

Duality from Distributionally Robust Learning to Gradient Flow Force-Balance

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics, Berlin

jj-zhu.github.io

Based on previous and on-going work with
students and collaborators

DP4ML Workshop, ICML 2023
Honolulu Hawaii, USA, July, 2023,



Weierstraß-Institut für
Angewandte Analysis und Stochastik



Duality of Distributionally Robust Learning

Distributional robustness, but what kind?

Distributional robustness, but what kind?

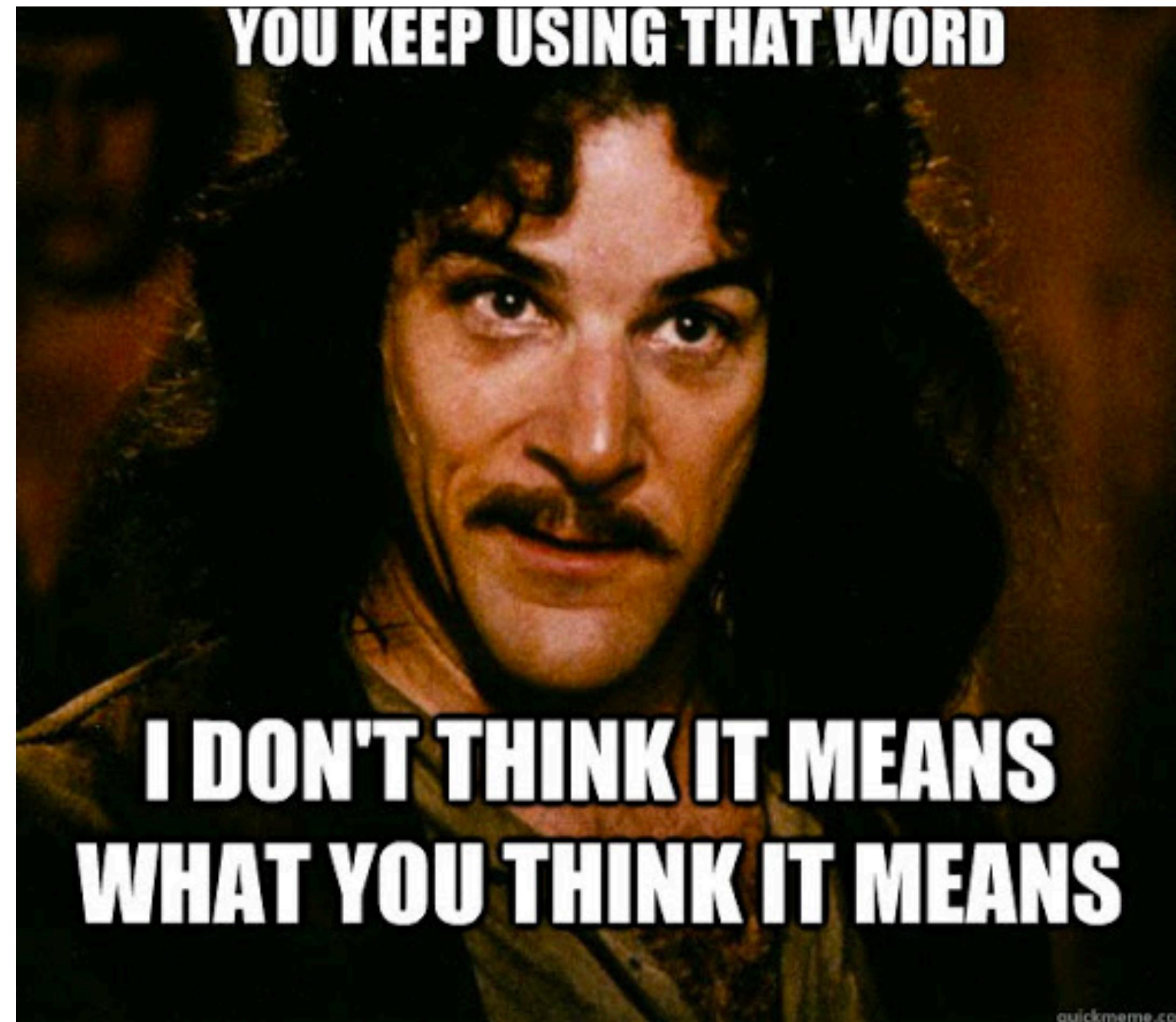


Figure credit: The Princess Bride,
a bedside story by your grandpa

From Statistical Learning to Distributionally Robust Learning

From Statistical Learning to Distributionally Robust Learning

Empirical Risk Minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

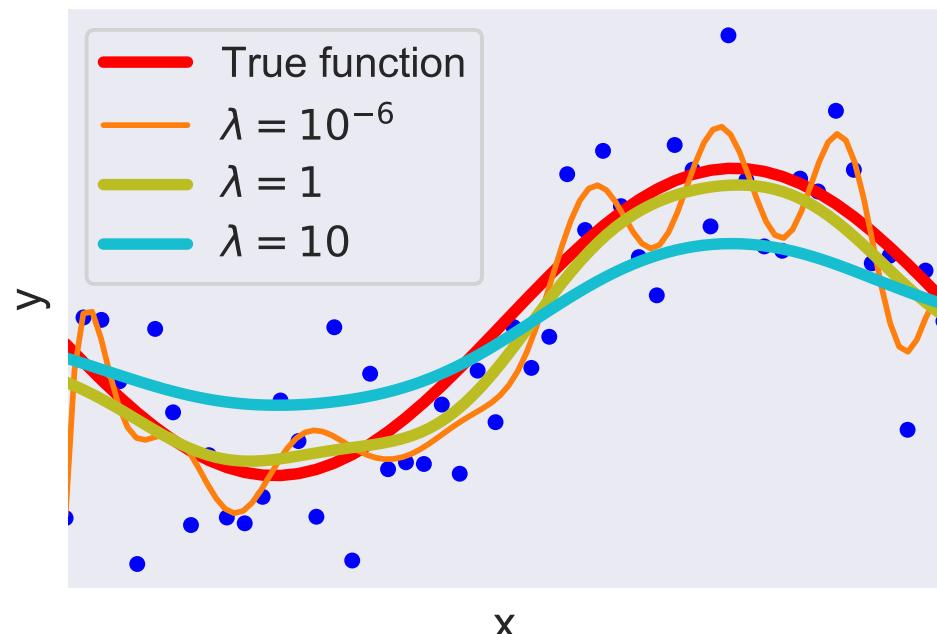
From Statistical Learning to Distributionally Robust Learning

Empirical Risk Minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$



From Statistical Learning to Distributionally Robust Learning

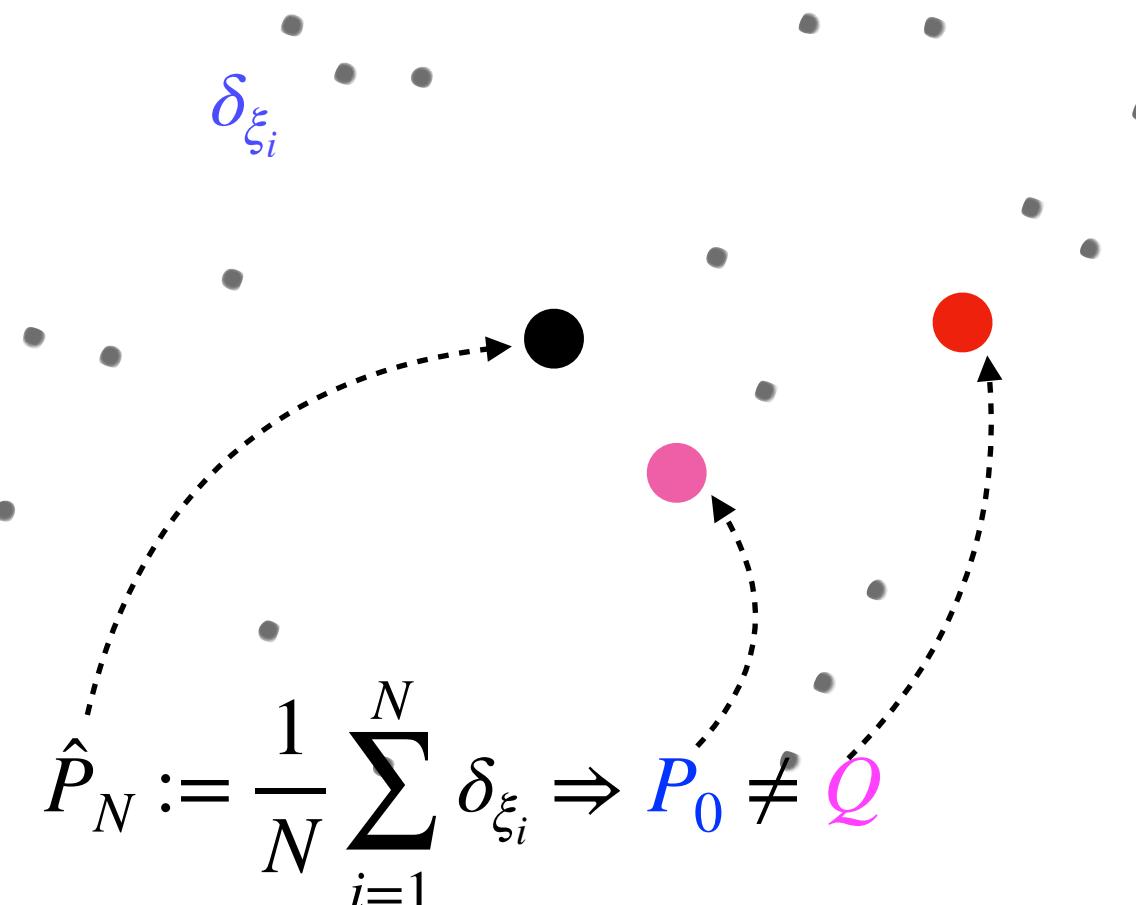
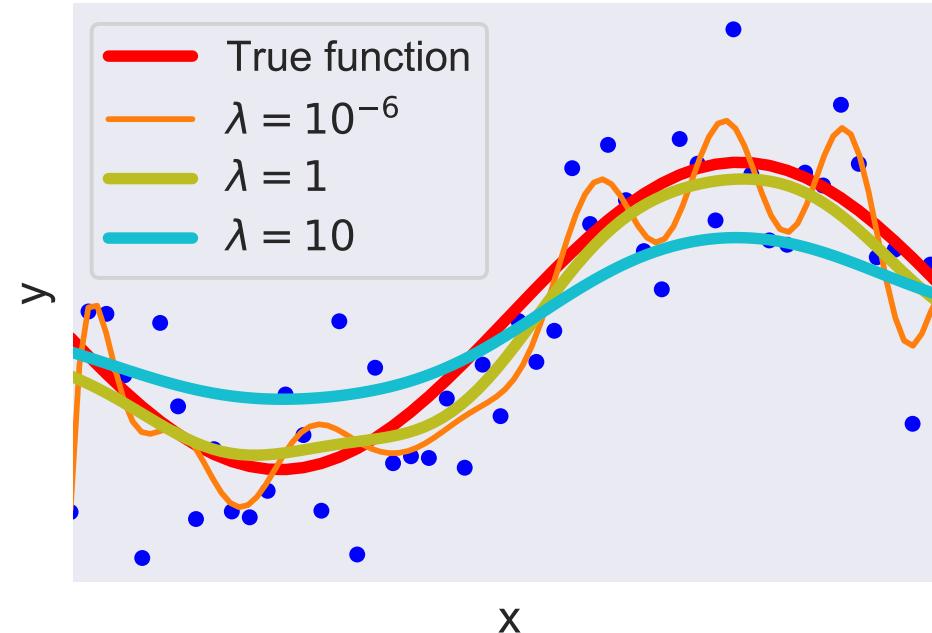
Empirical Risk Minimization

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts,
when Q ($\neq P_0$)



From Statistical Learning to Distributionally Robust Learning

Empirical Risk Minimization Distributionally Robust Optimization (DRO)

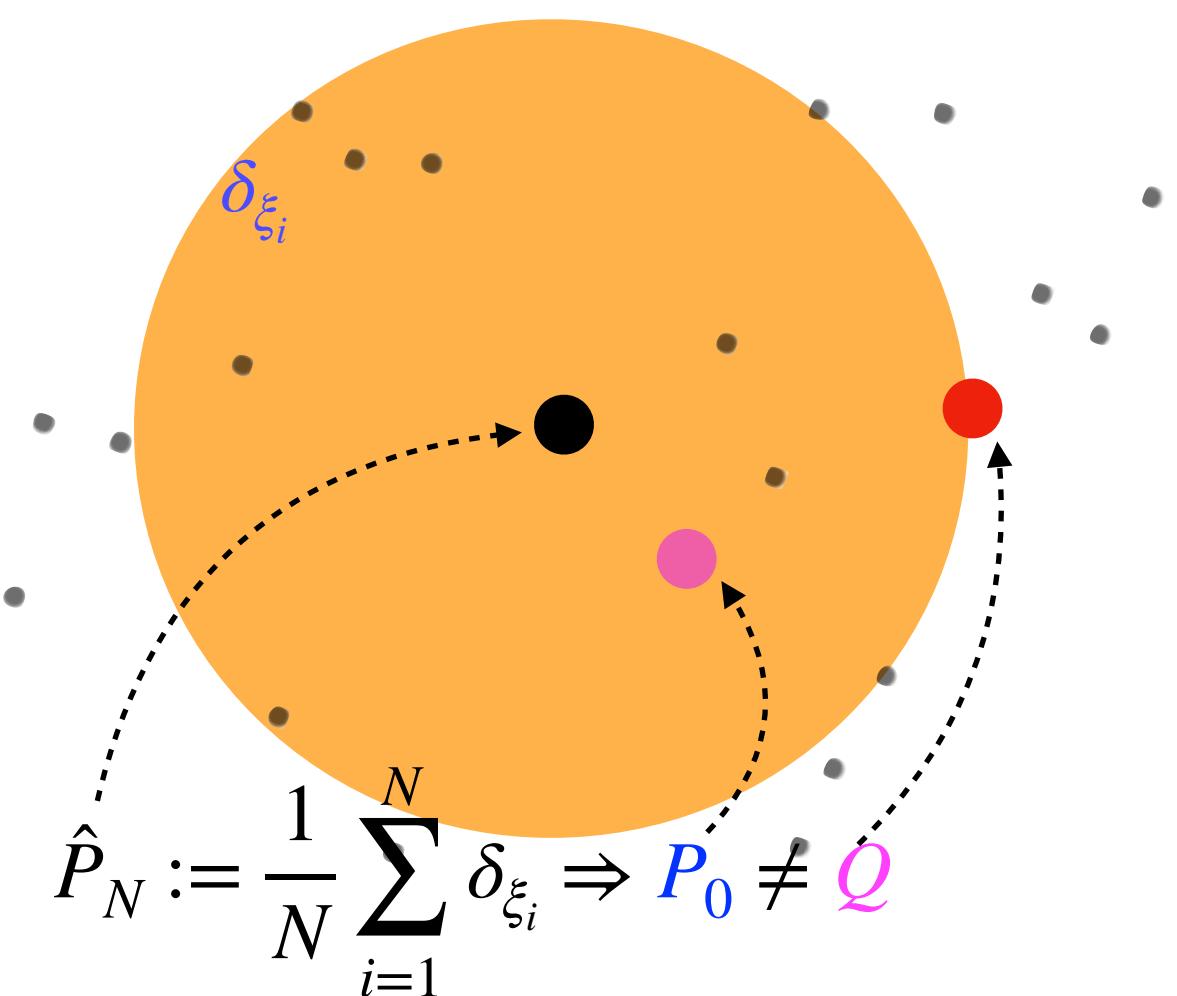
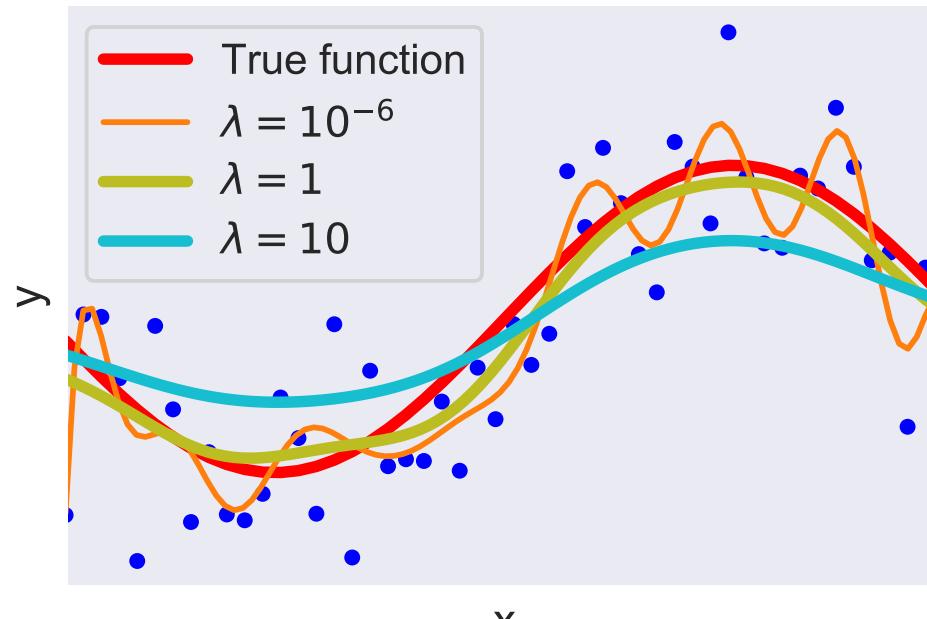
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

- “Robust” under statistical fluctuation

$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts,
when Q ($\neq P_0$)



From Statistical Learning to Distributionally Robust Learning

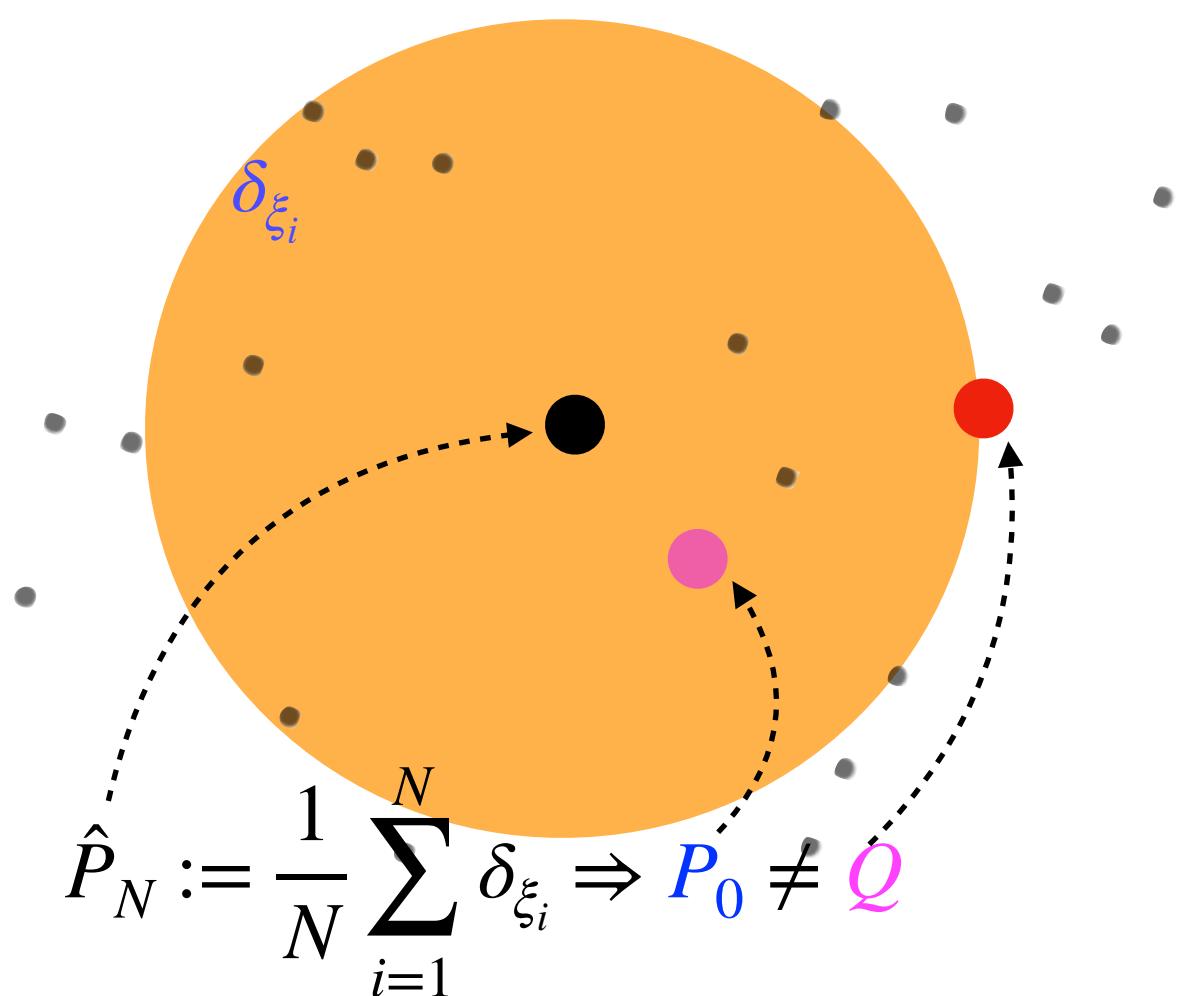
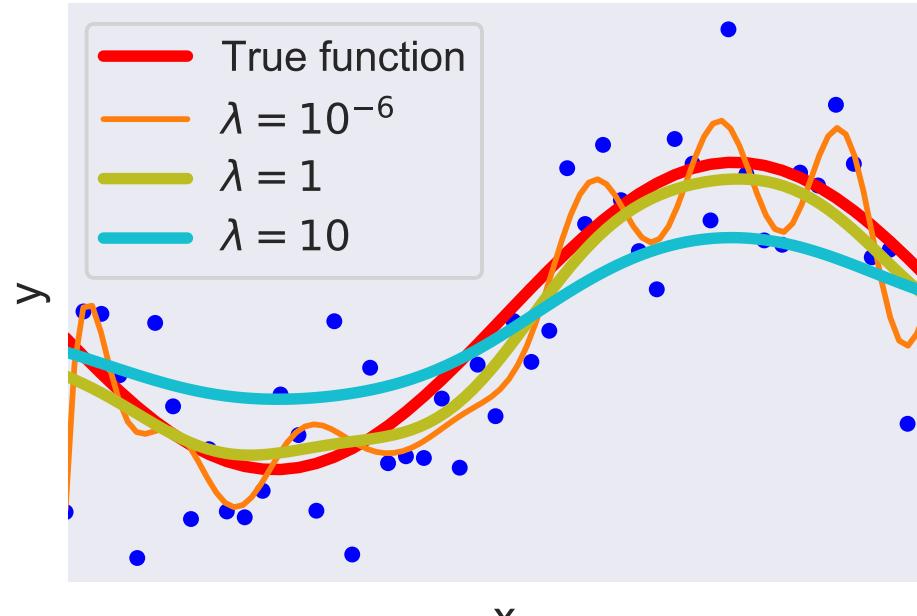
Empirical Risk Minimization Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts,
when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M}
 [Delage & Ye 2010] in certain geometry.

