Report - Introduction to Generative models : LEARNING ENERGY-BASED MODELS BY DIFFUSION RECOVERY LIKELIHOOD (2021)

Léo Bernouin leo.bernouin@ensae.fr
Mehdi El Kacemy mehdi.elkacemy@ensae.fr
Côme Nadler come.nadler@ensae.fr

1 Introduction

Energy-Based Models (EBMs) are a powerful class of probabilistic models that parameterize a probability distribution $p_\theta(x)$ using an energy function $f_\theta(x)$:

$$p_\theta(x) = \frac{1}{Z_\theta} \exp(-f_\theta(x)),$$

where $Z_\theta = \int \exp(-f_\theta(x))\, dx$ is the partition function, a normalization constant that is generally intractable for high-dimensional data. These models are particularly useful for tasks such as image synthesis, modeling complex data distributions, and unsupervised learning. Unlike classical generative models such as GANs (Generative Adversarial Networks), EBMs offer a more direct approach to modeling the probability density, allowing us to leverage the structural properties of the data.

1.1 Motivations

Energy-Based Models (EBMs), despite their potential, face significant challenges that hinder their practical applicability. One of the most prominent issues is the high computational cost associated with maximizing the marginal likelihood $\log p_\theta(x)$. This task requires estimating the partition function Z_θ, which is computationally expensive and often infeasible for high-dimensional data. Furthermore, EBMs rely on Markov Chain Monte Carlo (MCMC) methods to generate samples, but these methods frequently encounter convergence issues. Non-convergent chains lead to samples that fail to accurately represent the true data distribution, thereby compromising the estimation of the model's parameters. Finally, traditional methods often struggle to effectively capture the complex structures inherent in high-dimensional data, limiting their ability to model intricate distributions robustly.

To address these limitations, researchers have explored alternatives such as diffusion probabilistic models and score-matching approaches. While these techniques provide promising directions, they do not directly tackle the explicit energy density modeling challenges posed by EBMs, leaving room for methodological innovation.

1.2 Objectives of the paper

The primary objective of this paper is to introduce an innovative method called *Diffusion Recovery Likelihood*, which is designed to overcome the aforementioned challenges in training EBMs. This approach builds on three foundational ideas. First, it employs a *data diffusion* process, which adds Gaussian noise to the original data. This step simplifies the learning process and makes the resulting distributions easier to sample. Second, the method shifts its focus from directly learning the marginal distribution $p_\theta(x)$ to modeling a sequence of conditional distributions $p_\theta(x|x_t)$, where x_t represents a noisy version of the original data x. Finally, the approach incorporates a *progressive sampling* strategy, which reconstructs high-quality synthetic data by chaining these conditional distributions, starting from Gaussian white noise.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1.3 Contributions of the paper

This paper makes three key contributions that advance the state of the art in Energy-Based Models. Methodologically, the work introduces the concept of recovery likelihoods, which maximizes the conditional probability of data at varying noise levels. This innovation addresses long-standing issues in the stability and tractability of EBM training. Additionally, the proposed method demonstrates competitive results, surpassing many GAN-based and score-based approaches in image generation quality. It achieves state-of-the-art FID scores across datasets such as CIFAR-10, CelebA, and LSUN. Finally, the method provides theoretical and practical robustness by ensuring that MCMC chains converge to realistic samples, even in long-term scenarios, a challenge that has plagued traditional EBMs.

2 Problem definition, methodology and theoretical analysis

2.1 Task definition

As said in introduction, Energy-Based Models (EBMs) approximate the true data distribution $p_{\text{data}}(x)$ by parameterizing a probability density function using an energy function $f_\theta(x)$:

$$p_\theta(x) = \frac{1}{Z_\theta} \exp(f_\theta(x)),$$

where $Z_\theta = \int \exp(f_\theta(x))dx$ is the partition function that normalizes the distribution. The goal is to learn parameters θ such that $p_\theta(x)$ closely approximates $p_{\text{data}}(x)$.

Training EBMs conventionally involves maximizing the log-likelihood, i.e :

$$L(\theta) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log p_\theta(x)].$$

However, the computation of Z_θ is intractable in high-dimensional data spaces, and MCMC methods often fail to generate accurate samples due to convergence issues. To address these challenges, the proposed method introduces a noise-diffusion process and focuses on modeling a sequence of conditional distributions $p_\theta(x|x_t)$.

2.2 Presentation of the method

The method relies on three main components : a noise-diffusion process (i), a recovery likelihood objective (ii), and a progressive sampling strategy (iii).

Diffusion process

The noise-diffusion process incrementally adds Gaussian noise to the data, transforming the original distribution into a simpler one. Mathematically, the process is defined as :

$$x_t = \sqrt{1-\sigma_t^2}\, x_{t-1} + \sigma_t\, \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I),$$

where σ_t controls the noise magnitude at step t. Starting from clean data $x_0 \sim p_{\text{data}}(x)$, the process generates a sequence of increasingly noisy observations $\{x_t\}_{t=0}^T$, culminating in Gaussian white noise $x_T \sim \mathcal{N}(0, I)$.

