Abstract

This is a project for MATH 566: Optimal Transportation at UBC. My goal was to make some conjectures (without proof) of some results concerning what I term *regularized* or *congested* optimal transportation. The terminology is suggestive of its application in machine learning, where we will show this formulation is equivalent to a 'naïve' MAP estimator, in addition to its application in economics, where the additional term may be considered to be the disutility of congestion.

This project is entirely original, as I am unaware of any formulations of this problem in a mathematical setting. This is not to say that my work is without inspiration. On the machine learning side Cuturi [1] considers this problem with regularizer $w(\xi) = \lambda\xi\log\xi$ and proposes an algorithmic solver—and suggests the examination of other regularizers. On the economics side Carlier [2] considers a different formulation of congestion, which we will show is equivalent in the discrete case.

We take M^+, M^- to be metric spaces and recall the definition of the *Kantorovich Transportation* problem:

$$C(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int c(x,y)\,d\gamma(x,y). \tag{KT}$$

where $\Pi(\mu,\nu)$ is the set of measures γ with marginal measures $(\pi_1,\pi_2)_\#\gamma = (\mu,\nu)$—$d\gamma(x,y)$ can be thought of as a transportation plan identifying how much 'mass' is transported from a location x to location y. Thus KT can be thought of as the cheapest way to transform one distribution into another according to the cost function $c : M^+ \times M^- \to \mathbb{R}\cup\{\infty\}$ that tells us the difficulty of moving a unit from one location to another. A variety of constraints can be fixed on c to give the problem a certain structure (e.g. convexity/concavity, symmetry, triangle inequality, (semi-)continuity), that we will make reference to. Notably, the optimal measure(s) of KT are rarely absolutely continuous, rather they tend to concentrate all the mass from one location on a single location. This leads to problems with minimization algorithms.

In applications to economics the Kantorovich problem is impractical in that it does not account for the cost of congestion. We combat this problem by adding a *regularizing* term. Now, assume μ,ν are absolutely continuous measures, and consider *Congested Transportation* costs of the form

$$C(\mu,\nu) = \inf_{\gamma\in\Pi_\ell(\mu,\nu)} \int c(x,y)\,d\gamma + \int w\left(\frac{d\gamma}{d\ell}\right) d\ell(x,y) \tag{CT}$$

where ℓ is a measure on $M^-\times M^+$ such that $\mu\otimes\nu \ll \ell$, and $\Pi_\ell(\mu,\nu) = \{\gamma\in\Pi(\mu,\nu)|\gamma\ll\ell\}$ indicates transportation plans that are absolutely continuous with respect to ℓ (non-empty by our constraint on ℓ) and $w : \mathbb{R}^+ \to \mathbb{R}$ is a function of the Radon-Nikodym derivative of γ with respect to ℓ.

A natural choice for ℓ in machine learning is $\mu\otimes\nu$, however it may be more practical to consider it the Lebesgue measure when μ,ν are absolutely continuous or the counting

measure when they are both discrete—this is done by Cuturi [1]. In economics, ℓ might indicate some measure of the capacity of the infrastructure between x and y.

The w term has many different interpretations that yield different intuitive insights. We have already discussed its relevance to economics and machine learning (which we will examine further in conjectures 3 and 4, respectively), but its formulation is also remniscent of an internal energy. Lastly, it may also be seen as an approximation of (KT) as we will discuss in conjecture 2.

We begin with an existence result. This is notably distinct from the Kantorovich setting as $\Pi_\ell(\mu,\nu)$ is not weakly compact. We denote by Γ_K the set of minimizers of KT.

Conjecture 1 (Existence of Optimal Plan)**.** *There exists an optimal measure $\gamma_o \in \Pi_\ell(\mu,\nu)$ attaining the infimum of (CT) if there exists a $\gamma_K \in \Gamma_K$ such that $c(x,y)$ is upper semi-continuous on a neighbourhood of supp(γ_K) and w is strictly convex. Conversely, if w is concave and $\Gamma_K \cap \Pi_\ell(\mu,\nu) = \emptyset$, there does not exist an absolutely continuous minimizer of (CT).*

Remark. The continuity condition is inspired by [3], and ensures that we can approximate the cost on an open (hence not ℓ-null) set by costs on Γ_K to accuracy $\mathcal{O}(|(x,y)|)$, controlling the growth of the first term of eq. (CT). Meanwhile, the convexity of w ensures that the second term of eq. (CT) can be decreased by choosing a 'flatter' distribution (Jensen's inequality would illustrate this clearly). That there is a convex/concave condition may be intuited by the idea that convex costs encourage moderation while concave costs encourage 'putting all your eggs in one basket' (similar to the Kantorovich case with concave cost [4, §2.4.4]), regardless of the coercivity of w, and ends up concentrating the optimal measure around γ_K.

