Variational Learning of Gaussian Process Latent Variable Models through Stochastic Gradient Annealed Importance Sampling

Jian Xu[1] **Shian Du**[2] **Junmei Yang**[1] **Qianli Ma**[1] **Delu Zeng**[*1] **John Paisley**[3]

[1]South China University of Technology, Guangzhou, China
[2]Tsinghua University, Shenzhen, China
[3]Columbia University, NY, USA

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 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 VAIS-GPLVM, a variational Annealed Importance Sampling method that leverages time-inhomogeneous unadjusted Langevin dynamics to construct the variational posterior. 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 [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 and 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 and 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 and 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.

[*]Corresponding author: dlzeng@scut.edu.cn

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, as shown in Rainforth et al. [2018], Rudner et al. [2021].

To address these limitations, we propose VAIS-GPLVM, a variational Annealed Importance Sampling method that leverages time-inhomogeneous unadjusted Langevin dynamics to construct the variational posterior. Our method builds on the foundations of AIS, originally derived from nonequilibrium statistical mechanics Jarzynski [1997], and later extended in Crooks [1998], Neal [2001]. AIS remains a gold-standard technique for unbiased evidence estimation, as it explores a broader range of posterior distributions and gradually approaches 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 VAIS-GPLVM, a variational Annealed Importance Sampling method that uses time-inhomogeneous unadjusted Langevin dynamics to construct the variational posterior. This approach mitigates the issue of weight collapse in high-dimensional GPLVMs, yielding a tighter lower bound and improved variational approximation for complex, high-dimensional data.

- 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 and 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 and Lawrence [2010] proposed a Bayesian version of GPLVMs using sparse representations to reduce model complexity. We typically define the conditional distribution as $p(\boldsymbol{f}_d \mid \boldsymbol{u}_d, \mathbf{H}) = \mathcal{N}(\boldsymbol{f}_d; \boldsymbol{\mu}_d, Q_{nn})$, where $\boldsymbol{\mu}_d = K_{nm}K_{mm}^{-1}\boldsymbol{u}_d$, $Q_{nn} = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$, $\boldsymbol{u}_d$ is the inducing variable Titsias [2009]. Here, K_{nn} is the covariance matrix evaluated over latent inputs $\{\boldsymbol{h}_n\}_{n=1}^N$ using a user-defined positive-definite kernel function $k_\theta(\boldsymbol{h}, \boldsymbol{h}')$, parameterized by a shared set of kernel hyperparameters θ across all output dimensions D. The data likelihood is modeled as a Gaussian distribution, i.e.,

$$p(\mathbf{X} \mid \mathbf{F}, \mathbf{H}) = \prod_{n=1}^{N}\prod_{d=1}^{D} \mathcal{N}\left(x_{n,d}; \boldsymbol{f}_d\left(\boldsymbol{h}_n\right), \sigma^2\right) \tag{2}$$

where $\mathbf{F} = \{\boldsymbol{f}_d\}_{d=1}^D$, $\boldsymbol{x}_d$ is the d-th column of $\mathbf{X}$, and m is the number of inducing points. It is assumed that the prior is defined as $p(\boldsymbol{u}_d) = \mathcal{N}(\mathbf{0}, K_{mm})$ 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:

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

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

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

In Equation (4), $p\left(\boldsymbol{f}_d|\boldsymbol{u}_d\right)$ is a simplification of the traditional Sparse Gaussian Process (Sparse GP) approach. In the Sparse GP model, we typically assume $p\left(\boldsymbol{f}_d|\boldsymbol{u}_d, \boldsymbol{h}_n\right)$ as the conditional probability distribution of the latent variable $\boldsymbol{h}_n$, and we integrate over $\boldsymbol{h}_n$. Proofs can be seen in the Appendix.

2.2 IMPORTANCE-WEIGHTED VARIATIONAL INFERENCE

A main contribution of Salimbeni et al. [2019] is to propose a variational scheme for LV-GP models based on importance-weighted VI Domke and 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 in Vanilla VI with a sample average of K terms:

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

where $B_{n,d} = \underset{\boldsymbol{f}_d, \boldsymbol{h}_n}{\mathbb{E}} \log \frac{1}{K} \sum_k p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,k}\right) \frac{p(\boldsymbol{h}_{n,k})}{q(\boldsymbol{h}_{n,k})}$. Proofs can be seen in the Appendix.

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, as shown in Rainforth et al. [2018]. 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 widely. Related theoretical proofs can be seen in Domke and Sheldon [2018], Maddison et al. [2017], Rainforth et al. [2018].

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 variational 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] 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:

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

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

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

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.,

$$\int q_k(\mathbf{H}_{k-1}) \mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right) \mathrm{d}\mathbf{H}_{k-1} = q_k\left(\mathbf{H}_k\right). \tag{8}$$

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,

$$\begin{aligned}\log p(\mathbf{X}) &\geq \mathbb{E}_{q_{\mathrm{fwd}}(\mathbf{H}_{0:K})}\left[\log \frac{q_{\mathrm{bwd}}\left(\mathbf{H}_{0:K}\right)}{q_{\mathrm{fwd}}\left(\mathbf{H}_{0:K}\right)}\right] \\ &= \mathbb{E}_{q_{\mathrm{fwd}}(\mathbf{H}_{0:K})}[\log p\left(\mathbf{X}, \mathbf{H}_K\right) - \log q_0\left(\mathbf{H}_0\right) \\ &\quad - \sum_{k=1}^{K} \log \frac{\mathcal{T}_k(\mathbf{H}_k \mid \mathbf{H}_{k-1})}{\tilde{\mathcal{T}}_k\left(\mathbf{H}_{k-1} \mid \mathbf{H}_k\right)}].\end{aligned} \tag{9}$$

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

$$\begin{aligned}&\mathcal{L}_{\mathrm{AIS}}(\psi, \gamma) \\ =&\sum_{n=1}^{N}\sum_{d=1}^{D} \mathbb{E}_{q_{\mathrm{fwd}}(\boldsymbol{h}_{0:K})q(\boldsymbol{f_d})}\left[\log p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,K}\right)\right] \\ &+ \sum_{n=1}^{N} \mathbb{E}_{q_{\mathrm{fwd}}(\boldsymbol{h}_{0:K})}\left[\log p\left(\boldsymbol{h}_{n,K}\right) - \log q_0\left(\boldsymbol{h}_{n,0}\right)\right] \\ &- \sum_{k=1}^{K} \mathbb{E}_{q_{\mathrm{fwd}}(\mathbf{H}_{0:K})} \log \frac{\mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right)}{\tilde{\mathcal{T}}_k\left(\mathbf{H}_{k-1} \mid \mathbf{H}_k\right)} \\ &- \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \,\|\, p(\boldsymbol{u}_d)\right)\end{aligned} \tag{10}$$

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 $\log p\left(x_{n,d} \mid \boldsymbol{f}_d, \boldsymbol{h}_{n,K}\right)$ 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 and Welling [2013] to obtain an unbiased estimate of the expectation over $q_{\mathrm{fwd}}\left(\mathbf{H}_{0:K}\right)$ and $q\left(\boldsymbol{f_d}\right)$ (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 and Teh [2011], Heng et al. [2020], Wu et al. [2020], Marceau-Caron and Ollivier [2017] and jointly optimize ψ and γ by stochastic gradient descent. For visualization, we present our AIS method alongside the traditional IW method's graphical model in Fig. 1.

