Sampling via Measure Transport

Raffo Luca
387774 - luca.raffo@epfl.ch

Stochastic Simulation
MATH-414
Department of Mathematics

Ecole polytechnique fédérale de Lausanne
Department of Mathematics

Autumn Semester 2024

1 Sampling via Measure Transport

In many applications we encounter the need to sample from complex probability distributions, i.e., those with a high computational cost associated with evaluating their density function, or presenting multi-modality, very strong correlations, or many others.

A recent approach directed at alleviating this challenges is based upon the transport of measures. In this framework we start from a reference measure μ_{ref} that is easy to sample from and a target measure μ_{target} that presents some of the aforementioned issues; and we construct a transformation map T that transforms μ_{ref} into μ_{target}. In theory, to generate samples from μ_{target} we can just sample points from μ_{ref} and transform them using T.

Constructing a map T that *exactly* transforms μ_{ref} into μ_{target} is often out of reach, so we restrict ourselves to finding a good approximation of the map to improve other known sampling algorithm.

1.1 Push Forward of Measures

We start by recalling some fundamentals about measure theory and proving some theoretical results that will be useful for our approach. Throughout the survey, μ_{ref} and μ_{target} will be probability measures on $\mathcal{B}(\mathbb{R}^n)$.

Definition 1.1. We say that a map $T : \mathbb{R}^n \to \mathbb{R}^n$ *pushes forward* μ_{ref} to μ_{target} if

$$\mu_{target}(A) = \mu_{ref}(T^{-1}(A)) \text{ for any set } A \in \mathcal{B}(\mathbb{R}^n).$$

We then write $T_{\#}\mu_{ref} = \mu_{target}$.

In order to be able to actually use these maps we need an operative characterization.

Theorem 1.1. $T_{\#}\mu_{ref} = \mu_{target}$ if and only if $\int_{\mathbb{R}^n} \phi \, d\mu_{target} = \int_{\mathbb{R}^n} \phi \circ T \, d\mu_{ref}$ for all ϕ Borel measurable, non-negative and bounded.

Proof. Clearly, by denoting the indicator function of the set A as χ_A, we can see that the first condition is equivalent to

$$\int_{\mathbb{R}^n} \chi_A \, d\mu_{target} = \int_{\mathbb{R}^n} \chi_{T^{-1}(A)} \, d\mu_{\text{ref}} = \int_{\mathbb{R}^n} \chi_A \circ T \, d\mu_{ref}. \tag{$\bullet$}$$

Clearly, the second condition implies $\bullet$ by taking $\phi = \chi_A$. Moreover, $\bullet$ implies the second condition by monotone convergence: we just need to approximate ϕ from below by simple functions. □

Now, as an application of this result, we show that this characterization takes a nice form when applied to measures that are absolutely continuous with respect to the Lebesgue measure.

Theorem 1.2. If the probability measures μ_{ref} and μ_{target} admit corresponding densities f and g with respect to the Lebesgue measure, then

$$T_{\#}\mu_{\text{ref}} = \mu_{\text{target}} \text{ if and only if } g = f \circ T^{-1}|\det \nabla T^{-1}|.$$

Proof. The proof follows from a simple change of variables. Let us denote $\mu_{\text{ref}} = f(x)dx$ and $\mu_{\text{target}} = g(y)dy$, then the push forward condition is equivalent to

$$\int_{\mathbb{R}^n} \phi \circ T(x) f(x)\, dx = \int_{\mathbb{R}^n} \phi(y) g(y)\, dy, \tag{τ}$$

and if we make a change of variables with $y = T(x)$ on the right hand side, then τ becomes

$$\int_{\mathbb{R}^n} \phi \circ T(x) f(x)\, dx = \int_{\mathbb{R}^n} \phi \circ T(x) g(T(x)) |\det \nabla T(x)|\, dx,$$

which is equivalent to

$$f(x) = g(T(x))|\det \nabla T(x)|,$$

which, by using the inverse function theorem can be rewritten as

$$f(T^{-1}(y)) \left|\det(\nabla T^{-1}(y))\right| = g(y).$$

□

1.2 KL Divergence

As we anticipated, it is very difficult in general to find a map T that perfectly transports the measure. In order to find a transformation that can provide a good approximation $T\#\mu_{\text{ref}} \approx \mu_{\text{target}}$ we need to define a dissimilarity measure in the space of probability measures.

