NEO: Non Equilibrium Sampling on the Orbit of a Deterministic Transform

Achille Thin, Yazid Janati, Sylvain Le Corff, Charles Ollion, Arnaud Doucet, Alain Durmus, Eric Moulines, Christian Robert

NEO: Non Equilibrium Sampling on the Orbit of a Deterministic Transform

Achille Thin[†], Yazid Janati[‡], Sylvain Le Corff[‡], Charles Ollion[†], Arnaud Doucet[⊤], Alain Durmus[⋆], Éric Moulines[†], and Christian Robert[≀]

[†]CMAP, École Polytechnique, Institut Polytechnique de Paris, Palaiseau.

[‡]Samovar, Télécom SudParis, département CITI, TIPIC, Institut Polytechnique de Paris, Palaiseau.

[⊤]Department of Statistics, University of Oxford.

[⋆]CMLA, École Normale Supérieure Paris-Saclay.

[≀]Ceremade, Université Paris-Dauphine & Department of Statistics, University of Warwick.

Abstract

Sampling from a complex distribution π and approximating its intractable normalizing constant Z are challenging problems. In this paper, a novel family of importance samplers (IS) and Markov chain Monte Carlo (MCMC) samplers is derived. Given an invertible map T, these schemes combine (with weights) elements from the forward and backward Orbits through points sampled from a proposal distribution ρ. The map T does not leave the target π invariant, hence the name NEO, standing for Non-Equilibrium Orbits. NEO-IS provides unbiased estimators of the normalizing constant and self-normalized IS estimators of expectations under π while NEO-MCMC combines multiple NEO-IS estimates of the normalizing constant and an iterated sampling-importance resampling mechanism to sample from π. For T chosen as a discrete-time integrator of a conformal Hamiltonian system, NEO-IS achieves state-of-the art performance on difficult benchmarks and NEO-MCMC is able to explore highly multimodal targets. Additionally, we provide detailed theoretical results for both methods. In particular, we show that NEO-MCMC is uniformly geometrically ergodic and establish explicit mixing time estimates under mild conditions.

1 Introduction

Consider a target distribution of the form $\pi(x) \propto \rho(x)\mathrm{L}(x)$ where ρ is a probability density function (pdf) on $\mathbb{R}^d$ and L is a nonnegative function. Typically, in a Bayesian setting, π is a posterior distribution associated with a prior distribution ρ and a likelihood function L. An other situation of interest is generative modeling where π is the distribution implicitly defined by a Generative Adversarial Networks (GAN) discriminator-generator pair where ρ is the distribution of the generator and L is derived from the discriminator (Turner et al., 2019; Che et al., 2020). An other situation of interest is generative modeling where π is the distribution implicitly defined by a Variational Auto Encoder (VAE) encoder-decoder pair where ρ is the distribution output by the encoder and L is an importance weight between the distribution of the decoder and of the encoder (Kingma and Welling, 2013; Burda et al., 2016). We are interested in this paper in sampling from π and approximating its intractable normalizing constant $Z = \int \rho(x)\mathrm{L}(x)\mathrm{d}x$. These problems arise in many applications in statistics, molecular dynamics or machine learning, and remain challenging.

Many approaches to compute normalizing constants are based on Importance Sampling (IS) - see Agapiou et al. (2017); Akyildiz and Míguez (2021) and the references therein - and its variations, among others, Annealed Importance Sampling (AIS) (Neal, 2001; Wu et al., 2016; Ding and Freedman, 2019) and Sequential Monte Carlo (SMC) (Del Moral et al., 2006). More recently, Neural IS has also become very popular in machine learning; see e.g. El Moselhy and Marzouk (2012); Müller et al. (2019); Papamakarios et al. (2019); Prangle (2019); Wirnsberger et al. (2020). Neural IS is an adaptive IS which relies on an importance function obtained by applying a normalizing flow to a reference distribution. The parameters of this normalizing flow are chosen by minimizing a divergence between the proposal and the target (such as the Kullback-Leibler Müller et al. (2019) or the χ^2-divergence Agapiou et al. (2017)).

More recently, the *Non-Equilibrium IS* (NEIS) method has been introduced by Rotskoff and Vanden-Eijnden (2019) as an alternative to these approaches. Similar to Neural IS, NEIS consists in transporting samples $\{X^i\}_{i=1}^N$ from a reference distribution using a family of deterministic mappings. This family for NEIS is chosen to be an homogeneous differential flow $(\phi_t)_{t\in\mathbb{R}}$. In contrast to Neural IS, for any $i \in [N]$, the sample X^i is propagated both forward and backward in time along the orbits associated with $(\phi_t)_{t\in\mathbb{R}}$ until stopping conditions are met. Moreover, the resulting estimator of the normalizing constant is obtained by computing weighted averages of the whole orbit $(\phi_t(X^i))_{t\in[\tau_{+,i},\tau_{-,i}]}$, where $\tau_{+,i}, \tau_{-,i}$ are the resulting stopping times, and not only the endpoints $\phi_{\tau_{+,i}}(X^i), \phi_{\tau_{-,i}}(X^i)$. In Rotskoff and Vanden-Eijnden (2019), the authors provide an application of NEIS with $(\phi_t)_{t\in\mathbb{R}}$ associated with a conformal Hamiltonian dynamics, and reports impressive numerical results on difficult normalizing constants estimation problems, in particular for high-dimensional multimodal distributions.

We propose in this work NEO-IS which alleviates the shortcomings of NEIS. Similar to NEIS, samples are drawn from a reference distribution, typically set to ρ, and are propagated under the forward and backward orbits of a *discrete-time* dynamical system associated with an invertible transform T. An estimator of the normalizing constant is obtained by reweighting all the points on the whole orbits using the IS rule. Contrary to NEIS, the NEO-IS estimator of Z is unbiased under assumptions that are mild and easy to verify. It is more flexible than NEIS because it does not rely on the accuracy of the discretization of a continuous-time dynamical system.

We then show how it is possible to leverage the unbiased estimator of Z defined by NEO-IS to obtain NEO-MCMC, a novel massively parallel MCMC algorithm to sample from π. In a nutshell, NEO-MCMC relies on parallel walkers which each estimates the normalizing constant but are allowed to interact through a resampling mechanism. Our contributions can be summarized as follows.

- **(i)** We present a novel class of IS estimators of the normalizing constant Z referred to as NEO-IS. More broadly, a small modification of this algorithm also allows us to estimate integrals with respect to π. Both finite sample and asymptotic guarantees are provided for these two methodologies.

- **(ii)** We develop a new massively parallel MCMC method, NEO-MCMC. NEO-MCMC combines NEO-IS unbiased estimator of the normalizing constant with iterated sampling-importance resampling methods. We prove that it is π-reversible and ergodic under very general conditions. We derive also conditions which imply that NEO-MCMC is uniformly geometrically ergodic (with an explicit expression of the mixing time).

- **(iii)** We illustrate our findings using numerical benchmarks which show that both NEO-IS and NEO-MCMC outperform state-of-the-art (SOTA) methods in difficult settings.

Algorithm 1 NEO-IS Sampler

1. Sample $X^{1:N} \overset{\text{iid}}{\sim} \rho$ for $i \in [N]$.
2. For $i \in [N]$, compute the path $(\mathrm{T}^j(X^i))_{j=0}^K$ and weights $(w_j(X^i))_{j=0}^K$.
3. $I_{\varpi,N}^{\mathrm{NEO}}(f) = N^{-1} \sum_{i=1}^N \sum_{k\in\mathbb{Z}} w_k(X^i) f(\mathrm{T}^k(X^i))$.

2 NEO-IS algorithm

In this section, we derive the NEO-IS algorithm. The two key ingredients for this algorithm are (1) the reference distribution ρ and (2) a transformation T assumed to be a C^1-diffeomorphism with inverse T^{-1}. Write, for $k \in \mathbb{N}^* = \mathbb{N} \setminus \{0\}$, $\mathrm{T}^k = \mathrm{T} \circ \mathrm{T}^{k-1}$, $\mathrm{T}^0 = \mathrm{Id}_d$ and similarly $\mathrm{T}^{-k} = \mathrm{T}^{-1} \circ \mathrm{T}^{-(k-1)}$. For any $k \in \mathbb{Z}$, denote by $\rho_k : \mathbb{R}^d \to \mathbb{R}_+$ the pushforward of ρ by T^k, defined for $x \in \mathbb{R}^d$ by $\rho_k(x) = \rho(\mathrm{T}^{-k}(x))\mathbf{J}_{\mathrm{T}^{-k}}(x)$, where $\mathbf{J}_\Phi(x) \in \mathbb{R}^+$ is the absolute value of the Jacobian determinant of $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ evaluated at x. In line with multiple importance sampling *à la* Owen and Zhou Owen and Zhou (2000), we introduce the proposal density

$$\rho_{\mathrm{T}}(x) = \Omega^{-1} \sum\nolimits_{k\in\mathbb{Z}} \varpi_k \rho_k(x) \;, \tag{1}$$

where $\{\varpi_k\}_{k\in\mathbb{Z}}$ is a nonnegative sequence and $\Omega = \sum_{k\in\mathbb{Z}} \varpi_k$. Note that we assume in the sequel that the support of the weight sequence defined as $\{k \in \mathbb{Z} : \varpi_k \neq 0\}$ is finite. Thus, the mixture distribution in (1) is a **finite mixture**. Given $x \in \mathbb{R}^d$, $\rho_{\mathrm{T}}(x)$ is a function of the forward and backward orbit of T through x. For any nonnegative function f, the definition of ρ_{T} implies that

$$\int f(y)\rho_{\mathrm{T}}(y)\mathrm{d}y = \Omega^{-1} \int \sum_{k\in\mathbb{Z}} \varpi_k f(\mathrm{T}^k(x))\rho(x)\mathrm{d}x \;.$$

Assuming that $\varpi_0 > 0$, the ratio $\rho(x)/\rho_{\mathrm{T}}(x) \leq \varpi_0^{-1}\Omega < \infty$ is bounded. We can therefore apply the IS principle which allows to write the identity

$$\int f(x)\rho(x)\mathrm{d}x = \int \left(f(y)\frac{\rho(y)}{\rho_{\mathrm{T}}(y)} \right) \rho_{\mathrm{T}}(y)\mathrm{d}y = \int \sum_{k\in\mathbb{Z}} f(\mathrm{T}^k(x)) w_k(x)\rho(x)\mathrm{d}x \;, \tag{2}$$

where the weights are given by (see Appendix A.2 for a detailed derivation),

$$w_k(x) = \varpi_k \rho(\mathrm{T}^k(x))/\{\Omega\rho_{\mathrm{T}}(\mathrm{T}^k(x))\} = \varpi_k\rho_{-k}(x) \Big/ \sum\nolimits_{i\in\mathbb{Z}} \varpi_{k+i}\rho_i(x) \quad . \tag{3}$$

We assume in the sequel that $\varpi_0 > 0$. In particular, note that under this condition, the weights w_k are also upper bounded uniformly in x: for any $x \in \mathbb{R}^d$, $w_k(x) \leq \varpi_k/\varpi_0$. Eqs. (2) and (3) suggest to estimate the integral $\int f(x)\rho(x)\mathrm{d}x$ by $I_{\varpi,N}^{\mathrm{NEO}}(f) = N^{-1}\sum_{i=1}^N \sum_{k\in\mathbb{Z}} w_k(X^i) f(\mathrm{T}^k(X^i))$ where $\{X^i\}_{i=1}^N$ are i.i.d. samples from the proposal ρ, which is denoted by $X^{1:N} \overset{\text{iid}}{\sim} \rho$.

This estimator is obtained by a weighted combination of the elements of the independent forward and backward orbits $\{\mathrm{T}^k(X^i)\}_{k\in\mathbb{Z}}$ with $X^{1:N} \overset{\text{iid}}{\sim} \rho$. This estimator is referred to as NEO-IS. Choosing $f \equiv \mathrm{L}$ provides the NEO-IS estimator of the normalizing constant of π:

$$\widehat{Z}_{X^i} = \sum_{k\in\mathbb{Z}} \mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i) \;, \quad \widehat{Z}_{X^{1:N}} = N^{-1}\sum_{i=1}^N \widehat{Z}_{X^i} \;. \tag{4}$$

We now study the performance of the NEO-IS estimator. The following two quantities play a fundamental role in the analysis:

$$E_{\mathrm{T}}^{\varpi} = \mathbb{E}_{X\sim\rho}\big[\big(\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))/Z\big)^2\big], M_{\mathrm{T}}^{\varpi} = \sup_{x\in\mathbb{R}^d}\sum_{k\in\mathbb{Z}} w_k(x)\mathrm{L}(\mathrm{T}^k(x))/Z \;. \tag{5}$$

Theorem 1. *$\widehat{Z}_{X^{1:N}}$ is an unbiased estimator of Z. If $E_{\mathrm{T}}^{\varpi} < \infty$, then, $\mathbb{E}[|\widehat{Z}_{X^{1:N}}/Z - 1|^2] = N^{-1}(E_{\mathrm{T}}^{\varpi} - 1)$. If $M_{\mathrm{T}}^{\varpi} < \infty$, then, for any $\delta \in (0,1)$, with probability $1-\delta$, $\sqrt{N}\left|\widehat{Z}_{X^{1:N}}/Z - 1\right| \leq M_{\mathrm{T}}^{\varpi}\sqrt{\log(2/\delta)/2}$.*

The (elementary) proof is postponed to Appendix A.3. E_{T}^{ϖ} plays the role of the second-order moment of the importance weights $\mathbb{E}_{X\sim\rho}[\mathrm{L}^2(X)]$ which is key to the performance of IS algorithms Agapiou et al. (2017); Akyildiz and Míguez (2021). In addition, since the NEO-IS estimator $\widehat{Z}_{X^{1:N}}$ is unbiased, the Cauchy-Schwarz inequality shows that $\mathbb{E}_{X\sim\rho}\big[\big(\sum_{k\in\mathbb{Z}} w_k(X)\mathrm{L}(\mathrm{T}^k(X))\big)^2\big] \geq Z^2$ and hence that $E_{\mathrm{T}}^{\varpi} \geq 1$. Note that if $\|\mathrm{L}\|_\infty = \sup_{x\in\mathbb{R}^d}\mathrm{L}(x) < \infty$, then since the weights are uniformly bounded by $\Omega\varpi_0^{-1}$, we have $M_{\mathrm{T}}^{\varpi} \leq \|\mathrm{L}\|_\infty\Omega\varpi_0^{-1}/Z$.

Using the NEO-IS estimate $\widehat{Z}_{X^{1:N}}$ of the normalizing constant, we can construct a self-normalized IS estimate of $\int f(x)\pi(x)\mathrm{d}x$:

$$J_{\varpi,N}^{\mathrm{NEO}}(f) = N^{-1}\sum_{i=1}^N \frac{\widehat{Z}_{X^i}}{\widehat{Z}_{X^{1:N}}}\sum_{k\in\mathbb{Z}} \frac{\mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)}{\widehat{Z}_{X^i}} f(\mathrm{T}^k(X^i)) \;, \tag{6}$$

referred to as NEO-SNIS estimator. This expression may seem unnecessarily complicated but highlights the hierarchical structure of the estimator. We combine estimators $(\widehat{Z}_{X^i})^{-1}\sum_{k\in\mathbb{Z}}\mathrm{L}(\mathrm{T}^k(X^i))w_k(X^i)f(\mathrm{T}^k(X^i))$ evaluated on the forward and backward orbits through the points $\{X^i\}_{i=1}^N$ using weights $\{\widehat{Z}_{X^i}/\widehat{Z}_{X^{1:N}}\}_{i=1}^N$. Although the NEO-IS estimator is unbiased, the NEO-SNIS is in general biased. However, for bounded functions, both the bias and the variance of the NEO-SNIS estimator are $O(N^{-1})$, with constants proportional to E_{T}^{ϖ}. For g a π-integrable function, we set $\pi(g) = \int g(x)\pi(x)\mathrm{d}x$.

Theorem 2. *Assume that $E_{\mathrm{T}}^{\varpi} < \infty$. Then, for any function g satisfying $\sup_{x\in\mathbb{R}^d}|g(x)| \leq 1$ on $\mathbb{R}^d$, and $N \in \mathbb{N}$,*

$$\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[|J_{\varpi,N}^{\mathrm{NEO}}(g) - \pi(g)|^2\right] \leq 4\cdot N^{-1}E_{\mathrm{T}}^{\varpi} \;, \tag{7}$$

$$\left|\mathbb{E}_{X^{1:N}\overset{\mathrm{iid}}{\sim}\rho}\left[J_{\varpi,N}^{\mathrm{NEO}}(g) - \pi(g)\right]\right| \leq 2\cdot N^{-1}E_{\mathrm{T}}^{\varpi} \;. \tag{8}$$

If $M_{\mathrm{T}}^{\varpi} < \infty$, then for $\delta \in (0,1]$, with probability at least $1-\delta$,

$$\sqrt{N}|J_{\varpi,N}^{\mathrm{NEO}}(g) - \pi(g)| \leq \|g\|_\infty M_{\mathrm{T}}^{\varpi}\sqrt{32\log(4/\delta)} \;. \tag{9}$$

The proof is postponed to Appendix A.4. These results extend to NEO-SNIS estimators the results known for self-normalized estimators; see e.g., Agapiou et al. (2017); Akyildiz and Míguez (2021) and the references therein. The upper bounds stated in this result suggest it is good practice to keep $E_{\mathrm{T}}^{\varpi}/N$ small in order to obtain sensible approximations. For two pdfs p and q on $\mathbb{R}^d$, denote by $\mathrm{D}_{\chi^2}(p,q) = \int\{p(x)/q(x) - 1\}^2 q(x)\mathrm{d}x$ the χ^2-divergence between p and q.

Lemma 3. *For any nonnegative sequence $(\varpi_k)_{k\in\mathbb{Z}}$, we have $E_{\mathrm{T}}^{\varpi} \leq \mathrm{D}_{\chi^2}(\pi\|\rho_{\mathrm{T}}) + 1$.*

The proof is postponed to Appendix A.5. Lemma 3 suggests that accurate sampling requires N to scale linearly with the χ^2-divergence between the target π and the extended proposal ρ_{T}.

