FOLDED HAMILTONIAN MONTE CARLO FOR BAYESIAN GENERATIVE ADVERSARIAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Networks (GANs) can learn complex distributions over images, audio, and data that are difficult to model. We deploy a Bayesian formulation for unsupervised and semi-supervised GAN learning. We propose Folded Hamiltonian Monte Carlo (F-HMC) within this framework to marginalise the weights of the generators and discriminators. The resulting approach improves the performance by having suitable entropy in generated candidates for generator and discriminators' weights. Our proposed model efficiently approximates the high dimensional data due to its parallel composition, increases the accuracy of generated samples and generates interpretable and diverse candidate samples. We have presented the analytical formulation as well as the mathematical proof of the F-HMC. The performance of our model in terms of autocorrelation of generated samples on converging to a high dimensional multi-modal dataset exhibits the effectiveness of the proposed solution. Experimental results on high-dimensional synthetic multi-modal data and natural image benchmarks, including CIFAR-10, SVHN and ImageNet, show that F-HMC outperforms the state-of-the-art methods in terms of test error rates, runtimes per epoch, inception score and Frechet Inception Distance scores.

1 INTRODUCTION

Generative adversarial networks (GANs) [Goodfellow et al. (2014)] have received traction in the field of deep generative models. The development of GANs covers a wide range from multi-layer perceptrons to the BigGAN framework [Brock et al. (2019)] with residual blocks and self-attention layers (Zhang et al. (2019)) to synthesise realistic images. Despite GAN's effectiveness in generating realistic images, it experiences mode collapse, which occurs when the generator over-optimises for a particular discriminator and the discriminator never learns how to escape the trap. Recent work has focused on alternative metrics such as f-diversities [Nowozin et al. (2016)] or Wasserstein divergences [Arjovsky et al. (2017)] to substitute the Jensen-Shannon divergence inherent in traditional GAN training to alleviating several practical issues.

Saatci & Wilson (2017) recently proposed Bayesian GAN, a probabilistic framework for GANs based on Bayesian inference. It demonstrates how modelling the distribution of generators alleviates mode collapse and motivates the interpretability of learned generators. GAN training measures the full posterior distribution across network weights in a single-mode based on mini-max optimisation. Even if the generator does not recall training instances, samples from the generator are expected to be excessively compact compared with data distribution samples. In fact, as demonstrated by Bayesian GAN, a posterior distribution over the generators' parameters can be vast and highly multi-modal to model the real data distribution by fully reflecting the posterior distribution over the generator and discriminator parameters. In addition, He et al. (2019) proposed the probGAN, which is similar to the Bayesian GAN, iteratively learns a distribution over generators but with a carefully crafted prior. Learning is triggered by a tailored Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) to perform Bayesian inference.

In both Bayesian GAN and ProbGAN, SGHMC is deployed to marginalised the parameters of the generator and discriminators. The foundation of our idea in this paper is to utilise different samples for generator parameters in order to mitigate GAN collapse mode and more efficiently produce data samples with an adequate degree of entropy, especially with high-dimensional and highly correlated

data. To do so, we propose a Folded Hamiltonian Monte Carlo (F-HMC) to replace the SGHMC part of the Bayesian framework. This proposal shares several desired properties with SGHMC, such as 1) being experimentally well-adjusted in training the GAN due to its Hamiltonian dynamics and 2) directly importing parameters such as the learning rate from gradient descent into the sampler. Furthermore, it benefits from the following advantages:

1. F-HMC explores more accurately the target density, especially in the scenario with high-dimensional and highly correlated data.
2. F-HMC converges faster to the target density in terms of lag number.
3. F-HMC provides the practical advantage to the Bayesian GAN method by exploring a rich multi-modal distribution over the weight parameters of generators at an acceptable entropy level.
4. More importantly, because of the parallel composition, F-HMC has an efficient run time in high-dimensional data approximation.

Our main contributions are listed below. We will return to them in the experiment section to highlight each one of them.

1. We propose F-HMC to sample parameters of generators to create candidates from the multi-modal high-dimensional distribution.
2. We have mathematically verified the functionality of F-HMC.
3. Empirical results on a high-dimensional multi-modal synthetic dataset show that the generated samples from our proposed method cover target distribution with more similarity to the target density concerning entropy value.
4. We have shown that using our proposed model, the semi-supervised learning algorithms on natural image datasets (ImageNet, SVHN, and CIFAR10) outperform the state-of-the-art in terms of inception scores (IS) and Frechet Inception Distance scores (FID).

The structure of the paper is as follow: We begin with the problem formulation in Section 2 and enhance the framework by introducing the proposed F-HMC in Section 4. We also have provided the theoretical analysis of F-HMC to generate samples from the desired target. Then, we have demonstrated the performance of our proposed model in Section 5. Finally, Section 6 concludes the paper and discusses future work.

