \mathbb{N} \cup \{\infty\}$ be the maximal number n for which $\hat{\mathcal{G}}_\eta^{n+1}(v,w)$ is well defined, that is,

$$\tau(v,w) := \inf\{n \in \mathbb{N} : \hat{\mathcal{G}}_\eta^n(v,w) \notin V^*\}.^{10} \tag{3.3}$$

One can easily check that for $(v,w) \in V^*$ and $1 \le n \le \tau(v,w)$, we have

$$\hat{\mathcal{G}}_\eta^n(v,w) = \chi^{-1}(\mathcal{G}_\eta^n(\chi(v,w))). \tag{3.4}$$

The corresponding statement for SGD will be given in Section 3.5 once the appropriate notation has been introduced (see (3.20)).

Since all points in $\mathcal{M}$ are fixed points for $\hat{\mathcal{G}}_\eta, \hat{\mathcal{G}}_{\eta,1}, \dots \hat{\mathcal{G}}_{\eta,N}$ and by Lemma 3.1 (ii), we have $\hat{V} \cup (\mathbb{R}^{D-N} \times \{0\}) \subseteq V^*$ and

$$\hat{\mathcal{G}}_\eta(v,0) = (v,0) \qquad \text{and} \qquad \hat{\mathcal{G}}_{\eta,i}(v,0) = (v,0).$$

Furthermore, using the chain rule, equations (2.16) and (2.17) as well as Lemma 3.1 (iii), we can compute

$$\hat{\mathcal{G}}_\eta'(0,0) = \mathbb{1}_D - \frac{\eta}{N}\begin{pmatrix} 0 & 0 \\ 0 & G_{x^*} \end{pmatrix} \qquad \text{and} \qquad \hat{\mathcal{G}}_{\eta,i}'(0,0) = \mathbb{1}_D - \eta\begin{pmatrix} 0 & 0 \\ 0 & G_{x^*,[i]} \end{pmatrix}, \tag{3.5}$$

where the right-hand sides are both block matrices with dimensions $((D-N)+N) \times ((D-N)+N)$[11].

10. Of course, $\inf \emptyset := \infty$.
11. that is, the rows and the columns have been partitioned into two blocks of size $(D-N)$ and size N.

Lemma 3.2 *For every $\delta > 0$, there exists an open neighborhood $V_\delta \subseteq V^*$, such that for each $(v, w) \in V_\delta$ and each $i \in [N]$, we have*

$$\left\| \hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0) \right] \begin{pmatrix} v \\ w \end{pmatrix} \right\| \leq \delta \|w\|, \tag{3.6}$$

as well as

$$\left\| \hat{\mathcal{G}}_{\eta,i}(v, w) - \left[\hat{\mathcal{G}}'_{\eta,i}(0, 0) \right] \begin{pmatrix} v \\ w \end{pmatrix} \right\| \leq \delta \|w\|. \tag{3.7}$$

Note that this lemma would hold trivially if the right-hand sides of (3.6) and (3.7) were replaced by the term $\delta\|(v, w)\|$.

Proof It is sufficient to consider each of the functions $\hat{\mathcal{G}}_\eta, \hat{\mathcal{G}}_{\eta,1}, \ldots, \hat{\mathcal{G}}_{\eta,N}$ separately, find a neighborhood V_δ in which the corresponding inequality (3.6), respectively (3.7), holds and conclude the proof by choosing the intersection over all these V_δ. In the following, we will only consider $\hat{\mathcal{G}}_\eta$ and will show how to find an V_δ such that (3.6) holds, as the proofs for $\hat{\mathcal{G}}_{\eta,1}, \ldots, \hat{\mathcal{G}}_{\eta,N}$ work analogously.

Let $0 \in V_\delta \subseteq V^*$ be an open, convex neighborhood, such that

$$\left\| \partial_w \hat{\mathcal{G}}_\eta(v, w) - \partial_w \hat{\mathcal{G}}_\eta(0, 0) \right\| \leq \delta, \ \forall\, (v, w) \in V_\delta.$$

Here $\partial_w \hat{\mathcal{G}}_\eta(v, w) \in \mathbb{R}^{D \times N}$ denotes the partial Jacobian with respect to the latter N components. We have

$$\begin{aligned} \hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0) \right] \begin{pmatrix} v \\ w \end{pmatrix} &= \hat{\mathcal{G}}_\eta(v, 0) + \int_0^1 \left[\partial_w \hat{\mathcal{G}}_\eta(v, tw) \right] w \, \mathrm{d}t - (v, 0) - \left[\partial_w \hat{\mathcal{G}}_\eta(0, 0) \right] w \\ &= \int_0^1 \left[\partial_w \hat{\mathcal{G}}_\eta(v, tw) - \partial_w \hat{\mathcal{G}}_\eta(0, 0) \right] w \, \mathrm{d}t \end{aligned}$$

and for $(v, w) \in V_\delta$ in particular

$$\left\| \hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0) \right] \begin{pmatrix} v \\ w \end{pmatrix} \right\| \leq \int_0^1 \left\| \partial_w \hat{\mathcal{G}}_\eta(v, tw) - \partial_w \hat{\mathcal{G}}_\eta(0, 0) \right\| \|w\| \, \mathrm{d}t \leq \delta \|w\|.$$

This finishes the proof. ■

3.3 Gradient Descent - the Stable Case

In this section we will prove Theorem A (i).

Theorem A (i) *Let $x^* \in \mathcal{M}$ with $\mu(x^*) < 0$. Then $x^* \in \operatorname{supp}\left(X_{\lim}^{\mathrm{GD}}\right)$.*

In the following, we will let $\Pi_v \in \mathbb{R}^{(D-N) \times D}$ and $\Pi_w \in \mathbb{R}^{(D-N) \times N}$ be the matrices that project a vector $(v, w)^t \in \mathbb{R}^D$ onto its v-, respectively w-component, that is, in block-matrix form we set

$$\Pi_v := \begin{pmatrix} \mathbb{1}_{D-N} & 0 \end{pmatrix} \qquad \text{and} \qquad \Pi_w := \begin{pmatrix} 0 & \mathbb{1}_N \end{pmatrix}.$$

Proof [of Theorem A (i)] Suppose $\mu(x^*) < 0$ and let $U \subseteq \mathcal{M}$ be some neighborhood of x^*. Our goal is to prove $\mathbb{P}(X_{\lim}^{\mathrm{GD}} \in U) > 0$. Choose $\delta > 0$ such that $e^{\mu(x^*)} + \delta =: \gamma < 1$, let V_δ be the corresponding neighborhood given by Lemma 3.2. Let $R_\delta > 0$ be some radius, such that

$$\mathcal{B}_{\mathbb{R}^{D-N}}(R_\delta) \times \mathcal{B}_{\mathbb{R}^N}(R_\delta) \subseteq V_\delta.$$

We assume without loss of generality that U has the form

$$U = \chi\left(\mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}\right) \tag{3.8}$$

for some $0 < R < R_\delta$. Similarly to the definition of τ, we define a map $\tau_\delta : V_\delta \to \mathbb{N} \cup \{\infty\}$ by

$$\tau_\delta(v, w) := \inf\left\{n \in \mathbb{N} : \hat{\mathcal{G}}_\eta^n(v, w) \notin V_\delta\right\}. \tag{3.9}$$

Note that $\tau_\delta(v, w) \leq \tau(v, w)$.

As a consequence of Lemma 3.2 and the expression (2.18) for $\mu(x^*)$, for all $(v, w) \in V_\delta$, we have

$$\begin{aligned}
\left\|\Pi_w \hat{\mathcal{G}}_\eta(v, w)\right\| &\leq \left\|\left[\Pi_w \hat{\mathcal{G}}'_\eta(0, 0)\right]\begin{pmatrix} v \\ w \end{pmatrix}\right\| + \left\|\Pi_w\left(\hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0)\right]\begin{pmatrix} v \\ w \end{pmatrix}\right)\right\| \\
&\leq \left\|\left(\mathbb{1}_N - \frac{\eta}{N} G_{x^*}\right) w\right\| + \left\|\hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0)\right]\begin{pmatrix} v \\ w \end{pmatrix}\right\| \\
&\leq e^{\mu(x^*)}\|w\| + \delta\|w\| = \gamma\|w\|.
\end{aligned} \tag{3.10}$$

Using this bound inductively, for all $(v, w) \in V_\delta$ and $1 \leq n \leq \tau_\delta(v, w)$, we get

$$\left\|\Pi_w \hat{\mathcal{G}}_\eta^n(v, w)\right\| \leq \gamma^n \|w\|. \tag{3.11}$$

Recalling equation (3.5) and using Lemma 3.2 implies

$$\left\|\Pi_v \hat{\mathcal{G}}_\eta(v, w) - v\right\| = \left\|\Pi_v\left(\hat{\mathcal{G}}_\eta(v, w) - \left[\hat{\mathcal{G}}'_\eta(0, 0)\right]\begin{pmatrix} v \\ w \end{pmatrix}\right)\right\| \leq \delta\|w\|,$$

for all $(v, w) \in V_\delta$ and thus also for all $1 \leq n < \tau_\delta(v, w)$,

$$\left\|\Pi_v \hat{\mathcal{G}}_\eta^{n+1}(v, w) - \Pi_v \hat{\mathcal{G}}_\eta^n(v, w)\right\| \leq \delta\left\|\Pi_w \hat{\mathcal{G}}_\eta^n(v, w)\right\| \leq \delta\gamma^n\|w\|. \tag{3.12}$$

With this, we can bound

$$\begin{aligned}
\left\|\Pi_v \hat{\mathcal{G}}_\eta^n(v, w)\right\| &\leq \|v\| + \sum_{m=1}^{n}\left\|\Pi_v \hat{\mathcal{G}}_\eta^m(v, w) - \Pi_v \hat{\mathcal{G}}_\eta^{m-1}(v, w)\right\| \\
&\leq \|v\| + \sum_{m=1}^{n} \delta\gamma^{m-1}\|w\| \leq \|v\| + \frac{\delta}{1-\gamma}\|w\|,
\end{aligned} \tag{3.13}$$

for all $(v, w) \in V_\delta$ and $1 \leq n + 1 < \tau_\delta(v, w)$. Now set

$$R_v := \frac{R}{2} \qquad \text{and} \qquad R_w = \min\left(\frac{(1-\gamma)R}{2\delta}, R_\delta\right).$$

Suppose for some $(v, w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w)$, we have $\tau_\delta(v, w) < \infty$. Then by definition

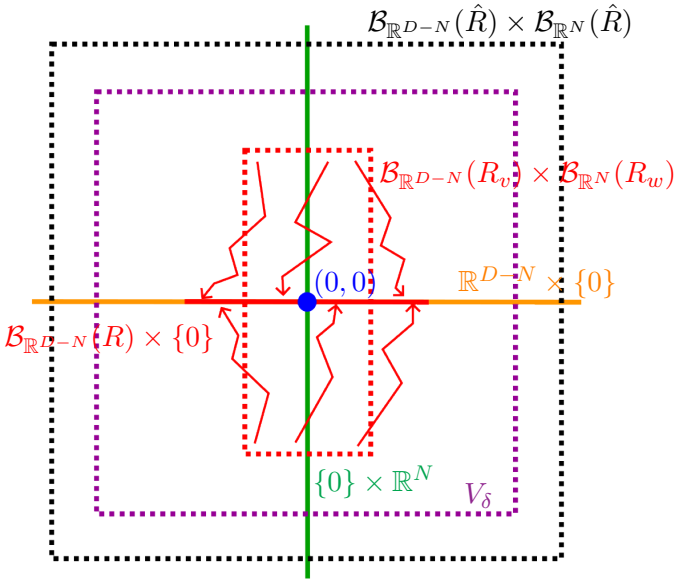

Figure 2: Schematic representation of the proof of Theorem A (i). The red arrows show possible sample trajectories.

of τ_δ, we have $\hat{\mathcal{G}}_\eta^{\tau_\delta(v,w)}(v,w) \notin V_\delta$, so by (3.8) in particular

$$\left\|\Pi_v \hat{\mathcal{G}}_\eta^{\tau_\delta(v,w)}(v,w)\right\| \geq R_\delta \qquad \text{or} \qquad \left\|\Pi_w \hat{\mathcal{G}}_\eta^{\tau_\delta(v,w)}(v,w)\right\| \geq R_\delta. \tag{3.14}$$

However, (3.13) implies

$$\left\|\Pi_v \hat{\mathcal{G}}_\eta^{\tau_\delta(v,w)}(v,w)\right\| \leq \|v\| + \frac{\delta}{1-\gamma}\|w\| < R_v + \frac{\delta}{1-\gamma} R_w \leq R \leq R_\delta$$

and (3.11) implies

$$\left\|\Pi_w \hat{\mathcal{G}}_\eta^{\tau_\delta(v,w)}(v,w)\right\| \leq \gamma^{\tau_\delta(v,w)}\|w\| \leq \|w\| < R_w \leq R_\delta,$$

contradicting (3.14). Thus $\tau_\delta(v,w) = \infty$, for all $(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w)$. Now (3.12) implies that $\left(\Pi_v \hat{\mathcal{G}}_\eta^n(v,w)\right)$ is a Cauchy sequence and (3.13) shows that

$$\left\|\lim_{n\to\infty} \Pi_v \hat{\mathcal{G}}_\eta^n(v,w)\right\| \leq \|v\| + \frac{\delta}{1-\gamma}\|w\| < R_v + \frac{\delta}{1-\gamma} R_w \leq R.$$

Furthermore, (3.11) shows $\Pi_w \hat{\mathcal{G}}_\eta^n(v,w) \to 0$. Thus, for each $(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w)$ the sequence $\left(\hat{\mathcal{G}}_\eta^n(v,w)\right)$ converges with

$$\lim_{n\to\infty} \hat{\mathcal{G}}_\eta^n(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}.$$

Let $\tilde{U} = \chi\left(\mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w)\right)$. Suppose $X_0^{\mathrm{GD}} \in \tilde{U}$. By (3.4) and continuity of χ we have

$$\begin{aligned} X_{\lim}^{\mathrm{GD}} &= \lim_{n\to\infty} \mathcal{G}_\eta^n(X_0^{\mathrm{GD}}) = \lim_{n\to\infty} \chi\left(\hat{\mathcal{G}}_\eta^n\left(\chi^{-1}(X_0^{\mathrm{GD}})\right)\right) \\ &= \chi\left(\lim_{n\to\infty} \hat{\mathcal{G}}_\eta^n\left(\chi^{-1}(X_0^{\mathrm{GD}})\right)\right) \in \chi(\mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}) = U, \end{aligned}$$

so $X_0^{\mathrm{GD}} \in \tilde{U}$ implies $X_{\lim}^{\mathrm{GD}} \in U$. Since $\tilde{U}$ is an open set, we have

$$\mathbb{P}(X_{\lim}^{\mathrm{GD}} \in U) \geq \mathbb{P}\left(X_0^{\mathrm{GD}} \in \tilde{U}\right) > 0,$$

completing the proof. ■

3.4 Gradient Descent - the Unstable Case

In this section, we will prove Theorem A (ii).

