Gabriele Steidl
Jannis Chemseddine
TU Berlin

Lecture 1: Optimal Transport
Lecture 2: Generative Flows
Lecture 3: Bayesian Inverse Problems
Lecture 4: Experiments

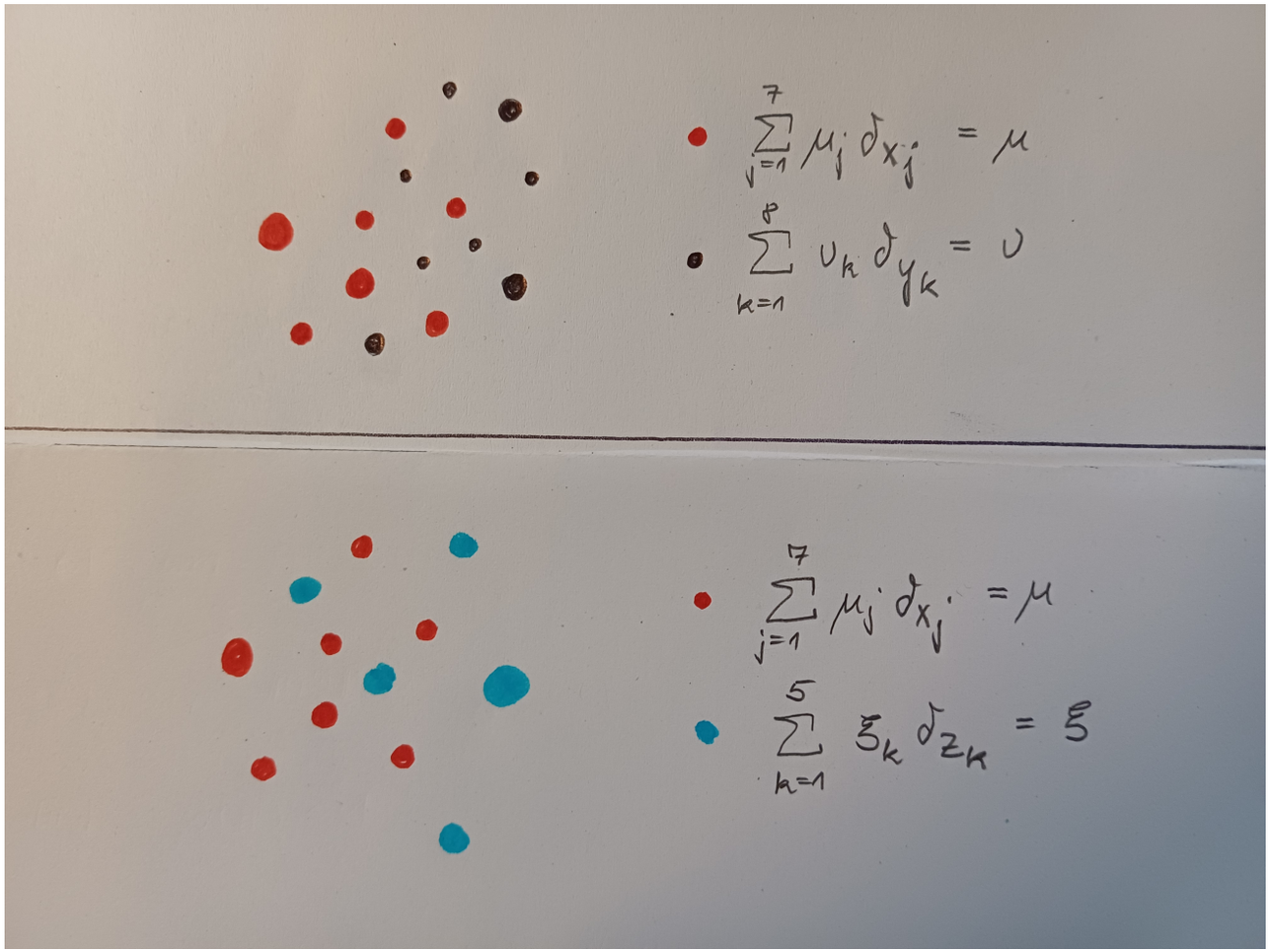
Berlin Mathematics Research Center MATH+

Funded under Germany's Excellence Strategy by DFG Deutsche Forschungsgemeinschaft

Motivation

Question:

- What is the best way to transport the mass from μ to ν and what costs such a transport?
- Is ν or ξ nearer to μ?



1. Optimal Transport (OT)

Technische Universität Berlin

- $\mathcal{M}(\mathbb{R}^d)$ is a Banach space with **total variation norm**

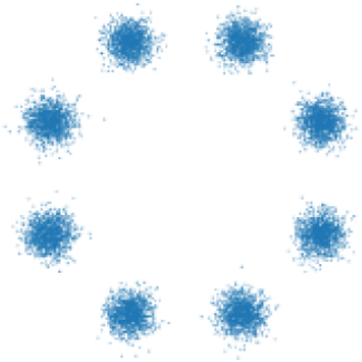
$$\|\mu\|_{TV} := |\mu|(\mathbb{R}^d) \quad \text{where} \quad |\mu|(A) := \sup_{A=\bigcup A_k, A_i \cap A_j = \emptyset} \sum_{k=1}^{n} |\mu(A_k)|$$

- pre-dual space of $\mathcal{M}(\mathbb{R}^d)$ is $C_0(\mathbb{R}^d)$, i.e. $C_0(\mathbb{R}^d)' = \mathcal{M}(\mathbb{R}^d)$

$$\|\mu\|_{TV} = \sup_{\|\varphi\|_\infty \leq 1} |\langle \varphi, \mu \rangle|, \quad \langle x, \mu \rangle := \int_{\mathbb{R}^d} \varphi(x)\, d\mu(x)$$

- $\mathcal{P}(\mathbb{R}^d)$ subset of probability measures on $\mathbb{R}^d$

Samples from a probability measure:



$d = 2$

$d = 28^2 = 784$

TV-norm not a good measure in $\mathcal{P}(\mathbb{R}^d)$

Examples: $\mathcal{P}(\mathbb{R}^d)$

1. Atomic measures (empirical measures if same weights)

$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \delta_x(A) := \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise} \end{cases}$$

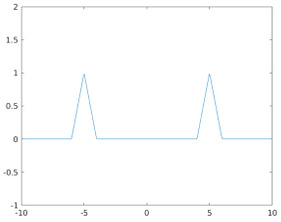
$$\|\mu\|_{TV} = |\mu_1| + \ldots + |\mu_m| = 1$$

2. Absolutely continuous measures $\mu \in \mathcal{M}(\mathbb{R}^d)$ with density $\varphi \in L_1(\mathbb{R}^d)$

$$\mu(A) = \int_A \varphi(x)\,\mathrm{d}x$$

$$\|\mu\|_{TV} = \int_{\mathbb{R}^d} |\varphi(x)|\,\mathrm{d}x = 1$$

- Also, if μ and ν have disjoint supports $\|\mu - \nu\|_{\mathrm{TV}} = 2$



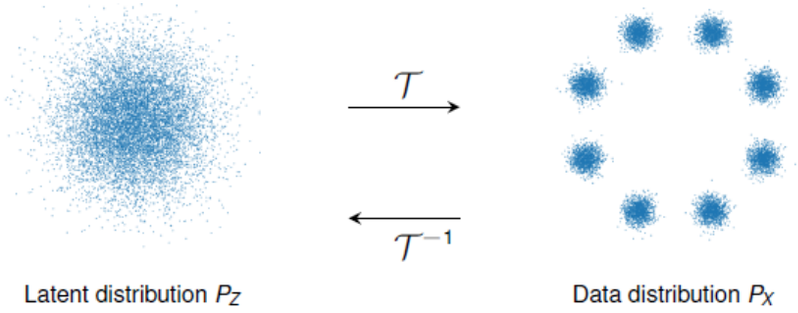
2. Monge Problem (1781)

Given $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\nu \in \mathcal{P}(\mathbb{R}^d)$
Find an optimal transport map $\hat{T} : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\hat{T} \in \underset{T\text{measurable}}{\operatorname{argmin}} \int_X c\left(x, T(x)\right) \, \mathrm{d}\mu(x) \quad \text{subject to} \quad \nu = T_{\#}\mu$$

with the push forward measure

$$T_{\#}\mu := \mu \circ T^{-1}$$



- $\int_{T^{-1}(A)} f(T(x))\mathrm{d}\mu(x) = \int_A f(y)\mathrm{d}\underbrace{(T_{\#}\mu)}_{\nu}(y)$
- in case of density $\mu = p_\mu \mathrm{d}\lambda$ and a diffeomorphism T:

$$p_{T_{\#}\mu}(y) = p_\mu\left(T^{-1}(y)\right)\left|\det \nabla T^{-1}(y)\right|$$

Discrete Monge Problem

Given

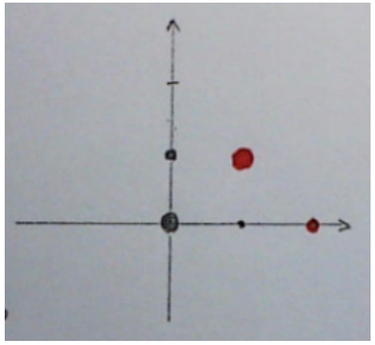
$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$$

Find an optimal transport map $\hat{T} : \{x_1, \ldots, x_m\} \to \{y_1, \ldots, y_n\}$ such that

$$\hat{T} \in \operatorname*{argmin}_T \sum_{i=1}^{N} c(x_i, T(x_i))\mu_i \quad \text{s.t.} \quad \nu_j = \sum_{T(x_i)=y_j} \mu_i$$

Example: $x_1 = (0,0)$, $x_2 = (1,0)$, $x_3 = (0,1)$ and $y_1 = (1,1)$, $y_2 = (2,0)$

Measures: $\mu = 3\delta_{x_1} + 1\delta_{x_2} + 2\delta_{x_3}$, $\nu = 4\delta_{y_1} + 2\delta_{y_2}$



Costs: $c(x,y) = \|x - y\|^2$
Only possible map: $T(x_1) = y_1$, $T(x_2) = y_1$, $T(x_3) = y_2$

Problem: Opposite task, i.e. changing the role of μ and ν, has no solution T

Continuous Monge Problem

Example:For two Gaussians $\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$ and $\nu = \mathcal{N}(m_\nu, \Sigma_\nu)$, where $\mathcal{N}(m, \Sigma)$ has the density

$$p(x) := (2\pi)^{-\frac{d}{2}} (\det \Sigma)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(x-m)^\mathsf{T}\Sigma^{-1}(x-m)},$$

the transport map is given by

$$T(x) = m_\nu + A(x - m_\mu), \quad A := \Sigma_\mu^{-\frac{1}{2}} \left(\Sigma_\mu^{\frac{1}{2}} \Sigma_\nu \Sigma_\mu^{\frac{1}{2}}\right) \Sigma_\mu^{-\frac{1}{2}},$$

Kantorovich Problem (1942)