From Statistical Learning to Distributionally Robust Learning

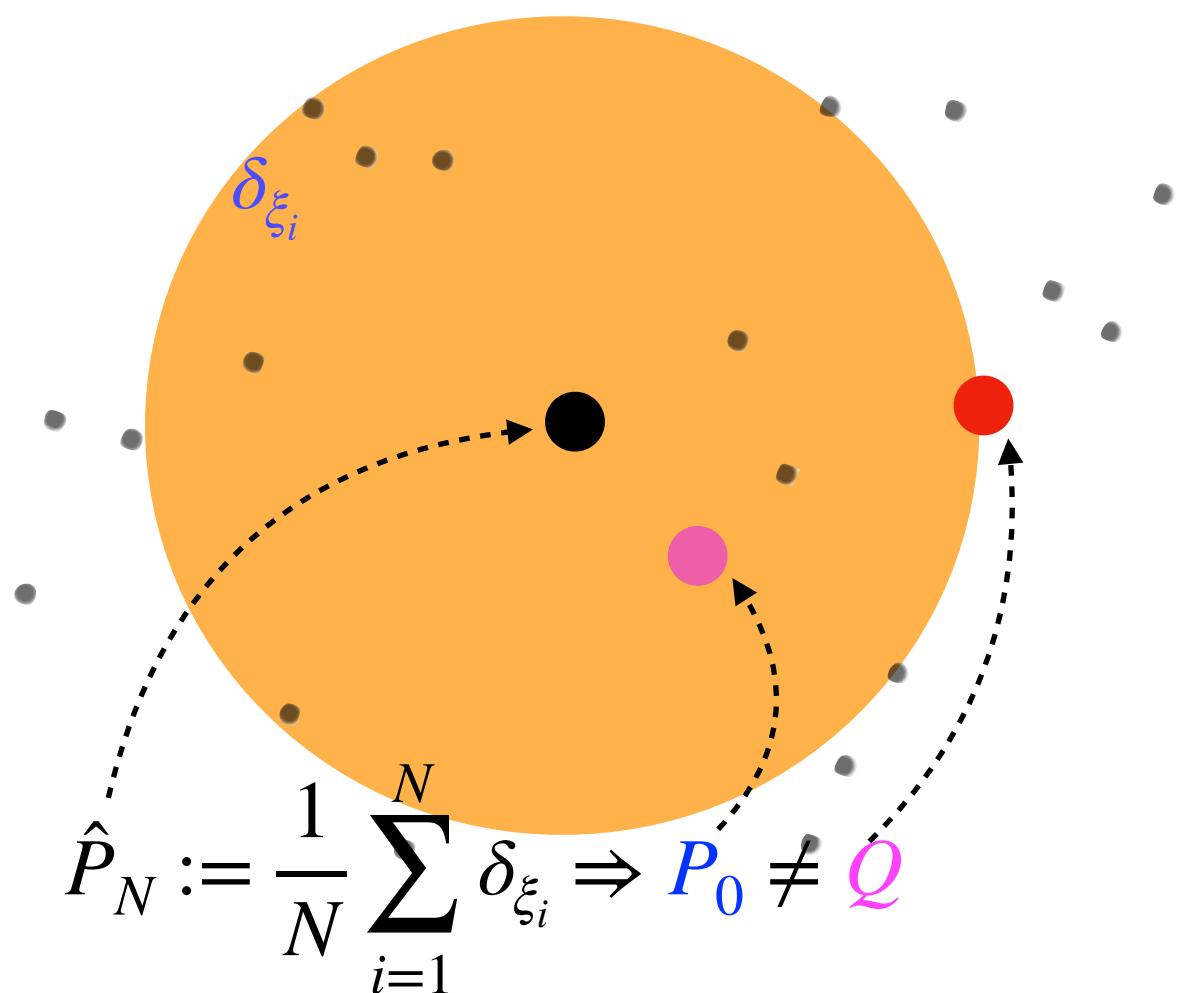
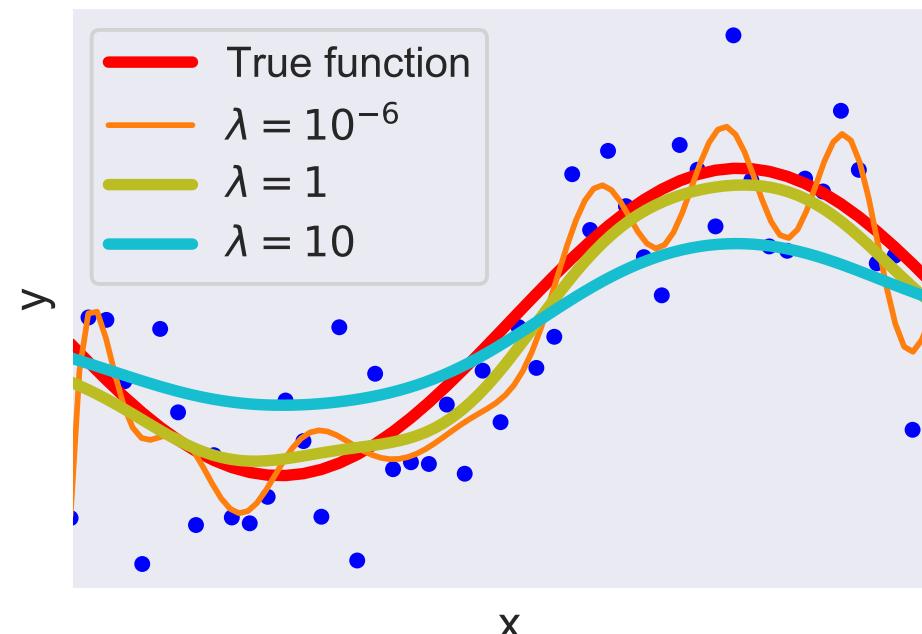
Empirical Risk Minimization Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

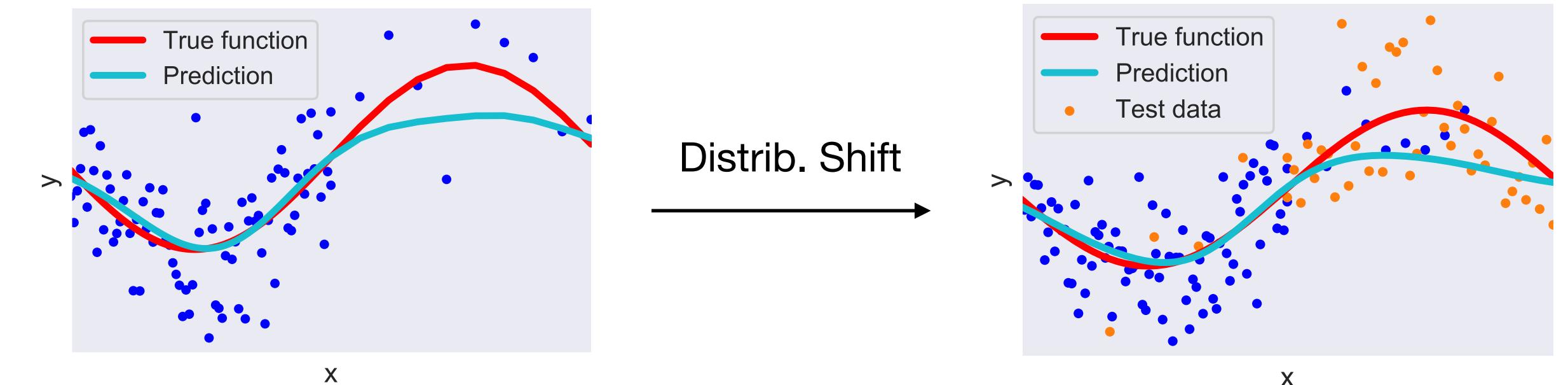
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts, when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M}
[Delage & Ye 2010] in certain geometry.



From Statistical Learning to Distributionally Robust Learning

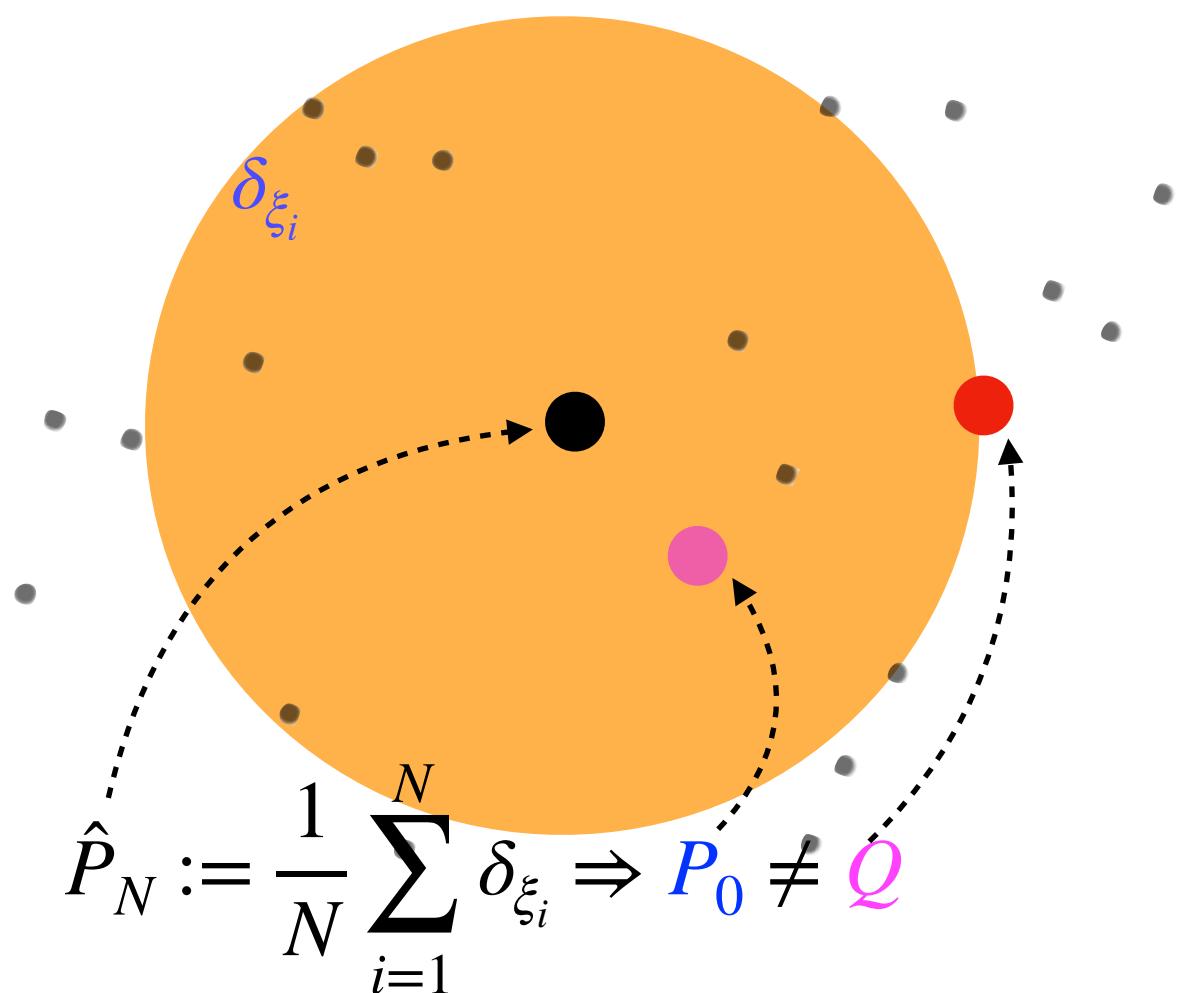
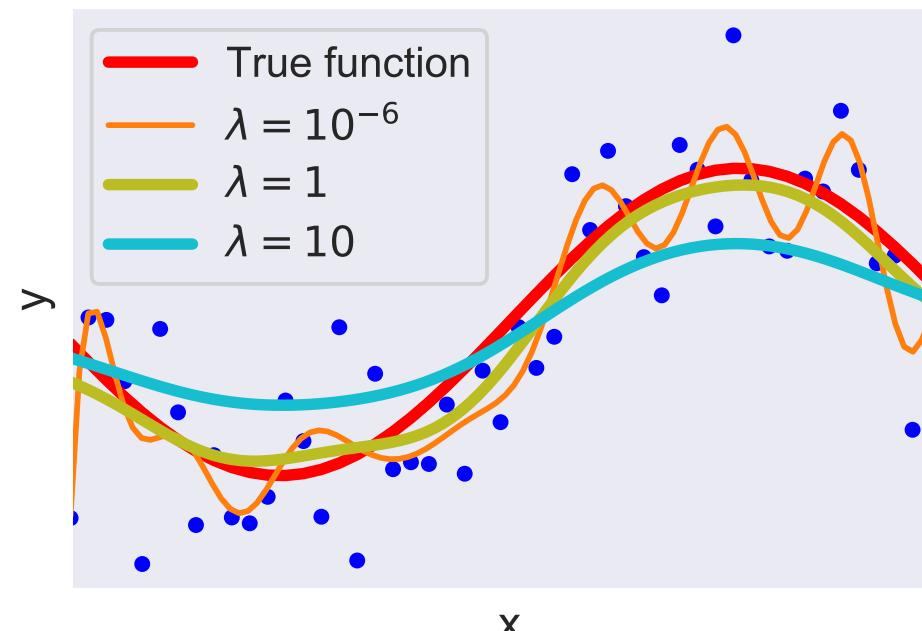
Empirical Risk Minimization Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

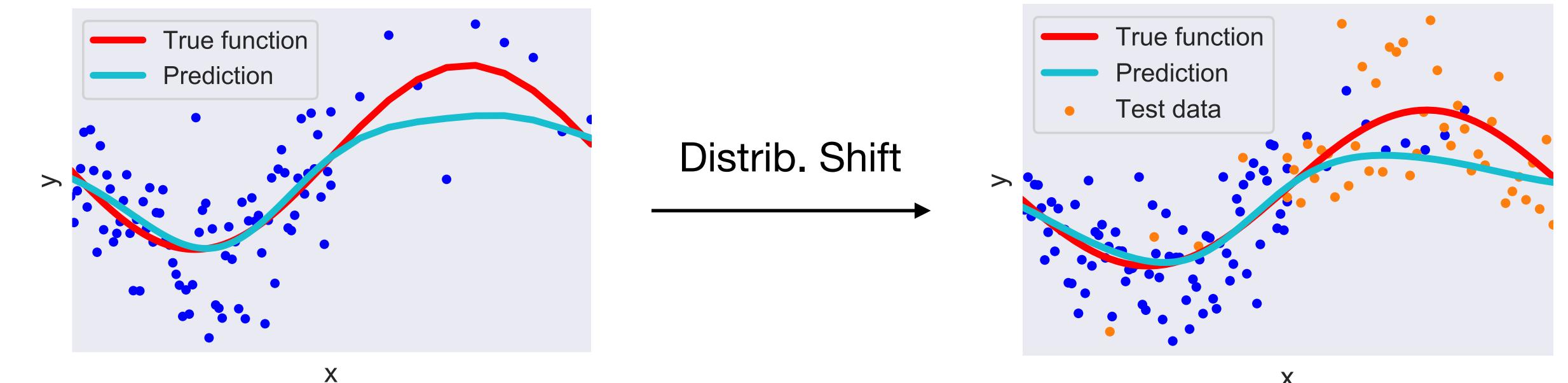
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts, when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M}
 [Delage & Ye 2010] in certain geometry.



Why study new geometry?

From Statistical Learning to Distributionally Robust Learning

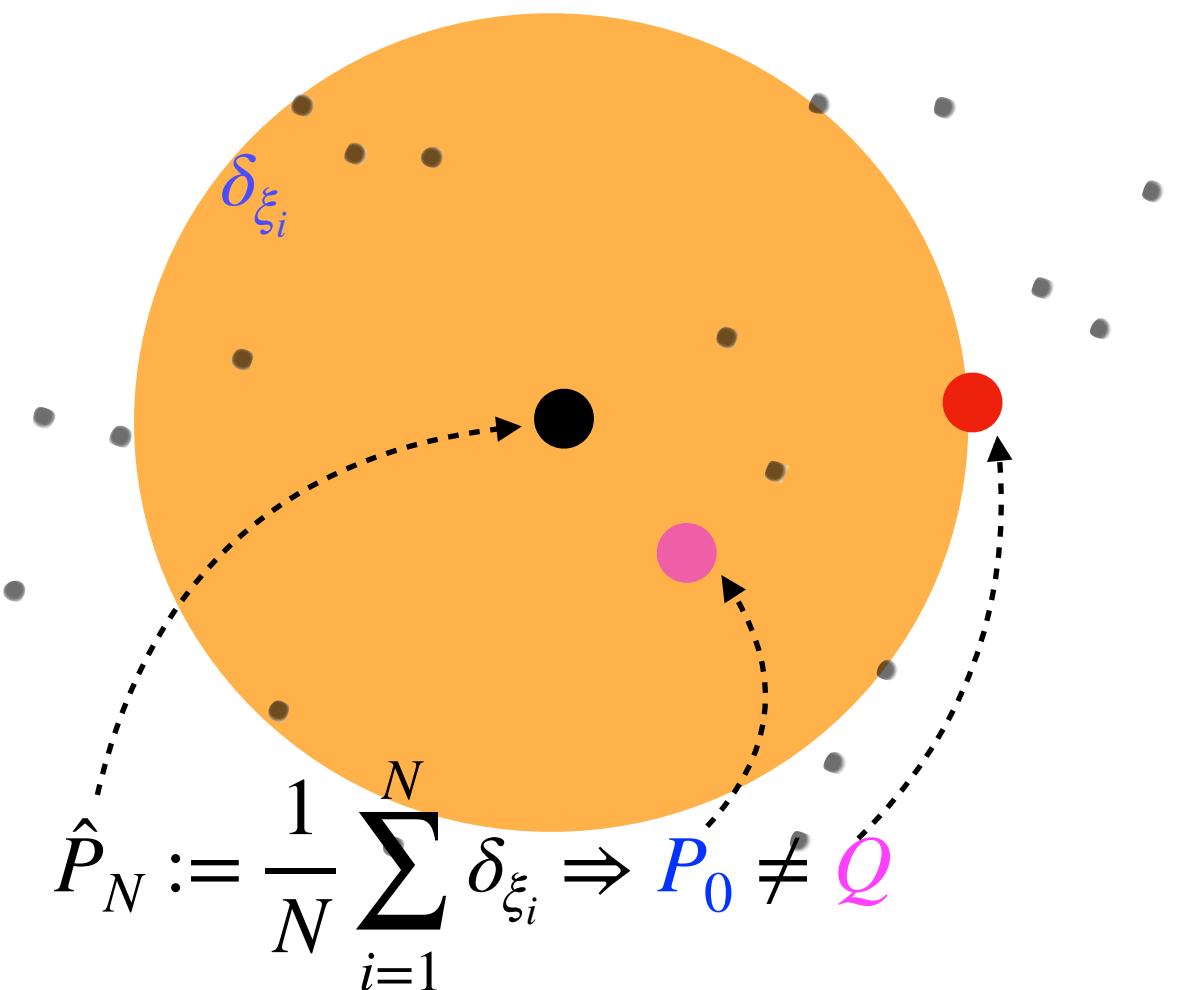
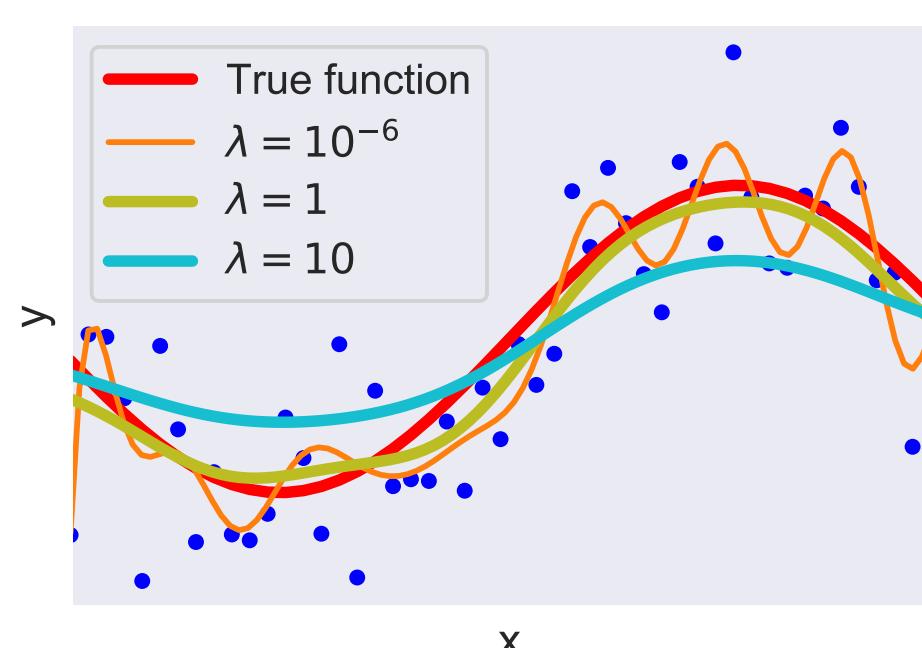
Empirical Risk Minimization Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

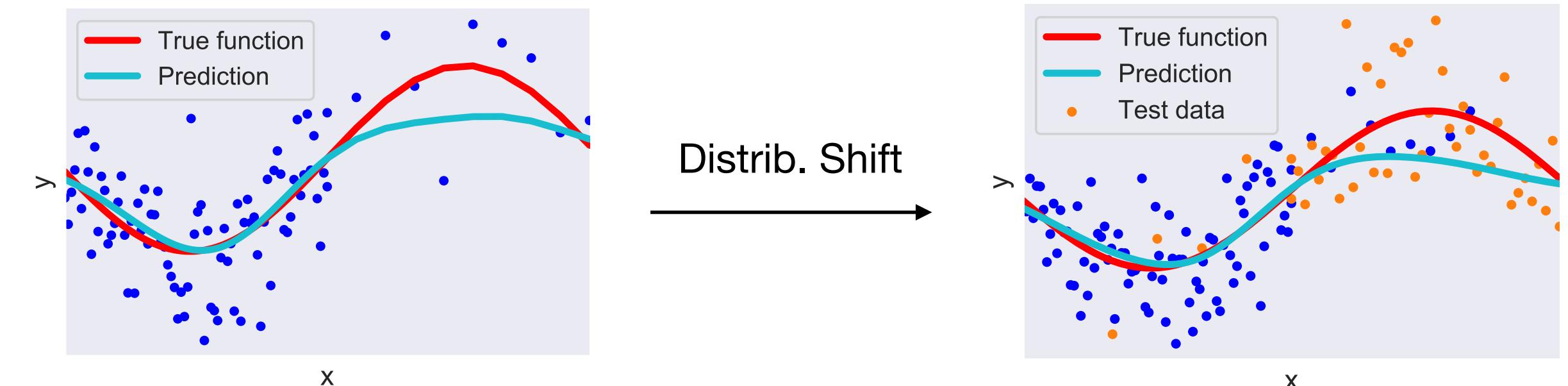
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts, when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M}
 [Delage & Ye 2010] in certain geometry.



