Stochastic Simulation

Autumn Semester 2024

Prof. Fabio Nobile *Assistant: Matteo Raviola*

Project - 11

Submission deadline: 16 January 2025

Sampling via measure transport

1 Introduction and background

The need to sample from complex probability distributions appears in many applications, i.e., those with a high computational cost associated with evaluating their probability density function, or presenting non-Gaussian features, multi-modality, very strong correlations, etc. This constitutes a challenge for many traditional sampling algorithms. A recent approach directed at alleviating these challenges, proposed by [1], is based on the theory of *transport maps*. Consider a reference measure μ_{ref}, such that obtaining samples from μ_{ref} is simple, and a target distribution μ_{target}, which exhibits some of the aforementioned complexities. In this case, μ_{ref} can be, e.g, a standard Gaussian measure. The idea behind transport maps is to construct a transformation T such that we can (easily) generate samples from μ_{target}, by first generating samples from μ_{ref} and then transforming them into samples from μ_{target} using T. Constructing such a map T that exactly transforms μ_{ref} into μ_{target} is often out of reach. Yet, the idea can be used to construct proposal distributions within a Metropolis Hastings algorithm

1.1 Construction of the map

Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra on $\mathbb{R}^n$, $\mu_{\text{ref}}, \mu_{\text{target}} : \mathcal{B}(\mathbb{R}^n) \to [0,1]$ be probability measures and $T : \mathbb{R}^n \to \mathbb{R}^n$ an invertible map. We say that a map T pushes forward μ_{ref} to μ_{target} if $\mu_{\text{target}}(A) = \mu_{\text{ref}}(T^{-1}(A))$ for any set $A \in \mathcal{B}(\mathbb{R}^n)$, which can be written compactly as

$$T_\sharp \mu_{\text{ref}} = \mu_{\text{target}}. \tag{1}$$

We shall refer to $T_\sharp \mu_{ref}$ as the push-forward distribution of μ_{ref}. Another way to characterize the push-forward measure $T_\sharp \mu_{ref}$ is to say that if $X \sim \mu_{ref}$ is a random variable with distribution μ_{ref}, then $Z = T(X)$ has distribution $T_\sharp \mu_{ref}$. Fig. 1 gives a pictorial representation, showing an i.i.d. sample $X^{(1)}, ..., X^{(M)} \overset{\text{i.i.d}}{\sim} \mu_{ref}$ and the transformed sample $Z^{(i)} = T(X^{(i)}), i = 1, ..., M$.

If the distributions μ_{ref} and μ_{target} admit corresponding densities η and π with respect to the Lebesgue measure, we can re-write (1) as $T_\sharp \eta = \pi$, which corresponds to:

$$\pi = \eta \circ T^{-1} \left| \det \nabla T^{-1} \right|, \tag{2}$$

where ∇T^{-1} denotes the Jacobian of the inverse of the map T. The transport map T satisfying (1) can be seen as a deterministic coupling between μ_{ref} and μ_{target}.

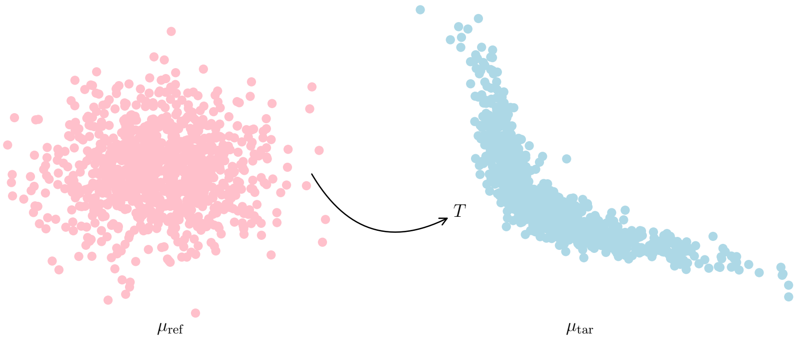

Figure 1: Example of a mapping between μ_{ref} and μ_{target} in $\mathbb{R}^2$ through the transport map T.

