# Quasi-Symplectic Langevin Variational Autoencoder

Zihao Wang, Hervé Delingette

# 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 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 an effective approach to tighten the ELBO for approximating the posterior distribution and Hamiltonian Variational Autoencoder (HVAE) is an effective MCMC inspired approach for constructing a low-variance ELBO that is amenable to the reparameterization trick. The HVAE adapted the Hamiltonian dynamic flow into variational inference that significantly improves the performance of the posterior estimation. We propose in this work a Langevin dynamic flow-based inference approach by incorporating the gradients information in the inference process through the Langevin dynamic which is a kind of MCMC based method similar to HVAE. Specifically, we employ a quasi-symplectic integrator to cope with the prohibit problem of the Hessian computing in naive Langevin flow. We show the theoretical and practical effectiveness of the proposed framework with other methods, which reaches the best negative log-likelihood on benchmark dataset. The Langevin-VAE is used for modeling a clinical medical image dataset.

## 1 INTRODUCTION

Variational Autoencoder (VAE) is a popular generative neural model, which is applied in a vast number of practical cases to perform unsupervised analysis or to modeling a 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). One of the key factors influencing the performance of VAE models is the quality of the marginal 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. To increase the flexibility of the posterior modeling, the Hamiltonian Variational Inference (HVI) is proposed to remove the requirement of an explicit formulation of the posterior distribution by forwarding sampling a Markov chain based on Hamiltonian dynamics (Salimans et al., 2015). 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 HVI algorithm. Further more, Caterini et al. (2018) first combined VAE and HVI in Hamiltonian Variational Autoencoders (HVAE) which include a dynamic phase space where momentum component $\rho$ and position component $z$ are integrated. The using of Hamiltonian flow for the latent distribution inference can introduce the target information (gradient flow) into the inference steps for improving the variational inference efficiency.

*Corresponding author.

In this work,we propose a novel inference framework named quasi-symplectic Langevin variational auto-encoder (Langevin-VAE) that leads to reversible Markov kernels and phase quasi-volume invariance. The major contributions of this paper are:

- The proposed MCMC method is a Langevin-flow-based asymptotic low variance unbiased lower bound estimator.
- Different from prior Langevin normalizing flow, this approach is a generalized Langevin flow-based inference framework, which avoids computing the Hessian.
- The method shows comparable performance through quantitative and qualitative comparison with conventional VAE and some prior inference frameworks on benchmark and real-world datasets.

## 2 PRELIMINARY

### 2.1 VARIATIONAL INFERENCE AND NORMALIZING FLOW

**Variational Inference** 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, which performs direct stochastic optimization of ELBO. Kingma & Welling proposed to use some multivariate Gaussian posterior distributions of latent variable $z$ generated by a universal function $\omega$, which makes reparameterization trick is possible.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) is an effective way to deal with this limitation, which constructs a mapping between the complex and simple distributions by gradients transform.

**Normalizing Flows** Rezende & Mohamed proposed the NFs as a way to deal with more general parametric posterior distributions that can still be efficiently optimized with amortized inference (Papamakarios et al., 2019). Briefly, NFs are a class of methods that use a series of invertible transformations $\mathcal{T}_I ... \circ ... \mathcal{T}_0$ to map a simple distribution $z_0$ into a complex one $z_i$: $z_i = \mathcal{T}_I ... \circ ... \mathcal{T}_0(z_0)$, By applying a cascade of transformations, the corresponding logarithm prior probability $p(z_i)$ of the transformed distribution becomes:

$$\log(p(z_i)) = \log(p(z_0)) - \Sigma_0^i \log \left| det \frac{\partial \mathcal{T}_i}{\partial z_{i-1}} \right| \tag{1}$$

where the non-zero Jacobian $|det \frac{\partial \mathcal{T}_i}{\partial z_{i-1}}|$ 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 $\mathcal{T}$ and consequently by the reversibility of normalizing flows.

