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

Boris Muzellec

ENSAE IP PARIS

Marco Cuturi

ENSAE IP PARIS

Google AI

Outline

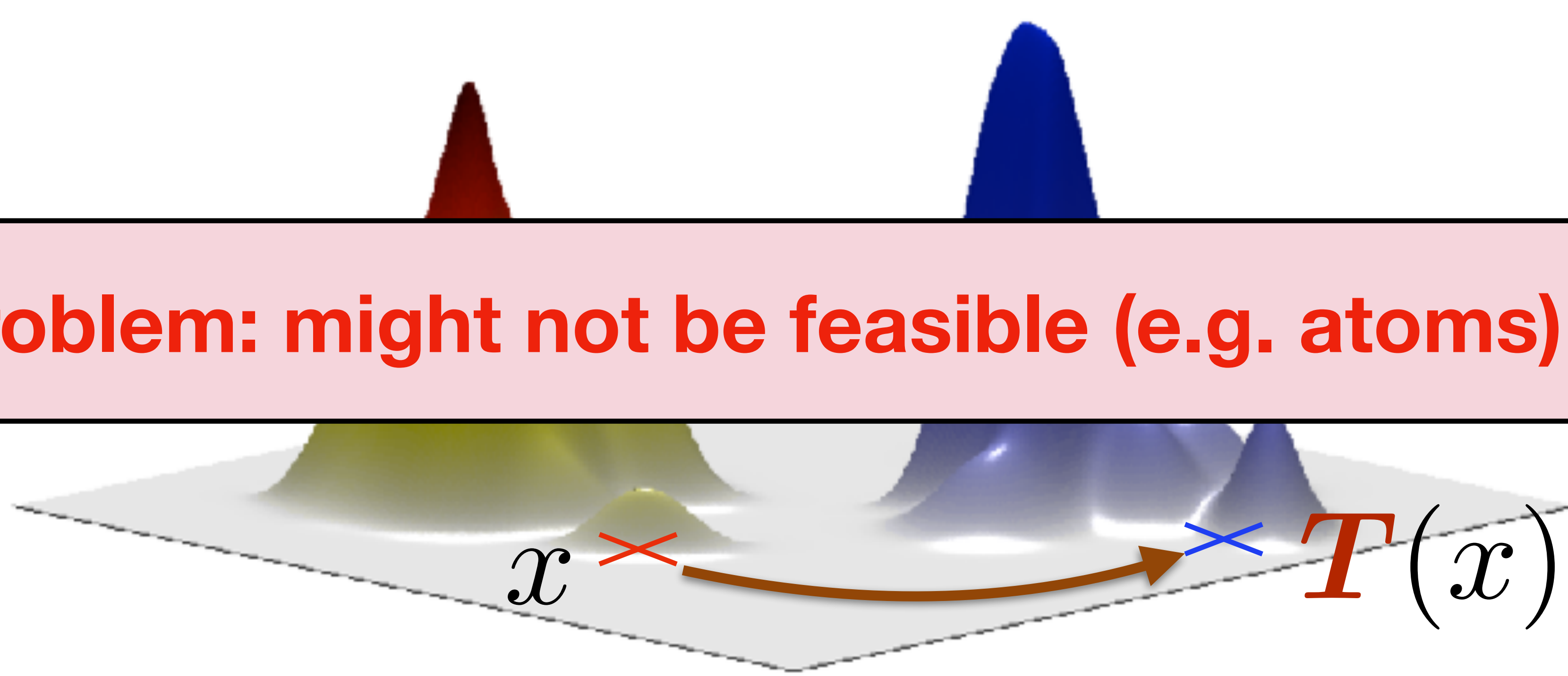
1. A (quick) intro to OT
2. Subspace-Optimal Transport
3. The Gaussian Case
4. Application: Semantic Mediation (NLP)

Monge Problem

Ω a probability space, $c : \Omega \times \Omega \to \mathbb{R}$.
μ, ν two probability measures in $\mathcal{P}(\Omega)$.

