From Gradient Flow Force-Balance to Robust Machine Learning

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics

Berlin, Germany

Basque Center for Applied Mathematics: BCAM
Bilbao, Basque Country, Spain. October 31st, 2023

Weierstraß-Institut für
Angewandte Analysis und Stochastik

Big picture:
measure optimization

Motivation: Langevin Monte-Carlo

Inference as measure optimization

Given density up to a constant $\pi(x) \propto \exp(-V(x))$

Generate samples from π (or estimate $\mathbb{E}_\pi \psi(X)$ for some ψ)

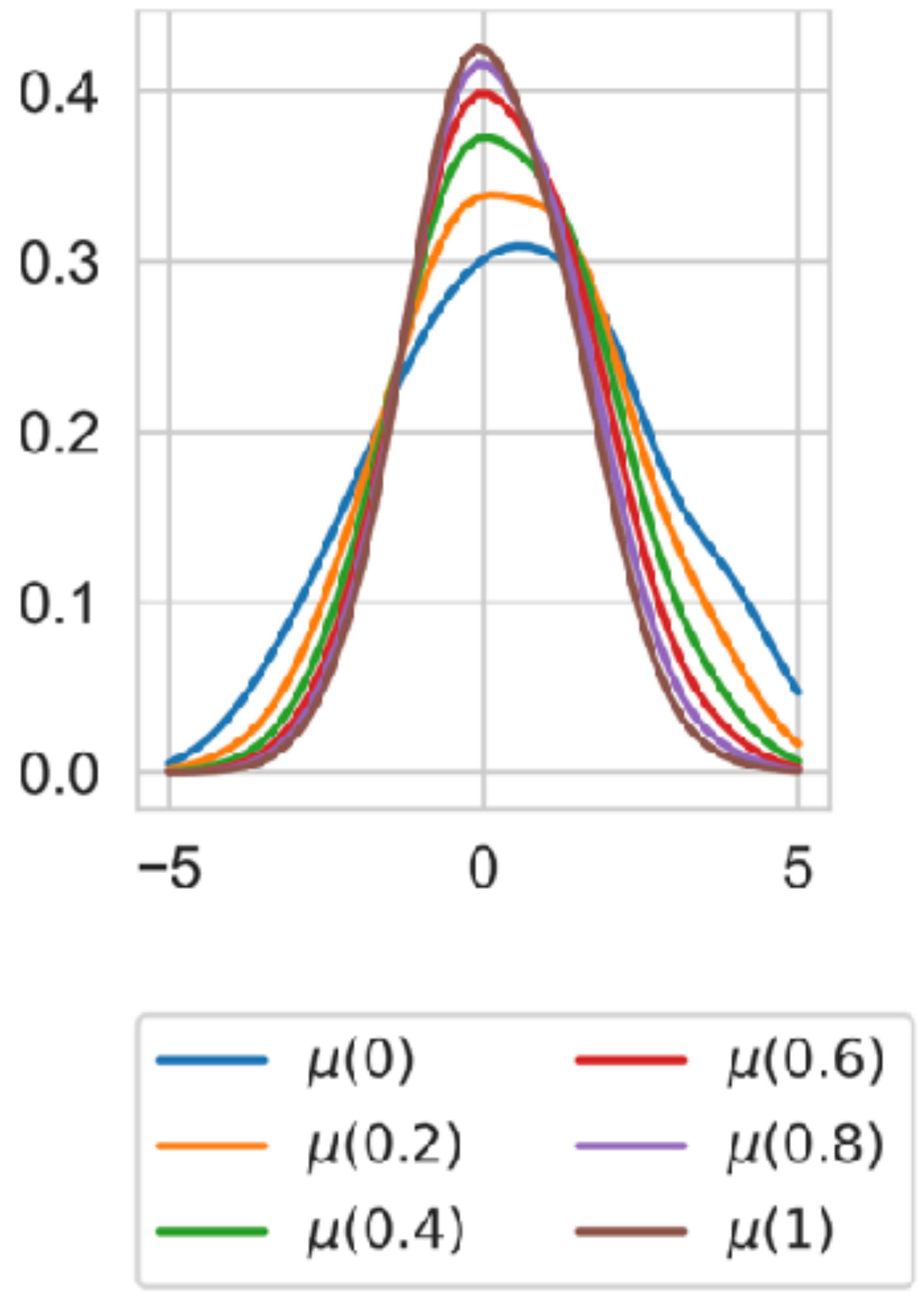
$$\inf_{\mu \in \mathcal{M}} \mathcal{D}_{\mathrm{KL}}(\mu \| \pi).$$

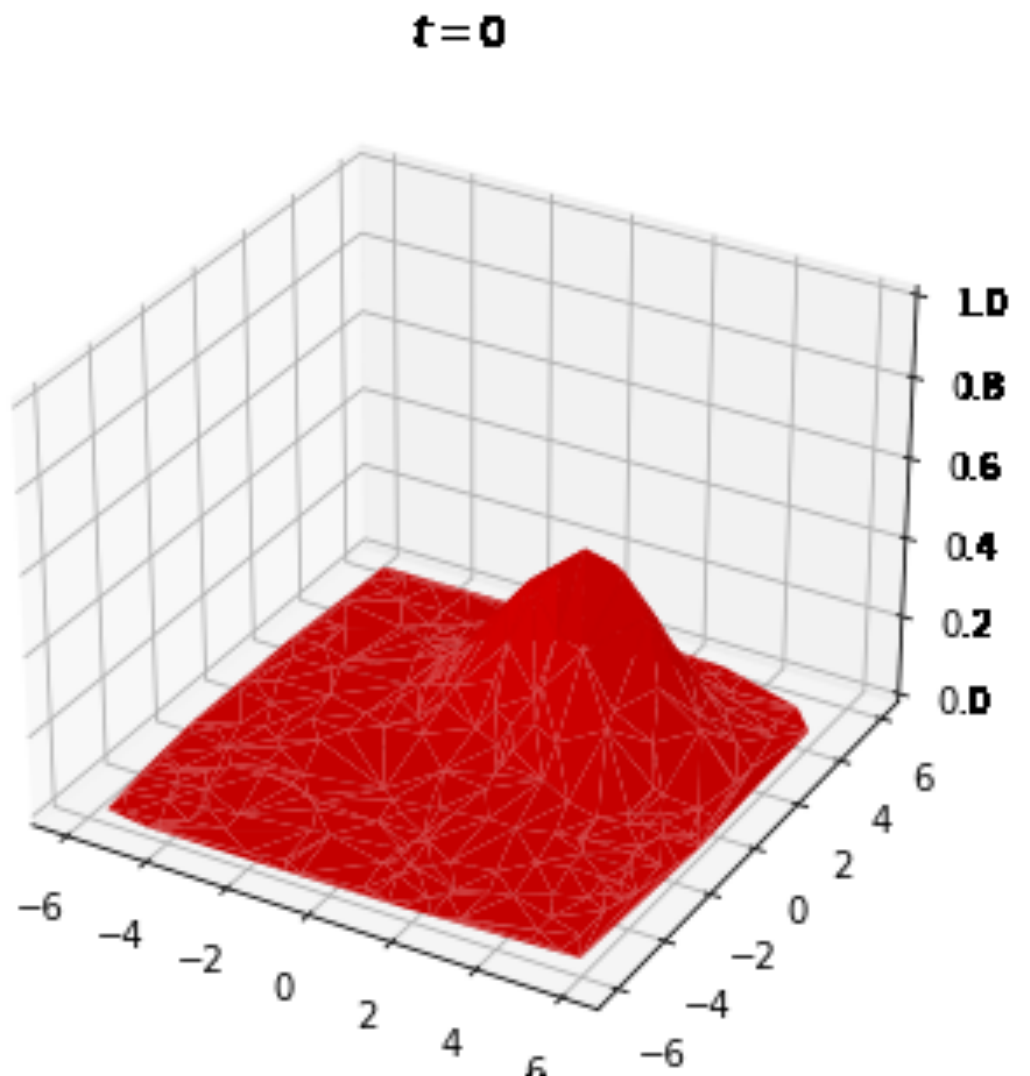
Monte-Carlo Sampling via Langevin SDE

$$X_{k+1} = X_k - \nabla V(X_k) \cdot \tau + \sqrt{2\tau} \Delta Z_k$$

where $\Delta Z_k \sim N(0,1)$, τ is the step size. The state distribution $X_T \sim \mu_T$ converges to π.

This **<u>stochastic</u>** dynamics is **equivalent** to the **<u>deterministic</u> PDE gradient flow in the Wasserstein space** [Otto 96].





Motivation: From statistical learning to distrib. 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

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

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

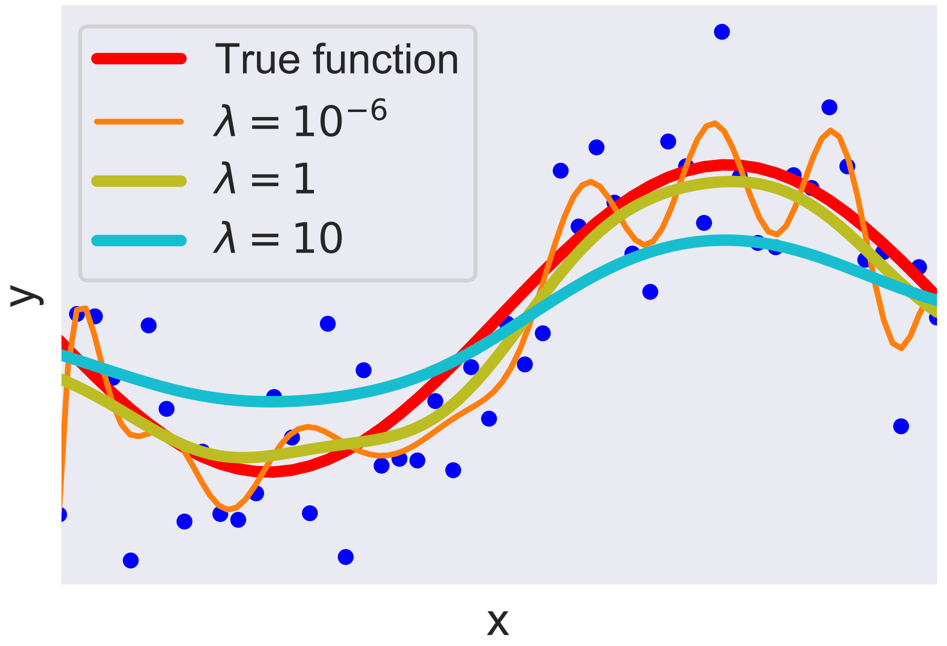
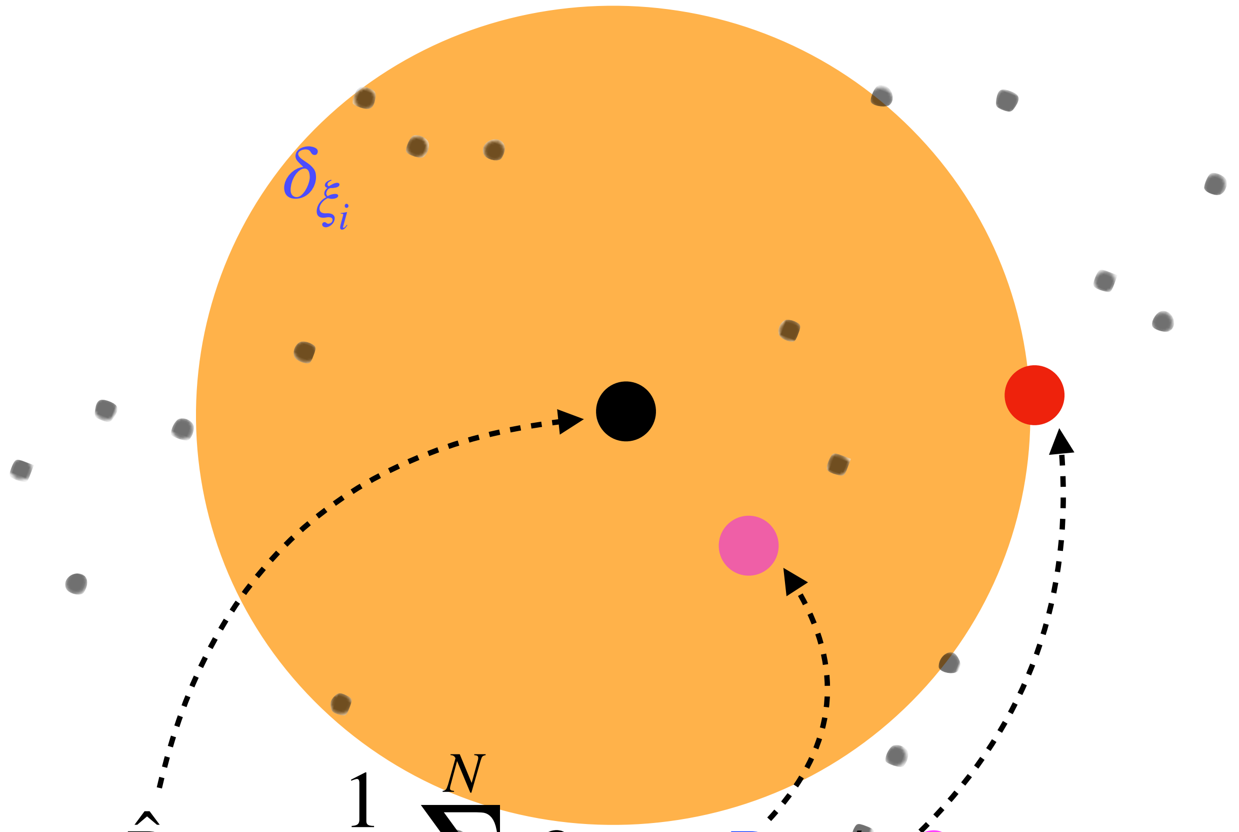


