# Адаптивные модификации алгоритма Метрополиса-Гастингса

Юрий Свирщевский
Научный руководитель: Сергей Самсонов

HDI Lab

Высшая школа экономики

NATIONAL RESEARCH
UNIVERSITY

28 Мая 2024

# Постановка задачи

- Есть некоторое сложное целевое распределение $\pi$ из которого мы хотим научиться семплировать.
- Известна ненормированная плотность этого распределения $\pi(x)$, также у нас уже есть небольшая выборка из $\pi$.
- Решение этой задачи позволяет, например, получать численные приближения многомерных интегралов.

# Монте-Карло с марковскими цепями (MCMC)

- Стандартный подход к семплированию из сложных распределений
- Получаемая выборка автокоррелирована, семплы не независимые
- Примеры: Алгоритм Метрополиса-Гастингса, семплирование по Гиббсу, Гамильтоново Монте-Карло

# Алгоритм Метрополиса-Гастингса

- Алгоритм Метрополиса-Гастингса позволяет семплировать из сложного целевого распределения с известной плотностью $\pi(x)$ путём сэмплирования из более простого предлагающего распределения $\eta(x)$ и принятием или отвержением каждого сэмпла.
- В глобальной версии марковская цепь $\{X_n\}$ строится по правилу по правилу

$$X_{n+1} = \begin{cases} X', & \text{с вероятностью } \alpha(X_n, X') \\ X_n, & \text{с вероятностью } 1 - \alpha(X_n, X') \end{cases}$$

  где $\alpha(X, X') = \frac{\pi(X')\eta(X)}{\pi(X)\eta(X')}$ — вероятность принятия, $X' \sim \eta(X)$.
- Можно показать, что при достаточно мягких условиях предельным распределением этой марковской цепи будет $\pi(x)$.
- От выбора предлагающего распределения сильно зависит качество алгоритма. Если оно недостаточно похоже на целевое, доля принятых семплов может быть очень близка к нулю.

# Использование генеративных моделей в М-Г

- Предлагающее распределение $\eta(x)$ хотелось бы моделировать с помощью современных глубоких генеративных моделей.
- Главная сложность заключается в том, что для подсчёта $\alpha(X, X')$ необходимо знать маргинальное правдоподобие (МП) модели.
- Есть классы т.н. явных генеративных моделей, специально спроектированных таким образом, что можно легко вычислять МП. Такими моделями являются, например, нормализующие потоки.
- Большинство же моделей, основанных на нейронных сетях, неявные.

# Идея исследования

- Обучить мощную генеративную модель, не позволяющую точного вычисления МП, на небольшой выборке из $\pi(x)$.
- Использовать её для моделирования предлагающего распределения в алгоритме Метрополиса-Гастингса.
- Использовать $\underline{\text{оценку маргинального правдоподобия}}$ $\hat{\eta}$ для вычилсения вероятности принятия $\hat{\alpha}(X, X') = \frac{\pi(X')\hat{\eta}(X)}{\pi(X)\hat{\eta}(X')}$.
- Сравнить с точным алгоритмом Метрополиса-Гастингса

# Вариационные автокодировщики

- Генеративная модель с явными ($x$) и латентными ($z$) переменными, параметризованная двумя нейросетями.
- Априорное распределение $p(z)$ фиксированная. Одна нейросеть (декодер) задаёт условное распределение $p(x|z)$. Другая (энкодер) задаёт $q_\varphi(z|x)$ — приближение настоящего апостериорного распределения $p(z|x)$.
- Маргинальное правоподобие $p(x) = \int_z p(x|z)p(z)dz$ нельзя посчитать точно. Интеграл не берётся. Более того, наивная Монте-Карло аппроксимация работает очень плохо.
- Но, можно использовать Importance Weighted оценку, основанную на семплах из энкодера

$$\hat{p}_L(x) = \frac{1}{L} \sum_{i=1}^{L} \frac{p(x, Z_i)}{q_\varphi(Z_i|x)},$$

где $Z_1, \ldots, Z_L \sim q_\varphi(\cdot|x)$ генерируются декодером независимо.

