

# From Gradient Flow Force-Balance to Distributionally Robust Learning

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics  
Berlin, Germany

European Conference on Computational Optimization (EUCO)  
Heidelberg University, Germany. September 26th, 2023



# Distributional robustness, but what kind?

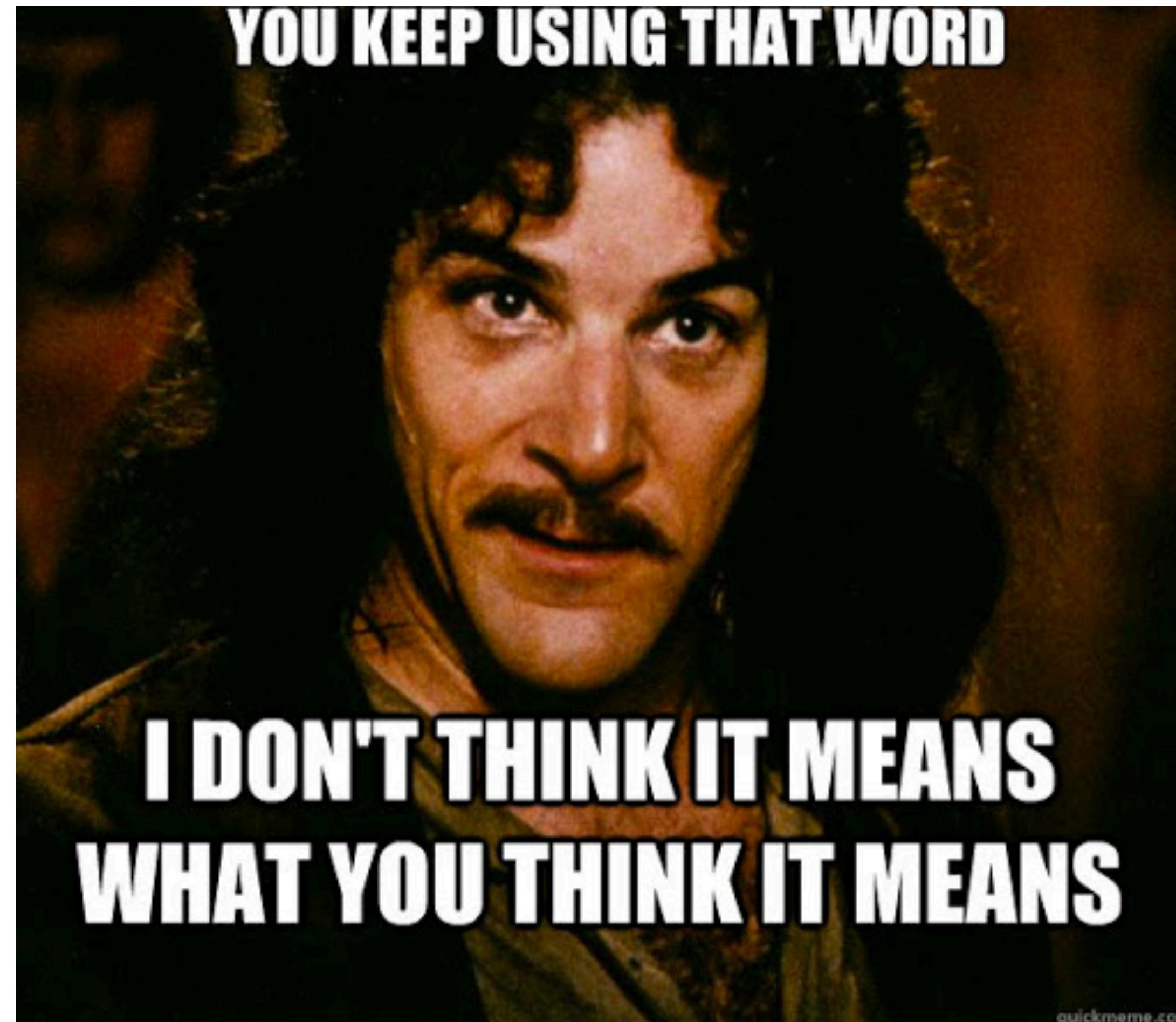


Figure credit: The Princess Bride,  
a bedside story by your grandpa

# Motivation: From Statistical Learning to 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  $Q (\neq P_0)$

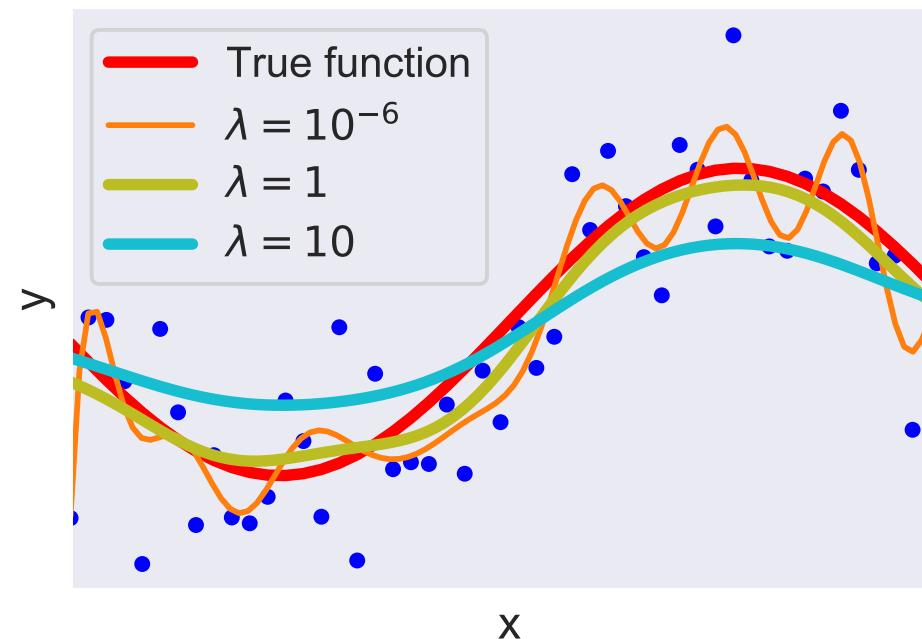
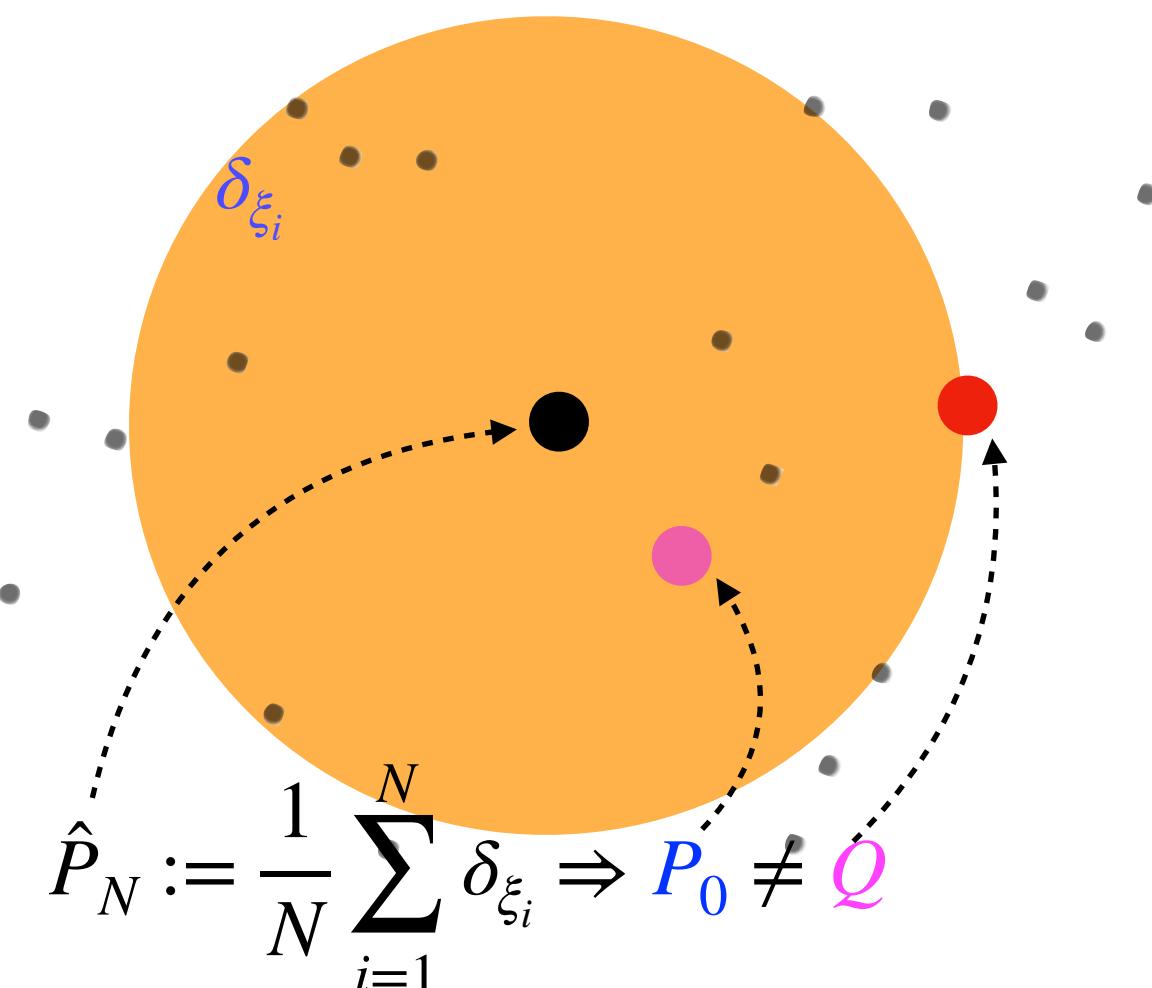


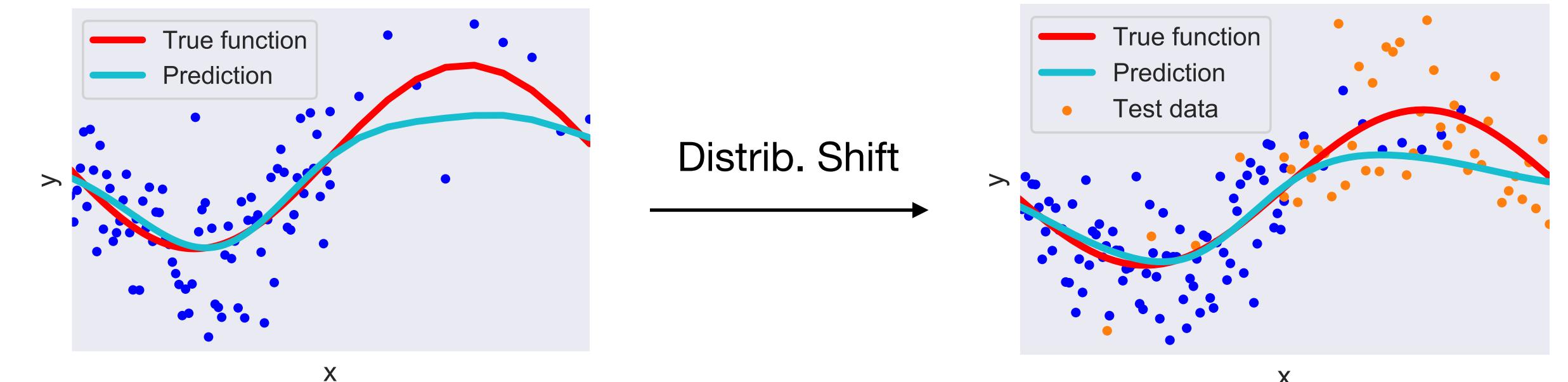
Figure credit: H. Kremer, J. Zhu



## Distributionally Robust Optimization (DRO)

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

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. descent in Fisher-Rao geometry

