The Bures-Wasserstein Distance for Machine Learning

Boris Muzellec

*Based joint work with **Marco Cuturi***

ENSAE IP PARIS

Outline

1. A (quick) intro to OT
2. The Bures-Wasserstein distance
3. Optimization with Bures distances
4. Applications

I. An intro to OT

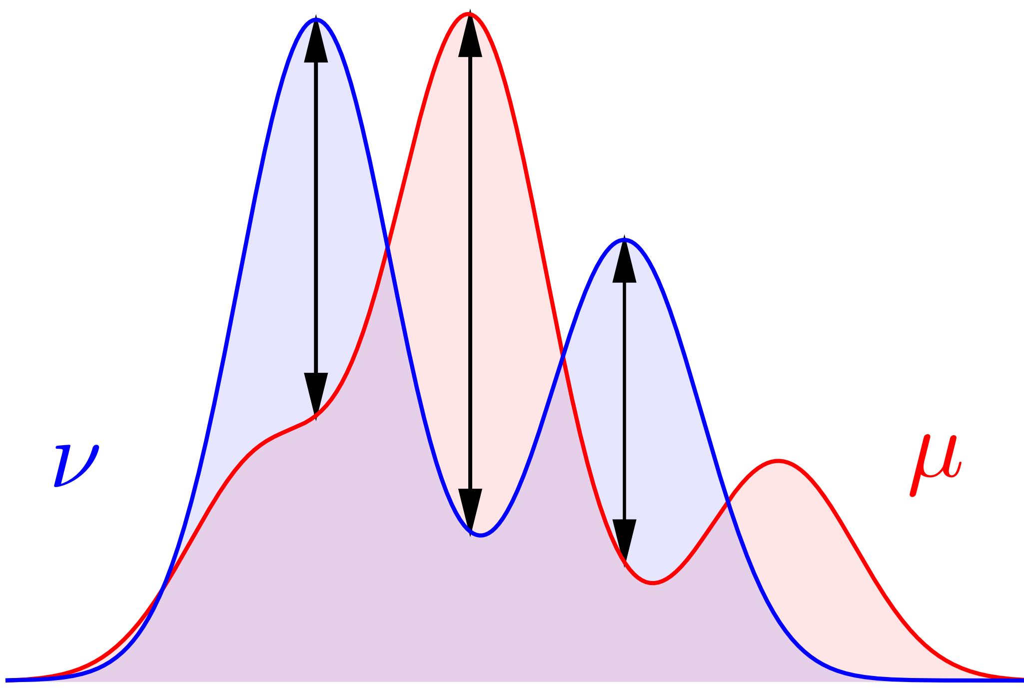
How to compare distributions?

How to compare distributions?

1. **"Vertically":**

   - **Look at differences between densities:**

$$|p(x) - q(x)| \quad \text{or} \quad \frac{p(x)}{q(x)}$$

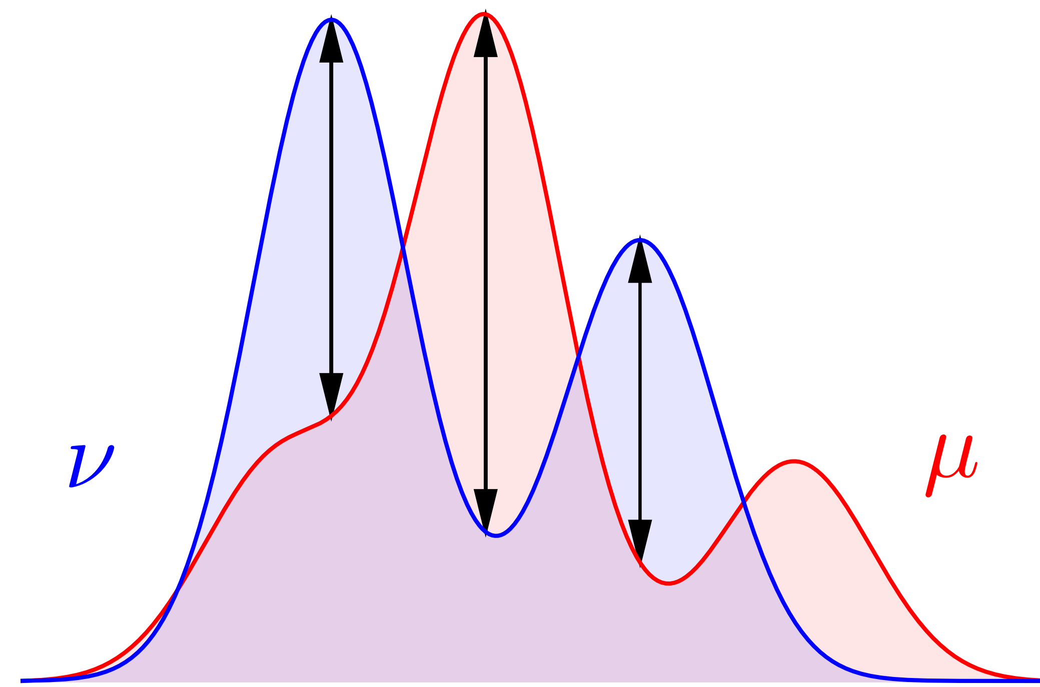


How to compare distributions?

1. **"Vertically":**

- **Look at differences between densities:**

$$|p(x) - q(x)| \quad \text{or} \quad \frac{p(x)}{q(x)}$$



- **Make something useful out of them:**

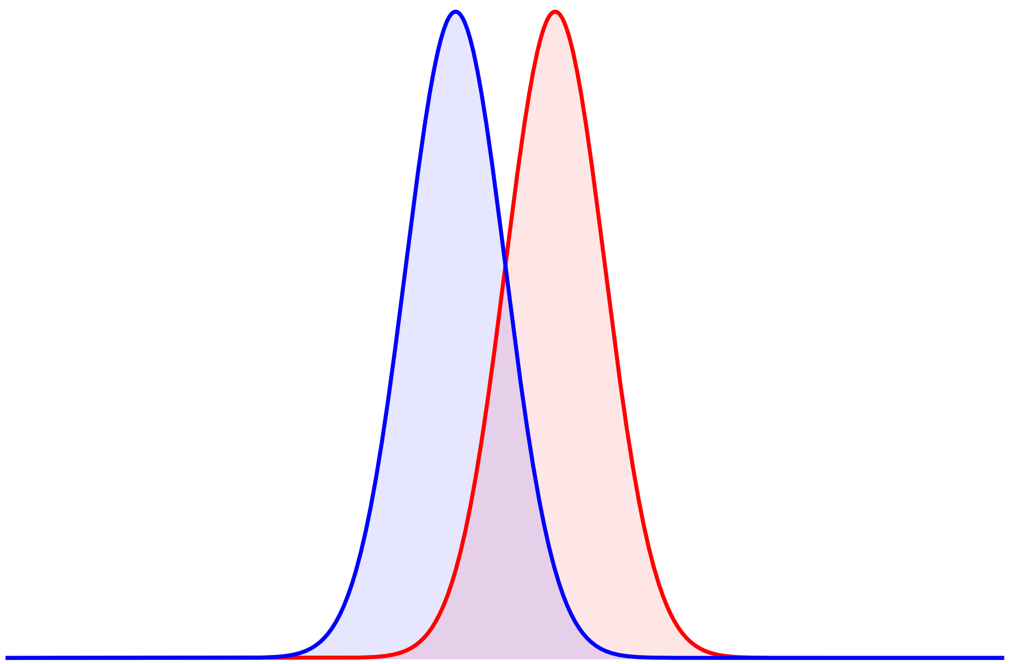
$$\mathrm{TV}(\mu, \nu) = \sup_{A \in \mathcal{B}} \left| \int \mathbb{1}_A(x) p(x) dx - \int \mathbb{1}_A(x) q(x) dx \right|$$ **(Total variation)**

$$D_{\mathrm{KL}}(\mu, \nu) = \int \log \frac{p(x)}{q(x)} p(x) dx$$ **(Kullback-Leibler)**

$$D_f(\mu, \nu) = \int f\left(\frac{p(x)}{q(x)}\right) q(x) dx$$ **(f-divergences)**

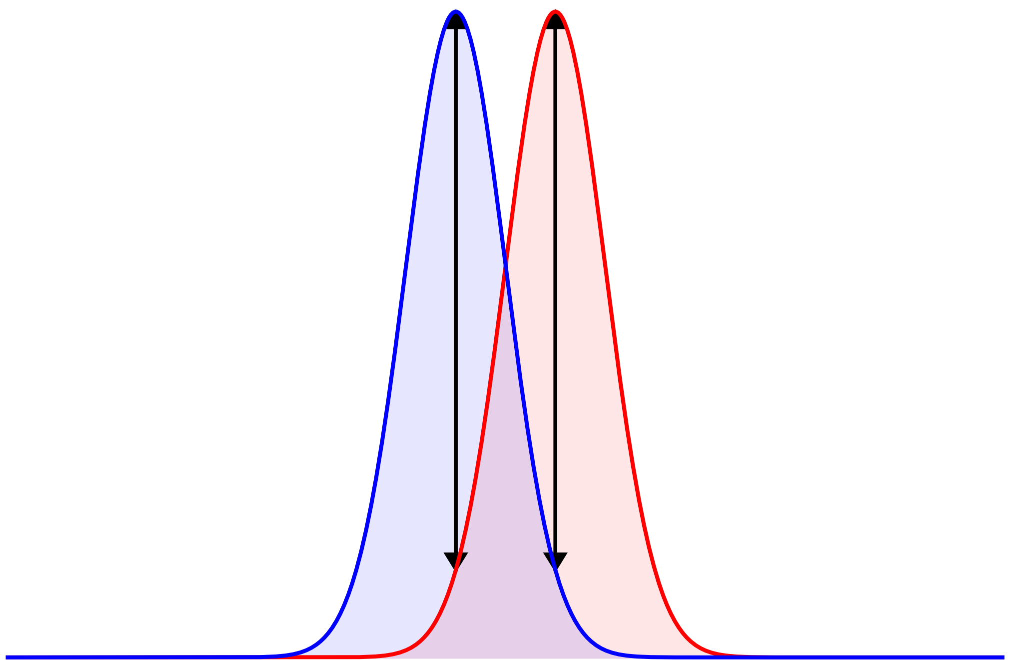
(…)

How to compare distributions?

What about  **?**

How to compare distributions?

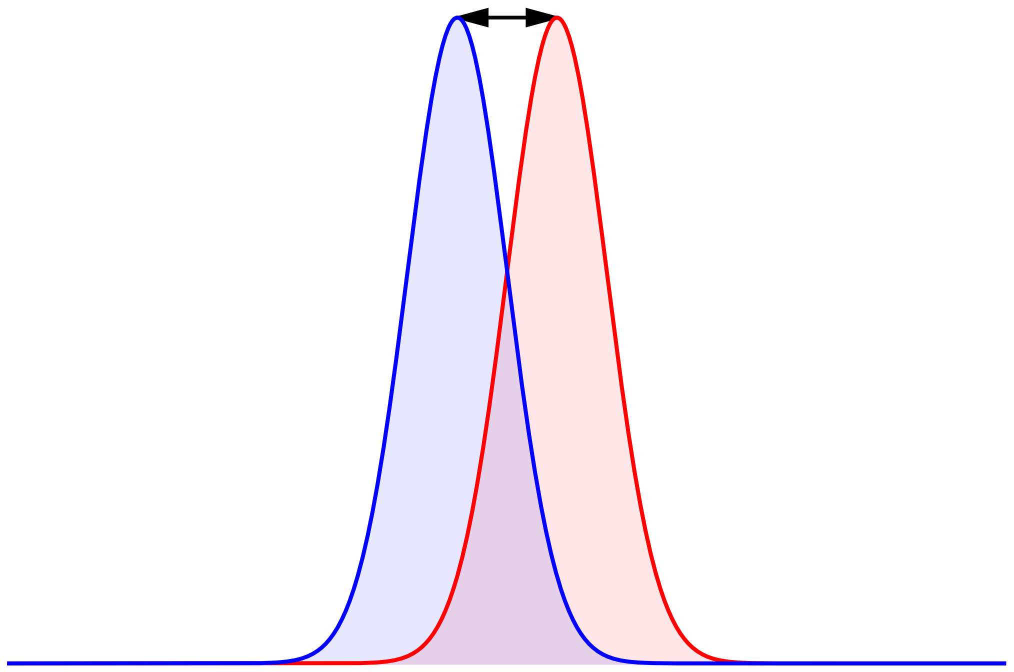
How to compare distributions?



