Is Stochastic Gradient Descent Effective? A PDE Perspective on Machine Learning processes

Davide Barbieri,[a,b] Matteo Bonforte,[a,c] Peio Ibarrondo[a,d]

Abstract

In this paper we analyze the behaviour of the stochastic gradient descent (SGD), a widely used method in supervised learning for optimizing neural network weights via a minimization of non-convex loss functions. Since the pioneering work of Li, Tai and E (2017), the underlying structure of such processes can be understood via parabolic PDEs of Fokker-Planck type, which are at the core of our analysis. Even if Fokker-Planck equations have a long history and a extensive literature, almost nothing is known when the potential is non-convex or when the diffusion matrix is (very) degenerate, and this is the main difficulty that we face in our analysis. This affects the long-time behaviour of solutions, a crucial point in understanding deep characteristics of the associated learning process.

Our main contribution is identifying two different regimes in the learning process. In the initial phase of SGD, the loss function drives the weights to concentrate around the nearest local minimum, which may not necessarily be optimal. We refer to this phase as the *drift regime* and we provide quantitative estimates that shed light on this concentration phenomenon. Next, we introduce the *diffusion regime*, which typically happens after some time, where stochastic fluctuations help the learning process to diffuse and so escape suboptimal local minima. We analyze the "Mean Exit Time" (MET), i.e. the time needed to escape a local minimum, and prove precise upper and lower bounds of the MET. Finally, we address the asymptotic convergence of SGD, tackling the complexities of non-convex cost functions together with the degeneracies in the diffusion matrix, that do not allow to use the standard approaches, and require new techniques. For this purpose, we exploit two different methods: duality and entropy methods.

This work provides new results about the dynamics and effectiveness of SGD, offering a deep connection between stochastic optimization and PDE theory. It also provides some answers and insights to basic questions in the Machine Learning processes: How long does SGD take to escape from a local minimum? Do neural network parameters converge using SGD? How do parameters evolve in the first stage of training with SGD?

Keywords: stochastic gradient descent, supervised learning, mass concentration, mean exit time, asymptotic behaviour, entropy method.
MSC2020 Classification. Pri: 35Q68, 68T07, 35K65, 35B40. Sec: 60J60, 35J70, 35K15

(a) Departamento de Matemáticas, Universidad Autónoma de Madrid,
ICMAT - Instituto de Ciencias Matemáticas, CSIC-UAM-UC3M-UCM,
Campus de Cantoblanco, 28049 Madrid, Spain.
(b) E-mail: `davide.barbieri@uam.es` Web-page: `http://www.uam.es/davide.barbieri`
(c) E-mail: `matteo.bonforte@uam.es` Web-page: `http://verso.mat.uam.es/~matteo.bonforte`
(d) E-mail: `peio.ibarrondo@uam.es`;

Contents

Acknowledgments

D.B., P.I. were partially supported by the Projects PID2022-142202NB-I00 and PID2019-105599GB-I0 (Spanish Ministry of Science and Innovation). M.B., P.I. were partially supported by the Projects MTM2017-85757-P, PID2020-113596GB-I00 and PID2023-150166NB-I00 (Spanish Ministry of Science and Innovation). D.B., M.B., P.I. acknowledge the financial support from the Spanish Ministry of Science and Innovation, through the "Severo Ochoa Programme for Centres of Excellence in R&D" (CEX2019-000904-S and CEX2023-001347-S) and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 777822. P.I. was partially funded by the FPU-grant FPU19/04791 from the Spanish Ministry of Universities.

1 Introduction

In recent decades, Machine Learning has gained significant attention across various scientific communities, thanks to its practical utility and wide range of applications. One of the main challenges in Machine Learning is choosing the weights or parameters of a neural network to minimize a non-convex loss function based on a given data set. This is a non-convex optimization problem formulated as follows:

$$\arg\min_{\theta\in\mathbb{R}^d} L(\theta) := \frac{1}{N}\sum_{i=1}^{N} L_i(\theta) \tag{1}$$

where $L, L_i : \mathbb{R}^d \to \mathbb{R}_+$ for $i = 1, \dots, N$ and $\theta \in \mathbb{R}^d$ are the parameters of the model. The function L represents the total loss function of the given data set with N samples, while L_i represents the loss associated to the i^{th} training sample. A classical approach to this problem is to perform a gradient descent, where given some initial weights $\theta_0 \in \mathbb{R}^d$ and a constant learning rate $\eta > 0$ the weights at step $n \geq 0$ are updated by

$$\theta_{n+1} = \theta_n - \eta\nabla L(\theta_n) \qquad \forall n \geq 0\,. \tag{GD}$$

This learning algorithm is computationally too expensive to implement since at each step we have to compute N gradients of loss functions L_i. Moreover, this method may lead the parameters to a local minimum of L instead of reaching the global minimum. In order to avoid these problems, the use of the stochastic gradient descent (SGD) has proven to be effective. Defining $\{\gamma_n\}_{n\geq 1}$ as i.i.d. uniform random variables with values on $\{1, \dots, N\}$, the SGD algorithm reads

$$\theta_{n+1} = \theta_n - \eta\nabla L_{\gamma_n}(\theta_n) \qquad \forall n \geq 0\,. \tag{SGD}$$

Implementing this method yields very good results in practice, but the mathematical theory behind remains poorly understood, see [20] for an overview. The goal of this manuscript is to shed light on this learning process and to present new research directions in this field.

In 2017, Li, Tai and E opened a new framework for analyzing the SGD learning process in [44], which they examined in more detail in [45]. They proved that the (SGD) iteration is the Euler-Maruyama discretization of an associated stochastic differential equation. See Section 1.2 for a discussion of this model and related work.

This manuscript is devoted to the study of the continuous counterpart of the learning process described by (SGD), as defined by the corresponding drift-diffusion PDE for the transition probability. For this purpose, we will examine the interplay between the drift term governed by the loss function and the degenerate diffusion provided by the randomness in (SGD). In the initial stages of the learning process, the dynamics of the parameters primarily follows the drift, tending to concentrate around the local minima of the loss function L. Nevertheless, under mild condition on the degeneracy of the diffusion coefficients, we estimate the training duration required for the neural network to escape from non-optimal local minima. In this manuscript, we present a quantitative analysis of the two regimes of motion in the (SGD): the drift regime in Section 2 and the diffusion regime in Section 3.

Additionally, in Section 4, we address the problem of the existence of steady states for the parameters distributions and the asymptotic convergence towards them. As a consequence of the high degeneracy of this problem, the differential operator describing the evolution of the transition probability resembles an intermediate case between a transport equation and a diffusion equation, see Section 4.4. By means of simple yet representative examples we analyze the extreme cases to get an idea of the possible scenarios. We provide two types of results. On one hand, we introduce a new variant of the learning process (SGD) and exploit the recent results by Porretta [58] about convergence to steady states in non-degenerate Fokker-Planck equations. This will also allow us to prove existence of steady measures for the general case. On the other hand, we review some results from Arnold and Erb [5] based on the Bakry-Émery entropy method, and see when they can apply to our situation. These results hold only for constant diffusion matrices, hence representing a local scenario around a minimum of the cost function. We provide a number of open questions and directions for future research.

Beside the interest in the Machine Learning applications, which has motivated this work, our analysis could be also of interest in the study of stochastic differential equations and their long time behaviour.

1.1 Main results

For more generality, in this paper we will consider the mini-batch Stochastic Gradient Descent, a variant of (SGD). Instead of selecting randomly one sample in each step, we choose randomly a batch of samples B_n of size $\mathscr{b} \geq 1$. Namely, the stochastic gradient descent with constant learning rate $\eta > 0$ and batch size $|B_n| = \mathscr{b} \geq 1$ reads as follows

$$\theta_{n+1} = \theta_n - \frac{\eta}{\mathscr{b}} \sum_{b_i \in B_n} \nabla L_{b_i}(\theta_n) \qquad \forall n \geq 0\,, \tag{SGD}$$

with $B_n = \{b_i\}_{i=1}^{\mathscr{b}} \subset \{1, 2, \ldots, N\}$ being a batch of indexes chosen with uniform distribution in each step n. In [45] and [26], it is deduced that we can approximate (SGD) with the continuous stochastic process defined by

$$\mathrm{d}X_t = -\nabla L(X_t)\,\mathrm{d}t + \sqrt{\frac{\eta}{\mathscr{b}} Q(X_t)}\,\mathrm{d}W_t\,, \tag{2}$$

with $Q(x) = \mathrm{cov}\,[\nabla L_{\gamma_n}(x)] = \frac{1}{N}\sum_{i=1}^{N} \nabla L_i(x) \otimes \nabla L_i(x) - \nabla L(x) \otimes \nabla L(x)$ being a non-negative matrix for any $x \in \mathbb{R}^d$. Observe that Q is generically not invertible. Indeed, if we write

$$Q = T^* T \qquad \text{with} \qquad T = \begin{pmatrix} \frac{1}{N}\nabla L_1 - \nabla L \\ \vdots \\ \frac{1}{N}\nabla L_N - \nabla L \end{pmatrix}\,,$$

then the rows of T sum zero at every $x \in \mathbb{R}^d$. Specifically, in the common case of overparametrization [15], we have that $Q(x) \in \mathbb{R}^{d\times d}$ satisfies the rank condition $\mathrm{rank}(Q) \leq N - 1 < d$. As a consequence, the noise in (2) is degenerate.

In addition, the ratio between the learning rate and the mini-batch is usually considered very small, which diminishes the noise. We call this ratio the *effective learning rate* and we denote it by

$$\varepsilon^2 := \frac{\eta}{2\mathscr{b}}\,.$$

One of the advantages of the continuous approximation (2) is that we can use the theory of PDEs to analyze the behaviour of the transition probability and its evolution in time. More specifically, the transition probability $\rho(t,x)$ associated to the process in (2) with an initial distribution $\rho_0(x)$ satisfies the following Fokker-Planck type equation:

$$\begin{cases} \partial_t \rho & = \nabla \cdot \Big(\varepsilon^2 \nabla \cdot (Q(x)\rho) + \rho \nabla L(x) \Big) \qquad \text{in} \qquad (0,\infty) \times \mathbb{R}^d\,, \\ \rho(0,x) & = \rho_0(x) \qquad\qquad\qquad\qquad\qquad\quad\;\; \text{in} \qquad \mathbb{R}^d\,, \end{cases} \tag{3}$$

where $\nabla \cdot A$, with A being a matrix, denotes the divergence taken columnwise. For the deduction of this equation from Ito's formula see [60, Section 4.5] and for the existence of solutions under suitable conditions on Q and L see [19]. The class of solutions that we consider lie on the probability space $\mathcal{P}_k(\mathbb{R}^d)$ of nonnegative normalized Radon measures with $k \geq 0$ finite moments.

Definition 1.1 (Weak solutions.)**.** We say that $\rho \in C\left([0,T), \mathcal{P}_k(\mathbb{R}^d)\right)$ is a weak solution of (3) starting from $\rho_0 \in \mathcal{P}_k(\mathbb{R}^d)$ if $\forall \varphi \in C^{1,2}\left([0,T] \times \mathbb{R}^d\right)$ with polynomial growth of order k in the spatial variable, it holds that

$$\begin{aligned} &\int_{\mathbb{R}^d} \varphi(T,x)\rho(T,x)\,\mathrm{d}x - \int_{\mathbb{R}^d} \varphi(0,x)\rho_0(x)\,\mathrm{d}x - \int_0^T \int_{\mathbb{R}^d} \partial_t \varphi(t,x)\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \int_0^T \int_{\mathbb{R}^d} \Big[\varepsilon^2 \mathrm{tr}\left(Q(x) D^2 \varphi(t,x) \right) - \nabla L(x) \cdot \nabla \varphi(t,x) \Big] \rho(t,x)\,\mathrm{d}x\,\mathrm{d}t \end{aligned} \tag{4}$$

The main challenge of equation (3) is the degenerate diffusion provided by the non-negative matrix Q. This equation is somehow an intermediate case between a diffusive Fokker-Planck equation and a pure transport equation. In the first case, when Q is uniformly elliptic, the diffusion regularizes and asymptotic convergence towards the unique steady state holds whenever L grows at infinity. However, in the latter case, where the diffusion matrix Q is zero, the mass tends to concentrate around critical points of L. Specifically, the solution will tend to a sum of Dirac's deltas weighted according to the mass distribution of the initial datum.

We address this problem by focusing on three main questions outlined below. The first question concerns the initial stage of the learning process described in (SGD):

Q1: How do parameters evolve in the first stage of training with SGD?

Due to the smallness of the effective learning rate $\varepsilon > 0$, we expect that the drift term, given by ∇L, will lead the dynamics of the parameters for small times. As detailed in Section 2, if the loss function is λ-convex with $\lambda > 0$ in a ball centered at a local minimum $x_0 \in \mathbb{R}^d$, then we aim to obtain that

$$\int_{B_{R(t)}(x_0)} \rho(t,x)\,\mathrm{d}x \geq \int_{B_{R_0}(x_0)} \rho_0(x)\,\mathrm{d}x - c(\varepsilon,t) \qquad \forall 0 < t < T_\varepsilon\,,$$

where $R(t)$ is a decreasing radius and $0 \leq c(\varepsilon,t) \xrightarrow{\varepsilon\to 0} 0$. In order to give a precise statement, we need to introduce the correct speed of decreasing of $R(t)$ and a suitable approximation of the characteristic function of a ball.

Theorem 1.2 (Local mass concentration)**.** *Assume that L is λ-convex in $B_{(1+\delta)R_0}(0)$ with a minimum at 0 and $\lambda > 0$. Let ρ be a weak solution of (3) with $0 \le Q(x) \le \sigma I_{d\times d}$ for every $x \in B_{(1+\delta)R_0}(0)$. Let us consider $\varphi(t,r) : [t_0,\infty) \times \mathbb{R}_+ \to \mathbb{R}_+$ the smooth cut-off function (see (24)) such that*

$$\varphi(t,r) \equiv 1 \quad \textit{if} \quad r \le R_0\, e^{-\frac{\lambda}{2}(t-t_0)} \qquad \textit{and} \qquad \varphi(t,r) \equiv 0 \quad \textit{if} \quad r > (1+\delta)R_0\, e^{-\frac{\lambda}{2}(t-t_0)}\,.$$

Then, given any $a > 0$ and $\alpha, \beta \in (0,1)$, there exists an $\varepsilon_0 > 0$ such that for every $\varepsilon \in (0, \varepsilon_0)$ it holds that

$$\int_{\mathbb{R}^d} \varphi(t,|x|)\rho(t,x)\,\mathrm{d}x \ge \int_{\mathbb{R}^d} \varphi(t_0,|x|)\rho(t_0,x)\,\mathrm{d}x - \beta\,.$$

for every $t_0 < t < T_\varepsilon + t_0$ with $T_\varepsilon := \frac{2}{\lambda}\log\left(\frac{R_0}{a\varepsilon^\alpha}\right)$.

The result above highlights the strong dependence on the initial parameter distribution during the first stage of the (SGD). Actually, it indicates that the mass of ρ_0 within the basins of attraction of L tends to stay concentrated around the local minima for a while.

Once we know that if the effective learning rate is small enough there is a concentration phenomenon around local minima, the next natural question is the following:

Q2: How long does SGD take to escape from a local minimum?

For this purpose, we approach the issue from the perspective of the Mean Exit Time (MET) problem associated to (2). Given a domain $\Omega \subset \mathbb{R}^d$, the first exit time for the stochastic process X_t starting at $x \in \Omega$ is given by

$$\tau_\Omega^x = \inf\{t > 0 : X_t \notin \Omega, X_0 = x\}\,.$$

The MET is the function

$$u(x) = \mathbb{E}[\tau_\Omega^x]\,,$$

which satisfies the following elliptic equation [57, Chapter 7.2]

$$\begin{cases} -\mathcal{A}u(x) = 1, & \text{in } \Omega \\ u(x) = 0, & \text{on } \partial\Omega \end{cases} \tag{5}$$

where

$$\mathcal{A}u(x) = \varepsilon^2 \mathrm{tr}\left(Q(x)D^2u(x)\right) - \nabla L(x)\cdot\nabla u(x).$$

We devote Appendices B and C to recall the connection between MET and PDEs and other classical results in the case of isotropic and non-degenerate diffusion, leading to the well-known Kramers' Law [13, 14, 40].

In order to cope with the diffusion matrix Q in (2), which is far from being uniformly elliptic, we provide new estimates for the MET associated to our problem. We begin with a lower bound.

Theorem 1.3 (Lower bound for MET)**.** *Assume $0 \le Q(x) \le \sigma I_{d\times d}$, L is λ-convex in $B_{R_0}(x_0)$ and let $0 < r \le R_0$. Let $x \in B_r(x_0)$ and let the Mean Exit Time $\mathbb{E}\big[\tau^x_{B_R(x_0)}\big]$ be a viscosity solution of (5). Then,*

$$\mathbb{E}\left[\tau^x_{B_R(x_0)}\right] \ge \frac{R_0^2 - r^2}{2\varepsilon^2\sigma\, d}\,. \tag{6}$$

Proving an upper bound for a general degenerate matrix Q is more challenging, since the MET will not be bounded if there is no diffusion. Nevertheless, assuming some non-degeneracy of the diffusion in just one direction is sufficient to provide an upper estimate.

Theorem 1.4 (Upper bound for MET)**.** *Let Ω be an open set, let $0 \le Q(x) \le \sigma I_{d\times d}$ for every $x \in \Omega \subset \mathbb{R}^d$. Assume that there exist $\beta, \Lambda > 0$, $i \in \{1,\dots,d\}$ and a vector $v \in \mathbb{S}^{d-1}$ such that*

$$v^T Q(x) v \ge \beta \qquad \textit{and} \qquad (v\cdot\nabla L(x))\,(v\cdot x) \le \Lambda (v\cdot x)^2 + \frac{\varepsilon^2\beta}{2} \qquad \forall x \in \Omega\,. \tag{7}$$

Assume that $R_v := \max\{|v\cdot x| : x \in \Omega\} < +\infty$ for $x \in B_r(x_0) \subset \Omega$ with $0 < r \le R_v$. If the Mean Exit Time $\mathbb{E}\left[\tau^x_{B_{R_0}(0)}\right]$ is a viscosity solution of (5), then we have that

$$\mathbb{E}\left[\tau^x_\Omega\right] \le \frac{2}{\Lambda}\left(e^{\frac{\Lambda R_v^2}{2\beta\varepsilon^2}} - e^{\frac{\Lambda r^2}{2\beta\varepsilon^2}}\right). \tag{8}$$

The last two Theorems will be proved in Section 3 along with more general results.

Remark (*About the condition (7) for the upper bounds*)**.** Note that the assumption on ∇L in Theorem 1.4 is unnecessary when $\Omega \subset \mathbb{R}^d$ is bounded and $L \in C^1(\mathbb{R}^d)$. In this case, the non-degeneracy of $Q(x)$ in the direction $v \in \mathbb{S}^{d-1}$ suffices to establish the upper bound of the MET. To clarify the assumptions involved in this result, we now present a few simple yet illustrative examples.

Example 1. *Polynomial loss.* Let us consider $\Omega = B_R(0)$ and

$$L(x_1,x_2) = \frac{1}{p}(x_1^2+x_2^2)^{\frac{p}{2}} \qquad \text{and} \qquad v = \begin{pmatrix}1\\0\end{pmatrix}.$$

Then, assumption (7) reads as

$$(v\cdot\nabla L(x))\,(v\cdot x) = \left(x_1^2+x_2^2\right)^{\frac{p-2}{2}} x_1^2 \le R^{p-2}\, x_1^2 = \Lambda\,(v\cdot x)^2 \qquad \forall x \in B_R(0)\,,$$

and hence, it is satisfied with $\Lambda = R^{p-2}$ and for any $\varepsilon > 0$. Note that when $p = 2$, we have $\Lambda = 1$ for any domain $\Omega \subset \mathbb{R}^d$, even if it is unbounded.

Example 2. *Exponential loss.* Let us consider $\Omega = B_R(0)$ with

$$L(x_1,x_2) = e^{x_1} + \frac{x_2^2}{2} \qquad \text{and} \qquad v = \begin{pmatrix}1\\0\end{pmatrix}.$$

In this case, the term

$$(v\cdot\nabla L(x))\,(v\cdot x) = x_1\, e^{x_1}\,,$$

tends to 0 linearly as x approaches to the origin. However, we can choose $0 < r \le R$ such that

$$x_1\, e^{x_1} \le \frac{\varepsilon^2\beta}{2} \qquad \forall x \in B_r(0)\,.$$

Therefore, assumption (7) holds with

$$(v\cdot\nabla L(x))\,(v\cdot x) = x_1\, e^{x_1} \le \frac{e^R}{r} x_1^2 + \frac{\varepsilon^2\beta}{2} = \Lambda\,(v\cdot x)^2 + \frac{\varepsilon^2\beta}{2}\,,$$

where $\Lambda = e^R/r$. A similar argument applies to any C^1 loss function, provided that Ω is bounded.

EXAMPLE 3. *Unbounded domain.* We present now the importance of assumption (7) whenever Ω is unbounded. Consider $\Omega = [-1,1] \times \mathbb{R}^d$ with

$$L(x_1, x_2) = \frac{x_1^2 x_2^2}{2} \qquad \text{and} \qquad v = \begin{pmatrix} 1 \\ 0 \end{pmatrix} .$$

Note that for any $\Lambda > 0$ and $x_1 \neq 0$ we can choose $x_2 = \sqrt{1 + \Lambda/2} \in \mathbb{R}^d$ such that

$$(v \cdot \nabla L(x))\,(v \cdot x) = 2x_1^2 x_2^2 > \Lambda x_1^2 ,$$

contradicting assumption (7).

The last question we consider in this manuscript concerns the asymptotic behaviour of the learning process. We aim to characterize and show the convergence of the parameter distribution for large times. Hence, we can state the last question as follows:

Q3: Does the distribution of neural network parameters converge using SGD?

To address this question we implement two different method relying on the known literature: a *duality method* and an *entropy method*.