2 Problem Formulation

Suppose having observed data $\mathscr{D} = \{x_i, y_i\}_{i=1}^N$ from an unknown probability distribution p_{data} where x describes the input and y denotes the corresponding label. We would like to estimate p_{data} which is a high-dimensional multi-modal distribution. Bayesian GAN [Saatci & Wilson (2017)] investigates the distributions over the weight parameters of the generators and creates distributions over an infinity space of generators and discriminators, corresponding to every conceivable setting of these weight vectors. We build upon the problem formulation in Bayesian GAN [Saatci & Wilson (2017)] and formulate the posterior as $p(y|f(x, \alpha))$ where $f(x, \alpha) = Gen(z, \hat{\alpha}_g)$. Here z represents white noise derived from $p(z)$, and $\hat{\alpha}_g$ represents distribution over generator parameters. We denote that parameter set α consisting of two sub parameter set $\hat{\alpha}_g$ related to the generator and $\hat{\alpha}_d$ associated with the discriminators.

We require the weight candidates of generators and discriminators to create candidates for posterior estimation. In this regard, we need to estimate posterior over $\hat{\alpha}_g$, $\hat{\alpha}_d$. Since generators and discriminators are performing min-max optimisation, the parameter over $\hat{\alpha}_g$ and $\hat{\alpha}_d$ are interdependent. First, generator weights $\hat{\alpha}_g$ are sampled from a prior $p(\hat{\alpha}_g|\beta_g)$, and a particular generative neural network is constructed conditioning on these samples. Then, white noise z derived from $p(z)$ is transformed through the network $Gen(z; \hat{\alpha}_g)$ to generate candidate data samples. Conversely, discriminator conditioned on its weights $Disc(.; \hat{\alpha}_d)$ produces the probability that these candidate samples are generated from the data distribution. This process can be stated as follow considering L as the likelihood:

$$p(\hat{\alpha}_g|z, \alpha_d) \propto exp\{L(Disc(Gen(z, \hat{\alpha}_g), \hat{\alpha}_d))\}p(\hat{\alpha}_g|\beta_g) \tag{1}$$

From the discriminator side, we need to form classification likelihood that classifies actual data from the generated samples and can be formulated as:

$$p(\hat{\alpha}_d|z, X, \alpha_g) \propto exp\{L(X, \hat{\alpha}_d)\} \times exp\{L(1 - Disc(Gen(z, \hat{\alpha}_g), \hat{\alpha}_d))\}p(\hat{\alpha}_d|\beta_d) \tag{2}$$

Here $p(\hat{\alpha}_d|\beta_d)$ indicates prior for $\hat{\alpha}_d$. By marginalising z from Equations 1 and 2, the equations get updated to $p(\hat{\alpha}_g|\hat{\alpha}_d) = \int p(\hat{\alpha}_g, z|\hat{\alpha}_d)dz$ and $p(\hat{\alpha}_d|\hat{\alpha}_g) = \int p(\hat{\alpha}_d|z, X, \hat{\alpha}_d)dz$. We can approximate the posterior over $\hat{\alpha}_g$ and $\hat{\alpha}_g$ by iteratively sampling from $p(\hat{\alpha}_g|\hat{\alpha}_d)$ and $p(\hat{\alpha}_d|\hat{\alpha}_g)$. Therefore we can have the corresponding generators and discriminators to generate candidate samples from the multi-modal high dimensional distribution (p_{data}). This paper proposes F-HMC as an efficient sampling strategy, especially when the target is high-dimensional and highly correlated.

3 BACKGROUND

This section briefly reviews data sampling basics, such as the Hamiltonian Monte Carlo sampler and Stochastic Hamiltonian Monte Carlo derived from [Chen et al. (2014)]

Suppose one wants to sample from the posterior distribution of X given a set of independent observations $x \in D$:

$$p(X|D) \propto \exp\left(-U(X)\right) \tag{3}$$

where the potential energy function U is given by

$$U = -\sum_{x \in D} \log p(x|X) - \log p(X) \tag{4}$$

Hamiltonian Monte Carlo (HMC) is a method for efficiently exploring the state space by proposing samples of X in a Metropolis-Hastings (MH) framework. By incorporating auxiliary momentum variables, V, these suggestions are generated from a Hamiltonian system. HMC creates samples from a joint distribution of (X, V) to sample from $p(X|D)$:

$$\pi(X, V) \propto \left(\exp(-U(X) - \frac{1}{2}V^T M^{-1} V\right) \tag{5}$$

The samples of X have a marginal distribution $p(X|D)$ if the resultant samples of V are simply discarded. Here M is a mass matrix that, coupled with V, defines a kinetic energy term. $H(X, V) = U(X) - \frac{1}{2}V^T M^{-1} V$ defines the Hamiltonian function. H intuitively calculates the total energy of a physical system by utilising position variables X and momentum variables V. To propose samples, HMC simulates Hamiltonian dynamics.

