Online Sinkhorn: Optimal Transport Distances from Sample Streams

A paper by Arthur Mensch and Gabriel Peyré (2020)

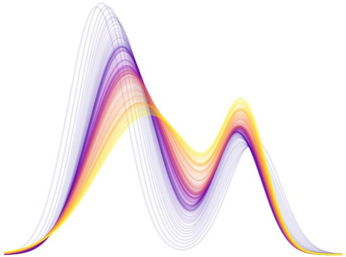
Presented by Maria Cherifa

October 19, 2026

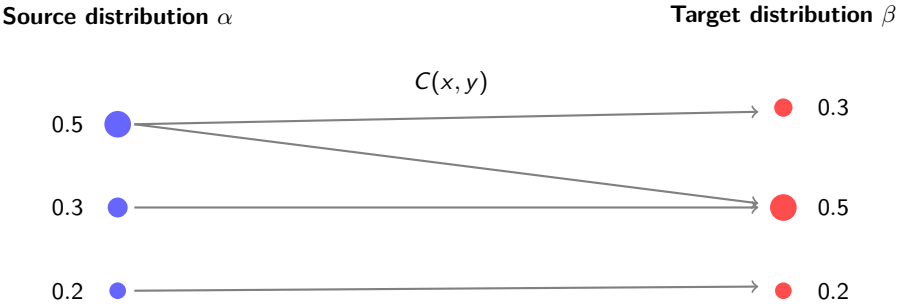
Research paper presentation.

0 Table of contents

1. Motivation and Introduction
2. Background on optimal transport
3. Online Sinkhorn and theoritical analysis
4. Experiments and discussions
5. Limitation and takeaways



1 Motivation: moving probability mass



How can we transport mass from α to β at minimum cost?

1 Motivation: definition of the optimal transport

- Let $(\mathcal{X}, d)$ be a metric space, and consider two probability distributions $\alpha, \beta \in \mathcal{P}(\mathcal{X})$.
- We want to compare them while taking into account the **geometry** of $\mathcal{X}$.
- Given a ground cost $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, find the cheapest way to transport mass from α to β.

Optimal Transport

$$W_C(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} C(x, y)\, d\pi(x, y).$$

where $\Pi(\alpha, \beta) = \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : \pi_1 = \alpha, \pi_2 = \beta\}$.

What if α and β are only accessible through streams of samples?

1 Motivation: optimal transport from sample streams

- The distributions α and β are unknown, but we can draw independent samples:

$$X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} \alpha, \qquad Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} \beta.$$

Research question

How can we estimate $W_C(\alpha, \beta)$ **progressively** as new samples arrive?

2 Optimal transport: three formulations

Let $\alpha, \beta \in \mathcal{P}(\mathcal{X})$ and $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

1. Continuous (Kantorovich) formulation

$$W_{C,0}(\alpha, \beta) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} C(x, y)\, d\pi(x, y).$$

2. Discrete formulation

For $\alpha = \sum_{i=1}^n a_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m b_j \delta_{y_j}$,

$$\min_{P \in \mathbb{R}_+^{n\times m}} \quad \sum_{i=1}^n \sum_{j=1}^m C_{ij} P_{ij}$$

$$\text{s.t.} \quad P\mathbf{1}_m = a, \qquad P^\top \mathbf{1}_n = b,$$

where $C_{ij} = C(x_i, y_j)$.

3. Probabilistic formulation

$$W_{C,0}(\alpha, \beta) = \inf_{\substack{(X,Y):\\ X\sim\alpha,\ Y\sim\beta}} \mathbb{E}\left[C(X, Y)\right].$$

2 Entropy-regularized optimal transport

The Kantorovich problem is linear in the transport plan. We introduce an **entropic regularization**:

Regularized optimal transport

$$W_{C,\varepsilon}(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \{\langle C,\pi\rangle + \varepsilon\,\mathrm{KL}(\pi\|\alpha\otimes\beta)\}\,.$$

where

$$\mathrm{KL}(\pi\|\alpha\otimes\beta) = \int \log\left(\frac{d\pi}{d(\alpha\otimes\beta)}\right) d\pi.$$

Why regularize?

