THE ASYMPTOTIC BEHAVIOR OF FRAUDULENT ALGORITHMS

BY MICHEL BENAÏM[1,a] AND LAURENT MICLO[2,b]

[1]*Department, Institut de Mathématiques*
Neuchâtel University, [a]*michel.benaim@unine.ch*

[2]*Department, Toulouse School of Economics*
Institut de Mathématiques de Toulouse
CNRS and University of Toulouse, [b]*miclo@math.cnrs.fr*

Let U be a Morse function on a compact connected m-dimensional Riemannian manifold, $m \geqslant 2$, satisfying $\min U = 0$ and let $\mathcal{U} = \{x \in M : U(x) = 0\}$ be the set of global minimizers. Consider the stochastic algorithm $X^{(\beta)} := (X^{(\beta)}(t))_{t \geqslant 0}$ taking values in M, whose generator is $U\triangle[\cdot] - \beta\langle \nabla U, \nabla[\cdot]\rangle$, where $\beta \in \mathbb{R}$ is a real parameter. We show that for $\beta > \frac{m}{2} - 1$, $X^{(\beta)}(t)$ converges a.s. as $t \to \infty$, toward a point $p \in \mathcal{U}$ and that each $p \in \mathcal{U}$ has a positive probability to be selected. On the other hand, for $\beta < \frac{m}{2} - 1$ and when the initial law does not charge $\mathcal{U}$, the law of $X^{(\beta)}(t)$ converges in total variation (at an exponential rate) toward the probability measure π_β having density proportional to $U(x)^{-1-\beta}$ with respect to the Riemannian measure.

1. Introduction. Global optimization is an important and difficult task in applied mathematics, so the development of corresponding algorithms has been the subject of a great deal of work. Specific assumptions on the function U to be optimized has led to very efficient approaches: e.g. gradient descent or Newton's method for convex optimization, moment method for polynomial optimization. There are also a few general algorithms, often using a certain amount of randomness, such as simulated annealing or interacting particle algorithms. Here we will consider the special situation where the minimum value of U is known and can be used by the algorithm. Such an algorithm is said to be **fraudulent**, since in general this value is not available and one might think that knowing it is equivalent to knowing a global minimum. However, there are natural situations where the minimum value is given but the corresponding minimizers are not.

Here are some simple illustrative examples. Consider a generic polynomial P of odd degree larger than 5 (assume e.g. that the coefficients are sampled according to a non-degenerate Gaussian law) and define $U = P^2$ on $\mathbb{R}$. We know that the minimal value of U is zero, but we cannot express its roots through radicals. For a more geometric example, consider a tangent continuous vector field V on a sphere $\mathbb{S}^m$ of even dimension $m \in 2\mathbb{N}$. From Brouwer's hairy ball theorem [6], there exists a point $x_0 \in \mathbb{S}^m$ such that $V(x_0) = 0$. Thus we know that the minimal value of the mapping $U : \mathbb{S}^m \ni x \mapsto \|V(x)\|^2$ is zero, without being able to point a global minimizer in general. Consider for instance a vector field V constructed from a $(m+1, m+1)$-Brownian sheet, i.e. defined from $[0, +\infty)^{m+1}$ to $\mathbb{R}^{m+1}$ (see e.g. page 269 of Walsh [22]), by restriction to the sphere centred at $(2, 2, ..., 2)$ of radius 1, and by projection on its tangent spaces. Below we will only consider smooth functions U, so V should also be regularized.

MSC2020 subject classifications: Primary 60J60; secondary 58J65, 90C26, 65C05, 60F15, 60J35, 35K10, 37A50.

Keywords and phrases: Global optimization, fraudulent stochastic algorithms, Morse functions, attractive and repulsive minimizers.

More important and current applications of fraudulent algorithms can be found in the field of statistical classification. Appendix A.1 provides some heuristics explaining why the stochastic algorithm $X^{(\beta)}$ introduced below can be seen as a toy model for the diffusion limit of mini-batch stochastic gradient descent algorithms extensively used in the theory of Machine Learning, see for instance Li, Tai and E [16], Wu, Wang and Su [25], Mori, Ziyin, Liu and Ueda [18] and Wojtowytsch [23, 24] as well as references therein.

Another instance of the usefulness of fraudulent procedures is when a global minimizer is known and that we are looking for all the other ones. For other interests of fraudulent algorithms, we refer to [17], where the term was coined. There the motivation for the stochastic algorithm $X^{(\beta)}$ comes from its approximation of the large-time limit behavior of the time-inhomogeneous swarm mean-field algorithm introduced in [5], which is itself non-fraudulent but uses its current distribution to estimate in real time the minimal value.

Let us present more precisely the framework considered here: U is a Morse function defined on a compact manifold M of dimension $m \geqslant 2$. The underlying stochastic process $X^{(\beta)} := (X^{(\beta)}(t))_{t\geqslant 0}$, takes values in M and comes with a real parameter β which can be tuned to increase the relative importance of U with respect to the injected randomness. More specifically, the generator of $X^{(\beta)}$ will have the form $U\triangle[\cdot] - \beta\langle\nabla U, \nabla[\cdot]\rangle$, i.e., if we were in an Euclidean context, the associated $X^{(\beta)}$ can be constructed as solution of the s.d.e.

$$dX^{(\beta)}(t) = -\beta\nabla U(X^{(\beta)}(t))dt + \sqrt{2U(X^{(\beta)}(t))}dB_t$$

where $B = (B_t)_{t\geqslant 0}$ is a standard Brownian motion on $\mathbb{R}^m$.

Two quantities $\beta_\vee \geqslant \beta_\wedge \in \mathbb{R}$ (depending explicitly on the eigenvalues of the Hessians of U at its global minima, see Remark 2.4 below) were introduced in [17] so that $\beta > \beta_\vee$ implies the a.s. convergence of $X^{(\beta)}(t)$ as $t \to \infty$, toward the global minima of U (and each of them attracts the algorithm with positive probability, when $X^{(\beta)}(0)$ is not a global minima), while for $\beta < \beta_\wedge$ the probability that $X^{(\beta)}(t)$ converges toward a global minimum of U is zero.

Our goal in the present paper is to sharpen these result and describe completely the long term behavior of $X^{(\beta)}$ for all $\beta \neq \beta_0$, where $\beta_0 := \frac{m}{2} - 1$ is a universal (i.e. independent of U) critical value, contrary to the results of [17], where $\beta_\wedge$ and $\beta_\vee$ depended on U (except in dimension 1). It thus follows from the current results that this dependence was artificial in dimension larger or equal to 2, $\beta_0 \in [\beta_\wedge, \beta_\vee]$ being the exact critical value for the following behaviors. We will show that for $\beta > \beta_0$, $X^{(\beta)}(t)$ a.s. converges toward a global minimizer of U and that each global minimizer has a positive probability to be selected (except when $X^{(\beta)}(0)$ is itself a global minima). On the other hand, for $\beta < \beta_0$, the process converges in distribution toward a (unique) invariant distribution whose density (with respect to the Riemannian measure) is explicit. The present results are thus sharper than those of [17] as soon as $\beta_0 > \beta_\wedge$ and this is equivalent (as it can be seen from (2) in [17]) to the fact there exists at least one global minimum $x \in \mathcal{U}$ such that the Hessian of U at x is not proportional to the identity. In practice this feature is quite generic, as soon as $m > 1$. Here our results will be obtained by an approach completely different from that of [17], which was based on comparisons with Bessel processes of various dimensions. Instead, below we will rely on the persistence/non-persistence approach presented in [4] and [2].

The paper is organized as follows. Section 2 sets the notation and presents the main results. Section 3 considers the situation where M is no longer a compact manifold but the Euclidean space $\mathbb{R}^m$. It allows to introduce the main ingredients of the proof in a simple setting. Section 4 is devoted to the proof of the main results. Certain additional points are discussed in appendix.

2. Notation and main result. We assume throughout that M is a compact connected Riemannian manifold having dimension $m \geqslant 2$ and $U : M \to \mathbb{R}$ is a smooth function such that (this is the fraudulent assumption):

$$\min_M U = 0.$$

The zero set of U,

$$\mathcal{U} := \{p \in M \,:\, U(p) = 0\},$$

is then the set of global minimizers. We furthermore assume that every $p \in \mathcal{U}$ is *non-degenerate*, meaning that the Hessian of U at p is non-degenerate. This assumption implies that $\mathcal{U}$ is finite. In particular, $N := M \backslash \mathcal{U}$ is a noncompact connected manifold.

Let L_β be the operator on $\mathcal{C}^2(M)$ defined as

$$L_\beta[\cdot] := U\triangle[\cdot] - \beta\langle \nabla U, \nabla[\cdot]\rangle \tag{1}$$

where $\triangle$, $\langle\cdot,\cdot\rangle$ and ∇ stand for the Laplacian, scalar product and gradient associated to the Riemannian structure of M, and $\beta \in \mathbb{R}$.

A *diffusion process generated by (1)*, is a continuous-time Feller Markov process on M, $X^{(\beta)} = (X^{(\beta)}(t))_{t\geqslant 0}$, with infinitesimal generator $\mathcal{L}_\beta$ and domain $D(\mathcal{L}_\beta) \subset \mathcal{C}^0(M)$ (see e.g. Le Gall [15], Section 6.2, for the definitions of Feller processes, domains and generators) such that for all $f \in \mathcal{C}^2(M)$:

$$f \in D(\mathcal{L}_\beta) \text{ and } \mathcal{L}_\beta f = L_\beta f.$$

Since the mapping ∇U and $\sqrt{U}$ are Lipschitzian, due to the non-degeneracy assumption of the zeroes of U for the latter, such a diffusion process exists. More details are given in the appendix. In addition, given the initial distribution of $X^{(\beta)}(0)$, say μ, the law of $X^{(\beta)}$, $\mathbb{P}_\mu^{(\beta)}$, is uniquely determined by μ and L_β. As usual, we write $\mathbb{P}_x^{(\beta)}$ for $\mathbb{P}_{\delta_x}^{(\beta)}$. By a mild (but convenient) abuse of notation we may write $\mathbb{P}_x(X^{(\beta)} \in \cdot)$ for $\mathbb{P}_x^{(\beta)}(\cdot)$. We also let $P^{(\beta)} = (P_t^{(\beta)})_{t\geqslant 0}$ denote the semi-group induced by $X^{(\beta)}$. It is defined, as usual, by

$$\forall\, t \geqslant 0,\, \forall\, x \in M, \qquad P_t^{(\beta)} f(x) := \mathbb{E}_x f(X^{(\beta)}(t))$$

for every measurable, bounded or nonnegative, map $f : M \to \mathbb{R}$.

The proof of the next proposition, which relies on classical results, is given in Appendix A.3.

PROPOSITION 2.1. *Here are some basic properties of $P^{(\beta)}$:*

(i) *$P^{(\beta)}$ leaves N and $\mathcal{U}$ invariant: for all $t \geqslant 0$, $P_t^{(\beta)}\mathbf{1}_N = \mathbf{1}_N$.*
(ii) *$P^{(\beta)}$ is Feller on M and strong-Feller on N:*
- *For all $t \geqslant 0$ and $f \in \mathcal{C}^0(M)$, $P_t^{(\beta)}(f) \in \mathcal{C}^0(M)$;*
- *For all $t > 0$ and $f : N \to \mathbb{R}$ bounded measurable, $P_t^{(\beta)}(f)$ is continuous on N.*

Note that $P^{(\beta)}$ is not strong Feller on M, as it can be seen by considering the indicator function of N. In order to state our main result we first associate, to each $p \in \mathcal{U}$, a certain Lyapunov exponent. Given a symmetric positive definite $m \times m$ real matrix A, and $\beta \in \mathbb{R}$, define the probability measure $\mu_{A,\beta}$ on $\mathbb{S}^{m-1}$, the unit sphere in $\mathbb{R}^m$, via

$$\forall\, \theta \in \mathbb{S}^{m-1}, \qquad \mu_{A,\beta}(d\theta) = \frac{1}{Z(A, 1+\beta)} \langle \theta, A\theta\rangle^{-1-\beta}\, \sigma(d\theta) \tag{2}$$

where, σ is the uniform probability measure on $\mathbb{S}^{m-1}$, $\langle \cdot, \cdot \rangle$ the Euclidean dot product (not to be confused with the Riemannian metric on M) and $Z(A, 1+\beta)$ is the normalization constant.

Define the *β-average eigenvalue* of A as

$$\Lambda(A,\beta) = \int_{\mathbb{S}^{m-1}} \langle \theta, A\theta \rangle \mu_{A,\beta}(d\theta) = \frac{Z(A,\beta)}{Z(A,1+\beta)}. \tag{3}$$

Let $\lambda_1(A) \leqslant \ldots \leqslant \lambda_m(A)$ be the eigenvalues of A. Observe that $\Lambda(A,\beta)$ only depends on these eigenvalues, because σ is invariant by orthogonal transformations and A is orthogonally conjugate to a diagonal matrix. Observe also that

$$\lambda_1(A) \leqslant \Lambda(A,\beta) \leqslant \lambda_m(A). \tag{4}$$

REMARK 2.2. Inequalities (4) are strict, except when $\lambda_1(A) = \lambda_m(A)$. Furthermore it can be shown (see the appendix Section A.4) that for all numbers $\lambda_- < \lambda < \lambda_+$, there exists, for m sufficiently large, a $m \times m$ definite positive matrix A such that $\lambda_1(A) = \lambda_-, \Lambda(A,\beta) = \lambda$ and $\lambda_m(A) = \lambda_+$.

Given $p \in \mathcal{U}$, we let A_p denote the diagonal matrix whose entries $0 < \lambda_1(p) \leqslant \ldots \leqslant \lambda_m(p)$ are the eigenvalues of the Hessian of U at p. Set

$$\beta_0 \coloneqq \frac{m}{2} - 1.$$

Our main result is the following.

THEOREM 2.3. *Let $x \in N$ and $\beta \in \mathbb{R}$.*

(i) *If $\beta > \beta_0$, then*

$$\sum_{p \in \mathcal{U}} \mathbb{P}_x \left[\limsup_{t \to +\infty} \frac{\ln(d(X^{(\beta)}(t), p))}{t} \leqslant -\Lambda(A_p,\beta)(\beta - \beta_0) \right] \;=\; 1,$$

where each term in the above sum is positive.

