Are Particle-Based Methods the Future of Sampling in Joint Energy Models? A Deep Dive into SVGD and SGLD

Vedant Rajiv Shah

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Ismini Lourentzou, Chair
Pinar Yanardag
Chris L Thomas

July 29th, 2024
Blacksburg, Virginia

Keywords: Stein Variational Gradient Descent, Joint Energy Models, Stochastic Gradient Langevin Dynamics, Wide Residual Networks, Energy Based Models

Are Particle-Based Methods the Future of Sampling in Joint Energy Models? A Deep Dive into SVGD and SGLD

Vedant Rajiv Shah

ABSTRACT

This thesis investigates the integration of Stein Variational Gradient Descent (SVGD) with Joint Energy Models (JEMs), comparing its performance to Stochastic Gradient Langevin Dynamics (SGLD). We incorporated a generative loss term with an entropy component to enhance diversity and a smoothing factor to mitigate numerical instability issues commonly associated with the energy function in energy-based models. Experiments on the CIFAR-10 dataset demonstrate that SGLD, particularly with Sharpness-Aware Minimization (SAM), outperforms SVGD in classification accuracy. However, SVGD without SAM, despite its lower classification accuracy, exhibits lower calibration error underscoring its potential for developing well-calibrated classifiers required in safety-critical applications. Our results emphasize the importance of adaptive tuning of the SVGD smoothing factor (α) to balance generative and classification objectives. This thesis highlights the trade-offs between computational cost and performance, with SVGD demanding significant resources. Our findings stress the need for adaptive scaling and robust optimization techniques to enhance the stability and efficacy of JEMs. This thesis lays the groundwork for exploring more efficient and robust sampling techniques within the JEM framework, offering insights into the integration of SVGD with JEMs.

Are Particle-Based Methods the Future of Sampling in Joint Energy Models? A Deep Dive into SVGD and SGLD

Vedant Rajiv Shah

GENERAL AUDIENCE ABSTRACT

This thesis explores advanced techniques for improving machine learning models with a focus on developing well-calibrated and robust classifiers. We concentrated on two methods, Stein Variational Gradient Descent (SVGD) and Stochastic Gradient Langevin Dynamics (SGLD), to evaluate their effectiveness in enhancing classification accuracy and reliability. Our research introduced a new mathematical approach to improve the stability and performance of Joint Energy Models (JEMs). By leveraging the generative capabilities of SVGD, the model is guided to learn better data representations, which are crucial for robust classification. Using the CIFAR-10 image dataset, we confirmed prior research indicating that SGLD, particularly when combined with an optimization method called Sharpness-Aware Minimization (SAM), delivered the best results in terms of accuracy and stability. Notably, SVGD without SAM, despite yielding slightly lower classification accuracy, exhibited significantly lower calibration error, making it particularly valuable for safety-critical applications. However, SVGD required careful tuning of hyperparameters and substantial computational resources. This study lays the groundwork for future efforts to enhance the efficiency and reliability of these advanced sampling techniques, with the overarching goal of improving classifier calibration and robustness with JEMs.

To my family and friends, whose constant support and belief in me have made this journey possible. I dedicate this work to you with deep appreciation and affection.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my research advisor, Dr. Ismini Lourentzou, for her invaluable guidance, insightful edits, and unwavering support throughout the development of this thesis. Her expertise and mentorship have been instrumental in shaping both my research and writing skills. I am also deeply grateful to Dr. Safa for her pioneering work with Stein Variational Gradient Descent (SVGD) as a sampler, which inspired my thesis. Her guidance in developing my understanding and implementation of the SVGD sampler was pivotal to the success of my research.

I would like to extend my appreciation to my committee members Dr. Pinar Yanardag and Dr. Chris Thomas who evaluated my work during the defense of my thesis. I am also thankful to Virginia Tech for providing me with the necessary resources, teaching experience, and research opportunities. The support from my institution played a significant role in my academic and professional development.

To my family and friends, your unwavering support and encouragement have been essential to the completion of my thesis. Your belief in my abilities and your patience throughout this journey have been my pillars of strength.

Thank you all for your contributions, support, and encouragement.

Contents

List of Figures

List of Tables

List of Abbreviations

EBM Energy Based Model

GAN Generative Adversarial Networks

GB Giga Bytes

JEM Joint Energy Model

MCMC Markov Chain Monte Carlo

RBF Radial Basis Function

SAM Sharpness Aware Minimization

SGLD Stochastic Gradient Langevin Dynamics

SVGD Stein Variational Gradient Descent

VAE Variational AuteEncoders

WRN Wide Residual Network

Chapter 1

Introduction

The exploration and advancement of perception in machine learning have been significantly driven by the quest to develop models capable of understanding and generating complex data representations. Among these, Energy-Based Models (EBMs) have garnered attention for their unique approach to representing probabilistic distributions [1, 2]. By defining an energy function over the space of inputs and outputs, EBMs offer a versatile and flexible approach to model a broad spectrum of tasks ranging from various applications in reinforcement learning and continual learning to image generation and compositional visual generation [2, 3, 4, 5, 6]. Despite their theoretical charm, EBMs face practical challenges that have limited their adoption, particularly in comparison to their discriminative and generative counterparts such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) [7]. These challenges are predominantly associated with the efficiency and stability of sampling techniques, stability of training procedures, and robustness of the models under various conditions [2, 8].

Joint Energy Models (JEM), a notable advancement in the realm of EBMs, are hybrid models that have shown promising results in unifying the capabilities of discriminative learning for classification and generative learning for data augmentation in a single network, unlike the adversarial component of GANs, by leveraging information sharing [9, 10]. JEMs have set a new benchmark for the potential of EBMs by capitalizing on their inherent generative nature yet maintaining their competitive classification performance. However, JEMs reliance

on Markov Chain Monte Carlo (MCMC) techniques such as Stochastic Gradient Langevin Dynamics (SGLD) for sampling, introduces limitations in terms of convergence rate, sample diversity, and computational efficiency which adversely affect both the model performance as well as its confidence [8].

In this thesis, we explore the potential of particle-based sampling techniques, specifically Stein Variational Gradient Descent (SVGD), as an alternative to traditional MCMC methods like SGLD. Theoretically, SVGD stands out for its efficiency in generating diverse samples through a deterministic process that iteratively updates particles to mimic the target distribution. However, our empirical findings indicate that SVGD does not surpass SGLD in performance within the JEM framework. Despite this, SVGD's theoretical properties and practical advantages warrant a detailed investigation into its application and limitations.

Our contributions are as follows:

1. **Integration of SVGD with JEMs for Improved Calibration**: We have integrated the SVGD sampler into Joint Energy Models (JEMs) to enhance the calibration of predicted outcomes while maintaining training stability. This integration leverages the unique advantages of SVGD's particle-based sampling technique and incorporates a smoothing factor α to address numerical instability issues. Our empirical results show that this integration improves the Expected Calibration Error (ECE) over SGLD with SAM optimization, indicating more reliable confidence estimates in predictions. This contribution also provides a robust framework for balancing classification and generative objectives, offering practical enhancement to the training of JEMs.

2. **Comprehensive Evaluation of SVGD**: Our work extensively evaluates the potential of SVGD as an alternative to traditional MCMC methods like SGLD within the JEM framework. Through empirical experiments, we identified key hyperparameters

such as the number of particles, the number of sampling steps, and the smoothing factor, and analyzed their impact on model performance. Even though empirically SGLD with SAM remains superior in performance, our detailed investigation highlights the conditions under which particle-based sampling techniques can be effectively utilized, providing valuable insights for future research and optimization in this domain.

The architectural modifications proposed in this thesis aim to explore the feasibility of particle-based sampling techniques as viable alternatives to expensive MCMC methods within Joint Energy Models (JEMs). By comparing the theoretical and empirical performance of SVGD and SGLD, this research provides valuable insights into the strengths and limitations of these techniques. Specifically, our findings contribute to understanding when and how particle-based methods can be effectively utilized, thereby advancing the development of efficient and robust sampling techniques for energy-based models.

The rest of the thesis is structured as follows: Chapter 2 lays the theoretical groundwork, detailing the principles underlying EBMs, JEMs, SGLD, and the SVGD sampling method. Chapter 3 describes the methodology outlining the integration of the SVGD sampler into our JEM, the rationale behind the traceable entropy term, as well as justification for other architectural modifications. Chapter 4 presents a comprehensive overview of the datasets, data augmentation strategies, and complete setup for all our experiments. Chapter 5 presents our results and discusses the implications of the findings, followed by Chapter 6 where we highlight the limitations of our study and provide future direction to expand on this work. Finally, Chapter 7 concludes the thesis, summarizing the contributions of our work.

Chapter 2

Literature Review

2.1 Energy Based Models

EBMs offer a versatile framework for representing complex distributions over data. At their core, EBMs define an energy function that maps configurations of variables (*e.g.*, images, text) to scalar values, representing the system's energy [11]. The fundamental principle guiding EBMs is that configurations with lower energy are more likely, encapsulating a probabilistic interpretation of the model's preferences [9, 11]. This is formalized through the concept of an energy function, which in its generalized form under Gibbs distribution is defined as follows:

$$P(Y \mid X) = e^{-E(Y,X)}/Z(X). \tag{2.1}$$

Here, we are expressing the probability of a configuration 'Y' given an input 'X', and the 'Z(X)' term represents a partition function that ensures that probabilities normalize to 1 [11]. However, in the context of machine learning, ideally, we would want to describe the statistical distribution of states in a system. This is achieved in physics using the Boltzmann distribution as follows [9]:

$$P_\theta(x) = e^{-E_\theta(x)}/Z(\theta). \tag{2.2}$$

Here, $P_\theta(x)$ represents the probability of observing state x under the model parameterized by θ, $E_\theta(x)$ represents the energy function associated with state x parameterized by θ, and

$Z(\theta)$ is the partition function that's typically computed over all possible states of the system.

