On the convergence of global-optimization fraudulent stochastic algorithms

Laurent Miclo*

Institut de Mathématiques de Toulouse, UMR 5219,
Toulouse School of Economics, UMR 5314
CNRS and University of Toulouse

Abstract

We introduce and analyse the almost sure convergence of a new stochastic algorithm for the global minimization of Morse functions on compact Riemannian manifolds. This diffusion process is called fraudulent because it requires the knowledge of minimal value of the function. Its investigation is nevertheless important, since in particular it appears as the limit behavior of non-fraudulent and time-inhomogeneous swarm mean-field algorithms used in global optimization.

Keywords: Global optimization, stochastic algorithms, diffusion processes on Riemannian manifolds, almost sure convergence, Morse functions, Bessel processes.

MSC2020: primary: 60J60, secondary: 58J65, 90C26, 65C05, 60F15, 60J35, 35K10.

*Fundings from the grants ANR-17-EURE-0010 and AFOSR-22IOE016 are acknowledged.

1 Introduction

To find the global minima of functions also admitting local minima is of great importance, both from a theoretical and practical point of view. Here the smooth context of Morse functions on compact Riemannian manifolds is considered. We introduce a time-homogeneous stochastic algorithm X called **fraudulent** because it requires the knowledge of the minimal value. In some sense it serves as an illustration of the folklore assertion that to know the minimal value of a function and to find a corresponding global minimum are equivalent problems. Nevertheless we found interesting to investigate this algorithm for four reasons:

- The process X is an approximation of the large-time limit behavior of the time-inhomogeneous swarm mean-field algorithm introduced in [1]. The latter algorithm is non-fraudulent since it uses its current distribution to estimate in real time the minimal value.
- The principle behind the convergence of X toward the global minima can be used to devise other non-fraudulent stochastic algorithms based on particles systems that learn adaptatively the minimal value.
- The stochastic algorithm X is useful to find other global minima, once one is known, because as soon as the dimension is larger than or equal to 2, all global minima attract X with a positive probability.
- The stochastic algorithm X is 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 [7], Wu, Wang and Su [13], Mori, Ziyin, Liu and Ueda [8] and Wojtowytsch [12] Mori, Ziyin, Liu and Ueda [8] and references therein.

More precisely, on a compact Riemannian manifold M, of dimension $m \geqslant 1$, let U be a Morse function satisfying $\min_M U = 0$. Recall that a smooth mapping is a Morse function if its Hessian is non-degenerate at each of its critical points (which are the points where the gradient vanishes).

Consider a diffusion $X \coloneqq (X(t))_{t\geqslant 0}$ associated to the generator

$$L_\beta \quad \coloneqq \quad U\Delta \cdot -\beta \langle \nabla U, \nabla \cdot \rangle$$

where Δ, $\langle \cdot, \cdot \rangle$ and ∇ stand for the Laplacian, scalar product and gradient coming from the Riemannian structure, and where β is a real number. Since we want to find the global minima of U, we are more interested in the case $\beta > 0$, where the drift has an attractive (respectively repulsive) effect with respect to the local minima (resp. maxima). But as we are to see, it is convenient to also consider non-positive values of β, where the effects of the drift are reversed.

Due to the Morse assumption on U and the fact that $\min_M U = 0$, we have that $\sqrt{U}$ is a Lipschitz mapping on M (for more details, see Remark 4 below). As a consequence it is possible to construct X, whatever the initial condition, as the unique strong solution to a stochastic differential equation driven by a m-dimensional Brownian motion $B \coloneqq (B(t))_{t\geqslant 0}$ (independent from $X(0)$), see for instance Ikeda and Watanabe [5]. Heuristically, this stochastic differential equation can be written under the Itô's form

$$dX(t) \quad = \quad -\beta \nabla U(X(t))\, dt + \sqrt{2U(X(t))}\, dB(t) \tag{1}$$

or under Stratonovich's form

$$dX(t) \quad = \quad -\left(\beta + \frac{1}{2\sqrt{2}}\right) \nabla U(X(t))\, dt + \sqrt{2U(X(t))} \circ dB(t)$$

where $dB(t)$ has to be isometrically interpreted in the tangent space above $X(t)$ through stochastic parallel displacements. At least when M is a flat torus, the writing (1) is perfectly rigorous.

Consider the set of global minima of U given by

$$\mathcal{U} \;\coloneqq\; \{x \in M \,:\, U(x) = 0\}$$

Note that if $X(0) \in \mathcal{U}$, then X does not move: for any $t \geqslant 0$, we have $X(t) = X(0)$. This observation can be strengthened into the attractiveness of $\mathcal{U}$, which is the main goal of this paper.

We need some additional notations. Under the Morse assumption, $\mathcal{U}$ consists of a finite set of points, say $y_1, y_2, ..., y_N$, with $N \in \mathbb{N}$. For each $n \in [\![N]\!] \coloneqq \{1, 2, ..., N\}$, denote $\lambda_1(n) \leqslant \lambda_2(n) \leqslant \cdots \leqslant \lambda_m(n)$ the positive eigenvalues of the Hessian of U at y_n. Introduce the condition

$$\beta \;>\; \max_{n \in [\![N]\!]} \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_1(n)} - 1 \tag{2}$$

Theorem 1 *Whatever the initial condition $X(0)$, under (2) the limit $X(\infty) \coloneqq \lim_{t \to +\infty} X(t)$ exists a.s. and belongs to $\mathcal{U}$. Furthermore, when $m \geqslant 2$, if X starts from a point $x_0 \notin \mathcal{U}$, we have for any $y \in \mathcal{U}$,*

$$\mathbb{P}_{x_0}[X(\infty) = y] \;>\; 0$$

When $m = 1$, denote y_1 and y_2 the boundary points of the connected component of $M \backslash \mathcal{U}$ containing x_0 (when $\mathcal{U}$ is a singleton, we get $y_1 = y_2$). Then we have

$$\mathbb{P}_{x_0}[X(\infty) = y_1] \;>\; 0, \quad \mathbb{P}_{x_0}[X(\infty) = y_2] \;>\; 0, \quad \forall\, y \in \mathcal{U} \backslash \{y_1, y_2\}, \quad \mathbb{P}_{x_0}[X(\infty) = y] = 0$$

Thus under (2) X is a time-homogeneous stochastic algorithm minimizing globally U and finding all the global minima as soon as $m \geqslant 2$. A.s. convergence was not considered in the previously cited recent works in Machine Learning, which rather concentrated on approximation properties (from discrete time stochastic gradient descent algorithms), on convergence in law and in the behaviour of the associated invariante measures.