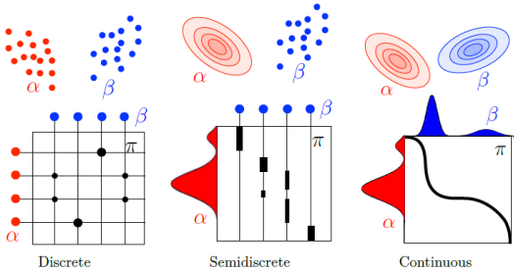
Given $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\nu \in \mathcal{P}(\mathbb{R}^d)$
Find an optimal transport plan $\hat{\alpha} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with given marginals μ and ν

$$\hat{\pi} \in \operatorname*{argmin}_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, \mathrm{d}\alpha(x,y)$$

$$\mathrm{OT}(\mu,\nu) = \min_{\alpha \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, \mathrm{d}\alpha(x,y)$$

where

- $\Pi(\mu,\nu) := \{\alpha \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : (P_1)_{\#}\alpha = \mu,\ (P_2)_{\#}\alpha = \nu$
- $P_1(x_1,x_2) = x_1,\ P_2(x_1,x_2) = x_2$



: Book Peyré/Cuturi: Computational optimal transport

- Existence of a minimizer is ensured if c is lsc and bounded from below

Discrete Kantorovich Problem

Given

$$\mu = \sum_{i=1}^{m} \mu_i \delta_{x_i}, \quad \nu = \sum_{j=1}^{n} \nu_j \delta_{y_j}$$

Find an optimal transport plan $\hat{\alpha} = \sum_{i,j=1}^{m,n} \hat{\alpha}_{ij} \delta_{x_i, y_j}$, i.e. $\alpha = (\alpha_{ij})_{i,j} \in \mathbb{R}_{\geq 0}^{m,n}$ such that

$$\hat{\alpha} \in \operatorname*{argmin}_{\alpha} \sum_{i=1}^{m} \sum_{j=1}^{n} c(x_i, y_j) \alpha_{ij} = \operatorname*{argmin}_{\alpha} \langle c, \alpha \rangle$$

subject to $\alpha_{ij} \geq 0$ and

$$\sum_{i=1}^{m} \alpha_{ij} = \nu_j, \quad j = 1, \ldots, n, \qquad \alpha^\top 1_m = \nu$$

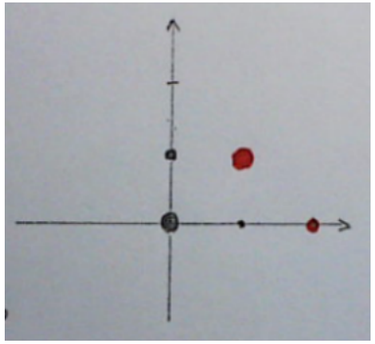
$$\sum_{j=1}^{n} \alpha_{ij} = \mu_i, \quad i = 1, \ldots, m \qquad \alpha 1_n = \mu$$

	ν_1	$\ldots$	ν_j	$\ldots$	ν_m
μ_1	$\alpha_{1,1}$	$\ldots$	$\alpha_{1,j}$	$\ldots$	$\alpha_{1,m}$
$\vdots$			$\vdots$		$\vdots$
μ_i	$\alpha_{i,1}$	$\ldots$	$\alpha_{i,j}$	$\ldots$	$\alpha_{i,m}$
$\vdots$			$\vdots$		$\vdots$
μ_n	$\alpha_{n,1}$	$\ldots$	$\alpha_{n,j}$	$\ldots$	$\alpha_{n,m}$

Examples with $c(x, y) = \|x - y\|^2$

Example 1: $x_1 = (0, 0)$, $x_2 = (1, 0)$, $x_3 = (0, 1)$ and $y_1 = (1, 1)$, $y_2 = (2, 0)$

Measures: $\mu = 3\delta_{x_1} + 1\delta_{x_2} + 2\delta_{x_3}$, $\nu = 4\delta_{y_1} + 2\delta_{y_2}$



Optimal map: $T(x_1) = y_1$, $T(x_2) = y_1$, $T(x_3) = y_2$

Optimal plan:

	y_1	y_2
x_1	2	4
x_1	1	1
x_3	1	5

costs

	4	2
3	2	1
1	0	1
2	2	0

optimal plan $\hat{\alpha}$

	4	2
3	3	0
1	1	0
2	0	2

$\alpha = (\mathrm{Id}, T)_{\#}\mu$

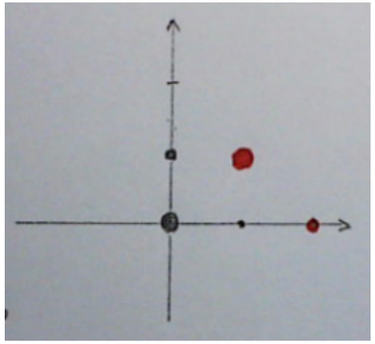
$\mathrm{OT}(\mu, \nu) = 11$

$\langle c, \alpha \rangle = 17$

Examples with $c(x, y) = \|x - y\|^2$

Example 2: $x_1 = (0, 0)$, $x_2 = (1, 0)$, $x_3 = (0, 1)$ and $y_1 = (1, 1)$, $y_2 = (2, 0)$

Measures: $\mu = 3\delta_{x_1} + 1\delta_{x_2} + 2\delta_{x_3}$, $\nu = 2\delta_{y_1} + 4\delta_{y_2}$



Optimal map: $T(x_1) = y_2$, $T(x_2) = y_2$, $T(x_3) = y_1$

Optimal plan:

	y_1	y_2
x_1	4	2
x_1	1	1
x_3	5	1

costs

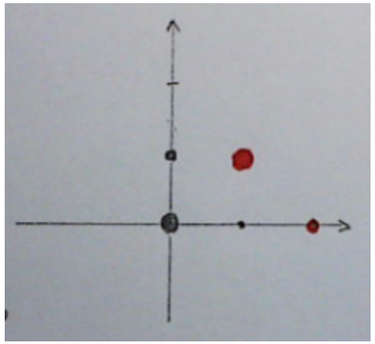
	2	4
3	0	3
1	0	1
2	2	0

optimal plan $\hat{\alpha} = (\mathrm{Id}, T)_{\#}\mu$

Examples with $c(x, y) = \|x - y\|^2$

Example 3: $x_1 = (1, 1)$, $x_2 = (2, 0)$ and $y_1 = (0, 0)$, $y_2 = (1, 0)$, $y_3 = (0, 1)$ and $y_1 = (1, 1)$, $y_2 = (2, 0)$

Measures: $\mu = 2\delta_{x_1} + 4\delta_{x_2}$, $\nu = 3\delta_{y_1} + 1\delta_{y_2} + 2\delta_{y_3}$



Optimal map: does not exists

Optimal plan:

	y_1	y_2	y_3
x_1	4	1	5
x_2	2	1	1

costs

	3	1	2
2	0	0	2
4	3	1	2

optimal plan $\hat{\alpha}$

Wasserstein Spaces

For $p \in [1, \infty)$, the space of measures with finite p-th moment is defined by

$$\mathcal{P}_p(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^p \mathrm{d}\mu(x) < \infty \right\}.$$

For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the Wasserstein p-distance is given by

$$W_p^p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \mathrm{d}\alpha(x, y),$$

which actually defines a metric.

The metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$ is called the p-th Wasserstein space.

Convergence of $(\mu_n)_n$, $\mu_n \in \mathcal{P}_2(\mathbb{R}^d)$: $W_2(\mu_n, \mu) \to 0$ as $n \to \infty$ if and only if we have weak/narrow convergence

$$\mu_n \overset{*}{\rightharpoonup} \mu \text{ in } C_b(X)'$$

and

$$\int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu_n(x) \quad \to \quad \int_{\mathbb{R}^d} \|x\|^2 \, \mathrm{d}\mu(\mathrm{x}) \quad \text{as } \mathrm{n} \to \infty$$

Brenier's Theorem

Theorem Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, where $\mu \ll \lambda$ and $c(x,y) = \|x-y\|^2$. Then

- Kantorovich problem has a unique solution $\hat{\pi}$
- $\hat{\pi} = (I, \hat{T})_{\#}\mu$, where $\hat{T} \in L^2_\mu(X,X)$ is the optimal transport map
- If ν has bounded support, then

$$\hat{T}(x) = x - \nabla\varphi(x) = \nabla\psi(x) \quad for\ \mu - a.e.\ x,$$

for some lower semi-continuous, convex, differentiable μ-a.e function ψ.

- Conversely, if ψ is lower semi-continuous, convex and differentiable μ-a.e. with $|\nabla\psi| \in L^2_\mu(X,X)$, then

$$T = \nabla\psi$$

is the optimal transport map from μ to $\nu = T_{\#}\mu \in \mathcal{P}_2(X)$.

Outline

1. Basic Notation: Spaces of Measures
2. Monge and Kantorovich Problem of OT
3. Dual OT
4. Regularized Optimal Transport

Discrete OT and its Dual

(P) $\operatorname{argmin}_\alpha \langle c, \alpha \rangle$ subject to $\alpha 1_n = \mu,\ \alpha^\mathsf{T} 1_m = \nu,\ \alpha \geq 0$ nm variables

(P) $\operatorname{argmin}_\alpha \langle c, \alpha \rangle$ subject to $\underbrace{(1_n \otimes I_m)}_{P_1} \alpha = \mu,\ \underbrace{(I_n \otimes 1_m)}_{P_2} \alpha = \nu, \alpha \geq 0$

$$P_1 = \begin{pmatrix} I_m & \dots & I_m \end{pmatrix} \in \mathbb{R}^{m \times nm}$$

$$P_2 = \begin{pmatrix} \underbrace{1 \cdots 1}_{m} & 0 & \cdots & 0 \\ 0 & \underbrace{1 \cdots 1}_{m} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underbrace{1 \cdots 1}_{m} \end{pmatrix} \in \mathbb{R}^{n \times nm},$$

(P) $\min_{\alpha \geq 0} \max_{\varphi, \psi} \langle c, \alpha \rangle + \langle \mu - P_1 \alpha, \varphi \rangle + \langle \nu - P_2 \alpha, \psi \rangle = \min_{\alpha \geq 0} \max_{\varphi, \psi} L(\alpha, \mu, \nu)$

(D) $\max_{\varphi, \psi} \min_{\alpha \geq 0} \langle c, \alpha \rangle + \langle \mu - P_1 \alpha, \varphi \rangle + \langle \nu - P_2 \alpha, \psi \rangle$

$$= \max_{\varphi, \psi} \min_{\alpha \geq 0} \langle c - P_1^\mathsf{T} \varphi - P_2^\mathsf{T} \psi, \alpha \rangle + \langle \mu, \varphi \rangle + \langle \nu, \psi \rangle =$$

$$= \max_{\varphi_i + \psi_j \leq c_{i,j}} \langle \mu, \varphi \rangle + \langle \nu, \psi \rangle \qquad n + m \text{ variables}$$

If $(c - P_1^\mathsf{T} \varphi - P_2^\mathsf{T} \psi)_{ij} = c_{ij} - \varphi_i - \psi_j < 0$, then first part $= -\infty$

If $(c - P_1^\mathsf{T} \varphi - P_2^\mathsf{T} \psi)_{ij} \geq 0$, then first $= 0$, where $\alpha_{ij} = 0$ if $c_{ij} > \varphi_i - \psi_j$

OT and its Dual (Continuous Setting)

Primal OT:

$$\mathrm{OT}(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} c\,\mathrm{d}\pi,$$

Dual OT:

$$\mathrm{OT}(\mu,\nu) = \max_{\substack{(\varphi,\psi)\in C(\mathbb{R}^d)^2 \\ \varphi(x)+\psi(y)\le c(x,y)}} \int_{\mathbb{R}^d} \varphi\,\mathrm{d}\mu + \int_{\mathbb{R}^d} \psi\,\mathrm{d}\nu.$$

$(\hat\varphi,\hat\psi) = (\hat\varphi,\hat\varphi^c)$ with *c-transformed function*

$$\varphi^c(y) = \min_{x\in\mathbb{R}^d}\{c(x,y)-\varphi(x)\}.$$

Special case: If $c(x,y)=|x-y|$, then $\varphi=-\psi$ is 1-Lipschitz (more general $c=d$ for metric space with distance d)

$$W_1(\mu,\nu) = \max_{\varphi\in\mathrm{Lip}_1(\mathbb{R}^d)} \int_{\mathbb{R}^d} \varphi\,\mathrm{d}\mu = \max_{\varphi\in\mathrm{Lip}_1(\mathbb{R}^d)} \langle\varphi,\mu\rangle$$

This plays a role in so-called Wasserstein GANs.

