D transforms

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi}\in L_1(\boldsymbol{\mu}),\boldsymbol{\psi}\in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x)+\boldsymbol{\psi}(y)\leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Imagine we choose a $\boldsymbol{\varphi}$. Can we find a good $\boldsymbol{\psi}$?

D transforms

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Imagine we choose a $\boldsymbol{\varphi}$. Can we find a good $\boldsymbol{\psi}$?

We need that $\boldsymbol{\psi}$ satisfies for *all* $\boldsymbol{x}, \boldsymbol{y}$

$$\boldsymbol{\varphi}(\boldsymbol{x}) + \boldsymbol{\psi}(\boldsymbol{y}) \leq \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y})$$

D transforms

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Imagine we choose a $\boldsymbol{\varphi}$. Can we find a good $\boldsymbol{\psi}$?

We need that $\boldsymbol{\psi}$ satisfies for *all* $\boldsymbol{x}, \boldsymbol{y}$

$$\boldsymbol{\varphi}(\boldsymbol{x}) + \boldsymbol{\psi}(\boldsymbol{y}) \leq \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y})$$

$$\boldsymbol{\psi}(\boldsymbol{y}) \leq \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x})$$

D transforms

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi} \in L_1(\boldsymbol{\mu}), \boldsymbol{\psi} \in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x) + \boldsymbol{\psi}(y) \leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

Imagine we choose a $\boldsymbol{\varphi}$. Can we find a good $\boldsymbol{\psi}$?

We need that $\boldsymbol{\psi}$ satisfies for *all* $\boldsymbol{x}, \boldsymbol{y}$

$$\boldsymbol{\varphi}(\boldsymbol{x}) + \boldsymbol{\psi}(\boldsymbol{y}) \leq \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y})$$

$$\boldsymbol{\psi}(\boldsymbol{y}) \leq \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x})$$

$$\boldsymbol{\psi}(\boldsymbol{y}) \leq \inf_{\boldsymbol{x}} \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x})$$

D transforms

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi}\in L_1(\boldsymbol{\mu}),\boldsymbol{\psi}\in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x)+\boldsymbol{\psi}(y)\leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

For given $\boldsymbol{\varphi}$, cannot get a better $\boldsymbol{\psi}$ than

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{D}^p(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

D transforms

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\substack{\boldsymbol{\varphi}\in L_1(\boldsymbol{\mu}),\boldsymbol{\psi}\in L_1(\boldsymbol{\nu}) \\ \boldsymbol{\varphi}(x)+\boldsymbol{\psi}(y)\leq \boldsymbol{D}^p(x,y)}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \boldsymbol{\psi} d\boldsymbol{\nu}.$$

DUAL

For given $\boldsymbol{\varphi}$, cannot get a better $\boldsymbol{\psi}$ than

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{D}^p(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}} \int \boldsymbol{\varphi} d\boldsymbol{\mu} + \int \overline{\boldsymbol{\varphi}} d\boldsymbol{\nu}.$$

SEMI-DUAL

D transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y D^p(x,y) - \psi(y).$$

D transforms

$$\overline{\boldsymbol{\varphi}}(\boldsymbol{y}) \stackrel{\text{def}}{=} \inf_{\boldsymbol{x}} \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\varphi}(\boldsymbol{x}).$$

$$\overline{\boldsymbol{\psi}}(\boldsymbol{x}) = \inf_{\boldsymbol{y}} \boldsymbol{D}^p(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\psi}(\boldsymbol{y}).$$

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \sup_{\boldsymbol{\varphi}} \int \overline{\overline{\boldsymbol{\varphi}}} d\boldsymbol{\mu} + \int \overline{\boldsymbol{\varphi}} d\boldsymbol{\nu}.$$

D transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y D^p(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

D transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y D^p(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

For all φ, we have $\overline{\overline{\overline{\varphi}}} = \overline{\varphi}$

D transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y D^p(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

For all φ, we have $\overline{\overline{\overline{\varphi}}} = \overline{\varphi}$

φ is D^p-concave if $\exists \phi : \varphi = \overline{\phi}$

φ is D^p-concave $\Rightarrow \overline{\overline{\varphi}} = \varphi$

D transforms

$$\overline{\varphi}(y) \stackrel{\text{def}}{=} \inf_x D^p(x,y) - \varphi(x).$$

$$\overline{\psi}(x) = \inf_y D^p(x,y) - \psi(y).$$

$$W_p^p(\mu,\nu) = \sup_\varphi \int \overline{\overline{\varphi}} d\mu + \int \overline{\varphi} d\nu.$$

$$W_p^p(\mu,\nu) = \sup_{\varphi \text{ is } D^p\text{-concave}} \int \varphi d\mu + \int \overline{\varphi} d\nu.$$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$$\overline{\varphi}_x(y) - \overline{\varphi}_x(y') = D(x,y) - D(x,y') \le D(y,y')$$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x, y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \leq D(x,y)$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.
$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$
$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y) \le -\overline{\varphi}(x)$

D transforms, W_1

Prop. If $c = D$, namely $p = 1$, then
φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

For given x, $\overline{\varphi}_x(y) \stackrel{\text{def}}{=} D(x,y) - \varphi(x)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) = \inf_x \overline{\varphi}_x(y)$ is 1-Lipschitz.

$\Rightarrow \overline{\varphi}(y) - \overline{\varphi}(x) \le D(x,y)$

$\Rightarrow -\overline{\varphi}(x) \le D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y)$

$\Rightarrow -\overline{\varphi}(x) \le \inf_y D(x,y) - \overline{\varphi}(y) \le -\overline{\varphi}(x)$

$\Rightarrow -\overline{\varphi}(x) \le \overline{\overline{\varphi}}(x) \le -\overline{\varphi}(x)$ and $\overline{\varphi}(x) = -\varphi(x)$

D transforms, W_1

$$W_1(\mu,\nu) = \sup_{\varphi \text{ is } D\text{-concave}} \int \varphi d\mu + \int \overline{\varphi} d\nu.$$

SEMI-DUAL

Prop. If $c = D$, then

φ is D-concave $\Leftrightarrow \overline{\varphi} = -\varphi$, φ is 1-Lipschitz

$$W_1(\mu,\nu) = \sup_{\varphi \text{ 1-Lipschitz}} \int \varphi(d\mu - d\nu).$$

W1

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

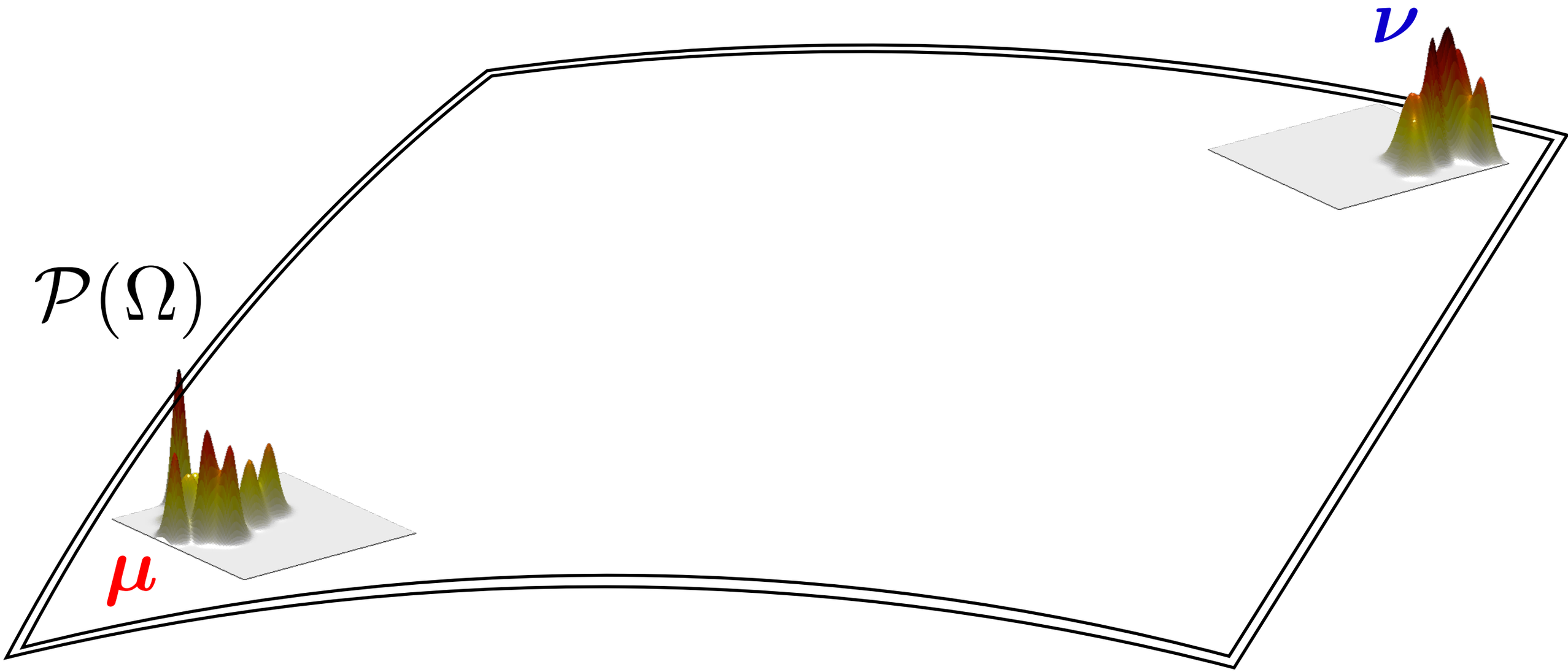
$$\boldsymbol{P}^\star := (\mathrm{Id}, T^\star)_\sharp \boldsymbol{\mu} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$$

is also *optimal* for the Kantorovich problem.

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

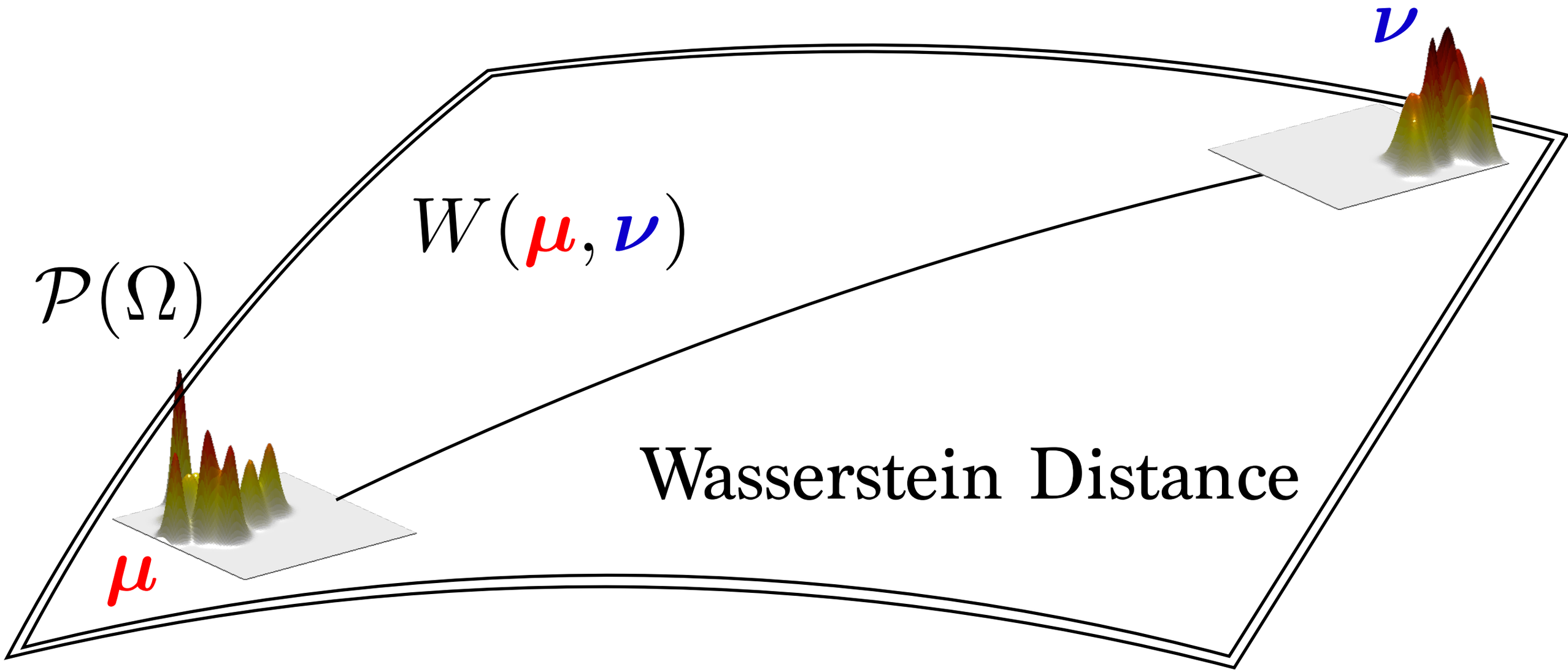
Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



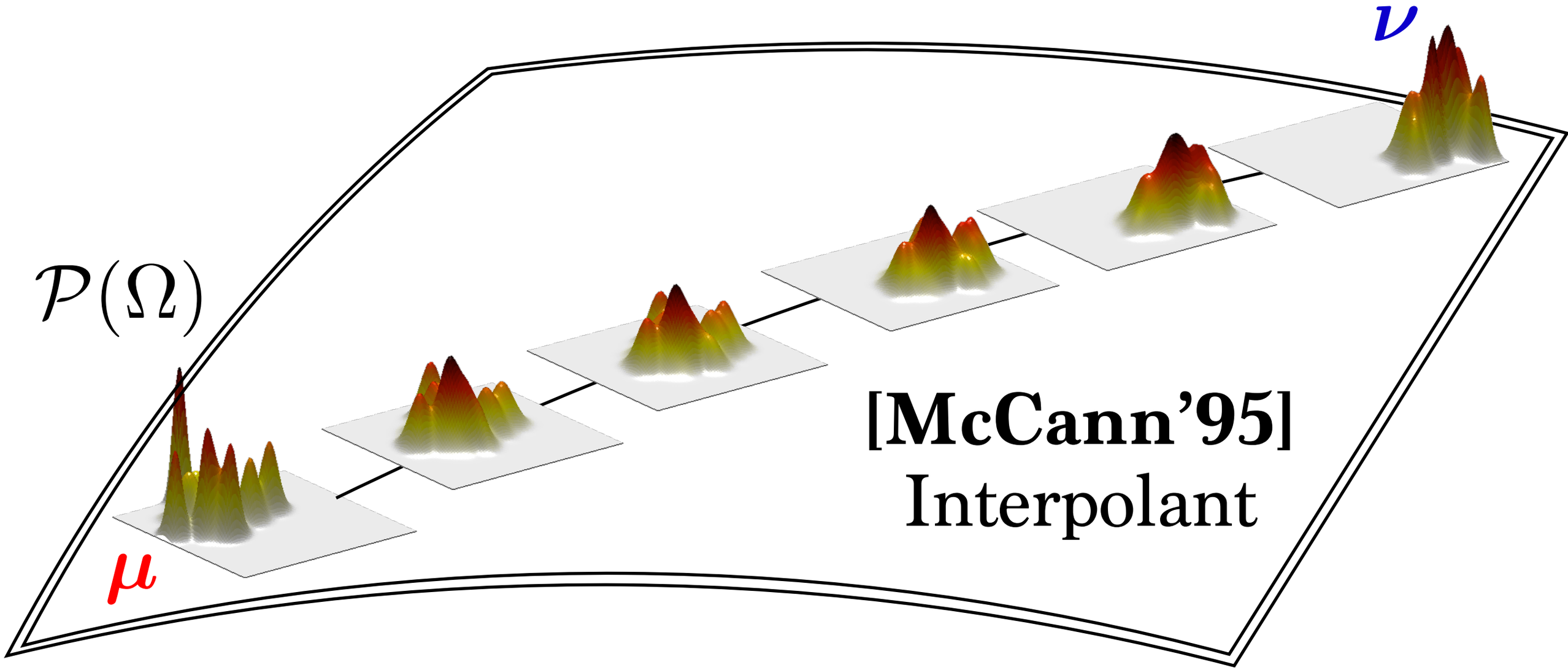
Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