Remark 1. We can extend NEO to non homogeneous flows, replacing the family $\{\mathrm{T}^k : k \in \mathbb{Z}\}$ with a collection of mappings $\{\mathsf{T}_k : k \in \mathbb{Z}\}$. This would allow us to consider further flexible classes of transformations such as normalizing flows; see e.g. Papamakarios et al. (2019). The χ^2-divergence $\mathrm{D}_{\chi^2}(\pi \| \rho_{\mathrm{T}})$ provides natural criteria for learning the transformation. We leave this extension to future work.

Conformal Hamiltonian transform The efficiency of NEO relies heavily on the choice of T. Intuitively, a sensible choice of T requires that (i) E_{T}^{ϖ} is small, i.e. ρ_{T} should be close to π by Lemma 3 (see (5)), (ii) the inverse T^{-1} and the Jacobian of T are easy to compute. Following Rotskoff and Vanden-Eijnden (2019), we use for T a discretization of a conformal Hamiltonian dynamics. Assume that $U(\cdot) = -\log \pi(\cdot)$ is continuously differentiable. We consider the augmented distribution $\tilde{\pi}(q,p) \propto \exp\{-U(q) - K(p)\}$ on $\mathbb{R}^{2d}$, where q is the position, p is the momentum, and $K(p) = p^T \mathrm{M}^{-1} p/2$ is the kinetic energy, with M a positive definite mass matrix. By construction, the marginal distribution of the momentum under $\tilde{\pi}$ is the target pdf $\pi(q) = \int \tilde{\pi}(q,p)\mathrm{d}p$. The conformal Hamiltonian ODE associated with $\tilde{\pi}$ is defined by

$$\begin{aligned} \mathrm{d}q_t/\mathrm{d}t &= \nabla_p H(q_t, p_t) = \mathrm{M}^{-1} p_t \ , \\ \mathrm{d}p_t/\mathrm{d}t &= -\nabla_q H(q_t, p_t) - \gamma p_t = -\nabla U(q_t) - \gamma p_t \ , \end{aligned} \tag{10}$$

where $H(q,p) = U(q) + K(p)$, and $\gamma > 0$ is a damping constant. Any solution $(q_t, p_t)_{t \geq 0}$ of (10) satisfies setting $H_t = H(q_t, p_t)$, $\mathrm{d}H_t/\mathrm{d}t = -\gamma p_t^T \mathrm{M}^{-1} p_t \leq 0$. Hence, all orbits converge to fixed points that satisfy $\nabla U(q) = 0$ and $p = 0$; see e.g. Franca et al. (2019); Maddison et al. (2018).

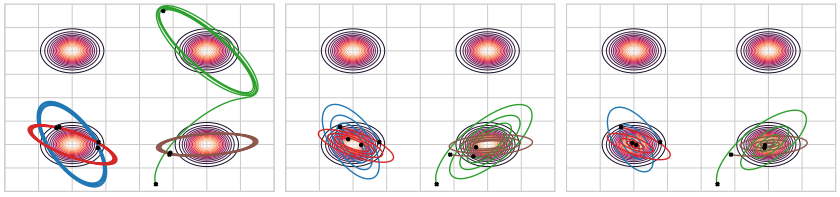

Figure 1: Left: $E_{\mathrm{T}_h}^{\mathbb{1}_{[K]}}(K)$ vs $E^{\mathrm{IS}}(K)$ (red) in $\log_{10}$-scale as a function of optimization step K. Second left to right: Corresponding orbits for $\gamma = 0.1, 1, 2$.

In the applications below, we consider the conformal version of the symplectic Euler (SE) method of (10), see Franca et al. (2019). This integrator can be constructed as a splitting of the two conformal and conservative parts of the system (10). When composing a dissipative with a symplectic operator, we set for all $(q,p) \in \mathbb{R}^{2d}$, $\mathrm{T}_h(q,p) = (q + h\mathrm{M}^{-1}\{\mathrm{e}^{-h\gamma}p - h\nabla U(q)\}, \mathrm{e}^{-h\gamma}p - h\nabla U(q))$, where $h > 0$ is a discretization stepsize. This transformation can be connected with classical momentum optimization schemes, see (Franca et al., 2019, Section 4). By (Franca et al., 2019, Section 3), for any $h > 0$ T_h is a C^1-diffeomorphism on $\mathbb{R}^{2d}$ with Jacobian given by $\mathbf{J}_{\mathrm{T}_h}(q,p) = \mathrm{e}^{-\gamma h d}$. In addition, its inverse is $\mathrm{T}_h^{-1}(q,p) = (q - h\mathrm{M}^{-1}p, \mathrm{e}^{\gamma h}\{p + h\nabla U(q - h\mathrm{M}^{-1}p)\})$. Therefore, the weight (3) of the NEO estimator is given by

$$w_k(q,p) = \frac{\varpi_k \tilde{\rho}(\mathrm{T}_h^k(q,p))\mathrm{e}^{-\gamma k h d}}{\sum_{j \in \mathbb{Z}} \varpi_{k+j} \tilde{\rho}(\mathrm{T}_h^j(q,p))\mathrm{e}^{-\gamma j h d}} \ ,$$

where $\tilde{\rho}(q,p) \propto \rho(q)\mathrm{e}^{-K(p)}$. Figure 1 displays for different values of γ, the bound $E_{\mathrm{T}_h}^{\mathbb{1}_{[0:K]}} - 1$ (5) as a function of K corresponding to the sequence of weights $(\varpi_k)_{k \in \mathbb{Z}} = (\mathbb{1}_{[0:K]}(k))_{k \in \mathbb{Z}}$ (i.e. only the $K+1$

Algorithm 2 NEO-MCMC Sampler

At step $n \in \mathbb{N}^*$, given the conditioning orbit point Y_{n-1}.

Step 1: Update the conditioning point

1. Set $X_n^1 = Y_{n-1}$ and for any $i \in \{2, \dots, N\}$, sample $X_n^i \overset{\text{iid}}{\sim} \rho$.
2. Sample the orbit index I_n with probability proportional to $(\widehat{Z}_{X_n^i})_{i\in[N]}$, (4).
3. Set $Y_n = X_n^{I_n}$

Step 2: Output a sample

4. Sample index K_n with probability proportional to $\{w_k(Y_n)\mathrm{L}(\mathrm{T}^k(Y_n))/\widehat{Z}_{Y_n}\}_{k\in\mathbb{Z}}$
5. Output $U_n = \mathrm{T}^{K_n}(Y_n)$.

first elements of the forward orbits are used and are equally weighted). For comparison, we also present on the same plot, the bounds achieved by averaging $K+1$ independent IS estimates, $E^{\mathrm{IS}}(K) - 1 = (K+1)^{-1}\mathbb{E}_{X\sim\rho}[\mathrm{L}(X)^2]$. Interestingly, Figure 1 shows that there is a trade-off in the choice of γ which controls the exploration of the state space by the Hamiltonian dynamics since the higher γ, the faster the orbits converge towards the modes.

3 NEO-MCMC algorithm

We now derive sampling methods based on the NEO-IS estimator. A natural idea consists in adapting the Sampling Importance Resampling procedure (SIR) (see for example Rubin (1987); Skare et al. (2003)) to the NEO framework. The SIR method to sample $J^{\mathrm{NEO}}_{\varpi,N}$ (see (6)) consists of 4 steps.

(SIR-1) Draw independently $X^{1:N} \overset{\text{iid}}{\sim} \rho$ and compute the associated forward and backward orbits $\{\mathrm{T}^k(X^i)\}_{k\in\mathbb{Z}}$ of the point.
(SIR-2) Compute the normalizing constants associated with each orbit $\{\widehat{Z}_{X^i}\}_{i=1}^N$.
(SIR-3) Sample an orbit index $I_N \in [N]$ with probability $\{\widehat{Z}_{X^i}/\sum_{j=1}^N \widehat{Z}_{X^j}\}_{i=1}^N$.
(SIR-4) Draw the iteration index K^N on the I^N-th orbit with probability $\{\mathrm{L}(\mathrm{T}^k(X^{I^N}))w_k(X^{I^N})/\widehat{Z}_{X^{I^N}}\}_{k\in\mathbb{Z}}$.

The resulting draw is denoted by $U^N = \mathrm{T}^{K^N}(X^{I^N})$. By construction, for any bounded function f, we get that $\mathbb{E}\left[f(U^N)\middle|X^{1:N}, I^N\right] = \{\widehat{Z}_{X^{I^N}}\}^{-1}\sum_{k\in\mathbb{Z}} w_k(X_{I^N})\mathrm{L}(\mathrm{T}^k(X_{I^N}))$ which implies $\mathbb{E}\left[f(U^N)\middle|X^{1:N}, I^N\right] = J^{\mathrm{NEO}}_{\varpi,N}(f)$ (see (6)). Using Theorem 2, we therefore obtain $|\mathbb{E}[f(U^N)] - \int f(z)\pi(z)\mathrm{d}z| \leq 10^{1/2}\|f\|_\infty E^{\varpi}_{\mathrm{T}} N^{-1}$, showing that the law of the random variable $\mu_N = \mathrm{Law}(U^N)$ converges in total variation to π as $N \to \infty$,

$$\|\mu_N - \pi\|_{\mathrm{TV}} = \sup_{\|f\|_\infty \leq 1} |\mu_N(f) - \pi(f)| \leq 10^{1/2} E^{\varpi}_{\mathrm{T}} N^{-1} \,. \tag{11}$$

Based on Andrieu et al. (2010), we now derive the NEO-MCMC procedure, which in a nutshell consists in iterating the SIR procedure while keeping a conditioning point (or equivalently, orbit); see Appendix C.

The convergence of NEO-MCMC does not rely on letting $N \to \infty$: the NEO-MCMC works as soon as $N \geq 2$, although as we will see below the mixing time decreases as N increases.

This procedure is summarized in Algorithm 2. The NEO-MCMC procedure is an iterated algorithm which produces a sequence $\{(Y_n, U_n)\}_{n\in\mathbb{N}}$ of points in $\mathbb{R}^d$. The n-th iteration of the NEO-MCMC algorithm consists in two main steps: 1) updating the conditioning point $Y_{n-1} \to Y_n$ 2) sampling U_n by selecting a point in the orbit $\{\mathrm{T}^k(Y_n)\}_{k\in\mathbb{Z}}$ of the conditioning point. Compared to SIR, only the generation of the points (step (SIR-1)) is modified: we set $X_n^1 = Y_{n-1}$ (the **conditioning point**), and then draw $X_n^{2:N} \overset{\text{iid}}{\sim} \rho$. The sequence $\{Y_n\}_{n\in\mathbb{N}}$ defined by Algorithm 2 is a Markov chain:

$$\mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_{0:n-1}\right) = \mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_{n-1}\right) = P(Y_n, \mathsf{A}) \;,$$

where

$$P(y, \mathsf{A}) = \int \delta_y(\mathrm{d}x^1) \prod_{j=2}^{N} \rho(x^j)\mathrm{d}x^j \sum_{i=1}^{N} \frac{\widehat{Z}_{x^i}}{\sum_{j=1}^{N} \widehat{Z}_{x^j}} \mathbb{1}_{\mathsf{A}}(x^i) \;, \quad y \in \mathbb{R}^d \;, \mathsf{A} \in \mathcal{B}(\mathbb{R}^d) \;. \tag{12}$$

Note that this Markov kernel describes the way, at stage $n+1$, the conditioning point Y_{n+1} is selected given Y_n, which **depends only on** the estimator of the normalizing constants associated with each orbit, **but not** on the sample U_n selected on the conditioning orbit. In addition, given the conditioning point Y_n at the n-th iteration, the conditional distribution of the output sample U_n is $\mathbb{P}\left(U_n \in \mathsf{B} \,|\, I_n, X_n^{1:N}\right) = \mathbb{P}\left(U_n \in \mathsf{B} \,|\, Y_n\right) = Q(Y_n, \mathsf{B})$ where

$$Q(y, \mathsf{B}) = \sum_{k\in\mathbb{Z}} \frac{w_k(y)\mathrm{L}(\mathrm{T}^k(y))}{\widehat{Z}_y} \mathbb{1}_{\mathsf{B}}(\mathrm{T}^k(y)) \;, \quad y \in \mathbb{R}^d \;, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d) \;. \tag{13}$$

With these notations, if the Markov chain is started at $Y_0 = y$, then for any $n \in \mathbb{N}$, the law of the n-th conditioning point is $\mathbb{P}\left(Y_n \in \mathsf{A} \,|\, Y_0 = y\right) = P^n(y, \mathsf{A})$ and the law of the n-th sample is $\mathbb{P}\left(U_n \in \mathsf{B} \,|\, Y_0\right) = P^n Q(y, \mathsf{B})$. Define $\tilde{\pi}$ the pdf given, for $y \in \mathbb{R}^d$, by

$$\tilde{\pi}(y) = \frac{\rho(y)}{Z} \sum_{k\in\mathbb{Z}} w_k(y)\mathrm{L}(\mathrm{T}^k(y)) = \frac{\rho(y)\widehat{Z}_y}{Z} \;. \tag{14}$$

The following theorem shows that, for any initial condition $y \in \mathbb{R}^d$, the distribution of the variable Y_n converges in total variation to $\tilde{\pi}$ and that the distribution of U_n converges to π.

Theorem 4. *The Markov kernel P is reversible with respect to the distribution $\tilde{\pi}$, ergodic and Harris positive, i.e., for all $y \in \mathbb{R}^d$, $\lim_{n\to\infty} \|P^n(y,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} = 0$. In addition, $\pi = \tilde{\pi}Q$ and $\lim_{n\to\infty} \|P^n Q(y,\cdot) - \pi\|_{\mathrm{TV}} = 0$. Moreover, for any bounded function g and any $y \in \mathbb{R}^d$, $\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} g(U_i) = \pi(g)$, $\mathbb{P}$-almost surely, where $\{U_i\}_{i\in\mathbb{N}}$ is defined in Algorithm 2 with $Y_0 = y$.*

The proof is postponed to Appendix A.6.

Remark 2. We may provide another sampling procedure of $\{Y_n\}_{n\in\mathbb{N}}$. Define the pdf on the extended space $[N] \times \mathbb{R}^{dN}$ by $\check{\pi}(i, x^{1:N}) = N^{-1}\tilde{\pi}(x^i) \prod_{j=1, j\neq i}^{N} \rho(x^j)$. Consider a Gibbs sampler targeting $\check{\pi}$ consisting in (a) sampling $X_n^{1:N\setminus\{I_{n-1}\}} | (I_{n-1}, X_{n-1}) \sim \prod_{j\neq I_{n-1}} \rho(x^j)$, (b) sampling $I_n | X_n^{1:N} \sim \mathrm{Cat}(\{\widehat{Z}_{X_n^i} / \sum_{j=1}^{N} \widehat{Z}_{X_n^j}\}_{i=1}^{N}$ and (c) set $Y_n = X_n^{I_n}$. This algorithm is a Gibbs sampler on $\check{\pi}$ and we easily verify that the distribution of $\{Y_n\}_{n\in\mathbb{N}}$ is the same as Algorithm 2.

The next theorem provides non asymptotic quantitative bounds on the convergence in total variation. The main interest of NEO-MCMC algorithm is motivated empirically from observed behaviour: the mixing time of the corresponding Markov chain improves as N increases. This behaviour is quantified theoretically in the next theorem. Moreover, this improvement is obtained with little extra computational overhead, since sampling N points from the proposal distribution ρ, computing the forward and backward orbits of the points and evaluating the normalizing constants $\{\widehat{Z}_{X_n^i}\}_{i=1}^N$ can be performed in parallel.

Theorem 5. *Assume that $M_{\mathrm{T}}^{\varpi} < \infty$, see* (5)*. Set $\epsilon_N = (N-1)/(2M_{\mathrm{T}}^{\varpi} + N - 2)$ and $\kappa_N = 1 - \epsilon_N$. Then, for any $y \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\|P^k(y,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} \leq \kappa_N^k$ and $\|P^k Q(y,\cdot) - \pi\|_{\mathrm{TV}} \leq \kappa_N^k$.*

Instead of sampling the new points $X_n^{2:N}$ independently from ρ (Step 1 in Algorithm 2), it is possible to draw the proposals $X_n^{1:N}$ conditional to the current point Y_{n-1}; see So (2006); Craiu and Lemieux (2007); Shestopaloff et al. (2018); Ruiz et al. (2020) for related works. Following Ruiz et al. (2020), we use a reversible Markov kernel with respect to the proposal ρ, i.e., such that $\rho(x)m(x,x') = \rho(x')m(x',x)$, assuming for simplicity that this kernel has density $m(x,x')$. If $\rho = \mathrm{N}(0, \sigma^2 \mathrm{Id}_d)$, an appropriate choice is an autoregressive kernel $m(x,x') = \mathrm{N}(x'; \alpha x, \sigma^2(1-\alpha^2)\mathrm{Id}_d)$. More generally, we can use a Metropolis–Hastings kernel with invariant distribution ρ. In this case, for each $i \in [N]$, define for $i \in [N]$,

$$r_i(x^i, x^{1:N\setminus\{i\}}) = \prod_{j=1}^{i-1} m(x^{j+1}, x^j) \prod_{j=i+1}^{N} m(x^{j-1}, x^j) \; . \tag{15}$$

Since m is reversible with respect to ρ, for all $i, j \in [N]$, $\rho(x^i) r_i(x^i, x^{1:N\setminus\{i\}}) = \rho(x^j) r_j(x^j, x^{1:N\setminus\{j\}})$. The only modification in Algorithm 2 is Step 1, which is replaced by: *Draw $U_n \in [N]$ uniformly, set $X_n^{U_n} = Y_{n-1}$ and sample $X_n^{1:N\setminus\{U_n\}} \sim r_{U_n}(X_n^{U_n}, \cdot)$.* The validity of this procedure is established in Appendix A.6.

