Subspace Detours: Building Transport Plans that are Optimal on Subspace Projections

Boris MUZELLEC (ENSAE, IP Paris) *Marco CUTURI (ENSAE, IP Paris & Google Brain)*

Goal: Alleviate OT *curse of dimensionality* using plans and maps between lower-dimensional projections of distributions.

Approach: Define OT plans and maps that take a "detour" by a lower-dimensional subspace.

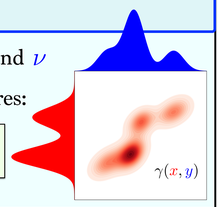
Results: Characterized all solutions. Gave two particular instances, MI and MK. Closed forms for Gaussians. Applications: color transfer, domain adaptation, and semantic mediation (NLP).

1. Optimal Transport

Wasserstein Distance

- $\Pi(\mu,\nu)$: set of couplings between μ and ν
- Distance between probability measures:

$$W_2^2(\mu,\nu) \stackrel{\text{def}}{=} \min_{\gamma\in\Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\left[\|X-Y\|^2\right]$$



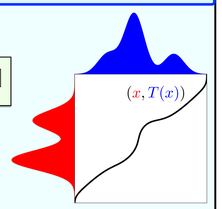
Monge maps and Kantorovich plans

- When there exists $T : T_\sharp\mu = \nu$:

$$W_2^2(\mu,\nu) = \min_{T:T_\sharp\mu=\nu} \mathbb{E}_{X\sim\mu}[\|X-T(X)\|^2]$$

- Link between maps and plans:

$$\gamma^\star = (\mathrm{Id}, T^\star)_\sharp\mu$$



Curse of Dimensionality

- OT sample complexity: $x_1,\dots,x_n\sim\mu$; $y_1,\dots,y_n\sim\nu$

$$|W_2(\mu_n,\nu_n) - W_2(\mu,\nu)| \sim (1/n)^{1/d}$$

2. Workaround: Project First, Transport Next

Sliced Wasserstein [Bonneel et al.'15]

- Average on 1D projections:

$$\mathrm{SW}_2^2(\mu,\nu) \stackrel{\text{def}}{=} \mathbb{E}_{\theta\in S^{d-1}}\left[W_2^2\left((p_\theta)_\sharp\mu,(p_\theta)_\sharp\nu\right)\right]$$

Subspace Robust Wasserstein [Paty & Cuturi'19]

- Projection on adversarially-selected k-D subspace:

$$P_k(\mu,\nu) \stackrel{\text{def}}{=} \max_{E:\dim(E)=k} W_2((p_E)_\sharp\mu,(p_E)_\sharp\nu)$$

$$S_k^2(\mu,\nu) \stackrel{\text{def}}{=} \min_{\gamma\in\Pi(\mu,\nu)} \max_{E:\dim(E)=k} \int \|p_E(x)-p_E(y)\|^2 d\gamma(x,y)$$

Projecting first, transporting next is promising, but it restricts OT to happen in the projected space. *Can we recover a plan in the entire space?*

3. Subspace Transport

Subspace-optimal plans

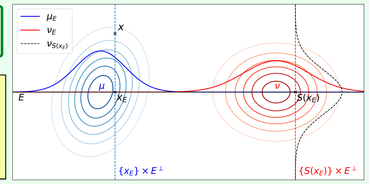
- k-D subspace E, Monge map $S : E\to E$ from μ_E to ν_E
- Marginals: $\mu_E \stackrel{\text{def}}{=} (p_E)_\sharp\mu$, $\nu_E \stackrel{\text{def}}{=} (p_E)_\sharp\nu$, $\gamma_E \stackrel{\text{def}}{=} (p_E,p_E)_\sharp\gamma$

$$\Pi_E(\mu,\nu) \stackrel{\text{def}}{=} \{\gamma\in\Pi(\mu,\nu) : \gamma_E = (\mathrm{Id}_E, S)_\sharp\mu_E\}$$

Degrees of freedom?

