APPENDIX:VARIATIONAL LEARNING OF GAUSSIAN PROCESS LATENT VARIABLE MODELS THROUGH STOCHASTIC GRADIENT ANNEALED IMPORTANCE SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Gaussian Process Latent Variable Models (GPLVMs) have become increasingly popular for unsupervised tasks such as dimensionality reduction and missing data recovery due to their flexibility and non-linear nature. An importance-weighted version Salimbeni et al. (2019) of the Bayesian GPLVMs has been proposed to obtain a tighter variational bound. However, this version of the approach is primarily limited to analyzing simple data structures, as the generation of an effective proposal distribution can become quite challenging in high-dimensional spaces or with complex data sets. In this work, we propose an Annealed Importance Sampling (AIS) approach to address these issues. By transforming the posterior into a sequence of intermediate distributions using annealing, we combine the strengths of Sequential Monte Carlo samplers and VI to explore a wider range of posterior distributions and gradually approach the target distribution. We further propose an efficient algorithm by reparameterizing all variables in the evidence lower bound (ELBO). Experimental results on both toy and image datasets demonstrate that our method outperforms state-of-the-art methods in terms of tighter variational bounds, higher log-likelihoods, and more robust convergence.

1 INTRODUCTION

Gaussian processes (GPs) Rasmussen (2003) have become a popular method for function estimation due to their non-parametric nature, flexibility, and ability to incorporate prior knowledge of the function. Gaussian Process Latent Variable Models (GPLVMs), introduced by Lawrence & Hyvärinen (2005), have paved the way for GPs to be utilized for unsupervised learning tasks such as dimensionality reduction and structure discovery for high-dimensional data. It provides a probabilistic mapping from an unobserved latent space $\mathbf{H}$ to data-space $\mathbf{X}$.

The work by Titsias & Lawrence (2010) proposed a Bayesian version of GPLVMs and introduced a variational inference (VI) framework for training GPLVMs using sparse representations to reduce model complexity. This method utilizes an approximate surrogate estimator $g(\mathbf{X}, \mathbf{H})$ to replace the true probability term $p(\mathbf{X})$, i.e. $\mathbb{E}_{q(\mathbf{H})}\left[g(\mathbf{X}, \mathbf{H})\right] = p(\mathbf{X})$. VI typically defines an evidence lower bound (ELBO) as the loss function for the model in place of $\log p(\mathbf{X})$. To describe the accuracy of this lower bound, we discuss a Taylor expansion of $\log p(\mathbf{X})$,

$$\mathbb{E}_{q(\mathbf{H})}\left[\log g(\mathbf{X}, \mathbf{H})\right] \approx \log p(\mathbf{X}) - \frac{1}{2}\mathrm{var}_{q(\mathbf{H})}\left[\frac{g(\mathbf{X}, \mathbf{H})}{p(\mathbf{X})}\right] \tag{1}$$

The formula has been discussed in numerous works, including Thin et al. (2020); Maddison et al. (2017); Domke & Sheldon (2018). Therefore, as the variance of the estimator decreases, the ELBO becomes tighter. Based on this formula and the basic principles of the central limit theorem, importance-weighted (IW) VI Domke & Sheldon (2018) seeks to reduce the variance of the estimator by repeatedly sampling from the proposal distribution $q(\mathbf{H})$, i.e., $g\left(\mathbf{X}, \mathbf{H}\right) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{p(\mathbf{X}, \mathbf{H}_k)}{q(\mathbf{H}_k)}\right]$, where $\mathbf{H}_k \sim q\left(\mathbf{H}_k\right)$. An importance-weighted version Salimbeni et al. (2019) of the Bayesian GPLVMs based on this has been proposed to obtain a tighter variational bound. While this method can obtain a

tighter lower bound than the classical VI, it is a common problem that the relative variance of this importance-sampling based estimator tends to increase with the dimension of the latent variable. Moreover, the generation of an effective proposal distribution can become quite challenging in high-dimensional spaces or with complex data sets.

The problem of standard importance sampling techniques is that it can be challenging to construct a proposal distribution $q(\mathbf{H})$ that performs well in high-dimensional spaces. To address these limitations, we propose a novel approach for variational learning of GPLVMs by leveraging Stochastic Gradient Annealed Importance Sampling (SG-AIS). AIS is derived from early work by Jarzynski (1997) and has been further developed by Crooks (1998); Neal (2001). This approach remains one of the 'gold standard' techniques to estimate the evidence unbiasedly because it explores a wider range of posterior distributions and gradually approach the target distribution Del Moral et al. (2006); Salimans et al. (2015); Grosse et al. (2013; 2015).

Specifically, our proposed approach leverages an annealing procedure to transform the posterior distribution into a sequence of intermediate distributions, which can be approximated by using a Langevin stochastic flow. This dynamic is a time-inhomogeneous unadjusted Langevin dynamic that is easy to sample and optimize. We also propose an efficient algorithm designed by reparameterizing all variables in the ELBO. Furthermore, we propose a stochastic variant of our algorithm that utilizes gradients estimated from a subset of the dataset, which improves the speed and scalability of the algorithm . Our experiments on both toy and image datasets show that our approach outperforms state-of-the-art methods in GPLVMs, demonstrating lower variational bounds, higher log-likelihoods, and more robust convergence.

Overall, our contributions are as follows:

- We propose a novel approach for variational learning of GPLVMs by leveraging Stochastic Gradient Annealed Importance Sampling (SG-AIS), which addresses the limitations of standard importance sampling techniques and allows for the estimation of the evidence unbiasedly, resulting in a tighter lower bound and a better variational approximation in complex data and high-dimensional space.
- We propose an efficient algorithm designed by reparameterizing all variables to further improve the estimation of the variational lower bounds. We also leverage stochastic optimization to maximize optimization efficiency.
- Our experiments on both toy and image datasets demonstrate that our approach outperforms state-of-the-art methods in GPLVMs, showing lower variational bounds, higher log-likelihoods, and more robust convergence.

