An Overview of Transport Map MCMC

Robert Zimmerman

Report for STA4519H - Optimal Transport: Theory & Algorithms

December 4, 2019

1 Introduction

Markov chain Monte Carlo (MCMC) algorithms constitute a broad class of methods which yield samples from probability distributions. These techniques have become ubiquitous in Applied Statistics since they were originally conceived in the early 1950s ([**Met+53**]). While other more direct sampling methods can be used to sample from simple univariate distributions, MCMC is essentially the only choice when the distribution from which we want to sample – hereafter referred to generally as the *target distribution* – is unwieldy.

The basic structure of an MCMC algorithm is very simple. Such an algorithm constructs a stochastic process $\{X_n\}$ by sampling, at the n'th iteration, a random variable Y_n from some *proposal distribution* (which is much easier to sample from) and accepting the proposal as $Y_n = X_n$ (otherwise, $X_n = X_{n-1}$) according to a rule that makes $\{X_n\}$ a Markov chain; under certain conditions, the stationary distribution of this Markov chain is exactly the target distribution. Approximate samples from the target can then be extracted as $X_B, X_{B+1}, \ldots$ for sufficiently large B.

While MCMC can theoretically be used to (approximately) sample from virtually any target distribution, sampling can still be arduous when the target is difficult to compute or is multimodal, particularly in high dimensions. The former scenario is often the case in the Bayesian setting, in which we are given some data assumed to follow some parametric distribution, as well as a prior distribution on one or more of the parameters, and we want to obtain draws from the posterior distribution. This requires us to compute the full likelihood function at every iteration of the algorithm, which can be prohibitively expensive when the sample size is large (and is compounded when the algorithm gets stuck in a mode and therefore rejects most proposals).

As such, the development of new MCMC techniques (which has blossomed since the since the early 1990s) has been driven by the need for faster algorithms, along with the necessary convergence theory to support it. Here, "faster" essentially means that the algorithm explores the support of the target distribution more quickly, which ultimately leads to faster convergence to the target distribution. One well-known set of techniques used to accomplish that are adaptive MCMC methods, which (at the cost of certain theoretical guarantees) are faster because they automatically "re-learn" better parameter values while they run ([**RR09**]).

Meanwhile, the field of optimal transport has been growing in popularity recently as well. The theory began with an (apparently simple) optimization problem posed in the late 1700s by Gaspard Monge in the late 1700s ([**Mon81**]), who was interested in the cheapest way to move one configuration of sand to another ([**San15**]). The problem attracted renewed interest the 1940s with the generalization of Monge's problem by Kantorovich ([**Kan58**]), and has seen a major development starting from the 1980s. It has found applications in diverse fields such as economics, imaging, traffic, urban planning ([**San10**]), and cosmology.

Matthew D. Parno, in his PhD thesis at MIT together with his advisor Youssef M. Marzouk (who we hereafter refer to as *the authors*), have integrated the theory of optimal transport into MCMC in order to produce a fast MCMC sampler. The algorithm and underlying theory were published in 2017 as the paper *Transport map accelerated Markov chain Monte Carlo* ([**PM18**]), which is the focus of this report. Their method lends itself well to parallelization, although the authors have not experimented with this in the paper and only briefly remarked on the potential. We discuss more in Section 5.

The basic idea at the heart of their method is to use transport maps to iteratively transform the usual Metropolis-based proposal distributions into non-Gaussian proposals that explore the target density more efficiently, which can be considered a type of adaptive MCMC approach. This approach constructs transport maps which are solutions of convex optimisation problems, and evaluates those maps in order to draw from the proposal distribution. Solutions to convex optimizations can be constructed efficiently, and the lower triangular structure of the maps simplifies their evaluation. Thus, the algorithm is efficient. Moreover, the authors prove that the resulting Markov chain converges (in a strong sense) to the target distribution under relatively weak assumptions – a desirable requirement of MCMC algorithms in general.

The remainder of this report is organized as follows. Section 2 lays out the technical background required to mathematically present the main results of the paper. Section 3 summarizes these results and the accompanying MCMC algorithm. Section 4 is devoted to the authors' proof that the algorithm is ergodic. Finally, Section 5 discusses future directions of research.

2 Technical Background

2.1 Markov Chain Monte Carlo

In the framework of MCMC, we are interested in obtaining samples $\theta^{(1)}, \theta^{(2)}, \ldots$ distributed according a target distribution defined by a measure μ_π which has a density $\pi(\theta)$ (more concisely, we say that say that we want to "sample from π"). Typically, π is difficult – or impossible – to sample from directly, usually because the domain is multidimensional and/or π is only known up to its normalizing constant. In MCMC, we construct a Markov chain $\{\theta^{(k)}\}_k$ whose stationary distribution is exactly π without facing these problems.

At heart, essentially every MCMC algorithm involves repeating a variant of the so-called *Metropolis accept/reject* step in order to produce the desired Markov chain. At each iteration of the algorithm a *proposal* random variable θ' is drawn according to a conditional proposal density $q(\cdot \mid \theta)$, where θ is the sample obtained from the previous iteration; usually, θ is used as a parameter of q. When this proposal density is symmetric in the sense that $q(\theta' \mid \theta) = q(\theta \mid \theta')$, the proposal θ' is accepted as a new sample with probability $\min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right)$; the resulting algorithm is called the *Metropolis algorithm*. On the other hand, when $q(\cdot \mid \theta)$ is not symmetric, then the acceptance probability can be modified as $\min\left(1, \frac{\pi(\theta')q(\theta'|\theta)}{\pi(\theta)q(\theta|\theta')})\right)$ and this algorithm is called the *Metropolis-Hastings algorithm*.

