Supplementary material for paper: A Swiss Army Knife for Minimax Optimal Transport

This supplementary material contains the proofs for the different theoretical results of the main paper, as well as more details on the experimental part provided for the sake of reproducibility.

Recall of the setting

Optimal transport problem For two probability measures μ_1 and μ_2, a cost function $c : (\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y} \mapsto c(\mathbf{x}, \mathbf{y}) \in \mathbb{R}_+$ and two complete metric spaces $\mathcal{X}$ and $\mathcal{Y}$, the Kantorovitch (Kantorovich, 1942) formulation of OT seeks for an optimal coupling γ between μ_1 and μ_2 which minimizes the following quantity:

$$\begin{aligned} W_c(\mu_1, \mu_2) &= \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \int_{\mathcal{X}\times\mathcal{Y}} c(\mathbf{x}, \mathbf{y}) d\gamma(\mathbf{x}, \mathbf{y}) \\ &= \inf_{\gamma \in \Pi(\mu_1,\mu_2)} \underset{(X,Y)\sim\gamma}{\mathbb{E}} \left[c(X, Y)\right], \end{aligned}$$

When c is the squared Euclidean distance, we write simply W_2. In the discrete version of the problem, *i.e.* when μ_1 and μ_2 are defined as empirical measures supported on vectors $\{\mathbf{x}_i\}_{i=1}^m$, $\{\mathbf{y}_j\}_{j=1}^n$ in $\mathbb{R}^d$ with probability vectors $\mathbf{r} \in \Delta_m$ and $\mathbf{c} \in \Delta_n$, the previous problem can be expressed as follows:

$$\mathbf{P}^* \in \underset{\mathbf{P}\in\Pi(\mathbf{r},\mathbf{c})}{\operatorname{argmin}} \langle \mathbf{P}, \mathbf{C}\rangle_F, \tag{1}$$

where $\langle \cdot, \cdot\rangle_F$ is the Frobenius dot product, $\mathbf{C} \in \mathbb{R}_+^{m\times n}$ is a cost matrix representing the pairwise costs of transporting $\mathbf{x}_i$ to $\mathbf{y}_j$ and $\mathbf{P}$ is a joint distribution given by a matrix of size $m \times n$ belonging to the transportation polytope $\Pi(\mathbf{r}, \mathbf{c})$ (also called Birkhoff polytope for $m = n$) defined as:

$$\Pi(\mathbf{r}, \mathbf{c}) = \{\mathbf{P} \in \mathbb{R}_+^{m\times n}; \mathbf{P}\mathbb{1}_n = \mathbf{c}; \mathbf{P}^T \mathbb{1}_m = \mathbf{r}\}.$$

Minimax OT (Paty & Cuturi, 2019) showed that one can see the OT problem, with c taken to be the squared Euclidean distance, as a trace minimization problem of the second-order displacement matrix defined for any $\gamma \in \Pi(\mu_1, \mu_2)$ as

$$\mathbf{V}_\gamma := \int_{\mathcal{X}\times\mathcal{Y}} (\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T d\gamma(\mathbf{x}, \mathbf{y})$$

or equivalently in the discrete case for any $\mathbf{P} \in \Pi(\mathbf{r}, \mathbf{c})$ as

$$(\mathbf{V}_\mathbf{P})_{ij} := \sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}(\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i - \mathbf{y}_j)^T.$$

Using this reformulation, we can equivalently rewrite Equation (1) as follows: $\mathbf{P}^* \in \operatorname{argmin}_{\mathbf{P}\in\Pi(\mathbf{r},\mathbf{c})} \operatorname{tr}(V_\mathbf{P})$.

Robust optimal transport Let $\mathcal{C}$ be a convex compact set of cost functions defined over $\mathcal{X} \times \mathcal{Y}$ with no particular constraints on the form of distances used to compute matrices belonging to $\mathcal{C}$ as long as a solution of the corresponding Kantorovich problem exists. We further denote the convex hull of a set S as $\operatorname{Conv}(S)$ that is the set of convex combinations of S's elements.

We now consider the following bilinear minimax problem:

$$\operatorname{RKP}(\Pi, \mathcal{C}) = \min_{\gamma\in\Pi} \max_{c\in\mathcal{C}} \underset{\mathbf{x},\mathbf{y}\sim\gamma}{\mathbb{E}} \left[c(\mathbf{x}, \mathbf{y})\right], \tag{2}$$

We denote the value at the solution of this problem by $\operatorname{RKP}(\Pi, \mathcal{C})$ and write $\operatorname{RKP}(\mathcal{P}, \mathcal{C})$ for any set $\mathcal{P}$ (even non convex) to denote $\operatorname{RKP}(\operatorname{conv}(\mathcal{P}), \mathcal{C})$. We also extend the notation W_c, presented before, to the case $c \in \mathcal{C}$ by defining

$$W_\mathcal{C}(\mu_1, \mu_2) = \operatorname{RKP}(\Pi, \mathcal{C}).$$

Proofs from Section 3.2

For any vector $\mathbf{u} = (u_1, ..., u_d) \in \mathbb{R}^d$ and matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$, we define their respective norms (Schatten p-norm in case of matrix) as

$$\|\mathbf{u}\|_p := \left(\sum_{1\le i\le d} |u_i|^p\right)^{\frac{1}{p}}, \|\mathbf{M}\|_p := \left(\sum_{1\le i\le d} \sigma_i^p(\mathbf{M})\right)^{\frac{1}{p}},$$

where $p \in [1, +\infty]$ and $\{\sigma_i(\mathbf{M})\}$ are $\mathbf{M}$'s singular values. In particular, if $\mathbf{M}$ is a symmetric, positive semi-definite matrix, then

$$\|\mathbf{M}\|_p = \operatorname{Tr}\{\mathbf{M}^p\}^{\frac{1}{p}}.$$

We also recall that the dual of a $p-$norm (respectively Schatten $p-$norm) is the $q-$norm (respectively the Schatten $q-$norm), with q equal to $\frac{p}{p-1}$ if $p > 1$, to ∞ if $p = 1$ and to 1 if $p = \infty$. We define $\mathcal{C}$ as a family of Mahalanobis distance cost matrices, indexed by bounded symmetric matrices $\mathbf{M} \in \mathcal{S}_d^+(\mathbb{R})$ (where $\mathcal{S}_d(R)$ is the set of symmetric PSD matrices):

$$\mathcal{C} = \{c^{\mathbf{M}} : (\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x}-\mathbf{y})^T \mathbf{M}(\mathbf{x}-\mathbf{y}); \|\mathbf{M}\|_p \le 1\}. \quad (3)$$

Claim (footnote 1 in the main paper) $\mathcal{C}$ defined in (3) is a convex compact set.

