Robust Optimization and Machine Learning under Distribution Shift

J.J. (Jia-Jie) Zhu
jj-zhu.github.io

WIAS, Berlin & MPI-IS, Tübingen

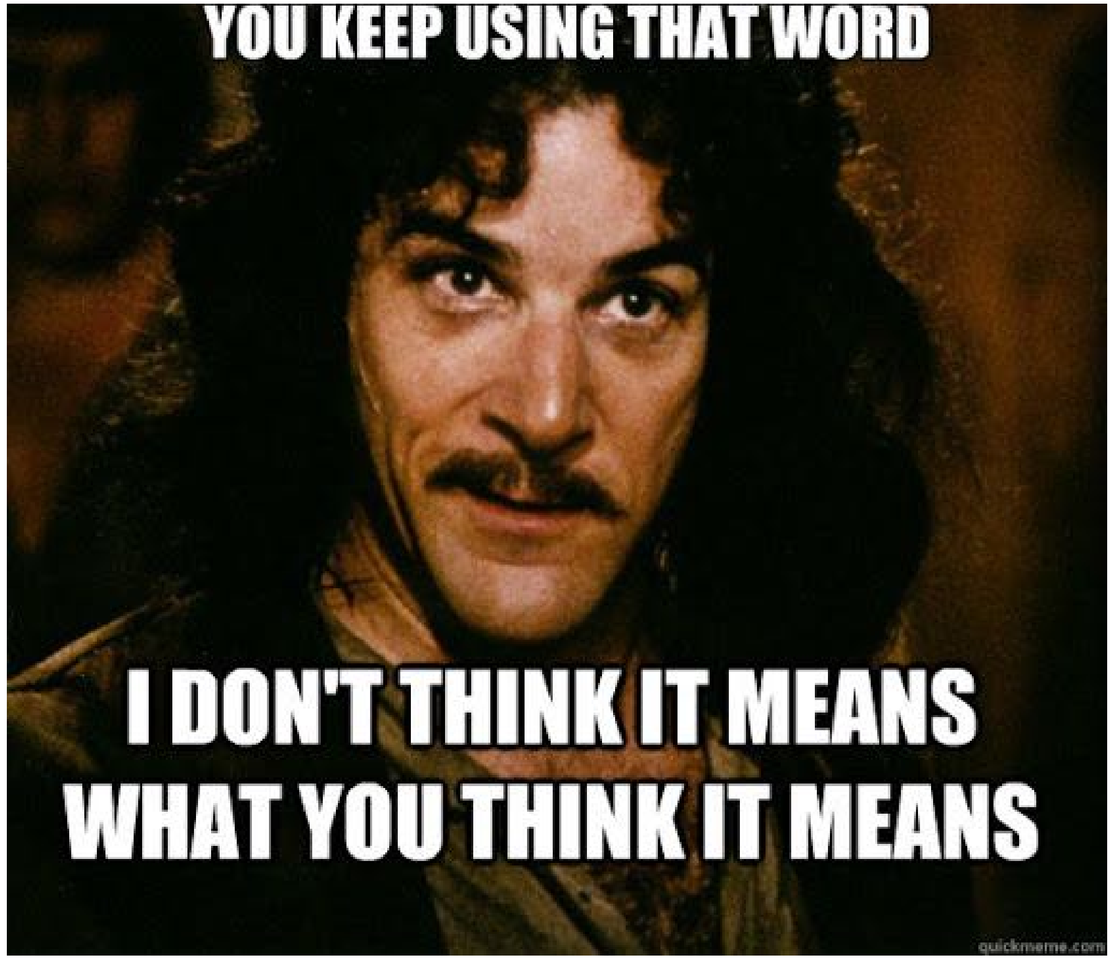
Leibniz MMS Summer School, Dagstuhl
August 26, 2021

