

Quasi-Symplectic Langevin Variational Autoencoder

Zihao Wang, Hervé Delingette

▶ **To cite this version:**

Zihao Wang, Hervé Delingette. Quasi-Symplectic Langevin Variational Autoencoder. 2021. hal-03024748v3

HAL Id: hal-03024748

https://inria.hal.science/hal-03024748v3

Preprint submitted on 19 Apr 2021 (v3), last revised 3 Jun 2021 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QUASI-SYMPLECTIC LANGEVIN VARIATIONAL AUTOENCODER

Zihao WANG *
Inria Sophia Antipolis, University Côte d'Azur
2004 Route des Lucioles, 06902 Valbonne
zihao.wang@inria.fr

Hervé Delingette
Inria Sophia Antipolis, University Côte d'Azur
2004 Route des Lucioles, 06902 Valbonne
herve.delingette@inria.fr

ABSTRACT

Variational autoencoder (VAE) is a very popular and well-investigated generative model vastly used in neural learning research. To leverage VAE in practical tasks dealing with a massive dataset of large dimensions it is required to deal with the difficulty of building low variance evidence lower bounds (ELBO). Markov Chain Monte Carlo (MCMC) is one of the effective approaches to tighten the ELBO for approximating the posterior distribution. Hamiltonian Variational Autoencoder (HVAE) is an effective MCMC inspired approach for constructing a the low-variance ELBO which is also amenable to the reparameterization trick. In this work, we propose a Quasi-symplectic Langevin Variational autoencoder (Langevin-VAE) by incorporating the gradients information in the inference process through Langevin dynamic. We shows the effectiveness of the proposed approach by toy and real world examples.

1 INTRODUCTION

Variational Autoencoders (VAE) are a popular generative neural model applied in a vast number of practical cases to perform unsupervised analysis and to generate specific dataset. It has the advantages of offering a quantitative assessment of generated model quality and being less cumbersome to train compared to Generative Adversarial Networks (GANs). The key factor influencing the performance of VAE models is the quality of the ,arginal likelihood approximation in the corresponding evidence lower bound (ELBO). A common method to make the amortized inference efficient is to constraint the posterior distribution of the latent variables to follow a given closed-form distribution, often multivariate Gaussian (Wolf et al., 2016). However, this severely limits the flexibility of the encoder. In (Salimans et al., 2015), the Hamiltonian variational inference is proposed to remove the requirement of an explicit formulation of the posterior distribution by forward sampling a Markov chain based on Hamiltonian dynamics. It can be seen as a type of normalizing flows (NFs) (Rezende & Mohamed, 2015) where repeated transformations of probability densities are replaced by time integration of space and momentum variables. To guarantee the convergence of HVI to the true posterior distribution, Wolf et al. proposed to add an acceptance step in Hamiltonian variational inference algorithm. In (Caterini et al., 2018), authors first combined VAE and HVI in Hamiltonian Variational Autoencoders (HVAE) which include a a dynamic phase space where momentum component ρ and position component z are integrated which introduces the target information into the flow. Briefly, HVAE employs a K steps Hamiltonian $\mathcal{H}_K$ transformation process to build an unbiased estimation of posterior $q(z)$ by extending $\tilde{p}(x, z)$ as $\tilde{p}(x, \mathcal{H}_K(z_0, \rho_0))$ leading to: $\tilde{p}(x) := \frac{\hat{p}(x, \mathcal{H}_K(z_0, \rho_0))}{q(\mathcal{H}_K(z_0, \rho_0))}$, where: $\hat{p}(x, z_K, \rho_K) = \hat{p}(x, \mathcal{H}_K(z_0, \rho_0)) = \hat{p}(x, z_K)\mathcal{N}(\rho_K|0, I)$. HVAE in particular, allows the use of the reparameterization trick during inference thus leading to efficient ELBO gradients computation.

In this work, as an exploration of the application of dynamic systems in the field of machine learning, we propose a novel inference framework named quasi-symplectic Langevin variational auto-encoder (Langevin-VAE) that leads to both reversible Markov kernels and phase quasi-volume invariance similarly to Hamiltonian flow (Caterini et al., 2018) while reducing the computation and memory requirements. The proposed method is a low-variance unbiased lower bound estimator for

*Corresponding author.

infinitesimal discretization steps but needs just one target Jacobian calculation and avoids computing the Hessian of Jacobian. The leapfrog integrator of Hamiltonian flows is replaced in our approach by the quasi-symplectic Langevin integration.We show that the Langevin-VAE is a generalized stochastic inference framework since the proposed Langevin-VAE becomes symplectic when the viscosity coefficient $\nu = 0$ is set to zero. The method is verified through quantitative and qualitative comparison with HVAE inference framework on a benchmark dataset.