?

How to compare distributions?

How to compare distributions?

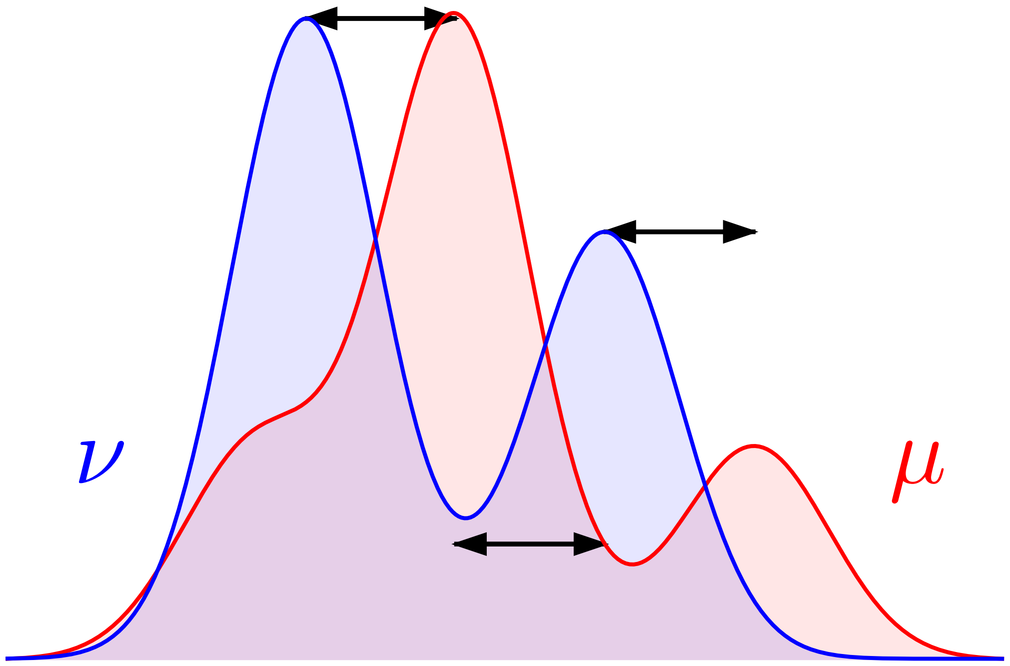
Or  ?

How to compare distributions?

How to compare distributions?

2. **"Horizontally":**

 - **Look at distances on the supports:**

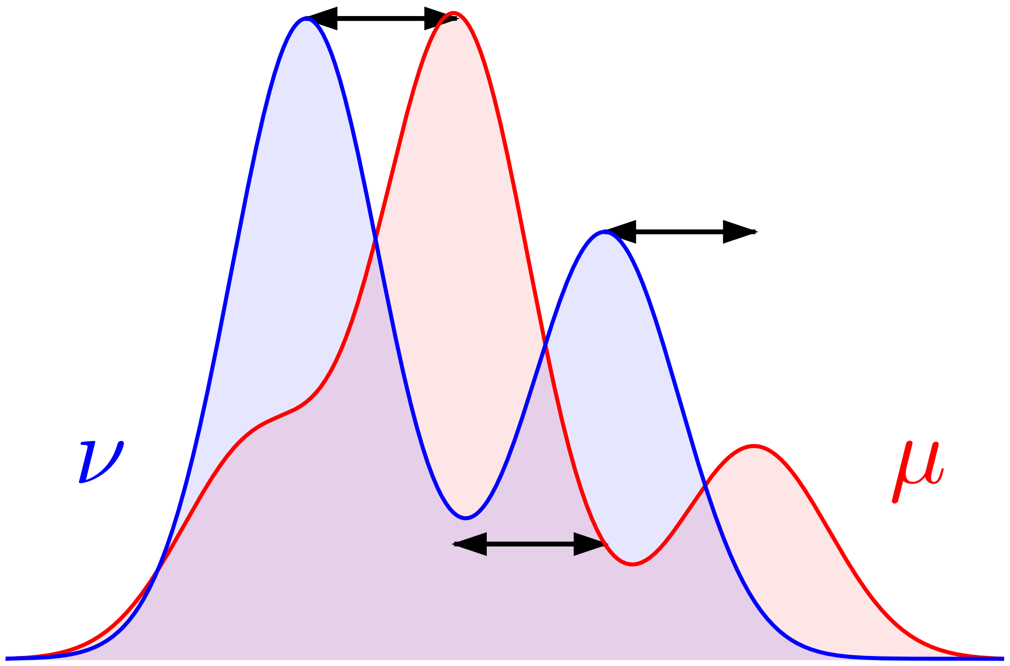


How to compare distributions?

2. **"Horizontally":**

   - **Look at distances on the supports:**

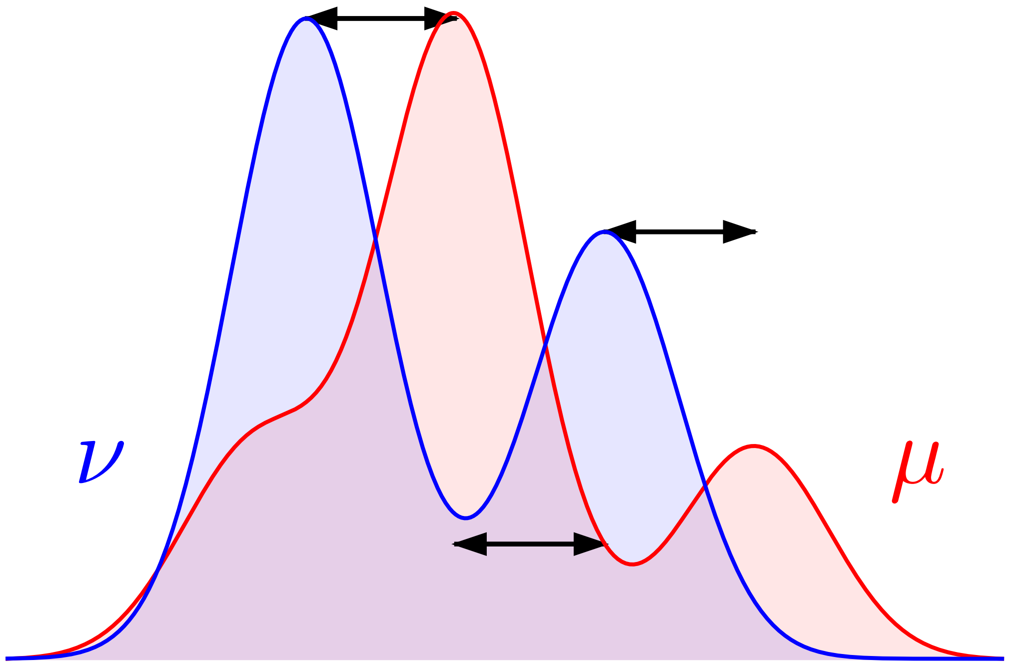
$$\int\int \|x-y\|^2 d\mu(x) d\nu(y) \quad \textbf{?}$$



How to compare distributions?

2. **"Horizontally":**

   - **Look at distances on the supports:**

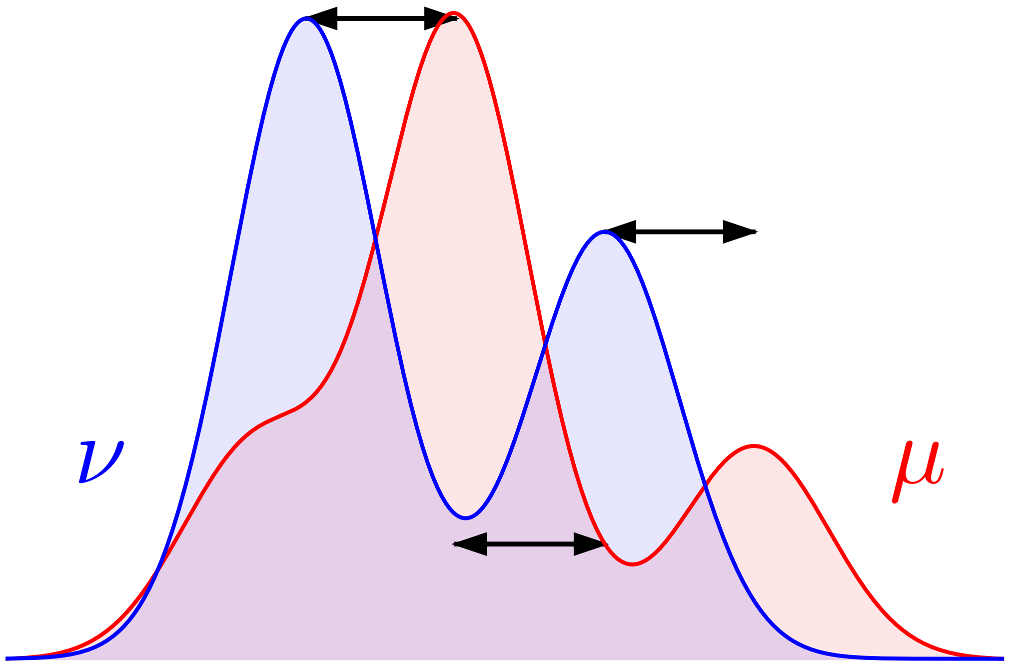


How to compare distributions?

2. "Horizontally":

- **Look at distances on the supports:**

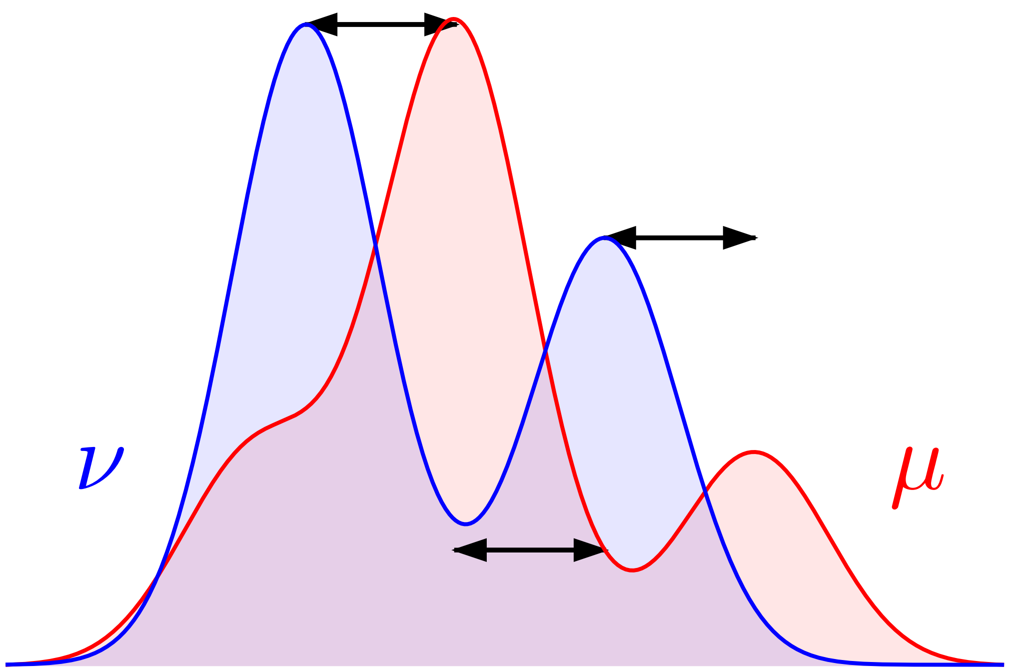
$$\inf_T \int \|x - T(x)\|^2 d\mu(x)$$



How to compare distributions?