Distributional Robustness

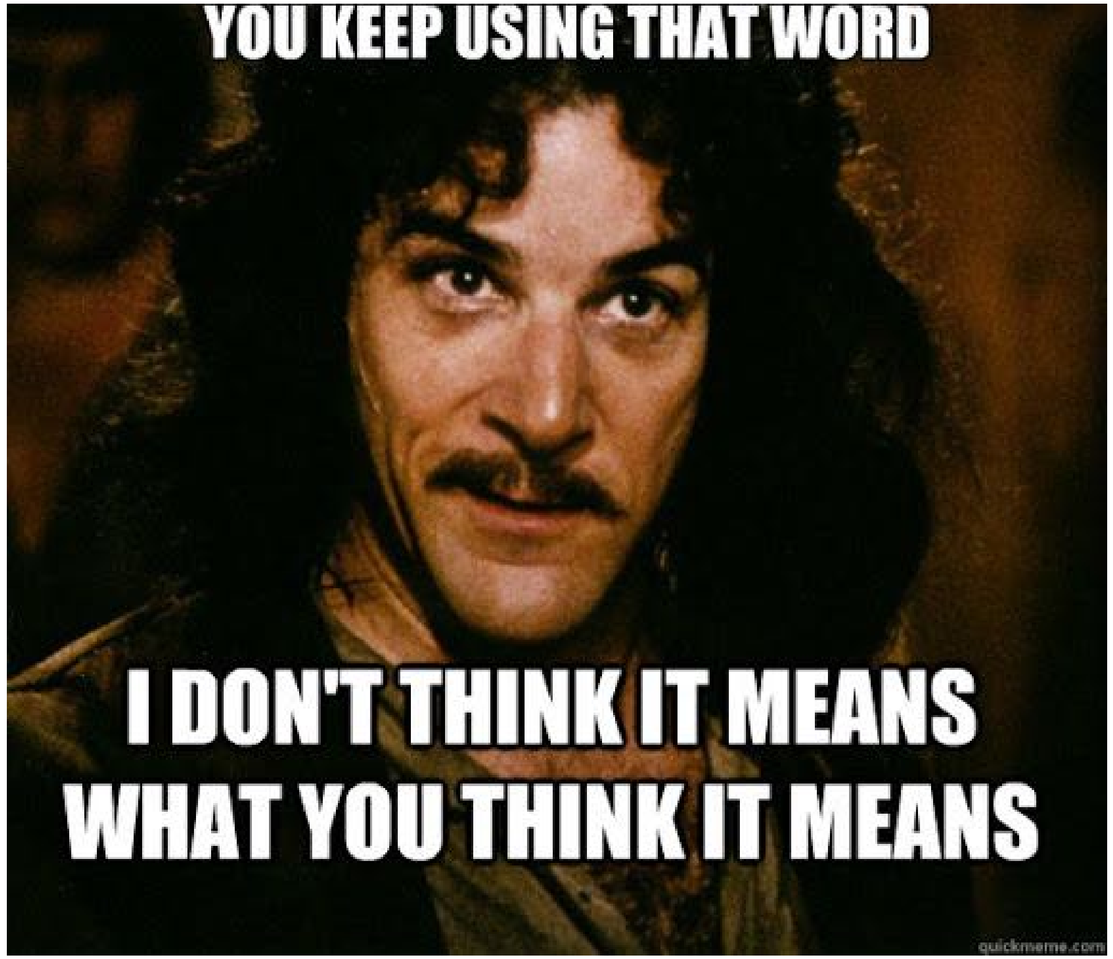
Distributional **Robustness**

What is robustness?

What is robustness?

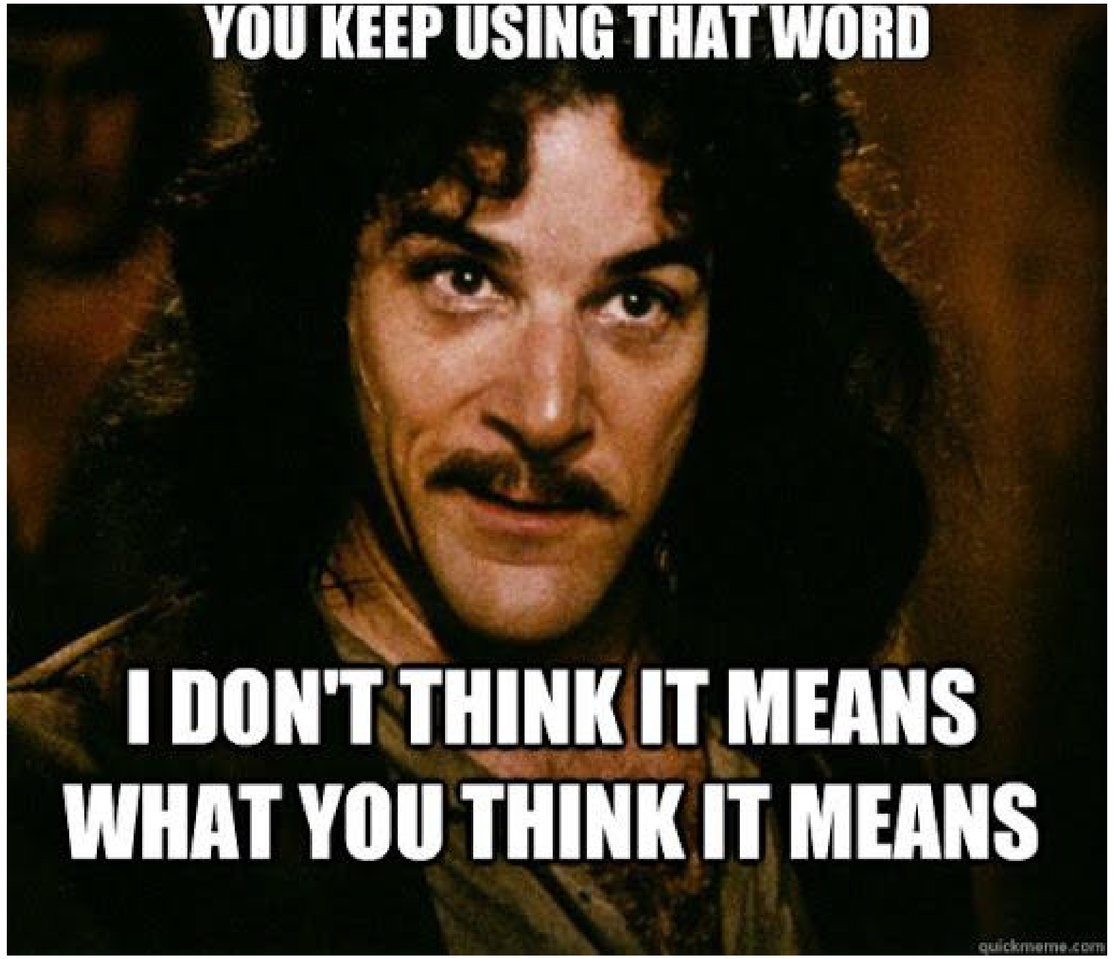


What is robustness?



