# Analysis of an Adaptive Biasing Force method based on self-interacting dynamics

Michel Benaïm, Charles-Edouard Bréhier, Pierre Monmarché

# Analysis of an Adaptive Biasing Force method based on self-interacting dynamics

Michel Benaïm [*,1], Charles-Edouard Bréhier [†,2], and Pierre Monmarché [‡,3]

[1]Institut de Mathématiques, Université de Neuchâtel, Switzerland
[2]Univ Lyon, CNRS, Université Claude Bernard Lyon 1, UMR5208, Institut Camille Jordan, F-69622 Villeurbanne, France
[3]Sorbonne-Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France

October 10, 2019

**Abstract**

This article fills a gap in the mathematical analysis of Adaptive Biasing algorithms, which are extensively used in molecular dynamics computations. Given a reaction coordinate, ideally, the bias in the overdamped Langevin dynamics would be given by the gradient of the associated free energy function, which is unknown. We consider an adaptive biased version of the overdamped dynamics, where the bias depends on the past of the trajectory and is designed to approximate the free energy.

The main result of this article is the consistency and efficiency of this approach. More precisely we prove the almost sure convergence of the bias as time goes to infinity, and that the limit is close to the ideal bias, as an auxiliary parameter of the algorithm goes to 0.

The proof is based on interpreting the process as a self-interacting dynamics, and on the study of a non-trivial fixed point problem for the limiting flow obtained using the ODE method.

## 1 Introduction

Let $\mu_\star$ be a probability distribution on the $d$-dimensional flat torus $\mathbb{T}^d$, of the type:

$$d\mu_\star(x) = \frac{e^{-\beta V(x)}}{Z(\beta)} dx \ , \quad Z(\beta) = \int_{\mathbb{T}^d} e^{-\beta V(x)} dx, \tag{1}$$

[*]michel.benaim@unine.ch
[†]brehier@math.univ-lyon1.fr
[‡]pierre.monmarche@sorbonne-universite.fr

where $dx$ is the normalized Lebesgue measure on $\mathbb{T}^d$. For applications in physics and chemistry (e.g. in molecular dynamics), $\mu_\star$ is referred to as the Boltzmann-Gibbs distribution associated with the potential energy function $V$ and the inverse temperature parameter $\beta > 0$. For applications in statistics (e.g. in Bayesian statistics), $-\beta V$ is referred to as the log-likelihood. In this article, the function $V : \mathbb{T}^d \to \mathbb{R}$ is assumed to be smooth.

In order to estimate integrals of the type $\int \varphi d\mu_\star$, with $\varphi : \mathbb{T}^d \to \mathbb{R}$, probabilistic methods are used, especially when $d$ is large. The Markov Chain Monte Carlo (MCMC) method consists in interpreting the integral as the (almost sure) limit

$$\int \varphi d\mu_\star = \lim_{T\to\infty} \frac{1}{T}\int_0^T \varphi(X_t^0)dt = \lim_{T\to\infty} \int \varphi d\mu_T^0,$$

where $\mu_t^0 = \frac{1}{T}\int_0^T \delta_{X_t^0} dt$ is the random empirical distribution associated with an ergodic Markov process $\left(X_t^0\right)_{t\geqslant 0}$, with unique invariant distribution $\mu_\star$. The choice of the Markov dynamics is not unique, and in this work we consider the overdamped Langevin dynamics

$$dX_t^0 = -\nabla V(X_t^0)dt + \sqrt{2\beta^{-1}}dW_t$$

where $\left(W_t\right)_{t\geqslant 0}$ is a $d$-dimensional Wiener process. In practice, discrete-time Markov processes, defined for instance using the Metropolis-Hastings algorithm, are employed.

The convergence to equilibrium requires that the Markov process explores the entire energy landscape, which may be a very slow process. Indeed, in practical problems, the dimension $d$, *i.e.* the number of degrees of freedom in the system, is very large, and the probability distribution $\mu_\star$ is multimodal: the function $V$ admits several local minima (interpreted as potential energy wells) and $\beta$ is large. In that situation, the Markov process is metastable: when it reaches an energy well, it tends to stay there for a long time (whose expectation goes to infinity when $\beta$ goes to infinity) before hopping to another energy well. Asymptotic results for the exit time from energy wells when $\beta \to \infty$ are given by Eyring-Kramers type formulas [14, 27]. The metastability of the process substantially slows down the exploration of the energy landscape, hence the convergence when $T \to \infty$ towards the target quantity $\int \varphi d\mu_\star$.

In the development of Monte-Carlo methods in the last decades, many techniques have been studied in order to efficiently sample multimodal distributions. The bottom-line strategy to enhance sampling consists in biasing the dynamics and in reweighting the averages: indeed, for any smooth function $\tilde{V} : \mathbb{T}^d \to \mathbb{R}$, one has

$$\int \varphi d\mu_\star = \frac{\int \varphi e^{-\beta V}}{\int e^{-\beta V}} = \frac{\int \varphi e^{-\beta(V-\tilde{V})}\ e^{-\beta\tilde{V}}}{\int e^{-\beta(V-\tilde{V})}\ e^{-\beta\tilde{V}}} = \lim_{t\to\infty} \frac{\int_0^t \varphi(\tilde{X}_s)e^{-\beta(V(\tilde{X}_s)-\tilde{V}(\tilde{X}_s)}ds}{\int_0^t e^{-\beta(V(\tilde{X}_s)-\tilde{V}(\tilde{X}_s)}ds},$$

where the biased dynamics is given by $d\tilde{X}_t = -\nabla\tilde{V}(\tilde{X}_t)dt + \sqrt{2\beta^{-1}}dW_t$. This is nothing but an Importance Sampling method, and choosing carefully the function $\tilde{V}$ may substantially reduce the computational cost. Indeed, if the distribution with density proportional to $e^{-\beta\tilde{V}(x)}$ is not multimodal, the biased process $\tilde{X}_t$ converges to equilibrium and explores the

state space faster than the unbiased process $X_t$. In the sequel, we explain how to choose $\tilde{V}$ in order to benefit from the importance sampling strategy.

From now on, in order to simplify the notation, $\beta = 1$. In addition, without loss of generality, assume that $\int_{\mathbb{T}^d} e^{-V(x)} dx = 1$.

Instead of treating the problem in an intractable full generality, we focus on the typical situation when some additional a priori knowledge on the system is available. Precisely, let $\xi : \mathbb{T}^d \to \mathbb{T}^m$ be a smooth function, which is referred to as the reaction coordinate (following the terminology employed in the molecular dynamics community). Let us stress that the identification of appropriate reaction coordinates is a delicate question, which depends on the system at hand. The problem of automatic learning of good reaction coordinates currently generates a lot of research, see for instance [13, 15] and references within. We do not consider this question in the sequel.

The biasing potential in the importance sampling schemes considered in this work will be of the type $\tilde{V}(x) = V(x) - A(\xi(x))$, where $A : \mathbb{T}^m \to \mathbb{R}$. In practice, the number of macroscopic variables $m$ is very small compared to the dimension $d$ of the model (which describes the full microscopic system). As will be explained below, without loss of generality, we assume that $\xi(x) = \xi(y, z) = z$ for all $x = (y, z) \in \mathbb{T}^{d-m} \times \mathbb{T}^m$. This expression for the reaction coordinate simplifies the presentation of the method, however considering more general reaction coordinates $\xi$ is possible up to adapting some definitions below. To explain the construction of the method and to justify its efficiency, we assume that the reaction coordinate is representative of the metastable behavior of the system: roughly, this means that only the exploration in the $z$ variable is affected by the metastability, whereas the exploration in the $y$ variable is much faster.

In this framework, the fundamental object is the free energy function $A_\star$ defined as follows: for all $z \in \mathbb{T}^m$,

$$A_\star(z) = -\log\Big(\int_{\mathbb{T}^{d-m}} e^{-V(y,z)} dy\Big). \tag{2}$$

For general considerations on the free energy and related computational aspects, we refer to [31, 32]. By construction, if $X = (Y, Z)$ is a random variable with distribution $\mu_\star$, then the marginal distribution of $Z$ is given by

$$d\nu_\star(z) = e^{-A_\star(z)} dz.$$

Introduce the notation $(Y_t^0, Z_t^0) = X_t^0$ for the solution of the overdamped Langevin dynamics

$$\begin{cases} dY_t^0 = -\nabla_y V(Y_t^0, Z_t^0) dt + \sqrt{2} dW_t^{(d-m)}, \\ dZ_t^0 = -\nabla_z V(Y_t^0, Z_t^0) dt + \sqrt{2} dW_t^{(m)}, \end{cases}$$

where $W_t = (W_t^{(d-m)}, W_t^{(m)})$. It $\nu_t^0 = \frac{1}{t}\int_0^t \delta_{Z_s^0} ds$ denotes the empirical distribution for the variable $Z^0$, then almost surely

$$\nu_t^0 \underset{t\to\infty}{\to} \nu_\star,$$

in the sense of weak convergence in the set $\mathcal{P}(\mathbb{T}^m)$ of probability distributions on $\mathbb{T}^m$. Since the reaction coordinate is representative of the metastability of the system, this convergence shares the same computational issues as when considering the full process $X^0$.

A much better performance can be attained considering the following biased dynamics, where $V(x)$ is replaced by $\tilde{V}_\star(x) = V(x) - A_\star(\xi(x))$:

$$\begin{cases} dY_t^\star = -\nabla_y V(Y_t^\star, Z_t^\star)dt + \sqrt{2}dW_t^{(d-m)}, \\ dZ_t^\star = -\nabla_z V(Y_t^\star, Z_t^\star)dt + \nabla A_\star(Z_t^\star)dt + \sqrt{2}dW_t^{(m)}. \end{cases}$$

Define the associated empirical measures on $\mathbb{T}^d$ and $\mathbb{T}^m$ respectively:

$$\mu_t^\star = \frac{1}{t}\int_0^t \delta_{X_s^\star} ds \ , \quad \nu_t^\star = \frac{1}{t}\int_0^t \delta_{Z_s^\star} ds,$$

where $X_s^\star = (Y_s^\star, Z_s^\star)$. As explained above, $\int \varphi d\mu_\star$ can then be computed by the reweighting procedure. Observe that by ergodicity for $\big(X_t^\star\big)_{t\geqslant 0}$ and the definition of $A_\star$, one has

$$\nu_t^\star \underset{t\to\infty}{\to} dz,$$

*i.e.* at the limit the distribution of $Z_t^\star$ is uniform on $\mathbb{T}^m$. This observation, which is referred to as the *flat histogram property* in the literature devoted to applications, means that the process $X^\star$ does not suffer from slow convergence to equilibrium due to energy barriers, compared to the process $X^0$.

In practive, the free energy function $A_\star$ is not known, thus the ideal approach described above is not applicable. In fact, in many applications, the real objective is the computation of the free energy function. One of the important features of many free energy computation algorithms, such as the one studied in this work, is to compute an approximation of the free energy function on-the-fly, and to use this approximation to enhance sampling. Checking that such adaptive algorithms are efficient and consistent requires careful mathematical analysis.

In this article, we consider a class of adaptive biasing methods, where the dynamics is of the form

$$\begin{cases} dY_t = -\nabla_y V(Y_t, Z_t)dt + \sqrt{2}dW_t^{(d-m)}, \\ dZ_t = -\nabla_z V(Y_t, Z_t)dt + \nabla A_t(Z_t)dt + \sqrt{2}dW_t^{(m)}, \end{cases} \tag{3}$$

where the function $A_t$ depends on time $t$, approximates $A_\star$ when $t \to \infty$, and is defined in terms of the empirical measure

$$\mu_t = \frac{1}{t}\int_0^t \delta_{X_s} ds. \tag{4}$$

The process $\big(X_t\big)_{t\geqslant 0} = \big(Y_t, Z_t\big)_{t\geqslant 0}$ is not a Markov process, instead it is a self-interacting diffusion process. The precise construction of the algorithm studied in this article is provided below.

This article is organized as follows. The construction of the algorithm (9) studied in this work is presented in Section 2 below. The main result, Theorem 2.3, is stated in Section 2.3, and a comparison with the literature is given. Section 3 gives a proof of the well-posedness of the self-interacting dynamics (9) (Proposition 2.2). Section 4 exhibits the limiting flow (obtained by applying the ODE method) and establishes the asymptotic pseudotrajectory property. Finally, Section 5 provides the final crucial ingredients for the proof of the main result, Theorem 2.3: a PDE estimate which provides some uniform bounds, and a global asymptotic stability property for the limiting flow.

# 2 The Adaptive Biasing Force algorithm

The objectives of this section are to define the Adaptive Biasing Force method [17] studied in this article, and to state the main results.

Recall the definitions (1) and (2) of the target distribution $\mu_\star$ and of the free energy $A_\star$ respectively. The potential energy function $V$ is assumed to be of class $\mathcal{C}^\infty$.

The reaction coordinate $\xi : \mathbb{T}^d \to \mathbb{T}^m$ satisfies $\xi(y,z) = z$ for all $x = (y,z) \in \mathbb{T}^d$. This expression substantially simplifies the presentation compared with a more general choice of $\xi : \mathbb{T}^d \to \mathbb{R}^m$. In applications, this is not restrictive, and consists in considering the so-called extended ABF algorithm [22]. Precisely, an auxiliary variable $\mathcal{Z}$ is added to the state space, the extended potential energy function for $\overline{X} = (X, \mathcal{Z})$ is given by $\overline{V}(\overline{X}) = V(X) + \frac{1}{2\sigma^2}|\xi(X) - \mathcal{Z}|^2$, where $\sigma > 0$ is a small parameter, and one sets $\overline{\xi}(\overline{X}) = \mathcal{Z}$.

