Controlled stochastic processes for simulated annealing

ESAIM: COCV 32 (2026) 24
https://doi.org/10.1051/cocv/2026005

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

CONTROLLED STOCHASTIC PROCESSES FOR SIMULATED ANNEALING

Vincent Molin*, Axel Ringh, Moritz Schauer and Akash Sharma

Abstract. Simulated annealing solves global optimization problems by means of a random walk in a cooling energy landscape based on the objective function and a temperature parameter. However, if the temperature is decreased too quickly, this procedure often gets stuck in suboptimal local minima. In this work, we consider the cooling landscape as a curve of probability measures. We prove the existence of a minimal norm velocity field which solves the continuity equation, a differential equation that governs the evolution of the aforementioned curve. The solution is the weak gradient of an integrable function, which is in line with the interpretation of the velocity field as a derivative of optimal transport maps. We show that controlling stochastic annealing processes by superimposing this velocity field would allow them to follow arbitrarily fast cooling schedules. Here we consider annealing processes based on diffusions and piecewise deterministic Markov processes. Based on convergent optimal transport-based approximations to this control, we design a novel interacting particle–based optimization method that accelerates annealing. We validate this accelerating behaviour in numerical experiments.

Mathematics Subject Classification. 90C26, 53B12, 35Q84, 49Q22, 65C35.

Received July 4, 2025. Accepted January 8, 2026.

1. Introduction

Stochastic exploration methods solve optimization problems by changing the argument of the optimization objective $U(x)$ randomly, for example by making a random walk in the energy landscape induced by the objective function. The magnitude of these random fluctuations can be cast as a temperature, which corresponds to different scalings of the energy landscape. Simulated annealing [1–3] is an important example of this. In its continuous-time diffusion-based version, introduced in [4, 5], the temperature $\beta(t)^{-1}$ is decreased over time to trap the random walker in an optimum. At a time t, the random walkers momentarily follow diffusive dynamics which, if kept unchanged, would mix towards the stationary measure μ_t with density function $\exp(-\beta(t)U(x))$ up to proportionality, under suitable conditions on U. If the temperature is decreased slowly enough, the random walker finds and becomes trapped in a *global* optimum as time goes to infinity [5]. If the temperature is decreased too quickly, the random walkers are not able to keep up with the changes in the energy landscape and get frozen in local minima.

Keywords and phrases: Simulated annealing, optimal transport, interacting particle system, global optimization, stochastic differential equations, piecewise deterministic Markov processes.

Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, Gothenburg, Sweden.

* Corresponding author: molinv@chalmers.se

In this work, we prove the existence of a vector field $v_t(x)$ which captures the changes in the cooling landscape, in the sense that it describes how particles in this landscape, for example, random walkers, have to be pushed to get arbitrarily close to a global optimum in bounded time. This vector field achieves this with minimal effort and arises as derivative $v_t(x) = \lim_{h\to 0} h^{-1}\left(T_{\mu_t\to\mu_{t+h}}(x) - x\right)$ where $T_{\mu_t\to\mu_{t+h}}$ is the Monge optimal transport map (in $\|\cdot\|_2^2$-sense) transporting probability mass distributed as μ_t to probability mass distributed as μ_{t+h}. Then, as an application, we design and implement a novel interacting particle system–based optimization method that accelerates annealing using the vector field $v_t(x)$ approximated by discrete optimal transport.

So let $U\colon \mathbb{R}^d \to \mathbb{R}$ be an objective function of interest and consider the optimization problem

$$\min_{x\in\mathbb{R}^d} U(x).$$

Although many algorithms have been developed to solve these types of problems, see, *e.g.*, one of the monographs [6, 7], it is still difficult to find globally optimal solutions if U is non-convex. One approach to such problems is named after the metallurgic technique called annealing, where a metal is heated and then slowly cooled to alter its physical properties by finding a more stable configuration. The classical continuous-time annealing process for optimization in $\mathbb{R}^d$, studied in, *e.g.*, [4, 5], is given by the stochastic differential equation

$$dX_t = -\nabla U(X_t)\,dt + \sqrt{2\beta^{-1}(t)}\,dB_t, \tag{1.1}$$

where B_t is a d-dimensional Brownian motion. Under fairly general conditions, this process is guaranteed to find a global minimum if the cooling schedule β has at most logarithmic growth [5, 8].

In this work, we will consider the sequence of cooling measures as a curve in a suitable probability space. To this end, for a fixed inverse temperature $\beta \in \mathbb{R}_+$ we define the Gibbs measure corresponding to U as

$$\mu_\beta(dx) := Z_\beta^{-1}\mathrm{e}^{-\beta U(x)}\,dx,$$

where $Z_\beta := \int \mathrm{e}^{-\beta U(x)}\,dx$ is the normalizing constant. For sufficiently regular functions U, the Gibbs measure μ_β concentrates on the set of global minima of U as $\beta\to\infty$, in the sense that μ_β converges weakly to a limiting measure μ_∞ supported on the set of global minima of U [9, 10]. Next, let $\beta(t)$ be a smooth function $\mathbb{R}_+ \to [1,\infty)$ and, by a slight abuse of notation, define the measure

$$\mu_t(dx) := \mu_{\beta(t)}(dx) = Z_{\beta(t)}^{-1}\mathrm{e}^{-\beta(t)U(x)}\,dx. \tag{1.2}$$

We will refer to the map $t\mapsto\mu_t$ as the Gibbs curve.

For a given Markov process with time-dependent generator $\mathcal{A}_t$ (in the sense of Stroock–Varadhan [11], Chap. 6), the Kolmogorov forward equation that governs the evolution of the law is given by

$$\partial_t\rho_t = \mathcal{A}_t^*\rho_t. \tag{1.3}$$

When $\mathcal{A}_t$ is the generator of the diffusion process in (1.1), this corresponds to the PDE

$$\partial_t\rho_t = \nabla\cdot(\rho_t\nabla U) + \frac{1}{\beta(t)}\Delta\rho_t, \tag{1.4}$$

where $\nabla\cdot$ denotes the divergence operator and $\partial_t\rho_t = \frac{\partial}{\partial t}\rho_t$. Further, and crucial to our point of view is now that, for this generator, it holds that

$$\mathcal{A}_t^*\mu_t = 0 \neq \partial_t\mu_t, \quad \text{if } \beta' \neq 0,$$

and hence the Gibbs curve, $t \mapsto \mu_t$, can not be a solution to Equation (1.3). Clearly, this holds for any time-inhomogeneous Markov process satisfying $\mathcal{A}_t^* \mu_t = 0$, which is normally the type of dynamics constructed in the simulated annealing literature, see, for instance, [4, 5, 12, 13]. A question, and the topic under investigation in this paper, is then, can we find a modified $\mathcal{A}_t^o$ such that

$$(\mathcal{A}_t^o)^* \mu_t = \partial_t \mu_t? \tag{1.5}$$

This would then yield that the law of the process undergoes a deformation equivalent to that of the energy landscape, alleviating the need for logarithmic cooling over infinite time. Specifically, we are going to look for deterministic velocity fields v_t such that

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{1.6}$$

with the end goal of superimposing these dynamics on the simulated annealing process. Viewing the Gibbs curve as a curve in an appropriate metric space of probability measures, one choice of v_t arises as limits of optimal transport problems, see, *e.g.*, [14], Chapter 8. We first ensure that the theoretical underpinnings guaranteeing the existence of this velocity are satisfied. We then take inspiration from this and design an interacting particle system algorithm, where we incorporate solving discrete optimal transport problems along the way.

1.1. Related work

It is possible to understand the simulated annealing process (1.1) in a number of other ways. For example, the Jordan–Kinderlehrer–Otto formalism identifies the Langevin dynamics in (1.1) as the Wasserstein-2 gradient flow of the functional

$$F_\beta(\rho) = \int U \, d\rho + \frac{1}{\beta} \int \log\left(\frac{d\rho}{dx}\right) d\rho \tag{1.7}$$

with dx the Lebesgue measure on $\mathbb{R}^d$, see, *e.g.*, [14, 15]. As $F_\beta(\rho) = \frac{1}{\beta} \mathrm{KL}(\rho \mid \mu_\beta) + C(\beta)$, this is the gradient flow of the Kullback–Leibler divergence. Hence, we can view the annealing process as an inhomogeneous gradient flow of the functionals $F_{\beta(t)}$. In general, this gradient flow converges at a speed that can be exponentially slow in β, see Section 3.3, which is related to the fact that $\beta(t)$ can grow at most logarithmically fast in order to guarantee convergence [5, 13].

Different perspectives on simulated annealing have lead to different ways of constructing new optimization methods. For instance, from the gradient flow formulation described above, [16] consider a new family of methods based on modifying the functional in (1.7). Orthogonally, [13] instead consider replacing the diffusive dynamics in (1.1) with piecewise deterministic Markov processes (PDMPs). PDMPs have also been used for representing solutions to (1.3) with a known time-homogeneous diffusion generator, [17].

Interacting particles-based methods have gained increasing attention due to their better capabilities to explore non-convex and anisotropic landscape (see *e.g.* [18–23]). From the algorithmic perspective, the method we design bears a resemblance to interacting particles-based approaches, such as consensus-based optimization [22, 24] and sequential Monte Carlo simulated annealing [18]. More precisely, consensus-based optimization orchestrates interaction among particles where each particle drifts towards a weighted average, where higher weights are assigned to particles if they currently are, relative to the other particles, in positions corresponding to better objective values. In contrast, we assign to each particle an individual target by first solving a population-wide discrete optimal transport problem. This introduces an inhomogeneous interaction between the particles. For ease of presentation, we defer further elaboration on this to Remark 5.4. Finally, while we will look for a velocity field that follows the Gibbs curve by continuously moving the particles, it is possible to understand sequential Monte Carlo simulated annealing [18] as achieving this curve-following by successively removing particles from low probability regions and resampling them in high probability regions. We elaborate on this in Remark 5.5.

A similar velocity field formulation can be found in, *e.g*, [25–28] (see also [29]). While we approximate the velocity field using a particle system and discrete optimal transport, in [25] the authors instead consider computing an inexact but numerically more tractable velocity field. Approximate optimal transport maps have also been investigated for use in Bayesian inference, see for instance [30, 31]. Both [27] and [28] derive controlled (overdamped) Langevin dynamics similar to ours (see (4.3)), the former *via* Girsanov theorem and latter through Fokker-Planck PDE manipulations. Both approaches subsequently train neural networks to approximate the required vector field corresponding to the control in the dynamics. Our work takes a distinct approach based on approximate transport map characterization of the vector field. We also make a theoretical contribution of existence of the required vector field in the form of the gradient of a scalar field. In addition, we also superimpose the controlled dynamics with popular piecewise deterministic Markov process, namely bouncy particle dynamics, resulting in a controlled PDMP whose law follows the Gibbs path. Although our focus is optimization, it is worth noting that the controlled PDMPs can also be used for sampling and Bayesian inference analogous to how controlled Langevin has been used in [27, 28].

We also note that there is an equivalence between minimal energy steering between two known marginals and the Schrödinger bridge problem, where the latter is to estimate the most likely evolution between these two marginals [32]. In order to sample from these bridges, ideas about steering (guiding) have been used in [33–35]. Here, we show that there exists a minimal energy steering that allows us to follow a prescribed curve of measures, *i.e.*, that allows us to follow given marginals for all time points.

As a final observation here, we note that our view on simulated annealing is reminiscent of interior points methods in optimization, following a barrier trajectory [36], [7], Chapter 14, [37], Chapter 11, [38], Part 3. To this end, note that we can identify the finite-dimensional, potentially non-convex, optimization problem as an infinite-dimensional linear optimization problem over the set of probability measures. For a continuously differentiable function U with a compact set of global minimizers (see Asm. 3.1), we have that

$$\min_{x\in\mathbb{R}^d} U(x) = \min_{\rho\in\mathscr{P}} \int U(x)\,\rho(dx),$$

where $\mathscr{P}$ denotes the set of probability measures on $\mathbb{R}^d$. The minimum at the right-hand side is attained in, *e.g.*, $\delta_{x^*}(dx)$, where $\delta_x(dx)$ denotes the Dirac measure at x and where x^* is any global minimizer of U. Now, let Π be the set of measures with strictly positive Lebesgue probability densities on $\mathbb{R}^d$, that is

$$\Pi := \{\rho \in \mathscr{P}(\mathbb{R}^d)\colon \rho(dx) = \pi(x)\,dx,\ \pi(x) > 0\ \forall x \in \mathbb{R}^d\}.$$

Note that any minimizer of the form $\delta_{x^*}(dx)$ is in the (weak) closure of Π, and therefore, with a slight abuse of notation, we have that

$$\min_{\rho\in\mathscr{P}} \int U(x)\,\rho(dx) = \inf_{\rho\in\Pi} \int U(x)\,\rho(dx) = \inf_{\pi\in\Pi} \int U(x)\pi(x)\,dx$$

Indeed, by introducing a regularization term parametrized by β we can consider the problem

$$\min_{\pi\in\Pi} \int U(x)\pi(x)\,dx + \frac{1}{\beta}\int \pi(x)\log\pi(x)\,dx,$$

which has the solution $\pi_\beta \propto \exp(-\beta U) \in \Pi$. While Π is not the interior of $\mathscr{P}$, following the Gibbs curve $t \mapsto \pi_t$ can still be seen as an analogue to following the barrier trajectory.

1.2. Outline

The outline of the paper is as follows: in Section 2 we introduce notation and give a brief outline of relevant results from the literature, in particular the metric side of the spaces of probability measures and curves in

such spaces. In Section 3, we show that the Gibbs curve $t \mapsto \mu_t$ is absolutely continuous, by establishing the existence of a velocity field v_t which solves (1.6). Using this velocity field, we then proceed in Section 4 to show that there exists both a diffusion process and a piecewise deterministic Markov process that at any time t have a law agreeing with μ_t, hence giving a definite answer to (1.5). In Section 5, we leverage these results to design interacting particle–based optimization algorithms. To do so, we show a fairly general convergence result for transport map estimation relying on self-normalized reweighting, and then estimate the velocity field by solving discrete optimal transport problems. Finally, in Section 6 we illustrate and evaluate our methods on a number of test problems. For improved readability, the paper also contains Appendix A to which we defer some of our proofs.

2. Background

In this section we introduce background material needed for the rest of the paper. For ease of reference, we first set up the notation used throughout.

2.1. Notation

For density functions f, g, $f \propto g$ means that there exists a constant $a > 0$ such that $f = ag$. We use dx as shorthand notation for the standard Lebesgue measure on $\mathbb{R}^d$, and read $\mu(dx) = \pi(x)\,dx$ as the measure μ with Radon–Nikodym derivative, or density, π with respect to the Lebesgue measure dx. For measures μ, ν then $\mu \ll \nu$ if and only if μ is dominated by ν. We denote by $|\cdot|$ the Euclidean norm, by $\langle \cdot, \cdot \rangle$ the Euclidean inner product, and by $\|\cdot\|_{\mathrm{Fr}}$ the Frobenius norm of a matrix.

In the following, we need a number of different spaces. By $C(E, Y)$ we denote the set of Y-valued continuous functions on E. When $Y = \mathbb{R}$ we write $C(E)$, and by $C^k(E)$ we denote the set of $\mathbb{R}$-valued k-times continuously differentiable functions on E. $C_c^k(E)$ is the set of compactly supported $f \in C^k(E)$. By $\mathscr{P}(\mathbb{R}^d)$ we denote the probability measures on $\mathbb{R}^d$, and for $\mu \in \mathscr{P}(\mathbb{R}^d)$, we write with a slight abuse of notation $L^p(\mu) = L^p(\mu; \mathbb{R}^d, \mathbb{R}^\ell)$ for the set of functions $f\colon \mathbb{R}^d \to \mathbb{R}^\ell$ such that $\|f\|_{L^p(\mu)} := \int_{\mathbb{R}^d} |f|^p \, d\mu < \infty$, with $\ell = 1$ or d depending on the context. $\mathscr{P}_p(E)$ is the set of probability measures on the metric space (E, d) with bounded p-th moments, *i.e.*, $\mu \in \mathscr{P}_p(E)$ if and only if there exists a point $x_0 \in E$ such that $\int d(x, x_0)^p \mu(dx) < \infty$. For a measure μ or a function f, we denote by $\mathrm{supp}(\mu)$, $\mathrm{supp}(f)$ the support of μ and f, respectively.

We define the (L^2-)Poincaré constant C_P of $\mu \in \mathscr{P}(\mathbb{R}^d)$ as

$$C_P(\mu) := \inf \left\{ C \geqslant 0 \colon \|f\|_{L^2(\mu)} \leqslant C \|\nabla f\|_{L^2(\mu)}, \quad \forall f \in C_c^\infty(\mathbb{R}^d) \right\}.$$

A Borel probability measure $\mu \in \mathscr{P}(\mathbb{R}^d)$ with a finite Poincaré constant, $C_P(\mu) < \infty$, defines a norm on the set of smooth compactly supported functions $C_c^\infty(\mathbb{R}^d)$ by the pairing

$$\langle f, g \rangle_\mu = \int f g \, d\mu + \int \langle \nabla f, \nabla g \rangle \, d\mu.$$

Denote by $H^1(\mu)$ the completion of $C_c^\infty(\mathbb{R}^d)$ with respect to the norm induced by $\langle \cdot, \cdot \rangle_\mu$. By the Meyers-Serrin theorem (see [39], Chap. 7), $H^1(\mu)$ is the space of L^2-functions with weak derivatives in $L^2(\mu)$, that is,

$$H^1(\mu) = \{ f \in L^2(\mu; \mathbb{R}^d, \mathbb{R}) \colon \nabla f \in L^2(\mu; \mathbb{R}^d, \mathbb{R}^d) \}.$$

We say that a sequence of probability measures (μ_n) converges weakly to a probability measure μ iff

$$\lim_{n \to \infty} \int f \, d\mu_n \to \int f \, d\mu$$

for every continuous bounded function f. For a probability measure μ and $f, g \in L^2(\mu)$ we write $\mathrm{Cov}_\mu(f,g) := \int (f - \int f\, d\mu)(g - \int g\, d\mu)\, d\mu$ and $\mathrm{Var}_\mu(f) := \mathrm{Cov}_\mu(f,f)$.

2.2. Optimal transport and curves in the metric space $(\mathscr{P}_p(\mathbb{R}^d), \mathcal{W}_p)$

Here we state definitions and results related to the metric side of spaces of probability measures, which are central to our arguments. We follow the comprehensive treatment of [14]. The space of probability measures $\mathscr{P}_p(\mathbb{R}^d)$ for $p \geqslant 1$ becomes a metric space when equipped with the p-th Wasserstein distance $\mathcal{W}_p$, defined *via* the optimal transport problem

$$\mathcal{W}_p^p(\mu,\nu) := \inf_{\substack{\gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) \\ \int \gamma(\cdot, dx) = \mu \\ \int \gamma(dx, \cdot) = \nu}} \mathbb{E}_{(X,Y)\sim\gamma}\left[|X - Y|^p\right]. \tag{2.1}$$

The set $\Gamma(\mu,\nu) := \{\gamma \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d) \colon \int \gamma(\cdot, dx) = \mu, \int \gamma(dx, \cdot) = \nu\}$ is the admissible couplings, or transport plans, between μ and ν. Intuitively, a joint distribution $\gamma^* \in \Gamma(\mu,\nu)$ achieving the infimum in (2.1) describes a minimal cost deformation of the probability mass distributed according to μ into ν. Endowing $\mathscr{P}_p(\mathbb{R}^d)$ with $\mathcal{W}_p$ yields a (complete and separable) metric space [40], Theorem 6.18.

If the measure we transport from has a Lebesgue density, then the optimal transport plan γ^* takes the form of a deterministic transport map T.

Proposition 2.1 (Existence of Monge transport map, [41], Thms. 2.44 and 2.50)**.** *Let $\mu, \nu \in \mathscr{P}_p(\mathbb{R}^d)$. If μ has a Lebesgue density then there exists a Monge map solving (2.1), that is, a measurable map $T \colon \mathbb{R}^d \to \mathbb{R}^d$ such that $\nu = T_{\#}\mu$ and*

$$\mathcal{W}_p(\mu,\nu) = \left(\int |x - T(x)|^p\, \mu(dx)\right)^{\frac{1}{p}}.$$

This corresponds to the deterministic optimal coupling γ^*, given by

$$\gamma^* = (\mathrm{Id}, T)_{\#}\mu.$$

Further, when $p > 1$, T is μ-almost everywhere unique and we denote by $T_{\mu\to\nu}$ the optimal transport map from μ to ν.

Definition 2.2 (Absolutely continuous curves and metric derivatives, [14], Sect. 1.1)**.** We call a curve[1] $\eta \colon I \to \mathscr{P}_p(\mathbb{R}^d)$, $t \mapsto \mu_t$ *absolutely continuous* if the *metric derivative* $|\mu'|(t)$, defined as the limit

$$|\mu'|(t) := \lim_{h\to 0} \frac{\mathcal{W}_p(\mu_t, \mu_{t+h})}{h},$$

exists for Lebesgue almost every t and belongs to $L^1(I)$.

We can now make the continuity equation (1.6) formal.

Theorem 2.3 (Absolutely continuous curves and velocity fields, [14], Thm. 8.3.1)**.** *Let $I \mapsto \mu_t$ be an absolutely continuous curve in $(\mathscr{P}_p(\mathbb{R}^d), \mathcal{W}_p)$. Then there exists a time-dependent velocity field $v_t(x) \colon I \times \mathbb{R}^d \to \mathbb{R}^d$ satisfying the continuity equation*

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{2.2}$$

[1] A curve is a continuous map of an interval $I \subset \mathbb{R}$ into a topological space.

in the sense that for all compactly supported smooth functions $f \in C_c^\infty(I \times \mathbb{R}^d)$

$$\int_I \int_{\mathbb{R}^d} (\partial_t f + \langle v_t, \nabla_x f \rangle)\, d\mu_t\, dt = 0. \tag{2.3}$$

Further, it holds that $\|v_t\|_{L^p(\mu_t, \mathbb{R}^d)} \in L^1(I)$.