Compare to the dual notation of the TV norm:

$$\|\mu\|_{TV} = \sup_{\|\varphi\|_\infty\le 1} |\langle\varphi,\mu\rangle|$$

1. Optimal Transport (OT)

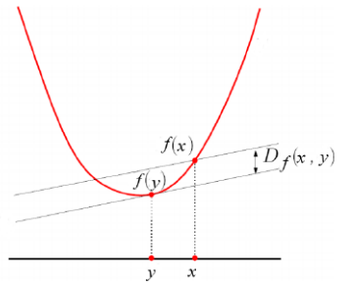
1. Spaces of Measures
2. Monge and Kantorovich Problem of OT
3. Discrete OT and its Dual
4. Continuous OT
5. Regularized Optimal Transport

Numerical computation of OT is time consuming for many samples $m = n$:

- Simplex algorithm $O(n^3 \log n)$
- Hungarian algorithm $O(n^3)$
- regularized Wasserstein $O(n^2 log n)$, critical for small regularization parameter ε
- Toolboxes:
  - Python Optimal Transport (POT, https://pythonot.github.io/): general toolbox for optimal transport including unbalanced optimal transport, sliced Wasserstein und Gromov-Wasserstein
  - Geomloss (https://www.kernel-operations.io/geomloss/): entropic optimal transport wit efficient GPU implementation and automatic differentiation

Bregman distance: $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ convex lsc and $\mathrm{dom} f \cap (0, +\infty) \neq \emptyset$

$$D_f(x, y) = f(x) - f(y) + \langle \nabla f(y), x - y \rangle$$



Function Name	$f(x)$	dom f	$D_f(x, y)$
Squared Norm	$\frac{1}{2}x^2$	$(-\infty, \infty)$	$\frac{1}{2}(x-y)^2$
Shannon Entropy	$x \log x$	$[0, \infty)$	$x \log \frac{x}{y} - x + y$
Bit Entropy	$x \log x + (1-x)\log(1-x)$	$[0, 1]$	$x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$
Burg Entropy	$-\log x$	$(0, \infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$

Examples:

1. $f(x) = \frac{1}{2}\|x\|^2$

$$D_f(x, y) = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 - \langle y, x - y \rangle = \frac{1}{2}\|x - y\|^2$$

2. $f(x) = \langle x, \log x \rangle$ (negative) Shannon entropy = maximal chaos

$$\begin{aligned} D_f(x, y) &= \langle x, \log x \rangle - \langle y, \log y \rangle - \langle \log y + 1, x - y \rangle \\ &= \langle x, \log x \rangle - \rangle x, \log y \rangle - \langle x, 1 \rangle + \langle y, 1 \rangle \\ &= \mathrm{KL}(x, y) \end{aligned}$$

Properties of Bregman Distances

Properties:

- $D_f(x, y) \geq 0$
- $D_f(x, y) = 0$ iff $x = y$ in case f is strictly convex.
- In general not symmetric and does not fulfill a triangular inequality
- Jointly convex, lsc.
- If f is strictly convex, D_f is strictly convex in the first argument.
- If expressions exist

$$\nabla_x D_f(x, y) = \nabla f(x) - \nabla f(y)$$

(P) $\quad \mathrm{OT}_\varepsilon = \min_{\alpha \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y) \mathrm{d}\alpha(x,y) + \varepsilon \mathrm{KL}(\alpha, \mu \otimes \nu)$

Alternatively, we can regularize with the entropy
$\mathrm{ent}(\alpha) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \alpha \log(\alpha) \, \mathrm{d}x(x,y)$

Discrete setting:

$$\begin{aligned}
\text{(P)} \qquad \mathrm{OT}_\varepsilon(\mu,\nu) &= \min_{\alpha \in \Pi(\mu,\nu)} \langle c, \alpha \rangle + \varepsilon \Big(\sum_{i,j} \alpha_{i,j} \log \alpha_{i,j} - \alpha_{i,j} \log(\mu_i \nu_j) - \alpha_{i,j} \Big) \\
&= \varepsilon \min_{\alpha \in \Pi(\mu,\nu)} \sum_{i,j} \alpha_{i,j} \log \alpha_{i,j} - \alpha_{i,j} \log(\mu_i \nu_j \mathrm{e}^{-c/\varepsilon}) - \alpha_{i,j} \\
&= \varepsilon \min_{\alpha \in \Pi(\mu,\nu)} \mathrm{KL}(\alpha, \underbrace{\mathrm{diag}(\mu) \, \mathrm{e}^{-c/\varepsilon} \, \mathrm{diag}(\nu)}_{K})
\end{aligned}$$

Rewriting this into the dual form leads to

$$\begin{aligned}
\text{(P)} \qquad & \min_{\alpha} \max_{\varphi,\psi} \mathrm{KL}(\alpha, K) + \langle \mu - P_1 \alpha, \varphi \rangle + \langle \nu - P_2 \alpha, \psi \rangle \\
\text{(D)} \qquad & \max_{\varphi,\psi} \min_{\alpha} \left\{ \mathrm{KL}(\alpha, K) - \langle \alpha, P_1^\mathsf{T} \varphi + P_2^\mathsf{T} \psi \rangle \right\} + \langle \mu, \varphi \rangle + \langle \nu, \psi \rangle \\
& = \max_{\varphi,\psi} \langle K, \mathrm{e}^{-P_1^\mathsf{T} \varphi - P_2^\mathsf{T} \psi} \rangle + \langle \mu, \varphi \rangle + \langle \nu, \psi \rangle
\end{aligned}$$

Sinkhorn Algorithm (Primal)

(P) $\quad \mathrm{OT}_\varepsilon(\mu,\nu) = \varepsilon \min_{\alpha\in\Pi(\mu,\nu)} \mathrm{KL}(\alpha, K)$

Sinkhorn algorithm wrt α:

$$\alpha^{(0)} := K,$$

$$\text{for } r = 0, 1, \ldots$$

$$\alpha^{(r+\frac{1}{2})} := \mathrm{diag}\left(\frac{\mu}{\alpha^{(r)}\mathbf{1}}\right)\alpha^{(2r)}$$

$$\alpha^{(r+1)} := \alpha^{(r+\frac{1}{2})}\mathrm{diag}\left(\frac{\nu}{(\alpha^{(r+\frac{1}{2})})^\mathsf{T}\mathbf{1}}\right)$$

	ν_1	$\ldots$	ν_j	$\ldots$	ν_m
μ_1	$\alpha_{1,1}$	$\ldots$	$\alpha_{1,j}$	$\ldots$	$\alpha_{1,m}$
$\vdots$			$\vdots$		$\vdots$
μ_i	$\alpha_{i,1}$	$\ldots$	$\alpha_{i,j}$	$\ldots$	$\alpha_{i,m}$
$\vdots$			$\vdots$		$\vdots$
μ_n	$\alpha_{n,1}$	$\ldots$	$\alpha_{n,j}$	$\ldots$	$\alpha_{n,m}$

- See also block-iterative SMART, mirror descent algorithm

Sinkhorn Algorithm (Dual)

(D) $\max_{\varphi,\psi} F(\varphi,\psi) = \max_{\varphi,\psi} \langle K, \mathrm{e}^{-P_1^\mathsf{T}\varphi - P_2^\mathsf{T}\psi}\rangle + \langle \mu, \varphi\rangle + \langle \nu, \psi\rangle$

$$P_1 = \begin{pmatrix} I_m & \dots & I_m \end{pmatrix} \in \mathbb{R}^{m\times nm}$$

$$P_2 = \begin{pmatrix} \underbrace{1\ \cdots\ 1}_{m} & 0 & \cdots & 0 \\ 0 & \underbrace{1\ \cdots\ 1}_{m} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \underbrace{1\ \cdots\ 1}_{m} \end{pmatrix} \in \mathbb{R}^{n\times nm},$$

Idea:

$$\nabla_\varphi F = -\mathrm{diag}(\mathrm{e}^{-\varphi})\, K\, \mathrm{diag}(\mathrm{e}^{-\psi})\mathbf{1} + \mu = 0$$

Sinkhorn algorithm with $u := \mathrm{e}^{-\varphi}$, $v := \mathrm{e}^{-\psi}$

$$\nabla_\varphi F = 0 : \mathrm{diag}(\mathrm{e}^{-\varphi})\, K\, \mathrm{diag}(\mathrm{e}^{-\psi^{(r-1)}})\mathbf{1} = \mu \quad \to \quad u^{(r)} = \frac{\mu}{Kv^{(r-1)}}$$

$$\nabla_\psi F = 0 : \left(\mathrm{diag}(\mathrm{e}^{-\varphi^{(r)}})\, K\, \mathrm{diag}(\mathrm{e}^{-\psi})\right)^\mathsf{T}\mathbf{1} = \nu \quad \to \quad v^{(r)} = \frac{\nu}{K^\mathsf{T}u^{(r)}}$$

Relation to α: (set gradient of Lagrangian wrt α to 0)

$$0 = \log\alpha - \log K - P_1^\mathsf{T}\varphi - P_2^\mathsf{T}\psi \quad \Longleftrightarrow \quad \log\alpha = \log K + P_1^\mathsf{T}\varphi + P_2^\mathsf{T}\psi$$

$$\alpha = \mathrm{diag}(u)\, K\, \mathrm{diag}(v)$$

Convergence of Sinkhorn Algorithm

On $\mathbb{R}^d_{>0}$ consider the equivalence relation

$$x \sim x' \iff \exists t > 0 \ : \ x = t\, x'.$$

The quotient space, called positive projective space

$$\mathbb{P}^d_{>0} := \mathbb{R}^d_{>0} / \sim$$

can be endowed with a vector space structure using the following operations:

- Addition (internal operation): $x + x := x \odot y$,
- Scalar multiplication (external operation): $\gamma \cdot x := x^\gamma$,

with component-wise product $\odot$. Using the Hilbert projective metric:

$$d_H(x, x) := \log\left(\frac{\max_i \ x_i/y_i}{\min_i \ x_i/y_i}\right), \qquad x, x \in \mathbb{P}^d_{>0}.$$

the space $\mathbb{P}^d_{>0}$ becomes a complete metric space.

