Free energy computations by minimization of Kullback-Leibler divergence: an efficient adaptive biasing potential method for sparse representations

I. Bilionis

Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

P.S. Koutsourelakis *

School of Civil and Environmental Engineering & Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, USA

Abstract

The present paper proposes an adaptive biasing potential for the computation of free energy landscapes. It is motivated by statistical learning arguments and unifies the tasks of biasing the molecular dynamics to escape free energy wells and estimating the free energy function, under the same objective. It offers rigorous convergence diagnostics even though history dependent, non-Markovian dynamics are employed. It makes use of a greedy optimization scheme in order to obtain sparse representations of the free energy function which can be particularly useful in multidimensional cases. It employs embarrassingly parallelizable sampling schemes that are based on adaptive Sequential Monte Carlo and can be readily coupled with legacy molecular dynamics simulators. The sequential nature of the learning and sampling scheme

enables the efficient calculation of free energy functions parametrized by the temperature. The characteristics and capabilities of the proposed method are demonstrated in three numerical examples.

1 Introduction

Free energy is a central concept in thermodynamics and in the study of several systems in biology, chemistry and physics [8]. It represents a rigorous way to coarse-grain systems consisting of very large numbers of atomistic degrees of freedom, to probe states not accessible experimentally, to characterize global changes as well as investigate relative stabilities. In most applications, a brute-force computation based on sampling the atomistic positions is impractical or infeasible as the free energy barriers to overcome are so large that the system remains trapped in metastable free energy sets [40,42,8,56].

Equilibrium techniques for computing free energy surfaces such as Thermodynamic Integration [29] or Adaptive Integration [49,21] require the simulation of very long atomistic trajectories in order to achieve equilibrium and lack convergence diagnostics. Techniques based on non-equilibrium path sampling [26,27,23,25] lack adaptivity and require the user to specify a particular

* Corresponding Author. Tel: 607-254-5441

  Email addresses: `ib227@cornell.edu` (I. Bilionis), `pk285@cornell.edu` (P.S. Koutsourelakis).

path on the reaction coordinate space connecting two energetically important free energy regions, which can be non-trivial a task [20]. Furthermore, sampling along these paths correctly might necessitate advanced and quite involved techniques [10]. More recently proposed adaptive biasing potential [2,55,32,1,18] and adaptive biasing force [11,12,41,53,24] techniques are capable of dynamically utilizing information obtained from the atomistic trajectories to bias the current dynamics in order to facilitate the escape from metastable sets [35]. They are able to automatically discover important regions of the reaction coordinate space. Since they rely on history-dependent, non-Markovian dynamics, it is not a priori clear, and in which sense, the system reaches a stationary state, although some work has been done along theses lines in [4] for Langevin-type systems and [39,35].

We propose an adaptive biasing potential technique where the two tasks of biasing the dynamics and estimating the free energy landscape are unified under the same objective of minimizing the Kullback-Leibler divergence between appropriately selected distributions on the extended space that includes atomic coordinates and the collective variables [37,38]. This framework provides a natural way for selecting the basis functions used in the approximation of the free energy and obtaining sparse representations which is critical when multi-dimensional collective variables are used. It allows the analyst to utilize and correct any prior information on the free energy landscape and provides an efficient manner of obtaining good estimates at various temperatures. The scheme proposed is embarrassingly parallelizable and relies on adaptive Sequential Monte Carlo procedures which enable efficient sampling from the high-dimensional and potentially multi-modal distributions of interest.

2 Methodology - A statistical learning approach for adaptively calculating free energies

For clarity of the presentation, we will first introduce our method for the so-called alchemical case and generalize it later for the reaction coordinate case. Consider a molecular system with generalized coordinates $\boldsymbol{q} \in \mathcal{M} \subset \mathbb{R}^N$ following a Boltzmann-like distribution which in turn depends on some parameters $\boldsymbol{z} \in \mathcal{D} \subset \mathbb{R}^d$

$$p(\boldsymbol{q}|\boldsymbol{z}) \propto \exp\left(-\beta V(\boldsymbol{q};\boldsymbol{z})\right) \tag{1}$$

where $V(\boldsymbol{q};\boldsymbol{z})$ is the potential energy of the system and β plays the role of inverse temperature. The free energy $A(\boldsymbol{z})$ is defined, up to an additive constant, by:

$$A(\boldsymbol{z}) = -\beta^{-1} \int \exp\left(-\beta V(\boldsymbol{q};\boldsymbol{z})\right) d\boldsymbol{q} \tag{2}$$

Our goal is to compute the function $A(\mathbf{z})$ over the whole domain $\mathcal{D}$.

Let $\hat{A}(\mathbf{z};\boldsymbol{\theta})$ be an estimate of $A(\mathbf{z})$ parametrized by $\theta \in \boldsymbol{\Theta} \subset \mathbb{R}^K$. We adopt a statistical perspective of learning $A(\mathbf{z})$ from simulation data. A popular approach to carrying out regression tasks and functional approximations relies on kernel models [28]. Kernel regression models have proven successful in high-dimensional scenario where d is in the order of 10 or 100 [51,52]. The unknown function is selected from a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ induced by a semi-positive definite kernel $K(\cdot,\cdot)$. We adopt representations with respect to a kernel function $K(\cdot,\cdot)$ [28]:

$$\hat{A}(\boldsymbol{z};\boldsymbol{\theta}) = \sum_{j=1}^{K} \theta_j K_j(\boldsymbol{z},\boldsymbol{z}_j), \quad \boldsymbol{z} \in \mathcal{D} \tag{3}$$

where $\boldsymbol{z}_j$ are points in $\mathcal{D}$ which are selected as described in the sequence,

In order to fix the additive constant, we select a point $\mathbf{z}_0 \in \mathcal{D}$ such that: $\hat{A}(\mathbf{z}; \boldsymbol{\theta}) = 0.$ [1]

In relevant literature different types of kernel functions have been used such as thin plate splines, multiquadrics, or Gaussians. While all these functions can be employed in the framework presented, we focus our discussion here on Gaussian kernels which also have an intuitive parametrization with regards to the *scale of variability* of $\hat{A}$ as quantified by the bandwidth parameters $\boldsymbol{\tau}_j = \{\tau_{j,l}\}_{l=1}^d$ in each dimension:

$$K_j(\boldsymbol{z}) = K(\boldsymbol{z}, \boldsymbol{z}_j; \boldsymbol{\tau}_j) = \exp\{-\sum_{l=1}^d \tau_{j,l}(z_l - z_{j,l})^2\} \tag{4}$$

Gaussian kernels in the context of free energy approximations have also been used in [32,37,18].

We define a joint probability distribution on the generalized coordinates $\boldsymbol{q}$ and the parameters $\mathbf{z}$ as follows:

$$p(\mathbf{q}, \boldsymbol{z} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} 1_{\mathcal{D}}(\boldsymbol{z}) e^{-\beta \left(V(\mathbf{q}, \boldsymbol{z}) - \hat{A}(\boldsymbol{z}; \boldsymbol{\theta})\right)} \tag{5}$$

where $1_{\mathcal{D}}(\boldsymbol{z})$ is the indicator function on $\mathcal{D}$ and $Z(\boldsymbol{\theta})$ is the normalization constant, i.e.:

$$Z(\boldsymbol{\theta}) = \int 1_{\mathcal{D}}(\boldsymbol{z}) e^{-\beta \left(V(\mathbf{q}, \boldsymbol{z}) - \hat{A}(\boldsymbol{z}; \boldsymbol{\theta})\right)} d\boldsymbol{z} d\boldsymbol{\theta} \tag{6}$$

It is noted that the first-order partial derivatives of the log of the normalization function give rise to the expectation of the respective kernel:

$$\frac{\partial \log Z}{\partial \theta_j} = \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial Z}{\partial \theta_j} = \beta E_{p(\boldsymbol{z}|\boldsymbol{\theta})} \left[K_j(\boldsymbol{z})\right] \tag{7}$$

---

[1] This is always possible by changing the kernels in Equation 3 to $K'_j(\mathbf{z}, \boldsymbol{z}_j) = K_j(\mathbf{z}) - K_j(\mathbf{z}_0, \boldsymbol{z}_j)$.

whereas the second-order derivatives, produce the covariance between the kernels:

$$\frac{\partial^2 \log Z}{\partial \theta_j \partial \theta_l} = -\frac{1}{Z^2(\boldsymbol{\theta})} \frac{\partial Z}{\partial \theta_j} \frac{\partial Z}{\partial \theta_l} + \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial^2 Z}{\partial \theta_j \partial \theta_l}$$

$$= \beta^2 E_{p(\boldsymbol{z}|\boldsymbol{\theta})} \left[(K_j(\boldsymbol{z}) - E_{p(\boldsymbol{z}|\boldsymbol{\theta})}[K_j(\boldsymbol{z})])(K_l(\boldsymbol{z}) - E_{p(\boldsymbol{z}|\boldsymbol{\theta})}[K_l(\boldsymbol{z})]) \right] \quad (8)$$

$$= \beta^2 Cov_{p(\boldsymbol{z}|\boldsymbol{\theta})} [K_j, K_l]$$

The expectations in the two equations above involve the *unknown* marginal density with respect to the parameters $\boldsymbol{z} \in \mathcal{D}$:

$$p(\boldsymbol{z} \mid \boldsymbol{\theta}) = \int p(\mathbf{q}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, d\mathbf{q}$$

$$= \frac{1}{Z(\boldsymbol{\theta})} 1_{\mathcal{D}}(\boldsymbol{z}) e^{-\beta \left(A(\boldsymbol{z}) - \hat{A}(\boldsymbol{z};\boldsymbol{\theta}) \right)} \quad (9)$$

which depends on the unknown free energy $A(\boldsymbol{z})$ of Equation (2).

The key property of $p(\boldsymbol{z} \mid \boldsymbol{\theta})$ is that it reduces to the uniform distribution for $\mathbf{z} \in \mathcal{D}$ if and only if the free energy estimate is exact i.e. $\hat{A}(\boldsymbol{z};\boldsymbol{\theta}) = A(\boldsymbol{z}), \; \boldsymbol{z} \in \mathcal{D}$.

