A minimax formulation to the Wasserstein Optimal Transport problem

Alexandre Milewski, Vaish Gajaraj [1]

NYU Courant Institute Of Mathematical Sciences

[1]Special thanks to Prof. Esteban Tabak, Prof. Debra Laefer

Introduction

- The original Optimal Transport problem was proposed by Gaspard Monge
- Monge was motivated by logistics: given a pile of sand, and a hole of the same volume; Find the most efficient transport map to fill the hole.

Monge Problem

Problem (Monge)

Given normalised mass distributions P and Q, we seek a transport map T satisfying

$$\int_A Q(x)\,dx = \int_{T^{-1}(A)} P(x)\,dx \text{ for any Borel subset } A \subset \mathbf{R}^d \quad (1)$$

which minimizes the quantity $\int \phi(x, T(x))P(x)\,dx$ for a given cost function ϕ

The condition (1) is called the push-forward condition, denoted $T_{\#}P = Q$. It is equivalent to the statement $X \sim P \Rightarrow T(X) \sim Q$

Optimal Transport in Change Detection

- Optimal transport can be useful in data science. In this case, we view P and Q are considered to be probability distributions

Optimal Transport in Change Detection

- Optimal transport can be useful in data science. In this case, we view P and Q are considered to be probability distributions
- By representing the initial and final states of a system by probability distributions, we can seek an optimal map, which may inform us of the behaviour of the system.

Optimal Transport in Change Detection

- Optimal transport can be useful in data science. In this case, we view P and Q are considered to be probability distributions
- By representing the initial and final states of a system by probability distributions, we can seek an optimal map, which may inform us of the behaviour of the system.
- An example of this may be change detection, in which we treat an initial and final image as samples of a distribution. From these samples we can infer an optimal map between them, which can be used to inform where change has occurred.

Optimal Transport in Change Detection

- Optimal transport can be useful in data science. In this case, we view P and Q are considered to be probability distributions
- By representing the initial and final states of a system by probability distributions, we can seek an optimal map, which may inform us of the behaviour of the system.
- An example of this may be change detection, in which we treat an initial and final image as samples of a distribution. From these samples we can infer an optimal map between them, which can be used to inform where change has occurred.
- Here the word image, can be viewed in a more abstract sense, such as 3-D point clouds. In particular LiDAR data is used to measure the topography of a surface.

Wasserstein Problem

In this talk we consider the cost function $\phi(x, y) = \|x - y\|^2$.

Problem (Wasserstein)

Find a mapping T, such that

$$T = \arg\min_{F} \left\{ \int \|x - F(x)\|^2 \, dP \;\middle|\; F_{\#}P = Q \right\}$$

Strategy

- This problem is not suitable for computation; the push-froward condition is hard to enforce, given that it must hold for every subset of the domain.
- Instead we reformulate this problem into a minimax problem, maximizing over Borel functions g and minimizing over all convex functions ψ
- From this perspective, we can view g as an indicator, highlighting the disparities between $F_{\#}P$ and Q.
- ψ is modified to negate the discrepancies highlighted by g.

A crucial theorem

We begin with existence and uniqueness of the Wasserstein problem, which are guaranteed by Brenier's Theorem.

Brenier's Theorem

Assume P and Q are in $\mathbf{R}^n$ and don't give mass to n-1 dimensional sets. Then there exists a unique solution of the form $T = \nabla\varphi$, where φ is convex.

A crucial theorem

We begin with existence and uniqueness of the Wasserstein problem, which are guaranteed by Brenier's Theorem.

Brenier's Theorem

Assume P and Q are in $\mathbf{R}^n$ and don't give mass to n-1 dimensional sets. Then there exists a unique solution of the form $T = \nabla\varphi$, where φ is convex.

- It also happens that the push-forward condition admits a unique solution of the form $\nabla\varphi$ where φ is convex.

A crucial theorem

We begin with existence and uniqueness of the Wasserstein problem, which are guaranteed by Brenier's Theorem.

Brenier's Theorem

Assume P and Q are in $\mathbf{R}^n$ and don't give mass to n-1 dimensional sets. Then there exists a unique solution of the form $T = \nabla\varphi$, where φ is convex.

- It also happens that the push-forward condition admits a unique solution of the form $\nabla\varphi$ where φ is convex.
- This implies that the solution to the Wasserstein problem is equivalent to finding such a φ.