- Disintegrations $\gamma_{(x_E,S(x_E))}, x_E\in E$

= full characterisation



Monge-Independent Plans

Extend γ_E with independent couplings $\mu_{x_E}\otimes\nu_{S(x_E)}$

$$\pi_{\mathrm{MI}} \stackrel{\text{def}}{=} (\mu_{x_E}\otimes\nu_{S(x_E)})\otimes(\mathrm{Id}_E, S)_\sharp\mu_E$$

Prop. MI is the limit of discrete E-optimal plans:

μ_n,ν_n empirical dist., $\pi_n\in\Pi_E(\mu_n,\nu_n), n\geq 0$

Then $\pi_n(\mu_n,\nu_n) \rightharpoonup \pi_{\mathrm{MI}}(\mu,\nu)$

Monge-Knothe Maps

Extend γ_E with optimal couplings between μ_{x_E} and $\nu_{S(x_E)}$

$$T_{\mathrm{MK}}(x_E,x_{E^\perp}) \stackrel{\text{def}}{=} (S(x_E), \hat{T}(x_E;x_{E^\perp}))\in E\oplus E^\perp$$

where $\hat{T}(x_E;\cdot) : E^\perp\to E^\perp$ is the Monge map from μ_{x_E} to $\nu_{S(x_E)}$

Prop. The Monge-Knothe plan is optimal in $\Pi_E(\mu,\nu)$

$$\pi_{\mathrm{MK}} \in \arg\min_{\gamma\in\Pi_E(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}[\|X-Y\|^2]$$

where $\pi_{\mathrm{MK}} \stackrel{\text{def}}{=} (\mathrm{Id}_{\mathbb{R}^d}, T_{\mathrm{MK}})_\sharp\mu$

Prop. Let $d_\varepsilon^2(x,y) \stackrel{\text{def}}{=} \sum_{i=1}^k (x_i-y_i)^2 + \varepsilon\sum_{j=k+1}^d (x_j-y_j)^2$

$T_\varepsilon(\mu,\nu)$ Monge map with cost $d_\varepsilon^2(x,y)$

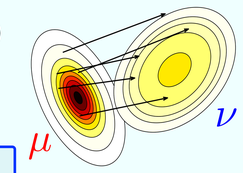
Then $T_\varepsilon(\mu,\nu) \xrightarrow[\varepsilon\to 0]{} T_{\mathrm{MK}}(\mu,\nu)$ in $L_2(\mu)$

4. Optimal transport for Gaussians

Monge Map

$$T_{\mu\nu} : x\to \mathbf{m}_\nu + \mathbf{T}^{\mathbf{AB}}(x-\mathbf{m}_\mu)$$

$$\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$



$(T_{\mu\nu})_\sharp\mu = \nu$

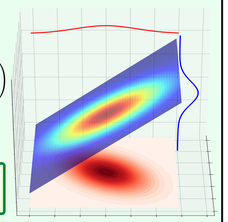
Wasserstein-Bures Distance

$$W_2^2(\mu,\nu) = \|\mathbf{m}_\mu-\mathbf{m}_\nu\|_2^2 + \mathfrak{B}^2(\mathrm{var}\mu,\mathrm{var}\nu)$$

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) = \mathrm{Tr}(\mathbf{A}+\mathbf{B}-2\mathbf{T}^{\mathbf{AB}}\mathbf{A})$$

5. Subspace Detours for Gaussians

- $\mu = \mathcal{N}(0_d,\mathbf{A})$, $\nu = \mathcal{N}(0_d,\mathbf{B})$
- $\mathbf{A} = \begin{pmatrix}\mathbf{A}_E & \mathbf{A}_{EE^\perp}\\ \mathbf{A}_{EE^\perp}^\top & \mathbf{A}_{E^\perp}\end{pmatrix}$, $\mathbf{B} = \begin{pmatrix}\mathbf{B}_E & \mathbf{B}_{EE^\perp}\\ \mathbf{B}_{EE^\perp}^\top & \mathbf{B}_{E^\perp}\end{pmatrix}$
- $(\mathbf{V}_E \quad \mathbf{V}_{E^\perp})$ O.N.B. of $E\oplus E^\perp$

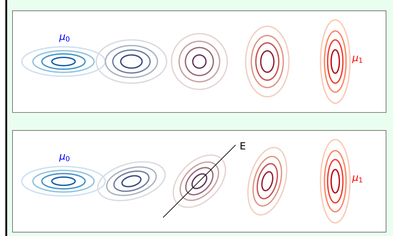


Monge-Independent Transport

Prop. Let $\mathbf{C} \stackrel{\text{def}}{=} (\mathbf{V}_E\mathbf{A}_E + \mathbf{V}_{E^\perp}\mathbf{A}_{EE^\perp}^\top)\,\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E}\,(\mathbf{V}_E\top + (\mathbf{B}_E)^{-1}\mathbf{B}_{EE^\perp}\mathbf{V}_{E^\perp}^\top)$

$\pi_{\mathrm{MI}}(\mu,\nu) = \mathcal{N}(0_{2d},\Sigma_{\mathrm{MI}}) \in \mathcal{P}(\mathbb{R}^d\times\mathbb{R}^d)$ with $\Sigma_{\mathrm{MI}} \stackrel{\text{def}}{=} \begin{pmatrix}\mathbf{A} & \mathbf{C}\\ \mathbf{C}^\top & \mathbf{B}\end{pmatrix}$

