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) as one of the well investigated generative model is very popular in nowadays neural learning research works. To leverage VAE in practical tasks which have high dimensions and massive dataset often facing the problem of low variance evidence lower bounds construction. Markov chain Monte Carlo (MCMC) is an effective approach to tight the evidence lower bound (ELBO) for approximating the posterior distribution. Hamiltonian Variational Autoencoder (HVAE) is one of those effective MCMC inspired approaches for constructing the low-variance ELBO which is also amenable for reparameterization trick. The solution significantly improves the performance of the posterior estimation, yet, a main drawback of HVAE is the leapfrog method need to access the posterior gradient twice which leads to bad inference efficiency and the GPU memory requirement is fair large. This flaw limited the application of Hamiltonian based inference framework for large scale networks inference. To tackle this problem, we propose a Quasi-symplectic Langevin Variational autoencoder (Langevin-VAE), which can be a significant improvement over resource usage efficiency. We qualitatively and quantitatively demonstrate the effectiveness of the Langevin-VAE compared to the state-of-art gradients informed inference framework.

1 INTRODUCTION

Variational autoencoder (VAE) as one of the mainstays of current generative neural models group have been applied in practical issues to generate target distributions dataset with advantages that quantitative assessment of model quality is possible and less finicky for training than generative adversarial networks (GANs). The key factor that influences the performance of VAE models is the quality of the likelihood approximation or the inference quality of the corresponding evidence lower bound (ELBO) when the latent variables exist. It is a common method which scarifies the flexibility of the model to estimate the parameter of latent distribution by specifying the variable distribution follow on a certain type of distribution (Wolf et al. (2016)). In the work of Salimans et al. (2015), the Hamiltonian variational inference is employed to approximate the variational lower bound of the log-likelihood. This Hamiltonian variational inference is a type of normalizing flows (NFs) (Rezende & Mohamed (2015)) that without necessary for presuming the posteriors follow one type of distribution. Salimans et al. (2015) demonstrated the possibility that integrates Hamiltonian Monte Carlo (HMC) with variational inference to improve the effectiveness of posterior distribution exploration. To eliminate the declination that leaves out the acceptance verification in HMC variational inference, Wolf et al. (2016) retrieved the acceptance step in Hamiltonian Monte Carlo algorithm which ensure the convergence of HMC to the target posterior distribution. Hamiltonian Variational Autoencoder (HVAE) (Caterini et al. (2018)) is a type of Hamiltonian flows based VAE model equipped with augmented dynamic phase space (momentum component ρ and position component θ) which can introduce the target information into the dynamic process. Briefly, HVAE employees a k steps Hamiltonian dynamic $\mathcal{H}_k$ transformation process to build an unbiased estimation of target distribution $p(x)$ through inferring distribution $q(z)$ by extending $\tilde{p}(x, \theta)$ as $\tilde{p}(x, \mathcal{H}_k(\theta_0, \rho_0))$ which leads: $\tilde{p}(x) := \frac{\hat{p}(x, \mathcal{H}_k(\theta_0, \rho_0))}{q(\mathcal{H}_k(\theta_0, \rho_0))}$, where: $\hat{p}(x, \theta_k, \rho_k) = \hat{p}(x, \mathcal{H}_k(\theta_0, \rho_0)) = \hat{p}(x, \theta_k)\mathcal{N}(\rho_k|0, I)$. By

*Corresponding author.

introducing the MCMC reverse kernel, the reparameterization trick is possible to be applied for ELBO gradients estimation to incorporate the model as HVAE. HVAE makes the inference process become directive by introducing explicit posterior information during the training period. However, the HVAE requires to access $\partial lnp(x,z)/\partial z$ by $2 \times k$ times which is quite expensive for computational complexity. This flaw becomes very serious when the number of steps k is large wherein complexity task a large k is necessary for sufficient inference. Another weakness of HVAE is the extended space of reverse transformation kernel introduce extra computational complexity and this will increase the memory requirement significantly in the leapfrog step and the inference time will increase doubly.

