A UNIFIED DUAL METHOD FOR MATCHING PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matching problems are ubiquitous in data science as they enable the alignment of structured objects and distributions. While existing solvers are often tailored to specific matching formulations, we unify a broad class of such problems within a common mathematical and optimization framework based on duality theory. Theoretically, we demonstrate that matching objectives decomposable as a difference of convex (DC) functions can be recast as implicit registration problems. This connection links matching to another well-studied class of objectives and yields a dual formulation amenable to natural optimization strategies. We then apply these findings to quadratic matching (QM) problems, which admit DC decompositions and for which we provide extensive convergence guarantees. Our framework applies to Gromov-Wasserstein (GW), as well as its unbalanced formulation and several variants, which are increasingly popular QM problems. Numerically, we implement our algorithms at scale for various data modalities such as graphs, point clouds, meshes, and word embeddings. Finally, our modular approach allows us to explore new formulations such as fracture matching, broadening the scope of problems that can be addressed within this framework.

1 INTRODUCTION

Many objects studied in data science and imaging, such as point clouds, shapes, graphs, physical variables, or probability distributions, are naturally spread across space. Comparing, aligning, and transferring information between such objects requires to establish reliable spatial correspondences between such inputs. To this end, two approaches are available: registration, which searches for optimal transforms of one source object to a target one, and matching, which draws pointwise correspondences between their constituent points. Although the former class of methods provides more regular output, it is also less flexible and is not applicable to all data modalities. The latter, on the other hand, is generally harder to solve algorithmically, and the lack of efficient solvers for arbitrary problems is still an important bottleneck for many downstream applications.

Arguably the most popular category of matching techniques, quadratic matching searches for matchings as solutions of a quadratic optimization problem, allowing the use of principled numeric routines while being sufficiently expressive to encode complex geometric objectives. Depending on the exact problem and the scientific community, quadratic matching takes different names: quadratic assignment (Lawler, 1963), Gromov–Wasserstein (Mémoli, 2011), graph matching (Livi & Rizzi, 2013), or quadratic-form optimal transport (Wang & Zhang, 2025). More custom QM instances have also been introduced for shape matching (Mandad et al., 2017; Brifault et al., 2025), evidencing the importance of this class of problems.

Yet, quadratic matching remains challenging to compute. Its non-convexity makes it NP-hard to minimize globally. Local optimization via first-order methods is possible (Frank & Wolfe, 1956), but it requires to manipulate large matching matrices explicitly which limits the scalability of this approach. A promising line of work exploit duality properties of quadratic matching to recast quadratic matching problems as a sequence of linear matching instances, or optimal transport (OT). OT is a popular and well-studied tool in data science, its success coming from the availability of efficient solvers (Peyré & Cuturi, 2019). Indeed, OT admits a dual formulation that implicitly encodes matching matrices in variables of lower size, removing its main computational bottleneck. OT is also a flexible framework, since regularizations can be added to the linear optimization problems to smooth

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or sparsify the final matching (Blondel et al., 2018), to ignore outliers (Chapel et al., 2020; Liero et al., 2018), or to enable fast and GPU-compatible algorithms (Cuturi, 2013). Building upon this connexion between quadratic matching and optimal transport, faster QM methods have been proposed in specific cases (Gold & Rangarajan, 1996; Peyré et al., 2016; Séjourné et al., 2023), and duality properties of some QM instances have been recently investigated (Rioux et al., 2024; Zhang et al., 2024; Sebbouh et al., 2024; Houry et al., 2026). However, these works still lack the generality required to solve any QM problem at scale.

Contribution. Focusing on the current limitations of the matching literature, we propose a new framework to solve any quadratic matching locally in an efficient and rigorous way. We first draw a conceptual link between matching and registration problems, revealing a hidden geometric structure in matching problems, and introduce a dual formula that reduce arbitrary matching problems to a sequence of OT computations. We then specialize these results to the quadratic case, allowing us to build an efficient solver with quantitative convergence guarantees that we apply in various practical settings. Beyond the methodological contributions, we release an optimized, backend-agnostic implementation that is modular by design and readily extensible to other regularized quadratic problems. While unbalanced Gromov-Wasserstein serves as our primary motivating application throughout the paper, the framework applies with minimal adaptation to a wider class of problems.

Notations. In this article, $\mathcal{X}$ and $\mathcal{Y}$ are two compact measurable metric spaces, $\mathcal{C}(\mathcal{X})$ denotes the space of continuous real-valued functions on $\mathcal{X}$, $\mathcal{M}(\mathcal{X})$ is the space of measures of $\mathcal{X}$, and $\mathcal{M}_+(\mathcal{X})$ is its restriction to positive measures. We recall that in the discrete setting, when $\mathcal{X} = \{x_1, \dots, x_{\mathrm{N}}\}$ and $\mathcal{Y} = \{y_1, \dots, y_{\mathrm{M}}\}$ are finite point sets, any measures $\alpha \in \mathcal{M}(\mathcal{X})$ and $\beta \in \mathcal{M}(\mathcal{Y})$ correspond to weight vectors of $\mathbb{R}^{\mathrm{N}}$ and $\mathbb{R}^{\mathrm{M}}$ and measures π of $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ correspond to matrices of $\mathbb{R}^{\mathrm{N}\times\mathrm{M}}$:

$$\alpha = \sum_{i=1}^{\mathrm{N}} \alpha_i \delta_{x_i}, \quad \beta = \sum_{j=1}^{\mathrm{M}} \beta_j \delta_{y_j} \quad \text{and} \quad \pi_{ij} = \sum_{i,j} \pi_{ij} \delta_{(x_i, y_j)}.$$

Let $A : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}_A$ be a continuous function with values in a Hilbert space $\mathcal{H}_A$. It naturally acts on $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ via its associated *integral operator*, defined as follows:

$$A : \pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \;\longmapsto\; A[\pi] := \int A(x,y)\, \mathrm{d}\pi(x,y) \in \mathcal{H}_A.$$

In the discrete case and when $\mathcal{H}_A = \mathbb{R}^{\mathrm{D}}$ is Euclidean, it becomes $[\sum_{ij} A(x_i, y_j)_d \pi_{ij}]_d \in \mathbb{R}^{\mathrm{D}}$. Given two Hilbert vectors $x \in \mathcal{H}_{\mathcal{X}}$ and $y \in \mathcal{H}_{\mathcal{Y}}$, their *tensor product* is the rank-1 linear operator:

$$xy^\top : z \in \mathcal{H}_{\mathcal{Y}} \longmapsto \langle y, z \rangle_{\mathcal{H}_{\mathcal{Y}}}\, x \;\; \in \mathcal{H}_{\mathcal{X}}.$$

The limit of sums of such operators forms a Hilbert space, the *Hilbert-Schmidt space* $\mathrm{HS}(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{X}})$. In the Euclidean case, the tensor product of two vectors $x \in \mathbb{R}^{\mathrm{D}}$ and $y \in \mathbb{R}^{\mathrm{E}}$ is $xy^\top = [x_d y_e]_{de} \in \mathbb{R}^{\mathrm{D}\times\mathrm{E}}$, and $\mathrm{HS}(\mathbb{R}^{\mathrm{D}}, \mathbb{R}^{\mathrm{E}})$ identifies with $\mathbb{R}^{\mathrm{D}\times\mathrm{E}}$ equipped with the Frobenius norm. Finally, given a convex function $\mathcal{F} : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, its *convex conjugate* $\mathcal{F}^*$ is defined as:

$$\mathcal{F}^* : c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y}) \longmapsto \sup_{\pi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})} \left(\int c(x,y) \mathrm{d}\pi(x,y) - \mathcal{F}(\pi) \right).$$

2 Measure Matching as an Implicit Registration Problem

To model both finite and continuous inputs in a common mathematical framework, our work focuses on matching from a measure-theoretical perspective. Given a source $\alpha \in \mathcal{M}_+(\mathcal{X})$ and a target $\beta \in \mathcal{M}_+(\mathcal{Y})$, we represent matchings from α to β as elements π of $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ – which we call *transport plans* by analogy with optimal transport. Each $\pi(x,y)$ indicates the amount of $\alpha(x)$ that is sent to $\beta(y)$. This framework is naturally compatible with matching relaxation methods where one point of α can be sent to several points of β, and conversely: one-to-one matching corresponds plans π that describe the graph of a bijective function of $\mathcal{X} \to \mathcal{Y}$, or, in the finite setting, when π is a permutation matrix. Despite the variety of existing methods, most matching algorithms can be represented as variational problems in this measure setting, taking the form:

$$\mathrm{M}(\alpha, \beta) \;=\; \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \mathcal{E}_{\mathcal{R}}(\pi) \qquad \text{with} \qquad \mathcal{E}_{\mathcal{R}}(\pi) := \mathcal{E}(\pi) + \mathcal{R}_{\alpha,\beta}(\pi). \tag{1}$$

The continuous energy term $\mathcal{E}$ encodes the objective of the matching problem; an ideal matching should reach $\mathcal{E}(\pi) = 0$. The convex constraint term $\mathcal{R}_{\alpha,\beta}$ forces the matching to *send* α to β and may contain data-specific regularizations; when there is no ambiguity, we will keep the dependency on α, β implicit and simply write $\mathcal{R} := \mathcal{R}_{\alpha,\beta}$. To clarify this formalism, we use Gromov–Wasserstein (GW) as a running example throughout our article. GW promotes isometric matching by penalizing distortions between matched pairs. In the finite case, these distorsions are quantified by symmetric matrices $C^{\mathcal{X}}$ and $C^{\mathcal{Y}}$ so that GW reads:

$$\min_{\pi \in \mathbb{R}_+^{\mathrm{N}\times\mathrm{M}}} \frac{1}{2} \sum_{ijkl} (C^{\mathcal{X}}_{ik} - C^{\mathcal{Y}}_{jl})^2 \pi_{ij} \pi_{kl} \quad \text{s.t.} \quad \sum_j \pi_{ij} = \alpha_i, \quad \text{and} \quad \sum_i \pi_{ij} = \beta_j.$$

After encoding the constraints in the term $\mathcal{R}$, GW takes the form of eq. (1) with:

$$\mathcal{E}(\pi) = \frac{1}{2} \sum_{ijkl} (C^{\mathcal{X}}_{ik} - C^{\mathcal{Y}}_{jl})^2 \pi_{ij} \pi_{kl} \quad \text{and} \quad \mathcal{R}(\pi) = \begin{cases} 0 & \text{if } \sum_j \pi_{ij} = \alpha_i \text{ and } \sum_i \pi_{ij} = \beta_j \\ +\infty & \text{otherwise.} \end{cases}$$

In the continuous case, we replace matrix operations with their measure-based counterparts:

$$\mathcal{E}(\pi) = \frac{1}{2} \int \big(c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y')\big)^2 \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y')$$
$$\text{and} \quad \mathcal{R}(\pi) = \begin{cases} 0 & \text{if } \pi_1 := \int \mathrm{d}\pi(x, \cdot) = \mathrm{d}\alpha(x) \text{ and } \pi_2 := \int \mathrm{d}\pi(\cdot, y) = \mathrm{d}\beta(y) \\ +\infty & \text{otherwise,} \end{cases}$$

where $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are continuous symmetric functions. Classical GW variants add regularizations to $\mathcal{R}$ or relax the margins constraints: we give typical examples in section B.1. Notably, entropic GW adds a KL-divergence term to $\mathcal{R}$, and unbalanced GW relaxes the marginals using:

$$\mathcal{R}(\pi) = \rho D_{\mathrm{KL}}(\pi_1, \|\alpha) + \rho D_{\mathrm{KL}}(\pi_2 \| \beta) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta), \quad \rho, \varepsilon \geq 0. \tag{2}$$

Matching problems are generally hard to solve numerically. They are usually non-convex, making global solutions difficult to compute. Besides, their objectives are defined on the set of plans $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$. In the finite case, these are encoded by large matrices, making even local minimization expensive to perform. In specific cases, recent works have alleviated these issues by reformulating matching objectives into easier problems (Rioux et al., 2024; Zhang et al., 2024; Sebbouh et al., 2024; Houry et al., 2026). We generalize these works to arbitrary matching problems, focusing on the cases where $\mathcal{E}$ is decomposed into convex and concave terms:

$$\mathcal{E}(\pi) = \mathcal{A}(\pi) - \mathcal{B}(\pi), \tag{3}$$

where $\mathcal{A}, \mathcal{B}$ are proper lower semi-continuous convex functions and continuous on their domains. This structure is known as difference of convex (DC). Under mild assumptions, any matching problem is approximable by a DC one, and in the finite case, exact DC decompositions always exist. Without loss of generality, we can also assume the convex conjugates $\mathcal{A}^*$, $\mathcal{B}^*$ to be continuous on $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$; we formalize these properties in section B.2. The DC structure isolates the hard part of the problem – the concave term – so we can handle it separately, allowing us to prove the equivalence:

Theorem 1. *If $\mathcal{E}$ admits the structure of eq.* (3)*, then eq.* (1) *is equivalent to:*

$$\mathrm{M}(\alpha, \beta) \;=\; \inf_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} - \int c_{\mathcal{B}}(x, y) \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}(c_{\mathcal{B}}), \tag{4}$$

where $\mathbf{R}(\pi) = \mathcal{A} + \mathcal{R}$ and $\mathbf{S} = \mathcal{B}^*$ are convex functions. If $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean spaces, then there exists two Hilbert-valued continuous embeddings $\Phi : \mathcal{X} \to \mathcal{H}$ and $\Psi : \mathcal{Y} \to \mathcal{H}$, a convex set $U \subset \mathrm{HS}(\mathcal{H}, \mathcal{H})$ and a convex function $\mathbf{S}'$ such that eq. (1) *becomes:*

$$\mathrm{M}(\alpha, \beta) \;=\; \inf_{\Gamma \in U} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} - \int \langle \Phi(x), \Gamma \Psi(y) \rangle \, \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}'(\Gamma). \tag{5}$$

This structure is identical to another class of methods seeking correspondences between geomeric objects called *registration*. The principle of registration is to find a deformation of the source shape β that minimizes its spatial discrepancy relative to the target shape α. Given a space of transformations $\mathcal{T}$ and a real-valued loss function $\mathcal{L}$, standard registration problems are phrased as:

$$\min_{\Gamma \in \mathcal{T}} \; \min_{\pi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})} \frac{1}{2} \int \mathcal{L}(x, \Gamma(y)) \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma), \tag{6}$$

where $\mathbf{R}$ is convex and $\mathbf{S}$ is an arbitrary regularizer. Many popular algorithms follow this structure, such as iterative closest points (Besl & McKay, 1992), robust point matching (Gold et al., 1994), coherent point drift (Myronenko & Song, 2010), correlation-based registration (Tsin & Kanade, 2004), functional maps (Ovsjanikov et al., 2012), Procrustes-Wasserstein (Grave et al., 2019) and other OT-based methods (Alvarez-Melis et al., 2019). Registration is also non-convex, but non-convexity is isolated within the outer minimization over Γ: it provides a cleaner framework for analyzing and exploring optimization landscapes. Registration objectives are also more robust: unlike plans π, optimal transformations Γ depend only on the global input geometry, making them intrinsically robust to empirical noise, point resampling, or permutations, and opening the door to geometrically-driven heuristics. Thanks to our reformulation, we can exploit these properties and adapt the numerous existing tools for registration to the matching context.

Our reformulation also reveals a nested minimization structure whose local solutions can be obtained by alternating optimization over Γ and π. Since the problem is convex in π, this parameter can be updated using any off-the-shelf solver. Yet, this approach still requires explicit optimization over the plan π. In classical registration, the regularization $\mathbf{R}$ is typically designed to make this optimization cheap. However, in our setting, $\mathbf{R}$ is prescribed by the original objective $\mathcal{E}$ so we need a method that handles arbitrary regularizations on π. Taking inspiration from optimal transport, we adapt duality results for DC functions (Toland, 1978) to avoid manipulating π explicitly:

Theorem 2. *Let us assume that $\mathcal{R}$ is a proper lower semi-continuous convex function. Let Φ be its convex set of dual variables and $\mathcal{D}_\mathcal{R}$ its dual objective that satisfies, for all $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$:*

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c \mathrm{d}\pi + \mathcal{R}(\pi) = \sup_{\phi \in \Phi} \mathcal{D}_\mathcal{R}(\phi;\, c, \alpha, \beta). \tag{7}$$

The matching problem of eq. (3) *is then equivalent to the following equations:*

$$\inf_{c_\mathcal{B} \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \sup_{c_\mathcal{A} \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \sup_{\phi \in \Phi} \mathcal{D}_\mathcal{R}\big(\phi;\, c_\mathcal{A} - c_\mathcal{B},\, \alpha, \beta\big) + \mathcal{B}^*(c_\mathcal{B}) - \mathcal{A}^*(c_\mathcal{A}), \tag{8}$$

$$\inf_{c_\mathcal{B} \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \sup_{c_\mathcal{A} \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (c_\mathcal{A} - c_\mathcal{B}) \mathrm{d}\pi + \mathcal{R}(\pi) + \mathcal{B}^*(c_\mathcal{B}) - \mathcal{A}^*(c_\mathcal{A}). \tag{9}$$

If $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean spaces, then there exists a sequence of continuous Hilbert embeddings $\Phi_\mathcal{A}, \Phi_\mathcal{B} : \mathcal{X} \to \mathcal{H}$ and $\Psi_\mathcal{A}, \Psi_\mathcal{B} : \mathcal{Y} \to \mathcal{H}$ such that these equations become:

$$\begin{aligned} \mathrm{M}(\alpha, \beta) \;=\; \inf_{\Gamma \in U} \sup_{\Xi \in U} \sup_{\phi \in \Phi} \mathcal{G}(\Gamma, \Xi, \phi) \;=\; \inf_{\Gamma \in U} \sup_{\Xi \in U} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \mathcal{F}(\Gamma, \Xi, \phi), \\ \textit{with} \quad \mathcal{G}(\Gamma, \Xi, \phi) := \mathcal{D}_\mathcal{R}\big(\phi;\, c_{\Gamma,\Xi},\, \alpha, \beta\big) + \mathbf{S}'(\Gamma) - \mathbf{S}''(\Xi) \\ \textit{and} \quad \mathcal{F}(\Gamma, \Xi, \phi) := \int c_{\Gamma,\Xi} \mathrm{d}\pi + \mathcal{R}(\pi) + \mathbf{S}'(\Gamma) - \mathbf{S}''(\Xi), \end{aligned} \tag{10}$$

where $U \subset \mathrm{HS}(\mathcal{H}, \mathcal{H})$ is a convex set, $\mathbf{S}'$ and $\mathbf{S}''$ are convex functions and $c_{\Gamma,\Xi}$ is the following effective cost (where the matrix notation indicates the direct sum, or concatenation, of its rows):

$$c_{\Gamma,\Xi}(x, y) = -\left\langle \begin{pmatrix} \Phi_\mathcal{A}(x) \\ \Phi_\mathcal{B}(x) \end{pmatrix}, \begin{pmatrix} -\Xi\Psi_\mathcal{A}(y) \\ \Gamma\Psi_\mathcal{B}(y) \end{pmatrix} \right\rangle. \tag{11}$$

For classical choices of $\mathcal{R}$, either $\mathcal{R}$-regularized linear problems ($\min_\pi \int c\mathrm{d}\pi + \mathcal{R}(\pi)$) or their duals ($\sup_\phi \mathcal{D}_\mathcal{R}(\phi;\, c,\, \alpha, \beta)$) admit efficient solvers for any $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$. We refer to these problems as $\mathcal{R}\mathrm{OT}$, as they have the structure of a $\mathcal{R}$-regularized optimal transport. If efficient $\mathcal{R}\mathrm{OT}$ solvers are available, theorem 2 provides a strategy to solve the original matching efficiently. As a counterpart, the embedding space on which matching is performed is now infinite-dimensional and structurally complex: our reformulation transfers the difficulty of matching problems into the embeddings $\Phi_\mathcal{A}$, $\Phi_\mathcal{B}$, $\Psi_\mathcal{A}$ and $\Psi_\mathcal{B}$, the problem structure being now agnostic to the matching objective $\mathcal{E}$.

3 An Explicit Dual for Regularized Quadratic Matching

To convert the results of the previous section into practical numerical algorithms, we need to explicitly have access to the embeddings $\Phi_\mathcal{A}$, $\Phi_\mathcal{B}$, $\Psi_\mathcal{A}$ and $\Psi_\mathcal{B}$ which are generally unknown and

high-dimensional, limiting the applicability of our method to arbitrary matching problems. However, in the quadratic matching (QM), the DC decomposition is often simple and explicit, bypassing these limitations. In QM, objectives are parametrized by a symmetric kernel $k \in \mathcal{C}((\mathcal{X} \times \mathcal{Y})^2)$:

$$\mathcal{E}(\pi) = \frac{1}{2} \int k(x, y, x', y') \mathrm{d}\pi(x, y) \mathrm{d}\pi(x', y'). \tag{12}$$

When $\mathcal{E}$ follows this structure, we denote by $\mathcal{R}\mathrm{QM}(\alpha, \beta)$ the $\mathcal{R}$-regularized quadratic matching of eq. (1). For instance, Gromov-Wasserstein is a QM with $k(x, y, x', y') = (c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y'))^2$.

In the finite setting and several continuous cases that we detail in section B.2, k admits an explicit decomposition as a difference of quadratic functions. We refer to such k as factorizable kernels:

Definition 1 (Factorizable kernel)**.** *We say that k is factorizable if its objective is decomposable as:*

$$\mathcal{E}(\pi) = \frac{1}{2} \left\| A[\pi] \right\|^2 - \frac{1}{2} \left\| B[\pi] \right\|^2, \tag{13}$$

where $A : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}_A$ and $B : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}_B$ are integral operators.

Under this hypothesis, we directly adapt the formulas of the previous section to the $\mathcal{R}\mathrm{QM}$ case:

Theorem 3. *Let assume that k is factorizable as in eq.* (3)*. Then, the following identity holds:*

$$\mathcal{R}\mathrm{QM}(\alpha, \beta) = \min_{\Gamma \in \mathrm{Im}(B)} \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} - \int \langle \Gamma, B(x, y) \rangle \, \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma), \tag{14}$$

with $\mathbf{R}(\pi) = \frac{1}{2} \left\| A[\pi] \right\|^2 + \mathcal{R}(\pi)$ and $\mathbf{S}(\Gamma) = \frac{1}{2} \left\| \Gamma \right\|^2$. Moreover:

$$\mathcal{R}\mathrm{QM}(\alpha, \beta) = \min_{\Gamma \in \mathrm{Im}(B)} \max_{\Xi \in \mathrm{Im}(A)} \sup_{\phi \in \Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\Gamma, \Xi},\, \alpha, \beta\big) - \frac{1}{2} \left\| \Xi \right\|^2 + \frac{1}{2} \left\| \Gamma \right\|^2, \tag{15}$$

with the effective cost $c_{\Gamma, \Xi}(x, y) := \langle \Xi, A(x, y) \rangle - \langle \Gamma, B(x, y) \rangle$.

To recover the registration-like structure of eq. (5), we must also assume that B is block-separable:

Definition 2 (Separable function)**.** *A function $B : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}_B$ is separable if there exist two embedding functions Φ and Ψ such that:*

$$\text{For all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}, \quad B(x, y) = \Phi(x) \Psi(y)^\top.$$

We say that B is block-separable if there is a collection of separable functions $B_1, \dots, B_{\mathrm{L}}$ such that $\left\| B[\pi] \right\|^2 = \left\| B_1[\pi] \right\|^2 + \cdots + \left\| B_{\mathrm{L}}[\pi] \right\|^2$ for all $\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$.

Then, $\mathcal{R}\mathrm{QM}$ admits the following form (we give a detailed version in the appendix, in theorem 7):

Theorem 4. *If B is block-separable, then let $\mathbf{R}$ and $\mathbf{S}$ be as in theorem 3. There exists embeddings $\Phi_{\mathcal{B}} : \mathcal{X} \to \mathcal{H}_\Phi$, $\Psi_{\mathcal{B}} : \mathcal{Y} \to \mathcal{H}_\Psi$ and a compact subspace $U \subset \mathrm{HS}(\mathcal{H}_\Phi, \mathcal{H}_\Psi)$ such that:*

$$\mathcal{R}\mathrm{QM}(\alpha, \beta) = \min_{\Gamma \in U} \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} - \int \langle \Phi_{\mathcal{B}}(x), \Gamma \Psi_{\mathcal{B}}(y) \rangle \, \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma). \tag{16}$$

If both A and B are block-separable, there exists embeddings $\Phi_{\mathcal{A}}, \Phi_{\mathcal{B}}, \Psi_{\mathcal{A}}$ and $\Phi_{\mathcal{B}}$ such that the effective cost $c_{\Gamma, \Xi}$ in eq. (10) *takes the scalar product form of eq.* (11) *and $\mathbf{S}', \mathbf{S}'' = \frac{1}{2} \left\| \cdot \right\|^2$.*

If we have access to a block-sparse decomposition of the kernel k, then the embeddings are explicitly computable and theorems 3 and 4 provide a computation-compatible version of theorem 2. In the finite case, this can be done by eigendecomposition of k. For GW kernels, explicit computations that we detail in section B.4 provide block embeddings directly without the need to diagonalize the whole kernel k, making their computation tractable on large inputs. Moreover, these embeddings are now related to the images of A and B: when these operators are low-rank, which is generally the case in practice, the optimization problem becomes low-dimensional and easier to solve.

4 Numerical Optimization of the $\mathcal{R}\mathrm{QM}$ Dual

We now exploit these theoretical results to develop a unified algorithm for $\mathcal{R}\mathrm{QM}$ with quantitative convergence guarantees. Up to a factorization of the kernel k, always possible in the finite case, our method applies to any QM instance and most standard choices of $\mathcal{R}$, making the proposed framework applicable in a broad range of settings. The algorithm proceeds by alternate minimization over the block of variables of eq. (10): Γ, Ξ and either Φ or π. Since our method is compatible with both primal and dual $\mathcal{R}\mathrm{OT}$ solvers, we can choose whether to optimize over Φ or π depending on which formulation is computationally more efficient. Each block is updated as follows:

ϕ/π-block: $\mathcal{R}$OT solve. For fixed (Γ, Ξ), maximizing $\mathcal{F}$ over ϕ or $\mathcal{G}$ over π amounts to solving a primal or dual $\mathcal{R}$OT problem with cost $c_{\Gamma,\Xi}$. For classical $\mathcal{R}$, any compatible solver can be applied.

Ξ-block: gradient ascent. When Γ is fixed, the function $h_\Gamma : \Xi \mapsto \min_\pi \mathcal{G}(\Gamma, \Xi, \pi)$ is 1-strongly concave. If the optimal $\pi_{\Gamma,\Xi}$ is unique, then h_Γ is differentiable at Ξ with $\nabla h_\Gamma(\Xi) = A[\pi_{\Gamma,\Xi}] - \Xi$. The block Ξ can then be optimized by alternating $\mathcal{R}$OT solves and gradient ascent steps.

Γ-block: closed form. Finally, if both Ξ and π are fixed, then eq. (14) shows that $\Gamma \mapsto \mathcal{G}(\Gamma, \Xi, \pi)$ is minimized at $\Gamma = B[\pi]$: the Γ-block is updated using this closed-form minimum.

By design, for well-chosen gradient steps in the Ξ-block, each update improves the current block objective value: this approach is guaranteed to converge. In practice, however, many $\mathcal{R}$OT solvers only provide an approximation of the true solution (ϕ, π), and gradient ascent on Ξ never reaches its true maximum in finite time: we must set tolerance thresholds to indicate when these computations should be stopped, yielding inexact block updates. Our alternating minimization scheme is summarized in algorithm 1. Despite the inexact updates, the stopping criterion imposed on Ξ still guarantees finite-time convergence under appropriate assumptions. We state this result informally here, providing full statements and detailed estimates in section E.

Theorem 5 (Convergence of algorithm 1, informal)**.** *Let $0 < \delta \leq 1$. **(i)** If the inner loops converge, then algorithm 1 terminates after $O(\delta^{-2})$ outer iterations. **(ii)** If the function h_Γ is L-smooth for all Γ, then the choice $\tau = 1/L$ and $\delta_{\mathcal{R}\mathrm{OT}} = \Theta(\delta^2/L^2)$ ensures that each inner loop converges in $\widetilde{O}(L)$ iterations, for a total of $\widetilde{O}(L\delta^{-2})$ cumulated $\mathcal{R}$OT solvings steps.*

Characterization of output plans. Theorem 5 ensures that our algorithm terminates but does not qualify the properties of the final plan $\widehat{\pi}$. By non-convexity of $\mathcal{R}$QM, we cannot expect $\widehat{\pi}$ to minimize the problem globally: as usual in DC programming (Le Thi & Pham Dinh, 2018), we will rather show that $\widehat{\pi}$ is almost critical. A plan π^* is *DC-critical* if it is optimal after linearization of the concave term $\pi \mapsto -\frac{1}{2}\|B[\pi]\|^2$ in the $\mathcal{R}$QM objective: that is, if for all $\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$,

$$\tfrac{1}{2}\|A[\pi^*]\|^2 - \langle B[\pi^*], B[\pi^*]\rangle + \mathcal{R}(\pi^*) \leq \tfrac{1}{2}\|A[\pi]\|^2 - \langle B[\pi^*], B[\pi]\rangle + \mathcal{R}(\pi), \quad \text{or equivalently :}$$
$$\tfrac{1}{2}\|A[\pi^*]\|^2 - \tfrac{1}{2}\|B[\pi^*]\|^2 + \mathcal{R}(\pi^*) \leq \tfrac{1}{2}\|A[\pi]\|^2 - \tfrac{1}{2}\|B[\pi]\|^2 + \mathcal{R}(\pi) + \tfrac{1}{2}\|B[\pi - \pi^*]\|^2.$$

The last inequality is approximately satisfied by $\widehat{\pi}$, its discrepancy controlled by the tolerance δ:

Theorem 6. *If $\delta_{\mathcal{R}\mathrm{OT}} \leq \delta^2/4$, and if the inner loops converge, the output of algorithm 1 is a δ-critical plan in the following sense (recalling that $\mathcal{E}_\mathcal{R} = \mathcal{E} + \mathcal{R}$ denotes the matching objective):*

$$\text{For all } \pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}), \quad \mathcal{E}_\mathcal{R}(\widehat{\pi}) \;\leq\; \mathcal{E}_\mathcal{R}(\pi) + \tfrac{1}{2}\|B[\pi - \widehat{\pi}]\|^2 + \delta\|B[\pi - \widehat{\pi}]\| + \delta^2/4.$$

Practical example: entropic unbalanced GW. Theorem 5 requires h_Γ to be L-smooth in order to ensure end-to-end convergence. This hypothesis does not hold for all choice of $\mathcal{R}$, but is satisfied whenever $\mathcal{R}$ contains an entropic term $\varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$, such as in the unbalanced case of eq. (2): then, $L = 1 + O(\varepsilon^{-1})$. Corresponding $\mathcal{R}$OT problems can be solved with the Sinkhorn algorithm, an iterative method that improves the dual variables ϕ at each step (Cuturi, 2013; Chizat et al., 2018). Relying on the Sinkhorn algorithm has several practical advantages. First, it is a dual solver that does not manipulate plans π explicitly, improving memory usage. Besides, Sinkhorn converges linearly: exploiting this property, we further prove in section F that algorithm 1 terminates after $\widetilde{O}((\rho + \varepsilon)/(\delta^2\varepsilon^2))$ single Sinkhorn iterations. The linear convergence of Sinkhorn also permits to replace the Ξ-updates by Nesterov's accelerated scheme with an inexact, but arbitrarily close, oracle (Devolder et al., 2013). This acceleration scheme can save a factor $\varepsilon^{-1/2}$ in the end-to-end complexity; see corollary 2 for details and conditions of use. Finally, since we have an estimate of the smoothness in this case, we can calibrate the step size τ automatically without having to tune this hyperparameter manually. This evidences the benefits of entropic regularization for QM, and allows us to build optimized end-to-end pipelines for unbalanced GW based on the Sinkhorn algorithm.

