Notes for the Optimal Transport class

Quentin Bertrand, Mathurin Massias, Titouan Vayer

Last updated: February 10, 2026

Contents

1 Review of OT, EOT, and the Sinkhorn Algorithm

Let $(\mathbf{x}_i)_{i=1}^n$ and $(\mathbf{y}_j)_{j=1}^m$ be two sets of points in $\mathbb{R}^d$. Let $\mathbf{a} \in \mathbb{R}^n$ and $\mathbf{P} \in \mathbb{R}^m$ be the mass vectors associated with these points, and let $\mathbf{C} = (d(\mathbf{x}_i, \mathbf{y}_j))_{i,j \in [n] \times [m]} \in \mathbb{R}^{n \times m}$ be the cost matrix between the points.

Definition 1.1 (OT)**.** *The optimal transport (OT) problem is written as:*

$$\min_{\mathbf{P} \geq 0} \langle \mathbf{C}, \mathbf{P} \rangle, \quad s.t. \quad \mathbf{P}\mathbf{1}_m = \mathbf{a}, \quad \mathbf{P}^T \mathbf{1}_n = \mathbf{P} \tag{1.1}$$

A dual formulation (Dual OT) is given by:

$$\max_{\mathbf{f}, \mathbf{g}} \langle \mathbf{a}, \mathbf{f} \rangle + \langle \mathbf{P}, \mathbf{g} \rangle, \quad s.t. \quad \forall (i,j)\ \mathbf{f}_i + \mathbf{g}_j \leq \mathbf{C}_{i,j} \tag{1.2}$$

OT is a linear problem but is nevertheless difficult to solve efficiently.

Definition 1.2 (EOT)**.** *The entropic regularized optimal transport problem (EOT) is written (for $\varepsilon > 0$) as:*

$$\min_{\mathbf{P} \geq 0} \langle \mathbf{C}, \mathbf{P} \rangle + \varepsilon \sum_{i,j} \mathbf{P}_{i,j} (\log \mathbf{P}_{i,j} - 1), \quad s.t. \quad \mathbf{P}\mathbf{1}_n = \mathbf{a}, \quad \mathbf{P}^T \mathbf{1}_m = \mathbf{P} \tag{1.3}$$

The main motivation for entropic regularization is to remove the constraints in the dual problem, which then takes the following form (Dual EOT):

$$\max_{\mathbf{f}, \mathbf{g}} \langle \mathbf{a}, \mathbf{f} \rangle + \langle \mathbf{P}, \mathbf{g} \rangle - \varepsilon \sum_{i,j} \exp \left(\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{C}_{i,j}}{\varepsilon} \right) \tag{1.4}$$

EOT may appear more difficult to solve than OT, but in fact it is simpler since there are no longer any constraints. One can use the Sinkhorn algorithm; however, the resulting transport plan is never exactly the same as that of OT due to the regularization. In particular, a major difference with OT is that the optimal transport plan P^* is dense, whereas there always exists a solution to OT with at most $m+n-1$ nonzero entries.

Definition 1.3 (Sinkhorn Algorithm)**.** *Denoting by* $\mathbf{K} = \exp(-\mathbf{C}/\varepsilon)$ *the* Gibbs kernel *(exponential being applied pointwise), the Sinkhorn algorithm consists in iterating the following two steps until convergence:*

$$\mathbf{u} \leftarrow \mathbf{a}/(\mathbf{K}\mathbf{v}), \quad \mathbf{v} \leftarrow \mathbf{P}/(\mathbf{K}^T\mathbf{u})$$

In practice, the algorithm converges very quickly (linear convergence). The dual potentials of the EOT dual problem are recovered via the relations $f = \varepsilon \log u$ *and* $g = \varepsilon \log v$*.*

Remark 1.4. *The Sinkhorn algorithm solves a* matrix scaling problem*, that is, it finds vectors* $\mathbf{u}$ *and* $\mathbf{v}$ *such that* $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$ *has rows summing to* $\mathbf{a}$ *and columns summing to* $\mathbf{P}$*.*
Step 1 of the algorithm updates $\mathbf{u}$ *to have correct row sums, while step 2 updates* $\mathbf{v}$ *to have correct column sums.*