The *crux* of the method lies then in constructing such map T. In general, there can be infinitely many such transformations T, and the study of how to find transport maps that are "optimal" in some given sense has been actively developed in recent years, both from a theoretical and a computational perspective. For simplicity, however, we will consider the following family of *parametric triangular maps* $T_{\boldsymbol{\alpha}_d} : \mathbb{R}^n \mapsto \mathbb{R}^n$

$$T_{\boldsymbol{\alpha}_d} = T(x; \boldsymbol{\alpha}_d) = \begin{bmatrix} T^1(x_1; \boldsymbol{\alpha}_d) \\ T^2(x_1, x_2; \boldsymbol{\alpha}_d) \\ \vdots \\ T^n(x_1, x_2, \dots, x_n; \boldsymbol{\alpha}_d) \end{bmatrix}, \tag{3}$$

that depends on the (unknown) parameters $\boldsymbol{\alpha}_d$, as follows:

$$T^1(x_1; \boldsymbol{\alpha}_{\boldsymbol{d}}) = \alpha_{1,0} + \int_0^{x_1} \exp\left(\sum_{i=0}^{d} \alpha_{1,i} w^i\right) \mathrm{d}w \tag{4}$$

$$T^2(x_1, x_2; \boldsymbol{\alpha}_{\boldsymbol{d}}) = \left(\sum_{i=0}^{d} \alpha_{2,i} x_1^i\right) + \int_0^{x_2} \exp\left(\sum_{0 \le i_1 + i_2 \le d} \alpha_{2,i_1 i_2} x_1^{i_1} w^{i_2}\right) \mathrm{d}w \tag{5}$$

$$\vdots$$

$$\begin{aligned} T^k(x_1, \dots, x_k; \boldsymbol{\alpha}_{\boldsymbol{d}}) = &\left(\sum_{0 \le i_1 + i_2 + \dots + i_{k-1} \le d} \alpha_{k,i_1 i_2 \dots i_{k-1}} x_1^{i_1} x_2^{i_2} \dots x_{k-1}^{i_{k-1}}\right) \\ &+ \int_0^{x_k} \exp\left(\sum_{0 \le i_1 + i_2 + \dots + i_k \le d} \alpha_{k,i_1 i_2 \dots i_k} x_1^{i_1} x_2^{i_2} \dots x_{k-1}^{i_{k-1}} w^{i_k}\right) \mathrm{d}w, \end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}_d = \left\{a_{1,0}, \{a_{1,i}\}_{i=0}^d, \dots, \{a_{k,i_1 \dots i_k}\}_{|\boldsymbol{i}_k| \le d}\right\}$, with $|\boldsymbol{i}_k| := i_1 + i_2 + \dots + i_k$, is a set of unknown coefficients. Notice that T^k is the k-th component of the map which only depends on the first k variables. The transport map is then constructed by finding $\boldsymbol{\alpha}_d$ such that if $X \sim \mu_{\text{ref}}$, the distribution of $Z = T(X, \boldsymbol{\alpha}_d)$ closely resembles μ_{target}. To do so, we first introduce a measure of "distance" between (equivalent) probability measures.

Definition: *Let P and Q be equivalent probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, having densities p, q with respect to the Lebesgue measure. We define the* Kullback-Leibler *(KL) divergence of Q from P as*

$$\mathcal{D}_{KL}(P \parallel Q) := \mathbb{E}_p\left[\log\left(\frac{p}{q}\right)\right] = \int_{\mathbb{R}^n} p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x. \tag{7}$$

Notice that the KL divergence is not, in general, symmetric, and as such it is not a proper distance. Moreover, $\mathcal{D}_{KL}(P \parallel Q) \geq 0$, with equality only when $p = q$, P-almost everywhere. Given an i.i.d. sample $X^{(1)}, ..., X^{(M)} \sim p$, the KL divergence can be approximated as

$$\mathcal{D}_{\text{KL}}(P \parallel Q) \approx \mathcal{D}^M_{\text{KL}}(P \parallel Q) := \frac{1}{M}\sum_{i=1}^{M}\log\left[\frac{p(X^{(i)})}{q(X^{(i)})}\right], \quad X^{(i)} \overset{\text{i.i.d}}{\sim} p. \tag{8}$$

