Supplementary materials

August 8, 2021

A Introduction of notation

We summarise below the (relatively standard) notation conventions which we adopt throughout this article.

- Matrices and tensors are denoted in upper case, e.g. X, U, V, in contrast to vectors which are written in lower case, e.g. x, u, v.
- For a matrix $X \in \mathbb{R}^{m\times n}$, we index its elements X_{ij} and write X_i for the ith column as a m-dimensional vector.
- For a tensor $X \in \mathbb{R}^{n_1\times\cdots\times n_d}$, we index its elements $X_{i_1,\ldots,i_d}$ and write $X_{(i)}$ for its matricisation [12] along mode i, which is a matrix of dimensions $n_i \times n_1\cdots n_{i-1}n_{i+1}\cdots n_d$.
- We write the inner product for vectors (x, y) as $\langle x, y\rangle = \sum_i x_i y_i$, for matrices (X, Y) as $\langle X, Y\rangle = \sum_{ij} X_{ij}Y_{ij}$ and so on for tensors.
- We denote elementwise multiplication by $\odot$, and outer product of vectors by $\otimes$. Unless otherwise specified, for x, y vectors and A a matrix of appropriate dimensions, by writing Ax we refer to the matrix-vector product, and by xy and x/y we refer to the elementwise product and quotient respectively.
- Following the notation of [12], we write the mode-k product of a tensor $A \in \mathbb{R}^{n_1\times\cdots\times n_d}$ and a matrix $B \in \mathbb{R}^{m\times n_k}$ as

$$\begin{aligned}(A \times_k B)_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_d} \\ = \sum_{i_k=1}^{n_k} A_{i_1,\ldots,i_d} B_{j,i_k},\end{aligned}$$

  and this is a tensor of dimensions $n_1 \times\cdots\times n_{k-1} \times m \times n_{k+1} \times\cdots\times n_d$.
- For non-negative matrices or tensors α and β, we write the entropy as $E(\alpha) = \langle\alpha, \log(\alpha) - 1\rangle$, the relative entropy $\mathrm{H}(\alpha|\beta) = \langle\alpha, \log(\alpha/\beta) - 1\rangle$ and the generalised Kullback-Leibler divergence as $\mathrm{KL}(\alpha|\beta) = \langle\alpha, \log(\alpha/\beta)\rangle - \langle\alpha, \mathbf{1}\rangle + \langle\beta, \mathbf{1}\rangle$.
- For a discrete metric space $\mathcal{X}$, we write $\mathcal{P}(\mathcal{X})$ and $\mathcal{M}_+(\mathcal{X})$ to be respectively the set of probability distributions and positive measures supported on $\mathcal{X}$.

B Wasserstein-NMF as a special case of WTF

From the framework introduced in Sections 3.3 and 3.4, we may recover the NMF method described by Rolet et al. [20] when we consider matrices as 2-way tensors. Let $X \in \mathbb{R}^{m \times n}$. In the context of NMF, columns of X correspond to observations, so we may take $\Phi(X, \hat{X})$ to be a Wasserstein loss along the columns of its arguments. We take $S_{i_1,i_2} = \delta_{i_1,i_2}$ to be fixed, and so for two factor matrices U, V we have

$$S[U, V] = \sum_{i=1}^{r} U_i \otimes V_i = UV^\top.$$

Thus, (14) becomes

$$\min_{U,V} \Phi(X, UV^\top) + \rho_1 E_{\Sigma_1}(U) + \rho_2 E_{\Sigma_2}(V),$$

which coincides (albeit with differing notation) with the Wasserstein-NMF problem introduced by Rolet et al. [20]. Proposition 2 gives the dual problem for the subproblems in factor matrices U and V respectively.

C Convex duality

For further background on the techniques involved, we refer the reader to the existing literature on variational problems involving optimal transport [5, 6, 8].

Definition 1 (Legendre transform)**.** *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper function, i.e. one that is not identically $+\infty$. The Legendre transform of f is defined for $u \in \mathbb{R}^n$ as*

$$f^*(u) = \sup_{x \in \mathbb{R}^n} \langle x, u \rangle - f(x).$$

Furthermore, $f^{**} = f$ if and only if f is convex and lower semicontinuous.

