On the Interplay of Optimal Transport and Distributionally Robust Optimization

Daniel Kuhn

Risk Analytics and Optimization Chair
École Polytechnique Fédérale de Lausanne
rao.epfl.ch

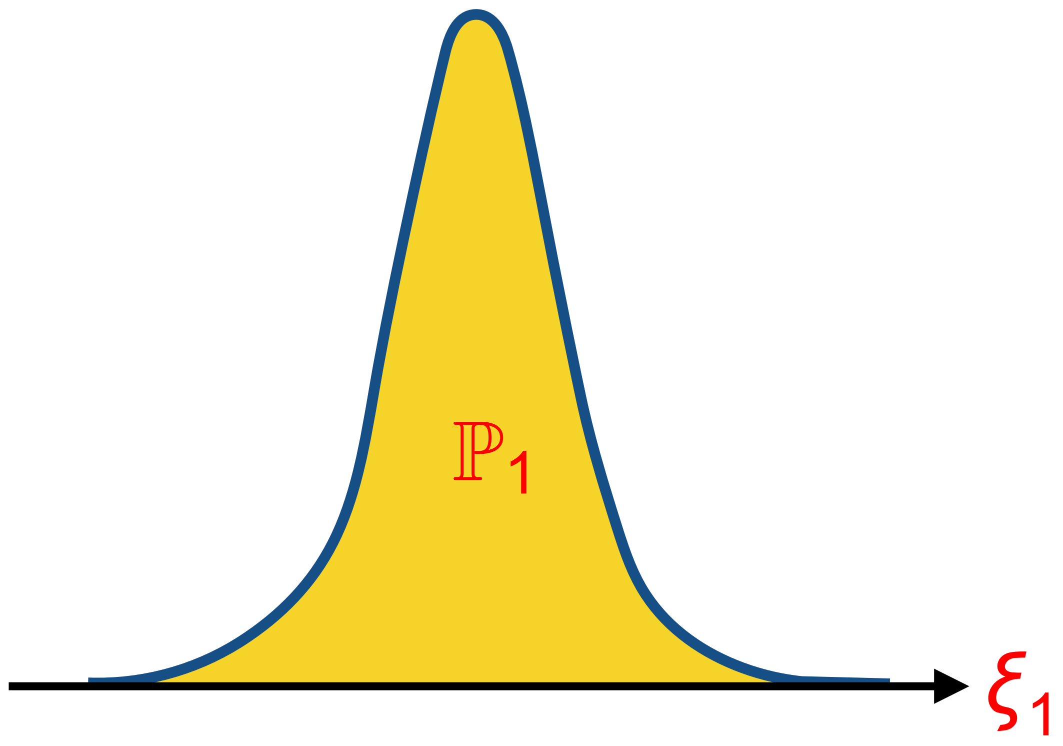
FNSNF
Swiss National Science Foundation

NCCR Automation

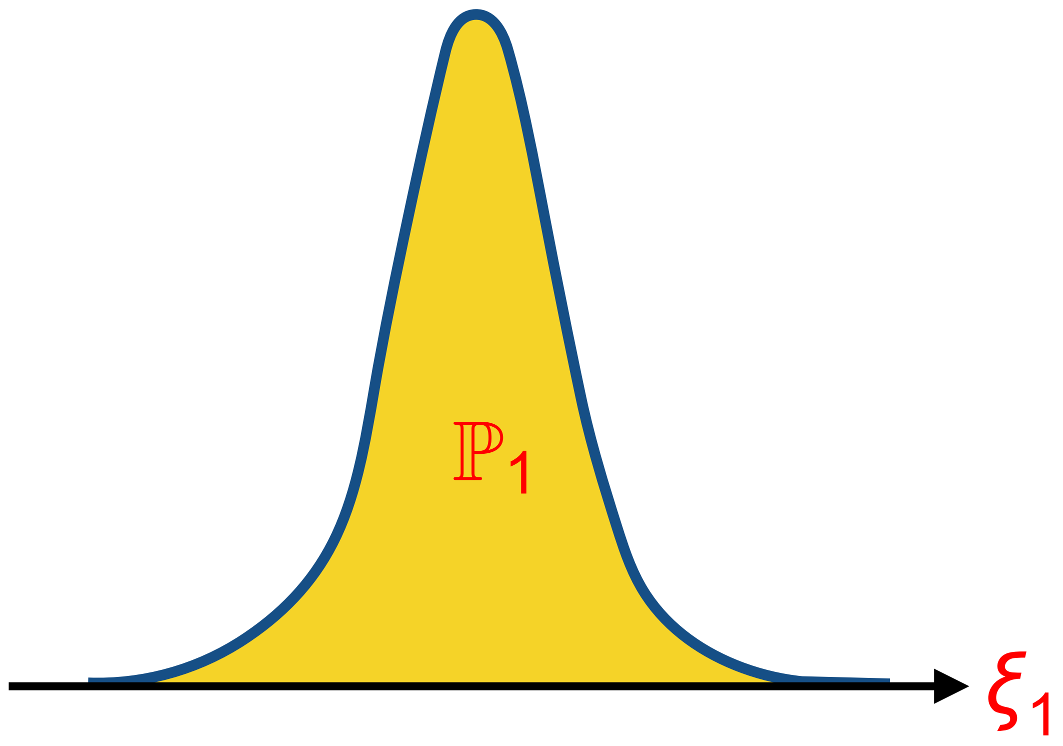
EPFL

Optimal Transport (OT)

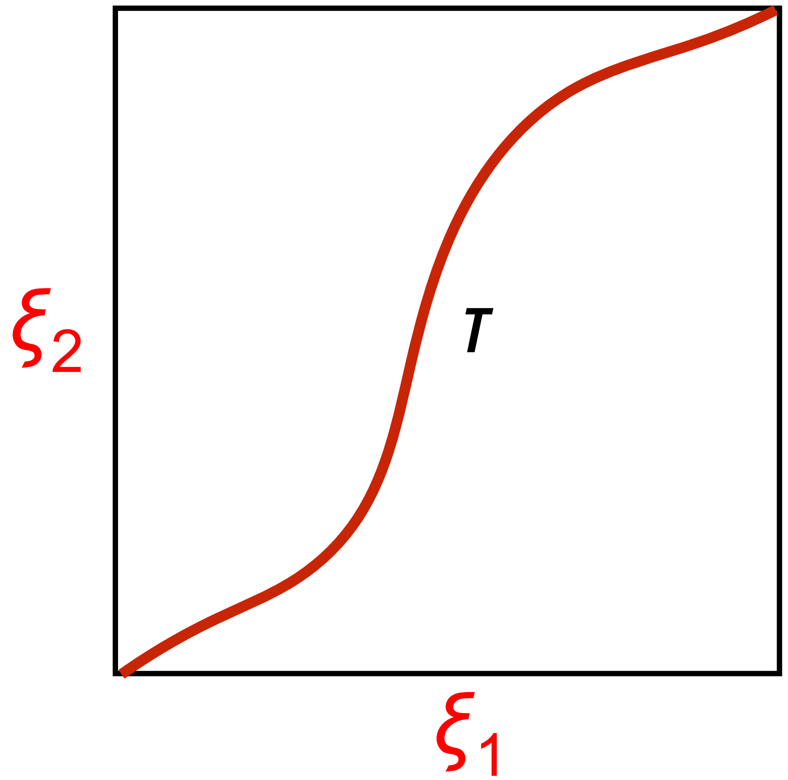
Pushforward Distribution



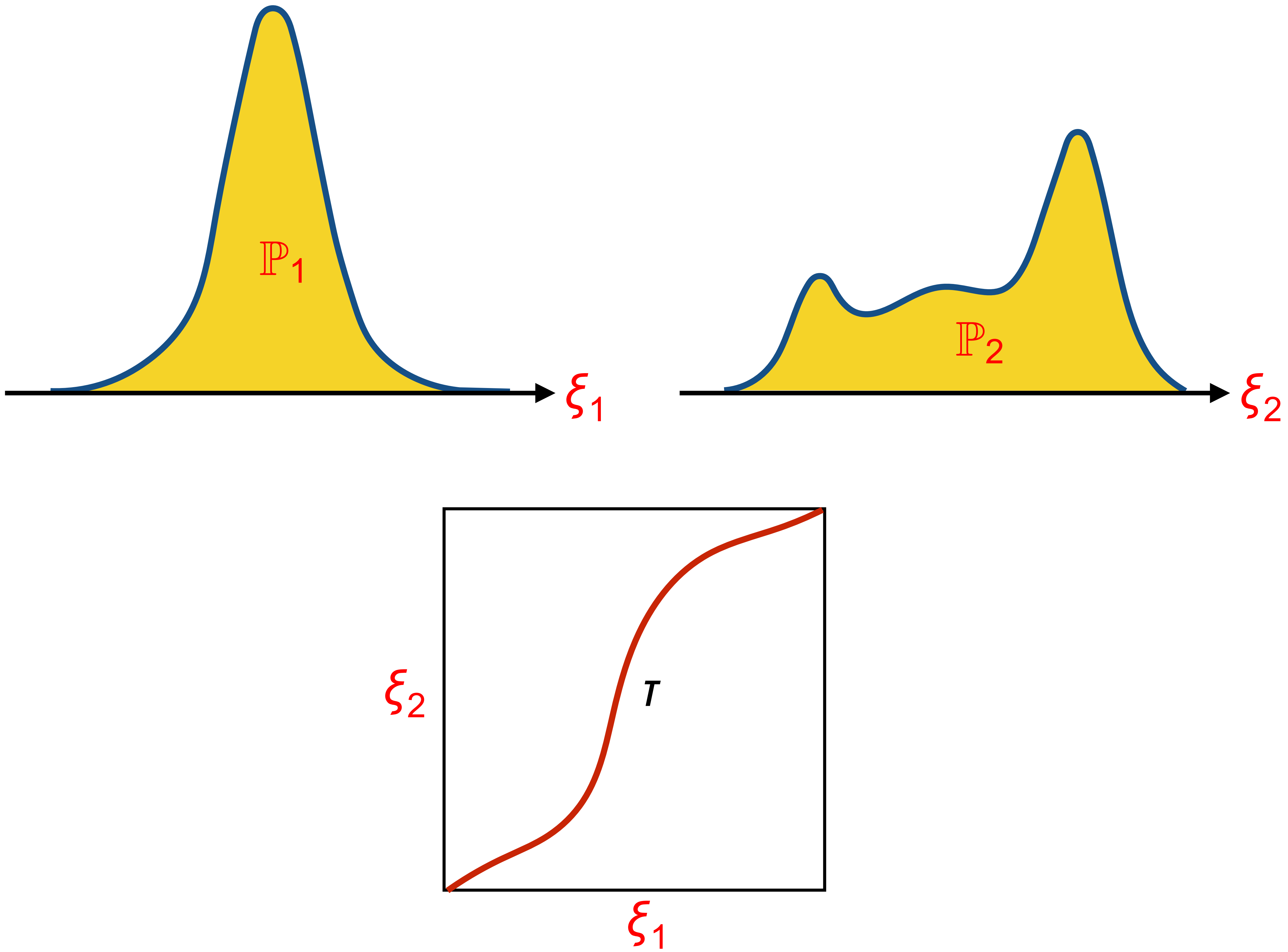
Pushforward Distribution



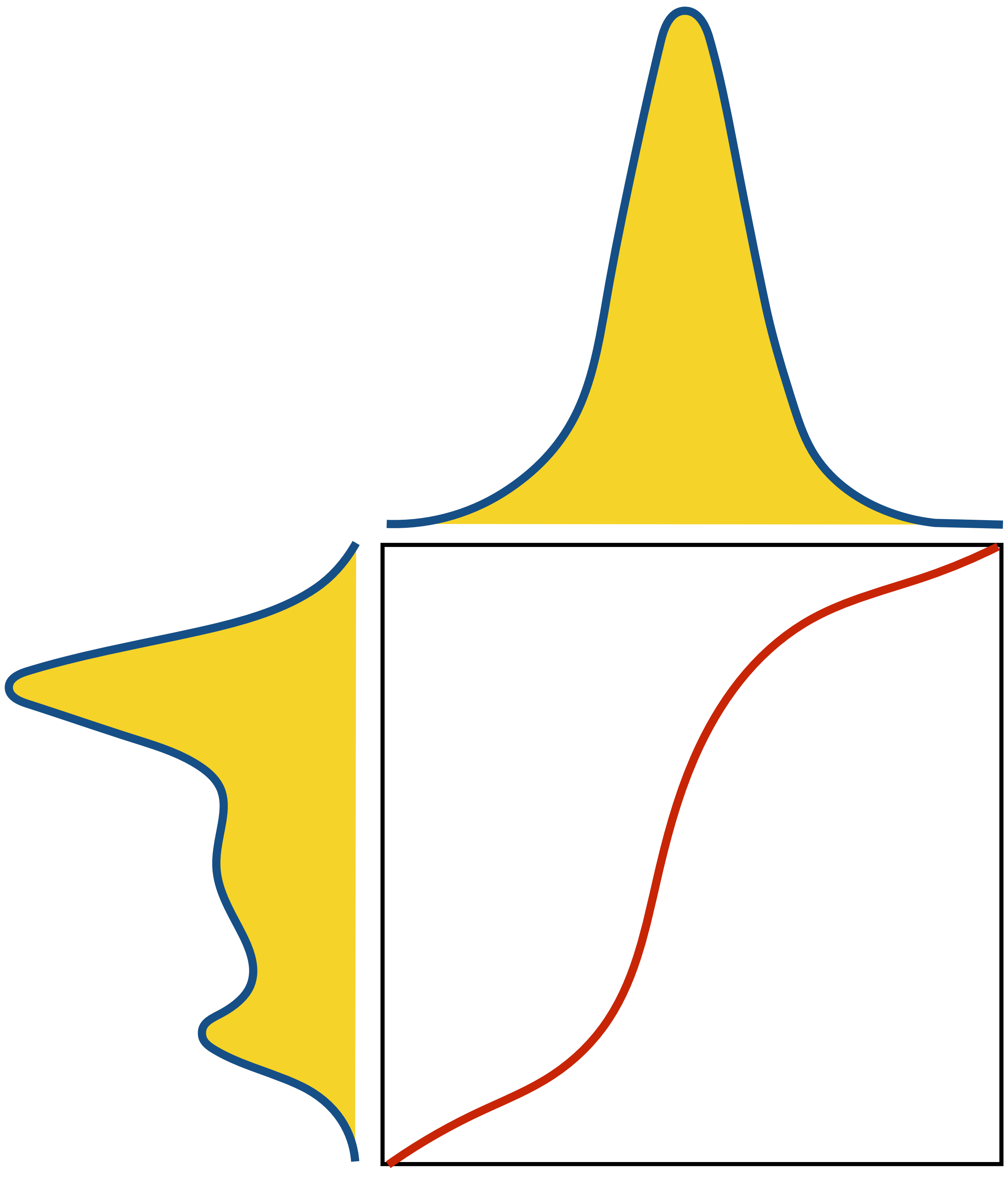
$$\xi_2 = T(\xi_1)$$



Pushforward Distribution

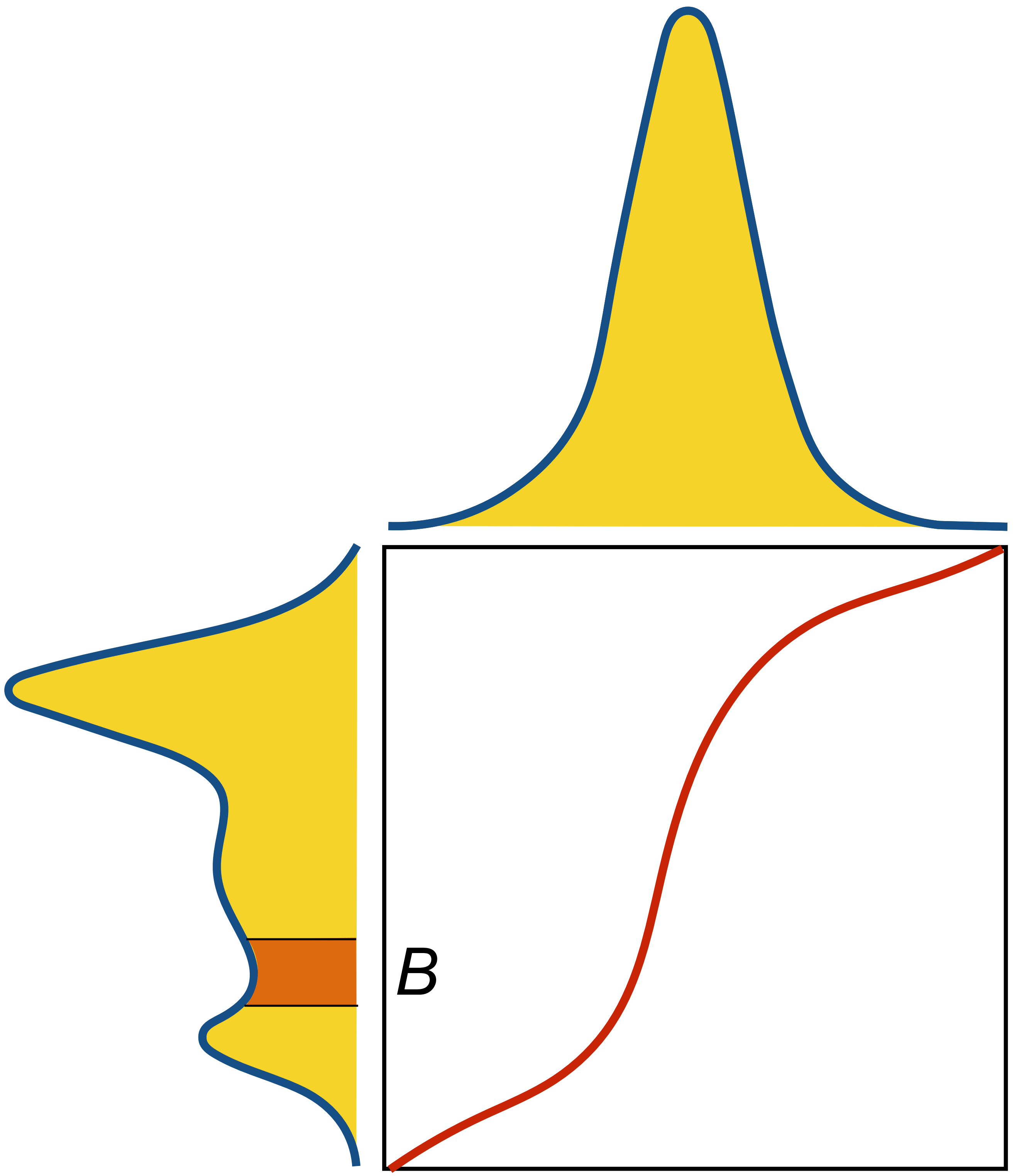


Pushforward Distribution



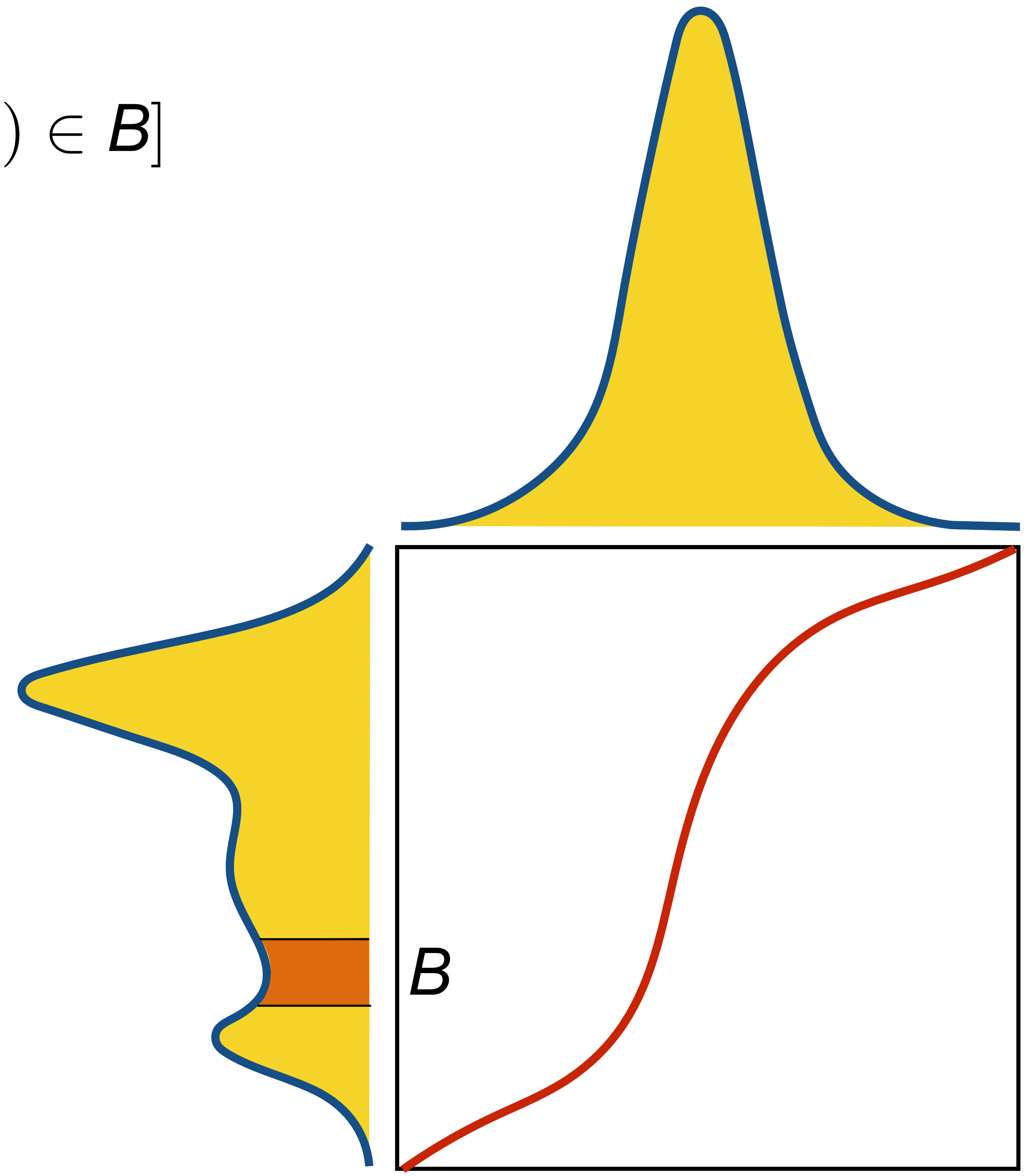
Pushforward Distribution

$\mathbb{P}_2[\xi_2 \in B]$



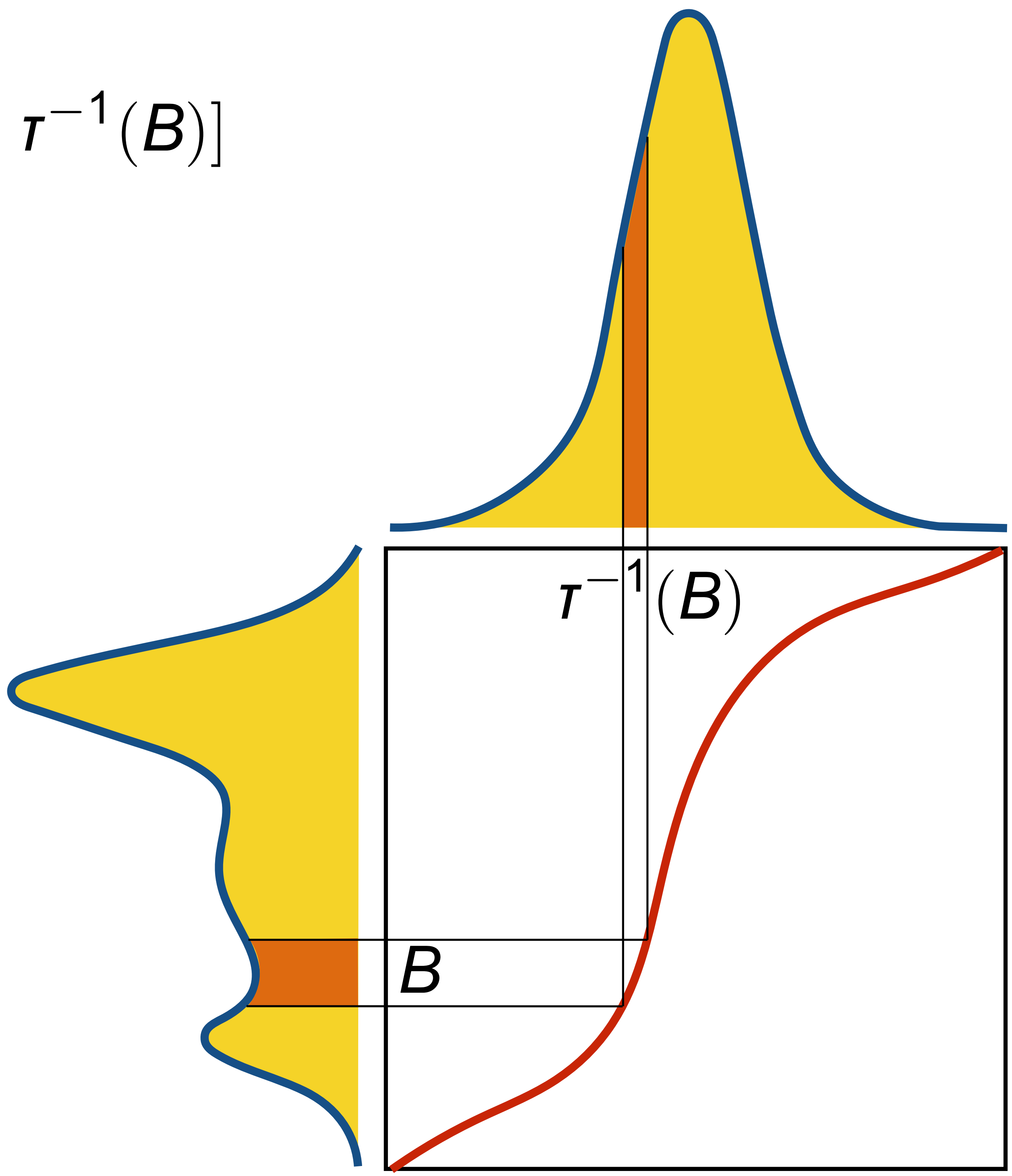
Pushforward Distribution

$$\mathbb{P}_2[\xi_2 \in B] = \mathbb{P}_1[\tau(\xi_1) \in B]$$



Pushforward Distribution

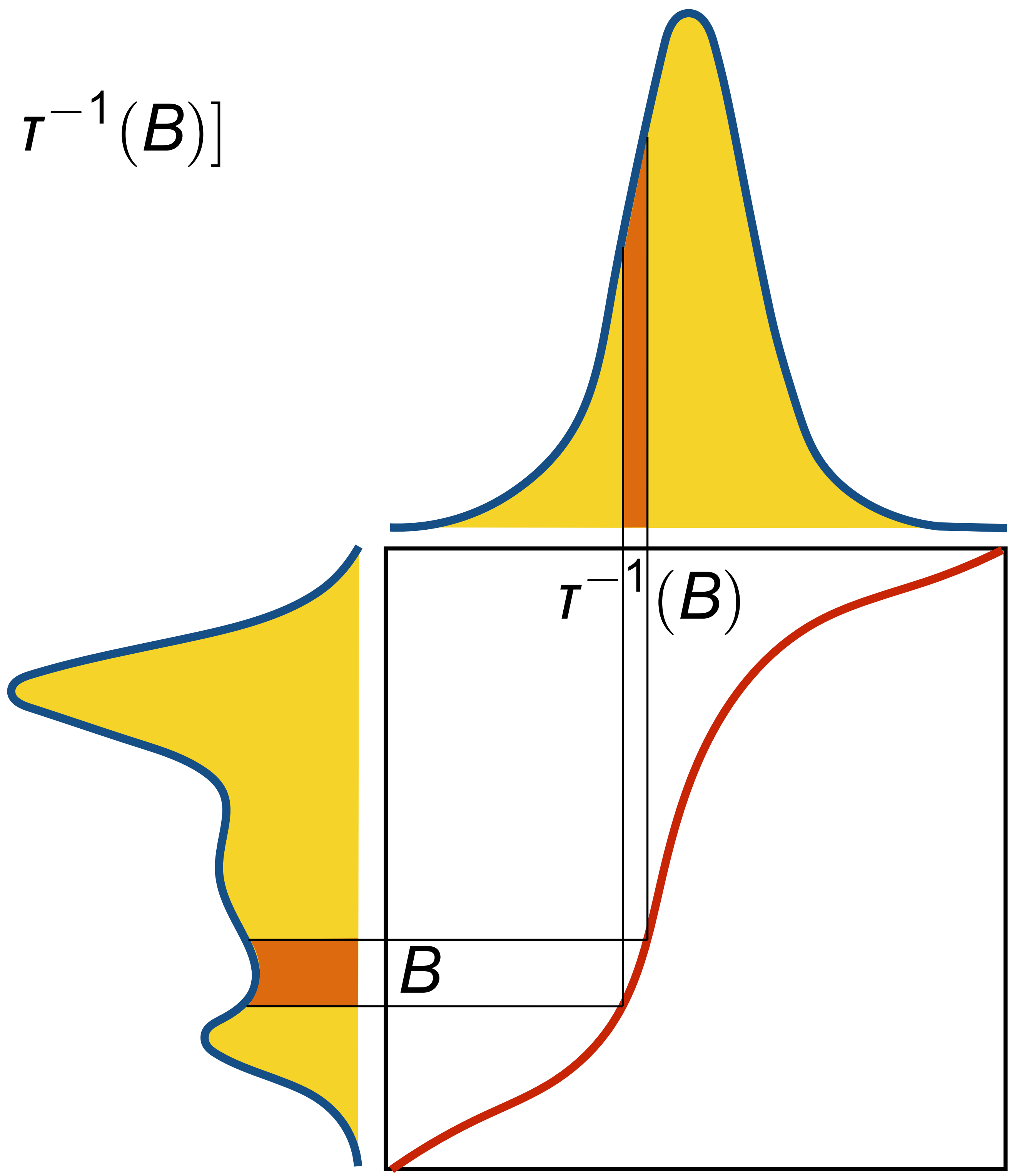
$$\mathbb{P}_2[\xi_2 \in B] = \mathbb{P}_1[\xi_1 \in \tau^{-1}(B)]$$



