Ferdinand Genans Boiteux
Project for the course "Simulation Methods for Generative Models"

Contents

1 Introduction

The authors of the paper "Transport map accelerated Markov chain Monte Carlo" [1] propose a new framework that combines transport of measure maps with the Metropolis-Hastings rule to generate non-Gaussian proposal distributions that can more effectively explore the target density.
Their method constructs a lower triangular transport map, which approximates the Knothe-Rosenblatt rearrangement, using information from previous MCMC states. This is achieved through the solution of a convex and separable optimization problem, which enables efficient and parallelizable adaptation of the map even for large numbers of samples.
The authors demonstrate that their approach produces an adaptive MCMC algorithm that is ergodic for the exact target distribution, even when using inexact or truncated maps. Their numerical experiments show that their method can produce order-of-magnitude speedups over standard MCMC techniques.

Our study conducted an examination of the main algorithm for efficient sampling from complex probability distributions, the underlying optimization problem that guides its design, and reformulations aimed at enhancing its computational efficiency, which also included a small exploration of transport theory.

2 Transport map

Since the paper we studied heavily depends on the concept of transport maps, it is crucial to define this notion and outline the necessary properties required for the transport map MCMC algorithm.

Transport map: Given two probability measures, μ and ν, defined on $\mathbb{R}^d$. A Borel map, $T : \mathbb{R}^d \to \mathbb{R}^d$, is considered a transport map between μ and ν if it satisfies the equality $\mu(T^{-1}(B)) = \nu(B)$ for every Borel set B. This is denoted as $T_{\#}\mu = \nu$, representing the pushforward of μ. The set of all transport maps between μ and ν is denoted as $\mathcal{T}(\mu, \nu)$.

Example: Lets consider $\mu, \nu \in \mathbb{P}(\mathbb{R})$ with: $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$.
We have: $\mathcal{T}(\mu, \nu) = \emptyset$ but $\mathcal{T}(\nu, \mu) = \{T : \text{Borel map}; T(1) = T(2) = 0\}$.
With this example, we see that $\mathcal{T}(\mu, \nu)$ can be empty or can contain infitintely many Borel maps. We also see that in general: $\mathcal{T}(\mu, \nu) \neq \mathcal{T}(\nu, \mu)$.

Main idea: we want a reference distribution μ that is easy to sample from and seek a transport map T such that $T_{\#}\mu = \nu$.
Indeed, if a transport map T was known, we would just have to sample from the source distribution and push those samples with T. Finding an exact transport map can be challenging. The paper we studies tries to overcome this difficulty by looking for approximate transport maps.

Notation: For the study of our article, we focus on $\mu, \nu \in \mathbb{P}(R^d)$ two absolutely continuous probabilities wrt the Lebesgue measure. We note there densities $\rho : \mathbb{R}^d \to \mathbb{R}^+$ resp. $\pi : \mathbb{R}^d \to \mathbb{R}^+$.

Remark: given $T \in \mathcal{T}(\mu, \nu)$ by the change-of-variables formula, we have:

$$\rho \circ T^{-1}(x) \det |\nabla T^{-1}(x)| = \pi(x) := T_{\#}\rho(x) \tag{1}$$

supposing ρ, π and T to be regular enough and that such a T exists.

2.1 Knothe-Rosenblatt rearrangement

As a special case of transport map between two measures, we will focus here on the Knothe-Rosenbaltt rearrangement or simply Knothe transport. This transport map will be usefull in a minimization problem that we will define later on. More details on this transport map can be find in [2].
First of all, we need to define the **increasing rearrangement formula**: Given $\mu, \nu \in \mathbb{P}(R^d)$ with cumulative distribution functions F resp. G and their right continuous inverses F^{-1} resp G^{-1}, we define :

$$T = G^{-1} \circ F$$

when μ is atomless, we have $T_{\#}\mu = \nu$ and it's an increasing transport map.

Now, we can define and construct the Knothe-Rosenblatt rearrangement:
By the increasing formula, we define $y_1 = T_1(x_1)$ by the increasing rearrangement formula of μ_1 into μ_1.