2 Background

2.1 GPLVM Variational Inference

In GPLVMs, we have a training set comprising of N D-dimensional real valued observations $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^N \in \mathbb{R}^{N\times D}$. These data are associated with N Q-dimensional latent variables, $\mathbf{H} = \{\boldsymbol{h}_n\}_{n=1}^N \in \mathbb{R}^{N\times Q}$ where $Q < D$ provides dimensionality reduction Titsias & Lawrence (2010). The forward mapping $\mathbf{H} \to \mathbf{X}$ is described by multi-output GPs independently defined across dimensions D. The work by Titsias & Lawrence (2010) proposed a Bayesian version of GPLVMs using sparse representations to reduce model complexity. The formula is described as,

$$
\begin{aligned}
p(\mathbf{H}) &= \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{h}_n; \mathbf{0}, I_Q\right) \\
p(\mathbf{F} \mid \mathbf{U}, \mathbf{H}) &= \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{f}_d; \boldsymbol{\mu_d}, Q_{nn}\right) \\
p(\mathbf{X} \mid \mathbf{F}, \mathbf{H}) &= \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(x_{n,d}; \boldsymbol{f}_d\left(\boldsymbol{h}_n\right), \sigma^2\right)
\end{aligned}
\tag{2}
$$

where $Q_{nn} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$, $\boldsymbol{\mu_d} = K_{nm}K_{mm}^{-1}\boldsymbol{u}_d$, $\mathbf{F} = \{\boldsymbol{f}_d\}_{d=1}^D$, $\mathbf{U} = \{\boldsymbol{u}_d\}_{d=1}^D$ is the inducing variable Titsias (2009), $\boldsymbol{x}_d$ is the d-th column of $\mathbf{X}$, and m is the number of inducing

points. K_{nn} is the covariance matrix corresponding to a user-chosen positive-definite kernel function $k_\theta(\boldsymbol{h}, \boldsymbol{h}')$ evaluated on latent points $\{\boldsymbol{h}_n\}_{n=1}^N$ and parameterized by hyperparameters θ. The kernel hyperparameters are shared across all dimensions D. It is assumed that the prior over $\mathbf{U}$ and $\mathbf{H}$ factorizes into $p(\boldsymbol{u}_d)$ and $p(\boldsymbol{h}_n)$, where $p(\boldsymbol{u}_d) = \mathcal{N}(\mathbf{0}, K_{mm})$ and $p(\boldsymbol{h}_n) = \mathcal{N}(\mathbf{0}, I_Q)$. Since $\boldsymbol{h}_n \in \mathbb{R}^Q$ is unobservable, we need to do joint inference over $\boldsymbol{f}(\cdot)$ and $\boldsymbol{h}$. Under the typical mean-field assumption of a factorized approximate posterior $q(\boldsymbol{f}_d)q(\boldsymbol{h}_n)$. We denote ψ as all variational parameters and γ as all GP hyperparameters. Thus, we arrive at the classical Mean-Field (MF) ELBO:

$$\begin{aligned} \text{MF-ELBO}(\gamma, \psi) = &\sum_{n=1}^{N}\sum_{d=1}^{D} \int q(\boldsymbol{f}_d) q\left(\boldsymbol{h}_n\right) \log p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_n\right) \mathrm{d}\boldsymbol{h}_n \, \mathrm{d}\boldsymbol{f}_d \\ &- \sum_{n=1}^{N} \mathrm{KL}\left(q\left(\boldsymbol{h}_n\right) \| p\left(\boldsymbol{h}_n\right)\right) - \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \| p(\boldsymbol{u}_d)\right), \end{aligned} \tag{3}$$

where we use the typical approximation to integrate out the inducing variable,

$$q\left(\boldsymbol{f}_d\right) = \int p\left(\boldsymbol{f}_d | \boldsymbol{u}_d\right) q\left(\boldsymbol{u}_d\right) d\boldsymbol{u}_d. \tag{4}$$

2.2 Importance-weighted Variational Inference

A main contribution of Salimbeni et al. (2019) is to propose a variational scheme for (Latent variable) LV-GP models based on importance-weighted VI Domke & Sheldon (2018) via amortizing the optimization of the local variational parameters. IWVI provides a way of lower-bounding the log marginal likelihood more tightly and with less estimation variance by Jensen's inequality at the expense of increased computational complexity. The IW-ELBO is obtained by replacing the expectation likelihood term (first term) in Vanilla VI with a sample average of K terms:

$$\text{IW-ELBO}(\gamma, \psi) = \sum_{n=1}^{N}\sum_{d=1}^{D} B_{n,d} - \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \| p(\boldsymbol{u}_d)\right), \tag{5}$$

where $B_{n,d} = \underset{\boldsymbol{f}_d, \boldsymbol{h}_n}{\mathbb{E}} \log \frac{1}{K} \sum_k p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,k}\right) \frac{p(\boldsymbol{h}_{n,k})}{q(\boldsymbol{h}_{n,k})}$. Although the IW objective outperforms classical VI in terms of accuracy, its effectiveness is contingent on the variability of the importance weights: $p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,k}\right) \frac{p(\boldsymbol{h}_{n,k})}{q(\boldsymbol{h}_{n,k})}$. When these weights vary widely, the estimate will effectively rely on only the few points with the largest weights. To ensure the effectiveness of importance sampling, the proposal distribution defined by $q\left(\boldsymbol{h}_{n,k}\right)$ must therefore be a fairly good approximation to $p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,k}\right) p\left(\boldsymbol{h}_{n,k}\right)$, so that the importance weights do not vary wildly. Related theoretical proofs can be seen in Domke & Sheldon (2018); Maddison et al. (2017).

