Meta Optimal Transport

arXiv preprint(2022) arXiv:2206.05262

Amos, Brandon and Cohen, Samuel and Luise, Giulia and Redko, Ievgen
Reporter: Fengjiao Gong

August 18, 2022

Outline

1. Meta Optimal Transport(OT)
2. Meta OT between discrete measures
3. Meta OT between continuous measures
4. Experiments

Meta Optimal Transport(OT)
Optimal Transport(OT)

Kantorovich problem:

$$\pi^\star(\alpha, \beta, \mathbf{c}) \in \arg\min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{c}(x, y)\mathrm{d}\pi(x, y) \tag{1}$$

where

- (α, β): two measures on domains $(\mathcal{X}, \mathcal{Y})$
- $\mathcal{U}(\alpha, \beta)$: a set of admissible couplings between α and β
- $\pi^\star$: optimal coupling, a joint distribution over the product space
- $\mathbf{c} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: ground cost between elements in $\mathcal{X}$ and elements in $\mathcal{Y}$

Optimal Transport(OT)

Kantorovich problem:

$$\pi^\star(\alpha, \beta, c) \in \arg\min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) \tag{1}$$

Standard (1) solver:

- once: computationally expensive
- repeatedly: re-solve the optimization problems from scratch
  ignore shared structure and information between different coupling problems

Meta Optimal Transport(OT)

- Meta OT:
  use **amortized optimization** to ***predict*** OT maps from the input measures

[reference]
Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. ***Learning to optimize: A primer and a benchmark.*** *arXiv preprint arXiv:2103.12828, 2021.*

Brandon Amos. ***Tutorial on amortized optimization for learning to optimize over continuous domains.*** *arXiv preprint arXiv:2202.00665, 2022.*

Meta Optimal Transport(OT)
Amortized optimization

Unconstrained Continuous optimization problems

$$z^\star(\phi) \in \arg\min_z J(z;\phi) \tag{2}$$

where

- J is the objective
- $z \in \mathcal{Z}$ is the domain
- $\phi \in \Phi$ is some context or parameterization
  conditions the objective but is not optimized over

Meta Optimal Transport(OT)
Amortized optimization

Unconstrained Continuous optimization problems

$$z^\star(\phi) \in \arg\min_z J(z;\phi) \tag{13}$$

- Learn a model $\hat{z}_\theta$ to approximate (13) with parameter θ

$$\hat{z}_\theta(\phi) \approx z^\star(\phi)$$

- J is differentiable: objective-based learning

$$\min_\theta \mathbb{E}_{\phi \sim \mathcal{P}(\phi)} J\left(\hat{z}_\theta(\phi); \phi\right) \tag{14}$$

where $\mathcal{P}(\phi)$ is a distribution over contexts

Meta Optimal Transport(OT)

Kantorovich problem:

$$\pi^\star(\alpha, \beta, \boldsymbol{c}) \in \arg\min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} \boldsymbol{c}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \tag{1}$$

Denote a joint meta-distribution over the input measures and costs

$$\mathcal{D}(\alpha, \beta, \boldsymbol{c})$$

Could we directly predict the primal solution to (1)?

$$\pi_\theta(\alpha, \beta, \boldsymbol{c}) \approx \pi^\star(\alpha, \beta, \boldsymbol{c}),\ (\alpha, \beta, \boldsymbol{c}) \sim \mathcal{D}$$

Not primal space!
optimal coupling is often a high-dimensional joint distribution with non-trivial marginal constraints

Meta OT between discrete measures

Outline:

- Entropic OT dual
- Recover primal solution from duals
- Mapping between duals
- Sinkhorn algorithm
- Meta OT between discrete measures