Why study new geometry?

New geometries leading to new fields of research and breakthroughs:

From Statistical Learning to Distributionally Robust Learning

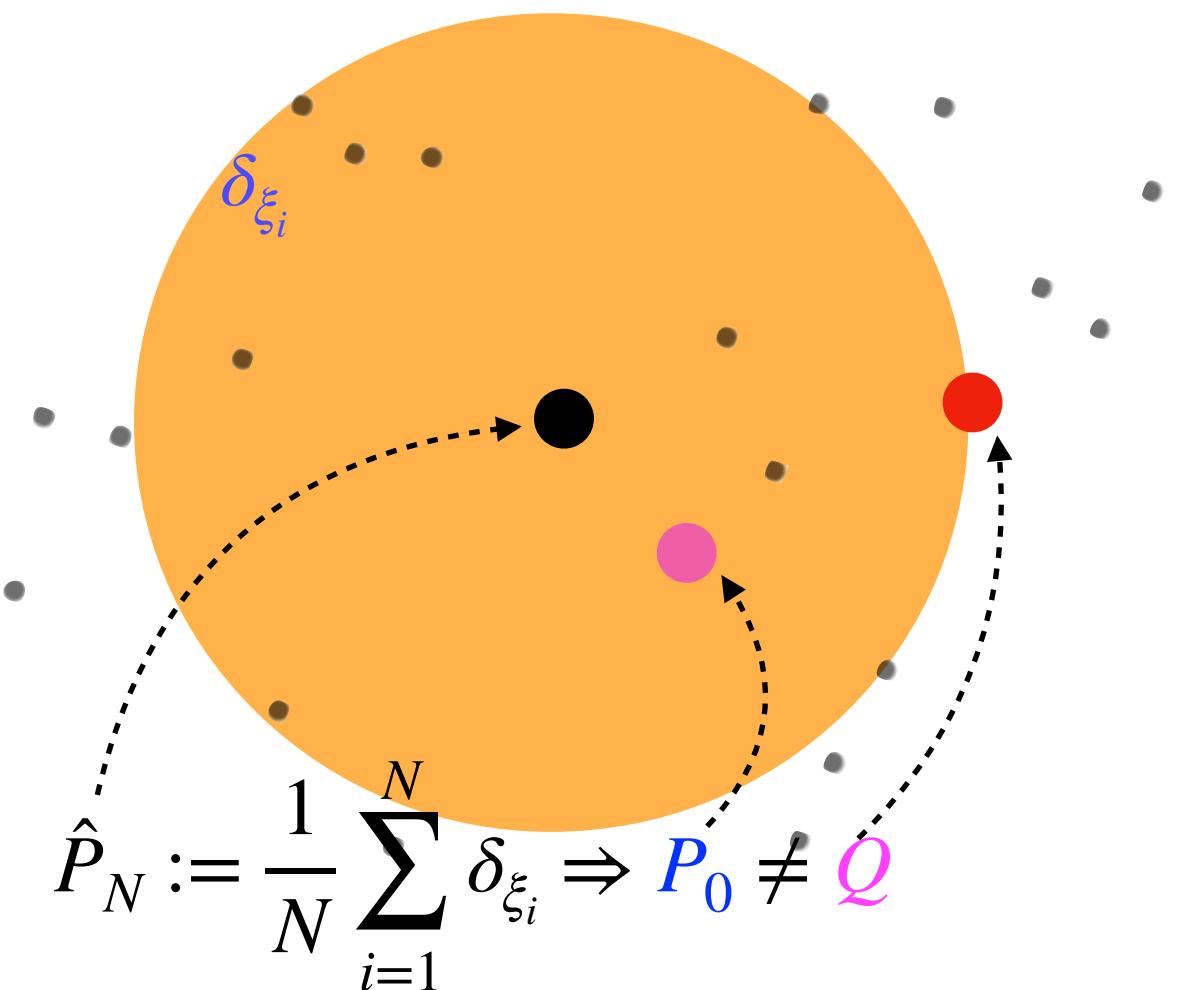
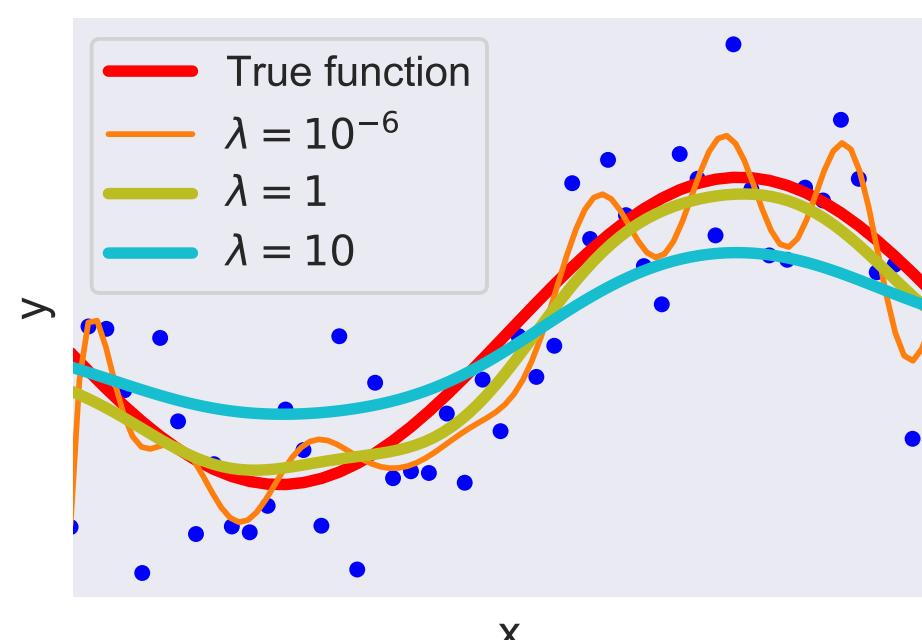
Empirical Risk Minimization Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

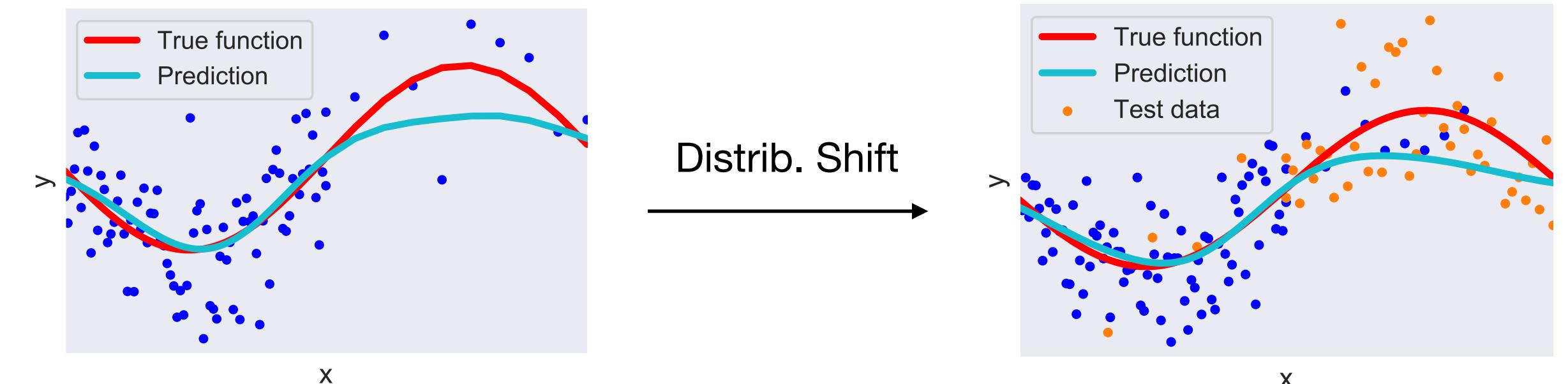
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts, when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M}
[Delage & Ye 2010] in certain geometry.



Why study new geometry?

New geometries leading to new fields of research and breakthroughs:

Information geometry [S. Amari et al.] e.g. natural gradient descent in Fisher-Rao geometry.

From Statistical Learning to Distributionally Robust Learning

Empirical Risk Minimization

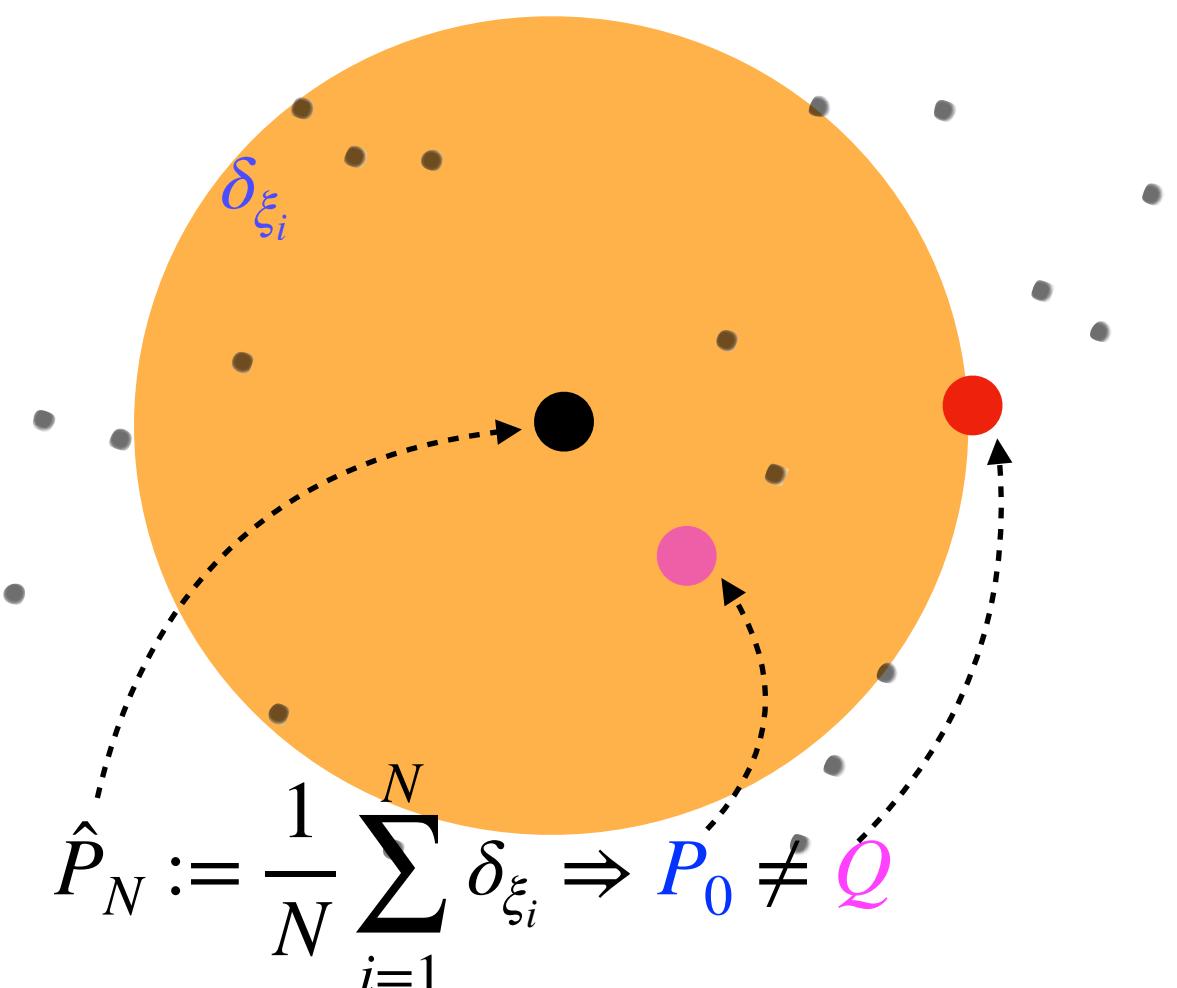
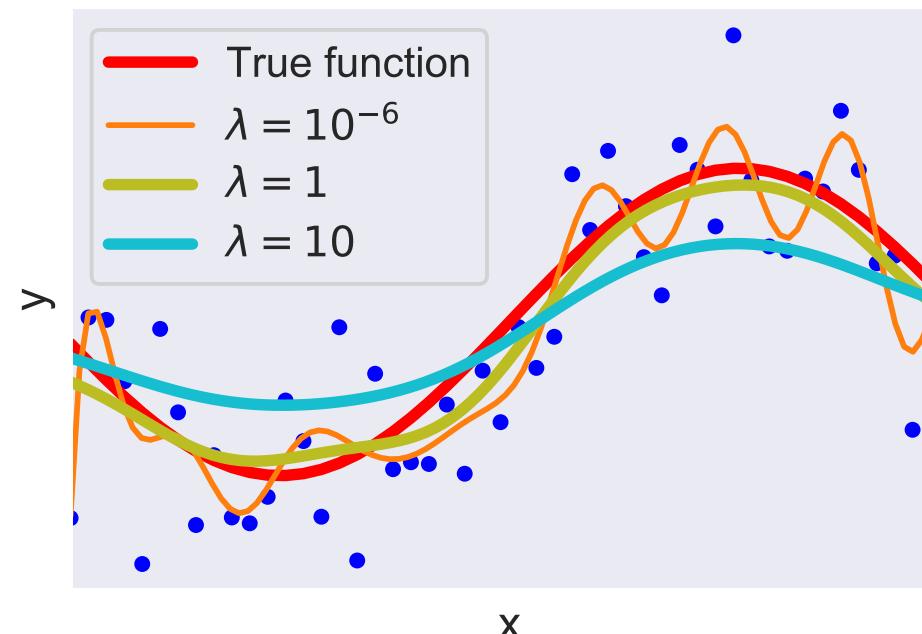
Distributionally Robust Optimization (DRO)

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^N l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- “Robust” under statistical fluctuation

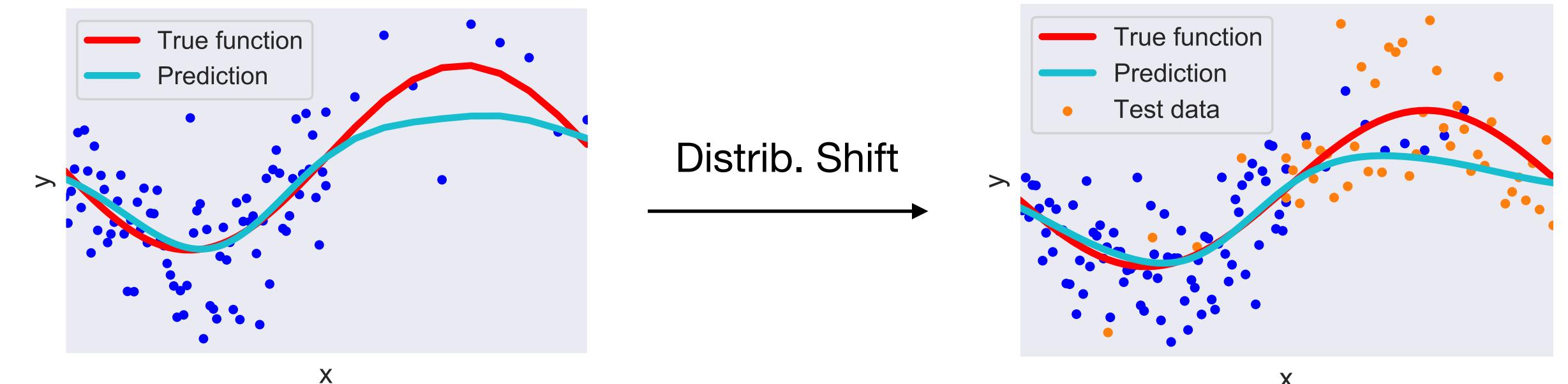
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N} \sum_{i=1}^N l(\hat{\theta}, \xi_i) + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$$

- Not robust under data distribution shifts, when $Q (\neq P_0)$



$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q l(\theta, \xi)$$

Worst-case distribution Q within the ambiguity set \mathcal{M} [Delage & Ye 2010] in certain geometry.



Why study new geometry?

New geometries leading to new fields of research and breakthroughs:

Information geometry [S. Amari et al.] e.g. natural gradient descent in Fisher-Rao geometry.

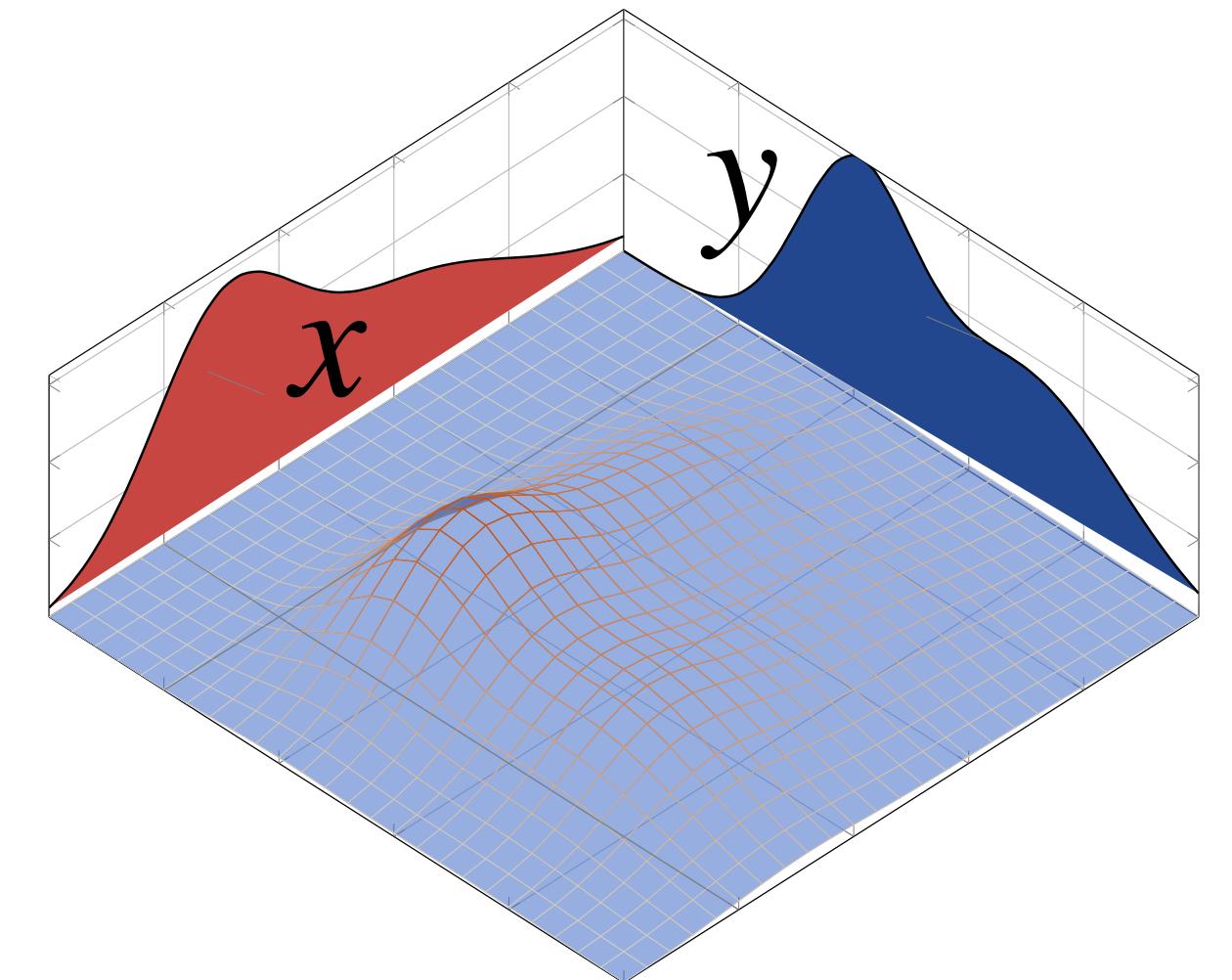
Wasserstein Gradient flow [F. Otto et al.] e.g. Fokker-Planck equation as GF in W_2

Background: Wasserstein Geometry

Background: Wasserstein Geometry

Definition. The p -**Wasserstein distance** between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