When $\boldsymbol{h}_{n,k}$ is high-dimensional, or the likelihood $p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,k}\right)$ is multi-modal, finding a good importance sampling distribution can be very difficult, limiting the applicability of the method. Unfortunately, original research by Salimbeni et al. (2019) only discusses the case when $\boldsymbol{h}_n$ is a one-dimensional latent variable, and they acknowledge that reliable inference for more complex cases is not yet fully understood or documented. To circumvent this issue, we provide an alternative for GPLVMs using Annealed Importance Sampling (AIS) Crooks (1998); Neal (2001); Wu et al. (2016), which defines state-of-the-art estimators of the evidence and designs efficient proposal importance distributions. Specially, we propose a novel ELBO, relying on unadjusted Langevin dynamics, which is a simple implementation that combines the strengths of Sequential Monte Carlo samplers and variational inference as detailed in Section 3. ,

3 Variational AIS Scheme in GPLVMs

3.1 Variational Inference via AIS

Annealed Importance Sampling (AIS)Neal (2001); Del Moral et al. (2006); Salimans et al. (2015); Zhang et al. (2021) is a technique for obtaining an unbiased estimate of the evidence $p(\mathbf{X})$. To achieve this, AIS uses a sequence of K bridging densities $\{q_k(\mathbf{H})\}_{k=1}^K$ that connect a simple base distribution $q_0(\mathbf{H})$ to the posterior distribution $p(\mathbf{H}|\mathbf{X})$. By gradually interpolating between these

distributions, AIS allows for an efficient computation of the evidence. This method is particularly useful when the posterior is difficult to sample from directly, as it allows us to estimate the evidence without evaluating the full posterior distribution directly. We can express this as follows:

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{H}) d\mathbf{H} = \mathbb{E}_{q_{\text{fwd}}(\mathbf{H}_{0:K})} \left[\frac{q_{\text{bwd}}(\mathbf{H}_{0:K})}{q_{\text{fwd}}(\mathbf{H}_{0:K})} \right] \tag{6}$$

where the variational distribution q_{fwd} and the target distribution q_{bwd} can be written as:

$$\begin{aligned} q_{\text{fwd}}(\mathbf{H}_{0:K}) &= q_0(\mathbf{H}_0)\, \mathcal{T}_1(\mathbf{H}_1 \mid \mathbf{H}_0) \cdots \mathcal{T}_K(\mathbf{H}_K \mid \mathbf{H}_{K-1}) \\ q_{\text{bwd}}(\mathbf{H}_{0:K}) &= p(\mathbf{X}, \mathbf{H}_K)\, \tilde{\mathcal{T}}_K(\mathbf{H}_{K-1} \mid \mathbf{H}_K) \cdots \tilde{\mathcal{T}}_1(\mathbf{H}_0 \mid \mathbf{H}_1). \end{aligned} \tag{7}$$

Here, we assume $\mathcal{T}_k$ is a forward MCMC kernel that leaves $q_k(\mathbf{H})$ invariant, which ensures that $\{\mathcal{T}_k\}_{k=1}^K$ are valid transition probabilities, *i.e.*, $\int q_k(\mathbf{H}_{k-1}) \mathcal{T}_k(\mathbf{H}_k \mid \mathbf{H}_{k-1}) \, \mathrm{d}\mathbf{H}_{k-1} = q_k(\mathbf{H}_k)$. And $\tilde{\mathcal{T}}_k$ is the "backward" Markov kernel moving each sample $\mathbf{H}_k$ into a sample $\mathbf{H}_{k-1}$ starting from a virtual sample $\mathbf{H}_K$. q_{fwd} represents the chain of states generated by AIS, and q_{bwd} is a fictitious reverse chain which begins with a sample from $p(\mathbf{X}, \mathbf{H})$ and applies the transitions in reverse order. In practice, the bridging densities have to be chosen carefully for a low variance estimate of the evidence. A typically method is to use geometric averages of the initial and target distributions to construct the sequence, i.e., $q_k(\mathbf{H}) \propto q_0(\mathbf{H})^{1-\beta_k} p(\mathbf{X}, \mathbf{H})^{\beta_k}$ for $0 = \beta_0 < \beta_1 < \cdots < \beta_K = 1$. AIS has been proven theoretically to be consistent as $K \to \infty$ Neal (2001) and achieves accurate estimate of $\log p(\mathbf{X})$ empirically with the asymptotic bias decreasing at a $1/K$ rate Grosse et al. (2013; 2015).

With this, we can derive the AIS bound,

$$\begin{aligned} \log p(\mathbf{X}) &\geq \mathbb{E}_{q_{\text{fwd}}(\mathbf{H}_{0:K})} \left[\log \frac{q_{\text{bwd}}(\mathbf{H}_{0:K})}{q_{\text{fwd}}(\mathbf{H}_{0:K})} \right] \\ &= \mathbb{E}_{q_{\text{fwd}}(\mathbf{H}_{0:K})} \left[\log p(\mathbf{X}, \mathbf{H}_K) - \log q_0(\mathbf{H}_0) - \sum_{k=1}^{K} \log \frac{\mathcal{T}_k(\mathbf{H}_k \mid \mathbf{H}_{k-1})}{\tilde{\mathcal{T}}_k(\mathbf{H}_{k-1} \mid \mathbf{H}_k)} \right]. \end{aligned} \tag{8}$$

This objective can be obtained by applying Jensen's inequality. For the GPLVM, we can naturally derive its AIS lower bound:

$$\begin{aligned} \mathcal{L}_{\text{AIS}}(\psi, \gamma) = &\sum_{n=1}^{N} \sum_{d=1}^{D} \mathbb{E}_{q_{\text{fwd}}(\boldsymbol{h}_{0:K}) q(\boldsymbol{f}_d)} \left[\log p(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,K}) \right] \\ &+ \sum_{n=1}^{N} \mathbb{E}_{q_{\text{fwd}}(\boldsymbol{h}_{0:K})} \left[\log p(\boldsymbol{h}_{n,K}) - \log q_0(\boldsymbol{h}_{n,0}) \right] \\ &- \sum_{k=1}^{K} \mathbb{E}_{q_{\text{fwd}}(\mathbf{H}_{0:K})} \log \frac{\mathcal{T}_k(\mathbf{H}_k \mid \mathbf{H}_{k-1})}{\tilde{\mathcal{T}}_k(\mathbf{H}_{k-1} \mid \mathbf{H}_k)} - \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \parallel p(\boldsymbol{u}_d) \right) \end{aligned} \tag{9}$$