4 Continuous-time version of NEO and NEIS

The NEO framework takes up and extends NEIS introduced in Rotskoff and Vanden-Eijnden (2019). NEIS focuses on normalizing constant estimation and should be therefore compared with NEO-IS. In Rotskoff and Vanden-Eijnden (2019), the authors do not consider possible extensions of these ideas to sampling problems. Proofs of the statements and detailed technical conditions are postponed to Appendix B. We first consider how NEO can be adapted to continuous-time dynamical system. Consider the Ordinary Differential Equation (ODE) $\dot{x}_t = b(x_t)$, where $b\colon \mathbb{R}^d \to \mathbb{R}^d$ is a smooth vector field. Denote by $(\phi_t)_{t\in\mathbb{R}}$ the flow of this ODE (assumed to be well-behaved). Under appropriate regularity condition $\mathbf{J}_{\phi_t}(x) = \exp(\int_0^t \nabla \cdot b(\phi_s(x))\mathrm{d}s)$; see Lemma 10. Let $\varpi : \mathbb{R} \to \mathbb{R}_+$ be a nonnegative smooth function with finite support, with $\Omega^c = \int_{-\infty}^{\infty} \varpi(t)\mathrm{d}t$. The continuous-time counterpart of the proposal distribution (1) is $\rho_{\mathrm{T}}^{\mathrm{c}}(x) = (\Omega^c)^{-1} \int_{-\infty}^{\infty} \varpi(t)\rho(\phi_{-t}(x))\mathbf{J}_{\phi_{-t}}(x)\mathrm{d}t$, which is a continuous mixture of the pushforward of the proposal ρ by the flow of $(\phi_s)_{s\in\mathbb{R}}$. Assuming for simplicity that $\rho(x) > 0$ for all $x \in \mathbb{R}^d$, then $\rho_{\mathrm{T}}^{\mathrm{c}}(x) > 0$ for all $x \in \mathbb{R}^d$, and using again the IS formula, for any nonnegative function f,

$$\int f(x)\rho(x)\mathrm{d}x = \int f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}^{\mathrm{c}}(x)}\rho_{\mathrm{T}}^{\mathrm{c}}(x)\mathrm{d}x = \int \left[\int_{-\infty}^{\infty} w_t^{\mathrm{c}}(x) f(\phi_t(x))\mathrm{d}t\right]\rho(x)\mathrm{d}x \; , \tag{16}$$

$$w_t^{\mathrm{c}}(x) = \varpi(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) \Big/ \int_{-\infty}^{\infty} \varpi(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s \; . \tag{17}$$

These relations are the continuous-time counterparts of (2). Eqs. (16)-(17) define a version of NEIS Rotskoff and Vanden-Eijnden (2019), with a finite support weight function ϖ; see Appendices B.2 and B.3 for weight functions with infinite support. This identity is of theoretical interest but must be discretized to obtain a computationally tractable estimator. For $h > 0$, denote by T_h an integrator with stepsize $h > 0$ of the ODE $\dot{x} = b(x)$. We may construct NEO-IS and NEO-SNIS estimators based on the transform $\mathrm{T} \leftarrow \mathrm{T}_h$ and weights $\varpi_k \leftarrow \varpi(kh)$. We might show that for any bounded function f and for any $x \in \mathbb{R}^d$, $\lim_{h \downarrow 0} \sum_{k \in \mathbb{Z}} w_k(x) f(\mathrm{T}_h^k(x)) = \int_{-\infty}^{\infty} w_t^c(x) f(\phi_t(x)) \mathrm{d}t$, where we omitted here the dependency in h of w_k. Therefore, taking $h \downarrow 0^+$, the NEO-IS converges to the continuous-version (16)-(17). There is however an important difference between NEO and the NEIS method in Rotskoff and Vanden-Eijnden (2019) which stems from the way (16)-(17) are discretized. Compared to NEIS, NEO-IS using $\mathrm{T} \leftarrow \mathrm{T}_h$ and weights $\varpi_k \leftarrow \varpi(kh)$ is unbiased for any stepsize $h > 0$. NEIS uses an approach inspired by the nested-sampling approach, which amounts to discretizing the integral in (16) also in the state-variable x; see Skilling (2006); Chopin and Robert (2010). This discretization is biased which prevents the use of this approach to develop MCMC sampling algorithm; see Appendix B.

5 Experiments and Applications

Normalizing constant estimation The performance of NEO-IS is assessed on different normalizing constant estimation benchmarks; see Jia and Seljak (2020). We focus on two challenging examples. Additional experiments and discussion on hyperparameter choice are given in the supplementary material, see Appendix D.1. **(1) Mixture of Gaussian `(MG25)`** : $\pi(x) = P^{-1} \sum_{i=1}^{P} \mathrm{N}(x; \mu_{i,j}, D_d)$, where $d \in \{10, 20, 40\}$, $D_d = \mathrm{diag}(0.01, 0.01, 0.1, \ldots, 0.1)$ and $\mu_{i,j} = [i, j, 0, \ldots, 0]^T$ with $i, j \in \{-2, \ldots, 2\}$. **(2) Funnel distribution `(Fun)`** $\pi(x) = \mathrm{N}(x_1; 0, a^2) \prod_{i=1}^{d} \mathrm{N}(x_i; 0, \mathrm{e}^{2bx_1})$ with $d \in \{10, 20, 40\}$, $a = 1$, and $b = 0.5$. In both case, the proposal is $\rho = \mathrm{N}(0, \sigma_\rho^2 \mathrm{Id}_d)$ with $\sigma_\rho^2 = 5$.

The NEO-IS estimator is compared with (i) the IS estimator using the proposal ρ, (ii) the Adaptive Importance Sampling (AIS) estimator of Tokdar and Kass (2010) and (iii) the Neural Importance Sampling (NIS)[1]. For NEO-IS, we use $\varpi_k = \mathbb{1}_{[K]}(k)$ with $K = 10$ (ten steps on the forward orbit), and conformal Hamiltonian dynamics $\gamma = 1$, $M = 5 \cdot \mathrm{Id}_d$ for dimensions $d = \{10, 20\}$, and $\gamma = 2.5$ for $d = 40$ (where γ is the damping factor, M the mass matrix, h is the stepsize of the integrator). The parameters of AIS are set to obtain a complexity comparable to NEO-IS; see Appendix D.1. For NIS, we use the default parameters. In `Fun`, we set $\gamma = 0.2$, $K = 10$, $M = 5 \cdot \mathrm{Id}_d$, and $h = 0.3$. The IS estimator was based on $5 \cdot 10^5$ samples, and NIS, NEO-IS and AIS were computed with $5 \cdot 10^4$ samples. Figure 2 shows that NEO-IS consistently outperforms the competing methods.

Sampling NEO-MCMC is assessed for the distributions `(MG25)` ($d = 40$) and `Fun` ($d = 20$). NEO-MCMC sampler is compared with (i) the No-U-Turn Sampler - Pyro library Bingham et al. (2019) - and (ii) i-SIR algorithm Ruiz et al. (2020). The proposal distribution is $\rho = \mathrm{N}(0, \sigma_\rho^2 \mathrm{Id}_d)$ with $\sigma_\rho^2 = 5$. Dependent proposals are used (see (15)) with $m(x, x') = \mathrm{N}(x'; \alpha x, \sigma_\rho^2(1 - \alpha^2) \mathrm{Id}_d)$ with $\alpha = 0.99$. For NUTS, the default parameters are used. For i-SIR, we use the same number of proposals $N = 10$, proposal distribution and dependent proposal as for NEO-MCMC. To make a fair comparison, we use the same clock time for all three algorithms. The number of iterations for correlated i-SIR, NEO-MCMC, and NUTS are $n = 4 \cdot 10^6$, $n = 4 \cdot 10^5$, and $n = 5 \cdot 10^5$, respectively. Figure 3 displays the empirical two-dimensional histograms of the two first coordinates of samples from the ground truth, i-SIR, NUTS and NEO-MCMC sampler. It is worthwhile to note that NEO-MCMC algorithm performs much better for `MG25` which is a very challenging

[1] We used the implementation provided by: `https://github.com/ndeutschmann/zunis`

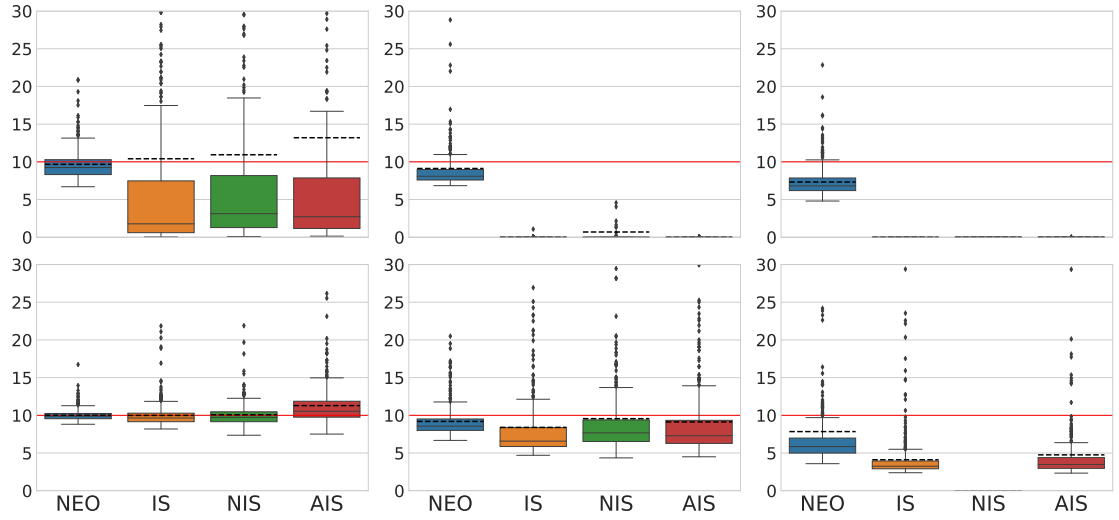

Figure 2: Boxplots of 500 independent estimations of the normalizing constant in dimension $d = \{10, 20, 45\}$ (from left to right) for `MG25` (top) and `Fun` (bottom). The true value is given by the red line. The figure displays the median (solid lines), the interquartile range, and the mean (dashed lines) over the 500 runs.

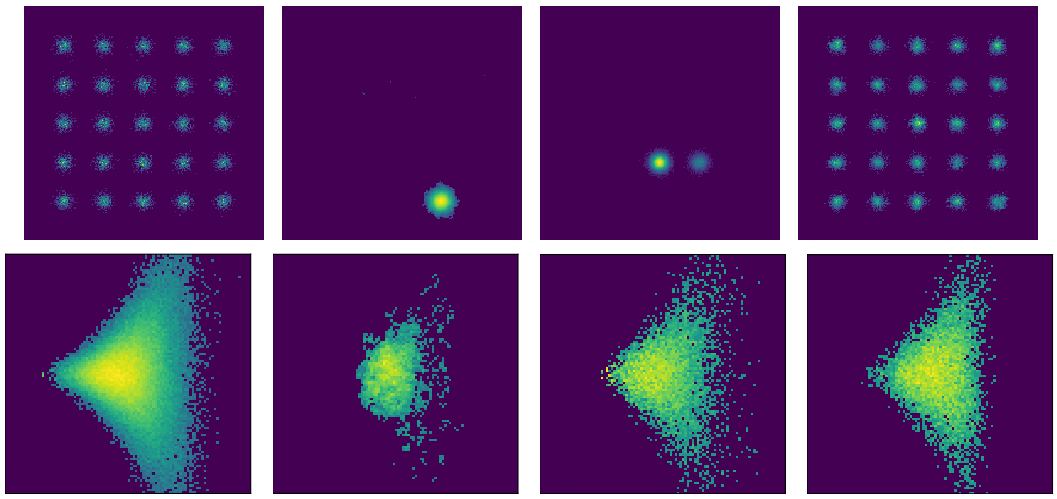

Figure 3: Empirical 2-D histogram of the samples of different algorithms targeting `MG25` example (top) and `Fun` (bottom). From left to right: samples from the target distribution, correlated i-SIR, NUTS, NEO-MCMC.

distribution, even for SOTA algorithm such as NUTS, which struggles to cross energy barriers between modes. For `Fun`, NEO-MCMC performs favourably with respect to NUTS, which is well adapted for this type of distributions.

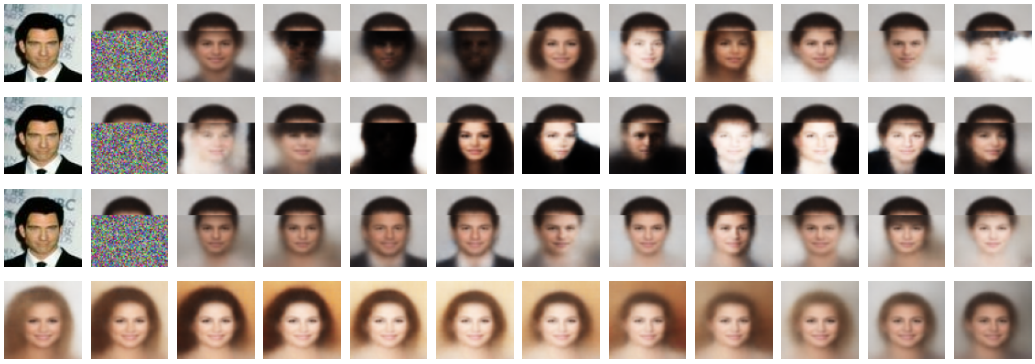

Figure 4: Gibbs inpainting for CelebA dataset. From top to bottom: i-SIR, HMC and NEO-MCMC: From left to right, original image, blurred image to reconstruct, and output every 5 iterations of the Markov chain. Last line: a forward orbit used in NEO-MCMC.

Block Gibbs Inpainting with Deep Generative models and NEO-MCMC We apply NEO-MCMC to the task of sampling the posterior of a deep latent variable model. The model consists of a latent variable $x \sim \mathrm{N}(0, \mathrm{Id}_d)$ and a conditional distribution $p(z \mid x)$ which generates an image $z = (z^1, \ldots, z^D) \in \mathbb{R}^D$. Given a family of parametric *decoders* $\{x \mapsto p_\theta(z \mid x), \theta \in \Theta\}$, and a training set $\mathcal{D} = \{z_i\}_{i=1}^M$, training involves finding the MLE $\theta^* = \arg\max_{\theta\in\Theta} p_\theta(\mathcal{D})$. As $p_\theta(z) = \int p_\theta(z \mid x)p(x)\mathrm{d}x$, the likelihood is intractable and to alleviate this problem, Kingma and Welling (2013) proposed to train jointly an approximate posterior $q_\phi(x|z)$ that maximizes a tractable lower-bound on the log-likelihood: $\mathrm{ELBO}(z, \theta, \phi) = \mathbb{E}_{X\sim q_\phi(\cdot|z)}[\log p_\theta(z, X)/q_\phi(X|z)] \leq p_\theta(z)$, where $q_\phi(x \mid z)$ is a tractable conditional distribution with parameters $\phi \in \Phi$. It is assumed in the sequel that conditional to the latent variable x, the coordinates are independent, *i.e.* $p_\theta(z \mid x) = \prod_{i=1}^D p_\theta(z^i|x)$.

It is possible to train VAE with the NEO algorithm using the unbiased estimate of the normalizing constant to construct an ELBO. This approach is described in the supplement Appendix E. It is assumed here that the VAE has been trained and we are only interested in the sampling problem. In our experiment, we use a VAE trained on CelebA dataset [2] Liu et al. (2018). We consider the Block Gibbs inpainting task introduced in (Levy et al., 2018, Section 5.2.2). We in-paint the bottom of an image using Block Gibbs sampling. Given an image z, denote by $[z^t, z^b]$ the top and the bottom half pixels. A two-stage Gibbs sampler amounts to (a) sampling $p_{\theta^*}(x|z^t, z^b)$ and (b) sampling $p_{\theta^*}(z^b|x, z^t) = p_{\theta^*}(z^b|x)$ (since z^b and z^t are independent conditional on x). Starting from an image z_0, we sample at each step $x_k \sim p_{\theta^*}(z \mid z_k)$ and then $\tilde{z}_k \sim p_{\theta^*}(z \mid x_k)$. We then set $z_{k+1} = (z_*^t, \tilde{z}_{k+1}^b)$. Stage (b) is elementary but stage (a) is challenging. We use the following decomposition of $p_{\theta^*}(x \mid z) = q_{\phi^*}^\beta(x \mid z)p_{\theta^*}(x, z)/(q_{\phi^*}^\beta(x \mid z)p_{\theta^*}(z))$ with $\beta \in (0, 1)$. We identify the *proposal* $\rho(x) \propto q_{\phi^*}^\beta(x \mid z)$, the *likelihood* $\mathrm{L} \leftarrow p_{\theta^*}(x, z)/q_{\phi^*}^\beta(x \mid z)$ and the normalizing constant $Z = p_{\theta^*}(z)$ and apply i-SIR and NEO-MCMC sampler for stage (a). We compare different algorithms in stage (a): i-SIR, HMC and NEO-MCMC, with the same computational complexity ($N = 10$, $K = 12$, $\gamma = 0.2$ for NEO-MCMC, $N = 120$ for i-SIR, and HMC is run with $K = 20$ leapfrog steps). For each algorithm, 10 steps are performed. Figure 8 displays the evolution of the resulting Markov chains. The samples clearly illustrate that NEO-MCMC mixes better than i-SIR and HMC. More are showcased in the supplementary.

[2] See `https://github.com/YannDubs/disentangling-vae/tree/master/results/betaH_celeba`

6 Conclusion

In this paper, we have proposed a new family of algorithms, NEO, for computing normalizing constants and sampling from complex distributions. This methodology comes with asymptotic and non-asymptotic convergence guarantees. For normalizing constant estimation, NEO-IS compares favorably to state-of-the-art algorithms on difficult benchmarks. NEO-MCMC is also very efficient for sampling complex distributions: it is particularly well-adapted to sampling multimodal distributions, thanks to its proposal mechanism which avoids being trapped in local modes. There are numerous potential extensions to this work. For example, it would be interesting to consider deterministic transformations other than conformal Hamiltonian dynamics integrators. These transformations could be trained, as for Neural IS, using a variation lower bound. It would also be interesting to further investigate the influence of the mixture weights $\{\varpi_k\}_{k\in\mathbb{Z}}$ on the efficiency of NEO.

References

Agakov, F. V. and Barber, D. (2004). An auxiliary variational method. In *International Conference on Neural Information Processing*, pages 561–566. Springer.

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Akyildiz, Ö. D. and Míguez, J. (2021). Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31(2):1–17.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C., Lee, A., Vihola, M., et al. (2018). Uniform ergodicity of the iterated conditional SMC and geometric ergodicity of particle Gibbs samplers. *Bernoulli*, 24(2):842–872.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2016). Importance weighted autoencoders. In *The 4th International Conference on Learning Representations (ICLR)*.