# Эксперименты I

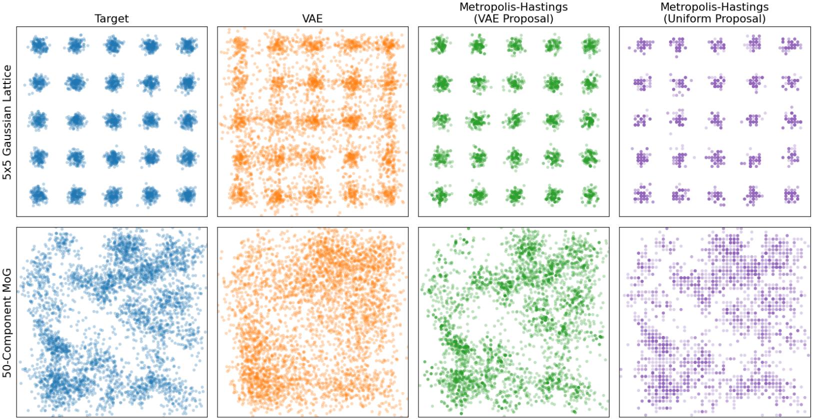


Figure: 2D синтетические примеры. 4000 точек

Доверительные интервалы для усреднённой метрики Вассерштейна по проекциям (расстояние до целевого распределения): VAE — [1.2e-2; 1.6e-2], Неточный М-Г — [1.0e-2; 1.3e-2], обычный М-Г —[1.5e-2; 2.0e-2].

