Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A two-scale approach for sampling local free energy minimisers

Antoine Leclerc
Mines Saint-Étienne
Pierre Monmarché
pierre.monmarche@univ-eiffel.fr

LAMA, Université Gustave Eiffel

Research Article

Keywords: mean-field particle systems, Wasserstein gradient flow, sampling, local minimisers, free energy

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11216222/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A two-scale approach for sampling local free energy minimisers

Antoine Leclerc[1] and Pierre Monmarché[2,3*]

[1]Mines de Saint-Etienne, France, 158, cours Fauriel, Saint-Étienne, 42023 , France.
[2]LAMA, Université Gustave Eiffel, 5, boulevard Descartes, Marne-la-Vallée, 77454, France.
[3]Institut Universitaire de France.

*Corresponding author(s). E-mail(s): pierre.monmarche@univ-eiffel.fr;

Abstract

When a free energy admits several local minimisers, the associated mean-field particle system exhibits a metastable behavior, undergoing random transitions between these minimisers. However, when the number of particles is large, these transitions are rare events and typically occur at a time-scale which is out of reach for simulations. In this work, we introduce a system of particles coupled with a slow macroscopic external variable which approximately follows a Langevin dynamics associated to a coarse-grained free energy. The empirical distribution of the particle system thus samples all local minimisers of the initial free energy, with transition times which are now independent from the number of particles. The construction is inspired by a theoretical tool: the two-scale approach for proving log-Sobolev inequalities for Gibbs measures.

Keywords: mean-field particle systems, Wasserstein gradient flow, sampling, local minimisers, free energy

1 Introduction

1.1 Overview

Motivation.

Consider a so-called free energy

$$\mathcal{F}(\rho) = \mathcal{E}(\rho) + \sigma^2 \mathcal{H}(\rho)\,, \tag{1}$$

defined over $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures ρ with finite second moment, where $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ is some energy functional, $\mathcal{H}(\rho) = \int_{\mathbb{R}^d} \rho \ln \rho$ is the entropy (equal to $+\infty$ if ρ doesn't have a Lebesgue density) and $\sigma^2 > 0$ is a temperature parameter, which we fix equal to 1 without loss of generality from now on (unless otherwise specified, when discussing specific models, particularly for numerical experiments). We are interested in the situation where $\mathcal{F}$ has several local minimisers, and our goal is to sample them. One of the main interest is to find, among these local minimisers, the global ones, but finding all local minimisers is also of interest. Indeed, these are stable solutions of the gradient flow associated to $\mathcal{F}$, and their study is related for instance to questions of phase transitions, clustering, localization in statistical physics [5, 24, 48, 52]. The question of minimizing free energies over $\mathcal{P}_2(\mathbb{R}^d)$ also arises in machine learning [26, 28, 38], optimisation [14, 45], mathematical biology [11], variational inference [3, 30] and sampling [32]. For instance, clustering is an important feature in the attention mechanism used in transformers [26, 50].

The gradient flow of $\mathcal{F}$ (with $\sigma^2 = 1$) is the solution of

$$\partial_t \rho_t = \nabla \cdot (\rho_t D\mathcal{E}) + \Delta \rho_t \tag{2}$$

with the intrinsic derivative $D\mathcal{E}$ given in (20) below. Similarly to the finite-dimensional situation, a way to sample different local minimiser is to follow (2) starting from different random initial conditions (for instance normal distributions with random means and covariance matrices). However, in order to find several minimisers of a function $f \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, an alternative to randomized initialisations is to perturbate the gradient descent with a Brownian noise, leading to the so-called overdamped Langevin diffusion process, solving

$$\mathrm{d}X_t = -\nabla f(X_t)\mathrm{d}t + \sqrt{2/\beta}\mathrm{d}B_t\,, \tag{3}$$

with β^{-1} a temperature parameter and B a d-dimensional Brownian motion. Under suitable mild conditions on f, this process is ergodic with respect to the Gibbs measure with density proportional to $\exp(-\beta f)$. In particular, given enough time, it will pass near all local minimisers. When β is small, the process behaves like a Brownian motion, and thus in high dimension it will take a very long time to actually get near a given local minimiser. When β is large, the process is metastable, i.e. it spends most of its time in the vicinity of local minimisers, jumping from one to the other along some rare transitions. A situation where (3) is particularly relevant is when the set of local minimisers of interest are all in the vicinity of a manifold of dimension much smaller

than the ambient space $\mathbb{R}^d$. This is for instance the case in molecular simulations, where x is the position of nuclei and f is the energy of the system. In this situation, $d = 3n$ with n the number of atoms (ranging from 100 to 10^9 in typical applications), but the configurations of atoms are very structured (for instance the interatomic bonds in a protein are fixed and prevent high variations of the corresponding interatomic distances) so that the region of low energies, containing all stable configurations of interest, is concentrated around a manifold of dimension much smaller. However, no parametrization of this manifold is available. In particular, the idea of following the gradient flow of f starting from random initial conditions is not practicable, as no good initial distribution is known; when resorting to uninformed distributions (e.g. isotropic Gaussian), with overwhelming probability, the gradient flow get stuck in a completely non-physical configuration (i.e. a local minimiser f with an extremely large value with respect to $\inf f$). However, provided that a single reasonable x_0 in the low-energy region is already known, we can use (3) to explore locally the manifold, taking β large enough to stay in its vicinity and small enough to undergo the energy barriers within the manifold. From this simple strategy, multi-scale variations can be considered, for instance by using a trajectory at high temperature β^{-1} to give initial conditions for trajectories at smaller temperature, by running a single trajectory with a temperature schedule $t \mapsto \beta_t^{-1}$ that decays with time (simulated annealing or sequential Monte Carlo [13, 16, 29]) or by having several trajectories exchanging their temperatures at random times (parallel tempering or replica exchange [37, 43]).

In view of this discussion, our goal is to provide an analogue of (3) when $(f, \mathbb{R}^d)$ is replaced by $(\mathcal{F}, \mathcal{P}_2(\mathbb{R}^d))$. This requires to perturbate the gradient flow (2) with a stochastic noise.

Related literature.

This question has already been tackled with various approaches: using the theory of Dirichlet forms on metric spaces [21, 44, 46, 47, 51, 53], working in a space of functions in bijection with (a subset of) $\mathcal{P}(\mathbb{R}^d)$, for instance repartition functions in one dimension [17, 18, 53] or densities in some Hilbert space with a trace-class Gaussian distribution as the reference measure [10], or perturbating the infinite dimensional equation (2) by a finite-dimensional noise [19, 25]. There are various motivations for these stochastic perturbations: the goal might be to define a canonical reversible process with respect to the entropy (which would thus really be an infinite-dimensional analogue of the Langevin dynamics (3)), or to exploit the regularization properties of the noise to get well-posed equations, as in [17], or to induce ergodicity (i.e. uniqueness of a stationary distribution) in cases where the deterministic flow has several stationary solutions, as in [2, 19, 25, 36, 49].

As we will see, by some aspects, the viewpoint taken in the present work is close to the one in [19, 25] since we will only need a finite-dimensional Brownian noise (which is convenient for practical implementation) and, as motivated above, our primary goal is to ensure the ergodicity of the dynamics, in order to visit all local minimisers of the free energy. However, in [19, 25], the fact that the initial equation is a gradient flow doesn't play any role, and the measure which is sampled by the stochastic process is completely unknown. Noise is simply there to allow for exploration. If the noise

is small enough then the stochastic dynamics can be expected to be attracted by stable solutions of the deterministic equation when it gets close to it, but the behavior of the process outside the vicinity of these stable solutions is not controlled (hence, neither is the time spent on average far from these stable solutions). By contrast, our stochastic perturbation will be designed to target a specific equilibrium measure, highly relying on the gradient flow structure of (2) and the specific form of the free energy $\mathcal{F}$. For this reason, it is less versatile than the approach in [19, 25] but closer to the Langevin process (3). In particular, the time spent by the process in the vicinity of local minimisers and the transition times between different minimisers is controlled as in the Langevin case (3) (in terms of β).

Organisation.

The rest of this work is organised as follows. In the rest of the introduction, we introduce our two-scale sampler (first, by presenting the mean-field particle system associated to (2) in Section 1.2, then describing the two-scale approach and motivating the sampler in a simple case with quadratic interaction in Section 1.3, finally extending this definition to more general cases in Section 1.4), and compare it to the family of enhanced-sampling methods based on collective variables in Section 1.5. Some general notations are gathered in Section 1.6. Two approximations are made in the definition of the process, since a finite number N of particles and a positive time-scale separation parameter ε are used: the theoretical results describing the convergence of the process to the idealized sampler as $N \to \infty$ and $\varepsilon \to 0$ are stated in Section 2. Section 3 is devoted to numerical experiments. Finally, the proofs of the theoretical results are provided in Section 4.

1.2 Mean-field interacting particle system

The system of N interacting particles associated to the free energy $\mathcal{F}$ is $\mathbf{X}_t = (X_t^1, \dots, X_t^N)$ solving

$$\mathrm{d}\mathbf{X}_t = -\nabla U_N(\mathbf{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}_t, \tag{4}$$

with $\mathbf{B}$ a dN-Brownian motion and

$$U_N(\mathbf{x}) = N\mathcal{E}\left(\pi_{\mathbf{x}}\right), \qquad \pi_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}\,.$$

Under suitable conditions on $\mathcal{E}$, it is well-known that for all $t \geqslant 0$, $\pi_{\mathbf{X}_t}$ converges almost surely to the solution ρ_t of (2), provided this holds at $t = 0$. This is referred to as the mean-field limit (associated to the propagation of chaos phenomenon, according to which X_t^1 and X_t^2 are asymptotically independent as $N \to \infty$ provided this holds at time $t = 0$). The fluctuations of this convergence have been studied in [15, 22] (see [9] for further more recent works) and can be informally described by the fact that, in some weak sense,

$$\mathrm{d}\pi_{\mathbf{X}_t} \underset{N\to\infty}{\simeq} -\nabla_{\mathcal{W}_2}\mathcal{F}(\pi_{\mathbf{X}_t})\mathrm{d}t + \frac{1}{\sqrt{N}}\sqrt{\pi_{\mathbf{X}_t}}\mathrm{d}\xi_t$$

for some Gaussian noise ξ, where $-\nabla_{\mathcal{W}_2}\mathcal{F}$ stands for the right hand side of (2). At this point, we may think that we have achieved our goal since the empirical distribution of particles evolves according to a stochastic perturbation of the gradient flow. However, we need N to be large to have a good approximation of the gradient flow, and in that case the noise intensity is very small. Transitions between different local minimisers occur at a time-scale of order e^{cN} for some $c > 0$ [39], and thus are never seen in practical simulations.

By contrast, we are aiming at defining a process where the noise persists in the limit $N \to \infty$.

At a fixed N, $(\mathbf{X}_t)_{t\geqslant 0}$ is an overdamped Langevin diffusion and, under suitable conditions, it is ergodic with respect to the Gibbs measure

$$\mu_N \propto \exp(-U_N)\,.$$

Under this measure, $\pi_{\mathbf{X}}$ is concentrated around the minimisers of $\mathcal{F}$ for large N. However, when $\mathcal{F}$ has several local minimisers, sampling this measure with (4) requires a time which is exponentially large with N since we have to wait for $\pi_{\mathbf{X}_t}$ to undergo several rare transitions, and is thus unfeasible in practice. This can be interpreted in terms of the log-Sobolev constant of μ_N, which is the largest constant $\lambda_N > 0$ such that

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^{dN}), \qquad \mathcal{H}\left(\nu|\mu_N\right) \leqslant \frac{1}{2\lambda_N}\mathcal{I}(\nu|\mu_N) \tag{5}$$

(with the relative entropy and Fisher information given in (21)). The long-time convergence rate of (4) to μ_N in relative entropy is λ_N, see [4, Theorem 5.2.1]. When $\mathcal{F}$ has several local minimisers, it vanishes exponentially fast with N [39] (see [31] for further details on this topic in finite dimension).

1.3 Two-scale approach

Our work is motivated by the two-scale approach used in [6, 7, 39, 40] (see references within for previous works) for establishing log-Sobolev inequalities (LSI) (5) for μ_N (with λ_N independent from N in these works). It relies on a representation of μ_N in the spirit of De Finetti's theorem, i.e. as a mixture of tensorised measures (or at least simpler Gibbs measures for which a uniform-in-N LSI is known by some convexity argument [41, 54]). Let us describe this in the simple case where

$$\begin{aligned}\mathcal{E}(\rho) &= \int_{\mathbb{R}^d} V(x)\rho(x)\mathrm{d}x + \frac{1}{4}\int_{\mathbb{R}^d} |x-y|^2\rho(x)\rho(y)\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathbb{R}^d} V_1(x)\rho(x)\mathrm{d}x - \frac{1}{2}\left(\int_{\mathbb{R}^d} x\rho(x)\mathrm{d}x\right)^2 \end{aligned} \tag{6}$$

for some $V \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ and $V_1(x) = V(x) + \frac{1}{2}|x|^2$. Then,

$$U_N(\mathbf{x}) = \sum_{i=1}^N V_1(x_i) - \frac{1}{2N}\left(\sum_{i=1}^N x_i\right)^2 .$$

Relying on the formula for the moment generating function of Gaussian distribution,

$$\mathbb{E}\left(\exp\left(z\cdot G\right)\right)=\exp\left(\frac{\kappa}{2}|z|^2\right)\,,\qquad z\in\mathbb{R}^d,\ G\sim\mathcal{N}\left(0,\kappa^{-1}I_d\right),$$

applied with $z=\sum_{i=1}^N x_i$ and $\kappa=1/N$, we can write

$$\int_{\mathbb{R}^{dN}}\varphi(\mathbf{x})e^{-U_N(\mathbf{x})}\mathrm{d}\mathbf{x}=\mathbb{E}\left(\int_{\mathbb{R}^{dN}}\varphi(\mathbf{x})e^{-\sum_{i=1}^N[V_1(x_i)-G\cdot x_i]}\mathrm{d}\mathbf{x}\right)$$

for any $\varphi\in L^\infty(\mathbb{R}^{dN})$. This amounts to the representation

$$\mu_N=\int_{\mathbb{R}^d}\mu_\theta^{\otimes N}\nu_N(\theta)\mathrm{d}\theta \tag{7}$$

with

$$\mu_\theta(x)=\frac{1}{Z(\theta)}e^{-V_1(x)+\theta\cdot x}\,,\qquad Z(\theta)=\int_{\mathbb{R}^d}e^{-V_1(x)+\theta\cdot x}\mathrm{d}x\,,$$

and

$$\nu_N(\theta)\propto\exp\left(-N\omega(\theta)\right),\qquad \omega(\theta)=\frac{1}{2}|\theta|^2-\ln Z(\theta)\,.$$

In other words, to sample μ_N, we can sample $\theta\sim\nu_N$ and then generate i.i.d. samples $X^1,\dots,X^N$ with law μ_θ, which can be done with

$$\mathrm{d}X_t^i=-\nabla V_1(X_t^i)\mathrm{d}t+\theta\mathrm{d}t+\sqrt{2}\mathrm{d}B_t^i\,. \tag{8}$$

Since these particles are independent, for a fixed θ, the long-time convergence of these toward $\mu_\theta^{\otimes N}$ is independent from N and given by the log-Sobolev constant of μ_θ. Assuming for instance that V_1 is the sum of a strongly convex and a bounded functions, it follows by standard arguments that this LSI constant can be bounded from below uniformly over θ (see [4] or the proof of Proposition 2 below for details). In other words, sampling $\mu_\theta^{\otimes N}$ can be done in a time independent from θ and N, contrary to μ_N.

All the sampling difficulty is now encoded in θ, the critical points of ω corresponding to those of $\mathcal{F}$. Indeed, on the one hand,

$$\nabla\omega(\theta)=\theta-\int_{\mathbb{R}^d}x\mu_\theta(x)\mathrm{d}x\,.$$

On the other hand, by classical considerations (see e.g. [52]), for the energy (6), a density ρ_* is a stationary solution of (2) if and only if

$$\rho_*(x)\propto\left(-V_1(x)+m_*\cdot x\right),\qquad m_*=\int_{\mathbb{R}}x\rho_*(x)\mathrm{d}x\,, \tag{9}$$

namely $\rho_*=\mu_{m_*}$. Multiplying this equation by x and integrating shows that $m_*=\int_{\mathbb{R}^d}x\mu_{m_*}(x)\mathrm{d}x$, which means that m_* is a critical point of ω. Conversely, if m_* is a

critical point of ω, it is straightforward to check that μ_{m_*} satisfies the self-consistency equation (9), hence is a critical point of $\mathcal{F}$.

We can go a step further and in fact see that θ_* is a local minimiser of ω if and only if μ_{θ_*} is a local minimiser of $\mathcal{F}$, cf. [42, Section 3.1.2]. In other words, in the situation where $\mathcal{F}$ has several local minimisers, this will be the case for ω and then trying to sample ν_N with

$$\mathrm{d}\Theta_t = -\nabla\omega(\Theta_t)\mathrm{d}t + \sqrt{2/N}\mathrm{d}B_t \tag{10}$$

would be unpracticable for large N due to metastability [31].

As a conclusion of this discussion, we can say that the question of sampling the local minimisers of $\mathcal{F}$ boils down to sampling the Gibbs measure ν_N associated to the coarse-grained/macroscopic potential ω, and the difficulty is that the temperature $1/N$ is vanishing as N increases.

Our approach is then simple: it consists in sampling θ according to $\nu_\beta \propto \exp(-\beta\omega)$ with $\beta > 0$ fixed independent from N. We postpone to Section 1.5 a discussion about similar ideas in different contexts.

The remaining problem is that $\nabla\omega$ cannot be computed in practice. Hence, we resort to a stochastic gradient approximation. Indeed, we have seen that sampling μ_θ with i.i.d. particles (8) is easy (in the sense that the time needed is independent from N and θ; due to V_1 possibly being non-convex, this can still be long in some cases), and then we can estimate

$$\nabla\omega(\theta) \simeq \theta - \frac{1}{N}\sum_{i=1}^{N} X_t^i \tag{11}$$

In fact, we sample both the macroscopic parameter θ_t and the particles simultaneously, while adding a small speed parameter $\varepsilon > 0$ to the dynamics of θ_t. Indeed, if θ_t moves slowly enough, we expect the particles to reach equilibrium as if θ_t were constant, so that the approximation (11) remains valid (an alternative would have been to have the particles influenced by an exponential moving time-average of Θ_t, in the spirit of [35]).