Warm starts and early stopping. The stopping criteria of algorithm 1 are theoretically-grounded but rather conservative: since (Γ, Ξ) varies slowly between consecutive $\mathcal{R}$OT solves, warm-starting the solver at its previous dual potentials makes a few solver iterations per Ξ- and ϕ-update – or even a fixed number of them – sufficient in practice. Empirically, in the entropic case, our method still

Algorithm 1: $\mathcal{R}$QM solver

Input: measures α, β; operators A, B; $\mathcal{R}$OT solver; tolerances δ, $\delta_{\mathcal{R}\mathrm{OT}}$; step size τ.
Output: a $(\delta, \delta^2/4)$-critical plan $\widehat{\pi}$ (theorem 5), with its variables (Γ, Ξ, ϕ).
Initialize Γ_0, Ξ and ϕ
for $t = 0, 1, 2, \dots$ **do**
  for $k = 0, 1, 2, \dots$ **do**
    $(\phi, \widehat{\pi}) \leftarrow \mathcal{R}$OT solver (cost $c_{\Gamma_t, \Xi}$, warm start at ϕ, accuracy $\delta_{\mathcal{R}\mathrm{OT}}$) `// ϕ-block`
    if $\frac{1}{2}\left\|A[\widehat{\pi}] - \Xi\right\|^2 + \delta_{\mathcal{R}\mathrm{OT}} \leq \delta^2/4$ **then break**
    $\Xi \leftarrow \Xi + \tau\left(A[\widehat{\pi}] - \Xi\right)$ `// Ξ-block`
  $\Gamma_{t+1} \leftarrow B[\widehat{\pi}]$ `// Γ-block`
  if $\|\Gamma_{t+1} - \Gamma_t\| \leq \delta$ **then return** $\widehat{\pi}$

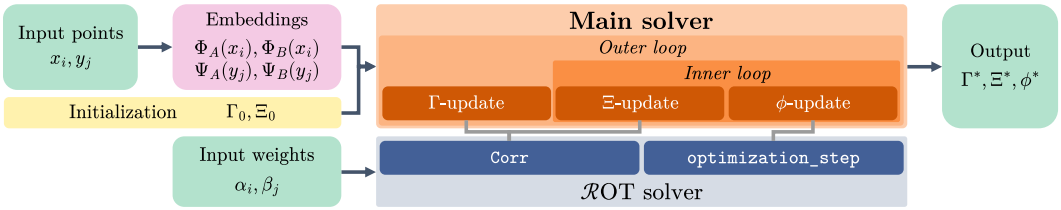

Figure 1: High-level description of our QM method for block-separable kernels. To apply our main solver to a specific $\mathcal{R}$QM problem, we only need to specifiy the block-separable embeddings of the objective (using eigendecomposition of k or explicit equations) and a $\mathcal{R}$OT solver for the chosen $\mathcal{R}$. This makes our approach fully modular and compatible with any state-of-the-art OT backend.

remains stable when only one Sinkhorn update is performed at each ϕ-block. This represents a large computational acceleration, since we do not solve a full $\mathcal{R}$OT problem before each Ξ update.

Efficient implementation in the block-separable case. To fully benefit from memory efficiency of dual $\mathcal{R}$OT solvers like Sinkhorn, we must replace the computations of $A[\widehat{\pi}]$ and $B[\widehat{\pi}]$ with cheaper ones. When A and B are block-separable, they are parametrized by their embeddings so that $A[\pi]$ and $B[\pi]$ take the form of a cross-correlation:

$$A[\pi] = \mathrm{Corr}_A[\pi] := \int \Phi_A(x)\Psi_A(y)^\top \mathrm{d}\pi(x, y). \tag{17}$$

These standard operations are efficiently implementable with Sinkhorn-like solvers. Moreover, the cost $c_{\Gamma,\Xi}$ of eq. (11) is an inner product, a cost function supported by most OT libraries: this makes our algorithm compatible with any choice of $\mathcal{R}$ as long as a $\mathcal{R}$OT solver is available. The block-wise structure makes it fully modular (fig. 1), and benefits from modern Sinkhorn backends (Charlier et al., 2021; Ye et al., 2026): our method is GPU-friendly, memory-efficient, and accelerable with multiscaling. Since matching problems are parametrized by embeddings, dimensionality reduction can be naturally applied to offer a substantial speed-up over exact resolution. We refer to section C for implementation details, acceleration techniques, and a discussion of initialization strategies.

5 Applications

We now highlight the practical utility of our new framework by presenting several applications on a wide variety of data types, all solved using our numerical routine presented in section 4. We detail our experimental setup in section A.2. These applications aim to demonstrate the versatility of our solver: we used the same algorithm to compute all these experiments, the only changes lying in the feature embeddings and the selected term $\mathcal{R}$.

Graph matching. Fig 2 evaluates our method on random graph matching using GW with shortest path as base costs. We compare our exact and dimension-reduced methods against the standard EGW solver of Peyré et al. (2016) and the low-rank (LR) solver of Scetbon et al. (2022). All methods

Table 1: Solving time and precision of word translation using balanced GW, following the method of Alvarez-Melis & Jaakkola (2018). Baseline results are directly reported from the original study.

SOLVER	DIM.	TIME	EN-ES		EN-FR		EN-IT		EN-DE	
			$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$	$\rightarrow$	$\leftarrow$
OURS	100	**36s**	75.4	75.8	79.1	79.3	71.7	70.6	58.4	59.6
OURS	200	70s	79.6	80.5	81.4	81.8	78.8	78.0	71.8	74.2
OURS	300	90s	80.7	**81.6**	**81.7**	**82.2**	**79.6**	**79.0**	**72.7**	**75.2**
BASE	300	2220s	**81.7**	80.4	81.3	78.9	78.9	75.2	71.9	72.8

perform similarly, except for LR which trades matching quality for computational speed. In sharp contrast, our method reaches an excellent accuracy with as few as $D = 10$ dimensions. Although EGW has a larger asymptotic complexity than our low-dimensional methods, its minimal overhead makes it competitive in this setting. However, EGW lacks theoretical convergence guarantees: in section A, we highlight simple cases where EGW fails to converge entirely. To our knowledge, our method is the first to robustly compute GW matchings on graphs with tens of thousands of nodes in a few seconds without suffering from such convergence issues.

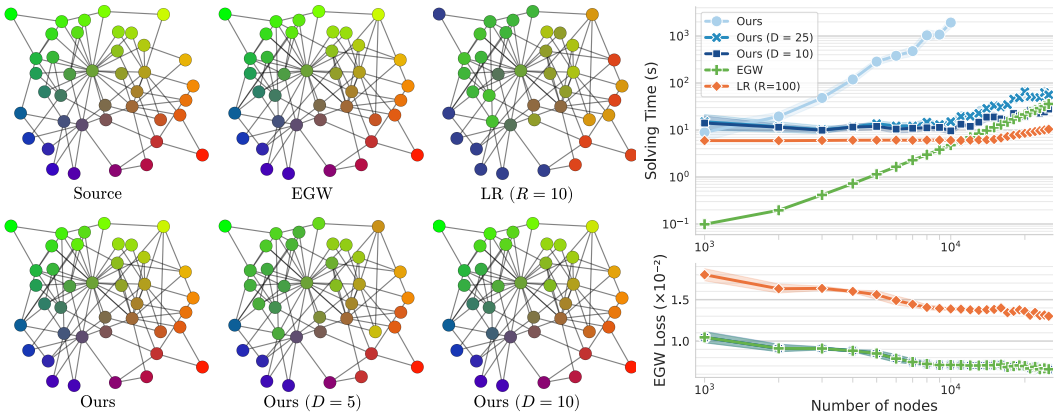

Figure 2: Gromov-Wasserstein methods for graph matching. *Left*: qualitative analysis; colors indicate which nodes are matched together. *Right*: quantitative results of speed and accuracy.

Word embedding alignment. To showcase our method in high dimension, we reproduce the experiment of (Alvarez-Melis & Jaakkola, 2018, Table 1) by aligning vocabularies of different languages. The MUSE dataset (Conneau et al., 2017) provides 300-dimensional embeddings of different language vocabularies (English, Spanish, French, Italian and German). We apply GW between the $N = 20k$ most frequent words of these languages, using a multiscale scheme to reduce solving time. We infer a translation for each source word in the target language and report total precision in table 1. We see that our solver obtains comparable results to the baseline while being almost $20\times$ faster. Using Nyström methods (Nakatsukasa, 2020) further accelerates our algorithm by reducing the embedding dimensions, providing a flexible trade-off between precision and running time.

Semi-balanced shape matching. In fig. 3, we match a source mesh to various partial targets from the SHREC'16 dataset (Cosmo et al., 2016) using semi-balanced GW. Semi-balancedness ensures that the whole partial shapes are matched to the source, while ignoring vertices of the full mesh that correspond to holes of the target. Our method converges in a few minutes and accurately preserves the aspect of the textures. We provide qualitative comparisons with other GW baselines in section A which highlight the limitation of concurrent unbalanced GW methods.

Fused matching of 3D point clouds. To demonstrate the large-scale applicability of our method, we apply our solver to the scene reconstruction dataset of Choi et al. (2015) which contains various scans of the same room encoded as 3D clouds with $100k$ to $200k$ points. We match these scans with



Figure 3: Texture transfer between source shape (*left*) and different targets using our method. The source contain $10k$ vertices; the targets, from $3k$ to $10k$.



Figure 4: Matching between 3D point clouds represented by 100 edges sampled from π^*.

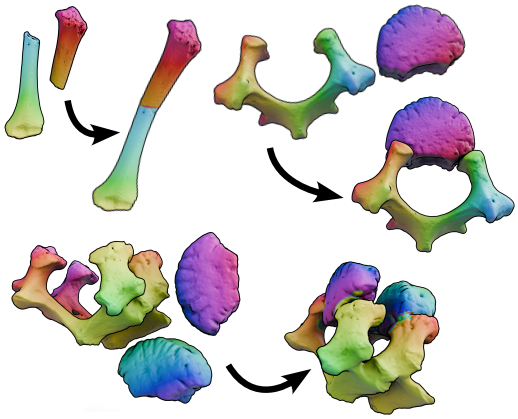

Figure 5: Color transfer between fractured and original objects using our method.

unbalanced GW, using RGB colors as additional linear features, following the *fused* idea of Vayer et al. (2020). The results are computed in about 30 min and are qualitatively satisfying, as depicted in fig. 4.

Fracture reconstruction. In fig. 5, we highlight our framework polyvalence by matching fractured objects with their original templates from the synthetic dataset of Li et al. (2025). We use a cluster-aware variant of GW that makes it invariant to isometries of the individual fragments; due to the high number of local minima, we selected the output as the best result over dozens of initializations. We combine this technique with multiscaling to reduce the computational burden of these computations. Since our input shapes contain $100k$ to $200k$ vertices each, an optimization still takes several hours. Yet, this preliminary work shows how our framework can help to tackle new matching problems with custom geometries without needing to redesign solvers from scratch.

6 Conclusion

Our work provides a common foundation for studying and solving a wide range of matching problems in a principled manner. Rather than requiring a dedicated solver for each formulation, it offers a flexible framework that can be readily adapted to new models. Built on duality theory, our framework encompasses Gromov–Wasserstein matching and its fused and unbalanced entropic variants within a unified algorithmic approach. For a broad class of regularized quadratic problems, we establish convergence rates, thereby bridging a gap in the existing literature. As illustrated by our fracture-matching model, the framework also enables new formulations to be explored without designing a solver from scratch. We therefore believe that generic solvers such as ours can facilitate rapid prototyping and testing, and help the community to investigate approaches beyond classical matching algorithms. More broadly, our theoretical results provide insight into which classes of matching problems are easier to solve and reveal connections between the many methods proposed in the literature.

AI USE STATEMENT

In this work, we used generative AI tools to assist with and polish writing, to retrieve and discover related work, and to assist with the development and verification of some mathematical proofs. We did not use generative AI tools for research ideation, the design or execution of the research, or the generation of synthetic datasets.

All AI-assisted content was carefully reviewed and independently verified by the authors. In particular, we checked the validity of all mathematical arguments and proofs, rewrote any AI-generated text ourselves, and verified the existence and relevance of all references retrieved with the assistance of LLMs. We take full responsibility for the final content of this work, including any text, claims, proofs, or other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility of our results, we provide extensive details of both theoretical and implementation aspects in the appendix. In section A.2, we describe the setup of our experiments and the preprocessing of the input data. In sections B and C, we provide a step-by-step description of all the computations and implementation tools necessary to reproduce our algorithm; we also attach our source code with simple examples of experiments as a file to the submission. Finally, in sections D and E, we rigorously prove all the theoretical results that we state in our article and discuss thoroughly convergence guarantees.

REFERENCES

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1881–1890, 2018.

David Alvarez-Melis, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1870–1879. PMLR, 2019.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Jean-Pierre Aubin. Chapter 3 compactness and continuity properties. In *Mathematical Methods of Game and Economic Theory*, volume 7 of *Studies in Mathematics and Its Applications*, pp. 75–102. Elsevier, 1979. doi: https://doi.org/10.1016/S0168-2024(09)70011-6. URL https://www.sciencedirect.com/science/article/pii/S0168202409700116.

Bachir Bekka, Pierre de La Harpe, and Alain Valette. Kazhdan's property (t). *preprint*, 11, 2007.

Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pp. 586–606. Spie, 1992.

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

Lucas Brifault, David Cohen-Steiner, and Mathieu Desbrun. Efficient and scalable spatial regularization of optimal transport. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

Benjamin Charlier, Jean Feydy, Joan Alexis Glaunes, François-David Collin, and Ghislain Durif. Kernel operations on the gpu, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. Robust reconstruction of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5556–5565, 2015.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Luca Cosmo, Emanuele Rodola, Michael M Bronstein, Andrea Torsello, Daniel Cremers, Y Sahillioğlu, et al. Shrec'16: Partial matching of deformable shapes. In *Eurographics Workshop on 3D Object Retrieval, EG 3DOR*, pp. 61–67. Eurographics Association, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods with inexact oracle: the strongly convex case. *CORE Discussion Paper*, 2013.

Jean Feydy, Alexis Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *Advances in Neural Information Processing Systems*, 33:14448–14462, 2020.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Ferdinand Genans. Fast and large-scale unbalanced optimal transport via its semi-dual and adaptive gradient methods. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2561–2600. PMLR, 29 Jun–03 Jul 2026.

Steven Gold and Anand Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on pattern analysis and machine intelligence*, 18(4):377–388, 1996.

Steven Gold, Chien-Ping Lu, Anand Rangarajan, Suguna Pappu, and Eric Mjolsness. New algorithms for 2d and 3d point matching: Pose estimation and correspondence. *Advances in neural information processing systems*, 7, 1994.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with wasserstein procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1880–1890. PMLR, 2019.

Aric Hagberg and Drew Conway. Networkx: Network analysis with python. *URL: https://networkx. github. io*, 1031, 2020.

Philip Hartman. On functions representable as a difference of convex functions. *Pacific Journal of Mathematics*, 1959.

J-B Hiriart-Urruty. Generalized differentiability/duality and optimization for problems dealing with differences of convex functions. In *Convexity and Duality in Optimization: Proceedings of the Symposium on Convexity and Duality in Optimization Held at the University of Groningen, The Netherlands June 22, 1984*, pp. 37–70. Springer, 1985.

Guillaume Houry, Jean Feydy, and François-Xavier Vialard. Gromov-Wasserstein at scale, beyond squared norms. In *Forty-third International Conference on Machine Learning*, 2026.

Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. Pytorch. In *Programming with TensorFlow: solution for edge computing applications*, pp. 87–104. Springer, 2021.

Eugene L Lawler. The quadratic assignment problem. *Management science*, 9(4):586–599, 1963.

Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257, 2016.

Sihang Li, Zeyu Jiang, Grace Chen, Chenyang Xu, Siqi Tan, Xue Wang, Irving Fang, Kristof Zyskowski, Shannon P McPherron, Radu Iovita, et al. Garf: Learning generalizable 3d reassembly for real-world fractures. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5711–5721. IEEE, 2025.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211 (3):969–1117, 2018.

Lorenzo Livi and Antonello Rizzi. The graph matching problem. *Pattern Analysis and Applications*, 16(3):253–283, 2013.

Manish Mandad, David Cohen-Steiner, Leif Kobbelt, Pierre Alliez, and Mathieu Desbrun. Variance-minimizing transport plans for inter-surface mapping. *ACM Transactions on Graphics (ToG)*, 36 (4):1–14, 2017.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

Andriy Myronenko and Xubo Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010.

Yuji Nakatsukasa. Fast and stable randomized low-rank matrix approximation. *arXiv preprint arXiv:2009.11392*, 2020.

Maks Ovsjanikov, Mirela Ben-Chen, Justin Solomon, Adrian Butscher, and Leonidas Guibas. Functional maps: a flexible representation of maps between shapes. *ACM Transactions on Graphics (ToG)*, 31(4):1–11, 2012.

Daniele Panozzo, Ilya Baran, Olga Diamanti, and Olga Sorkine-Hornung. Weighted averages on surfaces. *ACM Transactions on Graphics (TOG)*, 32(4):1–12, 2013.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–Wasserstein averaging of kernel and distance matrices. In *International conference on machine learning*, pp. 2664–2672. PMLR, 2016.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pp. 7673–7682. PMLR, 2020.

Gabriel Rioux, Ziv Goldfeld, and Kengo Kato. Entropic gromov-wasserstein distances: Stability and algorithms. *Journal of machine learning research*, 25(363):1–52, 2024.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time gromov wasserstein distances using low rank couplings and costs. In *International Conference on Machine Learning*, pp. 19347–19365. PMLR, 2022.

Meyer Scetbon, Michal Klein, Giovanni Palla, and Marco Cuturi. Unbalanced low-rank optimal transport solvers. *Advances in Neural Information Processing Systems*, 36:52312–52325, 2023.

Isaac J Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Structured transforms across spaces with cost-regularized optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pp. 586–594. PMLR, 2024.

Thibault Séjourné. *Generalized optimal transport, computation and applications*. PhD thesis, Université Paris sciences et lettres, 2022.

Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation. *Advances in Neural Information Processing Systems*, 34:8766–8779, 2021.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

John F Toland. Duality in nonconvex optimization. *Journal of Mathematical Analysis and Applications*, 66(2):399–415, 1978.

Yanghai Tsin and Takeo Kanade. A correlation-based approach to robust point set registration. In *European conference on computer vision*, pp. 558–569. Springer, 2004.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Fused gromov-wasserstein distance for structured objects. *Algorithms*, 13(9):212, 2020.

Ruodu Wang and Zhenyuan Zhang. Quadratic-form optimal transport. *Mathematical Programming*, pp. 1–54, 2025.

Felix X-F Ye, Xingjie Li, An Yu, Ming-Ching Chang, Linsong Chu, and Davis Wertheimer. Flashsinkhorn: Io-aware entropic optimal transport on gpu. *arXiv preprint arXiv:2602.03067*, 2026.

Zhengxin Zhang, Ziv Goldfeld, Youssef Mroueh, and Bharath K Sriperumbudur. Gromov–wasserstein distances: Entropic regularization, duality and sample complexity. *The Annals of Statistics*, 52(4):1616–1645, 2024.

A Additional Experiments

A.1 Other Baseline Comparisons

Failure cases of naive solver. We illustrate the relevance of our approach by identifying examples where our algorithm reaches a true EGW minimum whereas concurrent solvers (Peyré et al., 2016) fail to converge. As explained in Houry et al. (2026), this can happen when the costs are not CNT. We exhibit two typical examples of non-CNT configurations: shortest path distances on arbitrary graphs (fig. 6), and 2D point clouds for base costs equal to $\|\cdot\|^p$, with $p > 2$ (fig. 7). In both cases, our method manage to converge to the correct minimum whereas the standard baseline oscillates between two transport plans without reaching any critical point. Note that these examples represent typical inputs for GW applications and are not purely abstract pathological examples. Therefore, this type of behavior is to be expected in practice, and using our solver ensures that convergence is indeed reached in every configuration. Also note that, despite the symmetries of the inputs in fig. 6 and the symmetry of the initialization, our solver breaks the symmetry and reaches one of the global minima of the problem. Therefore, numerical noise is sufficient to prevent our algorithm from getting stuck in a critical point – a well-known property of gradient descent (Lee et al., 2016).

Partial shape matching using other baselines. We complement the results of fig. 3 with matching obtained using other Gromov-Wasserstein methods. Except from our method, only two other baselines are able to solve unbalanced GW problems: the UGW solver of Séjourné et al. (2021), and the low-rank relaxation of Scetbon et al. (2023). The first algorithm cannot scale properly to the inputs used in these experiments. Although faster, the low-rank solver rely on a relaxation that creates undesirable quantization artifacts, as we illustrate at the top of fig. 8. Therefore, to the best of our knowledge, our method is the only one which is able to compute accurate unbalanced Gromov-Wasserstein solutions on this input size in a reasonable time. As a reference, we also show

the matching obtained using a balanced Gromov-Wasserstein solver on these inputs at the bottom fig. 8. We observe large distorsion in the results, evidencing the need to rely on unbalanced versions of Gromov-Wasserstein to tackle this type of problems.

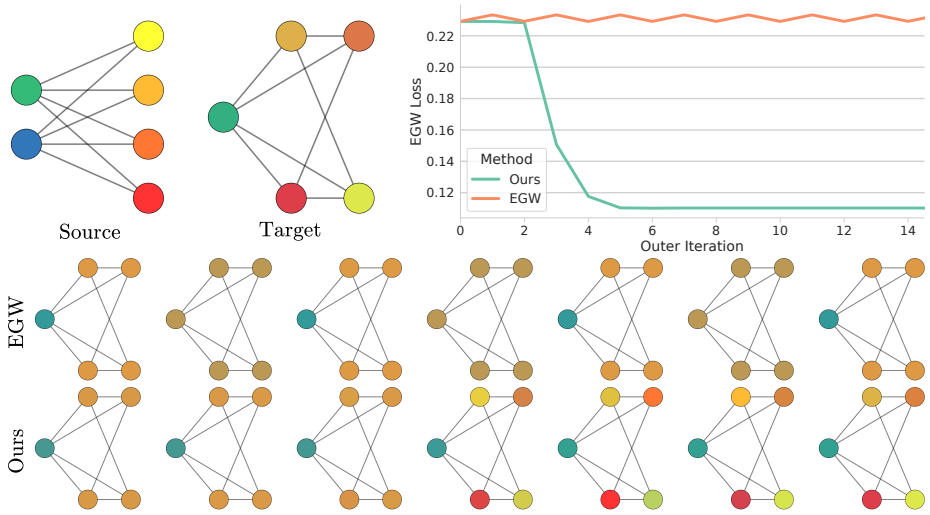

Figure 6: Graph matching obtained through the iterations of the EGW solver (Peyré et al. (2016)) and ours. EGW gets stuck between two configurations, whereas ours converge to the true optimum.

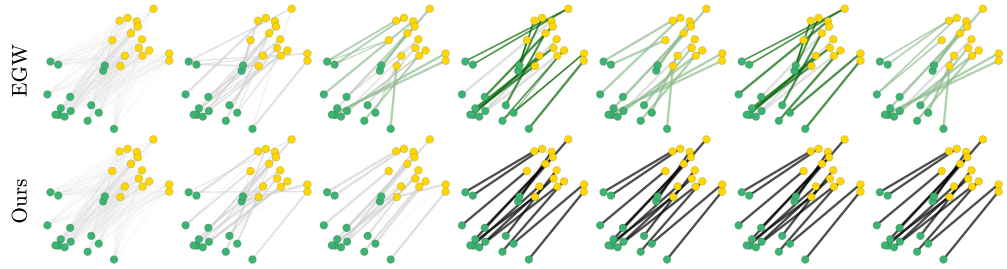

Figure 7: Matching obtained through iterations of the EGW solver and ours with $c_{\mathcal{X}}, c_{\mathcal{Y}} = \|\cdot\|^6$.



Figure 8: Texture transfer between the source of fig. 3 and different targets using the LR unbalanced solver of Scetbon et al. (2023) (*top*) and the balanced EGW solver of Peyré et al. (2016) (*bottom*).

A.2 Experimental Setup

All our algorithms were implemented in Pytorch Imambi et al. (2021) and executed on a RTX6000 GPU with 24GB of VRAM. To ensure a fair comparison with baselines, we reimplemented concurrent algorithms in Pytorch as well. In dimension reduction settings, we performed OT related operations using either the Keops (Charlier et al., 2021) or FlashSinkhorn (Ye et al., 2026) backends. Below, we provide additional details on the experiments we performed in this article.

Graph matching. We compute graph matching using the balanced entropic GW objective:

$$\min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (d_{\mathcal{X}}(x,x') - d_{\mathcal{Y}}(y,y'))^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') + \iota_{\Pi(\alpha,\beta)}(\pi) + \varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta),$$

where $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ are the shortest path distances of the input graphs $\mathcal{X}$ and $\mathcal{Y}$, and $\varepsilon = 2e-3$. The source graphs are generated randomly from the Barabási–Albert model using the corresponding function of the networkx library (Hagberg & Conway, 2020), with an edge parameter of 2. Targets graphs are obtained by adding 20% of noisy edges to the source, so that there is a ground-truth correspondences between the nodes of the source and the target. To apply our method, we normalize then diagonalize the shortest path distance matrices so that it takes the form of proposition 5. In our dimension reduction methods, we truncate this diagonalization to the D largest eigenvalues in absolute value. For the low-rank baseline, we also truncate the costs to the $D = 100$ first eigenspaces, necessary to ensure linear-time computations of this method. We use uniform weights α and β on both source and target graphs, and initialize the algorithms at the trivial plan – except for low-rank where we use a kmeans heuristics for initialization. The accuract reported in fig. 2, called *EGW loss*, refers to the value of eq. (12) computed using the matching π obtained after convergence.

Word embedding alignment. Results of table 1 rely on the balanced entropic GW problem:

$$\min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') + \iota_{\Pi(\alpha,\beta)}(\pi) + \varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta),$$

with cosine similarity as base costs:

$$c_{\mathcal{X}}(x,x') = \left\langle \frac{x}{\|x\|}, \frac{x'}{\|x'\|} \right\rangle \quad \text{and} \quad c_{\mathcal{Y}}(y,y') = \left\langle \frac{y}{\|y\|}, \frac{y'}{\|y'\|} \right\rangle.$$

These have a scalar structure, so we can directly apply the decomposition of proposition 3 on the spherical projection of the inputs. We use a multiscale scheme by solving a coarse problem containing only the $2k$ most frequent words of each language, using the coarse result as initialization for the fine-scale one. We then estimate a rigid alignment of the space from the optimal matching, as detailed in Alvarez-Melis & Jaakkola (2018). Finally, we match each word of the source language to a word of the target language using the CSLS matching method (Conneau et al., 2017). The reported precision corresponds to the proportion of source words whose matched target is a valid translation in the dictionaries provided by the MUSE dataset.

Semi-balanced shape matching. The explicit optimization problem we used for fig. 3 is:

$$\min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y') + \iota_{\alpha}(\pi_1) + \varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta),$$

with $\varepsilon = 5e-4$. The costs $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are sums of squared norms:

$$c_{\mathcal{X}}(x,x') = \|x-x'\|^2_{\text{geodesic}} + \|x-x'\|^2_{\text{exp}} \quad \text{and } c_{\mathcal{X}}(y,y') = \|y-y'\|^2_{\text{geodesic}} + \|y-y'\|^2_{\text{exp}}.$$

The norm $\|\cdot\|_{\text{geodesic}}$ correspond to Euclidean approximations of geodesic distances as in Panozzo et al. (2013). We follow the same method as detailed in Houry et al. (2026), yielding Euclidean embeddings in dimension $D = 8$. On the other hand, the terms $\|\cdot\|^2_{\text{exp}}$ represent squared-norm approximations in dimension $D = 20$ of the conditionally negative cost:

$$d_{\mathcal{X}}(x,x') = 2 - 2\exp\left(-\frac{\|x-x'\|}{0.5}\right) \quad \text{and} \quad d_{\mathcal{Y}}(y,y') = 2 - 2\exp\left(-\frac{\|y-y'\|}{0.5}\right),$$

using kernel PCA, to compute approximate Euclidean embeddings, as advocated in Houry et al. (2026). This way, $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ combine both global geodesic and local Euclidean information. We initialized the solver using landmarks and chose the weights α and β proportional to average area of triangles surrounding each vertex. For the low-rank baseline (in fig. 8), we used a rank of $R = 100$.

Fused matching of 3D point clouds. To compute the matchings of fig. 4, we minimize the fused unbalanced GW objective:

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (c_{\mathcal{X}}(x,x') - c_{\mathcal{Y}}(y,y'))^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y')$$
$$+ \lambda \int \|\mathrm{RGB}(x) - \mathrm{RGB}(y)\|^2 \,\mathrm{d}\pi(x,y) + \rho D_{\mathrm{KL}}(\pi_1\|\alpha) + \rho D_{\mathrm{KL}}(\pi_2\|\beta) + \varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta),$$

where $c_{\mathcal{X}}, c_{\mathcal{Y}}$ are 20-dimensional approximations of classical euclidean distances (following the kernel PCA method of Houry et al. (2026)). The fused features $\mathrm{RGB}(x) \in [0,1]^3$ correspond to the color of the point x. We chose $\varepsilon = 2e-3$, $\rho = 1e-1$, and $\lambda = 4e-4$. We applied multiscaling, sampling $10k$ random points from each input to build the coarse problem. We chose α and β as uniform weights and initialized the solver at the trivial plan.

