Kernel to Kinetics: From Gaussian Process to Hamiltonian Dynamics in Deep Probabilistic Models

Candidate Number 31326
The London School of Economics and Political Science

Abstract

This study investigates the scalability and adaptability of probabilistic machine learning and deep learning architecture across increasing data complexities and dimensions, transitioning from supervised to unsupervised learning frameworks. We begin with Gaussian Process (GP) classification, applied to the Ionosphere dataset, where labelled data from a phased array of 16 high-frequency antennas are modelled. However, this type of labelling is impractical for higher-dimensional, more complex datasets, prompting us to extend our analysis to the deep clustering of the MNIST dataset using unsupervised learning approaches. We first address the constraints of discrete latent space models such as Gaussian Mixture Models (GMM), establishing the need for models that capitalize on the representational power of neural networks. Variational Autoencoder (VAE) rise to this challenge by merging the representational power of neural networks with the robustness of variational inference to effectively navigate and model high-dimensional spaces. However, operational efficiency of VAE demands low-variance, unbiased estimators of the evidence lower bound (ELBO) and its gradients. We attempt to combine Hamiltonian Monte Carlo (HMC) with VAE (HMCVAE) and further proposed a novel Hamiltonian Importance Sampling Variational Auto-Encoder (HISVAE) to address this need. We benchmarked the disentanglement performance of HMCVAE and HISVAE against conventional VAE and Beta-VAE models, evaluating their ability to separate underlying factors of variation. Our empirical and theoretical contributions demonstrate how these models not only refine our understanding of latent structures but also significantly advance the frontier of machine learning research, inspiring theories in other domain that bridges the gap between human and artificial cognition.

1 Introduction

The narrative of machine learning evolution is fundamentally about enhancing our capacity to extract and comprehend latent patterns embedded in vast datasets. Our journey begins with Gaussian Processes (GPs), which excel in modelling nonlinear dependencies within moderate-dimensional signal spaces [1]. As the complexity of datasets increases, we transition to Gaussian Mixture Models (GMMs) for their improved scalability and adaptability[2]. Despite their theoretical appeal, GMMs often struggle under the burden of high dimensionality, a frequent challenge in complex datasets.

The advent of Variational Auto-Encoders (VAEs) marks a significant breakthrough, merging the representational power of neural networks with the robustness of variational inference[3]. This synthesis allows for effective navigation and modelling of high-dimensional spaces, proving pivotal not only for practical applications such as image and video generation but also enriching theoretical research[4]. Fields as diverse as human-like machine intelligence and theoretical neuroscience are impacted, with variational inference and VAEs inspiring novel approaches to understanding concepts

Code and Supplementary Materials Available at: https://github.com/edluyuan/HISVAE

like the free energy principle and active inference, offering profound insights into mechanisms underlying human-like reasoning and learning[5].

Operational efficiency in VAEs, however, necessitates precise, low-variance estimators for the Evidence Lower Bound (ELBO) and its gradients. Traditional mean-field parametrization typically lacks the required flexibility for detailed posterior approximations[6]. Addressing these challenges, we integrate Hamiltonian Monte Carlo (HMC) dynamics within the VAE framework, forging a dynamic model that incorporates explicit target information efficiently[7]. Contrary to approaches utilizing time-inhomogeneous dynamics, our model employs fixed step sizes and steps in the leapfrog integration, simplifying implementation while maintaining a balance between computational efficiency and the fidelity of posterior approximation. This method leverages a series of MCMC iterations within an augmented space, using both forward and reverse Markov chains to form an augmented target distribution that conserves the original posterior as a marginal distribution[8].

The resulting Hamiltonian Importance Sampling Variational Auto-Encoder (HISVAE) synthesizes efficient stochastic evolution with accurate posterior inference, substantially improving the interpretability and applicability of machine learning in handling complex, high-dimensional datasets. This work significantly furthers the discussion on representation disengagement, and simplifies certain dynamic components to prioritize computational tractability and reproducibility[9].

2 Gaussian Process Classification

Gaussian Processes (GPs) are a Bayesian non-parametric approach used primarily for regression and probabilistic classification. In classification, a GP is employed to infer a latent function f from which observations can be classified[10].