Figure credit: H. Kremer, J. Zhu

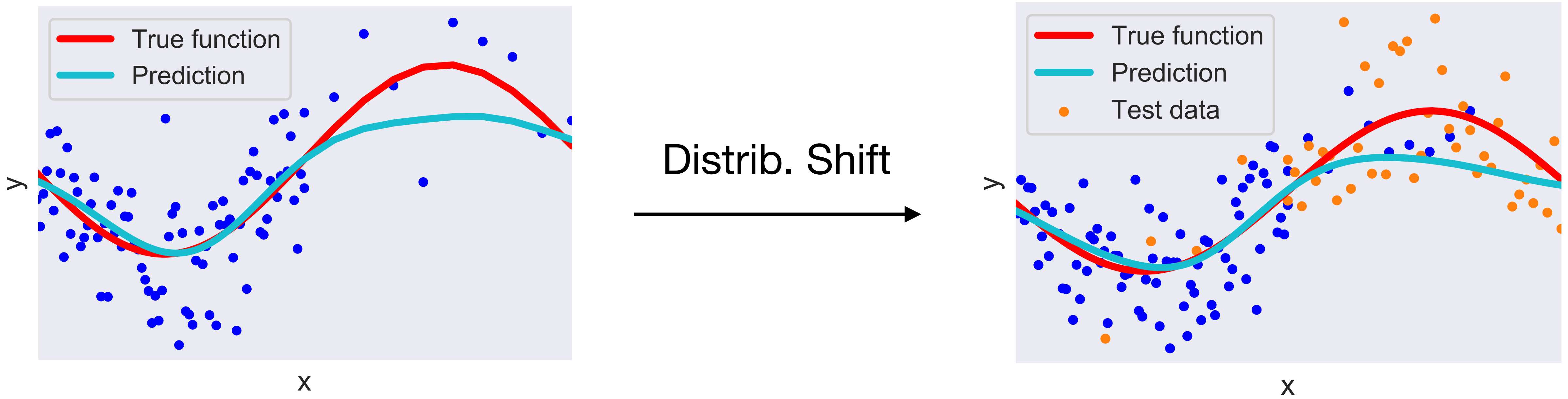


$$\hat{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

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.

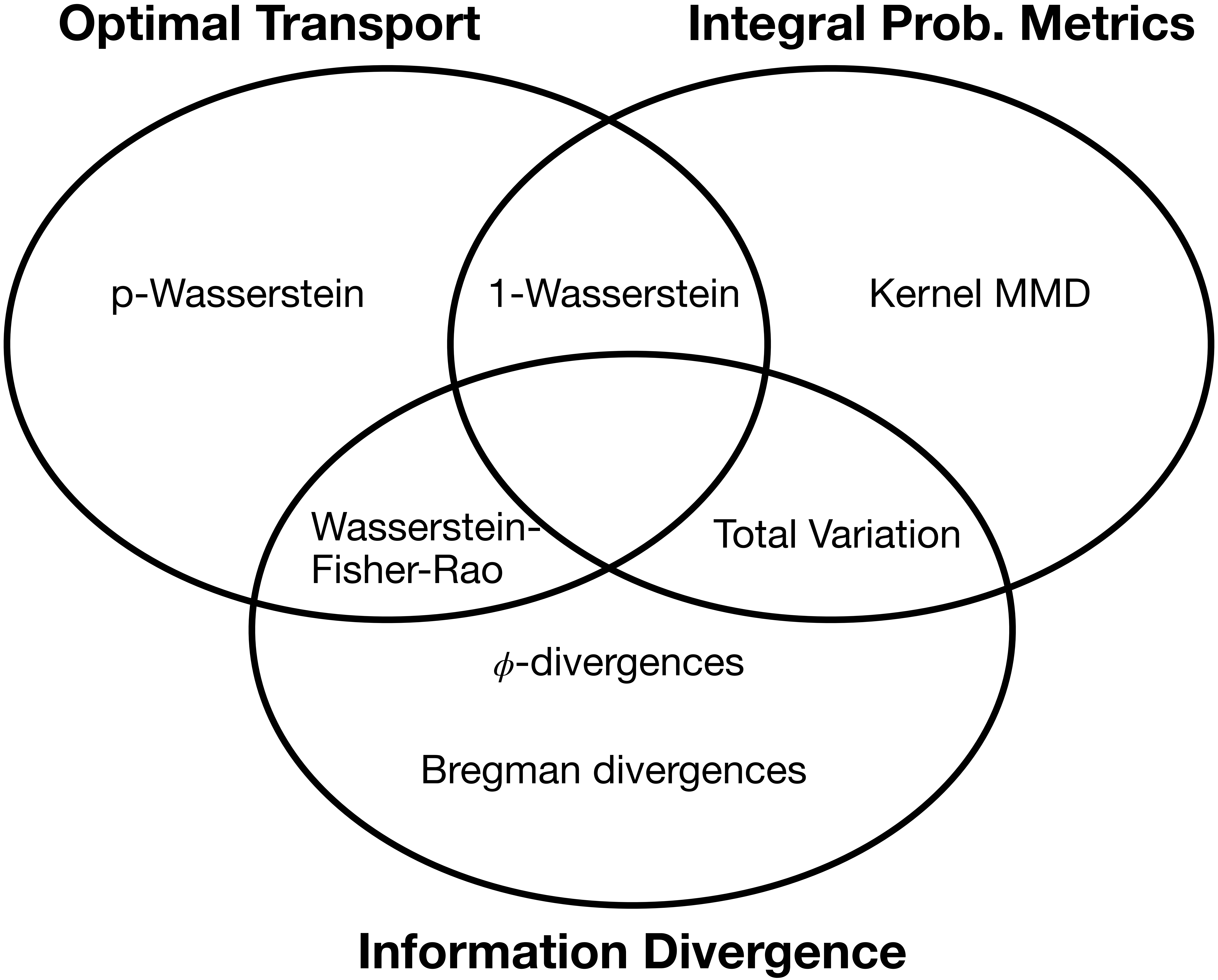


Why study new geometry?

New geometries leading to new principled fields of research and breakthroughs for computation

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



Optimal Transport

Integral Prob. Metrics

p-Wasserstein

1-Wasserstein

Kernel MMD

Wasserstein-Fisher-Rao

Total Variation

ϕ-divergences

Bregman divergences

Information Divergence

Figure credit: J. Zhu

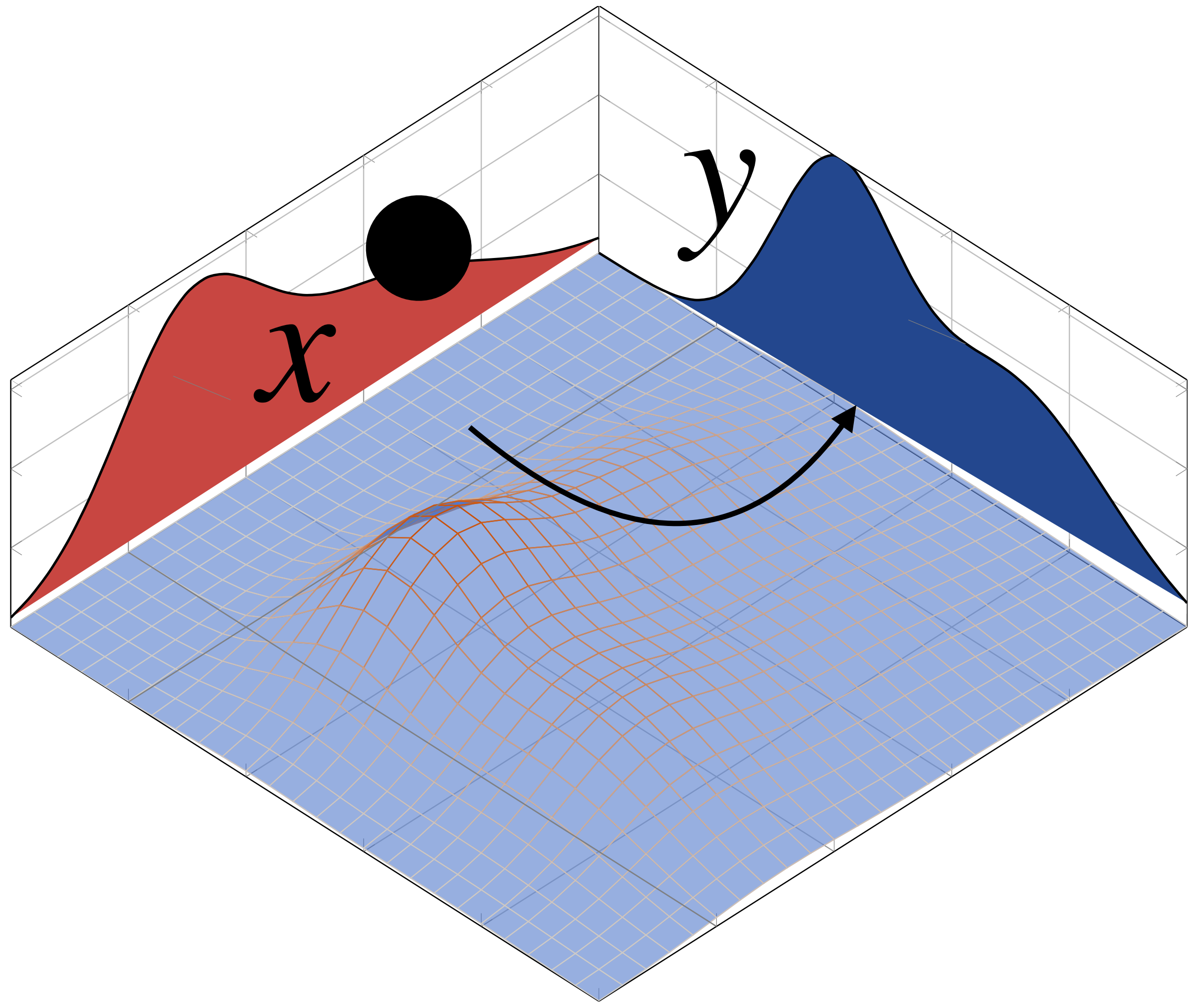
Kantorovich-Wasserstein geometry

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

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

(**Dual Kantorovich problem**)

$$= \sup \left\{ \int \psi_1(x)\,\mathrm{d}P(x) + \int \psi_2(y)\,\mathrm{d}Q(y) \,\middle|\, \psi_1(x) + \psi_2(y) \leq \; x - y \; ^p \right\}$$