[Monge'81] problem: find a map $T : \Omega \to \Omega$

$$\inf_{T_\sharp \mu = \nu} \int_\Omega c(x, T(x))\mu(dx)$$



Problem: might not be feasible (e.g. atoms)

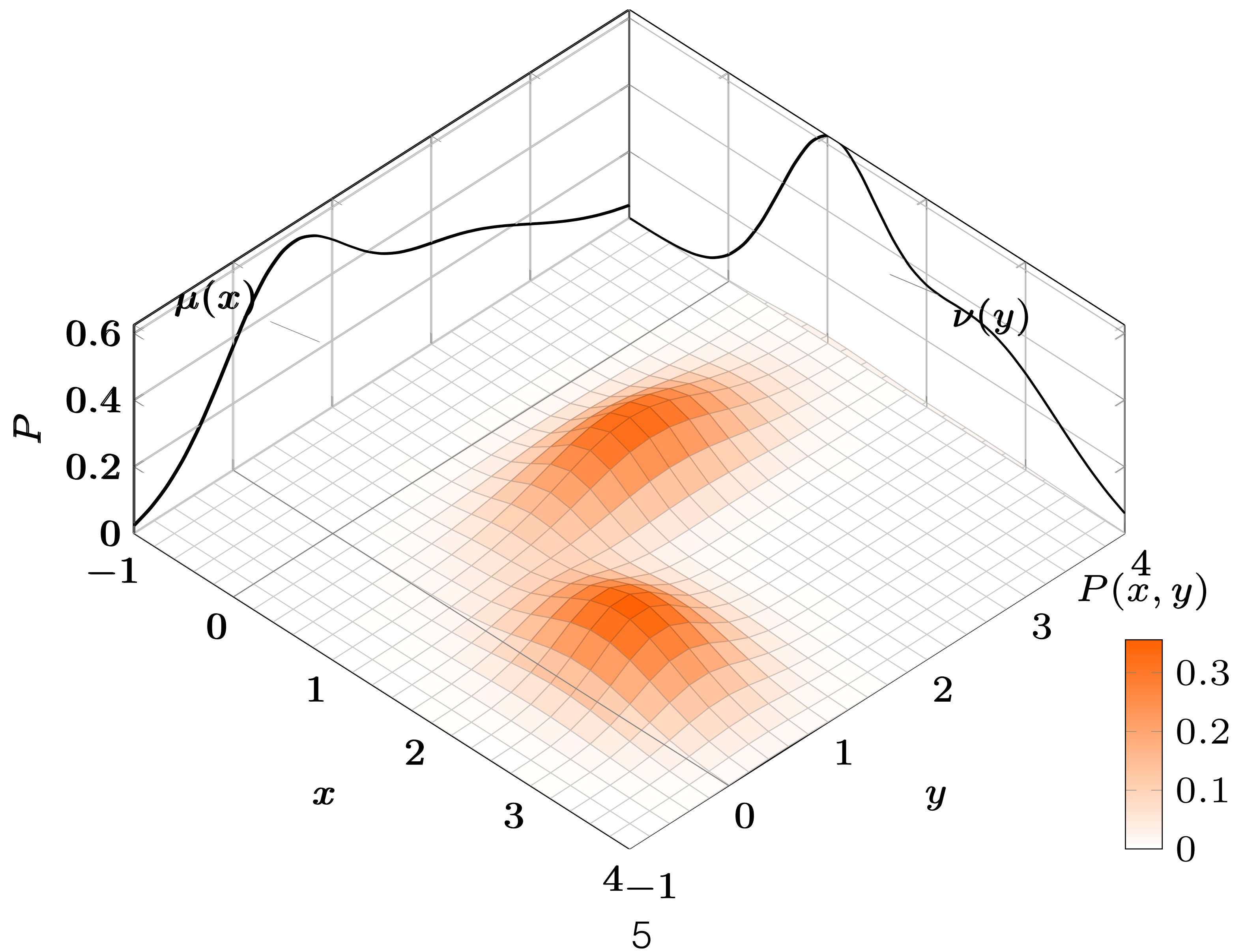
Kantorovich Relaxation

- Instead of maps $\boldsymbol{T} : \Omega \to \Omega$, consider probabilistic maps, i.e. **couplings** $\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega)$:

$$\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathcal{P}(\Omega \times \Omega) | \forall \boldsymbol{A}, \boldsymbol{B} \subset \Omega, \\ \boldsymbol{P}(\boldsymbol{A} \times \Omega) = \boldsymbol{\mu}(\boldsymbol{A}), \\ \boldsymbol{P}(\Omega \times \boldsymbol{B}) = \boldsymbol{\nu}(\boldsymbol{B})\}$$