Proof. Denoting $\mathcal{F}(\mathcal{X} \times \mathcal{Y}, \mathbb{R})$ the set of real valued functions on $\mathcal{X} \times \mathcal{Y}$, let:

$$\begin{aligned} \Phi : \mathbb{R}^{d \times d} &\mapsto \mathcal{F}(\mathcal{X} \times \mathcal{Y}, \mathbb{R}) \\ \mathbf{M} &\mapsto c^{\mathbf{M}} : (\mathbf{x}, \mathbf{y}) \mapsto (\mathbf{x} - \mathbf{y})^T \mathbf{M}(\mathbf{x} - \mathbf{y}) \end{aligned}$$

Notice that Φ is linear and its domain is a finite dimensional vector space, hence Φ is continuous.
Moreover, we have $\mathcal{C} = \Phi\left(\mathcal{B}_p^d\right)$ where

$$\mathcal{B}_p^d := \{\mathbf{M} \in \mathbb{R}^{d \times d}; \|\mathbf{M}\|_p \le 1\}$$

is the unit ball of norm $\|.\|_p$, which is compact and convex. As a result, $\mathcal{C}$ is a convex compact set. □

Proposition 1. *Let $\mathcal{C}$ be defined as in (3) for $\mathbf{M} \in \mathcal{S}_+^d(\mathbb{R})$. Then, $\mathcal{C}$ is a convex compact set of cost functions and the following holds:*

$$\max_{c \in \mathcal{C}} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} [c(\mathbf{x}, \mathbf{y})] = \max_{\mathbf{M} \in \mathcal{S}_+^d \ \|\mathbf{M}\|_p \le 1} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle_F = \|\mathbf{V}_\gamma\|_q$$

implying $\mathrm{RKP}(\Pi, \mathcal{C}) = \min_{\gamma \in \Pi} \|\mathbf{V}_\gamma\|_q$. *Furthermore, for any* $\gamma \in \Pi$,

$$\mathbf{M}^* = \mathop{\mathrm{argmax}}_{\mathbf{M} \in \mathcal{S}_+^d \ \|\mathbf{M}\|_p \le 1} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle_F = \left(\frac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|_q} \right)^{\frac{q}{p}} \quad (4)$$

verifies $\|\mathbf{M}^*\|_p = 1$. *In particular, for* $p = \infty$, $\min_{\gamma \in \Pi} \|\mathbf{V}_\gamma\|_1 = W_2^2(\mu_1, \mu_2)$*, i.e., we recover the classic 2-Wasserstein distance.*

Proof. With the notations used to prove that $\mathcal{C}$ is a convex compact, adding the PSD constraint on $\mathbf{M}$ can be done by considering the image of $\mathcal{B}_{p+}^d := \mathcal{B}_p^d \cap \mathcal{S}_+^d(\mathbb{R})$ by mapping Φ. $\mathcal{B}_{p+}^d$ is a convex compact set as it is the intersection of a convex compact set and a convex cone (the PSD cone). For fixed $\gamma \in \Pi$, we compute the maximum of $\mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} \left[c^{\mathbf{M}}(\mathbf{x}, \mathbf{y})\right]$ over $\mathbf{M} \in \mathcal{B}_{p+}^d$.

$$\begin{aligned} &\max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} [c(\mathbf{x}, \mathbf{y})] \\ &= \max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} \left[(\mathbf{x} - \mathbf{y})^T \mathbf{M}(\mathbf{x} - \mathbf{y})\right] \\ &= \max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} \left[\mathrm{Tr}\big((\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T \mathbf{M}\big)\right] \\ &= \max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} \left[\left\langle (\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T, \mathbf{M} \right\rangle_F\right] \\ &= \max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle_F . \end{aligned}$$

where we used properties of the trace operator, the linearity of the expectation and the definition of $\mathbf{V}_\gamma$.

This maximum is achieved for $\mathbf{M}^*$ verifying $\|\mathbf{M}^*\|_p = \|\mathbf{M}^*\|_p^p = \mathrm{Tr}\{(\mathbf{M}^*)^p\} = 1$. In fact, supposing this is not the case, i.e.$\|\mathbf{M}^*\|_p < 1$, then $\mathbf{M}^{**} = \frac{\mathbf{M}^*}{\|\mathbf{M}^*\|}$ verifies $\langle \mathbf{V}_\gamma, \mathbf{M}^{**} \rangle_F > \langle \mathbf{V}_\gamma, \mathbf{M}^* \rangle_F$, which contradicts $\mathbf{M}^*$'s optimality.
Using the equality case of the Hölder inequality for Schatten p-norms (Magnus, 1987, Theorem 5), the only PSD matrix achieving this maximum is:

$$\mathbf{M}^* = \left(\frac{\mathbf{V}_\gamma^q}{\mathrm{Tr}\{\mathbf{V}_\gamma^q\}} \right)^{\frac{1}{p}} = \left(\frac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|_q} \right)^{\frac{q}{p}}$$

and the value of the maximum is $\|\mathbf{V}_\gamma\|_q$. Taking the minimum over $\gamma \in \Pi$, we obtain:

$$\min_{\gamma \in \Pi} \max_{\mathbf{M} \in \mathcal{B}_{p+}^d} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \gamma} [c(\mathbf{x}, \mathbf{y})] = \min_{\gamma \in \Pi} \|\mathbf{V}_\gamma\|_q .$$

In particular, for $p = \infty$, the corresponding dual norm is $\|\cdot\|_1$, and we have:

$$\begin{aligned} \min_{\gamma \in \Pi} \|\mathbf{V}_\gamma\|_1 &= \min_{\gamma \in \Pi} \mathrm{Tr}\{\mathbf{V}_\gamma\} \\ &= \min_{\gamma \in \Pi} \mathop{\mathbb{E}}_{\mathbf{x},\mathbf{y} \sim \Pi} \left[\|\mathbf{x} - \mathbf{y}\|^2\right] = W_2^2(\mu_1, \mu_2). \end{aligned}$$

This concludes the proof. □

Corollary 1 (Euclidean norm case)**.** *Let $\mathcal{C}$ be defined with $p = 2$ in (3) and let $\mathbf{M}^* = \mathrm{argmax}_{\|\mathbf{M}\|_2 \le 1} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle$. Then $\mathbf{M}^* = \dfrac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|_2}$, thus $\mathbf{M}^*$ is PSD and $\|\mathbf{M}^*\|_2 = 1$.*

Proof. $\sup_{\|\mathbf{M}\|_2 \le 1} \langle \mathbf{V}_\gamma, \mathbf{M} \rangle_F$ is achieved, without imposing that $\mathbf{M}$ is PSD, for $\mathbf{M} = \frac{\mathbf{V}_\gamma}{\|\mathbf{V}_\gamma\|}$ (by the equality case of the Cauchy-Schwartz inequality). This matrix is PSD as $\mathbf{V}_\gamma$ is PSD, and has unit norm. □

Corollary 2. *With the assumptions from Proposition 1, the following inequality holds for any $p \in [1, +\infty]$:*

$$\frac{1}{d^{\frac{1}{p}}} W_2^2(\mu_1, \mu_2) \le W_{\mathcal{C}}(\mu_1, \mu_2) \le W_2^2(\mu_1, \mu_2). \quad (5)$$

Proof. Let $\gamma \in \Pi$. We have for any $p \geq 1$, $\|\mathbf{V}_\gamma\|_p \leq \|\mathbf{V}_\gamma\|_1$ by the monotonicity of Schatten $p-$norm, and we have $\min_{\gamma \in \Pi} \|\mathbf{V}_\gamma\|_1 = W_2^2(\mu_1, \mu_2)$. Taking the infimum over $\gamma \in \Pi$ yields the right hand side inequality in (5). To obtain the left hand side, notice that $\mathbf{A} := d^{-\frac{1}{p}} I_d$ verifies $\|\mathbf{A}\|_p \leq 1$, and that $\mathbf{A}$ is PSD, so that $c^{\mathbf{A}} \in \mathcal{C}$. Thus,

$$\min_{\gamma \in \Pi} \langle \mathbf{V}_\gamma, \mathbf{A} \rangle_F \leq W_{\mathcal{C}}(\mu_1, \mu_2).$$

Finally, notice that the left hand side in the previous inequality equals $\frac{1}{d^{\frac{1}{p}}} W_2^2(\mu_1, \mu_2)$, thus concluding the proof. □

Proofs from Section 3.3

Let $\mathcal{X}$ and $\mathcal{Y}$ are identified respectively with finite sets $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\mathbf{y}_j\}_{i=j}^n$, hence $\mathcal{C}$ is identified with a convex compact set of cost matrices with entries $(\mathbf{C})_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$. For any positive integer p, we denote $\Delta_p = \mathbf{v} \in \mathbb{R}_+^p; \mathbb{1}_p^T \mathbf{v} = 1$ the $p-$dimensional probability simplex.