DUALITY METHOD. In the non-degenerate case we exploit the dual equation of the Fokker-Planck equation (3), which is of Ornstein-Uhlenbeck type. We found that the recent results of Porretta [58] are especially useful in our setting, when the matrix is non-degenerate, and help to construct the stationary solutions by means of non-degenerate approximations. One of the key point in Porretta's approach is that the asymptotic stabilization of the solution ρ of (3) for large times, is equivalent to the oscillation decay of the solution of the dual equation. As far as we know, this method has a drawback: it requires the diffusion matrix Q to be uniformly elliptic. This is the reason why we introduce in Section 4.1 a variant of the (SGD) that we call Noisy Stochastic Gradient Descent (NSGD) by adding some gaussian noise in each iteration, something that kills the non-degeneracy. Namely,

$$\theta_{n+1} = \theta_n - \frac{\eta}{\mathscr{b}} \sum_{b_i \in B_n} \nabla L_{b_i}(\theta_n) + \eta Z_n \qquad \forall n \geq 0 , \tag{NSGD}$$

with $\{Z_n\}_{n\geq 0}$ being a family of i.i.d. gaussian processes satisfying for some $\delta > 0$

$$Z_n \sim \mathcal{N}(0, \delta I_{d\times d}) \qquad \forall n \geq 0 .$$

In this context, the diffusion matrix of the associated Fokker-Planck equation is uniformly elliptic, indeed, $Q_\delta(x) = \mathscr{b}^{-1}\Sigma^2(x) + \delta I_{d\times d}$. The results of [58] prove the convergence of the NSGD to a unique stationary measure, see Corollary 4.3. This convergence is qualitative in several aspects: on the one hand, it only works in non-degenerate cases, which are not realistic in ML applications. On the other hand the convergence rate towards equilibrium cannot be quantified, since it is obtained by a non-constructive proof. However, the above convergence for non-degenerate Fokker-Planck equations, gives us an approximation that allows to prove the existence of stationary measures for the "original" degenerate problem in the limit $\delta \to 0$. Indeed, in Section 4.2 we will prove the following theorem.

Theorem 1.5 (Existence of steady states)**.** *Assume that the diffusion matrix Q satisfies that there exist $\sigma_0, \sigma_1 > 0$ such that*

$$\|\sqrt{Q(\cdot)}\|_\infty \leq \sigma_0 \qquad \textit{and} \qquad \|\sqrt{Q(x)} - \sqrt{Q(y)}\| \leq \sigma_1 |x-y| \qquad \forall x, y \in \mathbb{R}^d\,. \quad (9)$$

Assume that the drift term $\nabla L(x)$ satisfies that there exist $\alpha, R > 0$ and $\gamma \geq 2$ such that

$$\nabla L(x) \cdot x \geq \alpha |x|^\gamma \qquad \forall x \in \mathbb{R}^d \quad \textit{with} \quad |x| \geq R, \tag{10}$$

and there exists $c_0 > 0$ such that

$$(\nabla L(x) - \nabla L(y)) \cdot (x-y) \geq -c_0 |x-y| \qquad \forall x, y \in \mathbb{R}^d\,. \tag{11}$$

Then there exists at least one invariant probability measure $\rho_\infty \in \mathcal{P}(\mathbb{R}^d)$ such that

$$\nabla \cdot \left(\varepsilon^2 \nabla \cdot (Q \rho_\infty) + \rho_\infty \nabla L \right) = 0\,, \tag{12}$$

in the sense of measures, that is, for all $\varphi \in C^{1,2}(\mathbb{R}^d)$ it holds that

$$\int_{\mathbb{R}^d} \left[\varepsilon^2 \mathrm{tr}\left(Q(x) D^2 \varphi(x) \right) - \nabla L(x) \cdot \nabla \varphi(x) \right] d\rho_\infty(x) = 0\,. \tag{13}$$

Entropy Method. When the invariant measure has a L^1-density, entropy methods allow for quantitative asymptotic results, like exponential decay towards equilibrium. We discuss this technique in detail in Section 4.4, where we adapt the classical Bakry-Émery approach [10] to our setting.

In the particular situations of a diffusion matrix that is degenerate but constant, and of a quadratic loss function $L(x) = \frac{1}{2} x^T C x$, equation (3) becomes

$$\partial_t \rho = \nabla \cdot (Q_0 \nabla \rho + \rho C x)\,. \tag{14}$$

In this case, the Bakry-Émery approach applies, as proven by Arnold and Erb [5], when the following two conditions are satisfied

(I) *Confinement potential:* C is positive stable, i.e. all eigenvalues has positive real part.

(II) *Hörmander's condition:* There are no eigenvectors of C in $\ker Q_0$.

The first assumption ensures the λ-convexity of L, while the second ensures that the diffusion is spread in the whole space by the drift.

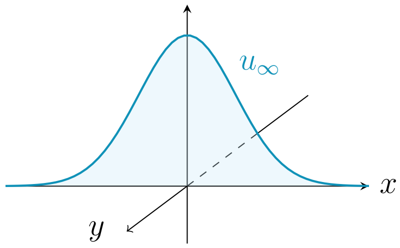
When condition (II) fails, the stationary state may not have an L^1 density, showing that our existence Theorem 1.5 is somehow sharp. We shall consider next an example in which exponential convergence in a suitable topology holds even when condition (II) is not met. More precisely, we will prove convergence for the degenerate non-Hörmander case which reads

$$\partial_t u = \nabla_{x,y} \cdot \left(\begin{pmatrix} Q_0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \nabla_x u \\ \nabla_y u \end{pmatrix} + u \begin{pmatrix} C_0 & 0 \\ 0 & C_3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right) \tag{15}$$

with $u : (0, \infty) \times \mathbb{R}^n \times \mathbb{R}^{d-n} \to \mathbb{R}^d$, assuming only that condition (II) holds for Q_0 and C_0, i.e. for the variables that we have called x.

In this framework, we prove the asymptotic convergence of (15) in the 2-Wasserstein distance and the exponential convergence of the second moments to a steady state of the form

$$u_\infty(x,y) = g_\infty(x)\,\delta_0(y)\,.$$



where $g_\infty(x) = c\,e^{-\frac{x^T K^{-1} x}{2}}$ and $K \in \mathbb{R}^{d\times d}$ is the unique positive definite matrix satisfying $2Q_0 = C_0 K + K C_0$.

Theorem 1.6 (Convergence in the Non-Hörmander case)**.** *Let us consider the equation (15) with $u_0 \in L^1(\mathbb{R}^d) \cap \mathcal{P}_2(\mathbb{R}^d)$ and with marginal on the x-variable in $L^2(\mathbb{R}^n, g_\infty^{-1}\,\mathrm{d}x)$. Let $Q_0, C_0 \in \mathbb{R}^{n\times n}$ satisfy assumptions (I) and (II) in $\mathbb{R}^n$. If $C_3 \in \mathbb{R}^{(d-n)\times(d-n)}$ is positive stable, then*

$$u(t) \xrightarrow{t\to\infty} u_\infty$$

in 2-Wasserstein. Moreover, the following convergence of the second moments holds

$$\left| \int_{\mathbb{R}^n} \int_{\mathbb{R}^{d-n}} (|x|^2 + |y|^2)(u(t,x,y) - u_\infty(x,y))\,\mathrm{d}x\,\mathrm{d}y \right| \le \kappa e^{-\lambda t}\,, \tag{16}$$

with $\lambda > 0$, and κ is a constant depending on u_0, Q_0, C_0.

We will prove an expanded version of this statement, with a quantitative rate λ, in Theorem 4.13.

This may not seem a realistic model for our ML problem. However, this may be considered as a description of the local behaviour of the Fokker-Planck equation (3), through a linearization of Q and L around the local minima of L. Indeed, in Section 4.5, we exploit this linearization argument by defining

$$u(t,y) = \rho(t, x_0 + \varepsilon z)$$

where $x_0 \in \mathbb{R}^d$ is a local minimum of L. Thus, the equation for our new function u reads

$$\partial_t u = \nabla \cdot \big(Q(x_0)\nabla u + u D^2 L(x_0) z \big) \tag{17}$$

up to an error of order ε.

We conclude this manuscript by presenting some open questions concerning the approximation of the steady state of (3) by

$$\rho_\infty(x) \approx \sum_{i=1}^{M} m_i(\infty)\, u_{i,\infty}\left(\frac{x - x_i}{\varepsilon}\right)\,, \tag{18}$$

where $u_{i,\infty}$ denotes the steady state of (17) linearized around the local minimum $x_i \in \mathbb{R}^d$, and $m_i(\infty)$ represent the mass partition such that $\sum_{i=1}^M m_i(\infty) = 1$.

Notation: Let $I_{d\times d}$ be the identity matrix of dimension $d \times d$. We denote by $\mathcal{M}(\mathbb{R}^d)$ the space of probability measures in $\mathbb{R}^d$. For these measures, we define the norm $\|\mu\|_{\mathcal{M}_k} = \|\mu\|_{\mathcal{P}_k} := \int_{\mathbb{R}^d} (1 + |x|^2)^{k/2} \mathrm{d}|\mu|(x)$ and the space $\mathcal{M}_k(\mathbb{R}^d) := \{\mu \in \mathcal{M}(\mathbb{R}^d) : \|\mu\|_{\mathcal{M}_k} < +\infty\}$. Analogously, we write $\mathcal{P}_k(\mathbb{R}^d)$ for the probability measures with k finite moments.

1.2 Related work on the SGD and the SDE approximation

Stochastic gradient descent (SGD) has a long history as an optimization technique in machine learning in general, and for neural networks in particular (see e.g. [8, 42]). In recent years, SGD has been the object of several theoretical and experimental studies, considering different aspects of the algorithm, due to its unique mixture of desirable properties of quality of the results, robustness, and generalization capabilities. Special attention has been devoted to the implicit bias, or implicit regularization, in the optimization (for recent reviews see [30, 65]), motivated by seminal works like [72] which show that neural networks are effectively trained by SGD for a number of parameters that exceeds the number of data point (overparametrized), and that results are qualitatively unaffected by explicit regularization. On the other hand, it has been observed that more efficient adaptive optimization methods may generalize worse than SGD [68, 73]. While implicit regularization in gradient descent algorithms can be attributed to multiple factors, involving interactions between the optimization algorithm and the properties of both the parametrization and the dataset [7, 12, 23, 32, 63], the stochasticity of SGD is believed to play a relevant role, in particular due to its state-dependent noise [18, 34]. Another crucial aspect of SGD design is the dependence of the results on the ratio between learning rate and batch size, also called linear scaling rule, which appears to be a key factor to obtain so-called flat minima, ensuring better generalization [29, 37] (but see also [52]).

The continuous stochastic approximation of SGD with an SDE that we consider in this paper was introduced and proved in [44]. However, the idea of considering a stochastic iterative algorithm as a discretization of a continuous time stochastic process, or, equivalently, that a continuous-time SDE could be obtained as a limiting procedure, is classical (see e.g. [41, Ch. 9]). More recently, an SDE model for SGD has been proposed in [50, 51], where the authors assume a gaussian state-independent covariance matrix as a model for a large number of samples, and in [37], where the authors emphasize that different learning rate with the same scaling to batch size give rise to the same continuous approximation and discuss the common practice of assuming the covariance to be approximately equal to the Hessian of the loss (sometimes called label noise). In [61] a different, but still state-independent, covariance matrix is considered, together with a time-dependent learning rate, which allow the authors to prove convergence to a stationary state with similar arguments to the ones used in the present paper for state-dependent diffusion matrices. A learning rate that decreases with time has also been considered in [55], to prove convergence of SGD for convex optimization problems, again in terms of the SDE continuous model. A general approach to the stochastic approximation with SDE of online learning in terms of the theory of semigroups was given in [26]. A different proof of the stochastic approximation of SGD was provided in [22], where the authors show experimentally the limiting behaviors for general state-depentent noise. Key properties of the diffusion that approximates SGD are given in [36], where the authors discuss the limiting behavior and the escape time for non-degenerate diffusion. We can compare in particular [36, Theorem 2] with our treatment of Question 2, where we provide quantitative bounds from above and below for the mean exit time in the case of degenerate diffusion, that is the generic overparametrized situation. Two regimes in the SDE model of SGD, that dominated by drift, and the diffusive one, were considered in [16] and applied to elementary architectures: these results should be compared directly with our theoretical study in Section 2, where we can obtain general quantitative results of mass concentration around minima in the drift regime.

Two observed features of the SGD that are reproduced by the SDE continuous model are the scaling law, which is intrinsic in the limiting procedure, and the implicit regularization. This second one has been studied for SDE in a long series of works. In [21], under the assumption of the existence and uniqueness of a steady state obtained by a divergence-free force, the authors show that the Fokker-Planck trajectories are monotone minimizer of the KL-divergence with respect to the steady state, hence justifying a Bayesian inference of the SGD, and show that such a state is not a minimum of the loss function itself but rather of a regularization in terms of the Shannon entropy. In [70] the authors assume the covariance of the SDE to be close to the Hessian of the loss function and deduce different escape time estimates from minima depending on the local spectral properties of the Hessian, hence defining an implicit bias towards flat minima. Still modeling the noise covariance with the Hessian of the loss, [46] discuss yet another form of implicit bias in SGD through a continuous SDE model: they show that in the presence of a manifold of minima, SGD does not randomly walks on that manifolds once it reaches it, but rather continues to minimize the trace of the Hessian, hence stabilizing around the flattest region of the minima. They obtain this result by first observing that the dynamics that is normal to the manifold approximately behaves like a constant-covariance Ornstein-Uhlenbeck process, and deduce asymptotically reaching the manifold of minima, and then study the tangent dynamics. The techiques that we introduce that allow us to obtain a rigorous proof of existence of stationary states for the SDE with the true covariance matrix, and a proof of the decay in the non degenerate case, may provide insight on how to consider the tangent dynamics for a general state dependent noise.

Limitations of the SDE model of SGD have been pointed out in several occasions. In [71] the whole approximation procedure is questioned in terms of heuristic arguments concerning the forcing of the scaling law in the limiting procedure, and the difference between the covariance matrix and the more commonly considered Hessian. More recently, [43] have considered the possible deviations of the SDE approximation of SGD for large values of the quotient between learning rate and batch size, by directly comparing trajectories of SGD with a fine discretization of the SDE. Large learning rates are indeed often used in practice to provide an implicit regularization bias [3, 29, 47].

Extensions of SDE modeling of SGD currently include proof of stochastic approximation for adaptive schemes such as ADAM [49], while [33, 62] have considered non-Gaussian noise.

Another notable mathematical approach to modeling the dynamics of neural network parameters trained via SGD relies on mean field limit techniques. When the number of neurons in a layer grows to infinity, the empirical distribution of the network parameters evolves according to a deterministic partial differential equation, expressed as a Wasserstein gradient flow, see for instance the survey [27]. This perspective allows one to bypass the complexities of individual parameter updates and instead analyze the global behavior of the parameter distribution. However, this approach is limited to shallow networks or to deep linear networks, where a well-posed infinite-width limit can be established. In particular, [59] rigorously proves a Law of Large Numbers and a Central Limit Theorem quantifying fluctuations around the mean field trajectory, while [23] shows that in a deep linear network, the infinite-width limit exhibits exponential convergence to a continuous-time limit.

2 Analysis in the drift regime

The dynamics of the parameters evolving according to the SGD present two different regimes of behaviour. On the first regime, which we call the drift regime, the diffusion is weaker than the drift term ∇L and the latter is the one who determines the behaviour of the learning process. This fact leads to a concentration phenomena around the local minima of the loss function L whenever the effective learning rate $\varepsilon > 0$ is small enough. As we will discuss in Section 3, once the concentration at small time has occurred, the diffusion tends to spread out the parameters. In this section, we are interested in quantifying this concentration behaviour at small times, which can be formulated as

Q1: How do parameters evolve in the first stage of training with SGD?

To this aim, we study the probability density function ρ of the stochastic process associated to the SGD. Let us recall that ρ satisfies the equation

$$\begin{cases} \partial_t \rho = \nabla \cdot \Big(\varepsilon^2 \nabla \cdot (Q(x)\rho) + \rho \nabla L(x) \Big) & \text{on} \quad (0,\infty) \times \mathbb{R}^d \,, \\ \rho(0,x) = \rho_0(x) & \text{in} \quad \mathbb{R}^d \,, \end{cases} \tag{19}$$

with $\nabla \cdot A = (\nabla \cdot A_{1\cdot}, \ldots, \nabla \cdot A_{d\cdot})$ and $A \in \mathbb{R}^{d\times d}$. In this framework, we will address the above question by obtaining lower bounds for the following quantity:

$$\int_{B_R(x_0)} \rho(t,x)\,\mathrm{d}x \,,$$

with $x_0 \in \mathbb{R}^d$ being a local minimum of L and $R > 0$ being certain radius. Nevertheless, before performing a local analysis, let us show a global concentration estimates when L is strictly convex in the whole space $\mathbb{R}^d$. This can be easily deduced from the following upper bound of the second moments of ρ.

Lemma 2.1. *Assume that L is λ-convex with $\lambda > 0$, L has a unique minimum in x_0 with $L(x_0) = 0$ and $0 \le \operatorname{tr}(Q(x)) \le \sigma$ for every $x \in \mathbb{R}^d$. If $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, then $\forall t > 0$*

$$\int_{\mathbb{R}^d} |x-x_0|^2 \rho(t,x)\,\mathrm{d}x \le \left(\int_{\mathbb{R}^d} |x-x_0|^2 \rho_0(x)\,\mathrm{d}x - \varepsilon^2 \frac{2\sigma}{\lambda} \right) e^{-\lambda t} + \varepsilon^2 \frac{2\sigma}{\lambda} \,.$$

Proof. Let us differentiate the second moment of the solution.

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} |x-x_0|^2 \rho(t,x)\,\mathrm{d}x &= \varepsilon^2 \int_{\mathbb{R}^d} \operatorname{tr}(QD^2(|x-x_0|^2))\rho(t,x)\,\mathrm{d}x \\ &\quad - \int_{\mathbb{R}^d} \nabla\left(|x-x_0|^2\right) \cdot \nabla L\, \rho(t,x)\,\mathrm{d}x \\ &= 2\varepsilon^2 \int_{\mathbb{R}^d} \operatorname{tr}(Q(x))\rho\,\mathrm{d}x - 2\int_{\mathbb{R}^d} (x-x_0)\cdot \nabla L(x) \rho\,\mathrm{d}x \end{aligned}$$

Note that since L is λ-convex and x_0 is any minimum, we have

$$(x-x_0)\cdot \nabla L(x) \ge L(x) - L(x_0) + \frac{\lambda}{2}|x-x_0|^2 \ge \frac{\lambda}{2}|x-x_0|^2 \,.$$

Using this estimate and $0 \leq \operatorname{tr}(Q(x)) \leq \sigma$ we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d} |x - x_0|^2 \rho(t, x)\, \mathrm{d}x \leq 2\varepsilon^2\sigma - \lambda \int_{\mathbb{R}^d} |x - x_0|^2\, \rho(t, x)\, \mathrm{d}x\,.$$

The result follows by integrating this expression in $[0, t]$. □

Once we have estimated the second moment of ρ, it is possible to obtain a concentration estimate by Chebyshev's inequality.

Corollary 2.2. *Under the assumptions of Lemma 2.1, we have that for every $R > 0$ and $t > 0$*

$$\int_{B_R(x_0)} \rho(t, x)\, \mathrm{d}x \geq 1 - \frac{1}{R^2} \left[\left(\int_{\mathbb{R}^d} |x - x_0|^2 \rho_0(x)\, \mathrm{d}x - \varepsilon^2 \frac{2\sigma}{\lambda} \right) e^{-\lambda t} + \varepsilon^2 \frac{2\sigma}{\lambda} \right]. \tag{20}$$

Proof. For estimate (20), we use Chebyshev's inequality: For any μ-measurable function f, $R > 0$ and $1 < p < \infty$ it holds that

$$\mu\left(\left\{x \in \mathbb{R}^d : |f(x)| > R\right\}\right) \leq \frac{1}{R^p} \int_{\mathbb{R}^d} |f(x)|^p \mathrm{d}\mu(x)\,.$$

Hence, considering $\mu = \rho(t)$, $f(x) = |x - x_0|$ and $p = 2$, we obtain

$$\int_{|x - x_0| > R} \rho(t, x)\, \mathrm{d}x \leq \frac{1}{R^2} \int_{\mathbb{R}^d} |x - x_0|^2 \rho(t, x)\, \mathrm{d}x\,.$$

Then, we use Lemma 2.1 and the fact that $\rho(t) \in \mathcal{P}(\mathbb{R}^d)$ to control the second moment and to conclude. □

Example 4: non degenerate case. We remark that Lemma 2.1 and Corollary 2.2 does not implies the convergence to $\delta_{x_0}(x)$ in any sense due to the error term $\varepsilon^2 \frac{2\sigma}{\lambda}$. However, this error seems to be unavoidable whenever there is a diffusion term. Let us consider the case $Q(x) = \frac{\sigma}{d} I$ and $L(x) = \frac{\lambda}{2}|x - x_0|^2$ for every $x \in \mathbb{R}^d$ with $\sigma, \lambda > 0$. In this case, the fundamental solution is smooth and reads as

$$\rho(t, x) = \frac{1}{\left(2\pi \frac{\varepsilon^2\sigma}{\lambda d}(1 - e^{-2\lambda t})\right)^{d/2}} \exp\left(-\frac{\lambda d |x - x_0|^2}{2\varepsilon^2 \sigma (1 - e^{-2\lambda t})} \right).$$

Moreover, we know that $\rho(t) \to \rho_\infty$ uniformly in compacts, in 2-Wasserstein distance and in $L^2(\mathbb{R}^d, \rho_\infty^{-1})$ with

$$\rho_\infty(x) = \frac{1}{\left(2\pi \frac{\varepsilon^2\sigma}{\lambda d}\right)^{d/2}} \exp\left(-\frac{\lambda d |x - x_0|^2}{2\varepsilon^2 \sigma} \right).$$

However, ρ_∞ does not concentrate in x_0 due to the exponential tails. Indeed, the second moment is

$$\int_{\mathbb{R}^d} |x - x_0|^2 \rho_\infty(x)\, \mathrm{d}x = \operatorname{tr}\left(\frac{\varepsilon^2 \sigma}{\lambda d} I \right) = \frac{\varepsilon^2 \sigma}{\lambda}\,.$$

One may think that this is not a representative example since the diffusion is non-degenerate. Nevertheless, let us consider the degenerate case where the operator is not even hypocoercive.