It is natural to wonder about the optimality of (2). In this direction, we will show:

Theorem 2 *Assume that*

$$\beta \;<\; \min_{n \in [\![N]\!]} \frac{\sum_{i \in [\![m]\!]} \lambda_i(n)}{2\lambda_m(n)} - 1 \tag{3}$$

Whatever the initial condition $X(0)$, we have

$$\mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] \;=\; 0$$

As a consequence, the value $m/2 - 1$ is critical for β when at each global minimum the Hessian is proportional to the identity:

Corollary 3 *Assume that for each $n \in \mathbb{N}$ we have $\lambda_1(n) = \lambda_m(n)$, then we have*

$$\beta > \frac{m}{2} - 1 \;\;\Rightarrow\;\; \mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] = 1$$
$$\beta < \frac{m}{2} - 1 \;\;\Rightarrow\;\; \mathbb{P}\left[\lim_{t \to +\infty} X(t) \text{ exists and belongs to } \mathcal{U}\right] = 0$$

In particular this result always applies in dimension 1 and we get $-1/2$ as the critical value for β (justifying our consideration of negative β).

In practice we may not have access to the eigenvalues of the Hessian at the global minima, so Condition (2) is difficult to check. An alternative to the appropriate choice of the coefficient β is to

let it depend on the current value of U. More precisely, consider $\zeta : (0,+\infty) \to (0,+\infty)$ a smooth function such that

$$\lim_{u\to 0_+} \zeta(u) = +\infty \tag{4}$$

$$\lim_{u\to 0_+} \sqrt{u}\zeta(u) = 0 \tag{5}$$

We replace the generator L_β by

$$L_\zeta := U\Delta \cdot -\zeta(U)\langle \nabla U, \nabla \cdot \rangle \tag{6}$$

with $\zeta(U(x))\nabla U(x) = 0$ for $x \in \mathcal{U}$ by continuity, due to (5) (this is in fact the only justification for this assumption). Heuristically this generator corresponds to the Itô stochastic differential equation

$$dX(t) = -\zeta(U(X(t))\nabla U(X(t))\, dt + \sqrt{2U(X(t))}\, dB(t)$$

The coefficients of this equation are not globally Lipschitz, nevertheless, we will see that whatever the initial distribution, there is a unique strong diffusion X associated to (6) in the sense of martingale problem. For this process X, Theorem 1 still holds, without Assumption (2). Of course in this situation we lose the critical phenomenon described by Corollary 3. But the advantage is that the diffusion X associated to (6) could be applied to deal with situations less regular than Morse functions.

In the above results, no estimate on the speed of convergence were provided, but it should be possible to remedy this by examining more quantitatively the following arguments. We hope to go further in future investigations.

The plan of the paper is as follows. In the next section we prove Theorem 1 under a stronger assumption than (2). Bessel processes with negative dimension play a pivotal role. In Section 3 the arguments are improved to lead to the desired results. An appendix succinctly recalls the swarm mean-field algorithm of [1] to explain how a fraudulent algorithm can appear from the investigation on non-fraudulent ones. Another simpler and illustrative example is given, even if it is probably less efficient than swarm algorithms.

Acknowledgments:

I would particularly like to thank Marc Arnaudon, Jérôme Bolte and Stéphane Villeneuve for the discussions we had about this paper, as well as referees for their observations about an earlier version of this paper.

2 Proof of a weaker version of Theorem 1

Here we prove Theorem 1 under the stronger assumption

$$\beta > 1 + \sup_{n\in[\![N]\!]} \frac{\sum_{i\in[\![m]\!]} \lambda_i(n)}{2\lambda_1(n)} \tag{7}$$

Its relaxation to (2) will be shown in the next section.

First let us give a sketch of the proof. For any given global minimum y_n, with $n \in [\![N]\!]$, we can find a small radius $r_n > 0$ such that inside the ball $B(y_n, r_n)$, the evolution of $U(X(t))$ is comparable to a time-changed Bessel process with negative dimension. Since a Bessel process with negative dimension a.s. converges to zero in finite time, X will stay in $B(y_n, r_n)$ forever with positive probability and then converge to y_n, if it happens to belong to an even smaller neighborhood V_n of y_n at some time. Nevertheless, the convergence is not expected to occur in finite time due to the time-change, see Example 7 below. It remains to remark that outside $\cup_{n\in[\![N]\!]} V_n$, the diffusion X is elliptic, so it will

end up entering $\cup_{n\in[\![N]\!]}V_n$. Since each times this happens X has a positive chance to converge to a point of $\mathcal{U}$, this event will end up occurring with probability 1. The last assertions of Theorem 1 are consequences of the ellipticity of X outside $\mathcal{U}$.

Let us now develop more precisely the above arguments.

We begin by recalling some general facts about the process X, for any fixed $\beta \in \mathbb{R}$. Its law on the set of continuous trajectories from $\mathbb{R}_+$ to M is uniquely determined by the initial distribution of $X(0)$ and the by fact for any smooth function $f : M \to \mathbb{R}$, the process $M^f := (M^f(t))_{t\geqslant 0}$ defined by

$$\forall\ t \geqslant 0, \qquad M^f(t) \ := \ f(X(t)) - f(X(0)) - \int_0^t L_\beta[f](X(s))\, ds$$

is a continuous martingale (with respect to the filtration generated by X). Furthermore, its bracket (the notation of which should not to be confused with the Riemannian scalar product) is given by

$$\forall\ t \geqslant 0, \qquad \left\langle M^f \right\rangle_t \ = \ 2\int_0^t U(X(s))\, \|\nabla f(X(s))\|^2\, ds \tag{8}$$

where $\|\cdot\|$ stands for the Riemannian norm.

When M is a flat torus, these observations can be deduced from Itô's formula, asserting that:

$$df(X(t)) \ = \ [U(X(t))\Delta f(X(t)) - \beta\langle \nabla U(X(t)), \nabla f(X(t))\rangle]\, dt + \sqrt{2U(X(t))}\, \langle \nabla f(X(t)), dB(t)\rangle$$

so that

$$\forall\ t \geqslant 0, \qquad M^f(t) \ = \ \int_0^t \sqrt{2U(X(s))}\, \langle \nabla f(X(s)), dB(s)\rangle$$

This formula can be generalized for general compact Riemannian manifolds, nevertheless the above martingale problem point of view is more synthetic and enables to avoid delicate geometric constructions. For extensive developments of the martingale problem approach, we refer to the books of Stroock and Varadhan [11] and Ethier and Kurtz [3].

Introduce

$$\sigma^f \ := \ \inf\{t \geqslant 0 \,:\, U(X(t))\, \|\nabla f(X(t))\|^2 = 0\}$$

with the convention that $\sigma^f = +\infty$ when the r.h.s. is the empty set. Assume that $U(X(0))\, \|\nabla f(X(0))\|^2 \neq 0$, so that $\sigma^f > 0$ a.s.

