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

Anonymous authors
Paper under double-blind review

A DERIVATION OF THE OVERDAMPED LANGEVIN PATH PROBABILITY RATIO

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

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

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

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

Thus we have,

$$\eta \nabla \log q_k \left(\mathbf{H}_{k-1}\right) + \sqrt{2\eta}\boldsymbol{\epsilon}_{k-1} = -\eta \nabla \log q_k \left(\mathbf{H}_k\right) - \sqrt{2\eta}\tilde{\boldsymbol{\epsilon}}_{k-1} \tag{3}$$

Then,

$$\tilde{\boldsymbol{\epsilon}}_{k-1} = -\sqrt{\frac{\eta}{2}} \left(\nabla \log q_k \left(\mathbf{H}_{k-1}\right) + \nabla \log q_k \left(\mathbf{H}_k\right)\right) - \boldsymbol{\epsilon}_{k-1} \tag{4}$$

Finally,

$$\log \frac{\mathcal{T}_k \left(\mathbf{H}_k \mid \mathbf{H}_{k-1}\right)}{\tilde{\mathcal{T}}_k \left(\mathbf{H}_{k-1} \mid \mathbf{H}_k\right)} = \log \frac{p\left(\boldsymbol{\epsilon}_{k-1}\right) \left|\det\left(\frac{\partial \mathbf{H}_k}{\partial \boldsymbol{\epsilon}_{k-1}}\right)\right|}{p\left(\tilde{\boldsymbol{\epsilon}}_{k-1}\right) \left|\det\left(\frac{\partial \mathbf{H}_{k-1}}{\partial \tilde{\boldsymbol{\epsilon}}_{k-1}}\right)\right|} = \log \frac{p\left(\boldsymbol{\epsilon}_{k-1}\right)}{p\left(\tilde{\boldsymbol{\epsilon}}_{k-1}\right)} = \frac{1}{2}\left(\|\tilde{\boldsymbol{\epsilon}}_{k-1}\|^2 - \|\boldsymbol{\epsilon}_{k-1}\|^2\right) \tag{5}$$

B A STOCHASTIC VARIANT OF SG-AIS

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. We can thus define an estimator of $\nabla \log p(\mathbf{X} \mid \cdot)$ in Eq. (12) as,

$$\nabla \log p(\mathbf{X} \mid \cdot) \approx \frac{N}{B} \nabla \log p(\mathbf{X}_J \mid \cdot) \tag{6}$$

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 $|\mathbf{X}_I| = B$, *i.e.*

$$p\left(\mathbf{X}, \mathbf{H}_K\right) \approx p\left(\mathbf{H}_K\right) p\left(\mathbf{X}_I \mid \mathbf{H}_K\right)^{\frac{N}{B}} \tag{7}$$

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.

C Related work

In this section, we would like to clarify the differences between our approach and the recently proposed Differentiable AIS (DAIS) method Zhang et al. (2021). DAIS was introduced to avoid the non-differentiability issue in AIS due to the Metropolis-Hastings correction step. To achieve differentiability and enable the use of gradient-based optimization for maximizing the marginal likelihood as the objective function, DAIS was proposed as a variant of AIS by bypassing the Metropolis-Hastings correction.

Furthermore, in Jankowiak & Phan (2022), DAIS was combined with variational inference for black-box inference in probabilistic programming frameworks. However, a significant difference from the algorithm discussed in this paper is that DAIS is not an inhomogeneous Unadjusted Langevin Algorithm (ULA), but a perturbed Hamiltonian system. It is known that Hamiltonian mechanics and Langevin dynamics belong to different dynamics formalisms, i.e., classical mechanics and Langevin statistical mechanics, respectively. For example, leapfrog integration is commonly used in Hamiltonian systems, while the ULA system is typically obtained by solving a reverse stochastic differential equation (SDE).

Our algorithm takes inspiration from nonequilibrium statistical mechanics Nilmeier et al. (2011) and applies it to the inference of the Bayesian Gaussian process latent variable model (GPLVM). It is important to note that the previous two methods do not discuss the model presented in this paper but provide preliminary evidence of the effectiveness of the method in Bayesian linear regression.