Then, for $1 < i \leq d$, we take the marginal on the first i variables and use the conditional probability formula:

$$\mu_i(dx_1 \cdots dx_i) = \mu_{i-1}(dx_1 \cdots dx_{i-1})\mu_i(dx_i|x_1, \cdots, x_{i-1})$$
$$\nu_i(dy_1 \cdots dy_i) = \nu_{i-1}(dy_1 \cdots dy_{i-1})\nu_i(dy_i|y_1, \cdots, y_{i-1}).$$

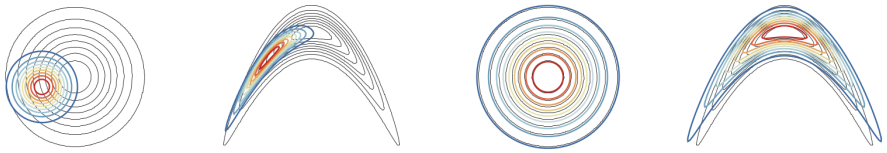
Next, we set $y_1 = T_1(x_1)$ and define $y_i = T_i(x_i; x_1, \cdots, x_{i-1})$ by the increasing rearrangement formula of $\mu_i(dx_i|x_1, \cdots, x_{i-1})$ into $\nu_i(dy_i|y_1, \cdots, y_{i-1})$.
Finally, we set $T(x) = (y_1, \cdots, y_n)$ to have the Knothe-Rosenblatt rearrangement.

By construction of the Knothe transport T, we have that $T_\#\mu = \nu$ and its Jacobian is upper triangular with positive entries on the diagonal.

3 Transport map MCMC

3.1 MCMC with a fixed transport map

Here, we suppose that with have a lower triangular transport map S such that $S_\#\nu \simeq \mu$, with the reference measure μ being a standard Gaussian. We will use this map in order to make a modified version of Metropolis-Hastings using this transport map.



(a) Local reference. **(b)** Local Target. **(c)** Global reference. **(d)** Local target.

Figure 1: Illustration of the transport map

This illustration taken from the article shows how the transport map can modify the proposal in order to make it more effective for our target space. Indeed, if we have a good proposal for our reference space, we also have a good proposol for the target space. More mathematically, let $q_r(r'|r)$ be a standard MH proposal in the reference space, we define:

$$q_\theta(\theta'|\theta) = q_r(S(\theta')|S(\theta)) \cdot |\det \nabla S(\theta')| \tag{2}$$

We will use q_θ as a proposal density in the MH algorithm. The acceptance rate is:

$$\alpha(\theta', \theta^{(k)}) = \min\left[1, \frac{\pi(\theta')q_\theta(\theta^{(k)}|\theta')}{\pi(\theta^{(k)})q_\theta(\theta'|\theta^{(k)})}\right] \tag{3}$$

The transport map makes it straightforward to evaluate this map, as $\theta' = S^{-1}(r')$ and $\theta^{(k)} = S^{-1}(\theta^{(k)})$. With the transport map, the loop in the Metropolis-Hastings algorithm proceeds as follows:

Algorithm 1 FixedTransportMCMC

for $k = 1$ **to** $K - 1$ **do**
    Compute the reference state: $r^{(k)} = S(\theta^{(k)})$
    Sample from the reference proposal: $r' \sim q_r(\cdot|r^{(k)})$
    Compute the target proposal: $\theta' = S^{-1}(r')$
    Accept the new proposal with probability α
    if the proposal is accepted **then**
        $\theta^{(k+1)} \leftarrow \theta'$
    else
        $\theta^{(k+1)} \leftarrow \theta^{(k)}$

The lower triangular structure of our map S facilitates computation of its inverse. Inverting S involves solving d one-dimensional inversion problems. For instance, each of these problems can be efficiently solved using Newton's algorithm.

