From Biased to Unbiased Dynamics: An Infinitesimal Generator Approach

Timothée Devergne
CSML & ATSIM, Istituto Italiano di Tecnologia
timothee.devergne@iit.it

Vladimir R. Kostic
CSML, Istituto Italiano di Tecnologia
University of Novi Sad
vladimir.kostic@iit.it

Michele Parrinello
ATSIM, Istituto Italiano di Tecnologia
michele.parrinello@iit.it

Massimiliano Pontil
CSML, Istituto Italiano di Tecnologia
AI Centre, University College London
massimiliano.pontil@iit.it

Abstract

We investigate learning the eigenfunctions of evolution operators for time-reversal invariant stochastic processes, a prime example being the Langevin equation used in molecular dynamics. Many physical or chemical processes described by this equation involve transitions between metastable states separated by high potential barriers that can hardly be crossed during a simulation. To overcome this bottleneck, data are collected via biased simulations that explore the state space more rapidly. We propose a framework for learning from biased simulations rooted in the infinitesimal generator of the process and the associated resolvent operator. We contrast our approach to more common ones based on the transfer operator, showing that it can provably learn the spectral properties of the unbiased system from biased data. In experiments, we highlight the advantages of our method over transfer operator approaches and recent developments based on generator learning, demonstrating its effectiveness in estimating eigenfunctions and eigenvalues. Importantly, we show that even with datasets containing only a few relevant transitions due to sub-optimal biasing, our approach recovers relevant information about the transition mechanism.

1 Introduction

Dynamical systems and stochastic differential equations (SDEs) provide a general mathematical framework to study natural phenomena, with broad applications in science and engineering. Langevin SDEs, the main focus of this paper, are widely used to simulate physical processes such as protein folding or catalytic reactions [see e.g. 47, and references therein]. A main objective is to describe the dynamics of the process, forecast its evolution from a starting state, ultimately gaining insights on macroscopic properties of the system.

In molecular dynamics, the motion of a molecule is sampled according to a potential energy $U(x)$, where the state vector x represents the positions of all the atoms. Specifically, the Langevin equation $dX_t = -\nabla U(X_t)dt + \sigma dW_t$ describes the stochastic behavior of the system at thermal equilibrium, where X_t is the random position of the state at time t, the scalar σ is a multiple of the square root of the system's temperature, and W_t is a vector random variable describing thermal fluctuations (Brownian motion). Most often, the atoms evolve in metastable states that are separated by barriers which can hardly be crossed during a simulation. For instance, for a protein the free energy barrier between the folded and unfolded states is larger than thermal agitation, making the transition between the

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2404

two states a rare event. Consequently, long trajectories need to be simulated before such interesting events are observed. In fact, one needs to observe many events to get the relevant thermodynamics (free energy) and kinetics (transition rates) information [34]. Beyond molecular dynamics, the slow mixing behavior of many systems modeled by SDEs is a major bottleneck in the study of rare events, and so designing methodologies which can accelerate the process is paramount.

A general idea to overcome the above problem is to perturb the system dynamics. One important approach which has been put in place in molecular dynamics is the so-called "bias potential enhanced sampling" [25, 44, 11]. The main idea is to add to the potential energy a bias potential V, thereby lowering the barrier and allowing the system state to be explored more rapidly. To make this approach tractable in large systems, V is often chosen as a function of a few wisely selected variables called collective variables (CVs). For instance, if a chemical reaction involves a bond breaking, physical intuition suggests to choose the distance between the reactive atoms [26, 29]. However, for complex processes, hand-crafted CVs might be "suboptimal", meaning that some of the degrees of freedom important for the transition are not taken into account, making the biasing process inefficient.

In recent years, machine learning approaches have been employed to find the most relevant CVs [8, 42, 14, 9, 10, 4, 27]. A key idea is to use available dynamical information to construct the CVs [41, 31, 7, 42]. For instance, if one can identify the slowest degrees of freedom of the system, one can accurately describe the transitions between metastable states. These approaches are based on learning the transfer operator of the system, which models the conditional expectation of a function (or observable) of the state at a future time, given knowledge of the state at the initial time. It is learned from the behavior of dynamical correlation functions at large lag times which reflects the slow modes of the system. The leading eigenfunctions of the learned transfer operator can then be used as CVs in biased simulations. Moreover, they provide valuable insights into the transition mechanism, such as the location of the transition state ensemble [48]. Still, this approach suffers from the same shortcoming described above, namely if the system is slowly mixing, long trajectories are needed to learn the transfer operator and extract good eigenfunctions.

More recently, there has been growing interest in learning the infinitesimal generator of the process [15, 1, 50, 20], which allows one to overcome the difficult choice of the lag-time. The statistical learning properties of generator learning have been addressed in [21], where an approach based on the resolvent operator has been proposed in order to bypass the unbounded nature of the generator. However the key difficulty of learning from biased simulations remains an open question. In this work, we prove that the infinitesimal generator is the adequate tool to deal with dynamical information from biased data. Leveraging on the statistical learning considerations in [23, 21], we introduce a novel procedure to compute the leading eigenpairs of the infinitesimal generator from biased dynamics, opening the doors to numerous applications in computational chemistry and beyond.

Contributions In summary, our main contributions are: **1)** We introduce a principle approach, based on the resolvent of the generator, to extract dynamical properties from biased data; **2)** We present a method to learn the generator from a prescribed dictionary of functions; **3)** We introduce a neural network loss function for learning the dictionary, with provable learning guarantees; **4)** We report experiments on popular molecular dynamics benchmarks, showing that our approach outperforms state-of-the-art transfer operator and recent generator learning approaches in biased simulations. Remarkably, even with datasets containing only a few relevant transitions due to sub-optimal biasing, our method effectively recovers crucial information about the transition mechanism.

Paper organization In Section 2, we introduce the learning problem. Section 3 explores limitations of transfer operator approaches. In Section 4, we review a recent generator learning approach [21] and adapt it to nonlinear regression with a finite dictionary of functions. Section 5 presents our method for learning from biased dynamics. Finally, in Section 6, we report our experimental findings.

2 Learning dynamical systems from data

In this section, we address learning stochastic dynamical systems from data. After introducing the main objects, we review existing data-driven approaches and conclude with practical challenges. We ground the discussion in the recently developed statistical learning theory, [22–24], contributing in particular to the existence of physical priors and feasibility of data acquisition for successful learning.

Stochastic differential equations (SDEs) and evolution operators While our observations in the paper naturally extend to general forms of SDEs [see e.g. 33], to simplify the exposition, we focus on the Langevin equation, which is most relevant to our discussion of biased simulations. Specifically, we consider the overdamped Langevin equation

$$dX_t = -\nabla U(X_t)dt + \sqrt{2\beta^{-1}}dW_t \quad \text{and} \quad X_0 = x, \tag{1}$$

describing dynamics in a (state) space $\mathcal{X} \subseteq \mathbb{R}^d$, governed by the *potential* $V : \mathbb{R}^d \to \mathbb{R}$ at the *temperature* $\beta^{-1} = k_B T$, where W_t is a $\mathbb{R}^d$-dimensional standard Brownian motion.

The SDE (1) admits a unique strong solution $X = (X_t)_{\geqslant 0}$ that is a Markov process to which we can associate the semigroup of Markov *transfer operators* $(\mathcal{T}_t)_{t\geq 0}$ defined, for every $t \geq 0$, as

$$[\mathcal{T}_t f](x) := \mathbb{E}[f(X_t)|X_0 = x], \;\; x \in \mathcal{X},\; f\colon \mathcal{X} \to \mathbb{R}. \tag{2}$$

For (1) the distribution of X_t converges to the *invariant measure* π on $\mathcal{X}$ called the Boltzmann distribution, given by $\pi(dx) \propto e^{-\beta V(x)}dx$. In such cases, one can define the semigroup on $L^2_\pi(\mathcal{X})$, and characterize the process by the *infinitesimal generator*

$$\mathcal{L} := \lim_{t\to 0^+}(\mathcal{T}_t - I)/t$$

defined on the Sobolev space $H^{1,2}_\pi(\mathcal{X})$ of functions in $L^2_\pi(\mathcal{X})$ whose gradient are also in $L^2_\pi(\mathcal{X})$. The transfer operator and the generator are linked one to another by the formula $\mathcal{T}_t = \exp(t\mathcal{L})$. Moreover, it can be shown (see Appendix A) that the generator $\mathcal{L}$ acts on $f\colon \mathcal{X} \to \mathbb{R}$ as

$$\mathcal{L}f = -\langle \nabla U, \nabla f\rangle + \beta^{-1}\Delta f, \tag{3}$$

which, integrating by parts, gives $\int(\mathcal{L}f)g\, d\pi = -\beta^{-1}\int\langle \nabla f, \nabla g\rangle\, d\pi = \int f(\mathcal{L}g)d\pi$, showing that $\mathcal{L}$ is self-adjoint. If $\mathcal{L}$ has only a discrete spectrum, one can solve (1) by computing the spectral decomposition

$$\mathcal{L} = \textstyle\sum_{i\in\mathbb{N}}\lambda_i f_i \otimes f_i, \tag{4}$$

Using (2) and the exponential relation between the transfer operator and the generator, one can write

$$[\mathcal{T}_t f](x) := \mathbb{E}[f(X_t)|X_0 = x] = \textstyle\sum_{i\in\mathbb{N}} e^{t\lambda_i} f_i(x)\langle f_i, f\rangle, \;\; x \in \mathcal{X},\; f\colon \mathcal{X} \to \mathbb{R} \tag{5}$$

where the timescales of the process appear as the inverses of the generator eigenvalues. Consequently, the eigenpairs of the generator offer valuable insight about the transitions within the studied system.

Learning from simulations The main difference underpinning the development of learning algorithms for the *transfer operator* and the *generator* lies in the nature of the data used. While for the transfer operator we can *only* observe a *noisy* evaluation of the output to learn a compact operator, in the case of the generator, knowing the drift and diffusion coefficients allows us to compute the output, albeit at the cost of learning an *unbounded differential operator*. Consequently, learning methods for the former align with vector-valued regression in function spaces [22], whereas methods for the latter, as discussed in the following section, are more akin to physics-informed regression algorithms. In both settings, we learn operators defined on a function (hypothesis) space, formed by the linear combinations of a prescribed set of basis functions (dictionary) $z_j : \mathcal{X} \to \mathbb{R}$, $j \in [m]$,

$$\mathcal{H} := \Big\{ h_u = \textstyle\sum_{j\in[m]} u_j z_j \;\big|\; u = (u_1, \dots, u_m) \in \mathbb{R}^m \Big\}. \tag{6}$$

The choice of the dictionary, instrumental in designing successful learning algorithms, may be based on prior knowledge on the process or learned from data [24, 30, 50]. The space $\mathcal{H}$ is naturally equipped with the geometry induced by the norm $\|h_u\|^2_{\mathcal{H}} := \sum_{j=1}^m u_j^2$. Moreover, every operator $A\colon \mathcal{H} \to \mathcal{H}$ can be identified with matrix $\mathrm{A} \in \mathbb{R}^{m\times m}$ by $Ah_u = z(\cdot)^\intercal \mathrm{A} u$. In the following, we will refer to A and A as the same object, explicitly stating the difference when necessary.

Transfer operator learning Learning the transfer operator $\mathcal{T}_t$ can be simply seen as the vector-valued regression problem [22], in which the action of $\mathcal{T}_t\colon L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X})$ on the domain $\mathcal{H} \subseteq L^2_\pi(\mathcal{X})$ is estimated by an operator $\widehat{T}_t\colon \mathcal{H} \to \mathcal{H}$. This aims to minimize the mean square error (MSE) w.r.t. the invariant distribution. Given a dataset $\mathcal{D}_n := (x_i, y_i = x_{i+1})_{i=1}^n$ of time-lag $t > 0$ consecutive states from a trajectory of the process, a common approach is to minimize the regularized empirical MSE, leading to the ridge regression (RR) estimator $\widehat{\mathrm{T}}_\gamma := \widehat{\mathrm{C}}_\gamma^{-1}\widehat{\mathrm{C}}_t$, where the empirical covariance matrices are $\widehat{\mathrm{C}} = \frac{1}{n}\sum_{i\in[n]} z(x_i)z(x_i)^\intercal$ and $\widehat{\mathrm{C}}_t = \frac{1}{n}\sum_{i\in[n]} z(x_i)z(y_i)^\intercal$. We then estimate the eigenpairs (λ_i, f_i) in (4) by the eigenpairs $(\widehat{\mu}_i, \widehat{u}_i)$ of $\widehat{\mathrm{T}}_\gamma$ as $\widehat{\lambda}_i := \ln(\widehat{\mu}_i/t)$ and $\widehat{f}_i := z(\cdot)^\intercal \widehat{u}_i$.

https://doi.org/10.52202/079017-2404

We stress that transfer operator approaches crucially relies on the definition of the time-lag t from which dynamics is observed. Setting this value is a delicate task, depending on the events one wants to study. If t is chosen too small, the cross-covariance matrices will be too noisy for slowly mixing processes. On the other hand, if t is too large, because the relevant phenomena occur at large time scales, a very long simulation is needed to compute the covariance matrices. In order to overcome this problem biased simulations can be used, which we discuss next.

