STANLEY: Stochastic gradienT ANisotropic LangEvin dYnamics for learning Energy-Based models

Anonymous ICCV submission

Paper ID 8604

Abstract

We propose in this paper, STANLEY*, a* ***S****tochastic gradien****T*** ***AN****isotropic* ***L****ang****E****vin d****Y****namics, for sampling high dimensional data. With the growing efficacy and potential of Energy-Based modeling, also known as non-normalized probabilistic modeling, for modeling a generative process of different natures of high dimensional data observations, we present an end-to-end learning algorithm for Energy-Based models (EBM) with the purpose of improving the quality of the resulting sampled data points. While the unknown normalizing constant of EBMs makes the training procedure intractable, resorting to Markov Chain Monte Carlo (MCMC) is in general a viable option. Realizing what MCMC entails for the EBM training, we propose in this paper, a novel high dimensional sampling method, based on an anisotropic stepsize and a gradient-informed covariance matrix, embedded into a discretized Langevin diffusion. We motivate the necessity for an anisotropic update of the negative samples in the Markov Chain by the nonlinearity of the backbone of the EBM, here a Convolutional Neural Network. Our resulting method, namely* STANLEY*, is an optimization algorithm for training Energy-Based models via our newly introduced MCMC method. We provide a theoretical understanding of our sampling scheme by proving that the sampler leads to a geometrically uniformly ergodic Markov Chain. Several image generation experiments are provided in our paper to show the effectiveness of our method.*

1. Introduction

The modeling of a data generating process is critical for many modern learning 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) [58, 30], 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 [35, 56, 57, 11], natural language processing [33, 7], density estimation [54, 47] and reinforcement learning [21].

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 [56]. Note that this choice is highly related to the type of the input data, as mentioned in [48].

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 cast 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 [48] 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) [32]. 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 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 [37], the authors develop a short-run MCMC as a flow-based generator mechanism despite its non conver-

gence 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 [24] and persistent Contrastive Divergence [50]. 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 change too, and samples from two different chains can be quite inconsistent. Several extensions varying the way the chain is initialized can be found in [52, 13, 11].

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 [32], that before convergence, MCMC samples are biased and thus correcting this bias while keeping a short and less expensive run is an appealing option. Several contributions aiming at removing this bias for improved MCMC training include coupling MCMC chains, see [39, 25] or simply estimating this bias and correct the chain afterwards, see [9].

In this work, we consider the case of a short-run MCMC for the training of an Energy-Based model. Rather than focusing 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, an Energy-Based model training method that embeds a newly proposed *convergent* and *efficient* MCMC sampling scheme, focusing on curvature informed metrics of the target distribution one wants to sample from.
- Based on an anisotropic stepsize, our method, which is an improvement of the Langevin Dynamics, achieves to obtain negative samples from the Energy-Based model data distribution and improves the overall optimization algorithm.
- We prove the geometric ergodicity uniformly on any compact set of our MCMC method assuming some regularity conditions on the target distribution and on the backbone of our EBM.
- We empirically assess the effectiveness of our method on several image generation tasks, both on synthetic and real datasets including the Oxford Flowers 102 dataset and CIFAR-10.

The rest of the paper is organized as follows. We introduce 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. The complete proofs of our theoretical results can be found in the supplementary material of this paper.

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 parameters vector 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) maximizing the marginal likelihood $p(\theta)$, *i.e.*,finding the vector θ^* such that for any $x \in \mathcal{X}$,

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

where the quantity of interest $p(\theta)$ is obtained by marginalizing over the input data $x \in \mathcal{X}$ and formally reads $p(\theta) := \int_{x \in \mathcal{X}} p(x, \theta) q(x) \mathrm{d}x$. We denote by $q(x)$ the true distribution of the input data x. The optimization task (2) is not tractable in closed form and requires an iterative procedure in order to be solved. The standard algorithm used to train EBMs is Stochastic Gradient Descent (SGD), see [40, 3]. SGD requires having access to the gradient of the objective function $\log p(\theta)$. This latter requires computing an intractable integral, due to the high nonlinearity of the generally utilized parameterized model $f_\theta(x)$. Given the general form defined in (1) we have that:

$$\begin{aligned} \nabla \log p(\theta) &= \int_{x \in \mathcal{X}} \nabla \log p(x, \theta) q(x) \mathrm{d}x \\ &= \mathbb{E}_{p(x,\theta)}[\nabla_\theta f_\theta(x)] - \mathbb{E}_{q(x)}[\nabla_\theta f_\theta(x)] \ , \end{aligned}$$

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

$$\nabla \log p(\theta) \approx \frac{1}{m} \sum_{j=1}^{m} \nabla_\theta f_\theta(x_j^p) - \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i^q) \ , \tag{3}$$

where $\{x_j^p\}_{j=1}^m$ are samples obtained from the EBM $p(x,\theta)$ and $\{x_i^q\}_{i=1}^n$ are drawn uniformly 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-art MCMC used in the EBM literature include Langevin Dynamics, see [19, 45] and Hamiltonian Monte Carlo (HMC), see [34]. Those methods are detailed in the sequel and are important concepts used throughout our paper.

Energy Based Models: Energy based models [30, 35] 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. Just as Variational Autotencoders (VAE) [28] or Generative Adversarial Networks (GAN) [17], EBMs are powerful tools for generative modeling tasks, as a building block for a wide variety of tasks. The main purpose of EBMs is to learn an energy function (1) that assigns low energy to a stream of observation and high energy values to other inputs. Learning, or Training, of such models is done via Maximum Likelihood (ML) [11] or Score Matching [49] or Noise Constrastive Estimation [14]. In several general applications, authors leverage the power of EBMs for classification purposes as in [18], deep regression [20] and also in Reinforcement Learning where [22] develop an energy-based optimal policy where the parameters of that energy function are provided by the reward of the overall system. Yet, unlike VAE or GANn Energy-Based models enjoy from a single structure requiring training (versus several networks) resulting in more stability. The use of implicit sampling techniques, such as MCMC, as detailed in the sequel, allows more flexibility trading of quality for computation time. Overall, the *implicit* property of the EBM, seen as an energy function, makes it a tool of choice as opposed to *explicit* generators that are limited to some design choice (see prior choices for VAE or both networks design in GAN).