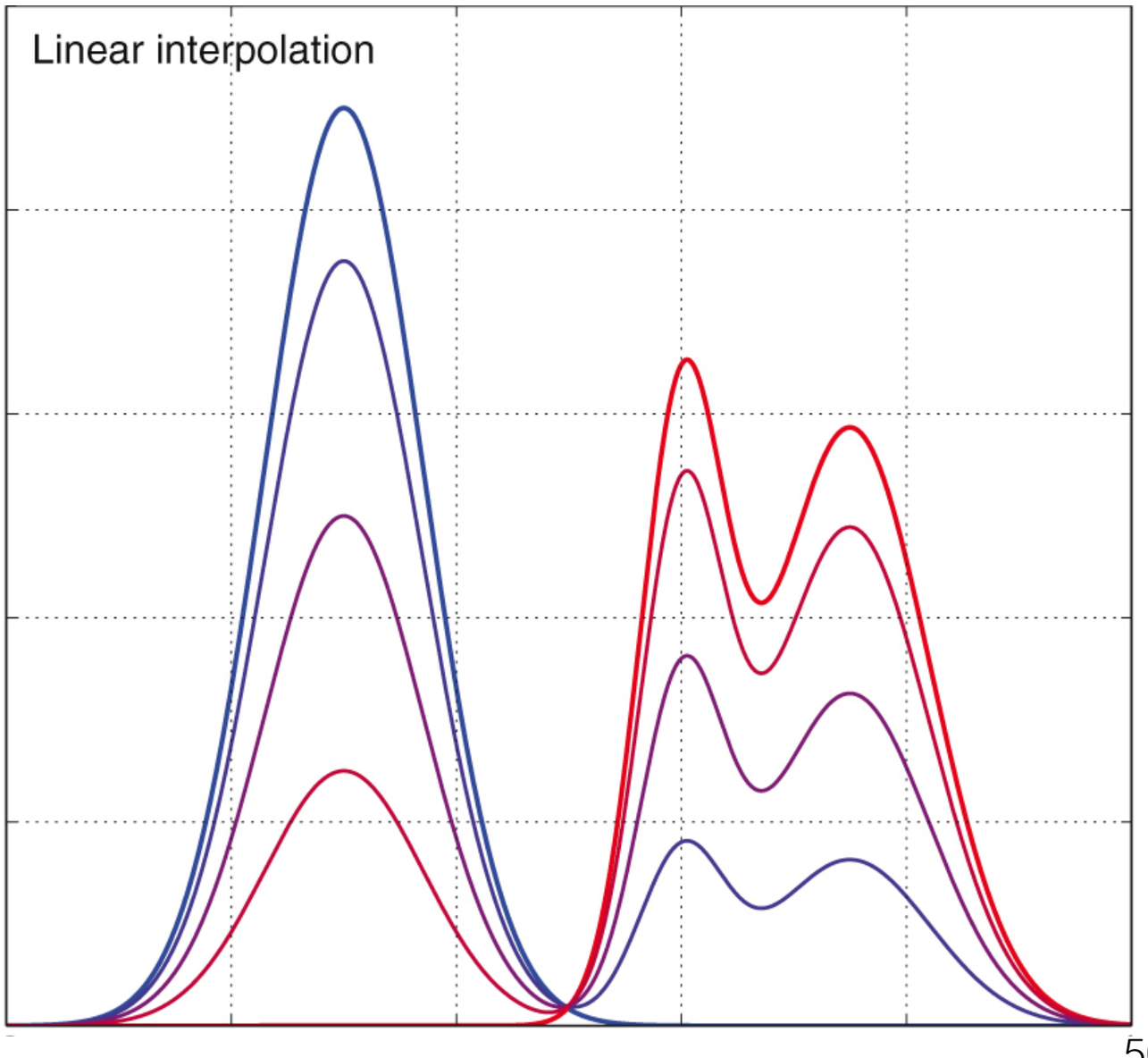
Optimal Transport Geometry

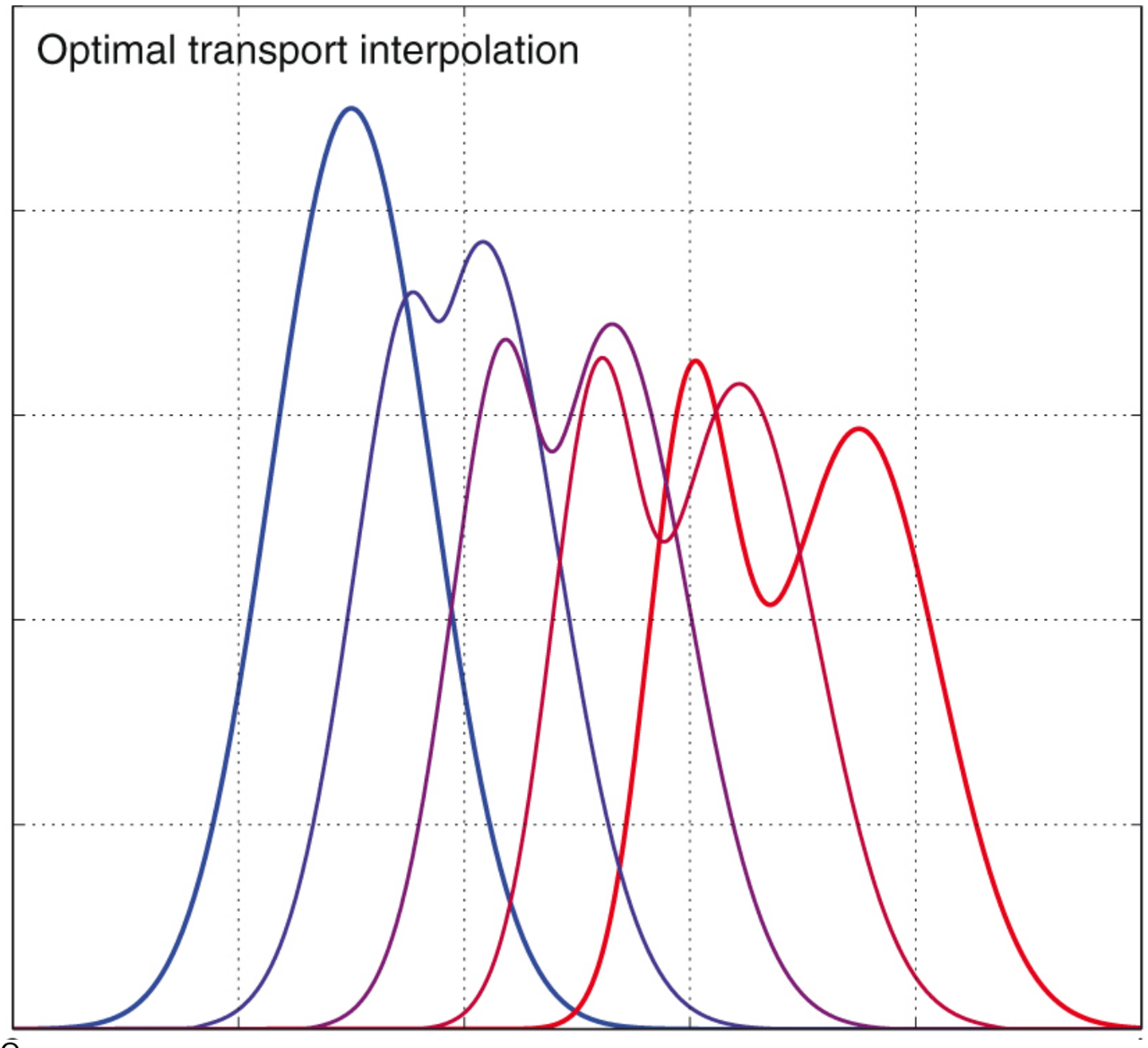
Very different geometry than standard information divergences (*KL*, Euclidean)



Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)





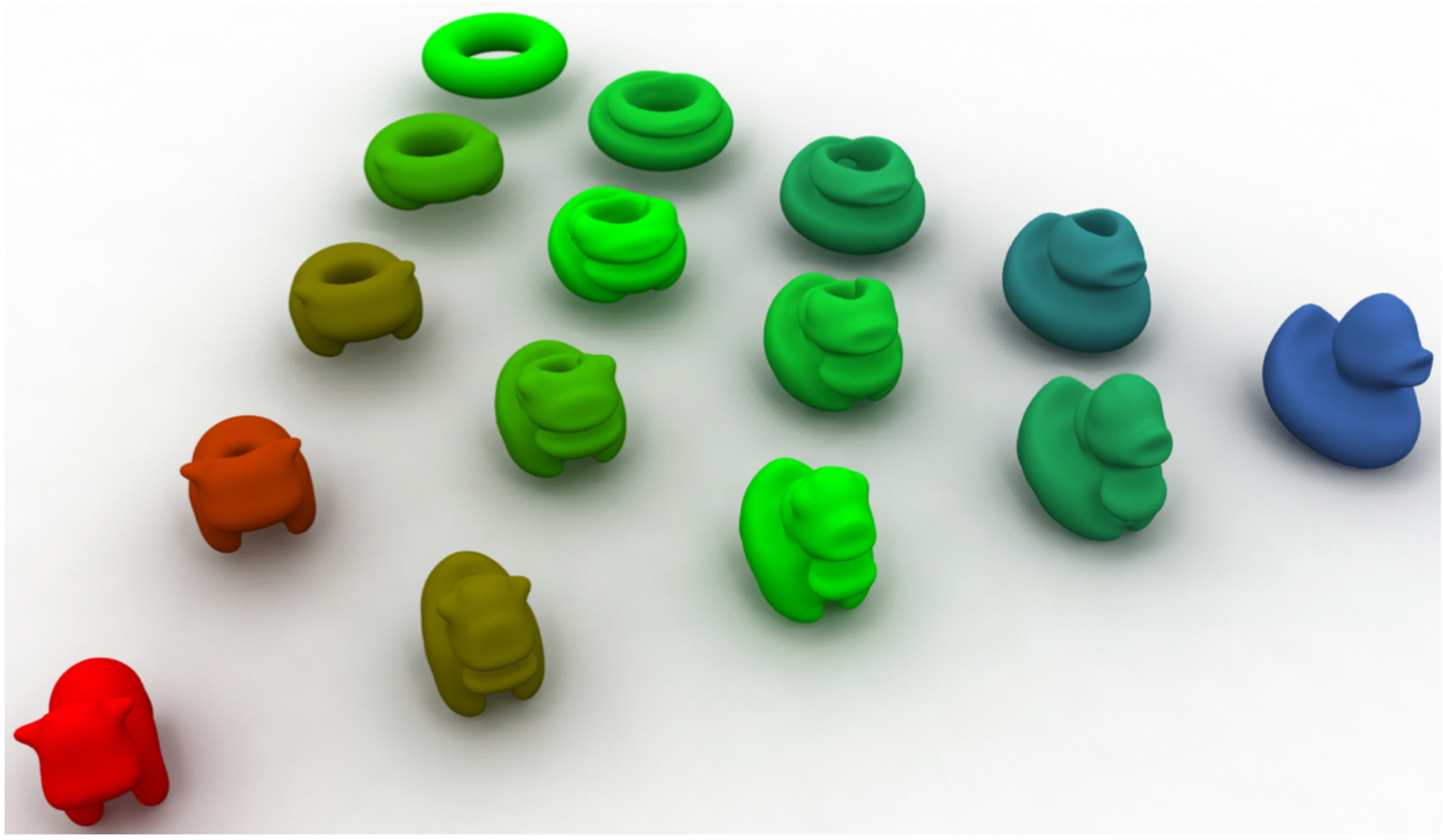
Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)

[SDPC..'15]

Optimal Transport Geometry

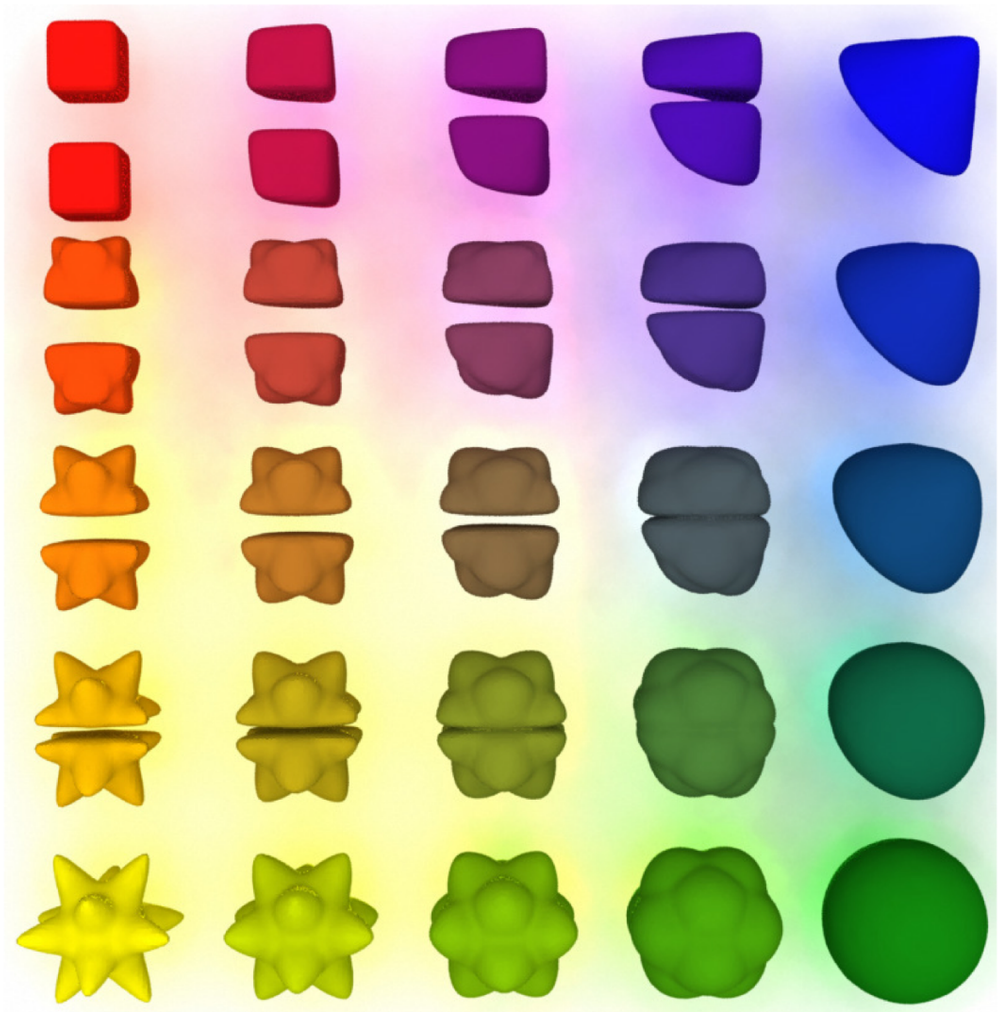
Very different geometry than standard information divergences (*KL*, Euclidean)



[SDPC..'15]

Optimal Transport Geometry

Very different geometry than standard information divergences (*KL*, Euclidean)



[SDPC..'15]

Variational OT Problems in ML

Up to 2010: OT solvers
used mostly for retrieval
in databases of histograms

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = ?$$

$$W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \cdots ?$$

OT is now used as a **loss** or **fidelity** term:

$$\operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{P}(\Omega)} F(W_p(\boldsymbol{\mu}, \boldsymbol{\nu_1}), W_p(\boldsymbol{\mu}, \boldsymbol{\nu_2}), \ldots, \boldsymbol{\mu}) = ?$$