Designing an EBM entails choosing an energy function capable of capturing the intricate relationships within data. The learning process in EBMs involves adjusting model parameters to minimize a loss function, thereby ensuring observed data is assigned lower energies than unobserved data. This is typically achieved through optimization methods that aim to reduce the discrepancy between the model's predictions and actual data distribution.

A critical aspect of working with EBMs is efficiently sampling from the modeled distribution or performing inference, both of which are often challenging due to the intractable partition function [12]. MCMC algorithms, which specialize in sampling from complex probability distributions where direct sampling is not feasible, have been applied here in the form of SGLD by several previous works [10, 13, 14]. SGLD facilitates approximate sampling and iteratively updates samples to converge toward the target distribution. In addition to sampling, EBMs face several challenges including training stability, and the computational complexity of computing the partition function [2, 12]. Recent innovations such as Joint Energy Models (JEMs) (refer to Section 2.2) and their variations have sought to overcome the aforementioned challenges as well as some of the limitations associated with traditional vanilla SGLD sampling such as requiring an infinitely large number of forward and backward passes through the network [10].

2.2 Joint Energy Models (JEMs)

JEMs represent a significant breakthrough in machine learning by harmonizing the generative and discriminative paradigms into a single cohesive framework [9]. The foundational principle of JEMs is based on the insight that conventional discriminative classifiers when interpreted through the lens of EBMs, can inherently possess generative capabilities [9]. JEM leverages

these insights to train a hybrid model that jointly maximizes the cross-entropy objective for classification and the maximum likelihood of the data, *i.e.*,

$$\mathcal{L}_{\text{JEM}}(\theta) = \mathbb{E}_{(x,y)}[\log p_\theta(y|x) + \log p_\theta(x)], \tag{2.3}$$

where the data maximum likelihood is learnt via contrastive divergence.

At their core, JEMs posit that a well-structured energy function can simultaneously govern the generation of new data instances and the classification of existing ones. This dual functionality is encapsulated within the model's energy landscape, delineated by a parameterized energy function, $E(x, y; \theta)$ where x denotes the input features, y denotes the labels if any for class conditioned generation, and θ denotes the model parameters. Here, the energy function is meticulously designed to allocate lower energy scores to more plausible (or observed) configurations of data and labels, mirroring the probabilistic inclinations of data distributions within the latent space. This allocation is quantitatively expressed through the Boltzmann distribution, where the probability of observing a particular data-label pair is inversely related to its energy (refer to Equation 2.2), thus forming a direct correlation between statistical likelihood and energy minimization.

JEMs essentially reinterpret the logits of a classification network as components of an energy function [9]. This in turn allows the same model to perform classification by direct inference and data generation through energy-based sampling methods like SGLD. While JEMs advance the frontier of what is possible within energy-based frameworks, they have a few critical limitations. Vanilla JEMs often struggle to achieve SOTA results in both generative and discriminative tasks simultaneously due to the inherent trade-offs in optimizing a single model for two distinct objectives [10, 13]. Additionally, training JEMs can be quite difficult as the energy landscape defined by the model can be complex and challenging to navigate

during the optimization process [15].

To address these limitations, significant advancements have been proposed in recent works [10, 13, 15]. Specifically notable papers like [10] introduce key innovations such as the use of proximal point SGLD instead of vanilla SGLD for more stable sampling. They also extend the 'You Only Propagate Once' (YOPO) [16] framework to speed up the training process by reducing redundant calculations and by incorporating an information initialization strategy for the sampling process. In another pivotal paper [13], the authors extend their previous work in [10] by introducing a Sharpness Aware Minimization (SAM) optimization framework to smooth the energy landscape and improve generalization [17]. By selectively omitting data augmentation in the maximum likelihood estimation process, SADA-JEM leads to a smoother energy landscape, which in turn leads to remarkable improvements in both image classification and generation [13].

However, despite the improvements made by [10, 13, 15], several challenges pertinent to sampling diversity and training stability remain unresolved. Notably, the efficacy of SGLD in exploring the model's energy landscape can be compromised due to the high variance in noisy gradients, which leads to slower mixing [18]. This high variance results in less effective exploration of the energy landscape, potentially producing samples that lack diversity. Consequently, this limitation hinders the model's ability to fully capture the essence of the underlying data distribution, affecting its overall performance and robustness.

Such limitations of SGLD highlight the necessity for alternative sampling approaches that can enhance the efficiency, stability, and diversity of generated samples within JEMs. Theoretically, SVGD is a promising candidate offering improvements in sampling diversity and potentially image quality by leveraging a particle-based approximation technique to iteratively update a set of particles in the direction that most rapidly decreases their energy.

2.3 Stein Variational Gradient Descent

SVGD is a novel approach to variational inference that leverages particle-based methods to approximate complex distributions. At its core, SVGD iteratively updates a set of particles to minimize the Kullback-Leibler (KL) divergence between the target distribution and the approximation provided by the particles' distribution [19]. This is analogous to gradient descent in optimization, with the significant difference being SVGD's use of particles to represent the distribution.

Mathematically, given a target distribution $p(x)$ and our set of particles at iteration t as $\{x_i^t\}_{i=1}^n$, according to the SVGD rule, these particles are updated as follows:

$$x_i^{t+1} = x_i^t + \epsilon_t \phi(x_i^t), \tag{2.4}$$

where ϵ_t is the step size, and $\phi(x)$ represents the optimal direction for moving the particle, computed based on both the gradient and log-probability of the target distribution. $\nabla \log p(x)$, and a reproducing kernel Hilbert space (RKHS) to ensure the smoothness of the approximation [19]. Mathematically, $\phi(x)$ is described as follows:

$$\phi(x_i) = \frac{1}{n} \sum_{j=1}^{n} \left[k(x_i, x_j) \nabla_{x_j} \log p(x_j) + \nabla_{x_j} k(x_i, x_j) \right], \tag{2.5}$$

where n represents the total number of particles and $k(x_i, x_j)$ denotes a kernel function (typically RBF) that assesses the similarity between x_i and x_j. $\nabla_{x_j} \log p(x_j)$ denotes the gradient of log probability toward the target distribution x_j and $\nabla_{x_j} k(x_i, x_j)$ is the gradient of kernel function with respect to x_j; this instigates a repulsive force that spreads the particles apart to explore the distribution space. This formulation ensures that each particle updates its position by considering the influence of all other particles, thereby capturing the complex

structure of our target distribution through a distributed representation. Put simply, the term $k(x_i, x_j)\nabla_{x_j} \log p(x_j)$ pulls particles towards the modes of $p(x)$, while $\nabla_{x_j} k(x_i, x_j)$ prevents particle clumping by introducing a repulsive force thereby promoting diversity in the particle approximation of the distribution [19].

JEMs are uniquely poised to benefit from SVGD's integration due to their inherent design around energy functions. Moreover, the strengths of SVGD, such as efficient posterior sampling without the need for explicit normalization and gradient utilization for particle updates, are particularly advantageous here [20]. When juxtaposed with other popular sampling techniques such as SGLD and Hamiltonian Monte Carlo (HMC), SVGD stands out for several reasons. Its deterministic update mechanism, rooted in computing RKHS gradients, often leads to more stable convergence compared to the stochastic, exploratory nature of SGLD [21]. Further, our implementation of the SVGD sampler allows for adaptive step sizes during the particle updates, taking into account the local structure of the energy landscape. This makes it much more flexible than SGLD that requires meticulous fine-tuning of step-size to balance exploration and stability [15], which can prove challenging in the context of JEMs where different regions of the energy landscape may have vastly different characteristics. SVGD sampling also has the unique characteristic of incorporating repulsive forces which may prevent mode collapse, a common issue in sampling methods where samples may converge to a few modes of target distribution, neglecting others [22, 23]. Moreover, SVGD being particle-based has the potential to adapt more fluidly to high-dimensional spaces, offering a more effective solution for complex distribution approximations.