Kantorovich Relaxation

$$\Pi(\mu,\nu) \stackrel{\text{def}}{=} \{P \in \mathcal{P}(\Omega \times \Omega) | \forall A, B \subset \Omega,$$
$$P(A \times \Omega) = \mu(A), P(\Omega \times B) = \nu(B)\}$$



Links between Monge & Kantorovich

Prop. For "well behaved" costs $\boldsymbol{c}$, if $\boldsymbol{\mu}$ has a density then an *optimal* Monge map $T^\star$ between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ must exist.

Prop. In that case

$$\boldsymbol{P}^\star := (\mathrm{Id}, T^\star)_\sharp \boldsymbol{\mu} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$$

is also *optimal* for the Kantorovich problem.

[Brenier'91] [Smith&Knott'87] [McCann'01]

Wasserstein Distances

Let $p \geq 1$. Let $\boldsymbol{c} := \boldsymbol{D}$, a metric.

Def. The p-Wasserstein distance between $\mu, \nu \in P(\Omega)$ is

$$W_p(\mu, \nu) \stackrel{\text{def}}{=} \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_\Omega D^p(x, y) d\gamma(x, y) \right)^{\frac{1}{p}}$$

Prop. When a Monge map T exists,

$$W_p(\mu, \nu) = \left(\inf_{T_\sharp \mu = \nu} \int_\Omega D^p(x, T(x)) \mu(dx) \right)^{\frac{1}{p}}$$

In the following : $p = 2, \quad c = \| \cdot \|$

Practical Issues

High-Dimensional issues:

- **Sampling complexity in** $\mathcal{O}\left(\frac{1}{n^{\frac{1}{d}}}\right)$ **[Dudley'69, Fournier & Guillin'15]**
- **Computational complexity**

(Partial) Solutions:

- **Regularization [Cuturi'13]**
- **Low-dimensional projections:**
  - **Sliced Wasserstein [Bonneel & al.'15]**
  - **Subspace Robust Wasserstein [Paty & Cuturi'19]**

Low-dimensional Approaches

- Sliced Wasserstein: 1D projections

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

- Subspace-Robust Wasserstein: adversarial kD projections

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

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

But how to reconstruct a transport map (or plan) in $\mathbb{R}^d$?

Subspace-Optimal Transport

Let E a subspace, $S : E \to E$ an (optimal) transport *on E*

Def. The class of E-optimal transport plans from μ to ν is

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

where $\mu_E \overset{def}{=} (p_E)_\sharp(\mu), \quad \nu_E \overset{def}{=} (p_E)_\sharp(\nu), \quad \gamma_E \overset{def}{=} (p_E, p_E)_\sharp(\gamma)$

A quick reminder

Def. Disintegration of μ on E : $(\mu_{x_E})_{x_E \in E}$ **s.t.**

$$\forall g \in C_b(E), x_E \to \int_{E^\perp} g\mu_{x_E} \text{ is Borel-measurable}$$

