Optimization and Dynamics: from Euclidean Gradient Descent to Wasserstein Gradient Flow

Jia-Jie Zhu
Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

International Workshop of Intelligent Autonomous Learning Systems 2023
Kleinwalsertal, Austria. August 15th, 2023

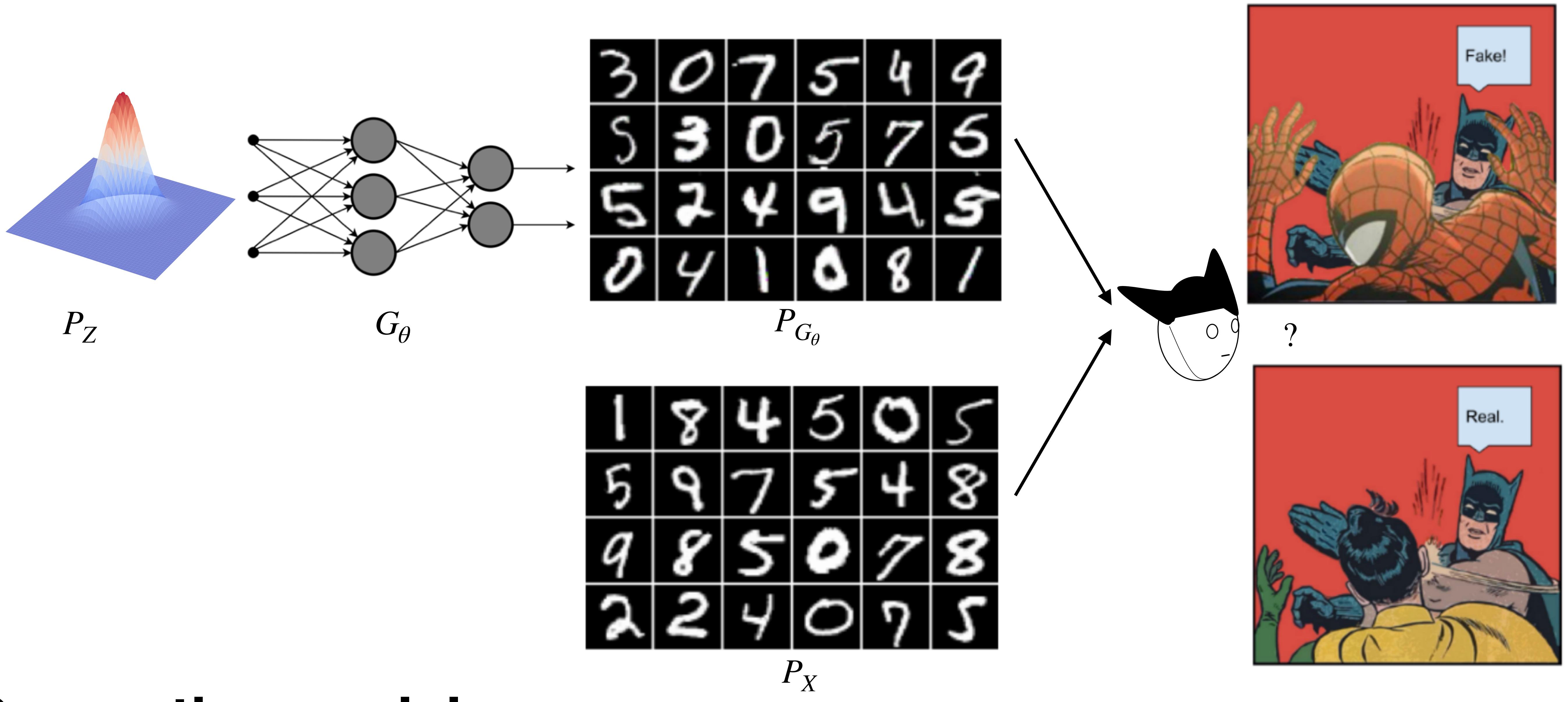
Weierstraß-Institut für
Angewandte Analysis und Stochastik

Motivation & Introduction

Kernel Methods for Robust Learning under **Distribution Shift**

Motivation: Generative Adversarial Nets

Generative models.



Generative models

$$\inf_{G_\theta} \mathscr{D}(P_{G_\theta}, P_X),$$

Here, P_{G_θ} is the distribution over the generated data $G_\theta(Z)$ where Z is sampled from a simple distribution such as $N(0,I)$.

Comparing two distributions

$\mathscr{D}$ is a (dis-)similarity measure on the *space of probability distributions*.

$$\underbrace{\mathscr{D}(P, Q)}$$

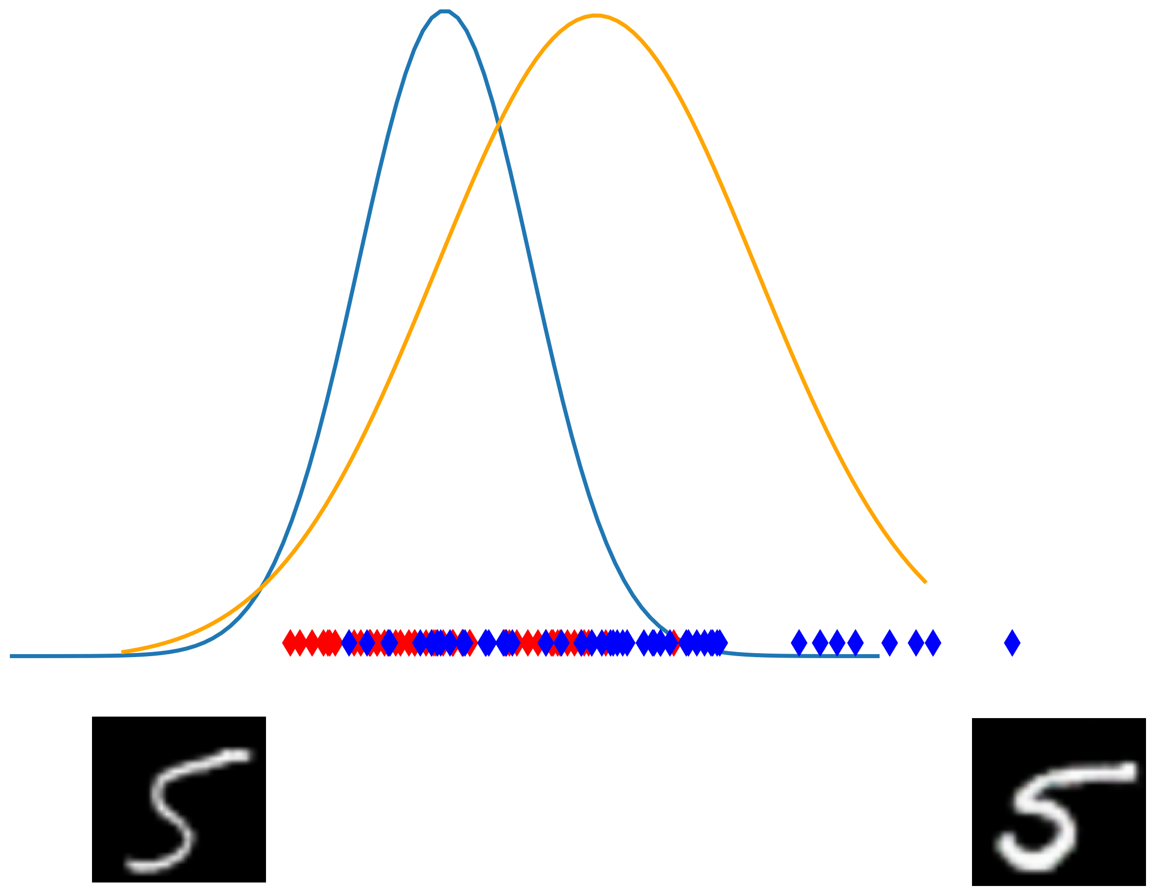


Figure credit: Z., W. Jitkrittum, internet meme

Motivation: Langevin Monte-Carlo

Inference as measure optimization

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

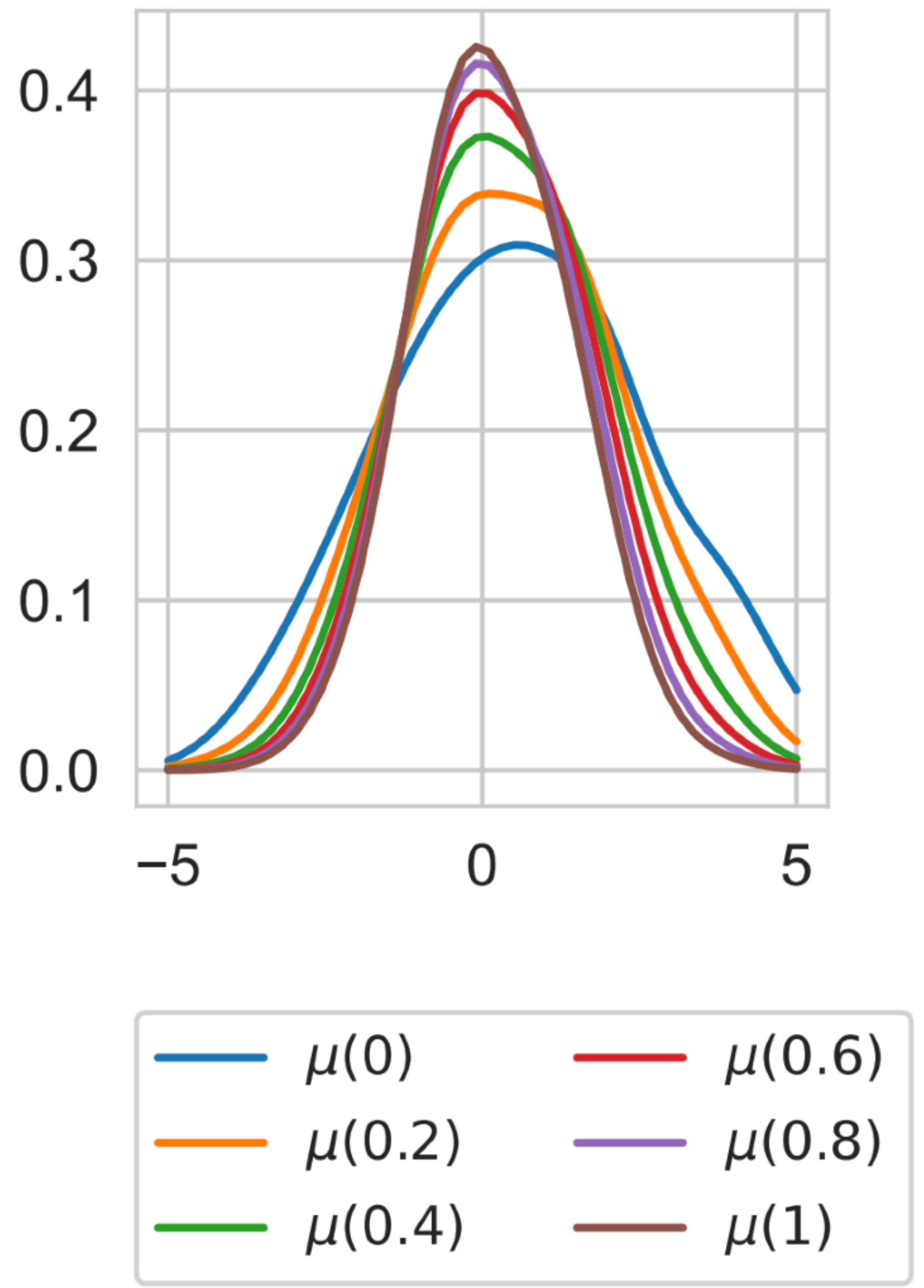
Generate samples from π (or estimate $\mathbb{E}_\pi \psi(X)$ for some ψ)

