Optimal Transport with Adaptive Regularisation

Hugues Van Assel[1*], **Titouan Vayer**[2], **Rémi Flamary**[3], **Nicolas Courty**[4]
[1]UMPA ENS Lyon , [2]LIP ENS Lyon, [3]CMAP École polytechnique,
[4]IRISA Université Bretagne Sud

Abstract

Regularising the primal formulation of optimal transport (OT) with a strictly convex term leads to enhanced numerical complexity and a denser transport plan. Many formulations impose a global constraint on the transport plan, for instance by relying on entropic regularisation. As it is more expensive to diffuse mass for outlier points compared to central ones, this typically results in a significant imbalance in the way mass is spread across the points. This can be detrimental for some applications where a minimum of smoothing is required per point. To remedy this, we introduce OT with Adaptive RegularIsation (OTARI), a new formulation of OT that imposes constraints on the mass going in or/and out of each point. We then showcase the benefits of this approach for domain adaptation.

1 Introduction

Optimal transport (OT) is a well-established framework to compare probability distributions with numerous applications in machine learning [1, 20, 21]. Discrete OT seeks a transportation plan that minimizes the total transportation cost between samples from the source and target distributions. In the absence of regularisation, this optimal OT plan is inherently sparse. Regularising OT with a strictly convex term is a widely adopted practical approach, leading to reduced numerical complexity and more diffuse OT plans [21]. As an illustration, the prominent entropic regularisation [9] leads to a dense plan. In some applications, the smoothing effect induced by the regularisation has a primary importance on its own. A key example is the construction of doubly stochastic affinity matrices for clustering and dimensionality reduction [15, 27], where smoothing enables connecting to neighbor points. Another is domain adaptation [8] where smoothed OT often results in enhanced performance when compared to non-regularised ones (see for instance Table 1). Many OT regularisation schemes on the primal formulation impose a constraint on the overall transport plan. Consequently, central data points tend to exhibit a denser (or more diffuse) transport plan compared to extreme (or outlier) data points, for which diffusion is more costly. As a result, the latter points receive limited benefits from the smoothing effect introduced by the regularizer as shown in the left side of Figure 1. This partly explains OT's significant sensitivity to outliers in many applications [19, 7]. To remedy this, one needs to constrain the transport plan in a pointwise manner. Note that this has recently been explored for constructing affinity matrices [25] (*i.e.* symmetric OT setting) leading to enhanced noise robustness and clustering abilities.

Contributions. In this work, we develop a new formulation of OT, called OT with Adaptive RegularIsation (OTARI), allowing to control, for any strictly convex function ψ, the value of ψ on each row and/or column of the OT plan. We then show the advantages of OTARI over usual regularised OT on domain adaptation tasks, focusing particularly on the negative entropy and the ℓ_2 norm respectively associated with entropic [9] and quadratic [4] optimal transport.

[*]hugues.van_assel@ens-lyon.fr

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

2 Regularised Optimal Transport

We first introduce the discrete OT problem before presenting regularised formulations and associated algorithms. All the notations are presented in Appendix A. Let $X_S = \{\boldsymbol{x}_i^S \in \mathbb{R}^d\}_{i=1}^{N_S}$ and $X_T = \{\boldsymbol{x}_i^T \in \mathbb{R}^d\}_{i=1}^{N_T}$ denote the sets of respectively source and target point locations. The discrete *Monge-Kantorovitch* problem [13] focuses on the optimal allocation strategy to transport the empirical measure $\mu_S = \sum_{i=1}^{N_S} a_i \delta_{\boldsymbol{x}_i^S}$ onto $\mu_T = \sum_{i=1}^{N_T} b_i \delta_{\boldsymbol{x}_i^T}$ where $\boldsymbol{a} \in \Delta^{N_S}$ and $\boldsymbol{b} \in \Delta^{N_T}$. It consists in computing a *coupling* $\mathbf{P} \in \mathbb{R}_+^{N_S \times N_T}$ *i.e.* a joint probability measure over $X_S \times X_T$ solving the linear program

$$\min_{\mathbf{P} \in \Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle \tag{OT}$$

where $\Pi(\boldsymbol{a}, \boldsymbol{b})$ is the transport polytope with marginals $(\boldsymbol{a}, \boldsymbol{b})$ and the *cost matrix* $\mathbf{C} \in \mathbb{R}_+^{N_S \times N_T}$ encodes the pairwise distances between the source and target samples. One can typically consider the squared Euclidean distance $C_{ij} = \|\boldsymbol{x}_i^S - \boldsymbol{x}_j^T\|_2^2$ or any Riemannian distance over a manifold [26].

To enable faster algorithmic resolution as well as smoother solutions, one can rely on a strictly convex regulariser $\psi : \mathbb{R}^{N_s} \to \mathbb{R}$. It amounts to solving $\min_{\mathbf{P} \in \Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle + \varepsilon^\star \sum_i \psi(\mathbf{P}_{i:})$ where $\varepsilon^\star > 0$. Interestingly, regularised OT can also be framed using a convex constraint as follows

$$\min_{\mathbf{P} \in \Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C} \rangle \quad \text{s.t.} \quad \mathbf{P} \in \overline{\mathcal{B}}(\eta) \tag{ROT}$$

where $\overline{\mathcal{B}}(\eta) \coloneqq \{\mathbf{P} \text{ s.t. } \sum_i \psi(\mathbf{P}_{i:}) \leq \eta\}$. Note that the previously introduced $\varepsilon^\star$ is the optimal dual variable associated with the constraint $\overline{\mathcal{B}}(\eta)$ in the above equivalent formulation. Throughout, we make the following assumption on ψ.

Assumption 1. *Let $\psi : \mathrm{dom}(\psi) \to \mathbb{R} \cup \{\infty\}$ be strictly convex and differentiable on the interior of its domain $\mathrm{dom}(\psi) \subset \mathbb{R}_+^{N_S}$.*

In what follows, we denote by $\psi(\mathbf{P}) = (\psi(\mathbf{P}_{1:}), ..., \psi(\mathbf{P}_{N_S:}))^\top$. We introduce $\psi^* := \mathbf{p} \to \sup_{\mathbf{q} \in \mathrm{dom}(\psi)} \langle \mathbf{p}, \mathbf{q} \rangle - \psi(\mathbf{q})$ the convex conjugate of ψ [23]. Note that when ψ is strictly convex, this supremum is uniquely achieved and from Danskin's theorem [10]: $\nabla \psi^*(\boldsymbol{p}) = \arg\max_{\mathbf{q} \in \mathrm{dom}(\psi)} \langle \mathbf{p}, \mathbf{q} \rangle - \psi(\mathbf{q})$. We show in Appendix B.1 that when $\varepsilon^\star > 0$, *i.e.* when the constraint $\mathbf{P} \in \overline{\mathcal{B}}(\eta)$ is active, (ROT) is solved for $\mathbf{P}^\star = \nabla \psi^*((\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)/\varepsilon^\star)$ [1] where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \varepsilon^\star)$ solve the following dual problem