(ii) *If $\beta < \beta_0$, then $X^{(\beta)}$ has, on N, a unique invariant probability distribution given by*

$$\pi_\beta(dx) \coloneqq \frac{1}{C_\beta} U(x)^{-1-\beta}(x) \ell(dx),$$

where C_β is a normalization constant and $\ell(dx)$ stands for the Riemannian measure. Furthermore:

(a) *$X^{(\beta)}$ is positive recurrent on N, meaning that for all $f \in L^1(\pi_\beta)$, $\mathbb{P}_x$ a.s.,*

$$\lim_{t \to +\infty} \frac{1}{t} \int_0^t f(X^{(\beta)}(s)) ds = \pi_\beta(f)$$

(b) *There exist positive constants a, b, χ (depending on β) with $\chi < \beta_0 - \beta$, such that for all $f : N \to \mathbb{R}$, measurable,*

$$|\mathbb{E}_x[f(X^{(\beta)}(t))] - \pi_\beta(f)| \leqslant \frac{ae^{-bt}}{d(x,\mathcal{U})^\chi} \|f\|_\chi,$$

where

$$\|f\|_\chi := \sup_{y \in N} |f(y)| d(y,\mathcal{U})^\chi.$$

(iii) *If $\beta = \beta_0$, then, for every neighborhood O of $\mathcal{U}$, $\mathbb{P}_x$ a.s.,*

$$\lim_{t\to+\infty} \frac{1}{t}\int_0^t \mathbf{1}_{\{X^{(\beta)}(s)\in O\}} ds = 1$$

REMARK 2.4. Theorem 2.3 is an improvement over the results of [17], which showed the a.s. convergence of $X^{(\beta)}$ toward elements of $\mathcal{U}$ (each being approached with a positive probability) only for $\beta > \beta_\vee \geqslant \beta_0$ with

$$\beta_\vee := \max_{p\in\mathcal{U}} \frac{\sum_{l\in[\![m]\!]} \lambda_l(p)}{2\lambda_1(p)} - 1, \tag{5}$$

and the a.s. non-convergence of $X^{(\beta)}$ toward elements of $\mathcal{U}$ for $\beta < \beta_\wedge \leqslant \beta_0$, with

$$\beta_\wedge := \min_{p\in\mathcal{U}} \frac{\sum_{l\in[\![m]\!]} \lambda_l(p)}{2\lambda_m(p)} - 1. \tag{6}$$

REMARK 2.5. Here we restrict our attention to dimensions $m \geqslant 2$, so that N is connected. The case $m = 1$ which corresponds to the circle is already treated in [17].

REMARK 2.6. While it is true that for $\beta > \beta_0$, whatever the initial point x_0 outside $\mathcal{U}$, the law of $X^{(\beta)}(t)$ will charge all the points of $\mathcal{U}$ asymptotically for large t, we also have that when we let x_0 converge toward a particular $x \in \mathcal{U}$, the asymptotical weight put by $X^{(\beta)}(t)$ on x converges toward 1. As a consequence, if we want to use $X^{(\beta)}$ to find all the elements of $\mathcal{U}$ with a non-neglectable chance, say already knowing a particular $x \in \mathcal{U}$, we should not initialize $X^{(\beta)}$ close to x. Or, as it was proposed by one of the referees, we should modify U in a neighborhood of x so that x is no longer a global minimum of the new function. Ideally, when have knowledge of an initial point x_0, such that all elements of $\mathcal{U}$ are approached with a more or less even probability, then we should repeat the algorithm a corresponding number of times to find all the elements of $\mathcal{U}$.

REMARK 2.7. By Theorem 2.3, the diffusion $X^{(\beta)}$ on N is transient for $\beta > \beta_0$ and positive recurrent if and only if $\beta < \beta_0$, due to the fact that $\int_N U^{-1-\beta}\, d\ell = +\infty$ for $\beta \geqslant \beta_0$. By standard results (see e.g. Kliemann [12], Theorem 3.2 applied with $C = N$), it is then either null recurrent or transient for $\beta = \beta_0$. It would be interesting to investigate this situation.

3. Euclidean computations. This section considers a situation where the state space M is no longer a compact manifold but the Euclidean space $\mathbb{R}^m$, with $m \geqslant 2$. We state a theorem (Theorem 3.1 below) analogous to Theorem 2.3 (i) . This result is interesting in itself, and its proof allows us to explain, in a simple framework, how to characterize the attractiveness/repulsivity of a global minimum. The main idea is to expand a critical point to a sphere, using polar decompositions, following [4].

Let $U : \mathbb{R}^m \to \mathbb{R}_+$ be a smooth function with $\min U = 0$. We assume that for each $p \in \mathcal{U} := U^{-1}(0)$, $\operatorname{Hess} U(p)$ is positive definite. In particular, points in $\mathcal{U}$ are isolated and $\mathcal{U}$ is therefore countable.

For any fixed $\beta \in \mathbb{R}$, as in (1), we are interested in the operator L_β defined on $\mathcal{C}^2(\mathbb{R}^m)$, via

$$\forall\, x \in \mathbb{R}^m, \qquad L_\beta[f](x) := U(x)\triangle f(x) - \beta\langle \nabla U, \nabla f\rangle(x) \tag{7}$$

where $\triangle$, $\langle\cdot,\cdot\rangle$ and ∇, respectively denote, the Euclidean Laplacian, scalar product and gradient. Throughout all this section $\|x\| = \sqrt{\langle x, x\rangle}$ denotes the Euclidean norm of x.

Associated to (7) is the stochastic differential equation

$$dX^{(\beta)}(t) = -\beta\nabla U(X^{(\beta)}(t))dt + \sqrt{2U(X^{(\beta)}(t))}dB_t \tag{8}$$

where $B = (B_t)_{t\geqslant 0}$ is a standard Brownian motion on $\mathbb{R}^m$.

By local Lipschitz continuity of ∇U and $\sqrt{U}$, there exists, for each $x \in \mathbb{R}^m$, a unique solution $X^{(\beta)} : [0, \tau^\infty) \to \mathbb{R}^m$ starting from x, (i.e. $X^{(\beta)}(0) = x$). Here, $0 < \tau^\infty \leqslant \infty$, denotes the explosion time of $X^{(\beta)}$ and is characterized by

$$\tau^\infty > t \Leftrightarrow \|X^{(\beta)}(t)\| < \infty.$$

The set $\mathbb{R}^m\backslash\mathcal{U}$ is invariant, in the sense that for all $t \geqslant 0, x \in \mathbb{R}^m\backslash\mathcal{U}$,

$$\mathbb{P}_x(X^{(\beta)}(t) \in \mathbb{R}^m\backslash\mathcal{U} \,|\tau^\infty > t) = 1.$$

The proof of this last point is the same as the proof of Proposition 2.1 (i) given in the appendix.

THEOREM 3.1. (i) *Suppose $\beta > \beta_0$. Then, for all $x \in \mathbb{R}^m\backslash\mathcal{U}$ and $p \in \mathcal{U}$,*

$$\mathbb{P}_x\left[\limsup_{t\to+\infty} \frac{\ln(\|X^{(\beta)}(t) - p\|)}{t} \leqslant -\Lambda(A_p, \beta)(\beta - \beta_0)\right] > 0 \tag{9}$$

where $A_p, \Lambda(A_p, \beta)$ are defined as in Section 2.

(ii) *Suppose $\beta > \beta_0$, and in addition, that there exist positive constants α, r (possibly depending on β) such that*

$$2\beta_0 U(x) - \beta\langle\nabla U(x), x\rangle \leqslant -\alpha\|x\|^2 \tag{10}$$

whenever $\|x\| \geqslant r$. Then, $\mathcal{U}$ is finite and for all $x \in \mathbb{R}^m\backslash\mathcal{U}$,

$$\sum_{p\in\mathcal{U}} \mathbb{P}_x\left[\limsup_{t\to+\infty} \frac{\ln(\|X^{(\beta)}(t) - p\|)}{t} \leqslant -\Lambda(A_p, \beta)(\beta - \beta_0); \tau^\infty = \infty\right] = 1. \tag{11}$$

(iii) *Suppose $\beta < \beta_0$. Then, for all $p \in \mathcal{U}$ and $x \in \mathbb{R}^m\backslash\{p\}$*

$$\mathbb{P}_x\left[\lim_{t\to\infty} X^{(\beta)}(t) = p\right] = 0.$$

REMARK 3.2. The condition (10) is given for its simplicity. However, the conclusion (11) holds true under the weaker assumption, implied by (10) (see Lemma 3.3 below), that $X^{(\beta)}$ almost surely never explodes (i.e $\tau^\infty = \infty$) and eventually enters a ball $B(0, r)$ containing $\mathcal{U}$ for some $r > 0$.

The remainder of this section is devoted to the proof of Theorem 3.1. We first recall some classical facts about diffusion operators, see e.g. Bakry, Gentil and Ledoux [1]. The **carré du champ** Γ_L associated to a Markov generator L defined on an algebra $\mathcal{A}(L)$ is the bilinear functional defined on $\mathcal{A}(L) \times \mathcal{A}(L)$ via

$$\forall\ f, g \in \mathcal{A}(L), \qquad \Gamma_L[f, g] \coloneqq L[fg] - fL[g] - gL[f]$$

(we will denote $\Gamma_L[f] \coloneqq \Gamma_L[f, f]$).

The generator L is said to be of **diffusion**, if $\mathcal{A}(L)$ is stable by composition with smooth functions and if we have

$$L[\varphi(f)] = \varphi'(f)L[f] + \frac{\varphi''(f)}{2}\Gamma_L[f] \tag{12}$$

for any $f \in \mathcal{A}(L)$ and any function φ smooth on the image of f.

In this situation we also have, with the same notations,

$$\Gamma_L[\varphi(f)] = (\varphi'(f))^2 \Gamma_L[f] \tag{13}$$

The Markov generator given in (7) is of diffusion with $\mathcal{A}(L_\beta) = \mathcal{C}^2(\mathbb{R}^m)$. The corresponding carré du champ is given by

$$\forall\ f \in \mathcal{C}^2(\mathbb{R}^m), \qquad \Gamma_{L_\beta}[f] = 2U\|\nabla f\|^2. \tag{14}$$

Our first goal is to show that, under condition (10), $X^{(\beta)}$ never explodes and always enter the ball $B(0,r)$. For all $s \geqslant 0$, we let

$$\tau_s = \inf\{t \geqslant 0\ : \|X^{(\beta)}(t)\| \leqslant s\} \text{ and } \tau^s = \inf\{t \geqslant 0\ : \|X^{(\beta)}(t)\| \geqslant s\}.$$

Note that these stopping times depend on β, but to shorten notation we omit this dependance in their definition.

LEMMA 3.3. *Under the condition (10),*

$$\mathbb{P}_x(\tau^\infty = \infty; \tau_r < \infty) = 1$$

for all $x \in \mathbb{R}^m$ and r is as in (10).

PROOF. Let $V : \mathbb{R}^m \to \mathbb{R}$ be a smooth function coinciding with $\ln(\|x\|^2)$ for $\|x\| \geqslant r$. Using the formulae (12) and (14) it comes that, for all $\|x\| \geqslant r$,

$$L_\beta(V)(x) = \frac{2}{\|x\|^2}\left(2\beta_0 U(x) - \beta\langle \nabla U(x), x\rangle\right) \leqslant -2\alpha.$$

In particular, for all $x \in \mathbb{R}^m$, $L_\beta(V)(x) \leqslant C$ where $C = \sup_{\{x\in\mathbb{R}^m\,:\|x\|\leqslant r\}} |L_\beta(V)(x)|$. Thus, by Ito's formulae, for all $k \geqslant 1$,

$$\begin{aligned}
\ln(k^2)\mathbb{P}_x(\tau^k \leqslant t) &\leqslant \mathbb{E}_x(V(X^{(\beta)}(t \wedge \tau^k)) \\
&= V(x) + \mathbb{E}_x\left[\int_0^{t\wedge\tau^k} L_\beta[V](X^{(\beta)}(s))ds\right] \\
&\leqslant V(x) + tC.
\end{aligned}$$

This shows that $\mathbb{P}_x(\tau^k \leqslant t) \to 0$, as $k \to \infty$. Hence $\mathbb{P}_x(\tau^\infty < \infty) = 0$.

Now, by Ito formulae again, the process $(M_t)_{t\geqslant 0}$ defined as

$$M_t := V(X^{(\beta)}(t \wedge \tau_r)) - \ln(r^2) - \int_0^{t\wedge\tau_r} L_\beta V(X^{(\beta)}(s))ds \geqslant 2\alpha(t \wedge \tau_r)$$

is a nonnegative $\mathbb{P}_x$ local martingale. A nonnegative local martingale may not be a martingale but is always a supermartingale (Le Gall [15], Proposition 4.7). Thus $2\alpha\mathbb{E}_x(t \wedge \tau_r) \leqslant \mathbb{E}_x(M_t) \leqslant V(x) - \ln(r^2)$. Hence $\mathbb{E}_x(\tau_r) < \infty$. □

Our next goal is to investigate the behavior of $X^{(\beta)}$ around a critical point $p \in \mathcal{U}$. Without loss of generality, we assume that $p = \{0\}$. We let $A = \mathrm{Hess}\, U(0)$. Fix $\epsilon \in (0,1)$ small enough so that $\mathcal{U} \cap B(0,\epsilon) = \{0\}$. Write any $x \in B(0,\epsilon)\backslash\{0\}$ under its polar decomposition $x = \rho\theta$ with $\rho \in (0,\epsilon)$ and $\theta \in \mathbb{S}^{m-1}$. This decomposition induces the mapping

$$Q : \mathcal{C}^2(B(0,\epsilon)) \ni f \mapsto Q[f] \in \mathcal{C}^2((0,\epsilon) \times \mathbb{S}^{m-1})$$

with

$$\forall\ (\rho,\theta) \in (0,\epsilon) \times \mathbb{S}^{m-1}, \qquad Q[f](\rho,\theta) \coloneqq f(\rho\theta) \tag{15}$$

Endow $\mathbb{S}^{m-1}$ with its usual Riemannian structure, inherited from $\mathbb{R}^m$, and denote $\langle \cdot, \cdot \rangle_\theta$, ∇_θ, div_θ and $\triangle_\theta$ the corresponding scalar product, gradient, divergence, and Laplace-Beltrami operator. Note that $\langle \cdot, \cdot \rangle_\theta$ is just the restriction of $\langle \cdot, \cdot \rangle$ to the tangent space of $\mathbb{S}^{m-1}$ at θ.

Classical computations in polar coordinates show that for any $f, g \in \mathcal{C}^2(B(0,\epsilon))$, we have on $(0,\epsilon) \times \mathbb{S}^{m-1}$,

$$Q[\langle \nabla f, \nabla g \rangle] = \partial_\rho Q[f] \partial_\rho Q[g] + \frac{1}{\rho^2} \langle \nabla_\theta Q[f], \nabla_\theta Q[g] \rangle_\theta ,$$

$$Q[\triangle f] = \partial_\rho^2 Q[f] + \frac{m-1}{\rho} \partial_\rho Q[f] + \frac{1}{\rho^2} \triangle_\theta Q[f].$$

It leads us to introduce the operator L_β on $\mathcal{C}^2((0,\epsilon) \times \mathbb{S}^{m-1})$ defined by

$$\mathsf{L}_\beta[\cdot] := \mathsf{U} \left(\partial_\rho^2 [\cdot] + \frac{m-1}{\rho} \partial_\rho [\cdot] + \frac{1}{\rho^2} \triangle_\theta [\cdot] \right) - \beta \left((\partial_\rho \mathsf{U}) \partial_\rho [\cdot] + \frac{1}{\rho^2} \langle \nabla_\theta \mathsf{U}, \nabla_\theta [\cdot] \rangle_\theta \right) \tag{16}$$

where $\mathsf{U} := Q[U]$. Indeed, on $\mathcal{C}^2(B(0,\epsilon))$, we have the intertwining relation

$$\mathsf{L}_\beta \circ Q = Q \circ L_\beta .$$

LEMMA 3.4. *The operator L_β extends to a diffusion operator, still denoted L_β, on $\mathcal{C}^2([0,\epsilon) \times \mathbb{S}^{m-1})$, whose associated diffusion process $\mathsf{X}^{(\beta)}$ leave $\{0\} \times \mathbb{S}^{m-1}$ invariant. On $\{0\} \times \mathbb{S}^{m-1}$, identified with $\mathbb{S}^{m-1}$, $\mathsf{X}^{(\beta)}$ admits for generator the operator G_β acting on $\mathcal{C}^2(\mathbb{S}^{m-1})$ via*

$$\forall\, \mathfrak{f} \in \mathcal{C}^2(\mathbb{S}^{m-1}), \qquad G_\beta[\mathfrak{f}] := \frac{1}{2} \Psi_A^{1+\beta} \mathsf{div}_\theta (\Psi_A^{-\beta} \nabla_\theta f) \tag{17}$$

where $\forall\ \theta \in \mathbb{S}^{m-1}, \Psi_A(\theta) = \langle \theta, A\theta \rangle$. Furthermore, G_β has a unique invariant probability measure on $\mathbb{S}^{m-1}$, given by $\mu_{A,\beta}$ (see Equation (2)).

