Simplifying Optimal Transport through Schatten-p Regularization

Anonymous authors
Paper under double-blind review

Abstract

We propose a new general framework for recovering low-rank structure in optimal transport using Schatten-p norm regularization. Our approach extends existing methods that promote sparse and interpretable transport maps or plans, while providing a unified and principled family of convex programs that encourage low-dimensional structure. The convexity of our formulation enables direct theoretical analysis: we derive optimality conditions and prove recovery guarantees for low-rank couplings and barycentric maps in simplified settings. To efficiently solve the proposed program, we develop a mirror-descent algorithm with convergence guarantees for $p \geq 1$. Experiments on synthetic and real data demonstrate the method's efficiency, scalability, and ability to recover low-rank transport structures.

1 Introduction

Optimal transport (OT) has emerged as a fundamental computational tool across many areas, including machine learning, computer vision, statistics, and biology (Arjovsky et al., 2017; Peyré and Cuturi, 2019; Schiebinger et al., 2019; Bonneel and Digne, 2023). It provides a principled framework for comparing probability distributions, and it has a rich mathematical history (Villani et al., 2008). While the combination of practical utility and deep mathematical theory has led to the broad adoption of OT ideas in mathematics, science, and engineering, finding ways to *scale* OT solutions and make them *interpretable* remains a fundamental research question (Cuturi et al., 2023; Khamis et al., 2024). In particular, OT typically suffers from the curse of dimensionality (Chewi et al., 2025), and regularized estimators may lack sparsity (Genevay et al., 2019).

A long line of work has focused on making OT scalable and interpretable through *regularization*. The most classical of these is entropic regularization, which yields a strictly convex program that can be solved via Sinkhorn scaling (Sinkhorn, 1967; Cuturi, 2013). More recent work has sought to increase efficiency and interpretability through quadratic regularization (Blondel et al., 2018; Lorenz et al., 2021), as well as low-rank factorizations (Forrow et al., 2019; Scetbon et al., 2021). These methods show promise in biological applications, particularly in single-cell RNA sequencing analysis (Klein et al., 2025).

Another closely related set of recent works attempts to include sparsity in the OT map using *elastic costs* Cuturi et al. (2023); Klein et al. (2024); Chen et al. (2025). In these works, using different cost modifications can be shown to encourage sparse or low-rank transport displacements. This leads to OT maps with simple, interpretable structures.

Except for entropic regularization, our work simultaneously generalizes all of the aforementioned methods in a unified framework. We believe that this unified picture can lead to more principled development of tailored regularization. Furthermore, the theory of OT has not yet fully leveraged the extensive literature on regularization for scaling and interpretability present in other fields, such as *compressed sensing*. In compressed sensing, the use of ℓ_1 or nuclear-norm penalties as proxies for rank minimization has yielded provably efficient algorithms (Eldar and Kutyniok, 2012; Wright and Ma, 2022). Our general formulation marries ideas from OT and compressed sensing, providing a bridge that we expect to be fruitful for developing sparse and low-rank optimal-transport models moving forward.

1.1 Contributions

In this work, we present Schatten-p regularized OT, which we call *Schatten OT*. This novel formulation is both general and amenable to direct theoretical analysis. We summarize the main contributions of our work:

- We demonstrate how the Schatten-OT program simultaneously generalizes a large portion of prior work on low-rank and sparse methods in OT, while also yielding new regularized formulations.
- We propose a general mirror-descent framework that efficiently solves the Schatten OT scale.
- For $p \geq 1$, the resulting optimization problem is convex, allowing convergence guarantees for mirror descent and analysis of low-rank couplings, low-rank transport displacements, and low-rank covariance structures.
- Experiments on synthetic and real data demonstrate the flexibility and effectiveness of Schatten OT.

1.2 Related Work

Regularized variants of OT have become increasingly important in current applied and theoretical research. The story of regularized OT begins with entropic regularization (Cuturi, 2013), which has roots in Schrödinger (1932). More recent regularizations include quadratic and sparse regularization (Blondel et al., 2018; Lorenz et al., 2021; González-Sanz and Nutz, 2024), which seek to encourage sparse structures in the transport plan.

Other work has studied low-rank factorizations in couplings to scale OT. Forrow et al. (2019) define a notion of factored couplings. Scetbon et al. (2021) use this notion of low-rank factorization of the coupling to develop an efficient Sinkhorn algorithm for factored couplings. Later, Lin et al. (2021) use multiple couplings to move through anchor points. Halmos et al. (2024) propose a new algorithm to optimize over the LC factorization, and Halmos et al. (2025) use hierarchical low-rank structures.

Another line of recent work has studied the regularization of displacements. Cuturi et al. (2023) introduce the notion of elastic OT costs and show how to construct maps with sparse or low-rank structure. Later, Klein et al. (2024) introduce learnable parameters into these costs, enabling greater flexibility in selecting the regularizer. Chen et al. (2025) use neural networks to learn maps in these settings.

We note that the incorporation of low-dimensional structure in OT displacements dates back to earlier subspace-robust notions of OT. Paty and Cuturi (2019) compute Wasserstein distances over worst-case subspaces in the ambient space. These methods have some practical statistical advantages (Niles-Weed and Rigollet, 2022).

We can broadly think of regularizing OT as encoding bias in the transport plan. However, there are many other ways the OT problem can be biased. For example, some works seek to encode biases by optimizing the ground cost used within OT. Alvarez-Melis et al. (2019) learn an OT with invariances using an alternating minimization procedure, and focus on optimization over Schatten-p balls. Sebbouh et al. (2024) learn a matrix $\boldsymbol{M}$ that defines an inner product cost between measures on different spaces. Jin et al. (2021) match distributions in different spaces using separate linear transformations. We note that these works implicitly regularize transport, as in subspace-robust OT.

In seeking a principled way to regularize and scale OT, we draw connections with *compressed sensing*. Compressed sensing focuses on recovering sparse structures from data. Original foundational works concentrate on recovering sparse vectors using ℓ_1 regularization (Donoho, 2006; Candes et al., 2006). These ideas were later extended to low-rank matrices (Fazel et al., 2008), which used nuclear norm regularization. The use of more general Schatten-p norms followed this (Nie et al., 2012). The extension of these regularizations to other settings has been fruitful. For example, Scarvelis and Solomon (2024) use it in the context of deep learning.

The ideas of compressed sensing are seeing a resurgence in the age of modern machine learning and AI. Sparse autoencoders have become a primary tool for practitioners studying mechanistic interpretability

(Huben et al., 2024). Sparse coding and rate reduction form a recent framework for training deep models to develop "white-box" methods (Yu et al., 2020; 2023). Compression as a general technique is effective at demonstrating intelligence in simple puzzles (Liao and Gu, 2025). These examples show the importance and practicality of developing theoretically principled compression techniques for machine learning and AI problems.

1.3 Notation

Bold lowercase letters are vectors and bold uppercase letters are matrices. We denote the set of integers $[n] := \{1, \dots, n\}$. For vectors, $\|\cdot\|$ is the standard ℓ_2 (Euclidean) norm. For matrices, $\|\cdot\|_{S_p}$ is the Schatten-p norm, i.e., the ℓ_p norm of the vector of singular values, and $\|\cdot\|_{S_2} = \|\cdot\|_F$ is the Frobenius norm. The set of probability measures over $\mathbb{R}^d$ with finite pth moment is $\mathcal{P}_p(\mathbb{R}^d)$, and the subset of absolutely continuous measures is $\mathcal{P}_{p,ac}(\mathbb{R}^d)$. The indicator function is $\mathbb{1}$.

1.4 Organization

First, in Section 2, we give the necessary background and outline our optimization program. Then, in Section 3, we provide our algorithmic framework for solving the Schatten OT problem. After this, Section 4 gives theoretical results about the structure of Schatten OT couplings. Finally, Section 5 presents experiments on synthetic and real data, highlighting the flexibility and advantages of our framework.

2 Background and Method

In this section, we first discuss background ideas in OT and compressed sensing and then the Schatten OT method. We begin in Section 2.1 by describing discrete OT and its common regularizations. Then, Section 2.2 discusses background on Schatten-p regularization in compressed sensing. After this, we define our Schatten-p norm regularized OT, Schatten OT, in Section 2.3.

2.1 OT and Regularization

For simplicity of presentation, we focus on the discrete case; in Appendix A, we show how these ideas extend to the continuous setting. Consider two discrete measures $\mu = \sum_{i=1}^m a_i \delta_{\boldsymbol{x}_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{\boldsymbol{y}_j}$, where $a_i, b_j \geq 0$ and $\sum_i a_i = \sum_j b_j = 1$. We let $\boldsymbol{X} \in \mathbb{R}^{d \times m}$ and $\boldsymbol{Y} \in \mathbb{R}^{d \times n}$ be matrices with the support points as columns. Without loss of generality, assume $n \geq m$. The transportation polytope $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$ is the set of $m \times n$ nonnegative matrices whose rows sum to $\boldsymbol{a} = [a_1, \dots, a_m]^\top$ and columns sum to $\boldsymbol{b} = [b_1, \dots, b_n]^\top$. We also refer to these matrices as couplings between μ and ν. In the transport problem, we are thinking of transporting μ to ν. Therefore, we refer to μ as the *source distribution* and ν as the *target distribution*.

We assume an $m \times n$ matrix of costs $\boldsymbol{C}$, where $\boldsymbol{C}_{ij} = c(\boldsymbol{x}_i, \boldsymbol{y}_j)$, for some function $c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$. The function c is typically called the *ground cost*. As a concrete example, we can use the pth power of the Euclidean distance,

$$\boldsymbol{C}_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^p. \tag{1}$$

OT seeks a minimum-cost coupling between the measures μ and ν. It is formulated as a linear program over the transportation polytope,

$$\mathsf{OT}(\mu, \nu) = \min_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, \boldsymbol{C} \rangle. \tag{2}$$

When the cost corresponds to a power of a metric on some underlying space, the resulting OT cost can be used to define a metric on $\mathcal{P}_p(\mathbb{R}^d)$. For the rest of this paper, we will assume that $c(\boldsymbol{x}_i, \boldsymbol{y}_j) = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2$. However, we note that our regularization can be applied to OT with any ground cost.

For a review of computational methods related to this linear program, see Peyré and Cuturi (2019). While there are many deep and interesting results related to OT, in practice, the direct use of OT in the form presented can suffer. In particular, OT suffers the curse of dimensionality; worst-case statistical rates of estimation for the p-Wasserstein distance are $O(n^{-1/d})$, assuming that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are i.i.d. samples from some

population measures. On the computational side, for large $n \asymp m$, the linear program incurs computational cost $O(n^3)$, and we must store an $O(n^2)$ variable in memory. To combat these issues, various regularizers have been considered, as we mentioned in the introduction. These regularized OT variants solve

$$\min_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \lambda R(\boldsymbol{P}),$$

where $R : \mathbb{R}^{m \times n} \to \mathbb{R}$ is the regularization function and λ is a tunable parameter. An example is the entropy function $R(\boldsymbol{P}) = \sum_{ij} \boldsymbol{P}_{ij}(\log(\boldsymbol{P}_{ij}) - 1)$ (Cuturi, 2013).

2.2 Schatten-p Regularization in Compressed Sensing

Schatten-p regularization in compressed sensing served as a way to generalize ℓ_p regularization for sparse vector recovery. In particular, using Schatten-p regularization is strictly more general than ℓ_p regularization because we can encode vectors as diagonal matrices, in which case $\|\boldsymbol{x}\|_p = \|\mathrm{diag}(\boldsymbol{x})\|_{S_p}$.

