Adaptive Metropolis-Hastings with Inexact Proposal Density Evaluation

Anonymous Authors[1]

Abstract

In this paper we introduce Approximate Metropolis-Hastings — a modification of the Metropolis-Hastings algorithm that uses an estimate of the proposal density when calculating acceptance probabilities instead of the exact density. This allows using proposals based on generative models with an intractable marginal likelihood, such as variational autoencoders. We provide a theoretical justification of the proposed algorithm using perturbation theory for Markov kernels and demonstrate its advantages using numerical experiments.

1. Introduction

Suppose that we are given a target distribution π on a measurable space $(\mathsf{X}, \mathcal{X})$, and we aim to sample from π or to estimate the integral of some function $f : \mathsf{X} \to \mathbb{R}^d$ with respect to π. In many problems of interest, for example in Bayesian statistics (Mira et al., 2013), π might only be known up to a normalizing constant. In such cases the standard solution is to apply an approach based on Markov Chain Monte Carlo (MCMC) (Andrieu et al., 2003), a family of algorithms which aim to construct a time-homogeneous Markov chain $\{X_k\}_{k\in\mathbb{N}}$, such that the distribution of X_k approaches π in a suitable metrics as k increases. Perhaps the most well-known MCMC method is the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which allows sampling from any target distribution with known unnormalized density. The main idea of the algorithm is to generate candidates from a proposal distribution, and then accept or reject each candidate. There is a vast amount of literature dedicated to different modifications of the Metropolis-Hastings procedure (Tjelmeland, 2004; Liu et al., 2000; Andrieu et al., 2010). The choice of proposal distribution is crucial, as the acceptance rate depends on how similar the proposal and target are. For high dimensional target distributions selecting a good proposal is challenging. More specifically, the acceptance rate tends to approach 0 as the number of dimensions increases. This motivates the development of adaptive modifications of the Metropolis-Hastings algorithm, see (Gabrié et al., 2022; Kobyzev et al., 2021), that choose the proposal distribution from some suitable parametric class. Some papers have experimented with using generative models specifically designed to allow analytic computation of the marginal likelihood, such as normalizing flows (Gabrié et al., 2022; Kobyzev et al., 2021) and Boltzmann generators (Noé et al., 2019), to model the proposal. However, the design constraint of having a tractable marginal likelihood can reduce the expressivity of a model. It is therefore natural to try using more powerful generative models with intractable marginal likelihoods to as proposals. We can leverage these models' greater flexibility; however, this comes at the cost of having to deal with marginal likelihood estimates, which can have high variance and be computationally expensive. In this paper we suggest an approach to adaptive MCMC based on Variational Autoencoders (Kingma & Welling, 2013) and compare its performance with the traditional approach based on generative models with tractable marginal likelihood.

2. Related Works

Parameterizing flexible probabilistic models with neural networks is popular in the adaptive MCMC literature, see (Song et al., 2017; Hoffman et al., 2019; Albergo et al., 2019; Nicoli et al., 2020; Hackett et al., 2021). However, a typical problem of such methods is that increasing the problem dimension causes standard likelihood-based models, such as normalizing flows, to model the target distribution, and especially its tails, with decreasing accuracy (Del Debbio et al., 2021; Grenioux et al., 2023). Some papers (Pompe et al., 2020; Gabrié et al., 2022; Samsonov et al., 2022) suggested mitigating the problem of inaccurate tail behavior by combining local and global proposals. However, the idea of using inexact proposals is not well studied in the modern literature on adaptive MCMC methods. At the same time, there are theoretical works focused on the properties of perturbations of ergodic Markov kernels, starting from the seminal paper (Breyer et al., 2001). Other contributions on the topic include papers (Bardenet et al., 2014; Korattikara et al., 2014; Chen et al., 2022) studying subsampling methods in the context of Bayesian problems. We refer the reader to an excellent recent paper (Rudolf et al., 2024), which contains a much more detailed review of this topic.