Since for any $t \in [0, \sigma^f)$, we have $U(X(t))\, \|\nabla f(X(t))\|^2 > 0$, we can consider the time change $(\tau_t^f)_{t\in[0,\varsigma)}$ uniquely defined through

$$\forall\ t \in [0, \varsigma^f), \qquad 2\int_0^{\tau_t^f} U(X(s))\, \|\nabla f(X(s))\|^2\, ds \ = \ t$$

where

$$\varsigma^f \ := \ 2\int_0^{\sigma^f} U(X(s))\, \|\nabla f(X(s))\|^2\, ds$$

Define the process Y^f via

$$(Y^f(t))_{t\in[0,\varsigma^f)} \ := \ (f(X(\tau_t^f)))_{t\in[0,\varsigma^f)}$$

Classical time-change theory, see for instance the first section of Chapter 5 of Revuz and Yor [9], Levy's characterization theorem and (8), enable us to construct a Brownian motion $(W(t))_{t\geqslant 0}$ (up to a possible enlargement of the underlying probability space), so that for any time $t \in [0, \varsigma)$,

$$dY^f(t) \ = \ \frac{1}{2}F^f(Y^f(t))\, dt + dW(t)$$

where for any $x \in M$ such that $U(x) \left\|\nabla f(x)\right\|^2 \neq 0$,

$$\begin{aligned} F^f(x) &:= \frac{L_\beta[f](x)}{U(x)\left\|\nabla f(x)\right\|^2} \\ &= \frac{\Delta f(x)}{\left\|\nabla f(x)\right\|^2} - \beta \frac{\langle \nabla U(x), \nabla f(x)\rangle}{U(x)\left\|\nabla f(x)\right\|^2} \end{aligned}$$

Now let us apply these considerations to a particular function f. The first idea is to take $f = U$ (in next section we will see that this is not optimal). To simplify the notations, all the superscripts f are removed in this case. Thus we have

$$\sigma = \inf\{t \geqslant 0 \,:\, X(t) \in \mathcal{C}\}$$

with $\mathcal{C} := \{x \in M \,:\, \nabla U(x) = 0\}$ being the set of critical points of U. Furthermore we consider the process Y given by

$$(Y(t))_{t\in[0,\varsigma)} := (U(X(\tau_t)))_{t\in[0,\varsigma)}$$

with

$$\forall\ t \in [0,\varsigma), \qquad 2\int_0^{\tau_t} U(X(s))\left\|\nabla U(X(s))\right\|^2 ds = t$$

and

$$\varsigma := 2\int_0^{\sigma} U(X(s))\left\|\nabla U(X(s))\right\|^2 ds$$

The evolution of Y is given by

$$dY(t) = \frac{1}{2}F(X(\tau_t))\,dt + dW(t) \tag{9}$$

where $(W(t))_{t\geqslant 0}$ is a Brownian motion and where for any $x \in M\backslash\mathcal{C}$,

$$F(x) := \frac{\Delta U(x)}{\left\|\nabla U(x)\right\|^2} - \frac{\beta}{U(x)}$$

The process $(U(X(t)))_{t\geqslant 0}$ is not Markovian in general, nevertheless we are to show that it can be conveniently compared to a Bessel process while $X(t)$ is close to an element of $\mathcal{U}$.

Indeed, fix a global minimum y_n, with $n \in [\![N]\!]$. Consider an exponential system of coordinates $(x_1, x_2, ..., x_m)$ on a neighborhood $\mathcal{N}_n$ of y_n with $x_1(y_n) = x_2(y_n) = \cdots = x_m(y_n) = 0$ and such that the vectors $(\partial_{x_i})_{i\in[\![m]\!]}$ forms an orthonormal basis of the tangent space at y_n consisting of eigenvectors of the Hessian of U respectively to the eigenvalues $(\lambda_i(n))_{i\in[\![m]\!]}$. Let be given $\epsilon \in (0,1)$, the value of which will be chosen more precisely below. We can find a small enough radius $r_n > 0$, such that the open ball $B(y_n, r_n)$ is included into $\mathcal{N}_n$ and such that for any $x \in B(y_n, r_n)$, identified with its coordinates $(x_1, x_2, ..., x_m)$, we have

$$(1-\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n) \leqslant \Delta U(x) \leqslant (1+\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n) \tag{10}$$

$$(1-\epsilon)\sum_{i\in[\![m]\!]}\lambda_i^2(n)x_i^2 \leqslant \left\|\nabla U\right\|^2(x) \leqslant (1+\epsilon)\sum_{i\in[\![m]\!]}\lambda_i^2(n)x_i^2 \tag{11}$$

$$\frac{1}{2}(1-\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n)x_i^2 \leqslant U(x) \leqslant \frac{1}{2}(1+\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n)x_i^2 \tag{12}$$

Indeed, the first and second estimates are obtained through the expressions of the Laplace-Beltrami and gradient operators in terms of the metric, taking into account that the first order expansion of the metric in a exponential (or geodesic) chart is constant, see for instance Section 10.1 page 58 of Chow, Lu and Ni [2].

Remark 4 For $x \in M\backslash\mathcal{U}$, we have $\nabla\sqrt{U} = \nabla U/(2\sqrt{U})$ and the above bounds enable to see the norm of this vector is bounded on $B(y_n, r_n)\backslash\{y_n\}$, for all $n \in [\![N]\!]$. Since $\left\|\nabla\sqrt{U}\right\|$ is clearly bounded on $M\backslash \cup_{n\in[\![N]\!]} B(y_n, r_n)$, we deduce that $\sqrt{U}$ is Lipschitzian, as announced in the introduction. □

In particular, y_n is the unique critical point of U in $B(y_n, r_n)$. Assume that $X(0) \in V_n$ with

$$V_n \;\coloneqq\; \{x \in B(y_n, r_n)\backslash\{y_n\} \,:\, U(x) < u_n/2\} \tag{13}$$

with

$$u_n \;\coloneqq\; \min_{z\in\partial B(y_n,r_n)} U(z) \tag{14}$$

where $\partial B(y_n, r_n)$ is the boundary of $B(y_n, r_n)$. Denote

$$\sigma_n \;\coloneqq\; \inf\{t \geqslant 0 \,:\, X(t) = y_n \text{ or } X(t) \notin B(y_n, r_n)\} \;\leqslant\; \sigma \tag{15}$$

(as usual, $\sigma_n = +\infty$ when the r.h.s. is the empty set), and

$$\varsigma_n \;\coloneqq\; 2\int_0^{\sigma_n} U(X(s))\,\|\nabla U(X(s))\|^2\,ds$$

Replacing ς by $\varsigma_n \leqslant \varsigma$ in the above considerations enables to investigate the process X while it stays in $B(y_n, r_n/2)\backslash\{y_n\}$. We are led to study $(Y(t))_{t\in[0,\varsigma_n]}$. The following result is important in this respect.