2 PRELIMINARY

2.1 VARIATIONAL INFERENCE AND NORMALIZING FLOW

One core problem in the Variational Inference (VI) task is to find a suitable replacement distributions $q_\theta(z)$ of the posterior distribution $p(z|x)$ for optimizing the ELBO: $\text{argmax}_\theta \, \mathbb{E}_q[\log p(x, z) - \log q_\theta(z)]$. To tackle this problem, Ranganath et al. proposed black box variational inference by estimating the noisy unbiased gradient of ELBO to perform direct stochastic optimization of ELBO. Kingma & Welling proposed to use some multivariate Gaussian posterior distributions of latent variable z generated by a universal function ω, which makes reparameterization trick is possible: $\text{argmax}_{\theta,\phi} \, \mathbb{E}_{q_\theta(z|x)}[\log p_\theta(x, z) - \log q_\phi(z|x)]$. To better approximate potentially complex posterior distributions of latent variables, the use of simple parametric distributions like multivariate Gaussian is a limitation. Yet only a few of distributions are compatible with the reparameterization trick. Normalizing Flows (NFs) Rezende & Mohamed was proposed as a way to deal with more general parametric posterior distributions that can still be efficiently optimized with amortized inference (Papamakarios et al., 2019). NFs are a class of methods that use a series of invertible transformation $T_K ... \circ ... T_0$ to map a simple distribution z_0 into a complex one z_k: $z_k = T_K ... \circ ... T_0(z_0)$, By applying a cascade of transformations, the corresponding logarithm prior probability $p(z_k)$ of the transformed distribution becomes:

$$ln(p(z_k)) = \log(p(z_0)) - \Sigma_0^k \log \left| det \frac{\partial T_k}{\partial z_{k-1}} \right| \tag{1}$$

where the non-zero Jacobian $|det \frac{\partial T_k}{\partial z_{k-1}}|$ of each transformation ensures the global volume invariance of the probability density. The positivity of each Jacobian terms is guaranteed by the invertibility of each transformation T and consequently by the reversibility of normalizing flows.

The Hamiltonian dynamics in HVAE can also be seen as a type of NFs, for which Eq: (1) also holds. The Hamiltonian flow samples the posterior distribution $p(z|x)$ as the trajectory of a particle evolving as an Hamiltonian dynamical system under the potential energy given by $U(x, z) = -\log p(x, z)$ given an initial momentum ρ_0. The Hamiltonian dynamics is such that the distribution of phase space z, ρ remains constant along each trajectory according to Liouville's theorem (symplectic) (Fassò & Sansonetto, 2007). When using the leapfrog integrator with step size l for discretizing the Hamiltonian dynamics, the Jacobian remains closed to 1 with $\lim_{l \to 0} |det \frac{\partial T_k}{\partial z_k}|_l^{-1} = 1$. This property simplifies the Jacobian calculations at each discretization step (Caterini et al., 2018). In HVAE, the posterior approximation is constructed by applying K steps of the Hamiltonian flow (Caterini et al., 2018): $q^K(\mathcal{H}_k(\theta_0, \rho_0)) = q^0(\mathcal{H}_k(\theta_0, \rho_0)) \prod_{k=1}^K |det \nabla \Phi^k(\mathcal{H}_k(\theta_0, \rho_0))|^{-1}$, where Φ^k represents the leapfrog discretization transform of Hamiltonian dynamics. When combined with the reparameterization trick, it allows to compute an unbiased estimator of the lower bound gradients $\nabla_\theta \mathbb{L}$. Yet, since the leapfrog integrator involves two split sub-steps to update the momentum (Girolami & Calderhead, 2011), the computation of the lower bound gradient requires $2 \times K$ computations of the Jacobian ∇U.

2.2 LANGEVIN MONTE-CARLO AND NORMALIZING FLOW

In molecular physics, Langevin dynamics is a stochastic process that models the diffusion of free particles within a potential field $U(x)$ through a Stochastic Differential Equation (SDE): $m \partial v / \partial t := -\nabla U - \gamma v + \eta$, involving the particle velocity v, its acceleration $\partial v / \partial t$, damping factor γ and environmental noise η. Overdamped Langevin dynamics is obtained when the damping term is much greater than the inertial one leading to a first order SDE : $\gamma v = -\nabla U + \eta$.

