
Moslin Supplementary Notes

Contents

1	Optimal transport (OT) to map cells across time points	2
1.1	The Kantorovich relaxation of OT	2
1.2	Entropic regularization and the Sinkhorn algorithm	3
1.3	OT extensions: unbalancedness and (Fused) Gromov-Wasserstein (FGW)	4

1 Optimal transport (OT) to map cells across time points

Following the success of previous methods¹ that link cells across time, moslin is based on optimal transport (OT) to reconstruct links between cross-sectional measurements. In this section, we give an introduction to OT² and highlight extensions that we use in moslin to incorporate lineage tracing data.

OT is an area of mathematics concerned with mapping distributions from one space to another; in our case, these distributions can be thought of as ensembles of sampled cells at different time points. The original notion of OT was introduced by the French mathematician Gaspard Monge in 1781 who considered the problem of moving a pile of sand to a hole where the shape of the pile and the hole are prescribed and a certain cost is associated with moving each grain of sand from source to target locations. This problem is known as the *Monge problem*; it leads to a constrained non-convex optimization problem that is not guaranteed to have a feasible solution in general². Kantorovich³ in 1942 famously proposed to relax the transport problem by allowing probability mass from a source destination to be split across several target destinations, thus evolving from deterministic *transport maps* to probabilistic *couplings*.

1.1 The Kantorovich relaxation of OT

For probability vectors (equivalently, *histograms*) $\mathbf{a} \in \Delta_N$ and $\mathbf{b} \in \Delta_M$, define the corresponding discrete measures α and β as

$$\alpha(\mathbf{x}) = \sum_{i=1}^N a_i \delta_{\mathbf{x}_i} \quad (1)$$

$$\beta(\mathbf{y}) = \sum_{j=1}^M b_j \delta_{\mathbf{y}_j}, \quad (2)$$

with respect to source and target locations, $(\mathbf{x}_i, \mathbf{y}_j) \in \mathcal{X} \times \mathcal{Y}$ for all $(i, j) \in \{1, \dots, N\} \times \{1, \dots, M\}$ where $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^D$. In our application, these represent N sampled cells at t_1 and M sampled cells at t_2 .

OT over discrete probability measures seeks to find a *coupling matrix* $P \in \mathbb{R}_+^{N \times M}$, transporting mass from α to β in a way that is optimal with respect to a cost function $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ where $c(\mathbf{x}_i, \mathbf{y}_j)$ is the cost associated with moving a piece of unit mass from location \mathbf{x}_i to \mathbf{y}_j . In our case, \mathbf{x}_i and \mathbf{y}_j can be thought of as cellular profiles in some embedded phenotypic space, e.g. the D -dimensional scVI⁴ or PCA latent space, and c quantifies biologically meaningful distance in that space.

The set of feasible couplings, given by those matrices P that satisfy the marginal constraints imposed though the probability vectors \mathbf{a} and \mathbf{b} , may be written as

$$U(\mathbf{a}, \mathbf{b}) := \left\{ P \in \mathbb{R}_+^{N \times M} : P\mathbf{1}_M = \mathbf{a}, P^\top \mathbf{1}_N = \mathbf{b} \right\}, \quad (3)$$

for constant-one vectors $\mathbf{1}_N$ and $\mathbf{1}_M$ of lengths N and M , respectively. In our case, the marginals \mathbf{a} and \mathbf{b} describe distributions over sampled cell states at early and late time points, respectively. They may be chosen to be uniform; alternatively, they can reflect any cell-level prior such as cellular growth and death.

With these definitions at hand, we can state the Kantorovich relaxation of the OT problem as follows:

$$\mathcal{L}_c(\alpha, \beta) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \langle C, P \rangle, \quad (4)$$

for cost matrix $C \in \mathbb{R}_+^{N \times M}$ with $C_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$ and $\langle C, P \rangle := \sum_{ij} C_{ij} P_{ij}$. Following Peyré et al.², we make the dependency of the OT problem \mathcal{L}_c on the cost function c explicit. In case the cost

function c is related to a proper distance metric d via $c(\mathbf{x}_i, \mathbf{y}_j) = d^p(\mathbf{x}_i, \mathbf{y}_j)$, i.e. the cost is given by the p -th power of a distance metric, we define the p -Wasserstein distance as

$$W_p(\alpha, \beta) = \mathcal{L}_{d^p}(\alpha, \beta)^{1/p}, \quad (5)$$

where \mathcal{L}_{d^p} is defined as in Equation (4). We refer to Peyré et al.² for an overview and to Chen et al.⁵ for a single-cell application of Wasserstein distances to describe patient-level variation in terms of single-cell gene expression.

