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) [Meyn and Tweedie, 2012]. The goal of this embedded MCMC procedure while training the Energy-based model is to synthesize new examples of the input data and use those new synthetic observations, in our case images, to approximate some expectations that we will describe later. The sampling phase is thus crucial for both the EBM training speed and its final accuracy in generating new samples.

The computational burden of those MCMC transitions at each iteration of the EBM training procedure is alleviated via different techniques in the literature. For instance, in [Nijkamp et al., 2019], the authors develop a short-run MCMC as a flow-based generator mechanism despite its non convergence property. A large class of solutions aiming at reducing the cost of running MCMC until convergence, which in practice can be unfeasible, is using Contrastive Divergence [Hinton, 2002] and persistent Contrastive Divergence [Tieleman, 2008]. This principled approach keeps in memory the final chain state under the previous global

model parameter and uses it as the initialization of the current chain. The heuristic of such approach is that along the EBM iterations, the conditional distributions, depending on the model parameter, are more and more similar and thus using a good sample from the previous chain is in general a good sample of the current one. Though, this method can be limited during the first iterations of the EBM training since when the model parameter changes drastically, the conditional distributions do too and samples from two different chains can be quite inconsistent. Several extensions varying the way the chain is initialized can be found in [Welling and Hinton, 2002, Gao et al., 2018, Du and Mordatch, 2019].

An interesting line of work in the realm of MCMC-based EBM tackles the biases induced by stopping the MCMC runs too early. Indeed, it is known, see [Meyn and Tweedie, 2012], that before convergence, MCMC samples are biased and thus correcting this biase while keep a short run and less expensive run is an appealing option. Several contributions aiming at removing this bias for improved MCMC training include coupling MCMC chains, see [Qiu et al., 2019, Jacob et al., 2020] or by simply estimating this bias and correct the chain afterwards, see [Du et al., 2020].

In this work, we consider the case of a short-run MCMC for the training of an Energy-Based Model but rather than focussing on debiasing the chain, we develop a new sampling scheme which purpose is to obtain better samples from the target distribution using less MCMC transitions. We consider that the shape of the target distribution, which highly inspires our proposed method, is of utmost importance to obtain such negative samples.

The contributions of our paper are as follows:

- We develop StAnLey, a EBM training method that embeds a newly proposed *convergent* and *efficient* MCMC sampling scheme, focussing on curvature informed metrics of the target distribution one wants to obtain samples from.
- Based on a anisotropic stepsize, our method, which is an improvement of the Langevin dynamics, achieves to obtain negative samples from the EBM data distribution.
- We prove the geometric ergodicity uniformly on any compact set of our method assuming some regularity conditions on the target distribution
- We empirically verify the relevance of our method on several image generation tasks.

The rest of the paper is organized as follows. We introduce in Section 2 the important notions of this paper regarding EBM and MCMC procedures. Section 3 develops the main algorithmic contribution of this paper, namely StAnLey. Section 4 introduces the main theoretical results of our paper and focuses on the ergodicity of our propose MCMC sampling method. Section 5 present several image generation experiments on a toy dataset and baseline deep image datasets. Section 6 concludes our work

2 On MCMC based Energy Based Models

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

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

where $\theta \in \Theta \subset \mathbb{R}^d$ 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). The natural way of fitting model (1) is to employ Maximum Likelihood Estimation (MLE) to maximize the marginal likelihood $p(\theta)$ and consisting of finding the vector of parameters θ^* such that for any $x \in \mathcal{X}$,

$$\theta^* = \arg\max_{\theta \in \Theta} \log p(\theta) \ . \tag{2}$$

The quantity of interest $p(\theta)$ is obtained by marginalizing over the input data $x \in \mathcal{X}$ and formally reads $p(\theta) = \int_{x \in xset} p(x, \theta) q(x) \mathrm{d}x$ where we note $q(x)$ the true distribution of the input data x. The optimization task (2) is not tractable in closed form and requires an iterative procedure to be solved. The standard algorithm used to train EBMs is Stochastic Gradient Descent (SGD), see [Robbins and Monro, 1951, Bottou

