A Dealing with different samples sizes and dimensions

In the general case where we consider collections of potentially different size ($n \neq m$) of vectors of potentially different dimensionality ($d_x \neq d_y$), we have that $\mathbf{X}$ and $\mathbf{Y}$ are matrices of size $d_x \times m$ and $d_y \times n$, respectively, while $\mathbf{P}$ is of size $d_x \times d_y$.

Note that for the transportation problem in (10), i.e.,

$$\max_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \langle \Gamma, \mathbf{X}^\top \mathbf{P} \mathbf{Y} \rangle,$$

the dimensions d_x, d_y are irrelevant. While the transportation coupling $\mathbf{\Gamma}$ is now of size $m \times n$, the problem is equally meaningful as before, i.e., DOT is well formulated and solved analogously in this case ($n \neq m$) as in the case of equal-sized marginals.

On the other hand, the transformation problem in (10), namely

$$\max_{\mathbf{P} \in \mathcal{F}} \langle \mathbf{X} \Gamma \mathbf{Y}^\top, \mathbf{P} \rangle,$$

is oblivious to n and m. However, when $d_x \neq d_y$, the problem no longer admits a closed form solution in general, so in this case this step requires optimization too. However, there exist various iterative algorithms to solve this problem—known as *unbalanced Procrustes*—efficiently (Gower and Dijksterhuis, 2004; Park, 1991; Viklands, 2006).

B The case $p = 1$

Recall that the Schatten ℓ_1-norm is the nuclear norm $\|A\|_* = \sum_{i=1}^n \sigma_i(A)$. Therefore, the invariance set of interest is now

$$\mathcal{F}_1 = \{\mathbf{P} \mid \|\mathbf{P}\|_* = d\}, \tag{17}$$

which, as before, contains the identity matrix. Note that adding either condition in Lemma 4.1 yields, again, the set of orthonormal matrices.[6] Therefore, in the case one wants to rely on Lemma 4.2 to solve the problem efficiently, this choice of invariance ends up being equivalent to the $p = \infty$ case described in Section 4.1. However, we remark that this equivalence is a consequence of the simplifying assumptions, and that one could still solve this problem with the Frank-Wolfe approach described in Section 4.3, in which case the two cases $p = \infty$ and $p = 1$ would indeed lead to different solutions.

[6] The intersection of the Schatten ℓ_2 and ℓ_∞ norm balls, defined in terms of that of the ℓ_2 and ℓ_∞ vector norm balls, occurs in the extreme points of the latter (see Fig. 1).

C Further Extensions

The framework proposed here can be further extended by considering other transformations that can be easily incorporated into the Procrustes problem framework. For example, scaling and translation can be added on top of orthogonal Procrustes and still yield a closed form solution (Gower and Dijksterhuis, 2004).

D Proofs

Lemma 4.1. *If any of the following conditions holds;*

1. $\forall \mathbf{P} \in \mathcal{F}$, $\mathbf{P}$ *is angle-preserving*
2. $\exists k \geq 0 : \|\mathbf{P}\|_F = k \quad \forall \mathbf{P} \in \mathcal{F}$ *and the matrix* $\mathbf{Y}$ *is* ν*-whitened (i.e.,* $\mathbf{Y} \operatorname{diag}(\mathbf{q})^2 \mathbf{Y}' = \mathbf{I}_d$*).*

then problem (9) *is equivalent to*

$$\max_{\Gamma \in \Pi(\mu,\mathbf{q})} \max_{\mathbf{P} \in \mathcal{F}} \langle \Gamma, \mathbf{X}' \mathbf{P} \mathbf{Y} \rangle = \max_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \max_{\mathbf{P} \in \mathcal{F}} \langle \mathbf{X} \Gamma \mathbf{Y}', \mathbf{P} \rangle \tag{18}$$