**Hamiltonian Flows** The Hamiltonian dynamics in HVAE can also be seen as a type of NFs, for which Eq: (1) also holds. Briefly, HVAE employs an $I$ steps Hamiltonian transformation process: $\mathcal{H}_I$ to build an unbiased estimation of posterior $q(z)$ by extending $\tilde{p}(x,z)$ as $\tilde{p}(x, \mathcal{H}_I(z_0, \rho_0))$ leading to: $\tilde{p}(x) := \frac{\hat{p}(x, \mathcal{H}_I(z_0,\rho_0))}{q(\mathcal{H}_I(z_0,\rho_0))}$, where: $\hat{p}(x, z_I, \rho_I) = \hat{p}(x, \mathcal{H}_I(z_0,\rho_0)) = \hat{p}(x, z_I)\mathcal{N}(\rho_I|0, I)$. In particular, the HVAE enables the use of the reparameterization trick during inference thus leading to efficient ELBO gradients computation. The Hamiltonian dynamics is such that the distribution of phase space $(z, \rho)$ [1] remains constant along each trajectory according to Liouville's theorem (symplectic) (Fassò & Sansonetto, 2007). When using the leapfrog integrator with step size $t$ for discretizing the Hamiltonian dynamics, the Jacobian remains to 1 (ignoring numerical rounding errors) with $|det \frac{\partial \mathcal{H}_i}{\partial z_i}|_t^{-1} = 1$. This property simplifies the Jacobian calculations at each discretization step (Caterini et al., 2018). In HVAE, the posterior approximation is constructed by applying $I$ steps of the Hamiltonian flow: $q^I(\mathcal{H}_i(\theta_0,\rho_0)) = q^0(\mathcal{H}_i(\theta_0,\rho_0)) \prod_{i=1}^I |det \nabla \Phi^i(\mathcal{H}_i(\theta_0,\rho_0))|^{-1}$, where $\Phi^i$ 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 $\nabla_\theta \mathbb{L}$.

[1] https://en.wikipedia.org/wiki/Phase_space

### 2.2 LANGEVIN MONTE-CARLO AND LANGEVIN FLOW

A Langevin dynamics describes a stochastic evolution of particles within 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., 2021) 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 $\Phi$ based on the discretization of the following *Langevin–Smoluchowski* SDE (Girolami & Calderhead, 2011) related to the overdamped Langevin dynamics :

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

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. 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 :

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

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 TRIVIAL JACOBIAN BY QUASI-SYMPLECTIC LANGEVIN TRANSFORM

To avoid the computation of Hessian matrices in Langevin normalizing flows, we propose to revert to 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 $\nu$ controls the amount of damping. We can see that the Langevin–Smoluchowski type SDE of Eq.:(2) is nothing but the special case of high friction motion (Sandev T., 2019) when Eq.: (4) has an over-damped frictional force (proof is in A.1).

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 $(K_i, \Phi_i)$ that are evolving according to the mapping $\Psi_\sigma(K_i, \Phi_i) = (K_{i+1}, \Phi_{i+1})$ where $\sigma$ 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 $t$:

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

where initial conditions are $K_0 = \kappa_0; \Phi_0 = \phi_0$.

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 :*

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

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 Langevin-VAE integrator leads to transformation kernels that are reversible and with trivial Jacobians. The proofs of those two properties can be found in appendix A.2 and more discussion about the quasi-symplectic integrator can be found in Milstein (2003). The advantage of the QSL flow compared to the regular Langevin flow is that it avoids computing the Hessian of the log probability, which is a major advantage given the complexity of the Hessian computation.

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

**Definition.** *An $I$ steps discrete quasi-symplectic Langevin normalizing flow $\mathcal{L}^I$ 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 $(\mathcal{B}, \sigma_{\mathcal{B}}, \mu_\beta)$:*

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

*where $\sigma_{(\cdot)}$ and $\mu_{(\cdot)}$ are the $\sigma$-algebra and probability measure for set $(\cdot)$ respectively, $\Psi_i$ 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 for a Langevin flow 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 $\Psi_0$ when the kernel stochastic factor $\sigma = 0$, and degenerates to a symplectic transition when $\nu = 0$.