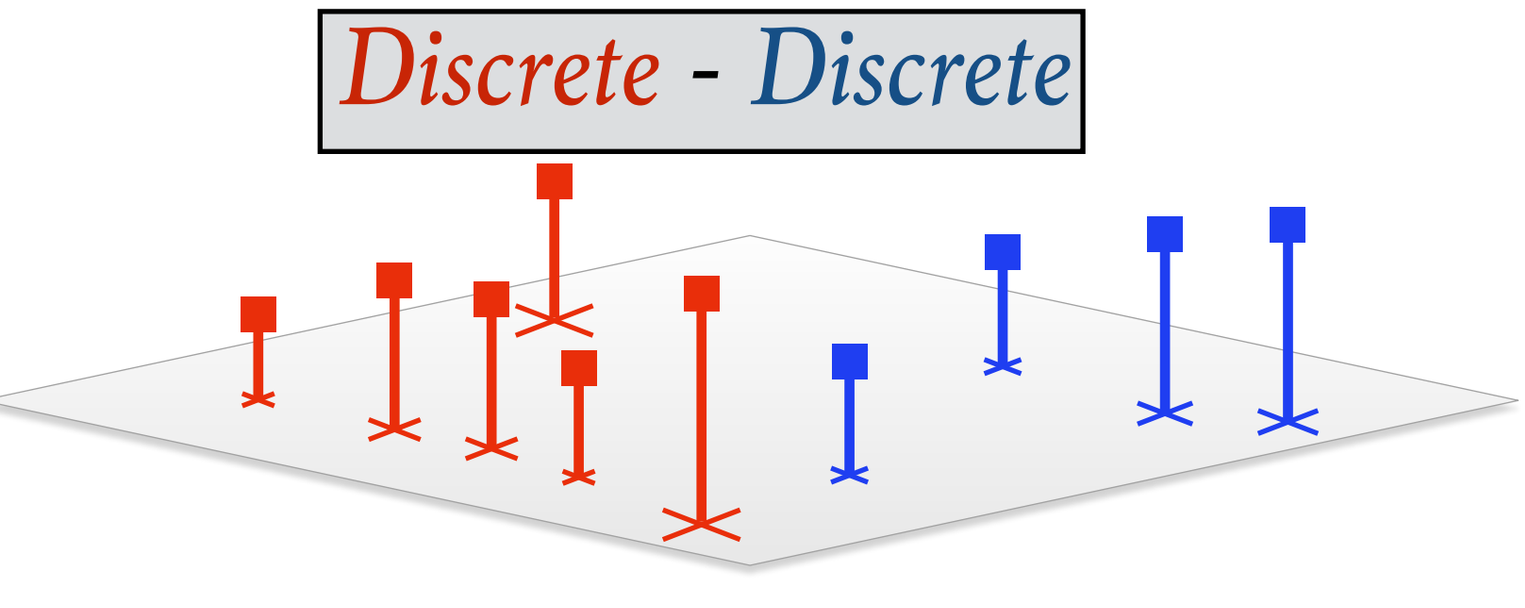
$$\text{“}\nabla_{\boldsymbol{\mu}}\text{”} W_p(\boldsymbol{\mu}, \boldsymbol{\nu}) = ?$$

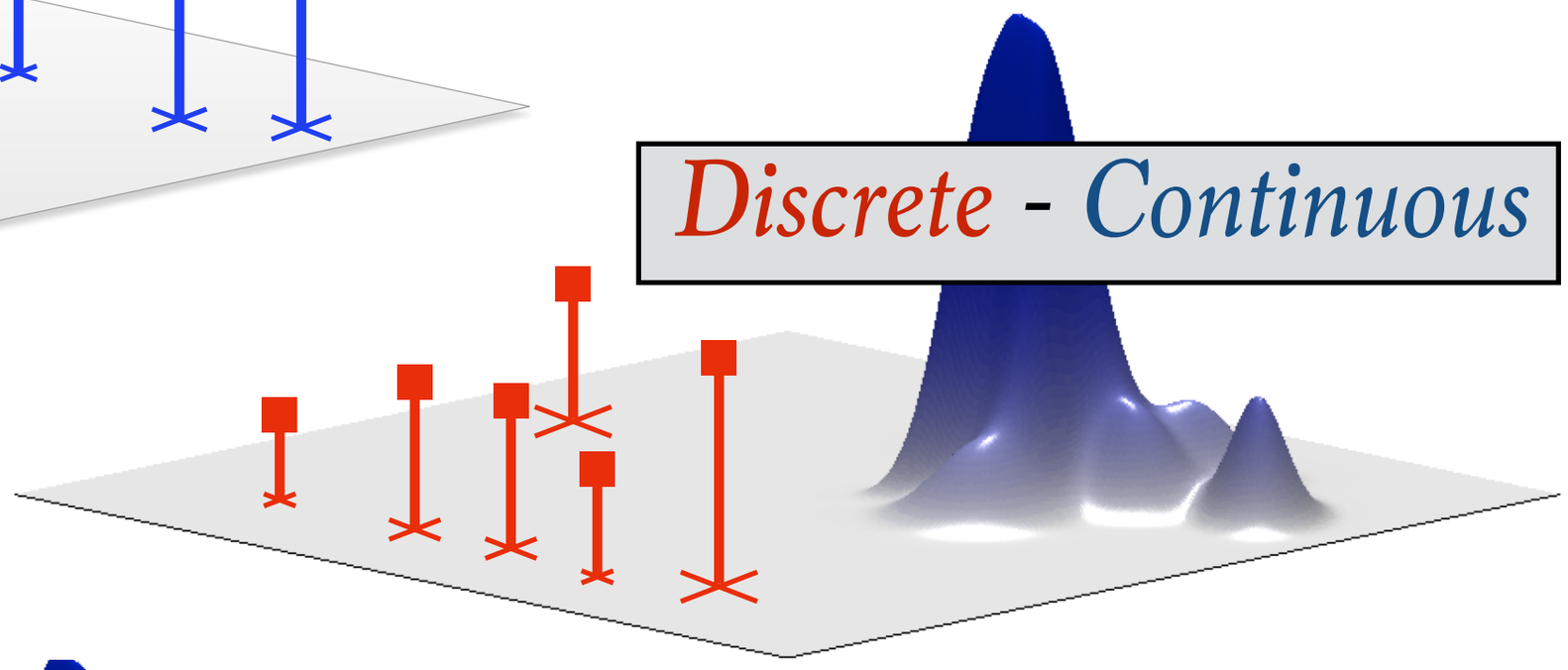
[JKO’98]
[AGS’06]

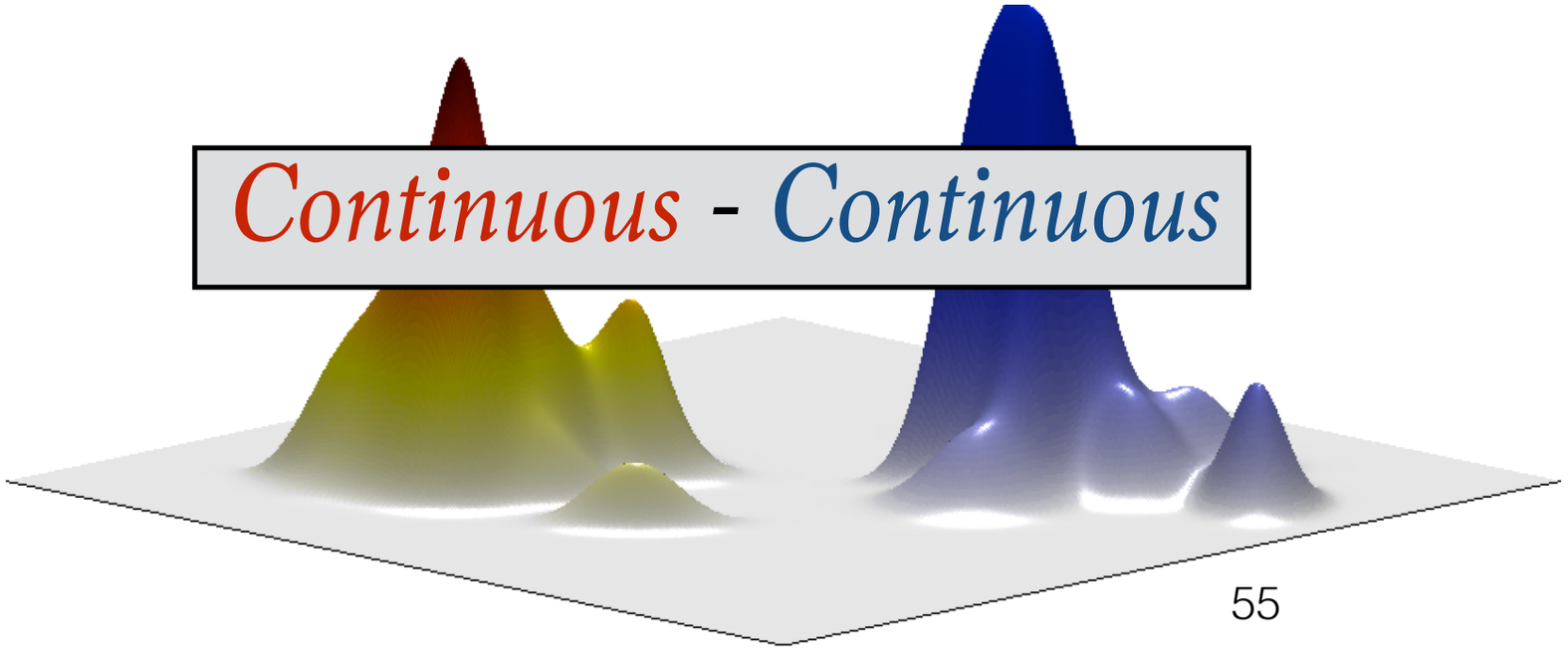
2. Computing OT exactly

- Typology: discrete/continuous problems
- Easy cases and exact solvers for discrete measures.

When can we compute OT?

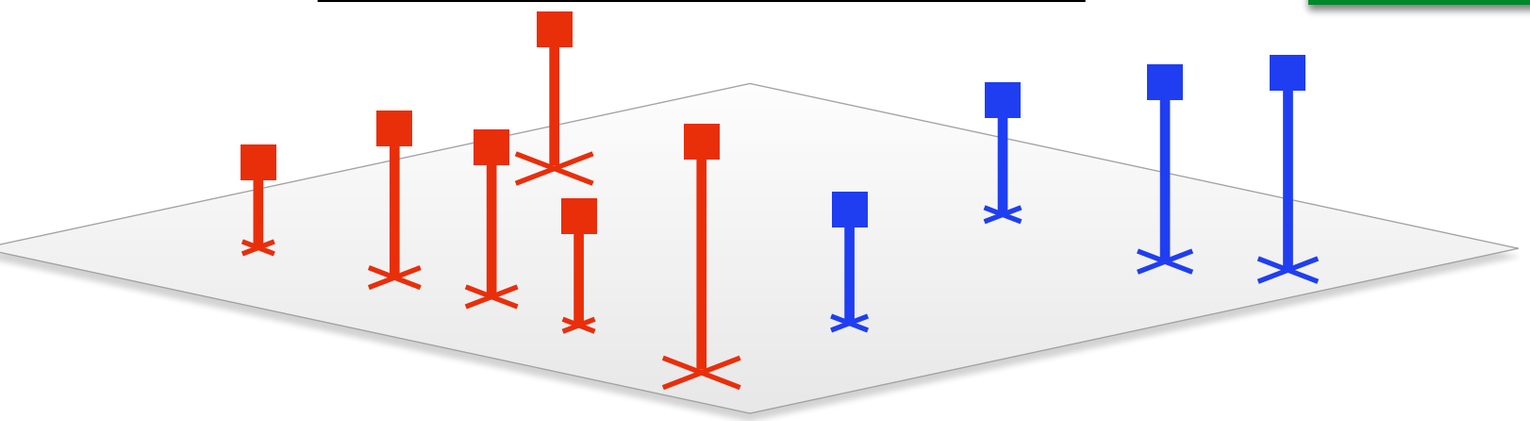






When can we compute OT?

Discrete - *Discrete*

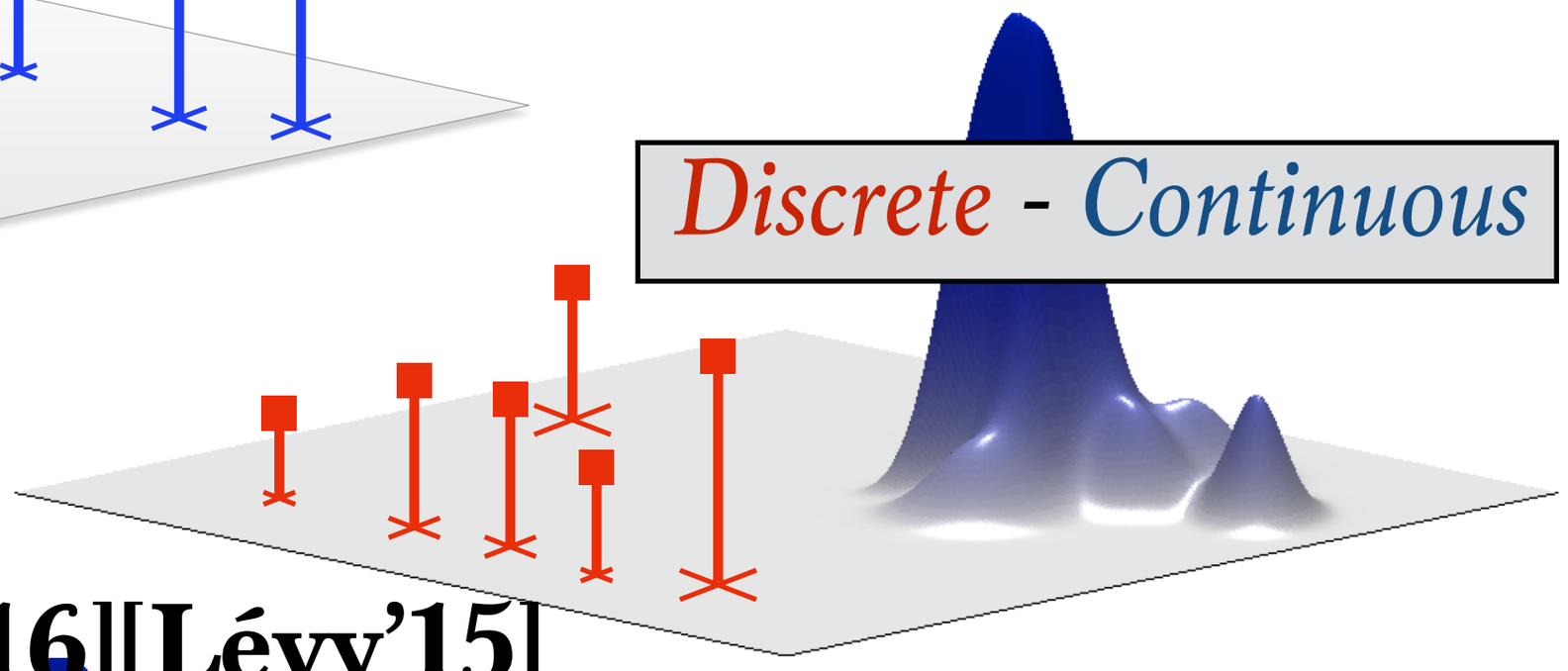


Network flow solvers

low dim.