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

What is robustness?



- Many fields: …robust statistics, robust control, robust optimization, adversarial robustness, robust learning…
- Robustness is a principle for decision-making

Optimization under undertainty

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

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

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost

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

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost → uncertain

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

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost

uncertain

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

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

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost

uncertain

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

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

- Do well on average

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost

uncertain

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

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

- Do well on average
- Strength: high-performance (optimal)

Optimization under undertainty

Empirical risk minimization (ERM)

(sample average approximation (SAA))

loss/cost

uncertain

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

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, sim2real, off-policy/offline RL

Optimization under undertainty

Empirical risk minimization (ERM)
(sample average approximation (SAA))

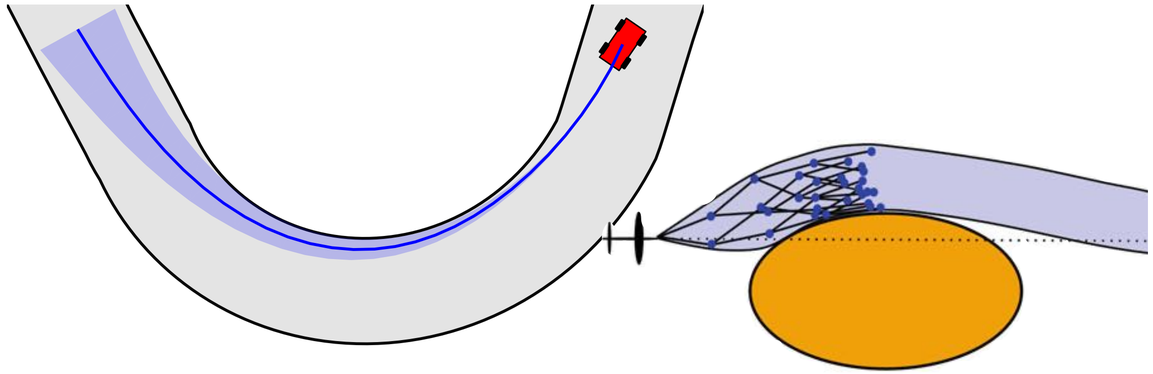
loss/cost → $l(\theta, \xi)$ ← uncertain

$$\min_{\theta} \; \mathbb{E}_{\xi \sim \hat{P}}$$

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, sim2real, off-policy/offline RL

Robust optimization (RO)
(robust control, games)



Optimization under undertainty

Empirical risk minimization (ERM)
(sample average approximation (SAA))

loss/cost

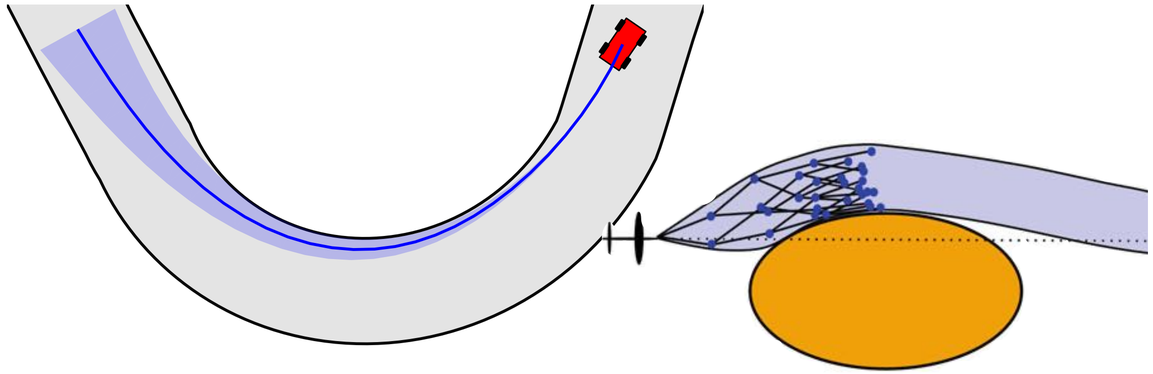
uncertain

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

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, sim2real, off-policy/offline RL

Robust optimization (RO)
(robust control, games)



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

Optimization under undertainty

Empirical risk minimization (ERM)
(sample average approximation (SAA))