$$\begin{cases} dX = M^{-1}V dt \\ dV = -\nabla\, U(X) dt \end{cases} \tag{6}$$

Furthermore, SGHMC is based on the idea of combining stochastic optimisation with a first-order Langevin dynamic MCMC technique, demonstrating that adding the "right amount" of noise to stochastic gradient ascent iterates results in samples from the target posterior as the step size is annealed. SGHMC accomplishes this by including a "friction" term in the momentum update:

$$\begin{cases} d\text{X}= M^{-1}V dt \\ dV = -\nabla\, U(X)\, dt \text{ - } BM^{-1}V dt + \mathcal{N}(0, 2Bdt) \end{cases} \tag{7}$$

4 F-HMC MODEL

This section proposes F-HMC as an efficient and scalable sampling strategy, particularly when the target is high-dimensional and highly correlated (Contribution 1). Additionally, we have presented the mathematical analysis of F-HMC (Contribution 2) in section 4.2 to prove that the F-HMC samples from the equivalent target distribution specified in Equation 7.

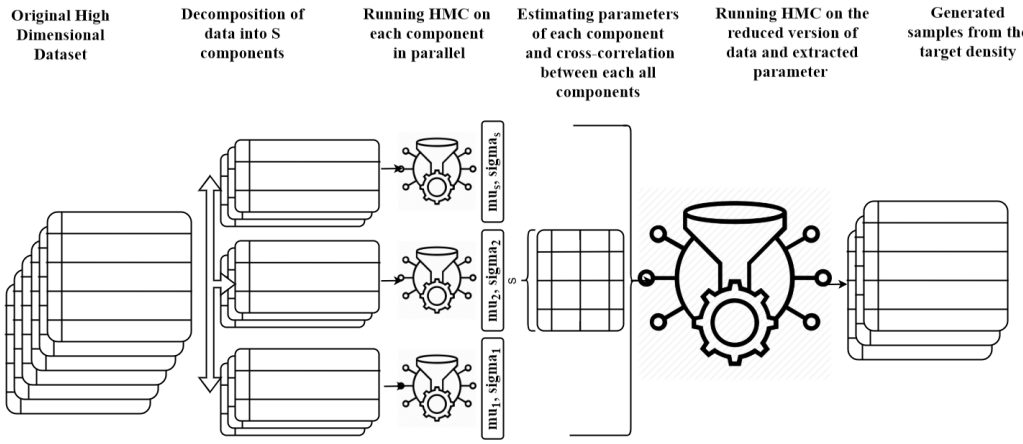

Figure 1: The pipeline of the F-HMC model which is consists of 4 stages. The first stage is decomposed into S components. The second stage is running the component in parallel and estimating their corresponding density parameter. The third stage is running another HMC on the reduced dimension data coming from the last stage with respect to the cross-correlation between components and finally generating data from the target distribution.

4.1 METHODOLOGY

In the scenario where the gradient elements are on dramatically different scales and highly correlated, the SGHMC cannot efficiently explore the target density. As a pragmatic approach, we propose F-HMC, a parallelised algorithm that uses parameter decomposition to divide the updating tasks into blocks. In particular, F-HMC decomposes the data into S components. These components run in parallel using an SGHMC model fit to their data distribution using the Gaussian likelihoods with a Laplacian prior, in order to finding the μ and σ as estimation parameters of the distribution of each component. Then, all the μs and σs that manifest lower-dimensional representation of original data (due to the decomposition) are given to another HMC (fold) with respect to the cross-correlation between all components.

Figure 1 demonstrates the functionality of the F-HMC; since the first part is running on parallel and the second fold is running on a reduced number of dimensions, the overall execution time, especially on high dimensional setup, improves. Furthermore, because two HMC samplers thoroughly examine the data, the system's entropy will be satisfactory.

Algorithm 1 shows one iteration of the Bayesian learning for the generator parameters using our proposed model. Here θ is the friction term for HMC, and η is the learning rate. K shows the number of iterations in Bayesian GAN, and I shows the number of F-HMC iterations, and S indicates the number of components in the F-HMC.

4.2 THEORETICAL ANALYSIS

This section aims to prove that the probability density of generated data using F-HMC is equivalent to the target distribution of SGHMC. Therefore we can verify that samples from F-HMC are mathematically valid (Contibution 2).