Theorem A (ii) *Let $x^* \in \mathcal{M}$ with $\mu(x^*) > 0$. Then $x^* \notin \operatorname{supp}\left(X_{\lim}^{\mathrm{GD}}\right)$.*

In the case $\mu(x^*) < 0$ considered in the previous section, all the eigenvalues of $\mathbb{1}_N - \eta G_{x^*}$ lie strictly within the unit circle. As a consequence, we were able to construct an open neighborhood of x^*, such that for every initial condition in that neighborhood, GD converges to some $X_{\lim}^{\mathrm{GD}} \in \mathcal{M}$ near x^*. In the case $\mu(x^*) > 0$ we consider in this section, we only know that some eigenvalue of $\mathbb{1}_N - \eta G_{x^*}$ lies strictly outside the unit circle while there could still be eigenvalues on or strictly inside the unit circle. As a consequence, one should not expect to be able to construct a neighborhood $U \subseteq \mathbb{R}^D$ of x^*, such that GD does not converge to some $X_{\lim}^{\mathrm{GD}} \in \mathcal{M}$ near x^* for any initial condition in U. Instead, we will construct a neighborhood $U \subseteq \mathbb{R}^D$ of x^*, such that GD does not converge to some $X_{\lim}^{\mathrm{GD}} \in \mathcal{M}$ near x^* for Lebesgue-almost any initial condition in U. In fact, we will show that the set of initial conditions in U for which GD does converge to some $X_{\lim}^{\mathrm{GD}} \in \mathcal{M}$ near x^* is contained in a lower dimensional manifold called the *center-stable manifold*[12]. While for the preceding and all subsequent sections it is convenient to consider vectors in $\mathbb{R}^D$ as consisting of a $D-N$-dimensional tangential part v and an N-dimensional transversal part w, here it will be more convenient to consider vectors in $\mathbb{R}^D$ as consisting of a center-stable part v_- corresponding to the eigenvalues strictly inside or on the unit circle and an unstable part v_+ corresponding to the eigenvalues strictly outside the unit circle.

For the remainder of this section, consider a fixed $x^* \in \mathcal{M}$ with $\mu(x^*) > 0$. Let $\bar{\mu}_1, \ldots, \bar{\mu}_D$ be the eigenvalues of be the eigenvalues of $\hat{\mathcal{G}}'(0,0)$ appearing according to their multiplicity[13] and ordered by their absolute values, that is,

$$|\bar{\mu}_1| \geq \cdots \geq |\bar{\mu}_D|.$$

Note that the eigenvalue 1 must appear at least with multiplicity $D-N$, corresponding to the tangential part and $\log|\bar{\mu}_1| = \mu(x^*) > 0$, which implies $|\bar{\mu}_1| > 1$. Let $D_+ > 1$ be the number of eigenvalues with absolute value greater than 1 and let $D_- = D - D_+$ be the number of eigenvalues with absolute value less or equal to 1, counted with multiplicity in both cases. Furthermore, let $\bar{S} \in \mathbb{R}^{D\times D}$ be an invertible matrix, such that

$$\hat{\mathcal{G}}'_\eta(0,0) = \bar{S}\begin{pmatrix} \bar{\mu}_1 & & \\ & \ddots & \\ & & \bar{\mu}_D \end{pmatrix}\bar{S}^{-1} = \bar{S}\begin{pmatrix} A_+ & 0 \\ 0 & A_- \end{pmatrix}\bar{S}^{-1},$$

12. While the set, we call center-stable manifold is indeed a C^1-manifold (see Theorem 6.2.8 in Katok and Hasselblatt, 1995), we will only show that it is the graph of a Lipschitz continuous function.
13. Since $\hat{\mathcal{G}}'_\eta(0,0)$ is symmetric, the algebraic and geometric multiplicities coincide.

where $A_+ \in \mathbb{R}^{D_+ \times D_+}$ and $A_- \in \mathbb{R}^{D_- \times D_-}$ are given by

$$A_+ := \begin{pmatrix} \bar{\mu}_1 & & \\ & \ddots & \\ & & \bar{\mu}_{D_+} \end{pmatrix} \qquad \text{and} \qquad A_- := \begin{pmatrix} \bar{\mu}_{D_++1} & & \\ & \ddots & \\ & & \bar{\mu}_{D} \end{pmatrix}.$$

Note that $\|A_-\| = 1$ and that A_+ is invertible with $\|A_+^{-1}\|^{-1} = |\bar{\mu}_{D_+}| > 1$, where $\bar{\mu}_{D_+}$ is the eigenvalue with the smallest absolute value that is still larger than 1. Finally, let $\bar{\mathcal{G}}_\eta : \mathbb{R}^{D_+} \times \mathbb{R}^{D_-} \to \mathbb{R}^{D_+} \times \mathbb{R}^{D_-}$ given by

$$\bar{\mathcal{G}}_\eta(v_+, v_-) = \bar{S}\hat{\mathcal{G}}_\eta\left(\bar{S}^{-1}\begin{pmatrix} v_+ \\ v_- \end{pmatrix}\right).$$

The following is a version of the center-stable manifold theorem.

Theorem 3.3 (Center-Stable Manifold) *There exist open neighborhoods $0 \in V_- \subseteq \mathbb{R}^{D_-}$, $0 \in V_+ \subseteq \mathbb{R}^{D_+}$ and a map $\beta^* : V_- \to V_+$ such that for any $(v_+, v_-) \in (V_+ \times V_-) \setminus \text{graph}(\beta^*)$, there exists some $n \in \mathbb{N}$, such that $\bar{\mathcal{G}}_\eta^n(v_+, v_-) \notin V_+ \times V_-$.*

If $\bar{\mathcal{G}}_\eta$ is a local diffeomorphism at 0, this result is standard (see, e.g., chapter 6.2 in Katok and Hasselblatt, 1995). However, in our setting, it is possible that $\bar{\mu}_i = 0$ for some $i \in [D]$. Still, Theorem 3.3 follows from the arguments in Katok and Hasselblatt (1995) with mild modifications. Alternatively, stable/unstable/center manifold theorems in non-locally diffeomorphic setting have also been obtained by reducing the problem to the locally diffeomorphic case by an abstract method called inverse limits (see, e.g., Ruelle and Shub, 1980 or Qian et al., 2009).

In order to prove Theorem A (ii), we require the following lemma. Since the equivalent statement will also be useful to prove the instability of SGD, we formulate it to cover both GD and SGD.

Lemma 3.4 *Let $A \subseteq \mathbb{R}^D$ be a Lebesgue-null set. Then*

$$\mathbb{P}\left(\exists n \in \mathbb{N}_0, \ \textit{s.t.} \ X_n^{\text{GD/SGD}} \in A\right) = 0.$$

Proof By Hypothesis (H2), we have $\mathbb{P}(X_0^{\text{GD/SGD}} \in A) = \nu(A) = 0$. By Hypothesis (H3), we also have $\mathbb{P}(X_n^{\text{GD/SGD}} \in A) = 0$, for every $n \geq 1$ and thus

$$\mathbb{P}\left(\exists n \in \mathbb{N}_0, \ \text{s.t.} \ X_n^{\text{GD/SGD}} \in A\right) = 0.$$

■

Proof [of Theorem A (ii)] Suppose $\mu(x^*) > 0$. Let $x^* \in U \subseteq \mathbb{R}^D$ be the open neighborhood given by

$$U := \chi(\bar{S}(V_+ \times V_-)),$$

where V_+ and V_- are the neighborhoods given by Theorem 3.3. Suppose $\omega \in \Omega$ is such that $X_{\lim}^{\text{GD}}(\omega) \in U \cap \mathcal{M}$. Then there must exist an $m \in \mathbb{N}$, such that $\mathcal{G}_\eta^n(X_m^{\text{GD}}(\omega)) = X_{m+n}^{\text{GD}(\omega)} \in U$ for all $n \in \mathbb{N}_0$. Equivalently, we must have $\bar{\mathcal{G}}_\eta^n(\bar{S}^{-1}\chi^{-1}(X_m^{\text{GD}}(\omega))) \in V_+ \times V_-$, for all $n \in \mathbb{N}_0$. By Theorem 3.3 this means that we can only have $X_{\lim}^{\text{GD}}(\omega) \in U \cap \mathcal{M}$, if there exists an $m \in \mathbb{N}$

such that $X_m^{\mathrm{GD}}(\omega) \in \chi(\bar{S}\,\mathrm{graph}(\beta^*))$. However, since χ is a diffeomorphism and $\bar{S}$ an invertible matrix, the set $\chi(\bar{S}\,\mathrm{graph}(\beta^*))$ has Lebesgue measure zero. By Lemma 3.4, this implies

$$\mathbb{P}(X_{\mathrm{lim}}^{\mathrm{GD}}(\omega) \in U \cap \mathcal{M}) = 0,$$

completing the proof. ■

3.5 Random Dynamical System Framework for SGD

While in sections 3.3 and 3.4 we studied the dynamics of gradient descent, the rest of the paper will be concerned with establishing analogous results for stochastic gradient descent. For this, we will formulate SGD as a *random dynamical system* (see Arnold, 1998 for a general introduction to the theory). Formally, for some appropriate probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we will introduce a map

$$\begin{aligned} \varphi : \mathbb{N}_0 \times \Omega \times \mathbb{R}^D &\to \mathbb{R}^D, \\ (n, \omega, x) &\mapsto \varphi_\omega^{(n)}(x), \end{aligned}$$

where evaluation of the map $\varphi_\omega^{(n)}$ should correspond to applying n iterations of SGD with ω serving as the seed for the random choices of training examples. We expect the map φ to satisfy

$$\varphi_\omega^{(n+m)}(x) = \varphi_{\theta^n\omega}^{(m)}\left(\varphi_\omega^{(n)}(x)\right), \tag{3.15}$$

where $\theta^n\omega$ is the seed for the same training examples, but shifted by n steps. Put more clearly, if ω generates the sequence of training examples $(\xi_1, \xi_2, \dots) \in [N]^{\mathbb{N}}$, then $\theta^n\omega$ should generate the sequence $(\xi_{n+1}, \xi_{n+2}, \dots) \in [N]^{\mathbb{N}}$. Since the training examples are chosen independently with identical distributions, the sequences $(\xi_1, \xi_2, \dots)$ and $(\xi_{n+1}, \xi_{n+2}, \dots)$ are equal in distribution. In other words, the map $\theta : \Omega \to \Omega$ leaves the probability measure $\mathbb{P}$ invariant, that is,

$$\mathbb{P}(E) = \mathbb{P}\left(\theta^{-1}(E)\right), \ \forall\, E \in \mathcal{F}. \tag{3.16}$$

A map $\theta : \Omega \to \Omega$ satisfying (3.16) is called *metric dynamical system*. Equation (3.15) is known as the *cocycle property* and a map φ satisfying it is called a *cocycle* over θ. A pair (θ, φ), consisting of a metric dynamical system and a cocycle over it, is called a random dynamical system (see Defintion 1.1.1 in Arnold, 1998).

Since Ω should encode the randomness both at initialization and during training, we let

$$\Omega := \mathbb{R}^D \times [N]^{\mathbb{N}} \qquad \text{and} \qquad \mathcal{F} := \mathfrak{B}(\mathbb{R}^D) \otimes \mathfrak{P}([N])^{\otimes\mathbb{N}},$$

where $\mathfrak{P}([N])$ denotes the power set of $[N]$. If β denotes the uniform measure on $[N]$ we can define the probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ by

$$\mathbb{P} := \nu \otimes \beta^{\otimes\mathbb{N}}.$$

Thus, the elements $\omega \in \Omega$ have the form $\omega = (\omega_{\mathrm{init}}, \omega_1, \omega_2, \dots)$ and the random variables $X_0^{\mathrm{SGD}}, \xi_1, \xi_2, \dots$ used in SGD (see Section 2.2) can be formally defined by

$$X_0^{\mathrm{SGD}}(\omega) := \omega_{\mathrm{init}} \qquad \text{and} \qquad \xi_n(\omega) := \omega_n.$$

The shift operator $\theta : \Omega \to \Omega$ can now be defined by

$$\theta : \omega = (\omega_{\text{init}}, \omega_1, \omega_2, \dots) \mapsto \theta\omega := (\omega_{\text{init}}, \omega_2, \omega_3, \dots).$$

Clearly θ satisfies (3.16) and thus defines a metric dynamical system on Ω. Furthermore θ is ergodic with respect to the sub-sigma algebra $\hat{\mathcal{F}} := \sigma(\xi_1, \xi_2, \dots)$, that is,

$$\theta^{-1}(E) = E \Rightarrow \mathbb{P}(E) \in \{0, 1\}, \ \forall\, E \in \hat{\mathcal{F}}.$$

Recall (see (2.8) and (2.11)) that the random variables $X_1^{\text{SGD}}, X_2^{\text{SGD}}, \dots$ can be defined recursively by

$$X_{n+1}^{\text{SGD}}(\omega) := \mathcal{G}_{\eta,\xi_{n+1}(\omega)}(X_n^{\text{SGD}}(\omega)), \ \forall\, n \geq 0.$$

The map $\varphi : \mathbb{N}_0 \times \Omega \times \mathbb{R}^D \to \mathbb{R}^D$ will be defined by

$$\varphi_\omega^{(0)}(x) := x \qquad \text{and} \qquad \varphi_\omega^{(n)}(x) := \left[\mathcal{G}_{\eta,\xi_n(\omega)} \circ \dots \circ \mathcal{G}_{\eta,\xi_1(\omega)}\right](x). \tag{3.17}$$

It is easy to check that φ satisfies (3.15). Thus φ is a cocylce over θ and (θ, φ) a random dynamical system. The random variables X_n^{SGD} can now be expressed as

$$X_n^{\text{SGD}}(\omega) = \varphi_\omega^{(n)}\left(X_0^{\text{SGD}}(\omega)\right).$$

Note that the points $x \in \mathcal{M}$ are fixed points of φ, that is,

$$\varphi_\omega^{(n)}(x) = x, \ \forall\, n \in \mathbb{N}_0, \omega \in \Omega, x \in \mathcal{M}. \tag{3.18}$$

In order to introduce an RDS framework for the linearization of SGD (see Section 2.3), for SGD in local coordinates (see Section 3.2) and for linearized SGD in local coordinates, we will again fix a global minimum $x^* \in \mathcal{M}$ for the remainder of this section. We define a map $\Phi : \mathbb{N}_0 \times \Omega \to \mathbb{R}^{D \times D}$ by

$$\Phi : (n, \omega) \mapsto \Phi_\omega^{(n)} := \mathrm{D}_x \varphi_\omega^{(n)}(x^*), \tag{3.19}$$

where $\mathrm{D}_x \varphi_\omega^{(n)}(x^*)$ denotes the Jacobian of the map $\varphi_\omega^{(n)}$. Applying the chain rule to (3.17) and using (3.18), we get

$$\Phi_\omega^{(0)} = \mathbb{1}_D \qquad \text{and} \qquad \Phi_\omega^{(n)} = \mathcal{G}'_{\eta,\xi_n(\omega)}(x^*) \dots \mathcal{G}'_{\eta,\xi_1(\omega)}(x^*).$$

From this, it is easy to see that Φ is a *matrix cocycle*, that is, it satisfies

$$\Phi_\omega^{(n+m)} = \Phi_{\theta^n\omega}^{(m)} \Phi_\omega^{(n)}.$$

Since the matrices $\mathcal{G}'_{\eta,1}, \dots, \mathcal{G}'_{\eta,N}$ all respect the splitting $\mathbb{R}^D = \mathcal{T}(x^*) \oplus \mathcal{N}(x^*)$ with $\mathcal{G}'_{\eta,i}|_{\mathcal{T}(x^*)} = \mathrm{Id}_{\mathcal{T}(x^*)}$ (see Section 2.3), the same holds true for $\Phi_\omega^{(n)}$ for all $\omega \in \Omega$ and all $n \in \mathbb{N}_0$.