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

# Background: Kantorovich-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

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

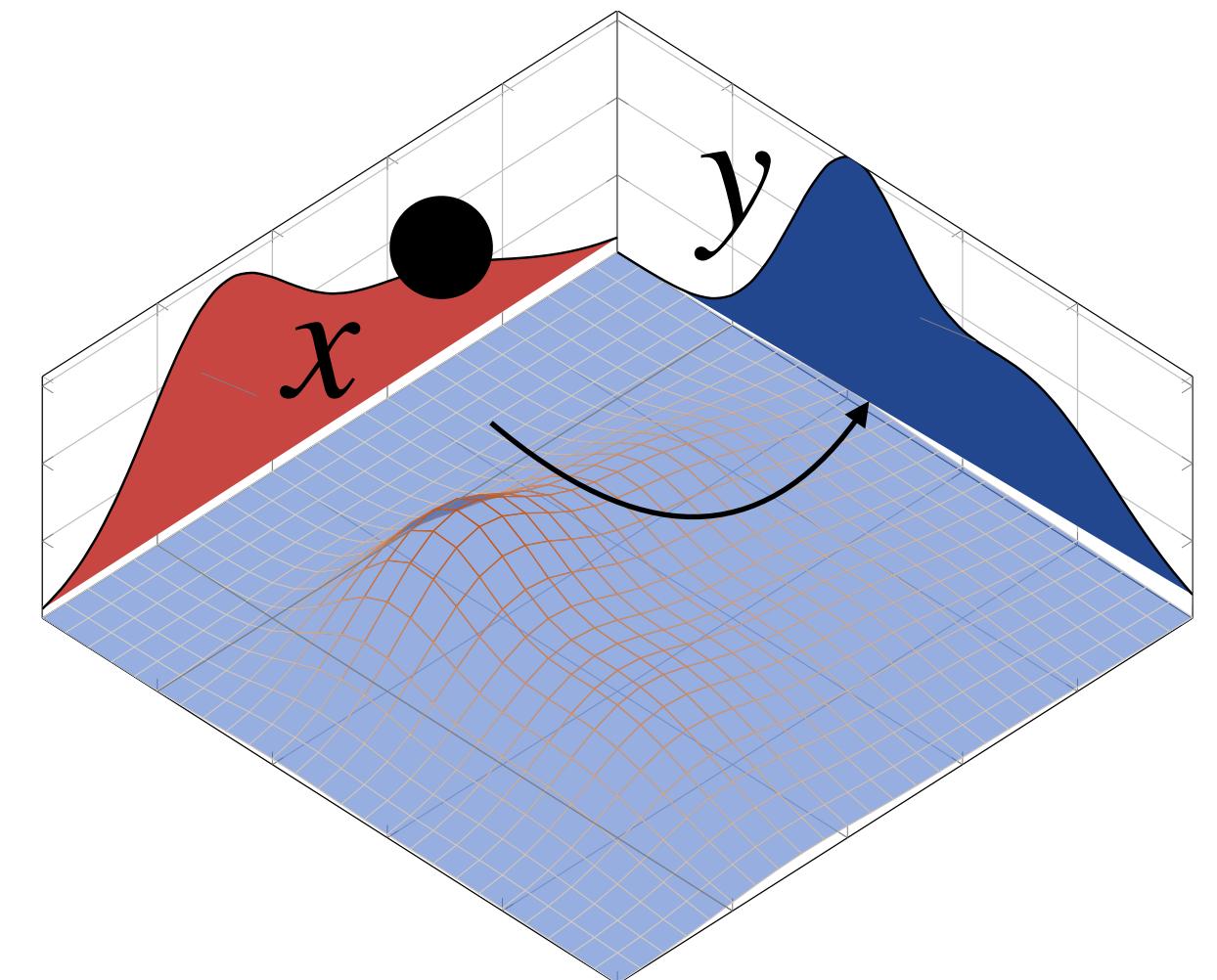
**(Dual Kantorovich problem)**

$$= \sup \left\{ \int \psi_1(x) dP(x) + \int \psi_2(y) dQ(y) \mid \psi_1(x) + \psi_2(y) \leq |x - y|^p \right\}$$

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

**Dynamic formulation: Benamou-Brenier**

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



# **Gradient Flow Force-Balance**

# Gradient flow facts

Otto's Gradient flow equation in the Wasserstein space

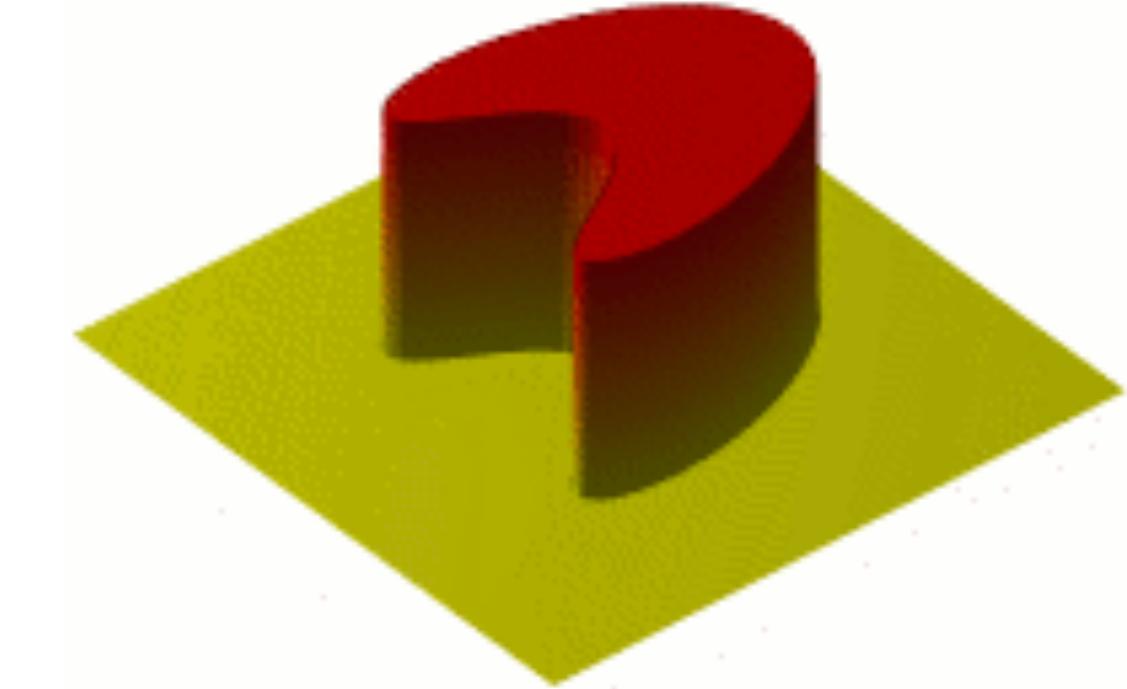
$$\partial_t \mu - \nabla \cdot (\mu \nabla \frac{\delta F}{\delta \mu}[\mu]) = 0$$

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

© 1998 Society for Industrial and Applied Mathematics

## THE VARIATIONAL FORMULATION OF THE FOKKER-PLANCK EQUATION\*

RICHARD JORDAN<sup>†</sup>, DAVID KINDERLEHRER<sup>‡</sup>, AND FELIX OTTO<sup>§</sup>



It describes the “steepest” dissipation of energy  $F$  in  $(\text{Prob}(\bar{X}), W_2)$ .  
[Otto et al 90s-2000s, Ambrosio 2005, ...]

In a different flavor, we can write it just like ODE gradient flow  
 $\dot{x} = -\nabla f(x)$  in the **primal rate-form**

$$\dot{\mu} = -\mathbb{K}_{\text{Otto}}(\mu) DF \quad (\text{DF is the (sub)diff., e.g., in the sense of Fréchet})$$

Time-discretization yields the *minimizing movement scheme* (MMS)

$$u_k \in \arg \inf_{u \in \mathcal{P}} F(u) + \frac{1}{2\tau} W_2^2(u, u_{k-1})$$

ODE flow: gradient descent

$$x^k \in \arg \min_{x \in \mathbb{R}^d} F(x) + \frac{1}{2\tau} \|x - x^{k-1}\|^2.$$

# Gradient flow facts

*Minimizing movement scheme* (MMS) for a generalized (Banach) gradient system

$$u_k \in \arg \inf_{u \in \mathcal{P}} F(u) + \tau \mathcal{R} \left( u_{k-1}, \frac{1}{\tau} (u - u_{k-1}) \right)$$

e.g.,  $\mathcal{R}$  can be the *squared “Wasserstein norm”*. We must specify the ingredients

Space : $\mathcal{P}$  or  $\mathcal{M}^+$ , Energy : $F$ , Dissipation Geometry : $\mathcal{R}$

“The merit of the right gradient flow formulation of a dissipative evolution equation is that it separates energetics and kinetics: The **energetics** endow the state space with a **functional**, the **kinetics** endow the state space with a (Riemannian) **geometry** via the metric tensor.” [Otto 2001]