EXAMPLE 5: DEGENERATE CASE. Assume $L(x) = \frac{\lambda}{2}|x|^2$ with global minimum at $x_0 = 0$ and the diffusion matrix

$$Q(x) = \begin{pmatrix} \sigma & 0\dots & 0 \\ 0 & 0\dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

with $\sigma, \lambda > 0$. Note that this operator is not hypocoercive since $\lambda I e_i = \lambda e_i$ and $e_i \in \ker(Q)$ for any $i = 2, \dots, d$. Given an initial datum $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$, the solution to

$$\partial_t \rho = \nabla \cdot \left(\varepsilon^2 Q \nabla \rho + \lambda x \rho \right) ,$$

is the following:

$$\rho(t,x) = \frac{e^{\lambda(d-1)t}}{\sqrt{2\pi \frac{\varepsilon^2 \sigma}{\lambda}(1 - e^{-2\lambda t})}} \int_{\mathbb{R}} \exp\left(-\frac{\lambda |x_1 - \tilde{x}_1 e^{-\lambda t}|^2}{2\varepsilon^2 \sigma (1 - e^{-2\lambda t})} \right) \rho_0\left(\tilde{x}_1, x' e^{\lambda t}\right) \mathrm{d}\tilde{x}_1 \,,$$

with $x' = (x_2, \dots, x_d) \in \mathbb{R}^{d-1}$. Due to the tail decay of $\rho_0 \in \mathcal{P}_2(\mathbb{R}^2)$, the stationary solution is

$$\rho_\infty(x) = \frac{1}{\sqrt{2\pi \frac{\varepsilon^2 \sigma}{\lambda}}} \exp\left(-\frac{\lambda |x_1|^2}{2\varepsilon^2 \sigma} \right) \delta_{0'}(x')$$

See Theorem 4.13 for a 2-Wasserstein convergence of $\rho(t) \to \rho_\infty$ when $t \to \infty$. Note that even in this very degenerated case the second moment of the invariant measure is given by

$$\int_{\mathbb{R}^d} |x|^2 \rho_\infty(x) \,\mathrm{d}x = \frac{1}{\sqrt{2\pi \frac{\varepsilon^2 \sigma}{\lambda}}} \int_{\mathbb{R}} |x_1|^2 \exp\left(-\frac{\lambda |x_1|^2}{2\varepsilon^2 \sigma} \right) \mathrm{d}x_1 = \frac{\varepsilon^2 \sigma}{\lambda}$$

Based on this simple example, we aim to estimate this error that the diffusion generates in the concentration phenomena around local minima. As in the example, the main influent parameters are the λ-convexity of the loss function in the basin of attraction, the effective learning rate $\varepsilon > 0$ and the L^∞-norm of the matrix Q, which quantify the diffusion. From an application perspective, these parameters establish the likelihood that the weights of the neural network will fall into the nearest local minimum.

2.1 Local mass concentration

In this section, we show that in the first steps of the learning process, the parameters tends to concentrate with high probability around the local minima of the loss function. Assuming without loss of generality that 0 is a local minimum of L, we would like to obtain an estimate on the following quantity

$$\mathbb{P}\{X_t \in B_\varepsilon(0)\} = \int_{B_\varepsilon(0)} \rho(t,x) \,\mathrm{d}x \,. \tag{21}$$

The main problem is that the variation of this quantity depends on the flux on the boundary of $B_\varepsilon(0)$ which is a priori unknown. Namely, if we try to differentiate this quantity using the equation of the probability density function, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_{B_\varepsilon(0)}\rho(t,x)\,\mathrm{d}x &= \int_{B_\varepsilon(0)}\nabla\cdot\Big(\varepsilon^2\nabla\cdot(\rho Q)+\rho\nabla L\Big)\,\mathrm{d}x\\ &= \int_{|x|=\varepsilon}\Big(\varepsilon^2\nabla\cdot(\rho Q)+\rho\nabla L\Big)\cdot\frac{x}{|x|}\,\mathrm{d}x\end{aligned}$$

Since this variation is complicated to estimate with (21) to close the differential inequality, we will consider a different quantity with a controllable variation. To motivate this new approach, let us consider the following continuity equation ($\varepsilon = 0$)

$$\begin{cases}\partial_t\mu = \nabla\cdot(\mu\nabla L)\\ \mu(0,x)=\mu_0(x)\end{cases}\tag{22}$$

with L being λ-convex with $\lambda > 0$ and a global minimum at 0. It is well known that if $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, then the solution $\mu(t)$ converge to δ_0 in the Wassertstein topology [1], but we are interested in the speed of concentration. Moreover, the associated flow dynamic is

$$\begin{cases}\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) & = -\nabla L(\Phi_t(x))\\ \Phi_0(x) & = x.\end{cases}$$

and the solution of (22) is the pushforward $\mu(t,x) = (\Phi_t)_{\#}\mu_0(x)$. In order to show the concentration velocity of $\Phi_t(x)$, let us compute the variation in time of the Euclidean distance from $\Phi_t(x)$ to the global minimum at 0:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|\Phi_t(x)-0|^2 &= 2\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x)\cdot(\Phi_t(x)-0) = -2\nabla L(\Phi_t(x))\cdot(\Phi_t(x)-0)\\ &\le -\lambda|\Phi_t(x)-0|^2\,,\end{aligned}$$

which implies that

$$|\Phi_t(x)| \le e^{-\frac{\lambda}{2}t}|x| \qquad \forall t>0.\tag{23}$$

Hence, the flow tends to concentrate particles starting from a point $x\in\mathbb{R}^d$ around 0 with exponential velocity. This phenomena can be understood in the sense of the following lemma.

Lemma 2.3. *Let L be a λ-convex function with $\lambda > 0$ and let μ be a solution to (22) in the sense of [1] with $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then, for every $t > 0$ it holds that*

$$\int_{B_{R(t)}(0)}\mu(t,x)\,\mathrm{d}x \ge \int_{B_{R_0}(0)}\mu_0(x)\,\mathrm{d}x$$

with $R(t) = R_0\,e^{-\frac{\lambda}{2}t}$.

Proof. Recall that $\mu(t,x) = (\Phi_t)_{\#}\mu_0(x)$. Then,

$$\int_{B_{R(t)}(0)}\mu(t,x)\,\mathrm{d}x = \int_{\Phi_t^{-1}(B_{R(t)}(0))}\mu_0(x)\,\mathrm{d}x\,.$$

Since $R(t) = R_0\,e^{-\frac{\lambda}{2}t}$ and (23) holds, we have that

$$\Phi_t(B_{R_0}(0)) \subseteq B_{R(t)}(0)\,,$$

and the result follows. $\square$

Considering the diffusion equation (19) for ρ. Compared with the result in the lemma above, we expect that the probability of a suitable shrinking ball should be non decreasing up to an error depending on ε and before certain time. Namely, our goal is to obtain a lower bound in terms of the initial datum ρ_0 of the form

$$\int_{B_{R(t)}(0)} \rho(t,x)\,\mathrm{d}x \geq \int_{B_{R_0}(0)} \rho_0(x)\,\mathrm{d}x - c(\varepsilon,t)$$

with an appropiate shrinking radius $R(t)$ and $c(\varepsilon,t)$ tending to 0 as ϵ goes to 0. Since we want to use the equation of ρ and test it with the characteristic function $\chi_{B_{\mathbb{R}(t)}(0)}(x)$, we use the following $C^{1,2}$ approximation cut-off function

$$\varphi(t,r) = \begin{cases} 1, & \text{if} \qquad r \leq R(t) \\ 1 - \frac{2}{\delta^2 R(t)^2}(r-R(t))^2, & \text{if} \qquad R(t) < r \leq (1+\frac{\delta}{2})R(t) \\ \frac{2}{\delta^2 R(t)^2}((1+\delta)R(t)-r)^2, & \text{if } (1+\frac{\delta}{2})R(t) < r \leq (1+\delta)R(t) \\ 0, & \text{if } (1+\delta)R(t) < r\,. \end{cases} \tag{24}$$

for some radius $R(t)$ to be chosen later and some small constant $\delta > 0$. In terms of this test function, the result reads as follows.

Theorem 2.4. *Assume that L is λ-convex in $B_{(1+\delta)R_0}(0)$ with a minimum at 0 and $\lambda > 0$. Let ρ be a weak solution of (3) with $0 \leq Q(x) \leq \sigma I_{d\times d}$ for every $x \in B_{(1+\delta)R_0}(0)$. Then, for any time dependent radius $R(t)$ satisfying*

$$R'(t) \geq -\frac{\lambda}{2} R(t) \quad \forall t > 0\,, \tag{25}$$

it holds that for every $T > 0$ and φ as in (24)

$$\int_{\mathbb{R}^d} \varphi(T,|x|)\rho(T,x)\,\mathrm{d}x \geq \int_{\mathbb{R}^d} \varphi(0,|x|)\rho_0(x)\,\mathrm{d}x - C\varepsilon^2 \int_0^T \frac{1}{R(t)^2}\,\mathrm{d}t \tag{26}$$

with $C > 0$ depending only on δ, σ and d.

Proof. We want to differentiate the quantity in (26), therefore let us compute the derivatives of the test function f.

$$\partial_t \varphi(t,r) = \begin{cases} \frac{4}{\delta^2 R(t)^2}(r-R(t))r\frac{R'(t)}{R(t)}, & \text{if} \qquad R(t) < r \leq (1+\frac{\delta}{2})R(t) \\ \frac{4}{\delta^2 R(t)^2}((1+\delta)R(t)-r)r\frac{R'(t)}{R(t)}, & \text{if } (1+\frac{\delta}{2})R(t) < r \leq (1+\delta)R(t) \\ 0, & \text{otherwise}\,. \end{cases}$$

$$\partial_r \varphi(t,r) = \begin{cases} -\frac{4}{\delta^2 R(t)^2}(r-R(t)), & \text{if} \qquad R(t) < r \leq (1+\frac{\delta}{2})R(t) \\ -\frac{4}{\delta^2 R(t)^2}((1+\delta)R(t)-r), & \text{if } (1+\frac{\delta}{2})R(t) < r \leq (1+\delta)R(t) \\ 0, & \text{otherwise}\,. \end{cases}$$

$$\partial_{rr} \varphi(t,r) = \begin{cases} -\frac{4}{\delta^2 R(t)^2}, & \text{if} \qquad R(t) < r \leq (1+\frac{\delta}{2})R(t) \\ \frac{4}{\delta^2 R(t)^2}, & \text{if } (1+\frac{\delta}{2})R(t) < r \leq (1+\delta)R(t) \\ 0, & \text{otherwise}\,. \end{cases}$$

Let us recall the gradient and the Hessian matrix of radial functions with respect to x.

$$\begin{aligned}
\nabla\varphi(t,|x|) &= \partial_r\varphi(t,|x|)\frac{x}{|x|}\,,\\
D^2\varphi(t,|x|) &= \partial_{rr}\varphi(t,|x|)P_0 + \frac{\partial_r\varphi(t,|x|)}{|x|}(I_{d\times d}-P_0)\,,
\end{aligned}$$

with $P_0 = \frac{x}{|x|}\otimes\frac{x}{|x|}$ being the projection matrix with respect to the vector $\frac{x}{|x|}$. Now, let us differentiate the desired quantity and use the equation of ρ:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(t,|x|)\rho(t,x)\,\mathrm{d}x &= \int_{\mathbb{R}^d}\partial_t\varphi(t,|x|)\rho(t,x)\,\mathrm{d}x\\
&+\varepsilon^2\int_{\mathbb{R}^d}\mathrm{tr}\big(Q(x)D^2\varphi(t,|x|)\big)\rho(t,x)\,\mathrm{d}x - \int_{\mathbb{R}^d}\nabla\varphi(t,|x|)\cdot\nabla L(x)\rho(t,x)\,\mathrm{d}x
\end{aligned}$$

Note that this differentiation is formal, however we can make it rigourous by integrating the expression above in time and using Definition 1.1 for weak solutions of (3).

Then, we have to use the previously calculated derivatives of φ. We will compensate the integral term of ∇L with the one with $\partial_t\varphi$ and the integral with ϵ will be the error. In order to simplify the computation we split the integrals in two regions, depending on the time dependent radius $R = R(t)$:

$$\begin{aligned}
&\int_{\mathbb{R}^d}\partial_t\varphi\,\rho\,\mathrm{d}x + \varepsilon^2\int_{\mathbb{R}^d}\mathrm{tr}\left(Q\,D^2\varphi\right)\rho\,\mathrm{d}x - \int_{\mathbb{R}^d}\nabla\varphi\cdot\nabla L\,\rho\,\mathrm{d}x\\
=&\int_{R\le|x|\le(1+\frac{\delta}{2})R}\left[\partial_t\varphi + \varepsilon^2\mathrm{tr}\left(Q\,D^2\varphi\right) - \nabla\varphi\cdot\nabla L\right]\rho\,\mathrm{d}x\\
&+\int_{(1+\frac{\delta}{2})R\le|x|\le(1+\delta)R}\left[\partial_t\varphi + \varepsilon^2\mathrm{tr}\left(Q\,D^2\varphi\right) - \nabla\varphi\cdot\nabla L\right]\rho\,\mathrm{d}x\\
=&I + II
\end{aligned}$$

Let us estimate the first term.

$$\begin{aligned}
I =& \int_{R\le|x|\le(1+\frac{\delta}{2})R}\left[\partial_t\varphi - \partial_r\varphi\frac{x}{|x|}\cdot\nabla L + \varepsilon^2\mathrm{tr}\left(Q\left(\partial_{rr}\varphi\,P_0 + \frac{\partial_r\varphi}{|x|}(I_{d\times d}-P_0)\right)\right)\right]\rho\,\mathrm{d}x\\
=&\frac{4}{\delta^2R^2}\int_{R\le|x|\le(1+\frac{\delta}{2})R}\left[(|x|-R)|x|\frac{R'}{R} + (|x|-R)\frac{x}{|x|}\cdot\nabla L\right]\rho\,\mathrm{d}x\\
&-\frac{4\varepsilon^2}{\delta^2R^2}\int_{R\le|x|\le(1+\frac{\delta}{2})R}\left[\mathrm{tr}\left(Q\,P_0\right) + \frac{(|x|-R)}{|x|}\mathrm{tr}\left(Q\left(I_{d\times d}-P_0\right)\right)\right]\rho\,\mathrm{d}x\,.
\end{aligned}$$

Using the λ-convexity of L and the bounds for the traces of the product of nonnegative matrices $0\le\mathrm{tr}(QP_0)\le\sigma d$ and $0\le\mathrm{tr}(Q(I_{d\times d}-P_0))\le\sigma d(d-1)$, we obtain

$$\begin{aligned}
I \ge& \frac{4}{\delta^2R^2}\int_{R\le|x|\le(1+\frac{\delta}{2})R}(|x|-R)|x|\left(\frac{R'}{R}+\frac{\lambda}{2}\right)\rho\,\mathrm{d}x\\
&-\frac{4\varepsilon^2}{\delta^2R^2}\int_{R\le|x|\le(1+\frac{\delta}{2})R}\left[\sigma d + \left(1-\frac{R}{|x|}\right)\sigma d(d-1)\right]\rho\,\mathrm{d}x\,.
\end{aligned}$$

Therefore, if the time dependent radius satisfies $R'(t) \geq -\frac{\lambda}{2}R(t)$, the first integral above is nonnegative. Moreover, since ρ is a probability measure we obtain the following lower bound for the second integral

$$I \geq -\frac{4\sigma d^2}{\delta^2}\frac{\varepsilon^2}{R^2(t)}\,. \tag{27}$$

For the integral on the other region, II, we follow the same argument. First, we use the explicit expressions of the derivatives of φ.

$$\begin{aligned}
II =& \int_{(1+\frac{\delta}{2})R\leq|x|\leq(1+\delta)R}\left[\partial_t\varphi - \partial_r\varphi\,\frac{x}{|x|}\cdot\nabla L + \varepsilon^2\mathrm{tr}\left(Q\left(\partial_{rr}\varphi\,P_0 + \frac{\partial_r\varphi}{|x|}(I_{d\times d}-P_0)\right)\right)\right]\rho\,\mathrm{d}x\\
=& \frac{4}{\delta^2R^2}\int_{(1+\frac{\delta}{2})R\leq|x|\leq(1+\delta)R}\left[\Big((1+\delta)R-|x|\Big)|x|\frac{R'}{R} + \Big((1+\delta)R-|x|\Big)\frac{x}{|x|}\cdot\nabla L\right]\rho\,\mathrm{d}x\\
&+\frac{4\varepsilon^2}{\delta^2R^2}\int_{(1+\frac{\delta}{2})R\leq|x|\leq(1+\delta)R}\left[\mathrm{tr}\left(Q\,P_0\right) - \frac{(1+\delta)R-|x|}{|x|}\mathrm{tr}\left(Q\left(I_{d\times d}-P_0\right)\right)\right]\rho\,\mathrm{d}x\,.
\end{aligned}$$

Then, we use the λ-convexity of L and $0 \leq \mathrm{tr}(QP_0) \leq \sigma d$ and $0 \leq \mathrm{tr}(Q(I_{d\times d}-P_0)) \leq \sigma d(d-1)$ to obtain that

$$\begin{aligned}
II \geq& \frac{4}{\delta^2R^2}\int_{(1+\frac{\delta}{2})R\leq|x|\leq(1+\delta)R}\Big((1+\delta)R-|x|\Big)|x|\left(\frac{R'}{R}+\frac{\lambda}{2}\right)\rho\,\mathrm{d}x\\
&-\frac{4\varepsilon^2}{\delta^2R^2}\int_{(1+\frac{\delta}{2})R\leq|x|\leq(1+\delta)R}\left(\frac{(1+\delta)R}{|x|}-1\right)\sigma d(d-1)\rho\,\mathrm{d}x\,.
\end{aligned}$$

Hence, the differential inequality for the time dependent radius $R'(t) \geq -\frac{\lambda}{2}R(t)$ implies that

$$II \geq -\frac{4(1+\delta)\sigma d^2}{\delta^2(1+\frac{\delta}{2})}\frac{\varepsilon^2}{R^2(t)}\,.$$

Finally, we conclude by combining both estimates of I and II,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\varphi(t,|x|)\rho(t,x)\,\mathrm{d}x = I + II \geq -C\frac{\varepsilon^2}{R(t)^2}\,,$$

with $C = \frac{4\sigma d^2}{\delta^2}(1+\frac{1+\delta}{1+\delta/2})$ and integrating the inequality in time on $[0,T]$. □

Note that the shrinking condition (25) of the radius $R(t)$ coincide with the concentration velocity of the flow map (23) associated to the continuity equation (22). The main idea is that if the radius decreases slower than the flow map concentrates, the mass in the shrinking balls will be almost preserved, up to an error depending on ε.

Despite the error term not being sharp, it is impossible to prove the estimate (26) with $C = 0$ due to the diffusive nature of the equation of ρ. Indeed, if we consider the case $Q(x) = I$ in $\mathbb{R}^2$ with a double well potential $L(x,y) = (x^2-1)^2 + y^2$, it is easy to see that even if the mass of the initial datum is concentrated in one of the local minimums, half of the mass will reach the neighbourhood of the other local minimum. A simple proof can be done studying the local mass of the unique steady state $\rho_\infty(x,y) = c\exp(-\frac{L(x,y)}{\varepsilon^2})$.

Now, we use Theorem 2.4 with a decreasing radius satisfying $R'(t) \geq -\frac{\lambda}{2}R(t)$ in order to obtain an explicit control of the error term. Thus, a natural choice $R(t)$ is the exponentially decreasing radius, which we have considered previously in the non diffusive case in Lemma 2.3 and satisfies condition (25) with equality.

Corollary 2.5. *Under assumptions of Theorem 2.4, if we choose the radius to be*

$$R(t) = R_0\, e^{-\frac{\lambda}{2}t}\,,$$

then for every $T > 0$ it holds that

$$\int_{\mathbb{R}^d} \varphi(T, |x|)\rho(T, x)\, \mathrm{d}x \geq \int_{\mathbb{R}^d} \varphi(0, |x|)\rho_0(x)\, \mathrm{d}x - \varepsilon^2 \frac{C}{R_0^2 \lambda} \left(e^{\lambda T} - 1\right)\,, \tag{28}$$

with $C > 0$ depending only on δ, σ and d.

Proof. Let us substitute the definition of the radius in the error term of (26).

$$-C \int_0^T \frac{\varepsilon^2}{R(t)^2}\, \mathrm{d}t = -\varepsilon^2 \frac{C}{R_0^2} \int_0^T e^{\lambda t}\, \mathrm{d}t = -\varepsilon^2 \frac{C}{R_0^2 \lambda} \left(e^{\lambda T} - 1\right)\,. \qquad \square$$

The above result gives us a lower bound for the mass concentrated around a ball centered in the local minimum of L. We are now in the position to prove Theorem 1.2.

Proof of Theorem 1.2. Letting $T = T_\varepsilon$ in (28), with $T_\varepsilon := \frac{2}{\lambda} \log\left(\frac{R_0}{a\varepsilon^\alpha}\right)$, gives that for every $0 < t < T_\varepsilon$ we have

$$\begin{aligned} \int_{\mathbb{R}^d} \varphi(t, |x|)\rho(t, x)\, \mathrm{d}x &\geq \int_{\mathbb{R}^d} \varphi(0, |x|)\rho_0(x)\, \mathrm{d}x - \varepsilon^2 \frac{C}{R_0^2 \lambda} \left(e^{\lambda T_\varepsilon} - 1\right) \\ &= \int_{\mathbb{R}^d} \varphi(0, |x|)\rho_0(x)\, \mathrm{d}x - \varepsilon^{2-2\alpha} \frac{C}{a^2 \lambda}\,. \end{aligned}$$

Clearly the error term can be made smaller than β by choosing $\varepsilon_0 > 0$ small enough, namely

$$\varepsilon_0^{2-2\alpha} \frac{C}{a^2 \lambda} \leq \beta\,.$$

This gives the desired estimate for all $\varepsilon \in (0, \varepsilon_0)$ and all $0 < t < T_\varepsilon$

$$\int_{\mathbb{R}^d} \varphi(t, |x|)\rho(t, x)\, \mathrm{d}x \geq \int_{\mathbb{R}^d} \varphi(0, |x|)\rho_0(x)\, \mathrm{d}x - \beta\,.$$

Finally, we notice that we can replace $t = 0$ by $t = t_0$ by the time shift invariance of the equation. $\square$

3 Analysis in the diffusion regime

After a first concentration phenomena near local minima of the loss function L, the diffusion will allow the parameters of the neural network to jump from one local minimum of L to another one. This property of the diffusion is the main advantage of using the SGD as a learning process with respect to the classical Gradient Descent, since it allows to avoid local minima of L which could be attractors at first steps. Thus, the main question that we want to address in this section is the following:

Q2: How long does SGD take to escape from a local minimum?

To this end, we focus on how the diffusion affects the dynamics of the learning process. In particular, we want to estimate the mean exit time from a neighborhood of a local minimum of the loss function. Lower bounds for the mean exit time can be obtained without further assumption. On the other hand, some non-degeneracy condition on the diffusion is needed in order to ensure that the mean exit time is finite. Indeed, as a counterexample, consider the case with no diffusion at all, presented in (22), which leads to the deterministic flow map

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = -\nabla L(\Phi_t) \cdot \Phi_t(x) \\ \Phi_0(x) = x\,. \end{cases}$$

In this case, all the mass splits and concentrates in the local minima of L, with no jump from one minimum to another one, i.e. the mean exit time is infinite.