2. **"Horizontally":**

   - **Look at distances on the supports:**

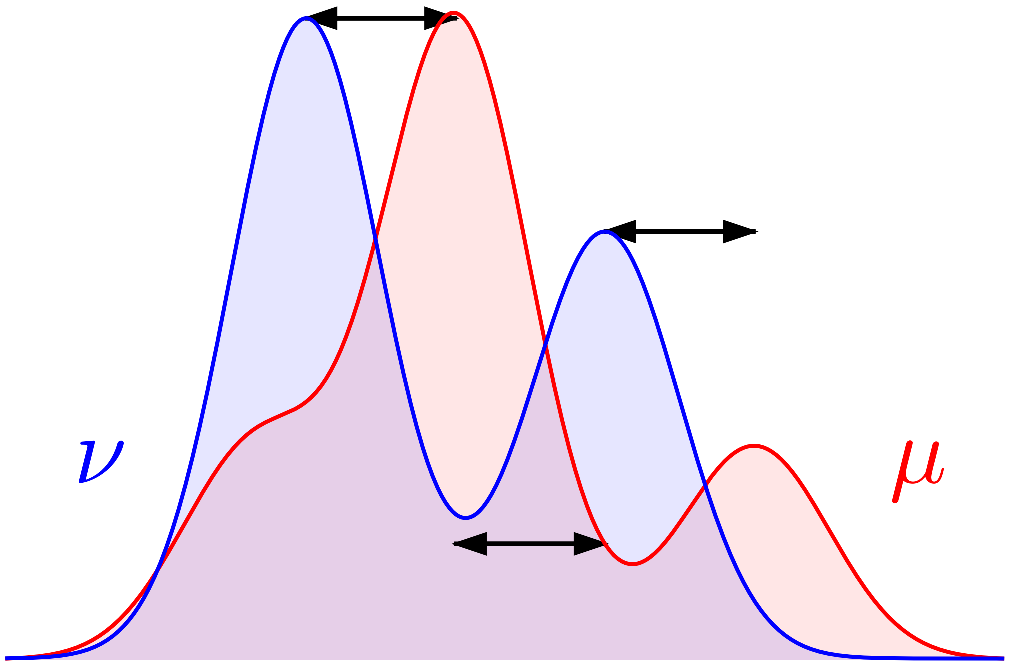


How to compare distributions?

2. **"Horizontally":**

- **Look at distances on the supports:**

$$\inf_{T:T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$



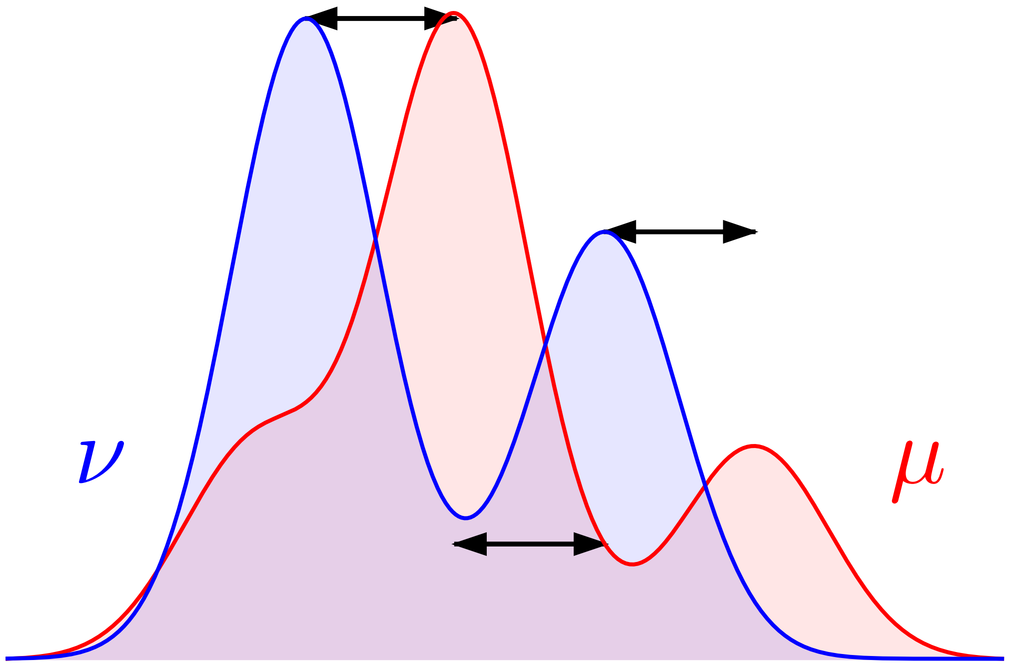
$$T_\sharp \mu = \nu \quad \textit{iff} \quad X \sim \mu \implies T(X) \sim \nu$$

How to compare distributions?

2. “Horizontally”:

- **Look at distances on the supports:**

$$\inf_{T : T_\sharp \mu = \nu} \int \|x - T(x)\|^2 d\mu(x)$$



$$T_\sharp \mu = \nu \quad \textit{iff} \quad X \sim \mu \implies T(X) \sim \nu$$

“T pushes forward μ to ν”

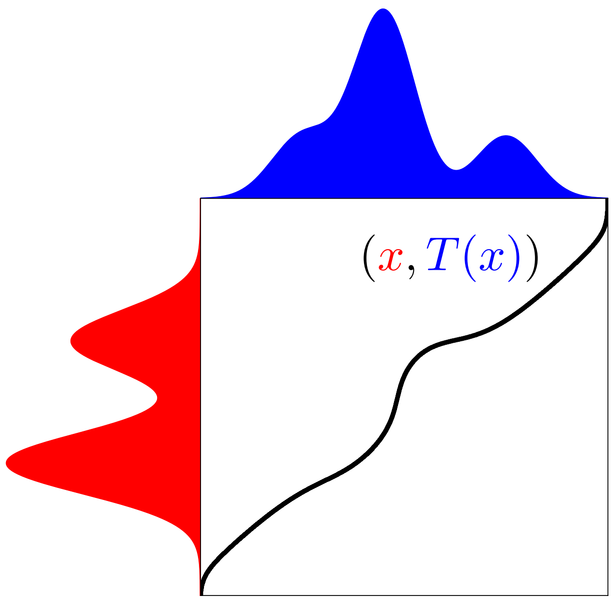
“T is a Monge map from μ to ν”

(2-)Wasserstein Distances

- **Monge version**

Prop. When a Monge map T exists,

$$W_2^2(\mu, \nu) = \inf_{T_\sharp \mu = \nu} \int_\Omega \|x - T(x)\|^2 \mu(dx)$$

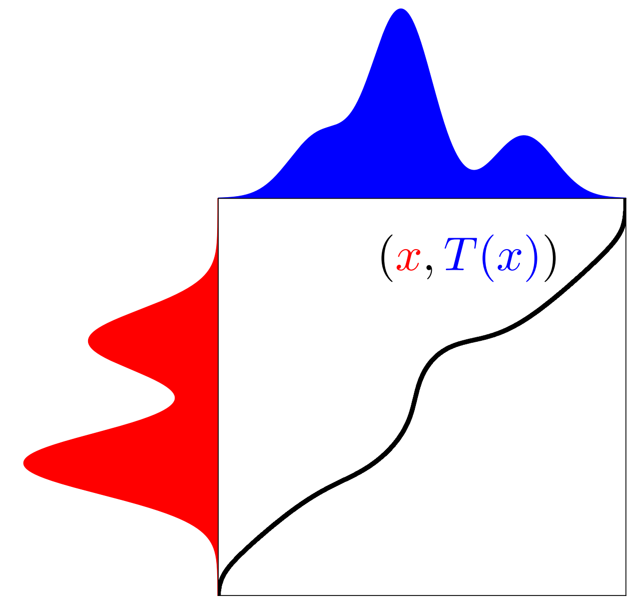


(2-)Wasserstein Distances

- **Monge version**

Prop. When a Monge map T exists,

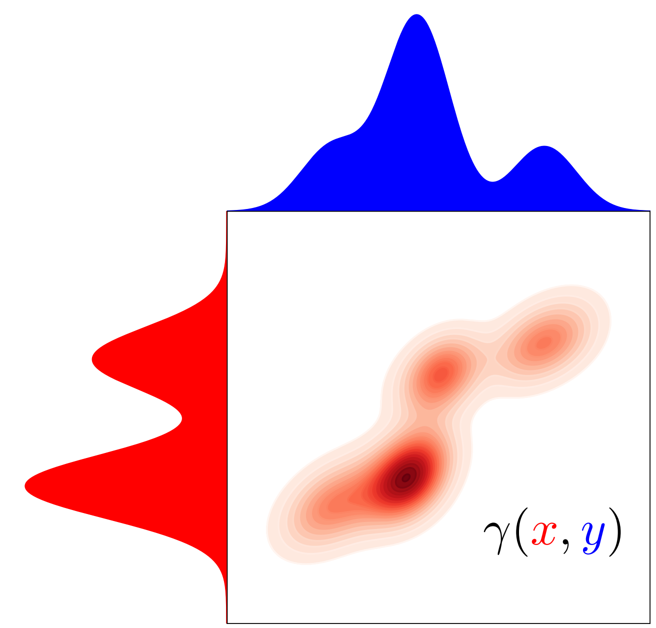
$$W_2^2(\mu,\nu) = \inf_{T_\sharp\mu=\nu} \int_\Omega \|x - T(x)\|^2 \mu(dx)$$



- **Kantorovich version**

Def. The 2-Wasserstein distance between $\mu, \nu \in P(\Omega)$ is

$$W_2^2(\mu,\nu) \stackrel{\text{def}}{=} \inf_{\gamma\in\Pi(\mu,\nu)} \int_\Omega \|x - y\|^2 d\gamma(x,y)$$



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

"Couplings"

"Kantorovich / transportation plans"

Monge maps: existence

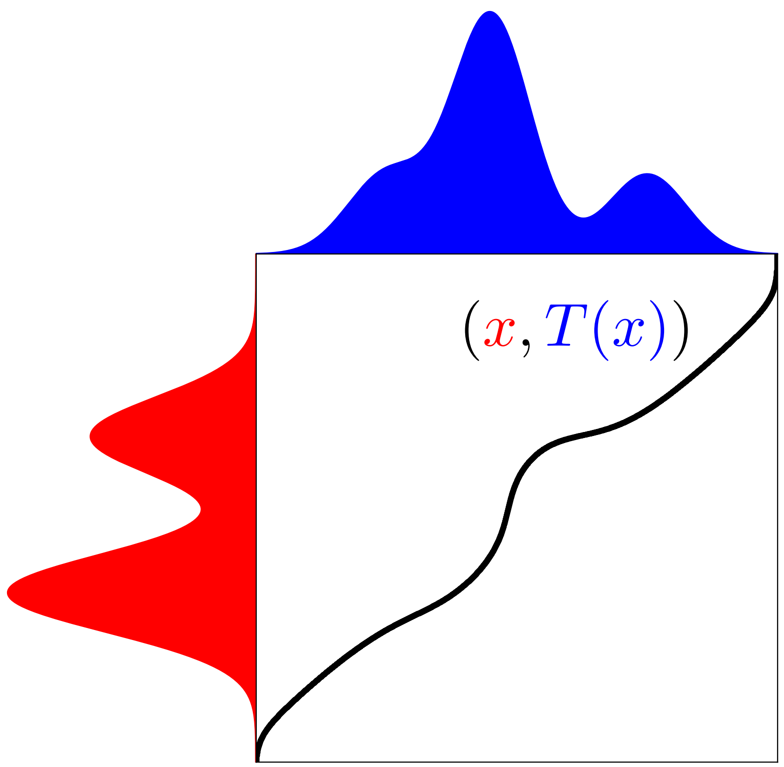
Prop. For "well behaved" costs c, if μ has a density then an *optimal* Monge map $T^\star$ between μ and ν must exist.

Monge maps: existence

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

- **Link between Monge maps and Kantorovich plans:**

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



How to compute Wasserstein distances?

How to compute Wasserstein distances?

- **Discrete/Discrete:**
  - **LP with** $O(n^3 \log n)$ **complexity using network simplex**
  - **Better with (entropic) regularization [Cuturi'13, Genevay et al.'16, Altschuler et al.'17…]**

How to compute Wasserstein distances?

- **Discrete/Discrete:**
  - **LP with $O(n^3 \log n)$ complexity using network simplex**
  - **Better with (entropic) regularization [Cuturi'13, Genevay et al.'16, Altschuler et al.'17…]**

- **Discrete/Continous:**
  - **Ok-ish… (Laguerre tesselations)**

How to compute Wasserstein distances?