Pushforward Distribution

$$\mathbb{P}_2[\xi_2 \in B] = \mathbb{P}_1[\xi_1 \in \tau^{-1}(B)]$$

$$\implies \mathbb{P}_2 = \tau_\# \mathbb{P}_1$$

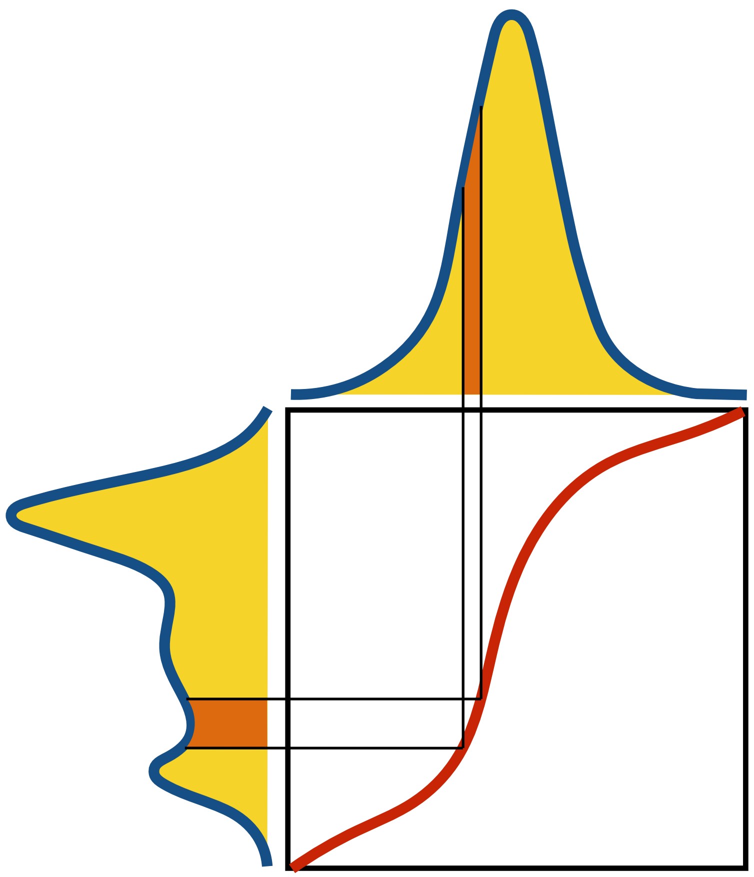


Monge's OT Problem

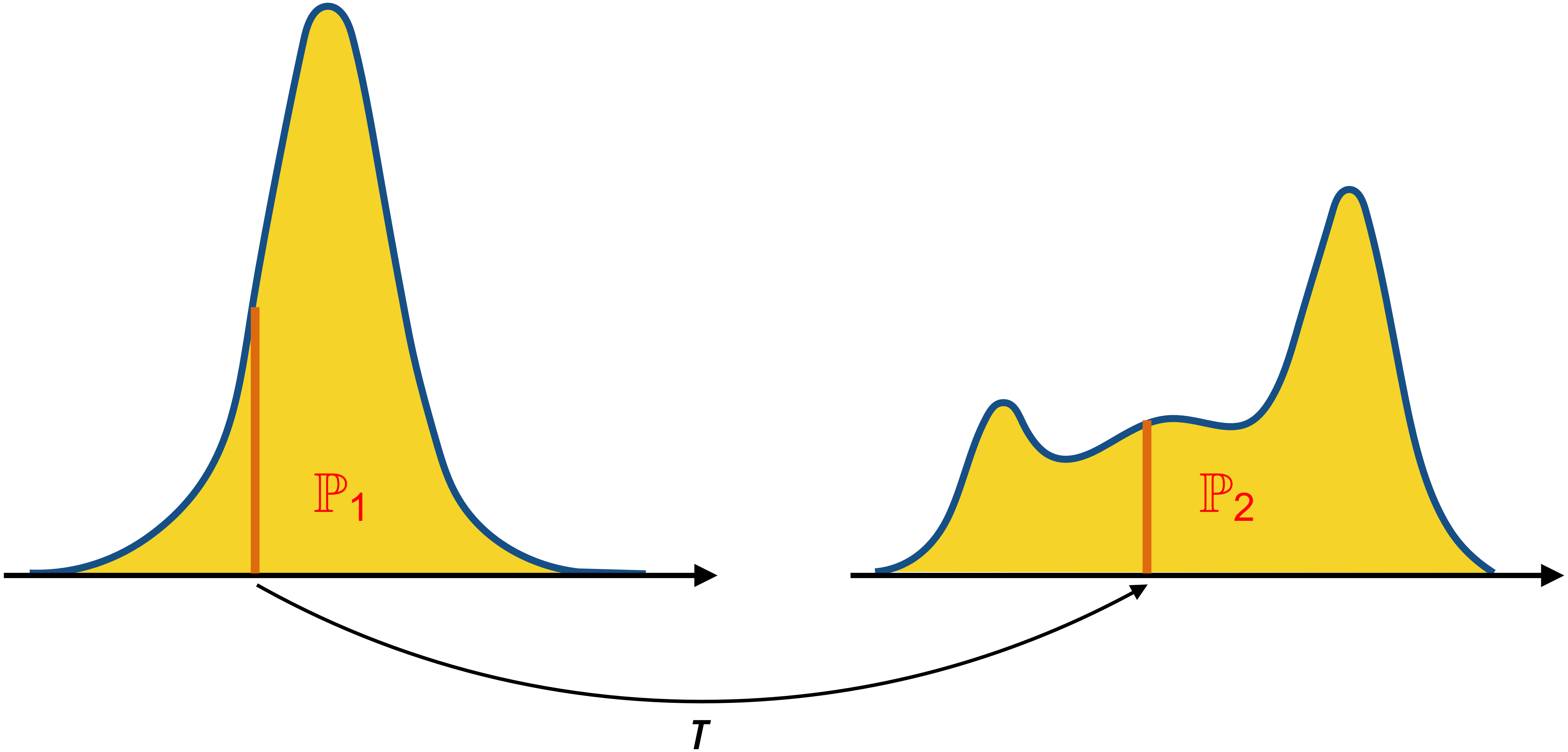
$$\inf_{T} \int c(\xi_1, T(\xi_1))\, \mathrm{d}\mathbb{P}_1(\xi_1)$$

$$\text{s.t.} \quad T_{\#}\mathbb{P}_1 = \mathbb{P}_2$$

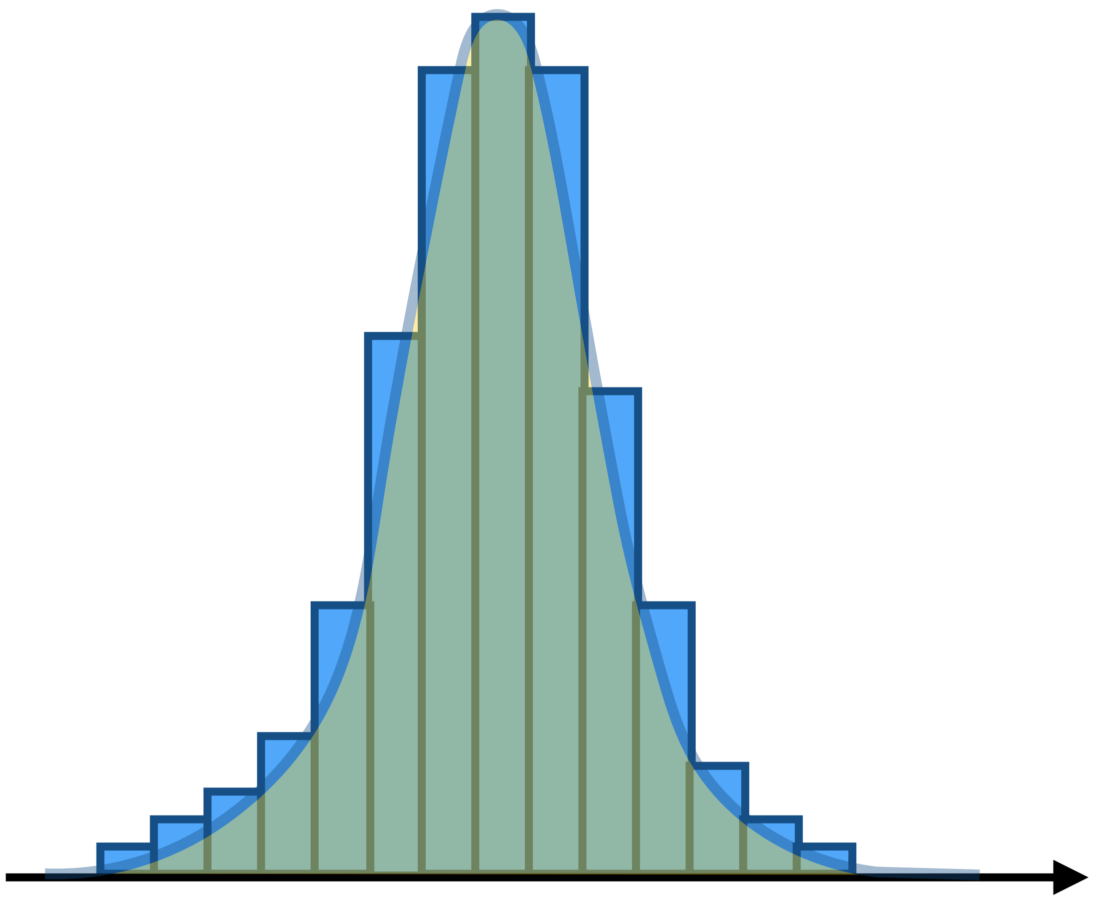
Gaspard Monge
1784

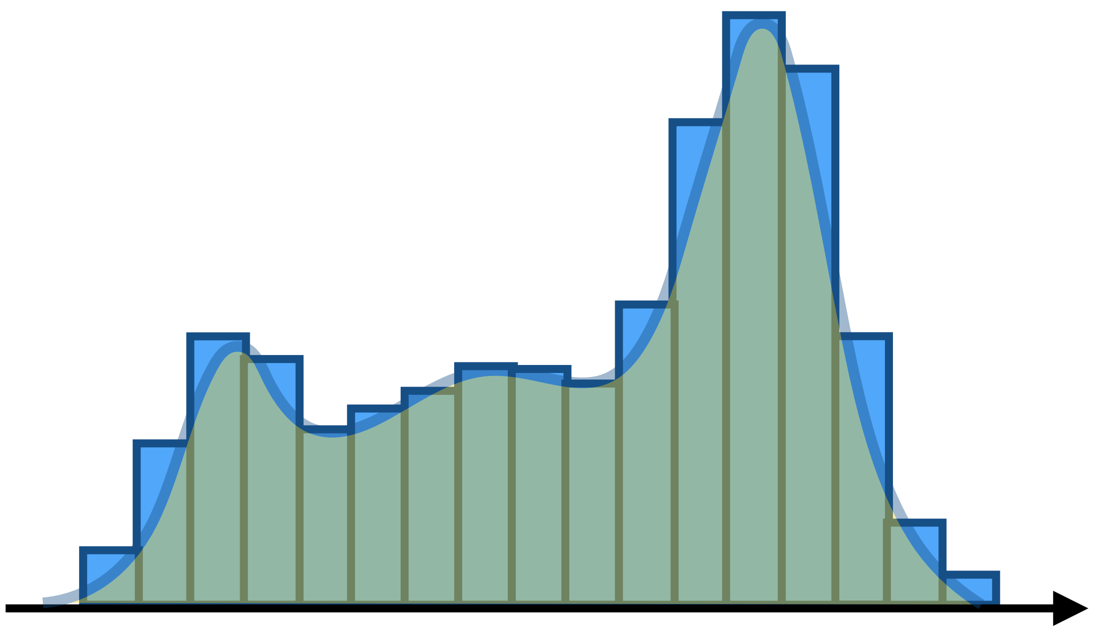


Transportation Maps vs Transportation Plans

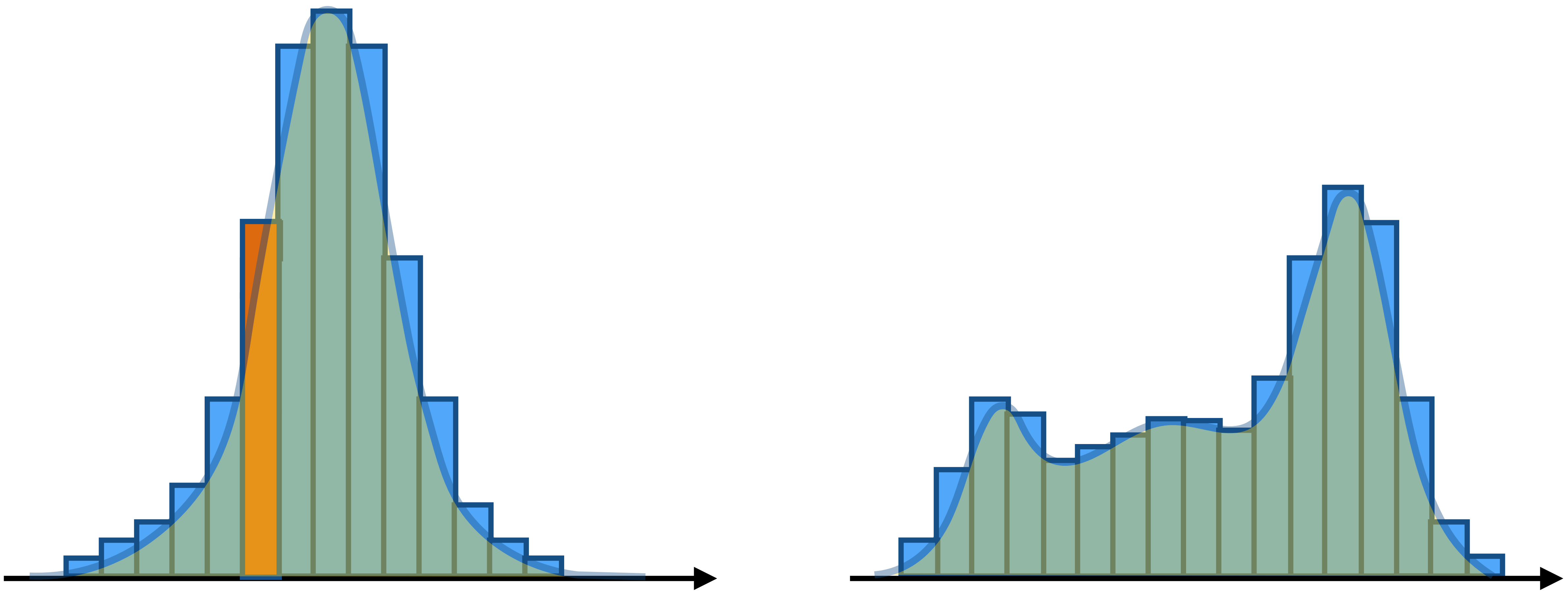


Transportation Maps vs Transportation Plans

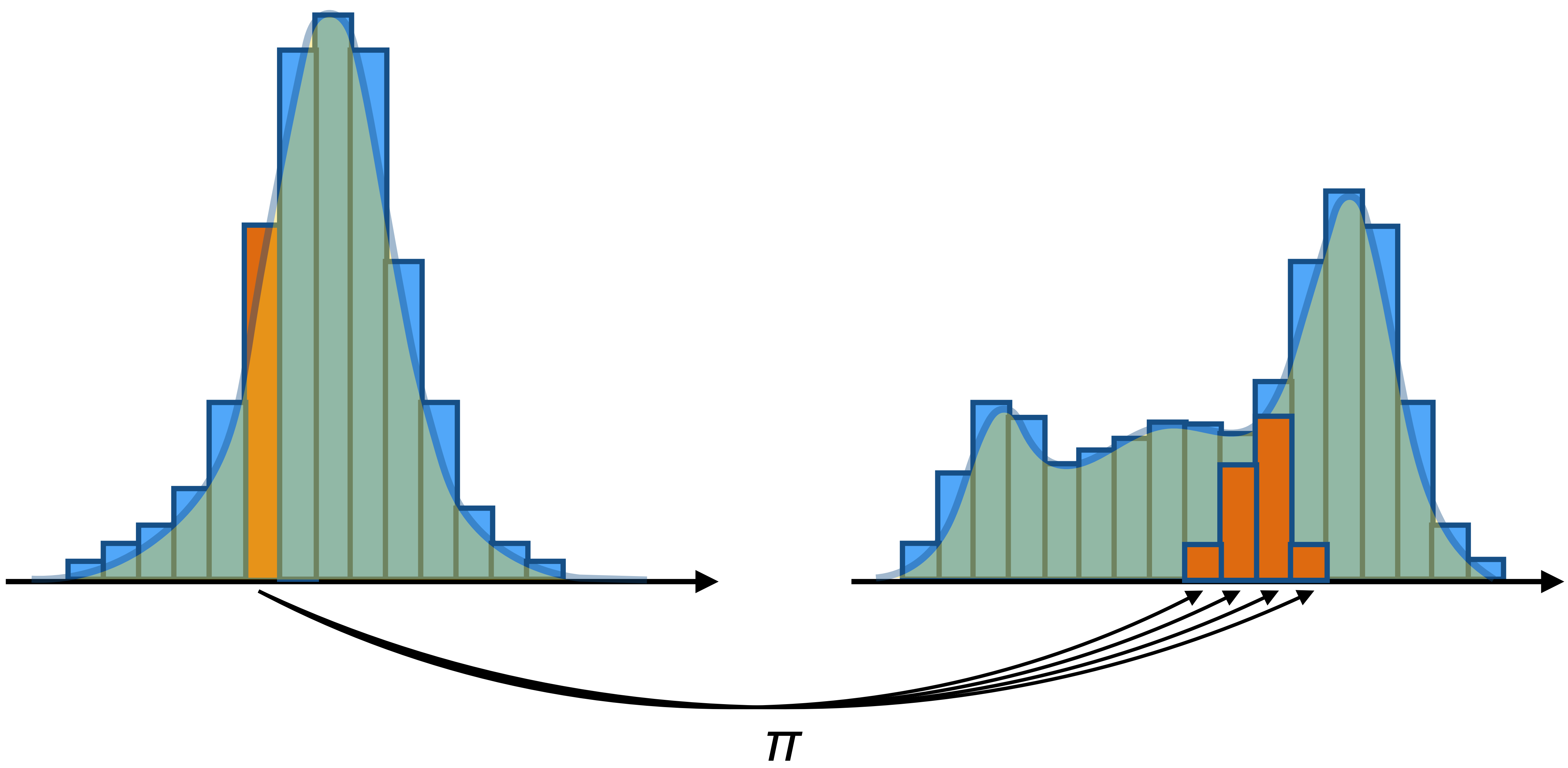




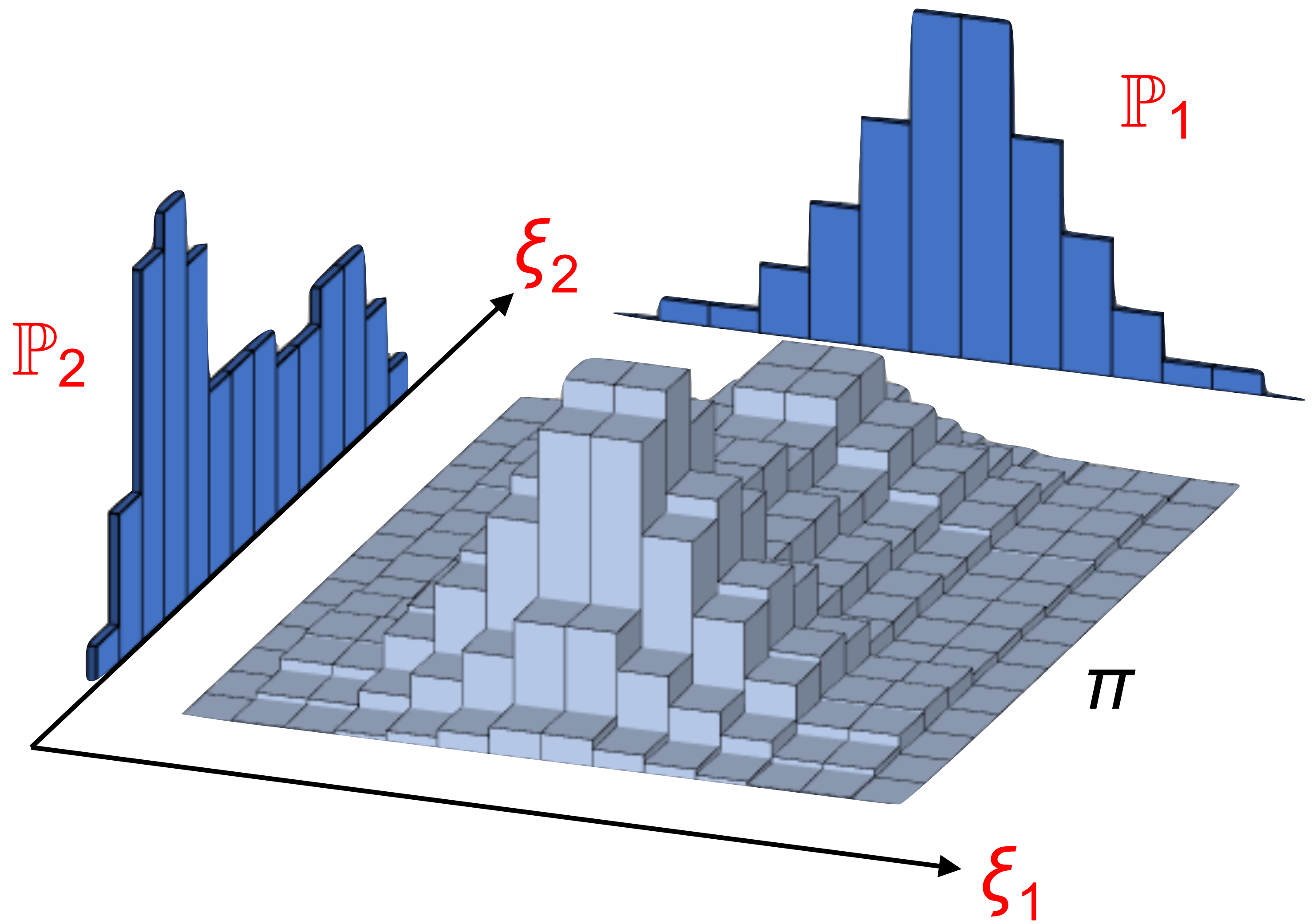
Transportation Maps vs Transportation Plans



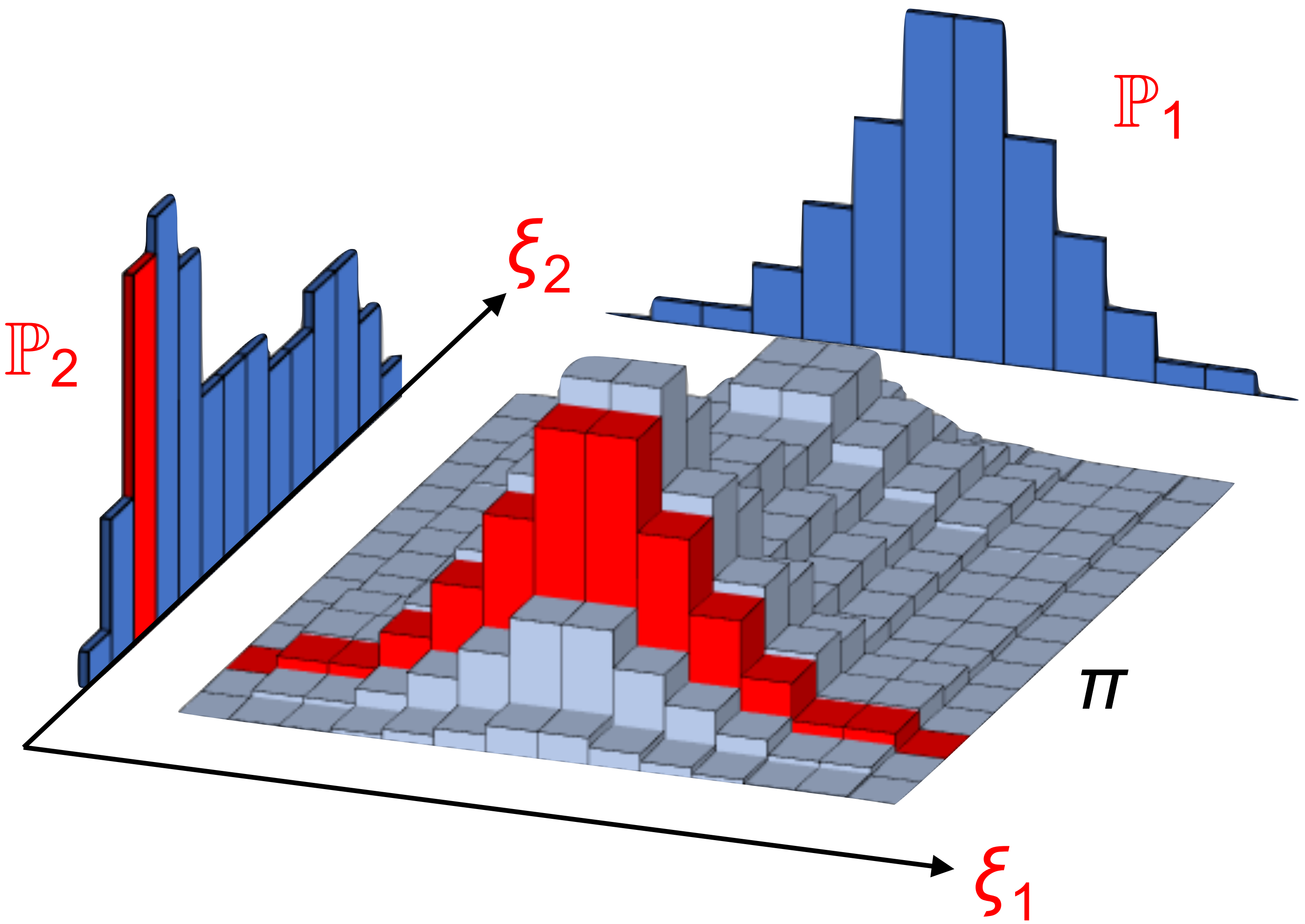
Transportation Maps vs Transportation Plans



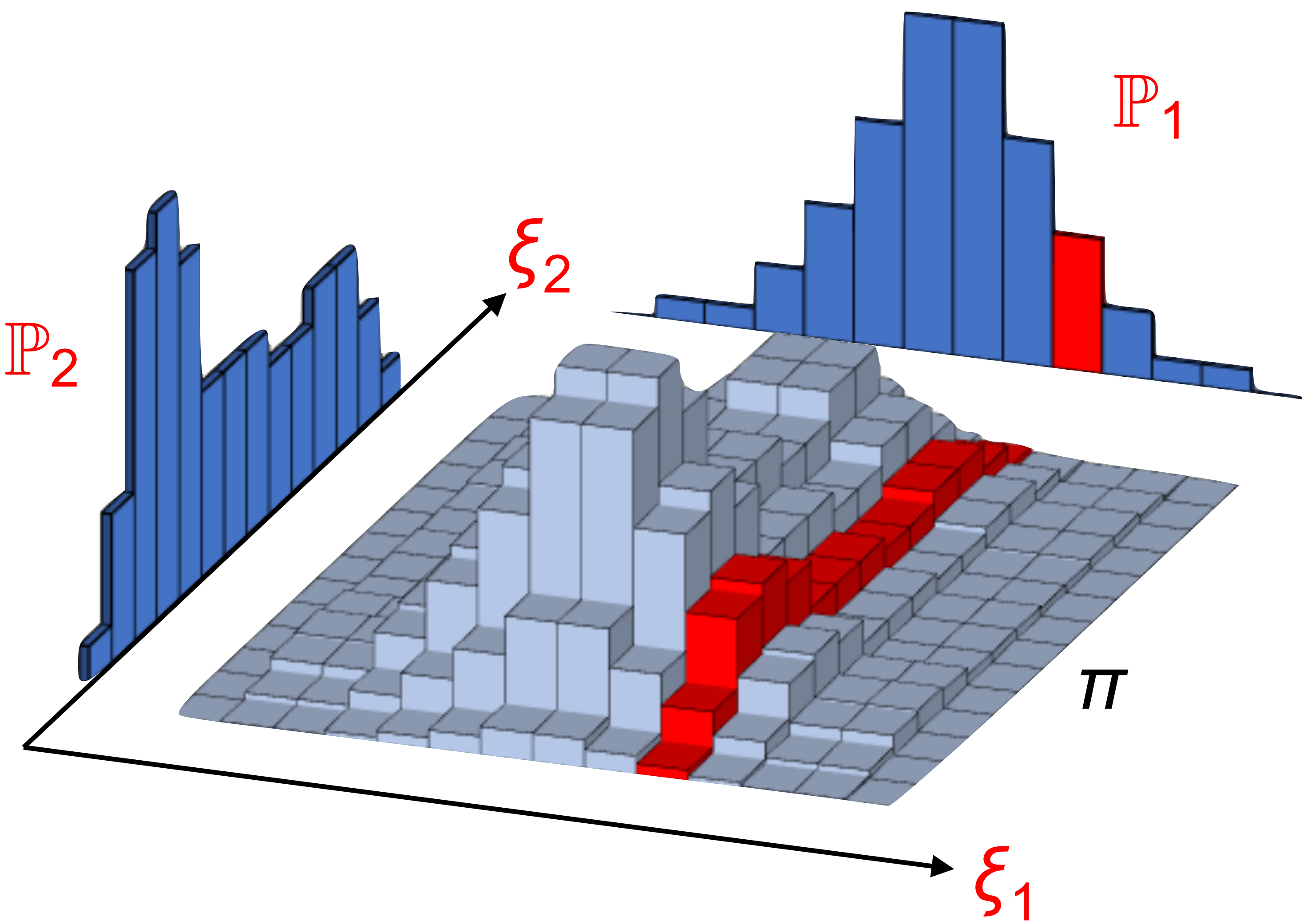
Transportation Maps vs Transportation Plans



Transportation Maps vs Transportation Plans

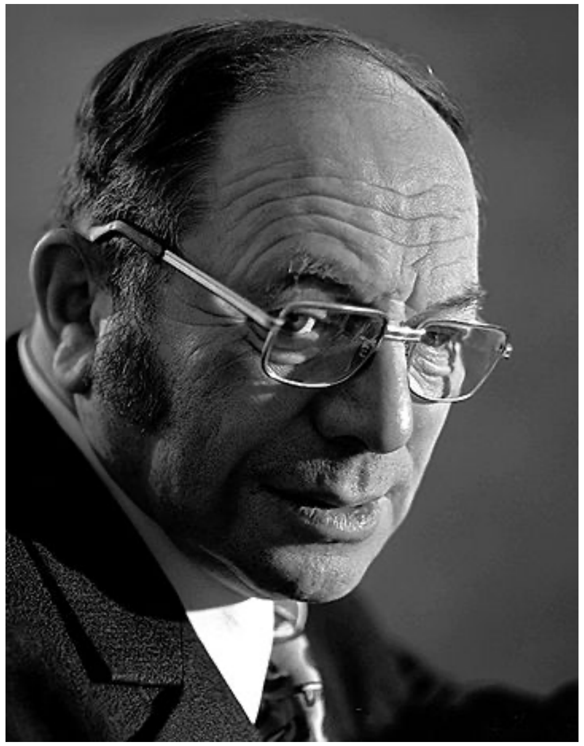


Transportation Maps vs Transportation Plans

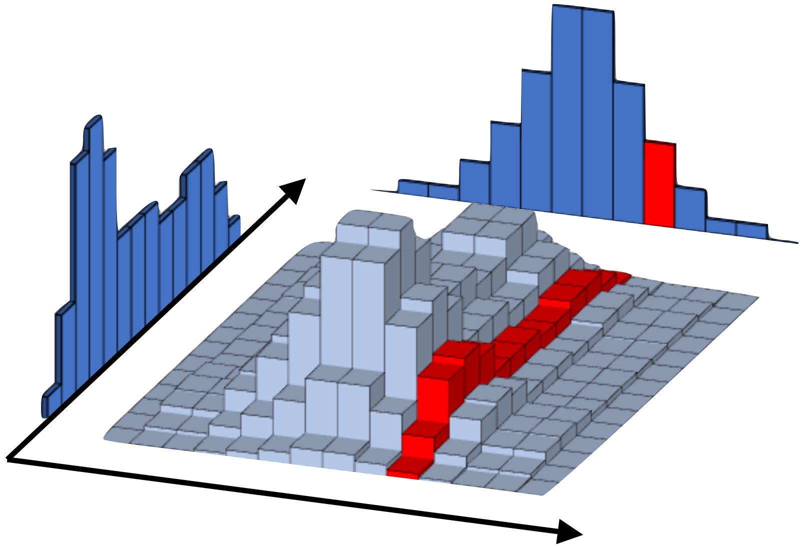


Kantorovich's OT Problem

$$\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int c(\xi_1, \xi_2)\, \mathrm{d}\pi(\xi_1, \xi_2)$$



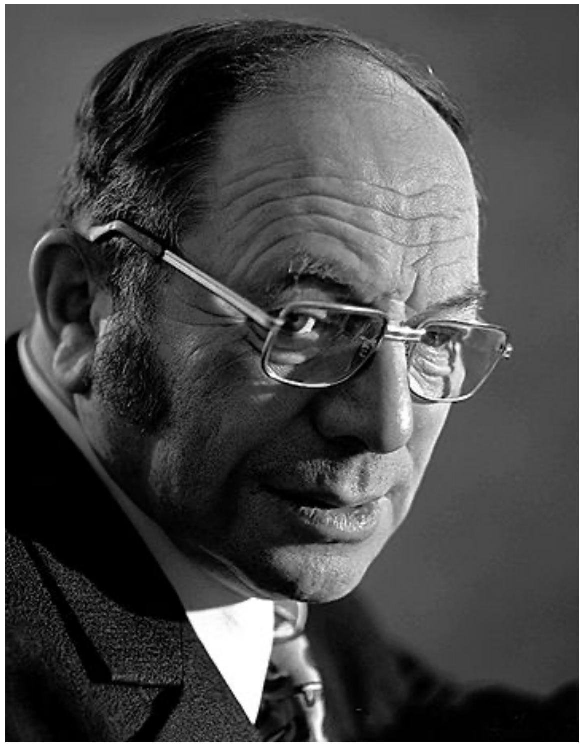
Leonid Kantorovich
1942



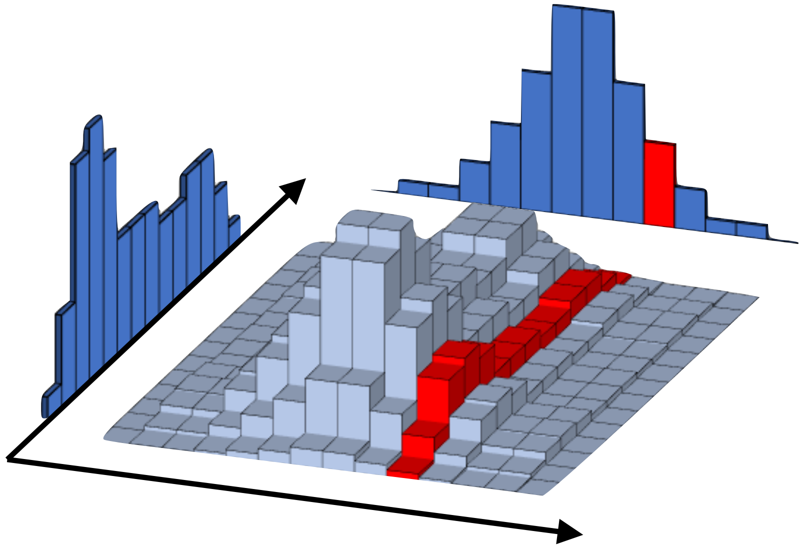
Kantorovich's OT Problem

$$\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int c(\xi_1, \xi_2)\, \mathrm{d}\pi(\xi_1, \xi_2)$$

$\Longrightarrow$ linear program



Leonid Kantorovich
1942



Wasserstein Distances

For $p \geq 1$, set

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \left(\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int d(\xi_1, \xi_2)^p \, \mathrm{d}\pi(\xi_1, \xi_2) \right)^{\frac{1}{p}}$$

Leonid Vaseršteĭn
1969

Wasserstein Distances

For $p \geq 1$, set

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \left(\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int \boxed{d(\xi_1, \xi_2)}^p \, \mathrm{d}\pi(\xi_1, \xi_2) \right)^{\frac{1}{p}}$$

metric on ξ-space

Leonid Vaseršteĭn
1969

Wasserstein Distances

For $p \geq 1$, set

$$W_p(\mathbb{P}_1, \mathbb{P}_2) = \left(\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int \boxed{d(\xi_1, \xi_2)}^p \, \mathrm{d}\pi(\xi_1, \xi_2) \right)^{\frac{1}{p}}$$

metric on ξ-space
usually $d(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|$

Leonid Vaseršteĭn
1969

Wasserstein Distances

For $p \geq 1$, set

$$\boxed{W_p(\mathbb{P}_1, \mathbb{P}_2)} = \left(\min_{\pi \in \Pi(\mathbb{P}_1, \mathbb{P}_2)} \int d(\xi_1, \xi_2)^p \, \mathrm{d}\pi(\xi_1, \xi_2) \right)^{\frac{1}{p}}$$

metric on $\mathbb{P}$-space

Leonid Vaseršteĭn
1969

Computing Wasserstein Distances

Theorem: Computing $W_p(\mathbb{P}_1, \mathbb{P}_2)$ is #P-hard even if $\mathbb{P}_1 \sim \mathcal{U}[0,1]^d$ and $\mathbb{P}_2$ is a two-point distribution.[1)]