Practical considerations. The objective function defined in Equation (4) is linear with constraints given by the $N + M$ equality constraints imposed through $U(\mathbf{a}, \mathbf{b})$, thus it defines a convex linear program whose solution is in general non-unique. Various strategies have been suggested to solve Equation (4), among them network flow solvers and the auction algorithm⁶, however, all of these remain limited to finding a solution in time $\mathcal{O}(N^3)$ for $M = N$, omitting logarithmic factors. This poses a scalability issue for applications in current single-cell genomics datasets which frequently contain hundreds of thousands of cells. Further, practical limitations include the difficulty to adapt these algorithms to run on GPUs and to be differentiable.

1.2 Entropic regularization and the Sinkhorn algorithm

To overcome these practical limitations, consider the following entropically regularized⁷ variant of the OT problem,

$$\mathcal{L}_c^\epsilon(\alpha, \beta) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \langle P, C \rangle - \epsilon H(P), \quad (6)$$

for regularization strength $\epsilon > 0$ and entropy term

$$H(P) := - \sum_{ij} P_{ij} (\log P_{ij} - 1). \quad (7)$$

In contrast to the unregularized problem, Equation (6) is ϵ -strongly convex and thus possesses a unique global optimum. Further, for $\epsilon \rightarrow 0$ and $\epsilon \rightarrow \infty$, Peyré et al.² and Cominetti and Martín⁸ show the following asymptotic results:

$$\mathcal{L}_c^\epsilon(\alpha, \beta) \rightarrow \mathcal{L}_c \text{ for } \epsilon \rightarrow 0, \quad (8)$$

$$P_\epsilon \rightarrow \mathbf{a}\mathbf{b}^\top \text{ for } \epsilon \rightarrow \infty, \quad (9)$$

where P_ϵ refers to the solution of the regularized problem with regularization parameter ϵ . In particular, these results show that for small ϵ , the regularized problem approximates the unregularized problem while for large ϵ , the coupling converges to the outer product of the two marginals, which is closely related to the *Maximum Mean Discrepancy* distance commonly used in generative models for distribution matching, see Lotfollahi et al.⁹ for a single-cell application. To solve the regularized OT problem of Equation (6), we reproduce the following proposition from Peyré et al.²:

Proposition 1.1 (Solution to the regularized OT problem). *The unique solution to the regularized OT problem introduced in Equation (6) can be written as*

$$P_{ij} = u_i K_{ij} v_j \quad \forall (i, j) \in \{1, \dots, N\} \times \{1, \dots, M\}, \quad (10)$$

for the associated Gibbs kernel $K_{ij} = \exp(-C_{ij}/\epsilon)$ and scaling variables $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^N \times \mathbb{R}_+^M$ to be inferred.

Proof. For a proof, see Peyré et al.². □

Sinkhorn algorithm. The form of the solution outlined in Proposition 1.1 can be used to construct an algorithm that iterates between scaling the rows and the columns of a candidate matrix. Impose therefore the marginal constraints of the feasible set $U(\mathbf{a}, \mathbf{b})$ on a solution in the form of Proposition 1.1,

$$\text{diag}(\mathbf{u}) K \text{diag}(\mathbf{v}) \mathbf{1}_M = \mathbf{a}, \quad (11)$$

$$\text{diag}(\mathbf{v}) K^\top \text{diag}(\mathbf{u}) \mathbf{1}_N = \mathbf{b}, \quad (12)$$

which can be simplified to give

$$\mathbf{u} \odot (K\mathbf{v}) = \mathbf{a}, \quad (13)$$

$$\mathbf{v} \odot (K^\top \mathbf{u}) = \mathbf{b}, \quad (14)$$

with \odot denoting elements-wise multiplication. Iteratively solving these equations gives rise to *Sinkhorn's algorithm*,

$$\mathbf{u}^{(l+1)} := \frac{\mathbf{a}}{K\mathbf{v}^{(l)}}, \quad (15)$$

$$\mathbf{v}^{(l+1)} := \frac{\mathbf{b}}{K^\top \mathbf{u}^{(l+1)}}, \quad (16)$$

where the division is applied element-wise. Yule¹⁰ originally suggested iterations of this form, Sinkhorn¹¹ proofed their convergence and Cuturi⁷ suggested applying the algorithm to solve entropically regularized OT problems which gives rise to a differentiable solution. Note that Sinkhorn's algorithm is well suited to run on GPUs since it only relies on matrix vector products.