Definition 1.2. Let μ and ν be probability measures having densities f and g with respect to the Lebesgue measure. We define the *Kullback-Leibler divergence* of μ and ν as

$$\mathcal{D}_{KL}(\mu\|\nu) := \mathbb{E}_\mu \left[\log\left(\frac{f}{g}\right)\right] = \int_{\mathbb{R}^n} f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

Clearly, the definition is not symmetrical and hence is not a proper distance. Anyway we can still use it as a dissimilarity measure. Moreover, we notice that $\mathcal{D}_{KL}(\mu\|\nu) \geq 0$, with equality only when $f = g$, μ-almost everywhere.

Given an i.i.d. sample $X^{(1)}, ..., X^{(M)} \sim \mu$, the KL divergence can be approximated as

$$\mathcal{D}_{KL}(\mu\|\nu) \approx \mathcal{D}^M_{KL}(\mu\|\nu) := \frac{1}{M}\sum_{i=1}^{M} \log\left[\frac{f(X^{(i)})}{g(X^{(i)})}\right].$$

Having defined a notion of divergence between probability measures, we can cast the construction of the transport map as the unconstrained minimization problem $\min_T \mathcal{D}^M_{KL}(T\#\mu_{\text{ref}}\|\mu_{\text{target}})$.

Actually, this quantity has some interesting properties:

Theorem 1.3. If $X^{(i)} \overset{\text{iid}}{\sim} \mu$, and μ and ν have densities f and g with respect to the Lebesgue measure, and $g = \frac{1}{Z}\tilde{g}$ with $Z = \int_{\mathbb{R}^n} \tilde{g}(y)\, d\mathcal{L}^n$, then

$$\begin{aligned}\arg\min_{T\in\mathbb{R}^d} D_{KL}^M(T\#\mu\|\nu) &= \arg\min_{T\in\mathbb{R}^d} D_{KL}^M(\mu\|T^{-1}\#\nu) \\ &= \arg\min_{T\in\mathbb{R}^d} \frac{1}{M}\sum_{i=1}^{M}\Big[-\log(\tilde{g}(T(X^{(i)}))) - \log|\det \nabla T(X^{(i)})|\Big].\end{aligned}$$

Proof. Firstly, notice that if $X^{(i)} \overset{\text{iid}}{\sim} \mu$ then $T(X^{(i)}) \overset{\text{iid}}{\sim} T\#\mu$ by definition of push forward measure; moreover let us call f_T the density of $T\#\mu$, which takes the form given by Theorem 1.2. Then we can write:

$$\mathcal{D}_{KL}^M(T\#\mu\|\nu) = \frac{1}{M}\sum_{i=1}^{N}\log\left(\frac{f_T(Y^{(i)})}{g(Y^{(i)})}\right) = \frac{1}{M}\sum_{i=1}^{N}\log\left(\frac{f(T^{-1}(y))|\det\nabla T^{-1}(y)|)}{g(Y^{(i)})}\right)$$

Recall the inverse function theorem: $|\det\nabla T^{-1}(y)| = \frac{1}{|\det\nabla T(x)|}$, then applying a change of variable $y = T(x)$, and by denoting with $g_{T^{-1}}$ the density of $T^{-1}\#\nu$ with respect to the Lebesgue measure, we can rewrite the RHS of the last equation as:

$$\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{f(X^{(i)})}{g(T(X^{(i)}))|\det\nabla T(X^{(i)})|}\right) = \frac{1}{M}\sum_{i=1}^{N}\log\left(\frac{f(X^{(i)})}{g_{T^{-1}}(X^{(i)})}\right) = \mathcal{D}_{KL}^M(\mu\|T^{-1}\#\nu)$$

where we have used again Theorem 1.2 in the form $g_{T^{-1}}(x) = g(T(x))|\det\nabla T(x)|$. Finally, notice that

$$\begin{aligned}&\arg\min_T \mathcal{D}_{KL}^M(\mu\|T^{-1}\#\nu) = \arg\min_T \frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{f(X^{(i)})}{g(T(X^{(i)}))|\det\nabla T(X^{(i)})|}\right) \\ &= \arg\min_T \frac{1}{M}\sum_{i=1}^{M}\Big[\log(f(X^{(i)})) - \log(g(T(X^{(i)}))) - \log|\det\nabla T(X^{(i)})|\Big] \\ &= \arg\min_T \frac{1}{M}\sum_{i=1}^{M}\Big[-\log\left(\frac{\tilde{g}(T(X^{(i)}))}{Z}\right) - \log|\det\nabla T(X^{(i)})|\Big] \\ &= \arg\min_T \frac{1}{M}\sum_{i=1}^{M}\Big[-\log(\tilde{g}(T(X^{(i)}))) - \log|\det\nabla T(X^{(i)})|\Big]\end{aligned}$$

as the T does not appear in neither $\log(f(X^{(i)}))$ and $\log(Z)$ and thus we can disregard them in the optimization purposes. □

