We recall the definition of the Kantorovich Transportation problem:

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

and define Γ_{K} to be the space of minimizing measures γ_{K} for this equation.

Now, assume μ,ν are absolutely continuous measures, and consider 'Congested Transportation' costs of the form

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

where ℓ is the $2n$-dimensional Lebesgue measure on $M^-\times M^+$, $\Pi_{\mathrm{ac}}(\mu,\nu) = \{\gamma|(\pi^+)_\#\gamma = \mu, (\pi^-)_\#\gamma = \nu, \gamma \ll \ell\}$ indicates absolutely continuous measures with marginal densities μ,ν and $w:\mathbb{R}^+\to\mathbb{R}$ is a function of the density of γ. Note that the discrete case may be formulated as (CT) in the case where M^+, M^- are separable.

The w term has many different interpretations that yield different intuitive insights. It is inspired by regularization strategies that make (CT) faster to compute than (KT) [1], sharing its formulation with that of an internal energy, but it may also be considered as the economic cost of congestion in a transportation plan [2]. A last application of (CT) that I will investigate is as a means of approximating (KT).

We denote γ_{K} the minimizer to the the Kantorovich transportation problem with cost $c(x,y)$. The question of existence of minimizing measures is notably distinct from the Kantorovich setting as $\Pi_{\mathrm{ac}}(\mu,\nu)$ is not weakly compact:

Conjecture 1 (Existence of Optimal Plan)**.** *There exists an optimal measure $\gamma_o\in\Pi_{ac}(\mu,\nu)$ attaining the infimum of (CT) if there exists a $\gamma_K\in\Gamma_K$ such that $c(x,y)$ is continuous on a neighbourhood of supp(γ_K) and w is strictly convex. Conversely, if w is concave, 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.

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 $\gamma_K\perp\ell$ is not absolutely continuous (e.g. hw1 problem 2), 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 $\gamma_n \to \gamma_K \in \Gamma_K$ in the weak-$*$ topology.

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 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 the convergence may not occur in practice, due to machine-precision errors that persist.

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 [6, §4.1]. In this case, conjecture 1 states that in cases where the utility is concave (e.g. descreasing marginal utility), and there are similar paths available, we may expect these paths to be taken.

However, this is notably dependent on the structure of our network: congestion costs may now be dependent on the traffic from other routes. (CT) only considers transportation networks of a bipartite graph composed of single connections between each $x \in \mathrm{supp}(\mu)$ and $y \in \mathrm{supp}(\nu)$. 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 [6, §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.

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 [7], this is advantageous because less sparse matrices reduce computation costs [1]. With the latter application in mind, I propose the following conjecture that establishes conditions under computation will be expedient:

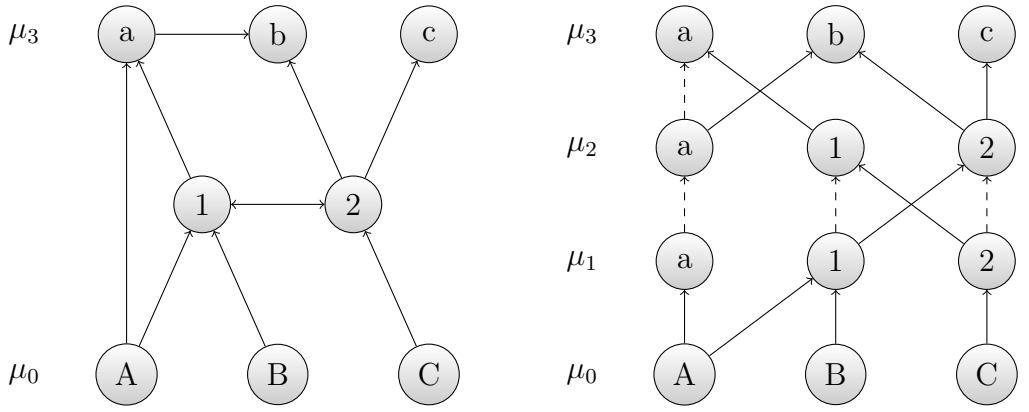

Figure 1: On the left, an example of the type of graph considered in [6, §4.1], on the right is the decomposed graph that Conjecture 3 proposes is equivalent. The cost $c(x, y)$ between two vertices not connected by an edge $x \to y$ is ∞, while the costs between connected $x \to y$ is as on the left, dashed edges indicated a cost of zero.

Conjecture 4 (Sparsity of Transportation)**.** *Let c be lower semicontinuous. Define the set of feasible couplings to be $A := supp(\mu \otimes \nu) \cap \{x, y | c(x, y) < \infty\}$. If 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*

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

This is a necessary condition in the case where $\sup_A c(x, y) = \infty$.

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, 7] is the (scaled) negentropy of a measure, in which case $w(\xi) = \lambda\xi \log \xi$ (where $\lambda \in \mathbb{R}^+$). This notably satisfies the conditions for both conjectures 1 and 4. This is often (eg. [8, §2.1]) justified by an appeal to Jaynes' principle of entropy maximization [9, §9.7].

The last thing I will show is how this term may be obtained from a Maximum a Posteriori estimation (where $w(\gamma) = -\log p(\gamma|\alpha)$ according to some prior distribution p parametrized by α), which is a consequence of conjecture 5. This term then corresponds to the prior assumption that each coupling is drawn from iid Poisson distributions, where we replace the factorial on the denomonator with Stirling's approximation to get the continuous analog:

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

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

$$\arg\max_\gamma p(\gamma|\alpha) = \arg\min_\gamma \sum_{ij} \gamma_{ij} \log(\gamma_{ij}) \tag{5.1}$$

Note that the λ factor can arise by scaling the probability distribution $\gamma \to \lambda\gamma$:

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

Additional questions:

- Do our results generalize to non-homogeneous density functions $w(\gamma, x, y)$? In an economic setting this reflects a non-uniform transportation capacity across the network, in machine learning this reflects prior knowledge that some couplings are more likely than others.
- Conjecture 3 extends our results to a discrete process where $\gamma = (\gamma_n)$, as considered in [6, §4.1]. However [2, 6] also consider the case of the continuous process. Can we do the same by considering $\gamma = \gamma_t \otimes dt$ (as a disintegration)?

- (Experimental) Does including more terms of Stirling's approximation to the regularizer prove advantageous? How does eq. (5.1) change if α is a vector? Does eq. (5.2) shed any light on the choice of λ?

References

[1] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances," *Neural Information Processing Systems (NIPS)*, p. 2292–2300, 2013. [Online]. Available: https://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf

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

[3] F. Santambrogio, "c-Cyclical monotonicity of the support of optimal transport plans." [Online]. Available: http://www.math.u-psud.fr/~santambr/SupportcCM.pdf

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

[5] C. Léonard, "From the schrödinger problem to the monge–kantorovich problem," *Journal of Functional Analysis*, vol. 262, no. 4, pp. 1879 – 1920, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022123611004253

[6] G. Carlier, "Optimal transportation and economic applications," in *SIMA, New Mathematical Models in Economics and Finance*, 2010.

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

[8] B. J. Brewer and M. J. Francis, "Entropic Priors and Bayesian Model Selection," in *American Institute of Physics Conference Series*, ser. American Institute of Physics Conference Series, P. M. Goggans and C.-Y. Chan, Eds., vol. 1193, Dec. 2009, pp. 179–186.

[9] E. Jaynes and G. Bretthorst, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. [Online]. Available: https://books.google.ca/books?id=tTN4HuUNXjgC