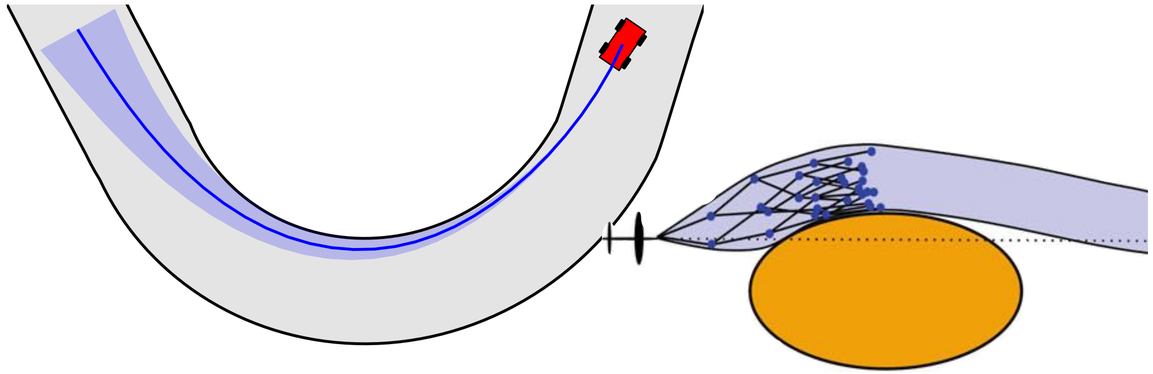
loss/cost → $l(\theta, \xi)$ ← uncertain

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

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, sim2real, off-policy/offline RL

Robust optimization (RO)
(robust control, games)



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

- Do well in the worst case

Optimization under undertainty

Empirical risk minimization (ERM)
(sample average approximation (SAA))

loss/cost

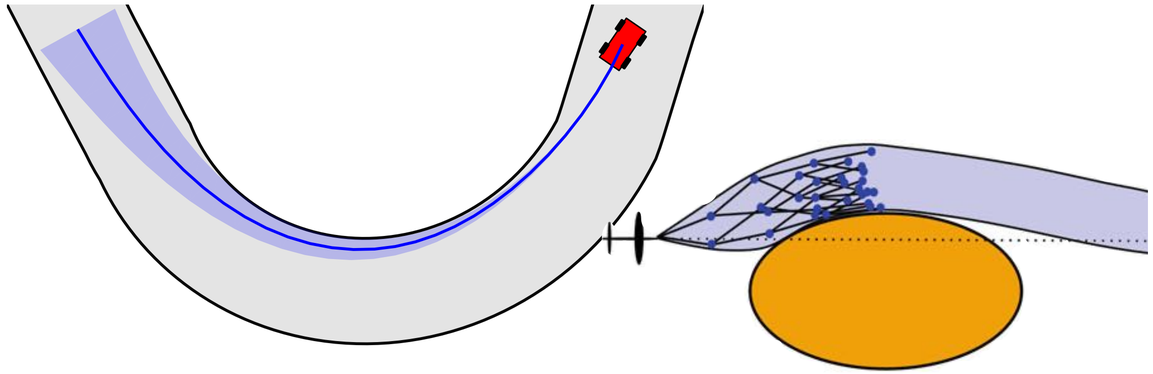
uncertain

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

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, sim2real, off-policy/offline RL

Robust optimization (RO)
(robust control, games)



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

- Do well in the worst case
- Strength: robustness

Optimization under undertainty

Empirical risk minimization (ERM)
(sample average approximation (SAA))

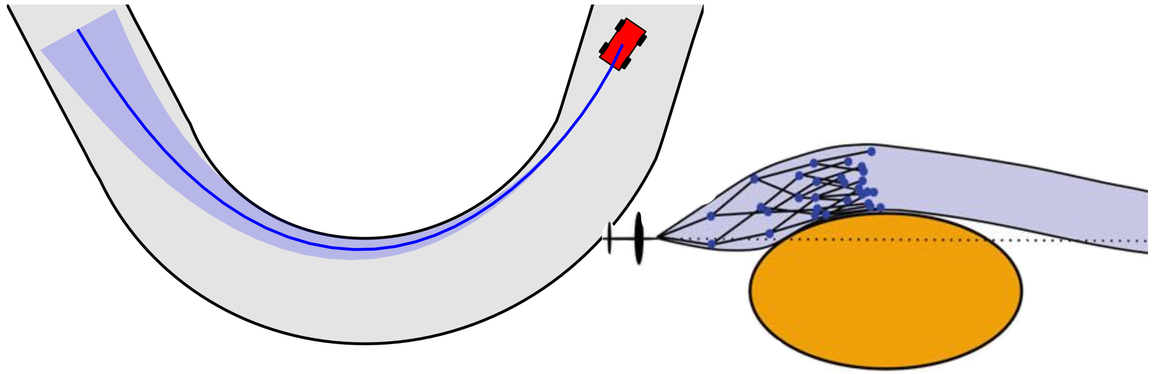
loss/cost → $l(\theta, \xi)$ ← uncertain

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

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, sim2real, off-policy/offline RL

Robust optimization (RO)
(robust control, games)



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

- 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

Various robustness formulation in optimization

Various robustness formulation in optimization

- Nominal constraint: $c(z) \leq 0$

Various robustness formulation in optimization

- Nominal constraint: $c(z) \leq 0$
- Robust constraint: $c(z, e) \leq 0, \forall e \in \mathscr{E}$

Various robustness formulation in optimization

- Nominal constraint: $c(z) \leq 0$
- Robust constraint: $c(z, e) \leq 0, \forall e \in \mathscr{E}$
- Chance constraint: $P(c(z, e) \leq 0) \geq 0.95$

Various robustness formulation in optimization

- Nominal constraint: $c(z) \leq 0$
- Robust constraint: $c(z, e) \leq 0, \forall e \in \mathscr{E}$
- Chance constraint: $P(c(z, e) \leq 0) \geq 0.95$
- Scenario optimization: $c(z, e_i) \leq 0, \text{for } i = 1,2,...,\text{N}$

Worst-case robust optimization (RO)

Worst-case robust optimization (RO)

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

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

- A simple case

$$\text{(P)} \min_{\theta} \sup_{\|p-\hat{p}\|_{\alpha} \leq \epsilon} p^T \theta$$

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

- A simple case

$$\text{(P)} \min_{\theta} \sup_{\|p-\hat{p}\|_\alpha \leq \epsilon} p^T \theta$$

- But this formulation actually contains infinitely many optimization problems!