# Эксперименты II

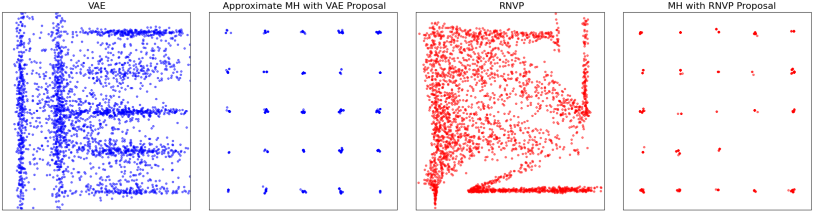


Figure: Сравнение VAE и нормализующего потока RNVP

# Эксперименты III

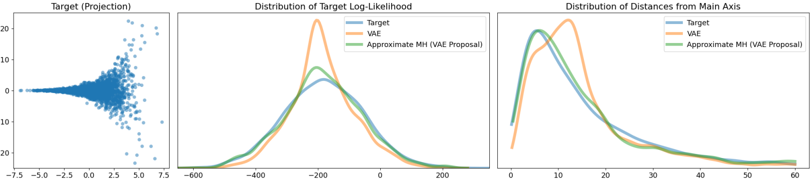


Figure: Наш алгоритм улучшает распределение признаков на 128-мерной воронке

# Эксперименты IV

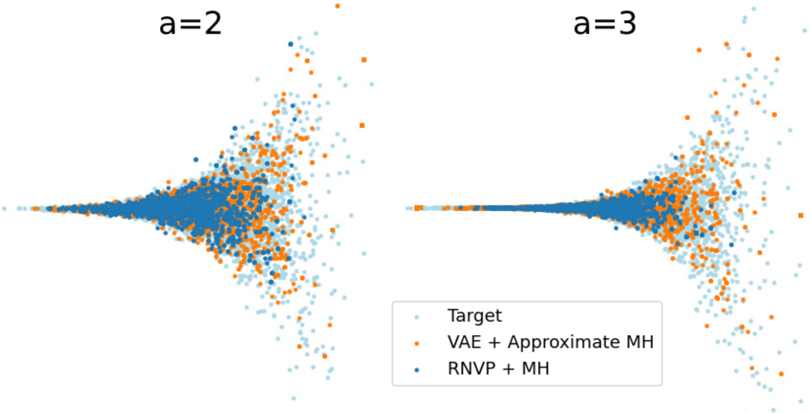


Figure: Наш алгоритм работает лучше чем существующие на 128-мерной воронке

# Эксперименты V

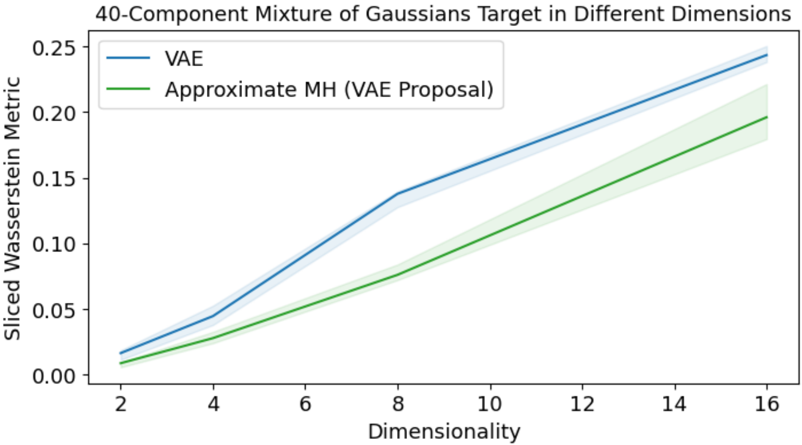
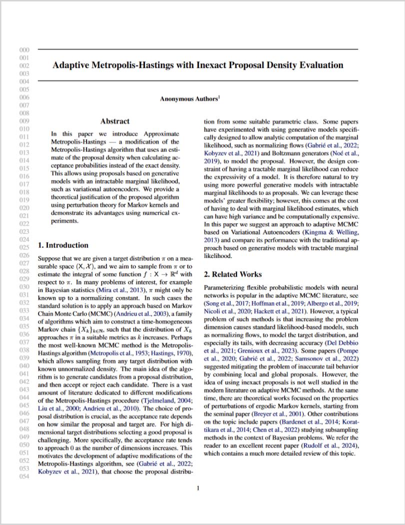


Figure: Наш алгоритм стабильно улучшает качество выборки

# Публикация

## Статья отправлена на воркшоп SPIGM конференции ICML 2024



# Adaptive Metropolis-Hastings with Inexact Proposal Density Evaluation

**Anonymous Authors**[1]

## Abstract

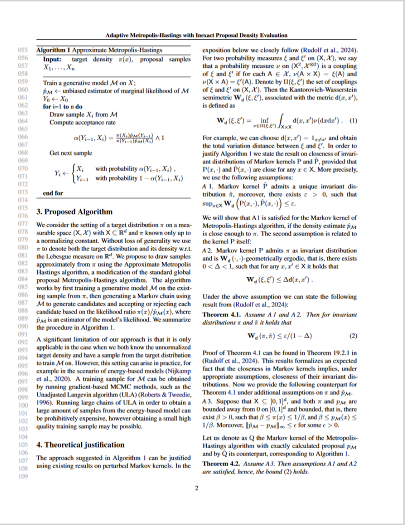
In this paper we introduce Approximate Metropolis-Hastings — a modification of the Metropolis-Hastings algorithm that uses an estimate of the proposal density when calculating acceptance probabilities instead of the exact density. This allows using proposals based on generative models with an intractable 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 $\pi$ on a measurable space $(\mathsf{X}, \mathcal{X})$, and we aim to sample from $\pi$ or to estimate the integral of some function $f : \mathsf{X} \to \mathbb{R}^d$ with respect to $\pi$. In many problems of interest, for example in Bayesian statistics (Mira et al., 2013), $\pi$ 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 $\pi$ 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 $\pi$ on a measurable space $(\mathsf{X}, \mathcal{X})$ with $\mathsf{X} \subseteq \mathbb{R}^d$ and $\pi$ known only up to a normalizing constant. Without loss of generality we use $\pi$ 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 $\pi$ 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 $\pi$, 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 $\xi$ and $\xi'$ on $(\mathsf{X}, \mathcal{X})$, we say that a probability measure $\nu$ on $(\mathsf{X}^2, \mathcal{X}^{\otimes 2})$ is a coupling of $\xi$ and $\xi'$ 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 $\xi$ and $\xi'$ 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 $\xi$ and $\xi'$. In order to justify Algorithm 1 we state the result on closeness of invariant distributions of Markov kernels $\mathrm{P}$ 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 $\pi$. The second assumption is related to the kernel $\mathrm{P}$ itself:

*A* 2. Markov kernel $\mathrm{P}$ admits $\pi$ 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 $\pi$ 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 $\pi$ and $\hat{p}_{\mathcal{M}}$.

*A* 3. Suppose that $\mathsf{X} \subseteq [0, 1]^d$, and both $\pi$ 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 $\mathrm{Q}$ 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.*

Вопросы?

Спасибо за внимание!