Proof. Suppose (1) holds, i.e., $\langle \mathbf{P}\mathbf{x}, \mathbf{P}\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$ for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Then, in particular $\|\mathbf{P}\mathbf{y}\|_2 = \|\mathbf{y}\|_2$ for every $\mathbf{y}^{(j)}$, and therefore:

$$\langle \mathbf{v}, \mathbf{q} \rangle = \sum_{j=1}^m \|\mathbf{P}\mathbf{y}^{(j)}\|_2 = \|\mathbf{y}^{(j)}\|_2$$

and therefore only the first term in (10) depends on $\mathbf{P}$ or Γ, from which the conclusion follows. On the other hand, suppose (2) holds, and let $\tilde{\mathbf{Y}} = \mathbf{Y} \operatorname{diag}(\mathbf{q})$, so that $\tilde{\mathbf{Y}} \tilde{\mathbf{Y}}' = \mathbf{I}_d$. We have:

$$\begin{aligned}
\langle \mathbf{v}, \mathbf{q} \rangle &= \sum_{i=1}^m q_j \|\mathbf{P}\mathbf{y}^{(j)}\|_2^2 \\
&= \sum_{j=1}^m \|\mathbf{P}\mathbf{y}^{(j)} q_j\|_2^2 \\
&= \|\mathbf{P}\tilde{\mathbf{Y}}\|_2^2 \\
&= \langle \mathbf{P}\tilde{\mathbf{Y}}, \mathbf{P}\tilde{\mathbf{Y}} \rangle = \langle \mathbf{P}, \mathbf{P}\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}' \rangle = \|\mathbf{P}\|_F^2 = k^2,
\end{aligned}$$

that is, $\langle \mathbf{v}, \mathbf{q} \rangle$ again does not depend on $\mathbf{P}$. This concludes the proof. □

Lemma 4.2. *Let* $\mathbf{M}$ *be a matrix with SVD decomposition* $\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}'$ *and let* $\Sigma = \operatorname{diag}(\boldsymbol{\sigma})$*, then*

$$\operatorname*{argmax}_{\mathbf{P}:\|\mathbf{P}\|_p \leq k} \langle \mathbf{P}, \mathbf{M} \rangle = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}' \tag{19}$$

where $\mathbf{s}$ *is such that* $\|\mathbf{s}\|_p \leq k$ *and attains* $\mathbf{s}'\boldsymbol{\sigma} = k\|\boldsymbol{\sigma}\|_q$*, for* $\|\cdot\|_q$ *the dual norm of* $\|\cdot\|_p$*.*

Proof. Suppose $\mathbf{P}$ is such that $\|\mathbf{P}\|_p \leq k$, and let $\mathbf{U}_\mathbf{P} \operatorname{diag}(\mathbf{s}) \mathbf{V}'_\mathbf{P}$ be its singular value decomposition. This implies that $\|\mathbf{s}\|_p = \|\mathbf{P}\| \leq k$. In addition,

$$\begin{aligned}\langle \mathbf{P}, \mathbf{M} \rangle &= \langle \mathbf{P}, \mathbf{U}\Sigma\mathbf{V}' \rangle \\ &= \langle \mathbf{U}'\mathbf{P}\mathbf{V}, \Sigma \rangle \\ &= \sum_{i=1}^{d} [\mathbf{U}'\mathbf{P}\mathbf{V}]_{ii} \sigma_i(\mathbf{M}) \\ &= \sum_{i=1}^{d} \mathbf{u}_i \mathbf{P} \mathbf{v}_i \sigma_i(\mathbf{M}) \leq \sum_{i=1}^{d} s_i \sigma_i(\mathbf{M}) = \langle \mathbf{s}, \boldsymbol{\sigma} \rangle\end{aligned}$$

Here, the inequality holds because, by definition of the SVD decomposition, for every i it must hold that $\|\mathbf{u}_i\|_2 = \|\mathbf{v}_i\|_2 = 1$ and

$$\mathbf{u}_i \mathbf{P} \mathbf{v}_i \leq \sup_{\substack{\mathbf{u} \perp \operatorname{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}\} \\ \mathbf{v} \perp \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_{i-1}\}}} \frac{\mathbf{u}'\mathbf{P}\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \leq \sigma_i(\mathbf{P}) = s_i \tag{20}$$

