Swarm gradient dynamics for global optimization: the mean-field limit case

Jérôme Bolte[*], Laurent Miclo[*][†], and Stéphane Villeneuve[*][‡]

November 28, 2023

Abstract

Using jointly geometric and stochastic reformulations of nonconvex problems and exploiting a Monge-Kantorovich (or Wasserstein) gradient system formulation with vanishing forces, we formally extend the simulated annealing method to a wide range of global optimization methods. Due to the built-in combination of a gradient-like strategy and particle interactions, we call them swarm gradient dynamics. As in the original paper by Holley-Kusuoka-Stroock, a functional inequality is the key to the existence of a schedule that ensures convergence to a global minimizer. One of our central theoretical contributions is proving such an inequality for one-dimensional compact manifolds. We conjecture that the inequality holds true in a much broader setting. Additionally, we describe a general method for global optimization that highlights the essential role of functional inequalities 'a la Łojasiewicz.

Contents

[*] Toulouse School of Economics, University of Toulouse Capitole, 1 esplanade de l'université, 31000 Toulouse, France, the authors acknowledge funding from ANR under grant ANR-17-EUR-0010 (Investissements d'Avenir program). Email: jerome.bolte@tse-fr.eu

[†] CNRS-IMT-TSE-R. Email: laurent.miclo@math.cnrs.fr

[‡] Corresponding Author, TSE-TSM-R. E-mail: stephane.villeneuve@tse-fr.eu

1 Introduction

The global minimization of a nonconvex function is one of the most challenging problems in modern optimization. There are few global optimization methods that provide reasonable convergence guarantees, the most famous is probably the simulated annealing, whose premises are found in [36], or the moment method [35], and their many variants[1]. On the other hand, metaheuristics methods are numerous and have some notable empirical success: they orchestrate interactions between local and global strategies, combine random and deterministic procedures, and often end up with methods using optimizing agents. Some examples of metaheuristics are inspired by analogies with biology, as evolutionary algorithms [25], ethology (e.g., ant colonies [20]), or particle swarms, see e.g., [34]. The goal of this paper is to introduce a new family of swarm methods through gradient descent in the Monge-Kantorovich (or Wasserstein) space and give general guarantees for their convergence to global minimizers.

Let us be more specific and consider the problem of solving

$$\min_M U, \tag{$\mathcal{P}$}$$

where $U : M \to \mathbb{R}$ is a differentiable function defined on a compact Riemannian manifold M. In order to introduce our swarm methods, we need first some considerations on simulated annealing.

Three views on simulated annealing Our starting point is indeed the famous simulated annealing method. In its time-continuous form, and when the state space is flat M (e.g., a flat torus), it is a solution $X \coloneqq (X_t)_{t\geq 0}$ of the time-inhomogeneous Langevin-like stochastic differential equation,

$$dX_t = -\beta_t \nabla U(X_t)\, dt + \sqrt{2} dB_t, \tag{a}$$

where $(B_t)_{t\geq 0}$ is a Brownian motion and $\beta_t \to +\infty$ [2] is a time-dependent parameter tuned so that the expectation of $U(X_t)$ tends to $\min_M U$. The formulation (a) can be extended to any compact Riemannian manifold M, but this requires more involved notations, see e.g., Ikeda and Watanabe [31] or Emery [23].

The intuitive interpretation is quite natural, the method combines local gradient search with a vanishing Brownian exploration of the feasible set M. Although the method is often used as a heuristic, its proof has been made rigorous in various frameworks via different approaches, the two main ones being based on large deviations, see e.g., Azencott et al. [2], and on functional inequalities, cf. Holley, Kusuoka and Stroock [29]. Key to the foundational approach of [29], is the establishment of a generalized log-Sobolev inequality followed by hypercontractivity arguments. This approach was then simplified by Miclo [39] via the

[1] In the case of simulated annealing and Langevin like dynamics, see also [29, 39, 27] and modern extensions [42, 46, 26].

[2] Often called the inverse temperature

identification of the relative entropy as a convenient Lyapunov function. In order to explain the role of the log-Sobolev inequality, the relative entropy, the mechanisms behind the convergence properties, and understand the scope of the method, we view simulated annealing along three complementary angles:

(a) the SDE form of the algorithm: the overdamped Langevin dynamics (a) above,

(b) the PDE counterpart of (a), which describes the time evolution of the density $t \mapsto \rho(t)$ of X_t. It assumes the form of a Fokker-Planck equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = \beta_t \operatorname{div}(\rho \nabla U) + \Delta \rho, \quad t \geq 0. \tag{b}$$

(c) Otto's formalism [33], which we recall briefly in the Appendix, allows one to interpret the latter as a gradient-like system in the space of probabilities on M endowed with Monge-Kantorovich metric:

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\rho(t) = -\mathrm{grad}_{\mathcal{W}}\,\mathcal{U}_{\beta_t}[\rho(t)], \tag{c}$$

with $\mathcal{U}_\beta[\rho] = \beta \int_M U\rho\, d\ell + \int_M \rho \log \rho\, d\ell$, where ℓ is the Riemannian measure of M. This quantity is also known as the relative entropy of ρ with respect to the Gibbs measure whose density is proportional to $\exp(-\beta U)$.

This triple perspective, mainly due to [33], is not new and has known a recent success in sampling [16, 21], optimization [38, 37] and machine learning [42, 15].

Depending on the form we adopt to study the dynamics, subsequent results or developments may be considerably easier to understand. Indeed, while (a) classically provides operational algorithms through discretization, (b) offers a tractable version amenable to classical PDE analysis methods as Lyapunov methods. As for the last angle, (c), it confers a sharp geometrical content to the method and allows to interpret the essential tools of convergence through classical intuitive geometric ideas. An essential fact about (c) is that functional inequalities, as the log-Sobolev inequality, may be seen as Łojasiewicz gradient inequalities. In our case it means that there exists an exponent $\gamma \in (0, 1)$ such that the slope of $(\mathcal{U}_\beta - \min \mathcal{U}_\beta)^\gamma$ is bounded away from zero (save at the stationary measure). This reparametrization sharpens the energy landscape while leaving unchanged level sets: this allows for a direct convergence analysis of the gradient method (c), see [6] and references therein. In the simulated annealing case the log-Sobolev inequality of Holley, Kusuoka and Stroock [29] turns out to be an instance of such an inequality, see [6].

Swarm gradient dynamics The triple-perspective (a)-(b)-(c) we used to describe the strategy of simulated annealing can be generalized to a much larger framework. For this, we adopt the angle (c) under which we observe that it is natural to consider more general convex functions φ than $\mathbb{R}_+ \ni r \mapsto r\ln(r)$ in the Boltzmann entropy,

$$\mathcal{B}[\rho] = \int_M \rho \log \rho\, d\ell.$$

Referring to the results in [1], we may indeed use a whole family of convex functionals

$$\mathcal{H}[\rho] = \int_M \varphi(\rho)\, d\ell,$$

leading to a penalized cost

$$\mathcal{U}_\beta[\rho] = \beta \int_M U\rho\, d\ell + \mathcal{H}[\rho]$$

and to the triplet of "equivalent" minimizing dynamics modeled on (c), (b), (a),

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -\mathrm{grad}_{\mathcal{W}}\,\mathcal{U}_{\beta_t}[\rho(t)] \tag{1}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = \beta_t \mathrm{div}(\rho\nabla U) + \mathrm{div}\rho\nabla\varphi'(\rho) \tag{2}$$

$$dX_t = -\beta_t \nabla U(X_t)\, dt + \sqrt{2}\alpha(\rho) dB_t, \tag{3}$$

with $\lim_{t\to\infty}\beta_t = +\infty$, φ, α are some positive functions to be specified, and ρ_t is the probability density of X_t for each t. The fact that a particle interacts with its law may be considered as a swarm effect, this is why we call these dynamics *swarm gradient dynamics*[3], see Section 4.2 for more insight. Principles and other considerations behind the above dynamics are described in Sections 2.4, 4.2, and in the Appendix.

The key to convergence: functional inequalities The central question is that of the convergence properties to a global minimizer. In particular, an essential question is: what are the assumptions ensuring that the global minimization $\min_M U$ problem is solved by the above?

In simulated annealing, the essential tool for convergence is the log-Sobolev inequality of Holley, Kusuoka and Stroock [29]. We also recalled that this inequality can be advantageously thought as a Łojasiewicz inequality when considering the problem along the gradient system angle (c). We follow, therefore, the same protocol but in a reverse way: we formally write Łojasiewicz inequalities using the Monge-Kantorovich formalism, which reveals, in turn, the functional inequalities we would like to have at our disposal. This leads us to consider:

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu_\beta)|^2 \rho\, d\ell \geq c(\beta)\,\Omega\left(\int_M \varphi(\rho) - \varphi(\mu_\beta) - \varphi'(\mu)(\rho - \mu_\beta)\, d\ell\right) \tag{4}$$

where $c, \Omega : (0, +\infty) \to \mathbb{R}_+$ are positive functions having specific properties and where μ_β is the unique stationary measure of $\mathcal{U}_\beta$. We are at the heart of this paper and our central result: proving such a functional inequality under adequate assumptions. Our result holds in compact one-dimensional manifolds for power-like potential functions φ. As a consequence, we obtain a full convergence result of our global methods on compact one-dimensional manifolds. We also evidence the general mechanisms of global convergence, and for completeness, we sketch the form that operational algorithms could take (but we postpone their study to another paper).

Apart from the interest of our work for optimization, we believe that it raises important questions and hopes on the validity domain of the family of inequalities in (4). Positive outcomes would lead to new results in optimization and in other fields.

[3] Since β is variable, they are actually time-dependent swarm dynamics.

Related works The quantitative comparison between entropy-type functional and its time derivative, often called the entropy production or dissipation, dates back to [39], where it was exploited through the logarithmic Sobolev inequality of [29]. Using Otto's formalism, this can be, in turn, reinterpreted as an approach à la Łojasiewicz [6].

Equation (2) may be seen as a formal generalization of porous media equation and fast diffusion equations – which corresponds to the case when φ is a potential. These have been studied by several authors using the Monge-Kantorovich framework, see, e.g., Otto [41], and Carrillo, McCann and Villani [13, 14]. Their asymptotic analysis is through the Bakry-Emery method [3]: it consists in the second-order time differentiation of the entropy. Contrary to [29, 39] and our current approach, this approach requires convexity, which makes it unsuitable for general global minimization.

The article of Iacobelli, Patacchini, and Santambrogio [30] is also connected to our approach since they consider ultrafast diffusion equations, which corresponds to a negative exponent m in (7). However, the hypocoercive bounds they obtain do not seem well-suited to extensions to time-inhomogeneous situations since they do not lead to a differential inequality satisfied by the entropy-like functionals.

The uniqueness of the stationary measure, i.e., of the minimizer of $\mathcal{U}_\beta$, is not a new result, it can be found in, e.g., Carrillo, Jüngel, Markowich, Toscani and Unterreiter [12]. For the sake of completeness, we provide a proof in the next section.

Let us conclude by mentioning a few works using nonlinear diffusion where nonlinearities generally affect the drift coefficient but not the diffusion term as here. In Eberle, Guillin, and Zimmer [22], the authors use coupling techniques, Carillo, Gvalani, Pavliotis and Schlichting [11] treat the case of interaction potentials while Delarue and Tse [18] consider chaos propagation.

2 Presentation of the problem

2.1 A family of relaxations in the probability space $\mathcal{P}(M)$

Consider the nonconvex minimization problem:

$$\text{Find a global minimizer of } U : M \to \mathbb{R} \text{ on a compact Riemannian manifold } M. \qquad (\mathcal{P})$$

Denote by d the distance on M, and ℓ the natural Riemannian measure. Up to a normalization factor, assume $\ell(M) = 1$. Let $\mathcal{P}(M)$ be the space of probability measures on M equipped with the Monge-Kantorovich distance defined through

$$\mathcal{W}_2^2(\mu,\nu) = \inf\left\{ \int_M d^2(x,y)\, p(dx,dy) : p \text{ is a coupling of } \mu \text{ and } \nu \text{ on } M^2 \right\},$$

for any μ, ν in $\mathcal{P}(M)$. The extreme values of U play a special role in our approach, one defines

$$\operatorname{osc}(U) := \max_M U - \min_M U, \tag{5}$$

which we may assume positive –since otherwise, the problem would be trivial.

We make the following regularity assumptions:

Assumption (A). The manifold M and the function $U : M \to [0, +\infty[$ are of class C^2.

The C^2 regularity assumptions are simple means to obtain existence results for the gradient evolution in the Monge-Kantorovich space as in [1, 24]

We embed our initial problem $\min_M U$ in $\mathcal{P}(M)$ and consider the relaxed minimization problem

$$\mathcal{U}[\rho] := \int_M U\rho.$$

Using the monotonicity of the integral, one easily sees that

$$\min_{\mathcal{P}(M)} \mathcal{U} = \min_M U.$$

Let $\beta > 0$, we introduce a *penalized relaxation* of U in the metric space $(\mathcal{P}(M), \mathcal{W}_2)$ through

$$\mathcal{U}_\beta[\rho] = \beta\mathcal{U}[\rho] + \mathcal{H}[\rho] = \beta \int_M U\rho\, d\ell + \mathcal{H}[\rho] \tag{6}$$

where

$$\mathcal{H}[\rho] = \begin{cases} \displaystyle\int_M \varphi(\rho)\, d\ell & \text{if } \rho \text{ is absolutely continuous w.r.t the Riemannian measure} \\ +\infty & \text{otherwise,} \end{cases}$$

with $\varphi : [0, +\infty) \to \mathbb{R}_+$ is strictly convex and C^2 on $(0, +\infty)$. Up to a multiplicative factor, the first term in (6) is the classical relaxation of U within the probability space over M. On the other hand, as in simulated annealing, the second term acts as a penalization forcing the minimizer to be unique and to have a density with respect to ℓ (see Lemma 1). As we shall see below, letting β go to infinity allows one to recover the initial problem.