2.1 Kernel Specification

The power of GPs lies in their use of kernels to implicitly map data to a high-dimensional space without having to compute the transformation explicitly:

2.1.1 RBF Kernel

The Radial Basis Function (RBF) kernel, also known as the squared exponential kernel, is defined as:

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right)$$

where l is the length-scale parameter which determines the smoothness of the function, and σ^2 is the variance parameter that controls the vertical variation.

2.1.2 White Kernel

The White kernel adds noise to the model[11], representing it as:

$$k_{\text{white}}(x, x') = \sigma_n^2 \delta_{xx'}$$

where σ_n^2 is the noise variance and $\delta_{xx'}$ is the Kronecker delta, 1 if $x = x'$ and 0 otherwise.

2.2 Model Optimization

The parameters of the Gaussian Process are optimized by maximizing the log marginal likelihood of the observed data:

$$\log p(y|X) = -\frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} y - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{n}{2} \log 2\pi$$

where K is the covariance matrix computed from the kernel function over all pairs of training instances, and I is the identity matrix[12].

2.3 Prediction and Evaluation

The predictive distribution at a new point x^* is given by a Bernoulli distribution parameterized by the sigmoid of the latent function:

$$p(y^* = 1|x^*, X, y) = \sigma(f(x^*))$$

where $f(x^*)$ is normally distributed with mean and variance derived from the GP posterior.

3 Gaussian Mixture Model Clustering

Gaussian Mixture Models (GMMs) are probabilistic models that assume all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters[13]. A Gaussian Mixture Model represents a composite distribution whose density function is given as a weighted sum of Gaussian components, it is expressed as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where: x is a data point. K is the number of Gaussian components in the mixture. π_k are the mixing weights satisfying $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. $\mathcal{N}(x|\mu_k, \Sigma_k)$ denotes the Gaussian distribution for the k-th component with mean μ_k and covariance matrix Σ_k. Mixture Weights (π_k): The probabilities associated with each Gaussian component in the mixture. Means (μ_k): The mean of each Gaussian component. Covariances (Σ_k): The covariance matrices of each Gaussian component, determining the spread and orientation of the component in the data space.

3.1 Expectation-Maximization Algorithm

The EM algorithm is used to find the maximum likelihood estimates of the parameters in a GMM[14], involving the following steps iteratively:

3.1.1 Expectation Step (E-step):

Calculate the responsibility $\gamma(z_{nk})$ that component k has for data point x_n:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)},$$

where $\gamma(z_{nk})$ denotes the responsibility of component k for data point x_n.

3.1.2 Maximization Step (M-step):

Update the parameters using the current responsibilities:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n,$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^\top,$$

$$\pi_k = \frac{N_k}{N},$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

4 Variational Autoencoders for Deep Clustering

VAEs are deep generative models that learn a probabilistic mapping of data points into an abstract latent space. The key components of a VAE include the encoder, the decoder, and the loss function, which incorporates both reconstruction loss and a regularization term derived from the Kullback-Leibler divergence By encoding data into a latent space, VAEs facilitate both efficient data compression and meaningful data generation[3, 15].

4.1 Model Architecture

4.1.1 Encoder

The encoder component of a VAE transforms input data x into a distribution in the latent space characterized by parameters μ and σ^2, representing the mean and variance:

$$q_\phi(z|x) = \mathcal{N}(z; \mu_\phi(x), \sigma^2_\phi(x))$$

4.1.2 Decoder

The decoder maps latent variables back to the data space, aiming to reconstruct the input:

$$p_\theta(x|z) = \text{Bernoulli}(x; \sigma_\theta(z))$$

4.1.3 Reparameterization Trick

This step introduces stochasticity essential for gradient-based optimization[16]:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

4.2 Loss Function

The ELBO, serving as the loss function, combines reconstruction fidelity with a regularization term:

$$\mathcal{L}(\phi, \theta; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{D}_{\text{KL}}(q_\phi(z|x) \| p(z))$$

where the first term is the expected log-likelihood of the observed data (reconstruction loss), and the second term is the Kullback-Leibler divergence between the encoded distribution and the prior distribution of the latent variables (regularization).