- **Discrete/Discrete:**
    - **LP with** $O(n^3 \log n)$ **complexity using network simplex**
    - **Better with (entropic) regularization** **[Cuturi'13, Genevay et al.'16, Altschuler et al.'17…]**

- **Discrete/Continous:**
    - **Ok-ish… (Laguerre tesselations)**

- **Continuous/Continous: ?**
    - **Closed form: elliptical distributions (next slides)**

II. The Wasserstein-Bures Distance

Elliptical Distributions

« Def. » Probability measures with densities

$$f(\mathbf{x}) = \frac{1}{\sqrt{|\mathbf{C}|}} h((\mathbf{x}-\mathbf{m})^\top \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}))$$

where $\int_{\mathbb{R}^d} h(\|\mathbf{x}\|^2)\mathrm{d}\mathbf{x} = 1, \quad \mathbf{C} \in S_n^+$

Examples:

- Multivariate normal distributions
- Elliptical uniform distributions
- (Multivariate) t-Student…

OT for Elliptical Distributions

[Gelbrich'90]

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

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

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

OT for Elliptical Distributions

[Gelbrich'90]

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

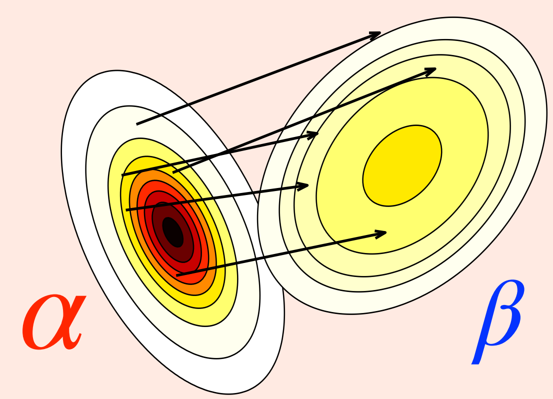
$$W_2^2(\alpha, \beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|_2^2 + \mathfrak{B}^2(\mathrm{cov}\alpha, \mathrm{cov}\beta)$$

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

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



$T_\sharp \alpha = \beta$

A lower bound

- **What if α, β are not elliptical?**

A lower bound

- **What if α, β are not elliptical?**

Prop. **Wasserstein-Bures is a lower bound of Wasserstein.**

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

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

A Lemma

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

Lemma. [Bhatia et al.'17]

$$F(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

$$= \max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\}$$

Lower bound

Prop.

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

With $\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \overset{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2F(\mathbf{A},\mathbf{B})$ $\quad F(\mathbf{A},\mathbf{B}) = \max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\}$

Proof.
(centered case)

Lower bound

Prop.

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

With $\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2F(\mathbf{A},\mathbf{B})$ $\quad F(\mathbf{A},\mathbf{B}) = \max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\}$

Proof.
(centered case)

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

$$= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2 \max_{\gamma\in\Pi(\mu,\nu)} \mathrm{Tr}[Cov_\gamma(X,Y)]$$

Lower bound

Prop.

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

With $\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2F(\mathbf{A},\mathbf{B})$ $\quad F(\mathbf{A},\mathbf{B}) = \max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\}$

Proof.
(centered case)

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

$$= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2 \max_{\gamma \in \Pi(\mu,\nu)} \mathrm{Tr}[Cov_\gamma(X,Y)]$$

But $\gamma \in \Pi(\mu,\nu) \implies \mathrm{cov}(\gamma) = \begin{pmatrix} \mathbf{A} & Cov_\gamma(X,Y) \\ Cov_\gamma(X,Y)^T & \mathbf{B} \end{pmatrix} \geq 0$

Lower bound

Prop.

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

With $\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A}+\mathrm{Tr}\mathbf{B}-2F(\mathbf{A},\mathbf{B})$ $\quad F(\mathbf{A},\mathbf{B}) = \max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\}$

Proof.
(centered case)

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

$$= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\max_{\gamma\in\Pi(\mu,\nu)} \mathrm{Tr}[Cov_\gamma(X,Y)]$$

But $\gamma \in \Pi(\mu,\nu) \implies \mathrm{cov}(\gamma) = \begin{pmatrix} \mathbf{A} & Cov_\gamma(X,Y) \\ Cov_\gamma(X,Y)^T & \mathbf{B} \end{pmatrix} \geq 0$

Hence $\forall\gamma \in \Pi(\mu,\nu), \quad \mathrm{Tr}[Cov_\gamma(X,Y)] \leq F(\mathbf{A},\mathbf{B}) \quad \blacksquare$

How tight is this bound?

- **Q: Is there an equality case?**

- **Q: (Matching) upper bound?**

How tight is this bound?

- **Q: Is there an equality case?**
  - **A: Yes —> Elliptical distributions**

- **Q: (Matching) upper bound?**

How tight is this bound?

- **Q: Is there an equality case?**
  - **A: Yes —> Elliptical distributions**

- **Q: (Matching) upper bound?**
  - **A: …   (independent coupling)**

$$W_2^2(\mu, \nu) \leq \|\mathbf{m}_\mu - \mathbf{m}_\nu\|_2^2 + \mathrm{Tr}\,\mathbf{A} + \mathrm{Tr}\,\mathbf{B}$$

Equality Case

Lemma. [Bhatia et al.'17]

$$\arg\max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\} = (\mathbf{A}\mathbf{B})^{\frac{1}{2}} = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$$

γ, μ, ν **such that** $Cov_\gamma(X, Y) = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$ **?**

Equality Case

Lemma. [Bhatia et al.'17]

$$\arg\max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\} = (\mathbf{A}\mathbf{B})^{\frac{1}{2}} = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$$

γ, μ, ν **such that** $Cov_\gamma(X, Y) = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$ **?**

$$\mathrm{rk}[\mathrm{cov}(\gamma)] = \mathrm{rk}\begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}} \\ \mathbf{T}^{\mathbf{A}\mathbf{B}}\mathbf{A} & \mathbf{B} \end{pmatrix} = d \quad (< 2d)$$

Equality Case

Lemma. [Bhatia et al.'17]

$$\arg\max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\} = (\mathbf{A}\mathbf{B})^{\frac{1}{2}} = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$$

γ, μ, ν **such that** $Cov_\gamma(X, Y) = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$ **?**

$$\mathrm{rk}[\mathrm{cov}(\gamma)] = \mathrm{rk}\begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}} \\ \mathbf{T}^{\mathbf{A}\mathbf{B}}\mathbf{A} & \mathbf{B} \end{pmatrix} = d \quad (< 2d)$$

- γ **is the law of** (X, Y) **with** $X \sim \mu, Y \sim \nu$ **and** $Y = \mathbf{T}^{\mathbf{A}\mathbf{B}}X$**.**

Equality Case

Lemma. [Bhatia et al.'17]

$$\arg\max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\} = (\mathbf{AB})^{\frac{1}{2}} = \mathbf{A}\mathbf{T}^{\mathbf{AB}}$$

γ, μ, ν **such that** $Cov_\gamma(X, Y) = \mathbf{A}\mathbf{T}^{\mathbf{AB}}$ **?**

$$\mathrm{rk}[\mathrm{cov}(\gamma)] = \mathrm{rk}\begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{T}^{\mathbf{AB}} \\ \mathbf{T}^{\mathbf{AB}}\mathbf{A} & \mathbf{B} \end{pmatrix} = d \quad (< 2d)$$

- γ **is the law of** (X, Y) **with** $X \sim \mu, Y \sim \nu$ **and** $Y = \mathbf{T}^{\mathbf{AB}}X$**.**
- **Implies** $\nu = (\mathbf{T}^{\mathbf{AB}})_\sharp\mu$ **and** $\mathbf{T}^{\mathbf{AB}}\mathbf{A}\mathbf{T}^{\mathbf{AB}} = \mathbf{B}$ **(Riccati equation).**

Equality Case

Lemma. [Bhatia et al.'17]

$$\arg\max\{\mathrm{tr}\mathbf{X} : \begin{pmatrix} \mathbf{A} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{B} \end{pmatrix} \geq 0\} = (\mathbf{A}\mathbf{B})^{\frac{1}{2}} = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$$

γ, μ, ν **such that** $Cov_\gamma(X, Y) = \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}}$ **?**

$$\mathrm{rk}[\mathrm{cov}(\gamma)] = \mathrm{rk}\begin{pmatrix} \mathbf{A} & \mathbf{A}\mathbf{T}^{\mathbf{A}\mathbf{B}} \\ \mathbf{T}^{\mathbf{A}\mathbf{B}}\mathbf{A} & \mathbf{B} \end{pmatrix} = d \quad (< 2d)$$

- γ **is the law of** (X, Y) **with** $X \sim \mu, Y \sim \nu$ **and** $Y = \mathbf{T}^{\mathbf{AB}}X$.
- **Implies** $\nu = (\mathbf{T}^{\mathbf{AB}})_\sharp \mu$ **and** $\mathbf{T}^{\mathbf{AB}}\mathbf{A}\mathbf{T}^{\mathbf{AB}} = \mathbf{B}$ **(Riccati equation).**
- **e.g.** μ, ν **are from the same** ***elliptical family*****.**

Elliptical Distributions

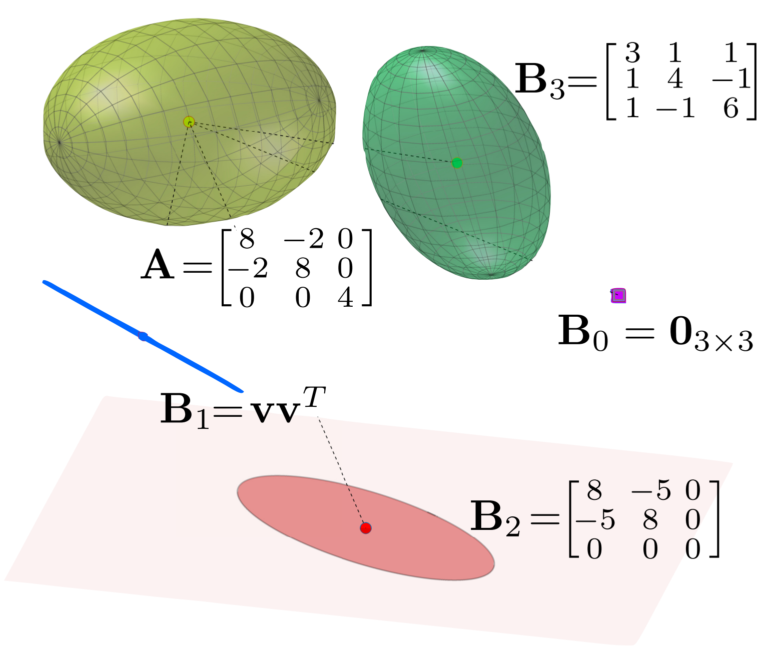
« Def. » Probability measures with densities

$$f(\mathbf{x}) = \frac{1}{\sqrt{|\mathbf{C}|}} h((\mathbf{x}-\mathbf{m})^\top \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}))$$

where $\int_{\mathbb{R}^d} h(\|\mathbf{x}\|^2)\mathrm{d}\mathbf{x} = 1, \quad \mathbf{C} \in S_n^+$

Examples:

- **Multivariate normal distributions**
- **Elliptical uniform distributions**
- **(Multivariate) t-Student…**



III. Working with the Bures distance

Issues

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

$$= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}\mathbf{B})^{\frac{1}{2}}$$

1. **How to compute matrix roots (in a scalable way)?**
2. **How to compute gradients?**
3. **Can I avoid projections on the PSD cone?**

How (not) to compute roots?

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

How (not) to compute roots?

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

- **Option 1: SVD**
    - $O(n^3)$ **complexity**
    - **Batched version?**

How (not) to compute roots?

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

How (not) to compute roots?

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

- **Option 2: Iterations? e.g.**
  - **Babylonian algorithm** $\quad \mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{X}_k^{-1}\mathbf{A}), \qquad \mathbf{X}_0 = \mathbf{A}$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}} \quad (\text{if } \max_{ij} \tfrac{1}{2}|1 - \lambda_i^{1/2}\lambda_j^{-1/2}| < 1)$$

How (not) to compute roots?