MCMC is not an ideal sampling method, and is only suitable when direct sampling is impossible. For example, the random variables produced by MCMC algorithms are correlated, and therefore not independent draws from any distribution. More importantly, the stationary distribution μ_θ is only reached asymptotically, so that the random variables produced are only approximately distributed according to μ_θ. Moreover, it is not even guaranteed in the first place that the Markov chain converges to μ_θ, where convergence in this sense refers to the chain being ergodic.

Definition 1 (Ergodicity)**.** Let $\{X_n\}$ be a Markov chain on some state space $\mathcal{X}$ with stationary measure μ_θ and n-step transition kernel $P^n(x, \cdot) = \mathbb{P}(X_n \in \cdot \mid X_0 = x)$. Then $\{X_n\}$ is said to be *ergodic* if P^n converges to μ_θ in total variation; that is, if

$$\lim_{n\to\infty} \sup_{A\subseteq\mathcal{X}} |P^n(x, A) - \mu_\theta(A)| = 0, \qquad \text{for } \mu_\theta\text{-a.e. } x \in \mathcal{X}.$$

When the state space $\mathcal{X}$ is finite, then the Markov chain is always ergodic ([**Ros95**]). More generally, the so-called "fundamental theorem of Markov chains" states that if a Markov chain is irreducible and aperiodic with μ_θ as a stationary measure, then the chain is ergodic. In Metropolis algorithms and Metropolis-Hastings

algorithms, the stationarity of μ_θ is essentially guaranteed by the construction of the Metropolis acceptance probability, and the other two criteria are usually satisfied as well.

While results like the fundamental theorem provide asymptotic guarantees, the practical ability of the Markov chain produced by an MCMC algorithm to reasonably approximate the stationary distribution is governed by the proposal distribution's size and spatial orientation, which are difficult to tune properly ([**HST01**]). The idea of *adaptive MCMC*, devised in the early 2000s, uses the history of the process to tune the proposal distribution on-the-fly. For example, it is known that for a standard Metropolis algorithm with a d-dimensional $\mathcal{N}(0, \Sigma)$ proposal, the optimal choice for Σ (under certain regularity conditions) is $\frac{2.38^2}{d}\Sigma_0$, where Σ_0 is the covariance of the target distribution μ_θ ([**RR01**]). Of course, Σ_0 is almost always unknown; a typical adaptive version of the algorithm therefore estimates Σ_0 by the empirical covariance matrix Σ_n based on the samples produced up to the n'th step of the algorithm.

Unfortunately, the chains produced by adaptive MCMC algorithms are non-Markovian, and hence do not preserve the stationarity of μ_π. As such, the fundamental theorem does not apply, and convergence to μ_π is not guaranteed. However, [**RR07**] show that convergence of the adaption is satisfied under certain conditions – namely, that the algorithm satisfies the diminishing adaptation property and that the family of transition kernels induced by the algorithm is simultaneously strongly aperiodically geometrically ergodic.

Definition 2 (Diminishing adaptation property)**.** An adaptive MCMC algorithm has the *diminishing adaptation* property when the total variation distance between successive transition kernels $P_{\gamma^{(n)}}$ and $P_{\gamma^{(n+1)}}$ converges to 0 in probability:

$$\sup_{x\in\mathbb{R}^n}\sup_{A\subseteq\mathcal{X}}|P_{\gamma^{(n+1)}}(x,A)-P_{\gamma^{(n)}}(x,A)|\xrightarrow{P}0.$$

Definition 3 (Simultaneously strongly aperiodically geometrically ergodic)**.** A family of transition kernels $\{P_\gamma\}$ parameterized by a vector of map parameters $\gamma\in\Gamma$ is *simultaneously strongly aperiodically geometrically ergodic* if there exists a Borel set $C\subseteq\mathbb{R}^n$, a drift function $V:\mathbb{R}^n\to[1,\infty)$ with $\sup_{x\in C}V(x)<\infty$, and scalars $\delta>0$, $\lambda<1$, and $b<\infty$ such that the following two conditions hold:

1. (Minorization) For each $\bar\gamma\in\Gamma$, there is a probability measure $\nu_{\bar\gamma}(\cdot)$ defined on C with $P_{\bar\gamma}(x,\cdot)\geq\delta\nu_{\bar\gamma}(\cdot)$ for all $x\in C$.
2. (Simultaneous drift) The inequality $\int_{\mathbb{R}^n}V(x)\,P_{\bar\gamma}(x,\mathrm{d}x)\leq\lambda V(x)+b\mathbb{1}_C(x)$ holds for all $\bar\gamma\in\Gamma$ and $x\in\mathbb{R}^n$.

The following theorem, which appears as Theorem 3 in [**RR07**], is used by the authors to prove that their transport map MCMC algorithm is ergodic:

Theorem 1. Consider an adaptive MCMC algorithm which satisfies diminishing adaptation and whose family of transition kernels is simultaneously strongly aperiodically geometrically ergodic with drift function V satisfying $\mathbb{E}\left[V(X_0)\right]<\infty$. Then the adaptive MCMC algorithm is ergodic.

2.2 Optimal Transport

The field of optimal transport, dating back to the work of Gaspard Monge in the late 18th century, was motivated by an ostensibly simple optimization problem: that of finding the cheapest way to move one configuration of sand to another ([**San15**]). In the modern formulation, we are concerned with the existence of a specific *transport map* between two Borel probability measures μ_θ and μ_r on $\mathbb{R}^n$ – that is, any transformation $T:\mathbb{R}^n\to\mathbb{R}^n$ such that $\mu_r=T_\sharp\mu_\theta$ (i.e., μ_r is the pushfoward measure of μ_θ).