## 2.1 Construction

The definition of the algorithm requires to make precise how in the evolution equation (3), the biasing potential function $A_t$, or its gradient $\nabla A_t$, is determined in terms of the empirical distribution $\mu_t$ given by (4). The algorithm is based on the following identity: the gradient $\nabla A_\star$ of the free energy function $A_\star$ defined by (2) is given by

$$\nabla A_\star(z) = \frac{\int_{\mathbb{T}^{d-m}} \nabla_z V(y,z) e^{-V(y,z)} dy}{\int_{\mathbb{T}^{d-m}} e^{-V(y,z)} dy} = \mathbb{E}_{\mu_\star}\big[\nabla_z V(Y,Z) \;\big|\; Z = z\big]. \tag{5}$$

More generally, let $A : \mathbb{T}^m \to \mathbb{R}$ be a smooth function, and let $d\mu_\star^A(x) \propto e^{A(z)} d\mu_\star(y,z)$ be the ergodic invariant distribution of

$$\begin{cases} dY_t^A = -\nabla_y V(Y_t^A, Z_t^A)dt + \sqrt{2} dW_t^{(d-m)}, \\ dZ_t^A = -\nabla_z V(Y_t^A, Z_t^A)dt + \nabla A(Z_t^A)dt + \sqrt{2} dW_t^{(m)}. \end{cases}$$

Then one has the identity

$$\nabla A_\star(z) \;\; = \;\; \mathbb{E}_{\mu_\star^A}\big[\nabla_z V(Y,Z) \;\big|\; Z = z\big]. \tag{6}$$

The expressions for the gradient of the free energy function in equations (5) and (6) are simpler than (for instance) the expressions (5) and (6) in [30] which hold for a general reaction coordinate mapping $\xi$, whereas we consider only the case $\xi(y,z) = z$.

The occupation measures $\mu_t$ defined by (4) are in general singular with respect to the Lebesgue measure on $\mathbb{T}^m$. In order to define the mapping $\mu_t \mapsto A_t$, we introduce a regularization kernel $K_\epsilon$, depending on the parameter $\epsilon \in (0,1]$, such that

$$\nabla A_\star(z) = \lim_{\epsilon \to 0} \frac{\iint_{\mathbb{T}^d} \nabla_z V(y,z') K_\epsilon(z',z) d\mu_\star(y,z')}{\iint_{\mathbb{T}^d} K_\epsilon(z',z) d\mu_\star(y,z')}.$$

Indeed, formally, the expression (5) for $\nabla A_\star$ is obtained with the kernel $K_\epsilon(z,z')$ replaced by a Dirac distribution $\delta(z-z')$. See Assumption 2.1 below for precise conditions on the kernel function $K_\epsilon$.

For every $\epsilon \in (0,1]$ and $\mu \in \mathcal{P}(\mathbb{T}^d)$, define the mapping $F^\epsilon[\mu] : \mathbb{T}^m \to \mathbb{R}^m$ as follows:

$$F^\epsilon[\mu](\cdot) = \frac{\iint \nabla_z V(y,z) K_\epsilon(z,\cdot) d\mu(y,z)}{\iint K_\epsilon(z,\cdot) d\mu(y,z)}. \tag{7}$$

Due to the action of the regularization kernel $K_\epsilon$, in general $F^\epsilon[\mu]$ cannot be written as a gradient. For instance if $m = 1$, a smooth function $F : \mathbb{T} \to \mathbb{R}$ is a gradient if and only if its average value is zero $\int F(z)dz = 0$; in general, $\int F^\epsilon[\mu](z)dz \neq 0$.

The last ingredient in the construction is a projection operator $\mathbf{P}$, such that one defines $\nabla A^\epsilon[\mu] = \mathbf{P}(F^\epsilon[\mu])$. More precisely, for every $\epsilon \in (0,1]$ and $\mu \in \mathcal{P}(\mathbb{T}^d)$, define the mapping $A^\epsilon[\mu]$ as follows:

$$A^\epsilon[\mu] = \underset{A \in H^1(\mathbb{T}^m), \int A(z)dz = 0}{\operatorname{argmin}} \int \left| F^\epsilon[\mu](z) - \nabla A(z) \right|^2 dz \,. \tag{8}$$

As will be explained below, $A^\epsilon[\mu]$ is solution of an elliptic PDE. Note that $F^\epsilon[\mu]$ and $A^\epsilon[\mu]$ are functions depending only on $z \in \mathbb{T}^m$, with a dimension $m$ much smaller than $d$ the total number of degrees of freedom of the system. Typically, one has $m \in \{1,2,3\}$, which makes it possible to use the algorithm in practice.

We are now in position to define the process considered in this article: it is the solution of the system

$$\begin{cases} dY_t = -\nabla_y V(Y_t, Z_t)dt + \sqrt{2} dW_t^{(d-m)}, \\ dZ_t = -\nabla_z V(Y_t, Z_t)dt + \nabla A_t(Z_t)dt + \sqrt{2} dW_t^{(m)}, \\ A_t = A^\epsilon[\mu_t], \\ \mu_t = \frac{1}{t}\int_0^t \delta_{(Y_s, Z_s)} ds. \end{cases} \tag{9}$$

Arbitrary (deterministic) initial conditions $Y_0 = y_0 \in \mathbb{T}^{d-m}$, $Z_0 = z_0 \in \mathbb{T}^m$, $\mu_0 = \delta_{(y_0,z_0)}$ and $A_0 = A^\epsilon[\mu_0]$ are provided. This process belongs to the class of self-interacting diffusions, see [9, 10, 11, 12] for standard references.

## 2.2 Well-posedness of the system (9)

Recall that $V : \mathbb{T}^d \to \mathbb{R}$ is assumed to be of class $\mathcal{C}^\infty$. Let us first state the assumptions satisfied by the kernel function $K_\epsilon$.

**Assumption 2.1.** *For any $\epsilon \in (0,1]$, the mapping $K_\epsilon : \mathbb{T}^m \times \mathbb{T}^m \to (0,\infty)$ is of class $\mathcal{C}^\infty$ and positive.*

*For all $z \in \mathbb{T}^m$, one has*

$$\int K_\epsilon(z,\cdot)dz = \int K_\epsilon(\cdot,z)dz = 1$$

*In addition, if $\psi : \mathbb{T}^d \to \mathbb{R}$ is a continuous and bounded function, one has*

$$\iint_{\mathbb{T}^d} \psi(y,z') K_\epsilon(z',z) dy dz' \underset{\epsilon \to 0}{\to} \int_{\mathbb{T}^{d-m}} \psi(y,z) dy \;, \quad \forall\; z \in \mathbb{T}^m.$$

*Finally, there exists $c_K \in (0,\infty)$, such that*

$$\sup_{z\in\mathbb{T}^m} \int_{\mathbb{T}^m} |z-z'|^2\big(K_\epsilon(z',z)+K_\epsilon(z,z')\big)dz' \leqslant c_K\epsilon.$$

Define $m_\epsilon = \min_{z,z'\in\mathbb{T}^m} K_\epsilon(z',z)$ and $M_\epsilon^{(k)} = \max_{z,z'\in\mathbb{T}^m} |\nabla_z^k K_\epsilon(z',z)| + \max_{z,z'\in\mathbb{T}^m} |\nabla_{z'}^k K_\epsilon(z',z)|$, where $k$ is a nonnegative integer and $\nabla^k$ denotes the derivative of order $k$. Owing to Assumption 2.1, one has $m_\epsilon > 0$ and $M_\epsilon^{(k)} < \infty$ for all $\epsilon \in (0,1]$, however these estimates are not uniform with respect to $\epsilon$, *i.e.* $\inf_{\epsilon\in(0,1]} m_\epsilon = 0$ and $\sup_{\epsilon\in(0,1]} M_\epsilon^{(k)} = \infty$.

Note that to establish the well-posedness of the system (9), where $\epsilon \in (0,1]$ is fixed, upper bounds are allowed to depend on $\epsilon$. However, it will be crucial in Section 5 to derive some upper bounds which are uniform with respect to $\epsilon$ in order to prove the convergence when $t$ goes to infinity of $\mu_t$ and $A_t$ (to a limit depending on $\epsilon$), see Proposition 5.3.

The exact form of the kernel function $K_\epsilon$ has no influence on the analysis below. Let us give an example: let $K_\epsilon(z^1,z^2) = \prod_{j=1}^m k_\epsilon\big(z_j^2 - z_j^1\big)$, where for all $z \in \mathbb{T}$,

$$k_\epsilon(z) = Z_\epsilon^{-1} \exp\big(-\frac{\sin^2(z/2)}{\epsilon^2/2}\big)$$

is the so-called von-Mises kernel.

Owing to Assumption 2.1, it is straightforward to check that $F^\epsilon[\mu]$ is of class $\mathcal{C}^\infty$, for any $\mu \in \mathcal{P}(\mathbb{T}^d)$. Then the mapping $A^\epsilon[\mu]$ is the solution of the elliptic linear partial differential equation

$$\Delta A^\epsilon[\mu] = \mathrm{div}(F^\epsilon[\mu])$$

and standard elliptic regularity theory implies that $A^\epsilon[\mu]$ is also of class $\mathcal{C}^\infty$. See Lemma 3.1 below for quantitative bounds (depending on $\epsilon$).

**Proposition 2.2.** *Under Assumption 2.1, for any initial conditions $x_0 = (y_0, z_0) \in \mathbb{T}^d$, the system (9) admits a unique solution, which is defined for all times $t \geqslant 0$.*

The proof of Proposition 2.2 is postponed to Section 3

## 2.3 Main result and discussion

Remark that the free energy can be defined up to an additive constant. Above, $A_\star$ has been normalized so that $\int_{\mathbb{T}^m} e^{-A_\star} dz = 1$, while $A_t$ is such that $\int_{\mathbb{T}^m} A_t dz = 0$. Denote $\bar{A}_\star = A_\star - \int_{\mathbb{T}^m} A_\star(z)dz$. The standard norm on the Sobolev space $W^{1,p}(\mathbb{T}^m)$, for $p \in [2,\infty)$, is denoted by $\|\cdot\|_{W^{1,p}}$.

**Theorem 2.3.** *Under Assumption 2.1, there exists $\epsilon_0 > 0$ and, for all $p \in [1,+\infty)$, there exists $C_p \in [0,+\infty)$ such that, for all $\epsilon \in (0,\epsilon_0]$, there exists a unique probability distribution $\mu_\infty^\epsilon \in \mathcal{P}(\mathbb{T}^d)$ which satisfies*

$$d\mu_\infty^\epsilon(x) = d\mu_\star^{A^\epsilon[\mu_\infty^\epsilon]}(x) \propto e^{A^\epsilon[\mu_\infty^\epsilon](z)} d\mu_\star(y,z).$$

*In addition, one has the error estimate*

$$\|\bar{A}_\star - A^\epsilon[\mu^\epsilon_\infty]\|_{W^{1,p}} \;\leqslant\; C_p\sqrt{\epsilon}\,,$$

*and, for any initial conditions $x_0 = (y_0, z_0) \in \mathbb{T}^d$, almost surely, one has the convergence*

$$\begin{aligned}\|A_t - A^\epsilon[\mu^\epsilon_\infty]\|_{W^{1,p}} &\underset{t\to\infty}{\longrightarrow} 0\\ \mu_t &\underset{t\to\infty}{\longrightarrow} \mu^\epsilon_\infty\,,\end{aligned}$$

*the latter in the sense of weak convergence in the set $\mathcal{P}(\mathbb{T}^d)$.*

The first identity in Theorem 2.3 means that the limit $\mu^\epsilon_\infty$ of $\mu_t$ is the fixed point of the mapping $\mu \mapsto \mu_\star^{A^\epsilon[\mu]}$, see Section 4. Equivalently, the limit $A^\epsilon[\mu^\epsilon_\infty]$ of $A_t$ is the fixed point of the mapping $A \mapsto A^\epsilon[\mu_\star^A]$, where we recall that $d\mu_\star^A(x) = e^{A(z)}d\mu_\star(y,z)$.

The almost sure convergence results of Theorem 2.3 may be loosely rephrased as follows

$$\lim_{\epsilon\to 0}\lim_{t\to\infty} A_t = A_\star\,, \quad \lim_{\epsilon\to 0}\lim_{t\to\infty} \mu_t = \mu_\star^{A_\star},$$

and implies that the empirical distribution $\nu_t = \frac{1}{t}\int_0^t \delta_{\xi(X_s)}ds$ satisfies the approximate asymptotic flat-histogram property

$$\lim_{\epsilon\to 0}\lim_{t\to\infty} \nu_t = dz.$$

We stress that $\mu^\epsilon_\infty$ is not close (when $\epsilon \to 0$) to the multimodal target distribution $\mu_\star$: with the notation above one has $\mu_\star = \mu_\star^0 \neq \mu_\star^{A_\star}$. However, the algorithm gives a way to approximate $\int \varphi d\mu_\star$ by reweighting: using the Cesaro Lemma, it is straightforward to check that one has

$$\lim_{t\to\infty} \frac{\int_0^t \varphi(X_s)e^{-A_s(Z_s)}ds}{\int_0^t e^{-A_s(Z_s)}ds} = \lim_{t\to\infty} \frac{\int_0^t \varphi(X_s)e^{-A^\epsilon[\mu^\epsilon_\infty](Z_s)}ds}{\int_0^t e^{-A^\epsilon[\mu^\epsilon_\infty](Z_s)}ds} = \frac{\int \varphi(y,z)e^{-A^\epsilon[\mu^\epsilon_\infty](z)}d\mu^\epsilon_\infty(y,z)}{\int e^{-A^\epsilon[\mu^\epsilon_\infty](z)}d\mu^\epsilon_\infty(y,z)} = \int \varphi d\mu_\star,$$

for any smooth $\varphi : \mathbb{T}^d \to \mathbb{R}$. Indeed, by the Sobolev embedding $W^{1,p}(\mathbb{T}^m) \subset \mathcal{C}^0(\mathbb{T}^m)$ if $p > m$, $A_t$ converges to $A^\epsilon[\mu^\epsilon_\infty]$ uniformly on $\mathbb{T}^m$.

Up to an error depending only on the width $\epsilon > 0$ of the kernel function $K_\epsilon$, the adaptive algorithm (9) is thus a consistent way to approximately compute $\int \varphi d\mu_\star$, as well as the free energy function $A_\star$. The approximate asymptotic flat-histogram property stated above shows that the sampling in the slow, macroscopic variable $z$ is enhanced, hence the efficiency of the approach. Such results are a mathematical justification for the use of the ABF method based on self-interating dynamics in practical computations.