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

Dynamic formulation: *à la Benamou-Brenier*

$$W_2^2(P, Q) = \min \left\{ \int_0^1 \int \; v_t \; ^2 \mathrm{d}\mu_t \mathrm{d}t \,\middle|\, \mu_0 = P, \mu_1 = Q, \partial_t \mu_t + \mathrm{div}(v_t \mu_t) = 0 \right\}$$

Kernel maximum-mean discrepancy

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

$$\mathrm{MMD}(P,Q) := \left\| \int k(x,\cdot\,)dP - \int k(x,\cdot\,)dQ \right\|_{\mathscr{H}}.$$

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

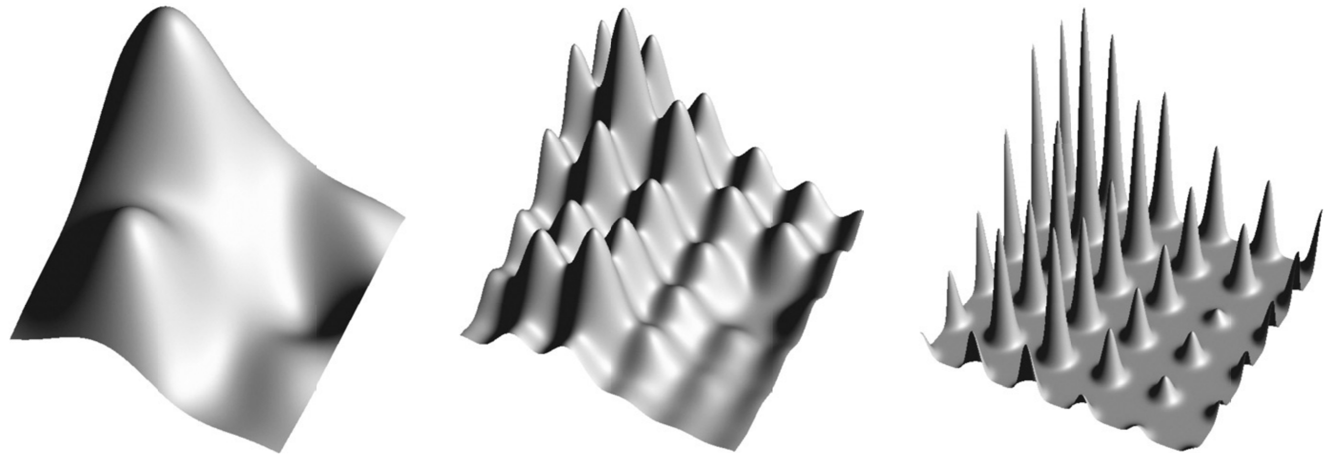
Dual formulation as an integral probability metric.

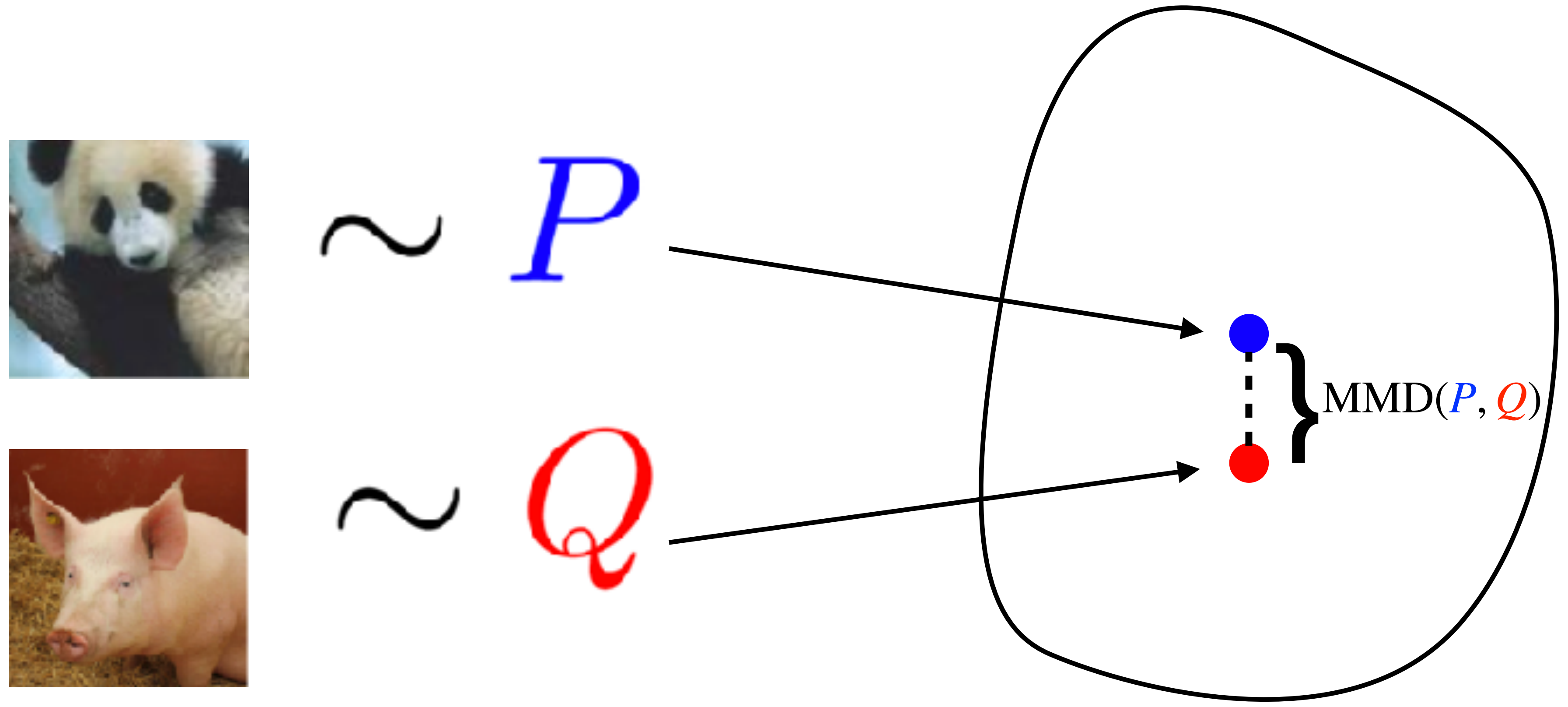
$$\mathrm{MMD}(P,Q) = \sup_{\|f\|_{\mathscr{H}} \le 1} \int f\, d(P-Q)$$

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

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

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





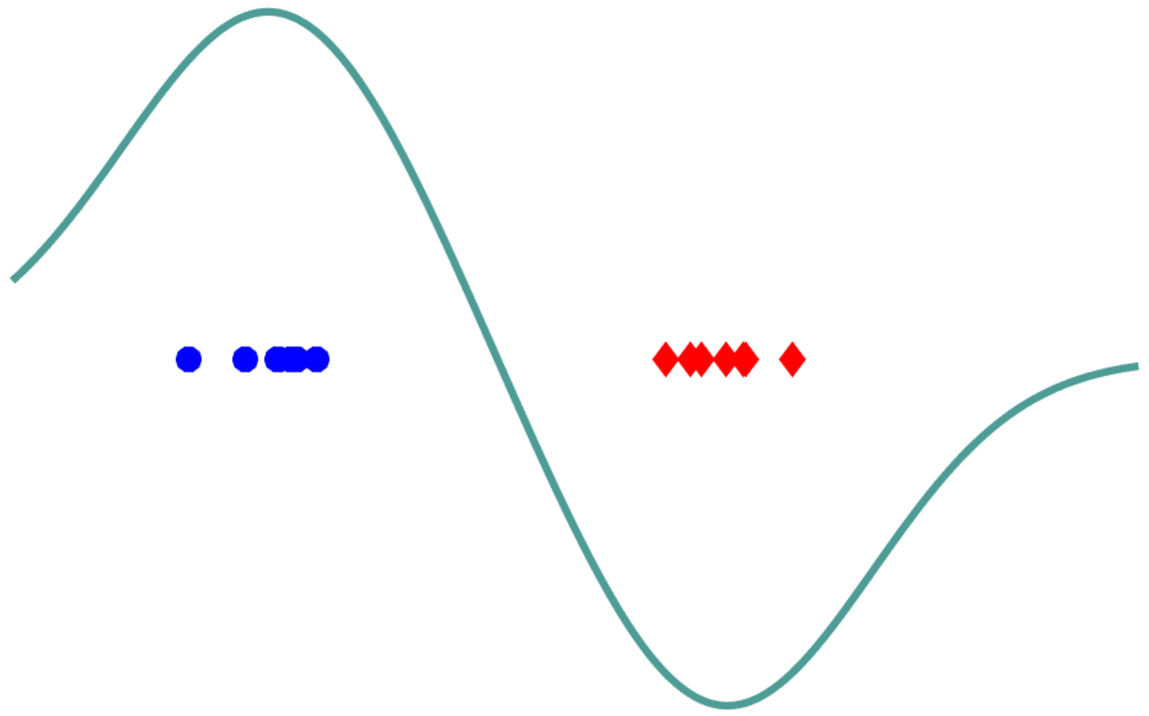


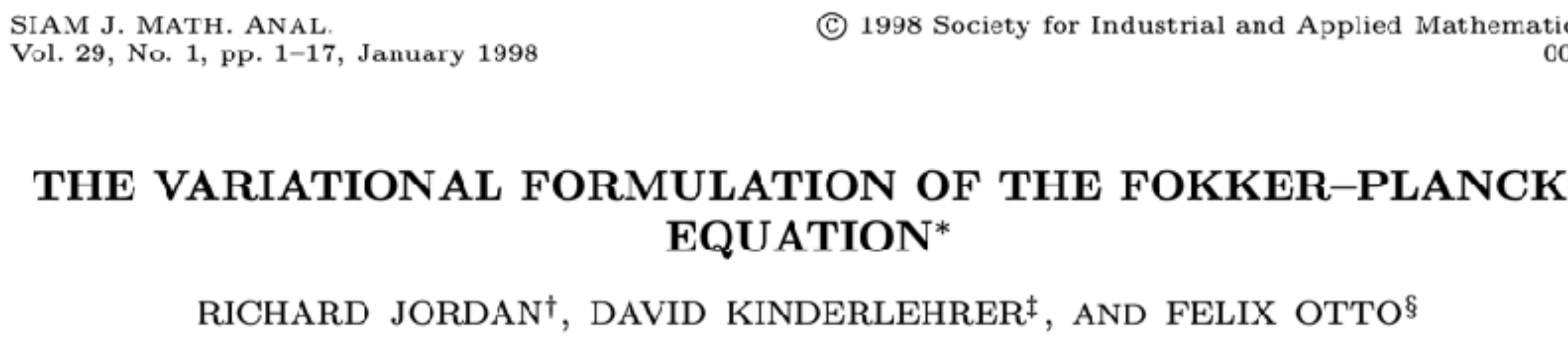
Figure credit: W. Jitkrittum, J. Zhu, H. Wendland