Algorithm 1 Approximate Metropolis-Hastings

Input: target density $\pi(x)$, proposal samples $X_1, \dots, X_n$

Train a generative model $\mathcal{M}$ on X;
$\hat{p}_{\mathcal{M}} \leftarrow$ unbiased estimator of marginal likelihood of $\mathcal{M}$
$Y_0 \leftarrow X_0$
for i=1 **to** n **do**
  Draw sample X_i from $\mathcal{M}$
  Compute acceptance rate

$$\alpha(Y_{i-1}, X_i) = \frac{\pi(X_i)\hat{p}_{\mathcal{M}}(Y_{i-1})}{\pi(Y_{i-1})\hat{p}_{\mathcal{M}}(X_i)} \wedge 1$$

  Get next sample

$$Y_i \leftarrow \begin{cases} X_i & \text{with probability } \alpha(Y_{i-1}, X_i) , \\ Y_{i-1} & \text{with probability } 1 - \alpha(Y_{i-1}, X_i) \end{cases}$$

end for

3. Proposed Algorithm

We consider the setting of a target distribution π on a measurable space $(\mathsf{X}, \mathcal{X})$ with $\mathsf{X} \subseteq \mathbb{R}^d$ and π known only up to a normalizing constant. Without loss of generality we use π to denote both the target distribution and its density w.r.t. the Lebesgue measure on $\mathbb{R}^d$. We propose to draw samples approximately from π using the Approximate Metropolis Hastings algorithm, a modification of the standard global proposal Metropolis-Hastings algorithm. The algorithm works by first training a generative model $\mathcal{M}$ on the existing sample from π, then generating a Markov chain using $\mathcal{M}$ to generate candidates and accepting or rejecting each candidate based on the likelihood ratio $\pi(x)/\hat{p}_{\mathcal{M}}(x)$, where $\hat{p}_{\mathcal{M}}$ is an estimator of the model's likelihood. We summarize the procedure in Algorithm 1.

A significant limitation of our approach is that it is only applicable in the case when we both know the unnormalized target density and have a sample from the target distribution to train $\mathcal{M}$ on. However, this setting can arise in practice, for example in the scenario of energy-based models (Nijkamp et al., 2020). A training sample for $\mathcal{M}$ can be obtained by running gradient-based MCMC methods, such as the Unadjusted Langevin algorithm (ULA) (Roberts & Tweedie, 1996). Running large chains of ULA in order to obtain a large amount of samples from the energy-based model can be prohibitively expensive, however obtaining a small high quality training sample may be possible.

4. Theoretical justification

The approach suggested in Algorithm 1 can be justified using existing results on perturbed Markov kernels. In the exposition below we closely follow (Rudolf et al., 2024). For two probability measures ξ and ξ' on $(\mathsf{X}, \mathcal{X})$, we say that a probability measure ν on $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ is a coupling of ξ and ξ' if for each $\mathsf{A} \in \mathcal{X}$, $\nu(\mathsf{A} \times \mathsf{X}) = \xi(\mathsf{A})$ and $\nu(\mathsf{X} \times \mathsf{A}) = \xi'(\mathsf{A})$. Denote by $\Pi(\xi, \xi')$ the set of couplings of ξ and ξ' on $(\mathsf{X}, \mathcal{X})$. Then the Kantorovich-Wasserstein semimetric $\mathbf{W}_{\mathsf{d}}(\xi, \xi')$, associated with the metric $\mathsf{d}(x, x')$, is defined as

$$\mathbf{W}_{\mathsf{d}}(\xi, \xi') = \inf_{\nu \in \Pi(\xi, \xi')} \int_{\mathsf{X} \times \mathsf{X}} \mathsf{d}(x, x') \nu(\mathrm{d}x \mathrm{d}x') . \quad (1)$$