Che, T., Zhang, R., Sohl-Dickstein, J., Larochelle, H., Paull, L., Cao, Y., and Bengio, Y. (2020). Your GAN is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint arXiv:2003.06060*.

Chopin, N. and Robert, C. P. (2010). Properties of nested sampling. *Biometrika*, 97(3):741–755.

Craiu, R. V. and Lemieux, C. (2007). Acceleration of the multiple-try metropolis algorithm using antithetic and stratified sampling. *Statistics and computing*, 17(2):109.

Cremer, C., Morris, Q., and Duvenaud, D. (2017). Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*.

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

Ding, X. and Freedman, D. J. (2019). Learning deep generative models with annealed importance sampling. *arXiv preprint arXiv:1906.04904*.

Douc, R., Garivier, A., Moulines, E., Olsson, J., et al. (2011). Sequential monte carlo smoothing for general state space hidden markov models. *Annals of Applied Probability*, 21(6):2109–2145.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018). *Markov Chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

El Moselhy, T. A. and Marzouk, Y. M. (2012). Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850.

Franca, G., Sulam, J., Robinson, D. P., and Vidal, R. (2019). Conformal symplectic and relativistic optimization. *arXiv preprint arXiv:1903.04100*.

Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

Hartman, P. (1982). *Ordinary Differential Equations: Second Edition*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

Jia, H. and Seljak, U. (2020). Normalizing constant estimation with Gaussianized bridge sampling. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–14. PMLR.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*.

Lawson, D., Tucker, G., Dai, B., and Ranganath, R. (2019). Energy-inspired models: Learning with sampler-induced distributions. *arXiv preprint arXiv:1910.14265*.

Levy, D., Hoffman, M. D., and Sohl-Dickstein, J. (2018). Generalizing Hamiltonian Monte Carlo with neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Lindsten, F., Douc, R., and Moulines, E. (2015). Uniform ergodicity of the particle Gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. (2018). Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*.

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. (2019). Neural importance sampling. *ACM Transactions on Graphics*, 38(145).

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*.

Prangle, D. (2019). Distilling importance sampling. *arXiv preprint arXiv:1910.03632*.

Rotskoff, G. and Vanden-Eijnden, E. (2019). Dynamical computation of the density of states and Bayes factors using nonequilibrium importance sampling. *Physical Review Letters*, 122(15):150602.

Rubin, D. B. (1987). Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):542–543.

Ruiz, F. J., Titsias, M. K., Cemgil, T., and Doucet, A. (2020). Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. *arXiv preprint arXiv:2010.01845*.

Shestopaloff, A. Y., Neal, R. M., et al. (2018). Sampling latent states for high-dimensional non-linear state space models with the embedded hmm method. *Bayesian Analysis*, 13(3):797–822.

Skare, Ø., Bølviken, E., and Holden, L. (2003). Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737.

Skilling, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859.

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88.

So, M. K. (2006). Bayesian analysis of nonlinear and non-gaussian state space models via multiple-try sampling methods. *Statistics and Computing*, 16(2):125–141.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60.

Turner, R., Hung, J., Frank, E., Saatchi, Y., and Yosinski, J. (2019). Metropolis-Hastings generative adversarial networks. In *International Conference on Machine Learning*, pages 6345–6353. PMLR.

Wirnsberger, P., Ballard, A. J., Papamakarios, G., Abercrombie, S., Racanière, S., Pritzel, A., Rezende, D. J., and Blundell, C. (2020). Targeted free energy estimation via learned mappings. *arXiv preprint arXiv:2002.04913*.

Wu, Y., Burda, Y., Salakhutdinov, R., and Grosse, R. (2016). On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273*.

A Proofs

A.1 Additional notation

By abuse of notation, we denote by ρ and $\tilde{\pi}$ the probability measures with density with respect to the Lebesgue measure ρ and $\tilde{\pi}$ respectively.

A.2 Proof of (3)

The second expression of w_k follows from $\mathbf{J}_{\mathrm{T}^{-j}}(\mathrm{T}^k(x)) = \mathbf{J}_{\mathrm{T}^{k-j}}(x)/\mathbf{J}_{\mathrm{T}^k}(x)$ which implies

$$\begin{aligned} w_k(x) &= \varpi_k \rho(\mathrm{T}^k(x))/\sum\nolimits_{j\in\mathbb{Z}} \varpi_j \rho(\mathrm{T}^{k-j}(x))\mathbf{J}_{\mathrm{T}^{-j}}(\mathrm{T}^k(x)) \ , \\ &= \varpi_k \rho(\mathrm{T}^k(x))\mathbf{J}_{\mathrm{T}^k}(x)/\sum\nolimits_{j\in\mathbb{Z}} \varpi_j \rho(\mathrm{T}^{k-j}(x))\mathbf{J}_{\mathrm{T}^{k-j}}(x) = \varpi_k \rho_{-k}(x) \Big/ \sum\nolimits_{i\in\mathbb{Z}} \varpi_{k+i}\rho_i(x) \quad . \end{aligned}$$

A.3 Proof of Theorem 1

The unbiasedness of $\widehat{Z}_{X^{1:N}}$ follows directly from (2). Moreover, as $\widehat{Z}_{X^{1:N}}$ is unbiased and $E_{\mathrm{T}}^{\varpi} < \infty$, we can write

$$\mathrm{Var}_\rho[\widehat{Z}_X/Z] = \mathbb{E}_\rho[(\widehat{Z}_X/Z)^2] - 1 = E_{\mathrm{T}}^{\varpi} - 1 \ . \tag{18}$$

As $X^{1:N} \overset{\text{iid}}{\sim} \rho$, $\mathrm{Var}_\rho[\widehat{Z}_{X^{1:N}}/Z] = N^{-1}\,\mathrm{Var}_\rho[\widehat{Z}_X/Z]$. Finally, if $M_{\mathrm{T}}^{\varpi} < \infty$, then Hoeffding's inequality applies and we can write for any $\epsilon > 0$,

$$\mathbb{P}(|\widehat{Z}_{X^{1:N}}/Z - 1| > \epsilon) \leq 2\exp(-2N\epsilon^2/(M_{\mathrm{T}}^{\varpi})^2) \ . \tag{19}$$

Writing $\delta = 2\exp(-2N\epsilon^2/(M_{\mathrm{T}}^{\varpi})^2)$, we identify $\log(2/\delta) = 2N\epsilon^2/(M_{\mathrm{T}}^{\varpi})^2$ and $\epsilon = M_{\mathrm{T}}^{\varpi}\sqrt{\log(2/\delta)/(2N)}$. Plugging this expression of ϵ in (19) concludes the proof.

A.4 Proof of Theorem 2

We preface the proof of Theorem 2 with two auxiliary lemmas.

Lemma 6. *Let A, B be two integrable random variables satisfying $|A/B| \leq M$ almost surely and denote $a = \mathbb{E}[A]$, $b = \mathbb{E}[B]$. Then,*

$$|\mathbb{E}[A/B] - a/b| \leq \frac{\sqrt{\mathrm{Var}(A/B)\,\mathrm{Var}(B)}}{b} \ , \tag{20}$$

$$\mathrm{Var}(A/B) \leq \mathbb{E}\left[|A/B - a/b|^2\right] \leq \frac{2}{B^2}\left(\mathbb{E}\left[|A_N - A|^2\right] + M^2\mathbb{E}\left[|B_N - B|^2\right]\right) \ . \tag{21}$$

Proof. Write first, using the Cauchy-Schwarz inequality,

$$\begin{aligned} \left|\mathbb{E}\left[\frac{A}{B}\right] - \frac{a}{b}\right| &= \left|\mathbb{E}\left[\frac{A}{B}\right] - \frac{\mathbb{E}\left[A\right]}{b}\right| = \left|\mathbb{E}\left[A\left(\frac{1}{B} - \frac{1}{b}\right)\right]\right| \ , \\ &= \left|\mathbb{E}\left[\frac{A}{B}\left(\frac{b-B}{b}\right)\right]\right| = \left|\mathbb{E}\left[\left(\frac{A}{B} - \mathbb{E}\left[\frac{A}{B}\right]\right)\left(\frac{B-b}{b}\right)\right]\right| \ , \\ &\leq \frac{\sqrt{\mathrm{Var}(A/B)}\sqrt{\mathrm{Var}(B)}}{b} \ . \end{aligned}$$

Moreover, using $|A/B| \leq M$ yields

$$\left|\frac{A}{B} - \frac{a}{b}\right| = \left|\frac{1}{b}(A-a) + A\left(\frac{1}{B} - \frac{1}{b}\right)\right| \leq \frac{1}{b}|A-a| + \frac{|A|}{Bb}|B-b| \,, \\ \leq \frac{1}{b}|A-a| + \frac{M}{b}|B-b| \,.$$

Therefore,

$$|A/B - a/b|^2 \leq \frac{2}{b^2}\left(|A-a|^2 + M^2|B-b|^2\right) \,,$$

Using that $\mathbb{E}\left[|A/B - a/b|^2\right] = \mathrm{Var}(A/B) + |\mathbb{E}[A/B] - a/b|^2$ concludes the proof. □

We get the following lemma from (Douc et al., 2011, Lemma 4).

Lemma 7. *Assume that A and B are random variables and that there exist positive constants b, M, C, K such that*

(i) $|A/B| \leq M$, $\mathbb{P}$*-a.s.*,

(ii) for all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left(|B-b| > \epsilon\right) \leq K\exp(-R\epsilon^2)$,

(iii) for all $\epsilon > 0$ *and all* $N \geq 1$, $\mathbb{P}\left(|A| > \epsilon\right) \leq K\exp\left(-R\epsilon^2/M^2\right)$,

then,

$$\mathbb{P}(|A/B| \geq \epsilon) \leq 2K\exp(-Rb^2\epsilon^2/4M^2) \,.$$

Proof. By the triangle inequality,

$$|A/B| = \left|\frac{A}{B}(b-B)b^{-1} + b^{-1}A\right| \,, \\ \leq b^{-1}\left|A/B\right|\left|b-B\right| + b^{-1}\left|A\right| \leq Mb^{-1}\left|b-B\right| + b^{-1}\left|A\right| \,.$$

Therefore,

$$\{|A/B| \geq \epsilon\} \subseteq \left\{|B-b| \geq \frac{\epsilon b}{2M}\right\} \cup \left\{|A| \geq \frac{\epsilon b}{2}\right\} \,.$$

Then, conditions (ii) and (iii) imply that

$$\mathbb{P}\left(|A/B| \geq \epsilon\right) \leq \mathbb{P}\left(|B-b| \geq \frac{\epsilon b}{2M}\right) + \mathbb{P}\left(|A| \geq \frac{\epsilon b}{2}\right) \,, \\ \leq 2K\exp(-Rb^2\epsilon^2/(4M^2)) \,.$$

□

Proof of Theorem 2. Let $g : \mathbb{R}^d \to \mathbb{R}$ such that $\sup_{x\in\mathbb{R}^d}|g|(x) \leq 1$ and denote $\pi(g) = \int g\mathrm{d}\pi$. We use Lemma 6 with $A = A_N$ and $B = \widehat{Z}_{X^{1:N}}$ where

$$A_N = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))g(\mathrm{T}^k(X^i)) \,,\ \widehat{Z}_{X^{1:N}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}} w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i)) \,. \tag{22}$$

By construction, since $\sup_{x\in\mathbb{R}^d}|g|(x)\leq 1$, almost surely $A_N/\widehat{Z}_{X^{1:N}}\leq 1$ and $\mathrm{Var}(\widehat{Z}_{X^{1:N}})=N^{-1}\mathrm{Var}(\widehat{Z}_{X^1})$. Then, using (2) with $a=\mathbb{E}[A_N]=Z\pi(g)$ and $b=\mathbb{E}[\widehat{Z}_{X^{1:N}}]=Z$, Lemma 6 implies

$$\left|J^{\mathrm{NEO}}_{\varpi,N}(g)-\pi(g)\right|=\left|\mathbb{E}[A_N/\widehat{Z}_{X^{1:N}}]-a/b\right|\leq N^{-1/2}\sqrt{\mathrm{Var}(A_N/\widehat{Z}_{X^{1:N}})\mathrm{Var}(\widehat{Z}_{X^1})}\,. \tag{23}$$

On the other hand,

$$\mathbb{E}\left[|A_N-a|^2\right]=N^{-1}\mathbb{E}_{X\sim\rho}\left[\left\{\sum_{k\in\mathbb{Z}}w_k(X)\mathrm{L}(\mathrm{T}^k(X))g(\mathrm{T}^k(X))-Z\pi(g)\right\}^2\right]\leq N^{-1}Z^2E^{\varpi}_{\mathrm{T}}\,.$$

These inequalities yield using $\mathrm{Var}(\widehat{Z}_{X^1})\leq E^{\varpi}_{\mathrm{T}}$ and Lemma 6 again:

$$\begin{aligned}\mathbb{E}\left[|J^{\mathrm{NEO}}_{\varpi,N}(g)-\pi(g)|^2\right]&\leq\frac{2}{N}(E^{\varpi}_{\mathrm{T}}+\mathrm{Var}(\widehat{Z}_{X^1}))\leq\frac{4}{N}E^{\varpi}_{\mathrm{T}}\,,\\ |\mathbb{E}\left[J^{\mathrm{NEO}}_{\varpi,N}(g)-\pi(g)\right]|&\leq\frac{\sqrt{2(E^{\varpi}_{\mathrm{T}}+\mathrm{Var}(\widehat{Z}_{X^1}))\mathrm{Var}(\widehat{Z}_{X^1})}}{N}\leq\frac{2E^{\varpi}_{\mathrm{T}}}{N}\,,\end{aligned}$$

which concludes the proof.

Define

$$\tilde{A}_N=N^{-1}\sum_{i=1}^{N}\sum_{k\in\mathbb{Z}}w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\left(g(\mathrm{T}^k(X^i))-\pi(g)\right)\,.$$

With this notation, the proof of (9) relies on the application of Lemma 7 to $A=\tilde{A}_N$ and $B=\widehat{Z}_{X^{1:N}}$, since

$$J^{\mathrm{NEO}}_{\varpi,N}(g)-\pi(g)=A_N/\widehat{Z}_{X^{1:N}}\,.$$

As $\sup_{x\in\mathbb{R}^d}|g|(x)\leq 1$, we get that $\tilde{A}_N/\widehat{Z}_{X^{1:N}}\leq 2$. By (2), $\mathbb{E}[\widehat{Z}_{X^{1:N}}]=Z$ and $\widehat{Z}_{X^{1:N}}=N^{-1}\sum_{i=1}^{N}W_i$ with $W_i=\sum_{k\in\mathbb{Z}}w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\leq M^{\varpi}_{\mathrm{T}}$. Then, by Hoeffding's inequality, for all $\varepsilon>0$,

$$\mathbb{P}(|B_N-Z|>\varepsilon)\leq 2\exp(-2N(\varepsilon/M^{\varpi}_{\mathrm{T}})^2)\,.$$

Similarly, A_N is centered and $A_N=N^{-1}\sum_{i=1}^{N}U_i$ with

$$U_i=\sum_{k\in\mathbb{Z}}w_k(X^i)\mathrm{L}(\mathrm{T}^k(X^i))\{g(\mathrm{T}^k(X^i))-\pi(g)\}$$

and $|U_i|\leq 2M^{\varpi}_{\mathrm{T}}$ almost surely. By Hoeffding's inequality, for all $\varepsilon>0$,

$$\mathbb{P}(|A_N|>\epsilon)\leq 2\exp(-N\varepsilon^2/(8(M^{\varpi}_{\mathrm{T}})^2))\,.$$

The assumptions of Lemma 7 are met so that

$$\mathbb{P}(|J^{\mathrm{NEO}}_{\varpi,N}(g)-\pi(g)|>\varepsilon)\leq 4\exp(-\varepsilon^2NZ^2/[32(M^{\varpi}_{\mathrm{T}})^2])\,,$$

which concludes the proof. □

A.5 Proof of Lemma 3

As $w_k(x) = \varpi_k \rho(\mathrm{T}^k(x))/\{\Omega \rho_{\mathrm{T}}(\mathrm{T}^k(x))\}$, by Jensen's inequality,

$$\begin{aligned} E_{\mathrm{T}}^{\varpi} = \int \left(\sum_{k \in \mathbb{Z}} w_k(x) \mathrm{L}(\mathrm{T}^k(x))/Z \right)^2 \rho(x) \mathrm{d}x &= \int \left(\sum_{k \in \mathbb{Z}} \frac{\varpi_k}{\Omega} \frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x) \mathrm{d}x \ , \\ &\leq \int \sum_{k \in \mathbb{Z}} \frac{\varpi_k}{\Omega} \left(\frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x) \mathrm{d}x \ , \\ &\leq \Omega^{-1} \sum_{k \in \mathbb{Z}} \varpi_k \int \left(\frac{\pi(\mathrm{T}^k(x))}{\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right)^2 \rho(x) \mathrm{d}x \ . \end{aligned}$$

Using the change of variables $y = \mathrm{T}^k(x)$ yields, by (1),

$$E_{\mathrm{T}}^{\varpi} \leq \Omega^{-1} \sum_{k \in \mathbb{Z}} \varpi_k \int \left(\frac{\pi(y)}{\rho_{\mathrm{T}}(y)} \right)^2 \rho(\mathrm{T}^{-k}(y)) \mathbf{J}_{\mathrm{T}^{-k}}(y) \mathrm{d}y \leq \int \left(\frac{\pi(y)}{\rho_{\mathrm{T}}(y)} \right)^2 \rho_{\mathrm{T}}(y) \mathrm{d}y \ .$$

A.6 Proofs of NEO MCMC sampler

Proof of Theorem 4. Note first that by symmetry, we have

$$P(y, \mathsf{A}) = N^{-1} \int \sum_{i=1}^{N} \delta_y(\mathrm{d}x^i) \prod_{j=1, j \neq i}^{N} \rho(x^j) \mathrm{d}x^j \sum_{k=1}^{N} \frac{\widehat{Z}_{x^k}}{\sum_{j=1}^{N} \widehat{Z}_{x^j}} \mathbb{1}_{\mathsf{A}}(x^k) \ . \tag{24}$$