[Mérigot'11][KMT'16][Lévy'15]

Discrete - *Continuous*

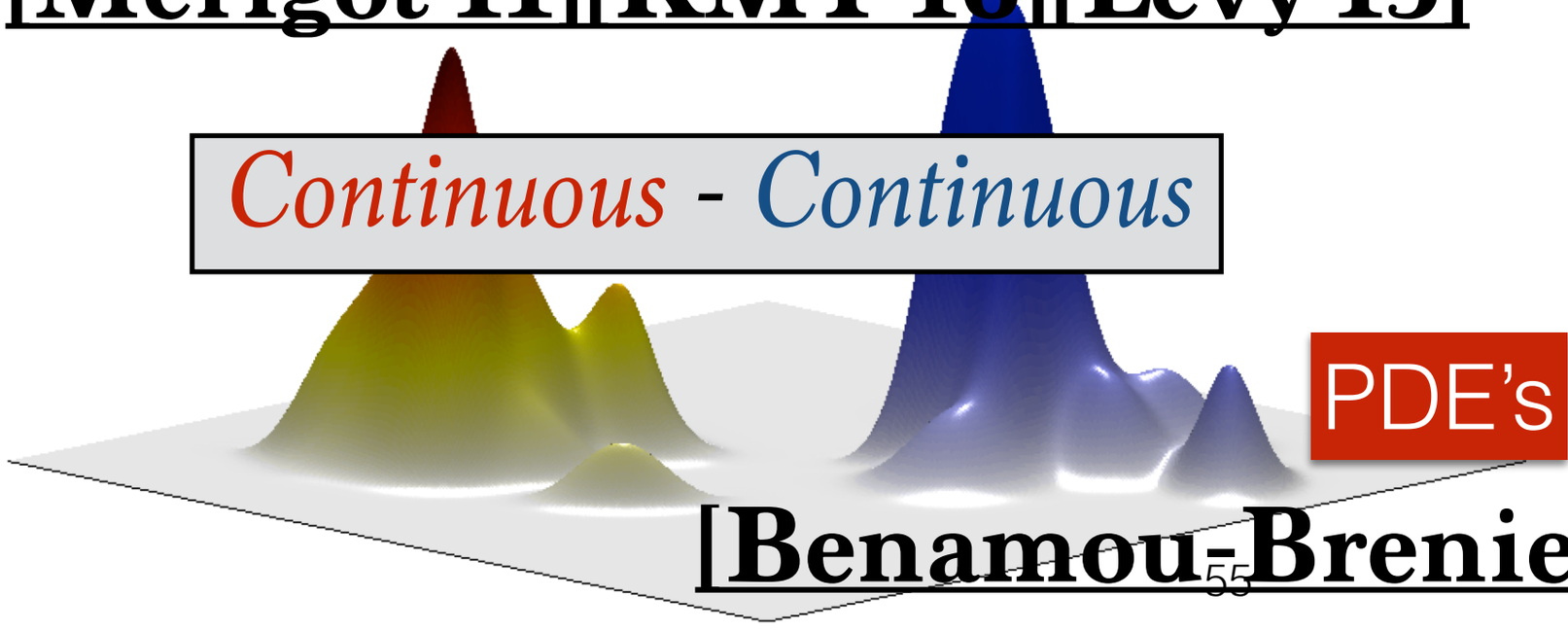


Stochastic Optimization

[GCPB'16]

[ACB'17]

Continuous - *Continuous*



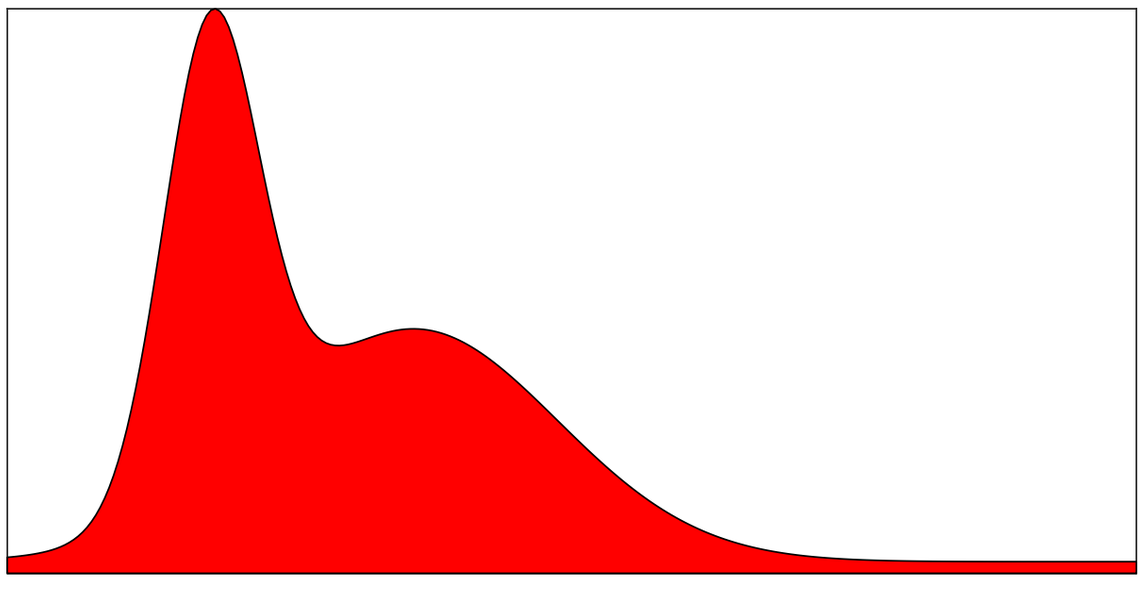
PDE's

[Benamou-Brenier'98]

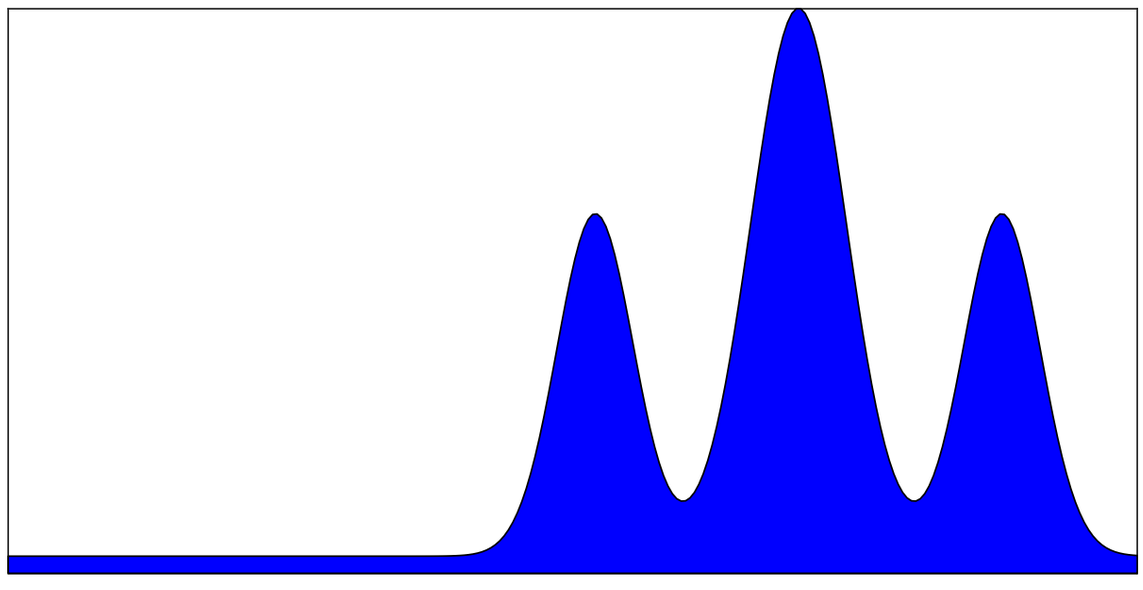
Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x,y) = \boldsymbol{c}(|x-y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$



0 $\boldsymbol{\mu}$ 10

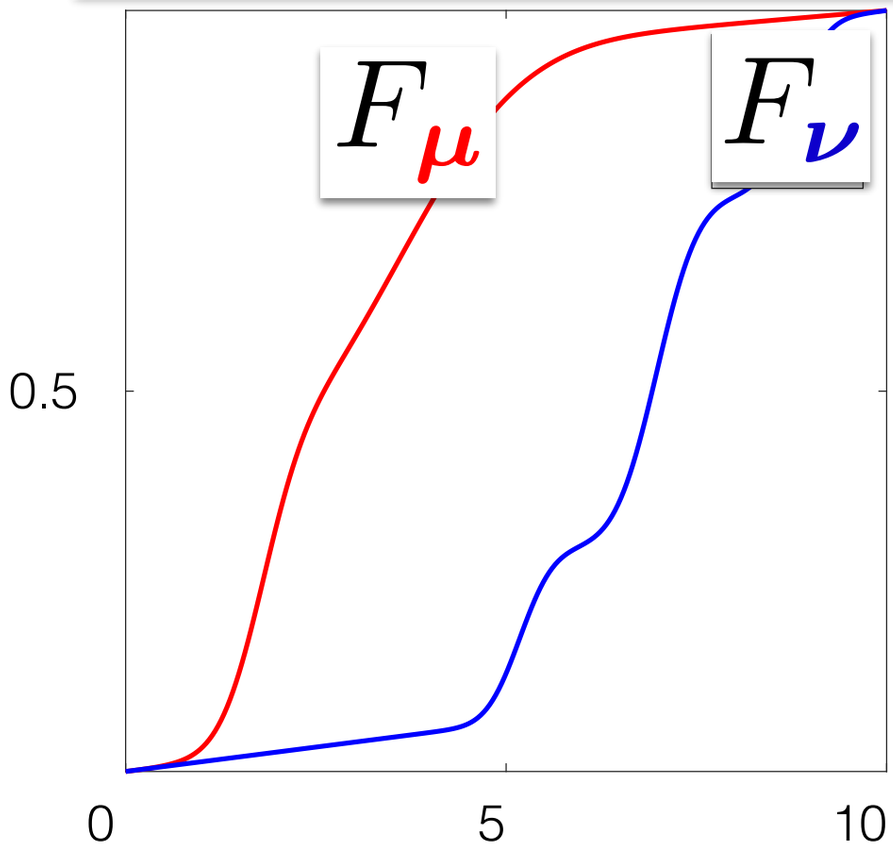


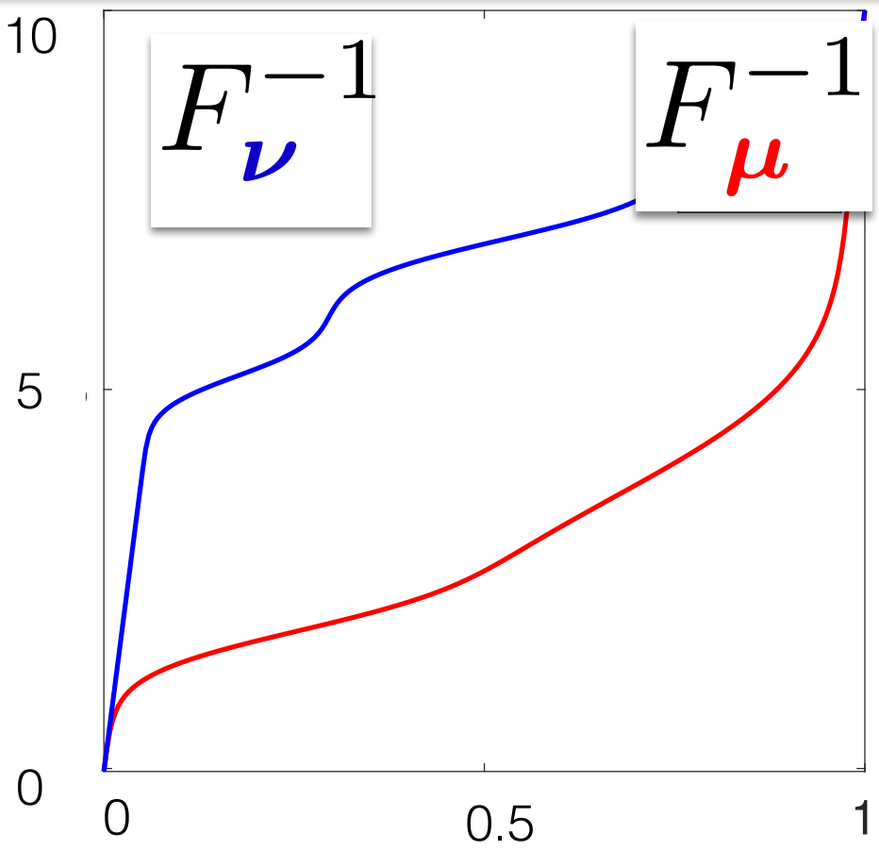
0 $\boldsymbol{\nu}$ 10

Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x,y) = \boldsymbol{c}(|x-y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$

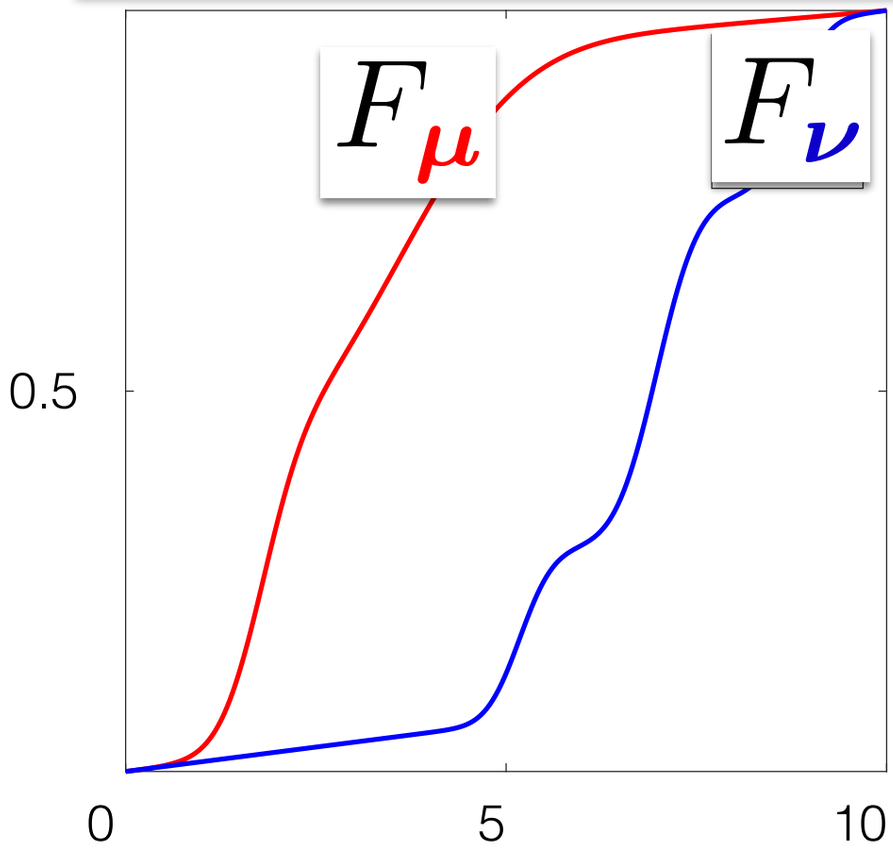


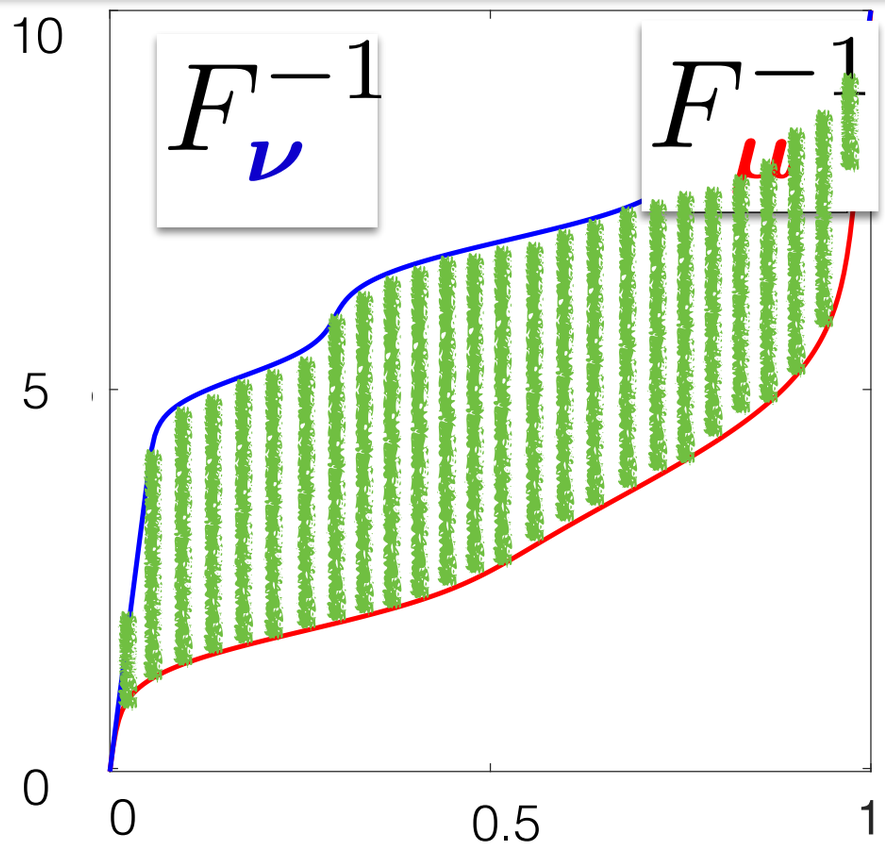


Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x,y) = \boldsymbol{c}(|x-y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$