- Strict convexity: uniqueness of the optimal coupling.
- Smoother, less concentrated transport plans.
- Exploitable structure: efficient Sinkhorn iterations.

$$\varepsilon \to 0 \quad\Longrightarrow\quad W_{C,\varepsilon}(\alpha,\beta) \to W_{C,0}(\alpha,\beta).$$

2 From primal to dual: the key structure

For discrete marginals a, b, consider

Constrained optimization

$$\min_{P \geq 0} \sum_{i,j} C_{ij} P_{ij} + \varepsilon \sum_{i,j} P_{ij} \log \frac{P_{ij}}{a_i b_j}$$
$$\text{s.t } P\mathbf{1} = a,\ P^\top \mathbf{1} = b$$

The Lagrangian is defined by,

$$\mathcal{L}(P, f, g) = \sum_{i,j} \left[C_{ij} P_{ij} + \varepsilon P_{ij} \log \frac{P_{ij}}{a_i b_j} \right] + \sum_i f_i \left(a_i - \sum_j P_{ij} \right) + \sum_j g_j \left(b_j - \sum_i P_{ij} \right).$$

First-order optimality: $\frac{\partial \mathcal{L}}{\partial P_{ij}} = 0$

$$P^*_{ij} = a_i b_j \exp\left(\frac{f^*_i + g^*_j - C_{ij}}{\varepsilon} \right)$$

The optimal coupling has an exponential structure.

2 Sinkhorn: alternating optimization in the dual

- Use strong duality and compute the dual potentials.

$$\max_{f,g \in C(\mathcal{X})} F_{\alpha,\beta}(f,g) = \langle f, \alpha \rangle + \langle g, \beta \rangle - \langle e^{f \oplus g - C}, \alpha \otimes \beta \rangle + 1.$$

Key idea: alternate maximization over the dual potentials f and g

- Step 1: $f^+ = \arg\max_f F_{\alpha,\beta}(f,g)$

$$f^+(x) = -\log \int e^{g(y)-C(x,y)} d\beta(y)$$

- Step 2: $g^+ = \arg\max_g F_{\alpha,\beta}(f^+,g)$,

$$g^+(y) = -\log \int e^{f^+(x)-C(x,y)} d\alpha(x)$$

Sinkhorn iterations

$$f_{t+1} = T_\beta(g_t), \qquad g_{t+1} = T_\alpha(f_{t+1}).$$

with $T_\mu(h)(x) = -\log \int e^{h(y)-C(x,y)} d\mu(y)$.

2 Estimating optimal transport from samples

In practice, the distributions α and β are often unknown; we only have access to i.i.d. samples.

- Empirical distributions

  Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \alpha$ and $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} \beta$.

Empirical distributions are defined by

$$\widehat{\alpha}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \quad \widehat{\beta}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}$$

- Empirical Sinkhorn

Construct the kernel matrix:

$$K_{ij} = e^{-C(X_i, Y_j)/\varepsilon}.$$

Apply Sinkhorn iterations to compute:

$$\boxed{\widehat{W}_n = W_{C,\varepsilon}(\widehat{\alpha}_n, \widehat{\beta}_n)}$$

Computational complexity

- Kernel construction: $O(n^2 d)$ for a typical cost on $\mathbb{R}^d$.
- Each Sinkhorn iteration: $O(n^2)$.
- Storing the dense kernel: $O(n^2)$ memory.

