Transport of Measure for Bayesian Computation

François-Xavier Briol
University of Warwick (Department of Statistics)
Imperial College London (Department of Mathematics)

Reading Group
SAMSI working group on Probabilistic Numerics

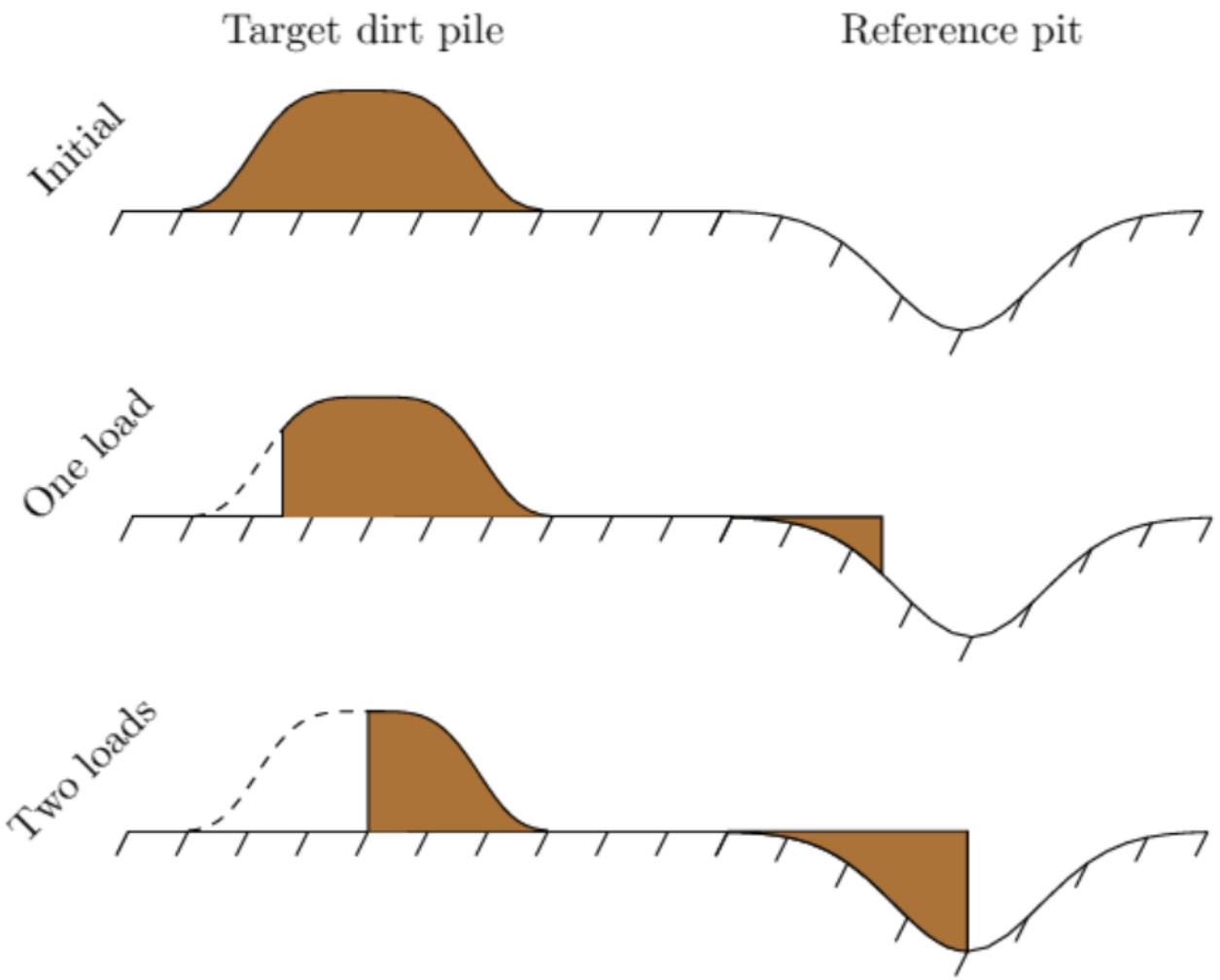
Introduction

Most of the work presented in this talk is due to Youssef Marzouk and some of his students/postdocs at MIT:

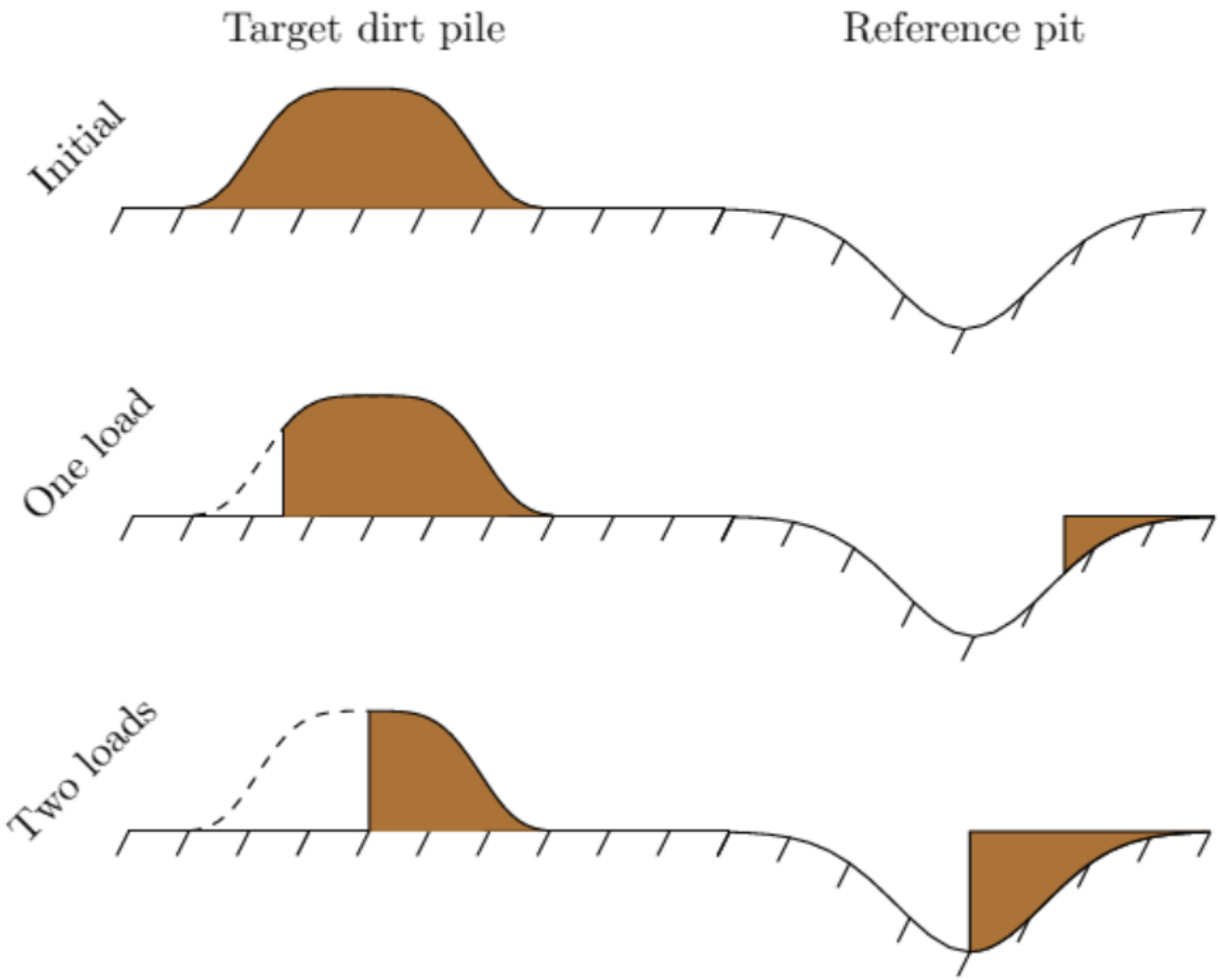
1. Marzouk, Y., Moselhy, T., Parno, M., & Spantini, A. (2016). An Introduction to Sampling via Measure Transport. In "Handbook of Uncertainty Quantification".
   - Survey paper of work in the area.
2. Parno, M. D. (2015). Transport Maps for Accelerated Bayesian Computation. PhD thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology.
   - PhD thesis by one of Marzouk's students.

(All of the figures in the slides come from these references.)

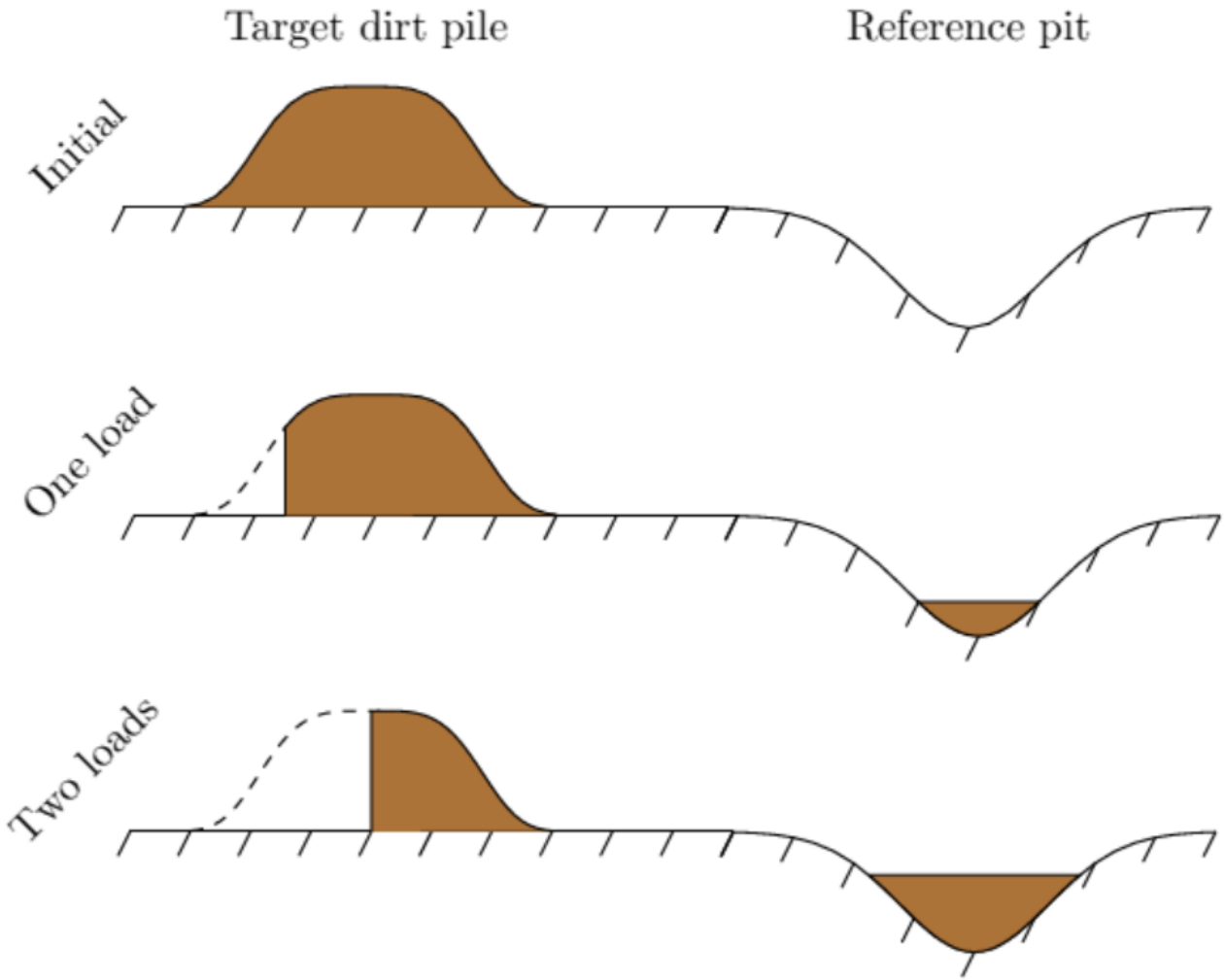
Monge's problem - Optimal transport of dirt [Parno, 2015]



Monge's problem - Optimal transport of dirt [Parno, 2015]



Monge's problem - Optimal transport of dirt [Parno, 2015]



Monge's Optimal Transport Problem

- Suppose $\mu_{\text{tar}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ is some **target** probability measure and let $\mu_{\text{ref}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ be some simple **reference** probability measure (think $\mu_{\text{tar}} =$ posterior and $\mu_{\text{ref}} =$ standard Gaussian).
- We want to find a (continuous) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mu_{\text{tar}}(A) \;=\; \mu_{\text{ref}}\big(T^{-1}(A)\big) \qquad \forall A \in \mathcal{B}(\mathbb{R}^d)$$

subject to minimising a total cost depending on some function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$C(T) := \int c\big(x, T(x)\big) \mathrm{d}\mu_{\text{ref}}(x)$$

Here T is called the **pushforward map** and we write $T_{\#}\mu_{\text{ref}} = \mu_{\text{tar}}$.