3.2 TIME-INHOMOGENEOUS UNADJUSTED LANGEVIN DIFFUSION

$\mathcal{T}_k$ can be constructed using a Markov kernel with an invariant density such as MH or 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,

$$\begin{aligned}&\mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right) \\ &= \mathcal{N}\left(\mathbf{H}_k; \mathbf{H}_{k-1} + \eta \nabla \log q_k\left(\mathbf{H}_{k-1}\right), 2\eta I\right)\end{aligned} \tag{11}$$

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:

$$\nabla \log q_k\left(\cdot\right) = \beta_k \nabla \log p(\mathbf{X}, \cdot) + \left(1 - \beta_k\right) \nabla \log q_0(\cdot). \tag{12}$$

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

$$\begin{aligned}\nabla \log p(\mathbf{X}|\cdot) = -\frac{1}{2}\sum_{d=1}^{D} \nabla(&\log\det\left(Q_{nn} + \sigma^2 I\right) \\ &+ \left(\boldsymbol{x}_d - \boldsymbol{\mu_d}\right)^T \left(Q_{nn} + \sigma^2 I\right)^{-1} \left(\boldsymbol{x}_d - \boldsymbol{\mu_d}\right)).\end{aligned} \tag{13}$$

Since Eq. (13) 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),

$$\mathbf{d}\mathbf{H}_t = \nabla \log q_t(\mathbf{H})\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{B}_t, \quad \mathbf{H}_0 \sim q_0 \tag{14}$$

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

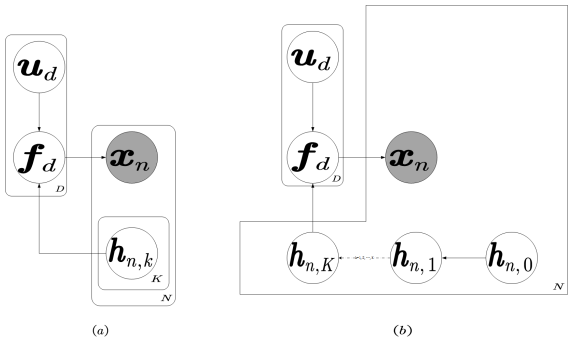

Figure 1: The graphical models of (a) IW and (b) our method. We leverages an annealing procedure to transform the posterior distribution into a sequence of intermediate distributions.

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 GPLVM 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 $\nabla \log q_k(\cdot) = \beta_k \nabla(\frac{N}{B} \log p(\mathbf{X}_J|\cdot) + \log p(\cdot)) + (1-\beta_k) \nabla \log q_0(\cdot)$
    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 descent on $\mathcal{L}(\psi, \gamma)$
until ψ, γ converge

measuring the law of $\mathbf{h}_T$ for such annealed diffusions have been showed in Andrieu et al. [2016], Tang and Zhou [2021], Fournier and Tardif [2021]. For ease of sampling, we define the corresponding Euler-Maruyama discretization as,

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

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,

$$\mathbf{H}_{k-1} = \mathbf{H}_k + \eta \nabla \log q_k(\mathbf{H}_k) + \sqrt{2\eta}\tilde{\boldsymbol{\epsilon}}_{k-1}, \quad (16)$$

where $\tilde{\boldsymbol{\epsilon}}_{k-1} = -\sqrt{\frac{\eta}{2}}[\nabla \log q_k(\mathbf{H}_{k-1}) + \nabla \log q_k(\mathbf{H}_k)] - \boldsymbol{\epsilon}_{k-1}$. Based on Eq. (11), the term related to $\mathcal{T}_k$ in Eq. (10) can be written explicitly as:

$$\begin{aligned} \sum_{k=1}^{K} R_{k-1} &= \sum_{k=1}^{K} \log \frac{\mathcal{T}_k(\mathbf{H}_k \mid \mathbf{H}_{k-1})}{\tilde{\mathcal{T}}_k(\mathbf{H}_{k-1} \mid \mathbf{H}_k)} \\ &= \sum_{k=1}^{K} \frac{1}{2}\left(\|\tilde{\boldsymbol{\epsilon}}_{k-1}\|^2 - \|\boldsymbol{\epsilon}_{k-1}\|^2\right). \end{aligned} \quad (17)$$

We abbreviate this probability ratio as R_{k-1}. Additional proofs can be seen in Appendix A.

3.3 REPARAMETERIZATION TRICK AND STOCHASTIC GRADIENT DESCENT

For ease of sampling, we consider a reparameterization version of Eq. (10) based on the Langevin mappings associated with q_k given by

$$T_k(\mathbf{H}_{k-1}) = \mathbf{H}_{k-1} + \eta \nabla \log q_k(\mathbf{H}_{k-1}) + \sqrt{2\eta}\boldsymbol{\epsilon}_{k-1}. \quad (18)$$

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,

$$\mathbf{H}_k = T_k(\mathbf{H}_{k-1}) = T_k \circ T_{k-1} \circ \cdots T_1(\mathbf{H_0}) \quad (19)$$

Moreover, for LVGP models, we also have a reparameterization version Salimbeni and Deisenroth [2017] of the posteriors of $\mathbf{H_0}$ and $\boldsymbol{f}_d$ in Eq. (10), that is,

$$\begin{aligned} &\boldsymbol{h}_{n,0} = \boldsymbol{a}_n + L_n \boldsymbol{\epsilon} \\ &\boldsymbol{f}_d = K_{nm} K_{mm}^{-1} \boldsymbol{m}_d \\ &+ \sqrt{K_{nn} - K_{nm} K_{mm}^{-1}\left(K_{mm} - \boldsymbol{S}_d{}^T \boldsymbol{S}_d\right) K_{mm}^{-1} K_{mn}}\boldsymbol{\epsilon}_{f_d} \end{aligned} \quad (20)$$