PROOF. Our assumptions on U imply that, uniformly over $\theta \in \mathbb{S}^{m-1}$,

$$\lim_{\rho \to 0_+} \frac{\mathsf{U}(\rho,\theta)}{\rho^2} = \frac{1}{2} \langle \theta, A\theta \rangle , \tag{18}$$

$$\lim_{\rho \to 0_+} \frac{\partial_\rho \mathsf{U}(\rho,\theta)}{\rho} = \langle \theta, A\theta \rangle , \tag{19}$$

$$\lim_{\rho \to 0_+} \frac{\nabla_\theta \mathsf{U}(\rho,\theta)}{\rho^2} = A\theta - \langle \theta, A\theta \rangle \theta . \tag{20}$$

Indeed, by the usual expansion of U around 0, we have

$$\begin{aligned} U(x) &= U(0) + \langle \nabla U(0), x \rangle + \frac{1}{2} \langle x, \mathrm{Hess}\, U(0) x \rangle + o(\langle x, x \rangle) \\ &= \frac{1}{2} \langle x, Ax \rangle + o(\langle x, x \rangle) \end{aligned}$$

which translates into

$$\mathsf{U}(\rho,\theta) = \frac{\rho^2}{2} \langle \theta, A\theta \rangle + o(\rho^2)$$

leading to the first announced limit (18). Similarly,

$$\begin{aligned}\nabla U(x) &= \nabla U(0) + \mathrm{Hess}\, U(0)x + o(\sqrt{\langle x, x\rangle}) \\ &= Ax + o(\sqrt{\langle x, x\rangle}).\end{aligned}$$

At $x = \rho\theta$ with $\rho > 0$, $\partial_\rho \mathsf{U}(\rho,\theta)\theta$ is the radial part of $\nabla U(x)$ and $\nabla_\theta \mathsf{U}(\rho,\theta)/\rho$ is the tangential part. It follows that

$$\begin{aligned}\partial_\rho \mathsf{U}(\rho,\theta) &= \langle \nabla U(x), \theta\rangle, \\ \frac{\nabla_\theta \mathsf{U}(\rho,\theta)}{\rho} &= \nabla U(x) - \partial_\rho \mathsf{U}(\rho,\theta)\theta,\end{aligned}$$

and we get

$$\begin{aligned}\frac{\partial_\rho \mathsf{U}(\rho,\theta)}{\rho} &= \langle \theta, A\theta\rangle + o(1), \\ \frac{\nabla_\theta \mathsf{U}(\rho,\theta)}{\rho^2} &= A\theta - \langle \theta, A\theta\rangle\theta + o(1),\end{aligned}$$

leading to the wanted second and third results (19) and (20).

It follows that for any $F \in \mathcal{C}^2([0,\epsilon) \times \mathbb{S}^{m-1})$, we have, uniformly over $\theta \in \mathbb{S}^{m-1}$,

$$\lim_{\rho \to 0_+} \mathsf{L}_\beta[F](\rho,\theta) = \frac{1}{2}\langle \theta, A\theta\rangle \triangle_\theta F(0,\theta) - \beta \langle A\theta - \langle \theta, A\theta\rangle\theta, \nabla_\theta F(0,\theta)\rangle_\theta \tag{21}$$

Denoting $\mathsf{L}_\beta[F](0,\theta)$ the r.h.s. enables us to see L_β as a diffusion operator on $[0,\epsilon) \times \mathbb{S}^{m-1}$, whose associated diffusion process $\mathsf{X}^{(\beta)}$ leaves $\{0\} \times \mathbb{S}^{m-1}$ invariant, and such that on $\{0\} \times \mathbb{S}^{m-1}$, identified with $\mathbb{S}^{m-1}$, its generator coincides with the operator defined by

$$G_\beta(\mathfrak{f}) := \langle \theta, A\theta\rangle \left(\frac{1}{2}\triangle_\theta \mathfrak{f}(\theta) - \beta \langle b(\theta), \nabla_\theta \mathfrak{f}\rangle_\theta \right)$$

where

$$\forall\, \theta \in \mathbb{S}^{m-1}, \qquad b(\theta) := \frac{A\theta - \langle \theta, A\theta\rangle\theta}{\langle \theta, A\theta\rangle} = \frac{1}{2}\nabla_\theta \ln(\langle \theta, A\theta\rangle).$$

It is easily checked that G_β can be rewritten under the divergence form given by (17). This divergence form implies that the probability measure $\mu_{A,\beta}$ defined in (2) is invariant. By ellipticity of G_β there is no other invariant probability measure.

□

LEMMA 3.5. *Suppose $\beta > \beta_0$ and $0 < \lambda < \Lambda(A,\beta)$. For all $0 < \eta \leqslant 1$, there exists $0 < \epsilon_1$ such that for all $\|x\| \leqslant \epsilon_1$,*

$$\mathbb{P}_x \left[\limsup_{t \to +\infty} \frac{\ln(\|X^{(\beta)}(t)\|)}{t} \leqslant -\lambda(\beta - \beta_0); \right] \geqslant 1 - \eta. \tag{22}$$

If now, $\beta < \beta_0$, then for all $x \in \mathbb{R}^m \backslash \{0\}$,

$$\mathbb{P}_x \left[\lim_{t \to +\infty} \|X^{(\beta)}(t)\| = 0 \right] = 0.$$

PROOF. The proof follows from the stochastic persistence approach used in [4], [2]. For reader's convenience it is presented in details in the appendix (Section A.2). Let V be the function defined on $(0,\epsilon)\times\mathbb{S}^{m-1}$ via

$$\mathsf{V}(\rho,\theta) := -\ln(\rho). \tag{23}$$

We claim that:

(a) $\mathsf{L}_\beta[\mathsf{V}]$ can be extended into a continuous function H_β on $[0,\epsilon)\times\mathbb{S}^{m-1}$;
(b) $\Gamma_{\mathsf{L}_\beta}[\mathsf{V}]$ is bounded on $(0,\epsilon)\times\mathbb{S}^{m-1}$; and
(c) $\mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)] = \Lambda(A,\beta)(\beta-\beta_0)$ (the l.h.s. is a shorthand for the integration of the mapping $\mathbb{S}^{m-1}\ni u\mapsto \mathsf{H}_\beta(0,u)$ with respect to $\mu_{A,\beta}$, defined in (2)) .

Using the form of L_β (equation (16)) and the equalities (18), (19), (a) holds true with

$$\mathsf{H}_\beta(0,\theta) = (\beta-\beta_0)\langle\theta, A\theta\rangle \tag{24}$$

and (c) directly follows from the definition of $\Lambda(A,\beta)$. For (b) , the definition of L_β and $\Gamma_{\mathsf{L}_\beta}$, lead to

$$\forall\ f\in\mathcal{C}^2((0,\epsilon)\times\mathbb{S}^{m-1}),\qquad \Gamma_{\mathsf{L}_\beta}[f] = 2\mathsf{U}\left((\partial_\rho f)^2+\frac{1}{\rho^2}|\nabla_\theta f|^2\right).$$

Thus,

$$\Gamma_{\mathsf{L}_\beta}[\mathsf{V}] = 2\frac{\mathsf{U}(\rho,\theta)}{\rho^2}$$

which is bounded in view of (18). This concludes the proof of the claim.

If $\beta>\beta_0, \mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)]>0$ and the first assertion of the lemma follows from Theorem A.13 (based on Theorem 5.4 in [4]). If $\beta<\beta_0, \mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)]<0$, the second assertion follows Proposition A.9 (i) in Appendix A.2 applied on the manifold $[0,\epsilon)\times\mathbb{S}^{m-1}$ with extinction set $\{0\}\times\mathbb{S}^{m-1}$.

□

We can now conclude the proof of Theorem 3.1. We start with assertion (ii) . Fix $\beta>\beta_0$. For $n\in\mathbb{N}$ sufficiently large (so that $\Lambda(A_p,\beta)>\frac{1}{n}$) and $p\in\mathcal{U}$, let $\mathcal{E}_n(p)$ be the event defined as

$$\mathcal{E}_n(p) = \left\{\limsup_{t\to+\infty}\frac{\ln(\|X^{(\beta)}(t)-p\|)}{t}\leqslant -(\Lambda(A_p,\beta)-\frac{1}{n})(\beta-\beta_0)\right\},$$

and let

$$\mathcal{E}_n = \bigcup_{p\in\mathcal{U}}\mathcal{E}_n(p).$$

The set $\mathcal{U}$ is finite, since (10) cannot be satisfied by a point $x\in\mathcal{U}$ and by consequence $\mathcal{U}$ is included into the compact ball centered at 0 and of radius r. Thus there exists, by Lemma 3.5, $\epsilon_1>0$ such that

$$\mathbb{P}_x(\mathcal{E}_n(p))\geqslant\frac{1}{2}$$

for all $x\in B(p,\epsilon_1)$ and all $p\in\mathcal{U}$. Let

$$\mathcal{U}_{\epsilon_1} := \bigcup_{p\in\mathcal{U}} B(p,\epsilon_1)$$

and $\tau_{\mathcal{U}_{\epsilon_1}} := \inf\{t \geqslant 0 \ : X^{(\beta)}(t) \in \mathcal{U}_{\epsilon_1}\}$. By ellipticity of L_β on $\mathbb{R}^m \backslash \mathcal{U}$, $\mathcal{U}_{\epsilon_1}$ is open and *accessible* from all $x \in \mathbb{R}^m$, in the sense that $\mathbb{P}_x(X^{(\beta)}(t_x) \in \mathcal{U}_{\epsilon_1}) > 0$ for some $t_x \geqslant 0$. Thus, by Feller continuity and compactness of $\overline{B(0,r)}$, there exists $\delta > 0$ such that

$$\mathbb{P}_x(\tau_{\mathcal{U}_{\epsilon_1}} < \infty) \geqslant \delta$$

for all $x \in \overline{B(0,r)}$. Combined with Lemma 3.3, this proves that $\mathbb{P}_x(\tau_{\mathcal{U}_{\epsilon_1}} < \infty) \geqslant \delta$ for all $x \in \mathbb{R}^m$. Thus,

$$\mathbb{P}_x(\mathcal{E}_n) \geqslant \delta/2$$

for all $x \in \mathbb{R}^m$. The strong Markov property, implies that $\mathbb{P}_x(\mathcal{E}_n) = 1$. Hence

$$\mathbb{P}_x(\bigcap_n \mathcal{E}_n) = 1.$$

This concludes the proof of (ii) .

We now pass to the proof of (i) . Fix $\beta > \beta_0$ and assume without loss of generality that $p = \{0\}$. Since $\mathbb{P}_x(\tau_{\epsilon_1} < \infty) > 0$ for all $x \in \mathbb{R}^m \backslash \{0\}$, the proof of (9) follows from Lemma 3.5 .

Finally (iii) is an immediate consequence of the second part of Lemma 3.5 (recall that 0 was an arbitrary point of $\mathcal{U}$, up to a translation).

4. Proof of Theorem 2.3.

4.1. *Proof of Theorem 2.3 (i)* . The proof is similar to that of Theorem 3.1. We begin by proving a Riemannian version of Lemma 3.5. The proof of Theorem 2.3 (i) will then follow by an argument similar to that given at the end of Section 3.

Let $y \in \mathcal{U}$ and let $B_M(y,\epsilon)$ be the Riemannian ball with center y and radius ϵ, where $\epsilon > 0$ is sufficiently small so that

- the only critical point for U in $B_M(y,\epsilon)$ is y,
- the exponential mapping $\exp_y : T_yM \to M$ is a diffeomorphism between the tangent ball $B(0,\epsilon)$ of T_yM and $B_M(y,\epsilon)$.

Recall that the exponential mapping $\exp_y : T_yM \to M$ associates to any tangent vector $v \in T_yM$ the point $x \in M$ which is the position at time 1 of the (constant speed) geodesic starting at time 0 from y with speed v.

Consider $(e_1, e_2, ..., e_m)$ an orthonormal basis of T_yM consisting of eigenvectors associated to the eigenvalues $(\lambda_1, \lambda_2, ..., \lambda_m)$ of the Hessian of U at the critical point y. A priori this Hessian is a bilinear form on T_yM, but the Euclidean structure of T_yM enables us to see it as a symmetric endomorphism on T_yM, and $(\lambda_1, \lambda_2, ..., \lambda_m)$ and $(e_1, e_2, ..., e_m)$ correspond to its spectral decomposition.

Let $(v_1, v_2, ..., v_m)$ be the coordinate system associated to $(e_1, e_2, ..., e_m)$ on $B(0,\epsilon)$. Such a coordinate system based on the exponential mapping is said to be a normal coordinate system. From now on and until the end of this section, we identify a map $f : B_M(y,\epsilon) \to \mathbb{R}$ with $f \circ \exp_y : B(0,\epsilon) \to \mathbb{R}$, and write $f(v)$ for $f \circ \exp_y(v)$. Under this identification, the matrix corresponding to the Hessian at y admits the classical form

$$(\partial^2_{k,l} U(0))_{k,l \in [\![m]\!]}$$

where ∂_k is a shorthand for $\frac{\partial}{\partial v_k}$. The introduction of the lecture notes of Pennec [19] is a convenient reference for these assertions (a more thorough exposition can be found in the book of Gallot, Hulin and Lafontaine [9]).

A first interest of the normal coordinate system $(v_1, v_2, ..., v_m)$ on $B(0,\epsilon)$ is that we can consider the corresponding polar decomposition as in the previous section: each $v = (v_1, v_2, ..., v_m) \in B(0,\epsilon)\backslash\{0\}$ can be uniquely written under the form $\rho\theta$ with $\rho \in (0,\epsilon)$ and $\theta \in \mathbb{S}^{m-1}$, where the basis $(e_1, e_2, ..., e_m)$ enables us to identify T_yM with $\mathbb{R}^m$.

Before going further, let us recall some other traditional notations and facts from Riemannian geometry. For any $v \in B(0,\epsilon)$, denote $g(v) := (g_{k,l}(v))_{k,l\in[\![m]\!]}$ the matrix of the pull-back of the Riemannian metric: for any vectors b and $\tilde{b}$ from $T_{\exp_y(v)}M$, identified with their coordinates $(b_k)_{k\in[\![m]\!]}$ and $(\tilde{b}_k)_{k\in[\![m]\!]}$ in the basis $(\partial_k)_{k\in[\![m]\!]}$, we have

$$\left\langle b, \tilde{b} \right\rangle_v = \sum_{k,l\in[\![m]\!]} g_{k,l}(v) b_k \tilde{b}_l$$

The determinant of $g(v)$ and the inverse matrix $g^{-1}(v)$ are respectively denoted $|g|(v)$ and $(g^{k,l}(v))_{k,l\in[\![m]\!]}$. For any smooth function f, the expressions of its gradient and Laplacian are given by

$$\begin{aligned}
\nabla f(v) &= \left(\sum_{l\in[\![m]\!]} g^{k,l}(v) \partial_l f(v) \right)_{k\in[\![m]\!]} \\
\triangle f(v) &= \frac{1}{\sqrt{|g|(v)}} \sum_{k,l\in[\![m]\!]} \partial_k \left(\sqrt{|g|} g^{k,l} \partial_l f \right)(v) \\
&= \sum_{k,l\in[\![m]\!]} g^{k,l}(v) \left(\partial^2_{k,l} f(v) - \sum_{j\in[\![m]\!]} \Gamma^j_{k,l}(v) \partial_j f(v) \right)
\end{aligned}$$

where $\Gamma^j_{k,l}(v)$ are the Christoffel symbols at v, see for instance the listing [21] (again we abuse notation in the r.h.s by identifying f with its formulation in the coordinate system $v = (v_1, v_2, ..., v_m)$). There should be no confusion between the traditional uses of the letter Γ both for the carré du champ (taking a generator in index) and for the Christoffel symbols (with two indices and one exponent).