MCMC procedures: Whether for sampling from a posterior distribution [41], or in general intractable likelihoods [8], several inference methods are available. Approximate inference is a partial solution to the inference problem and include techniques such as Variational Inference (VI) [51, 6] or Laplace Approximation [55, 46]. Those methods allow simplification of the intractable quantities and result in the collection of "good enough" samples. As seen in (3), training an EBM requires obtaining samples from the model itself. Given the nonconvexity of the structural model $f_\theta(\cdot)$ with respect to the model parameter θ, direct sampling is not an option. Besides, in order to update the model parameter θ, usually through gradient descent type of methods [3], exact samples from the EBM are needed in order to compute a good approximation of its (intractable) gradient, see (3). To do so, we generally have recourse to MCMC methods. MCMC are a class of inference algorithms that provide a principled iterative approach to obtain samples from any intractable distribution. While being exact, the samples generally represent a larger computation burden than methods such as VI. Increasing the efficiency of MCMC methods, by obtaining exact samples, in other words constructing a chain that converges faster, in fewer transitions is thus of utmost importance in the context of optimizing our EBM. Several attempts have been proposed for the standalone task of posterior sampling through the use of Langevin diffusion, see Unadjusted Langevin in [4] of MALA algorithm in [42, 44, 12] or leveraging Hamiltonian Dynamics, see HMC in [15]. We propose in the next section, an improvement of the Langevin diffusion with the ultimate goal of speeding the EBM training procedure. Our method includes this latter improvement in an end-to-end learning algorithms for Energy-Based models.

3. Gradient Informed Langevin Diffusion

We now introduce the main algorithmic contribution of our paper, namely STANLEY. STANLEY is a learning algorithm for EBMs, comprised of a novel MCMC method for sampling negative samples from the intractable model. We provide theoretical guarantees of our scheme in Section 4.

3.1. Preliminaries on Langevin MCMC based EBM

State of the art MCMC sampling algorithm, particularly used during the training procedure of EBMs, is the discretized Langevin diffusion, cast as Stochastic Gradient Langevin Dynamics (SGLD), see [53]. In particular, several applications using EBM and SGLD have thrived in image generation, Natural Language Processing or even biology [10]. Yet, the choice of the proposal, generally Gaussian, is critical for improving the performances of both the sampling step (inner loop of the whole procedure) and the EBM training. We recall the vanilla discretized Langevin diffusion used in the related literature as follows:

$$z_t = z_{t-1} + \frac{\gamma}{2}\nabla \log \pi_\theta(z_t) + \sqrt{\gamma} B_t \ ,$$

where $\pi_\theta(\cdot)$ is the target distribution, z represents the states of the chains, *i.e.,*the generated samples in the context of EBM, t is the MCMC iteration index and B_t is the Brownian motion, usually set as a Gaussian noise and can be written as $B_t :== \Sigma\xi$ where ξ is a standard Gaussian random variable. This method directs the proposed moves towards areas of high probability for the stationary distribution π_θ, for any $\theta \in \Theta$, using the gradient of $\log \pi_{theta}$ and has

been the object of several studies [15, 5]. In high dimensional and highly nonlinear settings, the burden of computing this gradient for a certain number of MCMC transitions leads to a natural focus: the improvement on the behaviour of such sampling scheme by assimilating information about the landscape of the target, also called the stationary, distribution, while keeping its ease of implementation.

3.2. STANLEY an Anisotropic Energy Based Modeling Approach

Given the drawbacks of current MCMC methods used for training EBMs, we introduce a new sampler based on the Langevin updates presented above in Step 4 of Algorithm 1.

Heuristic behind the efficacy of STANLEY: Some past modifications have been proposed in particular to optimize the covariance matrix of the proposal of the general MCMC procedure in order to better stride the support of the target distribution. Langevin Dynamics is one example of those improvements where the proposal is a Gaussian distribution where the mean depends on the gradient of the log target distribution and the covariance depends on some Brownian motion. For instance, in [2, 31], the authors propose adaptive and geometrically ergodic Langevin chains. Yet, one important characteristic of our EBM problem, is that for each model parameter through the EBM training iterations, the target distribution moves and the proposal should take that adjustment into account. The technique in [2, 31] does not take the whole advantage of changing the proposal using the target distribution. 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 towards the stationary distribution is reached, in other words it would make sense towards the end of the EBM training, as the target distribution from a model parameter to the next one are similar. 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.

Moreover, in [16], the authors consider the approximation of a constant. Even though this simplification leads to ease of implementation, the curvature metric chosen by the authors need to be inverted, step that can be a computational burden if not intractable. Especially in the case we are considering in our paper, *i.e.*,ConvNet-based EBM, where the high nonlinearity would lead to intractable expectations.

Therefore, in (4) and (5) of Algorithm 1, we propose a variant of Langevin Dynamics, in order to sample from a target distribution, using a full anisotropic covariance matrix based on the anisotropy and correlations of the target distribution, see the $\sqrt{\gamma_t}\mathsf{B}_t$ term.

Algorithm 1 STANLEY for Energy-Based model

1: **Input**: Total number of iterations T, number of MCMC transitions K and of samples M, sequence of global learning rate $\{\eta_t\}_{t>0}$ for the EBM model update, constant threshold th, sequence of MCMC stepsize $\gamma_{k\,k>0}$ for the Langevin transitions, 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:

$$\gamma_t = \frac{\mathsf{th}}{\max(\mathsf{th}, |\nabla f_{\theta_t}(z_{t-1}^m)|)} \ . \tag{4}$$

4: Draw M samples $\{z_t^m\}_{m=1}^M$ from the objective potential (1) via Langevin diffusion:
5: **for** $k = 1$ to K **do**
6: Construct the Markov Chain as follows:

$$z_k^m = z_{k-1}^m + \gamma_k/2 \nabla f_{\theta_t}(z_{k-1}^m) + \sqrt{\gamma_k}\mathsf{B}_k \ , \tag{5}$$

where B_t is the Brownian motion, drawn from a Normal distribution.
7: **end for**
8: Assign $\{z_t^m\}_{m=1}^M \leftarrow \{z_K^m\}_{m=1}^M$.
9: Sample m positive observations $\{x_i\}_{i=1}^m$ from the empirical data distribution.
10: Compute the gradient of the empirical log-EBM (1) as follows:

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

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

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

12: **end for**
13: **Output:** Vector of fitted parameters θ_{T+1}.

4. Geometric ergodicity of STANLEY 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.

4.1. Notations and Assumptions

For any chain state $z \in \mathcal{Z}$ we denote by $\Pi_\theta(z, \cdot)$ the transition kernel as defined in the STANLEY update 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 has a geometric dependence on k.

We now state the assumptions required for our analysis. The first one is related to the continuity of the gradient of the log posterior distribution and the unit vectors pointing in the direction of the sample z and in the direction of the gradient of the log posterior distribution at z:

H1. *The stationary distribution is positive and has continuous derivative such that for all $\theta \in \mathbb{R}^d$:*

$$\lim_{z\to\infty} \frac{z}{|z|} \nabla f_\theta(z) = -\infty \ ,$$
$$\limsup_{z\to\infty} \frac{z}{|z|} \frac{\nabla f_\theta(z)}{|\nabla f_\theta(z)|} < 0 \ .$$

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

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

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:

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

The definition of this norm allow us to establish a convergence rate for our sampling method by deriving an upper bound of $\|\Pi_\theta^k - \pi_\theta\|_V$ where $k > 0$ 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 at a given EBM model θ. 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 as:

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

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

$$V_\theta(z) := c_\theta \pi_\theta(z)^{-\beta} \ , \tag{6}$$

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 (precisely its supremum). Define the auxiliary functions

$$V_1(z) := \inf_{\theta\in\Theta} V_\theta(z) \quad \text{and} \quad V_2(z) := \sup_{\theta\in\Theta} V_\theta(z) \ , \tag{7}$$

and assume that

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

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

4.2. Convergence Results

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 [32, 1] 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 $(\Pi_\theta)_{\theta\in\Theta}$.