Fracture reconstruction. In fig. 5, we optimize the weighted GW objective:

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int \sum_k \mathbf{1}_{x\in C_k}\mathbf{1}_{x'\in C_k}(\|x-x'\|^2 - \|y-y'\|^2)^2 \mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y')$$
$$+ \varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta), \quad (18)$$

where $\mathbf{1}_{x\in C_k}$ is equal to 1 if x belongs to the fragment C_k and 0 otherwise. We take $\varepsilon = 10^{-4}$ and α, β as uniform weights. The coarse optimization is performed on random subsamples of the inputs with $3k$ points. The best coarse plan (in terms of the objective of eq. (18)) is selected over 30 random initializations. The final output is then obtained by computing the fine-scale problem initialized at this best coarse plan.

B Additional Results

The main part of our article describes a general strategy to tackle matching problems. This section provides additional ingredients to apply these results to classical matching instances, as well as existence results for the decompositions we use in our main theorem. In section B.1, we provide examples of constraint functions $\mathcal{R}$ and the associated duals. In section B.2, we justify the existence of DC decomposition for general matching, and factorizations of quadratic objectives in a difference of quadratic operators. In section B.3, we detail the formulation of our dual theorems when the quadratic operators are block-separable. Finally, in section B.4, we explicit the block-separable decomposition of GW objectives, which are the formula we ultimately implement in our solvers. Before that, we introduce additional and definitions:

Weak-* continuity. A sequence of measures $(\alpha_n) \in \mathcal{M}(\mathcal{X})$ converges weakly-* to $\alpha \in \mathcal{M}(\mathcal{X})$ (denoted by $\alpha_n \rightharpoonup \alpha$) if for any continuous function $f \in \mathcal{C}(\mathcal{X})$, $\int f \mathrm{d}\alpha_n \to \int f \mathrm{d}\alpha$. A function $\mathcal{F} : \mathcal{M}(\mathcal{X}\times\mathcal{Y}) \to V$ is weak-* continuous if $\mathcal{F}(\alpha_n) \to \mathcal{F}(\alpha)$ for all sequence $\alpha_n \rightharpoonup \alpha$. Bounded closed spaces of $\mathcal{M}(\mathcal{X}\times\mathcal{Y})$ are compact for the weak-* topology; in particular, if $\mathcal{E}$ is weak-* continuous and goes to infinity at infinity, the infimum:

$$\inf_{\pi\in\mathcal{M}(\mathcal{X}\times\mathcal{Y})} \mathcal{E}(\pi)$$

is always attained.

Trivial measure. Given $\alpha \in \mathcal{M}(\mathcal{X})$ and $\beta \in \mathcal{M}(\mathcal{Y})$, we denote by $\alpha\otimes\beta \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$ the *trivial* measure with marginals α and β:

$$\text{For all } c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y}), \quad \int c \mathrm{d}\alpha\otimes\beta = \int c(x,y)\mathrm{d}\alpha(x)\mathrm{d}\beta(y).$$

We will notably use this notation to simplify the expression of quadratic matching objectives:

$$\int k \mathrm{d}\pi\otimes\pi = \int k(x,y,x',y')\mathrm{d}\pi(x,y)\mathrm{d}\pi(x',y').$$

Generalized KL-divergence. In this work, we adopt the setting of Séjourné (2022) and use the following definition of KL-divergence:

$$D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta) := \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \otimes \beta}\right) \mathrm{d}\beta - \int \mathrm{d}\alpha \otimes \beta + \int \mathrm{d}\pi.$$

The terms $\int \mathrm{d}\pi$ and $\mathrm{d}\alpha \otimes \beta$ cancel when both measures have the same mass, but are required in unbalanced settings to avoid biasing matching objectives towards 0-mass plans.

Convex indicator function. Given a convex set A, we denote by ι_A its convex indicator function:

$$\iota_A(x) = \begin{cases} 0 & \text{if } x \in A, \\ +\infty & \text{otherwise.} \end{cases}$$

Adjoints of integral operators. Since an integral operator $A : \pi(\mathcal{X} \times \mathcal{Y}) \to \mathcal{H}_A$ is a continuous linear function, it admits an adjoint $A^\top : \mathcal{H}_A \to \mathcal{C}(\mathcal{X} \times \mathcal{Y})$ that is given by:

$$\text{For all} \quad \Xi \in \mathcal{H}_A \quad \text{and} \quad (x, y) \in \mathcal{X} \times \mathcal{Y}, \quad (A^\top \Xi)(x, y) = \langle \Xi, A(x, y) \rangle .$$

B.1 Classical choices of $\mathcal{R}$

In tables 2 and 3, we introduce the main constraint terms $\mathcal{R}$ encountered in both matching and registration literatures, along with their duals (following the notations of theorem 2). The two most standards are optimal transport and closest point. In the OT case, the matching π must belong to:

$$\Pi(\alpha, \beta) := \left\{ \pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}), \ \pi_1 := \int \mathrm{d}\pi(\cdot, y) = \alpha \ \text{ and } \ \pi_2 := \int \mathrm{d}\pi(x, \cdot) = \beta \right\}, \quad (19)$$

the set of *couplings* between α and β. Closest point matchings, on the other hand, are obtained by matching each point of α to its nearest neighbor in β based on a cost function c:

$$\pi_{\mathrm{CP}}(x, y) = \begin{cases} \alpha(x) & \text{if } y \in \arg\min_x c(x, y), \\ 0 & \text{otherwise.} \end{cases}$$

Such matchings are actually the solutions of a problem of the form $\min_\pi \int c \mathrm{d}\pi + \iota_\alpha(\pi_1)$, where $\iota_\alpha(\pi_1)$ only enforces the marginal constraint on α. Therefore, this popular method in registration is entirely compatible with our framework with the choice $\mathcal{R}(\pi) = \iota_\alpha(\pi_1)$. Other important cases combine marginal penalization and entropic (or KL-divergence) regularization:

$$\mathcal{R}(\pi) = \mathrm{D}_{\varphi_1}(\pi_1 \| \alpha) + \mathrm{D}_{\varphi_2}(\pi_2 \| \beta) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$$
$$\text{with} \quad \mathrm{D}_\varphi(\mu \| \nu) := \begin{cases} \int \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{if } \mu \ll \nu, \\ +\infty & \text{otherwise.} \end{cases} \quad (20)$$

The function D_φ is called the Csiszár divergence associated to the function φ: it is well-defined as soon as φ is convex and $\varphi(1) = 0$. As explained in Séjourné (2022), their dual problems admit the following form, the space of dual variables being $\Phi = \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})$:

$$\mathcal{D}_{\mathcal{R}}((f, g), c, \alpha, \beta) = -\int \varphi_1^*(-f) \mathrm{d}\alpha - \int \varphi_2^*(-g) \mathrm{d}\beta - \varepsilon \int (e^{(f \oplus g - c)/\varepsilon} - 1) \, \mathrm{d}\alpha \otimes \beta,$$

where φ_1^* and φ_2^* are the convex conjugates of φ_1 and φ_2. The unbalanced case of eq. (2) corresponds to $D_\varphi = D_{\mathrm{KL}}$ (with $\varphi(r) = r \log(r) - r + 1$).

B.2 General DC Decomposition of Matching Problems

Here, we motivate our study of DC matching problems by providing several theorems showing the existence of DC decompositions for general matching objectives.

Table 2: List of classical regularizers.

Regularizer	$\mathcal{R}(\pi)$
Classical OT	$\iota_{\Pi(\alpha,\beta)}(\pi)$
Closest Point	$\iota_{\alpha}(\pi_1)$
Entropic OT	$\iota_{\Pi(\alpha,\beta)}(\pi) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$
Semi-Balanced EOT	$\iota_{\alpha}(\pi_1) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$
Unbalanced OT	$\rho_1 D_{\mathrm{KL}}(\pi_1, \|\alpha) + \rho_2 D_{\mathrm{KL}}(\pi_2 \| \beta) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$
Partial OT	$\rho_1 \mathrm{TV}(\pi_1 \| \alpha) + \rho_2 \mathrm{TV}(\pi_2 \| \beta) + \varepsilon D_{\mathrm{KL}}(\pi \| \alpha \otimes \beta)$

Table 3: Dual functions of the regularizers.

Regularizer	$\mathcal{D}_{\mathcal{R}}((f,g),\alpha,\beta)$
Classical OT	$\int f \mathrm{d}\alpha + \int g \mathrm{d}\beta + \iota_{f \oplus g \leq c}(f,g)$
Closest Point	$\int f \mathrm{d}\alpha + \iota_{f(x)=\min_y c(x,y)}(f)$
Entropic OT	$\int f \mathrm{d}\alpha + \int g \mathrm{d}\beta - \varepsilon \int (e^{(f \oplus g - c)/\varepsilon} - 1)\, \mathrm{d}\alpha \otimes \beta$
Semi-Balanced EOT	$\int f \mathrm{d}\alpha - \varepsilon \int (e^{(f \oplus 0 - c)/\varepsilon} - 1) \mathrm{d}\alpha \otimes \beta$
Unbalanced OT	$\rho_1 \int (1 - e^{-f/\rho_1}) \mathrm{d}\alpha + \rho_2 \int (1 - e^{-g/\rho_2}) \mathrm{d}\beta - \varepsilon \int \big(e^{(f \oplus g - c)/\varepsilon} - 1\big) \mathrm{d}\alpha \otimes \beta$
Partial OT	$\int \min(\rho_1, f) \mathrm{d}\alpha + \int \min(\rho_2, g) \mathrm{d}\beta - \varepsilon \int \big(e^{(f \oplus g - c)/\varepsilon} - 1\big) \mathrm{d}\alpha \otimes \beta$

General matching approximation by DC functions. DC functions are dense in spaces of compactly-supported continuous functions: therefore, continuous matching objectives can always be approximated by a difference of convex functions on compact domains. If the matching objective grows arbitrarily at infinity, the minimization can always be restricted to a compact set and the class of DC functions adequately approximates any sufficiently regular matching problem. We formalize this property in the following proposition.

Proposition 1. *Let assume that $\mathcal{E}$ is continuous for the weak-* convergence and goes to infinity at infinity. Then, for any precision $\delta > 0$, there exists a bounded convex set $V \subset \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ containing the minima of $\mathcal{E}$ and two convex functions $\mathcal{A}_\delta$, $\mathcal{B}_\delta$ continuous on their domain such that:*

$$\textit{For all } \pi \in V, \quad |\mathcal{E}(\pi) - \mathcal{A}_\delta(\pi) + \mathcal{B}_\delta(\pi)| < \delta,$$

and $\mathcal{A}_\delta(\pi) - \mathcal{B}_\delta(\pi) = +\infty$ for $\pi \notin V$.

Besides, when $\mathcal{A}$ and $\mathcal{B}$ are superlinear, their convex conjugates are continuous:

Lemma 1. *Let assume that $\mathcal{A}$ is superlinear:*

$$\frac{\mathcal{A}(\pi_n)}{\mathrm{TV}(\pi_n)} \longrightarrow +\infty \quad \textit{for all } (\pi_n) \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \quad \textit{such that} \quad \mathrm{TV}(\pi_n) \longrightarrow \infty,$$

with $\mathrm{TV}(\pi)$ denoting the total variation of π. Then, the convex conjugate $\mathcal{B}^*$ is continuous on $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$ for the uniform norm.

Since adding the same term to $\mathcal{A}$ and $\mathcal{B}$ do not modify their difference, we can always enforce the DC decomposition to have superlinear terms by adding the function $\pi \mapsto \mathrm{TV}(\pi)^2$ to both $\mathcal{A}$ and $\mathcal{B}$. Therefore, we can always enforce the continuity of the convex conjugates $\mathcal{A}^*$ and $\mathcal{B}^*$ without any additional hypothesis – a crucial requirement for our reformulation theorems.

When $\mathcal{X}$ and $\mathcal{Y}$ are finite spaces, the set $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ is identified with a Euclidean matrix space: if $\mathcal{E}$ is of class $\mathcal{C}^2$, Hartman (1959); Hiriart-Urruty (1985) proved that it admits an exact DC decomposition over the whole space. In this case, the DC hypothesis is satisfied without any approximation. We refer to Hiriart-Urruty (1985) for other characterizations of DC decomposable functions.

Factorizability of quadratic operators. In the discrete case, the spectral theorem ensures that quadratic operators always admit a factorization like eq. (13) which is given by the decomposition of the kernel matrix $[k(x_i, y_j, x_k, y_l)]_{(ij),(kl)}$ in positive and negative eigenspaces. Similarly, since

the images of A and B are isomorphic to Euclidean spaces $\mathbb{R}^{\mathrm{D}}$ and $\mathbb{R}^{\mathrm{E}}$, the singular value decompositions of each matrix $[A(x_i, y_j)_d]_{ij}$ and $[B(x_i, y_j)_e]_{ij}$ (with $1 \leq d \leq \mathrm{D}$ and $1 \leq e \leq \mathrm{E}$) provides a block-separable factorization of QM objectives in the finite setting.

When α and β are continuous, this reasoning does not hold anymore. Indeed, spectral decompositions are notions related to Hilbert spaces, whereas $\mathcal{M}(\mathcal{X} \times \mathcal{Y})$ is only a Banach space. However, when using entropic regularization, the $\mathcal{R}\mathrm{QM}$ optimization can actually be restricted to the space of measures with squared-integrable density with respect to $\alpha \otimes \beta$:

$$\mathcal{M}^{L^2}(\alpha \otimes \beta) := \left\{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \quad \text{s.t. } \exists f \in L^2(\mathcal{X} \times \mathcal{Y}, \alpha \otimes \beta), \quad \pi = f \mathrm{d}\alpha \otimes \beta\right\}.$$

Since $L^2(\mathcal{X} \otimes \mathcal{Y}, \alpha \otimes \beta)$ is a Hilbert space, this allows us to diagonalize the kernel and prove:

Proposition 2. *If $\mathcal{R}$ contains an entropic term, k is continuous and α, β have compact support, then there exists two linear operators $A, B : \mathcal{M}_+^{L^2}(\alpha \otimes \beta) \to L^2(\mathcal{X} \times \mathcal{Y}, \alpha \otimes \beta)$ such that:*

$$\mathcal{R}\mathrm{QM}(\alpha, \beta) = \inf_{\pi \in \mathcal{M}_+^{L^2}(\alpha \otimes \beta)} \tfrac{1}{2} \|A[\pi]\|^2 - \tfrac{1}{2} \|B[\pi]\|^2 + \mathcal{R}(\pi).$$

with $\mathcal{M}_+^{L^2}(\alpha \otimes \beta)$ denoting the subset of positive measures of $\mathcal{M}^{L^2}(\alpha \otimes \beta)$.

The setting is slightly different from the hypotheses of our article, but our duality results can be adapted to this L^2 case. It evidences the naturality of our factorization property when working with entropic-regularized QM, suggesting that the reformulations we propose in section 3 are more than a numerical trick and display a fundamental property of quadratic matching.

B.3 Detail of the Block-Separable Case

Theorem 4 states the existence of embedding functions such that $\mathcal{R}\mathrm{QM}$ is equivalent to our new formula. In the separable case, these embeddings correspond to the functions Φ, Ψ of definition 2, but the block-separable case is more involved: we explain how to build them from the block-decomposition of the operators A and B. Following the notations of definition 2, we denote the embeddings of each block operator A_i and B_j as follows:

$$A_i(x, y) = \Phi_{A,i}(x)\Psi_{A,i}(y)^\top,$$
$$B_j(x, y) = \Phi_{B,j}(x)\Psi_{B,j}(y)^\top.$$

The finite-dimensional case. In the finite-dimensional case, the vectors $\Phi_B(x)$, $\Psi_B(y)$, $\Phi_A(y)$ and $\Psi_B(y)$ are obtained by concatenating the corresponding block embeddings as follows:

$$\begin{aligned} \Phi_A(x) &= \begin{pmatrix} \Phi_{A,1}(x) \\ \dots \\ \Phi_{A,\mathrm{K}}(x) \end{pmatrix}, \qquad \Psi_A(y) = \begin{pmatrix} \Psi_{A,1}(y) \\ \dots \\ \Psi_{A,\mathrm{K}}(y) \end{pmatrix}, \\ \Phi_B(x) &= \begin{pmatrix} \Phi_{B,1}(x) \\ \dots \\ \Phi_{B,\mathrm{L}}(x) \end{pmatrix}, \qquad \Psi_B(y) = \begin{pmatrix} \Psi_{B,1}(y) \\ \dots \\ \Psi_{B,\mathrm{L}}(y) \end{pmatrix}. \end{aligned} \tag{21}$$

The space of linear operators U corresponds to block-diagonal matrices $\Gamma = \mathrm{diag}(\Gamma_1, \dots, \Gamma_{\mathrm{L}})$, and the scalar product of theorem 4 coincides with the following block-wise matrix operation:

$$\langle \Gamma, B(x, y) \rangle = \langle \Phi_B(x), \Gamma \Psi_B(y) \rangle = \begin{pmatrix} \Phi_{B,1}(x) & \dots & \Phi_{B,\mathrm{L}}(x) \end{pmatrix} \begin{pmatrix} \Gamma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Gamma_{\mathrm{L}} \end{pmatrix} \begin{pmatrix} \Psi_{B,1}(y) \\ \vdots \\ \Psi_{B,\mathrm{L}}(y) \end{pmatrix}.$$

Similarly, the convex potentials take the form $\Xi = \mathrm{diag}(\Xi_1, \dots, \Xi_{\mathrm{K}})$, so that:

$$\langle \Xi, A(x, y) \rangle = \langle \Phi_A(x), \Xi \Psi_A(y) \rangle = \begin{pmatrix} \Phi_{A,1}(x) & \dots & \Phi_{A,\mathrm{K}}(x) \end{pmatrix} \begin{pmatrix} \Xi_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Xi_{\mathrm{K}} \end{pmatrix} \begin{pmatrix} \Psi_{A,1}(y) \\ \vdots \\ \Psi_{A,\mathrm{K}}(y) \end{pmatrix}.$$

Algebraic formalization. We rigorously define this finite-dimensional intuition of block-wise decomposition and generalizes it to Hilbert-valued operators A and B. Given a sequence of Hilbert vectors $u_1 \in \mathcal{H}_1, \dots, u_{\mathrm{L}} \in \mathcal{H}_{\mathrm{L}}$ living in distinct spaces, we use the following matrix notation to denote their direct sum (the natural formalization of vector concatenation):

$$\begin{pmatrix} u_1 \\ \dots \\ u_{\mathrm{L}} \end{pmatrix} := u_1 \oplus \cdots \oplus u_{\mathrm{L}} \quad \in \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_{\mathrm{L}}.$$

It comes with the natural scalar product:

$$\left\langle \begin{pmatrix} u_1 \\ \dots \\ u_{\mathrm{L}} \end{pmatrix}, \begin{pmatrix} v_1 \\ \dots \\ v_{\mathrm{L}} \end{pmatrix} \right\rangle = \langle u_1, v_1 \rangle + \cdots + \langle u_{\mathrm{L}}, v_{\mathrm{L}} \rangle .$$

For any $\mathcal{H}_{\mathcal{X}} := \mathcal{H}_{\mathcal{X},1} \oplus \cdots \oplus \mathcal{H}_{\mathcal{X},\mathrm{L}}$ and $\mathcal{H}_{\mathcal{Y}} := \mathcal{H}_{\mathcal{Y},1} \oplus \cdots \oplus \mathcal{H}_{\mathcal{Y},\mathrm{L}}$, let us denote the set of *block-wise* Hilbert-Schmidt operators from $\mathcal{H}_{\mathcal{Y}}$ to $\mathcal{H}_{\mathcal{X}}$ as:

$$\mathrm{HS}_{\oplus}(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{X}}) := \mathrm{HS}(\mathcal{H}_{\mathcal{Y},1}, \mathcal{H}_{\mathcal{X},1}) \oplus \cdots \oplus \mathrm{HS}(\mathcal{H}_{\mathcal{Y},n}, \mathcal{H}_{\mathcal{X},\mathrm{L}}) \subset \mathrm{HS}(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{X}}),$$

where any $\Gamma = \Gamma_1, \dots, \Gamma_n \in \mathrm{HS}_{\oplus}(\mathcal{H}_{\mathcal{Y}}, \mathcal{H}_{\mathcal{X}})$ acts linearly on $\mathcal{H}_{\mathcal{Y}}$ as follows:

$$\Gamma \begin{pmatrix} u_1 \\ \dots \\ u_{\mathrm{L}} \end{pmatrix} = \begin{pmatrix} \Gamma_1 u_1 \\ \dots \\ \Gamma_{\mathrm{L}} u_{\mathrm{L}} \end{pmatrix} \in \mathcal{H}_{\mathcal{X}}.$$

We now assume that A and B are block-separable in the sense of definition 2. First, note that block-separability is equivalent to the existence of separable functions $A_i : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}_{A,i}$ and $B_j : \mathcal{X} \times \mathcal{Y} \to \mathcal{H}_{B,j}$ such that, for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$:

$$A(x,y) = \begin{pmatrix} A_1(x,y) \\ \dots \\ A_{\mathrm{K}}(x,y) \end{pmatrix} \quad \text{and} \quad B(x,y) = \begin{pmatrix} B_1(x,y) \\ \dots \\ B_{\mathrm{L}}(x,y) \end{pmatrix}.$$

Since each A_i and B_j is separable, we can write for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and all $1 \le i \le \mathrm{K}$, $1 \le j \le \mathrm{L}$:

$$A_i(x,y) = \Phi_{A,i}(x)\Psi_{A,i}(y)^\top \quad \text{where} \quad \Phi_{A,i}(x) \in \mathcal{H}_{A,i}^{\Phi} \quad \text{and} \quad \Psi_{A,i}(y) \in \mathcal{H}_{A,i}^{\Psi},$$
$$B_j(x,y) = \Phi_{B,j}(x)\Psi_{B,j}(y)^\top \quad \text{where} \quad \Phi_{B,j}(x) \in \mathcal{H}_{B,j}^{\Phi} \quad \text{and} \quad \Psi_{B,j}(y) \in \mathcal{H}_{B,j}^{\Psi}.$$

Therefore, for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have:

$$A(x,y) = \begin{pmatrix} \Phi_{A,1}(x)\Psi_{A,1}(y)^\top \\ \dots \\ \Phi_{A,K}(x)\Psi_{A,K}(y)^\top \end{pmatrix} \quad \text{and} \quad B(x,y) = \begin{pmatrix} \Phi_{B,1}(x)\Psi_{B,1}(y)^\top \\ \dots \\ \Phi_{B,L}(x)\Psi_{B,L}(y)^\top \end{pmatrix}. \tag{22}$$

For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we define the concatenated embeddings $\Phi_A(x)$,$\Psi_A(y)$,$\Phi_B(x)$ and $\Psi_B(y)$ as in eq. (21) (where the matrix notation now represent the direct sum of the rows). With these definitions, theorem 4 becomes:

Theorem 7. *Let assume that A and B are block-separable with concatenated embeddings $\Phi_A \in \mathcal{H}_A^{\Phi}$, $\Psi_A \in \mathcal{H}_A^{\Psi}$, $\Phi_B \in \mathcal{H}_B^{\Phi}$ and $\Psi_B \in \mathcal{H}_B^{\Psi}$. Let $\mathbf{R}$ and $\mathbf{S}$ be as in theorem 3. Then,*

$$\mathcal{R}\mathrm{QM}(\alpha,\beta) \;=\; \min_{\Gamma \in \mathrm{HS}_{\oplus}(\mathcal{H}_B^{\Psi}, \mathcal{H}_B^{\Phi})} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \; -\int \langle \Phi_B(x), \Gamma\Psi_B(y) \rangle \, \mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma).$$

Similarly, with the notations of theorem 2, the following equivalences hold:

$$\mathcal{R}\mathrm{QM}(\alpha,\beta) \;=\; \inf_{\Gamma \in \mathrm{HS}_{\oplus}(\mathcal{H}_B^{\Psi}, \mathcal{H}_B^{\Phi})} \; \sup_{\Xi \in \mathrm{HS}_{\oplus}(\mathcal{H}_A^{\Psi}, \mathcal{H}_A^{\Phi})} \; \sup_{\phi \in \Phi} \; \mathcal{G}(\Gamma, \Xi, \phi)$$
$$\textit{with} \quad \mathcal{G}(\Gamma,\Xi,\phi) := \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\Gamma,\Xi},\, \alpha, \beta\big) + \tfrac{1}{2}\left\|\Gamma\right\|^2 - \tfrac{1}{2}\left\|\Xi\right\|^2$$

and

$$\mathcal{R}\mathrm{QM}(\alpha,\beta) \;=\; \inf_{\Gamma \in \mathrm{HS}_{\oplus}(\mathcal{H}_B^{\Psi}, \mathcal{H}_B^{\Phi})} \; \sup_{\Xi \in \mathrm{HS}_{\oplus}(\mathcal{H}_A^{\Psi}, \mathcal{H}_A^{\Phi})} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \; \mathcal{F}(\Gamma, \Xi, \phi),$$
$$\textit{with} \quad \mathcal{F}(\Gamma,\Xi,\phi) := \int c_{\Gamma,\Xi} \mathrm{d}\pi + \mathcal{R}(\pi) + \tfrac{1}{2}\left\|\Gamma\right\|^2 - \tfrac{1}{2}\left\|\Xi\right\|^2,$$

the effective cost $c_{\Gamma,\Xi}$ being equal to:

$$c_{\Gamma,\Xi}(x,y) = -\left\langle \begin{pmatrix} \Phi_A(x) \\ \Phi_B(x) \end{pmatrix}, \begin{pmatrix} -\Xi\Psi_A(y) \\ \Gamma\Psi_B(y) \end{pmatrix} \right\rangle.$$

We provide the proof of this result in section D, along with the proof of theorem 4. This formalism properly defines the block-wise operations that define the costs involved in this theorem. It notably highlights the fact that the space of transforms on which optimization is performed only contains a small fraction of all linear operators $\mathrm{HS}(\mathcal{H}_A^\Psi, \mathcal{H}_A^\Phi)$ and $\mathrm{HS}(\mathcal{H}_B^\Psi, \mathcal{H}_B^\Phi)$: this allows for efficient numerical implementations that we detail in section C.1.

B.4 Explicit decompositions of Gromov-Wasserstein kernels

As discussed in section B.2, for discrete inputs, eigendecomposition of the QM kernel is sufficient to obtain block-sparse embeddings. However, this process is costly and intractable on inputs with moderate size. For Gromov-Wasserstein and its variants, the kernel structure allows to compute embeddings without relying on its diagonalization, making GW particularly suited to computational applications. In some settings, the block-separability also holds in the infinite case, evidencing the naturality of the block-separable formalism when applied to Gromov-Wasserstein. In the following, we factorize GW kernels of the form:

$$k(x, y, x', y') = \big(c_\mathcal{X}(x, x') - c_\mathcal{Y}(y, y')\big)^2,$$

depending on the properties of the base costs $c_\mathcal{X}$ and $c_\mathcal{Y}$. Our computations rely on the following characterization lemma:

Lemma 2. *Let $k : \mathcal{C}((\mathcal{X} \times \mathcal{Y})^2)$ be symmetric and let A, B be two integral operators. The identity*

$$\frac{1}{2}\int k \,\mathrm{d}\pi \otimes \pi \;=\; \frac{1}{2}\left\|A[\pi]\right\|^2 - \frac{1}{2}\left\|B[\pi]\right\|^2$$

holds for all $\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ if and only if, for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$:

$$k(x, y, x', y') \;=\; \langle A(x, y), A(x', y')\rangle - \langle B(x, y), B(x', y')\rangle\,.$$

GW with inner products and squared norms. When $c_\mathcal{X}$ and $c_\mathcal{Y}$ are scalar products, k is a polynomial. After rearrangement of the monomials, we obtain the desired decomposition.

Proposition 3. *The scalar product kernel is equivalent to:*

$$\big(\langle x, x'\rangle^2 - \langle y, y'\rangle^2\big)^2 \;=\; \langle A(x, y), A(x', y')\rangle - \langle B(x, y), B(x', y')\rangle$$
$$\text{with} \quad A(x, y) = \begin{pmatrix} xx^\top \\ yy^\top \end{pmatrix} \text{ and } B(x, y) = \sqrt{2}\, xy^\top,$$

the matrix notation denoting the direct sum of the cells (as detailed in section B.3).

The block embeddings of A and B are obtained by separating the x and y terms of the formula:

$$\Phi_A(x) = \begin{pmatrix} xx^\top \\ 1 \end{pmatrix}, \quad \Psi_A(y) = \begin{pmatrix} 1 \\ yy^\top \end{pmatrix}, \quad \Phi_B(x) = \sqrt[4]{2}x \quad \text{and} \quad \Phi_B(y) = \sqrt[4]{2}y.$$

A similar decomposition can be obtained when costs are differences of scalar products:

Proposition 4. *Let $c_\mathcal{X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $c_\mathcal{Y} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ such that for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$:*

$$c_\mathcal{X}(x, x') = \langle x_+, x'_+\rangle - \langle x_-, x'_-\rangle \quad \textit{and} \quad c_\mathcal{Y}(y, y') = \langle y_+, y'_+\rangle - \langle y_-, y'_-\rangle\,.$$

Then,

$$\big(c_\mathcal{X}(x, x') - c_\mathcal{Y}(y, y')\big)^2 \;=\; \langle A(x, y), A(x', y')\rangle - \langle B(x, y), B(x', y')\rangle$$
$$\text{with} \quad A(x, y) = \begin{pmatrix} x_+x_+^\top \\ x_-x_-^\top \\ y_+y_+^\top \\ y_-y_-^\top \\ \sqrt{2}x_-y_+^\top \\ \sqrt{2}x_+y_-^\top \end{pmatrix} \quad \text{and} \quad B(x, y) = \begin{pmatrix} \sqrt{2}x_+x_-^\top \\ \sqrt{2}y_+y_-^\top \\ \sqrt{2}x_+y_+^\top \\ \sqrt{2}x_-y_-^\top \end{pmatrix}.$$

Once again, we can build the embeddings $\Phi_A(x)$, $\Psi_A(y)$, $\Phi_B(x)$ and $\Psi_B(y)$ by separating the terms in x and y adequately. Although difference of scalar products are never used explicitly in practice, most GW instances can be reduced to this cost structure. Notably, squared Euclidean norms can be rewritten in this form:

$$\|x - x'\|^2 = \langle x_+, x'_+ \rangle - \langle x_-, x'_- \rangle, \quad \text{with} \quad x_+ = \tfrac{1}{\sqrt{2}}(\|x\|^2 + 1) \text{ and } x_- = \begin{pmatrix} \sqrt{2}x \\ \tfrac{1}{\sqrt{2}}(\|x\|^2 - 1) \end{pmatrix}.$$

Therefore, the squared norm case (where $c_{\mathcal{X}}, c_{\mathcal{Y}} = \|\cdot\|^2$) admits an explicit block-separable factorization as well.