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

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-GPLVM	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)
		IWVI-GPLVM	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)
		VAIS-GPLVM (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-GPLVM	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)
		IWVI-GPLVM	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)
		VAIS-GPLVM (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)

bound in Eq. (10) can be rewritten as:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{AIS}}(\psi, \gamma) \\
&= \sum_{n=1}^{N} \sum_{d=1}^{D} \mathbb{E}_{p\left(\boldsymbol{\epsilon}_{f_d}\right) p(\boldsymbol{\epsilon}_{0:K-1}) p(\epsilon)} \left[\log p\left(x_{n,d} \mid \boldsymbol{\epsilon}_{f_d}, \boldsymbol{\epsilon}_{0:K-1}, \boldsymbol{\epsilon}\right)\right] \\
&\quad + \sum_{n=1}^{N} \mathbb{E}_{p(\boldsymbol{\epsilon}_{0:K-1}) p(\epsilon)} \left[\log p\left(\boldsymbol{h}_{n,K}\right) - \log q_0\left(\boldsymbol{h}_{n,0}\right)\right] \\
&\quad - \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{\epsilon}_{0:K-1}) p(\boldsymbol{\epsilon})} R_{k-1} - \sum_{d=1}^{D} \mathrm{KL}\left(q(\boldsymbol{u}_d) \parallel p(\boldsymbol{u}_d)\right),
\end{aligned} \tag{21}$$

where R_{k-1} is defined in Eq. (17) and $\mathbf{h}_{n,k}$ is reparameterized as $\mathbf{h}_{n,k} = T_k \circ T_{k-1} \circ \cdots T_1\left(\mathbf{h}_{n,0}\right) = \bigcirc_{i=1}^{k} T_i(\mathbf{a}_n + L_n \boldsymbol{\epsilon})$.

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 and Deisenroth [2017], Kingma and Welling [2013], Hoffman and Blei [2015], Naesseth et al. [2020] and stochastic gradient descent Welling and Teh [2011], Chen et al. [2014], Zou et al. [2019], Teh et al. [2016], Sato and Nakagawa [2014], Alexos et al. [2022] as described in Algorithm 1.

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 $p\left(\mathbf{X}, \mathbf{H}_K\right)$ 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 GPLVMs as described in Algorithm 1.

4 RELATED WORK

IWVI IWVI Domke and Sheldon [2018] demonstrated that importance weighting constitutes a form of augmented variational inference, thereby revealing the looseness inherent in previous variational objectives. This insight was later extended to the case of α-divergences Geffner and Domke, Daudel et al. [2023]. However, Rainforth et al. [2018] showed that tighter ELBOs can reduce the gradient estimator's signal-to-noise ratio (SNR), impairing inference network learning. Similarly, Salimbeni et al. [2019] addressed this in Deep Gaussian Processes (DGPs) by introducing an importance-weighted objective with latent noisy covariates, balancing accuracy and computational cost through analytic solutions.

Building on this, Rudner et al. [2021] found that increasing importance samples degrades gradient SNR for latent variable parameters, sometimes reducing gradients to pure noise. They mitigated this by adapting doubly-reparameterized gradient estimators to the DGP context, improving stability. In contrast, our method utilizes the structured intermediate distributions of AIS, inspired by Sequential Monte Carlo (SMC), to achieve a more stable and accurate variational approximation. This approach effectively avoids weight collapse, particularly in high-dimensional scenarios.

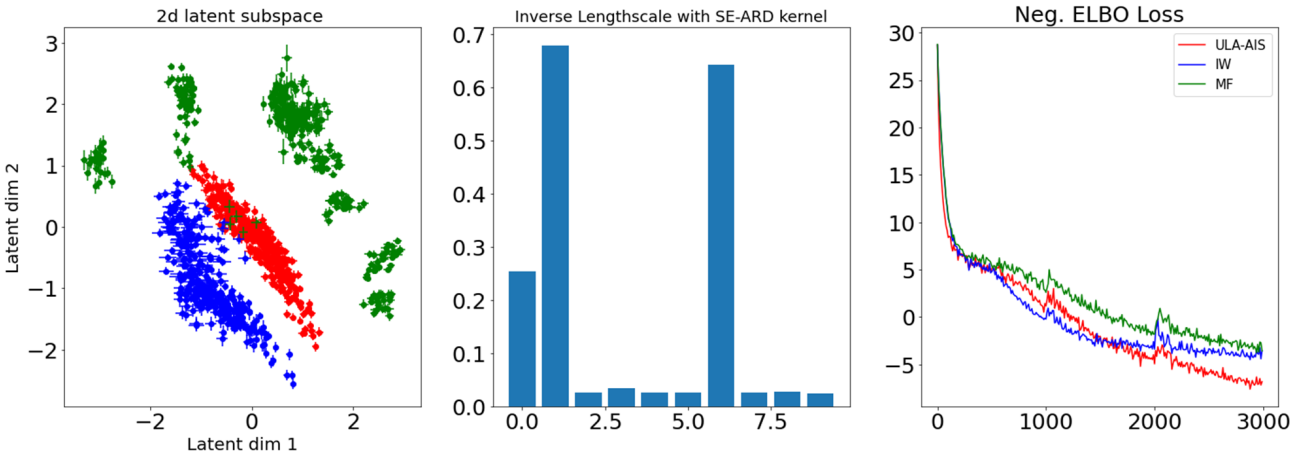

Figure 2: We lowered the data dimensionality using our proposed method in the multi-phase oilflow dataset and visualized a two-dimensional slice of the latent space that corresponds to the most dominant latent dimensions. The inverse lengthscales learnt with SE-ARD kernel for each dimension are depicted 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 for GPLVMs.

Figure 3: 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.

Differentiable AIS Our method builds upon a well-established line of research Neal [2001], Del Moral et al. [2006], Zhang et al. [2021], Xu and Campbell [2023], Chen et al. [2025], and is specifically designed to overcome the known limitations of IW methods through AIS. We highlight the key differences between our approach and the Differentiable AIS (DAIS) method proposed by Zhang et al. [2021]. DAIS circumvents the non-differentiability of traditional AIS by removing the Metropolis-Hastings correction, thereby enabling gradient-based optimization of the marginal likelihood. It has also been extended to black-box variational inference settings Jankowiak and Phan [2022]. However, unlike our method, DAIS relies on a perturbed Hamiltonian system, whereas we adopt an inhomogeneous Unadjusted Langevin Algorithm (ULA). These represent fundamentally different formalisms: Hamiltonian mechanics typically employ symplectic integrators such as leapfrog methods, while Langevin dynamics utilize reverse stochastic differential equations (SDEs). Moreover, our algorithm is grounded in nonequilibrium statistical mechanics Nilmeier et al. [2011] and is applied to Bayesian inference for Gaussian Process Latent Variable Models (GPLVM), in contrast to prior methods, which primarily focus on Bayesian linear regression.