Recall that the stochastic process describing the dynamics of the parameters when learning with SGD satisfies (2), that is,

$$\begin{cases} \mathrm{d}X_t &= -\nabla L(X_t)\,\mathrm{d}t + \sqrt{2\varepsilon^2 Q(X_t)}\mathrm{d}W_t \\ X_0 &= x\,, \end{cases}$$

with the degenerate but nonnegative matrix $Q(x) \geq 0$, i.e. $v^T Q(x) v \geq 0$ for every $v \in \mathbb{R}^d$ and every $x \in \mathbb{R}^d$. Due to the degeneracy of Q, obtaining estimates on the Mean Exit Time (MET) from a basin of attraction of L is not obvious. Recall also that the MET $u(x) = \mathbb{E}[\tau_\Omega^x]$ from an open set $\Omega \subset \mathbb{R}^d$ given an initial point $x \in \Omega$ satisfies (5), that is

$$\begin{cases} -\mathcal{A}u(x) = 1, & \text{in } \Omega \\ u(x) = 0, & \text{on } \partial\Omega \end{cases}$$

where

$$\mathcal{A}u(x) = \varepsilon^2 \mathrm{tr}\left(Q(x)D^2u(x)\right) - \nabla L(x) \cdot \nabla u(x).$$

This family of equations contains, as special cases, the non diffusive scenario ($Q(x) \equiv 0$, i.e. pure drift) and the non-degenerate diffusive scenario ($Q(x) \geq \delta I_{d\times d}$).

In the non-diffusive case, the mean exit time is $u(x) = +\infty$ for all $x \in \Omega$. Indeed, the flow map X_t is deterministic and describes the concentration of the parameters around the local minima of the loss function L. Hence, there is no possibility for X_t to escape from a neighbourhood of a local minimum of L in any time.

In the non-degenerate diffusive case, by Kramer's Law, $u(x) < +\infty$ for every $x \in \Omega$. Indeed, even if the drift term ∇L pushes X_t into the local minimum, the diffusion in every direction ensures the boundedness, see Appendix C.

In this section, we provide lower and upper bounds for the solution u to (5) under mild conditions on Q and L, allowing for degeneracies.

Theorem 3.1. *Let $\Omega = B_R(0) \subset \mathbb{R}^d$, let L be λ-convex in $B_R(0)$ with $\lambda > 0$ and a local minimum at 0, and assume that $0 \leq Q(x) \leq \sigma I_{d\times d}$ for every $x \in B_R(0)$. Then,*

$$w(x) = \frac{R^2 - |x|^2}{2\varepsilon^2 d\sigma}\,,$$

is a subsolution of (5), in the sense that $-\mathcal{A}w(x) \leq 1$ for all $x \in B_R(0)$.

Proof. Note that

$$\nabla w(x) = -\frac{1}{\varepsilon^2 \sigma d} x \qquad \text{and} \qquad D^2 w(x) = -\frac{1}{\varepsilon^2 \sigma d} I_{d\times d}\,.$$

Therefore,

$$-\varepsilon^2 \mathrm{tr}\left(QD^2 w\right) + \nabla L \cdot \nabla w = \frac{\mathrm{tr}\,(Q)}{\sigma d} - \frac{1}{\varepsilon^2 \sigma d} x \cdot \nabla L \leq \frac{\mathrm{tr}\,(Q)}{\sigma d} - \frac{\lambda}{2\varepsilon^2 \sigma d}|x|^2 \leq 1\,,$$

where we have use the λ-convexity of L in the first inequality and the bound $\mathrm{tr}(Q) \leq \sigma d$ in the second inequality. Since $w(x) = 0$ if $|x| = R$, we conclude that w is a subsolution of (5). □

In order to obtain an upper bound for the Mean Exit Time, some non-degeneracy in the diffusion is required. Indeed, it suffices to have a positive lower bound of Q in one direction and some smoothness condition in L. We begin by showing the simpler case with Theorem 3.2 in order to emphasize the main idea, and then prove the general case with Theorem 3.3. Note that the non degeneracy assumption on Q is essential, while the condition on L can be dropped in many cases, for instance on bounded domains.

Theorem 3.2. *Let Ω be an open set, let $L(x) \geq 0$ and $0 \leq Q(x) \leq \sigma I_{d\times d}$ for every $x \in \Omega$. Assume that there exits $\beta, \Lambda > 0$ and $i \in \{1, \dots, d\}$ such that*

$$(Q(x))_{ii} \geq \beta \qquad \textit{and} \qquad \partial_{x_i} L(x) x_i \leq \Lambda x_i^2 + \frac{\varepsilon^2 \beta}{2} \qquad \forall x \in \Omega\,.$$

Then, defining $R = \max\{|x_i| : x \in \Omega\}$ we have that

$$w(x) = \frac{2}{\Lambda}\left(e^{\frac{\Lambda R^2}{2\beta\varepsilon^2}} - e^{\frac{\Lambda x_i^2}{2\beta\varepsilon^2}}\right)\,,$$

is a supersolution of (5) in Ω, in the sense that $-\mathcal{A}w(x) \geq 1$ for all $x \in \Omega$.

Proof. Note that

$$\begin{aligned}
\partial_{x_i} w(x) &= -\frac{2}{\beta\varepsilon^2} x_i e^{\frac{\Lambda x_i^2}{2\beta\varepsilon^2}} \\
\partial_{x_i x_i} w(x) &= -\frac{2}{\beta\varepsilon^2} e^{\frac{\Lambda x_i^2}{2\beta\varepsilon^2}} \left(\frac{\Lambda}{\beta\varepsilon^2} x_i^2 + 1\right)\,.
\end{aligned}$$

Therefore, using the assumptions on Q and on L we obtain

$$\begin{aligned}
-\varepsilon^2 \mathrm{tr}\left(QD^2 w\right) + \nabla L \cdot \nabla w &= 2\frac{(Q(x))_{ii}}{\beta} e^{\frac{\Lambda x_i^2}{2\beta\varepsilon^2}} \left(\frac{\Lambda}{\beta\varepsilon^2} x_i^2 + 1\right) - \frac{2}{\beta\varepsilon^2} e^{\frac{\Lambda x_i^2}{2\beta\varepsilon^2}} \partial_{x_i} L(x) x_i \\
&\geq \frac{\Lambda}{\beta\varepsilon^2} x_i^2 + 2 - \frac{\Lambda}{\beta\varepsilon^2} x_i^2 - 1 = 1\,.
\end{aligned}$$

On the other hand, if $x \in \partial B_R(0)$ we have that $x_i \leq R$ and hence $w(x) \geq 0$. □

Note that we can generalize the result above whenever there exists a direction $v \in \mathbb{S}^{d-1}$ where Q is strictly elliptic in that direction.

Theorem 3.3. *Let Ω be an open set, let $0 \leq Q(x) \leq \sigma I_{d\times d}$ for every $x \in \Omega$. Assume that there exits $\beta, \Lambda > 0$, $i \in \{1, \dots, d\}$ and a vector $v \in \mathbb{S}^{d-1}$ such that*

$$v^T Q(x) v \geq \beta \qquad \text{and} \qquad (\nabla L(x) \cdot v)\,(v \cdot x) \leq \Lambda (v\cdot x)^2 + \frac{\varepsilon^2 \beta}{2} \qquad \forall x \in \Omega\,.$$

Then, defining $R = \max\{|v \cdot x| : x \in \Omega\}$ we have that

$$w(x) = \frac{2}{\Lambda}\left(e^{\frac{\Lambda R^2}{2\beta\varepsilon^2}} - e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} \right)\,,$$

is a supersolution of (5) in Ω.

Proof. Note that

$$\begin{aligned} \nabla w(x) &= -\frac{2}{\beta\varepsilon^2} e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} (v\cdot x) v \\ D^2 w(x) &= -\frac{2}{\beta\varepsilon^2} e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} \left(\frac{\Lambda}{\beta\varepsilon^2} (v\cdot x)^2 + 1 \right) v \otimes v\,. \end{aligned}$$

Therefore, using the assumptions on Q and on L along with $\mathrm{tr}(Qv\otimes v) = v^T Q v$, we obtain

$$\begin{aligned} -\varepsilon^2 \mathrm{tr}\left(QD^2 w\right) + \nabla L \cdot \nabla w =& 2\frac{\mathrm{tr}(Q(x) v\otimes v)}{\beta} e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} \left(\frac{\Lambda}{\beta\varepsilon^2}(v\cdot x)^2 + 1 \right) \\ & - \frac{2}{\beta\varepsilon^2} e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} (\nabla L(x)\cdot v)(v\cdot x)) \\ \geq & \frac{\Lambda}{\beta\varepsilon^2}(v\cdot x)^2 + 2 - \frac{\Lambda}{\beta\varepsilon^2}(v\cdot x)^2 - 1 = 1\,, \end{aligned}$$

since $e^{\frac{\Lambda (v\cdot x)^2}{2\beta\varepsilon^2}} \geq 1$. On the other hand, if $x \in \partial\Omega$ we have that $(v\cdot x)^2 \leq R^2$ and hence $w(x) \geq 0$. ☐

Thanks to the semigroup property of the Markovian process described by (2), we can use the subsolution and supersolution of Theorems 3.1 and 3.3 starting at any time $t_0 > 0$. This allows us to estimate the time that the learning process will spend after the concentration phenomena occurred in the drift regime.

We are now in the position to prove Theorem 1.3 and 1.4. Note that, in order to make use of the classical comparison principle [24] in the next proofs, we are requiring that the mean exit time is a viscosity solution of (5).

Proof of Theorem 1.3. Using the comparison principle for degenerate elliptic equations in [24, Theorem 3.3], we obtain that

$$u(x) = \mathbb{E}\left[\tau^x_{B_R(0)}\right] \geq w(x) = \frac{R_0^2 - |x|^2}{2\varepsilon^2 \sigma d}\,,$$

since u is a solution of (5) by construction and w is a subsolution of (5) by Theorem 3.1. Note that the Markovian property of the stochastic process (2) allows the time shifting of the initial datum. We conclude recalling that $|x| \leq r$. ☐

By similar a argument we can obtain an upper bound for the mean exit time.

Proof of Theorem 1.4 Using the comparison principle for degenerate elliptic equations in [24, Theorem 3.3], we obtain that

$$u(x) = \mathbb{E}\left[\tau_\Omega^x\right] \leq w(x) = \frac{2}{\Lambda}\left(e^{\frac{\Lambda R^2}{2\beta\varepsilon^2}} - e^{\frac{\Lambda(v\cdot x)^2}{2\beta\varepsilon^2}}\right),$$

since u is a solution of (5) by construction and w is a supersolution of (5) by Theorem 3.3. Note that the Markovian property of the stochastic process (2) allows the time shifting of the initial datum. We conclude recalling that $|v \cdot x| \leq |x| \leq r$. □

4 Asymptotic behaviour of the SGD

In this section, we analyze the asymptotic behaviour of the distribution of parameters of a neural network while training with SGD. Starting from an initial parameter configuration and a dataset, we aim to determine the final parameter probability distribution after prolonged SGD implementation. We will study this problem within the framework of the continuous stochastic process that approximates the SGD and its associated probability measure. The central question we will address is:

Q3: Does the distribution of neural network parameters converge using SGD?

To answer this question is delicate, and we propose two different method: duality and entropy methods. Both methods provide partial results about the convergence of the transition probability of the stochastic approximation to an invariant probability measure.

The duality method for a wide class of nondegenerate Fokker-Planck equations was introduced by Porretta in 2024 [58]. The first step consists of proving the time decay of oscillations of the solutions of the dual equation, often referred to as Orstein-Uhlenbeck or backward Kolmogorov. As a second step, this decay can be translated into a decay of the moments of the probability measure that solves the original Fokker-Planck equation. Finally, the third step consists of showing that the convergence of the moments implies the existence of an invariant probability measure and the convergence to it. The main drawback of this method is that, to the best of our knowledge, it only works with nondegenerate diffusion, i.e. $0 < \delta I \leq Q \leq \sigma I$ with $\sigma \geq \delta > 0$. In order to adapt this method to the present setting, we propose a variant of the SGD adding some random noise at each step of the learning process.

A nowadays more standard approach to study the asymptotic behavior of Fokker-Planck equations is the entropy method [10, 11], also known as the Bakry-Emery method. The relative entropy - a special Lyapunov function - quantifies how far is the probability at time t from the invariant probability measure. The first stability property, i.e. convergence to an invariant probability measure, can be proven by showing that the relative entropy decays to 0 as time goes to infinity. A more delicate analysis is then required to quantify the possible decay rates of the entropy. In general, even for the standard Fokker-Planck equation ($Q = I$ and $L(x) = |x|^2$), exponential decay holds only under additional assumptions, typically the boundedness of the first and second moment for the initial data (see e.g. [66, Sec. 4.1]). A quantitative exponential decay of the relative entropy can be proven by means of a weighted Poincaré type inequality which in turn is equivalent to the entropy-entropy production inequality.

In the present setting, the exponential decay follows by establishing a suitable Poincaré inequality obtained by using the results of [5] and [53]. The main limitations of this approach are that either the diffusion matrix must be non-degenerate with $Q \equiv \sigma I$ or, when Q is allowed to be degenerate, the loss function L must be quadratic.

4.1 Duality method for Noisy SGD

Let us consider a data set $\{(x_i, y_i)\}_{i=1}^N$ and the total loss function

$$L(\theta) = \frac{1}{N} \sum_{i=1}^{N} L_i(\theta)\,,$$

where each $L_i(\theta)$ for $i = 1, \dots, N$ is the partial loss function associated to the datum (x_i, y_i). Then, the stochastic gradient descent with constant learning rate $\eta > 0$ and batch size $|B_n| = \mathscr{b} > 0$ reads as follows

$$\theta_{n+1} = \theta_n - \frac{\eta}{\mathscr{b}} \sum_{b_i \in B_n} \nabla L_{b_i}(\theta_n) \qquad \forall n \geq 0\,, \tag{SGD}$$

with $B_n = \{b_i\}_{i=1}^{\mathscr{b}} \subset \{1, 2, \dots, N\}$ being a batch of indexes chosen with uniform distribution at each step n. We know by [45] and [26] that we can approximate (SGD) with the stochastic process defined by (2):

$$\mathrm{d}X_t = -\nabla L(X_t)\,\mathrm{d}t + \sqrt{\frac{\eta}{\mathscr{b}}}\Sigma(X_t)\mathrm{d}W_t\,,$$

where $Q(x) := \Sigma^2(x) = \mathrm{cov}\left[\nabla L_{b_i}(x)\right] \geq 0$ reads explicitly

$$Q(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla L_i \otimes \nabla L_i - \nabla L \otimes \nabla L.$$

Note that the noise in the above equation is typically degenerate: indeed, $\mathrm{rank}(\Sigma(x)) \leq N - 1$ and, in most real case scenarios, the number of parameters to be learned exceeds the size of the dataset, that is $N < d$.

In order to obtain a nondegenerate SDE, we propose another learning iteration with a suitable extra noise $\{Z_n\}_{n\geq 0}$, which we call Noisy Stochastic Gradient Descent:

$$\theta_{n+1} = \theta_n - \frac{\eta}{\mathscr{b}} \sum_{b_i \in B_n} \nabla L_{b_i}(\theta_n) + \eta Z_n. \tag{NSGD}$$

In what follows, our choice for $\{Z_n\}_{n\geq 0}$ is that of a family of i.i.d. Gaussian variables

$$Z_n \sim \mathcal{N}(0, \delta I_{d\times d}) \qquad \forall n \geq 0\,.$$

This extra noise does not depend on the characteristics of the learning process, and a priori may provide an invariant probability measure that is independent of the data set and the loss function. However, more sophisticated choices of noise could be considered, based on the loss function. We have chosen the simplest noise that allows to apply the results of [58], in the form of Theorem 4.2 below, indeed, our choice of noise leads us to the following nondegenerate SDE.

Lemma 4.1. *The process* (NSGD) *is the Euler-Maruyama approximation of*

$$dX_t = -\nabla L(X_t)\,\mathrm{d}t + \sqrt{\eta\, Q_\delta(X_t)}\,dW_t\,, \tag{29}$$

with $Q_\delta(x) := \mathscr{b}^{-1}Q(x) + \delta I_{d\times d}$.

Proof. The proof follows by a slight modification of the one in [44, Theorem 1]. See Appendix A for the details. □

By standard theory, we can associate to (29) the following Ornstein-Uhlenbeck type equation for $u(t,x) = \mathbb{E}[X_t]$

$$\begin{cases} \partial_t u = \frac{\eta}{2}\mathrm{tr}(Q_\delta(x)D^2u) - \nabla L(x)\cdot\nabla u =: -\mathcal{L}_\delta u \\ u(0,x) = u_0(x)\,, \end{cases} \tag{30}$$

Analogously, the probability density function of X_t satisfies the $L^2(\mathbb{R}^d)$-dual equation of (30), that is, the Fokker-Planck equation

$$\begin{cases} \partial_t \rho = \nabla\cdot\big(\frac{\eta}{2}\nabla\cdot(\rho\, Q_\delta(x)) + \rho\nabla L(x)\big) =: -\mathcal{L}_\delta^*\rho \\ \rho(0,x) = \rho_0(x)\,, \end{cases} \tag{31}$$

where $\nabla\cdot(\rho\, Q_\delta(x))$ is the divergence taken column-wise.

The goal is to prove convergence results for the distribution of the parameters θ in (NSGD). In order to do that, we prove that there exists a stationary state of equation (31) and the solution ρ converges to it as time goes to infinity. We are in the position to use the theory of [58] and establish the asymptotic behavior of the above non-degenerate Ornstein-Uhlenbeck and Fokker-Planck type equations, by exploiting the duality between (30) and (31), and proving that the weighted oscillation of u implies convergence of the moments of m. The weighted oscillation that we will consider is the following

$$[u]_{\langle\cdot\rangle^k} = \sup_{x,y\in\mathbb{R}^d} \frac{|u(x)-u(y)|}{\langle x\rangle^k + \langle y\rangle^k}\,,$$

with $\langle x\rangle^k = (1+|x|^2)^{k/2}$. This turns out to be a Lyapunov function of $\mathcal{L}_\delta$ (see [58]).

Theorem 4.2. [58, Theorem 3.5.] *Assume that the diffusion matrix* $Q_\delta(x) = \mathscr{b}^{-1}Q(x) + \delta I_{d\times d}$ *with* $\delta > 0$ *satisfies that there exist* $\sigma_0, \sigma_1 > 0$ *such that*

$$\|\sqrt{Q_\delta(\cdot)}\|_\infty \le \sigma_0 \text{ , and } \|\sqrt{Q_\delta(x)} - \sqrt{Q_\delta(y)}\| \le \sigma_1|x-y| \qquad \forall x,y\in\mathbb{R}^d\,. \tag{32}$$

Assume that the drift term $\nabla L(x)$ *satisfies that there exist* $\alpha, R > 0$ *and* $\gamma \ge 2$ *such that*

$$\nabla L(x)\cdot x \ge \alpha|x|^\gamma \qquad \forall x\in\mathbb{R}^d \quad \textit{with} \quad |x|\ge R, \tag{33}$$

and there exists $c_0 > 0$ *such that*

$$(\nabla L(x) - \nabla L(y))\cdot(x-y) \ge -c_0|x-y| \qquad \forall x,y\in\mathbb{R}^d\,. \tag{34}$$

If $\rho\in C([0,T):\mathcal{M}_k(\mathbb{R}^d))$ *is a viscosity solution of* (31)*, then*

$$\|\rho(t)\|_{\mathcal{M}_k} \le K_\delta\, e^{-\omega t}\|\rho_0\|_{\mathcal{M}_k} \qquad \forall t\in(0,T)\,, \tag{35}$$

provided that $\rho_0\in\mathcal{M}_k(\mathbb{R}^d)$ *and* $\int_{\mathbb{R}^d} d\rho_0(x) = 0$.

The above theorem is written for zero mean valued solutions. As shown in [58], it can be easily extended to the case of non-zero mean valued solutions with finite k-moments: the linearity of the equation allows to show both existence of a stationary solution ρ_∞^δ, and to show the convergence $\rho(t) \xrightarrow{t\to\infty} \rho_\infty^\delta$ in appropriate topologies, noticing that the mass of both ρ and ρ_∞^δ need to be the same.

Corollary 4.3. [58, Theorem 5.7] *Under the assumptions of Theorem 4.2, if $\rho_0 \in \mathcal{P}_k(\mathbb{R}^d)$, then there exists a unique stationary measure $\rho_\infty^\delta \in \mathcal{P}(\mathbb{R}^d)$ such that*

$$\mathcal{L}_\delta^* \rho_\infty^\delta = 0 \qquad in \quad \mathbb{R}^d . \tag{36}$$

Moreover, it holds that

$$\|\rho(t) - \rho_\infty^\delta\|_{\mathcal{P}_k} \le K_\delta\, e^{-\omega t} \qquad \forall t > 0 . \tag{37}$$

Remark (About convergence rates)**.**

(i) In the next section we will consider the limits as $\delta \to 0^+$. A careful examination of the proof of Theorem 4.2 reveals that the constant K_δ of formula (35) blows up when $\delta \to 0^+$. On the other hand, *the convergence rate $\omega > 0$ only depends on α and k, but not on δ.* However, it can not be computed explicitly because the proof is not constructive. Also, as far as we know, the above convergence results cannot be applied to the general framework of degenerate diffusion matrices, as described in [58].

(ii) Corollary 4.3 holds for the nondegenerate equation(31) and proves exponential convergence of solution $\rho(t)$ to the unique invariant measure ρ_∞ with the same mass. This holds whenever the growth of the potential L is (super)quadratic, namely $\gamma \ge 2$. A complementary result of [58], that we not consider here, allows to show the same convergence, but only with polynomial decay rates, in the case when $\gamma \in (0,2)$. Indeed, since we are considering approximations of Machine Learning processes, in practice, one can always modify L at "infinity" to have (super)quadratic growth.

(iii) To the best of our knowledge, convergence to a stationary measure in the case of degenerate Fokker-Planck type equations with a non-quadratic loss function L, has not been proved.

We conclude this section by recalling a technical lemma of [58], used in the proof of the above results, that we will need in what follows.

Lemma 4.4. *[58, Lemma 3.1.] If there exist $\alpha, \gamma, R > 0$ such that*

$$\nabla L(x) \cdot x \ge \alpha |x|^\gamma \qquad \forall x \in \mathbb{R}^d \quad with \quad |x| \ge R, \tag{38}$$

and there exists $\sigma < +\infty$ such that

$$\|Q_\delta(x)\|_\infty \le \sigma \qquad \forall x \in \mathbb{R}^d , \tag{39}$$

then, for every $\beta > 0$ there exists $K_\beta > 0$ such that

$$\mathcal{L}_\delta \langle x\rangle^k \ge (\alpha - \beta) k \,\langle x\rangle^{k+\gamma-2} - K_\beta \qquad \forall x \in \mathbb{R}^d \tag{40}$$

4.2 Existence of steady states: proof of Theorem 1.5

In this section we prove, under mild condition, the existence of at least one stationary measure for our Fokker-Planck type equations, as stated in Theorem 1.5, whose statement we recall here.

