Distributionally Robust Learning and Optimization in MMD Geometry

J.J. (Jia-Jie) Zhu

jj-zhu.github.io

Weierstraß-Institut für
Angewandte Analysis und Stochastik

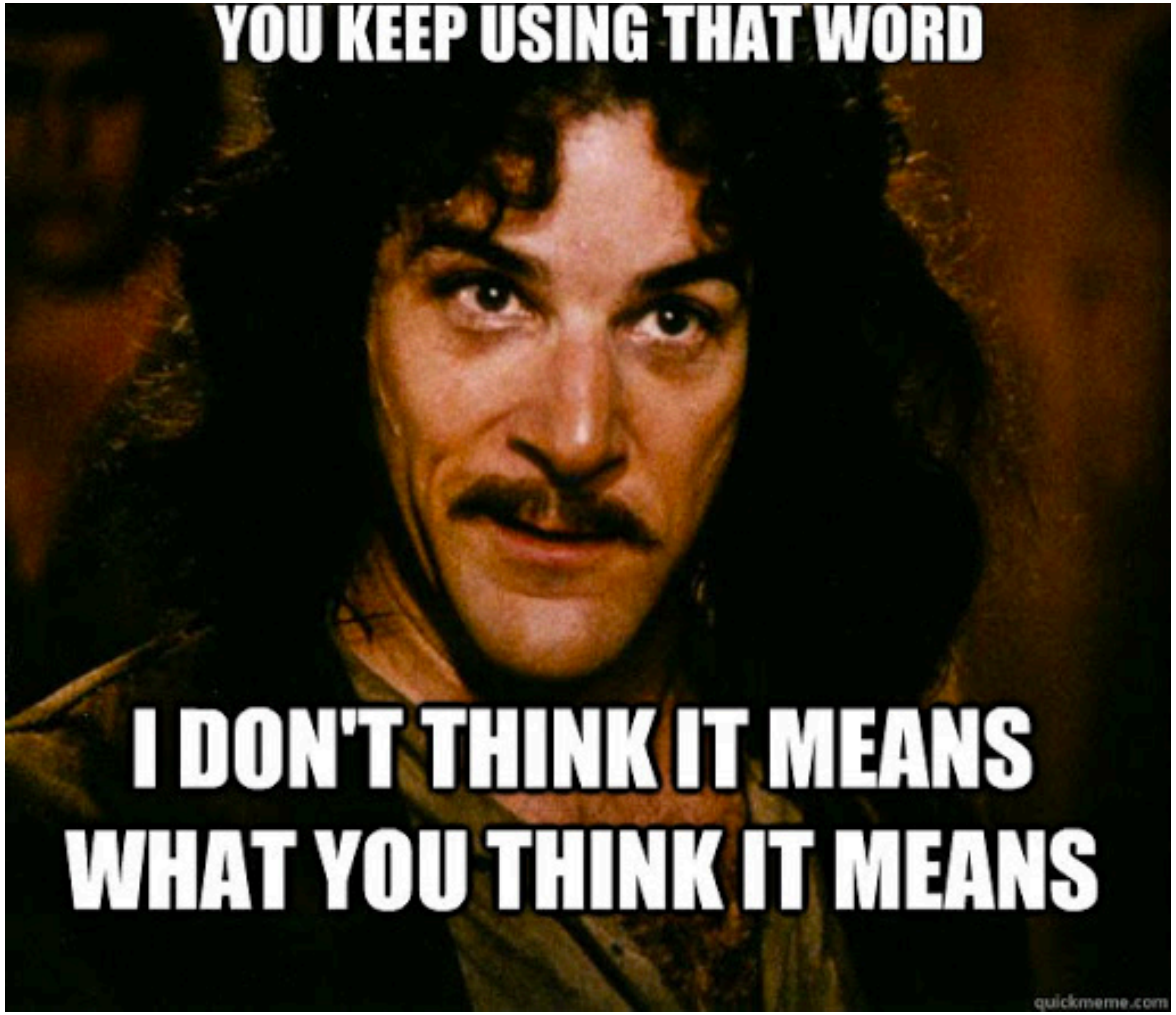
Weierstrass Institute for Applied Analysis and Stochastics, Berlin

Workshop on Optimal Transport, Statistics, Machine Learning
September 8th, 2022
TU Eindhoven

Distributional Robustness

Distributional Robustness

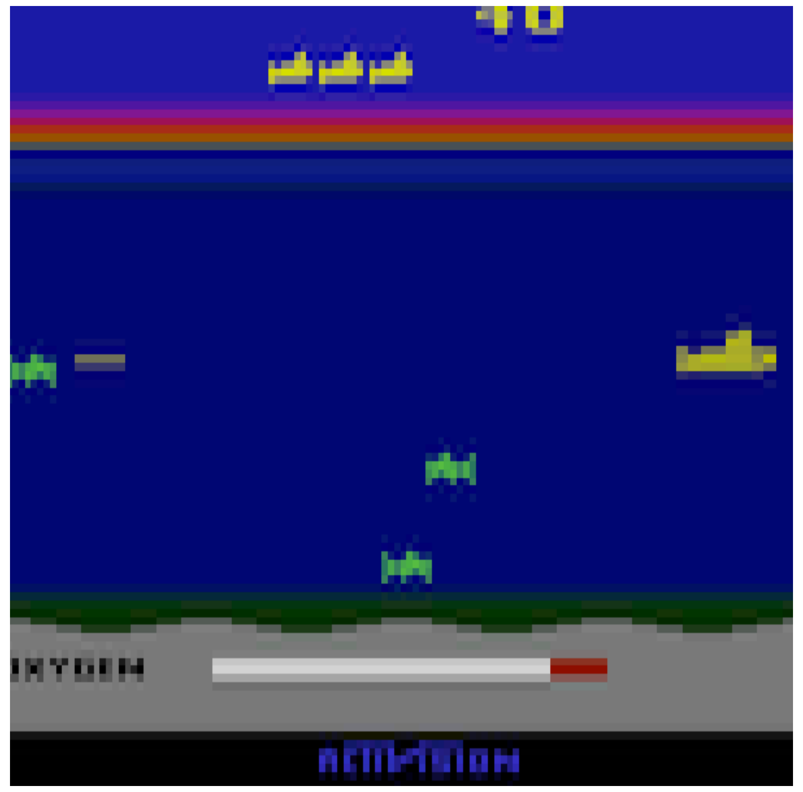
What is robustness?



- Many fields: …robust statistics, robust control, robust optimization, adversarial robustness, robust learning…

Robustness in Modern Machine Learning and Optimization

Modern machine learning



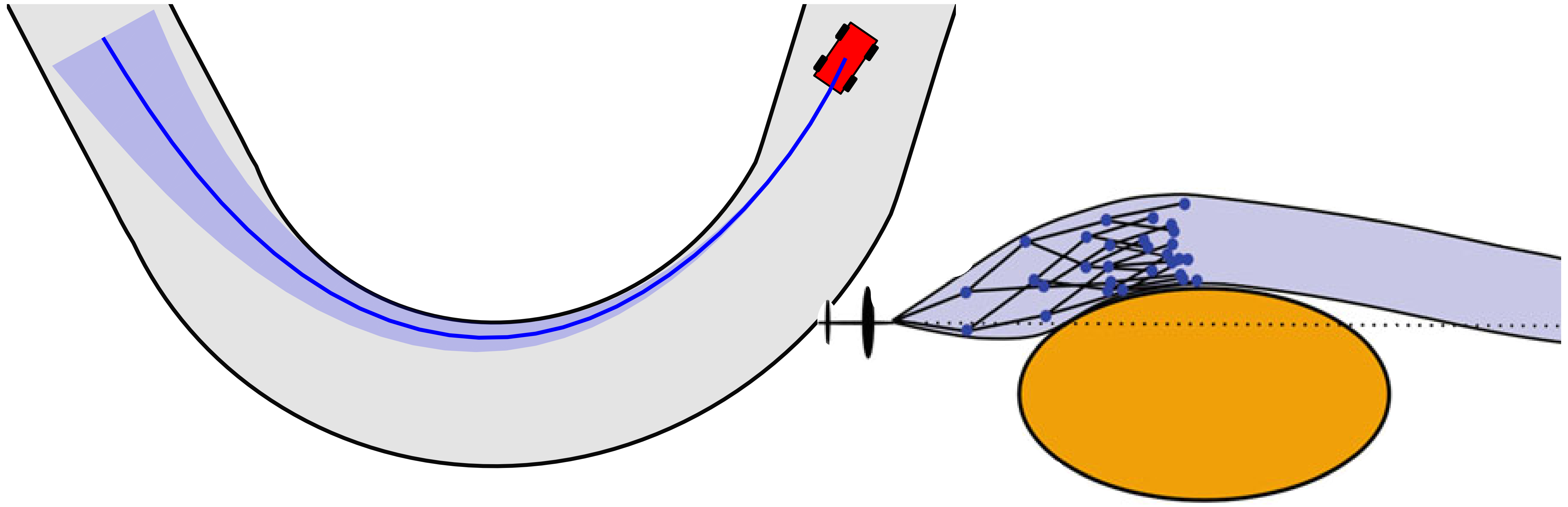
loss/cost → l; model → f_θ; data (random) → X, Y

$$\min_{\theta} \mathbb{E}_{[X,Y]\sim\hat{P}} \, l(f_\theta(X), Y)$$

Empirical dist. $\hat{P} = \sum_{i=1}^{N} \frac{1}{N}\delta_{\xi_i}$

- Do well on average
- Strength: high-performance (optimal)
- Weakness: fragile — adversarial attacks, off-policy RL, bias, fairness, causality