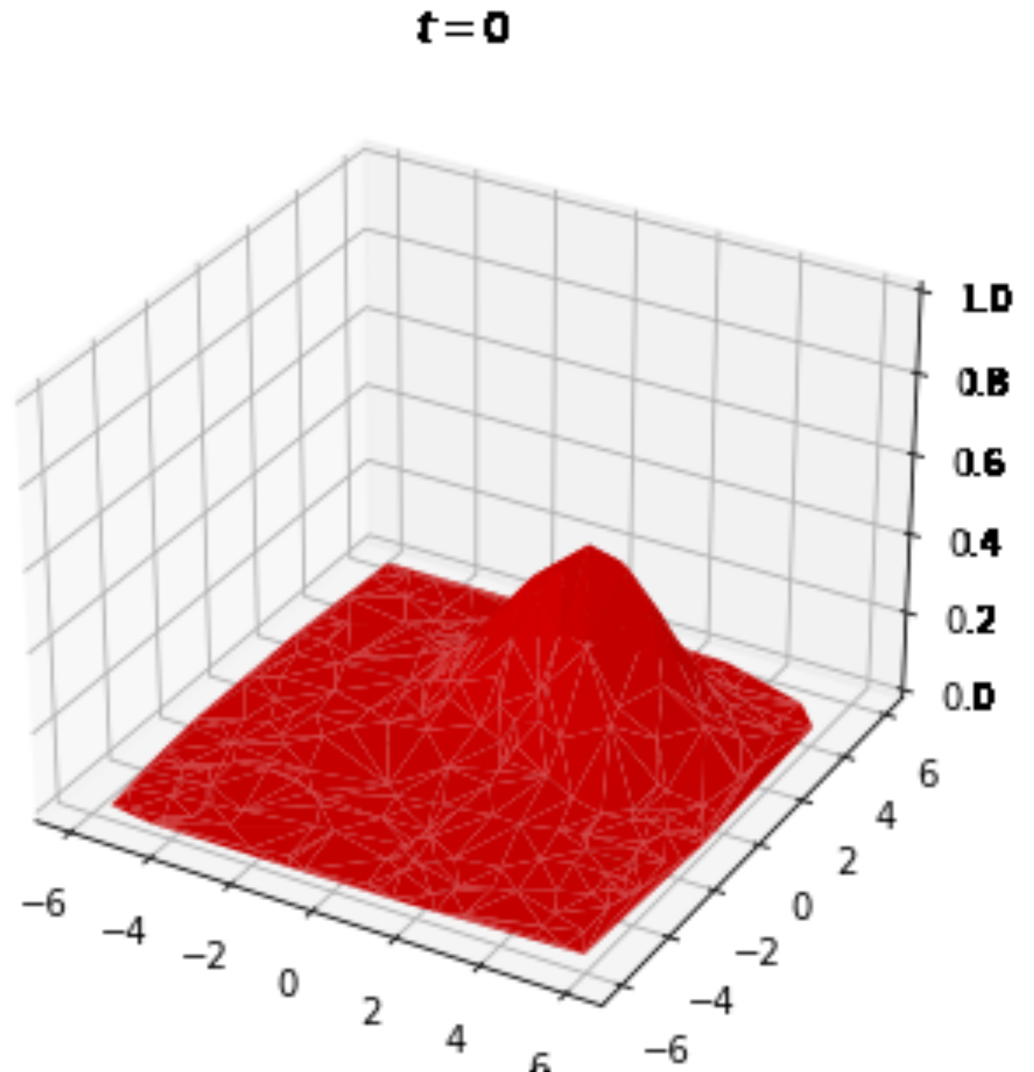
$$\inf_{\mu \in \mathcal{M}} \mathcal{D}_{\mathrm{KL}}(\mu \| \pi).$$

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. N-particle approximation results in the noisy SGD in optimization. This dynamics is **equivalent** to the **PDE gradient flow in the Wasserstein space** [Otto 96].

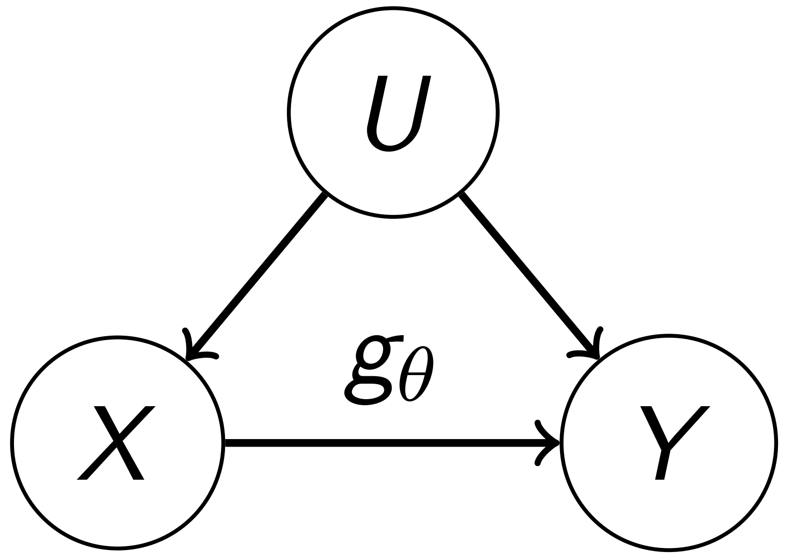




Motivation: Structured Distribution Shift
Causal Confounding

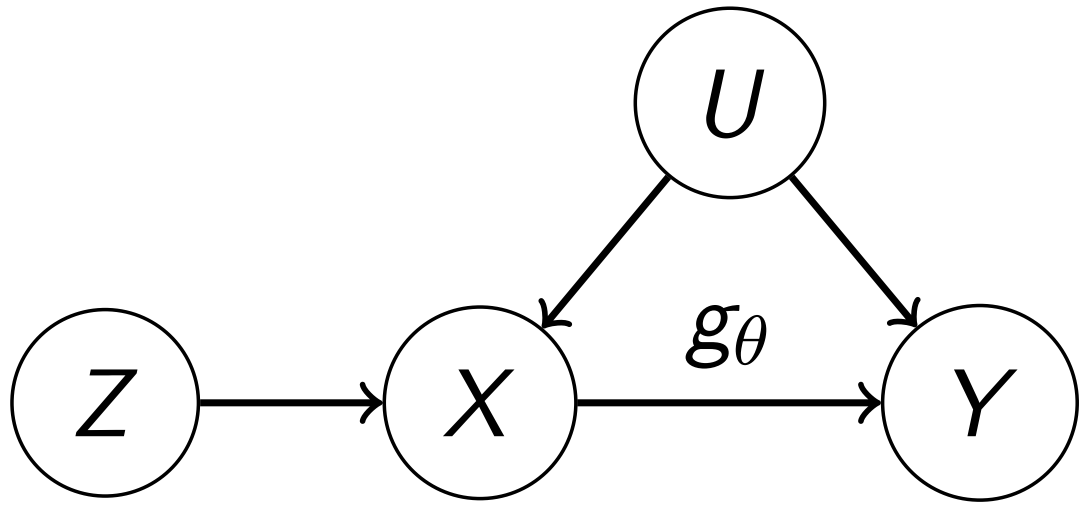
Causal confounding can lead to much stronger distribution shifts than those considered in (joint) distribution shift!

X: Smoking, Y: Cancer, U: Lifestyle



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

Take into account genetic predisposition for nicotine addiction Z



To estimate g_θ robustly, we use **conditional moment restriction**
$\mathbb{E}[\epsilon_U | Z] = \mathbb{E}[Y - g_\theta(X) | Z] = 0 \ \mathbb{P}_Z$-a.s.

Figure credit: H. Kremer

Motivation & Introduction

Kernel Methods for **Robust Learning under Distribution Shift**

Distributional <u>robustness</u>, but what kind?