Meta OT between discrete measures

Discrete OT:

$$P^\star(\alpha, \beta, c) \in \operatorname*{arg\,min}_{P \in U(a,b)} \langle C, P \rangle \tag{3}$$

$$U(a, b) := \left\{ P \in \mathbb{R}_+^{n \times m} : P 1_m = a, \quad P^\top 1_n = b \right\}$$

where

- P is a coupling matrix
- $P^\star(\alpha, \beta)$ is the optimal coupling
- $C \in \mathbb{R}^{m \times n}$ is discretized cost matrix with entries $C_{i,j} := c\left(x_i, y_j\right)$

$$\langle C, P \rangle := \sum_{i,j} C_{i,j} P_{i,j}$$

- $a \in \Delta_{m-1}, b \in \Delta_{n-1}$ in probability simplex
- $\alpha := \sum_{i=1}^m a_i \delta_{x_i}$ and $\beta := \sum_{i=1}^n b_i \delta_{y_i}$ discrete measures

Meta OT between discrete measures

Entropic OT:

$$P^\star(\alpha, \beta, \mathbf{c}, \epsilon) \in \arg\min_{P \in U(a,b)} \langle C, P \rangle - \epsilon H(P) \tag{4}$$

where

$$H(P) := -\sum_{i,j} P_{i,j} \left(\log\left(P_{i,j}\right) - 1\right)$$

is the discrete entropy of P.

[reference]

Roberto Cominetti and J San Martín. ***Asymptotic analysis of the exponential penalty trajectory in linear programming.*** *Mathematical Programming, 67(1):169–187, 1994.*

Marco Cuturi. ***Sinkhorn distances: Lightspeed computation of optimal transport****. Advances in neural information processing systems, 26:2292–2300, 2013.*

Meta OT between discrete measures

Entropic OT dual:

$$f^\star, g^\star \in \underset{f \in \mathbb{R}^n, g \in \mathbb{R}^m}{\arg\max} \langle f, a \rangle + \langle g, b \rangle - \epsilon \langle \exp\{f/\epsilon\}, K \exp\{g/\epsilon\} \rangle \tag{5}$$

$$K_{i,j} := \exp\left\{-C_{i,j}/\epsilon\right\}$$

where

- $K \in \mathbb{R}^{m \times n}$: the Gibbs kernel
- $f \in \mathbb{R}^n, g \in \mathbb{R}^m$: dual variables or potentials
- $f^\star(\alpha, \beta, \mathbf{c}, \epsilon), g^\star(\alpha, \beta, \mathbf{c}, \epsilon)$: optimal duals

[reference] Prop. 4.4

Gabriel Peyrë, Marco Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

Meta OT between discrete measures

- Recover primal solution from duals: given optimal duals $f^\star, g^\star$

$$P^\star_{i,j}(\alpha, \beta, c, \epsilon) := \exp\{f^\star_i/\epsilon\} K_{i,j} \exp\left\{g^\star_j/\epsilon\right\} \tag{6}$$

- Mapping between duals: first-order optimality conditions of (5)

$$g(f; b, c) := \epsilon \log b - \epsilon \log\left(K^\top \exp\{f/\epsilon\}\right) \tag{8}$$

It is sufficient to predict one of the potentials, e.g. f, and recover the other.

Meta OT between discrete measures
Sinkhorn algorithm

closed-form block coordinate ascent updates on (5)

Algorithm 1 Sinkhorn($\alpha, \beta, c, \epsilon, f_0 = 0, g_0 = 0$)

---

for iteration $i = 1$ to N **do**

$\quad f_i \leftarrow \epsilon \log a - \epsilon \log \left(K \exp\{g_{i-1}/\epsilon\}\right)$

$\quad g_i \leftarrow \epsilon \log b - \epsilon \log \left(K^\top \exp\{f_{i-1}/\epsilon\}\right)$

end for

Compute P_N from f_N, g_N using eq. (6)

return $P_N \approx P^\star$

[reference] Remark. 4.21

Gabriel Peyrë, Marco Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

Meta OT between discrete measures
Amortization objective

Re-formulate (5) to just optimize over f:

$$f^\star(\alpha, \beta, \boldsymbol{c}, \epsilon) \in \operatorname*{arg\,min}_{f \in \mathbb{R}^n} J(f; \alpha, \beta, \boldsymbol{c}) \tag{15}$$

where $-J(f; \alpha, \beta, \boldsymbol{c}) := \langle f, a \rangle + \langle g, b \rangle$ is the dual objective over f.

