Adaptation of the Independent Metropolis-Hastings Sampler with Normalizing Flow Proposals

James A. Brofos
Yale University

Marylou Gabrié
CDS, New York University
CCM, Flatiron Institute

Marcus A. Brubaker
York University
Vector Institute

Roy R. Lederman
Yale University

Abstract

Markov Chain Monte Carlo (MCMC) methods are a powerful tool for computation with complex probability distributions. However the performance of such methods is critically dependent on properly tuned parameters, most of which are difficult if not impossible to know a priori for a given target distribution. Adaptive MCMC methods aim to address this by allowing the parameters to be updated during sampling based on previous samples from the chain at the expense of requiring a new theoretical analysis to ensure convergence. In this work we extend the convergence theory of adaptive MCMC methods to a new class of methods built on a powerful class of parametric density estimators known as normalizing flows. In particular, we consider an independent Metropolis-Hastings sampler where the proposal distribution is represented by a normalizing flow whose parameters are updated using stochastic gradient descent. We explore the practical performance of this procedure on both synthetic settings and in the analysis of a physical field system, and compare it against both adaptive and non-adaptive MCMC methods.

1 INTRODUCTION

Markov Chain Monte Carlo (MCMC) methods are procedures for generating samples from probability distributions, typically given knowledge of the density of the distribution up to proportionality. These MCMC samplers often depend on parameters; for instance, in the random walk Metropolis procedure on $\mathbb{R}^n$, one may treat the covariance matrix of a normal proposal distribution as a parameter of the method; see, for instance, Haario et al. (2001). The performance of an MCMC procedure will depend on these parameters. It would be preferable if these parameters could be adapted during sampling *at every step of the chain*, however such adaptions can violate the Markov property of the chain and undermine its convergence to the desired target distribution.

An important variation of MCMC is the independent Metropolis-Hastings sampler. This method samples from a target distribution by first sampling from an auxiliary proposal distribution (independently from the current state of the chain) and accepts or rejects those proposals according to the Metropolis-Hastings criterion. The effectiveness of this algorithm depends on the ratio of the target density to the ratio of the proposal density (Robert and Casella, 2005): if the ratio is bounded over the support of the target distribution, the algorithm enjoys a powerful theory of geometric ergodicity. The independent Metropolis-Hastings algorithm is the focus of the present work.

Recently in the machine learning community, normalizing flows have emerged as a powerful mechanism for expressing complex densities, see (Kobyzev et al., 2020; Papamakarios et al., 2021) for recent reviews. Normalizing flows are defined by a parametric, smooth and invertible function which transforms a simple distribution (e.g., a Gaussian) into a more complex one (e.g., natural images) and uses the change-of-variables formula to exactly determine the resulting probability density function in the complex space. Provided that the family of normalizing flows under consideration is sufficiently expressive, any distribution can be constructed in theory this way. In practice, many normalizing flows exhibit a universal approximation property whereby, given suitable model capacity, they can approximate any distribution arbitrarily well, e.g., (Huang et al., 2018; Jaini et al., 2019). Indeed, normalizing flows are distinguished among parameteric

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

families of distributions by their expressiveness and tractability of sampling and log-density evaluation; the precise attributes that one requires for a proposal distribution in the independent Metropolis-Hastings sampler. By incorporating normalizing flows into the MCMC framework we seek to leverage their expressivity along with the ergodicity of the MCMC procedure in order to produce samples from a target distribution (see fig. 1). The principle computational challenge associated to normalizing flows is the identification of parameters that produce the best approximation of a target density. Therefore, a question of principle theoretical interest and practical importance is, "During the course of sampling, under what conditions can the parameters of the normalizing flow be adapted at every step of the chain?"

The outline of this paper is as follows. In section 2 we review important concepts from the analysis of Markov chains; we provide the independent Metropolis-Hastings algorithm and state the conditions under which it enjoys geometric ergodicity; we devise a metric space over transition kernels, which will be important for analyzing notions of continuity. We review recent experimental works that demonstrated the benefit of normalizing flow proposals in MCMCs and related theoretical literature in section 3. In section 4 we state our theories for the continual adaptation of Markov chains. We begin by considering *deterministic adaptations* wherein parameter updates are determined sequentially and deterministically without regard to the state of the chain; this case can be used to motivate the adaptation of normalizing flows as a gradient flow. We then proceed to consider *stochastic adaptations* wherein the state of the chain and the adaptation of the parameters of the normalizing flow at the n^{th} step are *not necessarily independent* given the history of the chain up to the $(n-1)^{\text{th}}$ step. This circumstance includes the case wherein the accepted proposal sampled from the normalizing flow is also used in the computation of the adaptation, as necessary for the "pseudo-likelihood" algorithm we examine numerically in section 5.

2 PRELIMINARIES

In giving an overview of Markov chains and their associated theory, we emulate the notation and presentation of Meyn and Tweedie (1993). Refer to appendix A for a review of total variation distances. Throughout, we let $\mathcal{X}$ denote a set which we equip with its Borel σ-algebra, denoted $\mathfrak{B}(\mathcal{X})$. We associate to $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ a measure $\mu : \mathfrak{B}(\mathcal{X}) \to [0, \infty)$ – satisfying $\mu(A) \geq 0$ for all $A \in \mathfrak{B}(\mathcal{X})$, $\mu(\emptyset) = 0$, and the condition of countable additivity – to create the measure space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu)$. A probability measure is a measure which satisfies Kolmogorov's axioms (Kolmogorov, 1960). A signed measure relaxes the condition of non-negativity. If X is an $\mathcal{X}$-valued random variable and Π is a probability measure on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ we write $X \sim \Pi(\cdot)$ to mean that for any $A \in \mathfrak{B}(\mathcal{X})$ we have $\Pr[X \in A] = \Pi(A)$. If a probability measure Π has a density with respect to a dominating measure μ, this means that for all $A \in \mathfrak{B}(\mathcal{X})$, $\Pi(A) = \int_A \pi(x)\ \mu(\mathrm{d}x)$. The support of a density π is $\mathrm{Supp}(\pi) = \{x \in \mathcal{X} : \pi(x) > 0\}$. When we turn our attention to the discussion of parameterizations of transition kernels, we will write $\mathcal{Y}$ as a generic parameter space and use the symbol $\theta \in \mathcal{Y}$ to refer to a particular parameterization. We denote the Dirac measure concentrated at $x \in \mathcal{X}$ by $\delta_x(\cdot)$.

2.1 Transition Kernels

In MCMC, we generate a sequence of $\mathcal{X}$-valued random variables, denoted $(X_0, X_1, \ldots)$ that satisfy the Markov property. The transition to state X_{n+1} given $X_n = x_n$ is formally captured by the notion of a transition kernel.

Definition 2.1 (Robert and Casella (2005)). A transition kernel on $\mathcal{X}$ is a function $\mathcal{X} \times \mathfrak{B}(\mathcal{X}) \ni (x, A) \mapsto K(x, A)$ that satisfies the following two properties: (i) For all $x \in \mathcal{X}$, $K(x, \cdot)$ is a probability measure and (ii) For all $A \in \mathfrak{B}(\mathcal{X})$, $K(\cdot, A)$ is $\mathfrak{B}(\mathcal{X})$-measurable.

Thus, the propagation of the state from step n to step $n+1$ is represented by $X_{n+1} \sim K(x_n, \cdot)$. When considering Markov chains, we will frequently be interested in the n-step transition probability measure from some initial state $X_0 = x_0$; we denote this probability measure by $K^n(x_0, \cdot) = \Pr[X_n \in \cdot | X_0 = x_0]$, which has the following expression:

$$K^n(x_0, A) = \underbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}_{(n-1)-\text{times}} K(x_0, \mathrm{d}x_1) K(x_1, \mathrm{d}x_2) \cdots K(x_{n-2}, \mathrm{d}x_{n-1}) K(x_{n-1}, A). \tag{1}$$

Of principle interest to the theory of Markov chains is the limiting behavior of the n-step transition probability measure.

Definition 2.2. The transition kernel K with n-step transition law K^n is ergodic for Π if, for every $x \in \mathcal{X}$, $\lim_{n\to\infty} \|K^n(x, \cdot) - \Pi(\cdot)\|_{\text{TV}} = 0$.

In the sequel, we will require continuity of sequences of transition kernels, which necessitates that we equip the space of transition kernels with a metric. A natural metric considers the worst-case total variation distance between kernels.

Definition 2.3. Two transition kernels K and K' on $\mathcal{X} \times \mathfrak{B}(\mathcal{X})$ are equal if $\sup_{x\in\mathcal{X}} \|K(x, \cdot) - K'(x, \cdot)\|_{\text{TV}} = 0$.

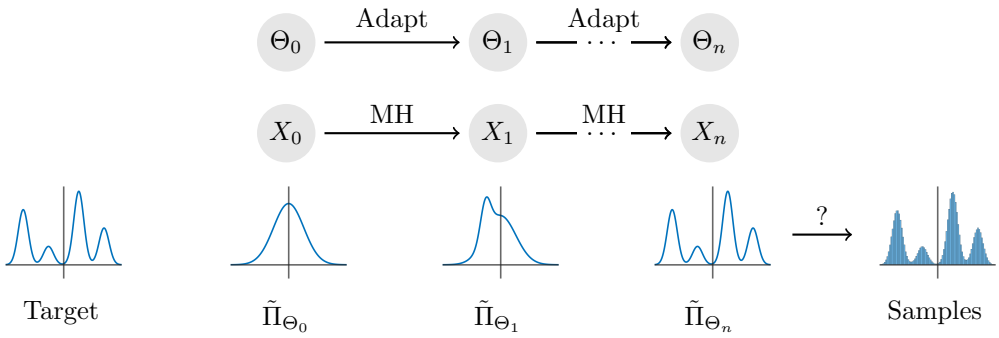


Figure 1: This work examines the convergence of adaptive Markov chain Monte Carlo algorithms using the independent Metropolis-Hastings algorithm when the proposal distribution is parameterized by a normalizing flow. In this illustration, we seek to draw samples from a target distribution. We begin with an initial parameter Θ_0 which parameterizes a simple proposal distribution, denoted $\tilde{\Pi}_{\Theta_0}$, which is a normalizing flow, and an initial state of the chain X_0; a sample from this proposal is accepted or rejected according to the Metropolis-Hastings criterion, yielding a transition to the state X_1. The parameters of the normalizing flow are thereafter adapted to produce a new proposal distribution $\tilde{\Pi}_{\Theta_1}$, which we hope is closer to the target distribution. Iterating this procedure we obtain both a sequence of states $(X_n)_{n\in\mathbb{N}}$ and a sequence of normalizing flow parameters $(\Theta_n)_{n\in\mathbb{N}}$. The principle question of this work is to establish when the sequence of states converges to the target density.

Proposition 2.4. *Let K and K' be transition kernels on $\mathcal{X}\times\mathfrak{B}(\mathcal{X})$. Then the function, $d(K,K') = \sup_{x\in\mathcal{X}}\|K(x,\cdot)-K'(x,\cdot)\|_{\mathrm{TV}}$ is a distance function on transition kernels.*

A proof is given in appendix D.

2.2 Independent Metropolis-Hastings

Definition 2.5. Let Π and $\tilde{\Pi}$ be two probability measures on $\mathfrak{B}(\mathcal{X})$ with densities with respect to some dominating measure μ given by π and $\tilde{\pi}$, respectively. Consider a Markov chain $(X_0, X_1, X_2, \ldots)$ constructed via the following procedure given an initial state of the Markov chain $X_0 = x_0$. First, randomly sample $\tilde{X}\sim\tilde{\Pi}$. Then set $X_{n+1} = \tilde{X}$ with probability $\min\left\{\frac{\pi(\tilde{X})\tilde{\pi}(X_n)}{\pi(X_n)\tilde{\pi}(\tilde{X})}, 1\right\}$ and otherwise set $X_{n+1} = X_n$. The Markov chain $(X_0, X_1, X_2, \ldots)$ is called the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}$.

Proposition 2.6. *Let K denote the transition kernel of the independent Metropolis-Hastings sampler. The stationary distribution of $(X_0, X_1, X_2, \ldots)$ is Π and if there exists a constant $M\geq 1$ such that $\frac{\pi(x)}{\tilde{\pi}(x)}\leq M,\ \forall\, x\in\mathrm{Supp}(\pi)$, then the independent Metropolis-Hastings sampler is uniformly ergodic in the sense that $\|K^n(x,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}}\leq 2\left(1-\frac{1}{M}\right)^n$.*

For a proof of these results, refer to Meyn and Tweedie (1993); Robert and Casella (2005). There is a question of when such a M as in proposition 2.6 will exist. Under a compactness condition and assumptions of continuity on both the proposal and target densities, then an affirmative existence result can be given.

Corollary 2.7. *If, in addition, $\mathcal{X}$ is a compact set and if π and $\tilde{\pi}$ are continuous on $\mathcal{X}$, and if $\mathrm{Supp}(\pi)\subseteq\mathrm{Supp}(\tilde{\pi})$ then there exists such an M as in proposition 2.6.*

A proof is given in appendix E. The transition kernel of the independent Metropolis-Hastings sampler has the form

$$K(x,\mathrm{d}x') = \min\left\{1, \frac{\pi(x')\tilde{\pi}(x)}{\pi(x)\tilde{\pi}(x')}\right\}\tilde{\pi}(x')\ \mu(\mathrm{d}x') + \left(1-\int_{\mathcal{X}}\min\left\{1,\frac{\pi(w)\tilde{\pi}(x)}{\pi(x)\tilde{\pi}(w)}\right\}\tilde{\pi}(w)\ \mu(\mathrm{d}w)\right)\delta_x(\mathrm{d}x'). \tag{2}$$

The first term in eq. (2) is the probability of an accepted transition from x to the region $\mathrm{d}x'$ whereas the second term is the probability of remaining at x, which only contributes if x lies in the region $\mathrm{d}x'$.

2.3 Adaptive Transition Kernels

As alluded to in section 1, the transition kernel may depend on parameters, denoted by θ and taking values in a set $\mathcal{Y}$. In this case, we express the dependency of the kernel K on its parameters by writing K_θ. In adaptive MCMC, given a target probability measure Π, we seek to strategically construct a sequence of transition kernels $(K_{\Theta_n})_{n\in\mathbb{N}}$ where $(\Theta_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{Y}$-valued random variables. Ideally, the sequence $(\Theta_n)_{n\in\mathbb{N}}$ will enable sampling from Π that becomes more effective with each step. In the adaptive MCMC framework, the one-step transition laws for X_{n+1} given $X_n = x_n$ and $\Theta_n = \theta_n$ is

$X_{n+1} \sim K_{\theta_n}(x_n, \cdot)$. The n-step transition law given $X_0 = x_0$ and $(\Theta_0 = \theta_0, \ldots, \Theta_{n-1} = \theta_{n-1})$ is

$$K^n_{(\theta_i)_{i=0}^{n-1}}(x_0, A) = \underbrace{\int_{\mathcal{X}} \cdots \int_{\mathcal{X}}}_{(n-1)-\text{times}} K_{\theta_0}(x_0, \mathrm{d}x_1) K_{\theta_1}(x_1, \mathrm{d}x_2) \cdots K_{\theta_{n-2}}(x_{n-2}, \mathrm{d}x_{n-1}) K_{\theta_{n-1}}(x_{n-1}, A) \tag{3}$$

Therefore, by the law of total expectation, the n-step transition law given $X_0 = x_0$ is $G^n(x_0, A) = \mathbb{E}_{(\Theta_0, \ldots, \Theta_{n-1})} K^n_{(\Theta_i)_{i=0}^{n-1}}(x_0, A)$, where the expectation is computed over the marginal distribution of the parameters. We now give a precise definition for what it means for an adaptive MCMC procedure to be ergodic.

Definition 2.8. The n-step transition law G^n is said to be ergodic for the probability measure Π if, for every $x \in \mathcal{X}$, $\lim_{n\to\infty} \|G^n(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} = 0$.

The principal theoretical tools of our analysis are the definitions of containment, simultaneous uniform ergodicity, and diminishing adaptation. Diminishing adaptation together with either containment or simultaneous uniform ergodicity implies ergodicity of the adaptive MCMC procedure in the sense of definition 2.8. The remainder of this section is a review of Roberts and Rosenthal (2007); Bai et al. (2011).

Definition 2.9. The sequence of Markov transition kernels $\{K_{\Theta_n}\}_{n\in\mathbb{N}}$ is said to exhibit diminishing adaptation if $\lim_{n\to\infty} d(K_{\Theta_{n+1}}, K_{\Theta_n}) = 0$ in probability.

Lemma 2.10 (Roberts and Rosenthal (2007))**.** *Suppose that $\Theta_{n+1} = \Theta_n$ w.p. $1 - \alpha_n$ and otherwise $\Theta_{n+1} = \Theta'_n$ where $\Theta'_n \in \mathcal{Y}$ is any other element of the index set. If $\lim_{n\to\infty} \alpha_n = 0$, then $(K_{\Theta_0}, K_{\Theta_1}, \ldots)$ exhibits diminishing adaptation.*

Definition 2.11. Define $W_\epsilon(x, K) = \inf\{n \geq 1 : \|K^n(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} < \epsilon\}$. The sequence $(\Theta_n)_{n\in\mathbb{N}}$ is said to exhibit containment if, for every $\epsilon > 0$, the sequence $(W_\epsilon(X_0, K_{\Theta_0}), W_\epsilon(X_1, K_{\Theta_1}), \ldots)$ is bounded in probability given $X_0 = x_0$ and $\Theta_0 = \theta_0$, where $X_{n+1} \sim K_{\Theta_n}(X_n, \cdot)$.