Remark 1 (The balanced case)**.** *When we restrict π to the set $\Pi(\alpha, \beta)$ of couplings (defined in eq.* (19)*), the kernel decompositions can be further simplified. Indeed, most terms of A and B depend only on x or on y and remain constant when the marginals of π are fixed. Since they do not play a role in the minimization, we can simplify the previous computations in the balanced case:*

$$\frac{1}{2}\int (\langle x, x' \rangle - \langle y, y' \rangle)^2 \mathrm{d}\pi \otimes \pi = \mathrm{Cste}(\alpha, \beta) - \left\| \int x y^\top \mathrm{d}\pi \right\|^2,$$

$$\frac{1}{2}\int (\|x - x'\|^2 - \|y - y'\|)^2 \mathrm{d}\pi \otimes \pi = \mathrm{Cste}(\alpha, \beta) - 4\left\| \int x y^\top \mathrm{d}\pi \right\|^2 - 2\int \|x\|^2 \|y\|^2 \mathrm{d}\pi,$$

giving the same factorization as in Rioux et al. (2024); Zhang et al. (2024). Similarly, the objective of proposition 4 becomes:

$$\frac{1}{2}\int (c_{\mathcal{X}}(x, x') - c_{\mathcal{Y}}(y, y'))^2 \mathrm{d}\pi \otimes \pi = \mathrm{Cste}(\alpha, \beta) + \left\| \int \begin{pmatrix} x_- y_+^\top \\ x_+ y_-^\top \end{pmatrix} \mathrm{d}\pi \right\|^2 - \left\| \int \begin{pmatrix} x_+ y_+^\top \\ x_- y_-^\top \end{pmatrix} \mathrm{d}\pi \right\|^2.$$

Embeddable costs. The previous decompositions are specific to scalar products or squared norms, but they can be extended to any costs that can be represented as a squared norm or scalar product in a certain embedding space. We refer to such costs as *embeddable costs*, that encompasses two important categories: costs of positive and conditionally of negative type. We briefly introduce them here, refering to (Bekka et al., 2007, Appendix C) for a detailed introduction to this rich theory.

Definition 3 (Cost of positive type)**.** *A cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is of positive type if it is symmetric and is such that for any finite collection of points $x_1, \ldots, x_n \in \mathcal{X}$ and coefficients $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$:*

$$\sum_{i=1}^n \lambda_i \lambda_j c(x_i, x_j) \geq 0.$$

Definition 4 (Cost conditionally of negative type)**.** *A non-negative cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is conditionally of negative type (CNT) if it is symmetric, satisfies $c(x, x) = 0$ for all $x \in \mathcal{X}$ and is such that for any finite collection of points $x_1, \ldots, x_n \in \mathcal{X}$ and coefficients $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$:*

$$\sum_{i=1}^n \lambda_i = 0 \implies \sum_{i=1}^n \lambda_i \lambda_j c(x_i, x_j) \leq 0\,.$$

Many important examples fall in these categories. First, note that these categories are stable by sum and rescaling by positive coefficients. Moreover, if two costs c_1, c_2 are positive, then $c_1 \cdot c_2$ is positive as well: in particular, any non-negative integer power c^k of a positive cost is positive. On the other hand, tree distances, hyperbolic geodesic distances, spherical distances and all (non-squared) Euclidean distances are CNT. If c is CNT, then c^p is CNT as well if $0 < p \leq 2$: in particular, $c(x, x') = \|x - x'\|^p$ if CNT for any positive p smaller than 2. Lastly, CNT and positive costs are connected by the following characterization:

Theorem 8 (Schoenberg (1938))**.** *Let $c \in \mathcal{C}(\mathcal{X} \times \mathcal{X})$ be symmetric and satisfies $c(x, x) = 0$ for all $x \in \mathcal{X}$. Let $x_0 \in \mathcal{X}$. The following conditions are equivalent:*

1. *The cost c is CNT,*
2. *The cost $\tilde{c} : x, x' \in \mathcal{X} \times \mathcal{X} \mapsto c(x, x_0) + c(x_0, x') - c(x, x')$ is positive,*
3. *The cost $e^{-\lambda c} : x, x' \in \mathcal{X} \times \mathcal{X} \mapsto \exp(-\lambda c(x, x'))$ is positive for every $\lambda \geq 0$.*

A remarkable property of these costs is that they are indeed embeddable:

Theorem 9 (Aronszajn (1950))**.** *Let $c \in \mathcal{C}(\mathcal{X} \times \mathcal{X})$ be symmetric. The cost c is positive if and only if there exists a real Hilbert space $\mathcal{H}$ with a continuous mapping $\varphi : \mathcal{X} \to \mathcal{H}$ such that:*

$$\text{For all } x, x' \in \mathcal{X}, \quad c(x, x') \;=\; \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} \,.$$

Theorem 10 (Schoenberg (1938))**.** *Let $c \in \mathcal{C}(\mathcal{X} \times \mathcal{X})$ be symmetric and satisfies $c(x, x) = 0$ for all $x \in \mathcal{X}$. The cost c is CNT if and only if there exists a real Hilbert space $\mathcal{H}$ and a continuous mapping $\varphi : \mathcal{X} \to \mathcal{H}$ such that:*

$$\text{For all } x, x' \in \mathcal{X}, \quad c(x, x') \;=\; \|\varphi(x) - \varphi(x')\|^2_{\mathcal{H}} \,.$$

Although these embeddings are potentially infinite-dimensional, when working with a finite number of points, we can always restrict them to a finite-dimensional subset of the initial Hilbert space. In practice, they can be computed by diagonalizing the cost matrix defined by $c(x, x')$ in the positive case or by $c(x, x_0) + c(x_0, x') - c(x, x')$ in the CNT case. For inputs of size N, these embeddings are N-dimensional. Their dimension can then be reduced using approximation techniques, as we explain in section C.2.

Arbitrary costs on finite spaces. When $\mathcal{X}$ and $\mathcal{Y}$ are finite, the spectral theorem allows us to reexpress any arbitrary cost as a difference of scalar product, allowing us to apply the computations we made in the Euclidean case:

Proposition 5. *Let assume that $\mathcal{X}$ is finite, and let $c : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$. Then, there exists two functions $\varphi_+ : \mathcal{X} \longrightarrow \mathbb{R}^{D_+}$ and $\varphi_- : \mathcal{X} \longrightarrow \mathbb{R}^{D_-}$ such that:*

$$\text{For all } x, x' \in \mathcal{X}, \quad c(x, x') = \langle \varphi_+(x), \varphi_+(x') \rangle - \langle \varphi_-(x), \varphi_-(x') \rangle \,.$$

Therefore, up to the diagonalization of the costs $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$, any finite instance of Gromov-Wasserstein fits in the setting of proposition 4. Unlike diagonalizing the operator k explicitely, diagonalizing the costs is numerically tractable and has a negligible computational cost compared to solving the matching problem: this result is the key to efficiently apply our solver to GW.

Weighted kernels. We finally provide an extension of our computations to reweighted kernels:

$$k_w(x, y, x', y') = \langle w_{\mathcal{X}}(x), w_{\mathcal{X}}(x') \rangle \, \langle w_{\mathcal{Y}}(y), w_{\mathcal{Y}}(y') \rangle \, k(x, y, x', y'),$$

where k is block-separable. The additional scalar terms adjust the matching penalization based on an affinity score between $x \leftrightarrow x'$ and $y \leftrightarrow y'$, offering some flexibility in the modeling of GW problems and allowing to prioritize local geometry more than plain GW. In notably models the case of frature matching of fig. 5, where $w_{\mathcal{X}}(x) = [\mathbf{1}_{x \in C_k}]_k$ is a hot encoding of the fragment C_k to which each point x belongs (and without weighting on the points of the non-fractured target $\mathcal{Y}$).

The embeddings of k_w can be obtained from those of k by taking the tensor product of each block coordinate with the weight vector $w_{\mathcal{X}}$ and $w_{\mathcal{Y}}$:

$$\Phi_A(x) = \begin{pmatrix} \Phi_{A,1}(x) \\ \dots \\ \Phi_{A,\mathrm{K}}(x) \end{pmatrix} \quad \Longrightarrow \quad \Phi_{A,w}(x) = \begin{pmatrix} \Phi_{A,1}(x) w_{\mathcal{X}}(x)^\top \\ \dots \\ \Phi_{A,\mathrm{K}}(x) w_{\mathcal{X}}(x)^\top \end{pmatrix},$$

$$\Phi_B(x) = \begin{pmatrix} \Phi_{B,1}(x) \\ \dots \\ \Phi_{B,\mathrm{L}}(x) \end{pmatrix} \quad \Longrightarrow \quad \Phi_{B,w}(x) = \begin{pmatrix} \Phi_{B,1}(x) w_{\mathcal{X}}(x)^\top \\ \dots \\ \Phi_{B,\mathrm{L}}(x) w_{\mathcal{X}}(x)^\top \end{pmatrix},$$

the embeddings $\Psi_A(y)$ and $\Psi_B(y)$ being defined similarly. More generally this construction also provides a recipe to compute the embeddings of product kernels $k' = k_1 k_2$, where both k_1 and k_2 are block-separable, since any block-separable kernel can be expressed as a sum of terms of the form $\langle w_{\mathcal{X}}(x), w_{\mathcal{X}}(x') \rangle \, \langle w_{\mathcal{Y}}(y), w_{\mathcal{Y}}(y') \rangle = \big\langle w_{\mathcal{X}}(x) w_{\mathcal{Y}}(y)^\top, w_{\mathcal{X}}(x') w_{\mathcal{Y}}(y')^\top \big\rangle$.

C Detailed Implementation

This section is dedicated to the practical implementation of our proposed solver. Specifically, we detail how the high-level description of fig. 1 translates into practical and modular algorithm, how

to perform computations as efficiently as possible, and how to initialize the potentials Γ_0, Ξ_0 in practice.

In the following, we assume that the inputs measures α and β are discrete with N and M points, respectively:

$$\alpha = \sum_{i=1}^{\mathrm{N}} \alpha_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{\mathrm{M}} \beta_j \delta_{y_j}.$$

We also assume that the Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ are Euclidean, since the spaces $\mathrm{Im}\,(A)$ and $\mathrm{Im}\,(B)$ are generated by the images of the finite-dimensional space $\mathcal{M}\,(\mathcal{X} \times \mathcal{Y}) \cong \mathbb{R}^{\mathrm{N}\times\mathrm{M}}$.

C.1 Fast Implementation of Block Updates

In the main article, we explain how block-separable kernels allow for efficient and modular implementations of our solver. We provide detail about this fact here, explaining how to exploit the block structure to keep computations as fast as possible.

The separable case. Before treating the general case, we first detail the algorithms when the operators A and B are separable to provide the intuition of this approach. If these operators satisfy:

$$A(x,y) = \Phi_A(x)\Psi_A(y)^\top \quad \text{and} \quad B(x,y) = \Phi_B(x)\Psi_B(y)^\top,$$

then for any plan π, $A[\pi] = \mathrm{Corr}_A[\pi]$ and $B[\pi] = \mathrm{Corr}_B[\pi]$ (following the notations of eq. (17)). Moreover, all $\mathcal{R}\mathrm{OT}$ computations take the scalar product form $c_{\Gamma,\Xi}(x,y) = -\langle x_{\text{eff}}, y_{\text{eff}}\rangle$ where x_{eff} and y_{eff} are given by:

$$x_{\text{eff}} = \begin{pmatrix} \Phi_{\mathcal{A}}(x) \\ \Phi_{\mathcal{B}}(x) \end{pmatrix} \quad \text{and} \quad y_{\text{eff}} = \begin{pmatrix} -\Xi\Psi_{\mathcal{A}}(y) \\ \Gamma\Psi_{\mathcal{B}}(y) \end{pmatrix}.$$

Algorithm 1 only interacts with the $\mathcal{R}\mathrm{OT}$ solver through these parameters x_{eff}, y_{eff} and through the cross-correlation operations involved in $A[\pi]$, $B[\pi]$. Therefore, we can encode all these operations in a black-box $\mathcal{R}\mathrm{OT}$ solver specialized to costs of the form $c(x,y) = -\langle x, y\rangle$ that we denote by `Sink` (by reference to the Sinkhorn algorithm that serves as our $\mathcal{R}\mathrm{OT}$ solver in our practical implementation). Our method then takes the form of algorithm 2, `Sink.optimization_step()` runs a few warm-started iterations of this solver and `Sink.Corr(`Φ, Ψ`)` returns $\int \Phi(x)\Psi(y)^\top \mathrm{d}\pi(x,y)$. Note that, in practice, OT solvers do not directly implement such cross-correlations, but rather transport integrals of the form:

$$v(x) = \int u(y)\mathrm{d}\pi(x,y) \quad \text{or} \quad v(y) = \int u(x)\mathrm{d}\pi(x,y). \tag{23}$$

Therefore, to compute the cross-correlation between two embedding vectors $\Phi(x) \in \mathbb{R}^{\mathrm{D}}$ and $\Psi(y) \in \mathbb{R}^{\mathrm{E}}$, two choices are available:

$$\int_{\mathcal{X}} \Phi(x) \left(\int_{\mathcal{Y}} (\Psi(y)^\top \mathrm{d}\pi(x,y)) \right) \quad \text{or} \quad \int_{\mathcal{Y}} \Psi(y)^\top \left(\int_{\mathcal{X}} (\Phi(x)\mathrm{d}\pi(x,y)) \right).$$

The inner integral bearing the main computational cost, it is more efficient to perform this integral on the vector with smallest dimension. We detail this procedure in algorithm 3, where `Sink.integrate` implements the integrals of eq. (23) (and is performed internally by the chosen OT backend). When using linear-memory OT backends, the cross-correlation computation has a complexity of $O(\mathrm{NM}\min(\mathrm{D},\mathrm{E}) + (\mathrm{N}+\mathrm{M})\mathrm{DE})$ in time and $O(\mathrm{ND}+\mathrm{ME})$ in memory.

Algorithm 3: `Sink.Corr` implementation.

Input: Sinkhorn solver `Sink`; $\Phi \in \mathbb{R}^{\mathrm{N}\times\mathrm{D}}$; $\Psi \in \mathbb{R}^{\mathrm{M}\times\mathrm{E}}$.
Output: $\Omega \in \mathbb{R}^{\mathrm{D}\times\mathrm{E}}$, cross-correlation of Φ and Ψ for the current transport plan of `Sink`.
if $\mathrm{D} \leq \mathrm{E}$ **then**
  $\overline{\Phi} \leftarrow$ `Sink.integrate(`Φ, axis $= 0$`)`
  $\Omega \leftarrow \sum_i \overline{\Phi}_i \Psi_i^\top$
else
  $\overline{\Psi} \leftarrow$ `Sink.integrate(`Ψ, axis $= 1$`)`
  $\Omega \leftarrow \sum_j \Phi_j \overline{\Psi}_j^\top$

Algorithm 2: $\mathcal{R}$QM solver (separable formulation)

Input: as in algorithm 1, with A, B given by Φ_A, Ψ_A and Φ_B, Ψ_B; inner iterations n_{inner}.
Output: (Γ, Ξ) and the solver `Sink`, whose potentials encode the plan $\widehat{\pi}$.
Initialize Γ_0, Ξ and the $\mathcal{R}$OT solver `Sink`
`Sink.x` $\leftarrow (\Phi_A(x), \Phi_B(x))^\top$
for $t = 0, 1, 2, \dots$ *until* $\|\Gamma_{t+1} - \Gamma_t\| \leq \delta$ **do**
  for $l = 1, \dots, n_{inner}$ **do**
    `Sink.y` $\leftarrow (-\Xi\Psi_A(y), \Gamma_t\Psi_B(y))^\top$
    `Sink.optimization_step()`
    $\Xi \leftarrow \Xi + \tau \left(\texttt{Sink.Corr}(\Phi_A(x), \Psi_A(y)) - \Xi\right)$
  $\Gamma_{t+1} \leftarrow \texttt{Sink.Corr}(\Phi_B(x), \Psi_B(y))$

Algorithm 2 shows that, in the separable case, all computations involve either standard matrix-vector operations or standard interactions with the black-box `Sink` solver, demonstrating the modularity of our method as advertised in the main article. The block-separable case yields heavier calculations, but the structure is similar. We explain it in the following, where we assume that A and B have the structure of eq. (22) and $D_{A,i}$, $E_{A,i}$, $D_{B,j}$, $E_{B,j}$ are the dimensions of $\Phi_{A,i}$, $\Psi_{A,j}$, $\Phi_{B,j}$, $\Psi_{B,j}$.

Block cross-correlations. The operators A and B can be expressed as block-wise cross-correlation operators:

$$A[\pi] = \text{Block-Corr}_A[\pi] := \begin{pmatrix} \mathrm{Corr_{A,1}}[\pi] \\ \dots \\ \mathrm{Corr_{A,K}}[\pi] \end{pmatrix} \quad \text{and } B[\pi] = \text{Block-Corr}_B[\pi] := \begin{pmatrix} \mathrm{Corr_{B,1}}[\pi] \\ \dots \\ \mathrm{Corr_{B,L}}[\pi] \end{pmatrix},$$

where for all i and j,

$$\mathrm{Corr_{A,i}}[\pi] = \int \Phi_{\mathrm{A,i}}(\mathrm{x})\Psi_{\mathrm{A,i}}(\mathrm{y})^\top \mathrm{d}\pi(\mathrm{x},\mathrm{y}) \quad \text{and} \quad \mathrm{Corr_{B,j}}[\pi] = \int \Phi_{\mathrm{B,j}}(\mathrm{x})\Psi_{\mathrm{B,j}}(\mathrm{y})^\top \mathrm{d}\pi(\mathrm{x},\mathrm{y}).$$

Therefore, each block of $A[\pi]$ and $B[\pi]$ can be computed using the same standard operation as for separable costs.

Effective cost parametrization. Similarly, the effective cost $c_{\Gamma,\Xi}$ can be computed block-wise. Rather than applying the potentials Ξ and Γ on only one of the embeddings $\Phi_A(x)$ or $\Psi_B(y)$ as in eq. (11), we can split the computation over both sides so that:

$$\langle \Xi, A(x,y) \rangle = \left\langle \begin{pmatrix} \Xi_{\mathcal{X},1}\Phi_{A,1}(x) \\ \dots \\ \Xi_{\mathcal{X},K}\Phi_{A,K}(x) \end{pmatrix}, \begin{pmatrix} \Xi_{\mathcal{Y},1}\Psi_{A,1}(y) \\ \dots \\ \Xi_{\mathcal{Y},K}\Psi_{A,K}(y) \end{pmatrix} \right\rangle$$

$$\text{for any} \quad \Xi_{\mathcal{X},i}, \Xi_{\mathcal{Y},i} \quad \text{such that} \quad \Xi = \begin{pmatrix} \Xi_{\mathcal{X},1}^\top \Xi_{\mathcal{Y},1} \\ \dots \\ \Xi_{\mathcal{X},K}^\top \Xi_{\mathcal{Y},K} \end{pmatrix},$$

with a similar formula for $\langle \Gamma, B(x,y) \rangle$. As before, we can write $c_{\Gamma,\Xi}(x,y) = -\langle x_{\text{eff}}, y_{\text{eff}} \rangle$, with

$$x_{\text{eff}} = \text{Block-Proj}_{\Gamma_\mathcal{X},\Xi_\mathcal{X}}(\Phi_A(x), \Phi_B(x)) := \begin{pmatrix} \Xi_{\mathcal{X},1}\Phi_{A,1}(x) \\ \dots \\ \Xi_{\mathcal{X},K}\Phi_{A,K}(x) \\ \Gamma_{\mathcal{X},1}\Phi_{B,1}(x) \\ \dots \\ \Gamma_{\mathcal{X},L}\Phi_{B,L}(x) \end{pmatrix}$$

$$\text{and} \quad y_{\text{eff}} = \text{Block-Proj}_{\Gamma_\mathcal{Y},\Xi_\mathcal{Y}}(\Psi_A(y), \Psi_B(y)) := \begin{pmatrix} -\Xi_{\mathcal{Y},1}\Psi_{A,1}(y) \\ \dots \\ -\Xi_{\mathcal{Y},K}\Psi_{A,K}(y) \\ \Gamma_{\mathcal{Y},1}\Psi_{B,1}(y) \\ \dots \\ \Gamma_{\mathcal{Y},L}\Psi_{B,L}(y) \end{pmatrix}.$$

The decomposition of Ξ and Γ on both x and y allows to minimize the dimension of the vectors x_{eff} and y_{eff}, since the cost of each Sinkhorn iteration scales linearly with this dimension. For each block, we therefore choose $\Xi_{\mathcal{X},i} = \Xi_i^\top$ and $\Xi_{\mathcal{Y},j} = 1$ if $D_{A,i} > E_{A,i}$, $\Xi_{\mathcal{X},i} = 1$ and $\Xi_{\mathcal{Y},j} = \Xi_i$ otherwise. Using these definitions, our method now takes the form of algorithm 4. The complexity of each operation is equal to $O(N^2(D_A \vee E_A + D_B \vee E_B))$ in time and $O(N(D_A \vee E_A + D_B \vee E_B))$ in memory, with:

$$D_A \vee E_A = \sum_j \min(D_{A,i}, E_{A,i}) \quad \text{and} \quad D_B \vee E_B = \sum_j \min(D_{B,j}, E_{B,j}),$$

avoiding the quadratic dependency in the dimension of naive implementations.

Algorithm 4: $\mathcal{R}\mathrm{QM}$ solver (block-separable formulation)

Input: as in algorithm 1, with A, B given by Φ_A, Ψ_A and Φ_B, Ψ_B; inner iterations n_{inner}.
Output: (Γ, Ξ) and the solver `Sink`, whose potentials encode the plan $\widehat{\pi}$.
Initialize Γ_0, Ξ and the $\mathcal{R}\mathrm{OT}$ solver `Sink`
for $t = 0, 1, 2, \dots$ *until* $\|\Gamma_{t+1} - \Gamma_t\| \leq \delta$ **do**
  for $l = 0, 1, 2, \dots n_{inner}$ **do**
    `Sink.x` $\leftarrow$ Block-Proj$_{\Gamma_k, \Xi_k}(\Phi_A(x), \Phi_B(x))$
    `Sink.y` $\leftarrow$ Block-Proj$_{\Gamma_k, \Xi_k}(\Psi_A(y), \Psi_B(y))$
    `Sink.optimization_step()`
    $\Xi \leftarrow \Xi + \tau \left(\texttt{Sink.Block-Corr}(\Phi_A(x), \Psi_A(y)) - \Xi\right)$
  $\Gamma_{k+1} \leftarrow$ `Sink.Block-Corr`$(\Phi_B(x), \Psi_B(y))$

Algorithm 5: `Sink.Block-Corr` implementation.

Input: Sinkhorn solver `Sink`; $\Phi = (\Phi_1, \dots, \Phi_K)$; $\Psi = (\Psi_1, \dots, \Psi_K)$.
Output: $\Omega = (\Omega_1, \dots, \Omega_K)$.
for $k = 0, \dots, K$ **do**
  $\Omega_k \leftarrow$ `Sink.Corr`(Φ_k, Ψ_k)
$\Omega \leftarrow (\Omega_1, \dots, \Omega_K)$

C.2 Advanced Implementation Techniques

Further memory optimization. In proposition 4, some blocks embeddings of A and B are terms of the form $f(x)g(x)^\top$ ($x_+x_+^\top$, $x_+x_-^\top$, etc). In algorithm 4, the only operations involving these embeddings take the form:

$$\int f(x)g(x)^\top \mathrm{d}\pi(x,y) = \int f(x)g(x)^\top \mathrm{d}\pi_1(x) \quad \text{and} \quad f(x)g(x)^\top \Xi = f(x)\big(g(x)^\top \Xi\big).$$

These are standard computations and efficiently implemented in scientific computing libraries, we can compute such integrals when needed without storing the tensor product explicitly, simply representing the corresponding embedding as a tuple $(f(x), g(x))$. This approach reduces the memory footprint of these embeddings from $\mathrm{N}\dim(f)\dim(g)$ to $\mathrm{N}(\dim(f)+\dim(g))$, significantly increasing scalability when f and g are high dimensional. Using the same techniques on y tems, this allows our Gromov-Wasserstein implementation to scale linearly in the input dimension for the Euclidean cases of propositions 3 and 4.

Low-dimensional embedding computations. The embeddings $\Phi_A(x)$, $\Psi_A(y)$, $\Phi_B(x)$ and $\Psi_B(y)$ of section B.4 are usually not known explicitly but can be obtained from the diagonalization of the cost matrices $[c_{\mathcal{X}}(x_i, x_k)]_{ik}$ and $[c_{\mathcal{Y}}(y_j, y_l)]_{jl}$. In general, these embeddings are high-dimensional and unsuited to large-scale applications. It is therefore beneficial to rely on dimension reduction techniques, such as truncated PCA. The KeOps library (Charlier et al., 2021) allows to perform truncated eigendecompositions efficiently and with linear memory footprint, preventing any memory overflow. Alternatively, approximations methods like Generalized Nyström (Nakatsukasa, 2020) provide low-dimensional approximations of these embeddings in linear time and memory. The cost of these methods is negligible compared to the rest of our pipeline.

Acceleration with multiscaling. The optimal potentials Γ and Ξ depend on the geometry of the inputs α and β, rather than their precise pointwise configuration. Therefore, optimal potentials (Γ_c^*, Ξ_c^*) obtained by replacing the inputs with coarse approximations $\alpha_c \approx \alpha$ and $\beta_c \approx \beta$ remain close to the true optimum (Γ^*, Ξ^*). This motivates a natural multiscale strategy: solve the coarse problem, then initialize the fine-scale problem using the coarse output (Γ_c^*, Ξ_c^*). Similarly, standard OT multiscaling techniques (Feydy et al., 2020) can be applied in order to extrapolate an initialization ϕ_0 from the coarse dual parameters ϕ_c^*. This multiscaling scheme substantially reduces the number of iterations needed for convergence. To construct the coarse approximation of the input, three choices are available. We can take a random subsample of the original points, use a k-means clustering to approximate the fine-scale measures, or use a predefined subset of the initial inputs.

Mask-based sparsification. In practice, optimal plans concentrate their mass on small fractions of all possible pairs, most of their coefficient being negligible. Some OT solvers exploit this behavior by accepting a sparsity mask that indicates which coefficients are expected to be nearly zero: they ignore them during computation, significantly fastening the resolution of the problem. In classical OT, two points that are far apart are unlikely to be matched, so that sparsity masks can be computed based on the position of points in the space. For quadratic matching, this technique does not apply directly since the two inputs generally live in uncomparable spaces. However, the multiscale strategy provides a natural alternative: since the coarse plan π_c^* approximates π^*, a near-zero entry $\pi_c^*(x_c, y_c) \approx 0$ indicates that any fine-scale pair (x, y) with $x \approx x_c$ and $y \approx y_c$ is also unlikely to be matched. This allows sparsity masks to be estimated from the coarse solution, further accelerating the fine-scale solution. We describe the exact procedure in algorithm 6.

Algorithm 6: Sparsity mask computation

Input: Coarse inputs x_i^c, y_j^c, fine inputs x_i, y_j; optimal coarse potentials Ξ^c, Γ^c; optimal coarse plan π^c; tolerance parameter $\delta > 0$.
Output: Sparsity mask $M \in \{0,1\}^{\mathrm{N}\times\mathrm{M}}$.

$\tilde{x} \leftarrow (\Phi_A(x), \Phi_B(x))^\top$
$\tilde{y} \leftarrow (-\Xi^c \Psi_A(y), \Gamma^c \Psi_B(y))^\top$
$\tilde{x}^c \leftarrow (\Phi_A(x^c), \Phi_B(x^c))^\top$
$\tilde{y}^c \leftarrow (-\Xi^c \Psi_A(y^c), \Gamma^c \Psi_B(y^c))^\top$
$I \leftarrow [\arg\min_u \|\tilde{x}_i - \tilde{x}_u^c\|^2]_i$
$J \leftarrow [\arg\min_v \|\tilde{y}_j - \tilde{y}_v^c\|^2]_j$
$M \leftarrow [\pi^c_{I(i)J(j)} > \delta]_{ij}$.

C.3 Initialization of potentials

The initialization of the solver is a crucial element for matching problems. Indeed, since $\mathcal{R}\mathrm{QM}$ is non-convex, the output of our solver depends on the initial potential $\Gamma_0 \in \mathcal{H}_B$. Moreover, although the initialization of Ξ_0 does not influence the local minimum reached by our method, it can significantly accelerate convergence and must also be taken into account. Our framework allows for various choices of initializations that we detail here. We provide the formulas for Γ_0, which can be straightforwardly adapted to initializations of Ξ_0. We focus on the separable case, since each block of a block-separable potential can be initialized using these formulas.

Trivial initialization. The classical choice in the literature is to initialize the optimization at the trivial plan $\pi = \alpha \otimes \beta$. We then choose Γ_0 as the corresponding cross-correlation:

$$\Gamma_0 = \left(\int \Phi(x) \mathrm{d}\alpha(x)\right)\left(\int \Psi(y) \mathrm{d}\beta(y)\right)^\top.$$

This choice has the advantage to not induce any bias in the optimization trajectory. Empirically, it generally provides satisfying results in the balanced setting but tends to be poorly suited to the unbalanced or in challenging settings (e.g. when inputs have many symmetries).

Landmarks. If some correspondences are already known between some remarkable points, or *landmarks*, we can exploit this information to initialize our algorithm more accurately. Given a

sequence $(x_1, y_1), \dots, (x_K, y_K)$ of landmarks, we simply set:

$$\Gamma_0 = \frac{1}{K} \sum_i \Phi(x_i)\Psi(y_i)^\top.$$

This initialization guides the optimization towards a plan that matches each x_i to a point close to y_i.