3 Learning from biased simulations

As discussed above, in molecular dynamics, the desired physical phenomena often cannot be observed within an affordable simulation time. To address this, one solution is to modify the potential,

$$U'(x) := U(x) + V(x), \quad x \in \mathcal{X}$$

where we assume that the introduced perturbation (a form of bias in the data) $V(x)$ is known. For example the bias potential V may be constructed from previous system states to promote transitions to not yet visited regions. One of the prototypical examples is metadynamics [25], where V is a sum of Gaussians built on the fly in order to reduce the barrier between metastable states. However, the bias potential alters the invariant distribution [12], making it challenging to recover the unbiased dynamics from biased data. Denoting the invariant measure of the perturbed process by π' and its generator by $\mathcal{L}' \colon H^{1,2}_{\pi'}(\mathcal{X}) \to H^{1,2}_{\pi'}(\mathcal{X})$, our principal objective is thus to:

> *Gather data from simulations generated by $\mathcal{L}'$ to learn the spectral decomposition of the unperturbed generator $\mathcal{L}$.*

To tackle this problem, we note that since the eigenfunctions of the generator $\mathcal{L}$ are also eigenfunctions of every transfer operator $\mathcal{T}_t = e^{t\mathcal{L}}$, we can address the related problem of learning the transfer operator from perturbed dynamics. Unfortunately, there is an inherent difficulty in doing so. While one *typically knows the perturbation* in the generator, that is $\mathcal{L}' = \mathcal{L} + \langle \nabla V, \nabla(\cdot) \rangle$, this knowledge is not easily transferred to the perturbation of the transfer operator. Indeed, recalling that $\mathcal{T} := \mathcal{T}_1 = e^{\mathcal{L}}$, and since the differential operator $\langle \nabla V, \nabla(\cdot) \rangle$ in general does not share the same eigenstructure of $\mathcal{L}$, one has that

$$\mathcal{T}' := e^{\mathcal{L}'} = e^{\mathcal{L} - \langle \nabla V, \nabla(\cdot) \rangle} \neq \mathcal{T} e^{-\langle \nabla V, \nabla(\cdot) \rangle}.$$

Simply put, the generator depends linearly on the bias, while the transfer operator does not. One strategy to overcome the data distribution change, is to *adapt* the notion of the risk. To discuss this idea, recall that the invariant distribution of overdampted Langevin dynamics is the Boltzmann distribution defined by the potential. Hence, we have that

$$\pi(dx) = \frac{e^{-\beta U(x)} dx}{\int e^{-\beta U(x)} dx}, \; \pi'(dx) = \frac{e^{-\beta U'(x)} dx}{\int e^{-\beta U'(x)} dx} \;\text{ and }\; \frac{d\pi}{d\pi'}(x) = \frac{e^{\beta V(x)}}{\int e^{\beta V(x)} \pi'(dx)} \tag{7}$$

where the last term is the Radon-Nikodym derivative, which exposes the data-distribution change. Consequently, we can express the covariance operators for the unperturbed process as weighted expectations of the perturbed data features

$$\mathrm{C} = \mathbb{E}_{X' \sim \pi'} \left[\tfrac{d\pi}{d\pi'}(X') z(X') z(X')^\intercal \right]. \tag{8}$$

However, since the transition kernel of the process $(X'_t)_{t \geq 0}$ generated by $\mathcal{L}'$ is different from that of the original process, the above reasoning does not hold for the cross-covariance matrix, that is,

$$\mathrm{C}_t := \mathbb{E}_{X_0 \sim \pi'} \left[\tfrac{d\pi}{d\pi'}(X_0)\, z(X_0) z(X_t)^\intercal \right] \neq \mathbb{E}_{X'_0 \sim \pi'} \left[\tfrac{d\pi}{d\pi'}(X'_0)\, z(X'_0) z(X'_t)^\intercal \right] =: \mathrm{C}'_t,$$

Consequently, the estimator $\widehat{T}_t$ obtained by minimizing the reweighed risk functional $\mathcal{R}'(\widehat{T}_t) := \mathbb{E}_{X_0 \sim \pi'} \left[\tfrac{d\pi}{d\pi'}(X_0) \, \| z(X'_t) - \widehat{\mathrm{T}}_t^\intercal z(X_0) \|_2^2 \right]$ *does not* minimize the true risk since $\mathcal{R}'(\widehat{T}_t) \neq \mathcal{R}(\widehat{T}_t)$. Despite this difference, whenever the perturbation is small or controlled and the *time-lag t is small enough*, estimating the true transfer operator of the process from the perturbed dynamics via reweighed covariance/cross-covariance operators has been systematically used as the state-of-the art approach in the field of atomistic simulations [7, 9, 31, 49]. The (limited) success of such approaches is based on a delicate balance of a small enough lag-time and biased potential, since for small $t > 0$ one can approximate C_t by C'_t and minimize $\mathcal{R}'(T) \approx \mathcal{R}(T)$ over $T \colon \mathcal{H} \to \mathcal{H}$.

4 Infinitesimal generator learning

In this section, we address generator learning. While there has been significant progress on this topic [24, 15, 35, 50, 20], we follow the recent approach in [21] for learning the generator $\mathcal{L}$ on an a priori fixed hypothesis space $\mathcal{H}$ through its resolvent. Leveraging on its strong statistically guarantees, we adapt it from kernel regression to nonlinear regression over a dictionary of basis functions, setting the stage for the development of our deep-learning method.

While transfer operator learning does not require any prior knowledge of the system's drift and diffusion, making use of this information helps learning the generator and avoids the need for setting the time lag parameter. We briefly discuss how to achieve this for over-damped Langevin processes when the constant diffusion term is known. We estimate the generator *indirectly* via its resolvent $(\eta I - \mathcal{L})^{-1}$, where $\eta > 0$ is a prescribed parameter. To this end, we observe that the action of the resolvent in $\mathcal{H}$ can be expressed as $((\eta I - \mathcal{L})^{-1} h_u)(x) = \chi_\eta(x)^\intercal u$, where χ_η is the embedding of the resolvent $(\eta I - \mathcal{L})^{-1}$ into $\mathcal{H}$, given by $\chi_\eta(x) = \int_0^\infty \mathbb{E}[z(X_t) e^{-\eta t} \,|\, X_0 = x] dt$, $x \in \mathcal{X}$, see [21]. We then aim to approximate $\chi_\eta(x) \approx \mathrm{G}^* z(x)$ by a matrix $\mathrm{G} \in \mathbb{R}^{m \times m}$. Unfortunately the embedding of the resolvent is not known in close form. To overcome this, we contrast the resolvent by defining a *regularized energy kernel* $\mathfrak{E}_\pi^\eta \colon H_\pi^{1,2}(\mathcal{X}) \times H_\pi^{1,2}(\mathcal{X}) \to \mathbb{R}$, given by $\mathfrak{E}_\pi^\eta[f,g] = \mathbb{E}_{x\sim\pi}\left[\eta f(x) g(x) - f(x)[\mathcal{L}g](x)\right]$, which using (3) becomes

$$\mathfrak{E}_\pi^\eta[f,g] = \mathbb{E}_{x\sim\pi}\left[\eta f(x) g(x) + f(x) \nabla U(x)^\intercal \nabla g(x) - \tfrac{1}{\beta} f(x) \Delta g(x)\right], \tag{9}$$

and, due to the identity $\int f \mathcal{L} g \, d\pi = -\beta^{-1} \int (\nabla f)^\intercal (\nabla g) d\pi$, also

$$\mathfrak{E}_\pi^\eta[f,g] = \mathbb{E}_{x\sim\pi}\left[\eta f(x) g(x) + \tfrac{1}{\beta} \textstyle\sum_{k\in[d]} \partial_k f(x) \partial_k g(x)\right]. \tag{10}$$

Since $\mathcal{L}$ is negative semi-definite, the above kernel induces the *regularized squared energy norm* $\mathfrak{E}_\pi^\eta \colon H_\pi^{1,2}(\mathcal{X}) \to [0, +\infty)$ by $\mathfrak{E}_\pi^\eta[f] := \mathfrak{E}_\pi^\eta[f,f] = \mathbb{E}_{x\sim\pi}\left[\eta f^2(x) - f(x)[\mathcal{L}f](x)\right]$. It counteracts the resolvent and balances the transient dynamics (energy) of the process with the invariant distribution π. In a nutshell, instead of using the mean square error of $f(x) := \|\chi_\eta(x) - \mathrm{G}^\intercal z(x)\|_2$ to define the risk, we "*fight fire with fire*" and penalize the energy to formulate the *generator regression problem*

$$\min_{G\colon \mathcal{H}\to\mathcal{H}} \mathcal{R}_\partial(G) \equiv \mathcal{R}_\partial(\mathrm{G}) := \mathfrak{E}_\pi^\eta\left[\|\chi_\eta(\cdot) - \widehat{\mathrm{G}}^\intercal z(\cdot)\|_2\right]. \tag{11}$$

Indeed, this risk overcomes the difficulty of not knowing χ_η. To show this, let us define the space $\mathcal{H}_\pi^\eta(\mathcal{X}) := \{f \in H_\pi^{1,2}(\mathcal{X}) \,|\, \mathfrak{E}_\pi^\eta[f] < \infty\}$ associated to the energy norm $\|f\|_{\mathcal{H}_\pi^\eta} := \sqrt{\mathfrak{E}_\pi^\eta[f]}$, and recalling that the operator $G\colon \mathcal{H} \to \mathcal{H}$ is identified with a matrix $\mathrm{G} \in \mathbb{R}^{m\times m}$ via $G h_u = z(\cdot)^\intercal (\mathrm{G} u)$, define the (injection) operator $\mathcal{Z}\colon \mathbb{R}^m \to \mathcal{H}_\pi^\eta$ by $\mathcal{Z} u = z(\cdot)^\intercal u$, for every $u \in \mathbb{R}^m$. Then, since $\mathrm{HS}(\mathbb{R}^m, \mathcal{H}_\pi^\eta) \equiv \mathrm{HS}(\mathcal{H}, \mathcal{H}_\pi^\eta)$, the norm is the sum of squared $\mathcal{H}_\pi^\eta$ norm over the standard basis in $\mathbb{R}^m$, and one obtains

$$\begin{aligned} \mathcal{R}_\partial(G) &= \|(\eta I - \mathcal{L})^{-1} - G\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}_\pi^\eta)} \\ &= \underbrace{\|P_\mathcal{H}(\eta I - \mathcal{L})^{-1} - G\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}_\pi^\eta)}}_{\text{projected problem}} + \underbrace{\|(I - P_\mathcal{H})(\eta I - \mathcal{L})^{-1}\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}_\pi^\eta)}}_{\text{representation error } \rho(\mathcal{H})}, \end{aligned} \tag{12}$$

where $P_\mathcal{H}$ is the orthogonal projector in $\mathcal{H}_\pi^\eta(\mathcal{X})$ onto $\mathcal{H}$. In learning theory $\rho(\mathcal{H})$ is known as the approximation error of the hypothesis space $\mathcal{H}$ [see e.g. 43]. While this error may vanish for infinite-dimensional spaces, when $\mathcal{H}$ is finite dimensional, controlling $\rho(\mathcal{H})$ is crucial to achieving statistical consistency. This can be accomplished by minimizing (11), which is equivalent to

$$\min_{G\colon \mathcal{H}\to\mathcal{H}} \|P_\mathcal{H}(\eta I - \mathcal{L})^{-1} - G\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}_\pi^\eta)} = \|\mathcal{Z}(\mathcal{Z}^*\mathcal{Z})^\dagger \mathcal{Z}^* (\eta I - \mathcal{L})^{-1} \mathcal{Z} - \mathcal{Z}\mathrm{G}\|^2_{\mathrm{HS}(\mathbb{R}^m,\mathcal{H}_\pi^\eta)} \tag{13}$$

where $(\cdot)^\dagger$ is the Moore-Penrose's pseudoinverse. Using the covariance matrices

$$\mathcal{Z}^*(\eta I - \mathcal{L})^{-1}\mathcal{Z} = \mathrm{C} = \left(\mathbb{E}_{x\sim\pi}[z_i(x) z_j(x)]\right)_{i,j\in[m]}, \quad \text{and} \quad \mathrm{W} = \mathcal{Z}^*\mathcal{Z} = \left(\mathfrak{E}_\pi^\eta[z_i, z_j]\right)_{i,j\in[m]}, \tag{14}$$

w.r.t. the invariant distribution and energy, respectively, gives the ridge regularized (RR) solution $\mathrm{G} = (\mathrm{W} + \gamma \mathrm{I})^{-1}\mathrm{C}$, $\gamma > 0$. The induced RR estimator of the resolvent, $G_{\eta,\gamma} : \mathcal{H} \to \mathcal{H}$ is given, for every $h_u \in \mathcal{H}$, by $G_{\eta,\gamma} h_u := \mathcal{Z}(\mathrm{W} + \gamma \mathrm{I})^{-1}\mathrm{C} u = z(\cdot)^\intercal (\mathrm{W} + \gamma \mathrm{I})^{-1}\mathrm{C} u$, and it can be estimated given data from π by replacing expectation and the energy in (14) with their empirical counterparts.

https://doi.org/10.52202/079017-2404

5 Unbiased learning of the infinitesimal generator from biased simulations

In this section, we present the main contributions of this work: approximating the leading eigenfunctions (corresponding to the slowest time scales) of the infinitesimal generator from biased data.While the general pipeline for the method can be found in figure 1, in the following, we first address regressing the generator on an a priori fixed hypothesis space $\mathcal{H}$. Then we introduce our deep-learning method to either build a suitable space $\mathcal{H}$, or even directly learn the eigenfunctions.