In this work, we propose a novel inference framework named quasi-symplectic Langevin variational auto-encoder (Langevin-VAE) that satisfy both reversibility of MCMC and phase quasi-volume invariance similar with Hamiltonian flow (Caterini et al. (2018)) but overcome its limitations of hardware overhead and computational time. The proposed method is a low-variance unbiased lower bound estimator in the case infinitesimal discretization step but with just once target Jacobian calculation and evade the Hessian accessing for computing the phase space volume variety between the inference steps. We show that the Langevin-VAE is a generalized stochastic inference framework since proposed Langevin-VAE is a symplectic approach when the viscosity coefficient $\nu = 0$. We reduce the computing burden of the flow based VAE inference by introducing the quasi-symplectic Langevin dynamic flow that abandon the leapfrog integrator for the flow inference. 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 of likelihood,

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_q[lnp(x,z) - lnq_\theta(z)] \tag{1}$$

Ranganath et al. proposed black box variational inference by deducing the noisy unbiased gradient of ELBO directly apply stochastic optimization to cracking the Eq: 1. Kingma & Welling (2014) proposed to use some normal distributions $\mathcal{N}$ to represent the distribution of latent variable z by a universal function ω, which makes reparameterization trick is possible:

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_q[lnp(x,\omega(\epsilon)) - lnq_\theta(\omega(\epsilon))], \quad \epsilon{\sim}\mathcal{N}(0,1) \tag{2}$$

To have sufficient representation power to approximate the complicate target posterior, the distribution of the latent parameters cannot be simple distributions. Nevertheless, only a few of distributions are amenable for reparameterization. Normalizing Flows (NFs) (Rezende & Mohamed (2015)) is one of the expressive approaches for distribution representation which is also adaptable for reparameterization trick (Papamakarios et al. (2019)). NFs is a general term for a class of methods that using a series of invertible transformation $T_K ... \circ ... T_0$ to transform naive distribution z_0 for complexity distribution z_k representation: $z_k = T_K ... \circ ... T_0(z_0)$, By applying a series of the transformation, the corresponding logarithm distribution density $q(z_k)$ of the transformed distribution is:

$$ln(p(z_k)) = ln(p(z_0)) - \Sigma_0^k ln|det\frac{\partial T_k}{\partial z_{k-1}}| \tag{3}$$

where the determinate Jacobian $|det\frac{\partial T_k}{\partial z_{k-1}}|$ ensures the global volume invariant property of every transformation steps asymptotically. This Jacobian item of invertible transformation T is the basis behind the necessary condition of reversibility for normalizing flows. As we mentioned in section 1 that the Hamiltonian dynamic is a type of NFs, thus the Eq: (3) is also hold for Hamiltonian flow. The Hamiltonian flow is a gradients trajectory which consisted by kinetic and potential pairs (Hamiltonian) follows the Liouville's theorem (symplectic) (Fassò & Sansonetto (2007)). This intrinsic attribution of field without sources leads to approximately unit Jacobian for leapfrog discretization with step size l of Hamiltonian normalizing flow: $\lim_{l\to 0}|det\frac{\partial T_k}{\partial z_k}|_l^{-1} = I$. This property makes the trivial Jacobian calculations for HVAE in each discretization steps (Caterini et al. (2018)). The Hamiltonian dynamics of HVAE in combine with reparameterization trick

was used to construct an unbiased esitimator of the lower bound gradients $\nabla_\theta \mathbb{L}$ that the approximation distribution: $q(\cdot)$ is constructed through the K steps 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 dynamic which need to compute the Jaconbian ∇U in two split sub-steps for momentum item (Girolami & Calderhead (2011)). The determinate Jacobian computational complexity is $\mathbb{O}(D^3)$ for D dimensional square matrix that is not adaptable to large scale flow inference of high dimensional problems. Hamiltonian flow clever use the Liouville's theorem to avoid the determinate computation. Yet, the Langevin flow is not inferring on Hamiltonian phase space destined that we cannot directly profit from the Liouville's theorem to simplify the calculation of Jacobian. To tackle this issue, we applied a quasi-symplectic Langevin flow that can easily infer the Jacobian through the explicit form.

2.2 Langevin Monte-Carlo and Normalizing Flow

In molecular physics, Langevin dynamics is a stochastic process that models the diffusion process 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 : $v = -\nabla U + \eta$.