Proposition 2. *Let $\mathcal{P}$ be a finite subset of Π. The problem* $\mathrm{RKP}(\mathcal{P}, \mathcal{C}) := \mathrm{RKP}(\mathrm{Conv}(\mathcal{P}), \mathcal{C})$ *has a saddle point* $(\mathbf{P}^*, \mathbf{C}^*)$ *verifying*

$$\langle \mathbf{P}^*, \mathbf{C}^* \rangle_F = \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \max_{\mathbf{C} \in \mathcal{C}} \langle \mathbf{P}, \mathbf{C} \rangle_F \tag{6}$$

$$= \max_{\mathbf{C} \in \mathcal{C}} \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F . \tag{7}$$

Moreover, solving $\mathrm{RKP}(\mathcal{P}, \mathcal{C})$ *is equivalent to solving*

$$\begin{aligned} &\mathbf{C}^* \in \mathrm{argmax}_{\mathbf{C} \in \mathcal{C}, \mu \geq 0} \ \mu \\ &\textit{s.t.} \ \langle \mathbf{P}, \mathbf{C} \rangle_F \geq \ \mu, \ \forall \mathbf{P} \in \mathcal{P}. \end{aligned} \tag{8}$$

Also, $\mathbf{P}^* = \sum_{l=1}^{|\mathcal{P}|} q_l \mathbf{P}_l$*, where* $\{q_l\}_{l=1}^{|\mathcal{P}|}, \sum_i q_i = 1$*, are dual variables of* (8)*. In particular, solving* $\mathrm{RKP}(\Pi, \mathcal{C})$ *can be done by setting $\mathcal{P}$ as the set of vertices of* Π.

Proof. Since the set $\mathcal{P}$ is finite, $\mathrm{Conv}(\mathcal{P})$ is a convex compact set. Also, by definition, $\mathcal{C}$ is a convex compact set. Moreover, we note that for any $(\mathbf{P}, \mathbf{C}) \in \Pi \times \mathcal{C}$, the functions $\langle \mathbf{P}, \cdot \rangle_F$ and $\langle \cdot, \mathbf{C} \rangle_F$ are linear. By applying Sion's min-max theorem (Sion, 1958), problem $\mathrm{RKP}(\mathcal{P}, \mathcal{C}) := \mathrm{RKP}(\mathrm{Conv}(\mathcal{P}), \mathcal{C})$ has at least a saddle point, and any saddle point $(\mathbf{P}^s, \mathbf{C}^s)$ verifies:

$$\langle \mathbf{P}^s, \mathbf{C}^s \rangle_F = \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \max_{\mathbf{C} \in \mathcal{C}} \langle \mathbf{P}, \mathbf{C} \rangle_F$$

$$= \max_{\mathbf{C} \in \mathcal{C}} \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \langle \mathbf{P}, \mathbf{C} \rangle_F . \tag{9}$$

However, for any fixed $\mathbf{C} \in \mathcal{C}$, the linearity of $\langle \cdot, \mathbf{C} \rangle_F$ implies that its minimum on $\mathrm{Conv}(\mathcal{P})$ is achieved on one of its vertices, i.e.:

$$\min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \langle \mathbf{P}, \mathbf{C} \rangle_F = \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F \, \forall \mathbf{C} \in \mathcal{C} \Rightarrow$$

$$\max_{\mathbf{C} \in \mathcal{C}} \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \langle \mathbf{P}, \mathbf{C} \rangle_F = \max_{\mathbf{C} \in \mathcal{C}} \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F . \tag{10}$$

Combining (9) and (10) yields Equation (7).
Moreover, by the saddle point's definition, we have: $\mathbf{C}^* \in \mathrm{argmax}_{\mathbf{C} \in \mathcal{C}} \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F$. Using the fact that $\mathcal{P}$ is finite, we obtain the equivalent Problem (8). What is left is computing $\mathbf{P}^*$'s value. To this end, let us introduce $I_{\mathcal{C}}$, the convex indicator function of set $\mathcal{C}$, defined by:

$$\begin{aligned} I_{\mathcal{C}} : \mathbf{C} \mapsto 0 \quad &\text{if} \quad \mathbf{C} \in \mathcal{C} \\ +\infty \quad &\text{otherwise} \end{aligned}$$

Also, notice that μ is nonnegative even without imposing this condition. In fact, assuming that the cost matrices in $\mathcal{C}$ have positive values, we have $\min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F \geq 0$, for all $\mathbf{C} \in \mathcal{C}$. If μ^*, the value of μ at the solution was negative, its maximality contradicts the condition $\min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C} \rangle_F \geq 0$. Hence, Problem (8) is equivalent to the following:

$$\begin{aligned} &\max_{\mathbf{C} \in \mathbb{R}^{m \times n}, \mu \in \mathbb{R}} \mu - I_{\mathcal{C}}(\mathbf{C}), \\ &\qquad \text{s.t} \ \langle \mathbf{P}, \mathbf{C} \rangle_F \geq \mu \quad \forall \mathbf{P} \in \mathcal{P}. \end{aligned}$$

The Lagrangian of the previous problem is:

$$\mathcal{L}(\mathbf{q}, \mathbf{C}, \mu) = \mu - I_{\mathcal{C}}(\mathbf{C}) + \sum_{l=1}^{|\mathcal{P}|} q_l (\langle \mathbf{P}_l, \mathbf{C} \rangle_F - \mu), \tag{11}$$

where l indexes the finite set of matrices $\mathcal{P}$, $q_l \geq 0$ for all $l \in \{1, ..., |\mathcal{P}|\}$ denote the dual variables of the constraints, and $\mathbf{q} = (q_1, ..., q_{|\mathcal{P}|})$. A known optimization result (Boyd & Vandenberghe, 2004, Section 5.4.2) implies that that the solution to the primal, $(\mathbf{C}^*, \mu^*)$ and the solution to the dual, $\mathbf{q}^* = (q_1^*, ..., q_l^*)$ form a saddle point of the Lagrangian, which implies:

$$\mathcal{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*) = \max_{\mathbf{C}, \mu} \mathcal{L}(\mathbf{q}^*, \mathbf{C}, \mu) \tag{12}$$

Deriving the Lagrangian with respect to μ yields:

$$\sum_l q_l^* = 1. \tag{13}$$

In addition to this condition, knowing that the value of the Lagrangian is finite at the solution, we have $I_{\mathcal{C}}(\mathbf{C}^*) = 0$. Substituting the last two conditions in Equation (12) yields:

$$\begin{aligned} \mathcal{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*) &= \langle \mathbf{P}^*, \mathbf{C}^* \rangle_F \\ &= \max_{\mathbf{C} \in \mathbb{R}^{m \times n}} \langle \mathbf{P}^*, \mathbf{C} \rangle_F - I_{\mathcal{C}}(\mathbf{C}) \\ &= \max_{\mathbf{C} \in \mathcal{C}} \langle \mathbf{P}^*, \mathbf{C} \rangle_F \end{aligned} \tag{14}$$

where $\mathbf{P}^*$ is defined as in the proposition. Also, Equation (13) implies that there is at least one $l' \in \{1, ..., |\mathcal{P}|\}$ verifying $q_{l'} > 0$, and hence

$$\mu^* = \langle \mathbf{P}_{l'}, \mathbf{C}^* \rangle_F = \min_{\mathbf{P} \in \mathcal{P}} \langle \mathbf{P}, \mathbf{C}^* \rangle_F = \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \langle \mathbf{P}, \mathbf{C}^* \rangle_F \tag{15}$$