As a conclusion, in the case of (6) with a quadratic interaction energy, the process that we consider in the present work is given by

$$\begin{aligned} \forall i \in [\![1, N]\!], \qquad \mathrm{d}X_t^i &= -\nabla V_1(X_t^i)\mathrm{d}t + \Theta_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i \\ \mathrm{d}\Theta_t &= -\varepsilon\left(\Theta_t - \frac{1}{N}\sum_{i=1}^{N} X_t^i\right)\mathrm{d}t + \sqrt{2\varepsilon/\beta}\mathrm{d}B_t\,. \end{aligned}$$

1.4 More general case

The previous discussion can be extended to more general situations. For instance, consider the energy

$$\mathcal{E}(\rho) = \int_{\mathbb{R}^d} V(x)\rho(x)\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^d} W(x,y)\rho(x)\rho(y)\mathrm{d}x\mathrm{d}y \tag{12}$$

with $V \in \mathcal{C}^1(\mathbb{R}^d)$, $W \in \mathcal{C}^1(\mathbb{R}^{2d})$ such that $W(x,y) = W(y,x)$ for all $x, y \in \mathbb{R}^d$. Under mild conditions, we can write a Mercer decomposition of W of the form

$$W(x,y) = W_0(x) + W_0(y) + \sum_{k \geqslant 0} m_k(x) m_k(y) - \sum_{k \geqslant 0} n_k(x) n_k(y)\,, \tag{13}$$

for some functions W_0, $(n_k, m_k)_{k \in \mathbb{N}}$. For instance, for $\alpha > 0$,

$$-\alpha e^{-\frac{(x-y)^2}{2}} = -\sum_{k \geqslant 0} \alpha e^{-\frac{x^2}{2}} \frac{x^k y^k}{k!} e^{-\frac{y^2}{2}}\,. \tag{14}$$

From the representation (13), the energy can be written

$$\mathcal{E}(\rho) = \int_{\mathbb{R}^d} V_1(x)\rho(x)\mathrm{d}x + \frac{1}{2}\sum_{k \geqslant 0}\left(\int_{\mathbb{R}} m_k(x)\rho(x)\mathrm{d}x\right)^2 - \frac{1}{2}\sum_{k \geqslant 0}\left(\int_{\mathbb{R}} n_k(x)\rho(x)\mathrm{d}x\right)^2\,,$$

with $V_1 = V + W_0$. The two first terms are convex as a function of ρ (along flat interpolations $t \mapsto (1-t)\rho_0 + t\rho_1$), and thus they will not be the problematic ones. Denoting $\varphi(x) = (n_k(x))_{k \geqslant 0}$, we get an expression of the form

$$\mathcal{E}(\rho) = \mathcal{E}_c(\rho) - \frac{1}{2}\|\rho(\varphi)\|_{\ell^2}^2\,,$$

with $\rho \mapsto \mathcal{E}_c(\rho)$ which is convex. The dependency of the last part as a function of ρ is quadratic since the energy (6) only involves pairwise interactions. A natural extension to more general interactions is given by energies of the form

$$\mathcal{E}(\rho) = \mathcal{E}_c(\rho) + R\left(\rho(\varphi)\right)\,,$$

with φ taking value in some Hilbert space $\mathbb{H}$ and $R : \mathbb{H} \to \mathbb{R}$. However, in this situation, a reasonable assumption is that the Hessian of R is lower bounded, meaning that $\psi \mapsto R(\psi) + L\|\psi\|^2$ is convex for some large L, and in that case we can write

$$\mathcal{E}_c(\rho) + R\left(\rho(\varphi)\right) = \widetilde{\mathcal{E}}_c(\rho) - \frac{1}{2}\|\rho(\tilde{\varphi})\|^2\,,$$

with a convex $\widetilde{\mathcal{E}}_c$ and $\tilde{\varphi}(x) = \sqrt{L}\varphi$, and we are back to the previous form.

As a conclusion, the general form of energy that we wish to cover is

$$\mathcal{E}(\rho) = \mathcal{E}_0(\rho) - \frac{1}{2}\|\rho(\varphi)\|^2 \tag{15}$$

for some function $\varphi : \mathbb{R}^d \to \mathbb{H}$ with $\mathbb{H}$ some Hilbert space, and $\mathcal{E}_0$ is some energy that we can think as convex or at least not too concave to cause difficulties, as we will state precisely later on (specifically, see (30)).

In that case, the associated mean-field potential is

$$U_N(\mathbf{x}) = N\mathcal{E}(\pi_{\mathbf{x}}) = V_N(\mathbf{x}) - \frac{1}{2N}\left(\sum_{i=1}^N \varphi(x_i)\right)^2$$

with $V_N(\mathbf{x}) = N\mathcal{E}_0(\pi_{\mathbf{x}})$. Reasoning as in the quadratic interaction case (possibly with some subtleties if the dimension of $\mathbb{H}$ is infinite, see [40], but we won't consider this case afterwards), we get for the Gibbs measure $\mu_N \propto \exp(-U_N)$ the representation

$$\mu_N = \int_{\mathbb{H}} \mu_\theta^N \nu_N(\theta)\mathrm{d}\theta\,,$$

with

$$\mu_\theta^N = \frac{1}{Z_N(\theta)} e^{-V_N(\mathbf{x})+\theta\cdot\sum_{i=1}^N \varphi(x_i)}\,, \qquad Z_N(\theta) = \ln \int_{\mathbb{R}^{dN}} e^{-V_N(\mathbf{x})+\theta\cdot\sum_{i=1}^N \varphi(x_i)}\mathrm{d}\mathbf{x}$$

(which in general is not a tensor product if $\mathcal{E}_0$ is not linear) and

$$\nu_N \propto \exp\left(-N\omega_N(\theta)\right)$$

with

$$\omega_N(\theta) = \frac{1}{2}|\theta|^2 - \frac{1}{N}\ln Z_N(\theta)\,. \tag{16}$$

From

$$\nabla\omega_N(\theta) = \theta - \int_{\mathbb{R}^{dN}} \pi_{\mathbf{x}}(\varphi)\mu_\theta^N(\mathbf{x})\mathrm{d}\mathbf{x}\,, \tag{17}$$

proceeding as in the quadratic case, we end up with the coupled system

$$\begin{cases} \forall i \in [\![1,N]\!], & \mathrm{d}X_t^i = -\nabla_{x_i} V_N(\mathbf{X}_t)\mathrm{d}t + \Theta_t\cdot\nabla\varphi(X_t^i)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i \\ & \mathrm{d}\Theta_t = -\varepsilon\left(\Theta_t - \frac{1}{N}\sum_{i=1}^N \varphi(X_t^i)\right)\mathrm{d}t + \sqrt{2\varepsilon/\beta}\mathrm{d}B_t\,. \end{cases} \tag{18}$$

The introduction of this process, which can be referred to as a two-scale sampler, is our main contribution. Its theoretical and empirical study is the main topic of the rest of this work.

In practice we only consider the finite dimensional case $\mathbb{H} = \mathbb{R}^m$ for some $m \geqslant 1$. From an infinite-dimensional situation, we have to consider an orthogonal decomposition $\varphi = (\varphi_1, \varphi_2) \in \mathbb{H}_1 \oplus \mathbb{H}_2$ with $\mathbb{H}_1$ of finite dimension and then the process is (18) but with φ replaced by φ_1 and V_N replaced by $N\mathcal{E}_1(\pi_{\mathbf{x}})$ with $\mathcal{E}_1(\rho) = \mathcal{E}_0(\rho) - \frac{1}{2}\|\rho(\varphi_2)\|^2$. In that case, $\mathcal{E}_1$ might not be convex but, provided $\|\varphi_2\|_\infty$ is small enough for instance, it will still have good properties for our purpose (see [41] or Assumption 2 below).

1.5 Relation to collective variable-based enhanced sampling

Metastability is a central difficulty for sampling in molecular simulations [33]. On the other hand, in this field, the problem is typically multi-scale, with fast oscillations of

the nuclei positions $x \in \mathbb{R}^d$ and slow motion of macroscopic configurations (described by much less degrees of freedom than d). The context is thus very similar to the one we are facing in our mean-field settings. In this situation, over the last twenty years, a broad family of adaptive biasing methods have been developed for molecular simulations, based on collective variables. A collective variable is described by a function $\xi : \mathbb{R}^d \to \mathbb{R}^m$ with $m \ll d$, often $m \leqslant 3$. Ideally, it should be chosen so that the value of $\xi(x)$ describes which macroscopic configuration the microscopic configuration x belongs to (i.e. $\xi(x)$ is an order parameter). In other words, denoting by U the energy of the system (so that we aim at sampling the Gibbs measure with density $\mu_\beta \propto \exp(-\beta U)$ with β the inverse temperature), the value of $\xi(x)$ should ideally characterise the local minimiser of U whose basin of attraction (for the gradient flow) contains x. Initially, the definition of suitable collective variables were based on expert knowledge. Over recent years, the question of their unsupervised learning has been a very active research field, see e.g. [8] and references within.

Now, suppose that we are given a collective variable ξ. Assuming that $\xi(X)$ has a density ν_β when $X \sim \mu_\beta$, we write $A(\psi) = -\beta^{-1} \ln \nu_\beta(\psi)$. This function is called the free energy associated to ξ. If the collective variable is "good" (whatever this precisely means), then the different modes of μ_β should give different modes of ν_β. In other words, the multimodality of μ_β should transfer to ν_β, and the local minimisers of A should correspond to stable macroscopic configurations. Moreover, if the different modes of μ_β are separated in ν_β, then there should not be any remaining multimodality in the conditional law of X given $\xi(X) = \psi$, whatever $\psi \in \mathbb{R}^m$. In this situation, the main issue is to sample correctly $\xi(X)$ (which is difficult since, relying on a standard Langevin dynamics for X, transitions of $\xi(X)$ between modes are rare events due to metastability), while sampling the conditional distributions is easy. This situation is very similar to the one discussed in Section 1.3 in our mean-field situation.

In this context, a variety of adaptive importance methods are based on the idea to target a measure $\tilde{\mu}_\beta$ such that:

1. The conditional law of X given $\xi(X) = \psi$, for any $\psi \in \mathbb{R}^m$, is the same under μ_β and $\tilde{\mu}_\beta$.
2. The marginal distribution of $\xi(X)$ when $X \sim \tilde{\mu}_\beta$ is $\nu_{\beta_1} \propto \exp(-\beta_1 A)$ for some $\beta_1 < \beta$.

At the end of the simulation, expectations with respect to μ_β are then computed by importance reweighting of samples distributed according to $\tilde{\mu}_\beta$. The motivation of the definition of $\tilde{\mu}_\beta$ is that increasing the temperature β_1^{-1} reduces the transition time between modes. In some applications, $\xi(x)$ is in a compact torus (for instance when it is given by some interatomic angles), and then we may simply take $\beta_1 = 0$ (i.e. $\xi(X)$ is sampled uniformly). Moreover, this tempering is only done on a small-dimensional marginal distribution, instead of simply sampling μ_{β_1}, which would be a reference measure very far from the target distribution μ_β due to the high dimension d, leading to a large variance of the importance sampling weights and a poor result.

Standard algorithms with wide-spread applications which target ν_β are for instance the adaptive biasing force (ABF) method, metadynamics or OPES, see the discussion and references in [32, Section 1.2].

As we see, the two-scale sampler introduced in the present work follows exactly the same objective as these methods. We aim at sampling $(\mathbf{X}, \Theta)$ distributed according to the density $\mu_\theta^N(\mathbf{x})\nu_N(\theta)$, but it is multi-modal and the corresponding Langevin dynamics is very metastable. As discussed in Section 1.3, in good situations, the multi-modality is entirely encoded in the marginal distribution ν_N of the collective variable $\xi(\mathbf{X}, \Theta) := \Theta$. Hence, we target a modified measure where the conditional law of $\mathbf{X}$ given $\xi(\mathbf{X}, \Theta) = \theta$ is unchanged (this is still μ_θ^N) while the collective variable Θ is sampled at a higher temperature (β^{-1} instead of N^{-1}).

However, if we naively apply standard collective variable-based enhanced sampling methods to our situation, then the initial target marginal distribution ν_N gets steeper and steeper as $N \to \infty$, and thus the adaptive drift required to bias it to ν_β with a fixed β becomes larger and larger, leading to unstable numerical schemes. The reason why this doesn't occur for the two-scale sampler (18) is mainly due to the stochastic gradient approximation (11), which explicitly rely on the mean-field asymptotic structure of the problem and on the time-scale separation between the macroscopic and microscopic evolutions.

1.6 Some definitions and notations

We write $\mathcal{P}_p(\mathbb{R}^d)$ the set of probability measures with finite p^{th} moment, and $\mathcal{W}_p$ the associated Wasserstein distance.

A linear functional derivative of $\mathcal{E} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ is a measurable function $\frac{\delta \mathcal{E}}{\delta m} : \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that, for all $\mu_1, \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ with $\mathcal{E}(\mu_0) + \mathcal{E}(\mu_1) < \infty$,

$$\mathcal{E}(\mu_1) - \mathcal{E}(\mu_0) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta \mathcal{E}}{\delta m}(t\mu_1 + (1-t)\mu_0, x)(\mu_1 - \mu_0)(\mathrm{d}x)\mathrm{d}t\,. \tag{19}$$

If it exists, it is unique up to an additive constant. If $x \mapsto \frac{\delta \mathcal{E}}{\delta m}(\mu, x)$ is $\mathcal{C}^1$ for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then we call

$$D\mathcal{E}(\mu, x) := \nabla_x \frac{\delta \mathcal{E}}{\delta m}(\mu, x) \tag{20}$$

the intrinsic derivative of $\mathcal{E}$.

For two probability measures ν, μ on $\mathbb{R}^d$ with $\nu \ll \mu$, the relative entropy and Fisher information of ν with respect to μ are defined by

$$\mathcal{H}(\nu|\mu) = \int_{\mathbb{R}^d} \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\mathrm{d}\nu\,, \qquad \mathcal{I}(\nu|\mu) = \int_{\mathbb{R}^d} \left|\nabla \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\mu}\right)\right|^2 \mathrm{d}\nu\,, \tag{21}$$

where $|\nabla \ln(\mathrm{d}\nu/\mathrm{d}\mu)|$ is a classical gradient norm if $\mathrm{d}\nu/\mathrm{d}\mu$ is differentiable and is always defined in general as an upper gradient (see [1, Definition 1.2.4]). If ν has no density with respect to μ, $\mathcal{H}(\nu|\mu) = \mathcal{I}(\nu|\mu) = +\infty$.

2 Theoretical results

There are two approximations between the scheme (18) and the target macroscopic dynamics: the particle one (depending on N) and the finite-speed one (depending on ε). To check the consistency of the method, we provide successively basic convergence guarantees for both errors.

2.1 Propagation of chaos

At fixed $\varepsilon > 0$, we expect propagation of chaos to hold conditionally to the macroscopic noise B as $N \to \infty$, so that $(\pi_{\mathbf{X}_t}, \theta_t)$ should converge to

$$\begin{cases} \partial_t \nu_t &= \nabla \cdot \left(\nu_t \left[D\mathcal{E}_0(\nu_t, \cdot) - \bar{\Theta}_t \cdot \nabla\varphi\right]\right) + \Delta\nu_t \\ \mathrm{d}\bar{\Theta}_t &= -\varepsilon\left(\bar{\Theta}_t - \nu_t(\varphi)\right) + \sqrt{2\varepsilon/\beta}\mathrm{d}B_t\,, \end{cases} \tag{22}$$

assuming that $\pi_{\mathbf{X}_0} \to \nu_0$, which for instance holds if $\mathbf{X}_0 \sim \nu_0^{\otimes N}$ (which we assume in Theorem 1 to have a simpler statement). We work under these regularity conditions:

Assumption 1

1. The intrinsic derivative $D\mathcal{E}_0$ exists. There exists $L > 0$ such that, first, $x \mapsto D\mathcal{E}_0(\nu, x)$ is L-one-sided Lipschitz continuous uniformly in ν, meaning that
$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d),\ \forall x, y \in \mathbb{R}^d, \qquad (x-y)\cdot(D\mathcal{E}_0(\nu,x) - D\mathcal{E}(\nu,y)) \leqslant L|x-y|^2\,, \tag{23}$$
and, second, $\mu \mapsto D\mathcal{E}_0(\mu, x)$ is L-Lipschitz continuous uniformly in x, i.e.
$$\forall \nu, \mu \in \mathcal{P}_2(\mathbb{R}^d),\ \forall x \in \mathbb{R}^d, \qquad |D\mathcal{E}_0(\nu,x) - D\mathcal{E}_0(\mu,x)| \leqslant L\mathcal{W}_2(\nu,\mu)\,. \tag{24}$$
2. The parameter $\varphi \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}^m)$ is Lipschitz continuous, and so is $\nabla\varphi$.

Under these regularity conditions, well-posedness for (18) and (22) follows from classical fixed-point arguments. Applying then a standard synchronous coupling arguments, we obtain the following (conditional) mean-field limit:

Theorem 1 *Under Assumption 1, fix a Brownian motion B on $\mathbb{R}^m$, an initial distribution $\nu_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p > 2$ and $\theta_0 \in \mathbb{R}^m$. Denote by $(\nu_t, \bar{\Theta}_t)_{t\geqslant 0}$ the corresponding solution of (22) (with $\bar{\Theta}_0 = \theta_0$). For all $N \geqslant 1$, let $\mathbf{X}_0 \sim \nu_0^{\otimes N}$ (independent from B), let $B^1, \dots, B^N$ be N independent d-dimensional Brownian motions independent from B and $\mathbf{X}_0$, and denote by $(\mathbf{X}_t, \Theta_t)_{t\geqslant 0}$ the corresponding solutions of (18) (with $\Theta_0 = \theta_0$). Then, for all $T > 0$ and $\delta > 0$,*

$$\sup_{t\in[0,T]} \mathbb{P}\left(\mathcal{W}_2(\pi_{\mathbf{X}_t}, \nu_t) + |\Theta_t - \bar{\Theta}_t| \geqslant \delta\right) \underset{N\to 0}{\longrightarrow} 0\,.$$

If we consider a convergence conditionally to the sigma algebra generated by B, then a more quantitative result is stated in Remark 1 below.

2.2 Averaging

In this section, with respect to (22), we accelerate time by a factor $1/\varepsilon$. For a fixed Brownian motion B and an initial condition (ν_0, Θ_0), we consider for all $\varepsilon > 0$ the process

$$\begin{cases} \partial_t \nu_t^\varepsilon &= \frac{1}{\varepsilon} \nabla \cdot (\nu_t^\varepsilon \left[D\mathcal{E}_0(\nu_t^\varepsilon, \cdot) - \Theta_t^\varepsilon \cdot \nabla \varphi \right]) + \frac{1}{\varepsilon} \Delta \nu_t^\varepsilon \\ \mathrm{d}\Theta_t^\varepsilon &= -\Theta_t^\varepsilon + \nu_t^\varepsilon(\varphi) + \sqrt{2/\beta} \mathrm{d}B_t \,. \end{cases} \tag{25}$$

We are interested in the limit of this process as ε vanishes. We expect to observe an averaging phenomenon in the nice situation where, for any $\theta \in \mathbb{R}^m$, the equation

$$\partial_t \rho_t = \nabla \cdot (\rho_t \left[D\mathcal{E}_0(\rho_t, \cdot) - \theta \cdot \nabla \varphi \right]) + \Delta \rho_t \tag{26}$$

admits a unique stationary solution μ_θ, globally attractive. In this case, as $\varepsilon \to 0$, Θ^ε should converge to the solution of

$$\mathrm{d}\Theta_t^0 = -\Theta_t^0 + \mu_{\Theta_t^0}(\varphi) + \sqrt{2/\beta} \mathrm{d}B_t \,, \tag{27}$$

with $\Theta_0^0 = \Theta_0$, while ν_t^ε should converge to $\mu_{\Theta_t^0}$. This is the content of Theorem 3. Before stating it, let us motivate its assumptions and explain why (27) is a Langevin process associated to some coarse-grained energy ω (as in (10), but now with temperature β^{-1} instead of N^{-1}), without trying to state precise formal results in general settings (we refer to [20, 27, 34, 39] for further details, and particularly to [40, Section 6.1]).