4.3 Hamiltonian Monte Carlo Variational Autoencoders (HMCVAE)

To overcome the limitations in sampling efficiency and the variational gap in standard VAEs, HMC-VAE integrates Hamiltonian Monte Carlo into the VAE framework[17]:

$$H(z, p) = \frac{1}{2} p^\top p + V(z), \quad V(z) = -\log p_\theta(x|z) + \text{D}_{\text{KL}}(q_\phi(z|x) \| p(z))$$

HMC enhances the exploration of the latent space by utilizing physical dynamics, which helps in drawing samples that are more representative of the target posterior[18]. This integration is aimed at refining the accuracy of the ELBO estimation by improving the quality of samples used in the expectation calculation:

$$\mathcal{L}(\phi, \theta; x) = \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x|z^{(l)}) - \text{D}_{\text{KL}}(q_\phi(z|x) \| p(z)) = \text{BCE}(x, \hat{x}) + \beta \cdot \mathbb{E}[H(z, p)]$$

where $z^{(l)}$ are samples generated through HMC dynamics, ensuring that these samples are more effectively distributed according to the true posterior.

4.4 Hamiltonian Importance Sampling Variational Autoencoders (HISVAE)

HISVAE advances this approach by incorporating importance sampling into the HMC framework which adjusts sample weights based on changes in the Hamiltonian to refine ELBO estimation[9]:

$$w = \exp(H(z_{\text{old}}, p_{\text{old}}) - H(z_{\text{new}}, p_{\text{new}}))$$

This is particularly critical in scenarios where the variational distribution cannot adequately approximate complex posteriors. The key modification involves adjusting the ELBO calculation by weighting each sample according to its importance:

$$\mathcal{L}(\phi, \theta; x) = \sum_{l=1}^{L} w_l \log p_\theta(x|z^{(l)}) - \sum_{l=1}^{L} w_l \log \frac{q_\phi(z^{(l)}|x)}{p(z^{(l)})},$$

where w_l are the weights computed based on the Hamiltonian dynamics, specifically accounting for the energy differences in the system states induced by the leapfrog steps[19].

Algorithm 1 Leapfrog Steps for Hamiltonian Monte Carlo

function LEAPFROG($z, p, \nabla U, \epsilon, L$)
  Input:
  z ▷ Current position in the latent space
  p ▷ Current momentum associated with z
  ∇U ▷ Gradient of the potential energy
  ϵ ▷ Step size for integration
  L ▷ Number of leapfrog steps to perform
  Output:
  z, p ▷ New position and momentum after L leapfrog steps
  for $l = 1$ to L **do**
    $p \leftarrow p - \frac{\epsilon}{2} \cdot \nabla U(z)$ ▷ Half-step update for momentum
    $z \leftarrow z + \epsilon \cdot p$ ▷ Full-step update for position
    $p \leftarrow p - \frac{\epsilon}{2} \cdot \nabla U(z)$ ▷ Half-step update for momentum
  end for
  return z, p
end function

5 Experiment

In the experimental section of our study, the Gaussian Process (GP) classification was performed using the GPy framework, For the deep clustering task, we constructed a GMM-EM model with scikit-learn. All VAEs were trained using PyTorch, .

5.1 Datasets

Our exploration begins with the Ionosphere dataset, which comprises features derived from radar signals collected to discriminate between structured and random disturbances in the ionosphere. This dataset is ideal for our GP classification task. We used the MNIST dataset to test the limits of GMM-EM as well as testing performance of VAEs, MNIST consists of 28x28 pixel grayscale images of handwritten digits. Both datasets are archetypal benchmarks in their respective domains: Ionosphere for classification and MNIST for deep clustering. Their lightweight nature makes them amenable for training a family of VAEs within our computational constraints.

5.2 GP Classification

The dataset used in this experiment consists of high-dimensional features. To reduce computational complexity, PCA is applied to reduce the feature space to two principal components. The reduced dataset is then utilized as the input for the GP classification. A GP model with a combination of RBF and White kernel is employed. The results are shown in Figure 1.