Consider the degenerate Fokker-Planck equation

$$\partial_t \rho = \nabla \cdot \left(\frac{\eta}{2\cancel{b}} \nabla \cdot (\rho Q(x)) + \rho \nabla L(x) \right) =: -\mathcal{L}^* \rho \,. \tag{41}$$

If (32), (33) and (34) hold, then there exists an invariant probability measure $\rho_\infty \in \mathcal{P}(\mathbb{R}^d)$ such that

$$\int_{\mathbb{R}^d} \mathcal{L}\varphi(x) d\rho_\infty(x) = 0 \qquad \forall \varphi \in C_b(\Omega) \,. \tag{42}$$

The idea of the proof is to consider the invariant measures of the nondegenerate equation with diffusion matrix $Q_\delta(x) = \cancel{b}^{-1} Q(x) + \delta I_{d\times d}$ and apply Prokhorov's Theorem. By Corollary (4.3), we know that for each $\delta > 0$ there exists $\rho_\infty^\delta \in \mathcal{P}(\mathbb{R}^d)$ such that

$$-\mathcal{L}_\delta^* \rho_\infty^\delta = \nabla \cdot \left(\frac{\eta}{2} \nabla \cdot (\rho_\infty^\delta Q_\delta(x)) + \rho_\infty^\delta \nabla L(x) \right) = 0 \,.$$

In order to prove that the sequence $\{\rho_\infty^\delta\}_{\delta>0}$ is tight we will use $\langle x\rangle^k$ as a Lyapunov function. Notice that for any $\varepsilon \in (0,1)$, by Lemma 4.4, it holds that

$$\mathcal{L}_\delta \langle x\rangle^k \geq k\langle x\rangle^{k-2}(\alpha |x|^\gamma - \bar{K}) \,,$$

with $\bar{K} := d(\sigma+1)(d+k-2)$. Hence, we can find two nonnegative functions $\phi, \chi \in C^2(\mathbb{R}^2)$ such that $\lim_{|x|\to\infty} \phi(x) = +\infty$, χ is compactly supported and

$$\mathcal{L}_\delta \langle x\rangle^k \geq \phi(x) - \chi(x) \qquad \forall \delta \in (0,1) \,. \tag{43}$$

Now, let us prove that $\{\rho_\infty^\delta\}_{0<\delta<1}$ is tight, that is, for any $\varepsilon > 0$ there exists $U_\varepsilon \subset \mathbb{R}^d$ such that $\int_{U_\varepsilon} \mathrm{d}\rho_\infty^\delta(x) > 1-\varepsilon$ for every $\delta \in (0,1)$. Using $\mathcal{L}_\delta^* \rho_\infty^\delta = 0$ and inequality (43), we obtain

$$0 = \int_{\mathbb{R}^d} \langle x\rangle^k \mathcal{L}_\delta^* \rho_\infty^\delta \, \mathrm{d}x = \int_{\mathbb{R}^d} \mathcal{L}_\delta \langle x\rangle^k \mathrm{d}\rho_\infty^\delta(x) \geq \int_{\mathbb{R}^d} (\phi(x) - \chi(x)) \mathrm{d}\rho_\infty^\delta(x) \,,$$

and hence

$$0 \leq \int_{\mathbb{R}^d} \phi(x) \mathrm{d}\rho_\infty^\delta(x) \leq \int_{\mathrm{supp}(\chi)} \chi(x) \mathrm{d}\rho_\infty^\delta(x) \leq \|\chi\|_\infty \qquad \forall \delta \in (0,1) \,.$$

On the other hand, since $\phi(x) \to +\infty$ when $|x| \to \infty$, for any $n > 0$ there exists $R_n > 0$ such that $\phi(x) \geq n$ for $x \in B_{R_n}$. Thus,

$$\int_{B_{R_n}^c} n \, \mathrm{d}\rho_\infty^\delta \leq \int_{B_{R_n}^c} \phi(x) \, \mathrm{d}\rho_\infty^\delta \leq \|\chi\|_\infty \qquad \Rightarrow \qquad \int_{B_{R_n}^c} \mathrm{d}\rho_\infty^\delta \leq \frac{\|\chi\|_\infty}{n} \,,$$

which implies

$$\int_{B_{R_n}} \mathrm{d}\rho_\infty^\delta > 1 - \frac{\|\chi\|_\infty}{n} \qquad \forall n > 0 \,.$$

Therefore, we can apply Prokhorov's Theorem to obtain that there exists $\rho_\infty \in \mathcal{P}(\mathbb{R}^d)$ such that $\rho_\infty^\delta \rightharpoonup \rho_\infty$ in $\mathcal{P}(\mathbb{R}^d)$ up to a subsequence.

It remains to prove that ρ_∞ solves the equation. For any $\varphi \in C_c^\infty(\mathbb{R}^d)$, it holds that

$$\begin{aligned} 0 &\le \left| \int_{\mathbb{R}^d} \mathcal{L}\varphi(x) \mathrm{d}\rho_\infty(x) \right| = \left| \int_{\mathbb{R}^d} \mathcal{L}\varphi(x) \mathrm{d}\rho_\infty(x) - \int_{\mathbb{R}^d} \mathcal{L}_\delta\varphi(x) \mathrm{d}\rho_\infty^\delta(x) \right| \\ &\le \left| \int_{\mathbb{R}^d} \mathcal{L}\varphi(x) \, \mathrm{d}(\rho_\infty - \rho_\infty^\delta)(x) \right| + \left| \int_{\mathbb{R}^d} (\mathcal{L}\varphi(x) - \mathcal{L}_\delta\varphi(x)) \mathrm{d}\rho_\infty^\delta(x) \right| \\ &\le \langle \mathcal{L}\varphi, \rho_\infty - \rho_\infty^\delta \rangle_{C_b(\mathbb{R}^d), \mathcal{M}(\mathbb{R})} + \|\mathcal{L}\varphi - \mathcal{L}_\delta\varphi\|_{C_b(\mathbb{R}^d)} \overset{\delta\to 0^+}{\longrightarrow} 0 \,, \end{aligned}$$

which concludes the proof, since $\|\mathcal{L}\varphi - \mathcal{L}_\delta\varphi\|_{C_b(\mathbb{R}^d)} \le C\,\delta \|\varphi\|_{C^2(\mathbb{R}^d)}$, for some $C > 0$. □

Existence of stationary states -possibly measures- for general Fokker-Plank equations nowadays is a classical topic, we refer for instance to Chapter 2 of [19] and references therein. In particular, similar results have been obtained in [19, Corollary 2.4.4], with slightly different assumptions and proofs. Uniqueness is in general a delicate issue and in general may turn out to be false.

4.3 The question of convergence to stationary measures

Once existence of stationary measures is established, the next natural question would be which solutions converge to it (for large times), and this will be explored in the next section. In general we cannot expect better than convergence in a topology compatible with measures, since stationary solutions may happen to be purely measures (not functions), in view of the Examples 6 and 7 below, see also Example 8 in the next section. Also, since the diffusion matrix can be highly degenerate, so that, heuristically, part of the flow can be driven by "pure transport", in which case, the stationary solution can be a "pure measure". This motivates the following example.

Example 6. Let us consider the following transport equation

$$\begin{cases} \partial_t \rho(t,x) &= \nabla \cdot (Cx\rho(t,x)) \qquad \text{in} \quad (0,\infty) \times \mathbb{R}^d \,, \\ \rho(0,x) &= \rho_0(x) \qquad\qquad \text{in} \quad \mathbb{R}^d \,, \end{cases} \tag{44}$$

where $C \in \mathbb{R}^{d\times d}$ with $C > 0$. Notice that $\rho_\infty(x) = \delta_0(x)$ is a stationary measure since if we multiply by a test function $\varphi \in C_c^1(\mathbb{R}^d)$ and we integrate, it holds that

$$\int_{\mathbb{R}^d} x \cdot \nabla\varphi(x) \, \mathrm{d}\delta_0(x) = x \cdot \nabla\varphi(x) \Big|_{x=0} = 0 \qquad\qquad \forall \varphi \in C_c^1(\mathbb{R}^d) \,.$$

The equation above has the following explicit solution $\rho(t,x) = e^{\mathrm{tr}(C)t}\rho_0(e^{Ct}x)$, since

$$\partial_t \rho(t,x) = \mathrm{tr}(C) e^{\mathrm{tr}(C)t} \rho_0(e^{Ct}x) + e^{\mathrm{tr}(C)t+t} \left((Cx) \cdot \nabla\rho_0(e^{Ct}x) \right) = \nabla \cdot (Cx\rho(t,x)) \,.$$

Notice that the solution $\rho(t,x)$ preserve the mass: using the change of variable

$$y = e^{Ct}x \qquad \text{and} \qquad \mathrm{d}y = \left|\det e^{Ct}\right| \, \mathrm{d}x = e^{\mathrm{tr}(C)t} \, \mathrm{d}x \,,$$

we obtain

$$\int_{\mathbb{R}^d} \rho(t,x) \, \mathrm{d}x = \int_{\mathbb{R}^d} e^{\mathrm{tr}(C)t} \rho_0(e^{Ct}x) \, \mathrm{d}x = \int_{\mathbb{R}^d} \rho_0(y) \, \mathrm{d}y \qquad \forall t > 0 \,.$$

Let us show the convergence of the solutions of (44) to δ_0 whenever $\rho_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Let us consider the equation above as a gradient flow for the second moments in the Wasserstein space, namely,

$$\begin{cases} \partial_t \rho(t,x) &= -\mathrm{grad}_{W_2}\mathcal{E}[\rho] := \nabla \cdot \left(\rho(t,x)\nabla\left(\frac{\delta\mathcal{E}}{\delta\rho}[\rho]\right)\right) \\ \rho(0,x) &= \rho_0(x) \in \mathcal{P}_2(\mathbb{R}^d)\,. \end{cases}$$

with

$$\mathcal{E}[\rho] := \frac{1}{2}\int_{\mathbb{R}^d} x^T\, C^T C\, x \,\mathrm{d}\rho(x) \qquad \& \qquad \frac{\delta\mathcal{E}}{\delta\rho}[\rho] = Cx\,.$$

Notice that the global minimum of the energy is attainted at $\mathcal{E}[\delta_0] = 0$. As it can be seen in [28, Lemma 4.4.3] since $\mathcal{E}$ is λ-convex with some $\lambda > 0$ (in particular, for any $\lambda \in (0, \lambda_1(C^TC))$, it holds that for every $\rho \in \mathcal{P}_2(\mathbb{R}^d)$

$$\frac{\lambda}{2}W_2^2(\rho,\delta_0) \le \mathcal{E}[\rho] - \mathcal{E}[\delta_0] \le \frac{1}{2\lambda}\langle \mathrm{grad}_{W_2}\mathcal{E}[\rho], \mathrm{grad}_{W_2}\mathcal{E}[\rho]\rangle_\rho\,.$$

This estimates above for the λ-convex energy are the analogous of

$$\varphi(x) - \varphi(x_0) \ge \frac{\lambda}{2}|x - x_0|^2 \qquad \& \qquad |\nabla\varphi(x)|^2 \ge 2\lambda\left(\varphi(x) - \varphi(x_0)\right)\,,$$

whenever $x_0 \in \mathbb{R}^d$ is the unique minimum of a λ-convex function $\varphi \in C^1(\mathbb{R}^d)$.

Hence, following equation (4.18) of [28] we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}[\rho] = \langle \mathrm{grad}_{W_2}\mathcal{E}[\rho], \partial_t\rho\rangle_{W_2} = -\langle \mathrm{grad}_{W_2}\mathcal{E}[\rho], \mathrm{grad}_{W_2}\mathcal{E}[\rho]\rangle_\rho \le -2\lambda\mathcal{E}[\rho]\,,$$

which implies the decay of the energy

$$\mathcal{E}[\rho] \le e^{-2\lambda t}\mathcal{E}[\rho_0]\,.$$

Indeed, since $\mathcal{E}[\delta_0] = 0$, we have

$$\frac{\lambda}{2}W_2^2(\rho,\delta_0) \le \mathcal{E}[\rho] - \mathcal{E}[\delta_0] = \mathcal{E}[\rho] \le e^{-2\lambda t}\mathcal{E}[\rho_0]\,.$$

Thus, there is exponential Wasserstein convergence to the Dirac's delta at zero. In conclusion, if we assume a condition on the decay of the initial datum $\rho_0(x)$ when $|x| \to \infty$, such as finite second moments, then the solution converge to δ_0.

A similar phenomena can also happen in the case of non-zero diffusion matrix (still very degenerate), as the next example shows. See also Example 8 in the next section for a generalization to higher dimensions.

Example 7. Let us analyze the fundamental solution of the following degenerate Fokker-Planck equation:

$$\partial_t u = \nabla \cdot \left(\begin{pmatrix} 1 & 0 \\ 0 & 0\end{pmatrix}\nabla u + u\begin{pmatrix} 1 & 0 \\ 0 & 1\end{pmatrix}\begin{pmatrix} x \\ y\end{pmatrix}\right) \qquad \text{in} \quad (0,\infty)\times\mathbb{R}^2\,. \tag{45}$$

In this case, condition (II) is not satisfied since

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}}_{Q_0} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{C} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

This orthogonality between the diffusion direction and the pure drift direction suggest that the fundamental solution should be of the form

$$H(t,x) = g(t,x)h(t,y)\,,$$

where g is the fundamental solution for the one dimensional problem

$$\partial_t g = \partial_{x_1}\left(\partial_x g + xg\right) \qquad \text{in} \quad (0,\infty)\times\mathbb{R}\,, \tag{46}$$

and h solves the one dimensional transport equation

$$\partial_t h = \partial_y\left(yh\right) \qquad \text{in} \quad (0,\infty)\times\mathbb{R}\,. \tag{47}$$

Using the explicit solutions of (46) and (47), we conclude that the fundamental solution of (45) is

$$H(t,x,y) = \underbrace{\frac{2}{\sqrt{2\pi}(1-e^{-2t})}\, e^{-\frac{2\,x^2}{(1-e^{-2t})}}}_{g(t,x)}\; \underbrace{e^t\delta_0(e^t y)}_{h(t,y)}\,.$$

Thus, if we want to obtain the solution of the Cauchy problem associated to (54) with initial datum u_0 regular enough, we have just to convolve u_0 with the fundamental solution:

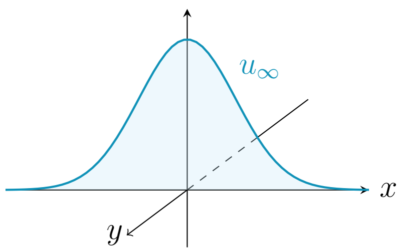
$$\begin{aligned} u(t,x,y) &= \big(H(t,\cdot,\cdot)*u_0\big)(x,y) = \int_{\mathbb{R}}\int_{\mathbb{R}} g(t,x-\tilde{x})h(t,y-\tilde{y})u_0(\tilde{x},\tilde{y})\mathrm{d}\tilde{x}\mathrm{d}\tilde{y} \\ &= e^t\int_{\mathbb{R}} g(t,x-\tilde{x})u_0(\tilde{x},e^t y)\mathrm{d}\tilde{x} \end{aligned}$$

Notice that for the asymptotic behaviour is not clear what should be the steady states. On the one hand, fixed $y\in\mathbb{R}$, $\big(g(t,\cdot)*u_0(\cdot,y)\big)(x)$ converge to the gaussian

$$g_\infty(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,.$$

But on the other hand, if we fix $x\in\mathbb{R}$, then $\big(h(t,\cdot)*u_0(x,\cdot)\big)(y)$ converge to $\delta_0(y)$ as a measure. Therefore, our educated guess for the steady state is

$$u_\infty(x,y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,\delta_0(y)\,.$$



Lemma 4.5. *Let u be a solution of (45) with initial datum $u_0 \in L^1(\mathbb{R}^2) \cap \mathcal{P}_2(\mathbb{R}^2)$. Then,*

$$u(t) \to u_\infty \qquad \text{as } t \to \infty\,,$$

in 2-Wasserstein sense.

Proof. As shown in [2, Theorem 2.7], we can prove the 2-Wasserstein convergence by

i) Convergence of second moments:

$$\int_{\mathbb{R}^2} (|x|^2 + |y|^2) u(t,x,y)\, \mathrm{d}x\, \mathrm{d}y \overset{t\to\infty}{\longrightarrow} \int_{\mathbb{R}^2} (|x|^2 + |y|^2) u_\infty(x,y)\, \mathrm{d}x\, \mathrm{d}y\,.$$

ii) Weak convergence of measures: $\forall \varphi \in \mathrm{Lip}(\mathbb{R}^2) \cap L^\infty(\mathbb{R}^2)$

$$\int_{\mathbb{R}^2} \varphi(x,y) u(t,x,y)\, \mathrm{d}x\, \mathrm{d}y \overset{t\to\infty}{\longrightarrow} \int_{\mathbb{R}^2} \varphi(x,y) u_\infty(x,y)\, \mathrm{d}x\, \mathrm{d}y\,.$$

Step 1. *Convergence of second moments.* Note that the second moment of u_∞ is

$$\int_{\mathbb{R}^2} \left(|x|^2 + |y|^2\right) u_\infty(x,y)\, \mathrm{d}x\, \mathrm{d}y = \int_{\mathbb{R}} |x|^2 g_\infty(x)\, \mathrm{d}x = 1\,.$$

Then, it suffices to prove that the following tends to 0:

$$\begin{aligned}
&\left|\int_{\mathbb{R}^2} \left(|x|^2 + |y|^2\right) u(t,x,y)\, \mathrm{d}x\, \mathrm{d}y - 1\right| \\
=&\left|\int_{\mathbb{R}^2} \left(|x|^2 + |y|^2\right) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) e^t u_0(\tilde{x}, e^t y) \mathrm{d}\tilde{x}\right) \mathrm{d}x\, \mathrm{d}y - 1\right| \\
\leq&\left|\int_{\mathbb{R}^2} |x|^2 \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) e^t u_0(\tilde{x}, e^t y) \mathrm{d}\tilde{x}\right) \mathrm{d}x\, \mathrm{d}y - 1\right| \\
&+ \left|\int_{\mathbb{R}^2} |y|^2 \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) e^t u_0(\tilde{x}, e^t y) \mathrm{d}\tilde{x}\right) \mathrm{d}x\, \mathrm{d}y\right| \\
\leq&\left|\int_{\mathbb{R}} \left(\int_{\mathbb{R}} |x|^2 g(t, x-\tilde{x})\, \mathrm{d}x\right) \left(\int_{\mathbb{R}} u_0(\tilde{x}, y)\, \mathrm{d}y\right) \mathrm{d}\tilde{x} - 1\right| \\
&+ e^{-2t} \left|\int_{\mathbb{R}^2} g(t, x-\tilde{x}) \left(\int_{\mathbb{R}} |y|^2 u_0(\tilde{x}, y)\, \mathrm{d}y\right) \mathrm{d}\tilde{x}\, \mathrm{d}x\right| \\
=&: I + II\,,
\end{aligned}$$

with

$$u_0^X(\tilde{x}) = \int_{\mathbb{R}} u_0(\tilde{x}, y)\, \mathrm{d}y\,.$$

In order to estimates I, note that

$$\int_{\mathbb{R}} |x|^2 g(t, x-\tilde{x})\, \mathrm{d}x = 1 + e^{-2t}(|\tilde{x}|^2 - 1)\,.$$

Since $u_0^X \in \mathcal{P}_2(\mathbb{R})$, it holds that

$$\begin{aligned}
I &= \left| \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |x|^2 g(t, x-\tilde{x})\,\mathrm{d}x \right) u_0^X(\tilde{x})\mathrm{d}\tilde{x} - 1 \right| \\
&= \left| \int_{\mathbb{R}} \left(1 + e^{-2t}(|\tilde{x}|^2 - 1)\right) u_0^X(\tilde{x})\mathrm{d}\tilde{x} - 1 \right| \\
&\le e^{-2t} \left(\int_{\mathbb{R}^d} (|x|^2 + |y|^2) u_0(x,y)\,\mathrm{d}x\,\mathrm{d}y + 1 \right) \overset{t\to\infty}{\longrightarrow} 0\,.
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
II &= e^{-2t} \left| \int_{\mathbb{R}} \left(\int_{\mathbb{R}} g(t, x-\tilde{x})\,\mathrm{d}x \right) \left(\int_{\mathbb{R}} |y|^2 u_0(\tilde{x}, y)\,\mathrm{d}y \right) \mathrm{d}\tilde{x} \right| \\
&= e^{-2t} \left| \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |y|^2 u_0(\tilde{x}, y)\,\mathrm{d}y \right) \mathrm{d}\tilde{x} \right| \\
&\le e^{-2t} \left(\int_{\mathbb{R}^2} (|\tilde{x}|^2 + |y|^2) u_0(\tilde{x}, y)\,\mathrm{d}y \right) \overset{t\to\infty}{\longrightarrow} 0\,,
\end{aligned}$$

and the convergence of the second moments follows.
Step 2. *Weak convergence of measures.* For any $\varphi \in \mathrm{Lip}(\mathbb{R}^2)$ we want to prove that the following quantity tends to 0 as t goes to ∞:

$$\left| \int_{\mathbb{R}^2} \varphi(x,y) \left(u(t,x,y) - u_\infty(x,y)\right)\,\mathrm{d}x\,\mathrm{d}y \right|.$$

By substituting the explicit expression of u and u_∞, we obtain

$$\begin{aligned}
&\left| \int_{\mathbb{R}^2} \varphi(x,y) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) e^t u_0(\tilde{x}, e^t y)\mathrm{d}\tilde{x} - g_\infty(x)\delta_0(y) \right) \mathrm{d}x\,\mathrm{d}y \right| \\
=& \left| \int_{\mathbb{R}^2} \left(\varphi(x, e^{-t}y) - \varphi(x,0) + \varphi(x,0)\right) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) u_0(\tilde{x}, y)\mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \right. \\
&\left. - \int_{\mathbb{R}} \varphi(x,0) g_\infty(x)\,\mathrm{d}x \right| \\
=& \left| \int_{\mathbb{R}^2} \left(\int_0^{e^{-t}y} \partial_r \varphi(x,r)\mathrm{d}r \right) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) u_0(\tilde{x}, y)\mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \right. \\
&\left. - \int_{\mathbb{R}} \varphi(x,0) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) u_0^X(\tilde{x})\mathrm{d}\tilde{x} - g_\infty(x) \right) \mathrm{d}x \right| \\
\le& A + B\,,
\end{aligned}$$

where

$$\begin{aligned}
A &= \left| \int_{\mathbb{R}^2} \left(\int_0^{e^{-t}y} \partial_r \varphi(x,r)\mathrm{d}r \right) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) u_0(\tilde{x}, y)\mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \right| \\
&\le e^{-t} \|\nabla\varphi\|_{L^\infty(\mathbb{R}^2)} \int_{\mathbb{R}^2} g(t, x-\tilde{x}) \left(\int_{\mathbb{R}} |y| u_0(\tilde{x}, y)\,\mathrm{d}y \right) \mathrm{d}\tilde{x}\,\mathrm{d}x \\
&\le e^{-t} \|\nabla\varphi\|_{L^\infty(\mathbb{R}^2)} \left(\int_{\mathbb{R}^2} (|x| + |y|) u_0(x,y)\,\mathrm{d}x\,\mathrm{d}y \right) \overset{t\to\infty}{\longrightarrow} 0
\end{aligned}$$

and

$$
\begin{aligned}
B &= \left| \int_{\mathbb{R}} \varphi(x,0) \left(\int_{\mathbb{R}} g(t, x-\tilde{x}) u_0^X(\tilde{x}) \mathrm{d}\tilde{x} - g_\infty(x) \right) \mathrm{d}x \right| \\
&\leq \left(\int_{\mathbb{R}} |\varphi(x,0)|^2 g_\infty(x) \, \mathrm{d}x \right)^{1/2} \left(\int_{\mathbb{R}} \left[u^X(t,x) - g_\infty(x) \right]^2 g_\infty^{-1}(x) \, \mathrm{d}x \right)^{1/2} \\
&= \left(\int_{\mathbb{R}} |\varphi(x,0)|^2 g_\infty(x) \, \mathrm{d}x \right)^{1/2} \| u^X(t) - g_\infty \|_{L^2(g_\infty^{-1})} \overset{t\to\infty}{\longrightarrow} 0 \,,
\end{aligned}
$$

with $u^X(t,x) = \int_{\mathbb{R}} g(t, x-\tilde{x}) u_0^X(\tilde{x}) \mathrm{d}\tilde{x}$ being the solution to (46) with initial datum u_0^X. $\square$

4.4 Entropy method

Once we have established the existence of a stationary measure (in the general case) we address the question of convergence towards it, possibly with rates. As already mentioned, we shall use entropy methods, and for this reason we shall assume that the stationary measure ρ_∞ is indeed an L^1 function, that we fix throughout this section. We explore possible adaptations to our problem of the classical entropy method introduced by Bakry and Émery in [10], see also [11]. Recall the stochastic approximation of order 1 of the SGD defined by

$$
\mathrm{d}X_t = -\nabla L(X_t) \, \mathrm{d}t + \sqrt{\frac{\eta}{\cancel{b}}} \Sigma(X_t) dW_t \,,
$$

together with the following Fokker-Planck equation for the associated probability density function ρ:

$$
\begin{cases}
\partial_t \rho &= \nabla \cdot \Big(\varepsilon^2 \nabla \cdot (Q(x)\rho) + \rho \nabla L(x) \Big) \\
\rho(0,\cdot) &= \rho_0 \in L^2(\mathbb{R}^d, \rho_\infty^{-1} \, \mathrm{d}x).
\end{cases}
\tag{48}
$$

where $\varepsilon^2 = \frac{\eta}{2\cancel{b}}$. Entropy methods aim at obtaining differential inequalities between the entropy functional and its time derivative (entropy production), by means of suitable functional inequalities, typically of weighted Poincaré type. There may be several possible entropies for the above equation, however the following choice seems the most appropriate.