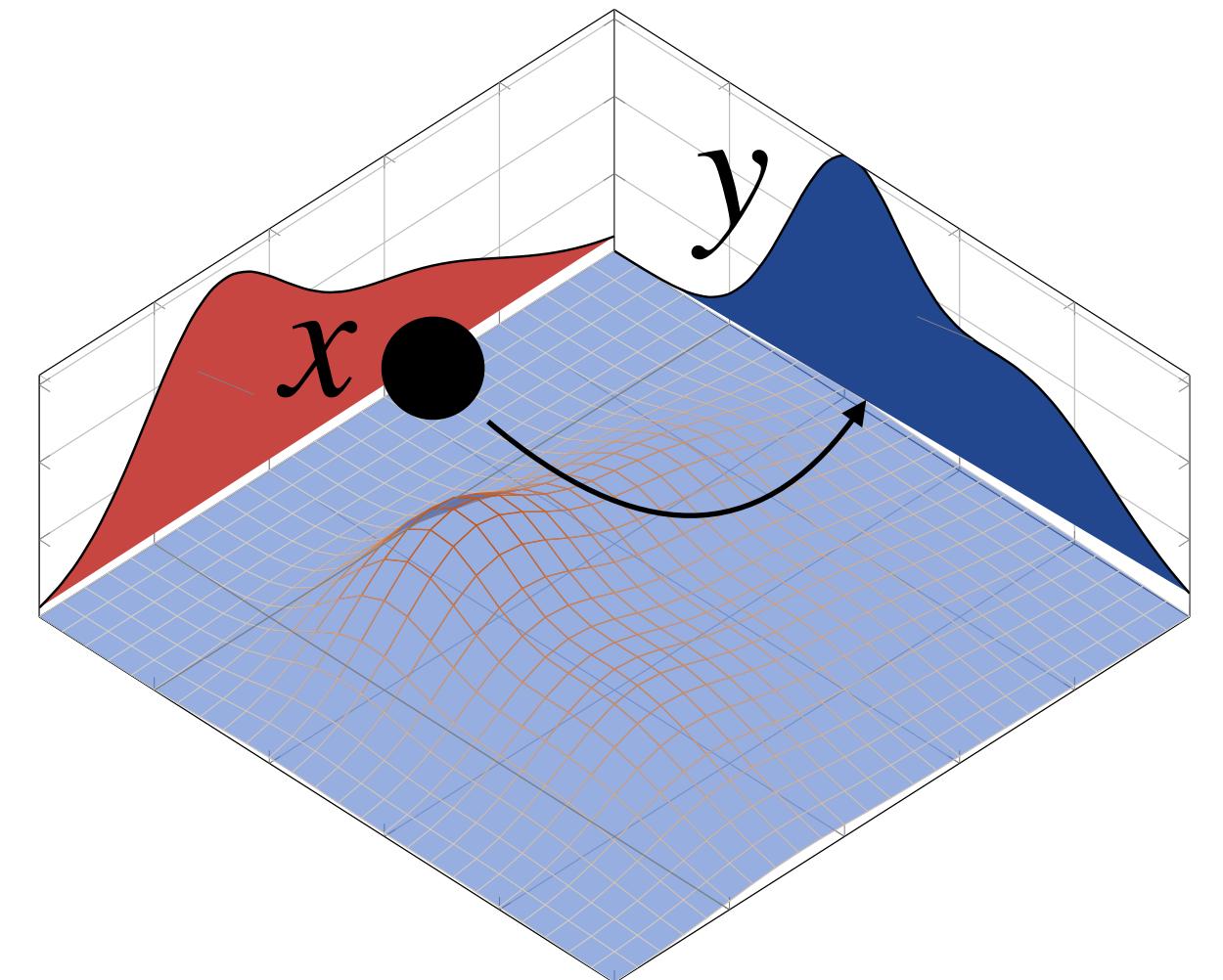
$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$



Background: Wasserstein Geometry

Definition. The p -Wasserstein distance between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

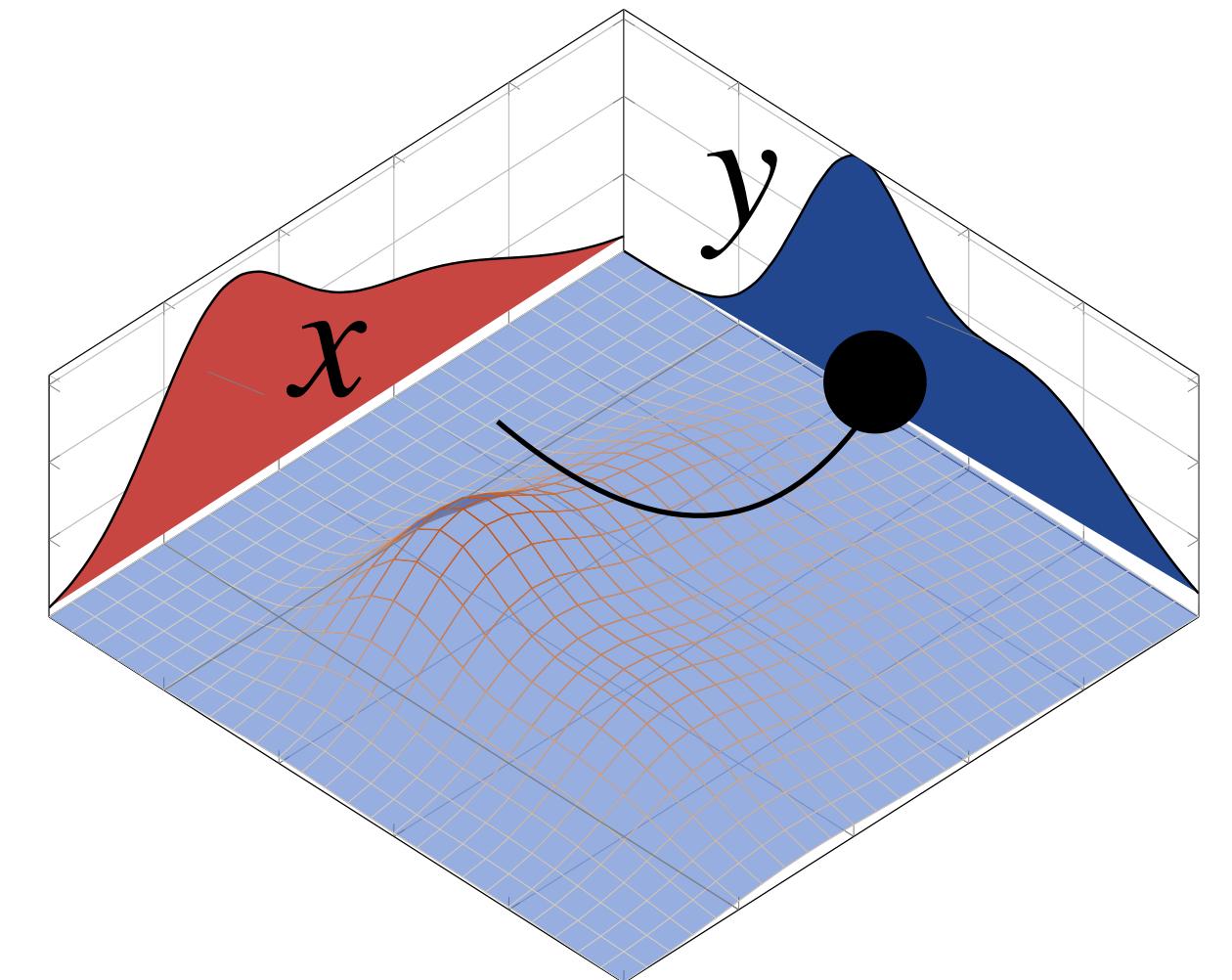
$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$



Background: Wasserstein Geometry

Definition. The p -**Wasserstein distance** between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$



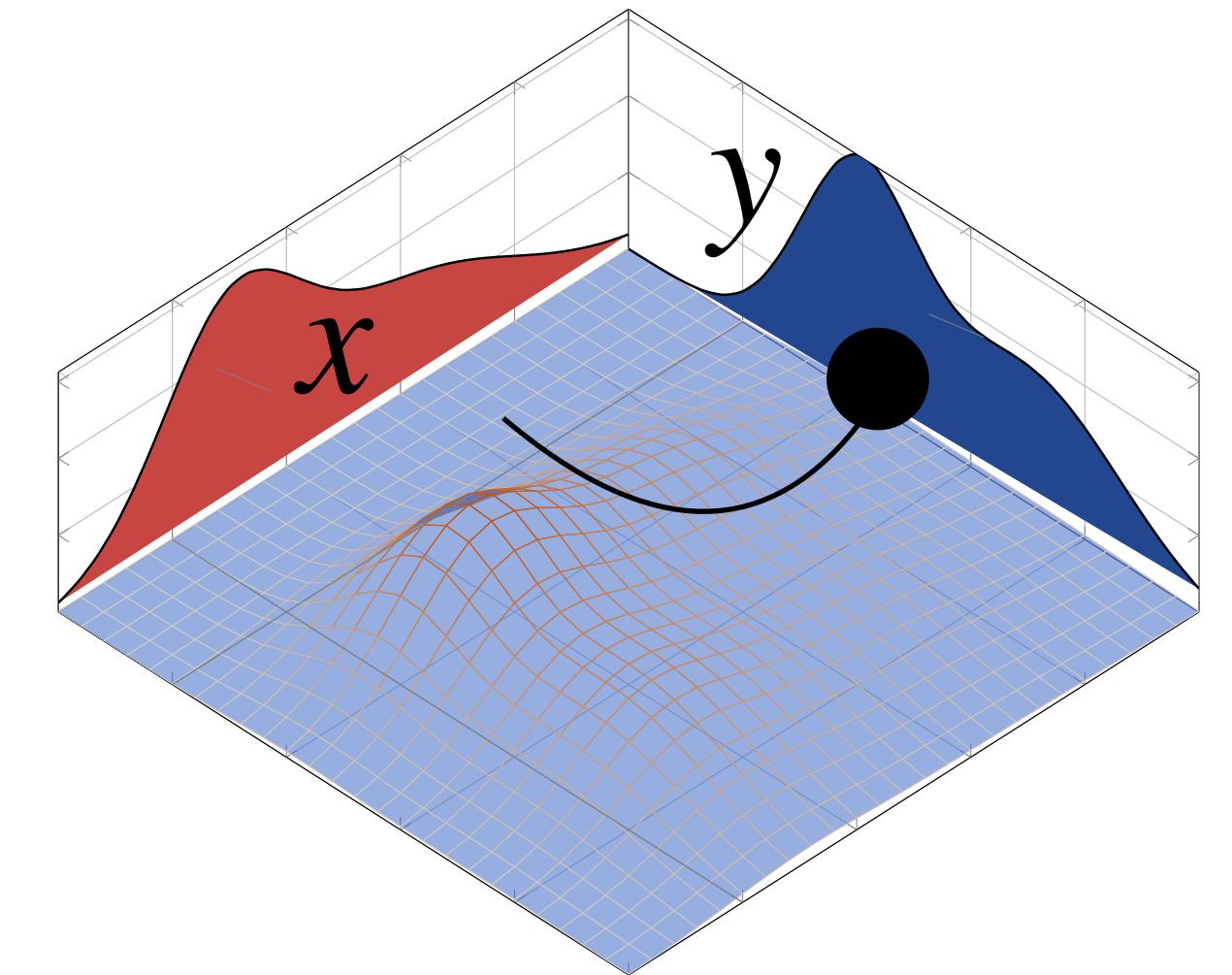
Background: Wasserstein Geometry

Definition. The p -Wasserstein distance between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$

$$\text{(Dual)} = \sup \left\{ \int \psi_1(x) dP(x) + \int \psi_2(y) dQ(y) \mid \psi_1 \oplus \psi_2 \leq c \text{ a.e.} \right\}$$

2-Wasserstein space $(\text{Prob}(\mathbb{R}^d), W_2)$ is a geodesic metric space.



Background: Wasserstein Geometry

Definition. The p -Wasserstein distance between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

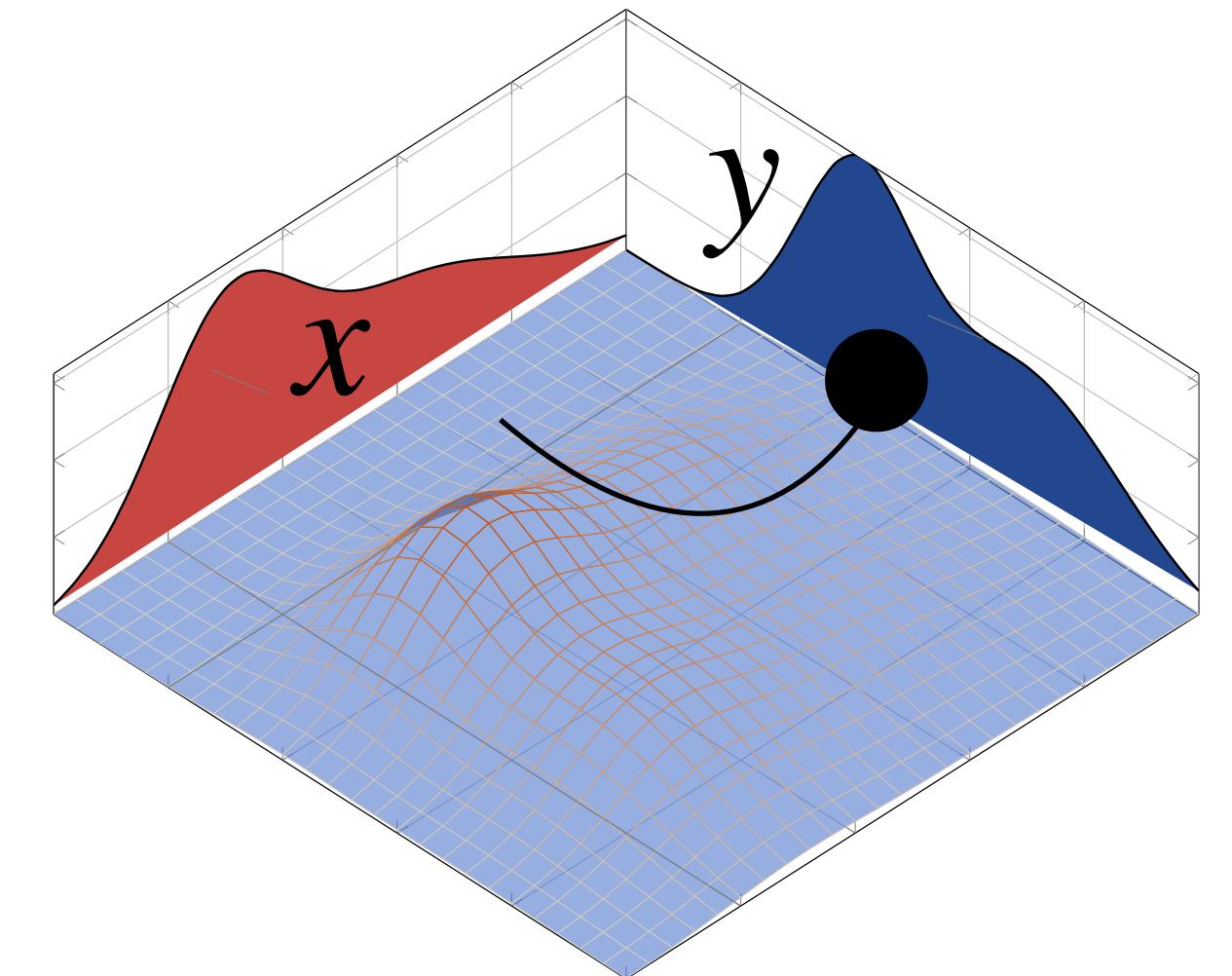
$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$

$$\text{(Dual)} = \sup \left\{ \int \psi_1(x) dP(x) + \int \psi_2(y) dQ(y) \mid \psi_1 \oplus \psi_2 \leq c \text{ a.e.} \right\}$$

2-Wasserstein space $(\text{Prob}(\mathbb{R}^d), W_2)$ is a geodesic metric space.

Dynamic formulation: Benamou–Brenier

$$W_2^2(P, Q) = \min \left\{ \int_0^1 \int |\nu_t|^2 d\mu_t dt \mid \mu_0 = P, \mu_1 = Q, \frac{d}{dt} \mu_t + \text{div}(\nu_t \mu_t) = 0 \right\}$$



Background: Wasserstein Geometry

Definition. The p -Wasserstein distance between probability measures P, Q on \mathbb{R}^d (with p finite moments, $p \geq 1$) is defined through the following Kantorovich problem

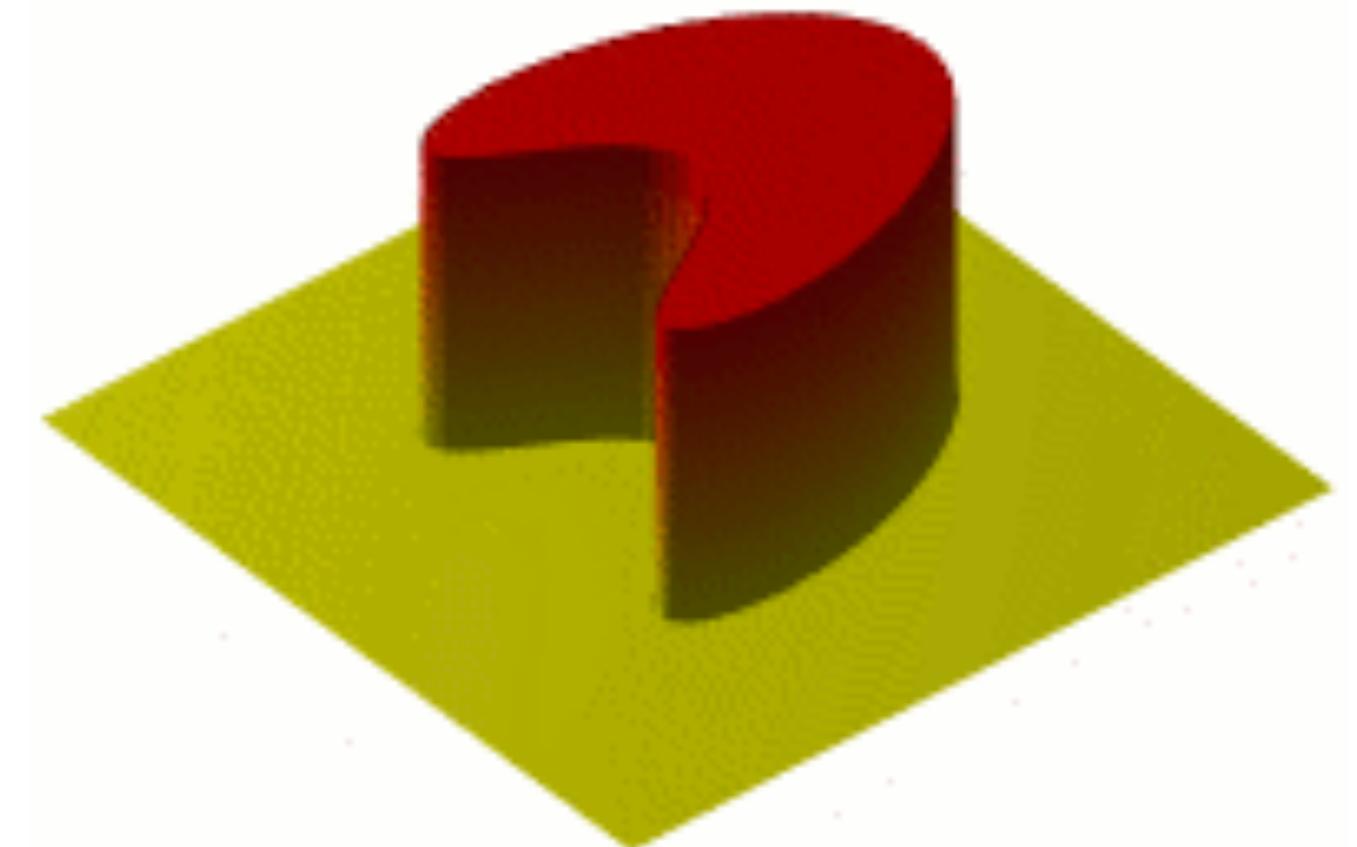
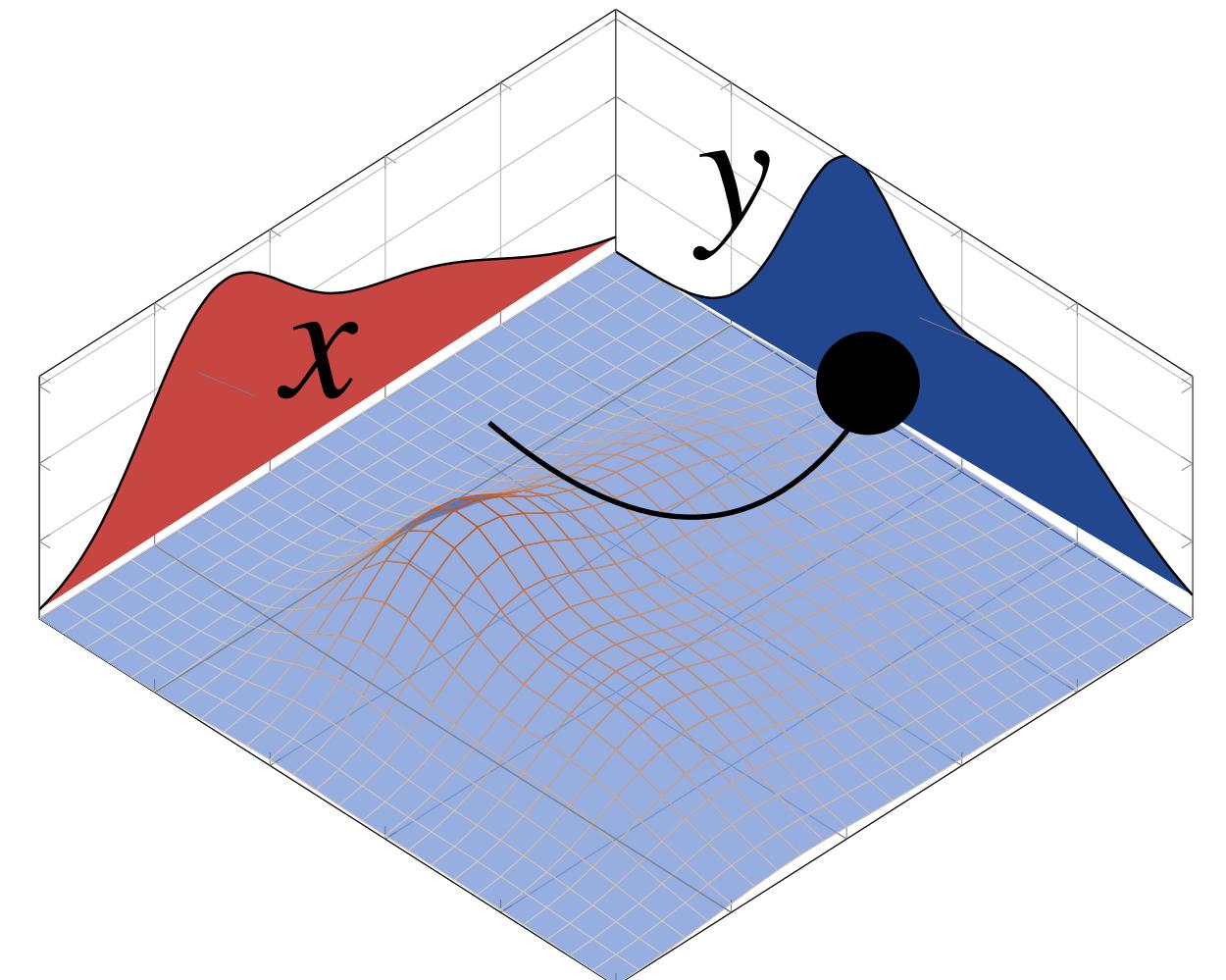
$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \mid \pi_{\#}^{(1)} \Pi = P, \pi_{\#}^{(2)} \Pi = Q \right\}$$

$$\text{(Dual)} = \sup \left\{ \int \psi_1(x) dP(x) + \int \psi_2(y) dQ(y) \mid \psi_1 \oplus \psi_2 \leq c \text{ a.e.} \right\}$$

2-Wasserstein space $(\text{Prob}(\mathbb{R}^d), W_2)$ is a geodesic metric space.