Random initialization. When no prior information is known about the optimal matching, the optimization landscape can be explored by sampling the space of possible Γ and keep the best solution over several random initializations. Let $\Phi(x) \in \mathbb{R}^{\mathrm{D}}$ and $\Phi(y) \in \mathbb{R}^{\mathrm{E}}$. Let $M(\Phi), M(\Psi)$ and $V(\Phi), V(\Psi)$ be the mean and coordinate-wise variance of Φ, Ψ for the measure α, β:

$$M(\Phi) = \int \Phi \mathrm{d}\alpha \quad \text{and} \quad V(\Phi) = \left[\int (\Phi_d - M(\Phi_d))^2 \mathrm{d}\alpha\right]_{1 \le d \le \mathrm{D}}$$
$$M(\Psi) = \int \Psi \mathrm{d}\beta \quad \text{and} \quad V(\Psi) = \left[\int (\Psi_e - M(\Psi)_e)^2 \mathrm{d}\beta\right]_{1 \le e \le \mathrm{E}}$$

For any $\Gamma \in \mathcal{H}_B$, if Γ is the cross-correlation matrix of a coupling $\pi \in \Pi(\alpha, \beta)$, then:

$$\text{For all } 1 \le d \le \mathrm{D} \quad \text{and} \quad 1 \le e \le \mathrm{E}, \quad |\Gamma_{de} - M(\Phi)_d M(\Psi)_e| \le \sqrt{V(\Phi)_d V(\Psi)_e}.$$

Therefore, we can sample each coordinate $[\Gamma_0]_{de}$ uniformly from the interval $[M(\Phi)_d M(\Psi)_e \pm \sqrt{V(\Phi)_d V(\Psi)_e}]$. These bounds hold in the balanced case $\pi \in \Pi(\alpha, \beta)$. They remain true in the unbalanced setting as well if we know that optimal plans have smaller margins than the inputs α, β, i.e. they can destroy mass but not create it.

D PROOFS

D.1 PROOFS OF SECTION 2

Theorem 1. *If $\mathcal{E}$ admits the structure of eq.* (3)*, then eq.* (1) *is equivalent to:*

$$\mathrm{M}(\alpha, \beta) \;=\; \inf_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} - \int c_{\mathcal{B}}(x, y) \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}(c_{\mathcal{B}}), \tag{4}$$

where $\mathbf{R}(\pi) = \mathcal{A} + \mathcal{R}$ and $\mathbf{S} = \mathcal{B}^*$ are convex functions. If $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean spaces, then there exists two Hilbert-valued continuous embeddings $\Phi : \mathcal{X} \to \mathcal{H}$ and $\Psi : \mathcal{Y} \to \mathcal{H}$, a convex set $U \subset \mathrm{HS}(\mathcal{H}, \mathcal{H})$ and a convex function $\mathbf{S}'$ such that eq. (1) *becomes:*

$$\mathrm{M}(\alpha, \beta) \;=\; \inf_{\Gamma \in U} \; \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} - \int \langle \Phi(x), \Gamma \Psi(y) \rangle \, \mathrm{d}\pi(x, y) + \mathbf{R}(\pi) + \mathbf{S}'(\Gamma). \tag{5}$$

Proof. Since $\mathcal{B}$ is a proper convex lower semi-continuous function, it is equal to its biconjugate: by the Fenchel-Moreau theorem, $\mathcal{B}^{**} = \mathcal{B}$. Therefore,

$$\text{For all } \pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}), \quad \mathcal{B}(\pi) = \sup_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \int c_{\mathcal{B}}(x, y) \mathrm{d}\pi(x, y) - \mathcal{B}^*(c_{\mathcal{B}}).$$

Consequently,

$$\mathcal{A}(\pi) - \mathcal{B}(\pi) + \mathcal{R}(\pi) = \inf_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} - \int c_{\mathcal{B}}(x, y) \mathrm{d}\pi(x, y) + \mathcal{A}(\pi) + \mathcal{R}(\pi) + \mathcal{B}^*(c_{\mathcal{B}}),$$

providing the first equation of the result. If $\mathcal{X} \subset \mathbb{R}^{\mathrm{D}}$ and $\mathcal{Y} \subset \mathbb{R}^{\mathrm{E}}$ are compact subset of Euclidean spaces, then $\mathcal{X} \times \mathcal{Y}$ is a compact subset of $\mathbb{R}^{\mathrm{D} \times \mathrm{E}}$.

To prove the second part of the theorem, we first use the fact that sums of separable functions $x, y \mapsto \sum_i f_i(x) g_i(y)$ are dense in $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$. Moreover, by the Stone–Weierstrass theorem, polynomial functions are dense in $\mathcal{C}(\mathcal{X})$ and in $\mathcal{C}(\mathcal{Y})$ for the uniform convergence. Since we assumed that $\mathcal{B}^*$ is continuous for the uniform convergence, we can restrict the infimum to polynoms of the form:

$$\text{For all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}, \quad p_\Gamma(x, y) = \sum_{d,e} \Gamma_{d,e} \Phi_d(x) \Psi_e(y),$$

where for all sequences of exponents $d = (d_1, \dots, d_{\mathrm{D}})$ and $e = (e_1, \dots, e_{\mathrm{E}})$, we define:

$$\Phi_d : x \in \mathcal{X} \mapsto x_1^{d_1} \dots x_{\mathrm{D}}^{d_{\mathrm{D}}} \in \mathbb{R} \quad \text{and} \quad \Psi_e : y \in \mathcal{Y} \mapsto y_1^{e_1} \dots y_{\mathrm{E}}^{e_{\mathrm{E}}} \in \mathbb{R},$$

and $\Gamma_{d,e}$ is a collection of real values with only a finite number of non-zero elements. Since the sequences involved in the previous decompositions are countable, and up to a reweighting of the terms $\Gamma_{d,e}$, we can write the sequences of monomials as sequences:

$$\Phi(x) = \left(\frac{1}{d_1! \dots d_{\mathrm{D}}!} \Phi_d(x) \right)_{d \in \mathbb{N}_0^{\mathrm{D}}} \quad \text{and} \quad \Phi(y) = \left(\frac{1}{e_1! \dots e_{\mathrm{E}}!} \Psi_e(y) \right)_{e \in \mathbb{N}_0^{\mathrm{E}}},$$

where the reweighting ensures that these sequences are squared-integrable. Therefore, for all x and y, $\Phi(x), \Phi(y) \in l^2(\mathbb{R})$ where $l^2(\mathbb{R})$ denotes the Hilbert space:

$$l^2(\mathbb{R}) := \left\{ (u_n)_{n \geq 0} \in \mathbb{R}^{\mathbb{N}} \quad \text{s.t.} \quad \sum_n u_n^2 < +\infty \right\}.$$

Similarly, if only a finite number of $\Gamma_{d,e}$ are non-zero, the following linear transform is a Hilbert-Schmidt operator of $l^2(\mathbb{R})$:

$$\Gamma : u \in l^2(\mathbb{R}) \longmapsto \Gamma u := \left(\sum_d \Gamma_{d,e} u_d \right)_e \in l^2(\mathbb{R}).$$

With these notations, the previous polynom writes:

$$p_\Gamma(x, y) = \langle \Phi(x), \Gamma \Psi(y) \rangle .$$

By change of variable $c_{\mathcal{B}} \leftarrow p_\Gamma$ in eq. (4), we obtain the desired formula. Since the function $\Gamma \mapsto p_\Gamma$ is linear, we also have the convexity of $\mathbf{S}'(\Gamma) = \mathbf{S}(p_\Gamma)$, proving the theorem. □

Theorem 2. *Let us assume that $\mathcal{R}$ is a proper lower semi-continuous convex function. Let Φ be its convex set of dual variables and $\mathcal{D}_{\mathcal{R}}$ its dual objective that satisfies, for all $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$:*

$$\min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \int c \mathrm{d}\pi + \mathcal{R}(\pi) = \sup_{\phi \in \Phi} \mathcal{D}_{\mathcal{R}}(\phi;\, c, \alpha, \beta). \tag{7}$$

The matching problem of eq. (3) *is then equivalent to the following equations:*

$$\inf_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \sup_{c_{\mathcal{A}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \sup_{\phi \in \Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\mathcal{A}} - c_{\mathcal{B}},\, \alpha, \beta\big) + \mathcal{B}^*(c_{\mathcal{B}}) - \mathcal{A}^*(c_{\mathcal{A}}), \tag{8}$$

$$\inf_{c_{\mathcal{B}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \sup_{c_{\mathcal{A}} \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \int (c_{\mathcal{A}} - c_{\mathcal{B}}) \mathrm{d}\pi + \mathcal{R}(\pi) + \mathcal{B}^*(c_{\mathcal{B}}) - \mathcal{A}^*(c_{\mathcal{A}}). \tag{9}$$

If $\mathcal{X}$ and $\mathcal{Y}$ are compact subsets of Euclidean spaces, then there exists a sequence of continuous Hilbert embeddings $\Phi_{\mathcal{A}}, \Phi_{\mathcal{B}} : \mathcal{X} \to \mathcal{H}$ and $\Psi_{\mathcal{A}}, \Psi_{\mathcal{B}} : \mathcal{Y} \to \mathcal{H}$ such that these equations become:

$$\begin{aligned} \mathrm{M}(\alpha, \beta) \;=\; & \inf_{\Gamma \in U} \sup_{\Xi \in U} \sup_{\phi \in \Phi} \mathcal{G}(\Gamma, \Xi, \phi) \;=\; \inf_{\Gamma \in U} \sup_{\Xi \in U} \min_{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})} \mathcal{F}(\Gamma, \Xi, \phi), \\ & \textit{with} \quad \mathcal{G}(\Gamma, \Xi, \phi) := \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\Gamma,\Xi},\, \alpha, \beta\big) + \mathbf{S}'(\Gamma) - \mathbf{S}''(\Xi) \\ & \textit{and} \quad \mathcal{F}(\Gamma, \Xi, \phi) := \int c_{\Gamma,\Xi} \mathrm{d}\pi + \mathcal{R}(\pi) + \mathbf{S}'(\Gamma) - \mathbf{S}''(\Xi), \end{aligned} \tag{10}$$

where $U \subset \mathrm{HS}(\mathcal{H}, \mathcal{H})$ is a convex set, $\mathbf{S}'$ and $\mathbf{S}''$ are convex functions and $c_{\Gamma,\Xi}$ is the following effective cost (where the matrix notation indicates the direct sum, or concatenation, of its rows):

$$c_{\Gamma,\Xi}(x, y) = -\left\langle \begin{pmatrix} \Phi_{\mathcal{A}}(x) \\ \Phi_{\mathcal{B}}(x) \end{pmatrix}, \begin{pmatrix} -\Xi \Psi_{\mathcal{A}}(y) \\ \Gamma \Psi_{\mathcal{B}}(y) \end{pmatrix} \right\rangle. \tag{11}$$

The proof is a consequence of the classical convex duality theorem:

Theorem 11 (Rockafellar–Fenchel Theorem)**.** *Let $\mathcal{F} : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{G} : \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R} \cup \{+\infty\}$ be proper convex lower semi-continuous functions. Let assume that there is a $\pi_0 \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ such that both $\mathcal{F}(\pi_0), \mathcal{G}(\pi_0)$ are finite and $\mathcal{G}$ is continuous at π_0. Then,*

$$\inf_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \mathcal{F}(\pi) + \mathcal{G}(\pi) \;=\; \sup_{c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})} -\mathcal{F}^*(-c) - \mathcal{G}^*(c).$$

Proof of theorem 2. We apply the Rockafellar–Fenchel theorem with:

$$\mathcal{F}(\pi) = -\int c_{\mathcal{B}} \mathrm{d}\pi + \mathcal{R}(\pi) + \iota_{\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})}(\pi) \quad \text{and} \quad \mathcal{G}(\pi) = \mathcal{A}(\pi).$$

By hypothesis on $\mathcal{A}$, $\mathcal{G}$ is continuous on its domain. Moreover, for any $\pi_0 \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$ such that $\mathcal{E}(\pi_0) + \mathcal{R}(\pi_0) < +\infty$, both $\mathcal{F}(\pi_0)$ and $\mathcal{G}(\pi_0)$ are finite. Therefore, the Rockafellar–Fenchel hypotheses are indeed satisfied. By definition of $\mathcal{D}_{\mathcal{R}}$, we have for all $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$:

$$-\mathcal{F}^*(-c) = \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c \mathrm{d}\pi - \int c_{\mathcal{B}} \mathrm{d}\pi + \mathcal{R}(\pi) = \sup_{\phi\in\Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c - c_{\mathcal{B}},\, \alpha, \beta\big).$$

Therefore:

$$\min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int c_{\mathcal{B}} \mathrm{d}\pi + \mathcal{R}(\pi) + \mathcal{A}(\pi) = \sup_{c_{\mathcal{A}}\in\mathcal{C}(\mathcal{X}\times\mathcal{Y})} \sup_{\phi\in\Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\mathcal{A}} - c_{\mathcal{B}},\, \alpha, \beta\big) - \mathcal{A}^*(c_{\mathcal{A}}),$$

providing the first identity. To prove the second one, we simply remark that, by definition of $\mathcal{D}_{\mathcal{R}}$:

$$\sup_{\phi\in\Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\mathcal{A}} - c_{\mathcal{B}},\, \alpha, \beta\big) = \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \int (c_{\mathcal{A}} - c_{\mathcal{B}}) \mathrm{d}\pi + \mathcal{R}(\pi).$$

Finally, we prove the last part of the theorem by using the same technique as in theorem 1, approximating continuous functions on $\mathcal{X}\times\mathcal{Y}$ by polynoms. □

D.2 Proofs of Section 3

The results of this section are based on the explicit convex conjugate of convex quadratic forms:

Lemma 3. *Let $A : \mathcal{M}(X\times Y) \to \mathcal{H}_A$ be an integral operator. The convex conjugate of the quadratic operator $\mathcal{A}(\pi) := \frac{1}{2}\|A[\pi]\|^2$ is*

$$\mathcal{A}^*(c) \;=\; \begin{cases} \frac{1}{2}\|A[\pi]\|^2 & \text{if } c = A^\top A\pi, \\ +\infty & \text{if } c \notin \operatorname{Im}\left(A^\top A\right). \end{cases}$$

Proof. Let $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$. By definition,

$$\mathcal{A}^*(c) = \sup_{\pi\in\mathcal{M}(X\times Y)} f(\pi) \quad \text{where} \quad f(\pi) = \int c \mathrm{d}\pi - \tfrac{1}{2}\|A[\pi]\|^2.$$

The function f is differentiable with $\nabla f(\pi) = c - A^\top A[\pi]$: its supremum is attained if and only if there is a π such that $\nabla f(\pi) = 0$, i.e. $c = A^\top A[\pi]$. In that case, a direct evaluation of f shows that:

$$\mathcal{A}^*(c) = \int (A^\top A[\pi]) \mathrm{d}\pi - \tfrac{1}{2}\|A[\pi]\|^2 = \tfrac{1}{2}\|A[\pi]\|^2.$$

Otherwise, the function f do not admit any maximum, and its supremum is $+\infty$. □

Theorem 3. *Let assume that k is factorizable as in eq.* (3)*. Then, the following identity holds:*

$$\mathcal{R}\mathrm{QM}(\alpha,\beta) = \min_{\Gamma\in\mathrm{Im}(B)} \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int \langle \Gamma, B(x,y)\rangle \,\mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma), \tag{14}$$

with $\mathbf{R}(\pi) = \frac{1}{2}\|A[\pi]\|^2 + \mathcal{R}(\pi)$ and $\mathbf{S}(\Gamma) = \frac{1}{2}\|\Gamma\|^2$. Moreover:

$$\mathcal{R}\mathrm{QM}(\alpha,\beta) = \min_{\Gamma\in\mathrm{Im}(B)} \max_{\Xi\in\mathrm{Im}(A)} \sup_{\phi\in\Phi} \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\Gamma,\Xi},\, \alpha, \beta\big) - \tfrac{1}{2}\|\Xi\|^2 + \tfrac{1}{2}\|\Gamma\|^2, \tag{15}$$

with the effective cost $c_{\Gamma,\Xi}(x,y) := \langle \Xi, A(x,y)\rangle - \langle \Gamma, B(x,y)\rangle$.

Proof. We apply eq. (4) to the quadratic case, plugging the convex conjugate of $\mathcal{B} := \frac{1}{2}\|B[\pi]\|^2$:

$$\min_{\pi'\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \min_{\pi\in\mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int (B^\top B[\pi']) \mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \frac{1}{2}\|B[\pi']\|^2,$$

where we used the fact that the expression is finite only if there exists a $\pi' \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ such that $c_{\mathcal{B}} = B^\top B[\pi']$. By change of variable $\Gamma \leftarrow B[\pi']$, we then write:

$$\min_{\Gamma \in \mathrm{Im}(B)} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int (B^\top \Gamma) \mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \frac{1}{2}\|\Gamma\|^2 .$$

Since $(B^\top\Gamma)(x,y) = \langle \Gamma, B(x,y)\rangle$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we obtain the desired formula. The second identity of the theorem is a consequence of theorem 2, using the same changes of variable as above ($\Xi \leftarrow A[\pi']$). □

Theorem 4. *If B is block-separable, then let $\mathbf{R}$ and $\mathbf{S}$ be as in theorem 3. There exists embeddings $\Phi_{\mathcal{B}} : \mathcal{X} \to \mathcal{H}_\Phi$, $\Psi_{\mathcal{B}} : \mathcal{Y} \to \mathcal{H}_\Psi$ and a compact subspace $U \subset \mathrm{HS}(\mathcal{H}_\Phi, \mathcal{H}_\Psi)$ such that:*

$$\mathcal{RQM}(\alpha,\beta) = \min_{\Gamma \in U} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int \langle \Phi_{\mathcal{B}}(x), \Gamma \Psi_{\mathcal{B}}(y)\rangle \,\mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma). \tag{16}$$

If both A and B are block-separable, there exists embeddings $\Phi_{\mathcal{A}}$,$\Phi_{\mathcal{B}}$,$\Psi_{\mathcal{A}}$ and $\Phi_{\mathcal{B}}$ such that the effective cost $c_{\Gamma,\Xi}$ in eq. (10) *takes the scalar product form of eq.* (11) *and $\mathbf{S}', \mathbf{S}'' = \frac{1}{2}\|\cdot\|^2$.*

We state a more precise version of theorem 4 in theorem 7:

Theorem 7. *Let assume that A and B are block-separable with concatenated embeddings $\Phi_A \in \mathcal{H}_A^\Phi$, $\Psi_A \in \mathcal{H}_A^\Psi$, $\Phi_B \in \mathcal{H}_B^\Phi$ and $\Psi_B \in \mathcal{H}_B^\Psi$. Let $\mathbf{R}$ and $\mathbf{S}$ be as in theorem 3. Then,*

$$\mathcal{RQM}(\alpha,\beta) \;=\; \min_{\Gamma \in \mathrm{HS}_\oplus(\mathcal{H}_B^\Psi, \mathcal{H}_B^\Phi)} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int \langle \Phi_B(x), \Gamma \Psi_B(y)\rangle \,\mathrm{d}\pi(x,y) + \mathbf{R}(\pi) + \mathbf{S}(\Gamma).$$

Similarly, with the notations of theorem 2, the following equivalences hold:

$$\mathcal{RQM}(\alpha,\beta) \;=\; \inf_{\Gamma \in \mathrm{HS}_\oplus(\mathcal{H}_B^\Psi, \mathcal{H}_B^\Phi)} \sup_{\Xi \in \mathrm{HS}_\oplus(\mathcal{H}_A^\Psi, \mathcal{H}_A^\Phi)} \sup_{\phi \in \Phi} \mathcal{G}(\Gamma,\Xi,\phi)$$

$$\textit{with} \quad \mathcal{G}(\Gamma,\Xi,\phi) := \mathcal{D}_{\mathcal{R}}\big(\phi;\, c_{\Gamma,\Xi},\, \alpha, \beta\big) + \tfrac{1}{2}\|\Gamma\|^2 - \tfrac{1}{2}\|\Xi\|^2$$

and

$$\mathcal{RQM}(\alpha,\beta) \;=\; \inf_{\Gamma \in \mathrm{HS}_\oplus(\mathcal{H}_B^\Psi, \mathcal{H}_B^\Phi)} \sup_{\Xi \in \mathrm{HS}_\oplus(\mathcal{H}_A^\Psi, \mathcal{H}_A^\Phi)} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} \mathcal{F}(\Gamma,\Xi,\phi),$$

$$\textit{with} \quad \mathcal{F}(\Gamma,\Xi,\phi) := \int c_{\Gamma,\Xi} \mathrm{d}\pi + \mathcal{R}(\pi) + \tfrac{1}{2}\|\Gamma\|^2 - \tfrac{1}{2}\|\Xi\|^2 ,$$

the effective cost $c_{\Gamma,\Xi}$ being equal to:

$$c_{\Gamma,\Xi}(x,y) = -\left\langle \begin{pmatrix} \Phi_A(x) \\ \Phi_B(x) \end{pmatrix}, \begin{pmatrix} -\Xi\Psi_A(y) \\ \Gamma\Psi_B(y) \end{pmatrix} \right\rangle .$$

Proof. In the separable case, the first identity is a direct consequence of eq. (14) and the identity:

$$\langle \Gamma, B(x,y)\rangle = \left\langle \Gamma, \Phi_B(x)\Psi_B(y)^\top \right\rangle = \langle \Phi_B(x), \Gamma\Psi_B(y)\rangle .$$

The block-separable case is treated analogously. Let us write $\|B\| = \sum_{i=1}^{\mathrm{L}} \|B_i\|$ with $B_i(x,y) = \Phi_{B,i}(x)\Psi_{B,i}(y)^\top$ for all $1 \le i \le \mathrm{L}$ and all $x \in \mathcal{X}$, $y \in \mathcal{Y}$. Then, eq. (14) is equivalent to:

$$\min_{\Gamma_1 \in \mathrm{Im}(B_1)} \cdots \min_{\Gamma_{\mathrm{L}} \in \mathrm{Im}(B_{\mathrm{L}})} \min_{\pi \in \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})} -\int \sum_{i=1}^{\mathrm{L}} \langle \Phi_{B,i}(x), \Gamma_i \Psi_{B,i}(y)\rangle \,\mathrm{d}\pi(x,y)$$

$$+ \mathbf{R}(\pi) + \tfrac{1}{2}\sum_{i=1}^{\mathrm{L}} \|\Gamma_i\|^2 .$$

Using the definitions of section B.3, we directly obtain the first identity theorem 7. The other identities follow from theorem 3, using the above identity for A. □

D.3 Proofs of the Appendix

Proposition 1. *Let assume that $\mathcal{E}$ is continuous for the weak-* convergence and goes to infinity at infinity. Then, for any precision $\delta > 0$, there exists a bounded convex set $V \subset \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ containing the minima of $\mathcal{E}$ and two convex functions $\mathcal{A}_\delta$, $\mathcal{B}_\delta$ continuous on their domain such that:*

$$\text{For all } \pi \in V, \quad |\mathcal{E}(\pi) - \mathcal{A}_\delta(\pi) + \mathcal{B}_\delta(\pi)| < \delta,$$

and $\mathcal{A}_\delta(\pi) - \mathcal{B}_\delta(\pi) = +\infty$ for $\pi \notin V$.

Proof. First, since $\mathcal{E}$ goes to infinity at infinity, all its minima are contained in a closed bounded ball V centered on the origin, which is a convex set of $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$. It is compact for the weak-* convergence, since it is bounded and closed. Let $\mathrm{DC}(V) \subset \mathcal{C}(V)$ be the set functions on V that are decomposable as a difference of continuous convex functions. It satisfies the hypotheses of the Stone-Weierstrass theorem:

- All constant functions belong to $\mathrm{DC}(V)$.
- Linear functions belong to $\mathrm{DC}(V)$: therefore, for all $\pi, \pi' \in V$, there exists $\mathcal{F} \in \mathrm{DC}(V)$ such that $\mathcal{F}(\pi) \neq \mathcal{F}(\pi')$.
- The set $\mathrm{DC}(V)$ is trivially stable by addition and multiplication by a scalar.
- If $\mathcal{F}_1$ and $\mathcal{F}_2$ are in $\mathrm{DC}(V)$, then the product $\mathcal{F}_1\mathcal{F}_2$ belongs to $\mathrm{DC}(V)$ as well.

To prove the last property, we first write:

$$\mathcal{F}_1\mathcal{F}_2 = \tfrac{1}{2}(\mathcal{F}_1 + \mathcal{F}_2)^2 - \tfrac{1}{2}(\mathcal{F}_1 - \mathcal{F}_2)^2.$$

We only need to prove that squared DC functions are also DC. To do so, we remark that if $\mathcal{F} \in \mathrm{DC}(V)$ we can write $\mathcal{F} = \mathcal{A} - \mathcal{B}$ where $\mathcal{A}$ and $\mathcal{B}$ are both *positive* continuous convex functions. Indeed, since they are continuous, these functions are bounded on the compact set V: if we add the same sufficiently large constant to $\mathcal{A}$ and $\mathcal{B}$, we ensure their positivity while keeping their difference identical. Then,

$$\mathcal{F}^2 = 2(\mathcal{A}^2 + \mathcal{B}^2) - (\mathcal{A} + \mathcal{B})^2 \in \mathrm{DC}(V),$$

since all the terms of the sum are convex as the square of positive convex functions.

Since V is compact for the weak-* convergence, the Stone-Weierstrass theorem states that $\mathrm{DC}(V)$ is dense in $\mathcal{C}(V)$. Therefore, for any weak-* continuous function $\mathcal{E}$ on $\mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$ and any $\delta > 0$, there are two convex functions $\mathcal{A}_\delta$ and $\mathcal{B}_\delta$ satisfying the desired approximation on V. By replacing $\mathcal{A}_\delta$ with $\mathcal{A}_\delta + \iota_V$, we obtain the desired result. □

Lemma 1. *Let assume that $\mathcal{A}$ is superlinear:*

$$\frac{\mathcal{A}(\pi_n)}{\mathrm{TV}(\pi_n)} \longrightarrow +\infty \quad \textit{for all } (\pi_n) \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}) \quad \textit{such that} \quad \mathrm{TV}(\pi_n) \longrightarrow \infty,$$

with $\mathrm{TV}(\pi)$ denoting the total variation of π. Then, the convex conjugate $\mathcal{B}^*$ is continuous on $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$ for the uniform norm.

Proof. Convex conjugates are lower-semi continuous as a limit of continuous functions $c \mapsto \int c\mathrm{d}\pi - \mathcal{A}(\pi)$. Lower-semi continuous functions on Banach spaces are continuous on the interior of their domain, as proven e.g. in (Aubin, 1979, Section 3.3, proposition 4). Since $(\mathcal{C}(\mathcal{X} \times \mathcal{Y}), \|\cdot\|_\infty)$ is a Banach space, we only need to prove that the domain of $\mathcal{A}^*$ is the whole space $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$ to conclude.

Let $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$. By duality between $(\mathcal{C}(\mathcal{X} \times \mathcal{Y}), \|\cdot\|_\infty)$ and $(\mathcal{M}(\mathcal{X} \times \mathcal{Y}), \mathrm{TV})$, the following inequality holds:

$$\text{For all } \pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y}), \quad \int c\mathrm{d}\pi \leq \|c\|_\infty \mathrm{TV}(\pi).$$

Therefore,

$$\mathcal{A}^*(c) = \sup_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \int c\mathrm{d}\pi - \mathcal{A}(\pi) \leq \sup_{\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})} \mathrm{TV}(\pi) \left(\|c\|_\infty - \frac{\mathcal{A}(\pi)}{\mathrm{TV}(\pi)} \right).$$

By superlinearity, the set of π such that $\mathcal{A}(\pi)/\mathrm{TV}(\pi) \leq \|c\|_\infty$ is bounded: therefore, the supremum can be restricted to a subset of π with bounded total variation, and the previous bound ensures that $\mathcal{A}^*(c) < +\infty$. Therefore, the domain of $\mathcal{A}^*$ is $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$: putting everything together, it demonstrates the continuity of $\mathcal{B}^*$ on $\mathcal{C}(\mathcal{X} \times \mathcal{Y})$ for the uniform norm. □

Proposition 2. *If $\mathcal{R}$ contains an entropic term, k is continuous and α, β have compact support, then there exists two linear operators $A, B : \mathcal{M}_+^{L^2}(\alpha \otimes \beta) \to L^2(\mathcal{X} \times \mathcal{Y}, \alpha \otimes \beta)$ such that:*

$$\mathcal{RQM}(\alpha, \beta) = \inf_{\pi \in \mathcal{M}_+^{L^2}(\alpha \otimes \beta)} \tfrac{1}{2} \|A[\pi]\|^2 - \tfrac{1}{2} \|B[\pi]\|^2 + \mathcal{R}(\pi).$$

with $\mathcal{M}_+^{L^2}(\alpha \otimes \beta)$ denoting the subset of positive measures of $\mathcal{M}^{L^2}(\alpha \otimes \beta)$.