Lemma 5 *We have for any $x \in B(y_n, r_n)\backslash\{y_n\}$,*

$$\left(\frac{(1-\epsilon)^2\sum_{i\in[\![m]\!]}\lambda_i(n)}{2(1+\epsilon)\lambda_m(n)} - \beta\right)\frac{1}{U(x)} \;\leqslant\; F(x) \;\leqslant\; \left(\frac{(1+\epsilon)^2\sum_{i\in[\![m]\!]}\lambda_i(n)}{2(1-\epsilon)\lambda_1(n)} - \beta\right)\frac{1}{U(x)}$$

Proof

For any $x \in B(y_n, r_n)\backslash\{y_n\}$, we have

$$\frac{(1-\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n)x_i^2}{(1+\epsilon)\sum_{i\in[\![m]\!]}\lambda_i^2(n)x_i^2} \;\leqslant\; \frac{2U(x)}{\|\nabla U(x)\|^2} \;\leqslant\; \frac{(1+\epsilon)\sum_{i\in[\![m]\!]}\lambda_i(n)x_i^2}{(1-\epsilon)\sum_{i\in[\![m]\!]}\lambda_i^2(n)x_i^2}$$

implying

$$\frac{(1-\epsilon)}{(1+\epsilon)\lambda_m(n)} \;\leqslant\; \frac{2U(x)}{\|\nabla U(x)\|^2} \;\leqslant\; \frac{(1+\epsilon)}{(1-\epsilon)\lambda_1(n)}$$

since for any real numbers x_i, for $i \in [\![m]\!]$, not all of them vanishing, we have

$$\frac{1}{\lambda_m(n)} \;\leqslant\; \frac{\sum_{i\in[\![m]\!]}\lambda_i(n)x_i^2}{\sum_{i\in[\![m]\!]}\lambda_i^2(n)x_i^2} \;\leqslant\; \frac{1}{\lambda_1(n)}$$

The announced follows, by writting

$$\forall\; x \in B(y_n, r_n)\backslash\{y_n\}, \qquad F(x) \;=\; \left(\Delta U(x)\frac{U(x)}{\|\nabla U(x)\|^2} - \beta\right)\frac{1}{U(x)}$$

■

Recall that β has been chosen so that

$$\beta \;>\; 1 + \sup_{n\in[\![N]\!]} \frac{\sum_{i\in[\![m]\!]} \lambda_i(n)}{2\lambda_1(n)}$$

so we can choose $\epsilon > 0$ so that

$$\delta \;\coloneqq\; \frac{(1+\epsilon)^2 \sum_{i\in[\![m]\!]} \lambda_i(n)}{2(1-\epsilon)\lambda_1(n)} + 1 - \beta$$

is negative.

We deduce from (9) that

$$\forall\; t \in [0, \varsigma_n), \qquad dY(t) \;\leqslant\; dW(t) + \frac{\delta - 1}{2Y(t)}\, dt \tag{16}$$

This inequality leads us to consider $\widetilde{Y} \coloneqq (\widetilde{Y}(t))_{t\geqslant 0}$ solution of the stochastic differential equation

$$\forall\; t \geqslant 0, \qquad d\widetilde{Y}(t) \;=\; dW(t) + \frac{\delta - 1}{2\widetilde{Y}(t)}\, dt \tag{17}$$

starting with $\widetilde{Y}(0) = u_n/2$ (defined in (14)).

The process $\widetilde{Y}$ is a Bessel process with negative dimension $\delta < 0$, for a recent account, see e.g. Le Gall [6]. It hits 0 in (a.s.) finite time and stays at 0 afterward, the strength of the drift not allowing it to escape from 0.

Define

$$\begin{aligned} \theta &\coloneqq \inf\{t \geqslant 0 \,:\, \widetilde{Y}(t) = 0\} \\ \widetilde{\theta} &\coloneqq \inf\{t \geqslant 0 \,:\, \widetilde{Y}(t) = u_n\} \\ p_n &\coloneqq \mathbb{P}[\theta < \widetilde{\theta}] \end{aligned} \tag{18}$$

Since $\widetilde{Y}$ is a Markov process and that $\lim_{t\to+\infty} \widetilde{Y}(t) = 0$ a.s., we have $p_n > 0$.

By comparison, we get

Lemma 6 *Assume that $X(0) \in V_n$ defined in* (13)*. Then we have*

$$\mathbb{P}\left[\forall\; t \geqslant 0,\; X(t) \in B(y_n, r_n) \text{ and } \lim_{t\to+\infty} X(t) = y_n\right] \;\geqslant\; p_n$$

Proof

For $\epsilon \in (0, Y(0))$, define

$$\theta_\epsilon \;\coloneqq\; \inf\{t \geqslant 0 \,:\, \widetilde{Y}(t) = \epsilon\}$$

Applying Theorem (3.7) of Revuz and Yor [9], we get that

$$\forall\; t \in [0, \theta_\epsilon \wedge \widetilde{\theta}), \qquad Y(t) \;\leqslant\; \widetilde{Y}(t)$$

Indeed, this a consequence of $Y(0) \leqslant u_n/2 = \widetilde{Y}(0)$ as well as of the comparison between (16) and (17). Note that in the proof of Theorem (3.7) of Revuz and Yor [9], we need $\widetilde{Y}$ to be Markovian and that its drift is Lipschitz, which is true as long as it belongs to the segment $[\epsilon, u_n]$. But Y is not required to be Markovian, it is sufficient that its drift is adapted.

Letting ϵ go to zero, we deduce

$$\forall\ t \in [0, \theta \wedge \widetilde{\theta}), \qquad Y(t) \ \leqslant\ \widetilde{Y}(t)$$

(one would have remarked that $Y(t)$ remains strictly below u_n and thus $X(\tau_t)$ remains in $B(y_n, r_n)$ as long as $t < \widetilde{\theta}$).

It follows that on the event $\{\theta < \widetilde{\theta}\}$, we have that Y hits 0 in finite time. Equivalently, $X(\tau_t)$ has to hit y_n in finite time, since $X(\tau_t)$ has also to remain in $B(y_n, r_n)$ on $\{\theta < \widetilde{\theta}\}$ and y_n is the unique point of this ball where U vanishes. Recall that if X hits y_n, then it has to stay there forever. Thus on the event $\{\theta < \widetilde{\theta}\}$, $X(t)$ always stays in $B(y_n, r_n)$ and $\lim_{t\to+\infty} X(t) = y_n$. It leads to the announced result. ■

Due to the time change, we cannot deduce the convergence of X to y_n in finite time from the corresponding convergence of $(X_{\tau_t})_{t\geqslant 0}$ (when it occurs). On the contrary, we believe that X never converges in finite time, as suggested by the following caricatural example.

Example 7 Let us momentarily leave the compact setting and consider the function $U : \mathbb{R} \ni x \mapsto x^2/2$. The corresponding diffusion X on $\mathbb{R}$ given by (1) solves the stochastic differential equation

$$\forall\ t \geqslant 0, \qquad dX(t) \ =\ -\beta X(t)dt + |X(t)|dB(t)$$

Assume that $X(0) > 0$ and consider

$$\sigma \ :=\ \inf\{t \geqslant 0 \,:\, X(t) = 0\}$$

While $t \in [0, \sigma)$, we have $X(t) > 0$, so that

$$\forall\ t \in [0, \sigma), \qquad dX(t) \ =\ -\beta X(t)dt + X(t)dB(t)$$

It leads us to introduce $(\widetilde{X}(t))_{t\geqslant 0} := (e^{\beta t}X(t))_{t\geqslant 0}$, satisfying

$$\forall\ t \in [0, \sigma), \qquad d\widetilde{X}(t) \ =\ \widetilde{X}(t)dB(t)$$

whose solution is well-known to be the exponential martingale

$$\forall\ t \in [0, \sigma), \qquad \widetilde{X}(t) \ =\ \widetilde{X}(0)\exp(B(t) - t/2)$$

It follows that

$$\forall\ t \in [0, \sigma), \qquad X(t) \ =\ X(0)\exp(B(t) - (1 + 2\beta)t/2)$$

By consequence we have $\sigma = +\infty$ and the process X does not hit the global minima of U in finite time. Note also that $\lim_{t\to+\infty} X(t) = 0$ as soon as $\beta > -1/2$ in accordance with the observation following Corollary 3.