Theorem 1 (Fenchel-Rockafellar theorem [19])**.** *Let E, F be (finite or infinite dimensional) real vector spaces, and E^*, F^* their respective topological dual spaces. Let $f : E \to (-\infty, \infty]$, $g : F \to (-\infty, \infty]$ be proper (not identically $+\infty$), convex, lower-semicontinuous (sublevel sets are closed) functions. Let $A : E \to F$ be a continuous linear operator. Consider the convex minimisation problem*

$$\min_{x \in E} f(x) + g(Ax) \tag{P}$$

Then (P) has a corresponding dual problem (P)*

$$\sup_{y \in F^*} -f^*(A^*y) - g^*(-y), \tag{P*}$$

where A^* is the adjoint of A, and $f^*(\cdot) = \sup_{x \in E} \langle x, \cdot \rangle - f(x)$, $g^*(\cdot) = \sup_{y \in F} \langle y, \cdot \rangle - g(y)$ are the Legendre transforms *of f and g, defined over E^* and F^* respectively.*

In general the dual problem (P*) provides a lower bound on the solution to the primal problem (P): $p^\star \geq d^\star$. However, the following simple condition is sufficient for equality to hold in a finite dimensional setting.

Theorem 2 (Condition for strong duality in finite dimensions, adapted from [19])**.** *If E, F are finite-dimensional and there exists some x in the relative interior of the feasible set, then $p^\star = d^\star$.*

D Proofs

Proof of Proposition 2. We write the primal problem as (16):

$$\min_{A^{(k)}} \Phi(X, S[A^{(1)}, \dots, A^{(d)}]) + \rho_k E_{\Sigma_k}(A^{(k)}).$$

Now let us substitute the definition of the Legendre transform of Φ, and we take formally an $\inf - \sup$ exchange:

$$\begin{aligned}
&\min_{A^{(k)}} \sup_{U} \left[\left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - \Phi^*(X, U) \right] + \rho_k E_{\Sigma_k}(A^{(k)}) \\
&= \sup_{U} -\Phi^*(X, U) + \min_{A^{(k)}} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle + \rho_k E_{\Sigma_k}(A^{(k)}) \\
&= \sup_{U} -\Phi^*(X, U) - \rho_k \max_{A^{(k)}} \left[\frac{-1}{\rho_k} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - E_{\Sigma_k}(A^{(k)}) \right].
\end{aligned}$$

We now note the identities $\langle A, B \times_k C \rangle = \langle A \times_k C^\top, B \rangle$ and $\langle A \times_k B, C \rangle = \langle B A_{(k)}, C_{(k)} \rangle$ so:

$$\begin{aligned}
\left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle &= \left\langle U \times_{j \geq k+1} (A^{(j)})^\top, S \times_{j \leq k} A^{(j)} \right\rangle \\
&= \left\langle U \times_{j \geq k+1} (A^{(j)})^\top, S \times_{j \leq k-1} A^{(j)} \times_k A^{(k)} \right\rangle \\
&= \left\langle A^{(k)} \left[S \times_{j \leq k-1} A^{(j)} \right]_{(k)}, \left[U \times_{j \geq k+1} (A^{(j)})^\top \right]_{(k)} \right\rangle \\
&= \left\langle A^{(k)}, \left[U \times_{j \geq k+1} (A^{(j)})^\top \right]_{(k)} \left[S \times_{j \leq k-1} A^{(j)} \right]_{(k)}^\top \right\rangle \\
&= \left\langle A^{(k)}, \Xi^{(k)}(U) \right\rangle,
\end{aligned}$$

where $\Xi^{(k)}(U) = \left[U \times_{j \geq k+1} (A^{(j)})^\top \right]_{(k)} \left[S \times_{j \leq k-1} A^{(j)} \right]_{(k)}^\top$. Thus,

$$\begin{aligned}
&\max_{A^{(k)}} \frac{-1}{\rho_k} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - E_{\Sigma_k}(A^{(k)}) \\
&= \max_{A^{(k)}} \left\langle A^{(k)}, \frac{-1}{\rho_k} \Xi^{(k)}(U) \right\rangle - E_{\Sigma_k}(A^{(k)}) \\
&= E^*_{\Sigma_k} \left(\frac{-1}{\rho_k} \Xi^{(k)}(U) \right).
\end{aligned}$$

Thus we have

$$\sup_{U} -\Phi^*(X, U) - \rho_k E^*_{\Sigma_k} \left(\frac{-1}{\rho_k} \Xi^{(k)}(U) \right).$$

Strong duality holds by application of the Fenchel-Rockafellar theorem.