Kullback-Leibler Divergence

In order to measure the difference between $F_{\#}P$ and Q, we use the Kullback-Leibler Divergence, which measures relative entropy.

Definition (Kullback-Leibler Divergence)

$$D_{\mathrm{KL}}(P\|Q) := \int \ln\left(\frac{dP}{dQ}\right)\, dP \geq 0$$

Since $D(P\|Q) = 0$ if and only if $P = Q$, it follows that

$$\varphi = \arg\min_{\psi} \{D_{\mathrm{KL}}(\psi_{\#}P\|Q) \mid \psi \text{ is convex}\}$$

Variational Formulation

The Kullback-Leibler divergence can be rewritten as a variational problem,

$$D_{\mathrm{KL}}(P\|Q) = 1 + \max_{g} \left\{ \int g(x)\, dP - \int e^{g(x)}\, dQ \right\}$$

Wherever $P \neq Q$, g will stray from 0, informing where the P can be modified to match Q.

Minimax

If $X \sim P$ and $Y \sim Q$, we can rewrite the integrals as expectations

$$D_{KL}(P\|Q) = 1 + \max_g \left\{ \mathbb{E}[g(X)] - \mathbb{E}[e^{g(Y)}] \right\}$$

Since $X \sim P$ implies $\nabla\psi(X) \sim \nabla\psi_{\#}P$,

$$D_{KL}(\nabla\psi_{\#}P\|Q) = 1 + \max_g \left\{ \mathbb{E}[g(\nabla\psi(X))] - \mathbb{E}[e^{g(Y)}] \right\}$$

Combining this with our earlier result, we have

$$\varphi = \arg\min_{\phi} \left\{ \max_g \left\{ \mathbb{E}[g(\nabla\psi(X))] - \mathbb{E}[e^{g(Y)}] \right\} \;\middle|\; \psi \text{ is convex} \right\}$$

Sample Based problem

This method can be very easily converted to accommodate optimal transport on distributions known only through the samples. Suppose $x_1, x_2, \dots x_n$ is a sample from X and $y_1, y_2, \dots y_m$ a sample from Y. Replacing expectation by empirical means yields

$$\varphi = \arg\min_{\phi} \left\{ \max_{g} \left\{ \frac{1}{n} \sum_i [g(\nabla\psi(x_i))] - \frac{1}{m} \sum_j [e^{g(y_j)}] \right\} \middle| \ \psi \text{ is convex} \right\}$$

A Weighted Formulation

$$\min_{\phi} \max_{g} L[\psi, g] \approx \arg\min_{\psi} \left\{ \max_{g} \left\{ \sum_{i} w_i^x g(\nabla\psi(x_i)) - \sum_{j} w_j^y e^{g(y_j)} \right\} \right\} \tag{2}$$

A direct approach could be used to solve this adversarial problem but would not be efficient:

- ψ and g must be rich enough
- Enforcing convexity is difficult

Solving Local Problems

$$T_{composed} = T_N \circ T_{N-1} \circ ... \circ T_1 \tag{3}$$

- Where each transported map $T_n = \nabla\psi$ as given by solving our minimax formulation.
- To make mapping between complex distributions easier, we solve our MINIMAX problem for many local maps and compose them to get a combined map that approximates our target.

A Step Back

We have now derived a formulation for local transport that can translate the "Push-Forward" condition into a data science problem. But what is the final problem we are seeking to solve?

What we seek to find between these local maps

Wasserstein Distance

The 2-wasserstein distance is defined as the minimum of the distance squared cost between two sets of points

$$W_2(x_{n-1}, x_n) = \arg\min_F \left\{ \int \|x_{n-1} - F(x_{n-1})\|^2 \, dx \right\}$$

One might think of these as finding the minimum distance between two nearby sets of points x_n and x_{n-1}.

A Global Minimum

While locally we can find Wasserstein Distances, we do not guarantee immediately that their composition is also a minimum of the distance-squared cost from our initial distribution of points to our final one.

A global minimum of local minimums

Minimize the total cost of the sum of each locally solved T

$$\min_{x_0,\ldots,x_N} \sum_{n=1}^{N} w^{x_n} W_2(x_{n-1}, x_n)$$

Assume we solved our problem

Assume we found the optimal $T_{composed}$ that can minimize our cost function from before. Such a map would move along the geodesic from the original set of points to the final set of points. Any intermediate map is therefore given by the following:

McCann's Interpolants

$$T_k \circ ... \circ T_2 \circ T_1(x_1) = (\frac{N-k}{k})x_1 + \frac{k}{N} T_{composed}(x_1)$$

Where N are our total number of intermediary points, k is the iteration we would like to look at, and x_1 is our original collection of points.