There may be infinitely many transport maps between two given probability measures, or there may be none at all; the latter case occurs when, for example, $\mu_\theta=\delta_a$ for some $a\in\mathbb{R}$ and μ_r is non-atomic, since

$T_\sharp \delta_a = \delta_{T(a)}$ ([**San15**]). In the former case, one can define a *cost function* $c : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and choose the particular transport map T which solves the Monge problem:

$$T = \underset{\mu_r = T_\sharp \mu_\theta}{\arg\min} \int_{\mathbb{R}^n} c(\theta, T(\theta))\, \mathrm{d}\mu_\theta(\theta).$$

The minimum may not exist for general cost functions. However, for the quadratic cost function $c(x, y) = \frac{1}{2}|x-y|^2$, Brenier's theorem ([**Bre91**]) asserts that such a map T exists, and is unique and monotone μ_θ-a.e. Maps which solve the Monge problem are called *optimal transport maps*. Other cost functions yield other maps; one particular example that we will use later is the Knothe-Rosenblatt rearrangement:

Definition 4 (Knothe-Rosenblatt rearrangement)**.** Let $\theta = (\theta_1, \ldots, \theta_n) \sim \mu_\theta$ and $r = (r_1, \ldots, r_n) \sim \mu_r$, and let T_t be the optimal transport maps corresponding to the cost function

$$c_t(\theta, r) = \sum_{j=1}^{n} t^{j-1} |\theta_i - r_i|^2. \tag{1}$$

The limiting map T_t as $t \to 0$ is called the Knothe-Rosenblatt arrangement between μ_θ and μ_r.

The Knothe-Rosenblatt arrangement exists and is uniquely defined when μ is absolutely continuous with respect to Lebesgue measure. Moreover, it features an extremely useful property: its Jacobian ∇T is lower triangular and has positive diagonal entries μ-a.e. The authors take advantage of this property, discussed further in Section 3.

3 Summary of Main Contribution

The authors integrate transport maps by choosing an additional measure μ_r that is easy to sample from, which they call a *reference measure*. For example, they typically take $\mu_r = \mathcal{N}(0, I)$. Then, they choose a transport map $\tilde{T}$ such that μ_r is *approximately* the pushforward measure $\tilde{T}_\sharp \mu_\theta$. Next, they construct an MCMC sampler which operates at each iteration in the following fashion: first, it samples a proposal r' from μ_r on the "reference space" using a freely-chosen proposal density $q_r(r' \mid r)$, and then computes the pullback of this proposal through $\tilde{T}$, producing a sample $\theta' = \tilde{T}^{-1}(r')$ distributed according to a proposal density $q_\theta(\theta' \mid \theta)$ on the "target space". This target-space proposal density is fully determined by q_r and $\tilde{T}$. The proposal θ' is then accepted using the standard Metropolis-Hastings accept/reject criterion. The algorithm is made adaptive by fine-tuning the transport map $\tilde{T}$ after every several iterations (by way of solving an optimization problem), based on the current sample. The full algorithm is presented in Algorithm 1; we highlight some of the technical foundations below.

Why use a transport map at all? Primarily because the choice of an appropriate map $\tilde{T}$ results in a proposal density q_θ which explores the target density π more efficiently, while still capturing the structure of μ_θ. Next we address the choice of map $\tilde{T}$. While the optimal transport map T_t induced by (1) in the Knothe-Rosenblatt rearrangement is appealing due to the convenient properties discussed in Section 2, the authors instead seek a practical approximation of T_t which shares the same desirable properties; namely, a map $\tilde{T}$ such that $\nabla \tilde{T}$ is lower triangular and for which $\mu_r \approx \tilde{T}\mu_\theta$. The authors justify the use of such an approximation with three considerations. First, there are computational issues in using the sequence of weights $\{t^i\}$ in (1). Second, finding an exact map T_t can be difficult, especially if the target contains many nonlinear dependencies that are not present in the reference distribution. Third, we can impose regularity conditions on an approximation (i.e., that it be a C^1-diffeomorphism as well as further constraints) which may not hold otherwise.

The question now becomes how to find a suitable approximation $\tilde{T}$ of the Knothe-Rosenblatt rearrangement. It is assumed that both μ_θ and μ_r are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$,

with densities π and ρ, respectively. The idea is to choose a density $\tilde{\pi}(\theta)$ which depends on some map $\tilde{T}$, and then to choose $\tilde{T}$ by minimizing the distance between $\tilde{\pi}(\theta)$ and $\pi(\theta)$. More specifically, with the requirement that $\tilde{T}$ be monotone and differentiable μ_θ-a.e., the authors choose $\tilde{\pi}(\theta)$ to be the density of the pullback of μ_r through $\tilde{T}$ (which the authors refer to as the *map-induced density*) which, by the change-of-variables formula (see Remark 2.6 of [**PC19**]), takes the form

$$\tilde{\pi}(\theta) = \rho(\tilde{T}(\theta)) \cdot |\det \nabla \tilde{T}(\theta)|. \tag{2}$$

They then choose $\tilde{T}$ as the map (living in some subspace of lower triangular functions from $\mathbb{R}^n$ to $\mathbb{R}^n$) that minimizes the Kullback-Leibler divergence between π and $\tilde{\pi}$, which is given by

$$D_{\mathrm{KL}}(\pi \| \tilde{\pi}) = \mathbb{E}_\pi \left[\log \pi(\theta) - \log \rho\big(\tilde{T}(\theta)\big) - \log \left| \det \nabla \tilde{T}(\theta) \right| \right].$$