Recovery likelihood

To recover clean data from its noisy versions, the model learns conditional distributions $p_\theta(x|x_t)$, parameterized as :

$$p_\theta(x|x_t) = \frac{1}{Z_\theta(x_t)} \exp\left(f_\theta(x) - \frac{1}{2\sigma_t^2}\|x - x_t\|^2\right),$$

where $Z_\theta(x_t)$ is the normalization constant. Training is conducted by maximizing the recovery likelihood:

$$J(\theta) = \mathbb{E}_{x\sim p_{\text{data}}, \epsilon\sim\mathcal{N}(0,I)}[\log p_\theta(x|x_t)].$$

The training procedure is described in Algorithm 1 :

Algorithm 1 Training algorithm for Diffusion Recovery Likelihood

Input: $\{x_i\}_{i=1}^n$: Dataset of n samples
Input: $\{\sigma_t\}_{t=0}^T$: Noise schedule
Input: η: Learning rate
Output: Optimized model parameters θ

Initialize: Energy model parameters θ
for $t = 1$ *to* T **do**
  Step 1 : Add noise to data

$$x_t = \sqrt{1-\sigma_t^2}\, x + \sigma_t\, \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I)$$

  Step 2 : Generate synthesized samples
  Refine samples x_{syn} using Langevin dynamics (see Algorithm 2).
  Step 3 : Compute gradients for energy function

$$\nabla_\theta \left[f_\theta(x) - f_\theta(x_{\text{syn}}) \right]$$

  Step 4 : Update model parameters

$$\theta \leftarrow \theta + \eta \nabla_\theta$$

end

return Optimized parameters θ

Sampling with Langevin dynamics

Samples from $p_\theta(x|x_t)$ are refined using Langevin dynamics, which iteratively adjusts samples towards high-probability regions of the learned distribution. The update rule is given by :

$$x^{k+1} = x^k + \frac{\delta^2}{2}\left(\nabla_x f_\theta(x^k) + \frac{1}{\sigma_t^2}(x_t - x^k) \right) + \delta \epsilon^k,$$

where δ is the step size, and $\epsilon^k \sim \mathcal{N}(0, I)$. The sampling procedure is described in Algorithm 2.

Algorithm 2 Langevin dynamics for sampling

Input: x_t: Noisy data sample
Input: δ: Step size, K: Number of iterations
Output: Refined sample x^K

Initialize: $x^0 \leftarrow x_t$
for $k = 1$ *to* K **do**
  Step 1 : Compute gradient

$$g^k = \nabla_x f_\theta(x^k) + \frac{1}{\sigma_t^2}(x_t - x^k)$$

  Step 2 : Update sample

$$x^{k+1} = x^k + \frac{\delta^2}{2} g^k + \delta \epsilon^k, \quad \epsilon^k \sim \mathcal{N}(0, I)$$

end

return Refined sample x^K

Progressive sampling strategy

Once the model is trained, it generates new samples by progressively reversing the diffusion process. Starting from $x_T \sim \mathcal{N}(0, I)$, the model iteratively applies Langevin dynamics to reconstruct x_0. The full procedure is outlined in Algorithm 3 :

Algorithm 3 Progressive Sampling Procedure

Input: $\{\sigma_t\}_{t=1}^T$: Noise schedule, Trained energy model f_θ
Output: Generated sample x_0

Initialize: $x_T \sim \mathcal{N}(0, I)$
for $t = T$ *to* 1 **do**
  Apply Langevin dynamics (Algorithm 2) to refine x_{t-1} from x_t
end

return x_0

2.3 Theoretical justification and guarantees

The proposed method provides theoretical guarantees. First, maximizing the recovery likelihood $J(\theta)$ aligns with the original objective of maximizing the marginal likelihood $L(\theta)$, ensuring consistency with EBM principles. Second, as $\sigma_t^2 \to 0$, the conditional distributions $p_\theta(x|x_t)$ approach a Gaussian approximation :

$$p_\theta(x|x_t) \approx \mathcal{N}(x_t + \sigma_t^2 \nabla_x f_\theta(x_t), \sigma_t^2 I).$$

Finally, the localized structure of $p_\theta(x|x_t)$ ensures the stability of MCMC chains, enabling robust sampling for high-dimensional data.

3 Experimental evaluation

This section evaluates the proposed *Diffusion Recovery Likelihood* method based on its ability to generate high-quality samples and validate the theoretical guarantees established in previous sections. The evaluation focuses on both the quantitative and qualitative aspects of the generated samples and includes comparisons with state-of-the-art methods.