# 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 $\phi_I$ and velocity $\kappa_I$ variables of dimension $\zeta : z = (\phi_I, \kappa_I)$. The objective of the autoencoder is to optimize its parameters as to maximize the evidence lower bound $\tilde{\mathbb{L}}$:

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

where $\Omega$ 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 I, damping $\nu$, time step $t$, 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* $\omega$ *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
    $\kappa_0 \sim \mathcal{N}(0, E_\zeta)$ ; // Sampling velocity from unit Gaussian.
    **for** $i = 1;\ i < I;\ i++$ **do**
      // Quasi-symplectic Langevin Transform
      $\kappa_{1,i} \leftarrow K_{II}(\frac{t}{2}, \kappa_i)$; $\phi_{1,i} \leftarrow \phi_i - t\frac{\partial \log(p(x,\phi_)}{2\partial \phi_i}$;
      $\kappa_{i+1} \leftarrow K_{II}(\frac{t}{2}, \kappa_{1,i})$; $\phi_{i+1} \leftarrow \phi_{1,i} + \frac{t}{2}\kappa_{1,i}$;
    **end**
    $p^*_\omega \leftarrow \hat{p}_{\omega_D}(x, \phi_I) \cdot \mathcal{N}(\kappa_I|0, E_\zeta)$;
    $q^*_\omega \leftarrow q^0_{\omega_E}(\phi_0) \cdot \mathcal{N}(\kappa_0|0, E_\zeta) exp(\nu t)$;
    $\tilde{\mathbb{L}}^*_j \leftarrow \log(p^*_\omega) - \log(q^*_\omega)$; ; // 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: $q_{\omega_E}(z|x) = q^I(\mathcal{L}^I(\phi_0, \kappa_0)|x) = q^0_{\omega_E}(\phi_0, \kappa_0|x) \prod_{i=1}^I |det \nabla \Psi_0(\phi_i, \kappa_i))|^{-1} = q^0_{\omega_E}(\phi_0, \kappa_0|x) \exp(I\nu t)$, where $q^0_{\omega_E}(\phi_0, \kappa_0|x)$ is an initial approximation parameterized by $\omega_E$ which can also be seen as the prior on random variables $(\phi_0, \kappa_0)$.

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

$$\tilde{\mathbb{L}} := \int_\Omega q_{\omega_E}(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\phi_0, k_0)) - log(q^0_{\omega_E}(\phi_0, k_0)) + I\nu t d\tilde{z} \tag{11}$$

## 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.

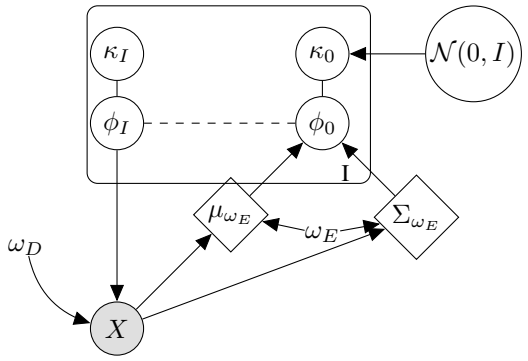


Figure 1: Graphical model of the Quasi-symplectic Langevin Variational Autoencoder. The multivariate Gaussian parameters $\mu_{\omega_E}, \Sigma_{\omega_E}$ defining the variational prior of latent variable $\phi_0$ are determined from the data $X$ and the parameter $\omega_E$ of the encoding network.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

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 $\omega_D$. 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:

$$\begin{aligned}\arg\max_{\omega_E,\omega_D} \tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0 \sim q^0_{\omega_E}(\cdot), \kappa_0 \sim \mathcal{N}_\zeta(\cdot)}(\log \hat{p}_{\omega_E,\omega_D}(x, \mathcal{L}^I(\phi_0, \kappa_0)) - \\ & \log(q^0_{\omega_E}(\phi_0, \kappa_0)) + K\nu t)\end{aligned} \tag{12}$$

The maximization of the lower bound (12) 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:

$$\begin{aligned}\arg\max_{\omega_E,\omega_D} \tilde{\mathbb{L}}^* =& \mathbb{E}_{\phi_0 \sim q^0_{\omega_E}(\cdot), \kappa_0 \sim \mathcal{N}_\zeta(\cdot)}(\log \hat{p}_{\omega_E,\omega_D}(x, \mathcal{L}^I(\phi_0, \kappa_0)) \\ & - \log(\hat{q}_{\omega_E}(\phi_0, \kappa_0)) + K\nu t - \frac{1}{2}\kappa_I^T \kappa_I) + \frac{\zeta}{2}; \quad \forall \phi_0, \kappa_0 \in \mathbb{R}^\zeta\end{aligned} \tag{13}$$

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 different methods in both qualitative and quantitative metrics.