where ψ and γ indicate the sets of all variational parameters and all GP hyperparameters, respectively. Our purpose is to evaluate this bound. First we note that the last KL term is tractable if we assume the variational posteriors of $\boldsymbol{u_d}$ are mean-field Gaussian distributions. So we concentrate on the terms in the expectation that we can evaluate relying on a Monte Carlo estimate. It is obvious that $\log p(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,K})$ is available in closed form as the conditional likelihood is Gaussian Titsias (2009). Therefore, the first three term can be computed by the popular "reparameterization trick" Rezende et al. (2014); Kingma & Welling (2013) to obtain an unbiased estimate of the expectation over $q_{\text{fwd}}(\mathbf{H}_{0:K})$ and $q(\boldsymbol{f_d})$ (detailed in Section 3.3). Afterwards, to evaluate expectation over q_{fwd}, we construct an MCMC transition operator $\mathcal{T}_k$ which leaves q_k invariant via a time-inhomogeneous unadjusted (overdamped) Langevin algorithm (ULA) as used in Welling & Teh (2011); Heng et al. (2020); Wu et al. (2020); Marceau-Caron & Ollivier (2017) and jointly optimize ψ and γ by stochastic gradient descent.

3.2 Time-inhomogeneous Unadjusted Langevin Diffusion

$\mathcal{T}_k$ can be constructed using a Markov kernel with an invariant density such as Metropolis-Hastings (MH) or Hamiltonian Monte Carlo (HMC), which enables q_{fwd} to converge to the posterior distribution of $\mathbf{H}$. For the sake of simplicity, we consider the transition density $\mathcal{T}_k$ associated to this

discretization,

$$\mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right)=\mathcal{N}\left(\mathbf{H}_k ; \mathbf{H}_{k-1}+\eta \nabla \log q_k\left(\mathbf{H}_{k-1}\right), 2 \eta I\right) \tag{10}$$

where $\eta>0$ is the step size and q_k is bridging densities defined in Section 3.1. Since we have $q_k(\mathbf{H}) \propto q_0(\mathbf{H})^{1-\beta_k} p(\mathbf{X}, \mathbf{H})^{\beta_k}$ in Section 3.1, the annealed potential energy is derived as:

$$\nabla \log q_k(\cdot)=\beta_k \nabla \log p(\mathbf{X}, \cdot)+\left(1-\beta_k\right) \nabla \log q_0(\cdot). \tag{11}$$

According to conditional probability formula $\log p(\mathbf{X}, \cdot)=\log p(\mathbf{X} \mid \cdot)+\log p(\cdot)$, the model log likelihood simplifies to:

$$\nabla \log p(\mathbf{X} \mid \cdot)=-\frac{1}{2} \sum_{d=1}^{D} \nabla\left(\log \det \left(Q_{nn}+\sigma^2 I\right)+\left(\boldsymbol{x}_d-\boldsymbol{\mu}_{\boldsymbol{d}}\right)^T\left(Q_{nn}+\sigma^2 I\right)^{-1}\left(\boldsymbol{x}_d-\boldsymbol{\mu}_{\boldsymbol{d}}\right)\right). \tag{12}$$

Since Eq. (12) is analytical, the gradient can be computed through automatic differentiation Baydin et al. (2018). The dynamical system propagates from a base variational distribution q_0 to a final distribution q_K which approximates the posterior density. Let $\eta:=T / K$, then the proposal q_{fwd} converges to the path measure of the following Langevin diffusion $\left(\mathbf{h}_t\right)_{t \in[0, T]}$ defined by the stochastic differential equation (SDE),

$$\mathbf{d} \mathbf{H}_t=\nabla \log q_t(\mathbf{H}) \mathrm{d} t+\sqrt{2}\, \mathrm{d} \mathbf{B}_t, \quad \mathbf{H}_0 \sim q_0 \tag{13}$$

where $\left(\mathbf{B}_t\right)_{t \in[0, T]}$ is standard multivariate Brownian motion and q_t corresponds to q_k in discrete-time for $t=t_k=k \eta$. For long times, the solution of the Fokker-Planck equations Risken (1996) tends to the stationary distribution $q_\infty(\mathbf{H}) \propto \exp (p(\mathbf{X}, \mathbf{H}))$. Additional quantitative results measuring the law of $\mathbf{h}_T$ for such annealed diffusions have been showed in Andrieu et al. (2016); Tang & Zhou (2021); Fournier & Tardif (2021). For ease of sampling, we define the corresponding Euler-Maruyama discretization as,

$$\mathbf{H}_k=\mathbf{H}_{k-1}+\eta \nabla \log q_k\left(\mathbf{H}_{k-1}\right)+\sqrt{2 \eta} \boldsymbol{\epsilon}_{k-1}, \tag{14}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, I)$, as done in Heng et al. (2020); Wu et al. (2020); Nilmeier et al. (2011). Since such process is reversible w.r.t. q_k, based on Nilmeier et al. (2011), the reversal $\tilde{\mathcal{T}}_k$ is typically realized by,

$$\mathbf{H}_{k-1}=\mathbf{H}_k+\eta \nabla \log q_k\left(\mathbf{H}_k\right)+\sqrt{2 \eta} \tilde{\boldsymbol{\epsilon}}_{k-1}, \tag{15}$$

where $\tilde{\boldsymbol{\epsilon}}_{k-1}=-\sqrt{\frac{\eta}{2}}\left[\nabla \log q_k\left(\mathbf{H}_{k-1}\right)+\nabla \log q_k\left(\mathbf{H}_k\right)\right]-\boldsymbol{\epsilon}_{k-1}$. Based on Eq. (10), the term related to $\mathcal{T}_k$ in Eq. (9) can be written explicitly as:

$$\sum_{k=1}^{K} R_{k-1}=\sum_{k=1}^{K} \log \frac{\mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right)}{\tilde{\mathcal{T}}_k\left(\mathbf{H}_{k-1} \mid \mathbf{H}_k\right)}=\sum_{k=1}^{K} \frac{1}{2}\left(\left\|\tilde{\boldsymbol{\epsilon}}_{k-1}\right\|^2-\left\|\boldsymbol{\epsilon}_{k-1}\right\|^2\right). \tag{16}$$

