AniLA: Anisotropic Langevin Dynamics for training Energy-Based Models

Belhal Karimi, Jianwen Xie, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA

Abstract

We develop in this paper

1 Introduction

The modeling of a data generating process is critical for many tasks. A growing interest in generative models within the realm of computer vision has led to multiple interesting solutions. In particular, Energy Based Models (EBM) [Zhu et al., 1998, LeCun et al., 2006], are a class of generative models that learns high dimensional and complex (in terms of landscape) representation/distribution of the input data. Since inception, EBMs have been used in several applications including computer vision [Ngiam et al., 2011, Xie et al., 2016, 2020, Du and Mordatch, 2019], natural language processing [Mikolov et al., 2013, Deng et al., 2020], density estimation [Wenliang et al., 2019, Song et al., 2020] and reinforcement learning [Haarnoja et al., 2017].

Formally, EBMs are built upon an unnormalized log probability, called the energy function, that is not required to sum to one, as standard log probability functions. This noticeable feature allows for more freedom in the way one parametrizes the EBM. For instance, Convolutional Neural Network (CNN) can be employed to parametrize the energy function, see [Xie et al., 2016]. Note that this choice is highly related to the type of the input data, as mentioned in [Song and Kingma, 2021].

The training procedure of such models consists of finding an energy function that assigns to lower energies to observations than unobserved points. This phase can be casted into an optimization task and several ways are possible to achieve it. In this paper, we will focus on training the EBM via Maximum Likelihood Estimation (MLE) and defer the readers to [Song and Kingma, 2021] for alternative procedures. Particularly, while using MLE to fit the EBM on a stream of observed data, the high non-convexity of the loss function leads to a non closed form maximization step. In general, gradient based optimization methods are thus used during that phase. Besides, given the intractability of the normalizing constant of our model, the aforementioned gradient, which is an intractable integral, needs to be approximated. A popular and efficient way to conduct such approximation is to use Monte Carlo approximation where the samples are obtained via Markov Chain Monte Carlo (MCMC).

2 On MCMC based Energy Based Models

Given a stream of input data noted $x \in \mathbb{R}^p$, the energy-based model (EBM) is a Gibbs distribution defined as follows:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp(f_\theta(x)) \tag{1}$$

where $\theta \in \mathbb{R}^d$ denotes the global vector parameters of our model and $Z(\theta) := \int_x \exp(f_\theta(x)) \mathrm{d}x$ is the normalizing constant (with respect to x).

Energy Based Models: Energy based models LeCun et al. [2006], Ngiam et al. [2011] are a class of generative models that leverages the power of Gibbs potential and high dimensional sampling techniques to produce high quality synthetic image samples. Training of such models occurs via Maximum Likelihood (ML).

MCMC procedures: MCMC are a class of inference algorithms

3 Gradient Informed Langevin Diffusion

3.1 Preliminaries and Bottlenecks of Langevin MCMC based EBM

State of the art MCMC sampling algorithm, particularly used during the training procedure of EBMs, is the discretized Langevin diffusion, casted as Stochastic Gradient Langevin Dynamics (SGLD), see Welling and Teh [2011].

3.2 Curvature informed MCMC

We introduce a new sampler based on the Langevin updates presented above.

Algorithm 1 StAnLeyfor Energy-Based Model

1: **Input**: Total number of iterations T, number of MCMC transitions K and of samples M learning rate η, initial values θ_0, initial chain states $\{z_0^m\}_{m=1}^M$ and n observations $\{x_i\}_{i=1}^n$.
2: **for** $t = 1$ to T **do**
3: Compute the anisotropic stepsize as follows:

$$\gamma_t = \frac{b}{\max(b, |\nabla f_{\theta_t}(z_{t-1}^m)|} \tag{2}$$

4: Draw m samples $\{z_t^m\}_{m=1}^M$ from the objective potential (1) via Langevin diffusion:

$$z_t^m = z_{t-1}^m + \gamma_t/2 \nabla f_{\theta_t}(z_{t-1}^m) + \sqrt{\gamma} \mathsf{B}_t \tag{3}$$

where B_t is the brownian motion, drawn from a Normal distribution.
5: Samples m positive observations $\{x_i\}_{i=1}^m$ from the empirical data distribution.
6: Compute the gradient of the empirical log-EBM (1) as follows:

$$\nabla \sum_{i=1}^m \log p_{\theta_t}(x_i) = \mathbb{E}_{p_{\text{data}}}\left[\nabla_\theta f_{\theta_t}(x)\right] - \mathbb{E}_{p_\theta}\left[\nabla_{\theta_t} f_\theta(z_t^m)\right] \approx \frac{1}{m}\sum_{i=1}^m \nabla_\theta f_{\theta_t}(x_i) - \frac{1}{m}\sum_{i=1}^m \nabla_\theta f_{\theta_t}(z_t^m) \tag{4}$$

7: Update the vector of global parameters of the EBM:

$$\theta_{t+1} = \theta_{t+1} + \eta \nabla \sum_{i=1}^m \log p_{\theta_t}(x_i) \tag{5}$$

8: **end for**
9: **Output:** Generated samples $\{z_T^m\}_{m=1}^M$

4 Geometric ergodicity of AniLA sampler

We will present in this section, our theoretical analysis for the Markov Chain constructed using Line 3-4.

Let Θ be a subset of $\mathbb{R}^d$ for some integer $d > 0$. We denote by $\mathcal{Z}$ the measurable space of $\mathbb{R}^\ell$ for some integer $\ell > 0$. We define a family of stationary distribution $(\pi_\theta(z))_{\theta\in\Theta}$, probability density functions with

respect to the Lebesgue measure on the measurable space $\mathcal{Z}$. This family of p.d.f. defines the stationary distributions of our newly introduced sampler.

Important Note: The stationary distributions are defined per $\theta \in \Theta$, *i.e.,*at each model update during the EBM optimization phase.

For any chain state $z \in \mathcal{Z}$ we denote by $\Pi_\theta(z, \cdot)$ the transition kernel as defined in the StAnLeyupdate in Line 4.

The objective of this section is to rigorously show that each transition kernel π_θ is uniformly geometrically ergodic and that this result is true uniformly in state s on any compact subset $\mathcal{C} \in \mathcal{Z}$. As a background note, a Markov chain, as built Line 4, is said to be geometrically ergodic when k iterations of the same transition kernel is converging to the stationary distribution of the chain and this convergence as a geometric dependence on k.

We begin with several usual assumptions for such results. The first one is related to the continuity of the gradient of the log posterior distribution and the unit vector pointing in the direction of the sample z and the unit vector pointing in the direction of the gradient of the log posterior distribution at z:

H 1. *(Continuity) The stationary distribution is positive and has continuous derivative such that for all* $\theta \in \mathbb{R}^d$*:*

$$\lim_{z\to\infty} \frac{z}{|z|} \nabla f_\theta(z) = -\infty \quad \textit{and} \quad \limsup_{z\to\infty} \frac{z}{|z|} \frac{\nabla f_\theta(z)}{|\nabla f_\theta(z)|} < 0 \tag{6}$$

We assume also some regularity conditions of the stationary distributions with respect to state s:

H2. *For all* $z \in \mathcal{Z}$*,* $\theta \to \pi_\theta$ *and* $\theta \to \nabla \log \pi_\theta$ *are continuous on* Θ*.*

For a positive and finite function noted $V : \mathcal{Z} \mapsto \mathbb{R}$, we define the V-norm distance between two arbitrary transition kernels Π_1 and Π_2 as follows:

$$\|\Pi_1 - \Pi_2\|_V := \sup_{z\in\mathcal{Z}} \frac{\|\Pi_1(z, \cdot) - \Pi_2(z, \cdot)\|_V}{V(z)} \tag{7}$$

The definition of this norm will allow us to establish a convergence rate for our sampling method by deriving an upper bound of the quantity $\|\Pi_\theta^k - \pi_\theta\|_V$ where k denotes the number of MCMC transitions. We also recall that Π_θ is the transition kernel defined by Line 4 and π_θ is the stationary distribution of our Markov chain. Then, this quantity characterizes how close to the target distribution, our chain is getting after a finite time of iterations and will eventually formalize *V-uniform ergodicity* of our method. We specify that strictly speaking π_θ is a probability measure, and not a transition kernel. However $\|\Pi_\theta^k - \pi_\theta\|_V$ is well-defined if we consider the the probability π_θ as a kernel by making the definition:

$$\pi(z, \mathcal{C}) := \pi(\mathcal{C}) \quad \text{for} \quad \mathcal{C} \in \mathcal{Z}, \quad z \in \mathcal{Z} \tag{8}$$