Conversely, assume that there exists a v_t *that solves* (2.2) *and such that* $\|v_t\|_{L^p(\mu_t, \mathbb{R}^d)} \in L^1(I)$. *Then* $t \mapsto \mu_t$ *is an absolutely continuous curve.*

Remark 2.4. Absolute continuity for a curve defined on an interval $[0, T]$ coincides with a finite curve length. More precisely, the length ℓ of a curve $t \mapsto \mu_t$ is defined by

$$\ell(t \mapsto \mu_t) := \int_0^T |\mu'|(t)\, dt = \inf_v \quad \int_0^T \|v_t\|_{L^2(\mu_t)}\, dt, \tag{2.4}$$

$$\text{s.t.} \quad \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0, \tag{2.5}$$

using the convention that the infimum is infinite if there are no velocity fields v satisfying the continuity equation.

For absolutely continuous curves of measures $t \mapsto \mu_t$ with Lebesgue densities, the existence of Monge transport maps in Proposition 2.1 gives another characterization of the velocity fields in Theorem 2.3.

Proposition 2.5 (Transport map characterization of v_t, [14], Prop. 8.4.6)**.** *Let* $t \mapsto \mu_t \in (\mathscr{P}_p(\mathbb{R}^d), \mathcal{W}_p)$ *be an absolutely continuous curve, with* $1 < p < \infty$. *If* μ_t *has a Lebesgue density for all* t, *then*

$$v_t = \lim_{h \to 0} h^{-1}(T_{\mu_t \to \mu_{t+h}} - \mathrm{Id}) \quad \text{in } L^2(\mu_t),$$

where $T_{\mu_t \to \mu_{t+h}}$ *is the unique Monge transport map from* μ_t *to* μ_{t+h}.

2.3. Simulated annealing on $\mathbb{R}^d$

The stochastic differential equation

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2\beta^{-1}}\, dB_t,$$

is referred to as the (overdamped) Langevin dynamics. Here, B_t is a Brownian motion on $\mathbb{R}^d$. This stochastic process has generator

$$\mathcal{L}_\beta f = \frac{1}{\beta} \Delta f - \langle \nabla f, \nabla U \rangle. \tag{2.6}$$

Under suitable conditions on U, this process has the Gibbs measure $\mu_\beta = Z_\beta^{-1} \mathrm{e}^{-\beta U(x)}\, dx$ as its unique invariant measure. Simulated annealing on $\mathbb{R}^d$ [2, 3, 5] is the stochastic process with time-dependent diffusion coefficient given by the Langevin-like SDE

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2\beta(t)^{-1}}\, dB_t. \tag{2.7}$$

For each potential U belonging to a fairly large class, there exists a constant $C_U > 0$ such that if $\beta(t) \leqslant C_U \log(1 + t)$, then, for any initial distribution, the solution of (2.7) converges weakly to μ_∞ as $t \to \infty$ [5].

Conversely, one can consider the rate at which the Gibbs curve itself approaches its limiting measure μ_∞. Assuming that U admits a unique global minimum x^* with a positive definite Hessian $\nabla^2 U(x^*)$, then [10],

Theorem 3.8 gives the following concentration rate for the Gibbs measure

$$\mathcal{W}_p(\mu_t, \mu_\infty) \leqslant K\beta(t)^{-1/2}, \tag{2.8}$$

for some positive constant K.

3. A velocity field solution to the continuity equation

In this section, we prove that the Gibbs curve, as a map from a finite interval $[0,T]$ to $\mathscr{P}_2(\mathbb{R}^d)$, is absolutely continuous. For this, we make the following assumptions on U.

Assumption 3.1. U is of class C^1 and satisfies $\inf_x U(x) > -\infty$. Moreover, $\{x \in \mathbb{R}^d \mid x = \inf_x U(x)\}$ is non-empty and compact. Without loss of generality, we also assume that $\inf_x U(x) = \min_x U(x) = 0$.

Assumption 3.2. There exist $\alpha > 0$, $R > 0$ such that $|x| > R \implies U(x) \geqslant \alpha|x|^2$.

Assumption 3.3. There exist $\alpha > 0, R > 0$ such that $|x| > R \implies \langle \nabla U(x), x\rangle \geqslant \alpha|x|^2$.

Given a potential U satisfying the above assumptions and a suitable cooling schedule β, we from this point on reserve the notation μ_t for the Gibbs curve, and π_t for the corresponding densities, that is,

$$\mu_t(dx) := \pi_t(x)\,dx, \tag{3.1}$$

$$\pi_t(x) := \frac{\mathrm{e}^{-\beta(t)U(x)}}{\int_{\mathbb{R}^d} \mathrm{e}^{-\beta(t)U(y)}\,dy}. \tag{3.2}$$

Remark 3.4. Note that, under Assumptions 3.1–3.3, μ_t concentrates on the global minimizers of U, see [9].

We state now the main result of this section.

Theorem 3.5. *Assume that $U \in C^1(\mathbb{R}^d, \mathbb{R})$ satisfies Assumptions 3.1–3.3. Let $\beta \in C^1([0,T_1), [1,\infty))$ be a differentiable cooling schedule. Then, for every $T \leqslant T_1$ such that $\lim_{t\uparrow T} \beta(t) < \infty$, the curve*

$$\begin{aligned} \eta\colon [0,T) &\to \mathscr{P}_2(\mathbb{R}^d) \\ t &\mapsto \mu_t \end{aligned}$$

is absolutely continuous. Further, for every $t \in [0,T)$ there exists a unique velocity field v_t of minimal $L^2(\mu_t)$-norm such that μ_t and v_t satisfies (2.2).

We first approach this theorem by an illustrative example, namely the case when U is a Gaussian potential, in the following section. After this, in Section 3.2, we turn our attention to the general statement of the theorem.

3.1. The velocity field in the Gaussian case

When U is a Gaussian potential the analysis is greatly simplified by the existence of explicit solutions to the intermediate transport problems. Let $U(x) = \frac{1}{2}x^T\Sigma^{-1}x$, with Σ positive definite, and let $\beta\colon [0,T) \to [1,\infty)$ be a smooth non-decreasing cooling schedule. Then, $t \mapsto \mu_t = \mathcal{N}(0, \beta(t)^{-1}\Sigma)$ is a curve of centred Gaussian measures in $(\mathscr{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ such that the variance tends to 0 as $\beta \to \infty$. For any two Gaussian distributions $\nu_i = \mathcal{N}(m_i, \Sigma_i)$, $i = 1, 2$, we have that

$$\mathcal{W}_2^2(\nu_1,\nu_2) = |m_1 - m_2|^2 + \operatorname{Trace}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right),$$

see, *e.g.*, [42–45]. Additionally, when $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$, we have

$$\mathcal{W}_2^2(\nu_1,\nu_2) = |m_1 - m_2|^2 + \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_{\mathrm{Fr}}^2,$$

and therefore

$$\begin{aligned}\mathcal{W}_2^2(\mu_t,\mu_{t+h}) &= \left\|(\beta(t)^{-1/2} - \beta(t+h)^{-1/2})\Sigma^{1/2}\right\|_{\mathrm{Fr}}^2 \\ &= \left|\beta(t)^{-1/2} - \beta(t+h)^{-1/2}\right|^2 \left\|\Sigma^{1/2}\right\|_{\mathrm{Fr}}^2.\end{aligned}$$

We verify that this is an absolutely continuous curve by verifying that the metric derivative $|\mu'|(t)$ is in $L^1([0,T])$. As β is differentiable and non-decreasing, we find that

$$\begin{aligned}|\mu'|(t) &= \lim_{h\downarrow 0} h^{-1}\mathcal{W}_2(\mu_t,\mu_{t+h}) \\ &= \lim_{h\downarrow 0} \|\Sigma^{1/2}\|_{\mathrm{Fr}}\, h^{-1}(\beta(t)^{-1/2} - \beta(t+h)^{-1/2}) \\ &= -\frac{d}{dt}(\|\Sigma^{1/2}\|_{\mathrm{Fr}}\,\beta(t)^{-1/2}) \\ &= \frac{\|\Sigma^{1/2}\|_{\mathrm{Fr}}}{2}\beta'(t)\beta(t)^{-3/2},\end{aligned}$$

which clearly exists for all t. Hence

$$\begin{aligned}\int_0^T |\mu'|(t)\,dt &= \int_0^T -\frac{d}{dt}(\|\Sigma^{1/2}\|_{\mathrm{Fr}}\,\beta(t)^{-1/2})\,dt \\ &= \|\Sigma^{1/2}\|_{\mathrm{Fr}}\,(\beta(0)^{-1/2} - \beta(T)^{-1/2}) < \infty,\end{aligned}$$

as long as $\beta(0) > 0$. We note that this is finite even when $\lim_{t\to T}\beta(t) = \infty$, and also in the case when $T \to \infty$.

The Monge map T between ν_1 and ν_2 is given by [45], Remark 2.31

$$T(x) = m_2 + \Sigma_1^{-1/2}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\Sigma_1^{-1/2}(x - m_1),$$

which in the commuting case $\Sigma_1\Sigma_2 = \Sigma_2\Sigma_1$ similarly reduces to $T(x) = m_2 + \Sigma_2^{1/2}\Sigma_1^{-1/2}(x - m_1)$. We find that $T_{\mu_t\to\mu_{t+h}}(x) = \frac{\beta(t+h)^{1/2}}{\beta(t)^{1/2}}x$, independently of Σ. We find by an application of Proposition 2.5 that

$$\begin{aligned}v_t(x) &= \lim_{h\to 0}\frac{T_{\mu_t\to\mu_{t+h}}(x) - x}{h} \\ &= \lim_{h\to 0}\frac{1}{h}\left(\frac{\beta(t+h)^{1/2}}{\beta(t)^{1/2}} - 1\right)x \\ &= -\frac{1}{2}\frac{\beta'(t)}{\beta(t)}x.\end{aligned}$$

Remark 3.6. We note that v_t above does not depend on the covariance Σ. Further, when $\beta'(t)/\beta(t)$ is constant, *i.e.*, $\beta(t)\propto \mathrm{e}^{ct}$ then

$$v_t(x) = -cx =: \nabla f(x).$$

Since this is time-homogeneous, the exponentially cooling Gaussian curve in $\mathscr{P}_2$ corresponds to the gradient flow of the convex functional $\mathcal{F}(\mu) = \int f(x)\,d\mu$ with $f(x) = -c|x|^2/2$, see for instance [46], Example 2.1.

Remark 3.7 (Unit speed parametrization)**.** Any absolutely continuous curve admits a unit speed parametrisation [14], Lemma 1.1.4. Finding the unit speed curve amounts to solving $|\mu'|(t) = 1$, which in this case reduces to finding a cooling schedule which satisfies

$$\beta'(t) = 2a^{-1}\beta(t)^{3/2}, \quad \beta(0) = 1,$$

with $a := \|\Sigma^{1/2}\|_{\mathrm{Fr}}$. The solution to this differential equation is $\beta(t) = a^2(a-t)^{-2}$, and the curve reaches the endpoint δ_0 at $t = a = \|\Sigma^{1/2}\|_{\mathrm{Fr}}$, *i.e.*, in finite time.

3.2. The velocity field in the general case

This section is devoted to proving Theorem 3.5, *i.e.*, showing that the Gibbs curve $t \mapsto \mu_t \in (\mathscr{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is absolutely continuous. To this end, we first note that due to the Gaussian tails, $\mu_t \in \mathscr{P}_p$ for any p.

Lemma 3.8. *Assume that U satisfies Assumptions 3.1 and 3.2. Then, for all $\beta \in (0,\infty)$ and $p \geqslant 1$, the Gibbs measure satisfies $\mu_\beta \in \mathscr{P}_p(\mathbb{R}^d)$.*

Proof. Let R be as in (3.2). Then $Z_\beta = \int_{\mathbb{R}^d} \mathrm{e}^{-\beta U(x)}\,dx \leqslant D + \int_{|x|>R} \mathrm{e}^{-\beta|x|^2}\,dx < \infty$ for some constant D and hence $\mu_\beta(dx) = Z_\beta^{-1}\mathrm{e}^{-\beta U(x)}\,dx$ is a well-defined probability measure. Similarly,

$$\int_{\mathbb{R}^d} |x|^p\,\mu_\beta(dx) \leqslant R^p + \frac{1}{Z_\beta}\int_{|x|>R} |x|^p \mathrm{e}^{-\beta|x|^2}\,dx < \infty. \qquad \square$$

The following lemma establishes a bound on the Poincaré constant of μ_t. We give a proof of this by constructing a Lyapunov function and relying on a result in [47]. For details, see Appendix A.

Lemma 3.9. *Let U satisfy Assumptions 3.1–3.3, and let $\beta\colon [0,T] \to (0,\infty)$. Then, for every t, μ_t satisfies a Poincaré inequality:*

$$\mathrm{Var}_{\mu_t}(f) = \int \left(f - \int f\,d\mu_t\right)^2 d\mu_t \leqslant C_P(\mu_t)\int |\nabla f|^2\,d\mu_t,$$

for all $f \in C_c^\infty(\mathbb{R}^d)$. Further, there exist positive constants A, B, K such that, for all $t \in [0,T]$

$$C_P(\mu_t) \leqslant C_t := A(1 + B\mathrm{e}^{\beta(t)K}).$$

Since the Poincaré constant $C_P(\mu_t)$ is bounded by the preceding lemma, the Sobolev space $H^1_t := H^1(\mu_t)$ is well-defined. Next, note that for the Gibbs curve, the continuity equation for the density π_t reads

$$\partial_t \pi_t(x) + \nabla\cdot(v_t(x)\pi_t(x)) = 0.$$

Differentiating π_t with respect to t, we find that

$$\partial_t \pi_t(x) = -\beta'(t)\left(U(x) - \int U\,d\mu_t\right)\pi_t(x). \tag{3.3}$$

As a brief aside, we note that $\int f(x)\,\partial_t\pi_t(x)\,dx = -\beta'(t)\,\mathrm{Cov}_{\mu_t}(f,U)$, and hence we can understand $\partial_t\mu_t$ as measuring the covariance of a function f against the potential U. By (3.3) we arrive at the following

equation for v_t

$$\nabla \cdot (v_t(x)\pi_t(x)) = \beta'(t)\left(U(x) - \int U\, d\mu_t\right)\pi_t(x). \tag{3.4}$$

In the proof of the following theorem, we will look for a solution to (3.4) in the space of gradients of functions in H^1. To this end, as μ_t satisfies a Poincaré inequality the bilinear map $\langle\cdot,\cdot\rangle_{\nabla\mu_t}\colon H^1_t \times H^1_t \to \mathbb{R}$, given by $\langle\cdot,\cdot\rangle_{\nabla\mu_t} := \langle\nabla\cdot,\nabla\cdot\rangle_{\mu_t}$, defines an inner product on the subspace of zero mean functions

$$\dot{H}^1_t = \left\{ f \in H^1_t \colon \int f\, d\mu_t = 0 \right\}.$$

To see this, first note that

$$\langle f, g\rangle_{\nabla\mu_t} = \int \langle \nabla f, \nabla g\rangle\, d\mu_t$$

is well-defined by Cauchy–Schwarz. Moreover, we have that $\langle f, f\rangle_{\nabla\mu_t} \geqslant 0$, and by the Poincaré inequality, equality holds if and only if $f = 0$. We can now prove Theorem 3.5 in the following form:

Theorem 3.10. *Assume that $\beta\colon [0,T) \to [1,\infty)$ is differentiable and that U satisfies Assumptions 3.1,–3.3. Then, for each $t \in [0,T)$ there exists a unique $h_t \in \dot{H}^1_t$ such that the velocity field $v_t = \nabla h_t \in L^2(\mu_t, \mathbb{R}^d, \mathbb{R}^d)$ solves the continuity equation*

$$-\nabla \cdot (v_t\mu_t) = \partial_t \mu_t$$

in the sense that

$$\int \langle v_t(x), \nabla f(x)\rangle\, \pi_t(x)\, dx = \int f(x)\, \partial_t \pi_t(x)\, dx, \quad \forall f \in C^\infty_c(\mathbb{R}^d). \tag{3.5}$$

This v_t is the vector field with minimal $\|\cdot\|_{L^2(\mu_t)}$-norm of all vector fields that satisfies (3.5). Further, $\int_0^T \|v_t\|_{L^2(\mu_t)}\, dt < \infty$, and hence the Gibbs curve is absolutely continuous.

Proof. Consider the linear functional $\phi\colon \dot{H}^1_t \to \mathbb{R}$ given by

$$\phi(f) = \int f(x)\, \partial_t \pi_t(x)\, dx = -\beta'(t) \int f(x) \left(U(x) - \int U \mathrm{d}\mu_t\right) \pi_t(x)\, dx. \tag{3.6}$$

By applying Hölder's inequality (Cauchy–Schwarz) we find that

$$\begin{aligned} |\phi(f)| &\leqslant |\beta'(t)| \left(\int |f(x)|^2 \pi_t(x)\, dx\right)^{1/2} \left(\int \left(U(x) - \int U\, d\mu_t\right)^2 \pi_t(x)\, dx\right)^{1/2} \\ &= |\beta'(t)| \sqrt{\mathrm{Var}_{\mu_t} f}\, \sqrt{\mathrm{Var}_{\mu_t} U}, \end{aligned}$$

In conjunction with the bound C_t on the Poincaré constant of μ_t in Lemma A.1 we find then the following estimate for the dual norm of ϕ:

$$\|\phi\|_{(\dot{H}^1_t)^*} = \sup_{f \in \dot{H}^1_t} \frac{|\phi(f)|}{\sqrt{\langle \nabla f, \nabla f\rangle_{\mu_t}}}$$

$$\begin{aligned}
&\leqslant \sup_{f\in\dot{H}^1_t} \frac{|\beta'(t)|\sqrt{\mathrm{Var}_{\mu_t} f}\sqrt{\mathrm{Var}_{\mu_t} U}}{\sqrt{\langle \nabla f, \nabla f\rangle_{\mu_t}}} \\
&= |\beta'(t)|\sqrt{C_P(\mu_t)}\sqrt{\mathrm{Var}_{\mu_t} U} \\
&\leqslant |\beta'(t)|\sqrt{C_t}\sqrt{\mathrm{Var}_{\mu_t} U}.
\end{aligned}$$

Since $\sqrt{\mathrm{Var}_{\mu_t} U}$ is bounded (Lem. A.2), the linear functional ϕ is continuous. By the Riesz representation theorem [48], p. 109 there exists a unique element h_t in the Hilbert space $(\dot{H}^1_t, \langle\cdot,\cdot\rangle_{\nabla\mu_t})$ such that

$$\phi(f) = \langle h_t, f\rangle_{\nabla\mu_t} \quad \text{and} \quad \|\phi\|_{(\dot{H}^1_t)*} = \|h_t\|_{\dot{H}^1_t}.$$

From the first of these equations we find in particular that, for all $f \in C_c^\infty \subset \dot{H}^1_t$,

$$\int f(x)\,\partial_t\pi_t(x)\,dx = \int \langle \nabla h_t(x), \nabla f(x)\rangle\,\pi_t(x)\,dx,$$

so $v_t := \nabla h_t$ solves the continuity equation in the prescribed sense, (3.5). Further,

$$\|v_t\|_{L^2(\mu_t)} = \|\nabla h_t\|_{L^2(\mu_t)} = \|h_t\|_{\dot{H}^1_t} = \|\phi\|_{(\dot{H}^1_t)*} \leqslant |\beta'(t)|\sqrt{C_t}\sqrt{\mathrm{Var}_{\mu_t} U}.$$

Since $\sup_{t\in[0,T]} |\beta'(t)|\sqrt{C_t}\sqrt{\mathrm{Var}_{\mu_t} U} < \infty$, $\|v_t\|_{L^2(\mu_t)} \in L^1([0,T])$ and by [14], Theorem 8.3.1, phrased as Theorem 2.3, the Gibbs curve is absolutely continuous.

Finally, the fact that the solution v_t is the element with minimal $\|\cdot\|_{L^2(\mu_t)}$-norm of all solutions to (3.5) follows from [14], Proposition 8.4.5, by noting that

$$v_t = \nabla h_t \in \overline{\{\nabla\varphi\colon \varphi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu_t)},$$

where $\overline{}^{L^2(\mu_t)}$ denotes the closure in the $L^2(\mu_t)$-norm. □

Remark 3.11. In the preceding proof, it is interesting to note that the ansatz $v_t = \nabla g_t$, for some $g_t \in H^1_t$, turns the continuity equation (3.4) into an elliptic PDE. To this end, let $\mathcal{L}_t$ be the operator defined by

$$\mathcal{L}_t f := \frac{1}{\beta(t)}\Delta f - \langle \nabla U, \nabla f\rangle.$$

As $\pi_t(x) > 0$ for all x, this allows us to rewrite (3.4) as

$$\mathcal{L}_t g_t(x) = \frac{\beta'(t)}{\beta(t)}\left(U(x) - \int U\,d\mu_t\right) =: \Psi_t(x). \tag{3.7}$$

Assuming that the preceding equation admits a sufficiently regular (classical) solution g_t, we see that this solution is only determined up to an additive constant. We then recover h_t as in Theorem 3.10 by noting that $h_t = g_t - \int g_t\,d\mu_t$, or, equivalently, by looking for a solution to (3.7) in $H^1_t/\mathbb{R} \simeq \dot{H}^1_t$. Moreover, Itô's formula yields a probabilistic representation of g_t by

$$g_t(x) = \int g_t\,d\mu_t - \lim_{S\to\infty} \mathbb{E}\left[\int_0^S \Psi_t(Y_s^x)\,ds\right], \tag{3.8}$$

where Y_s^x is a solution to the overdamped Langevin SDE at a fixed temperature $\beta(t)$ starting at x:

$$dY_s^x = -\nabla U(Y_s^x)\,ds + \sqrt{\frac{2}{\beta(t)}}\,dB_s, \quad Y_0^x = x.$$

Indeed, an application of Itô's formula yields that

$$g_t(Y_S^x) = g_t(x) + \int_0^S \mathcal{L}_t g_t(Y_s^x)\,ds + \frac{2}{\beta(t)}\int_0^S \nabla \cdot g_t(Y_s^x)\,dB_s.$$

We then find this probabilistic interpretation by first taking the expected value of both sides

$$\mathbb{E}\left[g_t(Y_S^x)\right] = g_t(x) + \mathbb{E}\left[\int_0^S \mathcal{L}_t g_t(Y_s^x)\,ds\right],$$

and then noting that, under Assumptions 3.1–3.3, the Markov process Y_s^x is *geometrically ergodic*, and specifically it holds that $\mathbb{E}\left[g_t(Y_S^x)\right] \to \int g_t\,d\mu_t$ as $S \to \infty$, for suitably regular g_t. Clearly, $\int \Psi_t\,d\mu_t = 0$, and again, by the geometric ergodicity of Y_s^x we find that $|\mathbb{E}\left[\Psi_t(Y_s^x)\right]| \leqslant c_1 \mathrm{e}^{-c_2 s}$, for some $c_1, c_2 > 0$ independent of s. Hence, as the right-hand side of (3.8) is finite, the claim follows.