A natural follow-up question is: what is the relation between the answers to (CT) and (KT)? An immediate insight is that (assuming uniqueness) $\mathrm{supp}(\gamma_\mathrm{K}) \subseteq \mathrm{supp}(\gamma_\mathrm{o})$.

Conjecture 2 (Relation between γ_K and γ_o)**.** *If w is concave and $\gamma_K \perp \ell$, the cost in (CT) can be approximated by a sequence of measures weakly converging to $\gamma_K \in \Gamma_K$.*

In general, if w_n is a sequence of functions such that $w_n'' \to 0$ pointwise on $(0,\infty)$, then the weak-$*$ limit points of the optimal measures γ_n are all contained in Γ_K.

Remark. The intuition for the first half of this conjecture is the idea that the cheapest transportation will be concentrated around the optimal Kantorovich transportation, with maximal density. The question remains if any weakly convergent sequence will work, or only a particular one.

The idea for the second half is that as the second derivative goes to zero, the convexity is weakened and the problem prefers more and more extreme tranportation measures. [5, Theorem 3.3] shows this convergence in the case where $w_n(\xi) = \frac{1}{n}\xi\log\xi$.
If $w_n \to 0$ as well, then we recover the Kantorovich cost. While convergence might hold in theory, [1, Figure 1] discusses how this convergence may not occur in practice, due to machine-precision errors that persist.

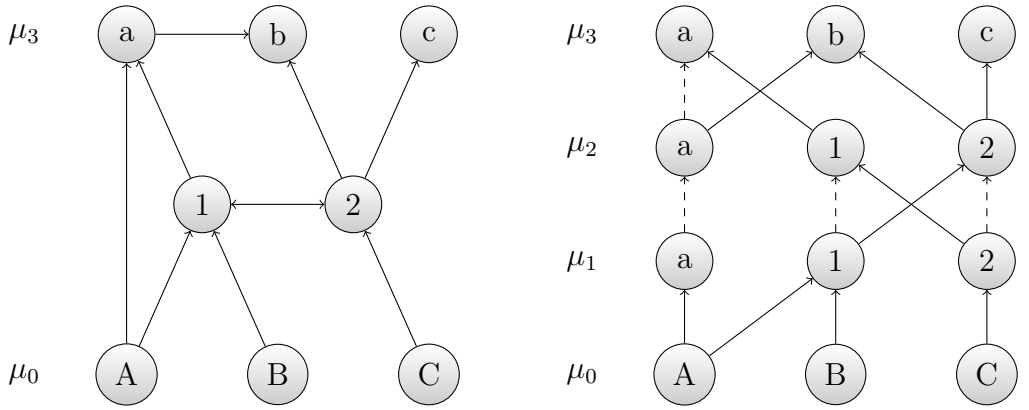

Figure 1: On the left, an example of the type of graph considered in [2, §4.1], on the right is the decomposed graph that Conjecture 3 proposes is equivalent. The cost $c(x, y)$ between two vertices not connected by an edge $x \to y$ is ∞, while the costs between connected $x \to y$ is as on the left, dashed edges indicated a cost of zero.

Conjecture 1 has a simple interpretation in economics. Here w is interpreted as the utility lost due to congestion in a *transportation network*—represented as a finite oriented graph as in fig. 1 [2, §4.1]. In this case, conjecture 1 states that in cases where the utility is concave (e.g. descreasing marginal utility), and there are multiple paths with similarly low costs available available, we may expect these paths to be taken.

However, this is notably dependent on the structure of our network: as now multiple paths may contribute congestion to the same road segment. Thus (CT) only considers transportation networks of a complete bipartite graph. A useful question is whether we can describe any other types of networks using this equation. An easy example: to remove the edge between any two vertices x, y, set $c(x, y) = \infty$, effectively barring the route.