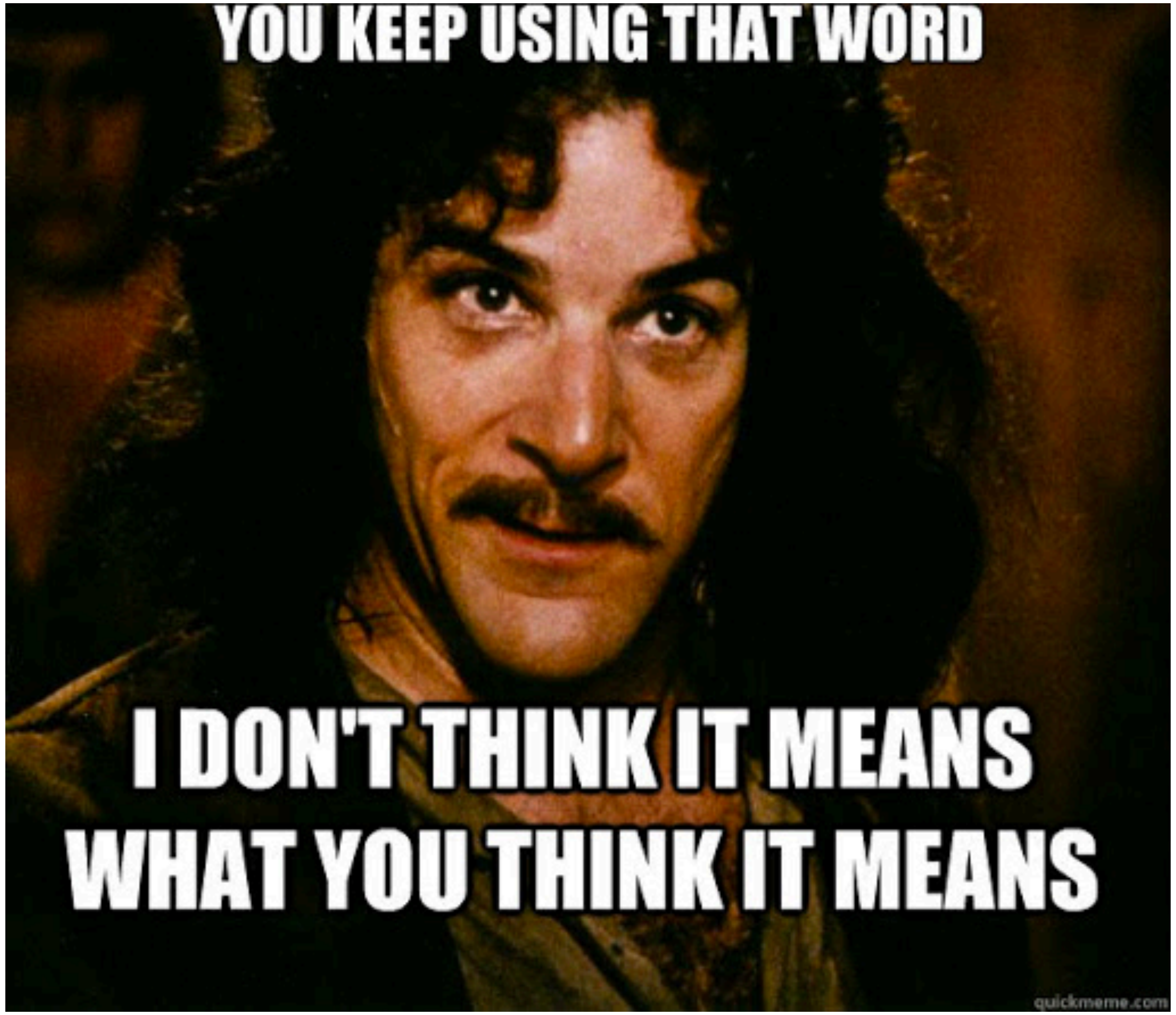


Figure credit: The Princess Bride,
a bedside story by your grandpa

From Statistical Learning to Distributionally Robust Learning

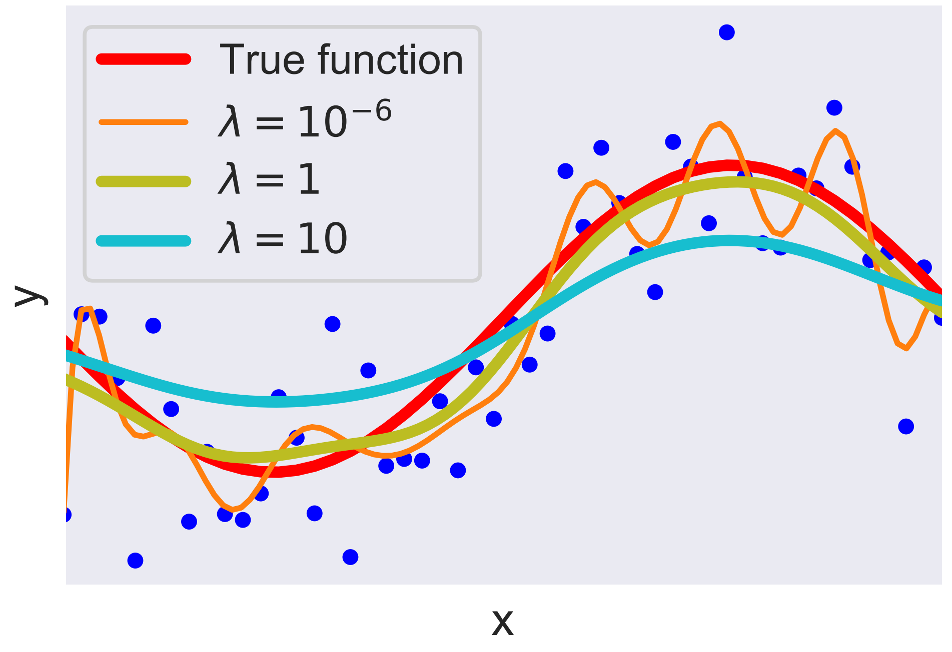
Empirical Risk Minimization

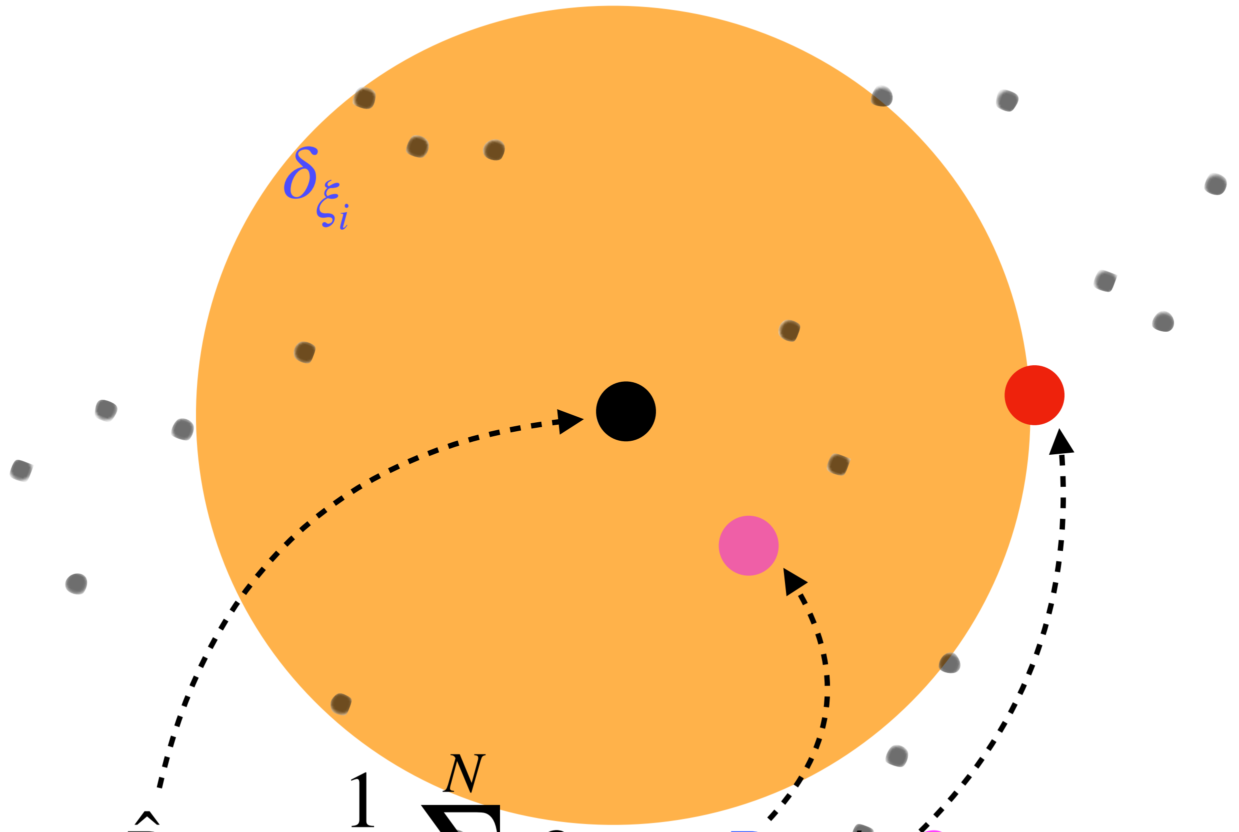
$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} l(\theta, \xi_i), \quad \xi_i \sim P_0$$

- "Robust" under statistical fluctuation

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

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



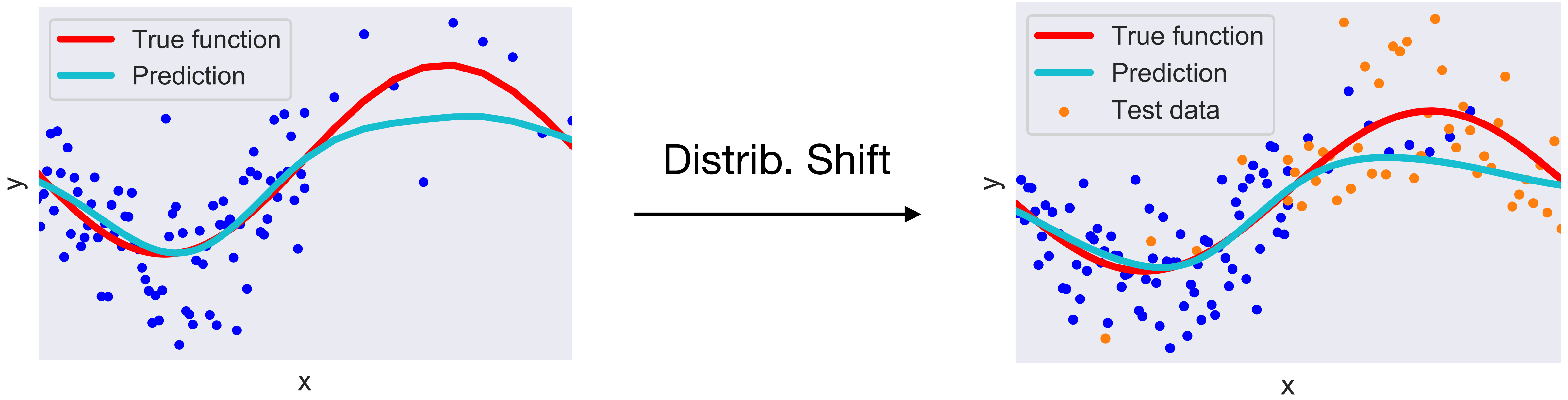


$$\hat{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

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 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

Figure credit: H. Kremer

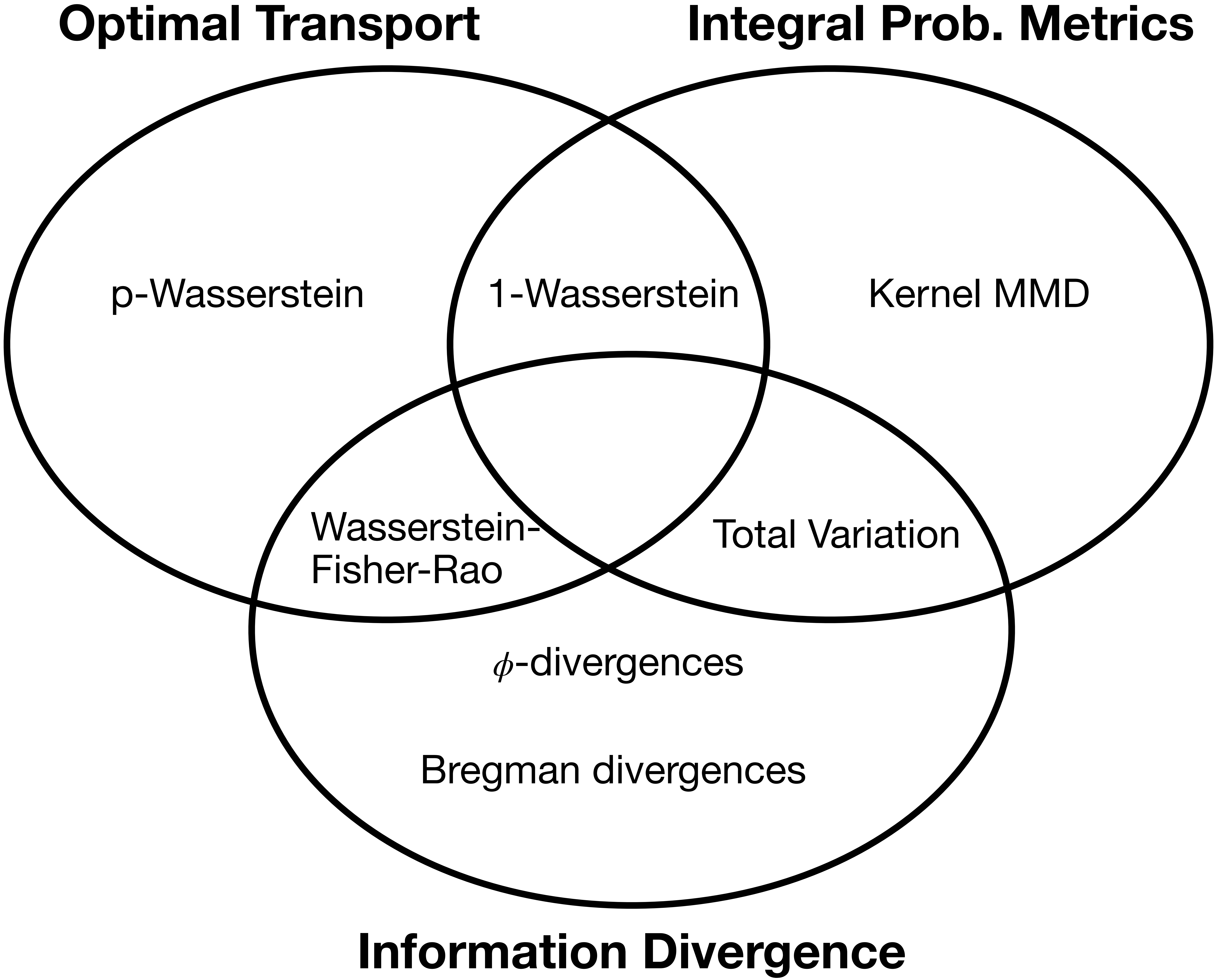


Figure credit: J. Zhu

Background: “Kernel Geometry”