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

- A simple case

$$\text{(P)} \min_{\theta} \sup_{\|p-\hat{p}\|_\alpha \leq \epsilon} p^T \theta$$

- But this formulation actually contains infinitely many optimization problems!
- We solve this RO using the (strong) *duality of convex optimization:*
  The primal (worst-case) RO problem (P) is equivalent to the dual problem

Worst-case robust optimization (RO)

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

- That looks robust, but how do I solve it?

- A simple case

$$\text{(P)} \min_{\theta} \sup_{\|p-\hat{p}\|_\alpha \leq \epsilon} p^T \theta$$

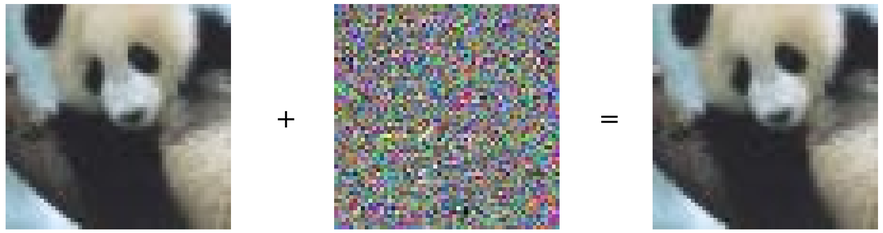
- But this formulation actually contains infinitely many optimization problems!
- We solve this RO using the (strong) *duality of convex optimization:*
  The primal (worst-case) RO problem (P) is equivalent to the dual problem

$$\text{(D)} \min_{\theta} \hat{p}^T \theta + \epsilon \|\theta\|_2$$

  i.e., (P) = (D).

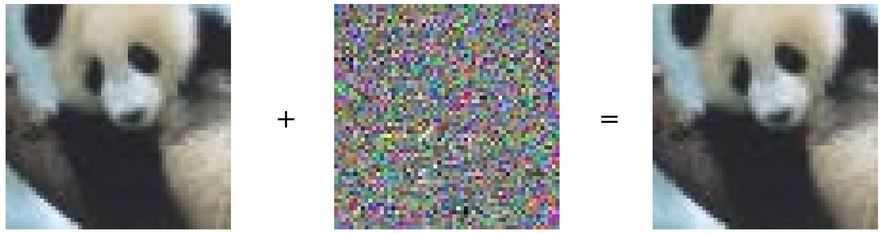
Application of worst-case RO: adversarial robustness of DNN

Application of worst-case RO: adversarial robustness of DNN



x $\quad$ δ

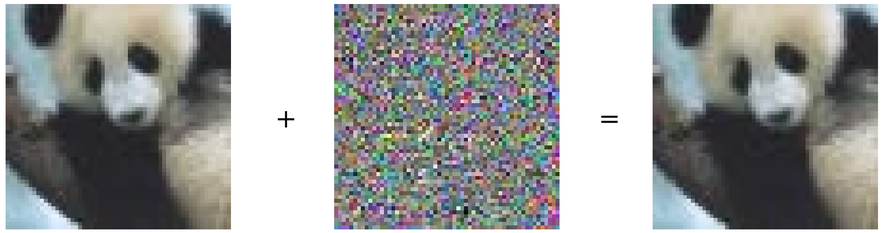
Application of worst-case RO: adversarial robustness of DNN



x δ

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \leq \epsilon} l(f_\theta(x_i + \delta), y_i)$$

Application of worst-case RO: adversarial robustness of DNN



x δ

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \leq \epsilon} l(f_\theta(x_i + \delta), y_i)$$

- Since deep model typically don't have the nice structure we have seen in the last slide, we are often left to performing SGD and hope for the alignment of the stars.

Application of worst-case RO: adversarial robustness of DNN


x

+


δ

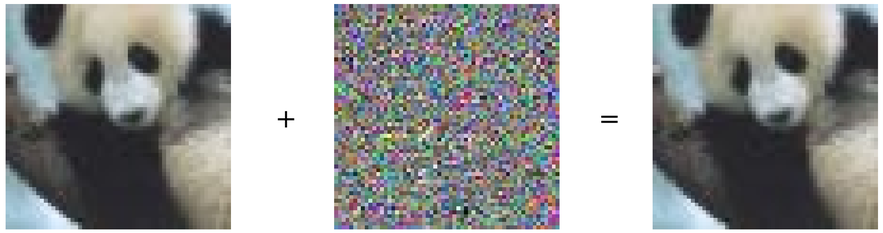
=

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \leq \epsilon} l(f_\theta(x_i + \delta), y_i)$$

- Since deep model typically don't have the nice structure we have seen in the last slide, we are often left to performing SGD and hope for the alignment of the stars.