1.3 Parametric Triangular Maps

In general, there can be infinitely many transformations T that can approximate our μ_{ref} well, but for the sake of simplicity we wil consider the family of *parametric triangular maps*, as defined

in Marzouk et al. (2016). These maps are particularly useful because they are always invertible.

Theorem 1.4. The map T_{α_d} is invertible for any choice of parameters α_d.

Proof. Notice that each component $T^k_{\alpha_d}$ only depends on $x_1, ..., x_k$, and therefore $J_{T_{\alpha_d}}$ is lower triangular. We know from linear algebra that a lower triangular matrix is invertible if and only if its diagonal elements are all different from zero. But from the definition of T_{α_d} and the fundamental theorem of calculus we know that:

$$\frac{\partial T^k}{\partial x_k} = \exp\left(\sum_{0 \leq i_1 + ... + i_k \leq d} \alpha_{k,i_1,...,i_k} x_1^{i_1} ... x_k^{i_k}\right) \geq 0$$

□

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. To ease notation, from now on we will denote both the measure and their density with respect to the Lebesgue measure with the same symbol.

Our task can be summarized as follows: we are given a known prior η, whose distribution is a bivariate standard normal. We aim to sample from the posterior

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

which is known up to the normalization constant (i.e., σ^2, y_i, t_i, B are known).

We will start by implementing a standard Random Walk Metropolis, which will be our benchmark. We will then try to find the parametrized transport map that minimizes the KL divergence between our RWM posterior π^y and the transported prior $T_{\tilde{\alpha}_d}\#\eta$, and then we will try to use the learnt map to create new algorithms to sample from the posterior.

2.1 Random Walk Metropolis

Our first approach will be a standard random walk Metropolis. It's a version of the classic Metropolis-Hastings algorithm (shown in Algorithm 1), with the feature of having a proposal density which has zero mean and σ as a scaling parameter (a.k.a. step size).
In our algorithm, our proposal density will be $q(x, y) = g_\sigma(y - x)$, i.e. the proposal density is a bivariate gaussian centered at the current state, with variance (which plays the role of the step size) σ. As a consequence of this choice, the acceptance probability takes the simplified form $\alpha(x, y) = \min\left\{\frac{f(y)}{f(x)}, 1\right\}$

Algorithm 1 Metropolis-Hastings

Require: λ (initial measure), q (proposal transition density), f (target density)
  Generate $X_0 \sim \lambda$
  for $n = 0, 1, \ldots$ **do**
    Generate $Y_{n+1} \sim q(X_n, \cdot)$ ▷ proposal state
    Generate $U \sim \mathcal{U}(0, 1)$
    if $U \leq \alpha(X_n, Y_{n+1})$ **then**
      Set $X_{n+1} = Y_{n+1}$ ▷ accept proposal
    else
      Set $X_{n+1} = X_n$ ▷ reject proposal
    end if
  end for

We simulated the algorithm to produce a chain with 25000 steps for different values of the step size.

Table 1: Step size and acceptance rates.

Step Size	0.005	0.01	0.05	**0.1**	0.5	1
Acceptance Rate	0.97	0.95	0.79	**0.61**	0.16	0.07

A trade-off has been found in a step size of 0.1, which gives us an acceptance rate of 0.60972 and a bidimensional histogram showing reasonable convergence. From now on this will be our step size.
In the figure below we show the bidimensional histogram (where darker pixels denote highers density) and the bidimensional traceplot with the selected step size (the starting point is highlighted in red).