Therefore, while there are some similarities in terms of using differentiable optimization and combining variational inference with AIS, our approach utilizes nonequilibrium dynamics and is tailored for the Bayesian GPLVM model.

D Practical Guidelines

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

Table 1: 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 ULA-AIS and in IW K denotes the number of repetitions of sampling .

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 & 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. We often use a linear schedule, where the temperature values are fixed and

Table 2: Comparison of running time among MF, IW, and AIS algorithms in one epoch

Datasets	Method	Time				
Frey Faces	MF	0.32s				
	IW	1.46s (K=5)	2.85s(K=10)	4.06s(K=15)	5.45s(K=20)	7.03s(K=25)
	AIS(Ous)	1.53s (K=5)	2.65s (K=10)	3.79s(K=15)	4.80s(K=20)	5.93s (K=25)

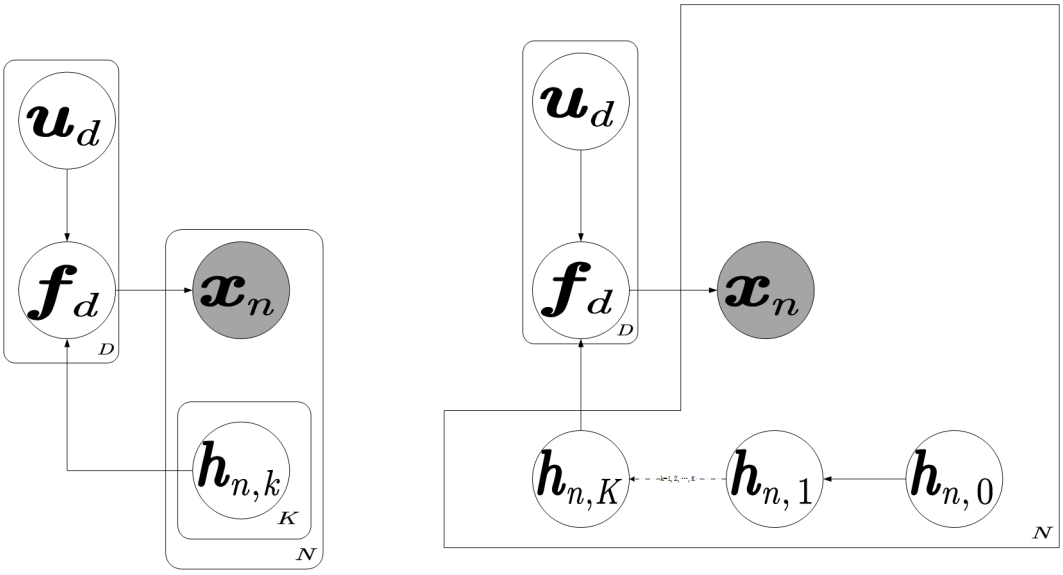

Figure 1: The graphical models of IW and our method. Unlike IW mehtod, only one sampling of latent variables to data is required for our AIS method.

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 1 [1].

E RUNTIME ANALYSIS.

In our experiments, we observed that the time complexity of Importance-weighted (IW) VI and SG-AIS almost linearly increases with K as K increases.

In the IW algorithm, the time complexity mainly stems from the K repeated samplings of latent variables to data, which is determined by the time complexity of the GPLVM model itself, $O(nm^2)$. As a result, as we increase the number of samples K, the frequency of repeated samplings increases, leading to a linear increase in time complexity.

In the AIS algorithm, only one sampling of latent variables to data is required, while the intermediate variable sampling is allocated to the annealing procedure, specifically the computation of Langevin stochastic flow. This sampling process is relatively less complex compared to the time complexity of the GPLVM model itself. As depicted in Fig 5.

Therefore, on Frey Faces dataset, as depicted in Table 4, compared to IW, the time complexity of AIS becomes lower as K reaches a certain threshold.