# Gradient Flow Force-Balance

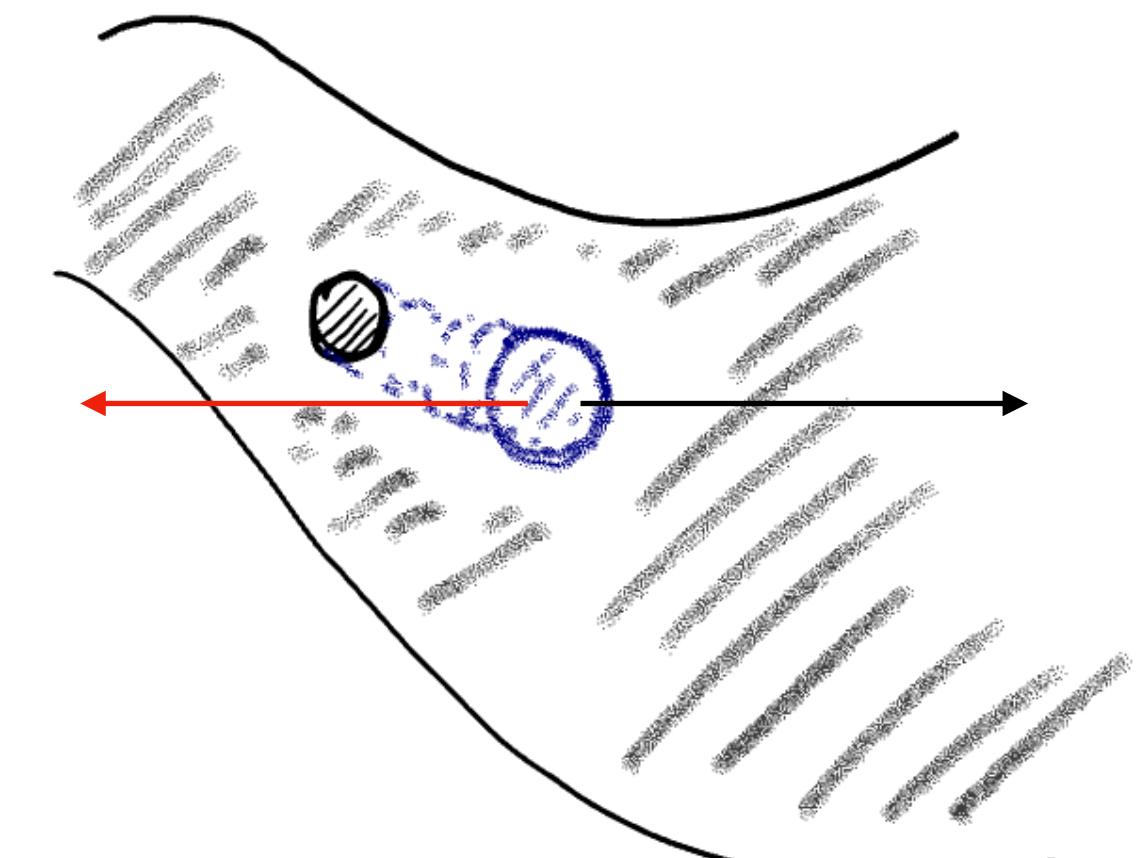
Force-balance in Wasserstein MMS  $u_k \in \arg \inf_{u \in \mathcal{P}} F(u) + \frac{1}{2\tau} W_2^2(u, u_{k-1})$

$$DF + \frac{\phi}{\tau} = \text{const.}, \text{ where } \phi \text{ is the Kantorovich potential}$$

Force: drive movement      Dissip. Geom.: resist movement, e.g., viscosity

In practice, approximate  $\phi$  based on data samples using **function approximators**, NN/ICNN/RKHS, e.g.,

$$\phi \approx f = \sum_{i=1}^n \alpha_i k(x_i, \cdot) \in \mathcal{H}.$$



Force-balance in ODE:

$$\nabla f(x_t) + \frac{x_t^\top - x_{t-1}^\top}{\tau} = 0 \in X^*$$

We will now see two applications of this force-balance relation to robust learning

# **Robust Learning under (Joint) Distribution Shift**

# Background: MMD and interaction force

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

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

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

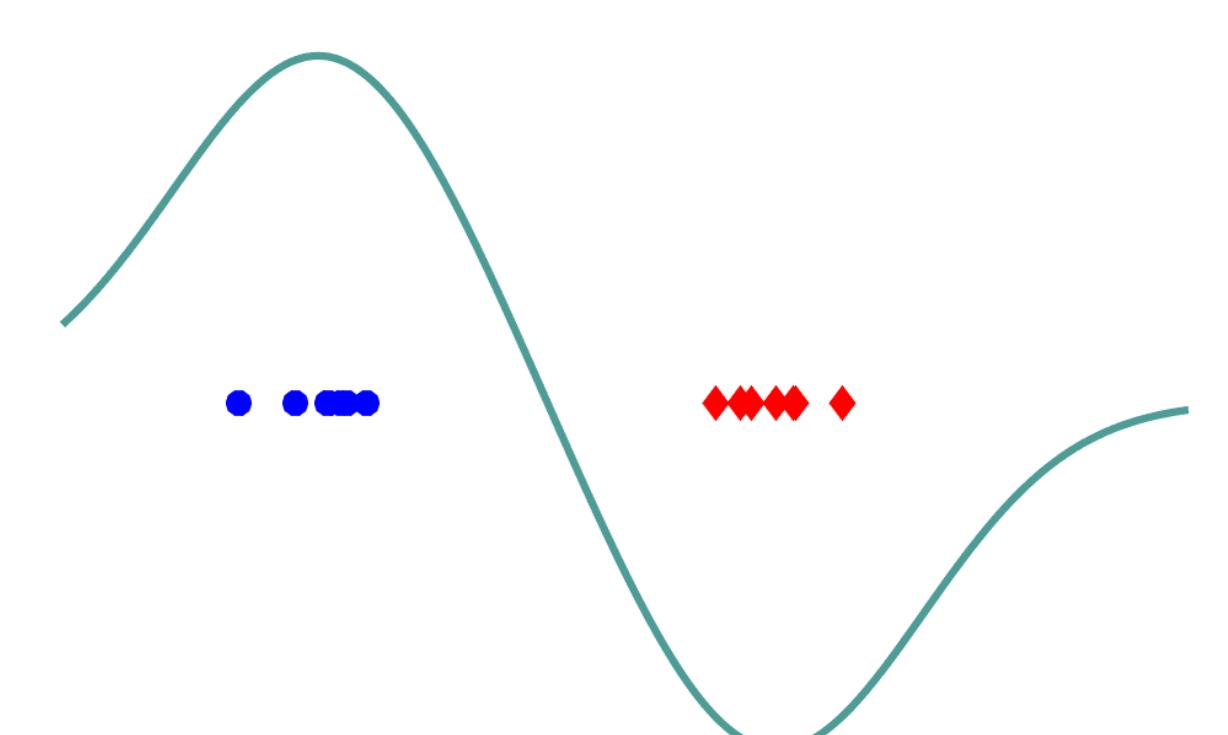
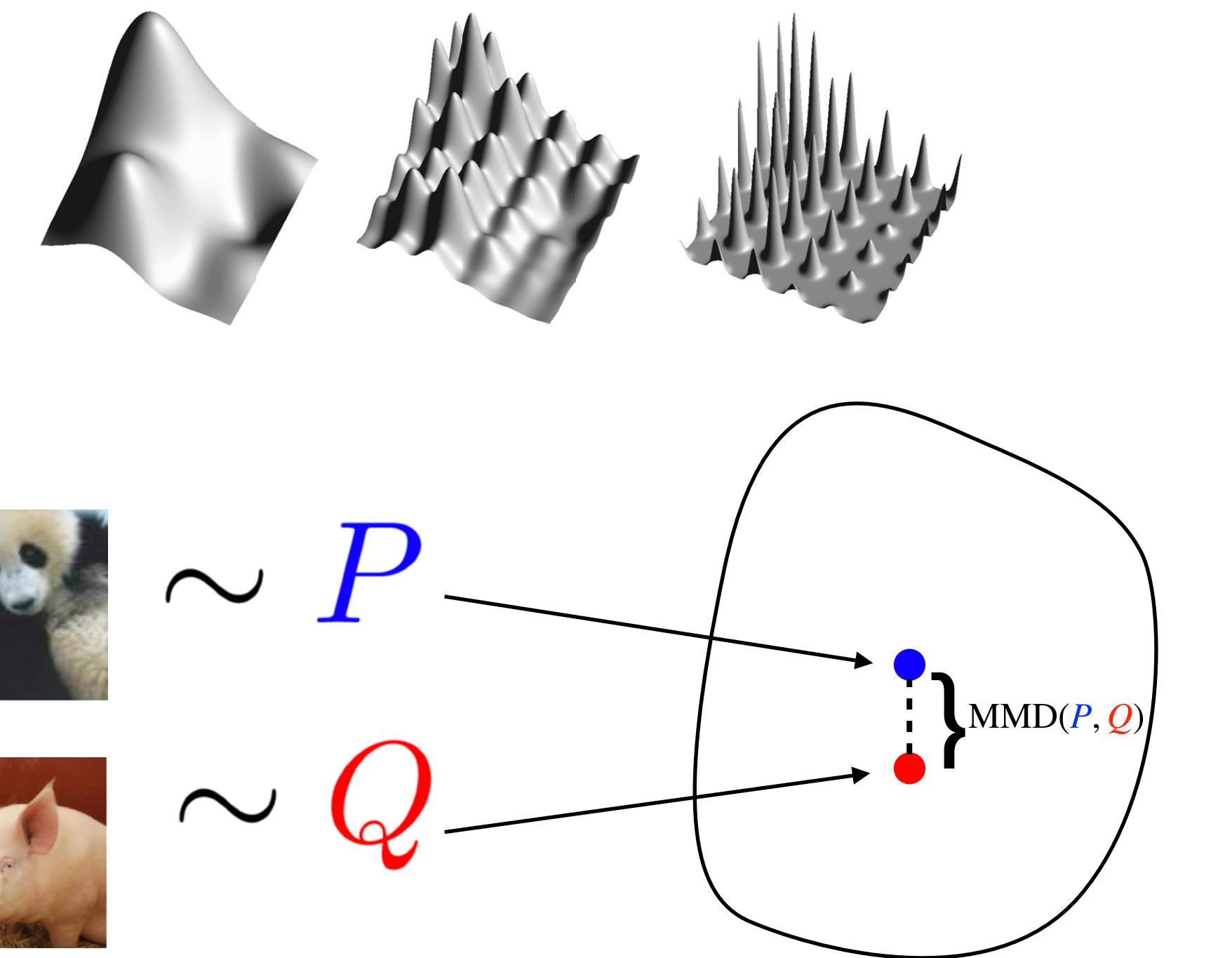
**Dual formulation as an integral probability metric.**

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

$\mathcal{H}$  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  $\mathcal{H}$ .