As Langevin dynamics describes a stochastic evolution of particles towards the minima of potential field $U(x)$, it has recently attracted a lot of attention in the machine learning communityCosta et al. (2015); Stuart et al. (2004); Girolami & Calderhead (2011); Welling & Teh (2011); Mou et al. (2020) for the stochastic sampling of posterior distributions $p_\Theta(z|x)$ in Bayesian inference. Langevin Monte-Carlo methods (Girolami & Calderhead (2011)) rely on the construction of Markov chains with stochastic paths parameterized by Θ based on the discretization of the following *Langevin–Smoluchowski* SDE (Girolami & Calderhead (2011)) related to the overdamped Langevin dynamics :

$$\delta\Theta(t) = \frac{1}{2}\nabla_\Theta ln(p_\Theta(z|x))\delta t + \delta\sigma(t) \tag{4}$$

where $\sigma(t)$ is a Wiener process and t represents the time. The stochastic flow in Eq (4) 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 (see section 3.2 in Rezende & Mohamed (2015)). To the best of our knowledge, little work aims to explore 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 Jacobian determinants of those mappings involves the Hessian matrix of $ln(p_\Theta(x))$ as follows :

$$ln|det\frac{\partial T_k}{\partial z_{k-1}}|^{-1} \sim \nabla\nabla ln(p_{z_k}(x)) + \mathbb{O}(z) \tag{5}$$

Yet, the Hessian matrix appearing in Eq (5) may be 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, for Langevin flow, the forward transform modelled by Fokker-Plank equation and the backward transform can be given by Kolmogorov's backward equation which is discussed in the work of Kobyzev et al. (2020) and are not elaborated in detail here.

2.3 Quasi-symplectic Langevin and Corresponding Flow

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:

$$\begin{aligned} \partial\Phi(t) &= Kdt \\ \partial K(t) &= -\frac{-\partial ln(p_\Phi(x))}{\partial\Phi}dt - \nu K(t) + \delta\sigma(t) \end{aligned} \tag{6}$$

where $K(t)$ is the stochastic velocity field, and ν controls the amount of damping. We can see that the Langevin–Smoluchowski type SDE: (4) is nothing but the special case of high friction motion (Sandev T. (2019)) when Eq: (6) have an over-damped frictional force $\nu K \gg \partial^2\phi/\partial t^2$ which leads to the omission of $\partial^2\Phi/\partial t^2$ (please refer to the appendix section A.1 for details). To get a trivial Jacobian for constructing the flow, we need to have a symplectic attribution for the transformation kernel. To achieve this, we introduce the quasi-symplectic Langevin method for building the flow Milstein et al. (2002). The quasi-symplectic Langevin is different from the Euler–Maruyama method which is a divergent method for Langevin SDE discretization, the quasi-symplectic Langevin method ensures the tractable Jacobian during the diffusion process and keep approximate symplectic structure for damping force and external potential items.

In our case, we use a determinate mapping Ψ ($\sigma = 0$) to ensure the reversibility of the dynamic flow construction, that is a *second order strong quasi-symplectic* method (7). The transformation Ψ is defined as:

$$\begin{aligned} K_{1,i} &= K_{II}(\frac{\tau}{2}, K_i);\ \ \Phi_{1,i} = \Phi_i + \tau\frac{-\partial ln(p_\Phi(x))}{2\partial\Phi} \\ K_{2,i} &= K_{1,i} + \sqrt{\tau}\sigma\xi_i;\ \ K_{i+1} = K_{II}(\frac{\tau}{2}, K_{2,i}) \\ \Phi_{i+1} &= \Phi_{1,i} + \frac{\tau}{2}K_{2,i};\ \ K_{II}(t,p) = pe^{-\nu t};\ \ \xi_i \sim N(0, I) \end{aligned} \tag{7}$$

where initial conditions are $K = \kappa_0; \Phi = \phi_0$. The above quasi-symplectic form satisfies the following two attributions (7),

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

Theorem 2 *Quasi-symplectic method (7) has Θ, K independent Jacobian:*

$$J = \frac{\partial\Phi_{i+1}}{\partial\Phi_i}\frac{\partial K_{i+1}}{\partial K_i} - \frac{\partial\Phi_{i+1}}{\partial K_i}\frac{\partial K_{i+1}}{\partial\Phi_i} = exp(-\nu\tau) \tag{8}$$

The theorem (1) shows that the VAE constructed based on the quasi-symplectic Langevin dynamic is energetically equivalent to HVAE once the $\nu = 0$. The theorem (2) leads to the Jaconbian J uncorrelated with Θ and K which help to trivial Jacobian calculation. Proofing of constructed transition (7) satisfy the two theorems can be found in the work of Milstein (2003).

By introducing the quasi-symplectic transformation Ψ: (I) We can be avoiding Hessian computing to get the Jacobian for Langevin dynamic flow which is quite consuming. (II) The gradients just need to be computed once in each steps in stead of twice as in leap-frog discretization of Hamiltonian flow. This two attributions will help us to reduce the burden of computing resources and improve computing time efficiency. We then give the formal definition of the quasi-symplectic Langevin normalizing flow.