For example, we can choose $\mathsf{d}(x, x') = \mathbb{1}_{x \neq x'}$ and obtain the total variation distance between ξ and ξ'. In order to justify Algorithm 1 we state the result on closeness of invariant distributions of Markov kernels P and $\hat{\mathrm{P}}$, provided that $\mathrm{P}(x, \cdot)$ and $\hat{\mathrm{P}}(x, \cdot)$ are close for any $x \in \mathsf{X}$. More precisely, we use the following assumptions:

A 1. Markov kernel $\hat{\mathrm{P}}$ admits a unique invariant distribution $\hat{\pi}$, moreover, there exists $\varepsilon > 0$, such that $\sup_{x \in \mathsf{X}} \mathbf{W}_{\mathsf{d}}\left(\mathrm{P}(x, \cdot), \hat{\mathrm{P}}(x, \cdot)\right) \leq \varepsilon$.

We will show that A 1 is satisfied for the Markov kernel of Metropolis-Hastings algorithm, if the density estimate $\hat{p}_{\mathcal{M}}$ is close enough to π. The second assumption is related to the kernel P itself:

A 2. Markov kernel P admits π as invariant distribution and is $\mathbf{W}_{\mathsf{d}}(\cdot, \cdot)$-geometrically ergodic, that is, there exists $0 < \Delta < 1$, such that for any $x, x' \in \mathsf{X}$ it holds that

$$\mathbf{W}_{\mathsf{d}}(\xi, \xi') \leq \Delta \mathsf{d}(x, x') .$$

Under the above assumption we can state the following result from (Rudolf et al., 2024):

Theorem 4.1. *Assume A 1 and A 2. Then for invariant distributions π and $\hat{\pi}$ it holds that*

$$\mathbf{W}_{\mathsf{d}}(\pi, \hat{\pi}) \leq \varepsilon / (1 - \Delta) \quad (2)$$

Proof of Theorem 4.1 can be found in Theorem 19.2.1 in (Rudolf et al., 2024). This results formalizes an expected fact that the closeness in Markov kernels implies, under appropriate assumptions, closeness of their invariant distributions. Now we provide the following counterpart for Theorem 4.1 under additional assumptions on π and $\hat{p}_{\mathcal{M}}$.

A 3. Suppose that $\mathsf{X} \subseteq [0, 1]^d$, and both π and $p_{\mathcal{M}}$ are bounded away from 0 on $[0, 1]^d$ and bounded, that is, there exist $\beta > 0$, such that $\beta \leq \pi(x) \leq 1/\beta$, and $\beta \leq p_{\mathcal{M}}(x) \leq 1/\beta$. Moreover, $\|\hat{p}_{\mathcal{M}} - p_{\mathcal{M}}\|_\infty \leq \epsilon$ for some $\epsilon > 0$.

Let us denote as Q the Markov kernel of the Metropolis-Hastings algorithm with exactly calculated proposal $p_{\mathcal{M}}$ and by $\hat{\mathrm{Q}}$ its counterpart, corresponding to Algorithm 1.

Theorem 4.2. *Assume A 3. Then assumptions A 1 and A 2 are satisfied, hence, the bound (2) holds.*