Recall that $\chi : \hat{V} \to \hat{U}$, introduced in Lemma 3.1, defines a diffeomorphism from a neighborhood $\hat{V}$ of the origin to a neighborhood $\hat{U}$ of x^*. Furthermore, we introduced a neighborhood $V^* \subseteq \hat{V}$ (see (3.2)), which allowed us to locally lift the maps $\mathcal{G}_{\eta,1}, \dots, \mathcal{G}_{\eta,n}$ via χ to the maps $\hat{\mathcal{G}}_{\eta,1}, \dots, \hat{\mathcal{G}}_{\eta,n} : V^* \to \hat{V}$ defined by

$$\hat{\mathcal{G}}_{\eta,i}(v, w) := \chi^{-1}(\mathcal{G}_{\eta,i}(\chi(v, w))).$$

Analogously to the definition of $\tau : V^* \to \mathbb{N} \cup \{\infty\}$ in Section 3.2 (see (3.3)), we introduce a map $\tau : \Omega \times V^* \to \mathbb{N} \cup \{\infty\}$ by

$$\tau_\omega(v, w) := \inf\left\{n \in \mathbb{N} : \left[\hat{\mathcal{G}}_{\eta,\xi_n(\omega)} \circ \dots \circ \hat{\mathcal{G}}_{\eta,\xi_1(\omega)}\right](v, w) \notin V^*\right\}.$$

This allows us to define a map $\psi : \mathbb{N}_0 \times \Omega \times V^* \supseteq D_\psi \to \hat{V}$, where

$$D_\psi := \{(n, \omega, v, w) \in \mathbb{N}_0 \times \Omega \times V^* : n \leq \tau_\omega(v, w)\},$$

by

$$\psi_\omega^{(n)}(v, w) := \left[\hat{\mathcal{G}}_{\eta,\xi_n(\omega)} \circ \cdots \circ \hat{\mathcal{G}}_{\eta,\xi_1(\omega)}\right](v, w).$$

Clearly, ψ satisfies the *local cocycle property*, meaning

$$\psi_\omega^{(n+m)}(x) = \psi_{\theta^n\omega}^{(m)}\left(\psi_\omega^{(n)}(x)\right)$$

whenever the left-hand side is well defined, that is, whenever $(n + m, \omega, v, w) \in D_\psi$. This turns the pair (θ, ψ) into a *local random dynamical system* (see Definition 1.2.1 in Arnold, 1998). The local cocycle ψ can be seen as the local lift of φ via χ, as one can easily check that

$$\psi_\omega^{(n)}(v, w) = \chi^{-1}\left(\varphi_\omega^{(n)}(\chi(v, w))\right), \tag{3.20}$$

whenever $n \leq \tau_\omega(v, w)$.

Similarly to the definition of Φ, we introduce a matrix cocycle $\Psi : \mathbb{N}_0 \times \Omega \to \mathbb{R}^{N\times N}$ by

$$\Psi_\omega^{(0)} := \mathbb{1}_N \qquad \text{and} \qquad \Psi_\omega^{(n)} = \left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \ldots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right). \tag{3.21}$$

Note that, as opposed to ψ, the matrix cocycle Ψ is defined globally. This is possible since $\tau_\omega(0, 0) = \infty$, for all $\omega \in \Omega$. Also by (3.5), we can express the Jacobian of $\psi_\omega^{(n)}$ at the origin by

$$\mathrm{D}_{(v,w)}\psi_\omega^{(n)}(0, 0) = \begin{pmatrix} \mathbb{1}_{D-N} & 0 \\ 0 & \Psi_\omega^{(n)} \end{pmatrix}.$$

Alternatively, this can also be seen by differentiating (3.21) and using Lemma 3.1, as well as (2.17). Furthermore, differentiating (3.20) at (0,0) yields the identity

$$\Phi_\omega^{(n)} = \begin{pmatrix} A & S_{x^*} \end{pmatrix} \begin{pmatrix} \mathbb{1}_{D-N} & 0 \\ 0 & \Psi_\omega^{(n)} \end{pmatrix} \begin{pmatrix} A & S_{x^*} \end{pmatrix}^{-1}, \tag{3.22}$$

where $\begin{pmatrix} A & S_{x^*} \end{pmatrix} \in \mathbb{R}^{D\times D}$ is the matrix from Lemma 3.1 (iii). The following corollary is a simple reformulation of Lemma 3.2.

Corollary 3.5 *For every $\delta > 0$, there exists a neighborhood $(0, 0) \in V_\delta \subseteq V^*$, such that for each $(v, w) \in V_\delta$, we have*

$$\left\|\psi_\omega^{(1)}(v, w) - \left(v, \Psi_\omega^{(1)} w\right)\right\| \leq \delta\|w\|.$$

Finally, we can reformulate (2.19) in terms of Ψ to get the expression

$$\lambda(x^*) = \inf_{n\in\mathbb{N}} \frac{1}{n}\mathbb{E}\left[\log\left\|\Psi_\omega^{(n)}\right\|\right] \tag{3.23}$$

for the Lyapunov exponent. By Kingman's sub-additive ergodic theorem[14], we have

$$\lambda(x^*) = \lim_{n\to\infty} \frac{1}{n} \log\left\|\Psi_\omega^{(n)}\right\|, \tag{3.24}$$

for almost every $\omega \in \Omega$.

14. Note that the fact that $\lambda(x^*)$ does not depend on ω requires θ to be ergodic, which it is not. However, the cocycle Ψ is measurable with respect to the sub-sigma algebra $\hat{\mathcal{F}} = \sigma(\xi_1, \xi_2, \ldots)$. Thus we may consider the probability space $(\Omega, \hat{\mathcal{F}}, \mathbb{P}|_{\hat{\mathcal{F}}})$, in which θ is ergodic, when applying the sub-additive ergodic theorem.

3.6 Stochastic Gradient Descent - the Stable Case

In this section we will prove Theorem B (i).

Theorem B (i) *Let $x^* \in \mathcal{M}$ with $\lambda(x^*) < 0$ and suppose*

$$\|\nabla_x \mathfrak{F}(x^*, y_i)\|^2 \neq \frac{1}{\eta}, \tag{3.25}$$

for every $i \in [N]$. Then $x^* \in \operatorname{supp}\left(X_{\lim}^{\mathrm{SGD}}\right)$.

Since this is the SGD equivalent of Theorem A (i), the proof will be similar. However, there is a major obstacle. A crucial ingredient in the proofs of Theorem A (i) is the expression $\|\mathbb{1}_N - \eta G_{x^*}\| = e^{\mu(x^*)}$ (see (3.10)), which allowed us to derive the bound (see (3.11))

$$\left\|\Pi_w \hat{\mathcal{G}}_\eta^n(v,w)\right\| \leq \gamma^n \|w\|. \tag{3.26}$$

For stochastic gradient descent, we only have a bound of the form

$$\left\|\Psi_\omega^{(n)}\right\| \leq \tilde{C}_\delta(\omega) e^{n(\lambda(x^*)+\delta)},$$

where for any $\delta > 0$, $\tilde{C}_\delta$ is a random variable which is finite almost surely. This can be seen as a consequence of (3.24). Random variables similar to $\tilde{C}_\delta$ are sometimes called *Oseledec regularity functions* in the literature. If $\tilde{C}_\delta(\omega)$ was almost surely bounded by some deterministic constant, it would be possible to derive the equivalent statement to (3.26) with an additional factor on the right-hand side. Unfortunately, the Oseledec regularity function $\tilde{C}_\delta$ will generally be unbounded. This is one of the defining features of so-called *non-uniform hyperbolicity*. Instead, we will derive the equivalent statement to (3.26) by essentially using the upper semi-continuity of the Lyapunov exponent (see Lemma 3.6 below). This allows for a proof of Theorem B (i) along the lines of the proof of Theorem A (i). It should be noted that this does not save the proof of Theorem *B* (ii). Thus an entirely different approach will be presented in sections 3.7-3.9.

Lemma 3.6 *Let $x^* \in \mathcal{M}$ be a point for which (3.25) is satisfied. For each $\gamma > e^{\lambda(x^*)}$, there exists a $\delta > 0$ and a random variable $C_\gamma : \Omega \to (0, \infty]$ such that $\mathbb{P}(C_\gamma(\omega) < \infty) = 1$ and*

$$\left\|\Pi_w \psi_\omega^{(n)}(v,w)\right\| \leq C_\gamma(\omega)\gamma^n\|w\|, \ \forall\, \omega \in \Omega,\ (v,w) \in V_\delta,\ n \leq \tau_{\delta,\omega}(v,w),$$

where V_δ is the neighborhood from Lemma 3.2/Corollary 3.5 and $\tau_{\delta,\omega} : V_\delta \to \mathbb{N} \cup \{\infty\}$ the map given by

$$\tau_{\delta,\omega}(v,w) := \inf\left\{n \in \mathbb{N} : \psi_\omega^{(n)}(v,w) \notin V_\delta\right\}.$$

Proof Let $x^* \in \mathcal{M}$ be such that (3.25) holds and $\gamma > e^{\lambda(x^*)}$. Choose an $\varepsilon > 0$ such that $e^{\lambda(x^*)+2\varepsilon} \leq \gamma$. By (3.23), there exists an $n^* \in \mathbb{N}$, such that

$$\mathbb{E}\left[\log\left\|\Psi_\omega^{(n^*)}\right\|\right] < n^*(\lambda(x^*) + \varepsilon). \tag{3.27}$$

Since (3.25) holds, the matrix $\Psi_\omega^{(1)}$ is invertible for every $\omega \in \Omega$ and we may define constants $0 < K_1 < K_2 < \infty$ by

$$K_1 := \inf_{\omega\in\Omega}\left\|\left(\Psi_\omega^{(1)}\right)^{-1}\right\|^{-1} = \min_{i\in[N]}\left\|\left(\mathbb{1}_N - \eta G_{x^*,[i]}\right)^{-1}\right\|^{-1}, \tag{3.28}$$

$$K_2 := \sup_{\omega\in\Omega}\left\|\Psi_\omega^{(1)}\right\| = \max_{i\in[N]}\left\|\mathbb{1}_N - \eta G_{x^*,[i]}\right\|. \tag{3.29}$$

By the cocycle property we have

$$K_1^n\|w\| \leq \left\|\Psi_\omega^{(n)} w\right\| \leq K_2^n\|w\|. \tag{3.30}$$

For some $\delta > 0$ to be determined later, let V_δ and $\tau_{\delta,\omega}$ be as described in the formulation of the Lemma. By Corollary 3.5, we have the naïve bound

$$\left\|\Pi_w \psi_\omega^{(1)}(v,w)\right\| \leq \left\|\Pi_w \psi_\omega^{(1)}(v,w) - \Psi_\omega^{(1)} w\right\| + \left\|\Psi_\omega^{(1)} w\right\| \leq (K_1+\delta)\|w\|$$

for all $\omega \in \Omega$ and $(v,w) \in V_\delta$. Thus we also have

$$\left\|\Pi_w \psi_\omega^{(n)}(v,w)\right\| \leq (K_2+\delta)^n\|w\|, \tag{3.31}$$

for all $\omega \in \Omega$, $(v,w) \in V_\delta$ and $n \leq \tau_{\delta,\omega}(v,w)$.