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 $\text{Dec}_{\omega_D}(\phi) \in [0, 1]^d$ that outputs $d$ Bernoulli parameters

from the latent variable $\phi \in \mathbb{R}^{\zeta}$ where $z = (\phi, \kappa)$. Then the conditional likelihood writes as : $p(x^i|z^i) = \prod_{j=1}^{d} \mathrm{Dec}_{\omega_D}(\phi)[j]^{x^i[j]} \ (1 - \mathrm{Dec}_{\omega_D}(\phi)[j])^{1-x^i[j]}$.

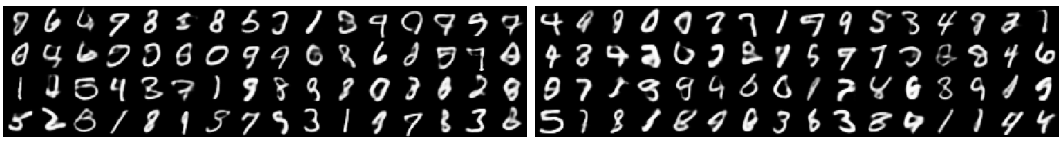

Figure 2: Quantitative result of Langevin VAE in comparison with HVAE. Left sub-figures are generated samples of HVAE. Right are samples of Langevin-VAE. In both methods, the number of steps in the flow computation is $K = 5$.

Table 1: Quantitative evaluation of the Langevin-VAE in comparison with the HVAE, IWAE, DBN, and DAN. It includes the comparison of the negative log likelihoods (NLL), the evidence lower bound (ELBO), and Inception Score (IS) (Borji, 2019)

| | **Langevin-VAE** | | **HVAE** | | **IWAE** | **DBN** | **DAN** |
|---|---|---|---|---|---|---|---|
| Flow steps | 2 | 5 | 2 | 5 | - | - | - |
| NLL | 82.95 | *82.40* | 83.10 | 82.75 | 82.90 | 84.55 | 84.13 |
| ELBO | -85.37 | *-84.81* | -85.70 | -85.29 | - | - | - |
| IS | 7.67 | *7.76* | 7.59 | 7.38 | - | - | - |

### 4.1 Quasi-symplectic Langevin VAE on binary image benchmark

#### 4.1.1 Implementation details

Following the classical VAE approach (Kingma & Welling, 2014), the encoder network parameterized by $\omega_E$ 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 set variant: $t \in [t_a, t_b]$. The training stage stops when the computed ELBO does not improve on a validation dataset after 200 steps or when the inference loop achieves 2000 epochs. The scale term $\sigma$ of Langevin dynamic was set as: $2\sqrt{T}$, where $T$ is the temperature.

Both tested models LVAE and HVAE share the same training and testing parameters except for unique parameters (detailed in A.4). 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 an NVIDIA GeForce GTX 1080 Ti GPU.

#### 4.1.2 Result on MNIST

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, IS scores for different Langevin-VAE and HVAE where Langevin and Hamiltonian flows are experimentally compared. In addition, we compare the negative log-likelihood of the two flow-based frameworks with Importance Weighted Autoencoder (IWAE) (Huang et al., 2019), Deep Belief Networks (DBNs) (Murray & Salakhutdinov, 2009), and Deep Autoregressive Networks (DANs) (Gregor et al., 2014). We observe an obvious number difference between flow-based and non-flow-based methods. What's more special is the IWAE, which achieved a comparable result (82.90 nats) against (82.40 nats) with Langevin-VAE. Yet, this was the best performance of IWAE (Huang et al., 2019), which achieved through a k-sample ($k = 5000$) importance weighting

MCMC. We notice that the Langivin-VAE reaches the best performance among all the evaluation metrics in comparison with the other methods on the MNIST benchmark.

One of the drawbacks of Jacobi flow-based methods is the time and space overhead of the gradient calculation. The HVAE requires $k + 1$ or $2 \times k$ times access of the Jacobi depends on with or without gradients reusing, which requires a compensation between the memory and time cost. In contrast, the Langevin-VAE relies on $k$ times of gradients computation, which is relatively more economical.