3.2 Adaptative transport map MCMC

Even though presenting MCMC with a fixed transport map was revelant in order to understand the main idea of this modified Metropolis Hastings algorithm, we don't have such a map in practice. Indeed, to construct a map T such that $T_{\#}\nu \simeq \mu$ we need in practive $(\theta^{(1)}, \cdots, \theta^{(K)})$ samples from our target distribution ν. To do so, the algorithm proposed is adapting the transport map T when new samples $(\theta'^{(1)}, \cdots, \theta'^{(K)})$ are given. We can translate the algorithm as followed:

Algorithm 2 Adaptive Transport MCMC

$T_0 \leftarrow$ Identity
generate initial samples using T_0
for $k \leftarrow 1$ to N_{loops} **do**
    FixedTransportMCMC(T_K)
    generate new samples using the fixed transport map and accept/reject criteria
    estimate a new transport map T_{K+1} using the new samples
return samples

3.3 Optimisation problem

Having explored the adaptation of the Metropolis-Hastings algorithm to incorporate a transport map, we now turn to the construction of such a map. To begin, we define the set over which we will operate to approximate a transport map.
A Borel map $T : \mathbb{R}^d \to \mathbb{R}^d$ is said to be lower triangular if:
$T(x) = [T_1(x_1), T_2(x_1, x_2), \ldots, T_d(x_1, \ldots, x_d)]^T$.
We set:
$\mathcal{T}_\triangle := \{T$ lower triangular, strictly increasing with T and differentiable everywhere$\}$.

While the Knothe-Rosenblatt rearrangement belongs to $\mathcal{T}_\triangle \cap \mathcal{T}(\mu, \nu)$, it can be computationally challenging to work with. Instead, we aim to approximate the pushforward equality $T \in \mathcal{T}_\triangle$ and $T_{\#}\mu \simeq \nu$. To accomplish this, we leverage the Kullback-Leibler

divergence to quantify the similarity between our two probability densities and formulate a corresponding minimization problem:

$$\underset{T\in\mathcal{T}_\triangle(\mu)}{\arg\min}\ D_{KL}(\nu||T_\#\mu) \tag{4}$$

Thanks to the Knothe transport, we know that the minimum of this Kullback-Leibler divergence is 0 on this set.
Given that $T_\#\mu = \rho\circ T^{-1}(x)\det|\nabla T^{-1}(x)|$ this is equivalent to

$$\underset{T\in\mathcal{T}_\triangle(\mu)}{\arg\min}\ \mathbb{E}_\nu\left[-\log\rho\circ T^{-1}(x) - \log|\det\nabla T^{-1}(x)|\right] \tag{5}$$

Remark: in what follows, we will replace T^{-1} with S. Usually, we refer as Ta if it is in $\mathcal{T}(\mu,\nu)$, and S if it is in $\mathcal{T}(\nu,\mu)$. It doesn't change the problem since the inverse is unique and it's more readable.

3.3.1 Making the optimization problem easier

Now that we have set our optimization problem, we would like to make it computationnaly easier to solve. Suppose that we have $(x^{(1)},\dots,x^{(K)}) \overset{\text{iid}}{\sim} \mu$, we can approximate our problem:

$$\underset{S\in\mathcal{T}_\triangle(\mu)}{\arg\min}\frac{1}{K}\sum_{k=1}^{K}\left[-\log\rho\circ S(x^{(k)}) - \log|\det\nabla S(x^{(k)})|\right] \tag{6}$$

Then, by choosing $\mu\sim\mathcal{N}(0,I)$ and knowing that $T\in\mathcal{T}_\nabla$, we have an easy formula for $\log\rho(x)$ and the Jacobian that leads to:

$$\underset{S\in\mathcal{T}_\triangle(\mu)}{\arg\min}\left[\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{1}{2}(S_i(x^{(k)})^2 - \log\frac{\partial S_i}{\partial x_i}(x^{(k)})\right] \tag{7}$$

In order to make each S_i inversible, we need to assure that for all k:

$$\frac{\partial S_i(\theta^{(k)})}{\partial\theta_i} > 0 \tag{8}$$

We add this constraint into $\mathcal{T}_\triangle$ without changing the space name. We will not discuss more details on the conditions for the map S that are necessary to proove the ergodicity of the algorithm proposed in the article. Nevertheless, we can notice that the conditions on the map S and the formulation above lead to a seprable optimisation problem over the dimension and that each problem is convex on a closed feasible domain. We refer to the article for more details.