Considering earlier setup outlined in Background section 3, Equation 7 converges to a stationary distribution [Chen et al. (2014)] given by the following formula:

$$V \sim \mathcal{N}(\underbrace{MB^{-1}\nabla U(X)}_{\hat{\mu}}, \overbrace{M}^{\sigma}) \tag{8}$$

Algorithm 1 One iteration of the generator in Bayesian GAN set up with our proposed F-HMC

Input: $\{\hat{\alpha}_g^{i,k}\}$ and $\{\hat{\alpha}_d^{i,k}\}$ from preceding iteration and number of θ as HMC friction term, η for the learning rate, and I as the number of F-HMC iterations.
for each $i \leftarrow 1$ to I **do**
  $z \sim p(z)$ ▷ sampling white noise z from its prior
  for each $s \leftarrow 1$ to S **do** ▷ running the decomposition
    $\mu_s \leftarrow HMC(logp(\hat{\alpha}_g^{i,k,s})|z, \hat{\alpha}_d^s)$
    append μ_s to the μ
  end for each
  cov $\leftarrow$ calculate covariance between S components
  $q \leftarrow HMC$ (μ, cov) ▷ merging all back to sample from the target density
end for each
$n \sim \mathcal{N}(0, 2\theta\eta I)$
$v \leftarrow (1-\theta)v + \eta q + n$
$\hat{\alpha}_g^{i,k} \leftarrow \hat{\alpha}_g^{i,k} + v$ ▷ updating $p(\hat{\alpha}_g)$ sample set
Output: generated samples for $\hat{\alpha}_g$

By taking the F-HMC methodology into account, the V in Equation 7 can be decomposed as the following matrix:

$$\begin{bmatrix} V_1 \\ V_2 \\ . \\ . \\ V_S \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{\mu_1} \\ \hat{\mu_2} \\ . \\ . \\ \hat{\mu_S} \end{bmatrix}, \begin{bmatrix} \sigma_{11} & . & . & \sigma_{1S} \\ . & . & . & . \\ . & . & . & . \\ \sigma_{S1} & . & . & \sigma_{SS} \end{bmatrix} \right) \tag{9}$$

Theorem: $\pi(X, V) \propto \exp(-H(X, V))$ *is the unique stationary distribution of the dynamics described in Equation 9.*

Proof: let $W = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$ and $R = \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix}$, Equation 7 along with the proposed decomposition in Equation 9 can be written in the following form:

$$\begin{aligned} d\begin{bmatrix} X \\ V \end{bmatrix} &= d\begin{bmatrix} X_1 & X_2 & . & . & X_S \\ V_1 & V_2 & . & . & V_S \end{bmatrix} = \\ &- \begin{bmatrix} 0 & -I \\ I & B \end{bmatrix} \begin{bmatrix} \nabla U(X_1) & \nabla U(X_2) & . & . & \nabla U(X_S) \\ M^{-1}V_1 & M^{-1}V_2 & . & . & M^{-1}V_S \end{bmatrix} + \mathcal{N}(0, 2Rdt) \\ &= -[W + R]\nabla H(X, V)dt + \mathcal{N}(0, 2Rdt) \end{aligned} \tag{10}$$

We employ the Fokker-Planck Equation (FPE) to describe the temporal evolution of the probability density function in relation with a Stochastic Differential Equation (SDE) that defines the development of the distribution on the random variable under specified stochastic dynamics. Under Hamiltonian dynamics, the random variables in our situation are position variable X and momentum variable V. Consider the following SDE:

$$d\omega = g(\omega)dt + \mathcal{N}(0, 2R(\omega)dt) \tag{11}$$

$\rho_t(\omega)$ is the distribution of ω governed by Eq. 11. we also Consider $J_i(.)$ as $\partial_i(.)$ by using FPE the equation evolves under the following formula:

$$J(\rho_t(\omega)) = -\sum_{i=1}^{n} J_{\omega_i}(g_i(\omega)\rho_t(\omega)) + \sum_{i=1}^{n}\sum_{j=1}^{n} J_{\omega_i}(J_{\omega_j}(R_{ij}(\omega)\rho_t(\omega))) \tag{12}$$

We can rewrite Equation 12 in the following compact form:

$$J(\rho(\omega)) = -\nabla^T[g(\omega)\rho_t(\omega)] + \nabla^T R(\omega)\nabla\rho_t(\omega)] \tag{13}$$

where $\nabla^T[g(\omega)\rho_t(\omega)] = \sum_{i=1}^{n} J_{\omega_i}(g_i(\omega)\rho_t(\omega))$, and $\nabla^T[R(\omega)\nabla\rho_t(\omega)] = \sum_{ij} J_{\omega_i}(R_{ij}(\omega)J_{\omega_j}(\rho_t(\omega)) = \sum_{ij} J_{\omega_i}(R_{ij}(\omega)J_{\omega_j}\rho_t(\omega)) + \sum_{ij} J_{\omega_i}(J_{\omega_j}(R_{ij}(\omega))\rho_t(\omega)) = \sum_{ij} J_{\omega_i}(J_{\omega_j}(R_{ij}(\omega)\rho_t(\omega)))$.