Important Note: The stationary distributions are defined per $\theta \in \Theta$, *i.e.,* at each model update during the EBM optimization phase. Thus uniformity of convergence of the chain is important in order to characterize the sampling phase *throughout the entire training phase*. Particularly at the beginning, the shape of the distributions one needs to sample from vary a lot from a parameter to another.

Theorem 1. *Assume H1-H3. For any $\theta \in \Theta$, there exists a drift function V_θ, a set $\mathcal{O} \subset \mathcal{Z}$, a constant $0 < \epsilon \leq 1$ such that*

$$\Pi_\theta(z, \mathcal{B}) \geq \epsilon \int_{\mathcal{B}} 1_{\mathcal{X}}(z) dy \ . \tag{8}$$

Moreover there exists $0 < \mu < 1$, $\delta > 0$ and a drift function V, now independent of θ such that for all $z \in \mathcal{Z}$:

$$\Pi_\theta V(z) \leq \mu V(z) + \delta 1_{\mathcal{O}}(z) \ . \tag{9}$$

Theorem 1 shows two important convergence results for our sampling method. First, it established the existence of a small set $\mathcal{O}$ leading to the crucially needed aperiodicity of the chain and ensuring that each transition moves toward a better state. Then, it provide a uniform ergodicity result of our sampling method in STANLEY, via the so-called *drift condition* providing the guarantee that our user-designed transition kernels $(\Pi_\theta)_{\theta\in\Theta}$ attracts the states into the small set $\mathcal{O}$.

Moreover, the independence on the EBM model parameter θ of V in (9) leads to *uniform* ergodicity as shown in the following Corollary. While Theorem 1 is critical for proving the aperiodicity and irreducibility of the chain, we now establish the geometric speed of convergence of the chain. Not only we show the importance of the *uniform* ergodicity of the chain, which makes it appealing for the EBM training since the model parameter θ is often updated, but we also derive a geometrical rate in the following:

Corollary 1. *Assume H1-H3. A direct consequence of Theorem 1 is that the family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$ are uniformly ergodic,i.e., for any compact $\mathcal{C}\subset\mathcal{Z}$, there exist constants $\rho\in]0,1[$ and $e>0$ such for any iteration $t>0$,we have:*

$$\sup_{z\in\mathcal{C}} \|\Pi_\theta^t f(\cdot) - \pi_\theta f(\cdot)\|_V \leq e\rho^k \|f\|_{V_\theta} \ , \qquad (10)$$

where V is the drift function used in Theorem 1.

In the following, we develop a sketch of proof of the main Theorem of our paper. We give the important details leading to the desired ergodicity results. Those various techniques are common in the MCMC literature and we refer the readers to several MCMC handbooks such as [34, 32] for more understanding.

4.3. Sketch of the Proof of Theorem 1

Notations for the proof: We denote by $z \to \mathrm{T}_\theta(z',z)$, the pdf of the Gaussian proposal of Line 3 for any current state of the chain $z'\in\mathcal{Z}$ and dependent on the EBM model parameter θ. The transition kernel from z to z' is denoted by $\Pi_\theta(z,z')$. $\mathcal{Z}$ is a subset of $\mathbb{R}^\ell$ and $\mathcal{B}$ is a Borel set of $\mathbb{R}^\ell$.

The proof of our results are divided into two main parts. We first prove the existence of a small set for our transition kernel Π_θ, noted $\mathcal{O}$ showing that for any state, the Markov Chain moves away from it. It constitutes the first step toward proving its irreducibility and aperiodicity. Then, we will establish the so-called *drift condition*, also known as the Foster-Lyapunov condition, crucial to prove the convergence of the chain. The drift condition ensures the recurrence of the chain as the property that a chain returns to its initial state within finite time, see [43] for more details. Uniform ergodicity is then established as a consequence of those drift conditions and thus proving (9).

(i) Existence of a small set: Let $\mathcal{O}$ be a compact subset of the state space $\mathcal{Z}$.. We recall the definition of the transition kernel in the case of a Metropolis adjustment and for any model parameter $\theta\in\Theta$ and state $z\in\mathcal{Z}$:

$$\Pi_\theta(z,\mathcal{B}) = \int_{\mathcal{B}} \alpha_\theta(z,y)\mathrm{T}_\theta(z,y)\mathrm{d}y + 1_{\mathcal{B}(z)}\int_{\mathcal{Z}}(1-\alpha_\theta(z,y))\mathrm{T}_\theta(z,y)\mathrm{d}y \ ,$$

where we have defined the Metropolis ratio between two states $z\in\mathcal{Z}$ and $y\in\mathcal{B}$ as $\alpha_\theta(z,y)=\min(1,\frac{\pi_\theta(z)\mathrm{T}_\theta(z,y)}{\mathrm{T}_\theta(y,z)\pi_\theta(y)})$. Under H1 and due to the fact that the threshold th leads to a symmetric positive definite covariance matrix with bounded non zero eigenvalues, then the following holds:

$$an_{\sigma_1}(z-y) \leq \mathrm{T}_\theta(z,y) \leq bn_{\sigma_2}(z-y) \quad \text{for all} \quad \theta\in\Theta \ ,$$

where σ_1 and σ_2 are the corresponding standard deviation of the two Gaussian distributions n_{σ_1} and n_{σ_2}. We denote by ρ_θ the ratio $\frac{\pi_\theta(z)\mathrm{T}_\theta(z,y)}{\mathrm{T}_\theta(y,z)\pi_\theta(y)}$ and define the quantity

$$\delta = \inf(\rho_\theta(z,y), \theta\in\Theta, \quad z\in\mathcal{O}) > 0 \ , \qquad (11)$$

where we have used H1 and H2. Then,

$$\Pi_\theta(z,\mathcal{B}) \geq \int_{\mathcal{B}\cap\mathcal{X}} \alpha_\theta(z,y)\mathrm{T}_\theta(z,y)\mathrm{d}y \geq \min(1,\delta)m\int_{\mathcal{B}} 1_{\mathcal{X}}(z)\mathrm{d}y \ .$$

According to (11), we can find a compact set $\mathcal{O}$ such that $\Pi_\theta(z,\mathcal{B})\geq\epsilon$ where $\epsilon=\min(1,\delta)m\mathbf{Z}$ where $\mathbf{Z}$ is the normalizing constant of the pdf $\frac{1}{\mathbf{Z}}1_{\mathcal{X}}(z)\mathrm{d}y$ and the proposal distribution is bounded from below by some quantity noted m. Thus proving (8), *i.e.,*the existence of a small set for our family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$.