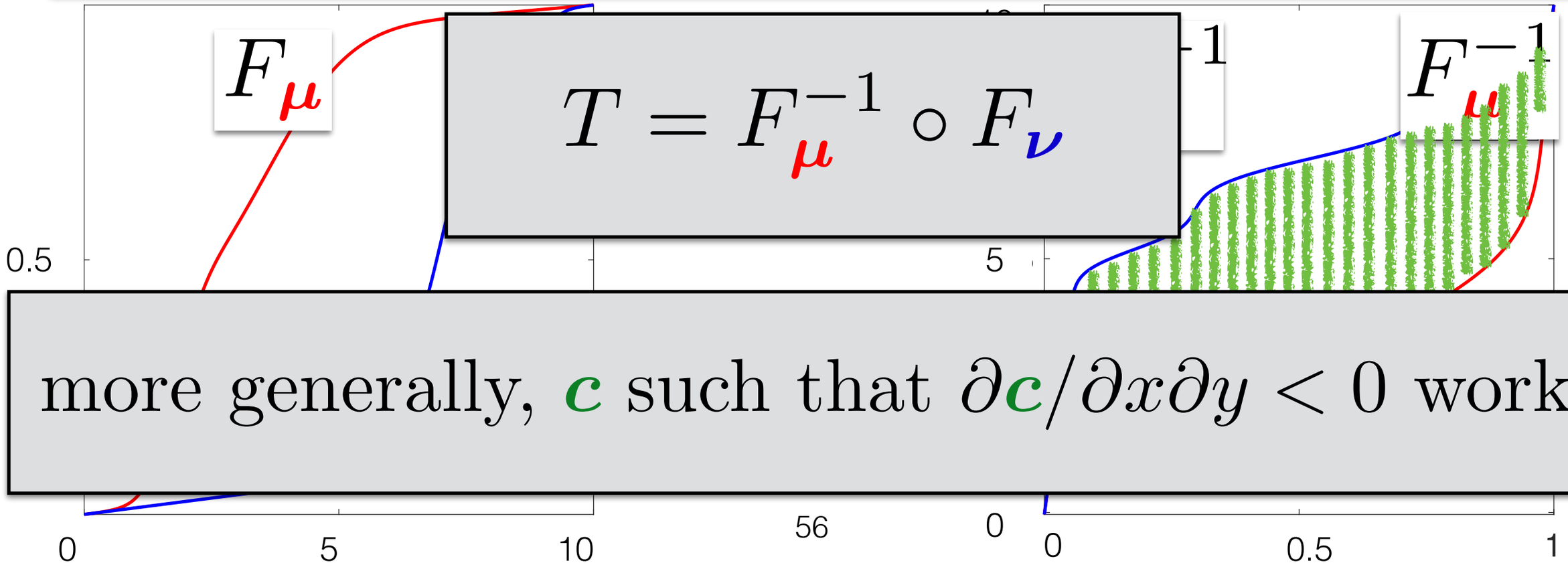


Easy (1): Univariate Measures

Remark. If $\Omega = \mathbb{R}$, $\boldsymbol{c}(x,y) = \boldsymbol{c}(|x-y|)$, $\boldsymbol{c}$ convex, $F_{\boldsymbol{\mu}}^{-1}, F_{\boldsymbol{\nu}}^{-1}$ quantile functions,

$$W(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_0^1 \boldsymbol{c}(|F_{\boldsymbol{\mu}}^{-1}(x) - F_{\boldsymbol{\nu}}^{-1}(x)|)dx$$

$$T = F_{\boldsymbol{\mu}}^{-1} \circ F_{\boldsymbol{\nu}}$$



more generally, $\boldsymbol{c}$ such that $\partial \boldsymbol{c}/\partial x \partial y < 0$ works

Easy and useful in ML

Sliced Wasserstein Distance [Rabin+'11]

$$SW(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbb{E}_{\theta \sim \mathcal{S}^{d-1}} \left[\int_0^1 \boldsymbol{c}(|F^{-1}_{\theta^T_\sharp \boldsymbol{\mu}}(x) - F^{-1}_{\theta^T_\sharp \boldsymbol{\nu}}(x)|) dx \right]$$

- Dodges the high-dimensionality curse, although it's not clear what it computes, certainly not OT.
- Works very well in practice, fast and easy.

Easy (2): Gaussian Measures

Remark. If $\Omega = \mathbb{R}^d$, $\boldsymbol{c}(x,y) = \|x-y\|^2$, and $\boldsymbol{\mu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \boldsymbol{\nu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})$ then

$$W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \|\mathbf{m}_{\boldsymbol{\mu}} - \mathbf{m}_{\boldsymbol{\nu}}\|^2 + B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2$$

where B is the Bures metric

$$B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2 = \mathrm{trace}(\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\boldsymbol{\nu}} - 2(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2})^{1/2}).$$

Easy (2): Gaussian Measures

Remark. If $\Omega = \mathbb{R}^d$, $\boldsymbol{c}(x, y) = \|x - y\|^2$, and $\boldsymbol{\mu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \boldsymbol{\nu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})$ then

$$W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \|\mathbf{m}_{\boldsymbol{\mu}} - \mathbf{m}_{\boldsymbol{\nu}}\|^2 + B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2$$

where B is the Bures metric

$$B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2 = \operatorname{trace}(\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\boldsymbol{\nu}} - 2(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2})^{1/2}).$$

The map $T : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$ is **optimal**, where $A = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}}$.

Easy (2): Gaussian Measures

Remark. If $\Omega = \mathbb{R}^d$, $\boldsymbol{c}(x,y) = \|x-y\|^2$, and $\boldsymbol{\mu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}), \boldsymbol{\nu} = \mathcal{N}(\mathbf{m}_{\boldsymbol{\nu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})$ then

$$W_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \|\mathbf{m}_{\boldsymbol{\mu}} - \mathbf{m}_{\boldsymbol{\nu}}\|^2 + B(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}, \boldsymbol{\Sigma}_{\boldsymbol{\nu}})^2$$

where B is the Bures

$$\text{trace}(\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_{\boldsymbol{\nu}} - 2(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{1/2})^{1/2}).$$

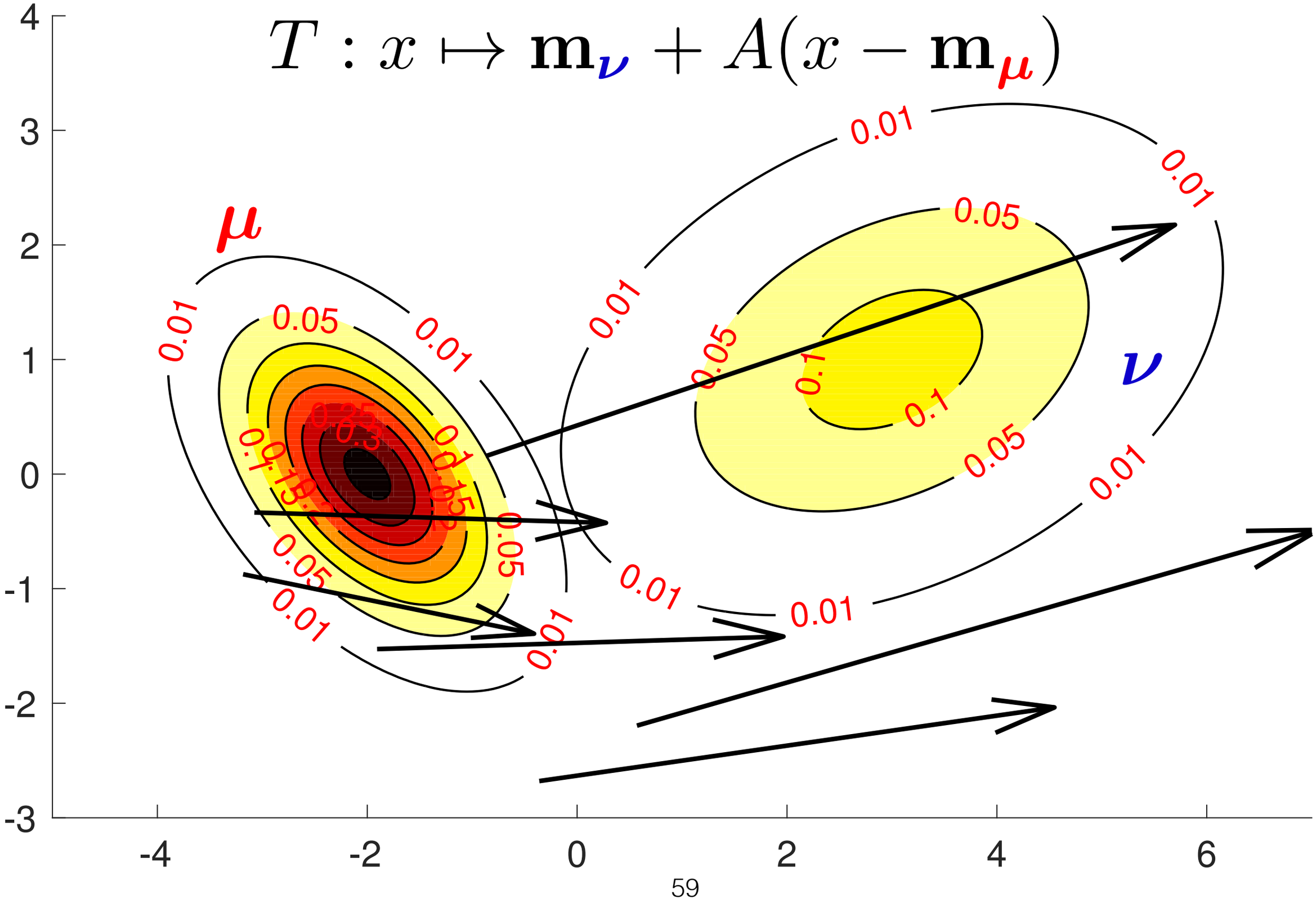
if $X \sim \mathcal{N}(m, \boldsymbol{\Sigma})$ then $Y := CX + b \sim \mathcal{N}(Cm + b, C\boldsymbol{\Sigma}C^T)$.

The map $T : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$ is **optimal**, where

$$A = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}} \left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\nu}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\frac{1}{2}} \right)^{\frac{1}{2}} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-\frac{1}{2}}.$$

Easy (2): Gaussian Measures

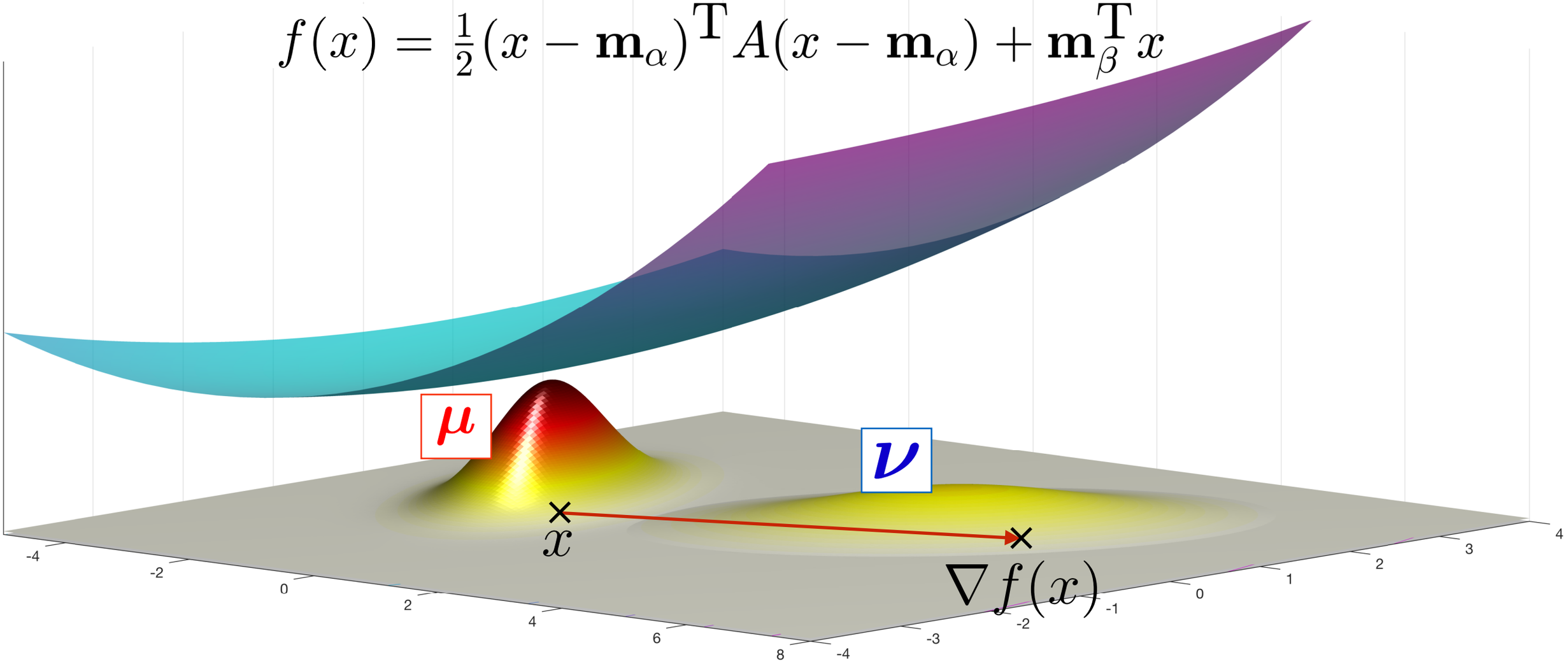
$$T : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$$



Easy (2): Gaussian Measures

$$T = \nabla\psi : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$$

$$f(x) = \tfrac{1}{2}(x - \mathbf{m}_\alpha)^{\mathrm{T}} A(x - \mathbf{m}_\alpha) + \mathbf{m}_\beta^{\mathrm{T}} x$$



Easy (3): Elliptical Distributions

$$T = \nabla\psi : x \mapsto \mathbf{m}_{\boldsymbol{\nu}} + A(x - \mathbf{m}_{\boldsymbol{\mu}})$$

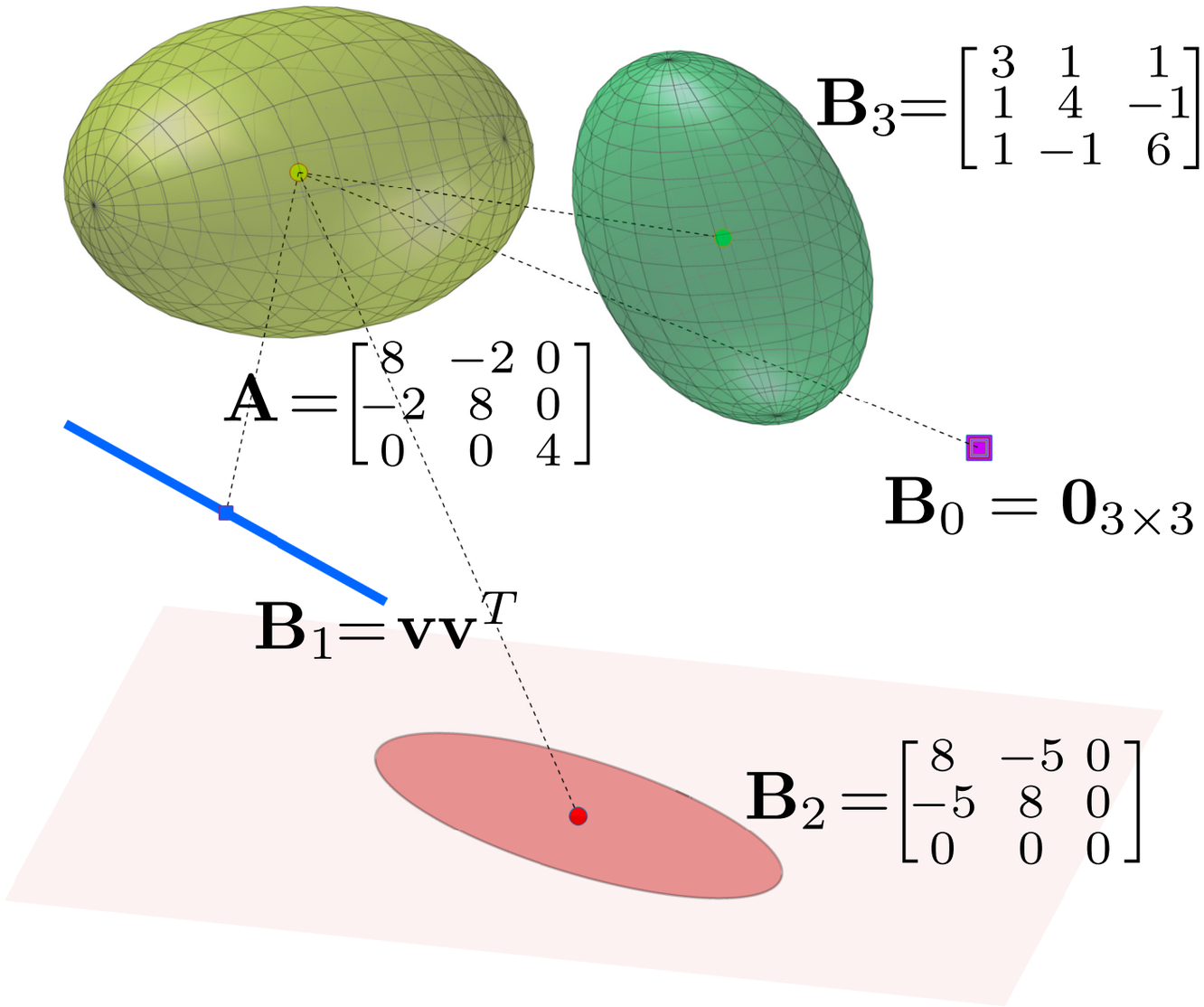
[Gelbrich'92] shows that the linear map T is also **optimal** for elliptically contoured distributions, *i.e.* distributions whose MGF are

$$\phi_X(\mathbf{t}) = \mathbb{E}\left[e^{\sqrt{-1}\mathbf{t}^T X}\right] = e^{\sqrt{-1}\mathbf{t}^T \mathbf{m}} g(\mathbf{t}^T \boldsymbol{C} \mathbf{t})$$