**Remark 2.4.** *From Theorem 2.3, we expect the following Central Limit Theorem to hold: for all bounded $\varphi$ on $\mathbb{T}^d$,*

$$\sqrt{t}\left(\int \varphi d\mu_t - \int \varphi d\mu^\epsilon_\infty\right) \underset{t\to\infty}{\overset{law}{\longrightarrow}} \mathcal{N}(0, \sigma_\varphi)$$

*where $\sigma_\varphi$ is the asymptotical variance obtained by considering the process with a constant bias $\nabla A^\epsilon[\mu^\epsilon_\infty]$. Nevertheless, the proof of such a result, extending [20, Theorem 4.III.5] at the cost of technical considerations, exceeds the scope of the present article.*

**Remark 2.5.** *The convergence of $A_t$ to $A^\epsilon[\mu_\infty^\epsilon]$ when $t \to \infty$ in fact holds for $\mathcal{C}^k$ norms, for all integers $k$. However, the convergence of $\bar{A}_\star - A^\epsilon[\mu_\infty^\epsilon]$ when $\epsilon \to 0$ can be obtained only in $W^{1,p}$, for all $p \in [2, \infty)$ (hence in $\mathcal{C}^0$ due to a Sobolev embedding, for $p > m$). In fact, higher-order derivatives of $F^\epsilon[\mu]$ (and of $A^\epsilon[\mu]$) are expected to explode when $\epsilon \to 0$.*

The ABF has originally been introduced in [18] in the molecular dynamics community, where it is widely used, see [23, 19, 17]. An example of application in statistics is developed in [16]. Another popular related biasing algorithm is the metadynamics algorithm [26],[4],[25],[8].

From a theoretical point of view, several variants of the ABF algorithm have been considered in various works. In a series of papers [30, 1, 29, 28], Lelièvre and his co-authors considered a process similar to (9) except that $\mu_t$ is replaced by the law of $X_t$. This corresponds to the mean-field limit of a system of $N$ interacting particles as $N$ goes to infinity [24]. The law of $X_t$ then solves a non-linear PDE, and long-time convergence is established through entropy techniques. In practice in fact, the bias $A_t$ is obtained both from interacting particles and from interaction with the past trajectories, so that $\mu_t$ is the empirical distribution of a system of $N$ replicas of the system $(X_t, Y_t)$ that contributes all to the same bias $A_t$.

The case of adaptive bias algorithm with a self-interacting process is addressed in [21] for the ABF algorithm and in [6, 7] for the related adaptive biasing potential (ABP) algorithm. We emphasize on the fact that in these works, $\mu_t$ is replaced by a weighted empirical measure $\bar{\mu}_t$ given, in the spirit of an importance sampling scheme, by

$$\bar{\mu}_t \;=\; \left(\int_0^t e^{-A_s(Z_s)} ds\right)^{-1} \int_0^t \delta_{X_s} e^{-A_s(Z_s)} ds\,.$$

Contrary to $\mu_t$ in Theorem 2.3, this weighted empirical measure converges toward $\mu_\star$. This makes the theoretical study simpler than in the present case. However, in practice, there should be no reason to use this weighting procedure for ABF due to the identity (6). Indeed, provided that $A_t$ converges to some $A_\infty$, in the idealized case where $K_\epsilon$ is a Dirac mass, then (6) implies that necessarily $A_\infty = A_\star$. This is no more true as soon as $\epsilon > 0$ (which is necessary for the well-posedness of the algorithm), and one of the main motivation of the present work was to determine whether the convergence of the natural (non re-weighted) version of ABF, which is the one used in practice, was robust with respect to the regularization step. Our results shows that this is true, provided $\epsilon$ is small enough.

## 2.4 Notation

Let $\mathbb{N} = \{1, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and let $k \in \mathbb{N}_0$ be a nonnegative integer. Let $\mathcal{C}^k(\mathbb{T}^{n_1}, \mathbb{R}^{n_2})$ be the space of functions of class $\mathcal{C}^k$ on $\mathbb{T}^{n_1}$ with values on $\mathbb{R}^{n_2}$. The derivative of order $k$ is denoted by $\nabla^k$. The space $\mathcal{C}^k(\mathbb{T}^{n_1}, \mathbb{R}^{n_2})$ is equipped with the norm $\|\cdot\|_{\mathcal{C}^k}$, defined by

$$\|\phi\|_{\mathcal{C}^k} = \sum_{\ell=0}^{k} \|\nabla^k \phi\|_{\mathcal{C}^0},$$

with $\|\phi\|_{\mathcal{C}^0} = \max_{z\in\mathbb{T}^{n_1}} \|\phi(x)\|$. To simplify, the dimensions $n_1$ and $n_2$ are omitted in the notation for the norm $\|\cdot\|_{\mathcal{C}^k}$.

If $\phi : \mathbb{T}^{n_1} \to \mathbb{R}^{n_2}$ is a Lipschitz continuous function, its Lipschitz constant is denoted by $\mathrm{Lip}(\phi)$.

The space $\mathcal{P}(\mathbb{T}^d)$ of probability distributions on $\mathbb{T}^d$ (equipped with the Borel $\sigma$-field) is equipped with the total variation distance $d_{\mathrm{TV}}$ and with the Wasserstein distance $d_{\mathcal{W}_1}$. Recall that one has the following characterizations:

$$d_{TV}(\mu_1,\mu_2) = \sup_{\psi:\mathbb{T}^d\to\mathbb{R},\|\psi\|_\infty\leqslant 1} \frac{1}{2}\Big|\int \psi d\mu_2 - \int \psi d\mu_1\Big|,$$
$$d_{\mathcal{W}_1}(\mu_1,\mu_2) = \sup_{\psi:\mathbb{T}^d\to\mathbb{R},\mathrm{Lip}(\psi)\leqslant 1} \Big|\int \psi d\mu_2 - \int \psi d\mu_1\Big|$$

where for the total variation distance the supremum is taken over bounded measurable functions $\psi$.

The space $\mathcal{P}(\mathbb{T}^d)$ is also equipped with the following distance, which generates the topology of weak convergence:

$$d_w(\mu_1,\mu_2) = \sum_{n\in\mathbb{N}} \frac{1}{2^n} \frac{\big|\int f_n d\mu_2 - \int f_n d\mu_1\big|}{1+\big|\int f_n d\mu_2 - \int f_n d\mu_1\big|},$$

where the sequence $\mathcal{S} = \{f_n\}_{n\in\mathbb{N}}$ is dense in $\mathcal{C}^0(\mathbb{T}^d,\mathbb{R})$, and, for all $n \in \mathbb{N}$, one has $f_n \in \mathcal{C}^\infty$ and $\|f_n\|_{\mathcal{C}^0} \leqslant 1$.

# 3 Proof of the well-posedness result Proposition 2.2

The objective of this section is to prove Proposition 2.2, which states that the system (9) is well-posed. Some auxiliary estimates are provided, where the upper bounds are allowed to depend on the parameter $\epsilon$. Lemma 3.1 provides estimates for $F^\epsilon[\mu]$ and $A^\epsilon[\mu]$, in $\mathcal{C}^k$, uniformly with respect to $\mu$. Lemma 3.2 provides some Lipschitz continuity estimates with respect to $\mu$, in total variation and Wasserstein distances.

## 3.1 Auxiliary estimates

**Lemma 3.1.** *For all $\epsilon \in (0,1]$ and $k \in \mathbb{N}_0$, there exists $C_{\epsilon,k} \in (0,\infty)$ such that one has*

$$\sup_{\mu\in\mathcal{P}(\mathbb{T}^d)} \Big(\|F^\epsilon[\mu]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{R}^m)} + \|A^\epsilon[\mu]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{R})}\Big) \leqslant C_{\epsilon,k}.$$

*Proof of Lemma 3.1.* Observe that

$$F^\epsilon[\mu] = \frac{F_{\mathrm{aux}}[\mu,\nabla_z V]}{F_{\mathrm{aux}}[\mu,1]},$$

where $F^\epsilon_{\mathrm{aux}}[\mu,\psi] = \iint \psi(y,z)K_\epsilon(z,\cdot)d\mu(y,z)$.

Owing to Assumption 2.1, one has

$$F^\epsilon_{\mathrm{aux}}[\mu,1] \geqslant m_\epsilon \int d\mu = m_\epsilon > 0,$$

for all $\mu \in \mathcal{P}(\mathbb{T}^d)$. In addition, for all $k \in \mathbb{N}_0$, one has

$$\nabla^k F^\epsilon_{\mathrm{aux}}[\mu,\psi] = \iint \psi(y,z)\nabla^k K_\epsilon(z,\cdot)d\mu(y,z),$$

thus, one obtains

$$\|F^\epsilon_{\mathrm{aux}}[\mu,\psi]\|_{\mathcal{C}^k} \leqslant \|\psi\|_{\mathcal{C}^0} M^{(k)}_\epsilon < \infty,$$

owing to Assumption 2.1.

Using the estimate above with $\psi = \nabla_z V$ and $\psi = 1$, it is then straightforward to deduce that

$$\|F^\epsilon[\mu]\|_{\mathcal{C}^k} = \|\frac{F_{\mathrm{aux}}[\mu,\nabla_z V]}{F_{\mathrm{aux}}[\mu,1]}\|_{\mathcal{C}^k} \leqslant C_{\epsilon,k}.$$

This concludes the proof of the estimates for $F^\epsilon[\mu]$. To prove the estimates for $A^\epsilon[\mu]$, observe that $\tilde{A}^\epsilon[\mu]$ solves the Euler-Lagrange equation associated with the minimization problem in (8),

$$\Delta \tilde{A}^\epsilon[\mu] = \mathrm{div}\big(F^\epsilon[\mu]\big).$$

Using the result proved above, and standard elliptic regularity theory and Sobolev embeddings, one obtains the required estimates for $\tilde{A}^\epsilon[\mu]$: for all $\epsilon \in (0,1]$ and $k \in \mathbb{N}_0$, there exists $C_{\epsilon,k} \in (0,\infty)$ such that for all $\mu \in \mathcal{P}(\mathbb{T}^d)$,

$$\|\tilde{A}^\epsilon[\mu]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{T})} \leqslant C_{\epsilon,k}.$$

Since $A^\epsilon[\mu]$ and $\tilde{A}^\epsilon[\mu]$ only differ by an additive constant, it only remains to prove that

$$\|A^\epsilon[\mu]\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{T})} \leqslant C_{\epsilon,0}.$$

This is a straightforward consequence of the estimate $\|\tilde{A}^\epsilon[\mu]\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{T})} \leqslant C_{\epsilon,0}$ and of (8).

This concludes the proof of Lemma 3.1. □

**Lemma 3.2.** *For all $\epsilon \in (0,1]$ and $k \in \mathbb{N}_0$, there exists $L_{\epsilon,k} \in (0,\infty)$ such that, for all $\mu_1,\mu_2 \in \mathcal{P}(\mathbb{T}^d)$, one has*

$$\|F^\epsilon[\mu_2] - F^\epsilon[\mu_1]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{R}^m)} + \|A^\epsilon[\mu_2] - A^\epsilon[\mu_1]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{R})} \leqslant L_{\epsilon,k}\big(d_{\mathrm{TV}}(\mu_1,\mu_2) \wedge d_{\mathcal{W}_1}(\mu_1,\mu_2)\big).$$

*Proof of Lemma 3.2.* First, observe that

$$\begin{aligned} F^\epsilon[\mu_2] - F^\epsilon[\mu_1] = {} & \frac{\iint \nabla_z V(y,z)K_\epsilon(z,\cdot)d(\mu_2-\mu_1)(y,z)}{\iint K_\epsilon(z,\cdot)d\mu_2(y,z)} \\ & - \frac{\iint \nabla_z V(y,z)K_\epsilon(z,\cdot)d\mu_1(y,z)\iint K_\epsilon(z,\cdot)d(\mu_2-\mu_1)(y,z)}{\iint K_\epsilon(z,\cdot)d\mu_1(y,z)\iint K_\epsilon(z,\cdot)d\mu_2(y,z)}. \end{aligned}$$

Using the characterizations of total variation and Wasserstein distances and the regularity properties of $V$ and $K_\epsilon$ (Assumption 2.1), proceeding as in the proof of Lemma 3.1 then yields

$$\|F^\epsilon[\mu_2] - F^\epsilon[\mu_1]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{T}^m)} \leqslant L_{\epsilon,k} d_(\mu_1,\mu_2),$$

for all $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{T}^d)$, with $L_{\epsilon,k} \in (0,\infty)$, with $d = d_{\mathcal{W}_1}$ and $d = d_{TV}$.

It remains to apply the same arguments as in the proof of Lemma 3.1 to obtain

$$\|\tilde{A}^\epsilon[\mu_2] - A^\epsilon[\mu_1]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{T})} + \|A^\epsilon[\mu_2] - A^\epsilon[\mu_1]\|_{\mathcal{C}^k(\mathbb{T}^m,\mathbb{T})} \leqslant L_{\epsilon,k} d(\mu_1,\mu_2),$$

which concludes the proof of Lemma 3.2. □

## 3.2 Well-posedness

Let $T \in (0,\infty)$ be an arbitrary positive real number. Introduce the Banach spaces

$$\mathcal{C}([0,T],\mathbb{T}^d)\ , \quad E = L^2\big(\Omega, \mathcal{C}([0,T],\mathbb{T}^d)\big),$$

equipped with the norms defined by

$$\|x\|_\alpha = \sup_{0\leqslant t\leqslant T} e^{-\alpha t}|x(t)|\ , \quad |\!|\!|X|\!|\!|_\alpha = \Big(\mathbb{E}\big[\|X\|_\alpha^2\big]\Big)^{\frac{1}{2}},$$

depending on the auxiliary parameter $\alpha \in (0,\infty)$. Let $\Phi : E \to E$ be defined as follows: for all $x = \big(y_t, z_t\big)_{t\geqslant 0}$, let $\mu_t^x = \frac{1}{1+t}\big(\mu_0 + \int_0^t \delta_{x_s} ds\big)$ and $A_t^x = A^\epsilon[\mu_t^x]$, for all $t \geqslant 0$. Then $X = \Phi(x)$ is the solution $X = \big(Y(t), Z(t)\big)_{t\geqslant 0}$ of

$$\begin{cases} dY(t) = -\nabla_y V(y_t, z_t)dt + \sqrt{2}dW^{(d-m)}(t), \\ dZ(t) = -\nabla_z V(y_t, z_t)dt + \nabla A_t^x(z_t)dt + \sqrt{2}dW^{(d)}(t), \end{cases}$$

with initial condition $(Y(0), Z(0)) = x_0 \in \mathbb{T}^d$, which is fixed.