Containment states that for a particular stochastic sequence of adaptations $(\Theta_n)_{n\in\mathbb{N}}$ there is, with arbitrarily high probability, a finite number of steps one may take with any of the parameters in the sequence in order to be arbitrarily close to the target distribution. The following theorems give the relationships between diminishing adaptation, simultaneous uniform ergodicity, containment, and ergodicity of the adaptive MCMC procedure. The proofs of these results may be found in (Roberts and Rosenthal, 2007).

Theorem 2.12. *Let $\{K_\theta\}_{\theta\in\mathcal{Y}}$ be a family of Markov chain transition kernels that are all stationary for the same distribution Π. Suppose that the family satisfies definition N.1 and that the sequence $(\Theta_0, \Theta_1, \ldots)$ satisfies definition 2.9. Then the chain whose transitions are governed by $X_{n+1} \sim K_{\Theta_n}(X_n, \cdot)$ is ergodic for the distribution Π.*

Theorem 2.13. *Let $\{K_\theta\}_{\theta\in\mathcal{Y}}$ be a family of Markov chain transition kernels that are all stationary for the same distribution Π. Suppose that the sequence $(\Theta_0, \Theta_1, \ldots)$ satisfies definitions 2.9 and 2.11. Then the chain whose transitions are governed by $X_{n+1} \sim K_{\Theta_n}(X_n, \cdot)$ is ergodic for the distribution Π.*

3 RELATED WORK

A series of works recently investigated the learning of a proposal distribution for the independent Metropolis-Hastings sampler with normalizing flows, in particular for statistical mechanics field theories. For such models, Albergo et al. (2019) followed by Nicoli et al. (2020, 2021) used stochastic independent adaptations models following the optimization of the *reverse* Kullback-Leibler divergence (KL), as in *Example* 2 of the next section. While this strategy is successful when the target is unimodal, it is known to yield underdispersed approximation of the target distribution and to be prone to mode collapse. Within the framework of variational inference, Naesseth et al. (2020) proposed to address these issues by optimizing instead an approximate *forward* KL using simple parametric families for the proposal. In this case, adaptations are stochastic and rely on the previous states of the chain to estimate gradients of the approximate *forward* KL, called "pseudo-likelihood" in example 3 of the present paper. Incorporating normalizing flows, Gabrié et al. (2021a) successfully sampled multimodal distributions using an initialization that echoes the containment property. In the context of statistical field theories, Hackett et al. (2021) also demonstrated the need for *forward* KL training to assist sampling of multimodal distributions while surveying strategies to obtain training samples different from the adaptive MCMC discussed here. Hoffman et al. (2019) focuses on using normalizing flows to adapt Hamiltonian Monte Carlo to unfavorable posterior geometry by transforming a complicated posterior into a isotropic Gaussian.

Among the works above, ergodicity was only tested numerically. One exception is Gabrié et al. (2021a) where a convergence argument based on a continuous time analysis is developed under the assumption of perfect adaptation. The present paper provides a theoretical framework to analyze for the ergodicity of the methods presented in the body of work above. Though our work has focused on establishing ergodicity via the mechanism of Roberts and Rosenthal (2007), we note the work of Andrieu and Moulines (2006), which may

be used to establish an ergodicity theory. We concur with the statement in Roberts and Rosenthal (2007) that Andrieu and Moulines (2006) "requir[es] other technical hypotheses which may be difficult to verify in practice" and that diminishing adaptation and containment are "somewhat simpler conditions." Holden et al. (2009) considered the case of *independent* adaptations of the independent Metropolis-Hastings algorithm; however, this technique requires that accepted and rejected states be treated identically in the adaptation procedure, so we do not consider it further. We also note that Parno and Marzouk (2018) investigated the ergodicity of an adaptive MCMC using invertible maps. These works have similar aims but differ in several key details. For instance Parno and Marzouk (2018) focuses on establishing an ergodicity theory of triangular transformations of a Gaussian base measure, representing a local proposal distribution, which in practice is accomplished by employing third-degree Hermite polynomials. Our work, on the other hand, employs normalizing flows as *global* proposal mechanisms (independent of the current state of the chain). This necessitates a somewhat different treatment in order to establish ergodicity of the adaptive chain. As in this work, a pseudo-likelihod objective (see example 3) is employed in order to inform adaptations, but their objective is concave due to the choice of Hermite polynomials, whereas in the case of neural networks, the objective is more complex. Parno and Marzouk (2018) also assumes that parameter space $\mathcal{Y}$ is compact, which is untrue for typical parameterizations of normalizing flows, and insists on enforcing diminishing adaptation probabilistically (lemma 2.10) whereas we allow parameters to converge in probability (lemma 4.2 and theorem 4.3).

4 ANALYTICAL APPARATUS

We now consider the principle problem of this paper: *When can the adaptive independent Metropolis-Hastings sampler with proposal distribution parameterized by a normalizing flow be given an ergodicity theory?* We separate our discussion into two components wherein the adaptations are either deterministic or not necessarily independent of the state of the chain.

4.1 Deterministic Adaptations

Theorem 4.1. *Let Π be a probability measure with density π. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Suppose that $(\theta_0, \theta_1, \ldots)$ is a deterministic $\mathcal{Y}$-valued sequence. Let $(K_{\theta_n})_{n\in\mathbb{N}}$ be an associated sequence of Markov transition kernels of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\theta_n}$. Let $K^n(x_0, A)$ denote the n-step transition probability from x_0 to $A \in \mathfrak{B}(\mathcal{X})$ with law eq. (3). Then Π is the stationary distribution for K^n. Suppose further that for each $n \in \mathbb{N}$ there exists M_n satisfying $\pi(x) \leq M_n \tilde{\pi}_{\theta_n}(x)$ for all $x \in \mathrm{Supp}(\pi)$ Then, $\|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq 2\prod_{i=0}^{n-1}\left(1 - \frac{1}{M_i}\right)$.*

A proof is given in appendix B. We note that theorem 4.1 permits great generality in how θ parameterizes Π_θ; indeed, our analysis here, and subsequently, applies to any parameterized family of distributions.

Example 1. Let Π be a probability measure with density π. Let $\mathcal{Y} = \mathbb{R}^m$ and suppose that every $\theta \in \mathcal{Y}$ smoothly parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$ for which $\mathrm{Supp}(\pi) = \mathrm{Supp}(\tilde{\pi}_\theta)$. Consider the initial value problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\theta(t) = -\nabla_\theta \mathbb{KL}(\tilde{\pi}_{\theta(t)} \| \pi), \qquad \theta(0) = \theta_0, \tag{4}$$

where $\theta_0 \in \mathcal{Y}$. Let $(t_0, t_1, \ldots)$ be a deterministic sequence of times and let $\theta_n = \theta(t_n)$ for $n \in \mathbb{N}$. Consider the family of Markov chain transition operators of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\theta_n}$ with transition kernels K_{θ_n}. Then Π is the stationary distribution of the Markov chain whose transitions satisfy $X_{n+1} \sim K_{\theta_n}(X_n, \cdot)$. From the condition $\mathrm{Supp}(\pi) = \mathrm{Supp}(\tilde{\pi}_\theta)$ it follows that Π is the stationary distribution for each K_θ. Since $(\theta_0, \theta_1, \ldots)$ is a deterministic sequence, it follows from theorem 4.1 that Π is the stationary distribution. The particular mechanism of producing a deterministic sequence was not important; however, the time derivative eq. (4) was chosen because it begins to imitate the evolution encountered in normalizing flow loss functions. ‖

4.2 Non-Independent Adaptations

Notice that the decision to make the adaptation and the subsequent state of the chain dependent is not artificial or contrived; in fact, if such a procedure can be equipped with an ergodicity theory, then the resulting algorithm would have an important computational advantage. Specifically, it would require fewer evaluations of the target density (or the gradient of the target density) than the corresponding procedure with independent adaptations. For instance, the following adaptation scheme does not fall into the category of independent adaptations.

Example 2. Let Π be a probability measure with density π on a space $\mathcal{X}$. Let $\mathcal{Y} = \mathbb{R}^m$ and suppose that every $\theta \in \mathcal{Y}$ smoothly parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$ for which $\mathrm{Supp}(\pi) = \mathrm{Supp}(\tilde{\pi}_\theta)$. Let $\tilde{X} \sim \tilde{\Pi}_{\theta_{n-1}}$ be the proposal produced by the independent Metropolis-Hastings sampler of Π

given $\tilde{\Pi}_{\theta_{n-1}}$. Consider the adaptation

$$\theta_n = \theta_{n-1} - \epsilon \nabla_\theta \log \frac{\tilde{\pi}_{\theta_{n-1}}(\tilde{X}(\theta_{n-1}))}{\pi(\tilde{X}(\theta_{n-1}))}, \tag{5}$$

which can be interpreted as the single-sample approximation of the gradient flow of $\mathbb{KL}(\tilde{\pi}_{\theta_{n-1}} \| \pi)$. ‖

This motivates us to explore this direction. Definition 2.9 and the continuous mapping theorem (see theorem D.1) leads immediately to the following result.

Lemma 4.2. *Suppose that the map $\theta \mapsto K_\theta$ is continuous and that the sequence $(\Theta_0, \Theta_1, \ldots)$ converges in probability in $\mathcal{Y}$. Then $(K_{\Theta_0}, K_{\Theta_1}, \ldots)$ exhibits diminishing adaptation.*

A proof is given in appendix D. We now consider the question of the continuity of the mapping $\theta \mapsto K_\theta$.

Theorem 4.3. *Let $(\theta_1, \theta_2, \ldots)$ be a $\mathcal{Y}$-valued sequence converging to θ. Let π be a probability density function on a space $\mathcal{X}$ and let $\tilde{\pi}_\theta$ be a family of density functions on $\mathcal{X}$ indexed by θ such that the map $\theta \mapsto \tilde{\pi}_\theta$ is continuous. Suppose further that $\mathrm{Supp}(\tilde{\pi}_\theta) = \mathcal{X}$ for every $\theta \in \mathcal{Y}$. Let Π be the probability measure on $\mathfrak{B}(\mathcal{X})$ with density π and let $\tilde{\Pi}_\theta$ be the probability measure on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Let K_θ be the transition kernel of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_\theta$. Then $\lim_{n\to\infty} K_{\theta_n} = K_\theta$ (i.e. the mapping $\theta \mapsto K_\theta$ is continuous).*

A proof is given in appendix D. When training normalizing flows, it is typical to apply stochastic gradient descent to the minimization of some loss function. The question of when the iterates of stochastic gradient descent converge is an important question that has been recently treated in the case of non-convex losses. We refer the interested reader to Bottou (1999); Mertikopoulos et al. (2020) for conditions and results guaranteeing the convergence of stochastic gradient descent. In practice, the convergence of the sequence of normalizing flow parameters can be further encouraged by a decreasing learning rate schedule. In appendix N we discuss simultaneous uniform ergodicity on compact spaces and give some examples of normalizing flows works in these cases. The condition for geometric ergodicity of the independent Metropolis-Hastings sampler is that there exists $M \geq 1$ such that $\pi(x) \leq M \cdot \tilde{\pi}(x)$ for all $x \in \mathrm{Supp}(\pi)$ where π is the density of the target distribution and $\tilde{\pi}$ is the proposal density. By taking the logarithm of both sides and rearranging we obtain the equivalent inequality, $\log \pi(x) - \log \tilde{\pi}(x) \leq \log M$ for all $x \in \mathrm{Supp}(x)$.

Proposition 4.4. *Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Let $(\Theta_0, \Theta_1, \ldots)$ be a sequence of $\mathcal{Y}$-valued random variables and consider the family of Markov chain transition operators of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\Theta_n}$ with transition kernels K_{Θ_n}. Suppose that for all $\delta > 0$, there exists $M \equiv M(\delta) \in [1, \infty)$ such that*

$$\Pr\left[\log \pi(x) - \log \tilde{\pi}_{\Theta_n}(x) < \log M \quad \forall \quad x \in \mathcal{X}\right] \geq 1 - \delta, \tag{6}$$

for all $n \in \mathbb{N}$. Then, $(\Theta_n)_{n\in\mathbb{N}}$ exhibits containment.

A proof is given in appendix F. Regarding the tail condition in eq. (6), we note that the tail behaviour of the most popular normalizing flow architectures can be explicitly controlled, as shown by Jaini et al. (2020). Specifically, with Lipschitz triangular bijections (including most affine coupling flow implementations) the tail behaviour remains identical to that of the base measure. Thus, to ensure heavy tails in a flow one can simply replace the typical Gaussian base measure with a heavier tailed one, e.g., a Laplace or Student-t. An even stronger condition than eq. (6) is that $\Pr\left[|\log \pi(x) - \log \tilde{\pi}_{\Theta_n}(x)| < \log M \quad \forall \quad x \in \mathcal{X}\right] \geq 1 - \delta$. Thus, we see that containment can be obtained for the transition kernels of the independent Metropolis-Hastings sampler if, for every n, $\log \tilde{\pi}_{\Theta_n}$ is within $\log M$ of $\log \pi$ with probability $1 - \delta$. Note that M does not need to even be close to unity (equivalently, $\log M$ need not be close to zero) in order for containment to hold; it is sufficient merely that, with high probability, the sequence $(\Theta_n)_{n\in\mathbb{N}}$ does not produce arbitrarily poor approximations of $\log \pi$.

The loss functions used in estimating normalizing flows are chosen to encourage closeness of the approximation and the target density. For instance, if one chooses to minimize $\mathbb{KL}(\tilde{\pi}_\theta \| \pi)$ as a function of $\theta \in \mathcal{Y}$ then $\mathbb{KL}(\tilde{\pi}_\theta \| \pi) = 0 \iff \tilde{\pi}_\theta = \pi$. The minimization of a loss function that encourages the closeness of the approximation and the target density is certainly no guarantee that eq. (6) holds; however, it gives an indication that eq. (6) *might* be true. We turn our attention in the next section to the empirical evaluation of adaptive samplers using normalizing flows. Some obstacles that could prevent the conditions of proposition 4.4 from holding are stated in appendix L.

Example 3. Recently, Gabrié et al. (2021a,b) proposed to sample from Boltzmann distributions and posteriors over the parameters of physical systems by alternating between an independence Metropolis-Hastings algorithm whose proposal is represented as a RealNVP normalizing flow and local updates computed by the Metropolis-adjusted Langevin algorithm (MALA). In Gabrié et al. (2021a) the authors "demonstrate the importance of initializing the training with some *a priori* knowledge of the relevant modes." This incorporation of prior knowledge is done to avert mode-collapse. We can connect knowledge of modes to the property

of containment: by ensuring that the proposal density of the independent Metropolis-Hastings sampler places sufficient mass on all modes with high probability, one satisfies containment by proposition 4.4. The specific training procedure used by these samplers is to adapt parameters of the normalizing flow as $\Theta_{n+1} = \Theta_n + \epsilon_n \frac{1}{n} \sum_{i=0}^{n} \nabla \log \tilde{\pi}_{\Theta_n}(X_i)$ where $(X_i)_{i=0}^n$ are the states of the chain to the n^{th} step and $(\epsilon_i)_{i=0}^{\infty}$ are a sequence of adaptation step-sizes. Because the states of the chain can only be regarded as approximate samples from the target distribution, we understand this update as seeking to update a "pseudo-likelihood." Diminishing adaptation of this procedure can be enforced using either lemma 2.10 or via convergence and continuity using lemma 4.2. When diminishing adaptation and containment are satisfied, this adaptative algorithm produces an ergodic chain by theorem 2.13. ‖

5 EXPERIMENTS

Here we evaluate the adaptive independent Metropolis-Hastings algorithm following the "pseudo-likelihood objective", with *non-independent* adaptations, summarized in algorithm 1 in appendix M. As a baseline adaptive MCMC technique, we consider the random walk Metropolis method of Haario et al. (2001); we also compare against Langevin dynamics. To assess the ergodicity of samplers, we compare MCMC samples against analytic samples drawn from the target density, except in the case of the physical system wherein we use domain knowledge to compare against Langevin dynamics. Specifically, we choose 10,000 random unit vectors and project the samples of the adaptive chain onto the vector space spanned by the chosen unit vector; we then compare these one-dimensional quantities to the projection of the baseline samples from the target distribution and compute the two-sample Kolmogorov-Smirnov (KS) test statistic (Smirnov, 1948; Cuesta-Albertos et al., 2006). In appendix J, we show how adaptation can actually degrade sample quality at finite time.

Code implementing the experiments in the brownian bridge, the two-dimensional multimodal example, and the experiments of appendix J can be found at `https://github.com/JamesBrofos/Adaptive-Normalizing-Flow-Chains`.