3.3.2 Map parametrization

With equation (7), we have a separable optimization problem. To solve it, the article proposes to reparametrize our maps. For example, we can reparametrize each map S_i with a multivariate polynomial expension.
As an example, lets take $\mu,\nu\in\mathbb{P}(\mathbb{R}^2)$ and the subsace $\mathbb{R}_2[X_1,X_2]$ of polynomials of $\mathbb{R}^2$ with a degree lower or equal to two. Each $P\in\mathbb{R}_2[X_1,X_2]$ is of the form:

$$P(X_1, X_2) = a_1 X_1^2 + a_2 X_2^2 + b_1 X_1 + b_2 X_2 + c X_1 X_2 + d$$

with $a_1, a_2, b_1, b_2, c, d \in \mathbb{R}$. Each polynomial of $\mathbb{R}_2[X_1, X_2]$ can be identified with its coefficient $\gamma = (a_1, a_2, b_1, b_2, c, d)$.
With this family of functions, we have;

$$S_{\gamma_1,1}(X_1, X_2) = a_{1,1} X_1^2 + b_{1,1} X_1 + c_1 X_1 X_2 + d \tag{9}$$

$$S_{\gamma_2,2}(X_1, X_2) = a_{1,2} X_1^2 + a_{2,2} X_2^2 + b_{1,2} X_1 + b_{2,2} X_2 + c_2 X_1 X_2 + d_2 \tag{10}$$

Where $\gamma_1 = (a_{1,1}, b_{1,1}, c_1)$ and $\gamma_2 = (a_{1,2}, a_{2,2}, b_{1,2}, b_{2,2}, c_2)$. With this reparametrization, the optimization problem (7) becomes:

$$\arg\min_{\gamma_1,\gamma_2} \left[\sum_{i=1}^{n} \sum_{k=1}^{K} \frac{1}{2} (S_{i,\gamma_i}(x^{(k)})^2 - \log \frac{\partial S_{i,\gamma_i}}{\partial x_i}(x^{(k)}) \right] \tag{11}$$

Here is an example with a 2d mixture of three gaussians:

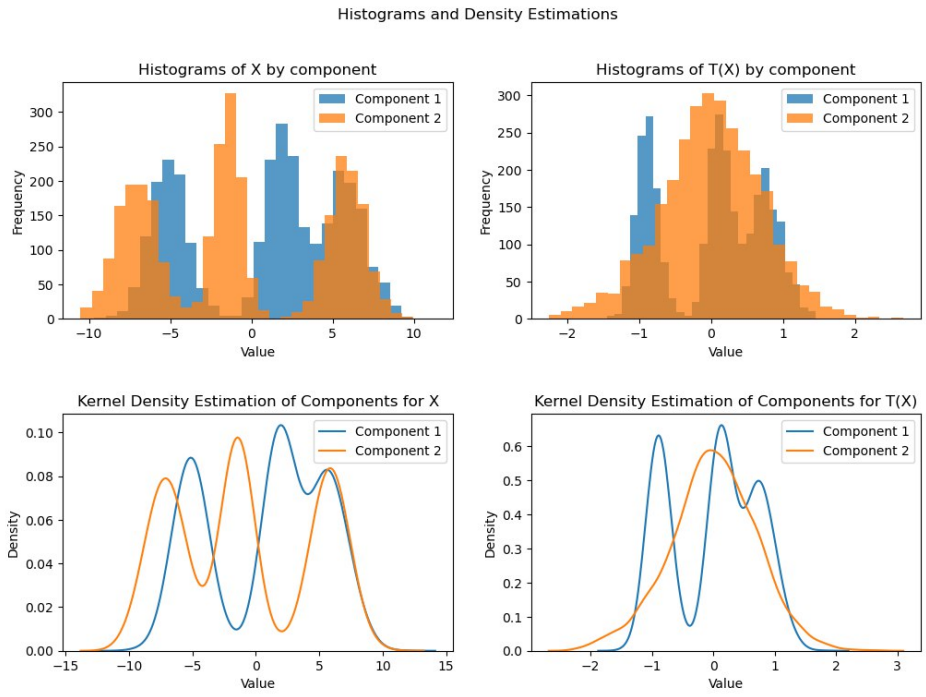


Figure 2: Effects of the transport map of a 2D gaussian mixture

The transformation resulted in a density that is much more similar to a Gaussian distribution. Specifically, the second component changed from a trimodal to unimodal distribution, while the variability in the first component decreased. Although a higher degree polynomial could provide a more accurate approximation, the partial derivative constraints restrict the T_1 function to almost linear behavior, even with a degree 3 polynomial.
It is clear that we can extend this example to higher dimensions by increasing the maximum degree of our multivariate polynomials or by using alternative bases such as orthogonal polynomials like Hermite or Legendre polynomials. An optimal choice

of basis would be one that results in linear S_i maps with respect to the expansion coefficients γ_i, as recommended in the paper. This property improves the efficiency of solving the minimization problem (1), further details can be found in the article.

4 Experimenting the algorithm

4.1 Transport map

In our experiments, we choosed to approximate our transport map with 2 degree polynomials like in section 3.3.2. We will illustrate here the modification applied to a 2D mixture of three gaussians with a transport map minimizing (11).
Our focus now is to examine the practical effectiveness of accelerating MCMC using transport maps. Our approach involves utilizing the Metropolis-Hastings algorithm and comparing it with the transport map variant. It's worth noting that transport maps can also be applied to other MCMC algorithms, such as Langevin Monte Carlo, making it a versatile tool.
Our example target distribution is the following two dimension gaussian mixture:

$$\frac{1}{3}\mathcal{N}([-3,0],I_2)+\frac{1}{3}\mathcal{N}([3,0],I_2)+\frac{1}{3}\mathcal{N}([0,3],I_2) \tag{12}$$

Although the target distribution is not particularly difficult, we will observe that the MH algorithm struggles to sample from it. Additionally, using a 2D example is advantageous for visualizing the problem, as increasing the number of dimensions would complicate our understanding.

For instance, here is what this target distribution looks like.

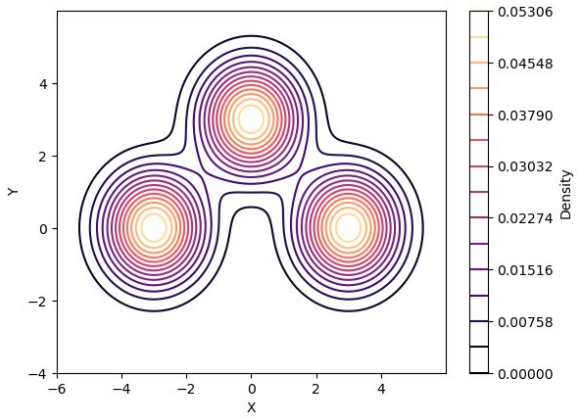


Figure 3: Contour map of our 2D gaussian mixture

We can also have alook on the map transformation used to make the target distribution closer to our source distriubtion, using 1000 samples from it tu have $T_{\#}\nu \simeq \mu$ for the KL divergence.

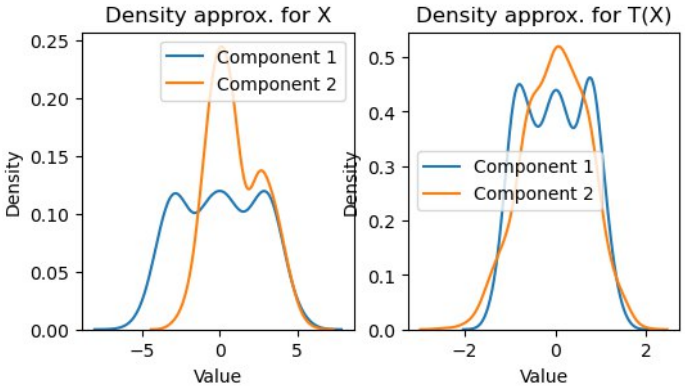
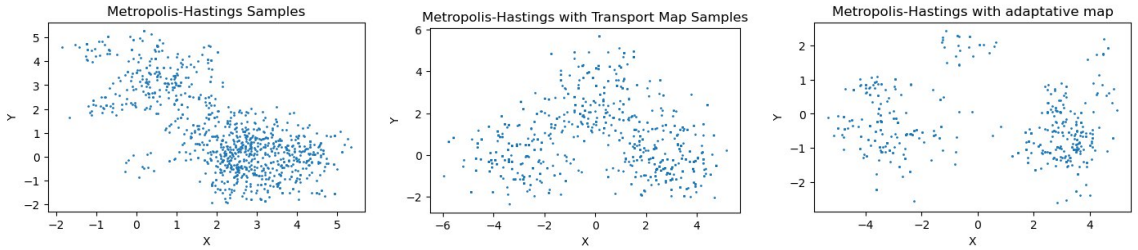