Perhaps the most popular regularization is Schatten-1 (nuclear norm) regularization. This is typically used to relax the rank of a matrix, and in a variety of settings, nuclear norm minimization has been shown to recover low-rank matrices (Recht et al., 2010; Candès and Tao, 2010; Candès et al., 2011).

These methods have been applied in a variety of settings. Some applications of these methods have included matrix completion and recommender systems (Nie et al., 2012), multitask learning (Zhang et al., 2018), high-dimensional covariance estimation (Gavish and Donoho, 2017), and image processing (Xie et al., 2016).

Optimization with nuclear norms, or more generally Schatten-p norms, involves a variety of algorithms. For example, nuclear norm minimization can involve saddle point or proximal algorithms Nesterov and Nemirovski (2013), the latter of which involves singular value thresholding (Cai et al., 2010). Other algorithmic paradigms include primal-dual methods (Chambolle and Pock, 2011) or ADMM (Yuan and Yang, 2009). To solve the more general case of Schatten-p regularized problems, for $0 < p \leq 2$, one typically resorts to Lagrangian-style methods (Nie et al., 2012) or iteratively reweighted nuclear norm-style methods (Lu et al., 2015). Another popular approach to nuclear norm minimization problems involves factor splitting to avoid SVD computations (Srebro et al., 2004; Fan et al., 2019).

2.3 Discrete Schatten-p Regularized OT

We now come to the main innovation of our work. We study a new variant of regularized OT problems using Schatten-p norms. We define the Schatten OT problem as

$$\textsf{Sch-OT}(\mu, \nu; \{(\lambda_i, p_i, q_i, \mathcal{A}_i)\}) := \min_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \sum_i \lambda_i \|\mathcal{A}_i(\boldsymbol{P})\|_{S_{p_i}}^{q_i}. \tag{3}$$

The idea of this program is to regularize toward simpler couplings with respect to the maps $\mathcal{A}_i$. Notice that the Schatten OT problem relies on three sets of parameters: the regularization strengths $\lambda_i \geq 0$, the Schatten powers and exponents $p_i, q_i > 0$, and maps $\mathcal{A}_i : \mathcal{U}(\boldsymbol{a}, \boldsymbol{b}) \to \mathbb{R}^{k_i \times l_i}$. Provided that $p_i, q_i \geq 1$ and $\mathcal{A}_i$ are affine, it is easily seen that Schatten OT is a convex program. With only one regularization term, this simplifies to $\textsf{Sch-OT}(\mu, \nu; (\lambda, p, q, \mathcal{A})) = \min_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \lambda \|\mathcal{A}(\boldsymbol{P})\|_{S_p}^q$.

We believe that convexity and generality are primary benefits of Schatten OT. It is general enough to cover many existing regularizations in the literature, as we will demonstrate shortly. The convexity of the problem leads to solutions that are easy to characterize, as we will show in our theory section. It also enables efficient solvers using convex optimization techniques.

While many past regularizations for OT fall into this framework, as we will demonstrate below, some do not. In particular, we cannot recover entropic regularization from (3). While entropic regularization cannot be realized by a Schatten norm of an affine function, one could add an entropic penalty to Schatten OT. This may be convenient from an algorithmic standpoint, but we leave the study of this additional regularization to future work. In Appendix A, we illustrate how to extend these ideas to the continuous setting.

Low-rank and Sparse Couplings: As a first example, consider the affine map $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{P}$. Then, depending on the choice of p, Schatten OT encourages low-rank or sparse couplings. In particular, choosing $q = p \leq 1$ encourages low-rank solutions. This yields a principled, optimization-based analog to the low-rank factorization pursued by works such as Forrow et al. (2019); Scetbon et al. (2021). On the other hand, $q = p = 2$ corresponds to the case of quadratically regularized OT (Blondel et al., 2018; Lorenz et al., 2021), since the Schatten-2 norm is just the Frobenius norm. This tends to encourage sparse solutions (González-Sanz and Nutz, 2024). Group sparsity (Blondel et al., 2018) can be achieved through proper choice of the affine maps in (3) and setting $p_i = 2$, $q_i = 1$.

Elastic costs: We can also recover some of the elastic cost regularizations of Cuturi et al. (2023); Klein et al. (2024); Chen et al. (2025). In particular, we can take $q = p = 1$ and let $\mathcal{A}(\boldsymbol{P})$ be the affine map

$$\boldsymbol{P} \mapsto \operatorname{diag}((\boldsymbol{P}_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j))_{i=1,j=1}^{m,n}.$$

Then, the Schatten OT penalty corresponds to the ℓ_1 elastic cost. Group-sparse elastic costs can be recovered from sums of Schatten regularizations. We can also recover the subspace elastic costs of Klein et al. (2024) by taking $q = p = 2$ and the affine map

$$\boldsymbol{P} \mapsto \operatorname{diag}((\boldsymbol{Q}_L \boldsymbol{P}_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j))_{i=1,j=1}^{n,m},$$

where $\boldsymbol{Q}_L$ is the projection onto the orthogonal complement of L. Note that, analogously to Klein et al. (2024), one could include an additional minimization over L in the Schatten OT formulation. This then defines a family of learnable Schatten OT problems. We discuss this possibility further in the appendix.

Barycentric projection maps and displacements: The formulation can be used to penalize map estimators directly. In particular, given a transport plan $\boldsymbol{P}$, one can estimate a transport map using the *barycentric projection*

$$T_{\boldsymbol{P}}(\boldsymbol{x}_i) = \frac{1}{a_i} \sum_{j=1}^{m} \boldsymbol{P}_{ij} \boldsymbol{y}_j = \frac{1}{\boldsymbol{a}_i} (\boldsymbol{Y}\boldsymbol{P}^\top)_{:,i}.$$

Notice that this map is linear in $\boldsymbol{P}$. Thus, we can penalize the barycentric projection map in our program by letting $\mathcal{A}(\boldsymbol{P}) = T_{\boldsymbol{P}}(\boldsymbol{X})\operatorname{diag}(\boldsymbol{A})^{-1/2}$, where we define $\boldsymbol{A} = \operatorname{diag}(\boldsymbol{a})$. As a further example, we could encourage displacements to be low-rank rather than the map itself. We call $T_{\boldsymbol{P}}(\boldsymbol{x}_i) - \boldsymbol{x}_i$ the *barycentric displacement*. Then, to encourage these to be simple, we could use $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{Y}\boldsymbol{P}^\top \boldsymbol{A}^{-1/2} - \boldsymbol{X}\boldsymbol{A}^{1/2}$. In both cases, the additional scaling of $\boldsymbol{A}^{1/2}$ allows the population limit of the program to be well defined.

Covariance regularization: All of the maps discussed so far include zeroth and first moments of the support points $(\boldsymbol{x}_i, \boldsymbol{y}_j)$. However, our formulation is flexible enough to include higher moments of our data distribution. For example, we can take $\mathcal{A}$ to be an affine function of the covariance induced by $\boldsymbol{P}$,

$$\boldsymbol{\Sigma}_{\boldsymbol{P}} = \sum_{ij} \boldsymbol{P}_{ij} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_j \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_i \\ \boldsymbol{y}_j \end{pmatrix}^\top,$$

which is linear in $\boldsymbol{P}$.

We could penalize the Schatten-1 norm of the cross-covariance $\sum_{ij} \boldsymbol{P}_{ij} \boldsymbol{x}_i \boldsymbol{y}_j^\top$, which is an affine function of $\boldsymbol{\Sigma}_{\boldsymbol{P}}$. If the vectors $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are whitened (i.e., they each have identity covariance), then the singular values of this matrix correspond to the canonical correlations. Minimizing the Schatten-1 norm in this case corresponds to minimizing the sum of canonical correlations, which seeks to increase independence between X and Y. Note that if we take the $\mathcal{A}(\boldsymbol{P}) = \sum_{ij} \boldsymbol{P}_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^\top$ and $p = 1$, then the regularization is just the quadratic OT cost, $\| \sum_{ij} \boldsymbol{P}_{ij}(\boldsymbol{x}_i - \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j)^\top \|_{S_1} = \sum_{ij} \boldsymbol{P}_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2$.

These illustrate just a few of the potential choices for adding covariance regularization to OT. In the appendix, we discuss covariance regularization in the context of Schatten OT for Gaussians. An in-depth study of these is left to future work.

3 A Mirror Descent Algorithm

The Schatten-OT program in (3) is convex whenever $p_i, q_i \geq 1$ and $\mathcal{A}_i$ are affine, but solving it directly with off-the-shelf convex solvers (e.g., CVXPY (Diamond and Boyd, 2016) or interior point methods) is only feasible for small problems, as the transportation polytope $\mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$ involves $\mathcal{O}(nm)$ variables and constraints. To address large-scale settings, we turn to first-order optimization methods. A particularly effective choice for optimization over the transport polytope is mirror descent with Kullback–Leibler (KL) geometry Kemertas et al. (2025). We use this algorithm for its simplicity and leave the analysis of more general methods, such as primal-dual algorithms or ADMM, to future work.

Following the approach of Kemertas et al. (2025), we develop mirror descent using the KL geometry on the transport polytope. This choice is natural, since using the KL geometry replaces a costly Euclidean projection with efficient Sinkhorn scaling. A few iterations of this method, followed by rounding, are effective at projecting to the polytope (Altschuler et al., 2017).

Assuming that $\mathcal{A}(\boldsymbol{P}) \neq 0$, we can compute a subgradient of the Schatten-p norm term in the Schatten OT problem as

$$q\|\mathcal{A}(\boldsymbol{P})\|_{S_p}^{q-p}\mathcal{A}^\star\left(\boldsymbol{U}\boldsymbol{\Sigma}^{p-1}\boldsymbol{V}^\top\right) \in \partial\|\mathcal{A}(\boldsymbol{P})\|_{S_p}^{q},$$

where ∂ denotes the subdifferential. This provides a computable subgradient of $F(\boldsymbol{P})$ at each iteration, provided that we can compute a singular value decomposition. For $p, q > 1$, this is a bona fide gradient.

The mirror descent iteration involves the following steps.

1. Form the SVD $\mathcal{A}(\boldsymbol{P}^k) = \boldsymbol{U}^k\boldsymbol{\Sigma}^k\boldsymbol{V}^{k\top}$ and the subgradient

$$\boldsymbol{G}^k = q\|\mathcal{A}(\boldsymbol{P}^k)\|_{S_p}^{q-p}\mathcal{A}^\star\big(\boldsymbol{U}^k(\boldsymbol{\Sigma}^k)^{p-1}\boldsymbol{V}^{k\top}\big)$$

2. Use multiplicative-weights form

$$\widehat{\boldsymbol{P}}_{ij} \propto \boldsymbol{P}_{ij}^k \exp(-\tau\boldsymbol{G}_{ij}^k)$$

3. Project back to the transport polytope

$$\boldsymbol{P}^{k+1} = \Pi_{\mathcal{U}(\boldsymbol{a},\boldsymbol{b})}^{\mathrm{KL}}(\widehat{\boldsymbol{P}}),$$

Here, $\tau^k > 0$ is a step size and $\Pi_{\mathcal{U}(\boldsymbol{a},\boldsymbol{b})}^{\mathrm{KL}}$ is the projection onto the transport polytope with respect to the KL divergence, which can be implemented via Sinkhorn scaling.