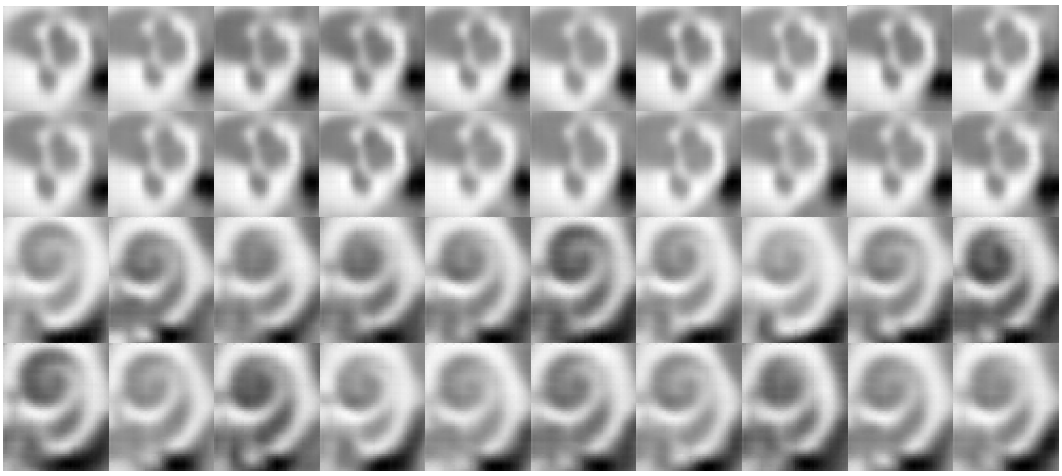

Figure 3: Qualitative assessment of the generated samples of Langevin-VAE. The number of steps in the flow computation is $K = 5$.

### 4.2 Quasi-symplectic Langevin VAE on Medical Image dataset

We employ the proposed method for inference the cochlea dataset to generate the cochlea CT images. The human cochlea is an auditory nerve organ with a spiral shape. Some severe hearing impairments can be treated with cochlear transplantation. The shape of the cochlea is of great significance to the formulation of preoperative and postoperative plans. The quantitative analysis of the shape of the cochlea needs to process a large amount of CT image data. In this experiment, we use the proposed method to model a cochlear CT images dataset.

#### 4.2.1 Dataset and Implementation details

The dataset includes 1080 patients 3D images that collected from the radiology department of *** University Hospital. The original slices sequences are with a spacing size of $0.185mm, 0.185mm, 0.25mm$. We used a reference image to register all the images to the cochlea region (FOI) by using an automatic pyramidal blocking-matching (APBM) framework (Ourselin et al., 2000; Toussaint et al., 2007). The FOI volumes are sampled into isometric spacing size of $0.2mm$ with volume of $(60, 50, 50)$.

The proposed 3D Langevin-VAE consists with two 3D CNN. The encoder takes tensors with shape of $(N_b = 10 \cdot N_c = 60, N_w = 50, N_h = 50)$ and processes the tensors by three 3D convolutional layers with softplus non-linear projections. The strides of all the convolutional layers are set as 2. At the end, The tensors are flattened to fully connected layer for get the mean and variance parameters. The outputs parameters are then applied with the Langevin flow presented in Alogorithm 1 for adding the target information. The decoder network accepts the parameters and using the deconvolutional operation to model the marginal likelihood $p(x|z)$. The decoder follows inverse operations as the encoder did to upscale the feature maps to the original tensor shape.

#### 4.2.2 Result on real dataset

Tab. 2 shows two inference metrics that represent the inference performance on the medical image dataset. We see that the Langevin-VAE outperforms the VAE method on the dataset abstraction ability

Table 2: Quantitative evaluation of the Langevin-VAE in comparison with the VAE

| | VAE | Langevin-VAE |
|---|---|---|
| **Avg. ELBO** | -85293.33 +/- 1.538 | **-85135.24 +/- 4.82** |
| **Avg. NLL** | 83204.7+/- 10.92 | **83159.44 +/- 5.31** |

(as the ELBO and NLL are all better than VAE). Fig. 3 shows 20 samples of generated fake cochlea CT images. We see that the Langevin-VAE learns the variance of the cochlea shapes and diversity of the intensity changes.

## 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. In comaprison with conventional VAE and other methods, the proposed method achieves better performance on both toy and real world problems. Specially, 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. To the best of our knowledge, the proposed approach is the first Langevin flow based method as an generative model for dataset modeling.

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.

Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2018.10.009. URL https://www.sciencedirect.com/science/article/pii/S1077314218304272.

Anthony L Caterini, Arnaud Doucet, and Dino Sejdinovic. Hamiltonian variational auto-encoder. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/3202111cf90e7c816a472aaceb72b0df-Paper.pdf.