3.3. On finiteness of $\beta(T)$ and an exponentially bad Poincaré constant

We elaborate here on the restriction of finite end inverse temperatures $\beta(T)$. First, we consider the strength of the previous results in the Gaussian setting of Section 3.1.

Example 3.12. Picking again $U(x) = |x|^2/2$ so that $\mu_t = \mathcal{N}(0, \beta(t)^{-1} I_d)$. Assuming that $\beta' \geqslant 0$, we now investigate the curve length bound

$$\ell \leqslant \int_0^T \left(\beta'(t)\sqrt{C_t}\sqrt{\mathrm{Var}_{\mu_t}(U)}\right) dt.$$

A straightforward computation yields

$$\begin{aligned}\mathrm{Var}_{\mu_t}(U) &= \mathrm{Var}_{\mu_t}(|x|^2/2)\\ &= \frac{1}{4}\mathrm{Var}\left(\sum_{i=1}^d \beta(t)^{-1} Y_i^2\right)\\ &= \frac{1}{4}\frac{1}{\beta(t)^2} 2d,\end{aligned}$$

where we denoted by Y_i independent standard Gaussian variables on $\mathbb{R}$. As $\sqrt{\mathrm{Var}_{\mu_t}(|x|^2/2)} = \sqrt{\frac{d}{2}}\frac{1}{\beta(t)}$ we have that

$$\ell \leqslant \frac{d}{2}\int_0^T \left(\frac{\beta'(t)}{\beta(t)}\sqrt{C_t}\right) dt,$$

from which we can see that if β diverges, one needs additional control on the Poincaré constant to ensure that this bound guarantees integrability. In this case, it is known that the optimal Poincaré constant is $C_P \propto \frac{1}{\sqrt{\beta}}$,

which yields that, for some constant $C > 0$,

$$\ell \leqslant C \int_0^T \beta'(t)\beta(t)^{-5/4}\, dt = 4C(\beta(0)^{-1/4} - \beta(T)^{-1/4}),$$

which converges even if $\lim_{t\to T} \beta(t) = \infty$.

As we just saw in the Gaussian case, the curve has finite length even as $\beta \to \infty$. This would in general be desirable; however, we note that for potentials satisfying Assumptions 3.1–3.3, the current proof technique cannot show this. In the proof of Theorem 3.10, we bound the curve length using the Poincaré inequality in conjunction with the exponential bound on the Poincaré constant in Lemma A.1. Moreover, it is in fact known that for a large class of objective functions U the bound on the Poincaré constant is not overly pessimistic, *i.e.*, the Poincaré constant does get exponentially bad [49]. Next, we present one simple example that illustrates this.

Example 3.13 (An exponentially bad Poincaré constant for a potential with a single global minima)**.** We now give an example of a potential U on $\mathbb{R}$ which illustrates a Poincaré constant with an exponential dependence on β for the corresponding Gibbs measure. Assuming that U has one deep, global minimum at 0 and a shallower local minimum at x_1, separated by a low probability barrier region, we can consider a test function which changes rapidly at the barrier.

For concreteness, let $U(x) = x^2 - 5\cos \pi x + 5$. This potential has a global optimum at $x = 0$, where $U(0) = 0$, as well as a shallower local well at $x \approx 2$ separated by a low probability, barrier region at $x \approx 1$. We can find a lower bound on $C_P(\mu_\beta)$ using the piecewise linear function

$$f = \begin{cases} 1 & \text{if } x \leqslant a, \\ 1 - \frac{x-a}{a-b} & \text{if } a < x < b, \\ 0 & \text{if } x \geqslant b. \end{cases}$$

Choosing the transition region $a = 1 - \delta$ and $b = 1 + \delta$ for f, we have that f is approximately Bernoulli distributed with $\mathbb{P}(f = 1) \approx 1 - \varepsilon$, with $\varepsilon \approx c\mathrm{e}^{-\beta U(2) - \beta U(0)} = c\mathrm{e}^{-4\beta}$, so that $\mathrm{Var}_\mu(f) \approx \varepsilon(1-\varepsilon) \approx \varepsilon \approx c\mathrm{e}^{-4\beta}$. For the denominator, we have that

$$\int (f')^2\, d\mu = \frac{1}{4\delta^2} \int_{1-\delta}^{1+\delta} \frac{1}{Z_\beta} \mathrm{e}^{-\beta U(x)}\, dx = \frac{1}{2\delta Z_\beta} \mathrm{e}^{-\beta U(1+h\delta)} \approx \frac{1}{2\delta Z_\beta} \mathrm{e}^{-11\beta}.$$

Hence, we find that

$$\frac{\mathrm{Var}_\mu f}{\int (f')^2\, d\mu} \approx C\mathrm{e}^{7\beta} \lesssim C_P(\mu_\beta).$$

To conclude, we note that this choice of U satisfies Assumptions 3.1–3.3,

4. Controlled stochastic annealing *via* superposition

The goal of this section is to establish that there exists stochastic processes driven by both the velocity field v_t and additional stochastic dynamics. To motivate this one can first consider the case of a curve in $\mathbb{R}^d$.

Example 4.1. Given the continuously differentiable curve $t \mapsto x_t \in \mathbb{R}^d$, starting at x_0 and ending in x_T, dynamics for following this curve is

$$dy_t = \dot{x}_t\, dt, \quad y_0 = x_0,$$

where $\dot{x}_t$ denotes the tangent vector of the curve at x_t. We can model uncertainty in estimates of $\dot{x}_t$ by some error process ε_t, yielding

$$dy_t = (\dot{x}_t + \varepsilon_t)\, dt.$$

To compensate for this error one can add a correction term attracting the dynamics towards the curve by

$$dy_t = (\dot{x}_t + \varepsilon_t)\, dt + (x_t - y_t)\, dt. \tag{4.1}$$

One can view the previous correction term, $(x_t - y_t)$, as a gradient of the current deviation, $-\nabla_y \frac{1}{2}\|x_t - y\|^2$.

Motivated by this, we set up similar schemes for following the Gibbs curve. First, we consider a superimposed stochastic dynamics,

$$dX_t = \underbrace{v_t\, dt}_{\text{I.}} \underbrace{- \nabla U(X_t)\, dt + \sqrt{2\beta(t)^{-1}}\, dB_t}_{\text{II.}},$$

for which the curve of marginal distributions evolve in a fashion analagous to Example 4.1. Here, the second term is the simulated annealing dynamics, which in the space of measures corresponds to the gradient of the Kullback–Leibler divergence as corrector, see (1.7). Dropping the velocity field v_t recovers simulated annealing, and corresponds to completely relying on correction to follow the curve in the aforementioned example.

Remark 4.2. We note that in (4.1), it is possible to scale the corrector term by any positive factor, which can be understood as different time scales for the first term and the second term. Analogously, one can run the stochastic dynamics at a different time scale $\tau(t)$,

$$dX_t = v_t(X_t)\, dt - \tau'(t)\nabla U(X_t)\, dt + \sqrt{2\tau'(t)\beta(t)^{-1}}\, dB_t,$$

where $\tau'(t) = \frac{d}{dt}\tau(t)$. As we shall see in the following section, the law of X_t always agrees with the Gibbs measure μ_t. Moreover, by absolute continuity of the Gibbs curve, any two pairs β_1, τ_1 and β_2, τ_2 of sufficiently regular functions such that $\frac{\beta_1'(t)}{\tau_1'(t)} = \frac{\beta_2'(t)}{\tau_2'(t)}$ for all t describe the same SDE, *i.e.*, time can be rescaled in any of the two SDEs to obtain the other. This can be deduced from the Fokker–Planck equation (4.4) below.

Throughout this section we assume that U satisfies Assumptions 3.1, 3.2, and 3.3. In order to simplify the arguments, we assume that the velocity field v_t satisfies the following growth condition.

Assumption 4.3 (Growth condition on v_t)**.** The velocity field v_t is locally bounded, and there exists a constant C such that

$$\sup_{0 \leqslant t \leqslant T} \langle x, v_t(x) \rangle \leqslant C\left(1 + |x|^2\right), \quad \text{for all } x \in \mathbb{R}^d.$$

As we saw in Section 3.1, this condition trivially holds for velocity fields arising from Gaussian potentials, as in this case $\langle x, v_t(x) \rangle \leqslant 0$. Indeed, we can understand the preceding assumption as a bound on the outward-pointingness of v_t for large $|x|$. The following result gives a condition under which the growth assumption holds. We postpone the proof as well as further discussion to Section 4.4.

Proposition 4.4. *Let $U \in C^2(\mathbb{R}^d)$ be a twice continuously differentiable potential with bounded second derivatives, satisfying Assumptions 3.1 and 3.2. Denote by $\lambda_{\min}(A)$ the smallest eigenvalue of the matrix A. If*

$$\int \lambda_{\min}\left(\nabla^2 U(x)\right) \mu_t(dx) > 0, \quad \textit{for all } 0 \leqslant t \leqslant T, \tag{4.2}$$

then Assumption 4.3 holds.

4.1. Controlled simulated annealing with diffusive dynamics

We will begin by investigating the case where particles are driven by both the velocity field and the diffusive Langevin-like dynamics (1.1), that is, the SDE

$$dX_t = v_t(X_t)\,dt - \nabla U(X_t)\,dt + \sqrt{\frac{2}{\beta(t)}}\,dB_t, \quad X_0 \sim \mu_0. \tag{4.3}$$

We show that such a process exists, and further, that the law of this process does not diverge from the Gibbs curve μ_t. It is instructive to first sketch the argument, postponing some of the technicalities to later. If it exists, the evolution of the law ρ_t of X_t in (4.3) is given by the Fokker–Planck equation

$$\partial_t \rho_t + \nabla \cdot (v_t \rho_t) + \nabla \cdot (-\nabla U \rho_t) + \frac{1}{\beta(t)} \Delta \rho_t = 0, \quad \rho_0 = \mu_0, \tag{4.4}$$

and we want to show that the Gibbs curve μ_t is the unique solution. To see that this is the case, recall that v_t and μ_t jointly satisfy the continuity equation

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0.$$

The other two terms, $\nabla \cdot (-\nabla U \rho_t) + \frac{1}{\beta(t)} \Delta \rho_t$, correspond to the μ_t-stationary diffusive dynamics $-\nabla U(X_s)\,ds + \sqrt{2\beta(t)^{-1}}\,dB_s$ and hence

$$\nabla \cdot (-\nabla U \mu_t) + \frac{1}{\beta(t)} \Delta \mu_t = 0. \tag{4.5}$$

Putting this together we find that, at least in a formal sense, the Gibbs curve solves (4.4), that is,

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) + \nabla \cdot (-\nabla U \mu_t) + \frac{1}{\beta(t)} \Delta \mu_t = 0.$$

We justify the preceding computations *via* the following two propositions. First, by showing that the SDE (4.3) admits a solution, Proposition 4.5 justifies the Fokker–Planck equation (4.4). Then, in Proposition 4.6, we show that this solution is unique and hence necessarily agrees with the Gibbs curve.

Proposition 4.5. *Let v_t be as in Theorem 3.10 and assume that v_t satisfies Assumption 4.3. Then the stochastic differential equation (4.3) has a weak solution $(X_t)_{t \geqslant 0}$.*

Proof. Define formally the operator $\mathcal{A}_t^o$ acting on test functions f by

$$\mathcal{A}_t^o f = \langle v_t, f \rangle + \mathcal{L}_t f = \langle v_t, f \rangle + \langle -\nabla U, \nabla f \rangle + \frac{1}{\beta(t)} \Delta f.$$

The *martingale problem* for $\mathcal{A}_t^o$ is well-posed if there, for all x, t, t_0, with $0 \leqslant t_0 \leqslant t \leqslant T$, exists a unique probability measure P_x on the path space $C([0,T], \mathbb{R}^d)$ such that $P_x[X_s = x, 0 \leqslant s \leqslant t_0] = 1$ and

$$f(X_t) - f(x) - \int_{t_0}^{t} (\mathcal{A}_s^o f)(X_s)\,ds$$

is a martingale. Assumption 3.3 and Assumption 4.3 yield, respectively,

$$\langle -\nabla U(x), x\rangle \leqslant 0, \quad \text{and} \quad \langle v_t(x), x\rangle \leqslant C(1+|x|^2),$$

for sufficiently large $|x|$. Hence the drift $b(t,x) = -\nabla U(x) + v_t(x)$ satisfies the conditions of [11], Theorem 10.2.2, namely

$$\langle x, b(t,x)\rangle \leqslant C(1+|x|^2),$$

which guarantees the existence of such a P_x. Further, by [50], Theorem 32.7, this is a weak solution to (4.3) (see also [51], p. 170). □

The above statement tells us that, for any initial point $X_0 = x_0$ there is a stochastic process $(X_t)_{t\geqslant 0}$ solving the SDE (4.3). We now want to ensure that this has the prescribed time marginals. With $(\mathcal{A}_t^o)^*$ the adjoint of $\mathcal{A}_t^o$, we reformulate the Fokker–Planck Equation (4.4) by

$$\partial_t \nu_t = (\mathcal{A}_t^o)^* \nu_t, \quad \nu_0 = \mu_0, \tag{4.6}$$

by which we mean that a family of measures (ν_t) on $\mathbb{R}^d$ solves (4.6) if and only if, for every t,

$$\int_{\mathbb{R}^d} f(t,x)\,\nu_t(dx) - \int_{\mathbb{R}^d} f(0,x)\,\mu_0(dx) = \int_0^t \int_{\mathbb{R}^d} \left(\partial_s f(s,x) + \mathcal{A}_s^o f(s,x)\right)\nu_s(dx)\,ds \tag{4.7}$$

for all $f \in C_b^{1,2}([0,T]\times\mathbb{R}^d)$ where $C_b^{1,2}([0,T]\times\mathbb{R}^d) := \{f(t,x)\colon [0,T]\times\mathbb{R}^d \to \mathbb{R} \mid f, \partial_t f, \partial_i f, \partial_i\partial_j f \in C_b([0,T]\times\mathbb{R}^d)\}$. Note that, by construction, the Gibbs curve μ_t solves (4.6).

Proposition 4.6. *Let $(X_t)_{t\geqslant 0}$ be a weak solution to (4.3) with $X_0 \sim \mu_0$. Then*

$$\mathrm{Law}(X_t) = \mu_t \quad \textit{for almost every } t \in [0,T].$$

Proof. By [50], Theorem 32.10 the measure $P_{\mu_0} = \int P_x \mu_0(dx)$ exists and is unique, and hence the law of X_t with $X_0 \sim \mu_0$ is unique. Write $\rho_t := \mathrm{Law}(X_t)$ and denote by $b(t,x) = -\nabla U(x) + v_t(x)$.

Let $f \in C_b^{1,2}([0,T]\times\mathbb{R}^d)$. By Itô's formula

$$\begin{aligned} df(t,X_t) &= \partial_t f(t,X_t)\,dt + \langle \nabla_x f(t,X_t), dX_t\rangle + \frac{1}{\beta(t)}\Delta_x f(t,X_t)\,dt \\ &= \partial_t f(t,X_t)\,dt + \langle \nabla_x f(t,X_t), b(t,X_t)\rangle\,dt + \frac{1}{\beta(t)}\nabla_x \cdot f(t,X_t)\,dB_t + \frac{1}{\beta(t)}\Delta_x f(t,X_t)\,dt. \end{aligned}$$

Since $\partial_i f$ is bounded, $\mathbb{E}\left[\int_0^T |\nabla_x \cdot f(t,X_t)\beta^{-1}(t)|^2\,dt\right] < \infty$, so $\mathbb{E}\left[\int_0^t \beta^{-1}(s)\nabla_x \cdot f(s,X_s)\,dB_s\right] = 0$. Now

$$\begin{aligned} \mathbb{E}[f(t,X_t)] - \mathbb{E}[f(0,X_0)] &= \mathbb{E}\left[\int_0^t \partial_s f(s,X_s)\,ds\right] + \mathbb{E}\left[\int_0^t \langle \nabla_x f(s,X_s), b(s,X_s)\rangle\,ds\right] \\ &\quad + \mathbb{E}\left[\int_0^t \frac{1}{\beta(s)}\Delta_x f(s,X_s)\,ds\right]. \end{aligned}$$

Identifying that $\mathbb{E}[f(t,X_t)] = \int f(t,x)\rho_t(dx)$, and similarly for the other terms, we find that

$$\int_{\mathbb{R}^d} f(t,x)\rho_t(dx) - \int_{\mathbb{R}^d} f(0,x)\mu_0(dx) = \int_0^t \int_{\mathbb{R}^d} \left(\partial_s f(s,x) + \mathcal{A}_s^o f(s,x)\right)\rho_s(dx)\,ds,$$

and in fact, by [52], Theorem 9.3.6 the solution $\rho = \rho_t(dx)\,dt$ is unique[2], and hence we conclude that the Gibbs space-time measure agrees with the law of the process, *i.e.*, $\mu = \mu_t(dx)\,dt = \mathrm{Law}(X_t)\,dt$. □

4.2. Superposition with other stochastic dynamics

As alluded to in the introduction, we now revisit the discussion at the beginning of the previous section, in light of the Kolmogorov forward equation (4.6). Starting this time from the continuity equation,

$$\partial_t \mu_t = -\nabla \cdot (v_t \mu_t),$$

we have seen that with $\mathcal{L}_t$ denoting the diffusion generator, also

$$\partial_t \mu_t = -\nabla \cdot (v_t \mu_t) + \mathcal{L}_t^* \mu_t. \tag{4.8}$$

At least formally, this is clear since $\mathcal{L}_t$ preserves μ_t in the sense that, for each fixed t, $\mathcal{L}_t^* \mu_t = 0$. In other words, we can think of the family $(\mathcal{L}_t)$ as a homogeneous solution to the continuity equation. Additionally, we saw that there indeed is a Markov process (X_t) with generator $\mathcal{A}_t^o f = \langle v_t, \nabla f \rangle + \mathcal{L}_t f$. It is tempting to then consider

$$\mathcal{A}_t^o f = \langle v_t, \nabla f \rangle + \mathcal{A}_t f$$

where, for each t, $\mathcal{A}_t$ is the generator *any* μ_t-stationary stochastic process, *i.e.* satisfying $\mathcal{A}_t^* \mu_t = 0$, since then formally

$$\partial_t \mu_t = -\nabla \cdot (v_t \mu_t) + \mathcal{A}_t^* \mu_t. \tag{4.9}$$

However, it is in general not trivial to show that the right-hand side of (4.9) defines a generator for a Markov process. Nevertheless, we next give another example of where this is indeed the case.

4.2.1. Controlled piecewise deterministic simulated annealing

We will now turn our attention to a case of non-diffusive stochastic dynamics. The Bouncy Particle Sampler (BPS) is a rejection-free Monte Carlo method introduced in [53]; see also [54]. Samples from a distribution of interest, $\mu(dx) = \pi(x)\,dx \propto \exp(-U(x))\,dx$, are produced by simulating a particular *piecewise deterministic Markov process* (PDMP). For an in-depth treatment on PDMPs, see [55]. If π is defined on $\mathbb{R}^d$, then the BPS process $(Z_t)_{t\geqslant 0} = ((X_t, Y_t))_{t\geqslant 0}$ takes values in an extended state space $\mathbb{R}^d \times S^{d-1}$, where $S^{d-1} \subset \mathbb{R}^d$ denotes the unit sphere. We think of X_t as a position process, and Y_t as a velocity process. In between events, the position X_t follows linear trajectories in the piecewise constant velocity Y_t,

$$\frac{d}{dt}\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} Y_t \\ 0 \end{pmatrix}.$$

Clearly, this ODE is solved by the smooth map $\xi_t(x, y) = (x + ty, y)$. At reflection events, arriving according to an inhomogeneous Poisson process with the instantaneous rate given by

$$\lambda(X_t, Y_t) = \max\{0, \langle Y_t, \nabla U(X_t)\rangle\}, \tag{4.10}$$

[2]This is verified by noting that $v. \in L^1(\mu_t(dx)\,dt, \mathbb{R}^d \times (0,T))$, since we have already shown that v is more regular, $\|v_t\|_{L^2(\mu_t)} \in L^1(dt, [0,T])$.

the velocity changes deterministically, given by a specular reflection against the level set of U in the current point. That is, for a process in state X_t and with velocity Y_t at a jump time, the velocity is changed to

$$Y_t' = R(X_t)Y_t, \quad \text{where } R(x) := I - 2\frac{\nabla U(x)\nabla U(x)^T}{|\nabla U(x)|^2} \tag{4.11}$$

and $R(x) = I$ when the latter is not defined. With $\psi(dy)$ the uniform measure on S^{d-1}, this process admits $\pi(x)\,dx\,\psi(dy)$ as a stationary measure. However, this measure might in general not be the only stationary measure if one does not impose additional velocity refreshment events [54]. At these events, the velocity Y_t is resampled from the velocity distribution. To achieve this, an additional jump kernel, with a constant rate of events λ_R, is superimposed. The generator of the process is then given by

$$\begin{aligned}\mathcal{A}f(x,y) = &\langle \nabla_x f(x,y), y\rangle \\ &+ \lambda(x,y)\left(f(x,R(x)y) - f(x,y)\right) \\ &+ \lambda_R \int (f(x,y') - f(x,y))\psi(dy'),\end{aligned}$$

satisfying

$$\iint \mathcal{A}f(x,y)\exp(-U(x))\,dx\,\psi(dy) = 0 \tag{4.12}$$

for a suitable class of test functions.