We begin with the proof of reversibility of P with respect to $\tilde{\pi}$. Let f, g be nonnegative measurable functions. By definition of P,

$$\begin{aligned} \int \tilde{\pi}(\mathrm{d}y) P(y, \mathrm{d}y') f(y) g(y') &= \frac{1}{NZ} \int \sum_{i=1}^{N} \rho(\mathrm{d}y) \widehat{Z}_y f(y) \delta_y(\mathrm{d}x^i) \prod_{l=1, l \neq i}^{N} \rho(\mathrm{d}x^l) \sum_{k=1}^{N} \frac{\widehat{Z}_{x^k}}{\sum_{j=1}^{N} \widehat{Z}_{x^j}} g(x^k) \ , \\ &= \frac{1}{NZ} \int \sum_{i=1}^{N} \widehat{Z}_{x^i} f(x^i) \prod_{l=1}^{N} \rho(\mathrm{d}x^l) \sum_{k=1}^{N} \frac{\widehat{Z}_{x^k}}{\sum_{j=1}^{N} \widehat{Z}_{x^j}} g(x^k) \ , \\ &= \frac{1}{NZ} \int \prod_{l=1}^{N} \rho(\mathrm{d}x^l) \frac{\sum_{i=1}^{N} \widehat{Z}_{x^i} f(x^i) \sum_{k=1}^{N} \widehat{Z}_{x^k} g(x^k)}{\sum_{j=1}^{N} \widehat{Z}_{x^j}} \ , \\ &= \int \tilde{\pi}(\mathrm{d}y) P(y, \mathrm{d}y') f(y') g(y) \ , \end{aligned}$$

which shows that P is $\tilde{\pi}$-reversible. We now establish that P is $\tilde{\pi}$-irreducible. We have for $y \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}
P(y, \mathsf{A}) &= \int \delta_y(\mathrm{d}x^1) \sum_{i=1}^N \frac{\widehat{Z}_{x^i}}{N \widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^N \rho(\mathrm{d}x^j) \\
&= \int \frac{\widehat{Z}_y}{\widehat{Z}_y + \sum_{j=2}^N \widehat{Z}_{x^j}} \mathbb{1}_{\mathsf{A}}(x) \prod_{j=2}^N \rho(\mathrm{d}x^j) + \int \sum_{i=2}^N \frac{\widehat{Z}_{x^i}}{\widehat{Z}_y + \sum_{j=2}^N \widehat{Z}_{x^j}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^N \rho(\mathrm{d}x^j) \\
&\geq \sum_{i=2}^N \int \frac{\widehat{Z}_{x^i}}{\widehat{Z}_y + \widehat{Z}_{x^i} + \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j}} \mathbb{1}_{\mathsf{A}}(x^i) \prod_{j=2}^N \rho(\mathrm{d}x^j) \\
&\geq \sum_{i=2}^N \int \tilde{\pi}(\mathrm{d}x^i) \mathbb{1}_{\mathsf{A}}(x^i) \int \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j}} \prod_{j=2, j\neq i}^N \rho(\mathrm{d}x^j) \, .
\end{aligned}$$

Since the function $f \colon z \mapsto (z+a)^{-1}$ is convex on $\mathbb{R}_+$ for $a > 0$, we get for $i \in \{2, \ldots, N\}$,

$$\begin{aligned}
\int \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j}} \prod_{j=2, j\neq i}^N \rho(\mathrm{d}x^j) &\geq \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + \int \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j} \prod_{j=2, j\neq i}^N \rho(\mathrm{d}x^j)} \\
&\geq \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + Z(N-2)} \, . \qquad (25)
\end{aligned}$$

Therefore, for $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ satisfying $\tilde{\pi}(\mathsf{A}) > 0$, we get $P(y, \mathsf{A}) > 0$ for any $y \in \mathbb{R}^d$ since $\widehat{Z}_x < \infty$ for any $x \in \mathbb{R}^d$. By definition, P is $\tilde{\pi}$-irreducible.

We show that P is Harris recurrent using (Tierney, 1994, Corollary 2). To this end, since P is $\tilde{\pi}$-irreducible, it is sufficient to show that P is a Metropolis type kernel. Define $\alpha(x^1, x^2) = (N-1) \int \prod_{j=3}^N \rho(\mathrm{d}x^j) \widehat{Z}_{x^2} / \sum_{j=1}^N \widehat{Z}_{x^j}$ for $x^1, x^2 \in \mathbb{R}^d$ and $\rho_{2:N}(\mathrm{d}x^{2:N}) = \{\prod_{j=2}^N \rho_{2:N}(x^j)\} \mathrm{d}x^{2:N}$. Then, by (12), we get with this notation, for $y \in \mathbb{R}^d$, $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$\begin{aligned}
&P(y, \mathsf{A}) \\
&= \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \sum_{i=2}^N \frac{\widehat{Z}_{x^i}}{N \widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^i) + \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{Z}_{x^1}}{N \widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^1) \\
&= \sum_{i=2}^N \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{Z}_{x^i}}{N \widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^i) + \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{Z}_{x^1}}{N \widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^1) \\
&= \sum_{i=2}^N \int \delta_y(\mathrm{d}x^1) \rho(\mathrm{d}x^i) \int \prod_{j=2, j\neq i}^N \rho(x^j) \mathrm{d}x^j \frac{\widehat{Z}_{x^i} \mathbb{1}_{\mathsf{A}}(x^i)}{N \widehat{Z}_{x^{1:N}}} + \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \frac{\widehat{Z}_{x^1} \mathbb{1}_{\mathsf{A}}(x^1)}{N \widehat{Z}_{x^{1:N}}} \\
&= \sum_{i=2}^N \int \frac{\alpha(y, x^i)}{(N-1)} \mathbb{1}_{\mathsf{A}}(x^i) \rho(\mathrm{d}x^i) + \int \delta_y(\mathrm{d}x^1) \rho_{2:N}(\mathrm{d}x^{2:N}) \left\{ 1 - \sum_{i=2}^N \frac{\widehat{Z}_{x^i}}{N \widehat{Z}_{x^{1:N}}} \right\} \mathbb{1}_{\mathsf{A}}(x^1) \\
&= \int_{\mathsf{A}} \alpha(y, y') \rho(y') \mathrm{d}y' + \left(1 - \int \alpha(y, y') \rho(y') \mathrm{d}y' \right) \delta_y(\mathsf{A}) \, . \qquad (26)
\end{aligned}$$

With the terminology of (Tierney, 1994, Corollary 2), P is Metropolis type kernel and therefore is Harris recurrent.

Note that Algorithm 2 defines a Markov chain $\{Y_i, U_i\}_{i\in\mathbb{N}}$ taking for U_0 an arbitrary initial point with Markov kernel denoted by $\tilde{P}$. By abuse of notation, we denote by $\{Y_i, U_i\}_{i\in\mathbb{N}}$ the canonical process on the canonical space $(\mathbb{R}^d \times \mathbb{R}^d)^{\mathbb{N}}$ endowed with the corresponding σ-field and denote by $\mathbb{P}_{y,u}$ the distribution associated with the Markov chain with kernel $\tilde{P}$ and initial distribution $\delta_y \otimes \delta_u$. Denote for any $y \in \mathbb{R}^d$ by $\mathbb{P}_y$ the marginal distribution of $\mathbb{P}_{y,u}$ with respect to $\{Y_i\}_{i\in\mathbb{N}}$, *i.e.* $\mathbb{P}_y(\mathsf{A}) = \mathbb{P}_{(y,u)}(\{Y_i\}_{i\in\mathbb{N}} \in \mathsf{A})$ for $u \in \mathbb{R}^d$, noting that by definition, $\mathbb{P}_{(y,u)}(\mathsf{A} \times (\mathbb{R}^d)^{\mathbb{N}})$ does not depend on u. In addition, under $\mathbb{P}_y$, $\{Y_i\}_{i\in\mathbb{N}}$ is a Markov chain associated with P. Therefore, since P is $\tilde{\pi}$-irreducible and Harris recurrent, we get by (Douc et al., 2018, Theorem 11.3.1) and (Tierney, 1994, Theorem 2, 3) for any $y \in \mathbb{R}^d$, $\lim_{k\to\infty} \|\delta_y P^k - \tilde{\pi}\|_{\mathrm{TV}} = 0$ and for any bounded and measurable function g,

$$n^{-1} \sum_{k=1}^{n} g(Y_k) = \tilde{\pi}(g) \;, \qquad \mathbb{P}_y\text{-almost surely} \;. \tag{27}$$

We now turn to proving the properties regarding Q. For any $\mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$, using (2), we obtain

$$\int \tilde{\pi}(y) Q(y, \mathsf{B}) \mathrm{d}y = Z^{-1} \int \rho(y) \sum_{k\in\mathbb{Z}} w_k(y) \mathrm{L}(\mathrm{T}^k(y)) \mathbb{1}_{\mathsf{B}}(\mathrm{T}^k(y)) \mathrm{d}y = \pi(\mathsf{B}) \;.$$

Using for all $y \in \mathbb{R}^d$, $\lim_{n\to\infty} \|P^n(y,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} = 0$, we get $\lim_{n\to\infty} \|P^n Q(y,\cdot) - \pi\|_{\mathrm{TV}} = 0$. It remains to show the stated Law of Large Numbers. Let $y, u \in \mathbb{R}^d$ and g be a bounded measurable function. Define for any $i \in \mathbb{N}^*$, $\tilde{U}_i = g(U_i) - Qg(Y_i)$. By definition, for any $i \in \mathbb{N}^*$, $\left|\tilde{U}_i\right| \leq 2 \sup_{x\in\mathbb{R}^d} |g(x)|$ and $\mathbb{E}_{(y,u)}[\tilde{U}_i | \mathcal{F}_{i-1}] = 0$, where $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ is the canonical filtration. Therefore, $\{\tilde{U}_i\}_{i\in\mathbb{N}^*}$ are $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$-martingale increments and $\{S_k = \sum_{i=1}^k \tilde{U}_i\}_{k\in\mathbb{N}}$ is a $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$-martingale. Using (Hall and Heyde, 1980, Theorem 2.18), we get

$$\lim_{n\to\infty} \{S_n/n\} = 0 \;, \quad \mathbb{P}_{(y,u)}\text{-almost surely} \;. \tag{28}$$

The proof is completed using that $\lim_{n\to\infty} \{n^{-1} \sum_{i=1}^n Qg(Y_i)\} = \tilde{\pi}(Qg) = \pi(g)$, $\mathbb{P}_y$-almost surely by (27) and therefore by definition, $\mathbb{P}_{(y,u)}$-almost surely. □

Proof of Theorem 5. We have for $(x, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$,

$$P(y, \mathsf{A}) \geq \sum_{i=2}^{N} \int \tilde{\pi}(\mathrm{d}x^i) \mathbb{1}_{\mathsf{A}}(x^i) \int \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j}} \prod_{j=2, j\neq i}^{N} \rho(\mathrm{d}x^j) \;.$$

Moreover, as for any $x \in \mathbb{R}^d$, $\widehat{Z}_x / Z \leq M_{\mathrm{T}}^{\varpi}$,

$$\int \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + \sum_{j=2, j\neq i}^N \widehat{Z}_{x^j}} \prod_{j=2, j\neq i}^{N} \rho(\mathrm{d}x^j) \geq \frac{Z}{\widehat{Z}_y + \widehat{Z}_{x^i} + Z(N-2)} \geq \frac{1}{2M_{\mathrm{T}}^{\varpi} + N - 2} \;.$$

We finally obtain the inequality

$$P(x, \mathsf{A}) \geq \tilde{\pi}(\mathsf{A}) \times \frac{N-1}{2M_{\mathrm{T}}^{\varpi} + N - 2} = \epsilon_N \tilde{\pi}(\mathsf{A}) \;. \tag{29}$$

The proof for P is concluded from (Douc et al., 2018, Theorem 18.2.4).

As $\|P^k(y,\cdot) - \tilde{\pi}\|_{\mathrm{TV}} \leq \kappa_N^k$, for any bounded function f, $\|f\|_\infty \leq 1$, we have $|P^k f(y) - \tilde{\pi}(f)| \leq \kappa_N^k$, by definition of the Total Variation Distance. Then, writing $f = Qg$ for any bounded function g, $\|g\|_\infty \leq 1$, we have $\|f\|_\infty \leq 1$ and

$$|P^k f(y) - \tilde{\pi}(f)| = |P^k Qg(y) - \tilde{\pi}Q(g)| = |P^k Qg(y) - \pi(g)| \leq \kappa_N^k \ . \tag{30}$$

□

Write now P the Markov kernel extending to correlated proposals: for $y \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$,

$$P(y, \mathsf{A}) = N^{-1} \int \sum_{i=1}^N \delta_y(\mathrm{d}x^i) r_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^N \frac{\widehat{Z}_{x^k}}{N\widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \ , \tag{31}$$

where the Markov kernels R_i are defined by $R_i(x^i, \mathrm{d}x^{1:N\setminus\{i\}}) = r_i(x^i, x^{1:N\setminus\{i\}})\mathrm{d}x^{1:N\setminus\{i\}}$ and r_i by (15).

Theorem 8. *P is $\tilde{\pi}$-invariant.*

Proof. Define the Nd-dimensional probability measure $\bar{\rho}_N(\mathrm{d}x^{1:N}) = \rho(\mathrm{d}x^1)R_1(x^1, \mathrm{d}x^{2:n})$. Let $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$. Then, we have

$$\begin{aligned}
\tilde{\pi}P(\mathsf{A}) &= N^{-1} \int \tilde{\pi}(\mathrm{d}y) \int \sum_{i=1}^N \delta_y(\mathrm{d}x^i) R_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^N \frac{\widehat{Z}_{x^k}}{N\widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (NZ)^{-1} \int \sum_{i=1}^N \rho(\mathrm{d}x^i) \widehat{Z}_{x^i} R_i(x^i, \mathrm{d}x^{1:n\setminus\{i\}}) \sum_{k=1}^N \frac{\widehat{Z}_{x^k}}{N\widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (NZ)^{-1} \int \bar{\rho}_N(\mathrm{d}x^{1:N}) \sum_{i=1}^N \widehat{Z}_{x^i} \sum_{k=1}^N \frac{\widehat{Z}_{x^k}}{N\widehat{Z}_{x^{1:N}}} \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (NZ)^{-1} \int \sum_{k=1}^N \widehat{Z}_{x^k} \bar{\rho}_N(\mathrm{d}x^{1:N}) \mathbb{1}_{\mathsf{A}}(x^k) \\
&= (NZ)^{-1} \int \sum_{k=1}^N \widehat{Z}_{x^k} \rho(\mathrm{d}x^k) \mathbb{1}_{\mathsf{A}}(x^k) = \tilde{\pi}(\mathsf{A}) \ .
\end{aligned}$$

□

B Continuous-time limit of NEO and NEIS

B.1 Proof for the continuous-time limit

Consider $\bar{h} > 0$ and a family $\{\mathrm{T}_h \ : \ h \in \left(0, \bar{h}\right]\}$ of C^1-diffeomorphisms. For $N \in \mathbb{N}^*$ and a bounded and continuous $f : \mathbb{R}^d \to \mathbb{R}$, write

$$I^{\mathrm{NEO}}_{\varpi,N,h}(f) = N^{-1} \sum_{i=1}^N \sum_{k\in\mathbb{Z}} w_{k,h}(X^i) f(\mathrm{T}_h^k(X^i)) \ , \tag{32}$$

where $\{X_i\}_{i=1}^N \overset{\text{iid}}{\sim} \rho$ and for some weight function $\varpi^{\mathrm{c}} : \mathbb{R} \to \mathbb{R}_+$ with bounded support (see **H**3), $k \in \mathbb{Z}$ and $h > 0$, setting $\varpi_{k,h} = \varpi^{\mathrm{c}}(kh)$,

$$w_{k,h}(x) = \varpi_{k,h}\rho_{-k}(x) \Big/ \sum\nolimits_{i\in\mathbb{Z}} \varpi_{k+i,h}\rho_i(x) \quad . \tag{33}$$

We show in this section the convergence of the sequence of NEO-IS estimators $\{I^{\mathrm{NEO}}_{\varpi,N,h}(f) \ : \ h \in \big(0, \bar{h}\big]\}$ as $h \downarrow 0$ to its continuous counterpart, the version (16) of NEIS Rotskoff and Vanden-Eijnden (2019), with weight function ϖ, in the case where for any $h \in \big(0, \bar{h}\big]$, T_h corresponds to one step of a discretization scheme with stepsize h of the ODE

$$\dot{x}_t = b(x_t) \ , \tag{34}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is a drift function. We are particularly interested in the case where (34) corresponds to the conformal Hamilonian dynamics (10) and $\{\mathrm{T}_h \ : \ h \in \big(0, \bar{h}\big]\}$ to its conformal symplectic Euler discretization: for all $(q, p) \in \mathbb{R}^{2d}$,

$$\mathrm{T}_h(q, p) = (q + h\mathrm{M}^{-1}\{\mathrm{e}^{-h\gamma}p - h\nabla U(q)\}, \mathrm{e}^{-h\gamma}p - h\nabla U(q)) \ . \tag{35}$$

We make the following conditions on b, ρ, ϖ^{c} and $\{\mathrm{T}_h \ : \ h \in \big(0, \bar{h}\big]\}$.

H1. *The function b is continuously differentiable and L_b-Lipschitz.*

Under **H**1, consider $(\phi_t)_{t\geq 0}$ the differential flow associated with (34), *i.e.* $\phi_t(x) = x_t$ where $(x_t)_{t\in\mathbb{R}}$ is the solution of (34) starting from x. Note that **H**1 implies that $(t, x) \mapsto \phi_t(x)$ is continuously differentiable on $\mathbb{R} \times \mathbb{R}^d$, see (Hartman, 1982, Theorem 4.1 Chapter V).

H1 is satisfied in the case of the conformal Hamiltonian dynamics if the potential U is continuously differentiable and with Lipschitz gradient, that is there exists $L_U \in \mathbb{R}_+^*$ such that for any $x_1, x_2 \in \mathbb{R}^d$, $\|\nabla U(x_1) - \nabla U(x_2)\| \leq L_U\|x_1 - x_2\|$.