If $\pi(\boldsymbol{z}) = 1_{\mathcal{D}}(\boldsymbol{z}) \frac{1}{|\mathcal{D}|}$ is the uniform density on $\mathcal{D}$ (whose volume is denoted by $| \mathcal{D} |$), then a natural strategy to estimate $A(\boldsymbol{z})$ is by minimizing a distance metric between $\pi(\mathbf{z})$ and $p(\mathbf{z} \mid \boldsymbol{\theta})$ in Equation (9) above. To that end, we propose employing the Kullback-Leibler (KL) divergence $\mathrm{KL}(\pi(\mathbf{z}) \parallel p(\mathbf{z} \mid \boldsymbol{\theta}))$ [50]:

$$\mathrm{KL}(\pi \parallel p) = \int \pi(\boldsymbol{z}) \log \frac{\pi(\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{\theta})} \, d\boldsymbol{z}$$

$$= \int \pi(\boldsymbol{z}) \log \pi(\boldsymbol{z}) d\mathbf{z} - \int \pi(\boldsymbol{z}) \log p(\boldsymbol{z} \mid \boldsymbol{\theta}) \, d\mathbf{z} \quad (10)$$

$$= -\log | \mathcal{D} | - \int \pi(\boldsymbol{z}) \log p(\boldsymbol{z} \mid \boldsymbol{\theta}) \, d\mathbf{z} \geq 0$$

The latter is not a metric in the mathematical sense, but it is frequently used as a measure of the distance between two probability distributions. It is always non-negative and becomes zero if and only if $\pi(\mathbf{z}) \equiv p(\mathbf{z} \mid \boldsymbol{\theta})$ or equivalently $\hat{A}(\mathbf{z}; \boldsymbol{\theta}) = A(\mathbf{z})$, $\mathbf{z} \in \mathcal{D}$ [2]. The aforementioned formulation offers a clear strategy for estimating the free energy by minimizing the following form with respect to $\boldsymbol{\theta}$:

$$I(\boldsymbol{\theta}) = -\int \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z} \tag{11}$$

Since the KL-divergence is always non-negative, the formulation above provides a lower bound on the objective function $I(\boldsymbol{\theta})$:

$$I(\boldsymbol{\theta}) \geq \log \mid \mathcal{D} \mid \tag{12}$$

which can be readily calculated and be used to monitor convergence as well as the quality of the approximation obtained.

Even though $I(\boldsymbol{\theta})$ depends on the unknown free energy $A(\mathbf{z})$ (from Equation (9)):

$$
\begin{aligned}
I(\boldsymbol{\theta}) &= -\int \pi(\mathbf{z}) \log p(\mathbf{z} \mid \boldsymbol{\theta}) d\mathbf{z} \\
&= \beta \int \pi(\mathbf{z}) \left(A(\mathbf{z}) - \hat{A}(\mathbf{z}; \boldsymbol{\theta}) \right) d\mathbf{z} + \log Z(\boldsymbol{\theta})
\end{aligned}
\tag{13}
$$

its partial derivatives $\boldsymbol{J}(\boldsymbol{\theta}) = \frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ do not , i.e. from Equation (7):

$$
\begin{aligned}
J_j(\boldsymbol{\theta}) &= \frac{\partial I(\boldsymbol{\theta})}{\partial \theta_j} \\
&= -\beta E_{\pi(\mathbf{z})} \left[\frac{\partial \hat{A}}{\partial \theta_j} \right] + \frac{\partial \log Z}{\partial \theta_j} \\
&= -\beta \left(E_{\pi(\mathbf{z})} \left[K_j(\mathbf{z}) \right] - E_{p(\mathbf{z}\mid\boldsymbol{\theta})} \left[K_j(\mathbf{z}) \right] \right)
\end{aligned}
\tag{14}
$$

---

[2] Of interest are free-energy *differences* and therefore perturbations of $A(\mathbf{z})$ or $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$ by a constant are ignored

where $E_{\pi(\mathbf{z})}[\cdot]$ implies an expectation with regards to $\pi(\mathbf{z})$.

It is important to note that according to Equation (8), the Hessian of the objective function $\boldsymbol{H}(\boldsymbol{\theta}) = \frac{\partial^2 I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ is proportional to the covariance between the kernels i.e.:

$$\frac{\partial^2 I}{\partial \theta_j \partial \theta_l} = \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_l}$$

$$= \beta^2 Cov_{p(\mathbf{z}|\boldsymbol{\theta})}[K_j, K_l] \tag{15}$$

Hence the objective function is convex with respect to $\boldsymbol{\theta}$ and there is a unique minimum.

Furthermore, the approximation of the free energy $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$, biases the potential of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ (Equation (5)) and allows the system to overcome free energy barriers [39]. As in [18], no binning is needed and the bias potential is non-local, providing information about the free energy landscape not only at the states visited but in their neighborhood as well. In contrast to other adaptive schemes, the proposed formulation connects the problems of estimating the free energy landscape and steering the atomistic dynamics beyond metastable wells, under a unified umbrella, and provides a clear convergence criterion [35].

From an algorithmic point of view, the proposed strategy poses two problems. The first involves the selection of the kernels to be used in the expansion of Equation (3). This is critical to the sparseness of the representation obtained, particularly for multidimensional $\mathbf{z}$. To that end, a greedy selection strategy is discussed in section 2.2 which progressively adds kernels (i.e. increases the cardinality K of the expansion in Equation (3)) as needed. The second problem involves the optimization of the objective function $I(\boldsymbol{\theta})$ which depends on the unknown free energy $A(\mathbf{z})$ and the intractable partition function $Z(\boldsymbol{\theta})$ (Equation (13)). An obvious approach is by gradient descent which is discussed

in detail in section 2.1. This requires the computation of expectations with respect to $p(z \mid \theta)$ (Equation (14)). The intractability of $p(z \mid \theta)$ necessitates the use of Monte Carlo sampling which must be carried out in the expanded space with respect to the joint density $p(q, z \mid \theta)$ (Equation (5)). This should nevertheless be able to capture multiple modes in the high-dimensional state space consisting of atomic degrees of freedom q and parameters z. To that end we propose performing this step by using non-equilibrium path sampling techniques based on *adaptive Sequential Monte Carlo* schemes discussed in section 2.3. Similar schemes for creating system replicas in parallel have been employed in [41,35]. We discuss a novel adaptive version that retains previous advantages while providing accurate estimates at reduced computational effort. These estimates can be readily updated as θ changes after each optimization step.

A discussion of each of the aforementioned algorithmic modules is contained in the ensuing sub-sections. The steps of the scheme proposed can be found in Algorithm 2.

2.1   *Optimization with noisy gradients*

We propose employing a gradient descent scheme in order to determine θ, although more involved procedures such as Improved Iterative Scaling [3,16], noisy conjugate gradients [46] can be also be employed. Second-order (quasi-)Newton techniques are also possible although the unavoidable Monte Carlo noise in the computation of the Hessian (i.e. the covariance in Equation (15)) can destroy its positive definiteness.

Let $\boldsymbol{\theta}_K$ denote the vector of kernel amplitudes (Equation (3)) when K such kernels are used. Let also $\boldsymbol{\theta}_{K,m}$ denote the estimate of $\boldsymbol{\theta}_K$ after m iterations of the gradient descent algorithm. Then at the $(m+1)-$iteration, the following update equation could be used:

$$\boldsymbol{\theta}_{K,m+1} = \boldsymbol{\theta}_{K,m} - \lambda \boldsymbol{J}(\boldsymbol{\theta}_{K,m}) \tag{16}$$

where λ is the learning rate.

Since only a noisy Monte Carlo estimate of the gradients $\boldsymbol{J}(\boldsymbol{\theta})$ (Equation (14)) is available, it is anticipated that the noise could impede convergence. For that purpose we propose employing a stochastic approximation variant of the Robbins & Monro scheme [43,7]. Rather than increasing the simulation size in order to reduce the variance, we compute a weighted average of the gradient's estimates at the current and previous iterations. By employing a decreasing sequence of weights, information from the earlier iterations gets discarded gradually and more emphasis is placed on the recent iterations. As it is shown in [17], this method converges with a fixed sample size. In particular, if $\hat{\boldsymbol{J}}(\boldsymbol{\theta}_{K,m})$ denotes the Monte Carlo estimate of the gradient (the details of this estimator are discussed in section 2.3) at the mth iteration, then we calculate:

$$\tilde{\boldsymbol{J}}_m = (1 - \eta_m)\tilde{\boldsymbol{J}}_{m-1} + \eta_m \hat{\boldsymbol{J}}(\boldsymbol{\theta}_{K,m}) \tag{17}$$

and, rather than Equation (16), we update $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{\theta}_{K,m+1} = \boldsymbol{\theta}_{K,m} - \lambda \tilde{\boldsymbol{J}}_m \tag{18}$$

where the sequence of weights $\{\eta_m\}$ is such that $\sum_{m=1}^{\infty} \eta_m = \infty$ and $\sum_{m=1}^{\infty} \eta_m^2 < \infty$ [3].

---

[3] A family of such sequences that was used in this work is $\eta_m = \eta m^{-p}$ with

2.2 Kernel selection - Sparse representation of free energy landscape

A critical objective in the proposed framework relates to the *sparseness* of the free energy approximation i.e. the cardinality K of the expansion in Equation (3). This is important in at least two ways. Firstly, because sparser representations can more clearly expose salient features of the free energy landscape, and as a consequence, of the atomistic ensemble considered. Secondly, because they reduce the number of parameters $\boldsymbol{\theta}$ with respect to which the optimization problem needs to be solved (section 2.1). Given a vocabulary of potentially overcomplete basis functions and a prescribed K, the problem amounts to identifying those kernels (Equation (3)) that best approximate the true free energy surface i.e. minimize the KL divergence for $\boldsymbol{z} \in \mathcal{D}$ (Equation (10)). This obviously implies an excessive computational effort since the aforementioned optimization problem would need to be solved for all possible $K-$sized combinations of basis functions.

For that purpose, we propose a hierarchical scheme that proceeds by adding a single kernel at each step. Similar greedy optimization procedures have been successfully applied in maximum entropy problems [16]. Without loss of generality, one can consider a vocabulary of functions that consists of the isotropic Gaussian kernels discussed in Equation (4). Each of these is parametrized by the location $\mathbf{z}_j$ of the kernel and its bandwidth $\boldsymbol{\tau}_j$. Given K such kernels, the corresponding parameters $\boldsymbol{\theta}_K = \{\theta_j\}_{j=1}^{K}$ (Equation (3)) that minimize $I(\boldsymbol{\theta})$ in Equation (11) and samples from the density $p(\mathbf{z} \mid \boldsymbol{\theta}_K)$ (Equation (9) or Equation (5)), we propose selecting the $(K+1)^{th}$ kernel by choosing

---

$1/2 < p \leq 1$.