Moreover, by the Lagrangian's definition, we have $\mu^* = \mathcal{L}(\mathbf{q}^*, \mathbf{C}^*, \mu^*)$. This latter equation combined with (15) and (14) yields:

$$\langle \mathbf{P}^*, \mathbf{C}^* \rangle_F = \max_{\mathbf{C} \in \mathcal{C}} \langle \mathbf{P}^*, \mathbf{C} \rangle_F = \min_{\mathbf{P} \in \mathrm{Conv}(\mathcal{P})} \langle \mathbf{P}, \mathbf{C}^* \rangle_F \tag{16}$$

i.e., $(\mathbf{P}^*, \mathbf{C}^*)$ is a saddle point of $\mathrm{RKP}(\mathcal{P}, \mathcal{C})$. □

Algorithm 1 Cutting set method for $\mathrm{RKP}(\Pi, \mathcal{C})$ with constraint elimination

```
Input: maxIt, $\mathcal{C}$, $\mathcal{P}_0 \subset \Pi$, thd1, thd2
$t, l \leftarrow 0$
$err, \mu_{-1} \leftarrow \infty$
while $t <$ maxIt and $err >$ thd1 and $\frac{\mu_{t-1}-\mu_t}{\mu_{t-1}} > \text{thd1}^2$ do
    Solve (8) to obtain $(\mu_t, \mathbf{C}_t), \mathbf{Q}$
    for $l$ in $\{0, ..., |\mathcal{P}_t| - 1\}$ do
        if $q_l \leq$ thd2 then
            $\mathcal{P}_t \leftarrow \mathcal{P}_t \setminus \{\mathbf{P}_l\}$
            $\mathbf{Q} \leftarrow \mathbf{Q} \setminus \{\mathbf{q}_l\}$
    Find $\mathbf{P}_t \in \mathrm{argmin}_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C}_t \rangle$
    $l \leftarrow \max(l, \langle \mathbf{P}_t, \mathbf{C}_t \rangle)$
    $err \leftarrow (\mu_t - l)/l$
    $\mathcal{P}_{t+1} = \mathcal{P}_t \cup \{\mathbf{P}_t\}$
    $t \leftarrow t + 1$
return $\sum_{l=0}^{|\mathcal{P}_t|-1} q_l \mathbf{P}_l, \mathbf{C}_t$
```

Proposition 3. *Let T be the number of iterations required by Algorithm 1 to reach error $err(T) \leq \text{thd1}$. Then,*

$$T \leq \left(\frac{\mathrm{diam}_\infty(\mathcal{C}) + \mathrm{RKP}(\mathcal{P}_0, \mathcal{C})}{2.\text{thd1}} + 1 \right)^{\dim(\mathcal{C})+1}$$

where $\mathrm{diam}_\infty(\mathcal{C}) := \sup_{\mathbf{C}^1, \mathbf{C}^2 \in \mathcal{C}, i, j} \left| \mathbf{C}^1_{ij} - \mathbf{C}^2_{ij} \right|$ *and* $\dim(\mathcal{C})$ *is the dimension of the affine hull of $\mathcal{C}$. Also, $\forall t \geq 0$, we have that* $0 \leq \mathrm{RKP}(\mathcal{P}_t, \mathcal{C}) - \mathrm{RKP}(\Pi, \mathcal{C}) \leq err(t)$.

Proof. In this proof, we use the notation $\|\mathbf{A}\|_1 = \sum_{ij} |\mathbf{A}_{ij}|$ and $\|\mathbf{A}\|_\infty = \sup_{ij} |\mathbf{A}|_{ij}$. We note that these notations are only used in this proof and do not apply to the rest of the paper, as they do not correspond to the Schatten-1 and ∞ norms.

We apply the result given in (Mutapcic & Boyd, 2009, Section 5.2) to our case. To this end, since our nominal problem corresponds to $\mathcal{P}_0$, we define its feasible set $\mathcal{F}_0$ as:

$$\mathcal{F}_0 = \{(\mu, \mathbf{C}) \in \mathbb{R}_+ \times \mathcal{C} | \mu \leq \min_{\mathbf{P} \in \mathcal{P}_0} \langle \mathbf{P}, \mathbf{C} \rangle_F\}.$$

Also, we define

$$\|(\mu, \mathbf{C})\|_\infty := |\mu| + \|\mathbf{C}\|_\infty. \tag{17}$$

For every $(\mu_1, \mathbf{C}^1), (\mu_2, \mathbf{C}^2) \in \mathcal{F}_0$ and for every constraint, i.e., for every $\mathbf{P} \in \mathcal{P}_0$, we have:

$$\begin{aligned}
&|(\langle \mathbf{P}, \mathbf{C}^1 \rangle_F - \mu_1) - (\langle \mathbf{P}, \mathbf{C}^2 \rangle_F - \mu_2)| \\
&\leq \left| \langle \mathbf{P}, \mathbf{C}^1 \rangle_F - \langle \mathbf{P}, \mathbf{C}^2 \rangle_F \right| + |\mu_1 - \mu_2| \qquad (18) \\
&\leq \|\mathbf{P}\|_1 \left\| \mathbf{C}^1 - \mathbf{C}^2 \right\|_\infty + |\mu_1 - \mu_2| \qquad (19) \\
&\leq \left\| \mathbf{C}^1 - \mathbf{C}^2 \right\|_\infty + |\mu_1 - \mu_2| \qquad (20) \\
&= \left\| (\mu_1, \mathbf{C}^1) - (\mu_2, \mathbf{C}^2) \right\|_\infty \qquad (21)
\end{aligned}$$

where $(\mu_1, \mathbf{C}^1) - (\mu_2, \mathbf{C}^2) := (\mu_1 - \mu_2, \mathbf{C}^1 - \mathbf{C}^2)$. (18) is due to the triangle inequality, followed by the Hölder inequality to obtain (19). Then, since $\mathcal{P}_0 \subset \Pi$ and any matrix in Π has all of its entries bounded by 1, we obtain (20). Lastly, we used definition (17) to obtain (21).

To establish the bound as done in (Mutapcic & Boyd, 2009), we also need to find the radius R of a ball that contains the feasible set $\mathcal{F}_0$, and we consider the affine hull of $\mathcal{C}$ instead of $\mathbb{R}^{m \times n}$ as the space containing $\mathcal{C}$. It is then sufficient to bound the diameter of $\mathcal{F}_0$, denoted $\mathrm{diam}_\infty(\mathcal{F}_0)$ and to take half of the bound for R. To this end, for any $(\mu_1, \mathbf{C}^1), (\mu_2, \mathbf{C}^2) \in \mathcal{F}_0$,

$$\begin{aligned}
\left\| (\mu_1, \mathbf{C}^1) - (\mu_2, \mathbf{C}^2) \right\|_\infty &= \left\| (\mu_1 - \mu_2, \mathbf{C}^1 - \mathbf{C}^2) \right\|_\infty \\
&= \left\| \mathbf{C}^1 - \mathbf{C}^2 \right\|_\infty + |\mu_1 - \mu_2| \\
&\leq \mathrm{diam}_\infty(\mathcal{C}) + |\mu_1 - \mu_2|.
\end{aligned}$$

We have:

$$\mu_1 - \mu_2 \leq \mu_1 \leq \min_{\mathbf{P} \in \mathcal{P}_0} \langle \mathbf{P}, \mathbf{C} \rangle_F \leq \mathrm{RKP}(\mathcal{P}_0, \mathcal{C}).$$