Cuturi et al.¹² recently used JAX¹³ to implement Sinkhorn's algorithm in their Optimal Transport Tools (OTT) software package; OTT thus allows for just-in-time compilation, GPU acceleration, online cost function evaluation, and automatic differentiation. Moslin is implemented in moscot, which connects OTT to the single-cell world. In contrast to the linear programming algorithms introduced above, Sinkhorn runs in time $\mathcal{O}(N^2)$ for $N = M$, omitting logarithmic factors. Thus, solving the regularized (Equation (6)) rather than the unregularized (Equation (4)) OT problem offers both practical and theoretical advantages including improved statistical convergence^{2,14}.

1.3 OT extensions: unbalancedness and (Fused) Gromov-Wasserstein (FGW)

The moslin model requires two further extensions of the regularized OT problem, we follow Peyré et al.² in their presentation. First, cells proliferate and die while they differentiate which we include in the marginal distributions \mathbf{a} and \mathbf{b} . However, as the rates of cellular growth and death are difficult to estimate based on scRNA-seq data, we relax the marginal constraints which leads to *unbalanced optimal transport*.

Second, lineage tracing information in ICE is only comparable within one time point as barcodes are generated independently in each replicate. Nevertheless, it is reasonable to assume that pairwise lineage similarity is conserved on average; across time points, cell pairs with similar lineage distances are more likely to belong together compared to cell pairs with very different lineage distances. Thus, we allow for pairwise comparisons of lineage information only within each time point which leads to *Gromov-Wasserstein OT* (GW).

Unbalanced OT. To relax the marginal constraints, consider a generalization of the regularized OT problem by adding divergences D_ϕ between marginal constraints and row/columns sums of P to the objective function,

$$\mathcal{L}_c^{\epsilon, \tau}(\alpha, \beta) := \min_{P \in \mathbb{R}_+^{N \times M}} \langle C, P \rangle + \rho_a D_\phi(P \mathbf{1}_M | \mathbf{a}) + \rho_b D_\phi(P^\top \mathbf{1}_N | \mathbf{b}) - \epsilon H(P), \quad (17)$$

where the parameters $\rho_a, \rho_b > 0$ control the weight given to the soft marginal constraints^{2,15}. We follow OTT¹² and use the parametrization $\rho_{a,b} := \frac{\epsilon\tau_{a,b}}{1-\tau_{a,b}}$ for $\tau_{a,b} \in (0, 1)$.

This is a generalization of the original entropic OT problem of Equation (6) as in the limit $\tau_1, \tau_2 \rightarrow 1$, i.e., $\rho_a, \rho_b \rightarrow \infty$, one recovers the original problem. A generalized version of the Sinkhorn algorithm may be applied to solve Equation (17); of particular importance in practical applications¹ has been the case $D_\phi(\cdot|\cdot) = \text{KL}[\cdot||\cdot]$ for which the Sinkhorn updates read

$$\mathbf{u}^{(l+1)} := \left(\frac{\mathbf{a}}{K\mathbf{v}^{(l)}} \right)^{\tau_a}, \quad (18)$$

$$\mathbf{v}^{(l+1)} := \left(\frac{\mathbf{b}}{K^\top \mathbf{u}^{(l+1)}} \right)^{\tau_b}. \quad (19)$$

We refer to Liero et al.¹⁵ for a treatment of the theory behind unbalanced OT and to Chizat et al.¹⁶ for the derivation of practical algorithms.

Unbalancedness for approximate growth rates and cellular sampling. The generalized Sinkhorn algorithm provides an efficient solution to unbalanced OT problems which arise in moslin when allowing for cellular growth and death at an uncertain rate g , or when considering the stochastic effects of sampling cells from a population (Methods).

Gromov-Wasserstein optimal transport (GW). In standard OT, we assume that point clouds $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \dots, \mathbf{y}_M\}$ corresponding to bins in the histograms \mathbf{a} and \mathbf{b} may be compared using the cost function $c(\cdot, \cdot)$, i.e. \mathcal{X} and \mathcal{Y} correspond to the same metric space, giving rise to the cost matrix C . To relax this assumption, consider a situation where vectors $\{\mathbf{x}_i\}_{i=1}^N$ may be compared using the cost function $c^\mathcal{X}$ and vectors $\{\mathbf{y}_i\}_{i=1}^M$ may be compared using the cost function $c^\mathcal{Y}$, but no direct comparisons of vectors in \mathcal{X} and \mathcal{Y} are possible. Using these cost functions, we define the entropically regularized GW problem,