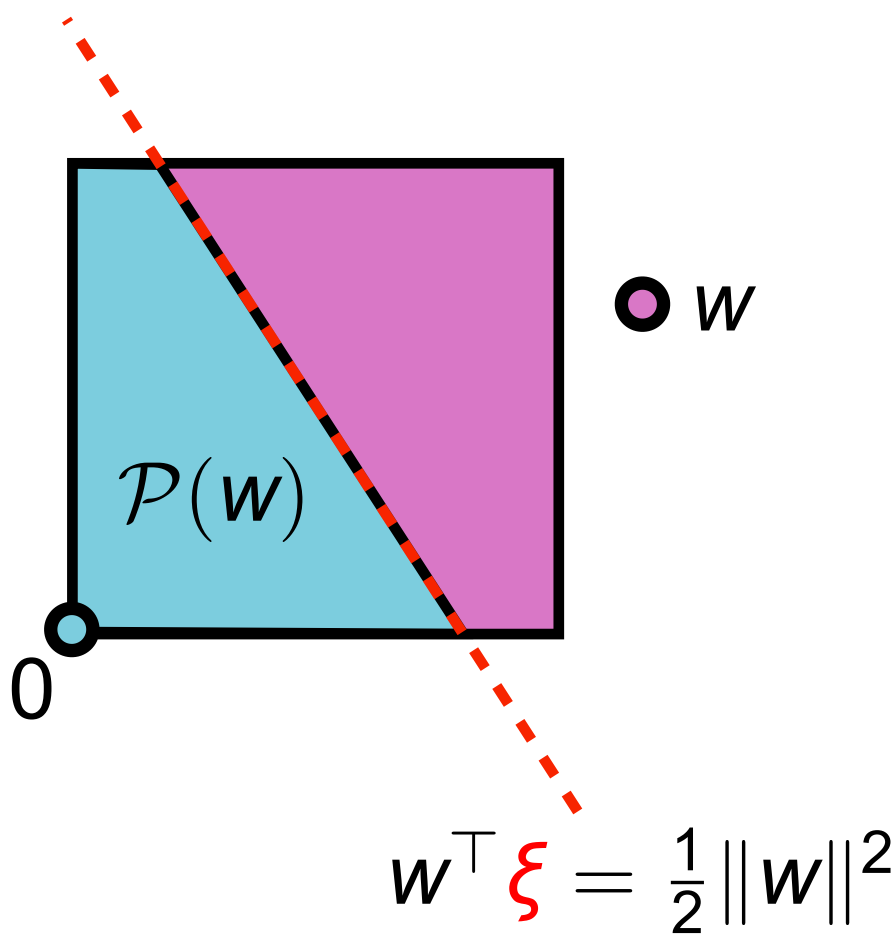
---

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Computing Wasserstein Distances

Theorem: Computing $W_p(\mathbb{P}_1, \mathbb{P}_2)$ is #P-hard even if $\mathbb{P}_1 \sim \mathcal{U}[0,1]^d$ and $\mathbb{P}_2$ is a two-point distribution.[1)]

Proof: Computing the volume of the knapsack polytope $\mathcal{P}(w)$ is #P-hard.[2)]



$$\mathcal{P}(w) = \left\{ \xi \in [0,1]^d : w^\top \xi \leq \tfrac{1}{2}\|w\|^2 \right\}$$

---

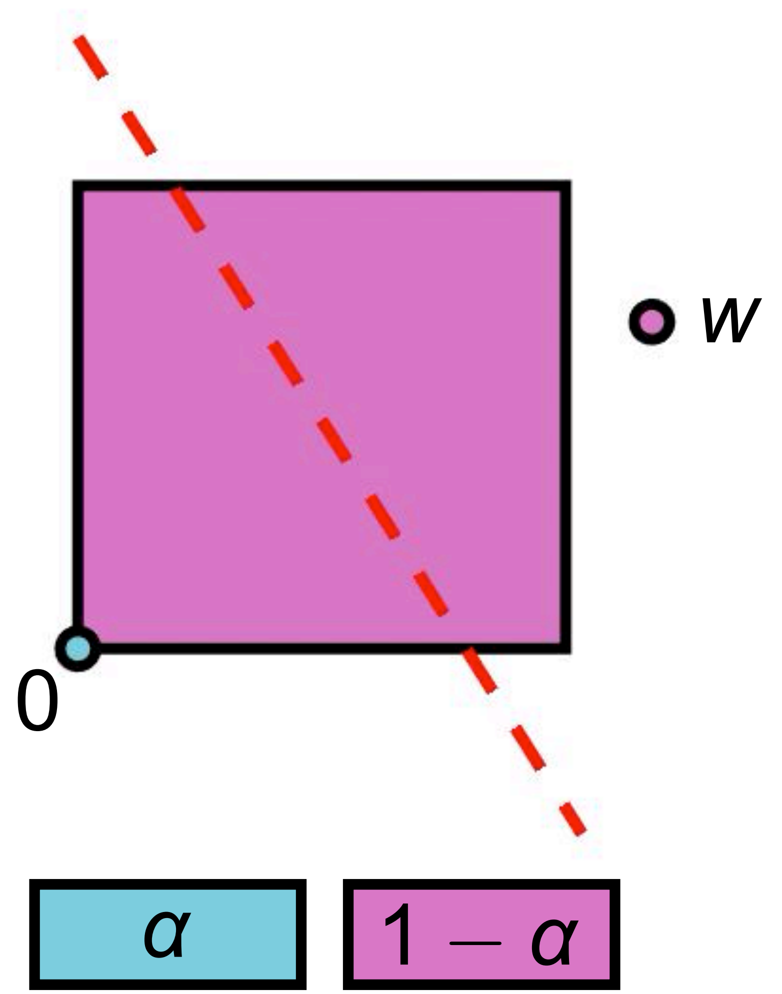
1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

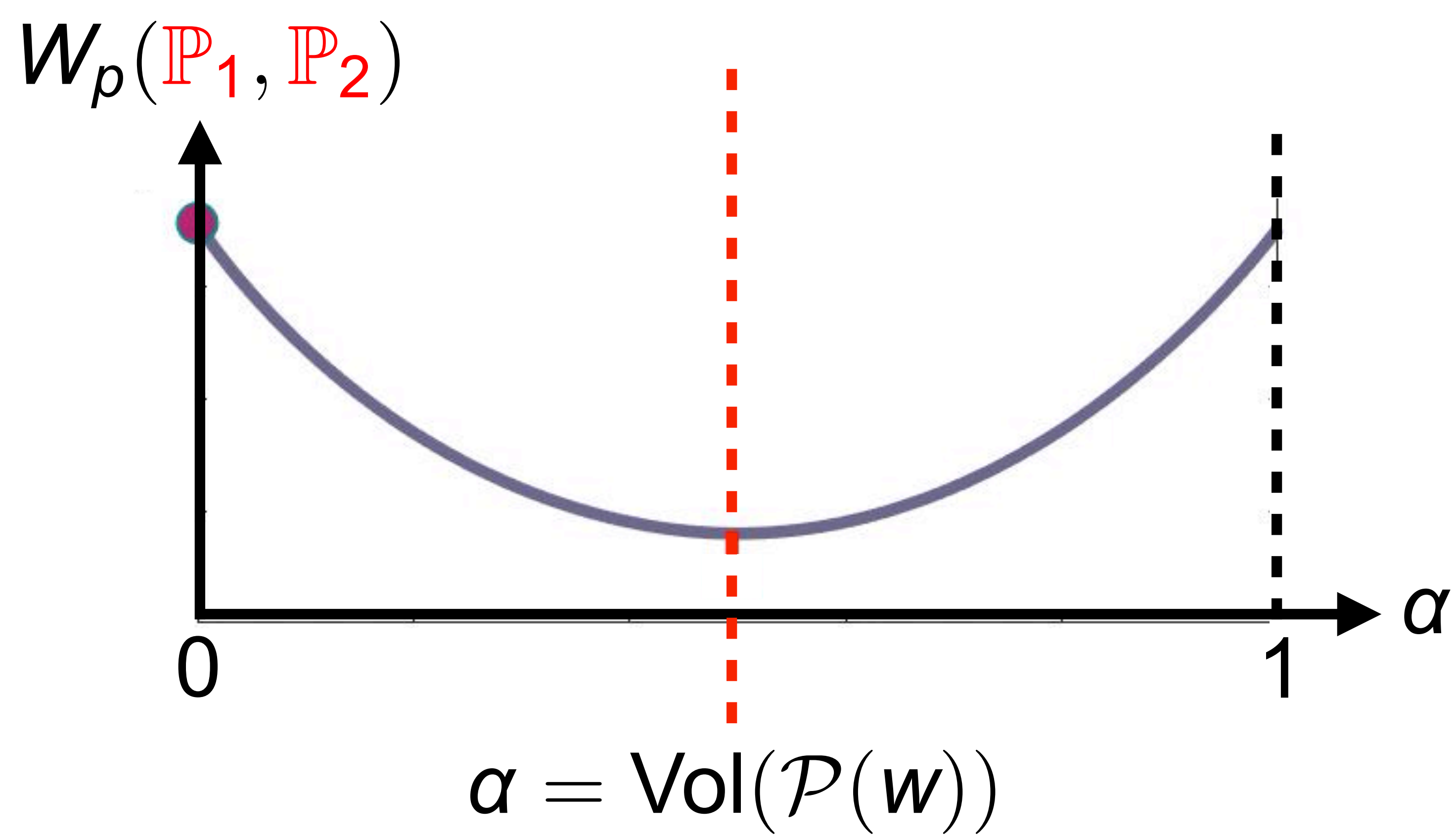
2) Dyer & Frieze, *SIAM J. Comp.*, 1988.

Computing Wasserstein Distances

Theorem: Computing $W_p(\mathbb{P}_1, \mathbb{P}_2)$ is #P-hard even if $\mathbb{P}_1 \sim \mathcal{U}[0,1]^d$ and $\mathbb{P}_2$ is a two-point distribution.[1)]

Proof: $\mathbb{P}_2 = \alpha \cdot \delta_0 + (1-\alpha) \cdot \delta_w$





1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Brenier's Theorem

Theorem (Brenier 1987): If $c(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|_2^2$, then

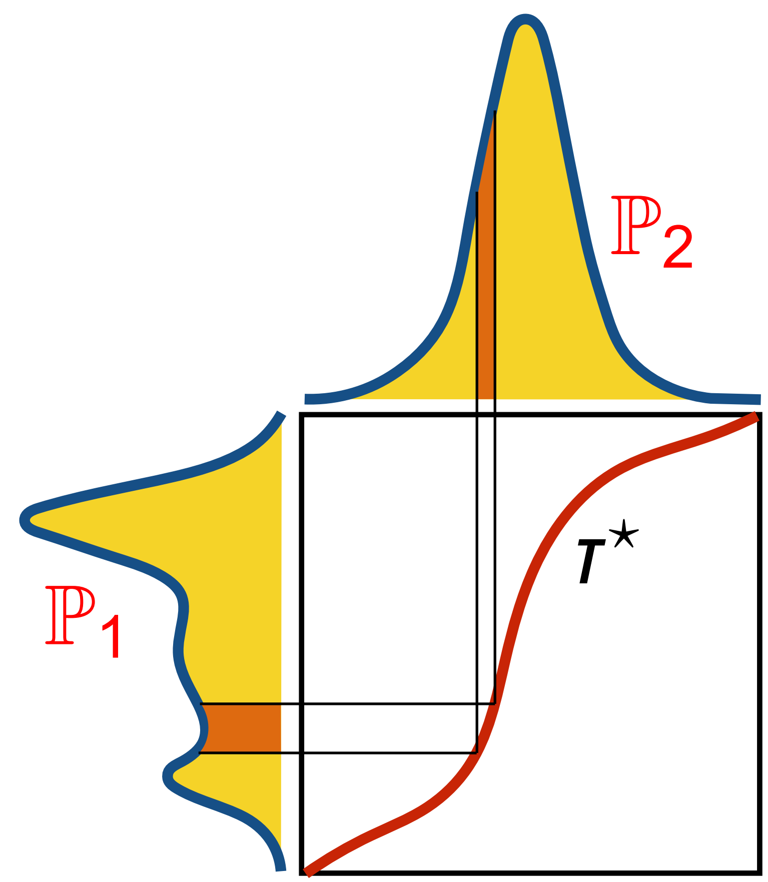
- $T^\star$ solves the Monge problem iff
- $T^\star$ is feasible and $\exists \varphi$ convex such that $T^\star = \nabla \varphi$.

Brenier's Theorem

Theorem (Brenier 1987): If $c(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|_2^2$, then

- $T^\star$ solves the Monge problem iff
- $\boxed{T^\star \text{ is feasible}}$ and $\exists \varphi$ convex such that $T^\star = \nabla \varphi$.

$\mathbb{P}_2 = T^\star_\# \mathbb{P}_1$

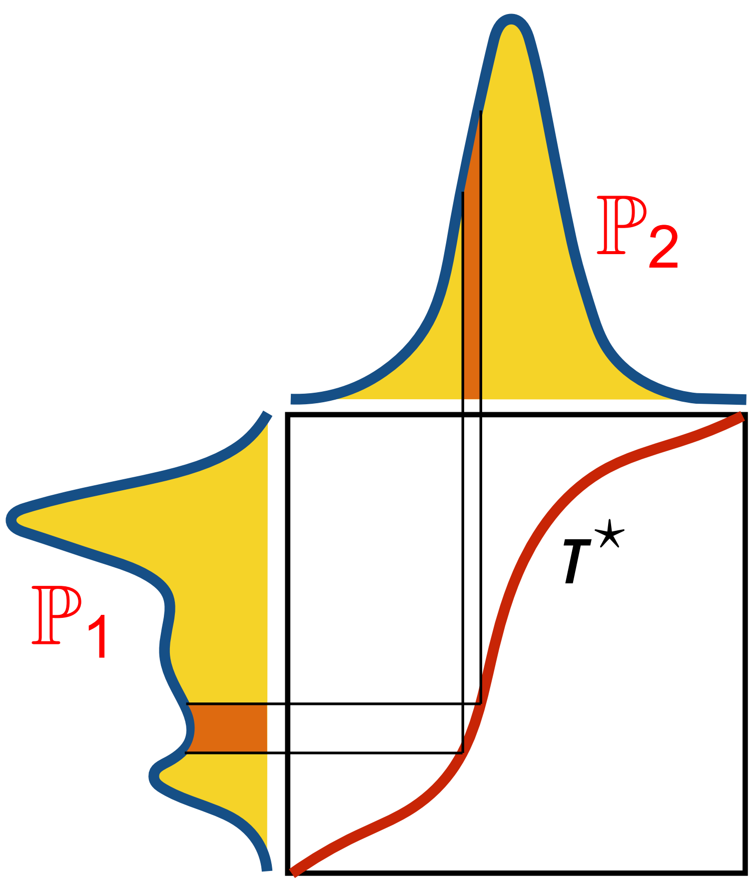


Brenier's Theorem

Theorem (Brenier 1987): If $c(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|_2^2$, then

- $T^\star$ solves the Monge problem iff
- $T^\star$ is feasible and $\boxed{\exists \varphi \text{ convex such that } T^\star = \nabla \varphi.}$

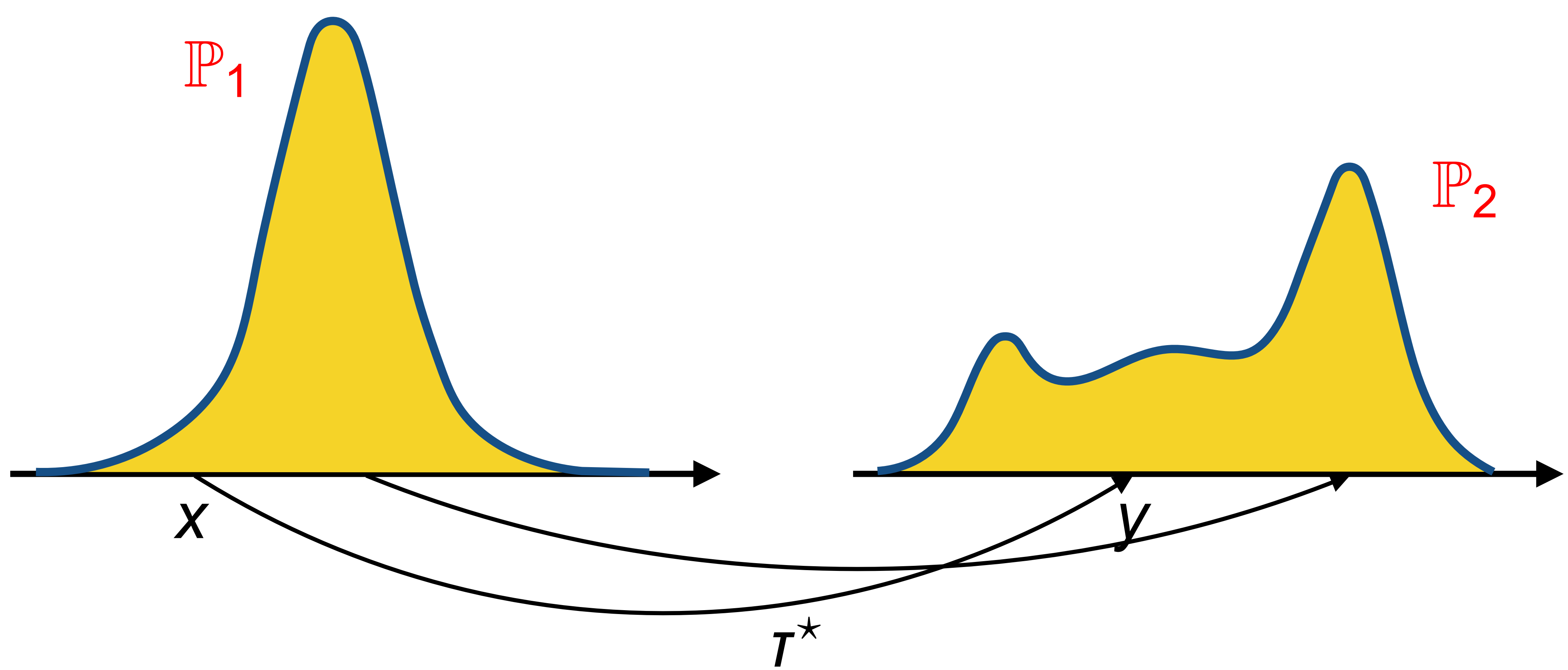
optimality



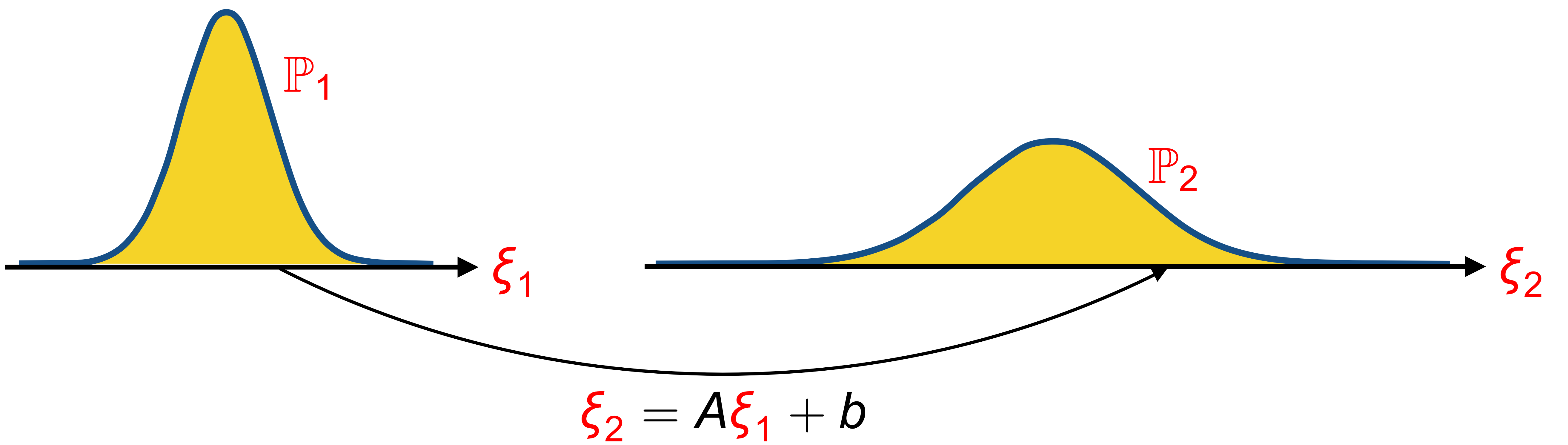
Brenier's Theorem

Theorem (Brenier 1987): If $c(\xi_1, \xi_2) = \|\xi_1 - \xi_2\|_2^2$, then

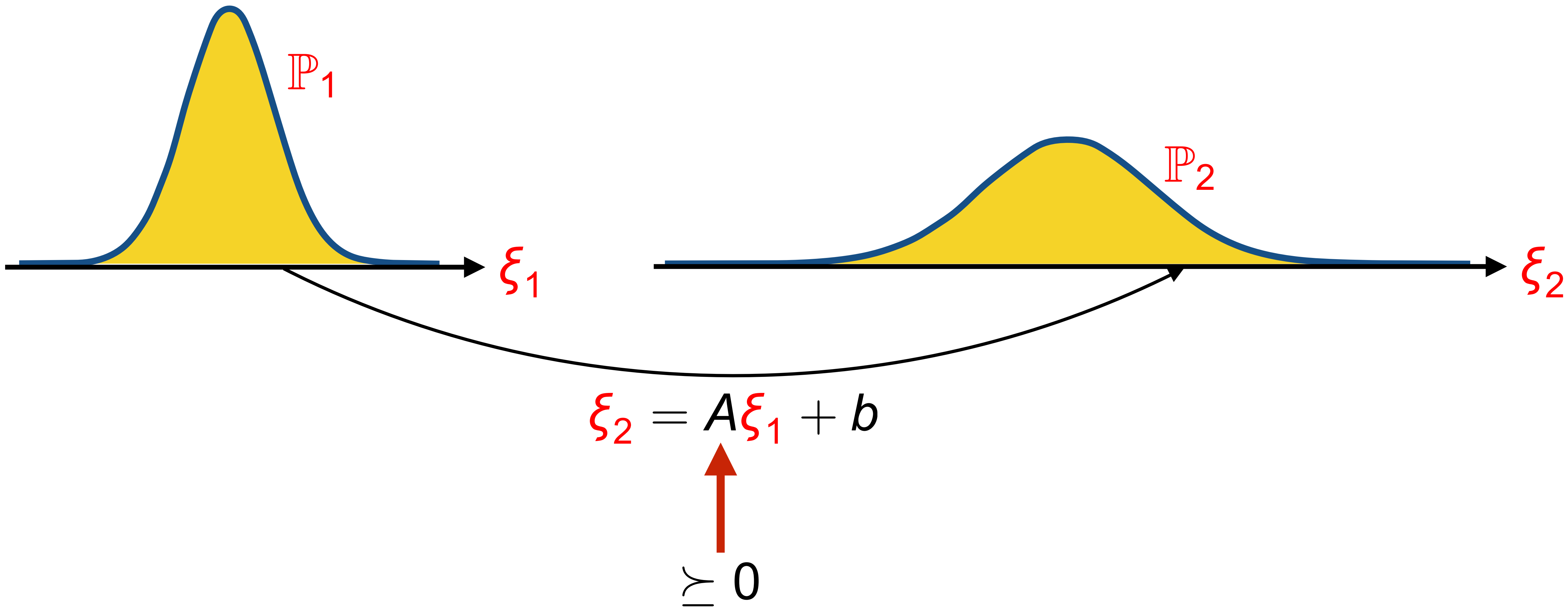
- $T^\star$ solves the Monge problem iff
- $T^\star$ is feasible and $\exists \varphi$ convex such that $T^\star = \nabla\varphi$.



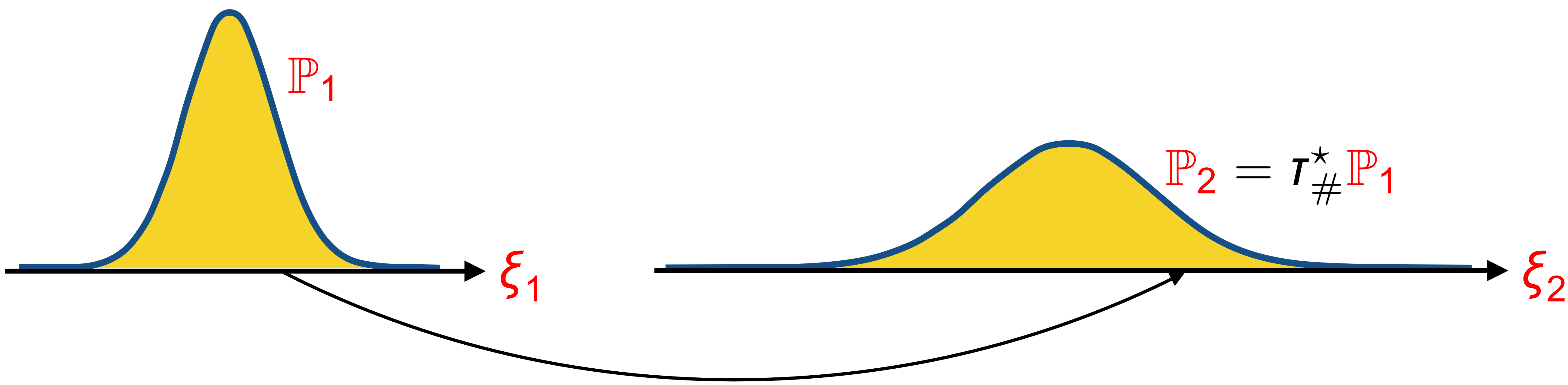
2-Wasserstein Distance in Closed Form



2-Wasserstein Distance in Closed Form

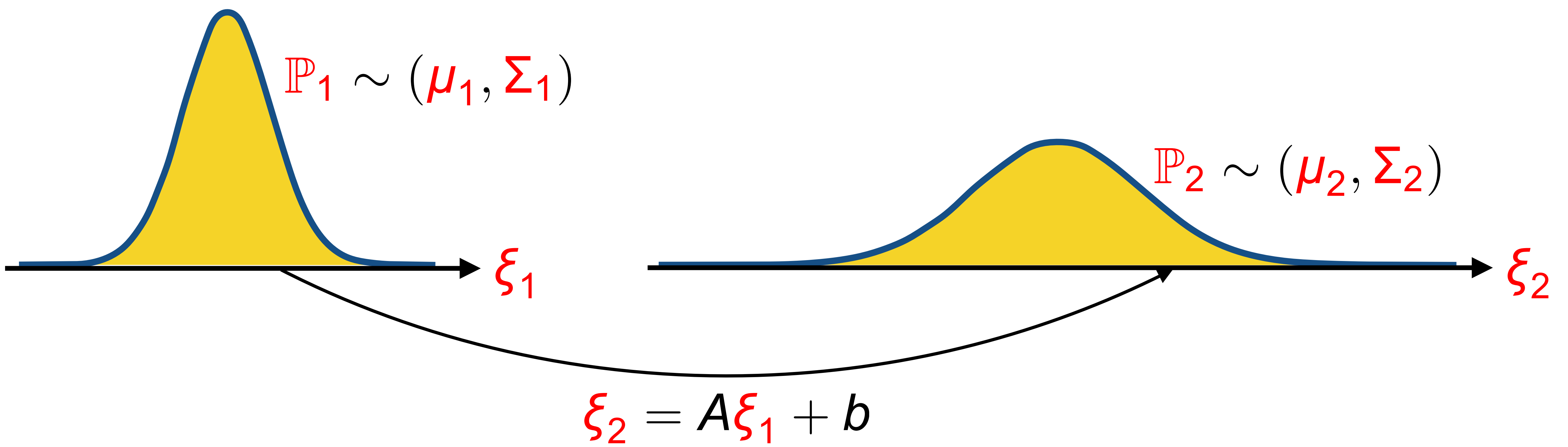


2-Wasserstein Distance in Closed Form

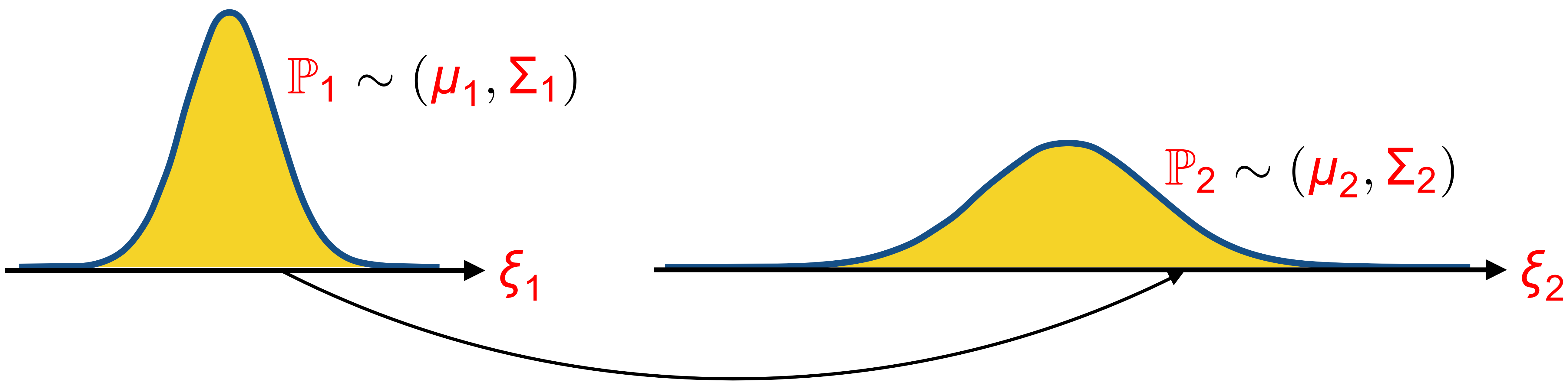


$$\xi_2 = \underbrace{A\xi_1 + b}_{= \tau^\star(\xi_1) \text{ by Brenier!}}$$

2-Wasserstein Distance in Closed Form



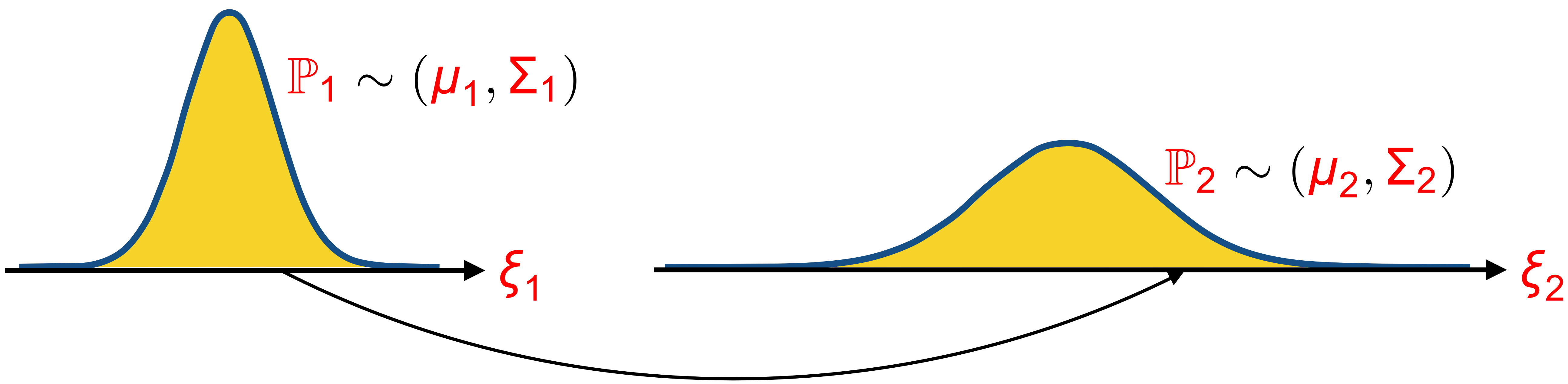
2-Wasserstein Distance in Closed Form



$$\xi_2 = A\xi_1 + b$$

$$\Sigma_1^{-\frac{1}{2}} \left(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}} \right)^{\frac{1}{2}} \Sigma_1^{-\frac{1}{2}}$$

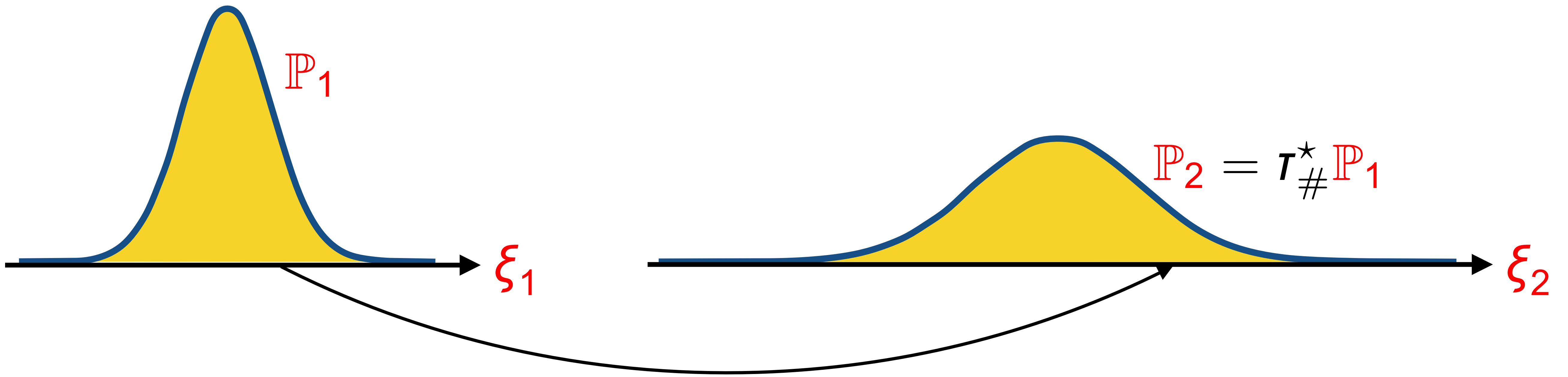
2-Wasserstein Distance in Closed Form



$$\xi_2 = A\xi_1 + b$$

$$\mu_2 - A\mu_1$$

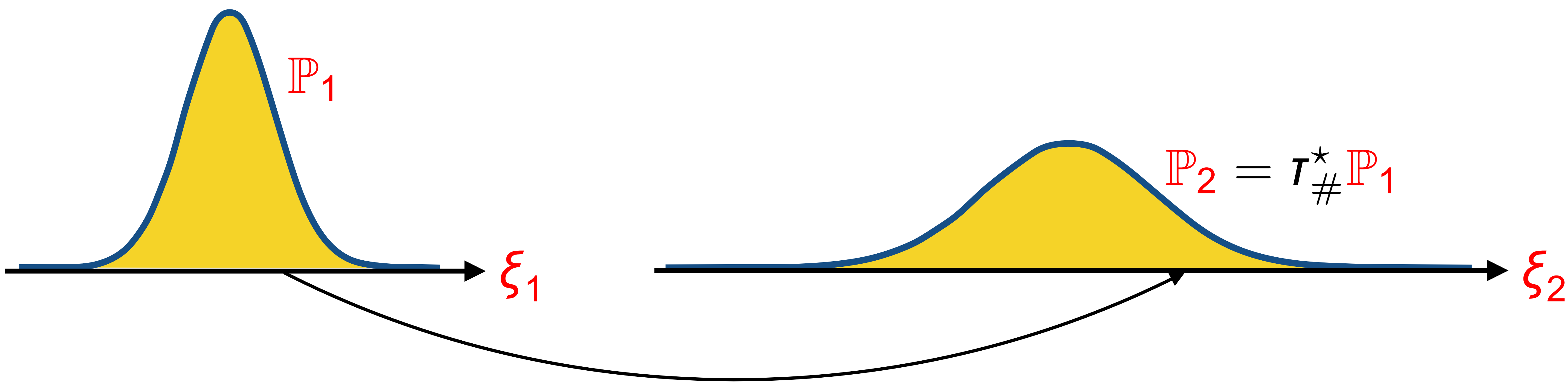
2-Wasserstein Distance in Closed Form



$$\xi_2 = \underbrace{A\xi_1 + b}_{= \tau^\star(\xi_1)}$$

$$\Longrightarrow \quad W_2^2(\mathbb{P}_1, \mathbb{P}_2) = \int \|\xi_1 - \tau^\star(\xi_1)\|_2^2 \, \mathrm{d}\mathbb{P}_1(\xi_1)$$

2-Wasserstein Distance in Closed Form



$$\xi_2 = \underbrace{A\xi_1 + b}_{= T^\star(\xi_1)}$$

$$\Longrightarrow \quad W_2^2(\mathbb{P}_1, \mathbb{P}_2) = \|\mu_1 - \mu_2\|_2^2 + \mathrm{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}} \Sigma_2 \Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]$$

2-Wasserstein Distance in Closed Form

Corollary 1: If $\mathbb{P}_2$ is a psd affine pushforward of $\mathbb{P}_1$, then

$$W_2^2(\mathbb{P}_1, \mathbb{P}_2) = \|\mu_1 - \mu_2\|_2^2 + \operatorname{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]$$

2-Wasserstein Distance in Closed Form

Corollary 2: If $\mathbb{P}_2$ and $\mathbb{P}_1$ are arbitrary distributions, then[1)]

$$W_2^2(\mathbb{P}_1, \mathbb{P}_2) \geq \|\mu_1 - \mu_2\|_2^2 + \mathrm{tr}\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]$$

---

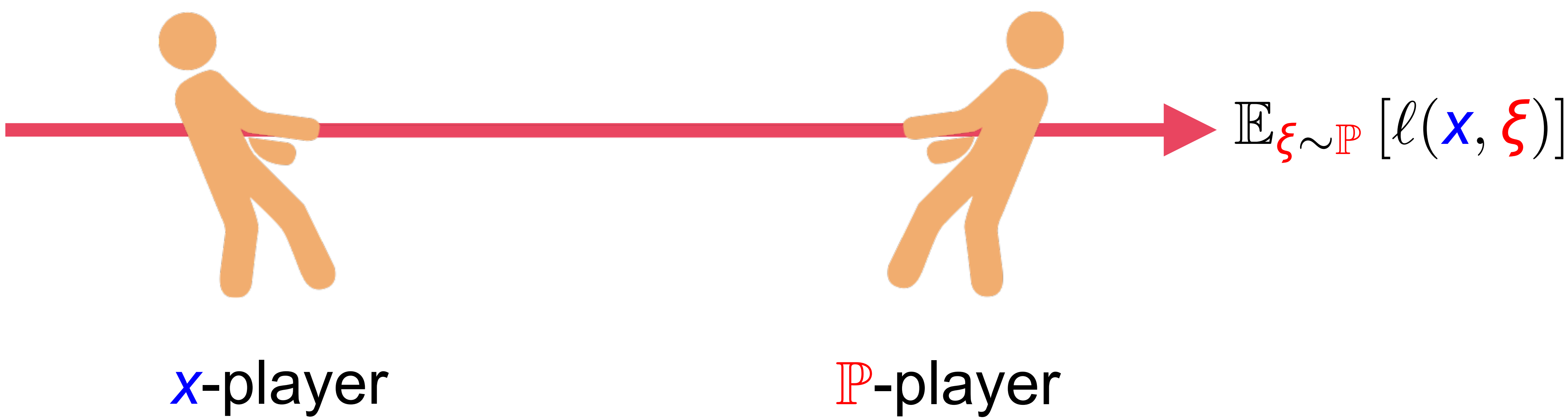
1) Gelbrich, *Mathematische Nachrichten*, 1990.