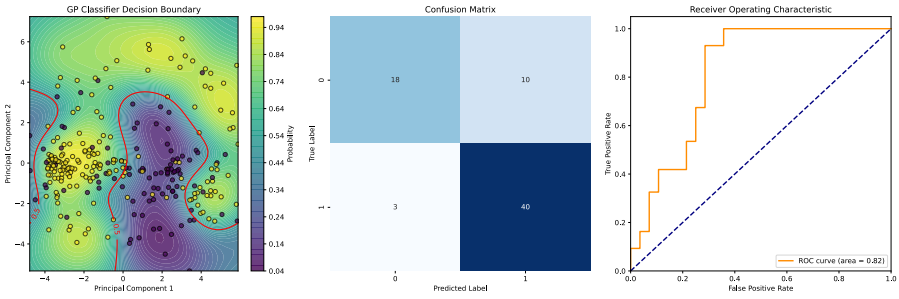

Figure 1: Gaussian Process Classifier Performance. Left: Decision boundary for the Gaussian Process classifier in PCA-reduced space, showing probability gradients and a p=0.5 contour, illustrating effective class separation. Center: Confusion matrix indicating correct predictions (diagonal) and errors, Right: ROC curve with an AUC of 0.82, demonstrating the classifier's good discriminative ability across thresholds.

Despite the GP classifier achieved a good discriminative ability, our experimental observations underscore the limitations inherent in traditional supervised learning models as datasets become increasingly complex, prompting a shift toward more adaptive, unsupervised learning methodologies.

5.3 GMM Clustering

we implement the EM algorithm for GMM to explore unsupervised learning capabilities in the dimensionally reduced space of MNIST, which was transformed via PCA to retain the top 50 principal components. The EM algorithm was initialized with ten Gaussian components, reflecting our hypothesis on the 10 digit categories.

Visual inspection of the clusters was facilitated by transforming the Gaussian means back to the original data space, allowing us to interpret the learned clusters in terms of recognizable digit images. Cluster centroids displayed characteristic features of the MNIST digits, confirming the model's ability to capture key aspects of the data structure. The results are shown in Figure 2.

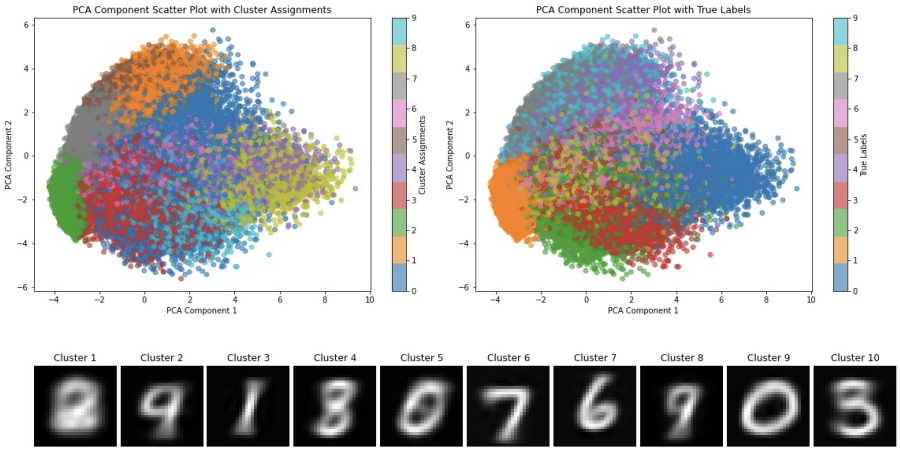

Figure 2: GMM Clustering on PCA-Reduced MNIST Dataset. Left: the PCA component scatter plot colored according to cluster assignments from GMM. Right: PCA scatter plot colored by the true labels . Bottom: centroid images for each cluster (Clusters 1 to 10), Cluster purity = 0.537

Despite achieving meaningful clustering, A Cluster purity of 0.537 shows the limited alignment of the model-generated clusters with the actual digit categories. Which indicates the model's reliance on the assumption of Gaussian-distributed data points restricts its ability to model more intricate distributions found in real-world datasets.

5.4 VAEs Clustering

We compare base VAE, Beta-VAE, HMCVAE and HISVAE with the same base model architecture as showed in Table 1. This setup aims to match the network structure in Kingma and Welling[3]. The number of latent dimensions is set to 2 across all models for a intuitive illustration of representation disengagement.