We can obtain this bound also for $\mu_2 - \mu_1$, hence for $|\mu_2 - \mu_1|$. Taking the supremum over $(\mu_1, \mathbf{C}^1), (\mu_2, \mathbf{C}^2) \in \mathcal{F}_0$ (the definition of a diameter), we obtain:

$$\mathrm{diam}_\infty(\mathcal{F}_0) \leq \mathrm{diam}_\infty(\mathcal{C}) + \mathrm{RKP}(\mathcal{P}_0, \mathcal{C}).$$

We can then set radius R as half of the previous upper bound, leading to the bound on the number of iterations T. For the second result of the proposition, for any $t \geq 0$, we have:

$$\min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C} \rangle_F \leq \mathrm{RKP}(\Pi, \mathcal{C}) \leq \mathrm{RKP}(\mathcal{P}_t, \mathcal{C}),$$

where the left inequality is due to taking the maximum over $\mathcal{C}$, while the right one is due to the set inclusion $\mathcal{P}_t \subset \Pi$. Thus,

$$0 \leq \mathrm{RKP}(\mathcal{P}_t, \mathcal{C}) - \mathrm{RKP}(\Pi, \mathcal{C}) \leq \mathrm{RKP}(\mathcal{P}_t, \mathcal{C}) - \min_{\mathbf{P} \in \Pi} \langle \mathbf{P}, \mathbf{C} \rangle_F.$$

In Algorithm 1, the right hand side is equal to $err(t)$, which yields the result. □

Proofs from Section 3.4

We first prove the following lemma that will be helpful in the following proofs.

Lemma 1. *Let c and d be two positive integers. The dual of the linear program*

$$\max_{\mathbf{p}\in\Delta_d,\mu\geq 0} \mu \quad \text{s.t. } \mathbf{G}\mathbf{p} \geq \mu\mathbb{1}_c, \tag{22}$$

is the linear program

$$\min_{\mathbf{q}\in\Delta_c,\nu\geq 0} \nu \quad \text{s.t.} \mathbf{G}^T\mathbf{q} \leq \nu\mathbb{1}_d,$$

Proof. We will transform (22) to a standard LP formulation. To this end, let $\mathbf{v} = (p_1, ..., p_d, \mu)$, i.e the concatenation of $\mathbf{p}$ and μ. Also, we transform the equality condition $\mathbb{1}_d^T\mathbf{p} = 1$ into the two inequalities $\mathbb{1}_d^T\mathbf{p} \leq 1$ and $-\mathbb{1}_d^T\mathbf{p} \leq -1$. We construct the following matrix:

$$\mathbf{F} = \begin{bmatrix} -\mathbf{G} & \mathbb{1}_c \\ \mathbb{1}_d^T & 0 \\ -\mathbb{1}_d^T & 0 \end{bmatrix}$$

Having $c+2$ rows and $d+1$ columns. Then, (22) can be re-written under the standard form:

$$\begin{aligned} \max \quad & \mathbf{e}_{d+1}^T\mathbf{v} \\ \text{s.t} \quad & \mathbf{F}\mathbf{v} \leq \mathbf{e}_{c+1} - \mathbf{e}_{c+2} \\ & \mathbf{v} \geq 0 \end{aligned}$$

where $\mathbf{e}_i$ denotes the vectors of $\mathbb{R}^{d+1}$'s canonical basis. This latter problem has the following dual:

$$\begin{aligned} \min \quad & (\mathbf{e}_{c+1} - \mathbf{e}_{c+2})^T\mathbf{w} \\ \text{s.t} \quad & \mathbf{F}^T\mathbf{w} \geq \mathbf{e}_{d+1} \\ & \mathbf{w} \geq 0 \end{aligned}$$

Using the fact that

$$\mathbf{F}^T = \begin{bmatrix} -\mathbf{G}^T & \mathbb{1}_d & -\mathbb{1}_d \\ \mathbb{1}_c^T & 0 & 0 \end{bmatrix}$$

and denoting $\mathbf{w} = (q_1, ..., q_c, \nu_1, \nu_2)$, and $\mathbf{q} = (q_1, ..., q_c)$, the dual is written:

$$\begin{aligned} \min \quad & \nu_1 - \nu_2 && (23) \\ \text{s.t} \quad & \mathbf{G}^T\mathbf{q} \leq (\nu_1 - \nu_2)\mathbb{1}_d && (24) \\ & \mathbb{1}_c^T\mathbf{q} \geq 1 && (25) \\ & \mathbf{q} \geq 0 && (26) \\ & \nu_1, \nu_2 \geq 0 && (27) \end{aligned}$$

Setting $\nu = \nu_1 - \nu_2$, from (24) and the fact that $\mathbf{G}$ has positive elements (Frobenius products between cost matrices and transport matrices), we have $\nu \geq 0$. Also, for ν^*, $\mathbf{q}^*$ the solution of the dual, we necessarily have $\mathbb{1}_c^T\mathbf{q}^* = 1$. In fact, assuming that $\mathbb{1}_c^T\mathbf{q}^* > 1$ and dividing (24) by $\mathbb{1}_c^T\mathbf{q}^*$, we see that $\nu^{**}, \mathbf{q}^{**}$ defined by $\mathbf{q}^{**} = \frac{\mathbf{q}^*}{\mathbb{1}_c^T\mathbf{q}^*}$ and $\nu^{**} = \frac{\nu^*}{\mathbb{1}_c^T\mathbf{q}^*}$ verify all the constraints, whereas $\nu^{**} < \nu^*$. This latter inequality contradicts the minimality of ν. Hence, the dual formulation is proven. □

Proposition 4 (Finite set $\mathcal{C}$)**.** *Let $\mathcal{C} = \text{Conv}(\{\mathbf{C}_1, ..., \mathbf{C}_d\})$. Then, for $t \geq 0$, solving the problem given in (8) over $\mathcal{P}_t \times \mathcal{C}$ is equivalent to the following linear program*

$$\min_{\mathbf{p}\in\mathbb{R}_+^d} \mathbb{1}_d^T\mathbf{p} \quad \textit{s.t. } \mathbf{G}\mathbf{p} \geq \mathbb{1}_{|\mathcal{P}_t|}, \tag{28}$$

where $\mathbf{G} \in \mathbb{R}^{|\mathcal{P}_t|\times d}$ is defined by $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle$. Moreover,

$$\mathbf{C}^* = \frac{\sum_{k=1}^d p_k^*\mathbf{C}_k}{\sum_{k=1}^d p_k^*}, \qquad \mathbf{P}^* = \frac{\sum_l^{|\mathcal{P}_t|} q_l^*\mathbf{P}_l}{\sum_l^{|\mathcal{P}_t|} q_l^*},$$

where $\mathbf{p}^*$ and $\mathbf{q}^*$ are optimal solutions of (28) and its dual.