Linear convergence of Sinkhorn algorithm:

$$d_H(u^{(r+1)}, u^*) \leq \lambda(K)^2 d_H(u^{(r)}, u^*)$$

with

$$\lambda(K) := \sup\left\{\frac{d_H(Kx, Ky)}{d_H(x, y)} : x \not\sim y\right\}$$

Further Topics

- Unbalanced optimal transport
- Sliced Wasserstein distances
- Gromov-Wasserstein distances
- ...

Unbalanced Optimal Transport

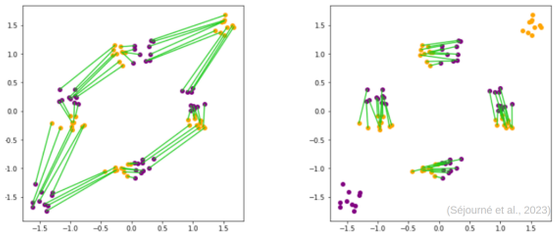
Given:

- positive measures $\mu, \nu \in \mathcal{M}_+(\mathbb{R}^d)$.
- cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$.
- divergence measure $D_f(\cdot|\cdot)$ (e.g., KL-divergence, Total Variation).
- regularization parameters $\tau_1, \tau_2 > 0$.

Find optimal coupling α:

$$\mathrm{UOT}(\mu,\nu) = \min_{\alpha \in \mathcal{M}_+(\mathbb{R}^d \times \mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\, d\alpha(x,y) \right.$$
$$\left. + \tau_1 D_f((P_1)_\# \alpha | \mu) + \tau_2 D_f((P_2)_\# \alpha | \nu) \quad tr\right.$$

- Compares measures with different mass.
- Robust to outliers.



(a) OT matching (b) Unbalanced OT matching

Sliced Wasserstein

Given: $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$

Define 1d Projection: $P_\theta(x) = \theta^\top x$ for $\theta \in \mathbb{S}^{d-1}$ (unit sphere)
$\Rightarrow P_{\theta,\#}\mu, P_{\theta,\#}\nu \in \mathcal{P}_p(\mathbb{R})$

Introduce:

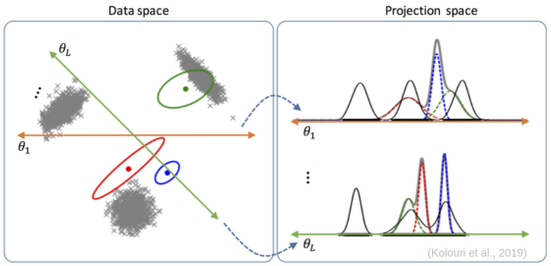
$$\mathrm{SW}_p^p(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(P_{\theta,\#}\mu, P_{\theta,\#}\nu)\, d\xi(\theta),$$

where $\xi \sim \mathcal{U}(\mathbb{S}^{d-1})$ is the uniform measure on the unit sphere $\mathbb{S}^{d-1}$.

Calculate Monte Carlo Estimate:

$$\mathrm{SW}_p^p(\mu,\nu) \approx \sum_{l=1}^{L} \mathrm{W}_p^p(P_{\theta_l,\#}\mu, P_{\theta_l,\#}\nu), \quad \theta_l \sim \mathcal{U}(\mathbb{S}^{d-1}).$$

- Fast 1d Wasserstein Computation via Sorting ($\mathcal{O}(n \log n)$)
- Fast SW approximation ($\mathcal{O}(L\, n \log n)$)



Gromov-Wasserstein Distance

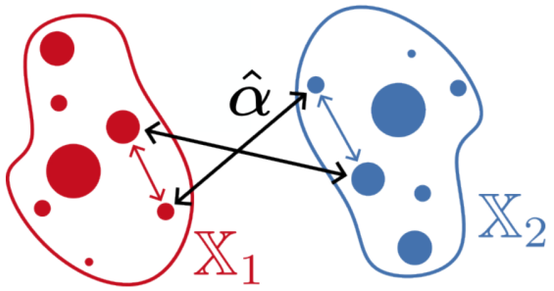
Given metric measure spaces:

$$\mathbb{X} = (X, d_X, \mu), \qquad (X, d_X) \text{ compact metric space, } \mu \in \mathcal{P}(X),$$
$$\mathbb{Y} = (Y, d_Y, \nu), \qquad (Y, d_Y) \text{ compact metric space, } \nu \in \mathcal{P}(Y).$$

Find optimal GW transport plan $\hat{\alpha} \in \mathcal{P}(X \times Y)$:

$$\hat{\alpha} \in \operatorname{argmin}_{\alpha \in \Pi(\mu,\nu)} \iint_{(X\times Y)^2} \left| d_X(x, x') - d_Y(y, y') \right|^2 d\alpha(x, y)\, d\alpha(x', y').$$

$$\mathrm{GW}^2(\mathbb{X}, \mathbb{Y}) = \iint_{(X\times Y)^2} \left| d_X(x, x') - d_Y(y, y') \right|^2 d\hat{\alpha}(x, y)\, d\hat{\alpha}(x', y').$$



→ GW vanishes only for isomorphic spaces and allows for interpolating geometries.

Berlin Mathematics Research Center

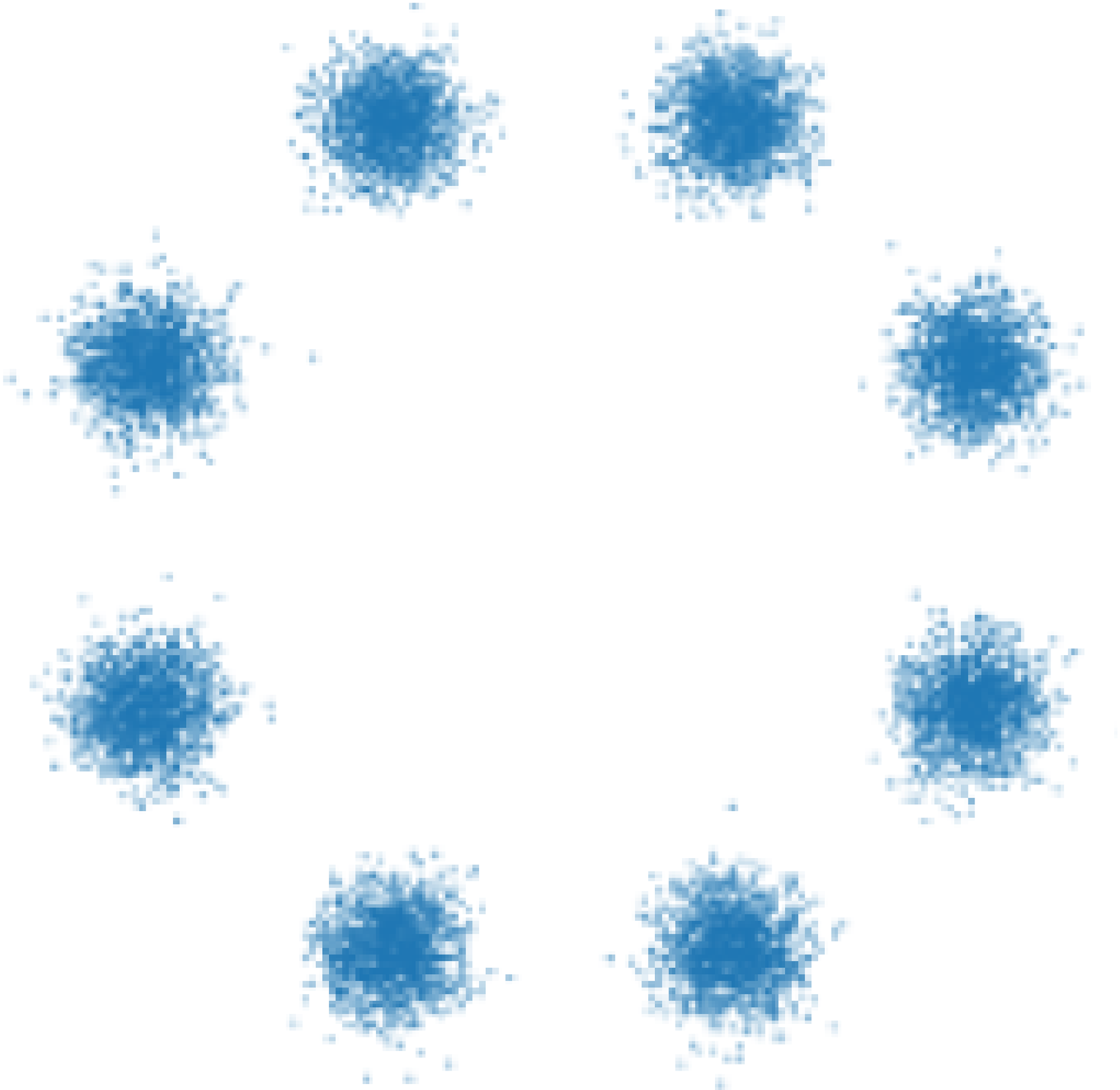
Funded under Germany's Excellence Strategy by
DFG Deutsche Forschungsgemeinschaft