- In terms of densities, this can be written as finding the change of variable T such that:

$$\pi_{\text{ref}}(x) \;=\; J_T(x)\pi_{\text{tar}}\big(T(x)\big)$$

Monge's Optimal Transport Problem

- Suppose $\mu_{\text{tar}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ is some **target** probability measure and let $\mu_{\text{ref}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ be some simple **reference** probability measure (think $\mu_{\text{tar}} =$ posterior and $\mu_{\text{ref}} =$ standard Gaussian).
- We want to find a (continuous) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mu_{\text{tar}}(A) \;=\; \mu_{\text{ref}}\big(T^{-1}(A)\big) \qquad \forall A \in \mathcal{B}(\mathbb{R}^d)$$

subject to minimising a total cost depending on some function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$C(T) := \int c\big(x, T(x)\big) \mathrm{d}\mu_{\text{ref}}(x)$$

Here T is called the **pushforward map** and we write $T_{\#}\mu_{\text{ref}} = \mu_{\text{tar}}$.

- In terms of densities, this can be written as finding the change of variable T such that:

$$\pi_{\text{ref}}(x) \;=\; J_T(x)\pi_{\text{tar}}\big(T(x)\big)$$

Monge's Optimal Transport Problem

- Suppose $\mu_{\text{tar}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ is some **target** probability measure and let $\mu_{\text{ref}} : \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ be some simple **reference** probability measure (think $\mu_{\text{tar}} =$ posterior and $\mu_{\text{ref}} =$ standard Gaussian).
- We want to find a (continuous) transport map $T : \mathbb{R}^d \to \mathbb{R}^d$:

$$\mu_{\text{tar}}(A) \;=\; \mu_{\text{ref}}\big(T^{-1}(A)\big) \qquad \forall A \in \mathcal{B}(\mathbb{R}^d)$$

subject to minimising a total cost depending on some function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

$$C(T) := \int c\big(x, T(x)\big)\mathrm{d}\mu_{\text{ref}}(x)$$

Here T is called the **pushforward map** and we write $T_{\#}\mu_{\text{ref}} = \mu_{\text{tar}}$.

- In terms of densities, this can be written as finding the change of variable T such that:

$$\pi_{\text{ref}}(x) \;=\; J_T(x)\pi_{\text{tar}}\big(T(x)\big)$$

Kantorovich Relaxation of the Optimal Transport Problem

- We want to find a **feasible joint measure** μ with density $\pi(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that:

$$\pi_{\text{ref}}(x) = \int \pi(x, y)\mathrm{d}y \qquad \pi_{\text{tar}}(y) = \int \pi(x, y)\mathrm{d}x$$

subject to minimising the cost functional:

$$C(\pi) = \int\int c(x, y)\pi(x, y)\mathrm{d}x\mathrm{d}y$$

- This problem is a relaxation because we do not have a one-to-one mapping anymore!
- The Monge and Kantorovich formulations can be shown to have the same unique solution under mild assumptions (strict convexity of c). Kantorovich's approach is linear in π, making it amenable to linear programming.

Kantorovich Relaxation of the Optimal Transport Problem

- We want to find a **feasible joint measure** μ with density $\pi(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that:

$$\pi_{\text{ref}}(x) = \int \pi(x, y)\mathrm{d}y \qquad \pi_{\text{tar}}(y) = \int \pi(x, y)\mathrm{d}x$$

subject to minimising the cost functional:

$$C(\pi) = \int\int c(x, y)\pi(x, y)\mathrm{d}x\mathrm{d}y$$

- This problem is a relaxation because we do not have a one-to-one mapping anymore!
- The Monge and Kantorovich formulations can be shown to have the same unique solution under mild assumptions (strict convexity of c). Kantorovich's approach is linear in π, making it amenable to linear programming.

Kantorovich Relaxation of the Optimal Transport Problem

- We want to find a **feasible joint measure** μ with density $\pi(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that:

$$\pi_{\text{ref}}(x) = \int \pi(x, y)\mathrm{d}y \qquad \pi_{\text{tar}}(y) = \int \pi(x, y)\mathrm{d}x$$

subject to minimising the cost functional:

$$C(\pi) = \int\int c(x, y)\pi(x, y)\mathrm{d}x\mathrm{d}y$$

- This problem is a relaxation because we do not have a one-to-one mapping anymore!
- The Monge and Kantorovich formulations can be shown to have the same unique solution under mild assumptions (strict convexity of c). Kantorovich's approach is linear in π, making it amenable to linear programming.

Applications of the Transport Problem to Statistics

- **Density estimation:** π_{tar} is some density which needs to be estimated, and π_{ref} is some simple density. We can then estimate T and obtain π_{tar} using the change of variable formula on previous slide.
- **Importance Sampling:** π_{ref} is some density which is simple to sample from and π_{tar} is the optimal importance distribution which we want to sample from.
- **Regression:** π_{ref} is some density over features and π_{tar} is some density over labels, both of which can be sample from. The aim is to estimate the map T which transports one into the other.
- Many other problems in business, economics, operations research, medical research, etc... can also be framed this way. Kantorovich got a Nobel prize in Economics for work relating to this problem

Applications of the Transport Problem to Statistics

- **Density estimation:** π_{tar} is some density which needs to be estimated, and π_{ref} is some simple density. We can then estimate T and obtain π_{tar} using the change of variable formula on previous slide.
- **Importance Sampling:** π_{ref} is some density which is simple to sample from and π_{tar} is the optimal importance distribution which we want to sample from.
- **Regression:** π_{ref} is some density over features and π_{tar} is some density over labels, both of which can be sample from. The aim is to estimate the map T which transports one into the other.
- Many other problems in business, economics, operations research, medical research, etc... can also be framed this way. Kantorovich got a Nobel prize in Economics for work relating to this problem