This example can be transferred on the compact space $\mathbb{R}/(2\pi\mathbb{Z})$ in the following way: consider on $\mathbb{R}/(2\pi\mathbb{Z})$ a Morse function $\widetilde{U}$ coinciding with the above U on $[-\pi/2, \pi/2]$ and such that 0 is a global minimum of $\widetilde{U}$. Let $\widetilde{X}$ be the diffusion evolving as (1), but with U replaced by $\widetilde{U}$. Assume that $\widetilde{X}(0) = \pi/4$ and by contradiction that the corresponding hitting time $\widetilde{\sigma}$ of 0 is finite with positive probability. Taking into account the Markov property, the process $\widetilde{X}$ then stays inside $(-\pi/2, \pi/2)$ and converges to 0 in finite time with a positive probability $p > 0$ (since if starting from $\pi/4$, $\widetilde{X}$ always hits $\pi/2$ before 0, it cannot converge to 0). It follows that if X also starts from $\pi/4$ and uses the same driving Brownian motion, then it coincides with $\widetilde{X}$ for all times with probability p. This implies that X converges to zero in finite time with positive probability, a contradiction.

The fact that the diffusion given by (1) does not hit $\mathcal{U}$ in finite time with positive probability is probably true under the Morse assumption of this paper.

□

The end of the proof of Theorem 1 follows the pattern sketched at the beginning of this section. More precisely, define

$$\begin{array}{rcl} \widetilde{A} & \coloneqq & M \setminus \cup_{n \in [\![N]\!]} B(y_n, r_n) \\ \widehat{A} & \coloneqq & \cup_{n \in [\![N]\!]} V_n \\ p & \coloneqq & \min\{p_n \,:\, n \in [\![N]\!]\} \;>\; 0 \end{array}$$

Whatever the initial distribution of $X(0)$, consider the sequences of stopping times $(\widetilde{\theta}_k)_{k \geqslant 0}$ and $(\widehat{\theta}_k)_{k \geqslant 0}$ defined iteratively from $\widetilde{\theta}_0 = 0$ via

$$\left\{ \begin{array}{rcl} \widehat{\theta}_k & \coloneqq & \inf\{t \geqslant \widetilde{\theta}_k \,:\, X(t) \in \widehat{A}\} \\ \widetilde{\theta}_{k+1} & \coloneqq & \inf\{t \geqslant \widehat{\theta}_k \,:\, X(t) \in \widetilde{A}\} \end{array} \right. \tag{19}$$

(with the usual convention that the infimum of the empty set is $+\infty$).

By the strong ellipticity of X on the compact set $M \setminus \widehat{A}$, for any $k \in \mathbb{N}$ such that $\widetilde{\theta}_k < +\infty$, we have a.s. $\widehat{\theta}_k < +\infty$. On the contrary, we deduce from Lemma 6 and from the strong Markov property that for any $k \in \mathbb{N}$ such that $\widehat{\theta}_k < +\infty$, we end up with $\widetilde{\theta}_{k+1} = +\infty$ with probability p at least. It follows that for any $k \in \mathbb{Z}_+$,

$$\mathbb{P}[\widetilde{\theta}_{k+1} < +\infty \,|\, \widetilde{\theta}_k < +\infty] \;\leqslant\; (1-p)$$

so by iteration we get

$$\forall\; k \in \mathbb{Z}^+, \qquad \mathbb{P}[\widetilde{\theta}_k < +\infty] \;\leqslant\; (1-p)^k$$

and finally

$$\mathbb{P}[\forall\; k \in \mathbb{Z}_+,\, \widetilde{\theta}_k < +\infty] \;=\; 0$$

The fact that a.s. there exists a random $k \in \mathbb{Z}_+$ such that $\widetilde{\theta}_k = +\infty$ ends the proof of the first statement of Theorem 1.

Concerning its second statement, note that when $X(0) \in V_n$, for some $n \in [\![N]\!]$, there is a positive probability that X exits $B(y_n, r_n) \setminus \{y_n\}$ via the boundary $\partial B(y_n, r_n)$.

When $m \geqslant 2$, assume that the $r_n > 0$, for $n \in [\![N]\!]$, have been furthermore chosen so small so that $M \setminus \sqcup_{n \in [\![N]\!]} V_n$ is connected and contains $\sqcup_{n \in [\![N]\!]} \partial B(y_n, r_n)$ (the $\sqcup$ meaning that it is a union of disjoint sets). It follows that if $X(0) \in \sqcup_{n \in [\![N]\!]} \partial B(y_n, r_n)$, then by ellipticity of L_β on the connected set $M \setminus \sqcup_{n \in [\![N]\!]} V_n$, for any $n \in [\![N]\!]$, we have $\mathbb{P}[X(\widehat{\theta}_1) \in B(y_n, r_n)] > 0$ for any given $n \in [\![N]\!]$. Taking into account that $\mathbb{P}[\widetilde{\theta}_2 = +\infty | X(\widehat{\theta}_1) \in B(y_n, r_n)] > 0$, we deduce that $\mathbb{P}[X(\infty) = y_n] > 0$.

When $m = 1$, a similar reasoning leads to the desired result, ending the proof of Theorem 1 under (7).

3 Extensions

After ending the proof of Theorem 1, we show Theorem 2 and next we present the extension to generators of the form (6) and discuss the Morse assumption.

3.1 Under Assumption (2)

We end the proof of Theorem 1. The overall approach is the same but instead of considering $f = U$, we take $f = U^a$, with $a \in (0, 1)$. This exponent being fixed, we remove all the superscripts f from the

notations, as in the previous section when f was equal to U. An immediate drawback with respect to the latter case is that a priori U^a is not smooth. But it is not difficult to go around this problem, since U^a is smooth on $M\backslash\mathcal{U}$ and we are only considering our process up to the time it may reach $\mathcal{U}$.

More precisely, we proceed as in the previous section, assuming that $X(0) \in V_n\backslash\{y_n\}$ for some $n \in [\![N]\!]$. Instead of considering σ_n given in (15), we introduce for any $\epsilon \in (0, U(X(0)))$ the stopping time

$$\sigma_{n,\epsilon} \;\coloneqq\; \inf\{t \geqslant 0 \,:\, X(t) \notin A_{n,\epsilon}\}$$

with

$$A_{n,\epsilon} \;\coloneqq\; \{x \in B(y_n, r_n) \,:\, U(x) \geqslant \epsilon\}$$

Remark that

$$\lim_{\epsilon \to 0_+} \sigma_{n,\epsilon} \;=\; \sigma_n$$

The advantage of $A_{n,\epsilon}$ is that there exist smooth functions f on M coinciding with U^a on $A_{n,\epsilon}$. The considerations at the beginning of Section 2 can be applied to such a function f, up to the time $\sigma_{n,\epsilon}$. By letting ϵ go to 0_+, we get the following result, analogous to (9).