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

- **Option 2: Iterations? e.g.**
  - **Babylonian algorithm** $\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{X}_k^{-1}\mathbf{A}), \qquad \mathbf{X}_0 = \mathbf{A}$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}} \quad (\text{if } \max_{ij} \tfrac{1}{2}|1-\lambda_i^{1/2}\lambda_j^{-1/2}| < 1)$$

  - **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{Y}_k^{-1}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

From DB to Newton-Schulz

- **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{Y}_k^{-1}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

From DB to Newton-Schulz

- **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{Y}_k^{-1}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

- **Inverse is costly. However, we expect** $\mathbf{Y}_k^{-1} \simeq \mathbf{X}_k$

From DB to Newton-Schulz

- **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{Y}_k^{-1}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

- **Inverse is costly. However, we expect $\mathbf{Y}_k^{-1} \simeq \mathbf{X}_k$**
    - **Approximate $\mathbf{Y}_k^{-1}$ using one Newton iteration for the inverse:**

$$\mathbf{Y}_k^{-1} \simeq 2\mathbf{X}_k - \mathbf{X}_k\mathbf{Y}_k\mathbf{X}_k$$

$$(f(x) = 1/x - y, \quad x_{n+1} = x_n - f(x)/f'(x) = x_n - \frac{1/x_n - y}{-1/x_n^2} = 2x_n - x_n^2 y)$$

From DB to Newton-Schulz

- **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \mathbf{Y}_k^{-1}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

- **Inverse is costly. However, we expect $\mathbf{Y}_k^{-1} \simeq \mathbf{X}_k$**
    - **Approximate $\mathbf{Y}_k^{-1}$ using one Newton iteration for the inverse:**

$$\mathbf{Y}_k^{-1} \simeq 2\mathbf{X}_k - \mathbf{X}_k\mathbf{Y}_k\mathbf{X}_k$$

$$\left(f(x) = 1/x - y, \quad x_{n+1} = x_n - f(x)/f'(x) = x_n - \frac{1/x_n - y}{-1/x_n^2} = 2x_n - x_n^2 y\right)$$

From DB to Newton-Schulz

- **Denman-Beavers**

$$\mathbf{X}_{k+1} = \frac{1}{2}(\mathbf{X}_k + \boxed{\mathbf{Y}_k^{-1}}), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(\mathbf{Y}_k + \mathbf{X}_k^{-1}), \qquad \mathbf{Y}_0 = \mathbf{I}$$

$$\lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$$

- **Inverse is costly. However, we expect $\mathbf{Y}_k^{-1} \simeq \mathbf{X}_k$**
    - **Approximate $\boxed{\mathbf{Y}_k^{-1}}$ using one Newton iteration for the inverse:**

$$\mathbf{Y}_k^{-1} \simeq 2\mathbf{X}_k - \mathbf{X}_k\mathbf{Y}_k\mathbf{X}_k$$

$$(f(x) = 1/x - y, \quad x_{n+1} = x_n - f(x)/f'(x) = x_n - \frac{1/x_n - y}{-1/x_n^2} = 2x_n - x_n^2 y)$$

- **Do the same thing with $\mathbf{X}_k^{-1} \simeq \mathbf{Y}_k$: Newton-Schulz algorithm (next slide).**

How to compute roots

- **Newton-Schulz square root iterations:**

$$\mathbf{X}_{k+1} = \frac{1}{2}\mathbf{X}_k(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k)\mathbf{Y}_k, \qquad \mathbf{Y}_0 = \mathbf{I}$$

How to compute roots

- **Newton-Schulz square root iterations:**

$$\mathbf{X}_{k+1} = \frac{1}{2}\mathbf{X}_k(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k), \qquad \mathbf{X}_0 = \mathbf{A}$$

$$\mathbf{Y}_{k+1} = \frac{1}{2}(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k)\mathbf{Y}_k, \qquad \mathbf{Y}_0 = \mathbf{I}$$

Prop. [Higham'08]

If $\|\mathbf{I} - \mathbf{A}\| < 1$**,** $\quad \lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$

with *quadratic* convergence.

How to compute roots

- **Newton-Schulz square root iterations:**

$$\mathbf{X}_{k+1} = \frac{1}{2}\mathbf{X}_k(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k), \qquad \mathbf{X}_0 = \mathbf{A}$$
$$\mathbf{Y}_{k+1} = \frac{1}{2}(3\mathbf{I} - \mathbf{Y}_k\mathbf{X}_k)\mathbf{Y}_k, \qquad \mathbf{Y}_0 = \mathbf{I}$$

Prop. [Higham'08]

If $\|\mathbf{I} - \mathbf{A}\| < 1$**,** $\quad \lim_{k\to\infty} \mathbf{X}_k = \mathbf{A}^{\frac{1}{2}}, \qquad \lim_{k\to\infty} \mathbf{Y}_k = \mathbf{A}^{-\frac{1}{2}}$

with *quadratic* convergence.

- **GPU friendly (batch matrix-matrix multiplications)**
- ***Gives simultaneously the square root and its inverse***

Issues

$$\mathfrak{B}^2(\mathbf{A},\mathbf{B}) \stackrel{\text{def}}{=} \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$
$$= \mathrm{Tr}\mathbf{A} + \mathrm{Tr}\mathbf{B} - 2\mathrm{Tr}(\mathbf{A}\mathbf{B})^{\frac{1}{2}}$$

1. **How to compute matrix roots (in a scalable way)?**
2. **How to compute gradients?**
3. **Can I avoid projections on the PSD cone?**

How to compute the Bures Gradient?

How to compute the Bures Gradient?

Option 1: Automatic differentiation

- **Has the same cost as computing** $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$
- **Gives the *exact* gradient of the *approximated* distance**

How to compute the Bures Gradient?

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇨ **we need** $\mathbf{T}^{\mathbf{AB}}$ **and** $\mathbf{T}^{\mathbf{BA}}$

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ➡ **we need** $\mathbf{T}^{\mathbf{AB}}$ **and** $\mathbf{T}^{\mathbf{BA}}$

The naive way: $\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}, \quad \mathbf{T}^{\mathbf{BA}} = \mathbf{B}^{-\frac{1}{2}}(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{B}^{-\frac{1}{2}}$

We need: $\{\mathbf{A}^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}\}, \quad \{\mathbf{B}^{\frac{1}{2}}, \mathbf{B}^{-\frac{1}{2}}\}, \quad \{(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\}, \quad \{(\mathbf{B}^{\frac{1}{2}}\mathbf{A}\mathbf{B}^{\frac{1}{2}})^{\frac{1}{2}}\}$

4 runs of Newton-Schulz

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇨ **we need $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$**

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇒ **we need $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$**

Prop.

$$\begin{aligned} \mathbf{T}^{\mathbf{AB}} &= \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}} \\ &= \mathbf{B}^{\frac{1}{2}} (\mathbf{B}^{\frac{1}{2}} \mathbf{A} \mathbf{B}^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{B}^{\frac{1}{2}} \end{aligned}$$

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ and $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇨ **we need $\mathbf{T}^{\mathbf{AB}}$ and $\mathbf{T}^{\mathbf{BA}}$**

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ➡ **we need** $\mathbf{T}^{\mathbf{AB}}$ **and** $\mathbf{T}^{\mathbf{BA}}$

The *better* way: $\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}, \quad \mathbf{T}^{\mathbf{BA}} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = (\mathbf{T}^{\mathbf{AB}})^{-1}$

We need: $\{\mathbf{A}^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}\}, \quad \{(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}, (\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}}\}$

2 runs of Newton-Schulz

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}$$

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇒ **we need** $\mathbf{T}^{\mathbf{AB}}$ **and** $\mathbf{T}^{\mathbf{BA}}$

The *better* way: $\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}} \mathbf{A}^{-\frac{1}{2}}, \quad \mathbf{T}^{\mathbf{BA}} = \mathbf{A}^{\frac{1}{2}} (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}} \mathbf{A}^{\frac{1}{2}} = (\mathbf{T}^{\mathbf{AB}})^{-1}$

We need: $\{\mathbf{A}^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}\}, \quad \{(\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}, (\mathbf{A}^{\frac{1}{2}} \mathbf{B} \mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}}\}$

0 if we computed $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **earlier**

How to compute the Bures Gradient?

$$\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}, \qquad \mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$$

- **[BM&Cuturi'18]**

In most applications, we need both $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **and** $\nabla_{\mathbf{B}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$

Option 2: Closed form & a nice hack

- $\nabla_{\mathbf{A}} \mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathbf{I} - \mathbf{T}^{\mathbf{AB}}$ ⇨ **we need** $\mathbf{T}^{\mathbf{AB}}$ **and** $\mathbf{T}^{\mathbf{BA}}$

The *better* way: $\mathbf{T}^{\mathbf{AB}} = \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}, \quad \mathbf{T}^{\mathbf{BA}} = \mathbf{A}^{\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = (\mathbf{T}^{\mathbf{AB}})^{-1}$

We need: $\{\mathbf{A}^{\frac{1}{2}}, \mathbf{A}^{-\frac{1}{2}}\}, \quad \{(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}, (\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{-\frac{1}{2}}\}$

0 if we computed $\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **earlier**

Issues

$$\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \stackrel{\text{def}}{=} \operatorname{Tr}\mathbf{A} + \operatorname{Tr}\mathbf{B} - 2\operatorname{Tr}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}$$

$$= \operatorname{Tr}\mathbf{A} + \operatorname{Tr}\mathbf{B} - 2\operatorname{Tr}(\mathbf{A}\mathbf{B})^{\frac{1}{2}}$$

1. **How to compute matrix roots (in a scalable way)?**
2. **How to compute gradients?**
3. **Can I avoid projections on the PSD cone?**

Can we avoid projections?

- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **is not necessarily PSD.**

Can we avoid projections?

- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A},\mathbf{B})$ **is not necessarily PSD.**

- **Classic workaround:** $\mathbf{A} = \Pi(\mathbf{L}_\mathbf{A}) \stackrel{\text{def}}{=} \mathbf{L}_\mathbf{A}\mathbf{L}_\mathbf{A}^\top$**. Effect on gradient methods?**

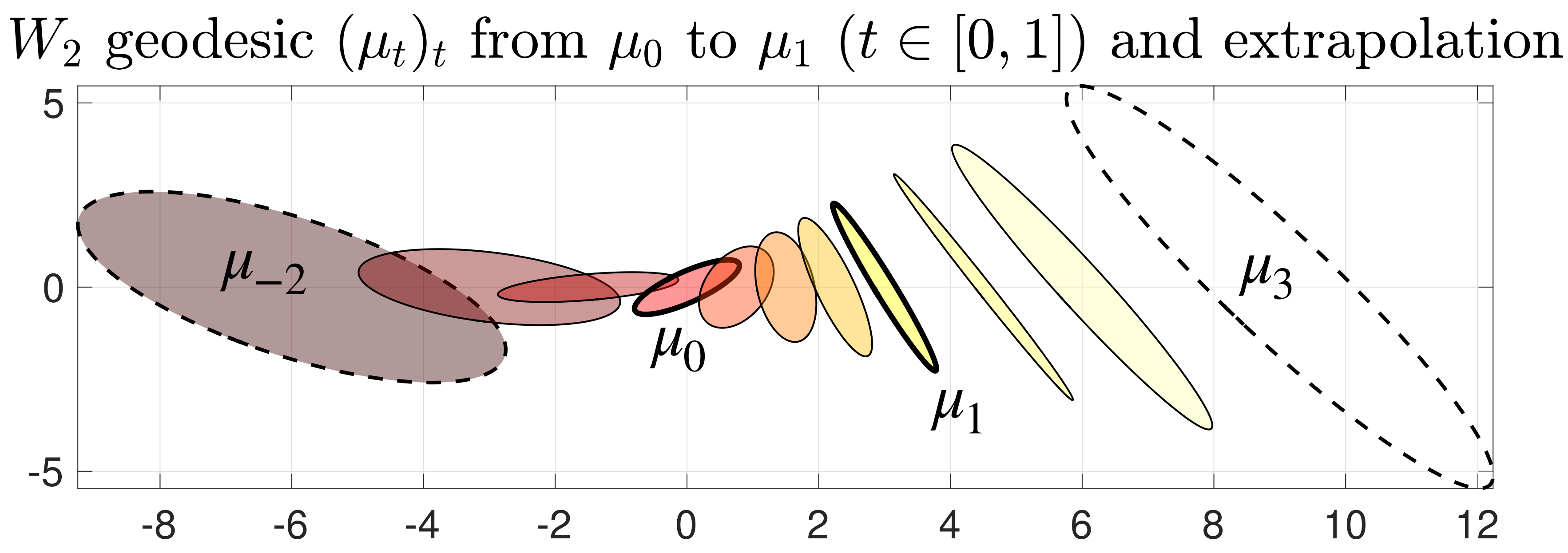
$$\nabla_{\mathbf{L}_\mathbf{A}} \tfrac{1}{2}\mathfrak{B}^2(\mathbf{L}_\mathbf{A}{\mathbf{L}_\mathbf{A}}^T, \mathbf{B}) = \left(\mathbf{I} - \mathbf{T}^{\mathbf{A}\mathbf{B}}\right)\mathbf{L}_\mathbf{A}$$

Can we avoid projections?

- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **is not necessarily PSD.**
- **Classic workaround:** $\mathbf{A} = \Pi(\mathbf{L}_{\mathbf{A}}) \stackrel{\text{def}}{=} \mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}^\top$. **Effect on gradient methods?**

$$\nabla_{\mathbf{L}_{\mathbf{A}}} \frac{1}{2}\mathfrak{B}^2(\mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}{}^T, \mathbf{B}) = \left(\mathbf{I} - \mathbf{T}^{\mathbf{AB}}\right)\mathbf{L}_{\mathbf{A}}$$

- **Riemannian geodesics:** $\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]$



Can we avoid projections?

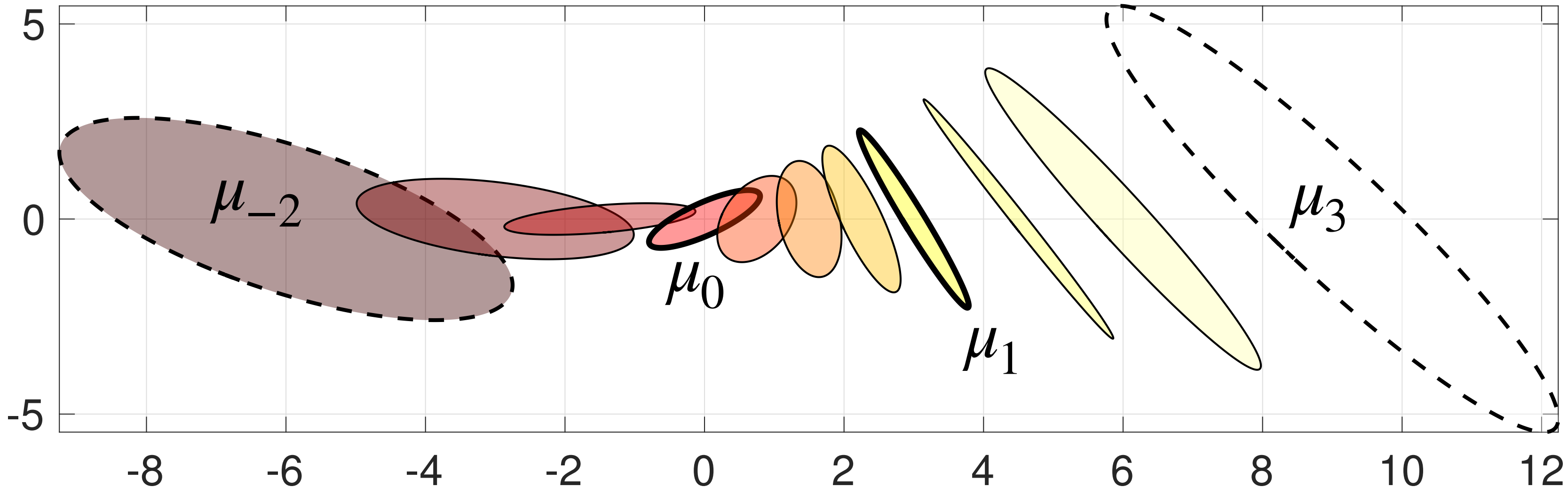
- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **is not necessarily PSD.**

- **Classic workaround:** $\mathbf{A} = \Pi(\mathbf{L}_\mathbf{A}) \overset{\text{def}}{=} \mathbf{L}_\mathbf{A}\mathbf{L}_\mathbf{A}^\top$. **Effect on gradient methods?**

$$\nabla_{\mathbf{L}_\mathbf{A}} \tfrac{1}{2}\mathfrak{B}^2(\mathbf{L}_\mathbf{A}\mathbf{L}_\mathbf{A}{}^T, \mathbf{B}) = \left(\mathbf{I} - \mathbf{T}^{\mathbf{AB}}\right)\mathbf{L}_\mathbf{A}$$

- **Riemannian geodesics:** $\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]$

W_2 geodesic $(\mu_t)_t$ from μ_0 to μ_1 ($t \in [0,1]$) and extrapolation



- **"$\Pi(\,\cdot\,)$ makes $\mathfrak{B}^2$ flat":** $\mathbf{L}_\mathbf{A} - t\nabla_{\mathbf{L}_\mathbf{A}}\frac{1}{2}\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \in \Pi^{-1}\{\mathbf{C}_{\mathbf{AB}}(t)\}$

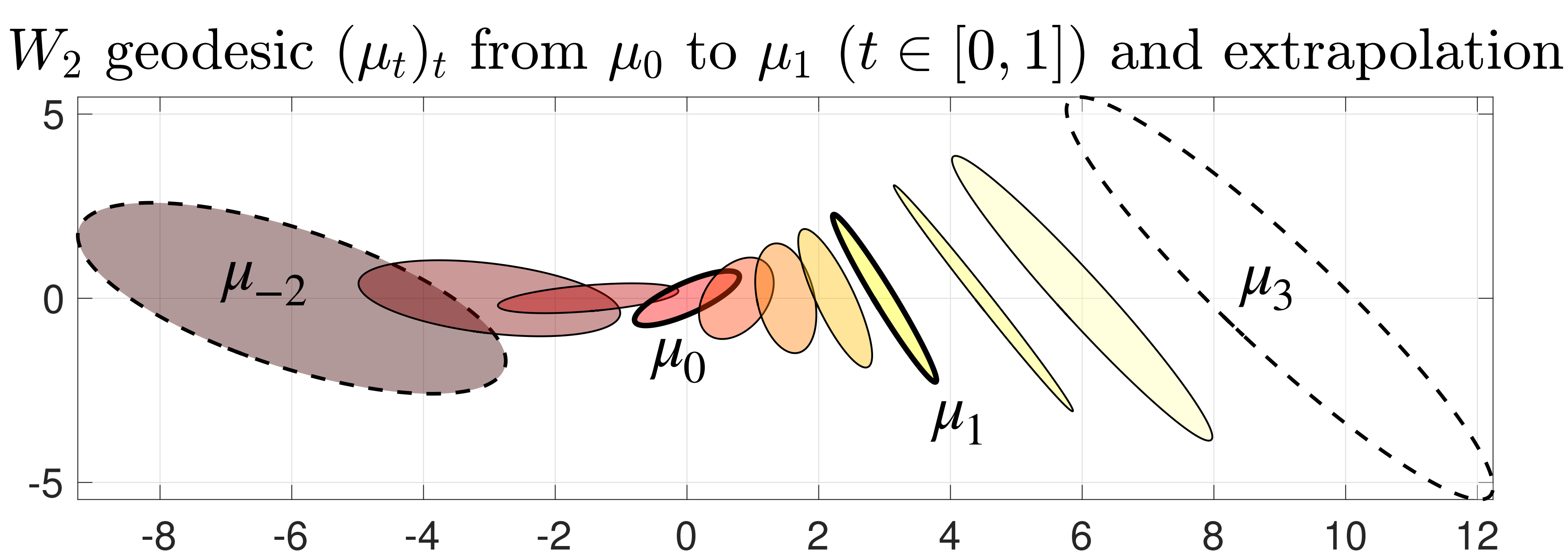
Can we avoid projections?

- $\mathbf{A} - t\nabla_{\mathbf{A}}\mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ **is not necessarily PSD.**

- **Classic workaround:** $\mathbf{A} = \Pi(\mathbf{L}_{\mathbf{A}}) \stackrel{\text{def}}{=} \mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}^{\top}$. **Effect on gradient methods?**

$$\nabla_{\mathbf{L}_{\mathbf{A}}} \frac{1}{2}\mathfrak{B}^2(\mathbf{L}_{\mathbf{A}}\mathbf{L}_{\mathbf{A}}{}^T, \mathbf{B}) = \left(\mathbf{I} - \mathbf{T}^{\mathbf{AB}}\right)\mathbf{L}_{\mathbf{A}}$$

- **Riemannian geodesics:** $\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]$