Gradient Flow Force-Balance

Gradient flow facts

Otto's Gradient flow equation in the Wasserstein space

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



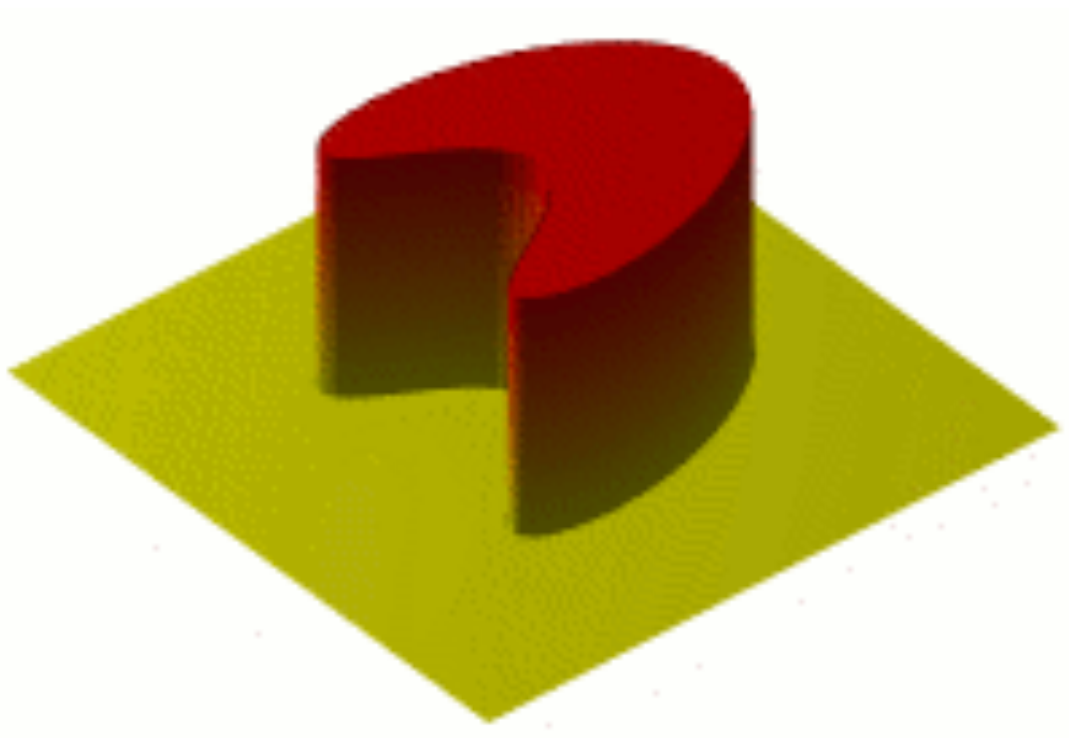
SIAM J. MATH. ANAL.
Vol. 29, No. 1, pp. 1–17, January 1998

© 1998 Society for Industrial and Applied Mathemati
00

THE VARIATIONAL FORMULATION OF THE FOKKER–PLANCK EQUATION*

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

e.g., diffusion, heat conduction, Fokker Planck equation
“steepest” dissipation of energy. [Otto et al 2000s, Ambrosio 2005, …]
The Wasserstein **gradient system** that generates the WGF is $(\mathrm{Prob}(\bar{X}), F, W_2)$



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

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

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

“JKO Scheme” $$u_k \in \arg \inf_{u \in \mathscr{P}} F(u) + \frac{1}{2\tau} W_2^2\left(u, u_{k-1}\right)$$

ODE flow: $(\mathbb{R}^d, F, \|\|_2)$
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 force-balance

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

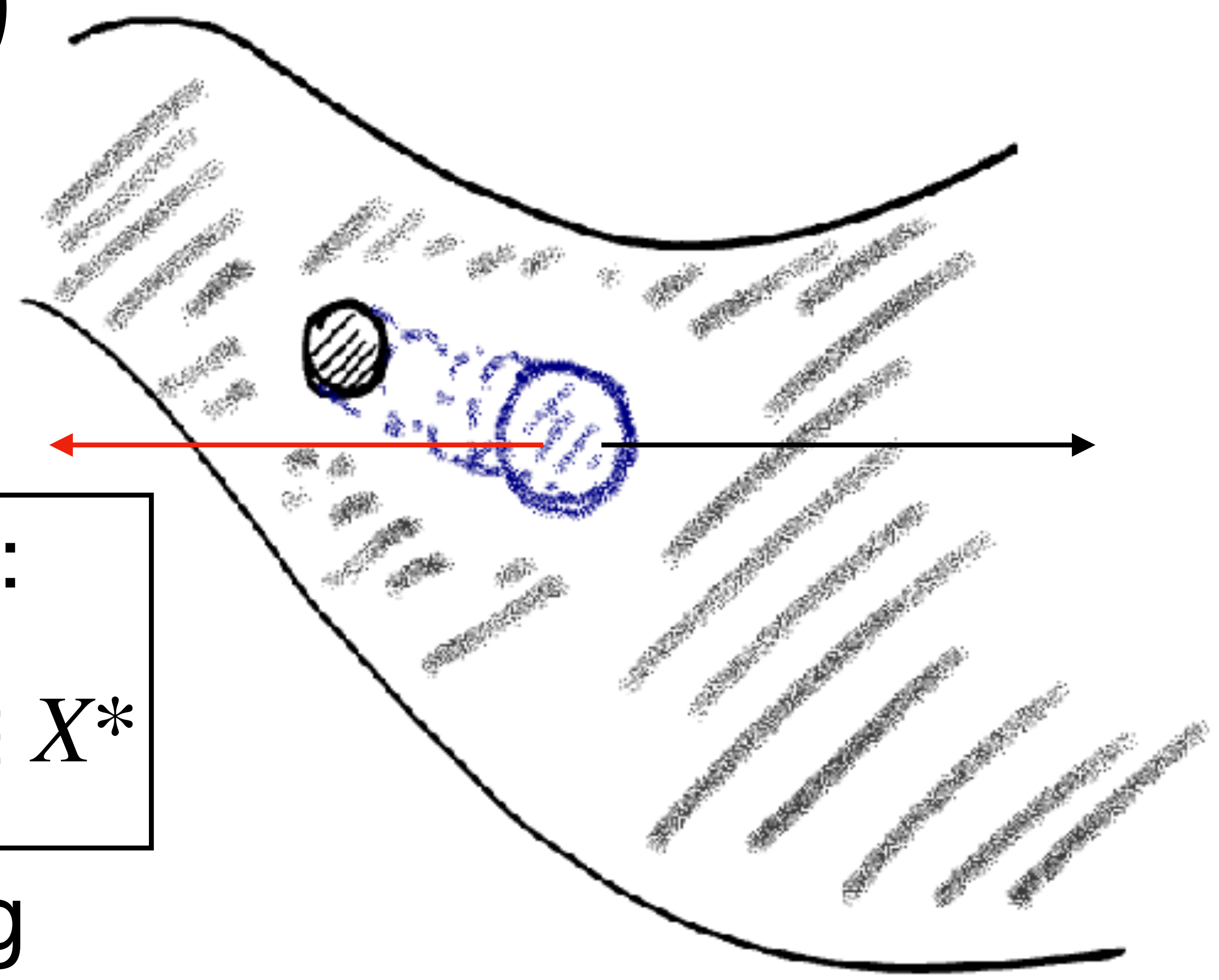
$$\mathrm{D}F + \frac{\phi}{\tau} = \mathrm{const}\,.\,,\ \phi : \text{“Kantorovich potential”}$$

Force:
drive movement
e.g., entropy

Dissipation Geometry:
resist movement
e.g., viscosity

Force-balance in ODE:

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



In practice, approximate ϕ (and hence $-\mathrm{D}F$) based on data samples using **function approximators** (force matching, score matching), NN/RKHS, e.g.,

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

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

Robust Learning

Distributional robustness, but what kind?