**As an interaction energy for Wasserstein GF** [Arbel et al.]

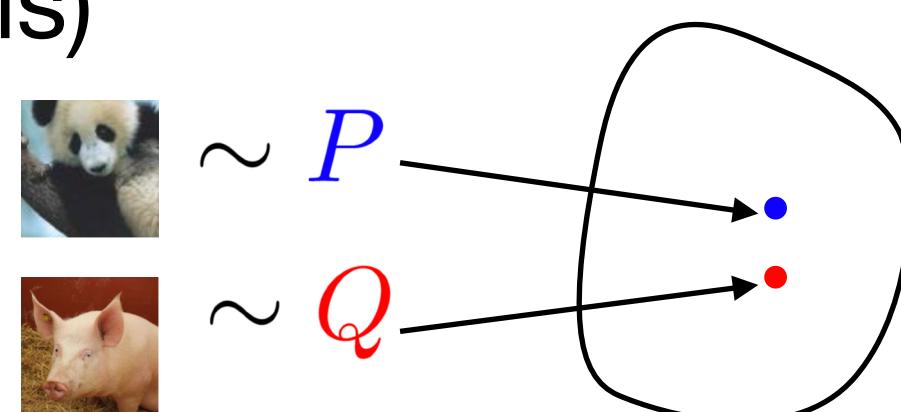
$$\text{MMD}^2(\mathbf{P}, \mathbf{Q}) = \iint k(x, y) d(\mathbf{P} - \mathbf{Q})(x) d(\mathbf{P} - \mathbf{Q})(y)$$



# Kernel DRO under distribution shift

**Primal DRO** (not solvable as it is)

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



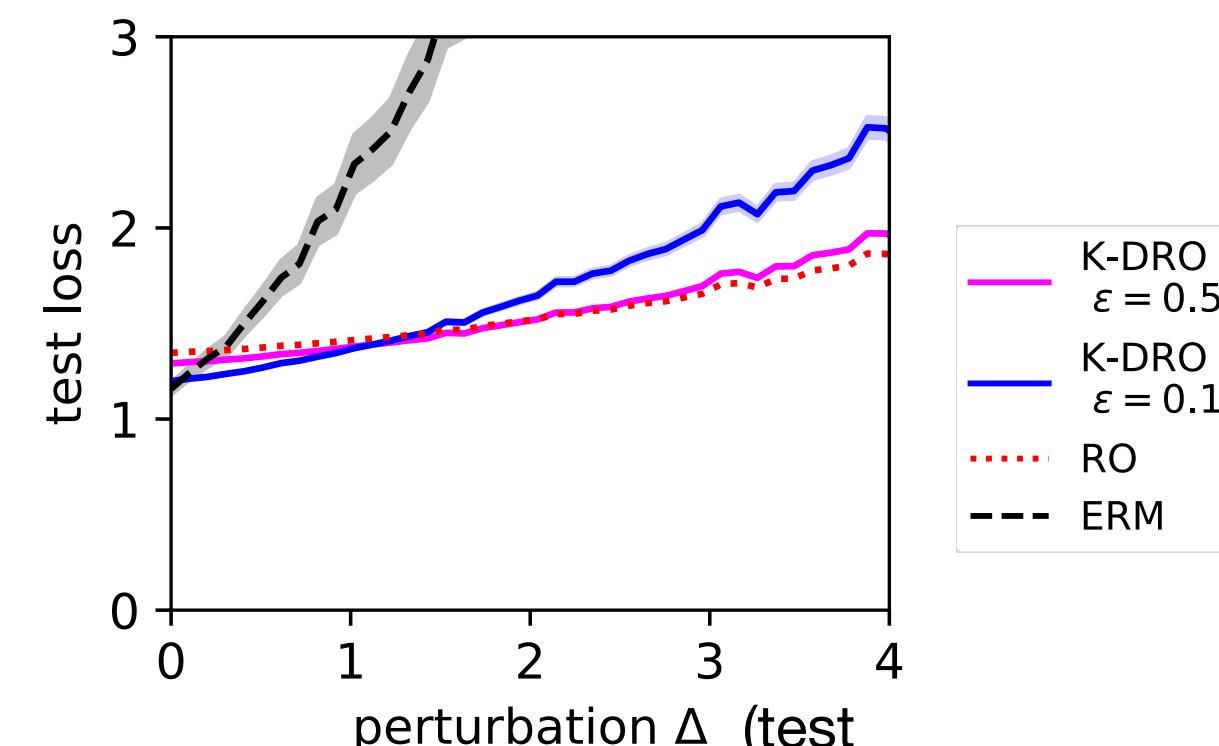
**Kernel DRO Theorem (simplified).** [Z. et al. 2021]

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

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

**Example. Robust least squares**

$$\min l(\theta, \xi) := \|A(\xi) \cdot \theta - b\|_2^2$$



**Entropy regularization (“interior point method”)**

$$\text{MMD}(Q, \hat{P}) + \lambda D_\phi(Q \parallel \omega) \leq \epsilon$$

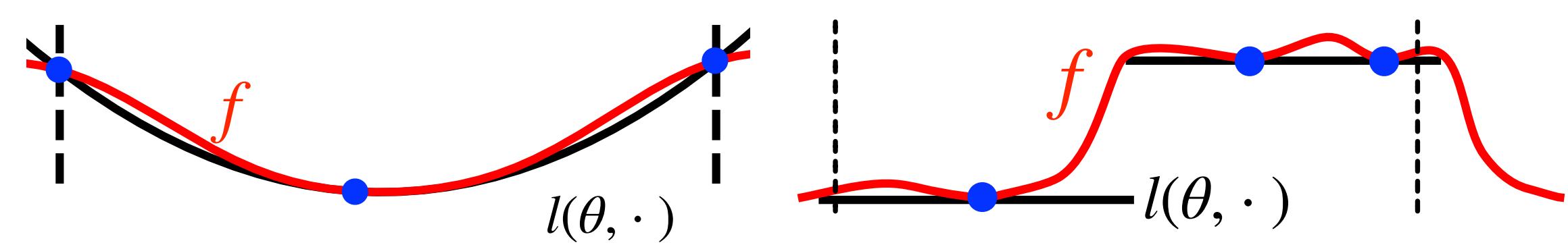
**Dual.** Adapted from [Kremer et al., Z. 2023]

$$\inf_{\theta, f \in \mathcal{H}} \left\{ \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathcal{H}} + \lambda \mathbb{E}_{\omega} \phi^* \left( \frac{-f + l}{\lambda} \right) \right\}$$

soft cons.  $\phi_{\text{KL}}^*(t) = \exp(t)$

log-barrier  $\phi_{\log}^*(t) = -\log(1-t)$

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



# Force-balance of Kernel DRO

Dual program:  $\min_{\theta, \mathbf{f} \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N \mathbf{f}(\xi_i) + \epsilon \|\mathbf{f}\|_{\mathcal{H}}$  s.t.  $l(\theta, \xi) \leq \mathbf{f}(\xi), \forall \xi$  a.e.

Lagrangian:  $\min_{\theta, \gamma \geq 0} \sup_{\mu \in \mathcal{P}} \mathbb{E}_{\mu} l(\theta, x) - \gamma \cdot \text{MMD}^2(\mu, \hat{\mu}_N) + \gamma \epsilon^2$

MMS in kernel-MMD

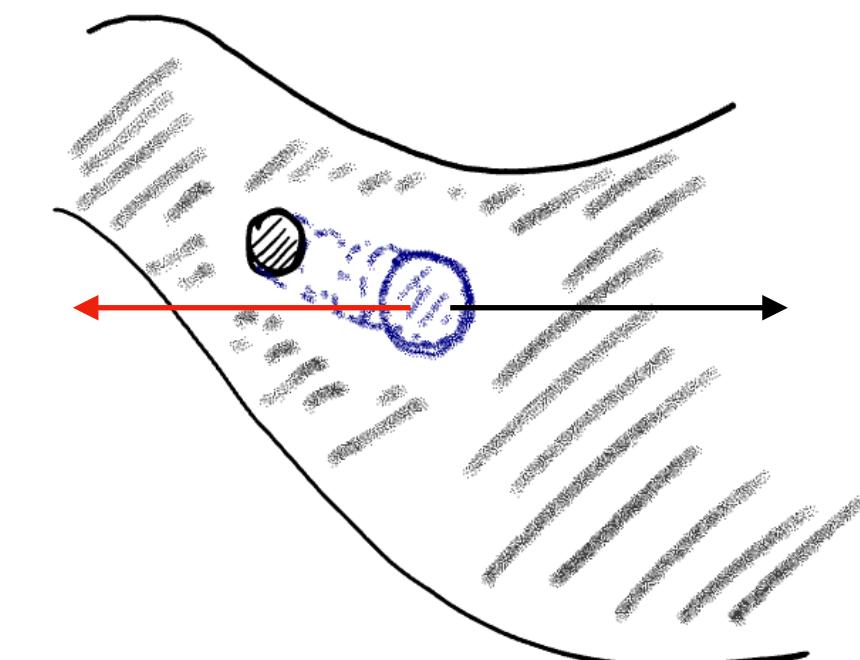
$$\inf_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} \text{MMD}^2(\mu, \mu^k) \implies -D^L F = \boxed{\frac{1}{\tau} \int k(x, \cdot) d(\mu - \mu^k)(x)} + \text{const.}$$

Force-balance using **function approximation** RKHS functions, e.g.,  $=: \mathbf{f} \in \mathcal{H}$

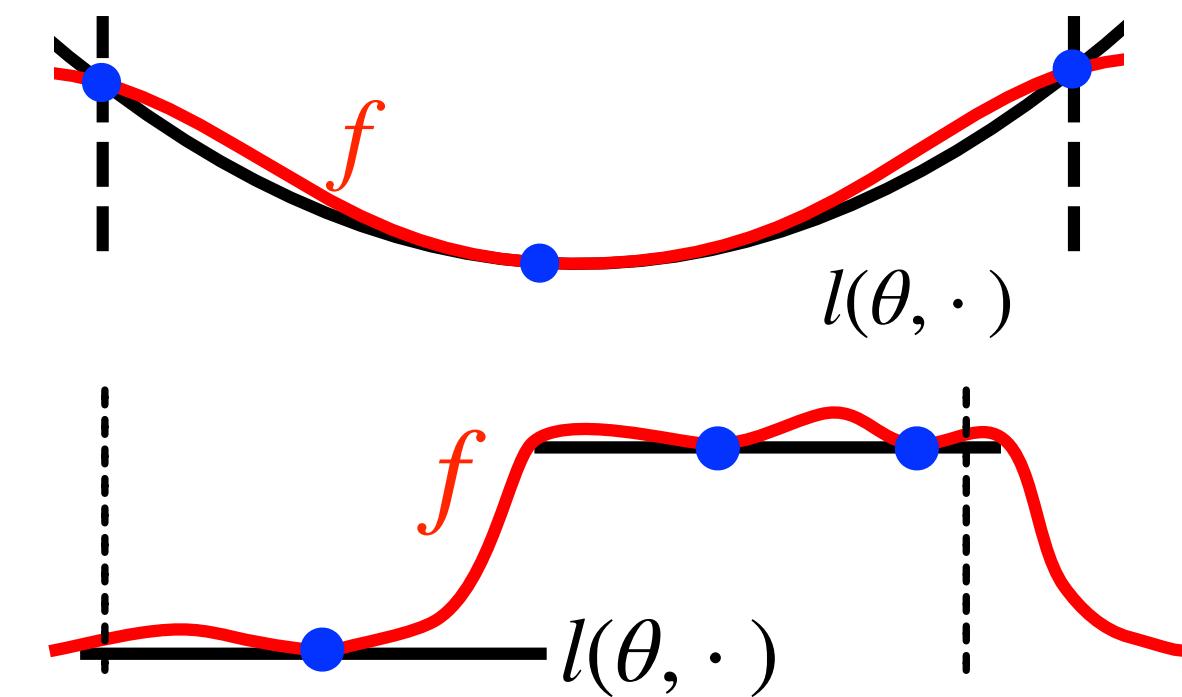
$$-DF = \mathbf{f} + f_0, \mathbf{f} = \sum_{i=1}^n \alpha_i k(x_i, \cdot) \in \mathcal{H}, f_0 \in \mathbb{R}$$

$DF = l(\theta, \cdot)$  under typical  $L^2$  first variation of probability measures  $\implies$  force-balance

$$l(\theta, \cdot) = \mathbf{f} + f_0 \text{ a.e.}$$



**Dual kernel function  $\mathbf{f}$**  as robust surrogate losses  
flatten the curve  $\rightarrow$  force balance



# **Robust Learning under Structured Distribution Shift**

# From Statistical Fluctuation to Structured Distribution Shift

Learning task  
What can go wrong?