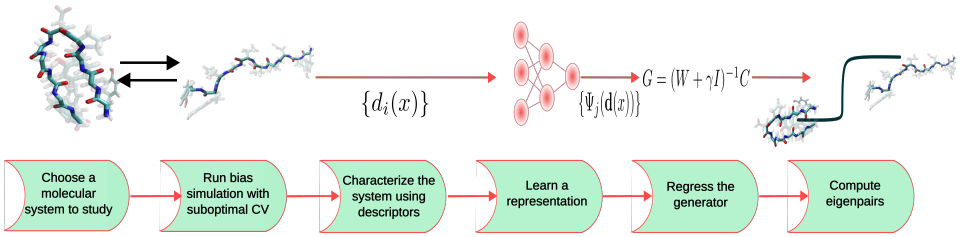

Figure 1: Pipeline of our method: from biased simulations to timescales and metastable states.

Unbiasing generator regression Whenever π is absolutely continuous w.r.t. π', the regularized energy kernel (9) satisfies the simple identity

$$\mathfrak{E}^{\eta}_{\pi}[f,g] = \mathfrak{E}^{\eta}_{\pi'}\big[f\sqrt{\tfrac{d\pi}{d\pi'}}, g\sqrt{\tfrac{d\pi}{d\pi'}}\big], \quad f,g \in H^{1,2}_{\pi}(\mathcal{X}), \tag{15}$$

which, recalling the rightmost equation in (7), implies that when the bias V and the diffusion coefficient β are known, the energy kernel can be empirically estimated through samples from π' via (10). Moreover, when the potential U is known too, we can use (9). Now, leveraging on (15) we directly obtain that

$$\mathcal{R}_{\partial}(G) \equiv \mathcal{R}_{\partial}(\mathrm{G}) = \mathfrak{E}^{\eta}_{x\sim\pi'}\big[\|\chi_\eta(x) - \widehat{\mathrm{G}}^{\mathsf{T}} z(x)\|_2 \sqrt{\tfrac{d\pi}{d\pi'}(x)}\big] \leq \kappa_V\, \mathfrak{E}^{\eta}_{x\sim\pi'} \|\chi_\eta(x) - \widehat{\mathrm{G}}^{\mathsf{T}} z(x)\|_2 \tag{16}$$

where $\kappa_V = \operatorname{ess\,sup}_{x\sim\pi'} \frac{d\pi}{d\pi'}(x)$, which recalling (7) is finite whenever the bias V is essentially bounded. Therefore, in sharp contrast to transfer operator learning, whenever the true embedding $\chi_\eta(x)$ can be estimated, one can derive principled estimators of the true generator $\mathcal{L}$'s dominant eigenpairs from the biased dynamics generated by $\mathcal{L}'$. This is established by the following proposition, the proof of which is presented in Appendix B.

Theorem 1. *Let $\mathcal{D}_n = (x'_i)_{i\in[n]}$ be the biased dataset generated from π'. Let $w(x) = e^{\beta V(x)}$ and define the empirical covariances w.r.t. the empirical distribution $\widehat{\pi}' = n^{-1}\sum_{i\in[n]} \delta_{x'_i}$ by*

$$\widehat{\mathrm{C}} = \big(\mathbb{E}_{x'\sim\widehat{\pi}'}[w(x') z_i(x') z_j(x')]\big)_{i,j\in[m]} \quad \text{and} \quad \widehat{\mathrm{W}} = \big(\mathfrak{E}^{\eta}_{\widehat{\pi}'}[\sqrt{w} z_i, \sqrt{w} z_j]\big)_{i,j\in[m]}. \tag{17}$$

Compute the eigenpairs $(\nu_i, v_i)_{i\in[m]}$ of the RR estimator $\widehat{\mathrm{G}}_{\eta,\gamma} = (\widehat{\mathrm{W}} + \eta\gamma \mathrm{I})^{-1}\widehat{\mathrm{C}}$, and estimate the eigenpairs in (4) as $(\widehat{\lambda}_i, \widehat{f}_i) = (\eta - 1/\nu_i, z(\cdot)^{\mathsf{T}} v_i)$. If the elements of $\mathcal{H}$ and their gradients are essentially bounded, and $\lim_{m\to\infty} \rho(\mathcal{H}) = 0$, then for every $\varepsilon > 0$, there exist $(m, n, \gamma) \in \mathbb{N}\times\mathbb{N}\times\mathbb{R}_+$, such that, for every $i \in [m]$, $|\lambda_i - \widehat{\lambda}_i| \leq \varepsilon$ and $\sin_{L^2_\pi}(\sphericalangle(f_i, \widehat{f}_i)) \leq \varepsilon$, with high probability.

Note that, due to the form of the estimator, the normalizing constant $\int w(x)dx$ does not need be computed. Moreover, relying on the upper bound in (16) we can alternatively compute $\widehat{\mathrm{C}}$ and $\widehat{\mathrm{W}}$ without the weights w and still ensure that the above result holds true.

Neural network based learning Theorem 1 guarantees successful estimation of the eigendecomposition of the generator in (4) whenever the energy-based *representation error* $\rho(\mathcal{H})$ in (12) is controlled. It is therefore natural to minimize $\rho(\mathcal{H})$ by choosing an appropriate basis function z_i's. Inspired by the recent work [24], we parameterize them by a neural network, and optimize them to span the leading invariant subspace of the generator.

Let $z^\theta = (z^\theta_i)_{i\in[m]} \colon \mathcal{X} \to \mathbb{R}^m$ be a neural network (NN) embedding parameterized by $\theta \in \Theta$ weights with continuously differentiable activation functions, and let λ^θ_i, $i \in [m]$, be real non-positive

(trainable) weights. We propose to optimize the NN to find the slowest time-scales λ_i^θ that solve the eigenvalue equation $\mathcal{L}z_i^\theta = \lambda_i^\theta z_i^\theta,\ i\in[m]$. Letting $\mathcal{Z}_\theta\colon \mathbb{R}^m \to \mathcal{H}_\pi^\eta(\mathcal{X})$ be the (parameterized) injection operator, given, for every $u \in \mathbb{R}^m$ by $\mathcal{Z}_\theta u = \sum_{i\in[m]} z_i^\theta u_i$, and denoting $\Lambda_\theta^\eta = (\eta I - \operatorname{diag}(\lambda_1^\theta,\dots,\lambda_m^\theta))^{-1}$, the eigenvalue equations for the resolvent then become $(\eta I - \mathcal{L})^{-1}\mathcal{Z}_\theta = \mathcal{Z}_\theta\Lambda_\theta^\eta$. In other words, we aim to find the best rank-m decomposition of resolvent $(\eta I-\mathcal{L})^{-1} \approx \mathcal{Z}_\theta\Lambda_\theta^\eta\mathcal{Z}_\theta^*$. Therefore, for some hyperparameter $\alpha \geq 0$ we introduce the loss

$$\mathcal{E}_\alpha(\theta) := \|(\eta I-\mathcal{L})^{-1} - \mathcal{Z}_\theta\Lambda_\eta^\theta\mathcal{Z}_\theta^*\|^2_{\mathrm{HS}(\mathcal{H}_\pi^\eta)} - \|(\eta I-\mathcal{L})^{-1}\|^2_{\mathrm{HS}(\mathcal{H}_\pi^\eta)} + \alpha \sum_{i,j\in[m]} (\langle z_i^\theta, z_j^\theta\rangle_{L^2_\pi} - \delta_{i,j})^2.$$

While the first term measures the approximation error in the energy space, it cannot be used as a loss, because the action of the resolvent is not known. To mitigate this, the second term is introduced, under the assumption that $(\eta I-\mathcal{L})^{-1} \in \mathrm{HS}\left(\mathcal{H}_\pi^\eta(\mathcal{X})\right)$ (see Appendix C for a discussion). The third term is optional; specifically, if the goal is not only to identify the proper invariant subspace of the generator ($\alpha = 0$), but also to optimize the neural network to extract eigenfunctions as features, then this last term ($\alpha > 0$) encourages the orthonormality of features in $L^2_\pi(\mathcal{X})$, an idea successfully exploited in machine learning and computational chemistry [see e.g. 24, and references therein].

Recalling (14) and denoting by C_θ and W_θ the covariance matrices associated to the parameterized features, after some algebra, we obtain that

$$\mathcal{E}_\alpha(\theta) = \operatorname{tr}\left[\mathrm{C}_\theta\Lambda_\theta^\eta\mathrm{W}_\theta\Lambda_\theta^\eta - 2\mathrm{C}_\theta\Lambda_\theta^\eta + \alpha(\mathrm{C}_\theta - \mathrm{I})^2\right]. \tag{18}$$

In turn, this can be estimated from biased data by two independent samples $\widehat{\pi}'_1$ and $\widehat{\pi}'_2$ as

$$\mathcal{E}_\alpha^{\widehat{\pi}'_1,\widehat{\pi}'_2}(\theta) = \operatorname{tr}\Big[(\widehat{\mathrm{C}}_\theta^1\Lambda_\theta^\eta\widehat{\mathrm{W}}_\theta^2\Lambda_\theta^\eta + \widehat{\mathrm{C}}_\theta^2\Lambda_\theta^\eta\widehat{\mathrm{W}}_\theta^1\Lambda_\theta^\eta)/2 - \widehat{w}^1\widehat{\mathrm{C}}_\theta^2\Lambda_\theta^\eta - \widehat{w}_2\widehat{\mathrm{C}}_\theta^1\Lambda_\theta^\eta + \alpha(\widehat{\mathrm{C}}_\theta^1 - \widehat{w}_1\mathrm{I})(\widehat{\mathrm{C}}_\theta^2 - \widehat{w}_2\mathrm{I})\Big], \tag{19}$$

where $\widehat{\mathrm{C}}_\theta^k$ and $\widehat{\mathrm{W}}_\theta^k$ are the empirical covariances given by (17) for distribution $\widehat{\pi}'_k$, while $\widehat{w}^k = \mathbb{E}_{x'\sim\widehat{\pi}'_k x} w(x')$, $k\in[2]$. Importantly, the computational complexity of the loss (19) is of the order $\mathcal{O}(nm^2d)$, where d is the state dimension and n the sample size, however it can be reduced to $\mathcal{O}(nmd)$ (see Appendix C) allowing its application to learn large dictionaries for high-dimensional problems with big amounts of (biased) data.

The following result, linked to controlling of the representation error as detailed in Theorem 1, provides theoretical guarantees for our approach. The proof and discussion are provided in Appendix B.

Theorem 2. *Given a compact operator $(\eta I-\mathcal{L})^{-1}$, $\eta > 0$, if $(z^\theta)_{i\in[m]} \subseteq \mathcal{H}_\pi^\eta(\mathcal{X})$ for all $\theta\in\Theta$, then*

$$\mathbb{E}\left[\mathcal{E}_\alpha^{\widehat{\pi}'_1,\widehat{\pi}'_2}(\theta)\right] = \overline{w}^2\,\mathcal{E}_\alpha(\theta) \geq -\textstyle\sum_{i\in[m]} \frac{\overline{w}^2}{(\eta-\lambda_i)^2}, \quad \textit{for all } \theta\in\Theta, \tag{20}$$

where $\overline{w} = \mathbb{E}_{x\sim\pi'}[w(x)]$. Moreover, if $\alpha > 0$ and $\lambda_{m+2} < \lambda_{m+1}$, then the equality holds if and only if $(\lambda_i^\theta, z_i^\theta) = (\lambda_i, f_i)$ π-a.e., up to the ordering of indices and choice of eigenfunction signs for $i\in[m]$.

This theorem provides a justification for minimizing the loss in (19), which can be achieved by stochastic optimization algorithms, to obtain an approximation of either the leading invariant subspace of the resolvent $(\eta I-\mathcal{L})^{-1}$ (without orthonormality loss, i.e. $\alpha = 0$), on which the estimator in Theorem 1 can be computed, or even the individual eigenpairs ($\alpha > 0$). A pseudocode of our method is provided below. The main advantage of this method is that it exploits the knowledge of the process. namely, if only the bias V and the diffusion coefficient β are known, recalling (10), the computation of loss relies just of the gradient of the features. On the other hand, the knowledge of the potential can also be exploited via (9). Finally, even if the neural network features are not perfectly learned, one can still resort to Theorem 1 to compute the approximate eigendecomposition of $\mathcal{L}$.