Here, for some $\beta \in]0, 1[$ we define the V function for all $z \in \mathcal{Z}$ as follows:

$$V_\theta(z) = c_\theta \pi_\theta(z)^{-\beta} \tag{9}$$

where c_θ is a constant, with respect to the chain state s, such that for all $z \in \mathcal{Z}$, $V_\theta(z) \geq 1$. Again, we note that the V norm is, in our case, function of the chain state noted z *and* of the global model parameter θ, estimated, and thus varying, through the optimization procedure. The convergence rate will thus be given for a particular model estimate (the supremum in fact). Define $V_1(z) := \inf_{\theta\in\Theta} V_\theta(z)$ and $V_2(z) := \sup_{\theta\in\Theta} V_\theta(z)$ and assume that:

H3. *There exists a constant* $a_0 > 0$ *such that for all* $\theta \in \Theta$ *and* $z \in \mathcal{Z}$*,* $V_2(z)$ *is integrable against the kernel* $\Pi_\theta(z, \cdot)$ *and*

$$\limsup_{a\to 0} \sup_{\theta\in\Theta, z\in\mathcal{Z}} \Pi_\theta V_2^a(z) = 1 \tag{10}$$

We will now give the main convergence result of our sampling method in StAnLey. The result consists of showing V-uniform ergodicity of the chain, the irreducibility of the transition kernels and their aperiodicity, ssee Meyn and Tweedie [2012] for more details. We also prove a drift condition which states that the transition kernels tend to bring back elements into a small set from which boils down V-uniform ergodicity of the transition kernels $(\Pi_\theta)_{\theta\in\Theta}$.

Theorem 1. *Assume H1-H3.*

5 Numerical Experiments

5.1 Application on Toy Example: Gaussian Mixture Model

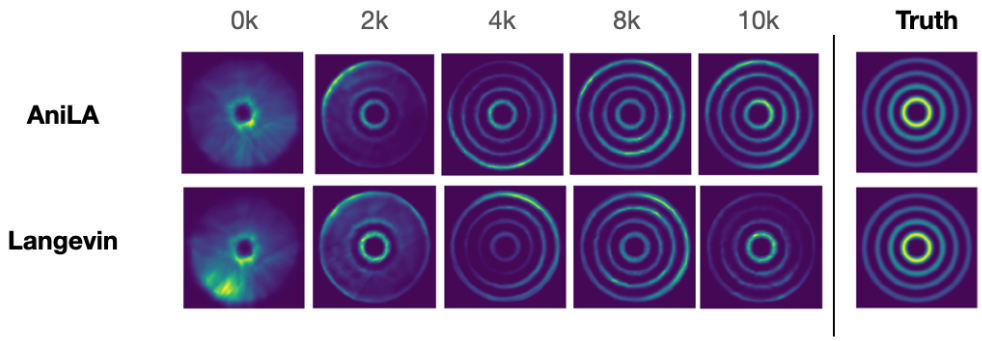

Figure 1: (Rings Toy Dataset)

5.2 Flowers Dataset

Figure 2: (Flowers Dataset). Left: Langevin Method. Right: AniLA method. After 100k iterations.

5.3 CIFAR Dataset

Figure 3: (CIFAR Dataset). Left: Langevin Method. Right: AniLA method. After 100k iterations.

6 Conclusion

References

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. 2019.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*, 2013.

Jiquan Ngiam, Zhenghao Chen, Pang W Koh, and Andrew Y Ng. Learning deep energy models. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1105–1112, 2011.

Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*, pages 574–584. PMLR, 2020.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

Li Wenliang, Dougal Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR, 2016.

Jianwen Xie, Zilong Zheng, Ruiqi Gao, Wenguan Wang, Song-Chun Zhu, and Ying Nian Wu. Generative voxelnet: Learning energy-based models for 3d shape synthesis and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998.