Since (26) is the Wasserstein gradient flow of the free energy

$$\mathcal{F}_\theta(\nu) = \mathcal{E}_0(\nu) - \theta \cdot \nu(\varphi) + \mathcal{H}(\nu) \,,$$

it admits a unique globally attractive stationary solution if $\mathcal{F}_\theta$ admits a unique global minimiser and satisfies a so-called Polak-?ojasiewicz (PL) inequality, namely if there exists $\lambda_\theta > 0$ such that

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d), \qquad \widetilde{\mathcal{F}}_\theta(\nu) \leqslant \frac{1}{2\lambda_\theta} \mathcal{I}_\theta(\nu) \,, \tag{28}$$

with

$$\widetilde{\mathcal{F}}_\theta(\nu) = \mathcal{F}_\theta(\nu) - \inf \mathcal{F}_\theta$$

and the free energy dissipation

$$\mathcal{I}_\theta(\nu) := \int_{\mathbb{R}^d} \left| \nabla \ln \nu + D\mathcal{E}_0(\nu) - \theta \cdot \nabla \varphi \right|^2 \nu$$

(equal to $+\infty$ if ν doesn't have a Lebesgue density). From Laplace-Varadhan lemma, as $N \to \infty$, the macroscopic potential ω_N in (16) converges to

$$\omega(\theta) = \frac{1}{2}|\theta|^2 - \inf_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}_\theta \,,$$

see e.g. [34, (3.30)]. Moreover, assuming suitable conditions, under the Gibbs measure μ_θ^N, $\pi_{\mathbf{x}}$ almost surely converges to μ_θ. As a consequence, letting $N \to \infty$ in (17) gives

$$\nabla\omega(\theta) = \theta - \mu_\theta(\varphi)\,, \tag{29}$$

meaning that (27) is indeed a Langevin process with potential ω and temperature β^{-1}.

As in Section (1.3), the critical points of ω (i.e. the fixed points of $\theta \mapsto \mu_\theta(\varphi)$) are in one-to-one correspondence with the stationary solutions of (2) (by $\theta \leftrightarrow \mu_\theta$), with local minimisers of ω corresponding to local minimisers of $\mathcal{F}$. In particular, $\mu_{\Theta_t^0}$ does sample efficiently these local minimisers (since the Gibbs density proportional to $e^{-\beta\omega}$ concentrates on the local minimisers of ω).

By contrast, when $\mathcal{F}_\theta$ has several critical points, then most of the nice properties listed above fail, see e.g. [39]. The limit behavior of (25) is then more intricate and we won't try to study this bad situation, as the algorithm is precisely not designed to work then (all the metastability should be incorporated in the macroscopic dynamics of Θ_t).

As a conclusion of this discussion, we work under the following conditions, which, as stated in Proposition 2 ensures that we are in the good situation discussed above.

Assumption 2

1. The energy $\mathcal{E}_0$ is semi-lower continuous and $D\mathcal{E}_0$ exists.
2. There exists $\eta \in [0, 1/4)$ such that for all $\nu, \nu' \in \mathcal{P}_2(\mathbb{R}^d)$ and $t \in [0, 1]$
$$\mathcal{E}_0(t\nu + (1-t)\nu') \leqslant t\mathcal{E}_0(\nu) + (1-t)\mathcal{E}_0(\nu') + t(1-t)\eta\|\nu - \nu'\|_{TV}^2\,. \tag{30}$$
3. There exists $c_0, C_0 > 0$ such that for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,
$$\mathcal{E}_0(\nu) \geqslant c_0 \int_{\mathbb{R}^d} |x|^2 \nu(\mathrm{d}x) - C_0\,. \tag{31}$$
4. There exist $\kappa, M, L > 0$ such that for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\frac{\delta\mathcal{E}_0}{\delta m}(\nu, \cdot)$ is the sum of three functions: a κ-strongly convex one, a M-bounded one and a L-Lipschitz one.
5. The parameter can be decomposed as $\varphi = (\varphi_\ell, \varphi_b) \in \mathbb{R}^{m_\ell} \times \mathbb{R}^{m_b}$ (possibly with $m_\ell = 0$ or $m_b = 0$) with φ_ℓ linear and φ_b bounded.

Proposition 2 *Under Assumption 2, there exists $\lambda > 0$ depending only on κ, L, M, η such that for all $\theta = (\theta_\ell, \theta_b) \in \mathbb{R}^{m_\ell} \times \mathbb{R}^{m_b}$, $\mathcal{F}_\theta$ admits a unique global minimiser $\mu_\theta \in \mathcal{P}_2(\mathbb{R}^d)$ and satisfies the PL inequality (28) with*

$$\lambda_\theta \geqslant \lambda e^{-|\theta_b|\|\varphi_b\|_\infty}\,. \tag{32}$$

The main result of this section is then the following:

Theorem 3 *Under Assumption 2, assume furthermore that $|\Theta_0|^2$ and $\mathcal{F}_{\Theta_0}(\nu_0)$ have finite expectation and that φ is Lipschitz continuous. Then, for any $T > t_0 > 0$ and $\delta > 0$,*

$$\sup_{t\in[0,T]} \mathbb{P}\left(|\Theta_t^\varepsilon - \Theta_t^0| \geqslant \delta\right) + \sup_{t\in[t_0,T]} \mathbb{P}\left(\mathcal{W}_2(\nu_t^\varepsilon, \mu_{\Theta_t^0}) \geqslant \delta\right) \underset{\varepsilon\to 0}{\longrightarrow} 0\,. \tag{33}$$

In the case where φ is linear (i.e. $m_b = 0$), we get a more quantitative result, cf. Remark 2. Moreover, in the case $m_b \neq 0$, this stronger result still holds for a truncated version of the process, given by (56), which can possibly be of interest in practice in order to sample the equilibria μ_θ with θ constrained in a ball.

2.3 Example

Consider an energy $\mathcal{E}$ with pairwise interaction of the form (12), where W admits a decomposition (13). For some $m \in \mathbb{N}$, set

$$\varphi(x) = (n_k(x))_{k \in [\![0,m]\!]} ,$$

and

$$\mathcal{E}_0(\rho) = \rho(V_1) + \frac{1}{2} \sum_{k \geqslant 0} (\rho(m_k))^2 - \frac{1}{2} \sum_{k > m} (\rho(n_k))^2 . \tag{34}$$

Then the energy $\mathcal{E}$ is decomposed as (15). In fact we can write

$$\mathcal{E}_0(\rho) = \int_{\mathbb{R}^d} V(x)\rho(\mathrm{d}x) + \frac{1}{2} \int_{\mathbb{R}^d} W_1(x,y)\rho(\mathrm{d}x)\rho(\mathrm{d}y),$$
$$W_1(x,y) = W(x,y) + \sum_{k=0}^{m} n_k(x) n_k(y) . \tag{35}$$

The conditions on φ required in Assumptions 1 and 2 are easily checked on specific models. For instance, for quadratic interaction as in Section 1.3, $\varphi(x) = ax$ for some $a \in \mathbb{R}$, hence φ is linear and a fortiori Lipschitz continuous (and so is $\nabla\varphi$). For the attractive Gaussian interaction (14)

$$n_k(x) = \sqrt{\alpha} e^{-\frac{x^2}{2}} \frac{x^k}{\sqrt{k!}} ,$$

so that φ is bounded with bounded derivatives of all orders.

We now turn to the conditions on $\mathcal{E}_0$. First,

$$D\mathcal{E}_0(\rho, x) = \nabla V_1(x) + \sum_{k \geqslant 0} \nabla m_k(x) \rho(m_k) - \sum_{k > m} \nabla n_k(x) \rho(n_k) .$$

Besides, it can be simpler to check the conditions by writing

$$D\mathcal{E}_0(\rho, x) = D\mathcal{E}(\rho, x) + \nabla\varphi(x) \cdot \rho(\varphi(x))$$
$$= \nabla V_1(x) + \int_{\mathbb{R}^d} \nabla_x W(x,y)\rho(\mathrm{d}y) + \nabla\varphi(x) \cdot \rho(\varphi) . \tag{36}$$

Assume that ∇V_1 satisfies a one-sided Lipschitz condition (23) (which is the case for instance if V_1 is convex outside a compact set). With the conditions we checked on φ,

we already have that $x \mapsto \nabla\varphi(x)\cdot\rho(\varphi)$ satisfies this condition (for a linear function f, $(\nabla f(x)-\nabla f(y))\rho(f)$ is zero, and for a bounded function f with second order derivative bounded, $|\nabla f(x)-\nabla f(y)||\rho(f)| \leqslant \|\nabla^2 f\|_\infty \|f\|_\infty |x-y|$). Then, if $\nabla_x^2 W$ is bounded (which is for instance the case for the Gaussian interaction (14)), the remaining term in (36) is $\|\nabla_x^2 W\|_\infty$-Lipschitz in x and finally (23) holds.

Similarly, assuming that $\nabla_{x,y}^2 W$ is bounded,

$$|D\mathcal{E}_0(\nu,x) - D\mathcal{E}_0(\mu,x)| \leqslant \left(\|\nabla_{x,y}^2 W\|_\infty + \|\nabla\varphi\|_\infty^2\right)\mathcal{W}_2(\nu,\mu)\,,$$

so that (24) holds.

Under the conditions discussed above, we have thus already checked the conditions in Assumption 1. Concerning Assumption 2, the lower-bounded (31) holds for instance if, in (35), we can decompose $W_1(x,y) = -|A(x-y)|^2 + W_2(x,y)$ for some matrix A and some bounded function W_2 (which corresponds to the assumption we have already made on the functions n_k, m_k) and if $V(x) \geqslant c_0|x|^2 + |Ax|^2 - C$ for some constants $c_0, C > 0$. Moreover, with this form,

$$\frac{\delta\mathcal{E}_0}{\delta m}(\rho,x) = V(x) - \int_{\mathbb{R}^d} |A(x-y)|^2\rho(\mathrm{d}y) + \int_{\mathbb{R}^d} W_2(x,y)\rho(\mathrm{d}y)\,.$$

Assume that $V = V_c + V_b$ where $x \mapsto V_c(x) - |Ax|^2$ is κ-strongly convex and V_b is M-bounded for some $\kappa, M > 0$. Then, for all $\rho \in \mathcal{P}_2(\mathbb{R}^d)$, $\frac{\delta\mathcal{E}_0}{\delta m}(\rho,x)$ is the sum of $V_c(x) - \int_{\mathbb{R}^d}|A(x-y)|^2\rho(\mathrm{d}y)$, which is κ-strongly convex as a function of x, and of $V_b + \int_{\mathbb{R}^d} W_2(\cdot,y)\rho(\mathrm{d}y)$, which is $M + \|W_2\|_\infty$-bounded. The fourth condition of Assumption 2 is thus satisfied.

Finally, since the two first terms in the right hand side of (34) are (flat) convex as a function of ρ, they do not contribute when checking (30), and we see that for all $\nu,\nu' \in \mathcal{P}_2(\mathbb{R}^d)$ and $t \in [0,1]$,

$$\begin{aligned}\mathcal{E}_0(t\nu + (1-t)\nu') - t\mathcal{E}_0(\nu) - (1-t)\mathcal{E}_0(\nu') &\leqslant t(1-t)\sum_{k>m}\left(\nu(n_k) - \nu'(n_k)\right)^2 \\ &= t(1-t)\sum_{k>m}\left(\mathbb{E}\left[n_k(Y) - n_k(Y')\right]\right)^2\,,\end{aligned}$$

with (Y,Y') a coupling of (ν,ν'). Taking an optimal coupling for the total variation distance, in the sense that $\|\nu-\nu'\|_{TV} = 2\mathbb{P}(Y \neq Y')$,

$$\begin{aligned}\sum_{k>m}\left(\mathbb{E}\left[n_k(Y) - n_k(Y')\right]\right)^2 &\leqslant \frac{1}{4}\|\nu-\nu'\|_{TV}^2 \sum_{k>m}\left(\mathbb{E}\left[n_k(Y) - n_k(Y') \mid Y \neq Y'\right]\right)^2 \\ &\leqslant \frac{1}{4}\|\nu-\nu'\|_{TV}^2 \mathbb{E}\left[\sum_{k>m}\left(n_k(Y) - n_k(Y')\right)^2 \,\middle|\, Y \neq Y'\right] \\ &\leqslant \|\nu-\nu'\|_{TV}^2 \left\|\sum_{k>m} n_k^2\right\|_\infty\,.\end{aligned}$$

This shows that (30) holds with

$$\eta = \left\| \sum_{k>m} n_k^2 \right\|_\infty .$$

If we can take m large enough so that $\eta < 1/2$ (which is the case in particular when $\sum_{k\geqslant 0} \|n_k\|_\infty^2 < \infty$) then Assumption 2 is satisfied.

3 Numerical experiments

In this section we consider models where the temperature σ^2 in (1) is not necessarily equal to 1. When implementing the coupled process (18), we apply it to the free energy $\sigma^{-2}\mathcal{F} = \sigma^{-2}\mathcal{E} + \mathcal{H}$, so that in (15) φ has to be replaced by φ/σ; we then rescale Θ into Θ/σ and ε into ε/σ^2, so that the final rescaled process reads

$$\begin{cases} \forall i \in [\![1, N]\!], & \mathrm{d}X_t^i = -\sigma^{-2}\nabla_{x_i} V_N(\mathbf{X}_t)\mathrm{d}t + \Theta_t \cdot \nabla\varphi(X_t^i)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i \\ & \mathrm{d}\Theta_t = -\varepsilon\left(\sigma^2\Theta_t - \frac{1}{N}\sum_{i=1}^N \varphi(X_t^i)\right)\mathrm{d}t + \sqrt{2\varepsilon/\beta}\mathrm{d}B_t \,. \end{cases} \tag{37}$$

In all experiments, it is simulated using an Euler–Maruyama scheme with step-size $h = 0.01$.

3.1 Double-well potential

We first consider a Curie–Weiss model corresponding to the quadratic energy (9) in dimension 1 with

$$V(x) = \frac{x^4}{4} - \frac{x^2}{2}\,, \qquad W(x-y) = \frac{\lambda}{2}\,|x-y|^2\,, \quad \lambda = 1\,,$$

so that the feature reduces to $\varphi(x) = x$ (hence $m = 1$ and $\Theta_t \in \mathbb{R}$). There is a critical temperature $\sigma_c^2 = 0.68^2 \approx 0.46$ such that: for $\sigma^2 > \sigma_c^2$ the free energy $\mathcal{F}$ admits a unique minimiser, a symmetric density centred at 0; while for $\sigma^2 < \sigma_c^2$ it admits two minimisers μ_+ and μ_-, symmetric of one another, given by the self-consistency equation (9) and with opposite barycentres $\pm m_*$ (numerically, $m_* \approx 0.79$ at the sub-critical temperature $\sigma^2 = 0.25$ used in the experiments below). Throughout, the reference measures $\mu_\pm$ are obtained by relaxing a standard Langevin dynamics (4) from Gaussian initial conditions $\mathcal{N}(\pm 1, 0.5)$.

The two stationary regimes.

Before applying the two-scale sampler, we illustrate the two stationary regimes relying on the plain mean-field particle system (4), that is, the standard interacting Langevin dynamics associated with $\mathcal{F}$ *without* the macroscopic variable Θ. For N particles it

reads

$$\mathrm{d}X_t^i = -V'(X_t^i)\mathrm{d}t - \lambda\left(X_t^i - \bar{X}_t\right)\mathrm{d}t + \sqrt{2\sigma^2}\,\mathrm{d}B_t^i\,, \qquad \bar{X}_t = \frac{1}{N}\sum_{j=1}^{N} X_t^j\,, \tag{38}$$

for $i \in [\![1, N]\!]$, where $V'(x) = x^3 - x$, the interaction strength is $\lambda = 1$, and σ^2 is the temperature. We integrate (38) with an Euler–Maruyama scheme of step $h = 0.01$, running it independently from two i.i.d. initial distributions, $X_0^i \sim \mathcal{N}(+1, 0.5)$ and $X_0^i \sim \mathcal{N}(-1, 0.5)$, with $N = 5000$ particles up to $T = 100$. For each run we record the final empirical distribution of the particles and the barycentre trajectory $t \mapsto \bar{X}_t$.

The outcome depends on whether the temperature lies below or above σ_c^2. In the sub-critical regime $\sigma^2 = 0.25 < \sigma_c^2$ (Figure 1), the two runs relax towards two distinct stationary solutions μ_+ and μ_-, symmetric of one another, whose barycentres settle at $\approx +0.79$ and ≈ -0.79: the initial condition alone selects which of the two minimisers is reached, and a trajectory started near one of them remains there. In the super-critical regime $\sigma^2 = 0.64 > \sigma_c^2$ (Figure 2), both runs relax, whatever their starting point, to the same symmetric solution centred at 0. As we see, only the sub-critical case corresponds to the metastable situation we are interested in: below σ_c^2 the free energy has several minimisers, and the plain dynamics cannot move from one to another on a reasonable time scale. We focus on this regime in the rest of the section.

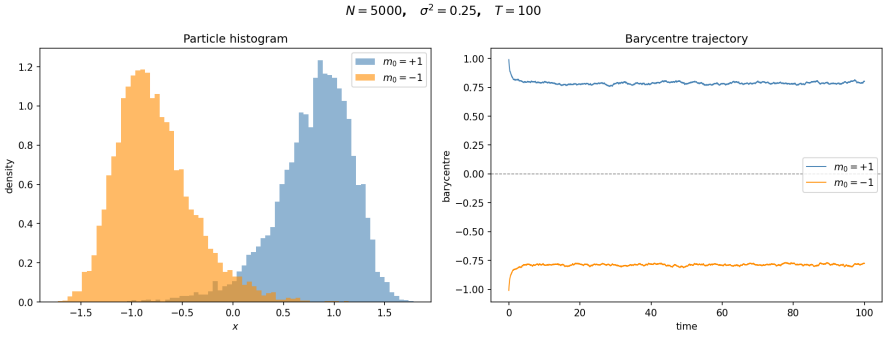


Fig. 1 Sub-critical regime, $\sigma^2 = 0.25 < \sigma_c^2$. Uncoupled system (38) with $N = 5000$, $T = 100$, started from $\mathcal{N}(\pm 1, 0.5)$. Left: final histograms of the particles for the two initial conditions $m_0 = \pm 1$, concentrating around the two stationary solutions $\mu_\pm$. Right: the two barycentre trajectories, settling at $\pm m_* \approx \pm 0.79$.

Transitions, Wasserstein distance and the barycentre as a proxy.

