Nonparametric Density Estimation using Optimal Transport

Rohan Buluswar [1]

[1]University of Chicago

The Problem

Let K be a compact subset of $\mathbb{R}^k$ for some $k \in \mathbb{N}$, and let π be a probability distribution supported on K. **Given S i.i.d. samples from π, we aim to estimate π without making any strict assumptions about its structure.**

For simplicity, we take K to be the unit square in $\mathbb{R}^2$, but the principle of the algorithm is the same in higher dimensions. Moreover, we assume that π admits a density. Because we cannot computationally handle an entire probability density function, we instead aim to output an empirical distribution that better approximates the underlying π.

Bicriteria Optimization

Given an input empirical distribution $\overline{\rho}$, there are two main considerations in constructing the output ρ:

- We want ρ to be 'similar' in some sense to $\overline{\rho}$, so that we do not deviate too far from the observed data. We measure this similarity by the Wasserstein-2 distance, $W_2(\overline{\rho}, \rho)^2$. This choice is motivated by the fact that the Wasserstein metric gives rise to a pseudo-manifold structure on the space of distributions.
- The empirical distribution $\overline{\rho}$ will be very 'spiky', but we expect π to be more smooth. Quantitatively, we want to output ρ with higher differential entropy, as defined by

$$-\int_K \rho(x) \log(\rho(x))\, dx$$

We therefore want to minimize both $W_2(\overline{\rho}, \rho)^2$ and $\int_K \rho \log(\rho)$. We thus introduce a weighting factor λ, solve the following optimization problem, and output $\hat{\rho}$:

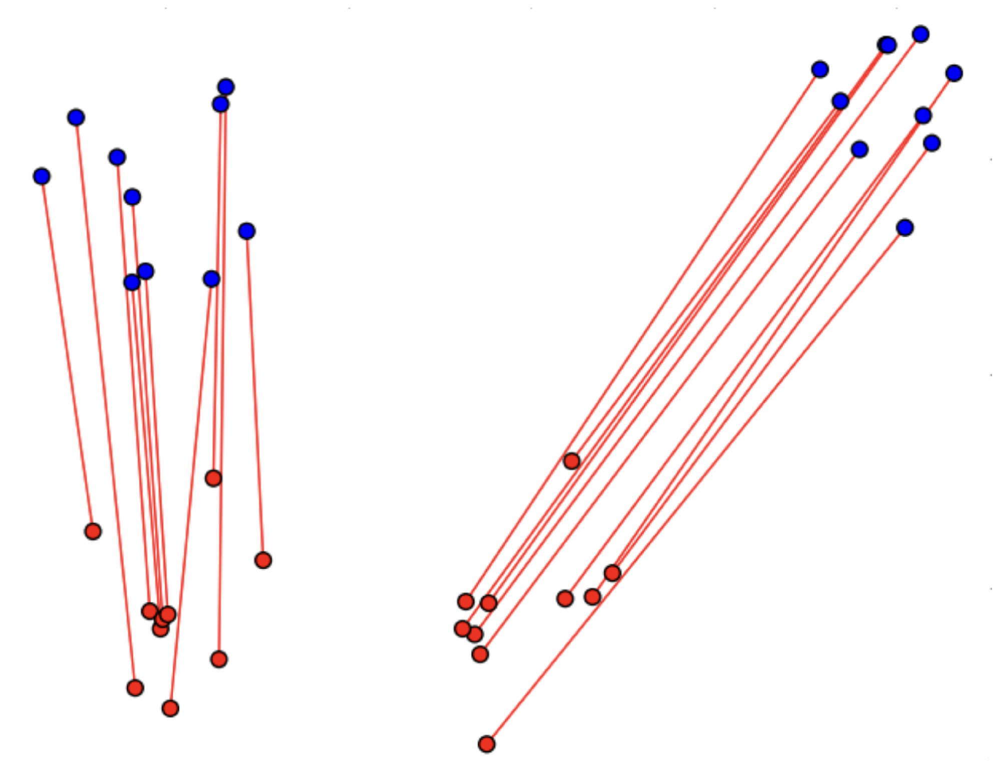
$$\hat{\rho} = \text{argmin } \mathcal{F}(\rho) := \text{argmin}_\rho W_2(\overline{\rho}, \rho)^2 + \lambda \int_K \rho(x) \log \rho(x) dx$$