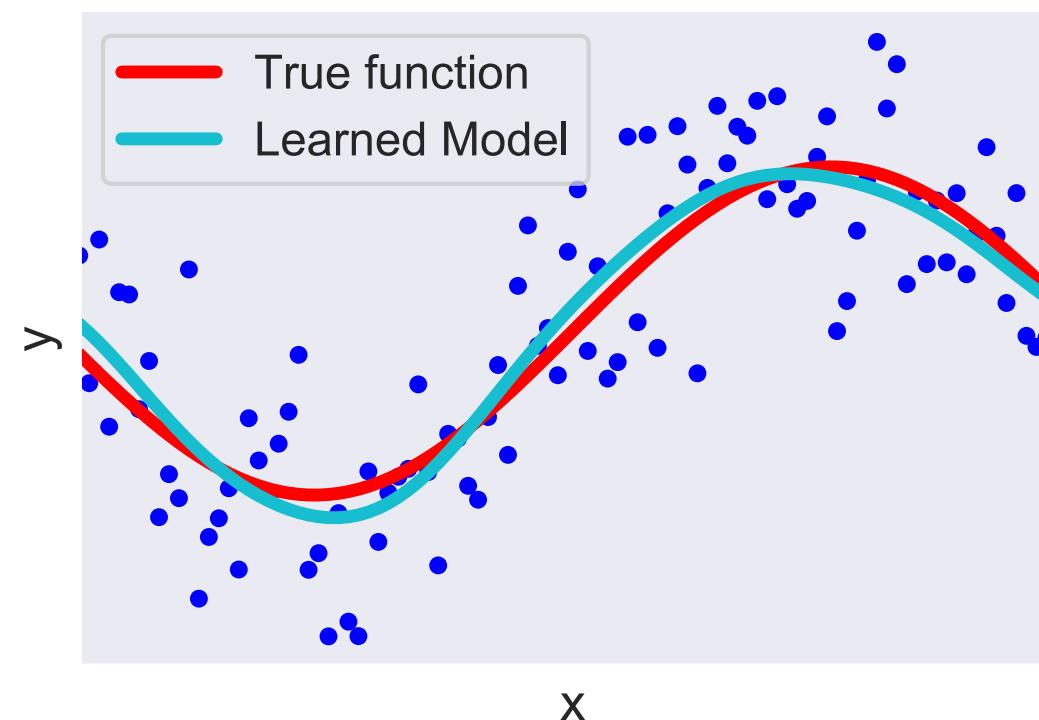
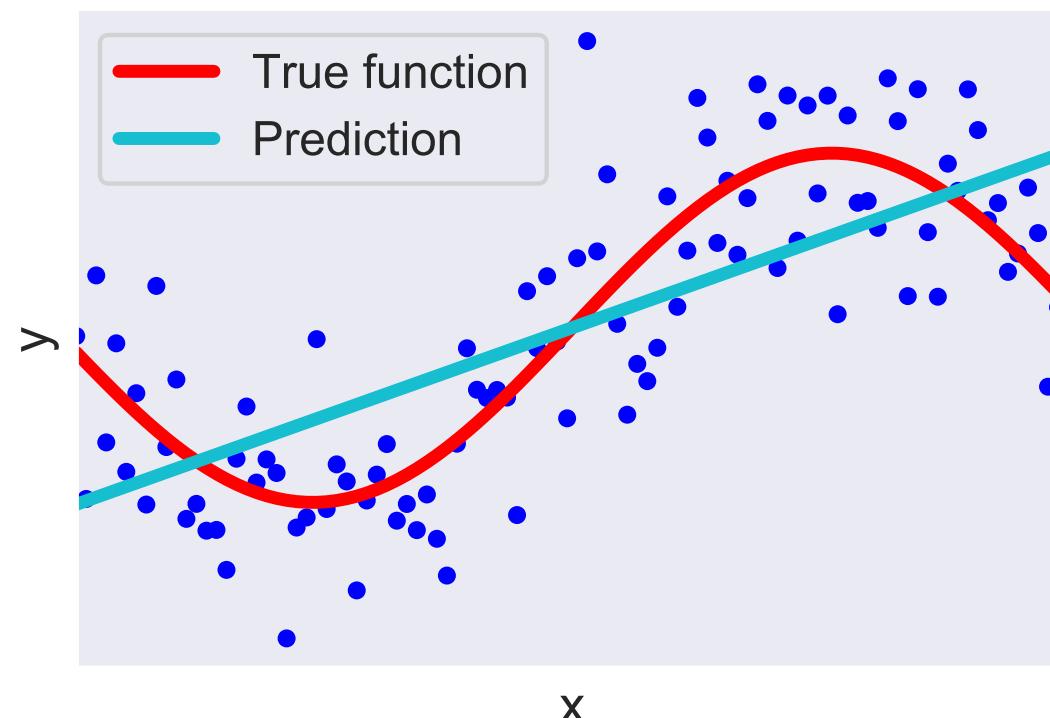


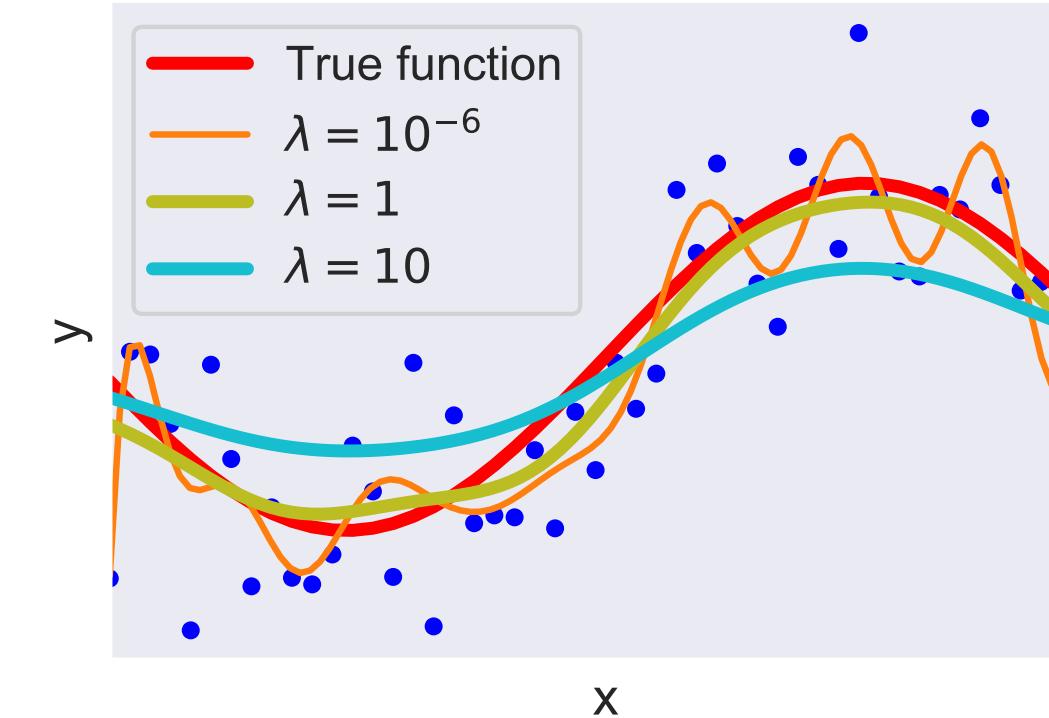
Figure credit: Heiner Kremer

Model mis-specification  
 $f_\theta \neq f \forall \theta \in \Theta$



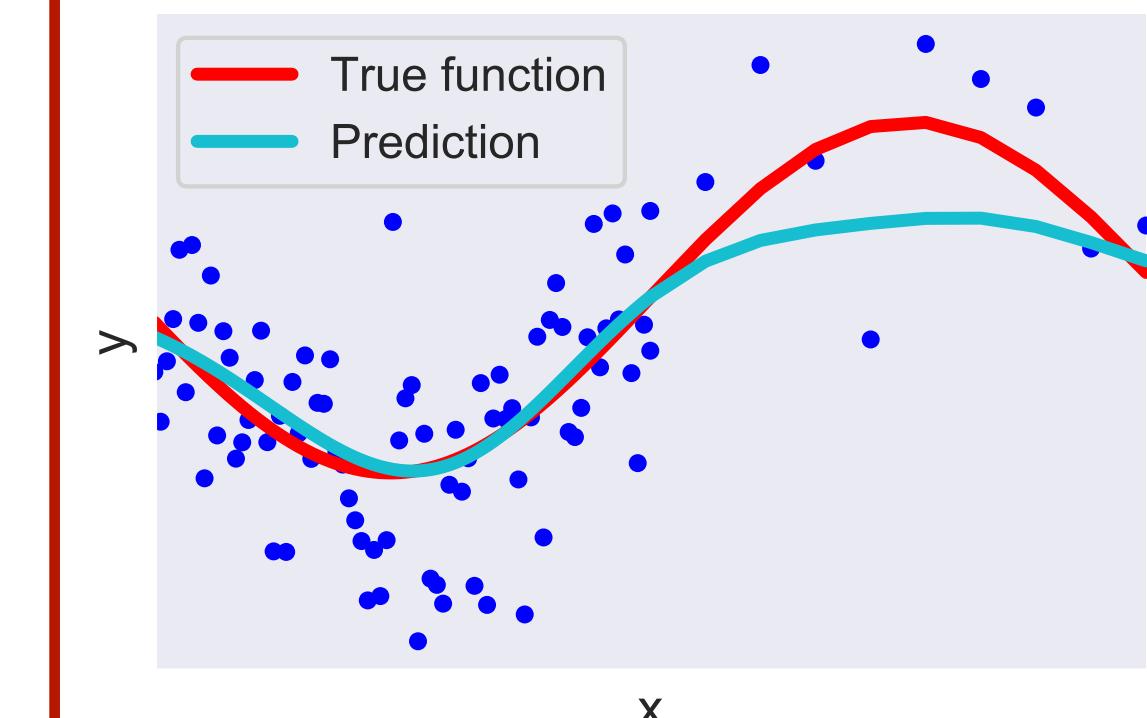
→ Use flexible models  
(NN/non-parametric)

Finite sample bias  
 $n \ll \infty$



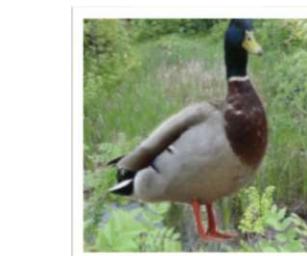
→ Statistical learning  
theory/regularization