Remark 1. (a) (Power-like penalizations) A strong focus will be put on the class of power-like functions. For any $m \in (0, +\infty) \setminus \{1\}$, define the convex function $\varphi_m : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$\forall\, r \geq 0, \qquad \varphi_m(r) \;\; := \;\; \frac{r^m - 1 - m(r-1)}{m(m-1)} \tag{7}$$

Let us observe that φ_m is a strictly convex function, C^2 on $(0, +\infty)$ and such that $\varphi_m(1) = 0$, $\varphi_m'(1) = 0$ and $\varphi_m''(1) = 1$ for every admissible m. By the Taylor-Lagrange formula, we deduce that φ_m is always positive, except at 1. The convex function $\varphi_1 : \mathbb{R}_+ \to \mathbb{R}_+$ defined by $\varphi_1(r) := r\ln(r) - (r-1)$ is recovered as the limit of φ_m when m goes to 1 and corresponds to the Boltzmann entropy. Hereafter, we will in particular consider functions φ that are constructed by gluing together two different functions φ_m at 1.

(b) (Regularity of the penalization) Observe as well, from [1, Theorem 9.3.9, p.212] and [1, Proposition 9.3.2, p.210], that the function $\mathcal{H}$ is lower for the Monge-Kantorovich distance, and geodesically convex in $\mathcal{P}(M)$.

The approach we adopt is through the one-parameter family of problems

$$\mathrm{val}(\mathcal{P}_\beta) := \inf_{\mathcal{P}(M)} \mathcal{U}_\beta \tag{$\mathcal{P}_\beta$}$$

where the parameter $\beta > 0$ is the inverse of a penalization parameter or the inverse of "the temperature" according to the simulated annealing literature. It ultimately tends to ∞, and one has an elementary but important fact:

Proposition 1 (A global optimization principle)**.** *Assume* (A) *and that* $\varphi : [0, +\infty) \to \mathbb{R}_+$ *is strictly convex and* C^2 *on* $(0, +\infty)$*. Then*

(i) $\displaystyle\lim_{\beta \to +\infty} \frac{1}{\beta} \operatorname{val}(\mathcal{P}_\beta) \;=\; \min_M U,$

(ii) if μ_β *is a sequence of solutions to* $(\mathcal{P}_\beta)$*, the weak* limit points of* μ_β *have a support concentrated on the set of minimizers of* U

Proof. As a first observation, it is clear that

$$\min_{\mathcal{P}(M)} \int_M U\rho = \min_M U.$$

Fix $\epsilon > 0$ and choose ρ_ϵ to be an ϵ-minimizer of $\mathcal{U}[\rho] = \int_M U\rho$. The above observation yields $\mathcal{U}[\rho_\epsilon] \leq \min_M U + \epsilon$. Since $\operatorname{dom} \mathcal{H}$ contains smooth densities, one can also assume that $\mathcal{H}(\rho_\epsilon)$ is finite. Take $\beta > 0$ and let $\mu_{\beta,\epsilon}$ be an ϵ-solution to $(\mathcal{P}_\beta)$, that is

$$\frac{1}{\beta}\mathcal{U}_\beta[\mu_{\beta,\epsilon}] = \mathcal{U}[\mu_{\beta,\epsilon}] + 1/\beta\mathcal{H}[\mu_{\beta,\epsilon}] \leq \mathcal{U}(\rho) + 1/\beta\mathcal{H}[\rho] + \epsilon$$

for all ρ. Thus choosing $\rho = \rho_\epsilon$ yields

$$\frac{1}{\beta}\mathcal{U}_\beta[\mu_{\beta,\epsilon}] \leq \min_M U + 1/\beta\mathcal{H}[\rho_\epsilon] + 2\epsilon.$$

Letting β goes to infinity yields

$$\limsup_{\beta \to \infty} \frac{1}{\beta}\mathcal{U}_\beta[\mu_{\beta,\epsilon}] \leq \min_M U + 2\epsilon.$$

Whence $\displaystyle\limsup_{\beta \to \infty, \epsilon \to 0} \frac{1}{\beta}\mathcal{U}_\beta[\mu_{\beta,\epsilon}] \leq \min_M U$. Since $\mathcal{U}_\beta \geq \min_M U$ by positivity of φ, (i) follows readily.
Let us prove (ii). Let ρ be a limit point of μ_β for the weak* topology. Since U is continuous, its mixed extension $\mathcal{U}$ is continuous for the weak* topology and thus $\lim_{\beta\to\infty} \mathcal{U}[\mu_\beta] = \mathcal{U}[\rho]$. On the other hand, by positivity of φ, one has $\mathcal{U} \leq \frac{1}{\beta}\mathcal{U}_\beta$, thus (i) gives $\limsup_{\beta\to\infty} \mathcal{U}[\mu_\beta] \leq \min_M \mathcal{U}$ whence $\mathcal{U}[\rho] \leq \min_M U$ and (ii) follows. □

Observe that

$$\operatorname{gap}(\beta) = \frac{1}{\beta} \inf_{\mathcal{P}(M)} \mathcal{U}_\beta - \min_M U \tag{8}$$

tends to zero when β tends to $+\infty$ by the previous result. The quantity $\operatorname{gap}(\beta)$ measures the global approximation abilities of the problem $(\mathcal{P}_\beta)$ with respect to the initial problem $\min_M U$.

2.2 Variational considerations and stationary measures

Let us denote by $\langle \cdot, \cdot \rangle_x$ the Riemannian metric on the tangent at x to M and $|\cdot|_x$ the corresponding norm. We generally omit the dependence in x.

Let us analyze the first-order conditions for the above problem $(\mathcal{P}_\beta)$ through the lenses of the Monge-Kantorovich metric. We shall freely use the definition of Monge-Kantorovich subgradients and related objects. As they are only central to our understanding but not to our proofs, we refer to [1] for details and to the Appendix for some elementary considerations. The subgradient of $\mathcal{U}_\beta$ with respect to the Monge-Kantorovich metric has a domain contained in $L^1(M)$ and is formally given by

$$\operatorname{grad}_{\mathcal{W}} \mathcal{U}_\beta\,[\rho] \quad = \quad -\operatorname{div}(\rho(\beta\nabla U + \nabla\varphi'(\rho)))$$

for any admissible ρ in $L^1(M)$. Stationary solutions of (2), with $\beta_t \equiv \beta$, are thus probability densities μ solution to

$$\operatorname{div}(\mu(\beta\nabla U + \nabla\varphi'(\mu))) \quad = \quad 0, \tag{9}$$

which is to be understood in the standard weak sense. By integration by parts, we have for $f \in C^2(M)$,

$$\begin{aligned}\int_M \operatorname{div}(\mu\nabla\varphi'(\mu))f\,d\ell \quad &= \quad -\int_M \left\langle \mu\nabla\varphi'(\mu), \nabla f\right\rangle\,d\ell \\ &= \quad -\int_M \mu\varphi''(\mu)\left\langle \nabla\mu, \nabla f\right\rangle\,d\ell \\ &= \quad -\int_M \left\langle \nabla(\mu\alpha(\mu)), \nabla f\right\rangle\,d\ell \\ &= \quad \int_M \mu\alpha(\mu)\triangle f\,d\ell \\ &= \quad \int_{\{\mu>0\}} \alpha(\mu)\triangle f\,\mu d\ell\end{aligned}$$

where the function $\alpha\,:\,(0,+\infty) \to \mathbb{R}_+$ is given by

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

On the other hand, we have

$$\int_M \operatorname{div}(\mu\beta\nabla U)f\,d\ell = -\int_M \left\langle \beta\nabla U, \nabla f\right\rangle\mu\,d\ell.$$

Finally, μ is a stationary solution to (9) if

$$\forall\ f \in \mathcal{C}^2(M), \qquad \int_{\{\mu>0\}} L_\mu[f]\,\mu d\ell \quad = \quad 0$$

with

$$L_\mu[f] \quad = \quad \alpha(\mu)\triangle f - \left\langle \beta\nabla U, \nabla f\right\rangle.$$

Remark 2 (Infinitesimal generator)**.** Observe that the choice $\varphi = \varphi_1$ leads to $\alpha(r) = 1$, that is to

$$L_\mu[f] = \triangle f - \left\langle \beta\nabla U, \nabla f\right\rangle.$$

This is the infinitesimal generator of the classical (overdamped) Langevin SDE. Up to a multiplicative constant, this is the only choice for the operator L_μ to be independent of μ.

Let us apply formally the above relationship to the function $f = \beta U + \varphi'(\mu)$, assuming here that the stationary density μ is smooth enough. We obtain[4]

$$\begin{aligned}\int_{\{\mu>0\}} |\nabla(\beta U + \varphi'(\mu))|^2 \mu \, d\ell &= - \int_{\{\mu>0\}} L_\mu[\beta U + \varphi'(\mu)] \, \mu d\ell \\ &= 0,\end{aligned} \tag{10}$$

where the last equality comes from the stationarity of μ. Therefore, $\beta U + \varphi'(\mu)$ is constant on every connected component of the set $\{\mu > 0\}$. We analyze this condition in the next paragraph.

2.3 Uniqueness of the stationary density

The main ingredient to study the uniqueness of the stationary density is the relation

$$\beta U + \varphi'(\mu) \quad = \quad c, \tag{11}$$

where the constant c depends on the considered connected component of the support $\{\mu > 0\}$.

Define $I \coloneqq \varphi'((0, +\infty))$ and denote by $\psi : I \to (0, +\infty)$ the inverse of φ'.

Lemma 1 (Existence and uniqueness of the minimizer of $(\mathcal{P}_\beta)$)**.** *Assume $\varphi'(0) = -\infty$, then there exists an unique stationary density μ_β solution to* (9)*. Moreover,*

(i) μ_β is positive everywhere on M and is characterized by the relation

$$\mu_\beta \quad = \quad \psi(c^* - \beta U), \tag{12}$$

where c^* is a normalization parameter characterized by the condition

$$\int_M \psi(c^* - \beta U) \, d\ell = 1.$$

(ii) μ_β is the global minimizer of $\mathcal{U}_\beta$.

Proof. Let us start with (i) and by showing that a stationary density μ is everywhere positive. Towards a contradiction, assume that the set $\{\mu = 0\}$ is nonempty and let M_1 be a connected component of the open set $\{\mu > 0\}$ with $\partial M_1 \neq \emptyset$. Let $(x_n)_{n \in \mathbb{N}}$ be a sequence of elements of M_1 converging to some point in the boundary ∂M_1. According to (11), we have for every $n \in \mathbb{N}$

$$\beta U(x_n) + \varphi'(\mu(x_n)) \quad = \quad c_{M_1}$$

where the left-hand side term converges to $-\infty$ which is absurd. Therefore, $M_1 = M$ and equation (11) is valid everywhere on M. Set $c \coloneqq c_{M_1}$. Being strictly convex, the function φ' is one-to-one and onto between $(0, +\infty)$, and by definition of ψ, equation (11) rewrites:

$$\mu \quad = \quad \psi(c^* - \beta U) \tag{13}$$

[4] One may observe that the first term in (10) is the squared norm of the Monge-Kantorovich gradient of $\mathcal{U}_\beta$ evaluated at μ.

As μ is a density function, we must have

$$\int_M \psi(c^* - \beta U)\, d\ell \quad = \quad 1. \tag{14}$$

Since ψ is strictly increasing and satisfies $\lim_{\inf I} \psi = \lim_{-\infty} \psi = 0$ and $\lim_{\sup I} \psi = +\infty$, there is a unique value $c^* \in \mathbb{R}$ that satisfies equation (14). We have

$$\mu = \psi(c^* - \beta U)$$

which ends not only the proof of the uniqueness of the stationary density μ but also gives its existence and its explicit form. To see (ii) and that μ is the global minimizer, we observe that

$$\begin{aligned}\mathcal{U}_\beta[\rho] - \mathcal{U}_\beta[\mu] &= \int_M (\varphi(\rho) - \varphi(\mu))\, d\ell + \int_M \beta U\, (\rho - \mu)\, d\ell \\ &= \int_M (\varphi(\rho) - \varphi(\mu))\, d\ell + \int_M (c^* - \varphi'(\mu))\, (\rho - \mu)\, d\ell \\ &= \int_M \big(\varphi(\rho) - \varphi(\mu) - \varphi'(\mu)\big)\ (\rho - \mu)\, d\ell,\end{aligned}$$

which is positive whenever $\rho \neq \mu$ by strict convexity of φ. □

When no confusion can occur we simply write μ for μ_β. We gather the assumptions we need regarding φ within

Assumption (B). $\varphi : [0, +\infty) \to \mathbb{R}$ is convex, twice differentiable on $(0, +\infty)$ with $\varphi'' > 0$, and satisfies $\varphi'(0) = -\infty$.