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon > 0} \langle \boldsymbol{\lambda}, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} \rangle + \varepsilon \Big(\sum_i \psi^*((\mathbf{C}_{i:} - \lambda_i \mathbf{1} - \boldsymbol{\mu})/\varepsilon) - \eta \Big) . \tag{Dual-ROT}$$

The above objective is concave thus the problem can be solved exactly using *e.g.* BFGS [17] or ADAM [14]. As a complementary view, one can also frame (ROT) as a ψ-Bregman projection over a convex set. The ψ-Bregman divergence is defined as $D_\psi(\mathbf{P}, \mathbf{Q}) := \psi(\mathbf{P}) - \psi(\mathbf{Q}) - \langle \mathbf{P} - \mathbf{Q}, \nabla \psi(\mathbf{Q}) \rangle$. The solution of (ROT) can then be expressed as $\mathbf{P}^\star = \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b}) \cap \overline{\mathcal{B}}(\eta)}(\mathbf{K}_\sigma)$ where $\mathbf{K}_\sigma \coloneqq \nabla \psi^*(-\mathbf{C}/\sigma)$ for any $\sigma < \varepsilon^\star$ (see Appendix B.1 for details). The key benefit of the above result is that it enables solving (ROT) with alternating Bregman projection schemes [3].

In this work, we focus specifically on certain Bregman divergences: the Kullback Leibler (KL) divergence and the squared Euclidean distance. The first reads $D_{\mathrm{KL}}(\mathbf{P}|\mathbf{Q}) = \langle \mathbf{P}, \log(\mathbf{P} \oslash \mathbf{Q}) - \mathbf{1}\mathbf{1}^\top \rangle$ with associated negative entropy $\psi_{\mathrm{KL}}(\mathbf{p}) = \langle \mathbf{p}, \log \mathbf{p} - 1 \rangle$. In this case, (ROT) boils down to entropic OT and solved for $\mathrm{Proj}^{\mathrm{KL}}_{\Pi(\boldsymbol{a},\boldsymbol{b})}(\mathbf{K}_{\varepsilon^\star})$ where $\mathbf{K}_{\varepsilon^\star} = \nabla \psi^*_{\mathrm{KL}}(-\mathbf{C}/\varepsilon^\star) = \exp(-\mathbf{C}/\varepsilon^\star)$ is a Gibbs kernel. This projection is well-known as the *static Schrödinger bridge* [16] referring to statistical physics where it first appeared [24], and can be computed efficiently using the Sinkhorn algorithm [9]. For the squared Euclidean distance, we define $\psi_2(\mathbf{p}) = \frac{1}{2}\|\mathbf{p}\|_2^2$. The associated problem (1) is usually referred to as quadratic OT [18] and can yield sparse OT plans unlike entropic.

3 Optimal Transport with Adaptive Regularisation

In this section, we present a new formulation of OT that imposes constraints on each row of the OT plan. We begin by introducing the set of matrices with *point-wise* constraints. To set the upper bound,

[1] We use the notation $\nabla \psi^*(\mathbf{P}) \coloneqq (\nabla \psi^*(\mathbf{P}_{1:}), ..., \nabla \psi^*(\mathbf{P}_{N_S:}))^\top$.

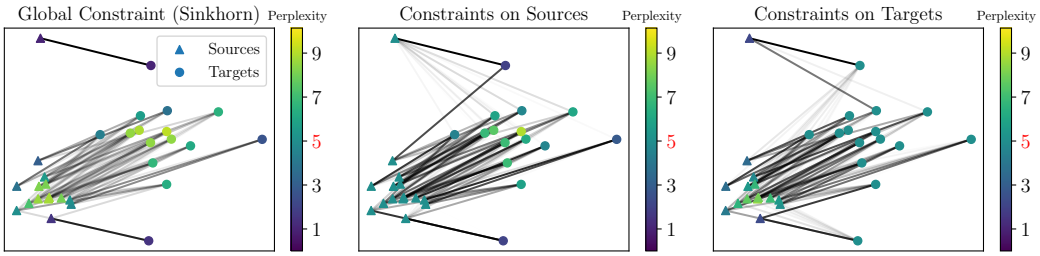

Figure 1: Entropic OT plans ($\xi = 5$) with global constraint, pointwise constraints on sources and then on targets. The three plans have the same global entropy. The color of each source (resp. target) point shows the perplexity (exponential of entropy) of the associated row (resp. column) of the OT plan.

we rely on the perplexity parameter ξ [25] that can be interpreted as the number of effective neighbors for each point. Concretely, we define $\mathbf{e}_\xi = \frac{1}{\xi}(\mathbb{1}_{i\leq\xi})_i$ the probability vector of ξ uniform arms and

$$\mathcal{B}_\psi(\xi) \coloneqq \{\mathbf{P} \geq \mathbf{0} \text{ s.t. } \forall i,\ \psi(\mathbf{P}_{i:}) \leq \psi(\mathbf{e}_\xi)\}\ . \tag{1}$$

Note that in particular $\psi_{\mathrm{KL}}(\mathbf{e}_\xi) = -(\log\xi + 1)$ and $\psi_2(\mathbf{e}_\xi) = 1/\xi$. We now define Optimal Transport with Adaptive RegularIsation (OTARI) as the generalization of (ROT) to the case where the strictly convex constraint is given by $\mathcal{B}_\psi(\xi)$. Similarly to Proposition 1 (Appendix B.1), we can frame OTARI as a ψ-Bregman projection of $\mathbf{K}_\sigma = \nabla\psi^*(-\mathbf{C}/\sigma)$ or solve it using dual ascent.