6 Experiments

In this section, we test the method described above on well-established [14, 9, 32, 36] molecular dynamics benchmarks, featuring biased simulations of increasing complexity. We first start by showing the efficiency of our method on a simple one dimensional double well potential. We then proceed to the Muller-Brown potential which is a 2D potential, where this time, sampling is

 https://doi.org/10.52202/079017-2404

Algorithm 1: From biased to unbiased dynamics via infinitesimal generator

1: **Parameters** $\eta > 0$ shift of the generator, m number of wanted eigenfunctions, K number of optimization steps, $\gamma > 0$ and $\alpha > 0$ regression and NN hyperparameters
2: **Inputs** Dataset $\mathcal{D}_n = (x_\ell)_{\ell\in[n]}$ gathered from a simulation with bias potential V
3: Compute weights $w(x_\ell) = \exp(\beta V(x_\ell))$, $\ell \in [n]$, to be used in line 7
4: **if** the dictionary of function z does not already exist **then**
5: **Initialization**: randomly initialize neural networks weights of Λ^θ and $(z_i^\theta)_{i\in[m]}$, set $k = 0$
6: **while** $k < K$ **do**
7: Compute $\widehat{\mathrm{C}}_\theta^j$ and $\widehat{\mathrm{W}}_\theta^j$, $j = 1, 2$, using (17) for two independent batches $\widehat{\pi}_1'$ and $\widehat{\pi}_2'$
8: Compute loss $\widehat{\mathcal{E}}^{\widehat{\pi}_1',\widehat{\pi}_2'}(\theta)$ using (19) and backpropagate
9: **end while**
10: **end if**
11: Compute $\widehat{\mathrm{C}}$ and $\widehat{\mathrm{W}}$ using (17) the datatset $\mathcal{D}_n$
12: Compute the eigenpairs $(\nu_i, v_i)_{i\in[m]}$ of $\widehat{\mathrm{G}}_{\eta,\gamma}=(\widehat{\mathrm{W}}+\eta\gamma\mathrm{I})^{-1}\widehat{\mathrm{C}}$
13: **Output** Estimated eigenpairs of $\mathcal{L}$ are $(\widehat{\lambda}_i, \widehat{f}_i)=(\eta-1/\nu_i, z^\theta(\cdot)^\top v_i)$, $i \in [m]$

accelerated by a bias potential built on the fly. Finally, we study the conformational landscape of alanine dipeptide. This small molecule is a classical testing ground for rare event methods. To showcase the efficiency of our method we analyse two different sets of data both generated in a metadynamics-like approach and showcase the efficiency of our approach, even with a small number of transitions in the training set. The codes used to train the models can be found in the following repository: https://github.com/DevergneTimothee/GenLearn

One dimensional double well potential We first showcase the efficiency of our method on a simple one dimensional toy model. We sample transitions from $U + V$, where U is a double well potential and V is a bias potential. The results are shown Figure 7 in the appendix, where our method clearly outperforms transfer operator approaches and recovers the true underlying dynamics.

Muller Brown potential with metadynamics biasing Muller Brown is a 2 dimensional potential presenting metastable states often studied in the context of enhanced sampling [19, 37, 14, 9]. It presents two minima, with one of them separated into two sub-basins. We thus expect two relevant eigenpairs: the slowest one corresponding to the transition between the two basins and the second slowest one describing the transition between the two sub-basins. However, at low temperature crossing the barrier occurs rarely. To expedite the rate of transition we use metadynamics and instead of having a predefined bias potential, as in the previous section, the bias is built on the fly using metadynamics [25]. The results of the training procedure are presented in Figure 2. We compare the results with deepTICA and a state of the art generator learning approach in [50]. From this figure, we see that we managed to accurately learn the dynamical behavior of the system despite the fact that the dynamics was performed using a bias potential. As expected, it is clearly outperforming transfer operator approaches. We achieve similar or slightly better results (particularly near the transition state) on the qualitative shape of the eigenfunctions. On the other hand, our method performs better than previous work on generator learning on the estimation of eigenvalues, and is the closest to the ground truth eigenvalues. This is likely to be due to the fact that the method in [50] requires the tuning of hyperparameters in the loss function, while in our case, these coefficients are trainable. It should be noted that here, the eigenfunctions were fitted with well-learned features. However, we present in the appendix results where the features are not perfectly learned, but we still manage to recover the eigenfunctions.

Alanine dipeptide with OPES biasing We next treat a more concrete molecular dynamics example with the study of the conformational change of alanine dipeptide in gas phase. It is a molecule containing 22 atoms, of which 10 are heavy. For the remaining of this study, we will only take into account the positions of the heavy atoms, making it a 30 dimensional system. This molecule has widely been used to test methods in enhanced sampling [9, 50, 42]: it presents a conformational change which is a rare event described by the angles ϕ and ψ. In the studies made on this system, the angle ϕ has been shown to be a good CV: the transition between the two states is very well described, and thus a bias potential can easily be built with this CV. On the other hand, the angle ψ misses most of the transition and is a non optimal CV. We generated biased dataset using a variation of

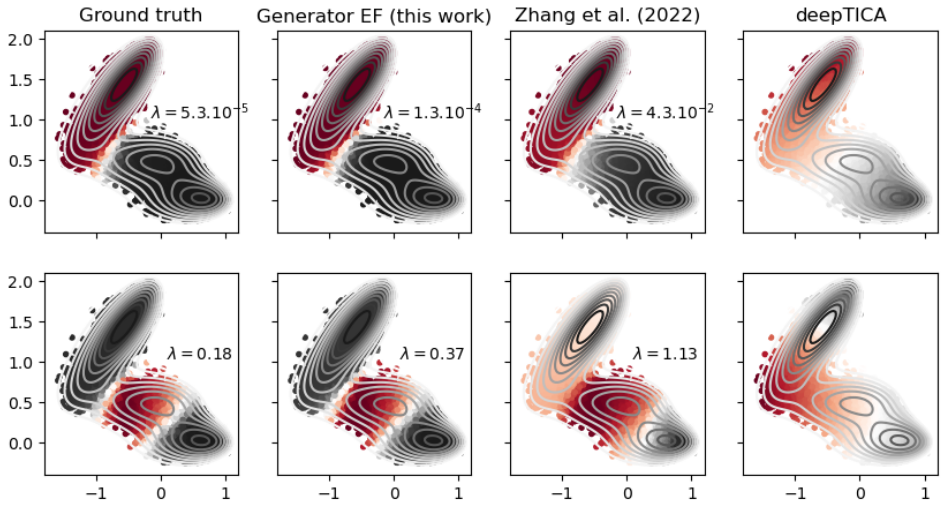

Figure 2: Muller Brown potential. Comparison of the ground truth two first relevant eigenfunctions of the potential (**first column**) with this work (**second column**), transfer operator approach deepTICA [7] (**third column**) and the work of Zhang et al. [50] (**fourth column**). x and y axis are the coordinates of the system and points are colored according to the value of the eigenfunction. The underlying potential is represented by the level lines in white. Associated eigenvalues λ are also reported.

metadynamics called on the fly probability enhanced sampling (OPES) [16], which allows a more extensive and faster exploration of the state space than metadynamics [25]:

Dataset 1: 800ns simulation, biasing on the ψ dihedral angle, with OPES leading to few transitions between the two states. The bias potential was built during the first 100ns of the simulation. For the remaining 700ns, the potential built during the first part was kept fixed to enhance transitions.
Dataset 2: 50ns simulation, biasing on ϕ dihedral angle, with OPES leading to many transitions between the two states. The bias potential was built during the first 20ns of the simulation. For the remaining 30ns, the potential built during the first part was kept to enhanced transitions.

Dataset 1 mimics situations where one has only a basic prior knowledge of the system: only a "suboptimal" CV is used yielding to only a few transitions between the metastable states within the affordable simulation time. In order to ensure translational and rotational invariant vectors, we use Kabsch [18] algorithm, which has been used in previous studies [50, 4, 10] to transform the positions of the atoms. The results are presented in Figure 3. Panels **a)** and **b)** display the first and second eigenfunctions learned by our method respectively. Notice that, even though only 2 transitions are present in dataset 1, the first eigenfunction separates the two metastable states, and the second identifies a faster transition in one metastable state. Panel **c)** showcases the good out-of-sample generalization ability of the method. It visualizes the first eigenfunction obtained as above, but this time visualized on points from dataset 2 and in the plane of dihedral angles ϕ and θ. Interestingly, we discover that a linear relationship is present in the transition region, in agreement with recent findings in the molecular dynamics literature [6, 19].

To further improve the description of the transition and to enhance the training set without any prior knowledge of the mechanism, one could perform biased simulations using the first eigenfunction. Nonetheless, this is not the scope of this paper. To push our method further and see its capabilities when training on a good dataset, we trained it on Dataset 2. One key quantity in molecular dynamics is the committor function for metastable states A and B, which is defined as the probability of, starting from A, going to B before going back to A. Theory tells us that the committor is linearly related to the first eigenfunction of the generator, a result going back to Kolmogorov [see 5, for a discussion]. This relation is exposed in panel **d)** of Figure 3, when comparing to the committor model obtained in [19] indicating the good performance of our method.

Chignolin miniprotein In this section, we report the results of our method obtained on a larger scale experiment: the folding/unfolding mechanism of the chignolin miniprotein. This system has

https://doi.org/10.52202/079017-2404

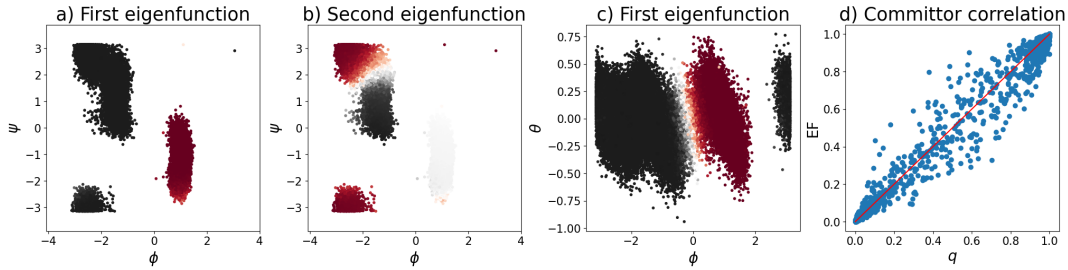

Figure 3: Alanine Dipeptide. Results of our method trained on Dataset 1 **a)** and **b)** first and second eigenfunctions represented on dataset 1, in the plane of the ϕ and ψ dihedral angles. **c)** first eigenfunction represented on dataset 2, in the plane of the ϕ and θ dihedral angles, indicating that our method is effective even when trained from poor CVs (see text for more discussion). On all three panels, points are colored according to the value of the eigenfunction. **d)** Comparison of our method with the committor (q) of [19]

extensively been studied [46, 19, 39, 7]. We first performed a 1 μs biased simulation using the deep-TDA collective variable [46, 37] to gather transitions. Then we chose descriptors as input of the neural networks that are known to describe well the folding process [7]. Finally, we trained the method described in the current work with this trajectory and compared it with the results obtained when training on a 106μs unbiased trajectory provided by D.E. Shaw research [28]. The results are presented in figure 4, showing a very good agreement between the training on an unbiased trajectory and on a biased one.

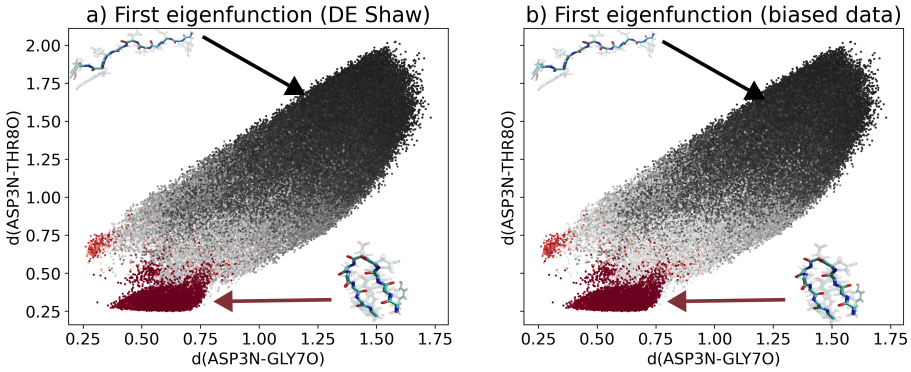

Figure 4: Our method for the chignolin miniprotein. The data points are represented in the plane of the distance between the nitrogen atom of the residue 3: ASP (ASP3N) and the oxygen atom of the residue 7: Gly (Gly7O) and the distance between ASP3N and the oxygen atom of residue 8: THR (THR8) which allow visualizing the folded and unfolded states.

7 Conclusions

We presented a method to learn the eigenfunctions and eigenvalues of the generator of Langevin dynamics from biased simulations, with strong theoretical guarantees. We contrasted this approach with those based on the transfer operator and a recent generator learning approach based on Rayleigh quotients. In experiments, we observed that our approach is effective even when trained from sub-optimal biased simulations. In the future our method could be applied to larger-scale simulations to discover rare events such as protein-ligand binding or catalytic processes. A main limitation of our method is that, in its current form, it is formulated for time-homogeneous bias potentials. However, the proposed framework could be naturally extended to time-dependent biasing, broadening its applicability in computational chemistry. Furthermore, given the quality of our results on alanine dipeptide, in the future, we can use our method to compute accurate eigenfunctions from old, possibly poorly converged, metadynamics simulations, thereby gaining novel and more accurate physical information.