By standard mirror descent theory (Beck and Teboulle, 2003; Nemirovsky and Yudin, 1983; Bubeck et al., 2015), the method achieves an $O(1/\sqrt{T})$ convergence rate for convex objectives ($p, q \geq 1$). The KL geometry ensures that nonnegativity is automatically preserved, and averaging can be used to guarantee convergence of the objective values.

The most expensive parts of the iteration are the SVD computation and the Sinkhorn projection. It is possible to use an adaptive low-rank approximation of $\mathcal{A}(\boldsymbol{P}^k)$ throughout the iterations to increase computational efficiency. It would also be interesting to attempt to use sketching methods to approximate low-rank solutions to this problem (Yurtsever et al., 2021).

The choice of step size is essential. In general, since the problem is convex but not smooth in general, one could take $\tau^k \propto 1/\sqrt{k}$, which yields the $O(1/\sqrt{T})$ convergence rate in objective value. In our experiments, we can observe faster convergence in specific settings. For example, when $p = q = 1$ and $\mathcal{A}$ has simple structure, for instance when $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{P}$ (low-rank couplings) or $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{Y}\boldsymbol{P}^\top\boldsymbol{A}^{-1}$ (low-rank barycentric maps), mirror descent can obtain faster convergence with a geometrically diminishing step size, the same schedule used in past work with sharp minima (Davis et al., 2018; Maunu et al., 2019).

4 Theory

In this section, we present our main theoretical results on the Schatten OT program. First, Section 4.1 uses convex optimization theory to outline the structure of solutions to the Schatten OT problem. After this, Section 4.2 uses this structure to prove two theorems that demonstrate Schatten OT's ability to recover low-rank couplings and barycentric displacements. Finally, in Section 4.3, we finish with a discussion of our theoretical results.

4.1 General Structural Theorems

Let $\boldsymbol{P}^\star$ be an optimal solution of the optimization problem (3) with $p, q \geq 1$. Since this is a constrained convex optimization problem in $\boldsymbol{P}$, we can appeal to standard theory. The solution is characterized by the KKT conditions, which state that there exists $\boldsymbol{G}^\star \in \partial \|\mathcal{A}(\boldsymbol{P}^\star)\|_{S_p}^q$ such that

$$\begin{aligned} &\boldsymbol{C} + \lambda \boldsymbol{G}^\star + \mathbf{1}_n \boldsymbol{u}^\top + \boldsymbol{v}\mathbf{1}_m^\top = \mathbf{0}, \\ &\boldsymbol{P}^\star \geq 0, \quad \boldsymbol{P}^\star \mathbf{1} = \boldsymbol{a}, \quad \boldsymbol{P}^{\star\top}\mathbf{1} = \boldsymbol{b}. \end{aligned}$$

Comparing these conditions to the optimality conditions for standard OT, we notice that the only difference is the inclusion of the $\lambda \boldsymbol{G}^\star$ in the first-order stationarity condition. Therefore, the optimality conditions for Schatten OT are precisely those for an OT problem with the *tilted cost*

$$\boldsymbol{S}(\lambda, \boldsymbol{G}^\star) := \boldsymbol{C} + \lambda \boldsymbol{G}^\star \in \mathbb{R}^{n \times m}. \tag{4}$$

where $\boldsymbol{G}^\star$ is some subgradient of $\|\mathcal{A}(\cdot)\|_{S_p}^q$ at $\boldsymbol{P}^\star$. The obstacle to our directly applying this result is that we do not know $\boldsymbol{G}^\star$, since that would require knowing $\boldsymbol{P}^\star$.

We can state this characterization as the following proposition.

Proposition 1. *The coupling $\boldsymbol{P}^\star$ is optimal for (3) if and only if there exists a subgradient $\boldsymbol{G}^\star$ of $\|\mathcal{A}(\boldsymbol{P}^\star)\|_{S_p}^q$ such that*

$$\boldsymbol{P}^\star \in \operatorname{argmin}_{\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{S}(\lambda, \boldsymbol{G}^\star), \boldsymbol{P} \rangle.$$

While we cannot apply this to the direct computation of $\boldsymbol{P}^\star$, we can use it to characterize solutions to the Schatten OT problem. In the following section, we develop this idea to prove the recovery of low-rank structure in the Schatten OT problem.

4.2 Discrete Recovery Theorems

In this section, we prove low-rank recovery theorems for Schatten OT. While these are restrictive toy examples, they represent the first such exact recovery results in the OT literature. We believe this is a first step towards applying compressed sensing ideas to regularized OT problems. It is an open question for future work to extend these ideas to the recovery of simple couplings in more complex settings. We begin in Section 4.2.1 with a recovery result for low-rank couplings. Then, Section 4.2.2 presents a consequence on the recovery of a low-rank set of barycentric displacements.

4.2.1 Low-Rank Coupling Recovery

We now assume that both the source and the target consist of R well-separated clusters, each with the same cardinality. We show that, for a nontrivial interval of regularization strengths λ, the nuclear-norm penalized OT problem recovers a rank-R, block-diagonal coupling that matches each source cluster uniformly to its corresponding target cluster. We assume uniform marginals $a_i = 1/m, b_j = 1/n$ for all i, j.

Our first assumption is on the clustered structure of μ and ν.

Assumption 2. *For two measures $\mu = \sum_{i=1}^m a_i \delta_{\boldsymbol{x}_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{\boldsymbol{y}_j}$, $n = Rg$, $m = Rg$ for integers $R, g \geq 1$. The source indices $[m]$ and target indices $[n]$ are partitioned into clusters $S_1, \ldots, S_R$ and $T_1, \ldots, T_R$, respectively, where $|S_t| = |T_t| = g$ for all t.*

Let $B(\boldsymbol{z}, \rho)$ denotes the closed Euclidean ball of radius $\rho > 0$ around $\boldsymbol{z} \in \mathbb{R}^d$. Our goal is to construct a setting where the cluster S_t is uniformly matched to T_t. We make the following assumptions about the locations of the source and target points.

Assumption 3. *For two measures $\mu = \sum_{i=1}^m a_i \delta_{\boldsymbol{x}_i}$ and $\nu = \sum_{j=1}^n b_j \delta_{\boldsymbol{y}_j}$,*

1. *The source points lie in disjoint balls, $\boldsymbol{x}_i \in B(\boldsymbol{c}_t, \rho)$ for $i \in S_t$, and the target points lie in disjoint balls, $\boldsymbol{y}_j \in B(\boldsymbol{d}_t, \rho)$ for $j \in T_t$.*
2. *Within matched clusters S_t and T_t, $\|\boldsymbol{x}_i - \boldsymbol{y}_j\| = \|\boldsymbol{x}_i - \boldsymbol{y}_{j'}\|$ for all $i \in S_t$ and $j, j' \in T_t$ for $t = 1, \ldots, R$.*
3. *The minimum inter-cluster distance $\Gamma := \min_{s \neq t} \|\boldsymbol{c}_t - \boldsymbol{d}_s\|$, and the maximum intra-cluster distance as $\gamma := \max_t \|\boldsymbol{c}_t - \boldsymbol{d}_t\|$ satisfy*

$$\Gamma \;>\; \gamma + 4\rho > 0. \tag{5}$$

Notice that, under our separation condition (5), the OT coupling actually respects the cluster structure, in the sense that it must match points in S_t to T_t. Furthermore, any plan that matches $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$ within clusters (when $i \in S_t$, $j \in T_t$) is optimal. However, these matched clusterings are not low-rank; they are full-rank. On the other hand, as we will show in the following theorem, a low-rank matching can be recovered from Schatten OT.

Theorem 4. *Let Assumptions 2 and 3 hold, and let the excess cost for an across-cluster matching be*

$$\Delta_{\min} := \min_{\substack{s \neq t \in [R] \\ i \in S_t, j \in T_s, j' \in T_t}} \Big\{ \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{y}_{j'}\|_2^2 \Big\}.$$

Then, for any regularization parameter λ satisfying

$$0 \;\leq\; \lambda \;<\; g \cdot \Delta_{\min}, \tag{6}$$

the minimizer of $\langle \boldsymbol{C}, \boldsymbol{P} \rangle + \lambda \|\boldsymbol{P}\|_{S_1}$ is a rank R coupling supported blockwise on $\bigcup_{t=1}^R (S_t \times T_t)$ that is uniform within clusters.

The essential idea of the proof is to ensure that 1) $\boldsymbol{P}^\star$ is the unique coupling that respects the cluster structure and also minimizes the nuclear norm, and 2) there is no way to make the nuclear norm term even smaller by using across-cluster matches without incurring more cost.

4.2.2 Low-Rank Displacement Recovery when $p = 1$

We now give a concrete example of how to recover a coupling with low-rank displacements. For our affine map, we consider the weighted barycentric displacement matrix $\mathcal{A}(\boldsymbol{P})$,

$$\mathcal{A}(\boldsymbol{P}) \;:=\; \big(T_{\boldsymbol{P}}(\boldsymbol{X}) - \boldsymbol{X}\big)\boldsymbol{A}^{1/2} \;= \boldsymbol{Y}\boldsymbol{P}^\top \boldsymbol{A}^{-1/2} - \boldsymbol{X}\boldsymbol{A}^{1/2}.$$

We also again assume that $p = q = 1$ in the Schatten OT formulation. For our recovery result, we make the following assumptions.

Assumption 5 (Symmetric two-target clusters with separation)**.** *Fix orthonormal vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and an integer $R \geq 2$. Let $0 < \mu_1 < \cdots < \mu_R$ be distinct scalars and put $\boldsymbol{m}_t := \mu_t \boldsymbol{u} \in \mathbb{R}^d$ for $t \in [R]$. Suppose:*

1. *The source set $[m]$ is partitioned into nonempty clusters $S_1, \ldots, S_R$ and there exists $\rho > 0$ such that $\boldsymbol{x}_i \in \boldsymbol{m}_t + [-\rho, \rho]\boldsymbol{u}$ for all $i \in S_t$.*
2. *For some $\varepsilon > 0$, the target support consists of the $2R$ points*

$$\boldsymbol{y}_{t,+} = \boldsymbol{m}_t + \varepsilon \boldsymbol{v}, \;\; \boldsymbol{y}_{t,-} = \boldsymbol{m}_t - \varepsilon \boldsymbol{v}, t \in [R],$$

with target masses $b_{t,+} = b_{t,-} = \frac{1}{2} \sum_{i \in S_t} a_i$.

3. The minimal separation between clusters is lower bounded

$$\Lambda \;:=\; \min_{s\neq t} |\mu_t - \mu_s| \;>\; 2\rho.$$

By symmetry, it is easy to see that all couplings that assign the source points in cluster S_t to $\boldsymbol{y}_{t,+}$ or $\boldsymbol{y}_{t,-}$ are optimal. In particular, it does not matter how the points are assigned within clusters. For $i \in S_t$ and $s \neq t$, it is also convenient to define the inter-cluster cost gap

$$\Delta_{i,s} \;:=\; \|\boldsymbol{x}_i - \boldsymbol{y}_{s,\pm}\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{y}_{t,\pm}\|_2^2 \;=\; (\mu_t - \mu_s)^2 + 2(\mu_t - \mu_s)\xi_i, \tag{7}$$

where $\boldsymbol{x}_i = \boldsymbol{m}_t + \xi_i \boldsymbol{u}$ and $|\xi_i| \leq \rho$.