2.4 Stochastic Gradient Langevin Descent

Stochastic Gradient Langevin Descent (SGLD) is a popular sampling method that combines the benefits of stochastic gradient descent (SGD) with Langevin dynamics to generate samples from a target distribution. SGLD extends the principles of SGD by incorporating a noise term derived from Langevin dynamics, rooted in the Langevin equation which describes the motion of a particle subjected to both deterministic and stochastic forces in a potential field. Theoretically, the SGLD update rule is formulated as :

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\nabla_\theta \log p(\theta \mid D) + \eta_t. \tag{2.6}$$

Here, θ_t and $\nabla_\theta \log p(\theta \mid D)$ represent the model parameters at time t and the gradient of the log posterior distribution, respectively, while $\eta_t \sim \mathcal{N}(0, \epsilon_t)$ denotes the Gaussian noise with variance proportional to the step size ϵ_t. However, in the context of large datasets, evaluating the gradient $\nabla_\theta \log p(\theta \mid D)$ over the entire dataset can be computationally prohibitive. SGLD typically mitigates this by using mini-batches to approximate the gradient as follows [24]:

$$\theta_{t+1} = \theta_t + \frac{\epsilon_t}{2}\left(\nabla_\theta \log p(\theta) + \frac{N}{n}\sum_{i=1}^{n}\nabla_\theta \log p(x_i \mid \theta)\right) + \eta_t \tag{2.7}$$

In addition to the notation of Equation 2.6, N and n denote the total number of data points and mini-batch size, respectively, while $\nabla_\theta \log p(\theta)$ represents the gradient of the log prior. The gradient of the log-likelihood for data point x_i is shown as $\nabla_\theta \log p(x_i \mid \theta)$. The step size ϵ_t is constrained under specific conditions such that $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\sum_{t=1}^{\infty} \epsilon_t^2 < \infty$. These conditions ensure that the algorithm explores the parameter space adequately and eventually converges to the target posterior distribution. The former condition guarantees that the algorithm can reach high-probability regions regardless of initial parameter values, while the latter ensures convergence by reducing step size over time [24]. This update

rule incorporates both gradient-based optimization and stochastic noise, striking a balance between exploring the parameter space effectively and exploiting known high-probability regions. One of the key advantages of incorporating this with JEMs is that SGLD integrates seamlessly with stochastic optimization frameworks, allowing for a scalable inference that captures parameter uncertainty. Moreover using mini-batches for gradient computation reduces the computational costs, especially in terms of compute memory as compared to full-batch MCMC methods making this a more suitable method for large datasets [25]. However, as expressed in Equation 2.7, the performance of the SGLD sampler is highly dependent on the choice of step size. Incorrect selection can lead to slow convergence or high variance in samples making it challenging to tune the algorithm for optimal performance. We were also able to verify this empirically. Additionally, SGLD sampling may still suffer from approximation errors, especially when the posterior distribution is highly non-Gaussian or has sharp modes [25, 26]. To address these specific challenges, one of the prominent optimization strategies used in tandem with SGLD sampling is SAM [17].

Chapter 3

Methodology

3.1 Method Overview

The primary objective of integrating SVGD is to enhance the sampling process by exploiting a set of interacting particles that evolve over iterations to approximate complex posterior distributions efficiently. Unlike SGLD, which relies on injecting noise into gradient updates and thus may suffer from high variance and inefficient exploration of the parameter space, SVGD uses a deterministic update rule based on a combination of repulsive forces and gradient-driven movements. Theoretically, this method not only ensures a more effective coverage of the distribution space but also stabilizes the convergence process, making it particularly advantageous for models where the energy landscape is intricate. In the context of generative models, entropy serves as a crucial measure of uncertainty and diversity within the model's predictive outputs [27]. By integrating an entropy term into the JEM framework, the system is designed to actively promote diversity in particle approximations. Mathematically, entropy is defined as follows:

$$H(x_i) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{3.1}$$

where $p(x)$ represents the model's estimated probability mass function, and x_i denotes the particles. Practically, maximizing the entropy term during SVGD updates should encourage

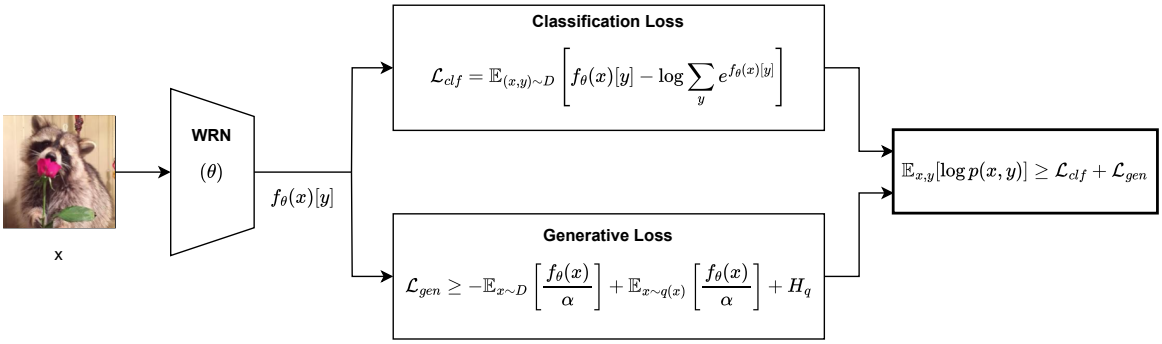


Figure 3.1: **Model Overview.** This model architecture diagram for JEMs showcases the flow of data through the WRN model, where an input image x with true label y is processed to compute both classification and generative losses. The Wide Residual Network (WRN) model, represented by θ parameters, outputs an energy score $f_\theta(x)$ used to calculate the classification loss $\mathcal{L}_{clf}$ via cross-entropy and the generative loss $\mathcal{L}_{gen}$ which accounts for samples generated via the SVGD sampler and the entropy term. The final model objective ensures that the joint probability $\mathbb{E}_{x,y}[\log p(x,y)]$ is at least as large as the sum of the classification and generative losses, optimizing both tasks concurrently.

the particles to explore more of the distribution space, thus preventing a common issue of particle degeneracy in which samples cluster around limited high-probability modes neglecting substantial areas of the distribution [28].

3.2 Model Configuration

The use of the Wide Residual Network (WRN) [29] as the backbone of our JEM as visualized in Figure 3.1 is a strategic choice, inherited from previous successful implementations detailed in [9, 10, 13]. This choice leverages WRNs proven capabilities in handling complex image datasets through its deep, yet efficient, architectural design [29, 30, 31]. The width and depth parameters of the WRN are inherited from [13], which optimize the trade-off between computational demand and the ability to capture intricate patterns in the data. The energy function is a critical component of the model, quantifying the state of the system in a way

that defines higher energy states as more probable, aligned with the practical implementation of the energy function. This function is integrated into the output layer of the WRN, which returns a scalar value that quantifies the energy associated with each input sample. The SVGD sampling procedure utilizes the gradients of this energy function to inform the updates to its particles. Specifically, it uses these gradients to push the particles towards regions of higher energy, effectively guiding them through the energy landscape defined by the model. Mathematically this is represented as follows:

$$E(x) = f_\theta(x), \tag{3.2}$$

where for an input sample x in the network, $E(x)$ is the energy and $f_\theta(x)$ is the output of a neural network parameterized by weights θ, indicative of the energy of x.

3.3 SVGD Implementation

This section delves into the details of how SVGD operates within our overarching framework, focusing on particle initialization, the kernel function's role, and the mechanism for updating the particles.

3.3.1 Particle Initialization

Particle initialization constitutes a fundamental aspect of our SVGD sampler, critically influencing the subsequent exploration of the distribution space by defining the initial conditions under which the model operates. Effective initialization is paramount to ensuring both the efficiency of convergence and the robustness of the SVGD algorithm's performance. Following established practices from prior research [13], our approach adopts a dual-phase strategy

for initializing particles.

Phase One: Establishing the Replay Buffer. Initially, the replay buffer is populated with particles drawn from a multivariate normal distribution. The parameters of this distribution—specifically its mean and covariance—are derived from the target dataset. This procedure ensures that the initial particle set is representative of the underlying data distribution, positioning the particles within a statistically coherent space relative to the target distribution.

Phase Two: Dynamic Update of the Replay Buffer. After initially populating the replay buffer, we implement a dynamic updating mechanism that selectively updates the buffer based on the particles exhibiting the highest energy, as determined by the SVGD sampler. This updating process employs a strategy, where we use the softmax function to calculate probabilities proportional to the energies of the particles. Empirically, we found that using the softmax strategy is more effective than selecting particles based on the maximum energy alone. By continually updating the buffer in this manner with the highest softmax-derived probabilities, we effectively create a feedback loop that refines the particle set towards increasingly probable states of the distribution. In subsequent iterations, the replay buffer allows the initial particle sampler to start sampling from these optimally adjusted positions, maintaining a trajectory that progressively aligns closer to the target distribution's mode. These modifications aim to increase the efficiency of the sampler by reducing unnecessary exploratory variance and focusing on more promising regions of the sample space.

3.3.2 Kernel Function

The kernel function in SVGD represents the interactions among particles and guides them toward a better approximation of the target distribution. Since the choice of kernel func-

tion is of utmost importance, we decided to use the 'Radial Basis Function' (RBF) for its effectiveness in smoothing particle interactions over the Euclidean space of the inputs [32]. Mathematically the RBF kernel is defined as follows:

$$k(x_i, x_j) = \exp\left(\frac{-||x_i - x_j||^2}{2\sigma^2}\right) \tag{3.3}$$

Here, σ denotes the bandwidth of the kernel and it controls the range and strength of influence each particle exerts on others. A smaller σ results in sharper interaction that is sensitive to close neighbors, whereas a larger σ smooths out interactions over a broader ranger. Given the importance of σ, it is crucial that such a parameter is flexible based on the distribution of particles at a given iteration. Thus, we introduce an adaptive bandwidth configuration in our kernel function to flexibly adapt to the density and distribution of particles across various stages of the distributions' exploration, especially as particles move to denser or sparser regions. Empirically we've found that the best heuristics for an adaptive bandwidth parameter are based on median distance among the particles.