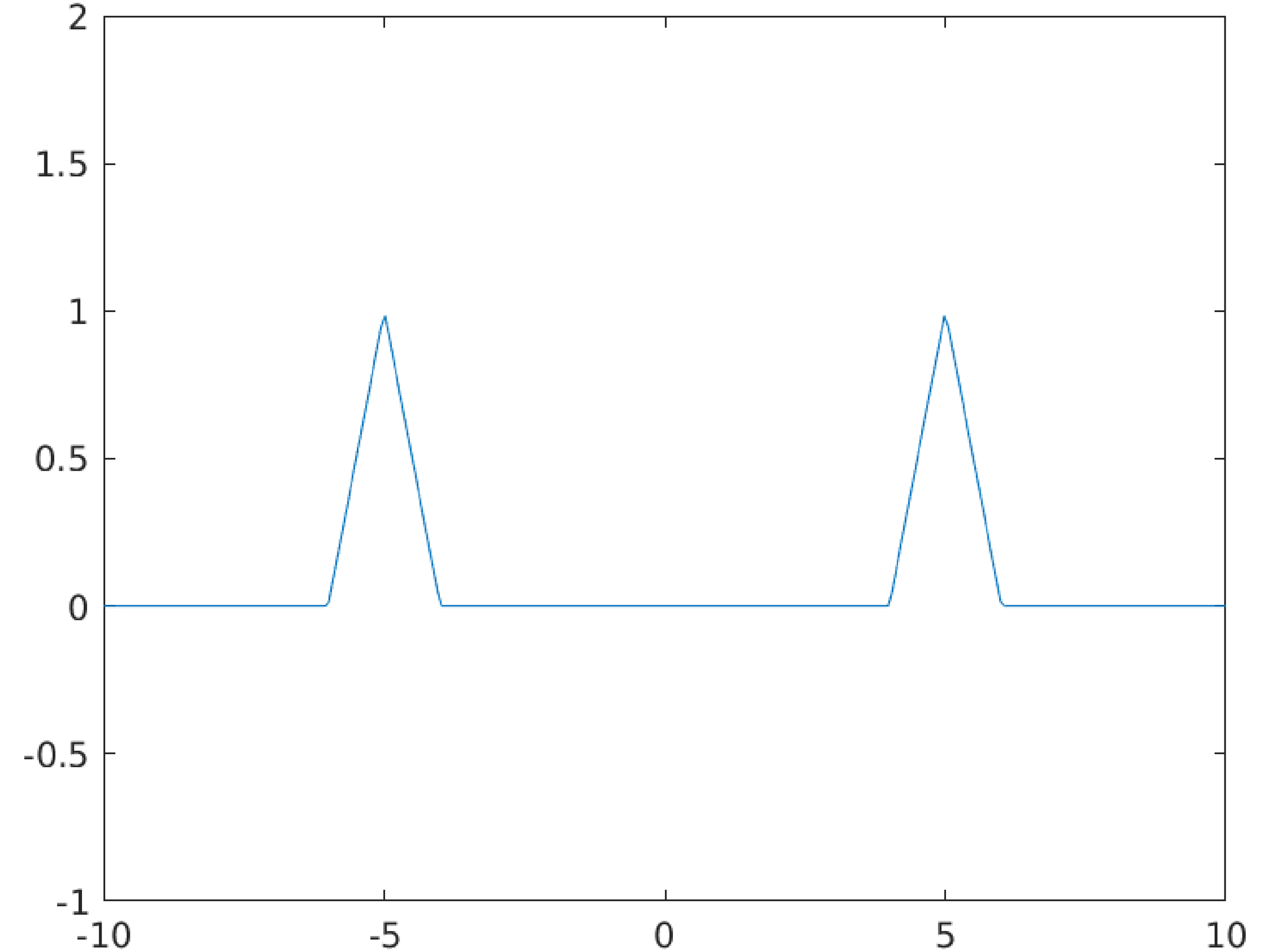
$$\sum_{j=1}^{7} \mu_j \delta_{x_j} = \mu$$

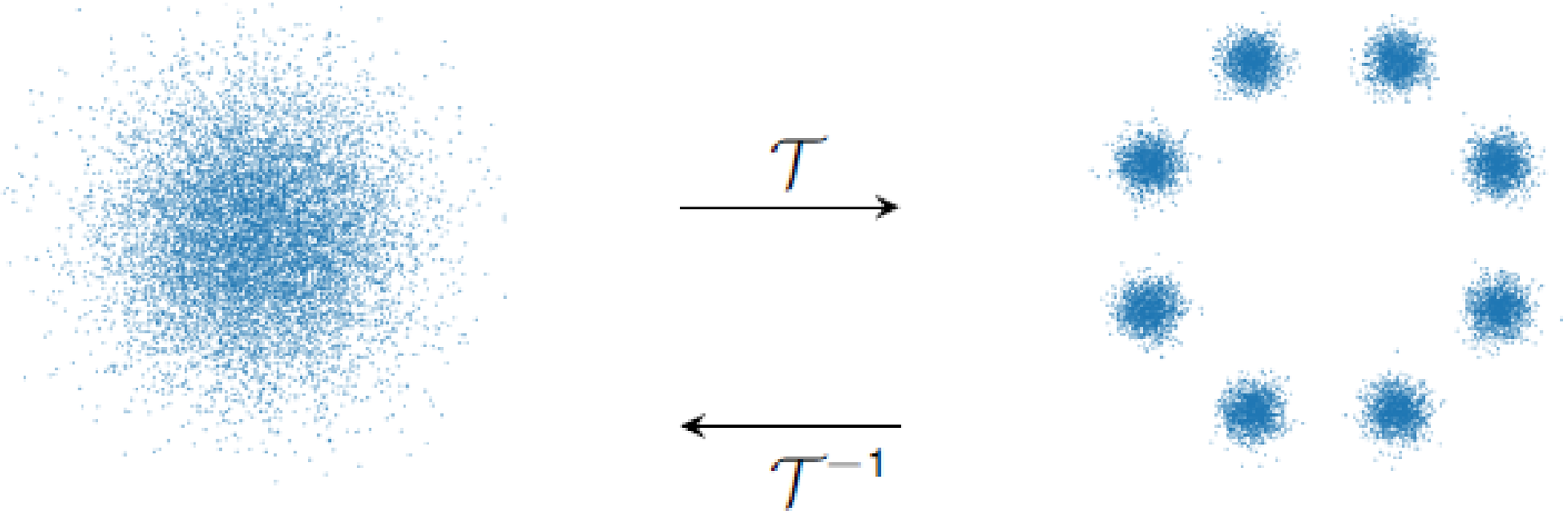
$$\sum_{k=1}^{8} \upsilon_k \delta_{y_k} = \upsilon$$

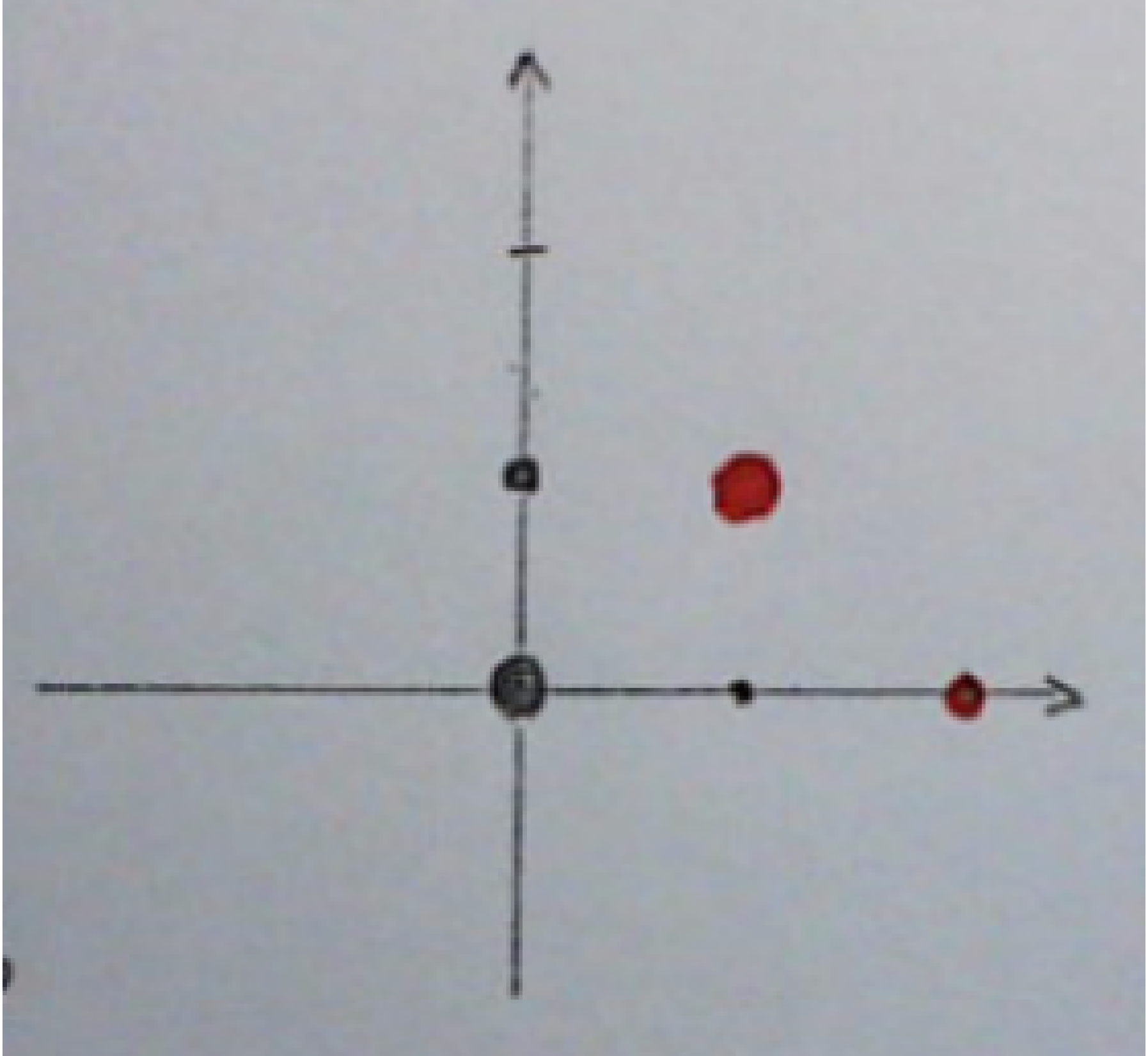
$$\sum_{j=1}^{7} \mu_j \delta_{x_j} = \mu$$

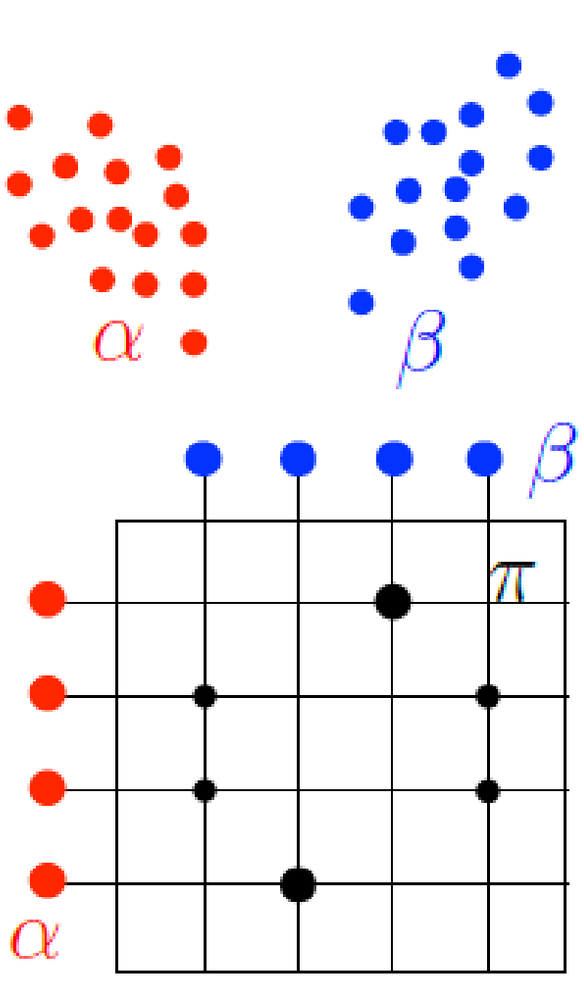
$$\sum_{k=1}^{5} \xi_k \delta_{z_k} = \xi$$