We now state the main recovery theorem for this setting. It says that, under our assumption, we can exactly recover a coupling with a rank-1 barycentric displacement.

Theorem 6. *Under Assumption 5 and the quadratic cost $\boldsymbol{C}_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2$. Define the explicit threshold*

$$\lambda_{\max} \;:=\; \Lambda - 2\rho \;>\; 0.$$

Then for every $\lambda \in [0, \lambda_{\max})$, the unique minimizer of (3) matches $\boldsymbol{x}_i$, for $i \in S_t$, to $\{\boldsymbol{y}_{t,\pm}\}$. Furthermore, it yields a rank-1 barycentric map.

4.3 Discussion

While we demonstrate low-rank recovery only in toy examples here, our methodology highlights the advantages of using convex formulations. In particular, it is easier to verify that the recovered solution is low-rank. In fact, these are the first guarantees of low-rank recovery within an OT problem in the literature. This is compared to nonconvex methods, which currently lack guarantees (Forrow et al., 2019; Klein et al., 2024).

In the future, it would be interesting to demonstrate exact recovery in more general settings using Schatten-p regularization when $p < 1$. It would also be interesting to develop more general recovery conditions to move beyond the toy examples considered here.

5 Experiments

In this section, we give some simulations on real data that demonstrate the advantages of the Schatten OT formulation. First, in Section 5.1, we demonstrate Schatten OT's ability to recover low-rank couplings and barycentric projection maps. Then, in Section 5.2, we examine the convergence rate of mirror descent to solve the Schatten OT problem. Finally, Section 5.3 gives an experiment on real data that demonstrates the ability of Schatten OT to recover simpler couplings with 4i perturbation data.

5.1 Low-rank Recovery

To first examine properties of the Schatten OT problem, we use CVXPY to solve the convex program exactly.

In our first experiment, we examine the Schatten OT's ability to recover low-rank couplings. To measure the quality of the recovered $\boldsymbol{P}^\star$, we use two metrics: effective rank and transport cost. The latter is defined as $\langle \boldsymbol{C}, \boldsymbol{P}^\star \rangle$. The former is the ratio of the nuclear norm to the operator norm:

$$\text{Effective Rank}(\boldsymbol{B}) = \frac{\|\boldsymbol{B}\|_{S_1}}{\|\boldsymbol{B}\|_{S_\infty}}.$$

In our experiments, the support of μ consists of two clusters centered at $(-2, 2)$ and $(-2, -2)$, and ν consists of two clusters centered at $(2, 2)$, and $(2, -2)$. The data within each cluster is Gaussian. We sample 10 points from each cluster, so that $n = m = 20$. The results are averaged over five randomly generated datasets.

In the first experiment displayed in the top row of Figure 1, we use a variance of 0.04 for each Gaussian component. In the left image, we show the effective rank versus regularization strength λ for $p = 1, 2,$

and $infty$. On the right, we display the transport cost. As we can see, nuclear norm regularization can significantly reduce the effective rank without substantially increasing the transport cost. Using the Schatten-2 norm can also reduce the effective rank, albeit more gradually.

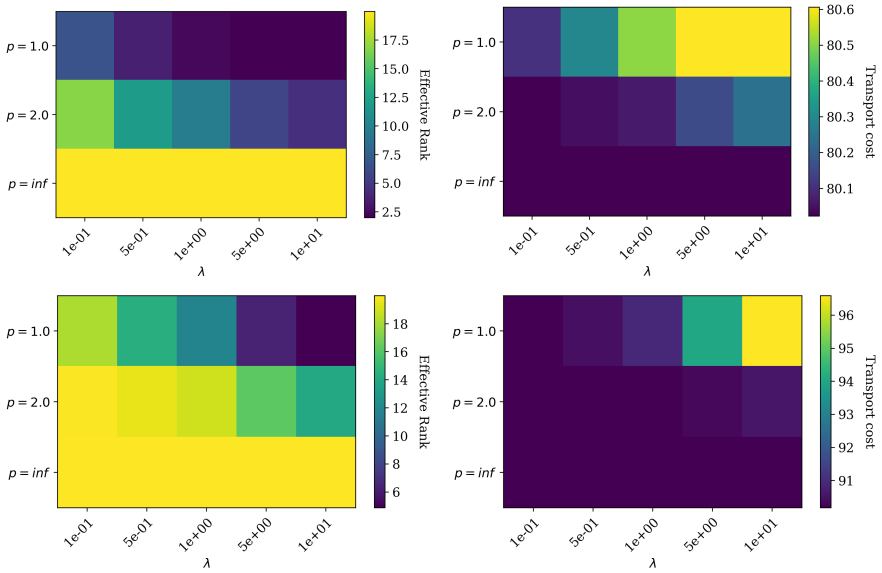

Figure 1: Solution quality of Schatten OT versus regularization parameter for mixture of Gaussian data. Top row: small variance. Bottom row: large variance. On the left, we show the effective rank of the found solution; on the right, we display its transport cost. As we can see, Schatten-1 regularization can greatly simplify the transport plan without substantially increasing transport costs.

We compare this experiment with a slight modification in the bottom row of Figure 1. Here, the data model is the same, except now the within-cluster variance is 2. As we can see, it is more challenging to find a low-rank transport plan, and when one is found, it increases the transport cost more substantially.

We include one more experiment in which we now wish to recover a low-rank displacement. We assume that the support of μ is standard Gaussian, and the support of ν is $\boldsymbol{y}_i = \boldsymbol{x}_i + \xi_i \boldsymbol{u}$, where $\boldsymbol{u}$ is a random unit vector and ξ_i is standard Gaussian. We now wish to recover a coupling with a low-rank barycentric map, and do the same experiment with a different affine map of $\boldsymbol{P}$. Figure 2 displays the results of this experiment. As we can see, Schatten-1 regularization again recovers a coupling with lower-rank displacements. However, the transport costs increase more substantially across the board with higher regularization strengths.

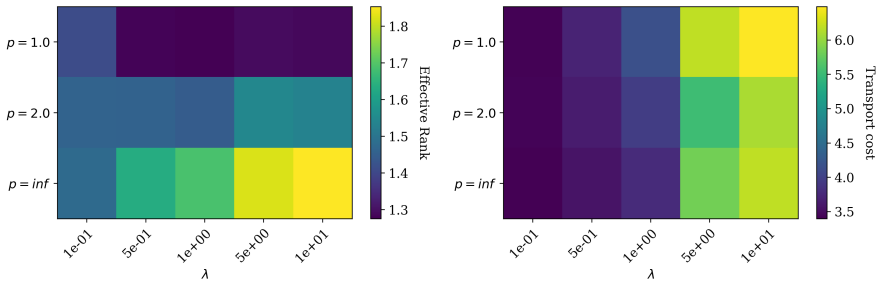

Figure 2: Solution quality of Schatten OT versus regularization parameter for Gaussian data with a low-rank perturbation. In the left display, we show the effective rank of the found barycentric displacements, and in the right display, we show the transport cost of the found coupling. As we can see, Schatten-1 regularization again simplifies the transport plan, but the transport cost increases substantially across the board.

5.2 Convergence Rates

In this section, we examine the convergence rate of the mirror descent algorithm with Schatten-1 regularization in two settings. In the first setting, we show sublinear convergence; in the second, linear convergence.

In the left display of Figure 3, we use a setting where we do not expect a low-rank coupling to be easy to find. We set $\lambda = .1$, and the data μ is a mixture of Gaussians with centers at $(-2, \pm 2)$ and ν is a mixture of Gaussians with centers at $(2, \pm 2)$. The variance is set to be 1, and $n = m = 20$. We observe slow sublinear convergence. Note the log scale on the x-axis.

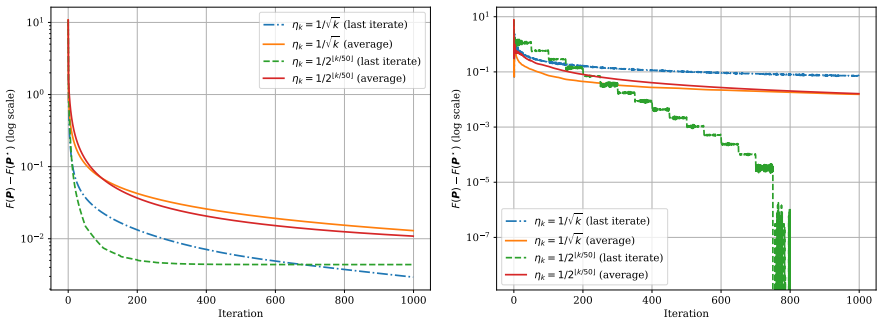

Figure 3: Plot of log excess cost versus iteration for the mirror descent algorithm on the Schatten OT problem. Left: In this experiment, the regularization parameter is small, and the variance of the Gaussian mixture components is large. This shows sublinear convergence of the algorithm, as is expected by the theory. Right: we reduce the variance and increase the regularization parameter, resulting in a low-rank optimal coupling. Here, we see that the geometrically diminishing step size converges linearly.

In the right display of Figure 3, we use the same setup as before, except now we set $\lambda = 10$ and let the clusters have variance 0.04. Now, since the recovered coupling is low-rank, mirror descent with a geometrically diminishing step size converges linearly. This implies that the objective is sufficiently sharp, which can be exploited by this step-size schedule.

5.3 4i Perturbation Example

In the experiment of Chen et al. (2025), the authors fit a displacement-sparse neural OT to 4i perturbation data. In the experiment, we see that dimensionality is reduced, but the error is higher, and the method has high variance because it requires fitting an input convex neural network (ICNN).

In the following, we show how Schatten OT can reduce the effective rank of couplings and barycentric projection maps. We use two perturbations within the CellOT data of Bunne et al. (2023). In particular, we follow Chen et al. (2025) and consider learning regularized couplings from the 4i perturbation data. The processed data is publicly available[1]. More details on our algorithmic setup for this experiment are given in Appendix D.

In Figure 4, we plot the effective rank against λ for two different affine maps $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{P}$ and $\mathcal{A}(\boldsymbol{P}) = \boldsymbol{Y}\boldsymbol{P}^\top \boldsymbol{A}^{-1/2}$. The color indicates the increase in transport cost relative to Sinkhorn with a regularization parameter of 1. We display the result for two different perturbations. For each, we average over five random subsamples of size 1000 from the control and perturbation distributions.

On the top row, we display the result for low-rank coupling recovery. As we can see, we can drastically reduce the complexity of transport plans without paying much more in transport costs. Next, in the bottom row of Figure 4, we repeat the previous experiment, focusing on recovering a low-rank barycentric projection map. We see again that Schatten-1 regularization can reduce the effective rank, although now the transport cost increases more substantially.