Taking the expectation over π is very beneficial, for if $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(K)} \overset{iid}{\sim} \pi$, then we can approximate the expectation with a sample average, ignore the $\log \pi(\theta)$ term (since π is fixed), and choose

$$\tilde{T} = \operatorname*{arg\,min}_{T \in \mathcal{T}} \frac{1}{K} \sum_{k=1}^{K} \left[- \log \rho\big(\tilde{T}(\theta^{(k)}\big) - \log \left| \det \nabla \tilde{T}(\theta^{(k)} \right| \right],$$

where $\mathcal{T}$ is some space of lower-triangular functions from $\mathbb{R}^n$ to $\mathbb{R}^n$. The choice of $\mathcal{T}$ is determined by the properties required for the map-induced density to exist (i.e., the maps must be differentiable and monotone), and the need for the resulting MCMC sampler to be ergodic, which requires that the maps be bi-Lipschitz; that is,

$$\lambda_{\min} \|\theta' - \theta\| \leq \|\tilde{T}(\theta') - \tilde{T}(\theta)\| \leq \lambda_{\max} \|\theta' - \theta\| \quad \text{for some } 0 < \lambda_{\min} \leq \lambda_{\max} < \infty. \tag{3}$$

The authors handle the two inequalities here separately. With the ultimate choice of $\tilde{T}$, the lower bound is equivalent to $\frac{\partial \tilde{T}_i}{\partial \theta_i} \geq \lambda_{\min}$ for $i = 1, \ldots, n$, which makes $\det \nabla \tilde{T}(\theta) > 0$. The equivalent condition can be replaced in practice by the weaker condition that $\frac{\partial \tilde{T}_i}{\partial \theta_i}\big|_{\theta^{(k)}} \geq \lambda_{\min}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. While many choices of $\tilde{T}$ can yield unbounded derivatives as $\|\theta\| \to \infty$, the upper bound can be enforced by choosing a sufficiently large $R > 0$, and constructing a map $\tilde{T}^R$ which is identical to $\tilde{T}$ inside the ball of radius R centered at 0, but linear outside it.

In practice, the set of candidate maps $\mathcal{T}$ must be finite-dimensional. Therefore, we can parameterize the i'th component of a candidate map $\tilde{T}(\theta)$ by a vector $\gamma_i \in \mathbb{R}^{M_i}$ and write the component as $\tilde{T}_i(\theta, \gamma_i)$. The complete map $\tilde{T}(\theta)$ is then parameterized by $\bar{\gamma} = [\gamma_1, \ldots, \gamma_n]$. The authors choose to parameterize each component $\tilde{T}_i$ with multivariate polynomial expansions (although any map which is linear in the coefficients $\bar{\gamma}$ will do). When the reference density $\mu_r = \mathcal{N}(0, I)$, the optimization problem reduces to

$$\operatorname*{arg\,min}_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\frac{1}{2} \tilde{T}_i^2(\theta^{(k)}) - \log \left. \frac{\partial \tilde{T}_i}{\partial \theta_i} \right|_{\theta^{(k)}} \right],$$

which separates into n individual convex optimization problems which can be solved in parallel, each given by

$$\min_{\gamma_i \in \mathbb{R}^{M_i}} \sum_{k=1}^{K} \left[\frac{1}{2} \tilde{T}_i^2(\theta^{(k)}; \gamma_i) - \log \left. \frac{\partial \tilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \right] \quad \text{s.t.} \quad \left. \frac{\partial \tilde{T}_i(\theta; \gamma_i)}{\partial \theta_i} \right|_{\theta^{(k)}} \geq \lambda_{\min} \ \ \forall k \in \{1, \ldots, K\}.$$

With this choice of $\tilde{T}$, the authors define a MCMC algorithm by taking the reference density ρ to be a standard Metropolis-Hastings proposal on the reference space $q_r(r' \mid r)$; the map-induced density defined by (2) is a density on the target space, which is given by

$$q_{\theta, \bar{\gamma}}(\theta' \mid \theta) = q_r\left(\tilde{T}(\theta') \mid \tilde{T}(\theta)\right) \cdot \left| \det \nabla \tilde{T}(\theta') \right|. \tag{4}$$

The authors go a step further by making the algorithm adaptive, in the sense of updating the maps $\tilde{T}$. The sampler is initialized by a simple map $\tilde{T}_0$, and updates the parameters γ_i after every K_U steps. This is performed efficiently by introducing a regularization term $g(\gamma_i)$ into the objective function, which ensures that the map does not prematurely get stuck in one particular region of the target space. The authors suggest a quadratic penalty term centered on the coefficients of the identity map: $g(\gamma_i) = k_R\|\gamma_i - \gamma_i^{\mathrm{Id}}\|^2$, where k_R is a user-defined regularization parameter. With this regularization term in place, each optimization problem (in dimension i) becomes

$$\min_{\gamma_i \in \mathbb{R}^{M_i}} k_R\|\gamma_i - \gamma_i^{\mathrm{Id}}\|^2 + \sum_{k=1}^{K}\left[\frac{1}{2}\tilde{T}_i^2(\theta^{(k)};\gamma_i) - \log\frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\bigg|_{\theta^{(k)}}\right] \quad \text{s.t.} \quad \frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\bigg|_{\theta^{(k)}} \geq \lambda_{\min} \;\; \forall k \in \{1,\dots,K\}. \tag{5}$$

Thus, the transport map based MCMC algorithm can be written out in full as Algorithm 1:

Algorithm 1: MCMC algorithm with adaptive map

Input : Initial state θ_0;
Initial vector of transport map parameters $\bar{\gamma}_0$;
Reference proposal $q_r(\cdot \mid r^{(k)})$;
Number of steps K_U between map adaptations;
Total number of steps L;
Regularization parameter k_R;
Minimum bi-Lipschitz threshold $\lambda_{\min}$