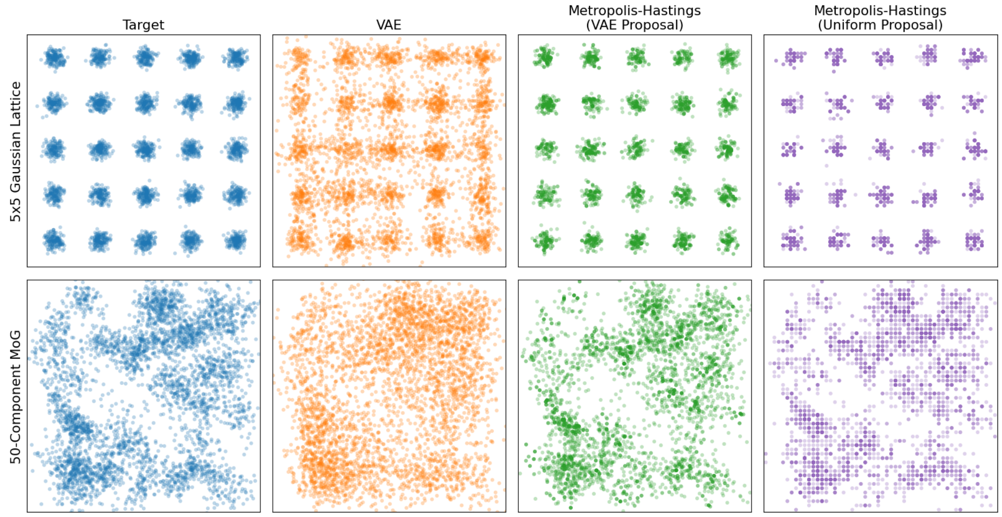

Figure 1. Performance of different sampling algorithms on 2D synthetic targets, 4000 samples

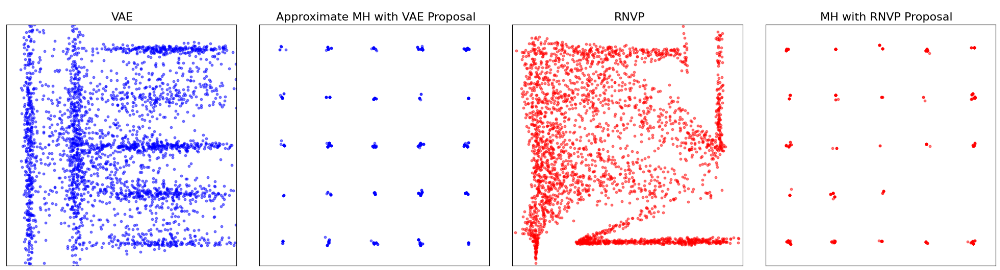

Figure 2. Behavior of VAE and RNVP on a $2d$ mixture of Gaussians. Mode dropping can be observed for the RNVP proposal.

Proof of Theorem 4.2 is given in Appendix A. Uniform geometric ergodicity of the Metropolis-Hastings sampler under assumption A3 follows from classical results in the literature, see e.g. (Johnson et al., 2013). Assumptions of Theorem 4.2 are of course restrictive and can be further relaxed. Obtaining variants of Theorem 4.2 under assumption more realistic than A3 is an interesting research direction for future work.

5. Particular Instance: VAE

Variational Autoencoders (VAE) (Kingma & Welling, 2013) are a natural choice of proposal model since they tend to generate out-of-distribution samples and are relatively resistant to mode collapse (Xiao et al., 2021). VAEs are latent-variable models that are parameterized by two neural networks. The prior $p(z)$ is non-parametric, one network (the encoder) defines the conditional distribution $p_\theta(x|z)$, another (the decoder) defines the posterior approximation $q_\phi(z|x)$ which tries to match the real posterior $p_\theta(z|x)$. Here, x and z denote the observed and latent variables respectively. The approximate posterior can be leveraged to derive effective marginal likelihood (ML) estimators. This is due to the fact that knowing the posterior $p_\theta(z|x)$ means knowing the ML, as $p_\theta(x) = p_\theta(x,z)/p_\theta(z|x)$. Development of unbiased ML estimates, including MCMC-based ones (Salimans et al., 2015), has been motivated by the fact that they can be used to construct ELBOs using Jensen's inequality(Mnih & Rezende, 2016) — optimization objectives of VAEs. In this work we focus on the L-sample importance weighted estimate

$$\hat{p}_L(x) = \tfrac{1}{L}\sum_{i=1}^{L} \tfrac{p_\theta(x,Z_i)}{q_\phi(Z_i|x)} \;, \tag{3}$$

where $Z_1, \ldots, Z_L \sim q_\phi(\cdot|x)$ are sampled independently from the decoder.