Acknowledgements

This work was partially funded by the European Union - NextGenerationEU initiative and the Italian National Recovery and Resilience Plan (PNRR) from the Ministry of University and Research (MUR), under Project PE0000013 CUP J53C22003010006 "Future Artificial Intelligence Research (FAIR)". We thank D.E Shaw research for providing the chignolin trajectory

References

[1] A Cabannes, V. and Bach, F. (2024). The Galerkin method beats graph-based approaches for spectral algorithms. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 451–459. PMLR.

[2] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25.

[3] Bakry, D., Gentil, I., and Ledoux, M. (2014). *Analysis and Geometry of Markov Diffusion Operators*. Springer.

[4] Belkacemi, Z., Gkeka, P., Lelièvre, T., and Stoltz, G. (2022). Chasing collective variables using autoencoders and biased trajectories. *Journal of Chemical Theory and Computation*, 18(1):59–78.

[5] Berezhkovskii, A. and Szabo, A. (2004). Ensemble of transition states for two-state protein folding from the eigenvectors of rate matrices. *The Journal of Chemical Physics*, 121(18):9186–9187.

[6] Bolhuis, P. G., Dellago, C., and Chandler, D. (2000). Reaction coordinates of biomolecular isomerization. *Proceedings of the National Academy of Sciences*, 97(11):5877–5882.

[7] Bonati, L., Piccini, G., and Parrinello, M. (2021). Deep learning the slow modes for rare events sampling. *Proceedings of the National Academy of Sciences*, 118(44):e2113533118.

[8] Bonati, L., Trizio, E., Rizzi, A., and Parrinello, M. (2023). A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar. *The Journal of Chemical Physics*, 159(1):014801.

[9] Chen, H., Roux, B., and Chipot, C. (2023). Discovering reaction pathways, slow variables, and committor probabilities with machine learning. *Journal of Chemical Theory and Computation*, 19(14):4414–4426.

[10] Chen, W., Tan, A. R., and Ferguson, A. L. (2018). Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *The Journal of Chemical Physics*, 149(7):072312.

[11] Comer, J., Gumbart, J. C., Hénin, J., Lelièvre, T., Pohorille, A., and Chipot, C. (2015). 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.

[12] Dama, J. F., Parrinello, M., and Voth, G. A. (2014). Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.*, 112:240602.

[13] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

[14] France-Lanord, A., Vroylandt, H., Salanne, M., Rotenberg, B., Saitta, A. M., and Pietrucci, F. (2024). Data-driven path collective variables. *Journal of Chemical Theory and Computation*, 20(8):3069–3084.

[15] Hou, B., Sanjari, S., Dahlin, N., Bose, S., and Vaidya, U. (2023). Sparse learning of dynamical systems in RKHS: An operator-theoretic approach. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 13325–13352. PMLR.

 https://doi.org/10.52202/079017-2404

[16] Invernizzi, M. and Parrinello, M. (2020). Rethinking metadynamics: From bias potentials to probability distributions. *The Journal of Physical Chemistry Letters*, 11(7):2731–2736.

[17] Invernizzi, M. and Parrinello, M. (2022). Exploration vs convergence speed in adaptive-bias enhanced sampling. *Journal of Chemical Theory and Computation*, 18(6):3988–3996.

[18] Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923.

[19] Kang, P., Trizio, E., and Parrinello, M. (2024). Computing the committor with the committor to study the transition state ensemble. *Nature Computational Science*, 4(6):451–460.

[20] Klus, S., Nüske, F., Peitz, S., Niemann, J.-H., Clementi, C., and Schütte, C. (2020). Data-driven approximation of the koopman generator: Model reduction, system identification, and control. *Physica D: Nonlinear Phenomena*, 406:132416.

[21] Kostic, V., Lounici, K., Halconruy, H., Devergne, T., and Pontil, M. (2024a). Learning the infinitesimal generator of stochastic diffusion processes. *Preprint, arXiv:2405.12940*.

[22] Kostic, V., Novelli, P., Maurer, A., Ciliberto, C., Rosasco, L., and Pontil, M. (2022). Learning dynamical systems via Koopman operator regression in reproducing kernel Hilbert spaces. In *Advances in Neural Information Processing Systems*.

[23] Kostic, V. R., Lounici, K., Novelli, P., and Pontil, M. (2023). Sharp spectral rates for koopman operator learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[24] Kostic, V. R., Novelli, P., Grazzi, R., Lounici, K., and Pontil, M. (2024b). Deep projection networks for learning time-homogeneous dynamical systems. In *The Twelfth International Conference on Learning Representations*.

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

[26] Leitold, C., Mundy, C. J., Baer, M. D., Schenter, G. K., and Peters, B. (2020). Solvent reaction coordinate for an sn2 reaction. *Journal of Chemical Physics*, 153(2).

[27] Lelièvre, T., Pigeon, T., Stoltz, G., and Zhang, W. (2024). Analyzing multimodal probability measures with autoencoders. *The Journal of Physical Chemistry B*, 128(11):2607–2631.

[28] Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science*, 334(6055):517–520.

[29] Magrino, T., Huet, L., Saitta, A. M., and Pietrucci, F. (2022). Critical assessment of data-driven versus heuristic reaction coordinates in solution chemistry. *The Journal of Physical Chemistry A*, 126(47):8887–8900.

[30] Mardt, A., Pasquali, L., Wu, H., and Noé, F. (2018). VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1).

[31] McCarty, J. and Parrinello, M. (2017). A variational conformational dynamics approach to the selection of collective variables in metadynamics. *The Journal of Chemical Physics*, 147(20):204109.

[32] Mori, Y., Okazaki, K.-i., Mori, T., Kim, K., and Matubayasi, N. (2020). Learning reaction coordinates via cross-entropy minimization: Application to alanine dipeptide. *The Journal of Chemical Physics*, 153(5):054115.

[33] Oksendal, B. (2013). *Stochastic Differential Equations: an Introduction with Applications*. Springer Science & Business Media.

[34] Pietrucci, F. (2017). Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Reviews in Physics*, 2:32–45.

[35] Pillaud-Vivien, L. and Bach, F. (2023). Kernelized diffusion maps. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5236–5259. PMLR.

[36] Plainer, M., Stark, H., Bunne, C., and Günnemann, S. (2023). Transition path sampling with Boltzmann generator-based MCMC moves. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*.

[37] Ray, D., Trizio, E., and Parrinello, M. (2023). Deep learning collective variables from transition path ensemble. *The Journal of Chemical Physics*, 158(20):204102.

[38] Reed, M. and Simon, B. (1980). *Functional Analysis (Modern Mathematical Physics, Volume 1)*. Academic Press.

[39] Rydzewski, J., Chen, M., Ghosh, T. K., and Valsson, O. (2022). Reweighted manifold learning of collective variables from enhanced sampling simulations. *Journal of Chemical Theory and Computation*, 18(12):7179–7192. PMID: 36367826.

[40] Salomon-Ferrer, R., Case, D. A., and Walker, R. C. (2013). An overview of the amber biomolecular simulation package. *WIREs Computational Molecular Science*, 3(2):198–210.

[41] Schwantes, C. R. and Pande, V. S. (2015). Modeling molecular kinetics with tica and the kernel trick. *Journal of Chemical Theory and Computation*, 11(2):600–608.

[42] Shmilovich, K. and Ferguson, A. L. (2023). Girsanov reweighting enhanced sampling technique (grest): On-the-fly data-driven discovery of and enhanced sampling in slow collective variables. *The Journal of Physical Chemistry A*, 127(15):3497–3517.

[43] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer New York.

[44] Torrie, G. and Valleau, J. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199.

[45] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2014). Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613.

[46] Trizio, E. and Parrinello, M. (2021). From enhanced sampling to reaction profiles. *The Journal of Physical Chemistry Letters*, 12(35):8621–8626. PMID: 34469175.

[47] Tuckerman, M. E. (2023). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford university press.

[48] Vanden-Eijnden, E. (2006). *Transition Path Theory*, pages 453–493. Springer Berlin Heidelberg, Berlin, Heidelberg.

[49] Yang, Y. I. and Parrinello, M. (2018). Refining collective coordinates and improving free energy representation in variational enhanced sampling. *Journal of Chemical Theory and Computation*, 14(6):2889–2894.

[50] Zhang, W., Li, T., and Schütte, C. (2022). Solving eigenvalue pdes of metastable diffusion processes using artificial neural networks. *Journal of Computational Physics*, 465:111377.

 https://doi.org/10.52202/079017-2404

Appendix

The appendix contains additional background on stochastic differential equations, proofs of the results omitted in the main body, and more information about our learning method and the numerical experiments.

notation	meaning	notation	meaning
$[\cdot]$	set $\{1, 2 \ldots, \cdot\}$	$[\cdot]_+$	nonnegative part of a number
$\mathcal{X}$	state space of the Markov process	$(X_t)_{t\geq 0}$	time-homogeneous Markov process
dW_t	Brownian motion	β	inverse temperature of the system
U	potential energy of the system	V	bias potential
π	Boltzmann distribution of potential U	π'	Boltzmann distribution of potential $U+V$
$\widehat{\pi}$	empirical version of π	$\widehat{\pi}'$	empirical version of π'
$d\pi/d\pi'$	density of π w.r.t. π'	w	exponential weights of $d\pi/d\pi'$
$L^2_\pi(\mathcal{X})$	L^2 space on $\mathcal{X}$ w.r.t. the measure π	$H^{1,2}_\pi(\mathcal{X})$	Sobolev space w.r.t. π on $\mathcal{X}$
$\mathcal{T}_t$	transfer operator with lag time t	$\hat{T}_t$	empirical estimator of $\mathcal{T}_t$
$\mathcal{L}$	generator of the true process	$\mathcal{L}'$	generator of the biased process
η	shift parameter	$(\eta I - \mathcal{L})^{-1}$	resolvent at η
$\mathcal{H}$	hypothetical Hilbert space	z	basis functions
$\mathfrak{E}^\eta[\cdot,\cdot]$	regularized energy kernel	$\mathfrak{E}^\eta[\cdot]$	regularized energy norm
$\mathcal{H}^\eta_\pi(\mathcal{X})$	space associated to the energy norm	$\mathcal{Z}$	injection operator
$\mathcal{R}_\partial$	risk functional	$\widehat{\mathcal{R}}_\partial$	empirical risk functional
C	covariance	$\widehat{\mathrm{C}}$	empirical covariance
W	covariance in energy space	$\widehat{\mathrm{W}}$	empirical covariance in energy space
λ_i	generator eigenvalue	f_i	generator eigenfunction
$\widehat{\lambda}_i$	empirical estimator of λ_i	$\widehat{f_i}$	empirical estimator of f_i
z^θ	neural network embedding	θ	neural network weights
$\mathcal{Z}_\theta$	neural network injection operator	Λ^η_θ	neural network weights
$\mathcal{E}_\alpha$	regularized loss function	$\widehat{\mathcal{E}}_\alpha$	regularized empirical loss function

Table 1: Summary of used notations.

A Background

In this work we consider stochastic differential equations (SDE) of the form

$$dX_t = a(X_t)dt + b(X_t)dW_t \quad \text{and} \quad X_0 = x. \tag{21}$$

The special case of Langevin equation considered in the main body of the paper corresponds to $a(x) = -\nabla U(x)$ and $b(x) = \sqrt{\frac{2}{\beta}}I$. Equation (21) describes the dynamics of the random vector X_t in the state space $\mathcal{X} \subseteq \mathbb{R}^d$, governed by the *drift* $a : \mathbb{R}^d \to \mathbb{R}^d$ and the *diffusion* $b : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ coefficients, where W_t is a $\mathbb{R}^p$-dimensional standard Brownian motion. Under the usual conditions [see e.g. 33] that a and b are globally Lispchitz and sub-linear, the SDE (21) admits an unique strong solution $X = (X_t)_{\geqslant 0}$ that is a Markov process to which we can associate the semigroup of Markov *transfer operators* $(\mathcal{T}_t)_{t\geq 0}$ defined, for every $t \geq 0$, as

$$[\mathcal{T}_t f](x) := \mathbb{E}[f(X_t)|X_0 = x], \quad x \in \mathcal{X},\ f\colon \mathcal{X} \to \mathbb{R}. \tag{22}$$

For stable processes, the distribution of X_t converges to an *invariant measure* π on $\mathcal{X}$, such that $X_0 \sim \pi$ implies that $X_t \sim \pi$ for all $t \geq 0$. In such cases, one can define the semigroup on $L^2_\pi(\mathcal{X})$, and characterize the process by the *infinitesimal generator* of the semi-group $(\mathcal{T}_t)_{t\geqslant 0}$,

$$\mathcal{L} := \lim_{t\to 0^+} \frac{\mathcal{T}_t - I}{t} \tag{23}$$

defined on the Sobolev space $H^{1,2}_\pi(\mathcal{X})$ of functions in $L^2_\pi(\mathcal{X})$ whose gradient are in $L^2_\pi(\mathcal{X})$, too, i.e. $\mathcal{L}\colon L^2_\pi(\mathcal{X}) \to L^2_\pi(\mathcal{X})$ and $\mathrm{dom}(\mathcal{L}) = H^{1,2}_\pi(\mathcal{X})$. The transfer operator and the generator are linked to each other by the formula $\mathcal{T}_t = \exp(t\mathcal{L})$.