A second interest of the normal coordinate system is that at 0, we recover the usual notions: $g(0)$ is the identity matrix and the Christoffel symbols all vanish at 0.

The above expressions lead to the following formulation of the generator L_β defined in (1):

$$L_\beta[\cdot] = U \sum_{k,l\in[\![m]\!]} g^{k,l} \left(\partial^2_{k,l}[\cdot] - \sum_{j\in[\![m]\!]} \Gamma^j_{k,l} \partial_j[\cdot] \right) - \beta \sum_{k,l\in[\![m]\!]} g^{k,l} \partial_k U \partial_l[\cdot] \tag{25}$$

Again we are slightly abusing notations by calling it L_β too, especially as we see it as only defined on $\mathcal{C}^2(B(0,\epsilon))$.

The associated carré du champ is given as

$$\Gamma_{L_\beta}[\cdot] = 2U \sum_{k,l\in[\![m]\!]} g^{k,l} \partial_k[\cdot] \partial_l[\cdot] \tag{26}$$

(this is a consequence of the algebraic relation $\Gamma_{\partial_k\partial_l}[\cdot] = 2\partial_k[\cdot]\partial_l[\cdot]$, even if $\partial_k\partial_l$ is not a Markov generator, i.e. when $k \neq l$).

Consider the mapping Q associated in (15) to the polar decomposition. Since Q is invertible from $\mathcal{C}^2(B(0,\epsilon))$ to $\mathcal{C}^2((0,\epsilon)\times\mathbb{S}^{m-1})$, there is a unique diffusion generator L_β acting on $\mathcal{C}^2((0,\epsilon)\times\mathbb{S}^{m-1})$ such that

$$\mathsf{L}_\beta \circ Q = Q \circ L_\beta$$

To compute L_β, let us write that for any $v \in B(0,\epsilon)\backslash\{0\}$,

$$\rho = \sqrt{\sum_{k\in[\![m]\!]} v_k^2}$$

$$\forall\, l \in [\![m]\!], \qquad \theta_l = \frac{v_l}{\rho}$$

It follows that for any $k \in [\![m]\!]$,

$$\partial_k \rho = \frac{v_k}{\rho} \;=\; \theta_k$$

$$\forall\, l \in [\![m]\!], \qquad \partial_k \theta_l = \frac{\delta_{k,l}}{\rho} - \frac{v_l}{\rho^2}\partial_k \rho \;=\; \frac{1}{\rho}(\delta_{k,l} - \theta_k\theta_l)$$

where $\delta_{k,l}$ is the Kronecker symbol.

It follows that

$$\partial_k = \theta_k \partial_\rho + \frac{1}{\rho}\sum_{l\in[\![m]\!]} (\delta_{k,l} - \theta_k\theta_l)\partial_{\theta_l} \tag{27}$$

and by composition, for any $k,l \in [\![m]\!]$, we can also write $\partial^2_{k,l}$ in terms of ∂_ρ, ∂^2_ρ, ∂_{θ_i} and $\partial^2_{\theta_i,\theta_j}$, for $i,j \in [\![m]\!]$. Replacing these expressions in (25), we get the formula for L_β in terms of differentiations of order 1 and 2, with respect to ρ and the θ_l, $l \in [\![m]\!]$.

In order to apply the general method of [4] as in Section 3, we need to check the three facts respectively listed in the following lemmas.

LEMMA 4.1. *For any $F \in \mathcal{C}^2([0,\epsilon)\times\mathbb{S}^{m-1})$, we have, uniformly over $\theta \in \mathbb{S}^{m-1}$,*

$$\lim_{\rho\to 0_+} \mathsf{L}_\beta[F(\cdot,\cdot)](\rho,\theta) = G_\beta[F(0,\cdot)](\theta)$$

where G_β is given in (17).

PROOF. For any $v \in B(0,\epsilon)$, define

$$\forall\, k,l \in [\![m]\!], \qquad \tilde{g}^{k,l}(v) := \delta_{k,l}$$

$$\forall\, j,k,l \in [\![m]\!], \qquad \tilde{\Gamma}^j_{k,l}(v) := 0$$

and in analogy with (25),

$$\tilde{L}_\beta[\cdot] = U\sum_{k,l\in[\![m]\!]} \tilde{g}^{k,l}\left(\partial^2_{k,l}[\cdot] - \sum_{j\in[\![m]\!]} \tilde{\Gamma}^j_{k,l}\partial_j[\cdot]\right) - \beta\sum_{k,l\in[\![m]\!]} \tilde{g}^{k,l}\partial_k U \partial_l[\cdot]$$

This operator coincides with the restriction of (7) to $B(0,\epsilon)$. It follows from (21) that uniformly over $\theta \in \mathbb{S}^{m-1}$,

$$\lim_{\rho\to 0_+} \tilde{\mathsf{L}}_\beta[F](\rho,\theta) = G_\beta[F(0,\cdot)](\theta)$$

where the operator $\tilde{\mathsf{L}}_\beta$ is such that $\tilde{\mathsf{L}}_\beta \circ Q = Q \circ \tilde{L}_\beta$.

Thus to get the wanted result, it is sufficient to show that

$$\lim_{\rho\to 0_+} (\mathsf{L}_\beta - \tilde{\mathsf{L}}_\beta)[F](\rho,\theta) = 0 \tag{28}$$

This convergence is a consequence of the writing

$$(L_\beta - \tilde{L}_\beta)[F] = U \sum_{k,l\in[\![m]\!]} (g^{k,l} - \tilde{g}^{k,l}) \left(\partial^2_{k,l} F - \sum_{j\in[\![m]\!]} \Gamma^j_{k,l} \partial_j F \right)$$
$$-U \sum_{k,l,j\in[\![m]\!]} \tilde{g}^{k,l} (\Gamma^j_{k,l} - \tilde{\Gamma}^j_{k,l}) \partial_j F - \beta \sum_{k,l\in[\![m]\!]} (g^{k,l} - \tilde{g}^{k,l}) \partial_k U \partial_l F$$

(where the restriction of F on $(0,\epsilon)\times\mathbb{S}^{m-1}$ was identified with $Q^{-1}[F]$ on $B(0,\epsilon)\backslash\{0\}$, with Q given in (15)), and of the following facts, valid uniformly in $\theta\in\mathbb{S}^{m-1}$ as ρ goes to 0_+:

- According to (27), for any $k,l\in[\![m]\!]$, $\partial_k F$ is of order $1/\rho$ and $\partial^2_{k,l}F$ is of order $1/\rho^2$.
- Due to the regularity of g and of the Christoffel symbols, for any $k,l\in[\![m]\!]$, $g^{k,l}-\tilde{g}^{k,l}$ and $\Gamma^j_{k,l}-\tilde{\Gamma}^j_{k,l}$ are of order ρ.
- By the assumption that y is a global minimum, U is of order ρ^2 and $\partial_k U$ is of order ρ, for any $k\in[\![m]\!]$.

□

We have seen in the previous section that G_β is reversible with respect to the probability measure $\mu_{A,\beta}$ defined in (2), where here $A := A_y$ is the diagonal matrix whose entries are the eigenvalues of the Hessian of U at $y\in\mathcal{U}$. To continue the method of [4], we also need the two following ingredients.

LEMMA 4.2. *Consider the function V defined on $(0,\epsilon)\times\mathbb{S}^{m-1}$ via*

$$\mathsf{V}(\rho,\theta) := -\ln(\rho).$$

The function $\Gamma_{\mathsf{L}_\beta}[\mathsf{V}]$ is bounded on $(0,\epsilon)\times\mathbb{S}^{m-1}$ and the function $\mathsf{L}_\beta[\mathsf{V}]$ can be extended into a continuous function H_β on $[0,\epsilon)\times\mathbb{S}^{m-1}$ satisfying (24) and thus $\mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)] = \Lambda(A,\beta)(\beta-\beta_0)$.

PROOF. We have $\Gamma_{\mathsf{L}_\beta}[\mathsf{V}] = Q[\Gamma_{L_\beta}[V]]$ with $V(v) = -\frac{1}{2}\ln(\sum_{k\in[\![m]\!]} v_k^2)$, so it is sufficient to see that $\Gamma_{L_\beta}[V]$ is bounded on $B_M(y,\epsilon)\backslash\{y\}$. Expanding $U(v)$ near 0 in the normal coordinate system $v=(v_1,v_2,...,v_d)$, we get for v small

$$U(v) \sim \frac{1}{2}\sum_{l\in[\![m]\!]} \lambda_l v_l^2$$

Hence, using (26),

$$\Gamma_{L_\beta}[V](v) \sim \frac{\sum_{l\in[\![m]\!]}\lambda_l v_l^2}{(\sum_{l\in[\![m]\!]} v_l^2)^2} \sum_{k,l\in[\![m]\!]} g^{k,l}(0) v_k v_l = \frac{\sum_{l\in[\![m]\!]}\lambda_l v_l^2}{\sum_{l\in[\![m]\!]} v_l^2}.$$

(see also [17]).

This proves the wanted boundedness.

For the wanted convergence, in view of the computations of the previous section, it is sufficient to see that (28) holds with F replaced by V. Note that when applied to a function only depending on ρ, as V, (27) reduce to $\partial_k = \theta_k\partial_\rho$. It follows that $\partial_k\mathsf{V}$ is of order $1/\rho$ and $\partial^2_{k,l}\mathsf{V}$ is of order $1/\rho^2$. This observation enables us to use the same arguments as in the end of the proof of Lemma 4.1 to conclude that (28) holds with F replaced by V. □

A Riemannian version of Lemma 3.5 follows directly from the preceding lemma, the proof being exactly the same as the proof of Lemma 3.5. The proof of Theorem 2.3 (i) then follows (almost) verbatim along the lines of the arguments given in the preceding section just after the proof of Lemma 3.5.

4.2. *Proof of Theorem 2.3 (ii)* . Let $V : N \to \mathbb{R}, x \mapsto \ln(U(x)^{-\beta})$. Observe that for all $f \in C^2(N)$,

$$\operatorname{div}(e^V \nabla f) = e^V (\langle \nabla V, \nabla f \rangle + \Delta f) = U^{-\beta-1} L_\beta f.$$

Let $C_c^2(N)$ be the set of $f \in C^2(N)$ having compact support. Then, for all $f \in C_c^2(N)$,

$$\int_N L_\beta f \, d\ell_\beta = 0,$$

where ℓ_β is the measure on N defined as

$$\ell_\beta(dx) := U(x)^{-(1+\beta)} \ell(dx).$$

Let $p \in \mathcal{U}$. By Morse's lemma, there is a smooth chart at p such that, in this chart system, U writes $x \mapsto \|x\|^2 = \sum_{i=1}^m x_i^2$. Since the map $x \mapsto \|x\|^{-2(\beta+1)}$ is locally integrable (i.e. in a neighborhood of $0_{\mathbb{R}^m}$) if and only if $2(\beta+1) < m$, it comes that $\int_N U(x)^{-(1+\beta)} \ell(dx) < \infty$ if and only if $2(\beta+1) < m$, that is $\beta < \beta_0$.

Assuming $\beta < \beta_0$, the probability measure

$$\pi_\beta(dx) := \frac{1}{C_\beta} \ell_\beta(dx)$$

(where C_β is a normalization constant) satisfies

$$\int_N L_\beta f \, d\pi_\beta = 0 \tag{29}$$

for all $f \in C_c^2(N)$. Observe that there is no evidence that the set $C_c^2(N)$ is a core for $\mathcal{L}_\beta$, so that we cannot immediately deduce from (29) that π_β is an invariant probability measure of $X^{(\beta)}$. However, by Theorem 9.17 page 248 in Ethier and Kurtz [8] (originally due to Echeverria [7]) the following properties (a) to (d) ensure that π_β is invariant:

(a) The space N is a separable locally compact metric space (for which the space $\hat{C}(N)$ of continuous function "vanishing at infinity" coincide with $\{f \in C^0(M) : f|_{\mathcal{U}} = 0\}$);
(b) The set $C_c^2(N)$ is an algebra dense in $\hat{C}(N)$;
(c) The operator $L_\beta : C_c^2(N) \to \hat{C}(N)$, satisfies the positive maximum principle;
(d) The martingale problem for $(L_\beta, C_c^2(N))$ is well-posed: for all $x \in N$, $\mathbb{P}_x^\beta$ (the law of $X^{(\beta)}$ starting from $X^{(\beta)}(0) = x$) is the unique probability on $D([0,\infty), N)$ such that $f(X(t)) - \int_0^t L_\beta f(X(s)) ds$ is a $\mathbb{P}_x^\beta$-martingale and $\mathbb{P}_x^\beta[X(0) = x] = 1$, where $(X(t))_{t \geqslant 0}$ is the canonical process on $D([0,\infty), N)$.

Properties (a) , (b) and (c) are easy to verify. Property (d) follows from, on one hand, that for any $\epsilon > 0$ sufficiently small, the stopped martingale problem on $N_\epsilon := \{x \in M : U(x) \geqslant \epsilon\}$ is well-posed by uniform ellipticity of $L^{(\beta)}$ on N_ϵ, and on the other hand, that these localized martingale problems can next be extended to the whole state space N. For instance, corresponding precise statements are found in Ethier and Kurtz [8], see Theorem 5.4 page 199, providing the existence of a solution of the stopped martingale problem on the N_ϵ, but also of the martingale problem on N, Theorem 4.1 page 182 for the uniqueness of

stopped martingale problems on the N_ϵ, and Theorem 6.2 page 217, for the deduction of the uniqueness of the solution of the martingale problem on N by localization.

• (ii) (a) : follows from the fact that a strong Feller process on a connected space having an invariant probability measure with full support, is positive recurrent (see e.g. [3], Corollary 7.10 for a statement on discrete time Markov chains and Proposition 4.58 (ii) for the application in continuous time). In particular, it is uniquely ergodic (i.e. its invariant probability measure is unique). Here the strong Feller property of $X^{(\beta)}$ on N follows from Proposition 2.1.

• (ii) (b) : The following lemma is a consequence of Lemma 4.2 and the stochastic persistence approach exposed in [2], [4].

LEMMA 4.3. *Assume $\beta < \beta_0$. Then, there exist a continuous map $W : N \to \mathbb{R}^+$, $0 \leqslant \rho < 1$, $\chi > 0$, $\kappa \geqslant 0$ and $T > 0$ such that*

(i) $W(x) = d(x,\mathcal{U})^{-\chi}$ *on a neighborhood of* $\mathcal{U}$,
(ii) $P_T^{(\beta)} W \leqslant \rho W + \kappa$.