Robust optimization & control



$$\min_{\theta} \sup_{\xi\in\mathcal{U}} l(\theta, \xi)$$

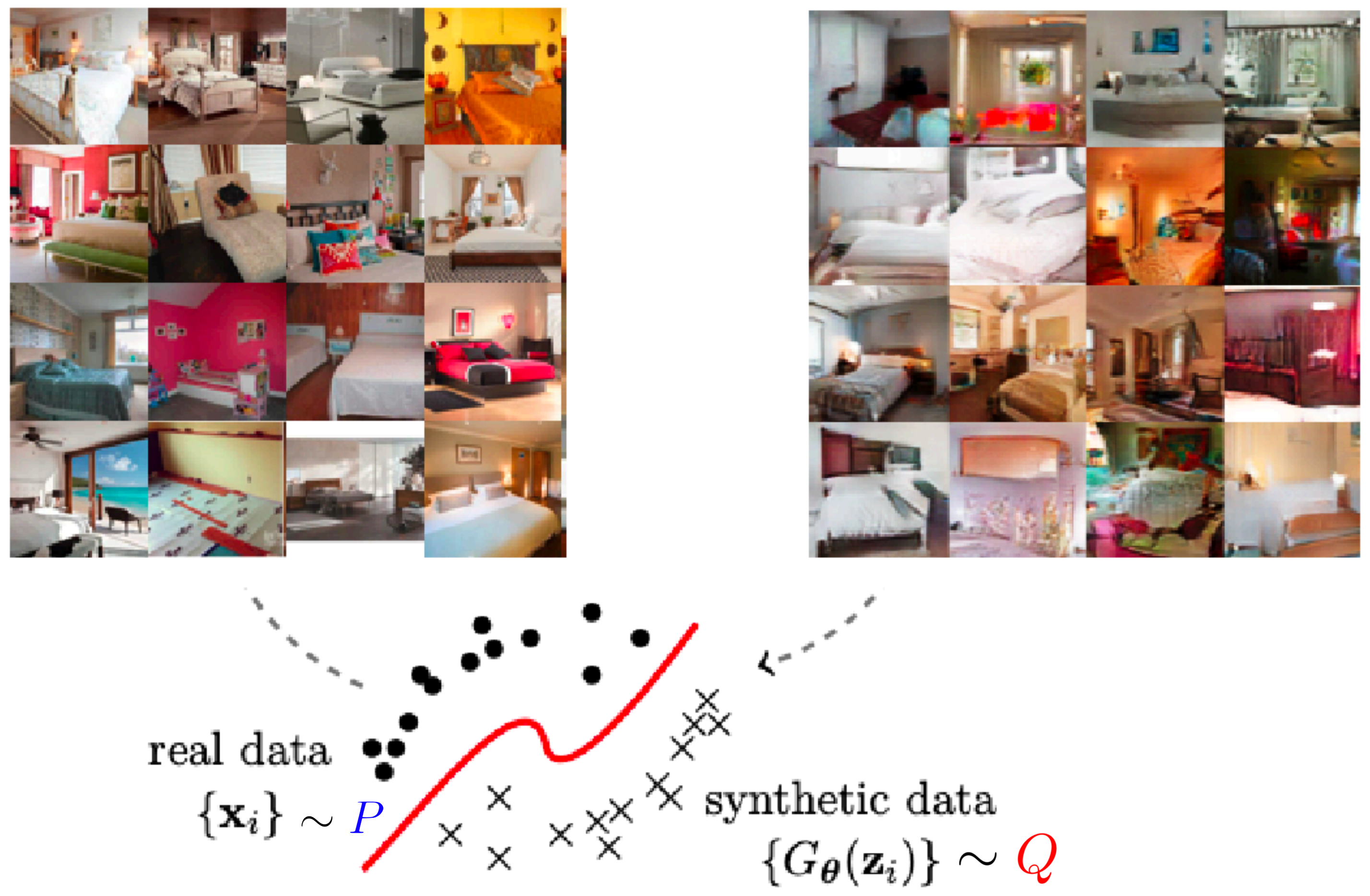
- Do well in the worst case
- Strength: robustness
- Weakness: conservative — worst case doesn't often happen

Image credit: Mnih'13, MuJuCo, Houska and Villanueva '19, Hewing et al.'18

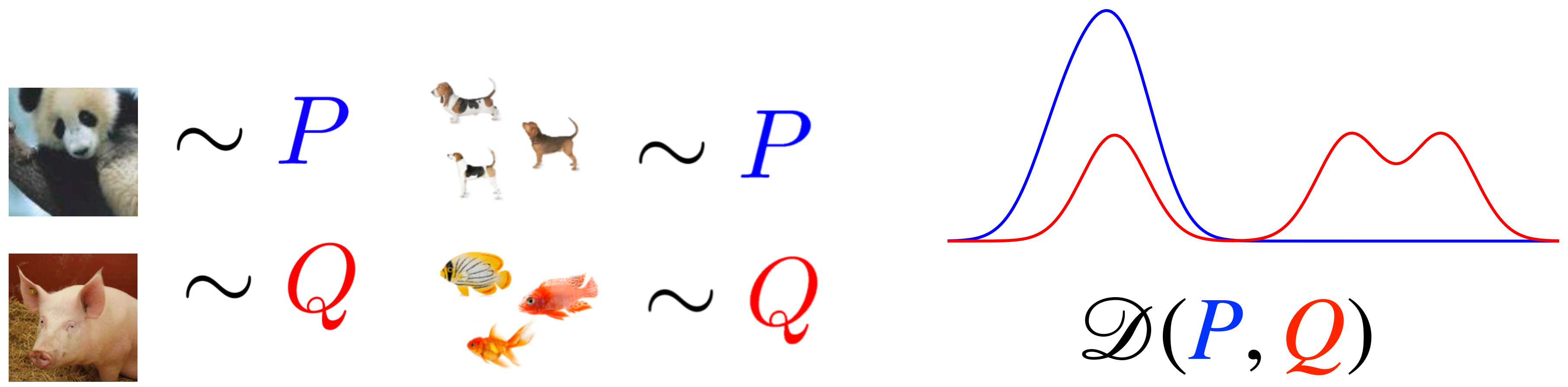
Distributional Robustness

Distribution Shift in Robust Machine Learning

Example. Generative modeling



We train a learning model to minimize the *distance between two (high-dimensional) data distributions* using kernel methods and optimal transport



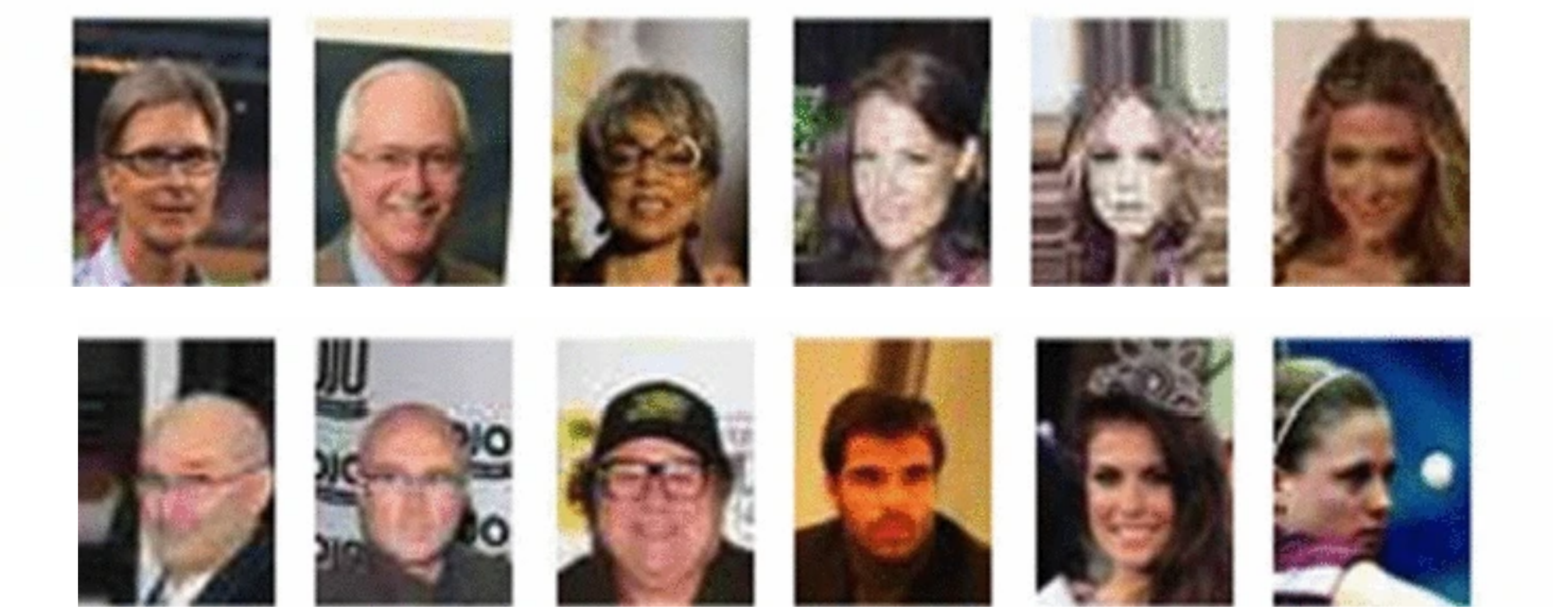
Example. Distributionally robust machine learning

Classify the presence of eyewear under adversarial attacks (cf. references)

$\hat{P}_{\text{train}} \neq Q_{\text{test}}$

Distribution shifts (slight) can break the system!