As Langevin dynamics describes a stochastic evolution of particles be acted in the particle interaction potential $U(x)$ that can be treated as a log probability density, it has recently attracted a lot of attention

in the machine learning community (Stuart et al., 2004; Girolami & Calderhead, 2011; Welling & Teh, 2011; Mou et al., 2020) for the stochastic sampling of posterior distributions $p_\Phi(z|x)$ in Bayesian inference. Langevin Monte-Carlo methods (Girolami & Calderhead, 2011) rely on the construction of Markov chains with stochastic paths parameterized by Φ based on the discretization of the following *Langevin–Smoluchowski* SDE (Girolami & Calderhead, 2011) related to the overdamped Langevin dynamics :

$$\delta\Phi(t) = \frac{1}{2}\nabla_\Phi \log(p(x,\Phi))\delta t + \delta\sigma(t) \tag{2}$$

where $\sigma(t)$ is a Wiener process and t represents the time. The stochastic flow in Eq (2) can be further exploited to construct Langevin dynamics based normalizing flow and its derived methods for posterior inference (Wolf et al., 2016; Kobyzev et al., 2020). The concept of Langevin normalizing flow was first briefly sketched by Rezende & Mohamed (2015) in their seminal work (in section 3.2). To the best of our knowledge, little work has explored practical implementations of Langevin normalizing flows. In (Gu et al., 2019), the authors proposed a Langevin normalizing flow where invertible mappings are based on overdamped Langevin dynamics discretized with the Euler–Maruyama scheme. The explicit computation of the Jacobians of those mappings involves the Hessian matrix of $\log(p_\Phi(x))$ as follows :

$$\log\left|det\frac{\partial T_k}{\partial z_{k-1}}\right|^{-1} \sim \nabla_z\nabla_z \log(p(x,z)) + \mathbb{O}(z) \tag{3}$$

Yet, the Hessian matrix appearing in Eq (3) is expensive to compute both in space and time and adds a significant overhead to the already massive computation of gradients. This makes the method of (Gu et al., 2019) fairly unsuitable for the inference of complex models.

In a more generic view, in the Langevin flow, the forward transform is modelled by the Fokker-Plank equation and the backward transform is given by Kolmogorov's backward equation which is discussed in the work of Kobyzev et al. and is not detailed here.

2.3 Quasi-symplectic Langevin and Corresponding Flow

2.3.1 Quasi-symplectic Langevin Transform

To avoid the computation of Hessian matrices in Langevin normalizing flows, we propose to revert to the undamped or generalized Langevin dynamic process as proposed in (Sandev T., 2019). It involves second order dynamics with inertial and damping terms:

$$\begin{aligned}\delta\Phi(t) &= K\delta t \\ \delta K(t) &= -\frac{\partial ln(p(x,\Phi))}{\partial\Phi}\delta t - \nu K(t) + \delta\sigma(t)\end{aligned} \tag{4}$$

where $\Phi(t)$ and $K(t)$ are the stochastic position and velocity fields, and ν controls the amount of damping. We can see that the Langevin–Smoluchowski type SDE: (2) is nothing but the special case of high friction motion (Sandev T., 2019) when Eq: (4) has an over-damped frictional force.

Indeed, since $\delta K(t) = \frac{\partial\phi(t)^2}{\partial t^2}$, in the case of an over-damped system, the frictional force νK overwhelms the inertial force : $\delta^2\theta/\delta t^2 \ll \nu K(t)$. According to the generalized Langevin diffusion equation, we have :

$$\frac{\frac{\partial\Phi(t)^2}{\partial t^2}}{\nu K(t)} = \frac{\frac{\partial\log(p(x,\Phi))}{\partial\Phi}dt}{\nu K(t)} - 1 + \frac{\delta\sigma(t)}{\nu K(t)}$$

Therefore, we get :

$$\nu K(t) = \nu\delta\Phi \approx \frac{\partial\log(p(x,\Phi))}{\partial\Phi}dt + \delta\sigma(t)$$

To get simple Jacobian expressions when constructing Langevin flow, we need to have a symplectic Langevin transformation kernel. To this end, we introduce a quasi-symplectic Langevin method for building the flow (Milstein et al., 2002). The quasi-symplectic Langevin differs from the Euler–Maruyama integrator method which diverges for the discretization of generalized Langevin SDE. Instead, the quasi-symplectic Langevin method makes the computation of the Jacobian tractable

during the diffusion process and keeps approximate symplectic properties for the damping and external potential terms.