Figure 4: Effect of the transport map on our mixture

As we can see, with the transformation, each component is way closer to a normalized 1D gaussian, as we wanted.

4.1.1 Comparing the algorithms



(a) Vanilla MH (b) Fixed map MH (c) Adaptative map MH

Figure 5: Samples comparison between algorithms

The fixed transport map algorithm demonstrate superior performance compared to the vanilla MH algorithm. Both MH and adaptative transport MH performed poorly with only 1 000 samples. Adaptive map's poor performance is expected, as it's based on MH samples and there is nearly no map updates here. However, in real-world cases, without a fixed map, the adaptive version is the only viable transport map option.

We will now compare adaptive transport map MH and vanilla MH using 5,000 samples for a clearer view of each algorithm's ergodic behavior.

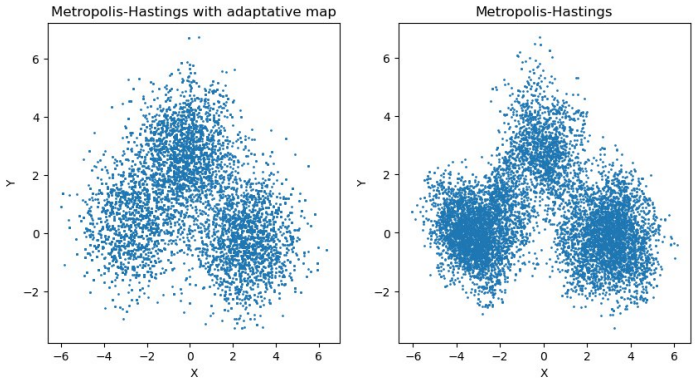


Figure 6: Comparison between MH and adaptative transport MH with 5 000 samples

From the results, it appears that the adaptive transport map algorithm has achieved convergence to the distribution using 5,000 samples, while one mode of the mixture appears to be lagging behind with the Metropolis-Hastings algorithm. However, we can see that the samples generated by the new algorithm are sparser compared to the vanilla MH algorithm, which produces samples that are much closer together. We could finetune the proposal covariance to correct this behavior.

5 Conclusion

Adaptive transport maps show potential for improving MCMC efficiency, though solving the optimization problem is computationally intensive. Our experiments didn't utilize parallelization or compare time efficiency. However, as the paper suggests, map update costs should decrease over time. As samples near the true target distribution, parameters converge to the true optimum, making a gradient-based solver's initial guess better for this convex problem.

Careful tuning of parameters, like map update frequency and basis functions, is crucial for successful adaptation. In practive, this tuning can be hard. Moreover, the map estimation relates strongly on samples generated to build them, making this algorithm dependant of non transport MCMC sammples to start with.

Some possible extensions for transport map MCMC methods could include adaptive basis functions (e.g., wavelets or neural networks) for better accuracy and flexibility, and incorporating domain knowledge to inform the choice of auxiliary distribution or transport map. For example, if the target distribution is known to have certain symmetries or constraints, incorporating this knowledge into the transport map could improve efficiency and accuracy.

References

[1] Parno Matthew and Marzouk Youssef. "Transport map accelerated Markov chain Monte Carlo". In: (Dec. 2014). arXiv: `1412.5492 [stat.CO]`.

[2] Santambrogio F. *Optimal Transport for Applied Mathematicians*. Calculus of Variations,PDEs, and Modeling. ISBN: 9783319208282.