3.3 Reparameterization Trick and Stochastic Gradient Descent

For ease of sampling, we consider a reparameterization version of Eq. (9) based on the Langevin mappings associated with q_k given by

$$T_k(\mathbf{H}_{k-1})=\mathbf{H}_{k-1}+\eta \nabla \log q_k\left(\mathbf{H}_{k-1}\right)+\sqrt{2 \eta} \boldsymbol{\epsilon}_{k-1}. \tag{17}$$

Based on the identity $\mathbf{H}_k=T_k(\mathbf{H}_{k-1})$, we have a representation of $\mathbf{H}_k$ by a stochastic flow,

$$\mathbf{H}_k=T_k\left(\mathbf{H}_{k-1}\right)=T_k \circ T_{k-1} \circ \cdots T_1(\mathbf{H_0}) \tag{18}$$

Moreover, for LVGP models, we also have a reparameterization version Salimbeni & Deisenroth (2017) of the posteriors of $\mathbf{H_0}$ and $\boldsymbol{f}_d$ in Eq. (9), that is,

$$\begin{aligned} \boldsymbol{h}_{n,0} &= \boldsymbol{a}_n+L_n \boldsymbol{\epsilon} \\ \boldsymbol{f}_d &= K_{nm} K_{mm}^{-1} \boldsymbol{m}_d+\sqrt{K_{nn}-K_{nm} K_{mm}^{-1}\left(K_{mm}-\boldsymbol{S}_d{}^T \boldsymbol{S}_d\right) K_{mm}^{-1} K_{mn}} \boldsymbol{\epsilon}_{f_d} \end{aligned} \tag{19}$$

Algorithm 1: Stochastic Unadjusted Langevin Diffusion(ULA) AIS algorithm for GPLVMs

Input: training data $\mathbf{X}$, mini-batch size B, sample number K, annealing schedule $\{\beta_k\}$, stepsizes η
Initialize all DGP hyperparameters γ, all variational parameters ψ
repeat
  Sample mini-batch indices $J \subset \{1, \ldots, N\}$ with $|J| = B$
  Draw $\boldsymbol{\epsilon}$ from standard Gaussian distribution.
  Set $\mathbf{H}_0 = \mathbf{a_n} + L_n \boldsymbol{\epsilon}$
  Set $\mathcal{L} = -\log q_0(\mathbf{H}_0)$
  for $k = 1$ **to** K **do**
    Draw $\boldsymbol{\epsilon}_k$ from standard Gaussian distribution.
    Set $\nabla \log q_k(\cdot) = \beta_k(\frac{N}{B} \log p(\mathbf{X}_J|\cdot) + \log p(\cdot)) + (1-\beta_k) \nabla \log q_0(\cdot)$
    Set $\mathbf{H}_k = \mathbf{H}_{k-1} + \eta \nabla \log q_k(\mathbf{H}_{k-1}) + \sqrt{2\eta}\boldsymbol{\epsilon}_{k-1}$
    Set $\tilde{\boldsymbol{\epsilon}}_{k-1} = \sqrt{\frac{\eta}{2}}[\nabla \log q_k(\mathbf{H}_{k-1}) + \nabla \log q_k(\mathbf{H}_k)] - \boldsymbol{\epsilon}_{k-1}$
    Set $R_{k-1} = \frac{1}{2}\left(\|\tilde{\boldsymbol{\epsilon}}_{k-1}\|^2 - \|\boldsymbol{\epsilon}_{k-1}\|^2\right)$
    Set $\mathcal{L} = \mathcal{L} - R_{k-1}$
  end for
  Sample mini-batch indices $I \subset \{1, \ldots, N\}$ with $|I| = B$
  Draw $\epsilon_{\boldsymbol{f}_d}$ from standard Gaussian distribution for $d = 1, 2, ..., D$.
  Set $\mathcal{L} = \mathcal{L} + \log p(\mathbf{H}_K) + \frac{N}{B} \log p(\mathbf{X}_I \mid \boldsymbol{\epsilon}_{f_d}, \boldsymbol{\epsilon}_{0:K-1}, \boldsymbol{\epsilon}) - \sum_{d=1}^{D} \mathrm{KL}(q(\boldsymbol{u}_d) \parallel p(\boldsymbol{u}_d))$
  Do gradient desent on $\mathcal{L}(\psi, \gamma)$
until ψ, γ converge

where vectors $\boldsymbol{a}_n \in \mathbb{R}^Q$, $\boldsymbol{m}_d \in \mathbb{R}^N$ and upper triangular matrixs L_n, $\boldsymbol{S}_d$ are the variational parameters, $\boldsymbol{\epsilon} \in \mathbb{R}^Q, \boldsymbol{\epsilon}_{f_d} \in \mathbb{R}^N$ are standard Gaussian distribution. After this reparameterization, a change of variable shows that AIS bound in Eq. (9) can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\mathrm{AIS}}(\psi, \gamma) = &\sum_{n=1}^{N}\sum_{d=1}^{D} \mathbb{E}_{p(\boldsymbol{\epsilon}_{f_d})p(\boldsymbol{\epsilon}_{0:K-1})p(\boldsymbol{\epsilon})}\left[\log p\left(x_{n,d} \mid \boldsymbol{\epsilon}_{f_d}, \boldsymbol{\epsilon}_{0:K-1}, \boldsymbol{\epsilon}\right)\right] \\ &+ \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{\epsilon}_{0:K-1})p(\boldsymbol{\epsilon})}\left[\log p\left(\boldsymbol{h}_{n,K}\right) - \log q_0\left(\boldsymbol{h}_{n,0}\right)\right] \\ &- \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{\epsilon}_{0:K-1})p(\boldsymbol{\epsilon})} R_{k-1} - \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \parallel p(\boldsymbol{u}_d)\right),\end{aligned} \tag{20}$$