Applications of the Transport Problem to Statistics

- **Density estimation:** π_{tar} is some density which needs to be estimated, and π_{ref} is some simple density. We can then estimate T and obtain π_{tar} using the change of variable formula on previous slide.
- **Importance Sampling:** π_{ref} is some density which is simple to sample from and π_{tar} is the optimal importance distribution which we want to sample from.
- **Regression:** π_{ref} is some density over features and π_{tar} is some density over labels, both of which can be sample from. The aim is to estimate the map T which transports one into the other.
- Many other problems in business, economics, operations research, medical research, etc... can also be framed this way. Kantorovich got a Nobel prize in Economics for work relating to this problem

Applications of the Transport Problem to Statistics

- **Density estimation:** π_{tar} is some density which needs to be estimated, and π_{ref} is some simple density. We can then estimate T and obtain π_{tar} using the change of variable formula on previous slide.
- **Importance Sampling:** π_{ref} is some density which is simple to sample from and π_{tar} is the optimal importance distribution which we want to sample from.
- **Regression:** π_{ref} is some density over features and π_{tar} is some density over labels, both of which can be sample from. The aim is to estimate the map T which transports one into the other.
- Many other problems in business, economics, operations research, medical research, etc... can also be framed this way. Kantorovich got a Nobel prize in Economics for work relating to this problem

Optimal Transport Problem for Bayesian Computation

In this talk, I will focus on the application of the optimal transport problem to Bayesian inference and computation.

- Suppose we knew the T which solves the transport problem between π_{ref} and π_{tar} for any cost functional c.
- Then, we could just sample $\{x_i\}_{i=1}^n$ exactly from the simple density π_{ref}, then apply the transport map T to obtain exact samples $\{T(x_i)\}_{i=1}^n$ from π_{tar}.
- This would be a very cheap way of doing exact sampling from π_{tar} and we could hence obtain a lot of such samples!

Optimal Transport Problem for Bayesian Computation

In this talk, I will focus on the application of the optimal transport problem to Bayesian inference and computation.

- Suppose we knew the T which solves the transport problem between π_{ref} and π_{tar} for any cost functional c.
- Then, we could just sample $\{x_i\}_{i=1}^n$ exactly from the simple density π_{ref}, then apply the transport map T to obtain exact samples $\{T(x_i)\}_{i=1}^n$ from π_{tar}.
- This would be a very cheap way of doing exact sampling from π_{tar} and we could hence obtain a lot of such samples!

Optimal Transport Problem for Bayesian Computation

In this talk, I will focus on the application of the optimal transport problem to Bayesian inference and computation.

- Suppose we knew the T which solves the transport problem between π_{ref} and π_{tar} for any cost functional c.
- Then, we could just sample $\{x_i\}_{i=1}^n$ exactly from the simple density π_{ref}, then apply the transport map T to obtain exact samples $\{T(x_i)\}_{i=1}^n$ from π_{tar}.
- This would be a very cheap way of doing exact sampling from π_{tar} and we could hence obtain a lot of such samples!

Optimal Transport Problem for Bayesian Computation

Important points to note:

- Finding a transport map T is quite a complicated task and we will therefore use an approximation: $\tilde{T}(x) \approx T(x)$ which is chosen from some suitable class of functions. This is clearly a variational problem, and the cost function c will also be chosen for convenience.
- Denote by $K = T^{-1}$ the **pullback map**, i.e. the map satisfying which brings back the target to the reference and is denoted $K^{\#}\mu_{\text{tar}} = \mu_{ref}$. An alternative approach could hence be to compute an approximation $\tilde{K}(x) \approx K(x)$ of the pullback.

Optimal Transport Problem for Bayesian Computation

Important points to note:

- Finding a transport map T is quite a complicated task and we will therefore use an approximation: $\tilde{T}(x) \approx T(x)$ which is chosen from some suitable class of functions. This is clearly a variational problem, and the cost function c will also be chosen for convenience.
- Denote by $K = T^{-1}$ the **pullback map**, i.e. the map satisfying which brings back the target to the reference and is denoted $K^{\#}\mu_{\text{tar}} = \mu_{ref}$. An alternative approach could hence be to compute an approximation $\tilde{K}(x) \approx K(x)$ of the pullback.

Triangular maps

Since we do not want to minimise a given cost c (i.e. we do not need a concept of optimality), we can just choose any map T which transports μ_{ref} to μ_{tar}:

- **Lower triangular map:** A map T of the form $T(x_1) = T^1(x_1)$, $T(x_2) = T^2(x_1, x_2)$, $T(x_3) = T^3(x_1, x_2, x_3)$, ...
- If the two measures are absolutely continuous, then there is a unique map of this form called the **Knothe-Rosenblatt re-arrangement** and it is optimal w.r.t the cost:

$$c(x, y) = \sum_{i=1}^{n} t^{i-1} \left| x_i - y_i \right|^2, \quad t > 0.$$

  when $t \to 0$. This is the map we will try to approximate.
- One advantage is that computing inverses of the map will be made easier.

Triangular maps

Since we do not want to minimise a given cost c (i.e. we do not need a concept of optimality), we can just choose any map T which transports μ_{ref} to μ_{tar}:

- **Lower triangular map:** A map T of the form $T(x_1) = T^1(x_1)$, $T(x_2) = T^2(x_1, x_2)$, $T(x_3) = T^3(x_1, x_2, x_3)$, ...
- If the two measures are absolutely continuous, then there is a unique map of this form called the **Knothe-Rosenblatt re-arrangement** and it is optimal w.r.t the cost:

$$c(x, y) = \sum_{i=1}^{n} t^{i-1} \left| x_i - y_i \right|^2, \quad t > 0.$$

  when $t \to 0$. This is the map we will try to approximate.
- One advantage is that computing inverses of the map will be made easier.