Lemma 4.6 (Entropy and Entropy production). *Let us consider the following relative entropy associated to ρ_∞,*

$$
\mathcal{E}(\mu \,|\, \rho_\infty) = \frac{1}{2} \int_{\mathbb{R}^d} \left(\frac{\mu(x)}{\rho_\infty(x)} - 1 \right)^2 \rho_\infty(x) \, \mathrm{d}x \,.
$$

Then, its derivative along the Fokker-Planck flow (48) is given by the Fisher information, or entropy production:

$$
\mathcal{I}(\mu \,|\, \rho_\infty) = \int_{\mathbb{R}^d} \nabla \left(\frac{\mu(x)}{\rho_\infty(x)} \right)^T Q(x) \nabla \left(\frac{\mu(x)}{\rho_\infty(x)} \right) \rho_\infty(x) \, \mathrm{d}x
$$

More precisely,

$$
\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(\rho(t) | \rho_\infty) = -\varepsilon^2 \mathcal{I}(\rho(t) \,|\, \rho_\infty) \,, \qquad \textit{for all } t > 0. \tag{49}
$$

Remark. An immediate consequence of the assumption $\rho_0 \in L^2(\mathbb{R}^d, \rho_\infty^{-1}\,\mathrm{d}x)$, is that the entropy is finite for all times, indeed $\mathcal{E}(\rho_0|\rho_\infty) < \infty$, hence the above lemma implies $\mathcal{E}(\rho(t)|\rho_\infty) < \infty$ for all $t > 0$.

Proof. Let us differentiate the entropy and use the equation of ρ to obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\rho(t)\,|\,\rho_\infty) &= \int_{\mathbb{R}^d} \partial_t \rho(t) \left(\frac{\rho(t)}{\rho_\infty} - 1\right) \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \nabla \cdot \left(\varepsilon^2 Q \nabla \rho(t) + \rho(t) \left(\nabla L + \varepsilon^2 \nabla \cdot Q\right)\right) \left(\frac{\rho(t)}{\rho_\infty} - 1\right) \mathrm{d}x \\
&= \int_{\mathbb{R}^d} \nabla \cdot \left(\rho_\infty \varepsilon^2 Q \nabla \left(\frac{\rho(t)}{\rho_\infty}\right)\right) \left(\frac{\rho(t)}{\rho_\infty} - 1\right) \mathrm{d}x \\
&\quad + \int_{\mathbb{R}^d} \nabla \cdot \left[\rho(t) \left(\varepsilon^2 Q \frac{\nabla \rho_\infty}{\rho_\infty} + \varepsilon^2 \nabla \cdot Q + \nabla L\right)\right] \left(\frac{\rho(t)}{\rho_\infty} - 1\right) \mathrm{d}x\,.
\end{aligned}
$$

After integration by parts in the second integral above, we have that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\rho(t)|\rho_\infty) &= -\,\varepsilon^2 \mathcal{I}(\rho(t)\,|\,\rho_\infty) - \int_{\mathbb{R}^d} \left(\varepsilon^2 Q \frac{\nabla \rho_\infty}{\rho_\infty} + \varepsilon^2 \nabla \cdot Q + \nabla L\right) \cdot \nabla \left(\frac{\rho(t)}{\rho_\infty}\right) \rho(t)\,\mathrm{d}x \\
&= -\,\varepsilon^2 \mathcal{I}(\rho(t)\,|\,\rho_\infty) - \frac{1}{2} \int_{\mathbb{R}^d} \left(\varepsilon^2 Q \frac{\nabla \rho_\infty}{\rho_\infty} + \varepsilon^2 \nabla \cdot Q + \nabla L\right) \cdot \nabla \left(\frac{\rho^2(t)}{\rho_\infty^2}\right) \rho_\infty \mathrm{d}x \\
&= -\,\varepsilon^2 \mathcal{I}(\rho(t)\,|\,\rho_\infty) - \frac{1}{2} \int_{\mathbb{R}^d} \left(\varepsilon^2 Q \nabla \rho_\infty + \rho_\infty \left(\varepsilon^2 \nabla \cdot Q + \nabla L\right)\right) \cdot \nabla \left(\frac{\rho^2(t)}{\rho_\infty^2}\right) \mathrm{d}x \\
&= -\,\varepsilon^2 \mathcal{I}(\rho(t)\,|\,\rho_\infty) + \frac{1}{2} \int_{\mathbb{R}^d} \nabla \cdot \left(\varepsilon^2 \nabla \cdot (Q \rho_\infty) + \rho_\infty \nabla L\right) \frac{\rho^2(t)}{\rho_\infty^2}\,\mathrm{d}x\,.
\end{aligned}
$$

Note that the second term vanishes since ρ_∞ is a steady state, satisfying equation (42). □

In order to establish a differential inequality for the entropy, we need a relationship between the entropy and its entropy production, which typically takes the form of a weighted Poincaré inequality associated to (48).

Definition 4.7 (Poincaré inequality)**.** Let $Q(x) \geq 0$ in the sense of matrices, for every $x \in \mathbb{R}^d$. We say that a *Poincaré inequality* holds with respect to $\rho_\infty \in \mathcal{P}(\mathbb{R}^d)$ if there exists $\lambda > 0$ such that, for all $f \in L^2(\mathbb{R}^d, \rho_\infty\,\mathrm{d}x)$

$$
\lambda \int_{\mathbb{R}^d} \left(f - \int_{\mathbb{R}^d} f \rho_\infty\,\mathrm{d}x\right)^2 \rho_\infty\,\mathrm{d}x \leq \varepsilon^2 \int_{\mathbb{R}^d} \left(\nabla f^T Q(x) \nabla f\right) \rho_\infty\,\mathrm{d}x. \tag{50}
$$

Indeed, letting $f = \mu/\rho_\infty$ with $\int_{\mathbb{R}^d} \mu(x)\,\mathrm{d}x = 1$, the above inequality reads as the entropy-entropy production inequality

$$
\mathcal{E}(\mu|\rho_\infty) \leq \frac{\varepsilon^2}{\lambda} \mathcal{I}\left(\mu\,|\,\rho_\infty\right). \tag{51}
$$

Note that $f \in L^2(\mathbb{R}^d, \rho_\infty\,\mathrm{d}x)$ if and only if $\mu \in L^2(\mathbb{R}^d, \rho_\infty^{-1}\,\mathrm{d}x)$.

The above Poincaré inequality is relevant because it allows to derive exponential decay of the entropy, and show convergence towards ρ_∞ with precise rates, given by $\lambda > 0$.

Proposition 4.8. *A Poincaré inequality rewritten in the form* (51) *holds for any* $\mu \in L^2(\mathbb{R}^d, \rho_\infty^{-1}\,\mathrm{d}x)$ *if and only if there is an exponential convergence (with rate* $\lambda > 0$*) of the relative entropy between the solution* $\rho(t)$ *of* (48) *starting at* $\rho_0 \in L^2(\mathbb{R}^d, \rho_\infty^{-1}\,\mathrm{d}x)$ *and the steady state* ρ_∞*, that is*

$$\mathcal{E}\left(\rho(t)\,|\,\rho_\infty\right) \leq e^{-\lambda t}\mathcal{E}\left(\rho_0\,|\,\rho_\infty\right) \qquad \forall t \geq 0\,. \tag{52}$$

Proof. One implication follows by using the Poincaré inequality with $f = \rho(t)/\rho_\infty$ in the equivalent form (51) to obtain a closed differential inequality

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\rho(t)\,|\,\rho_\infty) = -\varepsilon^2\mathcal{I}(\rho(t)\,|\,\rho_\infty) \leq -\lambda\mathcal{E}(\rho(t)\,|\,\rho_\infty)\,.$$

The result follows by integrating the above inequality in $[0,t]$.

The other implication, follows by differentiating in time (52) and using (49):

$$-\varepsilon^2\mathcal{I}(\rho(t)\,|\,\rho_\infty) = \frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\rho(t)\,|\,\rho_\infty) \leq -\lambda e^{-\lambda t}\mathcal{E}(\rho_0\,|\,\rho_\infty)\,,$$

Finally, the Poincaré inequality (50) for $f = \rho_0/\rho_\infty$, follows by choosing $t = 0$ and $\rho_0 = \mu$. ◻

Proving weighted Poincaré inequalities for a general nonnegative matrix Q and a general loss function L is a difficult challenge, in particular since the form of ρ_∞ is not explicit (except some special cases). To the best of our knowledge, only very partial results, under quite strong conditions, appear in the literature. We collect them hereafter, adapted to our notations.

Theorem 4.9. *[53] Assume that* $Q(x) = \sigma I_{d\times d}$ *for every* $x \in \mathbb{R}^d$ *and* L *is a nonnegative Morse function satisfying*

$$\liminf_{|x|\to\infty} |\nabla L(x)| \geq c_1\,,$$
$$\liminf_{|x|\to\infty} \left(|\nabla L(x)|^2 - \Delta L(x)\right) \geq -c_2\,,$$

for some $c_1, c_2 > 0$*. Then the unique invariant probability measure of* (48) *is*

$$\rho_\infty(x) = c\,e^{-\frac{L(x)}{\varepsilon^2\sigma}}\,,$$

with $c > 0$ *being a normalization constant, and the Poincaré inequality* (50) *holds.*

In the isotropic framework presented in Theorem 4.9, we can use Proposition 4.8 to conclude the following entropy convergence.

Corollary 4.10. *Under the conditions of Theorem 4.9, it holds that*

$$\mathcal{E}\left(\rho(t)\,|\,\rho_\infty\right) \leq e^{-\lambda t}\mathcal{E}\left(\rho_0\,|\,\rho_\infty\right) \qquad \forall t > 0\,. \tag{53}$$

If we want to consider more general diffusion matrices Q which can be degenerate, only few results are available. In this context of entropy methods, Fokker Planck equations with constant yet degenerate diffusion matrix Q_0 and quadratic loss function $L(x) = \frac{1}{2}x^T Cx$ are studied by Arnold and Erb in [5]. Namely,

$$\begin{cases} \quad \partial_t u = \nabla \cdot (Q_0 \nabla u + u\, Cx) & \text{in} \quad (0,\infty) \times \mathbb{R}^d\,, \\ u(0,x) = u_0(x) & \text{in} \quad \mathbb{R}^d\,, \end{cases} \tag{54}$$

For related Fokker Planck equations see [4, 6]. In [5], the authors adapted the entropy method for equation (54) under following assumptions:

(I) *Confinement potential:* C is positive definite.[1]

(II) *Hörmander condition:* There are no eigenvectors of C in $\ker Q_0$.

Assumption (I) ensures that the mass is not escaping at infinity along the flow. However, the idea behind Hörmander condition (II) in the equation (54) is that if there is a point where diffusion is not acting, then the drift term will push it to the diffusion regime, see the discussion of (54) in [35].

According to [5, Theorems 3.1 and 4.9], conditions (I) and (II) are equivalent to the existence and uniqueness of a steady state in $L^1(\mathbb{R}^d)$ and imply the convergence of the solution of (54) to it.

Theorem 4.11. *[5, Theorem 3.1] There exists a unique steady state $u_\infty \in L^1(\mathbb{R}^d)$ of (54) if and only if conditions (I) and (II) hold. Moreover, the steady state is a non-isotropic Gaussian of the form*

$$u_\infty(x) = c\, e^{-\frac{x^T K^{-1} x}{2}}\,,$$

with K being the unique, symmetric, and positive definite solution of the Lyapunov equation

$$2Q_0 = CK + KC\,.$$

We will denote by $\lambda_{\max}(C)$ and $\lambda_{\min}(C)$ the maximum and minimum eigenvalues of a nonnegative definite matrix C, respectively.

Theorem 4.12. *[5, Theorem 4.9] Let conditions (I) and (II) hold, and let u be the solution to (54) with $u_0 \in L^1(\mathbb{R}^d) \cap \mathcal{P}(\mathbb{R}^d)$ and $\mathcal{E}(u_0 \,|\, u_\infty) < \infty$. Then, for every $t > 0$*

$$\mathcal{E}\left(u(t) \,|\, u_\infty\right) \le c\, e^{-2\gamma t} \mathcal{E}(u_0 \,|\, u_\infty)$$

with $c > 1$ and $\gamma = \lambda_{\min}(C)$ if all the eigenvalues of C are different. If C has repeated eigenvalues, then $\gamma = \lambda_{\min}(C) - \delta$ for every $\delta > 0$.

There are interesting cases that fall out of the hypothesis of Theorem 4.12. Let us consider the following example where Hörmander's condition (II) does not hold on a given subspace. In this case there is still a steady state, but it will not be smooth anymore. As a consequence, the convergence to equilibrium can only happen in a suitable weak sense.

[1]Note that the authors in [5] address a slightly more general problem where C is not necessarily symmetric and hence they assume that C is positive stable, i.e. all eigenvalues has positive real part.

Example 8. Let us consider the equation for the function $u : (0,\infty) \times \mathbb{R}^n \times \mathbb{R}^{d-n}$ with $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{d-n}$ and matrices Q_0, C_0, C_1, C_2, C_3 with the corresponding dimensions:

$$\begin{cases} \partial_t u &= \nabla_{x,y} \cdot \left[\begin{pmatrix} Q_0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \nabla_x u \\ \nabla_y u \end{pmatrix} + u \begin{pmatrix} C_0 & C_1 \\ C_2 & C_3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \right] \qquad \text{in } (0,\infty) \times \mathbb{R}^n \times \mathbb{R}^{d-n} \\ u(0) &= u_0 \qquad \text{on } \mathbb{R}^n \times \mathbb{R}^{d-n} \end{cases} \tag{55}$$

In order to construct the solution of (55), we need the fundamental solution g defined in $(0,\infty) \times \mathbb{R}^n$ of the equation

$$\partial_t g = \nabla \cdot (Q_0 \nabla g + g C_0 x) \qquad \text{in } (0,\infty) \times \mathbb{R}^n .$$

If Q_0 and C_0 satisfy assumption (II), then (see e.g. [5, Lemma 2.5]) we have the explicit expression for g given by

$$g(t,x) = \frac{1}{(2\pi)^{d/2} \det(W(t))} \exp\left(-x^T W^{-1}(t) x\right) ,$$

where

$$W(t) = \int_0^t e^{C_0(s-t)} Q_0 e^{C_0^T(s-t)} \, \mathrm{d}s$$

is a positive definite matrix for all $t > 0$. In addition, if also condition (I) is satisfied there exists a unique steady state g_∞ as in Theorem 4.11.

Theorem 4.13. *Assume $Q_0, C_0 \in \mathbb{R}^{n\times n}$ satisfy (I) and (II) in $\mathbb{R}^n$. Let $C_1 = 0 \in \mathbb{R}^{n\times(d-n)}$, $C_2 = 0 \in \mathbb{R}^{(d-n)\times n}$ and $C_3 \in \mathbb{R}^{(d-n)\times(d-n)}$ be positive definite. Let us consider the equation (55) with $u_0 \in L^1(\mathbb{R}^d) \cap \mathcal{P}_2(\mathbb{R}^d)$ such that*

$$\int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^{d-n}} u_0(x,y) \, \mathrm{d}y \right)^2 g_\infty^{-1}(x) \, \mathrm{d}x < +\infty .$$

Let $\gamma = \gamma(Q_0, C_0)$ be as in Theorem 4.12 and let $\lambda = \min\{2\gamma, 2\lambda_{\max}(C_3)\}$. Then

i) The solution $u(t)$ converge exponentially fast to u_∞ weakly in measures:

$$\left| \int_{\mathbb{R}^d} \varphi(x,y) \left(u(t,x,y) - u_\infty(x,y)\right) \mathrm{d}x \, \mathrm{d}y \right| \le \kappa_1 e^{-\lambda t} \qquad \forall \varphi \in \mathrm{Lip}(\mathbb{R}^d) , \tag{56}$$

with κ_1 depending on u_0, φ and g_∞.

ii) There is exponential decay of the second moments:

$$\left| \int_{\mathbb{R}^n} \int_{\mathbb{R}^{d-n}} (|x|^2 + |y|^2)(u(t,x,y) - u_\infty(x,y)) \, \mathrm{d}x \, \mathrm{d}y \right| \le \kappa_2 e^{-\lambda t} , \tag{57}$$

with κ_2 depending on u_0 and g_∞.

As a consequence,

$$u(t,x,y) \longrightarrow u_\infty(x,y) := g_\infty(x) \delta_0(y)$$

in the 2-Wasserstein distance as $t \to +\infty$.

Proof. We will prove i) and ii) for the solution u which is given by

$$u(t,x,y) = e^{\operatorname{tr}(C_3)t} \int_{\mathbb{R}^n} g(t, x-\tilde{x}) u_0(\tilde{x}, e^{C_3 t} y) \mathrm{d}\tilde{x}\,,$$

see also Example 6 in Section 4.3.

Proof of i). By substituting the explicit expression of u and u_∞ and the change of variables $y' = e^{C_3 t} y$, we obtain

$$\begin{aligned}
&\left| \int_{\mathbb{R}^2} \varphi(x,y) \left(u(t,x,y) - u_\infty(x,y) \right) \mathrm{d}x\,\mathrm{d}y \right| \\
&= \left| \int_{\mathbb{R}^d} \varphi(x,y) \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x}) e^{\operatorname{tr}(C_3)t} u_0(\tilde{x_1}, e^{C_3 t} y) \mathrm{d}\tilde{x} - g_\infty(x) \delta_0(y) \right) \mathrm{d}x\,\mathrm{d}y \right| \\
&= \left| \int_{\mathbb{R}^d} \left(\varphi(x, e^{-C_3 t} y) - \varphi(x,0) + \varphi(x,0) \right) \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x}) u_0(\tilde{x}, y) \mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \right. \\
&\quad \left. - \int_{\mathbb{R}^n} \varphi(x,0) g_\infty(x)\,\mathrm{d}x \right| \\
&\le \int_{\mathbb{R}^d} \left| \varphi(x, e^{-C_3 t} y) - \varphi(x,0) \right| \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x}) u_0(\tilde{x}, y) \mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \\
&\quad + \left| \int_{\mathbb{R}^d} \varphi(x,0) \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x}) u_0(\tilde{x}, y) \mathrm{d}\tilde{x} - g_\infty(x) \right) \mathrm{d}x\,\mathrm{d}y \right| \\
&=:\ A + B\,.
\end{aligned}$$

Here,

$$\begin{aligned}
A &= \int_{\mathbb{R}^d} \left| \varphi(x, e^{-C_3 t} y) - \varphi(x,0) \right| \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x}) u_0(\tilde{x}, y) \mathrm{d}\tilde{x} \right) \mathrm{d}x\,\mathrm{d}y \\
&\le \|\nabla\varphi\|_{L^\infty(\mathbb{R}^d)} \int_{\mathbb{R}^d} |e^{-C_3 t} y| \left(\int_{\mathbb{R}^n} g(t,x-\tilde{x})\,\mathrm{d}x \right) u_0(x,y) \mathrm{d}\tilde{x}\,\mathrm{d}y \\
&\le e^{-\|C_3\| t} \|\nabla\varphi\|_{L^\infty(\mathbb{R}^d)} \int_{\mathbb{R}^n} \int_{\mathbb{R}^{d-n}} |y| u_0(x,y)\,\mathrm{d}x\,\mathrm{d}y \overset{t\to\infty}{\longrightarrow} 0\,,
\end{aligned}$$

with the induced norm for positive matrices $\|C_3\| = \lambda_{\max}(C_3)$. For the second term, let us define the marginal functions

$$u_0^X(x) = \int_{\mathbb{R}^{d-n}} u_0(x,y)\,\mathrm{d}y \qquad \text{and} \qquad u^X(t,x) = \int_{\mathbb{R}^n} g(t,x-\tilde{x}) u_0^X(\tilde{x}) \mathrm{d}\tilde{x}\,.$$

Then,

$$\begin{aligned}
B &= \left| \int_{\mathbb{R}^n} \varphi(x,0) \left(\int_{\mathbb{R}^d} g(t,x-\tilde{x}) u_0^X(\tilde{x}) \mathrm{d}\tilde{x} - g_\infty(x) \right) \mathrm{d}x \right| \\
&\le \int_{\mathbb{R}^n} \left| \varphi(x,0) \right| \left| u^X(t,x) - g_\infty(x) \right| \mathrm{d}x \\
&\le \left(\int_{\mathbb{R}^n} |\varphi(x,0)|^2 g_\infty(x)\,\mathrm{d}x \right)^{1/2} \left(\int_{\mathbb{R}^n} \left[u^X(t,x) - g_\infty(x) \right]^2 g_\infty^{-1}(x)\,\mathrm{d}x \right)^{1/2} \\
&= \left(\int_{\mathbb{R}^n} |\varphi(x,0)|^2 g_\infty(x)\,\mathrm{d}x \right)^{1/2} \left[\mathcal{E}\left(u^X(t) \,|\, g_\infty \right) \right]^{1/2} \\
&\le c\, e^{-\lambda t} \left(\int_{\mathbb{R}^n} |\varphi(x,0)|^2 g_\infty(x)\,\mathrm{d}x \right)^{1/2} \left[\mathcal{E}\left(u_0^X \,|\, g_\infty \right) \right]^{1/2} \overset{t\to\infty}{\longrightarrow} 0\,.
\end{aligned}$$

Note that in the second inequality above we have use Cauchy-Schwarz inequality and in the last inequality the entropy decay of Theorem 4.12. Moreover, this convergence implies the weak convergence in measure, see [1, Theorem 8.8].