The map that satisfies this formula also minimizes the global sum of Wasserstein costs, as proved in [Kuang, Tabak 2017]

Approach to get this optimal map

To do so, we iterate a global algorithm of optimal transport. Iterating on this process N times allows us to converge onto a composed map that gives a minimum global cost. Such a map will also best preserve the original structure of the initial data.

Initial State

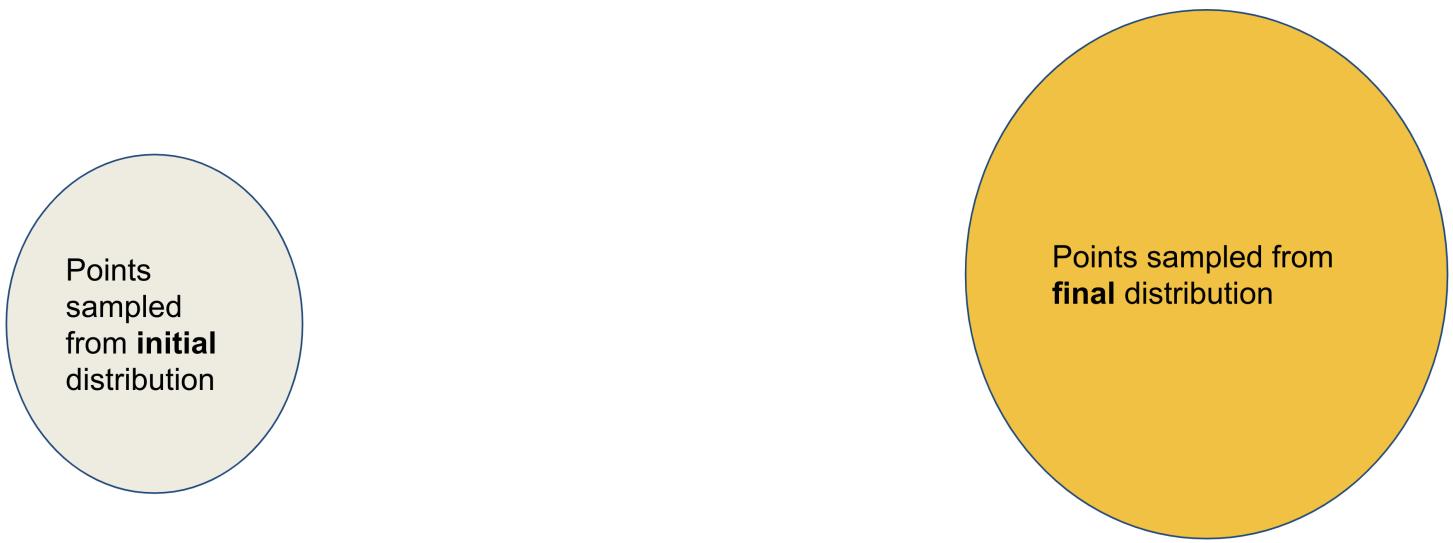

Figure: Starting distributions x and y of different spread

Assume they are both Gaussian and apply closed form solution.

Closed Form Solution

Closed Form Solution Between 2 Gaussians

$$x \Rightarrow \bar{y} + A(x - \bar{x}) \tag{4}$$

$$A = \mathrm{Cov}[x]^{-\frac{1}{2}}(\mathrm{Cov}[x]^{\frac{1}{2}}\mathrm{Cov}[y]\mathrm{Cov}[x]^{\frac{1}{2}})^{\frac{1}{2}}\mathrm{Cov}[x]^{-\frac{1}{2}} \tag{5}$$

Where x is the initial set of points and y is the final set of points

Preconditioning

Figure: Both set of points have been mapped onto each other

For each cloud we create, we first mix the initial or final data with its transported image, and then take a percentage of each of these two mixtures to make an interpolated cloud.

Mixed Interpolants

We pick N=3 intermediary clouds to create.