If $\alpha$ is sufficiently large, then the mapping $\Phi$ is a contraction, due to Lemma 3.3 stated below.

**Lemma 3.3.** *There exists $C \in (0,\infty)$ such that for all $\alpha \in (0,\infty)$, and for all $x^1, x^2 \in E$,*

$$|\!|\!|\Phi(x_2) - \Phi(x_1)|\!|\!|_\alpha \leqslant \frac{C}{\alpha} |\!|\!|x_2 - x_1|\!|\!|_\alpha.$$

*Proof of Lemma 3.3.* Let $x^1 = (y^1, z^1)$ and $x^2 = (y^2, z^2)$ be two elements of $E$, and set $X^1 = \Phi(x^1)$, $X^2 = \Phi(x^2)$. Then

$$\frac{d}{dt}\big(Y^2(t) - Y^1(t)\big) = \nabla_y V(y_t^1, z_t^1) - \nabla_y V(y_t^2, z_t^2)$$

and

$$\frac{d}{dt}\big(Z^2(t) - Z^1(t)\big) = \nabla_z V(y_t^1, z_t^1) - \nabla_z V(y_t^2, z_t^2) + \nabla A_t^2(z_t^2) - \nabla A_t^1(z_t^1),$$

where $A^i_t = A^\epsilon[\mu^i_t]$ and $\mu^i_t = \frac{1}{1+t}(\mu_0 + \int_0^t \delta_{x^i_s} ds)$.

First, since $V$ is of class $\mathcal{C}^2$, for all $t \geqslant 0$, one has the almost sure estimate

$$\begin{aligned} e^{-\alpha t}|Y^2(t) - Y^1(t)| &\leqslant Ce^{-\alpha t}\int_0^t \big(|y^2_s - y^1_s| + |z^2_s - z^1_s|\big)ds \\ &\leqslant Ce^{-\alpha t}\int_0^t e^{\alpha s} ds \|x^2 - x^1\|_\alpha \\ &\leqslant \frac{C}{\alpha}\|x^2 - x^1\|_\alpha. \end{aligned}$$

Second, similarly one has, for all $t \geqslant 0$,

$$\begin{aligned} e^{-\alpha t}|Z^2(t) - Z^1(t)| &\leqslant \frac{C}{\alpha}\|x^2 - x^1\|_\alpha + e^{-\alpha t}\int_0^t |\nabla A^2_s(z^2_s) - \nabla A^1_s(z^1_s)|ds \\ &\leqslant \frac{C}{\alpha}\|x^2 - x^1\|_\alpha + e^{-\alpha t}\int_0^t |\nabla A^2_s(z^2_s) - \nabla A^2_s(z^1_s)|ds + e^{-\alpha t}\int_0^t |\nabla A^2_s(z^1_s) - \nabla A^1_s(z^1_s)|ds \\ &\leqslant \frac{C}{\alpha}\|x^2 - x^1\|_\alpha + e^{-\alpha t}\int_0^t \|A^2_s - A^1_s\|_{\mathcal{C}^1} ds, \end{aligned}$$

owing to Lemma 3.1. In addition, owing to Lemma 3.2, one has

$$\begin{aligned} \|A^2_s - A^1_s\|_{\mathcal{C}^1} &= \|A^\epsilon[\mu^2_s] - A^\epsilon[\mu^1_s]\|_{\mathcal{C}^1} \\ &\leqslant L_{\epsilon,1} d_{\mathcal{W}_1}(\mu^1_s, \mu^2_s) \leqslant L_{\epsilon,1}\int_0^s |x^2(r) - x^1(r)|dr \\ &\leqslant L_{\epsilon,1}\int_0^s e^{\alpha r} dr \|x^2 - x^1\|_\alpha \\ &\leqslant \frac{L_{\epsilon,1}}{\alpha} e^{\alpha s}\|x^2 - x^1\|_\alpha. \end{aligned}$$

Finally, one obtains the almost sure estimate,

$$\|\Phi(x^2) - \Phi(x^1)\|_\alpha = \sup_{t\geqslant 0} e^{-\alpha t}|X^2(t) - X^1(t)| \leqslant \frac{C}{\alpha}\|x^2 - x^1\|_\alpha,$$

then taking expectation concludes the proof of Lemma 3.3. □

The proof of Proposition 2.2 is then straightforward.

*Proof of Proposition 2.2.* Observe that the following claims are satisfied.

- Owing to Lemma 3.1, for all $x \in E$, one has the almost sure estimate $\sup_{t\geqslant 0} \|\nabla A^x_t\|_{\mathcal{C}^0} \leqslant C_{\epsilon,0}$, and owing to Lemma 3.2, the mapping $t \mapsto A^x_t$ is Lipschitz continuous. Thus the mapping $\Phi$ is well-defined.
- The process $\big(Y(t), Z(t), A_t, \mu_t\big)_{t\geqslant 0}$ solves (9) if and only if $X = (Y, Z)$ is a fixed point of $\Phi$.

- The mapping $\Phi : E \to E$ is a contraction if $\alpha$ is sufficiently large, and admits a unique fixed point $X$, owing to Lemma 3.3.

Since the initial conditions $x_0$ and $\mu_0$, and the time $T \in (0, \infty)$ are arbitrary, these arguments imply that the global well-posedness of (9) and this concludes the proof. □

# 4 The limiting flow

Define the mapping $\Pi^\epsilon : \mu \in \mathcal{P}(\mathbb{T}^d) \mapsto \Pi^\epsilon[\mu] \in \mathcal{P}(\mathbb{T}^d)$, for $\epsilon \in (0, 1]$, as follows:

$$\Pi^\epsilon[\mu] = Z^\epsilon[\mu]^{-1} e^{-V(y,z)+A^\epsilon[\mu](z)} dydz,$$

with $Z^\epsilon[\mu] = \iint e^{-V(y,z)+A^\epsilon[\mu](z)} dydz$. The notation $V^\epsilon_\mu(y, z) = V(y, z) - A^\epsilon[\mu](z)$ is used in the sequel. The probability measure $\Pi^\epsilon[\mu]$ is the unique invariant distribution for the system

$$\begin{cases} dY^A_t = -\nabla_y V(Y^A_t, Z^A_t)dt + \sqrt{2}dW^{(d-m)}_t, \\ dZ^A_t = -\nabla_z V(Y^A_t, Z^A_t)dt + \nabla A(Z^A_t)dt + \sqrt{2}dW^{(m)}_t \end{cases}$$

with $A = A^\epsilon[\mu]$. With notations used above, $\Pi^\epsilon[\mu] = \mu_\star^{A^\epsilon[\mu]}$.

The objectives of this section are twofold. First, one proves that, for every $\pi \in \mathcal{P}(\mathbb{T}^d)$, there exists a unique solution $\big(\Phi^\epsilon(t, \pi)\big)_{t\geqslant 0}$ of the equation

$$\Phi^\epsilon(t, \pi) = e^{-t}\pi + \int_0^t e^{s-t}\Pi^\epsilon\big[\Phi^\epsilon(s, \pi)\big]ds.$$

In addition, $\pi^\epsilon_t = \Phi^\epsilon(t, \pi)$ solves, in a weak sense, the following ordinary differential equation

$$\dot\pi^\epsilon_t = \Pi^\epsilon[\pi^\epsilon_t] - \pi^\epsilon_t \ , \quad \pi^\epsilon_0 = \pi.$$

Second, one relates the properties of the empirical measure $\big(\mu_t\big)_{t\geqslant 0}$ in the regime $t \to \infty$, with the behavior of the limit flow, using the notion of Asymptotic Pseudo-Trajectories.

## 4.1 Well-posedness of the limiting flow

Let $M^\epsilon = \sup_{\mu\in\mathcal{P}(\mathbb{T}^d)} \|A^\epsilon[\mu]\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R})}$, and $M_\star = \|A_\star\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R})}$. Note that $M^\epsilon < \infty$ due to Lemma 3.1. Recall that $L_{0,\epsilon}$ is defined in Lemma 3.2.

**Lemma 4.1.** *Let $L(\epsilon) = 2L_{\epsilon,0}e^{4(M^\epsilon+M_\star)}$. Then for all $\mu^1, \mu^2 \in \mathcal{P}(\mathbb{T}^d)$, one has*

$$d_{TV}\big(\Pi^\epsilon[\mu^1], \Pi^\epsilon[\mu^2]\big) \leqslant L(\epsilon)d_{TV}(\mu^1, \mu^2).$$

*Proof of Lemma 4.1.*

$$\begin{aligned} d_{TV}\big(\Pi^\epsilon[\mu^1],\Pi^\epsilon[\mu^2]\big) &= \iint_{\mathbb{T}^d} e^{-V(y,z)}\Big|\frac{e^{A^\epsilon[\mu^1](z)}}{Z^\epsilon[\mu^1]}-\frac{e^{A^\epsilon[\mu^2](z)}}{Z^\epsilon[\mu^2]}\Big|dydz \\ &= \int_{\mathbb{T}^m} e^{-A_\star(z)}\Big|\frac{e^{A^\epsilon[\mu^1](z)}}{Z^\epsilon[\mu^1]}-\frac{e^{A^\epsilon[\mu^2](z)}}{Z^\epsilon[\mu^2]}\Big|dz \\ &\leqslant \int_{\mathbb{T}^m}\frac{e^{-A_\star(z)}}{Z^\epsilon[\mu^1]}\Big|e^{A^\epsilon[\mu^1](z)}-e^{A^\epsilon[\mu^2](z)}\Big|dz \\ &\quad+\int_{\mathbb{T}^m}\frac{e^{A^\epsilon[\mu^2](z)-A_\star(z)}}{Z^\epsilon[\mu^1]Z^\epsilon[\mu^2]}dz\Big|Z^\epsilon[\mu^1]-Z^\epsilon[\mu^2]\Big|. \end{aligned}$$

Using the lower bound

$$Z^\epsilon[\mu]=\iint_{\mathbb{T}^d} e^{-V(y,z)+A^\epsilon[\mu](z)}dydz=\int_{\mathbb{T}^m} e^{-A_\star(z)+A^\epsilon[\mu](z)}dz\geqslant e^{-M_\star-M^\epsilon},$$

and the upper bound

$$\big|Z^\epsilon[\mu^1]-Z^\epsilon[\mu^2]\big|\leqslant e^{M^\epsilon+M_\star}\int_{\mathbb{T}^m}|A^\epsilon[\mu^1](z)-A^\epsilon[\mu^2](z)|dz,$$

one obtains

$$\begin{aligned} d_{TV}\big(\Pi^\epsilon[\mu^1],\Pi^\epsilon[\mu^2]\big) &\leqslant 2e^{4(M^\epsilon+M_\star)}\int_{\mathbb{T}^m}|A^\epsilon[\mu^1](z)-A^\epsilon[\mu^2](z)|dz \\ &\leqslant 2e^{4(M^\epsilon+M_\star)}\|A^\epsilon[\mu^1]-A^\epsilon[\mu_2]\|_{\mathcal{C}^0} \\ &\leqslant 2L_{\epsilon,0}e^{4(M^\epsilon+M_\star)}d_{\mathrm{TV}}(\mu_1,\mu_2), \end{aligned}$$

where the last inequality follows from Lemma 3.2. This concludes the proof of Lemma 4.1. □

**Proposition 4.2.** *Let $\pi\in\mathcal{P}(\mathbb{T}^d)$. Then there exists a unique solution $\big(\Phi^\epsilon(t,\pi)\big)_{t\geqslant 0}$, with values in $\mathcal{C}\big([0,\infty),\mathcal{P}(\mathbb{T}^d)\big)$ (where $\mathcal{P}(\mathbb{T}^d)$ is equipped with the total variation distance $d_{TV}$), of the equation*

$$\Phi^\epsilon(t,\pi)=e^{-t}\pi+\int_0^t e^{s-t}\Pi^\epsilon[\Phi^\epsilon(s,\pi)]ds.$$

*Proof.* Uniqueness is a straightforward consequence of Lemma 4.1 and of Gronwall Lemma.

Existence is obtained using a Picard iteration argument. Precisely, introduce the mapping $\Psi:\mathcal{C}\big([0,\infty),\mathcal{P}(\mathbb{T}^d)\big)\to\mathcal{C}\big([0,\infty),\mathcal{P}(\mathbb{T}^d)\big)$, be defined by

$$\Psi(\pi)(t)=e^{-t}\pi+\int_0^t e^{s-t}\Pi^\epsilon[\pi_s]ds,$$

for $\pi = \big(\pi_t\big)_{t\geqslant 0}$.

Let $d_\alpha(\pi^1,\pi^2) = \sup_{t\geqslant 0} e^{-\alpha t} d_{TV}(\pi^1_t,\pi^2_t)$, where $\alpha > 0$ is chosen below. Then, using Lemma 4.1, one has

$$d_\alpha\big(\Psi(\pi^1),\Psi(\pi^2)\big) \leqslant \frac{L(\epsilon)}{\alpha} d_\alpha(\pi^1,\pi^2).$$

Choose $\alpha = 2L(\epsilon)$, and define

$$\pi^0 = \big(\pi^0_t = \pi\big)_{t\geqslant 0}\ , \quad \pi^{n+1} = \Psi(\pi^n),\ n\geqslant 0,$$

using the Picard iteration method. Let $T\in(0,\infty)$ be an arbitrary positive real number. Since $\mathcal{C}\big([0,T],\mathcal{P}(\mathbb{T}^d)\big)$ is a complete metric space (equipped with the distance $d_\alpha$), then $\big(\pi^n\big)_{n\in\mathbb{N}}$ converges when $n\to\infty$, and the limit $\pi^\infty$ solves the fixed point equation $\pi^\infty = \Psi(\pi^\infty)$, which proves the existence of a solution, and concludes the proof. □

By construction, the flow $\Phi^\epsilon : \mathbb{R}^+\times\mathcal{P}(\mathbb{T}^d)\to\mathcal{P}(\mathbb{T}^d)$ is continuous, when $\mathcal{P}(\mathbb{T}^d)$ is equipped with the total variation distance $d_{\mathrm{TV}}$. Adapting the proof of [9, Lemma 3.3], one checks that it is also a continuous mapping when $\mathcal{P}(\mathbb{T}^d)$ is equipped with the distance $d_w$.