Meta OT between discrete measures
Amortization model

Predict the solution to (15) with $\hat{f}_\theta(\alpha, \beta, \boldsymbol{c})$ parameterized by θ

- a computationally efficient approximation $\hat{f}_\theta \approx f^\star$
- model $\hat{f}_\theta$ depends on representations of the input measures and cost
- $\hat{f}_\theta$ as **a fully-connected MLP** mapping from the measures to the duals

Multilayer Perception - fully connected layer + vector input

Meta OT between discrete measures
Amortization loss

Apply (14) to (15)

$$\min_{\theta} \mathop{\mathbb{E}}_{(\alpha,\beta,\boldsymbol{c})\sim\mathcal{D}} J\left(\hat{f}_\theta(\alpha,\beta,\boldsymbol{c});\alpha,\beta,\boldsymbol{c}\right) \tag{16}$$

expectation of the dual objective

Algorithm 3 Training Meta OT

Initialize amortization model with θ_0
for iteration i **do**
  Sample $(\alpha, \beta, c) \sim \mathcal{D}$
  Predict duals $\hat{f}_\theta$ or $\hat{\varphi}_\theta$ on the sample
  Estimate the loss in eq. (16) or eq. (17)
  Update θ_{i+1} with a gradient step
end for

Meta OT between discrete measures
Amortization convergence

The model $\hat{f}_\theta$ distills information between the problem instances into the parameters θ.

Algorithm 4 Fine-tuning with Sinkhorn

Predict duals $\hat{f}_\theta(\alpha, \beta, c)$
Compute $\hat{g}$ from $\hat{f}_\theta$ using eq. (8)
return Sinkhorn$(\alpha, \beta, c, \epsilon, \hat{f}_\theta, \hat{g})$

Meta OT methods surpass standard algorithms by restricting the set of problems rather than considering the average-or worst-case performance.

Meta OT between discrete measures

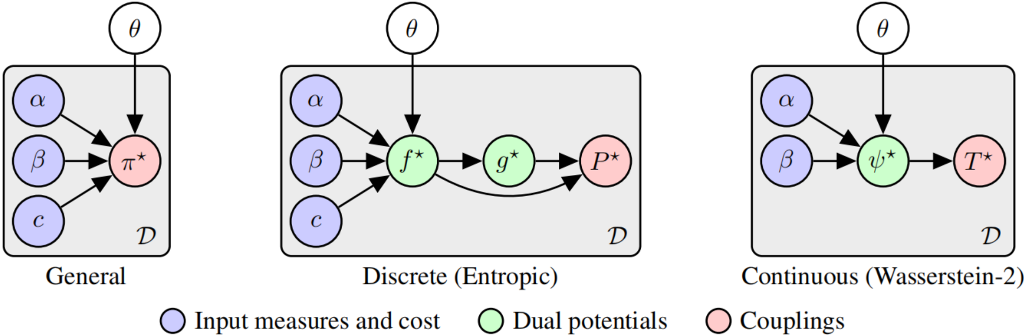

Figure 1: Meta OT uses objective-based amortization for optimal transport. In the general formulation, the *parameters* θ capture shared structure in the *optimal couplings* $\pi^\star$ between multiple input measures and costs over some *distribution* $\mathcal{D}$. In practice, we learn this shared structure over the *dual potentials* which map back to the coupling: $f^\star$ in discrete settings and $\psi^\star$ in continuous ones.

Meta OT between continuous measures

Outline:

- Wasserstein-2 distance
- Convex dual potentials
- Recover primal solution from dual
- Wasserstein-2 Generative Network(W2GN)
- Meta OT between continuous measures

Meta OT between continuous measures

Wasserstein-2 distance

$$W_2^2(\alpha, \beta) := \min_{\pi \in \mathcal{U}(\alpha,\beta)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|_2^2 \, \mathrm{d}\pi(x, y) = \min_T \int_{\mathcal{X}} \|x - T(x)\|_2^2 \, \mathrm{d}\alpha(x) \tag{9}$$

where

- #: pushforward operator, for all measurable set B

$$T_\# \alpha(B) := \alpha\left(T^{-1}(B)\right)$$

- T : transport map pushing α to β, denoted as $T_\# \alpha = \beta$
- α, β: continuous measures in Euclidean space $\mathcal{X}, \mathcal{Y} \in R^d$
- ground cost == squared Euclidean distance

difficulty of representing the coupling + satifying the constraints!