Proposition 2. *Let $(\boldsymbol{a}, \boldsymbol{b}, \xi)$ be such that $\Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \mathcal{B}_\psi(\xi)$ has an interior point and let $\mathbf{P}^\star$ solve*

$$\min_{\mathbf{P}\in\Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C}\rangle \quad \textit{s.t.} \quad \mathbf{P} \in \mathcal{B}_\psi(\xi)\ . \tag{OTARI-s}$$

Let $\boldsymbol{\varepsilon}^\star$ be the optimal dual variable associated with the constraint $\mathbf{P} \in \mathcal{B}_\psi(\xi)$. If $\boldsymbol{\varepsilon}^\star > \mathbf{0}$, then it holds $\mathbf{P}^\star = \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b})\cap\mathcal{B}_\psi(\xi)}(\mathbf{K}_\sigma)$ for any $0 < \sigma \leq \min_i \varepsilon_i^\star$. Moreover it holds $\mathbf{P}^\star = \nabla\psi^*\left(\mathrm{diag}(\boldsymbol{\varepsilon}^\star)^{-1}(\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)\right)$ where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\varepsilon}^\star)$ solve the following dual

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle + \left\langle \boldsymbol{\varepsilon}, \psi^*\left(\mathrm{diag}(\boldsymbol{\varepsilon})^{-1}(\mathbf{C} - \boldsymbol{\lambda} \oplus \boldsymbol{\mu})\right) - \psi(\mathbf{e}_\xi)\mathbf{1}\right\rangle\ . \tag{Dual-OTARI-s}$$

Algorithm 1 *Dykstra* for solving (OTARI-d)

Input: $\mathbf{C}$, $\psi(\cdot)$, ξ^a, ξ^b, ε, $\boldsymbol{a}$, $\boldsymbol{b}$
$(\mathbf{P}_b, \boldsymbol{\Xi}, \boldsymbol{\Theta}) \leftarrow (\nabla\psi^\star(-\mathbf{C}/\varepsilon), \mathbf{0}, \mathbf{0})$
while not converged **do**
  $\mathbf{P}_a \leftarrow \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a})}(\mathbf{P}_b)$
  $\overline{\mathbf{P}}_a \leftarrow \mathrm{Proj}^{D_\psi}_{\mathcal{B}_\psi(\xi^a)} \circ \nabla\psi^*(\nabla\psi(\mathbf{P}_a) + \boldsymbol{\Xi})$
  $\boldsymbol{\Xi} \leftarrow \boldsymbol{\Xi} + \nabla\psi(\mathbf{P}_a) - \nabla\psi(\overline{\mathbf{P}}_a)$
  $\mathbf{P}_b^\top \leftarrow \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{b})}(\overline{\mathbf{P}}_a^\top)$
  $\overline{\mathbf{P}}_b^\top \leftarrow \mathrm{Proj}^{D_\psi}_{\mathcal{B}_\psi(\xi^b)} \circ \nabla\psi^*((\nabla\psi(\mathbf{P}_b) + \boldsymbol{\Theta})^\top)$
  $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} + \nabla\psi(\mathbf{P}_b) - \nabla\psi(\overline{\mathbf{P}}_b)$
end while
Output: $\overline{\mathbf{P}}_b$

According to Proposition 2, one can solve (OTARI-s) using either alternating projections or dual ascent. When $\boldsymbol{\varepsilon}^\star > \mathbf{0}$, meaning that all constraints are active *i.e.* $\forall i,\ \psi(\mathbf{P}^\star_{i:}) = \psi(\mathbf{p}_\xi)$, dual ascent is usually faster. However, if $\boldsymbol{\varepsilon}^\star$ has null components, one can still rely on $\mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b})\cap\mathcal{B}_\psi(\xi)}(\mathbf{K}_\varepsilon)$ to provide an approximate solution as alternating Bregman projections are always guaranteed to converge.

Note that we can impose the pointwise constraint equivalently on the rows or the columns of the OT plan. Hence (OTARI-t) can be defined by imposing the constraint on the target samples *i.e.* $\mathbf{P}^\top \in \mathcal{B}_\psi(\xi)$. We also propose a doubly constrained formulation called (OTARI-d) that consists of projecting $\mathbf{K}_\sigma$ onto the nonempty set $\mathcal{B}_\psi(\xi^a) \cap \mathcal{B}_\psi^\top(\xi^b)$ where we defined $\mathcal{B}_\psi^\top(\xi) = \{\mathbf{P}^\top \in \mathcal{B}_\psi(\xi)\}$ thus ensuring sufficient smoothing for both rows and columns. Such projection can be computed using alternating Bregman projections, whose convergence has been well-studied [6, 3]. As we generally do not have access to a closed form for the projection onto the transport polytope $\Pi(\boldsymbol{a}, \boldsymbol{b})$, it is common to alternate projection onto $\Pi(\boldsymbol{a})$ and $\Pi(\boldsymbol{b})$ separately (see *e.g.* the seminal Sinkhorn algorithm [9]). We extend this scheme by adding projection steps into the pointwise constraints $\mathcal{B}_\psi(\xi)$ for both $\mathbf{P}$ and $\mathbf{P}^\top$. As this set is not affine, one needs to resort to the Dykstra procedure [11] that can be applied to a broad class of Bregman divergences [2], as shown in Algorithm 1. In Appendix C, we provide the form of the projections for ψ_{KL} and ψ_2. A key benefit of decoupling both row and column constraints is that projection onto $\mathcal{B}_\psi(\xi)$ exhibits a simple structure where the rows can be decoupled into independent subproblems.

4 Application to Domain Adaptation

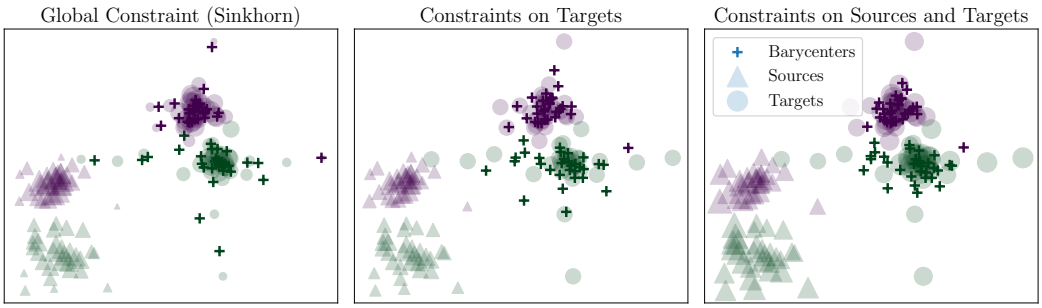

Figure 2: Toy domain adaptation scenario with entropic OT plans ($\xi = 10$) with various constraints. The size of the point is proportional to the associated entropy. When using Sinkhorn, barycentric mapping match outliers points since the OT plan is less diffuse for these points. In turn, using pointwise constraints concentrate the mapped points in high-density regions, thus giving more robust estimates for the mappings onto the target domain.