Dynamic formulation: Benamou–Brenier

$$W_2^2(P, Q) = \min \left\{ \int_0^1 \int |\nu_t|^2 d\mu_t dt \mid \mu_0 = P, \mu_1 = Q, \frac{d}{dt} \mu_t + \text{div}(\nu_t \mu_t) = 0 \right\}$$



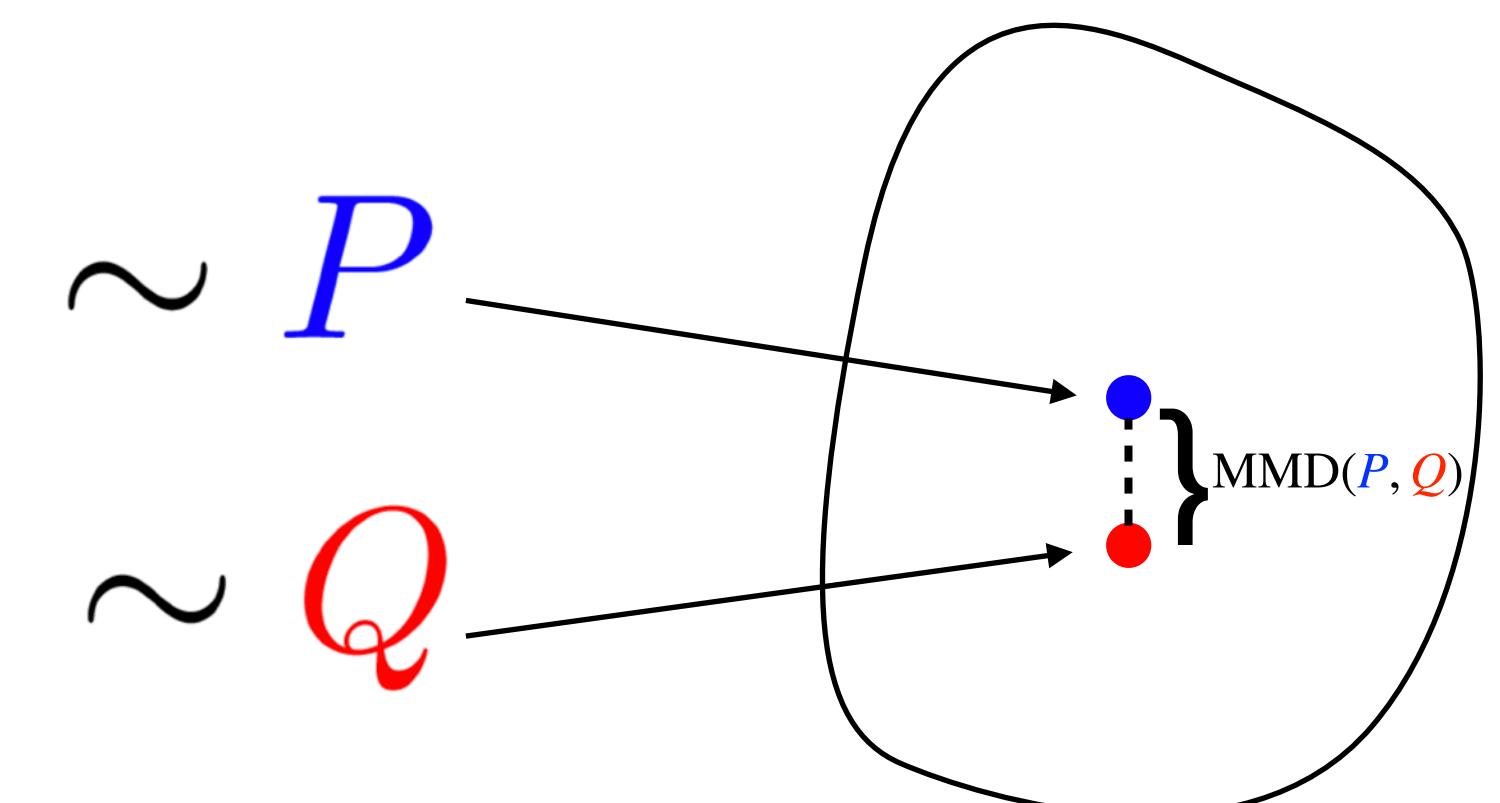
Background: “Kernel Geometry”

Background: “Kernel Geometry”

Definition. Kernel **Maximum-Mean Discrepancy (MMD)** associated with (PSD) kernel k (e.g., $k(x, x') := e^{-|x-x'|^2/\sigma^2}$)

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) := \left\| \int k(x, \cdot) d\mathcal{P} - \int k(x, \cdot) d\mathcal{Q} \right\|_{\mathcal{H}}.$$

$(\text{Prob}(\mathbb{R}^d), \text{MMD})$ is a (simple) metric space.



Background: “Kernel Geometry”

Definition. Kernel Maximum-Mean Discrepancy (MMD)
associated with (PSD) kernel k (e.g., $k(x, x') := e^{-|x-x'|^2/\sigma}$)

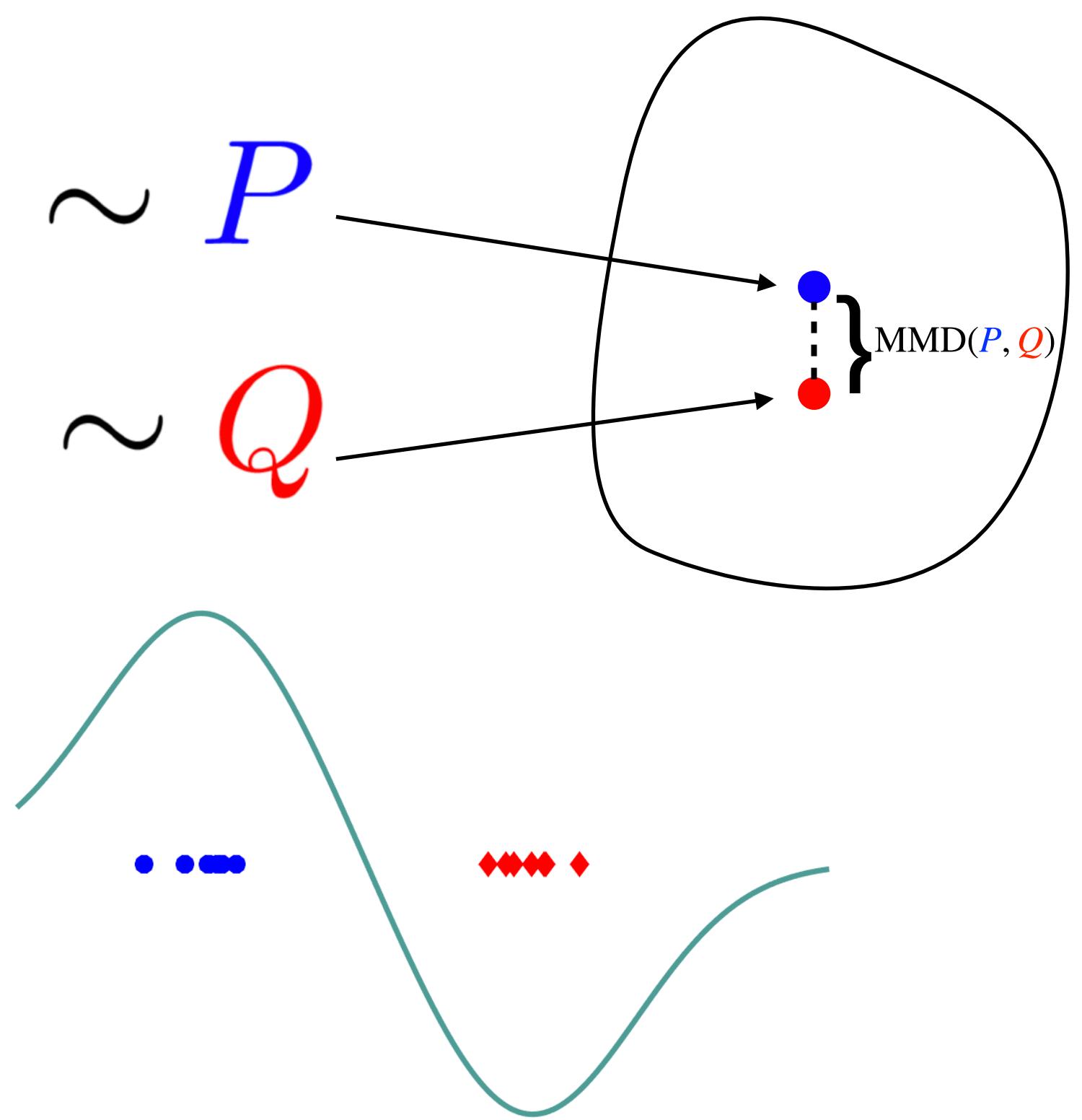
$$\text{MMD}(\mathcal{P}, \mathcal{Q}) := \left\| \int k(x, \cdot) d\mathcal{P} - \int k(x, \cdot) d\mathcal{Q} \right\|_{\mathcal{H}}.$$

$(\text{Prob}(\mathbb{R}^d), \text{MMD})$ is a (simple) metric space.

Dual formulation as an integral probability metric.

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f d(\mathcal{P} - \mathcal{Q})$$

\mathcal{H} is the **reproducing kernel Hilbert space** \mathcal{H} (RKHS),
which satisfies $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$,
 $\phi(x) := k(x, \cdot)$ is the canonical feature of \mathcal{H} .



Background: “Kernel Geometry”

Definition. Kernel Maximum-Mean Discrepancy (MMD)
associated with (PSD) kernel k (e.g., $k(x, x') := e^{-|x-x'|^2/\sigma}$)

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) := \left\| \int k(x, \cdot) d\mathcal{P} - \int k(x, \cdot) d\mathcal{Q} \right\|_{\mathcal{H}}.$$

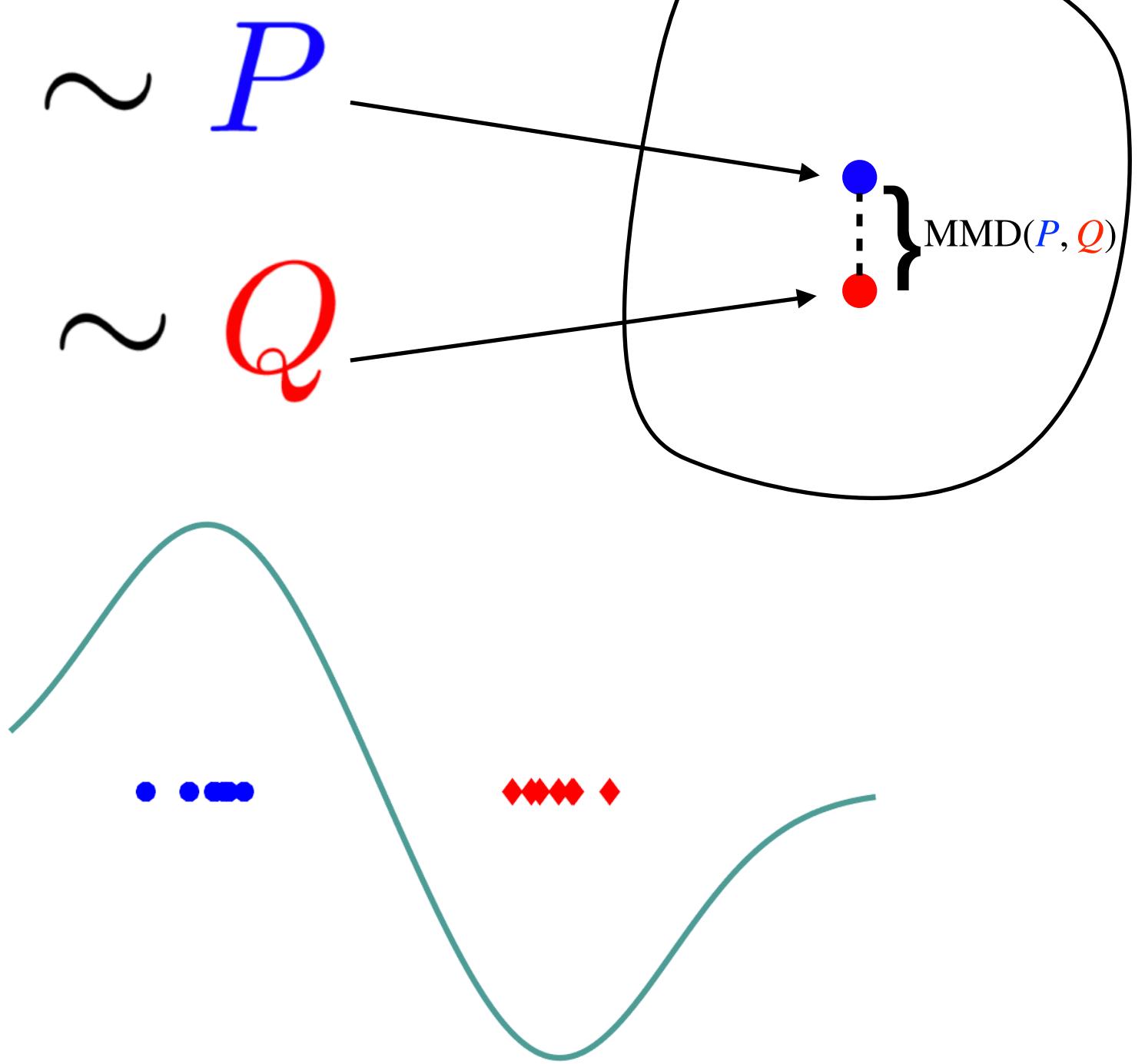
$(\text{Prob}(\mathbb{R}^d), \text{MMD})$ is a (simple) metric space.

Dual formulation as an integral probability metric.

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f d(\mathcal{P} - \mathcal{Q})$$

\mathcal{H} is the **reproducing kernel Hilbert space** \mathcal{H} (RKHS),
which satisfies $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$,
 $\phi(x) := k(x, \cdot)$ is the canonical feature of \mathcal{H} .

Example. Approx. nonlinear functions



Background: “Kernel Geometry”

Definition. Kernel Maximum-Mean Discrepancy (MMD)
associated with (PSD) kernel k (e.g., $k(x, x') := e^{-|x-x'|^2/\sigma^2}$)

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) := \left\| \int k(x, \cdot) d\mathcal{P} - \int k(x, \cdot) d\mathcal{Q} \right\|_{\mathcal{H}}.$$

$(\text{Prob}(\mathbb{R}^d), \text{MMD})$ is a (simple) metric space.

Dual formulation as an integral probability metric.

$$\text{MMD}(\mathcal{P}, \mathcal{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f d(\mathcal{P} - \mathcal{Q})$$

\mathcal{H} is the **reproducing kernel Hilbert space** \mathcal{H} (RKHS),
which satisfies $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$,
 $\phi(x) := k(x, \cdot)$ is the canonical feature of \mathcal{H} .

Example. Approx. nonlinear functions

$$\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N \|f(x_i) - y_i\|^2 + \lambda \|f\|_{\mathcal{H}}^2$$

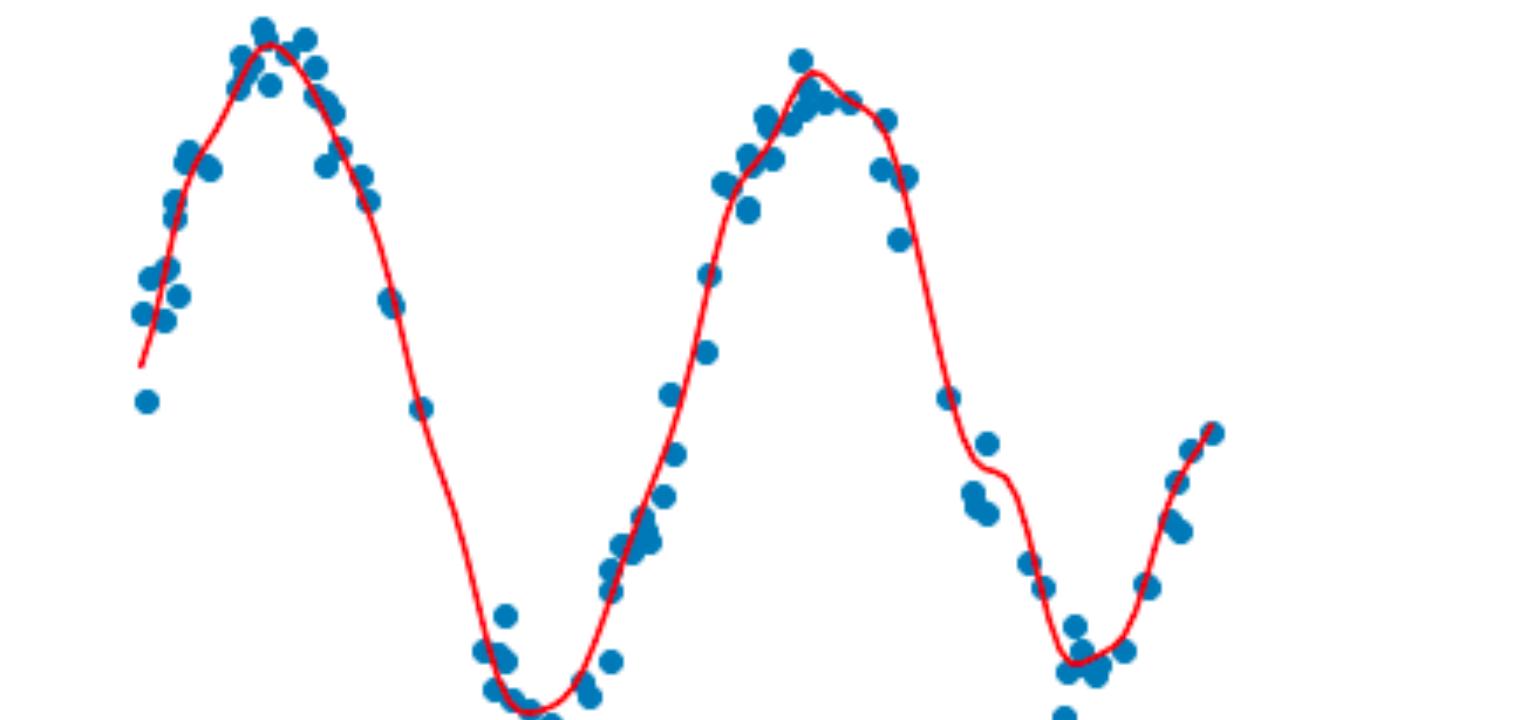
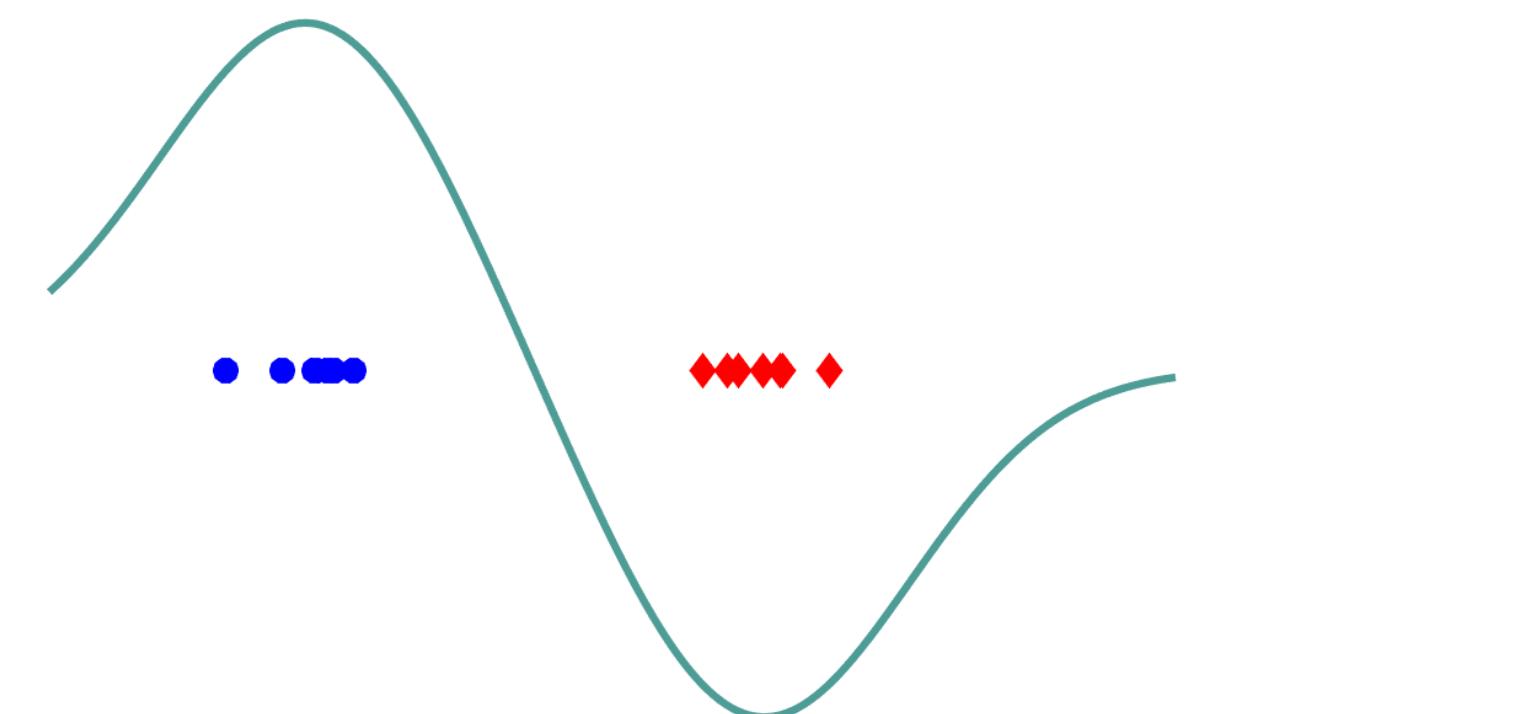
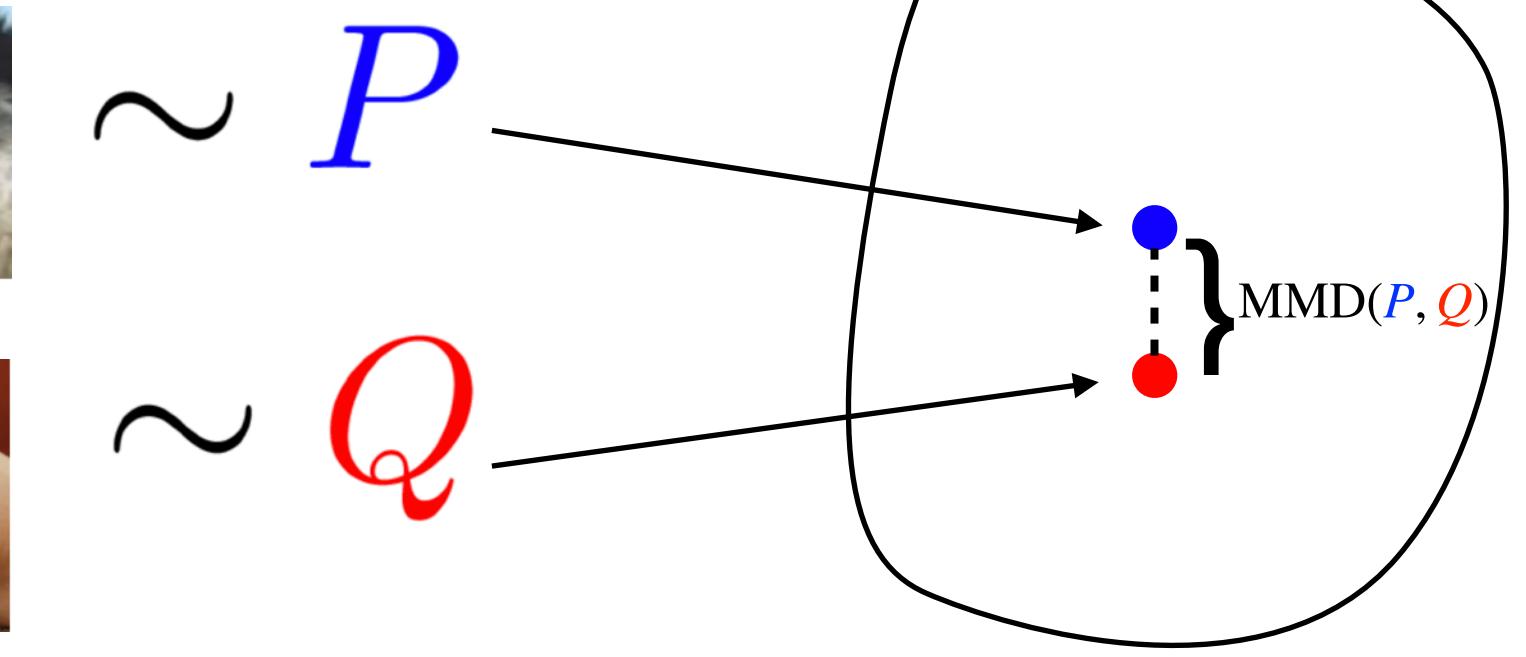