We now turn on the coupling between the particles and an additional macroscopic variable, simulating the two-scale system (18) for the double well. Taking $\varphi(x) = x$ (so $\Theta_t \in \mathbb{R}$) and $V_N = \sum_i V_1(X_t^i)$ with $V_1 = V + x^2/2$, and using $V_1'(x) = x^3$, the

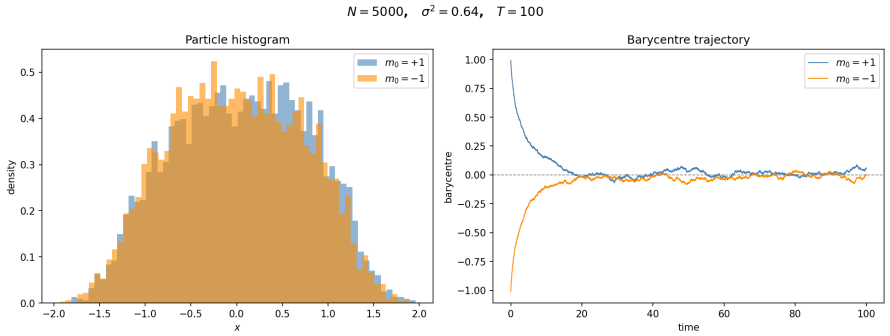


Fig. 2 Super-critical regime, $\sigma^2 = 0.64 > \sigma_c^2$. Same setting as Figure 1. From either initial condition the particles relax to the single stationary solution centred at 0, and the two barycentre trajectories converge to 0.

coupled system reads

$$\begin{aligned} \mathrm{d}X_t^i &= -\frac{(X_t^i)^3}{\sigma^2}\,\mathrm{d}t + \Theta_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^i\,, \\ \mathrm{d}\Theta_t &= -\varepsilon\left(\sigma^2\Theta_t - \bar{X}_t\right)\,\mathrm{d}t + \sqrt{2\varepsilon/\beta}\,\mathrm{d}W_t\,, \end{aligned} \tag{39}$$

for $i \in [\![1, N]\!]$, where the mean-field interaction of the previous paragraph is now *replaced* by the slow variable Θ_t, which tracks the barycentre $\bar{X}_t$ at speed ε. We integrate (39) with the same Euler–Maruyama scheme, and compare the particles to the two reference solutions μ_+ and μ_- obtained as above. To locate the transitions we use two indicators: the quadratic Wasserstein distances $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_\pm)$ from the empirical distribution of the particles to the references (which we approximate by the empirical distribution of the particles from Figure 1), and the barycentre $\bar{X}_t$ itself. Recall that, for two probability measures μ, ν on $\mathbb{R}^d$,

$$\mathcal{W}_2(\mu,\nu)^2 = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2\,\gamma(\mathrm{d}x,\mathrm{d}y)\,,$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν; in dimension $d = 1$, for two empirical measures carried by N points, this reduces to the sorted L^2 distance $\mathcal{W}_2(\pi_{\mathbf{x}}, \pi_{\mathbf{y}})^2 = \frac{1}{N}\sum_{i=1}^N (x_{(i)} - y_{(i)})^2$, with $x_{(1)} \leqslant \cdots \leqslant x_{(N)}$ and $y_{(1)} \leqslant \cdots \leqslant y_{(N)}$ the order statistics.

Figure 3 shows a run at $\sigma^2 = 0.25$, $\varepsilon = 0.07$, $\beta = 0.4$, $N = 3000$, $T = 500$, displaying four quantities: the histograms of the particles conditioned on the current basin (top left), the barycentre $\bar{X}_t$ (top right), the trajectory of Θ_t (bottom left), and the two distances $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_\pm)$ (bottom right). Here conditioning on the current basin means that, at each recorded time, the population is assigned to μ_+ or μ_- according to the sign of the barycentre $\bar{X}_t$, and the particle positions are pooled separately for the two cases; the two resulting histograms are then compared to the references $\mu_\pm$. Driven by Θ_t, the population moves back and forth between the two basins, spending here about 68% of the time in μ_+ and 32% in μ_-. The two conditional histograms only *resemble* the references $\mu_\pm$ and are noticeably broader: pooling every recorded

time spent in a basin describes the fluctuations (at temperature β^{-1}, not vanishing as $N \to \infty$) around the minimiser. The two $\mathcal{W}_2$ distances exchange cleanly at each transition. Crucially, the barycentre $\bar{X}_t$ carries exactly the same transition information as the pair of $\mathcal{W}_2$ distances, at a negligible cost: this validates the barycentre as a cheap proxy for detecting transitions, which is relevant in high dimension where $\mathcal{W}_2$ becomes prohibitive to compute.

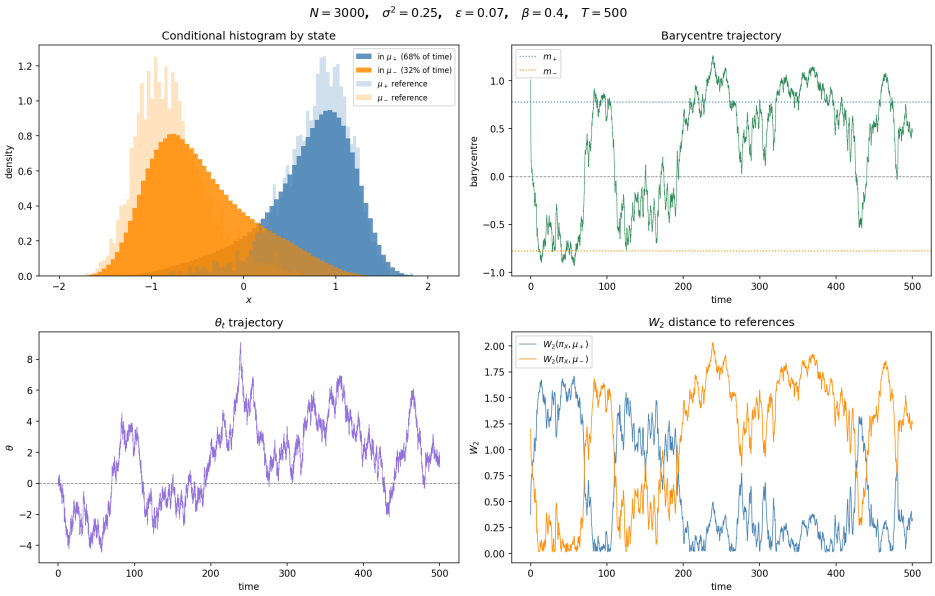


Fig. 3 Coupled two-scale dynamics (39) at $\sigma^2 = 0.25$, $\varepsilon = 0.07$, $\beta = 0.4$, $N = 3000$, $T = 500$. Top left: histograms of the particles conditioned on the current basin, compared to the references $\mu_\pm$ (light). Top right: barycentre $\bar{X}_t$, with the reference levels $m_\pm$. Bottom left: trajectory of Θ_t. Bottom right: distances $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_\pm)$. The barycentre and the $\mathcal{W}_2$ distances signal the same transitions.

Influence of β on the transitions.

The macroscopic variable Θ_t follows a Langevin dynamics at inverse temperature β in the coarse-grained potential: the noise intensity $\sqrt{2\varepsilon/\beta}$ decreases with β, so β controls the frequency of the transitions. Figure 4 compares three runs of (39) at fixed $\varepsilon = 0.03$, $N = 1500$, $T = 500$, $\sigma^2 = 0.25$, for $\beta \in \{0.05, 0.6, 2\}$. Each row shows the trajectory of Θ_t (left), its histogram together with the fraction of time spent on each side of $\Theta = 0$ (centre), and the distances $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_\pm)$ to the two references (right). On the Θ axis the two basins are separated by $\Theta = 0$ and centred at the deterministic fixed points $\theta^*_\pm = m_\pm/\sigma^2 \approx \pm 3.1$.

At small $\beta = 0.05$ the macroscopic noise is strong and Θ_t crosses very frequently between the two basins, visiting both in comparable proportion. As β increases the crossings become progressively rarer: at $\beta = 0.6$, and even more at $\beta = 2$, the trajectory dwells for long stretches in one basin and only switches occasionally. The number

of transitions thus decreases as β increases — the expected metastability/temperature trade-off, now carried by the one-dimensional variable Θ_t rather than by the full particle system — so that β tunes how aggressively the sampler explores the different minimisers.



Fig. 4 Influence of β at fixed $\varepsilon = 0.03$, $N = 1500$, $T = 500$, $\sigma^2 = 0.25$. From top to bottom: $\beta = 0.05$, 0.6, 2. Each row shows the trajectory of Θ_t (left), its histogram together with the fraction of time spent on each side of $\Theta = 0$ (centre), and the distances $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_\pm)$ (right). The number of transitions decreases as β increases.

Long-time distribution of Θ_t versus the target ν_β.

The previous runs show Θ_t crossing between the basins; we now check that its long-time law is the one predicted by the averaging principle. When ε is small, Θ_t evolves slowly enough for the particles to equilibrate at each frozen value of Θ, and its stationary law should be the coarse-grained Gibbs measure $\nu_\beta \propto e^{-\beta\omega}$. In the present quadratic case ω is explicit and we evaluate it by numerical quadrature,

$$\omega(\theta) = \frac{\sigma^2}{2\lambda}\,\theta^2 - \ln \int_{\mathbb{R}} \exp\left(-\frac{1}{\sigma^2}\left(V(x) + \frac{\lambda}{2}x^2\right) + \theta x\right) \mathrm{d}x\,,$$

with $\lambda = 1$. We integrate (39) over $M = 20$ independent trajectories, with i.i.d. initial conditions $X_0^i \sim \mathcal{N}(-1, 0.5)$ and $\Theta_0 = \bar{X}_0$, pool all the recorded values of Θ_t into a single histogram (the several trajectories improve the statistics, and the horizon is long enough for Θ_t to visit both basins), and compare it to ν_β. Figure 5 does this at $\sigma^2 = 0.3$, $\beta = 0.5$, $N = 1000$, for two values of the speed parameter, $\varepsilon = 0.05$ (horizon $T = 10^4$) and $\varepsilon = 0.1$ (horizon $T = 10^3$). In both cases the empirical distribution matches the target ν_β, and the agreement improves as ε decreases (at horizon $T \propto 1/\varepsilon$), as expected since the averaging approximation becomes exact in the limit $\varepsilon \to 0$.

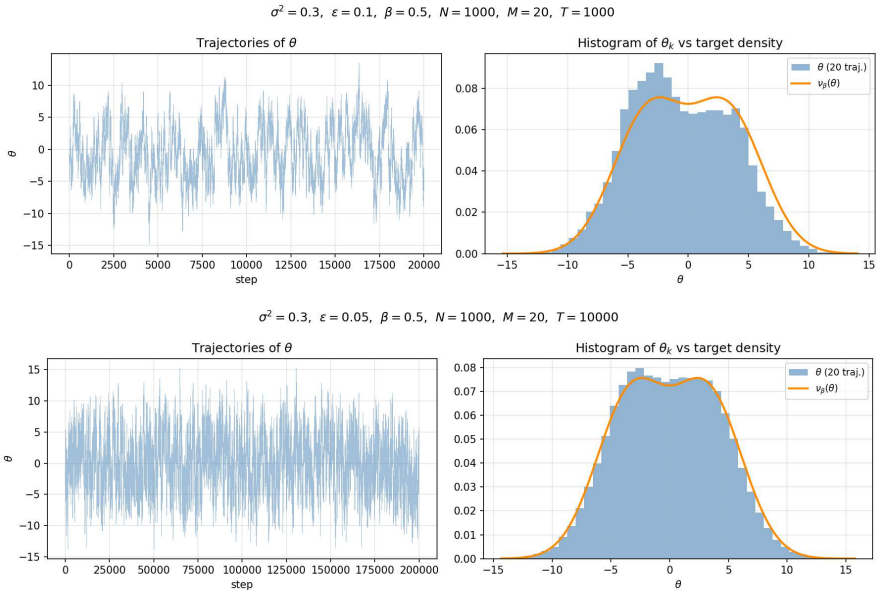


Fig. 5 Long-time distribution of Θ_t at $\sigma^2 = 0.3$, $\beta = 0.5$, $N = 1000$, aggregated over $M = 20$ trajectories. Top: $\varepsilon = 0.05$, $T = 10^4$. Bottom: $\varepsilon = 0.1$, $T = 10^3$. Left: the M trajectories of Θ_t, concatenated one after another. Right: histogram of Θ_t against the target density $\nu_\beta \propto e^{-\beta\omega}$.

3.2 Multi-well potential

We keep the quadratic interaction of Section 3.1 (feature $\varphi(x) = x$, so $\Theta_t \in \mathbb{R}$, and $\lambda = 1$), but replace the confining potential by the multi-well

$$V(x) = \cos(x) + \cos\left(\tfrac{x}{2}\right) + \frac{x^2}{100}\,.$$

The difference with the double well is one of complexity rather than of nature. In the double-well case the potential V already has two wells, and, below the critical temperature σ_c^2, the free energy has exactly two symmetric minimisers $\mu_\pm$, whereas above σ_c^2 it has a single one — the whole picture being governed by a single critical temperature. Here V has many wells of unequal depth, so the coarse-grained

potential ω can have many local minima and the free energy $\mathcal{F}$ several distinct localised stationary solutions μ_{θ^*}. On the range $[-30, 30]$ the wells of V sit at $x \approx \pm 3.57,\ \pm 8.73,\ \pm 15.89,\ \pm 20.97,\ \pm 28.21$ (ten wells), and more appear on a wider range. This is the genuinely multi-modal situation the two-scale sampler is designed for.

The specific choice of V is deliberate. The two cosines $\cos x$ and $\cos(x/2)$ have different periods (2π and 4π), so their sum is not a regular array of identical wells: the wells have *unequal depths* and *unequal spacings*, and are therefore separated by barriers of varying height and width. The quadratic term $x^2/100$ plays two roles: its growth at infinity makes the Gibbs measure well-defined, while its very small coefficient leaves a large number of wells within reach — ten on $[-30, 30]$, sixteen on $[-60, 60]$ — with the outer wells lying progressively higher and hence carrying less mass. This yields a stringent, non-symmetric test: the sampler must discover many minimisers, cross heterogeneous barriers, and recover the correct relative weights of wells of different depths. It also mirrors the molecular-simulation setting of the introduction, where a weakly confined low-energy region contains many metastable states of unequal importance. Moreover, although we know that there is a single free energy minimizer at high temperature and several localized ones at low temperature, the heterogeneity of the potential is such that we expect different localized mimizers to appear at different critical temperatures. For a given σ^2, the number of local minimizers is thus not clear at all, and this makes the sampling algorithm particularly useful: contrary to the double-well case considered in Section 3.1, if we wanted to find the minimizers by initializing standard particle systems (4) with various starting distributions, we wouldn't know where to look and when to stop.

Exploring the wells.

We simulate the coupled system (18) with this V (same convention as (39), with $V_1'(x) = -\sin x - \frac{1}{2}\sin\frac{x}{2} + \frac{x}{50} + x$):

$$\begin{aligned} \mathrm{d}X_t^i &= -\frac{V_1'(X_t^i)}{\sigma^2}\,\mathrm{d}t + \Theta_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^i\,, \\ \mathrm{d}\Theta_t &= -\varepsilon\left(\sigma^2\Theta_t - \bar{X}_t\right)\,\mathrm{d}t + \sqrt{2\varepsilon/\beta}\,\mathrm{d}W_t\,. \end{aligned} \tag{40}$$

Since ω now has several minima, Θ_t is metastable and hops between them, and the population μ_{Θ_t} correspondingly visits the associated stationary solutions. To identify which solution the particles currently sit in, we use a *fork descent*: at regular intervals we duplicate the particle cloud and let the copy evolve under the Θ-free mean-field Langevin dynamics (4) until its barycentre stabilises; the converged barycentre labels the well whose basin contains the current state. This detector needs no knowledge of ω and therefore carries over to the non-quadratic case of Section 3.3.

Figure 6 shows a run at $\sigma^2 = 0.4$, $\varepsilon = 0.03$, $\beta = 0.05$, $N = 300$, $T = 5000$, with four panels: the potential V, the trajectory of Θ_t, the barycentre $\bar{X}_t$ with the detected

minima coloured per well, and the coarse-grained potential ω computed by quadrature,

$$\omega(\theta) = \frac{\sigma^2}{2\lambda}\,\theta^2 - \ln \int_{\mathbb{R}} \exp\left(-\frac{1}{\sigma^2}\left(V(x) + \frac{\lambda}{2}x^2\right) + \theta x\right) \mathrm{d}x\,.$$

Over the horizon, Θ_t drives the population through several wells, and the barycentres returned by the fork descent cluster exactly on the minima of ω (dashed lines), confirming that the visited states are indeed the stationary solutions of $\mathcal{F}$.

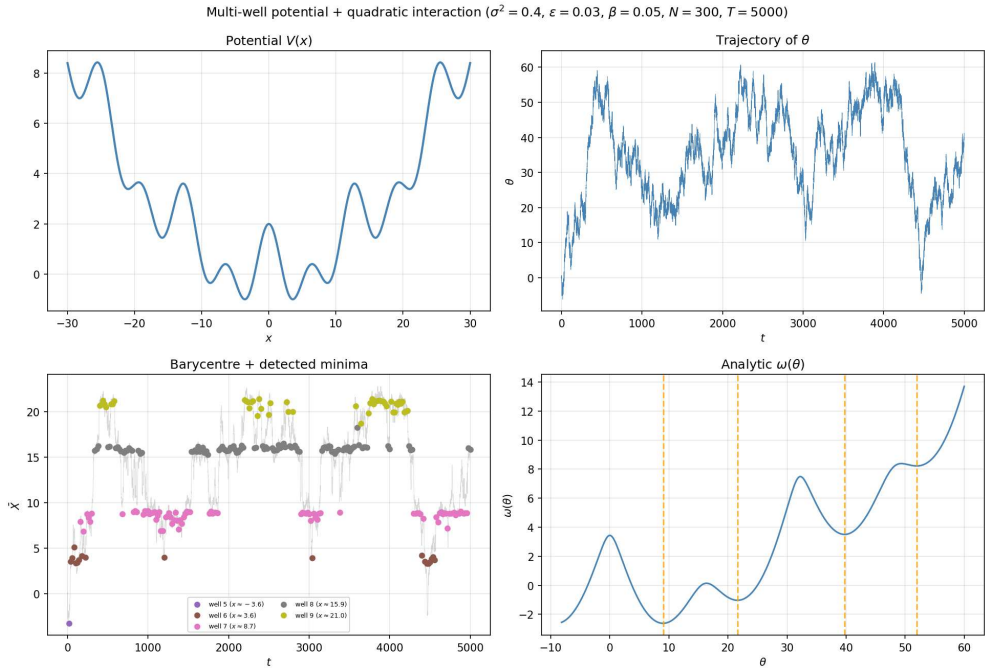


Fig. 6 Multi-well potential with quadratic interaction, at $\sigma^2 = 0.4$, $\varepsilon = 0.03$, $\beta = 0.05$, $N = 300$, $T = 5000$. Top left: the potential V. Top right: the trajectory of Θ_t. Bottom left: the barycentre $\bar{X}_t$ (grey) with the minima detected by the fork descent, coloured by well. Bottom right: the analytic coarse-grained potential $\omega(\theta)$, whose minima (dashed) match the detected barycentres.