Proof. Since $\mathcal{C}$ is the convex hull of matrices $\{\mathbf{C}_1, ..., \mathbf{C}_d\}$, i.e the set of their convex combinations, problem (8) can be formulated as follows:

$$\max_{\mathbf{p}\in\Delta_d,\mu\geq 0} \mu \quad \text{s.t. } \mu \leq \sum_l p_l \langle \mathbf{P}_k, \mathbf{C}_l \rangle_F \qquad \forall 1 \leq k \leq d$$

Let $\mathbf{G}_{kl}$ be the matrix whose elements are: $\mathbf{G}_{kl} = \langle \mathbf{P}_k, \mathbf{C}_l \rangle_F$. The previous problem can be re-written:

$$\max_{\mathbf{p}\in\Delta_d,\mu\geq 0} \mu \quad \text{s.t. } \mathbf{G}\mathbf{p} \geq \mu\mathbb{1}_{|\mathcal{P}_t|}, \tag{29}$$

Since the probability simplex Δ_d can be expressed as:

$$\Delta_d = \left\{ \frac{\mathbf{p}}{\mathbb{1}_d^T\mathbf{p}}; \mathbf{p} \in \mathbb{R}_+^d \setminus \{0\} \right\}$$

the previous problem is equivalent to

$$\max_{\mathbf{p}\in\mathbb{R}_+^d,\mu\geq 0} \mu \quad \text{s.t. } \mathbf{G}\mathbf{p} \geq \mu\mathbb{1}_d^T\mathbf{p}\mathbb{1}_{|\mathcal{P}_t|}.$$

By setting $\mu\mathbb{1}_d^T\mathbf{p} = 1$ (same technique used to derive primal SVM optimization problem as a constrained norm minimization problem), which proves formulation (28). Also, from the change of variables that we made on $\mathbf{p}$, we obtain

$$\mathbf{C}^* = \frac{\sum_{k=1}^d p_k^*\mathbf{C}_k}{\sum_{k=1}^d p_k^*},$$

where $\mathbf{p}^*$ is the solution of Problem (28).

Now we focus on the second part of the proof, to obtain the expression of $\mathbf{P}^*$, the other component of the saddle point. By the result in Lemma 1, denoting $\tilde{\mathbf{q}}^*$ the dual variables of Problem (29), $\tilde{\mathbf{q}}^*$ is a solution to the following dual problem:

$$\min_{\mathbf{q}\in\Delta_{|\mathcal{P}_t|},\nu\geq 0} \nu \qquad \text{subject to } \mathbf{G}^T\mathbf{q} \leq \nu\mathbb{1}_d, \tag{30}$$

By the same argument used to obtain the equivalent formulation (28), Problem (30) is equivalent to:

$$\max_{\mathbf{q}\in\mathbb{R}^d_+} \mathbb{1}^T_{|\mathcal{P}_t|}\mathbf{q} \qquad \text{s.t. } \mathbf{G}^T\mathbf{q} \leq \mathbb{1}_d, \tag{31}$$

where the components of the solution $\tilde{\mathbf{q}}^*$ by normalizing solution $\mathbf{q}^*$ of the previous problem, which yields the expression of $\mathbf{P}^*$. Finally, it is sufficient to notice that Problems (31) and (28) are each the dual of the other, to conclude the proof. □

To proceed for the next proposition, we recall the definition of the set

$$\mathcal{C}_{\mathbf{C}} = \{\mathbf{C} + \mathbf{E}^{\mathbf{M}} \in \mathbb{R}^{m\times n} \,|\, \mathbf{E}^{\mathbf{M}}_{ij} = (\mathbf{x}_i - \mathbf{y}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{y}_j); \\ \mathbf{M} \in \mathcal{S}^d_+(\mathbb{R}); \|\mathbf{M}\|_p \leq r\} \tag{32}$$

for given cost matrix $\mathbf{C}$ and radius $r > 0$.

Proposition 5 (Non centered family of Mahalanobis distances)**.** *For a fixed $\mathbf{C}$, let $\mathcal{C}_{\mathbf{C}}$ be defined as in (32). Then, for $t \geq 0$, solving (8) over $\mathcal{P}_t \times \mathcal{C}_{\mathbf{C}}$, is equivalent to solving the following convex program*

$$\min_{\mathbf{P}\in\mathrm{Conv}(\mathcal{P}_t)} r\|\mathbf{V}_{\mathbf{P}}\|_q + \sum_{ij}(\mathbf{P})_{ij}(\mathbf{C})_{ij}. \tag{33}$$

Moreover, if $\mathbf{P}^*$ is an optimal solution of (33), then $\mathbf{M}^*$ is given by (4) with γ replaced by $\mathbf{P}^*$.

Proof. $\mathcal{C}$ in this case is convex compact, as it is the same as $\mathcal{C}$ presented in Proposition 1, up to a translation by a matrix $\mathbf{C}$.

By Proposition 2, solving Problem (8) is equivalent to solving

$$\min_{\mathbf{P}\in\mathrm{Conv}(\mathcal{P}_t)} \max_{\mathbf{D}\in\mathcal{C}_{\mathbf{C}}} \langle \mathbf{P}, \mathbf{D}\rangle_F$$

However for any matrix $\mathbf{P} \in \mathrm{Conv}\,(\mathcal{P}_t)$, and for $r\mathcal{B}^d_{p+} = \{r\mathbf{M}, \mathbf{M} \in \mathcal{B}^d_{p+}\}$ ($\mathcal{B}^d_{p+}$ is defined as in the proof of Proposition 1), we have:

$$\begin{aligned}&\max_{\mathbf{D}\in\mathcal{C}_{\mathbf{C}}} \langle \mathbf{P}, \mathbf{D}\rangle_F \\ &= \max_{\mathbf{M}\in r\mathcal{B}^d_{p+}} \sum_{ij} \mathbf{P}_{ij}((\mathbf{x}_i - \mathbf{y}_j)\mathbf{M}(\mathbf{x}_i - \mathbf{y}_j) + (\mathbf{C})_{ij}) \\ &= \max_{\mathbf{M}\in\mathcal{B}^d_{p+}} r\sum_{ij} \mathbf{P}_{ij}((\mathbf{x}_i - \mathbf{y}_j)\mathbf{M}(\mathbf{x}_i - \mathbf{y}_j)) + \sum_{ij}\mathbf{P}ij(\mathbf{C})_{ij} \\ &= r\|\mathbf{V}_{\mathbf{P}}\|_q + \sum_{ij}\mathbf{P}_{ij}(\mathbf{C})_{ij}\end{aligned}$$

where in the last line, we used the developments done in Proposition 1, from which we also get the expression of $\mathbf{M}^*$. For the case $p = 2$, we use the result of Corollary 1, where the PSD constraint is not needed. □

For additional details on the experimental evaluations kindly proceed to the following page.

Experimental evaluations

In this section, we add the details needed to reproduce the experiments from the main paper using the code submitted in the supplementary material. We also provide more experimental results for the considered evaluation scenarios and full-size figures presented in a reduced size in the main paper. For all of the experiments, threshold thd2 used for constraint elimination is set to 10^{-12}.

Section 4.1: Convergence and execution time Convergence curves for first two plots are obtained for threshold value thd1= 0. The value for the first threshold is to let the algorithm perform all of the iterations, set to 100. As for the right figure, we set the maximum number of iterations to 1000 and thd1 to 10^{-8}, and we use MOSEK solver to solve the LP formulation, for which we set all tolerance values to 10^{-8}.

Section 4.2: Hypercube We set maxIter to 10, $\mathcal{P}_0$ is set to the uniform distribution and thd1= 10^{-8}. The experiment is reproduced 100 times.