Meta OT between continuous measures

Convex dual potentials:

$$\psi^\star(\cdot;\alpha,\beta) \in \underset{\psi \in \text{ convex}}{\arg\min} \int_{\mathcal{X}} \psi(x)\mathrm{d}\alpha(x) + \int_{\mathcal{Y}} \bar{\psi}(y)\mathrm{d}\beta(y) \tag{10}$$

where

- ψ: a convex function, a convex potential
- $\bar{\psi}(y)$: convex conjugate of ψ or Legendre-Fenchel transform

$$\bar{\psi}(y) := \max_{x\in\mathcal{X}} \langle x, y\rangle - \psi(x)$$

[reference]
Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In International Conference on Machine Learning, pages 6672–6681. PMLR, 2020.
Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. arXiv preprint arXiv:1902.07197, 2019.
Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. arXiv preprint arXiv:1909.13082, 2019.

Meta OT between continuous measures
Recover primal solution from the dual

Given optimal $\psi^\star$ for (10), an optimal map for (9) can be obtained with differentiation:

$$T^\star(x) = \nabla_x \psi^\star(x) \tag{11}$$

Potential ψ is often approximated with an input-convex neural network (ICNN)

[reference]
Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In International Conference on Machine Learning, pages 146–155. PMLR, 2017.
Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417, 1991.

Meta OT between continuous measures
Wasserstein-2 Generative Network(W2GN)

Model ψ_φ and $\overline{\psi_\varphi}$ with two separate ICNNs parameterized by φ.

$$\mathcal{L}(\varphi) := \underbrace{\mathbb{E}_{x\sim\alpha}\left[\psi_\varphi(x)\right] + \mathbb{E}_{y\sim\beta}\left[\langle \nabla\overline{\psi_{\not{\varphi}}}(y), y\rangle - \psi_\varphi(\nabla\overline{\psi_{\not{\varphi}}}(y))\right]}_{\text{Cyclic monotone correlations (dual objective)}} + \gamma \underbrace{\mathbb{E}_{y\sim\beta} \|\nabla\psi_\varphi \circ \nabla\overline{\psi_\varphi}(y) - y\|_2^2}_{\text{Cycle-consistency regularizer}}, \quad (12)$$

where

- $\not{\varphi}$ is a detached copy of the parameters
- first term: optimize the dual objective in (10)
- second term: estimate conjugate $\overline{\psi_\varphi}$

[reference]

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. arXiv preprint arXiv:1909.13082, 2019.

Meta OT between continuous measures
Wasserstein-2 Generative Network(W2GN)

Optimize the loss using samples from the measures:

Algorithm 2 W2GN(α, β, φ_0)

---

for iteration i = 1 to N **do**
  Sample from (α, β) and estimate $\mathcal{L}(\varphi_{i-1})$
  Update φ_i with approximation to $\nabla_\varphi \mathcal{L}(\varphi_{i-1})$
end for
return $T_N(\cdot) := \nabla_x \psi_{\varphi_N}(\cdot) \approx T^\star(\cdot)$

Meta OT between continuous measures

Meta ICNN

Meta ICNN predicts the parameters φ of an ICNN ψ_φ that approximates the optimal dual potentials.