Number of wells explored: influence of β.

How many distinct wells the sampler reaches over a given horizon is governed by the macroscopic temperature β^{-1}. Figure 7 reports a long run at $\beta = 0.05$ on the wide range $[-60, 60]$ (sixteen wells), $T = 20000$, aggregated over 10 simulations. The left panel shows the cumulative number of distinct wells visited as a function of time (for each simulation), which keeps growing and reaches 10 to 14 wells per trajectory; the right panel shows the visit frequency of each well: the central wells are visited most,

the outer ones progressively less, and a fraction of the fork descents do not converge in the allotted time (bar NC).

A sweep over β confirms the expected trade-off: for large β the sampler stays in one or two wells but every descent converges, whereas for very small β it touches almost every well but most descents fail to stabilise; an intermediate value $\beta \approx 0.05$–0.1 gives the broadest reliable exploration. This sweep is reported in Figure 8.

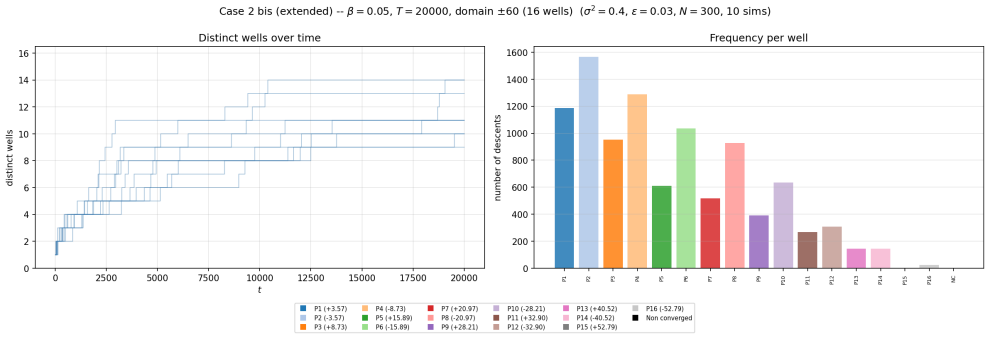


Fig. 7 Well exploration at $\beta = 0.05$ on the range $[-60, 60]$ (sixteen wells), $\sigma^2 = 0.4$, $\varepsilon = 0.03$, $N = 300$, $T = 20000$, over 10 simulations. Left: cumulative number of distinct wells visited over time (one step curve per simulation). Right: number of fork descents falling in each well $P_1, \ldots, P_{16}$, plus the non-converged ones (NC).

Long-time distribution of Θ_t versus ν_β.

Finally, as in the double-well case, we check that the long-time law of Θ_t matches the coarse-grained Gibbs measure $\nu_\beta \propto e^{-\beta\omega}$, with ω evaluated by quadrature. Figure 9 shows this comparison at $\sigma^2 = 2$, $\beta = 1$ for three increasing computational budgets. Even the coarsest one ($N = 10^3$, $\varepsilon = 0.1$, $T = 500$) already captures the shape of ν_β, and the match becomes very good for the richest one ($N = 10^4$, $M = 20$, $\varepsilon = 0.05$, $T = 10^4$), where the histogram reproduces the multi-modal target closely — in line with the averaging principle, the approximation improving as N and M increase, the horizon T grows and the speed ε decreases.

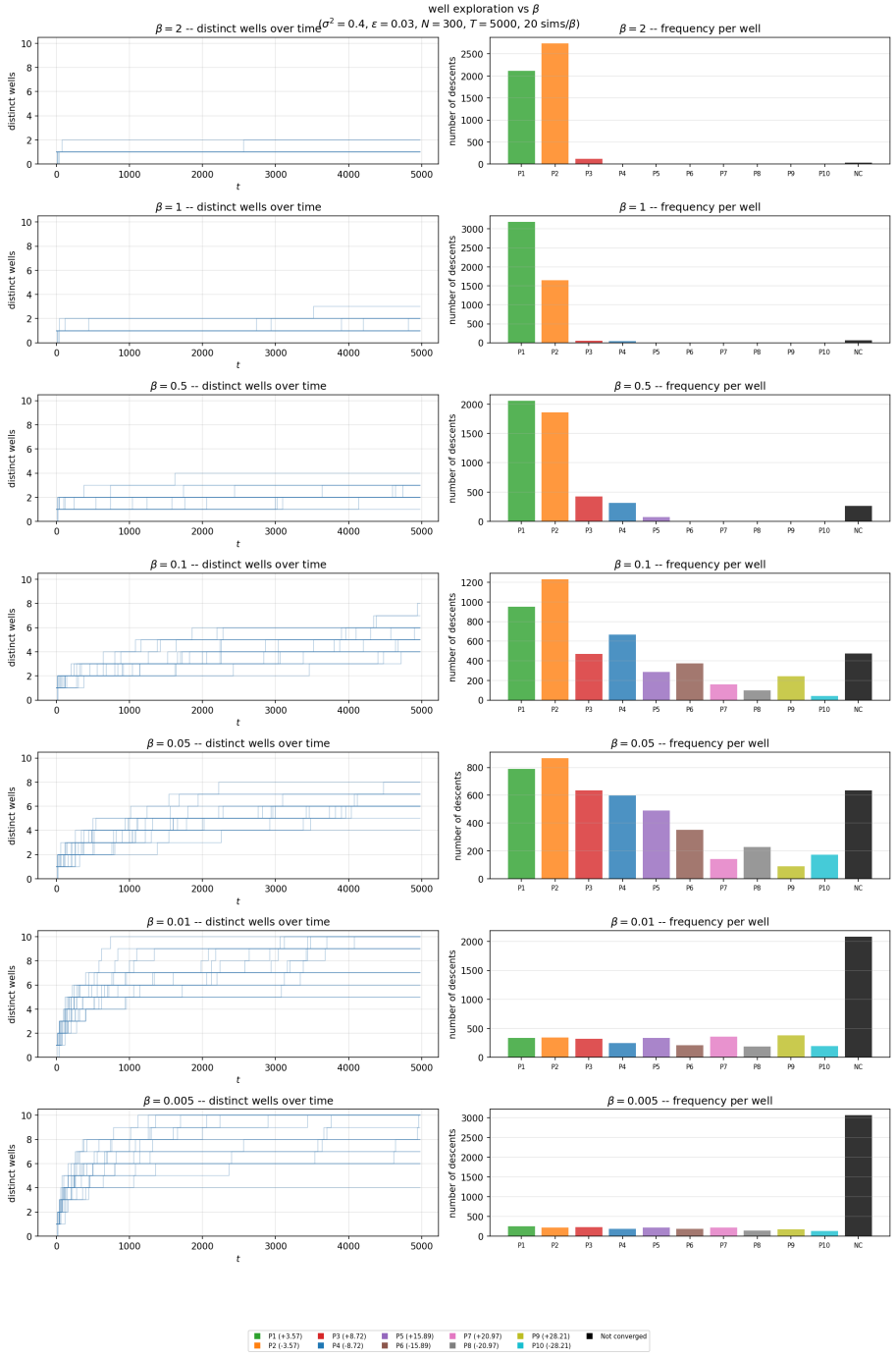


Fig. 8 Well exploration as a function of β on the range $[-30, 30]$ (ten wells), $\sigma^2 = 0.4$, $\varepsilon = 0.03$, $N = 300$, $T = 5000$, with 20 simulations per β. From top to bottom: $\beta = 2$, 1, 0.5, 0.1, 0.05, 0.01, 0.005. Left column: cumulative number of distinct wells visited over time (one step curve per simulation). Right column: number of fork descents falling in each well $P_1, \ldots, P_{10}$, plus the non-converged ones (NC). As β decreases the sampler reaches more wells, but the fraction of non-converged descents grows.

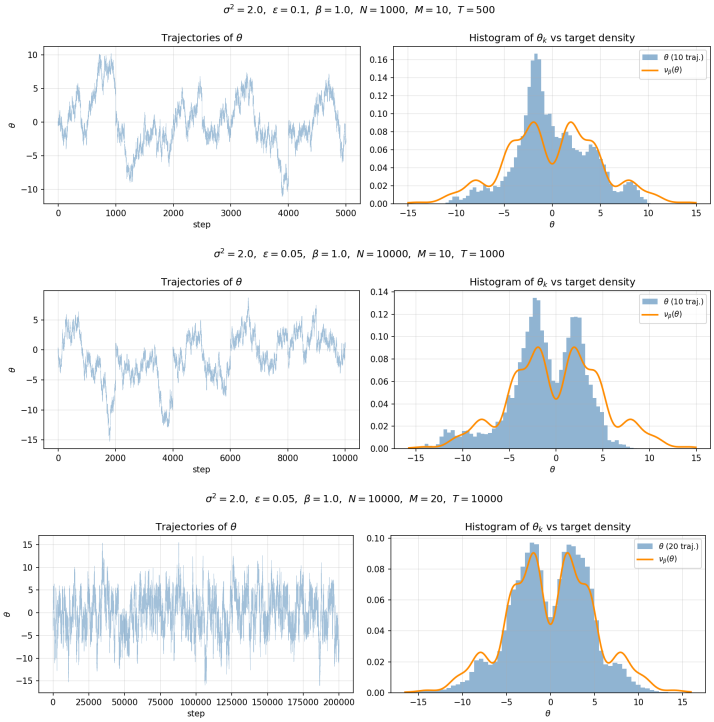


Fig. 9 Long-time distribution of Θ_t for the multi-well potential at $\sigma^2 = 2$, $\beta = 1$, with an increasing computational budget (top to bottom). Top: $N = 10^3$, $\varepsilon = 0.1$, $M = 10$, $T = 500$. Middle: $N = 10^4$, $\varepsilon = 0.05$, $M = 10$, $T = 10^3$. Bottom: $N = 10^4$, $\varepsilon = 0.05$, $M = 20$, $T = 10^4$. Left: concatenated trajectories of Θ_t; right: histogram of Θ_t against $\nu_\beta \propto e^{-\beta\omega}$. The match improves as N, M and T increase and ε decreases.

3.3 Non-quadratic interaction

We now leave the quadratic setting and consider the attractive Gaussian interaction

$$W(x-y) = -\lambda \exp\left(-\frac{|x-y|^2}{2\eta^2}\right), \tag{41}$$

in dimension $d = 1$, keeping the double-well confining potential $V(x) = x^4/4 - x^2/2$. Here $\lambda > 0$ sets the strength of the attraction and $\eta > 0$ its range. For the plain particle system (4) nothing changes — one simply uses this W — and the same kind of phase transition as in Section 3.1 occurs, now at a temperature depending on λ and η. Throughout this section, $\lambda = \eta = 1$ and $\sigma^2 = 0.15$.

Reduction to a finite macroscopic variable.

The two-scale construction applies once W is put in the general form (15), through the expansion

$$W(x-y) = -\lambda\, e^{-\frac{x^2+y^2}{2\eta^2}} \sum_{k\geqslant 0} \frac{(xy)^k}{\eta^{2k} k!} = -\lambda \sum_{k\geqslant 0} n_k(x)\, n_k(y)\,,$$
$$n_k(x) = e^{-\frac{x^2}{2\eta^2}} \frac{x^k}{\eta^k \sqrt{k!}}\,. \tag{42}$$

In fact, we will rather implement the two-scale sampler associated to the truncated potential

$$W_m(x-y) = -\lambda \sum_{k=0}^{m} n_k(x)\, n_k(y)\,,$$

for some $m \in \mathbb{N}$. In that case, we take the feature $\varphi = (n_k)_{k\in[\![0,m]\!]} : \mathbb{R} \to \mathbb{R}^{m+1}$. As in Section 2.3, it would have been possible to consider this feature (with a finite m) even with the full potential (42), by incorporating the remainder $-\frac{\lambda}{2}\sum_{k>m}\left(\int n_k \mathrm{d}\rho\right)^2 = \frac{1}{2}\rho((W - W_m)\star\rho)$ in the term $\mathcal{E}_0$. Instead, we neglect this remainder. We have two reasons to do this: first, it is computationally cheaper, since computing the pairwise interactions has a complexity $\mathcal{O}(N^2)$ when the potential is W (or $W - W_m$) but only $\mathcal{O}(Nm)$ with W_m (since we only need to compute $\frac{1}{N}\sum_{i=1}^N n_k(X_t^i)$ for $k \in [\![0, m]\!]$). The second reason is that, in order to assess the properties of the algorithm, we want to compare it to a clear reference. In particular, as in Figures 5 and 9 we want to compare the distribution sampled by the macroscopic variable with the theoretical macroscopic Gibbs measure ν_β. By truncating the potential, similarly to the previous experiments with quadratic interaction, we can compute ω over $\mathbb{R}^{1+m}$ by a simple quadrature for small values of m. As detailed below, the impact of the choice of m is discussed in Figure 10 and, after that, the experiments are conducted with $m = 3$ (the macroscopic variable Θ_t is thus a 4-dimensional vector).

Concretely, using the rescaling of (37) and writing $\bar{n}_{k,t} = \frac{1}{N}\sum_{i=1}^{N} n_k(X_t^i)$ for the empirical feature means, the coupled system reads

$$\begin{aligned} \mathrm{d}X_t^i &= -\frac{V'(X_t^i)}{\sigma^2}\,\mathrm{d}t + \sum_{k=0}^{m} \Theta_t^k\, n_k'(X_t^i)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t^i\,, \\ \mathrm{d}\Theta_t^k &= -\varepsilon\left(\frac{\sigma^2}{\lambda}\,\Theta_t^k - \bar{n}_{k,t}\right)\mathrm{d}t + \sqrt{2\varepsilon/\beta}\,\mathrm{d}W_t^k\,. \end{aligned} \tag{43}$$

Nevertheless, when proceeding to fork descents (as explained in Section 3.2) with the standard Langevin process (4), we still use the full potential (41) (with complexity $\mathcal{O}(N^2)$ at each step). In other words, the truncation from W to W_m is another approximation (in top of the approximation in N and ε), but we still use the resulting process as an algorithm to sample good initial conditions for a gradient descent converging to the local minimizers of the free energy corresponding to the Gaussian interaction (41).

Coarse-grained potential and stationary solutions.

The stationary solutions of $\mathcal{F}$ are obtained by a fixed-point iteration on the self-consistency equation $\rho \propto e^{-(V+W*\rho)/\sigma^2}$. As in the previous sections, for $m \in \mathbb{N}$, the coarse-grained potential is computed by quadrature,

$$\omega(\theta) = \frac{\sigma^2}{2\lambda}\,|\theta|^2 - \ln \int_{\mathbb{R}} \exp\left(-\frac{V(x)}{\sigma^2} + \theta\cdot\varphi(x)\right)\mathrm{d}x\,, \qquad \mu_\theta \propto \exp\left(-\frac{V}{\sigma^2} + \theta\cdot\varphi\right)\,.$$

Figure 10(a) shows the phase transition: for the exact interaction (41), $\mathcal{F}$ has two minimisers $\mu_\pm$, with barycentres $\pm m_*$, below $\sigma_c^2 \approx 0.275$, and a single symmetric one above; the truncated models (same $\mathcal{F}$ except that W is replaced by W_m) reproduce this bifurcation, and converge to the exact one as m increases. At $\sigma^2 = 0.15$, $m_* \approx 0.86$. As discussed in Section 1.3, the local minimisers of ω correspond to those of $\mathcal{F}$: Figure 10(b) shows that μ_{θ_*}, at the two minima $\theta_*^\pm$ of ω (computed with $m = 3$), coincides with the minimisers $\mu_\pm$ (up to an error $\mathcal{W}_2 \approx 0.03$ for $m = 3$, a distance which decreases geometrically in m, see the inset plot), and Figure 10(c) shows the profile of ω (for $m = 3$) along θ^1, the component conjugate to the odd feature n_1 that distinguishes the two wells: a symmetric double well with barrier $\Delta\omega \approx 0.29$, whose minima are indistinguishable in the graph from the values $\sigma^{-2}\mathcal{F}(\mu_\pm)$ with the full potential W (as determined by the fixed-point procedure).

Exploring the wells in the vector Θ.

Figure 11 shows a run of (43) at $\sigma^2 = 0.15$, $\lambda = \eta = 1$, $m = 3$, $\varepsilon = 0.3$, $\beta = 4$, $N = 300$, $T = 3000$. As in Section 3.2, every 20 time units we launch a fork descent, here with the exact Gaussian interaction (41). The component Θ_t^1 hops between the two minima $\pm\theta_*^1$ of ω, and the well detected by the fork descent follows its sign: the vector Θ_t drives the transitions between the two wells, exactly as the scalar Θ_t did in the quadratic case. All the descents converge to the stationary solutions $\mu_\pm$ (their barycentres are ± 0.85, against $\pm m_*$). Along the run, the particles are distributed

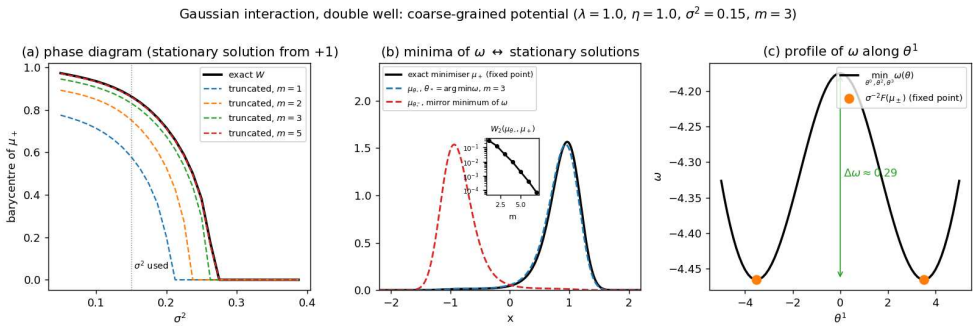


Fig. 10 Gaussian interaction, double-well potential, $\lambda = \eta = 1$, $\sigma^2 = 0.15$, computed on a grid. (a) Barycentre of the stationary solution obtained from $+1$ versus σ^2, for the exact interaction and for the truncations at order m. (b) Exact minimiser μ_+ versus $\mu_{\theta_*^\pm}$ at the two minima of ω, $m = 3$; inset: $\mathcal{W}_2(\mu_{\theta_*}, \mu_+)$ versus m. (c) Profile $\theta^1 \mapsto \min_{\theta^0,\theta^2,\theta^3} \omega(\theta)$, with the values $\sigma^{-2}\mathcal{F}(\mu_\pm)$ of the fixed-point minimisers (dots).

according to μ_{Θ_t}, as stated in Theorem 3: when Θ_t is in a well, $\mathcal{W}_2(\pi_{\mathbf{X}_t}, \mu_{\Theta_t})$ is at the level of the distance between N i.i.d. samples of μ_{Θ_t} and μ_{Θ_t} itself (medians 0.045 and 0.040), with a slight lag during the transitions (medians 0.072 and 0.051).

Transition times and number of particles.