Let the value of U at optimality be $U^\star$. Then the corresponding factor matrix $(A^{(k)})^\star$ must be the solution of

$$\max_{A^{(k)}} \left\langle A^{(k)}, \frac{-1}{\rho_k} \Xi^{(k)}(U^\star) \right\rangle - E_{\Sigma_k}(A^{(k)}).$$

If $\Sigma_k = \{\}$, then $E_{\Sigma_k}(x) = \langle x, \log(x) - 1 \rangle$. We are unconstrained and have the sum of an affine and a convex term. Differentiating, we find the first-order optimality condition

$$A^{(k)^\star} = \exp\left(\frac{-1}{\rho_k} \Xi^{(k)}(U^\star) \right).$$

If $\Sigma_k = \{A^{(k)} : \left\langle A^{(k)}, \mathbf{1} \right\rangle = 1\}$, then the problem is subject to a simplex constraint. At optimality, therefore, the gradient must be orthogonal to the simplex (parallel to $\mathbf{1}$):

$$\begin{aligned} \frac{-1}{\rho_k} \Xi^{(k)}(U^\star) - \log(A^{(k)^\star}) &= c\mathbf{1} \\ \implies A^{(k)^\star} &= \exp(-c) \exp\left(\frac{-1}{\rho_k} \Xi^{(k)}(U^\star) \right). \end{aligned}$$

Since $\left\langle A^{(k)}, \mathbf{1} \right\rangle = 1$, we conclude that $\exp(c) = \exp\left(\frac{-1}{\rho_k} \Xi^{(k)}(U^\star) \right)$. In the cases where Σ_k requires row or column normalisation, applying an identical argument row- or columnwise leads to an analogous result, where we normalise the output row- or columnwise. □

Proof of Proposition 3. As in the proof of Proposition 2, we introduce the Legendre transform of Φ and carry out an $\inf - \sup$ exchange.

$$\begin{aligned} &\min_S \Phi(X, S[A^{(1)}, \ldots, A^{(d)}]) + \rho_0 E_{\Sigma_0}(S) \\ &= \min_S \sup_U \left[\left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - \Phi^*(X, U) \right] + \rho_0 E_{\Sigma_0}(S) \\ &= \sup_U -\Phi^*(X, U) + \min_S \left[\left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle + \rho_0 E_{\Sigma_0}(S) \right] \\ &= \sup_U -\Phi^*(X, U) - \rho_0 \max_S \left[\frac{-1}{\rho_0} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - E_{\Sigma_0}(S) \right]. \end{aligned}$$

Now note that using the identities presented in the proof of Proposition 2:

$$\begin{aligned} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle &= \left\langle U \times_1 (A^{(1)})^\top \times \cdots \times_d (A^{(d)})^\top, S \right\rangle \\ &= \langle S, \Omega(U) \rangle \end{aligned}$$

where $\Omega(U) = U \times_1 (A^{(1)})^\top \times \cdots \times_d (A^{(d)})^\top$. Thus,

$$\begin{aligned} &\max_S \frac{-1}{\rho_0} \left\langle U, S \times_1 A^{(1)} \times \cdots \times_d A^{(d)} \right\rangle - E_{\Sigma_0}(S) \\ &= \max_S \left\langle S, \frac{-1}{\rho_0} \Omega(U) \right\rangle - E_{\Sigma_0}(S) \\ &= E^*_{\Sigma_0} \left(\frac{-1}{\rho_0} \Omega(U) \right). \end{aligned}$$

Thus, the dual problem is

$$\sup_U -\Phi^*(X,U) - \rho_0 E^*_{\Sigma_0}\left(\frac{-1}{\rho_0}\Omega(U)\right).$$

As before, strong duality holds by application of the Fenchel-Rockafellar theorem.

Let the value of U at optimality be $U^\star$. Then the corresponding core tensor $S^\star$ must be the solution of

$$\max_S \left\langle S, \frac{-1}{\rho_0}\Omega(U^\star)\right\rangle - E_{\Sigma_0}(S).$$

Following the previous argument given in the Proof of Proposition 2, we find that

$$S^\star = \begin{cases} \exp\left(\frac{-1}{\rho_0}\Omega(U^\star)\right), & \Sigma_0 = \{\}, \\ \frac{\exp\left(\frac{-1}{\rho_0}\Omega(U^\star)\right)}{\left\langle \exp\left(\frac{-1}{\rho_0}\Omega(U^\star)\right), \mathbf{1}\right\rangle}, & \Sigma_0 = \{S : \langle S, \mathbf{1}\rangle = 1\} \end{cases}.$$

□

Proof of Proposition 4. For marginal distributions p, q, we write an alternative form of the semi-balanced optimal transport problem (7) as

$$\mathrm{OT}^\lambda_\varepsilon(p,q) = \inf_{\gamma:\gamma\mathbf{1}=p} \varepsilon \mathrm{H}(\gamma|K) + \lambda \mathrm{KL}(\gamma^\top \mathbf{1}|q).$$