Having defined a notion of divergence between probability measures, we can cast the construction of the transport map as the following unconstrained minimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\alpha_d}\in\mathbb{R}^d} \mathcal{D}^M_{\text{KL}}(T_{\boldsymbol{\alpha}_d\sharp}\eta \parallel \pi) \\ =&\min_{\boldsymbol{\alpha_d}\in\mathbb{R}^d} \mathcal{D}^M_{\text{KL}}(\eta \parallel T^{-1}_{\boldsymbol{\alpha}_d\sharp}\pi) \\ =&\min_{\boldsymbol{\alpha_d}\in\mathbb{R}^d} \frac{1}{M}\sum_{i=1}^{M}\Big[-\log(T_{\boldsymbol{\alpha}_d}X^{(i)}) - \log|\det\nabla T_{\boldsymbol{\alpha}_d}(X^{(i)})|\Big], \quad X^{(i)} \overset{\text{i.i.d.}}{\sim} p \end{aligned} \tag{9}$$

Notice that it is enough to know the target density π up to a multiplicative constant as this does not change the minimizer in (9). Once such a transformation has been found, we can use the transport map to construct a proposal distribution for MCMC. We illustrate such an approach on the following simple Bayesian inverse problem (BIP).

1.2 Bayesian inference for a biochemical oxygen demand problem

We consider a Bayesian inference problem involving a model of biochemical oxygen demand (BOD) commonly used in water quality monitoring. Biochemical oxygen demand is the amount of dissolved oxygen needed (i.e. demanded) by aerobic biological organisms to break down organic material present in a given water sample at certain temperature over a specific time period. A simplified continuous time model for the BOD is given by

$$B(t;x) := a(x_1)(1 - e^{-b(x_2)t}),\ (x_1, x_2) =: x, \tag{10}$$

$$a(x_1) := \left[0.4 + 0.4\left(1 + \text{erf}\left(\frac{x_1}{\sqrt{2}}\right)\right)\right], \tag{11}$$

$$b(x_2) := \left[0.01 + 0.15\left(1 + \text{erf}\left(\frac{x_2}{\sqrt{2}}\right)\right)\right], \tag{12}$$

where t represents time, $x \in \mathbb{R}^2$ is an unknown random parameter, and $\operatorname{erf}(x)$ is the so-called error function[1], given by $\operatorname{erf}(x) := \frac{1}{\sqrt{2}} \int_{-x}^{x} e^{-t^2} \mathrm{d}t$. Suppose an array of data $y \in \mathbb{R}^5$ is collected at 5 different times $t_1 = 1, t_2 = 2, \dots, t_5 = 5$, where each measurement $y_i \in \mathbb{R}$ is assumed to be polluted by an additive Gaussian noise $\epsilon_i \in \mathbb{R}$, i.e., we assume that each measurement is modeled by

$$y_i = B(t_i; x) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, 2, \dots, 5.$$

Our goal is to characterize the distribution of the set of random parameters x conditioned on the measured data y, denoted by $\pi(x|y)$, usually called *posterior distribution*. Assuming that x is independent of ϵ, it follows from Bayes' theorem and the assumption of additive Gaussian noise that

$$\pi^y(x) := \underbrace{\pi(x|y)}_{\text{posterior}} \propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{5}(y_i - B(t_i; x))^2\right)}_{\text{likelihood}} \underbrace{\eta(x)}_{\text{prior}}, \tag{13}$$

where the likelihood measures the misfit between the observed data y and $\{B(t_i; x)\}_{i=1}^5$, and the prior models the randomness of x before y is observed. Notice that the posterior distribution is only known up to a normalization constant. One way of characterizing $\pi(x|y)$ is to sample from it. This can, in turn be done using Markov chain Monte Carlo. In this project, we will construct an approximate transport map from the prior distribution η to the posterior distribution π^y, which is then used to construct a proposal distribution within a MCMC algorithm. In particular we will assume that $\epsilon_i \sim \mathcal{N}(0, 10^{-3}),\ i = 1, \dots, 5,\ \eta = \mathcal{N}(0, I_{2\times2})$, where $I_{2\times2}$ is the identity matrix in $\mathbb{R}^2$ and that the recorded, noise-polluted data is given by:

$$y = [0.18, 0.32, 0.42, 0.49, 0.54]. \tag{14}$$

2 Goals of the project

1. Implement a random walk Metropolis (RWM) algorithm to sample from (13). Plot the (estimated) density π^y from the obtained chain, as well as the usual MCMC diagnostics, such as traceplots and autocorrelation functions. Report the effective sample size and the acceptance rate of your chain. The obtained chain will be the "reference " sample from π^y. Use your sample to estimate $\mathbb{E}_{\pi^y}[x_1]$ and $\mathbb{E}_{\pi^y}[x_2]$.

2. Show that the map $T_{\boldsymbol{\alpha}_d}$ is invertible for any choice of paramters $\boldsymbol{\alpha}_d$. Also, prove the equalities in Eq. (9).

3. Construct an approximate transport map from the prior η to the posterior distribution π^y by solving the optimization problem (9), using different polynomial degree $d = 1, 2, 3, 4$. You can replace the Monte Carlo appoximation in (9) by another quadrature formula, if you prefer. Plot the resulting KL divergence as a function of d.

   Hint: You can use the `scipy.optimize` package to perform the numerical minimization.

[1] You can evaluate the error function in Python using the Scipy function `scipy.special.erf`

4. Once such map has been constructed, we can use it to improve MCMC in one of the following ways:

   a) In the independent Sampler (IS) algorithm. Recall that in this case, the proposal distribution $q(X^{(i)}, X^*)$ does not depend on the current state of the chain $X^{(i)}$, i.e., $q(X^{(i)}, X^*) = q(X^*)$, and is here taken as $T_{\boldsymbol{\alpha}_d \sharp}\eta$. This version of MH is attractive from a computational point of view provided that the proposal distribution closely resembles the target distribution π^y, which will be hopefully the case if the transport map is sufficiently accurate.

      Implement a transport map-based independent sampler (TMIS) algorithm to obtain a sample size of $N = 25000$ from π^y, using maps of different polynomial degree. Compare your results to those obtained in Point 1.

   b) An alternative implementation is to combine a RWM with TMIS. The rationale behind this is to guarantee convergence to the target distribution π^y, in case that the constructed map is not very accurate. Thus, at each iteration of the MH algorithm, we do a step of RWM with probability γ, or a step of TMIS with probability $1-\gamma$, for some $\gamma \in (0,1)$, usually chosen *a priori*. Implement this approach to obtain a sample size of $N = 25000$ from π^y and compare your results with those from Point 1 and 4a.

5. Alternative to building the “forward” map T that transforms the prior into an approximation of the posterior, one could construct the “inverse” map $S = T^{-1}$ that transforms the posterior into an approximation of the prior. Samples from the posterior can be obtained from a preliminary MCMC. The inverse map S can be used to build a proposal for a Metropolis Hastings (MH) algorithm as follows: given the current state $X^{(i)}$ of the chain,

   - Compute $\hat{X}^{(i)} = S(X^{(i)})$ ($\hat{X}^{(i)}$ will have a distribution close to the prior).
   - Generate $\hat{Y}^{(i)}$ from a proposal kernel $Q(\hat{X}^{(i)}, \cdot)$
   - Generate the candidate state $Y^{(i)} = S^{-1}(\hat{Y}^{(i)})$.

   Write the MH acceptance probability for the candidate $Y^{(i)}$. Consider a Gaussian proposal kernel $Q(\hat{X}^{(i)}, \cdot) = \mathcal{N}(\hat{X}^{(i)}, \sigma^2 I)$. Implement this version of the MH algorithm and compare its performance with that of the algorithms of point 4.

6. The interesting feature of the algorithm in point 5 which uses the inverse transport map is that one could construct an adaptive version of it which, when new samples from the posterior become available, these can be used to improve the inverse map S. Explore possible adaptive versions of the algorithm in point 5.

References

[1] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via measure transport: An introduction. In *Handbook of Uncertainty Quantification*, pages 1–41. Springer, 2016.

[2] Matthew D Parno and Youssef M Marzouk. Transport map accelerated Markov Chain Monte Carlo. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):645–682, 2018.