For the plain particle system (4), the equilibrium of the macroscopic variable is $\nu_N \propto e^{-N\omega_N}$ (see (16)) and transitions between μ_+ and μ_- are rare events whose rate decays like $e^{-N\Delta\omega}$, while for the two-scale sampler Θ_t is distributed according to $\nu_\beta \propto e^{-\beta\omega}$ and the transition rate does not depend on N. Figure 12 compares both, counting the switches of the barycentre between the two wells (with the truncated interaction for the plain system, so that both target the same free energy). For the plain system the rate decays exponentially, from $7 \cdot 10^{-2}$ at $N = 4$ to 10^{-4} at $N = 28$, with a slope -0.28 close to $\Delta\omega$; for the sampler it stays between $2.6 \cdot 10^{-3}$ and $5.3 \cdot 10^{-3}$ from $N = 10$ to $N = 1000$.

Long-time distribution of Θ_t versus the target ν_β.

Finally, as in Sections 3.1 and 3.2, we check that the long-time law of Θ_t is the coarse-grained Gibbs measure $\nu_\beta \propto e^{-\beta\omega}$ predicted by the averaging principle. Since $\Theta_t \in \mathbb{R}^4$, we compare marginals: as before, ν_β is computed by quadrature, its marginals being obtained by integrating $e^{-\beta\omega}$ over the other components of θ. Figure 13 does this at $\beta = 4$, $N = 300$, for $\varepsilon \in \{1,\ 0.3,\ 0.1\}$ (horizon $T = 500/\varepsilon$, 8 trajectories started in the left well). Within each well, the law of Θ_t^1 matches that of ν_β for all three values of ε (total variation distance 0.02–0.04 after symmetrisation $\theta^1 \to -\theta^1$). The relative weight of the two wells, on the other hand, is estimated from a limited number of transitions (15–20 in total for each value of ε), and the fraction of time with $\Theta_t^1 > 0$ (0.25, 0.27 and 0.45) fluctuates around its target $1/2$; it improves with the number of trajectories or the horizon.

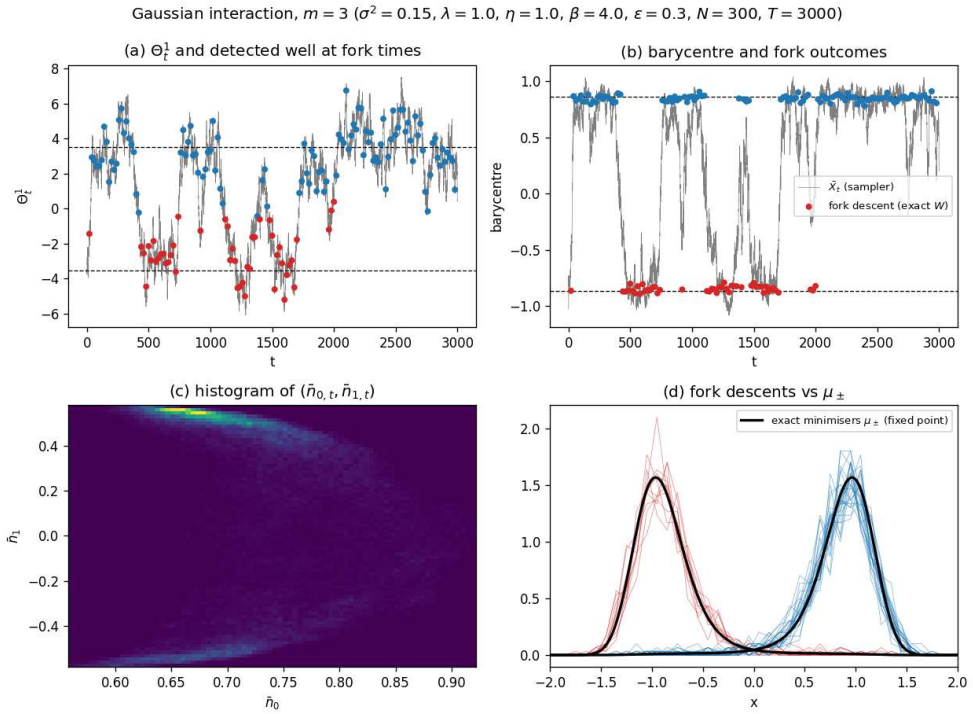


Fig. 11 Gaussian interaction, double-well potential, at $\sigma^2 = 0.15$, $\lambda = \eta = 1$, $m = 3$, $\varepsilon = 0.3$, $\beta = 4$, $N = 300$, $T = 3000$. (a) Trajectory of Θ^1_t, with the fork times coloured by the well detected during the descent (dashed: $\pm\theta^1_*$). (b) Barycentre $\bar{X}_t$ (grey) and barycentres reached by the fork descents with the exact Gaussian interaction (dots); dashed: $\pm m_*$. (c) Two-dimensional histogram of the feature means $(\bar{n}_{0,t}, \bar{n}_{1,t})$. (d) Final densities of the fork descents versus the stationary solutions $\mu_\pm$.

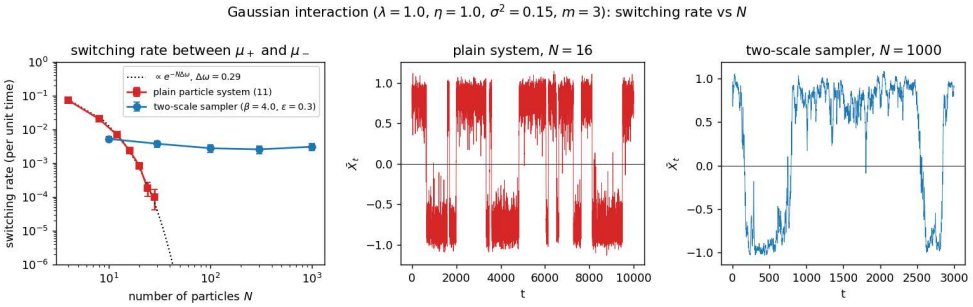


Fig. 12 Gaussian interaction, $\sigma^2 = 0.15$, $m = 3$. Left: switching rate between μ_+ and μ_- versus N, for the plain particle system (squares, 12 runs of length $T = 10^4$ per N) and for the two-scale sampler with $\beta = 4$, $\varepsilon = 0.3$ (circles, 8 runs of length $T = 3000$); error bars $\pm 2\sqrt{\#\text{switches}}$; dotted: $e^{-N\Delta\omega}$. Middle and right: barycentre trajectories of the plain system ($N = 16$) and of the sampler ($N = 1000$).

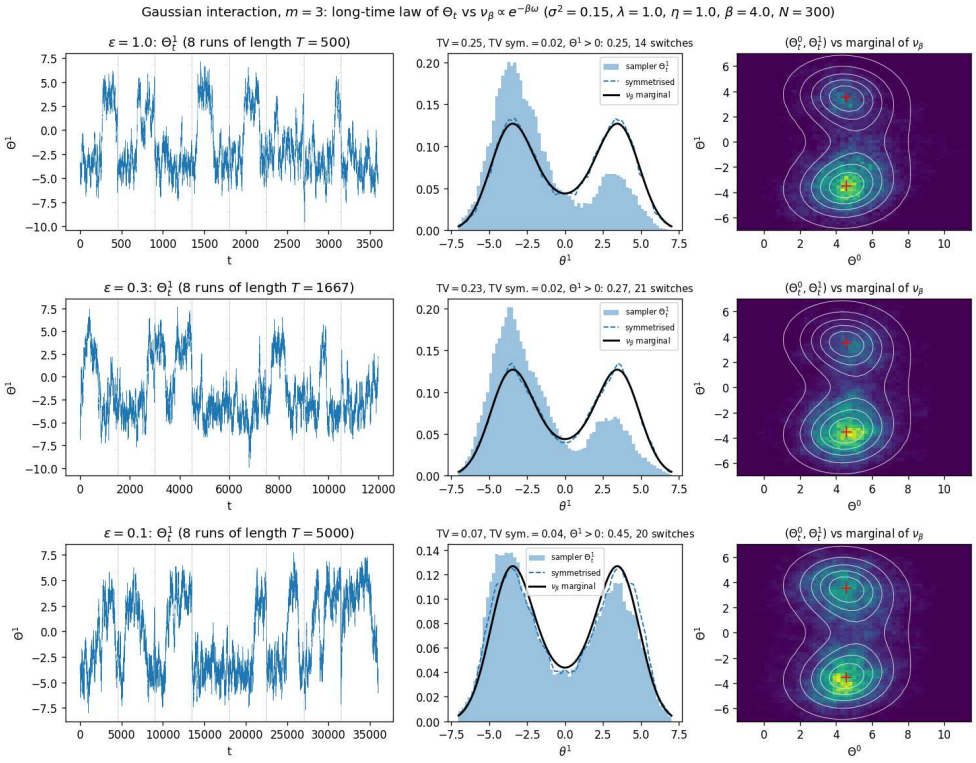


Fig. 13 Long-time distribution of Θ_t at $\sigma^2 = 0.15$, $m = 3$, $\beta = 4$, $N = 300$, aggregated over 8 trajectories. From top to bottom: $\varepsilon = 1,\ 0.3,\ 0.1$, with $T = 500/\varepsilon$. Left: the trajectories of Θ_t^1, concatenated one after another. Middle: histogram of Θ_t^1 (and its symmetrisation, dashed) against the marginal of $\nu_\beta \propto e^{-\beta\omega}$ computed by quadrature (black). Right: histogram of (Θ_t^0, Θ_t^1) with contours of the (θ^0, θ^1)-marginal of ν_β and the minima of ω (crosses).

4 Proofs

4.1 Propagation of chaos

This section is devoted to the proof of Theorem 1.

Proposition 4 *In the settings of Theorem 1, Equations (18) and (22) have unique strong solutions. Moreover, for any $T > 0$, $\sup_{t\in[0,T]} \int_{\mathbb{R}^d} |x|^p \nu_t(dx)$ is almost surely finite.*

Proof Since $\nabla_{x_i} V_N(\mathbf{x}) = D\mathcal{E}_0(\pi_{\mathbf{X}}, x_i)$, (23) implies that

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^{dN}, \qquad (\mathbf{x}-\mathbf{y}) \cdot (\nabla V_N(\mathbf{x}) - \nabla V_N(\mathbf{y})) \leqslant L|\mathbf{x}-\mathbf{y}|^2 .$$

Well-posedness for Equation (18) then follows from standard results for stochastic differential equations. Similarly, well-posedness for (22) follows from a standard fixed-point argument. We only detail the L^p moment for ν_t. Writing $f_p(x) = |x|^p$,

$$\begin{aligned} \mathrm{d}\int_{\mathbb{R}^d} f_p \mathrm{d}\nu_t &= \int_{\mathbb{R}^d} \left(\nabla f_p \cdot \left(-D\mathcal{E}_0(\nu_t, \cdot) + \bar{\Theta}_t \cdot \nabla\varphi\right) + \Delta f_p\right) \mathrm{d}\nu_t \mathrm{d}t \\ &\leqslant C\left(1 + |\bar{\Theta}_t|^p + \int_{\mathbb{R}^d} f_p \mathrm{d}\nu_t\right) \mathrm{d}t \,, \end{aligned}$$

for some $C > 0$, where we used Young inequality to bound $|\nabla f_p||\Theta_t| \leqslant C(f_p + |\Theta_t|^p)$, and we used (23) to get that

$$\begin{aligned} x \cdot D\mathcal{E}_0(\nu_t, x) &\leqslant L|x|^2 + x \cdot D\mathcal{E}_0(\nu_t, 0) \\ &\leqslant L|x|^2 + |x||D\mathcal{E}_0(\delta_0, 0)| + L|x|\mathcal{W}_2(\nu_t, \delta_0) \\ &= L|x|^2 + |x||D\mathcal{E}_0(\delta_0, 0)| + L|x|\sqrt{\int_{\mathbb{R}^d} f_2 \mathrm{d}\nu_t} \\ &\leqslant L|x|^2 + |x||D\mathcal{E}_0(\delta_0, 0)| + L|x|\left(\int_{\mathbb{R}^d} f_p \mathrm{d}\nu_t\right)^{1/p} \end{aligned}$$

(and then again Young inequality). Grönwall lemma then shows that, for any $T > 0$, there exists $C_T > 0$ such that, almost surely,

$$\sup_{t\in[0,T]} \int_{\mathbb{R}^d} f_p \mathrm{d}\nu_t \leqslant C_T \left(1 + \int_{\mathbb{R}^d} f_p \mathrm{d}\nu_0 + \int_0^T |\Theta_s|^p \mathrm{d}s\right) ,$$

which concludes the proof. □

Proof of Theorem 1 Given a solution of (22), denote by $\mathbf{Y}_t = (Y_t^1, \dots, Y_t^N)$ the solution of

$$\forall i \in [\![1, N]\!], \qquad \mathrm{d}Y_t^i = -D\mathcal{E}_0(\nu_t, Y_t^i)\mathrm{d}t + \bar{\Theta}_t \cdot \nabla\varphi(Y_t^i)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^i \,, \tag{44}$$

with initial conditions $\mathbf{Y}_0 = \mathbf{X}_0$. Write $(\mathcal{F}_t)_{t\geqslant 0}$ the filtration associated to $(B_t)_{t\geqslant 0}$. Conditionally to $\mathcal{F}_t$, $Y_t^1, \dots, Y_t^N$ are independent and distributed according to ν_t. In particular, for $i \in [\![1, N]\!]$, using that $\nabla_{x_i} V_N(\mathbf{x}) = D\mathcal{E}_0(\pi_{\mathbf{X}}, x_i)$,

$$\begin{aligned} \mathbb{E}\left(|D\mathcal{E}_0(\nu_t, Y_t^i) - \nabla_{x_i} V_N(\mathbf{Y}_t)|^2 \mid \mathcal{F}_t\right) &\leqslant L^2 \mathbb{E}\left(\mathcal{W}_2^2(\pi_{\mathbf{Y}_t}, \nu_t) \mid \mathcal{F}_t\right) \\ &\leqslant CN^{-\eta}\left(\int_{\mathbb{R}^d} |y|^p \nu_t(\mathrm{d}y)\right)^{2/p} , \end{aligned}$$

for some constant $C, \eta > 0$ thanks to [23] and Proposition 4. Then, since X^i and Y^i are driven by the same Brownian motion B^i,

$$\begin{aligned}
\mathrm{d}|\mathbf{X}_t - \mathbf{Y}_t|^2 &\leqslant 2(\mathbf{X}_t - \mathbf{Y}_t) \cdot (\nabla V_N(\mathbf{X}_t) - \nabla V_N(\mathbf{Y}_t)) + 2\sum_{i=1}^N (X_t^i - Y_t^i) \\
&\quad \times \left(|D\mathcal{E}_0(\nu_t, Y_t^i) - \nabla_{x_i} V_N(\mathbf{Y}_t)| + |\bar{\Theta}_t||\nabla\varphi(X_t^i) - \nabla\varphi(Y_t^i)| + |\Theta_t - \bar{\Theta}_t||\nabla\varphi(X_t^i)| \right) \mathrm{d}t \\
&\leqslant C(1 + |\bar{\Theta}_t|) \left(|\mathbf{X}_t - \mathbf{Y}_t|^2 + N|\Theta_t - \bar{\Theta}_t|^2 + \sum_{i=1}^N |D\mathcal{E}_0(\nu_t, Y_t^i) - \nabla_{x_i} V_N(\mathbf{Y}_t)|^2 \right) \mathrm{d}t \,,
\end{aligned}$$

for some $C > 0$. Similarly,

$$\begin{aligned}
\mathrm{d}|\bar{\Theta}_t - \Theta_t|^2 &\leqslant 2\varepsilon|\bar{\Theta}_t - \Theta_t| \left(|\bar{\Theta}_t - \Theta_t| + L\mathcal{W}_2(\nu_t, \pi_{\mathbf{X}_t}) \right) \mathrm{d}t \\
&\leqslant C \left(|\bar{\Theta}_t - \Theta_t|^2 + \frac{1}{N}|\mathbf{X}_t - \mathbf{Y}_t|^2 + \mathcal{W}_2^2(\nu_t, \pi_{\mathbf{Y}_t}) \right) \mathrm{d}t
\end{aligned}$$

for some $C > 0$. Writing

$$R_t = \mathbb{E}\left(|\bar{\Theta}_t - \Theta_t|^2 + \frac{1}{N}|\mathbf{X}_t - \mathbf{Y}_t|^2 \;\middle|\; \mathcal{F}_t \right),$$

combining the previous bounds (and relying again on [23]) gives

$$\mathrm{d}R_t \leqslant H_t \left(R_t + N^{-\eta} \right) \mathrm{d}t$$

for some $\mathcal{F}_t$-measurable random H_t, almost surely bounded on any finite time interval, and independent from N. By Grönwall lemma, since $R_0 = 0$, we get that

$$R_t \leqslant H_t' N^{-\eta}$$

for some $\mathcal{F}_t$-measurable random $H_t' < \infty$ (non-decreasing with t). Moreover, as we used above,

$$\mathcal{W}_2^2(\pi_{\mathbf{X}_t}, \nu_t) \leqslant \frac{2}{N}|\mathbf{X}_t - \mathbf{Y}_t|^2 + 2\mathcal{W}_2^2(\nu_t, \pi_{\mathbf{Y}_t}) \,,$$

which combined with the previous bound and [23] leads to

$$\mathbb{E}\left(\mathcal{W}_2^2(\pi_{\mathbf{X}_t}, \nu_t) + |\Theta_t - \bar{\Theta}_t|^2 \mid \mathcal{F}_t \right) \leqslant S_t N^{-\eta} \,, \tag{45}$$

for an $\mathcal{F}_t$-measurable random $S_t < \infty$ (non-decreasing with t). Then, for an arbitrary $K > 0$, we bound, for any $t \in [0, T]$ and $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\mathcal{W}_2(\pi_{\mathbf{X}_t}, \nu_t) + |\Theta_t - \bar{\Theta}_t| \geqslant \delta \right) &\leqslant \mathbb{P}\left(\mathcal{W}_2(\pi_{\mathbf{X}_t}, \nu_t) + |\Theta_t - \bar{\Theta}_t| \geqslant \delta,\ S_t \leqslant K \right) + \mathbb{P}\left(S_t > K \right) \\
&\leqslant \delta^{-2} K N^{-\eta} + \mathbb{P}\left(S_t > K \right) .
\end{aligned}$$

Taking the limsup as $N \to \infty$ and then letting $K \to \infty$ concludes the proof.

□

Remark 1 From (45), we get that, for all $T > 0$, almost surely,

$$\sup_{t \in [0,T]} \sup_{N \geqslant 1} N^\eta \mathbb{E}\left(\mathcal{W}_2^2(\pi_{\mathbf{X}_t}, \nu_t) + |\Theta_t - \bar{\Theta}_t|^2 \mid \mathcal{F}_t \right) < \infty \,,$$

where η is the rate provided by [23] (which depends explicitly on p and d; for instance, if $d > 4$ and $p = 4$, then $\eta = 2/d$).

4.2 Averaging

This section is devoted to the proof of Theorem 3.

4.2.1 PL inequalities and consequences

We start with some properties of the free energies $\mathcal{F}_\theta$, $\theta \in \mathbb{R}^m$.