Section 4.3: Stability and noise sensitivity The parameters used in this experiment for all additiional data sets are the same as for the MNIST 0-to-1 dataset and the two Gaussians. The maximum number of iterations of the cutting set method, maxIter is set to 10. $\mathcal{P}_0$ is set to the uniform distribution, thd1= 10^{-20}. The Mahalanobis ball has the radius $r = 0.01$. The 50 cost matrices are created with random Mahalanobis projections and different norms taking values in $(2, 3, 4, 5, 10)$. We also add the cost matrix associated with the squared Euclidean distance. Each cost matrix is divided by its Frobenius norm. The noise sensitivity is computed over 200 runs. In all examples of Figure 1, the sensitivity to noise is correlated to the

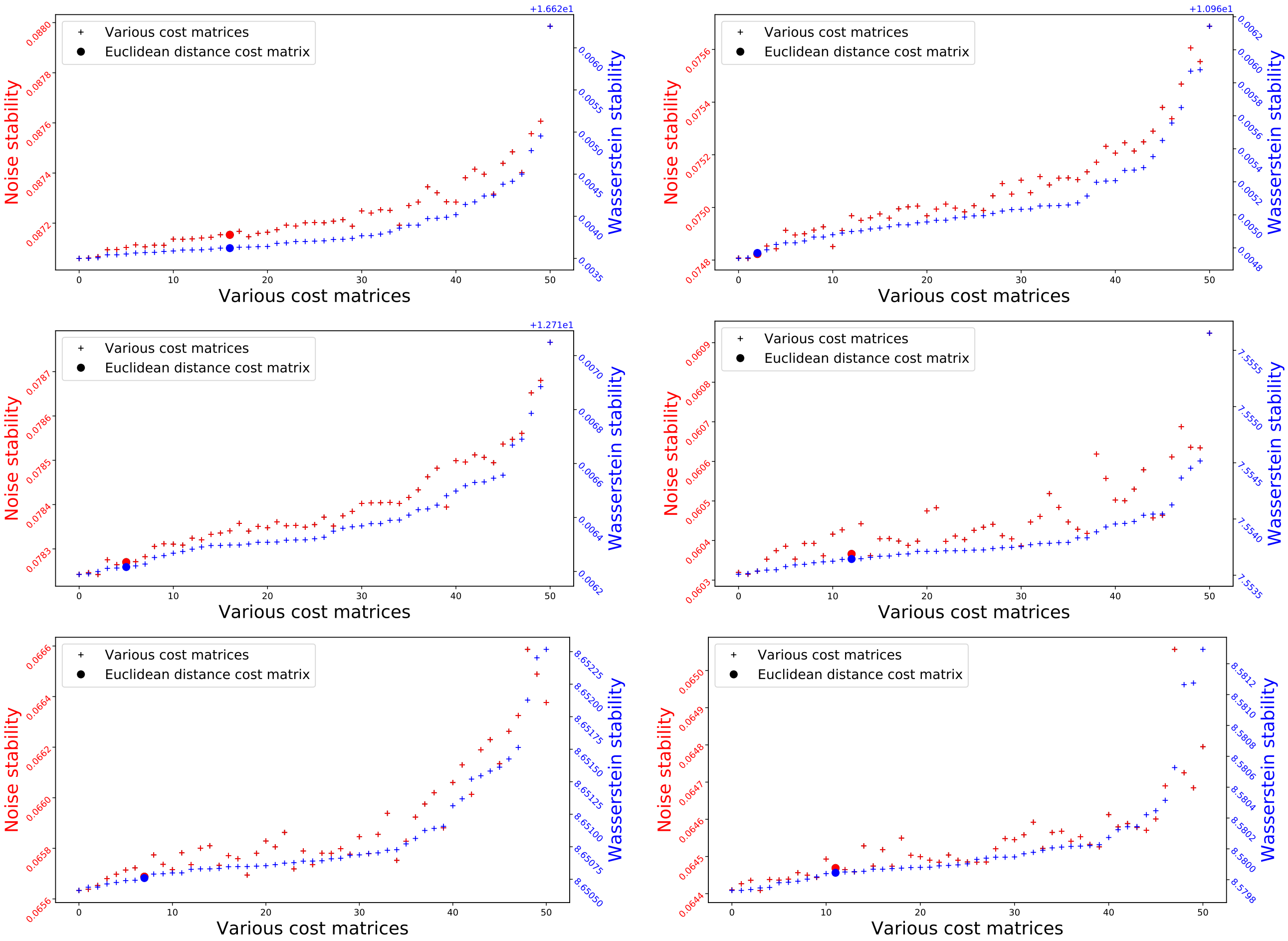

Figure 1. **Left to right, top to bottom**: Gaussians, MNIST 0-to-1, MNIST 3-to-0, MNIST 6-to-5, MNIST 7-to-1, MNIST 7-to-4 data sets. Y-axis (left) is the difference between the OT cost with $\mathbf{C}_i$ and $\mathbf{C}_i + \mathbf{E}^{\mathbf{M}}$. Y-axis (right), the Wasserstein stability defined in Section 3.5. Each column is a different cost matrix, the matrices are ordered by the Wasserstein stability.

stability of the cost matrix. The cost matrix associated with the squared Euclidean distance is often stable and robust to noise which is predictable as it is the most used distance in OT. However, it is never the best cost matrix in terms of our notion of stability.

Section 4.4: Color transfer We use the same setting as above with the following parameters: maxIter is set to 200, thd1$= 10^{-8}$, $r = 0.001$ and we divide each cost matrix element-wise by its corresponding transport cost. Below, we first provide images from the main paper in a bigger size in order to see more fine-grained details.