3.3.3 Particle Update Mechanism

The SVGD update mechanism is designed to iteratively refine the ensemble of particles such that the empirical distribution they form more closely approximates the target distribution. Theoretically, this process is fundamentally driven by minimizing the Kullback-Leibler (KL) divergence between the empirical and target distributions, which in turn enhances the model's sampling capabilities [19, 33]. In practice, at each iteration t, each particle $x_i^{(t)}$ is updated according to a carefully defined velocity field $\phi(x_i^{(t)})$, which is dependent on the

positions of all particles in the system. The update mechanism is represented as follows:

$$x_i^{(t+1)} = x_i^{(t)} + \epsilon_t \phi(x_i^{(t)}), \tag{3.4}$$

where ϵ_t is the step size, a crucial factor for controlling the learning rate of the particle updates. The velocity field $\phi(x_i)$ for each particle is computed by using a combination of kernel-induced drift, clamped gradients of the energies derived from the JEM, and the repulsive force (refer to Equation 2.5). The first term in the summation acts as the attraction force and uses the kernel function to weigh the influence of each particle on others based on their proximity or similarity, effectively creating a 'smoothed' attraction towards a high probability area. The clamped gradients limit the range of the gradient values ensuring that the updates are stable and excessively large steps that could lead to instability or divergence are avoided. The second term serves as a counterbalance and ensures that the particles don't collapse into a few modes of the distribution but rather explore it comprehensively.

3.4 Entropy Enhancement

In the SVGD framework, the concept of entropy plays a pivotal role (refer to Equation 3.5) to align more closely with the practical requirements of JEMs. The classical definition of entropy as seen in Equation 3.1, deeply rooted in statistics [34], measures the expected randomness in a distribution. However, for the purposes of variational inference using SVGD, we define entropy by focusing on the empirical distribution represented by a set of particles as follows:

$$H_q = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i). \tag{3.5}$$

This expression can be derived from considering the empirical distribution formed by the particles as an approximation of the true distribution $p(x)$. Each x_i contributes equally to this approximation, and the $\log p(x_i)$ provides a measure of how probable the particle x_i is under the target distribution. This modification computes the average log probability across particles, rather than their distribution over the model's entire probabilistic landscape. By focusing on $\log p(x)$, we prioritize improving the probabilistic endorsement of each particle's position directly. Theoretically, the use of the mean also makes this entropy robust against outliers and extreme values in particle positions. This stability is crucial for the iterative nature of SVGD, where each update influences the subsequent distribution of particles.

A well-documented and empirically confirmed challenge within JEMs lies in balancing classification efficacy with generative quality [10]. By integrating such a modified entropy formulation into the generative loss component, theoretically our framework benefits from a more consistent loss landscape, which facilitates the generation of more diverse and representative samples. Simultaneously, this integration should not compromise the model's classification accuracy. Thus, the inclusion of this entropy measure not only supports enhanced generative performance but also preserves, if not augments, the model's ability to classify accurately.

3.5 Loss Function

As discussed in Chapter 2 Section 2.1, EBMs inherently face the challenge of an intractable normalization constant, which complicates the direct computation of likelihoods. This characteristic necessitates the derivation of a tractable loss function, specifically one that accounts for the estimation of the normalization constant since its exact value cannot be computed. Based on established formulations in previous works we know that the loss function for a

JEM can be stated as follows [9]:

$$\log p_\theta(x, y) = \log p_\theta(y|x) + \log p_\theta(x). \tag{3.6}$$

Here, $\log p_\theta(y|x)$ denotes the cross-entropy objective for classification. The second term $\log p_\theta(x)$ represents the generative loss objective and can be optimized by minimizing the negative log-likelihood of the EBM. This term is defined as follows:

$$p_\theta(x) = \frac{e^{\frac{f_\theta(x)}{\alpha}}}{Z(\theta)} \qquad \text{-(i)}$$

Taking log on both sides we get:

$$\log p_\theta(x) = \frac{f_\theta(x)}{\alpha} - \log Z(\theta)$$

However, since we need to compute the negative log-likelihood over the entire data distribution (D) represented by $\mathcal{L}_{gen}$:

$$\begin{aligned} \mathcal{L}_{gen} &= -\mathbb{E}[\log p_\theta(x)] \\ &= -\mathbb{E}_{x\sim D}\left[\frac{f_\theta(x)}{\alpha}\right] + \log Z(\theta) \end{aligned} \qquad \text{-(ii)}$$

Let $q(x)$ be a proxy distribution for the SVGD sampler. Then by definition of KL Divergence we get:

$$\mathrm{KL}(q(x)||p_\theta(x)) = \mathbb{E}_{x\sim q(x)}\left[\log \frac{q(x)}{p_\theta(x)}\right] \geq 0 \qquad \text{-(iii)}$$

$$\mathrm{KL}(q(x)||p_\theta(x)) = \mathbb{E}_{x\sim q(x)}\left[\log q(x) - \frac{f_\theta(x)}{\alpha} + \log Z(\theta)\right] \qquad \text{from (i) and (iii)}$$

Rearranging for $\log Z(\theta)$:

$$\log Z(\theta) = \mathrm{KL}(q(x)||p_\theta(x)) - \mathbb{E}_{x\sim q(x)}[\log q(x)] + \mathbb{E}_{x\sim q(x)}\left[\frac{f_\theta(x)}{\alpha}\right]$$

Since we cannot directly compute the precise value of the KL divergence term over the entire data, we introduce an inequality and estimate a lower bound instead as follows:

$$\log Z(\theta) \geq -\mathbb{E}_{x\sim q(x)}[\log q(x)] + \mathbb{E}_{x\sim q(x)}\left[\frac{f_\theta(x)}{\alpha}\right] \qquad \text{-(iv)}$$

Substituting all known values back in the original generative loss term:

$$\mathcal{L}_{gen} \geq -\mathbb{E}_{x\sim D}\left[\frac{f_\theta(x)}{\alpha}\right] + \mathbb{E}_{x\sim q(x)}\left[\frac{f_\theta(x)}{\alpha}\right] + H_q \qquad \text{from (ii), (iv), (3.5)}$$

In theory, the probability density function of an EBM is given by Equation 2.2 as explained in Chapter 2 Section 2.1. This formulation, while elegant, can pose significant challenges in practice due to the potentially unbounded nature of the energy function $f_\theta(x)$, leading to numerical instability and difficulties in optimization. To address these challenges, we introduce a smoothing factor α into the energy function resulting in a modified formulation in step (i) of the derivation. Thus, our rigorous mathematical derivation presented here forms the backbone of our model's learning algorithm, enabling it to effectively learn the generative aspects of the data. Theoretically the generative loss objective can be stated as follows:

$$\max_q \min_\theta \left(-\mathbb{E}_{x\sim D}\left[\frac{f_\theta(x)}{\alpha}\right] + \mathbb{E}_{x\sim q(x)}\left[\frac{f_\theta(x)}{\alpha}\right] + H_q\right) \tag{3.7}$$

Minimizing this loss with respect to the model parameters θ is essential for aligning the model

with high-density regions of the data distribution, thereby improving both classification and generative tasks. While the loss formulation includes a maximization over q to better approximate the true distribution, our current focus is on optimizing θ to enhance classification performance. Theoretically, this minimization enhances the model's overall understanding of the data. This richer data representation supports better calibration with classification tasks, complementing the cross-entropy loss by refining the underlying data-driven features that both the classification and generative components rely on. This approach lays a strong foundation for a well-calibrated classifier, with future work planned to incorporate the maximization over q to fully leverage the model's generative capabilities. Thus in our current implementation the generative loss can be stated as follows:

$$\min_{\theta} \left(-\mathbb{E}_{x \sim D} \left[\frac{f_\theta(x)}{\alpha} \right] + \mathbb{E}_{x \sim q(x)} \left[\frac{f_\theta(x)}{\alpha} \right] + H_q \right) \tag{3.8}$$

Chapter 4

Experiments

In the previous Chapters, we have thoroughly discussed the theoretical underpinnings of Stochastic Gradient Langevin Dynamics (SGLD), Stein Variational Gradient Descent (SVGD), and Joint Energy Models (JEMs). Chapter 3 provided an in-depth analysis of how the SVGD sampler is integrated with the JEM framework, as defined in [13]. In this chapter, we will review the dataset and the experimental setup for empirically evaluating these different sampling techniques.

4.1 Experimental Setup

To ensure a comprehensive evaluation, we conducted experiments on the CIFAR-10 dataset [35], a widely used benchmark for image classification tasks. The experiments were designed to assess the performance of both the SVGD and the SGLD sampler with our JEM.

Dataset:

CIFAR-10: The dataset consists of 60,000 32×32 color images, divided into 10 classes, with 6,000 images per class. We used the standard split of 50,000 training images and 10,000 test images.

Evaluation Metrics:

Classification Accuracy: Measures the percentage of correctly classified images.

Total Loss: Combination of cross-entropy loss for smoothing the classification and custom loss functions for smoothing the landscape and minimizing negative log-likelihood under the target data distribution denoting the generative loss value (refer to Section 3.1).

Expected Calibration Error (ECE): Measures the discrepancy between predicted confidences and actual correctness of predictions, providing an indication of how well the model's predicted probabilities align with the true outcomes.

Experimental Configurations: We conducted two major types of experiments:

With SAM Optimization: Evaluating the performance of SVGD and SGLD samplers when combined with Sharpness-Aware Minimization (SAM) optimization has shown significant advances in selecting hyperparams leading to a smoother loss landscape [13, 17, 36].

Without SAM Optimization: Evaluating the baseline performance of SVGD and SGLD samplers without SAM optimization to understand the unabridged difference between the two different sampling techniques.

Table 4.1 below provides a detailed list of the key hyperparameters used in each of the four experimental configurations as well as the time taken for one epoch over the training set of the CIFAR-10 dataset.