2.4 Global minimization dynamics

Minimizing dynamics We are now in a position to provide dynamical systems meant to solve the problem $(\mathcal{P})$. Inspired by Holley, Kusuoka, and Stroock's approach to simulated annealing [29], as it was simplified in [39], and using as well the gradient view provided by Otto's formalism (see Appendix), we consider, formally, the gradient system

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = -\mathrm{grad}_{\mathcal{W}} \mathcal{U}_{\beta_t}\, [\rho(t)] \text{ a.e. on } \mathbb{R}_+, \quad \rho(0) = \rho_0, \tag{15}$$

where the term β_t is a C^1 positive time-varying parameter and where we use Newton's notation, $\frac{\mathrm{d}}{\mathrm{d}t}\rho$ here, for time derivatives. The initial distribution ρ_0 is chosen in the domain of $\mathcal{U}_\beta$, that is in the domain of $\mathcal{H}$. The time-dependent density ρ turns out to satisfy the following partial differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho \quad = \quad \mathrm{div}(\rho\, (\beta_t \nabla U)) + \mathrm{div}(\rho(\nabla \varphi'(\rho))), \quad \rho(0) = \rho_0. \tag{16}$$

The time-varying parameter β_t is traditionally interpreted as an inverse of a temperature which typically cools down, i.e.,

$$\lim_{t \to \infty} \beta_t = +\infty \tag{17}$$

Here we also interpret this parameter as the inverse of a penalty term echoing the static formula (6).

Remark 3. (a) (Simulated annealing) When $\varphi = \varphi_1$, (16) boils down to the famous simulated annealing dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho \quad = \quad \beta_t \mathrm{div}(\rho \nabla U) + \Delta\rho \tag{18}$$

which, by a famous "nonconvex" extension of the log-Sobolev inequality, due to Holley, Kusuoka and Stroock [29], is known to generate measures concentrating on the set of global minimizers of U whenever the temperature schedule is finely tuned.
(b) (Porous media) Taking U constant and $\varphi = \varphi_m$ in (16) with $m > 0$, the dynamic corresponds to the porous media equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho = m\Delta\left(\rho^m\right).$$

The case $m > 1$ refers to the slow diffusion case while the case $m < 1$, for which $\varphi'(0) = -\infty$ refers to the fast diffusion situation, (see Vazquez [45] and Otto [41]).

Existence results and evolution equations Following the pioneering work of [1], the non-autonomous theory for Monge-Kantorovich gradient flows has recently been developed in [24]. In the line of [24, Theorem 4.4, Theorem 5.4] and the existence results of [30], we *assume* that (15) and (16), have a common unique solution curve $t \mapsto \rho(t)$ in $(\mathcal{P}(M), \mathcal{W}_2)$, which satisfies in addition

$$t \mapsto \mathcal{U}_{\beta_t}[\rho(t)] \text{ and } t \mapsto \rho(t) \text{ are absolutely continuous,} \tag{19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}_{\beta_t}[\rho(t)] = \int_M (\varphi'(\rho) + \beta_t U)\frac{\mathrm{d}}{\mathrm{d}t}\rho \, d\ell + \frac{\mathrm{d}}{\mathrm{d}t}\beta_t \int_M U\rho \, d\ell, \tag{20}$$

where the time derivatives are taken for almost all times.

Functional inequalities Under hypothesis **(A), (B)** and some extra-assumptions on φ related to the geometry of the penalized cost, we intend to prove that the dynamics (15)-(16) has global optimizing properties, in the sense that the global cost

$$\mathcal{U}_{\beta_t}[\rho] = \int_M (\varphi(\rho) + \beta_t U\rho) \; d\ell \tag{21}$$

evaluated along the trajectory $t \mapsto \rho(t)$ given by (15)-(16) should converge to the value of $(\mathcal{P})$, i.e.,

$$\lim_{t \to +\infty} \mathcal{U}_{\beta_t}[\rho] = \mathrm{val}(\mathcal{P}) = \min_M U.$$

As it is customary in the analysis of PDEs the key to convergence is given by "entropy-energy" or "entropy-production" functional inequalities. In the "gradient or in the optimization world" , these can often be seen as Łojasiewicz type inequalities, see [6] and references therein. They connect the cost $\mathcal{U}_\beta$ to the norm of its gradient $\|\mathrm{grad}_{\mathcal{W}}\mathcal{U}_\beta\|$ and to the constant β:

$$\|\mathrm{grad}_{\mathcal{W}}\mathcal{U}_\beta\|^2 \geq c(\beta)\, \Omega\Big(\mathcal{U}_\beta(\rho) - \min \mathcal{U}_\beta\Big), \tag{22}$$

where $c, \Omega : (0, +\infty) \to \mathbb{R}_+$ are positive functions, with Ω being increasing and null at zero. Reexpressing $\mathcal{U}_\beta$ by means of its stationary density (12) gives

$$\begin{aligned}\mathcal{U}_\beta(\rho) - \mathcal{U}_\beta(\mu) &= \int_M \varphi(\rho) - \varphi(\mu)\, d\ell + \int_M \beta U\,(\rho - \mu)\, d\ell \\ &= \int_M \varphi(\rho) - \varphi(\mu)\, d\ell + \int_M (c^* - \varphi'(\mu))\,(\rho - \mu)\, d\ell \\ &= \int_M \left[\varphi(\rho) - \varphi(\mu) - \varphi'(\mu)\,(\rho - \mu)\right]\, d\ell.\end{aligned}$$

Because φ is convex, we obtain $\mathcal{U}_\beta(\rho) \geq \mathcal{U}_\beta(\mu)$ thus $\mathcal{U}_\beta(\mu) = \min \mathcal{U}_\beta$.
As a consequence, inequality (22) writes

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell \geq c(\beta)\, \Omega\left(\int_M \varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu)\, d\ell\right) \tag{23}$$

where $c, \Omega : (0, +\infty) \to \mathbb{R}_+$ are positive functions. A typical example is given by the log-Sobolev inequality of Holley, Kusuoka, and Stroock which can be written as

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell \geq C_{\mathrm{HKS}}(\beta)\left(\int_M \varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu)\, d\ell\right) \tag{24}$$

where $\varphi(r) = \varphi_1(r) = r\ln(r) - r + 1$ and

$$\lim_{\beta \to +\infty} \frac{1}{\beta} \ln(C_{\mathrm{HKS}}(\beta)) \quad \geq \quad -\mathrm{osc}(U)$$

(see [29] for the precise description of the l.h.s. in terms of the landscape of U).

When U is convex and φ is power-like, one can also recover Gagliardo-Nirenberg inequalities of [19], see [6] for connections with Łojasiewicz inequalities.

Convergence mechanisms for a fixed penalization parameter As previously mentioned, we adapt the approach of [39] developed in the Boltzmann entropy case ($\varphi = \varphi_1$) to our generalized class of relaxations.

Let us provide a first account of the general method through the constant parameter case. For $\rho \in \mathcal{P}(M)$ having a density with respect to ℓ, set

$$\mathcal{I}[\rho] = \mathcal{U}_\beta(\rho) - \min_{\mathcal{P}(M)} \mathcal{U}_\beta \tag{25}$$

$$\mathcal{J}[\rho] = \int_{\mathbb{T}} |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell \tag{26}$$

where the quantities may take infinite values and where μ is the unique stationary density (see Lemma 1), so that $\mathcal{I}[\mu] = 0$.

At this stage, we do not assume that $\beta > 0$ depends on time.

By time differentiation, using (19)-(20) and the evolution equation, we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\,\mathcal{I}[\rho(t)] &= \int_M \varphi'(\rho)\frac{\mathrm{d}}{\mathrm{d}t}\rho\, d\ell + \int \beta U \frac{\mathrm{d}}{\mathrm{d}t}\rho\, d\ell \\
&= \int_M (\varphi'(\rho) + \beta U)\frac{\mathrm{d}}{\mathrm{d}t}\rho\, d\ell \\
&= \int_M (\varphi'(\rho) + \beta U)\mathrm{div}(\rho(\beta\nabla U + \nabla\varphi'(\rho)))\, d\ell \\
&= -\int_M \nabla(\varphi'(\rho) + \beta U)(\beta\nabla U + \nabla\varphi'(\rho)))\, \rho d\ell \\
&= -\int_M |\nabla\varphi'(\rho) + \beta\nabla U|^2\, d\rho.
\end{aligned}$$

Whence, if we have some inequality à la Łojasiewicz like (23) (as for instance the log-Sobolev inequality of Holley, Kusuoka, and Stroock [29] when $\varphi = \varphi_1$), we derive a differential inequality for the one-variable function $\mathcal{I}[\rho]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathcal{I}[\rho(t)] \;\leq\; -c(\beta)\,\Omega(\mathcal{I}[\rho(t)]) \tag{27}$$

This implies in turn that $\mathcal{I}[\rho(t)]$ converges. If this limit was not zero, the fact that Ω is positive out of 0 would imply, through (27), that the decrease-rate would be perpetually lower than a negative constant, which is absurd. This allows proving that $\mathcal{I}[\rho(t)]$ tends to zero as $t \to \infty$. We thus have proved the first part of:

Theorem 1 (Convergence with a non-vanishing penalty parameter)**.** *Assume that* (**A**), (**B**) *are satisfied, and that there exist $c > 0$, $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ increasing, such that an inequality of the type*

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell \;\;\geq\;\; c\,\Omega\left(\int_M \varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu)\, d\ell\right), \tag{28}$$

holds true whenever ρ is measurable and the left hand side is finite. Then

(i) $\mathcal{U}_\beta[\rho(t)] \to \min \mathcal{U}_\beta$.

(ii) If moreover $\Omega(s) = \Theta(s^{2\theta})$ at 0, with $\theta \in (0,1)$, then $\rho(t)$ tends to μ_β for the Monge-Kantorovich metric, i.e., for the weak topology.*

Proof. Item (i) is already proved. For (ii), from $\Omega(s) = \Theta(s^{2\theta})$ holds, we deduce that the lower semicontinuous function $\mathcal{U}_\beta$ satisfies a Łojasiewicz inequality as in [6, Theorem 2], so that one may assert that the curve $t \mapsto \rho(t)$ has a finite Monge-Kantorovich length; convergence rates depending on θ are also available. □

What are the conditions for the above inequality (28) to be valid, is a delicate open question. The subject of the next section, and the central result of this paper, is to establish such inequalities for one-dimensional compact manifolds and a family of power-like potentials φ.

3 A functional inequality on the circle

3.1 Main theorem

For $m \in (0,1/2)$, set

$$\forall\, r \geq 0, \qquad \varphi_{m,2}(r) \;:=\; \begin{cases} \varphi_m(r) & \text{if } r \in (0,1], \\ \varphi_2(r) = \frac{(r-1)^2}{2} & \text{if } r \in (1,+\infty). \end{cases} \tag{29}$$

The function $\varphi_{m,2}$ is convex on $[0,+\infty)$ and $\mathcal{C}^2$ on $(0,+\infty)$. The latter property is a consequence of the fact that

$$\begin{aligned} \varphi_m(1) &= 0 \\ \varphi_m'(1) &= 0 \\ \varphi_m''(1) &= 1. \end{aligned}$$

Observe also that $\varphi_{m,2}'$ is concave on $(0,+\infty)$ with $\varphi_{m,2}'(0) = -\infty$, so that Lemma 1 applies. Let us recall that the unique solution of $(\mathcal{P}_\beta)$ is denoted by $\mu = \mu_\beta$ and that $\psi_{m,2} = [\varphi_{m,2}']^{-1}$.

This section is devoted to the proof of:

Theorem 2 (A new functional inequality on the circle)**.** *Assume that M is the circle $\mathbb{T} := \mathbb{R}/(L\mathbb{Z})$ of perimeter $L > 0$ endowed with its usual Riemannian structure. Then there exists a constant $c(\beta)$, depending on $\mu_{\min}$ and L, such that for any measurable density ρ on $\mathbb{T}$*

$$\int_{\mathbb{T}} |\nabla\varphi_{m,2}'(\rho) - \nabla\varphi_{m,2}'(\mu)|^2 \rho\, d\ell \geq c(\beta)\, \Omega\left(\int_{\mathbb{T}} \varphi_{m,2}(\rho) - \varphi_{m,2}(\mu) - \varphi_{m,2}'(\mu)(\rho-\mu)\, d\ell\right)$$

where, for β large,

$$\begin{aligned} c(\beta) &= O\left(\beta^{\frac{-3(2-m)}{1-2m}}\right) \\ \Omega(r) &= \begin{cases} r^{\frac{3}{2}} & \textit{if } r \in [0,1) \\ r^{\frac{1-2m}{2(1-m)}} & \textit{if } r \geq 1. \end{cases} \end{aligned}$$

Corollary 1 (An inequality à la Talagrand)**.** *Under the assumptions of the previous theorem, for any measurable density ρ on $\mathbb{T}$,*

$$\int_{\mathbb{T}} \varphi_{m,2}(\rho) - \varphi_{m,2}(\mu) - \varphi_{m,2}'(\mu)(\rho-\mu)\, d\ell \geq d(\beta)\, \omega\Big(\mathcal{W}_2(\rho,\mu)\Big)$$

where $d(\beta) = O\left(\beta^{\frac{-3(2-m)}{1-2m}}\right)$, and

$$\omega(r) \;=\; \begin{cases} 8/5\, r^{\frac{5}{8}} & \textit{if } r \in [0,1) \\ 4(1-m)/(3-2m)\, r^{\frac{3-2m}{4(1-m)}} & \textit{if } r \geq 1. \end{cases} \tag{30}$$

The rest of this section is devoted to the proof of this theorem (the corollary will follow easily using a generalization of Otto-Villani theorem [6]). Most of the intermediary results we provide are valid for a general compact C^2 manifold; thus, unless otherwise stated, we assume that M is arbitrary for the moment. In the remaining subsections of the present section, for the sake of simplicity, we shall often write $\varphi = \varphi_{m,2}$ and $\psi_{m,2} = \psi$.