$\left(\mathbf{z}_{K+1}, \boldsymbol{\tau}_{K+1} = \{\tau_{K+1,l}\}_{l=1}^d \right)$ that maximize:

$$(\mathbf{z}_{K+1}, \boldsymbol{\tau}_{K+1}) = arg \max_{(\mathbf{z}_{K+1}, \tau_{K+1})} \left| E_{\pi(\mathbf{z})} \left[K(\mathbf{z}, \mathbf{z}_{K+1}; \tau_{K+1}) \right] - E_{p(\mathbf{z}|\boldsymbol{\theta}_K)} \left[K(\mathbf{z}, \mathbf{z}_{K+1}; \tau_{K+1}) \right] \right|$$

(19)

Based on Equation (14), this suggests augmenting our expansion with the kernel that locally maximizes the gradient of $I(\boldsymbol{\theta})$. Intuitively this means that we incorporate the kernel function whose expected value with respect to the target, uniform distribution is worst approximated by the current density $p(\mathbf{z} \mid \boldsymbol{\theta}_K)$. It is obviously a suboptimal strategy, that is necessitated by reasons of computational cost. The maximization of the objective in Equation (19) can be readily carried out given samples from $p(\mathbf{z} \mid \boldsymbol{\theta}_K)$. The same formulation can be applied to any type of kernel or overcomplete basis employed (e.g. wavelets). The proposed strategy promotes sparseness and computational efficiency while offering a progressive resolution of the free energy landscape that naturally involves kernels with larger bandwidths (smaller $\boldsymbol{\tau}$) in the first steps, and successive unveiling of the finer details which can be captured by kernels of smaller bandwidths (i.e larger $\boldsymbol{\tau}$).

Furthermore, it offers a rigorous metric for monitoring convergence. In particular if $\hat{A}_K(\mathbf{z}; \boldsymbol{\theta}_K)$ and $\hat{A}_{K+1}(\mathbf{z}; \boldsymbol{\theta}_{K+1})$ denote the free-energy approximations obtained at successive steps using K and $K+1$ kernels respectively, and p_K and p_{K+1} the corresponding densities in Equation (5), then the improvement in terms of Kullback-Leibler divergence (Equation (10)), denoted by Δ_{K+1} can be assessed with:

$$0 \le \Delta_{K+1} = \text{KL}(\pi \parallel p_K) - \text{KL}(\pi \parallel p_{K+1})$$

(20)

$$= -\beta E_\pi \left[\hat{A}_{K+1}(\mathbf{z}; \boldsymbol{\theta}_{K+1}) - \hat{A}_K(\mathbf{z}; \boldsymbol{\theta}_K) \right] + \log \frac{Z(\boldsymbol{\theta}_{K+1})}{Z(\boldsymbol{\theta}_K)}$$

The expectation with respect to the uniform can in general be calculated analytically whereas the ratio of normalizing constants $\log \frac{Z(\boldsymbol{\theta}_{K+1})}{Z(\boldsymbol{\theta}_K)}$ (Equation (5)) is a direct output of the Sequential Monte Carlo sampling that is used to sample from the augmented densities and is discussed in the next section.

2.3 Adaptive Sequential Monte Carlo

The learning scheme proposed relies on efficient computations of the gradients appearing in Equation (14). These depend on expectations with respect to $p(\mathbf{z} \mid \boldsymbol{\theta})$ (Equation (9)) which is not available analytically since the actual free energy $A(\mathbf{z})$ is unknown. We resort to a Monte Carlo scheme that draws samples from the joint density $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ in Equation (5) which involves the atomic degrees of freedom $\mathbf{q}$. A brute-force approach would generally be inefficient as simulating atomistic trajectories suffers from well-known difficulties such as the high-dimensionality of $\mathbf{q}$, the disparity in scales between $\mathbf{q}$ and $\mathbf{z}$ and the presence of several metastable wells [22]. Furthermore, since $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$, new samples would have to drawn every time $\boldsymbol{\theta}$ changes after each iteration of the optimization routine.

For these reasons, we propose a parallelizable strategy that relies on Sequential Monte Carlo samplers (SMC, [15,13]). These offer a general statistical perspective that unifies a range of pertinent schemes that have been proposed in the context of non-equilibrium path sampling following Jarzynski's work [26,48]. We propose novel extensions that allow the algorithm to automatically adapt to the difficulties of the target density. They retain the ability to interact seamlessly with legacy, molecular dynamics simulators.

The proposed SMC schemes offer a flexible framework for sampling from a *sequence of unormalized probability distributions* and are therefore highly suited for the dynamic setting of the problem at hand where the target density $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ changes with $\boldsymbol{\theta}$. For a given $\boldsymbol{\theta}$, they approximate $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$ with a set of n random samples (or *particles*) $\{\mathbf{q}^{(i)}, \mathbf{z}^{(i)}\}_{i=1}^{n}$, which are updated using a combination of *importance sampling*, *resampling* and MCMC-based *rejuvenation* mechanisms [14]. Each of these particles is associated with an *importance weight $w^{(i)}$* which is proportional to $p(\mathbf{q}^{(i)}, \mathbf{z}^{(i)} \mid \boldsymbol{\theta})$. The weights are updated sequentially along with the particle locations in order to provide a particulate approximation:

$$p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \approx \sum_{i=1}^{n} W^{(i)} \, \delta_{\mathbf{q}^{(i)}}(\mathbf{q}) \delta_{\mathbf{z}^{(i)}}(\mathbf{z}) \tag{21}$$

where $W^{(i)} = w^{(i)} / \sum_{k=1}^{N} w^{(k)}$ are the normalized weights and $\delta_{\boldsymbol{x}^{(i)}}(.)$ is the Dirac function centered at $\boldsymbol{x}^{(i)}$. These particles and weights can be used to estimate expectations of any $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta})$-integrable function [13,9]. In particular for Equation (14):

$$\sum_{i=1}^{n} W^{(i)} K_j(\mathbf{z}) \to \int K_j(\mathbf{z}) \, p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \, d\mathbf{q} d\mathbf{z} = E_{p(\mathbf{z}|\boldsymbol{\theta})}\left[K_j(\mathbf{z})\right] \quad \text{(almost surely)}$$

$$\tag{22}$$

The proposed SMC algorithms will be used iteratively, after each step of the gradient descent algorithm. Given two successive estimates $\boldsymbol{\theta}_{K,m}$ and $\boldsymbol{\theta}_{K,m+1}$ (Equation (18)) and a particulate approximation of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m})$, the goal is to obtain new samples from $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m+1})$ (Algorithm 2) and compute the new expectations in Equation (14) based on Equation (22). The quality of the Monte Carlo estimates in Equation (22) depends on the proximity of the distributions $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m})$ and $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m+1})$. We propose building a path of intermediate, unormalized distributions that will bridge this gap based on

Equation (5) [4]:

$$\pi_\gamma(\mathbf{q}, \mathbf{z}) \propto p(\mathbf{q}, \mathbf{z} \mid (1-\gamma)\boldsymbol{\theta}_{K,m} + \gamma\boldsymbol{\theta}_{K,m+1})$$

$$= \exp\left\{-\beta\left(V(\mathbf{q}, \mathbf{z}) - \hat{A}(\mathbf{z}; \boldsymbol{\theta}_\gamma)\right)\right\}, \quad \gamma \in [0, 1] \tag{23}$$

where:

$$\boldsymbol{\theta}_\gamma = (1-\gamma)\boldsymbol{\theta}_{K,m} + \gamma\boldsymbol{\theta}_{K,m+1} \tag{24}$$

Clearly for $\gamma = 0$ one recovers $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m})$ and for $\gamma = 1$, $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m+1})$. The role of these auxiliary distributions is to provide a smooth transition path where importance sampling can be efficiently applied. Naturally, the more intermediate distributions are considered along this path, the higher the accuracy of the final estimates, but also the higher the computational cost. On the other hand too few intermediate distributions π_γ can adversely affect the overall accuracy of the approximation.

To that end we propose an *adaptive* SMC scheme that automatically determines the number of intermediate distributions needed [14,31]. In this process we are guided by the Effective Sample Size (ESS, [36]). In particular, let S be the total number of intermediate distributions (which is unknown a priori) and γ_s, $s = 1, 2, \ldots, S$ the associated reciprocal temperatures such that $0 = \gamma_1 < \gamma_2 < \ldots < \gamma_S = 1$, which are also unknown a priori. Let also $\{(\mathbf{q}_s^{(i)}, \mathbf{z}_s^{(i)}), W_s^{(i)}\}_{i=1}^N$ denote the particulate approximation of π_{γ_s} defined as in Equation (23) for $\gamma = \gamma_s$. The Effective Sample Size of these particles is then defined as $\text{ESS}_s = 1/\sum_{i=1}^N (W_s^{(i)})^2$ and provides a measure of the population variance. One extreme, i.e. when $\text{ESS}_s = 1$, arises when a single particle has a unit normalized weight whereas the rest have zero weights and as a result

[4] subscripts K and m indicating the number of kernels and optimization iterations respectively have been dropped

provide no information. The other extreme, i.e. $\text{ESS}_s = N$, arises when all the particles are equally informative and have equal weights $W_s^{(i)} = 1/N$.

If the next bridging distribution $\pi_{\gamma_{s+1}}$ is very similar to π_{γ_s} (ie. $\gamma_{s+1} \approx \gamma_s$), then ESS_{s+1} should not be that much different from ESS_s. On the other hand if that difference is pronounced then ESS_{s+1} could drop dramatically. Hence in determining the next auxiliary distribution, we define an acceptable reduction in the ESS, i.e. $\text{ESS}_{s+1} \geq \zeta \ \text{ESS}_s$ (where $\zeta < 1$) and prescribe γ_{s+1} (Equation (23)) accordingly.