Suppose that $\omega \in \Omega$ and $(v,w) \in V_\delta$ satisfy $\tau_{\delta,\omega}(v,w) \geq n^*$. By the cocycle properties for ψ and Ψ, the inequality (3.31), Corollary 3.5 and the inequality (3.30), we have

$$\begin{aligned}
&\left\|\Pi_w \psi_\omega^{(n^*)}(v,w) - \Psi_\omega^{(n^*)} w\right\| \\
&\leq \sum_{n=1}^{n^*} \left\|\Pi_w \psi_{\theta^{n+1}\omega}^{(n^*-n-1)}\left(\psi_{\theta^n\omega}^{(1)}\left(v, \Psi_\omega^{(n)} w\right)\right) - \Pi_w \psi_{\theta^{n+1}\omega}^{(n^*-n-1)}\left(v, \Psi_{\theta^n\omega}^{(1)}\Psi_\omega^{(n)} w\right)\right\| \\
&\leq \sum_{n=1}^{n^*} (K_2+\delta)^{n^*-n-1}\left\|\Pi_w \psi_{\theta^n\omega}^{(1)}\left(v, \Psi_\omega^{(n)} w\right) - \Psi_{\theta^n\omega}^{(1)}\Psi_\omega^{(n)} w\right\| \\
&\leq \sum_{n=1}^{n^*} \delta(K_2+\delta)^{n^*-n-1}\left\|\Psi_\omega^{(n)} w\right\| \\
&\leq \sum_{n=1}^{n^*} \delta(K_2+\delta)^{n^*-n-1} K_1^{n^*-n}\left\|\Psi_\omega^{(n^*)} w\right\|.
\end{aligned}$$

We now fix $\delta > 0$ to have a small enough value, s.t.

$$\sum_{n=1}^{n^*} \delta(K_2+\delta)^{n^*-n-1} K_1^{n^*-n} \leq \varepsilon n^*.$$

Thus we have

$$\left\|\Pi_w \psi_\omega^{(n^*)}(v,w) - \Psi_\omega^{(n^*)} w\right\| \leq \varepsilon n^* \left\|\Psi_\omega^{(n^*)} w\right\|,$$

for all $\omega \in \Omega$ and $(v,w) \in V_\delta$ with $\tau_{\delta,\omega}(v,w) \geq n^*$. In that case we also have

$$\begin{aligned}
\left\|\Pi_w \psi_\omega^{(n^*)}(v,w)\right\| &\leq \left\|\Psi_\omega^{(n^*)} w\right\| + \left\|\Pi_w \psi_\omega^{(n^*)}(v,w) - \Psi_\omega^{(n^*)} w\right\| \\
&\leq (1+\varepsilon n^*)\left\|\Psi_\omega^{(n^*)} w\right\| \leq \hat{C}(\omega)\|w\|,
\end{aligned} \tag{3.32}$$

where we define $\hat{C}(\omega) := (1+\varepsilon n^*)\left\|\Psi_\omega^{(n^*)}\right\|$. By (3.27), we have

$$\begin{aligned}
\mathbb{E}\left[\log\left(\hat{C}(\omega)\right)\right] &= \log(1+\varepsilon n^*) + \mathbb{E}\left[\log\left(\Psi_\omega^{(n^*)}\right)\right] \\
&\leq \varepsilon n^* + n^*(\lambda(x^*)+\varepsilon) = n^*(\lambda(x^*)+2\varepsilon) < n^*\log(\gamma).
\end{aligned} \tag{3.33}$$

Next we introduce another random variable $\tilde{C} : \Omega \to \mathbb{R} \cup \{\infty\}$ by

$$\tilde{C}(\omega) := \sup_{\ell \in \mathbb{N}} \sum_{k=0}^{\ell-1} \left[\log\left(\hat{C}(\theta^{kn^*}\omega) \right) - n^* \log(\gamma) \right].$$

Note that the random variable $\hat{C}$ is measurable with respect to the sigma algebra $\mathcal{F}_{n^*} := \sigma(\xi_1, \dots, \xi_{n^*})$ and that the random variables $\left(\hat{C}(\theta^{kn^*}\omega) \right)_{k \in \mathbb{N}_0}$ are thus independent and by θ-invariance (see (3.16)) identically distributed. By the strong law of large numbers, we have, for almost every $\omega \in \Omega$,

$$\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{k=0}^{\ell-1} \left[\log\left(\hat{C}(\theta^{kn^*}\omega) \right) - n^* \log(\gamma) \right] = \mathbb{E}\left[\log\left(\hat{C}(\omega) \right) \right] - n^* \log(\gamma) < 0$$

and thus in particular $\mathbb{P}(\tilde{C}(\omega) < \infty) = 1$. By definition, the random variable $\tilde{C}$ can be used to obtain the bound

$$\sum_{k=0}^{\ell-1} \log\left(\hat{C}(\theta^{kn^*}\omega) \right) \leq \tilde{C}(\omega) + \ell n^* \log(\gamma). \tag{3.34}$$

Combining (3.32) and (3.34)), we get

$$\begin{aligned} \left\| \Pi_w \psi_\omega^{(\ell n^*)}(v, w) \right\| &\leq \left[\prod_{k=0}^{\ell-1} \hat{C}\left(\theta^{kn^*}\omega \right) \right] \|w\| \\ &= \exp\left(\sum_{k=0}^{\ell-1} \log\left(\hat{C}\left(\theta^{kn^*}\omega \right) \right) \right) \|w\| \\ &\leq e^{\tilde{C}(\omega) + \ell n^* \log(\gamma)} \|w\| = e^{\tilde{C}(\omega)} \gamma^{\ell n^*} \|w\|, \end{aligned} \tag{3.35}$$

for all $\omega \in \Omega$, $(v, w) \in V_\delta$ and $\ell \in \mathbb{N}$ with $\tau_{\delta,\omega}(v, w) \geq \ell n^*$.

Now let $\omega \in \Omega$, $(v, w) \in V_\delta$ and $n \in \mathbb{N}$ with $\tau_{\delta,\omega}(v, w) \geq n$. Let $\ell \in \mathbb{N}_0$ and $0 \leq k \leq n^* - 1$ be such that $n = \ell n^* + k$. Using the cocycle property and the bounds (3.31) and (3.35), we obtain

$$\begin{aligned} \left\| \Pi_w \psi_\omega^{(n)}(v, w) \right\| &= \left\| \Pi_w \psi_{\theta^{\ell n^*}\omega}^{(k)} \left(\psi_\omega^{(\ell n^*)}(v, w) \right) \right\| \\ &\leq (K_2 + \delta)^k \left\| \Pi_w \psi_\omega^{(\ell n^*)}(v, w) \right\| \\ &\leq (K_2 + \delta)^k e^{\tilde{C}(\omega)} \gamma^{\ell n^*} \|w\| \\ &= \left(\frac{K_2 + \delta}{\gamma} \right)^k e^{\tilde{C}(\omega)} \gamma^{\ell n^* + k} \|w\| \\ &\leq \left(\frac{K_2 + \delta}{\gamma} \right)^{n^* - 1} e^{\tilde{C}(\omega)} \gamma^n \|w\|. \end{aligned}$$

Thus we can define the random variable $C_\gamma : \Omega \to \mathbb{R} \cup \infty$ by

$$C_\gamma(\omega) := \left(\frac{K_2 + \delta}{\gamma} \right)^{n^* - 1} e^{\tilde{C}(\omega)},$$

satisfying $\mathbb{P}(C_\gamma(\omega) < \infty) = 1$. ∎

With Lemma 3.6 in place, we can prove Theorem B (i) analogously to the proof of Theorem A (i) presented in Section 3.3.

Proof [of Theorem B (i)] Suppose $\lambda(x^*) < 0$ and let $U \subseteq \mathcal{M}$ be some neighborhood of x^*. Our goal is to prove $\mathbb{P}(X_{\lim}^{\mathrm{SGD}} \in U) > 0$. Choose $\gamma \in \mathbb{R}$, such that $e^{\lambda(x^*)} < \gamma < 1$ and let $\delta > 0$ and $C_\gamma : \Omega \to \mathbb{R} \cup \infty$ be such that the conclusion of Lemma 3.6 holds, that is, such that $\mathbb{P}(\mathbb{C}_\gamma(\omega) < \infty) = 1$ and such that we have

$$\left\|\Pi_w \psi_\omega^{(n)}(v,w)\right\| \leq C_\gamma(\omega)\gamma^n \|w\|, \tag{3.36}$$

for all $\omega \in \Omega$, $(v,w) \in V_\delta$ and $n \leq \tau_{\delta,\omega}(v,w)$. This bound will serve as the equivalent to (3.11) in the proof of Theorem A (i). Let $R_\delta > 0$ be some radius, such that

$$\mathcal{B}_{\mathbb{R}^{D-N}}(R_\delta) \times \mathcal{B}_{\mathbb{R}^N}(R_\delta) \subseteq V_\delta.$$

We may assume without loss of generality that U has the form

$$U = \chi\left(\mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}\right)$$

for some $R \leq R_\delta$. As a consequence of Corollary 3.5 we get

$$\left\|\Pi_v \psi_\omega^{(1)}(v,w) - v\right\| \leq \delta\|w\|,$$

for all $(v,w) \in V_\delta$ and thus also for all $1 \leq n < \tau_{\delta,\omega}(v,w)$,

$$\left\|\Pi_v \psi_\omega^{(n+1)}(v,w) - \Pi_v \psi_\omega^{(n)}(v,w)\right\| \leq \delta \left\|\Pi_w \psi_\omega^{(n)}(v,w)\right\| \leq \delta C_\gamma(\omega)\gamma^n\|w\|. \tag{3.37}$$

With this, we can bound

$$\begin{aligned}
\left\|\Pi_v \psi_\omega^{(n)}(v,w)\right\| &\leq \|v\| + \sum_{m=0}^{n-1} \left\|\Pi_v \psi_\omega^{(m+1)}(v,w) - \Pi_v \psi_\omega^{(m)}(v,w)\right\| \\
&\leq \|v\| + \sum_{m=0}^{n-1} \delta C_\gamma(\omega)\gamma^m\|w\| \\
&\leq \|v\| + \frac{\delta C_\gamma(\omega)}{1-\gamma}\|w\|,
\end{aligned} \tag{3.38}$$

for all $(v,w) \in V_\delta$ and $1 \leq n+1 < \tau_{\delta,\omega}(v,w)$. Now set

$$R_v := \frac{R}{2} \qquad \text{and} \qquad R_w(\omega) = \min\left(\frac{(1-\gamma)R}{2\delta C_\gamma(\omega)}, \frac{R_\delta}{C_\gamma(\omega)}\right).$$

Suppose for some $\omega \in \Omega$ with $C_\gamma(\omega) < \infty$ and some $(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w(\omega))$, we have $\tau_{\delta,\omega}(v,w) < \infty$. Then by definition $\psi_\omega^{(\tau_{\delta,\omega}(v,w))}(v,w) \notin V_\delta$, so in particular

$$\left\|\Pi_v \psi_\omega^{(\tau_{\delta,\omega}(v,w))}(v,w)\right\| \geq R_\delta \qquad \text{or} \qquad \left\|\Pi_w \psi_\omega^{(\tau_{\delta,\omega}(v,w))}(v,w)\right\| \geq R_\delta. \tag{3.39}$$

However, (3.38) implies

$$\left\|\Pi_v \psi_\omega^{(\tau_{\delta,\omega}(v,w))}(v,w)\right\| \leq \|v\| + \frac{\delta C_\gamma(\omega)}{1-\gamma}\|w\| < R_v + \frac{\delta C_\gamma(\omega)}{1-\gamma} R_w(\omega) \leq R \leq R_\delta$$

and (3.36) implies

$$\left\|\Pi_w \psi_\omega^{(\tau_{\delta,\omega}(v,w))}(v,w)\right\| \leq C_\gamma(\omega)\gamma^{\tau_{\delta,\omega}(v,w)}\|w\| \leq C_\gamma(\omega)\|w\| < C_\gamma R_w \leq R_\delta,$$

contradicting (3.39). Thus $\tau_{\delta,\omega}(v,w) = \infty$, for all $(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v(\omega)) \times \mathcal{B}_{\mathbb{R}^N}(R_w(\omega))$. Now (3.37) implies that $\left(\psi_\omega^{(n)}(v,w)\right)$ is a Cauchy sequence and (3.38) shows that

$$\left\|\lim_{n\to\infty} \Pi_v \psi_\omega^{(n)}(v,w)\right\| \leq \|v\| + \frac{\delta C_\gamma(\omega)}{1-\gamma}\|w\| < R_v + \frac{\delta C_\gamma(\omega)}{1-\gamma} R_w(\omega) \leq R.$$

Furthermore, (3.36) shows $\Pi_w \psi_\omega^{(n)}(v,w) \to 0$. Thus, for each $(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w(\omega))$ the sequence $\left(\psi_\omega^{(n)}(v,w)\right)$ converges with

$$\lim_{n\to\infty} \psi_\omega^{(n)}(v,w) \in \mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}.$$

Let $\tilde{U}(\omega) = \chi\left(\mathcal{B}_{\mathbb{R}^{D-N}}(R_v) \times \mathcal{B}_{\mathbb{R}^N}(R_w(\omega))\right)$ whenever $C_\gamma(\omega) < \infty$. Suppose $X_0^{\mathrm{SGD}}(\omega) \in \tilde{U}(\omega)$. By (3.20) and continuity of χ we have

$$\begin{aligned} X_{\lim}^{\mathrm{SGD}}(\omega) &= \lim_{n\to\infty} \varphi_\omega^{(n)}(X_0^{\mathrm{SGD}}(\omega)) = \lim_{n\to\infty} \chi\left(\psi_\omega^{(n)}\left(\chi^{-1}(X_0^{\mathrm{SGD}}(\omega))\right)\right) \\ &= \chi\left(\lim_{n\to\infty} \psi_\omega^{(n)}\left(\chi^{-1}(X_0^{\mathrm{SGD}}(\omega))\right)\right) \in \chi(\mathcal{B}_{\mathbb{R}^{D-N}}(R) \times \{0\}) = U, \end{aligned}$$

so $X_0^{\mathrm{SGD}}(\omega) \in \tilde{U}(\omega)$ implies $X_{\lim}^{\mathrm{SGD}}(\omega) \in U$. By construction $\tilde{U}$ is measurable with respect to $\sigma(\xi_1, \xi_2, \dots)$ and thus independent of X_0^{SGD}. Therefore we have

$$\mathbb{P}(X_{\lim}^{\mathrm{SGD}}(\omega) \in U) \geq \mathbb{P}\left(X_0^{\mathrm{SGD}}(\omega) \in \tilde{U}(\omega)\right) = \mathbb{E}\left[\nu\left(\tilde{U}(\omega)\right)\right].$$

Since $\tilde{U}$ is a non-empty open set almost surely, by Hypothesis (H2) we get

$$\mathbb{P}(X_{\lim}^{\mathrm{SGD}}(\omega) \in U) \geq \mathbb{E}\left[\nu\left(\tilde{U}(\omega)\right)\right] > 0,$$

completing the proof.

■

3.7 Generated Matrix semigroups

It remains to prove Theorem B (ii). The proof of Theorem B (i) in the previous section was of a *quenched* nature. One might expect that the best approach to proving Theorem B (ii) is to construct ω-wise center-stable manifolds similar to the ones constructed in Section 3.4. While an invariant manifold theory (Pesin theory) has been developed for random dynamical systems (see, e.g., Liu and Qian, 1995 or Arnold, 1998), these results only provide center-stable manifolds for single points $x \in \mathcal{M}$. For the argument in Section 3.4 it was crucial to have a center-stable manifold for an open subset of $\mathcal{M}$. The authors are not aware of any method to construct such a random center-stable manifold.

In the following, we will present an *annealed* argument. Instead of showing that, given $\lambda(x^*) > 0$, for almost every ω the points which converge to any $X_{\lim} \in \mathcal{M}$ near x^* form a

ν -null set, we will show that ν-almost every initial condition $x_0 \in \mathbb{R}^D$ does not converge to any $X_{\lim} \in \mathcal{M}$ near x^* almost surely. Both statements are equivalent to Theorem B (ii) by Fubini's theorem, but they are different in flavor. The former is a statement on the deterministic dynamics for a fixed ω, while the latter concerns the stochastic behavior of a Markov process. The advantage of the stochastic approach is that under the conditions we impose, namely regularity of x^* in the sense of Definition 2.2, the center-stable manifolds get "washed away" by the randomness: while for every ω there might exist a manifold of initial conditions which still converge to $\mathcal{M}$ near x^*, for every initial condition $X_0 \notin \mathcal{M}$, the probability of converging to some $X_{\lim} \in \mathcal{M}$ near x^* is zero.