Algorithm 5 Fine-tuning with W2GN

Predict dual ICNN parameters $\hat{\varphi}_\theta(\alpha, \beta, c)$
return W2GN$(\alpha, \beta, c, T, \hat{\varphi}_\theta)$

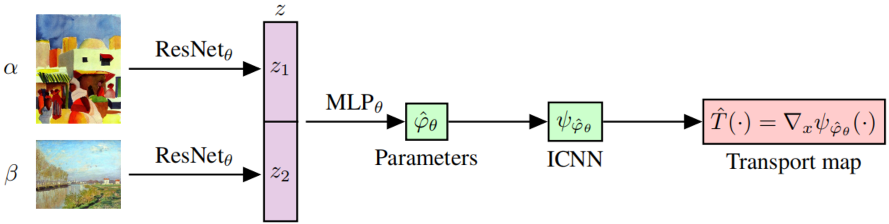

Figure 3: A Meta ICNN for image-based input measures. A shared ResNet processes the input measures α and β into latents z that are decoded with an MLP into the parameters φ of an ICNN dual potential ψ_φ. The derivative of the ICNN provides the transport map $\hat{T}$.

Meta OT between continuous measures

Amortization loss

Apply objective-based amortization (14) to W2GN loss in (12):

$$\min_{\theta} \mathop{\mathbb{E}}_{(\alpha,\beta)\sim\mathcal{D}} \mathcal{L}\left(\hat{\varphi}_{\theta}(\alpha,\beta);\alpha,\beta\right) \tag{17}$$

Here cost c is not included in meta-distribution $(\alpha,\beta) \sim \mathcal{D}(\alpha,\beta)$,as it remains fixed to the squared Euclidean cost everywhere.

Algorithm 3 Training Meta OT

Initialize amortization model with θ_0
for iteration i **do**
 Sample $(\alpha, \beta, c) \sim \mathcal{D}$
 Predict duals $\hat{f}_\theta$ or $\hat{\varphi}_\theta$ on the sample
 Estimate the loss in eq. (16) or eq. (17)
 Update θ_{i+1} with a gradient step
end for

Meta OT between continuous measures

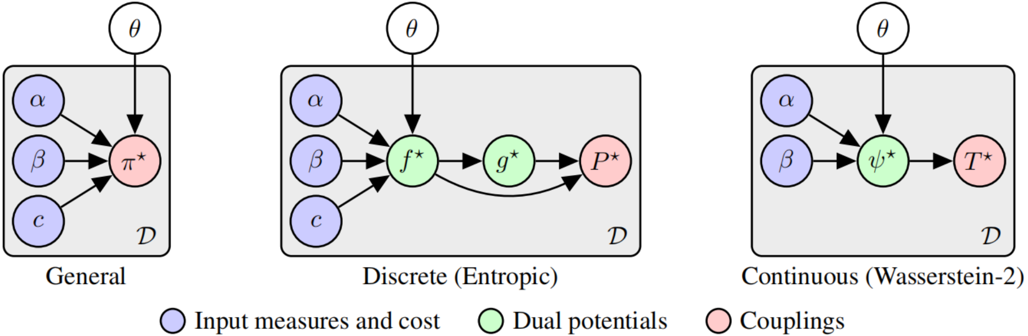

Figure 1: Meta OT uses objective-based amortization for optimal transport. In the general formulation, the *parameters* θ capture shared structure in the *optimal couplings* $\pi^\star$ between multiple input measures and costs over some *distribution* $\mathcal{D}$. In practice, we learn this shared structure over the *dual potentials* which map back to the coupling: $f^\star$ in discrete settings and $\psi^\star$ in continuous ones.

Experiments

Discrete

- Interpolation between MNIST test digits
  Goal: compute the optimal transport interpolation between two measures
- Supply-demand transport on spherical data
  supply and demands may change locations or quantities frequently, creating another Meta OT setting to be able to rapidly solve the new instances

Continuous

- Wasserstein-2 color transfer
  color transfer between two images: mapping the color palette of one image into the other one

Experiments - Discrete OT between MNIST digits

Given a pair of images α_0 and α_1, each grayscale image is cast as a discrete measure in 2-dimensional space where intensities define the probabilities of atoms.