Output: MCMC samples of the target distribution, $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(L)}\}$

Set state $\theta^{(1)} = \theta_0$;
Set parameters $\bar{\gamma}^{(1)} = \bar{\gamma}_0$;
for $k \leftarrow 1$ **to** $L-1$ **do**
  Compute the reference state, $r^{(k)} = \tilde{T}(\theta^{(k)}; \bar{\gamma}^{(k)})$;
  Sample the reference proposal, $r' \sim q_r(\cdot \mid r^{(k)}$;
  Compute the target proposal sample, $\theta' = \tilde{T}^{-1}(r'; \bar{\gamma}^{(k)})$;
  Compute the acceptance probability

$$\alpha_{\bar{\gamma}^{(k)}}(\theta', \theta) = \min\left\{1, \frac{\pi(\tilde{T}^{-1}(r';\bar{\gamma}^{(k)}))}{\pi(\tilde{T}^{-1}(r^{(k)};\bar{\gamma}^{(k)}))}\frac{q_r(r^{(k)} \mid r')}{q_r(r' \mid r^{(k)})}\frac{\det\nabla\tilde{T}^{-1}(r';\bar{\gamma}^{(k)})}{\det\nabla\tilde{T}^{-1}(r^{(k)};\bar{\gamma}^{(k)})}\right\} \tag{6}$$

  Set $\theta^{(k+1)} = \theta'$ with probability $\alpha(\theta', \theta)$; else set $\theta^{(k+1)} = \theta^{(k)}$;
  if $k \equiv 0 \pmod{K_U}$ **then**
    for $i \leftarrow 1$ **to** n **do**
      Update $\gamma_i^{(k+1)}$ by solving

$$\min_{\gamma_i \in \mathbb{R}^{M_i}} k_R\|\gamma_i - \gamma_i^{\mathrm{Id}}\|^2 + \sum_{j=1}^{k+1}\left[\frac{1}{2}\tilde{T}_i^2(\theta^{(j)};\gamma_i) - \log\frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\bigg|_{\theta^{(j)}}\right]$$
$$\text{s.t.} \quad \frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\bigg|_{\theta^{(j)}} \geq \lambda_{\min} \;\; \forall j \in \{1,\dots,k+1\}$$

  else
    Set $\bar{\gamma}^{(k+1)} = \bar{\gamma}^{(k)}$;
return Target samples $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(L)}\}$

4 Proof of Ergodicity

The primary theoretical contribution in [**PM18**] is that the Markov chain produced by Algorithm 1 is ergodic – that is, it converges to the target distribution $\pi(\theta)$. The proof, which we summarize in this section, relies on Theorem 1. In particular, the authors verify the three conditions of that theorem (i.e., diminishing adaptation, minorization, and simultaneous drift) in three separate propositions.

Theorem 2 (Ergodicity of Algorithm 1)**.** The Markov chain produced by Algorithm 1 is ergodic for the target distribution $\pi(\theta)$ when $\bar{\gamma}$ is constrained to a compact set within which $\tilde{T}(\theta; \bar{\gamma})$ is guaranteed to satisfy the bi-Lipschitz condition in 3 for all $\theta \in \mathbb{R}^n$.

The proof assumes that the target density $\pi(\theta)$ is finite, continuous, and super-exponentially light, where the latter term means that

$$\limsup_{\|\theta\| \to \infty} \frac{\theta}{\|\theta\|} \cdot \nabla \log \pi(\theta) = -\infty.$$

Note that the map-induced density $\tilde{\pi}$ need not satisfy this requirement. The authors further assume that the reference proposal density $q_r(r' \mid r)$ is Gaussian with bounded mean. They denote Γ for the space of map parameters $\bar{\gamma}$ such that $\tilde{T}(\theta; \bar{\gamma})$ satisfies the bi-Lipschitz condition (3), and write

$$P_{\bar{\gamma}}(\theta, A) = \int_A \left(\alpha_{\bar{\gamma}}(\theta', \theta) \cdot q_{\theta, \bar{\gamma}}(\theta' \mid \theta) + r(\theta) \cdot \delta_\theta(\theta') \right) \mathrm{d}\theta'$$

for the transition kernel of the chain at the k'th iteration, where $q_{\theta,\bar{\gamma}}(\theta' \mid \theta)$ is the map-induced density on the target space given by (4), $\alpha_{\bar{\gamma}}(\theta', \theta)$ is the Metropolis acceptance probability given by (6), and $r(\theta) = 1 - \int \alpha(\theta', \theta) q_{\theta,\bar{\gamma}}(\theta' \mid \theta) \, \mathrm{d}\theta'$ is the probability that the chain remains at θ. For any current value $x \in \mathbb{R}^n$ and parameterization $\bar{\gamma} \in \Gamma$, they write the set of proposals $y \in \mathbb{R}^n$ which are guaranteed to be accepted as $A_{\bar{\gamma}}(x) = \{y \in \mathbb{R}^n : \alpha_{\bar{\gamma}}(\theta', \theta) = 1\}$. Similarly, the authors write $R_{\bar{\gamma}}(x)$ for $A_{\bar{\gamma}}(x)^C$, the set of proposals $y \in \mathbb{R}^n$ which can possibly be rejected.

The following three technical lemmas, stated here without proof, are used in the subsequent proofs of Propositions 1, 2, and 3.