3.2 Some estimates

Let us define the positive quantities

$$\mu_{\min} \coloneqq \min_M \mu \ \text{ and } \ \mu_{\max} \coloneqq \max_M \mu. \tag{31}$$

Proposition 2 (Bounds for the stationary measure)**.** *We have, for any* $0 < m < 1$*,*

$$(1 + (1-m)\beta \mathrm{osc}(U))^{\frac{1}{m-1}} \leq \ \mu_{\min} \ \leq \ \mu_{\max} \ \leq \beta \mathrm{osc}(U) + 1.$$

where we recall that $\mathrm{osc}(U) = \max_M U - \min_M U$*.*

Proof. The stationary measure μ satisfies for every $x \in M$, $\mu(x) = \psi(c^* - \beta U(x))$ for some real constant c^* and with $\beta > 0$ (recall (13)). Because ψ is nondecreasing , we have

$$\forall x \in M,\ \psi(c^* - \beta \max_M U) \leq \mu(x) \leq \psi(c^* - \beta \min_M U).$$

Integrating with respect to the probability measure ℓ, we obtain

$$\psi(c^* - \beta \max_M U) \leq 1 \leq \psi(c^* - \beta \min_M U).$$

Because φ' is nondecreasing and $\varphi'(1) = 0$, we obtain the following bounds for the constant c^*

$$\beta \min_M U \leq c^* \leq \beta \max_M U.$$

Finally, for every $x \in M$, $\psi(-\beta \mathrm{osc}(U)) \leq \mu(x) \leq \psi(\beta \mathrm{osc}(U))$. Because $\beta \mathrm{osc}(U) \geq 0$, we have for any $m \in (0,1)$,

$$(1 - (m-1)\beta \mathrm{osc}(U))^{\frac{1}{m-1}} \leq \ \mu_{\min} \ \leq \ \mu_{\max} \ \leq 1 + \beta \mathrm{osc}(U),$$

which ends the proof. □

The following formal observation is essential. When the potential function φ is the entropy function φ_1, the density of ρ with respect to μ plays a pivotal role in the establishment of the log-Sobolev inequality (24), see [29]. In our case, the counterpart is the function

$$\psi(\varphi'(\rho) - \varphi'(\mu)).$$

It is also convenient to use the quantity

$$g := \ \varphi'(\rho) - \varphi'(\mu) \tag{32}$$

so that

$$\rho \ = \ \psi(g + \varphi'(\mu)). \tag{33}$$

An upper bound for the reduced cost $\mathcal{I}[\rho]$ This necessitates three steps.

Lemma 3. *For any measurable density ρ, we have*

$$\varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu) \quad \leq \quad \varphi(\psi(g + \varphi'(\mu_{\max}))) - \varphi(\mu_{\max}) - \varphi'(\mu_{\max})(\psi(g + \varphi'(\mu_{\max})) - \mu_{\max})$$

Proof. By definition of g, we have

$$\varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu) \quad = \quad \varphi(\psi(g + \varphi'(\mu))) - \varphi(\mu) - \varphi'(\mu)(\psi(g + \varphi'(\mu)) - \mu).$$

Fix $x \in M$ and consider the function F defined on $(0, +\infty)$ by

$$\forall\, r > 0, \qquad F(r) \quad := \quad \varphi(\psi(g(x) + \varphi'(r))) - \varphi(r) - \varphi'(r)(\psi(g(x) + \varphi'(r)) - r)$$

To prove the result, it is sufficient to show that F is nondecreasing . For $r > 0$, we compute

$$\begin{aligned}
F'(r) \quad &= \quad \varphi'(\psi(g(x) + \varphi'(r)))\psi'(g(x) + \varphi'(r)))\varphi''(r) - \varphi'(r) - \varphi''(r)(\psi(g(x) + \varphi'(r)) - r) \\
&\quad\; -\varphi'(r)[\psi'(g(x) + \varphi'(r)))\varphi''(r) - 1] \\
&= \quad \big[\varphi'(\psi(g(x) + \varphi'(r)))\psi'(g(x) + \varphi'(r)) - (\psi(g(x) + \varphi'(r)) - r) - \varphi'(r)\psi'(g(x) + \varphi'(r))\big]\, \varphi''(r) \\
&= \quad \big[(g(x) + \varphi'(r))\psi'(g(x) + \varphi'(r)) - (\psi(g(x) + \varphi'(r)) - r) - \varphi'(r)\psi'(g(x) + \varphi'(r))\big]\, \varphi''(r) \\
&= \quad \big[g(x)\psi'(g(x) + \varphi'(r)) - (\psi(g(x) + \varphi'(r)) - r)\big]\, \varphi''(r) \\
&= \quad \big[g(x)\psi'(\varphi'(s)) - (s - r)\big]\, \varphi''(r)
\end{aligned}$$

where we set $s := \psi(g(x) + \varphi'(r))$ in the last equality. Because $\psi = (\varphi')^{-1}$, we have

$$\psi'(\varphi'(s)) \quad = \quad \frac{1}{\varphi''(s)}$$

and we get for $r > 0$,

$$\begin{aligned}
F'(r) \quad &= \quad [g(x) - \varphi''(s)(s - r)]\frac{\varphi''(r)}{\varphi''(s)} \\
&= \quad [\varphi'(s) - \varphi'(r) - \varphi''(s)(s - r)]\frac{\varphi''(r)}{\varphi''(s)}.
\end{aligned}$$

Because, the function φ is convex and φ' is concave, we have $\varphi''(r)/\varphi''(s)$ is positive and the quantity $\varphi'(s) - \varphi'(r) - \varphi''(s)(s - r)$ is nonnegative, so that $F' \geq 0$ on $(0, +\infty)$ and F is thus nondecreasing . □

We define further $\xi_{\max}(s) := \psi(s + \varphi'(\mu_{\max}))$ for any real number s and set $\rho_{\max} := \xi_{\max}(g)$. Therefore, Lemma 3 can be rewritten

$$\varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu) \leq \varphi(\rho_{\max}) - \varphi(\mu_{\max}) - \varphi'(\mu_{\max})(\rho_{\max} - \mu_{\max}). \tag{34}$$

We are in position to give an upper bound for $\mathcal{I}[\rho]$.

Lemma 4 (An upper bound for the reduced cost $\mathcal{I}$)**.** *For every $\rho \in \mathcal{P}(M)$ such that $g \in L^2(\ell)$, we have*

$$\mathcal{I}[\rho] \leq \int_M g^2(x)\, \ell(dx).$$

Proof. Because φ is convex, we have

$$\varphi(\mu) - \varphi(\rho) - \varphi'(\rho)(\mu - \rho) \quad \geq \quad 0$$

and

$$\varphi(\mu_{\max}) - \varphi(\rho_{\max}) - \varphi'(\rho_{\max})(\mu_{\max} - \rho_{\max}) \quad \geq \quad 0.$$

Adding the latter positive quantity to the right-hand-side of equation (34) gives

$$\begin{aligned} 0 \leq \varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho - \mu) \quad &\leq \quad (\varphi'(\rho_{\max}) - \varphi'(\mu_{\max}))(\rho_{\max} - \mu_{\max}) \\ &= \quad g(\xi_{\max}(g) - \xi_{\max}(0)). \end{aligned}$$

Recalling that φ has been constructed by gluing φ_m and φ_2 at 1 (see equation (29)), we have that $\psi = (\varphi')^{-1}$ is convex and increasing with $0 \leq \psi' \leq 1$. Therefore, $\xi_{\max}$ is convex and we have

$$\psi'(\varphi'(\mu_{\max}))g \leq \xi_{\max}(g) - \xi_{\max}(0) \leq \psi'(g + \varphi'(\mu_{\max}))g.$$

Whence,

$$|\xi_{\max}(g) - \xi_{\max}(0)| \leq |g|,$$

which gives the desired result. □

A lower bound for the squared Monge-Kantorovich gradient $\mathcal{J}[\rho]$ Once more several steps are necessary to obtain a bound. Let us define the function $\theta : \mathbb{R} \to \mathbb{R}$ via

$$\forall\ r \in \mathbb{R}, \qquad \theta(r) \quad = \quad \int_0^r \sqrt{\psi(s + \varphi'(\mu_{\min}))}\, ds. \tag{35}$$

We observe first that:

Lemma 5 (Lower bound for θ). *Assume $0 < m < \frac{1}{2}$. For any $r \in \mathbb{R}$,*

$$|\theta(r)| \geq \frac{2}{3} \left(\min \left(\frac{1}{(C_0 + (1-m))^{\frac{3}{2} - \eta}}, \sqrt{\psi'(\varphi'(\mu_{\min}))} \right) \right) \min(|r|^{3/2}, |r|^{\eta}),$$

where $C_0 = 1 - (1-m)\varphi'(\mu_{\min})$ and $\eta := \frac{1-2m}{2(1-m)} \in (0, 1/2)$.

Proof. Assume first that $r > 0$. Because ψ is convex, we have for every s,

$$\psi(s + \varphi'(\mu_{\min})) \geq \mu_{\min} + \psi'(\varphi'(\mu_{\min}))s.$$

But $\mu_{\min} > 0$ thus

$$\begin{aligned} \theta(r) &\geq \sqrt{\psi'(\varphi'(\mu_{\min}))} \int_0^r \sqrt{s}\, ds \\ &= \frac{2}{3} \sqrt{\psi'(\varphi'(\mu_{\min}))} r^{3/2}. \end{aligned}$$

Now, assume that $r < 0$. By a change of variables, $t = -s$ and $\tau = -r > 0$, we get

$$-\theta(r) = \int_0^\tau \sqrt{\psi(\varphi'(\mu_{\min}) - t)}\, dt.$$

Remembering that ψ is both convex and positive, we get, for all t,

$$\psi(\varphi'(\mu_{\min}) - t) \geq \psi'(\varphi'(\mu_{\min}) - \tau)(\tau - t).$$

We deduce that

$$-\theta(r) \geq \frac{2}{3}\sqrt{\psi'(\varphi'(\mu_{\min}) - \tau)}r^{3/2}.$$

We shall now use the explicit form of the derivative of ψ given by

$$\psi'(\tau) \quad := \quad \begin{cases} (1 + (m-1)\tau)^{\frac{2-m}{m-1}} & \text{if } \tau \in (-\infty, 0) \\ 1 & \text{if } \tau \in (0, +\infty) \end{cases} \tag{36}$$

By definition of $\mu_{\min}$, we have $\mu_{\min} < 1$ whenever U is not constant over M. Therefore $\varphi'(\mu_{\min}) \leq 0 < \tau$, so that

$$\sqrt{\psi'(\varphi'(\mu_{\min}) - \tau)} = \big(1 - (1-m)\varphi'(\mu_{\min}) + (1-m)\tau\big)^{\frac{2-m}{2(m-1)}}.$$

Set

$$C_0 = 1 - (1-m)\varphi'(\mu_{\min}), \quad \alpha = \frac{2-m}{2(1-m)} \in (1, 3/2)$$

and

$$\eta = 3/2 - \alpha = (3(1-m) - (2-m))/(2(1-m)) = (1-2m)/2(1-m).$$

Therefore,

$$\begin{aligned} -\theta(r) \quad &\geq \quad \frac{2}{3}\frac{\tau^{3/2}}{(C_0 + (1-m)\tau)^{\alpha}} \\ &= \quad \frac{2}{3}\left(\frac{\tau^{3/2}}{(C_0 + (1-m)\tau)^{\alpha}}\mathbb{1}_{0<\tau\leq 1} + \frac{\tau^{3/2}}{(C_0 + (1-m)\tau)^{\alpha}}\mathbb{1}_{\tau>1}\right) \\ &\geq \quad \frac{2}{3}\left(\frac{\tau^{3/2}}{(C_0 + (1-m))^{\alpha}}\mathbb{1}_{0<\tau\leq 1} + \frac{\tau^{3/2}}{(C_0\tau + (1-m)\tau)^{\alpha}}\mathbb{1}_{\tau>1}\right) \\ &\geq \quad \frac{2}{3}\frac{1}{(C_0 + (1-m))^{\alpha}}\min(\tau^{3/2}, \tau^{\eta}). \end{aligned}$$

□

Remark 4 (On constants)**.** For reasons that will appear later during the study of our global optimization method, it is useful to have a compact expression for the inverse of the constant $2/(3(C_0 + (1-m))^{\alpha})$ appearing in Lemma 5. Using the equality $\psi'(\varphi'(\mu_{\min})) = 1/\varphi''(\mu_{\min})$, this inverse writes

$$C_1(\mu_{\min}) := \frac{3}{2}\max\left[(1 + (1-m)(1 - \varphi'(\mu_{\min})))^{\frac{2-m}{2(1-m)}}, \sqrt{\varphi''(\mu_{\min})}\right] > 1.$$

Observe that, as a function of $\mu_{\min}$, C_1 is decreasing and therefore is bounded above by

$$\frac{3}{2}\max\Big\{\Big[1 + (1-m)\Big[1 - \varphi'\Big(\big[1 + (1-m)\beta\mathrm{osc}(U)\big]^{1/(m-1)}\Big)\Big]\Big]^{\frac{2-m}{2(1-m)}}; \\ \sqrt{\varphi''((1 + (1-m)\beta\mathrm{osc}(U))^{1/(m-1)})}\Big\},$$

according to Proposition 2. When β goes to $+\infty$, this bound behaves as $O\left(\beta^{\frac{2-m}{2(1-m)}}\right)$. Finally, note that

$$C_1(\mu_{\min})|\theta(r)| \geq \min(|r|^{3/2}, |r|^{\eta}) \text{ for all } r. \tag{37}$$

We now turn to the desired lower bound for the squared Monge-Kantorovich gradient $\mathcal{J}[\rho]$.