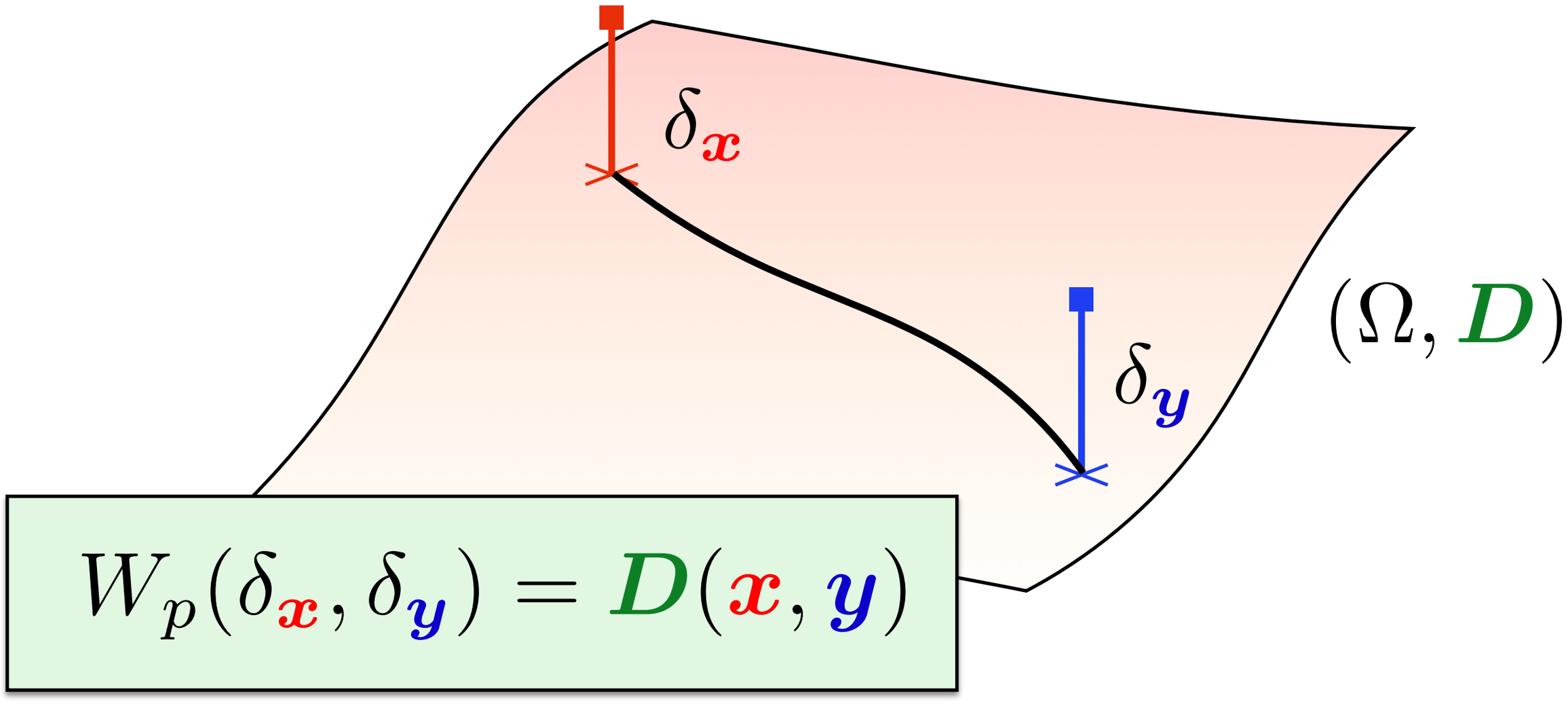
g of positive type.

Same formula applies, but variance is a factor (depends on g) of C, hence Bures factor is scaled.

Easy (3): Uniform Ellipses



Wasserstein Between Two Diracs

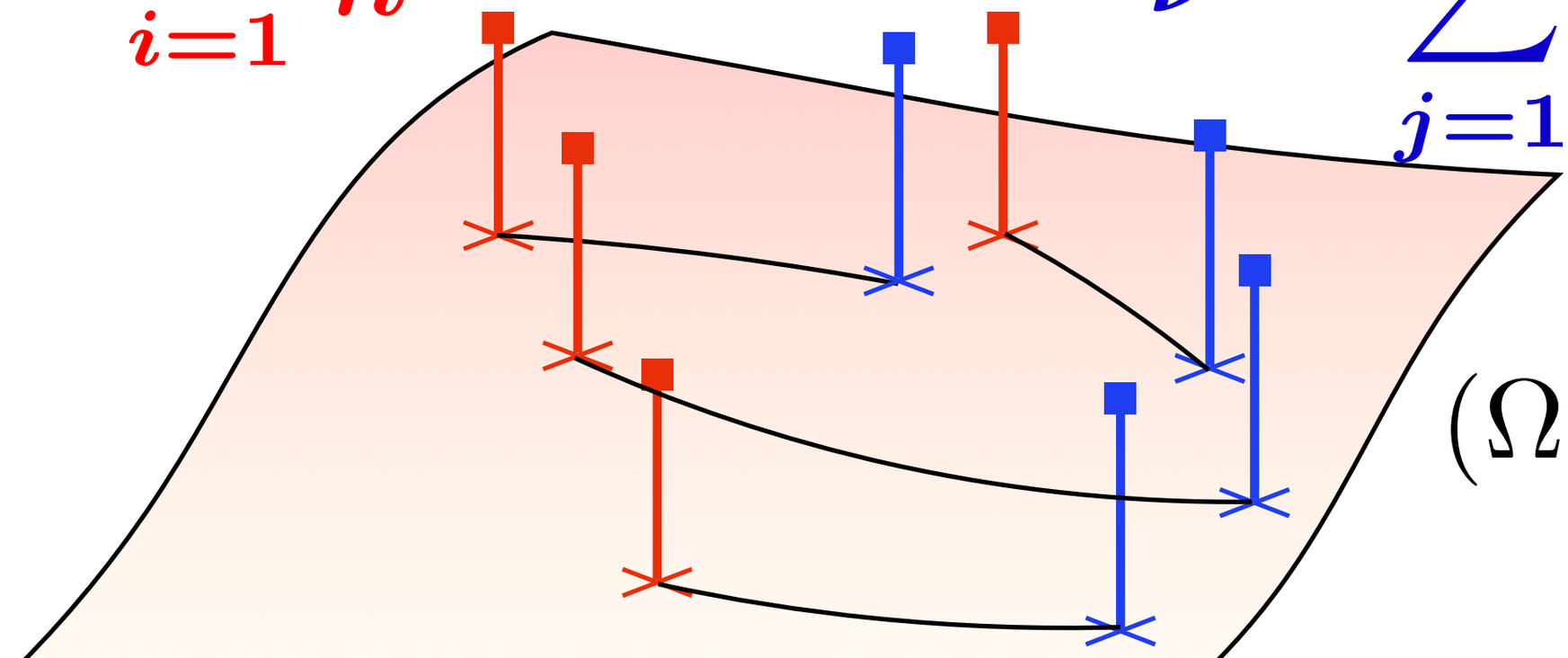


$$W_p(\delta_{\boldsymbol{x}}, \delta_{\boldsymbol{y}}) = \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{y})$$

Linear Assignment $\subset$ Wasserstein

$$\mu = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i}$$

$$\nu = \sum_{j=1}^{n} \frac{1}{n} \delta_{y_j}$$

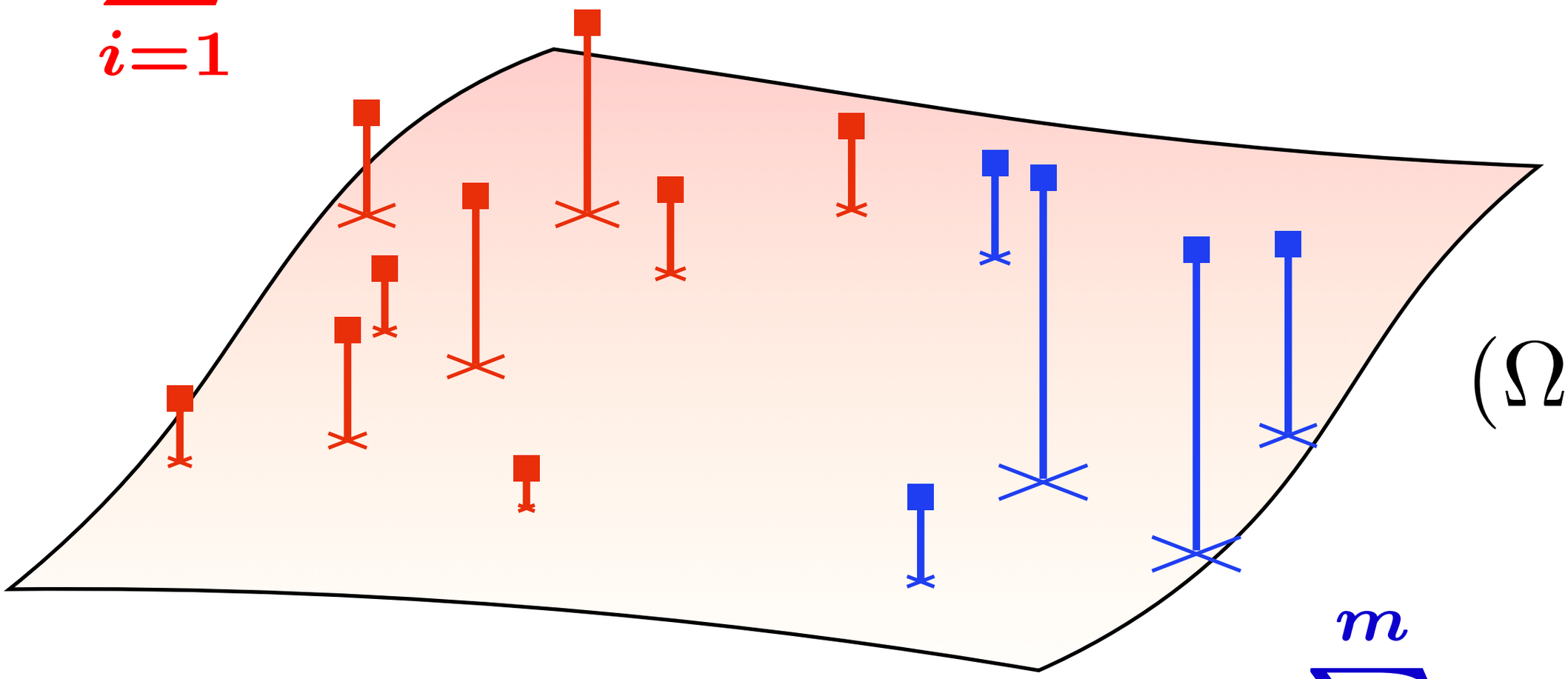


(Ω, D)

$$W_p^p(\mu, \nu) = \min_{\sigma \in S_n} \frac{1}{n} \sum_{i=1}^{n} D(x_i, y_{\sigma_i})^p$$

OT on Two Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$

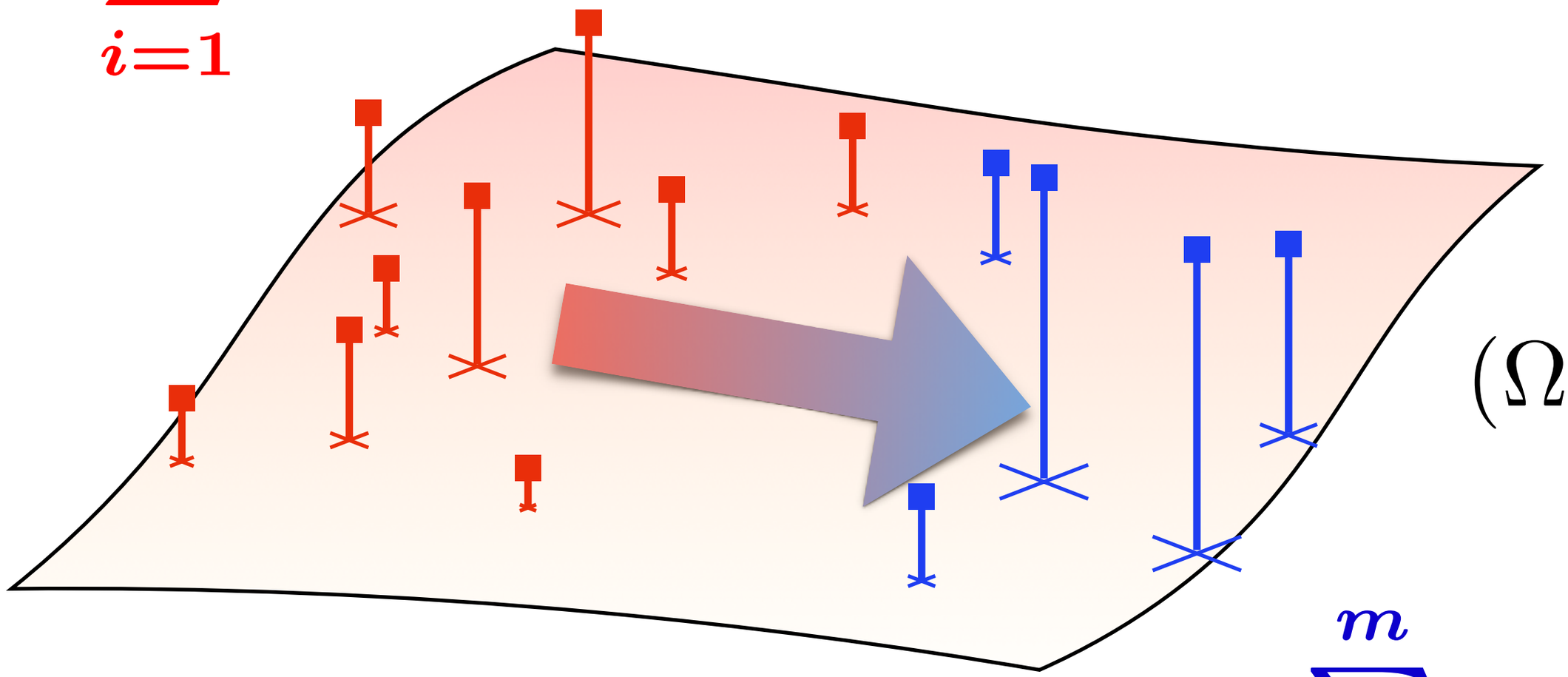


(Ω, D)

$$\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

OT on Two Empirical Measures

$$\mu = \sum_{i=1}^{n} a_i \delta_{x_i}$$



(Ω, D)

$$\nu = \sum_{j=1}^{m} b_j \delta_{y_j}$$

Wasserstein on Empirical Measures

Consider $\boldsymbol{\mu} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{\delta}_{\boldsymbol{x}_i}$ and $\boldsymbol{\nu} = \sum_{j=1}^{m} \boldsymbol{b}_j \boldsymbol{\delta}_{\boldsymbol{y}_j}$.

$$M_{\boldsymbol{X}\boldsymbol{Y}} \stackrel{\text{def}}{=} [D(\boldsymbol{x}_i, \boldsymbol{y}_j)^p]_{ij}$$

$$U(\boldsymbol{a}, \boldsymbol{b}) \stackrel{\text{def}}{=} \{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} | \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$$

Def. Optimal Transport Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}, M_{\boldsymbol{X}\boldsymbol{Y}} \rangle$$

Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu}, \boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P} \in \mathbb{R}_+^{n \times m} \\ \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{n\times m} \\ \boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Def. Dual OT problem

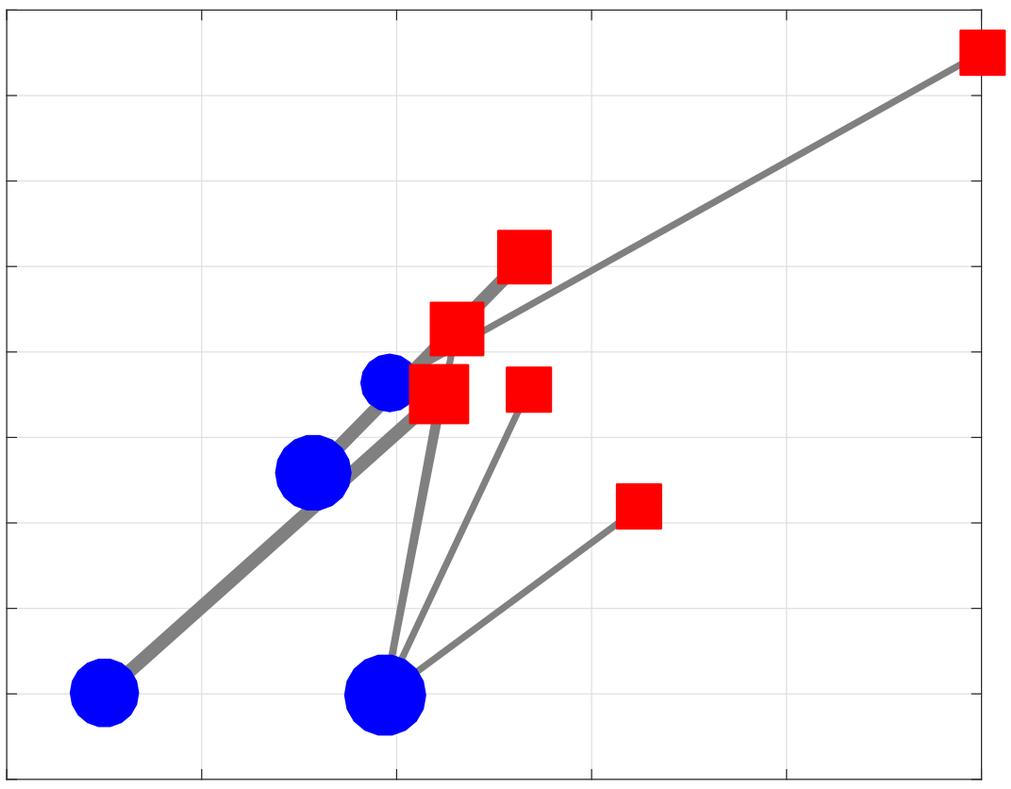
$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\alpha}_i+\boldsymbol{\beta}_j\leq D(\boldsymbol{x}_i,\boldsymbol{y}_j)^p}} \alpha^T\boldsymbol{a}+\beta^T\boldsymbol{b}$$

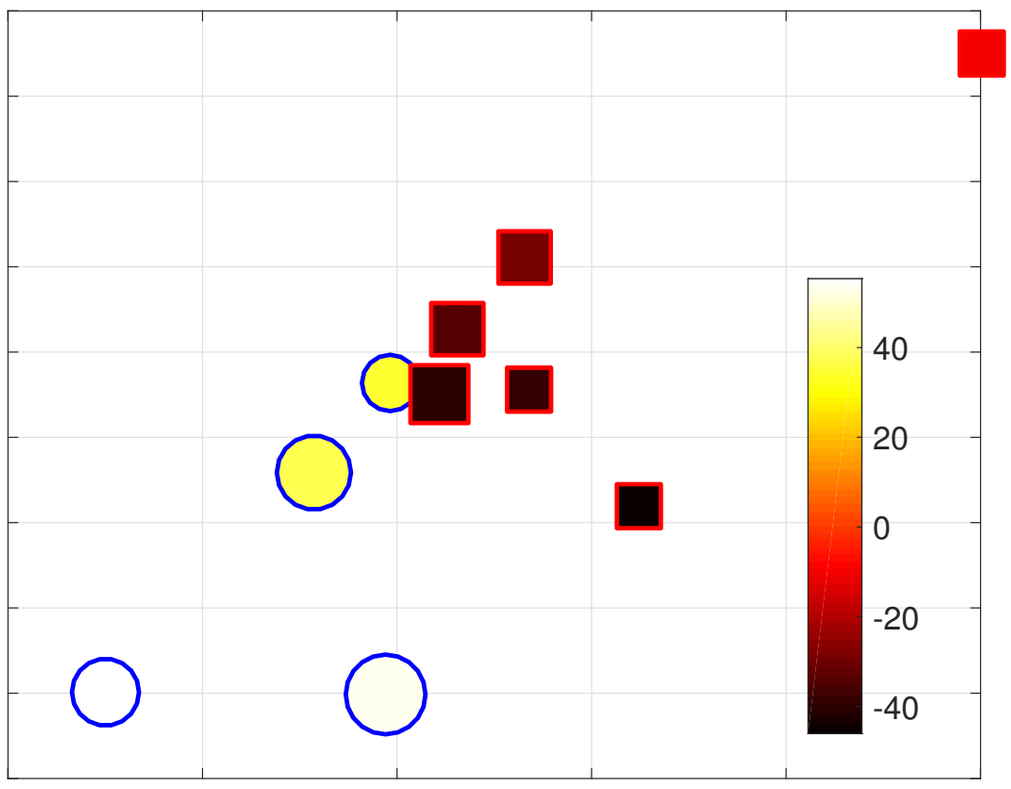
Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{n\times m} \\ \boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Def. Dual OT problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\alpha}_i+\boldsymbol{\beta}_j\leq D(\boldsymbol{x}_i,\boldsymbol{y}_j)^p}} \alpha^T\boldsymbol{a}+\beta^T\boldsymbol{b}$$





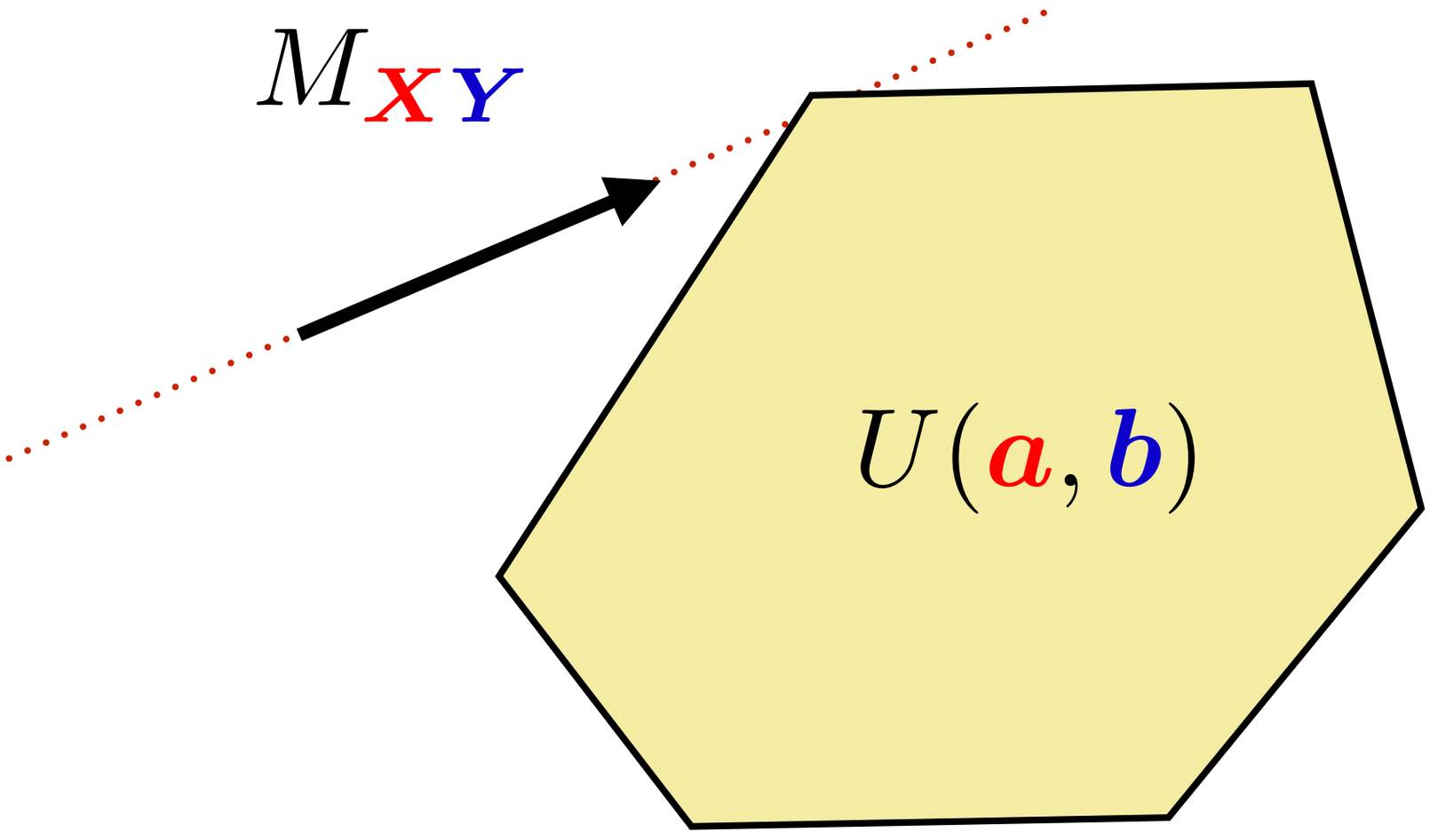
Dual Kantorovich Problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}_+^{n\times m} \\ \boldsymbol{P}\mathbf{1}_m=\boldsymbol{a},\boldsymbol{P}^T\mathbf{1}_n=\boldsymbol{b}}} \langle \boldsymbol{P}, M_{\boldsymbol{XY}} \rangle$$

Def. Dual OT problem

$$W_p^p(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^n,\boldsymbol{\beta}\in\mathbb{R}^m \\ \boldsymbol{\alpha}_i+\boldsymbol{\beta}_j\leq D(\boldsymbol{x}_i,\boldsymbol{y}_j)^p}} \alpha^T\boldsymbol{a}+\beta^T\boldsymbol{b}$$

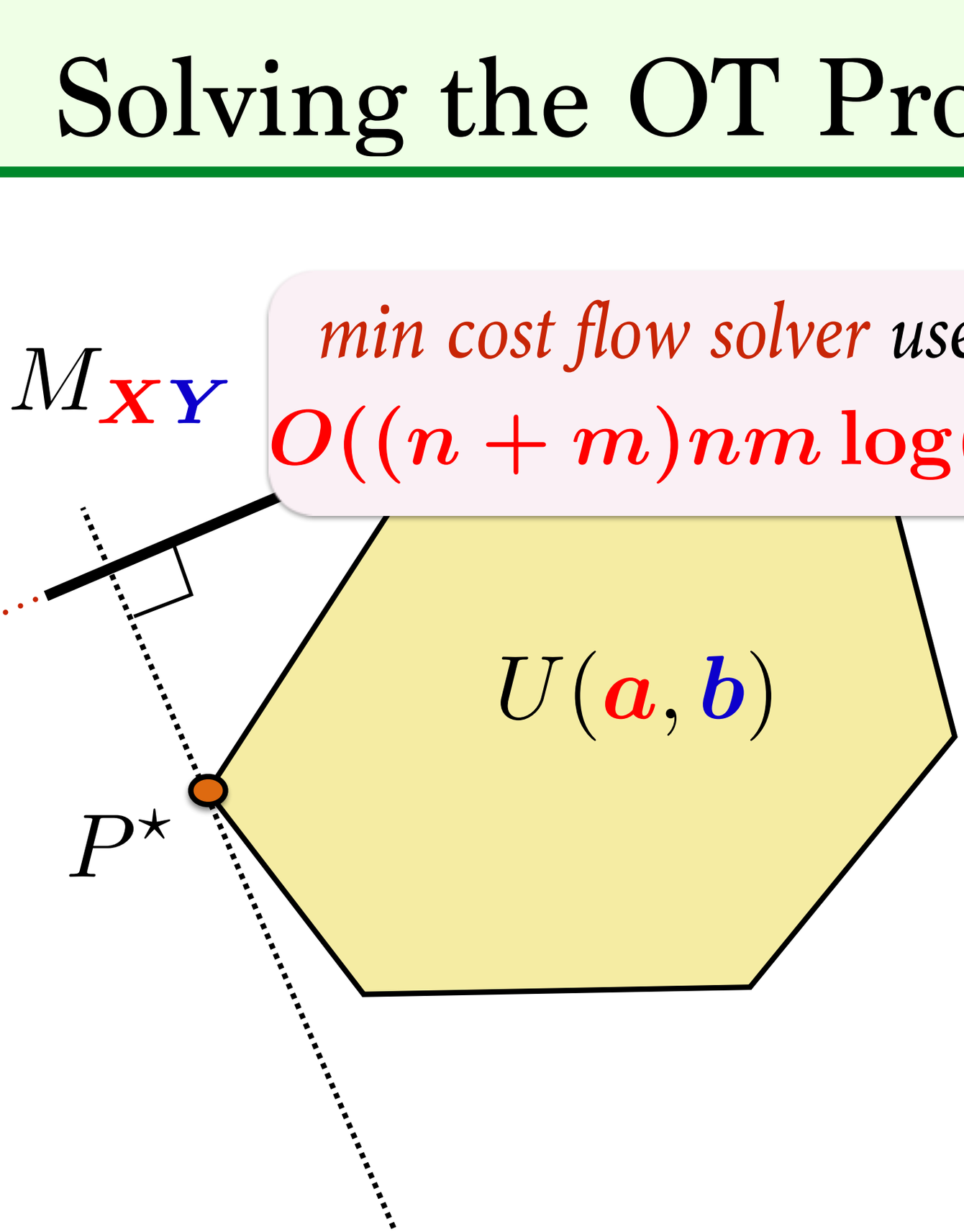
Solving the OT Problem



Solving the OT Problem

min cost flow solver used in practice.

$O((n+m)nm\log(n+m))$



[Tarjan'97]

Solving the OT Problem

min cost flow solver used in practice.

$$O((n+m)nm\log(n+m))$$

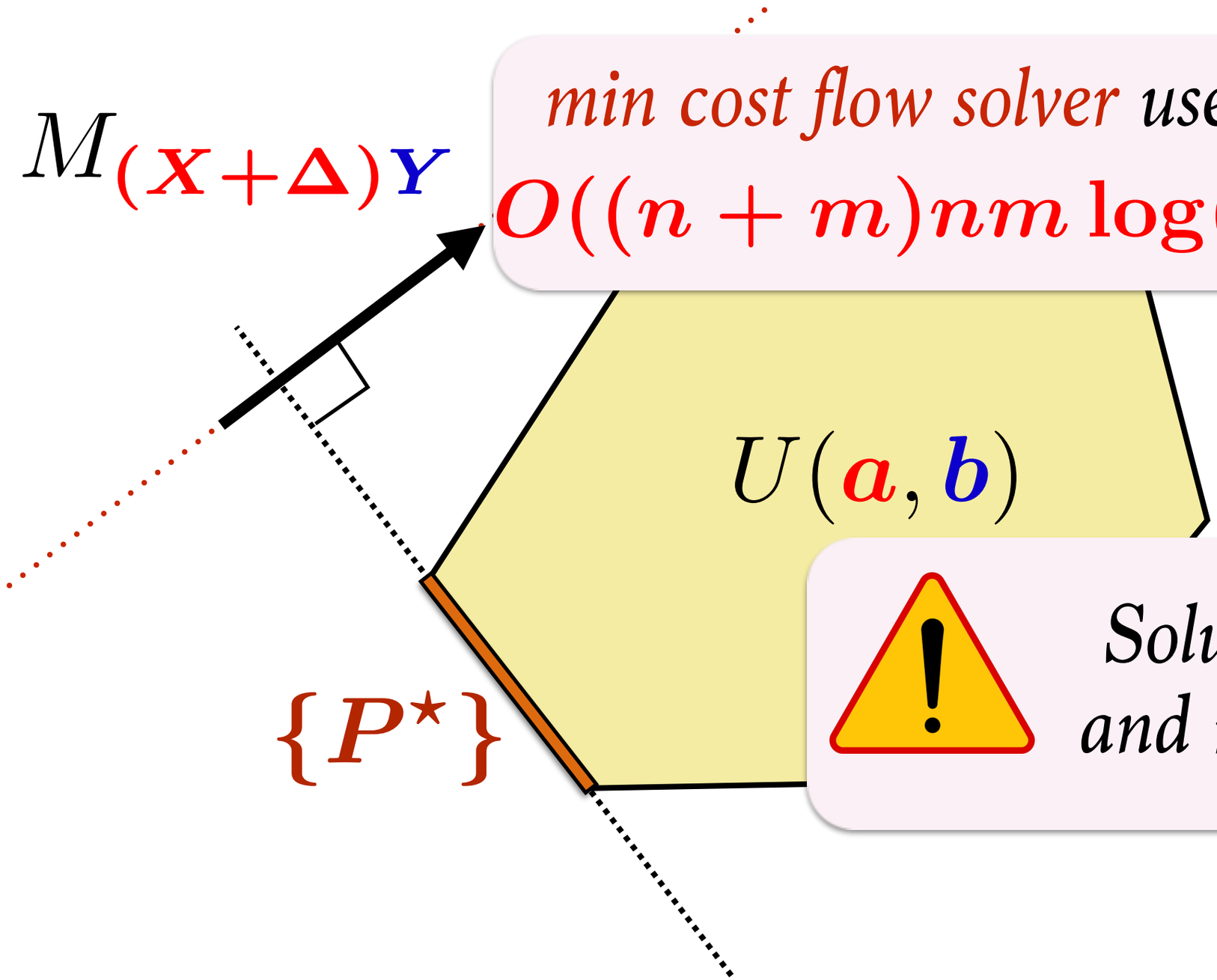
M_{XY}

$U(a, b)$

$P^\star$

Solution $P^\star$ unstable and not always unique.