5.1 Affine Flows in a Brownian Bridge

We consider sampling from a Gaussian process with the following mean and covariance functions: $\mu(t) = \sin(\pi t)$ and $\Sigma(t, s) = \min(t, s) - st$. For $0 < t, s < 1$, the covariance function identifies this distribution as a Brownian bridge whose mean is a sinusoid. We seek to sample this Gaussian process at 50 equally spaced times in the unit interval, yielding a fifty-dimensional target distribution. We estimate an affine normalizing flow from a Gaussian base distribution in order to sample from the target. Since the base distribution of the flow is Gaussian, and since affine transformations of Gaussian random variables remain Gaussian, in addition to the pseudo-likelihood training objective, we also consider gradient descent on the exact KL divergence between the target and the current proposal distribution. Minimization of the exact KL divergence is equivalent to maximum likelihood training, and therefore allows us to compare the efficiency lost by training with the pseudo-likelihood objective compared to the true likelihood. To enforce diminishing adaptation, we set a learning rate schedule for the gradient steps on the shift and scale of the affine transformation that converges to zero. In addition to Langevin dynamics, we also consider a preconditioned variant of the Metropolis-adjusted Langevin algorithm that uses the Hessian of the log-density to adapt proposals to the geometry of the target distribution (Girolami and Calderhead, 2011). Results shown in fig. 2 demonstrate the advantages of the adaptive independent Metropolis-Hastings samplers.

5.2 Two-Dimensional Examples

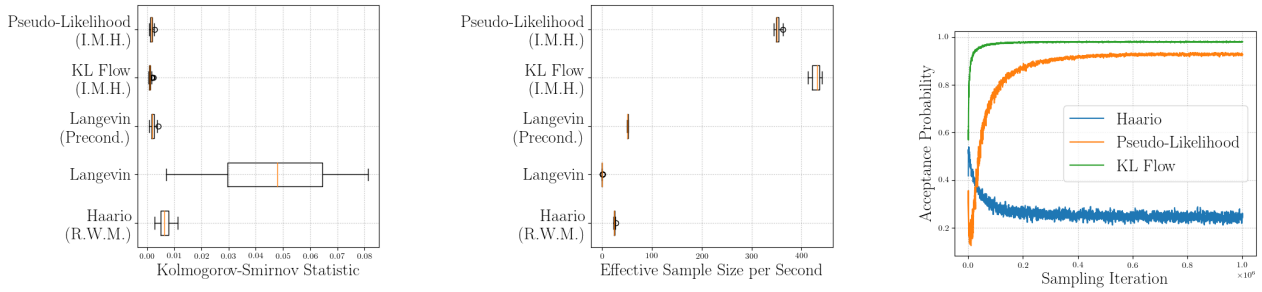
We use a RealNVP architecture to model a multimodal distribution and Neal's funnel distribution, both in $\mathbb{R}^2$. The multimodal density is a mixture of two Gaussians with a shared covariance structure given by $\Sigma = \text{diag}(1/100, 1/100)$. The two means of the component Gaussians are $(-2, 2)$ and $(2, -2)$. Neal's funnel distribution is defined by generating $v \sim \text{Normal}(0, 9)$ and $x \sim \text{Normal}(0, e^{-v})$, which defines a distribution in $\mathbb{R}^2$. To enforce diminishing adaptation, we set a learning rate schedule for the gradient steps on parameters of the RealNVP bijections that converges to zero. Results are shown in fig. 3. The expressivity of the RealNVP normalizing flow is key to building efficient proposals accommodating the distinct modes or the challenging multi-scale structure of Neal's funnel.

5.3 Analysis of a Physical Field System

We finally revisit a high-dimensional bi-modal example: the 1d-ϕ^4 system studied in Gabrié et al. (2021a). The statistics of a field $\phi : [0, 1] \to \mathbb{R}$ are given by the Boltzmann weight $e^{-\beta U}$ with the energy functional

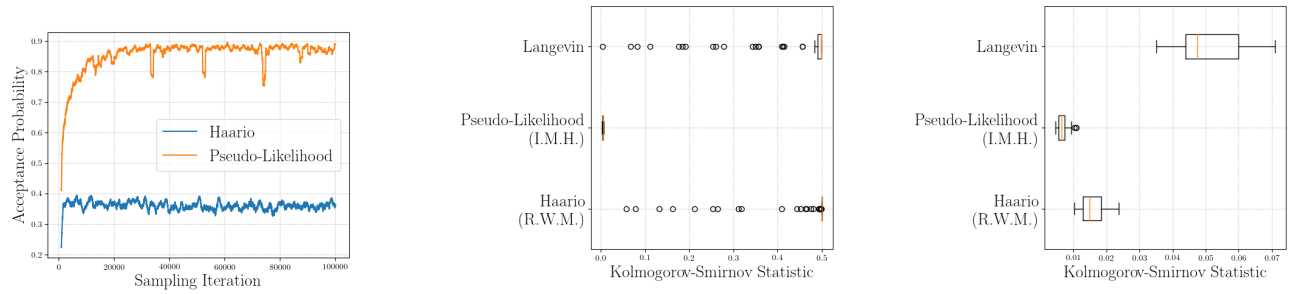
$$U(\phi) = \int_0^1 \left[\frac{a}{2} (\partial_s \phi)^2 + \frac{1}{4a} \left(1 - \phi^2(s)\right)^2 \right] \mathrm{d}s, \quad (7)$$

assuming boundary conditions $\phi(0) = \phi(1) = 0$. We discretize the field at 100 equally spaced locations between 0 and 1, for $a = 0.1$ and $\beta = 20$. Examples of



(a) Kolmogorov-Smirnov Statistics (b) ESS per Second (c) Acceptance Probability

Figure 2: Result of the Brownian bridge experiment. In assessing ergodicity according to the distribution of Kolmogorov-Smirnov statistics along random one-dimensional sub-spaces, the methods based on the independent Metropolis-Hastings algorithm and preconditioned Langevin dynamcis perform best. Langevin dynamics struggles in this posterior due to the multi-scale phenomena present in this distribution. In terms of the effective sample size per second of computation, the near-independent proposals and high acceptance rate of the independent Metropolis-Hastings sampler cause these algorithms to dominate. We also show the acceptance probability of the adaptive methods; we observe that the independent Metropolis-Hastings procedures enjoy adaptations that cause the acceptance probability to consistently improve over the course of learning.

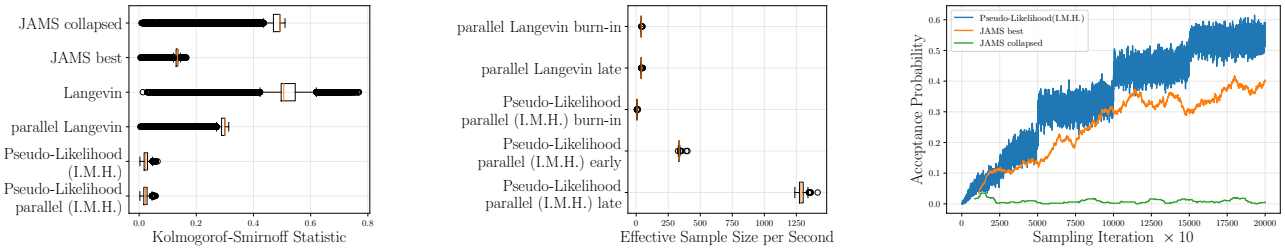


(a) (Multimodal) Acc. Prob. (b) (Multimodal) KS (c) (Neal Funnel) KS

Figure 3: Examination of the performance of MCMC methods on sampling from the multimodal mixture of Gaussians and Neal's funnel distribution. Both adaptive methods enjoy increasing acceptance rates in the multimodal distribution as a function of sampling iteration, but only the adaptive independent Metropolis-Hastings algorithm exhibits ergodicity for this distribution. Indeed, for the adaptive random walk and Langevin sampling methods, which are based on local updates, the multimodal distribution poses distinct challenges. In fact, both methods get stuck in one of the modes. By contrast, the adaptive independent Metropolis-Hastings samplers exhibit the best ergodicity of all methods considered. In Neal's funnel distribution, the adaptive independent Metropolis-Hastings algorithm possesses the best ergodicity.

states are plotted in fig. 6 of appendix O. The algorithm proposed in Gabrié et al. (2021a) is adapted with a learning rate schedule enforcing diminishing adaptations and a mixture transition kernel stochastically choosing from local Langevin updates or proposal sampling from the normalizing flow (appendix I shows that we can expect this mixture kernel to exhibit containment and diminishing adaption). Because the distribution is high-dimensional and multimodal, it is necessary[1] to run multiple parallel walkers initialized around the different modes. In this specific case, the energy and the distributions are even functions of ϕ. In the experiments, we initialize 100 walkers with uneven proportions in each mode (20-80) and test for the ergodicity of the parallel chains. Results are shown in fig. 4. Unlike the adaptive independent Metropolis-Hastings samplers, MALA single walkers are stuck in the mode they were initialized in and cannot recover the correct equal weights of the positive and negative mode. We also compare with the Jump Adaptive Multimodal Sampler (JAMS) of Pompe et al. (2020), using a MALA sampler for the local steps and

[1]This necessity can be lifted by employing an auxiliary fixed set of "training samples" featuring the two modes, in arbitrary proportions. These samples would drive the learning towards relevant regions, so that a random walker can then inform the adaption about the relative statistical weights of different modes



(a) Kolmogorov-Smirnov Statistics (b) ESS per Second (c) Acceptance Probability

Figure 4: Results of the ϕ^4 field experiment. As Langevin dynamics is unable to mix between the two modes, the better ergodicity of the independent Metropolis Hastings algorithm is reflected in Kolmogorov-Smirnov statistics as expected. The single chain Langevin has poorer ergodicity than its parallel chain equivalent, while for the I.M.H. a single chain approaches the ergodicity of the parallel setting. The Effective Sample Sizes are reported for chains of 1000 steps extracted at burn-in, after 4×10^4 iterations (early) and (late) when the NF proposal acceptance probability has reached 50%. Note that periodic jumps in acceptance correspond to iterations where learning rate was decreased.

an adaptive Gaussian mixture for the jumps and report results for the best of 10 runs and a typical failed run where the chains collapsed in one mode. We observe a KS mean statistic of 0.12 for one chain (best of 10 runs), compared to a KS mean statistic of 0.024 for the normalizing flow IMH, with the same number of iterations. The acceptance rate of jumps in JAMS reaches around 40% while the IMH gets to 55%. Additional details can be found in appendix O. Codes to reproduce this experiment can be found at https://github.com/marylou-gabrie/adapt-flow-ergo.

6 CONCLUSION

We have examined the question of when an adaptive independent Metropolis-Hastings sampler can be equipped with an ergodicity theory. We specifically consider the case wherein the proposal distribution is parameterized as a normalizing flow. We have considered the cases of deterministic adaptations, independent adaptations, and non-independent adaptations. For the non-independent adaptations case, we examine mechanisms by which to enforce the diminishing adaptation and containment conditions that together imply ergodicity.

Acknowledgements

M. G. would like to thank G. Rotskoff and E. Vanden-Eijnden for useful discussions about the physical field system experiments. We thank the Yale Center for Research Computing for use of the research computing infrastructure. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1752134. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. MAB was supported by an NSERC Discovery Grant and as part of the Vision: Science to Applications program, thanks in part to funding from the Canada First Research Excellence Fund. The work is also supported in part by NIH/NIGMS 1R01GM136780-01 and AFSOR FA9550-21-1-0317.

References

Albergo, M. S., Kanwar, G., and Shanahan, P. E. (2019). Flow-based generative models for Markov chain Monte Carlo in lattice field theory. *Physical Review D*, 100(3):034515.

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462 – 1505.

Bai, Y., Roberts, G., and Rosenthal, J. (2011). On the containment condition for adaptive markov chain monte carlo algorithms. *Adv. Appl. Stat.*, 21.

Berglund, N., Gesù, G. D., and Weber, H. (2017). An Eyring–Kramers law for the stochastic Allen–Cahn equation in dimension two. *Electronic Journal of Probability*, 22(none):1–27.

Bottou, L. (1999). *On-line Learning and Stochastic Approximations*, page 9–42. Publications of the Newton Institute. Cambridge University Press.

Cuesta-Albertos, J., Fraiman, R., and Ransford, T. (2006). Random projections and goodness-of-fit tests in infinite-dimensional spaces. *Bulletin Brazilian Mathematical Society*, 37:477–501.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2017).

Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. (2019). Reparameterizing distributions on lie groups. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 3244–3253. PMLR.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2021a). Adaptive monte carlo augmented with normalizing flows.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. (2021b). Efficient bayesian sampling using normalizing flows to assist markov chain monte carlo methods. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

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

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223 – 242.

Hackett, D. C., Hsieh, C.-C., Albergo, M. S., Boyda, D., Chen, J.-W., Chen, K.-F., Cranmer, K., Kanwar, G., and Shanahan, P. E. (2021). Flow-based sampling for multimodal distributions in lattice field theory. *arXiv preprint*, 2107.00734.

Hoffman, M., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. (2019). Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport.

Holden, L., Hauge, R., and Holden, M. (2009). Adaptive independent Metropolis–Hastings. *The Annals of Applied Probability*, 19(1):395 – 413.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. (2018). Neural autoregressive flows. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2078–2087. PMLR.

Jaini, P., Kobyzev, I., Yu, Y., and Brubaker, M. (2020). Tails of Lipschitz triangular flows. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4673–4681. PMLR.

Jaini, P., Selby, K. A., and Yu, Y. (2019). Sum-of-squares polynomial flow.

Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1.

Kolmogorov, A. N. (1960). *Foundations of the Theory of Probability*. Chelsea Pub Co, 2 edition.

Lebanon, G. (2017). *Probability*. Online manuscript.

Maire, F., Friel, N., Mira, A., and Raftery, A. E. (2019). Adaptive incremental mixture markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 28(4):790–805.

Mertikopoulos, P., Hallak, N., Kavis, A., and Cevher, V. (2020). On the almost sure convergence of stochastic gradient descent in non-convex problems. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1117–1128. Curran Associates, Inc.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Naesseth, C. A., Lindsten, F., and Blei, D. (2020). Markovian score climbing: Variational inference with KL(p——q). *Advances in Neural Information Processing Systems*, 2020-December(MCMC).

Nicoli, K. A., Anders, C. J., Funcke, L., Hartung, T., Jansen, K., Kessel, P., Nakajima, S., and Stornati, P. (2021). Estimation of Thermodynamic Observables in Lattice Field Theories with Deep Generative Models. *Physical Review Letters*, 126(3):32001.

Nicoli, K. A., Nakajima, S., Strodthoff, N., Samek, W., Müller, K. R., and Kessel, P. (2020). Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2).

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference.

Parno, M. D. and Marzouk, Y. M. (2018). Transport map accelerated markov chain monte carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Pompe, E., Holmes, C., and Łatuszyński, K. (2020). A framework for adaptive mcmc targeting multimodal distributions. *Annals of Statistics*, 48(5):2930–2952.

Rezende, D. J., Papamakarios, G., Racanière, S., Albergo, M. S., Kanwar, G., Shanahan, P. E., and Cranmer, K. (2020). Normalizing flows on tori and

spheres. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8083–8092. PMLR.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Roberts, G. and Rosenthal, J. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1.

Roberts, G. O. and Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of Applied Probability*, 44(2):458–475.

Royden, H. L. (1968). *Real analysis [by] H. L. Royden.* Macmillan, New York, 2d ed. edition.

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, 19(2):279 – 281.

Supplementary Material: Adaptation of the Independent Metropolis-Hastings Sampler with Normalizing Flow Proposals

A Review of Total Variation Distance

Similarity of probability measures can be assessed with respect to several criteria. A ubiquitous notion of distance between probability measures is given by the total variation norm of their difference.

Definition A.1. Let ν_1 and ν_2 be probability measures on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$. Then the total variation distance between ν_1 and ν_2 is defined by,

$$\|\nu_1(\cdot) - \nu_2(\cdot)\|_{\mathrm{TV}} = 2 \sup_{A \in \mathfrak{B}(\mathcal{X})} |\nu_1(A) - \nu_2(A)| \,. \tag{8}$$

The total variation distance is easily verified to be a proper distance in that it satisfies non-negativity, discernability, symmetry, and the triangle inequality. The total variation distance can therefore be understood as the largest possible disagreement between the probabilities assigned to any measurable set by ν_1 and ν_2. The total variation norm has the following equivalent representations which are occasionally useful.

Proposition A.2. *Within the context of definition A.1, the total variation distance between ν_1 and ν_2 is equivalently expressed as,*

$$\|\mu\|_{\mathrm{TV}} = \sup_{f \in \mathcal{M}} \left| \int_{\mathcal{X}} f(x) \; \nu_1(\mathrm{d}x) - \int_{\mathcal{X}} f(x) \; \nu_2(\mathrm{d}x) \right| \tag{9}$$

where $\mathcal{M} \overset{def.}{=} \{f : \mathcal{X} \to \mathbb{R} \; s.t. \; |f(x)| \leq 1 \; \forall \; x \in \mathcal{X}\}$.

For a proof of this result, and other equivalent characterizations of the total variation distance, we refer the interested reader to Roberts and Rosenthal (2004); Pollard (2001).