Distribution shifts  
 $\mathbb{P}_{\text{test}} \neq \mathbb{P}_{\text{train}}$



→ Robustness,  
Causality

$Q_{\text{test}}$



waterbird  
+ land

$\hat{P}_{\text{train}}$



landbird  
+ land



waterbird  
+ water

[Sagawa et al. 2020]

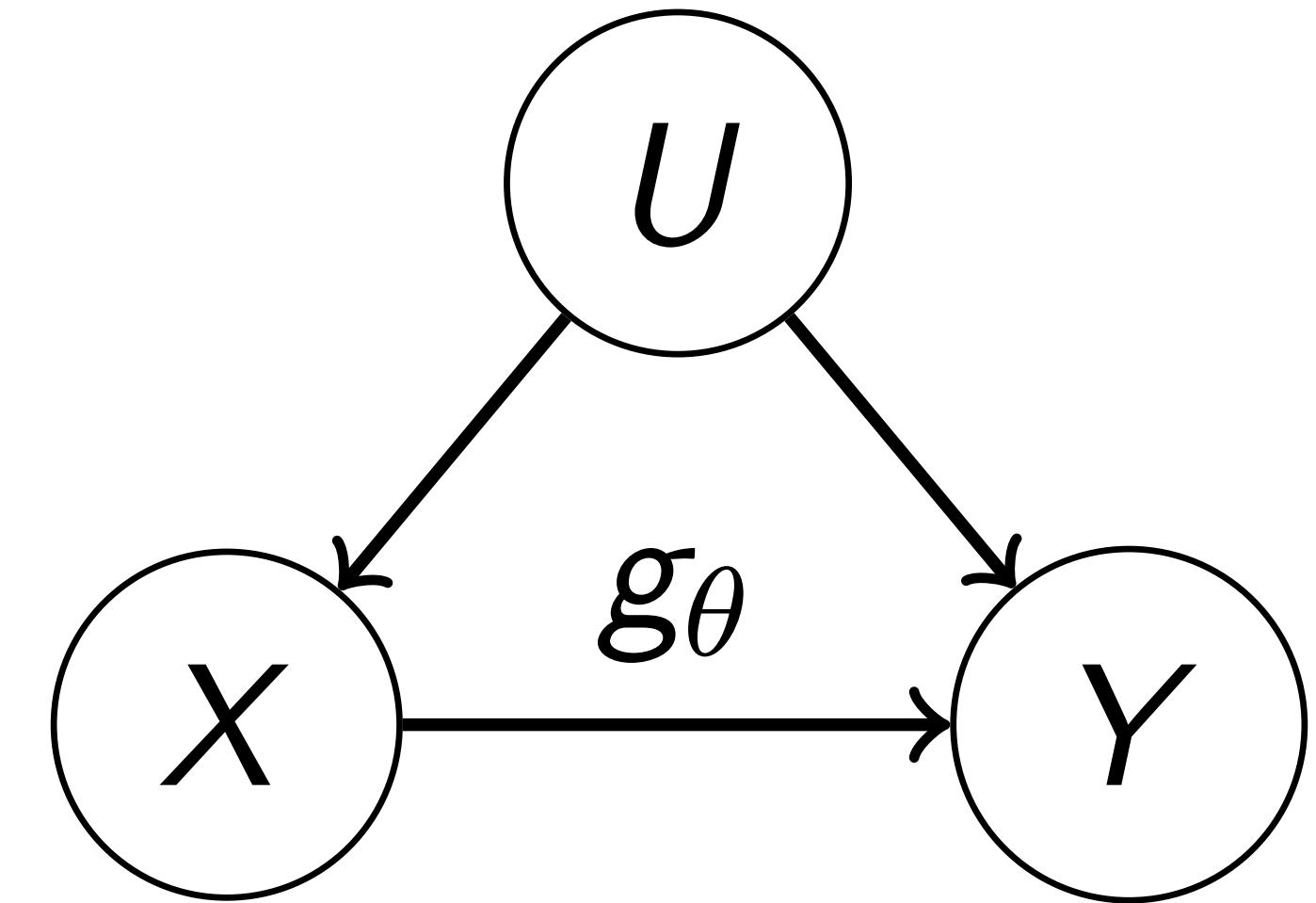
Wasserstein/Kernel DRO  
not suitable for (strong)  
structured distribution  
shifts !

# Structured Distribution Shift – Causal Confounding

**Causal confounding** can lead to much **stronger** distribution shifts than those considered in (**joint**) distribution shift, e.g., DRO, adversarial robustness.

$X$ : Smoking,  $Y$ : Cancer,  $U$ : Lifestyle

$$Y := g_\theta(X) + \epsilon_U, \quad \mathbb{E}[\epsilon_U] = 0, \text{ but } \mathbb{E}[\epsilon_U | X] \neq 0 \\ \implies g_\theta(x) \neq \mathbb{E}[Y | X = x]$$



Regression  $\min_{\theta} \mathbb{E}[\|Y - g_\theta(X)\|^2]$  or DRO does not work in this case.

# Kernel Method of Moment: conditional moment restriction for causal inference

Robustness against **structured distribution shifts** instead of (joint-)DRO. Estimating  $g_\theta$  via **conditional moment restriction (CMR)**

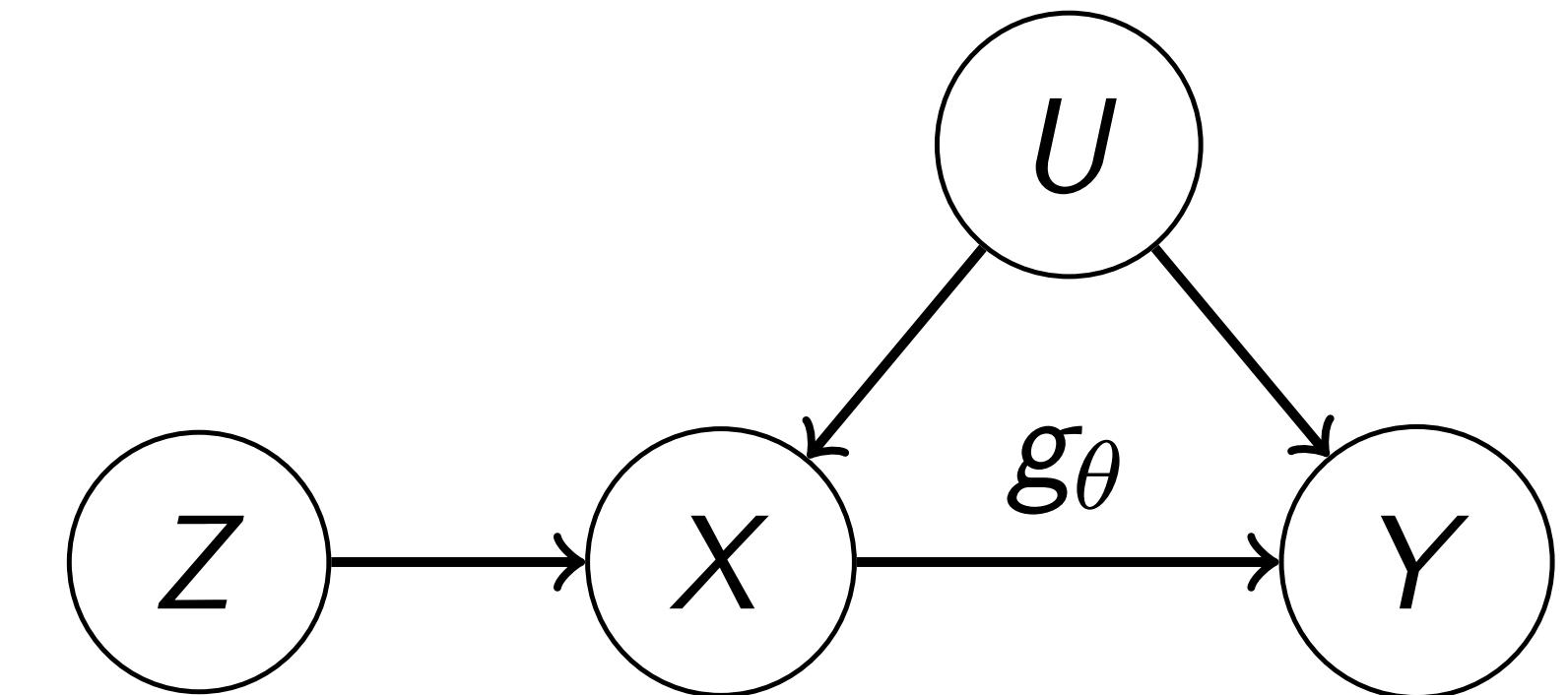
$$\mathbb{E}[Y - g_\theta(X) | Z] = 0 \text{ } \mathbb{P}_Z\text{-a.s.}$$

**Generalized Empirical likelihood** [Owen, 1988; Qin and Lawless, 1994] with **CMR** [Bierens, 1982]. Equivalently, generalized method of moment (GMM)

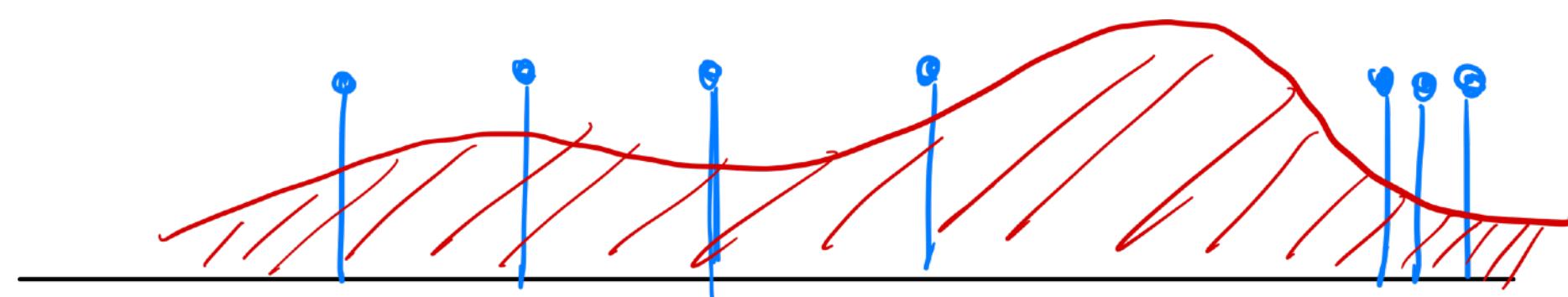
$$\inf_{\theta, Q \in \mathcal{P}} D_\phi(Q \parallel \hat{P}) \text{ s.t. } \mathbb{E}_Q \left[ (Y - g_\theta(X))^T h(Z) \right] = 0, \forall h \in \mathcal{H}$$

**Kernel MoM** [Kremer et al., 2023] with CMR

$$\inf_{\theta, Q \in \mathcal{P}} \frac{1}{2} \text{MMD}^2(Q, \hat{P}) \text{ s.t. } \mathbb{E}_Q \left[ (Y - g_\theta(X))^T h(Z) \right] = 0$$