6. Experiments

2D Multimodal Distributions. We visualize our algorithm for a 2D mixture-of-Gaussians target (see Figure 1). The samples generated by the VAE (second from the left) cover all the high-density regions of the target (leftmost). However, the proposal is blurry (the model produces a lot of artifacts). Approximate Metropolis-Hastings samples (second from the right) are less blurry and visibly more similar to the target. This example shows that applying our algorithm to VAEs is a reasonable idea, because they can produce artifacts even for relatively simple low-dimensional targets.

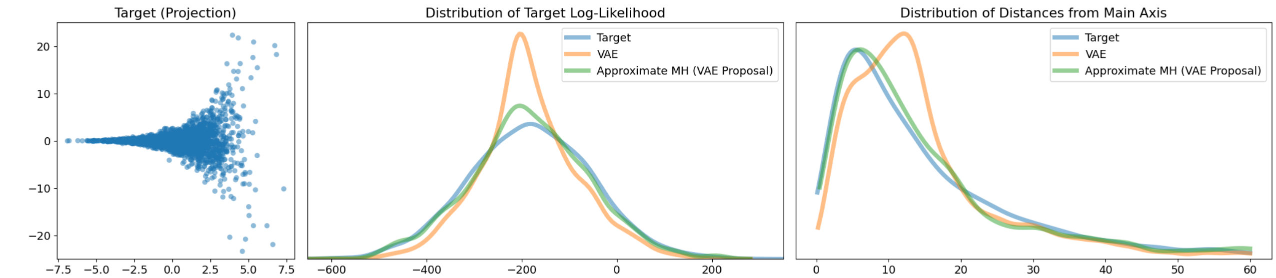

Figure 3. Approximate Metropolis-Hastings improves feature distributions for a 128-dimensional Funnel

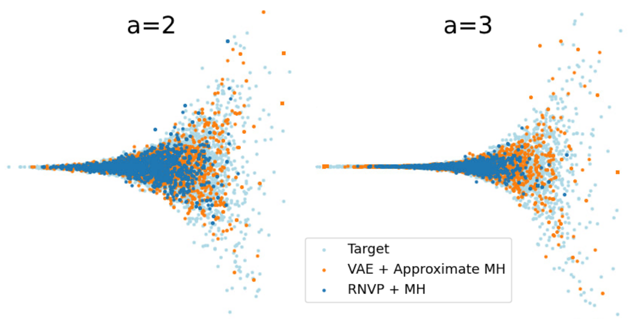

Figure 4. Comparison of Approximate Metropolis-Hastings and Exact Metropolis-Hastings on a 128-dimensional Funnel for different values of a

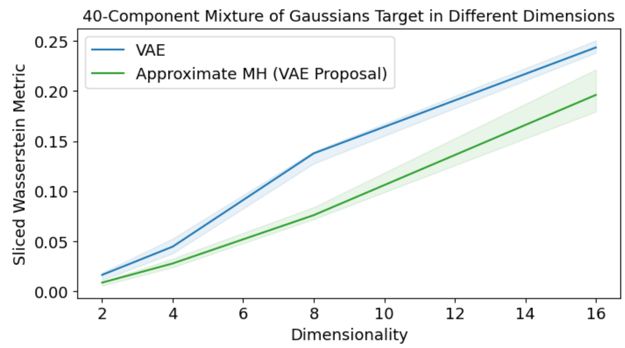

Figure 5. Scaling of the Approximate Metropolis-Hastings algorithm

The performance of non-adaptive Metropolis-Hastings with a uniform proposal (rightmost) is worse, but not that bad, because this is a 2D example. However, it does not scale to higher dimensions.

We explored the possible advantages of using a VAE proposal instead of a RealNVP (Dinh et al., 2016) proposal. RealNVP is a type of normalizing flow. They allow straightforward marginal likelihood calculation but, as we show, are not as flexible as VAEs. In Figure 2 Approximate Metropolis-Hastings with a VAE proposal and Metropolis-Hastings with a RNVP proposal are compared. Both proposals have approximately the same complexity and were trained for the same number of epochs. The RNVP proposal is uniform in most areas and this leads to mode loss. The quality of VAE proposal is more consistent across modes. This highlights the need for using more expressive models for proposals, since mode loss cannot be corrected by MH.