B Proofs Concerning Deterministic Adaptations

Proposition B.1 (Roberts and Rosenthal (2007)). *Suppose that $(\theta_0, \theta_1, \ldots)$ is a deterministic $\mathcal{Y}$-valued sequence. Let $(K_{\theta_n})_{n\in\mathbb{N}}$ be an associated sequence of Markov transition kernels. If Π is stationary for each K_{θ_n}, then Π is also the stationary distribution of the Markov chain whose transitions satisfy $X_{n+1} \sim K_{\theta_n}(X_n, \cdot)$.*

Proof. Let $A \in \mathfrak{B}(\mathcal{X})$ and suppose $X_n \sim \Pi$. We will show $X_{n+1} \sim \Pi$.

$$\begin{aligned}
\Pr[X_{n+1} \in A] &= \int_{\mathcal{X}} \Pr[X_{n+1} \in A | X_n = x] \cdot \Pr[X_n \in \mathrm{d}x] && (10)\\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \Pr[X_{n+1} \in A | X_n = x, \Theta_n = \theta] \cdot \delta_{\theta_n}(\mathrm{d}\theta) \cdot \Pr[X_n \in \mathrm{d}x] && (11)\\
&= \int_{\mathcal{X}} K_{\theta_n}(x, A) \cdot \Pr[X_n \in \mathrm{d}x] && (12)\\
&= \mathop{\mathbb{E}}_{x\sim\Pi}[K_{\theta_n}(x, A)] && (13)\\
&= \Pi(A). && (14)
\end{aligned}$$

□

Lemma B.2. *For each $n \in \mathbb{N}$,*

$$K_{\theta_n}(x, \mathrm{d}x') \geq \frac{\pi_\Pi(x')}{M_n}\mu(\mathrm{d}x'). \tag{15}$$

Proof. From eq. (2),

$$\begin{aligned}
K_{\theta_n}(x, \mathrm{d}x') &\geq \min\left\{1, \frac{\pi(x')\tilde{\pi}_{\theta_n}(x)}{\pi(x)\tilde{\pi}_{\theta_n}(x')}\right\} \tilde{\pi}_{\theta_n}(x')\ \mu(\mathrm{d}x') && (16)\\
&= \min\left\{\tilde{\pi}_{\theta_n}(x'), \frac{\pi(x')\tilde{\pi}_{\theta_n}(x)}{\pi(x)}\right\} \mu(\mathrm{d}x') && (17)\\
&\geq \min\left\{\tilde{\pi}_{\theta_n}(x'), \frac{\pi(x')}{M_n}\right\} \mu(\mathrm{d}x') && (18)\\
&= \frac{\pi(x')}{M_n}\ \mu(\mathrm{d}x') && (19)
\end{aligned}$$

□

Corollary B.3. *For any set $A \in \mathfrak{B}(\mathcal{X})$*

$$K_{\theta_n}(x, A) \geq \frac{1}{M_n}\Pi(A). \tag{20}$$

Proof. Integrate both sides of eq. (15) over the set A. □

Proof of Theorem 4.1. From corollary B.3 it follows that we may express the transition kernel at step $n \in \mathbb{N}$ as

$$\begin{aligned}
K_{\theta_n}(x, A) &= \frac{1}{M_n}\Pi(A) + \left(1 - \frac{1}{M_n}\right) \cdot \underbrace{\frac{K_{\theta_n}(x, A) - \frac{1}{M_n}\Pi(A)}{1 - \frac{1}{M_n}}}_{\tilde{K}_{\theta_n}(x, A)} && (21)\\
&= \frac{1}{M_n}\Pi(A) + \left(1 - \frac{1}{M_n}\right)\tilde{K}_{\theta_n}(x, A), && (22)
\end{aligned}$$

where $\tilde{K}_{\theta_n}(x, A)$ is another probability measure. With eq. (22), K_{θ_n} may be given the following interpretation: With probability $1/M_n$ generate the next state by the distribution Π and with probability $1 - 1/M_n$ generate the next state from the distribution $\tilde{K}_{\theta_n}$. Given an initial state $X_0 = x_0$, consider the Markov chain whose transitions

are generated according to $X_{n+1} \sim K_{\theta_n}(X_n, \cdot)$ with marginal laws $X_n \sim K^n(x_0, \cdot)$. From the representation in eq. (22) and proposition B.1, it follows that the total variation distance is zero as soon as we generate the next state from Π. Let T be the random variable representing the first step at which X_n is generated from Π. Then $K^n(x_0, \cdot) = \Pr[T \leq n]\,\Pi(\cdot) + \Pr[T > n]\,\tilde{K}^n(x_0, \cdot)$, where $\tilde{K}^n$ is the mixture component of K^n that is possibly not Π. Thus,

$$\begin{aligned}
\|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} &= \|\Pr[T \leq n]\,\Pi(\cdot) + \Pr[T > n]\,\tilde{K}^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} && (23)\\
&= \|\Pr[T > n]\,\tilde{K}^n(x_0, \cdot) - \Pr[T > n]\,\Pi(\cdot)\|_{\mathrm{TV}} && (24)\\
&= \Pr[T > n] \cdot \|\tilde{K}^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} && (25)\\
&\leq 2 \prod_{i=0}^{n-1} \left(1 - \frac{1}{M_i}\right), && (26)
\end{aligned}$$

since $T > n$ only if we generate the next state from $\tilde{K}_{\theta_i}$ for $i = 1, \ldots, n-1$, each of which occurs with probability $1 - 1/M_i$. □

Proposition B.4. *Suppose that $(\theta_0, \theta_1, \ldots)$ is a deterministic $\mathcal{Y}$-valued sequence. Let $(K_{\theta_n})_{n\in\mathbb{N}}$ be an associated sequence of Markov transition kernels. If Π is stationary for each K_{θ_n}, then*

$$\|K^{n+1}(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}}, \tag{27}$$

where K^n is defined in eq. (3).

Proof.

$$\begin{aligned}
\|K^{n+1}(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} &= \sup_{f\in\mathcal{M}} \left| \int_{\mathcal{X}} f(y) K^{n+1}(x_0, \mathrm{d}y) - \int_{\mathcal{X}} f(y) \Pi(\mathrm{d}y) \right| && (28)\\
&= \sup_{f\in\mathcal{M}} \left| \int_{\mathcal{X}} f(y) K^{n+1}(x_0, \mathrm{d}y) - \int_{\mathcal{X}} f(y) \int_{\mathcal{X}} K_{\theta_n}(w, \mathrm{d}y) \Pi(\mathrm{d}w) \right| && (29)\\
&= \sup_{f\in\mathcal{M}} \left| \int_{\mathcal{X}} f(y) \int_{\mathcal{X}} K^n(x_0, \mathrm{d}w) K_{\theta_n}(w, \mathrm{d}y) - \int_{\mathcal{X}} f(y) \int_{\mathcal{X}} K_{\theta_n}(w, \mathrm{d}y) \Pi(\mathrm{d}w) \right| && (30)\\
&= \sup_{f\in\mathcal{M}} \left| \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(y) K_{\theta_n}(w, \mathrm{d}y) \right) K^n(x_0, \mathrm{d}w) - \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(y) K_{\theta_n}(w, \mathrm{d}y) \right) \Pi(\mathrm{d}w) \right| && (31)\\
&\leq \sup_{f\in\mathcal{M}} \left| \int_{\mathcal{X}} f(w) K^n(x_0, \mathrm{d}w) - \int_{\mathcal{X}} f(w) \Pi(\mathrm{d}w) \right| && (32)\\
&= \|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}}. && (33)
\end{aligned}$$

□

Definition B.5. Let Π be a probability measure with density π. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Define the ergodic set of Π given $\mathcal{Y}$ as

$$\mathcal{Q} = \{\theta \in \mathcal{Y} : \text{ there exists } M_\theta < \infty \text{ such that } \pi(x) \leq M_\theta \tilde{\pi}_\theta(x) \ \ \forall \ \ x \in \mathrm{Supp}(\pi)\}. \tag{34}$$

The combination of proposition B.4 and definition B.5 allows one to give a slight generalization of theorem 4.1.

Corollary B.6. *Let Π be a probability measure with density π. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Suppose that $(\theta_0, \theta_1, \ldots)$ is a deterministic $\mathcal{Y}$-valued sequence. Let $(K_{\theta_n})_{n\in\mathbb{N}}$ be an associated sequence of Markov transition kernels of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\theta_n}$. Let $K^n(x_0, A)$ denote the n-step transition probability from x_0 to $A \in \mathfrak{B}(\mathcal{X})$. Then*

$$\|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq 2 \prod_{i=0}^{n-1} (1 - L_i), \tag{35}$$

where

$$L_i = \begin{cases} \frac{1}{M_i} & \text{if} \quad \theta_i \in \mathcal{Q} \\ 0 & \text{otherwise.} \end{cases} \tag{36}$$

Proof. The proof proceeds by induction. If $\theta_0 \in \mathcal{Q}$ then by argument in the proof of theorem 4.1 we have

$$\|K^1(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} = \|K_{\theta_0}(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} \tag{37}$$

$$\leq 2\left(1 - \frac{1}{M_0}\right). \tag{38}$$

If $\theta_0 \notin Q$ then we obtain the vacuously true inequality $\|K^1(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} \leq 1$. Now assume that $\|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} \leq \prod_{i=0}^{n-1}(1 - L_i)$. If $\theta_n \notin \mathcal{Q}$ then by proposition B.4 we have

$$\|K^{n+1}(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} \leq \|K^n(x_0, \cdot) - \Pi(\cdot)\|_{\text{TV}} \tag{39}$$

$$\leq 2\prod_{i=0}^{n-1}(1 - L_i) \tag{40}$$

$$= 2\prod_{i=0}^{n-1}(1 - L_i) \cdot 1 \tag{41}$$

$$= 2\prod_{i=0}^{n}(1 - L_i), \tag{42}$$

since $L_n = 0$. On the other hand, if $\theta_n \in \mathcal{Q}$ then, using the same argument as in the proof of theorem 4.1, the probability that none of $(X_0, \ldots, X_n)$ were drawn from Π is at most

$$\prod_{i=0}^{n-1}(1 - L_i). \tag{43}$$

Correspondingly, the probability that X_{n+1} is also not drawn from Π is $1 - 1/M_n$ so that the probability that none of $(X_0, \ldots, X_n, X_{n+1})$ is drawn from Π is at most

$$\left(1 - \frac{1}{M_n}\right)\prod_{i=0}^{n-1}(1 - L_i) = \prod_{i=0}^{n}(1 - L_i). \tag{44}$$

From this the conclusion follows. □

C Proofs Concerning Independent Adaptations

Theorem C.1. *Let Π be a probability measure with density π. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Suppose that $(\Theta_0, \Theta_1, \ldots)$ is a stochastic $\mathcal{Y}$-valued sequence. Let $(K_{\Theta_n})_{n\in\mathbb{N}}$ be an associated sequence of Markov transition kernels of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\Theta_n}$. Suppose that X_n and Θ_n are independent given the history of the chain to step $n-1$. Let $G^n(x_0, A)$ be the associated marginal transition law. Then*

$$\|G^n(x_0,\cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq 2 \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})} \left[\prod_{i=0}^{n-1} (1 - L_i)\right] \tag{45}$$

where $L_i = 1/M_i$ if $M_i < \infty$ and otherwise $L_i = 0$.

A proof is given in appendix C. This result was previously demonstrated in Holden et al. (2009), though we have offered a different proof procedure.

Example 4. Let Π be a probability measure with density π. Let $\mathcal{Y} = \mathbb{R}^m$ and suppose that every $\theta \in \mathcal{Y}$ smoothly parameterizes a probability measure $\tilde{\Pi}$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$ for which $\mathrm{Supp}(\pi) = \mathrm{Supp}(\tilde{\pi}_\theta)$. Consider the sequence of updates,

$$\theta_n = \theta_{n-1} - \epsilon \nabla_\theta \left(\frac{1}{s} \sum_{i=1}^{s} \log \frac{\tilde{\pi}_{\theta_{n-1}}(Y_s(\theta_{n-1}))}{\pi(Y_s(\theta_{n-1}))} \right) \tag{46}$$

where $Y_1, \ldots, Y_s \overset{\text{i.i.d.}}{\sim} \tilde{\Pi}_{\theta_{n-1}}$. This corresponds to the stochastic gradient approximation of example 1. Consider the family of Markov chain transition operators of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_{\theta_n}$ with transition kernels K_{θ_n}. Then Π is the stationary distribution of the Markov chain whose transitions satisfy $X_{n+1} \sim K_{\theta_n}(X_n, \cdot)$. To see this, let $\tilde{X}$ be a sample from $\tilde{\Pi}_{\theta_{n-1}}$ independent of $(Y_1, \ldots, Y_s)$ and let $U \sim \mathrm{Uniform}(0,1)$ be independent of both. Then $X_n = g(x_{n-1}, \theta_{n-1}, \tilde{X}, U)$ where g is given by,

$$g(x, \theta, \tilde{x}, u) = \begin{cases} \tilde{x} & \text{if } \; u < \min\left\{1, \frac{\pi(\tilde{x})\tilde{\pi}_\theta(x)}{\pi(x)\tilde{\pi}_\theta(\tilde{x})}\right\} \\ x & \text{otherwise} \end{cases} \tag{47}$$

and $\Theta_n = f(\theta_{n-1}, Y_1, \ldots, Y_s)$ where f is given by,

$$f(\theta, y_1(\theta), \ldots, y_s(\theta)) = \theta - \epsilon \nabla_\theta \left(\frac{1}{s} \sum_{i=1}^{s} \log \frac{\tilde{\pi}_\theta(y_s(\theta))}{\pi(y_s(\theta))} \right). \tag{48}$$

By lemma C.2, Θ_n and X_n are independent given the history of the chain to step $n-1$ and therefore, by proposition J.1, Π is the stationary distribution. ‖

Lemma C.2. *Suppose that $(X_1^{(a)}, \ldots, X_r^{(a)})$ and $(X_1^{(b)}, \ldots, X_s^{(b)})$ are two sets of random variables which are independent given the history of the chain to step $n-1$. Suppose that $\Theta_n = f(x_{n-1}, \theta_{n-1}, X_1^{(a)}, \ldots, X_r^{(a)})$ and $X_n = g(x_{n-1}, \theta_{n-1}, X_1^{(b)}, \ldots, X_s^{(b)})$ for two functions f and g. Then X_n and Θ_n are independent given the history of the chain to step $n-1$.*

Proof. The σ-algebra generated by Θ_n is a subset of the σ-algebra generated by $(X_1^{(a)}, \ldots, X_r^{(a)})$. Likewise, the σ-algebra generated by X_n is a subset of the σ-algebra generated by $(X_1^{(b)}, \ldots, X_s^{(b)})$. Since $(X_1^{(a)}, \ldots, X_r^{(a)})$ and $(X_1^{(b)}, \ldots, X_s^{(b)})$ are assumed independent given the history of the chain to step $n-1$, the conclusion follows. □

Proof of Theorem C.1.

$$\|G^n(x_0,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}} = \|\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})} K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}} \tag{49}$$

$$= 2\sup_{A\in\mathfrak{B}(\mathcal{X})}\left|\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})} K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \tag{50}$$

$$\leq \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})} 2\sup_{A\in\mathfrak{B}(\mathcal{X})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \tag{51}$$

$$= \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}} \tag{52}$$

$$\leq 2\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\left[\prod_{i=0}^{n-1}(1-L_i)\right], \tag{53}$$

where the first inequality can be deduced as follows: By Jensen's inequality,

$$\left|\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})} K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \leq \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right|. \tag{54}$$

Moreover,

$$\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \leq \sup_{A\in\mathfrak{B}(\mathcal{X})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \tag{55}$$

$$\implies \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \leq \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\sup_{A\in\mathfrak{B}(\mathcal{X})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \tag{56}$$

$$\implies \sup_{A\in\mathfrak{B}(\mathcal{X})}\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \leq \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\sup_{A\in\mathfrak{B}(\mathcal{X})}\left|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,A)-\Pi(A)\right| \tag{57}$$

The second inequality follows from corollary B.6 as follows:

$$\|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}} \leq 2\prod_{i=0}^{n-1}(1-L_i) \tag{58}$$

$$\implies \mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\|K^n_{(\Theta_0,\ldots,\Theta_{n-1})}(x_0,\cdot)-\Pi(\cdot)\|_{\mathrm{TV}} \leq 2\mathop{\mathbb{E}}_{(\Theta_0,\ldots,\Theta_{n-1})}\left[\prod_{i=0}^{n-1}(1-L_i)\right]. \tag{59}$$

□

D Proofs Concerning Continuity of Independent Metropolis-Hastings Transition Kernels

Theorem D.1. *Let f be a continuous function from the metric space $(\mathcal{X}, d_{\mathcal{X}})$ to the metric space $(\mathcal{Y}, d_{\mathcal{Y}})$. If $(X_0, X_1, \ldots)$ is a sequence of $\mathcal{X}$-valued random variables converging in probability to the random variable X then $(f(X_0), f(X_1), \ldots)$ is a sequence of $\mathcal{Y}$-valued converging in probability to $f(X)$.*

Proof of Lemma 4.2. Given that $\theta \mapsto K_\theta$ is continuous, if $(\Theta_0, \Theta_1, \ldots)$ converges in probability to Θ, we have immediately from theorem D.1 that $(K_{\Theta_0}, K_{\Theta_1}, \ldots)$ converges in probability to K_Θ. This means that for all $\epsilon > 0$ and $\delta > 0$, there exists $N(\epsilon, \delta) \in \mathbb{N}$ such that $\Pr\left[d(K_{\Theta_n}, K_\Theta) < \epsilon\right] \geq 1 - \delta$ for every $n \geq N$. For fixed $\epsilon > 0$ and $\delta > 0$, set $n \geq N(\epsilon/2, \delta)$ so that $\Pr\left[d(K_{\Theta_n}, K_\Theta) < \epsilon/2\right] \geq 1 - \delta$. Thus,

$$\begin{aligned}
\Pr\left[d(K_{\Theta_n}, K_{\Theta_{n+1}}) < \epsilon\right] &\geq \Pr\left[d(K_{\Theta_n}, K_\Theta) + d(K_{\Theta_{n+1}}, K_\Theta) < \epsilon\right] && (60)\\
&\geq \Pr\left[d(K_{\Theta_n}, K_\Theta) < \epsilon/2 \text{ and } d(K_{\Theta_{n+1}}, K_\Theta) < \epsilon/2\right] && (61)\\
&\geq \Pr\left[d(K_{\Theta_n}, K_\Theta) < \epsilon/2\right] + \Pr\left[d(K_{\Theta_{n+1}}, K_\Theta) < \epsilon/2\right] - 1 && (62)\\
&\geq 1 - \delta + 1 - \delta - 1 && (63)\\
&= 1 - 2\delta. && (64)
\end{aligned}$$