[1]We test all of our experiments on NVIDIA A100

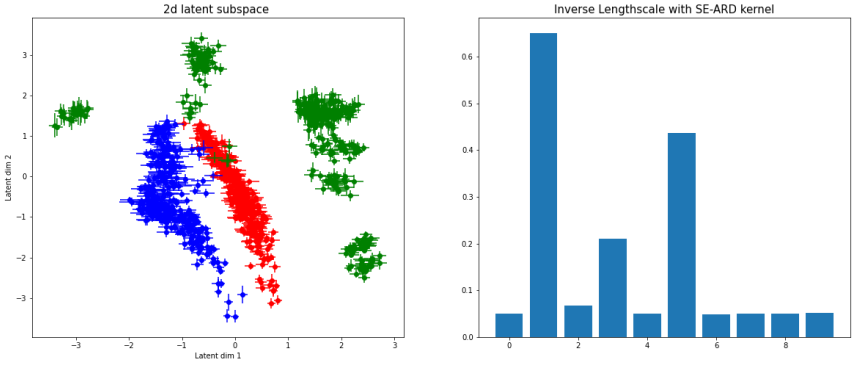

Figure 2: Dimensional Reduction Results for MF method

F 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,

$$\int \prod_{d\in a} \prod_{d\in o} p\left(\boldsymbol{x}_a, \boldsymbol{x}_o \mid \boldsymbol{f}_d, \mathbf{H}\right) d\boldsymbol{x}_a = \prod_{d\in o} p\left(\boldsymbol{x}_o \mid \boldsymbol{f}_d, \mathbf{H}\right) \tag{8}$$

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.

G 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.

H Limitations

There are limitations to our method. Firstly, it increases the time complexity by a factor of K, making it slower for larger datasets such as RGB images. Secondly, our method requires more hyperparameter tuning based on the data, such as the choice of annealing schedule, transition kernel, step size adjustment, and the length K of the transition chain. Without additional prior information, tuning these parameters can be challenging and lead to suboptimal results.

I Code Contributions

We have packaged the full code of this paper in the supplementary materials.

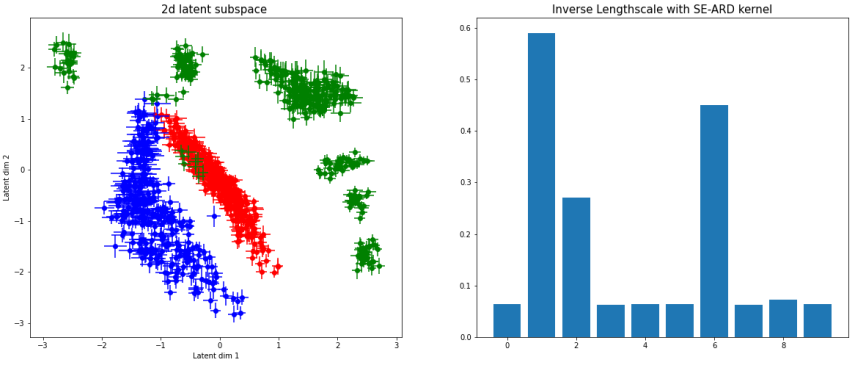

Figure 3: Dimensional Reduction Results for IW method.

Figure 4: 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

REFERENCES

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

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

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.