We will show this by constructing a *Lyapunov function* (see, e.g., Benaïm and Hurth, 2022) defined on a neighborhood of x^* which goes to infinity near $\mathcal{M}$. This is inspired by previous work on the instability of invariant subspaces for stochastic differential equations (see Baxendale, 1991; Baxendale and Stroock, 1988; Bedrossian et al., 2022; Blumenthal et al., 2023; Coti Zelati and Hairer, 2021). These works rely on conditions of Hörmander type to establish the existence of a spectral gap in the so-called projective process. Such a spectral gap is excluded by the discrete nature of our problem. Instead, we use an argument due to Le Page (1982) (see also Bougerol and Lacroix, 1985 for a survey in English) to find a spectral gap in a different topology. Le Page's argument needs the matrix semigroup on which the linear cocycle Ψ is supported to satisfy two algebraic properties, namely being contracting and strongly irreducible (see Definition 3.8 below). In this section, we will show that they follow from the regularity of x^*.

For some $x^* \in \mathcal{M}$, we denote the support of the matrix-valued random variable $\Psi_\bullet^{(n)}$ by $\mathcal{S}_n(x^*)$, that is,

$$\mathcal{S}_n(x^*) := \operatorname{supp}\left(\Psi_\bullet^{(n)}\right) \subset \mathbb{R}^{N\times N}.$$

From the definition of Ψ (3.21), one can readily see that

$$\mathcal{S}_n(x^*) = \left\{ \left(\mathbb{1}_N - \eta G_{x^*,[\xi_n]}\right) \dots \left(\mathbb{1}_N - \eta G_{x^*,[\xi_1]}\right) : (\xi_1, \dots, \xi_n) \in [N]^n \right\}.$$

Furthermore, we denote the total support of Ψ by $\mathcal{S}(x^*)$, that is,

$$\mathcal{S}(x^*) := \bigcup_{n=0}^{\infty} \mathcal{S}_n(x^*) \subset \mathbb{R}^{N\times N}.$$

Clearly, $\mathcal{S}(x^*)$ is a matrix semigroup with unity, generated by $\mathcal{S}_1(x^*)$, that is, $\mathcal{S}(x^*)$ contains exactly those matrices, which can be expressed as the product of an arbitrary number of elements in $\mathcal{S}_1(x^*)$, including the empty product, which is defined to be the identity matrix.

Recall from Definition 2.2, that a point $x^* \in \mathcal{M}$ is called regular, if for every $i \in \mathbb{N}$, we have

$$[G_{x^*}]_{i,i} = \|\nabla_x \mathfrak{F}(x^*, y_i)\|^2 \notin \left\{ \frac{1}{\eta}, \frac{2}{\eta} \right\},$$

and if there exists no proper subset $\emptyset \subsetneq \mathcal{A} \subsetneq [N]$, such that

$$[G_{x^*}]_{i,j} = \nabla_x \mathfrak{F}(x^*, y_i) \cdot \nabla_x \mathfrak{F}(x^*, y_j) = 0, \ \forall\, i \in \mathcal{A},\ j \in [N] \setminus \mathcal{A}.$$

Proposition 3.7 *Let $x^* \in \mathcal{M}$ be regular. Then $\mathcal{S}(x^*) \subseteq \mathrm{GL}(N, \mathbb{R})$.*

Proof Since $x^* \in \mathcal{M}$ is regular, in particular

$$\|\nabla_x \mathfrak{F}(x^*, y_i)\|^2 \neq \frac{1}{\eta}, \ \forall\, i \in [N].$$

Thus $\mathcal{S}_1(x^*) \subseteq \mathrm{GL}(N, \mathbb{R})$ and since $\mathcal{S}(x^*)$ is generated by $\mathcal{S}_1(x^*)$ also $\mathcal{S}(x^*) \subseteq \mathrm{GL}(N, \mathbb{R})$. ■

Definition 3.8 *A matrix semigroup of invertible matrices $\mathcal{S} \subseteq \mathrm{GL}(N, \mathbb{R})$ is called,*

(i) contracting*, if there exists a sequence $(M_n)_{n\in\mathbb{N}} \subset \mathcal{S}$, such that*

$$\lim_{n\to\infty} \frac{M_n}{\|M_n\|} = M,$$

for some rank-1 matrix $M \in \mathbb{R}^{N\times N}$.

(ii) strongly irreducible*, if for every proper linear subspace $\{0\} \subsetneq W \subsetneq \mathbb{R}^N$ the set of subspaces $\{MW : M \in \mathcal{S}\}$ contains infinitely many elements.*

Lemma 3.9 *Let $x^* \in \mathcal{M}$ be a regular point with $\lambda(x^*) > 0$. Then $\mathcal{S}(x^*) \subseteq \mathrm{GL}(N, \mathbb{R})$ is both contracting and strongly irreducible.*

Proof Let $x^* \in \mathcal{M}$ be a regular point with $\lambda(x^*) > 0$. For ease of notation, we will write $G := G_{x^*}$ and $G_{[i]} := G_{x^*,[i]}$. We start by proving that $\mathcal{S}(x^*)$ is contracting. From the original definition (2.12) of $\lambda(x^*)$, we get the inequality

$$\begin{aligned} 0 < \lambda(x^*) \leq \mathbb{E}\left[\log \left\|\mathcal{G}'_{\eta,\xi_1}(x^*)\right\|\right] &\leq \max_{i\in[N]} \log \left\|\mathcal{G}'_{\eta,i}(x^*)\right\| \\ &= \max_{i\in[N]} \log \left\|\left(\mathbb{1}_D - \eta \nabla_x \mathfrak{F}(x^*, y_i) \nabla_x \mathfrak{F}(x^*, y_i)^t\right)|_{\mathcal{N}(x^*)}\right\|. \end{aligned}$$

In particular, there must exist an $i^* \in [N]$ with

$$\left|1 - \eta \left\|\nabla_x \mathfrak{F}(x^*, y_{i^*})\right\|^2\right| = |1 - \eta G_{i^*,i^*}| > 1.$$

Without loss of generality, we assume $i^* = 1$. Consider the sequence $(M_n) \in \mathcal{S}(x^*)^{\mathbb{N}}$ given by

$$M_n := \left(\mathbb{1}_N - \eta G_{[1]}\right)^n.$$

Recall that $G_{[1]}$ is the matrix G with all but the first row replaced by zeros. Thus $[G_{[1]}]$ is a rank 1 matrix with non-trivial eigenvalue $G_{1,1}$ and the eigenvalues of $\mathbb{1}_N - \eta G_{[1]}$ are $\mu_1 = 1 - \eta G_{1,1}$ with multiplicity 1 and $\mu_2 = 1$ with multiplicity $N - 1$. Using basic finite-dimensional spectral theory, $\mathbb{1}_N - \eta G_{[1]}$ can be decomposed as $\mathbb{1}_N - \eta G_{[1]} = A + B$, where A is a rank-1 matrix with $A^n = \mu_1^{n-1} A$, $AB = BA = 0$ and

$$\lim_{n\to\infty} \frac{\|B^n\|}{\|A^n\|} = 0. \tag{3.40}$$

Now we can compute

$$M_n = \left(\mathbb{1}_N - \eta G_{[1]}\right)^n = (A + B)^n = A^n + B^n$$

and thus

$$\lim_{n\to\infty} \frac{M_n}{\|M_n\|} = \lim_{n\to\infty} \frac{A^n + B^n}{\|A^n + B^n\|} = \lim_{n\to\infty} \frac{\|A^n\|}{\|A^n + B^n\|} \frac{A^n}{\|A^n\|} + \frac{B^n}{\|A^n + B^n\|}.$$

As a consequence of (3.40), we have

$$\lim_{n\to\infty} \frac{\|A_n\|}{\|A^n + B^n\|} = 1 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{B^n}{\|A^n + B^n\|} = 0$$

and hence

$$\lim_{n\to\infty} \frac{M_n}{\|M_n\|} = \lim_{n\to\infty} \frac{A^n}{\|A^n\|} = \lim_{n\to\infty} \frac{\mu_1^{n-1} A}{\mu_1^{n-1} \|A\|} = \frac{A}{\|A\|},$$

which is indeed a rank-1 matrix.

In order prove that $\mathcal{S}(x^*)$ is strongly irreducible, consider a proper linear subspace $\{0\} \subsetneq W \subsetneq \mathbb{R}^N$ for wich we intend to show that $\{MW : M \in \mathcal{S}\}$ contains infinitely many elements. Consider the sets $\mathcal{A}, \mathcal{B} \subseteq [N]$ given by

$$\mathcal{A} := \{i \in [N] : e_i \in W\} \qquad \text{and} \qquad \mathcal{B} := \{i \in [N] : Ge_i \in W^\perp\},$$

where e_i denotes the i-th unit vector. Since G is positive definite[15], we have $e_i^t G e_i > 0$ for all i and thus $\mathcal{A} \cap \mathcal{B} = \emptyset$. Also, we have $G_{i,j} = e_i^t G e_j = 0$, for each $i \in \mathcal{A}$ and $j \in \mathcal{B}$. By the assumption that x^* is regular, this implies $\mathcal{B} \neq [N] \setminus \mathcal{A}$ and thus $\mathcal{A} \cup \mathcal{B} \neq [N]$. In other words, there must exists an $i^* \in [N]$, such that $e_{i^*} \notin W$ and $Ge_{i^*} \notin W^\perp$. The latter implies that there must exists some $w^* \in W$ such that $e_i^t G w^* = w^{*t} G e_i \neq 0$. Again, we assume without loss of generality that $i^* = 1$. Consider the sequence of subspaces $(W_n) \in \{MW : M \in \mathcal{S}\}^{\mathbb{N}_0}$ given by $W_n := (\mathbb{1}_N - \eta G_{[1]})^n W$ and let

$$\kappa_n := \sup_{w \in W_n \setminus \{0\}} \frac{\|G_{[1]} w\|}{\left\| w - \frac{1}{G_{1,1}} G_{[1]} w \right\|}.$$

Note that since the term in the supremum only depends on the direction of w and not on $\|w\|$, it is sufficient to take the supremum over the unit ball and by compactness the supremum must be attained. Since $G_{[1]} w$ is always a multiple of e_1 and $e_1 \notin W_0 = W$, the denominator is always non-zero for $n = 0$. Furthermore, $e_1^t G w^* \neq 0$ implies $G_{[1]} w^* \neq 0$ and we have $0 < \kappa_0 < \infty$. Also, using $G_{[1]}^2 = G_{i,i} G_{[1]}$, we get

$$G_{[1]}(\mathbb{1}_N - \eta G_{[1]}) w = (1 - \eta G_{1,1}) G_{[1]} w \quad \text{and}$$
$$(\mathbb{1}_N - \eta G_{[1]}) w - \frac{1}{G_{1,1}} G_{[1]} (\mathbb{1}_N - \eta G_{[1]}) w = w - \frac{1}{G_{1,1}} G_{[1]} w.$$

Applying this iteratively, allows us to compute

$$\begin{aligned}
\kappa_n &= \sup_{w \in W \setminus \{0\}} \frac{\|G_{[1]} (\mathbb{1}_N - \eta G_{[1]})^n w\|}{\left\| (\mathbb{1}_N - \eta G_{[1]})^n w - \frac{1}{G_{1,1}} G_{[1]} (\mathbb{1}_N - \eta G_{[1]})^n w \right\|} \\
&= \sup_{w \in W \setminus \{0\}} |1 - \eta G_{1,1}|^n \frac{\|G_{[1]} w\|}{\left\| w - \frac{1}{G_{1,1}} G_{[1]} w \right\|} \\
&= |1 - \eta G_{1,1}|^n \kappa_0.
\end{aligned}$$

15. Recall that G is the Gram matrix of the neural tangent kernel.

Since $G_{1,1} \neq \frac{2}{\eta}$, we have $|1 - \eta G_{1,1}| \neq 1$ and the sequence (κ_n) consists of pairwise distinct elements. Thus, in particular, the sequence (W_n) consists of pairwise distinct subspaces, completing the proof. ■

3.8 Construction of Lyapunov Functions

The goal of this section is, given a global minimum x^* with $\lambda(x^*) > 0$, to construct a Lyapunov function $F^* : \mathbb{R}^N \setminus \{0\} \to [0, \infty)$ which goes to infinity near 0 and such that

$$\mathbb{E}\left[F^*\left(\psi_\omega^{(1)}(v, w)_2\right)\right] \leq \tilde{\gamma} F^*(w), \tag{3.41}$$

for some $\tilde{\gamma} \in (0, 1)$ and every (v, w) in some neighborhood of $(0, 0)$. This means that the value of $F^*(w)$ must decrease on average along trajectories. Since F^* is large near the origin, points get "pushed away" from the set $\{w = 0\}$, which will allow us to prove a lack of convergence in the subsequent section.