More precisely, the quasi-symplectic Langevin integrator is based on the two state variables (K_i, Φ_i) that are evolving according to the mapping $\Psi_\sigma(K_i, \Phi_i) = (K_{i+1}, \Phi_{i+1})$ where σ is the kernel stochastic factor. It is known as the *second order strong quasi-symplectic* method (5) and is composed of the following steps for a time step τ:

$$\begin{aligned} &K_{II}(t,p) = pe^{-\nu t} \\ &K_{1,i} = K_{II}(\frac{\tau}{2}, K_i);\ \ \Phi_{1,i} = \Phi_i - \tau\frac{\partial \log(p(x,\Phi))}{2\partial\Phi} \\ &K_{2,i} = K_{1,i} + \sqrt{\tau}\sigma\xi_i;\ \ \xi_i \sim N(0, I) \\ &K_{i+1} = K_{II}(\frac{\tau}{2}, K_{2,i});\ \ \Phi_{i+1} = \Phi_{1,i} + \frac{\tau}{2}K_{2,i} \end{aligned} \tag{5}$$

where initial conditions are $K_0 = \kappa_0; \Phi_0 = \phi_0$. To guarantee the reversibility of the defined kernel, the mapping has to be deterministic and therefore we use $\sigma = 0$ in the remainder.

The above quasi-symplectic integrator satisfies the following two properties:

Property 1. *Quasi-symplectic method degenerates to a symplectic method when $\nu = 0$.*

Property 2. *Quasi-symplectic Langevin transform $\Psi_0(K_i, \Phi_i)$ (5) has constant Jacobian :*

$$|\Psi_0(K_i, \Phi_i)| = \frac{\partial\Phi_{i+1}}{\partial\Phi_i}\frac{\partial K_{i+1}}{\partial K_i} - \frac{\partial\Phi_{i+1}}{\partial K_i}\frac{\partial K_{i+1}}{\partial\Phi_i} = exp(-\nu\tau) \tag{6}$$

The first property shows that the VAE constructed based on the Quasi-Symplectic Langevin (QSL) dynamics is conceptually equivalent to a HVAE in the absence of damping $\nu = 0$. The second property implies that the QSL integrator leads to transformation kernels that are reversible and with trivial Jacobians. The proofs of those two properties can be found in Milstein (2003).

The advantage of the QSL flow compared to the regular overdamped Langevin flow is that it avoids computing the Hessian of the log probability to compute the Jacobian which is a major advantage given the complexity of the Hessian computation.

We give the below formal definition of the quasi-symplectic Langevin normalizing flow.

Definition. *A K steps discrete quasi-symplectic Langevin normalizing flow $\mathcal{L}^K$ is defined by a series of diffeomorphism, bijective and invertible mapping $\Psi_0 : \sigma_\mathcal{A} \to \sigma_\mathcal{B}$ between two measurable spaces $(\mathcal{A}, \sigma_\mathcal{A}, \mu_\alpha)$ and $(\mathcal{B}, \sigma_\mathcal{B}, \mu_\beta)$:*

$$\begin{aligned} \mathcal{L}^K\mu_\alpha(\mathcal{S}_\mathcal{A}) :&\Psi_i \circ \mu_\alpha(\mathcal{S}_\mathcal{A}) = \mu_\alpha(\Psi_{i-1}^{-1}(\mathcal{S}_\mathcal{B})), \\ &\forall\mathcal{S}_\mathcal{A} \in \sigma_\mathcal{A}, \mathcal{S}_\mathcal{B} \in \sigma_\mathcal{B}, i = \{1, ..., K\}. \end{aligned} \tag{7}$$

where $\sigma_{(\cdot)}$ and $\mu_{(\cdot)}$ are the σ-algebra and probability measure for set $(\cdot)$ respectively, Ψ_i is the i_{th} quasi-symplectic Langevin transform given by Eqs:(5).

2.3.2 Example for single step quasi-symplectic Langevin flow

We illustrate below definition 2.3.1 of a quasi-symplectic Langevin normalizing flow in case of a single transform applied on a single random variable. We consider a probability measure $p(x)$ of random variable set $x \in X$. Then a single step Langevin flow transforms the original random variable x to a new random variable $y = \Psi_0(x), y \in Y$. According to definition 2.3.1, the new probability measure $q(y)$ of random variable y is given by:

$$q(y) = \mathcal{L}^0 p(x) : \Psi_0 \circ p(x) = p(\Psi_0^{-1}(y)) \tag{8}$$

By Eq.(1), we conclude that:

$$q(y) = p(x) \cdot |det\frac{\partial\Psi_0}{\partial x}|^{-1} \tag{9}$$

The defined quasi-symplectic Langevin flow is a generalization of the Langevin flow with a quasi-symplectic structure for the parameters phase space. The quasi-symplectic Langevin normalizing flow has a deterministic kernel Ψ_0 when the kernel stochastic factor $\sigma = 0$, and degenerates to a symplectic transition when $\nu = 0$. With a deterministic kernel, the flow will be able to have the probability density evaluated.