Piecewise deterministic simulated annealing In [13] the BPS process is used as the basis for the simulated annealing process discussed in the previous section. Here, a theoretical analysis of the maximal cooling rate such that the law of the X_t-marginal converges to μ_∞ reveals that this process, as in the diffusion case, requires a logarithmic cooling schedule for general U. For a cooling schedule β, the time-inhomogeneous BPS is the Feller process on $\mathbb{R}^d \times S^{d-1}$ with generator given by

$$\begin{aligned}\mathcal{A}_t f(x,y) = &\langle \nabla_x f(x,y), y\rangle \\ &+ \beta(t)\lambda(x,y)\left(f(x,R(x)y) - f(x,y)\right) \\ &+ \lambda_R(t) \int (f(x,y') - f(x,y))\psi(dy')\end{aligned}$$

where $\psi = \mathrm{Unif}(S^{d-1})$ and $\lambda_R(t) \geqslant 0$ is some possibly time-dependent refreshment rate. By (4.12), $\mathcal{A}_t$ is μ_t-invariant.

Controlled piecewise deterministic simulated annealing We consider now the modified BPS on $(\mathbb{R}^d, S^{d-1})$ driven by the deterministic dynamics

$$\frac{d}{dt}\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \begin{pmatrix} Y_t + v_t(X_t) \\ 0 \end{pmatrix}.$$

To specify a PDMP with these characteristics, we need to be able to integrate these dynamics. To this end, we add an additional regularity assumption on the velocity field v_t.

Assumption 4.7. Assume $v_t(x)$ is continuous in time and Lipschitz in space, in the sense that there exists an $L > 0$, independent of t, such that

$$|v_t(x) - v_t(y)| \leqslant L|x - y| \quad \forall x, y, t.$$

The preceding assumption, strictly stronger than Assumption 4.3, guarantees that, for all s, x, y, the initial value problem

$$\frac{d}{dt}\zeta_t^s(x,y) = v_t(\zeta_t^s(x,y)) + y, \quad \zeta_0^s(x,y) = x \tag{4.13}$$

has a global (in time) solution ζ. For suitable functions f in the domain of the generator $\mathcal{A}_t^o$ of this process at time t, we have that

$$\mathcal{A}_t^o f(x,y) = \langle \nabla_x f(x,y), v_t(x)\rangle + \mathcal{A}_t f(x,y),$$

and as $\mathcal{A}_t$ preserves μ_t, we find that

$$\iint \mathcal{A}_t^o f(x,y)\pi_t(x)\,dx\,\psi(dy) = \iint f(x,y)\partial_t\pi_t(x)\,dx\,\psi(dy). \tag{4.14}$$

Under the conditions introduced so far, the following statement establishes results analogous to those in Section 4.1.

Proposition 4.8. *Assume that U satisfies Assumptions 3.1–3.3. Let v_t be as in Theorem 3.10 and assume that it further satisfies Assumption 4.7. Then there exists a PDMP $Z_t = (X_t, Y_t)$, taking values in $\mathbb{R}^d \times S^{d-1}$, with extended generator $\mathcal{A}_t^o$ such that, with $Z_0 \sim \pi_0(x)\,dx\,\psi(dy)$, it holds that*

$$\mathrm{Law}(Z_t) = \pi_t(x)\,dx\,\psi(dy),$$

where $\psi(dy)$ is the uniform measure on the sphere S^{d-1}.

Proof. Adding time as a coordinate, we introduce a homogeneous space-time PDMP $\mathcal{Z}_t = (t, X_t, Y_t)$ taking values in $\mathcal{M} := [0,T] \times \mathbb{R}^d \times S^{d-1}$. This process follows the homogeneous flow

$$\xi_t(s, X_s, Y_s) = (s + t, \zeta_t^s(x,y), 0), \quad \text{for } s, t \geqslant 0, s + t < T$$

until it reaches the boundary $s = T$, at which point it stops. By Assumption 4.7, ξ_t is as regular as need be. From [56], Proposition 4.2, we can verify that this process is Feller by showing that the flow ζ induced by v_t cannot bring the process arbitrarily fast toward the origin. Specifically, we need for every $t \geqslant s$ and y that

$$\lim_{|x_s|\to\infty} |x_t| = \infty.$$

Since v_t is Lipschitz, there exists constants c_0, c_1, such that, for all t,

$$\frac{d}{dt}|x_t| \geqslant -|v_t(x_t)| \geqslant -(c_0 + c_1|x_t|),$$

whereby Grönwall's inequality yields the estimate

$$|x_t| \geqslant |x_s|\mathrm{e}^{-c_1(t-s)} + \frac{c_0}{c_1}(\mathrm{e}^{-c_1(t-s)} - 1) \to \infty \quad \text{as} \quad |x_s| \to \infty.$$

Hence the space-time process $\mathcal{Z}$ is Feller and the space-time generator $\mathfrak{A} = \partial_t + \mathcal{A}_t^o$ admits $\mathcal{D} = C_c^\infty(\mathcal{M})$ as a core [56], Theorem 5.3. Let $\mu_0(dx)\psi(dy)$ be the initial law of (X_0, Y_0), which corresponds to $\mathcal{Z}_0 \sim \delta_0(dt)\mu_0(dx)\psi(dy)$. Since $\mathcal{Z}$ evolves purely deterministically in the time coordinate, it follows that then $\mathcal{Z}_t \sim \delta_t(dt)\rho_t(dx, dy)$ for $t \leqslant T$, and we define $\mathrm{Law}(X_t, Y_t) = \rho_t$. For any $f \in \mathcal{D}$ we find now by Dynkin's formula [55], Theorem 31.3 that, for any $0 \leqslant s \leqslant t \leqslant T$,

$$\begin{aligned}
\mathbb{E}[f(\mathcal{Z}_t)] - \mathbb{E}[f(\mathcal{Z}_0)] = \mathbb{E}\left[\int_0^t \mathfrak{A} f(\mathcal{Z}_s)\, ds\right] \iff \\
\int f(t,\cdot)\, d\rho_t - \int f(0,\cdot)\, d\rho_0 = \int_0^t \left(\int_{\mathbb{R}^d \times S^{d-1}} \frac{\partial f}{\partial s} + \mathcal{A}_s^o f\, d\rho_s\right) ds \\
= \int_0^t \left(\int_{\mathbb{R}^d \times S^{d-1}} \frac{\partial f}{\partial s} + \langle v_s, \nabla_x f\rangle\, d\rho_s\right) ds.
\end{aligned}$$

The above implies that the law of X_t satisfies the continuity equation $\nabla \cdot (v_t \mu_t) + \partial_t \mu_t = 0$. Additionally, the PDMP preserves the velocity distribution for any $t < T$ and the statement follows. □

4.3. Velocity field and diffusion on $\mathbb{R}$

In this section, we briefly consider the special case of a one-dimensional potential U. In this case, we can both show absolute continuity of the curve, and the existence of a diffusion process controlled by the velocity field, under weaker conditions. More specifically, we assume that U in this setting satisfies Assumptions 3.1 and 3.2, but we replace Assumption 3.3 with the following weaker condition:

Assumption 4.9. There exists an R and $\alpha > 0$ such that

$$|x| > R \implies x\left(\beta_0 U'(x) - \frac{U'(x)}{U(x)}\right) \geqslant \alpha |x|,$$

where β_0 is the lowest inverse temperature along the curve.

For $U : \mathbb{R} \to \mathbb{R}$ the continuity equation (2.5) reads

$$-\frac{d}{dx}(v_t \pi_t) = \partial_t \pi_t,$$

which is readily seen to be solved by

$$v_t(x) = -\frac{1}{\pi_t(x)} \int_{-\infty}^x \partial_t \pi_t(y)\, dy = \frac{\beta'(t)}{\pi_t(x)} \int_{-\infty}^x \left(U(y) - \int U\, d\mu_t\right) \pi_t(y)\, dy, \tag{4.15}$$

and the main result in this section amounts to verifying that this v_t is sufficiently regular.

Theorem 4.10. *Assume that $U : \mathbb{R} \to \mathbb{R}$ satisfies Assumptions 3.1, 3.2 and 4.9. Let $\beta : [0, T] \to [1, \infty)$ be a non-decreasing smooth cooling schedule. Then, for any $1 \leqslant p < \infty$, v_t defined as in (4.15) satisfies $\|v_t\|_{L^p(\mu_t)} \in L^1([0, T])$ and the curve*

$$[0, T] \ni t \mapsto \mu_t \in \mathscr{P}_p(\mathbb{R})$$

is absolutely continuous.

We give an elementary proof of the preceding theorem in Appendix A.2. In this, we note that $\int_{-\infty}^{x} \left(U(y) - \int U\, d\mu_t\right) \pi_t(y)\, dy \to 0$ when $|x| \to \infty$, and hence the assumptions on U forces v_t to point towards the origin in the tails. To conclude this section, we note that this control on the sign of v_t for large $|x|$ additionally allows us to drop the linear growth assumption used in Section 4.1, Assumption 4.3, yielding a stronger result as follows.

Proposition 4.11. *Assume that U and β are as in Theorem A.3 and let v_t be as in (4.15). Then the stochastic differential equation on $\mathbb{R}$ given by*

$$dX_t = (v_t(X_t) - \nabla U(X_t))\, dt + \sqrt{\frac{2}{\beta(t)}}\, dB_t$$

has a weak solution, and with $X_0 \sim \mu_0$ we have that $\mathrm{Law}(X_t) = \mu_t$ for almost every $t \in [0, T]$.

Sketch of proof. The proof follows analogously to those for Proposition 4.5 and Proposition 4.6. However, in the application of [11], Theorem 10.2.2 we instead use Assumption 4.9 and the sign of $v_t(x)$ for large $|x|$ to conclude that the conditions are satisfied. □

4.4. On Assumption 4.3

This section is mainly devoted to the proof of Proposition 4.12, restated here for convenience. We also make a few remarks related to both the assumption and the proposition. As a first such remark, we note that Assumption 4.3 is not needed in the case $d = 1$; see Section 4.3.

Proposition 4.12. *Let $U \in C^2(\mathbb{R}^d)$ be a twice continuously differentiable potential with bounded second derivatives, satisfying Assumptions 3.1 and 3.2. Denote by $\lambda_{\min}(A)$ the smallest eigenvalue of the matrix A. If*

$$\int \lambda_{\min}\left(\nabla^2 U(x)\right) \mu_t(dx) > 0, \quad \text{for all } 0 \leqslant t \leqslant T, \tag{4.16}$$

then Assumption 4.3 holds.

Proof. First, we fix now a reference temperature β such that (4.16) holds for the corresponding measure μ_β. We will now consider the velocity field v_β which arises at this specific temperature *via* the resolvent-like formulation in Remark 3.11.

From here on, the time index t is not related to the t in v_t, but refers to the time of the Langevin diffusion holding the temperature β fixed. Let X_t^x be the solution to the Langevin equation starting in $X_0 = x$, with semigroup denoted by (P_t). Recall the stochastic formulation of h_β from Remark 3.11, which equivalently can be written as

$$h_\beta(x) = -\int_0^\infty (P_t U(x) - \mu_\beta(U))\, dt.$$

As $v_\beta = \nabla h_\beta$, we have that

$$v_\beta = -\nabla_x \int_0^\infty (P_t U(x) - \mu_\beta(U))\, dt = -\int_0^\infty \nabla_x P_t U(x)\, dt$$

granted that we can exchange the order of integration and differentiation. As the noise is additive and U is C^2 with globally bounded hessian, the drift $-\nabla U$ is C^1 and Lipschitz, and we can therefor, *via* [57], Theorem 4.6.5

(see also [58], Thm. 6.13), express the derivative of the semigroup with respect to the initial condition x as

$$\nabla_x P_t U(x) = \nabla_x \mathbb{E}\left[U(X_t^x)\right] = \mathbb{E}\left[\nabla U(X_t^x) J_t^x\right],$$

where we denote by J_t^x the stochastic Jacobian matrix process $\frac{\partial X_t^x}{\partial x}$, given as the solution to

$$dJ_t^x = -\nabla^2 U(X_t^x) J_t^x \, dt, \quad J_0^x = I.$$

Applying Jensen and Cauchy–Schwarz,

$$\begin{aligned}\left|\mathbb{E}\left[\nabla U(X_t^x) J_t^x\right]\right|^2 &\leqslant \mathbb{E}\left[\left|\nabla U(X_t^x) J_t^x\right|^2\right] \leqslant \mathbb{E}\left[\left|\nabla U(X_t^x)\right|^2 \left\|J_t^x\right\|_{\mathrm{op}}^2\right] \\ &\leqslant \mathbb{E}\left[\left|\nabla U(X_t^x)\right|^2\right] \mathbb{E}\left[\left\|J_t^x\right\|_{\mathrm{op}}^2\right]\end{aligned}$$

We show now that $\left\|J_t^x\right\|_{\mathrm{op}}$ converges sufficiently fast to zero. So, let $y_0 \in S^{d-1}$ and consider $y_t = J_t^x y_0$. Then,

$$\begin{aligned}\frac{d}{dt}\frac{1}{2}\log\left(|y_t|^2\right) &= \frac{1}{2}\frac{1}{|y_t|^2}\frac{d}{dt}|y_t|^2 = \frac{y_t^T}{|y_t|^2}\frac{dJ_t^x}{dt} y_0 \\ &= -\frac{y_t^T}{|y_t|^2}\nabla^2 U(X_t^x) J_t^x y_0 \\ &= -\frac{y_t^T}{|y_t|}\nabla^2 U(X_t^x)\frac{y_t}{|y_t|} \\ &=: -\theta_t^T \nabla^2 U(X_t^x)\theta_t\end{aligned}$$

where we define $\theta_t := y_t/|y_t| \in S^{d-1}$. This means that we can write

$$\log(|y_t|) = \log(|y_0|) - \int_0^t \frac{d}{ds}\log(|y_s|)\, ds = \log(|y_0|) - \int_0^t \theta_s^T \nabla^2 U(X_s^x)\theta_s \, ds.$$

Now, note that $\theta_s^T \nabla^2 U(x)\theta_s \geqslant \lambda_{\min}\left(\nabla^2 U(x)\right)$ for all $x \in \mathbb{R}^d$, which means that for all $t > 0$

$$\frac{1}{t}\log(|y_t|) \leqslant \frac{1}{t}\log(|y_0|) - \frac{1}{t}\int_0^t \lambda_{\min}\left(\nabla^2 U(X_s^x)\right) ds.$$

In particular, it means that

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\log(|y_t|) &\leqslant \lim_{t\to\infty}\frac{1}{t}\log(|y_0|) - \frac{1}{t}\int_0^t \lambda_{\min}\left(\nabla^2 U(X_s^x)\right) ds \\ &= -\lim_{t\to\infty}\frac{1}{t}\int_0^t \lambda_{\min}\left(\nabla^2 U(X_s^x)\right) ds \\ &= -\int \lambda_{\min}\left(\nabla^2 U(x)\right)\mu_\beta(dx),\end{aligned}$$

where we in the last line replace the trajectory average by the space average, since the former converges to the latter by ergodicity of the process X_s^x [57], Theorem 1.3.12. Now, if $\int \lambda_{\min}\left(\nabla^2 U(x)\right)\mu_\beta(dx) = c > 0$, this means

that $\lim_{t\to\infty} \frac{1}{t}\log(|y_t|) \leqslant -c < 0$, where, in particular, c is independent of y_0. For large enough values of t, this means that

$$\frac{1}{t}\log(|y_t|) \leqslant -c + \frac{o(t)}{t},$$

where $o(t)$ is some function such that $\lim_{t\to\infty} \frac{o(t)}{t} = 0$, and thus that $|y_t| \leqslant \exp(-ct + o(t))$. This in turn means that $\|J_t^x\|_{\mathrm{op}} = \sup_{y_0\in S^{d-1}} |J_t^x y_0| = \sup_{y_0\in S^{d-1}} |y_t| \leqslant \sup_{y_0\in S^{d-1}} \exp(-ct+o(t)) = \exp(-ct+o(t))$ for large enough t. Hence there are constants C, C', c such that

$$\begin{aligned}
|v_\beta(x)| &= \left|\int_0^\infty \nabla_x P_t U(x)\,dt\right| \leqslant \int_0^\infty |\nabla_x P_t U(x)|\,dt = \int_0^\infty |\mathbb{E}\left[\nabla U(X_t^x(x)) J_t^x\right]|\,dt \\
&\leqslant \int_0^\infty \left(\mathbb{E}\left[|\nabla U(X_t^x)|^2\right]\right)^{1/2} \left(\mathbb{E}\left[\|J_t^x\|_{\mathrm{op}}^2\right]\right)^{1/2} dt \leqslant \int_0^\infty \left(\mathbb{E}\left[|\nabla U(X_t^x)|^2\right]\right)^{1/2} C\mathrm{e}^{-ct}\,dt \\
&\leqslant C'|\nabla U(x)|.
\end{aligned}$$

In conclusion, this gives $\langle x, v_\beta(x)\rangle \leqslant |x||v_\beta(x)| \leqslant C|x||\nabla U(x)| \leqslant C|x|(1+|x|) \leqslant C(1+|x|^2)$, where C is an appropriate constant that changes value between the inequalities. □

Remark 4.13 (Convex potentials)**.** A function $U \in C^2(\mathbb{R}^d)$ is strongly convex if and only if there exists some $m > 0$ such that $\lambda_{\min}(\nabla^2 U(x)) \geqslant m$ all $x \in \mathbb{R}^d$. From this perspective, (4.16) can be interpreted as U being "μ_t-on-average strongly convex", and the condition clearly holds for strongly convex functions. In fact, the condition holds for any convex function that has a positive definite Hessian in at least one point.

Remark 4.14 (Potentials with strictly positive Hessian in globally optimal solutions)**.** As mentioned in Remark 3.4, under Assumptions 3.1–3.3 we have that μ_t concentrates on the global minimizers of U as $\beta(t)$ increases[3]. In addition, assume that $U \in C^3$ and that the Hessian is strictly positive definite in all globally optimal solutions. This is true, *e.g.*, for Morse functions, and it means that set of globally optimal solutions must consists of isolated points. Using [9], Section 2 (see also [59], Chap. II) we can conclude that there exists a $\bar\beta > 0$ so that if $\beta(t) \geqslant \bar\beta$ for $t \in [t', T]$ for some $0 \leqslant t' < T$, then (4.16) is satisfied for $t \in [t', T]$.

5. Tractable approximations to controlled annealing processes

So far we have shown the existence of a velocity field which allows for arbitrarily fast cooling. In practice, computing the velocity field is difficult, as it is equivalent to solving the PDE (3.4) at every time t. Nonetheless, we will now introduce a series of approximations which in the end yields a tractable numerical scheme inspired by the idealized processes. In particular, we approximate the flow of the measures along the Gibbs curve using a finite number of interacting particles.

To this end, we first note that for any t, the limit in Proposition 2.5 yields that for any $\varepsilon > 0$ there exists some $h = h(\varepsilon) > 0$ such that

$$\left\|h^{-1}(T_{\mu_t\to\mu_{t+h}} - \mathrm{Id}) - v_t\right\|_{L^2(\mu_t)} < \varepsilon.$$

In other words, at any t, we could approximate v_t arbitrarily well by the finite difference quotient $h^{-1}(T_{\mu_t\to\mu_{t+h}} - \mathrm{Id})$. Computing the Monge map $T_{\mu_t\to\mu_{t+h}}$ is however still intractable. With access to independent random variables $X^1, \ldots, X^n \sim \mu_t$, $Y^1, \ldots, Y^n \sim \mu_{t+h}$ and denoting by μ_t^n, μ_{t+h}^n the corresponding empirical probability measures, one estimator for $T_{\mu_t\to\mu_{t+h}}$ would be the solution to the corresponding assignment problem. However, assuming that we are able to sample from μ_{t+h} means that we would already have solved the problem of

[3] In order to derive expressions for how the mass concentration looks like, *i.e.*, expressions for μ_∞, as well as rates for the concentration $\mu_t \to \mu_\infty$, stronger assumptions are needed; see [9] and [10], respectively.

following the Gibbs curve. Nevertheless, with access only to samples from μ_t, we can make μ_t^n resemble μ_{t+h}^n *via* importance sampling–like re-weighting, and then estimate the transport map by solving the transport problem between these discrete measures. In the next section, we establish that this procedure converges as $n \to \infty$, and in Section 5.2 we utilize this to design our numerical methods.

5.1. Self-normalized importance sampling estimation of optimal transport maps

Here, we briefly consider finite approximations to the minimizer T of the Monge problem

$$\inf_{T\colon T_{\#}\mu=\nu} \int |x - T(x)|^2 \, \mu(dx), \tag{5.1}$$

when $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ admit unnormalized Lebesgue densities p, q, and $\nu \ll \mu$. Specifically, we are interested in estimators to T from independent random variables $X^1, X^2, \ldots$ distributed according to μ. By the law of large numbers, the sequence of empirical distributions $\mu^n = n^{-1} \sum_{i=1}^n \delta_{X^i}$ converges weakly to μ. In fact, with normalized weights $w_i^n \propto q(X^i)/p(X^i)$, we additionally find that the empirical importance sampling measure $\nu^n = \sum_{i=1}^n w_i^n \delta_{X^i}$ converges weakly to ν, a standard result on importance sampling [60], Section 3.3.

As the Monge problem (5.1) is not in general solvable for measures supported on point clouds, we cannot *a priori* ask whether the solutions to the finite Monge problems from μ^n to ν^n converge. To remedy this, we instead consider conditional projections T^n of the optimal transport plans $\gamma^n \in \Gamma(\mu^n, \nu^n)$, defined as

$$T^n(x) := \mathbb{E}_{(X,Y)\sim\gamma^n}[Y \mid X = x] = \int y \, \gamma^n(x, dy).$$

In general, $T^n_{\#}\mu^n \neq \nu^n$, *i.e.*, T^n does not define an admissible map. Further, as T^n is not uniquely defined μ almost everywhere, we cannot hope for strong convergence in $L^2(\mu)$. It does however hold that $T^n \to T$ in a suitable weak sense and a suitable norm sense.