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

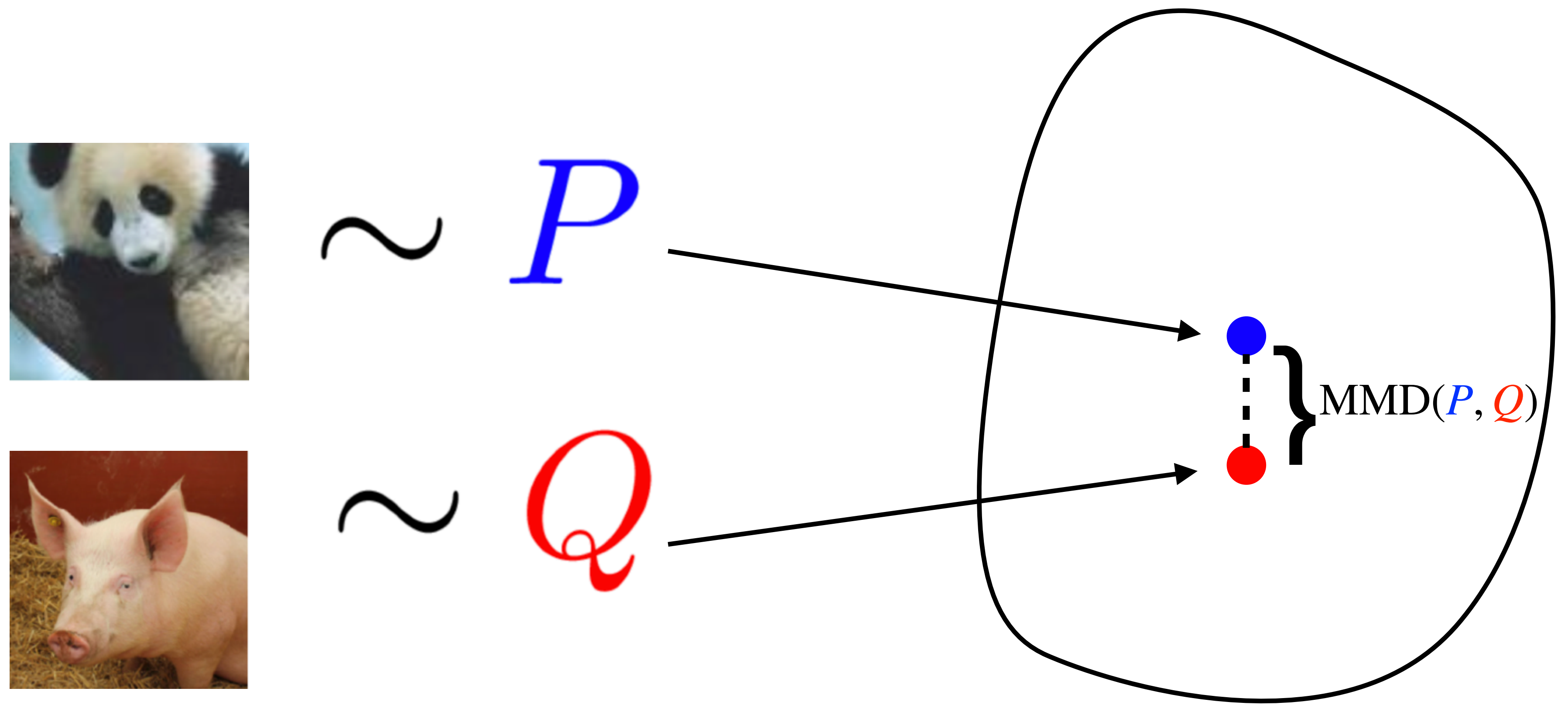
$$\mathrm{MMD}(P, Q) := \left\| \int k(x, \cdot\,) dP - \int k(x, \cdot\,) dQ \right\|_{\mathscr{H}}.$$

$(\mathrm{Prob}(\mathbb{R}^d), \mathrm{MMD})$ 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 $\mathscr{H}$.



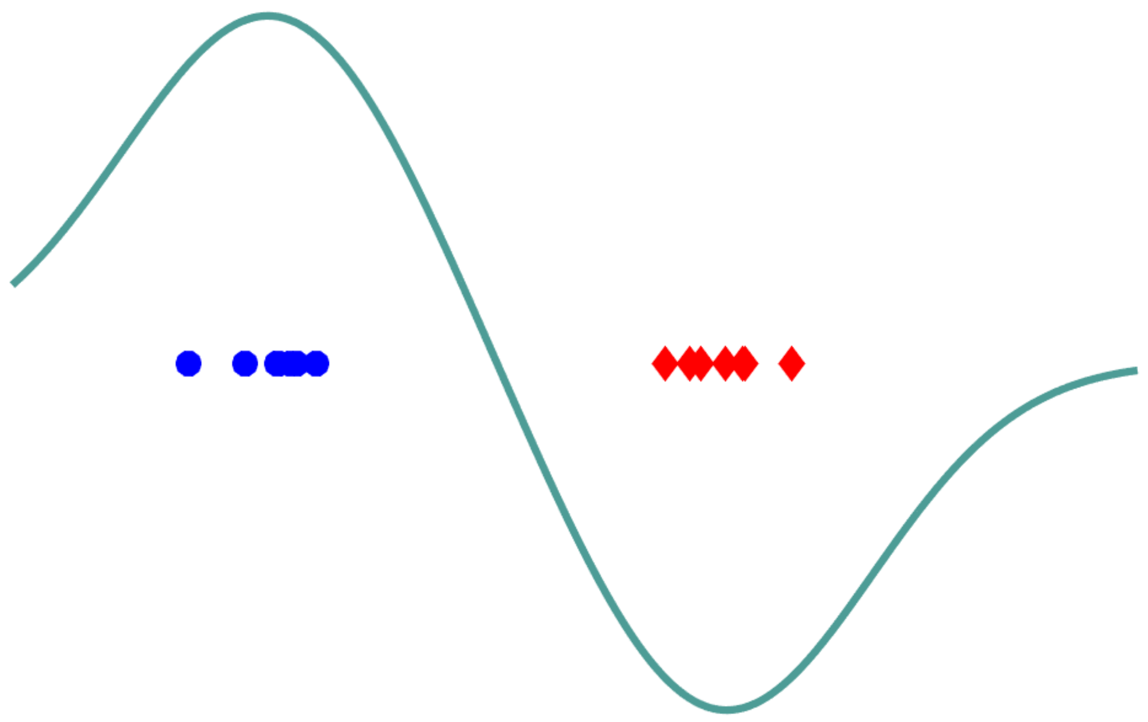
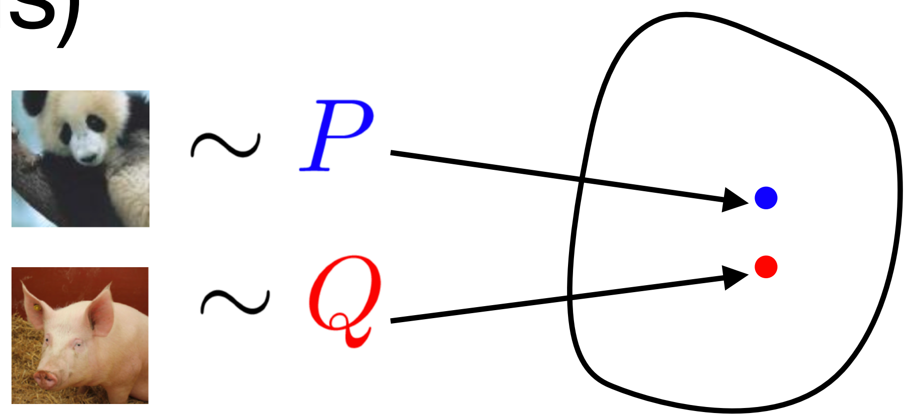


Figure credit: W. Jitkrittum, J. Zhu

Previous work: Kernel DRO

Primal DRO (not solvable as it is)

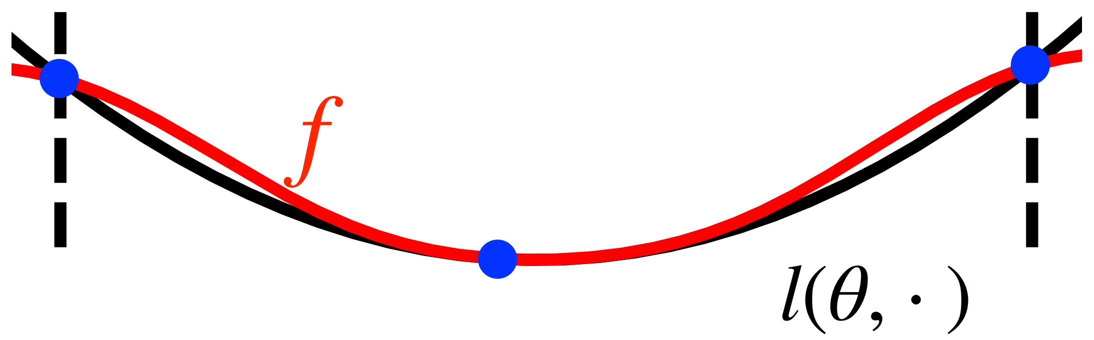
(DRO) $\min_{\theta} \sup_{\mathrm{MMD}(Q,\hat{P})\le\epsilon} \mathbb{E}_Q l(\theta,\xi)$

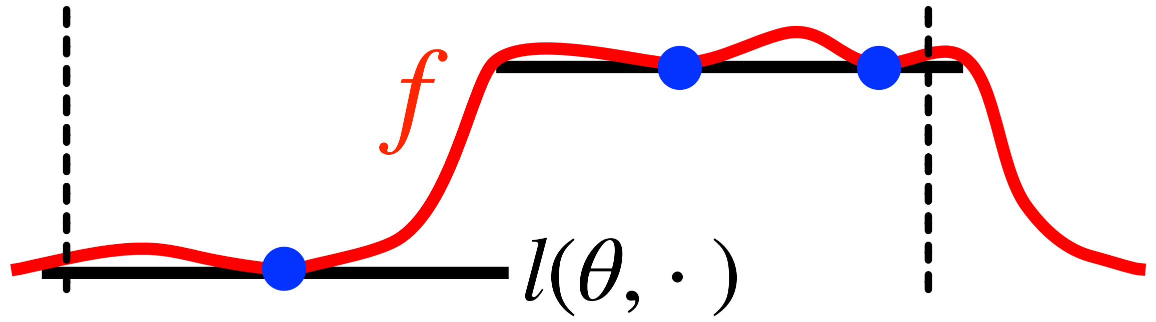


Kernel DRO Theorem (simplified). [Z. et al. 2021]
DRO problem is equivalent to the dual kernel machine learning problem, i.e., (DRO)=(K).

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

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

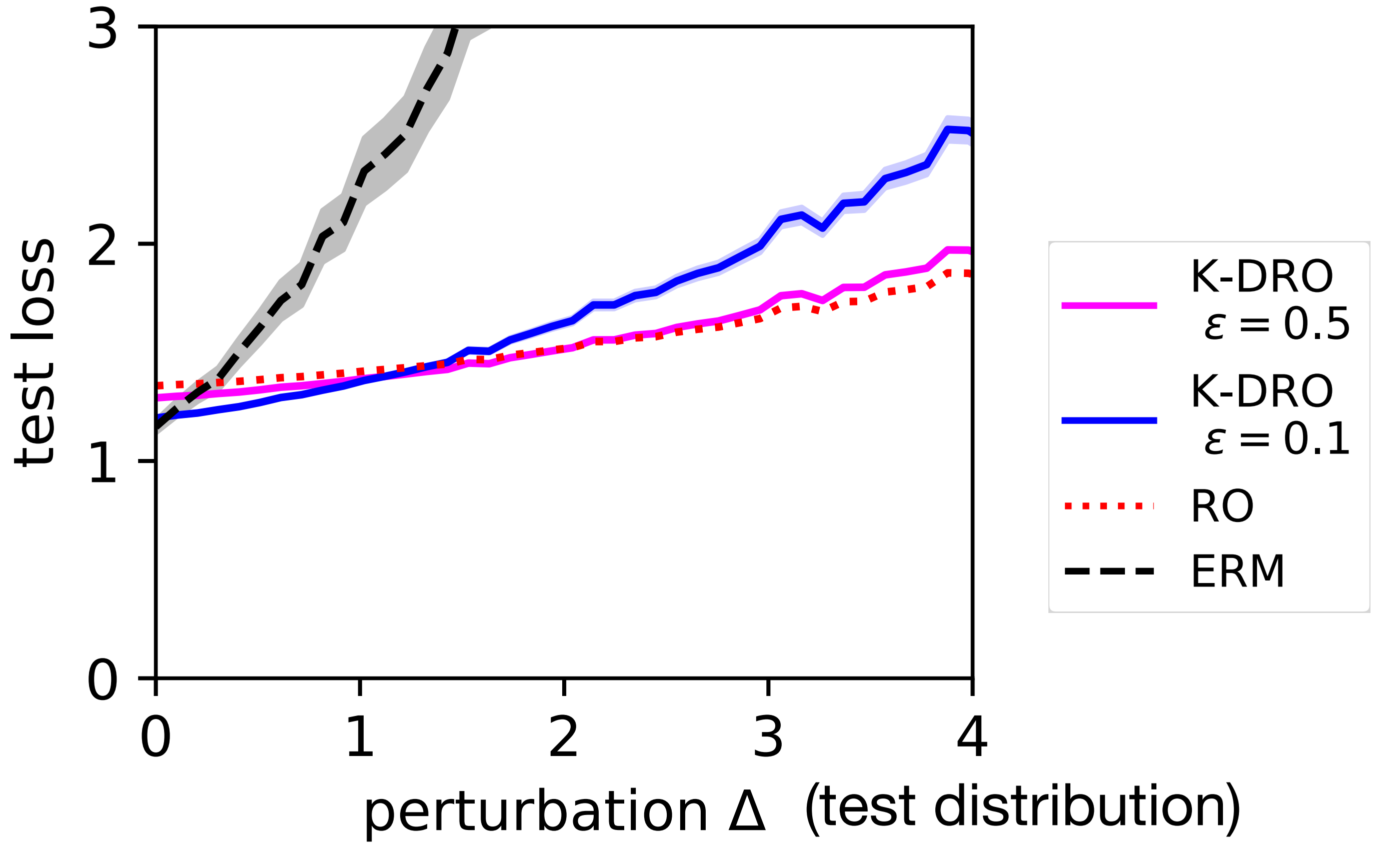




Example. Robust least squares
[El Ghaoui Lebret '97]