Table 1: Architecture for Variational Autoencoders

Layer	**Output Shape**	**Activation**
Input	784	-
Dense (Encoder)	512	ReLU
Dense (μ)	2	Linear
Dense ($\log \sigma^2$)	2	Linear
Reparameterize	2	-
Dense (Decoder)	512	ReLU
Dense (Output)	784	Sigmoid

Each model was trained using the Adam optimizer with a learning rate of 0.001. The training process continued for a predefined number of epochs or until convergence was observed based on the stability of the loss function (tolerance of 0.001). We also added early stopping by halting the training if there is no improvement in the loss on over 10 epochs.

To evaluate HMC and HISVAE after training, we estimate the out-of-sample ELBO which combines the BCE and the Kullback-Leibler (KL). This metric quantifies how well the model predicts unseen data, with both the accuracy of the reconstruction (through BCE) and the efficiency of the latent space encoding (through KL divergence). with lower values indicating better predictive performance. The reslut is shown in Table 2.

Table 2: Comparison of VAE models based on BCE, KL Divergence, and ELBO

Model	Total Epoch	Average BCE	Average KL Divergence	Average ELBO
VAE	36	142.22	6.03	-148.25
Beta-VAE	62	146.73	**4.17**	-150.90
HMCVAE	40	141.19	23.45	-164.64
HISVAE	113	**139.95**	19.75	**-139.94**

Notably, the HISVAE demonstrate significant improvements in ELBO and BCE. showning the intergation of HIS effectively balances the trade-offs between precision in reconstruction and the regularization imposed by the KL. This experiment underscore the potential of HISVAE in achieving high fidelity in data reconstruction while maintaining a principled approach to probabilistic latent representation learning. A intuitive Illustration of the latent space is presented in Figure 3. The reconstructions of the VAEs can be seen in Figure 4.

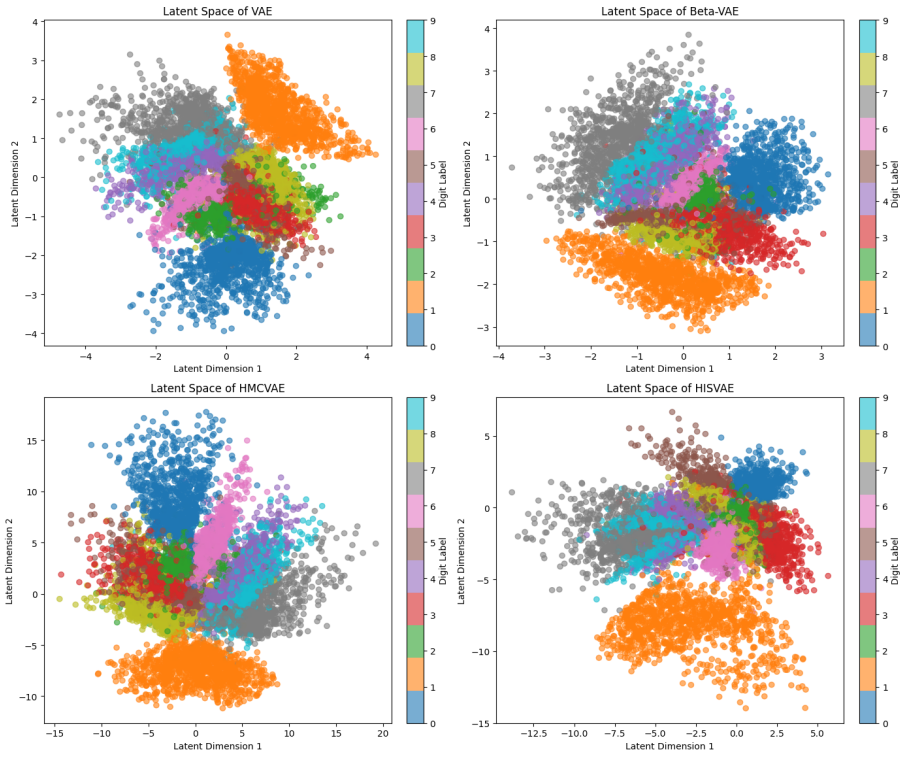

Figure 3: VAEs Clustering on MNIST. Top-Left: Latent Space of standard VAE. Top-Right: Latent Space of Beta-VAE. Bottom-Left: Latent Space of HMCVAE. Bottom-Right: Latent Space of HISVAE. Overall, all models demonstrated superior clustering compared to GMM, with HIS-based models achieving more refined clustering that captures detailed distinctions, such as overlapping curved-top '7s' with '9s'.