$$\forall x_E \in E, \mu_{x_E} \text{ is supported on } \{x_E\} \times E^\perp$$

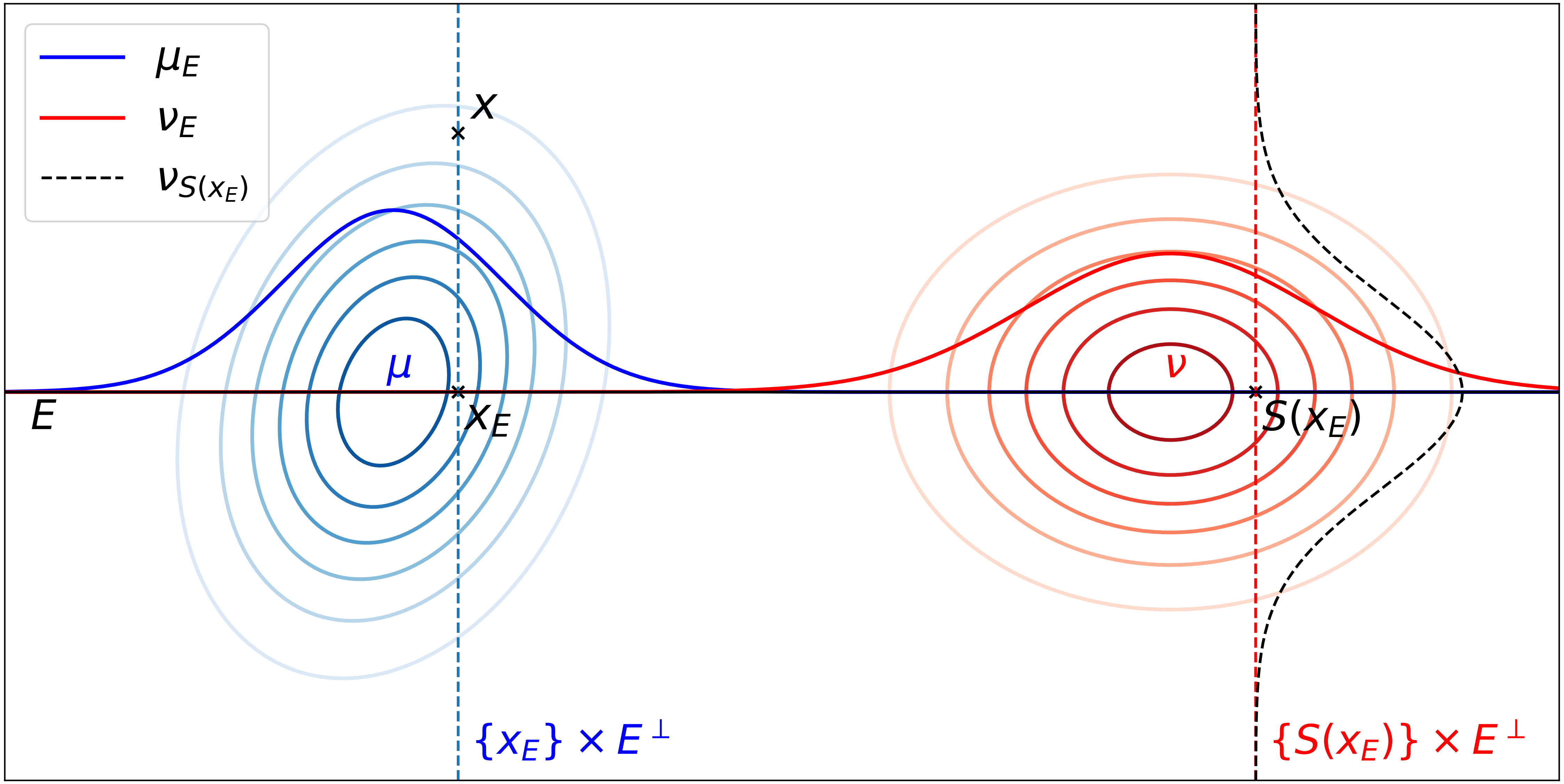
$$\forall f \in C_b(\mathbb{R}^d), \int f d\mu = \int \left(\int f(x_E, x_{E^\perp}) d\mu_{x_E}(x_{E^\perp}) \right) d\mu_E(x_E)$$

Notation: $\mu = \mu_{x_E} \otimes \mu_E$

Degrees of freedom in $\Pi_E(\mu, \nu)$?

- γ_E is supported on $\mathcal{G}(S) \stackrel{def}{=} \{(x_E, S(x_E)) : x_E \in E\}$

$\Longrightarrow$ γ is fully characterised by its disintegrations $\gamma_{(x_E, S(x_E))}, x_E \in E$



Monge-Independent Transport

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

Def. Monge-Independent (MI) transport plan:

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

where $\mu_E \stackrel{def}{=} (p_E)_\sharp(\mu)$, $\nu_E \stackrel{def}{=} (p_E)_\sharp(\nu)$, S **Monge map from** μ_E **to** ν_E, $\gamma_E = (\mathrm{Id}_E, S)_\sharp \mu_E$

Prop. Let $\mu, \nu \in P(\mathbb{R}^d)$ **be a.c. and compactly supported,**

$\mu_n, \nu_n, n \geq 0$ **uniform over** n **i.i.d samples,** $\pi_n \in \Pi_E(\mu_n, \nu_n), n \geq 0$

Then $\pi_n \rightharpoonup \pi_{\mathbf{MI}}$

MI is naturally obtained as the limit of discrete sampling.