3 QUASI-SYMPLECTIC LANGEVIN VAE

3.1 LOWER BOUND ESTIMATION WITH LVAE

In the quasi-symplectic Langevin VAE, we use an augmented latent space consisting of position ϕ_K and velocity κ_K variables of dimension $\zeta : z = (\phi_K, \kappa_K)$. The objective of the autoencoder is to optimize its parameters as to maximize the evidence lower bound $\tilde{\mathbb{L}}$:

$$\begin{aligned} &\log p(x) = \log \int_\Omega p(x,z)dz = \log \int_\Omega \tilde{p}(x)q(\tilde{z}|x)d\tilde{z} \\ &\geq \int_\Omega \log \tilde{p}(x)q(\tilde{z}|x)d\tilde{z} \equiv \tilde{\mathbb{L}} \end{aligned} \tag{10}$$

where Ω is the measure space of the latent variables and as $\tilde{p}(x)$ is an unbiased estimator for $p(x)$. The lower bound is equal to the evidence when the posterior approximation is equal to the true posterior. Thus maximizing the lower bound is equivalent to minimize the gap between the true posterior $p(z|x)$ and its approximation $q(z|x)$ (Blei et al., 2017).

Algorithm 1: Quasi-symplectic Variational Inference

Inputs: Data X, Inference steps K, damping ν, time step τ, prior $q^0_{\omega_E}(\phi_0)$
Output: Encoding and decoding parameters $\omega = (\omega_E, \omega_D)$
Initialize all parameters, variables;
Define: $K_{II}(t,p) = pe^{-\nu t}$;
while *NOT* ω *converged* **do**
  Get minibatch: $X_N \overset{N}{\leftarrow} X$;
  while *NOT j = N* **do**
    $x_j \overset{j}{\leftarrow} X_N$; // Get x_j in minibatch.
    $\phi_0 \sim q^0_{\omega_E}(\phi_0|x_j)$; // Sampling latent variable from variational prior
    $\kappa_0 \sim \mathcal{N}(0, I_\zeta)$; // Sampling velocity from unit Gaussian.
    for $i = 1;\ i < K;\ i++$ **do**
      // Quasi-symplectic Langevin Normalizing Flow
      $\kappa_{1,i} \leftarrow K_{II}(\frac{\tau}{2}, \kappa_i)$; $\phi_{1,i} \leftarrow \phi_i - \tau \frac{\partial \log(p(x,\phi_))}{2\partial \phi_i}$;
      $\kappa_{i+1} \leftarrow K_{II}(\frac{\tau}{2}, \kappa_{1,i})$; $\phi_{i+1} \leftarrow \phi_{1,i} + \frac{\tau}{2}\kappa_{1,i}$;
    end
    $p^*_\omega \leftarrow \hat{p}_{\omega_D}(x, \phi_K) \cdot \mathcal{N}(\kappa_K|0, I_\zeta)$;
    $q^*_\omega \leftarrow q^0_{\omega_E}(\phi_0) \cdot \mathcal{N}(\kappa_0|0, I_\zeta) exp(\nu\tau)$;
    $\tilde{\mathbb{L}}^*_j \leftarrow \ln(p^*_\omega) - \ln(q^*_\omega)$; ; // Quasi-symplectic Langevin ELBO
    $j \leftarrow j + 1$
  end
  $\tilde{\mathbb{L}}^* \leftarrow \sum_{i=1}^N \tilde{\mathbb{L}}^*_i / N$; // Minibatch average ELBO
  $\arg\max_{\omega \in \mathbb{R}^n} \tilde{\mathbb{L}}^*$; // Optimize average ELBO over parameters subset
end

The posterior approximation $q(z)$ is computed through a series of Langevin transformations which is the Langevin flow: $q_{\omega_E}(z|x) = q^K(\mathcal{L}^K(\phi_0, \kappa_0)) = q^0_{\omega_E}(\phi_0, \kappa_0|x) \prod_{k=1}^K |det \nabla \Psi_0(\phi_k, \kappa_k))|^{-1} = q^0_{\omega_E}(\phi_0, \kappa_0|x) \exp(K\nu\tau)$, where $q^0_{\omega_E}(\phi_0, \kappa_0|x)$ is an initial approximation parameterized by ω_E which can also be seen as the prior on random variables ϕ_0, κ_0.