(ii) Drift condition and ergodicity: We begin by proving that $(\Pi_\theta)_{\theta\in\Theta}$ satisfies a drift property. For a given EBM parameter $\theta\in\Theta$, we can see in [26] that the drift condition boils down to proving that

$$\sup_{z\in\mathcal{Z}}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < \infty \quad \text{and} \quad \limsup_{|z|\to\infty}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < 1 \ ,$$

where V_θ is the *drift function* defined in (6) Let denote the acceptation set, *i.e.,* $\rho_\theta\geq 1$ by $\mathcal{A}_\theta(z) := \{y\in\mathcal{Z}, \rho_\theta(z,y)\geq 1\}$ for any state $y\in\mathcal{B}$ and its complementary set $\mathcal{A}_\theta^*(z)$. The remaining of the proof is composed of three main steps. STEP (1) shows that for any $\theta\in\Theta$:

$$\limsup_{|z|\to\infty}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq 1 - \liminf_{|z|\to\infty}\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \ .$$

Then, using an important intermediary result, stated in Lemma 1, that initiates a relation between the set of accepted state noted $\mathcal{A}_\theta(z)$ and the cone $\mathcal{P}(z)$ designed so that it does not depend on the model parameter θ.

Lemma 1. *Define* $\mathcal{P}(z) := \{z - \ell\frac{z}{|z|} - \kappa\nu$, *with* $\kappa < a - \ell, \nu \in \{\nu \in \mathbb{R}^d, \|\nu\| < 1\}, |\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|} \leq \frac{\epsilon}{2}\}$ *and* $\mathcal{A}_\theta(z) := \{y \in \mathcal{Z}, \rho_\theta(z, y) \geq 1\}$. *Then for* $z \in \mathcal{Z}$, $\mathcal{P}(z) \subset \mathcal{A}_\theta(z)$.

Noting the limit inferior as $\underline{\lim}$, STEP (2) establishes that $1 - \underline{\lim}_{|z|\to\infty} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z, y)\mathrm{d}y \leq 1 - c$ where c is a constant, *independent of all the other quantities* towards showing uniformity of the final result. Finally, STEP (3) uses the inequality $\Pi_\theta V_\theta(z) \leq \bar{\mu} V_\theta(z) + \bar{\delta} 1_{\mathcal{O}}(z)$ dependent of θ and defines the V function, independent of θ, as $V(z) = V_1(z)^\alpha V_2(z)^{2\alpha}$ in order to establish the main result of Theorem 1, *i.e.,*

$$\Pi_\theta V(z) \leq \left(\frac{\bar{\mu}}{2\epsilon^2} + \frac{\epsilon^2}{1 + \bar{\mu}}\right) V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) \; .$$

Setting $\epsilon := \sqrt{\frac{\bar{\mu}(1+\bar{\mu})}{2}}$, $\mu := \sqrt{\frac{2\bar{\mu}}{1+\bar{\mu}}}$ and $\delta := \frac{\bar{\delta}}{2\epsilon^2}$ proves the uniformity of the inequality (9).

5. Numerical Experiments

We present in this section a collection of numerical experiments to show the effectiveness of our method. Both on synthetic and real datasets, we put our EBM to the test of data generation. After verifying the advantage of STANLEY on a Gaussian Mixture Model (GMM) retrieving the synthetic data observations, we next investigate its performance when learning a distribution over high-dimensional natural images such as pictures of flowers, see the Flowers dataset in [38], or general concepts displayed in CIFAR-10 [29]. For both methods, we use the Frechet Inception Distance (FID), as a reliable performance metrics as detailed in [23]. In the sequel, we tune the learning rates over a fine grid and report the best result for all methods. For our method STANLEY, the threshold parameter th, crucial for the implementation of the stepsize (4) is tuned over a grid search as well.

5.1. Toy Example: Gaussian Mixture Model

Datasets. We first demonstrate the outcomes of both methods including our newly proposed STANLEY for low-dimensional toy distributions. We generate synthetic 2D rings data points and use an EBM to learn the true data distribution and put it to the test of generating new faithful synthetic samples.

Methods and Settings. We consider two methods. Methods are ran with *nonconvergent* MCMC, *i.e.,*, we do not necessitate the convergence to the stationary distribution of the Markov chains. The number of transitions of the MCMC is set to $K = 100$ per EBM iteration. We use a standard deviation of 0.15 as in [36]. Both methods have a constant learning rate of 0.14. The value of the threshold th for our STANLEY method is set to $\mathsf{th} = 0.01$. The total number of EBM iterations is set to $T = 10\,000$. The global learning rate η is set to a constant equal to 0.0001.

Network architectures. For the backbone of the EBM model, noted $f_\theta(\cdot)$ in (1), we chose a CNN of 5 2D convolutional layers and Leaky ReLU activation functions, with the leakage parameter set to 0.05. The number of hidden neurons varies between 32 and 64.

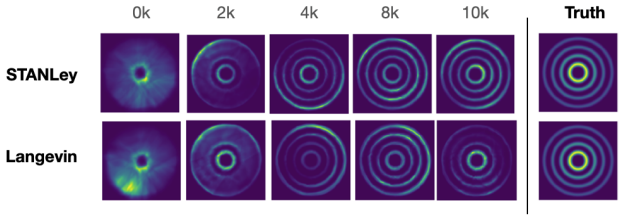

Figure 1. (Rings Toy Dataset) Top: our method, namely STANLEY Bottom: vanilla Langevin Dynamics. Both methods are used with the same backbone architecture. Generated samples are plotted through the iterations ever 2 000 steps.

Results. We observe Figure 1 the outputs of both methods on the toy dataset. While both methods achieve a great representation of the truth after a large number of iterations, we notice that STANLEY learns an energy that closely approximates the true density during the first thousands of iterations if the training process. The sharpness of the data generated by STANLEY in the first iterations shows an empirically better ability to sample from the 2D toy dataset.

5.2. Flowers Dataset

Datasets. We now compare the algorithms on *the Oxford Flowers 102* dataset [38]. The dataset is composed of 102 flower categories. Per request of the authors, the images have large scale, pose and light variations making the task of generating new samples particularly challenging.

Methods and Settings. Nonconvergent MCMC are also used in this experiments and the number of MCMC transitions is set to $K = 50$. Global learning parameters of the gradient update is set to 0.001 for both methods. We run each method during $T = 100\,000$ iterations and plot the results using the final vector of fitted parameters.

Network architectures. The backbone of the energy function for this experiment is a vanilla ConvNet composed

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

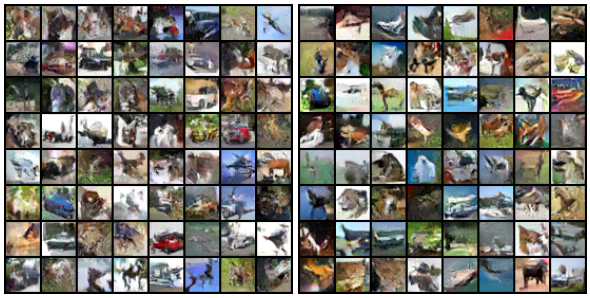

Figure 3. (CIFAR Dataset). Left: Langevin Method. Right: STANLEY method. After 100k iterations.

of 3×3 convolution layers with stride 1. 5 Convolutional Layers using ReLU activation functions are stacked.

