Online Sinkhorn: Optimal Transport Distances from Sample Streams

Yue Xiang

2021.2.5

Outlines

OT Distance

$(\mathcal{X}, d)$: a complete metric space
$C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$: cost function
α, β: probability distributions over the space $\mathcal{X}$
Find a plan $\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ to minimize the cost of moving α to β:

$$\mathcal{W}(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X})} \left\{ \langle C, \pi \rangle : \pi_1 = \alpha, \pi_2 = \beta \right\} \tag{1}$$

- $\langle C, \pi \rangle \triangleq \int C(x, y) \mathrm{d}\pi(x, y)$
- $\pi_1 = \int_{y \in \mathcal{X}} \mathrm{d}\pi(\cdot, y)$
- $\pi_2 = \int_{x \in \mathcal{X}} \mathrm{d}\pi(x, \cdot)$

Wasserstein(OT) distance allows to compare distributions with disjoint supports.

Challenges in Estimating OT Distance

- Yet OT algorithms handles discrete distributions only.
- Computing OT distances:
  sample once from $\alpha, \beta \rightarrow$ get $\hat{\alpha}, \hat{\beta}$ discrete realizations$\rightarrow$ solve a discrete linear program (LP).
  - numerically costly and statistically inefficient
  - can't adapt to ml settings where data is resampled continuously or accessed in an online manner

Entropy Penalty for Easy Computation

$$\mathcal{W}(\alpha, \beta) \triangleq \min_{\substack{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{X}) \\ \pi_1=\alpha, \pi_2=\beta}} \langle C, \pi \rangle + \varepsilon \mathrm{KL}(\pi \mid \alpha \otimes \beta) \tag{2}$$

where $\mathrm{KL}(\pi \mid \alpha \otimes \beta) \triangleq \int \log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}\right) \mathrm{d}\pi$.

Primal and Dual Problems

Primal problem on measures

$$\mathcal{W}(\alpha,\beta) \triangleq \min_{\pi\in\mathcal{U}(\alpha,\beta)} \left\{ \int_{x,y} C(x,y)\mathrm{d}\pi(x,y) + \int_{x,y} \log \frac{\mathrm{d}\pi}{\mathrm{d}\alpha\mathrm{d}\beta}(x,y)\mathrm{d}\pi(x,y) \right\} \tag{3}$$

Dual problem on functions

$$\mathcal{W}(\alpha,\beta) = \max_{f,g\in C(\mathcal{X})} \left\{ \langle \alpha, f\rangle + \langle \beta, g\rangle - \langle \alpha\otimes\beta, \exp(f\oplus g - C)\rangle + 1 \right\} \tag{4}$$

- $\langle \alpha, f\rangle \triangleq \int f(x)\mathrm{d}\alpha(x)$
- $(f\oplus g - C)(x,y) \triangleq f(x) + g(y) - C(x,y)$

Simultaneous Updates of f_t and g_t

Problem (4) can be solved by closed-form alternated maximization, which corresponds to Sinkhorn’s algorithm.
At iteration t, the updates are

$$f_{t+1}(\cdot) = T_{\beta}(g_t), \quad g_{t+1}(\cdot) = T_{\alpha}(f_{t+1}) \tag{5}$$

where $T_{\mu}(h) \triangleq -\log \int_{y \in \mathcal{X}} \exp(h(y) - C(\cdot, y)) \mathrm{d}\mu(y)$.

Sinkhorn Algorithm

When the input distributions are discrete, the function f_t and g_t need only to be evaluated on $(x_i)_t$ and $(y_i)_t$.
Let $\boldsymbol{u}_t \triangleq \left(e^{-f_t(x_i)}\right)_{i=1}^{n}$, $\mathbf{v}_t \triangleq \left(e^{-g_t(y_i)}\right)_{i=1}^{n}$, the iteration (5) becomes:

$$\boldsymbol{u}_{t+1} = \boldsymbol{K}\frac{1}{n\mathbf{v}_t} \quad \text{and} \quad \mathbf{v}_{t+1} = \boldsymbol{K}^\top \frac{1}{n\boldsymbol{u}_t} \tag{6}$$

where $\boldsymbol{K} = \left(e^{-C(x_i,y_i)}\right)_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$.

Complexity of Sinkhorn Algorithm

The Sinkhorn algorithm for OT operates in 2 phases:
1. Compute the kernel matrix K with a cost in $O(n^2 d)$(d: dimension of $\mathcal{X}$).
2. Each iterate of (6) costs $O(n^2)$.

Consistency and Bias

- Consistency
  Let $\hat{\alpha} = \frac{1}{n}\sum_i \delta_{x_i}$, $\hat{\beta} = \frac{1}{n}\sum_i \delta_{y_i}$. Consistency holds as $\mathcal{W}\left(\hat{\alpha}_n, \hat{\beta}_n\right) \to \mathcal{W}(\alpha, \beta)$.
- Bias
  The distance $\mathcal{W}(\hat{\alpha}, \hat{\beta})$ and optimal functions $f^*(\hat{\alpha}, \hat{\beta})$ are biased estimations.