From Equation 13 and considering the variable $\omega = (X, V)$ and $g(\omega) = -[R + W]\nabla H(X, V)$ and $R(\omega) = R(X, V) = R = \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix}$, the distribution evolution under dynamic system in Equation 10 can be written as follow:

$$J(\rho_t(X, V)) = \nabla^T\{[R + W][\rho_t(X, V)\nabla H(X, V) + \nabla\rho(X, V)]\} \tag{14}$$

Since $\pi(X, V) \propto \exp(-H(X, V))$, we can verify that the $\pi(X, V)$ is invariant under Equation 14 by calculating $[e^{-H(X,V)}\nabla H(X, V) + \nabla e^{-H(X,V)}] = 0$. Therefore, we can confirm that the dynamics given in Equation 9 have similar invariance properties to that of the original Hamiltonian dynamics of Equation 7.

5 Experiments

We implemented the proposed model using Pymc3 [1] and report its performance on generating samples from complex distributions in section 5.1. We have examined the model's performance in marginalising the generators' parameters on synthetic and natural image datasets such as SVHN [Netzer et al. (2011)], CIFAR 10 [Krizhevsky (2009)] and ImageNet [Deng et al. (2009)] in section 5.2 and 5.3, respectively. We have compared our results with WDCGAN [Arjovsky et al. (2017)], DCGAN, 10DCGAN (which is a fully supervised convolutional neural network composed of ten DCGANs constructed by ten random subsets with 80% of the size of the training set, [Radford et al. (2016)]), Bayesian GAN [Saatci & Wilson (2017)] and ProbGAN [He et al. (2019)] on supervised and semi-supervised tasks with four different numbers of labelled examples. For a fair comparison, each model has the same number of generators and discriminator with the same architecture.

5.1 F-HMC performance

To evaluate the performance of F-HMC and compare it against SGHMC, we have designed a set of experiments. First, we use Normalising Flows [Kobyzev et al. (2020); Rezende & Mohamed (2016)] as a rich family of distributions to examine F-HMC and SGHMC's abilities to explore complex distribution. Figure 2 shows the potential energy of the rich target distribution and the generated candidates using SGHMC and F-HMC. F-HMC successfully covers the target distribution (Advantage 1). Second, we have measured auto-correlation between the samples generated in each sampler as a metric to show the power of the sampler in exploring the target distribution. The more precise the sampler, the faster the auto-correlation reaches zero. Figure 3 shows auto-correlation in F-HMC drops faster to zero than SGHMC in terms of lag number (Advantage 2).

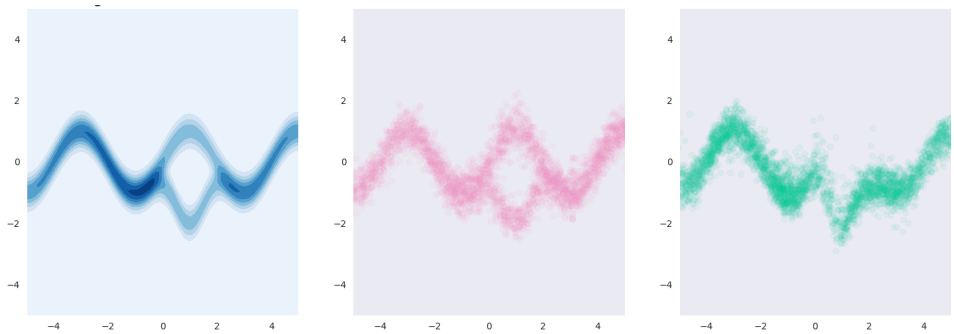

Figure 2: The graph on the left shows the potential distribution of the target. The middle graph shows F-HMC samples exploring the target, and the right graph shows the SGHMC samples. Both samplers converged to the target distribution, but the F-HMC covered the target more accurately than SGHMC.

[1] https://docs.pymc.io/

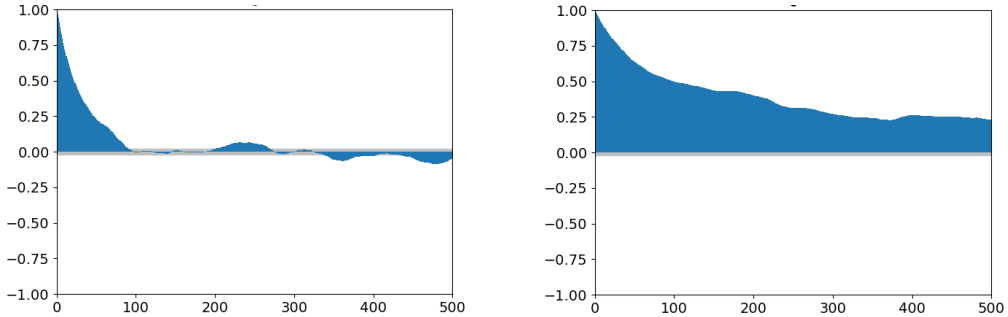

Figure 3: Auto-correlation between samples while the samplers explore the target distribution with 100 dimensions. The left graph show auto-correlation of samples using the F-HMC sampler while S is 25. The right graph shows the auto-correlation on the same setup using SGHMC. F-HMC converges faster than SGHMC to the target in terms of lag number