PROOF. For $y \in \mathcal{U}$, and $\epsilon > 0$ sufficiently small, let $V_y : M\backslash\{y\} \to \mathbb{R}^+$ be a smooth map such that

$$Q[V_y \circ \exp_y](\rho,\theta) = \mathsf{V}(\rho,\theta) := -\ln(\rho)$$

whenever $\rho < \epsilon$, where, using the notation of Section 4.1, $\mathsf{V} : (0,\epsilon) \times \mathbb{S}^{m-1} \to \mathbb{R}$ is as in Lemma 4.2 and Q is the mapping induced by the polar decomposition as in (15). Because $\Gamma_{\mathsf{L}_\beta}[\mathsf{V}]$ is bounded on $(0,\epsilon) \times \mathbb{S}^{m-1}$ and $\mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)] = \Lambda(A,\beta)(\beta - \beta_0) < 0$, it is possible, for ϵ sufficiently small, to find numbers $\chi, T > 0, \kappa$ and $0 \leqslant \rho < 1$ such that

$$P_T^{(\beta)}(e^{\chi V_y}) \leqslant \rho e^{\chi V_y} + \kappa$$

on $M\backslash\{y\}$. This follows from Proposition A.9 (i) in Appendix A.2 (based on [2], Proposition 8.2). The mapping $W : N \to \mathbb{R}^+$, defined as $W(x) = \sum_{y\in\mathcal{U}} e^{\chi V_y}$ satisfies the conditions of the lemma. □

By ellipticity of $L^{(\beta)}$ on N, every point $p \in N$ is an *accessible Doeblin* point for $P_T^{(\beta)}$. Combined with the preceding lemma this proves assertion (ii) (b) of Theorem 2.3 (see e.g. Theorem 8.15 in [3]).

4.3. *Proof of Theorem 2.3 (iii).* It follows from compactness of M and Feller continuity of $X^{(\beta)}$ that, with $\mathbb{P}_x$ probability one, every limit point (for the weak topology) of the family

$$\left\{ \frac{1}{t} \int_0^t \delta_{X_s^{(\beta)}} ds \right\}_{t\geqslant 0}$$

is an invariant probability of $X^{(\beta)}$ (see e.g. [3], Theorem 4.20 combined with Propositions 4.57 and 4.58). It then suffices to show that for $\beta = \beta_0$, every invariant probability of $X^{(\beta_0)}$ is supported by $\mathcal{U}$, or equivalently, that every ergodic probability measure of $X^{(\beta_0)}$ is a Dirac measure δ_p for some $p \in \mathcal{U}$. We proceed by contradiction. Suppose that there exists an ergodic probability measure of $X^{(\beta_0)}$, μ with $\mu(N) > 0$. Then $\mu(N) = 1$ (by invariance of N) and, by ellipticity of $X^{(\beta_0)}$ on N, μ is absolutely continuous with respect to $\ell(dx)$, hence also with respect to $\ell_{\beta_0}(dx)$. That is $\mu(dx) = f(x)\ell_{\beta_0}(dx)$ with $f \geqslant 0$ measurable and $\ell_{\beta_0}[f] = 1$. We claim that f is almost surely constant. This is in contradiction with the fact that $\ell_{\beta_0}(N) =$

∞. It remains to prove the claim. First assume that $\|f\|_\infty = \sup_{x\in N}|f(x)| < \infty$. Then, $f \in L^2(\ell_{\beta_0})$ because $\ell_{\beta_0}[f^2] = \mu[f] \leqslant \|f\|_\infty$. Thus,

$$\ell_{\beta_0}[(P_t^{\beta_0} f - f)^2] = \ell_{\beta_0}[(P_t^{\beta_0} f)^2 + g]$$

where $g \coloneqq f^2 - 2fP_t^{\beta_0} f \in L^1(\ell_{\beta_0})$ and $\ell_{\beta_0}[g] = -\mu[f]$. Thus,

$$\ell_{\beta_0}[(P_t^{\beta_0} f - f)^2] = \ell_{\beta_0}[(P_t^{\beta_0} f)^2] - \mu(f) = \ell_{\beta_0}[(P_t^{\beta_0} f)^2 - f^2] \leqslant 0$$

where the last inequality follows from Jensen's inequality. This shows that ℓ_{β_0}-almost surely, $P_t^{\beta_0} f = f$, and also μ-almost surely. By ergodicity f is μ-almost surely constant. Suppose now that $\|f\|_\infty = \infty$. Set $f_n = \min\{f, n\}$ and $\mu_n(dx) = f_n(x)\ell_{\beta_0}(dx)$. For every Borel set $A \subset N$,

$$\begin{aligned}\mu_n P_t^{\beta_0}(A) &= \mu_n P_t^{\beta_0}(A \cap \{f \leqslant n\}) + (\mu_n P_t^{\beta_0})(A \cap \{f > n\}) \\ &\leqslant (\mu P_t^{\beta_0})(A \cap \{f \leqslant n\}) + n(\ell_{\beta_0} P_t^{\beta_0})(A \cap \{f > n\}) \\ &= \mu(A \cap \{f \leqslant n\}) + n\ell_{\beta_0}(A \cap \{f > n\}) = \mu_n(A).\end{aligned}$$

This shows that μ_n is excessive, hence invariant because every finite excessive measure is invariant (see e.g. [3], Lemma 4.25). By what precedes, f_n is μ-almost surely constant. Thus f is μ-almost surely constant. This concludes the proof of the claim.

APPENDIX

A.1. On the over-parametrized model in Machine Learning and mini-batch stochastic gradient approximations. Here we present the heuristic reason why the fraudulent algorithm investigated in this paper can be seen as an idealized model of an asymptotic behavior encountered in the theory of Machine Learning. This exposition is based on the two papers of Wojtowytsch [23] and [24].

Assume we are given a finite family $(x_n, y_n)_{n\in[\![N]\!]}$ of feature-class couples from the product of Euclidean spaces $\mathbb{R}^k \times \mathbb{R}^l$, with $k, l \in \mathbb{N}$. We are looking for a function $f(\theta, \cdot)$ from $\mathbb{R}^k$ to $\mathbb{R}^l$, parametrized by some $\theta \in \Theta$, so that

$$U(\theta) \coloneqq \sum_{n\in[\![N]\!]} \|y_n - f(\theta, x_n)\|^2 \tag{30}$$

is minimal (where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^l$). The over-parametrized setting corresponds to families $(f(\theta, \cdot))_{\theta\in\Theta}$ sufficiently large so we are sure there exists at least one $\theta \in \Theta$ such that $U(\theta) = 0$. Thus we know a priori that $\min_\Theta U = 0$ and the goal is then to find a minimizing $\theta \in \Theta$.

To find such a global minimum of U, a first try is to consider the classical gradient descent algorithm, namely the dynamic system $(\theta(t))_{t\geqslant 0}$ whose evolution in Θ is given by

$$\forall\, t \geqslant 0, \qquad \frac{d\theta}{dt}(t) = -\nabla U(\theta(t)) \tag{31}$$

where we suppose that Θ is endowed with a Riemannian structure and that U is at least $\mathcal{C}^1$. We start from an arbitrary $\theta(0) \in \Theta$ (this will also be the case for all the subsequent evolutions).

Assuming furthermore that Θ is the Euclidean space $\mathbb{R}^m$, with $m \in \mathbb{N}$ (or the torus $(\mathbb{R}/\mathbb{Z})^m$ to be in a compact framework), for the purpose of implementing the above evolution on a

computer, it is preferable to replace it by an approximating time-discretization $(\theta(p))_{p\in\mathbb{Z}_+}$ such as

$$\forall\, p\in\mathbb{Z}_+,\qquad \theta(p+1)=\theta(p)-\eta\nabla U(\theta(p)) \tag{32}$$

where the positive $\eta>0$ is the time-step size (in the Machine Learning context, η is also called the learning rate and is often also depending on the time $p\in\mathbb{Z}_+$): the discrete time p in (32) rather corresponds to the continuous time ηp in (31).

From (30), we compute that the gradient $\nabla U(\theta)\coloneqq(\partial_{\theta_i}U(\theta))_{i\in[\![m]\!]}$ is given by

$$\forall\,\theta\in\Theta,\ \forall\, i\in[\![m]\!],\qquad \partial_{\theta_i}U(\theta)=-2\sum_{n\in[\![N]\!]}\langle y_n-f(\theta,x_n),\partial_{\theta_i}f(\theta,y_n)\rangle$$

where $\langle\cdot,\cdot\rangle$ is the scalar product in $\mathbb{R}^l$. Thus in (32), at each instant $p\in\mathbb{Z}_+$, a sum of N terms must be computed. To avoid having to make so many calculations, as in practice N is very large, one usually resorts to mini-batches. First the time domain is restricted to $[\![0,\sqrt{N}+1]\!]$ (to simplify notations, assume from now on that N is a square) on which the iteration (32) is modified via

$$\forall\, p\in[\![0,\sqrt{N}]\!],\qquad \theta(p+1)=\theta(p)-\eta\nabla U_p(\theta(p)) \tag{33}$$

where

$$\forall\, p\in[\![0,\sqrt{N}]\!],\ \forall\,\theta\in\Theta,\qquad U_p(\theta)\coloneqq\sum_{n\in[\![p\sqrt{N}+1,(p+1)\sqrt{N}]\!]}\|y_n-f(\theta,x_n)\|^2$$

Let us now assume that (x_n,y_n), for $n\in[\![N]\!]$, are independent samples from a law μ on $\mathbb{R}^k\times\mathbb{R}^l$. Define for any $n\in[\![\sqrt{N}]\!]$, the random variable $Z(n)\coloneqq(Z_i(n))_{i\in[\![m]\!]}$ taking values in $\mathbb{R}^m$ via

$$\forall\, i\in[\![m]\!],\qquad Z_i(n)\coloneqq-2\langle y_n-f(\theta(p),x_n),\partial_{\theta_i}f(\theta(p),x_n)\rangle$$

where $p\in[\![0,\sqrt{N}-1]\!]$ is such that $n\in[\![p\sqrt{N}+1,(p+1)\sqrt{N}]\!]$.

Fix $p\in[\![0,\sqrt{N}-1]\!]$. For $n\in[\![p\sqrt{N}+1,(p+1)\sqrt{N}]\!]$, the $Z(n)$ are independent and identically distributed according to the image of μ by the mapping

$$\mathbb{R}^k\times\mathbb{R}^l\ni(x,y)\mapsto-2(\langle y-f(\theta(p),x),\partial_{\theta_i}f(\theta(p),x)\rangle)_{i\in[\![m]\!]}\in\mathbb{R}^m$$

We compute that for any $i\in[\![m]\!]$,

$$\begin{aligned}\mathbb{E}[\partial_{\theta_i}U_p(\theta(p))]&=\sum_{n\in[\![p\sqrt{N}+1,(p+1)\sqrt{N}]\!]}\mathbb{E}[Z_i(n)]\\&=-2\sqrt{N}\int\langle y-f(\theta(p),x),\partial_{\theta_i}f(\theta(p),x)\rangle\,\mu(dx,dy)\\&=\partial_{\theta_i}\bar{U}(\theta(p))\end{aligned}$$

where the mapping $\bar{U}$ is defined by

$$\forall\,\theta\in\Theta,\qquad \bar{U}(\theta)\coloneqq\int\|y-f(\theta,x)\|^2\,\mu(dx,dy) \tag{34}$$

Writing

$$R(\theta(p))\coloneqq\nabla U_p(\theta(p))-\nabla\bar{U}(\theta(p))$$

we compute that the covariance matrix $A(\theta(p)) \coloneqq (A_{i,j}(\theta(p)))_{i,j\in[\![m]\!]}$ of $R(\theta(p))$ is given, for any $i,j \in [\![m]\!]$, by

$$\begin{aligned} A_{i,j}(\theta(p)) &= \sum_{n\in[\![p\sqrt{N}+1,(p+1)\sqrt{N}]\!]} \mathbb{E}[(Z_i(n)-\mathbb{E}[Z_i(n)])(Z_j(n)-\mathbb{E}[Z_j(n)])] \\ &= \sqrt{N}a_{i,j}(\theta(p)) \end{aligned}$$

where

$$a_{i,j}(\theta(p)) \coloneqq \left(\int F_{\theta(p),i}F_{\theta(p),j}\,d\mu - \left(\int F_{\theta(p),i},d\mu\right)\left(\int F_{\theta(p),j}\,d\mu\right)\right) \tag{35}$$

with, for any $\iota \in [\![m]\!]$, the mapping $F_{\theta(p),\iota}$ defined on $\mathbb{R}^k \times \mathbb{R}^l$ by

$$\forall\,(x,y)\in\mathbb{R}^k\times\mathbb{R}^l, \qquad F_{\theta(p),\iota}(x,y) \coloneqq -2\langle y - f(\theta(p),x), \partial_{\theta_\iota} f(\theta(p),x)\rangle \tag{36}$$

where we implicitly assumed that the mappings $F_{\theta(p),\iota}$ admits a moment of order two under μ.

The latter hypothesis allows us to apply the central limit theorem to get for N large the weak convergence of $N^{-1/4}R(\theta(p))$ toward a Gaussian distribution of mean 0 and covariance matrix $a(\theta(p)) \coloneqq (a_{i,j}(\theta(p)))_{i,j\in[\![m]\!]}$.

Thus we can, quite heuristically and for large N, rewrite (33) as

$$\begin{aligned} \forall\, p\in[\![0,\sqrt{N}]\!], \qquad \theta(p+1) &= \theta(p) - \eta(\nabla\bar{U}(\theta(p)) + R(\theta(p))) \\ &\simeq \theta(p) - \eta\nabla\bar{U}(\theta(p)) + \eta N^{1/4}\sigma(\theta(p))G(p) \end{aligned}$$

where $(G(p))_{p\in[\![0,\sqrt{N}]\!]}$ are independent standard Gaussian random variables in $\mathbb{R}^m$, and where $\sigma(\theta(p))$ is a (symmetric) matrix-square root of $a(\theta(p))$. This encourages us to take $\eta = 1/\sqrt{N}$, since we get

$$\forall\, p\in[\![0,\sqrt{N}]\!], \qquad \theta(p+1) \simeq \theta(p) - \eta\nabla\bar{U}(\theta(p)) + \sqrt{\eta}\sigma(\theta(p))G(p)$$

and, under appropriate regularity conditions, we recognize an Euler-Maruyama approximation (see for instance Section 9.1 of the book of Kloeden and Platen[13]) of the s.d.e.

$$\forall\, t\in[0,1], \qquad d\bar{\theta}(t) = -\nabla\bar{U}(\bar{\theta}(t))+\sigma(\bar{\theta}(t))\,dW(t) \tag{37}$$

where $(W(t))_{t\in[0,1]}$ is a standard Brownian motion on $\mathbb{R}^m$.

Here we end up with a s.d.e. on the time interval $[0,1]$, but to rather get $[0,T]$, for $T>0$, just consider mini-batches of length $\sqrt{N}/T$ and take $\eta = T/\sqrt{N}$.

Assume that the family $(f(\theta,\cdot))_{\theta\in\Theta}$ is still sufficiently large so there exists some $\theta\in\Theta$ so that $\bar{U}(\theta)=0$, as it was the case when the probability distribution μ was the empirical measure

$$\frac{1}{N}\sum_{n\in[\![N]\!]}\delta_{(x_n,y_n)}$$

(compare (30) with (34)).

Note that when $\theta_0\in\Theta$ is such that $\bar{U}(\theta_0)=0$, then $a(\theta_0)=\sigma(\theta_0)=0$. Indeed, we then have $y=f(\theta_0,x)$, μ-a.s. in $(x,y)\in\mathbb{R}^k\times\mathbb{R}^l$. It follows from (36) that $F_{\theta_0,\iota}=0$ μ-a.s. for all $\iota\in[\![m]\!]$, and we deduce from (35) that $a(\theta_0)=0$.