Diffusion models Our approach shares a fundamental connection with diffusion models Ho et al. [2020], Song et al. [2020], Li et al. [2023] through the use of nonequilibrium statistical mechanics. While diffusion models, especially in generative modeling Ruthotto and Haber [2021], Croitoru et al. [2023], use reverse stochastic processes to transform latent variables back to data space, they primarily focus on data generation rather than variational inference. In contrast to previous approaches, and similarly to Xu et al. [2024, 2025], our framework leverages the Unadjusted Langevin Algorithm (ULA) to better approximate posterior distributions by directly optimizing variational objectives. Additionally, our method models latent variable dynamics through forces driving the system toward equilibrium, drawing inspiration from nonequilibrium thermodynamics where systems relax to steady states via perturbative dynamics.

5 EXPERIMENTS

5.1 BASELINE METHODS

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

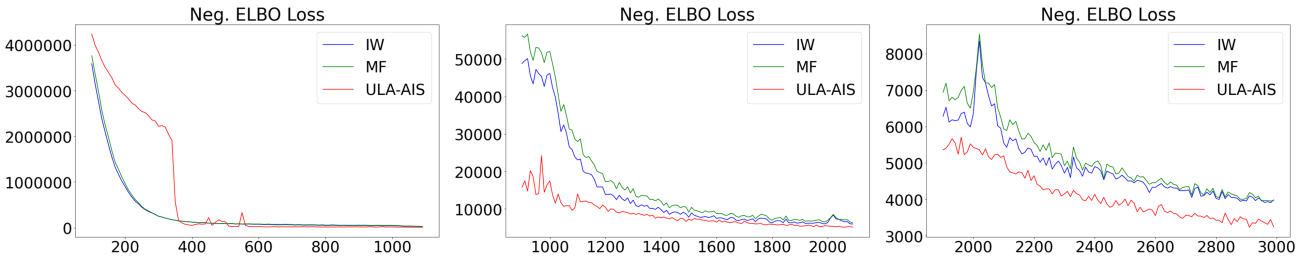

Figure 4: The negative ELBO convergence curves of the three methods on the Frey Faces dataset. It is noted that as the number of iterations increase, the y-axis scale gradually increases from left to right.

Table 2: Comparison of MF-GPLVM, IWVI-GPLVM, and VAIS-GPLVM 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-GPLVM	1000	48274 (443)	468 (9)	46027 (356)
			2000	6346 (20)	95 (1)	4771 (17)
			3000	3782 (15)	69 (0.2)	2822 (3)
		IWVI-GPLVM	1000	42396 (426)	394 (8)	39936 (312)
			2000	5643 (15)	76 (1)	4292 (13)
			3000	3596 (14)	63 (0.5)	2535 (4)
		VAIS-GPLVM (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-GPLVM	2000	-432.32(0.33)	0.27(0.004)	-552.87(0.28)
		IWVi-GPLVM	2000	-443.64(0.37)	**0.25**(0.003)	-567.13(0.31)
		VAIS-GPLVM (ours)	2000	**-453.18**(0.27)	**0.25**(0.002)	**-569.93**(0.26)

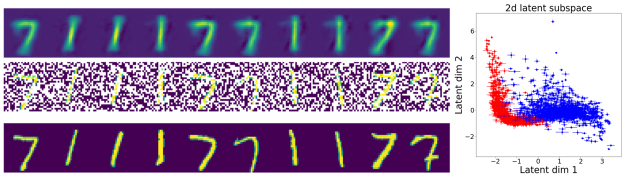

Figure 5: For MNIST with 75% missing pixels, we used digits 1 and 7. The bottom row shows ground truth, while the top row shows reconstructions from the 5D latent space. The 2D plot on the right visualizes the dimensions with the smallest lengthscales.

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.

We compare three different approaches: (a) Classical Sparse VI based on mean-field (MF) approximation Titsias and Lawrence [2010]; (b) Importance-weighted (IW) VI Salimbeni et al. [2019]; (c) The Unadjusted Langevin Diffusion Variational AIS model (hereinafter referred to as VAIS-GPLVM) is defined by the algorithm proposed 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.

5.2 DIMENSIONALITY REDUCTION

The multi-phase Oilflow data Bishop and 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 ten 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 experimental hyperparameters and compare the learning curves of two state-of-the-art models. The results are shown in Fig. 2. 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 observe that after sufficient training, our proposed method yields lower reconstruction loss and MSE than IWVI and MF methods. It is noted that our proposed method does not show an increase in time complexity compared to the baseline method IW. Therefore, even though we used a fixed number of iterations, we can ensure the fairness of the experiments.

5.3 MAKE PREDICTIONS IN UNSEEN DATA

We conducted reconstruction experiments on MNIST and Frey Faces to assess model uncertainty under missing structured inputs. For MNIST, we used digits 1 and 7 with a 5-dimensional latent space; each image is 784- dimensional. For Frey Faces Roweis and Saul [2000], we used the full dataset of 1965 images (20×28 pixels, 560-dimensional) with a 20-dimensional latent space. In both datasets, 5% of training samples had 75% of their pixels removed to test reconstruction. Results, shown in Fig. 3 and Fig. 5, reflect sampling from the learned latent distributions. Our setup follows prior work by Titsias and Lawrence [2010] and Gal et al. [2014]. Additional details are provided in the Appendix.

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.

5.4 EFFECTIVE SAMPLE SIZE (ESS) ANALYSIS

In our experiments, we observed clear evidence of weight collapse in IW-GPLVM, particularly as the dimensionality of the latent space increases. Below, we present additional results from the Brendan Faces reconstruction task, comparing IW-GPLVM and our proposed VAIS-GPLVM using standard diagnostic metrics:

Effective Sample Size (ESS) Rainforth et al. [2018] quantifies the number of samples that effectively contribute to the final estimate, despite using all K particles. It is defined as

$$\text{ESS} = \frac{1}{\sum_{k=1}^{K} \tilde{w}_k^2}, \tag{22}$$

where $\tilde{w}_k$ are the normalized importance weights. A low ESS indicates that only a few particles dominate the estimate, reflecting weight collapse.

Weight Entropy is defined as

$$H(\tilde{w}) = -\sum_{k=1}^{K} \tilde{w}_k \log \tilde{w}_k, \tag{23}$$

which measures the dispersion of the importance weights. Higher entropy suggests a more uniform distribution of weights and better utilization of available samples.

Table 3: Comparison of ESS and Weight Entropy for IWVI-GPLVM and VAIS-GPLVM (Ours) on the Brendan Faces Reconstruction Task. VAIS-GPLVM demonstrates a significant improvement in both Effective Sample Size (ESS) and Weight Entropy, indicating that it mitigates sample collapse and promotes a more uniform weight distribution.