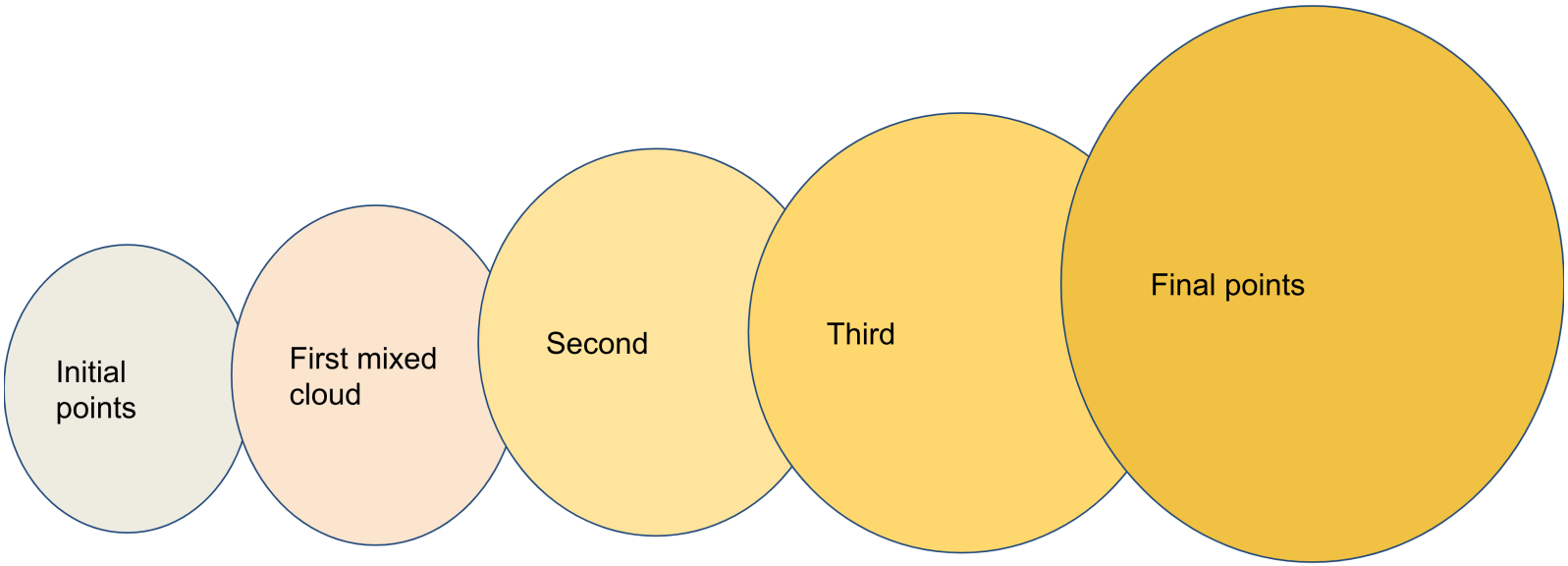

Figure: Our mixed interpolation creates many intermediary clouds on which to preform Local Transport