Consider the process Y given by

$$(Y(t))_{t\in[0,\varsigma_n)} \;\coloneqq\; (U(X(\tau_t)))_{t\in[0,\varsigma_n)}$$

with

$$\forall\; t \in [0, \varsigma_n), \qquad 2\int_0^{\tau_t} U(X(s)) \left\|\nabla U^a(X(s))\right\|^2 ds \;=\; t$$

and

$$\varsigma_n \;\coloneqq\; 2\int_0^{\sigma_n} U(X(s)) \left\|\nabla U^a(X(s))\right\|^2 ds$$

The evolution of Y is given by

$$dY(t) \;=\; \frac{1}{2} F(X(\tau_t))\, dt + dW(t)$$

where $(W(t))_{t\geqslant 0}$ is a Brownian motion and for any $x \in M\backslash\mathcal{C}$,

$$\begin{aligned} F(x) \;&\coloneqq\; \frac{\Delta U^a(x)}{\left\|\nabla U^a(x)\right\|^2} - \beta \frac{\langle \nabla U(x), \nabla U^a(x)\rangle}{U(x)\left\|\nabla U^a(x)\right\|^2} \\ &=\; \frac{\Delta U^a(x)}{a^2 U^{2a-2}(x)\left\|\nabla U\right\|^2} - \frac{\beta}{aU^a(x)} \end{aligned}$$

We compute that outside $\mathcal{C}$,

$$\Delta U^a \;=\; aU^{a-1}\Delta U + a(a-1)U^{a-2}\left\|\nabla U\right\|^2$$

so that outside $\mathcal{C}$,

$$F \;=\; \frac{1}{a}\left(\frac{U\Delta U}{\left\|\nabla U\right\|^2} + a - 1 - \beta\right)\frac{1}{U^a}$$

From Lemma 5, we get that a comparison is possible with a Bessel process of asymptotic dimension (i.e. when we let ϵ go to zero in Lemma 5)

$$\delta_a \quad = \quad 2 + \frac{1}{a}\left(\frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_1(n)} - 1 - \beta\right)$$

Thus under (2), by choosing $a > 0$ sufficiently small, we can get $\delta_a < 0$, ending the proof of Theorem 1.

Remark 8 Without resorting to U^a, Condition (2) can also be obtained through the following observation. Coming back to the argument of the previous section, what is important for our purpose is that the Bessel process (17) can hit 0 in finite time, since this implies the same behavior for the process Y. Indeed, once X has hit y_n, it cannot escape from it, so it closes our time-horizon (this is even more true when X does not hit y_n in finite time as suggested by Remark 7). It remains to recall that a Bessel process of dimension δ hits 0 in finite time if and only if $\delta < 2$. This leads to Condition (2). But this argument conceals there is a negative (even as negative as we want) dimension Bessel process behind the scene. □

3.2 Proof of Theorem 2

Up to now we have not taken into account the lower bound in Lemma 5. It can be used to prove there is no convergence toward $\mathcal{U}$. More precisely, assume that (3) holds and let us show we cannot have $\lim_{t\to+\infty} X(t) \in \mathcal{U}$ (except on a $\mathbb{P}$-negligible set).

Indeed, the lower bound in Lemma 5 enables to make a reverse comparison between Y and the Bessel process $\widetilde{Y}$ of (asymptotic) dimension

$$\delta \quad := \quad \min_{n\in[\![N]\!]} \frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_m(n)} + 1 - \beta$$

which is (strictly) larger than 2 under (3). Such a Bessel process diverges to $+\infty$ in large time without hitting 0 (starting from a positive value). It implies that the time σ_n always satisfies that $X(\sigma_n)$ belongs to the boundary of the ball $B(y_n, r_n)$. If we come back to (19), we get that for any $k \in \mathbb{Z}_+$, $\widetilde{\theta}_k < +\infty$. In particular the set $M \setminus \cup_{n\in[\![N]\!]} B(y_n, r_n)$ is visited again after any given time: the convergence to $\mathcal{U}$ is thus forbidden.

Corollary 3 is an immediate consequence of Theorems 1 and 2.

Remark 9 In (3), $\beta > 0$ is only possible for $m \geqslant 3$. This is because

$$\min_{n\in[\![N]\!]} \frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_m(n)} \quad > \quad 1 \tag{20}$$

Since for any $n \in [\![N]\!]$ we have

$$\begin{aligned}\frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_m(n)} \quad &\leqslant \quad \frac{m\lambda_m(n)}{2\lambda_m(n)} \\ &= \quad \frac{m}{2}\end{aligned}$$

condition (20) requires $m \geqslant 3$. □

A little more generally, denote for any $\beta \in \mathbb{R}$,

$$N_-(\beta) \quad := \quad \left\{n \in [\![N]\!] \,:\, \frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_m(n)} - 1 > \beta\right\}$$

$$N_+(\beta) \quad := \quad \left\{n \in [\![N]\!] \,:\, \frac{\sum_{i\in[\![m]\!]}\lambda_i(n)}{2\lambda_1(n)} - 1 < \beta\right\}$$

From the above considerations, we see that X cannot converge to y_n with $n \in N_-(\beta)$, while there is a positive probability that X converges to y_n for $n \in N_+(\beta)$. But this criterion does not strongly discriminate the elements of $\mathcal{U}$, in the sense we cannot find $\beta \in \mathbb{R}$ with $N_-(\beta) \neq \varnothing$ and $N_+(\beta) \neq \varnothing$, due to the inequalities

$$\forall\ n, n' \in [\![N]\!], \qquad \frac{\sum_{i\in[\![m]\!]} \lambda_i(n)}{2\lambda_m(n)} \ \leqslant\ \frac{m}{2} \ \leqslant\ \frac{\sum_{i\in[\![m]\!]} \lambda_i(n')}{2\lambda_1(n')}$$

We don't know if it is possible, in dimension $m \geqslant 2$, to find a Morse function U, $\beta \in \mathbb{R}$ and $n \neq n' \in [\![N]\!]$ such that

$$\mathbb{P}[X(\infty) = y_n] > 0 \quad \text{and} \quad \mathbb{P}[X(\infty) = y_{n'}] = 0$$

3.3 Further extensions

Let us first consider the case of the generator L_ζ defined in (6) under Assumptions (4) and (5). Its coefficients are not globally Lipschitz but they are Lipschitz on any open subset $\mathcal{U}_\epsilon := \{x \in M : U(x) > \epsilon\}$, with $\epsilon \in (0, \max_M U)$. This observation enables to construct by localization X until the first time it hits $\mathcal{U}$. Furthermore the law of the process obtained in this way is uniquely determined until it leaves $\mathcal{U}_\epsilon$, for any $\epsilon \in (0, \max_M U)$, due to the uniqueness of the solution of the corresponding martingale problem. The uniqueness of the law of X until it hits $\mathcal{U}$ follows. Taking into account that starting from $\mathcal{U}$, the process X cannot move (e.g. by using that $L_\zeta U = 0 = L_\zeta U^2$), we get the announced uniqueness of X in the sense of martingale problems.