where R_{k-1} is defined in Eq. (16) and $\mathbf{h}_{n,k}$ is reparameterized as $\mathbf{h}_{n,k} = T_k \circ T_{k-1} \circ \cdots T_1(\mathbf{h}_{n,0}) = \bigcirc_{i=1}^{k} T_i(\mathbf{a}_n + L_n\boldsymbol{\epsilon})$. In order to accelerate training and sampling in our inference scheme, we propose a scalable variational bounds that are tractable in the large data regime based on stochastic variational inference Hoffman et al. (2013); Salimbeni & Deisenroth (2017); Kingma & Welling (2013); Hoffman & Blei (2015); Naesseth et al. (2020) and stochastic gradient descent Welling & Teh (2011); Chen et al. (2014); Zou et al. (2019); Teh et al. (2016); Sato & Nakagawa (2014); Alexos et al. (2022). Instead of computing the gradient of the full log likelihood, we suggest to use a stochastic variant to subsampling datasets into a mini-batch $\mathcal{D}_J$ with $|\mathbf{X}_J| = B$, where $J \subset \{1, 2, .., N\}$ is the indice of any mini-batch. In the meantime, we replace the $p(\mathbf{X}, \mathbf{H}_K)$ term in Eq. (7) with another estimator computed using an independent mini-batch of indices $I \subset \{1, 2, .., N\}$ with $|\mathbf{X}_I| = B$. We finally derive a stochastic variant of the Stochastic Unadjusted Langevin Diffusion AIS algorithm for the LVGP models, as describe in Algorithm 1.

4 Experiments

4.1 Methods and Practical Guidelines

In the following section, we present two sets of experiments. In the first set of experiments, our aim is to demonstrate the quality of our model in unsupervised learning tasks such as data dimensionality reduction and clustering. This will allow us to evaluate the ability of our model to preserve the original information in the data. In the second set of experiments, we evaluate the expressiveness and efficiency of our model on the task of image data recovery.

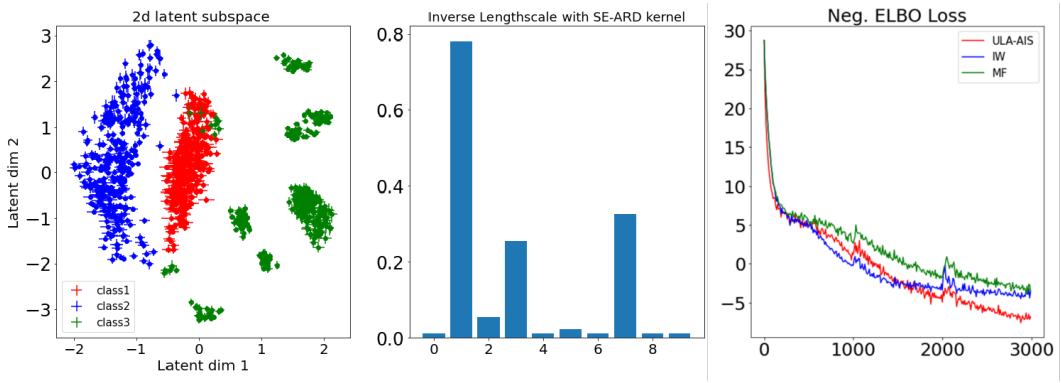

Figure 1: We lowered the data dimensionality using our proposed method in the multi-phase oilflow dataset and visualized a two-dimensional slice of the latent space that corresponds to the most dominant latent dimensions. The inverse lengthscales learnt with SE-ARD kernel for each dimension are depicted in the middle plot, and the negative ELBO learning curves are shown in the right plot. We set the same learning rate and compared the learning curves of two state-of-the-art models, MF and Importance Weighted VI within 3000 iterations.

We compare three different approaches: (a) Classical Sparse VI based on mean-field (MF) approximation Titsias & Lawrence (2010); (b) Importance-weighted (IW) VI Salimbeni et al. (2019); (c) ULA-AIS as given by the algorithm presented in this paper. We also provide guidelines on how to tune the step sizes and annealing schedules in Algorithm 1 to optimize performance. We conducted all our experiments on a Tesla A100 GPU. More details can be seen in Appendix C.

4.2 Dimensionality Reduction

Table 1: Comparison of MF, IW, and AIS under different number of iterations for two toy datasets

Dataset	Data Dim	Method	Iterations	Negative ELBO	MSE	Negative Expected Log Likelihood
Oilflow	(1000,12)	MF	1000	3.44 (0.25)	6.83 (0.27)	-1.42 (0.27)
			2000	-1.67 (0.17)	3.59 (0.13)	-8.38 (0.12)
			3000	-3.07 (0.12)	2.79 (0.11)	-11.24 (0.10)
		IW	1000	**0.01** (0.25)	**4.52** (0.28)	**-6.26** (0.26)
			2000	-3.19 (0.15)	2.77 (0.16)	-9.46 (0.15)
			3000	-4.13 (0.14)	2.60 (0.15)	-12.20 (0.12)
		AIS (ours)	1000	0.78 (0.24)	4.99 (0.23)	-4.01 (0.26)
			2000	**-5.04** (0.15)	**2.65** (0.15)	**-10.33** (0.16)
			3000	**-6.82** (0.12)	**2.16** (0.12)	**-13.06** (0.11)
Wine Quality	(1599,11)	MF	1000	32.69(0.13)	63.98(0.12)	31.71(0.15)
			2000	13.46(0.03)	48.95(0.05)	6.51(0.06)
			3000	11.59(0.03)	45.81(0.04)	4.07(0.05)
		IW	1000	**22.65**(0.07)	**50.77**(0.06)	**19.94**(0.09)
			2000	11.47(0.02)	40.86(0.03)	3.72(0.04)
			3000	10.73(0.03)	35.23(0.04)	2.71(0.03)
		AIS (ours)	1000	29.63(0.07)	57.49(0.05)	27.67(0.06)
			2000	**10.43**(0.03)	**34.60**(0.03)	**3.58**(0.04)
			3000	**8.86**(0.04)	**32.23**(0.04)	**2.47**(0.03)

The multi-phase Oilflow data Bishop & James (1993) consists of 1000, 12D data points belonging to three classes which correspond to the different phases of oil flow in a pipeline. We reduced the data dimensions to 10 while attempting to preserve as much information as possible. We report the reconstruction error and MSE with ± 2 standard errors over three optimization runs. Since the training is unsupervised, the inherent ground-truth labels were not a part of training. The 2D projections of the latent space for oilflow data clearly shows that our model is able to discover the class structure.