Experiment	Hyperparams	Time Per Epoch
SGLD with SAM	Sampler LR: 1, BS: 64, N_Steps: 10, N_GPUs: 1	$\sim$ 18 mins
SGLD without SAM	Sampler LR: 1, BS: 64, N_Steps: 10, N_GPUs: 1	$\sim$ 16 mins
SVGD with SAM	Sampler LR: 0.09, BS: 10, N_Steps: 10, N_GPUs: 4, N_particles: 30	$\sim$ 20 mins
SVGD without SAM	Sampler LR: 0.09, BS: 10, N_Steps: 10, N_GPUs: 4, N_particles: 30	$\sim$ 15 mins

Table 4.1: Key Hyperparameters of Experimental Configurations

In addition to the primary hyperparameters listed above, all experiments were run for 100 epochs in the interest of maximizing available resources. Other critical hyperparameters, including those for the optimizer and the JEM's learning rate, were directly adopted from [13]. This standardized approach ensures that our experimental focus remains on evaluating the attributes and performance of the sampling techniques themselves, rather than on the fine-tuning of specific model parameters within the JEM framework. All experiments were conducted using NVIDIA L40S GPUs (48GBs per GPU) and implemented with PyTorch version 1.10.2, ensuring a consistent computational environment with previous works [10, 13].

4.2 Data Augmentation

This section provides a detailed overview of the implementation specifics, particularly focusing on the data augmentation techniques employed and the different types of data loaders used for training our hybrid JEM for its dual classification-generation objective.

Intuitively when training an image classification model, the goal is to make the classifier robust to both the original image and its common variations. These variations are achieved through augmentations such as random cropping and horizontal flipping. In contrast, generative models are typically learning to generate the original image, treating each variant as a distinct image with its own generative space that needs to be learned separately. Given that JEM is a hybrid model, it must balance these two objectives — learning robust features for classification while also accurately generating images. This dual learning approach was empirically verified by [13], and thus, we have adopted a similar dual data loader strategy for training. The data loader plays a crucial role in preparing and feeding the data to the model during training and evaluation. In our implementation, we have used PyTorch's 'DataLoader' class to handle the batching, shuffling, and loading of our dataset. We have defined differ-

ent transformations for one of the data loaders in the training set which includes padding, random cropping, and random horizontal flipping to help the model generalize to different variants of the same image while the other training data loader is simply converted to a tensor and normalized to align with the generative objective. To leverage multi-GPU training, we utilized PyTorch's DistributedSampler along with DistributedDataParallel (DDP). This approach is preferred over the more commonly used DataParallel as it ensures efficient load balancing and synchronization across multiple GPUs (verified empirically), which is crucial for handling the large compute memory requirements of the SVGD sampler.

Chapter 5

Results and Analysis

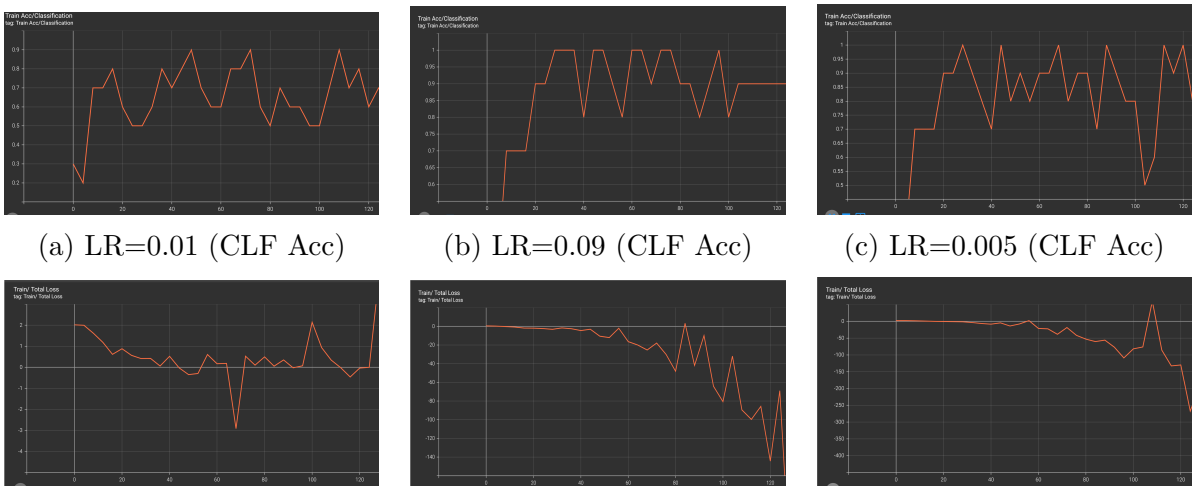
In the previous chapter, we outlined the experimental setup, including the dataset, evaluation metrics, and configurations used for empirical evaluations. In this chapter, we present and discuss the results of our experiments, focusing on three key metrics: classification accuracy, total loss, and expected calibration error. These metrics are crucial for assessing the performance of our JEM experiments integrated with the SGLD and SVGD samplers in its dual classification-generation objectives.

We conducted extensive experiments on the CIFAR-10 dataset [35] to evaluate the effectiveness of SVGD and SGLD samplers, both with and without SAM [17] optimization. The results are analyzed to provide insights into the strengths and weaknesses of each sampling technique, the impact of SAM optimization, and the overall performance of the samplers within the JEM framework. This chapter aims to interpret the empirical findings in a comprehensive manner, highlighting key observations and meaningful insights drawn from our exhaustive experiments.

5.1 Small-Scale Hyperparameter Tuning Experiments

Given the time-consuming nature of training on the entire CIFAR-10 dataset [35], we conducted preliminary experiments on a smaller subset of 10 images. These experiments aimed to understand the correlation between key SVGD hyperparameters, such as the number of

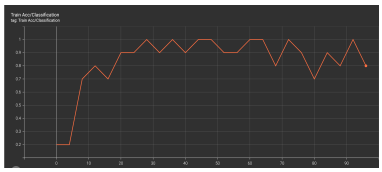
particles, number of sampling steps, sampler learning rate, and how they affect the exploration of the distribution space measured as the mean square difference between the initial and final positions of all particles in each epoch. Our central objective here was to ensure that we overfit on the small set and find an appropriate balance between the classification and generative loss. Additionally, these experiments also allowed us to determine the appropriate method between argmax and softmax to update our replay buffer as explained in Chapter 3 Subsection 3.3.1. We retain the same data augmentation strategy from Chapter 4, Section 4.2 and the key hyperparameters (refer to Table 4.1) for the full dataset run were decided in part based on these experiments. The accuracy and loss plots for the best of these experiments focusing on different hyperparameters at a time are shown in the Figures 5.1-5.4. To fairly assess one hyperparameter at a time, all other hyperparameters were frozen to a fixed value across the different runs as indicated in the figure captions.



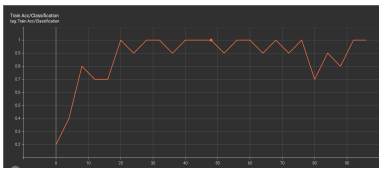
(a) LR=0.01 (CLF Acc) (b) LR=0.09 (CLF Acc) (c) LR=0.005 (CLF Acc)

(d) LR=0.01 (Total Loss) (e) LR=0.09 (Total Loss) (f) LR=0.005 (Total Loss)

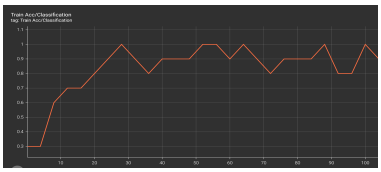
Figure 5.1: Metric Curves for Different Learning Rates at 43 Particles, $\alpha = 1$, and 10 Sampling Steps



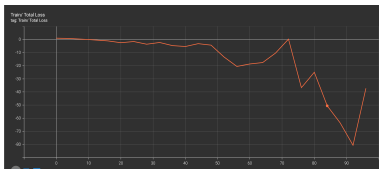
(a) N_Particles=15 (CLF Acc)



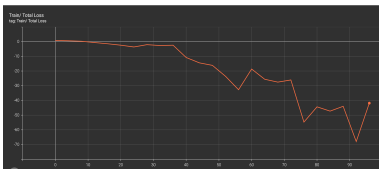
(b) N_Particles=23 (CLF Acc)



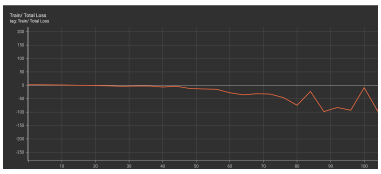
(c) N_Particles=40 (CLF Acc)



(d) N_Particles=15 (Total Loss)

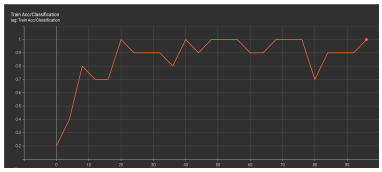


(e) N_Particles=23 (Total Loss)

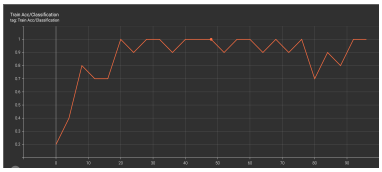


(f) N_Particles=40 (Total Loss)

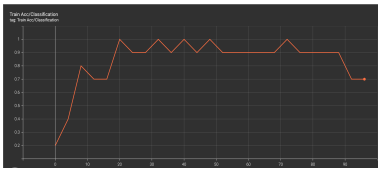
Figure 5.2: Metric Curves for Different Number of Particles at 0.09 Sampler Learning Rate, $\alpha = 1$, and 10 Sampling Steps



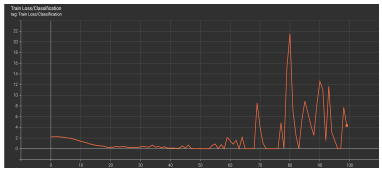
(a) N_Steps=5 (CLF Acc)