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.

Karol Gregor, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. Deep autoregressive networks. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1242–1250, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/gregor14.html.

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

Chin-Wei Huang, Kris Sankaran, Eeshan Dhekane, Alexandre Lacoste, and Aaron Courville. Hierarchical importance weighted autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2869–2878. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/huang19d.html.

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. *Journal of Machine Learning Research*, 22(42):1–41, 2021. URL http://jmlr.org/papers/v22/20-576.html.

Iain Murray and Russ R Salakhutdinov. Evaluating probabilities under high-dimensional latent variable models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou (eds.), *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2008/file/0d7de1aca9299fe63f3e0041f02638a3-Paper.pdf.

Sébastien Ourselin, A Roche, S Prima, and Nicholas Ayache. Block Matching : A General Framework to Improve Robustness of Rigid Registration of Medical Images. *Medical Image Computing and Computer-Assisted Intervention*, pp. 557–566, 2000.

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.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander (eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL http://proceedings.mlr.press/v33/ranganath14.html.

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.

Nicolas Toussaint, Jean-Christophe Souplet, and Pierre Fillard. MedINRIA: Medical Image Navigation and Research Tool by INRIA. In *Proc. of MICCAI'07 Workshop on Interaction in medical image analysis and visualization*, Brisbane, Australia, Australia, 2007. URL https://hal.inria.fr/inria-00616047.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

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

# A Appendix

## A.1 Over-damped form of the Generalized Langevin Diffusion

We consider a unit mass $m = 1$ evolving with a Brownian motion. The velocity part of the generalized Langevin type equation is:

$$\partial\Theta(t) = Kdt \quad \partial K(t) = \frac{\partial\Theta(t)^2}{\partial t^2} = \frac{\partial ln(p_\Theta(x))}{\partial\Theta}dt - \nu\Gamma K(t) + \delta\sigma(t) \tag{14}$$

In the case of an over-damped frictional force, the frictional force $\nu K$ overwhelms the inertial force $m \cdot \partial^2\theta/\partial t^2$, and thus $\frac{\frac{\partial\Theta(t)^2}{\partial t^2}}{\nu K(t)} \approx 0$. According to the generalized Langevin diffusion equation, we have :

$$\frac{\frac{\partial\Theta(t)^2}{\partial t^2}}{\nu K(t)} = \frac{\frac{\partial ln(p_\Theta(x))}{\partial\Theta}dt}{\nu K(t)} - \Gamma + \frac{\delta\sigma(t)}{\nu K(t)}$$

Therefore, we get :

$$\nu K(t)\Gamma \approx \frac{\partial ln(p_\Theta(x))}{\partial\Theta}dt + \delta\sigma(t)$$

which is the evolution given in Eq 4.

## A.2 Proof the integrator Eq. 5 is quasi-symplectic

**Proposition 1.** *Eq. 5 is asymptotic symplectic:* $\lim_{\nu\to 0} |\Psi_0(K_i, \Phi_i)| = exp(-\nu t)$

**Remark 1.** *Proposition [1] has propositional equivalences that the exterior power between two integration steps are equivalent as the Jacobian* $|\Psi_0(K_i, \Phi_i)|$ *is not dependent on the time step term* $t$*. Thus, to prove the proposition 1 is equivalent to proof that:* $dK_{i+1} \wedge d\Phi_{i+1} = dK_i \wedge d\Phi_i$*.*

*Proof:*

Let, $\nu \to 0$, the term $K_{II}$ of the composite integrator Eq. 5 goes to: $\lim_{\nu\to 0} K_{II}(t, \phi) = \phi$

Then,

$$\begin{aligned} dK_{i+1} &= dK_i + td(\frac{\partial \log(p(x, \Phi_i + \frac{t}{2}K_i))}{\partial \Phi_i}) \\ &= dK_i + d[\frac{t\partial \log(p(x, \Phi_i + \frac{t}{2}K_i))}{\partial \Phi_i}](d\Phi_i + \frac{t}{2}K_i) \\ d\Phi_{i+1} &= d\Phi_i + \frac{t}{2}dK_i + \frac{t}{2}d(K_i + \frac{t\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{\partial \Phi_i}) \\ &= d\Phi_i + tdK_i + d\frac{t^2\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{2\partial \Phi_i}(d\Phi_i + \frac{t}{2}K_i) \end{aligned} \tag{15}$$