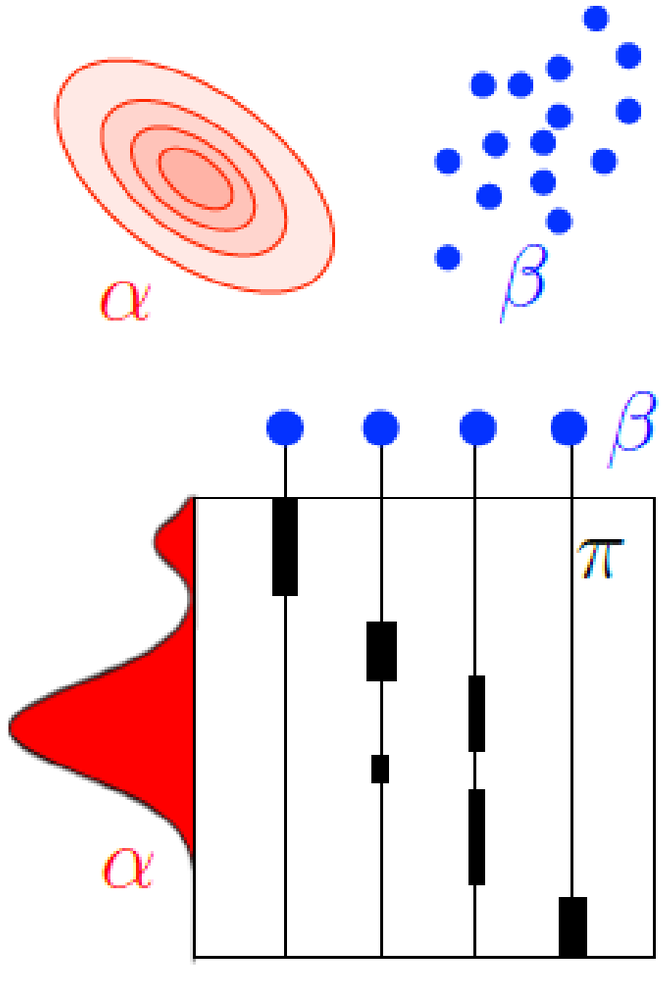

2
1.5
1
0.5
0
-0.5
-1
-10
-5
0
5
10


T
T^{-1}
Latent distribution P_Z
Data distribution P_X

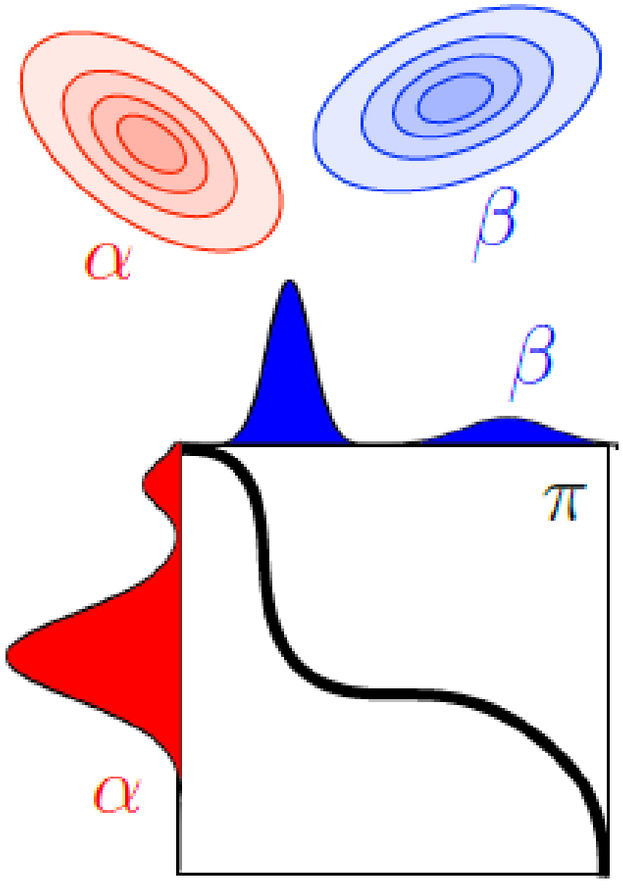




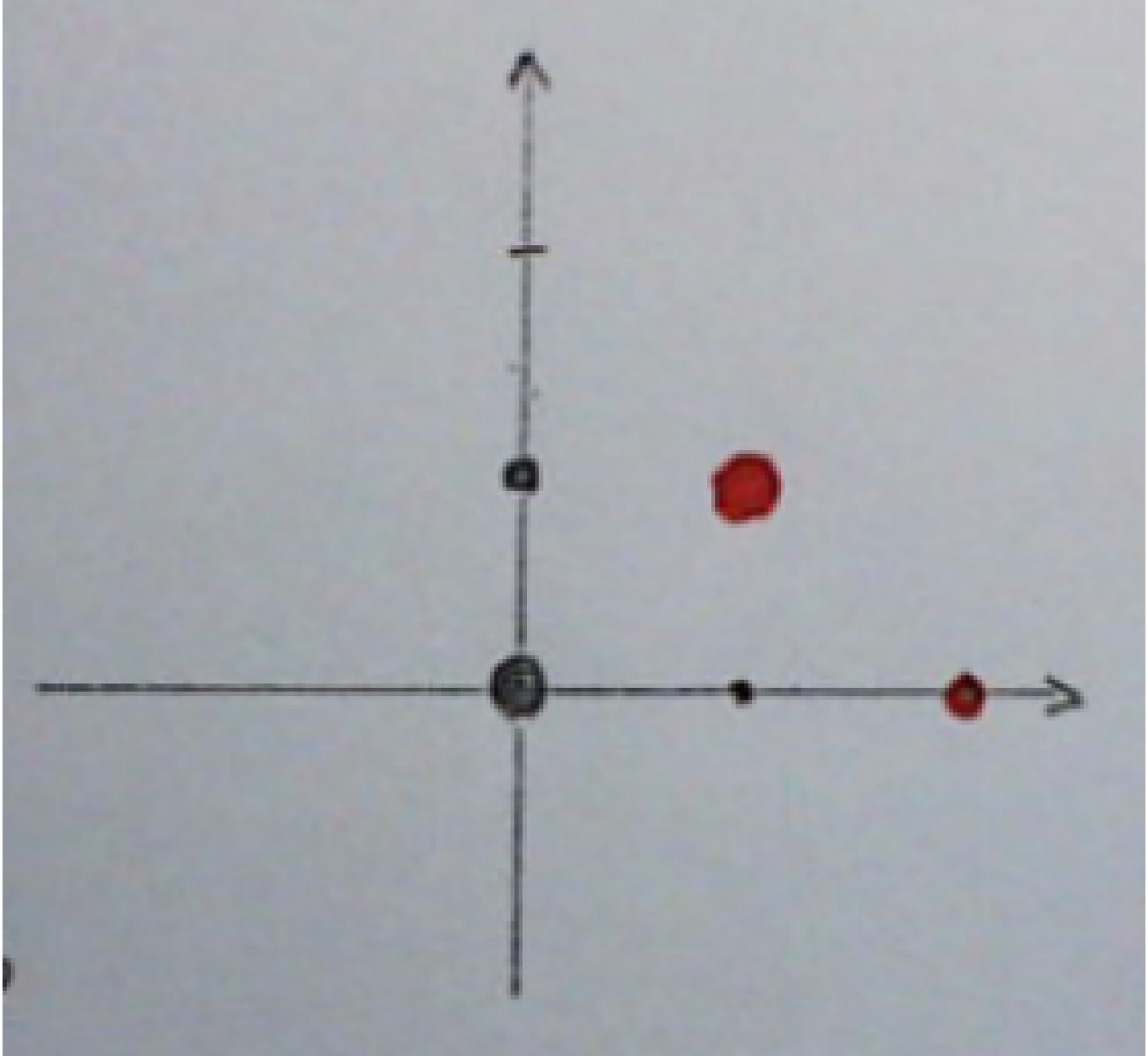
Discrete

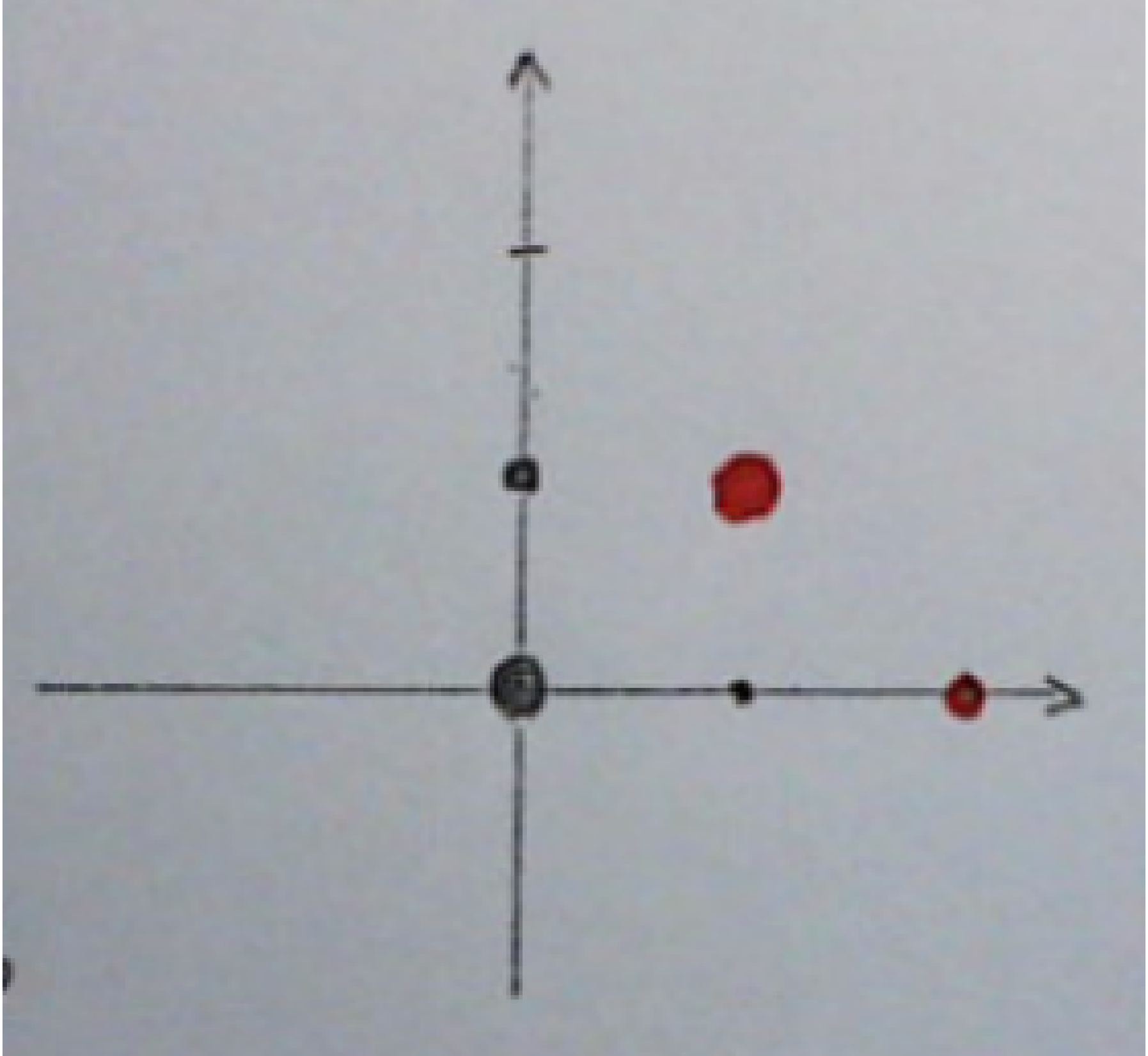


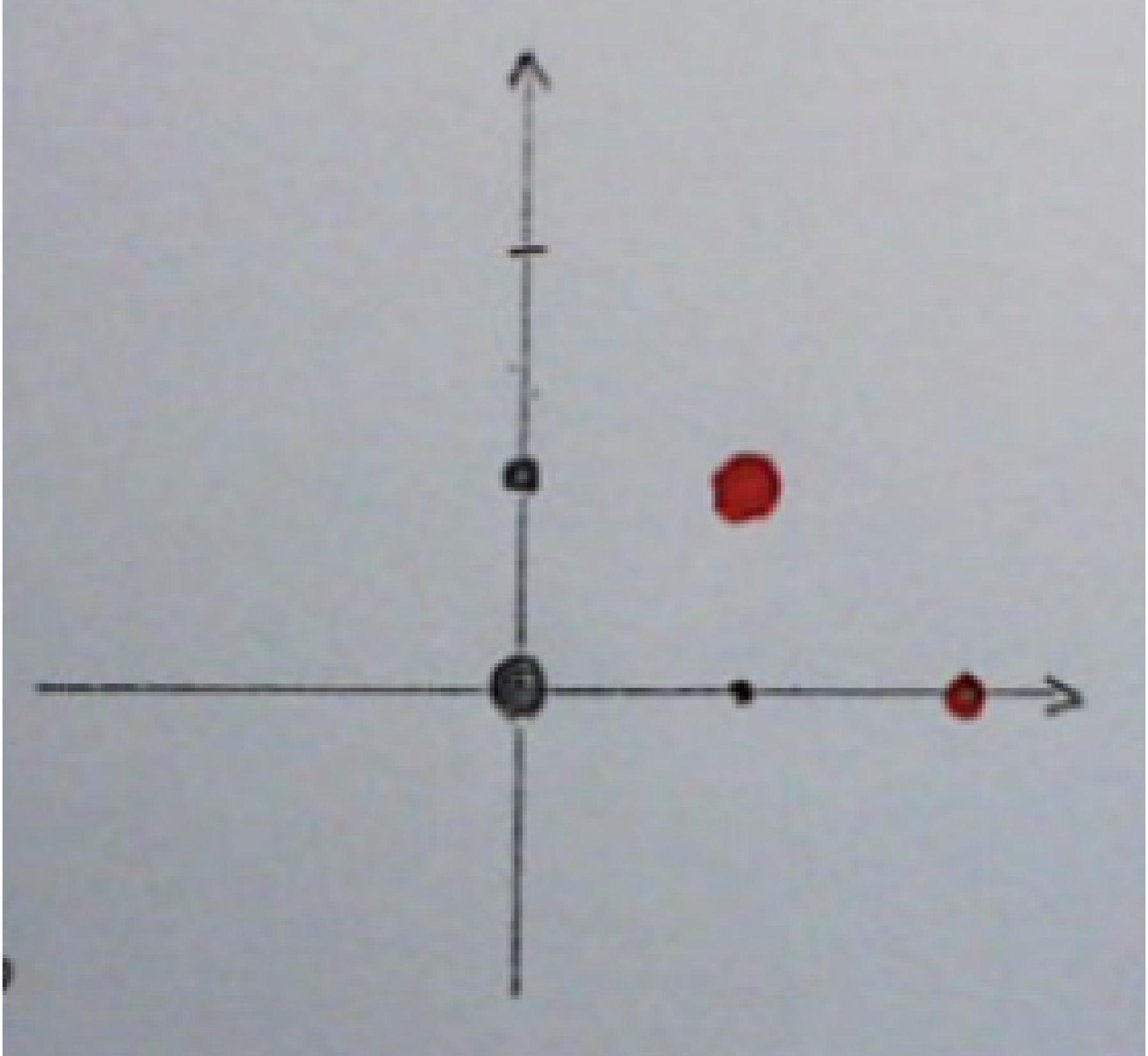
Semidiscrete

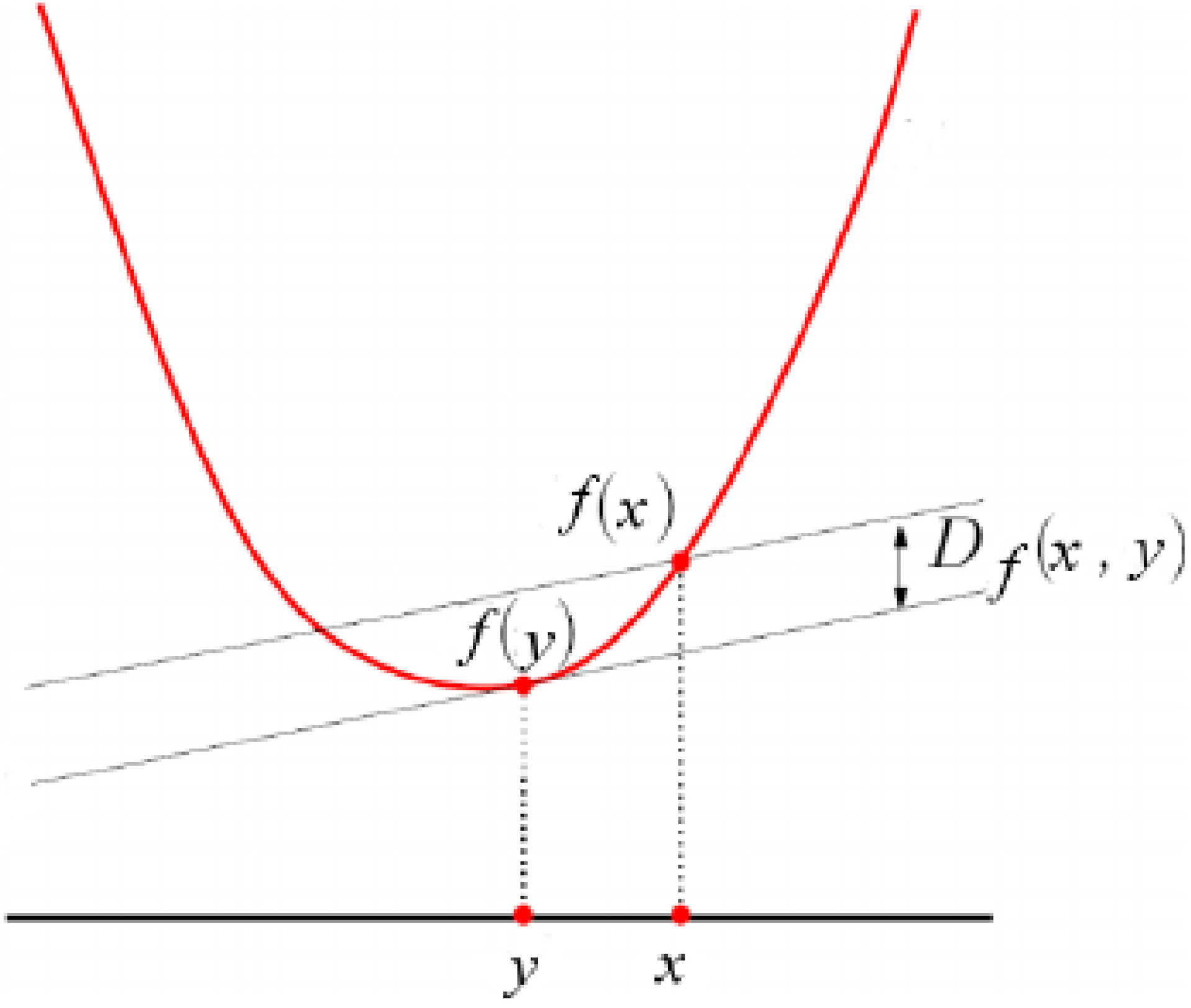


Continuous