Lemma 6 (Lower bound for the squared Monge-Kantorovich gradient)**.** *We have*

$$\mathcal{J}[\rho] = \int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell \quad \geq \quad \int_M |\nabla\theta(g)|^2\, d\ell.$$

Proof. Taking into account that both φ' and ψ are nondecreasing functions, we get

$$\begin{aligned} \rho = \psi(g + \varphi'(\mu)) &\geq \psi(g + \varphi'(\mu_{\min})) \\ &= (\theta'(g))^2. \end{aligned}$$

It ensues

$$\begin{aligned} \int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho\, d\ell &= \int_M |\nabla g|^2 \rho\, d\ell \\ &\geq \int_M |\nabla g|^2 (\theta'(g))^2\, d\ell \\ &= \int_M |\theta'(g)\nabla g|^2\, d\ell \\ &= \int_M |\nabla\theta(g)|^2\, d\ell \end{aligned}$$

□

3.3 Proof of Theorem 2 and its corollary

Assume for the moment that M is arbitrary.
In the previous section, we have proved two inequalities:

$$\mathcal{I}[\rho] \leq \int_M g^2\, d\ell \text{ and } \mathcal{J}[\rho] \geq \int_M |\nabla\theta(g)|^2\, d\ell.$$

To reach a conclusion, it suffices to relate the quantities

$$\int_M g^2\, d\ell \text{ and } \int_M |\nabla\theta(g)|^2\, d\ell.$$

Since $\eta \in (0, 1/2)$, Lemma 5 and (37) (see Remark 4) gives

$$C_1(\mu_{\min})|\theta(r)| \geq |r|^{3/2}, \text{ when } |r| < 1 \text{ and } C_1(\mu_{\min})|\theta(r)| \geq |r|^{\eta} \text{ otherwise.}$$

When β is large enough $C_1(\mu_{\min}) \geq 1$, we may thus write

$$g^2 \leq C_2(\mu_{\min}) \max(|h|^{\frac{4}{3}}, |h|^{\frac{2}{\eta}}),$$

where we have set $h := \theta(g)$ and $C_2 = C_1^{\frac{2}{\eta}}(\mu_{\min})$. Whence, taking the supremum and integrating yields

$$\int_M g^2 \, d\ell \leq C_2(\mu_{\min}) \max(||h||_\infty^{\frac{4}{3}}, ||h||_\infty^{\frac{2}{\eta}}).$$

Recall that the function $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ is such that

$$\Omega(r) \quad := \quad \begin{cases} r^{\frac{3}{2}} & \text{if } r \in [0,1) \\ r^\eta & \text{if } r \geq 1. \end{cases}$$

Let us prove that the increasing function Ω satisfies the inequality

$$\forall x > 1, \forall y > 0, \quad \Omega(xy) \leq x^{\frac{3}{2}} \Omega(y). \tag{38}$$

Let us consider two cases:
First case: $y \geq 1$. Because $x > 1$, this implies $xy > 1$. Thus,

$$\Omega(xy) = (xy)^\eta = x^\eta \Omega(y) \leq x^{\frac{3}{2}} \Omega(y).$$

Second case: $y < 1$. When $xy > 1$, the inequality follows as above. On the other hand, if $xy < 1$, we have

$$\Omega(xy) = (xy)^{\frac{3}{2}} = x^{\frac{3}{2}} \Omega(y).$$

Therefore, by using the fact that Ω is increasing and satisfies (38), we get

$$\Omega(\mathcal{I}[\rho]) \leq \Omega\left(\int_M g^2 \, d\ell\right) \leq (C_2(\mu_{\min}))^{\frac{3}{2}} ||h||_\infty^2.$$

To end the proof, it remains to compare

$$||h||_\infty^2 \text{ and } \int_M |\nabla h|^2 \, d\ell.$$

It is only at this point that we use the assumption about the dimension of M.

Let us start first by an observation regarding regularity and prove that if $\mathcal{J}$ is finite then ρ must be continuous. Observe first that for $\gamma > 0$ and any measurable ρ, we have

$$\begin{aligned} \mathcal{J}(\rho) &= \int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho \, d\ell & (39) \\ &= \int_M |\nabla\varphi'(\rho)|^2 \rho \, d\ell + \int_M |\nabla\varphi'(\mu)|^2 \, d\ell - 2\int_M \nabla\varphi'(\mu)\nabla\varphi'(\rho)\rho \, d\ell & (40) \\ &\geq \frac{1}{2}\int_M |\nabla\varphi'(\rho)|^2 \rho \, d\ell - \int_M |\nabla\varphi'(\mu)|^2 \rho \, d\ell & (41) \end{aligned}$$

where we have used $2|\nabla\varphi'(\mu)\nabla\varphi'(\mu)| \leq 2|\nabla\varphi'(\mu)|^2 + \frac{1}{2}|\nabla\varphi'(\rho)|^2$. Setting

$$v(r) = \int_1^r \sqrt{s}\varphi'(s) \, ds, \text{ for } r > 0$$

we see that

$$\int_M |\nabla \varphi'(\rho)|^2 \rho \, d\ell = \int_M |\nabla v(\rho)|^2 \, d\ell.$$

Thus the finiteness of $\mathcal{J}[\rho]$ implies that $\int_M |\nabla v(\rho)|^2 \, d\ell$ is finite too.

Assuming now that $\dim M = 1$, standard results ensures that $v(\rho)$ is absolutely continuous and thus so is ρ (so that we have furthermore $\int (v(\rho))^2 \, d\ell < +\infty$ and $v(\rho)$ belongs to the Sobolev space $W^{1,2}(M)$).

Observe that we also obtain that ρ is positive everywhere. Since the function $\rho - \mu$ is continuous and satisfies $\int_M \rho - \mu \, d\ell = 0$, there exists at least one point x_0 in M, such that $\rho(x_0) - \mu(x_0) = 0$. It follows from (32) that $g(x_0) = 0$ and from (35) that $h(x_0) = 0$ (where $h = \theta(g)$). For any $x \in \mathbb{R}/(L\mathbb{Z})$, denote by $[x_0, x]$ the shortest segment in $\mathbb{R}/(L\mathbb{Z})$ with boundary points $\{x_0, x\}$. Since h is absolutely continuous:

$$\begin{aligned} h^2(x) &= \left(\int_{[x_0,x]} h'(y) \, \ell(dy) \right)^2 \\ &\leq \ell([x_0,x]) \int_{[x_0,x]} (h'(y))^2 \, \ell(dy) \\ &\leq \frac{L}{2} \int_M |\nabla h|^2 \, d\ell. \end{aligned}$$

Gathering the previous results gives

$$\mathcal{J}[\rho] \geq c(\beta) \Omega(\mathcal{I}[\rho]),$$

with

$$c(\beta) = \frac{2}{L} C_2 (\mu_{\min})^{-\frac{3}{2}} \tag{42}$$

Since $2/\eta = 4(1-m)/(1-2m)$ and for β large enough, $C_1 = O\left(\beta^{\frac{2-m}{2(1-m)}}\right)$, one has $C_2 = O\left(\beta^{\frac{2(2-m)}{1-2m}}\right)$ and

$$c(\beta) = O\left(\beta^{\frac{-3(2-m)}{1-2m}}\right).$$

Let us conclude with the proof of Corollary 1. The key is to observe that $1/\sqrt{\Omega}$ is integrable at 0, so that the inequality is a Łojasiewicz gradient inequality. It suffices to use the generalization of Otto-Villani theorem provided in [6, Theorem 1(i)].

4 Time-dependent swarm gradient methods

Time dependence is key to obtain convergence to the actual global minimum: the penalty schedule β is tuned so that the exploratory forces embodied in $\mathcal{H}$ are sufficiently active at the beginning of the process while progressively loosing their influence on the dynamics as global goals have been achieved. In this second phase, as the diffusion process generated by $\mathcal{H}$ fades away, the gradient dynamics of U dominates and somehow terminates the process. As in the famous simulated annealing method, the presence of a functional inequality is fundamental for the dynamical system to converge.

In the remainder, unless otherwise stated we use a general potential φ.

4.1 Main convergence results

Convergence under a functional inequality The following general theorem shows the global optimization properties of (16), provided that a functional inequality is available (as in simulated annealing or as in Theorem 4 below).

Theorem 3 (Global optimization under a weak functional inequality)**.** *Assume the set of hypothesis A, B are met, and that β, U, satisfy a functional inequality of the type:*

$$\int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\mu)|^2 \rho \, d\ell \geq c(\beta)\, \Omega\left(\int_M \varphi(\rho) - \varphi(\mu) - \varphi'(\mu)(\rho-\mu)\, d\ell\right)$$

where $c, \Omega : (0, +\infty) \to \mathbb{R}$ are positive with Ω being nondecreasing. If the penalization schedule $t \mapsto \beta(t)$ satisfies

$$\lim_{t\to+\infty} \dot{\beta}(t)/c(\beta(t)) = 0 \tag{43}$$

$$\int_1^{+\infty} c(\beta(t))\, dt = +\infty \tag{44}$$

then

$$\mathcal{U}[\rho(t)] - \min_M U = \int_M U\rho(t) - \min_M U \ \leq\ \text{gap}(\beta_t) + o(\beta_t^{-1}), \tag{45}$$

where the quantity $\text{gap}(\beta_t) \to 0$ *was defined in* (8).

Remark 5 (Existence of a schedule)**.** Assume that the function c satisfies for large $\beta > 0$,

$$c(\beta) \quad = \quad O(\beta^{-\gamma})$$

for some exponent $\gamma > 0$ (as it is the case in Proposition 2). Then for any $\alpha \in (0, 1/(1+\gamma))$, any penalization schedule $t \mapsto \beta(t)$ such that for large enough $t > 0$, $\beta(t) = t^\alpha$ satisfies the assumptions of the convergence theorem, as it is readily checked.

Proof of Theorem 3

Proof. Recall that the evolution equation (16) writes

$$\forall\ t \geq 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\rho(t) = \mathrm{div}(\rho(\beta_t \nabla U + \nabla\varphi'(\rho))).$$

In view of (12), the curve of stationary measures $\nu_t \coloneqq \mu_{\beta_t}$, satisfies, for each fixed t

$$\beta_t \nabla U + \nabla\varphi'(\nu_t) \ = \ 0. \tag{46}$$

The functionals $\mathcal{I}$ and $\mathcal{J}$ used in Section 2.4 are adapted to the time-inhomogeneous case as follows

$$\mathcal{I}[t,\rho] \coloneqq \int_M \varphi(\rho)\, d\ell + \int_M \beta_t U\, \rho\, d\ell - \left(\int_M \varphi(\nu_t)\, d\ell + \int_M \beta_t U\, \nu_t\, d\ell\right),$$

and

$$\mathcal{J}[t,\rho] = \int_M |\nabla\varphi'(\rho) - \nabla\varphi'(\nu_t)|^2 \rho\, d\ell,$$

for any admissible ρ and time $t \geq 0$. Using regularity results (19)-(20), the differentiation of the objective along the evolution curve $t \mapsto \rho$ of (16) yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{I}[t,\rho(t)] &= \int \varphi'(\rho)\frac{\mathrm{d}}{\mathrm{d}t}\rho\, d\ell + \int \beta_t U \frac{\mathrm{d}}{\mathrm{d}t}\rho\, d\ell + \frac{\mathrm{d}}{\mathrm{d}t}\beta_t \int_M U(\rho-\nu_t)\, d\ell \\
&= \int (\varphi'(\rho)+\beta_t U)\mathrm{div}(\rho(\beta_t \nabla U + \nabla\varphi'(\rho)))\, d\ell + \dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell \\
&= -\int \nabla(\varphi'(\rho)+\beta_t U)(\beta_t \nabla U + \nabla\varphi'(\rho)))\, \rho d\ell + \dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell \\
&= -\int_M |\nabla\varphi'(\rho)+\beta_t\nabla U|^2\, \rho d\ell + \dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell \\
&= -\mathcal{J}[t,\rho(t)] + \dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell,
\end{aligned}$$

where we have used (46) in the first equality, the evolution equation in the second-one, and finally integration by parts. Let us set $v(t) = \mathcal{I}(t,\rho)$. Using the functional inequality gives

$$\dot{v}(t) \leq -c(\beta_t)\Omega(v(t)) + \dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell.$$

To give an upper bound of the second term, we write

$$\left|\dot{\beta}_t \int_M U(\rho-\nu_t)\, d\ell\right| = \left|\dot{\beta}_t\left(\int_M U\rho\, d\ell - \int_M U\nu_t\, d\ell\right)\right| \leq \mathrm{osc}(U)\left|\dot{\beta}_t\right|,$$

where the last inequality uses that ρ and ν_t are probability densities on M. Finally, we obtain

$$\dot{v}(t) \leq -c(\beta_t)\Omega(v(t)) + \mathrm{osc}(U)\left|\dot{\beta}_t\right|,$$

We are thus led to consider differential inequalities of the type

$$\dot{v} \quad \leq \quad -c(\beta)\Omega(v) + \delta\left|\dot{\beta}\right|$$

where $v : \mathbb{R}_+ \to \mathbb{R}_+$ is a nonnegative function, $d > 0$ is a positive constant and Ω is an nondecreasing function taking positive values on $(0,+\infty)$.