Instrument: Genetic predisposition for nicotine addiction  $Z$



Lift the restriction that  $Q$  is an atomic distribution

# Kernel MoM: duality and algorithm

$$\theta^{\text{KMM}} = \arg \min_{\theta} R(\theta)$$

$$R(\theta) := \inf_{Q \in \mathcal{P}} \frac{1}{2} \text{MMD}^2(Q, \hat{P}) \text{ s.t. } \mathbb{E}_Q \left[ (\psi(X; \theta))^T h(Z) \right] = 0$$

**Theorem.** [Kremer et al., Z. 2023] The MMD profile  $R(\theta)$  has the strongly dual form

$$R(\theta) = \sup_{\substack{f_0 \in \mathbb{R}, f \in \mathcal{F}, \\ h \in \mathcal{H}}} f_0 + \frac{1}{n} \sum_{i=1}^n f(x_i, z_i) - \frac{1}{2} \|f\|_{\mathcal{F}}^2$$

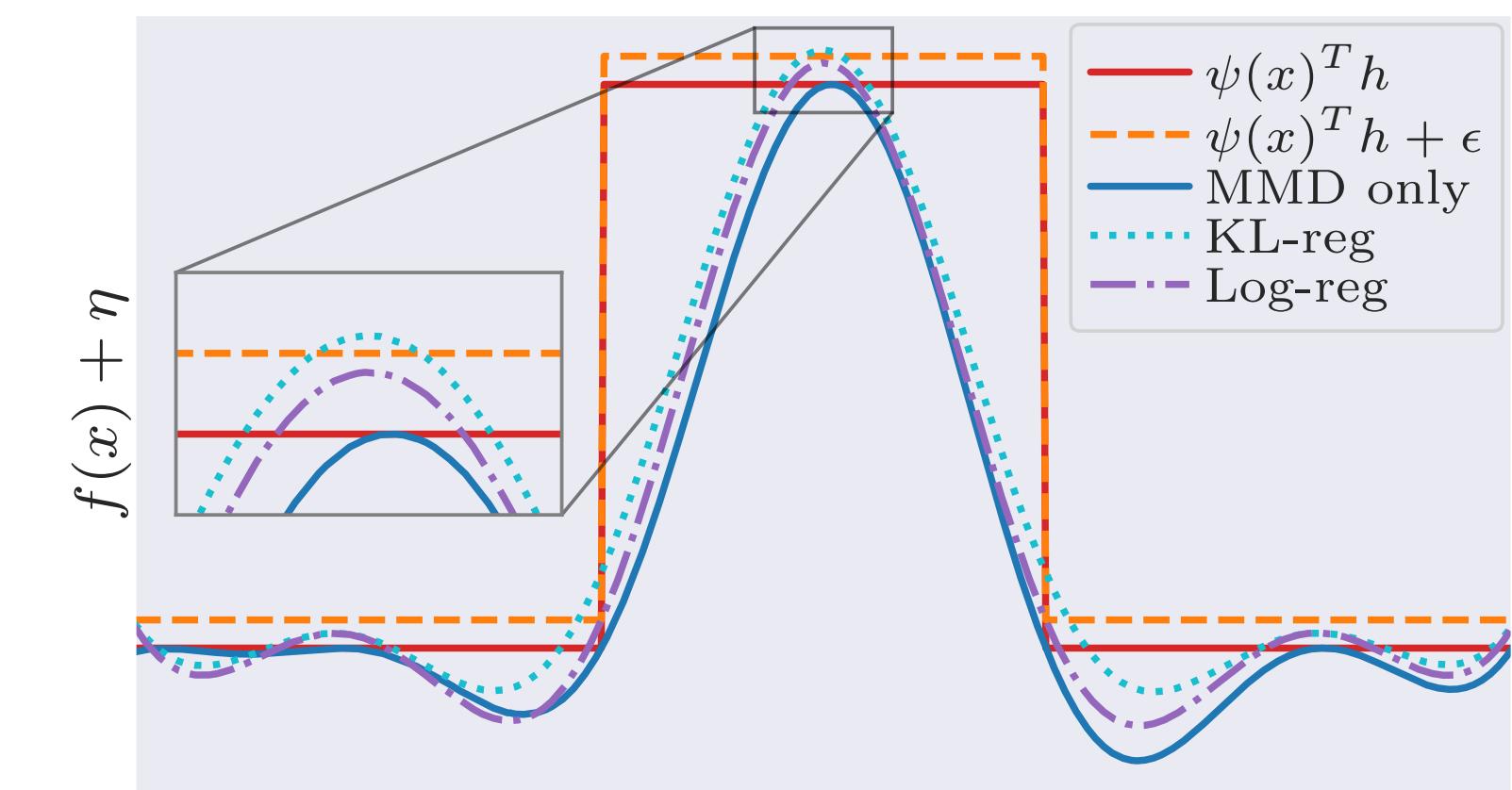
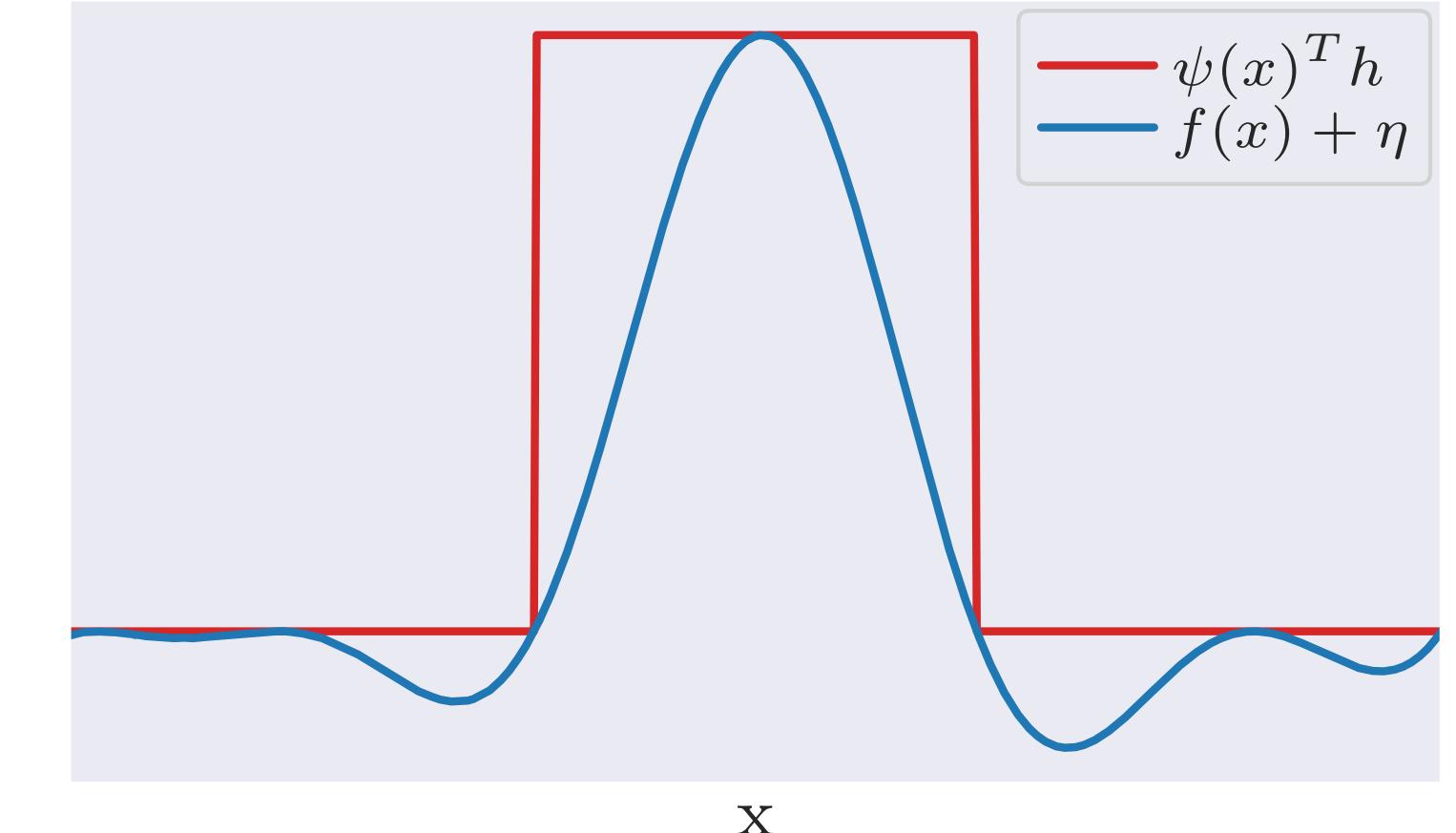
s.t.  $f_0 + f(x, z) \leq \psi(x; \theta)^T h(z) \quad \forall (x, z) \in \mathcal{X} \times \mathcal{Z}$ .

**Entropy regularization** Infinite constraint  $\rightarrow$  soft-constraint

$$\inf_{\theta, Q \in \mathcal{P}} \frac{1}{2} \text{MMD}^2(Q, \hat{P}) + \lambda D_\phi(Q \parallel \omega) \text{ s.t. } \mathbb{E}_Q \left[ \psi(X; \theta)^T h(Z) \right] = 0$$

results in an unconstrained dual

$$\mathbb{E}_{\hat{P}_n} [f_0 + f(X, Z)] - \frac{1}{2} \|f\|_{\mathcal{F}}^2 - \mathbb{E}_\omega \left[ \varphi_\epsilon^* (f_0 + f(X, Z) - \psi(X; \theta)^T h(Z)) \right]$$