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

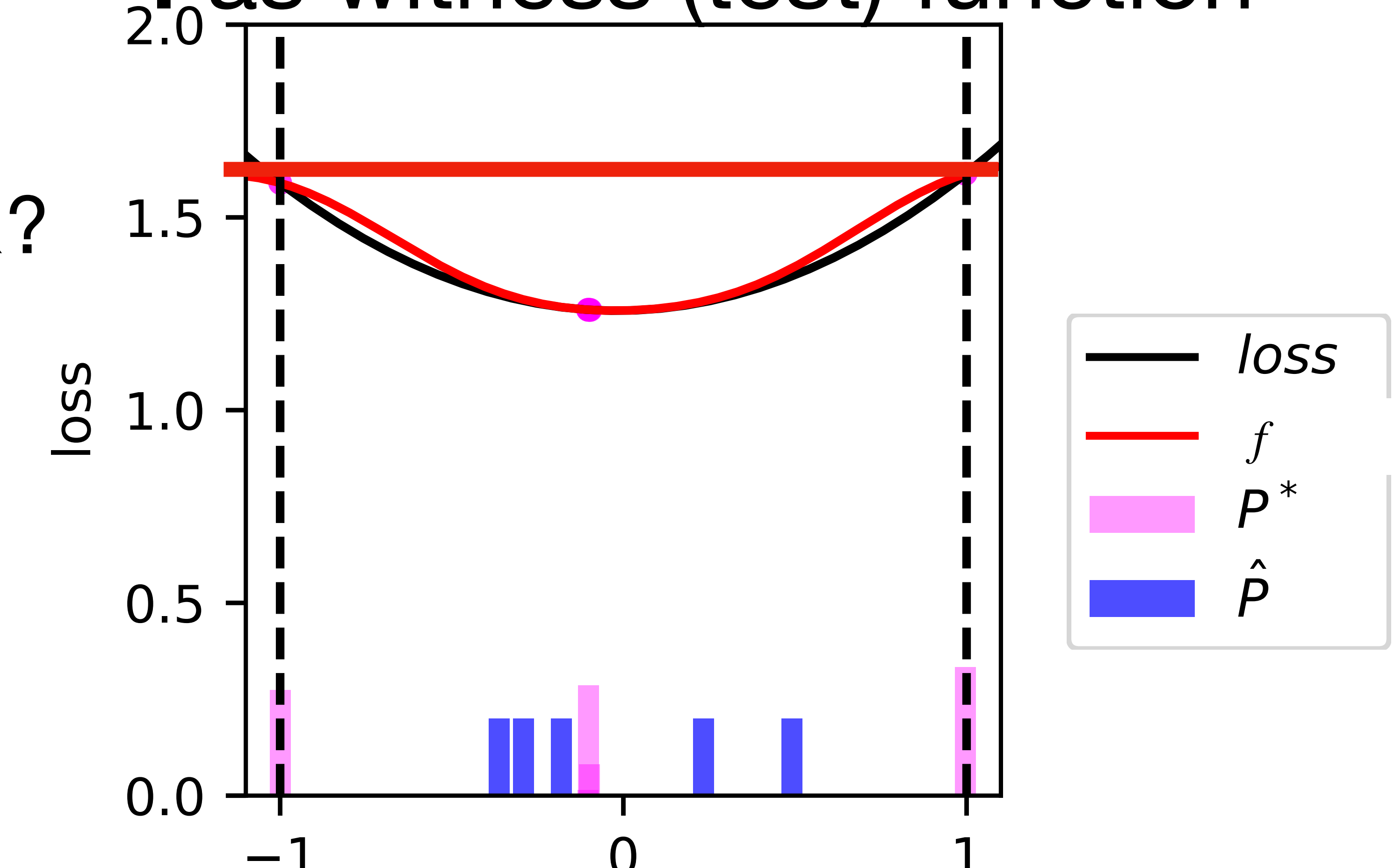
Given historical samples $\xi_1, \xi_2, \ldots, \xi_N$



Robustifying with DRO

f as witness (test) function

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



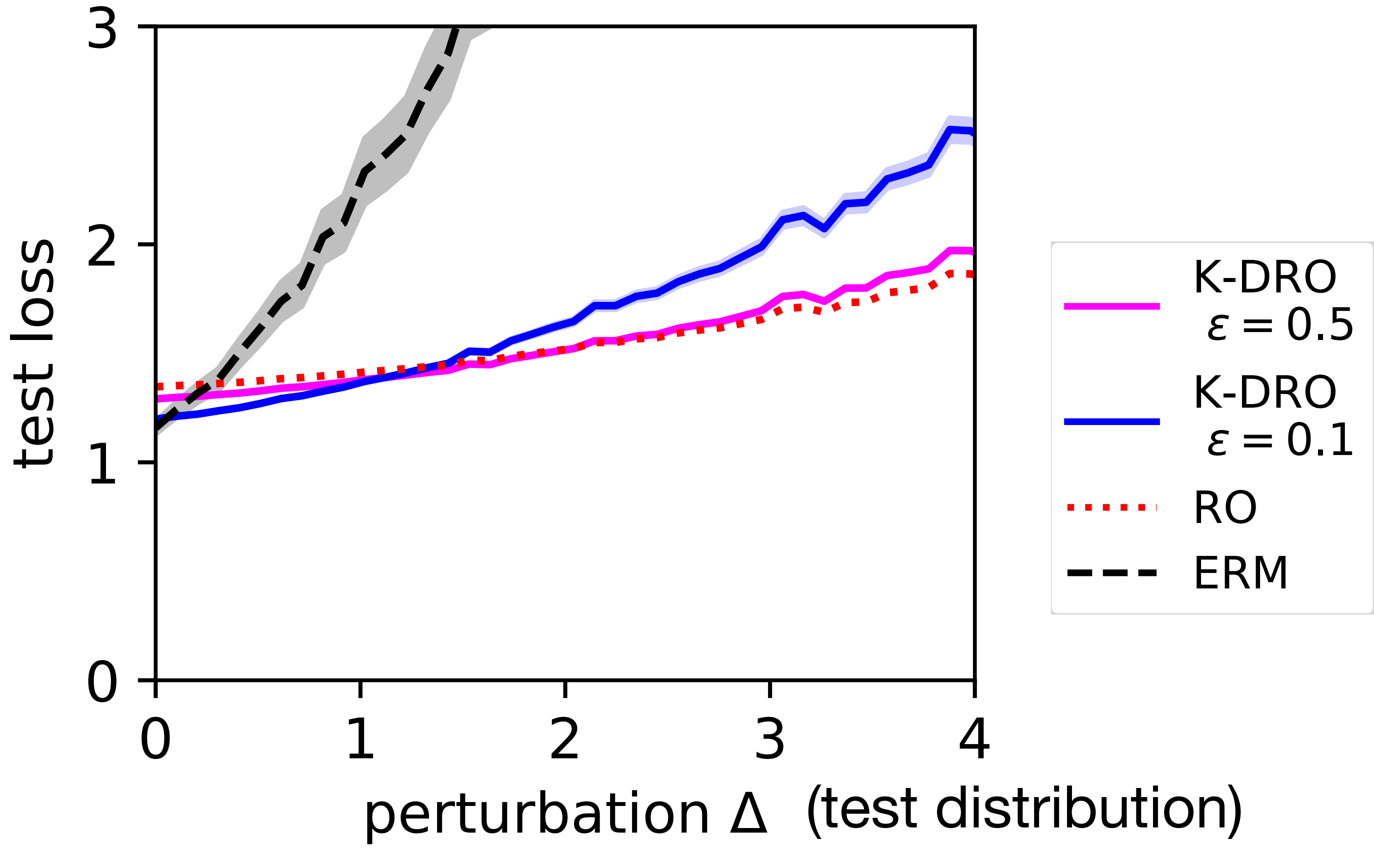
Experiments of robustness under distribution shifts

Toy example. Uncertain least squares

[El Ghaoui Lebret '97, **Z** et al. 2021]

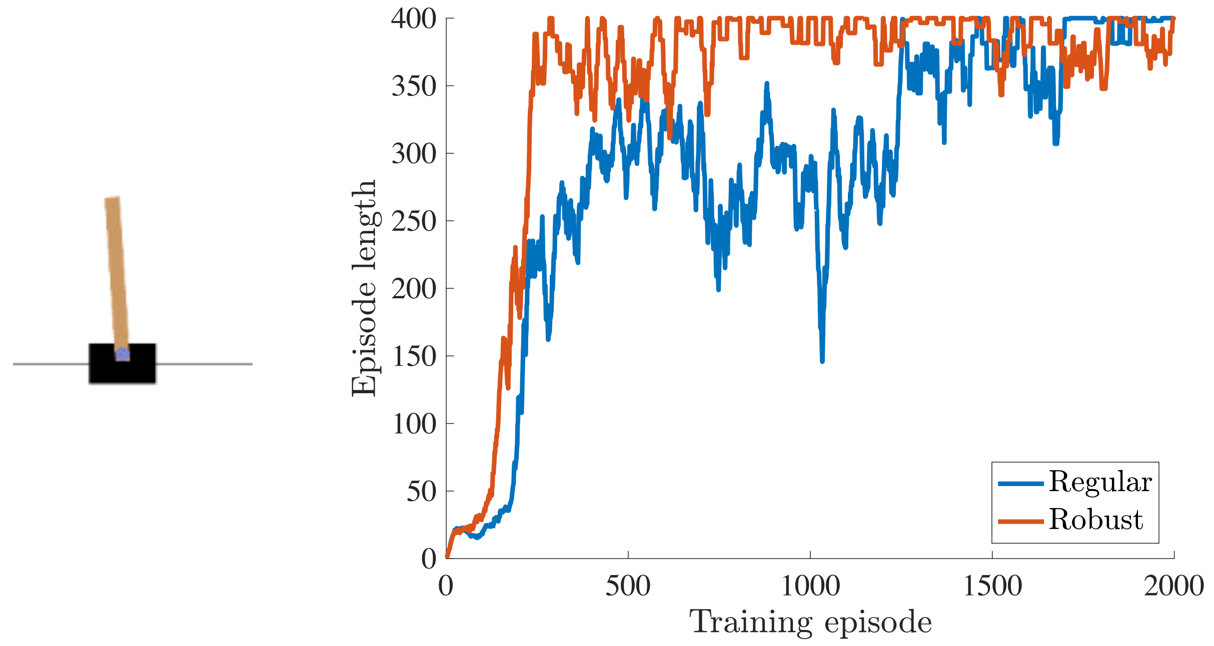
$$\text{minimize } l(\theta,\xi) := \|A(\xi)\cdot\theta - b\|_2^2$$