Proof. Let $\mu := \alpha \otimes \beta$. Any π such that $D_{\mathrm{KL}}(\pi \| \mu) < +\infty$ admits a density with respect to μ, i.e. $\pi(x, y) = f_\pi(x, y) \mathrm{d}\mu(x, y)$ with $f_\pi \in L^1(\mathcal{X} \times \mathcal{Y}, \mu)$. The space of squared-integrable functions $L^2(\mathcal{X} \times \mathcal{Y}, \mu)$ is dense in $L^1(\mathcal{X} \times \mathcal{Y}, \mu)$, so the original problem is equivalent to:

$$\mathcal{RQM}(\alpha, \beta) = \inf_{f \in L^2(\mathcal{X} \times \mathcal{Y}, \mu)} \frac{1}{2} \int f(v) \left(\int k(v, w) f(w) \mathrm{d}\mu(w) \right) \mathrm{d}\mu(v) + \mathcal{R}(f \mathrm{d}\mu).$$

Since k is continuous on the compact space $\mathcal{X} \times \mathcal{Y}$, it is bounded and we have the following:

$$\int k(v, w)^2 \mathrm{d}\mu(v) \mu(w) < +\infty.$$

Therefore, the operator $f \in L^2(\mathcal{X} \times \mathcal{Y}, \mu) \mapsto \mathcal{K} f := \int k(\cdot, w) f(w) \mathrm{d}\mu(w)) \in L^2(\mathcal{X} \times \mathcal{Y}, \mu)$ is an auto-adjoint Hilbert-Schmidt integral operator. In particular, it is compact and the spectral theorem applies: there exists an orthonormal basis $(e_i) \in L^2(\mathcal{X} \times \mathcal{Y}, \mu)$ and eigenvalues $(\lambda_i) \in \mathbb{R}$ such that,

$$\text{For all } f \in L^2(\mathcal{X} \times \mathcal{Y}, \mu), v \in \mathcal{X} \times \mathcal{Y}, \quad \mathcal{K} f(v) = \sum_i \lambda_i e_i(v) \int f(w) e_i(w) \mathrm{d}\mu(w).$$

We define $A, B : L^2(\mathcal{X} \times \mathcal{Y}, \mu) \to L^2(\mathcal{X} \times \mathcal{Y}, \mu)$ as:

$$A : f \longmapsto \int \sum_{\lambda_i > 0} |\lambda_i| \, e_i e_i(w) f(w) \mathrm{d}\mu(w),$$
$$B : f \longmapsto \int \sum_{\lambda_i < 0} |\lambda_i| \, e_i \int e_i(w) f(w) \mathrm{d}\mu(w).$$

Then, we have $\mathcal{K} f = A f - B f$ and:

$$\int k(v, w) f(v) f(w) \mathrm{d}\mu(v) \mathrm{d}\mu(w) = \|Af\|_{L^2}^2 - \|B[f]\|_{L^2}^2 .$$

Therefore, after identifying A and B with their action on measures $f \mathrm{d}\mu \in \mathcal{M}_+^{L^2}(\mu)$:

$$\mathcal{RQM}(\alpha, \beta) = \inf_{\pi \in \mathcal{M}_+^{L^2}(\mu)} \tfrac{1}{2} \|A[\pi]\| - \tfrac{1}{2} \|B[\pi]\|^2 + \mathcal{R}(\pi).$$

□

Lemma 2. *Let $k : \mathcal{C}((\mathcal{X} \times \mathcal{Y})^2)$ be symmetric and let A, B be two integral operators. The identity*

$$\frac{1}{2} \int k \, \mathrm{d}\pi \otimes \pi \;=\; \frac{1}{2} \|A[\pi]\|^2 - \frac{1}{2} \|B[\pi]\|^2$$

holds for all $\pi \in \mathcal{M}(\mathcal{X} \times \mathcal{Y})$ if and only if, for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$:

$$k(x, y, x', y') \;=\; \langle A(x, y), A(x', y') \rangle - \langle B(x, y), B(x', y') \rangle .$$

Proof. Given A and B, we define the function:

$$k_{A,B}(x,y,x',y') := \langle A(x,y), A(x',y')\rangle - \langle B(x,y), B(x',y')\rangle\,.$$

Let $k \in \mathcal{C}((\mathcal{X}\times\mathcal{Y})^2)$ be a symmetric function. If $k = k_{A,B}$ pointwise, then by bilinearity of the inner product,

$$\int k\,\mathrm{d}\pi\otimes\pi = \int k_{A,B}\,\mathrm{d}\pi\otimes\pi = \left\|\int A\,\mathrm{d}\pi\right\|^2 - \left\|\int B\,\mathrm{d}\pi\right\|^2.$$

Conversely, assume the integral identity holds for every $\pi \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$. Applying it to $\pi = \delta_{(x,y)}$ gives $k(x,y,x,y) = k_{A,B}(x,y,x,y)$. Finally, applying it to $\pi = \delta_{(x,y)} + \delta_{(x',y')}$ gives:

$$\begin{aligned} k(x,y,x,y) + k(x',y',x',y') + 2k(x,y,x',y') \\ = k_{A,B}(x,y,x,y) + k_{A,B}(x',y',x',y') + 2k_{A,B}(x,y,x',y'), \end{aligned}$$

which implies that $k(x,y,x',y') = k_{A,B}(x,y,x',y')$. □

E Congergence of the dual algorithm

E.1 Notation and setting

Objects. Let $\mathcal{P} := \mathcal{M}_+(\mathcal{X}\times\mathcal{Y})$. Marginal constraints, when present, are encoded in $\mathcal{R}$, so that $\mathrm{dom}\,\mathcal{R}$ is the set of admissible plans. Let $A : \mathcal{M}(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}_A$ and $B : \mathcal{M}(\mathcal{X}\times\mathcal{Y}) \to \mathcal{H}_B$ be the integral operators of eq. (13); they are linear, so that $B[\pi-\pi'] = B[\pi] - B[\pi']$ for $\pi,\pi' \in \mathcal{P}$. Their adjoints are $A^\top\Xi = \langle \Xi, A(\cdot,\cdot)\rangle$ and $B^\top\Gamma = \langle \Gamma, B(\cdot,\cdot)\rangle$; the variables Γ and Ξ range over $\mathcal{H}_B$ and $\mathcal{H}_A$. We use the objective of eq. (1), its linearization at fixed Γ (the inner problem of eq. (14)) and the value of that inner problem:

$$\mathcal{E}_\mathcal{R}(\pi) := \tfrac{1}{2}\|A[\pi]\|^2 - \tfrac{1}{2}\|B[\pi]\|^2 + \mathcal{R}(\pi), \qquad \mathcal{R}\mathrm{QM}(\alpha,\beta) = \inf_{\pi\in\mathcal{P}} \mathcal{E}_\mathcal{R}(\pi), \tag{24}$$

$$Q_\Gamma(\pi) := \tfrac{1}{2}\|A[\pi]\|^2 - \langle \Gamma, B[\pi]\rangle + \mathcal{R}(\pi), \qquad \Lambda(\Gamma) := \tfrac{1}{2}\|\Gamma\|^2 + \inf_{\pi\in\mathcal{P}} Q_\Gamma(\pi). \tag{25}$$

By theorem 3, $\mathcal{R}\mathrm{QM}(\alpha,\beta) = \inf_\Gamma \Lambda(\Gamma)$; Λ is the Lyapunov function of the outer loop. For a cost $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$ we write

$$\mathcal{R}\mathrm{OT}(c) := \inf_{\pi\in\mathcal{P}} \big\{\langle c,\pi\rangle + \mathcal{R}(\pi)\big\} \tag{26}$$

for the value of the $\mathcal{R}\mathrm{OT}$ problem eq. (7), seen as a function of its cost for fixed (α,β). With the effective cost $c_{\Gamma,\Xi} := A^\top\Xi - B^\top\Gamma$ of theorem 3, the reduced objective of section 4 is

$$h_\Gamma(\Xi) := -\tfrac{1}{2}\|\Xi\|^2 + \mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi}) = \sup_{\phi\in\Phi} \mathcal{D}_\mathcal{R}(\phi;\, c_{\Gamma,\Xi},\alpha,\beta) - \tfrac{1}{2}\|\Xi\|^2, \tag{27}$$

the second equality being the dual representation eq. (7).

An elementary identity. For all a, b, Ξ in a Hilbert space,

$$\tfrac{1}{2}\|a\|^2 - \tfrac{1}{2}\|b\|^2 - \langle \Xi, a-b\rangle = \tfrac{1}{2}\|a-\Xi\|^2 - \tfrac{1}{2}\|b-\Xi\|^2. \tag{28}$$

It follows by expanding the squares and is used repeatedly below.

Weak duality. Since $\langle \Xi, a\rangle - \frac{1}{2}\|\Xi\|^2 \le \frac{1}{2}\|a\|^2$ for all $a, \Xi \in \mathcal{H}_A$, for every Γ and Ξ

$$h_\Gamma(\Xi) = \inf_{\pi\in\mathcal{P}} \big\{\langle \Xi, A[\pi]\rangle - \tfrac{1}{2}\|\Xi\|^2 - \langle \Gamma, B[\pi]\rangle + \mathcal{R}(\pi)\big\} \le \inf_{\pi\in\mathcal{P}} Q_\Gamma(\pi),$$

hence

$$\sup_{\Xi\in\mathcal{H}_A} h_\Gamma(\Xi) \le \inf_{\pi\in\mathcal{P}} Q_\Gamma(\pi). \tag{29}$$

Equality is proved in proposition 6 when h_Γ is differentiable, and is not needed elsewhere.

Standing assumptions. Throughout the appendix we assume:

(S1) $\mathcal{R}$ is proper, convex and lower semicontinuous, and $\mathcal{R}(\pi_0) < \infty$ for some $\pi_0 \in \mathcal{P}$;

(S2) $\mathcal{R}\mathrm{QM}(\alpha, \beta) > -\infty$.

Under **(S1)–(S2)**, Λ is finite everywhere (lemma 5), and we set $\Delta^{\mathrm{out}} := \Lambda(\Gamma_0) - \mathcal{R}\mathrm{QM}(\alpha, \beta) \in [0, \infty)$ for the initial point Γ_0 of the algorithm.

$\mathcal{R}$OT oracle and certificate. A plan $\widehat{\pi} \in \mathcal{P}$ is $\delta_{\mathcal{R}\mathrm{OT}}$-*accurate* at cost c if

$$\langle c, \widehat{\pi} \rangle + \mathcal{R}(\widehat{\pi}) \le \mathcal{R}\mathrm{OT}(c) + \delta_{\mathcal{R}\mathrm{OT}}. \tag{30}$$

This is a primal notion: with a dual solver, $\widehat{\pi}$ is the plan recovered from the current potentials and $\delta_{\mathcal{R}\mathrm{OT}}$ is its primal–dual gap, which upper-bounds the left-hand side minus $\mathcal{R}\mathrm{OT}(c)$ by weak duality. It requires $\mathcal{R}(\widehat{\pi}) < \infty$; for KL or TV marginal penalties the recovered plan qualifies, while for a hard marginal constraint it must first be rounded onto the constraint set, e.g. by Altschuler et al. (2017, Alg. 2), the moments $A[\widehat{\pi}]$, $B[\widehat{\pi}]$ being then those of the rounded plan. For a pair $(\Xi, \widehat{\pi})$ with $\widehat{\pi}$ $\delta_{\mathcal{R}\mathrm{OT}}$-accurate at $c_{\Gamma,\Xi}$ we define the computable certificate

$$\mathrm{Cert}_\Gamma(\Xi, \widehat{\pi}) := \tfrac{1}{2} \left\| A[\widehat{\pi}] - \Xi \right\|^2 + \delta_{\mathcal{R}\mathrm{OT}}, \tag{31}$$

which is the quantity tested by the inner stopping rule of algorithm 1; lemma 4 shows that it bounds $Q_\Gamma(\widehat{\pi}) - \inf_\pi Q_\Gamma$.

Indices for algorithm 1. We analyze algorithm 1 with: constant accuracy $\delta_{\mathcal{R}\mathrm{OT}}$, constant step τ, certificate-based inner stopping rule. At outer iteration t, write $\Xi_{t,k}$ for the value of Ξ before the $(k+1)$-th $\mathcal{R}$OT solve (so $\Xi_{t,0} = \Xi_{t-1,K_{t-1}}$ is carried over from the previous outer iteration), $\widehat{\pi}_{t,k}$ for the $\delta_{\mathcal{R}\mathrm{OT}}$-accurate plan returned at cost $c_{\Gamma_t,\Xi_{t,k}}$, and

$$K_t := \min\left\{ k \ge 0 : \ \mathrm{Cert}_{\Gamma_t}(\Xi_{t,k}, \widehat{\pi}_{t,k}) \le \delta^2/4 \right\} \in \mathbb{N} \cup \{+\infty\}. \tag{32}$$

The inner loop of iteration t thus makes $K_t + 1$ $\mathcal{R}$OT solves and K_t updates $\Xi_{t,k+1} = \Xi_{t,k} + \tau(A[\widehat{\pi}_{t,k}] - \Xi_{t,k})$; it terminates if and only if $K_t < \infty$, in which case $\widehat{\pi}_t := \widehat{\pi}_{t,K_t}$, $\Gamma_{t+1} := B[\widehat{\pi}_t]$, and the algorithm returns $\widehat{\pi}_t$ when $\|\Gamma_{t+1} - \Gamma_t\| \le \delta$. When $\Gamma = \Gamma_t$ is fixed we drop the index t and write Ξ_k, $\widehat{\pi}_k$. For the entropic regularizers of table 3, $A[\widehat{\pi}] - \Xi$ is the gradient in Ξ of $\mathcal{D}_\mathcal{R}(\phi; c_{\Gamma,\Xi}, \alpha, \beta) - \frac{1}{2}\|\Xi\|^2$ at the potentials ϕ encoding $\widehat{\pi}$, which is why the Ξ-update is an (inexact) gradient step.

E.2 The fixed-Γ certificate: proof of Lemma 4

The following lemma justifies the inner stopping rule of algorithm 1.

Lemma 4 (Inexact supergradient and primal certificate)**.** *Fix $(\Gamma, \Xi) \in \mathcal{H}_B \times \mathcal{H}_A$, let $\widehat{\pi}$ be $\delta_{\mathcal{R}\mathrm{OT}}$-accurate at $c_{\Gamma,\Xi}$, and set $r := A[\widehat{\pi}] - \Xi$. Then, for every $\Xi' \in \mathcal{H}_A$,*

$$h_\Gamma(\Xi') \le h_\Gamma(\Xi) + \langle A[\widehat{\pi}] - \Xi, \Xi' - \Xi \rangle - \tfrac{1}{2}\|\Xi' - \Xi\|^2 + \delta_{\mathcal{R}\mathrm{OT}}, \tag{33}$$

and

$$Q_\Gamma(\widehat{\pi}) - \inf_{\pi \in \mathcal{P}} Q_\Gamma(\pi) \le \tfrac{1}{2}\|A[\widehat{\pi}] - \Xi\|^2 + \delta_{\mathcal{R}\mathrm{OT}} = \mathrm{Cert}_\Gamma(\Xi, \widehat{\pi}). \tag{34}$$

Proof. Let $\Xi' \in \mathcal{H}_A$. Since $c_{\Gamma,\Xi'} = c_{\Gamma,\Xi} + A^\top(\Xi' - \Xi)$, testing the infimum defining $\mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi'})$ at $\widehat{\pi}$ and using eq. (30),

$$\mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi'}) \le \langle c_{\Gamma,\Xi}, \widehat{\pi} \rangle + \mathcal{R}(\widehat{\pi}) + \langle A[\widehat{\pi}], \Xi' - \Xi \rangle \le \mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi}) + \langle A[\widehat{\pi}], \Xi' - \Xi \rangle + \delta_{\mathcal{R}\mathrm{OT}}.$$

Subtracting $\frac{1}{2}\|\Xi'\|^2 = \frac{1}{2}\|\Xi\|^2 + \langle \Xi, \Xi' - \Xi \rangle + \frac{1}{2}\|\Xi' - \Xi\|^2$ from both sides gives eq. (33).

For eq. (34), let $\pi \in \mathcal{P}$. The accuracy of $\widehat{\pi}$ tested against π reads

$$-\langle \Gamma, B[\widehat{\pi}] - B[\pi] \rangle + \mathcal{R}(\widehat{\pi}) - \mathcal{R}(\pi) \le -\langle \Xi, A[\widehat{\pi}] - A[\pi] \rangle + \delta_{\mathcal{R}\mathrm{OT}},$$

so that, by eq. (25) and then eq. (28) with $(a, b) = (A[\widehat{\pi}], A[\pi])$,

$$\begin{aligned} Q_\Gamma(\widehat{\pi}) - Q_\Gamma(\pi) &\le \tfrac{1}{2}\|A[\widehat{\pi}]\|^2 - \tfrac{1}{2}\|A[\pi]\|^2 - \langle \Xi, A[\widehat{\pi}] - A[\pi] \rangle + \delta_{\mathcal{R}\mathrm{OT}} \\ &= \tfrac{1}{2}\|A[\widehat{\pi}] - \Xi\|^2 - \tfrac{1}{2}\|A[\pi] - \Xi\|^2 + \delta_{\mathcal{R}\mathrm{OT}} \le \tfrac{1}{2}\|A[\widehat{\pi}] - \Xi\|^2 + \delta_{\mathcal{R}\mathrm{OT}}. \end{aligned}$$

Taking the infimum over π proves eq. (34). □

E.3 Outer convergence: proof of Theorem 5(i)

Lemma 5 (Square completion and outer Lyapunov function)**.** *For every* $(\Gamma, \pi) \in \mathcal{H}_B \times \mathcal{P}$*,*

$$\tfrac{1}{2}\|\Gamma\|^2 + Q_\Gamma(\pi) = \mathcal{E}_\mathcal{R}(\pi) + \tfrac{1}{2}\|\Gamma - B[\pi]\|^2. \tag{35}$$

Consequently, the function Λ *of eq.* (25) *satisfies*

$$\Lambda(\Gamma) = \inf_{\pi\in\mathcal{P}} \Big\{\mathcal{E}_\mathcal{R}(\pi) + \tfrac{1}{2}\|\Gamma - B[\pi]\|^2\Big\},$$

is finite under ***(S1)–(S2)****, and*

$$\Lambda(\Gamma) \geq \mathcal{R}\mathrm{QM}(\alpha,\beta) \quad \textit{for every } \Gamma \in \mathcal{H}_B, \qquad \inf_\Gamma \Lambda(\Gamma) = \mathcal{R}\mathrm{QM}(\alpha,\beta). \tag{36}$$

Proof. Expanding $\frac{1}{2}\|\Gamma - B[\pi]\|^2 = \frac{1}{2}\|\Gamma\|^2 - \langle\Gamma, B[\pi]\rangle + \frac{1}{2}\|B[\pi]\|^2$ and adding $\mathcal{E}_\mathcal{R}(\pi)$, the terms $\pm\frac{1}{2}\|B[\pi]\|^2$ cancel and give eq. (35); taking the infimum over π gives the expression of Λ.

Since the square is nonnegative, $\Lambda(\Gamma) \geq \inf_\pi \mathcal{E}_\mathcal{R}(\pi) = \mathcal{R}\mathrm{QM}(\alpha,\beta) > -\infty$ by **(S2)**, and $\Lambda(\Gamma) \leq \mathcal{E}_\mathcal{R}(\pi_0) + \frac{1}{2}\|\Gamma - B[\pi_0]\|^2 < \infty$ for the plan π_0 of **(S1)**, whose moments are finite since A and B are bounded. Finally, choosing $\Gamma = B[\pi]$ kills the square, so $\Lambda(B[\pi]) \leq \mathcal{E}_\mathcal{R}(\pi)$ for every π; together with the lower bound, this gives eq. (36). □

The identity eq. (35) shows that $\frac{1}{2}\|\Gamma\|^2 + Q_\Gamma$ majorizes $\mathcal{E}_\mathcal{R}$ with equality at $\Gamma = B[\pi]$, so that the Γ-update of algorithm 1 is a majorization–minimization (equivalently, DC) step. Its fixed points are the plans $\pi^\star$ with $\pi^\star \in \arg\min_\pi Q_{B[\pi^\star]}(\pi)$, the DC-critical plans of section 4; the algorithm produces them approximately, in the following sense.

Definition 5 ((δ,η)-criticality)**.** *We call* $\widehat{\pi} \in \mathcal{P}$ (δ,η)-critical*, with* $\delta, \eta \geq 0$*, if for every* $\pi \in \mathcal{P}$*,*

$$\mathcal{E}_\mathcal{R}(\widehat{\pi}) \leq \mathcal{E}_\mathcal{R}(\pi) + \tfrac{1}{2}\|B[\pi - \widehat{\pi}]\|^2 + \delta\|B[\pi - \widehat{\pi}]\| + \eta. \tag{37}$$

By eq. (35), $\widehat{\pi}$ is (δ,η)-critical if and only if it is an η-minimizer of the convex function $\pi \mapsto Q_{B[\widehat{\pi}]}(\pi) + \delta\|B[\pi - \widehat{\pi}]\|$; the case $(\delta,\eta) = (0,0)$ is DC criticality.

Lemma 6 (One outer step)**.** *Let* $\Gamma \in \mathcal{H}_B$*,* $\eta \geq 0$*, and let* $\widehat{\pi} \in \mathcal{P}$ *be an* η*-minimizer of* Q_Γ*, that is,*

$$Q_\Gamma(\widehat{\pi}) \leq \inf_{\pi\in\mathcal{P}} Q_\Gamma(\pi) + \eta. \tag{38}$$

Set $\Gamma^+ := B[\widehat{\pi}]$*. Then*

(i) $\Lambda(\Gamma) - \Lambda(\Gamma^+) \geq \frac{1}{2}\|\Gamma^+ - \Gamma\|^2 - \eta$*;*

(ii) $\widehat{\pi}$ *is* $\big(\|\Gamma^+ - \Gamma\|, \eta\big)$*-critical.*

Proof. We first prove (i). By eq. (38),

$$\Lambda(\Gamma) = \tfrac{1}{2}\|\Gamma\|^2 + \inf_\pi Q_\Gamma(\pi) \geq \tfrac{1}{2}\|\Gamma\|^2 + Q_\Gamma(\widehat{\pi}) - \eta. \tag{39}$$

Applying the square identity eq. (35) at $(\Gamma, \widehat{\pi})$, and using $B[\widehat{\pi}] = \Gamma^+$,

$$\tfrac{1}{2}\|\Gamma\|^2 + Q_\Gamma(\widehat{\pi}) = \mathcal{E}_\mathcal{R}(\widehat{\pi}) + \tfrac{1}{2}\big\|\Gamma - \Gamma^+\big\|^2,$$

so that eq. (39) becomes

$$\Lambda(\Gamma) \geq \mathcal{E}_\mathcal{R}(\widehat{\pi}) + \tfrac{1}{2}\big\|\Gamma - \Gamma^+\big\|^2 - \eta. \tag{40}$$

On the other hand, evaluating the infimum defining $\Lambda(\Gamma^+)$ at $\widehat{\pi}$ and applying eq. (35) at $(\Gamma^+, \widehat{\pi})$, where the square vanishes because $\Gamma^+ = B[\widehat{\pi}]$,

$$\Lambda(\Gamma^+) \leq \tfrac{1}{2}\big\|\Gamma^+\big\|^2 + Q_{\Gamma^+}(\widehat{\pi}) = \mathcal{E}_\mathcal{R}(\widehat{\pi}).$$

Subtracting this inequality from eq. (40) proves (i).

We now prove (ii). Fix an arbitrary $\pi \in \mathcal{P}$. By eq. (38), $Q_\Gamma(\widehat{\pi}) \le Q_\Gamma(\pi) + \eta$. Applying eq. (35) at $(\Gamma, \widehat{\pi})$ and at (Γ, π), this inequality reads

$$\mathcal{E}_{\mathcal{R}}(\widehat{\pi}) + \tfrac{1}{2}\left\|\Gamma - \Gamma^+\right\|^2 \le \mathcal{E}_{\mathcal{R}}(\pi) + \tfrac{1}{2}\left\|\Gamma - B[\pi]\right\|^2 + \eta. \tag{41}$$

Set $d := B[\pi - \widehat{\pi}] = B[\pi] - \Gamma^+$. Then $\Gamma - B[\pi] = (\Gamma - \Gamma^+) - d$, and expanding the square,

$$\tfrac{1}{2}\left\|\Gamma - B[\pi]\right\|^2 = \tfrac{1}{2}\left\|\Gamma - \Gamma^+\right\|^2 - \left\langle \Gamma - \Gamma^+, d\right\rangle + \tfrac{1}{2}\left\|d\right\|^2.$$

Substituting this into eq. (41), the term $\frac{1}{2}\left\|\Gamma - \Gamma^+\right\|^2$ cancels on both sides, and we obtain

$$\begin{aligned}\mathcal{E}_{\mathcal{R}}(\widehat{\pi}) &\le \mathcal{E}_{\mathcal{R}}(\pi) + \tfrac{1}{2}\left\|d\right\|^2 + \left\langle \Gamma^+ - \Gamma, d\right\rangle + \eta \\ &\le \mathcal{E}_{\mathcal{R}}(\pi) + \tfrac{1}{2}\left\|B[\pi - \widehat{\pi}]\right\|^2 + \left\|\Gamma^+ - \Gamma\right\|\ \left\|B[\pi - \widehat{\pi}]\right\| + \eta,\end{aligned}$$

where the last inequality is Cauchy–Schwarz. This is eq. (37) with (δ, η) replaced by $(\left\|\Gamma^+ - \Gamma\right\|, \eta)$, that is, $\widehat{\pi}$ is $(\left\|\Gamma^+ - \Gamma\right\|, \eta)$-critical. □

Theorem 12 (Outer convergence of algorithm 1; precise form of theorem 5(i))**.** *Assume **(S1)–(S2)** and let $\delta > 0$. Run algorithm 1 with any $\mathcal{R}\mathrm{OT}$ solver, any step size τ and any accuracy $\delta_{\mathcal{R}\mathrm{OT}}$ such that every inner loop terminates ($K_t < \infty$ for all t). Let*

$$T_\delta := \max\Big\{1, \Big\lceil \frac{4\Delta^{\mathrm{out}}}{\delta^2} \Big\rceil\Big\}. \tag{42}$$

Then:

(i) *for every outer iteration t, $Q_{\Gamma_t}(\widehat{\pi}_t) \le \inf_\pi Q_{\Gamma_t}(\pi) + \delta^2/4$,*

$$\Lambda(\Gamma_t) - \Lambda(\Gamma_{t+1}) \ge \tfrac{1}{2}\left\|\Gamma_{t+1} - \Gamma_t\right\|^2 - \tfrac{\delta^2}{4}, \tag{43}$$

and $\widehat{\pi}_t$ is $\big(\left\|\Gamma_{t+1} - \Gamma_t\right\|, \delta^2/4\big)$-critical;

(ii) *the stopping test $\left\|\Gamma_{t+1} - \Gamma_t\right\| \le \delta$ is met at some outer iteration $t_{\mathrm{stop}} < T_\delta$, and the returned plan $\widehat{\pi}_{t_{\mathrm{stop}}}$ is $(\delta, \delta^2/4)$-critical.*

Proof. Fix an outer iteration t. By definition eq. (32) of K_t, the plan $\widehat{\pi}_t = \widehat{\pi}_{t,K_t}$ is $\delta_{\mathcal{R}\mathrm{OT}}$-accurate at $c_{\Gamma_t, \Xi_{t,K_t}}$ and $\mathrm{Cert}_{\Gamma_t}(\Xi_{t,K_t}, \widehat{\pi}_t) \le \delta^2/4$. Hence lemma 4 gives $Q_{\Gamma_t}(\widehat{\pi}_t) \le \inf_\pi Q_{\Gamma_t}(\pi) + \delta^2/4$, which is the first claim of (i), and lemma 6 applied with $(\Gamma, \widehat{\pi}, \eta) = (\Gamma_t, \widehat{\pi}_t, \delta^2/4)$, for which $\Gamma^+ = B[\widehat{\pi}_t] = \Gamma_{t+1}$, gives the other two.

For (ii), suppose that the stopping test fails for every $t < T_\delta$. Summing eq. (43) over $t = 0, \dots, T_\delta - 1$ and using $\Lambda(\Gamma_{T_\delta}) \ge \mathcal{R}\mathrm{QM}(\alpha, \beta)$ from eq. (36),

$$\tfrac{1}{2}\sum_{t=0}^{T_\delta - 1}\left\|\Gamma_{t+1} - \Gamma_t\right\|^2 \le \Lambda(\Gamma_0) - \Lambda(\Gamma_{T_\delta}) + T_\delta\frac{\delta^2}{4} \le \Delta^{\mathrm{out}} + T_\delta\frac{\delta^2}{4}.$$

Therefore

$$\min_{0 \le t < T_\delta}\left\|\Gamma_{t+1} - \Gamma_t\right\|^2 \le \frac{1}{T_\delta}\sum_{t=0}^{T_\delta - 1}\left\|\Gamma_{t+1} - \Gamma_t\right\|^2 \le \frac{2\Delta^{\mathrm{out}}}{T_\delta} + \frac{\delta^2}{2} \le \delta^2,$$

because $T_\delta \ge 4\Delta^{\mathrm{out}}/\delta^2$. This contradicts the assumption that $\left\|\Gamma_{t+1} - \Gamma_t\right\| > \delta$ for all $t < T_\delta$, so the stopping test is met at some $t_{\mathrm{stop}} < T_\delta$. At that iteration, $\left\|\Gamma_{t_{\mathrm{stop}}+1} - \Gamma_{t_{\mathrm{stop}}}\right\| \le \delta$ and (i) show that $\widehat{\pi}_{t_{\mathrm{stop}}}$ is $(\delta', \delta^2/4)$-critical for some $\delta' \le \delta$, hence $(\delta, \delta^2/4)$-critical, because eq. (37) is monotone in both parameters. □

Theorem 12 is conditional only on the termination of the inner loops, which is observed at run time; corollary 1 gives a smoothness condition under which termination is guaranteed, with an explicit bound on K_t.

E.4 Inner-loop complexity: proof of Theorem 5(ii)

At fixed Γ, the inner loop of algorithm 1 is an inexact gradient ascent on h_Γ, with inexact gradient $\widehat{g}_k := A[\widehat{\pi}_k] - \Xi_k$ (lemma 4).

Lemma 7 (Strong concavity of the reduced objective)**.** *Assume **(S1)** and fix Γ. Then $h_\Gamma < +\infty$ on $\mathcal{H}_A$. If moreover $\mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi}) > -\infty$ for every $\Xi \in \mathcal{H}_A$, so that h_Γ is finite, then h_Γ is 1-strongly concave and upper semicontinuous on $\mathcal{H}_A$, and attains its maximum at a unique point $\Xi^\star_\Gamma$.*

Proof. The bound $h_\Gamma < +\infty$ is eq. (29) tested at the plan π_0 of **(S1)**. Assume h_Γ finite. For fixed π, $\Xi \mapsto \langle \Xi, A[\pi]\rangle - \langle \Gamma, B[\pi]\rangle + \mathcal{R}(\pi)$ is affine and continuous, so $\Xi \mapsto \mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi})$, an infimum of such functions, is concave and upper semicontinuous, and $h_\Gamma = \mathcal{R}\mathrm{OT}(c_{\Gamma,\cdot}) - \frac{1}{2}\|\cdot\|^2$ is 1-strongly concave and upper semicontinuous. A finite, upper semicontinuous, strongly concave function on a Hilbert space attains its supremum at exactly one point. □

Proposition 6 (Gradient of h_Γ and exact fixed point)**.** *Fix Γ and assume that h_Γ is finite on $\mathcal{H}_A$ and differentiable at $\Xi \in \mathcal{H}_A$. If $\pi_{\Gamma,\Xi}$ is any minimizer of $\pi \mapsto \langle c_{\Gamma,\Xi}, \pi\rangle + \mathcal{R}(\pi)$ over $\mathcal{P}$, then*

$$\nabla h_\Gamma(\Xi) = A[\pi_{\Gamma,\Xi}] - \Xi. \tag{44}$$

In particular, if h_Γ is differentiable on $\mathcal{H}_A$ and the $\mathcal{R}\mathrm{OT}$ problem at $c_{\Gamma,\Xi^\star_\Gamma}$ admits a minimizer $\pi^\star_\Gamma$, where $\Xi^\star_\Gamma$ is the maximizer of lemma 7, then

$$\Xi^\star_\Gamma = A[\pi^\star_\Gamma], \qquad \pi^\star_\Gamma \in \operatorname*{argmin}_{\pi\in\mathcal{P}} Q_\Gamma(\pi), \qquad \max_{\Xi\in\mathcal{H}_A} h_\Gamma(\Xi) = \min_{\pi\in\mathcal{P}} Q_\Gamma(\pi). \tag{45}$$

Proof. Write $v(\Xi) := \mathcal{R}\mathrm{OT}(c_{\Gamma,\Xi})$, so that $h_\Gamma = v - \frac{1}{2}\|\cdot\|^2$ and v is finite on $\mathcal{H}_A$.