Our goal is now to give conditions on the inverse temperature scheme $\beta : \mathbb{R}_+ \to (0,+\infty)$ ensuring that v converges to zero for large times:

Proposition 7 (Schedule conditions). *Assume that for large times,* (43) *and* (44) *hold, then*

$$\lim_{t\to+\infty} v(t) \quad = \quad 0.$$

The proof of Proposition 7 is based on the two following observations.

Lemma 8. *Under the assumptions of Proposition 7, we have*

$$\liminf_{t\to+\infty} v(t) \quad = \quad 0.$$

Proof. Towards a contradiction assume that there exist $\epsilon > 0$ and $T_0 \geq 0$ such that $v(t) \geq \epsilon$ for all $t \geq T_0$. Then, Ω being nondecreasing,

$$\forall\, t \geq T_0, \qquad \dot{v}(t) \;\leq\; -c(\beta(t))\Omega(\epsilon) + \delta \left|\dot{\beta}(t)\right|.$$

From the first condition (43), there exists $T_1 \geq T_0$ such that

$$\forall\, t \geq T_1, \qquad \left|\dot{\beta}(t)\right| \;\leq\; \frac{1}{2} c(\beta(t))\Omega(\epsilon)$$

and thus

$$\forall\, t \geq T_1, \qquad \dot{v}(t) \;\leq\; -\frac{1}{2} c(\beta(t))\Omega(\epsilon).$$

This implies

$$\forall\, t \geq T_1, \qquad v(t) \;\leq\; v(T_1) - \frac{1}{2}\Omega(\epsilon) \int_{T_1}^{t} c(\beta(s))\, ds,$$

so letting t go to infinity, due to the second condition (44), $\lim_{t\to+\infty} v(t) = -\infty$, in contradiction with the nonnegativity of v. □

Our second ingredient is:

Lemma 9. *Under the assumptions of Proposition 7, fix $\epsilon > 0$. Then there exists $T \geq 0$, such that*

$$\forall\, t \geq T, \quad \big(v(t) = \epsilon \;\;\Rightarrow\;\; \dot{v}(t) < 0\big).$$

Proof. Indeed, consider

$$T \;\coloneqq\; \inf\left\{\tau \geq 0 \,:\, \forall\, t \geq \tau,\; -\frac{1}{2} c(\beta(t))\Omega(\epsilon) + \delta \left|\dot{\beta}(t)\right| \leq 0\right\}$$

which is finite by assumption (43). Proceeding as in Lemma 8, for any $t \geq T$ such that $v(t) = \epsilon$, we may assert that

$$\dot{v}(t) \;\leq\; -\frac{1}{2} c(\beta(t))\Omega(\epsilon) \;<\; 0.$$

□

The proof of Proposition 7 is as follows: by Lemma 8, for any arbitrary small level $\epsilon > 0$, v will always go below ϵ at some point, but by Lemma 9 there exists a time T after which it will no longer be able to cross it upward. Whence v must be below ϵ for large times. This concludes the proof of Proposition 7.

Let us now come back to the proof of Theorem 3: since $v(t)$ tends to zero, we have for all $\epsilon > 0$ a T_0 such that

$$\mathcal{I}[t,\rho] \coloneqq \int_M \varphi(\rho)\, d\ell + \int_M \beta_t U \,\rho\, d\ell - \left(\int_M \varphi(\nu_t)\, d\ell + \int_M \beta_t U \,\nu_t\, d\ell\right) \leq \epsilon$$

that is

$$\int_M U\,\rho\,d\ell - \int_M U\,\nu_t\,d\ell \quad\leq\quad \beta_t^{-1}\left(\int_M \varphi(\nu_t)\,d\ell - \int_M \varphi(\rho)\,d\ell + \epsilon\right)$$

which implies in turn

$$\begin{aligned}\int_M U\,\rho\,d\ell - \min_M U \quad&\leq\quad \beta_t^{-1}\left(\int_M \varphi(\nu_t)\,d\ell + \epsilon\right) + \left(\int_M U\,\nu_t\,d\ell - \min_M U\right)\\ &\leq\quad \mathrm{gap}(\beta_t) + \epsilon\beta_t^{-1}.\end{aligned}$$

The value of ϵ being arbitrary, the results follows. □

One-dimensional global optimization We may now combine Theorem 3 with the functional inequality we obtained previously in Theorem 2.

Theorem 4 (Global optimization by swarm gradient in dimension 1)**.** *Assume that M is the circle $\mathbb{T} := \mathbb{R}/(L\mathbb{Z})$ endowed with its usual Riemannian structure, that the function φ is as in* (29)*, i.e., $\varphi = \varphi_{m,2}$ and that the schedule is given by*

$$\beta(t) = kt^{1/\gamma},\ \text{with } k > 0 \text{ and } \gamma = \tfrac{3(2-m)}{1-2m} \in [6, +\infty).$$

Then

$$\lim_{t\to\infty} \mathcal{U}[\rho(t)] = \min_M U \tag{47}$$

Proof. Theorem 2 provides a functional inequality as required by Theorem 3. From this inequality we may assume that $c(\beta) = \kappa\beta^{-\gamma}$ for some $\kappa > 0$. Now, using Remark 5, we choose $\beta(t) = c^{-1}(t) = \kappa^{-1}t^{\frac{1}{\gamma}}$. This is the choice made by assumption provided that $k = \kappa^{-1}$. One easily checks that the assumptions of Theorem 3 are satisfied, hence the result. □

4.2 A stochastic view on the dynamics: particles swarm optimization

The time discretization of Langevin diffusion and simulated annealing has a long history, see, e.g., [27, 8, 16, 21, 10]. It has experienced a revival with the advent of machine learning applications and the necessity of developing stochastic optimization algorithms with global minimization properties, see, for instance, [42, 46] for stochastic gradient descent (SGD) and variance reduction techniques, or [26] for momentum-based "acceleration" methods with Bayesian sampling methods. In this last section, we do not delve into the details of such a discretization process, but we nevertheless outline a possible approach for the swarm gradient dynamics. It will be properly developed in future works. It relies on the diffusion process associated with the evolution equation

$$\forall\ t \geq 0, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\rho \quad=\quad \mathrm{div}(\rho(\beta_t\nabla U + \nabla\varphi'(\rho))). \tag{48}$$

To evidence this link, let us use the formal integration by parts presented in Section 2.2 to obtain at any time $t \geq 0$ and for every $f \in \mathcal{C}^\infty(M)$,

$$\int_{\{\rho_t>0\}} L_{t,\rho}[f]\,\rho_t d\ell \quad=\quad \int_M f\mathrm{div}(\rho(\beta_t\nabla U + \nabla\varphi'(\rho)))\,d\ell$$

where

$$L_{t,\rho}[f] \quad = \quad \alpha(\rho)\triangle f - \langle \beta_t \nabla U, \nabla f\rangle \tag{49}$$

with $\alpha : (0, +\infty) \to \mathbb{R}_+$ given by

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

It is then natural to associate to equation (48) a Markov process $(X_t)_{t\geq 0}$ whose infinitesimal generator is given by (49) and whose law has density $\rho(t)$ for all $t \geq 0$. Due to the dependence of the evolution on the density, such a Markov process is said to be nonlinear. When M is a flat torus of dimension d, consider the stochastic differential equation

$$X_t = X_0 - \int_0^t \beta_s \nabla U(X_s)\, ds + \int_0^t \sqrt{\alpha(\rho(X_s))}\, dB_s, \tag{50}$$

where $(B_s)_{s\geq 0}$ is a Brownian motion of dimension d. If (X, ρ) is a solution of the stochastic differential equation (50) then an application of the Itô formula shows that ρ is a solution of equation (48). This observation can be extended to any compact Riemannian manifold.

Two particular choices for which the existence of (X, ρ) has been established in the literature are

(i) $\varphi(\rho) = \rho\log(\rho)$, we have $\alpha \equiv 1$, so $L_{t,\rho}$ does not depend on ρ and is the Langevin generator $\triangle\, \cdot - \beta_t \langle \nabla U, \nabla \cdot\rangle$. The existence and uniqueness of a strong solution to the extension of (50) on a compact Riemannian is well-known, as soon as ∇U is Lipschitz, which is true if U is smooth.

(ii) $\varphi(\rho) = \rho^m$ which corresponds to the equation of porous media. On $\mathbb{R}$, the existence and uniqueness of a strong solution to (50) is proven by Benachour, Chassaing, Roynette and Vallois [5] (see Belaribi and Russo [4] for more general functions φ).

The main difficulty in investigating the existence and uniqueness of (50) is the presence of the density ρ_s. It can be relaxed if it could be replaced by integrals of smooth functions with respect to $\rho_s(x)\,\ell(dx)$. It leads us to consider convolutions of ρ_s with respect to some smooth kernels, as those traditionally used in statistics for density estimation, cf. e.g. Silverman [43]. In our geometric setting, it seems natural to resort to the heat kernel (associated to the Laplace-Beltrami operator). To avoid the introduction of further notation, let us just consider the case of a flat torus $M = (\mathbb{R}/\mathbb{Z})^d$, endowed with its usual Riemannian structure. Let $K : (\mathbb{R}/\mathbb{Z})^d \to \mathbb{R}_+$ be a smooth function with support in $[-1/4, 1/4]^d$ (seen as a subset of $(\mathbb{R}/\mathbb{Z})^d$) and such that $\int_M K\, d\ell = 1$. For any probability density ρ on M and $h \in (0, 1)$, which is a bandwidth parameter, set

$$\forall\, x \in M, \qquad \rho_h(x) \quad \coloneqq \quad h^{-d}\int_M K((x-y)/h)\, \rho(y)\ell(dy). \tag{51}$$

When h is small, ρ_h is an approximation of ρ. Replacing in (50), ρ_s by $\rho_{h,s}$, we end up with the stochastic differential equation

$$X_t = X_0 - \int_0^t \beta_s \nabla U(X_s)\, ds + \int_0^t \sqrt{\alpha(\rho_h(X_s))}\, dB_s, \tag{52}$$

which is simpler to investigate, taking into account the usual mean field theory (see for instance Del Moral [17]).

Furthermore, (50) admits natural particles approximations. Indeed, (51) can be extended to any probability measure π on M, by replacing $\rho(y)\ell(dy)$ by $\pi(dy)$. In particular, it makes sense for the empirical measure of N particles, which is a crucial feature for particle approximations. More precisely, consider a system of N particles, $X_1, X_2, ..., X_N$ whose joint evolution is described by the stochastic differential equations,

$$\forall\, n \in [\![N]\!], \qquad dX_n(t) \;=\; -\beta_t \nabla U(X_n(t)) + \sqrt{\alpha(\rho_{N,h}(X_n(t)))}\, dB_n(t) \tag{53}$$

where the $(B_n(t))_{t\geq 0}$, for $n \in [\![N]\!]$, are independent Brownian motions of dimension d, and where

$$\forall\, x \in M, \qquad \rho_{N,h}(x) \;\coloneqq\; \frac{1}{N}\sum_{n=1}^{N}\frac{1}{h}K\left(\frac{x - X_n(t)}{h}\right)$$

namely $\rho_{N,h}$ is given by (51) when ρ is replaced by the empirical measure

$$\rho_N(t) \;\coloneqq\; \frac{1}{N}\sum_{n=1}^{N}\delta_{X_n(t)}$$

Resorting to mean field theory and chaos propagation, when N tends to infinity, the random evolution $(\rho_N(t))_{t\geq 0}$ converges in probability toward the dynamical system $(\rho(t))_{t\geq 0}$ (for related initial conditions), the law of $(X_1(t))_{t\geq 0}$ converges toward that of $(X_t)_{t\geq 0}$, solution of (52), and the trajectories of any fixed finite number of particles become asymptotically independent.

Putting together these observations, we believe that in addition to $(\beta_t)_{t\geq 0}$, schemes $(h_t)_{t\geq 0}$ and $(N_t)_{t\geq 0}$ can be found, with $\lim_{t\to+\infty} h_t = 0$ and $\lim_{t\to+\infty} N_t = \infty$, so that for large times t, the corresponding empirical measures ρ_{N_t} concentrate around the set of global minima of U. There are several ways to increase the number of particles, the most natural one might be to duplicate some of the current particles –but it is also possible to make them appear in independent random positions.

This procedure would provide a new stochastic algorithm for global minimization. At least up to the simulation of particle systems such as (53), but this can be done through the traditional Euler-Maruyama scheme.

5 On swarm gradient algorithms vs. simulated annealing

At the current stage of our understanding, we are not able to properly compare the theoretical or practical performances of both approaches. We merely provide some thoughts in that direction. One of the motivations of our study was to use approximating particles in order to improve the global knowledge of the energy landscape, with the hope that a swarm of interacting particles might do better than a swarm of independent ones (as in simulated annealing where particles follow simulated annealing dynamics with independent Brownian motions).

What follows is therefore informal: it is only meant to illustrate through rough approximations and simple dynamics that interacting swarm methods may be more advantageous.

Exit times of swarm methods: interacting particles vs. independent particles We examine exit times from wells for both the classical Langevin dynamics and our swarm method. For simplicity, we assume the schedule β_t to be constant (and large).