5.2 HIGH-DIMENSIONAL MULTI MODAL SYNTHETIC DATASET

We present experiments on a multi-modal synthetic dataset to test inferring a multi-modal posterior $p(\hat{\alpha}_g|D)$. This experiment shows F-HMC's ability to explore a set of generators' parameters with proper entropy and different complementary properties to encapsulate a rich data distribution. We fit a regular GAN, Bayesian GAN, and our proposed model to a dataset with D = 100 and 500. The generator for all models is a two-layer neural network: 10-1000-100, fully connected, with ReLU activation. The red samples in Figure 4 depict the target data, whereas the green samples depict the corresponding generated data. The experiments on D = 100 are shown in the first two rows, while the results on D = 500 are shown on two lower rows. The name of generating sampler is displayed on the right side of each row. We can ensure that both strategies (BGAN and F-HMC GAN) cover the intended distribution simply by comparing them visually. Even the visual comparison is insufficient to detect more outstanding performance in D = 100. Still, it is evident in D = 500 that F-HMC provides a more desirable match to the target distribution as the dimension increases (Advantage 3 and Contribution 3). Figure 5 shows the comparison of the performance of GAN, F-HMC GAN, and Bayesian GAN in terms of Jensen-Shannon divergence. The experiment estimates the similarity of the probability distribution of generated data to the original data and the level of entropy and confirms that The F-HMC exceeds other models (Advantage 3 and Contribution 3).

5.3 NATURAL IMAGE DATASET

Table 1: Supervised and semi-supervised learning results for image benchmarks. the N_s shows number of labelled examples.

N_s	Supervised	DCGAN10	W-DCGAN	BayesGAN	probGAN	F-HMC GAN
CIFAR-10						
500	65.1 ± 2.3	30.9 ± 2.7	55.8 ± 2.9	30.5 ± 2.3	30.1 ± 2.8	$\mathbf{30.0 \pm 3.1}$
1000	54.6 ± 2.1	29.1 ± 2.4	48.8 ± 3.2	$\mathbf{27.4 \pm 2.1}$	27.7 ± 3.1	27.6 ± 2.8
2000	52.4 ± 2.4	26.8 ± 3.3	37.9 ± 2.5	24.2 ± 1.9	28.3 ± 2.5	$\mathbf{23.9 \pm 2.3}$
4000	48.1 ± 1.0	24.7 ± 2.7	28.2 ± 2.9	22.3 ± 3.2	21.7 ± 2.8	$\mathbf{20.7 \pm 2.6}$
SVHN						
1000	55.1 ± 3.3	30.8 ± 2.3	30.1 ± 1.9	28.7 ± 3.1	$\mathbf{26.4 \pm 2.1}$	26.6 ± 2.2
2000	36.7 ± 2.63	17.9 ± 1.7	27.2 ± 2.6	14.2 ± 2.8	14.1 ± 2.6	$\mathbf{13.7 \pm 1.8}$
4000	28.2 ± 3.13	15.8 ± 1.4	25.1 ± 2.8	12.7 ± 2.9	13.5 ± 1.7	$\mathbf{11.7 \pm 1.4}$
8000	21.1 ± 2.2	15.1 ± 1.3	20.1 ± 1.9	9.2 ± 1.8	11.4 ± 1.8	$\mathbf{8.9 \pm 0.9}$
ImageNet						
1000	57.6 ± 4.2	53.4 ± 3.1	55.7 ± 3.7	48.9 ± 4.3	47.8 ± 4.6	$\mathbf{43.8 \pm 4.4}$
2000	42.3 ± 3.5	38.7 ± 2.5	40.6 ± 3.1	34.6 ± 4.6	34.5 ± 3.8	$\mathbf{33.6 \pm 3.7}$
4000	40.1 ± 3.6	31.8 ± 2.1	35.5 ± 2.9	27.8 ± 3.8	25.8 ± 3.2	$\mathbf{25.9 \pm 3.5}$
8000	36.8 ± 4.1	28.3 ± 1.8	34.3 ± 2.7	24.4 ± 3.1	24.1 ± 2.8	$\mathbf{22.7 \pm 2.7}$