	OT	EOT	EOTARI-s	EOTARI-t	EOTARI-d
MNIST $\to$ USPS ($\xi = 30$)	53.1(5.4)	64.2(2.8)	65.0(5.3)	66.4(3.5)	**67.4(2.9)**
MNIST $\to$ USPS ($\xi = 300$)	53.1(5.4)	68.8(3.1)	70.8(4.2)	70.2(3.4)	**72.6(5.1)**
USPS $\to$ MNIST ($\xi = 30$)	59.1(4.9)	60.8(5.4)	61.6(4.4)	**62.6(3.0)**	61.0(4.7)
USPS $\to$ MNIST ($\xi = 300$)	59.1(4.9)	59.8(1.6)	61.0(2.3)	**61.6(3.0)**	58.8(2.3)
	OT	**QOT**	**QOTARI-s**	**QOTARI-t**	**QOTARI-d**
MNIST $\to$ USPS ($\xi = 30$)	53.1(5.4)	68.3(3.9)	68.3(3.6)	**69.3(4.7)**	68.1(4.6)
MNIST $\to$ USPS ($\xi = 300$)	53.1(5.4)	60.7(1.5)	**67.0(2.4)**	65.5(2.3)	65.8(2.5)
USPS $\to$ MNIST ($\xi = 30$)	59.1(4.9)	60.4(3.5)	**62.8(3.7)**	59.6(2.7)	61.6(3.1)
USPS $\to$ MNIST ($\xi = 300$)	59.1(4.9)	59.2(3.4)	60.1(3.0)	**62.0(3.7)**	61.5(3.8)

Table 1: Domain adaptation 1-NN classification scores for OT (unregularised), EOT (entropic), EOTARI (entropic OTARI), QOT (quadratic), QOTARI (quadratic OTARI) for $\xi = 30$ and $\xi = 300$.

In this section, we evaluate OTARI on a domain adaptation task where the goal is to transport labeled data points to a target domain where a classifier is trained. Mapping onto the target domain is performed through a barycentric mapping of the form: for any $i \in [\![N_S]\!]$, $\hat{\boldsymbol{x}}_i = \frac{1}{a_i}\sum_j T_{ij}\boldsymbol{x}_j^T$. Looking at Figure 2, one can notice that using OTARI for domain adaptation yields a mapping that is concentrated in high-density (thus more faithful) regions of the target domain. On the opposite, when using globally constrained OT (left side of Figure 2), the barycentric mapping associated with an outlier is concentrated on the outlier's position. For the experiments, we take $\mathbf{C}$ as the squared Euclidean distance computed from raw images of the handwritten digit classification benchmark MNIST-USPS. Following the standard practice in OT-based domain adaptation [12], we map the source to the target samples and then train a 1-NN classifier on the barycentric mappings with source labels. We compute the outcomes across 10 independent trials. In each of these experiments, the target data is partitioned into a 90% training and 10% testing split, with OT barycentric mappings and 1-NN classifiers exclusively applied to the training set. Mean scores and standard deviations are displayed in Table 1. The latter shows that adaptive regularisation consistently outperforms global regularisation (set such that $\sum_i \psi(\mathbf{P}^\star_{i:}) = n\psi(\mathbf{e}_\xi)$ where $\mathbf{P}^\star$ solves ROT for a fair comparison) with significant performance gains in some settings.

5 Conclusion

In this work, we presented a versatile framework to control the value of any OT regulariser in source or/and target locations. We showed encouraging preliminary results for domain adaptation that will be investigated in upcoming works. One could also extend OTARI to continuous distributions and apply it to OT mapping estimation [22]. Other interesting directions include investigating optimization algorithms that can avoid quadratic memory complexity.

Acknowledgments

The authors are grateful to Aurélien Garivier and anonymous reviewers for insightful comments. This project was supported in part by the ANR projects AllegroAssai ANR-19-CHIA-0009 and OTTOPIA ANR-20-CHIA-0030. This work was also supported by the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] Mathieu Blondel, Vivien Seguy, and Antoine Rolet. Smooth and sparse optimal transport. In *International conference on artificial intelligence and statistics*, pages 880–889. PMLR, 2018.

[5] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] Yair Censor and Simeon Reich. The dykstra algorithm with bregman projections. *Communications in Applied Analysis*, 2(3):407–420, 1998.

[7] Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. InfoOT: Information maximizing optimal transport. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6228–6242. PMLR, 23–29 Jul 2023.

[8] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Neural Information Processing Systems (NeurIPS)*, 26, 2013.

[10] John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

[11] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[12] Rémi Flamary, Cédric Févotte, Nicolas Courty, and Valentin Emiya. Optimal spectral transportation with application to music transcription. *Neural Information Processing Systems (NeurIPS)*, 29, 2016.

[13] L Kantorovich. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, page 227, 1942.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Boris Landa, Ronald R Coifman, and Yuval Kluger. Doubly stochastic normalization of the gaussian kernel is robust to heteroskedastic noise. *SIAM journal on mathematics of data science*, 3(1):388–413, 2021.

[16] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

[17] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[18] Dirk A Lorenz, Paul Manns, and Christian Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

[19] Debarghya Mukherjee, Aritra Guha, Justin M Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *International Conference on Machine Learning*, pages 7850–7860. PMLR, 2021.

[20] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[21] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[22] Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

[23] R Tyrrell Rockafellar. *Convex analysis*, volume 11. Princeton university press, 1997.

[24] Erwin Schrödinger. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u . . . , 1931.

[25] Hugues Van Assel, Titouan Vayer, Rémi Flamary, and Nicolas Courty. Snekhorn: Dimension reduction with symmetric entropic affinities. *arXiv preprint arXiv:2305.13797*, 2023.

[26] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

[27] Ron Zass and Amnon Shashua. Doubly stochastic normalization for spectral clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Neural Information Processing Systems (NeurIPS)*. MIT Press, 2006.

A Notations

We adopt the conventions that $0/0 = 0$, $0\log(0) = 0$ and $x/0 = \infty$ for $x > 0$. exp, log applied to vectors/matrices are taken element-wise. $\mathbf{1}$ is the all-one vector whose size depends on the context. $\langle \cdot, \cdot \rangle$ is the standard inner product for matrices/vectors. P_{ij} denotes the entry at position (i, j) of a matrix $\mathbf{P}$ while $\mathbf{P}_{i:}$ and denotes the i-th row. $\mathbf{P} \geq \mathbf{0}$ means that for any (i, j), $P_{ij} \geq 0$. $\odot$ (*resp.* $\oslash$) stands for element-wise multiplication (*resp.* division) between vectors/matrices. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{a} \oplus \boldsymbol{b} \in \mathbb{R}^{n\times n}$ is $(a_i + b_j)_{ij}$. For $\alpha \in \mathbb{R}$, $\boldsymbol{p}^{\odot\alpha}$ and $\mathbf{P}^{\odot\alpha}$ denote element-wise exponentiation *i.e.* $[\boldsymbol{p}^{\odot\alpha}]_i = p_i^\alpha$. $[\mathbf{P}]_+$ is the element-wise positive part with $\max(0, P_{ij})$ in position (i, j). For $n \in \mathbb{N}$, Δ^n is the probability simplex $\{\boldsymbol{p} \in \mathbb{R}_+^n \text{ s.t. } \sum_i p_i = 1\}$. For $\boldsymbol{a} \in \Delta^{N_S}$ and $\boldsymbol{b} \in \Delta^{N_T}$, $\Pi(\boldsymbol{a}, \boldsymbol{b}) = \{\mathbf{P} \in \mathbb{R}_+^{N_S\times N_T} \text{ s.t. } \mathbf{P}\mathbf{1} = \boldsymbol{a} \text{ and } \mathbf{P}^\top\mathbf{1} = \boldsymbol{b}\}$ is the transport polytope with marginals $(\boldsymbol{a}, \boldsymbol{b})$ while $\Pi(\boldsymbol{a}) = \{\mathbf{P} \in \mathbb{R}_+^{N_S\times N_T} \text{ s.t. } \mathbf{P}\mathbf{1} = \boldsymbol{a}\}$ is the semi-relaxed transport polytope. For a set $\mathcal{E}$ and a divergence D, $\mathrm{Proj}_{\mathcal{E}}^{D}(\mathbf{K}) = \arg\min_{\mathbf{P}\in\mathcal{E}} D(\mathbf{P}|\mathbf{K})$.