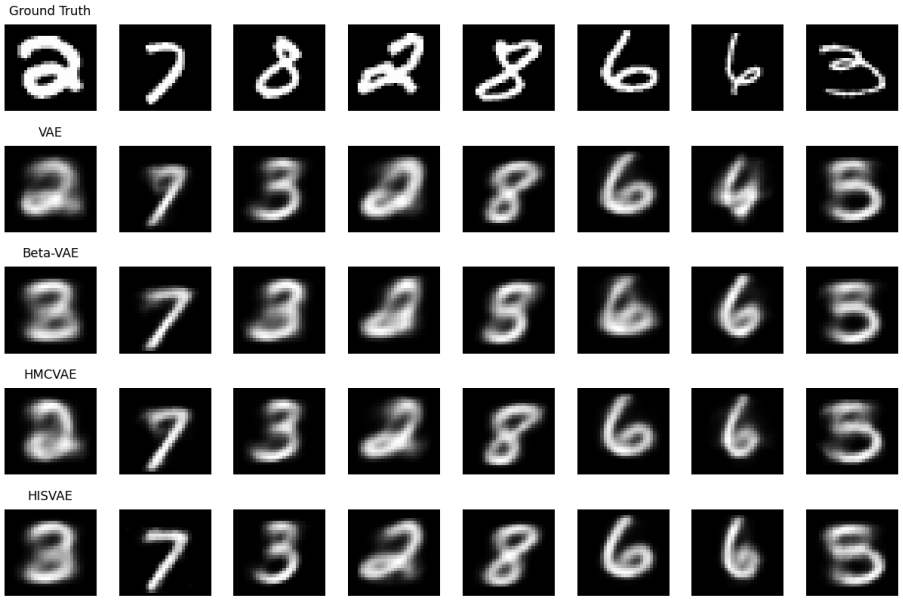

Figure 4: VAEs reconstruction. First Row: Ground truth. Second Row: Reconstruction of standard VAE. Third Row: Reconstruction of Beta-VAE, Fourth Row: Reconstruction of HMCVAE, Fifth Row: Reconstruction of HISVAE. All models capture the general shapes of the digits in the image, but with varying degrees of crispness. Hamiltonian MCMC models (HMCVAE and HISVAE) have an advantage in potentially achieving sharper reconstructions by incorporating dynamics from physics to explore the latent space more effectively.

6 Conclusion and Discussion

In this study we traced the evolution of machine learning models from GP to GMM and onto more sophisticated frameworks such as VAEs. Began with GPs, which proved effective in supervised settings with moderate-dimensional data, as demonstrated on the Ionosphere dataset, though the model's performance was constrained by the dimensionality and complexity of the data. Transitioning to unsupervised learning, GMMs were employed to explore deeper latent structures. Its performance was hindered by the assumption of Gaussian-distributed components, which is often too restrictive for complex real-world datasets, as seen in the MNIST clustering task.

To address these issues, we integrated Hamiltonian MCMC within the VAE framework, creating the Hamiltonian Monte Carlo Variational Autoencoder (HMCVAE). This model improved the exploration of the latent space by utilizing physical dynamics, which was beneficial for drawing samples that more accurately represented the target posterior. Despite these enhancements, the computational cost and complexity of implementing HMC dynamics remained a concern.

Further refinement led to the development of the HISVAE. The HISVAE incorporated importance sampling into the HMC framework, adjusting sample weights based on the Hamiltonian's changes. This approach significantly improved the fidelity of ELBO estimation, allowing for more precise and robust modeling of complex data distributions, it allows unbiaded estimation of ELBO as HISVAE allows the use of the reparameterization trick. The results demonstrated that HISVAE not only achieved better performance in terms of ELBO but also enhanced the interpretability and effectiveness of clustering on the MNIST dataset.