To construct F^* we employ a strategy that is inspired by recent advances in fluid dynamics (Bedrossian et al., 2022, 2025; Blumenthal et al., 2023). We construct the function F^* to be a Lyapunov function for the linearized dynamics, that is,

$$\mathbb{E}\left[F^*(\Psi_\omega^{(1)} w)\right] = \gamma F^*(w), \ \forall\, w \in \mathbb{R}^N \setminus \{0\}, \tag{3.42}$$

for some $\gamma \in (0, 1)$. From this we can conclude (Corollary 3.16 below) that (3.41) holds locally around the origin for any $\tilde{\gamma} \in (\gamma, 1)$. Making the ansatz

$$F^*(w) = \|w\|^{-p} f^*\left(\frac{w}{\|w\|}\right),$$

with $p > 0$ and $f^* : S^{N-1} \to [0, \infty)$, equation (3.42) is satisfied if and only if

$$\mathbb{E}\left[\left\|\Psi_\omega^{(1)} s\right\|^{-p} f^*\left(\frac{\Psi_\omega^{(1)} s}{\left\|\Psi_\omega^{(1)} s\right\|}\right)\right] = \gamma f^*(s), \ \forall\, s \in S^{N-1}.$$

We will study the family of linear operators $(\mathcal{P}_q : \mathcal{C}^0(S^{N-1}) \to \mathcal{C}^0(S^{N-1}))_{q \in \mathbb{R}}$ given by

$$[\mathcal{P}_q f](s) = \mathbb{E}\left[\left\|\Psi_\omega^{(1)} s\right\|^{q} f\left(\frac{\Psi_\omega^{(1)} s}{\left\|\Psi_\omega^{(1)} s\right\|}\right)\right]. \tag{3.43}$$

Here, $\mathcal{C}^0(S^{N-1})$ denotes the Banach space of real-valued continuous functions on S^{N-1}, equipped with the supremum norm. It can be readily seen that the operators $\mathcal{P}_q$ are bounded in this norm[16]. Furthermore, the operators $\mathcal{P}_q$ are positive, that is, they map non-negative functions to non-negative functions. We can reformulate (3.43) as the eigenvalue problem

$$\mathcal{P}_{-p} f^* = \gamma f^*, \quad f^* \geq 0, \ 0 < \gamma < 1, \ p > 0. \tag{3.44}$$

Let us pretend for now that the operators $\mathcal{P}_q$ are compact and strongly positive[17], that is, they map any non-negative function that is positive somewhere to a function that is positive

16. In fact $\|\mathcal{P}_q\| \leq K_2^q$ for $q \geq 0$ and $\|\mathcal{P}_q\| \leq K_1^{-q}$ for $q \leq 0$, with K_1 and K_2 defined by (3.28) and (3.29).
17. Neither of these properties actually hold.

everywhere. By the Krein-Rutman Theorem (Krein and Rutman, 1950), for each $q \in \mathbb{R}$, there would be a unique non-negative eigenfunction $f_q \geq 0$ whose corresponding eigenvalue, say $r(q)$ is simple, isolated, and principal. In this setting, the principal eigenvalue $r(q)$ is closely related to the q-th moment Lyapunov exponent Λ_q (see Definition A.2), by the identity $\Lambda_q = \log r(q)$ (see, e.g. Arnold and Kliemann, 1987, Lemma 2). To find a solution to (3.44), we therefore look for a $p > 0$ with $\Lambda_{-p} < 0$. In settings where the Krein-Rutman Theorem holds for the operators $\mathcal{P}_q$, the moment Lyapunov exponent Λ_q is known to satisfy $\Lambda_0 = 0$ and $\frac{\mathrm{d}}{\mathrm{d}q}\Lambda_q|_{q=0} = \lambda$ (Arnold and Kliemann, 1987, Theorem 1). If $\lambda > 0$, we can choose $0 < p \ll 1$ to have $\Lambda_{-p} < 0$ and thus a solution to (3.44).

If the operators $\mathcal{P}_q$ are not compact, a similar argument is still possible if one assumes that the operator $\mathcal{P}_0$ admits a spectral gap. By definition, the operator $\mathcal{P}_0$ is the Markov operator of the linearized dynamics projected onto the unit sphere. In particular, $\|\mathcal{P}_0\| = 1$ and the constant function $\mathbf{1} \in C^0(S^{N-1})$ is an eigenfunction with $\mathcal{P}_0\mathbf{1} = \mathbf{1}$. We say that $\mathcal{P}_0$ admits a spectral gap if 1 is a simple eigenvalue and the remaining spectrum is contained in some ball with radius less than 1. This is equivalent to the associated Markov process being uniformly geometrically ergodic (see, e.g., Hairer, 2021). Under the spectral gap assumption one can use tools from perturbation theory (Kato, 1995) to show that the principal eigenvalue $r(q)$ of $\mathcal{P}_q$ is analytic in q for a neighborhood of 0. This turns out to be sufficient to find a solution to the eigenvalue problem (3.44), whenever $\lambda > 0$ (see Bedrossian et al., 2022, 2025; Blumenthal et al., 2023).

Unfortunately, the discrete nature of our setting prohibits uniform geometric ergodicity and thus a spectral gap for $\mathcal{P}_0$, at least in $\mathcal{C}^0(S^{N-1})$. In order to circumvent this obstacle, we must consider a different Banach space. For $\alpha \in (0,1)$, we let $\mathcal{C}^\alpha(S^{N-1})$ denote the Banach space of α-Hölder continuous functions on the unit sphere $S^{N-1} \subset \mathbb{R}^N$, that is,

$$\mathcal{C}^\alpha(S^{N-1}) = \left\{ f : S^{N-1} \to \mathbb{R} : \exists\, h > 0 \text{ s.t. } |f(s_1) - f(s_2)| \leq h\|s_1 - s_2\|^\alpha,\ \forall\, s_1, s_2 \in S^{N-1} \right\}$$

and

$$\|f\|_{C^\alpha} = \|f\|_\infty + \sup_{s_1,s_2 \in S^{N-1}, s_1 \neq s_2} \frac{|f(s_1) - f(s_2)|}{\|s_1 - s_2\|^\alpha}.$$

As we will only consider the space of Hölder-continuous functions on S^{N-1} here, we will abbreviate $\mathcal{C}^\alpha = \mathcal{C}^\alpha(S^{N-1})$. Also, $L(\mathcal{C}^\alpha)$ denotes the space of bounded linear operators from $\mathcal{C}^\alpha$ to itself.

Lemma 3.10 (Proposition V.4.1 in Part A of Bougerol and Lacroix, 1985) *If $\mathcal{S}(x^*)$ is contracting and strongly irreducible, then there exists an $\alpha \in (0,1)$ such that*

(i) there exists an $\hat{q} > 0$, such that for $q \in (-\hat{q}, \hat{q})$ the operator $\mathcal{P}_p \in L(\mathcal{C}^0)$ restricts to a well-defined, bounded operator $\mathcal{P}_q \in L(\mathcal{C}^\alpha)$ and the map $\mathcal{P}_\bullet : (-\hat{q}, \hat{q}) \to L(\mathcal{C}^\alpha)$, $q \mapsto \mathcal{P}_q$ is analytic

(ii) and the operator $\mathcal{P}_0$ satisfies

$$\limsup_{n\to\infty} \|\mathcal{P}_0^n f - \kappa(f)\mathbf{1}\|_{\mathcal{C}^\alpha}^{\frac{1}{n}} < 1, \forall f \in \mathcal{C}^\alpha,$$

for some probability measure κ on S^{N-1}. Here, we let $\mathbf{1} \in \mathcal{C}^\alpha$ denote the constant function with value 1.

For the proof, we refer to Bougerol and Lacroix (1985). The integrability assumption is satisfied trivially in our setting, as $\Psi_\omega^{(1)}$ can only take finitely many values.

Lemma 3.10 (ii) says that 1 is an dominant eigenvalue of $\mathcal{P}_0$, that is, it is isolated and the spectral value with the largest absolute value. By classical perturbation theory (Kato, 1995) this implies that for q sufficiently close to 0 the largest spectral value of $\mathcal{P}_q$ is also an isolated eigenvalue and both the dominant eigenvalue and the corresponding Riesz projections are analytic in q. Put more precisely, we get the following corollary (see also Theorem V.4.3 in Part A of Bougerol and Lacroix, 1985). Here $\dot{\mathcal{C}}^\alpha$ denotes the dual space of $\mathcal{C}^\alpha$ and $\mathcal{Q}^* \in L(\dot{\mathcal{C}}^\alpha)$ the dual operator of an operator $\mathcal{Q} \in L(\mathcal{C}^\alpha)$.

Corollary 3.11 *In the setting of the previous lemma, there exists a $0 < \tilde{q} < \hat{q}$ and analytic maps $r : (-\tilde{q}, \tilde{q}) \to \mathbb{R}$, $f_\bullet : (-\tilde{q}, \tilde{q}) \to \mathcal{C}^\alpha$, $\kappa_\bullet : (-\tilde{q}, \tilde{q}) \to \dot{\mathcal{C}}^\alpha$ and $\mathcal{Q}_\bullet : (-\tilde{q}, \tilde{q}) \to L(\mathcal{C}^\alpha)$, such that*

$$\mathcal{P}_q f = \mathcal{Q}_q f + r(q)\langle \kappa_q, f\rangle f_q, \tag{3.45}$$

where $\mathcal{Q}_q f_q = 0$, $\mathcal{Q}_q^* \kappa_q = 0$, $\langle \kappa_q, f_q\rangle = 1$ and

$$\limsup_{n\to\infty} \left\|\mathcal{Q}_q^n\right\|^{\frac{1}{n}} < r(p).$$

Furthermore, $r(0) = 1$, $f_0 = \mathbf{1}$ and $\kappa_0 \in \dot{\mathcal{C}}^\alpha$ is given by $\langle \kappa_0, f\rangle = \kappa(f)$, where κ is the probability measure from Lemma 3.10 (ii).

The arguments for the rest of the section are similar to the ones made in chapter 4 of Bedrossian et al. (2022). Henceforth, let $\alpha \in (0, 1)$ be as in Lemma 3.10.

Theorem 3.12 *Let $x^* \in \mathcal{M}$ be regular with $\lambda(x^*) > 0$. There exist constants $p > 0$, $\gamma \in (0, 1)$ and $\alpha \in (0, 1)$ and a positive function $f^* \in \mathcal{C}^\alpha$ such that for all $s \in S^{N-1}$ we have*

$$\mathbb{E}\left[\left\|\Psi_\omega^{(1)} s\right\|^{-p} f^*\left(\frac{\Psi_\omega^{(1)} s}{\left\|\Psi_\omega^{(1)} s\right\|}\right)\right] = \gamma f^*(s). \tag{3.46}$$

Proof Note that (3.46) states that $\mathcal{P}_{-p} f^* = \gamma f^*$. In the following, we will show that $\frac{\mathrm{d}}{\mathrm{d}q} r(q)|_{q=0} = \lambda(x^*) > 0$. This will imply that for sufficiently small $p > 0$, we can set $\gamma := r(-p) < 1$ and $f^* := f_{-p}$. Since, for small p, the function f_{-p} is close[18] to $f_0 = \mathbf{1}$, it is indeed a positive function and by (3.45), we have $\mathcal{P}_{-p} f^* = \gamma f^*$.

In order to show $\frac{\mathrm{d}}{\mathrm{d}q} r(q)|_{q=0} = \lambda(x^*)$, note that for q sufficiently close to 0, we have $\langle \kappa_q, \mathbf{1}\rangle = \langle \kappa_q, f_0\rangle \neq 0$. By Corollary 3.11, this allows us to express $r(q)$ by

$$\log r(q) = \lim_{n\to\infty} \frac{1}{n} \log \|\mathcal{P}_q^n \mathbf{1}\|_{\mathcal{C}^\alpha}.$$

18. in the $\mathcal{C}^\alpha$-sense, but thus in particular in the $\mathcal{C}^0$-sense

Using Jensen's inequality, we can estimate

$$\begin{aligned}\log r(q) &= \lim_{n\to\infty} \frac{1}{n} \log \|\mathcal{P}_q^n \mathbf{1}\|_{\mathcal{C}^\alpha} \geq \lim_{n\to\infty} \frac{1}{n} \log \|\mathcal{P}_q \mathbf{1}\|_{\mathcal{C}^0} \\ &= \lim_{n\to\infty} \frac{1}{n} \log \left(\sup_{s\in S^{N-1}} \mathbb{E}\left[\left\| \Psi_\omega^{(1)} s \right\|^q \right] \right) \\ &= \lim_{n\to\infty} \frac{1}{n} \log \left(\sup_{s\in S^{N-1}} \mathbb{E}\left[e^{q \log \left\| \Psi_\omega^{(1)} s \right\|} \right] \right) \\ &\geq q \lim_{n\to\infty} \frac{1}{n} \sup_{s\in S^{N-1}} \mathbb{E}\left[\log \left\| \Psi_\omega^{(1)} s \right\| \right].\end{aligned}$$

By the Oseledec theorem (see Theorem 2.1) the limit in the last line is equal to $\lambda(x^*)$. Thus we get $r(q) \geq e^{q\lambda(x^*)}$, for sufficiently small q. Since we already know that r is differentiable in 0 from Corollary 3.11 and since $r(0) = 1$, this implies $\frac{\mathrm{d}}{\mathrm{d}q} r(q)|_{q=0} = \lambda(x^*)$, completing the proof. ■

We define a function $F^* : \mathbb{R}^N \setminus \{0\} \to \mathbb{R}_{>0}$ by

$$F^*(w) := \|w\|^{-p} f^*\left(\frac{w}{\|w\|} \right). \tag{3.47}$$

Corollary 3.13 *We have*

$$\mathbb{E}\left[F^*(\Psi_\omega^{(1)} w) \right] = \gamma F^*(w). \tag{3.48}$$

Proof This is a direct consequence of (3.46). Let $w \in \mathbb{R}^N \setminus \{0\}$ and set $s := \frac{w}{\|w\|} \in S^{N-1}$. Then

$$\begin{aligned}\mathbb{E}\left[F^*(\Psi_\omega^{(1)} w) \right] &= \mathbb{E}\left[\left\| \Psi_\omega^{(1)} w \right\|^{-p} f^*\left(\frac{\Psi_\omega^{(1)} w}{\|\Psi_\omega^{(1)} w\|} \right) \right] \\ &= \|w\|^{-p} \mathbb{E}\left[\left\| \Psi_\omega^{(1)} s \right\|^{-p} f^*\left(\frac{\Psi_\omega^{(1)} s}{\|\Psi_\omega^{(1)} s\|} \right) \right] \\ &= \|w\|^{-p} \gamma f^*(s) = \gamma F^*(w),\end{aligned}$$

showing the claim. ■

This establishes that F^* is a Lyapunov function for the linearized process induced by Ψ. For the remainder of this section, we will show that, in a neighborhood of the origin, F^* is also a Lyapunov function for the Markov process induced by the nonlinear cocycle ψ. Since f^* is continuous, positive, and has a compact domain, it is both bounded and bounded away from zero, that is, we can find constants $0 < C_- \leq C_+ < \infty$ such that

$$C_- \leq f^*(s) \leq C_+, \ \forall s \in S^{N-1}. \tag{3.49}$$

As a direct consequence, we also get the bound

$$C_- \|w\|^{-p} \leq F^*(w) \leq C_+ \|w\|^{-p}, \ \forall w \in \mathbb{R}^N \setminus \{0\}. \tag{3.50}$$

Lemma 3.14 *For every $\varepsilon > 0$, there exists a $\delta > 0$ such that for $w, \tilde{w} \in \mathbb{R}^N \setminus \{0\}$ we have*

$$\frac{\|w - \tilde{w}\|}{\|w\|} < \delta \ \Rightarrow \ \frac{|F^*(w) - F^*(\tilde{w})|}{F^*(w)} < \varepsilon. \tag{3.51}$$