After defining the infinitesimal generator for Markov processes by (23), we provide its explicit form for solution processes of equations like (21). Given a smooth function $f \in \mathcal{C}^2(\mathcal{X}, \mathbb{R})$, Itô's formula [see for instance 3, p. 495] provides for $t \in \mathbb{R}_+$,

$$\begin{aligned} f(X_t) - f(X_0) &= \int_0^t \sum_{i=1}^d \partial_i f(X_s) dX^i_s + \tfrac{1}{2} \int_0^t \sum_{i,j=1}^d \partial^2_{ij} f(X_s) d\langle X^i, X^j\rangle_s \\ &= \int_0^t \nabla f(X_s)^\intercal dX_s + \tfrac{1}{2} \int_0^t \mathrm{Tr}\big[X_s^\intercal (\nabla^2 f)(X_s) X_s\big] ds. \end{aligned}$$

Recalling (21), we get

$$\begin{aligned} f(X_t) = f(X_0) &+ \int_0^t \Big[\nabla f(X_s)^\intercal a(X_s) + \tfrac{1}{2}\mathrm{Tr}\big[b(X_s)^\intercal (\nabla^2 f(X_s)) b(X_s)\big]\Big] ds \\ &+ \int_0^t \nabla f(X_s)^\intercal b(X_s) dW_s. \end{aligned} \tag{24}$$

Provided f and b are smooth enough, the expectation of the last stochastic integral vanishes so that we get

$$\mathbb{E}[f(X_t)|X_0 = x] = f(x) + \int_0^t \mathbb{E}\Big[\nabla f(X_s)^\intercal a(X_s) + \tfrac{1}{2}\mathrm{Tr}\big[b(X_s)^\intercal (\nabla^2 f(X_s)) b(X_s)\big]\Big| X_0 = x\Big] ds$$

Recalling that $\mathcal{L} = \lim_{t\to 0^+} (\mathcal{T}_t f - f)/t$, we get for every $x \in \mathcal{X}$,

$$\begin{aligned} \mathcal{L}f(x) &= \lim_{t\to 0} \frac{\mathbb{E}[f(X_t)|X_0 = x] - f(x)}{t} \\ &= \lim_{t\to 0} \frac{1}{t}\Big[\int_0^t \mathbb{E}\Big[\nabla f(X_s)^\intercal a(X_s) + \tfrac{1}{2}\mathrm{Tr}\big[(X_s)^\intercal (\nabla^2 f(X_s)) b(X_s)\big]\Big] ds \Big| X_0 = x\Big] \\ &= \nabla f(x)^\intercal a(x) + \tfrac{1}{2}\mathrm{Tr}\big[b(x)^\intercal (\nabla^2 f(x)) b(x)\big], \end{aligned} \tag{25}$$

which provides the closed formula for the IG associated with the solution process of (21). In particular, for Langevin dynamics this reduces to (3).

Next, recalling that for a bounded linear operator A on some Hilbert space $\mathcal{H}$ the *resolvent set* of the operator A is defined as $\rho(A) = \{\lambda \in \mathbb{C} \mid A - \lambda I \text{ is bijective}\}$, and its *spectrum* $\mathrm{Sp}(A) = \mathbb{C}\backslash\{\rho(A)\}$, let $\lambda \subseteq \mathrm{Sp}(A)$ be the isolated part of the spectra, i.e. both λ and $\mu = \mathrm{Sp}(A) \setminus \lambda$ are closed in $\mathrm{Sp}(A)$. Then, the *Riesz spectral projector* $P_\lambda\colon \mathcal{H} \to \mathcal{H}$ is defined by

$$P_\lambda = \frac{1}{2\pi} \int_\Gamma (zI - A)^{-1} dz, \tag{26}$$

where Γ is any contour in the resolvent set $\mathrm{Res}(A)$ with λ in its interior and separating λ from μ. Indeed, we have that $P^2_\lambda = P_\lambda$ and $\mathcal{H} = \mathrm{Im}(P_\lambda) \oplus \mathrm{Ker}(P_\lambda)$ where $\mathrm{Im}(P_\lambda)$ and $\mathrm{Ker}(P_\lambda)$ are both invariant under A, and we have $\mathrm{Sp}(A_{|_{\mathrm{Im}(P_\lambda)}}) = \lambda$ and $\mathrm{Sp}(A_{|_{\mathrm{Ker}(P_\lambda)}}) = \mu$. Moreover, $P_\lambda + P_\mu = I$ and $P_\lambda P_\mu = P_\mu P_\lambda = 0$.

Finally if A is a *compact* operator, then the Riesz-Schauder theorem [see e.g. 38] assures that $\mathrm{Sp}(T)$ is a discrete set having no limit points except possibly $\lambda = 0$. Moreover, for any nonzero $\lambda \in \mathrm{Sp}(T)$, then λ is an *eigenvalue* (i.e. it belongs to the point spectrum) of finite multiplicity, and, hence, we can deduce the spectral decomposition in the form

$$A = \sum_{\lambda \in \mathrm{Sp}(A)} \lambda\, P_\lambda, \tag{27}$$

where the geometric multiplicity of λ, $r_\lambda = \mathrm{rank}(P_\lambda)$, is bounded by the algebraic multiplicity of λ. If additionally A is a normal operator, i.e. $AA^* = A^*A$, then $P_\lambda = P^*_\lambda$ is an orthogonal projector for each $\lambda \in \mathrm{Sp}(A)$ and $P_\lambda = \sum_{i=1}^{r_\lambda} \psi_i \otimes \psi_i$, where ψ_i are normalized eigenfunctions of A corresponding to λ and r_λ is both algebraic and geometric multiplicity of λ.

We conclude this section by stating the well-known Davis-Kahan perturbation bound for eigenfunctions of self-adjoint compact operators.

https://doi.org/10.52202/079017-2404

Proposition 1 ([13]). *Let A be compact self-adjoint operator on a separable Hilbert space $\mathcal{H}$. Given a pair $(\widehat{\mu}, \widehat{f}) \in \mathbb{C} \times \mathcal{H}$ such that $\|\widehat{f}\| = 1$, let λ be the eigenvalue of A that is closest to $\widehat{\mu}$ and let f be its normalized eigenfunction. If $\widehat{g} = \min\{|\widehat{\mu} - \lambda| \,|\, \lambda \in \mathrm{Sp}(A) \setminus \{\lambda\}\} > 0$, then $\sin(\sphericalangle(\widehat{f}, f)) \leq \|A\widehat{f} - \widehat{\mu}\widehat{f}\|/\widehat{g}$.*

B Unbiased generator regression

In this section, we prove Theorem 1 relying on recently developed statistical theory of generator learning [21]. To that end, let $\phi(x) := z(\cdot)^\intercal z(x) \in \mathcal{H}$ be a feature map of the RKHS space $\mathcal{H}$ of dimension $\dim(\mathcal{H}) = m$. Let $\mathrm{W}u_j = \sigma_j^2 u_j$ be the eigenvalue decomposition of W and let $v_j := u_j/\sigma_j$. This induces the SVD of the injection operator, $\mathcal{Z}u_j = \sigma_j \tilde{z}_j$ for $\tilde{z}_j := z(\cdot)^\intercal v_j$.

Since $\mathcal{H} \subseteq H^{1,\infty}_\pi(\mathcal{X})$, we have that

$$c_\tau = \operatorname*{ess\,sup}_{x\sim\pi} \textstyle\sum_{j\in\mathbb{N}} [\eta|\tilde{z}_j(x)|^2 - z_j(x)[\mathcal{L}\tilde{z}_j](x)] < +\infty,$$

and, denoting $\mathrm{W}_\gamma := \mathrm{W} + \eta\gamma\mathrm{I}$

$$\mathrm{tr}[\mathrm{W}_\gamma^{-1}\mathrm{W}_\gamma] \leq m.$$

In addition denote the empirical version of W_γ as $\widehat{\mathrm{W}}_\gamma := \widehat{\mathrm{W}} + \eta\gamma\mathrm{I}$.

Now, we can apply the following propositions from [21] to our setting, recalling the notation for normalizing constant $\overline{w} := \mathbb{E}_{x\sim\pi'}[w(x)]$ for which $w(\cdot)/\overline{w} = d\pi/d\pi'$.

Proposition 2. *Given $\delta > 0$, with probability in the i.i.d. draw of $(x_i)_{i=1}^n$ from π, it holds that*

$$\mathbb{P}\{\|\widehat{\mathrm{W}} - \mathrm{W}\| \leq \varepsilon_n(\delta)\} \geq 1 - \delta,$$

where

$$\varepsilon_n(\delta) = \frac{2\|\mathrm{W}\|}{3n}\mathcal{L}(\delta) + \sqrt{\frac{2\|\mathrm{W}\|}{n}\mathcal{L}(\delta)} \quad \textit{and} \quad \mathcal{L}(\delta) = \log\frac{4\,\mathrm{tr}(\mathrm{W})}{\delta\,\|\mathrm{W}\|}. \tag{28}$$

Proposition 3. *Given $\delta > 0$, with probability in the i.i.d. draw of $(x_i)_{i=1}^n$ from π, it holds that*

$$\mathbb{P}\left\{\|\mathrm{W}_\gamma^{-1/2}(\widehat{\mathrm{W}} - \mathrm{W})\mathrm{W}_\gamma^{-1/2}\| \leq \varepsilon_n^1(\gamma,\delta)\right\} \geq 1 - \delta, \tag{29}$$

where

$$\varepsilon_n^1(\gamma,\delta) = \frac{2c_\tau}{3n}\mathcal{L}^1(\gamma,\delta) + \sqrt{\frac{2\,c_\tau}{n}\mathcal{L}^1(\gamma,\delta)}, \tag{30}$$

and

$$\mathcal{L}^1(\gamma,\delta) = \ln\frac{4}{\delta} + \ln\frac{\mathrm{tr}(\mathrm{W}_\gamma^{-1}\mathrm{W})}{\|\mathrm{W}_\gamma^{-1}\mathrm{W}\|}.$$

Moreover,

$$\mathbb{P}\left\{\|\mathrm{W}_\gamma^{1/2}\widehat{\mathrm{W}}_\gamma^{-1}\mathrm{W}_\gamma^{1/2}\| \leq \frac{1}{1-\varepsilon_n^1(\gamma,\delta)}\right\} \geq 1 - \delta. \tag{31}$$

Proposition 4. *With probability in the i.i.d. draw of $(x_i)_{i=1}^n$ from π, it holds*

$$\mathbb{P}\left\{\|\mathrm{W}_\gamma^{-1/2}(\widehat{\mathrm{C}} - \mathrm{C})\|_F \leq \varepsilon_n^2(\gamma,\delta)\right\} \geq 1 - \delta,$$

where

$$\varepsilon_n^2(\gamma,\delta) = \frac{4\sqrt{2\,m\|\mathrm{W}\|}}{\eta}\,\ln\frac{2}{\delta}\sqrt{\frac{c_\beta}{n} + \frac{c_\tau}{n^2}}. \tag{32}$$

We are now ready to prove Theorem 1, which we restate here for convenience.

Theorem 1. *Let $\mathcal{D}_n = (x'_i)_{i\in[n]}$ be the biased dataset generated from π'. Let $w(x) = e^{\beta V(x)}$ and define the empirical covariances w.r.t. the empirical distribution $\widehat{\pi}' = n^{-1}\sum_{i\in[n]}\delta_{x'_i}$ by*

$$\widehat{\mathrm{C}} = \big(\mathbb{E}_{x'\sim\widehat{\pi}'}[w(x')z_i(x')z_j(x')]\big)_{i,j\in[m]} \quad \textit{and} \quad \widehat{\mathrm{W}} = \big(\mathfrak{E}^\eta_{\widehat{\pi}'}[\sqrt{w}z_i, \sqrt{w}z_j]\big)_{i,j\in[m]}. \tag{17}$$

Compute the eigenpairs $(\nu_i, v_i)_{i\in[m]}$ of the RR estimator $\widehat{\mathrm{G}}_{\eta,\gamma} = (\widehat{\mathrm{W}} + \eta\gamma \mathrm{I})^{-1}\widehat{\mathrm{C}}$, and estimate the eigenpairs in (4) as $(\widehat{\lambda}_i, \widehat{f}_i) = (\eta - 1/\nu_i, z(\cdot)^\intercal v_i)$. If the elements of $\mathcal{H}$ and their gradients are essentially bounded, and $\lim_{m\to\infty} \rho(\mathcal{H}) = 0$, then for every $\varepsilon > 0$, there exist $(m, n, \gamma) \in \mathbb{N}\times\mathbb{N}\times\mathbb{R}_+$, such that, for every $i \in [m]$, $|\lambda_i - \widehat{\lambda}_i| \leq \varepsilon$ and $\sin_{L^2_\pi}(\sphericalangle(f_i, \widehat{f}_i)) \leq \varepsilon$, with high probability.