and Bousquet, 2008]. SGD requires having access to the gradient of the objective function $\log p(\theta)$. This latter requires computing an intractable, due to the high nonlinearity of the parametrized model we use in general $f_\theta(x)$. Given the general form in (1) we have that:

$$\nabla \log p(\theta) = \int_{x \in xset} \nabla \log p(x, \theta) q(x) \mathrm{d}x = \mathbb{E}_{p(x,\theta)}[\nabla_\theta f_\theta(x)] - \mathbb{E}_{q(x)}[\nabla_\theta f_\theta(x)] \,, \tag{3}$$

and a simple Monte Carlo approximation of $\nabla \log p(\theta)$ yields

$$\nabla \log p(\theta) \approx \frac{1}{m} \sum_{j=1}^{m} \nabla_\theta f_\theta(x_j^p) - \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i^q) \,, \tag{4}$$

where are $\{x_j^p\}_{j=1}^m$ samples obtained from the EBM $p(x, \theta)$ and $\{x_i^q\}_{i=1}^n$ are samples obtained from the true data distribution $q(x)$.

While drawing samples from the data distribution is trivial, the challenge during the EBM training phase is to obtain good samples from the EBM distribution $p(x, \theta)$ for any model parameter $\theta \in \Theta$. This task is generally done using MCMC methods. State of the arts MCMC used in the EBM literature include Langevin dynamics, see [Grenander and Miller, 1994, Roberts et al., 1996] and Hamiltonian Monte Carlo (HMC), see [Neal et al., 2011]. Those methods are detailed in the sequel and are important concepts throughout our paper.

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).

TO COMPLETE

MCMC procedures: MCMC are a class of inference algorithms

TO COMPLETE, MCMC, Metropolis methods, detail Langevin Dynamics as baseline for our paper, HMC etc

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].

TO COMPLETE with disadvantage of vanilla Langevin

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:

$$\gamma_t = \frac{b}{\max(b, |\nabla f_{\theta_t}(z_{t-1}^m)|} \tag{5}$$

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

$$z_t^m = z_{t-1}^m + \gamma_t/2 \nabla f_{\theta_t}(z_{t-1}^m) + \sqrt{\gamma}\mathsf{B}_t \tag{6}$$

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:

$$\nabla \sum_{i=1}^m \log p_{\theta_t}(x_i) = \mathbb{E}_{p_{\text{data}}}\left[\nabla_\theta f_{\theta_t}(x)\right] - \mathbb{E}_{p_\theta}\left[\nabla_{\theta_t} f_\theta(z_t^m)\right] \approx \frac{1}{m}\sum_{i=1}^m \nabla_\theta f_{\theta_t}\left(x_i\right) - \frac{1}{m}\sum_{i=1}^m \nabla_\theta f_{\theta_t}\left(z_t^m\right) \tag{7}$$

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

$$\theta_{t+1} = \theta_{t+1} + \eta \nabla \sum_{i=1}^m \log p_{\theta_t}(x_i) \tag{8}$$

8: **end for**

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

TO COMPLETE with heuristic on why informing the stepsize with shape of the target is good.

Some modifications have been proposed in particular to optimize the covariance matrix of the proposal in order to better stride the support of the target distribution. In [Atchadé, 2006, Marshall and Roberts, 2012], the authors propose to construct adaptive Langevin chains for which they prove the geometric ergodicity of the chain uniformly on any compact subset of its parameters. Unfortunately, this technique does not take the whole advantage of changing the proposal using the target distribution (which is important and makes it an efficient method since for each model parameter through the EBM training iteration, the target moves) . In particular, the covariance matrix of the proposal is given by a stochastic approximation of the empirical covariance matrix. This choice seems completely relevant as soon as the convergence toward the stationary distribution is reached, in other words it would make sense towards the end of the EBM training). However, it does not provide a good guess of the variability during the first iterations of the chain since it is still very dependent on the initialization. This leads to chains that may be numerically trapped.