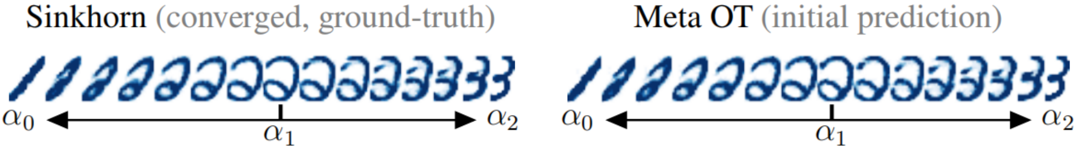

Figure 2: Interpolations between MNIST test digits using couplings obtained from (left) solving the problem with Sinkhorn, and (right) Meta OT model's initial prediction, which is ≈100 times computationally cheaper and produces a nearly identical coupling.

even without fine-tuning, Meta OT's predicted Wasserstein interpolations are close to the ground-truth interpolations obtained from Sinkhorn algorithm.

Experiments - Discrete OT

measures on 2-sphere:
$\mathcal{S}_2 := \{x \in \mathbb{R}^3 : \|x\| = 1\}$, i.e.
$\mathcal{X} = \mathcal{Y} = \mathcal{S}_2$

transport cost: the spherical distance $c(x, y) = \arccos(\langle x, y \rangle)$

Table 1: Discrete OT runtime (in seconds) to reach a marginal error of 10^{-3} and Meta OT's runtime.

	MNIST	Spherical
Sinkhorn	$3.3 \cdot 10^{-3} \pm 1.0 \cdot 10^{-3}$	1.5 ± 0.64
Meta OT + Sinkhorn	$2.2 \cdot 10^{-3} \pm 3.8 \cdot 10^{-4}$	$0.48 \pm .24$
Meta OT (Initial prediction)	$4.6 \cdot 10^{-5} \pm 2.8 \cdot 10^{-6}$	$4.4 \cdot 10^{-5} \pm 3.2 \cdot 10^{-6}$

improved runtime!

`http://github.com/facebookresearch/meta-ot`

Experiments - Discrete OT

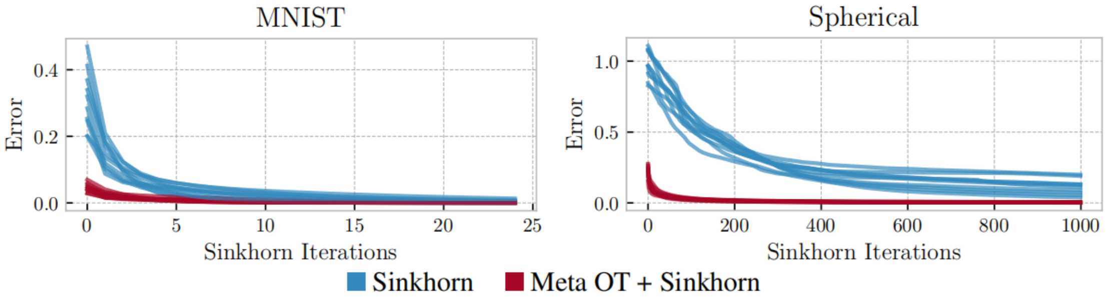

Figure 4: Sinkhorn convergence on test instances. Meta OT successfully predicts warm-start initializations that significantly improve the convergence of Sinkhorn iterations.

near-optimal predictions can be quickly refined in fewer iterations than running Sinkhorn with the default initialization.

Experiments - Discrete OT on spherical data

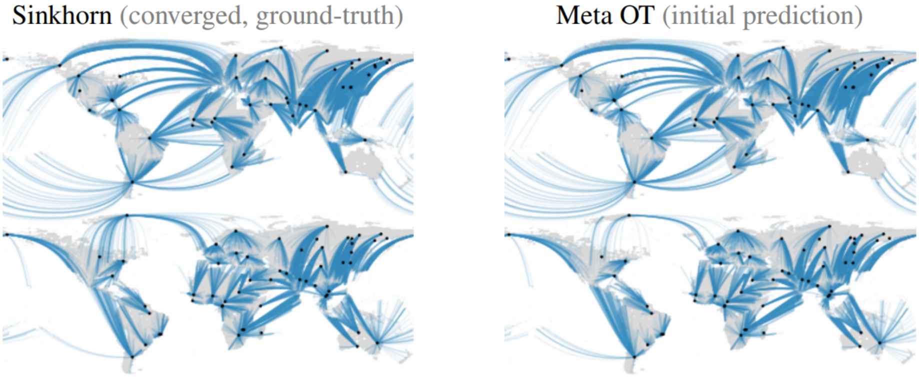

Figure 5: Test set coupling predictions of the spherical transport problem. Meta OT's initial prediction is ≈**37500** times faster than solving Sinkhorn to optimality. Supply locations are shown as black dots and the blue lines show the spherical transport maps T going to demand locations at the end. The sphere is visualized with the Mercator projection.