Lemma 1 (Bounded target space proposal)**.** For any map coefficents $\bar{\gamma} \in \Gamma$, the map-induced proposal $q_{\theta,\bar{\gamma}}(\theta' \mid \theta)$ is bounded as

$$k_L g_L(\theta' - \theta) \leq q_{\theta,\bar{\gamma}}(\theta' \mid \theta) \leq k_U g_U(\theta' - \theta),$$

where $k_L = k_1 \lambda_{\min}^n$, $k_U = k_2 \lambda_{\max}^n$, $g_L(x) = g_1(\lambda_{\max} x)$, and $g_U(x) = g_2(\lambda_{\min} x)$.

Lemma 2. Let $V(x) = c_V \pi^{-\alpha}(x)$ be a drift function defined for some $\alpha \in (0, 1)$, where the constant $c_V = \sup_x \pi^\alpha(x)$ is chosen so that $\inf_x V(x) = 1$. Then the following holds:

$$\limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, \mathrm{d}y)}{V(x)} < \limsup_{\|x\| \to \infty} \sup_{\bar{\gamma}} \int_{R_{\bar{\gamma}}(x)} q_{\theta,\bar{\gamma}}(y \mid x) \, \mathrm{d}y.$$

Lemma 3 (Nonzero acceptance probability)**.** The proposal has a nonzero probability of acceptance; equivalently,

$$\int_{R_{\bar{\gamma}}(x)} q_{\theta,\bar{\gamma}}(y \mid x) \, \mathrm{d}y < 1.$$

The following three propositions are key to the authors' proof of Theorem 2.

Proposition 1 (Diminishing adaptation of Algorithm 1)**.** Let the map parameters $\bar{\gamma}$ be restricted to a compact subset of Γ. Then the sequence of transition kernels defined by the update step in lines 9-13 of Algorithm 1 satisfies the diminishing adaptation condition.

Proof. When the MCMC chain is not at an adaption step, $\bar{\gamma}^{(k+1)} = \bar{\gamma}^{(k)}$. Thus, to show diminishing adaptation, we need to show that the difference between transition kernels at step K and $K + K_U$ decreases as $K \to \infty$. Mathematically, we require

$$\lim_{K\to\infty} \mathbb{P}\left(\sup_{x\in\mathbb{R}^n} \|P_{\bar{\gamma}^{(K)}}(x,\cdot) - P_{\bar{\gamma}^{(K+K_U)}}\|_{TV} \geq \delta_1\right) = 0$$

for any $\delta_1 > 0$. Because the maps are linear in $\bar{\gamma}$ for a fixed x, the mapping from $\bar{\gamma}$ to $P_{\bar{\gamma}}(x, A)$ is continuous for any $A \subseteq \mathcal{X}$. Combined with the fact that $q_{\theta,\bar{\gamma}}$ is bounded, it suffices to prove that

$$\lim_{K\to\infty} \mathbb{P}\left(\|\gamma_i^{(K+K_U)} - \gamma_i^{(K)}\| \geq \delta\right) = 0, \tag{7}$$

for any $\delta > 0$ and all $i \in \{1, 2, \ldots, n\}$.

To prove (7), the authors use the fact that $\bar{\gamma}^{(K)}$ is the minimizer of the objective function (5), which is based on a K-sample Monte Carlo approximation of the Kullback-Leibler divergence. With the convention $\log(0) = -\infty$, they define the objective function

$$f_i^{(k)}(\gamma_i) = \frac{1}{k} g(\gamma_i) + \frac{1}{k}\sum_{j=1}^{k}\left(\frac{1}{2}\tilde{T}_i^2(\theta^{(j)};\gamma_i) - \log\left.\frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\right|_{\theta^{(j)}}\right)$$

so that $\gamma_i^{(K)} = \arg\min_f {}_i^{(K)}(\gamma_i)$ and $\gamma_i^{(K+K_U)} = \arg\min_f {}_i^{(K+K_U)}(\gamma_i)$ for all $i \in \{1, 2, \ldots, n\}$. It follows that for all γ_i,

$$\begin{aligned}
\mathbb{P}\left(|f_i^{(K+K_U)}(\gamma_i) - f_i^{(K)}(\gamma_i)| \geq \delta_2\right) &= \mathbb{P}\left(\left|\frac{1}{K}\sum_{j=K+1}^{K+K_U}\left(\frac{1}{2}\tilde{T}_i^2(\theta^{(j)};\gamma_i) - \log\left.\frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\right|_{\theta^{(j)}}\right)\right| \geq \delta_2\right) \\
&\leq \frac{1}{K\delta_2}\mathbb{E}\left[\left|\sum_{j=K+1}^{K+K_U}\left(\frac{1}{2}\tilde{T}_i^2(\theta^{(j)};\gamma_i) - \log\left.\frac{\partial \tilde{T}_i(\theta;\gamma_i)}{\partial\theta_i}\right|_{\theta^{(j)}}\right)\right|\right] \\
&\xrightarrow{K\to\infty} 0.
\end{aligned} \tag{8}$$

Here the inequality is due to Markov's inequality; the expectation is finite because the map is bi-Lipschitz, the proposal density is bounded by Gaussian densities, and the map is linear for large $\|\theta\|$. It remains to show that this implies the convergence of $\|\gamma_i^{(K+K_U)} - \gamma_i^{(K)}\|$. To this end, let $\mathcal{C}_{\delta_2}^{(K)} = \{\gamma_i : f_i^{(K)}(\gamma_i) - \delta_2 \leq f_i^{(K)}(\gamma_i^{(K)}) + \delta_2\}$. Since $\gamma_i^{(K)} \in \mathcal{C}^{(K)}$ and $f_i^{(K)}$ is convex, as $\delta_2 \to 0$ we will have $\mathcal{C}_{\delta_2}^{(K)} \to \{\gamma_i^{(K)}\}$. Thus, for any $\delta > 0$ there exists a δ_2 such that

$$\sup_{\gamma_i,\gamma_i'\in\mathcal{C}_{\delta_2}^{(K)}} \|\gamma_i - \gamma_i'\| < \delta. \tag{9}$$

For any such $\delta_2 > 0$, (8) implies that

$$\lim_{K\to\infty} \mathbb{P}\left(\gamma_i^{(K+K_U)} \in \mathcal{C}_{\delta_2}^{(K)}\right) = 1,$$

which combined with (9) yields (7), as desired. □

Proposition 2 (Minorization condition for Algorithm 1)**.** There is a scalar δ and a set of probability measures $\nu_{\bar{\gamma}}$ defined on C such that $P_{\bar{\gamma}}(x,\cdot) \geq \delta\nu_{\bar{\gamma}}(\cdot)$ for all $x \in C$ and $\bar{\gamma} \in \Gamma$.