$\hat{P}_{\text{train}} \neq Q_{\text{test}}$



Learning with kernels and RKHSs

- A kernel is a symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.
- A p.d. k corresponds to a Hilbert space $\mathcal{H}$ (RKHS), which satisfies the **reproducing property** $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$, $\phi(x) := k(x, \cdot)$ is the **canonical feature** of $\mathcal{H}$.
- If $\mathcal{H}$ is a large *(dense in C_0 and $L_p(\mu)$, μ is a finite measure on $\mathbb{R}^d$)*, $\gamma_{\mathcal{H}}$ is a metric on $\mathcal{P}$. [Steinwart & Christmann 2008]
- Generalization to **integral probability metric** (IPM)

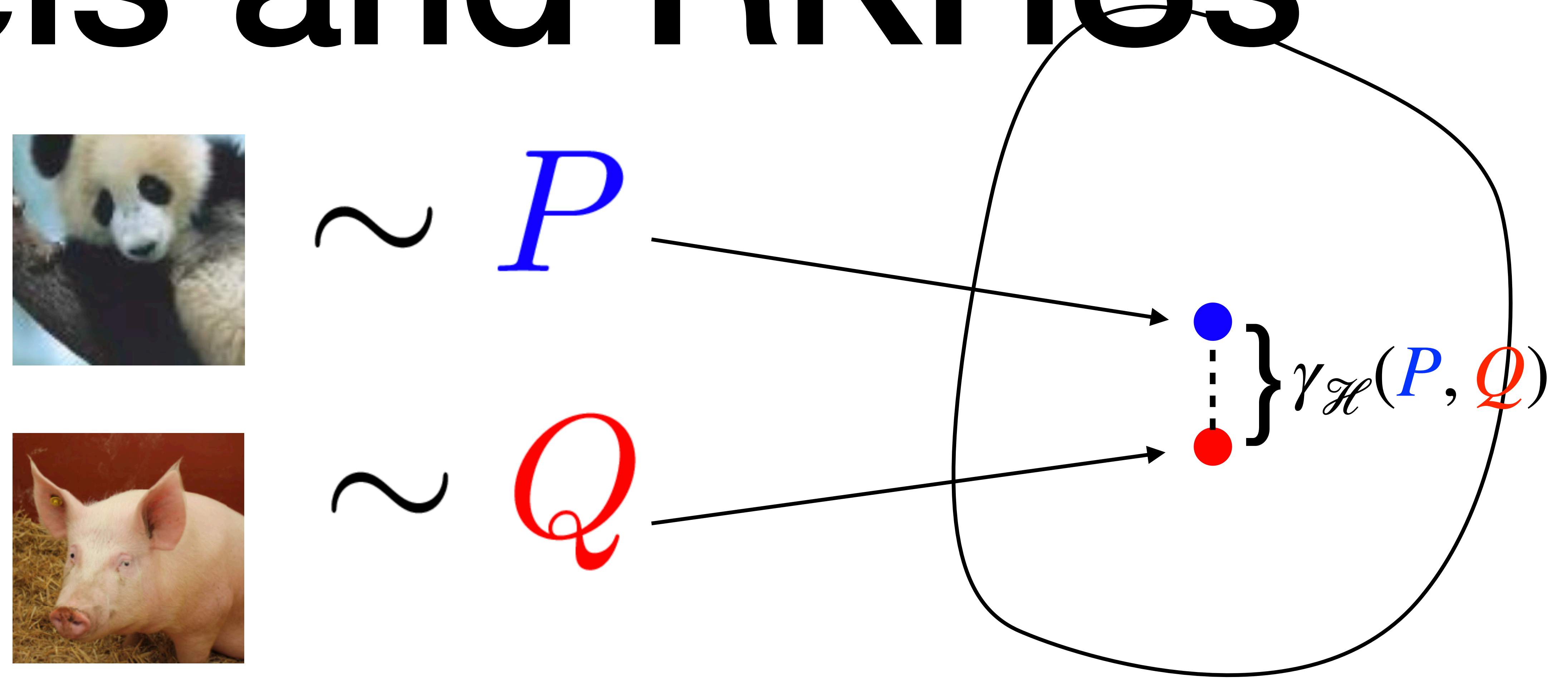
$$\text{IPM}(\mathcal{F}; P, Q) := \sup_{f \in \mathcal{F}} \int f \, d(P - Q).$$

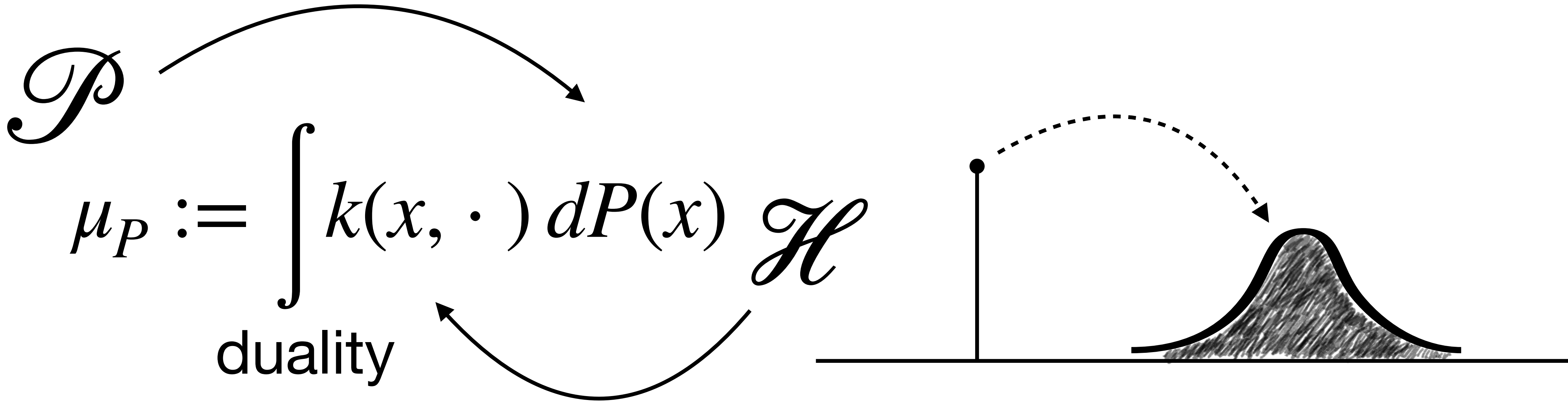
Special cases:

$\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$—> **Maximum Mean Discrepancy (MMD)**

$$\text{MMD}_{\mathcal{H}}(Q, P) := \sup_{\|f\|_{\mathcal{H}} \leq 1} \int f \, d(Q - P)$$
$$= \mathbb{E}_{x,x' \sim Q} k(x, x') + \mathbb{E}_{y,y' \sim P} k(y, y') - 2\mathbb{E}_{x \sim Q, y \sim P} k(x, y).$$

$\mathcal{F} = \{f : \|f\|_{\text{lip}} \leq 1\}$—> Wasserstein (type-1)





$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

μ can be viewed as a generalized moment vector
e.g., let $\phi(x) = [x, x^2]^\top$ (related: Lasserre moment-SOS)