$$\mathcal{L}_{L, c^\mathcal{X}, c^\mathcal{Y}}^\epsilon(\alpha, \beta) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \sum_{ijkl} L(C_{ij}^\mathcal{X}, C_{kl}^\mathcal{Y}) P_{ik} P_{jl} - \epsilon H(P), \quad (20)$$

for cost matrices $C^\mathcal{X} \in R_+^{N \times N}$, $C^\mathcal{Y} \in R_+^{M \times M}$ with $C_{ij}^\mathcal{X} = c^\mathcal{X}(\mathbf{x}_i, \mathbf{x}_j)$, $C_{kl}^\mathcal{Y} = c^\mathcal{Y}(\mathbf{y}_k, \mathbf{y}_l)$ and distance metric L . The solution to (the unregularized version of) Equation (20) defines the *Gromov-Wasserstein distance* between two metric spaces, each equipped with a probability distribution. This distance has been introduced by Mémoli¹⁷ as an extension to the Gromov-Hausdorff distance¹⁸, combined with entropic regularization by Peyré et al.¹⁹, Solomon et al.²⁰ and used in the single-cell field e.g. for data integration across modalities²¹. Equation (20) defines a non-convex, constrained, smooth optimization problem in P ; before discussing its optimization, we introduce one further generalization.

Fused Gromov-Wasserstein optimal transport. In the moslin model, while lineage information is only comparable within one time point, molecular similarity is comparable across time points. Thus, we encounter a problem that possesses characteristics of both OT (Equation (6)) and GW (Equation (20)). Problems of this kind require a combined objective function where some sampled features may be compared across spaces while others may only be compared within one space. This kind of problem is known as *entropically-regularized Fused Gromov-Wasserstein* (FGW), which has previously been employed in the single-cell field to reconstruct spatial cell arrangement²². The problem is defined via

$$\mathcal{L}_{L, c^\mathcal{X}, c^\mathcal{Y}, c}^{\epsilon, \alpha}(\alpha, \beta) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \alpha \sum_{ijkl} L(C_{ij}^\mathcal{X}, C_{kl}^\mathcal{Y}) P_{ik} P_{jl} + (1 - \alpha) \sum_{ik} C_{ik} P_{ik} - \epsilon H(P), \quad (21)$$

where $\alpha \in [0, 1]$ controls the weight given to the OT versus GW terms, the within-space cost functions $c^{\mathcal{X}}$ and $c^{\mathcal{Y}}$ are defined for these features $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_i\}_{i=1}^M$ which may not be compared across spaces and the across-space cost function c is defined for these features $(\mathbf{x}'_i, \mathbf{y}'_i)$ which may be compared across spaces²³. For $\alpha = 1$, we recover the GW problem introduced above.

Introducing the 4-tensor¹⁹,

$$\mathcal{T}(C^{\mathcal{X}}, C^{\mathcal{Y}})_{ijkl} := L\left(C_{ik}^{\mathcal{X}}, C_{jl}^{\mathcal{Y}}\right), \quad (22)$$

allows us to rewrite Equation (21) in shorter form,

$$\mathcal{L}_{L, c^{\mathcal{X}}, c^{\mathcal{Y}}, c}^{\epsilon, \alpha}(\alpha, \beta) := \min_{P \in U(\mathbf{a}, \mathbf{b})} \alpha \langle \mathcal{T}(C^{\mathcal{X}}, C^{\mathcal{Y}}) \otimes P, P \rangle + (1 - \alpha) \langle C, P \rangle - \epsilon H(P) \quad (23)$$

with tensor multiplication defined via

$$(\mathcal{T} \otimes P)_{ij} := \sum_{kl} \mathcal{T}_{ijkl} P_{kl}. \quad (24)$$

This tensor product may be computed in time $\mathcal{O}(N^3)$ for $M = N$ for a class of separable loss functions L , including l_2 loss and the KL divergence¹⁹.

FGW optimization. Following Peyré et al.¹⁹, we use projected gradient descent with iterations

$$P^{(l+1)} = \text{Proj}_{U(\mathbf{a}, \mathbf{b})}^{\text{KL}}\left(P^{(l)} \odot e^{-\tau \nabla J(P)|_{P^{(l)}}}\right), \quad (25)$$

where $\text{Proj}_{U(\mathbf{a}, \mathbf{b})}^{\text{KL}}(\tilde{P}) = \text{argmin}_{P \in U(\mathbf{a}, \mathbf{b})} \sum_{ij} P_{ij} \log\left(P_{ij}/\tilde{P}_{ij}\right)$ is a KL projection operator, τ is a step size, J is the FGW objective function defined in Equation (21) and \odot denotes element-wise multiplication. The gradient of the objective function may be written as

$$\nabla J = (1 - \alpha)C + \alpha \mathcal{T}(C^{\mathcal{X}}, C^{\mathcal{Y}}) \otimes P, \quad (26)$$

while the KL projection can be solved via an OT problem²⁴,

$$\text{Proj}_{U(\mathbf{a}, \mathbf{b})}^{\text{KL}}(\tilde{P}) = \text{argmin}_{P \in U(\mathbf{a}, \mathbf{b})} \left\langle -\epsilon \log \tilde{P}, P \right\rangle - \epsilon H(P). \quad (27)$$