Definition 2.1 *An I steps discrete quasi-symplectic Langevin normalizing flow $\mathcal{L}^I$ is defined by a series of diffeomorphism, bijective and invertible mapping $\Psi : \sigma_\mathcal{A} \to \sigma_\mathcal{B}$ between two measureable spaces $(\mathcal{A}, \sigma_\mathcal{A}, \mu_\alpha)$ and $(\mathcal{B}, \sigma_\mathcal{B}, \mu_\beta)$:*

$$\mathcal{L}^I\mu_\alpha(\mathcal{S}_\mathcal{A}) : \Psi_{i+1} \circ \mu_\alpha(\mathcal{S}_\mathcal{A}) = \mu_\alpha(\Psi_i^{-1}(\mathcal{S}_\mathcal{B})), \forall\mathcal{S}_\mathcal{A} \in \sigma_\mathcal{A}, \mathcal{S}_\mathcal{B} \in \sigma_\mathcal{B}, i = \{0, ..., I\}. \tag{9}$$

where $\sigma_{(\cdot)}$ and $\mu_{(\cdot)}$ are the σ-algebra and probability measure for set $(\cdot)$ respectively, Ψ_i is the quasi-symplectic Langevin transform given by Eqs:(7).

We show an example with single step flow transform for probability measure in the appendix section: (A.2) to explain the above definition in detail.

The defined quasi-symplectic Langevin flow is a generalization of the Langevin dynamic flow that equipped with a quasi-symplectic structure for the parameters phase space. The quasi-symplectic Langevin normalizing flow is a flow with a deterministic kernel Ψ when the kernel stochastic factor $\sigma = 0$, and degenerate to a symplectic transition as $\nu = 0$. In the case of deterministic the flow will be able to have the probability density evaluate directly.

3 QUASI-SYMPLECTIC LANGEVIN VAE

3.1 LOWER BOUND ESTIMATION WITH LVAE

In this section, we describe the evidence lower bound construction based on the forward presented quasi-symplectic Langevin flow. First, we recall that the basis ideal of Bayesian inference: as the latent variable $z \in \Omega$ is intractable directly, the variational evidence lower bound $\tilde{\mathbb{L}}$ provides an alternative approach to infer the loglikelihood $\mathbb{L}$,

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

where Ω is the measure space of the latent variables and as $\tilde{p}(x)$ is the unbiased estimator for p(x) which means parameters set of p(x): $\theta_p \in \Theta_p$ has unbiased estimation $\theta_{\tilde{p}} \in \Theta_{\tilde{p}}$. Obviously, the variational distribution $q(\tilde{z}|x)$ determinate the tight degree of the lower bound. The inequality takes the equal sign when $q(\cdot) \equiv p(\cdot)$ which leads to a group of methods (variational inference introduced in section 2.1) that attempt to minimize the gap between $q(\cdot)$ and $p(\cdot)$ in order to construct a tighter lower bound for the loglikelihood $\mathbb{L}$ (Blei et al. (2017)). We can build our approximation distribution $q(\cdot)$ though a series of transformation which is the Langevin flow: $q^K(\mathcal{L}^K(\phi_0,\cdot)) = q^0(\mathcal{L}(\phi_0,\cdot))\prod_{k=1}^{K}|det\nabla\Psi^k(\mathcal{L}^0(\phi_0,\cdot))|^{-1}$, where Ψ^k is the Langevin dynamic deterministic transform ($\sigma = 0$ in Eq:7).

We then introduce the lower bound for the symplectic Langevin VAE (please refer to the appendix section A.3 for details),

$$\tilde{\mathbb{L}} := \int_{\Omega} q(\tilde{z}|x) \cdot (\ln \hat{p}(x, \mathcal{L}^K(\phi_0, k_0)) - \ln(q^0(\phi_0, k_0)) + \sum_{k=1}^{K}(\nu\tau))d\tilde{z} \tag{11}$$