Theorem 1.5. *The Sinkhorn algorithm is equivalent to an alternating maximization scheme for the EOT dual problem.*

Proof. Let us denote $\operatorname{dual}(\mathbf{f},\mathbf{g}) = \langle \mathbf{a}, \mathbf{f}\rangle + \langle \mathbf{P}, \mathbf{g}\rangle - \varepsilon \sum_{i,j} \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{C}_{i,j}}{\varepsilon}\right)$.
If we seek to maximize this expression with respect to $\mathbf{f}$, the function is concave and it suffices to set the gradient to zero. We obtain

$$(\nabla_{\mathbf{f}} \operatorname{dual}(\mathbf{f},\mathbf{g}))_i = \mathbf{a}_i - \sum_j \exp\left(\frac{\mathbf{g}_j - \mathbf{C}_{i,j}}{\varepsilon}\right) \exp\left(\mathbf{f}_i/\varepsilon\right) = 0 \tag{1.5}$$

$$\iff \exp(\mathbf{f}_i/\varepsilon) = \mathbf{a}_i / \left(\sum_j \exp(\mathbf{g}_j/\varepsilon)\exp(\mathbf{C}_{i,j}/\varepsilon)\right) \tag{1.6}$$

$$\iff \mathbf{u}_i = \mathbf{a}_i/(\mathbf{K}\mathbf{v})_i \tag{1.7}$$

With $\mathbf{u} = \exp(\mathbf{f}/\varepsilon)$ and $\mathbf{v} = \exp(\mathbf{g}/\varepsilon)$. By applying the same reasoning while maximizing with respect to $\mathbf{g}$, we obtain $\mathbf{v} = \mathbf{P}/(\mathbf{K}^T\mathbf{u})$. □

2 Unbalanced Optimal Transport

Remark 2.1. *In the following, we denote* $i_{\{\mathbf{a}\}} : \mathbf{x} \mapsto \begin{cases} 0 & \textit{if } \mathbf{x} = \mathbf{a}, \\ +\infty & \textit{otherwise.} \end{cases}$ *the convex indicator function. Note that it differs from the classical indicator function. The idea is that its generalization to a set* $\mathcal{C}$ *(0 on* $\mathcal{C}$*,* $+\infty$ *elsewhere) is convex if and only if* $\mathcal{C}$ *is convex.*

There may be some problematic cases when solving OT. For instance, if the total mass to be transported differs on each side, $\sum a_i \neq \sum b_j$, or if some points are isolated. To address these issues, one can use unbalanced OT.

Definition 2.2 (Unbalanced OT)**.** *The unbalanced optimal transport (UOT) problem is:*

$$\min_{\mathbf{P}\geq 0} \; \langle \mathbf{C}, \mathbf{P}\rangle + D(\mathbf{P}\mathbf{1}_n, \mathbf{a}) + D(\mathbf{P}^T\mathbf{1}_m, \mathbf{P}) \tag{2.1}$$

where D is a kind of distance.

Remark 2.3. *We have simply replaced the mass conservation constraints present in OT by regularization terms that penalize discrepancies between the marginals and the mass vectors.*
Taking $D(\mathbf{x}, \mathbf{y}) = i_{\{\mathbf{0}\}}(\mathbf{x}-\mathbf{y})$, we recover OT.

D is generally a φ-divergence (such as the Kullback–Leibler divergence); these are functions that can be written as:

$$D_\varphi(\mathbf{x}, \mathbf{y}) = \int \varphi\left(\frac{d\mathbf{x}}{d\mathbf{y}}\right) d\mathbf{y} \quad \text{or } \sum_i \varphi\left(\frac{x_i}{y_i}\right) y_i \text{ in the discrete case.}$$

In our case, we fix $D(\mathbf{x}, \mathbf{y}) = \tau\,\mathrm{KL}(\mathbf{x}, \mathbf{y})$ with $\tau > 0$, and consider EOT.