## 4.2 The asymptotic pseudotrajectory property

Recall that a continuous function $\zeta:\mathbb{R}^+\to\mathcal{P}(\mathbb{T}^d)$ is an asymptotic pseudotrajectory for $\Phi^\epsilon$, if one has

$$\sup_{s\in[0,T]} d_w\big(\zeta(t+s),\Phi^\epsilon(s,\zeta(t))\big) \underset{t\to\infty}{\to} 0,$$

for all $T\in\mathbb{R}^+$. See for instance [5] for details.

The following result is the rigorous formulation of the link between the dynamics of the empirical measures $\mu_t$ in the ABF algorithm, and of the limit flow.

**Theorem 4.3.** *The process $\big(\mu_{e^t}\big)_{t\geqslant 0}$ is almost surely an asymptotic pseudotrajectory for $\Phi^\epsilon$.*

The proof requires auxiliary notations and results. For every $\epsilon>0$ and $\mu\in\mathcal{P}(\mathbb{T}^d)$, let

$$V^\epsilon_\mu(y,z) = V(y,z) - A^\epsilon[\mu](z),$$

and define the infinitesimal generator

$$\mathcal{L}^\epsilon_\mu = \Delta - \nabla V^\epsilon_\mu\cdot\nabla.$$

Introduce the projection operator defined by $K^\epsilon_\mu f = f - \int f d\Pi^\epsilon[\mu]$ and let $\big(P^{\epsilon,\mu}_t\big)_{t\geqslant 0}$ be the semi-group generated by $\mathcal{L}^\epsilon_\mu$ on $L^2(\mathbb{T}^d)$. Finally, let

$$Q^\epsilon_\mu = \int_0^\infty P^{\epsilon,\mu}_t K^\epsilon_\mu dt\,.$$

Then one has the following result.

**Lemma 4.4.** *For every $\epsilon > 0$, there exists $C_\epsilon \in (0, \infty)$, such that*

$$\|Q^\epsilon_\mu f\|_{\mathcal{C}^1} \leqslant C_\epsilon \|f\|_{\mathcal{C}^0}\,, \tag{10}$$

*for all $f \in \mathcal{C}^0(\mathbb{T}^d, \mathbb{R})$ and all $\mu \in \mathcal{P}(\mathbb{T}^d)$. Moreover, $\mathcal{L}^\epsilon_\mu K^\epsilon_\mu = -K^\epsilon_\mu$.*

*Proof.* Remark that, from Lemma 3.1, $V^\varepsilon_\mu \in \mathcal{C}^\infty(\mathbb{T}^d)$, from which it is classical to see that $P^{\epsilon,\mu}_t f \in \mathcal{C}^\infty(\mathbb{T}^d)$ for all $f \in \mathcal{C}^\infty(\mathbb{T}^d)$. In particular, $\mathcal{C}^\infty(\mathbb{T}^d)$ is a core for $\mathcal{L}^{\epsilon,\mu}$, see [3, Section 3.2] and thus it is enough to prove the result for $f \in \mathcal{C}^\infty(\mathbb{T}^d)$.

As a first step, for all $\epsilon \in (0, 1]$ there exists $R_\epsilon > 0$ such that for all $\mu \in \mathcal{P}(\mathbb{T}^d)$, $\Pi^\epsilon[\mu]$ satisfies a log-Sobolev inequality and a Sobolev inequality both with constant $R_\epsilon$, in the sense that for all positive $f \in \mathcal{C}^\infty(\mathbb{T}^d)$,

$$\begin{aligned}\int_{\mathbb{T}^d} f \ln f d\Pi^\epsilon[\mu] - \int_{\mathbb{T}^d} f d\Pi^\epsilon[\mu] \ln \int_{\mathbb{T}^d} f d\Pi^\epsilon[\mu] &\leqslant R_\epsilon \int_{\mathbb{T}^d} \frac{|\nabla f|^2}{f} d\Pi^\epsilon[\mu] \\ \|f\|^2_{L^p(\Pi^\epsilon[\mu])} &\leqslant R_\epsilon \|f\|^2_{H^1(\Pi^\epsilon[\mu])}\,,\end{aligned}$$

where $p = \frac{2d}{d-2}$. Indeed, from Lemma 3.1, the density of $\Pi^\epsilon[\mu]$ with respect to the Lebesgue measure is bounded above and below away from zero uniformly in $\mu \in \mathcal{P}(\mathbb{T}^d)$. The inequalities are then obtained by a perturbative argument from those satisfied by the Lebesgue measure, see [3, Proposition 5.1.6]).

As a second step, these inequalities imply the following estimates: for all $\epsilon \in (0, 1]$ there exists $R'_\epsilon > 0$ such that for all $\mathcal{P}(\mathbb{T}^d)$, $f \in \mathcal{C}^\infty(\mathbb{T}^d)$ and $t \geqslant 0$,

$$\begin{aligned}\|P^{\epsilon,\mu}_t K^\epsilon_\mu f\|_{L^2(\Pi^\varepsilon[\mu])} &\leqslant e^{-R_\epsilon t/2} \|K^\epsilon_\mu f\|_{L^2(\Pi^\varepsilon[\mu])} \\ \|P^{\epsilon,\mu}_t f\|_\infty &\leqslant \frac{R'_\epsilon}{\max(1, t^{d/2})} \|f\|_{L^2(\Pi^\varepsilon[\mu])} \\ \|\nabla P^{\epsilon,\mu}_t f\|_\infty &\leqslant \frac{R'_\epsilon}{\max(1, \sqrt{t})} \|f\|_\infty.\end{aligned}$$

Indeed, the first estimate is a usual consequence of the Poincaré inequality, implied by the log-Sobolev one (see [3, Theorem 4.2.5 and Proposition 5.1.3]). The second one, namely the ultracontractivity of the semi-group, is a consequence of the Sobolev inequality (see [3, Theorem 6.3.1]). The last one can be established thanks to the Bakry-Emery calculus (see [3, Section 1.16] for an introduction), by showing that $\mathcal{L}^\epsilon_\mu$ satisfies a curvature estimate. More precisely, denote

$$\begin{aligned}\Gamma^{\epsilon,\mu}(f, g) &= \frac{1}{2}\left(\mathcal{L}^\epsilon_\mu(fg) - f\mathcal{L}^\epsilon_\mu g - g\mathcal{L}^\epsilon_\mu f\right) \\ \Gamma^{\epsilon,\mu}_2(f) &= \frac{1}{2}\Gamma^{\epsilon,\mu}(f) - \Gamma^{\epsilon,\mu}(f, \mathcal{L}^\epsilon_\mu f),\end{aligned}$$

with $\Gamma^{\epsilon,\mu}(f) := \Gamma^{\epsilon,\mu}(f, f)$. Straightforward computations yield

$$\begin{aligned}\Gamma^{\epsilon,\mu}(f) &= |\nabla f|^2 \\ \Gamma^{\epsilon,\mu}_2(f) &\geqslant -|\nabla^2 V^\epsilon_\mu||\nabla f|^2 \geqslant -c_\epsilon \Gamma^{\epsilon,\mu}(f)\end{aligned}$$

for some $c_\epsilon > 0$ which does not depend on $\mu \in \mathcal{P}(\mathbb{T}^d)$ thanks to Lemma 3.1. According to [3, Theorem 4.7.2], this implies that

$$\Gamma^{\epsilon,\mu}(P_t^{\epsilon,\mu} f) \;\leqslant\; \left(\frac{1-e^{-c'_\epsilon t}}{c'_\epsilon}\right)^{-1} P_f^{\epsilon,\mu} f^2 \;\leqslant\; \left(\frac{1-e^{-c'_\epsilon t}}{c'_\epsilon}\right)^{-1} \|f\|_\infty^2,$$

which concludes the proof of the third estimate.

As a third step, we bound (using that $\|P_t^{\epsilon,\mu} f\|_\infty \leqslant \|f\|_\infty$ for all $t \geqslant 0$)

$$\begin{aligned}
\int_0^\infty \|P_t^{\epsilon,\mu} K_\mu^\epsilon f\|_\infty dt &\leqslant \int_0^1 \|K_\mu^\epsilon f\|_\infty dt + \int_1^\infty \|P_t^{\epsilon,\mu} K_\mu^\epsilon f\|_\infty dt \\
&\leqslant 2\|f\|_\infty + R'_\epsilon \int_1^\infty \|P_{t-1}^{\epsilon,\mu} K_\mu^\epsilon f\|_{L^2(\Pi^\epsilon[\mu])} dt \\
&\leqslant 2\|f\|_\infty + R'_\epsilon \int_0^\infty e^{-R_\epsilon s/2} \|K_\mu^\epsilon f\|_{L^2(\Pi^\epsilon[\mu])} dt \\
&\leqslant \left(2 + \frac{4R'_\epsilon}{R_\epsilon}\right) \|f\|_\infty ,
\end{aligned}$$

and similarly

$$\begin{aligned}
\int_0^\infty \|\nabla P_t^{\epsilon,\mu} K_\mu^\epsilon f\|_\infty dt &\leqslant \int_0^2 \frac{R'_\epsilon}{\max(1,\sqrt{t})} \|K_\mu^\epsilon f\|_\infty dt + R'_\epsilon \int_2^\infty \|P_{t-1}^{\epsilon,\mu} K_\mu^\epsilon f\|_\infty dt \\
&\leqslant 6R'_\epsilon \|f\|_\infty + R'^2_\epsilon \int_0^\infty e^{-R_\epsilon s/2} \|K_\mu^\epsilon f\|_{L^2(\Pi^\epsilon[\mu])} dt \\
&\leqslant \left(6R'_\epsilon + \frac{4R'^2_\epsilon}{R_\epsilon}\right) \|f\|_\infty ,
\end{aligned}$$

from which $Q_\mu^\epsilon f$ is well defined for $f \in \mathcal{C}^\infty(\mathbb{T}^d)$ and satisfies (10) for some $C_\epsilon$. Finally,

$$\begin{aligned}
\mathcal{L}_\mu^\epsilon Q_\mu^\epsilon f &= \int_0^\infty \mathcal{L}_\mu^\epsilon P_t^{\epsilon,\mu} K_\mu^\epsilon f dt \\
&= \int_0^\infty \partial_t \left(P_t^{\epsilon,\mu} K_\mu^\epsilon f\right) dt \;=\; -K_\mu^\epsilon f .
\end{aligned}$$

□

*Proof of Theorem 4.3.* First, note that the claim is equivalent to the following statement (see [9, Proposition 3.5]):

$$\sup_{s\in[0,T]} |\varepsilon_t(s) f| \underset{t\to\infty}{\to} 0,$$

for all $f \in \mathcal{S}$ and $T \in \mathbb{Q}^+$, where

$$\varepsilon_t(s) = \int_{e^t}^{e^{t+s}} \frac{\delta_{X_\tau} - \Pi^\epsilon[\mu_\tau]}{\tau} d\tau.$$

Using a Borel-Cantelli argument, and the fact that $\mathcal{S}$ is a countable set, it is sufficient to establish that there exists $C_\epsilon \in (0,\infty)$, such that

$$\mathbb{E}\big[\sup_{s\in[0,T]} |\varepsilon_t(s)f|^2\big] \leqslant C_\epsilon e^{-t}\|f\|^2_{\mathcal{C}^0},$$

for all $t \geqslant 0$ and $f \in \mathcal{S}$.

Let $f \in \mathcal{S}$ and introduce the function $F : (0,\infty)\times\mathbb{T}^d \to \mathbb{R}$ defined by $F(t,x) = t^{-1}Q^\epsilon_{\mu_t}f$. Then $F$ is of class $\mathcal{C}^{1,2}$ on $(0,\infty)\times\mathbb{T}^d$. Indeed, first, it is straightforward to check that $t \mapsto F^\epsilon[\mu_t] \in \mathcal{C}^k(\mathbb{T}^d,\mathbb{R}^m)$ is of class $\mathcal{C}^1$, for all $k \in \mathbb{N}_0$, since $t \mapsto \mu_t \in \mathcal{P}(\mathbb{T}^d)$ (equipped with the Wasserstein distance) is of class $\mathcal{C}^1$. Second, $A^\epsilon[\mu]$ is solution of the Euler-Lagrange equation $\Delta A^\epsilon[\mu] = \mathrm{div}(F^\epsilon[\mu])$, which establishes that $t \mapsto A^\epsilon[\mu_t] \in \mathcal{C}^k(\mathbb{T}^m,\mathbb{R})$ is also of class $\mathcal{C}^1$. Finally, it remains to apply standard arguments to establish the $\mathcal{C}^1$ regularity of $t \mapsto Q^\epsilon_{\mu_t}f$.