Background: Optimal Transport Theory

Optimal transport theory is a branch of mathematics that originally aimed to solve the following problem. Given locations of (indistinguishable) mines and factories in a space M, and a cost metric $c(x, y) : M \times M \to \mathbb{R}_{\geq 0}$, how can we move resources from the mines to the factories while minimizing the total cost incurred? By considering the minimum possible cost over all transport plans, we obtain a kind of distance between the distribution of mines and the distribution of factories. Typically, we will take $M = \mathbb{R}^n$ and c to be the Euclidean metric. The problem can be generalized to obtain a true metric on the space of continuous probability distributions (e.g. Wasserstein-2 distance) by choosing an optimal coupling instead of transport plan.

For example, we have the following optimal transport plan:

Figure 1. Example Optimal Transport Plan



Solution via Partial Differential Equation

To solve the optimization problem, we consider the gradient flow of the functional $\mathcal{F}$. We can equivalently define

$$\mathcal{F}(\rho) = \frac{1}{\lambda} W_2(\overline{\rho}, \rho)^2 + \int \rho(x) \log(\rho(x)) dx$$

In particular, we are interested in the Wasserstein gradient flow. According to the continuity equation, it is given by the following PDE (Chewi, 2024):

$$\partial_t \rho_t = \text{div}(\rho_t \nabla_{W_2} \mathcal{F}(\rho_t))$$

The Wasserstein gradient, ∇_{W_2}, can be computed using the first variation, δ. Specifically, $\nabla_{W_2}\mathcal{F}(\rho) = \nabla \delta \mathcal{F}(\rho)$ (Chewi, 2024). In our setting, if f_{ρ_t} is the Kantorovich potential associated with transport from ρ_t to $\overline{\rho}$, we have the following (justified) approximation:

$$\delta \mathcal{F}(\rho_t) \approx \log(\rho_t) + \frac{1}{\lambda} f_{\rho_t}$$

Moreover, the Kantorovich potential f_{ρ_t} is equal to $||x||^2 - 2\varphi_{\rho_t}(x)$, where $\nabla \varphi_{\rho_t}(x)$ is a transport plan from ρ_t to $\overline{\rho}$ (Chewi et al., 2024). Finally, we have

$$\partial_t \rho_t = \text{div}\left(\rho_t(\nabla \log(\rho_t) + \frac{1}{\lambda}(2x - 2\nabla \varphi_{\rho_t}))\right)$$

According to the above calculations, if we set $\rho_0 = \overline{\rho}$ and simulate this PDE, then as $t \to \infty$, $\rho_t \to$ argmin $\mathcal{F}(\rho)$. We therefore arrive at an algorithmic means of solving the optimization problem.

Equivalent Stochastic Differential Equation

We can rewrite the PDE as follows:

$$\partial_t \rho_t = \text{div}(\rho_t(\nabla \log(\rho_t) + \nabla \frac{f_{\rho_t}}{\lambda})) = \text{div}(\rho_t \nabla \log\left(\frac{\rho_t}{\exp(-\frac{f_{\rho_t}}{\lambda})}\right))$$