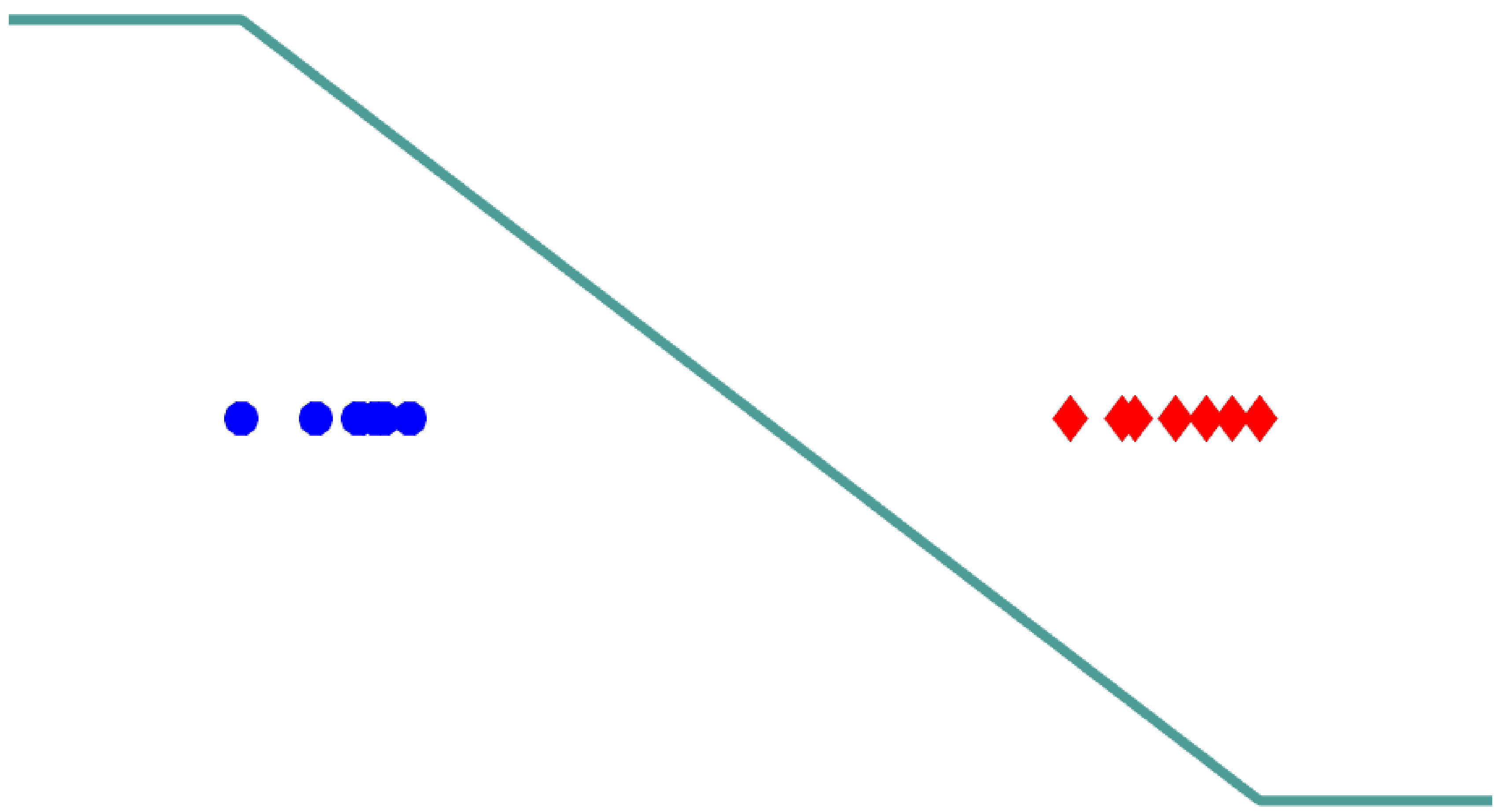
Illustration design inspired by Gretton, Sutherland, Jitkrittum NeurIPS 2019 tutorial

Duality: 1-Wasserstein vs. MMD-k

$$W_1(P, Q) = \sup_{||f||_L \le 1} E_P f(X) - E_Q f(Y).$$

$$||f||_L := \sup_{x \ne y} |f(x) - f(y)| \,/\, ||x - y||$$

W_1=0.88



$$MMD(P, Q) = \sup_{||f||_{\mathcal{F}} \le 1} E_P f(X) - E_Q f(Y).$$

MMD=1.8

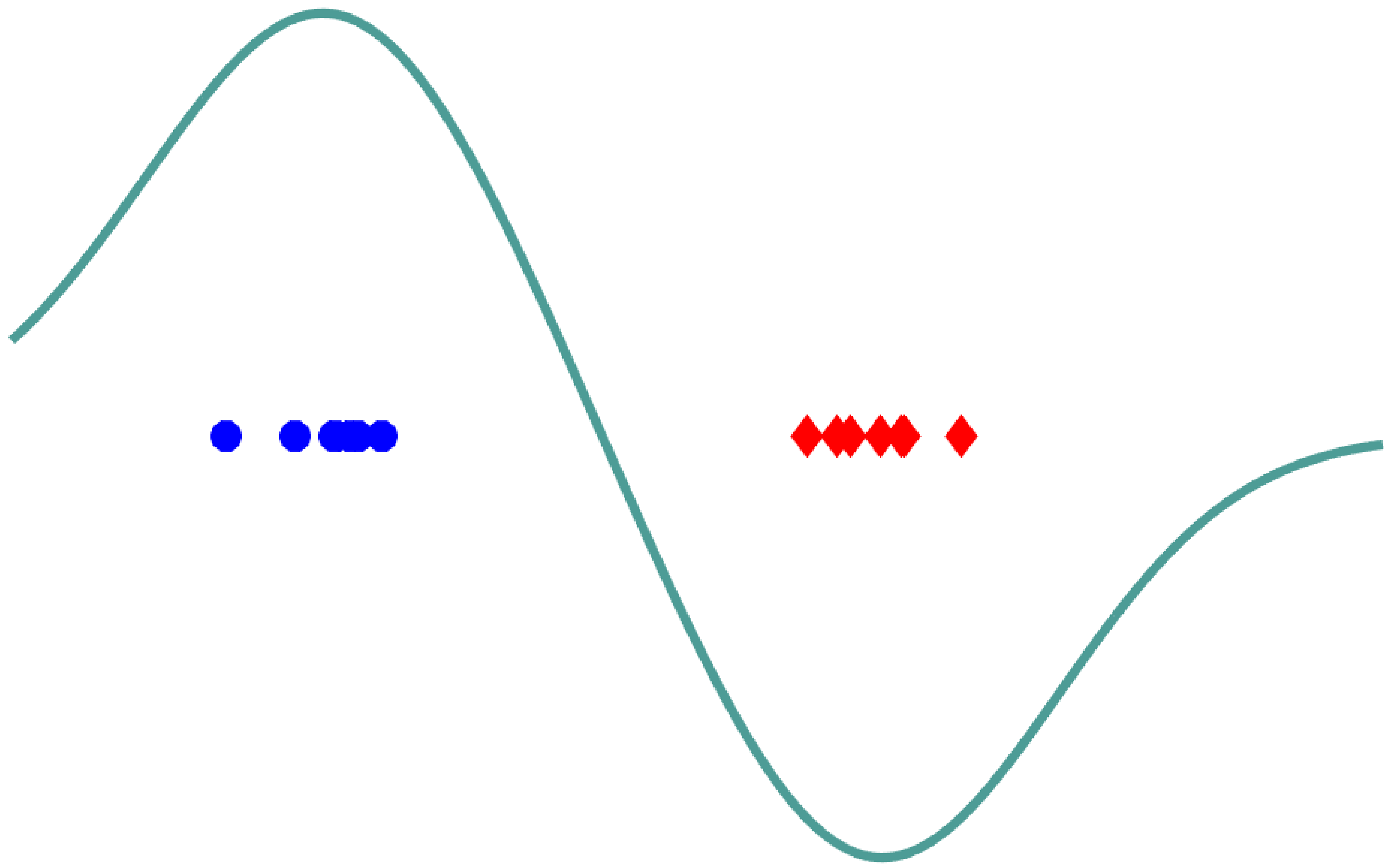


Illustration design inspired by Gretton, Sutherland, Jitkrittum NeurIPS 2019 tutorial

Distributional Robustness

Combine the strengths of ERM and RO: distributionally robust optimization (DRO)

(ERM) $\min_{\theta} \mathbb{E}_{\xi \sim \hat{P}} l(\theta, \xi)$

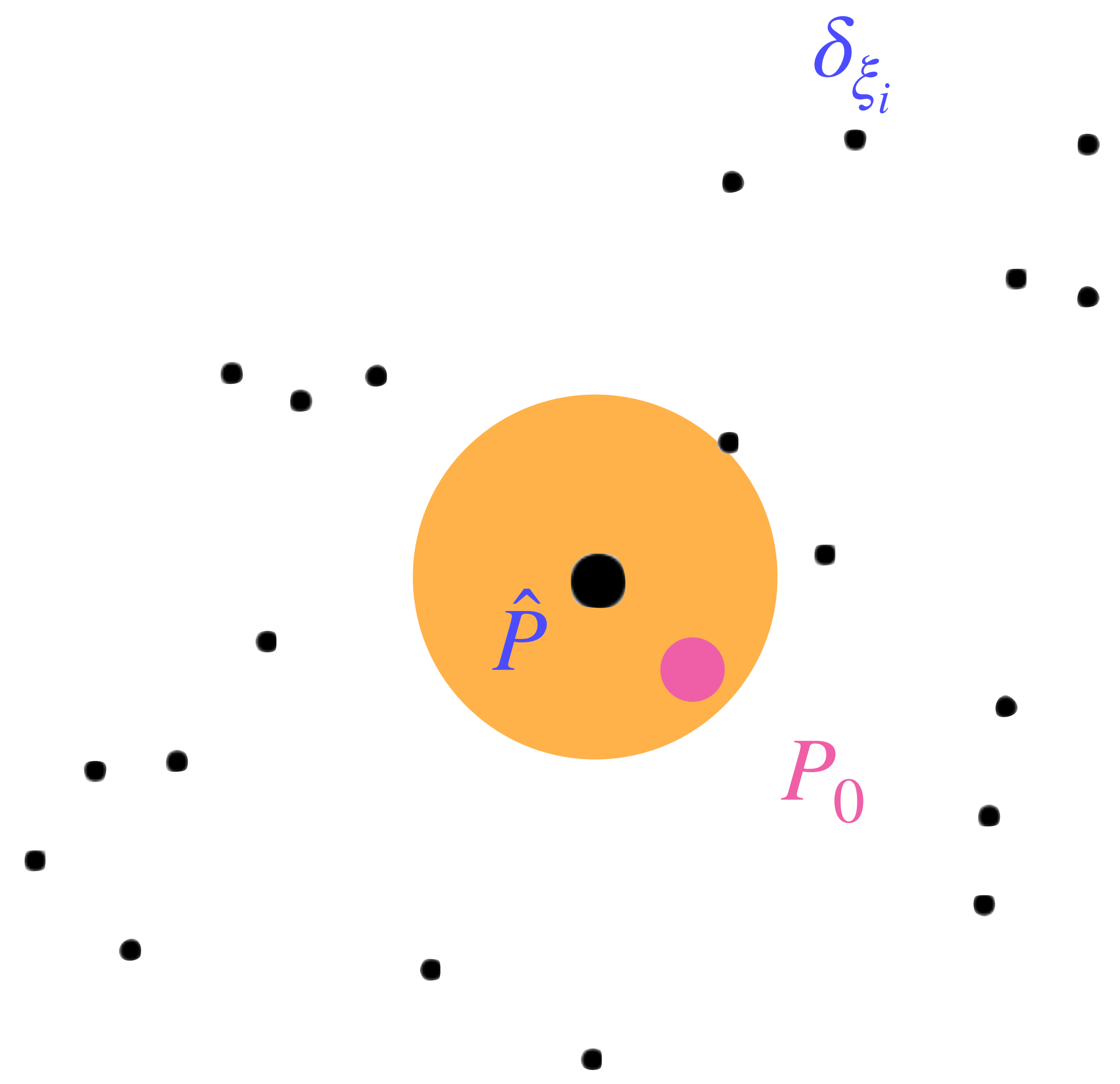
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_{Q} L(\theta, \xi)$$ (DRO)