Distributions with complex geometry. We tested our algorithm on a 128-dimensional Neal's Funnel (see Figure 3). The funnel is a multidimensional distribution with the first coordinate x_1 distributed as $\mathcal{N}(0, a^2)$, and the rest i.i.d. as $\mathcal{N}(0, e^{x_1})$, where a is a parameter. We found that the trained VAE's support mostly covers the target support, but the VAE cannot fully learn the target's shape. The Approximate Metropolis-Hastings brings the proposal distribution closer to the target, which can be seen by plotting the distributions of different sample features, such as target log-likelihood (middle) or distance from the main axis (right). This once again shows the usefulness of Approximate MH. We also compared our algorithm with RNVP-proposal classic MH on funnels with different values of a. As can be seen Figure 4, the VAE is better at modeling the tails of the target than RNVP, and Approximate MH samples are a better representation of the target than MH samples.

Scaling with dimensionality. We looked at how our algorithm scales with the number of dimensions in the case of a mixture-of-gaussians target (Figure 5). We measure sample quality using the sliced Wasserstein Metric between samples and the target. We see that applying Approximate Metropolis-Hastings reliably improves sample quality even as the problem dimension increases. Further experiment details can be found in Appendix C.

7. Further Work and Conclusion

In this work we have shown that Approximate Metropolis-Hastings can outperform adaptive Metropolis-Hastings algorithms based on normalizing flows, and also improve sample quality of vanilla Variational Autoencoders. One direction for further research is using different VAE architectures with more expressive posterior estimates, such as inverse autoregressive flows(Kingma et al., 2016). Better posterior estimates could improve the quality of generated samples, since it would make ML estimates more reliable. It would be natural to try using ML estimates other than the importance weighted estimate for the VAE. On the other hand, Approximate Metropolis-Hastings could also be applied to models other than VAE. Another possible extension of this work is a fully adaptive version of Approximate Metropolis-Hastings, where the proposal model is fine-tuned on generated samples while the algorithm is running.

References

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

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.

Andrieu, C., Doucet, A., and Holenstein, R. Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3): 269–342, 2010.

Bardenet, R., Doucet, A., and Holmes, C. Towards scaling up markov chain monte carlo: an adaptive subsampling approach. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 405–413, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/bardenet14.html.

Breyer, L., Roberts, G. O., and Rosenthal, J. S. A note on geometric ergodicity and floating-point roundoff error. *Statistics & probability letters*, 53(2):123–127, 2001.

Chen, N., Xu, Z., and Campbell, T. Bayesian inference via sparse hamiltonian flows. *Advances in Neural Information Processing Systems*, 35:20876–20888, 2022.

Del Debbio, L., Marsh Rossney, J., and Wilson, M. Efficient modeling of trivializing maps for lattice $\phi 4$ theory using normalizing flows: A first look at scalability. *Physical Review D*, 104(9), 2021. ISSN 24700029. doi: 10.1103/PhysRevD.104.094507. URL http://arxiv.org/abs/2105.12481.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018. ISBN 978-3-319-97703-4; 978-3-319-97704-1. doi: 10.1007/978-3-319-97704-1. URL https://doi.org/10.1007/978-3-319-97704-1.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Gabrié, M., Rotskoff, G. M., and Vanden-Eijnden, E. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences*, 119(10): e2109420119, 2022.

Grenioux, L., Oliviero Durmus, A., Moulines, E., and Gabrié, M. On sampling with approximate transport maps. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11698–11733. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/grenioux23a.html.

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. Flow-based sampling for multimodal distributions in lattice field theory. *arXiv preprint*, 2107.00734, 2021. URL http://arxiv.org/abs/2107.00734.