Distributionally Robust Optimization (DRO)

Distributionally Robust Optimization (DRO)

Zero-sum game against "nature":

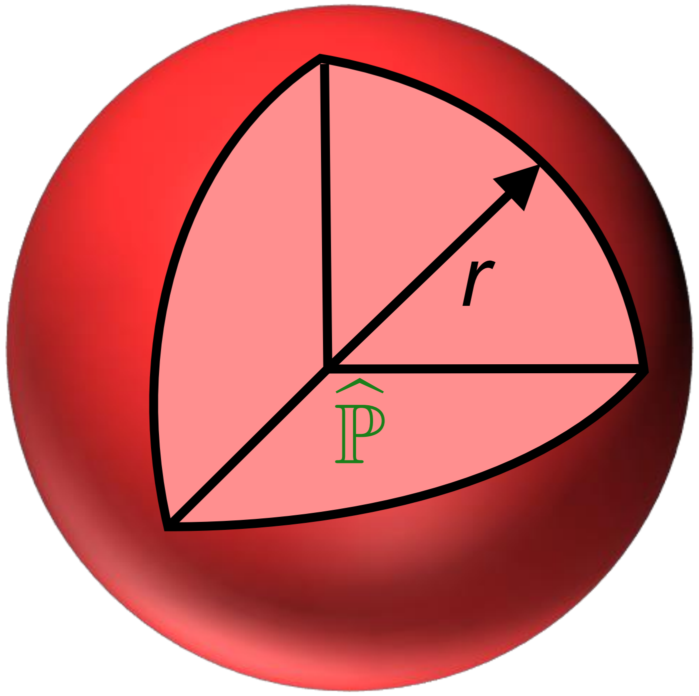
$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]$$



Wasserstein DRO

Zero-sum game against "nature":

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x, \xi) \right]$$

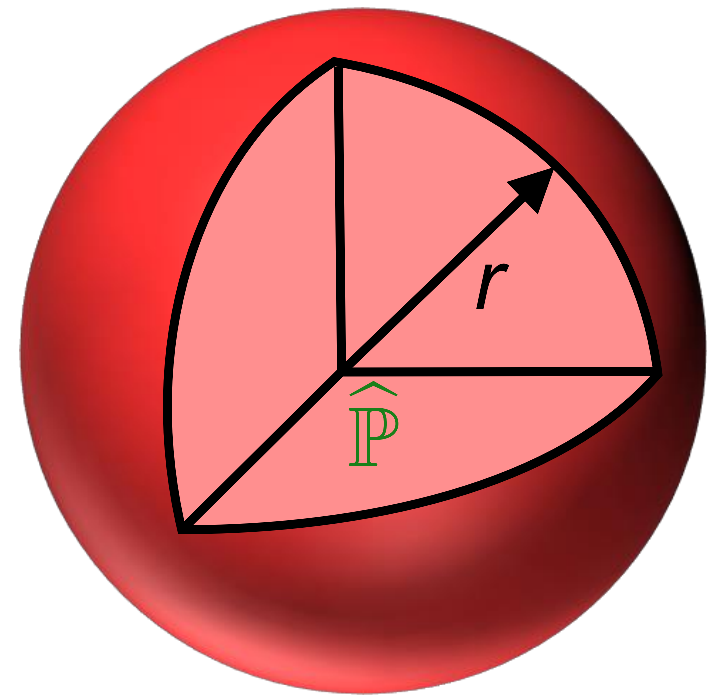


1) Pflug & Wozabal, *Quant. Finance*, 2007; Pflug & Pichler, *Springer*, 2014.

Wasserstein DRO

Zero-sum game against “nature”:

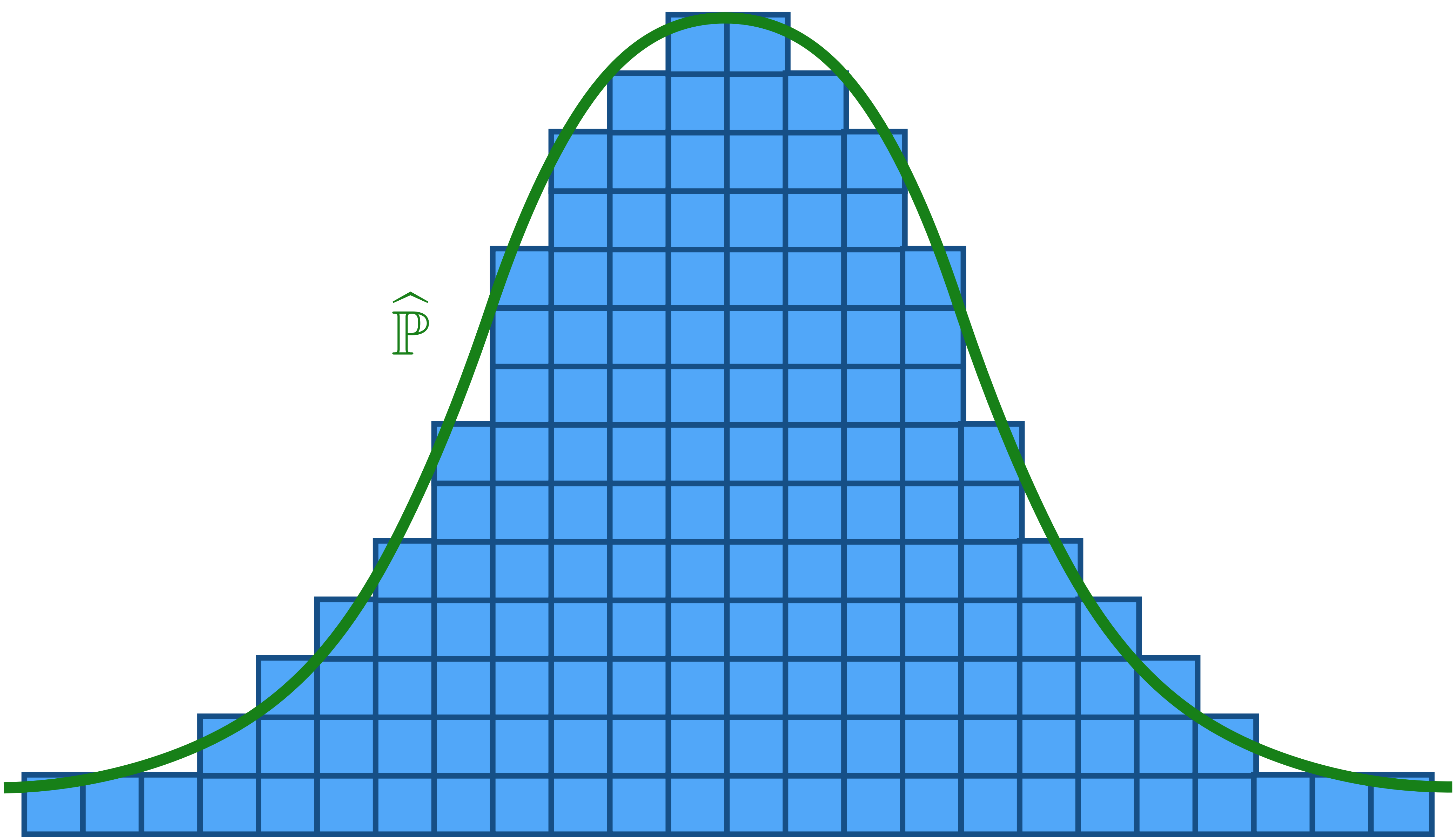
$$\inf_{x\in\mathcal{X}} \ \sup_{\mathbb{P}\in \boxed{\mathbb{B}_r(\widehat{\mathbb{P}})}} \ \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]$$



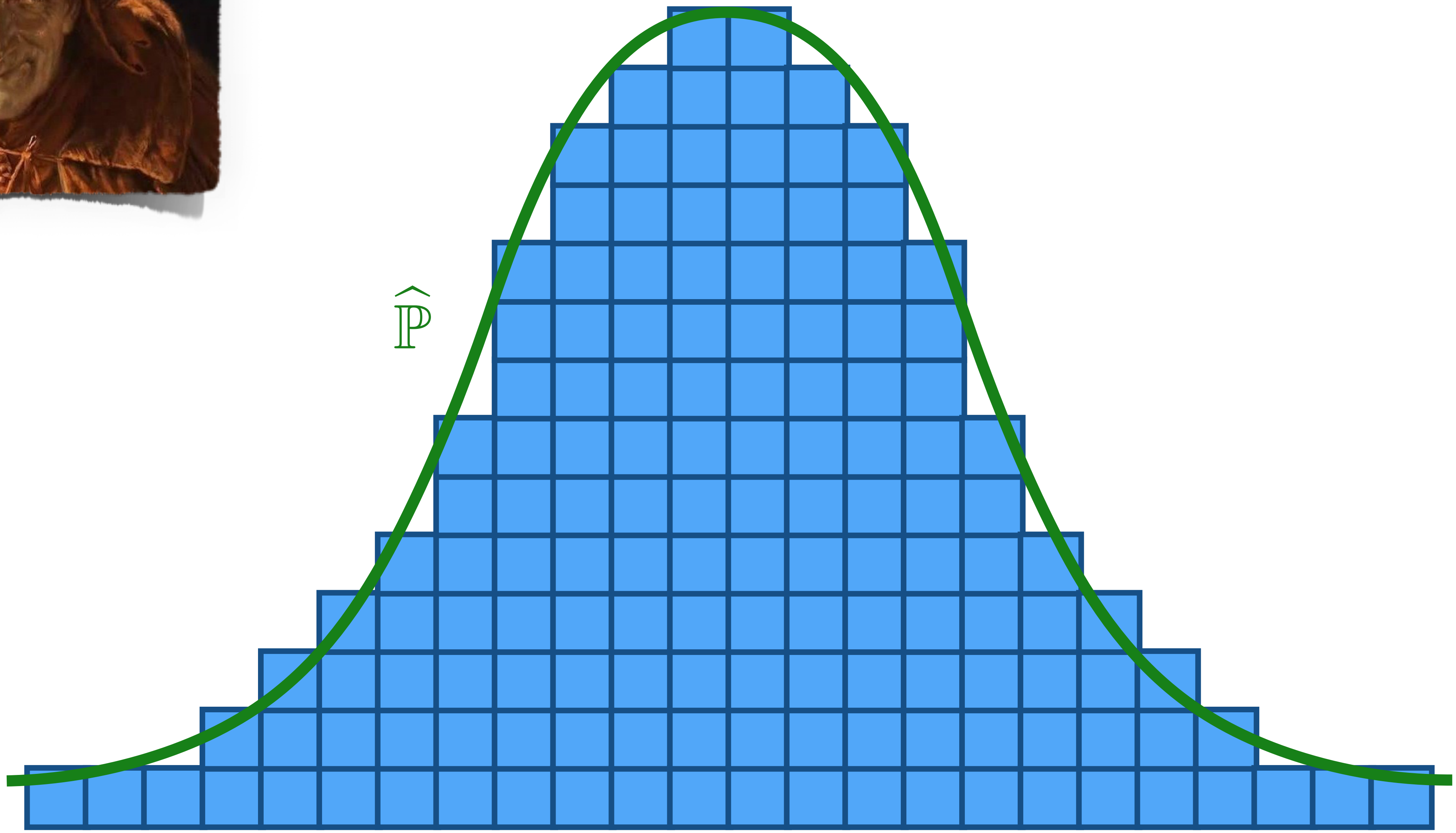
$$\mathbb{B}_r(\widehat{\mathbb{P}}) = \left\{ \mathbb{P} \text{ supported on } \Xi \;\middle|\; W_p(\mathbb{P},\widehat{\mathbb{P}}) \le r \right\}$$

[1] Pflug & Wozabal, *Quant. Finance*, 2007; Pflug & Pichler, *Springer*, 2014.