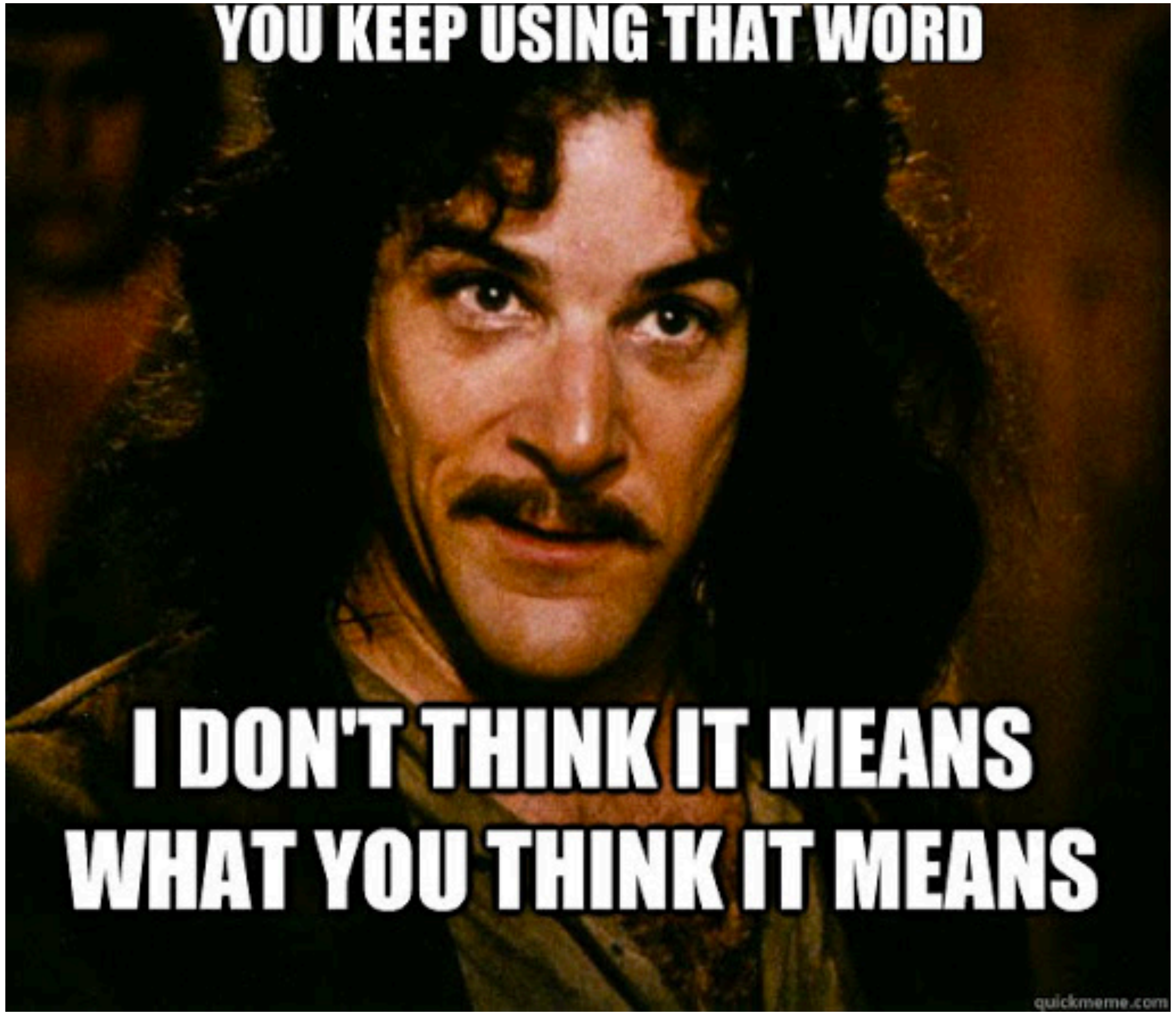


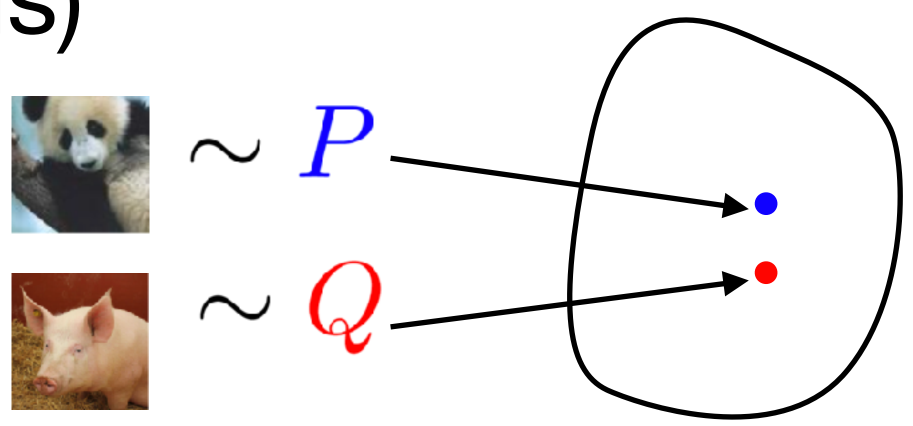
Figure credit: The Princess Bride,
a bedside story by your grandpa

Robust Learning under (Joint) Distribution Shift

Kernel DRO under distribution shift

Primal DRO (not solvable as it is)

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

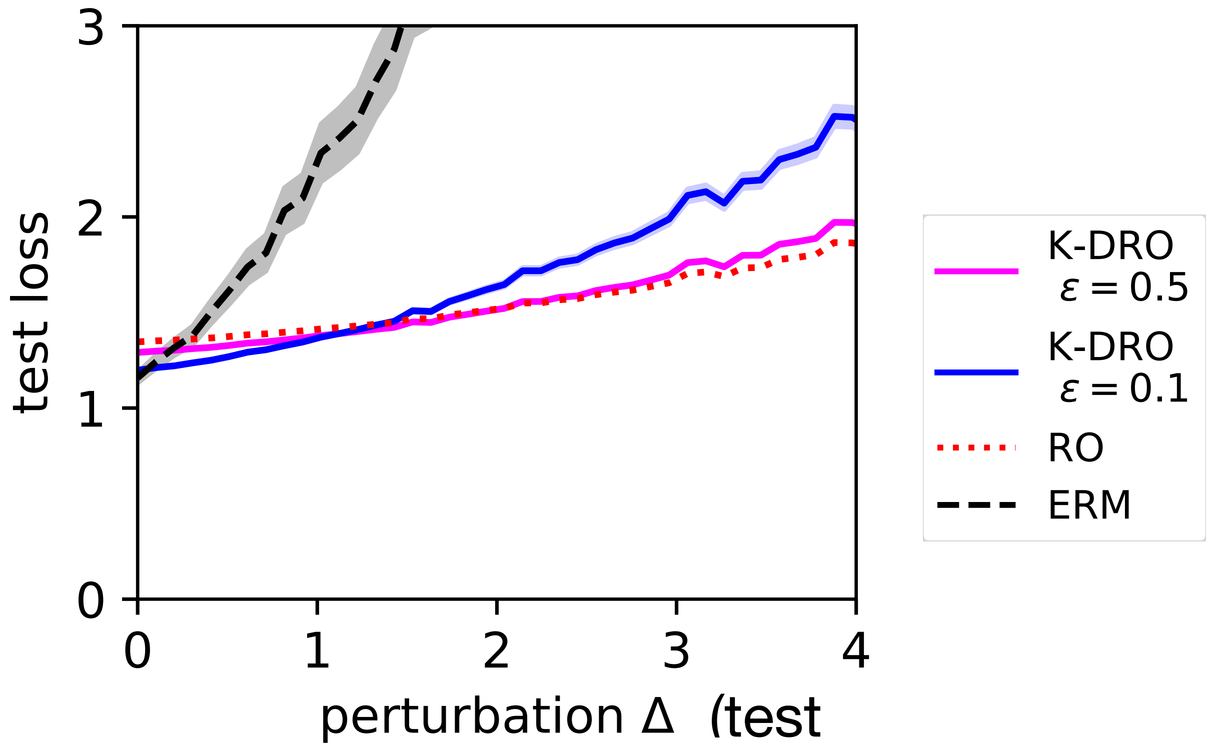


Kernel DRO Theorem (simplified). [Z. et al. 2021]
DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

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

Example. Robust least squares

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



Entropy regularization ("interior point method")

$$\mathrm{MMD}(Q,\hat{P}) + \lambda D_\phi(Q\|\omega) \le \epsilon$$

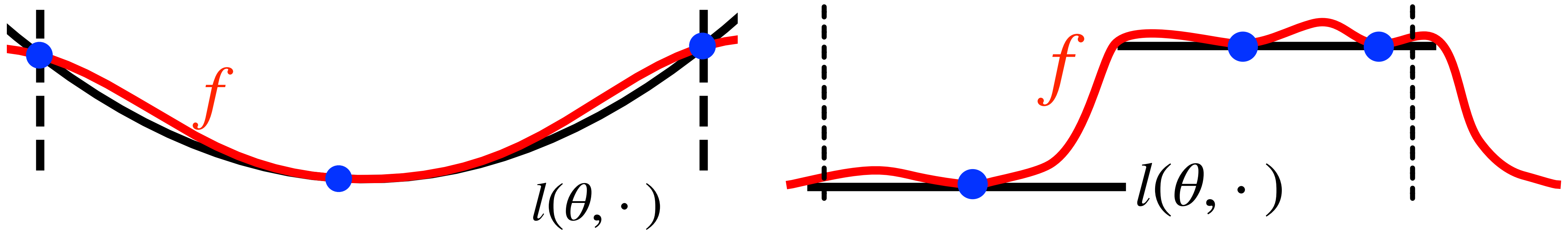
Dual. Adapted from [Kremer et al., **Z.** 2023]

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

soft cons. $\phi^*_{\mathrm{KL}}(t) = \exp(t)$

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

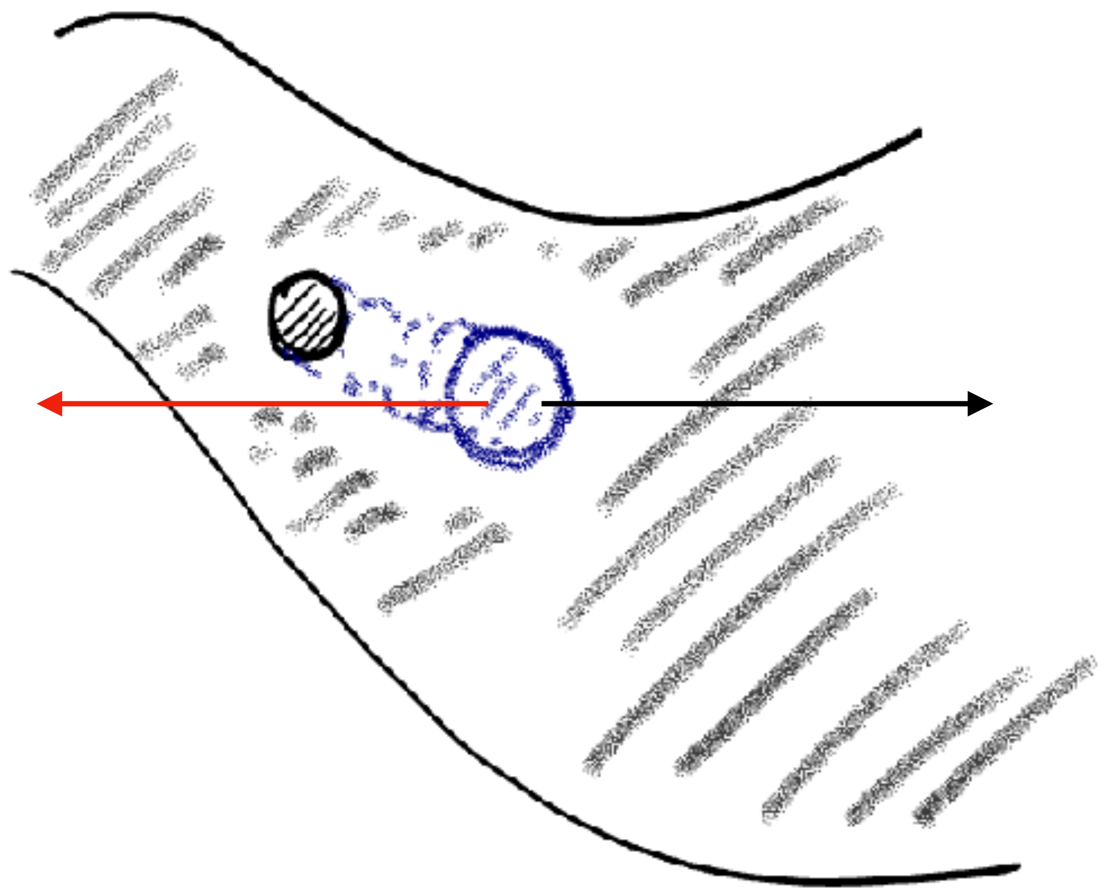
Geometric intuition: **dual kernel function f** as robust surrogate losses (<u>flatten the curve</u>)