Proof of Proposition 2 From the lower-bound (31), for any $\theta \in \mathbb{R}^m$,

$$\begin{aligned}\mathcal{F}_\theta(\nu) &\geqslant c_0 \int_{\mathbb{R}^d} |x|^2 \nu(\mathrm{d}x) - C_0 - |\theta| \left(|\varphi_\ell(0)| + \|\nabla \varphi_\ell\|_\infty \int_{\mathbb{R}^d} |x| \nu(\mathrm{d}x) + \|\varphi_b\|_\infty \right) + \mathcal{H}(\nu) \\ &\geqslant \frac{c_0}{2} \int_{\mathbb{R}^d} |x|^2 \nu(\mathrm{d}x) - C_0 - |\theta| \left(|\varphi_\ell(0)| + \|\varphi_b\|_\infty \right) - \frac{1}{2c_0} |\theta|^2 \|\nabla \varphi_\ell\|_\infty^2 + \mathcal{H}(\nu) \\ &= \frac{c_0}{4} \int_{\mathbb{R}^d} |x|^2 \nu(\mathrm{d}x) + \frac{c_0}{4} \mathcal{H}\left(\nu | \gamma\right) - C_1 (1 + |\theta|^2) \end{aligned} \tag{46}$$

for some constant C_1, where $\gamma = \mathcal{N}(0, (2c_0)^{-1} I)$. As a consequence, $\mathcal{F}_\theta$ is lower-bounded, and $\mathcal{F}_\theta(\nu)$ goes to infinity as $\int_{\mathbb{R}^d} |x|^2 \nu(\mathrm{d}x) \to \infty$. By semi-lower continuity, $\mathcal{F}_\theta$ admits a global minimiser μ_θ.

Since the term $\nu \mapsto \theta \cdot \nu(\varphi)$ is linear,

$$\mathcal{E}_\theta(\nu) := \mathcal{E}_0(\nu) - \theta \cdot \nu(\varphi)$$

satisfies the same condition (30) as $\mathcal{E}_0$. From [42, Lemma 3], for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\widetilde{\mathcal{F}}_\theta(\nu) \leqslant \mathcal{H}\left(\nu | \Gamma_\theta(\nu)\right) + \eta \|\nu - \mu_\theta\|_{TV}^2 \tag{47}$$

with the probability density

$$\Gamma_\theta(\nu) \propto \exp\left(-\frac{\delta \mathcal{E}_\theta}{\delta m}(\nu, \cdot) \right) = \exp\left(-\frac{\delta \mathcal{E}_0}{\delta m}(\nu, \cdot) + \theta \cdot \varphi \right) ,$$

and

$$\mathcal{H}(\nu | \mu_\theta) \leqslant \widetilde{\mathcal{F}}(\nu) + \eta \|\nu - \mu_\theta\|_{TV}^2 .$$

Combining Pinsker's inequality and the previous ones gives

$$\|\nu - \mu_\theta\|_{TV}^2 \leqslant 2\mathcal{H}(\nu | \mu_\theta) \leqslant 2\mathcal{H}\left(\nu | \Gamma_\theta(\nu)\right) + 4\eta \|\nu - \mu_\theta\|_{TV}^2 ,$$

so that

$$\|\nu - \mu_\theta\|_{TV}^2 \leqslant \frac{2}{1 - 4\eta} \mathcal{H}\left(\nu | \Gamma_\theta(\nu)\right) . \tag{48}$$

Applying this to ν being a critical point of $\mathcal{F}_\theta$ (i.e. a solution of $\nu = \Gamma_\theta(\nu)$) shows that $\nu = \mu_\theta$. In particular, μ_θ is the unique global minimiser of $\mathcal{F}_\theta$. Besides, combining (47) and (48) gives

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d), \qquad \widetilde{\mathcal{F}}_\theta(\nu) \leqslant \left(1 + \frac{2\eta}{1 - 4\eta} \right) \mathcal{H}(\nu | \Gamma_\theta(\nu)) . \tag{49}$$

Under Assumption 2, for any $\theta = (\theta_\ell, \theta_b) \in \mathbb{R}^{m_\ell} \times \mathbb{R}^{m_b}$ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\frac{\delta \mathcal{E}_\theta}{\delta m}(\nu, \cdot)$ is the sum of three functions: a κ-strongly convex one (including the term $\theta_\ell \cdot \varphi_\ell$) , a $M + |\theta_b| \|\varphi_b\|_\infty$-bounded one and a L-Lipschitz one. By the Bakry-Emery condition and the stability of log-Sobolev inequalities with respect to bounded and Lipschitz perturbations (see [4]), for any $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\Gamma_\theta(\nu)$ satisfies

$$\forall \nu, \mu \in \mathcal{P}_2(\mathbb{R}^d), \qquad \mathcal{H}(\mu | \Gamma_\theta(\nu)) \leqslant \frac{e^{|\theta_b| \|\varphi_b\|_\infty}}{\lambda'} \mathcal{I}\left(\mu | \Gamma_\theta(\nu)\right) ,$$

for some $\lambda' > 0$ depending only on κ, M and L. Applying this inequality with $\mu = \nu$, plugging it in (49) and noticing that

$$\mathcal{I}\left(\nu|\Gamma_\theta(\nu)\right) = \int_{\mathbb{R}^d} \left|\nabla \ln \nu + \frac{\delta \mathcal{E}_\theta}{\delta m}(\nu, \cdot)\right|^2 \nu = \mathcal{I}_\theta(\nu)$$

concludes the proof. □

Lemma 5 *In the settings of Proposition 2, assuming that φ is L_φ-Lipschitz continuous,*

$$\forall \theta, \psi \in \mathbb{R}^m, \qquad \mathcal{W}_2(\mu_\theta, \mu_\psi) \leqslant \frac{3L_\varphi}{\lambda_\psi}|\psi - \theta| \,. \tag{50}$$

Proof For a fixed $\theta \in \mathbb{R}^m$, along the flow (26), by a classical computation (see [1]), for all $t > 0$,

$$\partial_t \widetilde{\mathcal{F}}_\theta(\rho_t) = -\mathcal{I}_\theta(\rho_t) \,.$$

Keeping the same flow with this θ and considering another $\psi \in \mathbb{R}^m$,

$$\partial_t \widetilde{\mathcal{F}}_\psi(\rho_t) = -\mathcal{I}_\theta(\rho_t) + (\theta - \psi) \cdot \partial_t \rho_t(\varphi) \,.$$

Integrating by parts,

$$|\partial_t \rho_t(\varphi)| = \left|\int_{\mathbb{R}^d} \nabla\varphi \cdot \nabla \left[\ln \rho_t + D\mathcal{E}_0(\rho_t, \cdot) - \theta \cdot \nabla\varphi\right] \rho_t\right| \leqslant L_\varphi \sqrt{\mathcal{I}_\theta(\rho_t)}$$

and thus

$$\begin{aligned}
\partial_t \widetilde{\mathcal{F}}_\psi(\rho_t) &\leqslant -\frac{1}{2}\mathcal{I}_\theta(\nu_t) + \frac{1}{2}L_\varphi^2|\theta - \psi|^2 \\
&\leqslant -\frac{1}{4}\mathcal{I}_\psi(\nu_t) + 2L_\varphi^2|\theta - \psi|^2 \\
&\leqslant -\frac{\lambda_\psi}{2}\widetilde{\mathcal{F}}_\psi(\rho_t) + 2L_\varphi^2|\theta - \psi|^2 \,,
\end{aligned}$$

using the PL inequality (28). Applying this to $\rho_0 = \mu_\theta$ (so that $\rho_t = \mu_\theta$ for all $t \geqslant 0$) gives

$$\widetilde{\mathcal{F}}_\psi(\mu_\theta) \leqslant \frac{4L_\varphi^2}{\lambda_\psi}|\theta - \psi|^2 \,. \tag{51}$$

Moreover, the PL inequality is known to imply the transport inequality

$$\forall \nu \in \mathcal{P}_2(\mathbb{R}^d) \,, \qquad \mathcal{W}_2^2(\nu, \mu_\psi) \leqslant \frac{2}{\lambda_\psi}\widetilde{\mathcal{F}}_\psi(\nu) \,, \tag{52}$$

cf. [12, 20, 42]. Combining the two last inequalities establishes (50). □

Lemma 6 *Under the settings of Lemma 5, for all $\theta, \psi \in \mathbb{R}^m$ and $\rho \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\widetilde{\mathcal{F}}_\theta(\rho) \leqslant 2\widetilde{\mathcal{F}}_\psi(\rho) + \frac{4L_\varphi^2}{\lambda_\psi}|\psi - \theta|^2 \,.$$

Proof Using that μ_θ minimises $\mathcal{F}_\theta$,

$$\begin{aligned}
\widetilde{\mathcal{F}}_\theta(\rho) &= \mathcal{F}_\theta(\rho) - \mathcal{F}_\theta(\mu_\theta) \\
&= \mathcal{F}_\psi(\rho) + (\psi - \theta) \cdot (\rho(\varphi) - \rho(\mu_\theta)) - \mathcal{F}_\psi(\mu_\theta) \\
&\leqslant \widetilde{\mathcal{F}}_\psi(\rho) + L_\varphi|\psi - \theta| \left(\mathcal{W}_2(\rho, \mu_\psi) + \mathcal{W}_2(\mu_\theta, \mu_\psi)\right) \\
&\leqslant \widetilde{\mathcal{F}}_\psi(\rho) + \frac{\lambda_\psi}{2}\mathcal{W}_2^2(\rho, \mu_\psi) + \frac{4L_\varphi^2}{\lambda_\psi}|\psi - \theta|^2 \,.
\end{aligned}$$

where we used Lemma 5. The transport inequality (52) concludes the proof. □

4.2.2 Truncation

Since the PL constant λ_θ provided in (32) is not bounded from below uniformly in θ, we need to consider a truncated process for which this constant is uniform in time. We will then prove the averaging principle for the truncated process, for an arbitrary truncation parameter. Conclusion will follow by letting the truncation parameter go to infinity, using a control uniform on ε on the probability that the initial and modified processes split before a given time $T > 0$. This section is devoted to this point.

For an arbitrary $R \geqslant 1$ (considered as fixed until the very end of the proof of Theorem 3, where we will send it to infinity), introduce the function $g \in \mathcal{C}^2(\mathbb{R}^m, \mathbb{R}^m)$ given by

$$g(\theta_\ell, \theta_b) = (\theta_\ell, \chi(|\theta_b|/R)\theta_b) \,. \tag{53}$$

for some $\mathcal{C}^2$ cut-off function χ, non-increasing, with $\chi(r) = 1$ for $r \leqslant 1$ and $\chi(r) = 1/r$ for $r \geqslant 2$. Then g is 1-Lipschitz continuous and its second derivative is bounded by some constant $C_g > 0$ uniformly over $R \geqslant 1$. Moreover, Proposition 2 implies that

$$\lambda := \inf_{\theta \in \mathbb{R}^m} \lambda_{g(\theta)} > 0 \,. \tag{54}$$

This will enable to prove the averaging principle for the truncated process in the next section.

Since $\theta = g(\theta)$ as long as $|\theta_b| \leqslant R$, we will use the following to eventually let $R \to \infty$.

Lemma 7 *For any $T > 0$, decomposing the solutions of (25) and (27) as $(\Theta^\varepsilon_{\ell,t}, \Theta^\varepsilon_{b,t}) \in \mathbb{R}^{m_\ell} \times \mathbb{R}^{m_b}$ and $(\Theta^0_{\ell,t}, \Theta^0_{b,t}) \in \mathbb{R}^{m_\ell} \times \mathbb{R}^{m_b}$,*

$$\sup_{\varepsilon \in (0,1]} \mathbb{P}\left(\sup_{t \in [0,T]} |\Theta^\varepsilon_{b,t}| + |\Theta^0_{b,t}| \geqslant K \right) \underset{K \to \infty}{\longrightarrow} 0 \,.$$

Proof First,

$$\partial_t \mathbb{E}\left(|\Theta^\varepsilon_{b,t}|^2\right) \leqslant -\mathbb{E}\left(|\Theta^\varepsilon_{b,t}|^2\right) + \|\varphi_b\|^2_\infty + 4m_b\beta^{-1} \,,$$

which ensures that

$$\sup_{t \geqslant 0} \mathbb{E}\left(|\Theta^\varepsilon_{b,t}|^2\right) \leqslant M_b := \mathbb{E}\left(|\Theta^\varepsilon_{b,0}|^2\right) + \|\varphi_b\|^2_\infty + 4m_b\beta^{-1} \,. \tag{55}$$

The same goes for $\Theta^0_{b,t}$. Next,

$$\mathrm{d}|\Theta^\varepsilon_{b,t}|^2 \leqslant \left(-|\Theta^\varepsilon_{b,t}|^2 + \|\varphi_b\|^2_\infty + 4m_b\beta^{-1}\right)\mathrm{d}t + 2\sqrt{2/\beta}|\Theta^\varepsilon_{b,t}|\mathrm{d}B_t \,,$$

and thus, integrating in time,

$$|\Theta^\varepsilon_{b,t}|^2 \leqslant t\left(\|\varphi_b\|^2_\infty + 4m_b\beta^{-1}\right) + 2\sqrt{2/\beta}\int_0^t |\Theta^\varepsilon_{b,s}|\mathrm{d}B_s \,.$$

The Burkholder-Davis-Gundy inequality and (55) gives

$$\mathbb{E}\left(\sup_{t \in [0,T]} |\Theta^\varepsilon_{b,t}|\right) \leqslant 8\beta^{-1}TM_b,$$

for any $\varepsilon \in (0, 1]$, and similarly for $|\Theta^0_{b,t}|$. The Markov inequality concludes the proof. □

4.2.3 Averaging for the truncated process

In this section, for a fixed $R \geqslant 1$ and g given in (53), we consider the truncated processes

$$\begin{cases} \partial_t \rho_t^\varepsilon &= \frac{1}{\varepsilon} \nabla \cdot (\rho_t^\varepsilon \left[D\mathcal{E}_0(\rho_t^\varepsilon, \cdot) - g\left(\Psi_t^\varepsilon\right) \cdot \nabla \varphi \right]) + \frac{1}{\varepsilon} \Delta \rho_t^\varepsilon \\ \mathrm{d}\Psi_t^\varepsilon &= -\Psi_t^\varepsilon + \rho_t^\varepsilon(\varphi) + \sqrt{2/\beta} \mathrm{d}B_t \,, \end{cases} \tag{56}$$

and

$$\mathrm{d}\Psi_t^0 = -\Psi_t^0 + \mu_{g(\Psi_t^0)}(\varphi) + \sqrt{2/\beta} \mathrm{d}B_t \,, \tag{57}$$

with initial conditions $\Psi_0^\varepsilon = \Psi_0^0 = \Theta_0$ and $\rho_0^\varepsilon = \nu_0$. The goal of this section is to establish the following:

Proposition 8 *Under the settings of Theorem 3, there exists $\varepsilon_0 \in (0,1)$ such that for any $T > 0$, there exists $C_T > 0$ such that, for all $\varepsilon \in (0, \varepsilon_0]$, setting $t_\varepsilon := \frac{2\varepsilon}{\lambda} \ln \varepsilon^{-1}$,*

$$\sup_{t \in [0,T]} \mathbb{E}\left(|\Psi_t^\varepsilon - \Psi_t^0|^2 \right) + \sup_{t \in [t_\varepsilon, T]} \mathbb{E}\left(\mathcal{W}_2^2(\rho_t^\varepsilon, \mu_{g(\Psi_t^0)}) \right) \leqslant C_T \varepsilon \ln \varepsilon^{-1} \,. \tag{58}$$

Remark 2 In the case where φ is linear (i.e. $m_b = 0$) then $g(\theta) = \theta$ for all $\theta \in \mathbb{R}^m$, hence the processes (56) and (57) are equal to the original ones (25) and (27). In this situation, Proposition 8 is a stronger, more quantitative result than Theorem 3.

In order to prove Proposition 8, we start with some auxiliary results.

Lemma 9 *Under the settings of Theorem 3, there exists $\varepsilon_0 > 0$ such that, for all $T > 0$,*

$$\mathrm{M}_T := \sup_{t \in [0,T]} \sup_{\varepsilon \in (0, \varepsilon_0]} \mathbb{E}\left(|\Psi_t^\varepsilon|^2 + \widetilde{\mathcal{F}}_{g(\Psi_t^\varepsilon)}(\rho_t^\varepsilon) + |\rho_t^\varepsilon(\varphi)|^2 \right) < \infty \,. \tag{59}$$

Proof To alleviate notations we write $(\Psi, \rho) = (\Psi^\varepsilon, \rho^\varepsilon)$. Let λ be given by (54), and recall that g is 1-Lipschitz continuous and $|\Delta g| \leqslant C_g$. Applying Itō formula to $(\nu, \theta) \mapsto \mathcal{F}_{g(\theta)}(\nu)$,

$$\begin{aligned} \partial_t \mathbb{E}\left(\mathcal{F}_{g(\Psi_t)}(\rho_t) \right) &= \mathbb{E}\left(-\frac{1}{\varepsilon} \mathcal{I}_{g(\Psi_t)}(\rho_t) + \rho_t(\varphi) \cdot \nabla g(\Psi_t) \left(\rho_t(\varphi) - \Psi_t \right) + \rho_t(\varphi) \cdot \beta^{-1} \Delta g(\Psi_t) \right) \\ &\leqslant \mathbb{E}\left(-\frac{2\lambda}{\varepsilon} \widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t) + 3|\rho_t(\varphi)|^2 + |\Psi_t|^2 \right) + C_g^2 \beta^{-2} \,. \end{aligned} \tag{60}$$

Besides, writing L_φ the Lipschitz constant of φ,

$$\begin{aligned} |\rho_t(\varphi)|^2 &\leqslant 3|\rho_t(\varphi) - \mu_{g(\Psi_t)}(\varphi)|^2 + 3|\mu_{g(\Psi_t)}(\varphi) - \mu_0(\varphi)|^2 + 3|\mu_0(\varphi)|^2 \\ &\leqslant 3L_\varphi^2 \mathcal{W}_2^2(\rho_t, \mu_{g(\Psi_t)}) + 3L_\varphi^2 \mathcal{W}_2^2(\mu_{g(\Psi_t)}, \mu_0) + 3|\mu_0(\varphi)|^2 \\ &\leqslant 6L_\varphi^2 \lambda^{-1} \widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t) + 27 L_\varphi^4 \lambda^{-2} |\Psi_t|^2 + 3|\mu_0(\varphi)|^2 \,, \end{aligned} \tag{61}$$

where we used the transport inequality (52), Lemma 5 and that $|g(\psi)| \leqslant |\psi|$. Finally,

$$\partial_t \mathbb{E}\left(|\Psi_t|^2 \right) = \frac{4m}{\beta} + 2\mathbb{E}\left(\Psi_t(-\Psi_t + \rho_t(\varphi)) \right) \leqslant \frac{4m}{\beta} + \mathbb{E}\left(|\rho_t(\varphi)|^2 \right) .$$

Combining the three previous inequalities, we get a constant $C_0 > 0$ independent from ε such that

$$\partial_t \mathbb{E}\left(\mathcal{F}_{g(\Psi_t)}(\rho_t) + |\Psi_t|^2\right) \leqslant C_0 + \mathbb{E}\left(\left(C_0 - \frac{2\lambda}{\varepsilon}\right)\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t) + C_0|\Psi_t|^2\right).$$

Setting $\varepsilon_0 = 2\lambda/C_0$ we get, for any $\varepsilon \in (0, \varepsilon_0]$,

$$\partial_t \mathbb{E}\left(\mathcal{F}_{g(\Psi_t)}(\rho_t) + |\Psi_t|^2\right) \leqslant C_0 + C_0\mathbb{E}\left(|\Psi_t|^2\right).$$

From (46),

$$0 \leqslant \mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t)\right) \leqslant \mathbb{E}\left(\mathcal{F}_{g(\Psi_t)}(\rho_t) + C_1(1 + |\Psi_t|^2)\right)$$

and thus we get, for some $C_2 > 0$,

$$0 \leqslant \mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t) + |\Psi_t|^2\right) \leqslant \mathbb{E}\left(\mathcal{F}_{g(\Psi_0)}(\rho_0) + |\Psi_0|^2\right) + C_2 t + C_2\int_0^t |\Psi_s|^2 \mathrm{d}s$$

Grönwall lemma and (61) concludes the proof.