Therefore:

$$\begin{aligned}\sup_{\mathbf{P}: \|\mathbf{P}\|_p \leq k} \langle \mathbf{P}, \mathbf{M} \rangle &\leq \sup_{\mathbf{s}: \|\mathbf{s}\|_p \leq k} \langle \mathbf{s}, \boldsymbol{\sigma} \rangle \\ &= k \sup_{\mathbf{s}: \|\mathbf{s}\|_p \leq 1} \langle \mathbf{s}, \boldsymbol{\sigma} \rangle = k \|\boldsymbol{\sigma}\|_q\end{aligned}$$

where the last equality follows from the definition of dual norm for vectors.

Conversely, take any vector $\mathbf{s}$ with $\|\mathbf{s}\|_p = k$, and define $\tilde{\mathbf{P}}(\mathbf{s}) = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}'$. Clearly, $\|\tilde{\mathbf{P}}(\mathbf{s})\|_p = k$, so the supremum must satisfy:

$$\begin{aligned}\sup_{\mathbf{P}} \langle \mathbf{P}, \mathbf{M} \rangle &\geq \sup_{\mathbf{s}: \|\mathbf{s}\|_p \leq k} \langle \tilde{\mathbf{P}}(\mathbf{s}), \mathbf{M} \rangle \\ &= \sup_{\mathbf{s}: \|\mathbf{s}\|_p \leq k} \langle \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}', \mathbf{U}\Sigma\mathbf{V}' \rangle \\ &= \sup_{\mathbf{s}: \|\mathbf{s}\|_p \leq k} \langle \operatorname{diag}(\mathbf{s}), \Sigma \rangle = k \|\boldsymbol{\sigma}\|_q\end{aligned}$$

Therefore, we conclude that the optimal value of (19) is exactly $k\|\boldsymbol{\sigma}\|_q$.

Furthermore, (20) holds with equality if and only if $(\mathbf{u}_i, \mathbf{v}_i)$ coincide with the left and right singular vectors of $\mathbf{P}$. Thus, any $\mathbf{P}$ maximizing (19) must have the form $\mathbf{P} = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}'$, with $\|\mathbf{s}\|_p \leq k$ and $\langle \mathbf{s}, \boldsymbol{\sigma} \rangle = k\|\boldsymbol{\sigma}\|_q$, as stated.

□

Lemma 4.3. *Consider the Gromov-Wasserstein problem for discrete measures μ and ν (Peyré et al., 2016):*

$$\min_{\Gamma \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i,j,k,l} L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl}) \Gamma_{ij} \Gamma_{kl}, \tag{21}$$

where $(\mathbf{C}^x, \mathbf{p})$ and $(\mathbf{C}^y, \mathbf{q})$ are (intra-space) measured similarity matrices and L is a loss function. For the choice of cosine similarity and squared loss $L(a, b) = \frac{1}{2}|a - b|^2$, Problems (15) and (14) are equivalent.

Proof. For the choice of cosine metric, and assuming without loss of generality that the columns of $\mathbf{X}$ and $\mathbf{Y}$ are normalized, the similarity matrices are given by $\mathbf{C}^x = \mathbf{X}^\top \mathbf{X}$ and $\mathbf{C}^y = \mathbf{Y}^\top \mathbf{Y}$. In addition, let L be the ℓ_2 loss, i.e., $L(a, b) = |a - b|^2$. Then the objective in problem (21) becomes:

$$\begin{aligned}\mathcal{L}(\Gamma) &= \sum_{i,j,k,l} \left(\mathbf{C}^x_{ik} - \mathbf{C}^y_{jl}\right)^2 \Gamma_{ij} \Gamma_{kl} \\ &= \sum_{i,j,k,l} (\mathbf{C}^x_{ik})^2 \Gamma_{ij} \Gamma_{kl} - 2 \sum_{i,j,k,l} (\mathbf{C}^x_{ik} \mathbf{C}^y_{jl}) \Gamma_{ij} \Gamma_{kl} \\ &+ \sum_{i,j,k,l} \left(\mathbf{C}^y_{jl}\right)^2 \Gamma_{ij} \Gamma_{kl}\end{aligned}$$