- Projected gradient descent: $\delta \leftarrow \text{proj}(\delta + \alpha \nabla_\delta l)$

Application of worst-case RO: adversarial robustness of DNN

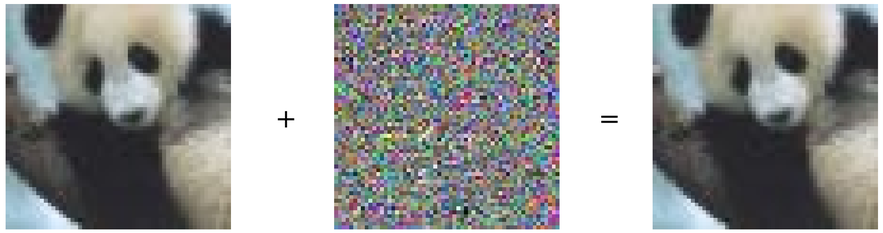


x δ

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \le \epsilon} l(f_\theta(x_i + \delta), y_i)$$

- Since deep model typically don't have the nice structure we have seen in the last slide, we are often left to performing SGD and hope for the alignment of the stars.
- Projected gradient descent: $\delta \leftarrow \text{proj}(\delta + \alpha \nabla_\delta l)$
  - sample (x_i, y_i)

Application of worst-case RO: adversarial robustness of DNN

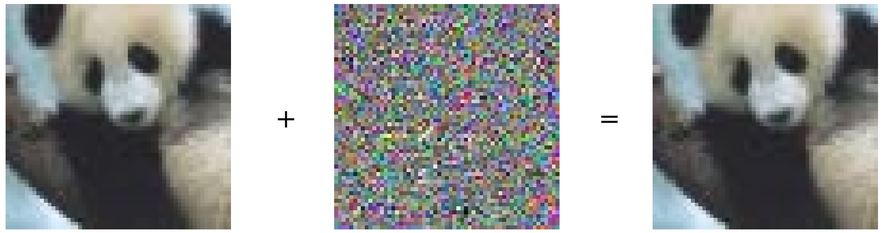


x δ

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \leq \epsilon} l(f_\theta(x_i + \delta), y_i)$$

- Since deep model typically don't have the nice structure we have seen in the last slide, we are often left to performing SGD and hope for the alignment of the stars.
- Projected gradient descent: $\delta \leftarrow \text{proj}(\delta + \alpha \nabla_\delta l)$
  - sample (x_i, y_i)
  - perform gradient ascent for the inner maximization problem over δ

Application of worst-case RO: adversarial robustness of DNN



x δ

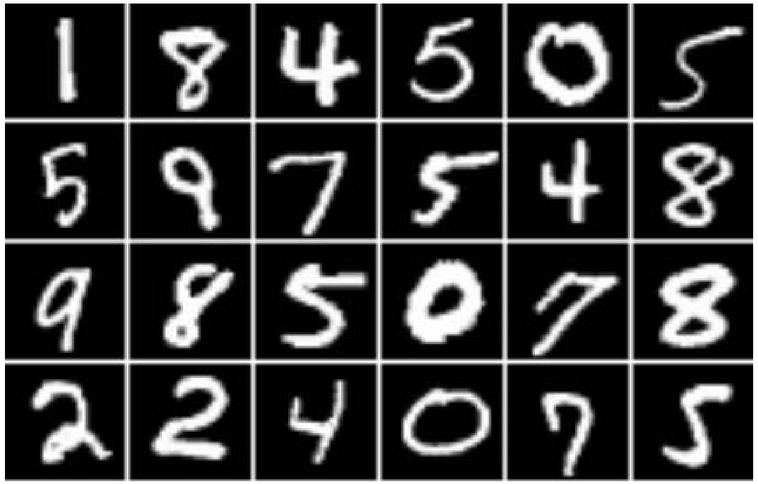
$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{\|\delta\| \leq \epsilon} l(f_\theta(x_i + \delta), y_i)$$

- Since deep model typically don't have the nice structure we have seen in the last slide, we are often left to performing SGD and hope for the alignment of the stars.
- Projected gradient descent: $\delta \leftarrow \text{proj}(\delta + \alpha \nabla_\delta l)$
  - sample (x_i, y_i)
  - perform gradient ascent for the inner maximization problem over δ
  - project back to the norm-ball

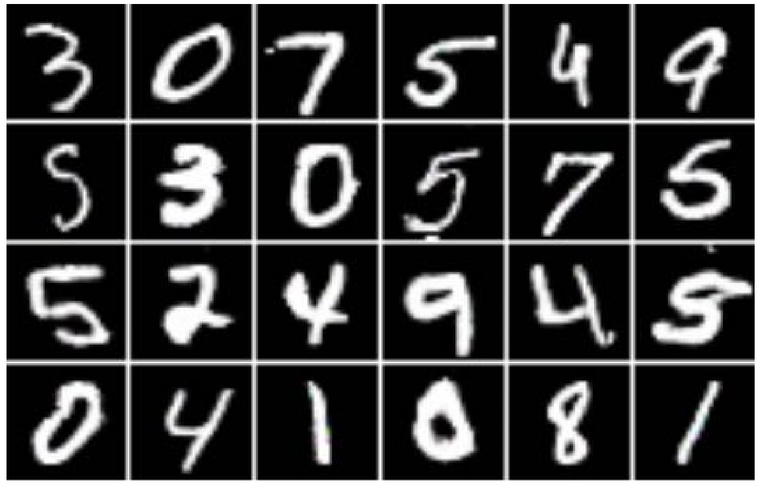
Image credit: Goodfellow et al. 2015

Distributional Robustness

How do we compare two sets of data?