[Delage and Ye 2010, Scarf 1958]

Find the worst-case distribution!
Problem of Moments [Stieltjes, Hausdorff, Hamburger, …]

- Robustifies against a set of probability measures $\mathcal{M}$ (***ambiguity set***), e.g.,
  - $\mathcal{M}$ can be a metric-ball centered at $\hat{P}$, e.g., using f-**divergences**, **optimal transport**, and **kernel methods**.
    - One way of constructing ambiguity region: one can quantify the empirical convergence rate $D(\hat{P}, P_0) \leq \epsilon$.



Robust learning under distribution shift

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

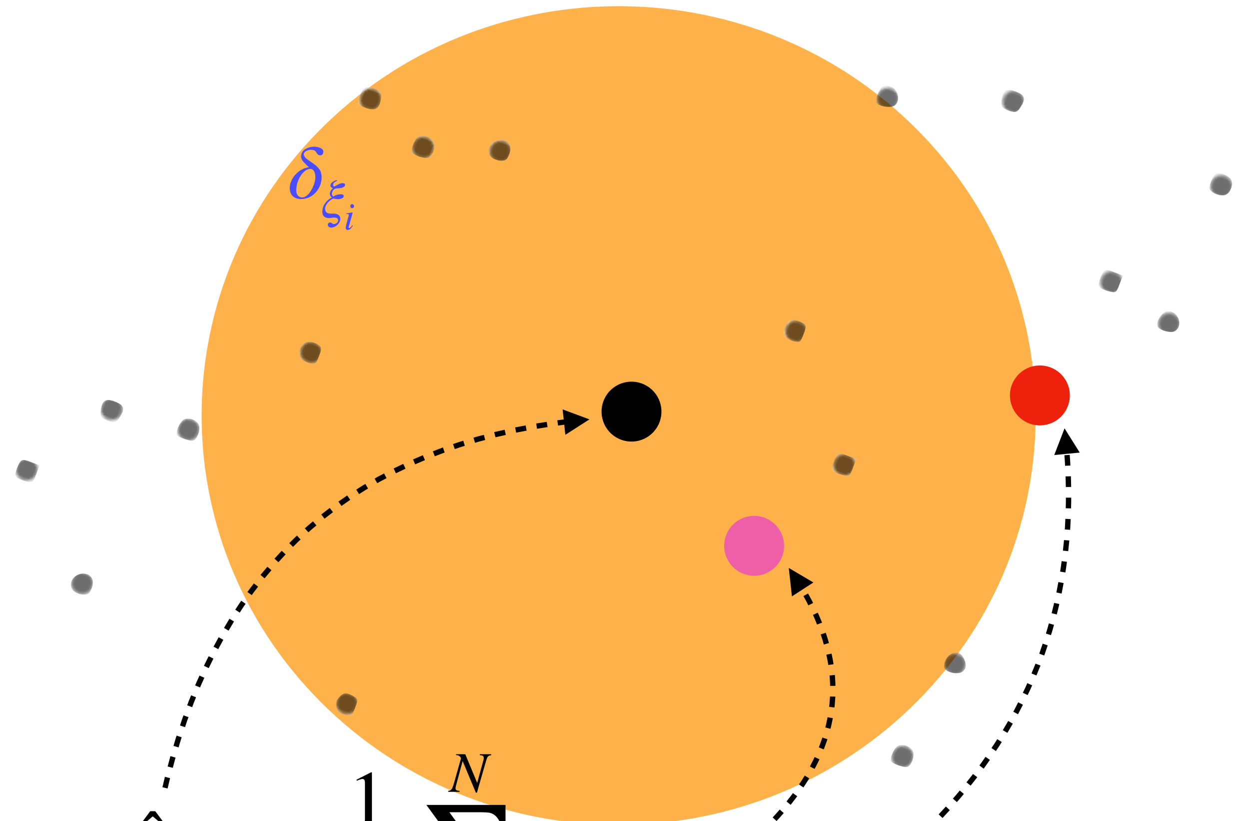
$$\mathbb{E}_{P_0} l(\hat{\theta}, \xi) \leq \frac{1}{N}\sum_{i=1}^{N} l(\hat{\theta}, \xi_i) + \mathcal{O}(\frac{1}{\sqrt{N}})$$

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

Distributionally Robust Learning

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

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathcal{M}$. Example: maximum mean discrepancy-ball $\{Q : \text{MMD}(Q, \hat{P}_N) \leq \rho\}$ or Wasserstein-ball
- **Question**: how do we actually solve an **MMD-constrained optimization problem**? (Non-trivial!)



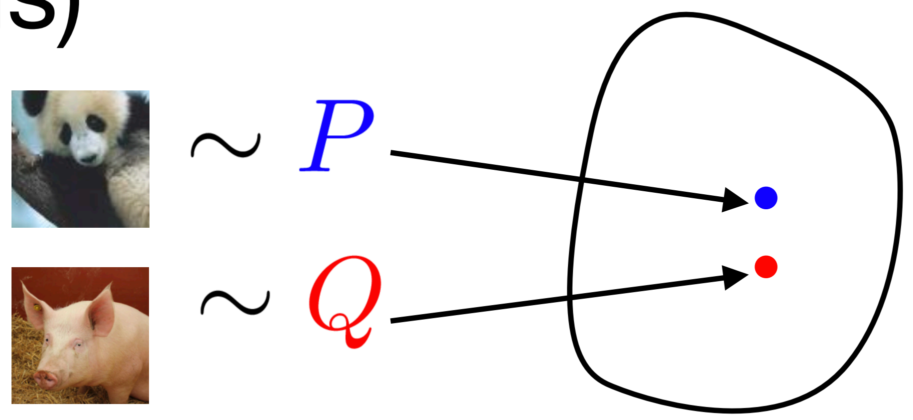
$$\hat{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

Distributional Robustness

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

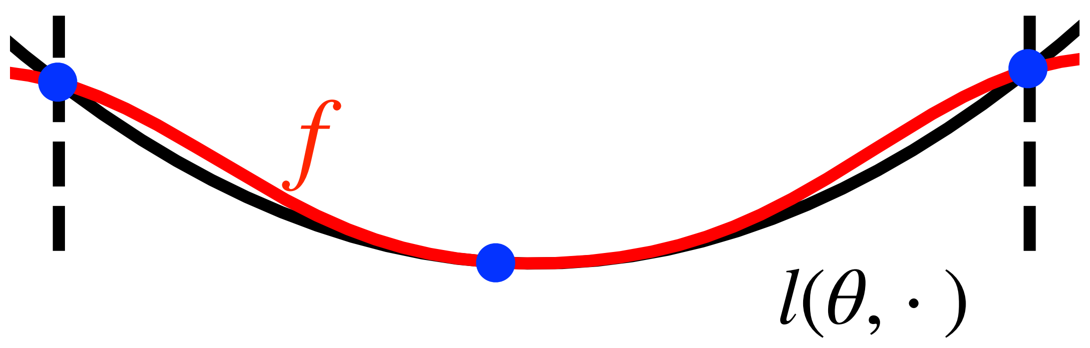


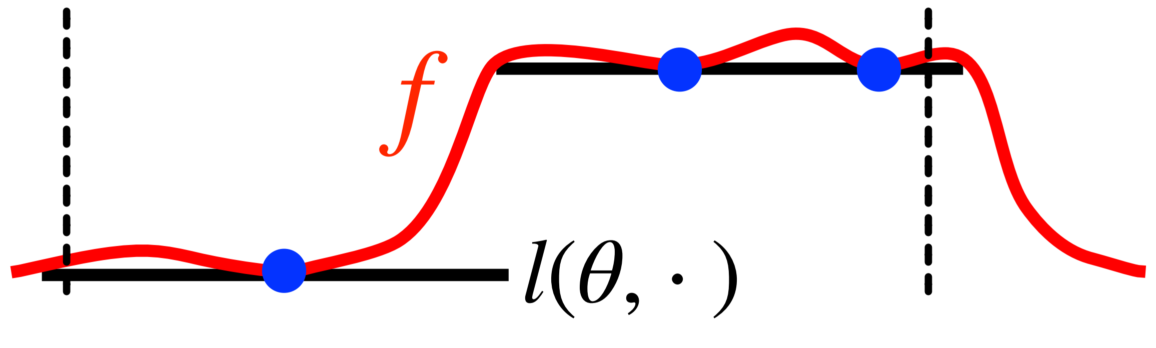
Kernel DRO Theorem (simplified). [Z. et al. 2021]
DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