Results. Visual results are provided in Figure 2 where we have used both methods to generate synthetic images of flowers. For each threshold iterations number (5 000 iterations) we sample 10 000 synthetic images from the EBM model under the current vector of parameters and use the same number of data observations to compute the FID similarity score as advocated in [23]. The evolution of the FID values are reported in Figure 4 (Left) through the iterations.

5.3. CIFAR Dataset

Datasets. For this third experiment we use the *CIFAR-10* dataset [29]. *CIFAR-10* is a popular computer-vision dataset of 50 000 training images and 10 000 test images, of size 32×32. It is composed of tiny natural images representing a wide variety of objects and scenes, making the task of self supervision supposedly harder.

Methods and Settings. We employ the same nonconvergent MCMC strategies for this experiment. The value of the threshold th for our STANLEY method is set to $\mathsf{th} = 0.0002$. The total number of EBM iterations is set to $T = 100\,000$. The global learning rate η is set to a constant equal to 0.0001. In this experiment, we slightly change the last step of our method described in Algorithm 1. Indeed, Step 11 is not a plain Stochastic Gradient Descent here but we rather use the ADAM optimizer [27]. The scaling factor of the Brownian motion is equal to 0.01.

Network architectures. We adopt a similar ConvNet as the one used in the Flowers example, composed of 3×3 convolution layers with stride 1. 5 Convolutional Layers using ReLU activation functions are stacked in our model.

Results. Visual results are provided in Figure 3 where we have used both methods to generate synthetic images of flowers. The FID values are reported in Figure 4 (Right) and have been computed using 10 000 synthetic images from each model. The similarity score is then evaluated every 5 000 iterations. While the FID curves for the Flowers dataset exhibits a superior performance of our method throughout the training procedure, we notice that in the case of CIFAR-10, vanilla method seems to be slightly better than STANLEY during the first iterations, *i.e.,* when the model is still learning the representation of the images. Yet, after a certain number of iterations, we observe that STANLEY leads to more accurate synthetic images. This behavior can be explained by the importance of incorporating curvature informed metrics into the training process when the parameter reaches a neighborhood of the optimal solution.

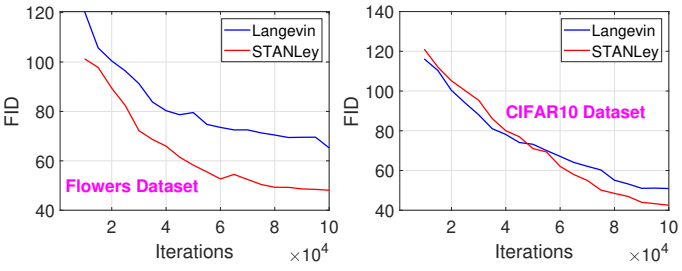

Figure 4. (FID values per method against 100k iterations elapsed). Left: Oxford Flowers dataset. Right: CIFAR-10 dataset.

6. Conclusion

Given the growing interest in self-supervised learning, we propose in this paper, an improvement of the so-called MCMC based Energy-Based models. In the particular case of a highly nonlinear structural model of the EBM, more precisely a Convolutional Neural Network in our paper, we tackle the complex task of sampling negative samples from the energy function. The multi-modal and highly curved landscape one must sample from inspire our technique called STANLEY, and based on a Stochastic Gradient Anisotropic Langevin Dynamics, that updates the Markov Chain using an anisotropic stepsize in the vanilla Langevin update. We provide strong theoretical guarantees for our novel method, including uniform ergodicity of the resulting chain. Our method is tested on several bench-

marks data and image generation tasks including toy and real datasets such as CIFAR-10.

References

[1] Stéphanie Allassonniere and Estelle Kuhn. Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Computational Statistics & Data Analysis*, 91:4–19, 2015. 5

[2] 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. 4

[3] 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. 2, 3

[4] Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted langevin algorithm. *arXiv preprint arXiv:1710.05559*, 2017. 3

[5] Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. Mcmc methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446, 2013. 4

[6] Nando de Freitas, Pedro Højen-Sørensen, Michael I Jordan, and Stuart Russell. Variational mcmc. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 120–127, 2001. 3

[7] 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. 1

[8] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000. 3

[9] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020. 2

[10] Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. *arXiv preprint arXiv:2004.13167*, 2020. 3

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 2, 3

[12] Alain Durmus, Gareth O. Roberts, Gilles Vilmart, and Konstantinos C. Zygalakis. Fast langevin based algorithm for mcmc in high dimensions. *Ann. Appl. Probab.*, 27(4):2195–2237, 08 2017. 3

[13] 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. 2

[14] Ruiqi Gao, Erik Nijkamp, Diederik P Kingma, Zhen Xu, Andrew M Dai, and Ying Nian Wu. Flow contrastive estimation of energy-based models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7518–7528, 2020. 3

[15] 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. 3, 4

[16] 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. 4

[17] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014. 3

[18] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020. 3

[19] 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. 3

[20] Fredrik K Gustafsson, Martin Danelljan, Goutam Bhat, and Thomas B Schön. Energy-based models for deep probabilistic regression. In *European Conference on Computer Vision*, pages 325–343. Springer, 2020. 3

[21] 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. 1

[22] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018. 3

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017. 7, 8

[24] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. 2

[25] 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. 2

[26] Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of metropolis algorithms. *Stochastic processes and their applications*, 85(2):341–361, 2000. 6, 13

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 8

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[29] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 7, 8

[30] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 1, 3

[31] Tristan Marshall and Gareth Roberts. An adaptive approach to langevin mcmc. *Statistics and Computing*, 22(5):1041–1057, 2012. 4

[32] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012. 1, 2, 5, 6

[33] 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. 1

[34] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011. 3, 6

[35] 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. 1, 3

[36] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 5272–5280. AAAI Press, 2020. 7

[37] 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. 1

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 7

[39] 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. 2

[40] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951. 2

[41] Christian P. Robert and George Casella. *Metropolis–Hastings Algorithms*, pages 167–197. Springer New York, New York, NY, 2010. 3

[42] G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *J. R. Statist. Soc. B*, 60:255–268, 1997. 3

[43] Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004. 6

[44] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996. 3

[45] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. 3

[46] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009. 3

[47] 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. 1

[48] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021. 1

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[50] 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. 2

[51] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, Jan. 2008. 3

[52] 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. 2

[53] 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. 3

[54] 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. 1

[55] Russ Wolfinger. Laplace's approximation for nonlinear mixed models. *Biometrika*, 80(4):791–795, 1993. 3