Since $\Gamma \in \Pi(\mathbf{p}, \mathbf{q})$, the first of these terms becomes

$$\sum_{i,k} (\mathbf{C}^x_{ik})^2 \sum_{j,l} \Gamma_{ij} \Gamma_{jl} = \sum_{i,k} (\mathbf{C}^x_{ik})^2 \mathbf{p}_i \mathbf{p}_k = \mathbf{p}^\top (\mathbf{C}^x)^2 \mathbf{p}$$

where $\mathbf{p}$ is the vector of probabilities in empirical distribution μ, and the last equation follows from the definition of the transportation polytope. Crucially, this term does not depend on Γ anymore. Analogously, the last term in $\mathcal{L}(\Gamma)$ does not depend on Γ either, so

$$\operatorname*{argmin}_{\Gamma \in \Pi(\mathbf{p}, \mathbf{q})} \mathcal{L}(\Gamma) = \operatorname*{argmax}_{\Gamma \in \Pi(\mathbf{p}, \mathbf{q})} \sum_{i,j,k,l} (\mathbf{C}^x_{ik} \mathbf{C}^y_{jl}) \Gamma_{ij} \Gamma_{kl} \tag{22}$$

On the other hand, consider problem (14). The objective it seeks to maximize is

$$\begin{aligned}\|\mathbf{X}\Gamma\mathbf{Y}^\top\|_F^2 &= \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{X}\Gamma\mathbf{Y}^\top \rangle \\ &= \langle \mathbf{X}^\top\mathbf{X}\Gamma, \Gamma\mathbf{Y}\mathbf{Y}^\top \rangle \\ &= \sum_{i=1}^{n} \sum_{l=1}^{m} \left[\mathbf{X}^\top\mathbf{X}\Gamma\right]_{il} \left[\Gamma\mathbf{Y}^\top\mathbf{Y}\right]_{il} \\ &= \sum_{i=1}^{n} \sum_{l=1}^{m} \left[\mathbf{C}^x\Gamma\right]_{il} \left[\Gamma\mathbf{C}^y\right]_{il} \\ &= \sum_{i=1}^{n} \sum_{l=1}^{m} \left(\sum_{k=1}^{n} \mathbf{C}^x_{ik} \Gamma_{kl}\right) \left(\sum_{j=1}^{m} \Gamma_{ij} \mathbf{C}^y_{jl}\right) \\ &= \sum_{i=1}^{n} \sum_{l=1}^{m} \sum_{k=1}^{n} \sum_{j=1}^{m} \mathbf{C}^x_{ik} \Gamma_{kl} \Gamma_{ij} \mathbf{C}^y_{jl}\end{aligned}$$

which is exactly the objective in (22). Hence, Problems (14) and (21) are indeed equivalent. □

E The Algorithm

Algorithm 1 Optimal Transport with Invariances

Inputs:
- Data matrices and histograms $(\mathbf{X}, \mathbf{p})$, $(\mathbf{Y}, \mathbf{q})$
- Order of invariance p and radius k_p
- Initial/final entropy regularization λ_0 and $\underline{\lambda}$, decay rate η