Observed MNIST handwritten digits. X.

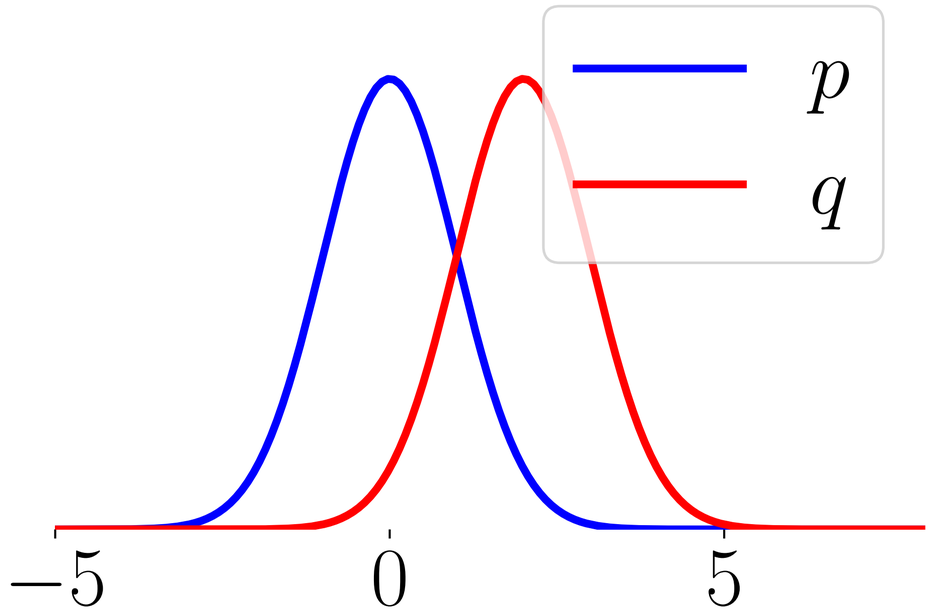


Generated images from a model. Y.

Is Y similar to X?

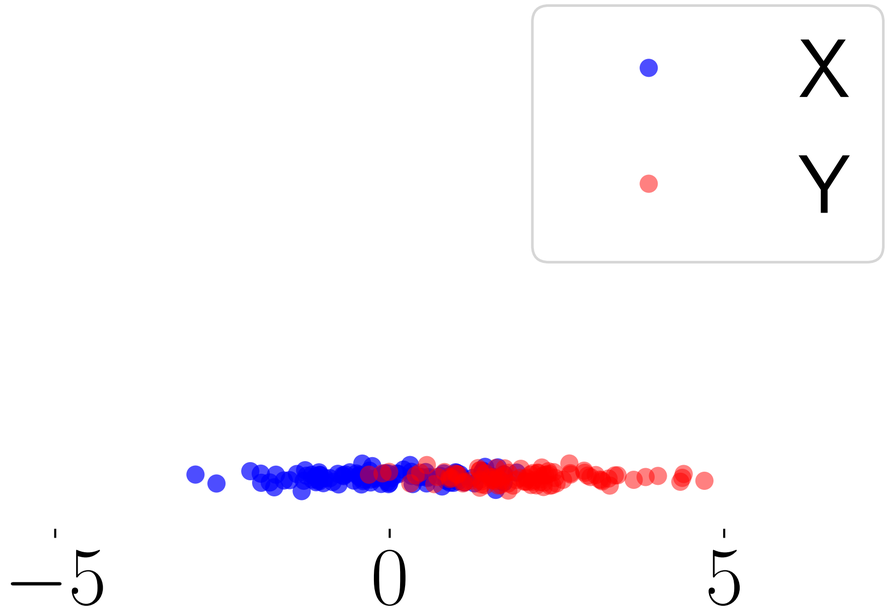
- Distance between distributions can be used to train generative models.