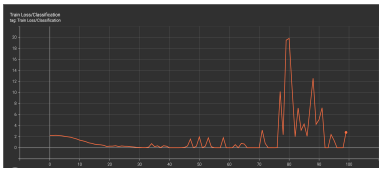
(b) N_Steps=10 (CLF Acc)



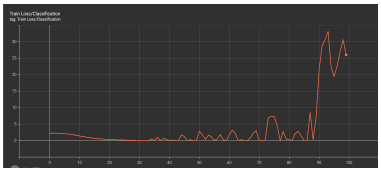
(c) N_Steps=20 (CLF Acc)



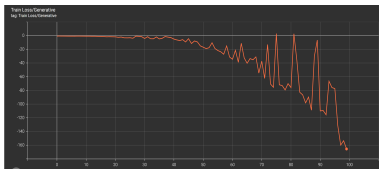
(d) N_Steps=5 (CLF Loss)



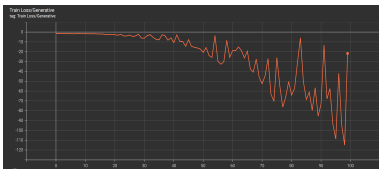
(e) N_Steps=10 (CLF Loss)



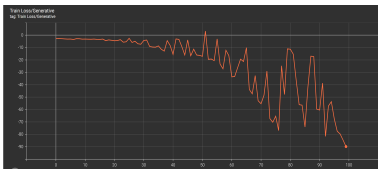
(f) N_Steps=20 (CLF Loss)



(g) N_Steps=5 (GEN Loss)

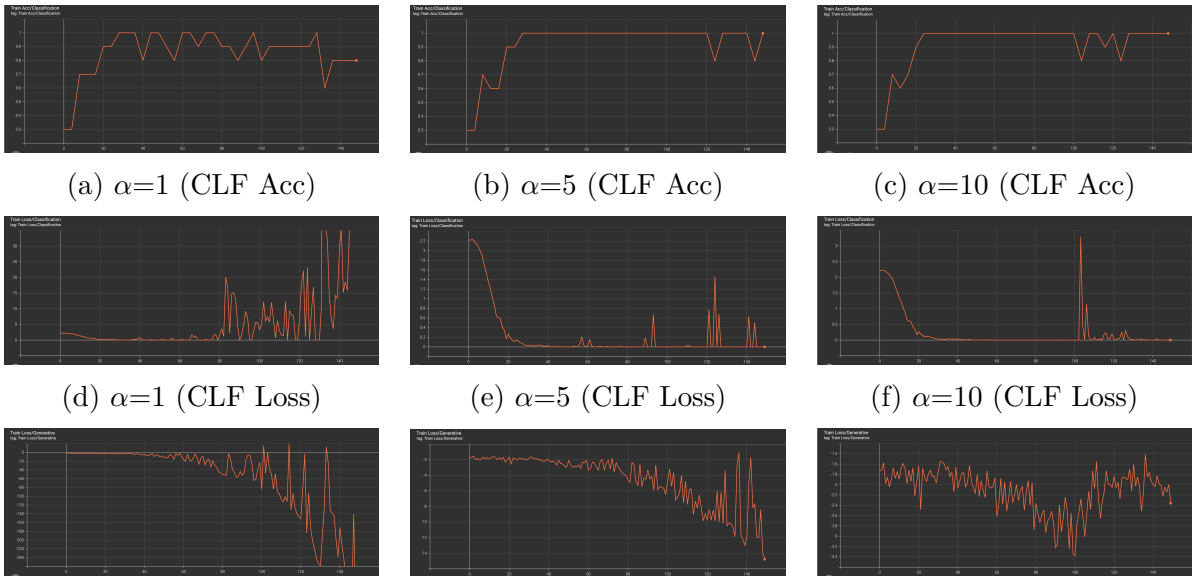


(h) N_Steps=10 (GEN Loss)



(i) N_Steps=20 (GEN Loss)

Figure 5.3: Metric Curves for Different Number of Steps at 0.09 Sampler Learning Rate, $\alpha = 1$, and 23 Particles



(a) α=1 (CLF Acc) (b) α=5 (CLF Acc) (c) α=10 (CLF Acc)

(d) α=1 (CLF Loss) (e) α=5 (CLF Loss) (f) α=10 (CLF Loss)

(g) α=1 (GEN Loss) (h) α=5 (GEN Loss) (i) α=10 (GEN Loss)

Figure 5.4: Metric Curves for Different Values of α at 0.09 Sampler Learning Rate, 43 Particles, and 10 Sampling Steps

5.1.1 Sampler Learning Rate

The sampler learning rate is pivotal in controlling the step size of updates during the SVGD process. It affects the stability and convergence of the training procedure, as shown in Figure 5.1. If this value is too high, *e.g.*, 0.01, it leads to instability and overshooting, causing poor model performance. On the other hand, if the value is too low, *e.g.*, 0.005, it slows down the convergence process and leads to sharp curves and highly negative total loss values. Thus, we choose a learning rate of 0.09 to ensure fast yet relatively smoother curves for accuracy and a more balanced range for total loss values.

5.1.2 Number of Particles

The number of SVGD particles is an essential hyperparameter in particle-based sampling methods. This parameter directly influences the diversity and representational capacity of

the particle set, which is vital for accurately modeling complex target distributions. Intuitively, increasing the number of particles should enhance the diversity of the generated samples, leading to more representative images and better approximation of the target distribution. However, our empirical observations indicate that an excessive number of particles can adversely affect the generative loss, consequently impairing the classification accuracy of the JEM, as shown in Figure 5.2.

Initially, we set the number of particles to 23 based on small-scale experiments, where this configuration provided a reasonable balance between computational efficiency and performance. However, during full dataset experiments, it became evident that over extended epochs, the generative loss and classification loss became unstable, and the classification accuracy did not exceed 30%. This behavior can be attributed to the inadequacy of a small number of particles in effectively capturing the intricacies of a complex target distribution.

To address this issue, we increased the number of particles to 30 for the full-scale run. Due to resource constraints, this was the maximum number of particles we could accommodate. This adjustment aimed to enhance the model's ability to explore the target distribution more thoroughly, thereby improving the stability and accuracy of the JEM's classification and generative tasks.

5.1.3 Number of Steps

The number of sampling steps is a crucial hyperparameter in the SVGD sampler as it determines the number of iterations each particle undergoes to approximate the target distribution during each update. Theoretically, the "more is better" principle might be expected to apply, suggesting that more sampling steps would lead to a more accurate and refined approximation of the target distribution.

However, empirical results reveal a more nuanced relationship. Specifically, we observed that increasing the number of sampling steps too drastically, such as to 20 steps, negatively impacts the classification loss. While this configuration results in smoother accuracy curves, the overall classification accuracy diminishes, as illustrated in Figure 5.3. This counterintuitive result can be attributed to the excessive refinement of the particles, which might lead to overfitting the generative aspects at the expense of the classification performance.

Additionally, increasing the number of steps significantly impacts the computational time required to sample from the distribution, thereby prolonging the total runtime of the experiments. This increase in computational load must be balanced against the potential performance gains. Based on this analysis, we determined that training with 10 sampling steps strikes an optimal balance as this decision is supported by both theoretical considerations and empirical evidence.

5.1.4 Smoothing Factor

The smoothing factor is one of the most vital hyperparameters in using the SVGD sampling technique with JEMs. As discussed in Chapter 3 Section 3.5, the generative loss term can become extremely negative due to its unbounded nature, disproportionately affecting the total loss value. Introducing this smoothing factor ensures that the generative loss does not dominate the overall loss, thereby maintaining the equilibrium essential for the JEM's dual objectives of classification and generation.

Additionally, the smoothing factor enables effective scaling of the energy values, preventing them from becoming excessively large and ensuring numerical stability during computations. This stabilization is critical for maintaining the performance and reliability of the model, especially during long training runs.

The selection of an appropriate smoothing factor is crucial, as later explained in Subsections 5.3.1 and 5.3.2. Empirically, our small-scale experiments indicated that as the value of this hyperparameter increases, the classification accuracy improves. This improvement occurs because the magnitude of the generative loss is reduced, resulting in a more balanced total loss. However, if the smoothing factor is increased excessively, the generative loss becomes too small, skewing the equilibrium towards the classification loss.

Based on the small-scale experiments, an optimal value for this hyperparameter was found to be 5, as shown in Figure 5.4. However, in the full dataset experiments, this low value was insufficient to effectively scale the magnitude of the generative loss. Consequently, for the full dataset, we selected a value of 50. This higher value helps to maintain relatively better balance between the generative and classification objectives over extended epochs, ensuring the stability and performance of the JEM framework.

5.2 Full-Scale Experiments

5.2.1 Metric Evaluation Results

The experiments on the full dataset reveal a clear hierarchy in terms of performance, with SGLD combined with SAM optimization achieving the highest classification accuracy, followed by SGLD without SAM, and lastly, the SVGD configurations. These results underscore the effectiveness of SAM optimization in fine-tuning model parameters. Despite the additional computational cost of back-propagating twice in each epoch, SAM optimization significantly improves accuracy. The trend in loss values mirrors the accuracy results, with the SGLD sampler both with and without SAM showing a more moderated range of loss values compared to SVGD configurations. Detailed results are provided in Table 5.5.