Applying Itô formula yields, for all $t \geqslant 0$ and $s \in [0,T]$, the equality

$$F(e^{t+s},X_{e^{t+s}}) = F(e^t,X_{e^t}) + \int_{e^t}^{e^{t+s}} \big(\partial_\tau + \mathcal{L}^\epsilon_{\mu_\tau}\big)F(\tau,X_\tau)d\tau + \sqrt{2}\int_{e^t}^{e^{t+s}} \langle \nabla F(\tau,X_\tau), dW(\tau)\rangle.$$

Observing that $\mathcal{L}^\epsilon_{\mu_\tau}F(\tau,X_\tau) = \tau^{-1}\mathcal{L}^\epsilon_{\mu_\tau}Q^\epsilon_{\mu_\tau}(X_\tau)f = -\tau^{-1}\big(f(X_\tau) - \int f d\Pi^\epsilon[\mu_\tau]\big)$, one obtains

$$\varepsilon_t(s)f = \varepsilon^1_t(s)f + \varepsilon^2_t(s)f + \varepsilon^3_t(s)f + \varepsilon^4_t(s)f,$$

where

$$\begin{aligned}
\varepsilon^1_t(s)f &= e^{-t}\Big(Q^\epsilon_{\mu_t}f - e^{-s}Q^\epsilon_{\mu_{t+s}}f\Big),\\
\varepsilon^2_t(s)f &= \int_{e^t}^{e^{t+s}} -\tau^{-2}Q^\epsilon_{\mu_t\tau}f(X_\tau)d\tau,\\
\varepsilon^3_t(s)f &= \int_{e^t}^{e^{t+s}} \tau^{-1}\frac{d}{d\tau}Q^\epsilon_{\mu_\tau}f(X_\tau)d\tau,\\
\varepsilon^4_t(s)f &= \sqrt{2}\int_{e^t}^{e^{t+s}} \tau^{-1}\langle\nabla Q^\epsilon_{\mu_\tau}f(X_\tau), dW(\tau)\rangle.
\end{aligned}$$

First, it is straightforward to check that the error terms $\varepsilon^1_t(s)f$ and $\varepsilon^2_t(s)f$ are upper estimated as follows: almost surely,

$$\sup_{0\leqslant s\leqslant T}|\varepsilon^1_t(s)f| + \sup_{0\leqslant s\leqslant T}|\varepsilon^2_t(s)f| \leqslant C_\epsilon e^{-t}\|f\|_\infty.$$

To treat the error term $\varepsilon^3_t(s)f$, it suffices to upper estimate the Lipschitz constant of $t \mapsto Q^\epsilon_{\mu_t}f$. Let $t_1,t_2 \in (0,\infty)$, then one has

$$\begin{aligned}
K^\epsilon_{\mu_{t_1}}f - K^\epsilon_{\mu_{t_2}}f &= \mathcal{L}^\epsilon_{\mu_{t_2}}Q^\epsilon_{\mu_{t_2}}f - \mathcal{L}^\epsilon_{\mu_{t_1}}Q^\epsilon_{\mu_{t_1}}f\\
&= \mathcal{L}^\epsilon_{\mu_{t_1}}\Big(Q^\epsilon_{\mu_{t_2}}f - Q^\epsilon_{\mu_{t_1}}f\Big) + \Big(\mathcal{L}^\epsilon_{\mu_{t_2}} - \mathcal{L}^\epsilon_{\mu_{t_1}}\Big)Q^\epsilon_{\mu_{t_2}}f,
\end{aligned}$$

thus one obtains

$$Q^\epsilon_{\mu_{t_2}} f - Q^\epsilon_{\mu_{t_1}} f = Q^\epsilon_{\mu_{t_1}} \delta^\epsilon_{t_1,t_2} f,$$

where the auxiliary function $\delta^\epsilon_{t_1,t_2} f$ is defined as

$$\delta^\epsilon_{t_1,t_2} f = K^\epsilon_{\mu_{t_1}} f - K^\epsilon_{\mu_{t_2}} f - \Big(\mathcal{L}^\epsilon_{\mu_{t_2}} - \mathcal{L}^\epsilon_{\mu_{t_1}}\Big) Q^\epsilon_{\mu_{t_2}} f,$$

and satisfies the centering condition $\int \delta^\epsilon_{t_1,t_2} f d\Pi^\epsilon[\mu_{t_1}] = \int \mathcal{L}^\epsilon_{\mu_{t_1}} \Big(Q^\epsilon_{\mu_{t_2}} f - Q^\epsilon_{\mu_{t_1}} f\Big) d\Pi^\epsilon[\mu_{t_1}] = 0$.

One has the estimate

$$\|Q^\epsilon_{\mu_{t_2}} f - Q^\epsilon_{\mu_{t_1}} f\|_\infty \leqslant C_\epsilon \|\delta^\epsilon_{t_1,t_2} f\|_\infty.$$

On the one hand, one has

$$\begin{aligned}
\|K^\epsilon_{\mu_{t_1}} f - K^\epsilon_{\mu_{t_2}} f\|_\infty &= \Big| \int f d\Pi^\epsilon[\mu_{t_1}] - \int f d\Pi^\epsilon[\mu_{t_2}] \Big| \\
&\leqslant \|f\|_\infty d_{\mathrm{TV}}(\Pi^\epsilon[\mu_{t_1}], \Pi^\epsilon[\mu_{t_2}]) \\
&\leqslant L(\epsilon) \|f\|_\infty d_{\mathrm{TV}}(\mu_{t_1}, \mu_{t_2}),
\end{aligned}$$

owing to Lemma 4.1.

On the other hand, one has

$$\begin{aligned}
\|\big(\mathcal{L}^\epsilon_{\mu_{t_2}} - \mathcal{L}^\epsilon_{\mu_{t_1}}\big) Q^\epsilon_{\mu_{t_2}} f\|_\infty &= \|\langle \nabla A^\epsilon[\mu_{t_2}] - \nabla A^\epsilon[\mu_{t_1}], \nabla_z Q^\epsilon_{\mu_{t_2}} f \rangle\|_\infty \\
&\leqslant \|A^\epsilon[\mu_{t_2}] - A^\epsilon[\mu_{t_1}]\|_{\mathcal{C}^1} \|Q^\epsilon_{\mu_{t_2}} f\|_{\mathcal{C}^1} \\
&\leqslant L_{1,\epsilon} C_\epsilon \|f\|_\infty d_{\mathrm{TV}}(\mu_{t_1}, \mu_{t_2}).
\end{aligned}$$

Finally, it is straightforward to check that

$$d_{\mathrm{TV}}(\mu_{t_1}, \mu_{t_2}) \leqslant \frac{2|t_2 - t_1|}{t_1 \wedge t_2},$$

using the identity $\dot{\mu}_t = \frac{1}{t+r}(\delta_{X_t} - \mu_t)$.

As a consequence, one obtains

$$\begin{aligned}
\sup_{0 \leqslant s \leqslant T} |\varepsilon^3_t(s) f| &\leqslant \int_{e^t}^{e^{t+T}} \tau^{-1} |\frac{d}{d\tau} Q^\epsilon_{\mu_\tau} f(X_\tau)| d\tau \\
&\leqslant C_\epsilon \int_{e^t}^{e^{t+T}} \tau^{-2} d\tau \|f\|_\infty \\
&\leqslant C_\epsilon e^{-t} \|f\|_\infty.
\end{aligned}$$

It remains to deal with the error term $\varepsilon^4_t(s) f$. Using Doob inequality implies

$$\begin{aligned}
\mathbb{E}\Big[\sup_{0 \leqslant s \leqslant T} |\varepsilon^4_t(s) f|^2\Big] &\leqslant C \int_{e^t}^{e^{t+T}} \tau^{-2} \mathbb{E}\Big[|\nabla Q^\epsilon_{\mu_\tau} f(X_\tau)|^2\Big] d\tau \\
&\leqslant C_\epsilon e^{-t} \|f\|^2_\infty.
\end{aligned}$$

This concludes the proof of the claim,

$$\mathbb{E}\Big[\sup_{s\in[0,T]}|\varepsilon_t(s)f|^2\Big]\leqslant C_\epsilon e^{-t}\|f\|^2_{\mathcal{C}^0},$$

for all $t\geqslant 0$ and $f\in\mathcal{S}$.

Applying a Borel-Cantelli argument then concludes the proof. □

# 5 Proof of Theorem 2.3

The objective of this section is to give a detailed proof of Theorem 2.3. There are two main ingredients. The first one is Proposition 5.3 below, which provides a uniform estimate over $\epsilon>0$ for $A^\epsilon[\mu]$, in the $\mathcal{C}^0$ norm (compare with Lemma 3.1 where the upper bound may depend on $\epsilon$). The second key ingredient is Proposition 5.7, which states a contraction property for the mapping $\Pi^\epsilon$, for an appropriate distance, for sufficiently small $\epsilon$, when restricted to an attracting set identified below (compare with Lemma 4.1 which is valid on the entire state space, but where no upper bound for $L(\epsilon)$ holds).

Combining these two ingredients provides a candidate for the limit as $t\to\infty$, using a standard Picard iteration argument. Using Theorem 4.3 (asymptotic pseudo-trajectory property) then proves the almost sure convergence of $\mu_t$ to this candidate limit.

## 5.1 Uniform estimate

The following PDE estimate is crucial for the analysis.

**Proposition 5.1.** *Let $m\in\mathbb{N}$. For every $p\in[2,\infty)$, there exists $C_p\in(0,\infty)$, such that the following holds: let $F:\mathbb{T}^m\to\mathbb{R}^m$ be a continuous function, then the solution $A$ of the elliptic PDE $\Delta A=\mathrm{div}(F)$, with the condition $\int A(z)dz=0$, satisfies*

$$\|A\|_{W^{1,p}(\mathbb{T}^m,\mathbb{R})}\leqslant C_p\|F\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R}^m)},$$

*and if $p>m$, then*

$$\|A\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R})}\leqslant C_p\|F\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R}^m)}.$$

*Proof.* The proof combines three arguments.

- If $p>m$, then by Sobolev embedding properties, one has $\|A\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R})}\leqslant C_p\|A\|_{W^{1,p}(\mathbb{T}^m,\mathbb{R})}$, with $C_p\in(0,\infty)$.
- By the Poincaré inequality (using the condition $\int A(z)dz=0$, one has $\|A\|_{W^{1,p}(\mathbb{T}^m,\mathbb{R})}\leqslant C_p\|\nabla A\|_{L^p(\mathbb{T}^m,\mathbb{R}^m)}$, with $C_p\in(0,\infty)$, see [2, Theorem 1.13].
- By elliptic regularity theory, one has $\|\nabla A\|_{L^p(\mathbb{T}^m,\mathbb{R}^m)}\leqslant C_p\|F\|_{L^p(\mathbb{T}^m,\mathbb{R}^m)}\leqslant C_p\|F\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R}^m)}$, with $C_p\in(0,\infty)$, see [2, Theorem 15.12].

□

**Remark 5.2.** *If $m = 1$, the proof is straighforward: indeed for all $z \in \mathbb{T}$, one has the identity $A(z) = \int_0^z F(z')dz' - z\int_0^1 F(z')dz'$.*

Using Proposition 5.1, one gets the following crucial estimate, which is uniform for $\epsilon > 0$ (contrary to those given in Lemmas 3.1,3.2 and 4.1 above).

**Proposition 5.3.** *One has the following estimate:*

$$M^0 = \sup_{\epsilon>0} \sup_{\mu\in\mathcal{P}(\mathbb{T}^d)} \|A^\epsilon[\mu]\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R})} < \infty.$$

*Proof.* Using Proposition 5.1 above, it suffices to check that

$$\sup_{\epsilon>0} \sup_{\mu\in\mathcal{P}(\mathbb{T}^d)} \|F^\epsilon[\mu]\|_{\mathcal{C}^0(\mathbb{T}^m,\mathbb{R}^m)} < \infty.$$

That estimate is a straightforward consequence of the definition 7, of the boundedness of $\nabla_z V$, and of the positivity of the kernel function $K_\epsilon$. □

## 5.2 Attracting set

Introduce the following notation: for all $B \in \mathcal{C}(\mathbb{T}^m, \mathbb{R})$, let

$$d\mu_B(y,z) = \mathcal{Z}_B^{-1} e^{-V(y,z)+B(z)} dydz \in \mathcal{P}(\mathbb{T}^d),$$

with $\mathcal{Z}_B = \iint e^{-V(y,z)+B(z)}dydz = \int e^{-A_\star(z)+B(z)}dz$.

First, for probability distribution of the form $\mu_B$, one has the following useful identity for $F^\epsilon[\mu_B]$.

**Lemma 5.4.** *For every $B \in \mathcal{C}(\mathbb{T}^m, \mathbb{R})$, one has*

$$F^\epsilon[\mu_B] = \frac{\int \nabla A_\star(z) K_\epsilon(z,\cdot) e^{B(z)-A_\star(z)}dz}{\int K_\epsilon(z,\cdot)e^{B(z)-A_\star(z)}dz}.$$

*Proof.* This is a straightforward consequence of the two identities below: for all $z \in \mathbb{T}^m$,

$$\int e^{-V(y,z)}dy = e^{-A_\star(z)},$$
$$\int \nabla_z V(y,z) e^{-V(y,z)}dy = -\nabla\left(\int e^{-V(y,z)}dy\right) = e^{-A_\star(z)}\nabla A_\star(z).$$

□

The set of the probability distribution of the type $\mu_B$ is an attractor for the dynamics of the limit flow, more precisely one has the following result.

**Proposition 5.5.** *One has the following result: for all $t \geqslant 0$,*

$$\sup_{\epsilon>0} \sup_{\mu\in\mathcal{P}(\mathbb{T}^d)} \inf_{B\in\mathcal{C}(\mathbb{T}^m,\mathbb{R})} d_{\mathrm{TV}}(\Phi^\epsilon(t,\mu),\mu_B) \leqslant 2e^{-t}.$$

*Proof.* For all $t \geqslant 0$ and $\mu \in \mathcal{P}(\mathbb{T}^d)$, one has

$$\Phi^\epsilon(t,\mu) = e^t\mu + \int_0^t e^{s-t}\Pi^\epsilon[\Phi^\epsilon(s,\mu)]ds = e^{-t}\mu + (1-e^{-t})\Psi^\epsilon(t,\mu),$$

where $\Psi^\epsilon(t,\mu) = \frac{1}{1-e^{-t}}\int_0^t e^{s-t}\Pi^\epsilon[\Phi^\epsilon(s,\mu)]ds = \mu_B$ for some $B \in \mathcal{C}(\mathbb{T}^m,\mathbb{R})$, owing to the definition of $\Pi^\epsilon$.

Then

$$\inf_{B\in\mathcal{C}(\mathbb{T}^m,\mathbb{R})} d_{\mathrm{TV}}(\Phi^\epsilon(t,\mu),\mu_B) \leqslant d_{\mathrm{TV}}(\Phi^\epsilon(t,\mu),\Psi^\epsilon(t,\mu)) \leqslant e^{-t}\|\mu - \Psi^\epsilon(t,\mu)\|_{\mathrm{TV}} \leqslant 2e^{-t}.$$

□

**Lemma 5.6.** *For every $p \in [2,\infty)$, there exists $C_p \in (0,\infty)$, such that for every $\epsilon > 0$, and every $B \in \mathcal{C}(\mathbb{T}^m,\mathbb{R})$, one has*

$$\|A^\epsilon[\mu_B] - \bar{A}_\star\|_{W^{1,p}(\mathbb{T}^m)} \leqslant C_p\sqrt{\epsilon}e^{2\left(\|B\|_{\mathcal{C}^0}+\|A_\star\|_{\mathcal{C}^0}\right)}. \tag{11}$$

Recall that $\bar{A}_\star = A_\star - \int_{\mathbb{T}^m} A_\star dz$.