2 Empirical Sinkhorn: limitations

1. Two sources of error:

Let $L = \widehat{W}_{n,k} - W_{C,\varepsilon}(\alpha, \beta)$ denote the estimation error after k Sinkhorn iterations.

$$L = \underbrace{\widehat{W}_{n,k} - W_{C,\varepsilon}(\widehat{\alpha}_n, \widehat{\beta}_n)}_{\substack{\text{Optimization} \\ \text{error}}} + \underbrace{W_{C,\varepsilon}(\widehat{\alpha}_n, \widehat{\beta}_n) - W_{C,\varepsilon}(\alpha, \beta)}_{\substack{\text{Statistical} \\ \text{error}}}.$$

- More iterations reduce optimization error.
- More samples reduce statistical error.

2. Fixed-size mini-batch estimation

Given independent batches of size b:

$$\overline{W}_T = \frac{1}{T} \sum_{t=1}^{T} W_{C,\varepsilon}(\widehat{\alpha}_t, \widehat{\beta}_t).$$

By the law of large numbers:

$$\overline{W}_T \xrightarrow[T \to \infty]{a.s.} \mathbb{E}\left[W_{C,\varepsilon}(\widehat{\alpha}_b, \widehat{\beta}_b) \right].$$

But in general:

$$\boxed{\mathbb{E}[W_{C,\varepsilon}(\widehat{\alpha}_b, \widehat{\beta}_b)] \neq W_{C,\varepsilon}(\alpha, \beta).}$$

Persistent statistical bias!

🤔 Can we **estimate population OT** by progressively updating the Sinkhorn potentials as new samples arrive?

3 Stochastic approximation: the Robbins–Monro idea

Goal: Find a root θ^* of an operator $F : \mathbb{R}^d \to \mathbb{R}^d$, such that $F(\theta^*) = 0$.

Challenge: We cannot evaluate F exactly, but have access to noisy and **unbiased** estimates $\widehat{F}_t$.

Robbins–Monro stochastic approximation

Starting from θ_0, update:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta_t \widehat{F}_t(\theta_t). \\ &= \underbrace{\theta_t - \eta_t F(\theta_t)}_{\text{Deterministic correction}} - \underbrace{\eta_t \xi_t}_{\text{Stochastic perturbation}},\end{aligned}$$

where $\xi_t = \widehat{F}_t(\theta_t) - F(\theta_t)$ and $\sum_t \eta_t = \infty, \quad \sum_t \eta_t^2 < \infty$.

Stochastic error: For an average of $n(t)$ independent samples with bounded variance,

$$\boxed{\mathbb{E}[\|\xi_t\| \mid \mathcal{F}_t] \leq \frac{\sigma}{\sqrt{n(t)}}.}$$

3 Online Sinkhorn: from fixed points to root finding

Key idea: introduce the exponentiated potentials $u = e^{-f}$, $v = e^{-g}$ and the kernel $\kappa(x, y) = e^{-C(x,y)}$.

First fixed-point equation

$$f^*(x) = -\log \int e^{g^*(y) - C(x,y)} d\beta(y)$$

$$u^*(x) = \int \frac{\kappa(x, y)}{v^*(y)} d\beta(y)$$

Second fixed-point equation

$$g^*(y) = -\log \int e^{f^*(x) - C(x,y)} d\alpha(x)$$

$$v^*(y) = \int \frac{\kappa(x, y)}{u^*(x)} d\alpha(x)$$

Root-finding formulation

Define $F : C_+(\mathcal{X})^2 \to C(\mathcal{X})^2$ by $F(u, v) = \begin{pmatrix} u - \int \frac{\kappa(\cdot, y)}{v(y)} d\beta(y) \\ v - \int \frac{\kappa(x, \cdot)}{u(x)} d\alpha(x) \end{pmatrix}$. Then the optimal potentials satisfy $\boxed{F(u^*, v^*) = 0}$.