Metric	IWVI-GPLVM	VAIS-GPLVM (Ours)
ESS ($K = 25$)	4.1	20.3
Weight Entropy ($K = 25$)	0.9	2.6

As shown in the Table 3, VAIS-GPLVM achieves substantially higher ESS and weight entropy, indicating more diverse and stable sampling behavior. In contrast, IW-GPLVM suffers from severe weight concentration, corroborating its known theoretical limitations and aligning with our earlier motivation that IWVI tends to experience weight collapse in high-dimensional settings.

6 CONCLUSION

In this paper, we propose VAIS-GPLVM, a novel variational approach for GPLVMs based on Annealed Importance Sampling. By leveraging annealing and unadjusted Langevin dynamics, our method estimates the ELBO via a sequence of tractable intermediate distributions. Empirical results on high-dimensional and structured datasets demonstrate improved variational bounds, faster convergence, and greater robustness. Notably, sharp drops in the loss curve further validate the effectiveness of our approach. Overall, VAIS-GPLVM offers a promising direction for variational learning in latent variable models.

ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Program of Guangdong, China (Grant No. 2023A1515011281). We would also like to express our sincere gratitude to the three reviewers and the meta-reviewers for their thorough and constructive feedback, which significantly helped improve the quality of this paper.

References

Antonios Alexos, Alex J Boyd, and Stephan Mandt. Structured stochastic gradient mcmc. In *International Conference on Machine Learning*, pages 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*, pages 1683–1691. PMLR, 2014.

Wei Chen, Shigui Li, Jiacheng Li, Junmei Yang, John Paisley, and Delu Zeng. Dequantified diffusion schr\" odinger bridge for density ratio estimation. *arXiv preprint arXiv:2505.05034*, 2025.

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.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 2023.

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

Kamélia Daudel, Joe Benton, Yuyang Shi, and Arnaud Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24 (243):1–83, 2023.

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.

Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. In *Third Symposium on Advances in Approximate Bayesian Inference*.

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.

Fengxiang He, Tongliang Liu, and Dacheng Tao. Why resnet works? residuals generalize. *IEEE transactions on neural networks and learning systems*, 31(12):5349–5362, 2020.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, (14):1303–1347, 2013.

Martin Jankowiak and Du Phan. Surrogate likelihoods for variational annealed importance sampling. In *International Conference on Machine Learning*, pages 9881–9901. PMLR, 2022.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

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

Shigui Li, Wei Chen, and Delu Zeng. Scire-solver: Accelerating diffusion models sampling by score-integrand solver with recursive difference. *arXiv preprint arXiv:2308.07896*, 2023.

Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI open*, 3:111–132, 2022.

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*, pages 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.

Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR, 2018.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 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*, 2014.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation*, pages 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 GJ Rudner, Oscar Key, Yarin Gal, and Tom Rainforth. On signal-to-noise ratio issues in variational inference for deep gaussian processes. In *International Conference on Machine Learning*, pages 9148–9156. PMLR, 2021.

Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2): e202100008, 2021.

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

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. *Advances in Neural Information Processing Systems*, 2017.

Hugh Salimbeni, Vincent Dutordoir, James Hensman, and Marc Deisenroth. Deep gaussian processes with importance-weighted variational inference. In *International Conference on Machine Learning*, pages 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*, pages 982–990. PMLR, 2014.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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*, 2009.

Michalis Titsias and Neil D Lawrence. Bayesian Gaussian process latent variable model. In *Artificial Intelligence and Statistics*, 2010.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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)*, pages 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.

Jian Xu, Delu Zeng, and John Paisley. Sparse inducing points in deep gaussian processes: Enhancing modeling with denoising diffusion variational inference. In *International Conference on Machine Learning*, pages 55490–55500. PMLR, 2024.

Jian Xu, Shian Du, Junmei Yang, Xinghao Ding, Delu Zeng, and John Paisley. Bayesian gaussian process odes via double normalizing flows. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Zuheng Xu and Trevor Campbell. Embracing the chaos: analysis and diagnosis of numerical instability in variational flows. *Advances in Neural Information Processing Systems*, 36:32360–32386, 2023.

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.

A DERIVATION OF EQUATION (3) AND (5)

A.1 DERIVATION OF EQUATION (3)

First, decompose the log evidence into a double summation form along the observation dimensions:

$$\log p(X) = \sum_{n=1}^{N} \sum_{d=1}^{D} \log p(x_{n,d}) \tag{24}$$

Term-wise Application of Jensen's Inequality Apply Jensen's inequality to each term $\log p(x_{n,d})$ and introduce the variational distribution to obtain:

$$\log p(x_{n,d}) \geq \mathbb{E}_{q(f_d,u_d)q(h_n)}[\log p(x_{n,d}|f_d, h_n)] - \mathrm{KL}(q(h_n)|p(h_n)) - \mathrm{KL}(q(f_d, u_d)|p(f_d, u_d)) \tag{25}$$

where $q(f_d, u_d) = p(f_d|u_d)q(u_d)$.

Sum the bounds of all terms to obtain the initial variational lower bound:

$$\text{MF-ELBO}_f(\gamma, \psi) = \sum_{n,d} \left[\mathbb{E}_{q(f_d,u_d)q(h_n)}[\log p(x_{n,d}|f_d, h_n)] - \mathrm{KL}(q(h_n)|p(h_n)) - \mathrm{KL}(q(f_d, u_d)|p(f_d, u_d))\right] \tag{26}$$

Sparse Variational Approximation:

$$q(f_d) = \int p(f_d|u_d)q(u_d)du_d \tag{27}$$

The KL term can then be simplified as:

$$\mathrm{KL}(q(f_d, u_d)|p(f_d, u_d)) = \int p(f_d|u_d)q(u_d) \log \frac{p(f_d|u_d)q(u_d)}{p(f_d|u_d)p(u_d)} du_d df_d = \mathrm{KL}(q(u_d)|p(u_d)) \tag{28}$$

and we have,

$$\mathbb{E}_{q(f_d,u_d)q(h_n)}[\log p(x_{n,d}|f_d, h_n)] = \mathbb{E}_{q(f_d)q(h_n)}[\log p(x_{n,d}|f_d, h_n)] \tag{29}$$

Substitute the simplified KL term and Equation (29) into Equation (26) to obtain the MF-ELBO consistent with the main text:

$$\text{MF-ELBO}(\gamma, \psi) = \sum_{n,d} \left[\mathbb{E}_{q(f_d)q(h_n)}[\log p(x_{n,d}|f_d, h_n)] - \mathrm{KL}(q(h_n)|p(h_n)) - \mathrm{KL}(q(u_d)|p(u_d))\right] \tag{30}$$

A.2 DERIVATION OF EQUATION (5)