[1] `https://www.research-collection.ethz.ch/handle/20.500.11850/609681`

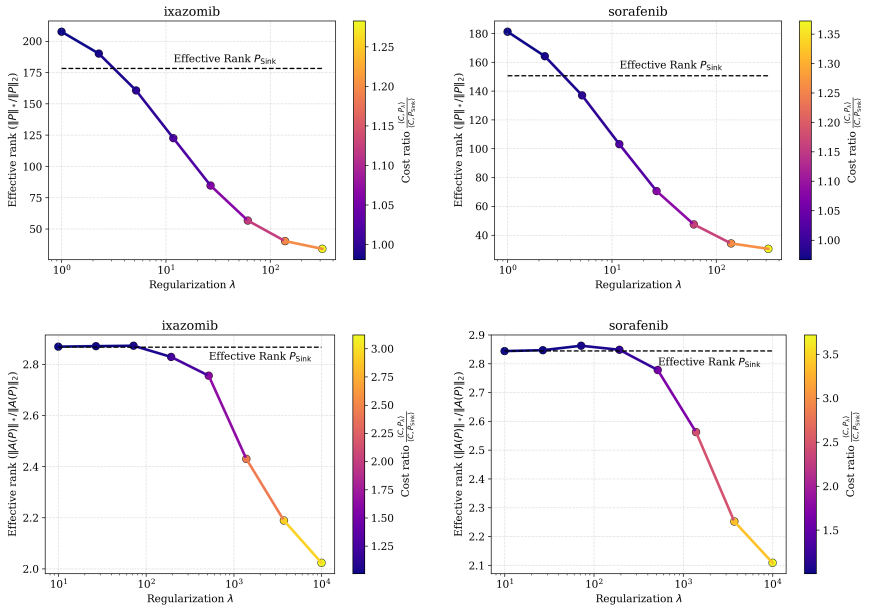

Figure 4: Plots of the performance of Schatten OT on the 4i perturbation data of Bunne et al. (2023). For reference, we compare in all plots with what one gets using a Sinkhorn coupling with a regularization parameter of 1. Top: We examine the performance of Schatten-1 regularization in recovering a low-rank coupling for two different perturbations. As we can see, Schatten OT can reduce the effective rank while not increasing the transport cost too much. Bottom: We now show the performance of Schatten-1 regularization in recovering a low-rank barycentric projection map for two different perturbations. Again, Schatten OT can reduce the effective rank of this map, though the transport cost now increases more.

6 Conclusion

We introduced Schatten-p regularized OT (Schatten OT), a unified convex framework for incorporating low-dimensional structure into OT problems. A key advantage of our formulation lies in its convexity and generality. Convexity allows us, for the first time, to provide provable recovery results in illustrative yet straightforward examples. Generality allows us to penalize any affine function of the coupling, thereby simultaneously encompassing many existing OT regularizations and enabling new ones.

Theoretically, we established the first recovery guarantees for low-rank couplings and low-rank barycentric displacements, bridging ideas from compressed sensing and OT theory. Algorithmically, we developed an efficient mirror-descent method to solve these regularized problems in practice. Empirically, this approach performs well and demonstrates practical utility on 4i cell-perturbation data. Our results show that Schatten OT recovers low-rank structure with only modest increases in transport cost, yielding simpler and more interpretable transport maps.

We believe this work paves the way for more interpretable and scalable OT methods. In particular, the Schatten OT framework may provide a foundation for connecting OT to broader advances in sparse modeling, compressed sensing, and interpretable machine learning.

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, Stefanie Jegelka, and Tommi S Jaakkola. Towards optimal transport with global invariances. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1870–

1879. PMLR, 2019.

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

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

Nicolas Bonneel and Julie Digne. A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library, 2023.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

Peter Chen, Yue Xie, and Qingpeng Zhang. Displacement-sparse neural optimal transport. *arXiv preprint arXiv:2502.01889*, 2025.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*. École d'Été de Probabilités de Saint-Flour. Springer, 2025.

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

Marco Cuturi, Michal Klein, and Pierre Ablin. Monge, bregman and occam: interpretable optimal transport in high-dimensions with feature-sparse maps. In *Proceedings of the 40th International Conference on Machine Learning*, pages 6671–6682, 2023.

Damek Davis, Dmitriy Drusvyatskiy, Kellie J MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Yonina C Eldar and Gitta Kutyniok. *Compressed sensing: theory and applications*. Cambridge university press, 2012.

Jicong Fan, Lijun Ding, Yudong Chen, and Madeleine Udell. Factor group-sparse regularization for efficient low-rank matrix recovery. *Advances in neural information processing Systems*, 32, 2019.

Maryam Fazel, Emmanuel Candes, Ben Recht, and Pablo Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047. IEEE, 2008.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

Matan Gavish and David L Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

Alberto González-Sanz and Marcel Nutz. Sparsity of quadratically regularized optimal transport: Scalar case. *arXiv preprint arXiv:2410.03353*, 2024.

Peter Halmos, Xinhao Liu, Julian Gold, and Benjamin Raphael. Low-rank optimal transport through factor relaxation with latent coupling. *Advances in Neural Information Processing Systems*, 37:114374–114433, 2024.

Peter Halmos, Julian Gold, Xinhao Liu, and Benjamin J Raphael. Hierarchical refinement: Optimal transport to infinity and beyond. *arXiv preprint arXiv:2503.03025*, 2025.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=F76bwRSLeK`.

Kun Jin, Chaoyue Liu, and Cathy Xia. Two-sided wasserstein procrustes analysis. In *IJCAI*, pages 3515–3521, 2021.

Mete Kemertas, Allan Douglas Jepson, and Amir-massoud Farahmand. Efficient and accurate optimal transport with mirror descent and conjugate gradients. *Transactions on Machine Learning Research*, 2025.

Abdelwahed Khamis, Russell Tsuchida, Mohamed Tarek, Vivien Rolland, and Lars Petersson. Scalable optimal transport methods in machine learning: A contemporary survey. *IEEE transactions on pattern analysis and machine intelligence*, 2024.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025.

Michal Klein, Aram-Alexandre Pooladian, Pierre Ablin, Eugène Ndiaye, Jonathan Niles-Weed, and Marco Cuturi. Learning elastic costs to shape monge displacements. *Advances in Neural Information Processing Systems*, 37:108542–108565, 2024.

Isaac Liao and Albert Gu. Arc-agi without pretraining, 2025. URL `https://iliao2345.github.io/blog_posts/arc_agi_without_pretraining/arc_agi_without_pretraining.html`.

Chi-Heng Lin, Mehdi Azabou, and Eva L Dyer. Making transport more robust and interpretable by moving data through a small number of anchor points. *Proceedings of machine learning research*, 139:6631, 2021.

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

Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2):829–839, 2015.

Tyler Maunu, Teng Zhang, and Gilad Lerman. A well-tempered landscape for non-convex robust subspace recovery. *Journal of Machine Learning Research*, 20(37):1–59, 2019.

Arkadij Semenovič Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for l1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013.

Feiping Nie, Heng Huang, and Chris Ding. Low-rank matrix recovery via efficient schatten p-norm minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 655–661, 2012.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International conference on machine learning*, pages 5072–5081. PMLR, 2019.

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

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Christopher Scarvelis and Justin M Solomon. Nuclear norm regularization for deep learning. *Advances in Neural Information Processing Systems*, 37:116223–116253, 2024.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.

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

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum-margin matrix factorization. *Advances in neural information processing systems*, 17, 2004.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

Yuan Xie, Shuhang Gu, Yan Liu, Wangmeng Zuo, Wensheng Zhang, and Lei Zhang. Weighted schatten p-norm minimization for image denoising and background subtraction. *IEEE transactions on image processing*, 25(10):4842–4857, 2016.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in neural information processing systems*, 33:9422–9434, 2020.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *Advances in Neural Information Processing Systems*, 36:9422–9457, 2023.

Xiaoming Yuan and Junfeng Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *preprint*, 12(2), 2009.

Alp Yurtsever, Joel A Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021.

Yu Zhang, Ying Wei, and Qiang Yang. Learning to multitask. *Advances in Neural Information Processing Systems*, 31, 2018.

A Extension to the Continuous Setting

Up until now, we have focused our attention on formulations in the discrete case. However, there is a direct extension of Schatten OT to the continuous setting by taking Schatten-p norms of appropriate linear operators over general Hilbert spaces. In this section, we let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ be general measures. The set of couplings between these measures is $\Pi(\mu, \nu)$.

To define our extension to the continuous case, we let $\mathcal{A} : \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \to \mathcal{B}(\mathcal{H})$ be a map from the space of couplings to the set of bounded linear operators on some Hilbert space $\mathcal{H}$. Let $\|\cdot\|_{S_p}$ now denote the Schatten-p norm over $\mathcal{B}(\mathcal{H})$, which is defined as $\|T\|_{S_p}^p = \mathrm{Tr}[(T^*T)^{p/2}]$. Then, we define the continuous Schatten OT problem

$$\textsf{Sch-OT}_p(\mu, \nu; (\lambda, p, q, \mathcal{A})) := \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\|X - Y\|^2 + \lambda \|\mathcal{A}(\pi)\|_{S_p}^q. \tag{8}$$

We note that, as in the discrete case, this notion depends on choices of λ, p, q, and $\mathcal{A}$. As before, choosing $p, q \geq 1$ and $\mathcal{A}$ an affine map makes the problem (8) convex.

Below, we give some examples of affine maps that extend our discrete examples. Throughout, we let $\rho = \mu \otimes \nu$ be the reference product measure.

Covariance regularization The most direct connection between the continuous and discrete cases is to penalize moments of the distribution. In the continuous case, this corresponds to regularizing the covariance of π. In this case, all of the regularizations discussed in Section 2.3 are the same except we regularize the linear operator over $\mathbb{R}^d$, $\boldsymbol{\Sigma}_\pi = \mathbb{E}_\pi \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}^\top$.

Continuous sparse and low-rank regularization: We now discuss the analogs of quadratic and low-rank regularization, the coupling matrix $\boldsymbol{P}$. Here, we can take $\mathcal{A}(\pi) = S_\pi : L^2(\nu) \to L^2(\mu)$ as the linear operator

$$(S_\pi f)(\boldsymbol{x}) = \int f(y) \frac{d\pi}{d\rho} d\nu(y) = \mathbb{E}_\pi[f(Y)|X = \boldsymbol{x}].$$

Note that T_π is affine in π. Then, regularizing the Schatten-p norm of S_π corresponds to using the Schatten-p norm of $\boldsymbol{P}$ as discussed earlier. The continuous quadratic case, where the continuous case has already been studied Lorenz et al. (2021), but not as a Schatten-2 norm of the operator S_π.

Elastic costs: We can recover the elastic costs by taking the Schatten norm of a specifically constructed operator. Choose a measurable partition of $\mathbb{R}^d \times \mathbb{R}^d$ given by $(E_k)_{k\in\mathbb{N}}$ with $\rho(E_k) > 0$. Define an orthonormal family in $L^2(\rho)$ by

$$\phi_k(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbb{1}((\boldsymbol{x}, \boldsymbol{y}) \in E_k)}{\sqrt{\rho(E_k)}}$$

Set the diagonal weights to be $s_k(\pi) = \int_{E_k} \|\boldsymbol{y} - \boldsymbol{x}\|_1 d\pi$, which are linear in π. Then, letting $(e_k)_{k\in\mathbb{N}}$ be a basis for ℓ_2, we can define the linear operator $\mathcal{A}(\pi) : \ell^2 \to L^2(\rho)$ by

$$A(\pi)e_k = s_k(\phi)\phi_k.$$

Notice that this is again linear in π, and furthermore $A(\pi)e_k$ are orthogonal with $\|A(\pi)e_k\| = s_n(\pi)$. Therefore, the singular values of $\mathcal{A}(\pi)$ are $s_n(\pi)$. Thus

$$\|\mathcal{A}(\pi)\|_{S_1} = \sum_k s_k(\pi) = \int \|\boldsymbol{y} - \boldsymbol{x}\|_1 d\pi.$$

A similar construction yields the subspace elastic costs discussed in Section 2.3. The general principle here is that elastic OT problems can be embedded as Schatten OT regularized problems over appropriate operators.