Solving the OT Problem

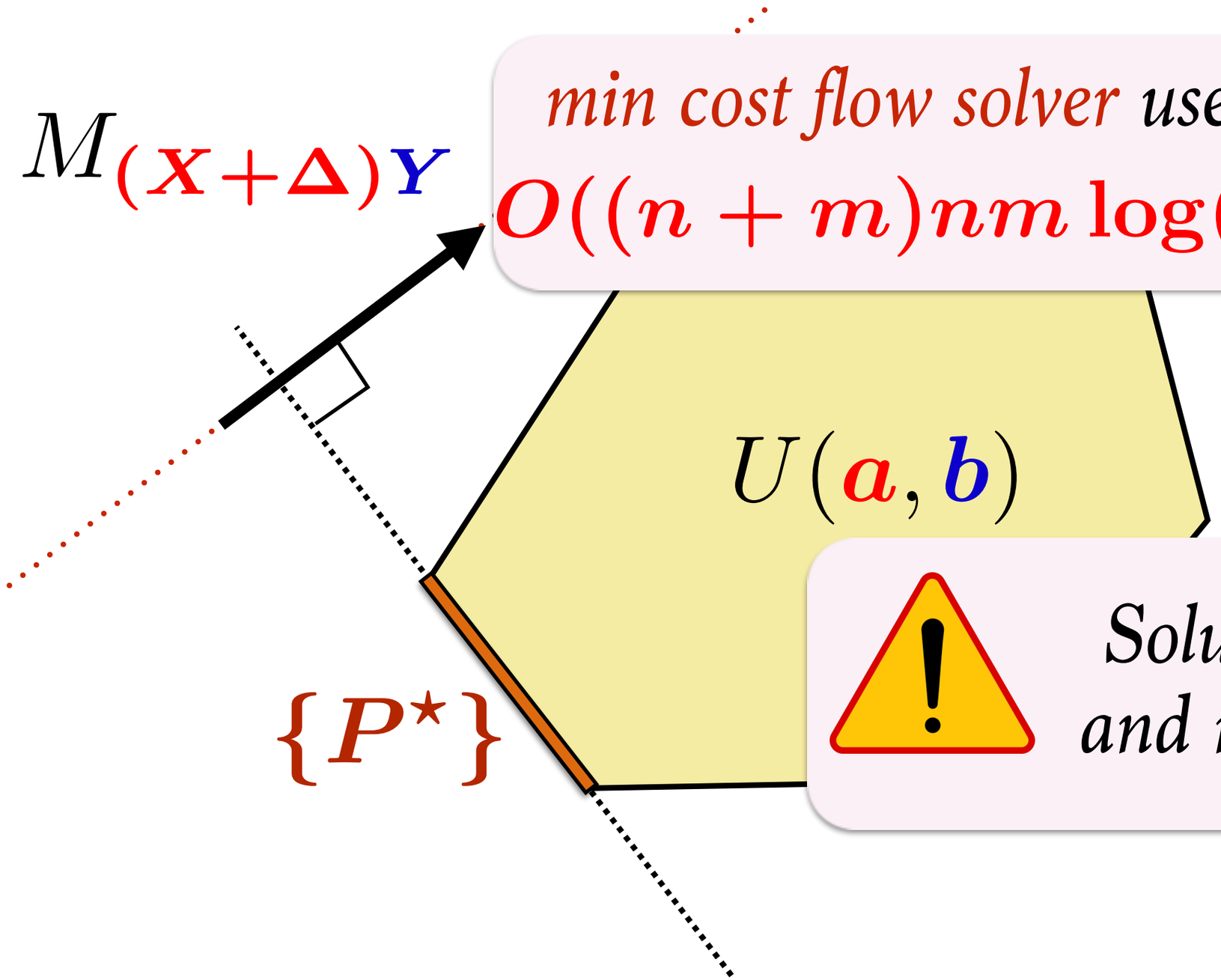


min cost flow solver used in practice.

$O((n+m)nm\log(n+m))$

Solution $P^\star$ unstable and not always unique.

Solving the OT Problem

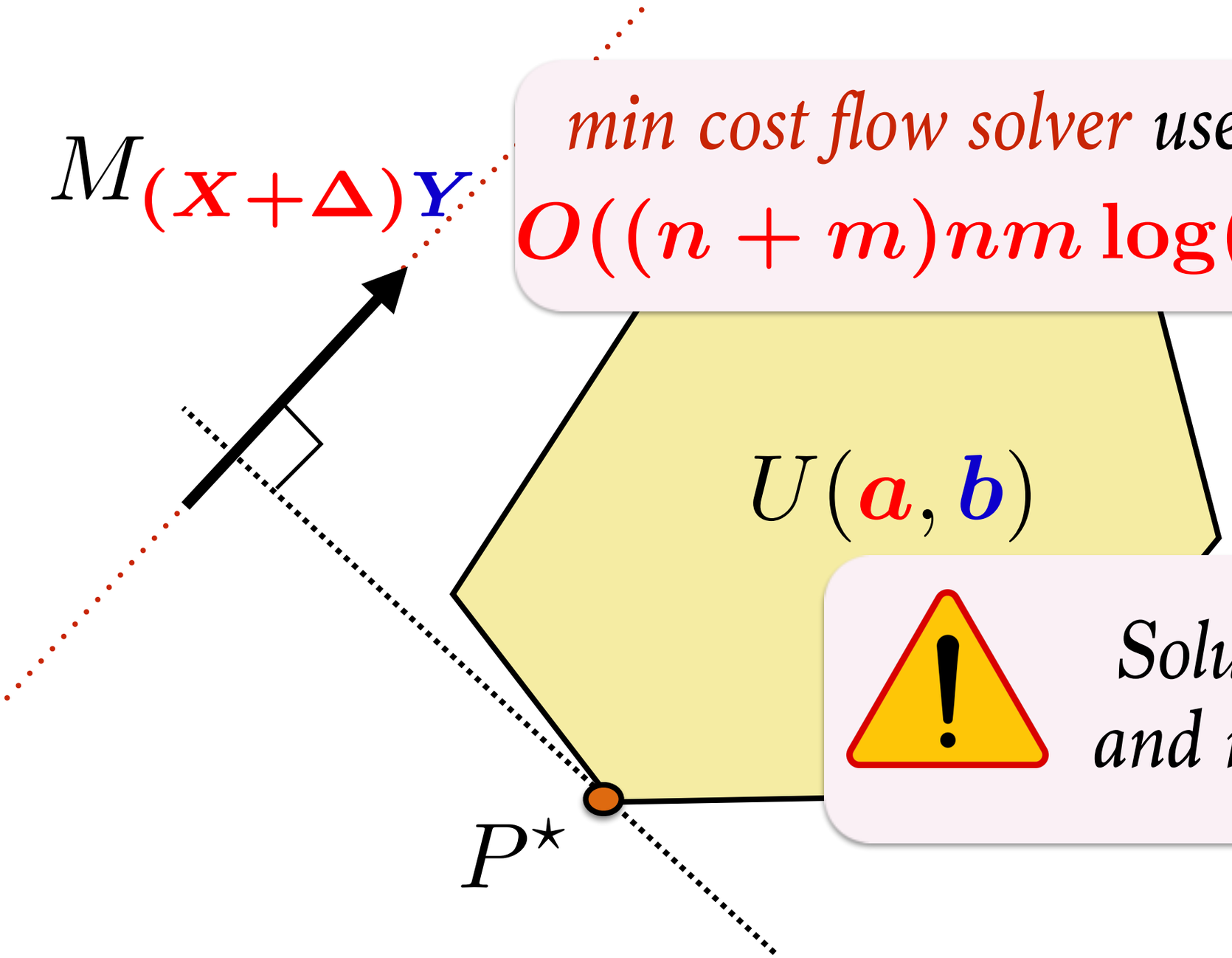


min cost flow solver used in practice.

$O((n+m)nm\log(n+m))$

Solution $P^\star$ unstable and not always unique.

Solving the OT Problem

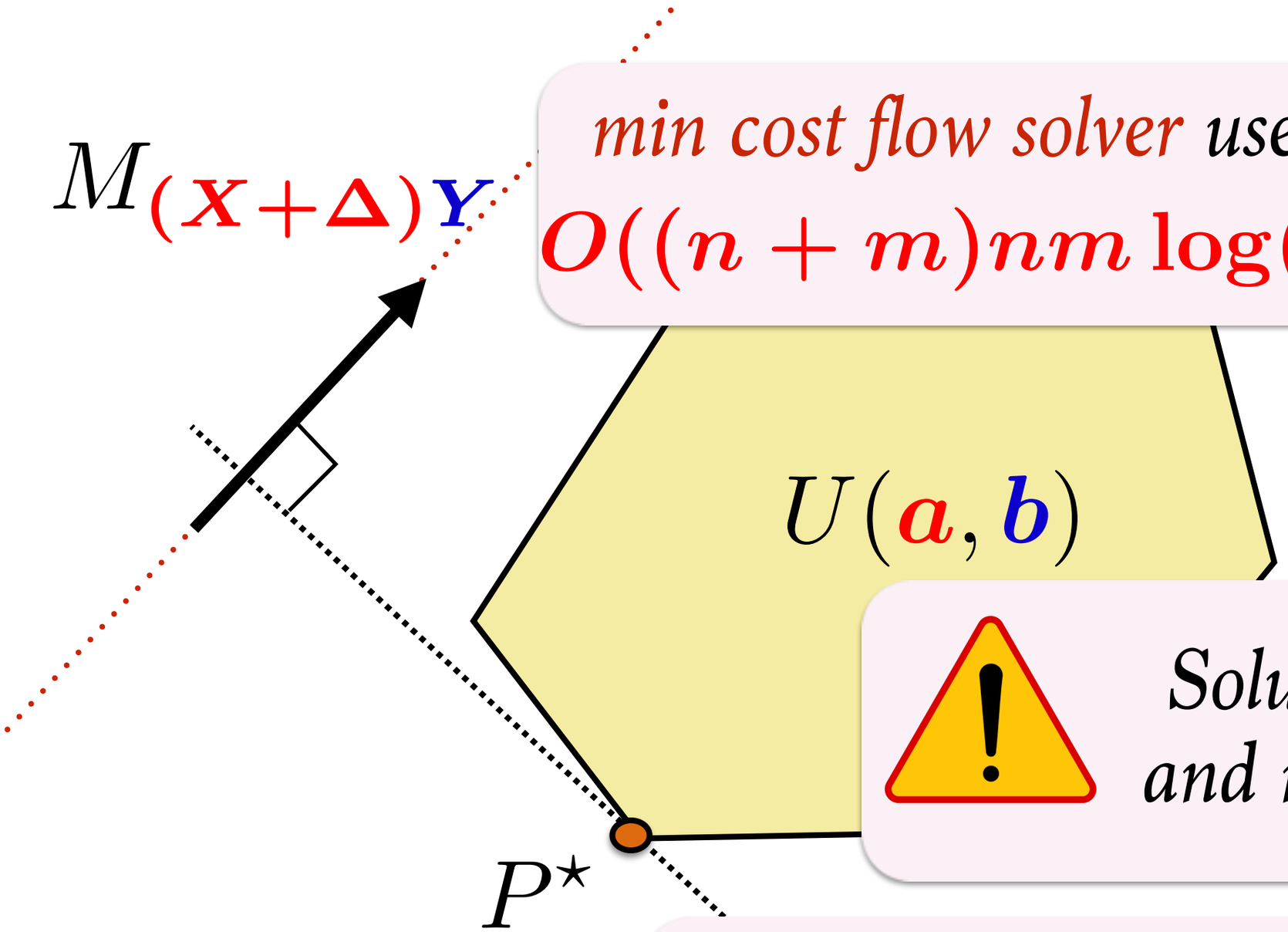


min cost flow solver used in practice.

$O((n+m)nm\log(n+m))$

Solution $P^\star$ unstable and not always unique.

Solving the OT Problem



min cost flow solver used in practice.

$$O((n+m)nm\log(n+m))$$

Solution $P^\star$ unstable and not always unique.

$W_p^p(\mu, \nu)$ not differentiable.

Discrete OT Problem

emd.c

emd.c:6:1 <No selected symbol>

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;                /* NEXT COLUMN */
  struct node2_t *NextR;                /* NEXT ROW */
} node2_t;


/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                              /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];               /* THE BASIC VARIABLES VECTOR */
```

Discrete OT Problem

emd.c

emd.c:6:1 <No selected symbol>

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;                /* NEXT COLUMN */
  struct node2_t *NextR;                /* NEXT ROW */
} node2_t;


/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                              /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];               /* THE BASIC VARIABLES VECTOR */
```

Discrete OT Problem

emd.c

emd.c:6:1 <No selected symbol>

```c
/*
    emd.c

    Last update: 3/14/98

    An implementation of the Earth Movers Distance.
    Based of the solution for the Transportation problem as described in
    "Introduction to Mathematical Programming" by F. S. Hillier and
    G. J. Lieberman, McGraw-Hill, 1990.

    Copyright (C) 1998 Yossi Rubner
    Computer Science Department, Stanford University
    E-Mail: rubner@cs.stanford.edu   URL: http://vision.stanford.edu/~rubner
*/

/*#include <stdio.h>
#include <stdlib.h>*/
#include <math.h>

#include "emd.h"

#define DEBUG_LEVEL 0
/*
 DEBUG_LEVEL:
   0 = NO MESSAGES
   1 = PRINT THE NUMBER OF ITERATIONS AND THE FINAL RESULT
   2 = PRINT THE RESULT AFTER EVERY ITERATION
   3 = PRINT ALSO THE FLOW AFTER EVERY ITERATION
   4 = PRINT A LOT OF INFORMATION (PROBABLY USEFUL ONLY FOR THE AUTHOR)
*/


#define MAX_SIG_SIZE1  (MAX_SIG_SIZE+1)  /* FOR THE POSIBLE DUMMY FEATURE */

/* NEW TYPES DEFINITION */

/* node1_t IS USED FOR SINGLE-LINKED LISTS */
typedef struct node1_t {
  int i;
  double val;
  struct node1_t *Next;
} node1_t;

/* node1_t IS USED FOR DOUBLE-LINKED LISTS */
typedef struct node2_t {
  int i, j;
  double val;
  struct node2_t *NextC;                /* NEXT COLUMN */
  struct node2_t *NextR;                /* NEXT ROW */
} node2_t;


/* GLOBAL VARIABLE DECLARATION */
static int _n1, _n2;                              /* SIGNATURES SIZES */
static float _C[MAX_SIG_SIZE1][MAX_SIG_SIZE1];/* THE COST MATRIX */
static node2_t _X[MAX_SIG_SIZE1*2];               /* THE BASIC VARIABLES VECTOR */
```