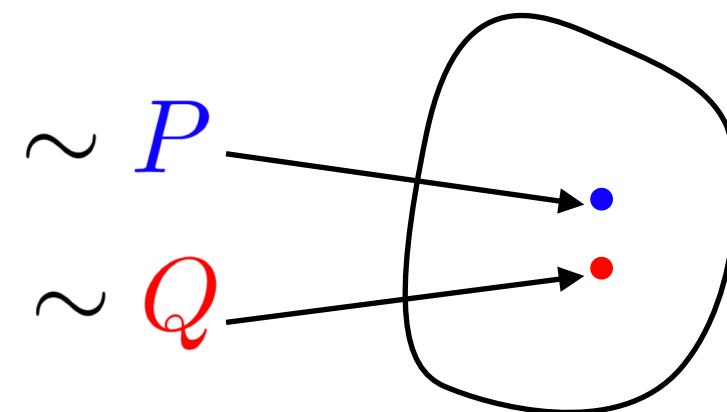
Figure credit: W. Jitkrittum, J. Zhu

Kernel distributionally robust optimization

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

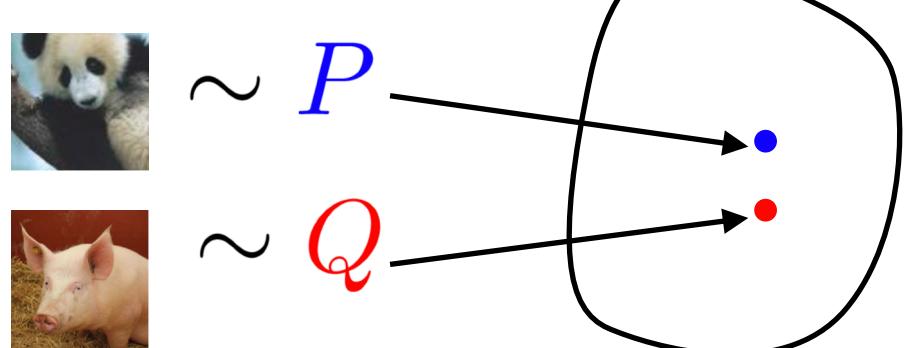
$$(\text{DRO}) \quad \min_{\theta} \sup_{\substack{\mathbb{E}_{\mathcal{Q}} l(\theta, \xi) \\ \text{MMD}(\mathcal{Q}, \hat{P}) \leq \epsilon}} \mathbb{E}_{\mathcal{Q}} l(\theta, \xi)$$



Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(\text{DRO}) \quad \min_{\theta} \sup_{\substack{\mathbb{E}_{\mathcal{Q}} l(\theta, \xi) \\ \text{MMD}(\mathcal{Q}, \hat{P}) \leq \epsilon}} \mathbb{E}_{\mathcal{Q}} l(\theta, \xi)$$



Kernel DRO Theorem (simplified). [Z. et al. 2021]

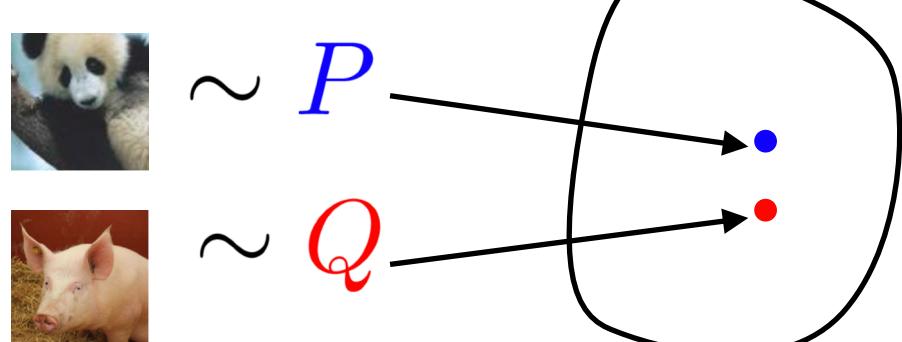
DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

$$(K) \quad \min_{\theta, \mathbf{f} \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N \mathbf{f}(\xi_i) + \epsilon \|\mathbf{f}\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq \mathbf{f}$$

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(\text{DRO}) \quad \min_{\theta} \sup_{\substack{\mathbb{E}_Q l(\theta, \xi) \\ \text{MMD}(Q, \hat{P}) \leq \epsilon}} \mathbb{E}_Q l(\theta, \xi)$$

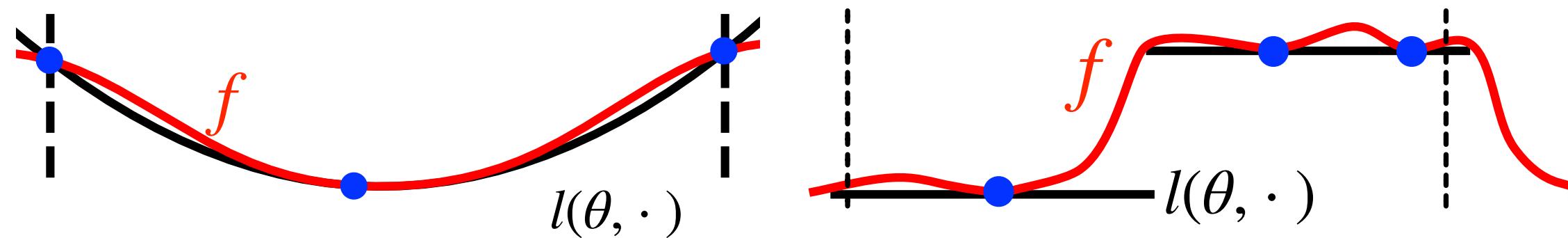


Kernel DRO Theorem (simplified). [Z. et al. 2021]

DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

$$(K) \quad \min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

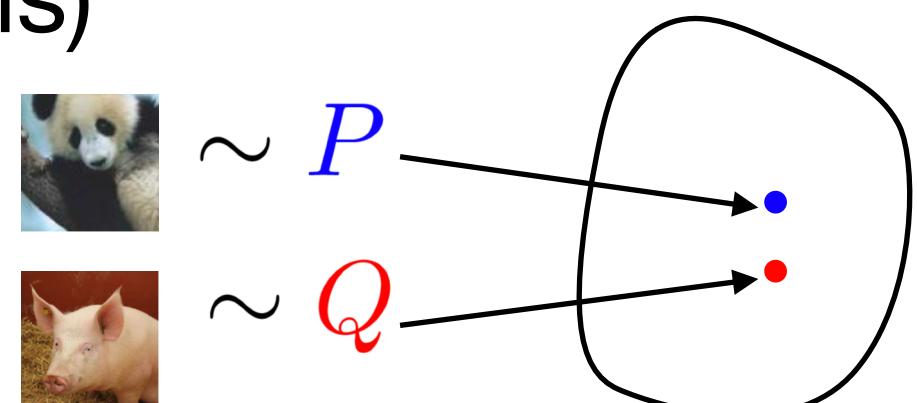
Geometric intuition: **dual kernel function f** as robust surrogate losses (flatten the curve)



Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(\text{DRO}) \quad \min_{\theta} \sup_{\substack{\text{MMD}(Q, \hat{P}) \leq \epsilon}} \mathbb{E}_Q l(\theta, \xi)$$



Kernel DRO Theorem (simplified). [Z. et al. 2021]

DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

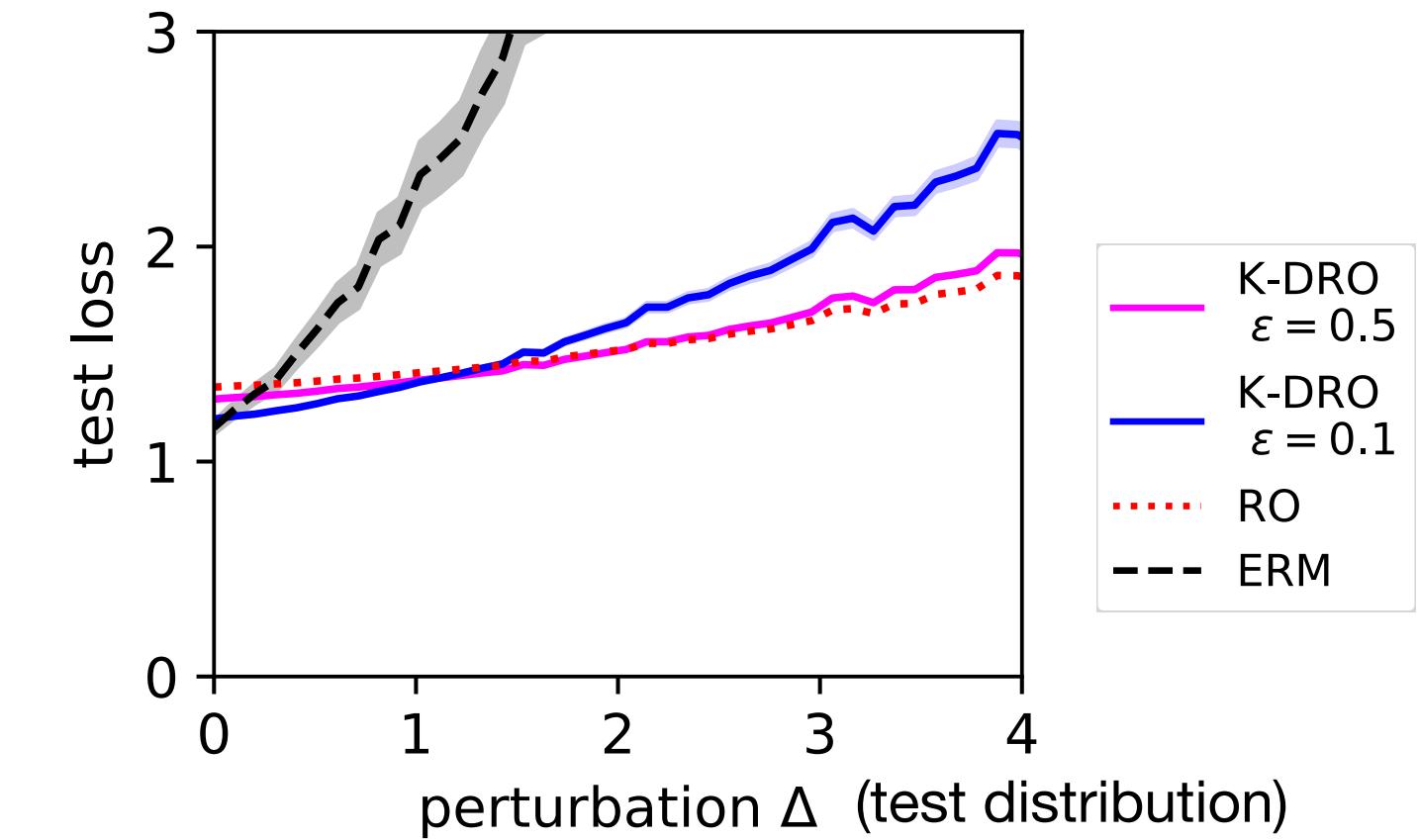
$$(K) \quad \min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

Example. Robust least squares

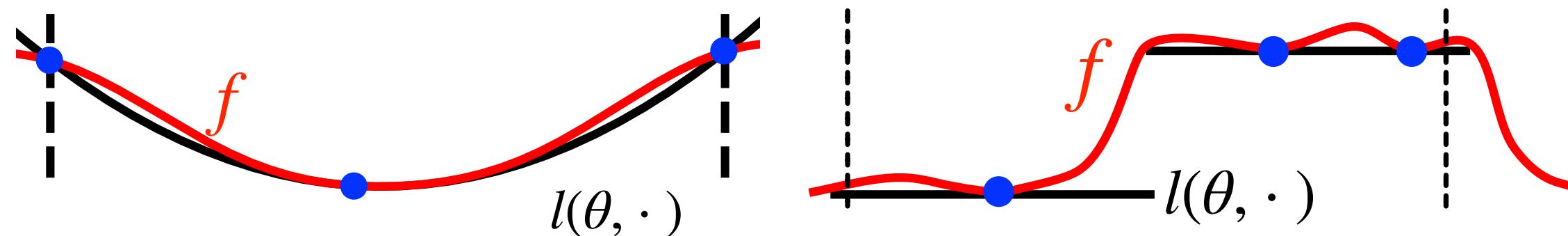
[El Ghaoui Lebret '97]

$$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$

Given historical samples $\xi_1, \xi_2, \dots, \xi_N$



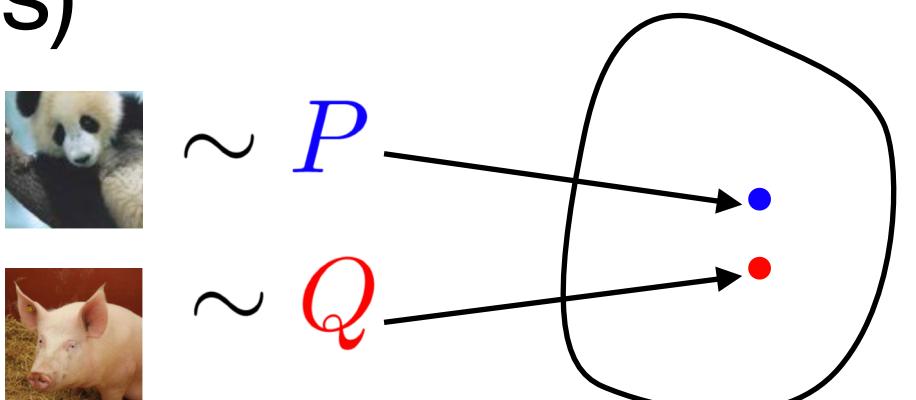
Geometric intuition: **dual kernel function f** as robust surrogate losses (flatten the curve)



Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(DRO) \min_{\theta} \sup_{\substack{\text{MMD}(Q, \hat{P}) \leq \epsilon}} \mathbb{E}_Q l(\theta, \xi)$$

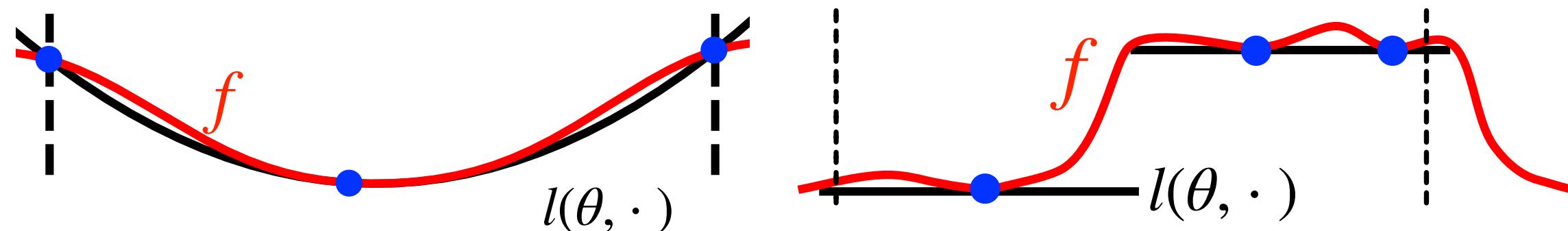


Kernel DRO Theorem (simplified). [Z. et al. 2021]

DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

$$(K) \min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

Geometric intuition: **dual kernel function f** as robust surrogate losses (flatten the curve)

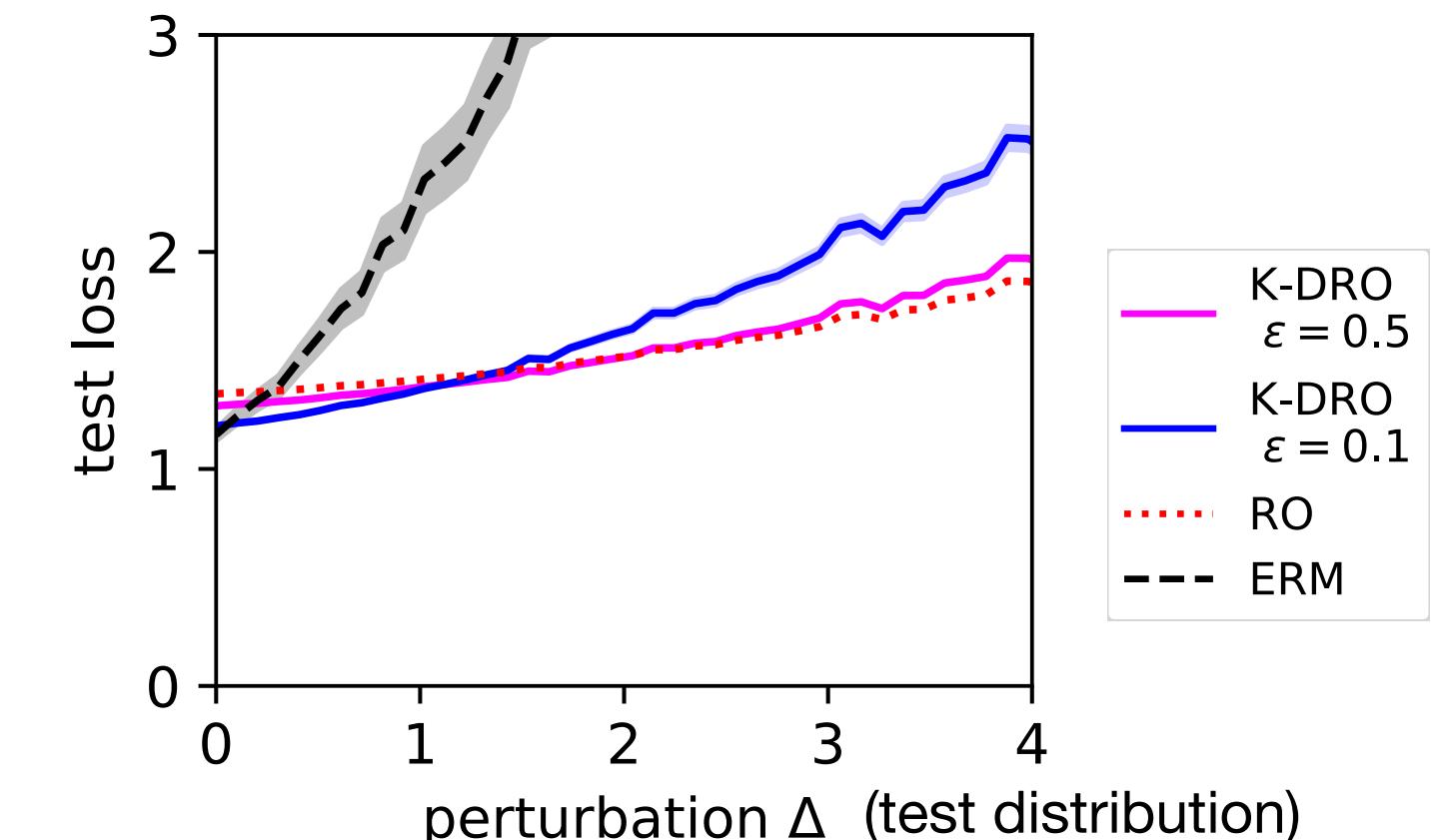


Example. Robust least squares

[El Ghaoui Lebret '97]