This establishes diminishing adaptation in the sense of definition 2.9. □

Proof of Proposition 2.4. To prove symmetry we write,

$$\begin{aligned}
d(K, K') &= \sup_{x \in \mathcal{X}} \|K(x, \cdot) - K'(x, \cdot)\|_{\mathrm{TV}} && (65)\\
&= \sup_{x \in \mathcal{X}} \|K'(x, \cdot) - K(x, \cdot)\|_{\mathrm{TV}} && (66)\\
&= d(K', K). && (67)
\end{aligned}$$

Identifiability follows from the definition of equality of Markov chain kernels given in definition 2.3. The triangle inequality is then proven as follows. Let K'' be another transition kernel on $\mathcal{X} \times \mathfrak{B}(\mathcal{X})$.

$$\begin{aligned}
d(K, K') &= \sup_{x \in \mathcal{X}} \|K(x, \cdot) - K'(x, \cdot)\|_{\mathrm{TV}} && (68)\\
&\leq \sup_{x \in \mathcal{X}} \left(\|K(x, \cdot) - K''(x, \cdot)\|_{\mathrm{TV}} + \|K''(x, \cdot) - K'(x, \cdot)\|_{\mathrm{TV}}\right) && (69)\\
&\leq \sup_{x \in \mathcal{X}} \|K(x, \cdot) - K''(x, \cdot)\|_{\mathrm{TV}} + \sup_{x \in \mathcal{X}} \|K''(x, \cdot) - K'(x, \cdot)\|_{\mathrm{TV}} && (70)\\
&= d(K, K'') + d(K'', K'). && (71)
\end{aligned}$$

□

In the sequel, we will limit our discussion to the transition kernel of the independent Metropolis-Hastings sampler. Recall that this transition kernel has the following form,

$$K_\theta(x, A) = \int_A \alpha_\theta(x, y)\tilde{\pi}_\theta(y)\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \alpha_\theta(x, w)\tilde{\pi}_\theta(w)\ \mu(\mathrm{d}w)\right) \mathbf{1}\left\{x \in A\right\}, \tag{72}$$

where

$$\alpha_\theta(x, y) = \min\left\{1, \frac{\pi(y)\tilde{\pi}_\theta(x)}{\pi(x)\tilde{\pi}_\theta(y)}\right\}. \tag{73}$$

Lemma D.2. *Let $(\theta_1, \theta_2, \ldots)$ be a $\mathcal{Y}$-valued sequence converging to θ. If for all $x \in \mathcal{X}$ and $A \in \mathfrak{B}(\mathcal{X})$ we have*

$$\begin{aligned}
\lim_{n \to \infty} \int_A \alpha_{\theta_n}(x, y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) &= \int_A \lim_{n \to \infty} \left[\alpha_{\theta_n}(x, y)\tilde{\pi}_{\theta_n}(y)\right]\ \mu(\mathrm{d}y) && (74)\\
&= \int_A \alpha_\theta(x, y)\tilde{\pi}_\theta(y)\ \mu(\mathrm{d}y). && (75)
\end{aligned}$$

then $\lim_{n \to \infty} K_{\theta_n} = K_\theta$.

Proof. By continuity of the distance function,

$$\lim_{n\to\infty} d(K_{\theta_n}, K_\theta) = d(\lim_{n\to\infty} K_{\theta_n}, K_\theta) \tag{76}$$

$$= \sup_{x\in\mathcal{X}} \sup_{A\in\mathfrak{B}(\mathcal{X})} \left|\lim_{n\to\infty} K_{\theta_n}(x,A) - K_\theta(x,A)\right|. \tag{77}$$

Therefore,

$$\lim_{n\to\infty} K_{\theta_n}(x,A) = \lim_{n\to\infty} \left(\int_A \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \alpha_{\theta_n}(x,w)\tilde{\pi}_{\theta_n}(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x\in A\}\right) \tag{78}$$

$$= \lim_{n\to\infty} \int_A \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) + \lim_{n\to\infty}\left(1 - \int_{\mathcal{X}} \alpha_{\theta_n}(x,w)\tilde{\pi}_{\theta_n}(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x\in A\} \tag{79}$$

$$= \int_A \lim_{n\to\infty} [\alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)]\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \lim_{n\to\infty}[\alpha_{\theta_n}(x,w)\tilde{\pi}_{\theta_n}(w)]\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x\in A\} \tag{80}$$

$$= \int_A \alpha_\theta(x,y)\tilde{\pi}_\theta(y)\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \alpha_\theta(x,w)\tilde{\pi}_\theta(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x\in A\} \tag{81}$$

$$= K_\theta(x,A). \tag{82}$$

Finally,

$$\lim_{n\to\infty} d(K_{\theta_n}, K_\theta) = \sup_{x\in\mathcal{X}} \sup_{A\in\mathfrak{B}(\mathcal{X})} \left|\lim_{n\to\infty} K_{\theta_n}(x,A) - K_\theta(x,A)\right| \tag{83}$$

$$= \sup_{x\in\mathcal{X}} \sup_{A\in\mathfrak{B}(\mathcal{X})} |K_\theta(x,A) - K_\theta(x,A)| \tag{84}$$

$$= 0. \tag{85}$$

□

The following result is called Scheffé's lemma; see Lebanon (2017); Pollard (2001).

Lemma D.3. *Let π_n be a sequence of probability densities that converge pointwise to another density π. Then, let $\Pi(A) = \int_A \pi(x)\ \mu(\mathrm{d}x)$ and $\Pi_n(A) = \int_A \pi_n(x)\ \mu(\mathrm{d}x)$ be the measures whose densities are π and π_n with respect to dominating measure μ, respectively. Then $\lim_{n\to\infty} \|\Pi(\cdot) - \Pi_n(\cdot)\|_{\mathrm{TV}} = 0$.*

We will also require the following theorem from Royden (1968, Page 270).

Theorem D.4. *Let $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ be a measurable space and let $(\Pi_n)_{n\in\mathbb{N}}$ be a sequence of probability measures converging to the probability measure Π. Let $\alpha_n : \mathcal{X} \to \mathbb{R}$ and $\beta_n : \mathcal{X} \to \mathbb{R}$ be two sequences of functions converging pointwise to the functions α and β, respectively. Suppose further that $|\alpha_n(x)| \leq \beta_n(x)$ for every $x \in \mathcal{X}$ and that,*

$$\lim_{n\to\infty} \int_{\mathcal{X}} \beta_n(x)\ \Pi_n(\mathrm{d}x) = \int_{\mathcal{X}} \beta(x)\ \Pi(\mathrm{d}x) < \infty. \tag{86}$$

Then,

$$\lim_{n\to\infty} \int_A \alpha_n(x)\ \Pi_n(\mathrm{d}x) = \int_A \alpha(x)\ \Pi(\mathrm{d}x), \tag{87}$$

for $A \in \mathfrak{B}(\mathcal{X})$.

Lemma D.5. *Suppose that for fixed $x \in \mathcal{X}$ the mapping $\theta \mapsto \tilde{\pi}_\theta(x)$ is continuous, that $y \in \mathcal{X}$, and that $\mathrm{Supp}(\tilde{\pi}_\theta) = \mathcal{X}$ for every $\theta \in \mathcal{Y}$. Let $(\theta_1, \theta_2, \ldots)$ be a $\mathcal{Y}$-valued sequence converging to θ. Then $\lim_{n\to\infty} \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y) = \alpha_\theta(x,y)\tilde{\pi}_\theta(y)$ pointwise.*

Proof.

$$\lim_{n\to\infty} \alpha_{\theta_n}(x,y) = \lim_{n\to\infty} \min\left\{1, \frac{\pi(y)\tilde{\pi}_{\theta_n}(x)}{\pi(x)\tilde{\pi}_{\theta_n}(y)}\right\} \tag{88}$$

$$= \left(\min\left\{1, \lim_{n\to\infty} \frac{\pi(y)\tilde{\pi}_{\theta_n}(x)}{\pi(x)\tilde{\pi}_{\theta_n}(y)}\right\}\right) \tag{89}$$

$$= \left(\min\left\{1, \frac{\pi(y)}{\pi(x)} \lim_{n\to\infty} \frac{\tilde{\pi}_{\theta_n}(x)}{\tilde{\pi}_{\theta_n}(y)}\right\}\right) \tag{90}$$

$$= \left(\min\left\{1, \frac{\pi(y)}{\pi(x)} \left(\lim_{n\to\infty} \tilde{\pi}_{\theta_n}(x)\right)\left(\lim_{n\to\infty} \frac{1}{\tilde{\pi}_{\theta_n}(y)}\right)\right\}\right) \tag{91}$$

$$= \left(\min\left\{1, \frac{\pi(y)}{\pi(x)} \left(\tilde{\pi}_{\theta}(x)\right)\left(\frac{1}{\lim_{n\to\infty} \tilde{\pi}_{\theta_n}(y)}\right)\right\}\right) \tag{92}$$

$$= \left(\min\left\{1, \frac{\pi(y)}{\pi(x)} \left(\tilde{\pi}_{\theta}(x)\right)\left(\frac{1}{\tilde{\pi}_{\theta}(y)}\right)\right\}\right) \tag{93}$$

$$= \alpha_{\theta}(x,y). \tag{94}$$

The assumption that $\mathrm{Supp}(\tilde{\pi}_\theta) = \mathcal{X}$ is used in eq. (92). □

Corollary D.6. *Let $(\theta_1, \theta_2, \ldots)$ be a $\mathcal{Y}$-valued sequence converging to θ. Let π be a probability density function on a compact space $\mathcal{X}$ and let $\tilde{\pi}_\theta$ be a family of density functions on $\mathcal{X}$ indexed by θ such that the map $\theta \mapsto \tilde{\pi}_\theta$ is continuous (i.e. $\pi_{\theta_n} \to \pi_\theta$). Assume further that $\mathrm{Supp}(\tilde{\pi}_\theta) = \mathcal{X}$ for every $\theta \in \mathcal{Y}$. Let $x \in \mathcal{X}$ be fixed and let $y \in \mathcal{X}$. Define*

$$\alpha_\theta(x,y) = \min\left\{1, \frac{\pi(y)\tilde{\pi}_\theta(x)}{\pi(x)\tilde{\pi}_\theta(y)}\right\}. \tag{95}$$

Then,

$$\lim_{n\to\infty} \int_A \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) = \int_A \alpha_\theta(x,y)\tilde{\pi}_\theta(y)\ \mu(\mathrm{d}y) \tag{96}$$

Proof. This follows immediate from theorem D.4 with $\beta_n(y) \equiv 1$, $\alpha_n(y) = \alpha_{\theta_n}(x,y)$ (which converges pointwise by lemma D.5) and the measures $\Pi_n(A) = \int_A \tilde{\pi}_{\theta_n}(x)\ \mu(\mathrm{d}x)$ and $\Pi(A) = \int_A \tilde{\pi}_\theta(x)\ \mu(\mathrm{d}x)$, which converge by lemma D.3. □

Proof of theorem 4.3. Fix $x \in \mathcal{X}$ and $A \in \mathfrak{B}(\mathcal{X})$. Thus,

$$\lim_{n\to\infty} K_{\theta_n}(x,A) = \lim_{n\to\infty} \left(\int_A \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \alpha_{\theta_n}(x,w)\tilde{\pi}_{\theta_n}(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x \in A\}\right) \tag{97}$$

$$= \lim_{n\to\infty} \int_A \alpha_{\theta_n}(x,y)\tilde{\pi}_{\theta_n}(y)\ \mu(\mathrm{d}y) + \lim_{n\to\infty} \left(1 - \int_{\mathcal{X}} \alpha_{\theta_n}(x,w)\tilde{\pi}_{\theta_n}(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x \in A\} \tag{98}$$

$$= \int_A \alpha_\theta(x,y)\tilde{\pi}_\theta(y)\ \mu(\mathrm{d}y) + \left(1 - \int_{\mathcal{X}} \alpha_\theta(x,w)\tilde{\pi}_\theta(w)\ \mu(\mathrm{d}w)\right)\mathbf{1}\{x \in A\} \tag{99}$$

$$= K_\theta(x,A). \tag{100}$$

where we have used corollary D.6 in eq. (99). The conclusion then follows from lemma D.2. □

E Proofs Concerning Simultaneous Uniform Ergodicity on Compact Spaces

Proof of Corollary 2.7. Because $\mathcal{X}$ is compact and π and $\tilde{\pi}$ are continuous, we know that π and $\tilde{\pi}$ attain maximum and minimum values on $\mathcal{X}$. Therefore, the ratio $\pi(x)/\tilde{\pi}(x)$ (i) does not diverge on $\text{Supp}(\pi)$ because $\text{Supp}(\pi) \subseteq \text{Supp}(\tilde{\pi})$ and (ii) is bounded by

$$\frac{\max_{x\in\mathcal{X}} \pi(x)}{\min_{x\in\mathcal{X}} \tilde{\pi}(x)}, \tag{101}$$

and M is at most this value, with equality if the maximum of π and the minimum of $\tilde{\pi}$ occur at the same point in $\mathcal{X}$. □

Proof of Proposition N.2. Define

$$M_\theta = \max_{x\in\text{Supp}(\pi)} \frac{\pi(x)}{\tilde{\pi}_\theta(x)} \tag{102}$$

and recall from corollary 2.7 that

$$\|K_\theta^n(x,\cdot) - \pi\|_{\text{TV}} \leq \left(1 - \frac{1}{M_\theta}\right)^n. \tag{103}$$

From eq. (101) and eq. (199), M_θ is bounded as

$$M_\theta \leq \frac{\max_{x\in\text{Supp}(\pi)} \pi(x)}{\min_{x\in\text{Supp}(\pi)} \tilde{\pi}_\theta(x)} \leq \frac{\max_{x\in\text{Supp}(\pi)} \pi(x)}{\delta} = M_\delta \tag{104}$$

The quantity M_δ does not depend on $\theta \in \mathcal{Y}$ and therefore we have, for all $\theta \in \mathcal{Y}$,

$$\|K_\theta^n(x,\cdot) - \pi\|_{\text{TV}} \leq \left(1 - \frac{1}{M_\delta}\right)^n. \tag{105}$$

Using this worst-case bound, we may find an n satisfying definition N.1 for all $\theta \in \mathcal{Y}$. □

Proof of Lemma N.3. Fix $\theta \in \mathcal{Y}$. Then

$$\begin{aligned}
\min_{x\in\text{Supp}(\pi)} \tilde{\pi}_\theta^*(x) &= \min_{x\in\text{Supp}(\pi)} (\beta\pi_{\Pi^*}^*(x) + (1-\alpha)\tilde{\pi}_\theta(x)) && (106)\\
&\geq \min_{x\in\text{Supp}(\pi)} \beta\pi_{\Pi^*}^*(x) && (107)\\
&= \delta. && (108)
\end{aligned}$$

The quantity δ is greater than zero since $\beta > 0$ and $\text{Supp}(\pi) \subseteq \text{Supp}(\pi_{\Pi^*}^*)$. Since θ was arbitrary, the conclusion follows. □

F Proofs Concerning Containment

Proof of Proposition 4.4.

$$\Pr\left[\log \pi(x) - \log \tilde{\pi}_{\Theta_n}(x) < \log M \quad \forall \quad x \in \mathcal{X}\right] \geq 1 - \delta \quad \forall \quad n \in \mathbb{N} \tag{109}$$

$$\implies \Pr\left[\frac{\pi(x)}{\tilde{\pi}_{\Theta_n}(x)} < M \quad \forall \quad x \in \mathcal{X}\right] \geq 1 - \delta \quad \forall \quad n \in \mathbb{N} \tag{110}$$

Then for all $\epsilon > 0$ there exists $N \equiv N(\epsilon, \delta) \in \mathbb{N}$ such that for all $x \in \mathcal{X}$,

$$\Pr\left[\|K^N_{\Theta_n}(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} < \epsilon\right] \geq 1 - \delta \quad \forall \quad n \in \mathbb{N} \tag{111}$$

$$\implies \Pr\left[W_\epsilon(x, K_{\Theta_n}) \leq N\right] \geq 1 - \delta \quad \forall \quad n \in \mathbb{N}. \tag{112}$$

□

G Adaptation via the Flow of the KL Divergence

Consider an adaptive sequence of parameters $(\theta_0, \theta_1, \ldots)$ which parameterize proposal densities $(\tilde{\pi}_{\theta_0}, \tilde{\pi}_{\theta_1}, \ldots)$. Given a target distribution Π with density π, the efficacy of the adaptive independent Metropolis-Hastings sampler is dictated by the ratio of the target density to the proposal density,

$$U(\theta) \stackrel{\text{def.}}{=} \sup_{x \in \text{Supp}(\pi)} \frac{\pi(x)}{\tilde{\pi}_\theta(x)}. \tag{113}$$

The smaller this upper bound, the better the mixing properties of the independent Metropolis-Hastings algorithm with proposal distribution $\tilde{\Pi}_{\theta_n}$. A central question is whether or not the sequence $(\theta_0, \theta_1, \ldots)$ actually produces improvements in these upper bounds; i.e. is $U(\theta_{n+1}) \leq U(\theta_n)$?