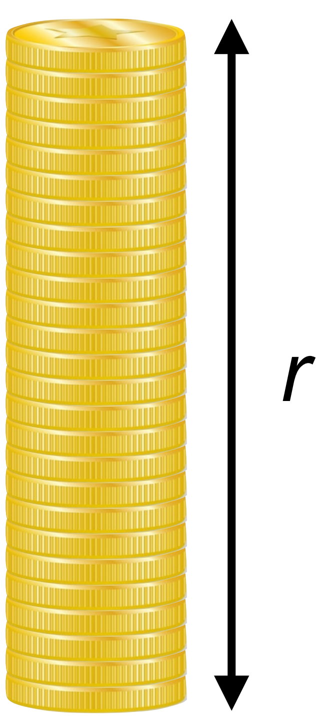
OT meets DRO

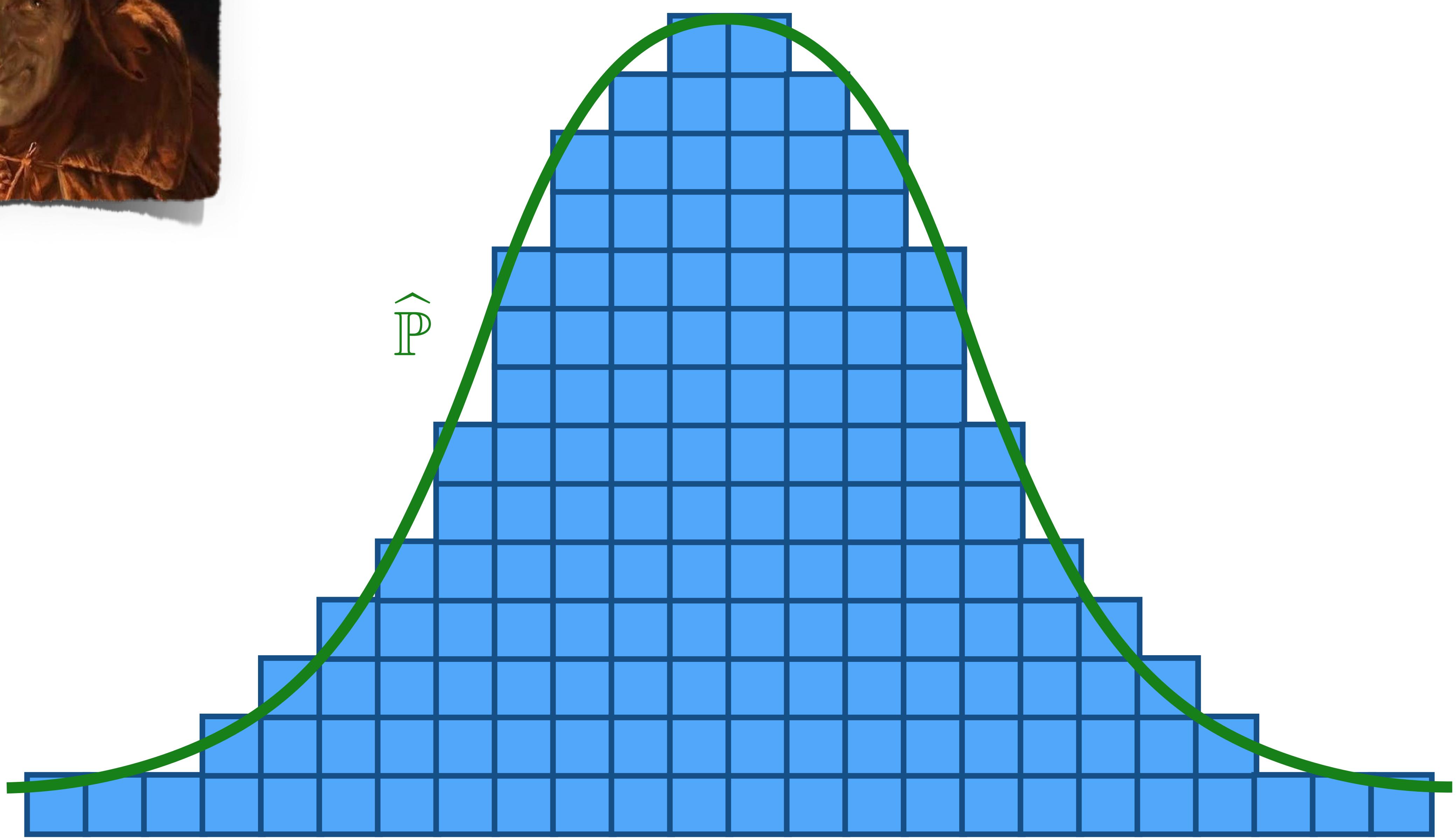


OT meets DRO

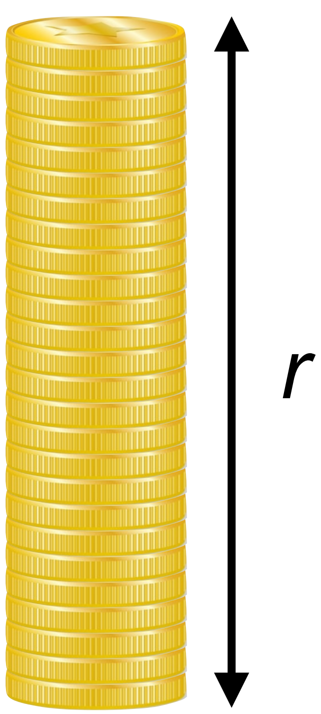


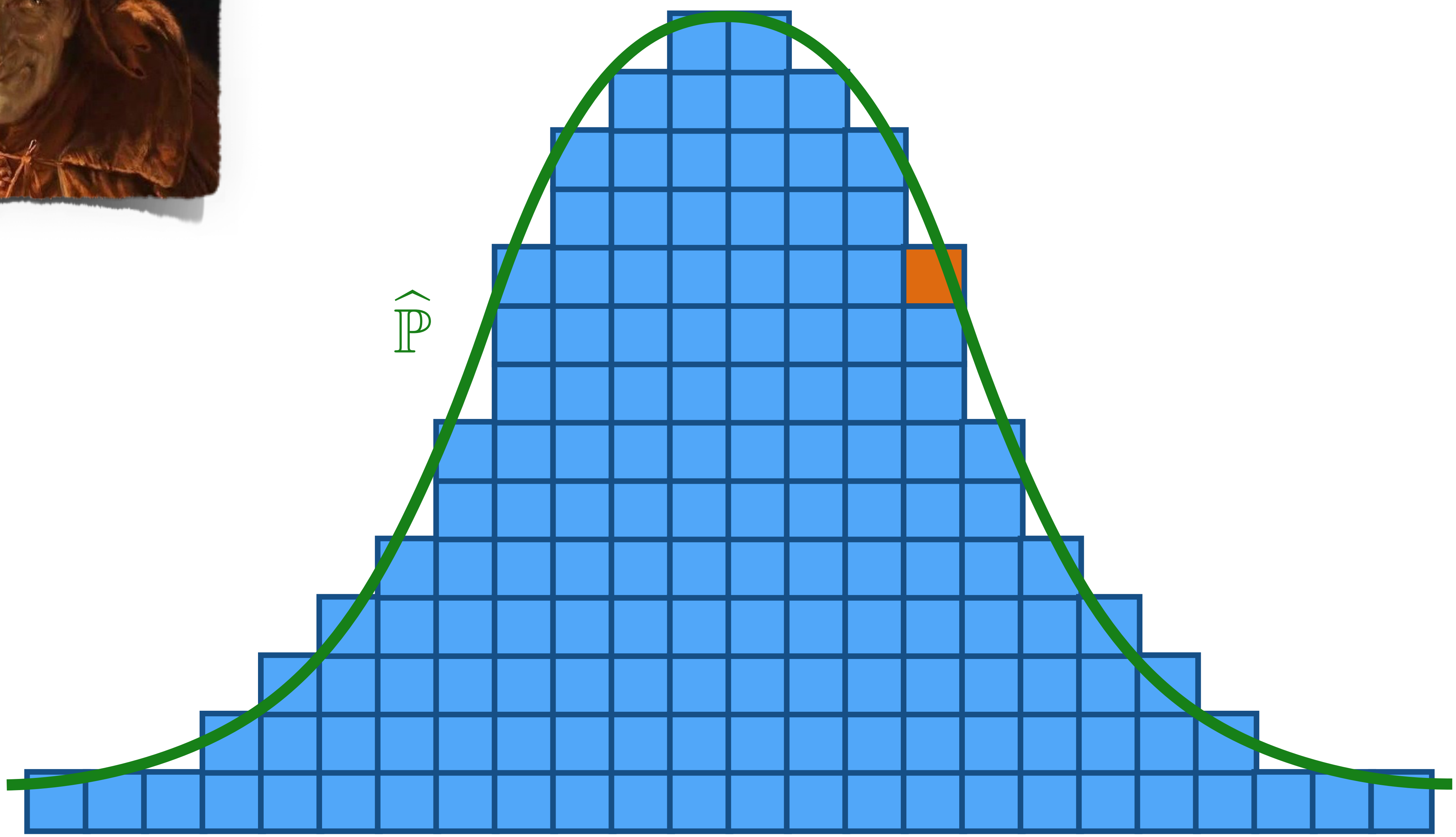
OT meets DRO



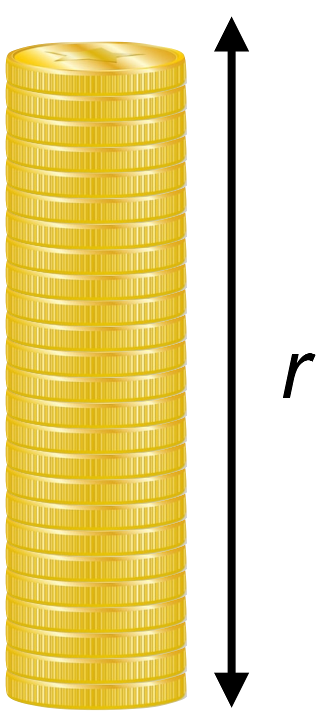


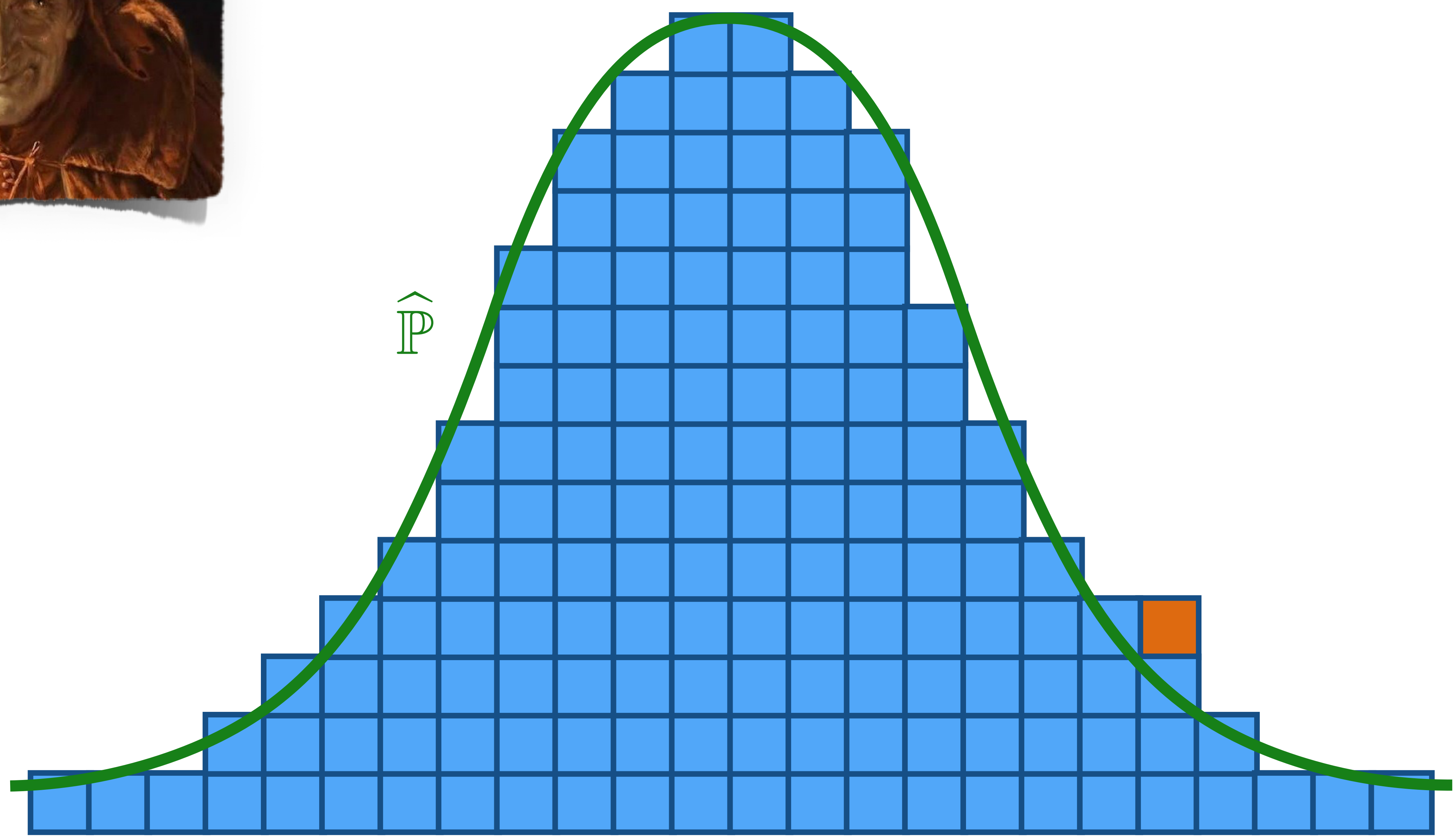
OT meets DRO



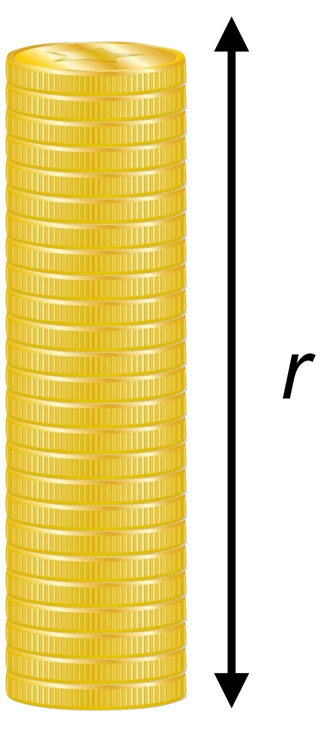


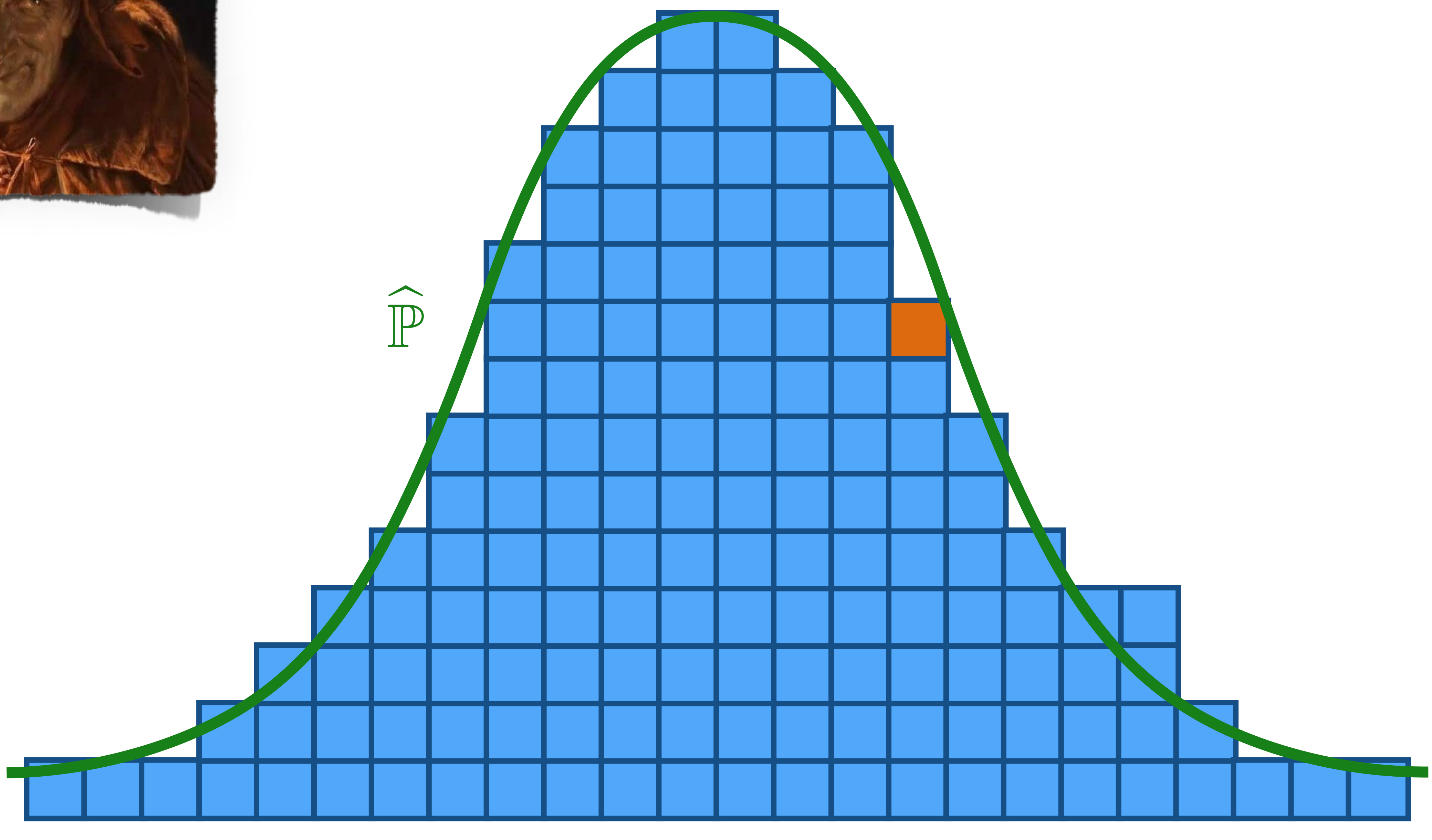
OT meets DRO



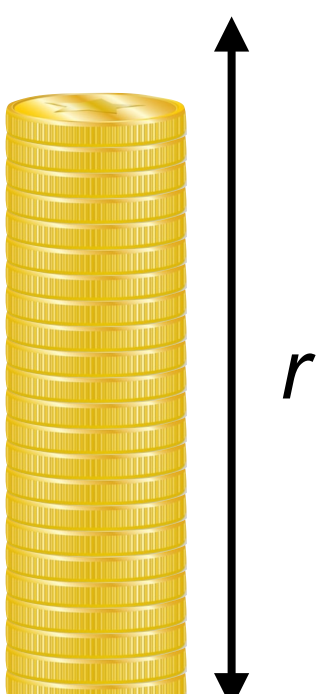


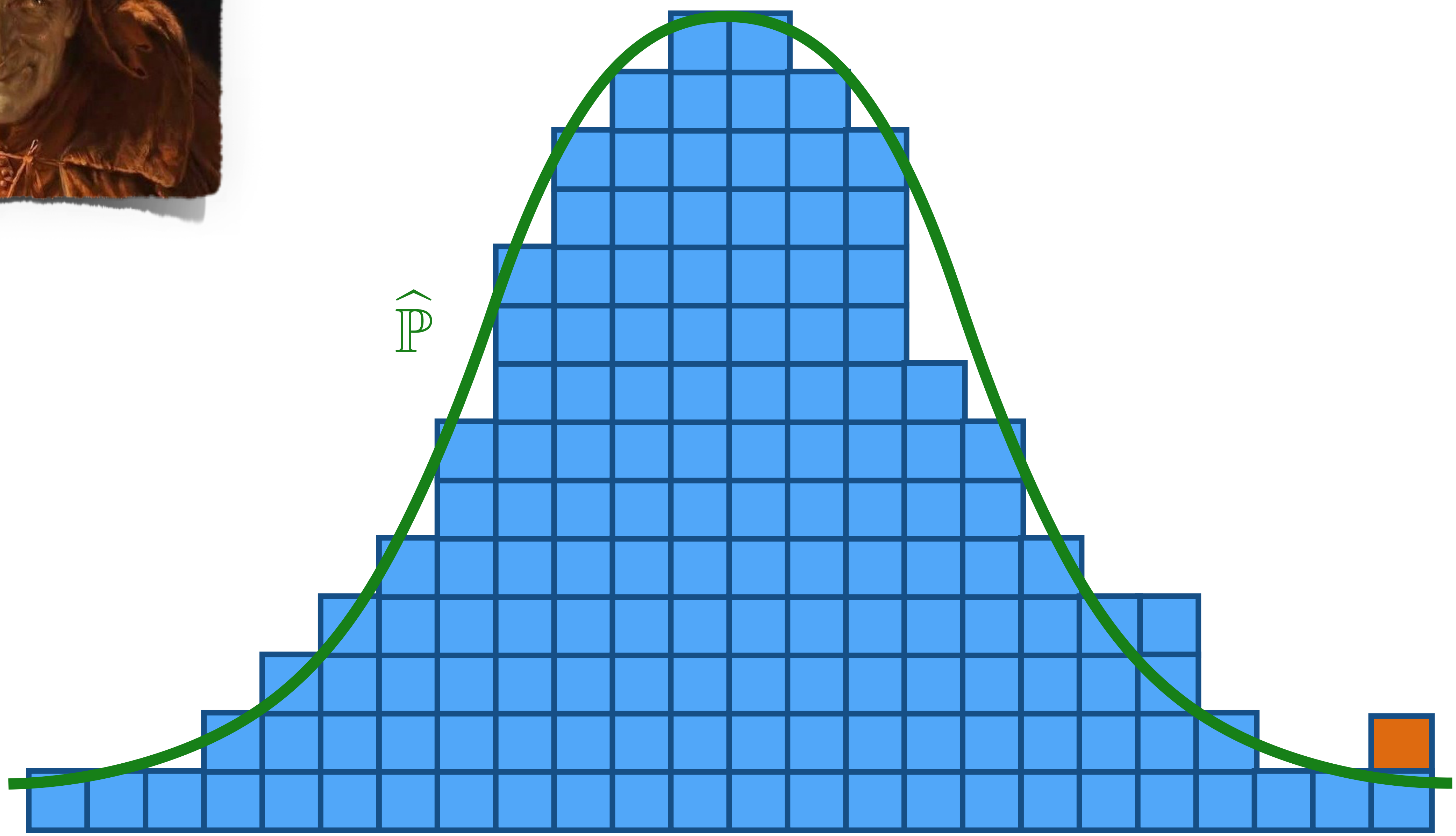
OT meets DRO



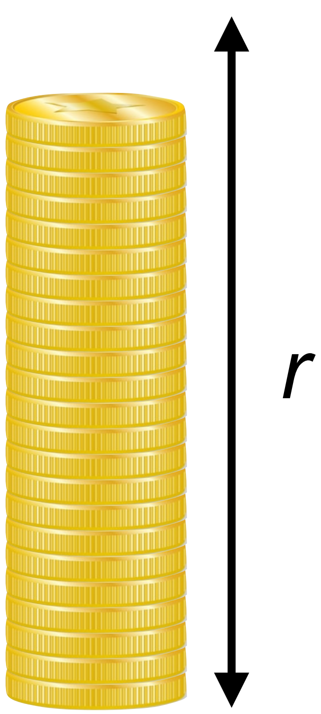


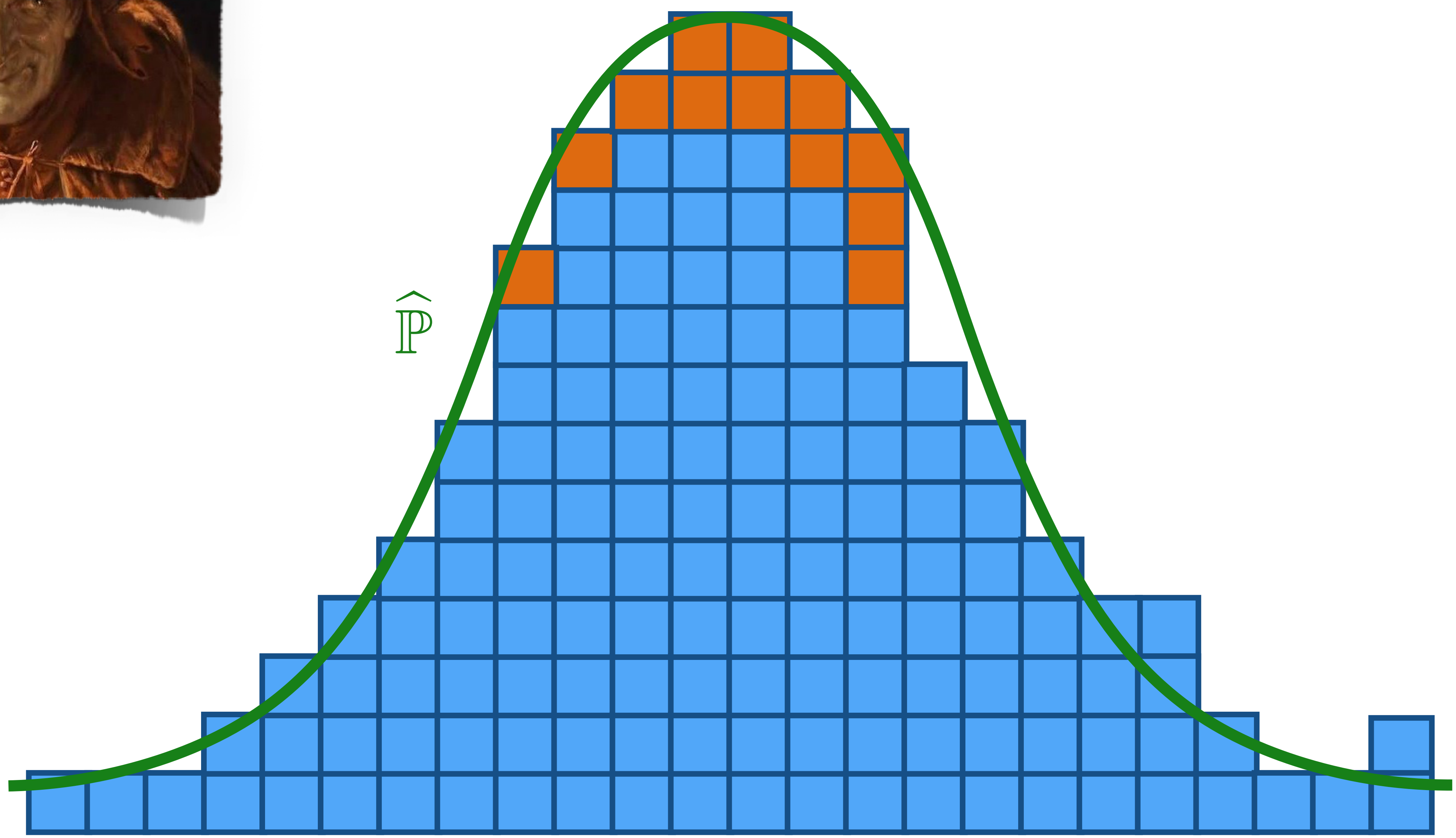
OT meets DRO



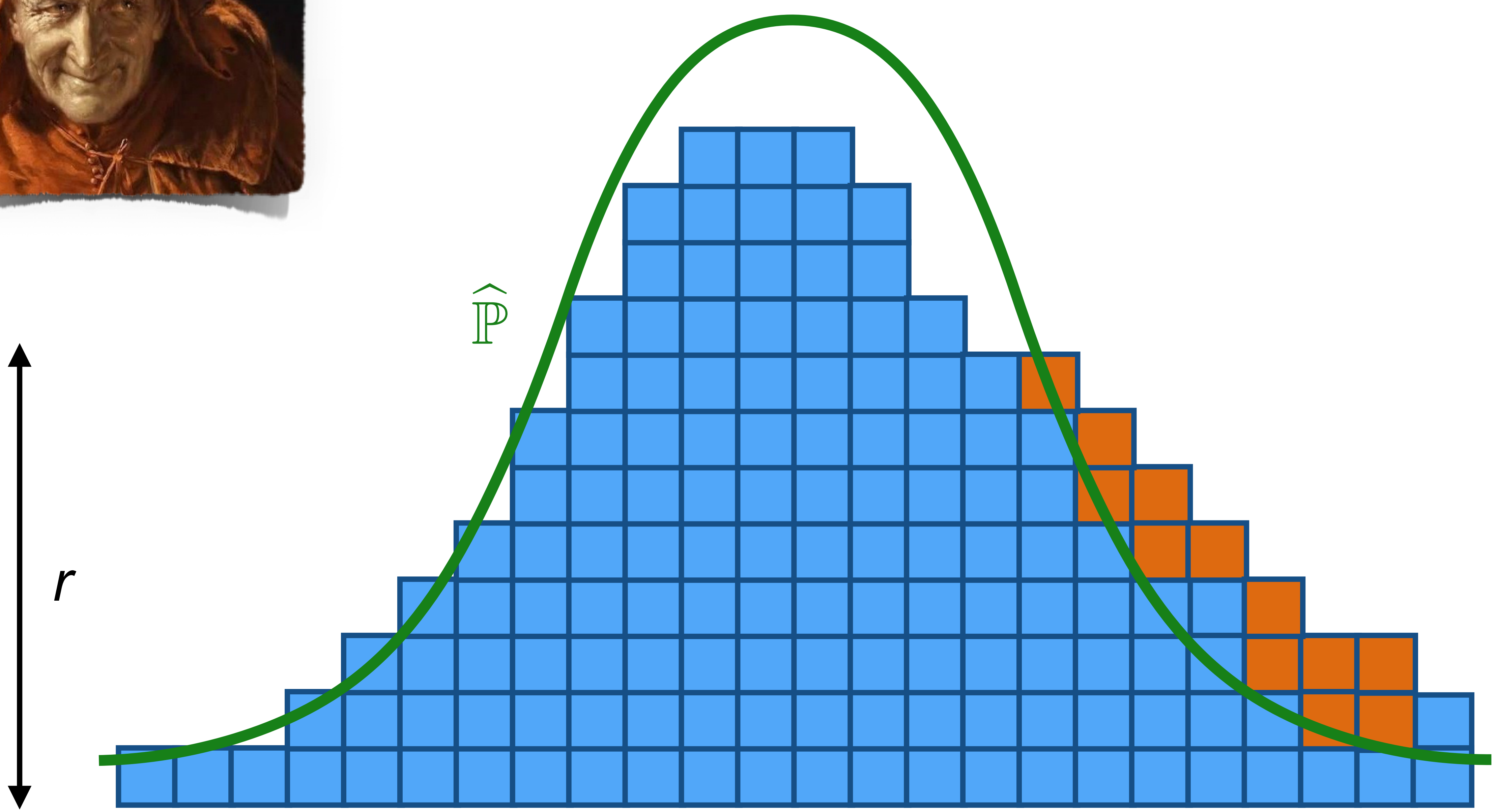


OT meets DRO

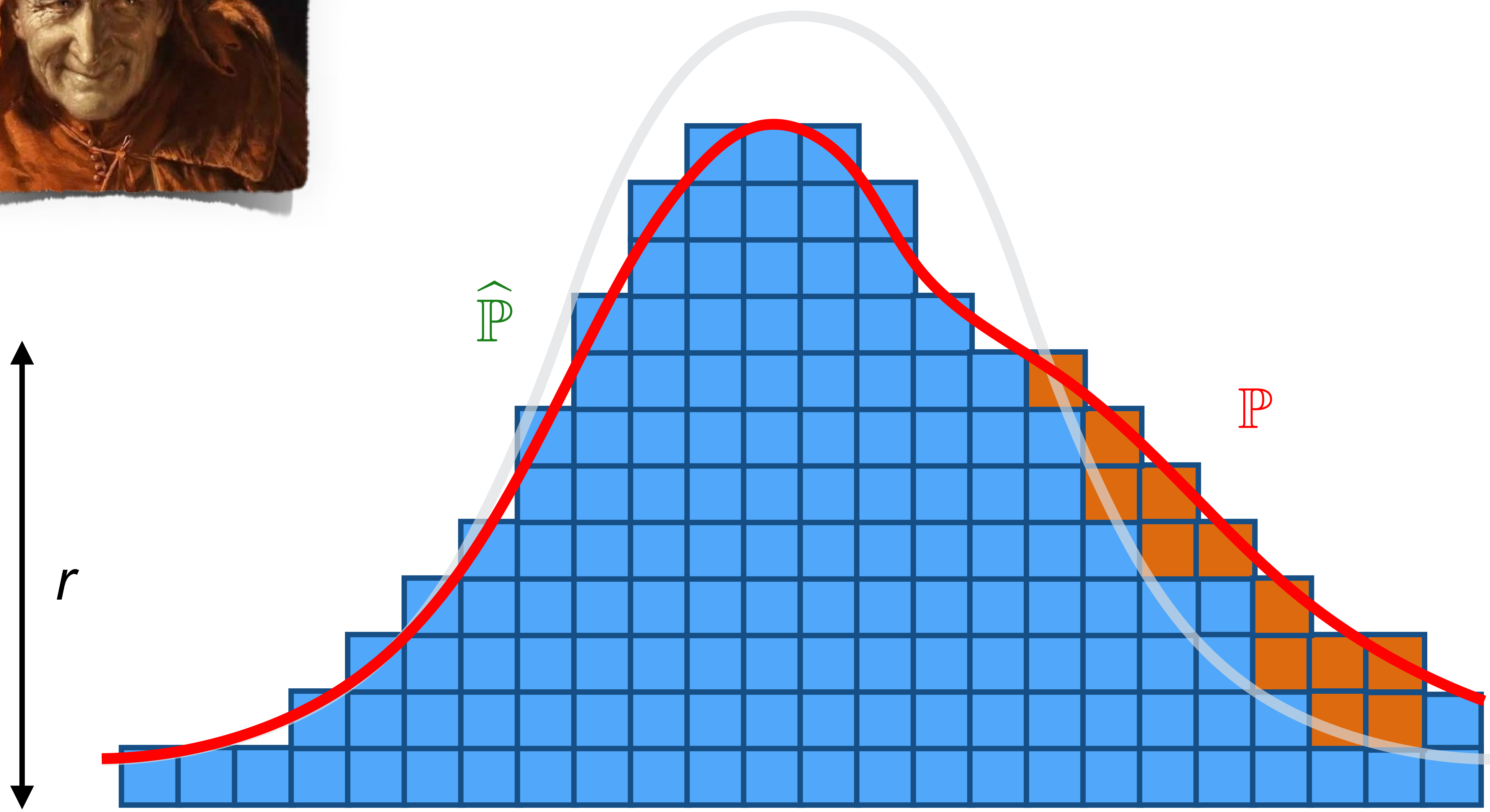




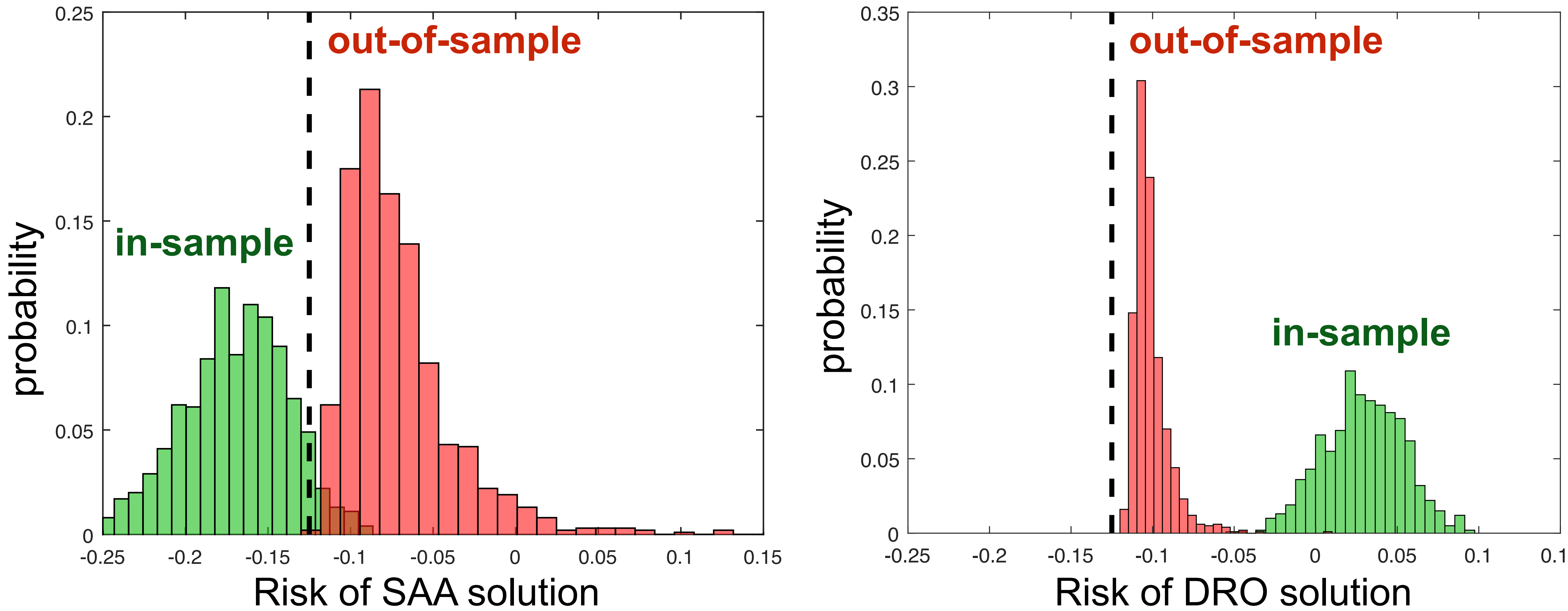
OT meets DRO



OT meets DRO

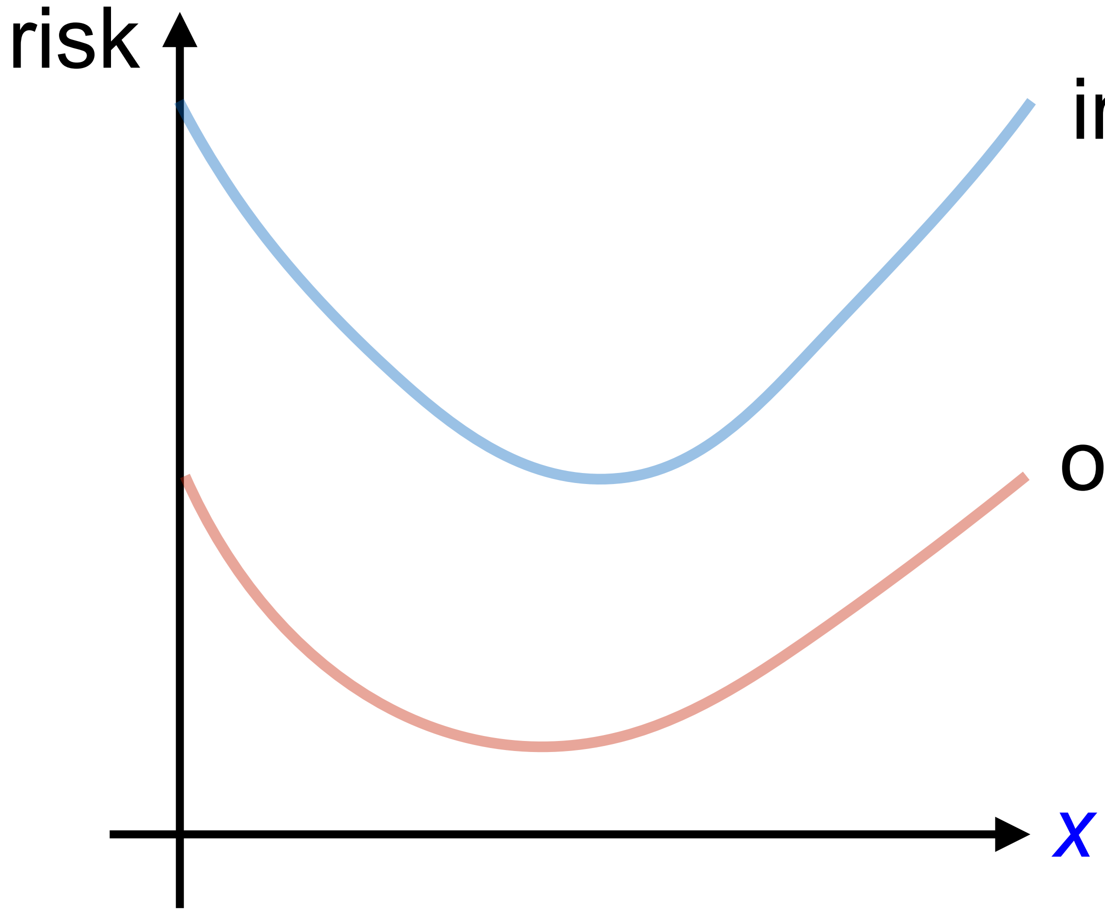


Underpromising and Overdelivering



1) Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018.

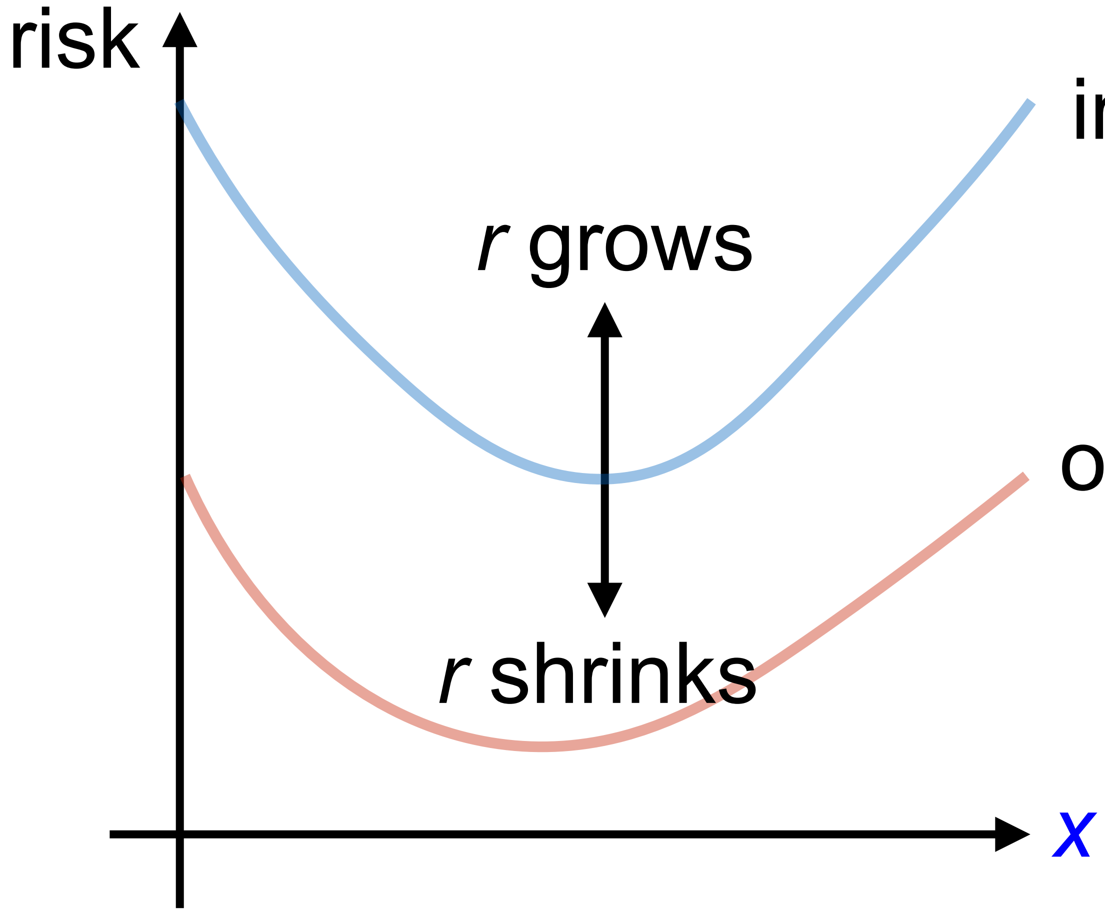
Statistical Guarantees



in-sample: $\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]$

out-of-sample: $\mathbb{E}_{\xi \sim \mathbb{P}^\star} [\ell(x, \xi)]$

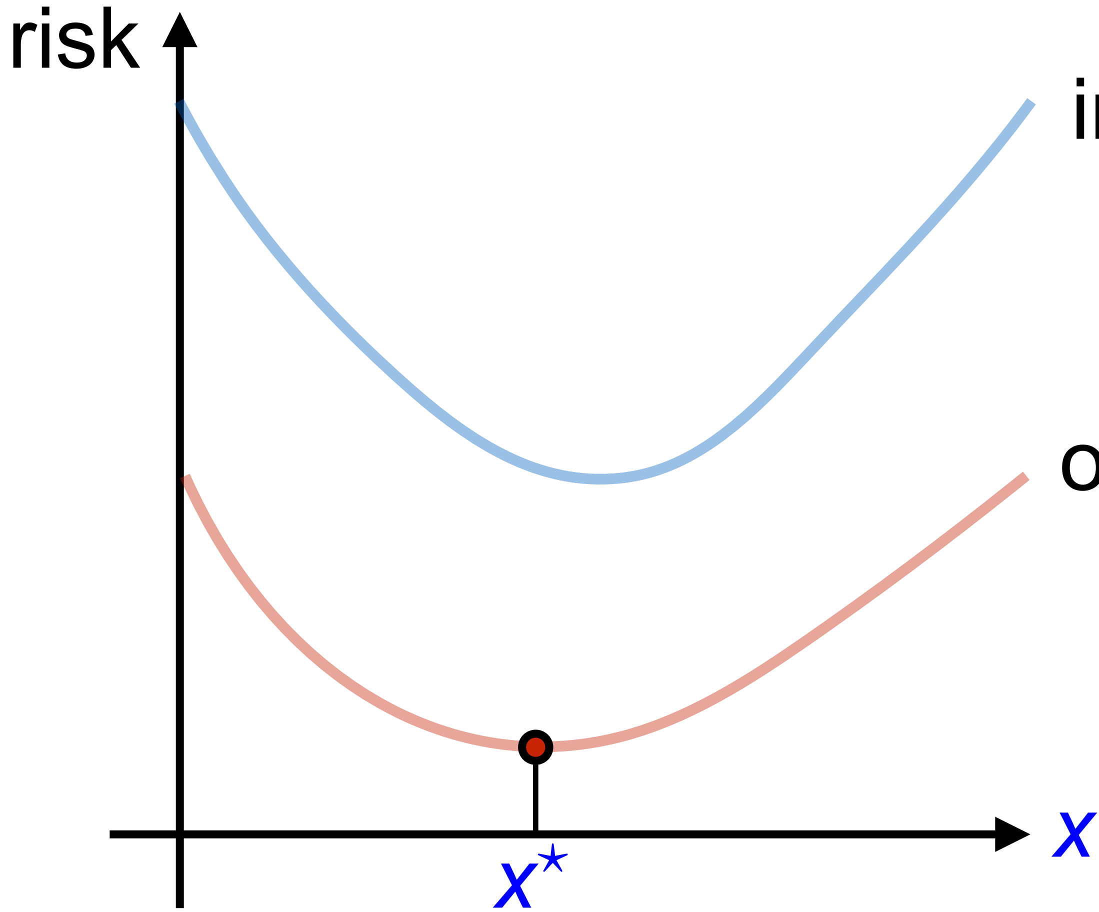
Statistical Guarantees



in-sample: $\sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]$

out-of-sample: $\mathbb{E}_{\xi\sim\mathbb{P}^\star}\left[\ell(x,\xi)\right]$

Statistical Guarantees



in-sample: $\sup\limits_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]$

out-of-sample: $\mathbb{E}_{\xi\sim\mathbb{P}^\star}\left[\ell(x,\xi)\right]$

Choice of radius:

- confidence set for[1] $\mathbb{P}^\star$: $r = O(N^{-1/d})$
- confidence set for[2] $x^\star$: $r = O(N^{-1/2})$
- confidence set for[3] $J^\star(x)$: $r = O(N^{-1/2})$

---

1) Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018.

2) Blanchet, Kang & Murthy, *J. Appl. Prob.*, 2019; Blanchet & Kang, *Oper. Res.*, 2021.

3) Gao, *Oper. Res.*, 2022.

Robustification vs Regularization

Theorem:[1] If $p \geq 1$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}}[\ell(x, \xi)] \leq \begin{cases} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N}[\ell(x, \xi)] \\ \cdots + \sum_{k=1}^{p-1} \frac{r^k}{k!} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N}\left[\|D_\xi^k \ell(x, \xi)\|^{q_k}\right]^{\frac{1}{q_k}} \\ \cdots + \frac{r^p}{p!} \sup_{\xi \in \Xi} \|D_\xi^p \ell(x, \xi)\|. \end{cases}$$

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p \geq 1$, then

nominal loss

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)] \leq \begin{cases} \boxed{\mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} [\ell(x, \xi)]} \\ \cdots + \sum_{k=1}^{p-1} \frac{r^k}{k!} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} \left[\| D_\xi^k \ell(x, \xi) \|^{q_k} \right]^{\frac{1}{q_k}} \\ \cdots + \frac{r^p}{p!} \sup_{\xi \in \Xi} \| D_\xi^p \ell(x, \xi) \|. \end{cases}$$

---

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p \geq 1$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)] \leq \begin{cases} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} [\ell(x, \xi)] \\ \cdots + \sum_{k=1}^{p-1} \boxed{\frac{r^k}{k!} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} \left[\| D_\xi^k \ell(x, \xi) \|^{q_k} \right]^{\frac{1}{q_k}}} \\ \cdots + \frac{r^p}{p!} \sup_{\xi \in \Xi} \| D_\xi^p \ell(x, \xi) \|. \end{cases}$$

variation regularization

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p \geq 1$, then

$$\sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right] \leq \begin{cases} \mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_N}\left[\ell(x,\xi)\right] \\ \cdots + \displaystyle\sum_{k=1}^{p-1} \frac{r^k}{k!} \mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_N}\left[\|D_\xi^k \ell(x,\xi)\|^{q_k}\right]^{\frac{1}{q_k}} \\ \cdots + \boxed{\dfrac{r^p}{p!} \displaystyle\sup_{\xi\in\Xi} \|D_\xi^p \ell(x,\xi)\|}. \end{cases}$$

Lipschitz regularization

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p \geq 1$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)] \leq \begin{cases} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} [\ell(x, \xi)] \\ \cdots + \sum_{k=1}^{p-1} \frac{r^k}{k!} \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} \left[\| D_\xi^k \ell(x, \xi) \|^{q_k} \right]^{\frac{1}{q_k}} \\ \cdots + \frac{r^p}{p!} \sup_{\xi \in \Xi} \| D_\xi^p \ell(x, \xi) \|. \end{cases}$$

Higher-order variation regularization used for:

- adversarial training of NNs (*e.g.*, Hein & Andriushchenko, *NeurIPS,* 2017)
- stabilizing training of GANs (*e.g.*, Roth *et al.*, *NeurIPS* 2017)
- regularization in imaging (*e.g.*, Bredies *et al.*, *SIAM J. Imaging Sci.*, 2010)

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p = 1$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)] \leq \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} [\ell(x, \xi)] + r \, \mathrm{lip}(\ell(x, \cdot)).$$

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Robustification vs Regularization

Theorem:[1] If $p = 1$, ℓ is convex in ξ and $\Xi = \mathbb{R}^d$, then

$$\sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right] = \mathbb{E}_{\xi\sim\widehat{\mathbb{P}}_N}\left[\ell(x,\xi)\right] + r\,\mathrm{lip}(\ell(x,\cdot)).$$

[1] Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018.

Robustification vs Regularization

Theorem:[1] If $p = 1$, ℓ is univariate convex and $\Xi = \mathbb{R}^d$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x^\top \xi) \right] = \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} \left[\ell(x^\top \xi) \right] + r \, \mathrm{lip}(\ell) \, \|x\|_* .$$

[1] Shafieeezadeh-Abadeh, Mohajerin Esfahani & Kuhn, *NeurIPS*, 2015; *JMLR* 2019.