B Proofs of the Optimal Transport Solutions

B.1 Optimal Transport with Global Constraint

Proposition 1. Let $\psi : \mathbb{R}^{N_S} \to \mathbb{R}$ satisfy Assumption 1. We define $\overline{\mathcal{B}}(\eta) := \{\mathbf{P} \text{ s.t. } \sum_i \psi(\mathbf{P}_{i:}) \leq \eta\}$. Let $(\boldsymbol{a}, \boldsymbol{b}, \eta)$ be such that $\Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \overline{\mathcal{B}}(\eta)$ has an interior point and let $\mathbf{P}^\star$ be a solution of

$$\min_{\mathbf{P}\in\Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C}\rangle \quad \text{s.t.} \quad \mathbf{P} \in \overline{\mathcal{B}}(\eta)\,. \tag{ROT}$$

Let $\varepsilon^\star$ be an optimal dual variable associated with $\mathbf{P} \in \overline{\mathcal{B}}(\eta)$. If $\varepsilon^\star > 0$, then for any $0 < \sigma \leq \varepsilon^\star$, it holds $\mathbf{P}^\star = \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b})\cap\overline{\mathcal{B}}(\eta)}(\mathbf{K}_\sigma)$ where $\mathbf{K}_\sigma := \nabla\psi^*(-\mathbf{C}/\sigma)$. One also has $\mathbf{P}^\star = \nabla\psi^*((\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)/\varepsilon^\star)$ where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \varepsilon^\star)$ solve

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\varepsilon>0} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle + \varepsilon\Big(\sum_i \psi^*((\mathbf{C}_{i:} - \lambda_i\mathbf{1} - \boldsymbol{\mu})/\varepsilon) - \eta\Big)\,. \tag{Dual-ROT}$$

Proof. We first show that $\mathbf{P}^\star = \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b})\cap\overline{\mathcal{B}}(\eta)}(\mathbf{K}_\varepsilon)$ before focusing on the dual problem.

Part I : Proof of the Bregman projection.

Simplifying the constant terms $\mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a},\boldsymbol{b})\cap\overline{\mathcal{B}}(\eta)}(\mathbf{K}_\varepsilon)$ boils down to the following problem

$$\min_{\mathbf{P}} \quad \sum_i \psi(\mathbf{P}_{i:}) - \langle \mathbf{P}, \nabla\psi(\mathbf{K}_\sigma)\rangle \tag{2}$$

$$\text{s.t.} \quad \sum_i \psi(\mathbf{P}_{i:}) \leq \eta \tag{3}$$

$$\mathbf{P}\mathbf{1} = \boldsymbol{a}, \quad \mathbf{P}^\top\mathbf{1} = \boldsymbol{b} \tag{4}$$

$$\mathbf{P} \in \mathbb{R}_+^{n\times n}\,. \tag{5}$$

This problem is convex and strictly feasible. Strong duality holds thanks to Slater's constraint qualification. Therefore the KKT conditions [5] are necessary and sufficient conditions for optimality. The Lagrangian can be expressed as

$$\mathcal{L}(\mathbf{P}, \nu, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_i \psi(\mathbf{P}_{i:}) - \langle \mathbf{P}, \nabla\psi(\mathbf{K}_\sigma)\rangle + \nu\Big(\sum_i \psi(\mathbf{P}_{i:}) - \eta\Big) \tag{6}$$

$$+ \langle \boldsymbol{\lambda}, \boldsymbol{a} - \mathbf{P}\mathbf{1}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} - \mathbf{P}^\top\mathbf{1}\rangle - \langle \boldsymbol{\Omega}, \mathbf{P}\rangle\,. \tag{7}$$

Any optimal primal-dual variables $(\mathbf{P}^\star, \nu^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\Omega}^\star)$ satisfies

$$\nabla_{\mathbf{P}}\mathcal{L}(\mathbf{P}^\star, \nu^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) = (\nu^\star + 1)\nabla\psi(\mathbf{P}^\star) - \nabla\psi(\mathbf{K}_\sigma) - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star - \boldsymbol{\Omega}^\star = \mathbf{0}\,. \tag{8}$$

Note that by definition $\mathbf{K}_\sigma = \nabla\psi^\star(-\mathbf{C}/\epsilon)$ and thus $\nabla\psi(\mathbf{K}_\sigma) = \nabla\psi[\nabla\psi^\star(-\mathbf{C}/\sigma)] = -\mathbf{C}/\sigma$ [23]. Thus we have

$$\mathbf{C} + \sigma(\nu^\star + 1)\nabla\psi(\mathbf{P}^\star) - \sigma\boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star - \sigma\boldsymbol{\Omega}^\star = \mathbf{0}\,. \tag{9}$$

Similarly, for the optimal transport problem (1)

$$\min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{C}\rangle \quad \text{s.t.} \quad \mathbf{T} \in \Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \overline{\mathcal{B}}(\eta)\,, \tag{10}$$

we get the optimality condition for optimal primal-dual variables $(\mathbf{T}^\star, \varepsilon^\star, \boldsymbol{\rho}^\star, \boldsymbol{\kappa}^\star, \boldsymbol{\Lambda}^\star)$

$$\mathbf{C} + \varepsilon^\star\nabla\psi(\mathbf{T}^\star) - \boldsymbol{\rho}^\star \oplus \boldsymbol{\kappa}^\star - \boldsymbol{\Lambda}^\star = \mathbf{0}\,. \tag{11}$$

Focusing on the original Bregman projection problem, we can then consider

$$\begin{cases} \boldsymbol{\lambda} &= \boldsymbol{\rho}^\star/\sigma \\ \boldsymbol{\mu} &= \boldsymbol{\kappa}^\star/\sigma \\ \nu &= \varepsilon^\star/\sigma - 1 \\ \boldsymbol{\Omega} &= \boldsymbol{\Lambda}^\star/\sigma \\ \mathbf{P} &= \mathbf{T}^\star\,. \end{cases} \tag{12}$$

With the above choice,

- $(\mathbf{P}, \nu, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Omega})$ satisfies the first order optimality condition.
- $\mathbf{P} = \mathbf{T}^\star \in \Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \mathcal{E}(\eta)$ thus the primal constraint is satisfied.
- $\sigma \leq \varepsilon^\star$ implies that $\nu \geq 0$ and $\boldsymbol{\Omega}$ has positive entries thereby dual constraints are satisfied.
- $\varepsilon^\star \neq 0$ thus by complementary slackeness $\psi(\mathbf{T}^\star) = \eta$ hence complementary slackness is also verified for $\mathbf{P}$ since $\mathbf{P} = \mathbf{T}^\star$.