Local Problem

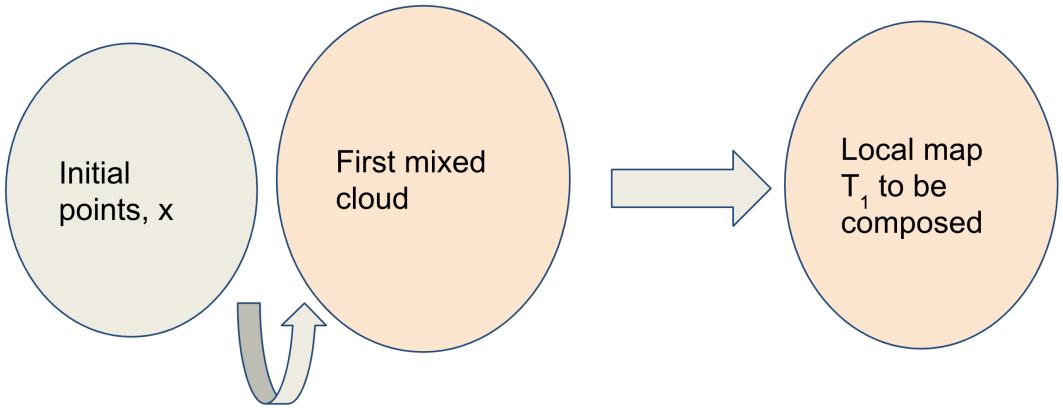

Figure: Between nearby points, we can solve local transport

Result of First Global Iteration

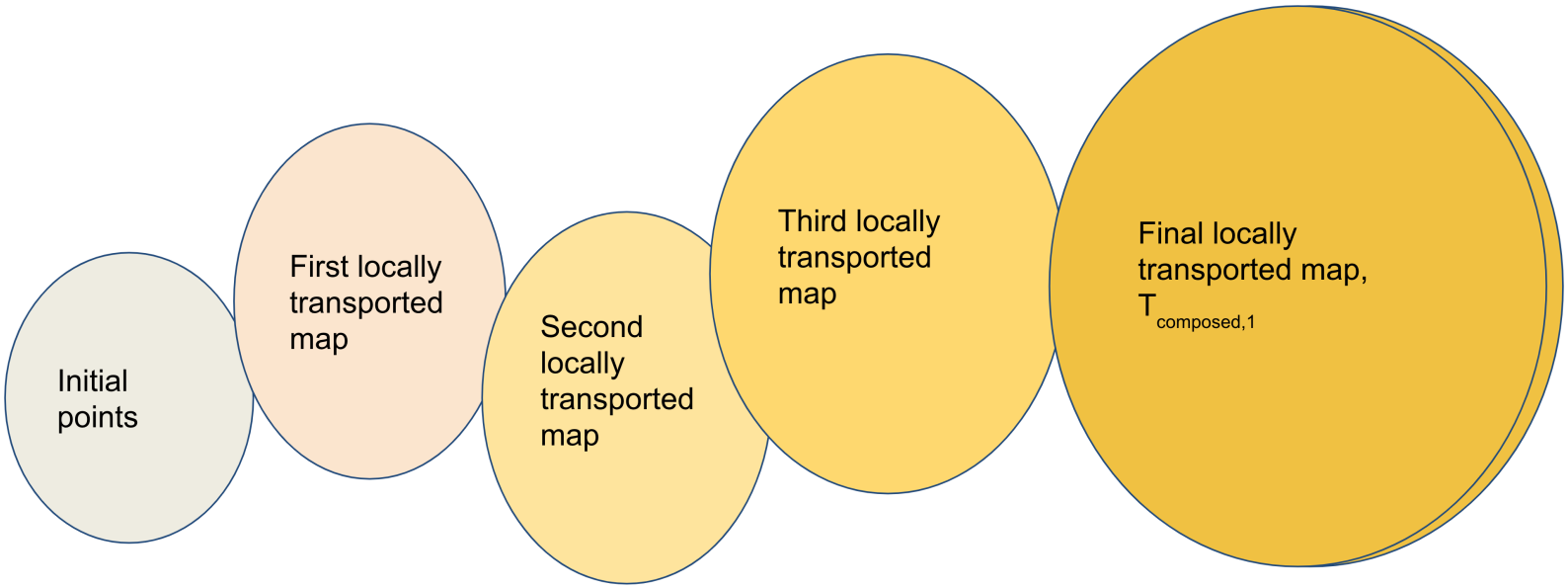

Figure: Our first global iteration gets close to our final distribution, yet its not optimal, therefore it might have some unnecessary perturbations from the original map