Proof of ii). First, note that

$$
\begin{aligned}
&\left|\int_{\mathbb{R}^d} \left(|x|^2+|y|^2\right)\left(u(t,x,y)-u_\infty(x,y)\right)\,\mathrm{d}x\,\mathrm{d}y\right| \\
=&\left|\int_{\mathbb{R}^d} \left(|x|^2+|y|^2\right)\left(e^{\operatorname{tr}(C_3)t}\int_{\mathbb{R}^n} g(t,x-\tilde{x})u_0(\tilde{x},e^{C_3t}y)\mathrm{d}\tilde{x}-g_\infty(x)\delta_0(y)\right)\,\mathrm{d}x\,\mathrm{d}y\right| \\
\le&\left|\int_{\mathbb{R}^d} |x|^2\left(e^{\operatorname{tr}(C_3)t}\int_{\mathbb{R}^n} g(t,x-\tilde{x})u_0(\tilde{x},e^{C_3t}y)\mathrm{d}\tilde{x}-g_\infty(x)\right)\,\mathrm{d}x\,\mathrm{d}y\right| \\
&+\left|\int_{\mathbb{R}^d} |y|^2\left(e^{\operatorname{tr}(C_3)t}\int_{\mathbb{R}^n} g(t,x-\tilde{x})u_0(\tilde{x},e^{C_3t}y)\mathrm{d}\tilde{x}\right)\,\mathrm{d}x\,\mathrm{d}y\right| \\
=:&A+B\,.
\end{aligned}
$$

Let us estimate both addends independently. Note that using Fubini-Tonelli and the change of variable $y'=e^{C_3t}y$ we obtain

$$
\begin{aligned}
A&=\left|\int_{\mathbb{R}^n}|x|^2\left(\int_{\mathbb{R}^n} g(t,x-\tilde{x})u_0^X(\tilde{x})\mathrm{d}\tilde{x}-g_\infty(x)\right)\,\mathrm{d}x\,\mathrm{d}y\right| \\
&=\left|\int_{\mathbb{R}^n}|x|^2\left(u^X(t,x)-g_\infty(x)\right)\,\mathrm{d}x\,\mathrm{d}y\right|\,,
\end{aligned}
$$

using the same notation for marginals functions as above. Then, using Cauchy-Schwarz inequality and the entropy decay of Theorem 4.12, it holds that

$$
\begin{aligned}
A&\le\left(\int_{\mathbb{R}^n}|x|^4 g_\infty(x)\,\mathrm{d}x\right)^{1/2}\left(\int_{\mathbb{R}^n}(u^X(t,x)-g_\infty(x))^2 g_\infty^{-1}(x)\,\mathrm{d}x\right)^{1/2} \\
&=\left(\int_{\mathbb{R}^n}|x|^4 g_\infty(x)\,\mathrm{d}x\right)^{1/2}\left[\mathcal{E}\left(u^X(t)\,|\,g_\infty\right)\right]^{1/2} \\
&\le\left(\int_{\mathbb{R}^n}|x|^4 g_\infty(x)\,\mathrm{d}x\right)^{1/2} c\,e^{-\lambda t}\left[\mathcal{E}\left(u_0^X\,|\,g_\infty\right)\right]^{1/2}\stackrel{t\to\infty}{\longrightarrow}0\,.
\end{aligned}
$$

For the second term B, consider again the change of variable $\tilde{y}=e^{C_3t}y$

$$
\begin{aligned}
B&=e^{-2\|C_3\|t}\left|\int_{\mathbb{R}^{d-n}}\int_{\mathbb{R}^n}|\tilde{y}|^2\left(\int_{\mathbb{R}^n} g(t,x-\tilde{x})\,\mathrm{d}x\right)u_0(\tilde{x},\tilde{y})\mathrm{d}\tilde{x}\mathrm{d}\tilde{y}\right| \\
&\le e^{-2\|C_3\|t}\int_{\mathbb{R}^n}\int_{\mathbb{R}^{d-n}}(|\tilde{x}|^2+|\tilde{y}|^2)u_0(\tilde{x},\tilde{y})\mathrm{d}\tilde{x}\mathrm{d}\tilde{y}\stackrel{t\to\infty}{\longrightarrow}0\,,
\end{aligned}
$$

since $|y|^2=e^{-2\|C_3\|t}|\tilde{y}|^2$ with the induced norm for positive matrices $\|C_3\|=\lambda_{\max}(C_3)$.

The convergence in Wasserstein metric follows by *i*) and *ii*), see the characterization [1, Theorem 8.8]. □

4.5 Some conclusions and open questions

In the previous section we have seen that in some cases, the convergence of the distribution of parameters towards a steady state can also be exponentially fast: this holds under quite restrictive conditions on Q and L. However, we think that this special situation is what happens to the distribution of the parameters near local minima of the loss function L. The heuristic reason is that L should be λ-convex close to a minimum, and also Q should be (practically) constant there. Let us be more precise: assuming all the necessary regularity for ρ, the behaviour of the stochastic process with probability density function ρ near the critical points of L can be understood as follow. Recall that ρ satisfies

$$\begin{cases} \partial_t \rho = \nabla \cdot \Big(\varepsilon^2 Q(x) \nabla \rho + \rho \left(\nabla L(x) + \varepsilon^2 \nabla \cdot Q(x) \right) \Big) & \text{on} \quad (0,\infty) \times \mathbb{R}^d \,, \\ \rho(0,x) = \rho_0(x) & \text{in} \quad \mathbb{R}^d \,, \end{cases} \tag{58}$$

with $\nabla L(x) = \frac{1}{N} \sum_{i=1}^N \nabla L_i(x)$ and

$$Q(x) = \frac{1}{N} \sum_{i=1}^N \nabla L_i(x) \otimes \nabla L_i(x) - \nabla L(x) \otimes \nabla L(x) \,.$$

Suppose that $x_0 \in \mathbb{R}^d$ is a local minimum of L, so that $\nabla L(x_0) = 0$. The goal is to obtain an asymptotic expansion in a neighbourhood of x_0 of (58) when ε is small. We consider

$$x = x_0 + \varepsilon z \,,$$

and assuming enough regularity of the loss functions L_i for $i = 1, \dots, N$ we can compute the Taylor expansion of ∇L and Q around x_0:

$$\begin{aligned} \nabla L(x) &= \nabla L(x_0) + \varepsilon D^2 L(x_0) z + \varepsilon^2 D^3 L(x_0)[z,z] + o\left(|\varepsilon z|^2\right) \\ Q(x) &= \frac{1}{N} \sum_{i=1}^N \nabla L_i(x_0) \otimes \nabla L_i(x_0) - \nabla L(x_0) \otimes \nabla L(x_0) \\ &\quad + \frac{\varepsilon^2}{N} \sum_{i=1}^N \Big[\big(D^2 L_i(x_0) z \big) \otimes \big(D^2 L_i(x_0) z \big) - \big(D^2 L(x_0) z \big) \otimes \big(D^2 L(x_0) z \big) \Big] + o(|\varepsilon z|^2) \,. \end{aligned}$$

Since $\nabla L(x_0) = 0$, the first order expansion reads as follows:

$$\begin{aligned} \nabla L(x) &= \varepsilon D^2 L(x_0) y + o(|\varepsilon z|) \\ Q(x) &= \frac{1}{N} \sum_{i=1}^N \nabla L_i(x_0) \otimes \nabla L_i(x_0) + o(\varepsilon) = Q(x_0) + o(|\varepsilon z|) \,. \end{aligned}$$

Now, let us consider the function $\tilde{u}(t,z) := \rho(t, x(z)) = \rho(t, x_0 + \varepsilon z)$. Assuming enough regularity for $\tilde{u}$, we have

$$\begin{gathered} \nabla_z \tilde{u}(t,y) = \varepsilon \nabla_x \rho(t, x(z)) \,, \\ \nabla_z \cdot \left(A \, \nabla_z \tilde{u}(t,z) \right) = \varepsilon^2 \nabla_x \cdot \left(A \, \nabla_x \rho(t, x(z)) \right) \,, \end{gathered}$$

for any matrix $A \in \mathbb{R}^{d \times d}$. As a consequence we can write the equation for $\tilde{u}$ as follows:

$$\partial_t \tilde{u} = \nabla \cdot \big(Q(x_0) \nabla \tilde{u} + \tilde{u} \, D^2 L(x_0) \, y \big) + \varepsilon^2 \, \nabla \cdot \left(O(1) \nabla \tilde{u} + \tilde{u} \, O(1) \right) \,, \tag{59}$$

This suggests that the behaviour of ρ near a critical point x_0 of L is governed by the following equation:

$$\begin{cases} \partial_t u = \nabla \cdot (Q_0 \nabla u + u\, C z) & \text{in} \quad (0,\infty) \times \mathbb{R}^d\,, \\ u(0,z) = \rho_0(x_0 + \varepsilon z) & \text{in} \quad \mathbb{R}^d\,, \end{cases} \tag{60}$$

and we observe that now, the diffusion matrix is constant and symmetric, indeed $Q_0 = Q(x_0), C = D^2 L(x_0) \in \mathbb{R}^{d\times d}$ and $Q(x_0) \geq 0$. Also, since x_0 is a local minimum we have $D^2 L(x_0) \geq 0$. This allows to apply the convergence results of Theorem 4.13, and get exponential decay towards a steady state.

The conclusion that we draw from the above discussion is that the behaviour of the parameters of the neural network during the learning process and their final distribution, can be understood through equation (60), more precisely, when Theorem 4.13 applies we obtain that there exists a steady state u_∞ so that weak convergence in measure is exponentially fast, that is (56) holds, and the second moments converge strongly and exponentially fast, namely (57) holds, i.e.,

$$\left| \int_{\mathbb{R}^n} \int_{\mathbb{R}^{d-n}} (|x|^2 + |y|^2)(u(t,x,y) - u_\infty(x,y))\, \mathrm{d}x\, \mathrm{d}y \right| \leq \kappa_2 e^{-\lambda t}\,,$$

for suitable $\kappa_2, \lambda > 0$.

Open Questions. We shall present a number of technical issues that we do not prove in this paper and would eventually lead to a rigorous proof of the above heuristic result.

i) **Regularity.** The first issue is undoubtedly the regularity of the solution ρ to problem (58), that can be obtained through the regularity of the "heat" kernel or fundamental solution to (58). We expect partial regularity in the "non degenerate" variables and we conjecture this to be enough to justify rigorously the above expansions.

ii) **Localization error estimates.** The next step would be to quantify the error between the original solution $\tilde{u}$ of (59) and the localized solution u of (60) (in each neighbourhood of the local minima) when $\varepsilon > 0$ is small (and fixed momentarily). We need suitable norm estimates for all $t > 0$, of the form

$$\|\tilde{u}(t) - u(t)\| \leq c(\varepsilon, t)\,.$$

The main issue would be to show that the error term $c(\varepsilon,t) \xrightarrow{\varepsilon \to 0^+} 0$ possibly with quantitative estimates of its behaviour.

iii) **Building a global approximated solution.** Assuming that the error term above can be controlled at each minima, it seems reasonable to approximate the solution ρ of (58) by a combination of the localized solutions of (60) around each local minimum.

Let us be more precise: let $\{x_1, \ldots, x_M\}$ denote the local minima of L, and u_i the corresponding solution to the localized equation (60) around the local minimum x_i (for any $1 \leq i \leq M$). We conjecture that the global behaviour of ρ should be given by

$$\rho(t,x) \approx \sum_{i=1}^{M} m_i(t)\, u_i\left(t, \frac{x - x_i}{\varepsilon}\right) \qquad \forall t > 0 \;\; \forall x \in \mathbb{R}^d\,, \tag{61}$$

where $m_i(t)$ represents the mass concentration/splitting around each local minimum x_i at time $t > 0$ and satisfy $\sum_{i=1}^{M} m_i(t) = 1$ for all $t > 0$.

Another challenge would be to show that the above approximation holds for all (sufficiently large) times. This would give information about the invariant probability measure (i.e., the global steady state), that should be well approximated by

$$\rho_\infty(x) \approx \sum_{i=1}^{M} m_i(\infty)\, u_{i,\infty}\left(\frac{x - x_i}{\varepsilon}\right) \qquad \forall x \in \mathbb{R}^d\,,$$

where $u_{i,\infty}$ are the stationary solutions as in Theorem 4.13, with diffusion matrix $Q(x_i)$ and drift matrix $D^2L(x_i)$.

iv) **Sharp mass displacement and splitting.** A crucial question in the above approximation would be to obtain sharp information about the portion of mass $m_i(t)$ that gradually concentrates around the minima x_i of L, together with its dependence on the initial distribution ρ_0. The local masses m_i represent the probability of the parameters of the neural network to be $x_i \in \mathbb{R}^d$ after training with SGD. As we have shown in Example 6, when $Q(x) = 0$ for every $x \in \mathbb{R}^d$, Equation (58) becomes a "pure transport" equation and the initial datum ρ_0 will determine completely the asymptotic behaviour of $\rho(t)$. Given a datum ρ_0, a delicate issue is to be able to estimate in the sharpest possible way which portion of its mass concentrate around each of the x_i, after we have entered the asymptotic regime. Another crucial issue would be to relate the asymptotic regime, i.e. the smallest time for which the approximation of point *iii*) is effective, with the mean exit times studied in Section 3.

Appendices

A Approximation of the Noisy SGD

Theorem A.1. *Let $\{\theta_n\}_{n\geq 0}$ denote the sequence given by (NSGD) and let X_t be the stochastic process defined by*

$$\begin{cases} dX_t = -\nabla L(X_t) + \sqrt{\eta\big(\not{b}^{-1}Q(X_t) + \delta I_{d\times d}\big)}\, dW_t \\ X_0 = \theta_0 \end{cases} \tag{62}$$

Then, assuming the same conditions of [44, Theorem 1], it holds that (62) is a weak approximation of order 1 of (NSGD), that is, for every function $g : \mathbb{R}^d \to \mathbb{R}$ with polynomial growth there exists $C > 0$ independent of η so that

$$\Big|\mathbb{E}\big[g(X_{n\eta})\big] - \mathbb{E}\big[g(\theta_n)\big]\Big| < C\,\eta \qquad \forall n \geq 0\,.$$

Proof. Let us follow the same steps of the proof of [44, Theorem 1]. First we recall the following Lemma.

Lemma A.2. *Let $0 < \eta < 1$. Consider a stochastic process $\{X_t\}_{t\geq 0}$, $t \geq 0$ satisfying*

$$dX_t = b(X_t)dt + \eta^{1/2}\sigma(X_t)dW_t,$$

with $X_0 = \theta_0 \in \mathbb{R}^d$. Define the one-step difference $\Lambda := X_\eta - \theta_0$, then we have

i) $\mathbb{E}[\Lambda_i] = b_i(\theta_0)\eta + \frac{\eta^2}{2}\left(\sum_{j=1}^d b_j(\theta_0)\partial_j b_i(\theta_0)\right) + \mathcal{O}(\eta^3)$.

ii) $\mathbb{E}\left[\Lambda_i\Lambda_j\right] = \left(b_i(\theta_0)b_j(\theta_0) + \sigma\sigma^T_{ij}(\theta_0)\right)\eta^2 + \mathcal{O}(\eta^3)$.

iii) $\mathbb{E}\left[\prod_{j=1}^s \Lambda_{i_j}\right] = \mathcal{O}(\eta^3)$ *for all* $s \geq 3$, $i_j = 1, \dots, d$.

Therefore, we apply Lemma A.2 with X_t as in (62) and we get

i) $\mathbb{E}\big[\Lambda_i\big] = -\eta\,\partial_i L(\theta_0) + \mathcal{O}(\eta^2)$.

ii) $\mathbb{E}\big[\prod_{j=1}^s \Lambda_{i_j}\big] = \mathcal{O}(\eta^2)$ for all $s \leq 2$, $i_j = 1, \dots, d$.

On the other hand, we consider the first step of the iteration (NSGD)

$$\bar{\Lambda} := \theta_1 - \theta_0 = -\frac{\eta}{\not{b}}\sum_{b_i\in B_1} \nabla L_{b_i}(\theta_0) + \eta Z_0\,,$$

and we proceed with the analogous computations:

i) $\mathbb{E}\big[\bar{\Lambda}_i\big] = -\eta\,\partial_i L(\theta_0) + \mathbb{E}\big[Z_{0,i}\big] = -\eta\,\partial_i L(\theta_0)\,.$

ii) $$\begin{aligned}\mathbb{E}\big[\bar{\Lambda}_i\bar{\Lambda}_j\big] &= \eta^2\,\partial_i L(\theta_0)\,\partial_j L(\theta_0) + \eta^2\mathbb{E}\big[Z_{0,i}\,Z_{0,j}\big] = \eta^2\,\partial_i L(\theta_0)\,\partial_j L(\theta_0) + \eta^2\mathrm{Cov}(Z_{0,i}\,,Z_{0,j}) \\ &= \eta^2\,\partial_i L(\theta_0)\,\partial_j L(\theta_0) + \eta^2\delta\,\delta_{ij} = \mathcal{O}(\eta^2)\,,\end{aligned}$$

where δ_{ij} is the Kronecker delta and we have used that $\mathbb{E}\big[Z_0\big] = 0$, $\mathrm{Var}(Z_0) = \delta I_{d\times d}$ together with the fact that θ_0 and Z_0 are independent.

Now, we will need a key result linking one step approximations to global approximations.

Theorem A.3 (Theorem 2 and Lemma 5 of [54])**.** *Let α be a positive integer and let L be smooth enough with controlled growth. Assume, in addition, that there exist two functions G_1, G_2 with polynomial growth so that*

$$\Big|\mathbb{E}\big[\prod_{j=1}^{s}\Lambda_{i_j}\big] - \mathbb{E}\big[\prod_{j=1}^{s}\bar{\Lambda}_{i_j}\big]\Big| \le G_1(x)\eta^{\alpha+1},$$

for $s = 1, 2, \dots, 2\alpha+1$ and

$$\mathbb{E}\big[\prod_{j=1}^{2\alpha+2}|\bar{\Lambda}_{i_j}|\big] \le G_2(x)\eta^{\alpha+1}.$$

Then, there exists a constant C such that for all g with polynomial growth we have

$$\big|\mathbb{E}[g(X_{n\eta})] - \mathbb{E}[g(x_n)]\big| \le C\eta^{\alpha} \qquad \forall n \ge 0\,.$$

Hence, we conclude the proof of Theorem A.1 by applying Theorem A.3 with $\alpha = 1$. $\square$

B Deduction of the Mean Exit Time problem

Let us recall some classical connections between stochastic processes and PDEs, and in order to focus on the main ideas, we keep it at the heuristic level. See [19, 56, 57, 60] for the precise statements with all the assumptions needed to make these results rigorous.

Consider a stochastic process in $\mathbb{R}^d$ defined by the following stochastic differential equation

$$\begin{cases} \mathrm{d}X_t &= b(X_t)\,\mathrm{d}t + a(X_t)\mathrm{d}W_t, \\ X_0 &= x \end{cases}$$

with sufficiently smooth functions $a : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $b : \mathbb{R}^d \to \mathbb{R}^d$, where W_t is the standard Brownian motion. The infinitesimal generator associated to this stochastic process is given by

$$\begin{aligned} \mathcal{A}f(x) &= \lim_{t\to 0}\frac{\mathbb{E}_x[f(X_t)] - f(x)}{t} \\ &= \frac{1}{2}\mathrm{tr}\Big(a(x)^T a(x) D^2 f(x)\Big) - b(x)\cdot\nabla f(x)\,. \end{aligned}$$

The two main differential equations related with this operator are the following:
Kolmogorov backward equation: Using Ito's Lemma one can deduce

$$\begin{cases} \partial_t u(t,x) &= \mathcal{A}u(t,x), & \text{in } \mathbb{R}_+ \times \mathbb{R}^d \\ u(0,x) &= f(x) & \text{in } \mathbb{R}^d\,. \end{cases}$$

Under suitable assumptions on the functions a, b it is known by Dynkin's formula [56, Chapters 7-8] that the solution represents the expected value of the function f applied to the stochastic process, namely,

$$u(t,x) = \mathbb{E}_x\left[f(X_t)\right] .$$

Fokker-Planck equation: Let us consider $\mathcal{A}^*$ to be the adjoint operator of $\mathcal{A}$ with respect to the $L^2(\mathbb{R}^d)$ scalar product. Then, the Fokker-Planck equation associated to the process X_t reads as follows

$$\begin{cases} \partial_t \rho(t,x) & = \mathcal{A}^* \rho(t,x), \qquad \text{in } \mathbb{R}_+ \times \mathbb{R}^d \\ \rho(0,x) & = \rho_0(x) \qquad\quad\;\; \text{in } \mathbb{R}^d . \end{cases}$$

The solution to this equation when $\rho_0 = \delta$ is the conditional probability $p(X_t = x | X_0 = x)$.

Beside these two classical parabolic differential equations, there are also elliptic differential equations associated to the operator $\mathcal{A}$ which describes certain properties of the stochastic process, such as the mean exit time of the stochastic process from a domain $\Omega \subset \mathbb{R}^d$.