*Proof.* Using Proposition 5.1, one has the following inequality:

$$\|A^\epsilon[\mu_B] - \bar{A}_\star\|_{W^{1,p}(\mathbb{T}^m,\mathbb{R})} \leqslant C_p\|F^\epsilon[\mu_B] - \nabla A_\star\|_{C^0(\mathbb{T}^m,\mathbb{R}^m)}.$$

Owing to Lemma 5.4 and using the Lipschitz continuity of $A_\star$, for all $z \in \mathbb{T}^m$, one has

$$\begin{aligned}
\left|F^\epsilon[\mu_B](z) - \nabla A_\star(z)\right| &\leqslant \left|\frac{\int\left(\nabla A_\star(z') - \nabla A_\star(z)\right)K_\epsilon(z',z)e^{B(z')-A_\star(z')}dz'}{\int K_\epsilon(z',z)e^{B(z')-A_\star(z')}dz'}\right| \\
&\leqslant C\frac{\int|z-z'|K_\epsilon(z',z)dz'e^{\|B\|_{\mathcal{C}^0}+\|A_\star\|_{\mathcal{C}^0}}}{\int K_\epsilon(z',z)dz'e^{-\|B\|_{\mathcal{C}^0}-\|A_\star\|_{\mathcal{C}^0}}} \\
&\leqslant C\sqrt{\epsilon}e^{2(\|B\|_{\mathcal{C}^0}+\|A_\star\|_{\mathcal{C}^0})},
\end{aligned}$$

owing to Assumption 2.1. This inequality concludes the proof. □

## 5.3 Contraction property on the attracting set

Let $M \in (0,\infty)$. Introduce the set

$$\mathcal{B}_M = \left\{B \in \mathcal{C}^0(\mathbb{T}^m,\mathbb{R}),\ \int B(z)dz = 0,\ \|B\|_{\mathcal{C}^0} \leqslant M\right\}.$$

Owing to Proposition 5.3, if $M \geqslant M^0$, then $A^\epsilon[\mu] \in \mathcal{B}_M$ for every $\mu \in \mathcal{P}(\mathbb{T}^d)$ and $\epsilon > 0$.

Introduce the notation

$$h_B(y,z) = \mathcal{Z}_B^{-1}e^{-V(y,z)+B(z)} \qquad \text{and} \qquad \tilde{\Pi}^\epsilon[h_B] = h_{A^\epsilon[\mu_B]},$$

so that $h_B$ and $\tilde{\Pi}^\epsilon[h_B]$ are the density with respect to the lebesgue measure of, respectively, $\mu_B$ and $\Pi^\epsilon[\mu_B]$.

To state the following result, the notation $\|h\|_2 = \left(\int h(x)^2dx\right)^{\frac{1}{2}}$ is used.

**Proposition 5.7.** *For every $M \in (0,\infty)$, there exists $C_M \in (0,\infty)$, such that for all $\epsilon > 0$ and all $B^1, B^2 \in \mathcal{B}_M$, one has*

$$\|\tilde{\Pi}^\epsilon[h_{B^1}] - \tilde{\Pi}^\epsilon[h_{B^2}]\|_2 \leqslant C_M\sqrt{\epsilon}\|h_{B^1} - h_{B^2}\|_2.$$

*Proof.* Let $B^1, B^2 \in \mathcal{B}_M$. Using Proposition 5.3, one has

$$\|\tilde{\Pi}^\epsilon[h_{B^1}] - \tilde{\Pi}^\epsilon[h_{B^2}]\|_2 = \|h_{A^\epsilon[\mu_{B^1}]} - h_{A^\epsilon[\mu_{B^2}]}\|_2 \leqslant C\|A^\epsilon[\mu_{B^1}] - A^\epsilon[\mu_{B^2}]\|_2.$$

In addition, using the Poincaré inequality and the definition of $A^\epsilon[\mu]$ as the orthogonal projection in $L^2$ of $F^\epsilon[\mu]$, one has

$$\|A^\epsilon[\mu_{B^1}] - A^\epsilon[\mu_{B^2}]\|_2 \leqslant C\|F^\epsilon[\mu_{B^1}] - F^\epsilon[\mu_{B^2}]\|_2.$$

Then, using Lemma 5.4, one obtains, for all $z \in \mathbb{T}^m$,

$$\begin{aligned}
|F^\epsilon[\mu_{B^1}](z) - F^\epsilon[\mu_{B^2}](z)| &= \Big|\frac{\int\big(\nabla A_\star(z') - \nabla A_\star(z)\big)K_\epsilon(z',z)e^{B^1(z')-A_\star(z')}dz'}{\int K_\epsilon(z',z)e^{B^1(z')-A_\star(z')}dz'} \\
&\qquad - \frac{\int\big(\nabla A_\star(z') - \nabla A_\star(z)\big)K_\epsilon(z',z)e^{B^2(z')-A_\star(z')}dz'}{\int K_\epsilon(z',z)e^{B^2(z')-A_\star(z')}dz'}\Big| \\
&\leqslant \Big|\frac{\int\big(\nabla A_\star(z') - \nabla A_\star(z)\big)K_\epsilon(z',z)\Big(e^{B^1(z')} - e^{B^2(z')}\Big)e^{-A_\star(z')}dz'}{\int K_\epsilon(z',z)e^{B^1(z')-A_\star(z')}dz'}\Big| \\
&\quad + \Big|\frac{\int\big(\nabla A_\star(z') - \nabla A_\star(z)\big)K_\epsilon(z',z)e^{B^2(z')-A_\star(z')}dz'\int K_\epsilon(z',z)\Big(e^{B^1(z')} - e^{B^2(z')}\Big)e^{-A_\star(z')}dz'}{\int K_\epsilon(z',z)e^{B^1(z')-A_\star(z')}dz'\int K_\epsilon(z',z)e^{B^2(z')-A_\star(z')}dz'}\Big| \\
&\leqslant Ce^{\|B^1\|_{\mathcal{C}^0(\mathbb{T},\mathbb{R})}}\int|z'-z|K_\epsilon(z',z)|e^{B^1(z')} - e^{B^2(z')}|dz' \\
&\quad + Ce^{\|B^1\|_{\mathcal{C}^0(\mathbb{T},\mathbb{R})}+2\|B^2\|_{\mathcal{C}^0(\mathbb{T},\mathbb{R})}}\int|z'-z|K_\epsilon(z',z)dz'\int K_\epsilon(z',z)|e^{B^1(z')} - e^{B^2(z')}|dz',
\end{aligned}$$

using Lipschitz continuity of $\nabla A_\star$, and the lower bound

$$\int K_\epsilon(z',z)e^{B^i(z')-A_\star(z')}dz' \geqslant e^{-\|B^i\|_{\mathcal{C}^0(\mathbb{T},\mathbb{R})}-\|A_\star\|_{\mathcal{C}^0(\mathbb{T})}}\int K_\epsilon(z',z)dz' = e^{-\|B^i\|_{\mathcal{C}^0(\mathbb{T},\mathbb{R})}-\|A_\star\|_{\mathcal{C}^0(\mathbb{T})}}.$$

One has $\|B^1\|_{\mathcal{C}^0} \leqslant M$ and $\|B^2\|_{\mathcal{C}^0} \leqslant M$. In addition, owing to Assumption 2.1, one has $\int|z'-z|K_\epsilon(z',z)dz' \leqslant C\sqrt{\epsilon}$. As a consequence, using the Jensen inequality (since $\int K_\epsilon(z',z)dz' = \int K_\epsilon(z,z')dz' = 1$ for all $z$), one obtains

$$\begin{aligned}
\|F^\epsilon[\mu_{B^1}] - F^\epsilon[\mu_{B^2}]\|_2 &\leqslant C_M\iint K_\epsilon(z',z)|z'-z|^2|e^{B_1(z')} - e^{B_2}(z')|^2dzdz' \\
&\qquad + C_M\epsilon\iint K_\epsilon(z',z)|e^{B_1(z')} - e^{B_2}(z')|^2dzdz' \\
&\leqslant C_M\epsilon\int|e^{B_1(z')} - e^{B_2(z')}|^2dz'.
\end{aligned}$$

It remains to check that

$$\int |e^{B_1(z')} - e^{B_2}(z')|^2 dz' \leqslant C\|h_{B_1} - h_{B_2}\|_2^2.$$

On the one hand,

$$\begin{aligned}\|h_{B_1} - h_{B_2}\|_2^2 &= \iint e^{-2V(y,z)} \Big| \frac{e^{B_1(z)}}{\int e^{B_1 - A_\star}} - \frac{e^{B_2}(z)}{\int e^{B_2 - A_\star}} \Big|^2 dydz \\ &\geqslant c \int \Big| \frac{e^{B_1(z)}}{\int e^{B_1 - A_\star}} - \frac{e^{B_2}(z)}{\int e^{B_2 - A_\star}} \Big|^2 dz,\end{aligned}$$

with $c > 0$. On the other hand, using Young inequality (with auxiliary parameter $\eta > 0$), one obtains

$$\begin{aligned}\int |e^{B_1(z')} - e^{B_2}(z')|^2 dz' &= \Big| \int e^{B_1 - A_\star} \frac{e^{B_1(z)}}{\int e^{B_1 - A_\star}} - \int e^{B_2 - A_\star} \frac{e^{B_2}(z)}{\int e^{B_2 - A_\star}} \Big|^2 dz \\ &\leqslant 2\eta^2 \int \Big| \frac{e^{B_2}(z)}{\int e^{B_2 - A_\star}} \Big|^2 dz \Big| \int e^{B_1 - A_\star} - \int e^{B_2 - A_\star} \Big|^2 \\ &\quad + \frac{2}{\eta^2} \big( \int e^{B_1 - A_\star} \big)^2 \int \Big| \frac{e^{B_1(z)}}{\int e^{B_1 - A_\star}} - \frac{e^{B_2}(z)}{\int e^{B_2 - A_\star}} \Big|^2 dz \\ &\leqslant 2C_M \eta^2 \int |e^{B_1(z')} - e^{B_2}(z')|^2 dz' + \frac{2C_M}{\eta^2} \|h_{B_1} - h_{B_2}\|_2^2.\end{aligned}$$

Choosing a sufficiently small parameter $\eta$ one finally obtains the claim above.

Gathering the estimates finally concludes the proof of the estimate

$$\|\tilde{\Pi}^\epsilon[h_{B^1}] - \tilde{\Pi}^\epsilon[h_{B^2}]\|_2 \leqslant C_M \sqrt{\epsilon} \|h_{B^1} - h_{B^2}\|_2.$$

□

## 5.4 Proof of the main result

The first part of this section is devoted to the construction of the candidate limits $\mu_\infty^\epsilon$ and $A_\infty^\epsilon = A^\epsilon[\mu_\infty^\epsilon]$, of $\mu_t$ and $A_t$ respectively, for small enough $\epsilon$.

Let $\bar{\epsilon}_0 = 1/(C_{M^{(0)}}^2 + 1)$, where $M = M^{(0)}$ is given by Proposition 5.3 and $C_M$ is given by Proposition 5.7.

Let $\epsilon \in (0, \bar{\epsilon}_0]$, and consider $A_{(0)} \in \mathcal{B}_{M^{(0)}}$. Define $\mu_{(0)} = \mu_{A_{(0)}}$, $h_{(0)} = h_{A_{(0)}}$, and by recursion, for all nonnegative integer $k$, let

$$\mu_{(k+1)} = \Pi^\epsilon[\mu_{(k)}] \ , \quad h_{(k+1)} = \tilde{\Pi}^\epsilon[h_{(k)}],$$

and let $A_{(k)} = A^\epsilon[\mu_{(k)}]$. Then one has $h_{(k)} = h_{A_{(k)}} \in \mathcal{B}_{M^{(0)}}$. We claim that $\big(\mu_{(k)}\big)_{k \geqslant 0}$ is a Cauchy sequence in the space $\mathcal{P}(\mathbb{T}^d)$ equipped with the total variation distance $d_{\mathrm{TV}}$. Indeed,

for all $k, \ell \geqslant 0$, one has

$$\begin{aligned} d_{\mathrm{TV}}(\mu_{(k)}, \mu_{(k+\ell)}) &\leqslant \|h_{(k)} - h_{(k+\ell)}\|_2 \\ &\leqslant \big(C_{M^{(0)}}\sqrt{\epsilon}\big)^k d_2(h^{(0)}, h^{(\ell)}) \\ &\leqslant C\rho^k, \end{aligned}$$

with $\rho \in (0,1)$. As a consequence, there exists $\mu^\epsilon_\infty$ such that $d_{\mathrm{TV}}(\mu_{(k)}, \mu^\epsilon_\infty) \underset{k\to\infty}{\to} 0$. Owing to Lemma 4.1, the mapping $\Pi^\epsilon$ is continuous on $\mathcal{P}(\mathbb{T}^d)$ equipped with $d_{\mathrm{TV}}$, thus $\mu^\epsilon_\infty = \Pi^\epsilon[\mu^\epsilon_\infty]$. This implies that $\mu^\epsilon_\infty = h_{A^\epsilon_\star}(x)dx$ where $A^\epsilon_\star = A^\epsilon[\mu^\epsilon_\infty] \in \mathcal{B}_{M^{(0)}}$.

It is then straightforward to check that $h^\epsilon_\infty = h_{A^\epsilon_\infty}$ is the unique fixed point of the mapping $\tilde{\Pi}^\epsilon$ (uniqueness is a consequence of Proposition 5.7).