It is important that the bound in eq. (113) is actually the *least* upper bound. This is because an arbitrary upper bound may decrease while the least upper bound decreases.

In estimating the parameters of normalizing flows, it is typical that parameters follow, at least approximately, the gradient flow of a prescribed loss function, such as a KL divergence. Since gradient flows are initial value problems with deterministic solutions, by examining the case wherein adaptations are obtained exactly by gradient flow allows us to bypass the added difficulty of contending with stochastic adaptations.

Proposition G.1. *Let $\mathcal{X}$ be a state space and let $\theta \in \mathbb{R}^m$ parameterize a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$. Given a target density π, consider the function $U : \mathbb{R}^m \to \mathbb{R}$ defined by eq. (113) and assume further that U is smooth with respect to its argument. Let $L : \mathbb{R}^m \to \mathbb{R}$ be a loss function and consider the gradient flow $\theta_t = -\nabla L(\theta_t)$ given an initial condition θ_0. A sufficient condition that $U(\theta_{t+s}) \leq U(\theta_t)$ is that*

$$\nabla U(\theta_{t'}) \cdot \nabla L(\theta_{t'}) \geq 0, \tag{114}$$

where $t' \in (t, t+s)$; i.e. $\nabla L(\theta_{t'})$ is an ascent direction of U at $\theta_{t'}$.

Proof. By applying the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}t} U(\theta_t) = \nabla U(\theta_t) \cdot \dot{\theta}_t \tag{115}$$

$$= \nabla U(\theta_t) \cdot -\nabla L(\theta_t). \tag{116}$$

By the fundamental theorem of calculus,

$$U(\theta_{t+s}) - U(\theta_t) = \int_t^{t+s} \left(\frac{\mathrm{d}}{\mathrm{d}t'} U(\theta_{t'}) \right) \Big|_{t'=t''} \mathrm{d}t'' \tag{117}$$

$$= -\int_t^{t+s} \nabla U(\theta_{t''}) \cdot \nabla L(\theta_{t''}) \, \mathrm{d}t'' \tag{118}$$

$$\leq 0 \tag{119}$$

Therefore, $U(\theta_{t+s}) \leq U(\theta_t)$. □

While verifying the conditions of proposition G.1 in general appears a daunting task, we can do some analysis in simple cases.

Example 5. Consider the problem of sampling Normal$(0, 1)$ by adapting a proposal of the form Normal(μ, σ^2). Assume further that $\sigma^2 > 1$. We can deduce an upper bound on the ratio of the target density to the proposal density as follows:

$$\max_{x \in \mathbb{R}} \frac{\exp(-x^2/2)/\sqrt{2\pi}}{\exp(-(x-\mu)^2/2\sigma^2)/\sqrt{2\pi\sigma^2}} = \sigma \max_{x \in \mathbb{R}} \exp\left(-\frac{x^2}{2} + \frac{(x-\mu)^2}{2\sigma^2} \right) \tag{120}$$

$$\leq \sigma \exp\left(\frac{\mu^2}{2(\sigma^2 - 1)} \right), \tag{121}$$

which can be deduced by maximizing $-\frac{x^2}{2} + \frac{(x-\mu)^2}{2\sigma^2}$ using calculus. The reverse KL divergence between the proposal distribution and the target distribution is seen to be,

$$\mathbb{KL}(\text{Normal}(\mu, \sigma^2) \| \text{Normal}(0,1)) = -\log\sigma + \frac{\sigma^2 + \mu^2}{2} - \frac{1}{2}. \tag{122}$$

Consider the gradient flow of the reverse KL divergence:

$$\begin{aligned}
\dot{\mu}_t &= -\frac{\partial}{\partial \mu}\mathbb{KL}(\text{Normal}(\mu, \sigma^2) \| \text{Normal}(0,1)) && (123)\\
&= -\mu && (124)\\
\dot{\sigma}_t &= -\frac{\partial}{\partial \sigma}\mathbb{KL}(\text{Normal}(\mu, \sigma^2) \| \text{Normal}(0,1)) && (125)\\
&= \frac{1}{\sigma_t} - \sigma_t. && (126)
\end{aligned}$$

When we specify initial conditions $\mu_0 \in \mathbb{R}$ and $\sigma_0 > 1$, this produces an initial value problem. To verify that adapting by following the gradient flow of KL divergence produces a provable improvement to the upper bound deduced in eq. (121), it suffices to check that the time derivative of the upper bound in decreasing under the postulated gradient flow dynamics. That is,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t \exp\left(\frac{\mu_t^2}{2(\sigma_t^2 - 1)}\right) &= \frac{\mu_t \sigma_t \exp(\mu_t^2/(2(\sigma_t^2-1)))}{\sigma_t^2 - 1} \cdot -\mu_t \\
&\quad + \frac{\exp(\mu_t^2/(2(\sigma_t^2-1)))(-(\mu_t^2+2)\sigma_t^2 + \sigma_t^4 + 1)}{(\sigma_t^2-1)^2} \cdot \left(\frac{1}{\sigma_t} - \sigma_t\right) && (127)\\
&= -\frac{(\sigma_t - 1)(\sigma_t + 1)\exp(\mu_t^2/(2(\sigma_t^2-1)))}{\sigma_t}. && (128)
\end{aligned}$$

It follows that this is a negative quantity if we can establish that $\sigma_t > 1$. From the initial condition $\sigma_0 > 1$, it follows that the positive solution of the differential equation $\dot{\sigma}_t = \frac{1}{\sigma_t} - \sigma_t$ is $\sigma_t = \sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}$, so we see, indeed, that $\sigma_t > 1$. Therefore, the upper bound is a decreasing function of t given the prescribed gradient flow dynamics. The differential equation $\dot{\mu}_t = -\mu_t$ also has an explicit solution given the initial condition μ_0, which is $\mu_t = \mu_0 e^{-t}$. These explicit solutions to the gradient flow of the KL divergence allow us to express the evolution of the upper bound concretely as,

$$\sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}\exp\left(\frac{\mu_0^2 e^{-2t}}{2e^{-2t}(\sigma_0^2 - 1)}\right) = \sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}\exp\left(\frac{\mu_0^2}{2(\sigma_0^2 - 1)}\right) \tag{129}$$

This is an intriguing formula since it suggests that although the sequence M_t is decreasing, it decreases only to a non-unit limit $\exp(\mu_0^2/(2(\sigma_0^2 - 1)))$; indeed, unless $\mu_0 = 0$, the limit of this upper bound does not approach one. For the purposes of MCMC, this may be acceptable, since uniform ergodicity can be obtained so long as the bound is finite; however, were the upper bound to equal one, this would be optimal. Connecting this back to the question of adaptation, choosing an increasing sequence of times $t_0 < t_1 < t_2 < \ldots$ and consider using $\text{Normal}(\mu_{t_n}, \sigma_{t_n}^2)$ as the proposal distribution at step n. The Doeblin coefficient at step n is therefore,

$$L_n = \frac{1}{\sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}\exp\left(\frac{\mu_0^2}{2(\sigma_0^2 - 1)}\right)}. \tag{130}$$

Finally, let us remark that the undesirable property that the upper bounds do not converge to unity can be easily corrected. The principle issue is that the factors of e^{-2t} cancel in the exponent. However, consider that instead of using (μ_t, σ_t) to inform adaptations one instead uses (μ_{2t}, σ_t) so that the mean value is further along in the solution to its initial value problem than the scale. Plugging this into the formula for the upper bound yields,

$$\sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}\exp\left(\frac{\mu_0^2 e^{-4t}}{2e^{-2t}(\sigma_0^2 - 1)}\right) = \sqrt{e^{-2t}(\sigma_0^2 - 1) + 1}\exp\left(\frac{\mu_0^2 e^{-2t}}{2(\sigma_0^2 - 1)}\right), \tag{131}$$

which converges to unity as $t \to \infty$ as desired.

Instead of the reverse KL divergence we may consider the forward KL divergence between the target distribution and the proposal.

$$\mathbb{KL}(\text{Normal}(0,1)\|\text{Normal}(\mu,\sigma^2)) = \log\sigma + \frac{1+\mu^2}{2\sigma^2} - \frac{1}{2}. \tag{132}$$

The forward KL divergence produces the following equations of motion.

$$\dot{\mu}_t = -\frac{\mu_t}{\sigma_t^2} \tag{133}$$

$$\dot{\sigma}_t = \frac{\mu_t^2+1}{\sigma_t^3} - \frac{1}{\sigma_t} \tag{134}$$

Applying the chain rule to eq. (121) with these equations of motion yields the following time derivative of the upper bound,

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma_t \exp\left(\frac{\mu_t^2}{2(\sigma_t^2-1)}\right) = \exp\left(\frac{\mu_t^2}{2(\sigma_t^2-1)}\right)\frac{-\mu_t^4\sigma^2 + \mu_t^2\sigma_t^4 - 2\mu_t^2\sigma_t^2 + \mu_t^2 - \sigma_t^6 + 3\sigma_t^4 - 3\sigma_t^2 + 1}{\sigma_t^3(\sigma_t^2-1)^2}. \tag{135}$$

This derivative is less than or equal to zero iff

$$-\mu_t^4\sigma^2 + \mu_t^2\sigma_t^4 - 2\mu_t^2\sigma_t^2 + \mu_t^2 - \sigma_t^6 + 3\sigma_t^4 - 3\sigma_t^2 + 1 \leq 0 \tag{136}$$

$$\iff \mu_t^2(\sigma_t^2-1)^2 \leq \mu_t^4\sigma_t^2 + (\sigma_t^2-1)^3, \tag{137}$$

which is true for $\sigma_t^2 > 1$.

‖

H Adaptation of a Kernel Density Proposal Distribution

A principle difficulty in using normalizing flows as proposal distributions is that it is unclear whether or not a given adaptation of the neural network parameters will provably decrease the upper bound on the ratio of the target density and the normalizing flow density. Analyzing this property theoretically in the case of normalizing flows does not appear to be forthcoming. Nevertheless, we have been able to analyze certain behaviors of proposals based on kernel density estimators. A version of this procedure (to use kernel density estimators as a proposal in independent Metropolis-Hastings) was previously pursued in Maire et al. (2019); however, they do not appear to have looked at the question of when the addition of a new component into the mixture is beneficial, which is the topic under consideration in the sequel.

Let $x_1, \ldots, x_n \in \mathbb{R}^m$. We consider the kernel density estimator computed by,

$$\tilde{K}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\mathrm{Vol}(\bar{B}_\epsilon(x_i))} \mathbf{1}\left\{x \in \bar{B}_\epsilon(x_i)\right\}. \tag{138}$$

where $\bar{B}_\epsilon(x)$ is the closed ball of radius ϵ centered at x. For notational convenience, we observe that $V = \mathrm{Vol}(\bar{B}_\epsilon(x)) = \mathrm{Vol}(\bar{B}_\epsilon(x'))$ for any $x, x' \in \mathbb{R}^m$ and we write $m_n(x) = \#\left\{i \in (1, \ldots, n) : x \in \bar{B}_\epsilon(x_i)\right\}$. Therefore, we have the simple expression for the kernel density estimator as $\tilde{K}_n(x) = m_n(x)/nV$.

Suppose that Π is a probability measure on $\mathbb{R}^m$ with compactly supported density $\pi : \mathbb{R}^m \to \mathbb{R}_+$. Define,

$$M_n = nV \sup_{x \in \mathrm{Supp}(\pi)} \frac{\pi(x)}{m_n(x)} \tag{139}$$

Hence $\pi(x)/\tilde{K}_n(x) \leq M_n$. Given a new observation $x_{n+1} \in \mathbb{R}^m$, we would like to understand conditions under which one can show $M_{n+1} \leq M_n$. This means that the inclusion of a new observation in the kernel density estimate reduces the upper bound on the ratio of the target density and the proposal density. To discuss this, we begin with two definitions.

Definition H.1. The inner bound is defined by,

$$M_n' = \sup_{x \in \bar{B}_\epsilon(x_{n+1}) \cap \mathrm{Supp}(\pi)} \frac{\pi(x)}{m_n(x)}. \tag{140}$$

Definition H.2. The outer bound is defined by,

$$M_n'' = \sup_{x \in \mathrm{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1})} \frac{\pi(x)}{m_n(x)}. \tag{141}$$

Lemma H.3. *Let A and B be sets. Then* $\max\{\sup A, \sup B\} = \sup A \cup B$.

Proof. Since $A \subset A \cup B$, it is immediate that $\sup A \leq \sup A \cup B$. Identical reasoning shows that $\sup B \leq \sup A \cup B$. Therefore, $\max\{\sup A, \sup B\} \leq \sup A \cup B$. Now suppose without loss of generality that $\sup B \geq \sup A$. Then for all $a \in A$ we have $a \leq \sup B$; moreover, for all $b \in B$, $b \leq \sup B$. Therefore, for all $x \in A \cup B$, $x \leq \sup B$. Thus, $\sup A \cup B \leq \sup B$, since $\sup A \cup B$ is by definition the least upper bound. Applying identical reasoning to the case $\sup A \geq \sup B$ reveals $\sup A \cup B \leq \max\{\sup A, \sup B\}$. Thus, we must have $\sup A \cup B = \max\{\sup A, \sup B\}$. □

Corollary H.4. *Since* $\mathrm{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1}) = \mathrm{Supp}(\pi) \cap \bar{B}_\epsilon(x_{n+1})^c$*, it follows from the distributive law of set relationships that,*

$$\left\{\mathrm{Supp}(\pi) \cap \bar{B}_\epsilon(x_{n+1})\right\} \cup \left\{\mathrm{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1})\right\} = \mathrm{Supp}(\pi). \tag{142}$$

Applying lemma H.3 shows that $nV \max\{M_n', M_n''\} = M_n$.

Lemma H.5. *The kernel density estimator* $\tilde{K}_{n+1}(x)$ *can be written as,*

$$\tilde{K}_{n+1}(x) = \begin{cases} \frac{m_n(x)}{(n+1)V} & \text{if } x \notin \bar{B}_\epsilon(x_{n+1}) \\ \frac{m_n(x)+1}{(n+1)V} & \text{otherwise.} \end{cases} \tag{143}$$

Proof. This follows immediately from the equation,

$$\tilde{K}_{n+1}(x) = \frac{1}{(n+1)V}\sum_{i=1}^{n+1}\mathbf{1}\left\{x \in \bar{B}_\epsilon(x_i)\right\} \tag{144}$$

$$= \frac{1}{(n+1)V}\left(m_n(x) + \mathbf{1}\left\{x \in \bar{B}_\epsilon(x_{n+1})\right\}\right). \tag{145}$$

□

Lemma H.6. *We have,*

$$\sup_{x\in\mathrm{Supp}(\pi)\setminus\bar{B}_\epsilon(x_{n+1})}\frac{\pi(x)}{\tilde{K}_{n+1}(x)} = (n+1)VM_n''. \tag{146}$$

Proof. For $x \in \mathrm{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1})$,

$$\frac{\pi(x)}{\tilde{K}_{n+1}(x)} = (n+1)V\frac{\pi(x)}{m_n(x)}. \tag{147}$$

Therefore,

$$\sup_{x\in\mathrm{Supp}(\pi)\setminus\bar{B}_\epsilon(x_{n+1})}\frac{\pi(x)}{\tilde{K}_{n+1}(x)} = (n+1)V\sup_{x\in\mathrm{Supp}(\pi)\setminus\bar{B}_\epsilon(x_{n+1})}\frac{\pi(x)}{m_n(x)} \tag{148}$$

$$= (n+1)VM_n''. \tag{149}$$

□

Lemma H.7. *We have,*

$$\sup_{x\in\mathrm{Supp}(\pi)\cap\bar{B}_\epsilon(x_{n+1})}\frac{\pi(x)}{\tilde{K}_{n+1}(x)} \leq nVQ_nM_n', \tag{150}$$

where

$$Q_n = \sup_{x\in\mathrm{Supp}(\pi)\cap\bar{B}_\epsilon(x_{n+1})}\frac{m_n(x)(n+1)}{(m_n(x)+1)n} \leq 1. \tag{151}$$

Proof. Within $\mathrm{Supp}(\pi)\cap\bar{B}_\epsilon(x_{n+1})$ we have the bound,

$$nV\frac{\pi(x)}{m_n(x)} \leq nVM_n' \tag{152}$$

$$\implies \frac{\pi(x)}{m_n(x)/nV}\cdot\frac{m_n(x)/n}{(m_n(x)+1)/(n+1)} \leq \frac{m_n(x)/n}{(m_n(x)+1)/(n+1)}nVM_n' \tag{153}$$

$$\implies \frac{\pi(x)}{\tilde{K}_{n+1}(x)} \leq nV\frac{m_n(x)/n}{(m_n(x)+1)/(n+1)}M_n' \tag{154}$$

Taking the supremum on both sides yields,

$$\sup_{x\in\mathrm{Supp}(\pi)\cap\bar{B}_\epsilon(x_{n+1})}\frac{\pi(x)}{\tilde{K}_{n+1}(x)} \leq nVQ_nM_n'. \tag{155}$$

Since $m_n(x) \leq n$, it follows that $Q_n \leq 1$. □

Proposition H.8. *If $M_n'' > M_n'$, then $M_{n+1} > M_n$. In this case, the inclusion of the new observation degrades the quality of the proposal distribution.*

Proof. Since $M_n'' > M_n'$, it follows that

$$\sup_{x \in \text{Supp}(\pi) \cap \bar{B}_\epsilon(x_{n+1})} \frac{\pi(x)}{\tilde{K}_{n+1}(x)} \leq nVQ_nM_n' \tag{156}$$
$$\leq nVM_n' \tag{157}$$
$$< nVM_n'' \tag{158}$$
$$\leq (n+1)VM_n'' \tag{159}$$
$$= \sup_{x \in \text{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1})} \frac{\pi(x)}{\tilde{K}_{n+1}(x)}. \tag{160}$$

Hence $M_{n+1} = (n+1)VM_n'' > nVM_n'' = M_n$. □

Proposition H.9. *It is necessary and sufficient that* $(1+1/n)M_n'' \leq M_n'$ *in order for* $M_{n+1} \leq M_n$.