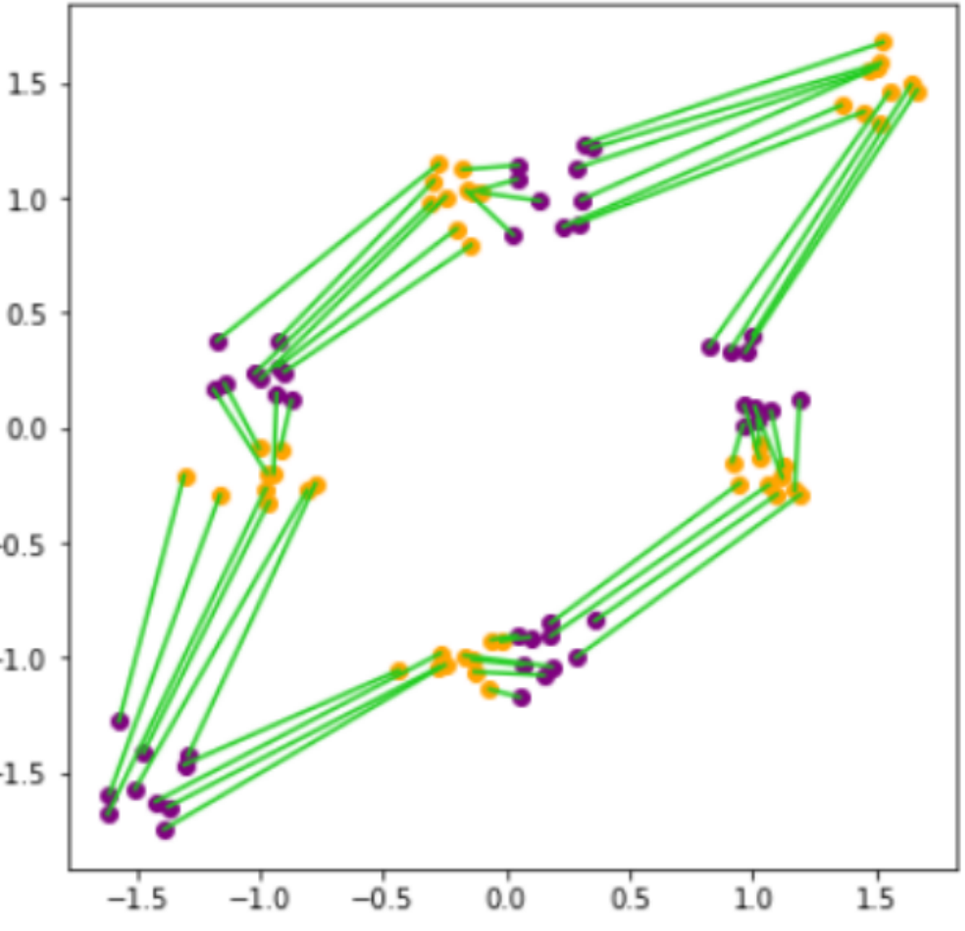




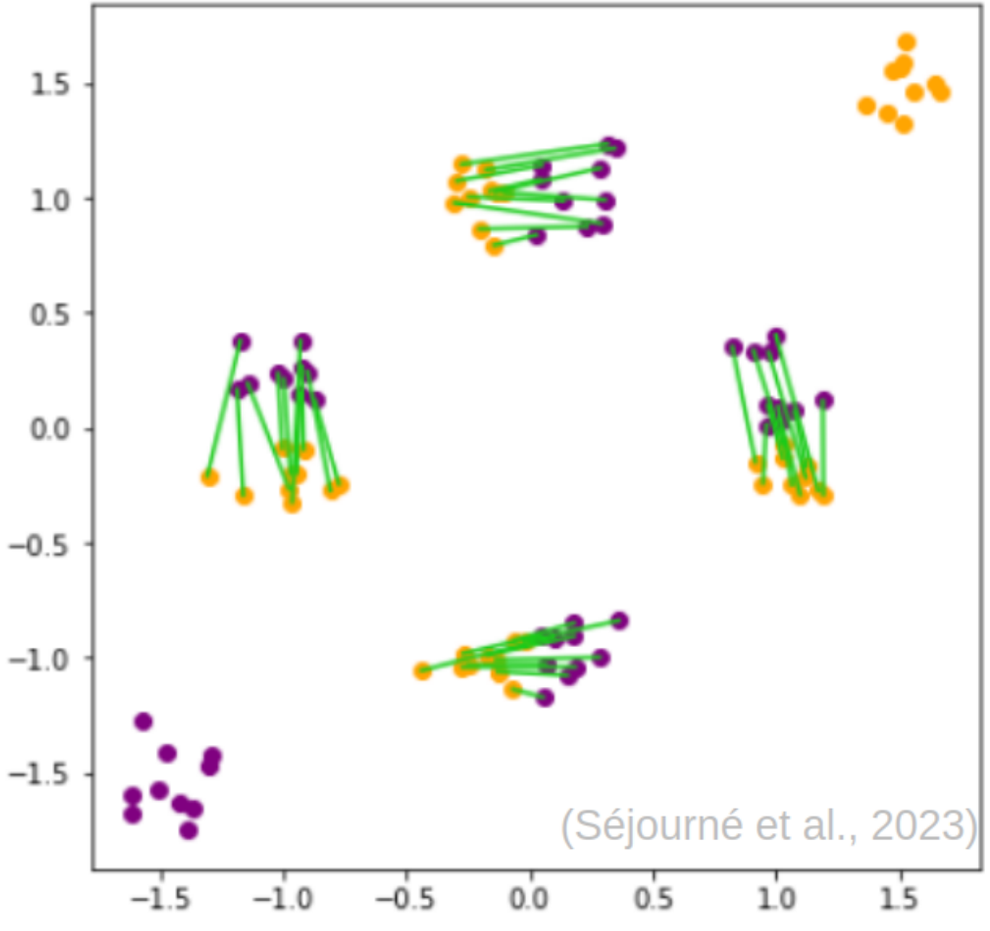



f(x)
D
f(x , y)
f(y)
y
x

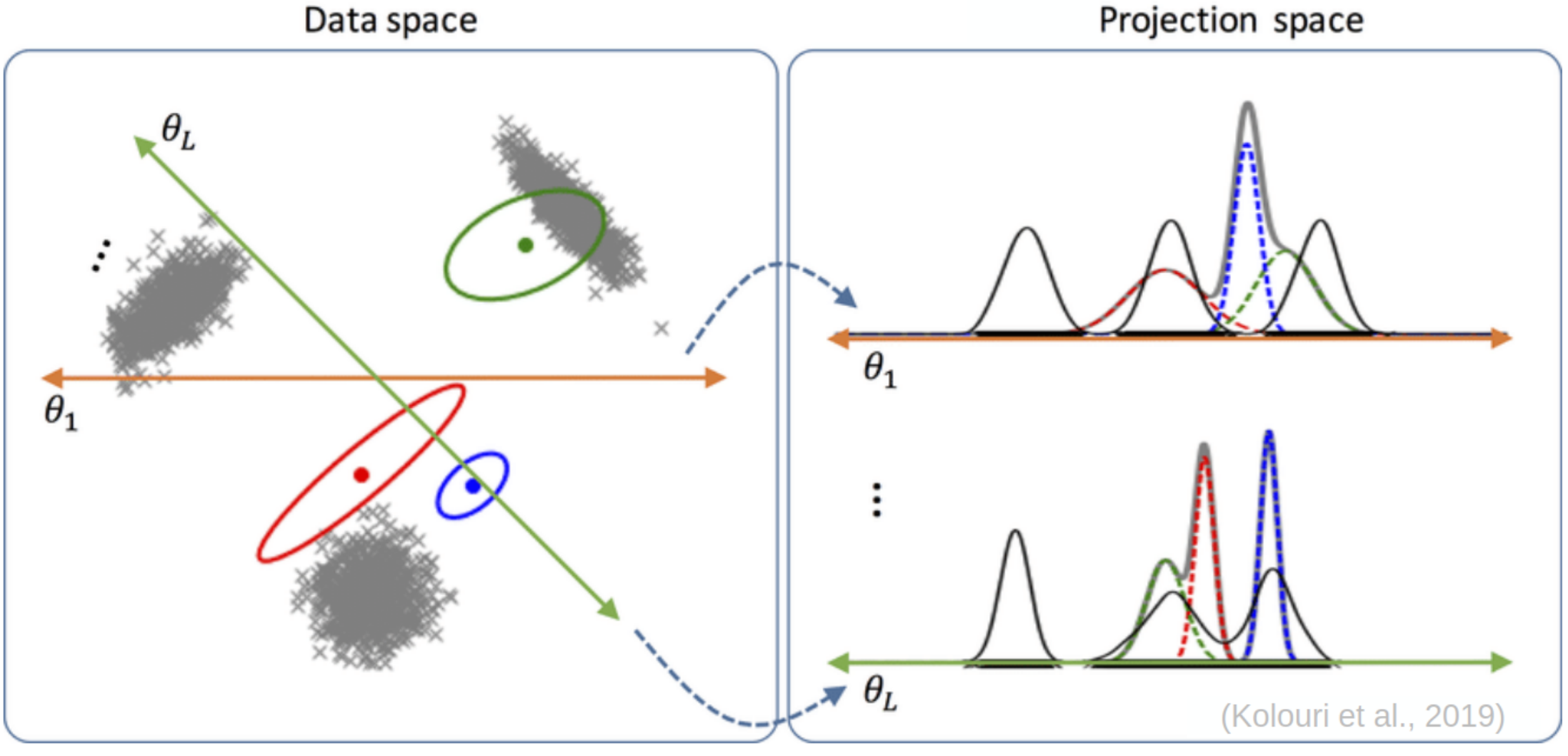
Function Name	$f(x)$	$\operatorname{dom} f$	$D_f(x, y)$
Squared Norm	$\frac{1}{2}x^2$	$(-\infty, \infty)$	$\frac{1}{2}(x-y)^2$
Shannon Entropy	$x \log x$	$[0, \infty)$	$x \log \frac{x}{y} - x + y$