Using Equation (26), Equation (27) and setting $\tau = 1/\epsilon$, we can re-write the update rule of Equation (25) as

$$P^{(l+1)} = \text{argmin}_{P \in U(\mathbf{a}, \mathbf{b})} \left\langle (1 - \alpha)C + \alpha \mathcal{T}(C^{\mathcal{X}}, C^{\mathcal{Y}}) \otimes P^{(l)}, P \right\rangle - \epsilon H(P), \quad (28)$$

which is the entropically regularized OT problem of Equation (6), solved efficiently at each iteration for an evolving cost matrix using the Sinkhorn algorithm^{7,19,22}. The algorithm outlined here is applicable to both GW ($\alpha = 1$) and FGW ($\alpha \in (0, 1)$) settings. The major computational bottleneck is the update of the tensor product of Equation (24) required at each update of Equation (28), which runs in time $\mathcal{O}(N^3)$ for $N = M$.

Unbalanced FGW optimal transport. We showed in Equation (17) how OT may be relaxed to allow deviations from the exact marginals \mathbf{a}, \mathbf{b} to accommodate uncertain cellular growth and death rates; we may use a similar trick for FGW by replacing the hard marginal constraints with soft quadratic ϕ divergences D_{ϕ}^{\otimes} . This has been explored by Séjourné et al.²⁵ where the authors introduce the *Unbalanced Gromov-Wasserstein* (UGW) divergence and show that it upper-bounds a proper distance metric between metric spaces equipped with a positive measure.

Combining unbalancedness with FGW leads to our final optimization problem which captures the various problem aspects:

- a Wasserstein term to compare cells across time points using e.g. molecular similarity¹.
- a Gromov-Wasserstein^{17,19} term to compare cells within time points using e.g. lineage information in ICE.
- unbalancedness^{16,26} to handle stochastic cell sampling, and uncertain growth- and death-rates.
- entropic regularization for improved computational^{2,7} and statistical properties¹⁴.

Thus, moslin's OT formulation allows the method to flexibly capture the various problem aspects of linking lineage-traced cells across time points.

References

- [1] Geoffrey Schiebinger, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.
- [2] Gabriel Peyré, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- [3] Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133(4): 1381–1382, 2006.
- [4] Romain Lopez, et al. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12): 1053–1058, 2018.
- [5] William S Chen, et al. Uncovering axes of variation among single-cell cancer specimens. *Nature methods*, 17(3):302–310, 2020.
- [6] Dimitri P Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1): 152–171, 1981.
- [7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- [8] Roberto Cominetti and J San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1):169–187, 1994.
- [9] Mohammad Lotfollahi, et al. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36(Supplement_2):i610–i617, 2020.
- [10] G Udny Yule. On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652, 1912.
- [11] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The annals of mathematical statistics*, 35(2):876–879, 1964.
- [12] Marco Cuturi, et al. Optimal transport tools (ott): A jax toolbox for all things wasserstein. *arXiv preprint arXiv:2201.12324*, 2022.
- [13] Roy Frostig, et al. Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, 2018.
- [14] Aude Genevay, et al. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.
- [15] Matthias Liero, et al. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.
- [16] Lenaïc Chizat, et al. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.
- [17] Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.
- [18] Mikhael Gromov, et al. *Metric structures for Riemannian and non-Riemannian spaces*, volume 152. Springer, 1999.
- [19] Gabriel Peyré, et al. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672. PMLR, 2016.
- [20] Justin Solomon, et al. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.
- [21] Pinar Demetci, et al. Scot: Single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.
- [22] Mor Nitzan, et al. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.
- [23] Titouan Vayer, et al. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

- [24] Jean-David Benamou, et al. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.
- [25] Thibault Séjourné, et al. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.
- [26] Lenaïc Chizat, et al. Unbalanced optimal transport: Dynamic and kantovich formulations. *Journal of Functional Analysis*, 274(11):3090–3123, 2018.