We seek the Legendre transform in the second argument q. Introduce u the dual variable of q and α the Lagrange multiplier for the constraint $\gamma\mathbf{1} = p$, then exchange the inf and sup.

$$\begin{aligned} \mathrm{OT}^{\lambda\,*}_\varepsilon(p,u) &= \sup_q \langle u, q\rangle - \inf_{\gamma\mathbf{1}=p} \left[\varepsilon \mathrm{H}(\gamma|K) + \lambda \mathrm{KL}(\gamma^\top\mathbf{1}|q)\right] \\ &= \inf_\alpha \sup_{q,\gamma} \langle u, q\rangle - \varepsilon \mathrm{H}(\gamma|K) \\ &\qquad - \lambda \mathrm{KL}(\gamma^\top\mathbf{1}|q) + \langle \alpha, \gamma\mathbf{1} - p\rangle. \end{aligned}$$

Use first order condition for q:

$$\begin{aligned} \frac{\partial}{\partial q}(\cdot) = u - \lambda\left(1 - \frac{\gamma^\top\mathbf{1}}{q}\right) = 0 \\ \Rightarrow q = (\gamma^\top\mathbf{1})\left(\frac{\lambda}{\lambda - u}\right), \end{aligned}$$

where multiplication is elementwise. Substituting back, we find that

$$\begin{aligned} \inf_\alpha \sup_\gamma \left\langle u, \left(\gamma^\top\mathbf{1} \odot \frac{\lambda}{\lambda - u}\right)\right\rangle - \varepsilon \mathrm{H}(\gamma|K) \\ - \lambda\left\langle \gamma^\top\mathbf{1}, \log\left(\frac{\lambda - u}{\lambda}\right) - \gamma^\top\mathbf{1} + q\right\rangle + \langle \alpha, \gamma\mathbf{1} - p\rangle. \end{aligned}$$

Differentiating with respect to γ, (and after some involved algebra) we find the first order condition for γ to be

$$\gamma_{ij} = \exp\left(\frac{\alpha_i}{\varepsilon}\right) K_{ij} \left(\frac{\lambda}{\lambda - u_j}\right)^{\lambda/\varepsilon}.$$

Substituting back and finally differentiating in α we find that

$$\alpha = \varepsilon \log\left(\frac{p}{K(\frac{\lambda}{\lambda-u})^{\lambda/\varepsilon}}\right) = \varepsilon \log\left(\frac{p}{Kf}\right),$$

where $f = \left(\frac{\lambda}{\lambda-u}\right)^{\lambda/\varepsilon}$ for brevity. With all this, the Legendre transform is

$$\begin{aligned}
\mathrm{OT}_\varepsilon^{\lambda^*}(p,u) &= -\left\langle p, \varepsilon \log\left(\frac{p}{Kf}\right)\right\rangle + \left\langle K^\top \frac{p}{Kf}, \varepsilon f\right\rangle \\
&= -\varepsilon\left\langle p, \log\left(\frac{p}{Kf}\right)\right\rangle + \varepsilon\langle p, \mathbf{1}\rangle.
\end{aligned}$$

The relationship between the primal and dual variables is therefore

$$\begin{aligned}
q &= \left(\frac{\lambda}{\lambda - u}\right) f \odot K^\top \frac{p}{Kf}, \\
\gamma &= \exp\left(\frac{\alpha}{\varepsilon}\right) Kf.
\end{aligned}$$

□

E Gibbs kernel convolution

Proposition 1 (Convolution with tensor-valued Gibbs kernel)**.** *In the context of Proposition 6, the Gibbs kernel $K = e^{-C/\varepsilon}$ is a $2d$-way tensor that decomposes multiplicatively:*

$$\begin{aligned}
K_{i_1,\ldots,i_d,j_1,\ldots,j_d} &= e^{-C_{i_1,\ldots,i_d,j_1,\ldots,j_d}/\varepsilon} \\
&= K^{(1)}_{i_1,j_1} \cdots K^{(d)}_{i_d,j_d}.
\end{aligned}$$

Furthermore, note that in the case of vector-valued input, the formula (25) *involves a matrix-vector convolution of the form $s \mapsto Ks$. In our setting, for $S \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ the corresponding operation is a convolution along all modes that has the following decomposition:*

$$\begin{aligned}
(KS)_{i_1,\ldots,i_d} &= \sum_{j_1,\ldots,j_d} K_{i_1,\ldots,i_d,j_1,\ldots,j_d} S_{j_1,\ldots,j_d} \\
&= S \times_{i=1}^{d} K^{(i)}.
\end{aligned}$$

Code

An implementation of the methods described in this paper is available at `https://github.com/zsteve/wtf`