The proposed adaptive SMC algorithm is summarized in Algorithm 1. It should be noted that unlike MCMC schemes, the particle perturbations in the *Rejuvenation* step do not require that the $P_s(.,.)$ is *ergodic* [15]. It suffices that it is a π_{γ_s}-invariant kernel, which readily allows adaptively changing its parameters in order to achieve better mixing rates. In the examples presented a Metropolized Gibbs scheme was used to sample $\boldsymbol{q}$ and $\boldsymbol{z}$ separately by employing a Metropolis-Adjusted Langevin Algorithm (MALA) for each set of coordinates [44]. In particular given $\left(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)} \right)$ these consist of:

- Updating $\boldsymbol{q}_{s-1}^{(i)} \rightarrow \boldsymbol{q}_s^{(i)}$:

$$
\begin{aligned}
\boldsymbol{q}_s^{(i)} - \boldsymbol{q}_{s-1}^{(i)} &= \tfrac{\Delta t_q}{2} \nabla_{\mathbf{q}} \pi_{\gamma_s}(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_q} \boldsymbol{r}_q \\
&= -\tfrac{\beta \Delta t_q}{2} \nabla_{\mathbf{q}} V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_q} \boldsymbol{r}_q
\end{aligned}
\tag{26}
$$

- Updating $\boldsymbol{z}_{s-1}^{(i)} \rightarrow \boldsymbol{z}_s^{(i)}$:

$$
\begin{aligned}
\boldsymbol{z}_s^{(i)} - \boldsymbol{z}_{s-1}^{(i)} &= \tfrac{\Delta t_z}{2} \nabla_{\mathbf{q}} \pi_{\gamma_s}(\boldsymbol{q}_s^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) + \sqrt{\Delta t_z} \boldsymbol{r}_z \\
&= -\tfrac{\beta \Delta t_z}{2} \left(\nabla_{\mathbf{z}} V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \nabla_{\mathbf{z}} \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_s}) \right) + \sqrt{\Delta t_q} \boldsymbol{r}_q
\end{aligned}
\tag{27}
$$

16

Algorithm 1 Adaptive SMC

Require: $s = 1$ and $\gamma_1 = 0$ and a population $\{(\mathbf{q}_1^{(i)}, \mathbf{z}_1^{(i)}),\ w_1^{(i)}\}_{i=1}^N$ which

approximate $\pi_{\gamma_1} \equiv p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m})$ in Equation (23)

Ensure: The final population $\{(\boldsymbol{\theta}_s^{(i)}, \boldsymbol{d}_s^{(i)}), w_s^{(i)}\}_{i=1}^N$ provides a particulate ap-

proximation of π_{γ_s} in the sense of Equations (21), (22).

  while $\gamma_s < 1$ **do**

    $s \leftarrow s + 1$

    {**Reweighting-Importance Sampling**}

    Let

$$
\begin{aligned}
w_s^{(i)}(\gamma_s) &= w_{s-1}^{(i)} \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)}, \mathbf{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)})} \\
&= w_{s-1}^{(i)} \frac{\exp\left\{-\beta(V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_s}))\right\}}{\exp\left\{-\beta(V(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) - \hat{A}(\boldsymbol{z}_{s-1}^{(i)}; \boldsymbol{\theta}_{\gamma_{s-1}}))\right\}} \qquad (25) \\
&= w_{s-1}^{(i)} \exp\left\{-\beta(\hat{A}(\boldsymbol{z}_{s-1}^{(i)}); (\gamma_s - \gamma_{s-1})(\boldsymbol{\theta}_{K,m+1} - \boldsymbol{\theta}_{K,m}))\right\}
\end{aligned}
$$

    be the *updated* weights as a function of γ_s. Determine $\gamma_s \in (\gamma_{s-1}, 1]$ so

    that $\mathrm{ESS}_s = \zeta\, \mathrm{ESS}_{s-1}$.

    {**Resampling**}

    if $\mathrm{ESS}_s \le \mathrm{ESS}_{min}$ **then**

      Resample

    end if

    {**Rejuvenation**}

    Use an MCMC kernel $P_s\left((\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}), (\boldsymbol{q}_s^{(i)}, \boldsymbol{z}_s^{(i)})\right)$ that leaves π_{γ_s} invariant

    to perturb each particle $(\boldsymbol{q}_{s-1}^{(i)}, \boldsymbol{z}_{s-1}^{(i)}) \to (\boldsymbol{q}_s^{(i)}, \boldsymbol{z}_s^{(i)})$.

  end while

---

where $\boldsymbol{r}_q$ and $\boldsymbol{r}_z$ are i.i.d standard Gaussian vectors. A Metropolis accept/reject

step with respect to the target invariant density $\pi_{\gamma_s}(.)$ was performed after

each update. Two different time steps were used Δt_q and Δt_z for the $\mathbf{q}$ and

z coordinates respectively. Their values were adjusted after each iteration s so as to retain an average acceptance ratio (over all particles n) between 50% and 80% [45]. The adaptivity afforded by the proposed scheme relies on the fact that ergodicity is not required from the rejuvenation step. As a result several MALA time steps can be performed in Equations (26) and (27) (at additional computational expense) or other molecular dynamics samplers can be employed which could potentially exhibit better mixing or fit more closely to the physics of the problem at hand [6].

Finally we note that the estimates of the ratio of normalization constants Z_s/Z_{s-1} between two successive unormalized densities $\pi_{\gamma_{s-1}}$ and π_{γ_s} can be obtained by averaging the unormalized updated weights in Equation (25) as a direct consequence of the importance sampling identity:

$$
\begin{aligned}
\frac{Z_s}{Z_{s-1}} &= \frac{\int \pi_{\gamma_s}(\mathbf{q},\mathbf{z})\ d\mathbf{q}d\mathbf{z}}{\int \pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})\ d\mathbf{q}d\mathbf{z}} \\[2mm]
&= \int \frac{\pi_{\gamma_s}(\mathbf{q},\mathbf{z})}{\pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})} \frac{\pi_{\gamma_{s-1}}(\mathbf{q},\mathbf{z})}{Z_{s-1}}\ d\mathbf{q}d\mathbf{z} \\[2mm]
&\approx \sum_{i=1}^n W_{s-1}^{(i)} \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)},\mathbf{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\mathbf{q}_{s-1}^{(i)},\mathbf{z}_{s-1}^{(i)})}
\end{aligned}
\tag{28}
$$

These estimators can be telescopically multiplied ([15,30]) in order to compute the ratio of normalization constants between any pair of distributions as required in Equation (20).

Given the preceding discussion in sections 2.1, 2.2 and 2.3, we summarize below the basic steps in the proposed free energy computation scheme: In the inner loop and for fixed K, gradient descent (subsection 2.1) is performed which makes use of the adaptive SMC scheme (subsection 2.3) in order to compute the expectations in the gradient. In the outer loop, the cardinality of the expansion K is increased by adding one kernel (i.e. $K \leftarrow K + 1$) based

on Equation (19). This is terminated when the KL gain (Equation (20)) does not exceed a prescribed tolerance.

Algorithm 2 Calculation of the free energy at a given temperature.

Require: $K = 0$, $\boldsymbol{\theta}_0 \equiv 0$ and a particulate approximation of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_0)$

(Equation (5)) at the desired temperature β.

  while true do

    Calculate Δ_K based on Equation (20).

    if $\Delta_K \leq$ tol **then**

      Break the loop.

    else

      Add the optimal $(K+1)^{th}$ kernel based on Equation (19) and set $K \leftarrow$ $K + 1$

      repeat

        Estimate gradient at $\boldsymbol{\theta}_{K,m}$ and calculate update $\boldsymbol{\theta}_{K,m+1}$ based on Equation (18)

        Use adaptive SMC (section 2.3) to construct particulate approxima-tion of $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m+1})$ from $p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}_{K,m})$.

      until Convergence criteria are met.

    end if

  end while

2.4 Obtaining the free energy landscape for various temperatures.

The methodology described in the previous sections is suitable for calculating the free energy as a function of $\mathbf{z}$ at a given temperature. However, one is often interested in the free energy landscape as a function of the temperature also. In order to achieve this goal we make use of the following two facts. Firstly,

the free energy landscape at higher temperatures is flatter and secondly that nearby temperatures have similar free energies landscapes. Based on these, we propose a natural extension to the sequential sampling framework of subsection 2.3 that can efficiently produce estimates of the free energy at various temperatures. The idea is to start from a higher temperature, compute the free energy as described before, then gradually move towards lower temperatures using the free energy of the previous temperature as an initial guess for the new one. In particular given the free energy estimate $\hat{A}_{\beta_1}(\boldsymbol{z}; \boldsymbol{\theta}(\beta_1))$ and the particulate approximation of the joint density $p_{\beta_1}(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}(\beta_1))$ at a temperature $1/\beta_1$, we propose employing the aforementioned adaptive SMC in order to obtain a particulate approximation of the following joint density at $\beta_2 > \beta_1$ (i.e. for lower temperature):

$$p_{\beta_2}(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}(\beta_1)) \propto \exp\left\{-\beta_2\left(V(\mathbf{q}, \mathbf{z}) - \hat{A}_{\beta_1}(\boldsymbol{z}; \boldsymbol{\theta}(\beta_1))\right)\right\} \qquad (29)$$

The iterations enumerated in Algorithm 2 can then be carried out in the same fashion by updating the existing $\boldsymbol{\theta}$ as well as adding new kernels if the convergence criteria are not satisfied.

The critical step involves building a sequence of distributions from $p_{\beta_1}(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}(\beta_1))$ to $p_{\beta_2}(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}(\beta_1))$ in Equation (29). For this purpose and similarly to a simulated annealing schedule we employ:

$$\pi_\gamma(\mathbf{q}, \mathbf{z}) \propto \exp\left\{-((1-\gamma)\beta_1 + \gamma\beta_2)\left(V(\mathbf{q}, \mathbf{z}) - \hat{A}_{\beta_1}(\boldsymbol{z}; \boldsymbol{\theta}(\beta_1))\right)\right\} \qquad (30)$$

The steps in Algorithm 1 should be adjusted to the aforementioned sequence of bridging distributions with the most striking difference in the *Reweighing*

step where the updated weights in Equation (25) should now be given by:

$$w_s^{(i)}(\gamma_s) = w_{s-1}^{(i)} \frac{\pi_{\gamma_s}(\mathbf{q}_{s-1}^{(i)}, \mathbf{z}_{s-1}^{(i)})}{\pi_{\gamma_{s-1}}(\mathbf{q}_{s-1}^{(i)}, \mathbf{z}_{s-1}^{(i)})}$$

$$= w_{s-1}^{(i)} \exp\left\{-(\gamma_s - \gamma_{s-1})(\beta_2 - \beta_1)\left(V(\mathbf{q}_{s-1}^{(i)}, \mathbf{z}_{s-1}^{(i)}) - \hat{A}_{\beta_1}(\mathbf{z}; \boldsymbol{\theta}(\beta_1))\right)\right\}$$

$$(31)$$

We demonstrate the efficacy of such an approach in the last example of section 3. It is finally noted that at the beginning of iterations at each new temperature, kernels with very small weights θ_j were removed if $\frac{\theta_j}{\max_i \theta_i} \leq 0.01$.