Triangular maps

Since we do not want to minimise a given cost c (i.e. we do not need a concept of optimality), we can just choose any map T which transports μ_{ref} to μ_{tar}:

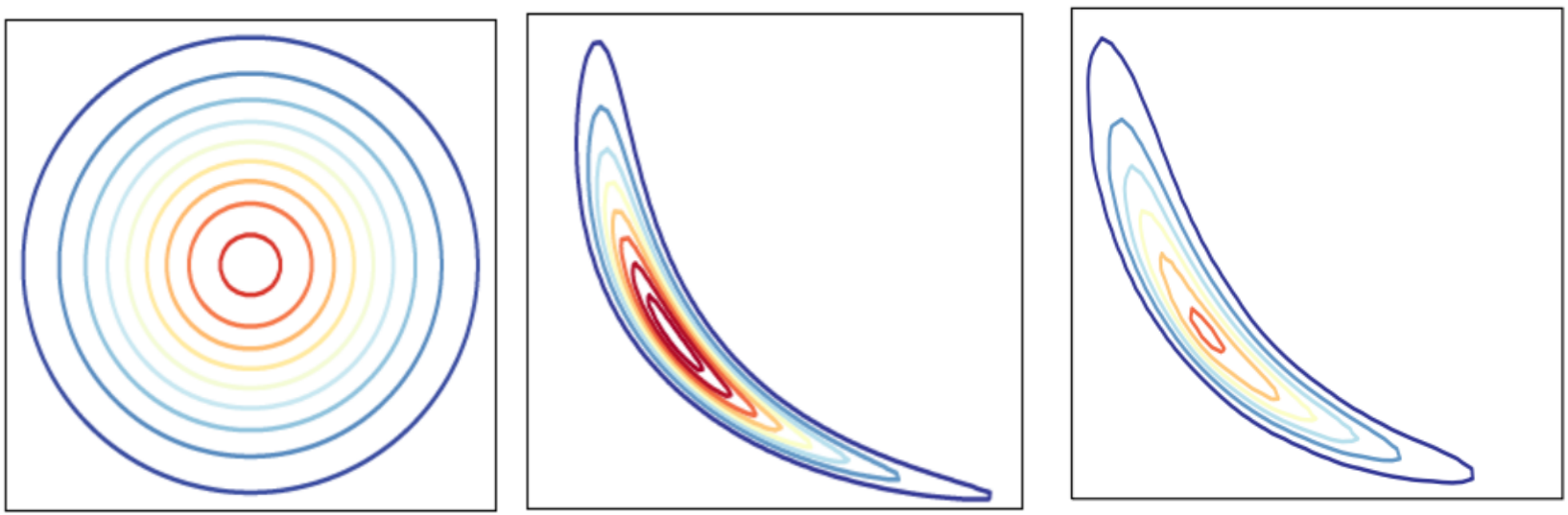
- **Lower triangular map:** A map T of the form $T(x_1) = T^1(x_1)$, $T(x_2) = T^2(x_1, x_2)$, $T(x_3) = T^3(x_1, x_2, x_3)$, ...
- If the two measures are absolutely continuous, then there is a unique map of this form called the **Knothe-Rosenblatt re-arrangement** and it is optimal w.r.t the cost:

$$c(x, y) = \sum_{i=1}^{n} t^{i-1} \left| x_i - y_i \right|^2, \quad t > 0.$$

  when $t \to 0$. This is the map we will try to approximate.
- One advantage is that computing inverses of the map will be made easier.

Example: Approximating the Pullback and Pushforward

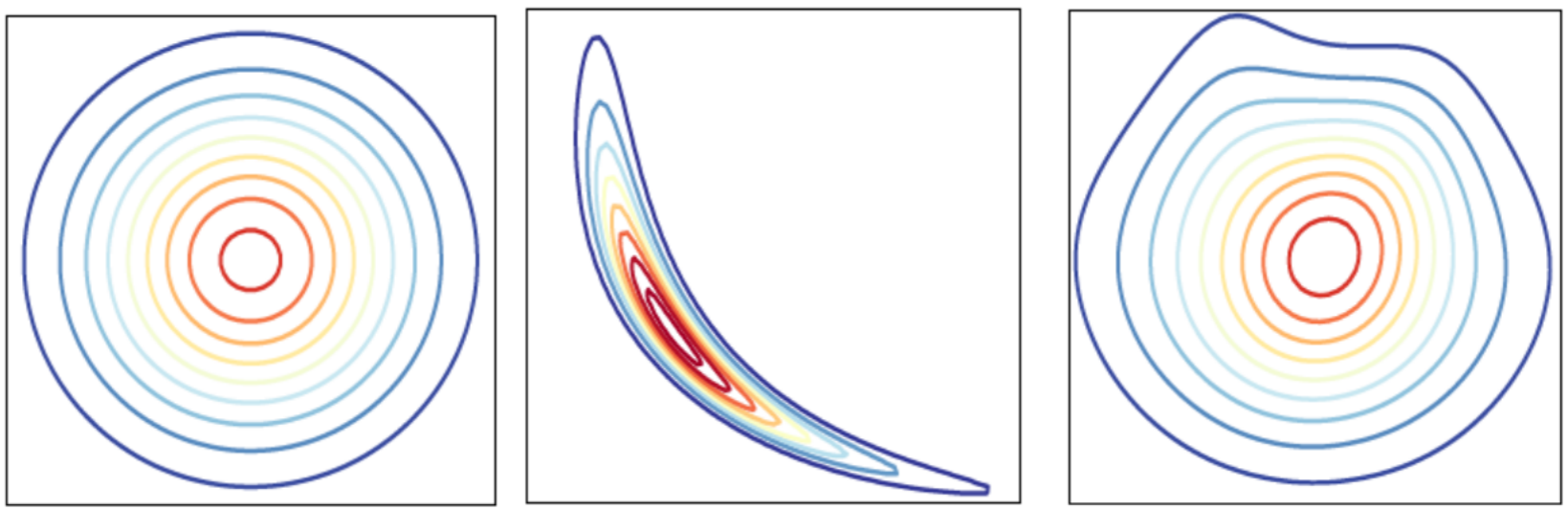
Approach 1: Approximate pushforward map $\tilde{T}$:



(a) Reference density (b) Target density (c) degree $p = 5$

Example: Approximating the Pullback and Pushforward

Approach 2: Approximate pullback map $\tilde{K}$:



(a) Reference density (b) Target density (c) Pullback $p = 5$

Two scenarios in Bayesian Computation

In this talk, we will look at two approaches for Bayesian computation based on transport maps. This will each correspond to a different scenario:

1. The posterior target density π_{tar} can be computed but only up to a normalising constant. This is a common scenario where MCMC is commonly used.