We can be a little more precise. For any $\theta\in\Theta$, denote $\alpha(\theta)$ the largest eigenvalue of the non-negative definite matrix $a(\theta)$. Define

$$K \coloneqq \sup\left\{\|\partial_{\theta_i} f(\theta,x)\| \,:\, \theta\in\Theta,\, i\in[\![m]\!],\, x\in\mathbb{R}^k\right\}$$

PROPOSITION A.4. *Assuming that $K < +\infty$, we have*

$$\forall\, \theta \in \Theta, \qquad \alpha(\theta) \leqslant 4K\bar{U}(\theta)$$

PROOF. Fix $\theta \in \Theta$ and $v := (v_i)_{i\in[\![m]\!]} \in \mathbb{R}^m$. We compute

$$\begin{aligned}
\sum_{i,j\in[\![m]\!]} v_i a_{i,j}(\theta) v_j &= \sum_{i,j\in[\![m]\!]} \left(\int F_{\theta,i} F_{\theta,j}\, d\mu - \left(\int F_{\theta,i}, d\mu \right) \left(\int F_{\theta,j}\, d\mu \right) \right) v_i v_j \\
&= \int \left(\sum_{i\in[\![m]\!]} v_i F_{\theta,i} \right)^2 d\mu - \left(\sum_{i\in[\![m]\!]} \int v_i F_{\theta,i}\, d\mu \right)^2 \\
&\leqslant \int \left(2 \left\langle y - f(\theta,x), \sum_{i\in[\![m]\!]} v_i \partial_{\theta_i} f(\theta,x) \right\rangle \right)^2 d\mu \\
&\leqslant 4 \int \|y - f(\theta,x)\|^2 \sum_{i\in[\![m]\!]} v_i^2 \left\|\partial_{\theta_i} f(\theta,x)\right\|^2 d\mu \\
&\leqslant 4K \left(\int \|y - f(\theta,x)\|^2 \, d\mu \right) \sum_{i\in[\![m]\!]} v_i^2 \\
&= 4K\bar{U}(\theta) \sum_{i\in[\![m]\!]} v_i^2
\end{aligned}$$

The wanted result follows. □

We deduce, in the sense of $m \times m$ matrices that

$$\forall\, \theta \in \Theta, \qquad \sigma(\theta) \leqslant 2\sqrt{K}\sqrt{\bar{U}(\theta)}\mathrm{Id}$$

where Id is the $m \times m$ identity matrix.

If the above inequality was an equality, we would recover the fraudulent algorithm studied in this paper, up to a linear time change and with $\beta = 1/\sqrt{2K}$.

Nevertheless, there is an important difference with our setting. As mentioned in Wojtowytsch [23], in the over-parametrized framework the set of global minima is usually a manifold (with high dimension and co-dimension) and do not consist of isolated points as under our Morse assumption. Furthermore the matrices $\sigma(\theta)$ are degenerate even for $\theta \in \Theta$ which is not a global minima of $\bar{U}$. Indeed, this paper should be seen as a first step toward the investigation of more general fraudulent algorithms, in particular hypo-elliptic ones. Discrimination between a.s. convergence toward the global minima and their avoidance will then be more involved. In view of the above implications, these extensions deserve to be investigated in future works and our results will serve as a reference of what can be obtained in the simplest, while non-trivial, situations

A.2. Stochastic persistence for diffusion processes. This section briefly presents stochastic persistence theory in the spirit of the papers [2], [4], in the specific case (sufficient for our purposes) of a diffusion process whose extinction set is compact. Although the results presented here can be deduced from those presented in these papers, diffusion properties and compactness of the extinction set simplify certain proofs. Therefore, for the reader's convenience, we have chosen to give a self-contained presentation.

Generalities. Let M be a metric space which can decomposed into $M = M_0 \sqcup M_+$, where

- The subset $M_0 \neq \varnothing$ is compact and is contained in the closure of M_+. It will play the role of the *extinction set*.
- The subset $M_+ \neq \varnothing$ is a n-dimensional smooth manifold, that will necessarily be non-compact, since M_0 is non-empty. It is called the *persistence set*.

On a filtered probability space $(\Omega, (\mathcal{F}_t)_{t\geqslant 0}, \mathbb{P})$, we are given a family of Markov processes $(X^x)_{x\in M}$, with $X^x = (X^x_t)_{t\geqslant 0}$, defined such that for all $x \in M$, $X^x_0 = x$, and $t \to X^x_t$ is continuous (i.e. X^x is a diffusion). Furthermore, denote $\mathbb{P}_x$ the law of X^x, and $(P_t)_{t\geqslant 0}$ the semigroup defined by

$$P_t f(x) = \mathbb{E}_x(f(X_t)),$$

for all f measurable and bounded. We assume that the semigroup $(P_t)_{t\geqslant 0}$ is weak Feller, meaning that $P_t(C_b(M)) \subset C_b(M)$ for all $t \geqslant 0$.

We make the hypothesis that the extinction set M_0 is *invariant*. That is, for all $t \geqslant 0$,

$$P_t \mathbf{1}_{M_0} = \mathbf{1}_{M_0}$$

Invariance of M_0 has the consequence (see e.g [2], Lemma 9.2) that for all $x \in M$

$$x \in M_+ \Leftrightarrow \{X^x_t,\, t \geqslant 0\} \subset M_+.$$

On M_+ the evolution of X^x is governed by a generator L which is a second order differential operator taking the form, in a local coordinate system,

$$L = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}(x)\partial_i\partial_j + \sum_{i=1}^{n} b_i(x)\partial_i \tag{38}$$

where the $a_{ij}(x)$ and $b_i(x)$ are smooth in x, and

$$a_{ij}(x) = \sum_{k=1}^{r} \sigma^i_k(x)\sigma^j_k(x),$$

with $\sigma(x) = (\sigma^i_k(x))_{k=1,\dots,r, i=1,\dots,n}$ is locally Lipschitz. The generator L is understood in the sense of martingale problems: For all $f \in C^2(M)$, $x \in M_+$ and $t \geqslant 0$, define

$$M^f_t(x) \coloneqq f(X^x_t) - f(x) - \int_0^t Lf(X^x_s)ds \tag{39}$$

The process $(M^f_t(x))_{t\geqslant 0}$ is a $(\mathbb{P}, (\mathcal{F}_t)_{t\geqslant 0})$-local martingale.

To control the behavior of the process at infinity, we assume throughout the existence of a proper continuous map $W : M \to \mathbb{R}^+$, which is C^2 on M_+ and satisfies the condition

$$LW \leqslant -aW + b$$

with $a, b > 0$. Note that when M is compact, this assumption is always verified, say with $W = 1$.

REMARK A.5. In the above framework, M_0 may not be regular, nevertheless the evolution of our Markov processes on M_0 can be approximated by the diffusions on M_+, due to the weak Feller assumption. A more regular setting is to assume that M is a n-dimensional smooth manifold with boundary M_0 and that the expression (38) of the generator L is also valid on M_0 (then the vectors $(b_i)_i$ and $(\sigma^i_k)_i$, for $k = 1, ..., r$, as well as their Lie brackets, have to be tangential to M_0 on M_0).

Another similar setting is to assume that M is a is a n-dimensional smooth manifold on which L can be written as

$$L = F_0 + \sum_{i=1}^{m} F_i^2$$

where F_0, F_i are smooth vector fields, and that M_0 is a compact set invariant under the flows induced by the $F_i, i = 0, \dots, m$.

In these cases the processes $(M_t^f(x))_{t\geqslant 0}$ can be defined for all $x \in M$ and $t \geqslant 0$ and are asked to be $(\mathbb{P}, (\mathcal{F}_t)_{t\geqslant 0})$-local martingales. The assumptions on $a(x), b(x)$, now assumed to be valid on M, and the existence of W, ensure that X^x exists as the unique and globally defined solution to a stochastic differential equation. Furthermore, by continuity with respect to the initial condition, $(P_t)_{t\geqslant 0}$ is then necessarily weak Feller. In the application in the main text, we will be in this regular situation. Nevertheless, we present here a generalized setting, as the arguments are in fact the same.

For f, C^2 in a neighborhood of $x \in M_+$, we let

$$\Gamma(f)(x) = (Lf^2)(x) - 2f(x)(Lf)(x) \geqslant 0$$

denote the carré du champ of f at x.

Extinction set and H exponents. We will review sufficient conditions ensuring that the process goes *extinct*, namely that $\mathbb{P}_x(X_t \underset{t\to+\infty}{\longrightarrow} M_0) = 1$ for some (or all) $x \in M$, where $X_t \underset{t\to+\infty}{\longrightarrow} M_0$ means that the distance of X_t to M_0 converges to zero for large $t \geqslant 0$; or *persists*, meaning that $\mathbb{P}_x(X_t \in \cdot)$ converges, as $t \to +\infty$, to some distribution on M_+ for all $x \in M_+$. This will be done under the following key hypothesis.

ASSUMPTION A.6. There exist a C^2 map $V : M_+ \to \mathbb{R}^+$ and a continuous map $H : M \to \mathbb{R}$ such that:

(i) $\lim_{x\to M_0} V(x) = \infty$;
(ii) For all $x \in M_+, LV(x) = H(x)$;
(iii) For some, hence for all, $\delta > 0$

$$\sup_{\{x\in M_+ \,:\, d(x,M_0)\leqslant\delta\}} \Gamma(V)(x) < \infty$$

where d stands for the distance on M.

The key point here is that although V is not defined on M_0, LV extends continuously to M_0.

REMARK A.7. Let (V, H) be as in Hypothesis A.6 and let $\tilde{V}$ be a C^2 function on M_+ which coincides with V on a neighborhood of M_0. Then $(\tilde{V}, \tilde{H})$ also satisfies Hypothesis A.6 with $\tilde{H} = H\mathbf{1}_{M_0} + L\tilde{V}\mathbf{1}_{M_+}$.

In view of this remark, we can (and will) always assume without loss of generality that V (hence H) is zero outside a compact neighborhood of M_0. As a consequence, we can replace (iii) of Hypothesis A.6 by

$$\sup_{\{x\in M_+\}} \Gamma(V)(x) < \infty$$

Here a first interest of Hypothesis A.6:

LEMMA A.8. *Under Hypothesis A.6, the invariance of M_0 (i.e $P_t\mathbf{1}_{M_0} = \mathbf{1}_{\mathbf{M_0}}$ for all $t \geqslant 0$) is equivalent to the apparently weaker condition that M_0 is stable:*

$$P_t\mathbf{1}_{M_0} \geqslant \mathbf{1}_{\mathbf{M_0}}$$

for all $t \geqslant 0$.

PROOF. Let

$$\mathcal{T} \coloneqq \inf\{t \geqslant 0 : X_t^x \notin M_+\} \tag{40}$$

Fix $x \in M_+$ and $T \geqslant 0$. For $\epsilon > 0$, define

$$\mathcal{T}_\epsilon \coloneqq \inf\{t \geqslant 0 : V(X_t) \geqslant 1/\epsilon\}$$
$$T_\epsilon \coloneqq T \wedge \mathcal{T}_\epsilon$$

Using the martingale problem, we can write

$$\begin{aligned} V(X_{T_\epsilon}^x) &= V(x) + \int_0^{T_\epsilon} LV(X_s^x)\,ds + M_{T_\epsilon}^f \\ &\leqslant V(x) + \|LV\|_\infty T_\epsilon + B_{\langle M^f\rangle_{T_\epsilon}} \end{aligned}$$

where $(B_t)_{t\geqslant 0}$ is the Brownian motion provided by Dambis-Dubins-Schwarz's theorem (see e.g. Theorem 5.13 of Le Gall [15]). We have

$$\begin{aligned} \left\langle M^f \right\rangle_{T_\epsilon} &= \int_0^{T_\epsilon} \Gamma[V](X_s^x)\,ds \\ &\leqslant \|\Gamma[V]\|_\infty T_\epsilon \\ &\leqslant \|\Gamma[V]\|_\infty T \end{aligned}$$

Thus we have

$$V(X_{T_\epsilon}^x) \leqslant V(x) + \|LV\|_\infty T + \max_{s\in[0,\|\Gamma[V]\|_\infty T]} B_s$$

Letting ϵ go to zero, we deduce

$$V(X_{T\wedge\mathcal{T}}^x) \leqslant V(x) + \|LV\|_\infty T + \max_{s\in[0,\|\Gamma[V]\|_\infty T]} B_s$$

Since the r.h.s. is finite, we get that $T < \mathcal{T}$. As this is true for all $T \geqslant 0$, it follows that $\mathcal{T} = +\infty$ (a.s.). □

Before introducing the H exponents, we recall a few definitions and facts on invariant measures. Let

$$G = \int_0^\infty e^{-t} P_t\,dt. \tag{41}$$

A probability μ on M is called invariant for $(P_t)_{t\geqslant 0}$ if $\mu P_t = \mu$ for all $t \geqslant 0$. Equivalently (see [3], Proposition 4.57) $\mu G = \mu$. A bounded measurable map g is called (G,μ) invariant if $Gg = g$ μ-almost surely. Invariant probability μ is called *ergodic* if every (G,μ) invariant map is μ-almost surely constant.

We let $P_{erg}(M_0)$ denote the set of ergodic probability measures supported by M_0.

We now can define the H-exponents as:

$$\Lambda^-(H) = -\sup\{\mu H \ : \mu \in P_{erg}(M_0)\}$$

and

$$\Lambda^+(H) = -\inf\{\mu H \ : \mu \in P_{erg}(M_0)\}$$

PROPOSITION A.9. (i) *Assume that* $\Lambda^-(H) > 0$. *Then, for every* $0 < \Lambda^- < \Lambda^-(H)$, *there exists* $T > 0, \theta > 0$, *and* U *a neighborhood of* M_0 *such that*

$$P_T(e^{\theta V})(x) \leqslant e^{\theta V(x)} e^{-T\Lambda^-}$$

for all $x \in U \backslash M_0$ *and* $\sup_{x \in M^+ \backslash M_0} P_T(e^{-\theta V})(x) < \infty$. *Furthermore, for all* $x \in U \backslash M_0$, $(X_t^x)_{t \geqslant 0}$ *eventually leaves* U.

(ii) *Assume that* $\Lambda^+(H) < 0$. *Then, for every* $0 > \Lambda^+ > \Lambda^+(H)$, *there exists* $T > 0, \theta > 0$, *and* U *a neighborhood of* M_0 *such that*

$$P_T(e^{-\theta V})(x) \leqslant e^{-\theta V(x)} e^{T\Lambda^+}$$

for all $x \in U \backslash M_0$.

This proposition is the key tool from which we will deduce persistence (Theorem A.12) and extinction (Theorem A.13) results. We now prove Proposition A.9.

LEMMA A.10. (i) *Let* Λ^- *be as in Proposition A.9* (i). *Then, there exists* $T > 0$ *and a neighborhood* U *of* M_0 *such that for all* $x \in U$,

$$\frac{1}{T} \int_0^T P_s H(x) ds < -\Lambda^- < 0.$$

(ii *) Similarly, let* Λ^+ *be as in Proposition A.9* (i). *Then, there exists* $T > 0$ *and a neighborhood* U *of* M_0 *such that for all* $x \in U$,

$$\frac{1}{T} \int_0^T P_s H(x) ds > -\Lambda^+ > 0.$$

PROOF. By continuity of $x \mapsto P_t H(x)$, it suffices to prove that for some $T > 0$, $\frac{\int_0^T P_s H(x) ds}{T} < -\Lambda^-$ for all $x \in M_0$. Suppose the contrary. Then, for some sequence $(x_n)_n$ included in M_0, we have for all n, $\mu_n H \geqslant -\Lambda^-$ where $\mu_n = \frac{\int_0^n \delta_{x_n} P_s ds}{n}$. By compactness of M_0, $(\mu_n)_n$ is tight (for the weak* topology) and every limit point μ of (μ_n) verifies $\mu H \geqslant -\Lambda^- > -\Lambda^-(H)$. Now it is easily seen that μ is an invariant probability on M_0. Hence, by the ergodic decomposition theorem, it satisfies $\mu H \leqslant -\Lambda^-(H)$. A contradiction. This concludes the proof of the first statement. The second is similar □

We now pass to the proof of Proposition A.9.