Definition 2.4 (Fenchel–Legendre Transform)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$; we define its Fenchel–Legendre transform by:*

$$\varphi^*(\mathbf{u}) = \sup_{\mathbf{x}} \; \langle \mathbf{x}, \mathbf{u}\rangle - \varphi(\mathbf{x})$$

Example 2.5.

1. *For $\varphi(\mathbf{x}) = \frac{a}{2}\|\mathbf{x}\|^2$ ($a > 0$), we have $\varphi^*(\mathbf{u}) = \frac{1}{2a}\|\mathbf{u}\|^2$ by setting the gradient to zero.*
2. *For $\varphi(\mathbf{x}) = \langle \mathbf{C}, \mathbf{x}\rangle$, we have $\varphi^*(\mathbf{u}) = i_{\{\mathbf{C}\}}(\mathbf{u})$.*
3. *For $\varphi(\mathbf{x}) = i_{\{\mathbf{C}\}}(\mathbf{x})$, we have $\varphi^*(\mathbf{u}) = \langle \mathbf{C}, \mathbf{u}\rangle$.*

Proposition 2.6.

1. *f^* is convex because it is the supremum of affine (hence convex) functions.*
2. *If f is convex (and lower semi-continuous), then $f^{**} = f$.*
3. *If f is convex and differentiable, then $\nabla f^* = (\nabla f)^{-1}$ up to a mild technical condition.*
4. *If f is separable, then f^* is separable (i.e. $f(\mathbf{x}) = \sum_i f_i(x_i) \to f*(\mathbf{u}) = \sum_i f_i^*(\mathbf{u})$).*

The Fenchel–Legendre transform is a very useful tool for solving optimization problems. It is somewhat analogous to the Fourier transform in signal processing.

Let us now see how to solve UOT. We will compute its dual using the Fenchel transform. Let us return to our UOT problem and introduce the functions $F(\mathbf{u}) = \tau \operatorname{KL}(\mathbf{u}, \mathbf{a})$ and $G(\mathbf{v}) = \tau \operatorname{KL}(\mathbf{v}, \mathbf{P})$ (the derivation also holds for generic penalizations).

$$
\begin{aligned}
&\min_{\mathbf{P}\geq 0} \langle \mathbf{C}, \mathbf{P}\rangle + \varepsilon \sum_{i,j} \mathbf{P}_{i,j}(\log \mathbf{P}_{i,j} - 1) + \tau F(\mathbf{P}\mathbf{1}_n) + \tau G(\mathbf{P}^T\mathbf{1}_m) \\
&= \min_{\mathbf{P}\geq 0, \mathbf{u}, \mathbf{v}} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) + \tau F(\mathbf{u}) + \tau G(\mathbf{v}) \qquad \text{s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{u}, \mathbf{P}^T\mathbf{1}_m = \mathbf{v} \\
&= \min_{\mathbf{P}\geq 0, \mathbf{u}, \mathbf{v}} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) + \tau F(\mathbf{u}) + \tau G(\mathbf{v}) + \max_{\mathbf{f},\mathbf{g}} \langle \mathbf{f}, \mathbf{u} - \mathbf{P}\mathbf{1}_n\rangle + \langle \mathbf{g}, \mathbf{v} - \mathbf{P}^T\mathbf{1}_m\rangle \\
&= \max_{\mathbf{f},\mathbf{g}} \min_{\mathbf{P}\geq 0, \mathbf{u}, \mathbf{v}} \langle \mathbf{C}, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) + \tau F(\mathbf{u}) + \tau G(\mathbf{v}) + \langle \mathbf{f}, \mathbf{u} - \mathbf{P}\mathbf{1}_n\rangle + \langle \mathbf{g}, \mathbf{v} - \mathbf{P}^T\mathbf{1}_m\rangle \\
&= \max_{\mathbf{f},\mathbf{g}} \min_{\mathbf{P}\geq 0} \langle \mathbf{C} - \mathbf{f}\mathbf{1}_n^T - \mathbf{1}_m\mathbf{g}^T, \mathbf{P}\rangle - \varepsilon H(\mathbf{P}) + \min_{\mathbf{u}} (\langle \mathbf{u}, \mathbf{f}\rangle + \tau F(\mathbf{u})) + \min_{\mathbf{v}} (\langle \mathbf{v}, \mathbf{g}\rangle + \tau G(\mathbf{v})) \\
&= \max_{\mathbf{f},\mathbf{g}} -\varepsilon \sum_{i,j} \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - \mathbf{C}_{i,j}}{\varepsilon}\right) - \tau F^*\left(\frac{-\mathbf{f}}{\tau}\right) - \tau G^*\left(\frac{-\mathbf{g}}{\tau}\right) \qquad \text{(Dual of UEOT)}
\end{aligned}
$$