2. The posterior target density π_{tar} cannot be computed, but it is possible to obtain samples $\{x_i\}_{i=1}^n$ from it. This is a common scenario where ABC methods are commonly used.

Both of these will replace the difficult task sampling, with the task of approximating a transport map. The argument here is that optimization is easier to do (especially in high dim) and is better understood theoretically (for example better convergence diagnostics).

Scenario 1: Density known up to normalising constant

- Any global minimiser of the following optimization problem is a valid transport map:

$$\begin{aligned} \min \; & D_{\mathrm{KL}}\left(T_{\#}\pi_{\mathrm{ref}} \,||\, \pi_{\mathrm{tar}}\right) \\ \text{s.t. } & \nabla T \succ 0,\, T \in \mathcal{T} \end{aligned}$$

  Here $D_{KL}(\pi_1||\pi_2) = \mathbb{E}_{\pi_1}(\log \pi_1/\pi_2)$ is the KL-divergence and $\mathcal{T}$ is an appropriate set of diffeomorphisms (smooth invertable fuction between manifolds with smooth inverse)

- The constraint $\nabla T \succ 0$ implies that all eigenvalues are real and positive. It ensures that the pushforward is stricly positive, that the KL evaluates to finite values, and does not exclude any transport map.

- The Knothe-Rosenblath rearrangement is the unique minimiser of this problem when T is lower triangular (i.e. $T \in \mathcal{T}_{\triangle}$). For tractability, we will need to consider an nested increasing sequence $\{\mathcal{T}_{\triangle}^{h}\}_h$.

Scenario 1: Density known up to normalising constant

- Any global minimiser of the following optimization problem is a valid transport map:

$$\begin{aligned} \min \; & D_{\mathsf{KL}}\left(T_{\#}\pi_{\mathsf{ref}} \,||\, \pi_{\mathsf{tar}}\right) \\ \text{s.t. } & \nabla T \succ 0, T \in \mathcal{T} \end{aligned}$$

  Here $D_{KL}(\pi_1||\pi_2) = \mathbb{E}_{\pi_1}(\log \pi_1/\pi_2)$ is the KL-divergence and $\mathcal{T}$ is an appropriate set of diffeomorphisms (smooth invertable fuction between manifolds with smooth inverse)

- The constraint $\nabla T \succ 0$ implies that all eigenvalues are real and positive. It ensures that the pushforward is stricly positive, that the KL evaluates to finite values, and does not exclude any transport map.

- The Knothe-Rosenblath rearrangement is the unique minimiser of this problem when T is lower triangular (i.e. $T \in \mathcal{T}_{\triangle}$). For tractability, we will need to consider an nested increasing sequence $\{\mathcal{T}_{\triangle}^{h}\}_h$.

Scenario 1: Density known up to normalising constant

- Any global minimiser of the following optimization problem is a valid transport map:

$$\begin{aligned} \min \; & D_{\mathrm{KL}}\left(T_{\#}\pi_{\mathrm{ref}} \,||\, \pi_{\mathrm{tar}}\right) \\ \text{s.t. } & \nabla T \succ 0,\, T \in \mathcal{T} \end{aligned}$$

  Here $D_{KL}(\pi_1||\pi_2) = \mathbb{E}_{\pi_1}(\log \pi_1/\pi_2)$ is the KL-divergence and $\mathcal{T}$ is an appropriate set of diffeomorphisms (smooth invertable fuction between manifolds with smooth inverse)

- The constraint $\nabla T \succ 0$ implies that all eigenvalues are real and positive. It ensures that the pushforward is stricly positive, that the KL evaluates to finite values, and does not exclude any transport map.

- The Knothe-Rosenblath rearrangement is the unique minimiser of this problem when T is lower triangular (i.e. $T \in \mathcal{T}_{\triangle}$). For tractability, we will need to consider an nested increasing sequence $\{\mathcal{T}_{\triangle}^{h}\}_h$.

Scenario 1: Density known up to normalising constant

- Denote by $\tilde{\pi}_{\text{tar}}$ the un-normalised version of the target density. Then, the KL-divergence can be written as:

$$D_{KL}\left(T_{\#}\pi_{\text{ref}} \,||\, \pi_{\text{tar}}\right) \;=\; \mathbb{E}_{\pi_{\text{ref}}}[-\log \tilde{\pi}_{\text{tar}} \circ T - \log\det \nabla T] + \mathfrak{C}$$

  where $\mathfrak{C}$ is a term independent of our optimization problem.

- The integral can be approximated using any standard quadrature method (Gaussian quadrature, MC, QMC,..., even BQ since we are in a PN reading group). μ_{ref} will usually be chosen to make this quadrature easy.

- We note that the KL-divergence will usually be non-convex, unless the target density is log-concave (e.g. log-Gaussian Cox processes).

Scenario 1: Density known up to normalising constant

- Denote by $\tilde{\pi}_{\text{tar}}$ the un-normalised version of the target density. Then, the KL-divergence can be written as:

$$D_{KL}\left(T_{\#}\pi_{\text{ref}} \,||\, \pi_{\text{tar}}\right) \;=\; \mathbb{E}_{\pi_{\text{ref}}}[-\log \tilde{\pi}_{\text{tar}} \circ T - \log\det \nabla T] + \mathfrak{C}$$

  where $\mathfrak{C}$ is a term independent of our optimization problem.

- The integral can be approximated using any standard quadrature method (Gaussian quadrature, MC, QMC,..., even BQ since we are in a PN reading group). μ_{ref} will usually be chosen to make this quadrature easy.

- We note that the KL-divergence will usually be non-convex, unless the target density is log-concave (e.g. log-Gaussian Cox processes).

Scenario 1: Density known up to normalising constant

- Denote by $\tilde{\pi}_{\text{tar}}$ the un-normalised version of the target density. Then, the KL-divergence can be written as:

$$D_{KL}\left(T_{\#}\pi_{\text{ref}} \,||\, \pi_{\text{tar}}\right) \;=\; \mathbb{E}_{\pi_{\text{ref}}}[-\log \tilde{\pi}_{\text{tar}} \circ T - \log\det \nabla T] + \mathfrak{C}$$

  where $\mathfrak{C}$ is a term independent of our optimization problem.