H2. *For any $h \in \big(0, \bar{h}\big]$, $\mathrm{T}_h : \mathbb{R}^d \to \mathbb{R}^d$ is a C^1-diffeomorphism. In addition, it holds:*

(i) there exist $C \geq 0$ and $\delta \in (0, 1]$ such that for any $x \in \mathbb{R}^d$,

$$\|\, \mathrm{T}_h(x) - (x + hb(x))\| \leq Ch^{1+\delta}(1 + \|x\|) \ ;$$

(ii) for any $x \in \mathbb{R}^d$ and $T \in \mathbb{R}_+^*$,

$$\lim_{h\downarrow 0} \max_{k\in[-\lfloor T/h\rfloor:\lfloor T/h\rfloor]} \|\mathbf{J}_{\phi_{kh}}(x) - \mathbf{J}_{\mathrm{T}_h^k}(x)\| = 0 \ .$$

Note that **H**2 is automatically satisfied for the conformal symplectic Euler discretization (35) of the conformal Hamiltonian dynamics. Indeed, in that case $\operatorname{div} b(\phi_t(x)) = \gamma d$, and therefore $\mathbf{J}_{\phi_t}(x) = e^{\gamma dt}$ for $t \in \mathbb{R}$, and for any $h > 0, k \in \mathbb{Z}$, $\mathbf{J}_{\mathrm{T}_h^k}(x) = e^{\gamma dhk}$; see Franca et al. (2019).

Define

$$\operatorname{support}(\varpi^{\mathrm{c}}) = \{t \in \mathbb{R} \ : \ \varpi^{\mathrm{c}}(t) \neq 0\} \ . \tag{36}$$

H3. *(i) ρ is continuous and positive on $\mathbb{R}^d$*

(ii) ϖ^{c} is piecewise continuous on $\mathbb{R}$, its support $\operatorname{support}(\varpi^{\mathrm{c}})$ is bounded and $\sup_{(s,t)\in\mathsf{A}_\varpi} \varpi^{\mathrm{c}}(t)/\varpi^{\mathrm{c}}(t+s) = m < \infty$ where

$$\mathsf{A}_\varpi = \{(s, t) \in \mathbb{R}^2; \ t \in \operatorname{support}(\varpi^{\mathrm{c}}), \ (s + t) \in \operatorname{support}(\varpi^{\mathrm{c}})\} \ .$$

(iii) Moreover, for any $x \in \mathbb{R}^d$, we have $\rho^c_{\mathrm{T}}(x) = \int \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t > 0$.

Note that **H**3 implies that $\sup_{t\in\mathbb{R}} |\varpi^{\mathrm{c}}(t)| < +\infty$. **H**3 is automatically satisfied for example in the case $\varpi^{\mathrm{c}} = \mathbb{1}_{[-T_1,T_2]}$ for $T_1, T_2 \geq 0$.

Theorem 9. *Assume **H**1, **H**2, **H**3. For any $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ continuous and bounded,*

$$\lim_{h\downarrow 0} \left| \sum_{k\in\mathbb{Z}} w_{k,h}(x) f(\mathrm{T}_h^k(x)) - \int_{-\infty}^{\infty} w_t^{\mathrm{c}}(x) f(\phi_t(x))\mathrm{d}t \right| = 0 \ ,$$

where $\{w_{k,h}\}_{k\in\mathbb{Z}}$ and w_t^{c} are defined in (33) *and* (17) *respectively, i.e. for $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$,*

$$w_t^{\mathrm{c}}(x) = \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) \Big/ \int_{-\infty}^{\infty} \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s \ . \tag{37}$$

Proof. Let f be a bounded continuous function, $x \in \mathbb{R}^d$. Setting

$$\begin{aligned} g_{k,h}(x) &= \rho(\mathrm{T}_h^k(x))\varpi^{\mathrm{c}}(kh)\mathbf{J}_{\mathrm{T}_h^k}(x) f(\mathrm{T}_h^k(x)) \\ h\Delta_{k,h}(x) &= h\sum_{i\in\mathbb{Z}} \rho(\mathrm{T}_h^i(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}_h^i(x)} \ , \end{aligned}$$

we have that

$$\sum_{k\geq 0} \frac{h g_{k,h}(x)}{h\Delta_{k,h}(x)} = \int_0^{T_\varpi} \frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)} g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t + \int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h} \frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)} g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t \ ,$$

as $g_{k,h}(x) = 0$ when $k > \lfloor T_\varpi/h\rfloor$. Therefore, we can consider the following decomposition,

$$\left| \sum_{k\geq 0} \frac{\rho(\mathrm{T}_h^k(x))\varpi^{\mathrm{c}}(kh)\mathbf{J}_{\mathrm{T}_h^k}(x) f(\mathrm{T}_h^k(x))}{\sum_{i\in\mathbb{Z}} \rho(\mathrm{T}_h^i(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}_h^i(x)}} - \int_0^{T_\varpi} \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))\mathrm{d}t}{\int \varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s} \right| \leq A + B$$

with

$$\begin{aligned} A = &\left| \int_0^{T_\varpi} \frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)} \left\{ g_{\lfloor t/h\rfloor,h}(x) - \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x)) \right\} \mathrm{d}t \right| \\ &+ \left| \int_{T_\varpi}^{h\lfloor T_\varpi/h\rfloor+h} \frac{1}{h\Delta_{\lfloor t/h\rfloor,h}(x)} g_{\lfloor t/h\rfloor,h}(x)\mathrm{d}t \right| \ , \end{aligned}$$

and

$$B = \int_0^{T_\varpi} \left| \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))\mathrm{d}t}{h\Delta_{\lfloor t/h\rfloor,h}(x)} - \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{\int \varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\mathrm{d}s} \right| \mathrm{d}t \ ,$$

We bound those terms separately. First of all, under **H**3-(ii), for any k such that $kh \in [0, T_\varpi]$, we have $h\Delta_{k,h}(x) \geq hm^{-1}\Delta_{0,h}(x)$. Second, as $\lim_{h\downarrow 0} h\Delta_{0,h}(x) = \int_0^{T_\varpi} \rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)\varpi^{\mathrm{c}}(s)\mathrm{d}s > 0$, there exists some $\tilde{h} > 0$ and $c > 0$ such that for all $k \in \mathbb{Z}$, $h < \tilde{h}$ implies

$$\int_0^{T_\varpi} \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t > c \ , \quad h\Delta_{k,h}(x) \geq hm^{-1}\Delta_{0,h}(x) > c \ . \tag{38}$$

Then, for $h < \tilde{h}$,

$$A \leq c^{-1} \int_0^{T_\varpi} |g_{\lfloor t/h \rfloor, h}(x) - \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))| \mathrm{d}t + c^{-1} \int_{T_\varpi}^{h\lfloor T_\varpi/h \rfloor + h} \left| g_{\lfloor t/h \rfloor, h}(x) \right| \mathrm{d}t \; .$$

By **H**1 and **H**3, the function $t \to \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))$ is continuous on the compact $[0, 2T_\varpi]$ and thus is bounded. Therefore, for any $h \in \left(0, \bar{h}\right)$,

$$\sup_{t \in [0, 2T_\varpi]} |\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))| \leq \sup_{t \in \mathbb{R}} |\varpi^{\mathrm{c}}| \sup_{x \in \mathbb{R}^d} |f(x)| \sup_{t \in [0, 2T_\varpi]} |\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)| < \infty \; . \quad (39)$$

Under **H**2, (39) and Lemma 13 imply that

$$\sup_{t \in [0, h\lfloor T_\varpi/h \rfloor + h)} g_{\lfloor t/h \rfloor, h}(x) \leq \sup_{t \in \mathbb{R}} |\varpi^{\mathrm{c}}(t)| \sup_{x \in \mathbb{R}^d} |f(x)| \sup_{t \in [0, h\lfloor T_\varpi/h \rfloor + h)} \rho(\mathrm{T}_h^{\lfloor t/h \rfloor}(x)) \mathbf{J}_{\mathrm{T}_h^{\lfloor t/h \rfloor}}(x) < \infty \; ,$$

Then, $\lim_{h \downarrow 0} \int_{T_\varpi}^{h\lfloor T_\varpi/h \rfloor + h} \left| g_{\lfloor t/h \rfloor, h}(x) \right| \mathrm{d}t = 0$. Finally, Lemma 14 implies that $\lim_{h \downarrow 0} A = 0$. Moreover, setting for $t \in [0, T_\varpi]$,

$$\begin{aligned} &\Delta_{t,h}^B(x) \quad (40) \\ &= \int |\rho(\phi_{h\lfloor s/h \rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h \rfloor + \lfloor t/h \rfloor))\mathbf{J}_{\phi_{h\lfloor s/h \rfloor}(x)} - \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x))| \mathbb{1}_{\mathsf{A}_\varpi}(s,t) \mathrm{d}s \\ &\quad + \int_{T_\varpi - h\lfloor t/h \rfloor}^{h(\lfloor T_\varpi/h \rfloor - \lfloor t/h \rfloor + 1)} |\rho(\phi_{h\lfloor s/h \rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h \rfloor + \lfloor t/h \rfloor))\mathbf{J}_{\phi_{h\lfloor s/h \rfloor}(x)}| \mathbb{1}_{\mathsf{A}_\varpi}(s,t) \mathrm{d}s \; , \end{aligned}$$

we have for $h < \tilde{h}$, by (38) and **H**3-(ii),

$$\begin{aligned} B &= \int_0^{T_\varpi} \left| \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{h\Delta_{\lfloor t/h \rfloor, h}(x)} - \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{\int \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x) \mathrm{d}s} \right| \mathrm{d}t \\ &\leq \int_0^{T_\varpi} \frac{\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))}{h\Delta_{\lfloor t/h \rfloor, h}(x) \int \varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x) \mathrm{d}s} \Delta_{t,h}^B(x) \mathrm{d}t \\ &\leq mc^{-2} \int_0^{T_\varpi} \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x)) \Delta_{t,h}^B(x) \mathrm{d}t \\ &\leq mc^{-2} \sup_{t \in \mathbb{R}} |\varpi^{\mathrm{c}}(t)| \sup_{x \in \mathbb{R}^d} |f(x)| \sup_{t \in [0, T_\varpi]} |\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| \int_0^{T_\varpi} \Delta_{t,h}^B(x) \mathrm{d}t \; . \end{aligned} \quad (41)$$

By **H**1 and **H**3, the function $s \to \rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)$ is continuous on the interval $[-T_\varpi, T_\varpi]$ and thus is bounded. Therefore, for any $h \in \left(0, \bar{h}\right)$,

$$\begin{aligned} &\sup_{(s,t) \in \mathsf{A}_\varpi} |\varpi^{\mathrm{c}}(h(\lfloor t/h \rfloor + \lfloor s/h \rfloor))\rho(\phi_{h\lfloor s/h \rfloor}(x))\mathbf{J}_{\phi_{h\lfloor s/h \rfloor}}(x)| \\ &\quad \leq \sup_{(s,t) \in \mathsf{A}_\varpi} |\varpi^{\mathrm{c}}(s+t)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| < T_\varpi \sup_{s \in \mathbb{R}} |\varpi^{\mathrm{c}}(s)| \sup_{s \in [-T_\varpi, T_\varpi]} |\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)| < \infty \; . \end{aligned} \quad (42)$$

This implies that

$$\lim_{h \downarrow 0} \int_{T_\varpi - h\lfloor t/h \rfloor}^{h(\lfloor T_\varpi/h \rfloor - \lfloor t/h \rfloor + 1)} |\rho(\phi_{h\lfloor s/h \rfloor}(x))\varpi^{\mathrm{c}}(h(\lfloor s/h \rfloor + \lfloor t/h \rfloor))\mathbf{J}_{\phi_{h\lfloor s/h \rfloor}(x)}| \mathrm{d}s = 0 \; .$$

Moreover, for any $t \in [0, T_\varpi]$, the function

$$s \mapsto |\varpi^{\mathrm{c}}(h(\lfloor t/h \rfloor + \lfloor s/h \rfloor))\rho(\phi_{h\lfloor s/h \rfloor}(x))\mathbf{J}_{\phi_{h\lfloor s/h \rfloor}}(x) - \varpi^{\mathrm{c}}(t+s)\rho(\phi_s(x))\mathbf{J}_{\phi_s}(x)|\mathbb{1}_{\mathsf{A}_\varpi}(s,t)$$

converges pointwise to 0 for almost all $s \in \mathbb{R}$ when $h \downarrow 0$ using **H**1, **H**3 and the continuity of $s \mapsto \phi_s(x)$. The Lebesgue dominated convergence theorem applies and by (40), for all $t \in [0, T_\varpi]$,

$$\lim_{h \downarrow 0} \Delta^B_{t,h}(x) = 0 \; .$$

Moreover, using $h\Delta_{k,h}(x) = h \sum_{i \in \mathbb{Z}} \rho(\mathrm{T}^i_h(x))\varpi^{\mathrm{c}}((k+i)h)\mathbf{J}_{\mathrm{T}^i_h(x)}$ and (42),

$$\sup_{t \in [0, T_\varpi]} \sup_{h \in \left(0, \bar{h}\right)} \Delta^B_{t,h}(x) < \infty \; .$$

The Lebesgue dominated convergence theorem and (41) show that $\lim_{h \downarrow 0} B = 0$ which concludes the proof. □

B.1.1 Supporting Lemmas

For $f \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R}^d)$, define $\mathfrak{J}_f(x)$ the Jacobian matrix of f evaluated at x and the divergence operator by $\operatorname{div} f(x) = \operatorname{tr}[\mathfrak{J}_f(x)]$.

Lemma 10. *Let b be a C^1 vector field in $\mathbb{R}^d$ and $(\phi_t)_{t \in \mathbb{R}}$ be the flow of the ODE* (34)*. For any $t \in \mathbb{R}$, the Jacobian of ϕ_t is given by*

$$\mathbf{J}_{\phi_t}(x) = \exp(\textstyle\int_0^t \operatorname{div} b(\phi_s(x))\mathrm{d}s) \; .$$

Proof. First, for $t \in \mathbb{R}$ and $x \in \mathbb{R}$, write $A(t,x) = \mathfrak{J}_{\phi_t}(x)$ the Jacobian matrix of ϕ_t evaluated at x. By Jacobi's formula, $\dot{\det} A(t,x) = \operatorname{tr}[\operatorname{adj}(A(t,x)) \cdot \dot{A}(t,x)]$, where $\operatorname{tr}[M]$ denotes the trace of a matrix M and $\operatorname{adj}(M)$ its adjugate, i.e. the transpose of the cofactor matrix of M such that $\operatorname{adj}(M)M = \det(M)\,\mathrm{Id}$. Since for all t and x, $\dot{A}(t,x) = \mathfrak{J}_{b \circ \phi_t}(x) = \mathfrak{J}_b(\phi_t(x)) \cdot A(t,x)$, then

$$\dot{\mathbf{J}}_{\phi_t}(x) = \operatorname{tr}[\operatorname{adj}(A(t,x)) \cdot \mathfrak{J}_b(\phi_t(x)) \cdot A(t,x)] = \operatorname{tr}[\mathfrak{J}_b(\phi_t(x))]\mathbf{J}_{\phi_t}(x) \; . \tag{43}$$

Integrating this ODE yields $\mathbf{J}_{\phi_t}(x) = \exp(\int_0^t \operatorname{div} b(\phi_s(x))\mathrm{d}s)$. □

Lemma 11. *Assume **H**1. Then, there exists $C > 0$ such that for any $x \in \mathbb{R}^d, t \in \mathbb{R}$, $k \in \mathbb{Z}$, $h > 0$,*

$$\|\phi_t(x)\| \leq Ce^{C|t|}(\|x\| + 1) \; ,$$
$$\|\,\mathrm{T}^k_h(x)\| \leq Ce^{C|kh|}(\|x\| + 1) \; .$$

This lemma follows from Gronwall's inequality and **H**1.