Proof Using α-Hölder continuity of f^* and the bounds in (3.49), we can estimate

$$\begin{aligned}|F^*(w) - F^*(\tilde{w})| &= \left| \|w\|^{-p} f^*\left(\frac{w}{\|w\|}\right) - \|\tilde{w}\|^{-p} f^*\left(\frac{\tilde{w}}{\|\tilde{w}\|}\right)\right| \\ &\leq \|w\|^{-p}\left| f^*\left(\frac{w}{\|w\|}\right) - f^*\left(\frac{\tilde{w}}{\|\tilde{w}\|}\right)\right| + \left|\|w\|^{-p} - \|\tilde{w}\|^{-p}\right| f^*\left(\frac{\tilde{w}}{\|\tilde{w}\|}\right) \\ &\leq F^*(w) C_-^{-1} \|f^*\|_{\mathcal{C}^\alpha} \left\| \frac{w}{\|w\|} - \frac{\tilde{w}}{\|\tilde{w}\|}\right\|^\alpha + F^*(w) C_+ C_-^{-1} \left| 1 - \left(\frac{\|\tilde{w}\|}{\|w\|}\right)^{-p}\right|.\end{aligned}$$

In order to complete the proof, we will show that for each $\varepsilon > 0$ there exists a $\delta > 0$, such that for $w, \tilde{w} \in \mathbb{R}^N \setminus \{0\}$ with $\|w\|^{-1}\|w - \tilde{w}\| < \delta$ we have

$$\left\| \frac{w}{\|w\|} - \frac{\tilde{w}}{\|\tilde{w}\|}\right\| \leq \left(\frac{C_- \varepsilon}{\|f^*\|_{\mathcal{C}^\alpha}}\right)^{\frac{1}{\alpha}} =: \varepsilon_1 \qquad \text{and} \qquad \left| 1 - \left(\frac{\|\tilde{w}\|}{\|w\|}\right)^{-p}\right| \leq \frac{C_+ \varepsilon}{C_-} =: \varepsilon_2. \tag{3.52}$$

We can bound

$$\begin{aligned}\left\| \frac{w}{\|w\|} - \frac{\tilde{w}}{\|\tilde{w}\|}\right\| &= \|w\|^{-1} \left\| w - \tilde{w} + \left(1 - \frac{\|w\|}{\|\tilde{w}\|}\right)\tilde{w}\right\| \\ &\leq \frac{\|w - \tilde{w}\|}{\|w\|} + \frac{\left|\|w\| - \|\tilde{w}\|\right|}{\|w\|} \leq 2\frac{\|w - \tilde{w}\|}{\|w\|}\end{aligned}$$

to see that the first inequality in (3.52) is satisfied for $w, \tilde{w}$ with $\|w\|^{-1}\|w - \tilde{w}\| < \delta_1 := \frac{1}{2}\varepsilon_1$. Since the map $t \mapsto t^{-p}$ is continuous at $t = 1 \mapsto 1$, there also exists a $\delta_2 > 0$ such that

$$\left| 1 - \frac{\|\tilde{w}\|}{\|w\|}\right| \leq \frac{\|w - \tilde{w}\|}{\|w\|} < \delta_2 \;\Rightarrow\; \left| 1 - \left(\frac{\|\tilde{w}\|}{\|w\|}\right)^{-p}\right| < \varepsilon_2.$$

Thus (3.51) will be satisfied for $\delta = \min(\delta_1, \delta_2)$. ∎

Lemma 3.15 *For every $\varepsilon > 0$, there exists a $\delta > 0$ such that*

$$F^*\left(\psi_\omega^{(1)}(v, w)_2\right) \leq (1 + \varepsilon) F^*\left(\Psi_\omega^{(1)} w\right), \tag{3.53}$$

for all $\omega \in \Omega$ and $(v, w) \in V_\delta$, where V_δ is the neighborhood given in Lemma 3.2/Corollary 3.5.

Proof Let $\varepsilon > 0$ and choose $\tilde{\delta} > 0$ such that the conclusion (3.51) of the previous lemma holds. Recall that (see (3.28))

$$K_1 := \inf_{\omega \in \Omega} \left\| \left(\Psi_\omega^{(1)}\right)^{-1}\right\|^{-1} > 0$$

and let $\delta = K_1 \tilde{\delta}$. Then, by Lemma 3.2, we have

$$\left\| \Psi_\omega^{(1)} w - \psi_\omega^{(1)}(v, w)\right\| \leq \delta \|w\| \leq K_1^{-1} \delta \left\| \Psi_\omega^{(1)} w\right\| = \tilde{\delta} \left\| \Psi_\omega^{(1)} w\right\|,$$

for all $\omega \in \Omega$ and $(v, w) \in V_\delta$. By (3.51), this implies

$$\frac{\left| F^*\left(\Psi_\omega^{(1)} w\right) - F^*\left(\psi_\omega^{(1)}(v, w)\right)\right|}{F^*\left(\Psi_\omega^{(1)} w\right)} < \varepsilon,$$

which implies (3.53). ∎

Corollary 3.16 *For every $\varepsilon > 0$, there exists a $\delta > 0$ s.t. for all $\omega \in \Omega$ and all $(v, w) \in V_\delta \setminus (\mathbb{R}^{D-N} \times \{0\})$ we have $\psi_\omega^{(1)}(v, w)_2 \neq 0$ and*

$$\mathbb{E}\left[F^*\left(\psi_\omega^{(1)}(v, w)_2\right)\right] \leq (\gamma + \varepsilon)\, F^*(w).$$

In particular, if we choose $0 < \varepsilon < 1 - \gamma$ this shows that F^* is indeed a Lyapunov function on a neighborhood of the origin.

Proof First note that if we choose $\delta < K_1$, using Corollary 3.5, we get

$$\left\|\psi_\omega^{(1)}(v, w)_2\right\| \geq \left\|\Psi_\omega^{(1)} w\right\| - \left\|\Psi_\omega^{(1)} w - \psi_\omega^{(1)}(v, w)_2\right\| \geq K_1\|w\| - \delta\|w\| > 0$$

for all $\omega \in \Omega$ and all $(v, w) \in V_\delta \setminus (\mathbb{R}^{D-N} \times \{0\})$. Let $\varepsilon > 0$ and choose $\delta > 0$ small enough, such that both $\delta < K_1$ and that the conclusion of Lemma 3.15 holds. Together with Corollary 3.13 this yields

$$\mathbb{E}\left[F^*(\psi_\omega^{(1)}(v, w)_2)\right] \leq (1 + \varepsilon)\mathbb{E}\left[F^*(\Psi_\omega^{(1)} w)\right] = (1 + \varepsilon)\gamma F^*(w) \leq (\gamma + \varepsilon)F^*(w),$$

for all $\omega \in \Omega$ and all $(v, w) \in V_\delta \setminus (\mathbb{R}^{D-N} \times \{0\})$. ■

3.9 Stochastic Gradient Descent - the Unstable Case

Theorem B (ii) *Let $x^* \in \mathcal{M}$ be regular with $\lambda(x^*) > 0$. Then $x^* \notin \operatorname{supp}(X_{\lim}^{\mathrm{SGD}})$.*

Proof [of Theorem B (ii)] Suppose $x^* \in \mathcal{M}$ is regular with $\lambda(x^*) > 0$. In the following, we will show that there exists a $\delta > 0$ such that the neighborhood $0 \in V_\delta \subseteq \mathbb{R}^D$ from Lemma 3.2/Corollary 3.5 satisfies

$$\mathbb{P}\left(\exists n \in \mathbb{N},\ \text{s.t. } \psi_\omega^{(n)}(v, w) \notin V_\delta\right) = 1,\ \forall, (v, w) \in V_\delta \setminus \left(\mathbb{R}^{D-N} \times \{0\}\right). \tag{3.54}$$

We first argue why this is sufficient to show $x^* \notin \operatorname{supp}(X_{\lim}^{\mathrm{SGD}})$. Let $x^* \in U \subseteq \mathcal{M}$ be the neighborhood given by

$$U := \chi(V_\delta) \cap \mathcal{M} = \chi\left(V_\delta \cap \left(\mathbb{R}^{D-N} \times \{0\}\right)\right).$$

Suppose for now that ω is such that $X_{\lim}^{\mathrm{SGD}} \in U$. By the openness of $\chi(V_\delta)$, there then either exists an $n \in \mathbb{N}_0$ s.t. $X_n^{\mathrm{SGD}} \in U$ or there exists an $m \in \mathbb{N}_0$, such that $X_{m+n}^{\mathrm{SGD}} \in \chi(V_\delta) \setminus \mathcal{M}$ for all $n \in \mathbb{N}_0$. By Lemma 3.4, the former happens with probability zero. The probability for the latter to happen can be estimated by

$$\begin{aligned}
&\mathbb{P}\left(\exists\, m \in \mathbb{N}_0,\ \forall\, n \in \mathbb{N}_0,\ X_{m+n}^{\mathrm{SGD}}(\omega) \in \chi(V_\delta) \setminus \mathcal{M}\right) \\
&\leq \sum_{m=1}^{\infty} \mathbb{P}\left(\forall\, n \in \mathbb{N}_0,\ \varphi_{\theta^m\omega}^{(n)}(X_m^{\mathrm{SGD}}(\omega)) \in \chi(V_\delta) \setminus \mathcal{M}\right) \\
&= \sum_{m=1}^{\infty} \mathbb{P}\left(\forall\, n \in \mathbb{N}_0,\ \psi_{\theta^m\omega}^{(n)}(\chi^{-1}(X_m^{\mathrm{SGD}}(\omega))) \in V_\delta \setminus (\mathbb{R}^{D-N} \times \{0\})\right).
\end{aligned}$$

Since X_m^{SGD} only depends on ω_{init} and $\omega_1, \ldots, \omega_m$ and the random map $\psi_{\theta^m\omega}^n$ only depends on $\omega_{m+1}, \ldots \omega_{m+n}$, they are independent. Thus, if (3.54) holds, this probability will also be zero.

It remains to find a $\delta > 0$ such that (3.54) holds. Let $p > 0$, $\gamma \in (0,1)$ and $f^* : S^{N-1} \to (0,\infty)$ be such that the conclusion of Theorem 3.12 holds. Also, let F^* be the function defined in (3.47), choose some $0 < \varepsilon < 1-\gamma$ and let δ be such that the conclusion of Corollary 3.16 holds. Without loss of generality, we assume $V_\delta \subseteq \mathbb{R}^{D-N} \times \mathcal{B}_{\mathbb{R}^N}(1)$[19]. Recall that $\tau_{\delta,\omega} : V_\delta \to \mathbb{N} \cup \{\infty\}$ is given by

$$\tau_{\delta,\omega}(v,w) := \inf\left\{n \in \mathbb{N} : \psi_\omega^{(n)}(v,w) \notin V_\delta\right\}.$$

We can reformulate (3.54) as

$$\mathbb{P}\left(\tau_{\delta,\omega}(v,w) < \infty\right) = 1,\ \forall (v,w) \in V_\delta \setminus \left(\mathbb{R}^{D-N} \times \{0\}\right).$$

Using Corollary 3.16 inductively, one can show that

$$\mathbb{E}\left[\mathbb{1}_{\tau_{\delta,\omega}(v,w)\geq n} F^*\left(\psi_\omega^{(n)}(v,w)_2\right)\right] \leq (\gamma+\varepsilon)^n F^*(w),$$

for all $n \in \mathbb{N}$ and $(v,w) \in V_\delta$. Since $V_\delta \subseteq \mathbb{R}^{D-N} \times \mathcal{B}_{\mathbb{R}^N}(1)$, we have

$$F^*(w) \geq C_- \|w\|^{-p} \geq C_-,\ \forall (v,w) \in V_\delta.$$

This allows us to compute

$$\begin{aligned}
\mathbb{P}(\tau_{\delta,\omega}(v,w) = \infty) &= \lim_{n\to\infty} \mathbb{P}(\tau_{\delta,\omega}(v,w) > n) \\
&\leq C_-^{-1} \lim_{n\to\infty} \mathbb{E}\left[\mathbb{1}_{\tau_{\delta,\omega}(v,w)>n} F^*\left(\psi_\omega^{(n)}(v,w)_2\right)\right] \\
&\leq C_-^{-1} \lim_{n\to\infty} (\gamma+\varepsilon)^n F^*(w) = 0,
\end{aligned}$$

for all $(v,w) \in V_\delta \setminus (\mathbb{R}^{D-N} \times \{0\})$. ■

Acknowledgements

The authors would like to thank Tim van Erven for his insightful feedback. Furthermore, the authors thank the DFG SPP 2298 for supporting their research. Both authors have been additionally supported by Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689), in the case of D.C. via the Berlin Mathematical School and in the case of M.E. via projects AA1-8, AA1-18 and EF45-5. Furthermore, M.E. thanks the DFG CRC 1114, the Einstein Foundation and the Dutch Research Council NWO (VI.Vidi.233.133) for support.

Appendix A. Comparison to second moment linear stability

A different notion of linear stability for SGD was introduced by Wu et al. (2018). Adapted to the notation introduced in Section 3.5, their condition can be expressed as follows.

Definition A.1 (Definition 2 in Wu et al., 2018) *A global minimum $x^* \in \mathcal{M}$ is called* second moment linearly stable*, if there exists a constant C such that*

$$\mathbb{E}\left[\left\|\Phi_\omega^{(n)} x\right\|^2\right] \leq C\|x\|^2,$$

19. If not, consider the intersection of V_δ and $\mathbb{R}^{D-N} \times \mathcal{B}_{\mathbb{R}^N}(1)$ henceforth.

for all $x \in \mathbb{R}^D$ and all $n \in \mathbb{N}$. Here $\Phi_\omega^{(n)}$ denotes the linearization around x^* of n steps of SGD with seed ω as defined in (3.19).

In contrast, we call a global minimum x^* with $\lambda(x^*) < 0$ *almost surely stable*. As will be argued below, second moment linear stability is almost a strictly stronger condition then almost sure stability. The fact that second moment stability and almost-sure stability are in general not equivalent, can already be seen for 1-dimensional linear stochastic processes. Let $(X_n)_{n\in\mathbb{N}_0}$ be the stochastic process given by $X_0 = 1$ and

$$X_{n+1} = Y_{n+1} X_n,$$

where $(Y_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence of real-valued random variables with $\mathbb{E}[\log_+ |Y_1|] < \infty$.[20] By the strong law of large numbers, we have

$$\lim_{n\to\infty} \frac{1}{n} \log |X_n| = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log |Y_n| = \mathbb{E}[\log |Y_1|] \in [-\infty, \infty), \text{ almost surely.}$$

Thus, if $\mathbb{E}[\log |Y_1|] < 0$, then the linear process (X_n) is almost-surely stable, that is, $X_n \to 0$ with probability 1. At the same time, since X_n and Y_{n+1} are independent we have

$$\mathbb{E}\left[|X_{n+1}|^2\right] = \mathbb{E}\left[|Y_{n+1}|^2\right] \mathbb{E}\left[|X_n|^2\right] = \mathbb{E}\left[|Y_1|^2\right] \mathbb{E}\left[|X_n|^2\right]$$

and therefore $\mathbb{E}\left[|X_n|^2\right] = \mathbb{E}\left[|Y_1|^2\right]^n$. Thus (X_n) is second moment stable, that is, there exists a $C > 0$ such that $\mathbb{E}\left[|X_n|^2\right] < C$, if and only if $\mathbb{E}\left[|Y_1|^2\right] \leq 1$. However, while $\mathbb{E}\left[|Y_1|^2\right] \leq 1$ implies $\mathbb{E}[\log |Y_1|] < 0$ by Jensen's inequality, the converse is clearly not true.[21]

Second moment stability is closely related to so-called moment Lyapunov exponents (see, e.g., Arnold, 1984), which have been linked to central limit theorems as well as large deviation theory for the convergence of finite-time Lyapunov exponents (see, e.g., Bougerol and Lacroix, 1985; Arnold et al., 1986; Arnold and Kliemann, 1987).