Force-balance of Kernel DRO

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

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



MMS in kernel-MMD

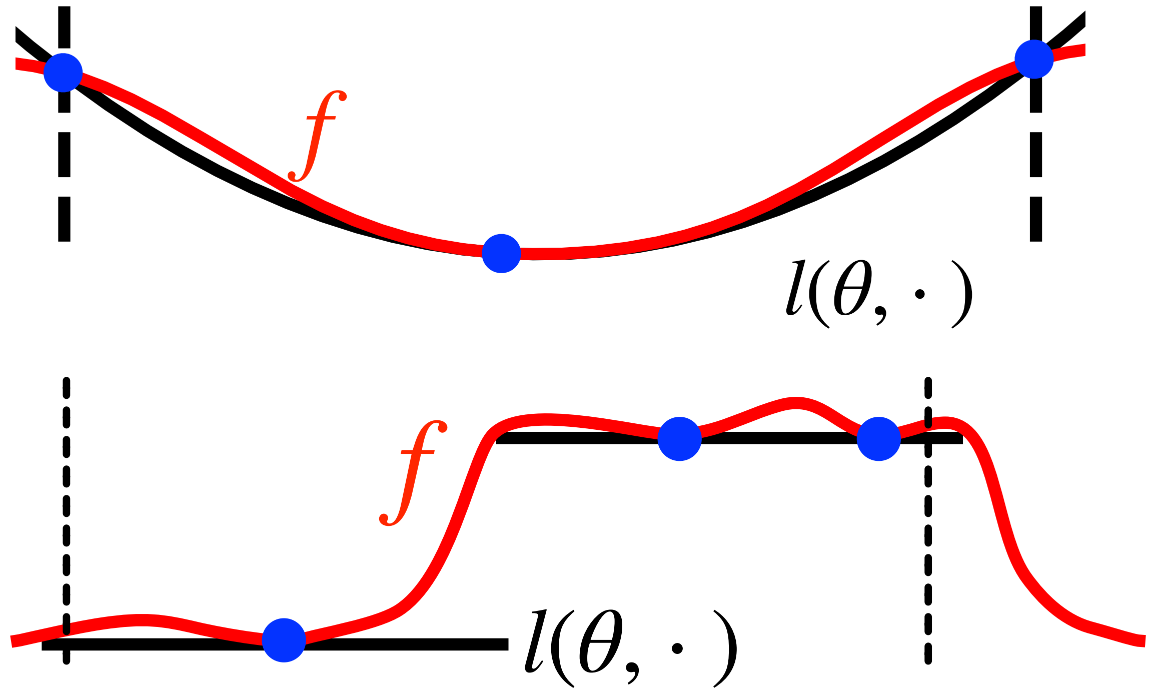
$$\inf_{\mu\in\mathscr{P}} F(\mu)+\frac{1}{2\tau}\mathrm{MMD}^2(\mu,\mu^k) \implies -\mathrm{D}^{L^2}F = \overset{=:\, f\in\mathscr{H}}{\boxed{\frac{1}{\tau}\int k(x,\cdot)\mathrm{d}(\mu-\mu^k)(x)}} + \mathrm{const}\,.$$

Dual kernel function f as robust surrogate losses

<u>flatten the curve → force balance</u>

Force-balance using **function approximation** RKHS functions, e.g.,

$$-\mathrm{D}F = f + f_0,\ f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot) \in \mathscr{H},\ f_0\in\mathbb{R}$$



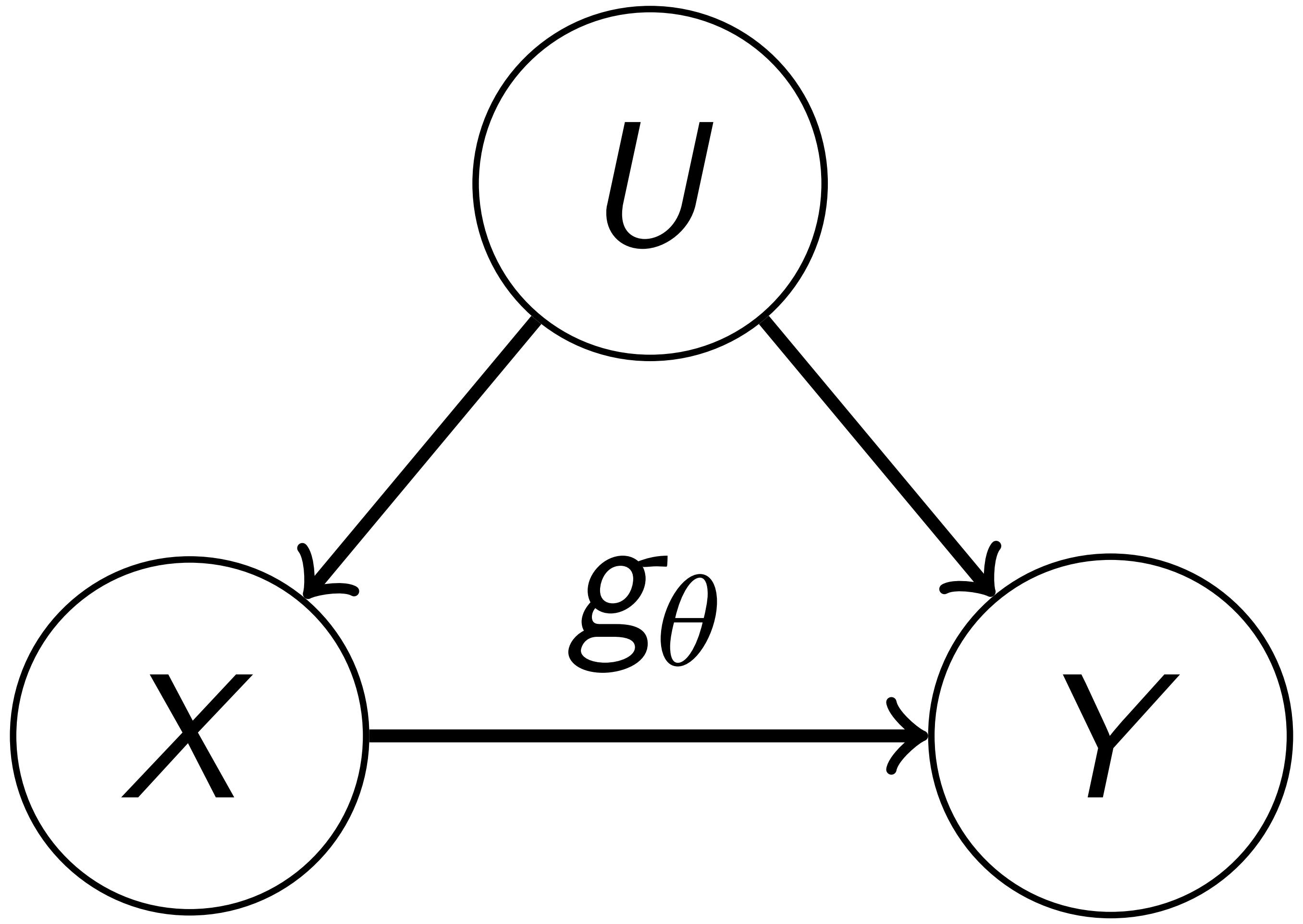
$\mathrm{D}^{L^2}F = l(\theta,\cdot) \implies$ force-balance relation: $l(\theta,\cdot) = f + f_0$ a.e.

(force matching, score matching)

Robust Learning under Structured Distribution Shift

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

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

$$\implies \quad g_\theta(x) \neq \mathbb{E}[Y \quad X = x]$$

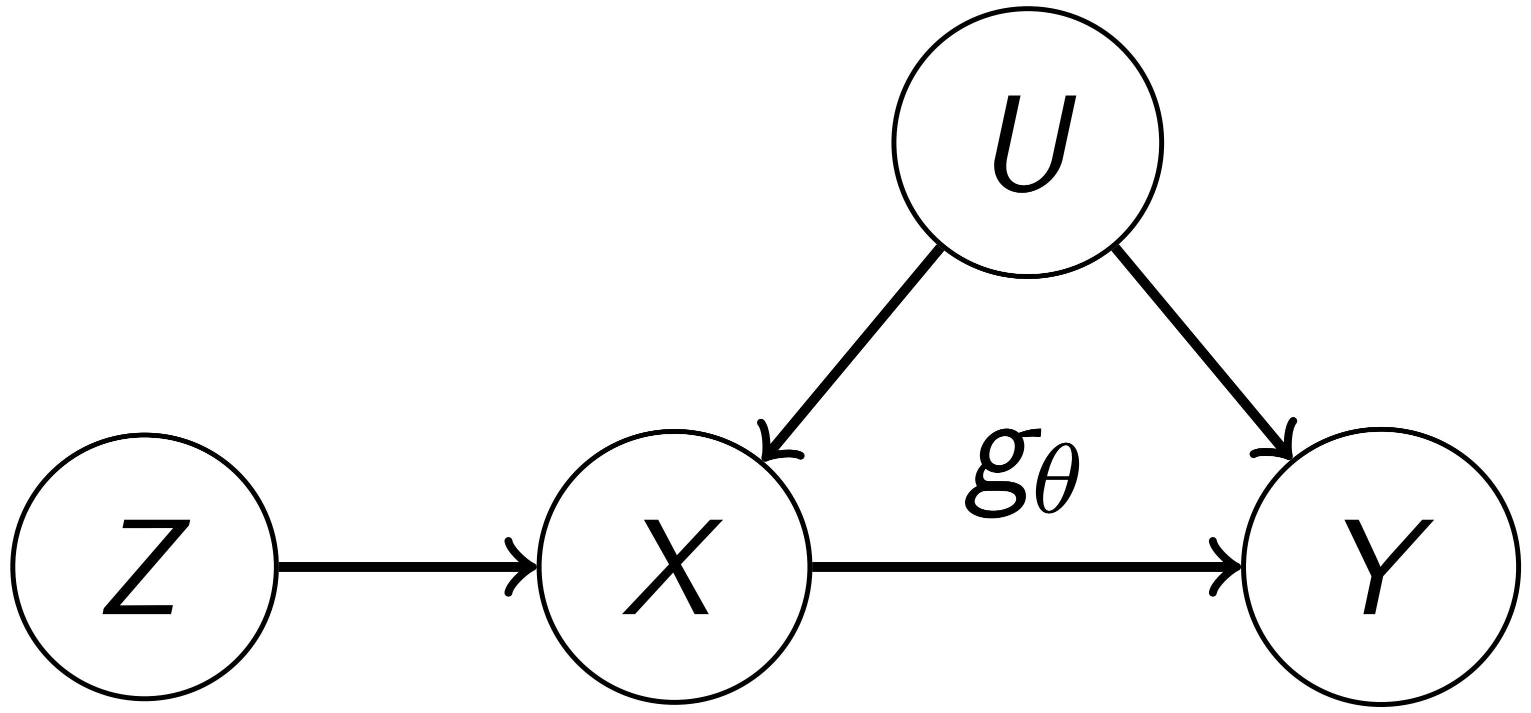
Mean regression $\min_\theta \mathbb{E}[\|Y - g_\theta(X)\|^2]$ and

(distributionally) robust optimization 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_θ via **conditional moment restriction (CMR)**

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



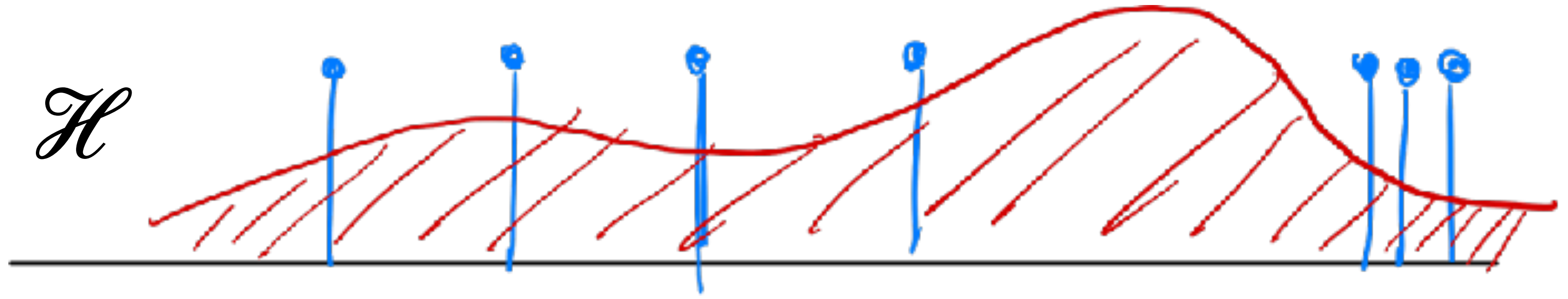
Instrument: Genetic predisposition for nicotine addiction Z