Notations

- $n_0 \triangleq 0$
- $n_{t+1} \triangleq n_t + n$
- n: size of mini-batch
- $\hat{\alpha}_t \triangleq \frac{1}{n} \sum_{i=n_t+1}^{n_{t+1}} \delta_{x_i}$
- $u_t \triangleq \exp(-f_t), \ v_t \triangleq \exp(-g_t)$
- $\kappa_y(\cdot) \triangleq \exp(-C(\cdot, y)), \ \kappa_x(\cdot) \triangleq \exp(-C(x, \cdot))$
- $\|f\|_{\mathrm{var}} \triangleq \max_x f(x) - \min_x f(x)$: variation norm

Using principles from SA, we cast the regularized OT problem as a root-finding problem of a function-valued operator $\mathcal{F} : \mathcal{C}_+(\mathcal{X}) \times \mathcal{C}_+(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$, for which we can obtained unbiased estimates. Optimal potentials are indeed exactly the roots of

$$\mathcal{F} : (u, v) \to \left(u(\cdot) - \int_{y \in \mathcal{X}} \frac{1}{v(y)} \kappa_y(\cdot) \mathrm{d}\beta(y), \quad v(\cdot) - \int_{x \in \mathcal{X}} \frac{1}{u(x)} \kappa_x(\cdot) \mathrm{d}\alpha(x) \right) \tag{7}$$

Using two empirical measures $\hat{\alpha}$ and $\hat{\beta}$ to estimate $\mathcal{F}$:

$$\hat{\mathcal{F}}_{\hat{\alpha},\hat{\beta}}(u, v) \triangleq \left(u(\cdot) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{v(y_i)} \kappa_{y_i}(\cdot) \quad v(\cdot) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{u(x_i)} \kappa_{x_i}(\cdot) \right) \tag{8}$$

Online Sinkhorn Iteration

Introduce a learning rate η_t in Sinkhorn iterations for finding roots of vector-valued functions:

$$
\begin{aligned}
(\hat{u}_{t+1}, \hat{v}_{t+1}) &\triangleq (1-\eta_t)\,(\hat{u}_t, \hat{v}_t) - \eta_t \hat{\mathcal{F}}_{\hat{\alpha}_t, \hat{\beta}_t}\,(\hat{u}_t, \hat{v}_t)\,, \qquad \text{i.e.} \\
e^{-\hat{f}_{t+1}} &= (1-\eta_t)\, e^{-\hat{f}_t} + \eta_t e_t^{-T_{\hat{\beta}}}\,(\hat{g}_t)
\end{aligned}
\tag{9}
$$

The estimates $\hat{u}_t$ and $\hat{v}_t$ are defined by weights $(p_{i,t}, q_{i,t})_{i \leqslant n_t}$ and positions $(x_i, y_i)_{i \leqslant n_t} \subseteq \mathcal{X}^2$:

$$
\begin{aligned}
e^{-\hat{f}_t(\cdot)} = \hat{u}_t(\cdot) \triangleq \textstyle\sum_{i=1}^{n_t} \exp\left(q_{i,t} - C(\cdot, y_i)\right) \\
e^{-\hat{g}_t(\cdot)} = \hat{v}_t(\cdot) \triangleq \textstyle\sum_{i=1}^{n_t} \exp\left(p_{i,t} - C(x_i, \cdot)\right).
\end{aligned}
\tag{10}
$$

p_i and q_i are updated in SA (9).

Online Sinkhorn Algorithm

Algorithm 1 Online Sinkhorn

Input: Dist. α and β, learning weights $(\eta_t)_t$, batch sizes $(n(t))_t$ **Set** $p_i = q_i = 0$ for $i \in (0, n_1]$
for $t = 0, \ldots, T-1$ **do**
  Sample $(x_i)_{(n_t,n_{t+1}]} \sim \alpha$, $(y_j)_{(n_t,n_{t+1}]} \sim \beta$.
  Evaluate $(\hat{f}_t(x_i))_{i=(n_t,n_{t+1}]}$, $(\hat{g}_t(y_i))_{i=(n_t,n_{t+1}]}$ using $(q_{i,t}, p_{i,t}, x_i, y_i)_{i=(0,n_t]}$ in (7).
  $q_{(n_t,n_{t+1}],t+1} \leftarrow \log \frac{\eta_t}{n} + (\hat{g}_t(y_i))_{(n_t,n_{t+1}]}$, $\quad p_{(n_t,n_{t+1}],t+1} \leftarrow \log \frac{\eta_t}{n} + (\hat{f}_t(x_i))_{(n_t,n_{t+1}]}$.
  $q_{(0,n_t],t+1} \leftarrow q_{(0,n_t],t} + \log(1-\eta_t)$, $\quad p_{(0,n_t],t+1} \leftarrow p_{(0,n_t],t} + \log(1-\eta_t)$.
Returns: $\hat{f}_T : (q_{i,T}, y_i)_{(0,n_T]}$ and $\hat{g}_T : (p_{i,T}, x_i)_{(0,n_T]}$