Proposition 5.1 ([30], Thm. 3.2)**.** *Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ be two probability measures admitting unnormalized Lebesgue densities p and q, such that $p(x) = 0 \implies q(x) = 0$. Let T be a Monge map from μ to ν solving (5.1). Let $X^1, X^2, \ldots$ be independent random variables such that $X^i \sim \mu(dx)$ for all i and define the empirical measures*

$$\mu^n := \frac{1}{n} \sum_{i=1}^n \delta_{X^i}, \quad \nu^n := \sum_{i=1}^n w_i^n \delta_{X^i}, \quad w_i := \frac{q(X^i)}{p(X^i)}, \quad w_i^n := \frac{w_i}{\sum_{j=1}^n w_j}.$$

Let γ^n be an optimal $\mathcal{W}_2$-transport plan of μ^n, ν^n,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, \gamma^n(dx, dy) = \inf_{\gamma \in \Gamma(\mu^n, \nu^n)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 \, \gamma(dx, dy),$$

and let $T^n(x) = \int y \, \gamma^n(x, dy)$ be the barycentric projection of γ^n. Then as $n \to \infty$, almost surely

$$T^n_{\#}\mu^n \to T_{\#}\mu = \nu \quad \textit{weakly}$$

and

$$\limsup_n \|T^n\|_{L^2(\mu_n)} \leqslant \|T\|_{L^2(\mu)}.$$

Proof. By the strong law of large numbers, both $\mu^n \to \mu$ and $\nu^n \to \nu$ weakly. As $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$, then, almost surely, $\mu^n, \nu^n \in \mathscr{P}_2(\mathbb{R}^d)$. Hence, by [14], Proposition 7.1.3, the sequence of optimal transport plans $(\gamma^n) \subset$

$\mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$ has weak limit points, and all such limit points γ satisfy

$$\gamma \in \operatorname*{argmin}_{\gamma' \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \, \gamma'(dx,dy).$$

As μ has a Lebesgue density, γ is unique. Therefore, the entire sequence γ^n in fact converges weakly to γ.

To show that T^n converges to T in the prescribed sense, we seek to apply [14], Theorem 5.4.4. To this end, we estimate

$$\begin{aligned}
\|T^n\|^2_{L^2(\mu^n)} &= \frac{1}{n} \sum_i \left| \int y \, \gamma^n(x_i, dy) \right|^2 \\
&\leqslant \frac{1}{n} \sum_i \int |y|^2 \, \gamma^n(x_i, dy) \\
&= \int |y|^2 \, \gamma^n(dx, dy) \\
&= \int |y|^2 \, \nu^n(dy) = \sum_{i=1}^n |x_i|^2 w_i,
\end{aligned}$$

using Jensen's inequality, and that γ^n is a coupling between μ^n and ν^n. As $\nu \in \mathscr{P}_2(\mathbb{R}^d)$, the strong law of large numbers yields that

$$\limsup_n \left| \sum_{i=1}^n |x_i|^2 w_i - \int |y|^2 \, \nu(dy) \right| = 0 \quad \text{almost surely.}$$

As $\int |y|^2 \, \nu(dy) = \int |T(x)|^2 \, \mu(dx) = \|T\|^2_{L^2(\mu)}$, we additionally find

$$\limsup_n \|T^n\|^2_{L^2(\mu^n)} \leqslant \|T\|^2_{L^2(\mu)}.$$

With $f \in C_c^\infty(\mathbb{R}^d)$

$$\begin{aligned}
\lim_n \int_{\mathbb{R}^d} f(x) T^n(x) \, \mu^n(dx) &= \lim_n \int_{\mathbb{R}^d} f(x) \left(\int_{\mathbb{R}^d} y \, \gamma^n(x, dy) \right) \mu^n(dx) \\
&= \lim_n \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x) y \, \gamma^n(dx, dy) \\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} f(x) y \, \gamma(dx, dy),
\end{aligned}$$

since $\gamma^n \to \gamma$ weakly. An application of [14], Theorem 5.4.4 now readily yields that, almost surely,

$$T^n_\# \mu^n \to T_\# \mu = \nu$$

as desired. □

5.2. Simulating approximately controlled annealing processes

We now describe tractable numerical approximation schemes of the processes in Section 4, relying on a finite difference approximation in time and a particle approximation in space, with parameters h, n respectively. The

final algorithms consist of simulating a population $\mathbf{X}_t$ of n particles,

$$\mathbf{X}_t = (X_t^1, ..., X_t^n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d,$$

and n piecewise constant velocity components

$$\mathbf{V}_t = (V_t^1, ..., V_t^n) \in \mathbb{R}^d \times ... \times \mathbb{R}^d.$$

Each particle X_t^i is driven by independent stochastic dynamics, and the velocity V_t^i, which is an approximation of the true velocity field. It is the velocities $\mathbf{V}_t$ that introduce interaction between the particles. At predetermined times, we compute $\mathbf{V}_t$ as follows: given that particle $X_t^i = x_i$ at time t for all i, we define the reweighted empirical measure $\hat{\mu}_{t+h}^n$ as in Proposition 5.1 by

$$\hat{\mu}_{t+h}^n := \sum_{i=1}^n w_i \delta_{x_i}, \quad w_i \propto \exp\left(-U(x_i)(\beta(t+h) - \beta(t))\right), \quad \sum_{i=1}^n w_i = 1. \tag{5.2}$$

With $C_{ij} := |x_i - x_j|^2$ and identifying that any $\gamma \in \Gamma(\hat{\mu}_t^n, \hat{\mu}_{t+h}^n)$ can be expressed as $\gamma = \sum_{ij} G_{ij} \delta_{x_i}(dx) \delta_{x_j}(dy)$, the transport problem between the two empirical measures is the finite-dimensional linear program

$$\begin{aligned} G^* := \underset{G \geqslant 0}{\operatorname{argmin}} \quad & \langle C, G \rangle_{\mathrm{Fr}} \\ \text{s.t.} \quad & G\mathbf{1} = \mathbf{1} \\ & G^T\mathbf{1} = n\mathbf{w}, \end{aligned} \tag{LP}$$

where $\mathbf{w} = (w_1, ..., w_n)^T$, $\langle C, G \rangle_{\mathrm{Fr}} := \sum_{ij} C_{ij} G_{ij}$ denotes the Frobenius inner product, and $\mathbf{1}$ denotes a vector of ones. By conditional projection of the solution G^* to this linear program, we find the transport map estimator

$$T(x_i) = \mathbb{E}_{(X,Y)\sim\gamma^*}[Y \mid X = x_i] = \sum_j G_{ij}^* x_j,$$

and set the components of $\mathbf{V}_t$ to

$$V_t^i = \frac{\left(\sum_j G_{ij}^* x_j\right) - x_i}{h}. \tag{5.3}$$

Barring any other dynamics, this velocity V_t^i would then move X_t^i to $T(X_t^i)$ in a straight line in h time.

Diffusive dynamics To integrate the diffusive dynamics, we use an Euler-Maruyama discretization with a fixed step length Δt, and we pick a velocity update interval $k\Delta t$ for some positive integer k. The resulting numerical scheme, described in Algorithm 5.1, is easily parallelisable over all steps except solving (LP).

Synchronous PDMP For n particles, we can formalize the algorithm as a single PDMP on $\mathbb{R}^{3\times d\times n}$ with scheduled events. For the process at time t, we simulate independently a random bounce time Δ^i for each particle and then advance the ensemble forward in time $\Delta = \min_i(\Delta^i)$ if $t + \Delta$ is before the next scheduled velocity estimation. If the next velocity computation is scheduled before $t + \Delta$, then the particles are instead propagated up to this time and the velocity is recomputed. To this end, let $\mathbf{Z}_t = (\mathbf{X}_t, \mathbf{Y}_t, \mathbf{V}_t)$ and define

$$\xi(s, t, \mathbf{Z}_t) = (\mathbf{X}_t + s(\mathbf{Y}_t + \mathbf{V}_t), \mathbf{Y}_t, \mathbf{V}_t),$$

Algorithm 5.1 Controlled Langevin diffusion for simulated annealing.

Given an objective function U, a cooling schedule β, initial states $\mathbf{X}_0 = \left(X_0^i\right)_{i=1}^n$, a time discretization Δt, a velocity update interval $h = k\Delta t$, and an end time T. Set $t = 0$.

1. *Estimate the velocity field*:
   (a) Compute the weights $\widetilde{w}_i = \exp(-(\beta(t+h) - \beta(t))U(X_t^i))$, the normalized weights $w_i = \widetilde{w}_i / \sum_j \widetilde{w}_j$, and the cost matrix $C_{ij} = |X_t^i - X_t^j|^2$.
   (b) Solve the discrete optimal transport problem $G^* = \operatorname{argmin}_{G_{ij} \geqslant 0} \{\langle G, C\rangle_{\mathrm{Fr}} \colon G\mathbf{1} = \mathbf{1}, G^T\mathbf{1} = n\mathbf{w}\}$.
   (c) Set $V_t^i = h^{-1}\left(\left(\sum_j G_{ij}^* X_t^j\right) - X_t^i\right)$ for $i = 1, ..., n$.
2. *Integrate the SDE*: For every particle i and $\ell = 0, ..., k-1$ sample independently $\zeta_\ell^i \sim \mathcal{N}(0, I_d)$ and set

$$X_{t+(\ell+1)\Delta t}^i = X_{t+\ell\Delta t}^i + \Delta t\left(V_t^i - \nabla U(X_{t+\ell\Delta t}^i)\right) + \sqrt{2\beta^{-1}(t+\ell\Delta t)\Delta t}\zeta_\ell^i.$$

Set $t = t + k\Delta t$ and repeat from Step 1 if $t < T$.

$$\lambda_b^i(t, \mathbf{Z}_t) = \max(0, \beta(t)\langle \nabla U(X_t^i), Y_t^i\rangle),$$
$$\mathcal{Q}_b^i(\mathbf{Z}_t, dz) = \delta_{(\mathbf{X}_t, (Y_t^1, ..., Y_t', Y_t^{i+1}, ..., Y_t^n), \mathbf{V}_t)}(dz), \quad Y_t' = R(X_t^i)Y_t^i.$$

Formally, we find a joint rate λ and a mixture kernel $\mathcal{Q}$

$$\lambda(t, \mathbf{Z}_t) = \lambda_r + \sum_i \lambda_b^i(t, \mathbf{Z}_t),$$
$$\mathcal{Q}(t, \mathbf{Z}_t, dz) = \frac{\lambda_r}{\lambda(t, \mathbf{Z}_t)}\mathcal{Q}_r(dz) + \sum_i \frac{\lambda_b^i(t, \mathbf{Z}_t)}{\lambda(t, \mathbf{Z}_t)}\mathcal{Q}_b^i(\mathbf{Z}_t, dz),$$

but in practice we understand this as sampling from the event that arrived first. The refreshment kernel $\mathcal{Q}_r$ resamples $\mathbf{Y}_t$ for one or more particles. With an update schedule $((t_k, \mathcal{Q}_k))_k$, $t_k = hk$ and $\mathcal{Q}_k$ a deterministic kernel supported on the velocity estimate from the barycentric projection of the solution of (LP), this scheme is implementable using the recursive construction in Algorithm 5.2.

Asynchronous PDMP In contrast to the discretization of the diffusive dynamics, where it was simple to propagate each particle forward in time in parallel, synchronization of the BPS dynamics comes at a considerable cost. Indeed, as the random dynamics affecting the particles in between velocity updates are completely independent, it is possible to let the particles evolve asynchronously in time until all of them have reached the boundary of the next scheduled velocity update. Similarly to the construction used in the proof of Proposition 4.8, we introduce the process

$$\mathcal{Z}_t = (\mathbf{T}_t, \mathbf{X}_t, \mathbf{Y}_t, \mathbf{V}_t) \in (\mathbb{R}_+ \times \mathbb{R}^d \times S^{d-1} \times \mathbb{R}^d)^n.$$

When the time coordinate T_t^i of a particle i lies in $[\tau_k, \tau_{k+1})$, it follows the deterministic dynamics

$$\xi^i(s, t, (T_t^i, X_t^i, Y_t^i, V_t^i)) = (T_t^i + s, X_t^i + s(Y_t^i + V_t^i), Y_t^i, V_t^i)$$

and upon reaching the boundary $T_t^i = \tau_{k+1}$ the value of T_t^i stops increasing until all particles have reached the boundary. When this happens, the velocity $\mathbf{V}_t$ is recomputed and the particles evolve forward again on $[\tau_{k+1}, \tau_{k+2})$.

Algorithm 5.2 Recursive construction of a PDMP with scheduled events.

Let Z_t be the state of a PDMP at time t, governed by an event rate λ, a flow ξ, a transition kernel $\mathcal{Q}$, and an event schedule $((t_k, \mathcal{Q}_k))_k$.

1. Let $t^* = t_{k*} = \min\{t_k \colon t_k > t\}$ be the arrival of the next scheduled event.
2. Simulate independently Δ as the first arrival time of an inhomogeneous Poisson process with rate $s \mapsto \lambda(t + s, \xi(s, t, Z_t))$, that is, a $[0, \infty)$-valued random variable with distribution function

$$F(\tau) = \exp\left(- \int_0^\tau \lambda(t + s, \xi(s, t, Z_t))\, ds \right), \quad \tau \geqslant 0.$$

3. Let $t' = \min(t^*, t + \Delta)$ and set for $s \in [0, t' - t)$

$$Z_{t+s} = \xi(s, t, Z_t).$$

4(a) *If* $t' = t + \Delta$: Draw

$$Z_{t'} \sim \mathcal{Q}(\xi(t' - t, t, Z_t), dz).$$

(b) *Else*, $t' = t^*$: Draw

$$Z_{t'} \sim \mathcal{Q}_{k*}(\xi(t' - t, t, Z_t), dz).$$

5. Repeat with $Z_{t'}$.

5.3. Remarks and connections to other methods

We end this section with a few remark, some of them about the connection between the algorithms we suggest and other algorithms in the literature.

Remark 5.2 (Computational complexity)**.** For a particle population of size n, the linear program (LP) can be solved in $\mathcal{O}(dn^2 + n^3)$ time using the network simplex algorithm [61]. This increases the computational complexity of Algorithm 5.1 and Algorithm 5.2, as compared to their respective non-controlled version, and evidently puts an eventual upper limit on the number of particles that can be simulated. However, it is noteworthy that the computational cost only scales linearly with respect to the dimension d of the underlying space[4]. Moreover, there are many regimes where the dominant computational constraint is levied by the evaluation of the objective function and its gradient.

Remark 5.3 (Sinkhorn and entropic transport maps)**.** While we in Remark 5.2 argue that the linear program (LP) can be efficiently solved for reasonable sizes of particle populations, there are (most likely) applications where the intermediate Gibbs measures are of more immediate interest, which in turn could warrant finer approximations of the flows *via* larger particle swarm sizes. The canonical choice for dealing with large scale transport problems is by smoothing the transport problem with an entropic regularization term, which allows for an efficient and highly parallelizable solution procedure *via* Sinkhorn iterations [45, 62]. It has also been suggested that, under a natural interpretation, the corresponding entropic transport map is more amenable to approximation [63]. In future works, it would be of interest to explore such approximations and see if they would improve the estimate of the velocity field.

[4]Computing the cost used in optimal transport problem requires computing the norm of $(n + 1)n$ vectors in dimension d.

Remark 5.4 (Consensus-based optimization)**.** It is natural to compare our numerical method to other ensemble-based optimization methods. In consensus-based optimization [22, 24], each particle X_t^i follows an SDE with a drift term $b^i(\mathbf{X}_t)$ which steers the particle to the current *consensus point* $c_\alpha(\mathbf{X}_t)$ of the ensemble. Here $c_\alpha(\mathbf{X}_t) = \int x \hat{\rho}_t^n(dx)$, where $\hat{\rho}_t^n$ is the probability measure defined by $\hat{\rho}_t^n \propto \sum_{i=1}^n \exp(-\alpha U(X_t^i))\delta_{X_t^i}$ and $b^i(\mathbf{X}_t) = \lambda(c_\alpha(\mathbf{X}_t) - X_t^i)$ for some chosen $\lambda > 0$, and the parameter α tunes the concentration towards the current best estimate. In contrast, our particles follow more inhomogeneous dynamics, as they are driven not towards a single global consensus point, but instead towards some individually assigned point such that the drift of the entire ensemble achieves a consensus effect.

Remark 5.5 (Birth-Death processes and Sequential Monte Carlo Simulated Annealing)**.** While we have focused on solving the Fokker–Planck equation (1.5) using velocity fields, it is also possible to instead consider other types of stochastic processes. We restrict now our attention to processes taking values in a compact subset $\mathcal{X} \subset \mathbb{R}^d$. Assume that the inverse temperature β is non-decreasing, $\beta' \geqslant 0$, and assume that U is continuous, in which case both $\sup_{x\in\mathcal{X}} U(x)$ and $\inf_{x\in\mathcal{X}} U(x)$ are bounded. Let $\hat{U}(x) = U(x) - \inf_{x\in\mathcal{X}} U(x) \geqslant 0$. The operator

$$\mathcal{Q}_t f(x) = |\beta'(t)\hat{U}(x)| \int (f(y) - f(x))\, \pi_t(y)\, dy$$

is the infinitesimal generator of a pure jump process [50], Proposition 17.2. This process jumps with the position-dependent rate $|\beta'(t)\hat{U}(x)| = \beta'(t)\hat{U}(x)$ to a random location independently sampled from μ_t. This process also formally solves the Fokker–Planck equation $\partial_t \mu_t = \mathcal{Q}_t^* \mu_t$ in the sense that

$$\begin{aligned}
\int \mathcal{Q}_t f(x)\pi_t(x)\, dx &= \int \left(|\beta'(t)\hat{U}(x)| \int (f(y) - f(x))\, \pi_t(y)\, dy \right) \pi_t(x)\, dx \\
&= \beta'(t) \left(\int f(y)\pi_t(y)\, dy \int \hat{U}(x)\pi_t(x)\, dx - \int \hat{U}(x) f(x)\pi_t(x)\, dx \right) \\
&= -\beta'(t)\, \mathrm{Cov}_{\mu_t}\left(f, \hat{U}\right) = -\beta'(t)\, \mathrm{Cov}_{\mu_t}\left(f, U\right) = \int f(x)\partial_t \pi_t(x)\, dx,
\end{aligned}$$

see (3.3). A natural way to numerically implement this process is by using a particle swarm approximation to μ_t, removing particles with the position-dependent rate $|\beta'(t)\hat{U}(x)|$, and recreating them at a new location, drawn independently from the locations of the pool of remaining particles. This can be viewed as a continuous-time version of Sequential Monte Carlo Simulated Annealing [18], a discrete time simulated annealing algorithm where particles are successively resampled over time according to suitably chosen weights.

Remark 5.6 (Predictor–corrector)**.** At the end of Section 1.1, we note that our view of simulated annealing is reminiscent of an interior point method. In this context, it is also interesting to note that our proposed algorithms could arguably be seen as predictor–corrector-type methods [64, 65]. More precisely, estimating and using the velocity field v_t can be interpreted as a predictor-type step, which estimates how to move the particles to follow along the curve. Analogously, the superposition with stochastic μ_t-invariant and ergodic dynamics can be interpreted as a corrector step, where the mixing moves the particles closer to a configuration that represents the current point on the curve (*cf.* Example 4.1). In relation to this, it also becomes clear why the speed of mixing should be high enough relative to the speed with which we move along the Gibbs curve, *cf.* Remark 4.2.

Remark 5.7 (Non-stochastic version)**.** While we have eluded to the role played by the stochastic dynamics, in particular in Example 4.1 and Remark 5.6, a natural question to ask is how an algorithm only using estimates of the velocity field (*i.e.*, without the stochastic dynamics) behaves. First, it is worth observing that in this case, the particles would never move outside of the convex hull of their initial positions, which means that a global minimum must be inside this convex hull for the algorithm to find it. Second, in the simplified case of only two particles, it is easy to see that the resulting dynamics would be that the particle initiated in a point with higher objective function value would move towards the other particle, and that the latter would not move. Moreover,

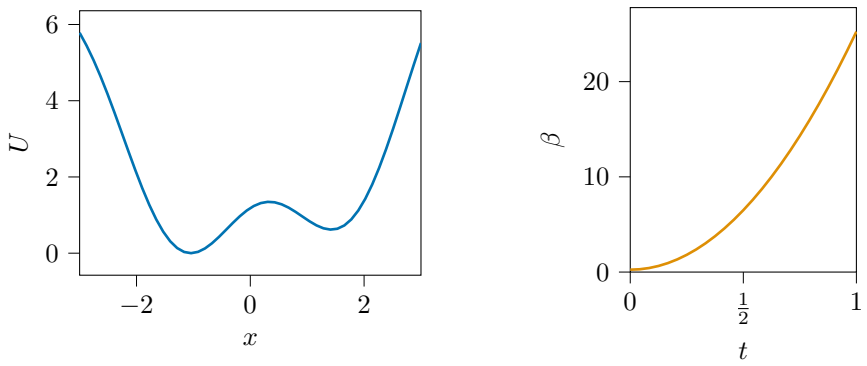


FIGURE 1. The double-well potential U, (left) and the cooling schedule β of Section 6.1.

when the first particle would reach a point with the same objective function value as the second particle, it would stop. This is because no reweighting occurs, and hence the estimated flow is zero. Analogous behaviour is observed in the case of multiple particles, and the particle population typically quickly gets stuck on a level set of the potential.