Experiment	Classification Accuracy	ECE
SGLD with SAM	93.48%	3.99%
SGLD without SAM	92.17%	1.98%
SVGD with SAM	60.24%	3.21%
SVGD without SAM	52.01%	2.36%

Table 5.5: Full-Scale Experimental Results

The Expected Calibration Error (ECE) [37] provides insights into model calibration. Interestingly, SVGD without SAM exhibits a low ECE despite its lower accuracy. This suggests that while the SVGD sampler without SAM optimization is less accurate overall, the probabilities it assigns to predictions are well-calibrated. This phenomenon indicates that SVGD without SAM maintains a better alignment between predicted confidences and actual outcomes, even if the overall classification accuracy is lower, which could be attributed to other factors explained in Subsection 5.3.1.

5.2.2 Impact of SAM Optimization

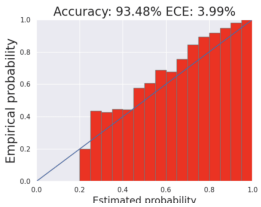
The effectiveness of SAM optimization in addressing unstable training scenarios has been thoroughly documented in numerous studies, highlighting its significant contributions to improving model robustness and stability [13, 17, 36]. The addition of SAM optimization significantly enhances the performance of the samplers as documented in Table 5.5.

Without SAM optimization, experiments with both samplers experienced higher loss values due to the inherent instability of JEM, as documented in previous works [9, 15]. Specifically, the SGLD sampler without SAM experienced exploding loss during training requiring restarts, whereas the SVGD sampler without SAM did not exhibit such instability. Despite this, the classification accuracy and ECE of the SGLD sampler without SAM, even with an exploding loss, were far superior to that of the SVGD sampler with SAM.

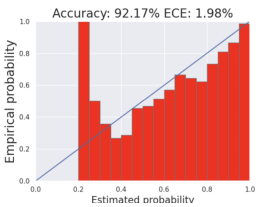
These findings highlight the critical role of SAM optimization in converging to optimal parameters more efficiently. In scenarios where SAM optimization is used, the models are able to reach optimal performance faster, which is reflected in the higher classification accuracy results. Conversely, without SAM, the experiments either struggle to achieve similar performance or the training becomes highly unstable leading to exploding losses which results in training the same model several times to reach a specific number of epochs.

Analyzing the ECE results further as visualized in Figures 5.5, it is evident the higher ECE values for SGLD, particularly with SAM, suggest that while these models may achieve higher accuracy, their confidence estimates are less reliable. On the other hand, the SVGD models, particularly without SAM, maintain good calibration (low ECE) despite lower accuracy. This suggests that while they are less accurate, the predictions they make are more reliable in terms of confidence estimation which can be vital for safety-critical operations.

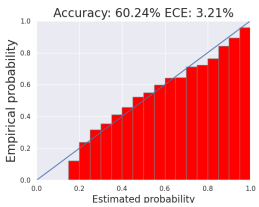
Overall, the use of SAM optimization improves accuracy, stability in training, and convergence on optimal parameters for JEMs regardless of the sampling techniques used.



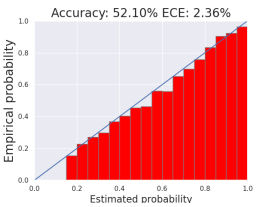
(a) SGLD w SAM



(b) SGLD w/o SAM



(c) SVGD w SAM



(d) SVGD w/o SAM

Figure 5.5: Calibration Results for SGLD & SVGD

5.3 Analysis of SVGD Performance

5.3.1 Challenges with Generative Loss

SVGD, despite its theoretical promise, exhibited significantly more negative total loss values in our experiments. This can be attributed to the inherent complexity and computational intensity of a particle-based sampler, as well as the empirical observation that the generative loss term in our SVGD sampler tends to dominate the total loss value. The generative loss term, as derived in Chapter 3 Section 3.5, represents a lower bound on the true loss function. This formulation indicates that without a carefully tuned hyperparameter α, this unbounded generative loss term could potentially decrease indefinitely, approaching $-\infty$, which in turn disproportionately influences the total loss value, making it extremely negative.

5.3.2 Necessity of Adaptive Scaling

The negative bias of the generative loss under SVGD suggests that the fixed smoothing factor α used may not be optimal. The incorrect selection of this fixed α likely led to the generative component overwhelming the classification objective, disrupting the equilibrium essential for our hybrid JEM's performance. Consequently, SVGD's empirical results highlight the critical need for adaptive tuning of the scaling factor to harness its full potential.

5.3.3 Calibration Analysis

An interesting observation from our experiments is that SVGD without SAM optimization exhibits lower Expected Calibration Error (ECE) [37] compared to its accuracy. This finding suggests that, while SVGD without SAM may not achieve high classification accuracy, it

produces well-calibrated probabilistic predictions. A well-calibrated model is one where the predicted confidence levels correspond accurately to the actual likelihood of correctness. For instance, if a model predicts an outcome with 70% confidence, that outcome should indeed occur 70% of the time.

The low ECE value for SVGD without SAM indicates that its prediction confidences are reliable, even if the predictions themselves are less accurate and unlike the SGLD sampler with SAM the model training paradigm is quite stable and the training loss does not explode. This insight into the calibration of SVGD without SAM suggests that further research should explore methods to improve its accuracy while maintaining or enhancing its calibration properties. Techniques such as adaptive smoothing factors or hybrid samplers that combine the strengths of SVGD and SGLD might offer pathways to achieve better overall performance.

Chapter 6

Limitations and Future Work

6.1 Limitations

6.1.1 Model Performance and Stability

A significant limitation in our study pertains to the performance and stability of the Joint Energy Model (JEM) when employing the Stein Variational Gradient Descent (SVGD) sampler. Although SVGD holds theoretical promise, it frequently resulted in more negative loss values compared to Stochastic Gradient Langevin Dynamics (SGLD). This phenomenon can be attributed largely to the generative loss term in SVGD, which tends to dominate the total loss as explained in Chapter 5 Subsection 5.2.1. Without a carefully tuned α, this generative loss can decrease indefinitely, disproportionately influencing the total loss. Despite extensive experiments, we were unable to identify the optimal α for the full CIFAR-10 dataset [35].

6.1.2 Computational Constraints

The high computational demands of SVGD, relative to SGLD, presented another substantial limitation. Particle-based methods like SVGD require significant computational resources in terms of both memory and processing power. The need for an increased number of particles and sampling steps to achieve a detailed approximation of the target distribution

resulted in higher computational costs. These demands significantly constrained the scope of our experiments, limiting our ability to explore a broader range of hyperparameters and configurations. This trade-off between computational cost and performance was particularly evident during full-scale experiments, where resource limitations restricted the number of particles we could use, potentially impacting overall performance.

6.1.3 Hyperparameter Tuning

Hyperparameter tuning, especially for the SVGD sampler, posed considerable challenges. Determining optimal values for the number of particles, sampling steps, and the smoothing factor α was largely empirical. Insights from small-scale experiments did not transfer straightforwardly to full-scale experiments. This was especially true for the smoothing factor α which required meticulous calibration to balance generative and classification losses, and inappropriate values led to a poor fit on the dataset as seen empirically.

6.1.4 Theoretical and Practical Gaps

Inherent theoretical and practical gaps in our study limit the broader applicability of our findings. While the theoretical assumptions underpinning EBMs are elegant, they do not always translate into practical performance improvements. The disparity between the theoretical promises of SVGD and its empirical performance in JEMs highlights the need for further research. Additionally, the potential unbounded nature of the energy function $f_\theta(x)$, despite the introduction of a smoothing factor, poses challenges in maintaining numerical stability and optimizing the model. Bridging these gaps is essential for developing more robust and practical EBMs.

6.2 Future Work

6.2.1 Enhancing Model Stability

While the current study has highlighted the stability issues associated with SVGD, particularly due to the negative bias of the generative loss term, future research should focus on developing advanced techniques to enhance model stability. One promising direction is the exploration of alternative particle-based sampling methods that inherently provide greater stability. For instance, integrating elements of Variational Inference with SVGD could offer more stable updates by leveraging variational bounds. Additionally, exploring other gradient-based sampling techniques that can better balance the generative and classification objectives within JEMs would be beneficial. Furthermore, a deeper theoretical investigation into the conditions that lead to instability could inform the design of more robust algorithms.

6.2.2 Advanced Hyperparameter Optimization

Given the challenges encountered with hyperparameter tuning, a promising future expansion of this work should incorporate automated hyperparameter optimization techniques. Methods such as Bayesian optimization, grid search, or random search can systematically identify optimal hyperparameter values, reducing the reliance on empirical tuning. Additionally, meta-learning approaches that adaptively adjust hyperparameters during training could be explored to dynamically optimize performance. Techniques such as adaptive learning rate schedules, second-order optimization methods, or momentum-based approaches could provide more stable and efficient training dynamics.

6.2.3 Bridging Theory and Practice

The gap between the theoretical promises of SVGD and its empirical performance with JEMs underscores the need for further research to bridge this divide. Future work should focus on developing new theoretical insights that can inform practical implementations. Additionally, integrating recent advancements in particle-based sampling and energy-based models can help refine current methods and improve their practical applicability. Continued theoretical studies should aim to provide a deeper understanding of the underlying mechanisms that drive the performance of these models.

6.2.4 Investigating Scalability and Efficiency

Scalability and efficiency are crucial aspects that need to be addressed in future work. Developing scalable algorithms that can efficiently handle large-scale datasets and complex models is essential for practical applications. Research should focus on optimizing the computational efficiency of JEMs, exploring techniques such as model compression, quantization, or distillation to reduce the computational footprint without compromising performance. Additionally, investigating the scalability of the proposed methods to distributed computing environments can facilitate their deployment in real-world scenarios.