Monge-Knothe Transport



Prop. MK maps between Gaussians are linear:

$$\mathbf{T}_{\mathrm{MK}} = \begin{pmatrix}\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E} & 0_{k\times(d-k)}\\ [\mathbf{B}_{EE^\perp}^\top(\mathbf{T}^{\mathbf{A}_E\mathbf{B}_E})^{-1} - \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)}\mathbf{A}_{EE^\perp}^\top](\mathbf{A}_E)^{-1} & \mathbf{T}^{(\mathbf{A}/\mathbf{A}_E)(\mathbf{B}/\mathbf{B}_E)}\end{pmatrix}$$

where $\mathbf{A}/\mathbf{A}_E \stackrel{\text{def}}{=} \mathbf{A}_{E^\perp} - \mathbf{A}_{EE^\perp}^\top\mathbf{A}_E^{-1}\mathbf{A}_{EE^\perp}$ (Schur Complement)

6. Applications: Leveraging lower-dimensional transport maps

Color Transfer

- KMeans quantization in 3D RGB space, 3000 clusters
- 1D OT in gray space $O(n\log n)$

- Extend with MK map (small 3D problems): $\sim\times 50$ speedup

Domain Adaptation with Gaussian Mixture Models

- *Office Home:* 4 domains, 65 classes
- Compute 2048D inception features
- Fit 1 GMM per domain, 1 component per class
- Compute pairwise MK distances, and OT between GMMs:

$$d_{\mathrm{GMM}}(\{a_i;\mu_i\},\{b_j;\nu_j\}) := \min_{\mathbf{P}\in U(\mathbf{a},\mathbf{b})} \langle \mathbf{P}, d_{\mathrm{MK}}^2(\mu_i,\nu_j)\rangle$$

- Map data with bar. proj.

$$\hat{T}_i := \sum_j \mathbf{P}_{ij}T_{\mathrm{MK}}^{ij}$$

- E is selected with PGD:

Algorithm 1 MK Subspace Selection

Input: $\mathbf{A},\mathbf{B}\in\mathrm{PSD}, k\in[1,d], \eta$
  $\mathbf{V}\leftarrow\mathrm{Polar}(\mathbf{AB})$
  while not converged **do**
    $\mathcal{L}\leftarrow\mathrm{MK}(\mathbf{V}^\top\mathbf{A}\mathbf{V},\mathbf{V}^\top\mathbf{B}\mathbf{V};k)$
    $\mathbf{V}\leftarrow\mathbf{V}-\eta\nabla_{\mathbf{V}}\mathcal{L}$
    $\mathbf{V}\leftarrow\mathrm{Polar}(\mathbf{V})$
  end while
Output: $E = \mathrm{Span}\{\mathbf{v}_1,..,\mathbf{v}_k\}$

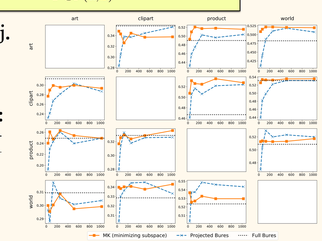


Figure 1: 1-NN accuracy scores v.s. dimension k. E is selected using Algorithm 1 between the source and target covariance matrices for $k = 32, 64, 128, 256, 512, 1024$. Rows: sources, Columns: targets.

Elliptical Word Embeddings: Semantic Mediation

Elliptical word embeddings [BM & MC'18]:

- Represent each word with a PSD matrix

Semantic mediation:

- MK between word embeddings
- E is the k first directions in the SVD of context c

Table 1: Sym. differences of 20-NN sets of $w|c_1$ minus $w|c_2$ using MK. Embeddings are 12×12 pretrained normalized covariance matrices. E is spanned by the 4 principal directions of the contexts. Words in increasing distance order.

Word	Context 1	Context 2	Difference
instrument	monitor	oboe	cathode, monitor, sampler, rca, watts, instrumentation, telescope, synthesizer, ambient
	oboe	monitor	tuned, trombone, guitar, harmonic, octave, baritone, clarinet, saxophone, virtuoso
windows	pc	door	netscape, installer, doubleclick, burner, installs, adapter, router, cpus
	door	pc	screwed, recessed, rails, ceilings, tiling, upvc, profiled, roofs
fox	media	hedgehog	Penny, quiz, Whitman, outraged, Tinker, ads, Keating, Palin, show
	hedgehog	media	panther, reintroduced, kangaroo, Harriet, fair, hedgehog, bush, paw, bunny

github.com/BorisMuzellec/SubspaceOT