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

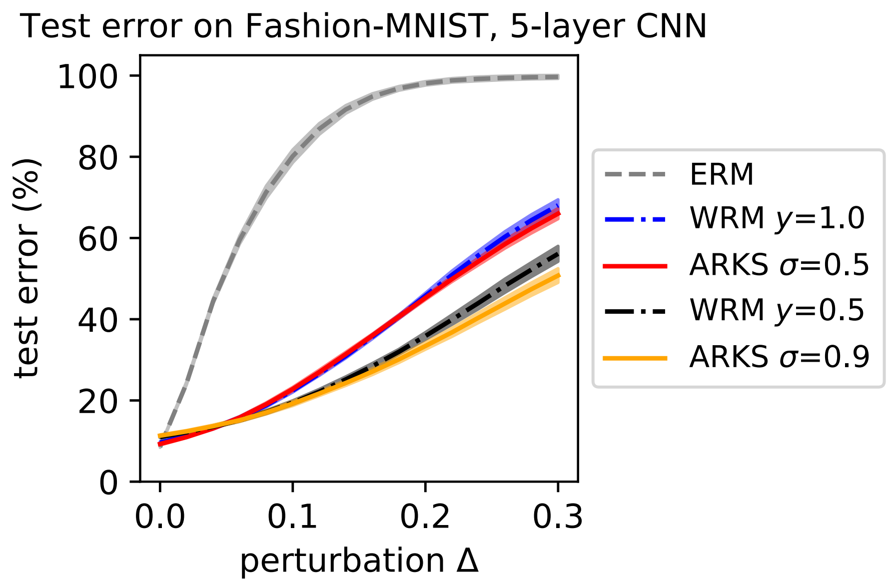


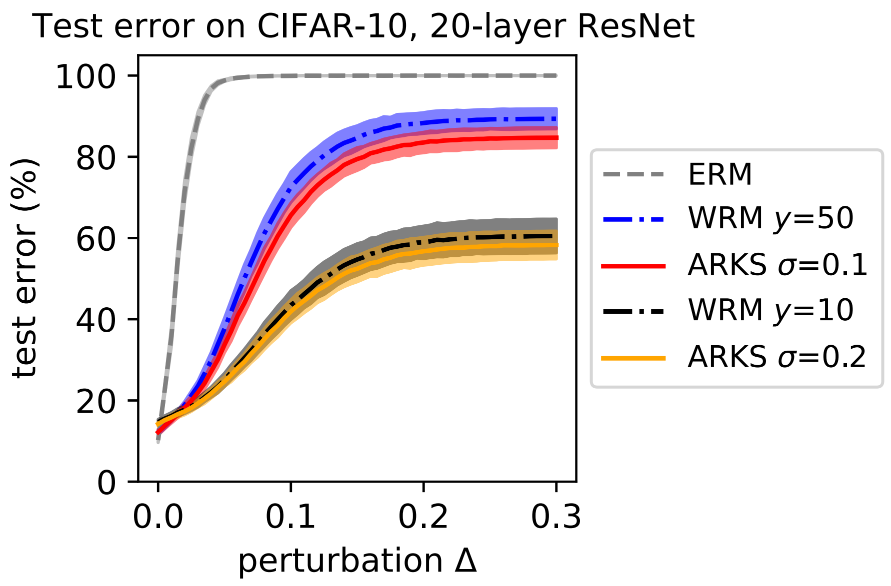
Robust reinforcement learning

[Sinha et al. 2017]

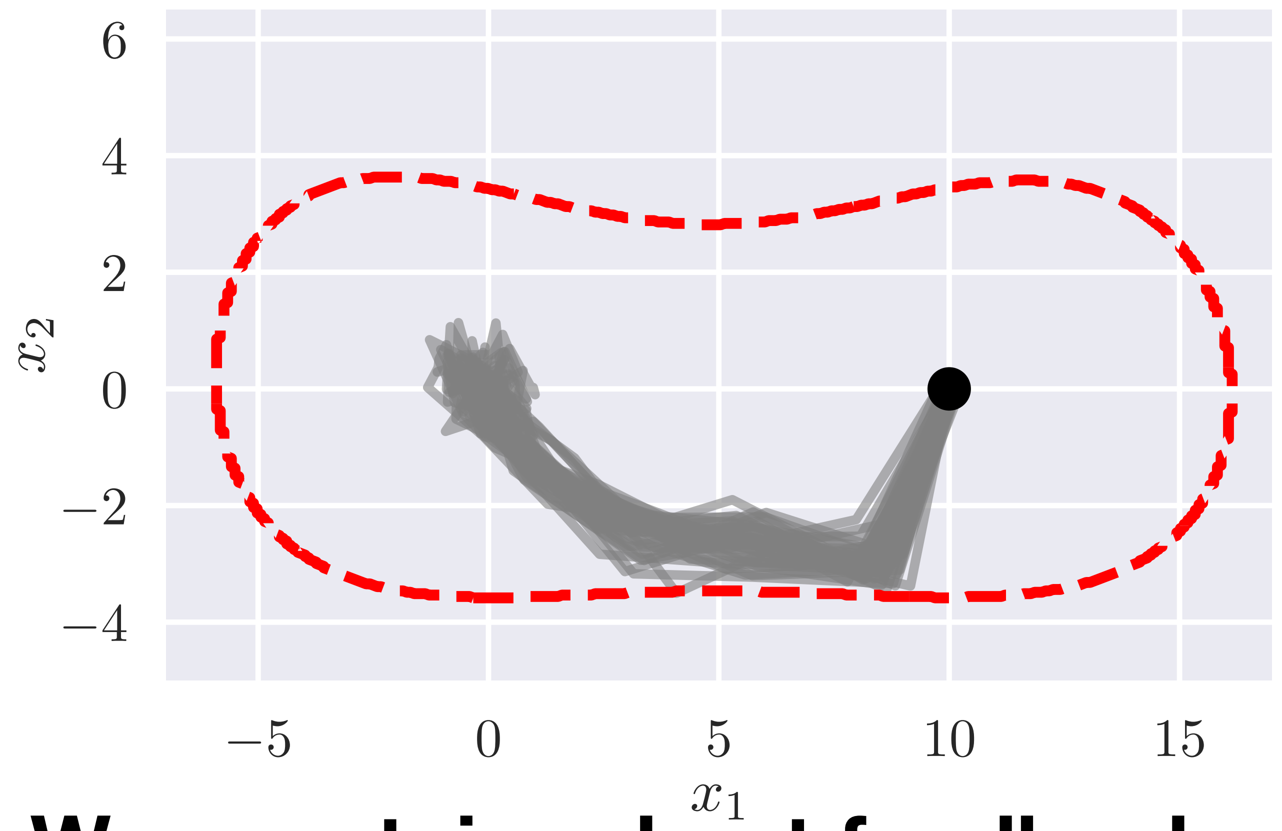


Certified adversarially robust deep learning (Classify the presence of glasses using a 20-layer DNN [Sinha et al. 2017; Z et al. 2022]

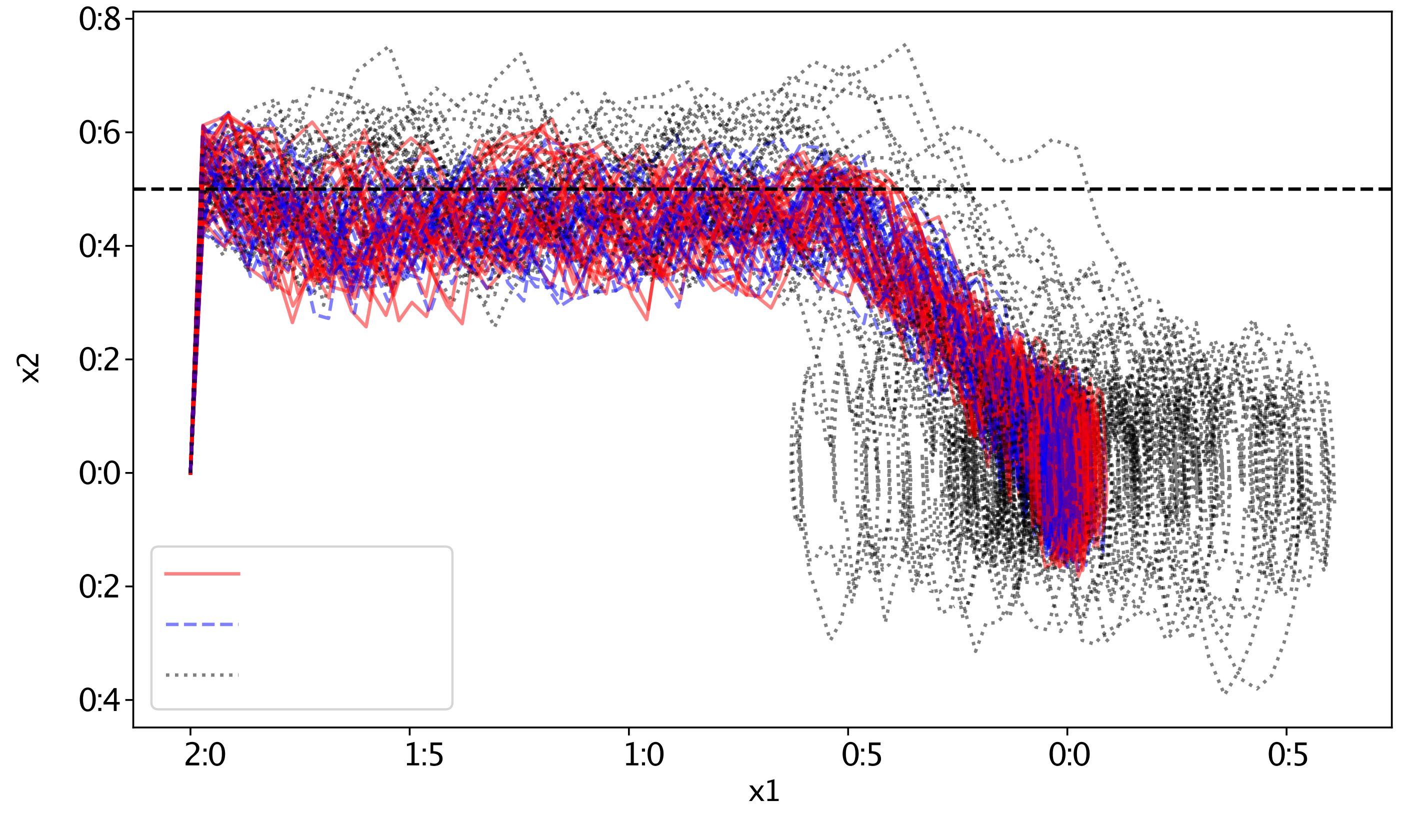




Kernel stochastic model predictive control (MPC) with nonlinear constraints [Nemmour et al. & **Z** 2022]



Wasserstein robust feedback optimal control of nonlinear dynamical systems [Zhong & **Z** 2023]



Entropy-MMD & Conditional independence

To relax the semi-infinite constraint in DRO reformulations

$$\min_{\theta} \sup_{\mathrm{MMD}(Q,\hat{P})+\lambda D_\phi(Q\|\omega)\le\epsilon} \mathbb{E}_Q l(\theta,\xi)$$