Barycentric projection maps and displacements: We can also consider Schatten-p regularization of the barycentric projection maps and displacements. In the continuous case, the barycentric projection map is $T_\pi(\cdot) = \mathbb{E}_{\pi(Y|\cdot)}Y$. Let the displacement map be $D_\pi(\cdot) = \mathbb{E}_{\pi(Y|\cdot)}Y - \cdot$. Then, we can formulate a Schatten-p penalization of this barycentric displacement map by viewing T_π or D_π as operators $T_\pi, D_\pi : \mathbb{R}^d \to L^2(\mu)$ given by $T_\pi v = \langle v, T_\pi(\cdot)\rangle$ and $D_\pi v = \langle v, D_\pi(\cdot)\rangle$.

B Supplementary Proofs

B.1 Proof of Theorem 4

Proof. Define, for each t, the cluster indicator mass vectors

$$\boldsymbol{\alpha}^{(t)} \in \mathbb{R}^n, (\boldsymbol{\alpha}^{(t)})_i = \begin{cases} \frac{1}{Rg}, & i \in S_t, \\ 0, & \text{otherwise,} \end{cases}$$

and similarly

$$\boldsymbol{\beta}^{(t)} \in \mathbb{R}^m, (\boldsymbol{\beta}^{(t)})_j = \begin{cases} \frac{1}{Rg}, & j \in T_t, \\ 0, & \text{otherwise.} \end{cases}$$

The rank R coupling we wish to recover is

$$\boldsymbol{P}^\star := \sum_{t=1}^{R} \boldsymbol{\alpha}^{(t)} {\boldsymbol{\beta}^{(t)}}^\top. \tag{9}$$

Notice that $\boldsymbol{P}^\star \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$ is block-diagonal with blocks $(S_t \times T_t)$ that are uniform (each entry equals $1/(Rg)^2$), and we can explicitly compute $\|\boldsymbol{P}^\star\|_{S_1} = \frac{1}{R}$. It will be convenient to define the normalized indicator vectors $\boldsymbol{u}^{(t)} = \boldsymbol{\alpha}^{(t)}/\|\boldsymbol{\alpha}^{(t)}\|_2$, $\boldsymbol{v}^{(t)} = \boldsymbol{\beta}^{(t)}/\|\boldsymbol{\beta}^{(t)}\|_2$, which we stack into matrices $\boldsymbol{U}^\star = [\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(R)}]$, $\boldsymbol{V}^\star = [\boldsymbol{v}^{(1)}, \ldots, \boldsymbol{v}^{(R)}]$. In this way, the canonical subgradient of $\|\cdot\|_{S_1}$ at $\boldsymbol{P}^\star$ is $\boldsymbol{G}^\star := \boldsymbol{U}^\star {\boldsymbol{V}^\star}^\top$.

Step 1: Lower bound on $\Delta_{\min}$. For any $\boldsymbol{x} \in B(\boldsymbol{c}_t, \rho)$, $\boldsymbol{y}_{\text{in}} \in B(\boldsymbol{d}_t, \rho)$, $\boldsymbol{y}_{\text{out}} \in B(\boldsymbol{d}_s, \rho)$ with $s \neq t$,

$$\|\boldsymbol{x} - \boldsymbol{y}_{\text{out}}\| \geq \|\boldsymbol{c}_t - \boldsymbol{d}_s\| - \|\boldsymbol{x} - \boldsymbol{c}_t\| - \|\boldsymbol{y}_{\text{out}} - \boldsymbol{d}_s\| \geq \Gamma - 2\rho,$$

and

$$\|\boldsymbol{x} - \boldsymbol{y}_{\text{in}}\| \leq \|\boldsymbol{c}_t - \boldsymbol{d}_t\| + \|\boldsymbol{x} - \boldsymbol{c}_t\| + \|\boldsymbol{y}_{\text{in}} - \boldsymbol{d}_t\| \leq \gamma + 2\rho.$$

Thus

$$\|\boldsymbol{x}-\boldsymbol{y}_{\text{out}}\|^2-\|\boldsymbol{x}-\boldsymbol{y}_{\text{in}}\|^2\geq(\Gamma-2\rho)^2-(\gamma+2\rho)^2,$$

which is positive when (5) holds.

Step 2: Across-block exclusion via tilted cost. By convexity of $\|\cdot\|_{S_1}$,

$$\|\boldsymbol{P}\|_{S_1}\geq\|\boldsymbol{P}^\star\|_{S_1}+\langle\boldsymbol{G}^\star,\boldsymbol{P}-\boldsymbol{P}^\star\rangle,\boldsymbol{G}^\star\in\partial\|\boldsymbol{P}^\star\|_{S_1},\ \boldsymbol{G}^\star=\boldsymbol{U}^\star{\boldsymbol{V}^\star}^\top.$$

Hence for any feasible P,

$$\langle\boldsymbol{C},\boldsymbol{P}\rangle+\lambda\|\boldsymbol{P}\|_{S_1}-(\langle\boldsymbol{C},\boldsymbol{P}^\star\rangle+\lambda\|\boldsymbol{P}^\star\|_{S_1})\geq\left\langle\boldsymbol{S}(\lambda,\boldsymbol{G}^\star),\boldsymbol{P}-\boldsymbol{P}^\star\right\rangle.$$

For $i\in S_t$ and $j\in T_s$ with $s\neq t$, one has $\boldsymbol{G}^\star_{ij}=0$, since the left and right singular vectors are block-supported and orthonormal across clusters. On the other hand, for $j'\in T_t$,

$$\boldsymbol{G}^\star_{ij'}=\langle\boldsymbol{u}^{(t)},\boldsymbol{e}_i\rangle\langle\boldsymbol{v}^{(t)},\boldsymbol{e}_{j'}\rangle=\frac{a_i}{\|\boldsymbol{\alpha}^{(t)}\|_2}\cdot\frac{b_{j'}}{\|\boldsymbol{\beta}^{(t)}\|_2}=\frac{1}{g}.$$

Therefore, for any $i\in S_t$, $s\neq t$, $j\in T_s$, and $j'\in T_t$,

$$\boldsymbol{S}_{ij}(\lambda,\boldsymbol{G}^\star)-\boldsymbol{S}_{ij'}(\lambda,\boldsymbol{G}^\star)=(\|\boldsymbol{x}_i-\boldsymbol{y}_j\|_2^2-\|\boldsymbol{x}_i-\boldsymbol{y}_{j'}\|_2^2)-\lambda\cdot\frac{1}{g}\ \geq\ \Delta_{\min}-\frac{\lambda}{g}.$$

If $\lambda<g\Delta_{\min}$, these gaps are strictly positive, so no $S(\lambda,\boldsymbol{G}^\star)$-optimal coupling can place mass across clusters. Any minimizer of the original problem must then be block-supported on $\bigcup_t(S_t\times T_t)$.

Step 3: Within-block tie-breaking via the nuclear norm. By the distance equality condition in Assumption 3, all within-cluster couplings have equal transport cost. Fix t. We can restrict any feasible coupling $\boldsymbol{P}\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})$ to the block (S_t,T_t), which we denote as $\boldsymbol{P}_{S_t,T_t}$. We note that this can be written as

$$\boldsymbol{P}_{S_t,T_t}=\mathbf{1}_g\mathbf{1}_g^\top/g^2+\boldsymbol{M}^{(t)},\ \text{where}\ \boldsymbol{M}^{(t)}\mathbf{1}=\mathbf{0},(\boldsymbol{M}^{(t)})^\top\mathbf{1}=\mathbf{0}.$$

In other words, we can represent it as rank-1 product coupling plus a perturbation with $\mathbf{0}$ row/column sums. Choose an orthonormal basis of $\mathbb{R}^g$ on the target side with first vector proportional to $\mathbf{1}_g$. Then $\boldsymbol{M}^{(t)}$ lives entirely in the orthogonal complement of $\mathbf{1}_g$. The standard inequality $\|\cdot\|_{S_1}\geq\|\cdot\|_{S_2}$ yields

$$\|\boldsymbol{P}_{S_t,T_t}\|_{S_1}\geq\|\boldsymbol{P}_{S_t,T_t}\|_{S_2}=\sqrt{\|\mathbf{1}_g\mathbf{1}_g^\top/g^2\|_{S_2}^2+\|\boldsymbol{M}^{(t)}\|_{S_2}^2}>\|\mathbf{1}_g\mathbf{1}_g^\top/g^2\|_{S_2}=\|\mathbf{1}_g/g\|_2\|\mathbf{1}_g/g\|_2,$$

whenever $\boldsymbol{M}^{(t)}\neq0$. Summing over t shows that among all block-supported couplings, the nuclear norm is uniquely minimized at $\boldsymbol{M}^{(t)}\equiv0$, i.e., at the uniform block $\boldsymbol{P}^\star$.

Combining these three steps proves the proposition. □

B.2 Proof of Theorem 6

Proof. We proceed in four steps. We will show that the coupling we recover, $\boldsymbol{P}^\star\in\Pi(\boldsymbol{a},\boldsymbol{b})$, satisfies the within-cluster equal split condition given by $\boldsymbol{P}^\star_{i,(t,+)}=\boldsymbol{P}^\star_{i,(t,-)}=\frac{1}{2}a_i$ if $i\in S_t$ otherwise $\boldsymbol{P}^\star_{i,(s,\sigma)}=0$ if $s\neq t$ or $i\notin S_t$.

Step 1: Feasibility and rank-1 structure. By construction, $\boldsymbol{P}^\star\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})$ and, for each $i\in S_t$,

$$T_{\boldsymbol{P}^\star}(\boldsymbol{x}_i)=\tfrac{1}{2}(\boldsymbol{y}_{t,+}+\boldsymbol{y}_{t,-})=\boldsymbol{m}_t.$$

Hence $T_{\boldsymbol{P}^\star}(\boldsymbol{x}_i)-\boldsymbol{x}_i=\boldsymbol{m}_t-\boldsymbol{x}_i=-\xi_i\boldsymbol{u}$ when $\boldsymbol{x}_i=\boldsymbol{m}_t+\xi_i\boldsymbol{u}$ for $|\xi_i|\leq\rho$, which is true by assumption. Writing $\boldsymbol{\gamma}\in\mathbb{R}^n$ such that $\gamma_i=-\xi_i\sqrt{a_i}$, we have

$$\mathcal{A}(\boldsymbol{P}^\star)=\boldsymbol{u}\boldsymbol{\gamma}^\top.$$

Therefore, $\mathrm{rank}\mathcal{A}(\boldsymbol{P}^\star) = 1$, and $\|\mathcal{A}(\boldsymbol{P}^\star)\|_{S_1} = \|\gamma\|$.