PROOF. (i). Using the notation of Lemma A.10, set

$$Y_T^x = \int_0^T H(X_s^x)\, ds - \int_0^T P_s H(x)\, ds,$$

and, up to reducing U,

$$\bar{H} = \sup_{x \in U} \int_0^T P_s H(x) ds < -\Lambda^- T.$$

Taking Lemma A.8 into account, we deduce that for all $\theta > 0$ and $x \in U \backslash M_0$,

$$\exp(\theta V(X_T^x)) = \exp(\theta V(x)) \exp(\theta \int_0^T P_s H(x) ds) \exp(\theta Y_T^x) \exp(\theta M_T^V(x))$$

$$< \exp(\theta V(x)) \exp(\theta T \bar{H}) \exp(\theta Y_T^x) \exp(\theta M_T^V(x)).$$

Thus, by taking the expectation and using Cauchy Schwarz inequality

$$\mathbb{E}(e^{\theta V(X_T^x)}) \leqslant \exp(\theta V(x)) \exp(\theta T \bar{H}) \sqrt{\mathbb{E}(\exp(2\theta Y_T^x))} \sqrt{\mathbb{E}(\exp(2\theta M_T^V(x)))}.$$

Let $\|H\|_\infty = \sup_{x \in M} |H(x)|, \|\Gamma(V)\|_\infty = \sup_{x \in M_+} \Gamma(V)(x)$ and

$$C(T) = \max(4T^2 \|H\|_\infty^2, 2T \|\Gamma(V)\|_\infty).$$

Since $\mathbb{E}(Y_T^x) = 0$ and $|Y_T^x| \leqslant 2T\|H\|_\infty$ a classical estimate (on the log-Laplace function $\theta \to \mathbb{E}(\ln(e^{\theta Y_T^x}))$ leads to

$$\mathbb{E}(\exp(2\theta Y_T^x)) \leqslant \exp(4\theta^2 T^2 \|H\|_\infty^2) \leqslant \exp(\theta^2 C(T)).$$

Now,

$$\mathbb{E}(\exp(2\theta M_T^V(x))) = \mathbb{E}(Z_T(\theta, x) \exp(2\theta^2 \langle M^V(x) \rangle_T)),$$

where

$$\langle M^V(x) \rangle_T = \int_0^T \Gamma(V)(X_s^x) ds \leqslant \frac{C(T)}{2},$$

and

$$Z_T(\theta, x) = \exp(2\theta M_T^V(x) - 2\theta^2 \langle M^V(x) \rangle_T).$$

By Novikov Theorem the local martingale $(Z_t(\theta, x))_{t \geqslant 0}$ is a true martingale. In particular $\mathbb{E}(Z_T(\theta, x)) = 1$. Thus

$$\mathbb{E}(\exp(2\theta M_T^V(x))) \leqslant \exp(\theta^2 C(T)).$$

Finally, we get that

$$P_T(e^{\theta V})(x) \leqslant \exp(\theta V(x)) \exp(\theta(T\bar{H} + \theta C(T))).$$

For θ small enough $T\bar{H} + \theta C(T) \leqslant -T\Lambda^-$. This concludes the proof of assertion (i) for $x \in U$. The same type of estimate (with $\|H\|_\infty$ in place of $\bar{H}$ also shows that $P_T(e^{\theta V})(x)$ is bounded outside of U. The fact that (X_t^x) eventually leaves U whenever $x \in U \backslash M_0$ relies on the following standard argument. Let $\tau = \min\{n \geqslant 0 : X_{nT}^x \notin U\}$. Then $\left(\exp\left(\theta(V(X_{(n \wedge \tau)T}^x) + (n \wedge \tau)T\Lambda^-\right)\right)_{n \geqslant 0}$ is a supermartingale. Since $V \geqslant 0$ it comes that $\mathbb{E}_x(\exp((n \wedge \tau)T\Lambda^-) \leqslant \exp \theta V(x)$, hence, by monotone convergence, $\mathbb{E}_x(\exp \tau(T\Lambda^-)) \leqslant \exp \theta V(x)$ proving that $\mathbb{P}_x(\tau < \infty) = 1$.

The proof of (ii) is similar to the proof of (i). □

The next two results are persistence and extinction consequences of this latter proposition.

Point $p \in M_+$ is called *accessible* from $x \in M_+$ if for every neighborhood U of p $G(x, U) > 0$. Equivalently $p \in \mathsf{supp}(G(x, \cdot))$ where supp denote the topological support of a measure. We say that p is accessible from M_+ if it is accessible from all $x \in M_+$. That is

$$p \in \bigcap_{x \in M_+} \mathsf{supp}(G(x, \cdot)).$$

We say that p is a *Doeblin point* if there exists a non trivial measure $\nu, t_0 > 0$ and a neighborhood O of p such that

$$P_{t_0}(x, \cdot) \geqslant \nu(\cdot)$$

for all $x \in O$. The following proposition follows from classical results

PROPOSITION A.11. *Assume that M_+ is connected and that L is elliptic on M_+, meaning that $a(x)$ is definite positive for all $x \in M_+$. Then every point $p \in M_+$ is Doeblin and accessible from M_+.*

THEOREM A.12. *Suppose $\Lambda^-(H) > 0$ and that there exists a Doeblin point $p \in M_+$ accessible from M_+. Then there exists a unique invariant probability Π such that $\Pi(M_+) = 1$. Furthermore, there exist positive constants C, α and θ, such that for every $f : M_+ \to \mathbb{R}$ measurable and $x \in M_+$*

$$|P_t f(x) - \Pi(f)| \leqslant Ce^{-\alpha t}(1 + \max(e^{\theta V(x)}, W(x)))\|f\|_\theta,$$

where

$$\|f\|_\theta = \sup_{x \in M_+} \frac{|f(x)|}{1 + \max(e^{\theta V(x)}, W(x))}$$

PROOF. The assumption that $LW \leqslant -aW + b$ (on M_+) implies that $P_T(W) \leqslant e^{-aT}W + \frac{b}{a}$ on M_+ (see for instance [3], Lemma 7.26). Consider $\tilde{W} \coloneqq \exp(\theta V) + W$ on M_+. Then, it comes that

$$P_T\tilde{W} \leqslant e^{-a'T}\tilde{W} + \beta$$

on M_+ with $a' = \min(a, \Lambda^-)$ and $\beta \geqslant 0$. The rest of the proof now follows from Theorem 8.15 in [3]. □

We say that M_0 is accessible from $x \in M_+$ if $\mathsf{supp}(G(x, \cdot)) \cap M_0 \neq \varnothing$, and accessible from M_+ if it is accessible from every $x \in M_+$. The following result is basically (up to a few details) the same as Theorem 5.4 in [4].

THEOREM A.13. *Let Λ^+ be as Proposition A.9 (ii). Let $\mathcal{A}$ be the event that $\liminf_{t\to\infty} \frac{V(X_t)}{t} \geqslant -\Lambda^+$.*

(i) *For every $0 < \eta \leqslant 1$, there exists a neighborhood $\tilde{U}$ of M_0 such that*

$$\mathbb{P}_x(\mathcal{A}) \geqslant 1 - \eta$$

for all $x \in \tilde{U}$.

(ii) *If M_0 is accessible from M_+, then*

$$\mathbb{P}_x(\mathcal{A}) = 1$$

for all $x \in M_+$.

PROOF. (i) For $\epsilon > 0$, let $U_\epsilon = \{x \in M_+ : e^{-\theta V(x)} < \epsilon\} \cup M_0$. Choose ϵ small enough so that $U_\epsilon \subset U$ and let $\tau = \min\{n \geqslant 0 : X^x_{nT} \notin U_\epsilon\}$, where U and T are as in Proposition A.9 (ii). In view of Proposition A.9 (ii), for all $x \in U_\epsilon$, the sequence $(e^{-\theta V(X^x_{(n\wedge\tau)T})})_{n\geqslant 0}$ is a $(\mathbb{P}, (\mathcal{F}_{nT})_{n\geqslant 0})$ supermartingale. Hence,

$$\mathbb{E}(e^{-\theta V(X^x_{(n\wedge\tau)T})}\mathbf{1}_{\tau<\infty}) \leqslant \mathbb{E}(e^{-\theta V(X^x_{(n\wedge\tau)T})}) \leqslant e^{-\theta V(x)}.$$

Letting $n \to \infty$ and using dominated convergence shows that

$$\epsilon\mathbb{P}_x(\tau < \infty) \leqslant \mathbb{E}(e^{-\theta V(X^x_{\tau T})}\mathbf{1}_{\tau<\infty}) \leqslant e^{-\theta V(x)}.$$

Thus

$$\mathbb{P}_x(\tau = \infty) \geqslant 1 - \eta$$

for all $0 < \eta \leqslant 1$ and $x \in \tilde{U} := U_{\epsilon\eta}$. We will now show that $\{\tau = \infty\} \subset \mathcal{A}$.

Let

$$\Delta_{n+1} := V(X^x_{(n+1)T}) - V(X^x_{nT}) - \int_0^T P_s H(X^x_{nT})ds$$

$$= \int_0^T H(X^x_{nT+s})ds + (M^V_{(n+1)T}(x) - M^V_{nT}(x)) - \int_0^T P_s H(X^x_{nT})ds.$$

The assumption that $\Gamma(V)$ is bounded makes the local martingale $(M^V_t(x))$ a true L^2 martingale (see e.g. Le Gall [15], Theorem 4.13) with quadratic variation

$$\langle M^V(x)\rangle_t = \int_0^t \Gamma(V)(X^x_s)ds \leqslant \|\Gamma(V)\|_\infty t.$$

Therefore,

$$\mathbb{E}(\Delta_{n+1}|\mathcal{F}_{nT}) = 0, \mathbb{E}(\Delta^2_{n+1}|\mathcal{F}_{nT}) \leqslant C(T)$$

with $C(T) = 2(T\|\Gamma(V)\|_\infty + 4T^2\|H\|^2_\infty)$, and consequently, by the strong law of large numbers for discrete time L^2 martingales (see e.g. [3], Theorem A8)

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \Delta_k}{n} = 0$$

almost surely. Therefore,

$$\liminf_{n\to\infty} \frac{V(X^x_{nT})}{nT} \geqslant -\Lambda^+$$

almost surely on the event $\tau = \infty$.

Now, for all $nT \leqslant t < (n+1)T$,

$$\mathbb{E}(|V(X^x_t) - V(X^x_{nT})|^2) \leqslant 2\left((T\|H\|_\infty)^2 + \mathbb{E}(\sup_{nT\leqslant t\leqslant(n+1)T}(M^V_t(x) - M^V_{nT}(x))^2)\right)$$

$$\leqslant 2(T\|H\|_\infty)^2 + 4\|\Gamma(V)\|_\infty T),$$

where the last inequality follows from Doob's inequality for continuous martingales. It then follows that

$$\lim_{n\to\infty} \sup_{nT\leqslant t\leqslant(n+1)T} \left|\frac{V(X^x_t) - V(X^x_{nT})}{nT}\right| = 0$$

almost surely. Thus

$$\liminf_{t\to\infty} \frac{V(X^x_t)}{t} \geqslant -\Lambda^+$$

almost surely on the event $\tau = \infty$.

(ii) Let $1 > \eta > 0$ and $\tilde{U}$ be as in (i). Let $R > 0$ be large enough so that for all $x \in M$, (X^x_t) eventually enters $W_R := \{y \in M : W(y) \leqslant R\}$. The existence of such an R follows from the assumption on W (see e.g Lemma 3.3). By Feller continuity, the map $x \to G(x, \tilde{U})$ is lower semi continuous. Hence by compactness of W_R and accessibility of M_0, there exists $\delta > 0$ such that $G(x, \tilde{U}) \geqslant \delta$ for all $x \in W_R$. It then follows that for all $x \in M$, $\mathbb{P}(\exists t \geqslant 0 : X^x_t \in \widetilde{U}) \geqslant \delta$. Combined with (i), and the strong Markov property, this shows that $\mathbb{P}_x(\mathcal{A}) \geqslant (1-\eta)\delta$. Now, for all $x \in M_+$, $\mathbb{P}_x$ almost surely,

$$\mathbf{1}_{\mathcal{A}} = \lim_{t\to\infty} \mathbb{P}_x(\mathcal{A}|\mathcal{F}_t) = \lim_{t\to\infty} \mathbb{P}_{X_t}(\mathcal{A}) \geqslant (1-\eta)\delta.$$

Thus $\mathbb{P}_x(\mathcal{A}) = 1$. □

A.3. The diffusion process generated by L_β and Proposition 2.1. Here we briefly explain how the diffusion $X^{(\beta)}$ can be constructed and give a proof of Proposition 2.1.

By Nash's embedding theorem, we can assume without loss of generality that M is a Riemannian submanifold of $\mathbb{R}^n$ (equipped with its Euclidean scalar product $\langle\,,\,\rangle$) for some n sufficiently large. For reasons that will become clear shortly, we write $\nabla_M, \triangle_M, \mathsf{div}_M$ the gradient, Laplacian, and divergence on M, and ∇, div, the gradient and divergence on $\mathbb{R}^n$. If F is a smooth vector field on M and $\tilde{F}$ a smooth globally integrable vector field on $\mathbb{R}^n$ such that $\tilde{F}|_M = F$, then $\tilde{F}$ and F, induce operators on $C^1(\mathbb{R}^n)$ and $C^1(M)$ respectively defined by:

$$\tilde{F}(\tilde{f})(x) = \langle \nabla \tilde{f}(x), \tilde{F}(x)\rangle = \frac{d(\tilde{f}\circ\Psi_t(x))}{dt}|_{t=0}$$

for all $\tilde{f} \in C^1(\mathbb{R}^n)$, and $x \in \mathbb{R}^n$;

$$F(f)(x) = \langle \nabla_M f(x), F(x)\rangle = \frac{d(f\circ\Psi_t(x))}{dt}|_{t=0}$$

for all $f \in C^1(M)$, and $x \in M$. In both formulae, $(\Psi^i_t)_{t\in\mathbb{R}}$ denotes the flow on $\mathbb{R}^n$ induced by $\tilde{F}$.

A direct consequence of the right hand side equalities is that

$$\tilde{F}(\tilde{f})|_M = F(f) \tag{42}$$

for every $f \in C^1(M)$ and $\tilde{f} \in C^1(\mathbb{R}^n)$ such that $f = \tilde{f}|_M$.