Proof. We first show that $\mathcal{R}_\partial(\widehat{\mathrm{G}}_{\eta,\gamma}) < \varepsilon$ for big enough $m, n \in \mathbb{N}$ and small enough $\gamma > 0$.

Observe that

$$\mathrm{W} = \mathbb{E}[\widehat{\mathrm{W}}]/\overline{w} \quad \text{and} \quad \mathrm{C} = \mathbb{E}[\widehat{\mathrm{C}}]/\overline{w} \tag{33}$$

and, hence

$$\mathrm{G}_{\eta,\gamma} := \mathrm{W}_\gamma^{-1}\mathrm{C} = (\mathbb{E}[\widehat{\mathrm{W}}_\gamma])^{-1}(\mathbb{E}[\widehat{\mathrm{C}}]),$$

due to cancellation of $\overline{w}$.

Given $\varepsilon > 0$, let $m \in \mathbb{N}$ be such that $\rho(\mathcal{H}) = \|(I - P_\mathcal{H})(\eta I - \mathcal{L})^{-1}\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}^\eta_\pi)} < \varepsilon/3$. Next, since

$$P_\mathcal{H}(\eta I - \mathcal{L})^{-1}\mathcal{Z} - \mathcal{Z}\mathrm{G}_{\eta,\gamma} = (I - \mathcal{Z}\mathrm{W}_\gamma^{-1}\mathcal{Z}^*)(\eta I - \mathcal{L})^{-1}\mathcal{Z} = \mathcal{Z}(\mathrm{W}^\dagger\mathrm{C} - \mathrm{W}_\gamma^{-1}\mathrm{C}),$$

we have that

$$\|P_\mathcal{H}(\eta I - \mathcal{L})^{-1}\mathcal{Z} - \mathcal{Z}\mathrm{G}_{\eta,\gamma}\|_{\mathrm{HS}(\mathcal{H},\mathcal{H}^\eta_\pi)} = \|\mathrm{W}^{1/2}(\mathrm{W}^\dagger\mathrm{C} - \mathrm{W}_\gamma^{-1}\mathrm{C})\|_F = \|\mathrm{W}^{1/2}(Wx^\dagger - \mathrm{W}_\gamma^{-1})\mathrm{C}\|_F \to 0,$$

as $\gamma \to 0$. Hence, let $\gamma > 0$ be such that $\|P_\mathcal{H}(\eta I - \mathcal{L})^{-1}\mathcal{Z} - \mathcal{Z}\mathrm{G}_{\eta,\gamma}\|_{\mathcal{H}\to\mathcal{H}^\eta_\pi} < \varepsilon/3$.

Finally, using the decomposition of the risk

$$\mathcal{R}_\partial(\widehat{\mathrm{G}}_{\eta,\gamma}) \leq \rho(\mathcal{H}) + \|P_\mathcal{H}(\eta I - \mathcal{L})^{-1}\mathcal{Z} - \mathcal{Z}\mathrm{G}_{\eta,\gamma}\|_{\mathrm{HS}(\mathbb{R}^m,\mathcal{H}^\eta_\pi)} + \|\mathcal{Z}(\widehat{\mathrm{G}}_{\eta,\gamma} - \widehat{\mathrm{G}}_{\eta,\gamma})\|_{\mathrm{HS}(\mathbb{R}^m,\mathcal{H}^\eta_\pi)}$$

it remains to show that for large enough n we have $\|\mathcal{Z}(\mathrm{G}_{\eta,\gamma} - \widehat{\mathrm{G}}_{\eta,\gamma})\|_{\mathrm{HS}(\mathbb{R}^m,\mathcal{H}^\eta_\pi)} \leq \varepsilon/3$.

To that end observe that

$$\begin{aligned}\mathrm{W}_\gamma^{1/2}(\widehat{\mathrm{G}}_{\eta,\gamma} - \mathrm{G}_{\eta,\gamma}) &= \mathrm{W}_\gamma^{1/2}\widehat{\mathrm{W}}_\gamma^{-1}(\widehat{\mathrm{C}} - \widehat{\mathrm{W}}_\gamma\mathrm{W}_\gamma^{-1}\mathrm{C} \pm \mathrm{C}) \\ &= \mathrm{W}_\gamma^{1/2}\widehat{\mathrm{W}}_\gamma^{-1}\mathrm{W}_\gamma^{1/2}\left(\mathrm{W}_\gamma^{-1/2}(\widehat{\mathrm{C}} - \mathrm{C}) - \mathrm{W}_\gamma^{-1/2}(\widehat{\mathrm{W}} - \mathrm{W})\mathrm{W}_\gamma^{-1/2}(\mathrm{W}_\gamma^{-1/2}\mathrm{C})\right).\end{aligned}$$

Thus, by multiplying the above expression by $\overline{w}$ and applying Propositions 3 and 4, we obtain that there exists $n \in \mathbb{N}$ such that $\|\mathcal{Z}(\mathrm{G}_{\eta,\gamma} - \widehat{\mathrm{G}}_{\eta,\gamma})\|_{\mathrm{HS}(\mathbb{R}^m,\mathcal{H}^\eta_\pi)} \leq \varepsilon/3$.

Next, assuming that $\|\widehat{\mathrm{W}} - \mathrm{W}\|$ is small, for the normalization of the estimated eigenfunctions we have that

$$\frac{\|v_j\|_2^2}{\|\widehat{f}_j\|^2_{\mathcal{H}^\eta_\pi}} = \frac{v_j^\intercal v_j}{v_j^\intercal \mathrm{W} v_j} \leq \frac{v_j^\intercal v_j}{v_j^\intercal \widehat{\mathrm{W}} v_j - v_j^\intercal(\mathrm{W} - \widehat{\mathrm{W}})v_j} \leq \frac{1}{\lambda^+_{\min}(\widehat{\mathrm{W}}) - \|\widehat{\mathrm{W}} - \mathrm{W}\|} \leq \frac{1}{\lambda_m(\mathrm{W}) - 2\|\widehat{\mathrm{W}} - \mathrm{W}\|}.$$

where we have that $\lambda_m(\mathrm{W}) > 0$ due to fact that (z_j) are linearly independent.

Therefore, to conclude the proof, we apply [21, Proposition 2] which directly relying on Proposition 1 yields the result. □

At last we remark, based on the observation that $\mathrm{W} = \mathbb{E}[\widehat{\mathrm{W}}]/\overline{w}$ and $\mathrm{C} = \mathbb{E}[\widehat{\mathrm{C}}]/\overline{w}$, one can readily obtain stronger version of Theorem 1 in the general RKHS setting of [21].

C Unbiased deep learning of spectral features

In this section, we provide details on our DNN method and prove Theorem 2. To that end, let us denote the terms in the loss as

$$\mathcal{E}_\gamma(\theta) := \underbrace{\|(\eta I - \mathcal{L})^{-1} - \mathcal{Z}_\theta\Lambda^\theta_\eta\mathcal{Z}^*_\theta\|^2_{\mathrm{HS}(\mathcal{H}^\eta_\pi)} - \|(\eta I - \mathcal{L})^{-1}\|^2_{\mathrm{HS}(\mathcal{H}^\eta_\pi)}}_{\text{expected loss } \mathcal{E}} + \gamma\underbrace{\textstyle\sum_{i,j\in[m]}(\langle z^\theta_i, z^\theta_j\rangle_{L^2_\pi} - \delta_{i,j})^2}_{\text{orthonormality loss } \mathcal{E}_{\mathrm{on}}}, \tag{34}$$

https://doi.org/10.52202/079017-2404

and recall that the injection operator is $\mathcal{Z}_\theta{=}(z^\theta)^\intercal(\cdot)\colon \mathbb{R}^m \to \mathcal{H}^\eta_\pi(\mathcal{X})$. It is easy to show from the definition of the adjoint that, for every $f \in \mathcal{H}^\eta_\pi(\mathcal{X})$, we have

$$\mathcal{Z}^*_\theta f = \mathbb{E}_{x\sim\pi}[z^\theta(x)((\eta I{-}\mathcal{L})f)(x)]. \tag{35}$$

Thus, $\mathcal{Z}^*_\theta\mathcal{Z}_\theta = \mathfrak{E}^\eta[z^\theta_i, z^\theta_j]_{i,j\in[m]}$ which we denote by W_θ, while $\mathcal{Z}^*_\theta(\eta I - \mathcal{L})\mathcal{Z}_\theta = \mathbb{E}_{x\sim\pi}[z^\theta_i(x)z^\theta_j(x)]_{i,j\in[m]}$ is denoted by C_θ.

Theorem 2. *Given a compact operator $(\eta I{-}\mathcal{L})^{-1}$, $\eta > 0$, if $(z^\theta)_{i\in[m]} \subseteq \mathcal{H}^\eta_\pi(\mathcal{X})$ for all $\theta{\in}\Theta$, then*

$$\mathbb{E}\big[\mathcal{E}^{\widehat{\pi}'_1,\widehat{\pi}'_2}_\alpha(\theta)\big] = \overline{w}^2\,\mathcal{E}_\alpha(\theta) \geq -\textstyle\sum_{i\in[m]}\frac{\overline{w}^2}{(\eta-\lambda_i)^2}, \quad \textit{for all } \theta \in \Theta, \tag{20}$$

where $\overline{w} = \mathbb{E}_{x\sim\pi'}[w(x)]$. Moreover, if $\alpha > 0$ and $\lambda_{m+2} < \lambda_{m+1}$, then the equality holds if and only if $(\lambda^\theta_i, z^\theta_i) = (\lambda_i, f_i)$ π-a.e., up to the ordering of indices and choice of eigenfunction signs for $i \in [m]$.

Proof. Let $P_k\colon \mathcal{H}^\eta_\pi(\mathcal{X}) \to \mathcal{H}^\eta_\pi(\mathcal{X})$ be spectral projector of $\mathcal{L}$ corresponding to the k largest eigenvalues. Now, consider

$$\mathcal{E}^k(\theta){=}\|P_k(\eta I{-}\mathcal{L})^{-1}{-}\mathcal{Z}_\theta\Lambda^\theta_\eta\mathcal{Z}^*_\theta\|^2_{\mathrm{HS}(\mathcal{H}^\eta_\pi)}{-}\|P_k(\eta I{-}\mathcal{L})^{-1}\|^2_{\mathrm{HS}(\mathcal{H}^\eta_\pi)}.$$

Due to Eckhart-Young theorem, we have that for every $\theta \in \Theta$, the best rank-m approximation of $(\eta I{-}\mathcal{L})^{-1}$ is $(\eta I{-}\mathcal{L})^{-1}P_m = (\eta I{-}\mathcal{L})^{-1}P_kP_m$, for $k > m$, and it holds

$$\mathcal{E}^k(\theta){\geq} \sum_{j=m+1}^{k}(\eta-\lambda_i)^{-1} - \sum_{j=1}^{k}(\eta-\lambda_i)^{-1} = -\sum_{j=1}^{m}(\eta-\lambda_i)^{-1}.$$

As before, expanding in $\mathcal{E}^k$ the HS norm via the trace, we obtain

$$\mathcal{E}^k(\theta){=}\|\mathcal{Z}_\theta\Lambda^\theta_\eta\mathcal{Z}^*_\theta\|^2_{\mathrm{HS}(\mathcal{H}^\eta_\pi)}{-}2\,\mathrm{tr}[\mathcal{Z}_\theta\Lambda^\theta_\eta\mathcal{Z}^*_\theta(\eta I{-}\mathcal{L})^{-1}P_k]{=}\mathcal{E}(\theta){+}2\,\mathrm{tr}[\mathcal{Z}_\theta\Lambda^\theta_\eta\mathcal{Z}^*_\theta(\eta I{-}\mathcal{L})^{-1}(I{-}P_k)],$$

and, hence, by Cauchy-Schwartz inequality, we have that

$$|\mathcal{E}(\theta) - \mathcal{E}^k(\theta)| \leq \|\mathcal{Z}_\theta\|^2_{\mathrm{HS}(\mathcal{H},\mathcal{H}^\eta_\pi)}\|\Lambda^\theta_\eta\|\|(\eta I{-}\mathcal{L})^{-1}(I{-}P_k)\| = \tfrac{1}{\eta}\sum_{i\in[m]}\mathfrak{E}^\eta_\pi[z^\theta_i](\eta{-}\lambda_{k+1})^{-1}.$$

Observing that $z^\theta_i \in \mathcal{H}^\eta_\pi(\mathcal{X})$, i.e. $\mathfrak{E}^\eta_\pi[z^\theta_i] < \infty$, we conclude that, for every $\theta \in \Theta$, $\lim_{k\to\infty}\mathcal{E}^k(\theta) = \mathcal{E}(\theta)$. Therefore, noting that $\mathcal{E}_\alpha(\theta) \geq \mathcal{E}(\theta)$, inequality in (20) is proven. Since the equality clearly holds for the leading eigenpairs of the generator, to prove the reverse, it suffices to recall the uniqueness result of the best rank-m estimator, which is given by $P_m(\eta I{-}\mathcal{L})^{-1}$, i.e. $(z^\theta_i)_{i\in[m]}$ span the leading invariant subspace $(f_i)_{i\in[m]}$ of the generator. So, if

$$\mathcal{E}_\alpha(\theta) = \mathcal{E}(\theta) = \sum_{j=1}^{m}(\eta-\lambda_i)^{-1}$$

and $\alpha > 0$, we have that $\mathcal{E}_{\mathrm{on}}(\theta) = 0$, implying that $(z_i)_{i\in[m]}$ is an orthonormal basis, and, hence $P_m(\eta I{-}\mathcal{L})^{-1}$, i.e. $(z^\theta_i)_{i\in[m]} = \mathcal{Z}_\theta\Lambda^\eta_\theta\mathcal{Z}^*_\theta$. The result follows.