3.1 Methodology

The evaluation methodology involves three key aspects : the criteria used to assess the generated samples, the datasets on which the experiments were conducted, and the experimental procedures adopted for training and evaluation.

To assess the quality of the generated samples, two widely-used metrics were employed. The first is the *Frechet Inception Distance* (FID), which quantifies the similarity between the distributions of real and generated data. FID is computed as the Wasserstein-2 distance between two multivariate Gaussian distributions, parameterized by their means $\mu_{\text{real}}, \mu_{\text{gen}}$ and covariances $\Sigma_{\text{real}}, \Sigma_{\text{gen}}$:

$$\text{FID} = \|\mu_{\text{real}} - \mu_{\text{gen}}\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{1/2}).$$

A lower FID score indicates a closer alignment between the real and generated distributions, and thus, higher-quality samples.

The second metric used is the *Inception Score* (IS), which evaluates both the diversity and quality of generated samples. IS is defined as:

$$\text{IS} = \exp\left(\mathbb{E}_{x \sim p_{\text{gen}}}\left[D_{\text{KL}}(p(y|x)\|p(y))\right]\right),$$

where $p(y|x)$ is the conditional label distribution of an image x as predicted by a pre-trained classifier, and $p(y)$ is the marginal distribution of labels. High IS values reflect high diversity and classifiability of the generated images.

The experiments were conducted on three benchmark datasets. The first is **CIFAR-10**, which contains 50,000 training images and 10,000 test images, all resized to 32×32. This dataset provides a standard benchmark for evaluating generative models. The second dataset, **CelebA**, consists of high-quality human face images resized to 64×64, and the third dataset, **LSUN**, contains high-resolution images (64×64 and 128×128) of churches and bedrooms, which allow for an assessment of the model's performance on complex, structured data.

The experimental procedure consisted of three stages. During training, the models were optimized using the conditional likelihood framework described in Section 2. The key hyperparameters, such

as the noise levels σ_t and the number of Langevin steps K, were carefully tuned to achieve optimal performance. Once trained, the models generated samples by progressively reversing the diffusion process, starting from Gaussian noise sampled from $\mathcal{N}(0, I)$. The FID scores were computed on 50,000 generated samples for a quantitative evaluation. Finally, the generated results were compared to those produced by state-of-the-art models, including GAN-based approaches (e.g., StyleGAN2) and score-based models (e.g., NCSN, DDPM).

3.2 Results

The quantitative results demonstrate the effectiveness of the proposed method in generating high-quality and diverse samples across all datasets.

On the **CIFAR-10** dataset, the proposed method achieves an FID score of 9.58, which is competitive with or superior to most existing methods. The Inception Score (IS) achieved is 8.30, reflecting both the diversity and the fidelity of the generated samples. For **CelebA**, the method produces an FID score of 5.98, which is comparable to state-of-the-art GANs. On the high-resolution **LSUN** dataset, the generated images exhibit remarkable fidelity, with coherent textures and well-preserved structural details, even at resolutions of 128×128.

The following table summarizes the comparative results on CIFAR-10, highlighting the competitiveness of the proposed approach in terms of both FID and IS :

Model	**FID** $\downarrow$	**IS** $\uparrow$
WGAN-GP	36.4	7.86 ± 0.07
SNGAN-DDLS	15.42	9.09 ± 0.10
StyleGAN2-ADA	3.26	9.74 ± 0.05
Ours (T=6, K=30)	**9.58**	$\mathbf{8.30 \pm 0.11}$

Table 1: Comparison of FID and IS scores on CIFAR-10.

Additionally, the stability of the MCMC chains was evaluated by running up to 100,000 steps during sampling. The results confirm that the generated samples remain realistic and that FID scores do not degrade significantly, validating the stability of the learned energy potentials over long runs.

3.3 Discussion

The experimental results validate the effectiveness of the proposed method across multiple dimensions. First, in terms of competitive quality, the proposed approach achieves state-of-the-art results on FID and IS scores when compared to explicit EBMs and many GAN-based models. This highlights its ability to generate samples that are both realistic and diverse. Second, the method demonstrates stability in MCMC sampling, overcoming a key limitation of traditional EBMs. Unlike existing methods, the proposed approach ensures that long-run MCMC chains remain stable, even for complex datasets, which is critical for generating high-quality samples in practice.

Despite these strengths, the method has certain limitations. While the FID scores are competitive, they fall short of GANs like StyleGAN2 on some datasets. Moreover, the computational cost remains high due to the iterative nature of Langevin sampling, which could limit its scalability for extremely large datasets or resolutions.

In conclusion, the experimental evaluation demonstrates that the proposed *Diffusion Recovery Likelihood* method effectively addresses the core challenges of EBM training. Its performance on benchmark datasets highlights its potential as a robust alternative to traditional generative models.