Step 2: A tilted-cost lower bound and across-cluster margin. Let $\boldsymbol{G}^\star$ be a canonical subgradient of the nuclear norm at $\boldsymbol{B}^\star := \mathcal{A}(\boldsymbol{P}^\star)$:

$$\boldsymbol{G}^\star \ \in \ \partial\|\boldsymbol{B}^\star\|_{S_1}, \boldsymbol{G}^\star = \boldsymbol{u}\boldsymbol{w}^\top, \text{ where } \boldsymbol{w} := \frac{\boldsymbol{\gamma}}{\|\boldsymbol{\gamma}\|_2}.$$

For any $\boldsymbol{P} \in \mathcal{U}(\boldsymbol{a}, \boldsymbol{b})$, by convexity of the nuclear norm,

$$\|\mathcal{A}(\boldsymbol{P})\|_{S_1} \geq \|\boldsymbol{B}^\star\|_{S_1} + \langle \boldsymbol{G}^\star, \mathcal{A}(\boldsymbol{P}) - \boldsymbol{B}^\star \rangle. \tag{10}$$

Using $\mathcal{A}(\boldsymbol{P}) - \boldsymbol{B}^\star = \boldsymbol{Y}(\boldsymbol{P} - \boldsymbol{P}^\star)^\top \boldsymbol{A}^{-1/2}$ and cyclicity of the trace,

$$\langle \boldsymbol{G}^\star, \mathcal{A}(\boldsymbol{P}) - \boldsymbol{B}^\star \rangle = \langle \boldsymbol{A}^{-1/2}\boldsymbol{G}^{\star\top}\boldsymbol{Y}, \boldsymbol{P} - \boldsymbol{P}^\star \rangle. \tag{11}$$

Combining (10) and (11) with the objective $F_\lambda(\boldsymbol{P}) := \langle \boldsymbol{C}, \boldsymbol{P} \rangle + \lambda\|\mathcal{A}(\boldsymbol{P})\|_{S_1}$ and the definition of the tilted cost $\boldsymbol{S}(\lambda, \boldsymbol{G})$ in (4) yields the lower bound

$$F_\lambda(\boldsymbol{P}) - F_\lambda(\boldsymbol{P}^\star) \geq \langle \boldsymbol{S}(\lambda, \boldsymbol{G}^\star), \boldsymbol{P} - \boldsymbol{P}^\star \rangle. \tag{12}$$

We can compute the tilted costs $\boldsymbol{S}(\lambda, \boldsymbol{G}^\star)$ explicitly: for any i and (t, σ),

$$(\boldsymbol{A}^{-1/2}\boldsymbol{G}^{\star\top}\boldsymbol{Y})_{i,(t,\sigma)} = \frac{1}{\sqrt{a_i}} w_i \langle \boldsymbol{u}, \boldsymbol{y}_{t,\sigma} \rangle = \frac{\gamma_i \mu_t}{\|\boldsymbol{\gamma}\|_2 \sqrt{a_i}} = -\frac{\xi_i \mu_t}{\|\boldsymbol{\gamma}\|_2}.$$

Therefore, assuming that $i \in S_t$ and for any $s \neq t$, $\sigma \in \{\pm\}$,

$$\boldsymbol{S}_{i,(s,\sigma)}(\lambda, \boldsymbol{G}^\star) - \boldsymbol{S}_{i,(t,\pm)}(\lambda, \boldsymbol{G}^\star) = \underbrace{\|x_i - y_{s,\sigma}\|^2 - \|x_i - y_{t,\pm}\|^2}_{=\Delta_{i,s}} + \lambda \left(-\frac{\xi_i}{\|\boldsymbol{\gamma}\|_2} \right) (\mu_s - \mu_t) \tag{13}$$

By assumption,

$$\Delta_{i,s} \ \geq \ |\mu_t - \mu_s|(|\mu_t - \mu_s| - 2\rho) \ \geq \ \Lambda(\Lambda - 2\rho) \ > \ 0. \tag{14}$$

Also, since $|\xi_i| \leq \rho$, we can bound $\|\gamma\|_2^2 = \sum_{k=1}^n a_k \xi_k^2 \leq \rho^2 \sum_{k=1}^n a_k = \rho^2$. This implies that $\frac{|\xi_i|}{\|\boldsymbol{\gamma}\|_2} \leq 1$. Thus we can extend the lower bound in (13)

$$\boldsymbol{S}_{i,(s,\sigma)}(\lambda, \boldsymbol{G}^\star) - \boldsymbol{S}_{i,(t,\pm)}(\lambda, \boldsymbol{G}^\star) \geq \Lambda(\Lambda - 2\rho) - \lambda). \tag{15}$$

Thus, at $\boldsymbol{G}^\star$, for every $\lambda \in [0, \Lambda - 2\rho)$, across-cluster tilted costs are strictly greater than within cluster tilted costs. Therefore any tilted cost optimal coupling must match $\boldsymbol{x}_i$ to $\{\boldsymbol{y}_{t,\pm}\}$ for $i \in S_t$.

Step 3: Within-cluster degeneracy and the nuclear-norm tie-break. Fix $t \in [R]$. For $i \in S_t$, any within-cluster move between the symmetric targets $(t, +)$ and $(t, -)$ has zero cost difference,

$$\|\boldsymbol{x}_i - \boldsymbol{y}_{t,+}\|^2 = \|\boldsymbol{x}_i - \boldsymbol{y}_{t,-}\|^2.$$

Moreover, the tilted cost is the same for $(t, +)$ and $(t, -)$, since $\langle \boldsymbol{u}, \boldsymbol{y}_{t,+} \rangle = \langle \boldsymbol{u}, \boldsymbol{y}_{t,-} \rangle = \mu_t$. Therefore, for any feasible $\boldsymbol{P}$ that sends mass within clusters (i.e., $\mathrm{supp}(\boldsymbol{P}) \subseteq \{(i, (t, \pm)) : \ i \in S_t\}$),

$$\langle \boldsymbol{S}(\lambda, \boldsymbol{G}^\star), \boldsymbol{P} - \boldsymbol{P}^\star \rangle = 0. \tag{16}$$

For such $\boldsymbol{P}$, we can write the within-cluster mass split by $p_i \in [0, 1]$ so that

$$\boldsymbol{P}_{i,(t,+)} = p_i a_i, \ \boldsymbol{P}_{i,(t,-)} = (1 - p_i) a_i$$

A direct computation gives

$$T_{\boldsymbol{P}}(\boldsymbol{x}_i) = p_i \boldsymbol{y}_{t,+} + (1-p_i)\boldsymbol{y}_{t,-} = \boldsymbol{m}_t + (2p_i - 1)\varepsilon \boldsymbol{v}, \varepsilon \geq 0.$$

Hence, with

$$\boldsymbol{\beta} \in \mathbb{R}^n,\ \beta_i := (2p_i - 1)\varepsilon\sqrt{a_i},$$

we obtain the rank-≤ 2 decomposition

$$\mathcal{A}(\boldsymbol{P}) = \boldsymbol{u}\boldsymbol{\gamma}^\top + \boldsymbol{v}\boldsymbol{\beta}^\top. \tag{17}$$

We claim that, for any $\boldsymbol{\beta} \neq 0$,

$$\|\boldsymbol{u}\boldsymbol{\gamma}^\top + \boldsymbol{v}\boldsymbol{\beta}^\top\|_{S_1} > \|\boldsymbol{u}\boldsymbol{\gamma}^\top\|_{S_1} = \|\boldsymbol{\gamma}\|. \tag{18}$$

Indeed, let $\boldsymbol{Q} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix whose first two columns are $\boldsymbol{u}$ and $\boldsymbol{v}$. Orthogonal invariance of singular values implies

$$\|\boldsymbol{u}\boldsymbol{\gamma}^\top + \boldsymbol{v}\boldsymbol{\beta}^\top\|_{S_1} = \|\begin{bmatrix}\boldsymbol{\gamma} & \boldsymbol{\beta} & 0 & \cdots & 0\end{bmatrix}\|_{S_1} = \sigma_1 + \sigma_2,$$

where $\sigma_1 \geq \sigma_2 \geq 0$. If $\boldsymbol{\beta}$ is not colinear with $\boldsymbol{\gamma}$, the matrix has rank 2, so $\sigma_2 > 0$, and $\sigma_1 \geq \|\boldsymbol{\gamma}\|_2$ (since $\|\boldsymbol{M}\|_2 \geq$ the Euclidean norm of any row). Hence $\sigma_1 + \sigma_2 > \|\boldsymbol{\gamma}\|_2$. If instead $\boldsymbol{\beta} = c\boldsymbol{\gamma}$ for some $c \neq 0$, then the matrix has rank 1 with singular value $\sqrt{\|\boldsymbol{\gamma}\|_2^2 + \|\boldsymbol{\beta}\|_2^2} = \sqrt{1+c^2}\|\boldsymbol{\gamma}\|_2 > \|\boldsymbol{\gamma}\|_2$. Thus (18) holds in all cases $\beta \neq 0$.

Combining (12) and (16), for any within-cluster feasible P,

$$F_\lambda(\boldsymbol{P}) - F_\lambda(\boldsymbol{P}^\star) \geq \lambda(\|\mathcal{A}(\boldsymbol{P})\|_{S_1} - \|\boldsymbol{B}^\star\|_{S_1}) = \lambda(\|\boldsymbol{u}\boldsymbol{\gamma}^\top + \boldsymbol{v}\boldsymbol{\beta}^\top\|_{S_1} - \|\boldsymbol{\gamma}\|_2), \tag{19}$$

which is strictly positive by (18) whenever $\beta \neq 0$, i.e., whenever some $p_i \neq \frac{1}{2}$.

Step 4: Optimality and uniqueness for $\lambda \in [0, \Lambda - 2\rho)$. Let $\lambda \in [0, \Lambda - 2\rho)$. For any feasible $\boldsymbol{P}$, decompose $\boldsymbol{P} - \boldsymbol{P}^\star$ into an across-cluster part and a within-cluster part. By (15),

$$\langle S(\lambda, \boldsymbol{G}^\star), \boldsymbol{P} - \boldsymbol{P}^\star\rangle > 0$$

if $\boldsymbol{P}$ sends any mass across clusters, and (12) implies $F_\lambda(\boldsymbol{P}) > F_\lambda(\boldsymbol{P}^\star)$ in that case. Therefore, any minimizer of F_λ must be supported within clusters. For within-cluster couplings, (19) implies $F_\lambda(\boldsymbol{P}) > F_\lambda(\boldsymbol{P}^\star)$ unless $p_i = \frac{1}{2}$ for all i, i.e., if $\boldsymbol{P} = \boldsymbol{P}^\star$. Consequently, $\boldsymbol{P}^\star$ is the *unique* minimizer of the Schatten OT problem for $\lambda \in [0, \Lambda - 2\rho)$. □

C The Gaussian Case

The previous section treated recovery of low-rank structures in discrete OT. We now discuss an application of the continuous Schatten regularization (8) for Gaussians.

We treat two Gaussian specializations of the Schatten-p programs discussed earlier: (i) a nuclear-norm penalty that promotes low-rank cross-covariance, and (ii) a nuclear-norm penalty that promotes low-rank transport. As emphasized in our general framework, the barycentric projection $x \mapsto \mathbb{E}_\pi[Y \mid X = x]$ is an *affine* map of the coupling π, so the induced Schatten-p penalty is convex in π for $p \geq 1$. The same holds for Schatten penalties applied to any affine image $A(\pi)$.