By localization also, we can construct small neighborhoods of the elements of $\mathcal{U}$, where the evolution of $U(X(t))$ can be compared with that of a Bessel process of negative dimension (as large as we wish, due to (4)), up to a time change. The arguments of Section 2 then show that Theorem 1 holds for this new diffusion X, without Assumption (2).

The Morse assumption on U was considered for simplicity, but it can be relaxed. For instance in the cases where $\mathcal{U}$ consists of finite number of connected and disjoint submanifolds, say $\mathcal{U}_1, \mathcal{U}_2, ..., \mathcal{U}_N$, with non-degenerate Hessians of U in the orthogonal directions. Assume furthermore that for any $n \in [\![N]\!]$, we can find orthogonal vector fields on $\mathcal{U}_n$ which are eigenvectors of the orthogonal Hessians. Then approximations such as (10), (11) and (12) are still valid in the corresponding exponential systems of coordinates, where m has to be replaced by the co-dimension m' of $\mathcal{U}_n$. More precisely, we can extend them into

$$\begin{aligned}
(1-\epsilon) \sum_{i\in[\![m']\!]} \lambda_i(y) \ \leqslant\ \Delta U(x) \ \leqslant\ (1+\epsilon) \sum_{i\in[\![m']\!]} \lambda_i(y) \\
(1-\epsilon) \sum_{i\in[\![m']\!]} \lambda_i^2(y) z_i^2 \ \leqslant\ \|\nabla U\|^2(x) \ \leqslant\ (1+\epsilon) \sum_{i\in[\![m']\!]} \lambda_i^2(y) z_i^2 \\
\frac{1}{2}(1-\epsilon) \sum_{i\in[\![m']\!]} \lambda_i(y) z_i^2 \ \leqslant\ U(x) \ \leqslant\ \frac{1}{2}(1+\epsilon) \sum_{i\in[\![m']\!]} \lambda_i(y) z_i^2
\end{aligned}$$

in a sufficiently small tubular neighborhood of $\mathcal{U}_n$, with x charted by (y, z), y standing for the Riemannian projection on $\mathcal{U}_n$ and $z := (z_i)_{i\in[\![m']\!]}$ for the coordinates in the exponential systems deduced from the eigenvectors (note that the gradient and the Laplacian of U in the directions of y are negligible).

More generally, when the eigenvectors cannot be defined globally on $\mathcal{U}_n$, one has to work in a finite number of charts, which increases the technicality of the proofs without changing the convergence result, when β is large enough in terms of the orthogonal eigenvalues of the Hessian (or when a generator of the form (6) is considered).

A Non-fraudulent algorithms

Two examples are given below showing how fraudulent algorithms can be related to the investigation of non-fraudulent algorithms.

We start with the time-inhomogeneous swarm mean-field algorithm introduced in [1], which was the initial motivation for this work. On a compact Riemannian manifold, consider the non-linear evolution equation

$$\frac{d}{dt}\rho_t \quad = \quad \mathrm{div}(\rho_t[\gamma_t \nabla U + \nabla \varphi'(\rho_t)]) \tag{21}$$

where

- ρ_t is the density with respect to the Riemannian probability ℓ of another probability on M,
- $(\gamma_t)_{t\geqslant 0}$ is an inverse temperature scheme, assumed to be smooth and to increase to $+\infty$ in large times,
- $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly convex function satisfying $\varphi(1) = 0$ and $\varphi'(0) = -\infty$ and is $\mathcal{C}^2$ on $(0, +\infty)$.

A non-linear diffusion $Y \coloneqq (Y(t))_{t\geqslant 0}$ is associated to this equation, whose evolution can be heuristically described as in (1) via

$$dY(t) \quad = \quad -\gamma_t \nabla U(Y(t)) + \sqrt{2\alpha(\rho_t(Y(t)))}\, dB(t)$$

where

- ρ_t is the density with respect to the Riemannian probability ℓ of the law of $Y(t)$,
- the function $\alpha : (0, +\infty) \to \mathbb{R}_+$ is given by

$$\forall\ r > 0, \qquad \alpha(r) \quad \coloneqq \quad \frac{1}{r}\int_0^r s\varphi''(s)\, ds$$

- $(B(t))_{t\geqslant 0}$ is a m-dimensional Brownian motion.

In [1] we considered the convex function defined, for given $m \in (0, 1/2)$, by

$$\forall\ r \geqslant 0, \qquad \varphi \quad \coloneqq \quad \begin{cases} \varphi_m(r) & \text{if } r \in (0, 1] \\ \varphi_2(r) & \text{if } r \in (1, +\infty) \end{cases}$$

with for any $m \in (0, +\infty)\backslash\{1\}$,

$$\forall\ r \geqslant 0, \qquad \varphi_m(r) \quad \coloneqq \quad \frac{r^m - 1 - m(r-1)}{m(m-1)}$$

It was shown through new functional inequalities that at least in dimension 1, we can find power schemes $(\gamma_t)_{t\geqslant 0}$, such that for any neigborhood $\mathcal{N}$ of $\mathcal{U}$, we have

$$\lim_{t\to+\infty} \rho_t(\mathcal{N}) \quad = \quad 1$$

namely that the above evolutions provide a theoretical global optimization procedure for U. In practice, the evolution $(\rho_t)_{t\geqslant 0}$ should be seen as a mean-field limit and approximated by swarm particle systems (each particle "counting" the number of particles around it to get an estimate of the local density and boosting the intensity of its own Brownian motion accordingly if they are too few or too many).

In the final discussion section of [1], an heuristic comparison was made with usual simulated annealing and it appeared that in large times and up to a time change, Y behaves like the diffusion X described by (1) with $\beta = m/(1-m)$.

Condition (2) may explain the restriction to dimension 1 considered in [1] and suggests the means to go beyond it, but the relation between the swarm and fraudulent algorithms should be first investigated more closely.

Another illustrative example consists of the following simpler algorithm, which is probably less efficient than the swarm particle algorithm due to the lack of interactions between the approximating particles, called $Z_1, Z_2, ..., Z_q$ below. The underlying idea is to estimate the quantity $U(x) - \min_M U$ (needed in a fraudulent setting) through

$$\forall\ x \in M, \qquad U(x) - \min_M U \quad = \quad -\lim_{\gamma \to +\infty} \frac{1}{\gamma} \ln(\mu_\gamma(x)) \tag{22}$$

where μ_γ is the density of the Gibbs measure associated to the potential U and to the inverse temperature $\gamma \geqslant 0$:

$$\forall\ x \in M, \qquad \mu_\gamma(x) \quad \coloneqq \quad \frac{\exp(-\gamma U(x))}{\int_M \exp(-\gamma U(y))\, \ell(dy)}$$

where we recall that ℓ is the Riemannian probability.