We claim that, for any initial condition of the type $\mu_B$, then $\Phi^\epsilon(t, \mu_B) \underset{t\to\infty}{\to} \mu^\epsilon_\infty$, more precisely one has exponential convergence to the fixed point $\mu^\epsilon_\infty$: there exists $c(\epsilon) \in (0, \infty)$ such that, for all $t \geqslant 0$, one has

$$\sup_{B \in \mathcal{B}_{M^{(0)}}} d_{\mathrm{TV}}(\Phi^\epsilon(t, \mu_B), \mu^\epsilon_\infty) \leqslant Ce^{-c(\epsilon)t}. \tag{12}$$

To prove this claim, observe that for all $t \geqslant 0$, the probability distribution $\Phi^\epsilon(t, \mu_B)$ can be written as $\mu_{B_t}$, where $B_t \in \mathcal{C}^0$, see Proposition 5.5, and without loss of generality $\int B_t(z)dz = 0$. In addition, $B_t \in \mathcal{B}_{M^{(1)}}$, for all $t \geqslant 0$, for some $M^{(1)} \in (0, \infty)$ depending only on $M^{(0)}$: indeed, the identity

$$h_{B_t} = e^{-t}h_{B_0} + \int_0^t e^{-(t-s)}\tilde{\Pi}^\epsilon[h_{B_s}]ds$$

implies, using Proposition 5.3, the bounds

$$0 < \inf_{t\geqslant 0} \inf_{x\in\mathbb{T}^d} h_{B_t}(x) \leqslant \sup_{t\geqslant 0} \sup_{x\in\mathbb{T}^d} h_{B_t}(x) < \infty,$$

and $B_t(z)$ is equal (up to an additive constant defined to respect the condition $\int B_t(z)dz = 0$) to $A_\star(z) + \log\big(\int e^{-V(y,z)}dy\big)$.

Let $\epsilon_0 = 1/(C^2_{M^{(1)}} + 1)$, and assume in the sequel that $\epsilon \in (0, \epsilon_0]$. Note that $M^{(1)} \geqslant M^{(0)}$, thus $\epsilon_0 \leqslant \bar{\epsilon}_0$.

Then $A^\epsilon_\infty$ is well-defined, $h^\epsilon_\infty$ is the unique fixed point of $\tilde{\Pi}^\epsilon$, and one obtains

$$\begin{aligned} d_{\mathrm{TV}}(\Phi^\epsilon(t, \mu_B), \mu^\epsilon_\infty) &\leqslant \|h_{B_t} - h^\epsilon_\infty\|_2 \\ &\leqslant e^{-t}\|h_B - h^\epsilon_\infty\|_2 + \int_0^t e^{-(t-s)}\|\tilde{\Pi}^\epsilon[h_{B_s}] - \tilde{\Pi}^\epsilon[h^\epsilon_\infty]\|_2 ds \\ &\leqslant e^{-t}\|h_B - h^\epsilon_\infty\|_2 + C_{M^{(1)}}\sqrt{\epsilon}\|h_{B_s} - h^\epsilon_\infty\|_2 ds, \end{aligned}$$

with $C_{M^{(1)}}\sqrt{\epsilon} < 1$. Applying the Gronwall Lemma, one obtains

$$d_{\mathrm{TV}}(\Phi^\epsilon(t, \mu_B), \mu^\epsilon_\infty) \leqslant \|h_{B_t} - h^\epsilon_\infty\|_2 \leqslant e^{-(1-C_{M^{(1)}}\sqrt{\epsilon})t}\|h_B - h^\epsilon_\infty\|_2,$$

and it is straightforward to check that sup $\{\|h_B - h^\epsilon_\infty\|_2,\ B \in \mathcal{B}_{M^{(0)}}\} < \infty$. This concludes the proof of the claim (12).

We are now in position to prove give the proof of Theorem 2.3. It is sufficient to focus on the question of convergence when $t \to \infty$, indeed the estimate for $\|A^\epsilon_\infty - \nabla A_\star\|_{W^{1,p}}$ is a straightforward consequence of Lemma 5.6, combined with Proposition 5.3, since $A^\epsilon_\infty$ is a fixed point of the mapping $A \mapsto A^\epsilon[\mu_A]$.

The idea of the proof, using concepts and tools developed in [5] may be described as follows. Since almost surely $\big(\mu_t\big)_{t\geqslant 0}$ is an asymptotic pseudo-trajectory for the semi-flow $\Phi^\epsilon$, one has the following property: the limit set $L(\mu)$ is an attractor free set for the semi-flow $\Phi^\epsilon$ in $\mathcal{P}(\mathbb{T}^d)$, in particular it is invariant, *i.e.* for all $t \geqslant 0$ one has $\Phi^\epsilon(t, L(\mu)) = L(\mu)$. Let us check that $L(\mu) = \{\mu^\epsilon_\infty\}$. First, introduce the set $\mathcal{M} = \{\mu_B\}$. Then Proposition 5.5 provides the inclusion $L(\mu) \subset \mathcal{M}$. Indeed, let $\nu \in L(\mu)$ and let $t \geqslant 0$ be arbitrary, then by invariance there exists $\tilde\nu \in L(\mu)$ such that $\nu = \Phi^\epsilon(\tilde\nu)$, thus $d(\nu, \mathcal{M}) = d(\Phi^\epsilon(\tilde\nu), \mathcal{M}) \leqslant 2e^{-t} \underset{t\to\infty}{\to} 0$. Similarly, let $\nu \in L(\mu) \subset \mathcal{M}$, and let $t \geqslant 0$ be arbitrary, then $\nu = \Phi^\epsilon(\tilde\nu)$ for some $\tilde\nu \in \mathcal{M}$. Thus $d(\nu, \mu^\epsilon_\infty) = d(\Phi^\epsilon(t, \tilde\nu), \Phi^\epsilon(t, \mu^\epsilon_\infty)) \leqslant Ce^{-ct} \underset{t\to\infty}{\to} 0$.

Let us now provide a detailed proof using only the results presented above.

*Proof of Theorem 2.3.* Let $T_1, T_2 \in (0, \infty)$ be arbitrary positive real numbers, and $T = T_1 + T_2$. For every $t \geqslant T$, one has

$$d_w\big(\mu_{e^t}, \mu^\epsilon_\infty\big) \leqslant d_w\big(\mu_{e^t}, \Phi^\epsilon(T, \mu_{e^{t-T}}\big) + d_w\big(\Phi^\epsilon(T, \mu_{e^{t-T}}), \mu^\epsilon_\infty\big).$$

Owing to Theorem 4.3, for any fixed $T_1, T_2$, one has, almost surely,

$$d_w\big(\mu_{e^t}, \Phi^\epsilon(T, \mu_{e^{t-T}})\big) \underset{t\to\infty}{\to} 0.$$

Observe that $d_w(\cdot,\cdot) \leqslant C d_{\mathrm{TV}}(\cdot,\cdot)$. In addition, for all $B \in \mathcal{C}(\mathbb{T}, \mathbb{R})$, using Lemma 4.1 and the claim (12) above, one has

$$\begin{aligned} d_{\mathrm{TV}}\big(\Phi^\epsilon(T, \mu_{e^{t-T}}), \mu^\epsilon_\infty\big) &\leqslant d_{\mathrm{TV}}\big(\Phi(T_1, \Phi(T_2, \mu_{e^{t-T}})), \Phi(T_1, \mu_B)\big) + d_{\mathrm{TV}}\big(\Phi(T_1, \mu_B), \mu^\epsilon_\infty\big) \\ &\leqslant e^{L(\epsilon)T_1} d_{\mathrm{TV}}\big(\Phi(T_2, \mu_{e^{t-T}}), \mu_B\big) + Ce^{-c(\epsilon)T_1} \end{aligned}$$

This implies that

$$\begin{aligned} d_{\mathrm{TV}}\big(\Phi^\epsilon(T, \mu_{e^{t-T}}), \mu^\epsilon_\infty\big) &\leqslant e^{L(\epsilon)T_1} \sup_{B\in\mathcal{C}(\mathbb{T},\mathbb{R})} d_{TV}(\Phi(T_2, \mu_{e^{t-T}}), \mu_B) + Ce^{-c(\epsilon)T_1} \\ &\leqslant 2e^{L(\epsilon)T_1} e^{-T_2} + 2e^{-c(\epsilon)T_1}, \end{aligned}$$

owing to Proposition 5.5.

$$\limsup_{t\to\infty}\ d_{\mathrm{TV}}\big(\Phi^\epsilon(T, \mu_{e^{t-T}}), \mu^\epsilon_\star\big) \leqslant 2e^{L(\epsilon)T_1} e^{-T_2} + 2e^{-c(\epsilon)T_1}.$$

Since $T_1$ and $T_2$ are arbitrary, letting first $T_2 \to \infty$, then $T_1 \to \infty$, one has almost surely

$$\limsup_{t\to\infty}\ d_w(\mu_{e^t}, \mu^\epsilon_\star) = 0,$$

which concludes the proof of the weak convergence of $\mu_t$ to $\mu_\infty^\epsilon$.

It remains to check that $A_t = A^\epsilon[\mu_t]$ converges to $A_\infty^\epsilon = A^\epsilon[\mu_\infty^\epsilon]$, in $\mathcal{C}^k$, for all $k \in \mathbb{N}$. This is a consequence of the regularity properties of $K_\epsilon$ and of $V$, which proves that $\mu \in (\mathcal{P}(\mathbb{T}^d), d_w) \mapsto F^\epsilon[\mu] \in \mathcal{C}^k$ is continuous for all $k \in \mathbb{N}$.

Using Sobolev embedding properties, as in the proof of Lemma 3.1, then concludes the proof. □

# Acknowledgments

The authors would like to thank Pierre-Damien Thizy for pointing out the relevant reference [2], for the PDE estimate presented in Proposition 5.1. This work has been partially supported by the Project EFI ANR-17-CE40-0030 of the French National Research Agency and by the SNF grant 200021 - 175728.

# References

[1] Houssam Alrachid and Tony Lelièvre. Long-time convergence of an adaptive biasing force method: variance reduction by Helmholtz projection. *SMAI J. Comput. Math.*, 1:55–82, 2015.

[2] Luigi Ambrosio. Lecture notes on elliptic partial differential equations. *Unpublished lecture notes. Scuola Normale Superiore di Pisa*, 30, 2015.

[3] 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.

[4] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

[5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999.

[6] Michel Benaïm and Charles-Édouard Bréhier. Convergence of adaptive biasing potential methods for diffusions. *C. R. Math. Acad. Sci. Paris*, 354(8):842–846, 2016.

[7] Michel Benaïm and Charles-Edouard Bréhier. Convergence analysis of adaptive biasing potential methods for diffusion processes. *Commun. Math. Sci.*, 17(1):81–130, 2019.

[8] Michel Benaïm, Ioana Ciotir, and Carl-Erik Gauthier. Self-repelling diffusions via an infinite dimensional approach. *Stoch. Partial Differ. Equ. Anal. Comput.*, 3(4):506–530, 2015.

[9] Michel Benaïm, Michel Ledoux, and Olivier Raimond. Self-interacting diffusions. *Probab. Theory Related Fields*, 122(1):1–41, 2002.

[10] Michel Benaïm and Olivier Raimond. Self-interacting diffusions. II. Convergence in law. *Ann. Inst. H. Poincaré Probab. Statist.*, 39(6):1043–1055, 2003.

[11] Michel Benaïm and Olivier Raimond. Self-interacting diffusions. III. Symmetric interactions. *Ann. Probab.*, 33(5):1717–1759, 2005.

[12] Michel Benaïm and Olivier Raimond. Self-interacting diffusions IV: Rate of convergence. *Electron. J. Probab.*, 16:no. 66, 1815–1843, 2011.

[13] Andreas Bittracher, Ralf Banisch, and Christof Schütte. Data-driven computation of molecular reaction coordinates. *The Journal of Chemical Physics*, 149(15):154103, 2018.

[14] Anton Bovier and Frank den Hollander. *Metastability*, volume 351 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2015. A potential-theoretic approach.

[15] Simon Brandt, Florian Sittel, Matthias Ernst, and Gerhard Stock. Machine learning of biomolecular reaction coordinates. *The Journal of Physical Chemistry Letters*, 9(9):2144–2150, 2018. PMID: 29630378.

[16] Nicolas Chopin, Tony Lelièvre, and Gabriel Stoltz. Free energy methods for Bayesian inference: efficient exploration of univariate Gaussian mixture posteriors. *Stat. Comput.*, 22(4):897–916, 2012.

[17] Jeffrey Comer, James C Gumbart, Jérôme Hénin, Tony Lelieèvre, Andrew Pohorille, and Christophe Chipot. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *The Journal of Physical Chemistry B*, 119(3):1129–1151, 2014.

[18] Eric Darve and Andrew Pohorille. Calculating free energies using average force. *The Journal of Chemical Physics*, 115(20):9169–9183, 2001.

[19] Bradley M Dickson. Survey of adaptive biasing potentials: comparisons and outlook. *Current Opinion in Structural Biology*, 43:63–67, 2017.

[20] Marie Duflo. *Algorithmes stochastiques*, volume 23 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996.

[21] Virginie Ehrlacher, Tony Lelièvre, and Pierre Monmarché. Increasing the number of reaction coordinates in adaptive biasing algorithms via tensor approximation. *to appear*.

[22] Haohao Fu, Xueguang Shao, Christophe Chipot, and Wensheng Cai. Extended adaptive biasing force algorithm. an on-the-fly implementation for accurate free-energy calculations. *Journal of Chemical Theory and Computation*, 12(8):3506–3513, 2016. PMID: 27398726.

[23] Jérôme Hénin and Christophe Chipot. Overcoming free energy barriers using unconstrained molecular dynamics simulations. *The Journal of chemical physics*, 121(7):2904–2914, 2004.

[24] Benjamin Jourdain, Tony Lelièvre, and Raphaël Roux. Existence, uniqueness and convergence of a particle approximation for the adaptive biasing force process. *M2AN Math. Model. Numer. Anal.*, 44(5):831–865, 2010.

[25] Benjamin Jourdain, Tony Lelièvre, and Pierre-André Zitt. Convergence of metadynamics: discussion of the adiabatic hypothesis. *arXiv preprint arXiv:1904.08667*, 2019.

[26] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, 2002.

[27] Tony Lelièvre. Two mathematical tools to analyze metastable stochastic processes. In *Numerical mathematics and advanced applications 2011*, pages 791–810. Springer, Heidelberg, 2013.

[28] Tony Lelièvre and Kimiya Minoukadeh. Long-time convergence of an adaptive biasing force method: the bi-channel case. *Arch. Ration. Mech. Anal.*, 202(1):1–34, 2011.

[29] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Computation of free energy profiles with parallel adaptive dynamics. *The Journal of chemical physics*, 126(13):134111, 2007.

[30] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Long-time convergence of an adaptive biasing force method. *Nonlinearity*, 21(6):1155–1181, 2008.

[31] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. *Free energy computations: A mathematical perspective*. Imperial College Press, London, 2010.

[32] Tony Lelièvre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.*, 25:681–880, 2016.