2.5 The reaction coordinate case.

The proposed method was described for the alchemical case. However, it is straightforwardly generalized to cover also the general reaction coordinate case. Let $\boldsymbol{\xi} : \mathcal{M} \to \mathcal{D}$ be a function of the system coordinates. This function is called a reaction coordinate [33]. It is evident that $\boldsymbol{q}$ can be viewed in a probabilistic framework as a random variable and so:

$$\mathbf{z} = \boldsymbol{\xi}(\boldsymbol{q}) \tag{32}$$

is also a random variable. The probability distribution of $\mathbf{z}$ can be found by integrating out the $\boldsymbol{q}$:

$$p(\mathbf{z} \mid \beta) = \int p(\boldsymbol{q})\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q} \propto \int \exp\left(-\beta V(\boldsymbol{q})\right)\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q} \tag{33}$$

The free energy $A(\mathbf{z})$ with respect to the reaction coordinate $\boldsymbol{\xi}(\boldsymbol{q})$ is defined to be the *effective potential* of $\mathbf{z} = \boldsymbol{\xi}(\boldsymbol{q})$

$$p(\mathbf{z}) \propto \exp\left(-\beta A(\mathbf{z})\right) \tag{34}$$

21

Combining these two equations we see that:

$$A(\mathbf{z}) = -\beta^{-1} \log \int \exp\left(-\beta V(\boldsymbol{q})\right) \delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z}) d\boldsymbol{q} \qquad (35)$$

If $\hat{A}(\mathbf{z}; \boldsymbol{\theta})$ is an estimate of $A(\mathbf{z})$, we define a new probability distribution over $\boldsymbol{q}$ by:

$$p(\boldsymbol{q}|\boldsymbol{\theta}) \propto 1_{\mathcal{D}}(\boldsymbol{\xi}(\boldsymbol{q})) \exp\left(-\beta(V(\boldsymbol{q}) - \hat{A}(\xi(\boldsymbol{q}); \boldsymbol{\theta}))\right) \qquad (36)$$

It is straight forward to see that under this new distribution for $\boldsymbol{q}$, the pdf of $\mathbf{z}$ becomes:

$$p(\mathbf{z}|\boldsymbol{\theta}) = \int p(\boldsymbol{q}|\boldsymbol{\theta})\delta(\boldsymbol{\xi}(\boldsymbol{q}) - \mathbf{z})d\boldsymbol{q} \propto 1_{\mathcal{D}}(\mathbf{z}) \exp\left(-\beta(A(\mathbf{z}) - \hat{A}(\mathbf{z}; \boldsymbol{\theta}))\right) \qquad (37)$$

This coincides with the expression in Equation (9) and therefore the ensuing derivations hold identically. From a practical point of view, sampling need only performed in the $\mathbf{q}$ space and therefore the adaptive SMC schemes are employed to obtain particulate approximations of the density in Equation (36). The only difference appears in the MCMC-based Rejuvenation step where the MALA sampler is employed only with regards to $\mathbf{q}$. In particular the update of Equation (26) now becomes:

$$\boldsymbol{q}_s^{(i)} - \boldsymbol{q}_{s-1}^{(i)} = \frac{\Delta t_q}{2} \nabla_{\mathbf{q}} \pi_{\gamma_s}(\boldsymbol{q}_{s-1}^{(i)}) + \sqrt{\Delta t_q} \boldsymbol{r}_q$$

$$= -\frac{\beta \Delta t_q}{2}\left(\nabla_{\mathbf{q}} V(\boldsymbol{q}_{s-1}^{(i)}) - \frac{\partial \hat{A}}{\partial \mathbf{z}} \nabla_{\mathbf{q}} \boldsymbol{\xi}(\mathbf{q})\right) + \sqrt{\Delta t_q} \boldsymbol{r}_q \qquad (38)$$

It is noted that, in contrast to some ABF methods which require second-order derivatives of $\boldsymbol{\xi}$ [24], the proposed technique only needs first-order derivatives. Finally, we point out that the ability of the proposed approach to provide efficiently estimates of parametrized free energy surfaces (as in section 2.4 with respect to the temperature β), can also be exploited in the reaction

coordinate case by defining a joint density:

$$p(\mathbf{q}, \mathbf{z} \mid \boldsymbol{\theta}) \propto \exp\left\{-\beta\left(V(\mathbf{q}) + \frac{\mu}{2} \parallel \mathbf{z} - \boldsymbol{\xi}(\mathbf{q}) \parallel^2 - \hat{A}_\mu(\mathbf{z}; \boldsymbol{\theta})\right)\right\} \quad (39)$$

where as in [37] an artificial spring with stiffness μ has been added. Clearly for $\mu \to \infty$ one recovers the aforementioned description, but for all other values of μ the formulation reduces to that of Equation (5) where in place of $V(\mathbf{q}, \mathbf{z})$ we now have $V(\mathbf{q}) + \frac{\mu}{2} \parallel \mathbf{z} - \boldsymbol{\xi}(\mathbf{q}) \parallel^2$. One can therefore obtain free energy surfaces for various μ values. For smaller μ the free energy would be flatter and in the extreme case of $\mu = 0$ it would be constant. As μ increases, the complexities of the free energy surface would become pronounced. Hence by exploiting the idea of section 2.4, a sequence of problems parametrized by μ rather than β, can be constructed to gradually move to larger μ values by using the free energy of the previous μ as an initial guess for the new one. The adaptive SMC scheme would ensure a smooth enough transition while retaining a good level accuracy for the approximations obtained.

3 Numerical Examples

3.1 Two-Dimensional Toy Example

Consider a two-dimensional system [54,34] with a single parameter z, interacting with potential energy

$$V(q; z) = \cos(2\pi z)(1 + d_1 q) + d_2 q^2$$

Assume that q given z and β is distributed according to

$$p(q|z, \beta) \propto \exp\left(-\beta V(q; z)\right)$$

23

where β is also a fixed parameter that plays the role of an inverse temperature. We wish to calculate an approximation $\hat{A}(z)$ of the free energy $A(z)$ on an interval $D = [-0.5, 0.5]$. The true free energy can be found analytically to be

$$A(z) = \cos(2\pi z) - \frac{d_1^2 \cos(2\pi z)^2}{4d_2} + c$$

where c is a constant that depends upon the specific choice of the fixed parameters. In what follows, we choose c so that $A(-0.5) = 0$.

To demonstrate our method in this simple example we used $d_1 = 2, d_2 = 30$. The potential energy $V(q; z)$ for this choice of the parameters is depicted in Figure 1(a). We fix the inverse temperature to $\beta = 10$. As shown in Figure 1(b), the distribution is bimodal with a big region of practically zero probability separating the two modes. Hence, metastability along the parameter z is apparent. The performance of the proposed method with respect to the number of particles used in the adaptive SMC scheme is depicted in Figures 2 and 3 which show the evolution of the estimated free energy landscape with $n = 100$ and $n = 10,000$ particles respectively. In both cases the method is capable of capturing the correct characteristics of the reference solution and as expected the noise in the computations is less when the number of particles is larger. In both cases the Robbins-Monro learning series is picked to be $\eta_m = m^{-p}$ with $p = 0.6$ and the learning rate $\lambda = 0.1$.

Figure 4(a) shows the first three kernels selected by the greedy scheme of section 2.2. Figure 4(b) depicts the log-values of the kernel weights $\{\theta_j\}_{j=1}^K$ which clearly demonstrate the ability of the proposed approach to provide sparse approximations. The first kernel selected has the greatest weight and hence it contains the majority of the information about the free energy curve. The rest of the kernels are progressive corrections of the estimate given by the
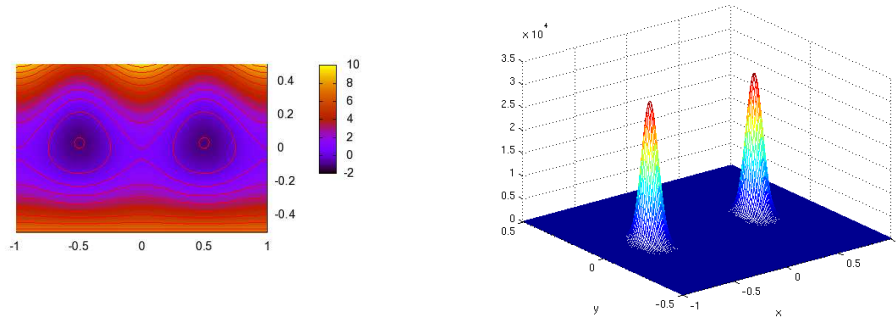
(a) The potential energy $V(q_1, q_2)$ for $d_1 = 2, d_2 = 30$

(b) The probability distribution $p(q_1, q_2 | \beta)$ for $d_1 = 2, d_2 = 30, \beta = 10$

Fig. 1. Potential energy and pdf for the toy example of section 3.1

first kernel. This conclusion is also supported by the result of Figure 5 which shows the evolution of the reduction in the KL divergence with respect to the total number of iterations as quantified by adding the Δ_{K+1} in Equation (20). Clearly the first kernel offers the greatest KL gain (Δ_1) and further kernel additions offer progressively smaller reductions in the KL divergence.