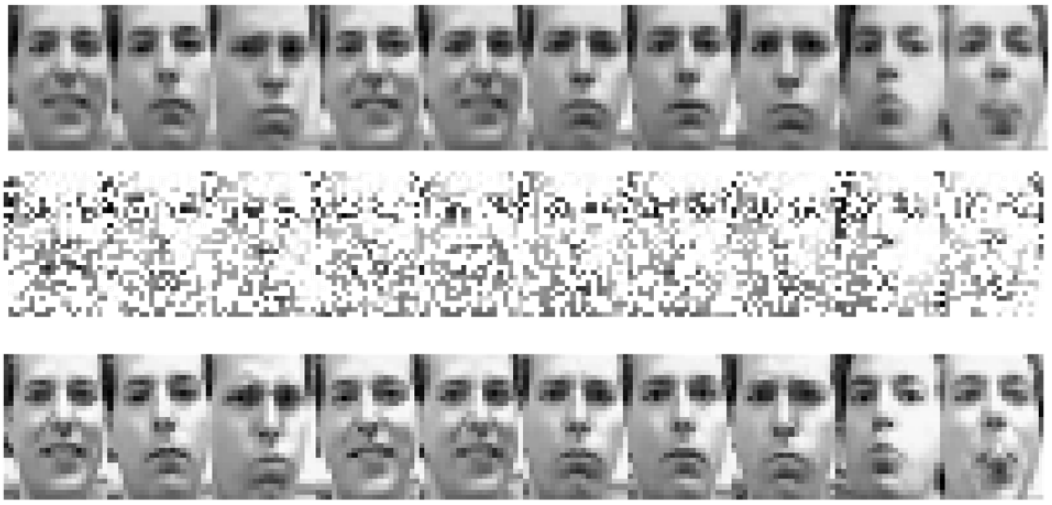

Figure 5: Missing Data Recovery Results for IW method. The bottom row represents the ground truth data and the top row showcases the reconstructions from the 20-dimensional latent distribution

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.

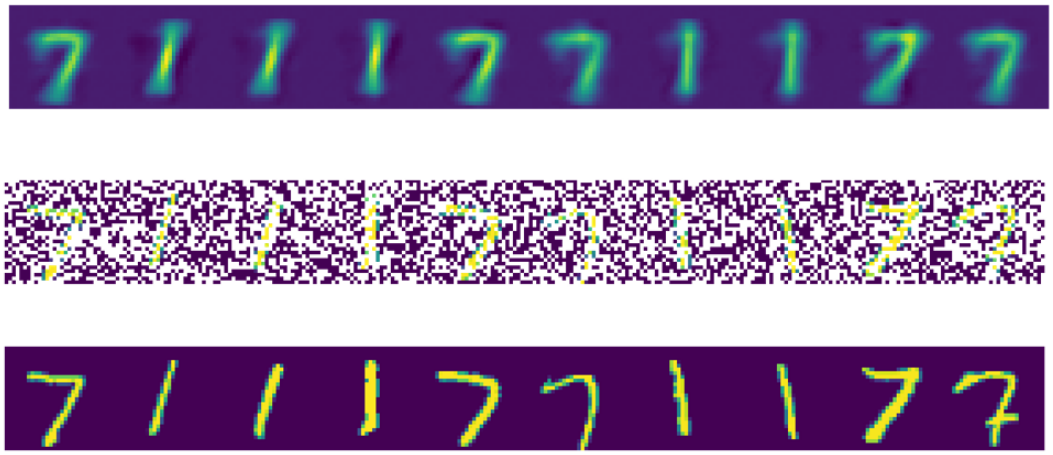

Figure 6: Missing Data Recovery Results for MF method. The top row represents the ground truth data and the bottom row showcases the reconstructions from the 5-dimensional latent distribution

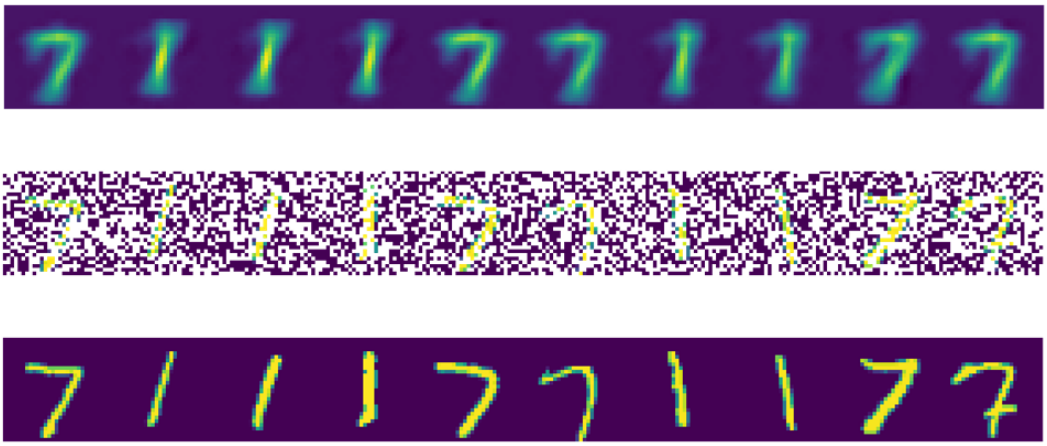

Figure 7: Missing Data Recovery Results for IW method. The top row represents the ground truth data and the bottom row showcases the reconstructions from the 5-dimensional latent distribution