Bit Entropy	$x \log x + (1-x) \log(1-x)$	$[0, 1]$	$x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}$
Burg Entropy	$-\log x$	$(0, \infty)$	$\frac{x}{y} - \log \frac{x}{y} - 1$

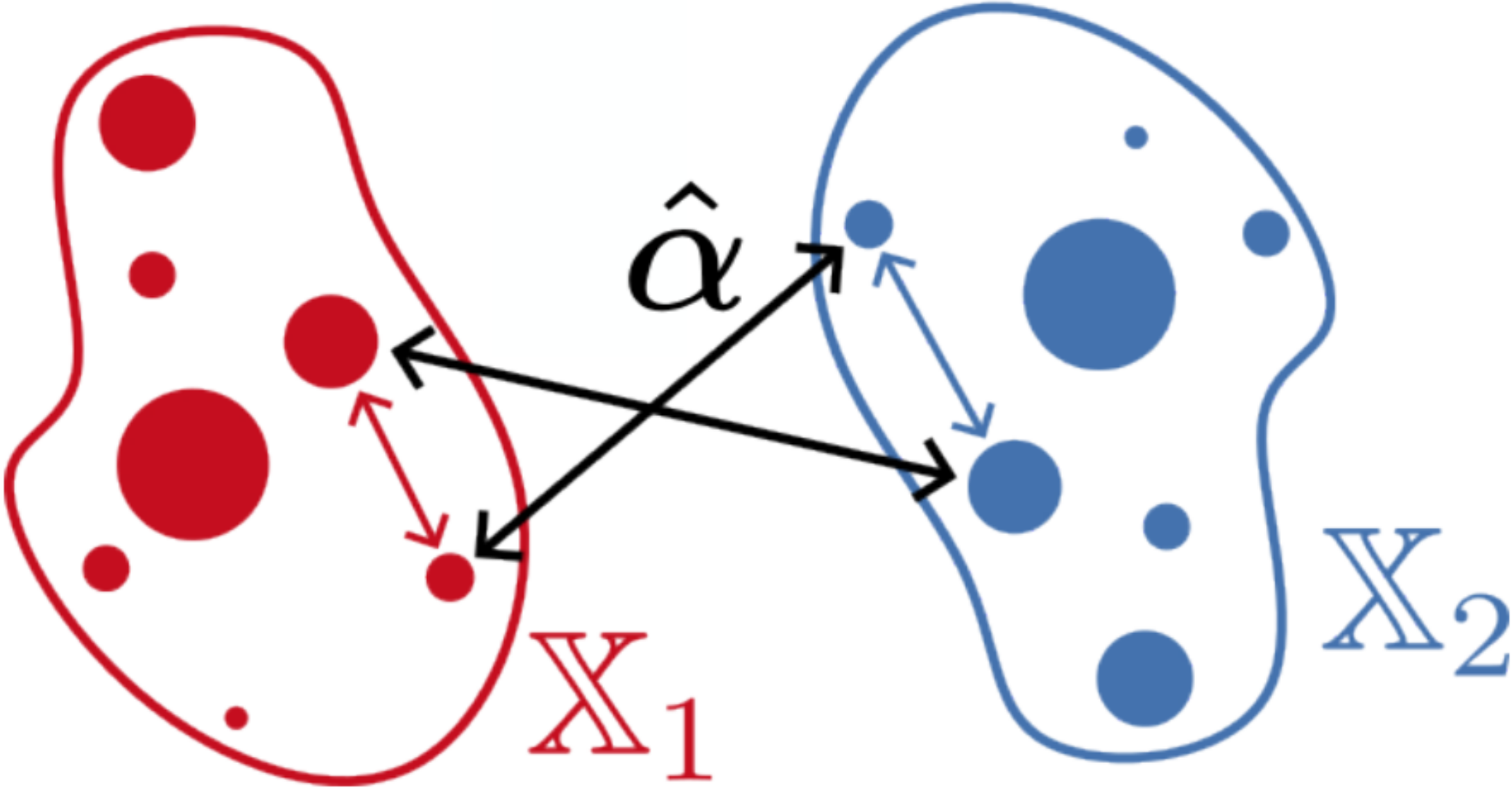


(a) OT matching



(b) Unbalanced OT matching


Data space
Projection space
θ_L
θ_1
θ_1
θ_L
(Kolouri et al., 2019)


$\hat{\alpha}$
$\mathbb{X}_1$
$\mathbb{X}_2$

(Beier & Beinert, 2025)