Hastings, W. K. Monte carlo sampling methods using markov chains and their applications. 1970.

Hoffman, M. D., Sountsov, P., Dillon, J. V., Langmore, I., Tran, D., and Vasudevan, S. NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport. In *1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–5*, 2019. URL http://arxiv.org/abs/1903.03704.

Johnson, A. A., Jones, G. L., and Neath, R. C. Component-Wise Markov Chain Monte Carlo: Uniform and Geometric Ergodicity under Mixing and Composition. *Statistical Science*, 28(3):360 – 375, 2013. doi: 10.1214/13-STS423. URL https://doi.org/10.1214/13-STS423.

Kingma, D. and Welling, M. Auto-encoding variational bayes. *ICLR*, 12 2013.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4743–4751, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Kobyzev, I., Prince, S. J., and Brubaker, M. A. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, November 2021. ISSN 1939-3539.

Korattikara, A., Chen, Y., and Welling, M. Austerity in mcmc land: Cutting the metropolis-hastings budget. In

Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 181–189, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/korattikara14.html.

Liu, J. S., Liang, F., and Wong, W. H. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449): 121–134, 2000.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Mira, A., Solgi, R., and Imparato, D. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23:653–662, 2013.

Mnih, A. and Rezende, D. Variational inference for monte carlo objectives. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2188–2196, New York, New York, USA, 20–22 Jun 2016. PMLR.

Nicoli, K. A., Nakajima, S., Strodthoff, N., Samek, W., Müller, K. R., and Kessel, P. Asymptotically unbiased estimation of physical observables with neural samplers. *Physical Review E*, 101(2), 2020. ISSN 24700053. doi: 10.1103/PhysRevE.101.023304.

Nijkamp, E., Hill, M., Han, T., Zhu, S.-C., and Wu, Y. N. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5272–5280, 2020.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Pompe, E., Holmes, C., and Łatuszyński, K. A framework for adaptive MCMC targeting multimodal distributions. *Annals of Statistics*, 48(5):2930–2952, 2020. ISSN 21688966. doi: 10.1214/19-AOS1916. URL https://arxiv.org/abs/1812.02609.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Rudolf, D., Smith, A., and Quiroz, M. Perturbations of Markov Chains. *arXiv preprint arXiv:2404.10251*, 2024.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1218–1226, Lille, France, 07–09 Jul 2015. PMLR.

Samsonov, S., Lagutin, E., Gabrié, M., Durmus, A., Naumov, A., and Moulines, E. Local-Global MCMC kernels: the best of both worlds. *Advances in Neural Information Processing Systems*, 35:5178–5193, 2022.

Song, J., Zhao, S., and Ermon, S. A-NICE-MC: Adversarial training for MCMC. In *Advances in Neural Information Processing Systems*, pp. 5140–5150, 2017.

Tjelmeland, H. Using all Metropolis–Hastings proposals to estimate mean values. Technical report, 2004.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

A. Proof of Theorem 4.2

Note that for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ the Markov kernels Q of MH algorithm and $\hat{\mathrm{Q}}$ of its perturbed version are defined, respectively, as

$$\begin{aligned}\mathrm{Q}(x,\mathsf{A}) &= \int_{y\in\mathsf{A}} \left(\frac{\pi(y)p_{\mathcal{M}}(x)}{\pi(x)p_{\mathcal{M}}(y)} \wedge 1\right) p_{\mathcal{M}}(y)\mathrm{d}y + \mathbb{1}_{x\in\mathsf{A}} \int_{y\in\mathsf{X}} \left(1 - \left(\frac{\pi(y)p_{\mathcal{M}}(x)}{\pi(x)p_{\mathcal{M}}(y)} \wedge 1\right)\right) p_{\mathcal{M}}(y)\mathrm{d}y \\ \hat{\mathrm{Q}}(x,\mathsf{A}) &= \int_{y\in\mathsf{A}} \left(\frac{\pi(y)\hat{p}_{\mathcal{M}}(x)}{\pi(x)\hat{p}_{\mathcal{M}}(y)} \wedge 1\right) p_{\mathcal{M}}(y)\mathrm{d}y + \mathbb{1}_{x\in\mathsf{A}} \int_{y\in\mathsf{X}} \left(1 - \left(\frac{\pi(y)\hat{p}_{\mathcal{M}}(x)}{\pi(x)\hat{p}_{\mathcal{M}}(y)} \wedge 1\right)\right) p_{\mathcal{M}}(y)\mathrm{d}y \;.\end{aligned} \tag{4}$$