6. Numerical experiments

We perform a number of numerical experiments to illustrate the behaviour and to examine the practical applicability of the proposed method. Before this, we note that the velocity estimates and the independent stochastic dynamics can run on different timescales. For any potential U, we fix the time interval to $[0, 1]$ and a cooling schedule β. After choosing a time discretization Δt for the Langevin dynamics, we can freely choose any parameter $\lambda > 0$ and instead integrate

$$dX_t = -\lambda \nabla U(X_t)\, dt + \lambda \sqrt{2\beta(t)^{-1}}\, dB_t,$$

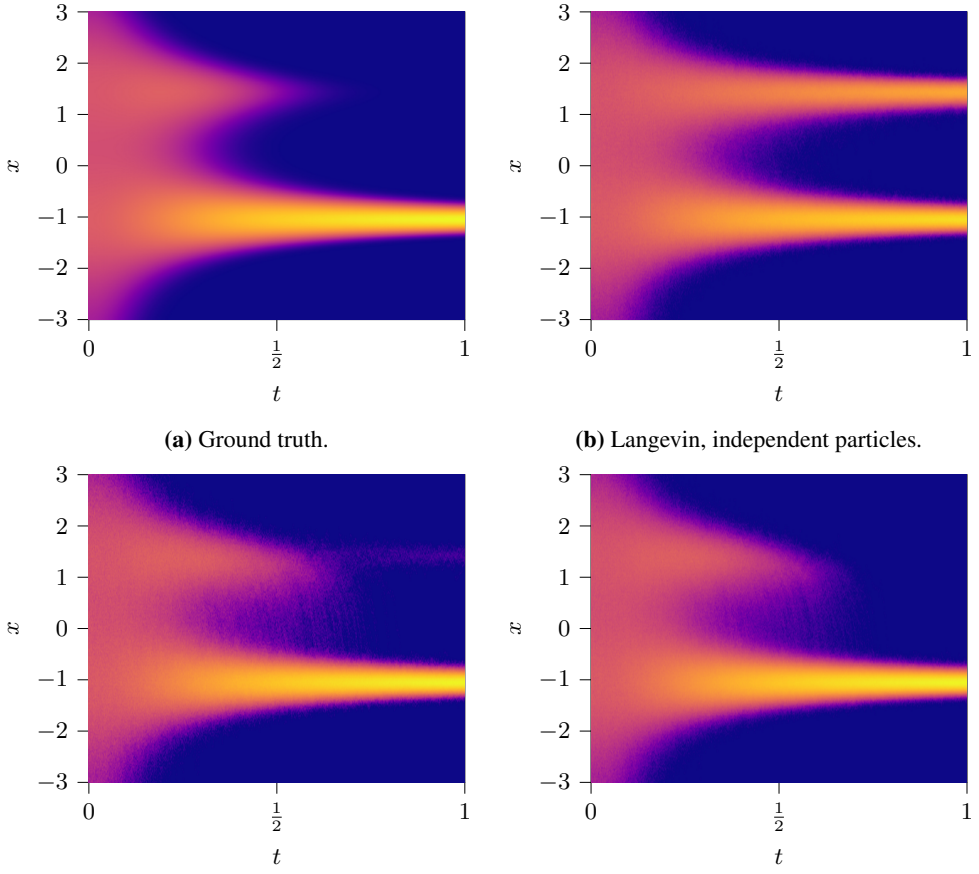
which, in practice, amounts to finding a stable gradient descent step length $\lambda \Delta t$. In the PDMP version, this corresponds to choosing a scaling $\lambda > 0$ for the BPS velocity component $Y_t \in \lambda S^{d-1}$. For the simulations we have performed, we present the normalized time interval $[0, 1]$ and give the corresponding choice of $\lambda \Delta t$ or λ, respectively. For brevity, we abbreviate the PDMP-based piecewise deterministic simulated annealing method as PDSA. In the sequel, we have implemented the synchronous version of the controlled PDSA, which lends itself well to the qualitative studies we perform. We provide a reference implementation and code for reproducing the results presented in this section at https://github.com/vincentmolin/controlled_annealing.

6.1. A double-well potential on $\mathbb{R}$

We first consider the double-well potential $U \colon \mathbb{R} \to \mathbb{R}$ given by

$$U(x) = \frac{1}{2}x^2 + \cos\left(2x - \frac{1}{2}\right) + C, \tag{6.1}$$

where C is chosen such that $\min U = 0$. This potential has a unique global minimum close to $x = -1$, and a local suboptimal minimum in the vicinity of $x = 1.5$. Our Gibbs curve is then defined by choosing the cooling schedule given by $\beta \colon t \mapsto \frac{1}{2} + 25t^2$ for $t \in [0, 1]$. The potential and the cooling schedule are shown in Figure 1. In particular, we initialize each particle by sampling from μ_0. Here, as β increases quickly, the independent stochastic dynamics of particles cause them to stick in the suboptimal well. Simulating the particles instead



(a) Ground truth.

(b) Langevin, independent particles.

(c) Controlled Langevin, 5 particles.

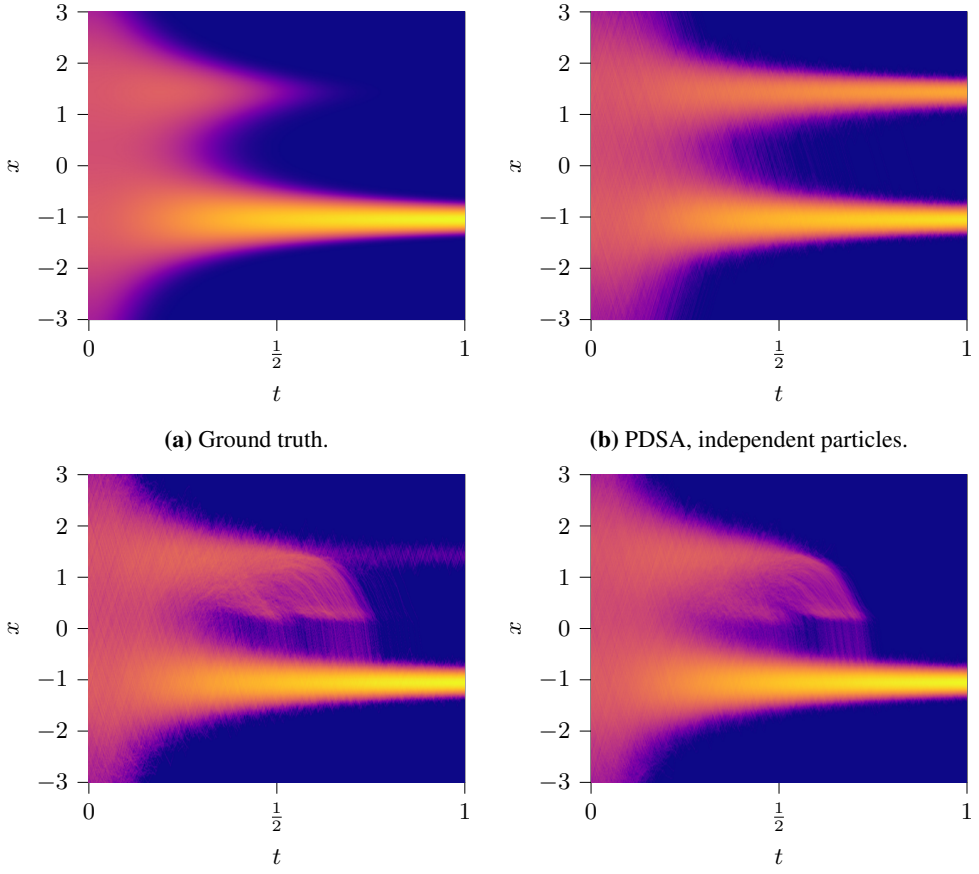
(d) Controlled Langevin, 10 particles.

FIGURE 2. Heat maps of empirical probability densities for the Gibbs measures corresponding to the double-well potential U in (6.1) and the quadratic cooling schedule $\beta(t) = \frac{1}{2} + 25t^2$, for $0 \leqslant t \leqslant 1$. The Langevin versions are simulated with 1000 steps of length $\lambda\Delta t = 25 \cdot 10^{-3}$. For the controlled versions we set $h = 2 \cdot 10^{-2}$, that is, we compute a new velocity estimate 50 times. From the upper right panel, we see that the fact that β increases quickly causes particles obeying the independent Langevin dynamics to stick in the suboptimal well. In contrast, from the two panels at the bottom row we see that the interactions in the proposed method allows particles to escape the local minima. Histograms are binned averages over 10,000 particle trajectories, coloured with an inverse hyperbolic sine colour scale.

in groups according to the controlled algorithms allows more flow of probability mass to the global optimum. Figure 2 illustrates the phenomenon for the diffusion-based method, as Figure 3 does for the PDMP version.

The most dramatic flow of mass happens close to the centre of the time interval $[0, 1]$. In Figure 4 we quantify this by computing the ground truth metric derivative, which has a fairly sharp peak in the middle. By also estimating the $\mathcal{W}_2$-distance of the average marginal law of the particles and the ground truth, we see that this is the region where the laws diverge from the Gibbs curve. As can be seen in the figure, increasing the population of interacting particles counteracts this to some extent.

We then estimate the convergence of the laws to the Gibbs curve as we increase the number of particles and the frequency of the velocity updates, keeping the other simulation parameters specified in Figure 2 and Figure 3 fixed. The results, shown in Figure 5, suggest that the convergence of the controlled PDSA marginals



(a) Ground truth.

(b) PDSA, independent particles.

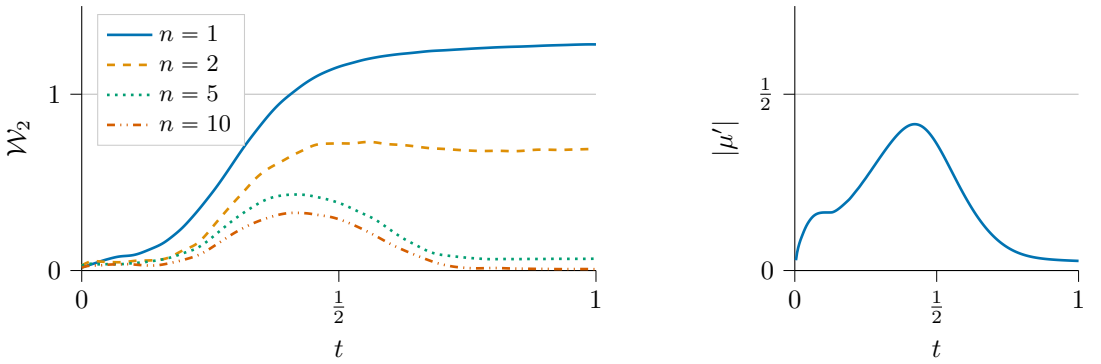
(c) Controlled PDSA, 5 particles.

(d) Controlled PDSA, 10 particles.

FIGURE 3. Heat maps of empirical probability densities for the Gibbs measures corresponding to the double-well potential U in (6.1) and the quadratic cooling schedule $\beta(t) = \frac{1}{2} + 25t^2$, $0 \leqslant t \leqslant 1$. The PDSA versions are simulated with a speed scale $\lambda = 25$, and for the controlled versions we set $h = 2 \cdot 10^{-2}$, that is, we compute a velocity estimates 50 times. As in the Langevin case in Figure 2, the independent PDSA dynamics of particles stick in the suboptimal well. In contrast, the interactions in the proposed method allows particles to escape the local minima. Histograms are binned averages over 10,000 particle trajectories, coloured with an inverse hyperbolic sine scale.

to the Gibbs curve approaches a power law with respect to the number of particles, while we do not observe this scaling in the Langevin setting.

Finally, we examine the robustness of the procedure by sampling the starting positions of the particles not from μ_0 but instead a sequence of shifted Gaussian measures. In Figure 6 we track the $\mathcal{W}_2$-distance of the estimated law to the Gibbs curve of the Langevin version as well as the controlled counterpart with populations of 5 or 10 particles, for initial distributions $\mathcal{N}(i, 1)$, $i = 0, 1, 2, 3$. With a high initial temperature β_0^{-1}, the high diffusivity of the particles mixes the law rapidly towards the Gibbs curve and the offset in initial positions has little effect on the final distributions. Barring the initial relatively high distance to μ_0, the behaviour closely matches that of Figure 4. As it is more difficult for the particles to explore the potential landscape in the colder regime, we also simulate the evolution of these offset initial laws starting further along the same cooling Gibbs



(a) $\mathcal{W}_2$-distance to the Gibbs curve μ_t. **(b)** The speed of the Gibbs curve in $\mathscr{P}_2$.

FIGURE 4. In the left panel, the distance of the time marginals from the Langevin simulations in Figure 2 is shown. Here n is the population size, and $n = 1$ corresponds to independent dynamics, while $n = 2, 5, 10$ denotes controlled versions. The largest deviation for the controlled methods happens around $t = 0.4$, at which point the Gibbs curve has the largest rate of change, as quantified by the metric derivative $|\mu'|(t) = \|v_t\|_{L^2(\mu_t)}$, see the right panel.

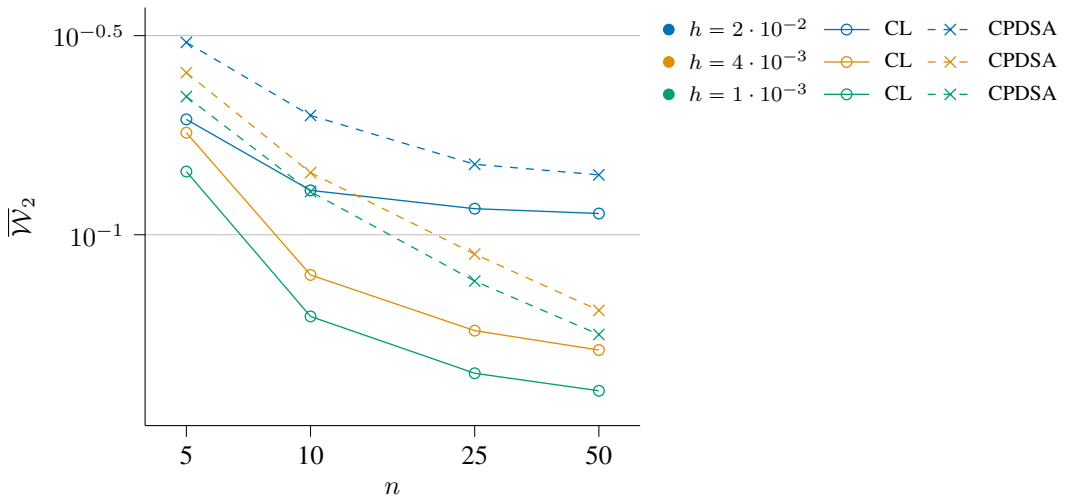
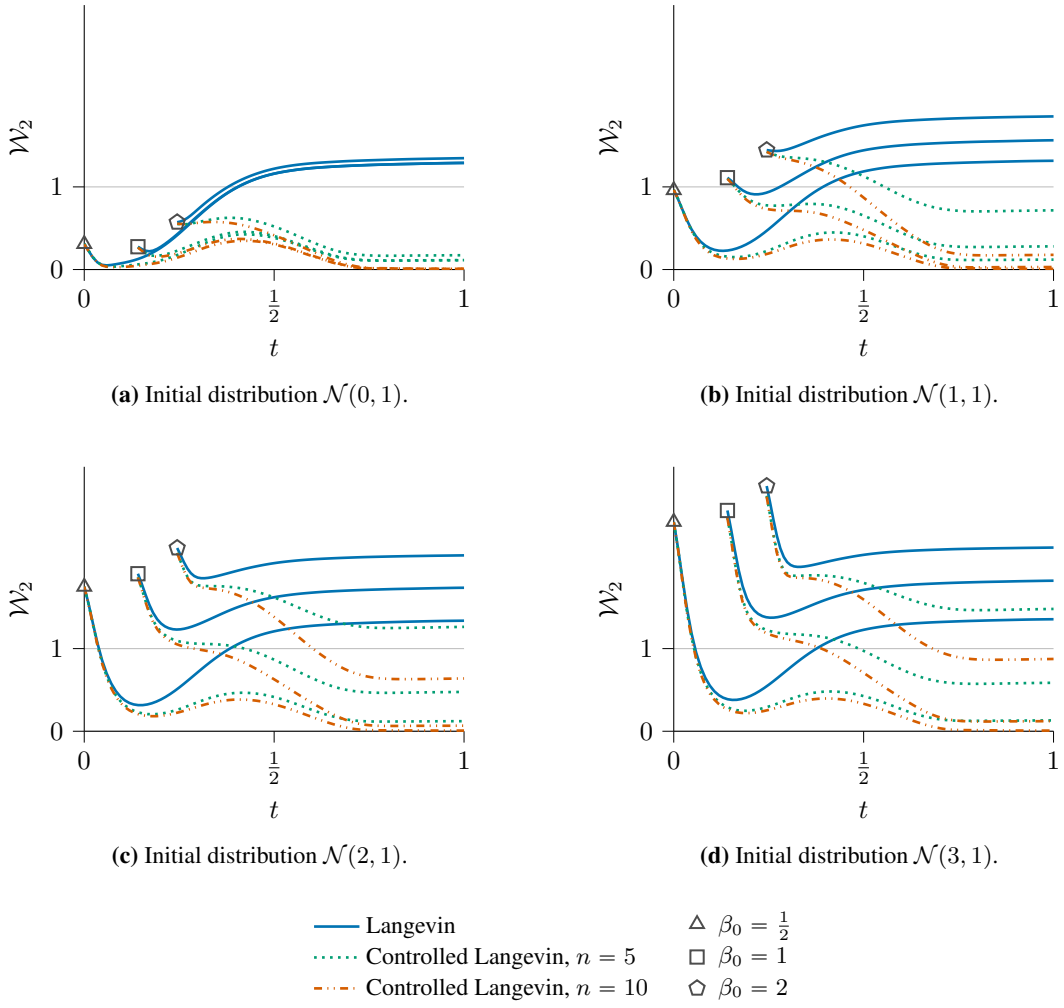


FIGURE 5. Convergence to the double-well Gibbs curve as the number of particles n increases and the velocity update interval h decreases. Varying the number of particles n (horizontal axis) and the velocity update interval h (colour), we average 1,000 simulations to estimate the time marginals $\rho_t^{(n,h)}$ of the simulations, and compute the time averaged $\mathcal{W}_2$-distance, $\overline{\mathcal{W}}_2 = \int_0^1 \mathcal{W}_2(\rho_t^{(n,h)}, \mu_t)\, dt$. Both axes are logarithmically scaled. For smaller h, the average distance of the controlled PDSA marginals seems to approach a linear law in this scaling. This suggests that the average distance of the controlled PDSA marginals approaches a power law with respect to the number of particles, while we do not observe this in the Langevin setting.

curve, beginning instead at $\beta(t_k) = k$, for $k = 1, 2$. Shown in the same figure, the estimated velocity fields leads to a better agreement with the Gibbs curve.



(a) Initial distribution $\mathcal{N}(0, 1)$.

(b) Initial distribution $\mathcal{N}(1, 1)$.

(c) Initial distribution $\mathcal{N}(2, 1)$.

(d) Initial distribution $\mathcal{N}(3, 1)$.

FIGURE 6. For the double-well potential we track the distance in $\mathcal{W}_2$ to the Gibbs curve for the empirical law of particles started not in μ_0 but $\mathcal{N}(i, \sigma = i)$, for $i = 0, 1, 2, 3$, keeping the other simulation parameters as in Figure 2. The repeated lines correspond to the laws of particles started from the same initial distribution but at a lower initial temperature, $\beta_0 = 1$ and $\beta_0 = 2$ respectively. For all initial distributions above, the particles started at $\beta_0 = 1/2$ quickly approaches the Gibbs curve, and we recover the behaviour in Figure 4. For the colder starts, the velocity field estimations leads to a better tracking of the Gibbs curve, an effect that increases with the number of particles.

6.2. Optimization problems

We now evaluate the effect of the velocity estimations by considering two versions of common benchmark problems in optimization on $\mathbb{R}^d$. Specifically, we set $d = 10$ and consider the Rosenbrock function

$$U_1(x) := \sum_{i=1}^{d-1} 5\left(x_{i+1} - x_i^2\right)^2 + (1 - x_i)^2, \tag{6.2}$$

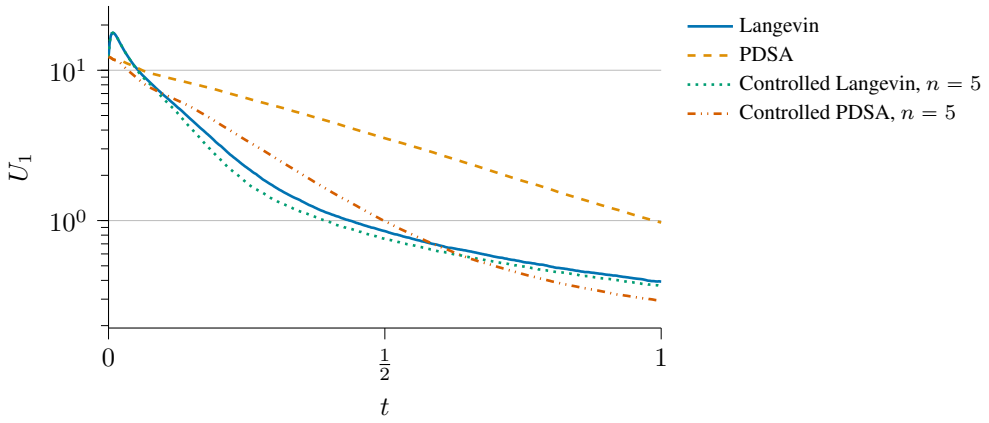


FIGURE 7. Results for the simulations on the Rosenbrock potential U_1 in (6.2). Out of 2000 total simulations with 5 particles, we report the median best of $k = 5$ objective value, $\min_{i \leqslant k} U_1(X_t^i)$, simulated independently (Langevin, PDSA) or in groups of 5.

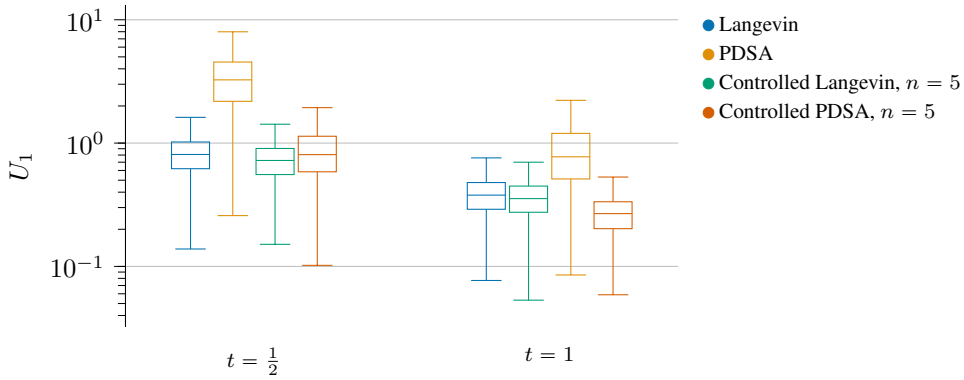


FIGURE 8. Box plots of the best objective values at $t = \frac{1}{2}$ and 1, showing the run-to-run variance of the results from the Rosenbrock simulations in Figure 7.

and the Rastrigin function given by

$$U_2(x) := |x|^2 + \sum_{i=1}^{d} (1 - \cos(2\pi x_i)), \tag{6.3}$$

both scaled as in [66]. Moreover, and in contrast to the previous experiment, we now initialize the particles away from the Gibbs curve to model a more realistic scenario.