To highlight the strength of our model, we set the same learning rate and other experimental hyperparameters and compare the learning curves of two state-of-the-art models. The results are shown in Fig. 1. We also tested our model performance on another toy dataset, Wine Quality Cortez et al. (2009), where we used the white variant of the Portuguese "Vinho Verde" wine. From table 1, we

Figure 2: In the Brendan faces reconstruction task with 75% missing pixels, the top row represents the ground truth data and the bottom row showcases the reconstructions from the 20-dimensional latent distribution.

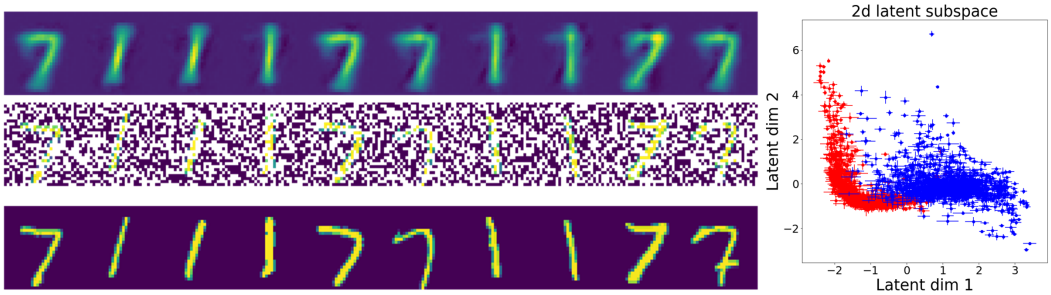

Figure 3: For the MNIST reconstruction task with 75% missing pixels, we chose digits 1 and 7. The bottom row represents the ground truth data and the top row showcases the reconstructions from the 5-dimensional latent distribution.The left side shows the reconstruction task, while the right side displays the 2-dimensional latent space corresponding to the smallest lengthscales.

observe that after sufficient training, our proposed method yields lower reconstruction loss and MSE than IWVI and MF methods.

4.3 Make Predictions in Unseen Data

Table 2: Comparison of MF, IW, and AIS under different number of iterations for two image datasets

Dataset	Data Dim	Method	Iterations	Negative ELBO	MSE	Negative Expected Log Likelihood
Frey Faces	(1965,560)	MF	1000	48274 (443)	468 (9)	46027 (356)
			2000	6346 (20)	95 (1)	4771 (17)
			3000	3782 (15)	69 (0.2)	2822 (3)
		IW	1000	42396 (426)	394 (8)	39936 (312)
			2000	5643 (15)	76 (1)	4292 (13)
			3000	3596 (14)	63 (0.5)	2535 (4)
		AIS (ours)	1000	**12444** (451)	**121** (9)	**10543** (322)
			2000	**5031** (16)	**66** (1)	**3130** (15)
			3000	**3249** (12)	**57**(0.3)	**2226** (3)
MNIST	(2163,784)	MF	2000	-432.32(0.33)	0.27(0.004)	-552.87(0.28)
		IW	2000	-443.64(0.37)	**0.25**(0.003)	-567.13(0.31)
		AIS (ours)	2000	**-453.18**(0.27)	**0.25**(0.002)	**-569.93**(0.26)

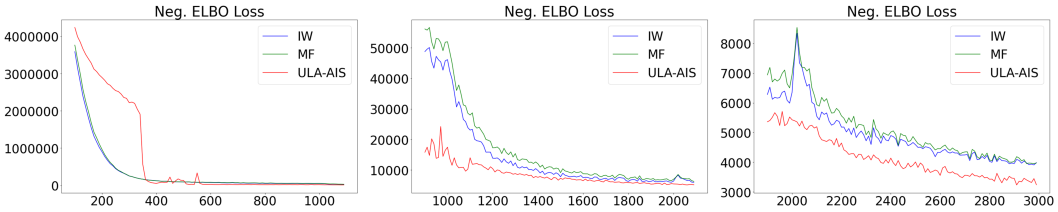

Figure 4: The negative ELBO convergence curves of the three methods on the Frey Faces dataset. It is noted that as the number of iterations increase, the y-axis scale gradually increases from left to right.

We conducted a reconstruction experiment on the MNIST and Frey Faces Data, focusing on how models capture uncertainty when training with missing data in structured inputs. For MNIST, we selected digits 1 and 7 with a latent variable dimensionality of 5. Each image has 784 pixels yielding a 784d data space[1]. For Frey Faces Data, we used the entire dataset with a latent variable dimensionality of 20. The image data set Roweis & Saul (2000) contains 1965 images of a face taken from sequential frames of a short video. Each image is of size 20×28 yielding a 560d data space. In both cases, we chose 5% of the training set as missing data samples and removed 75% of their pixels, seeking to recover their original appearance. Fig. 2 and Fig. 3 summarize the samples generated from the learned latent distribution. This reconstruction experiment is similar to the related work by Titsias & Lawrence (2010) and Gal et al. (2014). More details can be seen in Appendix D.

To demonstrate the effectiveness of our method in producing more accurate likelihoods and tighter variational bounds on image datasets, we present in Table 2 the negative ELBO, negative log-likelihood, and mean squared error (MSE) for reconstructed images on the Frey Faces and MNIST datasets, comparing with state-of-the-art methods. Our results show that our method achieves lower variational bounds and converges to higher likelihoods, indicating superior performance in high-dimensional and multi-modal image data. This suggests that adding Langevin transitions appears to improve the convergence of the traditional VI methods.

We also present in Fig. 4 a comparison of the negative ELBO convergence curves for Frey Faces datasets between our method and two other state-of-the-art methods. To better illustrate our lower convergence values, we gradually increase the y-axis scale from left to right. An interesting observation is that, compared to the IW and MF methods, our proposed method sometimes exhibits sudden drops in the loss curve, as shown in the leftmost plot of Fig. 4. This can be attributed to the fact that, by adding Langevin transitions, the algorithm's variational distribution gradually moves from the current distribution towards the true posterior distribution, resulting in sudden drops in the loss function when reaching the target distribution. Thus, such phenomena can be regarded as a common feature of annealed importance sampling and it becomes even more obvious in high-dimensional datasets.