Conjecture 3 (Adaption to Transportation Networks)**.** *The precise case explored in [2, §4.1], that is transportation between measures μ_0, μ_N concentrated on the vertices of a finite oriented simple graph $G = (V, E)$, can be realized as*

$$C_G(\mu_0, \mu_N) = \min \left\{ \sum_{i=0}^{N-1} C(\mu_i, \mu_{i+1}) \middle| \mu_i \in P(V) \right\} \tag{3.1}$$

where N is the length of the longest (acyclical) path, and $C(\mu_i, \mu_{i+1})$ is as in (CT), with $c(x, y) = \infty$ for $(x, y) \notin E$ (that is x and y do not share an edge).

Remark. Note that this in conjunction with conjecture 1 implies that the minimizer of eq. (3.1) can be decomposed into a sequence of optimal transportation plans concentrated on neighbouring vertices (see fig. 1). This conjecture is important as it allows us to extend results that hold for (CT) to more general settings. Indeed, it offers a more natural

example of conjecture 1 than could be given before: if there is decreasing marginal utility associated with congestion (ie. w is convex), then paths with similar costs will have similar congestions—it's not winner take all.

In machine learning, (CT) is important as w may be used as a regularizer that discourages sparsity of the transportation matrix in discrete transportation models [6]. This is advantageous because, counter intuitively, less sparse optimal matrices reduce computation costs [1]. With the latter application in mind, I propose the following conjecture that establishes conditions under computation will be expedient:

Conjecture 4 (Sparsity of Transportation)**.** *Let c be lower semicontinuous. Define the set of feasible couplings to be $A := supp(\mu \otimes \nu) \cap \{x, y | c(x, y) < \infty\}$. Assume w is C^1 on $\left(0, \frac{1}{\lambda(A)}\right]$ and continuous at 0, then the optimal transport plan will have minimal sparsity—that is $supp(\gamma_o) = A$—when*

$$\lim_{\xi \searrow 0} w'(\xi) = -\infty \tag{4.1}$$

The converse holds when $\sup_A c(x, y) = \infty$.

Remark. In a machine learning context, where the transportation coupling is often viewed as probabilistic, this states the conditions under which no feasible coupling will be ruled out. In an economic context, this provides the conditions under which every possible route will be traversed, which can be extended to more complex networks via conjecture 3. Notably, conjecture 4 posits that the values of w' away from zero are irrelevant.

The intuition behind this result parallels that of cyclical monotonicity, and indeed is corollary to the following conjected lemma.

Lemma 5 (w-Congested c-Equivalence)**.** *Let c be lower semi-continuous, $w \in C^1$. Consider the problem of congested transport between M^- and M^+, with optimal plan γ. Let $(x_0, y_0), (x_1, y_1) \in A$ with $\rho_{ij} := \gamma(x_i, y_j)$, and $c_{ij} = c(x_i, y_j)$. Then*

$$c_{11} + w'(\rho_{11}) + c_{22} + w'(\rho_{22}) = c_{12} + w'(\rho_{12}) + c_{21} + w'(\rho_{21}) \tag{wCcE}$$

Remark. Like c-cyclical monotonicity, this can be intuited from the discrete case (where it may be proven by contradiction—if eq. (wCcE) is not fulfilled, then more mass can be transported along the plan with lower marginal cost), and likely may be extended to the continuous case using similar methods [3]. This result reinforces conjectures 1 and 2, as it suggests that the optimal plan is invariant if an affine term is added to w: ie. $w(\xi) \to w(\xi) + a\xi + b$, hence emphasizing the importance of w'' and convexity.

A common regularizer [1, 6] is the (scaled) negentropy of a measure, in which case $w(\xi) = \lambda\xi \log \xi$ (where $\lambda \in \mathbb{R}^+$) with ℓ Lebesgue. This notably satisfies the conditions for both conjectures 1 and 4. This regularizer is often (eg. [7, §2.1]) justified by an appeal to Jaynes' principle of entropy maximization [8, §9.7].