Robustification vs Regularization

Theorem:[1] If $p = 1$, ℓ is univariate convex and $\Xi = \mathbb{R}^d$, then

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x^\top \xi) \right] = \mathbb{E}_{\xi \sim \widehat{\mathbb{P}}_N} \left[\ell(x^\top \xi) \right] + r \, \mathrm{lip}(\ell) \, \|x\|_*.$$

Offers robustness interpretation for:

- lasso regularization
- basis pursuit denoising
- regularized least absolute deviation regression
- the Dantzig selector

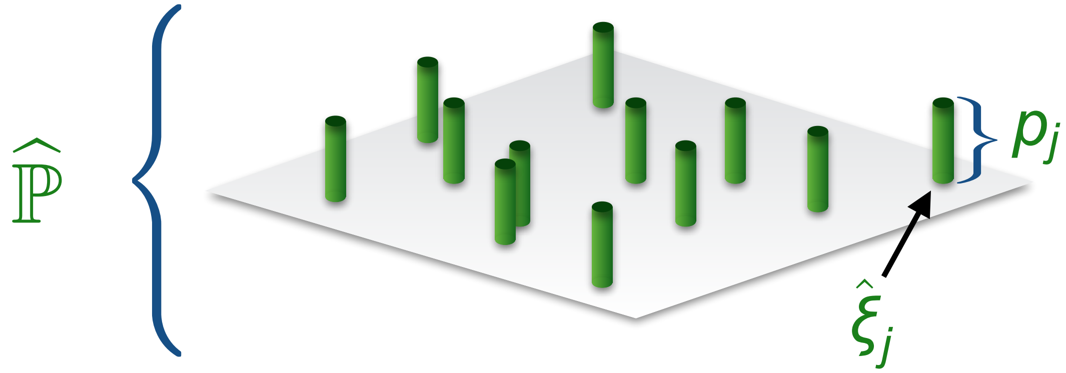
[1] Shafieeezadeh-Abadeh, Mohajerin Esfahani & Kuhn, *NeurIPS*, 2015; *JMLR* 2019.

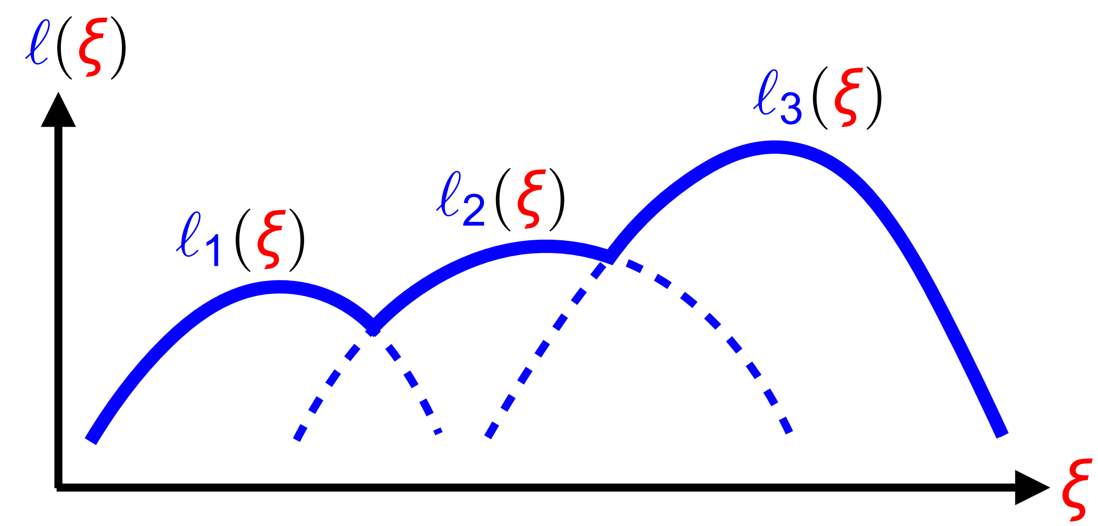
Computational Tractability

Primal DRO problem:

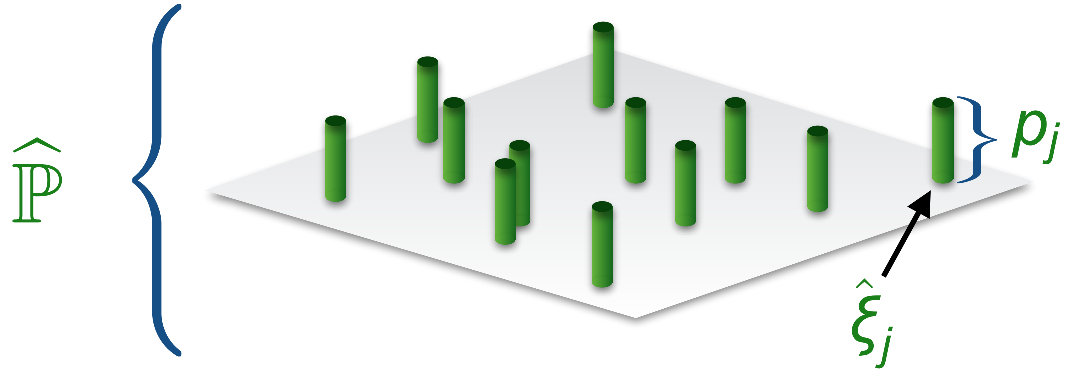
$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x, \xi) \right]$$

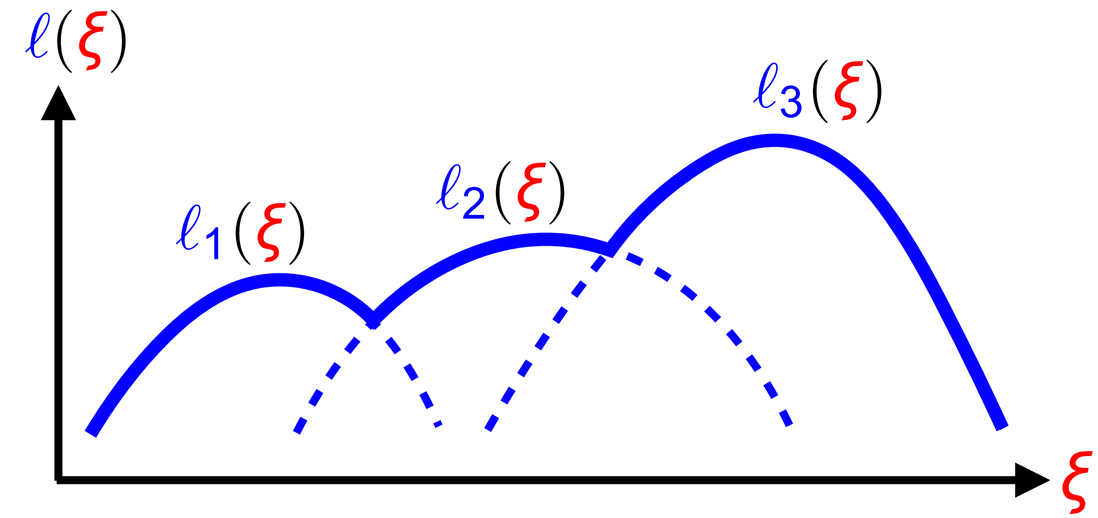
Computational Tractability





Computational Tractability





Theorem:[1] The **primal** DRO problem is equivalent to

$$
\begin{array}{lll}
\inf & \lambda r + \sum_j p_j s_j & \\
\text{s.t.} & x \in \mathcal{X},\ \lambda, \tau_{ij} \in \mathbb{R}_+,\ s_j \in \mathbb{R},\ \zeta_{ij}^{\ell}, \zeta_{ij}^{c}, \zeta_{ij}^{\Xi} \in \mathbb{R}^d & \forall i,j \\
& (-\ell_i)^{*2}(x, \zeta_{ij}^{\ell}) + \lambda c^{*1}(\zeta_{ij}^{c}/\lambda, \hat{\xi}_j) + \tau_{ij}\sigma_{\Xi}(\zeta_{ij}^{\Xi}/\tau_{ij}) \le s_j & \forall i,j \\
& \zeta_{ij}^{\ell} + \zeta_{ij}^{c} + \zeta_{ij}^{\Xi} = 0 & \forall i,j
\end{array}
$$

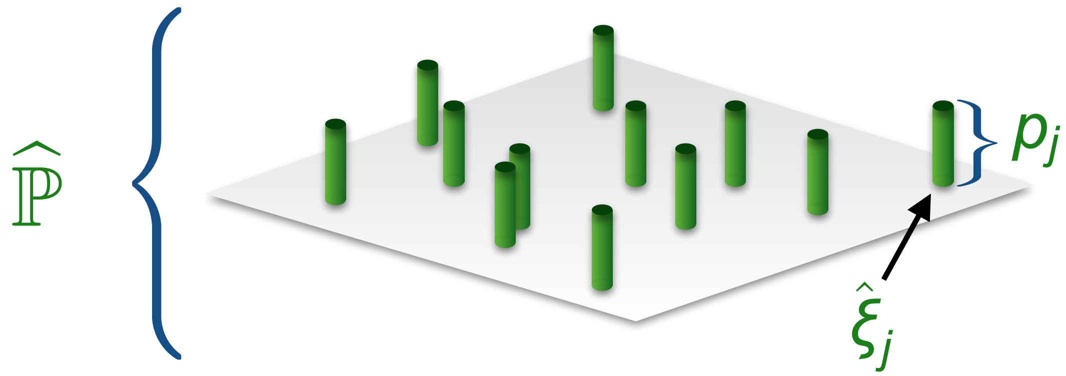
[1] Mohajerin Esfahani & Kuhn, *Math. Program.*, 2018; Zhao & Guan, *Oper. Res. Lett.*, 2018; Blanchet & Murthy, *Math. Oper. Res.*, 2019; Gao & Kleytweg, *Math. Oper. Res.*, 2022; Chen, Kuhn & Wiesemann, *Oper. Res.*, 2023.

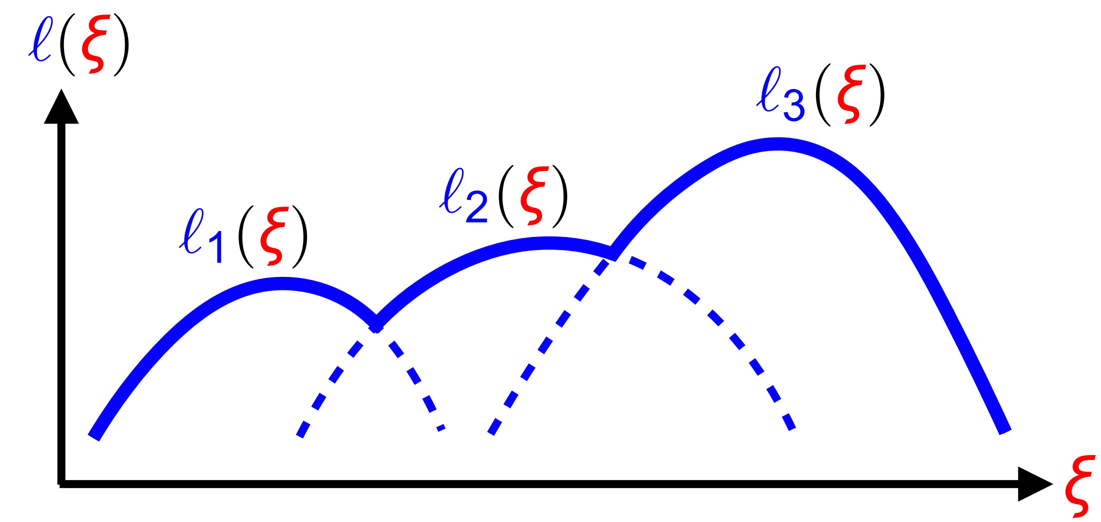
Computational Tractability

Dual DRO problem:

$$\sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \ \inf_{x \in \mathcal{X}} \ \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x, \xi) \right]$$

Computational Tractability





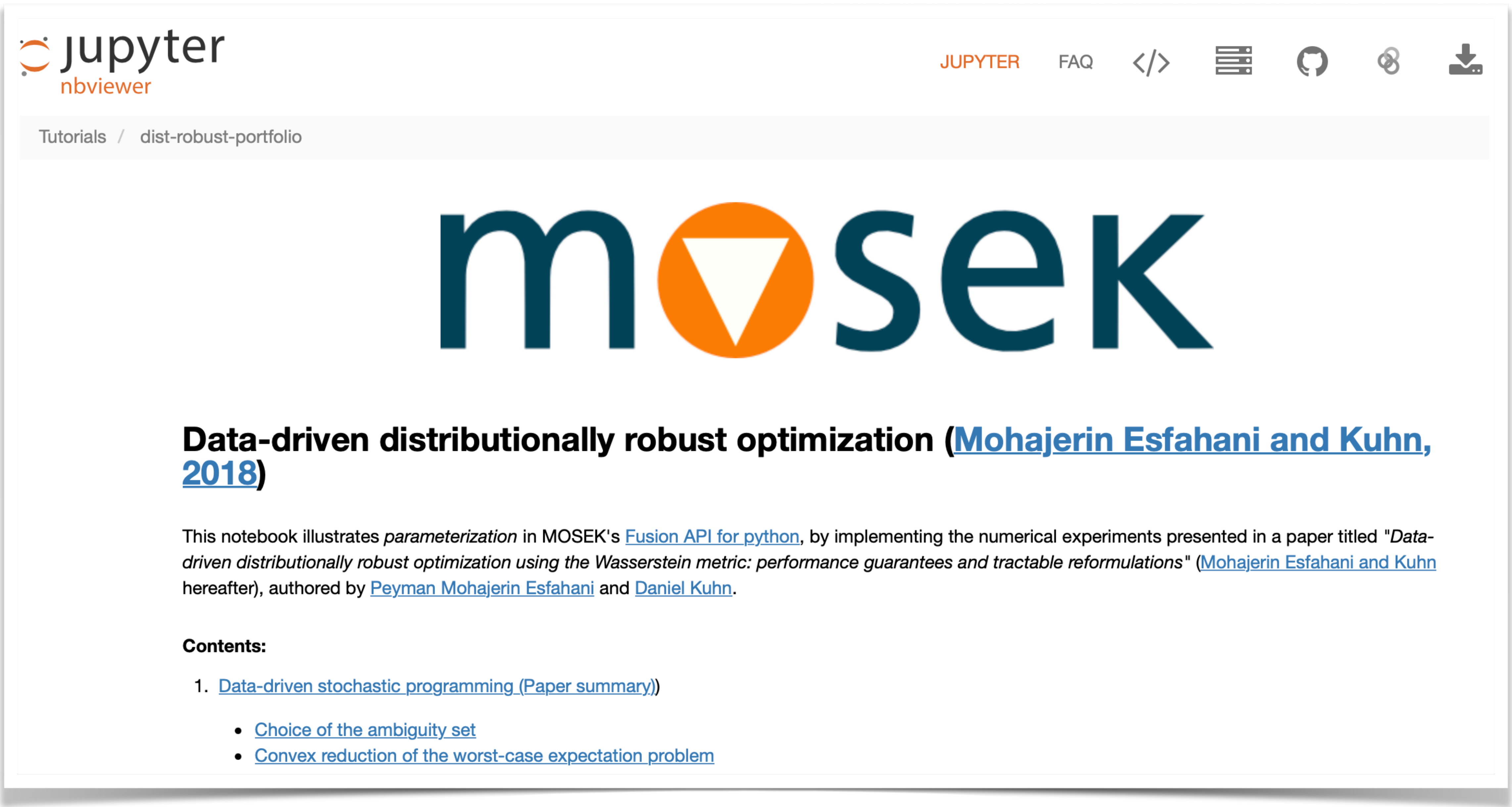
Theorem:[1] The **dual** DRO problem is equivalent to

$$
\begin{array}{lll}
\sup & -\sum_i \sum_j q_{ij} \ell_i^{*1}(\alpha_{ij}, \hat{\xi}_j + \xi_{ij}/q_{ij}) - v\, \sigma_{\mathcal{X}}(\beta/v) & \\
\text{s.t.} & q_{ij}, v \in \mathbb{R}_+, \ \xi_{ij} \in \mathbb{R}^d, \ \alpha_{ij}, \beta \in \mathbb{R}^n & \forall i, j \\
& q_{ij}\, \delta_{\Xi}(\hat{\xi}_j + \xi_{ij}/q_{ij}) \leq 0 & \forall i, j \\
& \sum_i \sum_j \alpha_{ij} + \beta = 0, \ \sum_i q_{ij} = p_j & \forall j \\
& \sum_i \sum_j q_{ij}\, c(\hat{\xi}_j + \xi_{ij}/q_{ij}, \hat{\xi}_j) \leq r &
\end{array}
$$

[1] Shafieezadeh-Abadeh, Aolaritei, Dörfler & Kuhn, *Working Paper*, 2023.

Computational Tractability

Jupyter notebook based on Mosek's Fusion API for Python:[1)]



jupyter nbviewer

JUPYTER FAQ

Tutorials / dist-robust-portfolio

Data-driven distributionally robust optimization (Mohajerin Esfahani and Kuhn, 2018)

This notebook illustrates *parameterization* in MOSEK's Fusion API for python, by implementing the numerical experiments presented in a paper titled "*Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations*" (Mohajerin Esfahani and Kuhn hereafter), authored by Peyman Mohajerin Esfahani and Daniel Kuhn.

Contents:

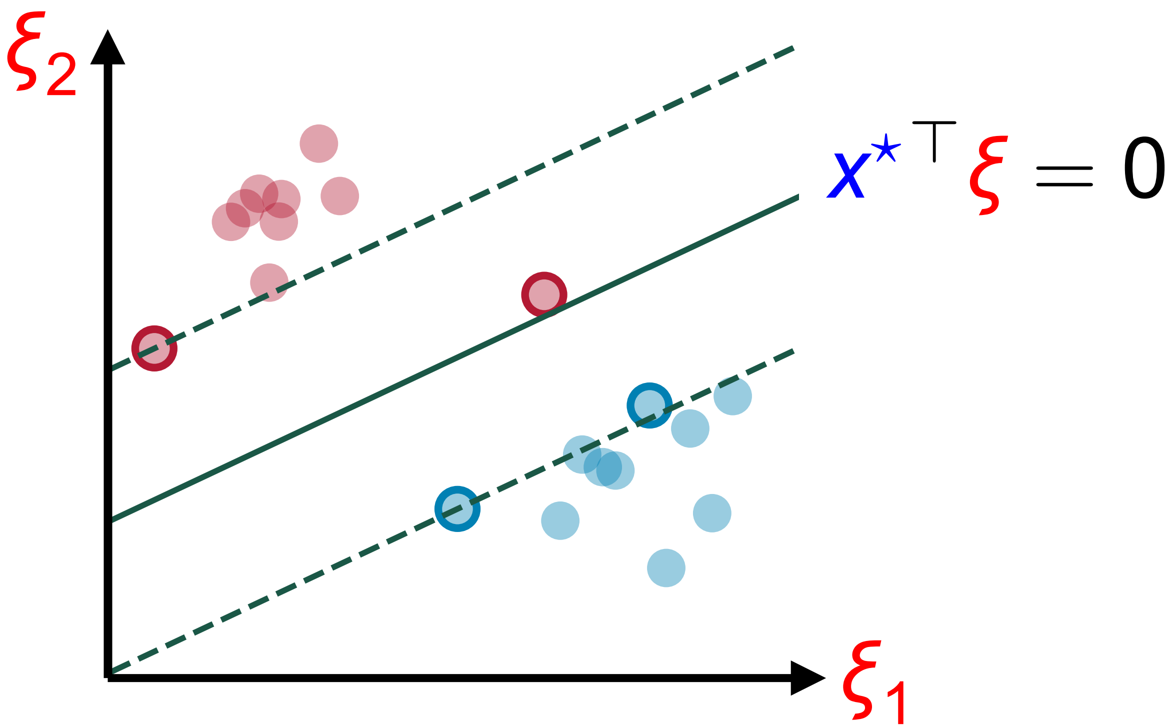
1. Data-driven stochastic programming (Paper summary))
   - Choice of the ambiguity set
   - Convex reduction of the worst-case expectation problem

---

1) https://nbviewer.jupyter.org/github/MOSEK/Tutorials

Distributionally Robust SVM

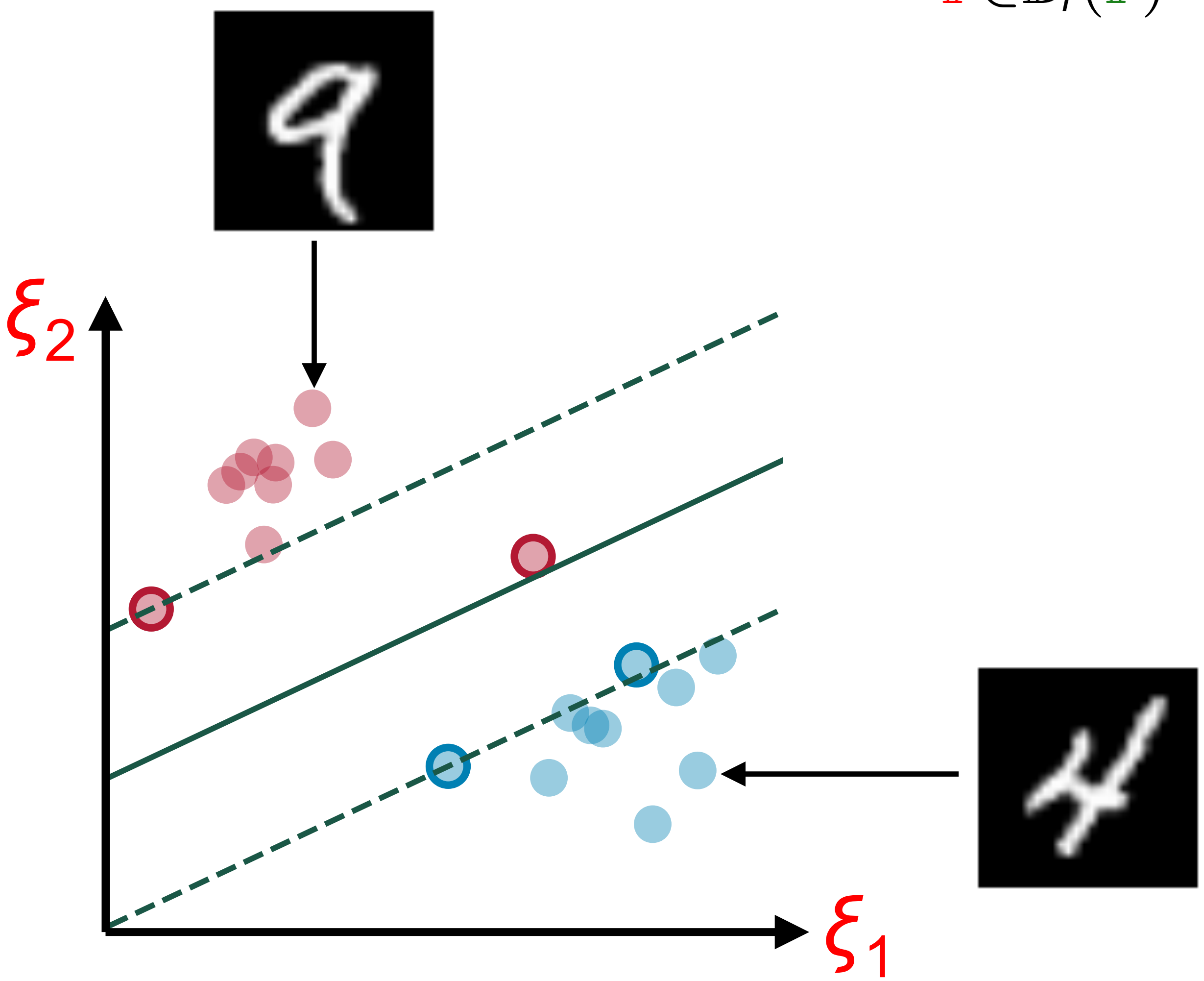
$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]$$



empirical distribution $\widehat{\mathbb{P}}$

Distributionally Robust SVM

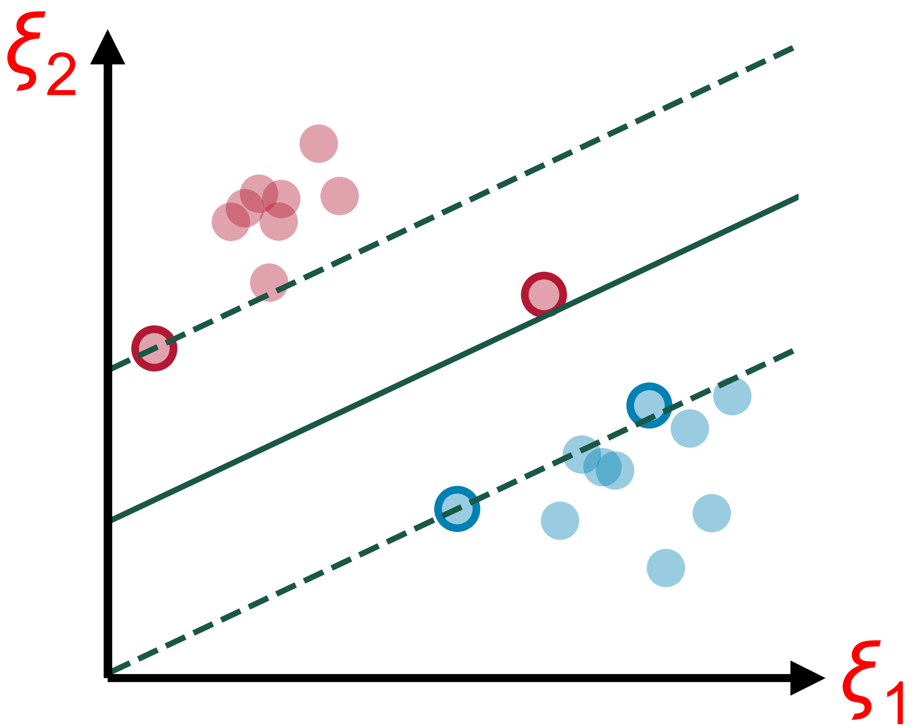
$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]$$



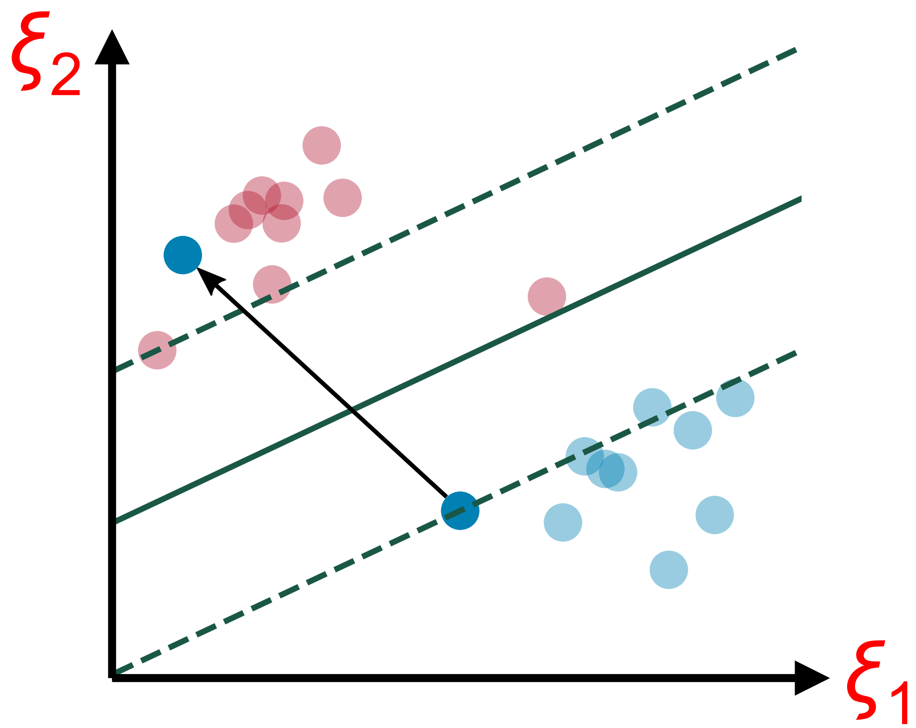
empirical distribution $\widehat{\mathbb{P}}$

Distributionally Robust SVM

$$\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x, \xi) \right]$$



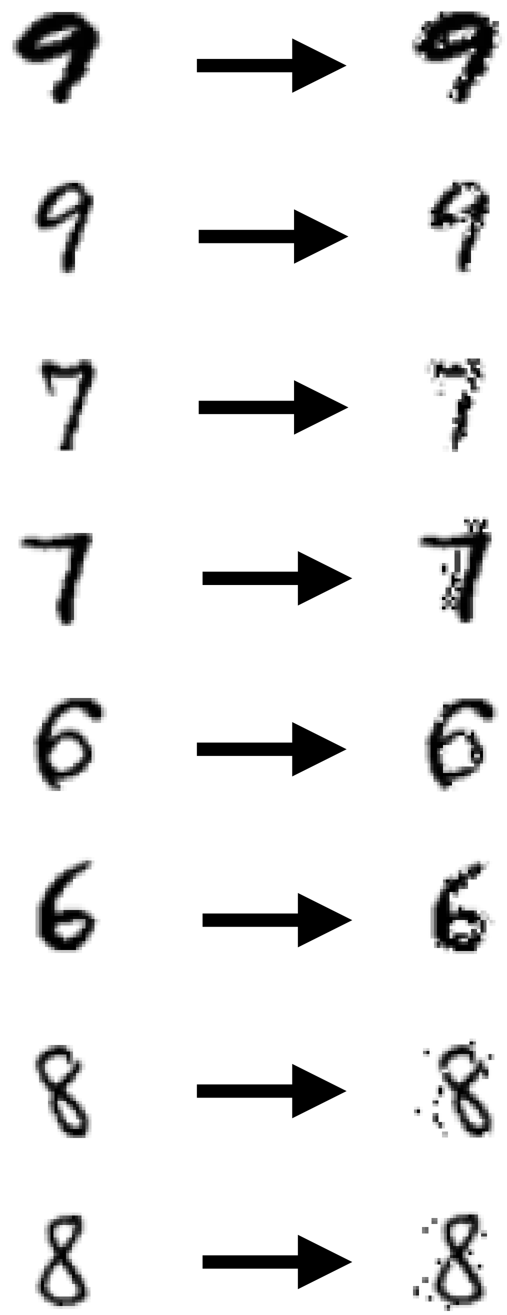
empirical distribution $\widehat{\mathbb{P}}$



worst-case distribution $\mathbb{P}^\star$

Adversarial Examples

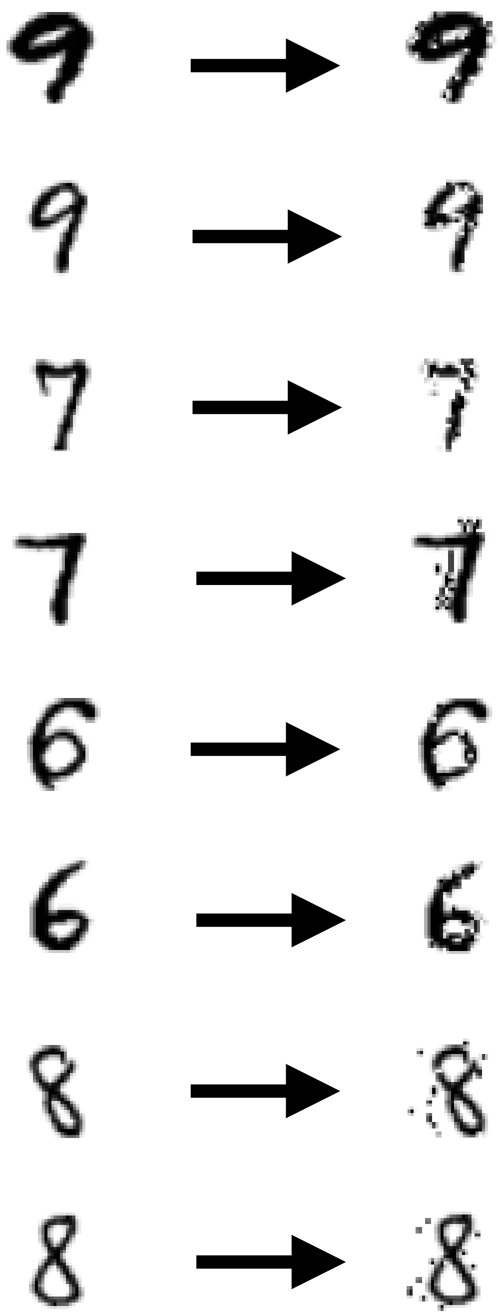
$$\mathbb{P}^\star \in \underset{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})}{\text{argmax}} \; \mathbb{E}_{\xi \sim \mathbb{P}} \left[\ell(x^\star, \xi) \right]$$

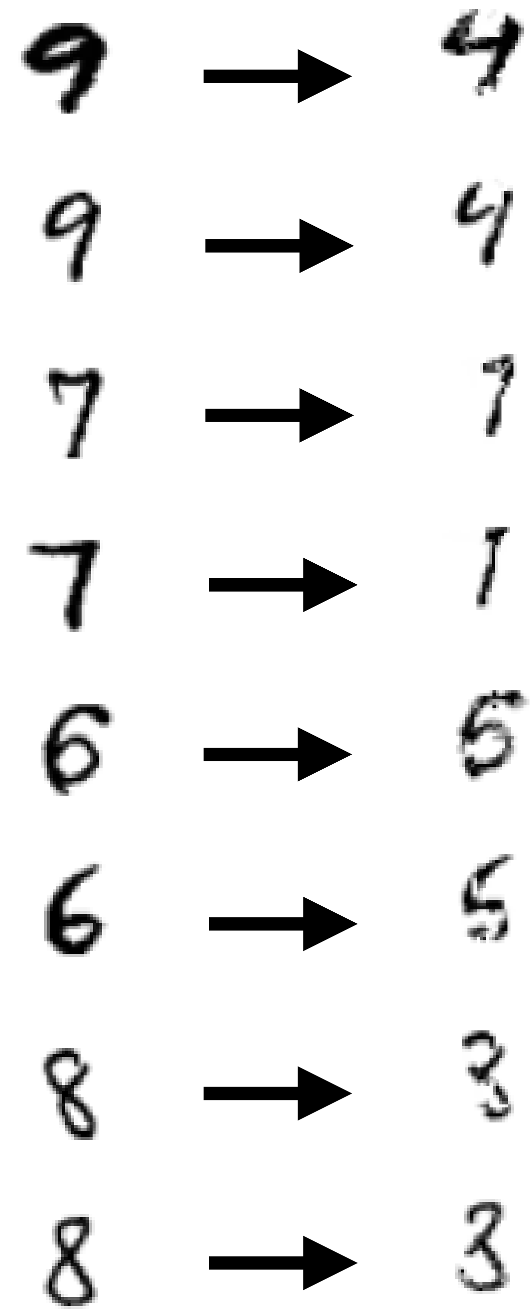


Adversarial Examples

$$\mathbb{P}^\star \in \operatorname*{argmax}_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x^\star, \xi)\right]$$

$$\mathbb{P}^\star \in \operatorname*{argmax}_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \inf_{x\in\mathcal{X}} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x, \xi)\right]$$

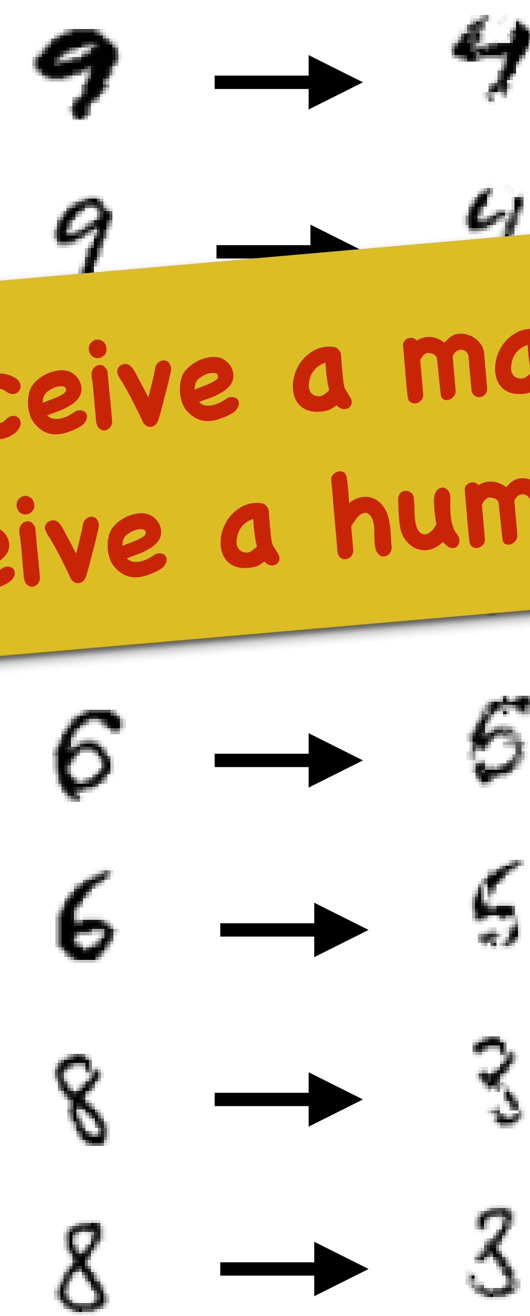




Adversarial Examples

$$\mathbb{P}^\star \in \operatorname*{argmax}_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x^\star,\xi)\right]$$

$$\mathbb{P}^\star \in \operatorname*{argmax}_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \inf_{x\in\mathcal{X}} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]$$



Best response to $x^\star$ can deceive a machine.
Nash strategy can deceive a human!