Lemma 12. *Assume **H**1 and **H**2-(i). There exists $C > 0$ such that for any $x \in \mathbb{R}^d, h \in \left(0, \bar{h}\right)$,*

$$\|\,\mathrm{T}_h(x) - \phi_h(x)\| \leq C\{1 + \|x\|\}\|h^{1+\delta} \; . \tag{44}$$

Proof. Under **H**1 and **H**2-(i), we have

$$\| \mathrm{T}_h(x) - \phi_h(x)\| \leq \|x + hb(x) - \phi_h(x)\| + C_F h^{1+\delta}(1 + \|x\|) \ ,$$

and as $\phi_h(x) = x + \int_0^h b(\phi_s(x))\mathrm{d}s$,

$$\begin{aligned}\|x + hb(x) - \phi_h(x)\| = \|hb(x) - \textstyle\int_0^h b(\phi_s(x))\| &\leq hL_b \sup_{s\in[0,h]} \|\phi_s(x) - x\| \\ &\leq L_b h^2 \{L_b \sup_{s\in[0,h]} \phi_s(x) + \|b(0)\|\} \ . \quad (45)\end{aligned}$$

The proof is completed using Lemma 11. □

Lemma 13. *Assume **H**1 and **H**2-(i). There exists $C > 0$ such that for any $x \in \mathbb{R}^d, k \in \mathbb{N}, h \in \big(0, \bar{h}\big)$, $kh \leq T_\varpi$,*

$$\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \leq Ce^{khC}(1 + \|x\|)h^\delta \ . \quad (46)$$

Proof. Using Lemma 12, **H**1 and **H**2-(i), there exist $C_1, C_2, C_3 > 0$ such that for any $x \in \mathbb{R}^d, k \in \mathbb{N}, h \in \big(0, \bar{h}\big)$, $kh \leq T_\varpi$,

$$\begin{aligned}
&\| \mathrm{T}_h^{k+1}(x) - \phi_{(k+1)h}(x)\| \leq \| \mathrm{T}_h^{k+1}(x) - \mathrm{T}_h \circ \phi_{kh}(x)\| + \| \mathrm{T}_h \circ \phi_{kh}(x) - \phi_{(k+1)h}(x)\| \\
&\leq (1 + hL_b)\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \\
&\qquad + h^{1+\delta} C_1\{2 + \| \mathrm{T}_h^k(x)\| + \|\phi_{kh}(x)\|\} + \| \mathrm{T}_h \circ \phi_{kh}(x) - \phi_{(k+1)h}(x)\| \\
&\leq (1 + hL_b)\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| + h^{1+\delta} 2C_1C_2e^{C_2T_\varpi}\{1 + \|x\|\} + C_3\{1 + \|\phi_{kh}(x)\|\}h^{1+\delta} \\
&\leq (1 + hL_b)\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \\
&\qquad + h^{1+\delta} 2C_1C_2e^{C_2T_\varpi}\{1 + \|x\|\} + C_3\{1 + C_2(1 + \|x\|)\}h^{1+\delta}e^{C_2T_\varpi} \\
&\leq (1 + hL_b)\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| + A_T\{1 + \|x\|\}h^{1+\delta} \ ,
\end{aligned}$$

with $A_T = (2C_1C_2 + C_3(1 + C_2))e^{C_2T_\varpi}$. A straightforward induction yields

$$\| \mathrm{T}_h^k(x) - \phi_{kh}(x)\| \leq \frac{(1 + hL_b)^k}{L_b} A_T(1 + \|x\|)h^\delta \ .$$

□

Lemma 14. *Assume **H**1, **H**2, **H**3 . For any $x \in \mathbb{R}^d$, and $f : \mathbb{R}^d \to \mathbb{R}^d$ bounded and continuous,*

$$\lim_{h\downarrow 0} \int_0^{T_\varpi} \left| \varpi^{\mathrm{c}}(h\,\lfloor t/h\rfloor)\rho(\mathrm{T}_h^{\lfloor t/h\rfloor}(x))\mathbf{J}_{\mathrm{T}_h^{\lfloor t/h\rfloor}}(x) f(\mathrm{T}_h^{\lfloor t/h\rfloor}(x)) - \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))\right| \mathrm{d}t = 0 \ .$$

Proof. Let $x \in \mathbb{R}^d$. Consider the following decomposition, for any $h < \bar{h}$,

$$\begin{aligned}
&\int_0^{T_\varpi} \left| \varpi^{\mathrm{c}}(h\,\lfloor t/h\rfloor)\rho(\mathrm{T}_h^{\lfloor t/h\rfloor}(x))\mathbf{J}_{\mathrm{T}_h^{\lfloor t/h\rfloor}}(x) f(\mathrm{T}_h^{\lfloor t/h\rfloor}(x)) - \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x))\right| \mathrm{d}t \\
&\leq \tfrac{h}{T_\varpi} \textstyle\sum_{k\in\mathbb{Z}} \varpi^{\mathrm{c}}(kh)|\rho(\mathrm{T}_h^k(x))\mathbf{J}_{\mathrm{T}_h^k}(x) f(\mathrm{T}_h^k(x)) - \rho(\phi_{kh}(x))\mathbf{J}_{\phi_{kh}}(x) f(\phi_{kh}(x))| \\
&+ \textstyle\int_0^{T_\varpi} |\varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x) f(\phi_t(x)) - \varpi^{\mathrm{c}}(h\,\lfloor t/h\rfloor)\rho(\phi_{h\lfloor t/h\rfloor}(x))\mathbf{J}_{\phi_{h\lfloor t/h\rfloor}}(x) f(\phi_{h\lfloor t/h\rfloor}(x))|\mathrm{d}t \ .
\end{aligned}$$

The first term converges to 0 by Lemma 13 and **H**2-(ii) as $\varpi^{\mathrm{c}}(kh) = 0$ for $kh > T_\varpi$. By **H**1 and **H**3, the function $t \to \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))$ is continuous on the compact $[0, T_\varpi]$ and thus is bounded. Therefore, for any $h \in \left(0, \bar{h}\right)$,

$$\sup_{t\in[0,T_\varpi]} |\varpi^{\mathrm{c}}(h\lfloor t/h\rfloor)\rho(\phi_{h\lfloor t/h\rfloor}(x))\mathbf{J}_{\phi_{h\lfloor t/h\rfloor}}(x)f(\phi_{h\lfloor t/h\rfloor}(x))| \\ \leq \sup_{t\in\mathbb{R}} |\varpi^{\mathrm{c}}| \sup_{x\in\mathbb{R}^d} |f(x)| \sup_{t\in[0,T_\varpi]} |\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)| < \infty \,. \quad (47)$$

Moreover, $t \mapsto \varpi^{\mathrm{c}}(h\lfloor t/h\rfloor)\rho(\phi_{h\lfloor t/h\rfloor}(x))\mathbf{J}_{\phi_{h\lfloor t/h\rfloor}}(x)f(\phi_{h\lfloor t/h\rfloor}(x))$ converges pointwise when $h \downarrow 0$ to $t \to \varpi^{\mathrm{c}}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)f(\phi_t(x))$ by continuity, using **H**1 and **H**3. The Lebesgue dominated convergence theorem applies and the second term goes to 0 as $h \downarrow 0$. □

B.2 NEIS algorithm after Rotskoff and Vanden-Eijnden (2019)

Non Equilibrium Importance Sampling (NEIS) has been introduced in the pioneering work of Rotskoff and Vanden-Eijnden (2019). NEIS relies on the flow of the ODE $\dot{x}_t = b(x_t)$ and the introduction of a set $\mathsf{O} \subset \mathbb{R}^d$. As in Appendix B, we assume **H**1 holds and denote by $(\phi_t)_{t\in\mathbb{R}}$ the flow of this ODE.

Define for $x \in \mathsf{O}$, the exit times $\tau^+(x) \geq 0$ (resp. $\tau^-(x) \leq 0$) satisfying

$$\tau^+(x) = \inf\{t \geq 0 \,:\, \phi_t(x) \notin \mathsf{O}\} \,,\; \tau^-(x) = \inf\{t \leq 0 \,:\, \phi_t(x) \notin \mathsf{O}\} \,. \quad (48)$$

The validity of NEIS relies on the following assumption.

H4. *The average time of an orbit in* O *is finite, i.e.*

$$Z_\tau = \int_{\mathsf{O}} (\tau^+(x) - \tau^-(x))\rho(x)\mathrm{d}x < \infty \,. \quad (49)$$

Under **H**4, we can define the proposal distribution

$$\rho_{\mathrm{T}}(x) = Z_\tau^{-1} \int_{\mathsf{O}} \mathbb{1}_{[\tau^-(x),\tau^+(x)]}(t)\rho(\phi_t(x))\mathbf{J}_{\phi_t}(x)\mathrm{d}t \,. \quad (50)$$

Under **H**4, (Rotskoff and Vanden-Eijnden, 2019, Equation (8)) derive the following estimator of $\rho(f)$, closely related to (16), in the case $\varpi \equiv 1$, on the restricted set $\mathsf{O} \subset \mathbb{R}^d$:

$$I_N^{\mathrm{NEIS}}(f) = \frac{1}{N}\sum_{i=1}^N \int_{\tau^-(X^i)}^{\tau^+(X^i)} w_t(X^i)f(\phi_t(X^i))\mathrm{d}t \quad (51)$$

$$w_t(x) = \frac{\rho(\phi_t(x))\mathbf{J}_{\phi_t(x)}}{\int_{\tau^-(x)}^{\tau^+(x)} \rho(\phi_t(x))\mathbf{J}_{\phi_t(x)}\mathrm{d}t} \,. \quad (52)$$

Note that in practice, in order for **H**4 to be verified, one typically requires that O be bounded, as discussed in Rotskoff and Vanden-Eijnden (2019).

Following Rotskoff and Vanden-Eijnden (2019), consider a d-dimensional system with position $q \in \mathbb{R}^d$, momentum $p \in \mathbb{R}^d$ and Hamiltonian $H(p,q) = (1/2)\|p\|^2 + U(q)$ where $U(q)$ is a potential assumed to be bounded from below. Denote by $V(E)$ the volume of the phase-space below some threshold energy E,

$$V(E) = \int \mathbb{1}_{\{H(p,q)\leq E\}}\mathrm{d}p\mathrm{d}q \,. \quad (53)$$

To calculate (53), we set $x = (p, q)$, define $\mathsf{O} = \{x; H(x) \leq E_{\max}\}$ for some $E_{\max} < \infty$, and use the dissipative Langevin dynamics with $b(x) = (p, -\nabla U(q) - \gamma p)$, *i.e.*

$$\dot{q} = p \ , \quad \dot{p} = -\nabla U(q) - \gamma p \ ,$$

for some friction coefficient $\gamma > 0$. With this choice, $\mathbf{J}_{\phi_t}(x) = \mathrm{e}^{-d\gamma t}$. Taking ρ to be the uniform distribution on the (bounded) set O, write the estimator for $E \leq E_{\max}$, $V(E)/V(E_{\max}) = \int \mathbb{1}_{\{H(p,q)\leq E\}}\rho(p,q)\mathrm{d}p\mathrm{d}q$, where $\rho(p,q) = \mathbb{1}_{\mathsf{O}}(p,q)/V(E_{\max})$, we get

$$\begin{aligned} V(E)/V(E_{\max}) &= \frac{1}{N}\sum_{i=1}^{N} \frac{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)} \mathbb{1}_{\{H(\phi_t(X^i))\leq E\}} \mathrm{d}t}{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)} \mathrm{d}t} \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{\int_{\tau^E(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)} \mathrm{d}t}{\int_{\tau^-(X^i)}^{\tau^+(X^i)} \mathbf{J}_{\phi_t(X^i)} \mathrm{d}t} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{e}^{-d\gamma(\tau^E(X^i)-\tau^-(X^i))} \ , \end{aligned} \tag{54}$$

where $\tau^E(x)$ denotes the (possibly infinite) time for a trajectory initiated at $x = (p, q)$ to reach the energy $E \leq E_{\max}$.

Finally, to estimate the normalizing constant, Rotskoff and Vanden-Eijnden (2019) discretize the energy levels $\{E_0, \ldots, E_P\}$ and write their estimator as

$$\widehat{Z}^{\mathrm{NEIS}}_{X^{1:N}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{P} \mathrm{e}^{-d\gamma(\tau^E_\ell(X^i)-\tau^-(X^i))}(E_\ell - E_{\ell-1}) \ , \tag{55}$$

using an approximation of the identity

$$Z = \int_{\mathsf{O}}\int_0^\infty \mathbb{1}_{\{\mathrm{L}(x)>L\}}\rho(x)\mathrm{d}L\mathrm{d}x = \int_0^\infty \mathbb{P}_{X\sim\rho}(\mathrm{L}(X) > L)\mathrm{d}L \ ,$$

which is at the core of nested sampling Chopin and Robert (2010).

B.3 NEO with exit times

Consider $\mathsf{O} \subset \mathbb{R}^d$ and let T be a C^1-diffeomorphism on $\mathbb{R}^d$. We introduce here an estimator based on the forward and backward orbits in O associated with T. Define the exit times $\tau^+ : \mathbb{R}^d \to \mathbb{N}$ and $\tau^- : \mathbb{R}^d \to \mathbb{N}_-$, given, for all $x \in \mathbb{R}^d$, by

$$\tau^+(x) = \inf\{k \geq 1 \ : \ \mathrm{T}^k(x) \notin \mathsf{O}\} \ , \tag{56}$$

$$\tau^-(x) = \sup\{k \leq -1 \ : \ \mathrm{T}^k(x) \notin \mathsf{O}\} \ , \tag{57}$$

with the convention $\inf\emptyset = +\infty$ and $\sup\emptyset = -\infty$, and set

$$\mathrm{I} = \{(x, k) \in \mathsf{O} \times \mathbb{Z} \ : \ k \in [\tau^-(x) + 1 : \tau^+(x) - 1]\} \ . \tag{58}$$

For any $k \in \mathbb{Z}$, define $\rho_k : \mathbb{R}^d \to \mathbb{R}_+$ by

$$\rho_k(x) = \rho(\mathrm{T}^{-k}(x))\mathbf{J}_{\mathrm{T}^{-k}}(x)\mathbb{1}_{\mathrm{I}}(x, -k) \ . \tag{59}$$

The density ρ_k is the push-forward of $\mathbb{1}_{\mathrm{I}}(x,k)\rho(x)$ by T^k, *i.e.* for any $k \in \mathbb{Z}$ and any bounded function $g : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathsf{O}} g(y)\rho_k(y)\mathrm{d}y = \int_{\mathsf{O}} g(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)\rho(x)\mathrm{d}x \; . \tag{60}$$

Consider the following assumption:

H5. *The nonnegative sequence* $(\varpi_k)_{k\in\mathbb{Z}}$ *satisfies* $\varpi_0 > 0$ *and*

$$Z_{\mathrm{T}}^{\varpi} = \int_{\mathsf{O}} \sum_{k\in\mathbb{Z}} \varpi_k \rho_k(x)\mathrm{d}x = \int_{\mathsf{O}} \sum_{k\in\mathbb{Z}} \varpi_k \rho(\mathrm{T}^k(x))\mathbf{J}_{\mathrm{T}^k}(x)\mathbb{1}_{\mathrm{I}}(x,k)\mathrm{d}x < \infty \; . \tag{61}$$

Consider the pdf

$$\rho_{\mathrm{T}}(x) = \frac{1}{Z_{\mathrm{T}}^{\varpi}} \sum_{k\in\mathbb{Z}} \varpi_k \rho_k(x) \; , \tag{62}$$

where Z_{T}^{ϖ} is the normalizing constant. This is a *non-equilibrium* distribution, since ρ_{T} is not invariant by T in general. Using ρ_{T} as an importance distribution to obtain an unbiased estimator of $\int f(x)\rho(x)\mathrm{d}x$ is feasible since as $\varpi_0 > 0$, $\sup_{x\in\mathsf{O}} \rho(x)/\rho_{\mathrm{T}}(x) \leq Z_{\mathrm{T}}/\varpi_0 < \infty$, hence

$$\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x = \int_{\mathsf{O}} \left(f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}(x)} \right) \rho_{\mathrm{T}}(x)\mathrm{d}x \; .$$

From (60), the right hand side can be computed using the following key result.

Theorem 15. *For any* $f : \mathbb{R}^d \to \mathbb{R}$ *measurable bounded function, we have*

$$\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x = \int_{\mathsf{O}} \sum\nolimits_{k\in\mathbb{Z}} f(\mathrm{T}^k(x))w_k(x)\rho(x)\mathrm{d}x \; , \tag{63}$$

where, for any $x \in \mathbb{R}^d$ *and* $k \in \mathbb{Z}$,

$$w_k(x) = \varpi_k\rho_{-k}(x) \Big/ \sum\nolimits_{j\in\mathbb{Z}} \varpi_{j+k}\rho_j(x) \quad . \tag{64}$$

Proof. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable bounded function. By (60), writing $g \leftarrow f\rho/\rho_{\mathrm{T}}$,

$$\begin{aligned}\int_{\mathsf{O}} f(x)\rho(x)\mathrm{d}x &= \int_{\mathsf{O}} \left(f(x)\frac{\rho(x)}{\rho_{\mathrm{T}}(x)} \right) \rho_{\mathrm{T}}(x)\mathrm{d}x \\ &= \int_{\mathsf{O}} \sum_{k\in\mathbb{Z}} \left(f(\mathrm{T}^k(x))\frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{Z_{\mathrm{T}}^{\varpi}\rho_{\mathrm{T}}(\mathrm{T}^k(x))} \right) \rho(x)\mathrm{d}x \; .\end{aligned}$$

We now need to prove:

$$\frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{Z_{\mathrm{T}}^{\varpi}\rho_{\mathrm{T}}(\mathrm{T}^k(x))} = \frac{\varpi_k\rho(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(x,k)}{\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x))} = \frac{\varpi_k\rho_{-k}(x)}{\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho_j(x)} = w_k(x) \quad ,$$

with the convention $0/0 = 0$. We thus need to show that for any $x \in \mathsf{O}$, $k \in \mathbb{Z}$,

$$\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x)) = \frac{\mathbb{1}_{\mathrm{I}}(x,k)}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho_j(x) \; .$$

Using the identity $\mathbf{J}_{\mathrm{T}^{-i+k}}(x) = \mathbf{J}_{\mathrm{T}^{-i}}(\mathrm{T}^k(x))\mathbf{J}_{\mathrm{T}^k}(x)$, we obtain

$$\begin{aligned}
\mathbb{1}_{\mathrm{I}}(x,k)\sum_{i\in\mathbb{Z}}\varpi_i\rho_i(\mathrm{T}^k(x)) &= \sum_{i\in\mathbb{Z}}\mathbb{1}_{\mathrm{I}}(x,k)\varpi_i\rho(\mathrm{T}^{-i}(\mathrm{T}^k(x)))\mathbf{J}_{\mathrm{T}^{-i}}(\mathrm{T}^k(x))\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-i) \\
&= \frac{1}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{i\in\mathbb{Z}}\mathbb{1}_{\mathrm{I}}(x,k)\varpi_i\rho(\mathrm{T}^{-i+k}(x))\mathbf{J}_{\mathrm{T}^{-i+k}}(x)\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-i) \\
&= \frac{1}{\mathbf{J}_{\mathrm{T}^k}(x)}\sum_{j\in\mathbb{Z}}\varpi_{j+k}\rho(\mathrm{T}^{-j}(x))\mathbf{J}_{\mathrm{T}^{-j}}(x)\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x),-j-k)\mathbb{1}_{\mathrm{I}}(x,k)
\end{aligned}$$

Note that if $(x,k)\in\mathrm{I}$, we have $(x,-j)\in\mathrm{I}$ if and only if $(\mathrm{T}^k(x),-j-k)\in\mathrm{I}$ by definition of I (58). The proof is concluded by noting that:

$$\mathbb{1}_{\mathrm{I}}(\mathrm{T}^k(x)),-j-k)\mathbb{1}_{\mathrm{I}}(x,k) = \mathbb{1}_{\mathrm{I}}(x,-j)\mathbb{1}_{\mathrm{I}}(x,k)\ .$$

□

C Iterated SIR

Let us recall the principle of the Sampling Importance Resampling method (SIR; Rubin (1987); Smith and Gelfand (1992)) whose goal is to approximately sample from the target distribution π using samples drawn from a proposal distribution ρ.

In SIR, a N-i.i.d. sample $X^{1:N}$ is first generated from the proposal distribution ρ. A sample X^* is approximately drawn from the target π by choosing randomly a value in $X^{1:N}$ with probabilities proportional to the importance weights $\{\mathrm{L}(X^i)\}_{i=1}^N$, where $\mathrm{L}(x) = \pi(x)/\rho(x)$. Note that the importance weights are required to be known only up to a constant factor.