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

cf. Kantorovich duality in optimal transport (OT) and Moreau-Yosida regularization in convex analysis

Geometric intuition: using kernel approximations as robust surrogate losses (flatten the curve)

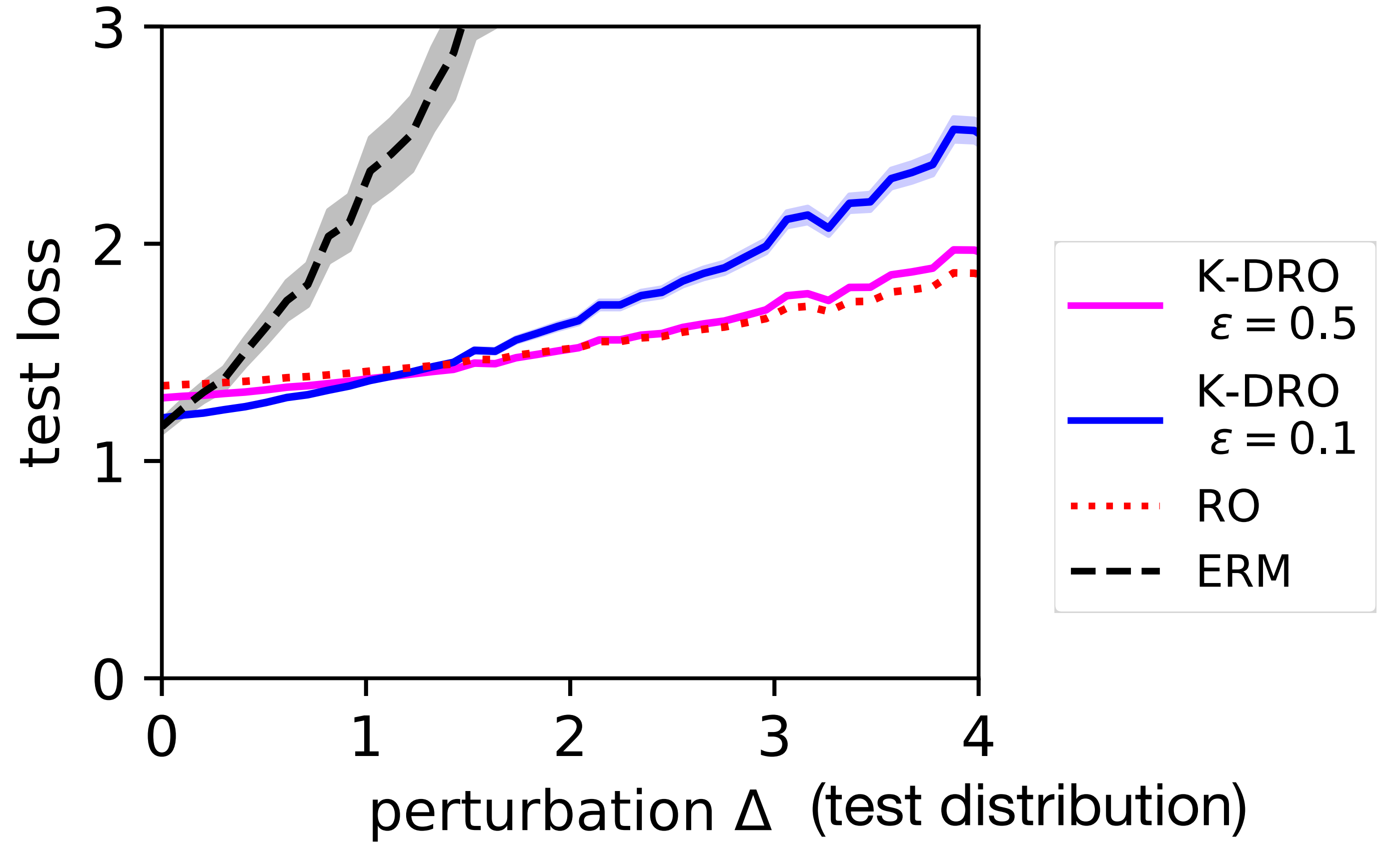




Example. Robust least squares
[El Ghaoui Lebret '97]

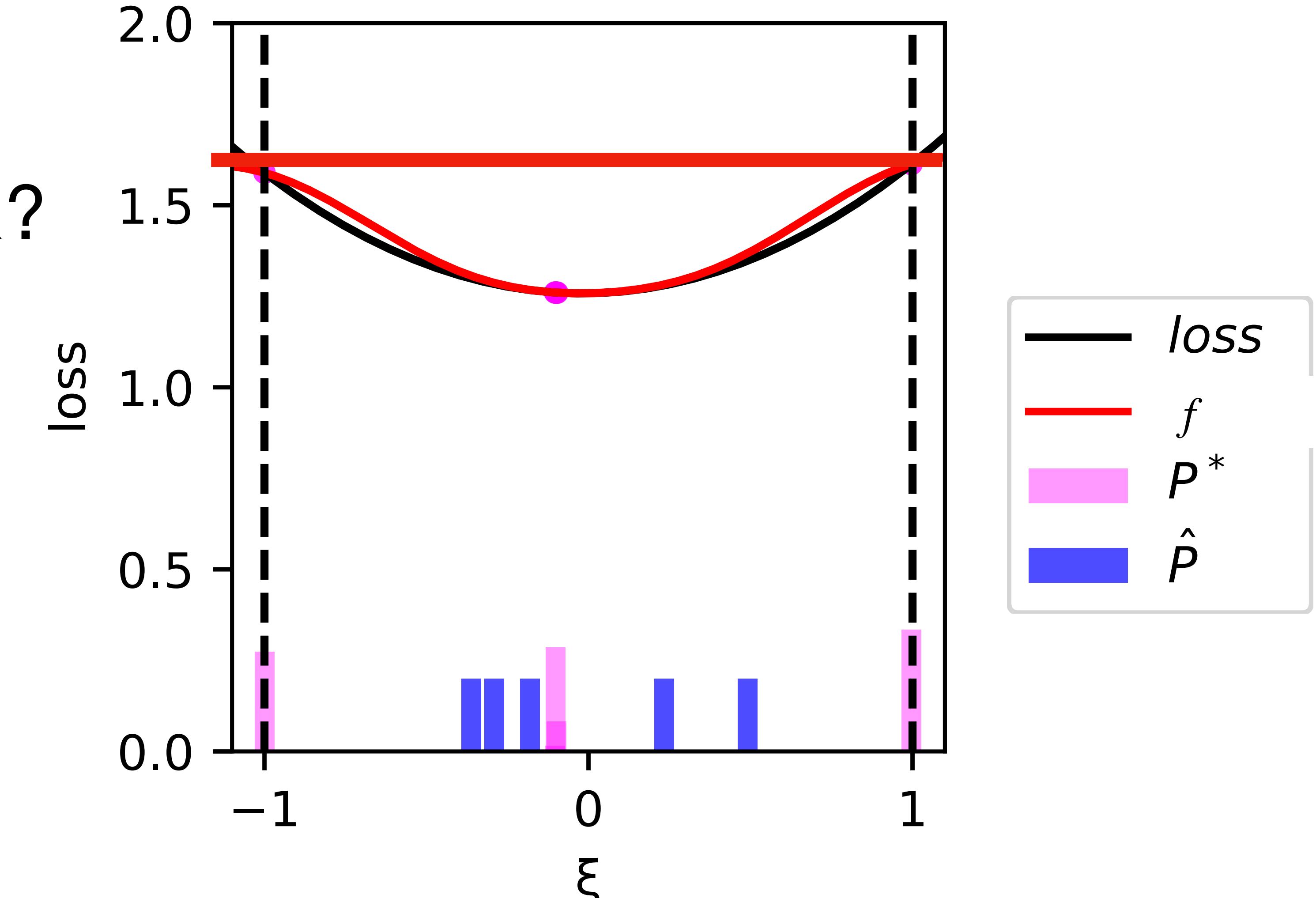
$$\text{minimize } l(\theta,\xi) := \|A(\xi)\cdot\theta - b\|_2^2$$

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



Robustifying with DRO

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



Comparing the "potentials"

2-Wasserstein DRO

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

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

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

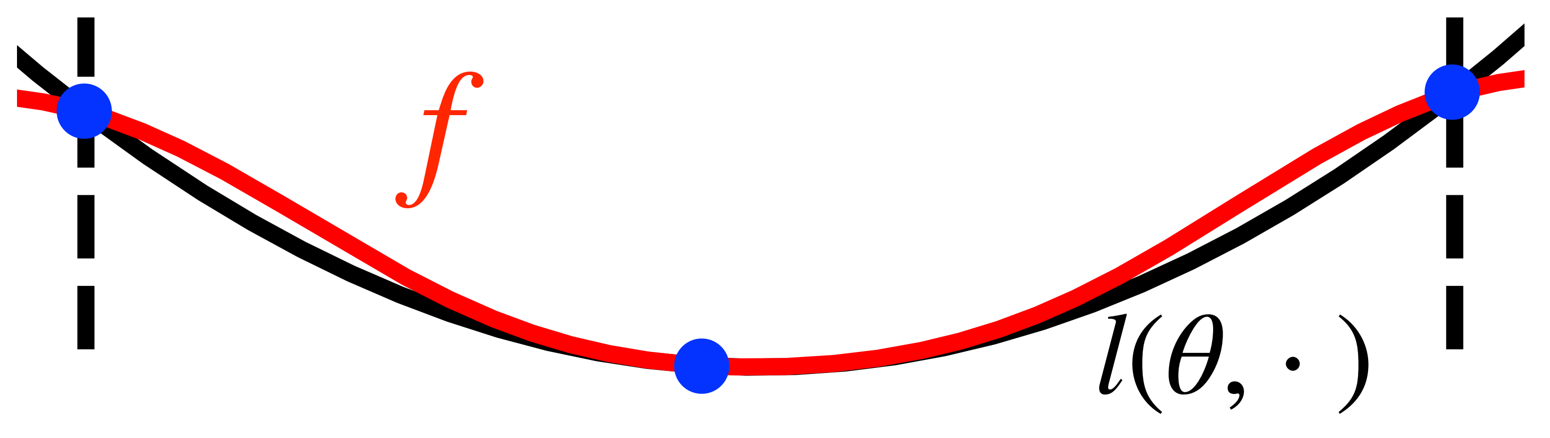
Q: what if the l is the loss for a **nonlinear** model (such as deep neural nets)?