Let $(e_1, \dots, e_n)$ be the canonical basis of $\mathbb{R}^n$. For $i = 1, \dots, n$ and $x \in M$, let $E_i(x) \in T_xM$ be the orthogonal projection of e_i onto T_xM. Let $\tilde{E}_i$ be a smooth vector field on $\mathbb{R}^n$, having compact support, such that $\tilde{E}_i|_M = E_i$. It is not hard to show that such a vector field exists. One can, for example, proceed as follows. Let $\mathcal{M} \subset \mathbb{R}^n$ be a normal tubular neighborhood of M. Every point $y \in \mathcal{M}$ writes uniquely $y = x + v$ with $x \in M$ and $v \in T_xM^\perp$. The map $r : \mathcal{M} \ni x + v \mapsto x \in M$, is a smooth retraction. It suffices to set $\tilde{E}_i(x) = \eta(x)E_i(r(x))$ if $x \in \mathcal{M}$ and $\tilde{E}_i(x) = 0$ otherwise, where $0 \leqslant \eta \leqslant 1$ is a smooth function with compact support in $\mathcal{M}$ such that $\eta|_M = 1$.

The following, key property, is proved in Stroock [20], Section 4.2.1. For the reader's convenience we provide an alternative short proof.

LEMMA A.14. *For every $f \in C^2(M)$ and $\tilde{f} \in \mathcal{C}^2(\mathbb{R}^n)$, such that $f = \tilde{f}|_M$, one has*

$$\sum_{i=1}^N \tilde{E}_i{}^2(\tilde{f})|_M = \triangle_M(f)$$

PROOF. Let F be a $\mathcal{C}^1$ vector field on M, and $\tilde{F}$ a $\mathcal{C}^1$ vector field on $\mathbb{R}^n$ such that $\tilde{F}|_M = F$. For all $x \in \mathbb{R}^n$ $\mathsf{div}\tilde{F}(x)$ equals the trace of the Jacobian matrix $D\tilde{F}(x)$, while for all $x \in M$, $\mathsf{div}_M F(x)$ equals the trace of the $d \times d$ matrix $(\langle D\tilde{F}(x)u_i, u_j\rangle)_{i,j}$ where $u_1, \dots, u_d$ is an (arbitrary) orthonormal basis of T_xM. This has the interesting consequence that

$$\mathsf{div}_M(F) = \mathsf{div}(F\circ r)|_M$$

where $r : \mathcal{M} \to M$ is the retraction defined above. Let $f \in \mathcal{C}^2(M)$. Then,

$$\nabla_M f = \sum_{i=1}^n \langle \nabla_M f, e_i\rangle e_i = \sum_{i=1}^n \langle \nabla_M f, E_i\rangle e_i = \sum_{i=1}^n E_i(f)e_i.$$

Thus,

$$\triangle_M f := \mathsf{div}_M(\nabla_M f) = \mathsf{div}(\nabla_M(f)\circ r)|_M = \sum_{i=1}^n \mathsf{div}[(E_i(f)\circ r)e_i]|_M$$

$$= \sum_{i=1}^n \langle \nabla(E_i(f)\circ r)|_M, e_i\rangle = \sum_{i=1}^n \langle \nabla_M E_i(f), e_i\rangle = \sum_{i=1}^n E_i^2(f).$$

Here we have used the fact that $\nabla(f\circ r)|_M = \nabla_M f$ for all $f\in\mathcal{C}^1(M)$. □

Now, let $\tilde{U}:\mathbb{R}^n\to\mathbb{R}_+$ be a smooth function such that $\tilde{U}|_M = U$, $\sqrt{\tilde{U}}$ is Lipschitz and $\nabla\tilde{U}$ has compact support. For instance $\tilde{U}(x) = \eta(x)U(r(x)) + 1 - \eta(x)$ for $x\in\mathcal{M}$ and $\tilde{U}(x) = 1$ otherwise, where η, r are as above. Here, the Lipschitz continuity of $\sqrt{\tilde{U}}$ follows from the fact that r is smooth and that, by assumption, the zeroes of U are non-degenerate.

Consider the stochastic differential equation on $\mathbb{R}^n$ defined by

$$\begin{aligned} dX(t) &= (-\beta - \frac{1}{2})\nabla\tilde{U}(X(t))dt \\ &\quad + \sum_{i=1}^n \left(\frac{1}{2}\langle\nabla\tilde{U}(X(t)), \tilde{E}_i(X(t)\rangle\tilde{E}_i(X(t)) + \tilde{U}(X(t))D\tilde{E}_i(X(t))\cdot\tilde{E}_i(X(t))\right)dt \\ &\quad + \sqrt{2\tilde{U}(X(t))}\sum_{i=1}^n \tilde{E}_i(X(t))dB^i(t) \end{aligned} \tag{43}$$

where $B = (B^1(t),\dots,B^n(t))_{t\geqslant 0}$ is a n-dimensional Brownian motion with $B(0) = 0$.

Since the coefficients of (43) are globally Lipschitz and bounded, the following properties (a) , (b) and (c) are classical (see e.g. Le Gall [15], Theorems 8.3 and 8.7 for (a) and (b) and Kunita [14], Theorem 4.5.1 for (c)) :

(a) For all $x\in\mathbb{R}^n$, there is a unique strong solution $\mathbb{R}_+\ni t\mapsto X^{(\beta,x)}(t)$ to (43) such that $X^{(\beta,x)}(0) = x$,

(b) The process $X^{(\beta)} := (X^{(\beta,x)})_{x\in\mathbb{R}^n}$ is a Feller Markov process on $\mathbb{R}^n$ whose generator $\tilde{\mathcal{L}}_\beta$ contains $\mathcal{C}_c^2(\mathbb{R}^n)$, the set of compactly supported $\mathcal{C}^2$ functions, in its domain and such that for all $\tilde{f}\in\mathcal{C}_c^2(\mathbb{R}^n)$,

$$\begin{aligned} \tilde{\mathcal{L}}_\beta(\tilde{f}) &= -\beta\langle\nabla\tilde{U},\nabla\tilde{f}\rangle - \frac{1}{2}\langle\nabla\tilde{U},\nabla\tilde{f}\rangle + \frac{1}{2}\sum_{i=1}^n \tilde{E}_i[\tilde{U}]\tilde{E}_i[\tilde{f}] + \tilde{U}\sum_{i=1}^n \tilde{E}_i^2(\tilde{f}) \\ &= -\beta\nabla\tilde{U}(\tilde{f}) - \frac{1}{2}\nabla\tilde{U}(\tilde{f}) + \frac{1}{2}\sum_{i=1}^n \tilde{E}_i[\tilde{U}]\tilde{E}_i[\tilde{f}] + \tilde{U}\sum_{i=1}^n \tilde{E}_i^2(\tilde{f}) \end{aligned} \tag{44}$$

(c) The map $x\mapsto X^{(\beta,x)}(t)$ is an homeomorphism. In particular,

$$\forall t\geqslant 0,\ X^{(\beta,x)}(t)\in\mathbb{R}^n\backslash\mathcal{U} \Leftrightarrow \exists t\geqslant 0,\ X^{(\beta,x)}(t)\in\mathbb{R}^n\backslash\mathcal{U}.$$

Set $S_i(x) = \sqrt{2\tilde{U}(x)}\tilde{E}_i(x)$. On $\mathbb{R}^n\backslash\mathcal{U}$, (43) can be rewritten, using Stratonovich formalism, as

$$\begin{aligned} dX(t) &= \left((-\beta - \frac{1}{2})\nabla\tilde{U}(X(t)) + \frac{1}{2}\sum_{i=1}^n DS_i(X(t))S_i(X(t))\right)dt + \sum_{i=1}^n S_i(X(t))dB^i(t) \\ &= (-\beta - \frac{1}{2})\nabla\tilde{U}(X(t)) + \sum_{i=1}^n S_i(X(t))\circ dB^i(t). \end{aligned} \tag{45}$$

The vector fields $\nabla\tilde{U}$ and S_i's being tangent to N, this latter expression shows that N (hence M) is invariant for $X^{(\beta)}$. That is:

$$\forall t \geqslant 0,\ X^{(\beta,x)}(t) \in N(\text{ resp. } M) \Leftrightarrow \exists t \geqslant 0,\ X^{(\beta,x)}(t) \in N(\text{ resp. } M).$$

It then follows that $X^{(\beta)} := (X^{(\beta,x)})_{x\in M}$ is a Feller Markov process on M, leaving N invariant, whose generator $\mathcal{L}_\beta$ contains $\mathcal{C}^2(M)$ in its domain and such that $\mathcal{L}_\beta f = \tilde{\mathcal{L}}_\beta \tilde{f}|_M = L_\beta f$ for all $f \in \mathcal{C}^2(M)$ and $\tilde{f} \in C^2(\mathbb{R}^n)$ such that $\tilde{f}|_M = f$. The last equalities follows from Lemma A.14 and (44), since on M we have

$$\langle \nabla\tilde{U}, \nabla\tilde{f}\rangle = \sum_{i=1}^{n} \tilde{E}_i[\tilde{U}]\tilde{E}_i[\tilde{f}]$$

The strong Feller property on N follows from the ellipticity of L_β on N (see e.g. Ichihara and Kunita [10, 11], Lemma 5.1).

A.4. On Remark 2.2. Given $0 < \lambda_- < \lambda_+$, and $m \geqslant 2$, let $D(\lambda_-, \lambda_+, m)$ be the set of diagonal matrices with entries $\lambda_- = \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_{m-1} \leqslant \lambda_m = \lambda_+$. The set $\{\Lambda(A,\beta)\ : A \in D(\lambda_-,\lambda_+,m)\}$ is a compact interval $[\lambda_-(m,\beta), \lambda_+(m,\beta)]$ (as the image by a continuous map of the compact connected set $D(\lambda_-,\lambda_+,m)$) contained in $[\lambda_-,\lambda_+]$.

Let $A \in D(\lambda_-,\lambda_+,m)$ be the matrix with entries $\lambda_1 = \ldots \lambda_{m-1} = \lambda_-$ and $\lambda_m = \lambda_+$. Then

$$Z(\beta,A) = \int [\lambda_+\theta_m^2 + \lambda_-(1-\theta_m^2)]^{-\beta}\sigma(d\theta) = \mathbb{E}\left[\left(\frac{\lambda_+ X_m^2 + \lambda_-(\sum_{i=1}^{m-1} X_i^2)}{\sum_{i=1}^{m} X_i^2}\right)^{-\beta}\right],$$

where $X_1, \ldots, X_m$ are i.i.d. $\mathcal{N}(0,1)$ random variables. By the strong law of large numbers and dominated convergence, this quantity converges, as $m \to \infty$, toward $\lambda_-^{-\beta}$. Thus $\lim_{m\to\infty} \lambda_-(m,\beta) = \lambda_-$. Similarly, $\lim_{m\to\infty} \lambda_+(m,\beta) = \lambda_+$.

A.5. On spherical integrals. In (23) we could have considered another function V. Indeed, our first choice was

$$\tilde{\mathsf{V}} := -\ln(U)$$

since it seemed somewhat more "intrinsic" with respect to U. It can be shown similarly that the points [a] and [b] following (23) equally hold, with V replaced by $\tilde{\mathsf{V}}$ and H_β by $\tilde{\mathsf{H}}_\beta$ given on $\{0\} \times \mathbb{S}^{m-1}$ by

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad \tilde{\mathsf{H}}_\beta(0,\theta) := -\mathrm{tr}(A) + 2(1+\beta)\frac{\langle\theta, A^2\theta\rangle}{\langle\theta, A\theta\rangle}$$

where we recall that $A := \mathrm{Hess}\, U(0)$.

he sign of the quantity $\mu_{A,\beta}[\tilde{\mathsf{H}}_\beta(0,\cdot)]$ can then be used to discriminate between the attractiveness and repulsivity of 0. In particular $\mu_{A,\beta}[\mathsf{H}_\beta(0,\cdot)]$ and $\mu_{A,\beta}[\tilde{\mathsf{H}}_\beta(0,\cdot)]$ must have the same sign. We tried to prove directly (without success!) that

$$2(1+\beta)\mu_{A,\beta}[\phi_A] > \mathrm{tr}(A) \Leftrightarrow \beta > \frac{m}{2} - 1 \tag{46}$$

$$2(1+\beta)\mu_{A,\beta}[\phi_A] < \mathrm{tr}(A) \Leftrightarrow \beta < \frac{m}{2} - 1 \tag{47}$$

with

$$\forall\ \theta \in \mathbb{S}^{m-1}, \qquad \phi_A(\theta) := \frac{\langle\theta, A^2\theta\rangle}{\langle\theta, A\theta\rangle}$$

A by-product of our computations is thus to show the validity of (46) and (47), which look as natural bounds on the corresponding spherical integrals for any given definite positive matrix A.

Funding. The first author was supported in part by SNF 200020-219913.
The second author was supported in part by ANR-17-EURE-0010 and AFOSR-22IOE016.

REFERENCES

[1] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.
[2] Michel Benaïm. Stochastic Persistence. *arXiv*, 1806.08450, 2018.
[3] Michel Benaïm and Tobias Hurth. *Markov Chains on Metric Spaces A Short Course*, Universitext. Springer, 2022.
[4] Michel Benaïm and Edouard Strickler. Random switching between vector fields having a common zero. *Ann. Appl. Probab.*, 29(1):326–375, 2019.
[5] Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the density case. ArXiv preprint 2204.01306, 2022.
[6] L. E. J. Brouwer. Über Abbildung von Mannigfaltigkeiten. *Math. Ann.*, 71:97–115, 1912.
[7] Pedro Echeverria E. A criterion for invariant measures of Markov processes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete.*, 61:1–16, 1982.
[8] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes, Characterization and Convergence* John Wiley and Sons, 1986.
[9] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*. Universitext. Springer-Verlag, Berlin, third edition, 2004.
[10] Kanji Ichihara, and Hiroshi Kunita. *A classification of the second order degenerate elliptic operators and its probabilistic characterization*, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete.*,30:235–254, 1974.
[11] Kanji Ichihara and Hiroshi Kunita. Supplements and corrections to the paper: A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Z. Wahrscheinlichkeitstheor. Verw. Geb.*, 39:81–84, 1977.
[12] Wolfgang Kliemann. *Recurrence and invariant measures for degenerate diffusions*, *Ann. Probab.*, 15(2):690–707, 1987.
[13] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations.*, volume 23 of *Appl. Math. (N. Y.)*. Berlin: Springer, 4th corrected printing edition, 2010.
[14] Hiroshi Kunita. *Stochastic flows and stochastic differential equations*, volume 24 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 1990
[15] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*, Graduate Texts in Mathematics, Springer, Cham, 2016.
[16] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms. I: Mathematical foundations. *J. Mach. Learn. Res.*, 20:47, 2019. Id/No 40.
[17] Laurent Miclo. On the convergence of global-optimization fraudulent stochastic algorithms. Preprint available at `https://hal.science/hal-04094950`, April 2023.
[18] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of SGD. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15959–15975. PMLR, 17–23 Jul 2022.
[19] Xavier Pennec. Probabilities and statistics on Riemannian manifolds : A geometric approach. Technical Report RR- 5093, inria-00071490, INRIA, 2004.
[20] Daniel W. Stroock *An Introduction to the Analysis of Paths on a Riemmanian Manifold*, Mathematical Surveys and Monographs, Vol 74 American Mathematical Society, 2000.
[21] Wikipedia contributors. List of formulas in Riemannian geometry — Wikipedia, the free encyclopedia, 2023. [Online; accessed 15-October-2023].
[22] John B. Walsh. An introduction to stochastic partial differential equations. École d'été de probabilités de Saint-Flour XIV - 1984, Lect. Notes Math. 1180, 265-437 (1986)., 1986.
[23] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part I: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023.
[24] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part II: Continuous time analysis. *Journal of Nonlinear Science*, 34(1):16, 2023.
[25] Lei Wu, Mingze Wang, and Weijie J Su. The alignment property of SGD noise and how it helps select flat minima: A stability analysis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.