Other ways of approximating the l.h.s. of (22) can be devised, thus in [1] the main role is played by the stationary measure of (21) for a fixed parameter γ_t, which is next sent to infinity. The convergence (22) is more classical and leads to the following natural procedure. Nevertheless other alternative approaches would deserve to be investigated.

The Gibbs measure $\mu_\gamma\, d\ell$ is the invariant (and even reversible) probability measure associated to the Markov generator $\Delta \cdot -\gamma \langle \nabla U, \nabla \cdot \rangle$, namely to the diffusion $Y \coloneqq (Y(t))_{t \geqslant 0}$ heuristically described as in (1) through

$$\forall\ t \geqslant 0, \qquad dY(t) \quad = \quad -\gamma \nabla U(Y(t))\, dt + \sqrt{2}\, dW(t)$$

where $W \coloneqq (W(t))_{t \geqslant 0}$ is a m-dimensional Brownian motion.

It is well-known that the law of $Y(t)$ converges to μ_γ as t goes to infinity. To get γ going to infinity, we consider a time-inhomogeneous version $Z \coloneqq (Z(t))_{t \geqslant 0}$ of Y associated to a scheme $\gamma : \mathbb{R}_+ \ni t \mapsto \gamma_t$ via

$$\forall\ t \geqslant 0, \qquad dZ(t) \quad = \quad -\gamma_t \nabla U(Z(t))\, dt + \sqrt{2}\, dW(t)$$

From the theory of simulated annealing, see Holley, Kusuoka and Stroock [4], it is known that for large times $t \geqslant 0$, the law $\mathcal{L}(Z(t))$ of $Z(t)$ becomes closer and closer to $\mu_{\gamma_t} d\ell$ if the inverse temperature scheme γ has a sufficiently slow logarithmic growth. More precisely, for a scheme of the form

$$\forall\ t \geqslant 0, \qquad \gamma_t \quad = \quad k^{-1} \ln(1 + t) \tag{23}$$

the relative entropy of $\mathcal{L}(Z(t))$ with respect to μ_{γ_t} goes to zero for large time $t \geqslant 0$ when $k > c$, where $c \geqslant 0$ the largest height of a well not containing a given element of $\mathcal{U}$ (the constant c does not depend on this fixed element).

To get an approximation of the density $\mu_{\gamma_t}(x)$, we first consider a finite sequence $Z_1, ..., Z_Q$, $Q \in \mathbb{N}$ of independent copies of Z, namely satisfying

$$\forall\ q \in [\![Q]\!],\ \forall\ t \geqslant 0, \qquad dZ_q(t) \quad = \quad -\gamma_t \nabla U(Z_q(t))\, dt + \sqrt{2}\, dW_q(t)$$

where $W_q \coloneqq (W_q(t))_{t \geqslant 0}$ are independent m-dimensional Brownian motions for $q \in [\![Q]\!]$. For Q large enough, by the law of the large numbers, we can expect that for fixed $t \geqslant 0$, the empirical measure of the $Z_q(t)$, $q \in [\![Q]\!]$ is an approximation of the law of $Z(t)$ and thus is close to μ_{γ_t} if the time t has been chosen large enough and the inverse temperature schedule according to (23) with $k > c$.

Next let be given a kernel approximation of the Dirac masses: it is a family $(K_h(\cdot,\cdot))_{h>0}$ of smooth mappings on M^2 such that for any $x \in M$, $K_h(x,y)\ell(dy)$ is a probability measure weakly converging toward δ_x as h goes to 0_+ (a geometrical example is the heat kernel on M at small times).

Through classical density approximation, see e.g. the book of Silverman [10], we hope that for $t \geqslant 0$ and Q large enough and $h > 0$ small enough,

$$\frac{1}{Q} \sum_{q \in [\![Q]\!]} K_h(Z_q(t), x)$$

is a good approximation of $\mu_{\gamma_t}(x)$ for any $x \in M$. In view of (22), we are led to consider a process $X \coloneqq (X(t))_{t \geqslant 0}$ defined by

$$dX(t) \;=\; -\beta \nabla U(X(t))\, dt + \sqrt{2V(\gamma_t, X(t), Z_{[\![Q_t]\!]}(t))}\, dB(t)$$

where β satisfies (2) and the Brownian motion B is independent of the Brownian motions W_q, for $q \in \mathbb{N}$, where

$$V(\gamma_t, X(t), Z_{[\![Q_t]\!]}(t)) \;\coloneqq\; \frac{\left|\ln\left(\sum_{q \in [\![Q_t]\!]} K_{h_t}(Z_q(t), X(t))/Q\right)\right|}{\gamma_t}$$

and where $Q : \mathbb{R}_+ \ni t \mapsto Q_t \in \mathbb{N}$ and $h : \mathbb{R}_+ \ni t \mapsto h_t$ are respectively non-decreasing and increasing evolutions, such that

$$\lim_{t \to +\infty} Q_t \;=\; +\infty \quad \text{and} \quad \lim_{t \to +\infty} h_t \;=\; 0$$

We can assume that the potential upward jumps $Q_t - Q_{t-}$ are either 0 or of size 1 and that furthermore when a jump of Q does occur at time $t \geqslant 0$, then a particle is chosen among the current Q_{t-} ones and gives rise to a new particle at the same place. Afterward the two particles at this same place evolve independently.

More precisely, in addition to (23) with $k > c$, we are looking for appropriate schedules Q and h such that X turns out to be a global minimizer of U, in the sense that $X(t)$ converges a.s. for large times $t \geqslant 0$ toward an element of $\mathcal{U}$, which is expected to be random if $\mathcal{U}$ is not a singleton, according to Theorem 1.

References

[1] Jérôme Bolte, Laurent Miclo, and Stéphane Villeneuve. Swarm gradient dynamics for global optimization: the density case. ArXiv preprint 2204.01306, 2022.

[2] Bennett Chow, Peng Lu, and Lei Ni. *Hamilton's Ricci flow*, volume 77 of *Grad. Stud. Math.* Providence, RI: American Mathematical Society (AMS), 2006.

[3] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986. Characterization and convergence.

[4] Richard A. Holley, Shigeo Kusuoka, and Daniel W. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2):333–347, 1989.

[5] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library.* North-Holland Publishing Co., Amsterdam, second edition, 1989.

[6] Jean-François Le Gall. Bessel processes, the Brownian snake and super-Brownian motion. In *Séminaire de Probabilités XLVII*, pages 89–105. Cham: Springer, 2015.

[7] 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.

[8] 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.

[9] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, third edition, 1999.

[10] Bernard W. Silverman. *Density estimation for statistics and data analysis*. CRC Press, Boca Raton, FL, 1986.

[11] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional diffusion processes*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1997 edition.

[12] Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part ii: Continuous time analysis, 2021.

[13] 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.

miclo@math.cnrs.fr

Toulouse School of Economics,
1, Esplanade de l'université
31080 Toulouse cedex 06, France

Institut de Mathématiques de Toulouse
Université Paul Sabatier, 118, route de Narbonne
31062 Toulouse cedex 9, France