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

$$\inf_{\theta, Q \in \mathscr{P}} D_\phi(Q \| \hat{P}) \ \text{ s.t. } \ \mathbb{E}_Q[Y - g_\theta(X) \quad Z] = 0 \ \mathbb{P}_Z\text{-a.s.}$$

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

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



Lift the restriction that Q is an atomic distribution

Kernel MoM: duality and algorithm

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

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

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

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

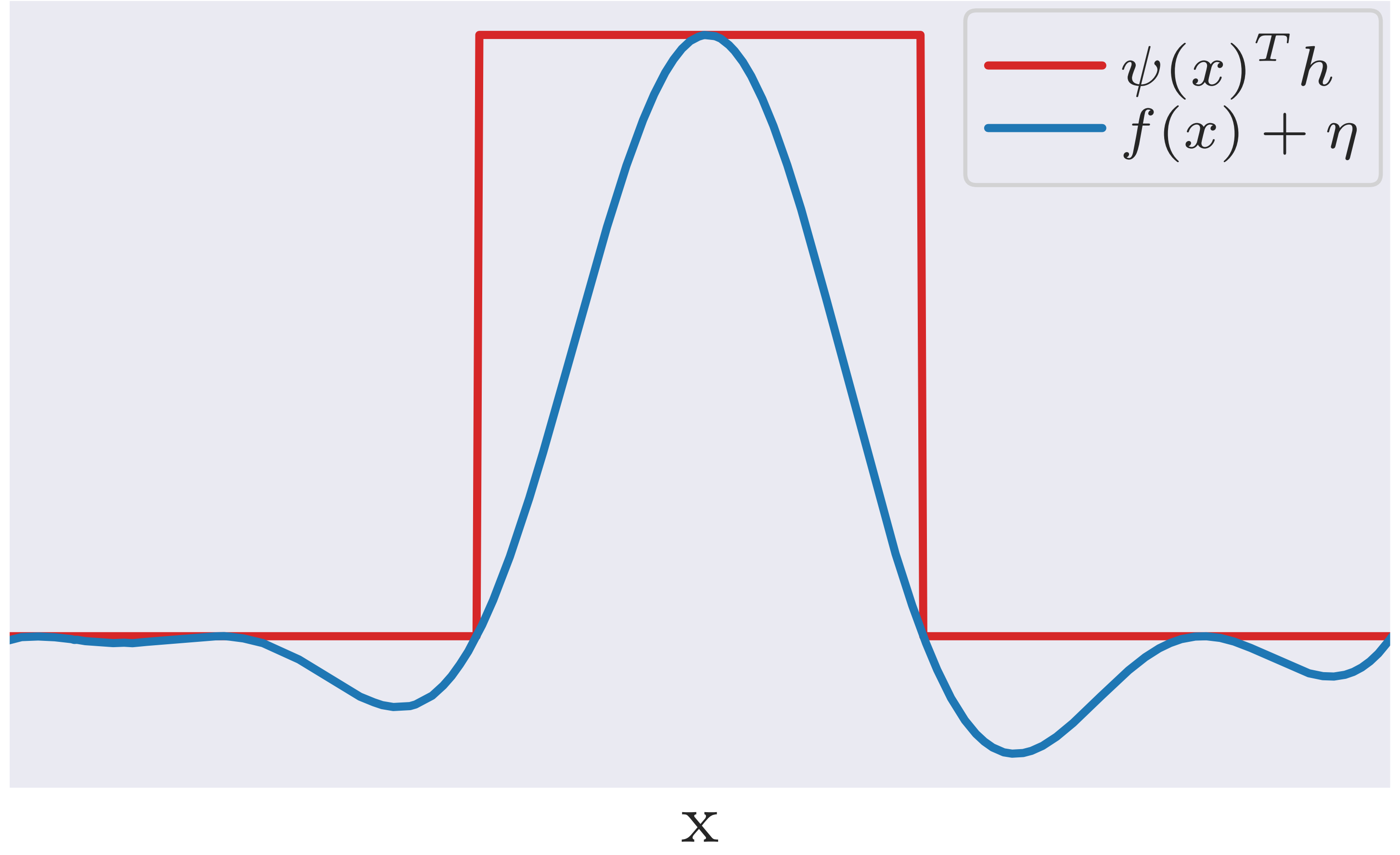
$$\text{s.t.} \quad f_0 + f(x,z) \leq \psi(x;\theta)^T h(z) \quad \forall (x,z) \in \mathcal{X} \times \mathcal{Z}\,.$$

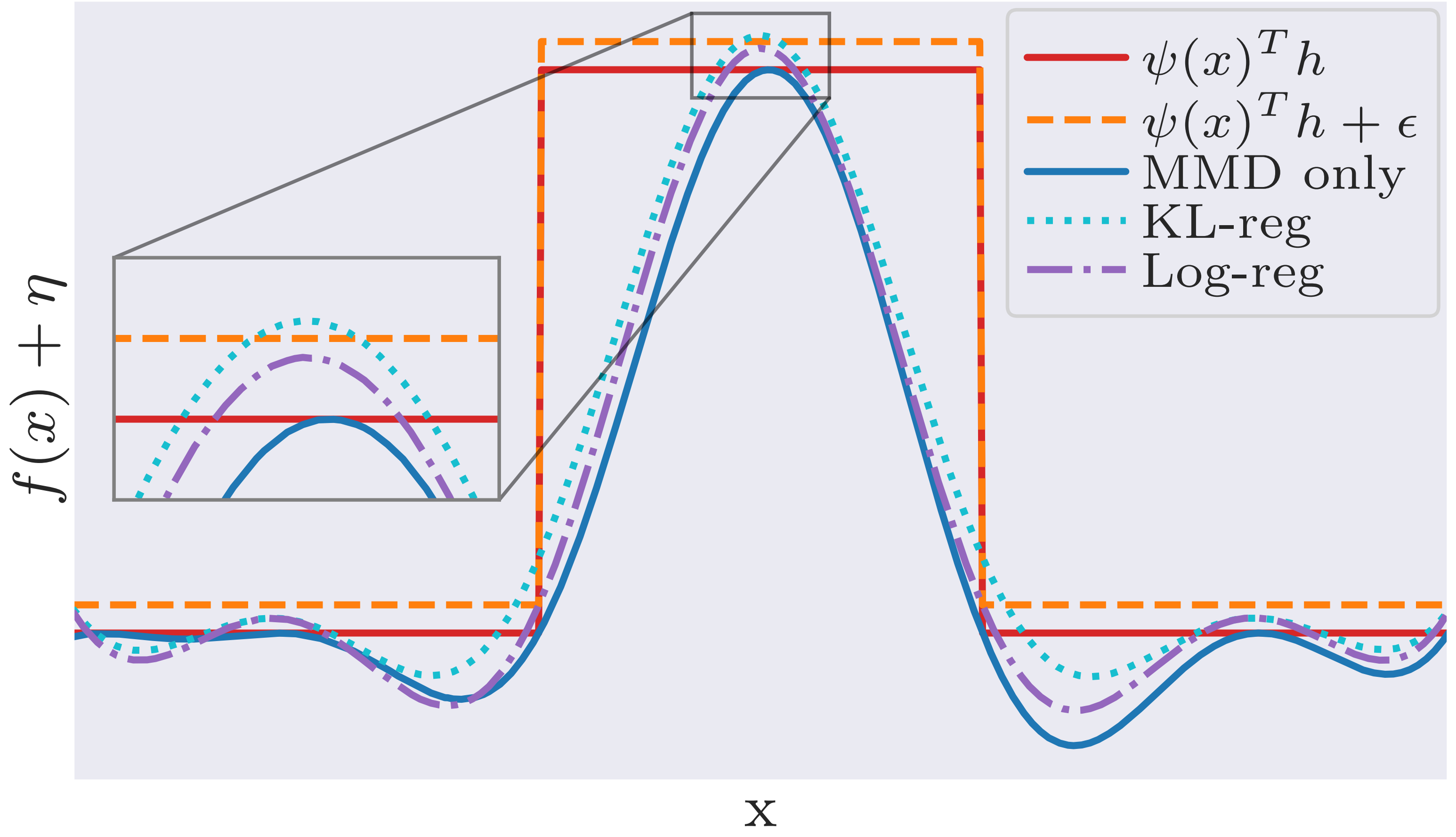
Entropy regularization Infinite constraint → soft-constraint

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

results in an unconstrained dual

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





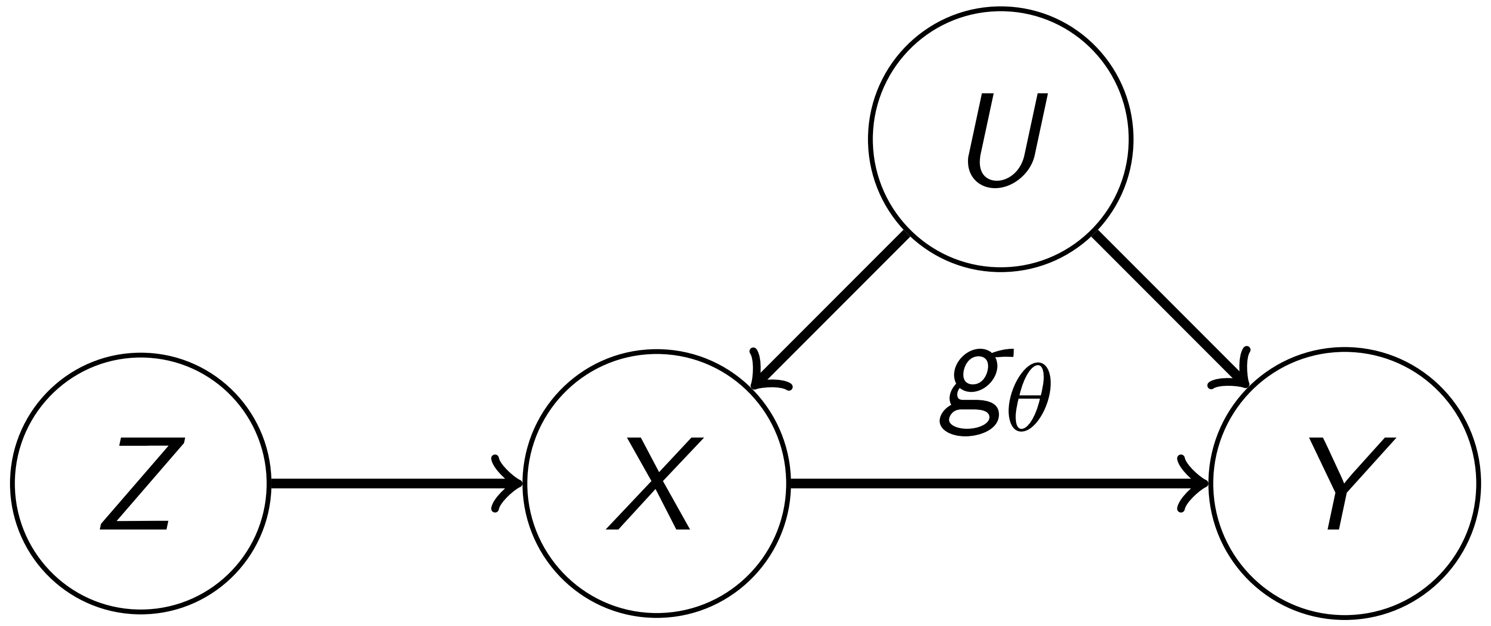
soft cons. $\phi_{\mathrm{KL}}^*(t) = \exp(t)$