Monge-Knothe Transport

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

Let $\forall x_E \in \hat{T}(x_E; \cdot\,) : E^\perp \to E^\perp$ **be the Monge map from** μ_{x_E} **to** $\nu_{S(E)}$

Def. Monge-Knothe (MK) transport map:

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

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

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

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

Monge-Knothe Transport *cont'd*

MK is the limit OT of « split » costs of the type

$$d^2(x,y) := \sum_{i=1}^{k} (x_i - y_i)^2 + \epsilon \sum_{j=k+1}^{d} (x_j - y_j)^2, \quad \epsilon \to 0$$

Prop. Let $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ **be two a.c. probability measures, and**

$\forall \epsilon > 0, \mathbf{P}_\epsilon \overset{def}{=} \mathbf{V}_E \mathbf{V}_E^\top + \epsilon \mathbf{V}_{E^\perp} \mathbf{V}_{E^\perp}^\top$**, and** T_ϵ **the OT map for the cost** $d^2_{\mathbf{P}_\epsilon}(x,y) \overset{def}{=} (x-y)^\top \mathbf{P}_\epsilon (x-y)$

Then $T_\epsilon \to T_{MK}$ **in** $\mathrm{L}_2(\mu)$

OT for Gaussian Distributions

[Gelbrich'90]

Prop. If $\alpha, \beta \in P(\mathbb{R}^d)$ **are elliptical distributions, then**

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

$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \overset{\text{def}}{=} \mathrm{Tr}(\mathbf{A} + \mathbf{B} - 2(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}})$ **is the (squared)** ***Bures*** **distance**

Prop. If $\alpha, \beta \in P(\mathbb{R}^d)$ **are elliptical distributions with** $\mathrm{var}\alpha = \mathbf{A}, \mathrm{var}\beta = \mathbf{B}$**, then**

$T(\mathbf{x}) = \mathbf{m}_\beta + \mathbf{T}^{\mathbf{AB}}(\mathbf{x} - \mathbf{m}_\alpha)$ **is the optimal Monge map**

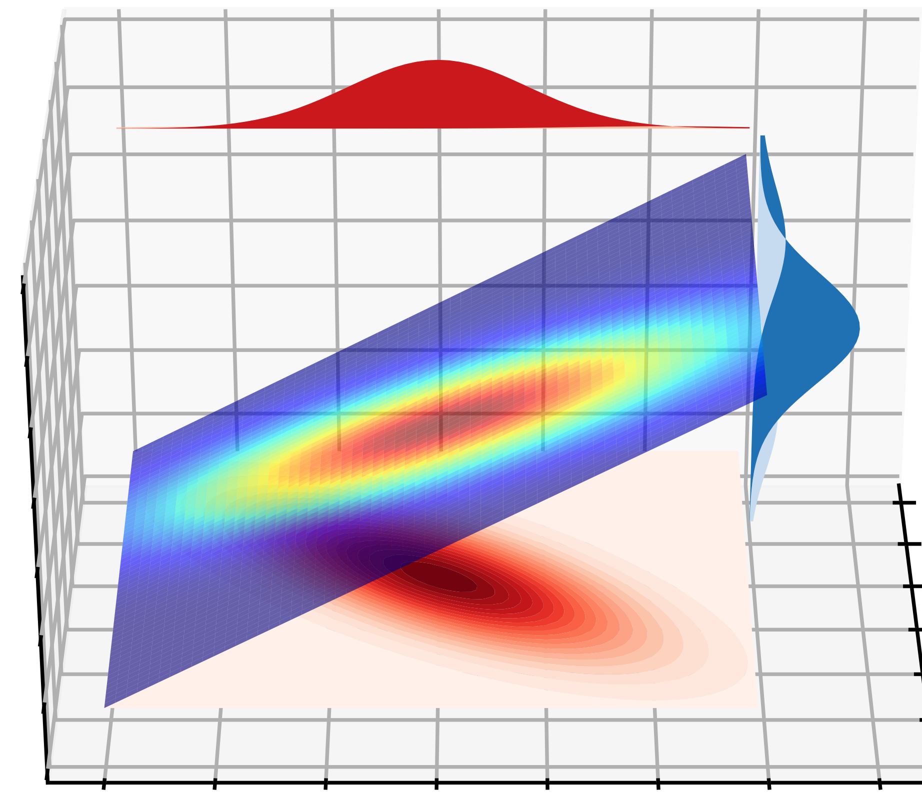
where $\mathbf{T}^{\mathbf{AB}} \overset{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$ **is such that** $\mathbf{T}^{\mathbf{AB}}\mathbf{A}\mathbf{T}^{\mathbf{AB}} = \mathbf{B}$ **and** $\mathbf{T}^{\mathbf{AB}} \in \mathrm{PSD}$