Following the IWAE approach, we apply importance-weighted variational inference to the latent variable h_n (the initial steps align with mean-field variational inference and are thus omitted). The MF-ELBO MF-ELBO(γ, ψ) is rewritten in the previous step as:

$$\sum_{n,d} \left[\mathbb{E}_q \left[\log \left(\frac{p(x_{n,d} \mid f_d, h_n)p(h_n)}{q(h_n)}\right)\right] - \mathrm{KL}(q(u_d)|p(u_d))\right]. \tag{31}$$

Sampling: Independently draw K samples $h_{n,1}, \ldots, h_{n,K}$ from $q(h_n)$. Estimation Construction: Approximate the likelihood term using importance weighting:

$$\frac{p(x_{n,d} \mid f_d, h_n)p(h_n)}{q(h_n)} \approx \frac{1}{K} \sum_{k=1}^{K} \frac{p(x_{n,d} \mid f_d, h_{n,k})p(h_{n,k})}{q(h_{n,k})}. \tag{32}$$

This estimator satisfies consistency in expectation:

$$\mathbb{E}_{q(h_{n,1:K})} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{p(x_{n,d} \mid f_d, h_{n,k})p(h_{n,k})}{q(h_{n,k})}\right] = p(x_{n,d} \mid f_d). \tag{33}$$

When $K = 1$, it reduces to the mean-field case:

$$\mathbb{E}_{q(h_{n,1})}\left[\frac{p(x_{n,d} \mid f_d, h_{n,1})p(h_{n,1})}{q(h_{n,1})}\right] = p(x_{n,d} \mid f_d). \tag{34}$$

Substitute the importance-weighted estimator into the ELBO to obtain Equation 5:

$$\log p(x_{n,d}) \geq E_{q(f_d)q(h_n)}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(x_{n,d} \mid f_d, h_{n,k})p(h_{n,k})}{q(h_{n,k})}\right)\right] - \mathrm{KL}(q(u_d)|p(u_d)). \tag{35}$$

This is the IW-ELBO in Equation (5) with $\sum_{n,d}$.

B DERIVATION OF THE OVERDAMPED LANGEVIN PATH PROBABILITY RATIO

For ease of sampling, we define the corresponding Euler-Maruyama discretization as,

$$\mathbf{H}_k = \mathbf{H}_{k-1} + \eta\nabla \log q_k\left(\mathbf{H}_{k-1}\right) + \sqrt{2\eta}\boldsymbol{\epsilon}_{k-1}, \tag{36}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, I)$. Based on results by Nilmeier et al. [2011], the backward step is realized by

$$\mathbf{H}_{k-1} = \mathbf{H}_k + \eta\nabla \log q_k\left(\mathbf{H}_k\right) + \sqrt{2\eta}\tilde{\boldsymbol{\epsilon}}_{k-1}, \tag{37}$$

Thus we have,

$$\eta\nabla \log q_k\left(\mathbf{H}_{k-1}\right) + \sqrt{2\eta}\boldsymbol{\epsilon}_{k-1} = -\eta\nabla \log q_k\left(\mathbf{H}_k\right) - \sqrt{2\eta}\tilde{\boldsymbol{\epsilon}}_{k-1} \tag{38}$$

Then,

$$\tilde{\boldsymbol{\epsilon}}_{k-1} = -\sqrt{\frac{\eta}{2}}\left(\nabla \log q_k\left(\mathbf{H}_{k-1}\right) + \nabla \log q_k\left(\mathbf{H}_k\right)\right) - \boldsymbol{\epsilon}_{k-1} \tag{39}$$

Finally,

$$\begin{aligned}
\log\frac{\mathcal{T}_k\left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right)}{\tilde{\mathcal{T}}_k\left(\mathbf{H}_{k-1} \mid \mathbf{H}_k\right)} &= \log\frac{p\left(\boldsymbol{\epsilon}_{k-1}\right)\left|\det\left(\frac{\partial \mathbf{H}_k}{\partial \boldsymbol{\epsilon}_{k-1}}\right)\right|}{p\left(\tilde{\boldsymbol{\epsilon}}_{k-1}\right)\left|\det\left(\frac{\partial \mathbf{H}_{k-1}}{\partial \tilde{\boldsymbol{\epsilon}}_{k-1}}\right)\right|} \\
&= \log\frac{p\left(\boldsymbol{\epsilon}_{k-1}\right)}{p\left(\tilde{\boldsymbol{\epsilon}}_{k-1}\right)} \\
&= \frac{1}{2}\left(\left\|\tilde{\boldsymbol{\epsilon}}_{k-1}\right\|^2 - \left\|\boldsymbol{\epsilon}_{k-1}\right\|^2\right)
\end{aligned} \tag{40}$$

C A STOCHASTIC VARIANT OF VAIS-GPLVM

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 $|\mathbf{X}_J \mid = B$, where $J \subset \{1, 2, .., N\}$ is the indice of any mini-batch. We can thus define an estimator of $\nabla \log p(\mathbf{X} \mid \cdot)$ in Eq. (12) as,

$$\nabla \log p(\mathbf{X} \mid \cdot) \approx \frac{N}{B}\nabla \log p(\mathbf{X}_J \mid \cdot) \tag{41}$$

In the meantime, we replace the $p\left(\mathbf{X}, \mathbf{H}_K\right)$ term in Eq. (7) with another estimator computed using an independent mini-batch of indices $I \subset \{1, 2, .., N\}$ with $|\mathbf{X}_I \mid = B$, *i.e.*

$$p\left(\mathbf{X}, \mathbf{H}_K\right) \approx p\left(\mathbf{H}_K\right) p\left(\mathbf{X}_I \mid \mathbf{H}_K\right)^{\frac{N}{B}} \tag{42}$$

With jointly using the reparameterization trick and stochastic gradient descent, we finally derive a stochastic variant of the Stochastic Unadjusted Langevin Diffusion AIS algorithm for the LVGP models as describe in Algorithm 1. Thanks to GPU acceleration, we can extend the proposed algorithm to larger datasets, such as image-based visual tasks.

D PRACTICAL GUIDELINES

In the context of this paper, the posterior distribution refers to the distribution of the latent variables given the observed data. This distribution is often intractable and challenging to sample from directly. VAIS-GPLVM aims to approximate this posterior distribution by transforming it into a sequence of intermediate distributions, which can be more tractable and easier to sample from.

The annealing process gradually transforms the posterior distribution by introducing a temperature parameter β . By annealing from $\beta = 0$ to $\beta = 1$, we move from an initial distribution, where the posterior is approximated by a simpler distribution to the target posterior distribution itself. The key idea behind annealing is it allows for a smoother exploration of the posterior space. At each intermediate distribution, we can use importance sampling to estimate the evidence by sampling from the proposal distribution and reweighting the samples using the ratios of the target and proposal distributions.

As the annealing process progresses, the samples from the proposal distribution gradually become more representative of the target distribution. This means that the exploration of the posterior space is not limited to a specific region but covers a wider range of possible configurations of the latent variables.