Proof. The proof follows Lemma 6.1 in [**Atc06**]. For $a > 0$, let g_a be the density of the d-dimensional Normal distribution with mean 0 and covariance matrix aI_d. Because the drift of the algorithm is bounded by δ and $\bar{\gamma} \in \Gamma$, we can find $\varepsilon_1 > 0$ and $k_1 > 0$ such that $\inf_{\gamma\in\Gamma} q_{\theta,\gamma}(\theta' \mid \theta) \geq k_1 g_{\varepsilon_1}(\theta' - \theta)$. We take $R > 0$ and $C = B(0, R)$, and define $\tau = \min_{\gamma\in\Gamma}\min_{\theta'-\theta,\theta\in C}\frac{\pi(\theta')q_{\theta,\gamma}(\theta'|\theta)}{\pi\theta q_{\theta,\bar{\gamma}}(\theta|\theta')} > 0$. Choosing $\varepsilon = \tau k_1$ and $\nu_{\bar{\gamma}}(A) = \frac{\int_{A\cap C} g_{\varepsilon_1}(z)\,\mathrm{d}z}{\int_C g_{\varepsilon_1}(z)\,\mathrm{d}z}$, we have

$$P_{\bar{\gamma}}(x, A) \geq \inf_{\gamma\in\Gamma} P_\gamma(x, A) \geq \varepsilon\nu_{\bar{\gamma}}(A)$$

for all $x \in C$ and $\bar{\gamma} \in \Gamma$. □

Proposition 3 (Simultaneous drift of Algorithm 1)**.** For all points $x \in \mathbb{R}^n$ and all feasible map parameters $\bar{\gamma} \in \Gamma$, there are scalars λ and b such that $\int_{\mathbb{R}^n} V(x) P_{\bar{\gamma}}(x, \mathrm{d}x) \leq \lambda V(x) + b \mathbb{1}_C(x)$.

Proof. By the proof of Lemma 6.2 in [**Atc06**], it suffices to show that

$$\limsup_{\|x\| \to \infty} \sup_{\bar{\gamma} \in \Gamma} \frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, \mathrm{d}y)}{V(x)} < 1, \tag{10}$$

and

$$\sup_{x \in \mathbb{R}^n} \sup_{\bar{\gamma} \in \Gamma} \frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, \mathrm{d}y)}{V(x)} < \infty. \tag{11}$$

Combining Lemma 2 and Lemma 3 immediately yields (10). With the choice of drift function $V(x) = c_V \pi^{-\alpha}(x)$ for $\alpha \in (0, 1)$, the authors show that

$$\frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, \mathrm{d}y)}{V(x)} \leq 1 + \int_{A_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta, \bar{\gamma}}(y \mid x) \, \mathrm{d}y + \int_{R_{\bar{\gamma}}(x)} \frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} \frac{\pi(y) q_{\theta, \bar{\gamma}}(x \mid y)}{\pi(x) q_{\theta, \bar{\gamma}}(y \mid x)} q_{\theta, \bar{\gamma}}(y \mid x) \, \mathrm{d}y. \tag{12}$$

Within the region of possible rejection $R_{\bar{\gamma}}(x)$, the Metropolis acceptance probability $\frac{\pi(y)}{\pi(x)} < 1$, so the integrand of the right-most term is bounded above by $\frac{\pi^{-\alpha}(y)}{\pi^{-\alpha}(x)} q_{\theta, \bar{\gamma}}(y \mid x)$. Two applications of the upper bound from Lemma 1 reduce (12) to

$$\begin{aligned}
\frac{\int_{\mathbb{R}^n} V(y) P_{\bar{\gamma}}(x, \mathrm{d}y)}{V(x)} &\leq 1 + k_U^2 \int_{A_{\bar{\gamma}}(x)} g_U(x - y) \, \mathrm{d}y + k_U^2 \int_{R_{\bar{\gamma}}(x)} g_U(y - x) \, \mathrm{d}y \\
&\leq 1 + \int g_U(x - y) \, \mathrm{d}y \\
&= 1 + k_U^2 \\
&< \infty.
\end{aligned}$$

Taking the supremum over all $x \in \mathbb{R}^n$ and $\bar{\gamma} \in \Gamma$ yields (11). □

Proof of Theorem 2. The transport MCMC algorithm satisfies the diminishing adaptation property by Proposition 1. It also satisfies the minorization condition and simultaneous drift conditions by Propositions 2 and 3, respectively, and its induced family of transition kernels is therefore simultaneously strongly aperiodically geometrically ergodic. Since the drift function $V(x)$ specified in Lemma 2 satisfies $\sup_{x \in C} V(x) < \infty$, we can initialize the algorithm by sampling X_0 from some distribution supported on C to ensure that $\mathbb{E}[V(X_0)] < \infty$. Hence, by Theorem 1, the transport MCMC algorithm is ergodic. □

5 Future Directions

Improving the efficiency of MCMC algorithms is a highly active field of research, due in particular to the new prominence of Bayesian problems featuring extremely large datasets, for which standard MCMC samplers are prohibitively expensive. A standard method for reducing large computation times is to use a divide-and-conquer strategy in which the computation is *parallelized*; that is, to split up the computations among a group of solvers working independently, and then to aggregate the results of the computations in some fashion.