Complexity: Each iteration:

- Computation cost: $\mathcal{O}(n_t^2)$
- Memory cost: $\mathcal{O}(n_t)$

Assumption 1. The cost $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is L-Lipschitz, and $\mathcal{X}$ is compact.
Assumption 2. $(\eta_t)_t$ is such that $\sum \eta_t = \infty$ and $\sum \eta_t^2 < \infty, 0 \leqslant \eta_t \leqslant 1$ for all $t > 0$.
Assumption 3. For all $t > 0, n(t) = \frac{B}{w_t^2} \in \mathbb{N}$ and $0 \leqslant \eta_t \leqslant 1$. $\sum w_t \eta_t < \infty$ and $\sum \eta_t = \infty$.
Proposition 1: Under Assumption 1 and 3, the online Sinkhorn algorithm converges almost surely:

$$\left\|\hat{f}_t - f^\star\right\|_{\text{var}} + \left\|\hat{g}_t - g^\star\right\|_{\text{var}} \to 0 \tag{11}$$

(The online Sinkhorn algorithm converges almost surely with slightly increasing batch-size $n(t)$.)

Proposition 2. *Under Assumption 1 and 2, the online Sinkhorn algorithm (Algorithm 1) yields a sequence (f_t, g_t) that reaches a ball centered around $f^\star, g^\star$ for the variational norm $\|\cdot\|_{\text{var}}$. Namely, there exists $T > 0$, $A > 0$ such that for all $t > T$, almost surely*

$$\|f_t - f^\star\|_{\text{var}} + \|g_t - g^\star\|_{\text{var}} \leqslant \frac{A}{\sqrt{n}}.$$

Faster Estimation of OT Distances for a Given Budget

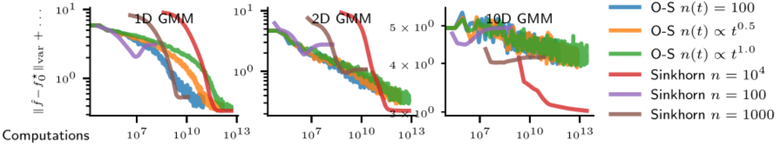


Figure 1: Online Sinkhorn consistently estimate the true regularized OT potentials. Convergence here is measured in term of distance with potentials evaluated on a "test" grid of size $n = 10^4$. Online-Sinkhorn can estimate potentials faster than sampling then scaling the cost matrix.

Consistent Estimation of Continuous OT Distances

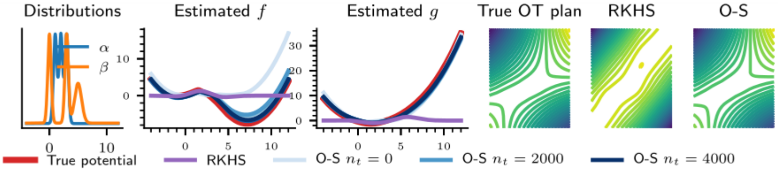


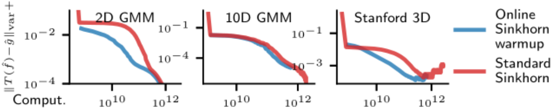
Figure 2: Online Sinhkorn finds the correct potentials over all space, unlike SGD over a RKHS parametrization of the potentials. The plan is therefore correctly estimated everywhere.

Consistent estimation of f^* and $g^\star$: for $N_t \to_{t\to\infty} +\infty$,

$$\left\|f_t - f^\star\right\|_\infty \to 0, \quad \left\|g_t - g^\star\right\|_\infty \to 0, \quad w_t \to \mathcal{W}(\alpha, \beta)$$

Efficient Warmup of Discrete Sinkhorn

Instead of computing the matrix $\left(\exp\left(-C\left(x_i, y_j\right)\right)\right)_{i,j}$ then scale it. fill the matrix while updating sketch potentials with online Sinkhorn.



$-\log \varepsilon$	4	3	2	1
Stanford	5.3x	4.4x	2.3x	1.2x
10d GMM	1.4x	1.5x	1.3x	1.2x
2d GMM	17x	3.7x	1.3x	2.0x

Figure 3: Online Sinkhorn allows to warmup Sinkhorn during the evaluation of the cost matrix, and to speed discrete optimal transport. Table 1: Speed-ups provided by OS vs S to reach a 10^{-3} precision.