Therefore the KKT conditions are met hence $(\mathbf{P}, \nu, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\Omega}) = (\mathbf{P}^\star, \nu^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\Omega}^\star)$ and $\mathbf{P}^\star = \mathbf{T}^\star$. □

<u>Part II : Proof of dual ascent.</u>

The optimal dual variables $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \varepsilon^\star)$ solve the following problem

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon > 0} \min_{\mathbf{P} \geq \mathbf{0}} \langle \mathbf{P}, \mathbf{C}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{a} - \mathbf{P}\mathbf{1}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} - \mathbf{P}^\top\mathbf{1}\rangle + \varepsilon\Big(\sum_i \psi(\mathbf{P}_{i:}) - \eta\Big) \tag{13}$$

$$= \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon > 0} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle - \varepsilon\eta + \min_{\mathbf{P} \geq \mathbf{0}} \langle \mathbf{P}, \mathbf{C} - \boldsymbol{\lambda}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\mu}^\top\rangle + \varepsilon\sum_i \psi(\mathbf{P}_{i:}) \tag{14}$$

$$= \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon > 0} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle - \varepsilon\eta + \min_{\mathbf{P} \geq \mathbf{0}} \sum_i \langle \mathbf{P}_{i:}, \mathbf{C}_{i:} - \lambda_i\mathbf{1} - \boldsymbol{\mu}\rangle + \varepsilon\psi(\mathbf{P}_{i:}) \tag{15}$$

$$\overset{(\star)}{=} \max_{\boldsymbol{\lambda}, \boldsymbol{\mu}, \varepsilon > 0} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle + \varepsilon\Big(\sum_i \psi^*((\mathbf{C}_{i:} - \lambda_i\mathbf{1} - \boldsymbol{\mu})/\varepsilon) - \eta\Big)\,. \tag{Dual-ROT}$$

In $(\star)$ we have used that $\psi^*(\boldsymbol{x}) = \sup_{\boldsymbol{y} \geq \mathbf{0}} \langle \boldsymbol{x}, \boldsymbol{y}\rangle - \psi(\boldsymbol{y})$. From Danskin's theorem [10], one can recover the solution of the primal

$$\forall i,\ \mathbf{P}^\star_{i:} = \nabla\psi^*((\mathbf{C}_{i:} - \lambda_i^\star\mathbf{1} - \boldsymbol{\mu}^\star)/\varepsilon^\star)\,. \tag{16}$$

Using matrix notations yields $\mathbf{P}^\star = \nabla\psi^*((\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)/\varepsilon^\star)$ where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \varepsilon^\star)$ are the solution of the dual problem (Dual-ROT).

B.2 OT with Pointwise Constraints on Either Sources or Targets (OTARI-s and OTARI-t) : proof of Proposition 2

Proposition 2. Let $(\boldsymbol{a}, \boldsymbol{b}, \xi)$ be such that $\Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \mathcal{B}_\psi(\xi)$ has an interior point and let $\mathbf{P}^\star$ solve

$$\min_{\mathbf{P} \in \Pi(\boldsymbol{a}, \boldsymbol{b})} \langle \mathbf{P}, \mathbf{C}\rangle \quad \text{s.t.} \quad \mathbf{P} \in \mathcal{B}_\psi(\xi)\,. \tag{OTARI-s}$$

Let $\boldsymbol{\varepsilon}^\star$ be the optimal dual variable associated with the constraint $\mathbf{P} \in \mathcal{B}_\psi(\xi)$. If $\boldsymbol{\varepsilon}^\star > \mathbf{0}$, then it holds $\mathbf{P}^\star = \mathrm{Proj}^{D_\psi}_{\Pi(\boldsymbol{a}, \boldsymbol{b}) \cap \mathcal{B}_\psi(\xi)}(\mathbf{K}_\varepsilon)$ for any $0 < \varepsilon \leq \min_i \varepsilon_i^\star$. Moreover it holds

$\mathbf{P}^\star = \nabla\psi^*\left(\mathrm{diag}(\boldsymbol{\varepsilon}^\star)^{-1}(\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)\right)$ where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\varepsilon}^\star)$ solve the following dual

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle + \left\langle \boldsymbol{\varepsilon}, \psi^*\left(\mathrm{diag}(\boldsymbol{\varepsilon})^{-1}(\mathbf{C} - \boldsymbol{\lambda} \oplus \boldsymbol{\mu})\right) - \psi(\mathbf{e}_\xi)\mathbf{1}\right\rangle . \quad \text{(Dual-OTARI-s)}$$

Proof. Again we break down the proof, focusing on the primal and then on the dual approach.

Part I : Proof of the Bregman projection.

The proof is almost identical to the one in Proposition 1. We use the same notations for simplicity. The only difference brought by the pointwise constraint is that $\boldsymbol{\nu}^\star$ is now vectorial. The first-order optimality condition for the Bregman projection problem reads

$$\mathbf{C} + \sigma(\mathrm{diag}\,(\boldsymbol{\nu}^\star) + \mathbf{I}_{N_S})\nabla\psi(\mathbf{P}^\star) - \sigma\boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star - \sigma\boldsymbol{\Omega}^\star = \mathbf{0}\,. \tag{17}$$

Again using the same notations as before, the first order KKT condition for problem (ROT) reads

$$\mathbf{C} + \mathrm{diag}(\boldsymbol{\varepsilon}^\star)\nabla\psi(\mathbf{T}^\star) - \boldsymbol{\rho}^\star \oplus \boldsymbol{\kappa}^\star - \boldsymbol{\Lambda}^\star = \mathbf{0}\,. \tag{18}$$

We end the proof by following the same reasoning as for Proposition 1, choosing for any i, $\nu_i = \varepsilon_i^\star/\sigma - 1 \geq 0$.