where $q(\tilde{z}|x)$ is the posterior distribution (the model). Intuitively, Ψ^k can be any discretizator (Runge-Kutta methods, Euler-Maruyama method etc. Li et al. (2019)) which are suitable for Langevin dynamic. However, none of those methods ensure the symplectic attribution for simplify Jaconbian computing. Although for those cases, the integrator Ψ^k can be another smooth mapping which even not guarantee the Jacobian invariant. Then the Hessian matrix of the logarithm probability $p(x)$ are needed to get the inverse Jacobian determinate (Gu et al. (2019)). Yet, the Hessian matrix increases the computational complexity with $O(n^2)$ which is intolerable for high dimensional latent status problems. In Langevin flow the Jaconbian just need to compute once in each step which is better than Hamiltonian approach. With the proposed approach, the Jacobian determinate is trivial for computing through a *second order strong quasi-symplectic* structure construction which given the Jaconbian explicitly by therome (2) without necessary to access the Hessian. For quasi-symplectic Langevin the symplectic integrator Ψ^k is given through equations: (7).

It should be pointed out that both the proposed estimator and the estimator of HVAE are Markov chain based sampling of the likelihood $p(x)$. The prior works that introducing this group of sampling methods for likelihood estimator construction including the works of Salimans et al. (2015); Wu et al. (2019); Wolf et al. (2016).

3.2 QUASI-SYMPLECTIC LANGEVIN VAE

In this section, we employee the quasi-symplectic Langevin lower bound $\tilde{\mathbb{L}}$ for stochastic inference of a variational auto-encoder. Given a set of dataset $X : \{x^i \in X; i \in \mathbb{N}_+\}$, we want to inference a model that well describes the generative attributions of the dataset, the quasi-symplectic Langevin VAE learns a distribution $\hat{p}_\omega(x)$ parameterized by ω to have the generative distribution as consistent as the true data distribution $p(x)$. This can be inferred through maximizing the quasi-symplectic Langevin lower bound with respect to the model parameters ω:

$$\arg\max_{\omega\in\mathbb{R}^n} \tilde{\mathbb{L}}^* = \mathbb{E}_{z_0\sim\bar{q}_\omega(z|x)}(\ln \hat{p}_\omega(x, \mathcal{L}^K(\phi_0,\kappa_0)) - \ln(q^0_\omega(\phi_0,\kappa_0)) - \sum_{k=1}^{K}\ln(|det\nabla\Psi_k^{-1}(\phi_0,k_0))|)) \tag{12}$$

where $\bar{q}(\cdot)$ is the approximation distribution of $\bar{q}(z|x)$. To maximize the lower bound (12), we considering the gradient operation of the lower bound that can be extracted to the outside of the integral operation in conjunction with the reparameterization trick in Eq.2. To have a fair comparison in the evaluation between different methods we also apply Rao-Blackwellization as literature (Caterini et al. (2018)) did for reducing the variance of the ELBO of our quasi-symplectic Langevin VAE:

$$\begin{aligned}\arg\max_{\omega\in\mathbb{R}^n}\tilde{\mathbb{L}}^* =&\mathbb{E}_{\phi_0\sim\bar{q}_\omega(\phi|x),\kappa_0\sim\mathcal{N}_\zeta}(\ln\hat{p}_\omega(x,\mathcal{L}^K(\phi_0))-\ln(\hat{q}_\omega(\phi_0,\kappa_0))-\\ &\sum_{k=1}^{K}\ln|det\nabla\Psi_k^{-1}(\phi_0,\kappa_0))|-\frac{1}{2}\kappa_K^T\kappa_K)+\frac{\zeta}{2};\quad\forall\phi,\kappa\in\mathbb{R}^\zeta\end{aligned}\tag{13}$$

Algorithm 1: Quasi-symplectic Variational Inference

Inputs: Data X, Inference steps K
Output: Model parameters ω
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.
    $\phi_0 \sim \bar{q}_\omega(\phi|x)$; // Sampling latent states from approximation distributions.
    $\kappa_0 \sim \mathcal{N}_\zeta$; // Sampling ζ dimensions viscosity states from standard normal distributions.
    for $i=1;\ i<K;\ i++$ **do**
      // Quasi-symplectic Langevin Normalizing Flow
      $\kappa_{1,i}\leftarrow K_{II}(\frac{\tau}{2},\kappa_i);\ \phi_{1,i}\leftarrow\phi_i+\tau\frac{-\partial ln(\partial p_{\phi_i}(x))}{2\partial\phi_i}$;
      $\xi_i\sim N(0,I);\ \kappa_{2,i}\leftarrow\kappa_{1,i}+\sqrt{\tau}\sigma\xi_i$;
      $\kappa_{i+1}\leftarrow K_{II}(\frac{\tau}{2},\kappa_{2,i});\ \phi_{i+1}\leftarrow\phi_{1,i}+\frac{\tau}{2}\kappa_{2,i}$;
      $|det\nabla\Phi^k(\mathcal{H}_k(\phi_0,\rho_0))|^{-1}\leftarrow exp(-\nu\tau)^{-1}$; // Theorem 2
    end
    $p^*_\omega\leftarrow\hat{p}_{\omega,K}(\phi_K)\cdot p_{\mathcal{N}(0,I_\zeta)}(\kappa_K)$;
    $q^*_\omega\leftarrow\hat{q}_{\omega,K}(\phi_0)\cdot p_{\mathcal{N}(0,I_\zeta)}(\kappa_0)\cdot|det\nabla\Phi^k(\mathcal{H}_k(\phi_0,\rho_0))|^{-1}$;
    $\tilde{\mathbb{L}}^*_j\leftarrow\ln(p^*_\omega)-\ln(q^*_\omega)$; ; // 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 avg ELBO over parameters subset
end