[56] 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. 1

[57] 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. 1

[58] 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. 1

A. Proofs of the Theoretical Results

A.1. Proof of Theorem 1

Theorem. *Assume H1-H3. For any $\theta \in \Theta$, there exists a drift function V_θ, a set $\mathcal{O} \subset \mathcal{Z}$, a constant $0 < \epsilon \leq 1$ such that*

$$\Pi_\theta(z, \mathcal{B}) \geq \epsilon \int_{\mathcal{B}} 1_{\mathcal{X}}(z) dy \; . \tag{12}$$

Moreover there exists $0 < \mu < 1$, $\delta > 0$ and a drift function V, now independent of θ such that for all $z\mathcal{Z}$:

$$\Pi_\theta V(z) \leq \mu V(z) + \delta 1_{\mathcal{O}}(z) \; . \tag{13}$$

Proof. We list the notations used throughout this proof in the following table:

Π_θ	$\triangleq$	Transition kernel of the MCMC defined by (5)
$\mathcal{O}$	$\triangleq$	Subset of $\mathbb{R}^p$ and small set for kernel Π_θ
$\mathcal{A}_\theta(z)$	$\triangleq$	Acceptance set at state $\in \mathcal{Z}$ such that $\rho_\theta \geq 1$
$\mathcal{A}^*_\theta(z)$	$\triangleq$	Complementary set of $\mathcal{A}_\theta(z)$
$\mathrm{T}_\theta(z', z)$	$\triangleq$	Probability density function of the Gaussian proposal
$\Pi_\theta(z, z')$	$\triangleq$	Transition kernel from state z to state z'

The proof of our results are divided into two parts. We first prove the existence of a set noted $\mathcal{O}$ as a small set for our transition kernel Π_θ. Proving a small set is important to show that for any state, the Markov Chain does not stay in the same state, and thus help in proving its irreducibility and aperiodicity.

Then, we will prove the drift condition towards a small set. This condition is crucial to prove the convergence of the chain since it states that the kernels tend to attract elements into that set. finally, uniform ergodicity is established as a consequence of those drift conditions.

(i) Existence of small set: Let $\mathcal{O}$ be a compact subset of the state space $\mathcal{Z}$. We also denote the probability density function (pdf) of the Gaussian proposal of Line 3 as $z \to \mathrm{T}_\theta(z', z)$ for any current state of the chain $z' \in \mathcal{Z}$ and dependent on the EBM model parameter θ. Given STANLEY's MCMC update, at iteration t, the proposal is a Gaussian distribution of mean $z^m_{t-1} + \gamma_t / 2 \nabla f_{\theta_t}(z^m_{t-1})$ and covariance $\sqrt{\gamma_t} \mathrm{B}_t$.

We recall the definition of the transition kernel in the case of a Metropolis adjustment and for any model parameter $\theta \in \Theta$ and state $z \in \mathcal{Z}$:

$$\Pi_\theta(z, \mathcal{B}) = \int_{\mathcal{B}} \alpha_\theta(z, y) \mathrm{T}_\theta(z, y) \mathrm{d}y + 1_{\mathcal{B}(z)} \int_{\mathcal{Z}} (1 - \alpha_\theta(z, y)) \mathrm{T}_\theta(z, y) \mathrm{d}y \tag{14}$$

where we have defined the Metropolis ratio between two states $z \in \mathcal{Z}$ and $y \in \mathcal{B}$ as $\alpha_\theta(z, y) = \min(1, \frac{\pi_\theta(z) \mathrm{T}_\theta(z, y)}{\mathrm{T}_\theta(y, z) \pi_\theta(y)})$. Thanks to Assumption H1 and due to the fact that the threshold th leads to a symmetric positive definite covariance matrix with bounded non zero eigenvalues implies that the proposal distribution can be bounded by two zero-mean Gaussian distributions as follows:

$$a n_{\sigma_1}(z - y) \leq \mathrm{T}_\theta(z, y) \leq b n_{\sigma_2}(z - y) \quad \text{for all} \quad \theta \in \Theta \tag{15}$$

where σ_1 and σ_2 are the corresponding standard deviation of the distributions and a and b are some scaling factors.

We denote by ρ_θ the ratio $\frac{\pi_\theta(z) \mathrm{T}_\theta(z, y)}{\mathrm{T}_\theta(y, z) \pi_\theta(y)}$ and define the quantity

$$\delta = \inf(\rho_\theta(z, y), \theta \in \Theta, \quad z \in \mathcal{O}) > 0 \tag{16}$$

given the assumptions H1 and H2. Likewise, the proposal distribution is bounded from below by some quantity noted m. Then,

$$\Pi_\theta(z, \mathcal{B}) \geq \int_{\mathcal{B} \cap \mathcal{X}} \alpha_\theta(z, y) \mathrm{T}_\theta(z, y) \mathrm{d}y \geq \min(1, \delta) m \int_{\mathcal{B}} 1_{\mathcal{X}}(z) \mathrm{d}y \tag{17}$$

Then, given the definition of (16), we can find a compact set $\mathcal{O}$ such that $\Pi_\theta(z,\mathcal{B}) \geq \epsilon$ where $\epsilon = \min(1,\delta)m\mathbf{Z}$ where $\mathbf{Z}$ is the normalizing constant of the pdf $\frac{1}{\mathbf{Z}}\mathbf{1}_{\mathcal{X}}(z)\mathrm{d}y$. Thus proving (8), *i.e.,*the existence of a small set for our family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$.

(ii) Drift condition and ergodicity: We first need to prove the fact that our family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$ satisfies a drift property.

For a given EBM model parameter $\theta \in \Theta$, we can see in [26] that the drift condition boils down to proving that for the drift function noted V_θ and defined in (6), we have

$$\sup_{z\in\mathcal{Z}} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < \infty \quad \text{and} \quad \lim \sup_{|z|\to\infty} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < 1 \tag{18}$$

Throughout the proof, the model parameter is set to an arbitrary $\theta \in \Theta$. Let denote the acceptation set, *i.e.,* $\rho_\theta \geq 1$ by $\mathcal{A}_\theta(z) := \{y \in \mathcal{Z}, \rho_\theta(z,y) \geq 1\}$ for any state $y \in \mathcal{B}$ and its complementary set $\mathcal{A}^*_\theta(z)$.

STEP (1): Following our definition of the drift function in (6) we obtain:

$$\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} = \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{V_\theta(y)}{V_\theta(z)}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \frac{\pi_\theta(y)\mathrm{T}_\theta(y,z)}{\pi_\theta(z)\mathrm{T}_\theta(z,y)}\mathrm{T}_\theta(z,y)\frac{V_\theta(y)}{V_\theta(z)}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} (1 - \frac{\pi_\theta(y)\mathrm{T}_\theta(y,z)}{\pi_\theta(z)\mathrm{T}_\theta(z,y)})\mathrm{T}_\theta(z,y)\mathrm{d}y \tag{19}$$

$$\overset{(a)}{\leq} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \tag{20}$$

where (a) is due to (6).