Proof. To establish sufficiency, we have:

$$M_{n+1} \leq \max\{(n+1)VM_n'', nVQ_nM_n'\} \tag{161}$$
$$\leq \max\{nVM_n', nVQ_nM_n'\} \tag{162}$$
$$= nVM_n' \tag{163}$$
$$= M_n \tag{164}$$

To show that this is actually necessary, consider $M_n'' \leq M_n' < (1+1/n)M_n''$. Then, from lemma H.6 we know,

$$\sup_{x \in \text{Supp}(\pi) \setminus \bar{B}_\epsilon(x_{n+1})} \frac{\pi(x)}{\tilde{K}_{n+1}(x)} = (n+1)VM_n''. \tag{165}$$

But from lemma H.7 we have,

$$\sup_{x \in \text{Supp}(\pi) \cap \bar{B}_\epsilon(x_{n+1})} \frac{\pi(x)}{\tilde{K}_{n+1}(x)} \leq nVQ_nM_n' \tag{166}$$
$$< (n+1)VQ_nM_n'' \tag{167}$$
$$\leq (n+1)VM_n'' \tag{168}$$

Now $M_n = nVM_n' < (n+1)VM_n'' = M_{n+1}$. □

Proposition H.8 informs us that if the worst case bound is outside of $\bar{B}_\epsilon(x_{n+1})$, then the adaptation has actually made the bound worse. This is because the density outside of $\bar{B}_\epsilon(x_{n+1})$ behaves in a predictable manner: it is decreased by a factor of $n/(n+1)$. Therefore, if the worst case bound on the ratio occurs outside of $\bar{B}_\epsilon(x_{n+1})$, the ratio can only get worse. At the same time, proposition H.9 informs us that the worst-case bound on the ratio inside $\bar{B}_\epsilon(x_{n+1})$ must be greater than the worst-case bound outside of $\bar{B}_\epsilon(x_{n+1})$ by at least a factor of $(1+1/n)$ in order for the inclusion of the additional observation x_{n+1} to improve the worst-case bound of the ratio of the target density to the kernel density estimator.

I Diminishing Adaptation and Containment for Mixture Kernels

Proposition I.1. *Let Π be a probability measure with density π with respect to measure μ. Let $\theta \in \mathcal{Y}$ and suppose that θ parameterizes a transition kernel K_θ. Let $(\Theta_0, \Theta_1, \ldots)$ be a sequence of $\mathcal{Y}$-valued random variables. Suppose that with probability $1-\delta$ there exists $M \equiv M(\delta) \in \mathbb{N}$ such that $K_{\Theta_n}(x, \mathrm{d}x') \geq \frac{\pi(x')}{M}\mu(\mathrm{d}x')$ for every $n = 1, 2, \ldots$. Then $(K_{\Theta_0}, K_{\Theta_1}, \ldots)$ exhibits containment.*

Proof. By assumption, with probability $1-\delta$, there exists $M \equiv M(\delta) \in \mathbb{N}$ such that $K_{\Theta_n}(x, \mathrm{d}x') \geq \frac{\pi(x')}{M}\mu(\mathrm{d}x')$ for every $n = 1, 2, \ldots$ and any $x \in \mathcal{X}$. Therefore, with probability $1-\delta$, there exists $M \equiv M(\delta) \in \mathbb{N}$ such that,

$$\|K^m_{\Theta_n}(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \left(1 - \frac{1}{M}\right)^m \tag{169}$$

for any $m \in \mathbb{N}$, every $n = 1, 2, \ldots$, and $x \in \mathcal{X}$. Let $\epsilon > 0$ be arbitrary. Then, with probability $1-\delta$, there exists $N = N(\epsilon, \delta) \in \mathbb{N}$ such that

$$\|K^N_{\Theta_n}(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon. \tag{170}$$

for every $n = 1, 2, \ldots$ and any $x \in \mathcal{X}$. Namely, the choice

$$N(\epsilon, \delta) = \left\lceil \frac{\log \epsilon}{\log\left(1 - \frac{1}{M(\delta)}\right)} \right\rceil \tag{171}$$

suffices. Thus as a special case, with probability $1-\delta$, there exists $N = N(\epsilon, \delta) \in \mathbb{N}$ such that

$$\|K^N_{\Theta_n}(X_n, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon. \tag{172}$$

for every $n = 1, 2, \ldots$. Define the function,

$$W_\epsilon(x, \theta) = \inf\left\{n \in \mathbb{N} : \|K^n_\theta(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon\right\}. \tag{173}$$

Hence, with probability $1-\delta$, there exists $N = N(\epsilon, \delta) \in \mathbb{N}$ such that

$$W_\epsilon(X_n, \Theta_n) \leq N \tag{174}$$

for every $n = 1, 2, \ldots$. This is the containment condition. □

Proposition I.2. *Let Π be probability measure with density π with respect to measure μ. Let $\theta \in \mathcal{Y}$ and suppose that θ parameterizes a probability measure $\tilde{\Pi}_\theta$ with density $\tilde{\pi}_\theta$. Let K_θ be the transition kernel of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_\theta$:*

$$K_\theta(x, \mathrm{d}x') = \min\left\{1, \frac{\pi(x')\tilde{\pi}_\theta(x)}{\pi(x)\tilde{\pi}_\theta(x')}\right\}\tilde{\pi}_\theta(x')\ \mu(\mathrm{d}x') + \left(1 - \int_{\mathcal{X}} \min\left\{1, \frac{\pi(w)\tilde{\pi}_\theta(x)}{\pi(x)\tilde{\pi}_\theta(w)}\right\}\tilde{\pi}_\theta(w)\ \mu(\mathrm{d}w)\right)\delta_x(\mathrm{d}x'). \tag{175}$$

Let K' be another transition and consider the transition kernel that is formed by the mixture of K_θ and K': $\hat{K}_\theta(x, A) = \alpha K_\theta(x, A) + (1-\alpha)K'(x, A)$ for $x \in \mathcal{X}$ and $A \in \mathfrak{B}(\mathcal{X})$. Let $(\Theta_0, \Theta_1, \ldots)$ be a sequence of $\mathcal{Y}$-valued random variables. If $(K_{\Theta_0}, K_{\Theta_1}, \ldots)$ exhibits diminishing adaptation then so does $(\hat{K}_{\Theta_0}, \hat{K}_{\Theta_1}, \ldots)$. Furthermore, suppose that with probability at least $1-\delta$ there exists $M \equiv M(\delta)$ such that $K_{\Theta_n}(x, \mathrm{d}x') \geq \frac{\pi(x')}{M}\mu(\mathrm{d}x')$ for every $n = 1, 2, \ldots$. Then $(\hat{K}_{\Theta_0}, \hat{K}_{\Theta_1}, \ldots)$ exhibits containment.

Proof.

$$d(\hat{K}_{\Theta_n}, \hat{K}_{\Theta_{n+1}}) = \sup_{x \in \mathcal{X}} \|\hat{K}_{\Theta_n}(x, \cdot) - \hat{K}_{\Theta_{n+1}}(x, \cdot)\|_{\mathrm{TV}} \tag{176}$$

$$= \sup_{x \in \mathcal{X}} \|\alpha K_{\Theta_n}(x, \cdot) - \alpha K_{\Theta_{n+1}}(x, \cdot)\|_{\mathrm{TV}} \tag{177}$$

$$= \alpha \sup_{x \in \mathcal{X}} \|K_{\Theta_n}(x, \cdot) - K_{\Theta_{n+1}}(x, \cdot)\|_{\mathrm{TV}} \tag{178}$$

$$= \alpha d(K_{\Theta_n}, K_{\Theta_{n+1}}). \tag{179}$$

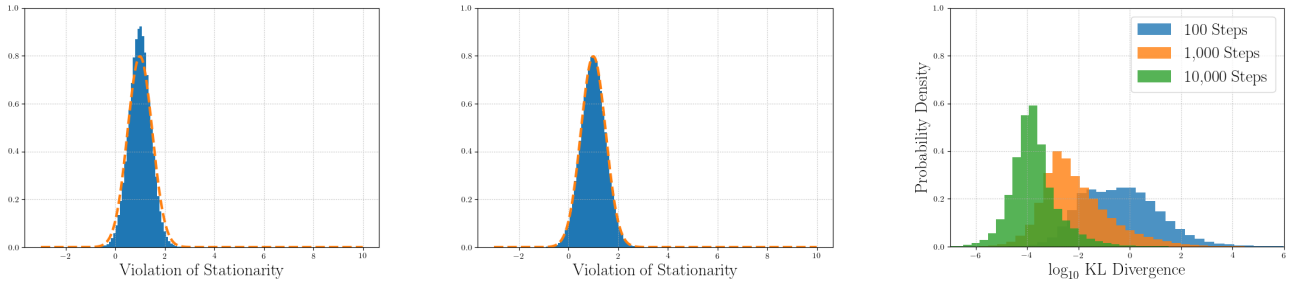
Hence, if $d(K_{\Theta_n}, K_{\Theta_{n+1}})$ converges in probability to zero then so does $d(\hat{K}_{\Theta_n}, \hat{K}_{\Theta_{n+1}})$.

To show containment, observe that with probability at least $1 - \delta$ there exists $M \equiv M(\delta)$ such that,

$$\hat{K}_\theta(x, \mathrm{d}x') \geq \alpha K_\theta(x, \mathrm{d}x') \tag{180}$$

$$\geq \alpha \frac{\pi(x')}{M} \mu(\mathrm{d}x'). \tag{181}$$

Containment then follows from proposition I.1. □



(a) 100 Steps (b) 10,000 Steps (c) Forward KL divergence by Step

Figure 5: Examinations of the violations of stationarity that result by maximizing the pseudo-likelihood of the accepted samples as an adaptation mechanism. After one-hundred steps of adaptation, one clearly perceives that the distribution of states does *not* follow the target distribution. However, after ten-thousand steps, the distribution of state is closer to the target distribution. We also show the KL divergence between the target distribution at the proposal distribution according to the number of steps of the chain. Results are computed over one-million random simulations.

J Violations of Stationarity

In general, adaptation of the parameters of the transition kernel will destroy stationarity. However, if the adaptations and the state of the chain enjoy a prescribed independence condition, then stationarity of the target distribution can be conserved.

Proposition J.1. *Suppose that* $(\Theta_0, \Theta_1, \ldots)$ *is a stochastic* $\mathcal{Y}$*-valued sequence. Let* $(K_{\Theta_n})_{n\in\mathbb{N}}$ *be an associated sequence of Markov transition kernels which produces an* $\mathcal{X}$*-valued chain as* $X_{n+1} \sim K_{\Theta_n}(X_n, \cdot)$*. Suppose further that* Θ_n *and* X_n *are independent given the history of the chain to step* $n-1$*. If* Π *is stationary for each* K_{Θ_n}*, then* Π *is also the stationary distribution of* $(X_n)_{n\in\mathbb{N}}$*.*

We first give an illustration of why maximizing the pseudo-likelihood objective may not always be beneficial. In particular, we look for evidence of violations of stationarity; violations of stationarity mean that if one begins with a sample from the target distribution and transforms it according to several steps of the transition kernel with adaptations, then the final state may not be distributed according to the target distribution. This can be interpreted as an undesirable form of sample degradation wherein applications of an adaptive transition kernel move exact samples further from the target distribution.

As a simple example, we consider sampling Normal$(1, 1/2)$ using a proposal distribution Normal(μ, σ^2); the proposal distribution can be interpreted as a simple normalizing flow consisting of a shift and scale applied to a standard normal base distribution. We consider adapting the parameters of the proposal distribution by computing the maximum likelihood estimates of the mean and standard deviation using the accepted samples. Specifically, let $(X_0, X_1, \ldots, X_n)$ denote the states of the chain to step n; then the parameters of the proposal distribution at step $n+1$ are $\mu_{n+1} = (n+1)^{-1}\sum_{i=0}^{n} X_i$ and $\sigma_{n+1} = \sqrt{(n+1)^{-1}\sum_{i=0}^{n} X_i^2 - \mu_{n+1}^2}$. Results for this adaptation mechanism are shown in fig. 5; after one-hundred steps of adaptation, there is a clear violation of stationarity, but which has nearly vanished by the ten-thousandth step. We observe that adaptations do tend to reduce the forward KL divergence between the target and the proposal distribution, with the closeness improving as the number of adaptation steps increases.

K Independence and Product Transition Kernels

Let Π be a probability measure of $\mathbb{R}^m$ with density π. Let $\hat{\Pi}(A_1, \ldots, A_n) = \prod_{i=1}^n \Pi(A_i)$ be the product probability measure on $(\mathbb{R}^m)^n = \prod_{i=1}^n \mathbb{R}^m$ (the product space of n copies of $\mathbb{R}^m$). Let $(x_1, \ldots, x_n) \in (\mathbb{R}^m)^n$ and define the i^{th} transition kernel by,

$$K_i((x_1, \ldots, x_n), (\mathrm{d}y_1, \ldots, \mathrm{d}y_n)) = [\alpha_i(x_i, y_i)\ \mathrm{d}y_i + \beta_i(x_i)\ \delta_{x_i}(\mathrm{d}y_i)] \cdot \prod_{j=1, j\neq i}^{n} \delta_{x_j}(\mathrm{d}y_j) \tag{182}$$

$$\alpha_i(x, y) = \min\left\{1, \frac{\pi(y)\tilde{\pi}_i(x|y)}{\pi(x)\tilde{\pi}_i(y|x)}\right\} \tilde{\pi}_i(y|x) \tag{183}$$

$$\beta_i(x) = \left(1 - \int_{\mathbb{R}^m} \alpha_i(x, w)\ \mathrm{d}w\right) \tag{184}$$

Thus we can understand K_i as a transition kernel that applies a Metropolis-Hastings accept-reject decision to the i^{th} dimension of the posterior using a proposal density $\tilde{\pi}_i(\cdot|x_i)$ on $\mathbb{R}^m$ and leaves all other dimensions unchanged. Because of the product structure of the joint distribution, these Metropolis-Hastings updates preserve the joint distribution even though the accept-reject decision is computed only on the i^{th} marginal.

As an example, consider K_1 and K_2. What is the composition transition kernel generated by first applying K_1 and subsequently applying K_2? By the Chapman-Kolmogorov formula, it is,

$$(K_2 \circ K_1)((x_1, \ldots, x_n), (A_1, \ldots, A_n)) \tag{185}$$

$$= \int_{\mathbb{R}^m} \cdots \int_{\mathbb{R}^m} K_2((y_1, \ldots, y_n), (A_1, \ldots, A_n)) K_1((x_1, \ldots, x_n), (\mathrm{d}y_1, \ldots, \mathrm{d}y_n)) \tag{186}$$

$$\begin{aligned} &= \int_{\mathbb{R}^m} \cdots \int_{\mathbb{R}^m} \left(\left[\int_{A_2} \alpha_2(y_2, w)\ \mathrm{d}w + \beta_2(y_2)\ \delta_{y_2}(A_2) \right] \cdot \prod_{j=1, j\neq 2}^{n} \delta_{y_j}(A_j) \right) \\ &\quad \times \left([\alpha_1(x_1, y_1)\ \mathrm{d}y_1 + \beta_1(x_1)\ \delta_{x_1}(\mathrm{d}y_1)] \cdot \prod_{j=1, j\neq 1}^{n} \delta_{x_j}(\mathrm{d}y_j) \right) \end{aligned} \tag{187}$$

$$= \left[\int_{A_1} \alpha_1(x_1, w)\ \mathrm{d}w + \beta_1(x_1)\delta_{x_1}(A_1) \right] \cdot \left[\int_{A_2} \alpha_2(x_2, w)\ \mathrm{d}w + \beta_2(x_2)\delta_{x_2}(A_2) \right] \cdot \prod_{j=3}^{n} \delta_{x_j}(A_j). \tag{188}$$

Notice that the composition kernel assumes a factorized form. Suppose that $(x_1', \ldots, x_n') \sim (K_2 \circ K_1)((x_1, \ldots, x_n), \cdot)$. Drawing a sample from this composition kernel can be achieved by setting $x_j' = x_j$ for $j = 3, \ldots, n$ and sampling x_1' and x_2' independently from the distributions

$$\Pr[x_i' \in A | x_i] = \int_A \alpha_i(x_i, w)\ \mathrm{d}w + \beta_i(x_i)\delta_{x_i}(A), \tag{189}$$

for $i \in \{1, 2\}$.