4 Experiment and Result

We will first examine the performance of quasi-symplectic Langevin VAE on the MNIST dataset (LeCun et al. (2010)) from different metrics: batch average inference efficiency (AVGT), GPU memory usage, evidence lower bound (ELBO) tightness, negative log-likelihood(NLL), Inception Score (IS) and Fréchet Inception(FID). Caterini et. al. have reported that the Hamiltonian based stochastic variational inference outperform than Planar Normalizing Flow, mean-field based Variational Bayes in terms of model parameters inference error and validated the performance of HVAE(Caterini et al. (2018)). Here, we compare the proposed method with HVAE on MNIST dataset implemented with *TensorFlow 2.0* and *TensorFlow Probability* framework to evaluate the proposed approach in both qualitative and quantitative metrics.

4.1 QUASI-SYMPLECTIC LANGEVIN VAE ON BENCHMARK

For a set of training data $X : \{x^i \in X; i \in \mathbb{N}_+\}$ we assume the data are manipulated by a set of latent variables $\Phi : \{\phi^i \in \Phi; i \in \mathbb{N}_+\}$ that are following Normal distributions $\phi_i \sim \mathcal{N}(0, I)$. The pixel x_i are determined by a joint Bernoulli distributions: $\phi_i \sim \prod_{j=1}^{J} \mathcal{B}(\hat{x_j}|Decoder(\phi_i))$, which are parameterized by the output of the decoder network. The encoder network modeling the mapping between data x and parameters of latent parameters distribution$\{\mu_i, \sigma_i\}$ which is a typical framework of VAE.Kingma & Welling (2014) The inference flow of the quasi-symplectic Langevin VAE is shown in figure 1 with a pseudo Bayesian probability graph. The decoder and encoder neural

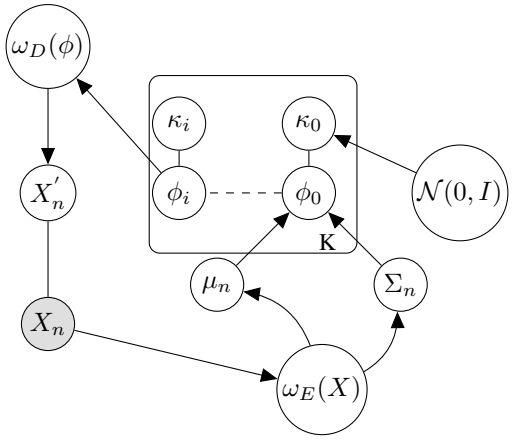

Figure 1: Graphical model of the Quai-symplectic Langevin inference framework. The dataset for inference is observable as $X_n \equiv X_n'$. The Gaussian distributions parameters μ, Σ that modelling the distribution of latent parameters ϕ_0 are determined by encoder neural network with parameters ω_E. The frictional term κ_0 are sampled from the standard normal distribution to be paired with μ_0. A K steps of Quai-symplectic Langevin sampling is then performed on the energy pairs with final status:$\{\phi_{i=K}, \kappa_{i=K}\}$. Decoder neural network with parameters ω_D then generate the data X_n' by inferencing the final status $\{\phi_{i=K}, \kappa_{i=K}\}$ of the quasi-symplectic Langevin sampling.

network structures are similar to the HVAE Caterini et al. (2018) and MCMCVAE (Salimans et al. (2015)), both have three layers 2D convolutional neural network for encoder and decoder respectively. The encoder network accepts a batch of data with shape of $(nb \times 28 \times 28)$ where nb is the batch size: $nb = 1000$ of training data. The dimension of latent variables is set as $\zeta = 64$. The training convergence criterion is the ELBO on validation dataset do not improve after 100 steps or achieve 2000 steps. Both of the models are set with the same conditions to be trained. We train the neural network by an Adamax optimizer with the learning rate $lr = 0.0005$. The experiments of network training and time consumption evaluation were implemented on one NVIDIA GeForce GTX 1080 Ti GPU. For memory usage evaluation the experiment was performed on an NVIDIA Quadro M2200 GPU.