Part II : Dual Problem of (OTARI-s)**.**

The optimization problem (OTARI-s) writes

$$\min_{\mathbf{P}\in\Pi(\boldsymbol{a},\boldsymbol{b})} \langle \mathbf{P}, \mathbf{C}\rangle \quad \text{s.t.} \quad \mathbf{P} \in \mathcal{B}_\psi(\xi)\,. \tag{19}$$

Introducing the dual variables $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\boldsymbol{\mu} \in \mathbb{R}^n$ for the marginals and $\boldsymbol{\varepsilon} \in \mathbb{R}_+^n$ for the constraint $\mathbf{P} \in \mathcal{B}_\psi(\xi)$. The problem can be formulated as

$$\min_{\mathbf{P}\geq\mathbf{0}} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}\geq\mathbf{0}} \langle \mathbf{P}, \mathbf{C}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{a} - \mathbf{P}\mathbf{1}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} - \mathbf{P}^\top\mathbf{1}\rangle + \langle \boldsymbol{\varepsilon}, \psi(\mathbf{P}) - \psi(\mathbf{e}_\xi)\mathbf{1}\rangle\,. \tag{20}$$

When $\boldsymbol{\varepsilon}^\star > \mathbf{0}$, relying on strong duality to invert the min and max operators, the problem reduces to

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \min_{\mathbf{P}\geq\mathbf{0}} \langle \mathbf{P}, \mathbf{C}\rangle + \langle \boldsymbol{\lambda}, \boldsymbol{a} - \mathbf{P}\mathbf{1}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b} - \mathbf{P}^\top\mathbf{1}\rangle + \langle \boldsymbol{\varepsilon}, \psi(\mathbf{P}) - \psi(\mathbf{e}_\xi)\mathbf{1}\rangle \tag{21}$$

$$= \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle - \langle \boldsymbol{\varepsilon}, \psi(\mathbf{e}_\xi)\mathbf{1}\rangle + \min_{\mathbf{P}\geq\mathbf{0}} \langle \mathbf{P}, \mathbf{C} - \boldsymbol{\lambda}\mathbf{1}^\top - \mathbf{1}\boldsymbol{\mu}^\top\rangle + \langle \boldsymbol{\varepsilon}, \psi(\mathbf{P})\rangle \tag{22}$$

$$= \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle - \langle \boldsymbol{\varepsilon}, \psi(\mathbf{e}_\xi)\mathbf{1}\rangle + \min_{\mathbf{P}\geq\mathbf{0}} \sum_i \langle \mathbf{P}_{i:}, \mathbf{C}_{i:} - \lambda_i\mathbf{1} - \boldsymbol{\mu}\rangle + \varepsilon_i\psi(\mathbf{P}_{i:}) \tag{23}$$

$$= \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle - \langle \boldsymbol{\varepsilon}, \psi(\mathbf{e}_\xi)\mathbf{1}\rangle + \sum_i \varepsilon_i\psi^*((\mathbf{C}_{i:} - \lambda_i\mathbf{1} - \boldsymbol{\mu})/\varepsilon_i) \tag{24}$$

$$\overset{(\star)}{=} \max_{\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\varepsilon}>\mathbf{0}} \langle \boldsymbol{\lambda}, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}, \boldsymbol{b}\rangle + \left\langle \boldsymbol{\varepsilon}, \psi^*\left((\mathbf{C} - \boldsymbol{\lambda} \oplus \boldsymbol{\mu}) \oslash \boldsymbol{\varepsilon}\mathbf{1}^\top\right) - \psi(\mathbf{e}_\xi)\mathbf{1}\right\rangle . \quad \text{(Dual-OTARI-s)}$$

In $(\star)$ we used that $\psi^*(\mathbf{X}) = (\psi^*(\mathbf{X}_{1:}), ..., \psi^*(\mathbf{X}_{N:}))^\top$. From Danskin's theorem [10], one can recover the solution of the primal

$$\mathbf{P}^\star = \nabla\psi^*\left(\mathrm{diag}(\boldsymbol{\varepsilon}^\star)^{-1}(\mathbf{C} - \boldsymbol{\lambda}^\star \oplus \boldsymbol{\mu}^\star)\right) \tag{25}$$

where $(\boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star, \boldsymbol{\varepsilon}^\star)$ solve (Dual-OTARI-s). □

C ψ-Bregman Projections for ψ_{KL} and ψ_2

In this section, we detail the expressions of the projections used in the alternating Bregman projection approach.

C.1 KL Projections

Proposition 3. When D_ψ is the KL divergence D_{KL}, one has for a matrix $\mathbf{K} \in \mathbb{R}_+^{N_S\times N_T}$,

$$\mathrm{Proj}^{\mathrm{KL}}_{\Pi(\boldsymbol{a})\cap\mathcal{B}_{\mathrm{KL}}(\xi)}(\mathbf{K}) = \mathrm{diag}(\boldsymbol{\Lambda}\mathbf{1})^{-1}\boldsymbol{\Lambda} \quad \text{with} \quad \boldsymbol{\Lambda} = \exp\left(\mathrm{diag}(\mathbf{1} + \boldsymbol{\gamma}^\star)^{-1}\log\mathbf{K}\right) \tag{26}$$

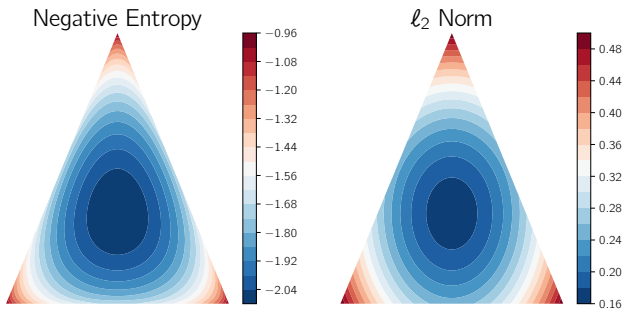

Figure 3: $\sum_i \psi(p_i)$ plotted over the 3 dimensional probability simplex for ψ_{KL} (negative Shannon entropy) and $\psi_2 : \boldsymbol{x} \to \frac{1}{2}\|\boldsymbol{x}\|_2^2$. Unlike ψ_{KL}, the level sets of ψ_2 intercept with the boundaries of the simplex thus leading to potentially sparse solutions when used to regularize OT.

where $\boldsymbol{\gamma}^\star \geq \mathbf{0}$ is the optimal dual variable associated with the constraint $\mathcal{B}_{\mathrm{KL}}(\xi)$.