For simplicity, we consider the mean zero case. Let $\mu = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$ and $\nu = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$ on $\mathbb{R}^d$ with $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1 \succ 0$. A *Gaussian coupling* is a joint Gaussian $\pi = \mathcal{N}\big(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_0 & \boldsymbol{K}\\ \boldsymbol{K}^\top & \boldsymbol{\Sigma}_1\end{bmatrix}\big)$, parameterized by a cross-covariance $\boldsymbol{K} \in \mathbb{R}^{d\times d}$ satisfying the feasibility constraint

$$\begin{bmatrix}\boldsymbol{\Sigma}_0 & \boldsymbol{K}\\ \boldsymbol{K}^\top & \boldsymbol{\Sigma}_1\end{bmatrix} \succeq 0 \iff \|\boldsymbol{\Sigma}_0^{-1/2}\boldsymbol{K}\boldsymbol{\Sigma}_1^{-1/2}\|_2 \leq 1. \tag{20}$$

We denote the set of all such $\boldsymbol{K}$ as $\mathcal{K}(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)$. Equivalently, we denote the set of all Gaussian couplings between μ and ν as $\Pi_g(\mu, \nu)$.

For the quadratic cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$, the transport cost under π is $\mathbb{E}_\pi\|X - Y\|^2 = \mathrm{tr}(\boldsymbol{\Sigma}_0) + \mathrm{tr}(\boldsymbol{\Sigma}_1) - 2\mathrm{tr}(\boldsymbol{K})$. Moreover the barycentric map induced by π is

$$T_\pi(\boldsymbol{x}) = \boldsymbol{A}_\pi \boldsymbol{x}, \;\; \boldsymbol{A}_\pi := \boldsymbol{K}^\top \boldsymbol{\Sigma}_0^{-1}. \tag{21}$$

C.1 Gaussian Low-rank Cross-Covariance

Consider the Gaussian Schatten OT problem

$$\min_{\boldsymbol{K} \in \mathcal{K}(\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1)} \mathrm{tr}(\boldsymbol{\Sigma}_0) + \mathrm{tr}(\boldsymbol{\Sigma}_1) - 2\mathrm{tr}(\boldsymbol{K}) + \lambda \|\boldsymbol{K}\|_{S_1}. \tag{22}$$

This is a semidefinite program.

We can solve this problem in closed form. Let $\boldsymbol{S} := \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_0^{1/2}$ and write its SVD $\boldsymbol{S} = \boldsymbol{U}\mathrm{diag}(\sigma_1, \ldots, \sigma_d)\boldsymbol{V}^\top$ with $\sigma_1 \geq \cdots \geq \sigma_d > 0$. Feasible $\boldsymbol{K}$ can be written as $\boldsymbol{K} = \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{M}\boldsymbol{\Sigma}_1^{1/2}$ with $\|\boldsymbol{M}\|_2 \leq 1$. By von Neumann's trace inequality, $\mathrm{tr}(\boldsymbol{K}) = \mathrm{tr}(\boldsymbol{S}\boldsymbol{M}) \leq \sum_i \sigma_i s_i$ where s_i are the singular values of $\boldsymbol{M}$, with equality when $\boldsymbol{M}$ shares singular vectors with $\boldsymbol{S}$. At optimum we may thus take $\boldsymbol{M} = \boldsymbol{U}\mathrm{diag}(s_1, \ldots, s_d)\boldsymbol{V}^\top$ with $0 \leq s_i \leq 1$ and $\|\boldsymbol{K}\|_{S_1} = \|\boldsymbol{M}\|_{S_1} = \sum_i s_i$. The objective in (22) reduces (up to constants) to

$$\min_{0 \leq s_i \leq 1} \sum_{i=1}^d \big(\lambda - 2\sigma_i\big) s_i = \sum_{i=1}^d \min_{0 \leq s \leq 1} (\lambda - 2\sigma_i) s,$$

which is separable and linear in each s_i. Hence we arrive at the following proposition.

Proposition 7 (Hard spectral selection)**.** *The unique minimizer of* (22) *is obtained by* hard thresholding *the singular spectrum of* $\boldsymbol{S}$*:*

$$s_i^\star = \mathbf{1}\{\sigma_i > \lambda/2\}, \qquad \boldsymbol{K}_\lambda = \boldsymbol{\Sigma}_0^{1/2}\boldsymbol{U}\mathrm{diag}(s_i^\star)\boldsymbol{V}^\top\boldsymbol{\Sigma}_1^{1/2}. \tag{23}$$

In particular, $\mathrm{rank}(\boldsymbol{K}_\lambda) = \#\{i : \sigma_i > \lambda/2\}$.

We include some examples to illustrate this hard thresholding rule.

(i) *Isotropic case:* $\boldsymbol{\Sigma}_0 = a^2\boldsymbol{I}, \boldsymbol{\Sigma}_1 = b^2\boldsymbol{I}$ gives $\boldsymbol{S} = ab\boldsymbol{I}$ and hence either $\boldsymbol{K}_\lambda = ab\boldsymbol{I}$ if $\lambda < 2ab$, or $\boldsymbol{K}_\lambda = 0$ if $\lambda \geq 2ab$.

(ii) *Commuting covariances:* If $\boldsymbol{\Sigma}_0 = \boldsymbol{U}\mathrm{diag}(\boldsymbol{a})\boldsymbol{U}^\top$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{U}\mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^\top$, then $\sigma_i = \sqrt{a_i b_i}$ and $\boldsymbol{K}_\lambda = \boldsymbol{U}\mathrm{diag}\big(\sqrt{\boldsymbol{a} \odot \boldsymbol{b}}\mathbf{1}\{\sqrt{\boldsymbol{a} \odot \boldsymbol{b}} > \lambda/2\}\big)\boldsymbol{U}^\top$. Here, $\odot$ is the Hadamard (elementwise) product.

We note that the inclusion of a Schatten-2 penalty in this program results in soft thresholding.

C.2 Gaussian Barycentric Displacements

We now penalize the (weighted) barycentric displacement. In this case, the resulting convex program is

$$\min_{\pi \in \Pi_g(\mu,\nu)} \mathrm{tr}(\boldsymbol{\Sigma}_0) + \mathrm{tr}(\boldsymbol{\Sigma}_1) - 2\mathrm{tr}(\boldsymbol{K}) + \lambda \big\|(\boldsymbol{A}_\pi - \boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}\big\|_{S_1}, \tag{24}$$

with $\boldsymbol{A}_\pi$ as in (21).

We can give a closed form when $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ commute, which already reveals a clear structure. Suppose there exists an orthogonal $\boldsymbol{U}$ with $\boldsymbol{\Sigma}_0 = \boldsymbol{U}\mathrm{diag}(\boldsymbol{a})\boldsymbol{U}^\top$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{U}\mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^\top$. Any feasible $\boldsymbol{K}$ aligned with $\boldsymbol{U}$ takes the form $\boldsymbol{K} = \boldsymbol{U}\mathrm{diag}(\sqrt{\boldsymbol{a} \odot \boldsymbol{b}} \odot \boldsymbol{m})U^\top$ with $0 \leq m_i \leq 1$. Then $\boldsymbol{A}_\pi = \boldsymbol{K}^\top\boldsymbol{\Sigma}_0^{-1}$ is diagonal in the same basis with entries

$$\alpha_i := \frac{\sqrt{a_i} m_i \sqrt{b_i}}{a_i} = m_i \sqrt{\frac{b_i}{a_i}},$$

which implies that the diagonal values of $(\boldsymbol{A}_\pi - I)\boldsymbol{\Sigma}_0^{1/2}$ are $m_i\sqrt{b_i} - \sqrt{a_i}$ in this basis as well. Hence $\|(\boldsymbol{A}_\pi - \boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}\|_{S_1} = \sum_i |m_i\sqrt{b_i} - \sqrt{a_i}|$, and

$$\mathrm{tr}(\boldsymbol{K}) = \sum_i \sqrt{a_i} m_i \sqrt{b_i}.$$

Consequently, (24) decouples into d scalar problems over $m_i \in [0, 1]$:

$$\min_{0 \le m \le 1} \phi_{\boldsymbol{a},\boldsymbol{b},\lambda}(m) := -2\sqrt{ab}m + \lambda|m\sqrt{b} - \sqrt{a}|. \tag{25}$$

Theorem 8. *Fix $(\boldsymbol{a}, \boldsymbol{b}, \lambda)$ with $\boldsymbol{a}, \boldsymbol{b} > 0$. The unique minimizer of (25) is*

$$m^\star = \begin{cases} 1, & \text{if } b \le a, \\ 1, & \text{if } b > a \text{ and } \lambda < 2\sqrt{a}, \\ \sqrt{a/b}, & \text{if } b > a \text{ and } \lambda \ge 2\sqrt{a}. \end{cases}$$

Equivalently, the coupling that solves (24) has barycentric projection map $T_\pi(\cdot) = \boldsymbol{A}_\lambda \cdot$, where

$$\boldsymbol{A}_\lambda = \boldsymbol{U}\mathrm{diag}(\boldsymbol{\alpha}^\star)\boldsymbol{U}^\top, \qquad \alpha_i^\star = \begin{cases} \sqrt{b_i/a_i}, & \text{if } b_i \le a_i, \\ \sqrt{b_i/a_i}, & \text{if } b_i > a_i \text{ and } \lambda < 2\sqrt{a_i}, \\ 1, & \text{if } b_i > a_i \text{ and } \lambda \ge 2\sqrt{a_i}. \end{cases} \tag{26}$$

The proof follows from analyzing the 1-dimensional optimization problem (25).

In words, the regularizer does not suppress contracting directions ($b_i \le a_i$), and it prunes expanding directions $b_i > a_i$ back to the identity once λ crosses the sharp threshold $2\sqrt{a_i}$. Thus, $\mathrm{rank}(\boldsymbol{A}_\lambda - \boldsymbol{I})$ equals the number of contracting eigendirections plus the number of expanding eigendirections with $\lambda < 2\sqrt{a_i}$.

As an example, consider the isotropic case. If $\boldsymbol{\Sigma}_0 = \sigma_0^2\boldsymbol{I}, \boldsymbol{\Sigma}_1 = \sigma_1^2\boldsymbol{I}$. If $\sigma_1 \le \sigma_0$, then $\boldsymbol{A}_\lambda = (\sigma_1/\sigma_0)\boldsymbol{I}$ for all λ. On the other hand, if $\sigma_1 > \sigma_0$, then $\boldsymbol{A}_\lambda = (\sigma_1/\sigma_0)\boldsymbol{I}$ for $\lambda < 2\sigma_0$, and $\boldsymbol{A}_\lambda = I$ for $\lambda \ge 2\sigma_0$.

C.3 Discussion

In the Gaussian case, for $p > 1$, the programs remain convex. In the commuting case, we can find similar separable one-dimensional convex problems as we found in the previous sections. For larger p, we would observe smooth shrinkage of the spectrum m_i rather than hard thresholds at $p = 1$. We leave a detailed analysis of the noncommuting case to future work.

D Details on 4i Experiment Setup

For reproducibility, we give more details on the 4i experiment here. We subsample source and target points to form measures with 1000 points each. We then run mirror descent with the desired regularization, where the initial step size is $\eta_0 = 0.1$ for low-rank coupling recovery and $\eta_0 = 10^{-4}$ for low-rank barycentric map recovery. A diminishing step size of $\eta_k = \eta_0/\sqrt{k}$ is used. We run mirror descent for a maximum of 50 iterations, where the Sinkhorn projection at each iteration is run for 500 iterations unless the marginal error of 10^{-12} is reached. This is followed by the rounding procedure of Altschuler et al. (2017) to ensure iterates remain in the transport polytope. We return the averaged iterate as our proposed solution to the Schatten OT problem.