Kernel DRO

Primal: $$\min_{\theta} \sup_{\mathrm{MMD}(P,\hat{P})\le\epsilon} \mathbb{E}_P\, l(\theta,\xi)$$

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

$$\text{s.t. } l(\theta,\xi) \le f(\xi), \forall\xi \text{ a.e.}$$



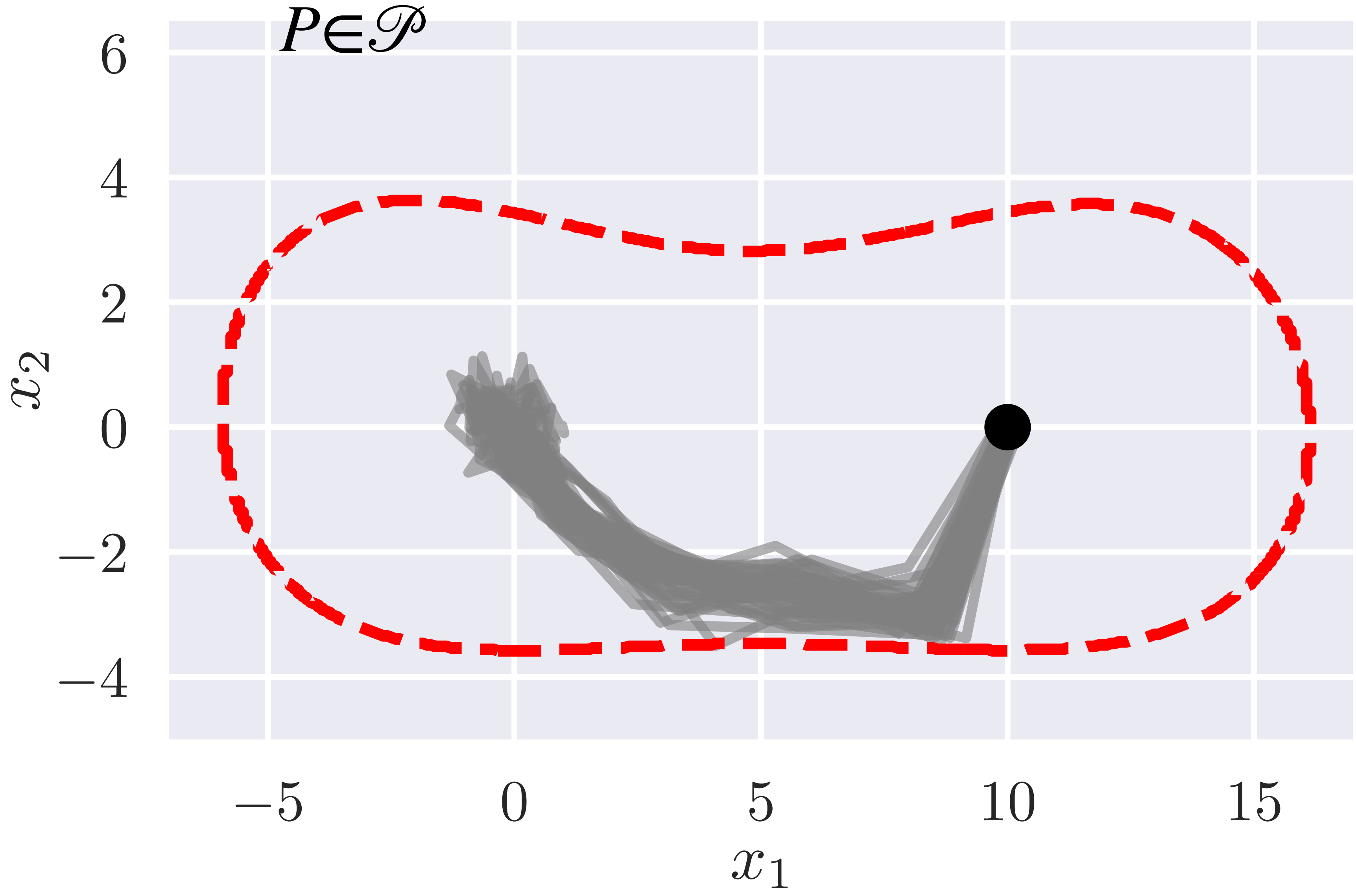
Applications: Distributionally Robust Deep Learning and Control

Application. Certified adversarially robust deep learning (Classify the presence of glasses using a **20-layer DNN** model)

Sinha et al. 2017; **Z** et al. 2022

Application. Distributional robust chance-constrained stochastic control with Bootstrapped ambiguity

$$\inf_{P \in \mathscr{P}} P(f(x, \xi) \le 0) \ge 95\,\%$$



Nemmour et al. 2022

Variational problem of dynamical systems

We can evolve the discrete time dynamical system by solving the variational problem (Jordan et al. 1998)

$$\rho_{t+1} = \mathrm{argmin}_{\varrho} F(\varrho) + \frac{1}{2\tau} W_2^2(\varrho, \varrho_t)$$

- **Question**: What if we don't know the physical law that governs the evolution of the system, e.g., F unknown?

SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 1–17, January 1998

THE VARIATIONAL FORMULATION OF THE FOKKER–PLANCK EQUATION*

RICHARD JORDAN†, DAVID KINDERLEHRER‡, AND FELIX OTTO§

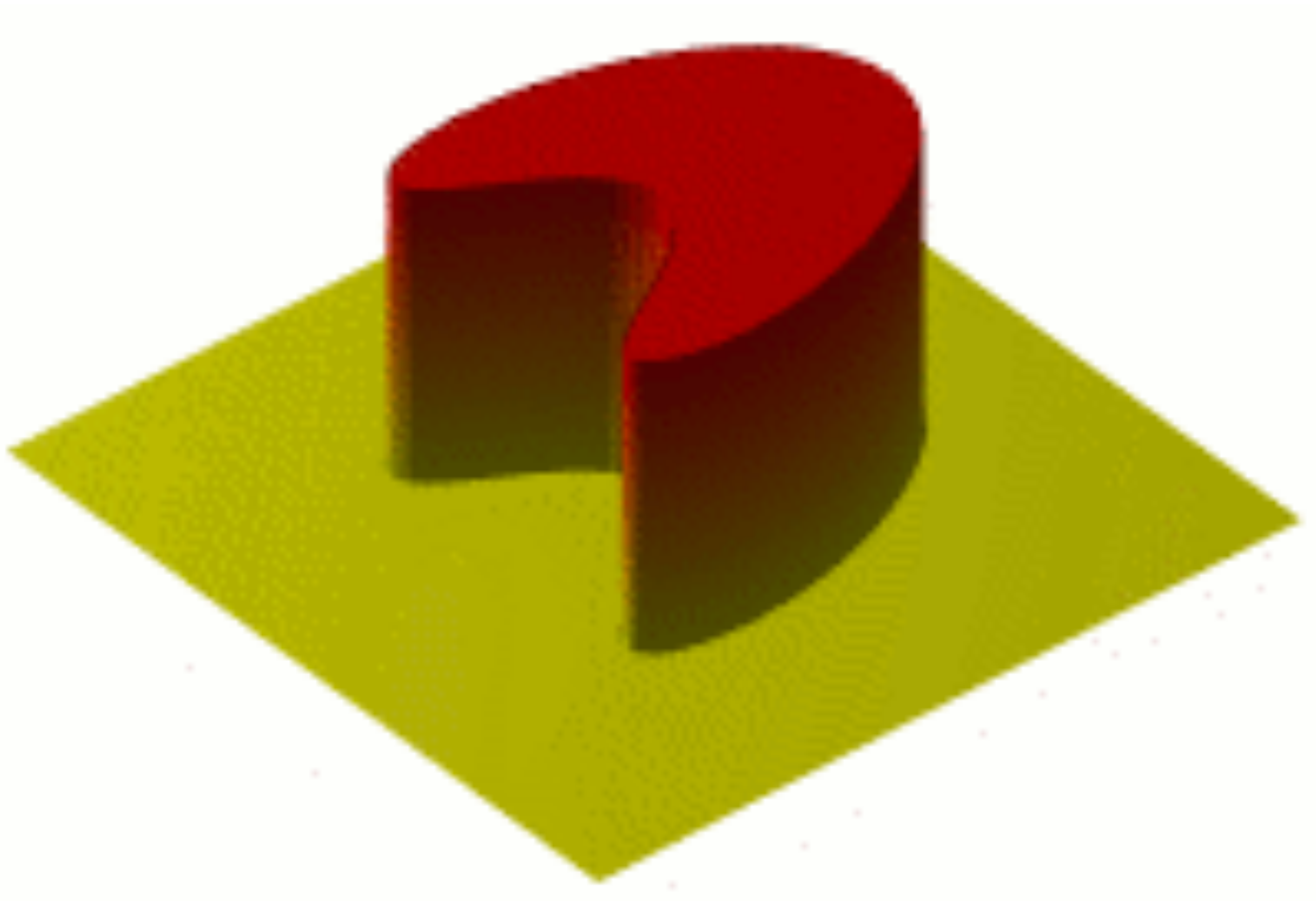
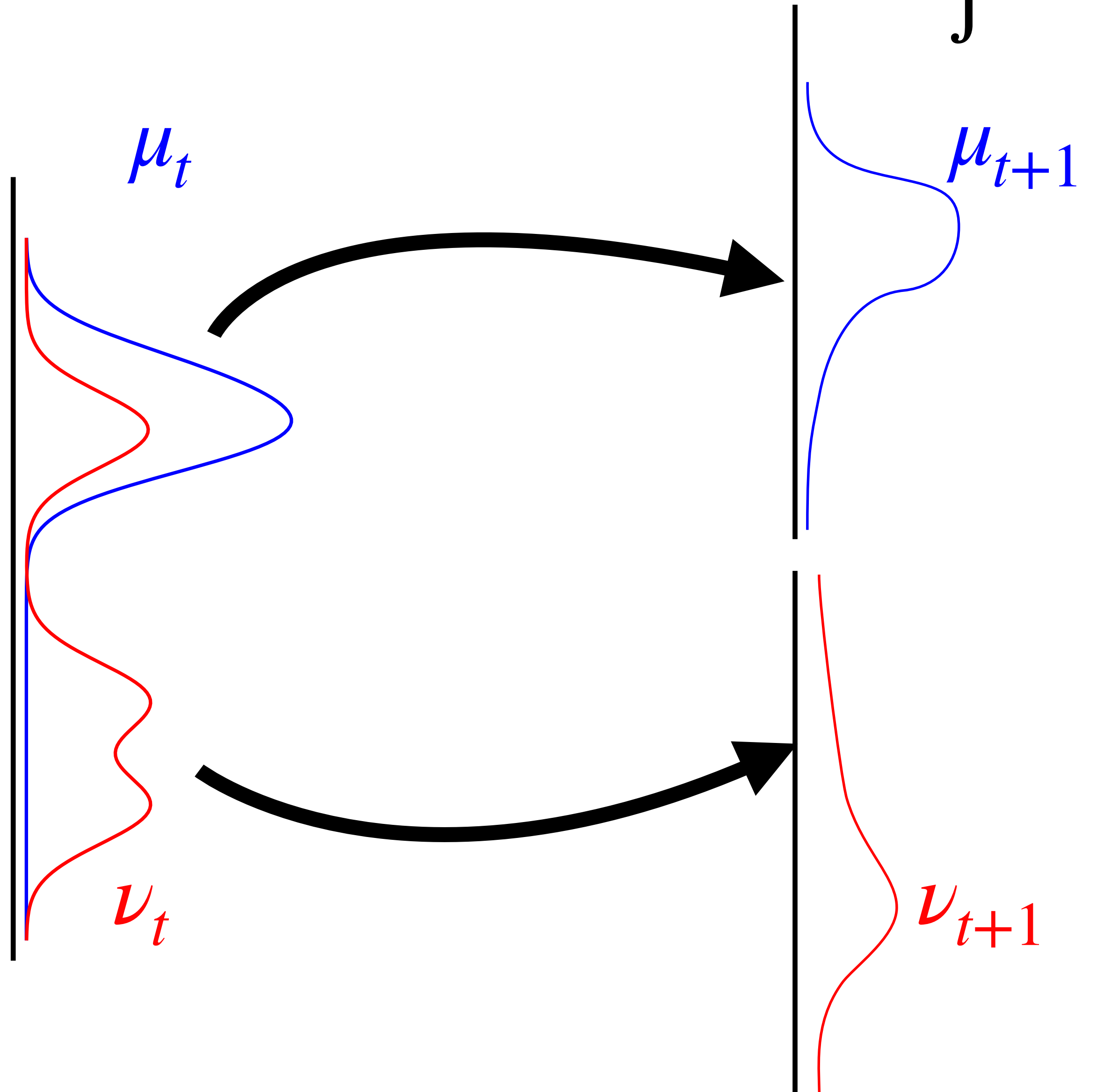


Image credit: https://en.wikipedia.org/wiki/Heat_equation

MMD Motivation: data-driven modeling of dynamical systems

We can use a data-driven model to model the unknown/uncertain dynamical systems from data/observation (Koopman theory, conditional embedding, etc.)

$$\mu_{t+1} = \mathscr{K}\mu_t, \quad \mathscr{K} := \mathscr{C}_{XY}(\mathscr{C}_{XX})^{-1}, \; \mu_P := \int k(x, \cdot)\, dP(x)$$



Unlike the gradient flow in W_2, the distance between the evolving data-driven dynamics models can be conveniently measured in the Hilbert norm

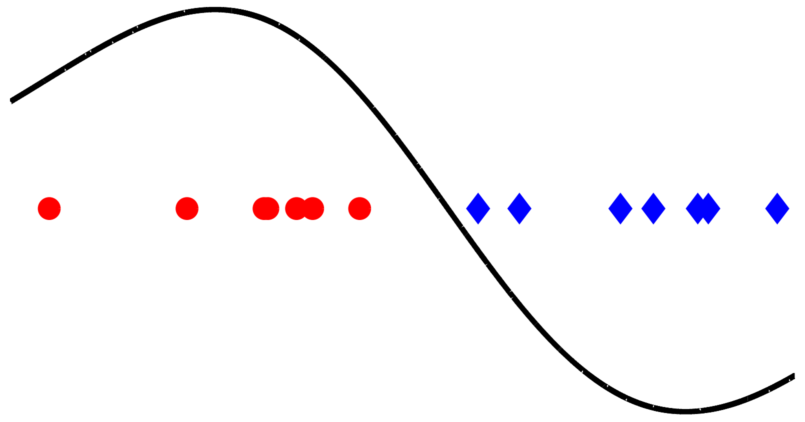
$$\|\mu_{t+1} - \nu_{t+1}\|_{\mathscr{H}} = \|\mathscr{K}\mu_t - \mathscr{K}\nu_t\|_{\mathscr{H}}$$
$$\leq \|\mathscr{K}\| \|\mu_t - \nu_t\|_{\mathscr{H}}$$

This motivates us to use this *Hilbert norm (i.e. MMD) as a natural tool for working with such data-driven models*.

Song et al. 2009, Fukumizu et al. 2011, Williams et al. 2015, Klus et al. 2018 etc.

Summary of Kernel DRO

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

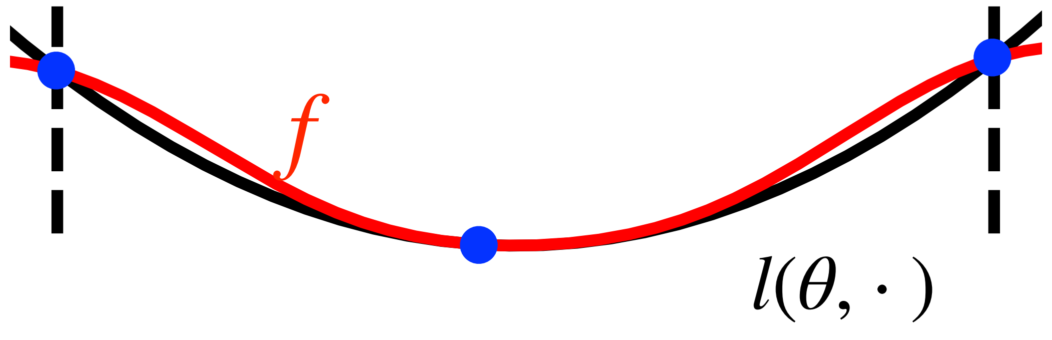


Kernel DRO

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

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

$$\text{s.t.} \; l(\theta, \cdot) \leq f \text{ a.e.}$$



A generalized dual algorithm for solving DRO with probability metric-balls, for nonlinear (non-convex) loss function

✓ Flatten the curve, smooth is robust

Some works on this topic

- **Zhu**, J.-J., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. **AISTATS 2021**
- **Zhu**, J.-J., Kouridi, C., Nemmour, Y. & Schölkopf, B. Adversarially Robust Kernel Smoothing. **AISTATS 2022**
- Nemmour, Y. , Kremer, H., Schölkopf, B. & **Zhu**, J.-J. Maximum Mean Discrepancy Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee. **IEEE CDC 2022**; Journal version WIP
- Kremer, H., **Zhu**, J.-J., Muandet, K. & Schölkopf, B. Functional Generalized Empirical Likelihood Estimation for Conditional Moment Restrictions. **ICML 2022**

Website: jj-zhu.github.io

Positions available in Berlin (PhD & postdoc)