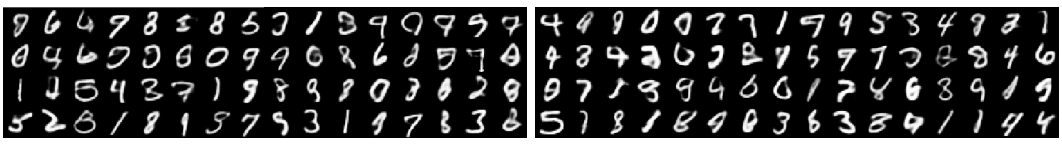

Figure 2: Quantitative result of Langevin VAE in comparison with HVAE. Left sub-figures are generated samples of HVAE. Right are samples of Langevin-VAE. Both of the two methods are set the flow steps K=5.

Table 1: Quantitative evaluation of the Langevin-VAE in comparison with the HVAE.

	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

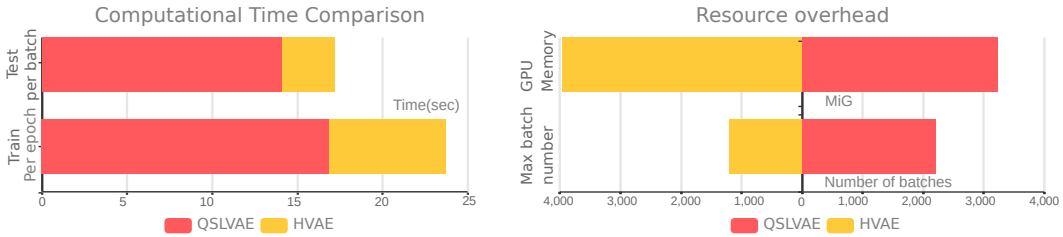

Figure 3: Comparison of the time and memory cost of the QSLVAE and HVAE. Left bar figure shows the time consumption of the test (per batch) and training (per epoch) stages. Right figure shows the memory overhead of 1200 batches feed data (above bar) and maximum batch number can be feed to the 4GiB GPU (below bar) for the two approaches.

4.2 Result

Both qualitative and quantitative results are studied. We show the generated samples of Langevin-VAE and HVAE in Fig:2. We see qualitatively that the samples diversity and image quality of the generated images are guaranteed for both the two models. Quantitatively, Table:1 shows the performance of the NLL score and the ELBO for Langevin-VAE and HVAE where two different flow inference steps are experimentally compared. To exam the hardware resource efficiency of the proposed framework, we follow a similar evaluation as the experiment of Anil et al. Anil et al. (2019) We study the effectiveness of the Langevin-VAE from two different views. First, the computational time of the proposed approach is outperformed than the HVAE in both the training and testing stages. As shown in Fig:3 (left), the time consumption for Langevin-VAE is around 17 seconds per epoch against 24 seconds for HVAE. This phenomenon is the empirical evidence of the Langevin-VAE can save more time since just once gradients computation step is necessary. The time gap for testing between the two approaches is lighter in comparison with the training time since no backpropagation is necessary. The ease of computational burden for the gradients $\nabla_\theta ln(p_\theta(x))$ can also reducing the memory usage of the computational instruments. In the right sub-figure of Fig: 3 we see that the Langevin-VAE can reduce almost 23% of GPU memory overhead vis-à-vis the HVAE which allow us to double the number of batches to train all at once for Langevin-VAE.

5 Conclusion

In this paper, we propose a new flow-based Bayesian inference framework by introducing the quasi-symplectic Langevin dynamic for building a stochastic ELBO. The advantage of using Hamiltonian inference is the deterministic dynamic flow helps avoid the construction of reverse kernel. However, the flaw of these kinds of gradients leaded probability transform methods is the gradient computing need extra time and the GPU memory for computing the Jacobian $\nabla log(P)$ of the dynamic energy can be fair large. The proposed approach can reduce significantly the GPU memory requirement for training this deterministic dynamic flow-based inference method. The proposed methods towards both memory-efficient and computing time-efficient which will be effective on very large scale models inference task.