Indeed $\min_{\mathbf{u}} \langle \mathbf{u}, \mathbf{f}\rangle + \tau F(\mathbf{u}) = -\sup_{\mathbf{u}} \langle \mathbf{u}, -\mathbf{f}\rangle - \tau F(\mathbf{u}) = -\tau \sup_{\mathbf{u}} \langle \mathbf{u}, -\frac{\mathbf{f}}{\tau}\rangle - F(\mathbf{u}) = -\tau F^*(-\frac{\mathbf{f}}{\tau})$, and similarly $\min_{\mathbf{v}} \langle \mathbf{v}, \mathbf{g}\rangle + \tau G(\mathbf{v}) = -\tau G^*(-\frac{\mathbf{g}}{\tau})$.

Remark 2.7 (Sanity check)**.** *If we take $F(\mathbf{u}) = i_{\{\mathbf{a}\}}(\mathbf{u})$ and $G(\mathbf{v}) = i_{\{\mathbf{P}\}}(\mathbf{v})$, then $F^*(\mathbf{u}) = \langle \mathbf{a}, \mathbf{u}\rangle$ and $G^*(\mathbf{v}) = \langle \mathbf{P}, \mathbf{v}\rangle$. We recover the dual problem of EOT.*

Since the function to be maximized is concave and smooth, we proceed as for EOT, namely we alternate maximization with respect to $\mathbf{f}$ and $\mathbf{g}$:

$$
\begin{aligned}
&\nabla = -\exp(\mathbf{f}/\varepsilon) \odot (\mathbf{K}\exp(\mathbf{g}/\varepsilon)) + \nabla F^*(-\mathbf{f}/\tau) = 0 \\
&\iff -\mathbf{f}/\tau = \nabla F(\mathbf{u} \odot \mathbf{K}\mathbf{v})
\end{aligned}
$$

In the case where $F(\mathbf{u}) = \tau \operatorname{KL}(\mathbf{u}, \mathbf{a}) = \tau \sum_i \mathbf{u}_i \log(\mathbf{u}_i/\mathbf{a}_i) - \mathbf{u}_i + \mathbf{a}_i$, we have $\nabla F(\mathbf{u}) = \sum_i \log(\mathbf{u}_i/\mathbf{a}_i)$.

Thus, $\mathbf{f}_i = -\tau \log(\mathbf{u}_i(\mathbf{K}\mathbf{v})_i/\mathbf{a}_i)$, which yields after computation

$$
\mathbf{u}_i = \exp(\mathbf{f}_i/\varepsilon) = \left(\frac{\mathbf{a}_i}{(\mathbf{K}\mathbf{v})_i}\right)^{\frac{\tau}{\tau+\varepsilon}}.
$$

Definition 2.8 (Sinkhorn for UOT)**.** *The Sinkhorn algorithm for UOT consists in iterating the following two steps until convergence:*

$$
\mathbf{u} \leftarrow \left(\frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}\right)^{\frac{\tau}{\tau+\varepsilon}}, \quad \mathbf{v} \leftarrow \left(\frac{\mathbf{P}}{\mathbf{K}^T\mathbf{u}}\right)^{\frac{\tau}{\tau+\varepsilon}}
$$

Remark 2.9. *As $\tau \to \infty$, the Sinkhorn algorithm for UOT converges to the Sinkhorn algorithm for EOT.*

References