6.1 Limitations

However, HISVAE, like other VAEs, often makes assumptions about the geometry of the posterior distribution, typically approximating it as Gaussian. While the introduction of HMC helps in exploring more complex distributions, there might be mismatches between the model's assumptions and the true underlying distributions, especially in cases where the data exhibits multi-modal characteristics or other complex statistical properties[20].

6.2 Applications and Future Directions

Integrating HISVAE with Generative Adversarial Networks (GANs). The integration of HISVAE with GANs can enhance the stability and diversity of the generative process[21]. Specifically, the Hamiltonian dynamics in HISVAE facilitate a more thorough exploration of the latent space, potentially addressing common challenges in GANs such as mode collapse and the vanishing gradient problem[22]. The importance sampling mechanism of HISVAE improves the efficiency of sampling by weighting the contributions of different samples based on their dynamical properties, which can lead to higher quality generations. This capability is especially beneficial in applications where the diversity and accuracy of generated samples are critical, such as in high-resolution image generation or complex scenario simulations in gaming and simulations.

In the context of theoretical neuroscience, particularly in relation to the free energy principle and active inference, HISVAE provides a sophisticated computational framework. The free energy principle suggests that biological systems minimize a bound on the surprise represented by their sensory inputs, akin to the minimization of the Evidence Lower Bound (ELBO) in variational autoencoders[23]. HISVAE optimizes this bound through Hamiltonian dynamics coupled with importance sampling, offering a robust model for understanding how neural processes optimize internal states to effectively minimize free energy. Extending to active inference, HISVAE is adept at modeling both perceptual and action-based decision-making processes. It simulates how an agent might balance exploratory behavior (via Hamiltonian dynamics) with the exploitation of known strategies (through importance sampling adjustments)[24]. It opens up pathways for understanding the workings of the human brain. Future researche can design artificial systems that mimic certain aspects of human cognition, such as learning, adaptation, and even consciousness. These systems could potentially operate under similar principles to the human brain, optimizing their internal states and actions based on predictions and sensory feedback[25, 15].

6.3 Conclusion

In conclusion, this study has detailed the progression of machine learning models from Gaussian Processes and Gaussian Mixture Models to more advanced frameworks like Variational Autoencoders, culminating in the development of Hamiltonian Importance Sampling Variational Autoencoders (HISVAE). Each step in this evolution addressed specific limitations of the predecessors—ranging from the handling of data complexity to the exploration of latent spaces. HISVAE, in particular, has demonstrated notable improvements in the fidelity of ELBO estimation and the interpretability of clustering, propelled by its innovative integration of Hamiltonian dynamics and importance sampling. Despite some remaining challenges related to model assumptions and computational complexity, the potential applications of HISVAE—from enhancing GANs to modeling cognitive processes in theoretical neuroscience—point to promising directions for future research. These developments not only push the boundaries of what machine learning models can achieve but also bridge computational methodologies to deep theoretical concepts, offering insights that could one day inform the creation of artificial systems that emulate human cognitive functions.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[4] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, H. Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *ArXiv*, abs/1512.09300, 2015.

[5] Cristian Meo and Pablo Lanillos. Multimodal vae active inference controller. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2693–2699. IEEE, 2021.

[6] Tim Salimans and Diederik P. Kingma. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2014.

[7] Christopher Wolf, Maximilian Karl, and Patrick van der Smagt. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.

[8] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[9] Radford M Neal. Hamiltonian importance sampling. In *talk presented at the Banff International Research Station (BIRS) workshop on Mathematical Issues in Molecular Dynamics*, 2005.

[10] Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008.

[11] Felipe Tobar, Thang D Bui, and Richard E Turner. Learning stationary time series using gaussian processes with nonparametric kernels. *Advances in neural information processing systems*, 28, 2015.

[12] Hyun-Chul Kim and Zoubin Ghahramani. Bayesian gaussian process classification with the em-ep algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959, 2006.

[13] Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.

[14] Naonori Ueda, Ryohei Nakano, Zoubin Ghahramani, and Geoffrey E Hinton. Split and merge em algorithm for improving gaussian mixture density estimates. *Journal of VLSI signal processing systems for signal, image and video technology*, 26:133–140, 2000.