Mean Exit Time equation:[57, Chapter 7.2] Since the mean exit time is time-independent, let us denote it by

$$u(x) = \mathbb{E}[\tau_\Omega^x] ,$$

with

$$\tau_\Omega^x := \inf\{t > 0 : X_t \notin \Omega, X_0 = x\} .$$

This mean exit time satisfies the following elliptic equation

$$\begin{cases} \mathcal{A}u(x) = -1, & \text{in } \Omega \\ u(x) = 0, & \text{otherwise.} \end{cases} \tag{63}$$

For the sake of clarity, let us show a heuristic proof of the deduction of this equation.

Proof. Let $u : \mathbb{R}^d \to \mathbb{R}$ be the solution to (63) and let us consider the stochastic process $u(X_t)$ with $X_0 = x$. By Ito's Lemma, we have that

$$\mathrm{d}\left(u(X_t)\right) = \mathcal{A}u(X_t)\,\mathrm{d}t + a(X_t)\nabla u(X_t)\mathrm{d}W_t .$$

Integrating in $[0, t]$ we obtain

$$u(X_t) - u(x) = \int_0^t \mathcal{A}u(X_s)\mathrm{d}s + \int_0^t a(X_s)\nabla u(X_s)\mathrm{d}W_s .$$

Since above identity holds for every time $t > 0$, let us choose t to be equal to

$$\tau_T = \min\{T, \tau_\Omega^x\} ,$$

in order to obtain

$$u(X_{\tau_T}) - u(x) = \int_0^{\tau_T} \mathcal{A}u(X_s)\,\mathrm{d}s + \int_o^{\tau_T} a(X_s)\nabla u(X_s)\mathrm{d}W_s . \tag{64}$$

By construction of $\tau_{_T}$ we know that

$$X_s \in \Omega \qquad \forall x \in \Omega \quad \forall 0 < s < \tau_{_T}\,,$$

and hence

$$\mathcal{A}u(X_s) = -1 \qquad \forall 0 < s < \tau_{_T}\,.$$

Using this property in (64), we obtain that

$$u(X_{\tau_{_T}}) - u(x) = -\tau_{_T} + \int_0^{\tau_T} a(X_s)\nabla u(X_s)\mathrm{d}W_s\,.$$

Next step is to take the expectation in both sides of the equality. For this purpose, note that under mild assumptions on a we have

$$\mathbb{E}\left[\int_0^{\tau_T} a(X_s)\nabla u(X_s)\mathrm{d}W_s\right] = 0\,,$$

since $\mathbb{E}\left[\tau_{_T}\right] \leq T < +\infty$. Thus, we obtain that

$$\mathbb{E}\left[u(X_{\tau_{_T}})\right] = u(x) - \mathbb{E}\left[\tau_{_T}\right]\,.$$

Since $\mathbb{E}\left[\tau_{_T}\right] \leq \mathbb{E}\left[\tau_\Omega^x\right] < +\infty$ and the trajectories of X_s are continuous, we can tend $T \to +\infty$ using Dominated Convergence Theorem,

$$\mathbb{E}\left[u(X_{\tau_\Omega^x})\right] = u(x) - \mathbb{E}\left[\tau_\Omega^x\right]\,.$$

Finally, note that by construction $X_{\tau_\Omega^x} \in \partial\Omega$ and hence $u(X_{\tau_\Omega^x}) = 0$. Then, we conclude

$$u(x) = \mathbb{E}\left[\tau_\Omega^x\right]\,. \qquad \square$$

C Kramers' Law

The simplest example of the Mean Exit Time problem presented before is the one associated to the stochastic process

$$\mathrm{d}X_t = -\nabla L(X_t)\,\mathrm{d}t + \sqrt{2\varepsilon^2}\mathrm{d}W_t\,. \tag{65}$$

The main advantage of this process is that the invariant probability measure is unique an explicit, namely, its density is given by

$$\rho_\infty(x) = c\,e^{-L(x)/\varepsilon^2}\,,$$

with a normalizing constant $c > 0$. If the function L has different local minima, for example $x_1, x_2 \in \mathbb{R}^d$, a natural question to ask is what is the mean time that needs X_t to reach a neighbourhood of x_2 if $X_0 = x_1$. In 1940, Kramers addressed this problem from a physical point of view in [40], where he developed what it is now known as the Kramers' Law to describe the mean transition time of an overdamped Brownian particle between local minima in a potential landscape. If we define

$$\tau_{x_2}^{x_1} = \min\{t > 0 : X_t \in B_R(x_2), X_0 = x_1\}\,,$$

Kramers Law in the one-dimensional case, d=1, reads

$$\mathbb{E}\left[\tau_{x_2}^{x_1}\right] \backsimeq \frac{2\pi}{\sqrt{L''(x_1)|L''(z)|}} e^{(L(z)-L(x_1))/\varepsilon^2} \,. \tag{66}$$

Here, $z \in \mathbb{R}^d$ is call the *relevant saddle* point and it is the maximum point of L among all the paths from x_1 to x_2, that is, z is the point where the *communication height*

$$H(x_1, x_2) = \inf_{\varphi: x_1 \to x_2} \left(\sup_{y \in \varphi} L(y) \right)$$

is attained. In the multidimensional case, a similar result holds assuming that the Hessian $D^2 L(z)$ has a single negative eigenvalue. Thus, Kramers's Law for $d \geq 2$ reads

$$\mathbb{E}\left[\tau_{x_2}^{x_1}\right] \backsimeq \frac{2\pi}{|\lambda_1(z)|} \sqrt{\frac{|\det(D^2 L(z))|}{\det(D^2 L(x_1))}} \, e^{(L(z)-L(x_1))/\varepsilon^2} \,, \tag{67}$$

with $\lambda_1(z) < 0$ being the unique negative eigenvalue of $D^2 L(z)$. Despite Kramers' Law being a well-known fact in physics since the mid-20th century, the rigorous mathematical proof did not arrive until 2004, thanks to Belgrund and Gentz in [14]. They showed that (67) is an equality up to an error of order $\varepsilon |\log(\varepsilon^2)|^{3/2}$ by using analytical tools based on estimates on the Green function and the capacity, see [13, Theorem 3.3].

However, another approach to this problem is the theory of large deviation, which gives a mathematically rigorous framework to the path-integral method used in physics. Considering the stochastic process in (65), the large deviation principle states that for small ε, the probability of sample paths being close to a set Γ of function $\varphi : [0, T] \to \mathbb{R}^d$ behaves like

$$\lim_{\varepsilon \to 0} 2\varepsilon^2 \log \mathbb{P}\{(X_t)_{0 \leq t \leq T} \in \Gamma\} = -\inf_{\varphi \in \Gamma} I(\varphi) \,,$$

with I being the *action function*

$$I(\varphi) = \frac{1}{2} \int_0^T \left| \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t) + \nabla L(\varphi(t)) \right|^2 \mathrm{d}t \,.$$

This large deviation principle can be stated roughly as

$$\mathbb{P}\{(X_t)_{0 \leq t \leq T} \in \Gamma\} \backsimeq e^{-\inf_{\Gamma} I / 2\varepsilon^2} \,.$$

Note that the action function presented above can be written as

$$\begin{aligned}
I(\varphi) &= \frac{1}{2} \int_0^T \left| \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t) + \nabla L(\varphi(t)) \right|^2 \mathrm{d}t \\
&= \frac{1}{2} \int_0^T \left| \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t) - \nabla L(\varphi(t)) \right|^2 \mathrm{d}t + 2 \int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t) \cdot \nabla L(\varphi(t)) \, \mathrm{d}t \\
&= \frac{1}{2} \int_0^T \left| \frac{\mathrm{d}}{\mathrm{d}t} \varphi(t) - \nabla L(\varphi(t)) \right|^2 \mathrm{d}t + 2 \Big[L(\varphi(t)) - L(\varphi(0)) \Big] \,.
\end{aligned}$$

Hence, if φ satisfies the time-reversed system $\frac{\mathrm{d}}{\mathrm{d}t} \varphi = \nabla L(\varphi)$, the first term of the action function above vanishes. Connecting a local minimum x_1 to a point in the basin of attraction of x_2 with such a solution is possible if one allows for an arbitrarily long time. Therefore, the quasipotential is given by

$$\bar{L} = 2\left[\inf_{\partial\Omega} L - L(x_1)\right] ,$$

with $\Omega \subset \mathbb{R}^d$ being a neighbourhood of x_1. In the case of double-well potential, if Ω is chosen such that it is cointained in the basin of attraction of x_1 and its boundary is close to the relevant saddle point z, the large deviation principle reads

$$\lim_{\varepsilon\to 0} \varepsilon^2 \log \mathbb{E}\left[\tau^{x_1}_{B_R(x_2)}\right] = L(z) - L(x_1) .$$

It is possible to generalize this large deviation approach to stochastic processes of the form

$$\begin{cases} \mathrm{d}X_t &= -\nabla L(X_t)\,\mathrm{d}t + \sqrt{2\varepsilon}a(X_t)\mathrm{d}W_t \\ X_0 &= x\,, \end{cases} \tag{68}$$

with a non-degenerate matrix $a(x) \in \mathbb{R}^{d\times d}$. In this case, the invariant probability measure is not explicit in general, nevertheless it is possible to write the action function as

$$I_a(\varphi) = \int_0^T \left| a(\varphi(t))^{-1}\left(\frac{\mathrm{d}}{\mathrm{d}t}\varphi(t) + \nabla L(\varphi(t))\right)\right|^2 \mathrm{d}t\,,$$

if φ is absolutely continuous, $\frac{\mathrm{d}}{\mathrm{d}t}\varphi$ is square integrable and $\varphi(0) = x$; in any other case $I_a(\varphi) = +\infty$. Under restrictive condition on L and a, a large deviation principle for (68) was proven in [25, 64].

Theorem C.1. *[67, Theorem 12.1] Assume that ∇L and a are bounded and Lipschitz continuous. If $a^T(x)a(x)$ is uniformly elliptic, then the following large deviation principle holds for every $t > 0$:*

i) Upper bound. For any closed set $\Gamma_c \subset (C([0,T]))^d$,

$$\lim_{\varepsilon\to 0} \varepsilon^2 \log \mathbb{P}\{(X_t)_{0\le t\le T} \in \Gamma_c\} \le -\inf_{\varphi\in\Gamma_c} I_a(\varphi)$$

ii) Lower bound. For any open set $\Gamma_o \subset (C([0,T]))^d$,

$$\lim_{\varepsilon\to 0} \varepsilon^2 \log \mathbb{P}\{(X_t)_{0\le t\le T} \in \Gamma_o\} \ge -\inf_{\varphi\in\Gamma_o} I_a(\varphi)$$

Hence, roughly speaking, for the stochastic process defined in (68) we have that

$$\mathbb{P}\{(X_t)_{0\le t\le T} \in \Gamma\} \backsimeq e^{-\inf\limits_{\Gamma} I_a/2\varepsilon^2} .$$

This approach provides a method to compute likelihood of rare events in stochastic systems. For a practical explanation of the techniques used to derive sharp asymptotic estimates, refer to [31].

Conflict of interest statement. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data availability statement. All data generated or analysed during this study are included in the published article.

References

[1] L. Ambrosio, E. Brué and D. Semola. *Lectures on Optimal Transport.* Springer, Cham, 2021.

[2] L. Ambrosio and N. Gigli. *A User's Guide to Optimal Transport.* In: Modelling and Optimisation of Flows on Networks. Lecture Notes in Mathematics, vol 2062. Springer, Berlin, Heidelberg, 2013.

[3] M. Andriushchenko, A. Varre, L. Pillaud-Vivien, and N. Flammarion *SGD with Large Step Sizes Learns Sparse Features.* Proceedings of the 40th International Conference on Machine Learning, PMLR 202, 2023.

[4] F. Achleitner, A. Arnold and D. Stürzer. *Large-time behavior in non-symmetric Fokker-Planck equations.* Riv. Math. Univ. Parma (N.S.), 6(1):1–68, 2015.

[5] A. Arnold, J. Erb. *Sharp entropy decay for hypocoercive and non-symmetric Fokker-Planck equations with linear drift.* ASC Report 29/2014, 1-45, Institute of Analysis and Scientific Computing, TU Wien, 2014.

[6] A. Arnold, P. Markowich, G. Toscani and A. Unterreiter *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations.* Comm. Partial Differential Equations 26 43–100, 2001.

[7] S. Arora, N. Cohen, W. Hu, and Y. Luo *Implicit Regularization in Deep Matrix Factorization.* 33rd Conference on Neural Information Processing Systems, 2019.

[8] F. Bach, E. Moulines *Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning.* Advances in neural information processing systems 24, 2011.

[9] D. Bakry, F. Barthe and P. Cattiaux. *A simple proof of the Poincaré inequality for a large class of probability measures including the log-concave case.* Electronic Communications in Probability, 13, 2008.

[10] D. Bakry, M. Émery. *Diffusions hypercontractives.* Séminare de Probabilités XIX 1983/84, 177-206, 1985.

[11] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators.* Vol. 348. Grundlehren der Mathematischen Wissenschaften, Fundamental Principles of Mathematical Sciences. Springer, Cham, 2014.

[12] D. Barrett, B. Dherin *Implicit gradient regularization.* International Conference on Learning Representations, 2021.

[13] N. Belgrund. *Kramers' law: validity, derivations and generalisations.* Markov Process. Relat. Fields 19, 459–490, 2013.

[14] N. Berglund and B. Gentz. *Noise-Induced Phenomena in Slow-fast Dynamical Systems.* Probability and its Applications (New York). Springer-Verlag London Ltd., London, 2006.

[15] M. Belkin, C. Liu and L. Zhu. *Loss landscapes and optimization in over-parameterized non-linear systems and neural networks.* Applied and Computational Harmonic Analysis, vol 59, 85-116, 2022.

[16] G. Ben Arous, R. Gheissari, A. Jagannath *High-dimensional limit theorems for SGD: Effective dynamics and critical scaling.* 36th Conference on Neural Information Processing Systems, 2022.

[17] C. Bianca, C. Dogbe.*On the existence and uniqueness of invariant measure for multidimensional stochastic processes.* Nonlinear Studies - The International Journal, Cambridge Scientific Publishers, hal-02151779, 2017.

[18] G. Blanc, N. Gupta, G. Valiant, and P. Valiant *Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process.* Proceedings of Machine Learning Research vol 125:1–31, 2020.

[19] V.I. Bogachev, N.V. Krylov, M. Röckner, S.V. Shaposhnikov. *Fokker-Planck-Kolmogorov Equations.* Amer. Math. Soc., Providence, Rhode Island, 2015.

[20] L. Bottou, F.E. Curtis and J. Nocedal. *Optimization Methods for Large-Scale Machine Learning.* SIAM Review 60, 2016.

[21] P. Chaudhari, S. Soatto *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks.* In 2018 Information Theory and Applications Workshop (ITA), pp. 1–10.

[22] X. Cheng, D. Yin, P. Bartlett, and Michael Jordan. Stochastic gradient and Langevin processes. In International Conference on Machine Learning, pp. 1810–1819. PMLR 119, 2020.

[23] L. Chizat, M. Colombo, X- Fernández-Real, A. Figalli *Infinite-width limit of deep linear neural networks.* Communications on Pure and Applied Mathematics, 2024.

[24] M.G. Crandall, H. Ishii, P.L. Lions. *User's guide to viscosity solutions of second order partial differential equations.* Bull. Amer. Math. Soc. 27, 1-67, 1992.

[25] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, 2nd ed., Applications of Mathematics (New York), vol. 38, Springer-Verlag, New York, 1998.

[26] Y. Feng, L. Li, J.-G. Liu. *Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations.* Commun. Math. Sci. 16 (3), 2018.

[27] X. Fernández-Real, A. Figalli *The Continuous Formulation of Shallow Neural Networks as Wasserstein-Type Gradient Flows*. In "Analysis at large - Dedicated to the life and work of J. Bourgain", A. Avila, M. Th. Rassias, Y. Sinai (Eds.), Springer 2022.

[28] A. Figalli and F. Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows.* EMS Press, Berlin, 2021.

[29] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He *Accurate, large minibatch sgd: Training imagenet in 1 hour*. arXiv preprint arXiv:1706.02677 (2017)

[30] P. Goyal, Y. Bengio *Inductive biases for deep learning of higher-level cognition*. Proc. R. Soc. A 478:20210068, 2022.

[31] T. Grafke, T. Schäfer and E. Vanden-Eijnden. *Sharp Asymptotic Estimates for Expectations, Probabilities, and Mean First Passage Times in Stochastic Systems with Small Noise.* Commun. Pure Appl. Math. 2023.

[32] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro *Implicit Bias of Gradient Descent on Linear Convolutional Networks*. 32nd Conference on Neural Information Processing Systems, 2018,

[33] M. Gurbuzbalaban, U. Simsekli, L. Zhu *The Heavy-Tail Phenomenon in SGD*. Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

[34] J. Z. HaoChen, C. Wei, J. D. Lee, T. Ma *Shape Matters: Understanding the Implicit Bias of the Noise Covariance*. Proceedings of Machine Learning Research 134:1–43, 2021.

[35] L. Hörmander. *Hypoelliptic second order differential equations*. Acta Math. 119, 147-171, 1967.

[36] W. Hu, C. Junchi Li, L. Li, and J.-G. Liu *On the diffusion approximation of nonconvex stochastic gradient descent*. Annals of Mathematical Sciences and Applications 4:3–32, 2019.

[37] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey *Three factors influencing minima in sgd*. In International Conference on Artificial Neural Newtorks 2018.

[38] S. Jastrzebski, M. Szymczak, S. Fort, D. Arpit, J. Tabor, K. Cho, and K. Geras *The break-even point on optimization trajectories of deep neural networks*. In International Conference on Learning Representations, 2021.

[39] Kloeden P E and Platen E. *Numeric Solution of Stochastic Differential Equations.* Springer, Berlin, 395–401, 1992.

[40] H. A. Kramers. *Brownian motion in a field of force and the diffusion model of chemical reactions*. Physica 7, 284–304, 1940.

[41] H. J. Kushner, G. Yin *Stochastic approximation and recursive algorithms and applications.* Springer, 2003.

[42] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller Efficient BackProp. In G. Orr, K. Müller (eds) "Neural networks: tricks of the trade". Springer 1998.

[43] Z. Li, S. Malladi, and S. Arora *On the Validity of Modeling SGD with Stochastic Differential Equations (SDEs).* Advances in Neural Information Processing Systems 34, 2021.

[44] 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, 2101–2110, 2017.

[45] Q. Li, C. Tai, and W. E. *Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations.* The Journal of Machine Learning Research, 20(1):1474–1520, 2019.

[46] Z. Li, T. Wang, and S. Arora *What happens after SGD reaches zero loss? - A mathematical framework.* International Conference on Learning Representations, 2022.

[47] Y. Li, C. Wei, T. Ma *Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks.* 33rd Conference on Neural Information Processing Systems, 2019.

[48] Fine-Tuning Language Models with Just Forward Passes.

[49] S. Malladi, K. Lyu, A. Panigrahi, S. Arora *On the SDEs and Scaling Rules for Adaptive Gradient Algorithms.* 36th Conference on Neural Information Processing Systems, 2022.

[50] S. Mandt, M. Hoffman, and D. Blei *A variational analysis of stochastic gradient algorithms.* Proceedings of The 33rd International Conference on Machine Learning 48:354-363, 2016.

[51] Mandt, M. D. Hoffman, and D. M. Blei *Stochastic gradient descent as approximate Bayesian inference.* The Journal of Machine Learning Research 18:4873–4907, 2017.

[52] D. Masters, C. Luschi *Revisiting small batch training for deep neural networks.* Preprint at https://arxiv.org/abs/1804.07612 (2018)

[53] G. Menz, A. Schlichting. *Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape.* Ann. Probab. 42 (5) 1809-1884, 2014.

[54] G.N. Milstein. *Numerical integration of stochastic differential equations,* volume 313. Springer Science & Business Media, 1995.

[55] C. Moucer, A. Taylor, and F. Bach *A systematic approach to Lyapunov analyses of continuous-time models in convex optimization.* SIAM Journal on Optimization, 33:1558–1586, 2023.

[56] B. Øksendal. *Stochastic differential equations.* Universitext. Springer-Verlag, Berlin, 2003.

[57] G.A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations.* Springer, New York, 2014.

[58] A. Porretta. *Decay rates of convergence for Fokker-Planck equations with confining drift.* Advances in Mathematics, vol. 436, 109393, 2024.

[59] G. Rotskoff, E. Vaden-Eijnden *Trainability and accuracy of neural networks: an interacting particle system approach*. Communications on Pure and Applied Mathematics LXXV:1889–1935, 2022.

[60] Z. Schuss. *Theory and Applications of Stochastic Differential Equations.* Applied Math. Science, vol. 170, 2010.

[61] B. Shi, W. Su, M. Jordan *On Learning Rates and Schrödinger Operators*. The Journal of Machine Learning Research 24:1-53, 2023.

[62] U. Simsekli, L. Sagun, M. Gürbüzbalaban *A Tail-Index Analysis of Stochastic Gradient Noise in Deep Neural Networks*. Proceedings of the 36th International Conference on Machine Learning, PMLR 97, 2019.

[63] D. Soudry, E. Hoffer, M. Shpigel Nacson, S. Gunasekar, and N. Srebro *The Implicit Bias of Gradient Descent on Separable Data*. The Journal of Machine Learning Research 19:2822–2878, 2018.

[64] S. R. S. Varadhan. *Large deviations and applications.* CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 46, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1984.

[65] G. Vardi *On the Implicit Bias in Deep-Learning Algorithms*. Communications of the ACM 66:86-93 (2023).

[66] J. L. Vázquez, *Asymptotic behaviour methods for the Heat Equation. Convergence to the Gaussian.* arXiv preprint https://arxiv.org/abs/1706.10034

[67] W. E, T. Li and E. Vanden-Eijnden. *Applied Stochastic Analysis.* American Mathematical Society, Graduate Studies in Mathematics, 2021.

[68] A. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht *The Marginal Value of Adaptive Gradient Methods in Machine Learning*. 31st Conference on Neural Information Processing System, 2017.

[69] S. Wojtowytsch. *Stochastic gradient descent with noise of machine learning type. Part II: Continuous time analysis*. Journal of Nonlinear Science, vol. 34, 2021.

[70] Z. Xie, I. Sato, and M. Sugiyama *A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima*. International Conference on Learning Representations, 2021.

[71] S. Yaida *Fluctuation - dissipation relations for stochastic gradient descent* International Conference on Learning Representations, 2019.

[72] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals *Understanding deep learning requires rethinking generalization*. International Conference on Learning Representations, 2017.

[73] P. Zhou, J. Feng, C. Ma, C. Xiong, S. HOI, W. E *Towards Theoretically Understanding Why SGD Generalizes Better Than A DAM in Deep Learning*. 34th Conference on Neural Information Processing Systems, 2020.