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

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

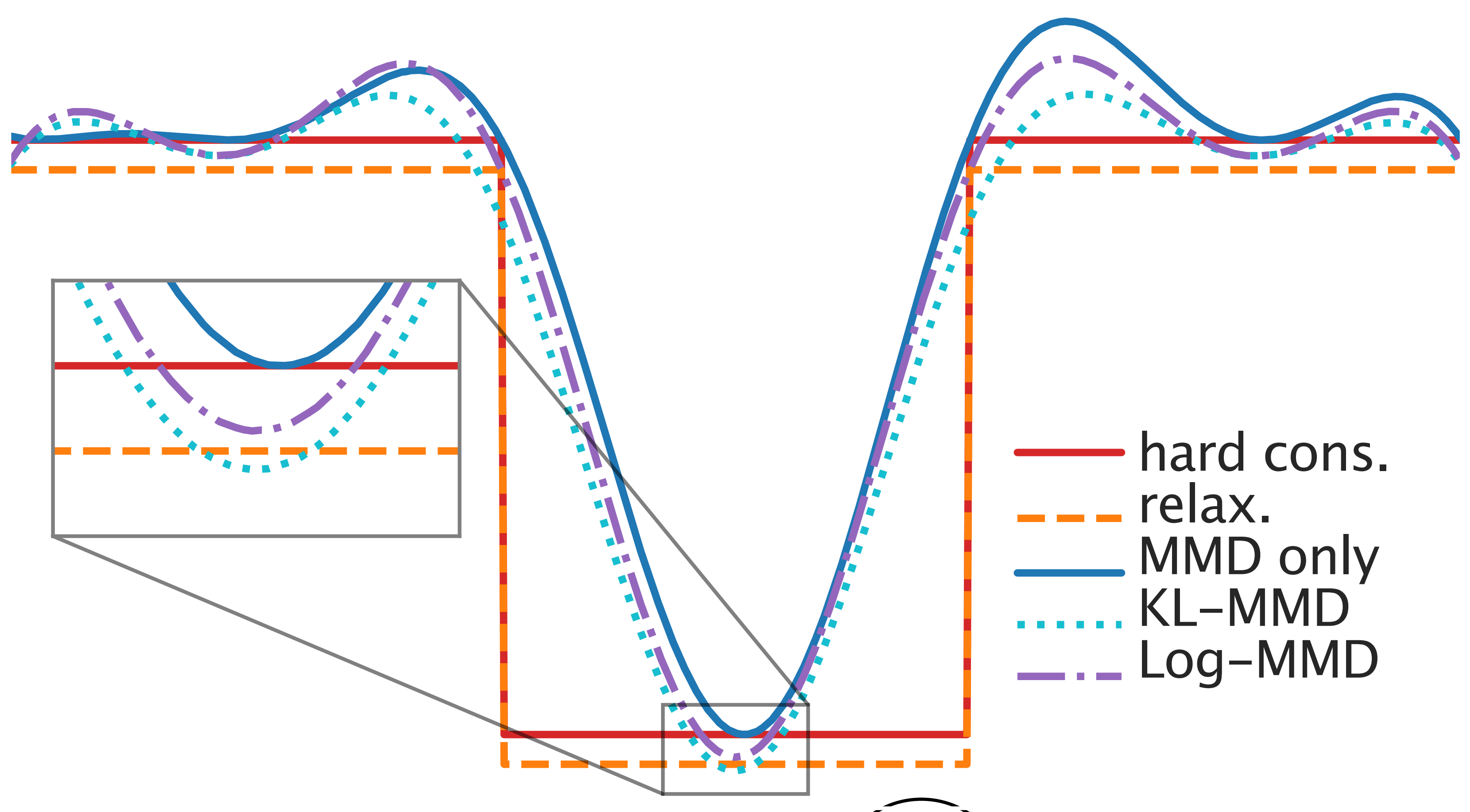
soft cons. $\phi^*_{\mathrm{KL}}(t) = \exp(t)$, log-barrier $\phi^*_{\log}(t) = -\log(1-t)$

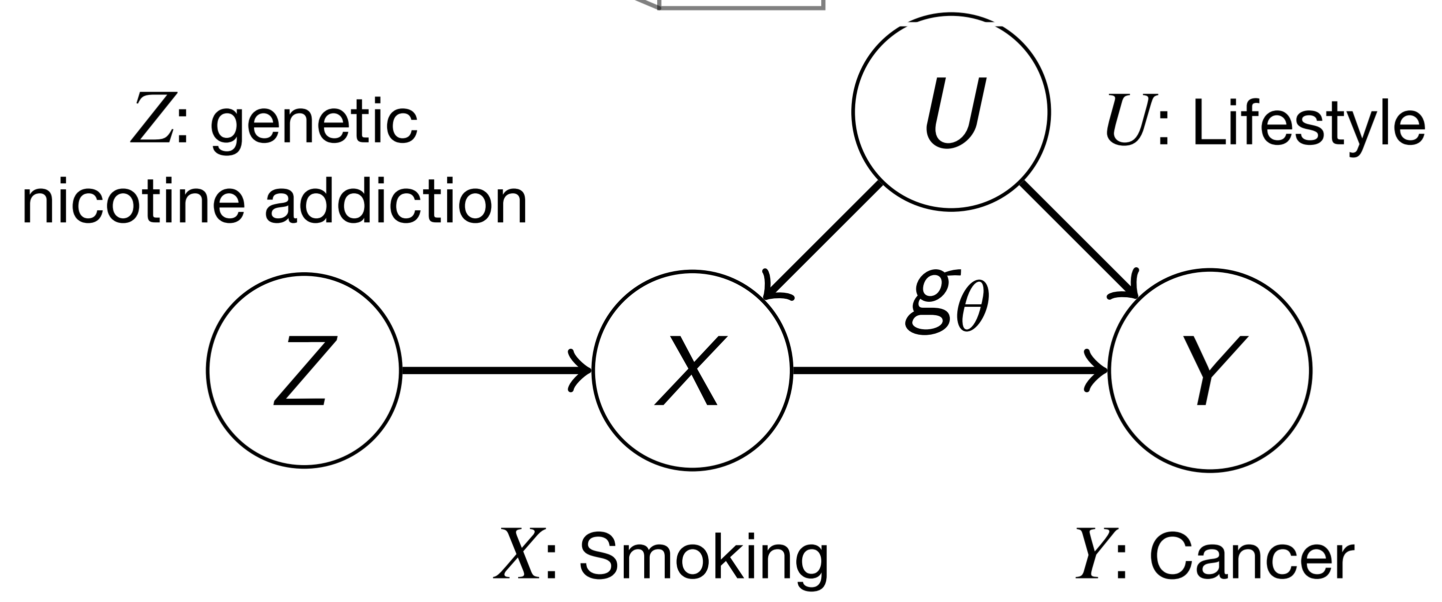
Example (Causality). Robustness against **structured distribution shifts** instead of (joint-)DRO. [Kremer et al., **Z.** 2023]. Estimating g_θ via **conditional moment restriction**

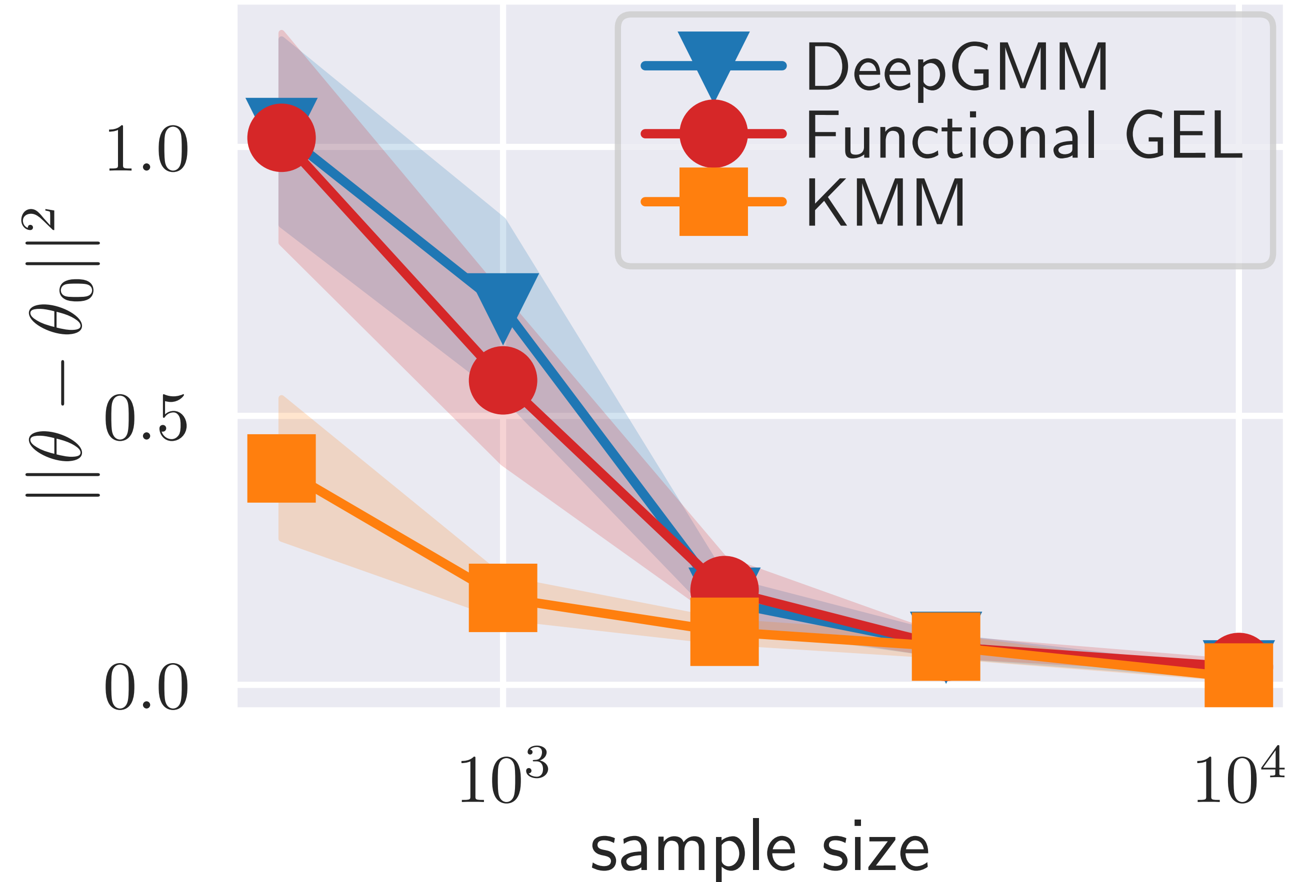
$$\mathbb{E}[Y - g_\theta(X)\,|\,Z] = 0 \;\; \mathbb{P}_Z\text{-a.s.}$$

Empirical likelihood formulation + similar techniques: $\forall h \in \mathcal{H}$,

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







DR Probablistic Robust Optimization

Example. Nonlinear DR-chance constrained opt. [Nemmour et al., **Z.** 2022]

$$\min_{x \in \mathcal{X}} c^T x \quad \mathrm{s.t.} \inf_{D(P, \hat{P}_N) \le \epsilon} P(f(x, \xi) \le 0) \ge 1 - \alpha$$

$f(x, \xi)$ **nonlinear** in uncertainty.