Monge-Independent: Gaussian Distributions

From now on: $\mu = \mathcal{N}(0_d, \mathbf{A})$, $\nu = \mathcal{N}(0_d, \mathbf{B})$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_E & \mathbf{A}_{EE^\perp} \\ \mathbf{A}_{EE^\perp}^\top & \mathbf{A}_{E^\perp} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{B}_E & \mathbf{B}_{EE^\perp} \\ \mathbf{B}_{EE^\perp}^\top & \mathbf{B}_{E^\perp} \end{pmatrix}$$

$(\mathbf{V}_E \quad \mathbf{V}_{E^\perp})$ orthonormal basis of $E \oplus E^\perp$



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

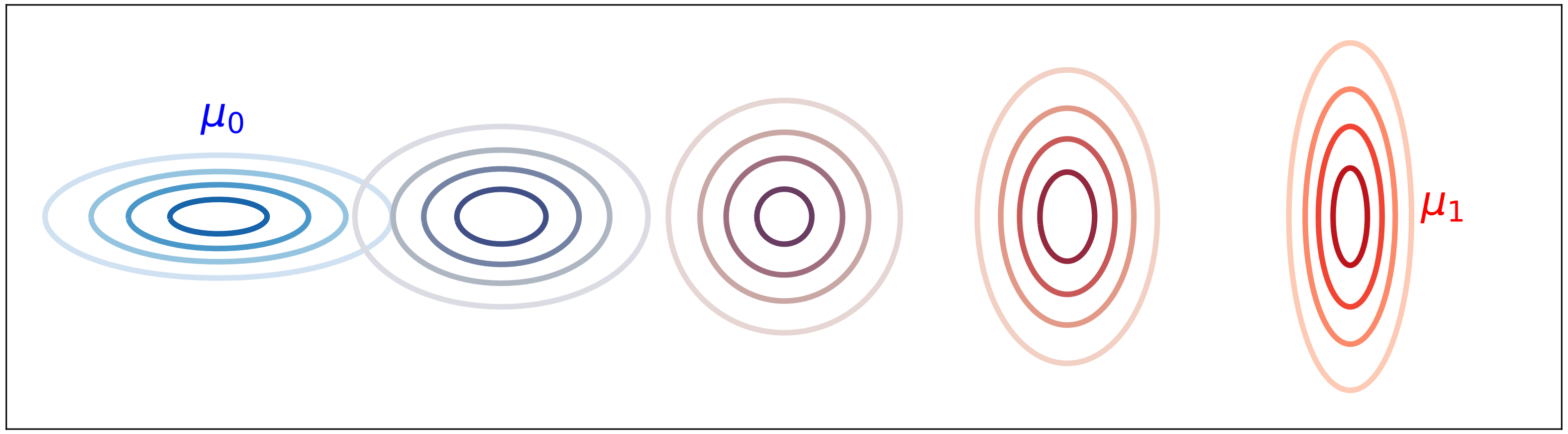
Then $\pi_{MK}(\mu, \nu) = \mathcal{N}(0_{2d}, \Sigma) \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$

