Optimal Transport and Sinkhorn Algorithm

Thomas Wynn, Austine Do
Presentation Date: April 10th

1 Motivation

When comparing two probability distributions, common metrics include the KL-Divergence or L-p norm.

$$D_{\mathrm{KL}}(P \parallel Q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx \qquad \|f-g\|_p = \left(\int |f(x)-g(x)|^p \, dx \right)^{\frac{1}{p}}$$

However these metrics rely on overlapping support. For instance, KL-Divergence is undefined when one distribution assigns zero probability to an event that the other does not. Similarly, L^p norms measure differences in probability densities but are not suitable for distributions with disjoint supports.

In scenarios where distributions have disjoint supports, a more tractable approach comes from optimal transport theory. Instead of measuring pointwise differences between the distributions, optimal transport theory gives us tools to measure the distances of probability measures with non-overlapping supports.

2 Optimal Transport Theory

2.1 Monge's Formulation

The canonical interpretation of optimal transport theory is concerned with the idea of "moving" mass from one distribution to another. The most well-known formulation in optimal transport theory is known as Monge's formulation:

$$\inf_T \int_X c(x, T(x))\, d\mu(x) \qquad \text{subject to } T_{\#}\mu = \nu$$

Here, X and Y are metric spaces. μ and ν are some probability measures on X and Y respectively. Canonically, μ represents the "source" distribution and ν represents the "target" distribution. $T : X \to Y$ is a "transport map", which assigns each $x \in X$ to a destination $T(x) \in Y$. $c(c, T(x))$ is some cost function, which represents the cost needed to move mass from x to $T(x)$. $T\#\mu = \nu$ is the pushforward constraint which means that $\nu(B) = \mu(T^{-1}(B))$ For any measurable set $B \subset Y$

Intuitively, we can analogize Monge's Formulation to moving a pile of sand to another pile of sand. We can denote the source pile as μ and ν as the destination pile. Each $d\mu(x)$ can be approximated as a infinitesimally small portion of sand at x, and $c(x, T(x))$ is the cost of moving this grain of sand given the transport map. The most intuitive cost function in this case would be the kinetic energy of moving one grain of sand to the destination pile (in physical contexts the squared Euclidean distance is often interpreted as being related to kinetic energy).

However, the problem with Monge's Formulation is that it assumes a one-to-one mapping of mass. However, there often does not exist a deterministic transport map such as in cases where mass must be split between multiple destinations.

2.2 Kantorovich Relaxation

$$\min_{\gamma} \int_{X \times Y} c(x,y)\, d\gamma(x,y) \tag{1}$$

$$\text{subject to} \quad \int_{Y} d\gamma(x,y) = d\mu(x), \int_{X} d\gamma(x,y) = d\nu(y). \tag{2}$$

Here, Kantorovich introduced a "relaxed" formulation of our problem. Instead of a deterministic map T, Knatorovich allowed for transport to be described by $\gamma(x, y)$, where γ describes the probability of transporting mass from $x \in X$ to $y \in Y$. The constraints ensure conservation of mass while transporting X to Y. This formulation allows us to split mass at each point, which allows us to address cases in which a single mass point in X may need to be distributed over multiple locations in Y

2.3 Entropy Regularization and Sinkhorn's algorithm

While Kantorovich's formulation generalizes Monge's problem, it is a linear program, so its optimal solutions are not necessarily unique. Entropic regularization and the resulting Sinkhorn's algorithm can make the problem more tractable.

$$\min_{\gamma} \int_{X \times Y} c(x,y)\, d\gamma(x,y) + \epsilon H(\gamma)$$

The entropy regularization term $H(\gamma)$ is introduced to Kantorovich's relaxation. In effect, the entropy term penalizes overly "concentrated" transport plans, which makes the solution more stable and less sensitive to small variations in data (in our sand pile example, the entropy term would penalize a bulldozer moving all of our mass from μ to ν, which could lead mass concentrated in one place). ϵ is a parameter that controls the strength of the entropy regularization parameter. Sinkhorn's algorithm computes the solution by iteratively updating the rows and columns of the transport matrix ν such that the marginal constraints are met:

$$\int_{Y} \gamma(x,y)\, dy = \mu(x), \quad \int_{X} \gamma(x,y)\, dx = \nu(y)$$

Essentially, the Sinkhorn algorithm starts with an initial guess for γ, and iterates through each row and column of γ scaling them by a factor of $\alpha(x) = \frac{\mu(x)}{\sum_y \gamma(x,y)}, \quad \beta(y) = \frac{\nu(y)}{\sum_x \gamma(x,y)}$ respectively. Iterations continue until convergence. During our presentation, we will further discuss the intuition and applications of the Sinkhorn algorithm.

References

[1] P. Ewald, P. Hoyos *Optimal transport in deep learning.*

[2] J. Solomon Shape Analysis (Lecture 19): Optimal transport YouTube, 2021 URL: `https://www.youtube.com/watch?v=MSbvkhAR0VY&t=2974s&ab_channel=JustinSolomon`.