Similarly to the HVAE (Caterini et al., 2018), an unbiased estimator of $p(x)$ is given by :

$$\tilde{p}(x) = \frac{\hat{p}(x, \mathcal{L}^K(\theta_0, k_0))}{q^0_{\omega_E}(\theta_0, \kappa_0)} \tag{11}$$

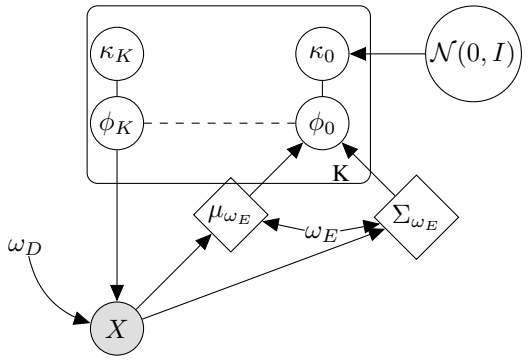


Figure 1: Graphical model of the Quasi-symplectic Langevin Variational Autoencoder. The multivariate Gaussian parameters $\mu_{\omega_E}, \Sigma_{\omega_E}$ defining the variational prior of latent variable ϕ_0 are determined from the data X and the parameter ω_E of the encoding network. The initial velocity latent variable κ_0 has a unit Gaussian prior and is paired by initial latent variable ϕ_0. After iterating K times the quasi-symplectic Langevin transform, the latent pair $\{\phi_K, \kappa_K\}$ is obtained from the initial variables $\{\phi_0, \kappa_0\}$. The decoder network with parameters ω_D is then used to predict the data from latent variables ϕ_K through the conditional likelihood $p(x|\phi_K)$. Variables in diamonds are deterministically computed. Network parameters ω_E, ω_D are optimized to maximize the ELBO.

We then give the lower bound for the quasi-symplectic Langevin VAE as :

$$\begin{aligned}\tilde{\mathbb{L}} := \int_{\Omega} q_{\omega_E}(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^K(\phi_0, k_0)) - \\ \log(q^0_{\omega_E}(\phi_0, k_0)) + K\nu\tau d\tilde{z}\end{aligned} \tag{12}$$

3.2 Quasi-symplectic Langevin VAE

The quasi-symplectic Langevin lower bound $\tilde{\mathbb{L}}$ lays the ground for the stochastic inference of a variational auto-encoder. Given a set of dataset $X : \{x^i \in X; i \in \mathbb{N}_+\}$, we aim to learn a generative model of that dataset from a latent space with the quasi-symplectic Langevin inference. The generative model $p(x, z)$ consists of a prior on initial variables $z_0 = (\phi_0, \kappa_0)$, $q^0_{\omega_E}(\phi_0, \kappa_0|x) = q^0_{\omega_E}(\phi_0|x) \cdot \mathcal{N}(\kappa_0|0, I_\zeta)$ and conditional likelihood $p_{\omega_D}(x|z)$ parameterized by ω_D. The Gaussian unit prior $\mathcal{N}(\kappa_0|0, I_\zeta)$ is the canonical velocity distribution from which the initial velocity of the Langevin diffusion will be performed. The distribution $q^0_{\omega_E}(\phi_0|x)$ is the variational prior that depends on the data x^i. Thus the generative model $p_{\omega_E,\omega_D}(x, z)$ is parameterized by both encoders and decoders and the quasi-symplectic Langevin lower bound writes as:

$$\begin{aligned}\arg\max_{\omega_E,\omega_D} \tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0 \sim q^0_{\omega_E}(), \kappa_0 \sim \mathcal{N}_\zeta()}(\ln \hat{p}_{\omega_E,\omega_D}(x, \mathcal{L}^K(\phi_0, \kappa_0)) - \\ & \ln(q^0_{\omega_E}(\phi_0, \kappa_0)) + K\nu\tau)\end{aligned} \tag{13}$$

The maximization of the lower bound (13), is can be performed efficiently with the reparameterization trick depending on the choice of the variational prior $q^0_{\omega_E}(\phi_0)$. To have a fair comparison with prior work(Caterini et al., 2018), we also perform Rao-Blackwellization for reducing the variance of the ELBO in the quasi-symplectic Langevin VAE:

$$\begin{aligned}\arg\max_{\omega_E,\omega_D} \tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0 \sim q^0_{\omega_E}(), \kappa_0 \sim \mathcal{N}_\zeta()}(\ln \hat{p}_{\omega_E,\omega_D}(x, \mathcal{L}^K(\phi_0, \kappa_0)) \\ & - \ln(\hat{q}_{\omega_E}(\phi_0, \kappa_0)) \\ & + K\nu\tau - \frac{1}{2}\kappa_K^T\kappa_K) + \frac{\zeta}{2}; \quad \forall \phi_0, \kappa_0 \in \mathbb{R}^\zeta\end{aligned} \tag{14}$$

The resulting algorithm is described in Alg.1.