From Distributions to Moments

Disadvantages of the Empirical Distribution

Limited scalability: $$\underbrace{\inf_{x\in\mathcal{X}} \sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]}$$

$\cong$ convex program of size $\mathcal{O}(N)$

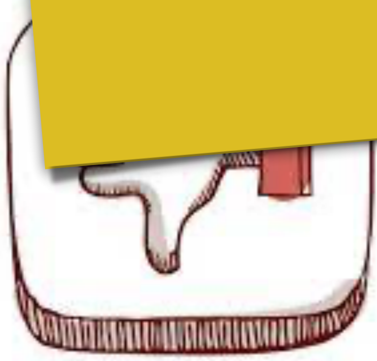
Slow convergence to true distribution

Disadvantages of Working with Distributions

Limited scalability: $\underbrace{\inf_{x \in \mathcal{X}} \sup_{\mathbb{P} \in \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]}$

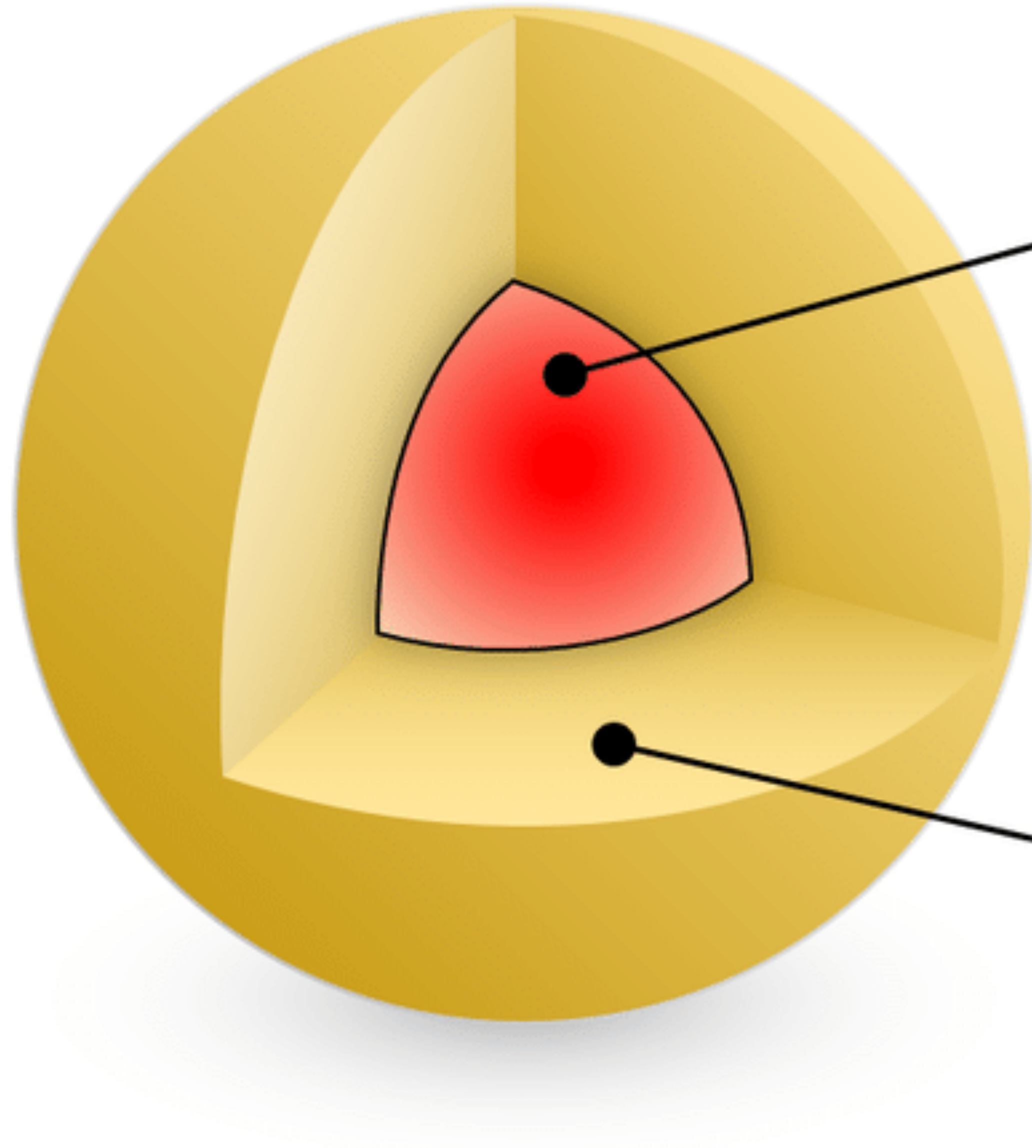
$\simeq$ convex

Focus on 1st and 2nd moments!



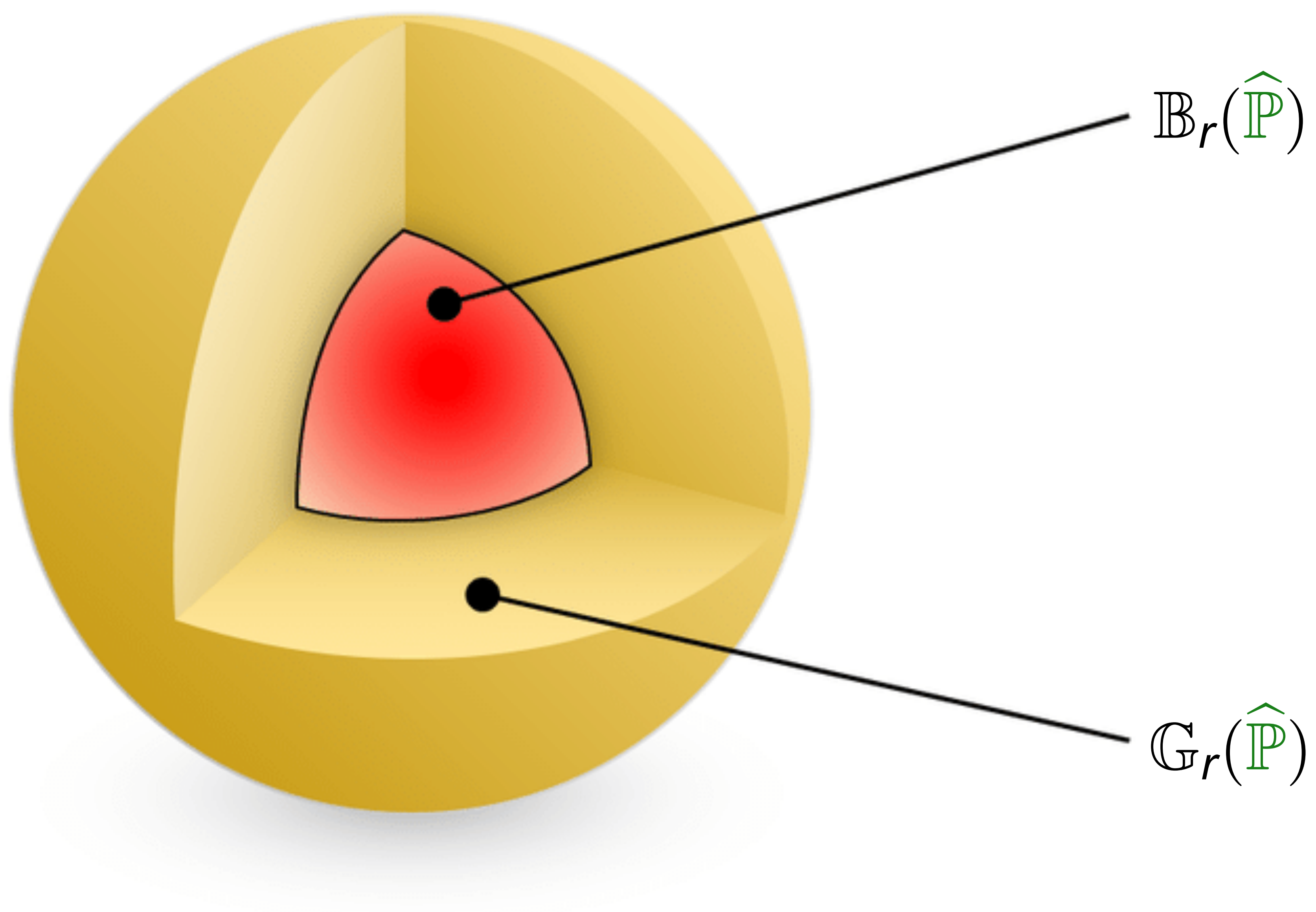
Statistical guarantees difficult to obtain.

Gelbrich Ambiguity Set



$$\mathbb{G}_r(\widehat{\mathbb{P}}) = \left\{ \mathbb{P} \sim (\mu, \Sigma) \;\middle|\; \|\mu - \hat{\mu}\|_2^2 + \mathrm{tr}\left[\Sigma + \hat{\Sigma} - 2\left(\Sigma^{\frac{1}{2}} \hat{\Sigma} \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right] \leq r^2 \right\}$$

Gelbrich Ambiguity Set

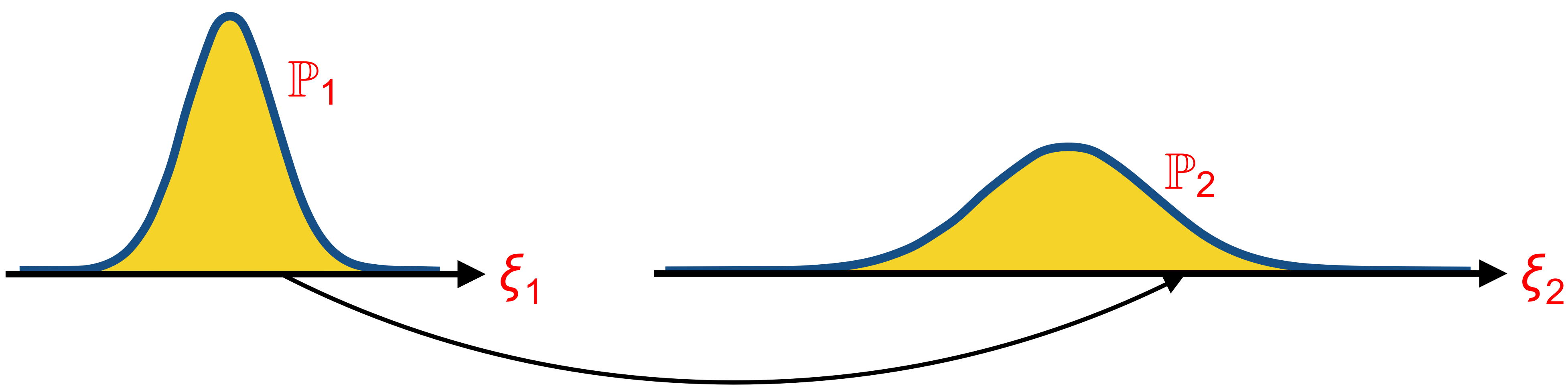


$$\mathbb{G}_r(\widehat{\mathbb{P}}) = \left\{ \mathbb{P} \sim (\mu, \Sigma) \;\middle|\; \underbrace{\|\mu - \hat{\mu}\|_2^2 + \mathrm{tr}\left[\Sigma + \hat{\Sigma} - 2\left(\Sigma^{\frac{1}{2}} \hat{\Sigma} \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]}_{\leq W_2^2(\mathbb{P}, \widehat{\mathbb{P}})} \leq r^2 \right\}$$

Structural Information

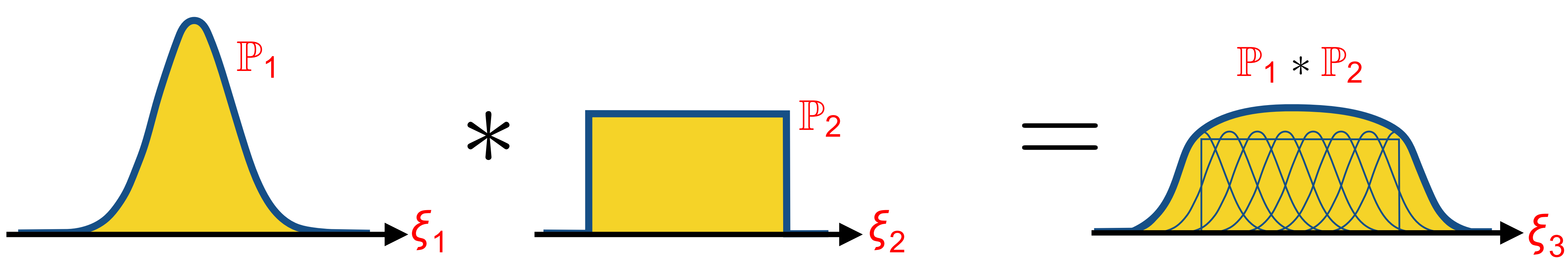
Stable ambiguity set $\mathcal{S}$: Closed under

- affine pushforwards



$\xi_2 = A\xi_1 + b$

- convolutions



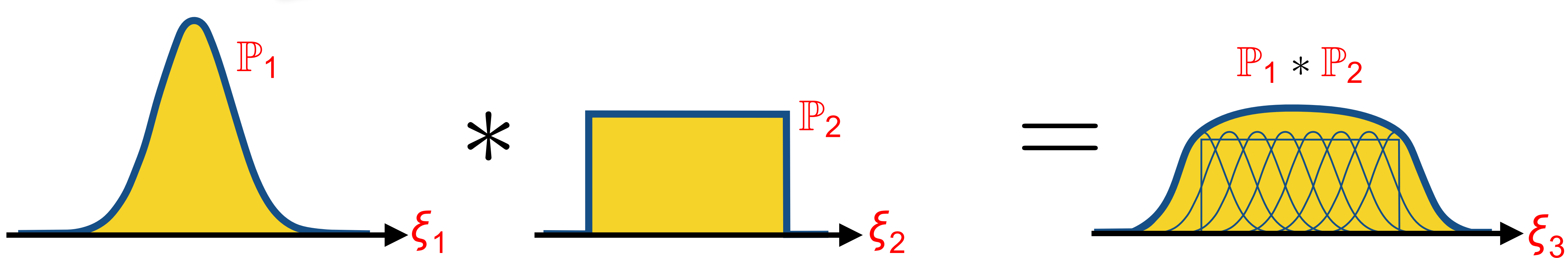
Structural Information

Stable ambiguity set $\mathcal{S}$: Closed under

- affine pushforwards
- convolutio

Examples: All distributions that are…

- symmetric
- symmetric & linear unimodal
- log-concave
- Gaussian



Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \alpha\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[\boxed{-\xi^\top x}\right] = -\hat{\mu}^\top x + \alpha\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

portfolio loss

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = \underbrace{\boxed{-\hat{\mu}^\top x}}_{\text{nominal loss}} + \alpha\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \alpha\boxed{\sqrt{x\hat{\Sigma}x}} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

nominal standard deviation

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \alpha\sqrt{x\hat{\Sigma}x} + \boxed{r\sqrt{1+\alpha^2}\,\|x\|_2}.$$

regularization term

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \boxed{\alpha}\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\boxed{\alpha}^2}\,\|x\|_2\,.$$

Standard risk coefficient: $\alpha = \sup_{\substack{\mathbb{P}\in\mathcal{S}\\ \mathbb{P}\sim(\mu,\Sigma)}} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\frac{(\xi-\mu)^\top x}{\sqrt{x^\top\Sigma x}}\right]$

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \boxed{\alpha}\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\boxed{\alpha}^2}\,\|x\|_2\,.$$

Standard risk coefficient: $\alpha = \sup_{\substack{\mathbb{P}\in\mathcal{S}\\ \mathbb{P}\sim(\mu,\Sigma)}} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\frac{(\xi-\mu)^\top x}{\sqrt{x^\top\Sigma x}}\right]$

independent of x, μ and Σ

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \alpha\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

Proposition: If $\mathcal{R}_{\xi\sim\mathbb{P}} = \beta\text{-VaR}_{\xi\sim\mathbb{P}}$, then

$$\alpha = \begin{cases} \sqrt{(1-\beta)/\beta} & \mathcal{S} = \{\text{all distributions}\} \\ 1/\sqrt{2\beta} & \mathcal{S} = \{\text{all symmetric distributions}\} \\ 1/(3\sqrt{2\beta}) & \mathcal{S} = \{\text{all symm. lin. unimodal distributions}\} \\ \Phi^{-1}(1-\beta) & \mathcal{S} = \{\text{all Gaussian distributions}\} \end{cases}$$

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Mean-Covariance Robust Portfolios

Theorem:[1] If $\mathcal{S}$ is stable and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is translation-invariant, scale-invariant and law-invariant, then

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathcal{R}_{\xi\sim\mathbb{P}}\left[-\xi^\top x\right] = -\hat{\mu}^\top x + \alpha\sqrt{x\hat{\Sigma}x} + r\sqrt{1+\alpha^2}\,\|x\|_2\,.$$

α available analytically if $\mathcal{S} = \{\text{all distributions}\}$ and $\mathcal{R}_{\xi\sim\mathbb{P}}$ is

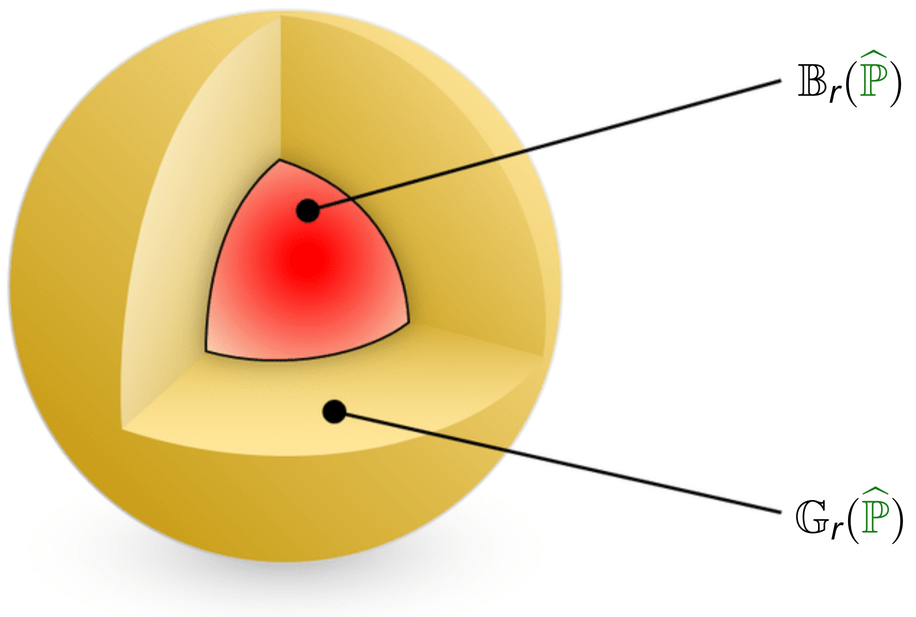
- any **spectral** risk measure;
- any risk measure with a **Kusuoka representation**;
- any **distortion** risk measure.

[1] Nguyen, Shafieezadeh-Abadeh, Filipovic & Kuhn, *Working Paper*, 2022.

Scalable Wasserstein DRO

Conservative approximation:

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right] \leq \sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]$$

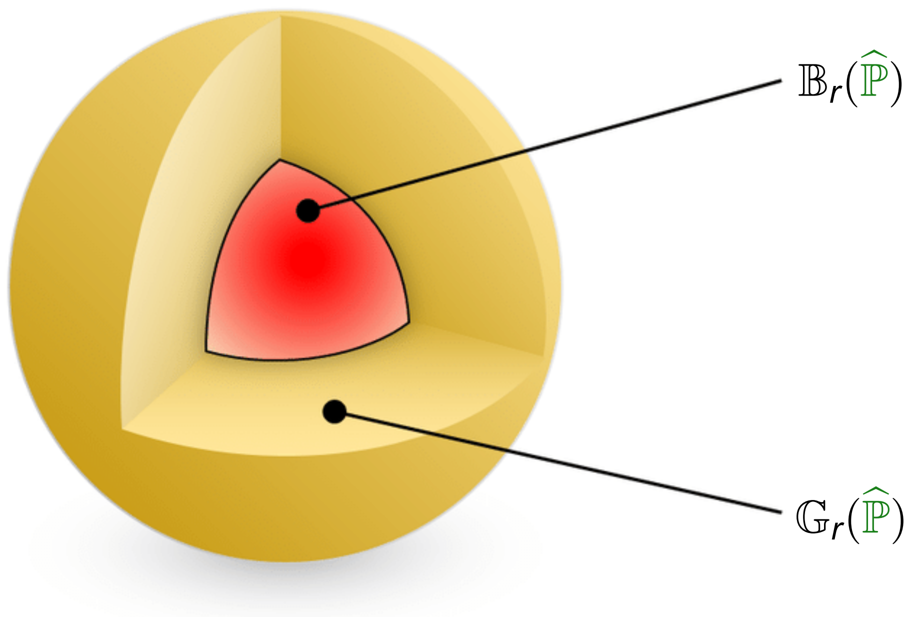


Scalable Wasserstein DRO

Conservative approximation:

$$\underbrace{\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right] \leq \sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]}$$

$$\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right] \leq \underbrace{\sup_{\mathbb{P}\in\mathcal{S}\cap\mathbb{G}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi\sim\mathbb{P}}\left[\ell(x,\xi)\right]}_{\text{independent of sample size}}$$



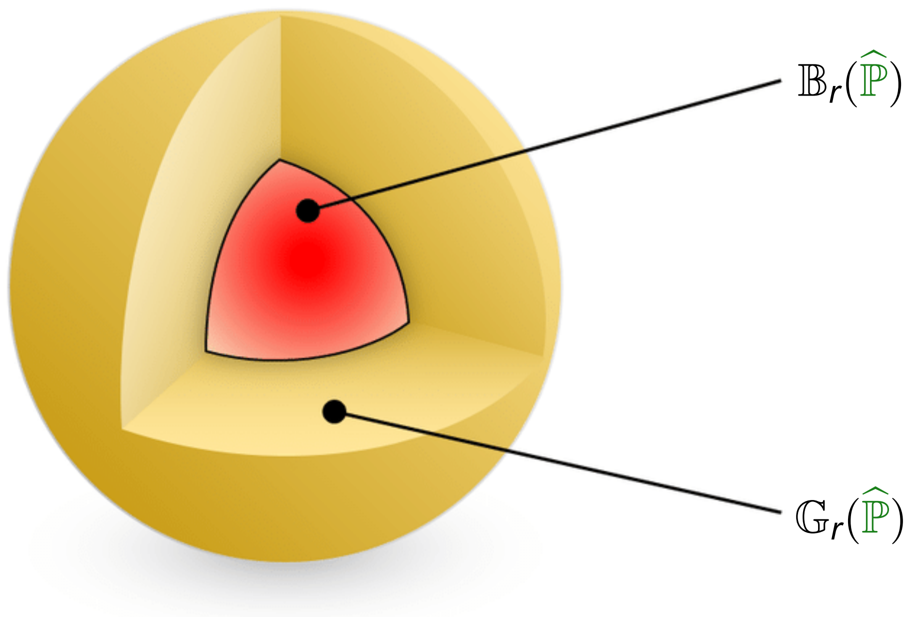
Scalable Wasserstein DRO

Conservative approximation:

$$\sup_{\mathbb{P} \in \mathcal{S} \cap \mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)] \leq \sup_{\mathbb{P} \in \mathcal{S} \cap \mathbb{G}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\xi \sim \mathbb{P}} [\ell(x, \xi)]$$

exact if…

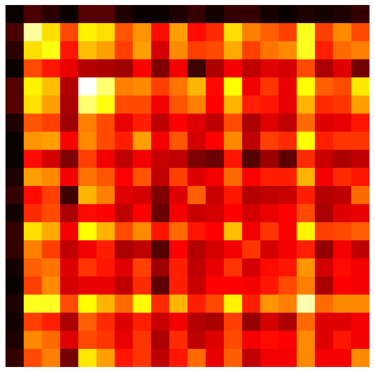
- $\ell(x, \xi)$ is quadratic in ξ or
- $\mathcal{S} = \{$all Gaussian distributions$\}$



Wasserstein DRO with Quadratic Loss

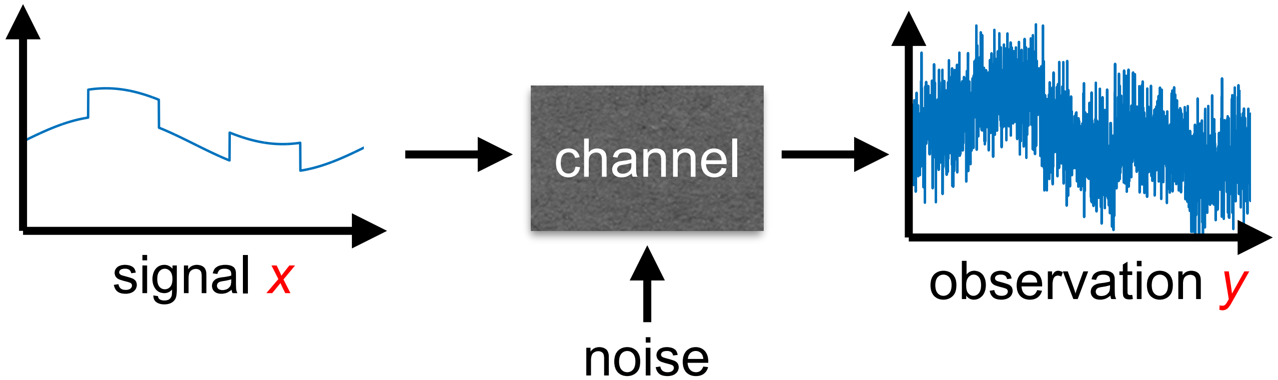
Inverse covariance estimation:[1)]

$$\inf_{X \succ 0} \; -\log\det X + \sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}\left[\xi^\top X \xi\right]$$

$X =$ 

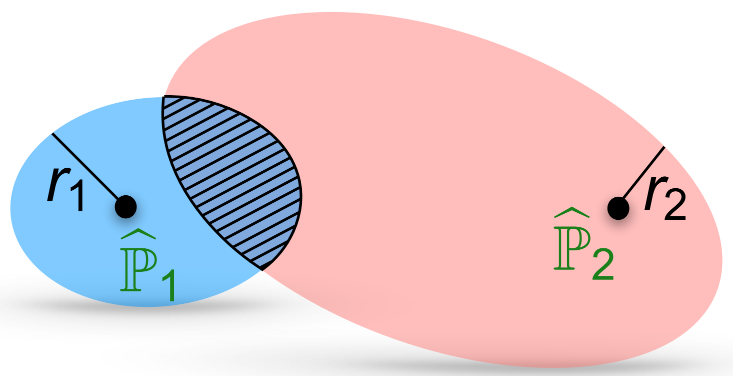
Signal processing & Kalman filtering:[2)]

$$\inf_{\psi(\cdot)} \sup_{\mathbb{P}\in\mathbb{B}_r(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{P}}\left[\|x-\psi(y)\|_2^2\right]$$



Domain adaptation:[3)]

$$\inf_{\beta} \sup_{\mathbb{P}\in\cap_k\mathbb{B}_{r_k}(\widehat{\mathbb{P}}_k)} \mathbb{E}_{\mathbb{P}}\left[(\beta^\top x - y)^2\right]$$



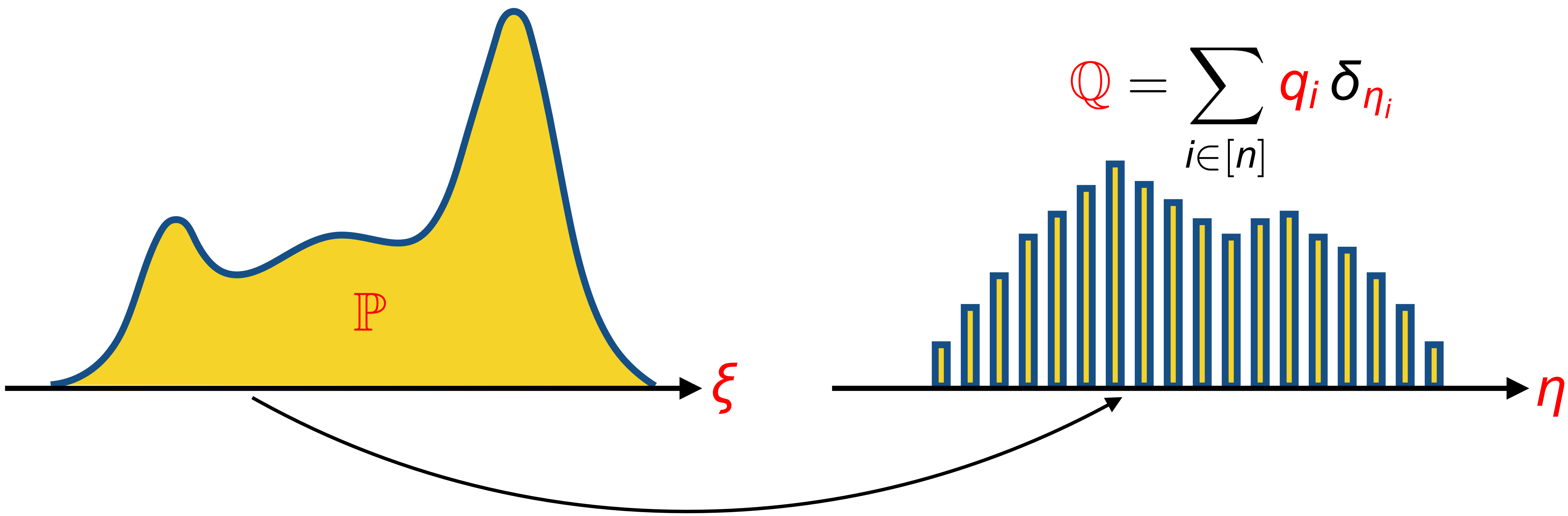
1) Nguyen, Kuhn & Mohajerin Esfahani, *Oper. Res.*, 2022.
2) Nguyen, Shafieezadeh-Abadeh, Kuhn & Mohajerin Esfahani, *Math. Oper. Res.*, 2023.
3) Taskesen, Yue, Blanchet, Kuhn & Nguyen, *ICML*, 2021.

Ask not what OT can do for DRO —
ask what DRO can do for OT.