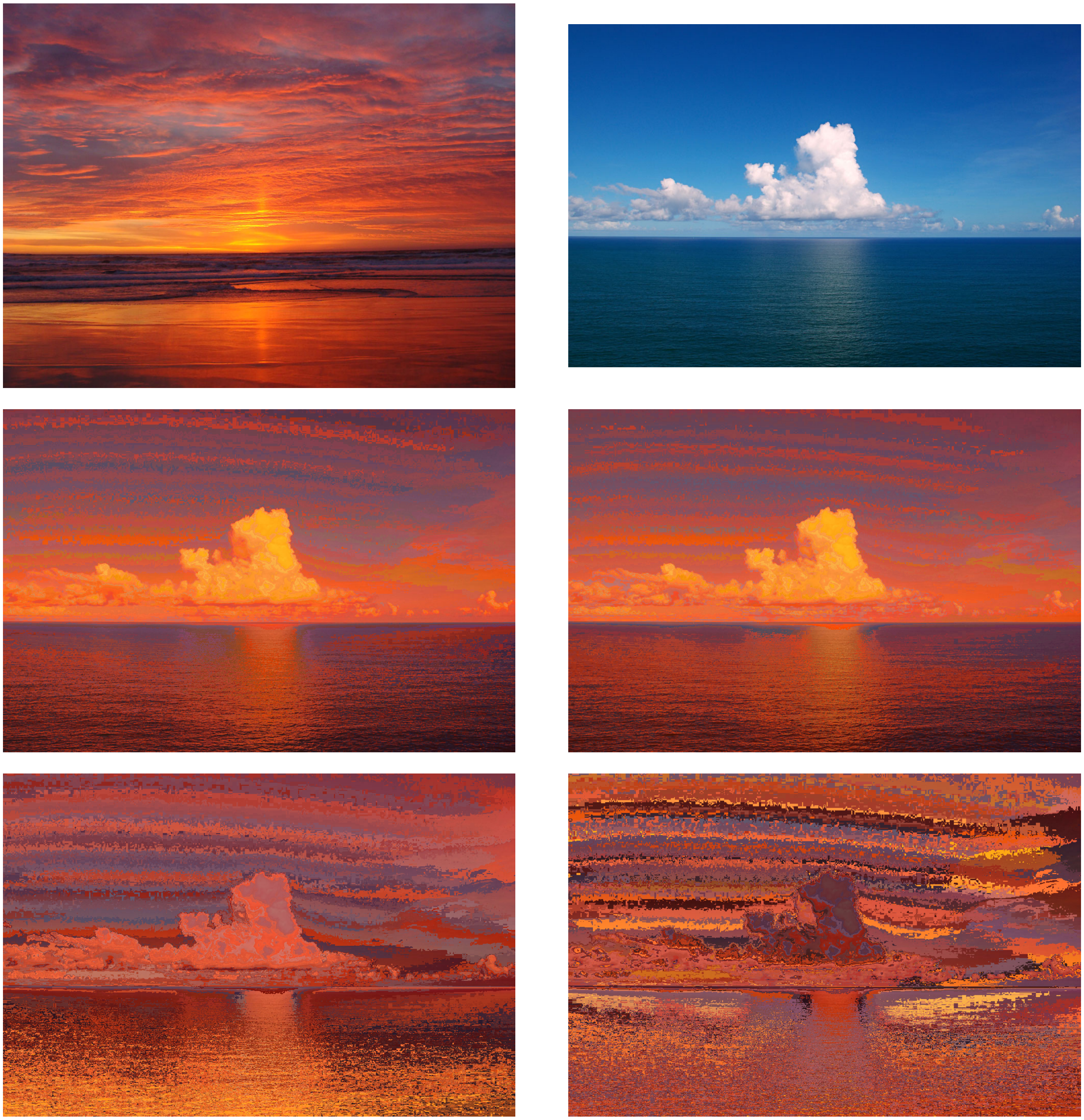

Figure 2. **Top row**: Original images of ocean sunset and ocean sky. **Middle row**: (**left**) most stable cost matrix, (**right**) squared Euclidean based cost matrix. Bottom row: (**left**) least stable Mahalanobis cost matrix, (**right**) least stable cost matrix. Notice the quality difference between the most stable matrix and the squared Euclidean based one in the area just under the cloud.

Figure 3 presents additional visualizations for a new pair of images. The obtained results are in line with experiments shown in the main paper and exhibit similar behaviour.

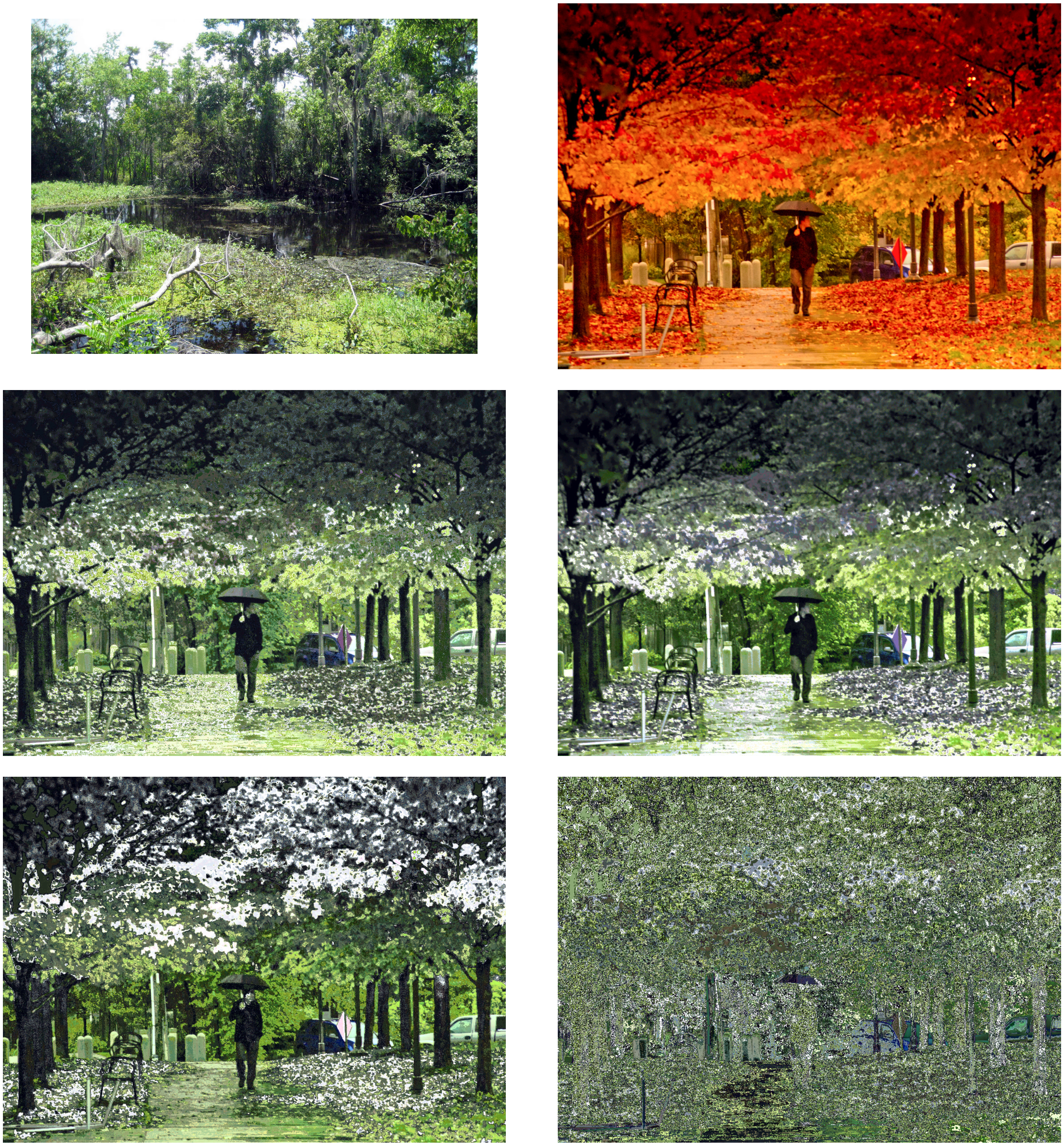

Figure 3. **Top row**: Original images of woods and autumn. **Middle row**: (**left**) most stable cost matrix, (**right**) Euclidean based cost matrix. Bottom row: (**left**) least stable Mahalanobis cost matrix, (**right**) least stable cost matrix.

References

Boyd, S. P. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, 2004. ISBN 978-0-521-83378-3.

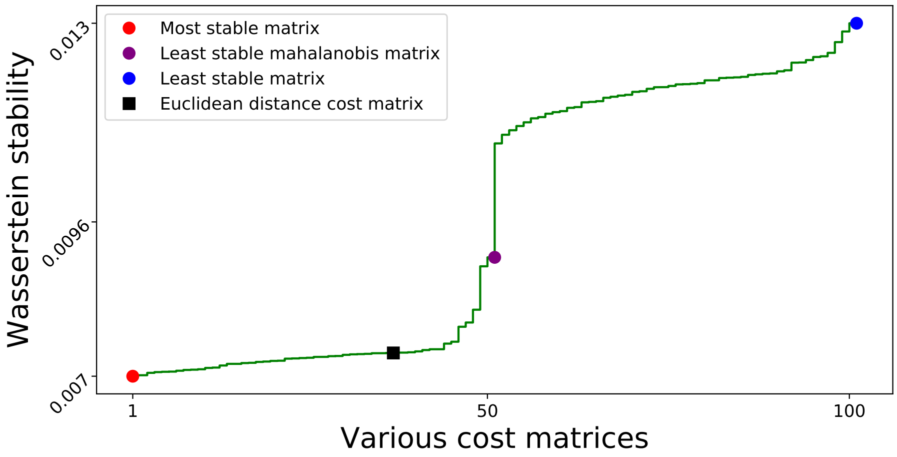

Figure 4. Cost matrices sorted by the Wasserstein stability. The first 50 are Mahalanobis cost matrices, while the last 50 are random cost matrices.

Kantorovich, L. On the translocation of masses. *Doklady of the Academy of Sciences of the USSR*, 37:199–201, 1942.

Magnus, J. R. A representation theorem for (trAp)1/p. Other publications TiSEM, Tilburg University, School of Economics and Management, 1987.

Mutapcic, A. and Boyd, S. P. Cutting-set methods for robust convex optimization with pessimizing oracles. *Optimization Methods and Software*, 24:381–406, 2009.

Paty, F. and Cuturi, M. Subspace robust wasserstein distances. In *ICML*, pp. 5072–5081, 2019.

Sion, M. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.