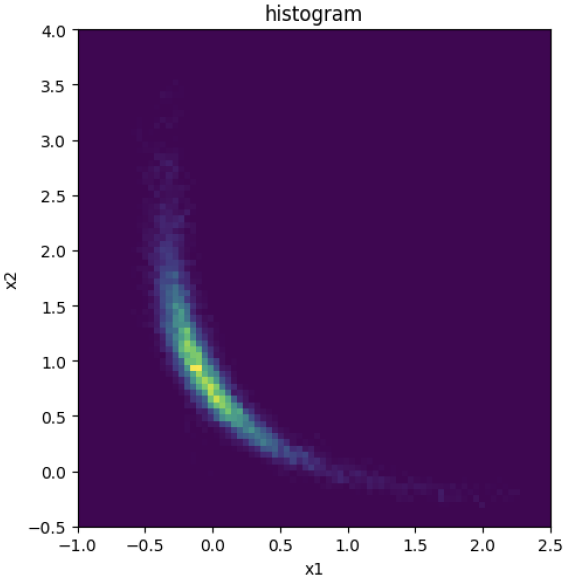

Figure 1: Bidimensional histogram.

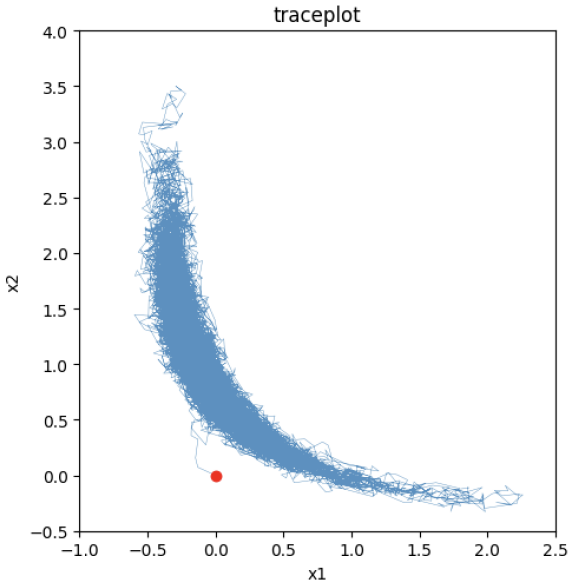

Figure 2: Bidimensional traceplot.

Within our distribution, we can compute approximations of both $\mathbb{E}_{\pi^y}(x_1)$ and $\mathbb{E}_{\pi^y}(x_2)$.
To this end, we make use of the CLT for Metropolis Hastings Markov Chains: under suitable regularity conditions (here satisfied),

$$\sqrt{N}(\mathbb{E}_{\pi^y}^{mcmc}(\phi(x)) - \mathbb{E}_{\pi^y}(\phi(x))) \xrightarrow{d} N(0, \sigma_{mcmc}^2),$$

where $\mathbb{E}_{\pi^y}^{mcmc}(\phi(x))$ denotes the empirical mean and

$$\sigma_{mcmc}^2 := \mathbb{V}ar_{\pi^y}(\phi(X_0)) + 2\sum_{l=1}^{+\infty} \mathbb{C}ov_{\pi^y}(\phi(X_0), \phi(X_l))$$

denotes the asymptotic variance. To estimate the latter with the *initial positive sequence estimator*:

$$\tilde{\sigma}_{mcmc}^2 \approx -\hat{c}_n(0) + 2\sum_{k=1}^{K} \big(\hat{c}_n(2k) + \hat{c}_n(2k+1)\big),$$

where K is the largest integer such that

$$\hat{c}_n(2k) + \hat{c}_n(2k+1) > 0 \quad \text{for all } k = 1, \dots, K,$$

and $\hat{c}_n$ denote empirical covariances.
After simulating we get

coordinate	$\mathbb{E}^{mcmc}_{\pi^y}$	Confidence Interval (95%)
x_1	0.010	$[-0.070, 0.050]$
x_2	0.956	$[0.833, 1.079]$

From now on, π^y_{mcmc} will be our benchmark, and we will try to improve it via transport map enhanced algorithms.

2.2 Finding the Optimal Map

In this subsection, we aim at finding the map $T_{\tilde{\alpha}_d}$ such that it minimizes $\mathcal{D}^M_{KL}(\pi^y \| T_{\alpha_d} \# \eta)$. In order to do so, we recall that from Theorem 1.3 that we only need to minimize the third formulation, which is easily implementable.

In the code, in order to build our parametric triangular map, we made use of the library `mpart`, built to deal with this sort of multi-index structures. The optimization code can be found in the appendix.

We show below the KL divergences for different degrees on the left, and a bivariate histogram of the transported samples from the prior with the map that attains the smallest KL divergence.