The idea of parallelizing MCMC in the Bayesian setting was pioneered in 2014 by Scott et. al. ([**Sco+16**]), and the problem has been tackled repeatedly since then. Perhaps surprisingly, splitting up the MCMC iterations among a series of "submachines" – in which each submachine produces draws from a modified target

distribution – is not difficult at all. Rather, the difficulty lies in *recombining* the submachine draws into draws from the true target distribution. In recent years, creative statisticians have combined the theory of MCMC with other areas of statistics in pursuit of this goal, leading to parallelized MCMC-based algorithms which integrate neural networks ([**MBK19**]), decision trees ([**Wan+15**]) and Gaussian processes ([**NS+18**]), to name only a few.

The above approaches parallelize their algorithms by splitting up the dataset of observations into subsets and essentially performing standard MCMC on each machine. In contrast, in [**PM18**] the authors suggest that their transport map MCMC algorithm can be parallelized in a different way; namely, by solving independently each of the n optimization problems defined by (5) – one for each dimension of the parameter space. They apply their parallelized algorithm to three Bayesian problems: a biochemical oxygen demand model, a predator-prey system, and an inference problem based on maple sap exudation). Their results compare favourably with those obtained from several other adaptive MCMC algorithms. However, the datasets used in these examples appear to be quite small – for example, the dataset used in the first problem contains a mere 20 observations.

We can summarize the situation as such: the previously-described MCMC algorithms parallelize well with respect to big data, while transport map MCMC parallelizes well with respect to high-dimensional data. This naturally leads us to wonder whether the latter scheme can be parallelized in the big data space as well.

A first step in answering this question would be to empirically compare the transport map MCMC algorithm to other algorithms – both parallelized and otherwise – when running on a Bayesian problem with a large dataset. The authors have made their implementations of transport map MCMC available online at their MUQ ("MIT Uncertainty Quantification") Library; however, their implementation is written in C++. Therefore, for practical purposes we should first implement transport map MCMC in the `R` statistical programming language, in order to compare the method more directly with other recent techniques.

If the transport map MCMC performs at least as well as those recent parallelized MCMC schemes, this would be a "win" for transport map MCMC, because most of those other schemes – especially the ones that integrate machine learning algorithms – have to date provided almost no theoretical guarantees. On the other hand, if transport map MCMC does appear to suffer when the number of observations becomes very large, then we propose to explore ways to modify the scheme to allow us to parallelize in the traditional sense (i.e., with the dataset split across submachines) as well as in the dimensionality of the parameter space.

References

[**Mon81**] Gaspard Monge. "Mémoire sur la théorie des déblais et des remblais". In: *Histoire de l'Académie Royale des Sciences de Paris* (1781).

[**Met+53**] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[**Kan58**] Leonid Kantorovitch. "On the translocation of masses". In: *Management Science* 5.1 (1958), pp. 1–4.

[**Bre91**] Yann Brenier. "Polar factorization and monotone rearrangement of vector-valued functions". In: *Communications on pure and applied mathematics* 44.4 (1991), pp. 375–417.

[**Ros95**] Jeffrey S Rosenthal. "Convergence rates for Markov chains". In: *SIAM Review* 37.3 (1995), pp. 387–405.

[**HST01**] Heikki Haario, Eero Saksman, and Johanna Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* 7.2 (2001), pp. 223–242.

[**RR01**] Gareth O Roberts and Jeffrey S Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms". In: *Statistical Science* 16.4 (2001), pp. 351–367.

[**Atc06**] Yves F Atchade. "An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift". In: *Methodology and Computing in Applied Probability* 8.2 (2006), pp. 235–254.

[**RR07**] Gareth O Roberts and Jeffrey S Rosenthal. "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms". In: *Journal of Applied Probability* 44.2 (2007), pp. 458–475.

[**RR09**] Gareth O Roberts and Jeffrey S Rosenthal. "Examples of adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 349–367.

[**San10**] Filippo Santambrogio. "Models and applications of optimal transport in economics, traffic and urban planning". In: *arXiv preprint arXiv:1009.3857* (2010).

[**San15**] Filippo Santambrogio. "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55 (2015), pp. 58–63.

[**Wan+15**] Xiangyu Wang et al. "Parallelizing MCMC with random partition trees". In: *Advances in Neural Information Processing Systems*. 2015, pp. 451–459.

[**Sco+16**] Steven L Scott et al. "Bayes and big data: The consensus Monte Carlo algorithm". In: *International Journal of Management Science and Engineering Management* 11.2 (2016), pp. 78–88.

[**NS+18**] Christopher Nemeth, Chris Sherlock, et al. "Merging MCMC subposteriors through Gaussian-process approximations". In: *Bayesian Analysis* 13.2 (2018), pp. 507–530.

[**PM18**] Matthew D Parno and Youssef M Marzouk. "Transport Map Accelerated Markov Chain Monte Carlo". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (2018), pp. 645–682.

[**MBK19**] Diego Mesquita, Paul Blomstedt, and Samuel Kaski. "Embarrassingly parallel MCMC using deep invertible transformations". In: *arXiv preprint arXiv:1903.04556* (2019).

[**PC19**] Gabriel Peyré and Marco Cuturi. "Computational optimal transport". In: *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.