3 Online Sinkhorn: estimating the operator from samples

At iteration t, draw independent mini-batches: $X_1^{(t)}, \dots, X_{n(t)}^{(t)} \overset{\text{i.i.d.}}{\sim} \alpha$, and $Y_1^{(t)}, \dots, Y_{n(t)}^{(t)} \overset{\text{i.i.d.}}{\sim} \beta$, then $F(u, v)$ is estimated by,

Empirical operator

$$\widehat{F}_t(u, v) = \begin{pmatrix} u - \dfrac{1}{n(t)} \displaystyle\sum_{j=1}^{n(t)} \dfrac{\kappa(\cdot, Y_j^{(t)})}{v(Y_j^{(t)})} \\ v - \dfrac{1}{n(t)} \displaystyle\sum_{i=1}^{n(t)} \dfrac{\kappa(X_i^{(t)}, \cdot)}{u(X_i^{(t)})} \end{pmatrix}$$

Key property: conditional unbiasedness

For new batches independent of the current iterates,

$$\boxed{\mathbb{E}\left[\widehat{F}_t(u_t, v_t) \mid \mathcal{F}_t\right] = F(u_t, v_t),}$$

where $\mathcal{F}_t$ contains the information available before drawing batch t.

3 Online Sinkhorn: stochastic approximation

Idea: apply Robbins–Monro stochastic approximation to solve $F(u^*, v^*) = 0$ using the unbiased operator $\widehat{F}_t$.

$$\boxed{\begin{pmatrix} u_{t+1} \\ v_{t+1} \end{pmatrix} = \begin{pmatrix} u_t \\ v_t \end{pmatrix} - \eta_t \widehat{F}_t(u_t, v_t).}$$

where the updates are given by,

$$u_{t+1}(x) = (1-\eta_t)u_t(x) + \frac{\eta_t}{n(t)} \sum_{j=1}^{n(t)} \frac{\kappa(x, Y_j^{(t)})}{v_t(Y_j^{(t)})}.$$

$$v_{t+1}(y) = (1-\eta_t)v_t(y) + \frac{\eta_t}{n(t)} \sum_{i=1}^{n(t)} \frac{\kappa(X_i^{(t)}, y)}{u_t(X_i^{(t)})}.$$

Interpretation

- $(1-\eta_t)$ retains information from past batches.
- η_t controls the contribution of new samples.
- Both potentials are updated simultaneously.

3 Convergence analysis of Online Sinkhorn

Goal: Establish conditions on the learning rates η_t and mini-batch sizes $n(t)$ ensuring convergence to the population optimal dual potentials (f^*, g^*):

$$\boxed{\|f_t - f^*\|_{\mathrm{var}} + \|g_t - g^*\|_{\mathrm{var}} \xrightarrow[t\to\infty]{a.s.} 0.}$$

Key mathematical ingredient: Contraction of the soft C-transform in the variation seminorm:

$$\boxed{\|T_\mu(h) - T_\mu(h')\|_{\mathrm{var}} \le \kappa \|h - h'\|_{\mathrm{var}}, \qquad 0 < \kappa < 1.}$$

where

$$\|h\|_{\mathrm{var}} = \sup_{x\in\mathcal{X}} h(x) - \inf_{x\in\mathcal{X}} h(x).$$

Proof strategy

Use the contraction property to derive a **recursive bound on the estimation error** in the presence of stochastic perturbations.

3 Convergence analysis: controlling stochastic errors

Step 1: Measure the distance to the population optimal dual potentials:

$$e_t = \|\hat{f}_t - f^*\|_{\mathrm{var}} + \|\hat{g}_t - g^*\|_{\mathrm{var}}.$$

Step 2: Combining the Online Sinkhorn updates with contraction gives the key bound:

$$e_{t+1} \leq \underbrace{(1-(1-\kappa)\eta_t)e_t}_{\text{Contraction}} + \underbrace{\eta_t \xi_t}_{\text{Sampling error}}.$$

where ξ_t measures the approximation error of the empirical soft C-transforms at (f^*, g^*).