Chapter 7

Conclusion

This thesis has explored the integration of Stein Variational Gradient Descent (SVGD) within the framework of Joint Energy Models (JEMs), contrasting its performance with the more traditional Stochastic Gradient Langevin Dynamics (SGLD) sampler. We investigated the theoretical underpinnings, practical implementations, and empirical performance of these techniques, particularly in the context of dual classification-generation objectives of JEMs.

7.1 Summary of Contributions

Our research has made several key contributions to the field of energy-based models:

1. **Integration of SVGD Sampler for Improved Calibration**: We integrated the SVGD sampler with JEMs to enhance the calibration of predicted outcomes. By incorporating a smoothing factor α, we addressed numerical stability issues, achieving better alignment between predicted confidences and actual outcomes, as indicated by lower Expected Calibration Error (ECE) values.

2. **Comprehensive Evaluation of SVGD**: Through extensive theoretical and empirical analysis, we evaluated SVGD as an alternative to traditional MCMC methods. Our experiments identified critical hyperparameters, including the number of particles, number of sampling steps, sampler learning rate, and smoothing factor, highlighting

their impact on model performance.

7.2 Empirical Findings

Our extensive experiments on the CIFAR-10 dataset revealed several important insights:

- **Performance Comparison**: SGLD consistently outperformed SVGD in classification accuracy. With SAM optimization, SGLD achieved a classification accuracy of approximately 93%, significantly higher than SVGD, which struggled with optimal convergence issues.
- **Impact of SAM Optimization**: SAM optimization was crucial in stabilizing the training process. Without SAM, both samplers experienced higher loss values, with SGLD being particularly prone to instability in training. SAM allowed both samplers to coverge to more optimal parameters however, it also resulted in models being less calibrated suggesting it may not be the most optimal choice for models applicable in safety-critical applications.
- **Challenges with SVGD**: Despite its theoretical promise, SVGD faced challenges in practical implementation. The generative loss term often dominated the total loss, leading to significantly negative values. This highlighted the necessity of adaptive tuning for the smoothing factor to balance the dual objectives effectively as discussed in Subsection 5.3.2.

7.3 Theoretical Implications

The findings from our empirical investigations have several theoretical implications:

- **Generative Loss Dynamics**: The behavior of the generative loss with SVGD underscores the importance of carefully tuning the smoothing factor α. An inappropriate α can lead to an unbounded decrease in the generative loss, disproportionately influencing the total loss.

- **Adaptive Scaling Necessity**: The necessity for adaptive scaling of the smoothing factor is evident. Fixed scaling factors do not suffice in maintaining equilibrium between generative and classification losses, especially in extended training scenarios. Moreover, they may not generalize for JEMs which can have different energy landscapes for different datasets.

- **Trade-off Between Computational Efficiency and Performance**: The trade-off between computational demands and performance gains of particle-based methods is critical. While particle-based methods like SVGD promise richer sample diversity, they require substantial computational resources, necessitating a balance between efficiency and accuracy.

7.4 Practical Implications

From a practical perspective, our research offers several valuable insights:

- **Hyperparameter Tuning**: The importance of meticulous hyperparameter tuning cannot be overstated. Parameters such as the number of particles, sampling steps, and sampler learning rate play a pivotal role in the performance of SVGD and require empirical validation through small-scale experiments before full-scale implementation.

- **Computational Resources**: The implementation of SVGD demands significant computational resources. Our experiments were constrained by available resources, limiting

the number of particles and sampling steps we could feasibly explore.

- **Stability with SAM**: SAM optimization emerges as a vital component in stabilizing JEMs. Its ability to mitigate instability and enhance performance underscores its practical value in training JEMs.

7.5 Final Thoughts

In conclusion, this thesis has made significant strides in understanding and enhancing JEMs through the integration of SVGD by rigorously exploring the theoretical and practical aspects of integrating particle-based sampling techniques within Joint Energy Models. While SGLD remains the superior sampler in our experiments, the insights gained from a comprehensive evaluation of SVGD highlight its potential especially for safety-critical applications. Moreover, our findings underscore the importance of continued exploration and refinement of sampling techniques to enhance the performance and stability of JEMs. Our contributions lay the groundwork for future advancements in EBMs, paving the way for more robust and efficient sampling techniques which in-turn would lead to well-calibrated models required for safety-critical tasks.

Bibliography

[1] M. Arbel, L. Zhou, and A. Gretton, "Generalized energy based models," in *International Conference on Learning Representations*, 2021.

[2] Y. Du and I. Mordatch, "Implicit generation and modeling with energy based models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] N. Liu, S. Li, Y. Du, J. Tenenbaum, and A. Torralba, "Learning to compose visual relations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23166–23178, 2021.

[4] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. S. Grathwohl, "Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc," in *International Conference on Machine Learning*, pp. 8489–8510, PMLR, 2023.

[5] T. Parshakova, J.-M. Andreoli, and M. Dymetman, "Distributional reinforcement learning for energy-based sequential models," *arXiv preprint arXiv:1912.08517*, 2019.

[6] W. Nie, A. Vahdat, and A. Anandkumar, "Controllable and compositional generation with latent-space energy-based models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13497–13510, 2021.

[7] T. Schröder, Z. Ou, J. Lim, Y. Li, S. Vollmer, and A. Duncan, "Energy discrepancies: a score-independent loss for energy-based models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] Q. Guo, C. Ma, Y. Jiang, Z. Yuan, Y. Yu, and P. Luo, "Egc: Image generation and classification via a diffusion energy-based model," in *IEEE/CVF International Conference on Computer Vision*, pp. 22952–22962, 2023.

[9] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," *arXiv preprint arXiv:1912.03263*, 2019.

[10] X. Yang and S. Ji, "Jem++: Improved techniques for training jem," in *IEEE/CVF International Conference on Computer Vision*, pp. 6494–6503, 2021.

[11] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, *et al.*, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[12] C.-C. Lin and A. McCarthy, "On the uncomputability of partition functions in energy-based sequence models," in *International Conference on Learning Representations (ICLR)*, 2022.

[13] X. Yang, Q. Su, and S. Ji, "Towards bridging the performance gaps of joint energy-based models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15732–15741, 2023.

[14] K. Makhtidi, A. Bustamam, and R. Adnan, "Training deep energy-based models through cyclical stochastic gradient langevin dynamics," in *International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 654–659, IEEE, 2022.

[15] M. Sustek, S. Sadhu, L. Burget, H. Hermansky, J. Villalba, L. Moro-Velazquez, and N. Dehak, "Stabilized training of joint energy-based models and their practical applications," *arXiv preprint arXiv:2303.04187*, 2023.

[16] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong, "You only propagate once: Accelerating adversarial training via maximal principle," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[18] K. A. Dubey, S. J Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing, "Variance reduction in stochastic gradient langevin dynamics," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[19] Q. Liu and D. Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[20] T. Pinder, C. Nemeth, and D. Leslie, "Stein variational gaussian processes," *arXiv preprint arXiv:2009.12141*, 2020.

[21] A. Korba, A. Salim, M. Arbel, G. Luise, and A. Gretton, "A non-asymptotic analysis for stein variational gradient descent," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4672–4682, 2020.

[22] Q. Liu, "Stein variational gradient descent as gradient flow," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] D. A. De Souza, D. Mesquita, S. Kaski, and L. Acerbi, "Parallel mcmc without embarrassing failures," in *International Conference on Artificial Intelligence and Statistics*, pp. 1786–1804, PMLR, 2022.

[24] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *International Conference on Machine Learning*, pp. 681–688, Citeseer, 2011.

[25] D. Zou, P. Xu, and Q. Gu, "Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling," in *Uncertainty in Artificial Intelligence*, pp. 1152–1162, PMLR, 2021.

[26] U. Simsekli, "Posterior sampling with stochastic gradient langevin dynamics,"

[27] M. Sensoy, L. Kaplan, F. Cerutti, and M. Saleki, "Uncertainty-aware deep classifiers using generative models," in *AAAI conference on Artificial Intelligence*, vol. 34, pp. 5620–5627, 2020.

[28] C. Gong, J. Peng, and Q. Liu, "Quantile stein variational gradient descent for batch bayesian optimization," in *International Conference on Machine Learning*, pp. 2347–2356, PMLR, 2019.

[29] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[30] X. Zhong, O. Gong, W. Huang, L. Li, and H. Xia, "Squeeze-and-excitation wide residual networks in image classification," in *IEEE International Conference on Image Processing (ICIP)*, 2019.

[31] J. Shi, Z. Li, S. Ying, C. Wang, Q. Liu, Q. Zhang, and P. Yan, "Mr image super-resolution via wide residual networks with fixed skip connection," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[32] P. González-Rodelas, H. M. Idais, M. Yasin, and M. Pasadas, "Optimal centers' allocation in smoothing or interpolating with radial basis functions," *Mathematics*, vol. 10, no. 1, p. 59, 2021.

[33] T. Liu, P. Ghosal, K. Balasubramanian, and N. Pillai, "Towards understanding the dy-

namics of gaussian-stein variational gradient descent," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[34] D. W. Robinson, "Entropy and uncertainty," *Entropy*, vol. 10, no. 4, pp. 493–506, 2008.

[35] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 (canadian institute for advanced research)," *URL http://www. cs. toronto. edu/kriz/cifar. html*, vol. 5, no. 4, p. 1, 2010.

[36] D. Bahri, H. Mobahi, and Y. Tay, "Sharpness-aware minimization improves language model generalization," in *Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2022.

[37] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, pp. 1321–1330, PMLR, 2017.