Moreover, the Langevin diffusion process, defined by the SDE $dZ_t = -\nabla V(Z_t) + \sqrt{2}dB_t$ for some potential function V, satisfies

$$\partial_t \mu_t = \text{div}(\mu_t \nabla \log(\frac{\mu_t}{\pi}))$$

where $\pi \propto \exp(-V)$ and $Z_t \sim \mu_t$ (Chewi, 2024). Hence, instantaneously at each t, ρ_t evolves like the law of a particle undergoing Langevin diffusion with $V = \frac{1}{\lambda} f_{\rho_t}$. We can therefore say that ρ_t evolves like the law of a particle undergoing the following SDE:

$$dZ_t = -\frac{1}{\lambda}(\nabla f_{\rho_t})dt + \sqrt{2}dB_t$$

$$dZ_t = \frac{1}{\lambda}\nabla(2\varphi_{\rho_t}(x) - ||x||^2) + \sqrt{2}dB_t = \frac{1}{\lambda}(2\nabla\varphi_{\rho_t}(x) - 2x) + \sqrt{2}dB_t$$

Up to a change of units of time, this is the same as

$$dZ_t = (2\nabla\varphi_{\rho_t}(x) - 2x) + \lambda\sqrt{2}dB_t$$

The Algorithm

We can now independently subject each 'particle' in the initial sample to the above SDE. The law of each particle evolves according to the PDE, so their overall distribution will, approximately. We thus have the following algorithm, which discretizes the SDE:

1. For each particle in the sample, take N copies, for some fixed $N \in \mathbb{N}$
2. For $t = 1, 2, \ldots, m$:
   1. Estimate the transport map $\nabla\varphi_{\rho_t}$ from the current empirical distribution to the original distribution, $\overline{\rho}$
   2. For each particle p, independently sample a Brownian motion $B(dt)$ and compute its new location by
   $$Z_{t+1}(p) = Z_t(p) + (2\nabla\varphi_{\rho_t}(Z_t(p)) - 2Z_t(p))dt + \lambda\sqrt{2}B(dt)$$
   3. Store the new empirical distribution ρ_{t+1} according to the new locations of the particles

There are several hyperparameters: λ, N, m (the number of steps), and the time interval dt used in discretizing the SDE. Several variants can also be constructed by relaxing the optimal transport problem in the computation of $\nabla\varphi_{\rho_t}$ to an entropic or unbalanced optimal transport.

Experimental Results

The variants of the algorithm were tested on samples from known underlying distributions π, and as a heuristic, success was measured by the decrease in optimal transport distance to π from $\overline{\rho}$ to $\hat{\rho}$, i.e.

$$\frac{W_2(\overline{\rho}, \pi) - W_2(\hat{\rho}, \pi)}{W_2(\overline{\rho}, \pi)}$$

Num. Steps	Standard	Entropic	Unbalanced
2	47.302 %	53.879 %	28.823 %
3	50.019 %	57.029 %	28.023 %
5	47.575 %	58.834 %	16.747 %
6	42.819 %	55.406 %	16.870 %
7	47.256 %	57.245 %	13.439 %
8	51.020 %	59.204 %	18.574 %

Table 1. % OT distance decrease for different variants of the algorithm, with Gaussian underlying, computed as average over five independent trials.

Below is an example of an input distribution and the resulting output distribution, using the standard version of the algorithm:

Figure 2. Input sample for underlying spiral distribution.

Figure 3. Standard algorithm output given input in Figure 2.

Acknowledgements

First, I would like to thank my research mentor, Prof. Lorenzo Orecchia, for his support throughout the course of the year. Second, I would like to acknowledge that this work has been funded by the Quad Undergraduate Research Scholars Program, and I am grateful for the support of the CCRF in my pursuit of my academic and professional goals.

References

1. Flamary, Rémi. *Discrete Optimal Transport*. https://remi.flamary.com/demos/transport.html.
2. Chewi, Sinho. "Log-Concave Sampling." Unpublished manuscript, 2024.
3. Chewi, Sinho, Jonathan Niles-Weed, and Philippe Rigollet. "Statistical Optimal Transport." Preprint, arXiv:2407.18163, 2024. https://arxiv.org/abs/2407.18163.

rbuluswar@uchicago.edu