The benefit of this exploration is that it allows for a more accurate estimation of the evidence, which corresponds to a tighter lower bound in the variational learning framework. By gradually annealing the temperature and exploring different distributions, VAIS-GPLVM can capture more complex structures in the posterior distribution, leading to better variational approximations in complex data and high-dimensional spaces.

When using the Unadjusted Langevin Diffusion method for sampling, one key challenge is to determine an appropriate step size η_k A fixed step size may work well for some samples but may be suboptimal for others. To address this issue, we can use the Adagrad Kingma and Ba [2014] optimizer to adaptively adjust the step size based on the historical gradient information. Specifically, for each dimension of the sampled variables, we divide the initial step size by the square root of the sum of squared gradient values for that dimension up to a noise. This technique can help achieve better performance and faster convergence, especially when dealing with complex and high-dimensional distributions where finding an appropriate step size is challenging. The adaptive step size adjustment can be implemented in combination with other techniques, such as early stopping, to further improve the sampling efficiency.

$$\eta_k = 0.9\eta_{k-1} + 0.1\frac{\eta_0}{\sqrt{G_k + \epsilon}}$$

where G_k is the sum of squared gradient values up to step k in Eq. (17), ϵ is a small smoothing term to avoid division by zero,and η_0 is the initial step size.

In the context of Annealed Importance Sampling (AIS), choosing an optimal temperature schedule β_k is a challenging task. When choosing an appropriate annealing schedule for Stochastic Gradient Annealed Importance Sampling, there are several trade-offs and considerations to keep in mind:

- Computational Efficiency: The annealing schedule should be carefully designed to balance the computational resources required for estimating the evidence. Too many bridging densities can lead to excessive computational burden, while too few densities may result in less accurate estimates.
- Exploration vs Exploitation: The annealing schedule should strike a balance between exploration and exploitation of the posterior distribution. An aggressive schedule that moves quickly from the base distribution to the posterior may lead to exploration limitations, while a slow schedule may lead to insufficient exploration and inefficiency.
- Smoothness of Transition: The annealing schedule should ensure a smooth transition between bridging densities. Abrupt changes in the densities can result in high-variance importance weights, which may lead to inaccurate estimates. Smooth transitions can be achieved by gradually adjusting the temperature or using appropriate interpolation functions.

We often use a linear schedule, where the temperature values are fixed and regularly spaced between 0 and 1. However, this approach may not always work well in practice, as the search space is complex and high-dimensional.

Alternatively, we can try to learn the temperature values β_k directly as additional inference parameters ϕ. This can be done using various techniques, such as gradient-based optimization. By doing so, we can obtain a temperature schedule that is tailored to the specific problem at hand and achieve better sampling performance. Additional experimental information can be seen in Table 4.

Dataset	Task	N	D	Z	Q	LR	K
Oilflow	Dimensionality Reduction	1000	12	50	10	0.02	5
Wine Quailty	Dimensionality Reduction	1599	11	50	9	0.02	5
Frey Face	Missing Data Recovery	1965	560	50	20	0.02	25
MNIST	Missing Data Recovery	2163	784	50	5	0.02	25

Table 4: Training experimental configuration where N and D denote the number of data points and data space dimensions, Z denotes the number of inducing inputs shared across dimensions, Q denotes the dimesionality of the latent space, LR denotes the learning rate, K denotes the length of the transition chain in VAIS-GPLVM and in IW K denotes the number of repetitions of sampling .

E DETAILS FOR IMPLEMENTING ON MISSING DATA TASKS

Specially, our training procedure leverages the marginalisation principle of Gaussian distributions and the fact that the data dependent terms of the ELBO factorise across data points and dimensions. This means we can trivially marginalise out the missing dimensions $\boldsymbol{x}_a$, because each individual data point $\boldsymbol{x}$ is modelled as a joint Gaussian. Consider a high-dimensional point $\boldsymbol{x}$ which we split into observed, $\boldsymbol{x}_o$ and unobserved $\boldsymbol{x}_a$ dimensions,

$$\int \prod_{d\in a} \prod_{d\in o} p\left(\boldsymbol{x}_a, \boldsymbol{x}_o \mid \boldsymbol{f}_d, \mathbf{H}\right) d\boldsymbol{x}_a = \prod_{d\in o} p\left(\boldsymbol{x}_o \mid \boldsymbol{f}_d, \mathbf{H}\right) \tag{43}$$

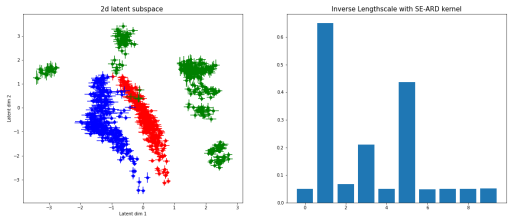

Figure 6: Dimensionality Reduction Results for MF method.

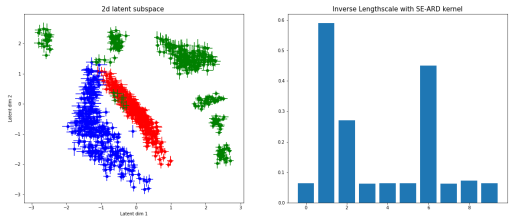

Figure 7: Dimensionality Reduction Results for IW method.

In this formula, the indices of missing and observed dimensions are denoted by a and o respectively, where $D = a \cup o$ represents all dimensions in the data. The marginal distributions $\boldsymbol{f}_d \in \mathbb{R}^N$ are defined in Eq. (4).The latent variables $\boldsymbol{h}_n$ for each data point are informed only by the observed dimensions. Furthermore, we can easily reconstruct the missing dimensions during training by constructing a variational latent distribution $q(\mathbf{H})$, as described in Section 4. This approach enables us to efficiently handle missing dimensions in high-dimensional datasets without requiring major modifications to the overall training process.

E.1 COMPARED TO STANDARD GPLVM

We have also conducted additional experiments comparing our proposed approach to the Standard GPLVM Lawrence [2003] in Table 5. We performed experiments 10 times and averaged the results, analyzing the performance (in terms of MSE and NLL) on four different datasets. Due to limited computational resources, we were only able to run the Standard GPLVM on a subset of the image datasets. For the image reconstruction task, we randomly selected 300 images as the training set and used consistent hyperparameters for the other experiments.