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

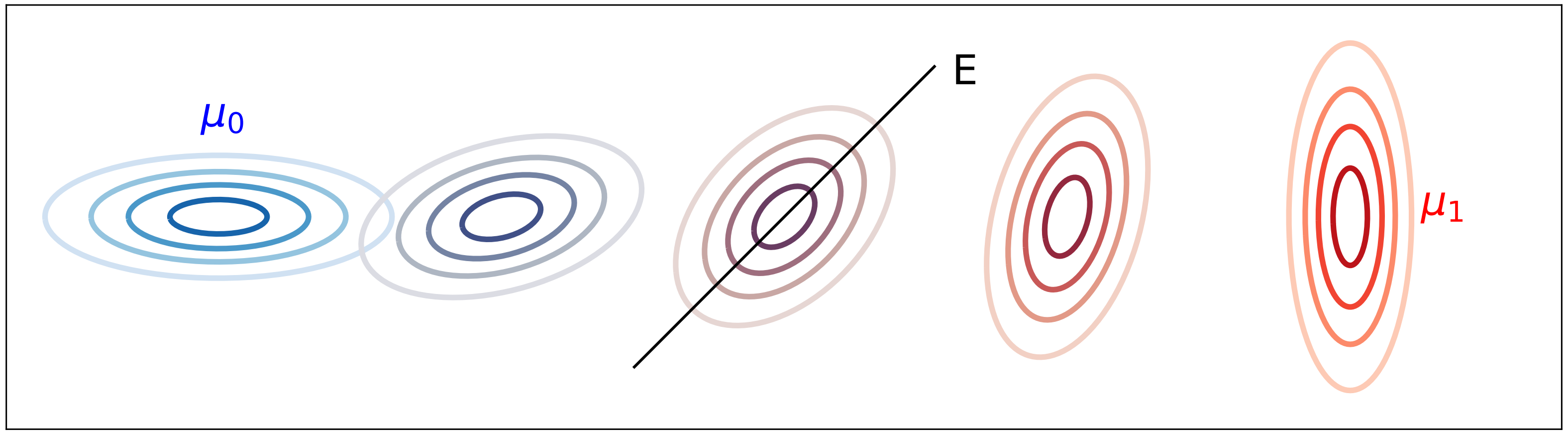
Monge-Knothe: Gaussian Distributions

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

where $\mathbf{A}/\mathbf{A}_E \overset{def}{=} \mathbf{A}_{E^\perp} - \mathbf{A}_{EE^\perp}^\top \mathbf{A}_E^{-1}\mathbf{A}_{EE^\perp}$ **is the Schur complement of** $\mathbf{A}$ **w.r.t.** $\mathbf{A}_E$ **and** $\mathbf{T}^{\mathbf{AB}} \overset{\mathrm{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$