predicted transport maps are close to the optimal maps obtained from Sinkhorn

Experiments - Continuous Wasserstein-2 color transfer

Table 2: Color transfer runtimes and values.

	Iter	Runtime (s)	Dual Value
Meta OT + W2GN	None	$3.5 \cdot 10^{-3} \pm 2.7 \cdot 10^{-4}$	$0.90 \pm 6.08 \cdot 10^{-2}$
	1k	$0.93 \pm 2.27 \cdot 10^{-2}$	$1.0 \pm 2.57 \cdot 10^{-3}$
	2k	$1.84 \pm 3.78 \cdot 10^{-2}$	$1.0 \pm 5.30 \cdot 10^{-3}$
W2GN	1k	$0.90 \pm 1.62 \cdot 10^{-2}$	$0.96 \pm 2.62 \cdot 10^{-2}$
	2k	$1.81 \pm 3.05 \cdot 10^{-2}$	$0.99 \pm 1.14 \cdot 10^{-2}$

$\approx$ 200 public domain images from WikiArt (`https://www.wikiart.org/`)

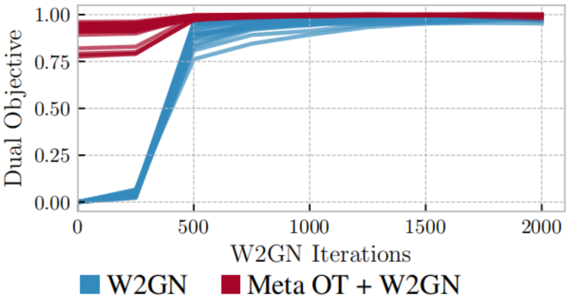

Figure 7: Convergence on color transfer test instances using W2GN. Meta ICNNs predicts warm-start initializations that significantly improve the (normalized) dual objective values.

Experiments - Continuous Wasserstein-2 color transfer

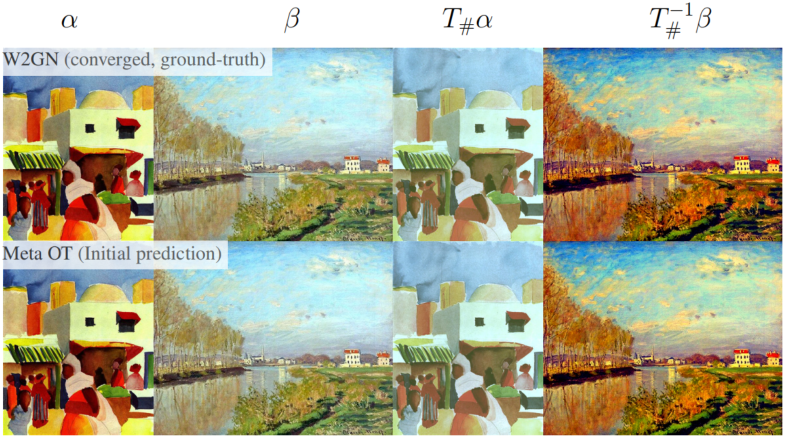

Figure 6: Color transfers with a Meta ICNN on test pairs of images. The objective is to optimally transport the continuous RGB measure of the first image α to the second β, producing an invertible transport map T. Meta OT's prediction is $\approx$**1000** times faster than training W2GN from scratch. The image generating α is Market in Algiers by August Macke (1914) and β is Argenteuil, The Seine by Claude Monet (1872), obtained from WikiArt.

Thanks!