For the Rosenbrock function, we use the linear cooling schedule $\beta(t) = 0.1 + 25t$. We draw the initial configurations independently from $\mathcal{N}(m_1, \frac{1}{20}I)$, centred at $m_1 = (-1.5, 1, 1, ..., 1)$. For the Langevin versions, we use a total of 500 steps of length $\lambda\Delta t = 5 \cdot \frac{1}{500}$. We run the PDSA versions on a timescale of $\lambda = 10$. We simulate controlled versions of both with 5 particles and 25 velocity updates, and for comparison, we also simulate the independent versions in groups of 5. In all simulations, we track the current best objective value $\min_{i=1,...,5} U_1(X_t^i)$ of any particle. In Figure 7 we plot the median of this metric across 2000 simulations, and in Figure 8 we illustrate the observed variance between these simulations in the form of box plots. For the Langevin versions, we find essentially no gain in performance from the velocity estimates. While there is a stronger effect in the PDSA case, this is mainly due to the poor performance of the independent PDSA algorithm.

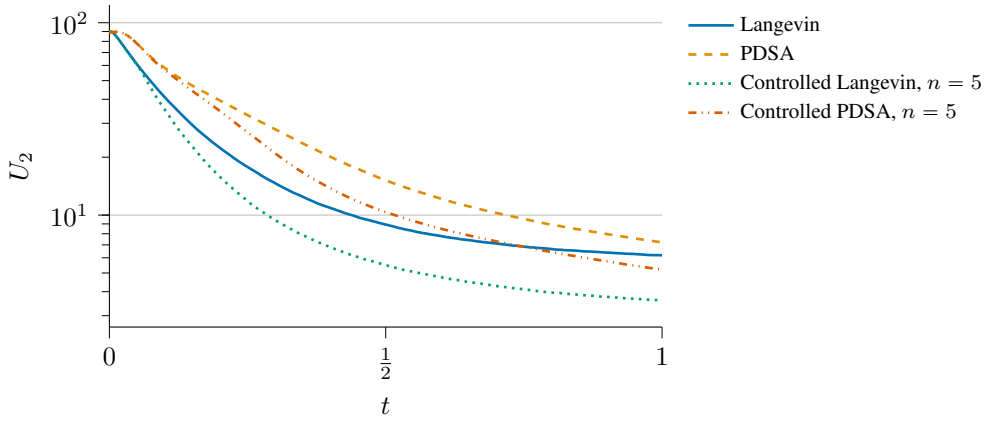


FIGURE 9. Results for the simulations on the Rastrigin potential U_2 in (6.3). Out of 2000 total simulations with k particles, we report the median best of $k = 5$ objective value, $\min_{i \leqslant k} U_2(X_t^i)$, simulated independently or in groups of 5. Here, the controlled Langevin version performs better than the other methods.

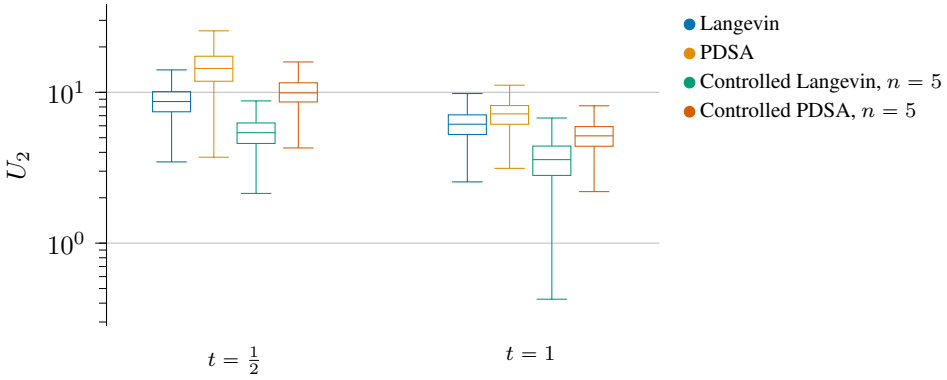


FIGURE 10. Box plots of the best objective values at $t = \frac{1}{2}$ and 1, showing the run-to-run variance of the results from the Rastrigin simulations in Figure 9. The controlled Langevin version performs best.

For the highly multimodal Rastrigin potential, we use the linear cooling schedule $\beta(t) = 0.1 + 5t$. We initialize the particles in the local minima around $m_2 = (3, ..., 3)$, drawing initial configurations from $\mathcal{N}(m_2, \frac{1}{20} I)$. We run the Langevin versions for 500 steps of length $\lambda \Delta t = \frac{5}{2} \cdot \frac{1}{500}$. We run the PDSA versions at a timescale of $\lambda = 50$, and we estimate the velocity 25 times for the controlled versions. In this experiment we compute similar, best of $k = 5$, objective values, shown in Figure 9 and Figure 10. The results show that the controlled Langevin version performs best, and in Figure 11 we find that simulating this controlled version with 5 particles performs similar to taking the best of 50 independent trials.

For further comparison, we assign a similar computational budget to the independent and the controlled Langevin methods on the Rastrigin problem. Solving a transport problem and then taking 20 Langevin steps takes nearly twice as long compared to only taking the Langevin steps for 5 independent particles. Therefore, we assign 1000 Langevin steps per particle to the independent version and 500 Langevin steps with 25 velocity field estimations to the controlled version, yielding an equivalent elapsed wall-clock time. We tune the linear cooling schedule and the Euler–Maruyama step length Δt for each method by searching for the parameters that yields the average best trajectory minimum using Optuna [67]. Additionally, we also tune the scale of the

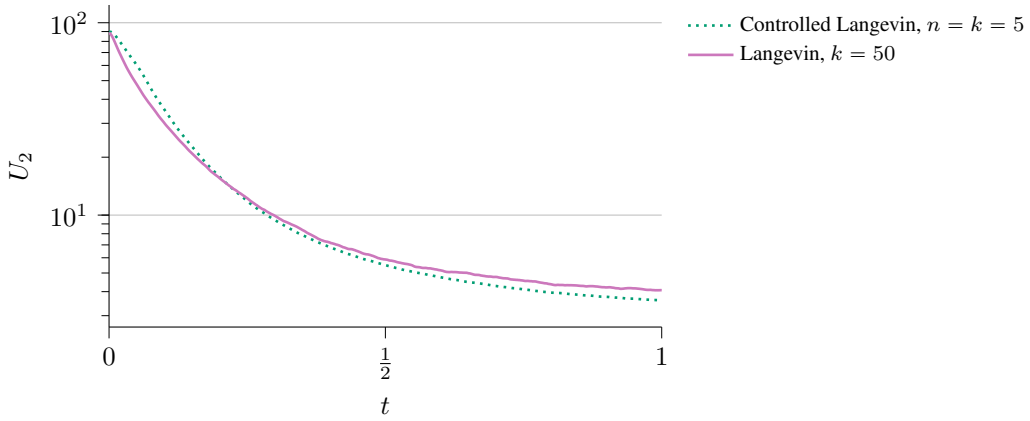


FIGURE 11. We once again compare the controlled Langevin version of Figure 9, simulated in groups of 5, to independent simulations. We plot the median best of 5 for the controlled version, to the median best of 50 independent simulations. In this case, superimposing 25 velocity estimations performs similar to a tenfold increase in independent starts.

TABLE 1. Hyperparameters selected for the performance comparison of both Langevin based methods on the Rastrigin potential in Figure 13 and Figure 12. Using the Optuna [67] framework, we tune the linear cooling schedule and the Euler–Maruyama step length Δt for each method by searching for parameters that leads to low trajectory minima, performing 200 trials per method, each averaged over 250 trajectories. We also tune the scale of the finite differencing h, $h = \Delta t \cdot n_{\text{steps per v}}$, in the velocity estimations, allowing it to vary from $0.5h$ to $2.5h$. The number of steps is fixed for each method, as well as the update interval for the controlled variant.

Hyperparameter	Independent	Controlled
$\beta_0 \in (0.1, 1)$	0.9	0.25
$\beta_T \in (\beta_0, 10)$	8.9	7.75
$\Delta t \in (0.005, 0.1)$	0.04	0.025
$h_{\text{v}} \in (0.5h, 2.5h)$	–	$2h$
Total steps	1000	500
Steps per v	–	20

finite differencing h, $h = \Delta t \cdot n_{\text{steps per v}}$, in the velocity estimations, allowing it to vary from $0.5h$ to $2.5h$. The chosen parameters can be found in Table 1. In Figure 12 we report the median best of 5 particles, showing essentially equal performance per elapsed real time. In Figure 13 we give the comparison in terms of gradient evaluations, where the controlled version outperforms the independent one. We note that our implementation is not designed for measuring the true computational efficiency of the proposed method on this potential. The medians showed in Figure 13 and Figure 12 are computed from 10 000 dense trajectories, each with 500 or 1000 timesteps, taking roughly four microseconds per solve on consumer grade hardware.

For the Rosenbrock problem, we found that the effect of estimating the flow was negligible compared to gradient descent dynamics. We expect that this is due to the Rosenbrock problem being a locally difficult problem, in the sense that its difficulty lies in its poor scaling and conditioning. However, the most important effect of the (true) velocity field is the global flow between modes. In contrast to the Rosenbrock problem, the Rastrigin potential is highly multimodal. In this case, we find that superimposing the velocity estimations yields a strong increase in performance for the Langevin version. Specifically, we find that solving a few transport

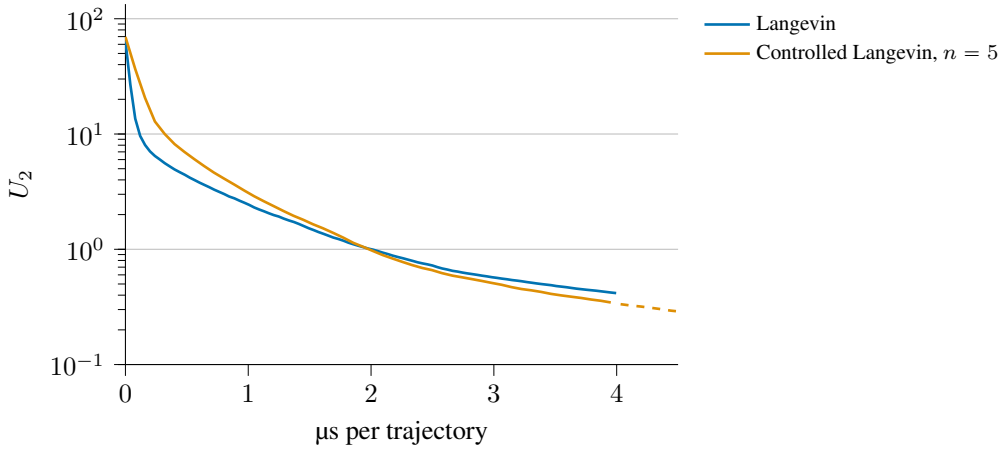


FIGURE 12. Comparison of performance on the Rastrigin potential U_2, measured in terms of the median of the cumulative best objective function value of 5 particles, plotted against equivalent wall-clock time. The cumulative median, computed from 50 000 total particle trajectories, show essentially equal performance. The simulation parameters were chosen based on a hyperparameter search procedure and can be found in Table 1. The dashed line represent steps taken outside the assigned computational budget, with parameters extrapolated from the search procedure, and are present for comparison with the results in Figure 13.

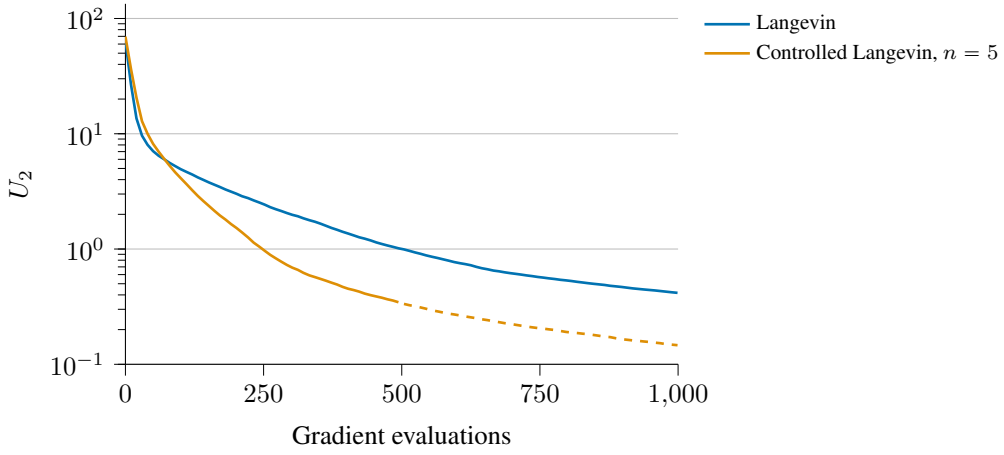


FIGURE 13. The median of the cumulative best Rastrigin function value of 5 particles, plotted against the number of gradient evaluations per particle. Here, the velocity field estimations leads to better performance with respect to the number of gradient steps. The simulation parameters were chosen based on a hyperparameter search procedure and can be found in Table 1. The dashed line represent steps taken outside the assigned computational budget, with parameters extrapolated from the search procedure.

problems along the way gave an improvement comparable to a tenfold increase in restarts, illustrating that there is some increased communication between modes as a result of the introduced interactions.

TABLE 2. In the left table we list the parameters for the system of ordinary differential equations which defines the minigolf potential U_{golf}. We solve the differential equations using the JAX-based [69] differential equation library Diffrax [68], which also allows us to compute gradients out of the box. We show a contour plot of the resulting potential in Figure 14. The right table shows the hyperparameters for the diffusion based methods we apply to U_{golf}.

Potential parameters	Symbol	Value
Turbulence	θ_1	0.7
Friction	θ_2	0.2
Confinement	θ_3	0.3
Integration end time	t_{end}	5
Small number	ε	0.1
Initial position	y_{tee}	$-[2,1]^T$
Target position	y_{goal}	$[1,2]^T$
ODE solver	Tsit5	
Step size	Constant	0.05

Langevin hyperparameters		Independent	Controlled
Initial	β_0	0.5	0.5
Final	β_T	7.5	7.5
Step size	Δt	0.025	0.025
Total steps		1000	100
Initial distribution		$\mathcal{N}(-[1,1]^T, \frac{1}{4}I)$	
Population size		–	5
Steps per v		–	20

6.3. A minigolf potential

We consider now a low dimensional but computationally challenging potential, the evaluation of which involves integrating an ordinary differential equation. So, let $U_{\text{golf}}\colon \mathbb{R}^2 \to \mathbb{R}$ be the function which assigns to each initial momenta $p_0 = x$ a value which measures how close to stopping at a given goal point y_{goal} a particular system (q, p) is at time t_{end} in the following way

$$U_{\text{golf}}(x) = \frac{1}{2}\left|q(t_{\text{end}}, x) - y_{\text{goal}}\right|^2 + \frac{1}{2}\left|p(t_{\text{end}}, x)\right|^2 + \frac{1}{2}\left|x\right|^2.$$

For a suitable set of scalar paramaters θ_i, we define $q(t, x)$ and $p(t, x)$ as the solution to the initial value problem

$$\begin{aligned} \frac{dq}{dt} &= p, & q_0 &= y_{\text{tee}}, \\ \frac{dp}{dt} &= \theta_1 \frac{f_{\text{wind}}(t, q)}{|f_{\text{wind}}(t, q)| + \varepsilon} - \theta_2 p - \theta_3 q, & p_0 &= x, \end{aligned}$$

where f_{wind} as the vector field on $\mathbb{R}^2$ given by $f_{\text{wind}}(q_1, q_2) = -\left[\sin\left(\pi q_2\right), \cos\left(\pi q_1/2\right)\right]^T$. To compute the $q(t_{\text{end}}, x)$ and $p(t_{\text{end}}, x)$ we solve the preceding differential equations using the differential equation library Diffrax [68], which allows us to compute gradients out of the box. The integration parameters for the potential are presented in Table 2, and we present a contour plot of the resulting potential in Figure 14. We evaluate the 5 particle controlled Langevin diffusion and the independent version on U_{golf}. The results are shown in Figure 15. Performance is in this case relatively similar, with a slight edge to the controlled version. Importantly, since we differentiate the golf potential in reverse, through the ODE solver, the cost of the gradient evaluations leads to a nearly indistinguishable difference, where we find that the 50 optimal transport solves in this case leads to a one percent increase of the total elapsed runtime.

To go full circle, we, again, estimate the Wasserstein-2 distance for the empirical laws of our particle systems in Figure 16, showing that the 5 interacting particles seem to fail to track the Gibbs curve in this case. Since the transport problems are cheap in this case, and since they can be computed without differentiating the potential,

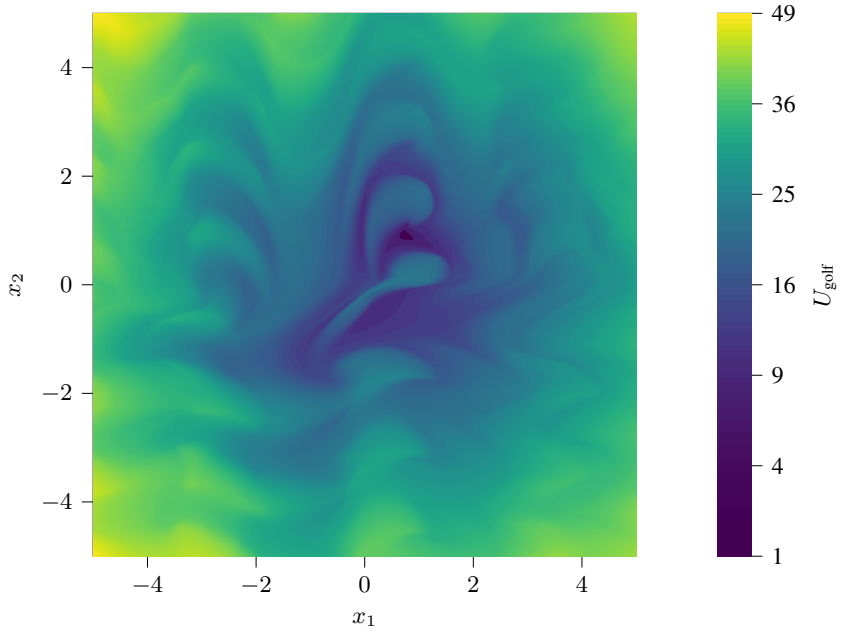


FIGURE 14. A contour plot for the potential U_{golf}, with the potential parameters specified in Table 2, reveals a complex energy landscape. Colors are normalized using a square root norm.

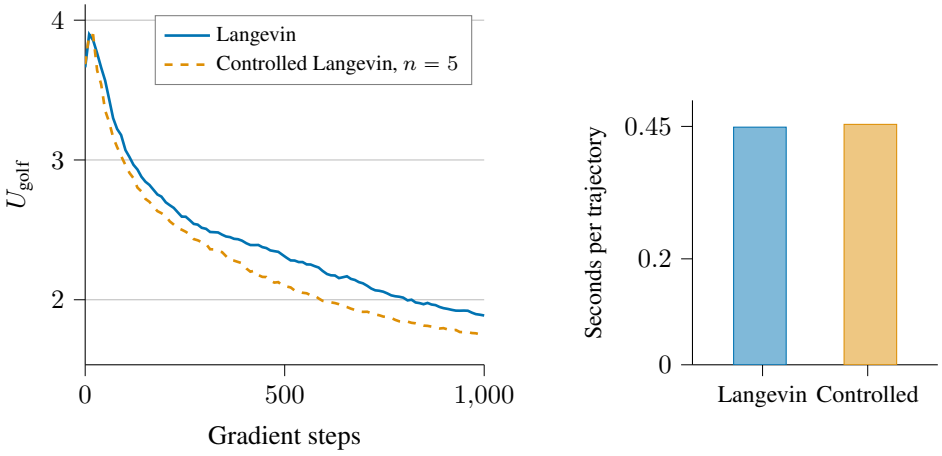


FIGURE 15. Median golf potential values for the Langevin and the controlled version with 5 particles, computed from 8000 trajectories using the parameters specified in Table 2. The average trajectory times, 0.449 seconds for the independent version and 0.454 when we superimpose 50 transport problems, differ by 1%. While the performance gain in this case is slight, the transport problems also have negligible effect on the elapsed wall-clock time.

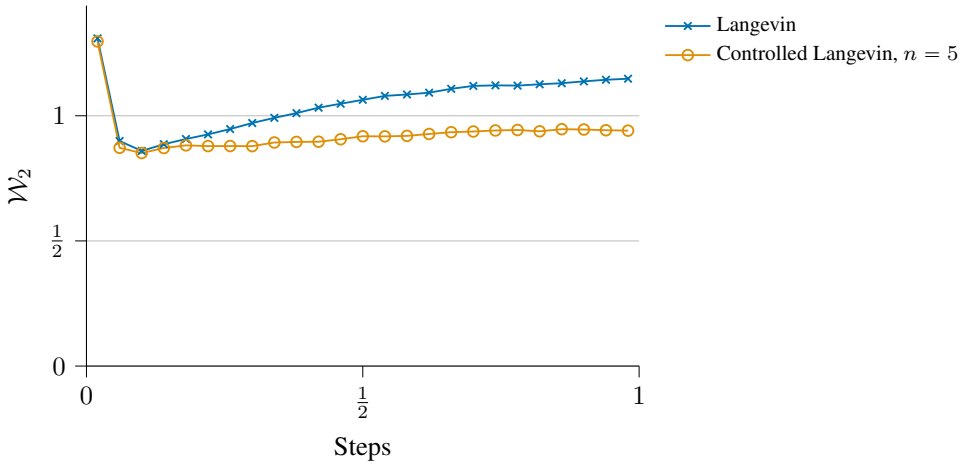


FIGURE 16. Estimated $\mathcal{W}_2$ distance to the golf Gibbs curve for the laws of the particles simulated according to the parameters in Table 2. The distances are approximated by solving the transport problem between histograms with a uniform 80 by 80 grid over $[-4, 4]^2$, against a ground truth histogram over the same grid. Histograms are computed as averages over 8000 trajectories, and the plot shows that the law of the controlled particles stays a bit closer, while it does not seem to approach the Gibbs curve.

and given the experiments on the other potentials, we believe it likely that there are parameters which leads to better performance on this potential.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and partially supported by the Chalmers AI Research (CHAIR) project "Stochastic Continuous-Depth Neural Networks". The authors would like to thank Adrien Corenflos for pointing out reference [30], and the anonymous reviewers for their numerous helpful comments that helped us improve the paper.