Monge interpolation

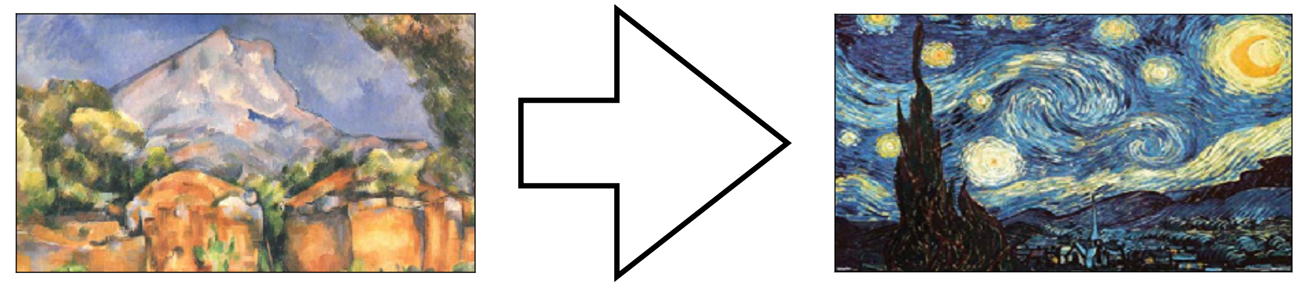


MK interpolation

Application: Color Transfer

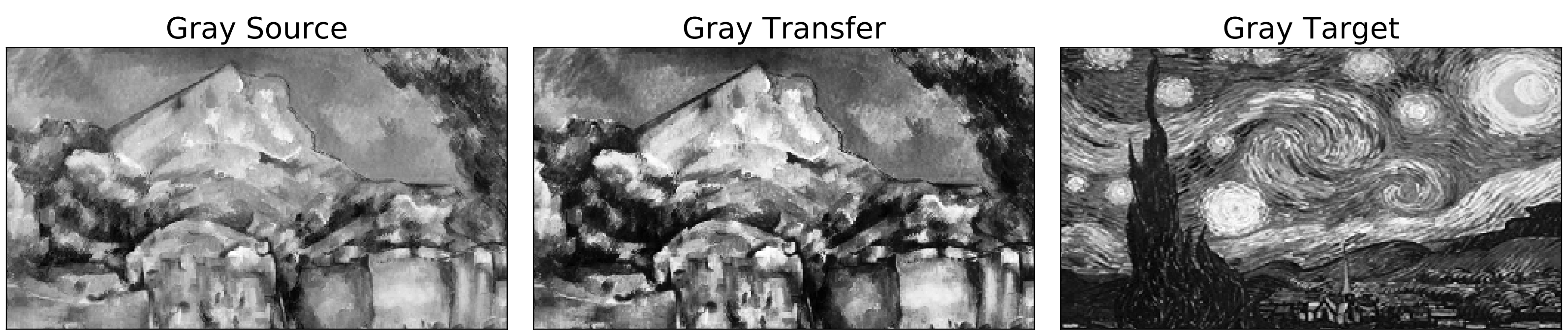
Transform a source image's color palette into that of a target image

- Use an OT map on pixel values

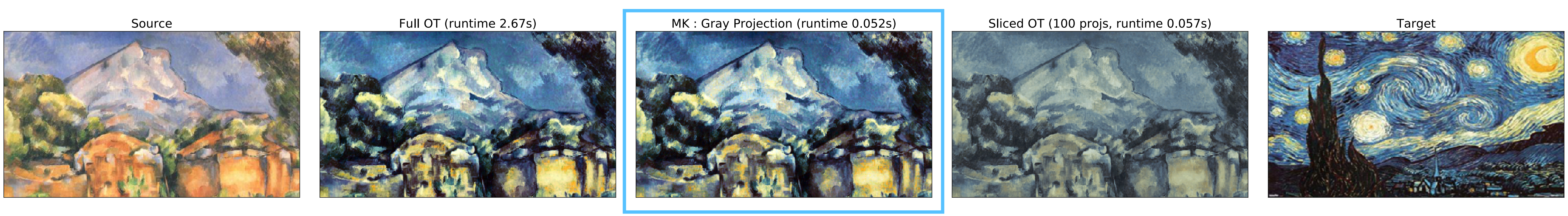


MK approach:

- Compute 1D OT map between grayscale images



- Then extrapolate a full transport map:



Elliptical Word Embeddings

« **Skipgram-like** » **model :**

- **Sliding window of size 10, extract positive pairs** $(w, c) \in \mathscr{R}$
- **Sample negative pairs** $(w, c') \notin \mathscr{R}$

ALL MODELS	ARE	WRONG BUT	SOME ARE USEFUL
ALL	MODELS ARE	WRONG	BUT SOME ARE USEFUL
ALL MODELS	ARE WRONG	BUT	SOME ARE USEFUL

- **Optimize**

$$\min \sum_{(w,c)\in\mathscr{R}} \left[M - \left([\mu_w, \mu_c]_{\mathfrak{B}} - [\mu_w, \mu_{c'}]_{\mathfrak{B}}\right)\right]_+$$

where $[\alpha, \beta]_{\mathfrak{B}} := \langle \mathbf{a}, \mathbf{b} \rangle + \mathrm{Tr}\left(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right)^{\frac{1}{2}}$ **is a Bures generalization of the dot product**

- **Train over *Wackypedia* + *UkWac* : 3 billion tokens**

Application: Semantic Mediation (NLP)

Elliptical word embeddings from [BM&MC'18]:

- each word is represented with a mean vector $\mathbf{m}$ and a PSD matrix Σ

Semantic mediation:

- MK between words *w1*, *w2*, *E* = the k first directions of the SVD of context *c*

Influence of context *c* on the nearest neighbours - Symmetric differences:

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