3.2 WCA Dimer

We consider $N = 16$ atoms in a two-dimensional fully periodic box of side l which interact with a purely repulsive WCA pair potential [34]:

$$
V_{\mathrm{WCA}}(r) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right] + \epsilon & , \text{if } r \geq r_0 \\ 0 & , \text{otherwise} \end{cases}
$$

where σ and ϵ give the length and energy scales respectively. Two of these atoms (say atoms 1 and 2) are assumed to form a dimer and their interaction

25

(a) K=1

(b) K=2

(c) K=3

(d) K=8
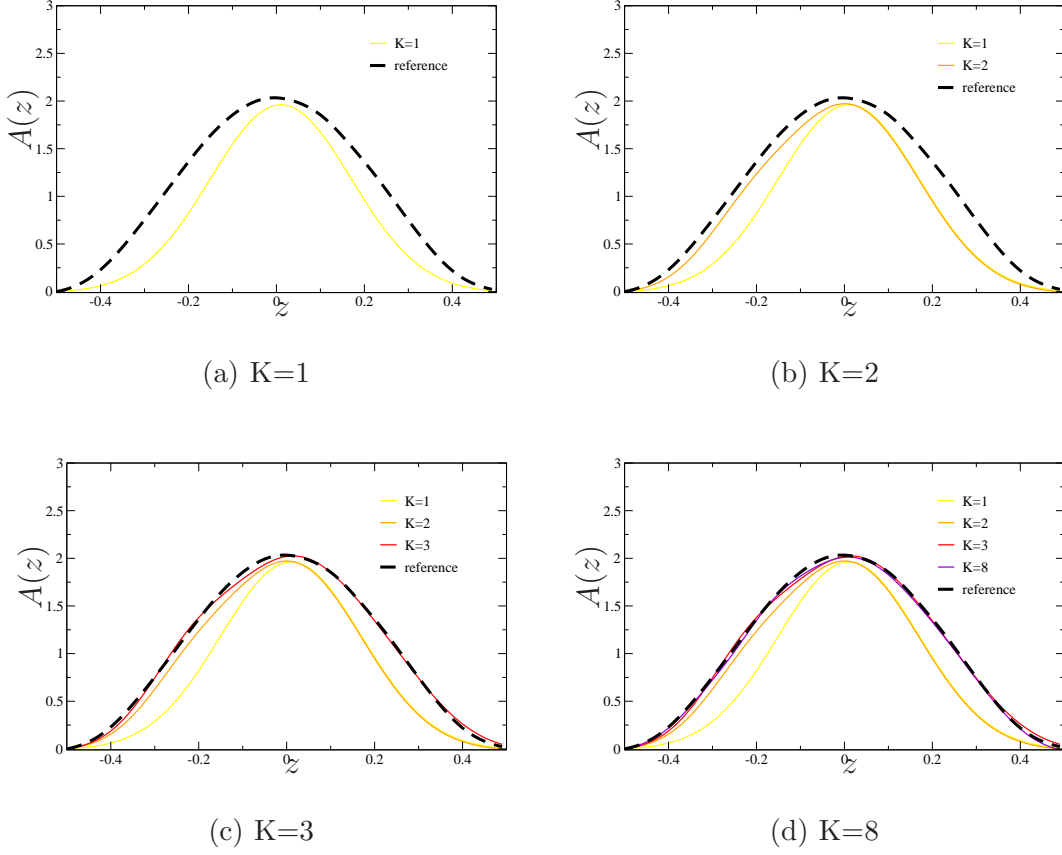
Fig. 2. Evolution of free energy estimates at various kernel numbers K when using $n = 100$ particles in the adaptive SMC scheme

is described instead with a double well potential:

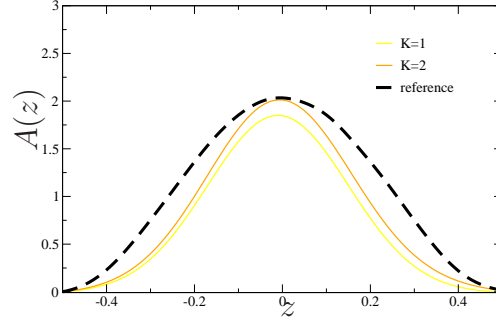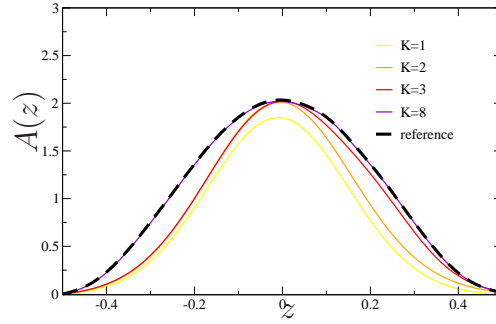$$V_S(r) = h \left[1 - \frac{(r - r_0 - w)^2}{w^2} \right]$$

where h, w, r_0 are fixed parameters and r the distance between them. This potential has two equilibrium points r_0 and $r_0 + 2w$. The barrier separating the two equilibria is h.

Let $\boldsymbol{q} = (\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N)$ with $\boldsymbol{q}_i \in \mathbb{R}^2$ denoting the position of atom i. The potential energy of the system is:

$$V(\boldsymbol{q}) = V_S(|\boldsymbol{q}_1 - \boldsymbol{q}_2|) + \sum_{i=1}^{2} \sum_{j=3}^{N} V_{\text{WCA}}(|\boldsymbol{q}_i - \boldsymbol{q}_j|) + \sum_{2 < i < j} V_{\text{WCA}}(|\boldsymbol{q}_i - \boldsymbol{q}_j|)$$
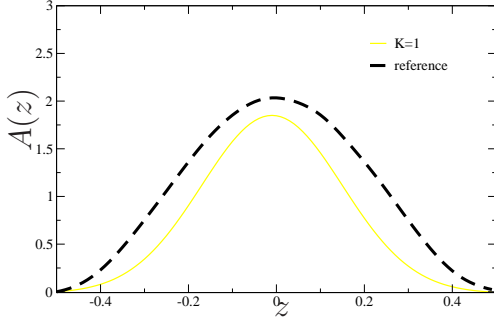
(a) K=1

(b) K=2
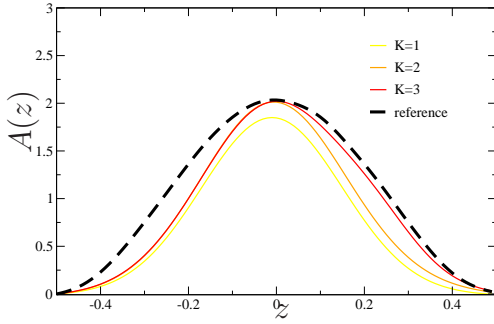
(c) K=3

(d) K=8

Fig. 3. Evolution of free energy estimates at various kernel numbers K when using $n = 10,000$ particles in the adaptive SMC scheme

We consider an NVT ensemble (the volume V is determined by the side of the box l). The probability distribution of the atomic positions $\boldsymbol{q}$ is

$$p(\boldsymbol{q}|\beta) \propto \exp\left(-\beta V(\boldsymbol{q})\right)$$

where $\beta = \frac{1}{k_B T}$, k_B is the Boltzman constant and T the temperature of the system. Under these assumptions atoms 1 and 2 will form a dimer with two equilibrium lengths. An *effective potential* of the dimer length in the presence of the other atoms is given by the free energy $A(r)$ with respect to the reaction coordinate

$$z = \xi(\boldsymbol{q}) = \| \boldsymbol{q}_1 - \boldsymbol{q}_2 \|_2$$

where $|| \cdot ||_2$ is the Euclidean norm of $\mathbb{R}^2$.
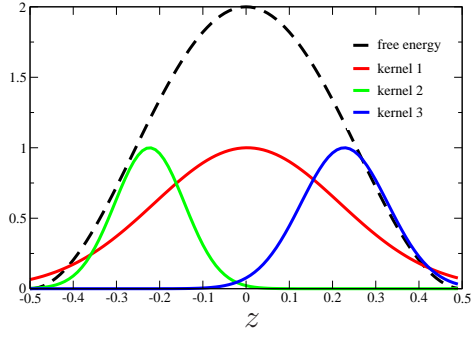
(a) First three kernels $K_j(.)$ picked by the greedy optimization scheme to illustrate selection of location and bandwidth parameters ($n = 10,000$)
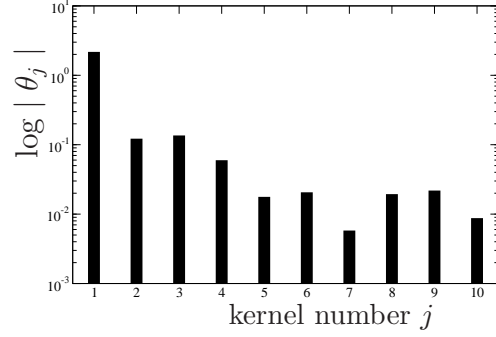
(b) Log absolute weights ($\log |\theta_j|$) of the first 10 kernels added. The θ value of the first kernel is over one order of magnitude larger than the rest

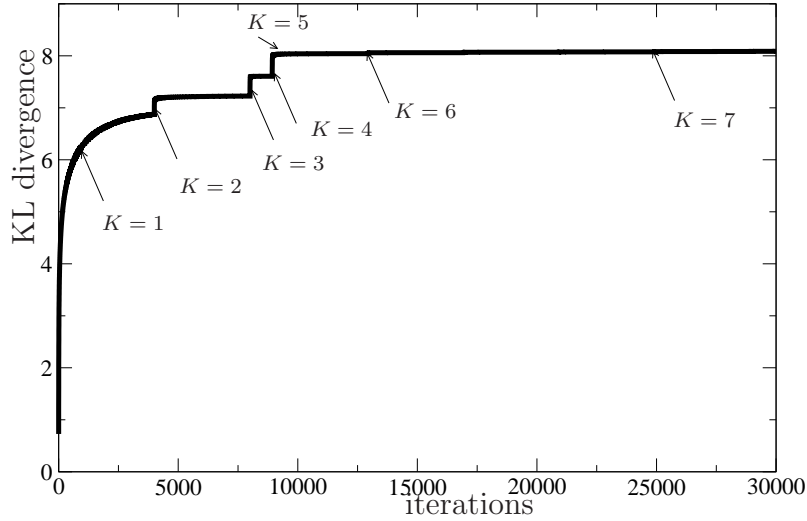Fig. 4. Kernels selected and kernel weights obtained with $n = 10,000$ particles

Fig. 5. Evolution of the reduction in the KL divergence at various K with respect tot the total number of iterations performed. Upon addition of each kernel a reduction Δ_K (Equation (20)) is achieved which becomes progressively smaller.