The fact that any composition kernel $K_k \circ \cdots \circ K_1$ has this product distribution form can be established via induction. Assume

$$(K_k \circ \cdots \circ K_1)((x_1, \ldots, x_n), (\mathrm{d}y_1, \ldots, \mathrm{d}y_n)) = \prod_{i=1}^{k} [\alpha_i(x_i, y_i)\ \mathrm{d}y_i + \beta_i(x_i)\ \delta_{x_i}(\mathrm{d}y_i)] \prod_{j=k+1}^{n} \delta_{x_j}(\mathrm{d}y_j) \tag{190}$$

Then,

$$(K_{k+1} \circ \cdots \circ K_1)((x_1, \ldots, x_n), (A_1, \ldots, A_n)) \tag{191}$$

$$= \int_{(\mathbb{R}^m)^n} K_{k+1}((y_1, \ldots, y_n), (A_1, \ldots, A_n))(K_k \circ \cdots \circ K_1)((x_1, \ldots, x_n), (\mathrm{d}y_1, \ldots, \mathrm{d}y_n)) \tag{192}$$

$$\begin{aligned} &= \int_{(\mathbb{R}^m)^n} \left[\left(\int_{A_{k+1}} \alpha_{k+1}(y_{k+1}, w)\ \mathrm{d}w + \beta_{k+1}(y_{k+1}) \delta_{y_{k+1}}(A_{k+1}) \right) \prod_{j=1, j\neq k+1}^{n} \delta_{y_j}(A_j) \right] \\ &\quad \left[\prod_{i=1}^{k} [\alpha_i(x_i, y_i)\ \mathrm{d}y_i + \beta_i(x_i)\ \delta_{x_i}(\mathrm{d}y_i)] \prod_{j=k+1}^{n} \delta_{x_j}(\mathrm{d}y_j) \right] \end{aligned} \tag{193}$$

$$= \left[\prod_{i=1}^{k+1} \left(\int_{A_i} \alpha_i(y_i, w)\ \mathrm{d}w + \beta_i(y_i) \delta_{x_i}(A_i) \right) \right] \left[\prod_{j=k+2}^{n} \delta_{x_j}(A_j) \right]. \tag{194}$$

This verifies that the composition kernel has the desired product structure. The special case of $k = n$ implies,

$$(K_n \circ \cdots \circ K_1)((x_1, \ldots, x_n), (\mathrm{d}y_1, \ldots, \mathrm{d}y_n)) = \prod_{i=1}^{n} [\alpha_i(x_i, y_i)\ \mathrm{d}y_i + \beta_i(x_i)\ \delta_{x_i}(\mathrm{d}y_i)]. \tag{195}$$

To sample $(x_1', \ldots, x_n') \sim (K_n \circ \cdots \circ K_1)((x_1, \ldots, x_n), \cdot)$, one may simply sample independently from the distributions,

$$\Pr[x_i' \in A | x_i] = \int_A \alpha_i(x_i, w)\ \mathrm{d}w + \beta_i(x_i) \delta_{x_i}(A) \tag{196}$$

for $i = 1, \ldots, n$. Each of these samples may be drawn in parallel because the i^{th} sample depends only on x_i.

L Reasons for Violations of Containment

Inexpressive Family

If the family of proposal densities is not sufficiently expressive, it may be that there does not exist any $M \geq 1$ satisfying $\log \pi(x) - \log \tilde{\pi}_\theta(x) < \log M$ for any θ. For instance, on $\mathbb{R}^n$, if the proposal densities have tails that vanish exponentially quickly and the target density's tails diminish only polynomially, then the family of proposal densities is not sufficiently expressive for proposition 4.4 to hold; see Jaini et al. (2020) for details on the tail behavior of normalizing flows. On the other hand, if the family of proposal distributions has a universality property (see, *inter alia* Kobyzev et al. (2020)), then this concern can be alleviated.

Mode Collapse

Proposal densities obtained through the minimization of certain loss functions, including $\mathbb{KL}(\tilde{\pi}_\theta \| \pi)$, may result in the modes of π not being properly represented in the proposal density $\tilde{\pi}_\theta$. In the case of $\mathbb{KL}(\tilde{\pi}_\theta \| \pi)$, the mode-seeking behavior of the loss function can cause the mode-collapse phenomenon, which can invalidate the assumption of proposition 4.4. One could alleviate this concern by targeting a tempered version of the target density or if one has confidence that mode collapse will not occur (for instance if the target density is unimodal).

Unstable Loss

If the adaptations produced by attempting to minimize the loss function are ill-behaved (for instance if the step-size is too large, leading to divergent adaptations), then the sequence $(\Theta_n)_{n\in\mathbb{N}}$ may parameterize poor proposal densities which cause the failure of eq. (6). If the algorithm relies of the convergence of the sequence $(\Theta_n)_{n\in\mathbb{N}}$, then ill-behaved adaptations may lead to a violation of ergodicity due to a failure of both containment and diminishing adaptation.

Inadequate Prior Knowledge

Sampling procedures can be developed so as to incorporate additional knowledge about the posterior, such as symmetries. When such information is available, developing a Bayesian inference procedure that incorporates it can lead to improvements. On the other hand, in complicated distributions, a lack of understanding about the posterior can lead to missed modes, resulting in a failure of containment.

Algorithm 1 Algorithm for sampling from a target distribution by adapting a normalizing flow proposal distribution in the independent Metropolis-Hastings algorithm using the pseudo-likelihood objective.

Input: A sequence of step-sizes $(\epsilon_0, \epsilon_1, \ldots)$; a sequence of adaptation probabilities $(\alpha_0, \alpha, \ldots)$ an initial state x_0 and initial parameter θ_0; a target distribution with density $\pi : \mathcal{X} \to \mathbb{R}$.
for $n = 0, 1, 2, \ldots$ **do**
  Sample a proposal state from the current proposal distribution $\tilde{x}_{n+1} \sim \tilde{\Pi}_{\theta_{n-1}}$.
  Generate $u \sim \text{Uniform}(0, 1)$ and compute the Metropolis-Hastings accept-reject decision.

$$a \leftarrow u < \min\left\{1, \frac{\pi(\tilde{x}_{n+1})\tilde{\pi}_{\theta_n}(x_n)}{\pi(x_n)\tilde{\pi}_{\theta_n}(\tilde{x}_{n+1})}\right\} \tag{197}$$

  if a **then**
    Accept the proposal $x_{n+1} \leftarrow \tilde{x}_{n+1}$.
  else
    Remain at current state $x_{n+1} \leftarrow x_n$.
  end if
  Generate $u' \sim \text{Uniform}(0, 1)$.
  if $u' < \alpha_n$ **then**
    Update the parameters

$$\theta_{n+1} \leftarrow \theta_n + \epsilon_n \nabla \log \tilde{\pi}_{\theta_n}(x_k) \tag{198}$$

    where $k \sim \text{Uniform}(\{0, 1 \ldots, n+1\})$.
  else
    Otherwise, keep the current parameters $\theta_{n+1} \leftarrow \theta_n$.
  end if
end for

M Pseudo-Likelihood Algorithm

The pseudo-likelihood training algorithm is given in algorithm 1.

N Simultaneous Uniform Ergodicity on Compact Spaces

Definition N.1. A family of transition kernels $\{K_\theta : \theta \in \mathcal{Y}\}$ is said to exhibit simultaneous uniform ergodicity if, for all $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $\|K_\theta^n(x, \cdot) - \Pi(\cdot)\|_{\mathrm{TV}} \leq \epsilon$ for all $\theta \in \mathcal{Y}$ and $x \in \mathcal{X}$.

Simultaneous uniform ergodicity is a strong condition which states that, no matter which parameter $\theta \in \mathcal{Y}$ one selects, there is a finite number of steps one can take with that fixed transition kernel in order to become arbitrarily close to the target distribution in total variation. We will see later that this condition can be made to hold for compactly supported target distributions.

We now turn our attention to the question of simultaneous uniform ergodicity.

Proposition N.2. *Let $\mathcal{X}$ and π satisfy the conditions of corollary 2.7. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ whose density $\tilde{\pi}_\theta$ is continuous and satisfies $\mathrm{Supp}(\pi) \subseteq \mathrm{Supp}(\tilde{\pi}_\theta)$. Suppose further that for all $\theta \in \mathcal{Y}$, there exists $\delta > 0$ such that*

$$\delta \leq \min_{x \in \mathrm{Supp}(\pi)} \tilde{\pi}_\theta(x) \tag{199}$$

Then the family of Markov chain transition operators of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}_\theta$ satisfies the simultaneous uniform ergodicity property.

A proof is given in appendix E. Perhaps the most straight-forward mechanism to guarantee eq. (199) is to consider mixture distributions with a fixed distribution that shares the same support as π.

Lemma N.3. *Suppose that Π^* is a distribution on $\mathcal{X}$ with continuous density $\pi^*_{\Pi^*}$ such that $\mathrm{Supp}(\pi) \subseteq \mathrm{Supp}(\pi^*_{\Pi^*})$. Suppose that every $\theta \in \mathcal{Y}$ parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ whose density $\tilde{\pi}_\theta$ is continuous. Consider probability measures $\tilde{\Pi}^*_\theta$ whose densities are constructed from mixtures,*

$$\tilde{\pi}^*_\theta(x) = \beta \pi^*_{\Pi^*}(x) + (1 - \beta)\tilde{\pi}_\theta(x), \tag{200}$$

where $\beta \in (0, 1)$. Then $\tilde{\pi}^*_\theta$ satisfies eq. (199) with

$$\delta = \beta \min_{x \in \mathrm{Supp}(\pi)} \pi^*_{\Pi^*}(x). \tag{201}$$

A proof is given in appendix E. A natural choice of Π^* would be the uniform distribution on $\mathcal{X}$. It is conceivable that one could consider adapting β in the same way that one adapts θ. However, in order to guarantee that eq. (201) is greater than zero, one will require the condition that $\beta \in (\beta_*, 1)$ where $\beta_* > 0$.

Example 6. Let Π be a probability measure with density π on a compact space $\mathcal{X}$. Let $\mathcal{Y} = \mathbb{R}^m$ and suppose that every $\theta \in \mathcal{Y}$ smoothly parameterizes a probability measure $\tilde{\Pi}_\theta$ on $\mathfrak{B}(\mathcal{X})$ with density $\tilde{\pi}_\theta$ for which $\mathrm{Supp}(\pi) = \mathrm{Supp}(\tilde{\pi}_\theta)$. Let $\tilde{\Pi}^*_\theta$ be as in lemma N.3. Let $(\alpha_0, \alpha_1, \ldots)$ be a sequence, bounded between zero and one, converging to zero. Consider the sequence of updates,

$$\theta_n = \begin{cases} \theta_{n-1} - \epsilon \nabla_\theta \log \frac{\tilde{\pi}_{\theta_{n-1}}(\tilde{X}(\theta_{n-1}))}{\pi(\tilde{X}(\theta_{n-1}))} & \text{w.p. } 1 - \alpha_{n-1} \\ \theta_{n-1} & \text{otherwise.} \end{cases} \tag{202}$$

where $\tilde{X} \sim \tilde{\Pi}^*_{\theta_{n-1}}$. Consider the family of Markov chain transition operators of the independent Metropolis-Hastings sampler of Π given $\tilde{\Pi}^*_{\theta_n}$ with transition kernels K_{θ_n} where the proposal at step n is $\tilde{X}$. Then by theorem 2.12 the distribution of $X_{n+1} \sim K_{\theta_n}(X_n, \cdot)$ converges to Π. ‖

Examples of compact spaces on which normalizing flows have been applied include the torus, the sphere, the special orthogonal group, and the Stiefel manifold (Rezende et al., 2020; Falorsi et al., 2019).

The reason we were required to invoke a mixture distribution in the adaptation was because it prevented any sequence from becoming arbitrarily ill-suited to sampling the target distribution; the fact that there was a global limit to how bad any proposal distribution could be allowed us to invoke simultaneous uniform ergodicity of the family of distributions.

O Experimental details on field experiment

We provide additional details on the experiment presented in section 5.3.

Field distribution. The ϕ^4 field model is a popular model used to study phase transition in statistical mechanics (see for example (Berglund et al., 2017)). Here we focus on its $1d$ version, where its field is defined on the segment $[0, 1]$, and impose Dirichlet boundary conditions $\phi(0) = \phi(1) = 0$. The energy function is an intergral over the segment of two terms:

$$U(\phi) = \int_0^1 \left[\frac{a}{2}(\partial_s \phi)^2 + \frac{1}{4a}\left(1 - \phi^2(s)\right)^2\right] \mathrm{d}s. \tag{203}$$

The *coupling term* $\frac{a}{2}(\partial_s \phi)^2$ encourages the smoothness of the field, while the local potential term $\frac{1}{4a}\left(1 - \phi^2(s)\right)^2$ favors fields taking values close to 1 or -1 over the segment. For large values of the parameter a, fields with significant statistical weights will take values close to 0 over the entire segment $[0, 1]$. As a decreases, the system undergoes a *phase transition* and two distinct modes forms concentrating around either $+1$ or -1.

Note that here the energy function (203) is symmetric under the symmetry $\phi \to -\phi$. We exploit this symmetry to provide high-quality reference samples in the experiments described next. Note however that as a biasing term is added to the energy, the statistical weights of either of the modes become unknown.

The numerical experiments described next shows that the adaptive sampler with normalizing flow proposals can recover the relative statistics thanks to efficient mixing, at least at the level of discretization described. Conversely, the energy barrier between the two modes prevents a Langevin sampler from mixing in a reasonable time.

Numerics. We sample the field at 100 equally spaced locations between 0 and 1. The RealNVP flow (Dinh et al., 2017) we optimize has 5 pairs of affine coupling layers updating each half of the 100 field variables. The scaling and translation transformations of each coupling layer is a 2-hidden-layer perceptron with relu activations and 100 units per layer.

The algorithm minimizing the "pseudo-likelihood" objective, defined in example 3, follows largely the lines of algorithm 1. Being more specific, we collect the states of 100 parallel walkers every 10 sampling iterations and take a gradient step with the corresponding 1000-sample batch. The initial learning rate of 10^{-3} is halved every 5000 gradient steps.

We initialize 100 chains: 20 at the uniform value of 1 and 80 at the uniform value of -1. Thanks to the adaptation of the normalizing flow, leading to good acceptance as reported in the main text, these chains can easily mix between modes and recover the proper statistical weights of 50/50.

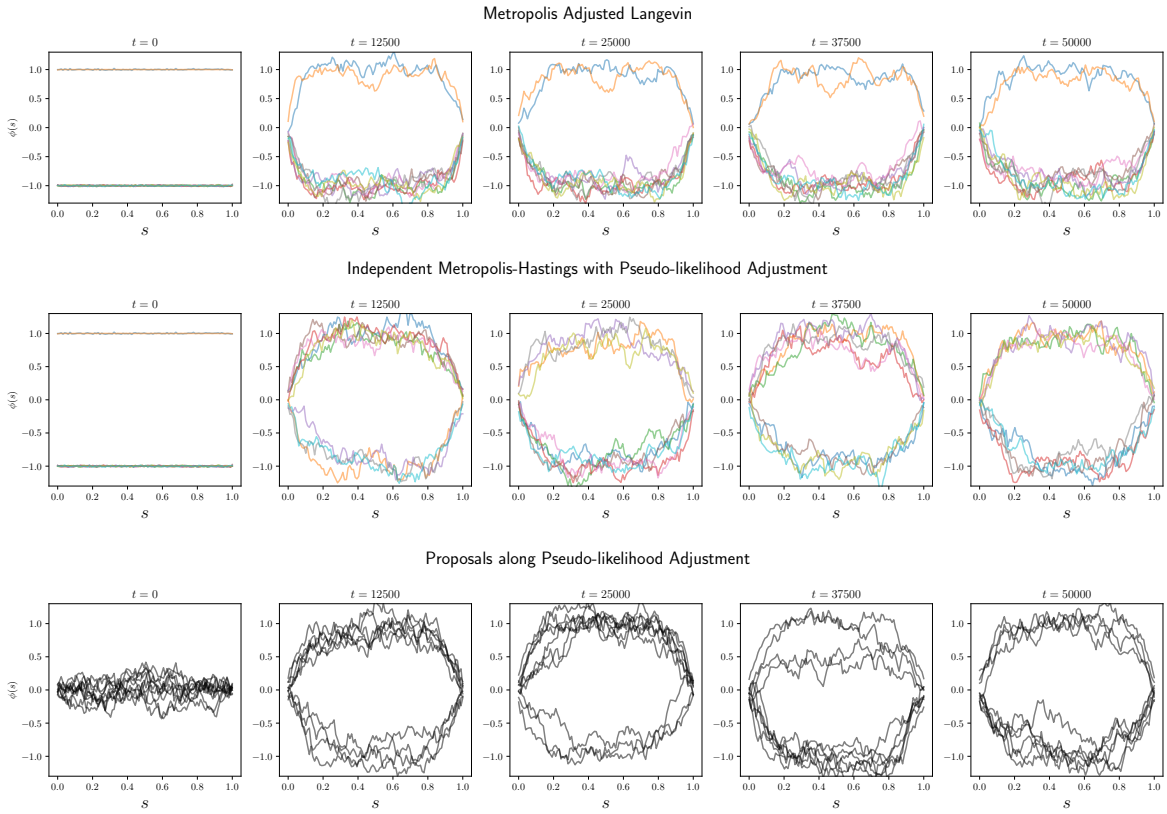


Figure 6: Samples and proposals in the ϕ^4 field experiment.