According to (15), we thus have that, for any state z in the acceptance set $\mathcal{A}_\theta(z)$:

$$\int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y \leq b\int_{\mathcal{A}_\theta(z)} n_{\sigma_2}(y-z)\mathrm{d}y\ . \tag{21}$$

For any state z in the complementary set of the acceptance set, noted $\mathcal{A}^*_\theta(z)$, we also have the following:

$$\int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}\mathrm{d}y \leq \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)^{1-\beta}\mathrm{T}_\theta(y,z)^{\beta}\mathrm{d}y \leq b\int_{\mathcal{A}^*_\theta(z)} n_{\sigma_2}(z-y)\mathrm{d}y\ . \tag{22}$$

While we can define the level set of the stationary distribution π_θ as $\mathcal{L}_{\pi_\theta(y)} = \{z \in \mathcal{Z}, \pi_\theta(z) = \pi_\theta(y)\}$ for some state $y \in \mathcal{B}$, a neighborhood of that level set is defined as $\mathcal{L}_{\pi_\theta(y)}(p) = \{z \in \mathcal{L}_{\pi_\theta(y)}, z + t\frac{z}{|z|}, |t| \leq p\}$. H1 ensures the existence of a radial r such that for all $z \in \mathcal{Z}, |z| \geq r$, then $0 \in \mathcal{L}_{\pi_\theta(y)}$ with $\pi_\theta(z) > \pi_\theta(y)$. Since the function $y \to n_{\sigma_2}(y-z)$ is smooth, it is known that there exists a constant $a > 0$ such that for $\epsilon > 0$, we have that

$$\int_{B(z,a)} n_{\sigma_2}(y-z)\mathrm{d}y \geq 1-\epsilon \quad \text{and} \quad \int_{B(z,a)\cap\mathcal{L}_{\pi_\theta(y)}(p)} n_{\sigma_2}(y-z)\mathrm{d}y \leq \epsilon \tag{23}$$

for some p small enough and where $B(z,a)$ denotes the ball around $z \in \mathcal{Z}$ of radius a. Then combining (21) and (23) we have that:

$$\int_{\mathcal{A}_\theta(z)\cap B(z,a)\cap\mathcal{L}_{\pi_\theta(y)}(p)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y \leq b\epsilon \tag{24}$$

Conversely, we can define the set $\mathcal{A} = \mathcal{A}_\theta(z) \cap B(z,a) \cap \mathcal{L}^+$ where $u \in \mathcal{L}^+$ if $u \in \mathcal{L}_{\pi_\theta(y)}(p)$ and $\phi_\theta(u) > \pi_\theta(p)$. Then using the second part of H1, there exists a radius $r' > r + a$, such that for $z \in \mathcal{Z}$ with $|z| \geq r'$ we have

$$\int_{\mathcal{A}} (\frac{\pi_\theta(y)}{\pi_\theta(z)})^{1-\beta}\mathrm{T}_\theta(y,z)\mathrm{d}y \leq \mathsf{d}(p,r')^{1-\beta}b\int_{\mathcal{A}_\theta(z)} n_{\sigma_2}(y-z)\mathrm{d}y \leq b\mathsf{d}(p,r')^{1-\beta} \tag{25}$$

where $\mathsf{d}(p,r') = \sup_{|z|>r'} \frac{\pi_\theta(z+p\frac{z}{|z|})}{\pi_\theta(z)}$. Note that H1 implies that $\mathsf{d}(p,r') \to 0$ when $r' \to \infty$. Likewise with $\mathcal{A} = \mathcal{A}_\theta(z) \cap B(z,a) \cap \mathcal{L}^-$ we have

$$\int_{\mathcal{A}} (\frac{\pi_\theta(y)}{\pi_\theta(z)})^{-\beta}\mathrm{T}_\theta(z,y)\mathrm{d}y \leq b\mathsf{d}(p,r')^{\beta} \tag{26}$$

Same arguments can be obtained for the second term of (19), *i.e.*, $\mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}$ and we obtain, plugging the above in (19) that:

$$\lim\sup_{|z|\to\infty}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq \lim\sup_{|z|\to\infty}\int_{\mathcal{A}_\theta^*(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \tag{27}$$

Since $\mathcal{A}_\theta^*(z)$ is the complementary set of $\mathcal{A}_\theta(z)$, the above inequality yields

$$\lim\sup_{|z|\to\infty}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq 1-\lim\inf_{|z|\to\infty}\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \tag{28}$$

STEP (2): The final step of our proof consists in proving that $1-\lim\inf_{|z|\to\infty}\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \leq 1-c$ where c is a constant, independent of all the other quantities.

Given that the proposal distribution is a Gaussian and using assumption H1 we have the existence of a constant c_a depending on a as defined above (the radius of the ball $B(z,a)$) such that

$$\frac{\pi_\theta(z)}{\pi_\theta(z-\ell\frac{z}{|z|})} \leq c_a \leq \inf_{y\in B(z,a)}\frac{\mathrm{T}_\theta(y,z)}{\mathrm{T}_\theta(z,y)} \quad \text{for any } z\in\mathcal{Z}, |z|\geq r^* \tag{29}$$

Then for any $|z|\geq r^*$, we obtain that $z-\ell\frac{z}{|z|}\in\mathcal{A}_\theta(z)$. A particular subset of $\mathcal{A}_\theta(z)$ used throughout the rest of the proof is the cone defined as

$$\mathcal{P}(z) := \{z-\ell\frac{z}{|z|}-\kappa\nu, \text{ with } \quad i<a-\ell, \nu\in\{\nu\in\mathbb{R}^d, \|\nu\|<1\}, |\nu-\frac{z-\ell\frac{z}{|z|}}{|z-\ell\frac{z}{|z|}|}\leq\frac{\epsilon}{2}\} \tag{30}$$

Using Lemma 1, we have that $\mathcal{P}(z)\subset\mathcal{A}_\theta(z)$

Then,

$$\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \overset{(a)}{\geq} \int_{\mathcal{A}_\theta(z)} a n_{\sigma_1}(y-z)\mathrm{d}y \overset{(b)}{\geq} a\int_{\mathcal{P}(z)} n_{\sigma_1}(y-z)\mathrm{d}y \tag{31}$$

where we have used (15) in (a) and applied Lemma 1 in (b).