We calculate $A(z)$ using our scheme for two different box sizes (densities): $l = 4$ (high density) and $l = 12$ (low density). The parameters are set to $N = 16$ atoms, $\beta = 1, \epsilon = 1, \sigma = 1, h = 1, w = 0.5$. We employed $n = 500$

28

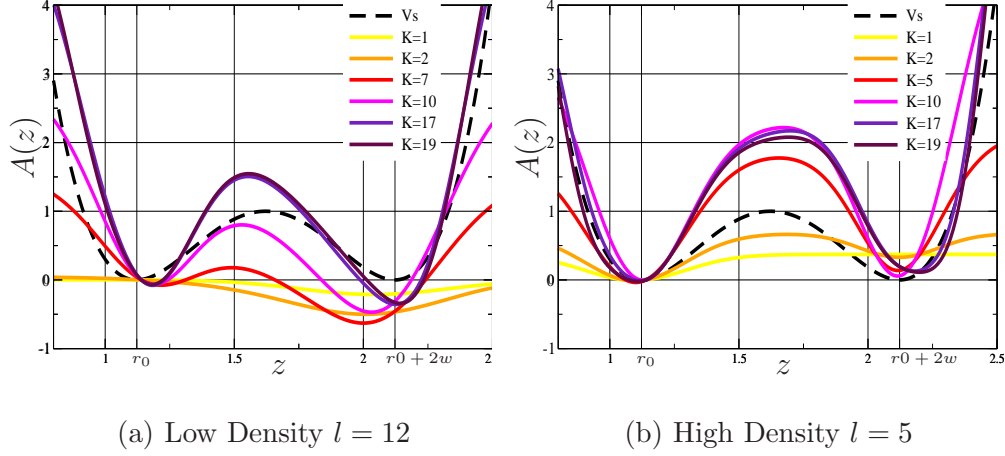(a) Low Density $l = 12$ (b) High Density $l = 5$

Fig. 6. The free energy of the dimer at two different densities compared with $V_S(r)$. Notice that at low density (a) the right well becomes the most probable. This situation is reversed at high density (b).

particles and the Robbins-Monroe learning series is again $\eta_m = m^{-p}$ with $p = 0.501$ and $\lambda = 0.1$. The resulting free energy curves at various stages of the estimation process with increasing number of kernels are shown in Figure 6. We notice that at *low density* i.e. when the box size is $l = 12$ (Figure 6(a)), the equilibria move to the right with the well closest to $r_0 + 2w$ becomes the most probable. Furthermore the free energy barrier is slightly decreased as compared to the *high density* case when $l = 5$ (Figure 6(b)). Under these conditions the equilibria move to the left and the well closest to r_0 becomes the most probable.

3.3 38-Atom Lennard-Jones Cluster (LJ$_{38}$)

We consider a 38-atom cluster in 3-dimensional space with pairwise interactions given by the Lennard-Jones potential:

$$V_{\text{LJ}}(r) = \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{40}$$

with ϵ and σ playing the role of energy and length scale respectively. Let the Cartesian coordinates of the system be

$$\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_{38}), \mathbf{q}_i \in \mathbb{R}^3 \tag{41}$$

Then the potential energy of the system is

$$V(\mathbf{q}) = \frac{1}{2} \sum_{i<j} V_{\text{LJ}} \left(|\mathbf{q}_i - \mathbf{q}_j| \right)$$

Finally we assume that the particles follow an NVT distribution of the form

$$p(\mathbf{q}|\beta) \propto \exp \left\{ -\beta V(\mathbf{q}) \right\}$$

where $\beta = 1/k_B T$. At zero temperature the system is known to have a global minimum yielding an FCC truncated octahedron (Figure 7(b)). The second and third lower energies give incomplete Mackey icosahedra. Furthermore there is a big number of liquid-like local minima ([19,5]).

Consider the family of order parameters initially introduced in [47]:

$$Q_l = \left(\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |\bar{Q}_{lm}|^2 \right)^2 \tag{42}$$

with

$$\bar{Q}_{lm} = \frac{1}{N_b} \sum_{r_{ij} < r_0} Y_{lm}(\theta_{ij}, \phi_{ij})$$

where the sum is over all the N_b pairs of atoms with $r_{ij} = |\mathbf{q}_i - \mathbf{q}_j| < r_0$, $Y_{lm}(\theta, \phi)$ is a spherical harmonic, while θ_{ij} and ϕ_{ij} are the polar and azimuthal

(a) $Q_4 \approx 0.01$ (b) $Q_4 \approx 0.19$ (truncated octahedron)

Fig. 7. Indicative metastable states corresponding to the two wells of the free energy landscape with respect to order parameter Q_4 (Equation (42))

angles of a bond vector with respect to an arbitrary coordinate system. In [5] it is shown that for $l = 4$, Q_4 can distinguish the FCC structure but not the icosahedral and liquid-like minima (Figure 7(a)). However, if one also considers the energy the two structures are well-separated. Hence, we define the two dimensional reaction coordinate:

$$\xi(\mathbf{q}) = (Q_4(\mathbf{q}), V(\mathbf{q}))$$

we compute the free energy

$$A(Q_4, E) = \beta^{-1} \int \exp\left\{-\beta V(\mathbf{q})\right\} \delta(Q_4 - Q_4(\mathbf{q})) \delta(E - V(\mathbf{q})) d\mathbf{q}$$

over the domain

$$D = [0, 0.2] \times [-175\epsilon, -145\epsilon]$$

for a temperature range $k_B T = 0.21$ to $k_B T = 0.091$ using the tempering scheme described in Section 2.4. We employ $n = 100$ particles and 10 MCMC/Rejuvenation steps per particle. At each $\beta = k_B T$, the Robbins-

31

Monro learning series was adjusted to $\eta_m = m^{-p}$ with $p = 0.501$ and a learning rate $\lambda = 0.1/\beta$. The adaptive SMC scheme automatically determined 260 intermediate steps/distributions in order to cover the whole range of the aforementioned temperatures. While the time step Δt_q employed in the MALA sampler was adaptively adjusted as discussed previously and took values between 10^{-4} (low temperatures) and 7×10^{-4} (high temperatures). The very first step, at $T = 0.21$ ($\beta = 4.76$) required $12,000$ optimization iterations to converge with a cost of approximately 7.2×10^5 time steps per particle. It is emphasized that due to the embarrassingly parallelizable nature of the SMC scheme employed , each particle can be simulated in a different CPU, largely independently of the rest. The sequence of intermediate β's determined automatically by the scheme discussed in section 2.4 is depicted in Figure 8. The similarity of the free energy surfaces at neighboring temperatures allowed us to converge with, on average, 800 optimization iterations at each intermediate β. The overall cost was 2.4×10^4 time steps per particle, i.e. equivalent to $1/30$ of the cost for calculating the free energy from scratch at the initial β. The free energy surfaces computed are depicted in Figure 9 at four indicative temperatures. The number of kernels selected by the algorithm varied between 90 and 120. As it has been reported in previous studies [5], we identified two metastable states at $Q_4 \approx 0.01$ which corresponds to the truncated octahedron and at $Q_4 \approx 0.19$ which corresponds to the icosahedron. The latter becomes more pronounced at lower temperatures.

In order to assess the quality of the results in two dimensions we also calculated the free energy profile using only Q_4 as the reaction coordinate (Figure 10). We also performed a numerical integration of the two-dimensional free energy surface i.e. by computing $A(Q_4) = -\beta^{-1} \log \int e^{-\beta A(Q_4, E)} dE$ with respect to
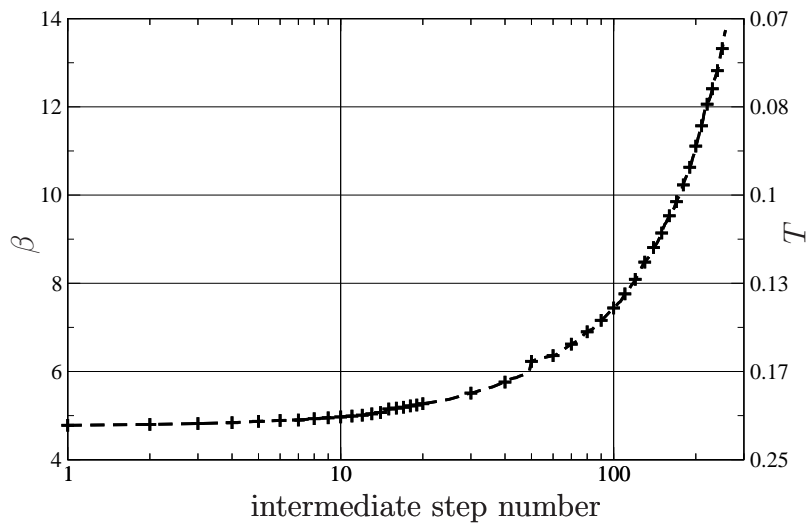
Fig. 8. Sequence of intermediate β's identified by the scheme discussed in section 2.4 for the LJ_{38} cluster. The free energy landscape was calculated at each of these temperatures by efficiently updating the free energy surface at the previous step.

the second reaction coordinate. The two free-energy curves are depicted in Figure 10 where good agreement is observed at two different temperatures.

4   Conclusions

In summary, the proposed method provides a unifying framework for estimating the free energy function simultaneously with biasing the dynamics. The minimization of the Kullback-Leibler divergence in the extended space provides rigorous convergence bounds and diagnostics. It requires minimal adjustment of parameter values a priori (basically only the learning rate λ and convergence tolerances) as it is adaptive and automatically promotes sparse representations of the free energy surface. It offers several possibilities for further improvements by considering different optimization schemes (e.g. noisy conjugate gradients) and employing different basis functions (e.g. wavelets).
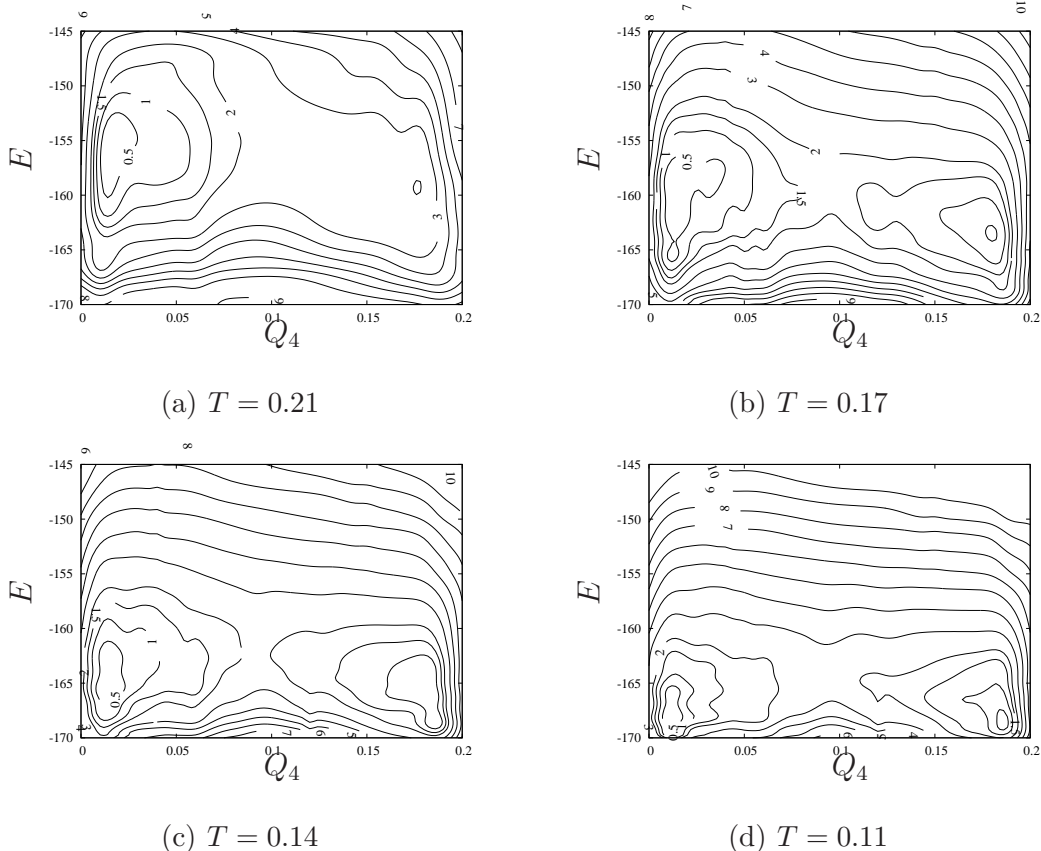
(a) $T = 0.21$