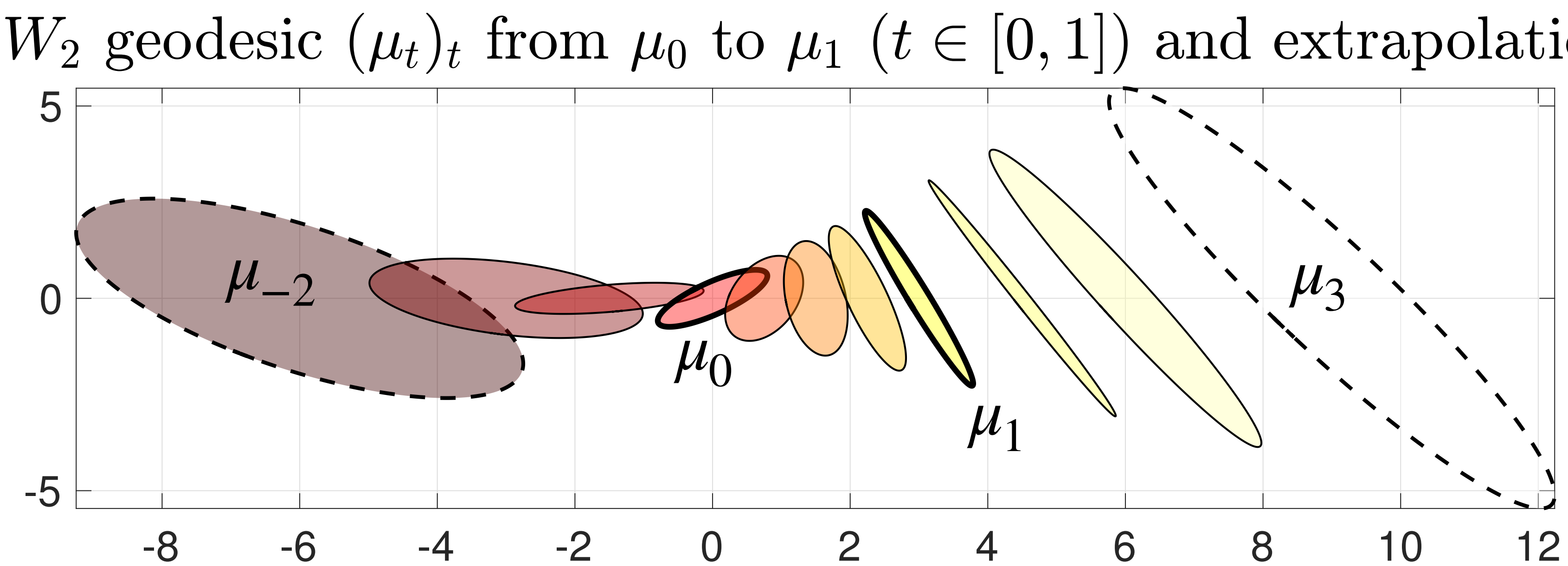


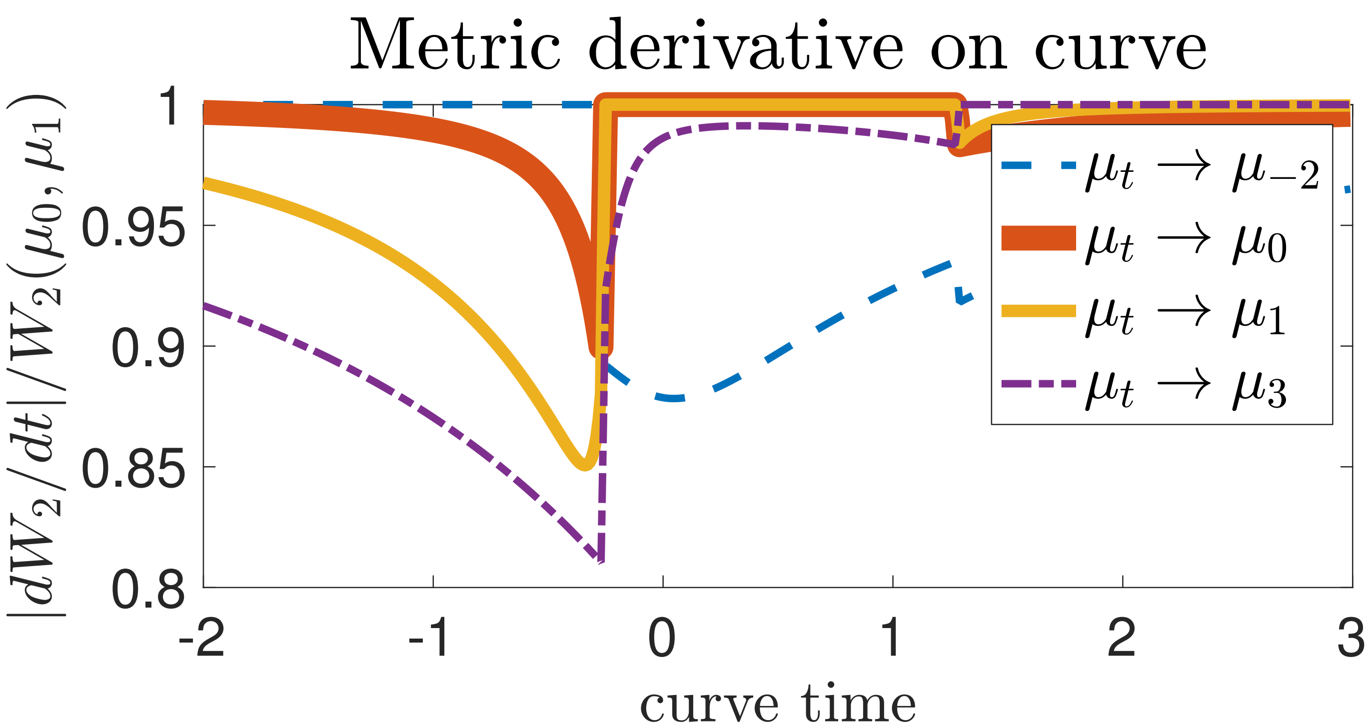
- **"$\Pi(\cdot)$ makes $\mathfrak{B}^2$ flat":** $\mathbf{L}_{\mathbf{A}} - t\nabla_{\mathbf{L}_{\mathbf{A}}}\frac{1}{2}\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) \in \Pi^{-1}\{\mathbf{C}_{\mathbf{AB}}(t)\}$

[BM&Cuturi'18]

Extrapolation

- **Riemannian geodesics:** $\mathbf{C}_{\mathbf{AB}}(t) = [(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]\mathbf{A}[(1-t)\mathbf{I} - t\mathbf{T}^{\mathbf{AB}}]$





IV. Applications

Elliptical Word Embeddings

- **[BM&Cuturi'18]**

« 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	

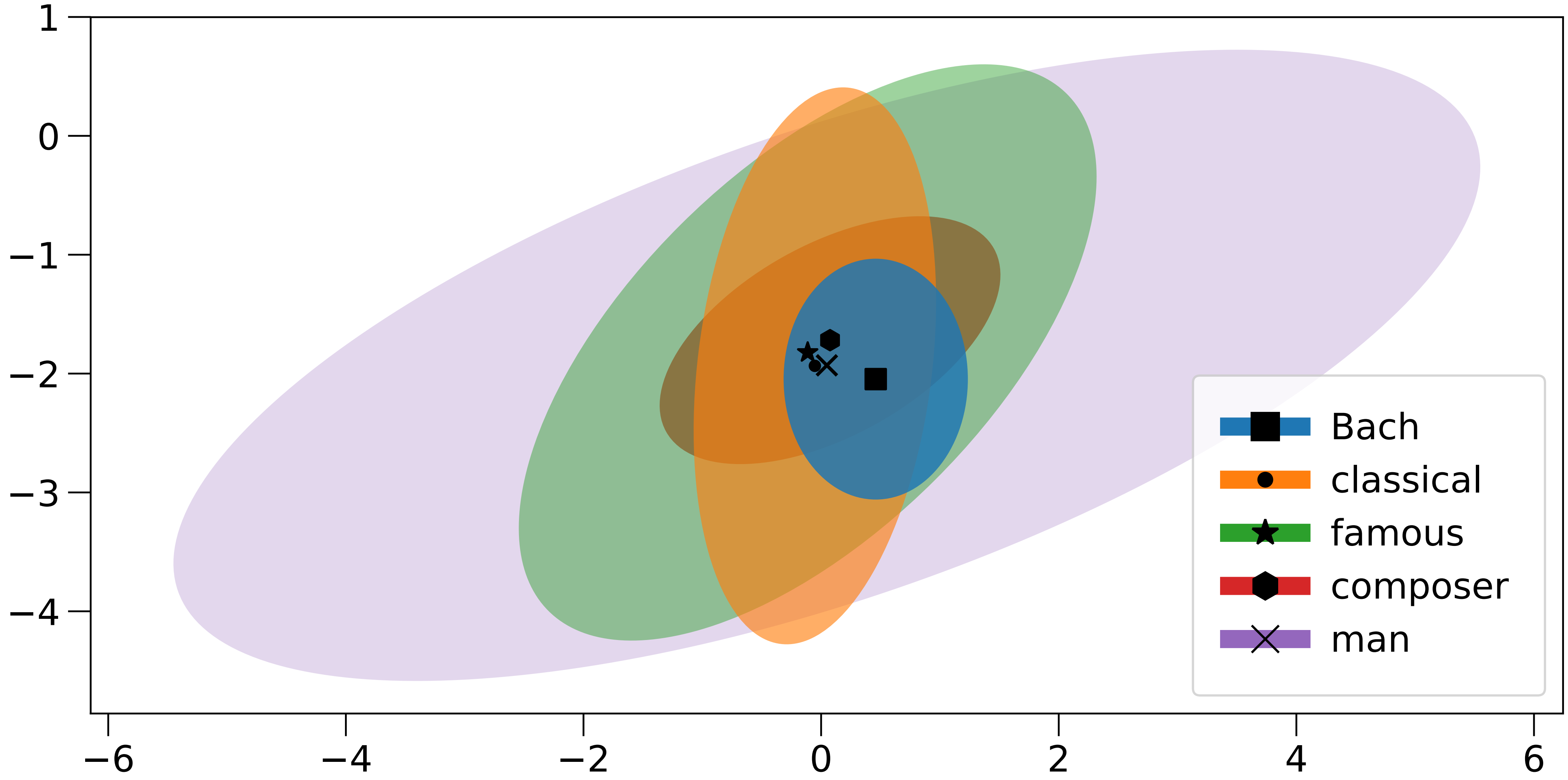
- **Optimize**

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

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**

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

Word Embeddings: visualization



Word Embeddings: Similarity Evaluation

Dataset	W2G/45/C	Ell/12/BC
SimLex	**33.28**	24.09
WordSim	62.52	**66.02**
WordSim-R	69.37	**71.07**
WordSim-S	57.56	**60.58**
MEN	61.5	**65.58**
MC	**79.5**	65.95
RG	**67.6**1	65.58
YP	20.86	**25.14**
MT-287	**61.71**	59.53
MT-771	**58.11**	56.78
RW	**30.62**	29.04

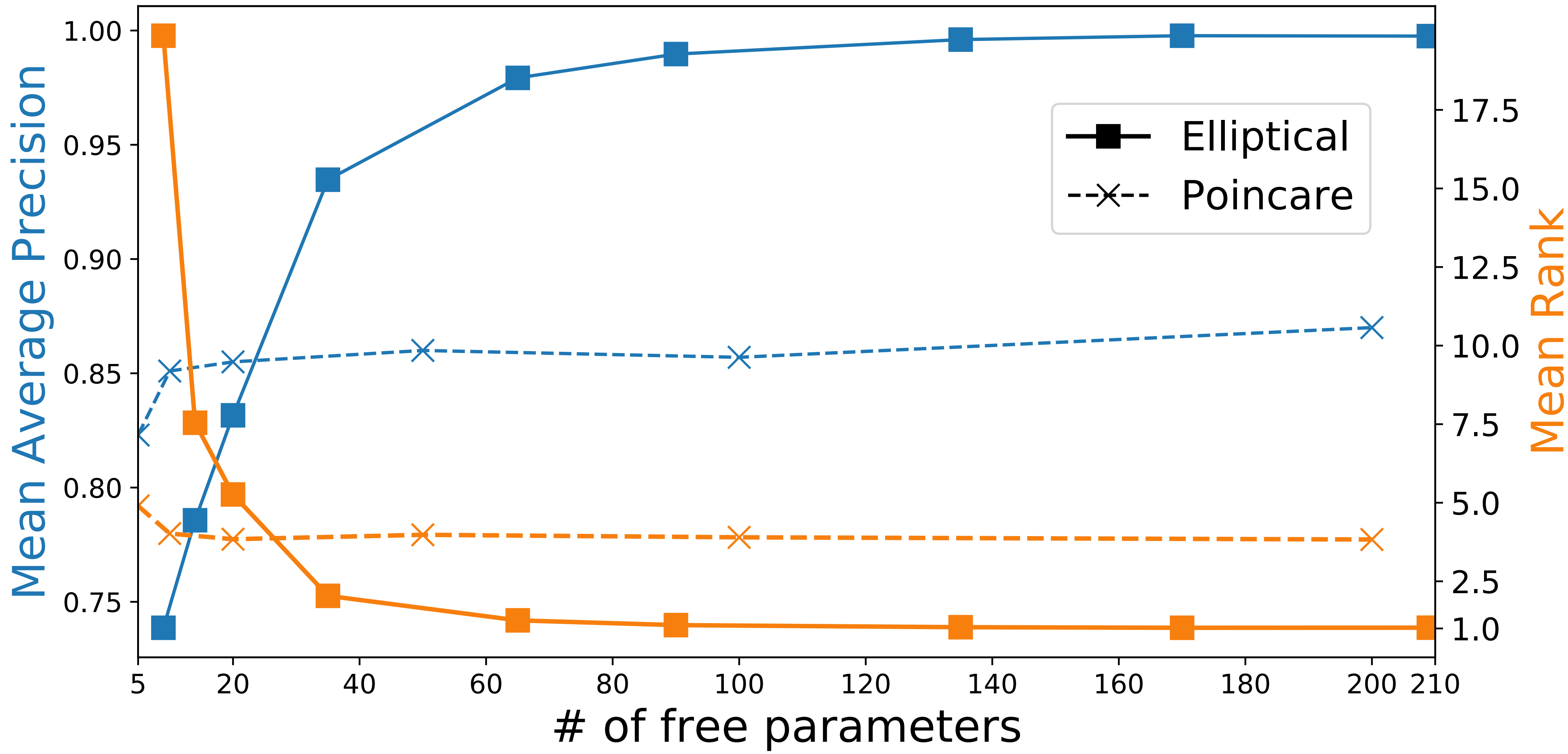
Spearman rank correlation with human scores

Comparison with [Vilnis & McCallum'15]

Hypernymy embeddings

A is a *hypernym* of B if every B is an A

- **Ex: ‘*mammal*’ > ‘*dog*’**
- **WordNet Dataset: 743,251 relations, 82,115 distinct nouns**



Comparison with [Nickel & Kiela’17]

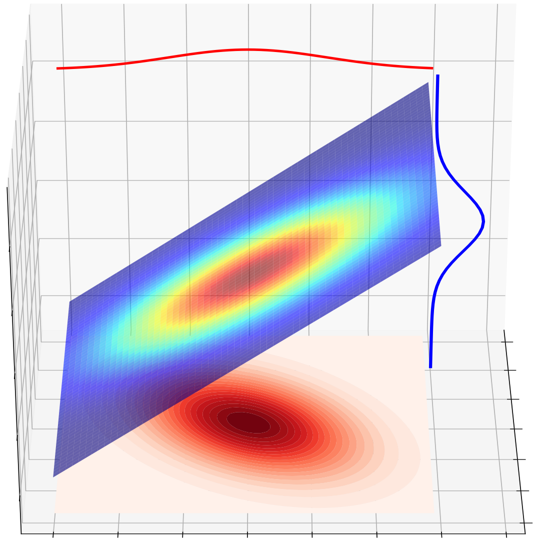
Other applications

- **Robust (min/max) estimation of inverse covariance matrices [Nguyen et al.'18]**