Lastly, I hope to give some more intuition to the choice of ℓ term in CT by showing how this term may be obtained from a Maximum a Posteriori estimation (where $w(\gamma) = -\log p(\gamma|\ell)$ according to some prior distribution p parametrized by ℓ), which is a consequence of conjecture 5. This term then corresponds to the prior assumption that each coupling is drawn from (naïvely) independent Poisson distributions with mean ℓ_{ij}. Notably this sheds some light on why $\ell = \mu \otimes \nu$ is a natural choice, as it seems like a good prior for the transportation between x and y would be proportional to both $d\mu(x)$ and $d\nu(y)$.

Applying Stirling's approximation to our poisson distributions, we get:

$$\begin{aligned} p(\gamma|\ell) \propto& \prod_{ij} \frac{\ell_{ij}^{\gamma_{ij}}}{\gamma_{ij}^{\gamma_{ij}} e^{-\gamma_{ij}}} \\ \log p(\gamma|\ell) =& C_\ell + \sum_{ij} \gamma_{ij} \left(\log \ell_{ij} + 1\right) - \gamma_{ij} \log\left(\gamma_{ij}\right) \end{aligned}$$

where C_ℓ is a normalizing constant. By lemma 5, the minimizing γ is invariant under the addition of linear terms, meaning

$$\arg\max_\gamma p(\gamma|\ell) = \arg\min_\gamma \sum_{ij} \gamma_{ij} \log\left(\frac{\gamma_{ij}}{\ell_{ij}}\right) = \arg\min_\gamma \mathrm{KL}(\gamma \parallel \ell) \tag{5.1}$$

where KL is the Kullback–Leibler divergence. Note that the λ factor can arise by scaling the probability distribution $\gamma \to \lambda\gamma$:

$$\begin{aligned} \log p(\lambda\gamma|\ell) =& C_\ell + \sum_{ij} \lambda\gamma_{ij} \left(\log \ell_{ij} + 1\right) - \lambda\gamma_{ij} \log\left(\lambda\gamma_{ij}\right) \\ =& C_\ell + \sum_{ij} \lambda\gamma_{ij} \left(\log \ell_{ij} + 1 - \log\lambda\right) - \lambda\gamma_{ij} \log\left(\gamma_{ij}\right) \\ \arg\max_\gamma p(\lambda\gamma|\ell) =& \arg\min_\gamma \sum_{ij} \lambda\gamma_{ij} \log\left(\frac{\gamma_{ij}}{\ell_{ij}}\right) \end{aligned} \tag{5.2}$$

Additional questions:

- Conjecture 3 extends our results to a discrete process where $\gamma = (\gamma_n)$, as considered in [2, §4.1]. However [2, 9] also consider the case of the continuous process. Can we do the same by considering $\gamma = \gamma_t \otimes dt$ (as a disintegration)?
- (Experimental) In what cases is the naïve independence assumption in our MAP estimation realistic/problematic? Similarly, does using $\ell = \mu \otimes \nu$ provide any advantages over a Lebesgue ℓ?

References

[1] M. Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. *Neural Information Processing Systems (NIPS)*, page 2292–2300, 2013. URL https://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optima

[2] Guillaume Carlier. Optimal transportation and economic applications. In *SIMA, New Mathematical Models in Economics and Finance*, 2010.

[3] Filippo Santambrogio. c-Cyclical monotonicity of the support of optimal transport plans. URL http://www.math.u-psud.fr/~santambr/SupportcCM.pdf.

[4] C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003. ISBN 9780821833124.

[5] Christian Léonard. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879 – 1920, 2012. ISSN 0022-1236. doi: http://dx.doi.org/10.1016/j.jfa.2011.11.026. URL http://www.sciencedirect.com/science/article/pii/S0022123611004253.

[6] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2016.

[7] B. J. Brewer and M. J. Francis. Entropic Priors and Bayesian Model Selection. In P. M. Goggans and C.-Y. Chan, editors, *American Institute of Physics Conference Series*, volume 1193 of *American Institute of Physics Conference Series*, pages 179–186, December 2009. doi: 10.1063/1.3275612.

[8] E.T. Jaynes and G.L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN 9780521592710. URL https://books.google.ca/books?id=tTN4HuUNXjgC.

[9] G. Carlier, C. Jimenez, and F. Santambrogio. Optimal transportation with traffic congestion and Wardrop equilibria. *ArXiv Mathematics e-prints*, December 2006.