Since the function $x \wedge 1$ is 1-Lipschitz, we get from the previous formula with the simple algebra that

$$\left|\mathrm{Q}(x,\mathsf{A}) - \hat{\mathrm{Q}}(x,\mathsf{A})\right| \leq 2\epsilon/\beta^5 \;.$$

Moreover, we have

$$\mathrm{Q}(x,\mathsf{A}) \geq (1/\beta^4)\nu(\mathsf{A}) \;,$$

where we have defined $\nu(\mathsf{A}) = \int_{y\in\mathsf{A}} p_{\mathcal{M}}(y)\mathrm{d}y$. Hence, the whole space X is $(1, 1/\beta^4)$-small (see (Douc et al., 2018), Chapter 9) in case of Q and $(1, 1/(3\beta^4))$-small in case of $\hat{\mathrm{Q}}$. This means that both kernels are uniformly geometrically ergodic, that is, they satisfy (2) with $\mathsf{d}(x, x') = \mathbb{1}_{x\neq x'}$ and $\Delta = 1 - 1/\beta^4$. Hence, both Q and $\hat{\mathrm{Q}}$ admit unique invariant distributions. Thus, A1 and A2 hold.

B. Metrics

The Wasserstein metric between two samples $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\mathbf{W}_p\left(\mathbf{x}, \mathbf{y}\right) = \min_{\gamma\in\mathbb{R}_+^{m\times n}} \left(\sum_{i,j} \gamma_{ij} ||x_i - y_j||_p\right)^{\frac{1}{p}}$$

with the minimum being taken over positive-valued matrices γ whose rows and columns all sum to 1. In this paper we report $\mathbf{W}_2\left(\cdot,\cdot\right)$. A sliced metric between 2 samples is calculated as the mean value of a 1D metric between random projections of those samples. To compute the Wasserstein metric in 1D we use Python Optimal Transport (Flamary et al., 2021).

C. Experiment Details

For all experiments with Approximate Metropolis-Hastings we use a Variational Autoencoders as the proposal and a 512-sample Importance weighted marginal likelihood estimate. The encoder and decoder are symmetric and both consist of fully-connected layers with batch normalization layers in-between. All proposals are trained on 2^{14} samples from the target distribution.

C.1. 2D Multimodal Distributions

Figure 1. Reported confidence intervals are for the MoG target in the second row of Figure 1. The intervals boundaries are the 25th and 75th quantiles of the sliced Wasserstein metric over 10 runs of 5000 samples each.

C.2. Distributions with Complex Geometry

The density of Neal's funnel with parameter a is

$$p_{\text{funnel}}(x) = Z^{-1} \exp\left(-\frac{x_1^2}{2a^2} - \frac{1}{2}e^{-x_1}\sum_{i=2}^{d}\left[x_i^2 + x_1\right]\right), \quad d \geq 2,$$

where Z is the normalizing constant.

The VAE used in Figure 3 and has 3 hidden layers of 128 neurons in both the encoder and decoder. The density plots are based on 5000 samples each.

In Figure 4 the VAE used is the same as in Figure 3. The flow proposal is 7 RealNVP layers. The flow proposal has approximately twice as many parameters as the VAE proposal, yet still performs worse. Modifying the training time and amount of layers did not significantly help RNVP. In general we found that training RNVPs is less reliable than training VAEs, as they are prone to exploding gradients and mode collapse.