Semi-Discrete OT

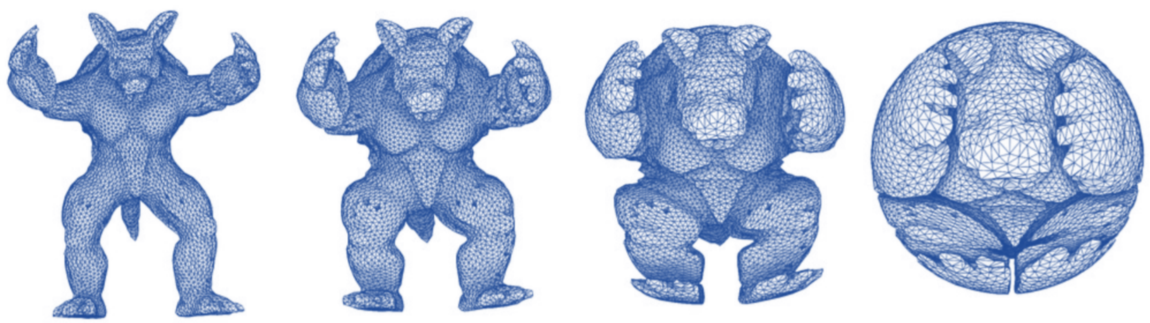
$$\min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(\xi, \eta) \sim \pi} \left[c(\xi, \eta) \right]$$

$$\mathbb{Q} = \sum_{i \in [n]} q_i \, \delta_{\eta_i}$$

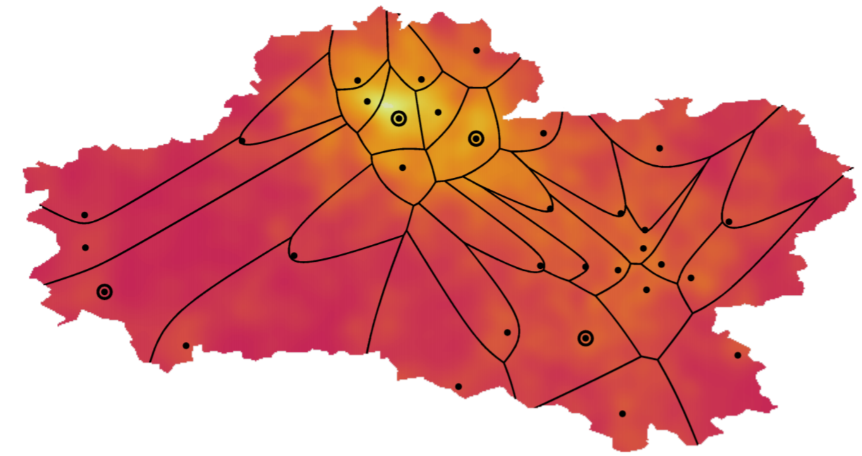


Applications of Semi-Discrete OT

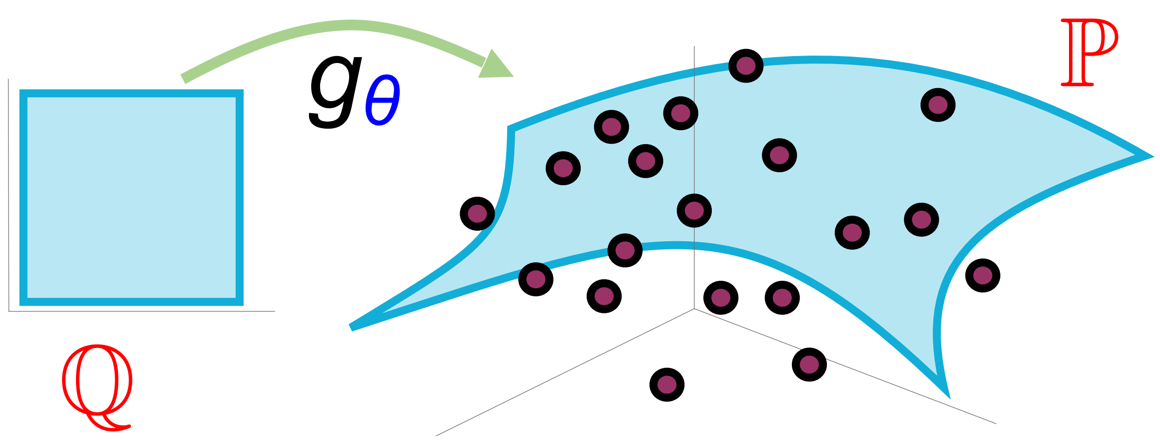
3D morphing[1)]



Resource allocation[2)]

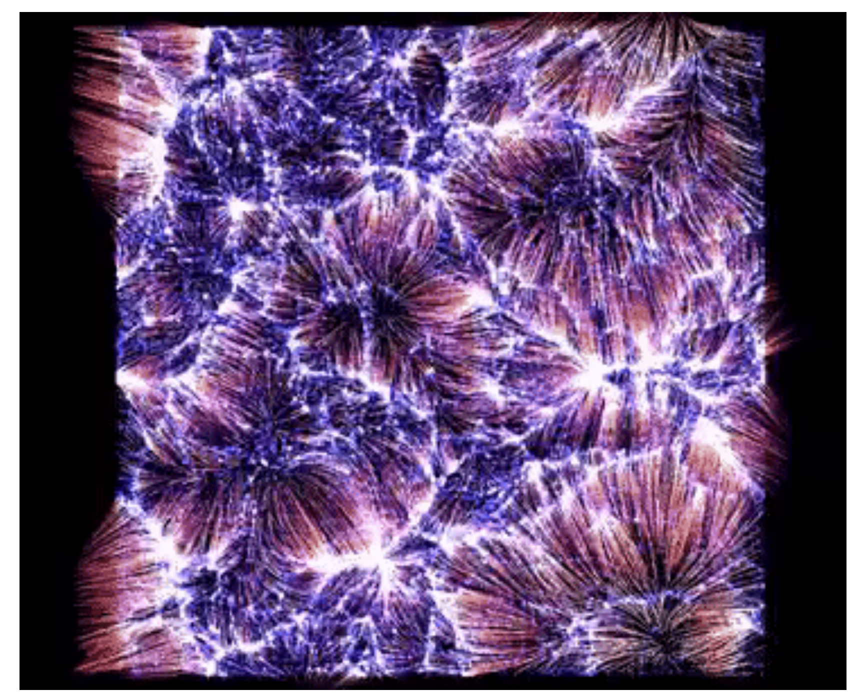


Generative models[3)]



$$\min_{\theta} W(\mathbb{P}, g_{\theta\#}\mathbb{Q})$$

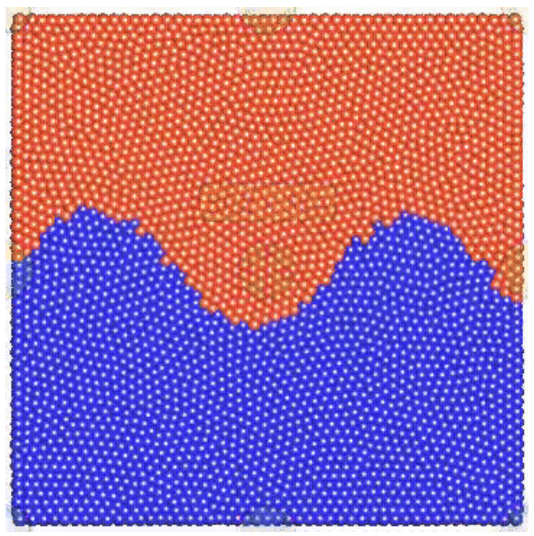
Reconstruction of the early universe[4)]



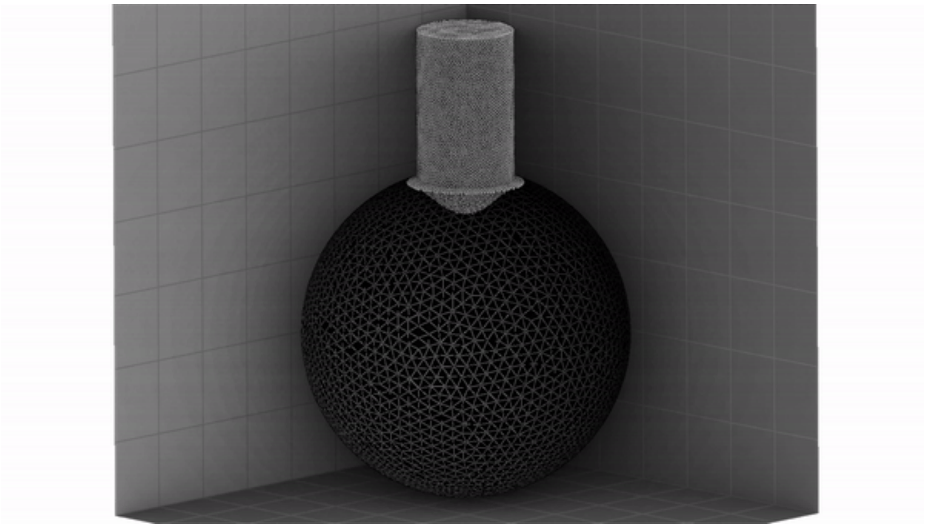
[1)] Lévy, *ESAIM: M2AN,* 2014.
[2)] Hartmann & Schuhmacher, *Math. Meth. Oper. Res.,* 2020.
[3)] Arjovsky, Chintala & Bottou, *ICML,* 2017.
[4)] Lévy, Mohayaee, von Hausegger & Natarajan, *MNRAS*, 2021.

Applications of Semi-Discrete OT

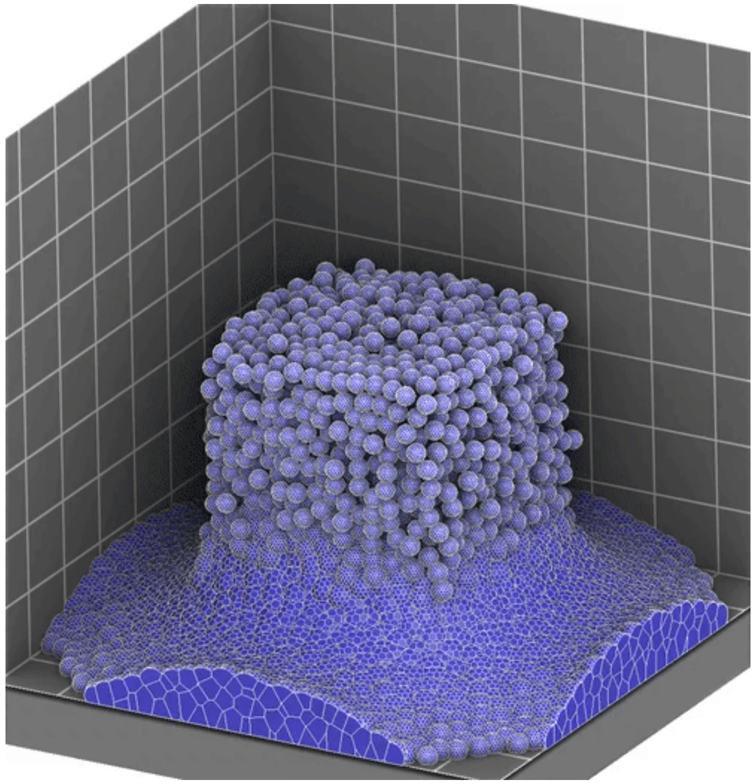
Taylor-Rayleigh instability using the Gallouet-Merigot scheme[1)]



Simulation of an incompressible bi-phasic flow in a bottle[1)]



Free-surface fluid simulation with Gallouet-Merigot scheme[1)]



1) Goes, Wallez & Huang, *ACM Trans. Graph.*, 2015; Gallouet & Merigot, *Found. Comput. Math.*, 2017; Lévy, *arXiv*, 2018.

Duality

primal OT: $$\min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(\xi, \eta) \sim \pi} \left[c(\xi, \eta) \right]$$

Duality

primal OT: $$\min_{\pi \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(\xi, \eta) \sim \pi} [c(\xi, \eta)]$$

dual OT: $$\sup_{\phi \in \mathbb{R}^n} \sum_{i \in [n]} q_i \phi_i - \mathbb{E}_{\xi \sim \mathbb{P}} [\psi_c(\phi, \xi)]$$

Duality

<u>primal OT</u>:

$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

<u>dual OT</u>:

$$\sup_{\phi \in \mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\boxed{\psi_c(\phi,\xi)}\right]$$

$$\psi_c(\phi,\xi) = \max_{i\in[n]} \phi_i - c(\xi,\eta_i)$$

Relation to Discrete Choice Theory

primal OT: $$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

dual OT: $$\sup_{\phi \in \mathbb{R}^n} \sum_{i \in [n]} q_i \phi_i - \mathbb{E}_{\xi \sim \mathbb{P}} \boxed{\psi_c(\phi,\xi)}$$

Luis Thurstone
1927

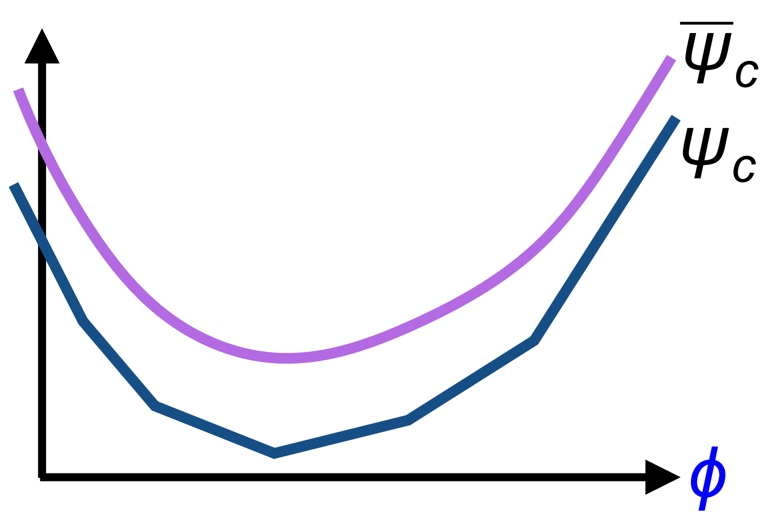
Daniel McFadden
1978

$$\underbrace{\psi_c(\phi,\xi) = \max_{i \in [n]} \phi_i - c(\xi,\eta_i)}_{\text{deterministic discrete choice model}}$$

Relation to Discrete Choice Theory

primal OT:
$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

dual OT:
$$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\psi_c(\phi,\xi)\right]$$



$$\underbrace{\overline{\psi}_c(\phi,\xi) = \sup_{\theta\in\Theta} \mathbb{E}_{z\sim\theta}\left[\max_{i\in[n]} \phi_i - c(\xi,\eta_i) + z_i\right]}_{\text{distributionally robust discrete choice problem}^{1)}}$$

distributionally robust
discrete choice problem[1)]

1) Natarajan, Song & Teo, *Manag. Sci.*, 2009.

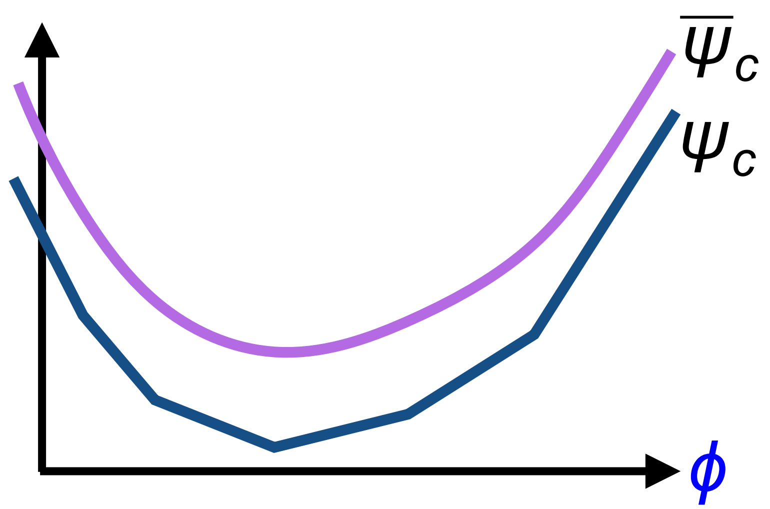
Relation to Discrete Choice Theory

primal OT:

$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

dual OT:

$$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\psi_c(\phi,\xi)\right]$$



$$\overline{\psi}_c(\phi,\xi) = \sup_{\theta\in\boxed{\Theta}} \mathbb{E}_{z\sim\theta}\left[\max_{i\in[n]} \phi_i - c(\xi,\eta_i) + z_i\right]$$

Frechet ambiguity set[1)]

$$\Theta = \{\theta \mid \theta(z_i \le s) \equiv F_i(s)\ \forall i \in [n]\}$$

---

1) Natarajan, Song & Teo, *Manag. Sci.*, 2009.

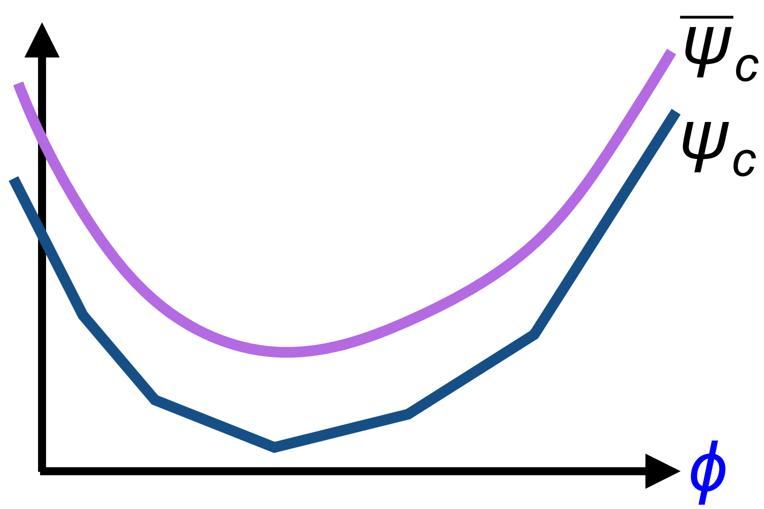
Smoothing the Dual Objective

primal OT:

$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \; \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

dual OT:

$$\sup_{\phi \in \mathbb{R}^n} \; \sum_{i\in[n]} q_i \phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\psi_c(\phi,\xi)\right]$$



$$\overline{\psi}_c(\phi,\xi) = \sup_{\theta\in\Theta} \; \mathbb{E}_{z\sim\theta}\left[\max_{i\in[n]} \phi_i - c(\xi,\eta_i) + z_i\right]$$

Smoothing the Dual Objective

primal OT:

$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right]$$

smooth dual OT:

$$\sup_{\phi \in \mathbb{R}^n} \sum_{i\in[n]} q_i \phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$

Regularizing the Primal Objective

<u>regularized primal OT</u>:

$$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right] + D_f(\pi \| \mathbb{P}\otimes\mathbb{Q})$$

<u>smooth dual OT</u>:

$$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$

Duality Revisited

regularized primal OT: $$\min_{\pi\in\Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right] + D_f(\pi\|\mathbb{P}\otimes\mathbb{Q})$$

smooth dual OT: $$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$

Theorem:[1)] Smooth dual OT is equivalent to regularized primal OT if $F_i(s) = \mathrm{proj}_{[0,1]}(1 - q_i F(-s))$ for all $i \in [n]$ and $f(s) = \int_0^s F^{-1}(t)\,\mathrm{d}t$.

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Duality Revisited

<u>regularized primal OT</u>: $$\min_{\pi \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(\xi,\eta)\sim\pi}\left[c(\xi,\eta)\right] + D_f(\pi \| \mathbb{P}\otimes\mathbb{Q})$$

<u>smooth dual OT</u>: $$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$

Theorem:[1] Smooth dual OT is equivalent to regularized primal OT if $F_i(s) = \mathrm{proj}_{[0,1]}(1 - q_i \boxed{F}(-s))$ for all $i \in [n]$ and $f(s) = \int_0^s \boxed{F}^{-1}(t)\,\mathrm{d}t$.

marginal generator

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Unification of Regularization Schemes

F	Regularizer
Exponential	KL divergence[1)]
Uniform	χ^2-divergence[2)]
Pareto	Tsallis divergence[3)]
Hyperbolic cosine	Hyperbolic divergence
t-distribution	Chebychev

[1)] Cuturi, *NeurIPS*, 2013; Genevay, Cuturi, Peyré & Bach, *NeurIPS*, 2016.
[2)] Blondel, Seguy & Rolet, *AISTATS*, 2018; Seguy, Damodaran, Flamary *et. al.*, *ICLR*, 2018.
[3)] Muzellec, Nock, Patrini & Nielsen, *AAAI*, 2017.

Inexact Averaged SGD[1)]

smooth dual OT: $$\sup_{\phi \in \mathbb{R}^n} \sum_{i \in [n]} q_i \phi_i - \mathbb{E}_{\xi \sim \mathbb{P}} \left[\overline{\psi}_c(\phi, \xi) \right]$$

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Inexact Averaged SGD[1)]

smooth dual OT: $$\sup_{\phi \in \mathbb{R}^n} \sum_{i \in [n]} q_i \phi_i - \mathbb{E}_{\xi \sim \mathbb{P}} \boxed{\overline{\psi}_c(\phi, \xi)}$$

unbiased stochastic gradient ⇔ optimal choice probabilities

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Inexact Averaged SGD[1)]

smooth dual OT: $\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}} \boxed{\overline{\psi}_c(\phi, \xi)}$

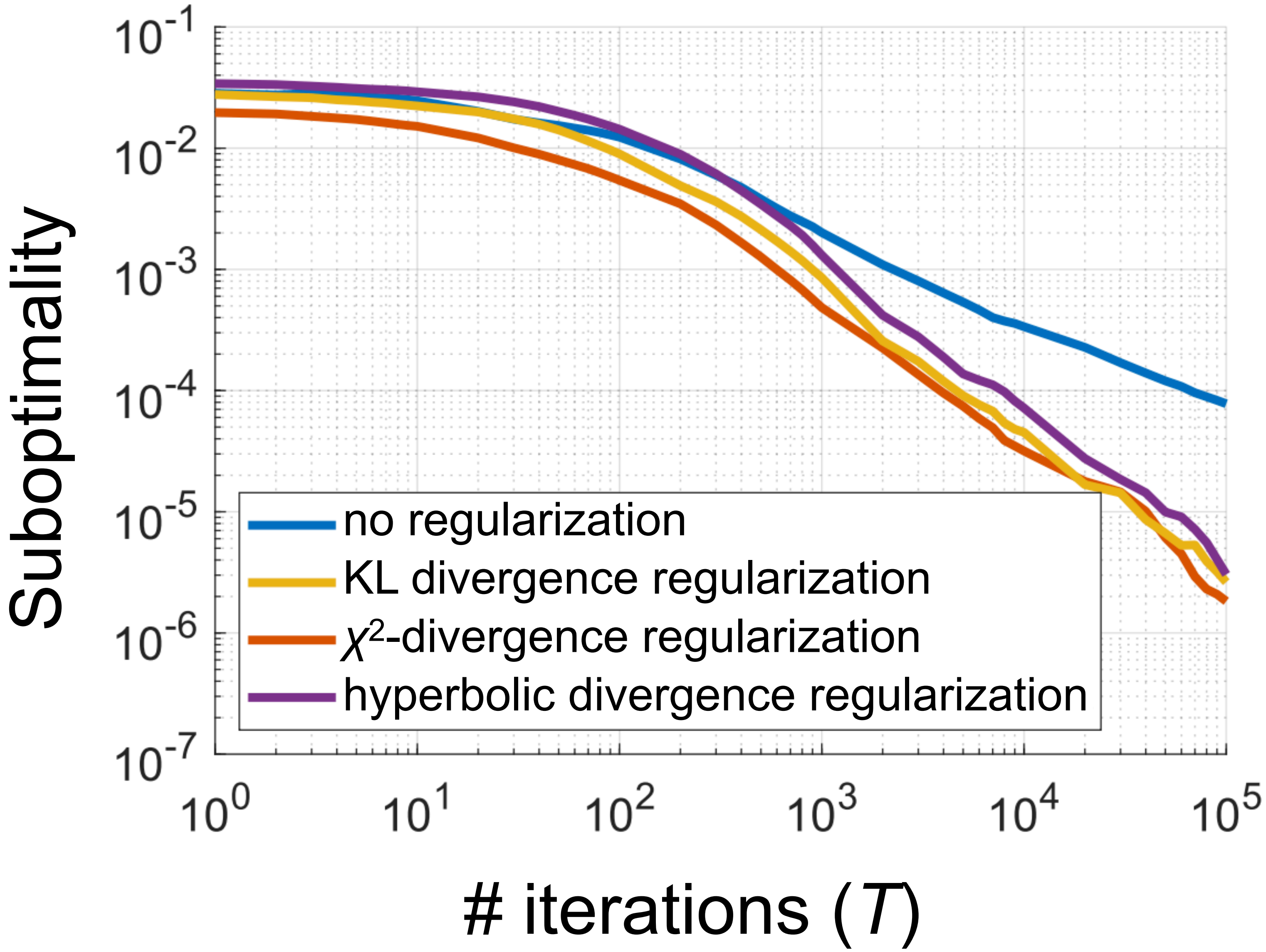
unbiased stochastic gradient ⇔ optimal choice probabilities

⇒ benefit from rich literature on discrete choice theory!

1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Inexact Averaged SGD[1)]

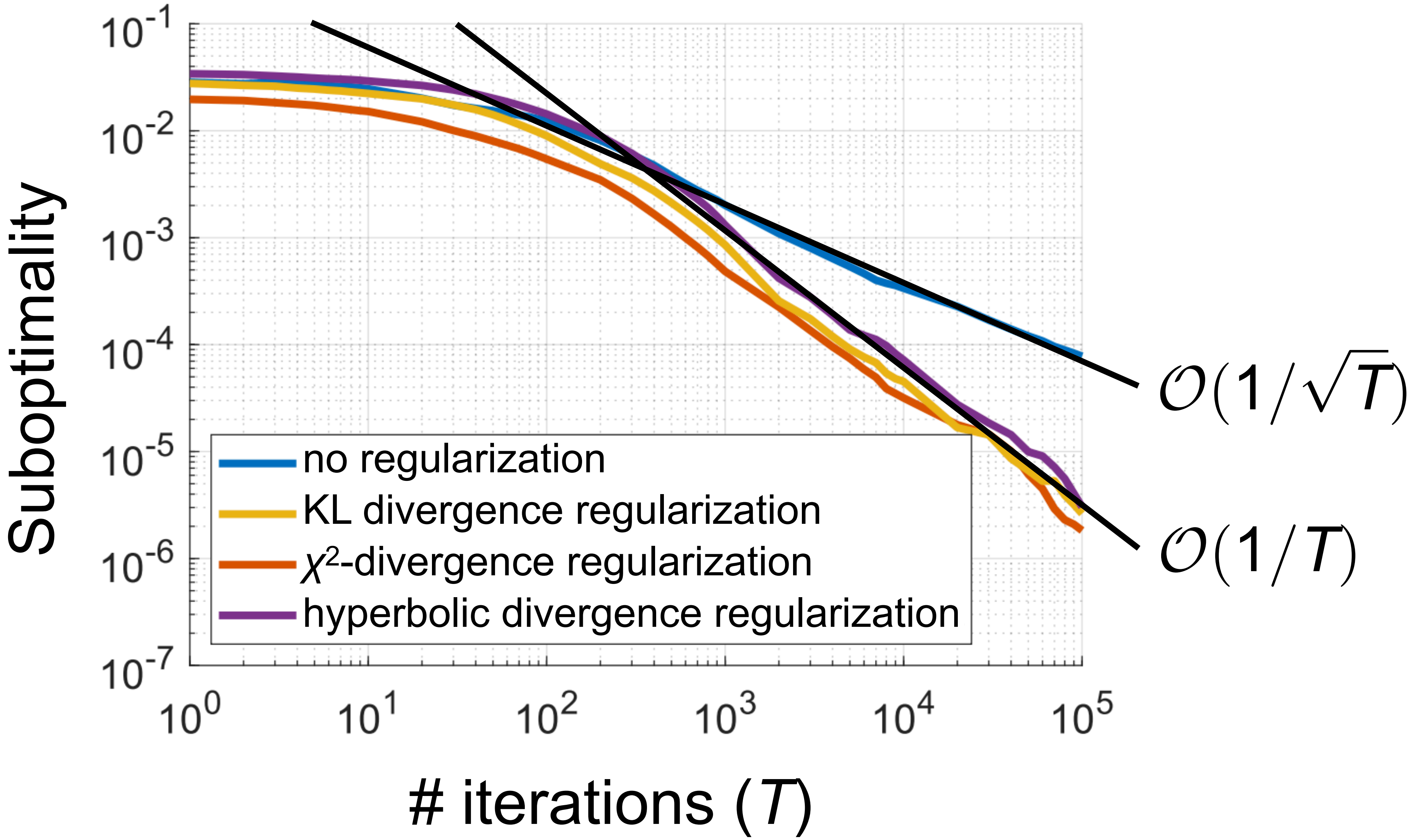
smooth dual OT: $$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$



1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Inexact Averaged SGD[1)]

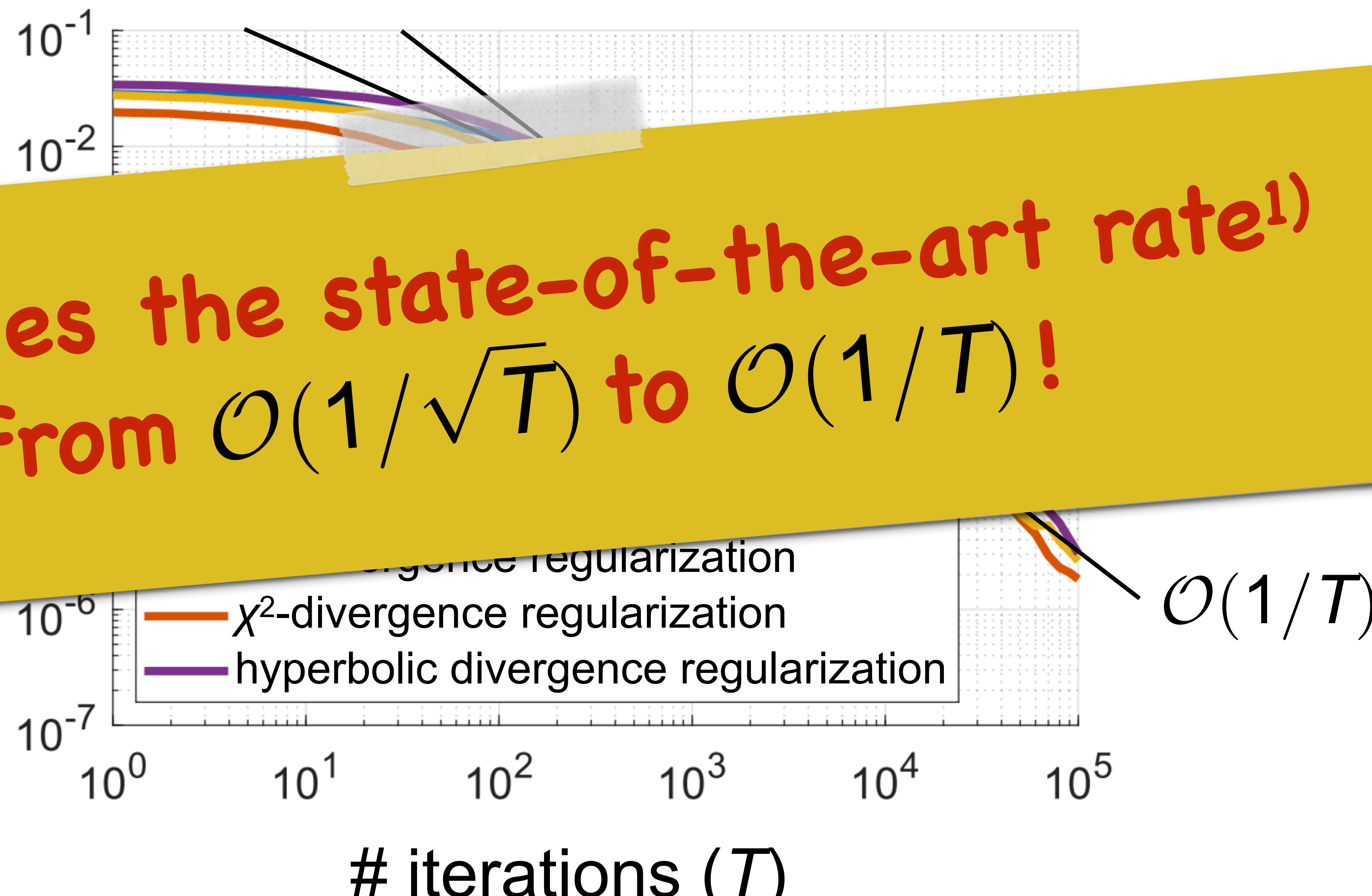
<u>smooth dual OT</u>: $$\sup_{\phi \in \mathbb{R}^n} \sum_{i \in [n]} q_i \phi_i - \mathbb{E}_{\xi \sim \mathbb{P}} \left[\overline{\psi}_c(\phi, \xi) \right]$$



1) Taskesen, Shafieezadeh-Abadeh & Kuhn, *Math. Program.*, 2023.

Inexact Averaged SGD

smooth dual OT: $$\sup_{\phi\in\mathbb{R}^n} \sum_{i\in[n]} q_i\phi_i - \mathbb{E}_{\xi\sim\mathbb{P}}\left[\overline{\psi}_c(\phi,\xi)\right]$$



Improves the state-of-the-art rate[1)] from $\mathcal{O}(1/\sqrt{T})$ to $\mathcal{O}(1/T)$!

1) Genevay, Cuturi, Peyré & Bach, *NeurIPS*, 2016.

Summary

- **Wasserstein DRO**
  - offers **statistical guarantees**
  - **tractable** in generic settings
  - provides probabilistic justification of **regularization**
  - leads to **adversarial examples** that can deceive humans

- **Gelbrich DRO**
  - highly **scalable approximation** using 1^{st} and 2^{nd} moments
  - exact for **quadratic** loss functions or **Gaussian** distributions

- **Semi-Discrete OT**
  - **#P-hard** but amenable to **fast SGD algo.** by adding noise
  - **dual smoothing** equivalent to **primal regularization**

This Talk is Based on…

[1] D. Kuhn, P. Mohajerin Esfahani, V. Nguyen & S. Shafieezadeh-Abadeh. **Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning**. *INFORMS TutORials in Operations Research*. 2019.

[2] P. Mohajerin Esfahani & D. Kuhn. **Data-Driven Distributionally Robust Optimization using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations**. *Mathematical Programming* 171(1–2), 115–166, 2018.

[3] V. Nguyen, D. Kuhn & P. Mohajerin Esfahani. **Distributionally Robust Inverse Covariance Estimation: The Wasserstein Shrinkage Estimator**. *Operations Research* 70(1), 490–515, 2018.

[4] V. Nguyen, S. Shafieezadeh-Abadeh, D. Filipovic & D. Kuhn. **Mean-Covariance Robust Risk Measurement**. Working paper, 2021.

[5] V. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn & P. Mohajerin Esfahani. **Bridging Bayesian and Minimax Mean Square Error Estimation via Wasserstein Distributionally Robust Optimization**. *Mathematics of Operations Research* 48 (1), 1-37, 2023.

[6] S. Shafieezadeh-Abadeh, L. Aolaritei, F. Dörfler & D. Kuhn. **Optimal Transport for Distributionally Robust Optimization: Nash Equilibria, Regularization, and Computation**. Working paper, 2023.

[7] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani & D. Kuhn. **Distributionally Robust Logistic Regression**. *Neural Information Processing Systems*, 2015.

[8] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani & D. Kuhn. **Regularization via Mass Transportation**. *Journal of Machine Learning Research* 20(103), 1–68, 2019.

[9] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, V. Nguyen & D. Kuhn. **Wasserstein Distributionally Robust Kalman Filtering**. *Neural Information Processing Systems*, 2018.

[10] B. Taşkesen, M.-C. Yue, J. Blanchet, D. Kuhn & V. Nguyen. **Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts**. *International Conference on Machine Learning*, 2021.

[11] B. Taşkesen, S. Shafieezadeh-Abadeh & D. Kuhn. **Semi-Discrete Optimal Transport: Hardness, Regularization and Numerical Solution**. *Mathematical Programming* 199 (1-2), 1033-1106, 2023.

[12] J. (T.) Zhen, D. Kuhn & W. Wiesemann. **A Unified Theory of Robust and Distributionally Robust Optimization via the Primal-Worst-Equals-Dual-Best Principle**. *Operations Research*, 2023.