- **Distributionally robust Kalhman filtering [Abadeh et al.'18]**

- **GANs: Fréchet Inception Distance (FID) [Heusel et al.'17]**

- **Extension to the subspace constraints: [BM&Cuturi'19]**



Extensions

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.**

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

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

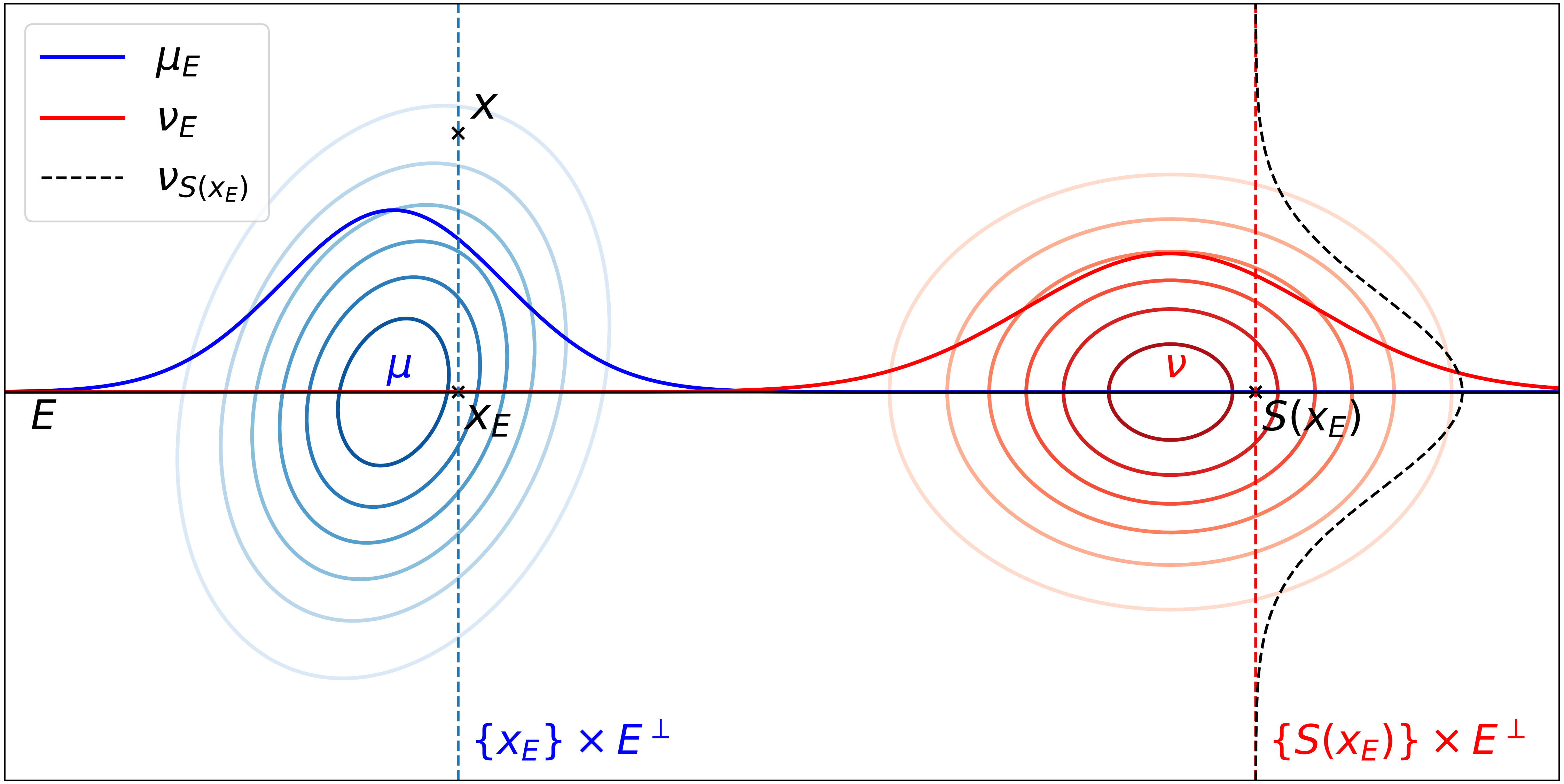
$$\forall f \in C_b(\mathbb{R}^d), \int f d\mu = \int \left(\int f(x_E, x_{E^\perp}) d\mu_{x_E}(x_{E^\perp}) \right) d\mu_E(x_E)$$

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

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

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

$\implies$ 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 $\mu_{x_E} \otimes \nu_{S(x_E)}$

Monge-Independent Transport

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

Def. Monge-Independent (MI) transport plan:

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

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

Monge-Independent Transport

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

Def. Monge-Independent (MI) transport plan:

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

where $\mu_E \overset{def}{=} (p_E)_\sharp(\mu)$, $\nu_E \overset{def}{=} (p_E)_\sharp(\nu)$, 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 $\pi_n \rightharpoonup \pi_{\textbf{MI}}$

Monge-Independent Transport

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

Def. Monge-Independent (MI) transport plan:

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

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

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** $\nu_{S(E)}$

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:

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

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:

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

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

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

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

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**

OT for Gaussian Distributions

[Gelbrich'90]

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

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

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

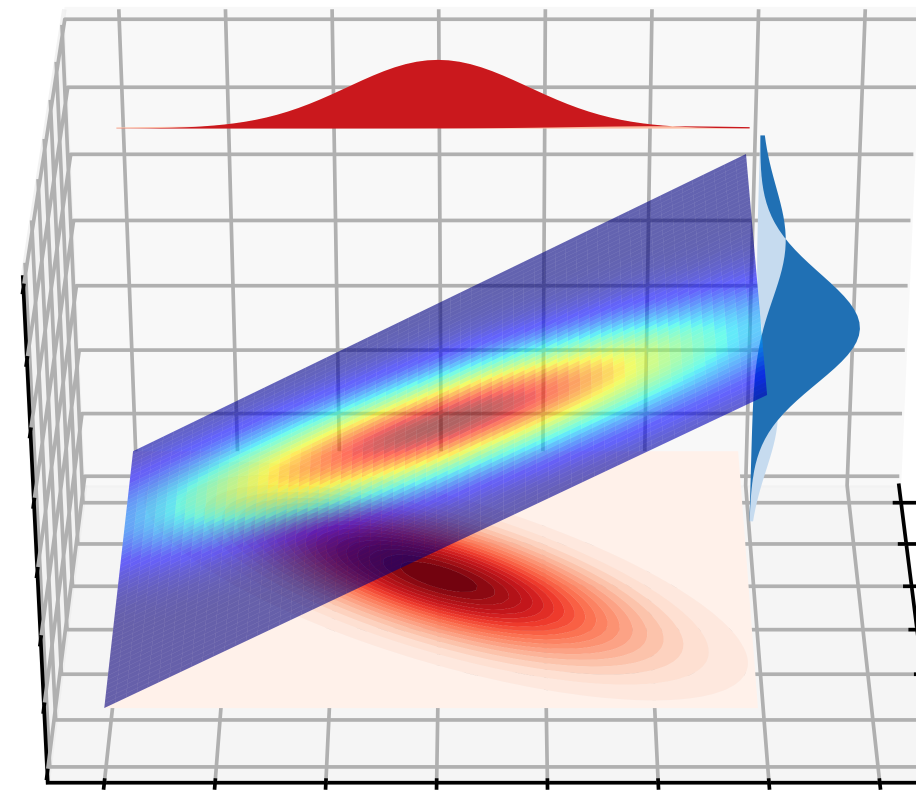
Prop. If $\alpha, \beta \in P(\mathbb{R}^d)$ **are elliptical distributions with** $\mathrm{var}\alpha = \mathbf{A}, \mathrm{var}\beta = \mathbf{B}$**, then**

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

where $\mathbf{T}^{\mathbf{AB}} \stackrel{\text{def}}{=} \mathbf{A}^{-\frac{1}{2}}(\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}})^{\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}}$ **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 $\Sigma \stackrel{def}{=} \begin{pmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{B} \end{pmatrix}$

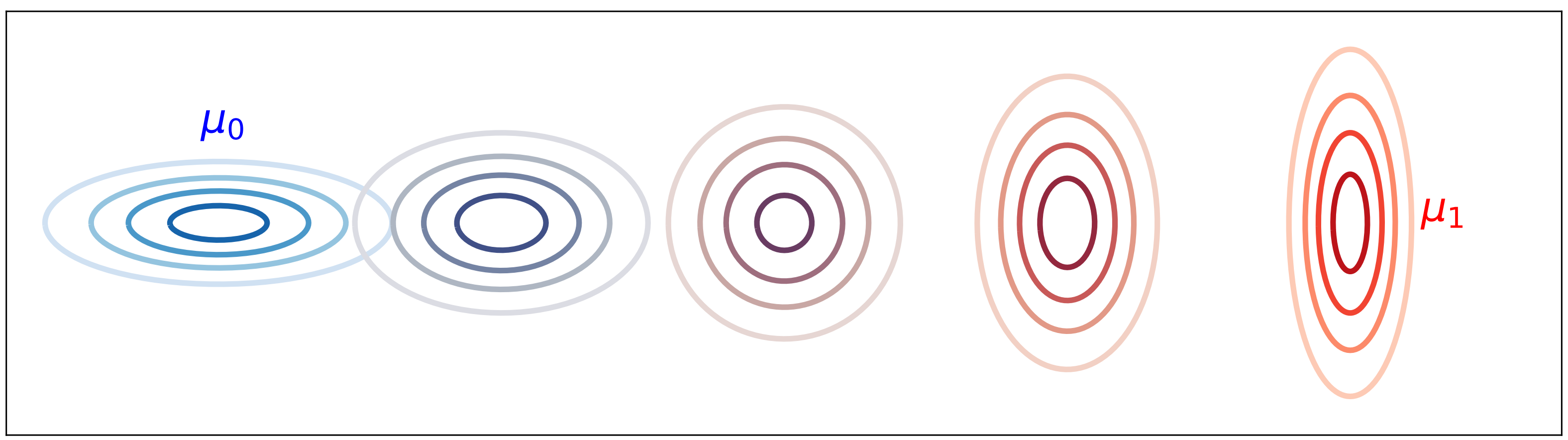
Then $\pi_{MK}(\mu, \nu) = \mathcal{N}(0_{2d}, \Sigma) \in \mathscr{P}(\mathbb{R}^d \times \mathbb{R}^d)$

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

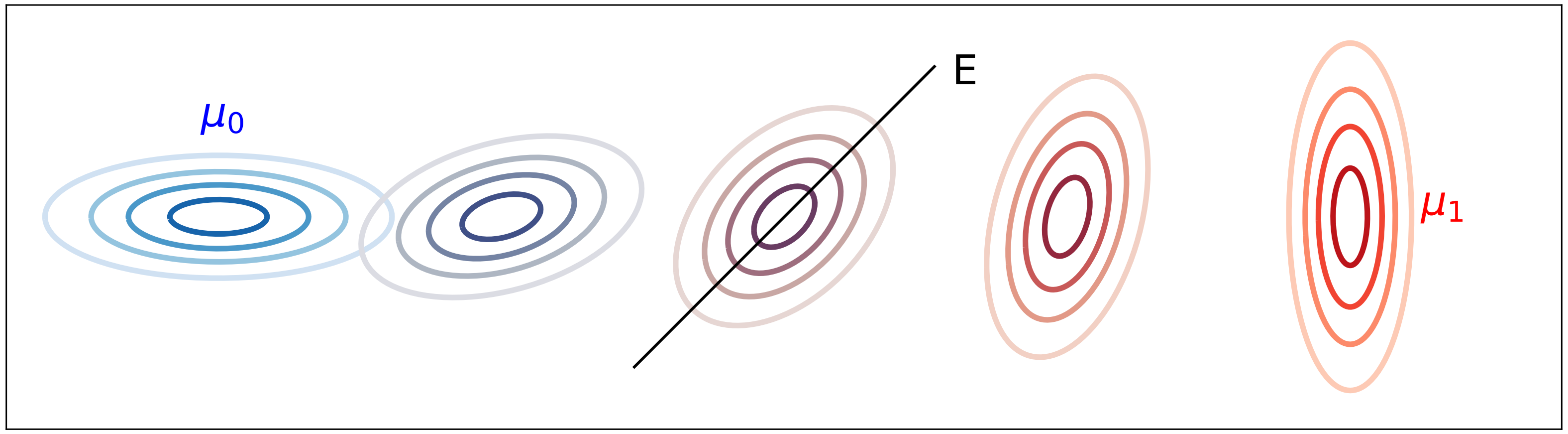
Monge-Knothe: Gaussian Distributions

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



Monge interpolation



MK interpolation

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

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*

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