If we define the translation of vector $z\in\mathcal{Z}$ by the operator $\mathcal{I}\subset\mathbb{R}^d\to T_z(\mathcal{I})$, then

$$\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y \geq a\int_{\mathcal{P}(z)} n_{\sigma_1}(y-z)\mathrm{d}y = \int_{T_z(\mathcal{P}(z))} n_{\sigma_1}(y-z)\mathrm{d}y \tag{32}$$

Recalling the objective of STEP (2) that is to find a constant c such that $1-\lim\inf_{|z|\to\infty}\int_{\mathcal{A}_\theta(z)}\mathrm{T}_\theta(z,y)\mathrm{d}y\leq 1-c$, we see from (32) that since the set $\mathcal{P}(z)$ does not depend on the EBM model parameter θ and that once translated by z the resulting set $T_z(\mathcal{P}(z))$ is independent of z (but depends on ℓ, see definition (30), then the integral $\int_{T_z(\mathcal{P}(z))} n_{\sigma_1}(y-z)\mathrm{d}y$ in (32) is independent of z thus concluding on the existence of the constant c such that $\lim\sup_{|z|\to\infty}\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)}\leq 1-c$. Thus proving the second part of (18) which is the main drift condition we ought to demonstrate. The first part of (18) can be proved by observing that $\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)}$ is smooth on $\mathcal{Z}$ according to H2 and by construction of the transition kernel. Smoothness implies boundedness on the compact $\mathcal{Z}$.

STEP (3): We now use the main proven equations in (18) to derive the second result (9) of Theorem 1.

We will begin by showing a similar inequality for the drift function V_θ, thus not having uniformity, as an intermediary step. The Drift property is a consequence of STEP (2) and (32) shown above. Thus, there exists $0<\bar{\mu}<1$, $\bar{\delta}>0$ such that for all $z\mathcal{Z}$:

$$\Pi_\theta V_\theta(z) \leq \bar{\mu}V_\theta(z) + \bar{\delta}1_{\mathcal{O}}(z)\ , \tag{33}$$

where V_θ is defined by (6).

Using the two functions defined in (7), we define for $z\in\mathcal{Z}$, the V function independent of θ as follows:

$$V(z) = V_1(z)^\alpha V_2(z)^{2\alpha}\ , \tag{34}$$

where $0 < \alpha < \min(\frac{1}{2\beta}, \frac{a_0}{3})$, a_0 is defined in H3 and β is defined in (6). Thus for $\theta \in \Theta$, $z \in \mathcal{Z}$ and $\epsilon > 0$:

$$\begin{aligned}\Pi_\theta V(z) &= \int_{\mathcal{Z}} \Pi_\theta(z,y) V_1(y)^\alpha V_2(y)^{2\alpha} \mathrm{d}y \\ &\overset{(a)}{\leq} \frac{1}{2}\int_{\mathcal{Z}} \Pi_\theta(z,y)\left(\frac{1}{\epsilon^2} V_1(y)^{2\alpha} + \epsilon^2 V_2(y)^{4\alpha}\right)\mathrm{d}y \\ &\overset{(b)}{\leq} \frac{1}{2\epsilon^2}\int_{\mathcal{Z}} \Pi_\theta(z,y) V_\theta(y)^{2\alpha} + \frac{\epsilon^2}{2}\int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha}\mathrm{d}y \end{aligned} \tag{35}$$

where we have used the Young's inequality in (a) and the definition of V_1, see (7), in (b). Then plugging (33) in (35), we have

$$\Pi_\theta V(z) \leq \frac{1}{2\epsilon^2}(\bar{\mu} V_\theta(z)^{2\alpha} + \bar{\delta} 1_{\mathcal{O}}(z)) + \frac{\epsilon^2}{2}\int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha}\mathrm{d}y \tag{36}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{2}\int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha}\mathrm{d}y \tag{37}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{2}\sup_{\theta\in\Theta, z\in\mathcal{Z}}\int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha}\mathrm{d}y \tag{38}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{1+\bar{\mu}} V(z) \tag{39}$$

$$\leq \left(\frac{\bar{\mu}}{2\epsilon^2} + \frac{\epsilon^2}{1+\bar{\mu}}\right) V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) \tag{40}$$

where we have used (34) and the assumption H3 in the last inequality, ensuring the existence of such exponent α.

Setting $\epsilon := \sqrt{\frac{\bar{\mu}(1+\bar{\mu})}{2}}$, $\mu := \sqrt{\frac{2\bar{\mu}}{1+\bar{\mu}}}$ and $\delta := \frac{\bar{\delta}}{2\epsilon^2}$ proves the uniform ergodicity in (9) and concludes the proof of Theorem 1. □

A.2. Proof of Lemma 1

Lemma. *Define* $\mathcal{P}(z) := \{z - \ell\frac{z}{|z|} - \kappa\nu$, *with* $\kappa < a - \ell, \nu \in \{\nu \in \mathbb{R}^d, \|\nu\| < 1\}, |\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|} \leq \frac{\epsilon}{2}\}$ *and* $\mathcal{A}_\theta(z) := \{y \in \mathcal{Z}, \rho_\theta(z,y) \geq 1\}$. *Then for* $z \in \mathcal{Z}$, $\mathcal{P}(z) \subset \mathcal{A}_\theta(z)$.

Proof. In order to show the inclusion of the set $\mathcal{P}(z)$ in $\mathcal{A}_\theta(z)$ we start by selecting the quantity $y = z - \ell\frac{z}{|z|} - \kappa\nu$ for $z \in \mathcal{Z}$ and $\kappa < a - \ell$ where a is the radius of the ball used in (23) such that $y \in \mathcal{P}(z)$. We will now show that $y \in \mathcal{A}_\theta(z)$.

By the generalization of Rolle's theorem applied on the stationary distribution π_θ, we guarantee the existence of some κ^* such that:

$$\nabla\pi_\theta(z - \ell\frac{z}{|z|} - \kappa^*\nu) = \frac{\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|})}{y - (z - \ell\frac{z}{|z|})} = -\frac{\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|})}{\kappa\nu}$$

Expanding $\nabla\pi_\theta(z - \ell\frac{z}{|z|} - \kappa^*\nu)$ yields:

$$\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|}) = -\kappa\nu\frac{z - \ell\frac{z}{|z|} - \kappa^*\nu}{|z - \ell\frac{z}{|z|} - \kappa^*\nu|}|\nabla\pi_\theta(z - \ell\frac{z}{|z|} - \kappa^*\nu)| \tag{41}$$

Yet, under H1, there exists ϵ such that

$$\frac{\nabla f_\theta(z)}{|\nabla f_\theta(z)|}\frac{z}{|z|} \leq -\epsilon$$

and for any $y \in \mathcal{P}(z)$ we note that $\frac{y}{|y|} - \frac{z}{|z|}| \leq \frac{\epsilon}{2}$, by construction of the set. Thus,

$$\frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\nu = \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}(\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|}) + \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}(\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|} - \frac{y}{|y|}) + \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\frac{y}{|y|} \leq 0 \tag{42}$$

where ν is used in the definition of $\mathcal{P}(z)$. Also note that $\frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\nu$ denotes the vector multiplication between the normalized gradient and ν.

Then plugging (42) into (41) leads to $\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|}) \geq 0$. Then $y \in \mathcal{P}(z)$ implies, using (29), that $\pi_\theta(y) \geq \pi_\theta(z - \ell\frac{z}{|z|}) \geq \frac{1}{c_a}\pi_\theta(z)$. Finally $y \in \mathcal{P}(z)$ implies that $y \in \mathcal{A}_\theta(z)$, concluding the proof of Lemma 1.

□