Moreover, if we consider the constant curvature simplification suggested in [Girolami and Calderhead, 2011], one still needs to invert the metric which may be neither explicit nor computationally tractable. Note that these constrains are common with other application fields such as pharmacology, where models are often complex and especially in convent based EBM as we are dealing with in this paper.

Therefore, we propose to sample from a proposal distribution using a full anisotropic covariance matrix based on the anisotropy and correlations of the target distribution. The expectation is obtained as the sum of the current iterate plus a drift which is proportional to the gradient of the logarithm of the target distribution.

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* $\theta \in \mathbb{R}^d$*:*

$$\lim_{z\to\infty} \frac{z}{|z|}\nabla f_\theta(z) = -\infty \quad \textit{and} \quad \limsup_{z\to\infty} \frac{z}{|z|}\frac{\nabla f_\theta(z)}{|\nabla f_\theta(z)|} < 0 \tag{9}$$

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:

$$\|\Pi_1 - \Pi_2\|_V := \sup_{z\in\mathcal{Z}} \frac{\|\Pi_1(z, \cdot) - \Pi_2(z, \cdot)\|_V}{V(z)} \tag{10}$$

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:

$$\pi(z, \mathcal{C}) := \pi(\mathcal{C}) \quad \text{for} \quad \mathcal{C} \in \mathcal{Z}, \quad z \in \mathcal{Z} \tag{11}$$

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

$$V_\theta(z) = c_\theta \pi_\theta(z)^{-\beta} \tag{12}$$

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*

$$\limsup_{a\to 0} \sup_{\theta\in\Theta, z\in\mathcal{Z}} \Pi_\theta V_2^a(z) = 1 \tag{13}$$

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, see 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

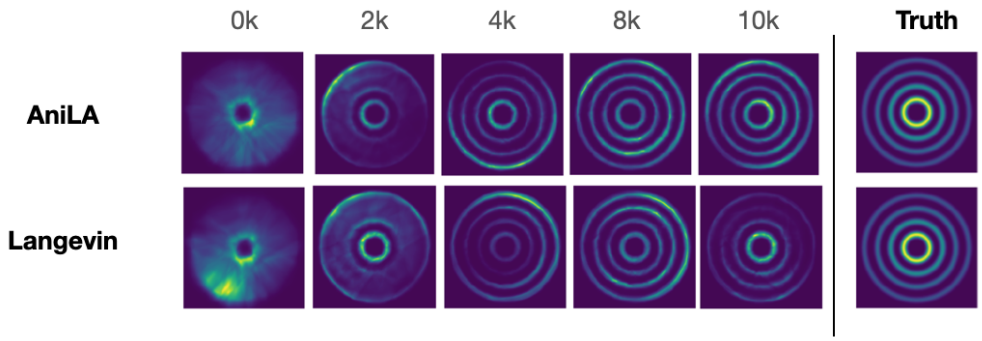

Figure 1: (Rings Toy Dataset)

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

Yves F Atchadé. An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodology and Computing in applied Probability*, 8(2):235–254, 2006.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc., 2008.

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.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020.

Ruiqi Gao, Yang Lu, Junpei Zhou, Song-Chun Zhu, and Ying Nian Wu. Learning generative convnets via multi-grid modeling and sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9155–9164, 2018.

Mark Girolami and Ben Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

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.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Pierre E Jacob, John O Leary, and Yves F Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):543–600, 2020.

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

Tristan Marshall and Gareth Roberts. An adaptive approach to langevin mcmc. *Statistics and Computing*, 22(5):1041–1057, 2012.

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.

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

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.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *arXiv preprint arXiv:1904.09770*, 2019.

Yixuan Qiu, Lingsong Zhang, and Xiao Wang. Unbiased contrastive divergence algorithm for training energy-based latent variable models. In *International Conference on Learning Representations*, 2019.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22: 400–407, 1951.

Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

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.

Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pages 1064–1071, 2008.

Max Welling and Geoffrey E Hinton. A new learning algorithm for mean field boltzmann machines. In *International Conference on Artificial Neural Networks*, pages 351–357. Springer, 2002.

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.