To show that

$$\mathcal{E}_\gamma(\theta) = \mathbb{E}\big[\mathcal{E}^{\widehat{\pi}'_1,\widehat{\pi}'_2}_\gamma(\theta)\big]/\overline{w}^2,$$

we rewrite (18) to encounter the distribution change, noting that the empirical covariances are reweighted but not normaliezed by $\overline{w}$. So, we have that

$$\mathcal{E}_\gamma(\theta) = \mathrm{tr}\left[\overline{w}^{-2}\mathbb{E}[\widehat{\mathrm{C}}_\theta]\Lambda^\eta_\theta\mathbb{E}[\widehat{\mathrm{W}}_\theta]\Lambda^\eta_\theta - 2\overline{w}^{-1}\mathbb{E}[\mathrm{C}_\theta]\Lambda^\eta_\theta + \overline{w}^{-2}\alpha(\mathbb{E}[\widehat{\mathrm{C}}_\theta] - \overline{w}\mathrm{I})^2\right].$$

But, since

$$\mathbb{E}_{x\sim\pi}[f(x)g(x)]{=}\mathbb{E}_{x'\sim\pi'}\big[\tfrac{d\pi}{d\pi'}(x')\,f(x')\,g(x')\big] {=}\mathbb{E}_{x'_1\sim\pi'}\big[\sqrt{\tfrac{d\pi}{d\pi'}(x'_1)}f(x'_1)\big]\,\mathbb{E}_{x'_2\sim\pi'}\big[\sqrt{\tfrac{d\pi}{d\pi'}(x'_2)}g(x'_2)\big],$$

where x'_1 and x'_2 are two independent r.v. with a law π', the proof is completed. □

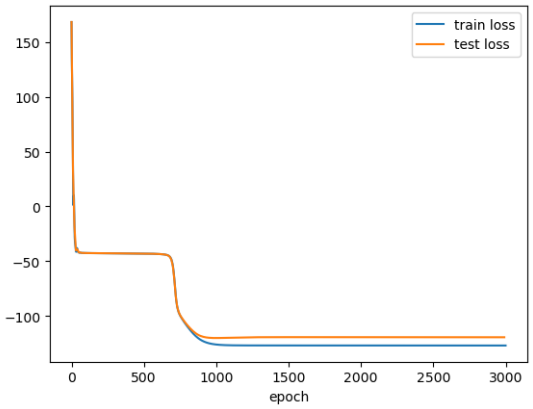

Figure 5: Typical behavior of the loss function during a training.

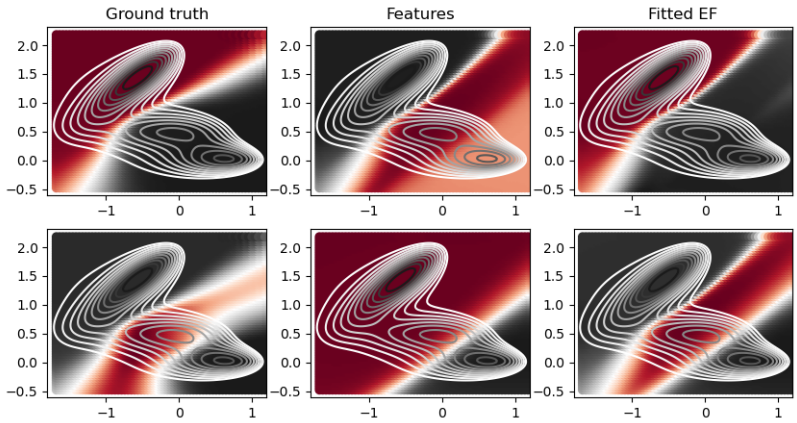

Figure 6: Comparison on Muller Brown potential with ground truth, learned features and fitted eigenfunctions

D Training of a neural network model

D.1 Evolution of the loss with eigenfunctions

It is interesting to note that during the training process, the loss reaches progressively lower plateaus. This is due to the fact that the NN has found a novel eigenfunction orthogonal to the previously found ones, starting with the constant one. Then during the plateau phase, the subspace is being explored until a new relevant direction is found. Typical behavior is shown in Figure 5. It was obtained for the case of a double well potential (see appendix below), but the same behavior was observed in all the training sessions. This nice property is a handy tool in properly optimizing the loss and understanding the proper stopping time.

D.2 Training with imperfect features

One of the main advantages of our method is that even with features that are not eigenfunctions of the generator, but that were trained with our method, we can recover the good eigenfunction estimates as proven in Theorem 1. In Figure 6, we illustrate such situation on a simulation of the Muller Brown potential: the trained features do not represent the ground truth, however, using our fitting method on the same dataset, we managed to recover eigenfunctions close to the ground truth. This is to the best of our knowledge the first time this kind of "learn and fit" method has been applied to the learning of the infinitesimal generator.

https://doi.org/10.52202/079017-2404

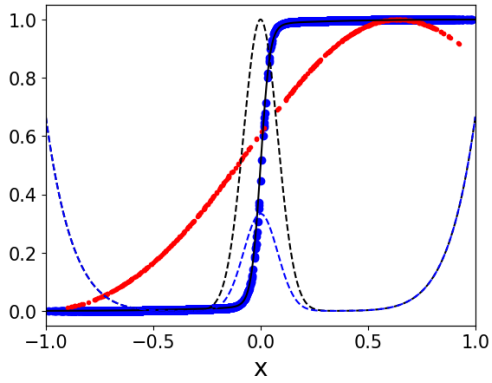

Figure 7: Simulation of the double well potential (black dashed lines) under an effective biased potential (blue dashed lines). Our method (blue points) compared to ground truth (black line) and transfer operator based approach (red points)

D.3 Activation functions and structure of the neural network

In all of our experiments we used the hyperbolic tangent activation function. This choice was made because it is a widely used, bounded function with continuous derivative. It thus satisfies all the criteria needed for this method. Finally, when looking for m eigenpairs, instead of having a neural network with m outputs, we choose to have m neural networks with one output.

D.4 Hyperparameters

Besides common hyperparameters such as learning rate, neural network architecture and activation function, our method requires only two hyperparameters: η and α. Other methods such as [50] do not require η, but on the other hand requires one weight per searched eigenfunction.

E Experiments

For all the experiments we used pytorch 1.13, and the optimizations of the models were performed using the ADAM optimizer. The version of python used is 3.9.18. All the experiments were performed on a workstation with a AMD® Ryzen threadripper pro 3975wx 32-cores processor and an NVIDIA Quadro RTX 4000 GPU. In all the experiments, the datasets were randomly split into a training and a validation dataset. The proportion were set to 80% for training and 20% for validation. The training of deepTICA models was performed using the mlcolvar package [8].

E.1 One dimensional double well potential

In this subsection, we showcase the efficiency of our method on a simple one dimensional toy model.

The target potential we want to sample has the form $U_{\mathrm{tg}}(x) = 4(-1.5\exp(-80x^2) + x^8)$, which has a form of two wells separated by a high barrier which can hardly be crossed during a simulation. In order to observe more transitions between the two wells and efficiently sample the space, we lower the barrier by running simulations under the following potential: $U_{\mathrm{sim}} = 4(-0.5\exp(-80x^2) + x^8)$, which thus makes a bias potential: $V_{\mathrm{bias}}(x) = U_{\mathrm{sim}}(x) - U_{\mathrm{tg}}(x) = -4(\exp(-80x^2))$. In Figure 7, we compare our method based on kernel methods (infinite dimensional dictionary of functions) with the ground truth and transfer operator baselines, namely deepTICA [7] which is a state-of-the-art method for molecular dynamics simulations. For this experiment we have used a Gaussian kernel of lengthscale 0.1, $\eta = 0.1$ and a regularization parameter of 10^{-5}

E.2 Muller Brown potential

For this experiment, the dataset was generated using an in-house code implementing the Euler-Maruyama scheme to discretize the overdamped Langevin equation. The simulation was performed at

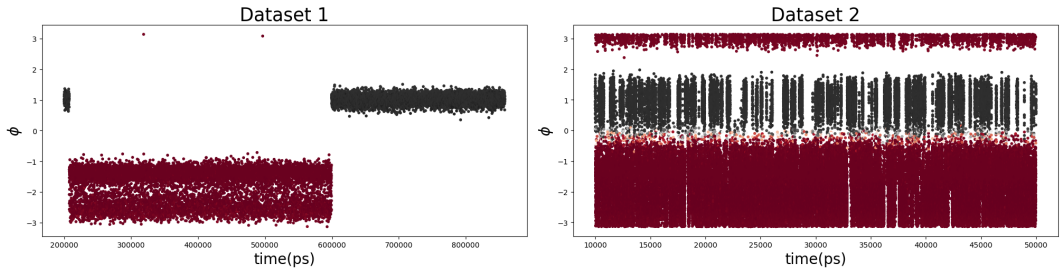

Figure 8: Time evolution of the ϕ angle in both datasets. Points are colored according to the value of the first nontrivial eigenfunction

a temperature of 1 (arbitrary unit) and a timestep of 10^{-3}. The simulation was run for 10^7 timesteps. The bias potential is built according to the following equation

$$U(x,t) = h \sum_{i=1}^{N_t} e^{-\frac{\|x-x_i\|^2}{2\sigma^2}} \tag{36}$$

where the centers x_i are built on the fly: every 500 timestep one more center is added to this list. This kind of bias potential is called metadynamics [25] and allows reducing the height of the barrier for a better exploration of the space. In order to have a static potential, no more centers are added after 300000 timesteps. We use a learning rate of 5.10^{-3}, the architecture of the neural network used is a multilayer perceptron with layers of size 2 (inputs), 20, 20 and 1. The parameter η was chosen to be 0.05.

E.3 Alanine dipeptide

Simulation details All the simulations are run with GROMACS 2022.3 [2] and patched with plumed 2.10 [45] in order to perform enhanced sampling simulations. We used the Amber99-SB [40] force field. The Langevin dynamics was sampled with a timestep of 2fs with a damping coefficent $\gamma_i = m_i/0.05$ at a target temperture of 300K. For both dataset, in order to make proper comparison, we used the explore version of OPES [17], with a barrier parameter of 20 kJ/mol and a pace of deposition of 500 timesteps.

Neural networks training For our neural networks training, we assume overdamped Langevin dynamics with the same value of the friction coefficient as in the simulation, as done in other works [50]. We use a learning rate of 10^{-3}, and the architecture of the neural network used is a multilayer perceptron with layers of size 30 (inputs), 20, 20 and 1. The parameter η was chosen to be 0.1.

E.4 Chignolin

Simulation details The simulations are run with GROMACS 2022.3 [2] and patched with plumed 2.10 [45]. They share the same setup used for the D.E Shaw trajectory [28] used in the main text . For the same reason, we kept the simulation condition consistent with that work. All simulations were performed with an integration time step of 2 fs and sampling NVT ensemble at 340K. The deepTDA model used for biasing the simulation is the one obtained in ref [46]

Neural networks training For our neural networks training, we assume overdamped Langevin dynamics. We use a learning rate of 5.10^{-4}, and the architecture of the neural network used is a multilayer perceptron with layers of size 210 (inputs), 50, 50 and 1. The parameter η was chosen to be 0.2.

https://doi.org/10.52202/079017-2404

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Justification: in Section 6 we provide experimental proofs of our claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] ,

   Justification: see the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: see Theorem 1 and 2 1; proofs are provided in the Appendix C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: the experimental setting is reported in Section 6 and expanded in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-2404

Answer: [Yes]

Justification: The experimental setting is reported in Section 6 and expanded in the appendix. The code is also shared on a github repository. The codes used to run the simulations are all open source.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: please see Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] .

   Justification: whenever appropriate we have reported these.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: please see Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We enjoy research and we respect other people work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We briefly present in the conclusions possible broader impacts of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-2404

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    his paper is a proof of concept, therefore we do not use large scale data or do not need any safeguard

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used open-source codes and algorithms, such as Gromacs and Plumed which are duly cited. D.E. Shaw has been acknowledged for providing the chignolin trajectory

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The github repository is provided.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2404