E.2 RUNTIME ANALYSIS

We observed that the runtime of Importance-Weighted (IW) VI and VAIS-GPLVM increases almost linearly with K. For IW, this is due to the K repeated samplings of latent variables, each with a complexity of $O(nm^2)$ from the GPLVM model. As K increases, the repeated samplings dominate the runtime. In contrast, VAIS-GPLVM requires only one such sampling,

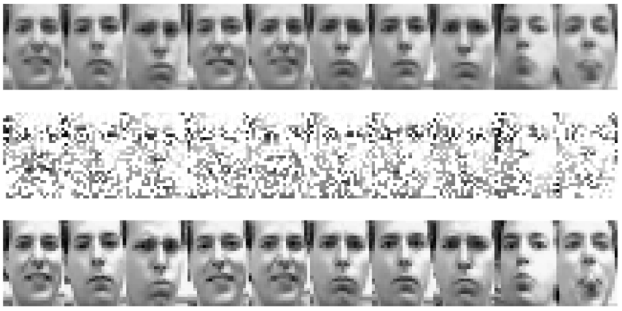

Figure 8: Missing Data Recovery Results for MF method. The bottom row represents the ground truth data and the top row showcases the reconstructions from the 20-dimensional latent distribution.

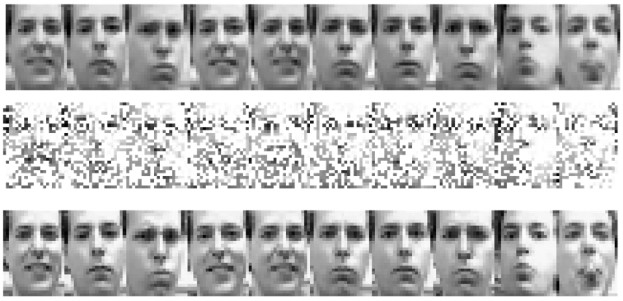

Figure 9: Missing Data Recovery Results for IW method. The bottom row represents the ground truth data and the top row showcases the reconstructions from the 20-dimensional latent distribution.

Dataset (Size)	**Method**	**MSE**	**NLL**
Oilflow (1000, 12)	Standard GPLVM	2.45 (0.05)	-12.42 (0.07)
	VAIS-GPLVM (Ours)	**1.71** (0.04)	**-15.81** (0.04)
Wine Quality (1599, 11)	Standard GPLVM	30.53 (0.03)	2.82 (0.02)
	VAIS-GPLVM (Ours)	30.79 (0.04)	**2.42** (0.03)
Frey Faces (300, 560)	Standard GPLVM	130.00 (7.00)	2632.00 (6.00)
	VAIS-GPLVM (Ours)	**115.00** (6.00)	**2417.00** (5.00)
MNIST (300, 784)	Standard GPLVM	0.36 (0.01)	-484.00 (3.00)
	VAIS-GPLVM (Ours)	**0.31** (0.01)	**-496.00** (2.00)

Table 5: Comparison of MSE and NLL between Standard GPLVM and our VAIS-GPLVM across four datasets.

with additional computations focused on the lighter Langevin stochastic flow during annealing. As shown in Table 6, AIS becomes more efficient than IW when K exceeds a certain threshold on the Frey Faces dataset.

Method	$K = 5$	$K = 10$	$K = 15$	$K = 20$	$K = 25$
IWVI-GPLVM	1.46s	2.85s	4.06s	5.45s	7.03s
VAIS-GPLVM (Ours)	1.53s	2.65s	3.79s	4.80s	5.93s

Table 6: Comparison of running time between IWVI-GPLVM and VAIS-GPLVM in one epoch for Frey Faces

E.3 ADDITIONAL RESULTS

In this section, we will demonstrate the visual effects of the MF and IW methods on three datasets: Oilflow, MINIST, and Frey Faces. These visualizations will be used for comparison with the main text. There results can be seen in Fig. 6, Fig.7, Fig.8, Fig.9, Fig.10, Fig.11.

From the visual appearance, it may seem that all three methods produce similar reconstructions. However, upon closer inspection, we can observe differences in certain details such as brightness and contrast. While these differences may be difficult to discern with the naked eye, we have quantified them using the mean squared error (MSE) between the

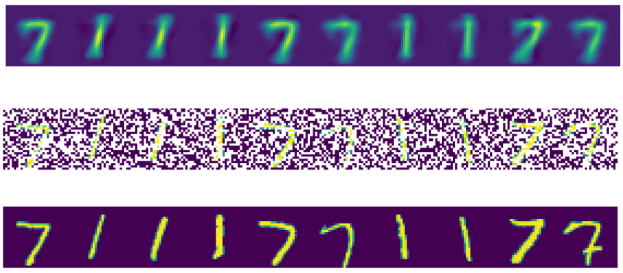

Figure 10: Missing Data Recovery Results for MF method. The top row represents the ground truth data and the bottom row showcases the reconstructions from the 5-dimensional latent distribution.

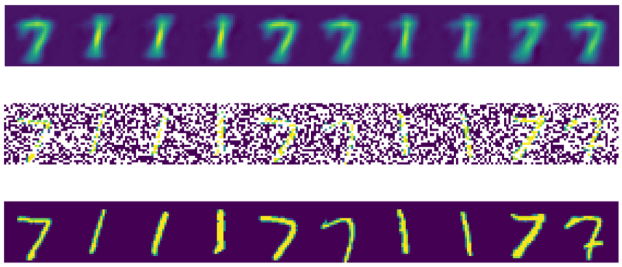

Figure 11: Missing Data Recovery Results for IW method. The top row represents the ground truth data and the bottom row showcases the reconstructions from the 5-dimensional latent distribution.

reconstructed images and the ground truth. The MSE results for all three methods on the test set are reported in Tables 2 in the main text.

F LIMITATIONS AND FUTURE WORK

One potential limitation could be the scalability of the method. As the size of the dataset increases, the computational resources required for estimating the evidence using VAIS-GPLVM may become more demanding. This is particularly true for large-scale datasets such as ImageNet, which contain millions of images. Running experiments on such massive datasets might pose challenges in terms of computational efficiency and memory requirements. Given that ImageNet involves higher-dimensional data, it may be more appropriate to combine GPLVM with other deep learning tools, such as convolutional neural networks (CNNs) Li et al. [2021], He et al. [2020] and transformers Vaswani [2017], Lin et al. [2022]. Broader application scenarios are currently being explored to incorporate these tools effectively and leave room for future work.

Additionally, the annealing schedule plays a crucial role in the exploration of the posterior distribution. Designing an appropriate annealing schedule may require domain knowledge or trial and error experimentation. It might be necessary to tune the schedule to ensure a balance between exploration and exploitation, as well as a smooth transition between bridging densities.

Regarding the applicability of VAIS-GPLVM in real-world applications, its performance may depend on the specific characteristics and requirements of the domain. Different datasets and applications may exhibit unique challenges, such as data sparsity, high dimensionality, or non-linear relationships, which could affect the effectiveness of SG-AIS. Evaluating the performance of SG-AIS in different domains and addressing these challenges would require further experimentation and investigation.