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

Kernel MoM: Nonlinear Instrumental Variable Regression

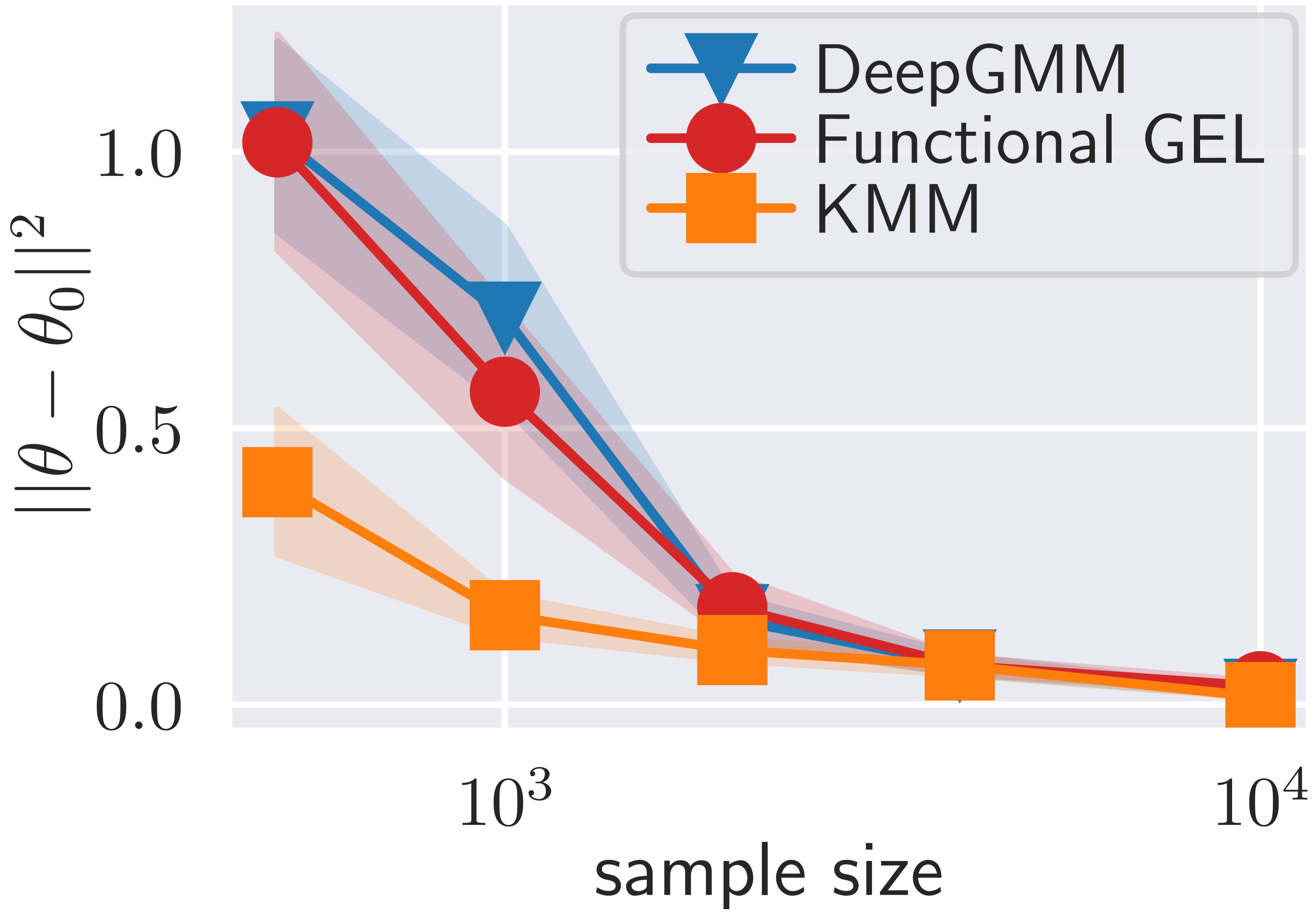
$$Y := g(X;\theta_0) + \nu(U) + \epsilon_1$$
$$X := \eta(Z) + \mu(U) + \epsilon_2 \quad ,$$
$$Z \sim P_Z, \quad \epsilon_{1/2} \sim \mathcal{N}(0,\sigma)$$

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



Estimate θ using Kernel MoM with CMR

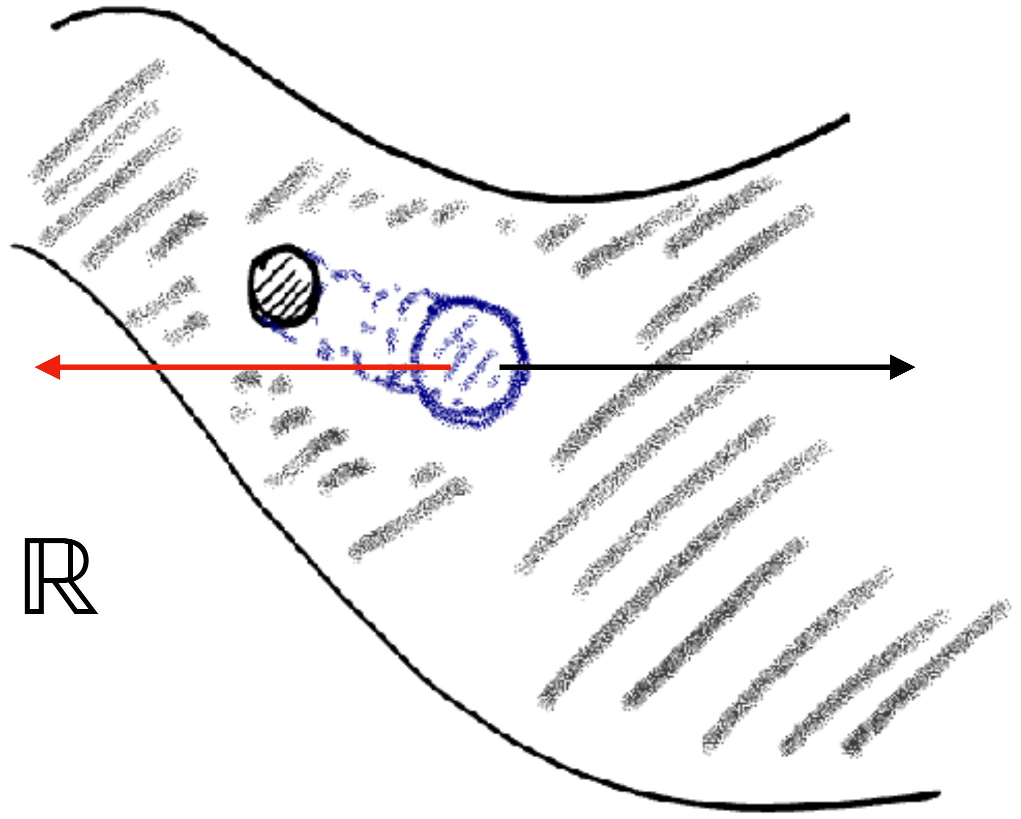
Takeaway. (Strong) structured distribution shifts (e.g., causal confounding) can be accounted for using the Kernel MoM + CMR, but not (joint) DRO, adversarial robustness, …



Force-balance of Kernel MoM

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

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



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

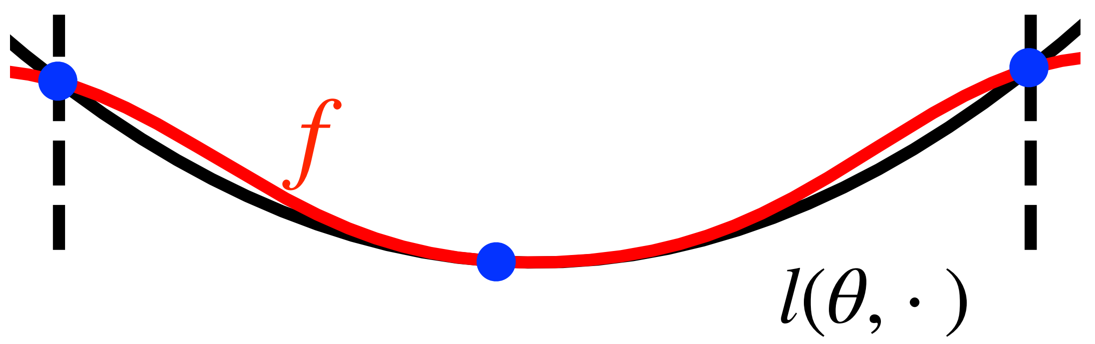
$$-\mathrm{D}F = f + f_0,\quad f=\frac{1}{\tau}\sum_{i=1}^{n}\alpha_i k([x_i,y_i,z_i],\cdot\,)\in\mathcal{H}, f_0\in\mathbb{R}$$

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

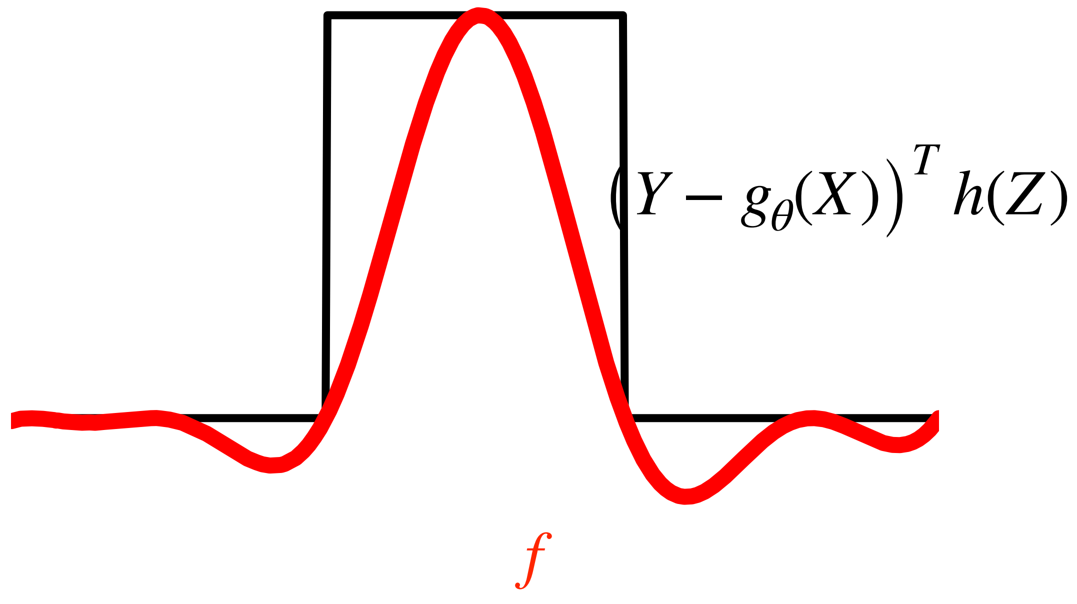
$$f+f_0 = \big(Y-g_\theta(X)\big)^T h(Z) \text{ a.e.}$$

Summary

- We exploited **explicitly parametrized dual force functions** for **robust learning** under **joint** and **structured distribution shifts**.
- 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 function



This talk is based on:

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

Website for slides, code
https://jj-zhu.github.io/

Beginner tutorial
on gradient flows

Workshop on Optimal Transport - OPT & ML
Berlin, March 2024