A supergradient of v. Let $\pi := \pi_{\Gamma,\Xi}$ and let $\Xi' \in \mathcal{H}_A$. Since π is admissible in the infimum defining $v(\Xi')$, and since $c_{\Gamma,\Xi'} = c_{\Gamma,\Xi} + A^\top(\Xi' - \Xi)$,

$$v(\Xi') \le \langle c_{\Gamma,\Xi'}, \pi\rangle + \mathcal{R}(\pi) = \langle c_{\Gamma,\Xi}, \pi\rangle + \mathcal{R}(\pi) + \langle A[\pi], \Xi' - \Xi\rangle = v(\Xi) + \langle A[\pi], \Xi' - \Xi\rangle\,, \tag{46}$$

where the last equality is the optimality of π at the cost $c_{\Gamma,\Xi}$. Thus $A[\pi]$ is a supergradient of the concave function v at Ξ.

Identification of the gradient. As v is concave and differentiable at Ξ, its only supergradient there is $\nabla v(\Xi)$; hence $\nabla v(\Xi) = A[\pi]$ and $\nabla h_\Gamma(\Xi) = A[\pi] - \Xi$.

Exact fixed point. If h_Γ is differentiable on $\mathcal{H}_A$, then $\nabla h_\Gamma(\Xi^\star_\Gamma) = 0$ at its maximizer, so eq. (44) gives $\Xi^\star_\Gamma = A[\pi^\star_\Gamma]$. The plan $\pi^\star_\Gamma$ is 0-accurate at $c_{\Gamma,\Xi^\star_\Gamma}$ with $r = A[\pi^\star_\Gamma] - \Xi^\star_\Gamma = 0$, so eq. (34) yields $\pi^\star_\Gamma \in \arg\min_\pi Q_\Gamma(\pi)$. Finally, by eq. (27), the optimality of $\pi^\star_\Gamma$ and $\Xi^\star_\Gamma = A[\pi^\star_\Gamma]$,

$$h_\Gamma(\Xi^\star_\Gamma) = -\tfrac{1}{2}\|\Xi^\star_\Gamma\|^2 + \langle \Xi^\star_\Gamma, A[\pi^\star_\Gamma]\rangle - \langle \Gamma, B[\pi^\star_\Gamma]\rangle + \mathcal{R}(\pi^\star_\Gamma) = Q_\Gamma(\pi^\star_\Gamma),$$

so eq. (29) is an equality, attained on both sides. □

Differentiability of h_Γ is part of the smoothness assumption **($\mathbf{H}_L$)** below, and holds for the entropic regularizers on finite spaces, where the $\mathcal{R}\mathrm{OT}$ minimizer is moreover unique and C^1 in Ξ (proposition 9).

Smoothness assumption. We now assume that the reduced objective is smooth, in one of two forms.

- **($\mathbf{H}_L$)** h_Γ is finite on $\mathcal{H}_A$, and for some $L \ge 1$ either (*global form*) h_Γ is L-smooth on $\mathcal{H}_A$, or (*local form*) there is $R > 0$ such that $\|\Xi^\star_\Gamma\| \le R$, the inner iterates satisfy $\|\Xi_k\| \le R$ for all k, $\delta_{\mathcal{R}\mathrm{OT}} \le \frac{1}{2}$, and h_Γ is twice differentiable with $-LI \preceq \nabla^2 h_\Gamma \preceq -I$ on the ball $\mathcal{B}_\Gamma := \{\|\Xi\| \le 3R + 1\}$.

The local form is the one met in practice (proposition 9). The proofs below use L-smoothness only through the descent lemma

$$\left|h_\Gamma(\Xi') - h_\Gamma(\Xi) - \langle \nabla h_\Gamma(\Xi), \Xi' - \Xi\rangle\right| \le \tfrac{L}{2}\|\Xi' - \Xi\|^2\,, \tag{47}$$

which holds whenever ∇h_Γ is L-Lipschitz along the segment $[\Xi, \Xi']$, and they apply it only along four families of segments:

(a) $[\Xi_k, \Xi_k + d]$ with $\|d\| \le \sqrt{2\delta_{\mathcal{ROT}}/L} \le 1$ (lemma 9);

(b) $[\Xi_k, \Xi_{k+1}]$ and (c) $[\Xi_k, \Xi_\Gamma^\star]$ (proposition 7);

(d) $[\Xi_k, \Xi_k + \frac{1}{L}\nabla h_\Gamma(\Xi_k)]$ (proposition 7).

In the local form, both endpoints of (b) and (c) lie in $\{\|\Xi\| \le R\}$ and both endpoints of (a) lie in $\{\|\Xi\| \le R+1\}$; since balls are convex, these segments lie in $\mathcal{B}_\Gamma$, where $\nabla^2 h_\Gamma \succeq -LI$, so eq. (47) holds along them. The following lemma handles (d).

Lemma 8 (The exact gradient step stays in the ball)**.** *In the local form of* ***($\mathbf{H}_L$)****, for every inner iterate Ξ_k,*

$$\left\|\Xi_k + \tfrac{1}{L}\nabla h_\Gamma(\Xi_k) - \Xi_\Gamma^\star\right\| \le \|\Xi_k - \Xi_\Gamma^\star\| \le 2R,$$

so that the segment $[\Xi_k, \Xi_k + \frac{1}{L}\nabla h_\Gamma(\Xi_k)]$ lies in $\{\|\Xi\| \le 3R\} \subset \mathcal{B}_\Gamma$.

Proof. Set $d := \Xi_k - \Xi_\Gamma^\star$, so that $\|d\| \le 2R$ and the segment $[\Xi_\Gamma^\star, \Xi_k]$ lies in $\{\|\Xi\| \le R\}$, where $-LI \preceq \nabla^2 h_\Gamma \preceq -I$. Since $\nabla h_\Gamma(\Xi_\Gamma^\star) = 0$ (h_Γ is differentiable and maximal there),

$$\nabla h_\Gamma(\Xi_k) = -Hd, \qquad H := -\int_0^1 \nabla^2 h_\Gamma(\Xi_\Gamma^\star + ud)\,\mathrm{d}u,$$

and H is symmetric with $I \preceq H \preceq LI$, hence $0 \preceq I - \frac{1}{L}H \preceq (1 - \frac{1}{L})I$. Therefore

$$\left\|\Xi_k + \tfrac{1}{L}\nabla h_\Gamma(\Xi_k) - \Xi_\Gamma^\star\right\| = \left\|(I - \tfrac{1}{L}H)\,d\right\| \le \|d\| \le 2R,$$

so the exact gradient step lies in $\{\|\Xi\| \le 3R\}$, and so does the segment joining it to Ξ_k, by convexity. □

Hence every argument below is valid in both forms of $(\mathbf{H}_L)$, and we do not distinguish them.

Lemma 9 (OT accuracy controls the gradient error)**.** *Assume* ***($\mathbf{H}_L$)*** *at Γ. Let $\widehat{\pi}_k$ be a $\delta_{\mathcal{ROT}}$-accurate plan at c_{Γ,Ξ_k} and define $\widehat{g}_k := A[\widehat{\pi}_k] - \Xi_k$. Then*

$$\|\widehat{g}_k - \nabla h_\Gamma(\Xi_k)\| \le \sqrt{2L\delta_{\mathcal{ROT}}}. \tag{48}$$

Proof. Set $e_k := \widehat{g}_k - \nabla h_\Gamma(\Xi_k)$. Combining eq. (33) (without its quadratic term) with the descent lemma eq. (47) along $[\Xi_k, \Xi_k + d]$ gives, for every d with $\|d\| \le 1$ (family (a)),

$$-\langle e_k, d\rangle \le \tfrac{L}{2}\|d\|^2 + \delta_{\mathcal{ROT}}.$$

If $e_k \neq 0$, take $d = -s\,e_k/\|e_k\|$ with $s \in (0,1]$: then $s\|e_k\| \le \frac{L}{2}s^2 + \delta_{\mathcal{ROT}}$. For $\delta_{\mathcal{ROT}} > 0$ the choice $s = \sqrt{2\delta_{\mathcal{ROT}}/L} \le 1$ gives $s\|e_k\| \le 2\delta_{\mathcal{ROT}}$, i.e. $\|e_k\| \le \sqrt{2L\delta_{\mathcal{ROT}}}$; for $\delta_{\mathcal{ROT}} = 0$, letting $s \downarrow 0$ gives $e_k = 0$. □

Proposition 7 (Inner loop: geometric convergence and certification)**.** *Assume* ***(S1)–(S2)*** *and* ***($\mathbf{H}_L$)*** *at Γ, and run the inner loop of algorithm 1 at Γ with step $\tau = 1/L$ and accuracy $\delta_{\mathcal{ROT}}$. Let $\widehat{g}_k := A[\widehat{\pi}_k] - \Xi_k$, $\Delta_k := h_\Gamma(\Xi_\Gamma^\star) - h_\Gamma(\Xi_k)$ and $q := 1 - 1/L$. Then, for every $k \ge 0$:*

(i) $\Delta_{k+1} \le q\,\Delta_k + \delta_{\mathcal{ROT}}$*, hence*

$$\Delta_k \le q^k \Delta_0 + L\,\delta_{\mathcal{ROT}}; \tag{49}$$

(ii) *the certificate satisfies*

$$\mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) = \tfrac{1}{2}\|\widehat{g}_k\|^2 + \delta_{\mathcal{ROT}} \le 2L\Delta_k + (2L+1)\delta_{\mathcal{ROT}}; \tag{50}$$

(iii) *for $\eta \in (0,1]$, if $\delta_{\mathcal{ROT}} = \eta/(2C_L)$ with $C_L := 2L^2 + 2L + 1$, then $\mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) \le \eta$ for every*

$$k \ge N_\Xi(\eta) := \left\lceil L \log \max\left\{1, \frac{4L\Delta_0}{\eta}\right\}\right\rceil, \tag{51}$$

so that the inner loop stopped at level η makes at most $N_\Xi(\eta) + 1$ $\mathcal{ROT}$ solves. The same holds with Δ_0 replaced by any upper bound, and $\Delta_0 \le \frac{L}{2}\|\Xi_0 - \Xi_\Gamma^\star\|^2$ ($\le 2LR^2$ in the local form).

Proof. Throughout, let $\psi := -h_\Gamma$. By lemma 7, ψ is 1-strongly convex on $\mathcal{H}_A$, and by **($\mathbf{H}_L$)** the descent lemma eq. (47) holds along the segments (a)–(d) listed after **($\mathbf{H}_L$)**. We write

$$g_k := \nabla h_\Gamma(\Xi_k) = -\nabla\psi(\Xi_k) \qquad \text{and} \qquad e_k := \widehat{g}_k - g_k$$

for the exact gradient and the gradient error, so that the Ξ-update of algorithm 1 with $\tau = 1/L$ reads

$$\Xi_{k+1} = \Xi_k + \tfrac{1}{L}\,\widehat{g}_k = \Xi_k + \tfrac{1}{L}\,(g_k + e_k), \tag{52}$$

and lemma 9 gives

$$\|e_k\|^2 \le 2L\,\delta_{\mathcal{R}\mathrm{OT}}. \tag{53}$$

One inexact gradient step. By eq. (52), $\Xi_{k+1} - \Xi_k = \frac{1}{L}(g_k + e_k)$. The descent lemma eq. (47) for ψ along the segment $[\Xi_k, \Xi_{k+1}]$ (family (b)) gives

$$\begin{aligned}\psi(\Xi_{k+1}) &\le \psi(\Xi_k) + \langle \nabla\psi(\Xi_k), \Xi_{k+1} - \Xi_k\rangle + \tfrac{L}{2}\,\|\Xi_{k+1} - \Xi_k\|^2 \\ &= \psi(\Xi_k) - \tfrac{1}{L}\,\langle g_k, g_k + e_k\rangle + \tfrac{1}{2L}\,\|g_k + e_k\|^2 .\end{aligned}$$

We expand the last two terms:

$$\begin{aligned}-\tfrac{1}{L}\,\langle g_k, g_k + e_k\rangle &= -\tfrac{1}{L}\,\|g_k\|^2 - \tfrac{1}{L}\,\langle g_k, e_k\rangle\,, \\ \tfrac{1}{2L}\,\|g_k + e_k\|^2 &= \tfrac{1}{2L}\,\|g_k\|^2 + \tfrac{1}{L}\,\langle g_k, e_k\rangle + \tfrac{1}{2L}\,\|e_k\|^2 .\end{aligned}$$

The two cross terms $\mp\frac{1}{L}\,\langle g_k, e_k\rangle$ cancel, and we are left with

$$\psi(\Xi_{k+1}) \le \psi(\Xi_k) - \tfrac{1}{2L}\,\|g_k\|^2 + \tfrac{1}{2L}\,\|e_k\|^2 . \tag{54}$$

Hence, the exact part of the step decreases ψ by $\frac{1}{2L}\,\|g_k\|^2$, and the error can cost at most $\frac{1}{2L}\,\|e_k\|^2$.

The gradient dominates the gap. By 1-strong convexity of ψ, for every $\Xi' \in \mathcal{H}_A$,

$$\psi(\Xi') \ge \psi(\Xi_k) + \langle \nabla\psi(\Xi_k), \Xi' - \Xi_k\rangle + \tfrac{1}{2}\,\|\Xi' - \Xi_k\|^2 .$$

The right-hand side is a quadratic function of Ξ'; it is minimized at $\Xi' = \Xi_k - \nabla\psi(\Xi_k)$, where its value is $\psi(\Xi_k) - \frac{1}{2}\,\|\nabla\psi(\Xi_k)\|^2$. Hence the inequality holds a fortiori with this minimal value on the right, for every Ξ'; taking $\Xi' = \Xi^\star_\Gamma$ on the left,

$$\psi(\Xi^\star_\Gamma) \ge \psi(\Xi_k) - \tfrac{1}{2}\,\|\nabla\psi(\Xi_k)\|^2 ,$$

that is,

$$\|g_k\|^2 = \|\nabla\psi(\Xi_k)\|^2 \ge 2\big[\psi(\Xi_k) - \psi(\Xi^\star_\Gamma)\big] = 2\Delta_k.$$

We record this as

$$\|g_k\|^2 \ge 2\Delta_k. \tag{55}$$

The recursion (i). Subtract $\psi(\Xi^\star_\Gamma)$ from both sides of eq. (54); since $\Delta_j = \psi(\Xi_j) - \psi(\Xi^\star_\Gamma)$ for every j, this gives

$$\Delta_{k+1} \le \Delta_k - \tfrac{1}{2L}\,\|g_k\|^2 + \tfrac{1}{2L}\,\|e_k\|^2 .$$

We now use eq. (55) to bound $-\frac{1}{2L}\,\|g_k\|^2 \le -\frac{1}{2L}\cdot 2\Delta_k = -\frac{1}{L}\Delta_k$, and eq. (53) to bound $\frac{1}{2L}\,\|e_k\|^2 \le \frac{1}{2L}\cdot 2L\delta_{\mathcal{R}\mathrm{OT}} = \delta_{\mathcal{R}\mathrm{OT}}$. Therefore

$$\Delta_{k+1} \le \Delta_k - \tfrac{1}{L}\Delta_k + \delta_{\mathcal{R}\mathrm{OT}} = q\,\Delta_k + \delta_{\mathcal{R}\mathrm{OT}}, \qquad q = 1 - \tfrac{1}{L}.$$

Applying this inequality k times, from Δ_0 to Δ_k, we obtain

$$\Delta_k \le q\Delta_{k-1} + \delta_{\mathcal{R}\mathrm{OT}} \le q^2\Delta_{k-2} + q\,\delta_{\mathcal{R}\mathrm{OT}} + \delta_{\mathcal{R}\mathrm{OT}} \le \cdots \le q^k\Delta_0 + \delta_{\mathcal{R}\mathrm{OT}}\sum_{j=0}^{k-1} q^j.$$

Since $0 \le q < 1$, the geometric sum satisfies $\sum_{j=0}^{k-1} q^j \le \sum_{j=0}^{\infty} q^j = \frac{1}{1-q} = L$, which proves eq. (49).

The certificate bound (ii). By definition, $\mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) = \frac{1}{2}\|\widehat{g}_k\|^2 + \delta_{\mathcal{R}\mathrm{OT}}$, so we need to bound $\|\widehat{g}_k\| = \|g_k + e_k\|$. We first bound the exact gradient by the gap. The descent lemma eq. (47) for h_Γ along the segment $[\Xi_k, \Xi_k + \frac{1}{L}g_k]$ (family (d)) gives

$$\begin{aligned} h_\Gamma\big(\Xi_k + \tfrac{1}{L}g_k\big) &\geq h_\Gamma(\Xi_k) + \big\langle g_k, \tfrac{1}{L}g_k\big\rangle - \tfrac{L}{2}\big\|\tfrac{1}{L}g_k\big\|^2 \\ &= h_\Gamma(\Xi_k) + \tfrac{1}{L}\|g_k\|^2 - \tfrac{1}{2L}\|g_k\|^2 = h_\Gamma(\Xi_k) + \tfrac{1}{2L}\|g_k\|^2. \end{aligned}$$

Since $\Xi_\Gamma^\star$ maximizes h_Γ, the left-hand side is at most $h_\Gamma(\Xi_\Gamma^\star)$, and we get

$$\tfrac{1}{2L}\|g_k\|^2 \leq h_\Gamma(\Xi_\Gamma^\star) - h_\Gamma(\Xi_k) = \Delta_k, \qquad \text{that is,} \qquad \|g_k\| \leq \sqrt{2L\Delta_k}.$$

Together with $\|e_k\| \leq \sqrt{2L\delta_{\mathcal{R}\mathrm{OT}}}$ from eq. (53), the triangle inequality gives

$$\|\widehat{g}_k\| \leq \|g_k\| + \|e_k\| \leq \sqrt{2L\Delta_k} + \sqrt{2L\delta_{\mathcal{R}\mathrm{OT}}}.$$

Squaring and using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\tfrac{1}{2}\|\widehat{g}_k\|^2 \leq \tfrac{1}{2}\big(2 \cdot 2L\Delta_k + 2 \cdot 2L\delta_{\mathcal{R}\mathrm{OT}}\big) = 2L\Delta_k + 2L\delta_{\mathcal{R}\mathrm{OT}}.$$

Adding $\delta_{\mathcal{R}\mathrm{OT}}$ on both sides proves eq. (50).

The iteration count (iii). Fix $\eta \in (0, 1]$ and set $\delta_{\mathcal{R}\mathrm{OT}} = \eta/(2C_L)$. Inserting the bound eq. (49) on Δ_k into the certificate bound eq. (50),

$$\begin{aligned} \mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) &\leq 2L\big(q^k\Delta_0 + L\delta_{\mathcal{R}\mathrm{OT}}\big) + (2L+1)\delta_{\mathcal{R}\mathrm{OT}} \\ &= 2L\,q^k\Delta_0 + \big(2L^2 + 2L + 1\big)\delta_{\mathcal{R}\mathrm{OT}} \\ &= 2L\,q^k\Delta_0 + C_L\,\delta_{\mathcal{R}\mathrm{OT}}. \end{aligned}$$

By the choice of $\delta_{\mathcal{R}\mathrm{OT}}$, the second term is exactly $C_L \cdot \eta/(2C_L) = \eta/2$. It therefore suffices to show that the first term is at most $\eta/2$ for $k \geq N_\Xi(\eta)$.

Since $\log(1-x) \leq -x$ for $x \in [0, 1)$, we have $q^k = (1 - \frac{1}{L})^k = e^{k\log(1-1/L)} \leq e^{-k/L}$ when $L > 1$ (and $q^k \leq e^{-k/L}$ trivially when $L = 1$, since then $q = 0$). Hence $2Lq^k\Delta_0 \leq \eta/2$ holds as soon as

$$2L\,e^{-k/L}\Delta_0 \leq \frac{\eta}{2} \quad\Longleftrightarrow\quad e^{-k/L} \leq \frac{\eta}{4L\Delta_0} \quad\Longleftrightarrow\quad k \geq L\log\frac{4L\Delta_0}{\eta}$$

(if $\Delta_0 = 0$ there is nothing to prove). If $4L\Delta_0/\eta \leq 1$, the right-hand side is nonpositive and the condition holds for every $k \geq 0$; otherwise it holds for every $k \geq L\log(4L\Delta_0/\eta)$. Both cases are summarized by

$$k \geq L\log\max\Big\{1, \frac{4L\Delta_0}{\eta}\Big\},$$

and since k is an integer, by $k \geq N_\Xi(\eta)$ with $N_\Xi(\eta)$ as in eq. (51). This proves $\mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) \leq \eta$ for every $k \geq N_\Xi(\eta)$.

Three remarks complete the proof of (iii).

- *Upper bounds on* Δ_0. If Δ_0 is replaced by any larger number $\overline{\Delta}_0$ in the argument above, the conclusion still holds, since $2Lq^k\overline{\Delta}_0 \geq 2Lq^k\Delta_0$; hence $N_\Xi(\eta)$ computed with $\overline{\Delta}_0$ is also a valid count.
- *Number of solves.* The inner loop of algorithm 1 stops at the first index k with $\mathrm{Cert}_\Gamma(\Xi_k, \widehat{\pi}_k) \leq \eta$. Since the certificate holds at $k = N_\Xi(\eta)$, this first index is at most $N_\Xi(\eta)$, and the loop makes at most $N_\Xi(\eta) + 1$ $\mathcal{R}\mathrm{OT}$ solves (one per index $k = 0, \ldots, N_\Xi(\eta)$).
- *The bound* $\Delta_0 \leq \frac{L}{2}\|\Xi_0 - \Xi_\Gamma^\star\|^2$. The descent lemma eq. (47) for h_Γ along the segment $[\Xi_\Gamma^\star, \Xi_0]$ (family (c) with $k = 0$), together with $\nabla h_\Gamma(\Xi_\Gamma^\star) = 0$, gives

  $$\begin{aligned} h_\Gamma(\Xi_0) &\geq h_\Gamma(\Xi_\Gamma^\star) + \langle \nabla h_\Gamma(\Xi_\Gamma^\star), \Xi_0 - \Xi_\Gamma^\star\rangle - \tfrac{L}{2}\|\Xi_0 - \Xi_\Gamma^\star\|^2 \\ &= h_\Gamma(\Xi_\Gamma^\star) - \tfrac{L}{2}\|\Xi_0 - \Xi_\Gamma^\star\|^2, \end{aligned}$$

  that is, $\Delta_0 \leq \frac{L}{2}\|\Xi_0 - \Xi_\Gamma^\star\|^2$. In the local form, $\|\Xi_0 - \Xi_\Gamma^\star\| \leq \|\Xi_0\| + \|\Xi_\Gamma^\star\| \leq 2R$, so $\Delta_0 \leq \frac{L}{2}(2R)^2 = 2LR^2$.

Finally, the local form of **($\mathbf{H}_L$)** requires $\delta_{\mathcal{R}\mathrm{OT}} \le \frac{1}{2}$; this is satisfied, because $\eta \le 1$ and $C_L = 2L^2 + 2L + 1 \ge 5$ for $L \ge 1$ give $\delta_{\mathcal{R}\mathrm{OT}} = \eta/(2C_L) \le \frac{1}{10}$. □

Corollary 1 (Oracle complexity of algorithm 1; precise form of theorem 5(ii))**.** *Assume **(S1)–(S2)**, let $\delta \in (0, 1]$, and suppose that **($\boldsymbol{H}_L$)** holds at $\Gamma = \Gamma_t$ for every $t < T_\delta$ with the same constants L (and R). Let*

$$\overline{\Delta} := \sup_{t<T_\delta} \Big[h_{\Gamma_t}(\Xi^\star_{\Gamma_t}) - h_{\Gamma_t}(\Xi_{t,0}) \Big] \le \frac{L}{2} \sup_{t<T_\delta} \big\| \Xi_{t,0} - \Xi^\star_{\Gamma_t} \big\|^2$$

be the largest initial inner gap. Run algorithm 1 with $\tau = 1/L$ and $\delta_{\mathcal{R}\mathrm{OT}} = \delta^2/(8C_L)$, where $C_L = 2L^2 + 2L + 1$. Then every inner loop terminates, with

$$K_t \le N_\Xi(\delta^2/4) = \Big\lceil L \log \max\Big\{1, \frac{16L\overline{\Delta}}{\delta^2}\Big\} \Big\rceil,$$

the conclusions of theorem 12 hold, and the total number of $\mathcal{R}\mathrm{OT}$ solves is at most

$$T_\delta\big(N_\Xi(\delta^2/4) + 1\big) = O\Big(\frac{\Delta^{\mathrm{out}} L}{\delta^2} \log \frac{L\overline{\Delta}}{\delta^2}\Big).$$

Proof. Fix an outer iteration $t < T_\delta$ and apply proposition 7(iii) at $\Gamma = \Gamma_t$ with the level $\eta := \delta^2/4 \in (0, 1]$. Its accuracy requirement $\delta_{\mathcal{R}\mathrm{OT}} = \eta/(2C_L)$ is exactly the value prescribed here, since

$$\frac{\eta}{2C_L} = \frac{\delta^2/4}{2C_L} = \frac{\delta^2}{8C_L},$$

and its iteration threshold, computed with the upper bound $\overline{\Delta} \ge \Delta_{t,0}$ allowed by proposition 7(iii), is

$$N_\Xi(\eta) = \Big\lceil L \log \max\Big\{1, \frac{4L\overline{\Delta}}{\eta}\Big\} \Big\rceil = \Big\lceil L \log \max\Big\{1, \frac{4L\overline{\Delta}}{\delta^2/4}\Big\} \Big\rceil = \Big\lceil L \log \max\Big\{1, \frac{16L\overline{\Delta}}{\delta^2}\Big\} \Big\rceil,$$

which is the quantity $N_\Xi(\delta^2/4)$ of the statement. Hence the certificate of iteration t is at most $\delta^2/4$ for every inner index $k \ge N_\Xi(\delta^2/4)$, so that the first certified index satisfies $K_t \le N_\Xi(\delta^2/4) < \infty$.

Since every inner loop terminates, theorem 12 applies: the run stops at some $t_{\mathrm{stop}} < T_\delta$, and it makes $\sum_{t \le t_{\mathrm{stop}}} (K_t + 1) \le T_\delta\big(N_\Xi(\delta^2/4) + 1\big)$ $\mathcal{R}\mathrm{OT}$ solves. Finally, $T_\delta \le 1 + 4\Delta^{\mathrm{out}}/\delta^2$ and $N_\Xi(\delta^2/4) + 1 \le 2 + L \log \max\{1, 16L\overline{\Delta}/\delta^2\}$, whose product is $O\big(\frac{\Delta^{\mathrm{out}} L}{\delta^2} \log \frac{L\overline{\Delta}}{\delta^2}\big)$. The bound on $\overline{\Delta}$ in the statement is the last claim of proposition 7(iii), applied at each t. □

E.5 Accelerated versions.

The previous analysis uses plain gradient ascent for the reduced problem in Ξ. The same framework immediately allows acceleration whenever the $\mathcal{R}\mathrm{OT}$ oracle converges geometrically.

Proposition 8 (Acceleration with a geometric $\mathcal{R}\mathrm{OT}$ oracle)**.** *Fix Γ and let h_Γ be L-smooth on a convex region containing its maximizer and all iterates. Since h_Γ is 1-strongly convex, set $\kappa :=$. Assume that a ζ-accurate $\mathcal{R}\mathrm{OT}$ solve can be computed in*

$$N_{\mathcal{R}\mathrm{OT}}(\zeta) = O\Big(\log \frac{1}{\zeta}\Big) \tag{56}$$

iterations, up to constants independent of the target accuracy.

Then, for any target inner accuracy $\eta > 0$, an accelerated inexact-gradient method applied to $-h_\Gamma$ reaches

$$\mathrm{Cert}_\Gamma(\Xi, \widehat{\pi}) \le \eta$$

using

$$O\Big(\sqrt{\kappa}\, \log \frac{L\Delta_0}{\eta}\Big) \tag{57}$$

$\mathcal{R}\mathrm{OT}$ calls, provided each call is solved to sufficiently small accuracy, for instance

$$\zeta = O\Big(\frac{\eta}{L\sqrt{\kappa}}\Big). \tag{58}$$

Consequently, the total number of iterations of the underlying $\mathcal{R}$OT solver is

$$O\left(\sqrt{\kappa}\,\log\frac{L\Delta_0}{\eta}\,\log\frac{L\sqrt{\kappa}}{\eta}\right)=\widetilde{O}(\sqrt{\kappa}). \tag{59}$$

In particular, arbitrarily accurate $\mathcal{R}$OT solves preserve, up to logarithmic factors, the usual accelerated dependence on the condition number.

Proof. A ζ-accurate $\mathcal{R}$OT solution provides an inexact first-order oracle for h_Γ; in particular, by lemma 9,

$$\|\widehat{g}-\nabla h_\Gamma(\Xi)\|\le\sqrt{2L\zeta}.$$

Standard accelerated methods for smooth strongly convex optimization with an inexact oracle converge in $O(\sqrt{\kappa}\log(1/\eta))$ iterations, with an error floor proportional to $\sqrt{\kappa}\,\zeta$ (Devolder et al., 2013). Choosing $\zeta=O(\eta/(L\sqrt{\kappa}))$ makes this error negligible at the accuracy required by the certificate. Since the $\mathcal{R}$OT solver is geometrically convergent, this increased oracle accuracy changes the cost of each call only logarithmically. □

F Entropic unbalanced Gromov–Wasserstein: proof of Theorem 13

We now specialize the preceding analysis to entropically regularized unbalanced Gromov–Wasserstein with KL penalties on both marginals. The UOT subproblems are solved by generalized Sinkhorn iterations Chizat et al. (2018); Pham et al. (2020). For standard facts on entropic UOT we follow Séjourné et al. (2023, Section 4), whose normalization (entropy relative to $\alpha\otimes\beta$, potentials (f,g)) coincides with ours. We focus on the additional regularity induced by the Gromov–Wasserstein DC structure and the end-to-end convergence of our algorithm for this specific case.

We denote by $\varepsilon>0$ the entropic regularization parameter, by $\rho_1,\rho_2>0$ the marginal penalties (table 2), by $\delta_{\mathcal{R}\mathrm{OT}}>0$ the accuracy of the UOT oracle eq. (30), and by $\delta>0$ the DC-criticality tolerance of theorem 12.