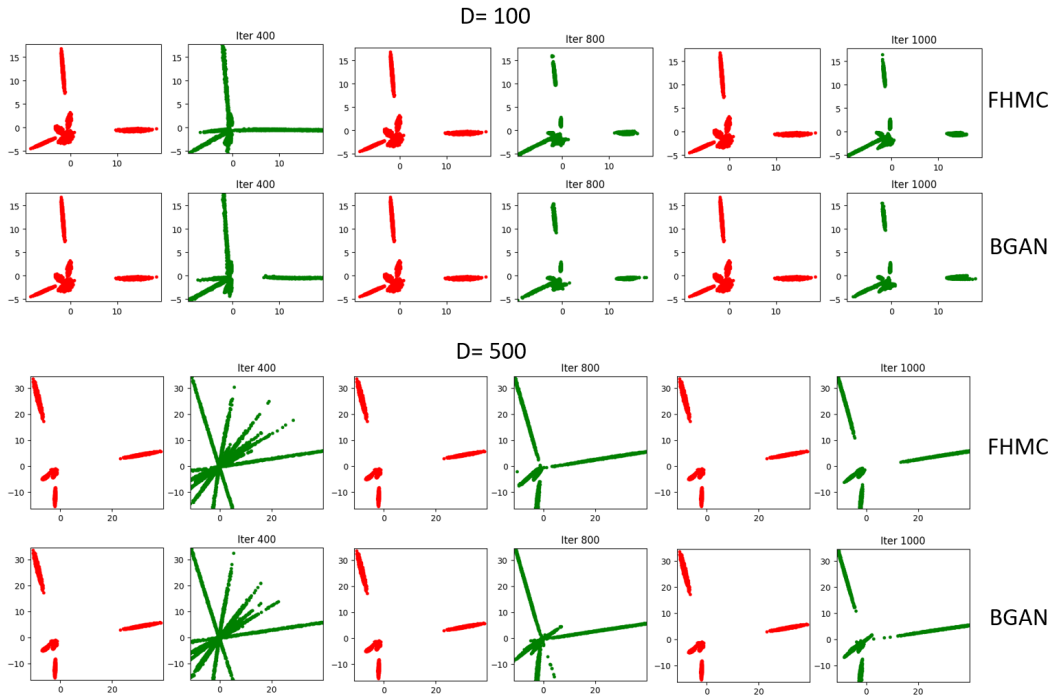

Figure 4: samples drawn from $p_{data}(x)$ and visualised in 2D. the red colour graph shows real data, and the green colour graph shows generated samples. The first two upper rows show the experiment on D=100, and the two lower rows show D=500. The name of generating sampler is shown on the right side of each row. The graphs generated by F-HMC have more visual similarity to the actual target, especially when the D = 500.

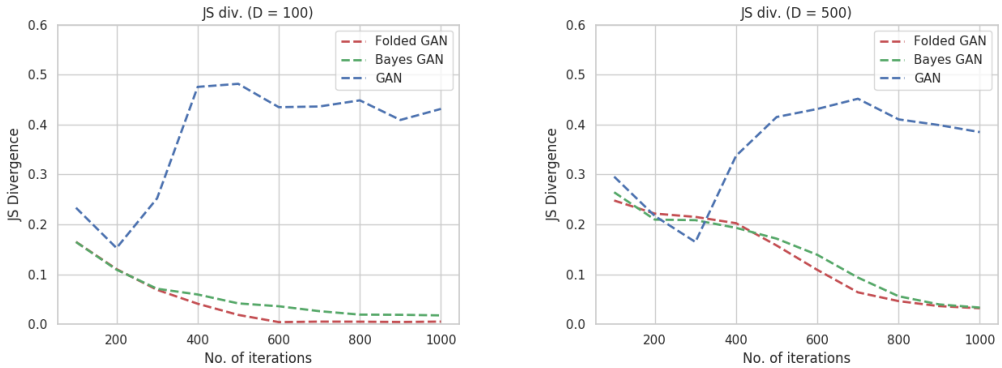

Figure 5: The Jensen-Shannon divergence between $p_{data}(x)$ and the number of iteration of model training. The left graph shows the experiment while D=100, and the right graph shows the same experiment with D = 500. We can confirm the superiority of F-HMC over other models in generating data more similar to the target concerning entropy.

We used a 5-layer network architecture for GAN's generator in all experiments on the natural images datasets. The corresponding discriminator for supervised GAN is a 5-layer 2-class DCGAN, and we have used a 5-layer, K + 1 class DCGAN for a semi-supervised GAN performing classification over K classes (see Saatci & Wilson (2017) for further details about Bayesian GAN structure)). To evaluate the performance of our proposed model, we have employed experiments in three measurement levels: 1-performance metric in supervised and semi-supervised learning using test error rate. 2-run

time per epoch in minutes by running all the models on a single GPU. 3-quality of generated images in terms of IS and FID scores. FID [Heusel et al. (2017)] is a measure that calculates the distance between vectors derived for real and synthetic images. IS [Salimans et al. (2016)] is an objective measure for assessing the quality and diversity of generated images.