Consider the d-dimensional torus $\mathbb{T}^d$ and the stochastic differential equation:

$$dZ_t \;=\; \sqrt{2}dB_t - \beta\nabla U(Z_t)\,dt. \tag{54}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion and U a Morse mapping, i.e., a C^2 function whose critical points are non-degenerated, we also assume. Given a time $t \geq 0$, a local minimizer b of U over M, a given height $h > 0$, denote by W the connected component of $\{x \in \mathbb{T}^d \,:\, U(x) \leq U(b) + h\}$ containing b. Shrinking h if necessary we assume that $U(b) = \min_W U$. We assume Z_t to be close to b and consider the *exit time* τ_t^{Lang} *from* W:

$$\tau_t^{\mathrm{Lang}} \;\coloneqq\; \inf\{s \geq 0 \,:\, Z_{s+t} \notin W\}$$

For the "swarm-like" Markov process, we consider $X \coloneqq (X_t)_{t\geq 0}$ whose evolution is described by (3) with the choice of $\varphi = \varphi_{m,2}$ defined in (29), we proceed similarly and consider the exit time

$$\tau_t^{\mathrm{swarm}} \;\coloneqq\; \inf\{s \geq 0 \,:\, X_{t+s} \notin W\}.$$

As it will be pleaded below through heuristic considerations, when β is large enough, we expect:

— when b is a non-global minimizer:

$$\tau_t^{\mathrm{swarm}} << \tau_t^{\mathrm{Lang}}, \tag{55}$$

— when b is a global minimizer:

$$\tau_t^{\mathrm{swarm}} >> \tau_t^{\mathrm{Lang}}. \tag{56}$$

In other words, in the presence of interaction forces as in (3), particles escape faster from non-global minimizers and stay longer in the vicinity of global ones.

Heuristic elements for the comparison For large $\beta > 0$, the time t and the position $Z_t = z$ being fixed, τ_t^{Lang} is of order $\exp(h\beta)$ in the following sense:

$$\lim_{\beta\to+\infty} \frac{1}{\beta}\ln(\mathbb{E}_z[\tau_t^{\mathrm{Lang}}]) \;=\; h \tag{57}$$

Actually, finer results are available, Eyring-Kramers formula provides the pre-exponential factors in terms of the Hessian of U at b and on the boundary of W, see, e.g., Chapter 11 of Bovier and den Hollander [9].

Let us come back to our swarm dynamics and the Markov process $X \coloneqq (X_t)_{t\geq 0}$ whose evolution is described by (3) (with $\varphi = \varphi_{m,2}$ as in (29)). Suppose that at some time $t \geq 0$, X_t is close to b and consider the exit time τ_t^{swarm}. This random time is more complex to apprehend than the exit time of the Langevin dynamic – since the evolution of $(X_s)_{s\geq 0}$ uses its marginal distributions. Yet, as we will see, several natural approximations heuristically suggest that it is generally shorter than τ_t^{Lang} when b is a local but non-global minimizer.

Let us investigate the approximations we mentioned. The diffusive coefficient of X_s at time $s \geq 0$ within (3) is $\alpha(\rho_s(X_s))$. According to the previous sections, for large $s \geq 0$, the law ρ_s of X_s is close to the invariant probability measure $\mu_\beta = \psi_m(c(\beta) - \beta U)$ with $\psi_m = (\varphi'_{m,2})^{-1}$. When t is large, it is thus natural to approximate the swarm dynamics by the process $(\widehat{X}_s)_{s\geq 0}$ defined by

$$\forall\ s \geq 0, \qquad d\widehat{X}_s \;=\; \sqrt{2\alpha(\mu_\beta(\widehat{X}_s))}dB_s - \beta \nabla U(\widehat{X}_s)\, ds$$

and denote

$$\widehat{\tau}_t \;\coloneqq\; \inf\{s \geq 0\ :\ \widehat{X}_{t+s} \notin W\}$$

assuming once more that $\widehat{X}_t$ is close to b.

We now provide two lemmas suggesting an even simpler approximation and an estimation of the exit time. To simplify the notation, we assume, with no loss of generality, that $\min_{\mathbb{T}^d} U = 0$.

Lemma 10 (An approximation of the diffusive coefficient)**.** *There exists a constant $C > 0$ such that for $\beta \geq 1$ large enough,*

$$c(\beta) \;\leq\; C\beta^{\frac{d}{d+2}}.$$

Furthermore, for any $x \in \mathbb{T}^d$ such that $U(x) > 0$,

$$\lim_{\beta\to+\infty} \frac{1}{\beta}\alpha(\mu_\beta(x)) \;=\; \frac{1-m}{m}U(x).$$

Proof. Denote x_0 a global minimum of U, namely satisfying $U(x_0) = 0$. Due to the regularity of U, there exists $r > 0$ small enough and $\widetilde{C} > 0$ big enough so that

$$\forall\ x \in B(x_0, r), \qquad U(x) \;\leq\; \widetilde{C}\,\|x - x_0\|^2$$

where $B(x_0, r)$ is the ball centered at x_0 and of radius r in $\mathbb{T}^d$. We deduce:

$$\begin{aligned}
1 \;&\geq\; \int_{B(x_0,r)} \psi(c(\beta) - \beta U(x))\,\ell(dx) \\
&\geq\; \int_{B(x_0,r)} (c(\beta) - \beta U(x))\;\ell(dx) \\
&\geq\; \int_{B(x_0,r)} \left(c(\beta) - \widetilde{C}\beta\,\|x - x_0\|^2\right)\;\ell(dx) \\
&=\; \ell(B(x_0,r))c(\beta) - \widetilde{C}\beta\int_{B(x_0,r)} \|x - x_0\|^2\;\ell(dx)
\end{aligned}$$

where in the second bound, we used that

$$\forall\ s \in \mathbb{R}, \qquad \psi_m(s) \;\geq\; s$$

due to the fact that

$$\forall\ s \in \mathbb{R}_+, \qquad \varphi'_{m,2}(s) \;\leq\; s$$

by our choice of the convex function $\varphi_{m,2}$. Thus we have

$$c(\beta) \quad \leq \quad \frac{1}{\ell(B(x_0,r))}\left(1+\widetilde{C}\beta\int_{B(x_0,r)}\|x-x_0\|^2\,\ell(dx)\right)$$

Note that for small $r>0$, $\ell(B(x_0,r))$ is of order r^d and $\int_{B(x_0,r)}\|x-x_0\|^2\,\ell(dx)$ of order r^{d+2}. Thus taking $r=\beta^{-\frac{1}{d+2}}$ for $\beta\geq 1$ large enough, we end up with the first announced result. Recall that for r small enough,

$$\alpha(r)=\frac{1}{m}r^{m-1}.$$

Because μ_β is the invariant measure, for any $x\in\mathbb{T}^d$ with $U(x)>0$, we have $\mu_\beta(x)$ close to 0 for large $\beta>0$ (recall indeed that we assumed that $\min_{\mathbb{T}^d}U=0$) and thus

$$\alpha(\mu_\beta(x))=\frac{1}{m}(\psi_m(c(\beta)-\beta U(x)))^{m-1}.$$

On the other hand, we have for large β,

$$\psi_m(c(\beta)-\beta U(x))=(1+(m-1)(c(\beta)-\beta U(x)))^{1/(m-1)}.$$

Thus, we deduce that

$$\alpha(\mu_\beta(x))=\frac{1}{m}+\frac{m-1}{m}(c(\beta)-\beta U(x)),$$

which implies the second announced result. □

The above lemma suggests that the swarm process may be seen, for large times, close to the process $X^{\mathrm{app}}:=(X^{\mathrm{app}}_s)_{s\geq 0}$ given by

$$\forall\ s\geq 0,\qquad dX^{\mathrm{app}}_t \quad = \quad \sqrt{2\frac{1-m}{m}\beta U(X^{\mathrm{app}}_s)}dB_s-\beta\nabla U(X^{\mathrm{app}}_s)\,ds. \tag{58}$$

The corresponding exit time is:

$$\tau^{\mathrm{app}}_t \quad := \quad \inf\{s\geq 0\ :\ X^{\mathrm{app}}_{t+s}\notin W\}$$

assuming that X^{app}_t is close to b. In order to estimate this quantity, we shall use the following:

Lemma 11 (On exit times and parametrization). *Consider the diffusion $\widetilde{Z}^{(\beta)}:=(\widetilde{Z}^{(\beta)}_s)_{s\geq t}$ satisfying*

$$d\widetilde{Z}^{(\beta)}_s \quad = \quad \sqrt{2\beta S(\widetilde{Z}^{(\beta)}_s)}dB_s-\beta\nabla U(\widetilde{Z}^{(\beta)}_s)\,ds \tag{59}$$

where S is a positive smooth function on $\mathbb{T}^d$. Then, assuming $\widetilde{Z}^{(\beta)}_t=z$ for some time t, we have

$$\forall\ \beta>0,\exists\, C>0\qquad \mathbb{E}_z[\widetilde{\tau}_{t,\beta}] \quad = \quad \frac{C}{\beta} \tag{60}$$

where

$$\widetilde{\tau}_{t,\beta}=\inf\{s\geq 0\ :\ \widetilde{Z}^{(\beta)}_{t+s}\notin W\}$$

Proof. Let us define the Brownian motion $\widehat{B} = (\widehat{B}_s)_{s\geq 0}$ as $\widehat{B}_s = B_{t+s} - B_t$. We consider for $s \geq 0$ the solution of (59),

$$d\widehat{Z}^{(\beta)}_s = \sqrt{2\beta S(\widehat{Z}^{(\beta)}_s)} d\widehat{B}_s - \beta \nabla U(\widehat{Z}^{(\beta)}_s)\, ds,$$

starting with $\widehat{Z}^{(\beta)}_0 = z$, so that $(\widehat{Z}^{(\beta)}_s)_{s\geq 0}$ has the same law as $(\widetilde{Z}^{(\beta)}_{t+s})_{s\geq 0}$ conditioned by $\widetilde{Z}^{(\beta)}_t = z$. Consider furthermore the solution of (59) with $\beta = 1$,

$$d\widehat{Z}_s = \sqrt{2S(\widehat{Z}_s)} d\widehat{B}_s - \nabla U(\widehat{Z}^t_s)\, ds,$$

starting with $\widehat{Z}_0 = z$. We define

$$\widehat{\tau} = \inf\{s \geq 0 \,:\, \widehat{Z}_s \notin W\}.$$

By the scaling property of the Brownian motion, the process $Y = (Y_u)_{u\geq 0}$ defined by $Y_u = \sqrt{\beta}\widehat{B}_{u/\beta}$ is a Brownian motion and we have

$$d\widehat{Z}^{(\beta)}_{u/\beta} = \sqrt{2S(\widehat{Z}^{t,(\beta)}_{u/\beta})} dY_u - \nabla U(\widehat{Z}^{(\beta)}_{u/\beta})\, du.$$

Therefore, $(\widehat{Z}^{(\beta)}_{u/\beta})_{u\geq 0}$ has the same law as $(\widehat{Z}_u)_{u\geq 0}$, so the stopping time $\widetilde{\tau}_{t,\beta}$ has the same law as $\widehat{\tau}/\beta$. In particular $\widetilde{\tau}_{t,\beta}$ is of order $1/\beta$ in the sense that

$$\forall\ \beta > 0, \qquad \mathbb{E}_z[\widetilde{\tau}_{t,\beta}] \quad = \quad \frac{\mathbb{E}_z[\widehat{\tau}]}{\beta} \tag{61}$$

□

Let us now proceed to a discussion on the exit times of interacting and independent particles process from a well with bottom b.

— The local minimum b is not global. In this case, we may "infer" from (61) that τ^{app}_t is of order $1/\beta \ll 1$, implying that X^{app} exits very rapidly from W and in the interval $[t, t + \tau^{\mathrm{app}}_t]$. Due to our approximations, we hope for the same behavior for τ^{swarm}_t: X should indeed escapes from W in a time of order $1/\beta$.
On the other hand, when at a large time $t \geq 0$, the position of the simulated annealing algorithm X_t happens to be close to b, then the exit time τ^{swarm}_t is of order $\exp(h\beta)$.

This is the reason behind the exit time comparison (55).

— The local minimum b is global. In that case, Lemma 11 does not apply since U approaches 0 (recall that $\min_M U = 0$). But this also means that the diffusive coefficient vanishes as one is closer to b, so that gradient forces dominate, yielding a trapping effect near b. This trapping effect is very likely to be superior to the case when the diffusion is constant as for the Langevin diffusion.

This suggests the exit time comparison (56).

Gathering these observations, and assuming our heuristic considerations have some validity, we expect the swarm algorithm to converge faster (and, hopefully, in a stronger sense) towards the global minima than the classical simulated annealing method.

6 Appendix

6.1 On the differential calculus in the probability space $\mathcal{P}(M)$

In this section, we provide some ingredients for an understanding of the equivalence between (a), (b), and (c).

We recall beforehand that the *pushforward* of a measure μ through a Borel map $T : M \mapsto M$, denoted $T\#\mu$, is defined by $(T\#\mu)(A) = \mu(T^{-1}(A))$ for all Borel sets $A \subset M$. We recall that for $f : M \mapsto \mathbb{R}$ bounded, we have

$$\int_M f(x)(T\#\mu)(d\ell) = \int_M f(T(x))\,\mu(d\ell).$$

Let us sketch some essential ideas to view $(\mathcal{P}(M), \mathcal{W}_2)$ through a "Riemannian space" perspective. To do this, we use, sometimes very informally, elements of [1, Chapter 8].