5 Conclusion

In this paper, we introduce a novel method for GPLVM through Stochastic Gradient Annealed Importance Sampling. Our approach leverages annealing to transform the posterior distribution into a sequence of tractable intermediate distributions, and utilizes unadjusted Langevin dynamics to estimate the Evidence Lower Bound (ELBO). We observe convincing evidence of the superiority of our method, particularly in high-dimensional or complex structured datasets, including lower variational bounds and more robust convergence. Furthermore, we also observe certain features in the loss curve of our method, such as steep drops, which further support our claims.

Overall, our results show that the proposed method achieves superior performance in both accuracy and robustness, indicating its potential as an effective tool for the variational learning of latent-variable GP Models.

[1]Since the MNIST dataset converges within 2000 iterations, we report the performance of several methods at convergence

REFERENCES

Antonios Alexos, Alex J Boyd, and Stephan Mandt. Structured stochastic gradient mcmc. In International Conference on Machine Learning, pp. 414–434. PMLR, 2022.

Christophe Andrieu, James Ridgway, and Nick Whiteley. Sampling normalizing constants in high dimensions using inhomogeneous diffusions. arXiv preprint arXiv:1612.07583, 2016.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. Journal of Marchine Learning Research, 18:1–43, 2018.

Christopher M Bishop and Gwilym D James. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 327(2-3):580–593, 1993.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In International conference on machine learning, pp. 1683–1691. PMLR, 2014.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. Decision support systems, 47(4): 547–553, 2009.

Gavin E Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. Journal of Statistical Physics, 90(5):1481–1487, 1998.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):411–436, 2006.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. Advances in neural information processing systems, 31, 2018.

Nicolas Fournier and Camille Tardif. On the simulated annealing in rd. Journal of Functional Analysis, 281(5):109086, 2021.

Yarin Gal, Mark Van Der Wilk, and Carl Edward Rasmussen. Distributed variational inference in sparse gaussian process regression and latent variable models. Advances in neural information processing systems, 27, 2014.

Roger B Grosse, Chris J Maddison, and Russ R Salakhutdinov. Annealing between distributions by averaging moments. Advances in Neural Information Processing Systems, 26, 2013.

Roger B Grosse, Zoubin Ghahramani, and Ryan P Adams. Sandwiching the marginal likelihood using bidirectional monte carlo. arXiv preprint arXiv:1511.02543, 2015.

Jeremy Heng, Adrian N Bishop, George Deligiannidis, and Arnaud Doucet. Controlled sequential monte carlo. The Annals of Statistics, 48(5):2904–2929, 2020.

Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In Artificial Intelligence and Statistics, pp. 361–369, 2015.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. Journal of Machine Learning Research, 2013.

Christopher Jarzynski. Nonequilibrium equality for free energy differences. Physical Review Letters, 78(14):2690, 1997.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Neil Lawrence and Aapo Hyvärinen. Probabilistic non-linear principal component analysis with gaussian process latent variable models. Journal of machine learning research, 6(11), 2005.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. Advances in Neural Information Processing Systems, 30, 2017.

Gaétan Marceau-Caron and Yann Ollivier. Natural langevin dynamics for neural networks. In International Conference on Geometric Science of Information, pp. 451–459. Springer, 2017.

Christian Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: Variational inference with kl (p|| q). Advances in Neural Information Processing Systems, 33:15499–15510, 2020.

Radford M Neal. Annealed importance sampling. Statistics and computing, 11(2):125–139, 2001.

Jerome P Nilmeier, Gavin E Crooks, David DL Minh, and John D Chodera. Nonequilibrium candidate monte carlo is an efficient tool for equilibrium simulation. Proceedings of the National Academy of Sciences, 108(45):E1009–E1018, 2011.

Carl Edward Rasmussen. Gaussian processes in machine learning. In Summer school on machine learning, pp. 63–71. Springer, 2003.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In International conference on machine learning, pp. 1278–1286. PMLR, 2014.

Hannes Risken. Fokker-planck equation. In The Fokker-Planck Equation, pp. 63–95. Springer, 1996.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. science, 290(5500):2323–2326, 2000.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In International conference on machine learning, pp. 1218–1226. PMLR, 2015.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. Advances in neural information processing systems, 30, 2017.

Hugh Salimbeni, Vincent Dutordoir, James Hensman, and Marc Deisenroth. Deep gaussian processes with importance-weighted variational inference. In International Conference on Machine Learning, pp. 5589–5598. PMLR, 2019.

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In International Conference on Machine Learning, pp. 982–990. PMLR, 2014.

Wenpin Tang and Xun Yu Zhou. Simulated annealing from continuum to discretization: a convergence analysis via the eyring–kramers law. arXiv preprint arXiv:2102.02339, 2021.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. Journal of Machine Learning Research, 17, 2016.

Achille Thin, Nikita Kotelevskii, Jean-Stanislas Denain, Leo Grinsztajn, Alain Durmus, Maxim Panov, and Eric Moulines. Metflow: A new efficient method for bridging the gap between markov chain monte carlo and variational inference. arXiv preprint arXiv:2002.12253, 2020.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In Artificial intelligence and statistics, pp. 567–574. PMLR, 2009.

Michalis Titsias and Neil D Lawrence. Bayesian gaussian process latent variable model. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp. 844–851. JMLR Workshop and Conference Proceedings, 2010.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 681–688, 2011.

Hao Wu, Jonas Köhler, and Frank Noé. Stochastic normalizing flows. Advances in Neural Information Processing Systems, 33:5933–5944, 2020.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. arXiv preprint arXiv:1611.04273, 2016.

Guodong Zhang, Kyle Hsu, Jianing Li, Chelsea Finn, and Roger B Grosse. Differentiable annealed importance sampling and the perils of gradient noise. Advances in Neural Information Processing Systems, 34:19398–19410, 2021.

Difan Zou, Pan Xu, and Quanquan Gu. Stochastic gradient hamiltonian monte carlo methods with recursive variance reduction. Advances in Neural Information Processing Systems, 32, 2019.