□

In the rest of this section, ε_0 is as given by Lemma 9.

Lemma 10 *Under the settings of Theorem 3, for all $T > 0$, there exists $\mathrm{K}_T > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$, $t \in [0, T]$ and $s \in (0, 1]$,*

$$\mathbb{E}\left(|\Psi_t^\varepsilon - \Psi_{t+s}^\varepsilon|^2\right) \leqslant \mathrm{K}_T s\,.$$

Proof This directly follows from Lemma 9 and

$$\mathbb{E}\left(|\Psi_t^\varepsilon - \Psi_{t+s}^\varepsilon|^2\right) \leqslant 2s\int_t^{t+s} \mathbb{E}\left(|\Psi_u^\varepsilon|^2 + |\rho_u(\varphi)|^2\right)\mathrm{d}u + \frac{4}{\beta}\mathbb{E}\left(|B_{t+s} - B_t|^2\right).$$

□

Proposition 11 *Under the settings of Theorem 3, for all $T > 0$, there exists $\mathrm{K}'_T > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$, setting $t_\varepsilon := \frac{2\varepsilon}{\lambda}\ln\varepsilon^{-1}$ (with λ given by (54)),*

$$\sup_{t\in[t_\varepsilon, T]} \mathbb{E}\left(\mathcal{W}_2^2(\mu_{g(\Psi_t^\varepsilon)}, \rho_t^\varepsilon)\right) \leqslant \mathrm{K}'_T\varepsilon\ln\varepsilon^{-1}\,.$$

Proof Again, we simply write $(\Psi_t^\varepsilon, \rho_t^\varepsilon) = (\Psi_t, \rho_t)$. Proceeding as in the proof Lemma 5,

$$\begin{aligned}
\partial_s \mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_{t+s})\right) &= \partial_{|u=0}\mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_{t+s})}(\rho_{t+s+u}) + (\Psi_{t+s} - \Psi_t)\cdot\rho_{s+t+u}(\varphi)\right)\\
&= \mathbb{E}\left(-\frac{1}{\varepsilon}\mathcal{I}_{g(\Psi_{t+s})}(\rho_{t+s}) + (\Psi_{t+s} - \Psi_t)\cdot\partial_s\rho_{s+t}(\varphi)\right)\\
&\leqslant \mathbb{E}\left(-\frac{1}{4\varepsilon}\mathcal{I}_{g(\Psi_{t+s})}(\rho_{t+s}) + \frac{2L_\varphi^2}{\varepsilon}|\Psi_{t+s} - \Psi_t|^2\right).
\end{aligned}$$

Using the PL inequality (28) and Lemma 10 gives, for $t \in [0, T]$ and $s \in [0, 1]$;

$$\partial_s \mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_{t+s})\right) \leqslant -\frac{\lambda}{2\varepsilon}\mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_{t+s})\right) + \frac{2\mathrm{K}_T L_\varphi^2}{\varepsilon}s\,,$$

which we integrate over $s \in [0, \delta]$ for some $\delta \in (0, 1]$ to get

$$\mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_{t+\delta})\right) \leqslant e^{-\frac{\delta\lambda}{2\varepsilon}}\mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_t)\right) + \frac{1-e^{-\frac{\delta\lambda}{2\varepsilon}}}{\lambda/(2\varepsilon)} \times \frac{2\mathrm{K}_T L_\varphi^2}{\varepsilon}\delta \leqslant \mathrm{M}_T e^{-\frac{\delta\lambda}{2\varepsilon}} + \frac{4\mathrm{K}_T L_\varphi^2}{\lambda}\delta\,.$$

Then, applying Lemma 6,

$$\begin{aligned}\mathbb{E}\left(\widetilde{\mathcal{F}}_{g(\Psi_{t+\delta})}(\rho_{t+\delta})\right) &\leqslant \mathbb{E}\left(2\widetilde{\mathcal{F}}_{g(\Psi_t)}(\rho_{t+\delta}) + \frac{4L_\varphi^2}{\lambda}|\Psi_{t+\delta}-\Psi_t|^2\right)\\ &\leqslant 2\mathrm{M}_T e^{-\frac{\delta\lambda}{2\varepsilon}} + \frac{12L_\varphi^2}{\lambda}\mathrm{K}_T\delta\,.\end{aligned}$$

Applying this for $\delta = t_\varepsilon$ and using (52) concludes the proof. □

Proof of Proposition 8 Since the processes are driven by the same Brownian motion B,

$$\Psi_t^\varepsilon - \Psi_t^0 = \int_0^t \left(\Psi_s^0 - \Psi_s^\varepsilon + \rho_s^\varepsilon(\varphi) - \mu_{g(\Psi_s^0)}(\varphi)\right)\mathrm{d}s\,.$$

Using Lemma 5 and that g is 1-Lipschitz continuous,

$$\begin{aligned}|\Psi_t^\varepsilon - \Psi_t^0| &\leqslant \int_0^t \left((1+3L_\varphi\lambda^{-1})|\Psi_s^0-\Psi_s^\varepsilon| + |\rho_s^\varepsilon(\varphi) - \mu_{g(\Psi_s^\varepsilon)}(\varphi)|\right)\mathrm{d}s\\ &\leqslant \int_0^t \left((1+3L_\varphi\lambda^{-1})|\Psi_s^0-\Psi_s^\varepsilon| + L_\varphi\mathcal{W}_2(\rho_s^\varepsilon, \mu_{g(\Psi_s^\varepsilon)})\right)\mathrm{d}s\,.\end{aligned}$$

Taking the square, the expectation and applying Grönwall's lemma, for any $T > 0$, there exists $C_T > 0$ such that

$$\sup_{t\in[0,T]}\mathbb{E}\left(|\Psi_t^\varepsilon - \Psi_t^0|^2\right) \leqslant C_T\int_0^T \mathbb{E}\left(\mathcal{W}_2^2(\rho_s^\varepsilon, \mu_{g(\Psi_s^\varepsilon)})\right)\mathrm{d}s\,.$$

Thanks to Lemma 9 (with (52)) and Proposition 11, distinguishing whether $t \gtrless t_\varepsilon$,

$$\int_0^T \mathbb{E}\left(\mathcal{W}_2^2(\rho_s^\varepsilon, \mu_{g(\Psi_s^\varepsilon)})\right)\mathrm{d}s \leqslant 2\lambda^{-1}\mathrm{M}_T t_\varepsilon + T\mathrm{K}'_T\varepsilon\ln\varepsilon^{-1}\,,$$

which concludes the proof of the bound for the first term in the left-hand side of (33). For the second term, it remains to bound, for $t \geqslant t_\varepsilon$,

$$\mathbb{E}\left(\mathcal{W}_2^2(\rho_t^\varepsilon, \mu_{g(\Psi_t^0)})\right) \leqslant 2\mathbb{E}\left(\mathcal{W}_2^2(\rho_t^\varepsilon, \mu_{g(\Psi_t^\varepsilon)}) + 9L_\varphi^2\lambda^{-2}|\Psi_t^\varepsilon - \Psi_t^0|^2\right)\,.$$

The bound already proven for the first term of (33) and Proposition 11 concludes the proof of Proposition 8. □

4.2.4 Conclusion of the proof

Proof of Theorem 3 Fix $T > t_0 > 0$. For an arbitrary $R \geqslant 1$, consider the event

$$\mathcal{A}_R = \left\{\sup_{t\in[0,T]}|\Theta_{b,t}^\varepsilon| + |\Theta_{b,t}^0| \leqslant R\right\}\,,$$

and let g be given by (53) and $\Psi^\varepsilon, \rho^\varepsilon, \Psi^0$ as in (56) and (57). Under the event $\mathcal{A}_R$, for all $t \in [0,T]$, $(\Psi_t^\varepsilon, \rho_t^\varepsilon, \Psi_t^0) = (\Theta_t^\varepsilon, \nu_t^\varepsilon, \Theta_t^0)$. For $t \in [0,T]$, we bound

$$\mathbb{P}\left(|\Theta_t^\varepsilon - \Theta_t^0| \geqslant \delta\right) \leqslant \mathbb{P}\left(|\Psi_t^\varepsilon - \Psi_t^0| \geqslant \delta\right) + \mathbb{P}(\mathcal{A}_R^c)\,.$$

By the Markov inequality and Proposition 8,

$$\limsup_{\varepsilon\to 0} \sup_{t\in[0,T]} \mathbb{P}\left(|\Theta_t^\varepsilon - \Theta_t^0| \geqslant \delta\right) \leqslant \mathbb{P}(\mathcal{A}_R^c)\,.$$

Using that, for any $t_0 > 0$, for any $R \geqslant 1$, $t_\varepsilon \leqslant t_0$ for ε small enough (with t_ε as in Proposition 8; it depends on R through λ), we get with the same argument that

$$\limsup_{\varepsilon\to 0} \sup_{t\in[t_0,T]} \mathbb{P}\left(\mathcal{W}_2(\nu_t^\varepsilon, \mu_{\Theta_t^0}) \geqslant \delta\right) \leqslant \mathbb{P}(\mathcal{A}_R^c)\,.$$

Letting $R \to \infty$ thanks to Lemma 7 concludes the proof. □

Acknowledgements. The research of P. Monmarché is supported by the project CONVIVIALITY (ANR-23-CE40-0003) of the French National Research Agency. Generative AI tools have been used only for English language in Section 3.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[2] Letizia Angeli, Julien Barré, Martin Kolodziejczyk, and Michela Ottobre. Well-posedness and stationary solutions of mckean-vlasov (s) pdes. *Journal of Mathematical Analysis and Applications*, 526(2):127301, 2023.

[3] Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in neural information processing systems*, 32, 2019.

[4] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

[5] Kaveh Bashiri and Anton Bovier. Gradient flow approach to local mean-field spin systems. *Stochastic Processes and their Applications*, 130(3):1461–1514, 2020.

[6] Roland Bauerschmidt and Thierry Bodineau. A very simple proof of the lsi for high temperature spin systems. *Journal of Functional Analysis*, 276(8):2582–2588, 2019.

[7] Roland Bauerschmidt, Thierry Bodineau, and Benoît Dagallier. A criterion on the free energy for log-sobolev inequalities in mean-field particle systems. *arXiv preprint arXiv:2503.24372*, 2025.

[8] Zineb Belkacemi, Paraskevi Gkeka, Tony Lelievre, and Gabriel Stoltz. Chasing collective variables using autoencoders and biased trajectories. *Journal of chemical theory and computation*, 18(1):59–78, 2022.

[9] Armand Bernou and Mitia Duerinckx. Uniform-in-time estimates on corrections to mean field for interacting brownian particles: A. bernou, m. duerinckx. *Probability Theory and Related Fields*, 194(3):1601–1710, 2026.

[10] Nawaf Bou-Rabee and Andreas Eberle. Two-scale coupling for preconditioned hamiltonian monte carlo in infinite dimensions. *Stochastics and Partial Differential Equations: Analysis and Computations*, 9:207–242, 2021.

[11] Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

[12] José A. Carrillo, Robert J. McCann, and Cédric Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Revista Matemática Iberoamericana*, 19(3):971 – 1018, 2003.

[13] Omar Chehab, Anna Korba, Austin Stromme, and Adrien Vacher. Provable convergence and limitations of geometric tempering for langevin dynamics. In *International Conference on Learning Representations*, volume 2025, pages 5451–5493, 2025.

[14] Lénaïc Chizat. Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*, 2022.

[15] Donald A Dawson. Critical dynamics and fluctuations for a mean-field model of cooperative behavior. *Journal of Statistical Physics*, 31(1):29–85, 1983.

[16] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.

[17] François Delarue and William RP Hammersley. Rearranged stochastic heat equation. *Probability Theory and Related Fields*, pages 1–62, 2024.

[18] François Delarue and William RP Hammersley. Rearranged stochastic heat equation: ergodicity and related gradient descent on the space of probability measures. *arXiv preprint arXiv:2403.16140*, 2024.

[19] François Delarue, Etienne Tanré, and Raphaël Maillet. Ergodicity of some stochastic fokker-planck equations with additive common noise. *arXiv preprint arXiv:2405.09950*, 2024.

[20] Matías G Delgadino, Rishabh S Gvalani, Grigorios A Pavliotis, and Scott A Smith. Phase transitions, logarithmic sobolev inequalities, and uniform-in-time propagation of chaos for weakly interacting diffusions. *Communications in*

Mathematical Physics, pages 1–49, 2023.

[21] Lorenzo Dello Schiavo. The dirichlet–ferguson diffusion on the space of probability measures over a closed riemannian manifold. *The Annals of Probability*, 50(2):591–648, 2022.

[22] Begona Fernandez and Sylvie Méléard. A hilbertian approach for fluctuations on the mckean-vlasov model. *Stochastic processes and their applications*, 71(1):33–53, 1997.

[23] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

[24] NJ Gerber, Rishabh S Gvalani, Martin Hairer, Grigorios A Pavliotis, and André Schlichting. Formation of clusters and coarsening in weakly interacting diffusions. *Nonlinearity*, 39(7):075023, 2026.

[25] Pierre Germain and Pierre Monmarché. Stochastic moments dynamics: a flexible finite-dimensional random perturbation of wasserstein gradient descent. *arXiv preprint arXiv:2505.07448*, 2025.

[26] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

[27] Arnaud Guillin, Wei Liu, Liming Wu, and Chaoen Zhang. Uniform Poincaré and logarithmic Sobolev inequalities for mean field particle systems. *The Annals of Applied Probability*, 32(3):1590 – 1614, 2022.

[28] Kaitong Hu, Zhenjie Ren, David Šiška, and Łukasz Szpruch. Mean-field Langevin dynamics and energy landscape of neural networks. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 57(4):2043 – 2065, 2021.

[29] Lucas Journel and Pierre Monmarché. Convergence of the kinetic annealing for general potentials. *Electronic Journal of Probability*, 27(none):1 – 37, 2022.

[30] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447, 2022.

[31] Tony Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical Mathematics and Advanced Applications 2011: Proceedings of ENUMATH 2011, the 9th European Conference on Numerical Mathematics and Advanced Applications, Leicester, September 2011*, pages 791–810. Springer, 2012.

[32] Tony Lelièvre, Xuyang Lin, and Pierre Monmarché. Convergence rates for an adaptive biasing potential scheme from a wasserstein optimization perspective. *Nonlinearity*, 39(4):045016, 2026.

[33] Tony Lelievre, Mathias Rousset, and Gabriel Stoltz. *Free energy computations*. Imperial College Press, London, 2010.

[34] Wei Liu and Liming Wu. Large deviations for empirical measures of mean-field gibbs measures. *Stochastic Processes and their Applications*, 130(2):503–520, 2020.

[35] Yulong Lu and Pierre Monmarché. Convergence of Time-Averaged Mean Field Gradient Descent Dynamics for Continuous Multi-Player Zero-Sum Games. *arXiv e-prints*, page arXiv:2505.07642, May 2025.

[36] Raphael Maillet. A note on the long-time behaviour of stochastic mckean-vlasov equations with common noise. *arXiv preprint arXiv:2306.16130*, 2023.

[37] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new monte carlo scheme. *EPL (Europhysics Letters)*, 19(6):451–458, 1992.

[38] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[39] Pierre Monmarché. Long-time propagation of chaos and exit times for metastable mean-field particle systems. *arXiv e-prints*, page arXiv:2503.00157, February 2025.

[40] Pierre Monmarché. Equivalence between N-particle log-Sobolev inequalities and non-linear Łojasiewicz inequalities for a class of mean field systems. *arXiv e-prints*, page arXiv:2609.06574, September 2026.

[41] Pierre Monmarché. Uniform log-Sobolev inequalities for mean field particles beyond flat-convexity. *to appear in Annales de la Faculté des Sciences de Toulouse*, 2026.

[42] Pierre Monmarché and Julien Reygner. Local convergence rates for wasserstein gradient flows and mckean-vlasov equations with multiple stationary solutions. *Probability Theory and Related Fields*, pages 1–59, 2025.

[43] Radford M Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.

[44] Ludger Overbeck, Michael Röckner, and Byron Schmuland. An analytic approach to fleming-viot processes with interactive selection. *The Annals of Probability*, pages 1–36, 1995.

[45] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

[46] Panpan Ren, Feng-Yu Wang, and Simon Wittmann. Diffusion processes on p-wasserstein space over banach space. *arXiv preprint arXiv:2402.15130*, 2024.

[47] Panpan Ren, Feng-Yu Wang, and Simon Wittmann. Markov processes and stochastic extrinsic derivative flows on the space of absolutely continuous measures. *arXiv preprint arXiv:2408.15687*, 2024.

[48] Etienne Sandier and Sylvia Serfaty. Gamma-convergence of gradient flows with applications to ginzburg-landau. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(12):1627–1672, 2004.

[49] Gaëtan Serré, Pierre Germain, Samuel Gruffaz, and Argyris Kalogeratos. Enhancing Exploration in Global Optimization by Noise Injection in the Probability Measures Space. *arXiv e-prints*, page arXiv:2601.22753, January 2026.

[50] Anna Shalova and André Schlichting. Solutions of stationary mckean–vlasov equation on a high-dimensional sphere and other riemannian manifolds. *Advances in Nonlinear Analysis*, 15(1):20250141, 2026.

[51] Jinghai Shao. A New Probability Measure-Valued Stochastic Process with Ferguson-Dirichlet Process as Reversible Measure. *Electronic Journal of Probability*, 16(none):271 – 292, 2011.

[52] Julian Tugaut. On the steady states for the granular media equation: existence, local uniqueness, local stability and rate of convergence. *Stochastics*, 98(1):17–34, 2026.

[53] Max-K. von Renesse and Karl-Theodor Sturm. Entropic measure and Wasserstein diffusion. *The Annals of Probability*, 37(3):1114 – 1191, 2009.

[54] Songbo Wang. Uniform log-sobolev inequalities for mean field particles with flat-convex energy. *arXiv preprint arXiv:2408.03283*, 2024.