soft cons.  $\phi_{\text{KL}}^*(t) = \exp(t)$

log-barrier  $\phi_{\text{log}}^*(t) = -\log(1-t)$

# Kernel MoM: Nonlinear Instrumental Variable Regression

$$Y := g(X; \theta_0) + \nu(U) + \epsilon_1$$

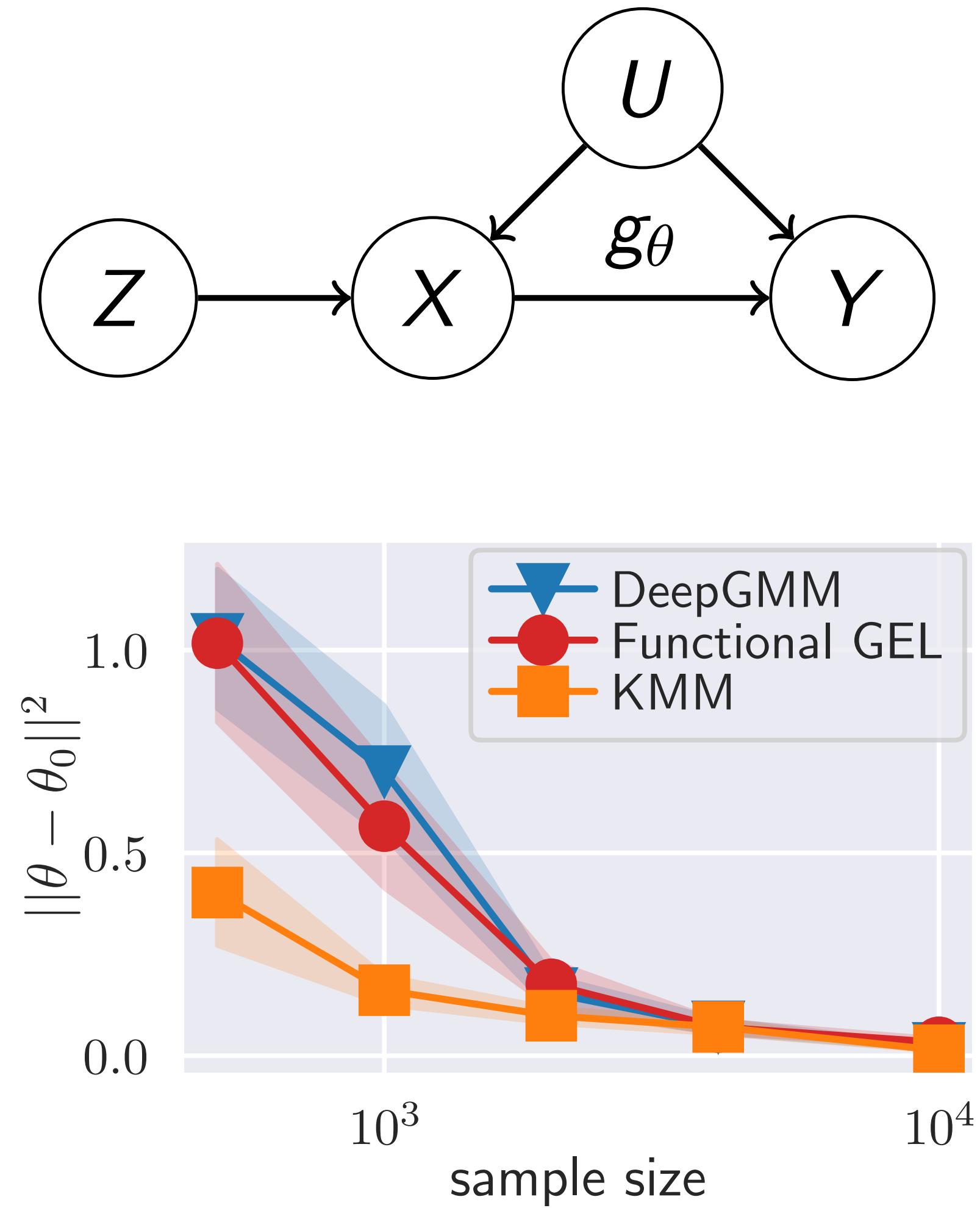
$$X := \eta(Z) + \mu(U) + \epsilon_2 ,$$

$$Z \sim P_Z, \quad \epsilon_{1/2} \sim \mathcal{N}(0, \sigma)$$

$g(x; \theta)$  is nonlinear in both  $x, \theta$ .

Estimate  $\theta$  using Kernel MoM with CMR

**Takeaway.** (Strong) Structured distribution shifts (e.g., causal confounding) can be accounted for using the Kernel MoM + CMR, but not (joint) DRO



# Force-balance of Kernel MoM

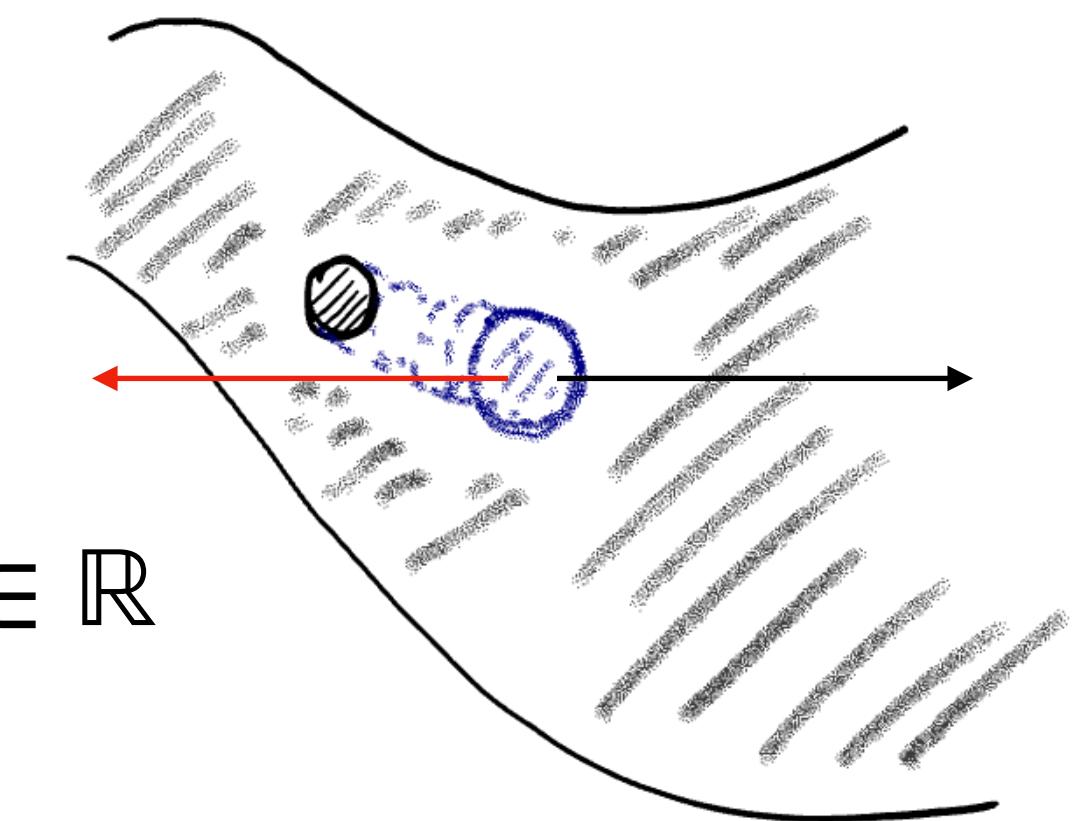
Lagrangian:

$$\sup_{\gamma \in \mathbb{R}, h \in \mathcal{H}} \inf_{\mathcal{Q}} \frac{1}{2} \text{MMD}^2(\mathcal{Q}, \hat{\mathcal{P}}) + \gamma \cdot \mathbb{E}_{\mathcal{Q}} \left[ (Y - g_{\theta}(X))^T h(Z) \right]$$

Minimizing movement scheme (MMS) in MMD  $\inf_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\gamma} \text{MMD}^2(\mu, \mu^k)$

Force balance using **function approximation**, e.g., kernel functions

$$-DF = f + f_0, \quad f = \frac{1}{\tau} \sum_{i=1}^n \alpha_i k([x_i, y_i, z_i], \cdot) \in \mathcal{H}, f_0 \in \mathbb{R}$$

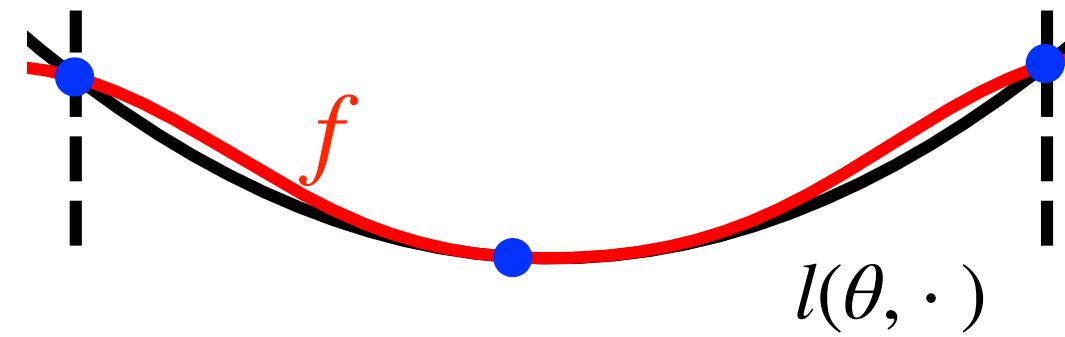


Since  $DF = (Y - g_{\theta}(X))^T h(Z)$ , the optimal force function approximates the moment function

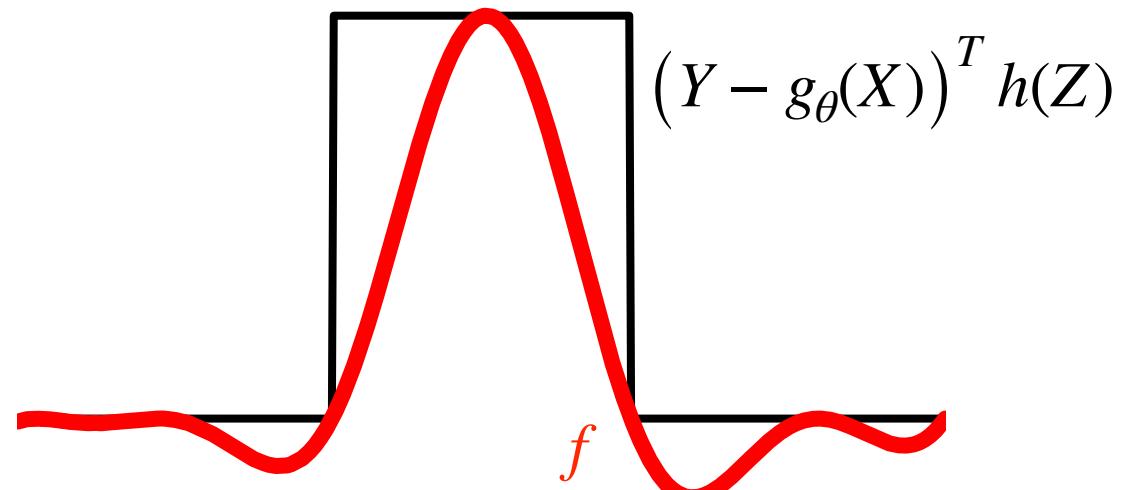
$$f + f_0 = (Y - g_{\theta}(X))^T h(Z) \text{ a.e.}$$

# Summary

- We exploited explicitly parametrized **dual force functions** for **robust learning under joint and structured distribution shifts**. This is inspired by **generalized force** in gradient flows, optimal transport, and mechanics.
- The gradient flow force-balance eqns give insights for constructing robust learning algorithms.
  - **Kernel DRO**: force gives the robustified surrogate loss



- **Kernel MoM**: force gives the robustified moment condition



This talk is mainly based on:

1. Z., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. AISTATS 2021
2. Kremer, H., Nemmour, Y., Schölkopf, B. & Z. Estimation Beyond Data Reweighting: Kernel Method of Moments. ICML 2023

Slides available: [jj-zhu.github.io](http://jj-zhu.github.io)



Postdoc position opening in Berlin: data-driven dynamics modeling for medical imaging. Contact for info.