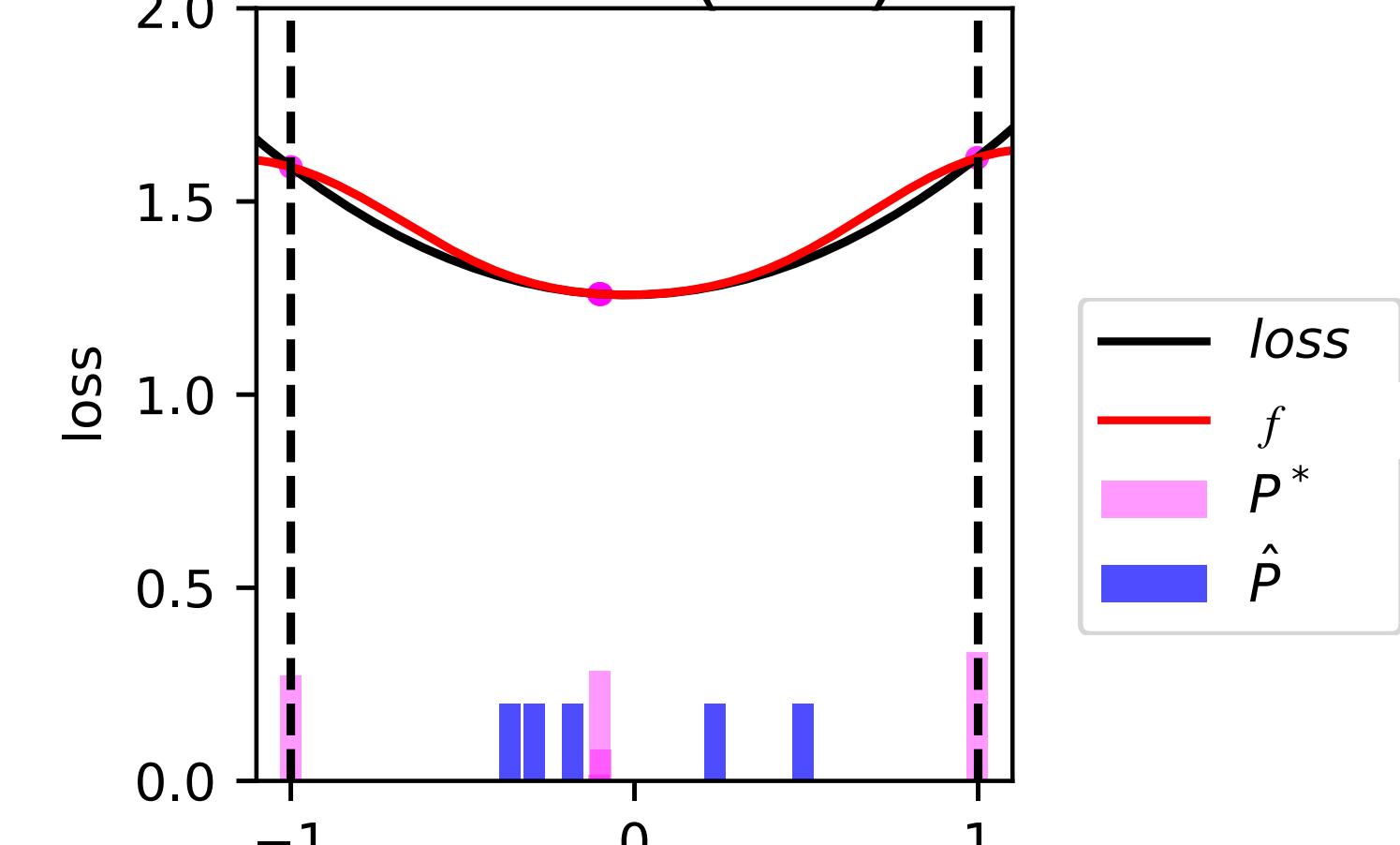
$$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$

Given historical samples $\xi_1, \xi_2, \dots, \xi_N$



Robustifying with DRO

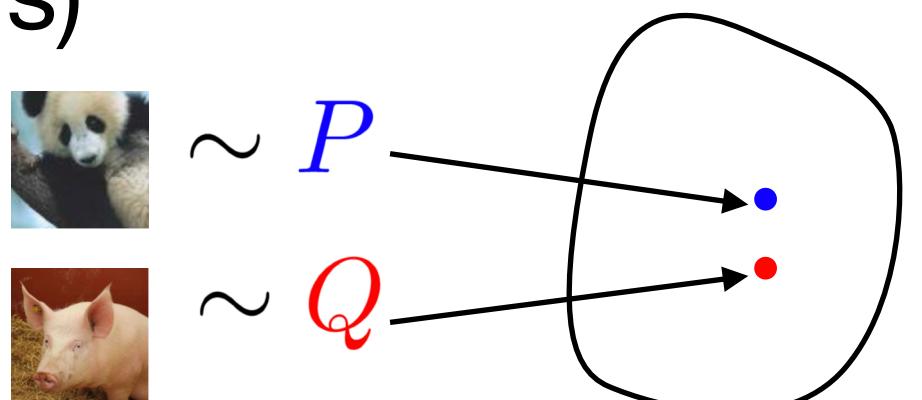
f as witness (test) function



Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(DRO) \min_{\theta} \sup_{\substack{\text{MMD}(Q, \hat{P}) \leq \epsilon}} \mathbb{E}_Q l(\theta, \xi)$$

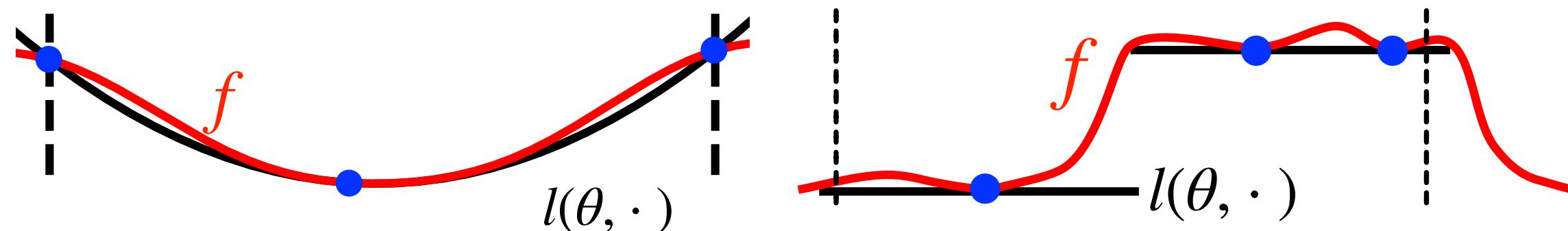


Kernel DRO Theorem (simplified). [Z. et al. 2021]

DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

$$(K) \min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

Geometric intuition: **dual kernel function f** as robust surrogate losses (flatten the curve)

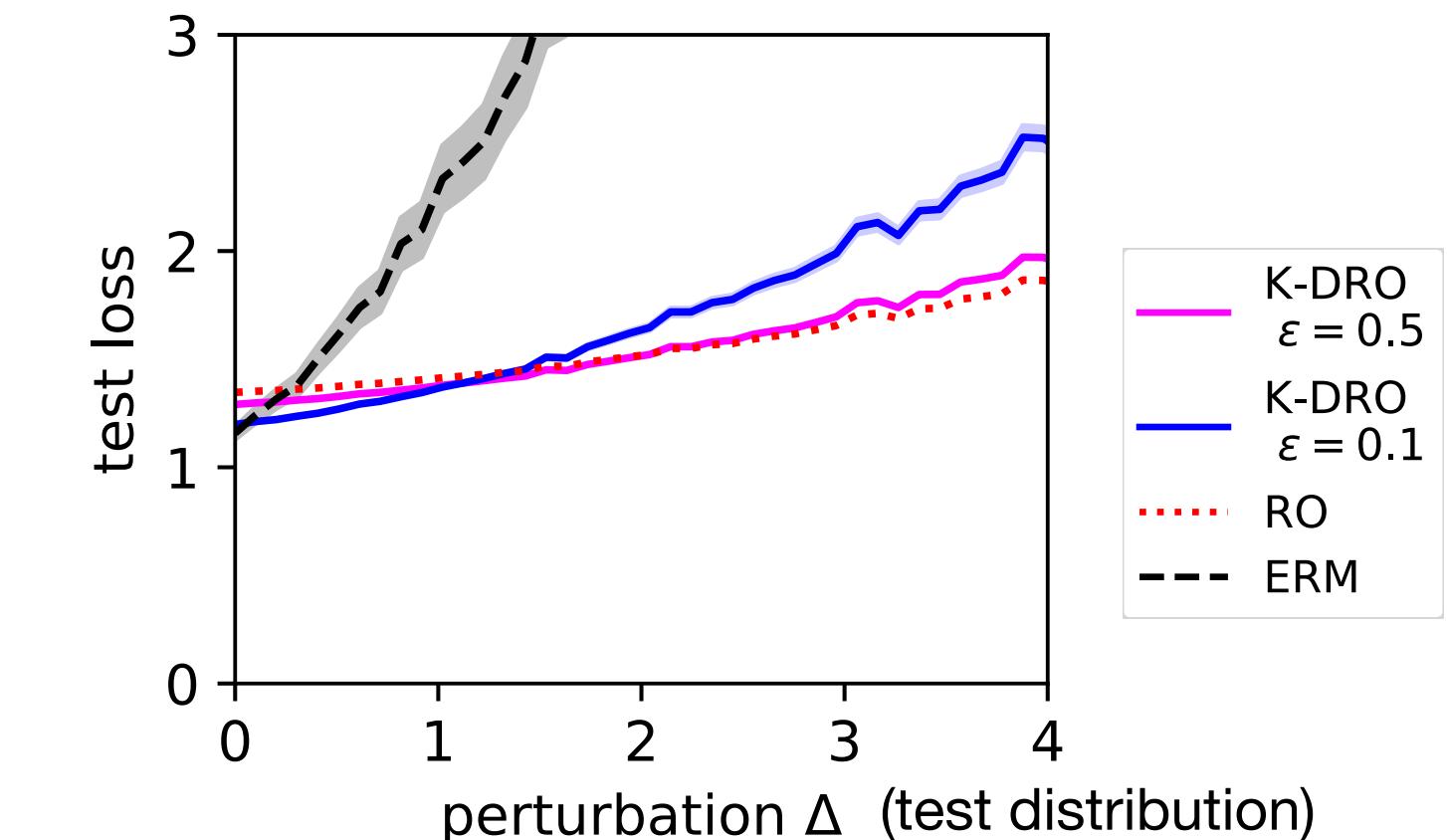


Example. Robust least squares

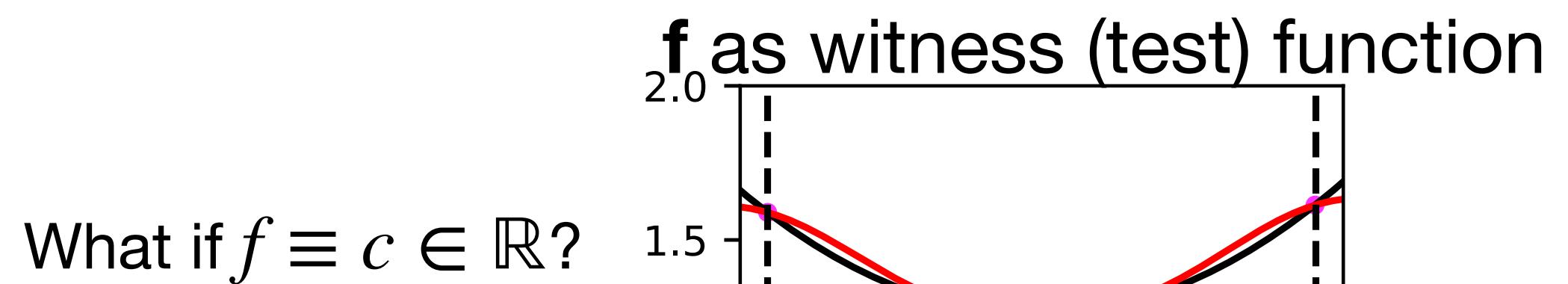
[El Ghaoui Lebret '97]

$$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$

Given historical samples $\xi_1, \xi_2, \dots, \xi_N$



Robustifying with DRO

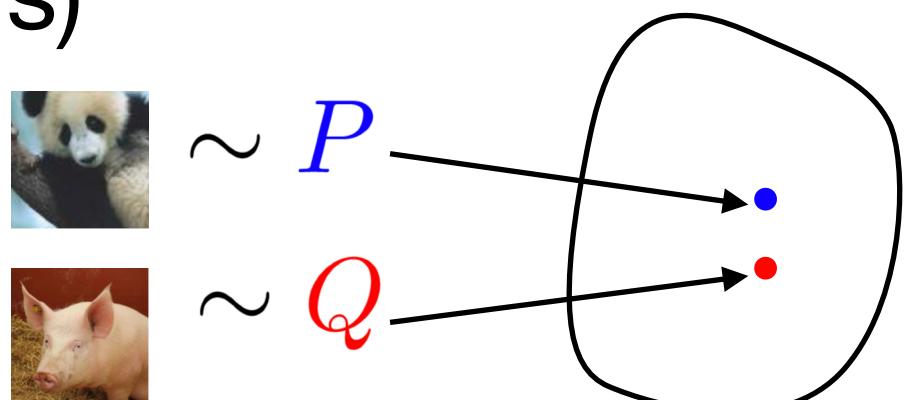


What if $f \equiv c \in \mathbb{R}$?

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

$$(DRO) \min_{\theta} \sup_{\substack{\text{MMD}(Q, \hat{P}) \leq \epsilon}} \mathbb{E}_Q l(\theta, \xi)$$

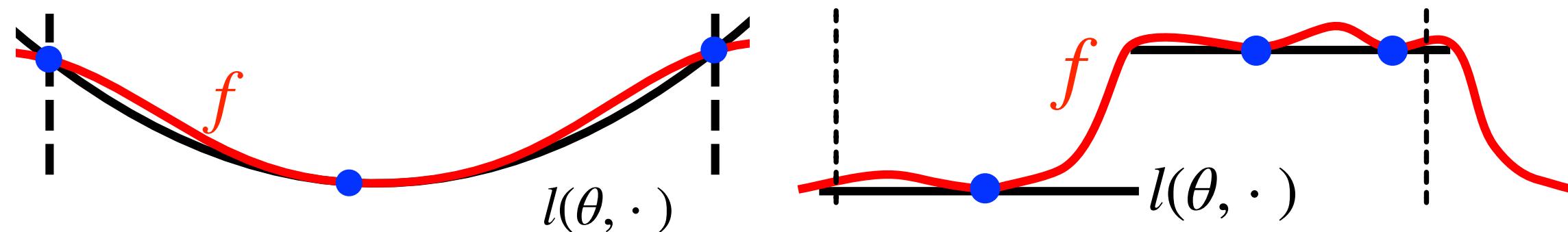


Kernel DRO Theorem (simplified). [Z. et al. 2021]

DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

$$(K) \min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot) \leq f$$

Geometric intuition: **dual kernel function** f as robust surrogate losses (flatten the curve)

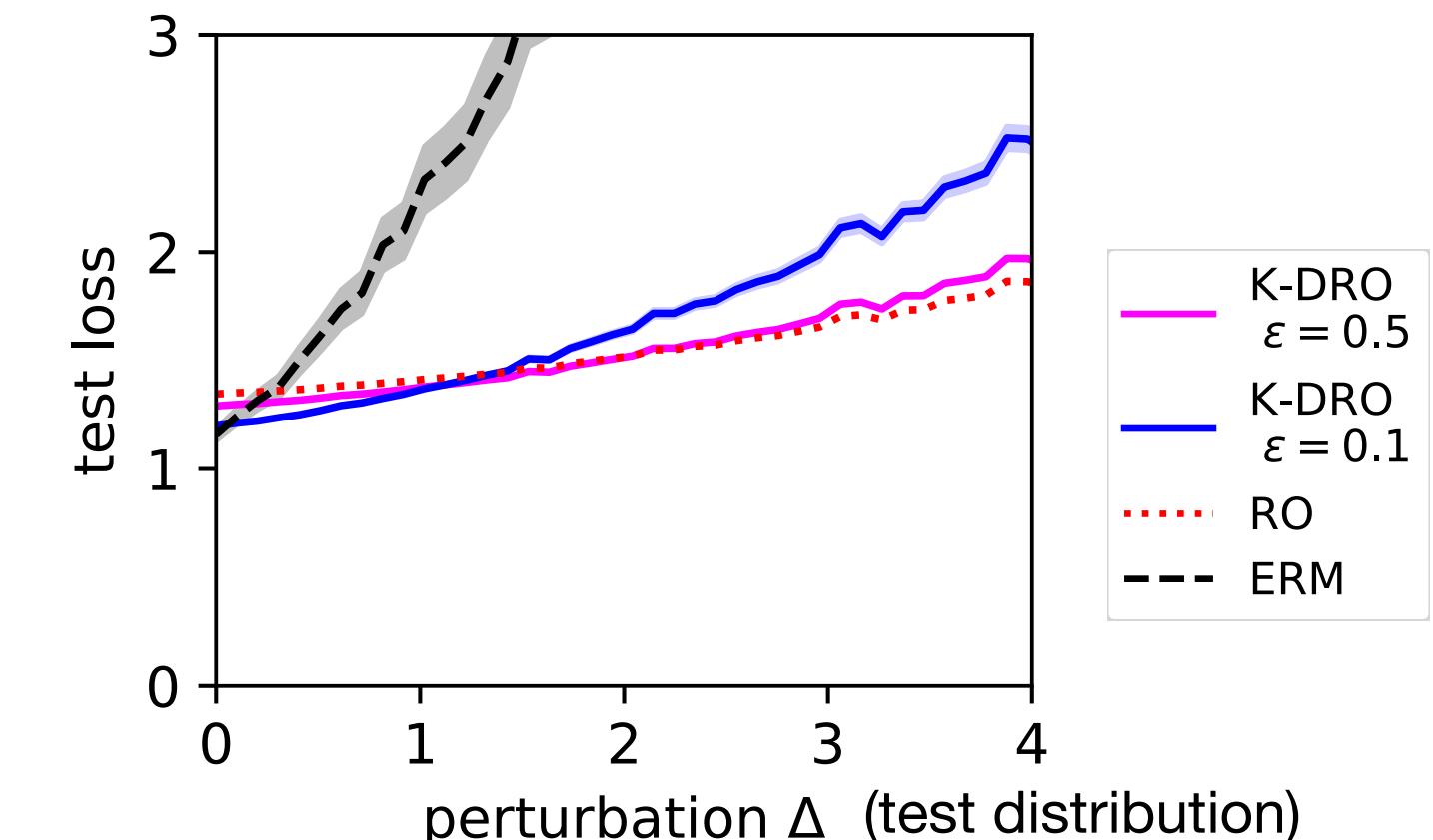


Example. Robust least squares

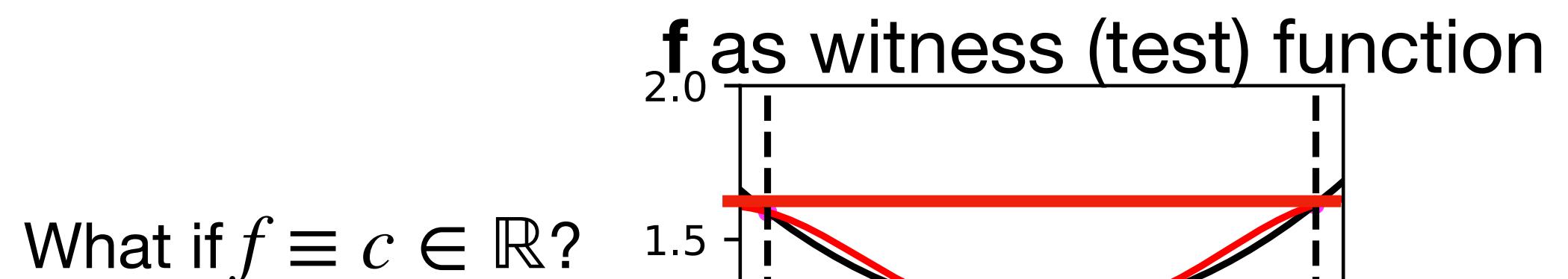
[El Ghaoui Lebret '97]

$$\text{minimize } l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$

Given historical samples $\xi_1, \xi_2, \dots, \xi_N$



Robustifying with DRO



What if $f \equiv c \in \mathbb{R}$?