First, we need to have a notion of an absolutely continuous path on $\mathcal{P}(M)$. Given a family of sufficiently regular vector fields[5] on M denoted by $b_t, t \in (-\eta, \eta)$ for $\eta > 0$, we consider the ordinary differential equation

$$\partial_t \phi_t(x) = b_t(\phi_t(x)), \quad \phi_0(x) = x \in M. \tag{62}$$

This defines a family of diffeomorphisms ϕ_t on M; for any $\mu \in \mathcal{P}(M)$, we define $\mu_t = \phi_t\#\mu$. If we consider vector fields on M such that

$$\int_{-\eta}^{\eta} \int_M |b_t(x)|^2 \, d\mu_t \, dt < +\infty,$$

one may derive, using integration by parts, that μ_t satisfies the continuity equation

$$\partial_t \mu_t + \mathrm{div}_M(\mu_t b_t) = 0$$

in the sense of distribution. Roughly speaking, curves $(\mu_t)_{t\in(-\eta,\eta)}$ defined by the above equation are the absolutely continuous curves on $\mathcal{P}(M)$ (see [1, Theorem 8.3.1.]). When one uses the derivative $\partial_t \mu_t\,_{|t=0}$ to define a "tangent space" to $\mathcal{P}(M)$ at μ, we might proceed to the following identification:

$$\bar{T}_\mu \mathcal{P}(M) = \{-\mathrm{div}_M(\mu b) : b \text{ vector field on } M, \int_M |b(x)|^2 \, d\mu < +\infty\}$$

or

$$\tilde{T}_\mu \mathcal{P}(M) = \{b \text{ vector field on } M : \int_M |b(x)|^2 \, d\mu < +\infty\} = L^2(M; \mu).$$

This provides a "differential structure" to $\mathcal{P}(M)$, so we can now turn to the "Riemannian" interpretation of the Monge-Kantorovich distance by Otto to understand the equivalence between the views (b) and (c). For this, we may use the two following "metrics":

$$\langle\langle \mathrm{div}_M(\mu b), \mathrm{div}_M(\mu b') \rangle\rangle = \int_M \langle b(x), b'(x) \rangle \, d\mu.$$

[5] See, e.g., [1, Lemma 8.1.4]

Consider a functional $\mathcal{J} : \mathcal{P}(M) \mapsto \mathbb{R}$, a probability μ and assume that for every $b \in L^2(M;\mu)$, the following quantity exists

$$d\mathcal{J}[\mu](b) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \Big(\mathcal{J}(\phi^b_\epsilon \# \mu) - \mathcal{J}(\mu) \Big)$$

where ϕ^b_t is the flow associated to the ODE (62) with vector field b. Then by using Riesz representation idea, we may formally associate two “concepts" of gradients to this differential mapping, $\mathrm{grad}_{\mathcal{W}} \mathcal{J}[\mu] \in \bar{T}_\mu \mathcal{P}(M)$ and $\nabla_{\mathcal{W}} \mathcal{J}[\mu] \in \tilde{T}_\mu \mathcal{P}(M)$

$$d\mathcal{J}[\mu](b) = \langle\langle \mathrm{grad}_{\mathcal{W}} \mathcal{J}[\mu], \mathrm{div}_M(\mu b') \rangle\rangle = - \int_M \langle \nabla_W \mathcal{J}[\mu], b(x) \rangle \, d\mu,$$

where b is an arbitrary L^2 vector field. Once gradients are defined, one can consider gradient curves for $\mathcal{J}$ through:

$$\partial \rho_t = \mathrm{div}_M(\rho_t \nabla_{\mathcal{W}} \mathcal{J}[\rho_t]) = -\mathrm{grad}_{\mathcal{W}} \mathcal{J}[\rho].$$

The interested reader may consult [1] for a proper derivation of these concepts and the use of subgradient curves which are actually necessary to define properly the gradient systems we use in this paper.

6.2 On the “equivalence" between (a), (b), (c)

We start by connecting (b) to (c) thanks to the previous paragraph. For this, one formally applies the previous considerations to the relative entropy functional

$$\mathcal{U}_\beta(\mu) = \beta \int_M U(x) \mu(x) d\ell(x) + \int_M \rho(x) \log(\rho(x)) \, d\ell(x).$$

Even though $\mathcal{J} = \mathcal{U}_\beta$ is not differentiable (it is not actually defined on the whole of $\mathcal{P}(M)$), we proceed *formally* and we obtain

$$\nabla_{\mathcal{W}} \mathcal{U}_\beta[\mu] = \beta \nabla U + \frac{\nabla \rho}{\rho}.$$

The corresponding Monge-Kantorovich gradient equation therefore writes

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho(t) = -\mathrm{grad}_{\mathcal{W}} \mathcal{U}_{\beta_t}[\rho(t)],$$

that is

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho = \beta_t \, \mathrm{div}_M(\rho \nabla U) + \Delta \rho, \quad t \geq 0.$$

which is exactly the Fokker-Planck equation (b). This shows the equivalence between (b) and (c). Similar considerations apply to the general swarm methods considered in Section 2.

We now establish the connection between the points of view (a) and (b). We consider the solution $(X_t)_t$ to the time-inhomogenous Langevin-like stochastic differential equation where β_t is assumed to be continuous. For every smooth function ϕ on M, Itô’s formula gives for $t \geq 0$ and $h > 0$,

$$\mathbb{E}[\phi(X_{t+h})] = \mathbb{E}[\phi(X_t)] + \mathbb{E} \left[\int_t^{t+h} (-\beta_s \langle \nabla U(X_s), \nabla \phi(X_s) \rangle + \Delta \phi(X_s)) \, ds \right].$$

Denoting by $p(t,x)$ the density function of the probability distribution of X_t, we deduce

$$\int_M \phi(x)(p(t+h,x)-p(t,x))\,d\ell = \int_t^{t+h}\int_M (-\beta_s\langle\nabla U(x),\nabla\phi(x)\rangle + \Delta\phi(x))p(s,x)\,d\ell\,ds.$$

Dividing by h and letting h tend to zero, we obtain

$$\int_M \phi(x)\partial_t p(t,x)\,d\ell = \int_M (-\beta_t\langle\nabla U(x),\nabla\phi(x)\rangle + \Delta\phi(x))p(t,x)\,d\ell.$$

Integrating by parts the right-hand side, we have for every smooth function ϕ,

$$\int_M \phi(x)\partial_t p(t,x)\,d\ell = \int_M \phi(x)\left(\beta_t \mathrm{div}_M(p(t,x)\nabla U(x)) + \Delta p(t,x)\right)\,d\ell,$$

which shows that the density function of X_t satisfies the Fokker-Planck equation (b).

Acknowledgements. JB acknowledges the support of ANR-3IA Artificial and Natural Intelligence Toulouse Institute, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, FA8655-22-1-7012, ANR MasDol and Centre Lagrange. JB, LM, SV acknowledges the support of ANR Chess, grant ANR-17-EURE-0010. LM and SV thank Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA8655-22-1-7016.

Funding and/or Conflicts of interests/Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Ambrosio, N. Gigli and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[2] R. Azencott, editor. *Simulated annealing. Parallelization techniques.* Chichester: John Wiley & Sons Ltd., 1992.

[3] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de probabilités, XIX, 1983/84*, volume 1123 of *Lecture Notes in Math.*, p. 177–206. Springer, Berlin, 1985.

[4] N. Belaribi and F. Russo. Uniqueness for Fokker-Planck equations with measurable coefficients and applications to the fast diffusion equation. *Electron. J. Probab.*, 17, p.1-28, 2012.

[5] S. Benachour, P. Chassaing, B. Roynette and P. Vallois. Processus associés à l'équation des milieux poreux. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 4e série, 23(4), p.793-832, 1996.

[6] A. Blanchet and J. Bolte. A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275 (7), p. 1650-1673, 2018.

[7] A. Blanchet, M. Bonforte, J. Dolbeault, G. Grillo, and J. L. Vàzquez. Asymptotics of the fast diffusion equation via entropy estimates. Arch. Rational Mech. Anal., 191 (2), p. 347-385, 2009.

[8] V. S. Borkar, and S. K. Mitter. A strong approximation theorem for stochastic recursive algorithms. Journal of Optimization Theory and Applications, 100, 499-513, 1999.

[9] A. Bovier and F. den Hollander. *Metastability*, volume 351 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].*

[10] P. Bras and G. Pages: Convergence of Langevin-Simulated Annealing algorithms with multiplicative noise, Arxiv:2109.11669, 2021.

[11] J. A. Carrillo, R. S. Gvalani, G. A. Pavliotis and A. Schlichting. Long-time behaviour and phase transitions for the McKean-Vlasov equation on the torus. *Arch. Ration. Mech. Anal.*, 235 (1), p. 635-690, 2020.

[12] J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani and A. Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatsh. Math.*, 133(1),p. 1–82, 2001.

[13] J. A. Carrillo, R. J. McCann and C. Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Rev. Mat. Iberoam.*, 19(3),p. 971–1018, 2003.

[14] J. Carrillo, R. J. McCann and C. Villani. Contractions in the 2-Wasserstein length space and thermalization of granular media. *Arch. Ration. Mech. Anal.*, 179(2),p.217–263, 2006.

[15] X. Cheng and P. Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Algorithmic Learning Theory* , Proceeding of Machine Learning, 83, p. 186-211, 2018.

[16] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3), p. 651-676, 2017.

[17] P. Del Moral. *Mean field simulation for Monte Carlo integration*, volume 126 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, FL, 2013.

[18] F. Delarue and A. Tse. Uniform in time weak propagation of chaos on the torus. arXiv 2104.14973, 2021.

[19] M. Del Pino and J. Dolbeault. Best constants for Gagliardo-Nirenberg inequalities and applications to nonlinear diffusions, Journal de Mathématiques Pures et Appliquées, 81, p. 847-875, 2002.

[20] M. Dorigo and C. Blum. Ant Colony optimization theory: a survey. *Theoretical Computational Science.*, 344, p. 243–278, 2005.

[21] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. The Annals of Applied Probability, 27(3), p. 1551-1587, 2017.

[22] A. Eberle, A. Guillin and R. Zimmer. Quantitative Harris type theorems for diffusions and McKean-Vlasov processes. *Transactions of the American Mathematical Society*, 371(10), p.7135–7173, 2019.

[23] M. Émery. *Stochastic calculus in manifolds.* Universitext. Springer-Verlag, Berlin, 1989 With an appendix by P.-A. Meyer.

[24] L.C. Ferreira and J.C. Valencia-Guevara. Gradient flows of time-dependent functionals in metric spaces and applications to PDEs. Monatshefte für Mathematik, 185(2), p. 231-268, 2018.

[25] D.B. Fogel. Evolutionary computation towards a new philosophy of machine intelligence. IEEE Press, NJ, Second Ed., 2000.

[26] X. Gao, M. Gürbüzbalaban and L. Zhu. Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration, *Operations Research*, Vol 70 (5), 2932-2947, 2022.

[27] S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$. *SIAM Journal on Control and Optimization*, 29(5):999?1018, 1991.

[28] T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition, Springer in Statistics, 2009.

[29] R. Holley, S. Kusuoka, and D. Stroock. Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.*, 83(2), p. 333-347, 1989.

[30] M. Iacobelli, F. S Patacchini, and F. Santambrogio. Weighted ultrafast diffusion equations: from well-posedness to long-time behaviour. Arch. Rational Mech.Anal., 232, p. 1165-1206, 2019.

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

[32] R. Jordan and D. Kinderlehrer. *An extended variational principle. In Partial Differential Equations and Applications*: Collected Papers in Honor of Carlo Pucci, volume 177 of Lecture Notes in Pure and Applied Mathematics, chapter 18, p. 187-200. CRC Press, 1996.

[33] R. Jordan, D. Kinderlehrer, and F. Otto. *The variational formulation of the Fokker–Planck equation*, SIAM journal on mathematical analysis, 29, p. 1-17, 1998.

[34] J. Kennedy and R. Eberhart. Particle Swarm optimization. *In Proceedings of Int. Conf. on neural networks.*, 4:1942–1946, 1995.

[35] J.-B. Lasserre, Global optimization with polynomials and the problem of moments. SIAM Journal on optimization, vol. 11, no 3, p. 796-817, 2001.

[36] N. Metropolis, A. W. Rosenbluth, M. N., Rosenbluth, A. H. Teller and E. Teller. Equation of state calculations by fast computing machines, The Journal of Chemical Physics 21 (6), p. 1087-1092, 1953.

[37] Y. A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett and M. I. Jordan, Is there an analog of Nesterov acceleration for MCMC? arXiv 1902.00996, 2019.

[38] Y.A. Ma, Y. Chen, C. Jin, N. Flammarion and M.I. Jordan, Sampling can be faster than optimization. Proceedings of the National Academy of Sciences, 116(42), p. 20881-20885, 2019.

[39] L. Miclo. Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre [Simulated annealing on $\mathbf{R}^n$. Study of the free energy evolution]. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2), p. 235-266, 1992.

[40] L. Miclo. Une étude des algorithmes de recuit simulé sous-admissibles. *Ann. Fac. Sci. Toulouse Math. (6)*, 4(4):819–877, 1995.

[41] F. Otto. *The geometry of dissipative evolution equations: the porous medium equation*, Communications in Partial Differential Equations, 26, p. 101-174, 2001.

[42] M. Raginsky, A. Rakhlin and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a non-asymptotic analysis. In Conference on Learning Theory, PLMR, p. 1674-1703, 2017.

[43] B. Silverman. *Density estimation for statistics and data analysis.* CRC Press, Boca Raton, FL, 1986.

[44] A. Taghvaei and P.G. Mehta. Accelerated flow for probability distributions. Proceedings of the 36th Int. Conf. on Machine Learning, 97, p. 6076-6085, 2019.

[45] J.L. Vazquez, The Porous Medium Equation. Mathematical Theory, Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford (2007).

[46] P. Xu, J. Chen, D. Zou and Q. Gu. Global convergence of Langevin dynamics-based algorithms for nonconvex optimization. Advances in Neural Information Processing Systems (31), (2018)