Step 3: A uniform Monte Carlo bound gives: $\mathbb{E}[\xi_t] \leq \frac{A}{\sqrt{n(t)}}$. Therefore,

$$\mathbb{E}[e_{t+1}] \leq (1-(1-\kappa)\eta_t)\mathbb{E}[e_t] + \frac{A\eta_t}{\sqrt{n(t)}}.$$

Convergence depends on the balance between contraction and sampling error.

3 Online Sinkhorn: fixed mini-batch sizes

$$\mathbb{E}[e_{t+1}] \leq (1 - (1 - \kappa)\eta_t)\mathbb{E}[e_t] + \frac{A\eta_t}{\sqrt{n}}.$$

1. Randomized Sinkhorn (informal)

$$\eta_t = 1$$

Each iteration **replaces** the previous estimate using the new mini-batch.

Proposition 1:(Mensch et al. 2020)
The iterates converge in distribution to a stationary law.

2. Online Sinkhorn (informal)

$$\eta_t \to 0$$

Each iteration **combines** the previous estimate with the new mini-batch.

Proposition 2:(Mensch et al. 2020)
Under suitable step-size conditions, the iterates eventually satisfy:

$$\limsup_{t \to \infty} e_t \leq \frac{A'}{\sqrt{n}} \quad \text{a.s.}$$

Can increasing $n(t)$ eliminate the residual error and guarantee exact convergence?

3 Online Sinkhorn: almost-sure convergence

Idea: Increase the mini-batch size $n(t)$ over time to progressively reduce the stochastic approximation error.

Assumptions:

- Let $w_t > 0$ and choose $n(t) = \left\lceil \frac{B}{w_t^2} \right\rceil, \quad B > 0.$
- Assume $0 \leq \eta_t \leq 1$ and $\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t w_t < \infty.$

Proposition 3: (Mensch et al. 2020) (informal)

Under the regularity assumptions of the paper, the Online Sinkhorn potentials satisfy

$$\boxed{\|\hat{f}_t - f^*\|_{\mathrm{var}} + \|\hat{g}_t - g^*\|_{\mathrm{var}} \xrightarrow[t\to\infty]{a.s.} 0.}$$

Why does it work?

$$\underbrace{\sum_t \eta_t = \infty}_{\text{Sustained contraction}} \qquad \underbrace{\sum_t \frac{\eta_t}{\sqrt{n(t)}} < \infty}_{\text{Controlled sampling errors}}$$

3 Online Sinkhorn: convergence rates

Question: How fast do the estimated dual potentials converge as the number of observed samples increases?

Proposition 4: (Mensch et al. 2020) (informal)

For $\iota \in (0,1)$, choose the schedules: $\eta_t = \frac{S}{t^{1-\iota}}$, and $n(t) = \left\lceil Bt^{4\iota} \right\rceil$, where $S > 0$ and $B > 0$.
Let N be the total number of observed samples and define $\delta_N = \|\hat{f}_{t(N)} - f^*\|_{\mathrm{var}} + \|\hat{g}_{t(N)} - g^*\|_{\mathrm{var}}$. Then, for sufficiently large N,

$$\boxed{\delta_N \leq \frac{D}{N^{\frac{1-\iota}{1+4\iota}}}.}$$

Online Sinkhorn achieves near-$O(N^{-1})$ potential estimation rates, while the corresponding OT cost estimator approaches an $O(N^{-1/2})$ rate.

4 Experiments: estimating continuous OT potentials

Objective: Compare Online Sinkhorn (OS) with classical Sinkhorn on fixed empirical distributions.

Experimental setting:

- Gaussian mixtures in dimensions $d \in \{1, 2, 10\}$.
- Classical Sinkhorn with $N \in \{10^2, 10^3, 10^4\}$ samples.
- Reference potentials computed with 10^4 independent samples.