Beginning of n^{th} Global Iteration

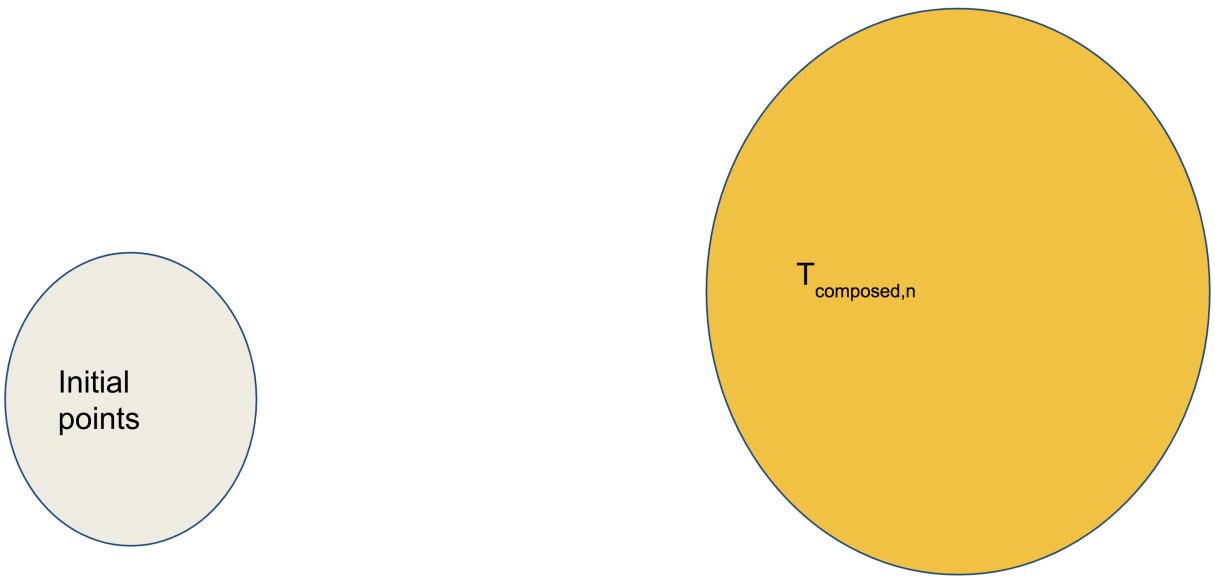

Figure: We start the next iteration using the transported map from last time as the new starting distribution

Our Solved $T_{composed}$

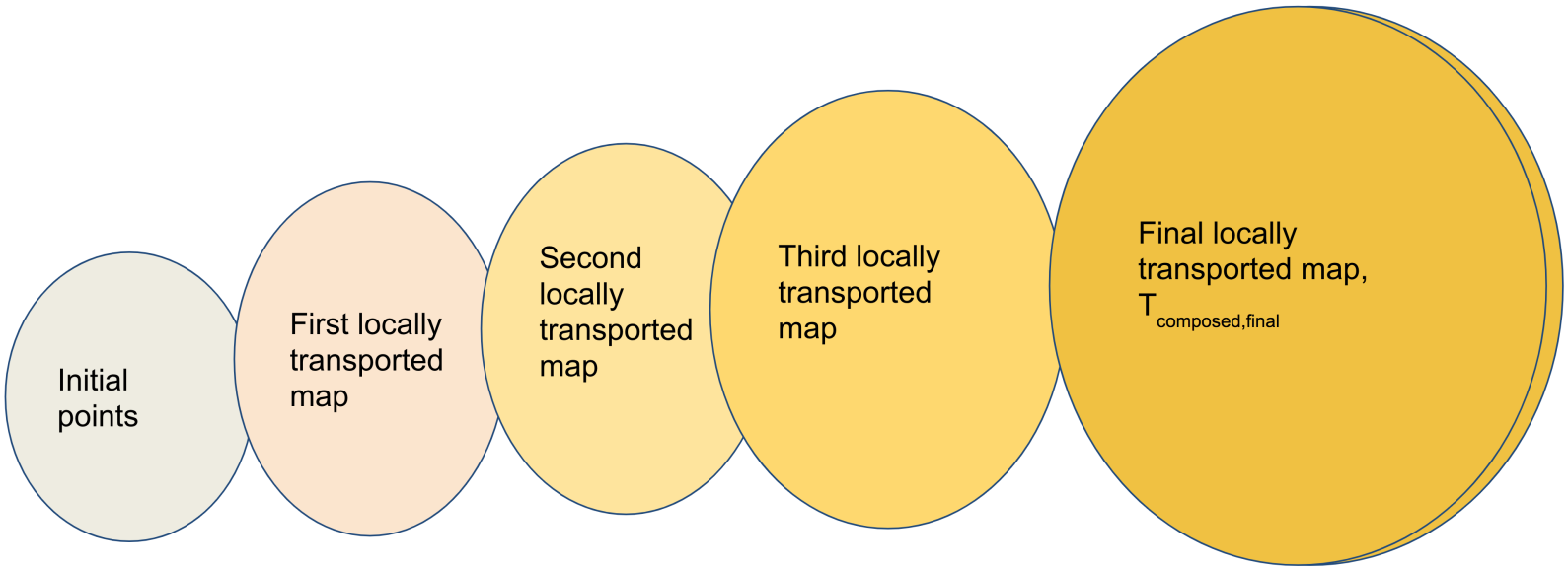

Figure: We iterate this process N times for our final map and its interpolates to be along the geodesic