Proof. The KL projection of a matrix $\mathbf{K} \in \mathbb{R}_+^{N_S \times N_T}$ onto $\Pi(\boldsymbol{a}) \cap \mathcal{B}_1(\xi)$ is the following problem.

$$\min_{\mathbf{P} \in \mathbb{R}_+^{N_S \times N_T}} \quad \mathrm{KL}(\mathbf{P}|\mathbf{K}) = \langle \mathbf{P}, \log(\mathbf{P} \oslash \mathbf{K}) - \mathbf{1}\mathbf{1}^\top \rangle \tag{27}$$

$$\text{s.t.} \quad \forall i \in [\![N_S]\!],\ \mathrm{H}(\mathbf{P}_{i:}) \geq \eta \tag{28}$$

$$\mathbf{P}\mathbf{1} = \boldsymbol{a}\,. \tag{29}$$

where for $\boldsymbol{p} \in \mathbb{R}_+^{N_S}$, $\mathrm{H}(\boldsymbol{p}) = -\langle \boldsymbol{p}, \log \boldsymbol{p} - \mathbf{1} \rangle$ is the Shannon entropy. The associated Lagrangian writes

$$\mathcal{L}(\mathbf{P}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \langle \mathbf{P}, \log \mathbf{P} - \log \mathbf{K} - \mathbf{1}\mathbf{1}^\top \rangle + \langle \boldsymbol{\gamma}, \eta \mathbf{1} - \mathrm{H}(\mathbf{P}) \rangle + \langle \boldsymbol{\lambda}, \boldsymbol{a} - \mathbf{P}\mathbf{1} \rangle\,. \tag{30}$$

Strong duality holds hence any optimal primal-dual variables $(\mathbf{P}^\star, \boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star)$ must satisfy the KKT conditions. The first-order optimality condition gives

$$\nabla_{\mathbf{P}} \mathcal{L}(\mathbf{P}^\star, \boldsymbol{\gamma}^\star, \boldsymbol{\lambda}^\star) = \log\left(\mathbf{P}^\star \oslash \mathbf{K}\right) + \mathrm{diag}(\boldsymbol{\gamma}^\star) \log \mathbf{P}^\star - \boldsymbol{\lambda}^\star \mathbf{1}^\top = \mathbf{0}\,. \tag{31}$$

Isolating $\mathbf{P}^\star$ yields

$$\forall (i,j) \in [\![N_S]\!] \times [\![N_T]\!], \quad P_{ij}^\star = \frac{1}{u_i} \exp\left((\log K_{ij})/(1+\gamma_i^\star)\right) \tag{32}$$

where $u_i = \exp\left(-\lambda_i/(1+\gamma_i^\star)\right)$. Given the marginal constraint, we have

$$u_i = a_i^{-1} \sum_{j \in [\![N_T]\!]} \exp\left((\log K_{ij})/(1+\gamma_i^\star)\right)\,. \tag{33}$$

We are now left with $\mathbf{P}^\star$ as a function of $\boldsymbol{\gamma}$. Plugging $\mathbf{P}^\star$ in $\mathcal{L}$ yields the dual function $\boldsymbol{\gamma} \mapsto \mathcal{G}(\boldsymbol{\gamma})$. This function is concave (property of the dual problem) and its gradient reads:

$$\nabla_{\boldsymbol{\gamma}} \mathcal{G}(\boldsymbol{\gamma}) = (\log \xi + 1)\mathbf{1} - \mathrm{H}(\mathbf{P}^\star(\boldsymbol{\gamma}))\,. \tag{34}$$

Similarly to [25], one can show that the above gradient is canceled for a unique $\overline{\boldsymbol{\gamma}}$. The optimal dual variable is then given by $\boldsymbol{\gamma}^\star = [\overline{\boldsymbol{\gamma}}]_+$.

□

C.2 Euclidean Projections

For the Euclidean case, we break down the projection into ℓ_2 norm and marginal projections. Starting with the marginal projection, we have the following expression [18]

$$\mathrm{Proj}_{\Pi(\boldsymbol{a})}^{\ell_2}(\mathbf{K}) = [\boldsymbol{\lambda}^\star \mathbf{1}^\top + \mathbf{K}]_+ \tag{35}$$

where $\boldsymbol{\lambda}^\star$ is such that $[\boldsymbol{\lambda}^\star \mathbf{1}^\top + \mathbf{K}]_+ \in \Pi(\boldsymbol{a})$.

Focusing on the ℓ_2 norm we have the following result.

Proposition 4. One has

$$\mathrm{Proj}^{\ell_2}_{\mathcal{B}_2(\xi)}(\mathbf{K}) = \mathrm{diag}(\boldsymbol{\gamma}^\star)^{-1}\mathbf{K} \tag{36}$$

where for any i, $\gamma_i^\star = \max\left(\xi^{1/2}\|\mathbf{K}_{i:}\|_2, 1\right)$.

Proof. The D_2 projection of a matrix $\mathbf{K} \in \mathbb{R}_+^{N_S \times N_T}$ onto $\mathcal{B}_2(\xi)$ reduces to

$$\min_{\mathbf{P}\in\mathbb{R}_+^{N_S\times N_T}} \quad D_2(\mathbf{P}|\mathbf{K}_\varepsilon) = \frac{1}{2}\left\langle \mathbf{P}^{\odot 2}, \mathbf{1}\right\rangle - \langle \mathbf{P}, \mathbf{K}\rangle \tag{37}$$

$$\text{s.t.} \quad \forall i \in [\![N_S]\!],\ \|\mathbf{P}_{i:}\|_2^2 \leq (1/\xi)\,. \tag{38}$$

Introducing the dual variable $\boldsymbol{\omega} \in \mathbb{R}_+^n$, the Lagrangian writes:

$$\mathcal{L}(\mathbf{P}, \boldsymbol{\omega}, \boldsymbol{\Omega}) = \frac{1}{2}\left\langle \mathbf{P}^{\odot 2}, \mathbf{1}\right\rangle - \langle \mathbf{P}, \mathbf{K}\rangle + \frac{1}{2}\sum_i \omega_i \left(\|\mathbf{P}_{i:}\|_2^2 - (1/\xi)\right)\,. \tag{39}$$

$\mathbf{P}^\star$ solves the primal problem if and only if there exists $\boldsymbol{\omega}^\star$ that satisfies the KKT conditions. The first-order condition yields

$$\nabla_{\mathbf{P}}\mathcal{L}(\mathbf{P}^\star, \boldsymbol{\omega}^\star, \boldsymbol{\Omega}^\star) = -\mathbf{K} + \mathrm{diag}(\boldsymbol{\omega}^\star + \mathbf{1})\mathbf{P}^\star = \mathbf{0}\,. \tag{40}$$

Hence it follows

$$\mathbf{P}^\star = \mathrm{diag}(\boldsymbol{\omega}^\star + \mathbf{1})^{-1}\mathbf{K}\,. \tag{41}$$

To satisfy the KKT conditions, one has to find the root $\boldsymbol{\omega}^\star$ of the following independent problems

$$\forall i, \quad (\omega_i + 1)^2 = \xi\|\mathbf{K}_{i:}\|_2^2\,. \tag{42}$$

Thus we have $\omega_i + 1 = \xi^{1/2}\|\mathbf{K}_{i:}\|_2$ and taking $\omega_i \geq 0$ into account yields the result. □