Definition A.2 *For any $p \in \mathbb{R}$, the p-th moment Lyapunov exponent $\Lambda_p(x^*)$ of a global minimum $x^* \in \mathcal{M}$ is given by*

$$\Lambda_p(x^*) := \sup_{w\in\mathbb{R}^N} \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\left\|\Psi_\omega^{(n)} w\right\|^p\right].$$

Convergence for each $w \in \mathbb{R}^N$ follows from a subadditivity argument. The precise relation between second moment Lyapunov exponents and second order stability can be stated as follows.

Proposition A.3 *In order for a global minimum $x^* \in \mathcal{M}$ to be second moment linearly stable, it is*

(i) necessary to have $\Lambda_2 \leq 0$ and

(ii) sufficient to have $\Lambda_2 < 0$.

20. Here $\log_+ |Y_1| = \max(0, \log |Y_1|)$.

21. Suppose for example that Y_1 takes the values $\frac{1}{3}$ and 2 with probability $\frac{1}{2}$ each. A simple calculation shows $\mathbb{E}[\log |Y_1|] = \frac{1}{2}(\log(2) - \log(3)) < 0$, while $\mathbb{E}\left[|Y_1|^2\right] = \frac{19}{9} > 1$.

Proof Note that by (3.22), a global minimum x^* is second moment linearly stable if and only if there exists a $C > 0$ such that

$$\mathbb{E}\left[\left\|\Psi_\omega^{(n)} w\right\|^2\right] \leq C\|w\|^2,$$

for all $w \in \mathbb{R}^N$ and all $n \in \mathbb{N}$. Now (i) follows directly from the definition of the second moment Lyapunov exponent. In order to prove (ii), suppose $\Lambda_{2,w} < 0$ for all $w \in \mathbb{R}^N$ and thus in particular for the unit vectors $e_1, \ldots, e_N$. Thus

$$\lim_{n\to\infty} \mathbb{E}\left[\left\|\Psi_\omega^{(n)} e_i\right\|^2\right] = 0, \ \forall\, i \in [N].$$

Let $C > 0$ be given by

$$C = \sup_{i\in[N],\, n\in\mathbb{N}} \mathbb{E}\left[\left\|\Psi_\omega^{(n)} e_i\right\|^2\right] < \infty.$$

Now for $w = w_1 e_1 + \cdots + w_N e_N$, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\Psi_\omega^{(n)} w\right\|^2\right] &= \mathbb{E}\left[\left\|\Psi_\omega^{(n)}\left(\sum_{i=1}^N w_i e_i\right)\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^N w_i \Psi_\omega^{(n)} e_i\right\|^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^N w_i \left\|\Psi_\omega^{(n)} e_i\right\|\right)^2\right] \leq \mathbb{E}\left[N \sum_{i=1}^N w_i^2 \left\|\Psi_\omega^{(n)} e_i\right\|^2\right] \\
&\leq NC \sum_{i=1}^N w_i^2 = NC\|w\|^2.
\end{aligned}$$

■

Proposition A.4 *For each $p \in \mathbb{R}$, we have $\Lambda_p(x^*) \geq p\lambda(x^*)$. In particular $\Lambda_2(x^*) \leq 0$ implies $\lambda(x^*) \leq 0$.*

Proof As a consequence of Oseledec theorem (see Theorem 2.1), we have

$$\lambda(x^*) = \sup_{w\in\mathbb{R}^N} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\log\left\|\Psi_\omega^{(n)} w\right\|\right].$$

The proposition follows from Jensen's inequality. ■

To summarize, if $x^* \in \mathcal{M}$ is second moment stable in the sense of Wu et al. (2018), it satisfies $\Lambda_2(x^*) \leq 0$ and thus $\lambda(x^*) \leq 0$. Most of these points should be regular in the sense of Definition 2.2 and even satisfy $\lambda(x^*) < 0$. By Theorem B, these global minima lie in the support of $X_{\lim}^{\mathrm{SGD}}$. On the other hand the one dimensional example demonstrated above suggests that it is possible to have global minima $x^* \in \mathcal{M}$ with $\lambda(x^*) < 0$ and $\Lambda_2(x^*) > 0$. Given that they are regular, these global minima will be in the support of $X_{\lim}^{\mathrm{SGD}}$, but will not be second moment linearly stable in the sense of Wu et al. (2018). Notably the paper by Wu et al. (2018) contains a sufficient condition for second order linear stability (see Therorem 1 in Wu et al., 2018). By the argument above, their condition is also a sufficient condition for $\lambda(x^*) \leq 0$.

In Ma and Ying (2021), a more general notion of "k-th order linear stability" is introduced. If $p = k$ is an even natural number this is equivalent to the following generalization of second moment linear stability (see Remark 1 in Ma and Ying, 2021).

Definition A.5 *For $p > 0$, a global minimum $x^* \in \mathcal{M}$ is called p-th moment linearly stable, if there exists a constant C such that*

$$\mathbb{E}\left[\left\|\Phi_\omega^{(n)} x\right\|^p\right] \leq C\|x\|^p,$$

for all $x \in \mathbb{R}^D$ and all $n \in \mathbb{N}$.

Proposition A.3 can be extended to p-th moment linear stability mutatis mutandis.

References

A. Andreyev and P. Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd, 2025. Available electronically via: `https://arxiv.org/abs/2412.20553`.

L. Arnold. A formula connecting sample and moment stability of linear stochastic systems. *SIAM Journal on Applied Mathematics*, 44(4):793–802, 1984. Available electronically via: `https://www.jstor.org/stable/2101254`.

L. Arnold. *Random dynamical systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 1998. doi: `10.1007/978-3-662-12878-7`.

L. Arnold and W. Kliemann. Large deviations of linear stochastic differential equations. In *Stochastic Differential Systems*, pages 115–151. Springer Berlin Heidelberg, 1987. doi: `10.1007/BFb0038929`.

L. Arnold, W. Kliemann, and E. Oeljeklaus. Lyapunov exponents of linear stochastic systems. In L. Arnold and V. Wihstutz, editors, *Lyapunov Exponents*. Springer Berlin Heidelberg, 1986. doi: `10.1007/BFb0076835`.

S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 948–1024. PMLR, 2022. Available electronically via: `https://proceedings.mlr.press/v162/arora22a`.

A. Avila, A. Eskin, and M. Viana. Continuity of the Lyapunov exponents of random matrix products, 2023. Available electronically via: `https://arxiv.org/abs/2305.06009`.

P. H. Baxendale. Statistical equilibrium and two-point motion for a stochastic flow of diffeomorphisms. In *Spatial stochastic processes*, volume 19 of *Progr. Probab.*, pages 189–218. Birkhäuser Boston, Boston, MA, 1991. doi: `10.1007/978-1-4612-0451-0_9`.

P. H. Baxendale and D. W. Stroock. Large deviations and stochastic flows of diffeomorphisms. *Probab. Theory Related Fields*, 80(2):169–215, 1988. doi: `10.1007/BF00356102`.

J. Bedrossian, A. Blumenthal, and S. Punshon-Smith. Almost-sure exponential mixing of passive scalars by the stochastic Navier-Stokes equations. *Ann. Probab.*, 50(1):241–303, 2022. doi: `10.1214/21-AOP1533`.

J. Bedrossian, A. Blumenthal, and S. Punshon-Smith. Non-uniqueness of stationary measures for stochastic systems with almost surely invariant manifolds, 2025. Available electronically via: `https://arxiv.org/abs/2505.22903`.

M. Benaïm and T. Hurth. *Markov chains on metric spaces—a short course*. Springer, Cham, 2022. doi: 10.1007/978-3-031-11822-7.

A. Blumenthal, M. Coti Zelati, and R. S. Gvalani. Exponential mixing for random dynamical systems and an example of Pierrehumbert. *Ann. Probab.*, 51(4):1559–1601, 2023. doi: 10.1214/23-AOP1627.

P. Bougerol and J. Lacroix. *Products of random matrices with applications to Schrödinger operators*, volume 8 of *Progress in Probability and Statistics*. Birkhäuser Boston, Inc., Boston, MA, 1985. doi: 10.1007/978-1-4684-9172-2.

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*, 2021. Available electronically via: https://arxiv.org/abs/2103.00065.

Y. Cooper. Global minima of overparameterized neural networks. *SIAM J. Math. Data Sci.*, 3 (2):676–691, 2021. doi: 10.1137/19M1308943.

M. Coti Zelati and M. Hairer. A noise-induced transition in the Lorenz system. *Comm. Math. Phys.*, 383(3):2243–2274, 2021. doi: 10.1007/s00220-021-04000-6.

J.-P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Modern Phys.*, 57(3):617–656, 1985. doi: 10.1007/978-0-387-21830-4_17.

B. Fehrman, B. Gess, and A. Jentzen. Convergence rates for the stochastic gradient descent method for non-convex objective functions. *Journal of Machine Learning Research*, 21(136): 1–48, 2020. Available electronically via: https://jmlr.org/papers/v21/19-636.html.

B. Gess, S. Kassing, and V. Konarovskyi. Stochastic modified flows, mean-field limits and dynamics of stochastic gradient descent. *Journal of Machine Learning Research*, 25(30):1–27, 2024. Available electronically via: https://www.jmlr.org/papers/v25/23-0220.html.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2017. Available electronically via: https://arxiv.org/abs/1706.02677.

M. Gurbuzbalaban, U. Simsekli, and L. Zhu. The heavy-tail phenomenon in sgd. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3964–3975. PMLR, 2021. Available electronically via: https://proceedings.mlr.press/v139/gurbuzbalaban21a.

M. Hairer. Convergence of Markov processes, 2021. Available electronically via: https://hairer.org/notes/Convergence.pdf.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9:1–42, 1997. doi: 10.1162/neco.1997.9.1.1.

L. Hodgkinson and M. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4262–4274. PMLR, 2021. Available electronically via: https://proceedings.mlr.press/v139/hodgkinson21a.

E. Hoffer, I. Hubara, and D. Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. Available electronically via: https://proceedings.neurips.cc/paper_files/paper/2017/hash/a5e0ff62be0b08456fc7f1e88812af3d-Abstract.html.

L. Hörmander. Hypoelliptic second order differential equations. *Acta Math.*, 119:147–171, 1967. doi: 10.1007/BF02392081.

K. Hornik, M. B. Stinchcombe, and H. L. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. Available electronically via: https://proceedings.neurips.cc/paper_files/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. doi: 10.1007/978-3-642-66282-9.

A. Katok and B. Hasselblatt. *Introduction to the modern theory of dynamical systems*, volume 54 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1995.

N. S. Keskar and R. Socher. Improving generalization performance by switching from adam to sgd, 2017. Available electronically via: https://arxiv.org/abs/1712.07628.

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. Available electronically via: https://arxiv.org/abs/1609.04836.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2014. Available electronically via: https://arxiv.org/abs/1412.6980.

J. F. C. Kingman. The ergodic theory of subadditive stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):499–510, 1968. Available electronically via: https://www.jstor.org/stable/2984253.

M. G. Krein and M. A. Rutman. Linear operators leaving invariant a cone in a Banach space. *Amer. Math. Soc. Translation*, 1950(26):128, 1950.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. Available electronically via: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

E. Le Page. Théorèmes limites pour les produits de matrices aléatoires. In *Probability measures on groups (Oberwolfach, 1981)*, volume 928, pages 258–303. Springer, Berlin-New York, 1982. Available electronically via: https://www.numdam.org/item/PSMIR_1980___1_A4_0.pdf (French).

Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2101–2110. PMLR, 2017. Available electronically via: https://proceedings.mlr.press/v70/li17f.html.

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, 2019. Available electronically via: https://jmlr.org/papers/v20/17-526.html.

P.-D. Liu and M. Qian. *Smooth ergodic theory of random dynamical systems*, volume 1606. Springer-Verlag, Berlin, 1995. doi: 10.1007/BFb0094308.

C. Ma and L. Ying. On linear stability of SGD and input-smoothness of neural networks. In *Advances in Neural Information Processing Systems*, 2021. Available electronically via: https://papers.nips.cc/paper_files/paper/2021/hash/8c26d2fad09dc76f3ff36b6ea752b0e1-Abstract.html.

J. W. Milnor. *Topology from the differentiable viewpoint*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997.

V. I. Oseledec. A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. *Trudy Moskov. Mat. Obšč.*, 19:179–210, 1968.

M. Qian, J.-S. Xie, and S. Zhu. *Smooth ergodic theory for endomorphisms*, volume 1978. Springer-Verlag, Berlin, 2009. doi: 10.1007/978-3-642-01954-8.

D. Ruelle and M. Shub. Stable manifolds for maps. In *Global theory of dynamical systems (Proc. Internat. Conf., Northwestern Univ., Evanston, Ill., 1979)*, volume 819, pages 389–392. Springer, Berlin, 1980. doi: 10.1007/BFb0087002.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0.

M. Sandri. Numerical calculation of lyapunov exponents. *Anthology The Mathematica Journal Year*, 6(3):78–84, 1996.

R. M. Schmidt, F. Schneider, and P. Hennig. Descending through a crowded valley - benchmarking deep learning optimizers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 9367–9376. PMLR, 2021. Available electronically via: https://proceedings.mlr.press/v139/schmidt21a.

J. M. Steele. Kingman's subadditive ergodic theorem. *Ann. Inst. H. Poincaré Probab. Statist.*, 25 (1):93–98, 1989. Available electronically via: https://www.numdam.org/item/AIHPB_1989__25_1_93_0.pdf.

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. Curran Associates, Inc., 2018. Available electronically via: http://proceedings.neurips.cc/paper_files/paper/2018/hash/6651526b6fb8f29a00507de6a49ce30f-Abstract.html.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. doi: 10.1145/3446776.