[15] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[16] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

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

[18] Paul Adrien Maurice Dirac. Generalized hamiltonian dynamics. *Canadian journal of mathematics*, 2:129–148, 1950.

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

[20] Clément Chadebec, Clément Mantoux, and Stéphanie Allassonnière. Geometry-aware hamiltonian variational auto-encoder. *arXiv preprint arXiv:2010.11518*, 2020.

[21] Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch vae-gan: Generating diverse videos from a single sample. *Advances in Neural Information Processing Systems*, 33:16761–16772, 2020.

[22] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Improving gan with neighbors embedding and gradient matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5191–5198, 2019.

[23] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[24] Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

[25] Zafeirios Fountas, Noor Sajid, Pedro Mediano, and Karl Friston. Deep active inference agents using monte-carlo methods. *Advances in neural information processing systems*, 33:11662–11675, 2020.

Supplementary Materials

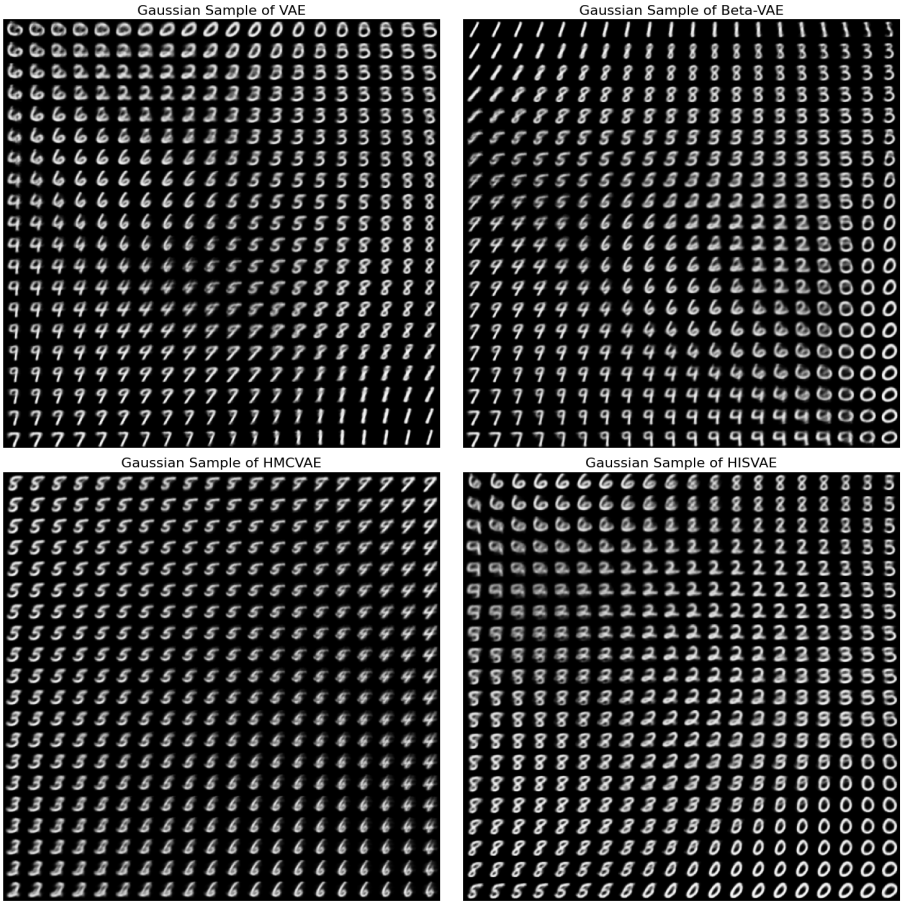

Figure 5: Gaussian Samples of VAE latent Space. Top-Left: Sample Space of standard VAE. Top-Right: Sample Space of Beta-VAE. Bottom-Left: Sample Space of HMCVAE. Bottom-Right: Sample Space of HISVAE. OverallObservations indicate that the HMCVAE and HISVAE produce clearer and more diverse digit representations compared to Hstandard VAE and Beta-VAE.