- The integral can be approximated using any standard quadrature method (Gaussian quadrature, MC, QMC,..., even BQ since we are in a PN reading group). μ_{ref} will usually be chosen to make this quadrature easy.

- We note that the KL-divergence will usually be non-convex, unless the target density is log-concave (e.g. log-Gaussian Cox processes).

Scenario 1: Density known up to normalising constant

The final variational problem is given by:

$$\min \ \frac{1}{n}\sum_{i=1}^{n}\Big(-\log \tilde{\pi}_{\text{tar}}(T(x_i)) - \sum_{k=1}^{d}\log \partial_k T^k(x_i)\Big)$$
$$\text{s.t. } \partial_k T^k > 0 \forall k, \quad T \in \mathcal{T}_{\triangle}^{h} \subset \mathcal{T}_{\triangle}, \quad \{x_i\}_{i=1}^{n} \sim \pi_{\text{ref}}$$

- We can approximate the error from the approximation of T as follows:

$$D_{\text{KL}}(T_{\#}\pi_{\text{ref}}||\pi_{\text{tar}}) \approx \frac{1}{2}\mathbb{V}ar_{\pi_{\text{ref}}}[\log \pi_{\text{ref}} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}]$$

- Furthermore, an estimate of normalising constant of the target density.

$$\beta = \frac{\tilde{\pi}_{\text{tar}}}{\pi_{\text{tar}}} = \exp \mathbb{E}_{\pi_{\text{ref}}}\Big[\log \pi_{ref} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}\Big]$$

Scenario 1: Density known up to normalising constant

The final variational problem is given by:

$$\min \ \frac{1}{n}\sum_{i=1}^{n}\Big(-\log\tilde{\pi}_{\text{tar}}(T(x_i)) - \sum_{k=1}^{d}\log\partial_k T^k(x_i)\Big)$$
$$\text{s.t. } \partial_k T^k > 0 \forall k, \quad T \in \mathcal{T}_{\triangle}^{h} \subset \mathcal{T}_{\triangle}, \quad \{x_i\}_{i=1}^{n} \sim \pi_{\text{ref}}$$

- We can approximate the error from the approximation of T as follows:

$$D_{\text{KL}}(T_{\#}\pi_{\text{ref}}||\pi_{\text{tar}}) \ \approx \ \frac{1}{2}\mathbb{V}ar_{\pi_{\text{ref}}}[\log\pi_{\text{ref}} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}]$$

- Furthermore, an estimate of normalising constant of the target density.

$$\beta \ = \ \frac{\tilde{\pi}_{\text{tar}}}{\pi_{\text{tar}}} \ = \ \exp\mathbb{E}_{\pi_{\text{ref}}}\Big[\log\pi_{ref} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}\Big]$$

Scenario 1: Density known up to normalising constant

The final variational problem is given by:

$$\min \ \frac{1}{n}\sum_{i=1}^{n}\Big(-\log \tilde{\pi}_{\text{tar}}(T(x_i)) - \sum_{k=1}^{d}\log \partial_k T^k(x_i)\Big)$$
$$\text{s.t. } \partial_k T^k > 0 \forall k, \quad T \in \mathcal{T}_{\triangle}^{h} \subset \mathcal{T}_{\triangle}, \quad \{x_i\}_{i=1}^{n} \sim \pi_{\text{ref}}$$

- We can approximate the error from the approximation of T as follows:

$$D_{\text{KL}}(T_{\#}\pi_{\text{ref}}||\pi_{\text{tar}}) \ \approx \ \frac{1}{2}\mathbb{V}ar_{\pi_{\text{ref}}}[\log \pi_{\text{ref}} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}]$$

- Furthermore, an estimate of normalising constant of the target density.

$$\beta \ = \ \frac{\tilde{\pi}_{\text{tar}}}{\pi_{\text{tar}}} \ = \ \exp \mathbb{E}_{\pi_{\text{ref}}}\Big[\log \pi_{ref} - \log T_{\#}^{-1}\tilde{\pi}_{\text{tar}}\Big]$$

Scenario 1: Density known up to normalising constant

Generally, we are interested in computing $\mathbb{E}_{\pi_{\text{tar}}}[g]$ for some g. We now make several important points concerning the estimators obtained:

- Our transport map $\tilde{T}$ is often approximate (i.e. $\tilde{T} \neq T$) and we can only sample from the approx. $\hat{\pi}_{\text{tar}} = \tilde{T}_{\#}\pi_{\text{ref}}$. The bias satisfies:

$$\left\|\mathbb{E}_{\pi_{\text{tar}}}[g] - \mathbb{E}_{\hat{\pi}_{\text{tar}}}[g]\right\| \leq C(g, \pi_{\text{tar}}, \hat{\pi}_{\text{tar}})\sqrt{D_{\text{KL}}(\hat{\pi}_{\text{tar}} \mid\mid \pi_{\text{tar}})}$$

  where $C(g, \pi_{\text{tar}}, \hat{\pi}_{\text{tar}}) = \sqrt{2}(\mathbb{E}_{\pi_{\text{tar}}}[\|g\|^2] + \mathbb{E}_{\hat{\pi}_{\text{tar}}}[\|g\|^2])^{1/2}$

- Minimising the KL-divergence corresponds to getting rid of bias.
- The MSE will usually be dominated by the bias, since the variance occurs from approximating the expectation with samples from the reference measure, which can be produced very cheaply.
- We could instead sample from the pullback density: $\tilde{T}^{\#} = \tilde{T}_{\#}^{-1} = \pi \circ \tilde{T} |\det \nabla \tilde{T}|$ via MCMC then map them through $\tilde{T}_{\#}$.

Summary

- We have set a variational problem whose (approximate) solution can be used as a transport map and only requires evaluating $\tilde{\pi}_{\text{tar}}$. Here, once we have a map, sampling from an approximate target will be very cheap but this is not an exact method in general.
- We are sometimes interested in sampling exactly from the target, or in scenarios where the target cannot be evaluated. We will now go through scenario 2, which only requires having access to samples from the target.
- We haven't discussed how to choose $\{\mathcal{T}_{\triangle}^{h}\}_{h}$. We will first discuss scenario 2, then get back to this problem for both scenarios.
- Questions?