Let $U' = \frac{\partial \log(p(x, \Phi_i + \frac{t}{2}K_i)}{2\partial \Phi_i}$, thus,

$$\begin{aligned} dK_{i+1} \wedge d\Phi_{i+1} = dK_i \wedge d\Phi_i + dK_i \wedge tdK_i + dK_i \wedge \frac{t^2}{2}dU'd\Phi_i + \\ dK_i \wedge \frac{t^3}{4}dU'dK_i + tdU'(d\Phi_i + \frac{t}{2}dK_i) \wedge d\Phi_i + tdU'(d\Phi_i + \\ \frac{t}{2}dK_i) \wedge tdK_i + tdU'(d\Phi_i + \frac{t}{2}dK_i) \wedge \frac{t^2}{2}dU'(d\Phi_i + \frac{t}{2}dK_i) \end{aligned} \tag{16}$$

According the property of exterior product, therefore:

$$dK_i \wedge tdK_i = tdU'd\Phi_i \wedge d\Phi_i = tdU'\frac{t}{2}dK_i \wedge tdK_i = 0 \tag{17}$$

Simplifying Eq. 16:

$$\begin{aligned} dK_{i+1} \wedge d\Phi_{i+1} &= dK_i \wedge d\Phi_i + t^2dU'(dK_i \wedge d\Phi_i) + t^2dU'(d\Phi_i \wedge dK_i) + \\ &\frac{t^4dU'^2}{4}(d\Phi_i \wedge dK_i) + \frac{t^4dU'^2}{4}(dK_i \wedge d\Phi_i) \\ &= dK_i \wedge d\Phi_i + t^2dU'(dK_i \wedge d\Phi_i) - t^2dU'(dK_i \wedge d\Phi_i) + \\ &\frac{t^4dU'^2}{4}(d\Phi_i \wedge dK_i) - \frac{t^4dU'^2}{4}(d\Phi_i \wedge dK_i) \\ &= dK_i \wedge d\Phi_i \end{aligned} \tag{18}$$

Q.E.D.

### A.3 Evidence lower bound of Langevin Flow

We consider the log-likelihood: $\log p(x)$ with latent variables $z$, based on Jensen's inequality:

$$\log p(x) \geq \int_\Omega \log \tilde{p}(x) q(\tilde{z}|x) d\tilde{z} \tag{19}$$

The data prior is given through the Langevin flow where $\mathcal{L}^I(\theta_0, k_0)$ are the $K$ steps Langevin flows with initialization states $(\theta_0, k_0)$:

$$\tilde{p} = \frac{\hat{p}(x, \mathcal{L}^I(\theta_0, k_0))}{q^0(\mathcal{L}^0(\theta_0, k_0))} \tag{20}$$

Therefore, we can get the Langevin flow lower bound:

$$\begin{aligned} \tilde{\mathbb{L}} :&\geq \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log q^0(\mathcal{L}^0(\theta_0, k_0)))d\tilde{z} \\ &= \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0) \prod_{k=1}^{I} |det \nabla \Psi_i^{-1} I(\theta_0, k_0))|))d\tilde{z} \\ &= \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0)) - \sum_{k=1}^{I} \log(|det \nabla \Psi_I^{-1}(\theta_0, k_0))|))d\tilde{z} \\ &= \int_\Omega q(\tilde{z}|x) \cdot (\log \hat{p}(x, \mathcal{L}^I(\theta_0, k_0)) - \log(q^0(\theta_0, k_0)) + \sum_{k=1}^{I} (\nu t))d\tilde{z} \end{aligned} \tag{21}$$

Table 3: Models hyper parameters and training/testing parameters setting.)

| | $\nu$ | $t_a, t_b$ | $T_0$ | $l_r$ | $adamax_\epsilon$ |
|---|---|---|---|---|---|
| Langevin VAE | 1e-2 | [1e-2, 5e-1] | 1.5 | 5e-4 | 1e-7 |
| Hamiltonian VAE | - | [1e-2, 5e-1] | 1.5 | 5e-4 | 1e-7 |

### A.4 PARAMETERS OF THE EXPERIMENT SETTING

Tab. 3 shows the parameters used for the experiment. Except for the parameter $\nu$ that is unique for the Langevin-VAE, all the other parameters are the same as the Hamiltonian-VAE.