Table 2: IS (higher is better), FID (lower is better) both trained with WGAN objective and run time (epochs in minutes) results on natural images datasets.

Dataset	Score	10DCGAN	BayesGAN	probGAN	F-HMC GAN
CIFAR10	IS	7.78	7.69	7.72	**7.79**
	FID	23.81	24.75	24.63	**23.73**
	Runtime	143	**91**	94	93
SVHN	IS	8.34	8.27	8.19	**8.31**
	FID	49.61	51.78	52.32	**47.21**
	Runtime	151	98	**89**	94
ImageNet	IS	8.41	8.51	8.56	**8.59**
	FID	30.2	29.78	28.12	**27.83**
	Runtime	671	358	349	**336**

Table 1 demonstrates supervised and semi-supervised learning results for all image benchmarks. Our proposed model mainly outperforms BayesGAN, probGAN, W-DCGAN, and 10-DCGAN in terms of test error rate (Contribution 4). F-HMC shows its substantial impact when running on higher-dimensional data (ImageNet) due to the parallel composition of F-HMC; it can efficiently explore higher dimension data.

Table 2 shows the generated images' quality and the run time of the models. The quality of images increases by using F-HMC. It is observed that the run time enhances when running the model on a higher dimension. We perceive that in lower-dimensional data (CIFAR, SVHN), the Bayesian GAN and probGAN's run time is more satisfying. Once we run the models on higher-dimensional data (ImageNet), the run time improves in F-HMC. The parallel composition of F-HMC makes the over epoch run time less than exploring the whole dimensions at once in Bayesian GAN and probGAN (Advantage 4).

6 Conclusion

Folded Hamiltonian Monte Carlo is presented in this paper as a scalable strategy in sampling high-dimensional, highly correlated data to improve Bayseain GAN in producing synthetic images/generating data by marginalising the weights of the generators and discriminators. We demonstrated that F-HMC converges faster and adapts to higher-dimensional inputs with more significant similarity to the target data. The theoretical and mathematical analysis of F-HMC is presented which confirms its functionality. F-HMC outperforms the state-of-the-art in terms of test error rates, runtimes per epoch, IS, and FID scores when evaluated on synthetic high-dimensional multi-modal data and natural image benchmarks, such as CIFAR-10, SVHN, and ImageNet. Despite the notable improvement that F-HMC can bring in sampling target density, its hyperparameters affect its enforcement. Nested sampling methods [Betancourt et al. (2011)] sample from the likelihood space instead of sample space and are more likely to land on a better set of parameters. Future directions include the adoption of a nested sampling method, e.g. restricted Hamiltonian Monte Carlo methods, to help the improvement of the outcomes.

7 Reproducibility

The experiments presented in the paper are designed to be reproducible and easy to extend. The notebook provides instructions on how to install the packages as well as the code required to run the experiments and the link to download the necessary datasets.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Michael Betancourt, Ali Mohammad-Djafari, Jean-Francois Bercher, and Pierre Bessiere. Nested sampling with constrained hamiltonian monte carlo. 2011. doi: 10.1063/1.3573613. URL http://dx.doi.org/10.1063/1.3573613.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. *31st International Conference on Machine Learning, ICML 2014*, 5, 02 2014.

Jia Deng, R. Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, volume 00, pp. 248–255, 06 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/abstract/document/5206848/.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014. doi: 10.1145/3422622.

Hao He, Hao Wang, Guang-He Lee, and Yonglong Tian. Bayesian modelling and monte carlo inference for GAN. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1l7bnR5Ym.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pp. 6629–6640, 2017.

Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020. ISSN 1939-3539. doi: 10.1109/tpami.2020.2992934. URL http://dx.doi.org/10.1109/TPAMI.2020.2992934.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yuval Netzer, T. Wang, A. Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. F-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 271–279, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.

Yunus Saatci and A. G. Wilson. Bayesian gan. In *NIPS*, 2017.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7354–7363. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhang19d.html.

A APPENDIX

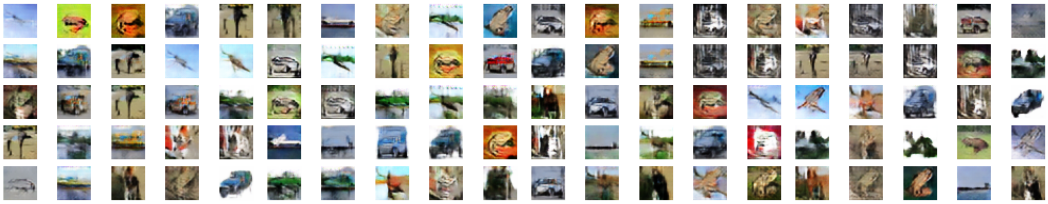

Figure 6: Data samples for the CIFAR10 generated using F-HMC strategy

Figure 7: Data samples for the SVHN generated using F-HMC strategy