Summary

- We have set a variational problem whose (approximate) solution can be used as a transport map and only requires evaluating $\tilde{\pi}_{\text{tar}}$. Here, once we have a map, sampling from an approximate target will be very cheap but this is not an exact method in general.
- We are sometimes interested in sampling exactly from the target, or in scenarios where the target cannot be evaluated. We will now go through scenario 2, which only requires having access to samples from the target.
- We haven't discussed how to choose $\{\mathcal{T}_{\triangle}^{h}\}_{h}$. We will first discuss scenario 2, then get back to this problem for both scenarios.
- Questions?

Scenario 2: Density known through samples

- We consider this time the inverse transport map S which satisfies $\mu_{\text{ref}} = \mu_{\text{tar}} \circ S^{-1}$.
- Our variational problem is now given by:

$$\begin{aligned} \min \; & D_{\text{KL}}\left(S_{\#}\pi_{\text{tar}} \,||\, \pi_{\text{ref}}\right) \\ \text{s.t. } & \nabla S \succ 0, S \in \mathcal{T}_{\triangle} \end{aligned}$$

- ... and can again be simplified to:

$$\begin{aligned} \min \; & \frac{1}{n}\sum_{i=1}^{n} -\log \pi_{\text{ref}}(S(z_i)) - \log\det \nabla S(z_i) \\ \text{s.t. } & \partial_k S^k > 0, S \in \mathcal{T}_{\triangle} \end{aligned}$$

Scenario 2: Density known through samples

- We consider this time the inverse transport map S which satisfies $\mu_{\text{ref}} = \mu_{\text{tar}} \circ S^{-1}$.
- Our variational problem is now given by:

$$\begin{aligned} \min \; & D_{\text{KL}}\left(S_{\#}\pi_{\text{tar}} \,||\, \pi_{\text{ref}}\right) \\ \text{s.t. } & \nabla S \succ 0, S \in \mathcal{T}_{\triangle} \end{aligned}$$

- ... and can again be simplified to:

$$\min \; \frac{1}{n}\sum_{i=1}^{n} -\log \pi_{\text{ref}}(S(z_i)) - \log\det \nabla S(z_i)$$

$$\text{s.t. } \partial_k S^k > 0, S \in \mathcal{T}_{\triangle}$$

Scenario 2: Density known through samples

- We consider this time the inverse transport map S which satisfies $\mu_{\text{ref}} = \mu_{\text{tar}} \circ S^{-1}$.
- Our variational problem is now given by:

$$\begin{aligned} \min \; & D_{\text{KL}}\left(S_{\#}\pi_{\text{tar}} \,||\, \pi_{\text{ref}}\right) \\ \text{s.t. } & \nabla S \succ 0, S \in \mathcal{T}_{\triangle} \end{aligned}$$

- ... and can again be simplified to:

$$\begin{aligned} \min \; & \frac{1}{n}\sum_{i=1}^{n} -\log \pi_{\text{ref}}(S(z_i)) - \log\det \nabla S(z_i) \\ \text{s.t. } & \partial_k S^k > 0, S \in \mathcal{T}_{\triangle} \end{aligned}$$

Scenario 2: Density known through samples

There are several notable differences with Scenario 1:

- The problem is convex if π_{ref} is log-concave, which is easily enforced by choosing a convenient reference measure. This can therefore be satisfied for any target!
- The objective function does not require the derivatives of the log-target density anymore!
- If the reference density is a product of marginals, then the optimization problem is fully parallelisable. For example, if it is a standard Gaussian then:

$$\min \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} \frac{1}{2} (S^k)^2(z_i) - \log \partial_k S^k(z_i)$$

$$\text{s.t. } \partial_k S^k > 0, S \in \mathcal{T}_{\triangle}$$

Scenario 2: Density known through samples

There are several notable differences with Scenario 1:

- The problem is convex if π_{ref} is log-concave, which is easily enforced by choosing a convenient reference measure. This can therefore be satisfied for any target!
- The objective function does not require the derivatives of the log-target density anymore!
- If the reference density is a product of marginals, then the optimization problem is fully parallelisable. For example, if it is a standard Gaussian then:

$$\min \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} \frac{1}{2} (S^k)^2(z_i) - \log \partial_k S^k(z_i)$$

$$\text{s.t. } \partial_k S^k > 0, S \in \mathcal{T}_{\triangle}$$

Scenario 2: Density known through samples

There are several notable differences with Scenario 1:

- The problem is convex if π_{ref} is log-concave, which is easily enforced by choosing a convenient reference measure. This can therefore be satisfied for any target!
- The objective function does not require the derivatives of the log-target density anymore!
- If the reference density is a product of marginals, then the optimization problem is fully parallelisable. For example, if it is a standard Gaussian then:

$$\min \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{d} \frac{1}{2} (S^k)^2(z_i) - \log \partial_k S^k(z_i)$$

$$\text{s.t. } \partial_k S^k > 0, S \in \mathcal{T}_\triangle$$

Parameterization of Approximate Maps

We want to pick an expressive family of parameterised map which satisfy the monotonicity constraints $\partial_k T^k > 0$ or $\partial_k S^k > 0$.

1. **Option 1:** We can choose a popular family of maps, such as a polynomial or radial basis function expansions, then enforce local monotonicity at a bunch of samples.

2. **Option 2:** We can consider a family which directly encodes these constraints. For example:

$$T^k(x_1, \ldots, x_k) = a_k(x_1, \ldots, x_{k-1}) + \int_0^{x_k} \exp(b_k(x_1, \ldots, x_{k-1}, w))\mathrm{d}w$$

for some parameterised functions $a_k : \mathbb{R}^{k-1} \to \mathbb{R}$ and $b_k : \mathbb{R}^k \to \mathbb{R}$.

Related work

1. El Moselhy, T. A., & Marzouk, Y. M. (2012). Bayesian Inference with Optimal Maps. Journal of Computational Physics, 231(23), 78157850.
2. Heng, J., Doucet, A., & Pokern, Y. (2015). Gibbs Flow for Approximate Transport with Applications to Bayesian Computation. arXiv:1509.08787.
3. Bernton, E., Jacob, P. E., Gerber, M., & Robert, C. P. (2017). Inference in generative models using the Wasserstein distance. arXiv:1701.05146.