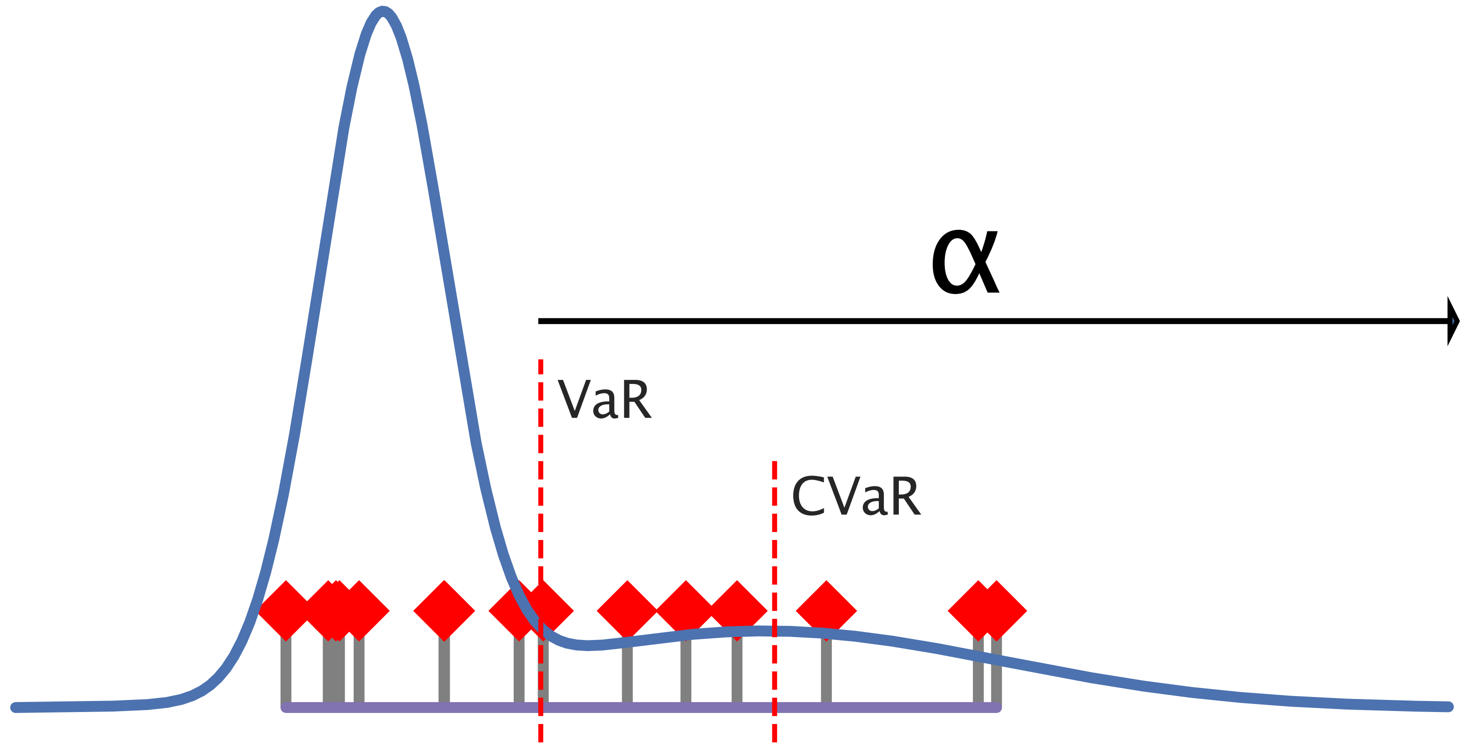
Idea: Rewrite $\mathbb{E}$ using the indicator function (or CVaR)

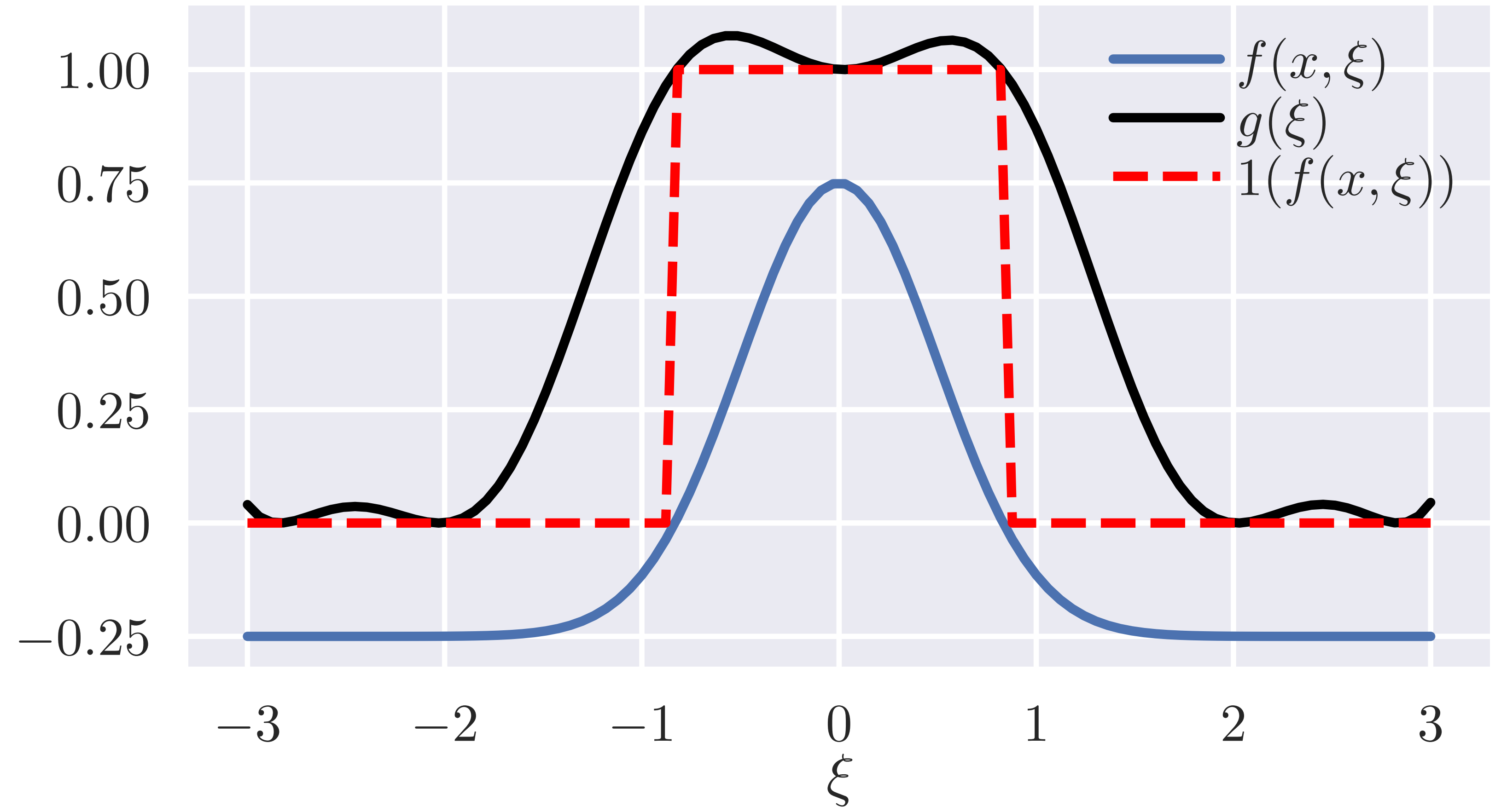
$$\sup_{D(P, \hat{P}_N) \le \epsilon} \mathbb{E}_P[\mathbb{I}(f(x, \xi) \ge 0))] \le \alpha$$

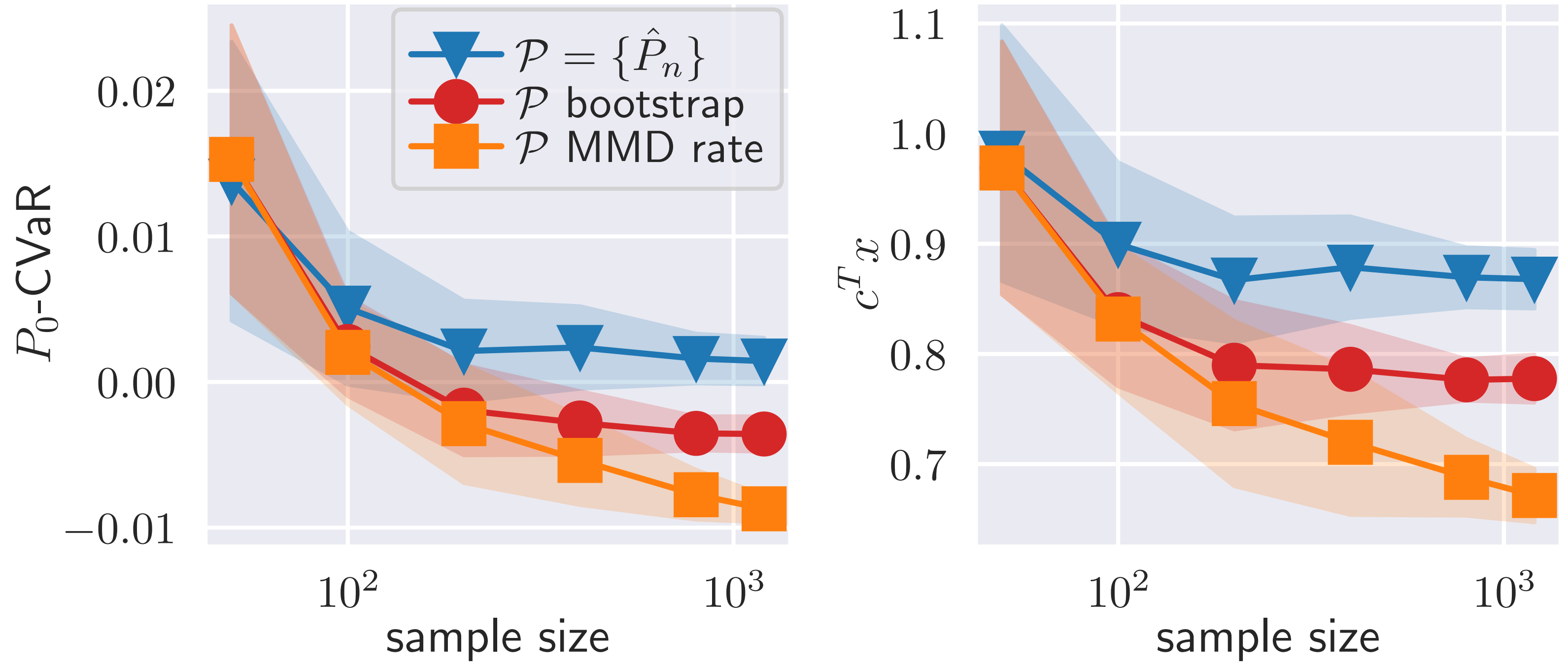
Apply the Kernel DRO Theorem and algorithm, informally

$$P\left(f(x, \xi) \le \mathcal{O}(\frac{1}{\sqrt{N}})\right) \ge 1 - \alpha,$$

with large probability. The big-O term depends on the class of f.







More ML problems we can use dual functions for measure optimization

Primal-dual optimization problems

$$\inf_{\mu \in \mathcal{M}} F(\mu) = \sup_{f \in \mathcal{F}} \mathcal{E}(f)$$

Examples in ML [Dvurechensky, **Z.**]

Generative models

$$\inf_{G_\theta} \mathbb{E}_Z \mathcal{D}(P, G_\theta(Z)) = \inf_{\mu \in \mathcal{M}} \sup_{f \in \mathcal{F}} \left\{ \int f(x) dP(x) - \mathbb{E}_{\theta \sim \mu} \int f(g_\theta(z)) dQ(z) \right\}$$

Distributionally robust optimization

$$\inf_{\theta} \sup_{\mathrm{MMD}(\mu, \hat{\mu}) \le \epsilon} \mathbb{E}_\mu [l(\theta; x)] = \inf_{\theta \in \mathbb{R}^d, f \in \mathcal{H}} \sup_{\mu \in \mathcal{M}} \mathbb{E}_\mu (l - f) + \frac{1}{N} \sum_{i=1}^{N} f(x_i) + \epsilon \|f\|_{\mathcal{H}}.$$

Wasserstein barycenter

$$\min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} \alpha_i \left[W_p(\mu, \nu_i) \right] = \min_{\mu \in \mathcal{M}} \sum_{i=1}^{n} \alpha_i \sup_{f_i \in \Psi_c} \left\{ \int f_i^c \mathrm{d}\mu + \int f_i \mathrm{d}\nu_i \right\},$$

Summary

- There are many important uses of the **dual** (kernel) function for **measure optimization**: causal inference, barycenter problems, conditional moments, (robust) control and RL
- However, optimization over measures is a mathematically non-trivial topic. We will now learn the *optimization perspective of **gradient flow***

This talk is mainly based on:

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

Nemmour, Y. , Kremer, H., Schölkopf, B. & **Z**. MMD Distributionally Robust Nonlinear Chance-Constrained Optimization with Finite-Sample Guarantee. IEEE CDC 2022

Z., Kouridi, C., Nemmour, Y. & Schölkopf, B. Adversarially Robust Kernel Smoothing. AISTATS 2022

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

Slides will be available
Website: jj-zhu.github.io