We consider discrete measures

$$\alpha=\sum_{i=1}^{\mathrm{N}}\alpha_i\delta_{x_i},\qquad \beta=\sum_{j=1}^{\mathrm{M}}\beta_j\delta_{y_j},$$

with strictly positive weights, and set

$$M_\alpha:=\sum_i\alpha_i,\qquad M_\beta:=\sum_j\beta_j.$$

Writing $a_{ij}:=A(x_i,y_j)$ and $b_{ij}:=B(x_i,y_j)$, we have

$$A[\pi]=\sum_{ij}\pi_{ij}a_{ij},\qquad B[\pi]=\sum_{ij}\pi_{ij}b_{ij},$$

and we let

$$R_A:=\max_{i,j}\|a_{ij}\|,\qquad R_B:=\max_{i,j}\|b_{ij}\|.$$

The objective is eq. (24) with

$$\mathcal{R}_{\mathrm{UOT}}(\pi):=\varepsilon D_{\mathrm{KL}}(\pi\|\alpha\otimes\beta)+\rho_1D_{\mathrm{KL}}(\pi_1\|\alpha)+\rho_2D_{\mathrm{KL}}(\pi_2\|\beta), \tag{60}$$

where $D_{\mathrm{KL}}(p\|q):=\sum_l\big[p_l\log(p_l/q_l)-p_l+q_l\big]$ is the generalized Kullback–Leibler divergence between nonnegative vectors. This is the entropic UOT functional of Séjourné et al. (2023, Definition 7) with $D_{\varphi_i}=\rho_iD_{\mathrm{KL}}$. Sinkhorn rates are stated in the symmetric case $\rho_1=\rho_2=\rho$.

At fixed (Γ,Ξ), define

$$P_{\Gamma,\Xi}(\pi):=\langle c_{\Gamma,\Xi},\pi\rangle+\mathcal{R}_{\mathrm{UOT}}(\pi),\qquad \mathrm{UOT}(c_{\Gamma,\Xi}):=\inf_{\pi\ge0}P_{\Gamma,\Xi}(\pi), \tag{61}$$

with

$$c_{\Gamma,\Xi,ij}=\langle\Xi,a_{ij}\rangle-\langle\Gamma,b_{ij}\rangle. \tag{62}$$

Thus $\mathrm{UOT}(c)$ is the value $\mathcal{R}\mathrm{OT}(c)$ of eq. (26) for $\mathcal{R}=\mathcal{R}_{\mathrm{UOT}}$, and the corresponding reduced objective is

$$h_\Gamma(\Xi):=-\frac{1}{2}\|\Xi\|^2+\mathrm{UOT}(c_{\Gamma,\Xi}). \tag{63}$$

Regularity with respect to the GW variable.

Proposition 9 (Regularity of the reduced UGW problem)**.** *Fix Γ and $R > 0$, and set*

$$C_\Gamma(R) := R_A R + R_B \|\Gamma\|,$$
$$M_\Gamma(R) := M_\alpha^{\frac{\varepsilon+\rho_1}{\varepsilon+\rho_1+\rho_2}} M_\beta^{\frac{\varepsilon+\rho_2}{\varepsilon+\rho_1+\rho_2}} \exp\left(\frac{C_\Gamma(R)}{\varepsilon+\rho_1+\rho_2}\right). \tag{64}$$

Then:

(i) *for every Ξ, the UOT problem eq.* (61) *has a unique minimizer $\pi_{\Gamma,\Xi}$, which is entrywise positive and depends continuously differentiably on Ξ; the function h_Γ is finite and twice continuously differentiable on $\mathcal{H}_A$, with*

$$\nabla h_\Gamma(\Xi) = A[\pi_{\Gamma,\Xi}] - \Xi; \tag{65}$$

(ii) *if $\|\Xi\| \le R$, then*

$$\|\pi_{\Gamma,\Xi}\|_1 \le M_\Gamma(R); \tag{66}$$

(iii) *h_Γ is 1-strongly concave on $\mathcal{H}_A$, and on $\{\|\Xi\| \le R\}$ its Hessian satisfies $-L_h I \preceq \nabla^2 h_\Gamma \preceq -I$ with*

$$L_h := 1 + \frac{R_A^2 M_\Gamma(R)}{\varepsilon}. \tag{67}$$

Proof. (i). Since $\varepsilon D_{\mathrm{KL}}(\cdot\|\alpha\otimes\beta)$ is strictly convex, $P_{\Gamma,\Xi}$ has at most one minimizer. The dual of eq. (61) (Séjourné et al., 2023, Proposition 2) is the maximization over $(f,g) \in \mathbb{R}^{\mathrm{N}} \times \mathbb{R}^{\mathrm{M}}$ of a smooth, strictly concave and coercive function. It therefore has a unique maximizer $(f^\star, g^\star)$. The minimizer of eq. (61) is then given by the primal–dual relation (Séjourné et al., 2023, Equation (21)):

$$\pi_{\Gamma,\Xi,ij} = \alpha_i\beta_j \exp\Big(\frac{f_i^\star + g_j^\star - c_{\Gamma,\Xi,ij}}{\varepsilon}\Big) > 0. \tag{68}$$

In particular, $\mathrm{UOT}(c_{\Gamma,\Xi})$ and h_Γ are finite.

On the open set $\{\pi > 0\}$, $P_{\Gamma,\Xi}$ is smooth and $\pi_{\Gamma,\Xi}$ is characterized by the stationarity condition

$$\nabla_\pi P_{\Gamma,\Xi}(\pi) = c_{\Gamma,\Xi} + \nabla\mathcal{R}_{\mathrm{UOT}}(\pi) = 0. \tag{69}$$

Its Jacobian with respect to π is $H_\pi := \nabla^2\mathcal{R}_{\mathrm{UOT}}(\pi) \succeq \varepsilon\,\mathrm{diag}(1/\pi_{ij}) \succ 0$, because the marginal KL terms are convex. By the implicit function theorem, $\Xi \mapsto \pi_{\Gamma,\Xi}$ is continuously differentiable. Since $c_{\Gamma,\Xi} = A^\top\Xi - B^\top\Gamma$, differentiating eq. (69) gives

$$D_\Xi \pi_{\Gamma,\Xi} = -H_\pi^{-1}A^\top, \qquad \text{with } H_\pi := H_{\pi_{\Gamma,\Xi}}. \tag{70}$$

Finally, $\mathrm{UOT}(c_{\Gamma,\Xi}) = P_{\Gamma,\Xi}(\pi_{\Gamma,\Xi})$, where $(\pi,\Xi) \mapsto P_{\Gamma,\Xi}(\pi) = \langle\Xi, A[\pi]\rangle - \langle\Gamma, B[\pi]\rangle + \mathcal{R}_{\mathrm{UOT}}(\pi)$ is smooth on $\{\pi > 0\} \times \mathcal{H}_A$. By the chain rule, and since the π-gradient vanishes at $\pi_{\Gamma,\Xi}$ by eq. (69) (envelope theorem), we get $\nabla_\Xi \mathrm{UOT}(c_{\Gamma,\Xi}) = A[\pi_{\Gamma,\Xi}]$. This gives eq. (65), in accordance with proposition 6 and eq. (44). Since the right-hand side is C^1, h_Γ is C^2.

(ii). Let $\|\Xi\| \le R$, and write $\pi := \pi_{\Gamma,\Xi}$ and $m := \|\pi\|_1 > 0$. Since π is a minimizer, $t \mapsto P_{\Gamma,\Xi}(t\pi)$ is minimal at $t = 1$. Setting its derivative to zero, as in Pham et al. (2020, Lemma 4), gives

$$0 = \langle c_{\Gamma,\Xi}, \pi\rangle + \varepsilon\sum_{ij}\pi_{ij}\log\frac{\pi_{ij}}{\alpha_i\beta_j} + \rho_1\sum_i(\pi_1)_i\log\frac{(\pi_1)_i}{\alpha_i} + \rho_2\sum_j(\pi_2)_j\log\frac{(\pi_2)_j}{\beta_j}.$$

We first bound the cost term. By Cauchy–Schwarz and the definitions of R_A and R_B,

$$c_{\Gamma,\Xi,ij} = \langle\Xi, a_{ij}\rangle - \langle\Gamma, b_{ij}\rangle \quad \ge -\|\Xi\|\,\|a_{ij}\| - \|\Gamma\|\,\|b_{ij}\| \ge -R_A R - R_B\|\Gamma\| = -C_\Gamma(R).$$

Since $\pi_{ij} \ge 0$, summing this entrywise bound against π gives

$$\langle c_{\Gamma,\Xi}, \pi\rangle = \sum_{ij} c_{\Gamma,\Xi,ij}\pi_{ij} \qquad \ge -C_\Gamma(R)\sum_{ij}\pi_{ij} = -C_\Gamma(R)m.$$

Recall that the marginals of π are $(\pi_1)_i = \sum_j \pi_{ij}, (\pi_2)_j = \sum_i \pi_{ij}$. Hence,both have the same total mass as π:

$$\sum_i (\pi_1)_i = \sum_j (\pi_2)_j = \sum_{i,j} \pi_{ij} = m.$$

We can therefore apply the log-sum inequality to each entropy term. Since

$$\sum_{i,j} \alpha_i \beta_j = M_\alpha M_\beta,$$

it gives

$$\sum_{i,j} \pi_{ij} \log \frac{\pi_{ij}}{\alpha_i \beta_j} \geq m \log \frac{m}{M_\alpha M_\beta},$$

$$\sum_i (\pi_1)_i \log \frac{(\pi_1)_i}{\alpha_i} \geq m \log \frac{m}{M_\alpha},$$

$$\sum_j (\pi_2)_j \log \frac{(\pi_2)_j}{\beta_j} \geq m \log \frac{m}{M_\beta}.$$

Combining these inequalities with $\langle c_{\Gamma,\Xi}, \pi \rangle \geq -C_\Gamma(R)m$ yields

$$0 \geq -C_\Gamma(R)m + \varepsilon m \log \frac{m}{M_\alpha M_\beta} + \rho_1 m \log \frac{m}{M_\alpha} + \rho_2 m \log \frac{m}{M_\beta}.$$

Dividing by $m > 0$ and collecting the logarithmic terms gives

$$0 \geq -C_\Gamma(R) + (\varepsilon + \rho_1 + \rho_2) \log m \qquad -(\varepsilon + \rho_1) \log M_\alpha - (\varepsilon + \rho_2) \log M_\beta.$$

which is eq. (66) after rearranging and exponentiating.

(iii). Strong concavity is lemma 7, since h_Γ is finite. Let $\|\Xi\| \leq R$. Differentiating eq. (65) and using eq. (70),

$$\nabla^2 h_\Gamma(\Xi) = -I - A H_\pi^{-1} A^\top. \tag{71}$$

From $H_\pi \succeq \varepsilon \operatorname{diag}(1/\pi_{ij})$ we get $H_\pi^{-1} \preceq \varepsilon^{-1} \operatorname{diag}(\pi_{ij})$. Hence, for $u \in \mathcal{H}_A$, by Cauchy–Schwarz, $\|a_{ij}\| \leq R_A$ and eq. (66),

$$0 \leq \left\langle u, A H_\pi^{-1} A^\top u \right\rangle \leq \frac{1}{\varepsilon} \sum_{ij} \pi_{ij} \langle u, a_{ij} \rangle^2 \leq \frac{R_A^2 \|\pi_{\Gamma,\Xi}\|_1}{\varepsilon} \|u\|^2 \leq \frac{R_A^2 M_\Gamma(R)}{\varepsilon} \|u\|^2 .$$

With eq. (71), this gives $-L_h I \preceq \nabla^2 h_\Gamma(\Xi) \preceq -I$. □

Remark 2 (No exponential deterioration in $1/\varepsilon$)**.** *The mass bound eq.* (64) *contains* $\exp\big(C_\Gamma(R)/(\varepsilon + \rho_1 + \rho_2)\big)$ *rather than* $\exp(C/\varepsilon)$. *Thus, for fixed positive marginal penalties and uniformly bounded* R *and* $\|\Gamma\|$, $M_\Gamma(R) = O(1)$ *and* $L_h = O(\varepsilon^{-1})$ *as* $\varepsilon \downarrow 0$.

Sinkhorn iterations. Assume $\rho_1 = \rho_2 = \rho$ and fix (Γ, Ξ). The dual of the UOT problem eq. (61) is solved by the unbalanced Sinkhorn algorithm (Chizat et al., 2018), (Séjourné et al., 2023, Definition 10). For KL penalties it alternates the two half-steps (Séjourné et al., 2023, Section 4.2)

$$f_i \leftarrow -\frac{\rho\varepsilon}{\rho + \varepsilon} \log \sum_j \beta_j \, e^{(g_j - c_{\Gamma,\Xi,ij})/\varepsilon}, \qquad g_j \leftarrow -\frac{\rho\varepsilon}{\rho + \varepsilon} \log \sum_i \alpha_i \, e^{(f_i - c_{\Gamma,\Xi,ij})/\varepsilon}, \tag{72}$$

and encodes the current plan as $\pi_{ij} = \alpha_i \beta_j e^{(f_i + g_j - c_{\Gamma,\Xi,ij})/\varepsilon}$. Up to the change of cost $C_{ij} = c_{\Gamma,\Xi,ij} - \varepsilon \log(\alpha_i \beta_j)$, this is Algorithm 1 of Pham et al. (2020) with $(\tau, \eta) = (\rho, \varepsilon)$. The optimal potentials $(f^\star, g^\star)$ are the fixed points of eq. (72) (Séjourné et al., 2023, Proposition 3). The log-sum-exp maps in eq. (72) are 1-Lipschitz for the sup norm. Hence each half-step is a contraction of factor $\rho/(\rho + \varepsilon)$ towards the optimum, for any current potentials:

$$\|f^{\text{new}} - f^\star\|_\infty \leq \frac{\rho}{\rho + \varepsilon} \|g - g^\star\|_\infty , \qquad \|g^{\text{new}} - g^\star\|_\infty \leq \frac{\rho}{\rho + \varepsilon} \|f - f^\star\|_\infty . \tag{73}$$

This is the key step in the proof of Theorem 1 of Pham et al. (2020), derived there from their Lemmas 1–2. See also Séjourné et al. (2023, Theorem 2), which states the resulting rate $(\rho/(\rho+\varepsilon))^2$ per complete iteration. We start from $g^{(0)}$ and call one *complete iteration* the update $g^{(s-1)} \to f^{(s)} \to g^{(s)}$ obtained by applying the two half-steps eq. (72) in this order; $\pi^{(s)}$ denotes the plan encoded by $(f^{(s)}, g^{(s)})$, and

$$D_s := \max\Big\{\Big\|f^{(s)} - f^\star\Big\|_\infty, \Big\|g^{(s)} - g^\star\Big\|_\infty\Big\} \quad (s \geq 1), \qquad D_0 := \Big\|g^{(0)} - g^\star\Big\|_\infty$$

measure the distance to the optimal potentials. Applying eq. (73) $2s-1$ times and then $2s$ times,

$$\Big\|f^{(s)} - f^\star\Big\|_\infty \leq \Big(\frac{\rho}{\rho+\varepsilon}\Big)^{2s-1} D_0, \qquad \Big\|g^{(s)} - g^\star\Big\|_\infty \leq \Big(\frac{\rho}{\rho+\varepsilon}\Big)^{2s} D_0, \tag{74}$$

and therefore

$$D_s \leq \Big(\frac{\rho}{\rho+\varepsilon}\Big)^{2s-1} D_0 \qquad \text{for all } s \geq 1. \tag{75}$$

Lemma 10 (Sinkhorn as a primal UOT oracle)**.** *Assume $\rho_1 = \rho_2 = \rho$, fix Γ and Ξ with $\|\Xi\| \leq R$, and let $M_\Gamma(R)$ be as in proposition 9. Then:*

(i) *for every $s \geq 1$ such that $D_s \leq \varepsilon/2$,*

$$P_{\Gamma,\Xi}(\pi^{(s)}) - \mathrm{UOT}(c_{\Gamma,\Xi}) \leq \frac{8\,(\varepsilon + 2\rho)\,M_\Gamma(R)}{\varepsilon^2}\, D_s^2; \tag{76}$$

(ii) *consequently, with the constant*

$$C_{\mathrm{S}} := \max\Big\{1,\ \frac{2D_0}{\varepsilon},\ \frac{8(\varepsilon+2\rho)M_\Gamma(R)}{\varepsilon^2} D_0^2\Big\}, \tag{77}$$

the plan $\pi^{(s)}$ is $\delta_{\mathcal{R}\mathrm{OT}}$-accurate for every $\delta_{\mathcal{R}\mathrm{OT}} \leq 1$ as soon as

$$s \geq N_{\mathrm{S}}(\delta_{\mathcal{R}\mathrm{OT}}) := \Big\lceil \frac{1}{2} + \frac{\log(C_{\mathrm{S}}/\delta_{\mathcal{R}\mathrm{OT}})}{2\log(1+\varepsilon/\rho)} \Big\rceil; \tag{78}$$

moreover

$$N_{\mathrm{S}}(\delta_{\mathcal{R}\mathrm{OT}}) \leq \frac{3}{2} + \frac{\rho+\varepsilon}{2\varepsilon} \log \frac{C_{\mathrm{S}}}{\delta_{\mathcal{R}\mathrm{OT}}}. \tag{79}$$

Proof. Let $\pi^\star := \pi_{\Gamma,\Xi}$. Since $\nabla P_{\Gamma,\Xi}(\pi^\star) = 0$, the primal gap is the Bregman divergence of $P_{\Gamma,\Xi}$ at $\pi^\star$. Using the linearity of the marginal maps,

$$P_{\Gamma,\Xi}(\pi) - \mathrm{UOT}(c_{\Gamma,\Xi}) = \varepsilon D_{\mathrm{KL}}(\pi\|\pi^\star) + \rho D_{\mathrm{KL}}(\pi_1\|\pi_1^\star) + \rho D_{\mathrm{KL}}(\pi_2\|\pi_2^\star). \tag{80}$$

(i). By the Gibbs representation of $\pi^{(s)}$ and $\pi^\star$,

$$\left|\log \frac{\pi_{ij}^{(s)}}{\pi_{ij}^\star}\right| \leq \frac{2D_s}{\varepsilon} =: \theta.$$

The same bound holds for the ratios of both marginals, since each marginal ratio is a weighted average of the corresponding entrywise ratios. If $\theta \leq 1$, the elementary bound

$$u\log u - u + 1 \leq 2\theta^2, \qquad e^{-\theta} \leq u \leq e^\theta,$$

therefore gives

$$D_{\mathrm{KL}}(\pi^{(s)}\|\pi^\star), \quad D_{\mathrm{KL}}(\pi_1^{(s)}\|\pi_1^\star), \quad D_{\mathrm{KL}}(\pi_2^{(s)}\|\pi_2^\star) \leq 2\theta^2\,\|\pi^\star\|_1\,.$$

Using $\|\pi^\star\|_1 \leq M_\Gamma(R)$ and $\theta = 2D_s/\varepsilon$ in eq. (80) yields

$$P_{\Gamma,\Xi}(\pi^{(s)}) - \mathrm{UOT}(c_{\Gamma,\Xi}) \leq \frac{8(\varepsilon+2\rho)M_\Gamma(R)}{\varepsilon^2} D_s^2.$$

(ii). Set $q := \frac{\rho}{\rho+\varepsilon}$. By eq. (75),

$$D_s \le q^{2s-1} D_0.$$

If $s \ge N_{\mathrm{S}}(\delta_{\mathcal{R}\mathrm{OT}})$, then

$$q^{2s-1} \le \frac{\delta_{\mathcal{R}\mathrm{OT}}}{C_{\mathrm{S}}}.$$

By the definition of C_{S} and $\delta_{\mathcal{R}\mathrm{OT}} \le 1$, this implies both $D_s \le \varepsilon/2$ and

$$P_{\Gamma,\Xi}(\pi^{(s)}) - \mathrm{UOT}(c_{\Gamma,\Xi}) \le \frac{8(\varepsilon+2\rho)M_\Gamma(R)}{\varepsilon^2} D_0^2 q^{2(2s-1)} \qquad \le C_{\mathrm{S}} q^{2s-1} \le \delta_{\mathcal{R}\mathrm{OT}}.$$

This proves eq. (78). Finally, eq. (79) follows from $\lceil x \rceil \le x+1$ and

$$\log(1+x) \ge \frac{x}{1+x}.$$

□

End-to-end rates. For the UGW problem considered here, hypotheses **(S1)**–**(S2)** hold since $\mathcal{R}_{\mathrm{UOT}}$ is proper, convex and lower semicontinuous, while both the GW energy and $\mathcal{R}_{\mathrm{UOT}}$ are nonnegative. Fix $\delta \in (0,1]$ and bounds $\overline{R}, \overline{\Gamma}, \overline{G} > 0$, and set

$$\overline{M} := \sup_{\|\Gamma\| \le \overline{\Gamma}} M_\Gamma(3\overline{R}+1), \tag{81}$$

$$\overline{L} := 1 + \frac{R_A^2 \overline{M}}{\varepsilon}, \qquad \overline{C}_L := 2\overline{L}^2 + 2\overline{L} + 1, \tag{82}$$

$$\overline{\delta}_{\mathcal{R}\mathrm{OT}} := \frac{\delta^2}{8\overline{C}_L}, \tag{83}$$

$$\overline{N}_\Xi := \left\lceil \overline{L} \log \max\left\{1, \frac{32\overline{L}^2\overline{R}^2}{\delta^2}\right\} \right\rceil. \tag{84}$$

Let $\overline{N}_{\mathrm{S}}$ denote the bound of lemma 10(ii) at accuracy $\overline{\delta}_{\mathcal{R}\mathrm{OT}}$, with $M_\Gamma(R)$ and D_0 replaced by their uniform bounds $\overline{M}$ and $\overline{G}$.

Theorem 13 (End-to-end complexity for entropic UGW)**.** *Assume $\rho_1 = \rho_2 = \rho$, and run algorithm 1 with $\tau = 1/\overline{L}$ and UOT accuracy $\overline{\delta}_{\mathcal{R}\mathrm{OT}}$, solving each UOT subproblem by Sinkhorn with at most $\overline{N}_{\mathrm{S}}$ complete iterations. Suppose that, throughout the run,*

$$\|\Gamma_t\| \le \overline{\Gamma}, \qquad \max\left\{\left\|\Xi^\star_{\Gamma_t}\right\|, \|\Xi_{t,k}\|\right\} \le \overline{R}, \qquad \left\|g^{(0)}_{t,k} - g^\star_{t,k}\right\|_\infty \le \overline{G}. \tag{85}$$

Then every inner loop terminates after at most $\overline{N}_\Xi$ iterations, and the algorithm terminates within T_δ outer iterations with a $(\delta, \delta^2/4)$-critical plan. In particular,

$$N^{\mathrm{tot}}_{\mathrm{Sink}} \le T_\delta(\overline{N}_\Xi + 1)\overline{N}_{\mathrm{S}}. \tag{86}$$

If ρ is fixed and $\overline{R}, \overline{\Gamma}, \overline{G} = O(1)$ as $\varepsilon, \delta \to 0$, then

$$N^{\mathrm{tot}}_{\mathrm{Sink}} = \widetilde{O}\left(\frac{\Delta^{\mathrm{out}}(\rho+\varepsilon)}{\delta^2\varepsilon^2}\right). \tag{87}$$

If moreover $\mathrm{N} = \mathrm{M} = n$ and A, B take values in fixed-dimensional Euclidean spaces, the total arithmetic complexity is

$$\widetilde{O}\left(\frac{n^2\Delta^{\mathrm{out}}(\rho+\varepsilon)}{\delta^2\varepsilon^2}\right). \tag{88}$$

Proof. Under the assumed bounds, proposition 9(iii) gives uniformly

$$-\overline{L}I \preceq \nabla^2 h_{\Gamma_t}(\Xi) \preceq -I \qquad \text{for all } \|\Xi\| \le 3\overline{R}+1. \tag{89}$$

Hence the local form of **($\mathbf{H}_L$)** holds with $L = \overline{L}$.

Moreover, lemma 10 and the bounds

$$M_{\Gamma_t}(\overline{R}) \le \overline{M}, \qquad \left\| g^{(0)}_{t,k} - g^{\star}_{t,k} \right\|_\infty \le \overline{G} \tag{90}$$

show that every UOT call reaches accuracy $\overline{\delta}_{\mathcal{R}\mathrm{OT}}$ within $\overline{N}_{\mathrm{S}}$ complete Sinkhorn iterations.

It remains to control the initial inner gap. Since $\nabla h_{\Gamma_t}(\Xi^\star_{\Gamma_t}) = 0$ and h_{Γ_t} is $\overline{L}$-smooth on the segment joining $\Xi_{t,0}$ and $\Xi^\star_{\Gamma_t}$,

$$h_{\Gamma_t}(\Xi^\star_{\Gamma_t}) - h_{\Gamma_t}(\Xi_{t,0}) \le \frac{\overline{L}}{2} \left\| \Xi_{t,0} - \Xi^\star_{\Gamma_t} \right\|^2 \tag{91}$$

$$\le 2\overline{L}\,\overline{R}^2. \tag{92}$$

Therefore corollary 1, with oracle accuracy

$$\frac{\delta^2}{8\overline{C}_L} = \overline{\delta}_{\mathcal{R}\mathrm{OT}},$$

gives

$$K_t \le \overline{N}_\Xi,$$

while theorem 12 gives termination within T_δ outer iterations and a $(\delta, \delta^2/4)$-critical plan.

There are at most $\overline{N}_\Xi + 1$ UOT calls per outer iteration and at most $\overline{N}_{\mathrm{S}}$ complete Sinkhorn iterations per call, which proves eq. (86).

Finally, for fixed ρ and bounded $\overline{R}, \overline{\Gamma}, \overline{G}$,

$$\overline{M} = O(1), \qquad \overline{L} = O(\varepsilon^{-1}), \qquad \overline{N}_\Xi = \widetilde{O}(\varepsilon^{-1}), \qquad \overline{N}_{\mathrm{S}} = \widetilde{O}\left(\frac{\rho+\varepsilon}{\varepsilon}\right). \tag{93}$$

Together with

$$T_\delta = O\left(\frac{\Delta^{\mathrm{out}}}{\delta^2}\right),$$

this yields eq. (87). When $\mathrm{N} = \mathrm{M} = n$, each Sinkhorn iteration, cost evaluation and moment computation costs $O(n^2)$, which gives the stated arithmetic complexity. □

Accelerated convergence rate. Previously, we established the convergence rate of our solver for the entropic unbalanced Gromov–Wasserstein problem. As discussed in section E.5, when h_Γ is smooth, the inner optimization can instead be accelerated using inexact first-order methods. This is particularly effective when the underlying $\mathcal{R}\mathrm{OT}$ solver converges geometrically, so that achieving accuracy δ requires only $O(\log(1/\delta))$ iterations. This is precisely the case here for generalized Sinkhorn, and we make the resulting accelerated complexity explicit below. Note also that, if the marginal penalties in $\mathcal{R}$ are chosen to be χ^2 rather than D_{KL}, one may similarly use the accelerated-gradient method of Genans (2026), which also enjoys geometric convergence.

Corollary 2 (Accelerated complexity for entropic UGW)**.** *Assume $\rho_1 = \rho_2 = \rho$ and the uniform bounds of theorem 13, and let*

$$\overline{L} = 1 + \frac{R_A^2 \overline{M}}{\varepsilon}.$$

At each outer iteration t, apply the accelerated inexact-gradient method of Devolder et al. (2013) to $\psi_t := -h_{\Gamma_t}$ on the ball $\{\|\Xi\| \le \overline{R}\}$, using at every oracle query a plan returned by generalized Sinkhorn with UOT accuracy

$$\overline{\delta}^{\mathrm{acc}}_{\mathcal{R}\mathrm{OT}} := \frac{\delta^2}{104\,\overline{L}^{3/2}}. \tag{94}$$

Then each inner problem reaches $\mathrm{Cert}_{\Gamma_t}(\Xi, \widehat{\pi}) \le \frac{\delta^2}{4}$ after $\widetilde{O}(\sqrt{\overline{L}})$ UOT calls.

Consequently, the outer algorithm terminates within

$$T_\delta = O\left(\frac{\Delta^{\mathrm{out}}}{\delta^2}\right)$$

iterations with a $(\delta, \delta^2/4)$-critical plan. If ρ is fixed and $\overline{R}, \overline{\Gamma}, \overline{G} = O(1)$ as $\varepsilon, \delta \to 0$, then $\overline{L} = O(\varepsilon^{-1})$ and

$$N^{\mathrm{tot}}_{\mathrm{Sink,acc}} = \widetilde{O}\left(\frac{\Delta^{\mathrm{out}}(\rho+\varepsilon)}{\delta^2 \varepsilon^{3/2}}\right). \tag{95}$$

Proof. A $\delta_{\mathcal{R}\mathrm{OT}}$-accurate UOT solve defines a $(2\delta_{\mathcal{R}\mathrm{OT}}, 2\overline{L}, 1)$ inexact oracle for ψ_t. Hence the accelerated bound of Devolder et al. (2013) gives

$$\psi_t(y_k) - \psi_t^\star \le C_0 \exp\left(-\frac{k}{2\sqrt{2\overline{L}}}\right) + 2(1+\sqrt{2\overline{L}})\,\delta_{\mathcal{R}\mathrm{OT}}.$$

Using eq. (50) at y_k,

$$\begin{aligned}\mathrm{Cert}_{\Gamma_t}(y_k, \widehat{\pi}) \le{}& 2\overline{L} C_0 \exp\left(-\frac{k}{2\sqrt{2\overline{L}}}\right) \\ &+ \left[4\overline{L}(1+\sqrt{2\overline{L}}) + 2\overline{L} + 1\right]\delta_{\mathcal{R}\mathrm{OT}}.\end{aligned}$$

Since $\overline{L} \ge 1$,

$$4\overline{L}(1+\sqrt{2\overline{L}}) + 2\overline{L} + 1 \le 13\overline{L}^{3/2}.$$

With eq. (94), the second term is therefore at most $\delta^2/8$. Taking $k = \widetilde{O}(\sqrt{\overline{L}})$ makes the first term at most $\delta^2/8$, proving the required certificate.

Finally, lemma 10 gives

$$N_{\mathrm{S}}(\zeta) = \widetilde{O}\left(\frac{\rho+\varepsilon}{\varepsilon}\right),$$

because its dependence on ζ is logarithmic. Thus the smaller UOT accuracy in eq. (94) affects only logarithmic factors. Multiplying the $\widetilde{O}(\sqrt{\overline{L}})$ UOT calls per outer iteration by T_δ and using $\overline{L} = O(\varepsilon^{-1})$ yields eq. (95). □