4 EXPERIMENT AND RESULT

We examine the performance of quasi-symplectic Langevin VAE on the MNIST dataset (LeCun et al., 2010) based on various metrics. Caterini et al. (2018) have reported that the Hamiltonian based stochastic variational inference outperforms that of Planar Normalizing Flow, mean-field based Variational Bayes in terms of model parameters inference error and quantitatively shown that the HVAE outperforms the naive VAE method in terms of Negative Log-likelihood (NLL) score and ELBO. Here, we compare the proposed LVAE with the HVAE on MNIST dataset. The experiments were implemented with *TensorFlow 2.0* and *TensorFlow Probability* framework to evaluate the proposed approach in both qualitative and quantitative metrics.

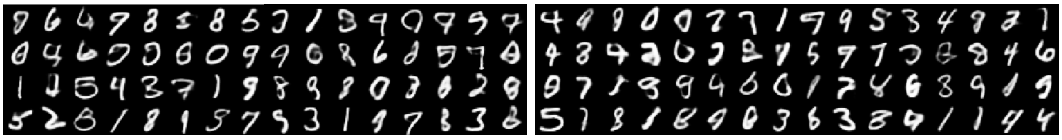


Figure 2: Quantitative result of Langevin VAE in comparison with HVAE. Left sub-figures are generated samples of HVAE. Right are samples of Langevin-VAE. In both methods, the number of steps in the flow computation is $K = 5$.

Table 1: Quantitative evaluation of the Langevin-VAE in comparison with the HVAE. It includes the comparison of the negative log likelihoods (NLL), the evidence lower bound (ELBO), the Fréchet Inception (FID) and Inception Score (IS)

	Langevin-VAE			HVAE		
Flow steps	1	5	10	1	5	10
NLL	**89.41**	***88.15***	**89.63**	89.60	88.21	89.69
ELBO	**-91.74**	***-90.14***	**-92.03**	-91.91 $\pm$ 0.01	-90.41	-92.39 $\pm$ 0.01
FID	***52.70***	**52.95**	**53.13**	53.12	53.26	53.21
IS	6.42	***6.49***	**6.30**	**6.57**	6.42	6.11

4.1 QUASI-SYMPLECTIC LANGEVIN VAE ON BINARY IMAGE BENCHMARK

Given a training dataset $X : \{x^i \in X; i \in \mathbb{N}_+\}$ consisting of binary images of size d, $x^i \in \{0, 1\}^d$, we define the conditional likelihood $p(x|z)$ as a product of d Bernoulli distributions. More precisely, we consider a decoder neural network $\mathrm{Dec}_{\omega_D}(\phi) \in [0, 1]^d$ that outputs d Bernoulli parameters from the latent variable $\phi \in \mathbb{R}^\zeta$ where $z = (\phi, \kappa)$. Then the conditional likelihood writes as : $p(x^i|z^i) = \prod_{j=1}^{d} \mathrm{Dec}_{\omega_D}(\phi)[j]^{x^i[j]} \, (1 - \mathrm{Dec}_{\omega_D}(\phi)[j])^{1-x^i[j]}$.

Following the classical VAE approach (Kingma & Welling, 2014), the encoder network parameterized by ω_E outputs multivariate Gaussian parameters : $\mu_{\omega_E}(x) \in \mathbb{R}^\zeta$ and $\Sigma_{\omega_E}(x) \in \mathbb{R}^\zeta$, such that the variational prior is a multivariate Gaussian $q^0_{\omega_E}(\phi_0|x) = \mathcal{N}(\phi_0|\mu_{\omega_E}(x), \Sigma_{\omega_E}(x))$ with diagonal covariance matrix. This choice obviously makes the reparameterization trick feasible to estimate the lower bound. The related graphical model of the quasi-symplectic Langevin VAE is displayed in Fig. 1.