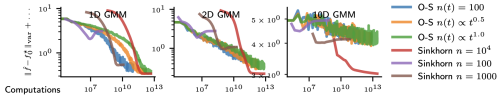


Figure 1 – Mensch & Peyre (2020)

Evaluation metric:

$$\delta_t = \|\hat{f}_t - f_{\mathrm{ref}}\|_{\mathrm{var}} + \|\hat{g}_t - g_{\mathrm{ref}}\|_{\mathrm{var}}.$$

Main observation: Online Sinkhorn estimates population potentials earlier for comparable computational budgets, especially in low dimensions.

4 Experiments: quality of the dual potentials

SGD with RKHS potentials

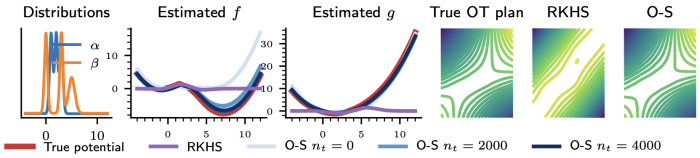
$$f(x) = \sum_{i=1}^{N} a_i k_\sigma(x, x_i).$$

- Generic kernel representation.
- Optimize the dual objective using SGD.
- Requires kernel bandwidth selection.

Online Sinkhorn

$$e^{-\hat{f}_t(x)} = \sum_{j=1}^{N_t} w_{j,t} e^{-C(x, Y_j)}.$$

- Kernel adapted to entropic OT.
- Stochastic fixed-point updates.
- No independent kernel bandwidth.



Main observation: Online Sinkhorn better recovers the shape of the dual potentials, particularly in low-probability regions.

4 Experiments: accelerating discrete Sinkhorn

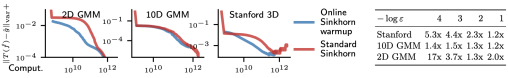
Idea: Use Online Sinkhorn as a warm-start strategy while progressively constructing the cost matrix.

Classical Sinkhorn:

1. Compute the full cost matrix $C \in \mathbb{R}^{N \times N}$.
2. Initialize dual potentials.
3. Run Sinkhorn iterations.

Online Sinkhorn warmup:

1. Compute costs progressively.
2. Update potentials during this phase.
3. Switch to classical 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 – Mensch & Peyre (2020)

Main observation: A better initialization reduces the computation needed to reach a prescribed accuracy, especially for small regularization.

4 Computational complexity and practical considerations

Implementation: Online Sinkhorn maintains non-parametric expansions of the exponentiated potentials:

$$u_t(x) = \sum_{j=1}^{N_t} w_{j,t} e^{-C(x, Y_j)}, \qquad v_t(y) = \sum_{i=1}^{N_t} z_{i,t} e^{-C(X_i, y)}.$$

Computational cost

- Online update: $O(N_t b_t)$
- Total computation: $O(N^2)$
- Memory: $O(N)$

Practical considerations

- No full cost matrix required.
- Growing kernel representation.
- Log-space implementation for numerical stability.

Main trade-off

Incremental computation and linear memory, but quadratic total computation and growing storage requirements.

5 Limitations and key takeaways

Key contributions

- **Streaming OT estimation:** Estimate population optimal transport quantities directly from sample streams.
- **Stochastic approximation:** Reformulate Sinkhorn using unbiased operator estimates.
- **Convergence guarantees:** Almost-sure convergence of dual potentials under increasing mini-batch sizes.
- **Practical benefits:** Incremental updates and warm-starting of classical Sinkhorn.

Limitations

- **Computational complexity:** $O(N^2)$ total computation.
- **Growing memory:** $O(N)$ storage.
- **Entropic regularization:** Theoretical constants deteriorate as $\varepsilon \to 0$.
- **Stationary distributions:** Theory assumes fixed α, β, excluding distribution drift.

Thank you 😀🙏

6 Bibliography I

Mensch, Arthur and Gabriel Peyré (2020). "Online Sinkhorn: Optimal Transport distances from sample streams". In: *arXiv: Optimization and Control*. URL: https://api.semanticscholar.org/CorpusID:220302079.