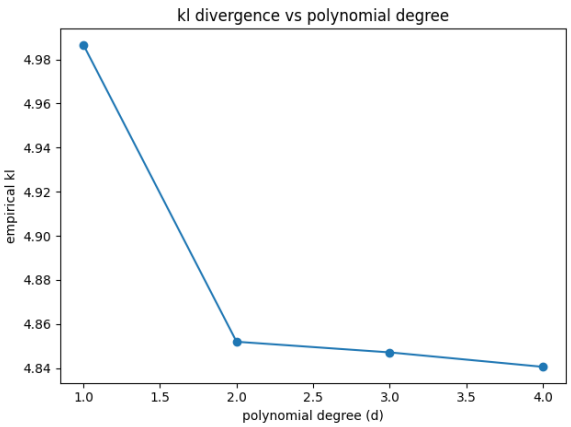

Figure 3: KL divergence vs polynomial degree.

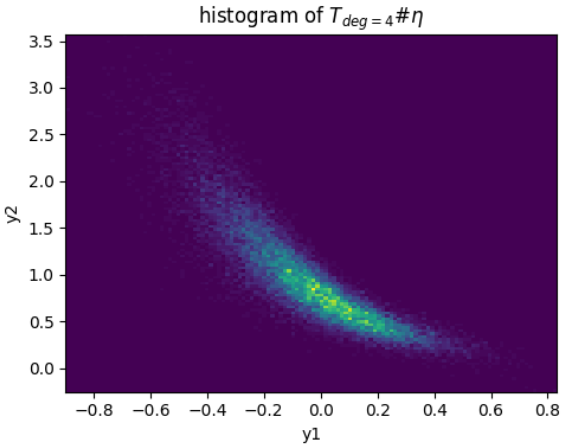

Figure 4: Bidimensional histogram of $T_{\tilde{\alpha}_4}\#\eta$.

2.3 Independence Sampler via Measure Transport

Now that we have learnt a transport map, our idea is to use it as a proposal density g in the Independence Sampler algorithm.

Algorithm 2 Independence sampler Metropolis-Hastings

Require: $X_0 \sim \lambda$, $\text{supp}(\lambda) \subset D_f$
for $n = 0, 1, \ldots$ **do**
Generate $Y_{n+1} \sim g$
Compute $\alpha(X_n, Y_{n+1}) = \min\left\{\frac{f(Y_{n+1})}{f(X_n)}\frac{g(X_n)}{g(Y_{n+1})}, 1\right\}$
Generate $U \sim \mathcal{U}(0,1)$ and set

$$X_{n+1} = \begin{cases} Y_{n+1}, & \text{if } U \leq \alpha(X_n, Y_{n+1}) \\ X_n, & \text{otherwise} \end{cases}$$

end for

And in order to have the probability density g, we apply the usual change of variable formula from Theorem 1.2.
Let the learned triangular map be

$$T_1(x_1) = e^{s_1}\, x_1 + m_1, \qquad T_2(x_1, x_2) = e^{s_2(x_1)}\, x_2 + m_2(x_1),$$

with s_2, m_2 polynomials in x_1. Since T is lower–triangular with positive diagonal, it is invertible and

$$x_1 \;=\; \frac{y_1 - m_1}{e^{s_1}}, \qquad x_2 \;=\; \frac{y_2 - m_2(x_1)}{e^{s_2(x_1)}}.$$

The Jacobian of T is lower–triangular, hence

$$\det \nabla T(x) = \frac{\partial T_1}{\partial x_1}\,\frac{\partial T_2}{\partial x_2} = e^{s_1}\, e^{s_2(x_1)}, \qquad \log\left|\det \nabla T(x)\right| = s_1 + s_2(x_1).$$

Take the reference η to be the standard normal on $\mathbb{R}^2$; its (unnormalized) log–density is $\log \tilde{\eta}(x) = -\frac{1}{2}\|x\|^2$. By the change–of–variables formula, for any $y \in \mathbb{R}^2$,

$$g(y) = \eta\big(T^{-1}(y)\big)\,\left|\det \nabla T^{-1}(y)\right| = \left.\frac{\eta(x)}{\left|\det \nabla T(x)\right|}\right|_{x = T^{-1}(y)}.$$

Therefore we can evaluate g (up to a constant) without any numerical density estimate:

$$\boxed{\log \tilde{g}(y) = \log \tilde{\eta}\big(T^{-1}(y)\big) - \log\left|\det \nabla T\right|\big|_{x=T^{-1}(y)} = -\tfrac{1}{2}\big(x_1^2 + x_2^2\big) - \big(s_1 + s_2(x_1)\big),}$$

where x_1, x_2 are the inverse coordinates given above.
Below we attach the practical recipe (used in the algorithm). Given y:

1. Compute $x_1 = (y_1 - m_1)/e^{s_1}$.

2. Evaluate $s_2(x_1)$ and $m_2(x_1)$ (polynomials in x_1).
3. Compute $x_2 = (y_2 - m_2(x_1))/e^{s_2(x_1)}$.
4. Return $\log \tilde{g}(y) = -\frac{1}{2}(x_1^2 + x_2^2) - (s_1 + s_2(x_1))$.