The decoder and encoder neural network architectures are similar to the HVAE (Caterini et al., 2018) and MCMCVAE (Salimans et al., 2015), both having three layers of 2D convolutional neural networks for encoder and decoder, respectively. The encoder network accepts a batch of data of size $(N_b \times 28 \times 28)$ with $N_b = 1000$. The dimension of latent variables is set as $\zeta = 64$ and the damping factor is $\nu = 0$. The discretization step is $\tau = 1e - 2$. The training stage stops when the computed ELBO does not improve on a validation dataset after 100 steps or when the inference loop achieves 2000 steps.

Both tested models LVAE and HVAE share the same training parameters. The stochastic ascent of the ELBO is based on the Adamax optimizer with a learning rate $lr = 5e - 5$. All estimation of computation times were performed on one NVIDIA GeForce GTX 1080 Ti GPU. For memory usage evaluation the experiment was performed on a NVIDIA Quadro M2200 GPU.

4.2 RESULT

Both qualitative and quantitative results are studied. The generated samples of Langevin-VAE and HVAE are shown in Fig: (2). We qualitatively see that the quality and diversity of the sampled images are guaranteed for both autoencoder models. Quantitatively, Table 1 shows the performance in terms of the NLL, ELBO, FID, IS scores for Langevin-VAE and HVAE where Langevin and Hamiltonian flows are experimentally compared.

5 CONCLUSION

In this paper, we propose a new flow-based Bayesian inference framework by introducing the quasi-symplectic Langevin flow for the stochastic estimation of a tight ELBO.

By introducing the quasi-symplectic Langevin dynamics, we also overcome the limitation of the Langevin normalizing flow (Gu et al., 2019) which requires to provide the Hessian matrix $\nabla\nabla log(p(x, \phi))$ to compute the Jacobian. Potential improvements of the quasi-symplectic Langevin inference can arise by investigating the manifold structure of the posterior densities of the latent variables (Girolami & Calderhead, 2011; Barp et al., 2017; Livingstone & Girolami, 2014) to improve the inference efficiency.

ACKNOWLEDGMENTS

This work was partially funded by the French government through the UCA JEDI "Investments in the Future" project managed by the National Research Agency (ANR) with the reference number ANR-15-IDEX-01, and was supported by the grant AAP Santé 06 2017-260 DGA-DSH.

REFERENCES

Alessandro Barp, Francois-Xavier Briol, Anthony Kennedy, and Mark Girolami. Geometry and dynamics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 5, 05 2017. doi: 10.1146/annurev-statistics-031017-100141.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

Anthony L. Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. In *NeurIPS*, 2018.

Francesco Fassò and Nicola Sansonetto. Integrable almost-symplectic hamiltonian systems. *Journal of Mathematical Physics*, 48(9):092902, 2007. doi: 10.1063/1.2783937. URL https://doi.org/10.1063/1.2783937.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011. doi: 10.1111/j.1467-9868.2010.00765.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00765.x.

Minghao Gu, Shiliang Sun, and Yan Liu. Dynamical sampling with langevin normalization flows. *Entropy*, 21:1096, 11 2019. doi: 10.3390/e21111096.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

I. Kobyzev, S. Prince, and M. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Samuel Livingstone and Mark A. Girolami. Information-geometric markov chain monte carlo methods using diffusions. *Entropy*, 16:3074–3102, 2014.

G. Milstein. Quasi-symplectic methods for langevin-type equations. *IMA Journal of Numerical Analysis*, 23:593–626, 10 2003. doi: 10.1093/imanum/23.4.593.

G. N. Milstein, Yu. M. Repin, and M. V. Tretyakov. Symplectic integration of hamiltonian systems with additive noise. *SIAM Journal on Numerical Analysis*, 39(6):2066–2088, 2002. doi: 10.1137/S0036142901387440. URL https://doi.org/10.1137/S0036142901387440.

Wenlong Mou, Yi-An Ma, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm, 2020.

George Papamakarios, Eric T. Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *ArXiv*, abs/1912.02762, 2019.

R. Ranganath, Sean Gerrish, and D. Blei. Black box variational inference. In *AISTATS*, 2014.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/rezende15.html.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1218–1226, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/salimans15.html.

Tomovski Ž. Sandev T. Generalized langevin equation. *Fractional Equations and Models. Developments in Mathematics*, 61, 2019. URL https://doi.org/10.1007/978-3-030-29614-8_6.

Andrew M. Stuart, Jochen Voss, and Petter Wilberg. Conditional path sampling of sdes and the langevin mcmc method. *Commun. Math. Sci.*, 2(4):685–697, 12 2004. URL https://projecteuclid.org:443/euclid.cms/1109885503.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Christopher Wolf, Maximilian Karl, and Patrick van der Smagt. Variational inference with hamiltonian monte carlo, 2016.