Duality perspective

Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, \lambda > 0} \frac{1}{N} \sum_{i=1}^N l_{\theta}^{\lambda \|\cdot\|^2}(\xi_i) + \lambda \epsilon^2$$

where Moreau envelope

$$l_{\theta}^{\lambda \|\cdot\|^2}(x) := \sup_u l(\theta, u) - \lambda \|u - x\|^2$$

Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, \lambda > 0} \frac{1}{N} \sum_{i=1}^N l_{\theta}^{\lambda \|\cdot\|^2}(\xi_i) + \lambda \epsilon^2$$

where Moreau envelope

$$l_{\theta}^{\lambda \|\cdot\|^2}(x) := \sup_u l(\theta, u) - \lambda \|u - x\|^2$$

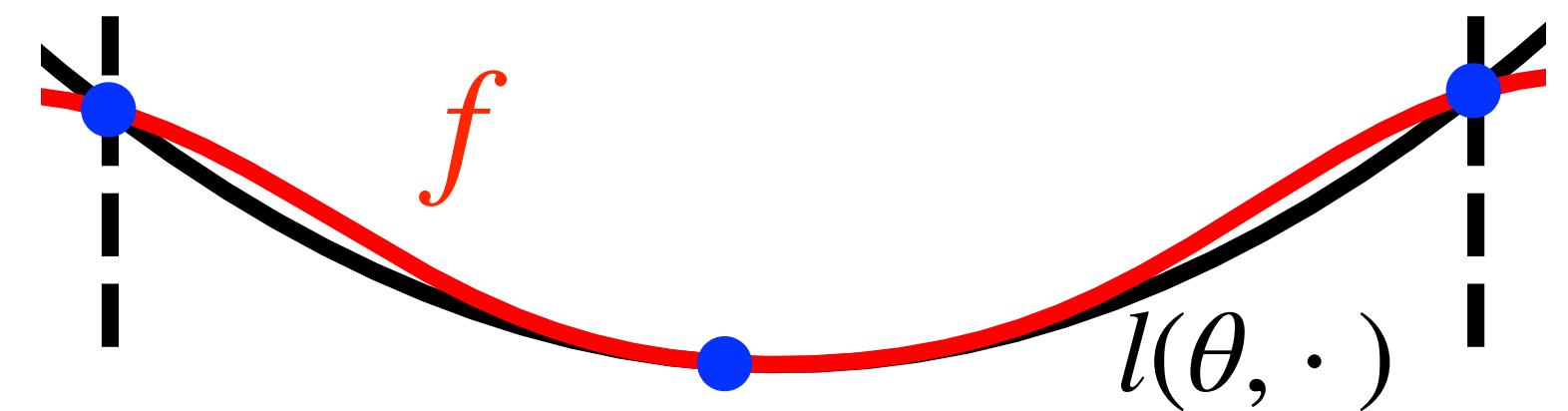
Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$$

s . t. $l(\theta, \xi) \leq f(\xi), \forall \xi$ a.e.



Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, \lambda > 0} \frac{1}{N} \sum_{i=1}^N l_{\theta}^{\lambda \|\cdot\|^2}(\xi_i) + \lambda \epsilon^2$$

where Moreau envelope

$$l_{\theta}^{\lambda \|\cdot\|^2}(x) := \sup_u l(\theta, u) - \lambda \|u - x\|^2$$

Considerations from WGF theory

- l is **nonlinear** (e.g., DNN)?
- Nonlinear (in measure) energies?

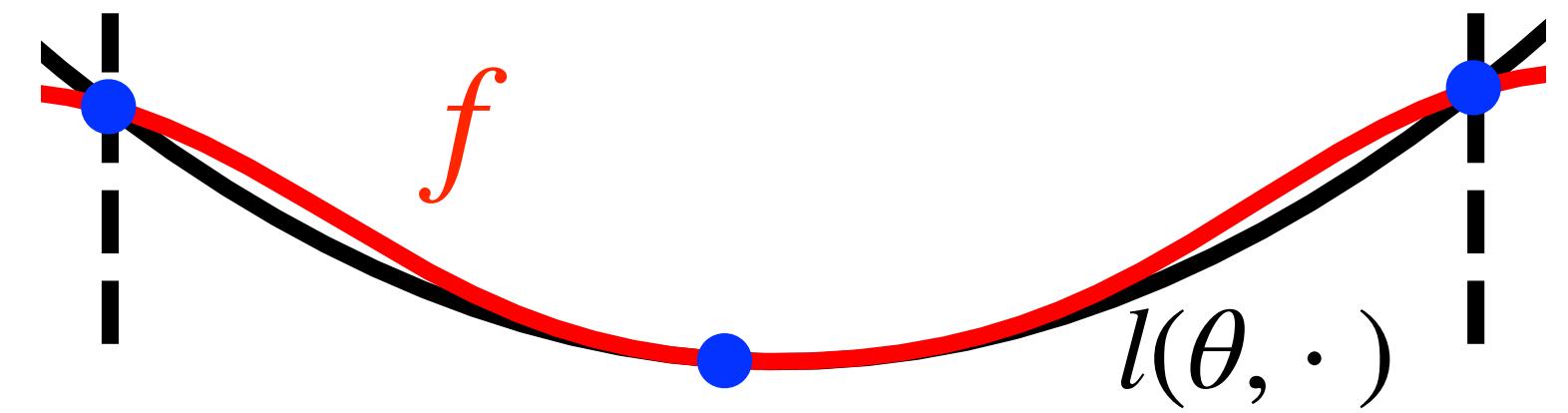
Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$$

s . t. $l(\theta, \xi) \leq f(\xi), \forall \xi$ a.e.



Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, \lambda > 0} \frac{1}{N} \sum_{i=1}^N l_{\theta}^{\lambda \|\cdot\|^2}(\xi_i) + \lambda \epsilon^2$$

where Moreau envelope

$$l_{\theta}^{\lambda \|\cdot\|^2}(x) := \sup_u l(\theta, u) - \lambda \|u - x\|^2$$

Considerations from WGF theory

- l is **nonlinear** (e.g., DNN)?
- Nonlinear (in measure) energies?

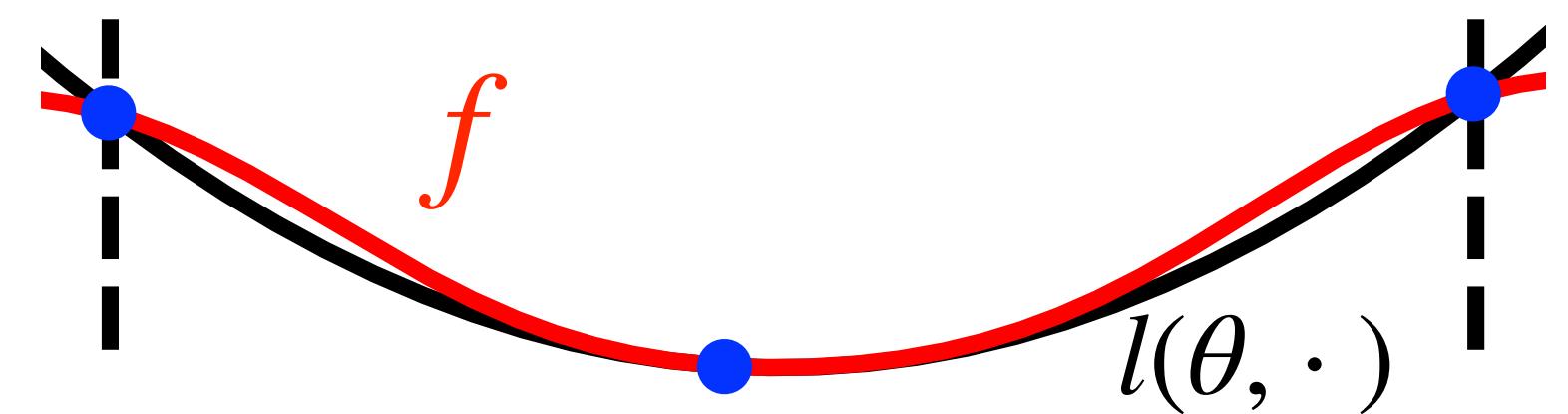
Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$$

s . t. $l(\theta, \xi) \leq f(\xi), \forall \xi$ a.e.



Duality perspective

2-Wasserstein

Kernel DRO [z. et al. 2021]

Primal:

$$\min_{\theta} \sup_{W_2(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, \lambda > 0} \frac{1}{N} \sum_{i=1}^N l_{\theta}^{\lambda \|\cdot\|^2}(\xi_i) + \lambda \epsilon^2$$

where Moreau envelope

$$l_{\theta}^{\lambda \|\cdot\|^2}(x) := \sup_u l(\theta, u) - \lambda \|u - x\|^2$$

Considerations from WGF theory

- l is **nonlinear** (e.g., DNN)?
- Nonlinear (in measure) energies?

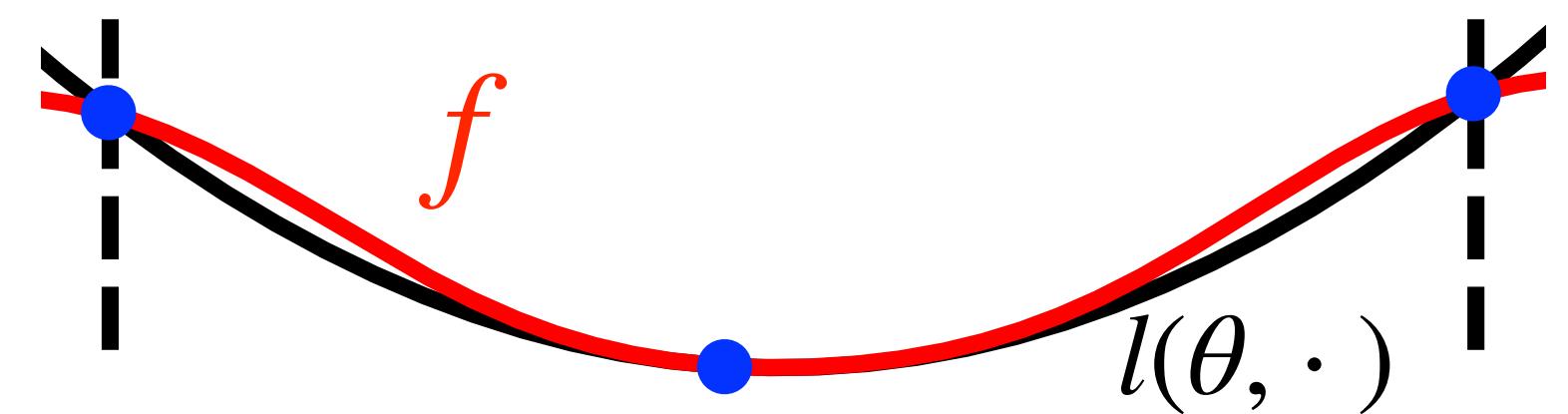
Primal:

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P l(\theta, \xi)$$

Dual:

$$\min_{\theta, f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N f(\xi_i) + \epsilon \|f\|_{\mathcal{H}}$$

s . t. $l(\theta, \xi) \leq f(\xi), \forall \xi$ a.e.



Nonlinear kernel approx. as robust surrogate losses (flatten the curve)

Duality of Gradient Flow Force-Balance

From static DRO to JKO scheme for gradient flows

DRO's Wasserstein measure optimization is not new.

$$\min_{\theta} \sup_{\substack{P \\ W_2(P, \hat{P}) \leq \epsilon}} \mathbb{E}_P I(\theta, \xi)$$

$$\min_{\theta} \sup_P \mathbb{E}_P I(\theta, \xi) - \gamma \cdot W_2^2(P, \hat{P})$$

From static DRO to JKO scheme for gradient flows

DRO's Wasserstein measure optimization is not new.

$$\min_{\theta} \sup_{\substack{P \\ W_2(P, \hat{P}) \leq \epsilon}} \mathbb{E}_P I(\theta, \xi)$$

$$\min_{\theta} \sup_P \mathbb{E}_P I(\theta, \xi) - \gamma \cdot W_2^2(P, \hat{P})$$

Wasserstein gradient flow [Otto et al. 90s-2000s]. The Fokker-Planck equation

$$\partial_t \mu + \nabla \cdot (\mu \nabla \frac{\delta F}{\delta \mu}[\mu]) = 0$$

is the gradient-flow equation of energy F in $(\text{Prob}(\bar{X}), W_2)$.

From static DRO to JKO scheme for gradient flows

DRO's Wasserstein measure optimization is not new.

$$\min_{\theta} \sup_{\substack{P \\ W_2(P, \hat{P}) \leq \epsilon}} \mathbb{E}_P I(\theta, \xi)$$

$$\min_{\theta} \sup_P \mathbb{E}_P I(\theta, \xi) - \gamma \cdot W_2^2(P, \hat{P})$$

Wasserstein gradient flow [Otto et al. 90s-2000s]. The Fokker-Planck equation

$$\partial_t \mu + \nabla \cdot (\mu \nabla \frac{\delta F}{\delta \mu}[\mu]) = 0$$

is the gradient-flow equation of energy F in $(\text{Prob}(\bar{X}), W_2)$.

Jordan-Kinderlehrer-Otto (JKO) scheme or Minimizing Movement Scheme (MMS):

$$\mu^{k+1} \in \inf_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu^k)$$

generalizes the DRO dual reformulation of DRO to **nonlinear-in-measure** F .

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

$-\nabla f(x(t)) \in X^*$ provides the **(thermodynamic) force** (can't see; shadow price)

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

$-\nabla f(x(t)) \in X^*$ provides the **(thermodynamic) force** (can't see; shadow price)

The equation should actually be written in the **force-balance** form

$$\mathbb{I}_R \dot{x}(t) = -\nabla f(x(t)) \in X^*, \quad \mathbb{I}_R : X \rightarrow X^* \text{ is the Riesz isomorphism.}$$

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

$-\nabla f(x(t)) \in X^*$ provides the **(thermodynamic) force** (can't see; shadow price)

The equation should actually be written in the **force-balance** form

$$\mathbb{I}_R \dot{x}(t) = -\nabla f(x(t)) \in X^*, \quad \mathbb{I}_R : X \rightarrow X^* \text{ is the Riesz isomorphism.}$$

Energy dissipation balance (equality) via **Fenchel(-Young) duality and optimality**

$$\frac{d}{dt} f(x(t)) = X^* \langle \nabla f(x(t)), \dot{x} \rangle_X = -\|\nabla f(x(t))\|^2 = -\left(\frac{1}{2}\|\dot{x}\|^2 + \frac{1}{2}\|\nabla f(x)\|^2\right)$$

Duality in gradient flow dynamics: ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

$-\nabla f(x(t)) \in X^*$ provides the **(thermodynamic) force** (can't see; shadow price)

The equation should actually be written in the **force-balance** form

$$\mathbb{I}_R \dot{x}(t) = -\nabla f(x(t)) \in X^*, \quad \mathbb{I}_R : X \rightarrow X^* \text{ is the Riesz isomorphism.}$$

Energy dissipation balance (equality) via **Fenchel(-Young) duality and optimality**

$$\frac{d}{dt} f(x(t)) = X^* \langle \nabla f(x(t)), \dot{x} \rangle_X = -\|\nabla f(x(t))\|^2 = -\left(\frac{1}{2}\|\dot{x}\|^2 + \frac{1}{2}\|\nabla f(x)\|^2\right)$$

Energy does not necessarily decrease along non-solutions, i.e., only inequality

$$\langle \nabla f(z(t)), \dot{z} \rangle_X \geq -\left(\frac{1}{2}\|\dot{z}\|^2 + \frac{1}{2}\|\nabla f(z(t))\|^2\right).$$

Duality in the Wasserstein gradient flow

In $(\text{Prob}(\bar{X}), F, W_2)$, **Fenchel(-Young) duality** yields the **Energy dissipation balance** (equality) [Ambrosio et al. 2007]

$$\frac{d}{dt} F(\mu(t)) = -\frac{1}{2} |\mu'|_{W_2}(t)^2 - \frac{1}{2} |\nabla^- F|_{W_2}(\mu(t))^2$$

Duality in the Wasserstein gradient flow

In $(\text{Prob}(\bar{X}), F, W_2)$, **Fenchel(-Young) duality** yields the **Energy dissipation balance** (equality) [Ambrosio et al. 2007]

$$\frac{d}{dt} F(\mu(t)) = -\frac{1}{2} |\mu'|_{W_2}(t)^2 - \frac{1}{2} |\nabla^- F|_{W_2}(\mu(t))^2$$

For (Boltzmann) entropy $F(u) = \rho \log \rho$, the *metric slope* is

$$|\nabla^- F|_{W_2}(\mu(t))^2 = \int |\nabla \log \rho|^2 \rho \, dx$$

Duality in the Wasserstein gradient flow

In $(\text{Prob}(\bar{X}), F, W_2)$, **Fenchel(-Young) duality** yields the **Energy dissipation balance** (equality) [Ambrosio et al. 2007]

$$\frac{d}{dt} F(\mu(t)) = -\frac{1}{2} |\mu'|_{W_2}(t)^2 - \frac{1}{2} |\nabla^- F|_{W_2}(\mu(t))^2$$

For (Boltzmann) entropy $F(u) = \rho \log \rho$, the *metric slope* is

$$|\nabla^- F|_{W_2}(\mu(t))^2 = \int |\nabla \log \rho|^2 \rho \, dx$$

However, for some **nonlinear (in measure) energy** (e.g., in variational inference)

$$F(\mu) = D_{\text{KL}}(\mu \| \pi), \frac{\delta F}{\delta \mu} [\mu] = \log \rho - \log \pi,$$

the density $\rho := \frac{d\mu}{d\mathcal{L}}$ and hence the force field $\frac{\delta F}{\delta \mu} [\mu]$ are **not accessible** for atomic distributions μ .

Kernel gradient flow as dual space force-balance

Motivated by the “Kernel DRO-type” derivation in [Zhu et al.’21, Kremer et al.’23],

Kernel gradient flow as dual space force-balance

Motivated by the “Kernel DRO-type” derivation in [Zhu et al.’21, Kremer et al.’23],

Proposition(informal). The gradient flow equation for $(\mathcal{P}(\bar{X}), F, \text{MMD})$ is given by the **dual space (force-balance) kernel gradient flow**

$$k * \dot{\mu} = -g, \quad \text{where } \nabla g = \nabla \frac{\delta F}{\delta \mu} [\mu] \quad \mu\text{-a.e.}$$

Notation of convolution $k * \mu := \int k(x, \cdot) \mu(dx)$.

Kernel gradient flow as dual space force-balance

Motivated by the “Kernel DRO-type” derivation in [Zhu et al.’21, Kremer et al.’23],

Proposition(informal). The gradient flow equation for $(\mathcal{P}(\bar{X}), F, \text{MMD})$ is given by the **dual space (force-balance) kernel gradient flow**

$$k * \dot{\mu} = -g, \quad \text{where } \nabla g = \nabla \frac{\delta F}{\delta \mu} [\mu] \quad \mu\text{-a.e.}$$

Notation of convolution $k * \mu := \int k(x, \cdot) \mu(dx)$. ∇g “matches the score” .

Kernel gradient flow as dual space force-balance

Motivated by the “Kernel DRO-type” derivation in [Zhu et al.’21, Kremer et al.’23],

Proposition(informal). The gradient flow equation for $(\mathcal{P}(\bar{X}), F, \text{MMD})$ is given by the **dual space (force-balance) kernel gradient flow**

$$k * \dot{\mu} = -g, \quad \text{where } \nabla g = \nabla \frac{\delta F}{\delta \mu} [\mu] \quad \mu\text{-a.e.}$$

Notation of convolution $k * \mu := \int k(x, \cdot) \mu(dx)$. ∇g “matches the score” .

Compared with the Wasserstein GF of entropy, our kernel scheme approximates the (unavailable) “score function” $\nabla g = \nabla \log \rho$ in a principled geometry. This gives the interpretation of the dual kernel function in dynamics

g is the approximate (thermodynamic) force field.

Back to (kernel) robust learning

Motivated by our insight so far, in “Kernel DRO” [Zhu et al. 2021]

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P I(\theta, \xi),$$

Back to (kernel) robust learning

Motivated by our insight so far, in “Kernel DRO” [Zhu et al. 2021]

$$\min_{\theta} \sup_{\substack{\mathbb{E}_P I(\theta, \xi), \\ \text{MMD}(P, \hat{P}) \leq \epsilon}} \quad$$

the distribution shift (a.k.a. adversarial attack) is modeled by the dynamical system of the dual force-balance kernel gradient flow

$$k * \dot{\mu} = -g, \quad \mu(0) = \hat{P}, \mu(1) = P.$$

Back to (kernel) robust learning

Motivated by our insight so far, in “Kernel DRO” [Zhu et al. 2021]

$$\min_{\theta} \sup_{\text{MMD}(P, \hat{P}) \leq \epsilon} \mathbb{E}_P I(\theta, \xi),$$

the distribution shift (a.k.a. adversarial attack) is modeled by the dynamical system of the dual force-balance kernel gradient flow

$$k * \dot{\mu} = -g, \quad \mu(0) = \hat{P}, \mu(1) = P.$$

On-going work: a **general-purpose measure optimization algorithm** motivated by this (see also the preprint on kernel mirror prox. [Dvurechensky & Zhu]).

Summary

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions or nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions or nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

- Role of the **dual kernel function** in this talk

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions or nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

- Role of the **dual kernel function** in this talk
 - **robust surrogate loss**

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions** or **nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

- Role of the **dual kernel function** in this talk
 - **robust surrogate loss**
 - **optimal test fcn. for two-sample test**

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions or nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

- Role of the **dual kernel function** in this talk
 - **robust surrogate loss**
 - **optimal test fcn. for two-sample test**
 - **approx. force field**

Summary

- **Consequence for robust learning/DRO:** our kernel scheme is designed to treat
 - Energy that's the **expectation of nonlinear functions or nonlinear in measures**

$$F(\mu) = \int V \, d\mu, \quad F(\mu) = \int \phi(\rho) \quad (\mu = \rho \cdot \mathcal{L})$$

which are challenging for computation using the Wasserstein GF (complication due to W-geodesics).

- Role of the **dual kernel function** in this talk
 - **robust surrogate loss**
 - **optimal test fcn. for two-sample test**
 - **approx. force field**
- Other important uses of the dual kernel function: Causal inference, IV, conditional moments, (Robust) control and RL

This talk is based on previous and on-going joint works with many co-authors.
Special thanks to Alexander Mielke for insightful discussions.

Z., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. AISTATS 2021

Kremer, H., Nemmour, Y., Schölkopf, B. & **Z.**. Estimation Beyond Data Reweighting: Kernel Method of Moments. ICML 2023

P. Dvurechensky, **Z.**, Kernel Mirror Prox and RKHS Gradient Flow for Mixed Functional Nash Equilibrium. WIAS Preprint

Z., Unbalanced Kernel Transport Problem. Working paper

Download the slides:



Website: jj-zhu.github.io