DECLARATIONS

CONFLICTS OF INTEREST

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.

DATA AVAILABILITY STATEMENT

The code associated with this article is available under the reference [70].

REFERENCES

[1] S. Kirkpatrick, C.D. Gelatt Jr. and M.P. Vecchi, Optimization by simulated annealing. *Science* **220** (1983) 671–680.
[2] P.J.M. Van Laarhoven and E.H.L. Aarts, Simulated annealing: Theory and applications. Springer, Dordrecht, The Netherlands (1987).
[3] A. Dekkers and E. Aarts, Global optimization and simulated annealing. *Math. Program.* **50** (1991) 367–393.
[4] S. Geman and C.-R. Hwang, Diffusions for global optimization. *SIAM J. Control Optim.* **24** (1986) 1031–1043.
[5] T.-S. Chiang, C.-R. Hwang and S.J. Sheu, Diffusion for global optimization in $\mathbb{R}^n$. *SIAM J. Control Optim.* **25** (1987) 737–753.

[6] D.P. Bertsekas, Nonlinear Programming, 2nd edn. Athena Scientific, Belmont, MA, USA (1999).
[7] J. Nocedal and S.J. Wright, Numerical Optimization, 2nd edn. Springer, New York, NY, USA (2006).
[8] R.A. Holley, S. Kusuoka and D.W. Stroock, Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.* **83** (1989) 333–347.
[9] C.-R. Hwang, Laplace's method revisited: weak convergence of probability measures. *Ann. Probab.* **8** (1980) 1177–1182.
[10] M. Hasenpflug, D. Rudolf and B. Sprungk, Wasserstein convergence rates of increasingly concentrating probability measures. *Ann. Appl. Probab.* **34** (2024) 3320–3347.
[11] D.W. Stroock and S.R. Srinivasa Varadhan, Multidimensional Diffusion Processes. Springer, Berlin, Heidelberg (2006).
[12] R.A. Holley and D.W. Stroock, Simulated annealing via Sobolev inequalities. *Commun. Math. Phys.* **115** (1988) 553–569.
[13] P. Monmarché, Piecewise deterministic simulated annealing. *ALEA Latin Am. J. Probab. Math. Stat.* **13** (2016) 357–398.
[14] L. Ambrosio, N. Gigli and G. Savaré, Gradient Flows in Metric Spaces and in the Space of Probability Measures. Birkhäuser, 2nd ed., (2008).
[15] R. Jordan, D. Kinderlehrer and F. Otto, The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29** (1998) 1–17.
[16] J. Bolte, L. Miclo and S. Villeneuve, Swarm gradient dynamics for global optimization: the mean-field limit case. *Math. Program.* **205** (2024) 661–701.
[17] P. Fearnhead, J. Bierkens, M. Pollock and G.O. Roberts, Piecewise deterministic Markov processes for continuous-time Monte Carlo. *Stat. Sci.* **33** (2018) 386–412.
[18] E. Zhou and X. Chen, Sequential Monte Carlo simulated annealing. *J. Global Optim.* **55** (2013) 101–124.
[19] Q. Liu and D. Wang, Stein variational gradient descent: a general purpose Bayesian inference algorithm. *Adv. Neural Inform. Process. Syst.* **29** (2016) 2378–2386.
[20] A. Garbuno-Inigo, F. Hoffmann, W. Li and A.M. Stuart, Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn. Syst.* **19** (2020) 412–441.
[21] A. Ringh and A. Sharma, Kalman–Langevin dynamics: exponential convergence, particle approximation and numerical approximation. arXiv preprint arXiv:2504.18139 (2025).
[22] J.A. Carrillo, Y.-P. Choi, C. Totzeck and O. Tse, An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.* **28** (2018) 1037–1066.
[23] J.A. Carrillo, F. Hoffmann, A.M. Stuart and U. Vaes, Consensus-based sampling. *Stud. Appl. Math.* **148** (2022) 1069–1140.
[24] D. Kalise, A. Sharma and M.V. Tretyakov, Consensus-based optimization via jump-diffusion stochastic differential equations. *Math. Models Methods Appl. Sci.* **33** (2023) 289–339.
[25] J. Heng, A. Doucet and Y. Pokern, Gibbs flow for approximate transport with applications to Bayesian computation. *J. Roy. Stat. Soc. B: Stat. Methodol.* **83** (2021) 156–187.
[26] Y. Tian, N. Panda and Y.T. Lin. Liouville Flow Importance Sampler. Proceedings of the 41st International Conference on Machine Learning in Proceedings of Machine Learning Research **235** (2024) 48186–48210.
[27] F. Vargas, S. Padhy, D. Blessing and N. Nüsken, Transport meets variational inference: Controlled Monte Carlo Diffusions, in *International Conference on Representation Learning*, vol. 2024, edited by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan and Y. Sun (2024) 55236–55278.
[28] M.S. Albergo and E. Vanden-Eijnden, Nets: a non-equilibrium transport sampler. Michael Samuel Albergo, Eric Vanden-Eijnden Proceedings of the 42nd International Conference on Machine Learning, *PMLR* **267** (2025) 1026–1055.
[29] S. Vaikuntanathan and C. Jarzynski, Escorted free energy simulations: improving convergence by reducing dissipation. *Phys. Rev. Lett.* **100** (2008) 190601.
[30] S. Reich, A nonparametric ensemble transform method for Bayesian inference. *SIAM J. Sci. Comput.* **35** (2013) A2013–A2024.
[31] M.D. Parno and Y.M. Marzouk, Transport map accelerated Markov chain Monte Carlo. *SIAM/ASA J. Uncertainty Quantif.* **6** (2018) 645–682.

[32] Y. Chen, T.T. Georgiou and M. Pavon, Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Rev.* **63** (2021) 249–313.
[33] M. Schauer, F. van der Meulen and H. van Zanten, Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli* **23** (2017) 2917–2950.
[34] M. Corstanje, F. van der Meulen and M. Schauer, Conditioning continuous-time Markov processes by guiding. *Stochastics* **95** (2023) 963–996.
[35] E. Bernton, J. Heng, A. Doucet and P.E Jacob, Schrödinger bridge samplers. arXiv preprint arXiv:1912.13170 (2019).
[36] A. Forsgren, P.E. Gill and M.H. Wright, Interior methods for nonlinear optimization. *SIAM Rev.* **44** (2002) 525–597.
[37] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, Cambridge, UK (2004).
[38] R.J. Vanderbei, Linear Programming, 3rd edn. Springer, Cham, Switzerland (2008).
[39] E.H. Lieb and M. Loss, Analysis, 2nd edn. American Mathematical Society, Providence, RI, USA (2001).
[40] C. Villani, Optimal Transport: Old and New. Springer Berlin Heidelberg (2008).
[41] C. Villani, Topics in Optimal Transportation. American Mathematical Society, Providence, RI, USA (2003).
[42] D.C. Dowson and B.V. Landau. The Fréchet distance between multivariate normal distributions. *J. Multivariate Anal.* **12** (1982) 450–455.
[43] C.R. Givens and R.M. Shortt, A class of Wasserstein metrics for probability distributions. *Michigan Math. J.* **31** (1984) 231–240.
[44] M. Knott and C.S. Smith, On the optimal mapping of distributions. *J. Optim. Theory Appl.* **43** (1984) 39–49.
[45] G. Peyré and M. Cuturi, Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** (2019) 355–607.
[46] A. Kazeykina, Z. Ren, X. Tan and J. Yang, Ergodicity of the underdamped mean-field Langevin dynamics. *Ann. Appl. Probab.* **34** (2024) 3181–3226.
[47] D. Bakry, F. Barthe, P. Cattiaux and A. Guillin, A simple proof of the Poincaré inequality for a large class of probability measures. *Electron. Commun. Probab.* **13** (2008) 60–66.
[48] D.G. Luenberger, Optimization by Vector Space Methods. John Wiley & Sons, Inc., New York, NY, USA (1969).
[49] G. Menz and A. Schlichting, Poincaré and lorarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.* **42** (2014) 1809–1884.
[50] O. Kallenberg, Foundations of Modern Probability, 3rd edn. Springer (2021).
[51] L.C.G. Rogers and D. Williams, *Diffusions, Markov Processes and Martingales: Itô Calculus*, Vol. 2. Cambridge University Press, Cambridge, UK (2000).
[52] V.I. Bogachev, N.V. Krylov, M. Röckner and S.V. Shaposhnikov, Fokker–Planck–Kolmogorov Equations. American Mathematical Society, Providence, RI, USA (2022).
[53] E.A.J.F. Peters and G. de With, Rejection-free Monte Carlo sampling for general potentials. *Phys. Rev. E* **85** (2012) 026703.
[54] A. Bouchard-Côté, S.J. Vollmer and A. Doucet, The bouncy particle sampler: a nonreversible rejection-free Markov chain Monte Carlo method. *J. Am. Stat. Assoc.* **113** (2018) 855–867.
[55] M.H.A. Davis, Markov Models & Optimization. Chapman & Hall, Boca Raton, FL, USA (1993).
[56] P. Holderrieth, Cores for piecewise-deterministic Markov processes used in Markov chain Monte Carlo. *Electron. Commun. Probab.* **26** (2021) 1–12.
[57] H. Kunita, Stochastic Flows and Stochastic Differential Equations. Cambridge University Press (1990).
[58] F. Baudoin, Diffusion Processes and Stochastic Calculus. European Mathematical Society (2014).
[59] C.-R. Hwang, *Frozen Patterns and Minimal Energy States*. PhD thesis, Brown University (1978).
[60] C.P. Robert and G. Casella, Monte Carlo Statistical Methods. Springer, New York, NY, USA (2004).
[61] N. Bonneel, M. Van De Panne, S. Paris and W. Heidrich, Displacement interpolation using Lagrangian mass transport, in *Proceedings of the 2011 SIGGRAPH Asia Conference* (2011) 1–12.
[62] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport. *Adv. Neural Inform. Process. Syst.* **26** (2013) 2292–2300.
[63] A.-A. Pooladian and J. Niles-Weed, Entropic estimation of optimal transport maps. arXiv preprint arXiv:2109.12004 (2021).
[64] S. Mehrotra, On the implementation of a primal-dual interior point method. *SIAM J. Optim.* **2** (1992) 575–601.

[65] S. Mizuno, M.J. Todd and Y. Ye, On adaptive-step primal-dual interior-point algorithms for linear programming. *Math. Oper. Res.* **18** (1993) 964–981.
[66] T. Guilmeau, E. Chouzenoux and V. Elvira, Simulated annealing: a review and a new scheme, in *2021 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE (2021) 101–105.
[67] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019).
[68] P. Kidger, *On Neural Differential Equations*. PhD thesis, University of Boxford (2021).
[69] J. Bradbury, R. Frostig, P. Hawkins, M.J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne and Q. Zhang, JAX: composable transformations of Python+NumPy programs (2018).
[70] V. Molin, Code for "Controlled stochastic processes for simulated annealing". https://github.com/vincentmolin/controlled_annealing (2025).

Appendix A. Deferred proofs

A.1 Proofs of Section 3.2

We verify that the Poincaré inequality used in the proof of Theorem 3.10 is bounded on $[0,T]$. We can convince ourselves that this is true via a Theorem in [47], but we repeat their construction of the Lyapunov function used for completeness.

Lemma A.1. *Let U satisfy Assumptions 3.1–3.3 and let $\beta\colon [0,T] \to (0,\infty)$. Then, for every t, μ_t satisfies a Poincaré inequality:*

$$\mathrm{Var}_{\mu_t}(f) = \int \left(f - \int f\, d\mu_t\right)^2 d\mu_t \leqslant C_P(\mu_t) \int |\nabla f|^2\, d\mu_t,$$

for all $f \in C_c^\infty(\mathbb{R}^d)$. Further, there exist positive constants A, B, K such that, for all $t \in [0,T]$

$$C_P(\mu_t) \leqslant C_t := A(1 + B\mathrm{e}^{\beta(t)K}).$$

Proof. By [47] we have that, if there exists a Lyapunov function $W \geqslant 1$ such that, for all x,

$$\mathcal{L}_t W(x) \leqslant -\theta W(x) + b\mathbf{1}_{B_R}(x)$$

then

$$\mathrm{Var}_{\mu_t}(f) \leqslant \frac{1}{\theta}(1 + b\kappa_t) \int |\nabla f|^2\, d\mu_t,$$

where κ_t is the Poincaré constant of μ_t restricted to the ball $B_R := \{x\colon |x| < R\}$ and $\theta = \gamma(\beta(t)\alpha - \gamma - \frac{n-1}{R}) > 0$ for some $\gamma > 0$. For any R, the following is a finite bound on the Poincaré constant of μ_t restricted to B_R

$$\kappa_t \leqslant \exp\left(\beta(t)\left(\sup_{x\in B_R} U(x) - \inf_{x\in B_R} U(x)\right)\right) R/\pi.$$

We couple the simple Lyapunov function $W(x) = |x|^2 + 1$ with Assumption 3.3. Then

$$\begin{aligned}
\mathcal{L}_t W(x) &= -2\langle \nabla U(x), x\rangle + 2n \\
&= \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|>R} \\
&\quad + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R} \\
&\leqslant \left(-2\alpha|x|^2 + 2n\right) I_{|x|>R} \\
&\quad + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R} \\
&= 2\left(-\alpha + \frac{\alpha+n}{|x|^2+1}\right) W(x) I_{|x|>R} \\
&\quad + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R} \\
&\leqslant 2\left(-\alpha + \frac{\alpha+n}{R^2+1}\right) W(x) I_{|x|>R} \\
&\quad + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R}.
\end{aligned}$$

Possibly increasing $\widetilde{R} := \max(R, \sqrt{\frac{n}{\alpha}} + 1)$ and setting $\theta = 2\left(\alpha - \frac{n+\alpha}{\widetilde{R}^2+1}\right) > 0$ we find

$$\begin{aligned}
\mathcal{L}_t W &\leqslant -\theta W(x) I_{|x|>R} + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R} \\
&= -\theta W(x)\left(1 - I_{|x|\leqslant R}\right) + \left(-2\langle \nabla U(x), x\rangle + 2n\right) I_{|x|\leqslant R} \\
&= -\theta W(x) + \left(\theta W(x) - 2\langle \nabla U(x), x\rangle + \frac{2n}{\beta(t)}\right) I_{|x|\leqslant \tilde{R}}
\end{aligned}$$

Then

$$b_t = \max_{|x|\leqslant R} \theta + \theta|x|^2 - 2\langle \nabla U(x), x\rangle + \frac{2n}{\beta(t)} < \infty.$$

Since $b_0 > b_t$ for all t we can pick $b := b_0$. With $K = \sup_{x\in B_R} U(x) - \inf_{x\in B_R} U(x)$ we then have

$$\operatorname{Var}_{\mu_t}(f) \leqslant \frac{1}{\theta}(1 + b\mathrm{e}^{\beta(t)K}\widetilde{R}/\pi)\int |\nabla f|^2 \, d\mu_t$$

for all t. □

Additionally, U admits moments of any order.

Lemma A.2 (Bounded variance of U)**.** *Assume that U satisfies Assumptions 3.1–3.3. Then, for all $\beta > 0$, U admits μ_β-moments of any order $n \geqslant 0$ in the sense that*

$$\int_{\mathbb{R}^d} (U(x))^n \pi_\beta(x)\, dx < \infty.$$

Specifically, it has bounded variance

$$\operatorname{Var}_{\mu_\beta}(U) = \int_{\mathbb{R}^d} \left(U(x) - \int U \, d\mu_\beta\right)^2 \pi_\beta(x)\, dx < \infty.$$

Proof. The moment-generating function of the random variable $U = U(X), X \sim \mu_\beta$ is given by

$$M_U(s) = \mathbb{E}_U\left[\mathrm{e}^{sU}\right] = \mathbb{E}_X\left[\mathrm{e}^{sU(X)}\right] = \int_{\mathbb{R}^d} \frac{\mathrm{e}^{(s-\beta)U(x)}}{Z_\beta}\, dx.$$

Since $\beta > 0$ there exists $\delta > 0$ such that $(s-\beta) < 0$ for $|s| < \delta$. By the proof of Lemma 3.8, $M_U(s)$ is then well-defined for all $s \in (-\delta, \delta)$ which guarantees the existence of all moments. □

A.2 Proofs of Section 4.3

Here, we establish that the Gibbs curve is absolutely continuous on $\mathbb{R}$. We first recall the definition of v_t in (4.15),

$$v_t(x) = -\frac{1}{\pi_t(x)} \int_{-\infty}^{x} \partial_t \pi_t(y)\, dy = \frac{\beta'(t)}{\pi_t(x)} \int_{-\infty}^{x} \left(U(y) - \int U\, d\mu_t \right) \pi_t(y)\, dy.$$

Theorem A.3. *Assume that $U\colon \mathbb{R} \to \mathbb{R}$ satisfies Assumption 3.1, 3.2 and 4.9. Let $\beta\colon [0,T] \to [1,\infty)$ be a non-decreasing smooth cooling schedule. Then, for any $1 \leqslant p < \infty$, v_t defined as in (4.15) satisfies $\|v_t\|_{L^p(\mu_t)} \in L^1([0,T])$ and the curve*

$$[0,T] \ni t \mapsto \mu_t \in \mathscr{P}_p(\mathbb{R})$$

is absolutely continuous.

The following elementary estimate controls the tails of v_t.

Lemma A.4. *Under Assumption 4.9 the following estimates hold*

$$x < -R \implies \int_{-\infty}^{x} U(z)\pi_t(z)\, dz \leqslant \frac{1}{\alpha}\pi_t(x)U(x),$$
$$x > R \implies \int_{x}^{\infty} U(z)\pi_t(z)\, dz \leqslant \frac{1}{\alpha}\pi_t(x)U(x).$$

Proof. Without loss of generality, assume that Assumption 4.9 holds with $\alpha = 1$. Then, since

$$\frac{d}{dx}(U\mathrm{e}^{-\beta U}) = \left(-\beta U' + \frac{U'}{U}\right) U\mathrm{e}^{-\beta U},$$

we find that, for $z < -R$, $\left(-\beta U' + \frac{U'}{U}\right) \geqslant 1$, so

$$\begin{aligned}\int_x^\infty U(z)\mathrm{e}^{-\beta U(z)}\, dz &\leqslant \int_{-\infty}^{x} \left(-\beta U'(z) + \frac{U'(z)}{U(z)}\right) U(z)\mathrm{e}^{-\beta U(z)}\, dz \\ &= U(x)\mathrm{e}^{-\beta U(x)}.\end{aligned}$$

Similarly, for $z > R$ we have that $-\left(-\beta U' + \frac{U'}{U}\right) \geqslant 1$ and hence

$$\begin{aligned}\int_x^\infty U(z)\mathrm{e}^{-\beta U(z)}\, dz &\leqslant -\int_{x}^{\infty} \left(-\beta U'(z) + \frac{U'(z)}{U(z)}\right) U(z)\mathrm{e}^{-\beta U(z)}\, dz \\ &= U(x)\mathrm{e}^{-\beta U(x)},\end{aligned}$$

and multiplying each side with Z_t^{-1} finishes the argument. □

Equipped with the preceding lemma, we are now ready to prove the main result of Section 4.3.

Proof of Theorem A.3. Writing $\bar{U}_t := \int U\, d\mu_t$ and $I_t(x) := \int_{-\infty}^{x} \left(U(z) - \bar{U}_t\right) \pi_t(z)\, dz$ so that $v_t(x) = \pi_t(x)^{-1} I_t(x)$. We note that $I(x) \to_{|x|\to\infty} 0$ and for sufficiently large $|x|$, $U(x) - \bar{U}_t > 0$. Hence

$$I'(x) = (U(x) - \bar{U}_t)\pi_t(x) > 0$$

so $I(x)$ is strictly increasing towards 0 when $x \to \infty$ and decreasing when $x \to -\infty$.

For $x > R$ we find then

$$\begin{aligned}
|I(x)| = -I(x) &= \int_{-\infty}^{x} \bar{U}_t \pi_t(z)\, dz - \int_{-\infty}^{x} U(z)\pi_t(z)\, dz \\
&= \bar{U}_t \int_{-\infty}^{x} \pi_t(z)\, dz - \left(\bar{U}_t - \int_{x}^{\infty} U(z)\pi_t(z)\, dz\right) \\
&= \bar{U}_t \left(\int_{-\infty}^{x} \pi_t(z)\, dz - 1\right) + \int_{x}^{\infty} U(z)\pi_t(z)\, dz \\
&= -\bar{U}_t \int_{x}^{\infty} \pi_t(z)\, dz + \int_{x}^{\infty} U(z)\pi_t(z)\, dz \\
&\leqslant \int_{x}^{\infty} U(z)\pi_t(z)\, dz \leqslant U(x)\pi_t(x),
\end{aligned}$$

where we applied Lemma A.4 in the last step. Similarly, for $x < -R$ we find that

$$\begin{aligned}
|I(x)| = I(x) &= \int_{-\infty}^{x} (U(z) - \bar{U}_t)\pi_t(z)\, dz \\
&\leqslant \int_{-\infty}^{x} U(z)\pi_t(z)\, dz \leqslant U(x)\pi_t(x).
\end{aligned}$$

Then $|x| > R \implies |v_t(x)| \leqslant |\beta'(t)| U(x)$ and

$$\begin{aligned}
\|v_t\|^p_{L^p(\mu_t,\mathbb{R})} &= \int_{\mathbb{R}} |v_t(x)|^p \pi_t(x)\, dx \\
&= |\beta'(t)|^p \int_{-R}^{R} (\pi_t(x))^{-p} |I(x)|^p \pi_t(x)\, dx + \int_{|x|>R} |v_t(x)|^p \pi_t(x)\, dx \\
&\leqslant |\beta'(t)|^p \underbrace{\left(\int_{-R}^{R} (\pi_t(x))^{-p} |I(x)|^p \pi_t(x)\, dx + \int_{|x|>R} (U(x))^p \pi_t(x)\, dx\right)}_{:=M_t} \\
&= |\beta'(t)|^p M_t < \infty.
\end{aligned}$$

As $t \mapsto M_t$ is continuous, $\sup_{0<t<T} M_t < \infty$ and the curve is absolutely continuous. □