(b) $T = 0.17$

(c) $T = 0.14$

(d) $T = 0.11$

Fig. 9. Free energy contours $A(Q_4, E)$ with respect to the two reaction coordinates Q_4 (x-axis) and E (y-axis) at various temperatures for LJ_{38}

Its sequential nature allows the efficient computation of a family of free energy surfaces at different temperatures. We believe that these features make the proposed approach suitable to calculate the free energy of systems more physically challenging than the ones discussed in this paper.
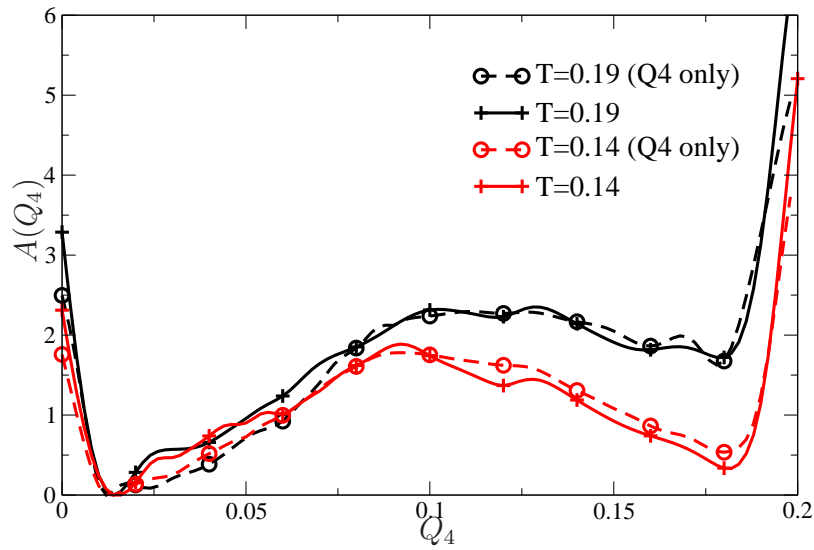
Fig. 10. Free energy profiles $A(Q_4)$ obtained using the proposed scheme while using only reaction coordinate (dashed lines) and by integrating the two dimensional free energy surface i.e. $A(Q_4) = -\beta^{-1} \log \int e^{-\beta A(Q_4,E)} dE$ (solid lines) for two temperatures $T = 0.19$ and $T = 0.14$

References

[1] Y.F. Atchade and J.S. Liu. The Wang-Landau algorithm for Monte Carlo computation in general state spaces. *Statistica Sinica*, 2(1983):1–25.

[2] Bernd Berg and Thomas Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Physical Review Letters*, 68(1):9–12, January 1992.

[3] A Berger. The Improved Iterative Scaling Algorithm: A Gentle Introduction. Technical report, School of Computer Science, Carnegie Mellon University, 1997.

[4] Giovanni Bussi, Alessandro Laio, and Michele Parrinello. Equilibrium Free Energies from Nonequilibrium Metadynamics. *Physical Review Letters*, 96(9):10–13, 2006.

[5] F. Calvo, J. P. Neirotti, David L. Freeman, and J. D. Doll. Phase changes in 38-atom Lennard-Jones clusters. II. A parallel tempering study of equilibrium and dynamic properties in the molecular dynamics and microcanonical ensembles. *The Journal of Chemical Physics*, 112(23):10350, 2000.

[6] E Cances, F Legoll, and G Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *edpsciences.org*, 2005.

[7] O Cappé, E Moulines, and T Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[8] C. Chipot and A. Pohorille. *Free energy calculations: theory and applications in chemistry and biology, front-matter*. Springer Berlin, 2006.

[9] N Chopin. Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *ANNALS OF STATISTICS*, 32(6):2385–2411, 2004.

[10] G. Ciccotti, T. Lelievre, and E. Vanden-Eijnden. Sampling Boltzmann-Gibbs distributions restricted on a manifold with diffusions: Application to free energy calculations. *Rapport de recherche du CERMICS*, 309:2006, 2006.

[11] Eric Darve. Vector free energy calculation with adaptive biasing force.

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

[13] P Del Moral. *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer New York, 2004.

[14] P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo for Bayesian Computation (with discussion) . In *Bayesian Statistics 8*. Oxford University Press, 2006.

[15] P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society B*, 68(3):411–436, 2006.

[16] S Della Pietra, V Della Pietra, and J Lafferty. Inducing Features of Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[17] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27:94–128, 1999.

[18] Bradley M Dickson, Frédéric Legoll, Tony Lelièvre, Gabriel Stoltz, and Paul Fleurat-Lessard. Free energy calculations: an efficient adaptive biasing potential method. *The journal of physical chemistry. B*, 114(17):5823–5830, May 2010.

[19] JPK Doye, MA Miller, and DJ Wales. The double-funnel energy landscape of the 38-atom Lennard-Jones cluster. *The Journal of Chemical Physics*, 110(14):6896, 1999.

[20] W N E, P G Ming, and P W Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *JOURNAL OF THE AMERICAN MATHEMATICAL SOCIETY*, 18(1):121–156, 2005.

[21] Marc Fasnacht, Robert Swendsen, and John Rosenberg. Adaptive integration method for Monte Carlo simulations. *Physical Review E*, 69(5):1–15, 2004.

[22] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Academic Pr, 2002.

[23] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185, May 1998.

[24] Jerome Henin, Giacomo Fiorin, Christophe Chipot, and Michael L Klein. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *Journal of Chemical Theory and Computation*, 6(1):35–47, January 2010.

[25] John E. Hunter, William P. Reinhardt, and Thomas F. Davis. A finite-time variational method for determining optimal paths and obtaining bounds on free energy changes from computer simulations. *The Journal of Chemical Physics*, 99(9):6856, 1993.

[26] C Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: a master-equation approach. *Rev. E*, 56:5018–5035, 1997.

[27] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, April 1997.

[28] G Kimeldorf and G Wahba. Some results on {Tchebycheffian} spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[29] J.G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3:300, 1935.

[30] P S Koutsourelakis. Design of complex systems in the presence of large uncertainties: A statistical approach. *Computer Methods in Applied Mechanics and Engineering*, 197(49-50):4092–4103, 2008.

[31] P S Koutsourelakis. Accurate uncertainty quantification using inaccurate models. *SIAM Journal of Scientific Computing*, 31(5):3274–3300, 2009.

[32] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562, 2002.

[33] T Leliévre, M Rousset, and G Stoltz. *FREE ENERGY COMPUTATIONS: A Mathematical Perspective*. Imperical College Press, 2010.

[34] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Computation of free energy dierences through nonequilibrium stochastic dynamics: The reaction coordinate case . *The Journal of chemical physics*, 222:) 624–643, 2007.

[35] 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.

[36] J S Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.

[37] L MARAGLIANO and E VANDEN-EIJNDEN. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical Physics Letters*, 426(1-3):168–175, July 2006.

[38] Luca Maragliano and Eric Vanden-Eijnden. Single-sweep methods for free energy calculations. *The Journal of chemical physics*, 128(18):184110, May 2008.

[39] Simone Marsili, Alessandro Barducci, Riccardo Chelli, Piero Procacci, and Vincenzo Schettino. Self-healing umbrella sampling: a non-equilibrium approach

for quantitative free energy calculations. *The journal of physical chemistry. B*, 110(29):14011–14013, July 2006.

[40] Hagai Meirovitch. Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. *Current opinion in structural biology*, 17(2):181–6, 2007.

[41] Paolo Raiteri, Alessandro Laio, Francesco Luigi Gervasio, Cristian Micheletti, and Michele Parrinello. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *The journal of physical chemistry. B*, 110(8):3533–9, March 2006.

[42] J M Rickman and R LeSar. Free-Energy Calculations in Materials Research. *Annual Review of Materials Research*, 32(1):195–217, 2002.

[43] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400407, 1951.

[44] C P Robert and G Casella. *Monte Carlo Statistical Methods*. Springer New York, 2nd edition, 2004.

[45] G O Roberts and J S Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16:351–367, 2001.

[46] N Schraudolph and T Graepel. Towards Stochastic Conjugate Gradient Methods. In *Proceedings of the 9th International Conference on Neural Information Processing, IEEE*, Singapore, 2002.

[47] PJ Steinhardt, DR Nelson, and M Ronchetti. Bond-orientational order in liquids and glasses. *Physical Review B*, 1983.

[48] G Stoltz. *Some mathematical methods for molecular and multiscale simulation*. PhD thesis, Ecole Nationale de Ponts et Chaussees, 2007.

[49] R Swendsen, M Fasnacht, and J Rosenberg. The adaptive integration method for calculating general free energy functions. *Computer Physics Communications*, 169(1-3):274–276, 2005.

[50] Cover T. and J Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

[51] M E Tipping. The Relevance Vector Machine. In S A Solla, T K Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press, 2000.

[52] M E Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research,*, 1:211–244, 2001.

[53] E. Vanden-Eijnden. Some recent techniques for free energy calculations. *Journal of Computational Chemistry*, 30(11), 2009.

[54] A F Voter. A method for accelerating the molecular dynamics simulation of infrequent events. *Journal of Chemical Physics*, 106(11):4665–4677, 1997.

[55] F Wang and D Landau. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. *Rev. E*, 64:56101, 2001.

[56] F Marty Ytreberg, Robert H Swendsen, and Daniel M Zuckerman. Comparison of free energy methods for molecular systems. *The Journal of chemical physics*, 125(18):184114, 2006.