For SIR, as $N\to\infty$, the sample X^* is *asymptotically* distributed according to π; see Smith and Gelfand (1992).

A subsequent algorithm is the *iterated SIR* (i-SIR) (Andrieu et al., 2010). Here, N is not necessarily large ($N\geq 2$), the whole process of sampling a set of proposals, computing the importance weights, and picking a candidate, is iterated. At the n-th step of i-SIR, the active set of N proposals $X_n^{1:N}$ and the index $I_n\in[N]$ of the conditioning proposal are kept. First i-SIR updates the active set by setting $X_{n+1}^{I_n} = X_n^{I_n}$ (keep the conditioning proposal) and then draw independently $X_{n+1}^{1:N\setminus\{I_n\}}$ from ρ. Then it selects the next proposal index $I_{n+1}\in[N]$ by sampling with probability proportional to $\{\tilde{w}(X_{n+1}^i)\}_{i=1}^N$. As shown in Andrieu et al. (2010), this algorithm defines a partially collapsed Gibbs sampler (PCG) of the augmented distribution

$$\bar{\boldsymbol{\pi}}(x^{1:N},i) = \frac{1}{N}\pi(x^i)\prod_{j\neq i}\rho(x^j) = \frac{1}{N}\tilde{w}(x^i)\prod_{j=1}^N\rho(x^j)\ .$$

The PCG sampler can be shown to be ergodic provided that ρ and π are continuous and ρ is positive on the support of π. If in addition the importance weights are bounded, the Gibbs sampler can be shown to be uniformly geometrically ergodic (Lindsten et al., 2015; Andrieu et al., 2018). It follows that the distribution of the conditioning proposal $X_n^* = X_n^{I_n}$ converges to π as the iteration index n goes to infinity. Indeed, for any integrable function f on $\mathbb{R}^d$, with $(X_{1:N},I)\sim\bar{\boldsymbol{\pi}}$,

$$\mathbb{E}[f(X^I)] = \int\sum_{i=1}^N f(x^i)\bar{\boldsymbol{\pi}}(x^{1:N},i)\mathrm{d}x^{1:N} = N^{-1}\sum_{i=1}^N\int f(x^i)\pi(x^i)\mathrm{d}x_i = \int f(x)\pi(x)\mathrm{d}x\ .$$

When the state space dimension d increases, designing a proposal distribution ρ guaranteeing proper mixing properties becomes more and more difficult. A way to circumvent this problem is to use dependent proposals, allowing in particular *local moves* around the conditioning orbit. To implement this idea, for each $i \in [N]$, we define a proposal transition, $r_i(x^i; x^{1:N\setminus\{i\}})$ which defines the the conditional distribution of $X^{1:N\setminus\{i\}}$ given $X^i = x^i$. The key property validating i-SIR with dependent proposals is that all one-dimensional marginal distributions are equal to ρ, which requires that for each $i, j \in [N]$,

$$\rho(x^i) r_i(x^i; x^{1:N\setminus\{i\}}) = \rho(x^j) r_j(x^j; x^{1:N\setminus\{j\}}) \; . \tag{65}$$

The (unconditional) joint distribution of the particles is therefore defined as

$$\rho_N\left(x^{1:N}\right) = \rho(x^1) r_1(x^1; x^{1:N\setminus\{1\}}) \; . \tag{66}$$

The resulting modification of the i-SIR algorithm is straightforward: $X^{1:N\setminus\{I_n\}}$ is sampled jointly from the conditional distribution $r_{I_n}(X_n^{I_n}, \cdot)$ rather than independently from ρ.

D Additional Experiments

D.1 Normalizing constant estimation

We consider here the problem of the estimation of the normalizing constant of Cauchy mixtures. The Cauchy distribution with scale σ has a pdf defined by $\mathrm{Cauchy}(x; \mu, \sigma) = [\pi\sigma(1 + \{(x-\mu)/\sigma\}^2]^{-1}$. The target distribution is a product of mixtures of two Cauchy distributions,

$$\pi(x) = \prod_{i=1}^{n} \frac{1}{2} \left[\mathrm{Cauchy}\left(x_i; \mu, \sigma\right) + \mathrm{Cauchy}\left(x_i; -\mu, \sigma\right)\right], \quad \mu = 5, \sigma = 1 \; .$$

NEO-IS is compared with IS estimator using the same proposal ρ. We also compare NEO-IS to Neural IS Müller et al. (2019) with a Cauchy as base distribution.

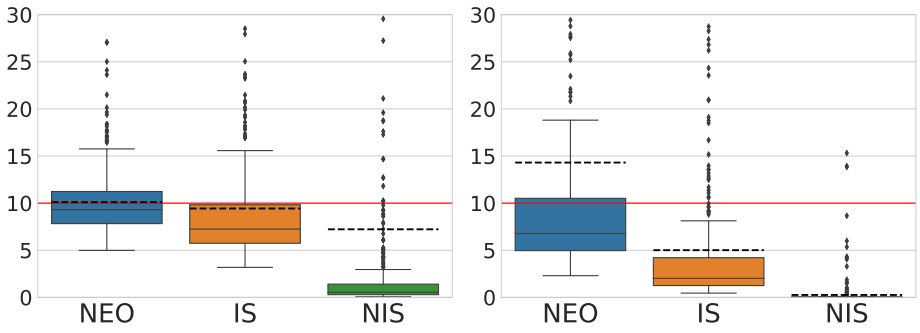

Figure 5: Boxplots of 500 independent estimations of the normalizing constant of the Cauchy mixture in dimension $d = 10, 15$ (top, bottom). The true value is given by the red line. The figure displays the median (solid lines), the interquartile range, and the mean (dashed lines) over the 500 runs

Model	VAE, $d = 32$	VAE, $d = 16$	IWAE, $d = 32$	IWAE, $d = 16$
IS	-90.17	-90.44	-88.76	-90.13
AIS	-89.67	-89.97	-88.30	-89.61
NEO-IS	-88.81	-89.17	-87.46	-88.99

Table 1: Evaluation of the log-likelihood (normalizing constant) of different Variational Auto Encoders.

Finally, we compare NEO-IS with NEIS[3]. We consider here `MG25` in dimension 5 and 10, where all the covariances of the Gaussian distributions are diagonal and equal to $0.005\,\mathrm{Id}$. NEIS and NEO-IS are run for the same computational time. We add an IS scheme as a baseline for comparison. All algorithms (NEO-IS, NEIS, IS) are run for 7.20s and 11.30s wall clock time respectively for $d = 5$ and $d = 10$. For NEO-IS, we use a conformal transform with $h = 0.1$, $K = 10$ and $\gamma = 1$. For NEIS, we choose $\gamma = 1$ and consider a stepsize $h = 10^{-4}$ corresponding to an optimal trade-off between the discretization bias inherent to NEISand its computational budget. We can observe that NEO-IS always outperforms NEIS, which suffers from a non-negligeable bias if the stepsize h is not chosen small enough.

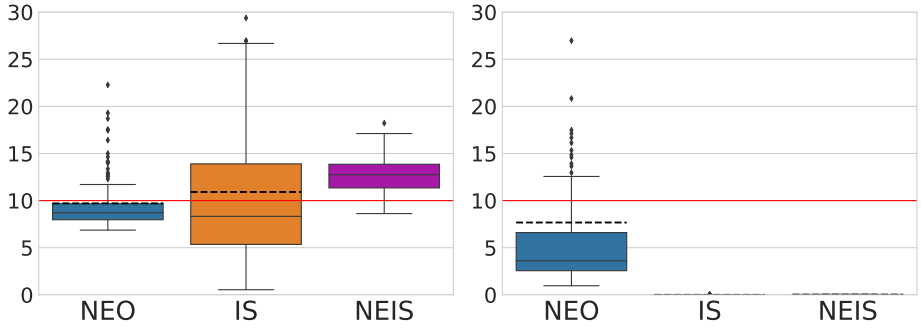

Figure 6: NEO v. NEIS. 25 GM with $\sigma^2 = 0.005$, $d = 5$. 500 runs each.

D.2 Log-likelihood estimation

We finally present the evaluation of the log-likelihood of the VAE introduced in Section 5: given a test set $\mathcal{T} = \{z_i\}_{i=1}^{M_\mathcal{T}}$, we estimate $\sum_{i=1}^{M_\mathcal{T}} \log p_{\theta^*}(z_i)$. We also estimate similarly the log-likelihood of an Importance Weighted Auto Encoder (IWAE) Burda et al. (2016). Following Wu et al. (2016), we compare IS, AIS, and NEO-IS. As previously, AIS, IS, and NEO-IS are given a similar computational budget, choosing here $K = 12$, $N = 5 \cdot 10^3$. For NEO, we choose $\gamma = 1.$ and $h = 0.2$. Similarly, the stepsize of HMC transitions in AIS is $h = 0.1$ in order to achieve an acceptance ratio of around 0.6 in the HMC transitions. We report in Table 1 the log-likelihood computed on the test set for VAE, IWAE with latent dimension in $\{16, 32\}$. For the same computational budget, NEO-IS yields consistently better values for the estimation of the log-likelihood of the VAE.

[3]The code from Rotskoff and Vanden-Eijnden (2019) we run is available at `https://gitlab.com/rotskoff/trajectory_estimators/-/tree/master`.

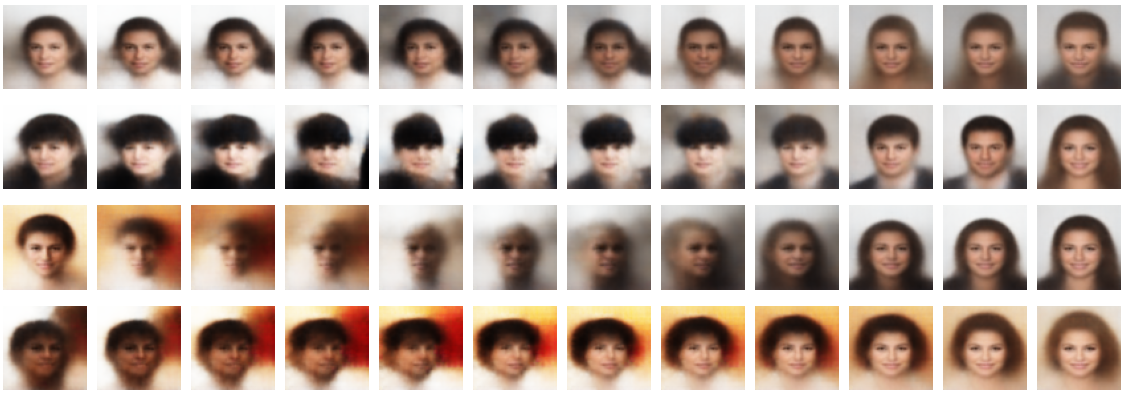

Figure 7: Forward orbits of NEO-MCMC.

D.3 Gibbs inpainting

We display here additional results for the Gibbs inpainting experiment presented in Section 5. We emphasize that the starting images are chosen at random in the test set.

E NEO-VAE

Denote by $p_\theta(x, z)$ the joint distribution of the observation $z \in \mathbb{R}^p$ and the latent variable $x \in \mathbb{R}^d$. The marginal likelihood is given, for $z \in \mathbb{R}^p$ by $p_\theta(z) = \int p_\theta(x, z)\mathrm{d}x$. Given a training set $\mathcal{D} = \{z_i\}_{i=1}^M$, the objective is to estimate θ by maximizing the likelihood, *i.e.* maximizing $\log p_\theta(\mathcal{D}) = \sum_{i=1}^M \log p_\theta(z_i)$. Variational inference (VI) provides us with a tool to simultaneously approximate the intractable posterior $p_\theta(x|z)$ and maximize the marginal likelihood $p_\theta(\mathcal{D})$ in the parameter θ. This is achieved by introducing a parametric family $\{q_\phi(x|z), \phi \in \Phi\}$ to approximate the posterior $p_\theta(x|z)$ and maximizing the Evidence Lower Bound (ELBO) (see Kingma and Welling (2019)) $\mathcal{L}_{\mathrm{ELBO}}(\mathcal{D}, \theta, \phi) = \sum_{i=1}^M \mathcal{L}_{\mathrm{ELBO}}(z_i, \theta, \phi)$ where

$$\begin{aligned} \mathcal{L}_{\mathrm{ELBO}}(z, \theta, \phi) &= \int \log\left(\frac{p_\theta(x, z)}{q_\phi(x \mid z)}\right) q_\phi(x \mid z)\mathrm{d}x \\ &= \log p_\theta(z) - \mathrm{KL}(q_\phi(\cdot \mid z) \| p_\theta(\cdot \mid z)) \,, \end{aligned} \tag{67}$$

and KL is the Kullback–Leibler divergence. In the sequel, we set $\rho(x) = q_\phi(x \mid z)$ and $\mathrm{L}(x) = p_\theta(x, z)/q_\phi(x \mid z)$. In such a case, $\pi(x) = \rho(x)\mathrm{L}(x)/Z = p_\theta(x \mid z)$ and $Z = p_\theta(z)$ (in these notations, the dependence in the observation z is implicit).

We follow the the auxiliary variational inference framework (AVI) provided by Agakov and Barber (2004). We consider a joint distribution $\bar{p}_\theta(x, u, z)$ which is such that $p_\theta(z) = \int p_\theta(x, u, z)\mathrm{d}x\mathrm{d}u$ where $u \in \mathsf{U}$ is an auxiliary variable (the auxiliary variable can both have discrete and continuous components; when u has discrete components the integrals should be replaced by a sum). Then as the usual VI approach, we consider a parametric family $\{\bar{q}_\phi(x, u|z), \phi \in \Phi\}$. Introducing auxiliary variables loses the tractability of (67) but they allow for their own ELBO as suggested in Agakov and Barber (2004); Lawson et al. (2019)

Figure 8: Gibbs inpainting for CelebA dataset. From top to bottom: i-SIR, HMC and NEO-MCMC: From left to right, original image, blurred image to reconstruct, and output every 5 iterations of the Markov chain.

by minimizing

$$\mathrm{KL}(\bar{q}_\phi(\cdot \mid z) \| \bar{p}_\theta(\cdot \mid z)) = \int \bar{q}_\phi(x, u|z) \log \left(\frac{\bar{p}_\theta(x, u, z)}{\bar{q}_\phi(x, u|z)} \right) \mathrm{d}x\mathrm{d}u \ . \tag{68}$$

The auxiliary variable u is naturally associated with the extended target $\bar{\boldsymbol{\pi}}$ defined similar to Remark 2,

$$\bar{\boldsymbol{\pi}}_N([x, x^{1:N\setminus\{i\}}], i) = \frac{\widehat{Z}_x}{NZ} \rho_N(x^{1:N}) \tag{69}$$

with $(x, u) = ([x, x^{1:N\setminus\{i\}}], i)$, a shorthand notation for a N-tuple $x^{1:N}$ with $x^i = x$, and, with r_i defined in (15),

$$\rho_N(x^{1:N}) = \rho(x^1) r_1(x^1, x^{2:N}) = \rho(x^j) r_j(x^j, x^{1:N\setminus\{j\}}) \ , \quad j \in \{1, \dots, N\} \ . \tag{70}$$

An extended proposal playing the role of $\bar{q}_\phi(x, u|z)$ is derived from the NEO MCMC sampler, i.e.

$$\bar{\rho}_N([x, x^{1:N\setminus\{i\}}], i) = \frac{\widehat{Z}_x}{N \widehat{Z}_{x^{1:N}}} \rho_N(x^{1:N}) \ . \tag{71}$$

where $\widehat{Z}_{x^{1:N}}$ is the NEO estimator (4) of the normalizing constant. Note that, by construction,

$$\sum_{i=1}^{N} \bar{\rho}_N(x^{1:N}, i) = \rho_N(x^{1:N}) \tag{72}$$

showing that this joint proposal can be sampled by drawing the proposals $x^{1:N} \sim \rho_N$, then sampling the path index $i \in [N]$ with probability proportional to $(\widehat{Z}_{x^i})_{i=1}^N$ (with $\widehat{Z}_x$ defined in (4)). The ratio of (69) over (71) is

$$\bar{\boldsymbol{\pi}}_N(x^{1:N}, i) \big/ \bar{\rho}_N(x^{1:N}, i) = \widehat{Z}_{x^{1:N}} \big/ Z \ . \tag{73}$$

Thus, we write the augmented ELBO (68)

$$\mathcal{L}_{\mathrm{NEO}} = \int \rho_N(x^{1:N}) \log \widehat{Z}_{x^{1:N}} \mathrm{d}x^{1:N} = \log Z - \mathrm{KL}(\bar{\rho}_N | \bar{\boldsymbol{\pi}}_N) \ , \tag{74}$$

where we have used (72) and that the ratio $\bar{\boldsymbol{\pi}}_N(x^{1:N}, i) \big/ \bar{\rho}_N(x^{1:N}, i)$ does not depend on the path index i. When $\varpi_k = \delta_{k,0}$, where $\delta_{i,j}$ is the Kronecker symbol, and $\rho_N(x^{1:N}) = \prod_{j=1}^N \rho(x^j)$, we exactly retrieve the Importance Weighted AutoEncoder (IWAE); see e.g. Burda et al. (2016) and in particular the interpretation in Cremer et al. (2017).

Choosing the conformal Hamiltonian introduced in Section 2 allows for a family of invertible flows that depends on the parameter θ which itself is directly linked to the target distribution. Table 2 displays the estimated NLL of all models provided by IS and the NEO method. It is interesting to note here again that NEO improves the training of the VAE when the dimension of the latent space is small to moderate.

Table 2: Negative Log Likelihood estimates for VAE models for different latent space dimensions.

	$d = 4$		$d = 8$		$d = 16$		$d = 50$	
model	IS	NEO	IS	NEO	IS	NEO	IS	NEO
VAE	115.01	113.49	97.96	97.64	90.52	90.42	88.22	88.36
IWAE, $N = 5$	113.33	111.83	97.19	96.61	89.34	89.05	87.49	87.27
IWAE, $N = 30$	111.92	110.36	96.81	95.94	88.99	88.64	86.97	86.93
NEO VAE, $K = 3$	109.14	107.47	94.50	94.26	89.03	88.92	88.14	88.16
NEO VAE, $K = 10$	110.02	107.90	94.63	94.22	89.71	88.68	88.25	86.95