// Initialize feasible transformation in $\mathcal{F}_p$
$\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \text{SVD}(\textsc{RandomMatrix}(d \times d))$
$\boldsymbol{\sigma} \leftarrow \text{diag}\,(\mathbf{\Sigma})$
$\mathbf{s} \leftarrow k_p \cdot \boldsymbol{\sigma} / \|\boldsymbol{\sigma}\|_p$
$\mathbf{P} = \mathbf{U}\,\text{diag}\,(\mathbf{s})\,\mathbf{V}^\top$
$\lambda \leftarrow \lambda_0$
while not converged **do**
  // Compute distances w.r.t. current mapping $\mathbf{P}$
  $\mathbf{C_P} \leftarrow \textsc{PairwiseDistances}(\mathbf{X}, \mathbf{PY})$
  // Solve regularized OT via Sinkhorn iterations
  $\mathbf{b} \leftarrow \mathbb{1}, \quad \mathbf{K} \leftarrow \exp\{-\mathbf{C_P}/\lambda\}$
  while not converged **do**
    $\mathbf{a} \leftarrow \mathbf{p} \oslash \mathbf{Kb}$
    $\mathbf{b} \leftarrow \mathbf{q} \oslash \mathbf{K}^\top \mathbf{a}$
  end while
  $\mathbf{\Gamma} \leftarrow \text{diag}\,(\mathbf{a})\,\mathbf{K}\,\text{diag}\,(\mathbf{b})$
  // Solve generalized Procrustes problem
  $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^\top \leftarrow \text{SVD}(\mathbf{X\Gamma Y}^\top)$
  $\boldsymbol{\sigma} \leftarrow \text{diag}\,(\mathbf{\Sigma})$
  $q \leftarrow \frac{p}{p-1}$
  $\mathbf{s} \leftarrow k_p \cdot \boldsymbol{\sigma}^{q-1} / \|\boldsymbol{\sigma}^{q-1}\|_p$
  $\mathbf{P} = \mathbf{U}\,\text{diag}\,(\mathbf{s})\,\mathbf{V}^\top$
  // Anneal entropy regularization
  $\lambda \leftarrow \max\{\lambda * \eta, \underline{\lambda}\}$
end while
return $\mathbf{\Gamma}, \mathbf{P}$

F Solving very large problems

While direct application of Algorithm 1 leads to high-quality solutions for small and mid-sized problems, scaling up to very large sets of points—e.g., hundreds of thousands of word embeddings in the word translation application—can be prohibitive.

We address this issue by dividing the problem into two phases. In the first stage, we solve a smaller problem (by taking a subsample of k points on each domain thus leading to smaller Γ and faster OT solution, but same size of $\mathbf{P}$). Once the first phase reaches convergence, we use the solution $\mathbf{P}^*$ of the first stage to initialize the full-size problem. Note that while this might resemble other approaches that also consider a reduced set of points in their initialization step (Conneau et al., 2018; Grave et al., 2018), a crucial difference is that here we rely on the same optimization problem (16) in both stages, although with different problem sizes.

We experimented with various choices of parameter k, and observed that the algorithm is remarkably robust to the choice of this parameter. We conjecture that the ordering in which word embeddings are provided (higher-frequency words first, in every language) helps ensure that the solution of the initial problem of reduced size is consistent with the full-size problem.[7]

While the end performance is consistent regardless of the choice of sub-sample size k, there is naturally a trade-off in run time of the two stages. While solving a smaller initial problem is obviously faster, we observed that in such cases the second stage required more iterations to converge, suggesting that the initial $\mathbf{P}^*$ fed into the second stage was of lower quality (further from the optimal for the full-size problem). In the results presented in Section 5.2, we take k as large as possible while keeping the time-per-iteration reasonable: $k = 5000$.

Note that this strategy of *bootstrapping* solutions of smaller problems can be applied repeatedly, to increasingly grow the problem size over multiple stages. While we did not require to do so in our experiments, it might be an appealing approach for solving extremely large problems.

[7] This, in fact, points to an issue mostly ignored in previous work on this task: the order of the word embeddings *leaks* important—albeit noisy—correspondence information, which various methods presented as 'fully-unsupervised' seem to rely on one way or another, yet rarely acknowledge it.