These steps give $\log \tilde{g}$ exactly; the normalizing constant cancels in the Metropolis–Hastings acceptance ratio.

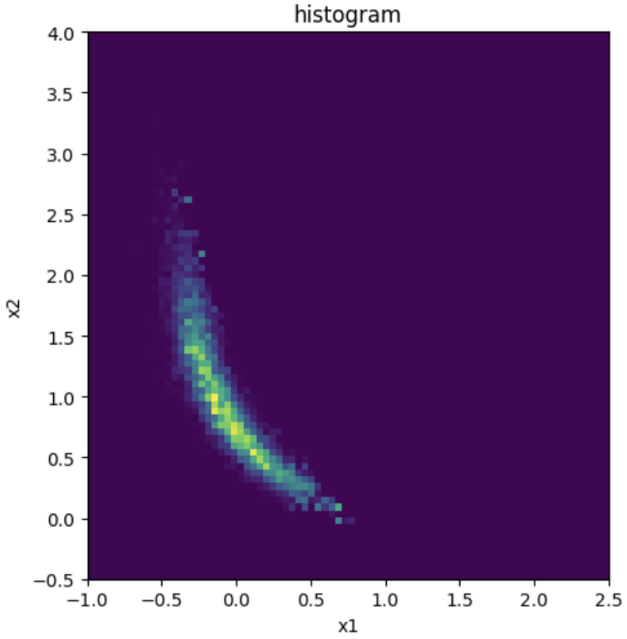

Figure 5: Indepence Sampler via Measure Transport histogram

This time the acceptance rate is 0.711. The means of the coordinates match those that we found previously.

coordinate	$\mathbb{E}_{\pi^y}^{mcmc}$	Confidence Interval (95%)
x_1	0.001	$[-0.038, 0.037]$
x_2	0.916	$[0.870, 0.963]$

2.4 RWM and Independence Sampler

It is also possible to mix the two previous approaches: we can choose a parameter $\gamma \in (0, 1)$ a priori, and then at each iteration of the MH algorithm choose to do a step of RWM with probability $\gamma = 0.75$, or a step of TMIS with probability $1 - \gamma$. We get an acceptance rate of 0.582. We compare again the usuale diagnostics, checking for similar results.

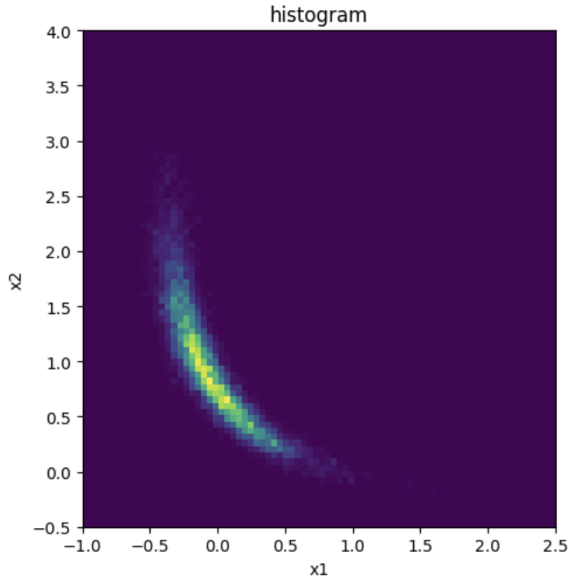

Figure 6: RWM and TMIS

coordinate	$\mathbb{E}_{\pi^y}^{mcmc}$	Confidence Interval (95%)
x_1	0.046	$[-0.067, 0.025]$
x_2	0.974	$[0.930, 1.020]$

References

Marzouk, Youssef et al. (2016). "Sampling via Measure Transport: An Introduction". In: *Handbook of Uncertainty Quantification*. Springer International Publishing, pp. 1–41. ISBN: 9783319112596. DOI: 10.1007/978-3-319-11259-6_23-1. URL: http://dx.doi.org/10.1007/978-3-319-11259-6_23-1.