By introducing the quasi-symplectic Langevin dynamic, we overcome the weakness that the Langevin normalizing flow (Gu et al. (2019)) which needs to provide the Hessian matrix $\nabla\nabla log(P)$ to get the determinate of Jacobin. We show that the proposed method can halve the memory requirement for gradients caching and reducing the computational efficiency vis-à-vis the Hamiltonian variational

inference (Caterini et al. (2018)) without inference accuracy loss. There are still potential improvements for quasi-symplectic Langevin inference by investigating the manifold structure of the target densities (see: Girolami & Calderhead (2011); Barp et al. (2017); Livingstone & Girolami (2014)) for inference efficiency further improving which would be our future work.

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

Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 9749–9758. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9168-memory-efficient-adaptive-optimization.pdf.

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.

Sueli I.R. Costa, Sandra A. Santos, and João E. Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59 – 69, 2015. ISSN 0166-218X. doi: https://doi.org/10.1016/j.dam.2014.10.004. URL http://www.sciencedirect.com/science/article/pii/S0166218X14004211. Distance Geometry and Applications.

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.

Xuechen Li, Yi Wu, Lester Mackey, and Murat A Erdogdu. Stochastic runge-kutta accelerates langevin monte carlo and beyond. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 7748–7760. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8990-stochastic-runge-kutta-accelerates-langevin-monte-carlo-and-beyond.pdf.

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.

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.

Anqi Wu, Sebastian Nowozin, Edward Meeds, Richard E. Turner, Jose Miguel Hernandez-Lobato, and Alexander L. Gaunt. Deterministic variational inference for robust bayesian neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1l08oAct7.

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 νK 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 6.

A.2 Example for single step quasi-symplectic Langevin flow

We consider a probability measure $p(x)$ of random variable set $x \in X$. Let Ψ as a quasi-symplectic Langevin mapping. Then a single step Langevin flow transform the original random variable x to a new random variable $y = \Psi(x), y \in Y$. According to the definition 2.1, the new probability measure $q(y)$ of random variable y is given by:

$$q(y) = \mathcal{L}^0 p(x) : \Psi_0 \circ p(x) = p(\Psi_0^{-1}(y)) \tag{15}$$

By Eq.(3), we conclude:

$$q(y) = p(x) \cdot |det\frac{\partial \Psi_0}{\partial x}|^{-1} \tag{16}$$

A.3 Evidence lower bound of Langevin Flow

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

$$\ln p(x) \geq \int_\Omega \ln \tilde{p}(x) q(\tilde{z}|x) dz \tag{17}$$

The data prior is given through the Langevin flow where $\mathcal{L}^K(\theta_0, k_0)$ are the K steps Langevin flows with initialization states (θ_0, k_0):

$$\tilde{p} = \frac{\hat{p}(x, \mathcal{L}^K(\theta_0, k_0))}{q^0(\mathcal{L}^0(\theta_0, k_0))} \tag{18}$$

Therefore, we can get the Langevin flow lower bound:

$$\begin{aligned}
\tilde{\mathbb{L}} :&\geq \int_\Omega q(\tilde{z}|x) \cdot (\ln \hat{p}(x, \mathcal{L}^K(\theta_0, k_0)) - \ln q^0(\mathcal{L}^0(\theta_0, k_0)))dz \\
&= \int_\Omega q(\tilde{z}|x) \cdot (\ln \hat{p}(x, \mathcal{L}^K(\theta_0, k_0)) - \ln(q^0(\theta_0, k_0) \prod_{k=1}^{K} |det \nabla \Psi_k^{-1} k(\theta_0, k_0))|))dz \\
&= \int_\Omega q(\tilde{z}|x) \cdot (\ln \hat{p}(x, \mathcal{L}^K(\theta_0, k_0)) - \ln(q^0(\theta_0, k_0)) - \sum_{k=1}^{K} \ln(|det \nabla \Psi_k^{-1}(\theta_0, k_0))|))dz \\
&= \int_\Omega q(\tilde{z}|x) \cdot (\ln \hat{p}(x, \mathcal{L}^K(\theta_0, k_0)) - \ln(q^0(\theta_0, k_0)) + \sum_{k=1}^{K} (\nu\tau))dz
\end{aligned} \tag{19}$$