Case 1: Simple Mean Shift in 1D



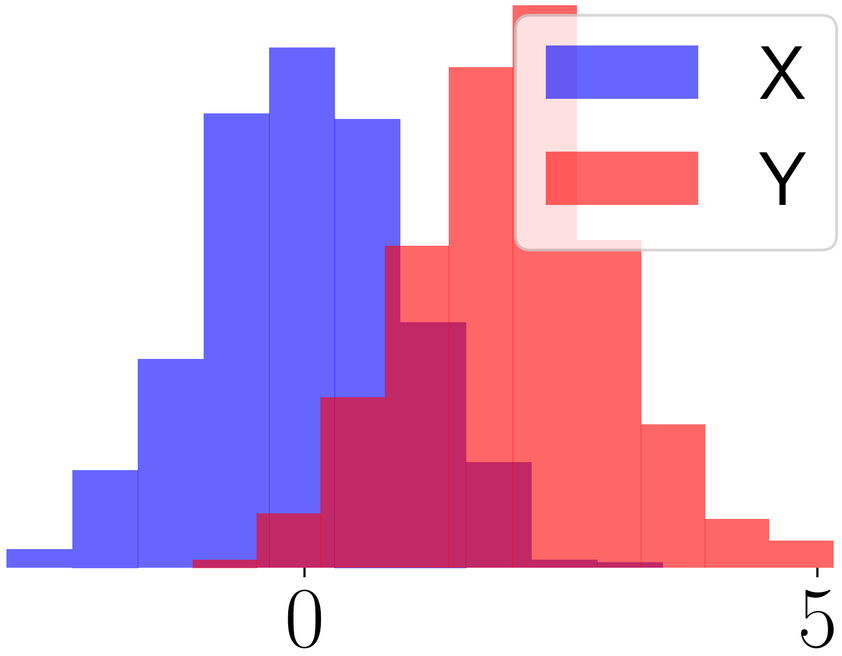
- Two Gaussian distributions.

Case 1: Simple Mean Shift in 1D



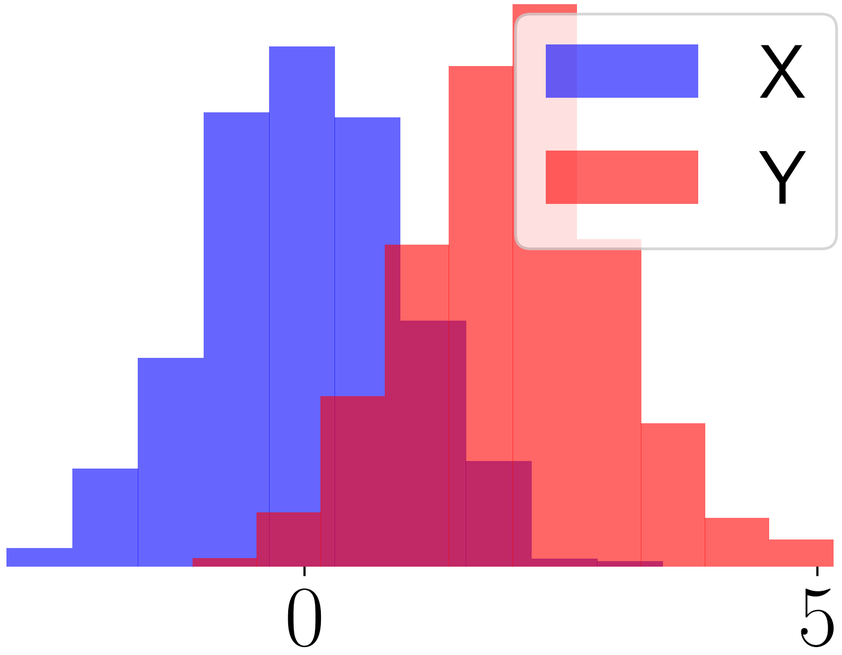
- We have only samples $\mathsf{X} \sim p$ and $\mathsf{Y} \sim q$.
- $\mathsf{X} = \{x_1, \ldots, x_n\}$ and $\mathsf{Y} = \{y_1, \ldots, y_n\}$. Sets of numbers.

Case 1: Simple Mean Shift in 1D



- We have only samples $\mathsf{X} \sim p$ and $\mathsf{Y} \sim q$.
- $\mathsf{X} = \{x_1, \ldots, x_n\}$ and $\mathsf{Y} = \{y_1, \ldots, y_n\}$. Sets of numbers.

Case 1: Simple Mean Shift in 1D

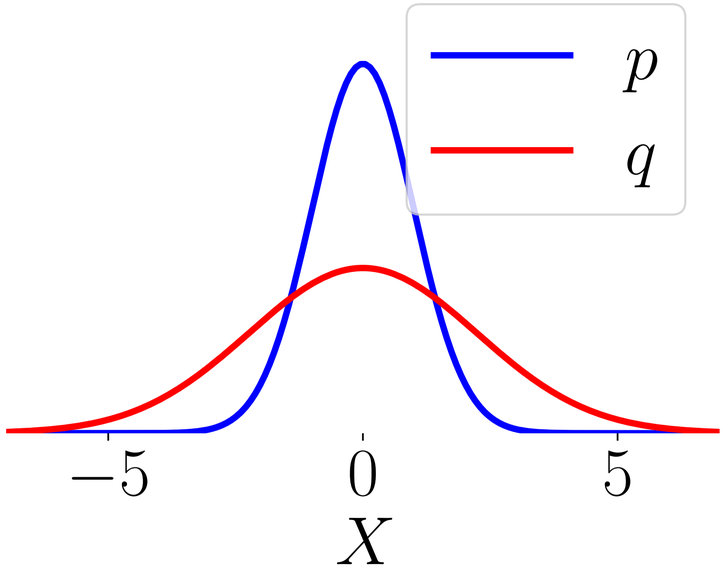


- Assume no differece in high-order moments.
- "Distance" = difference in the means. T-test.

$$\text{(population) } D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$$

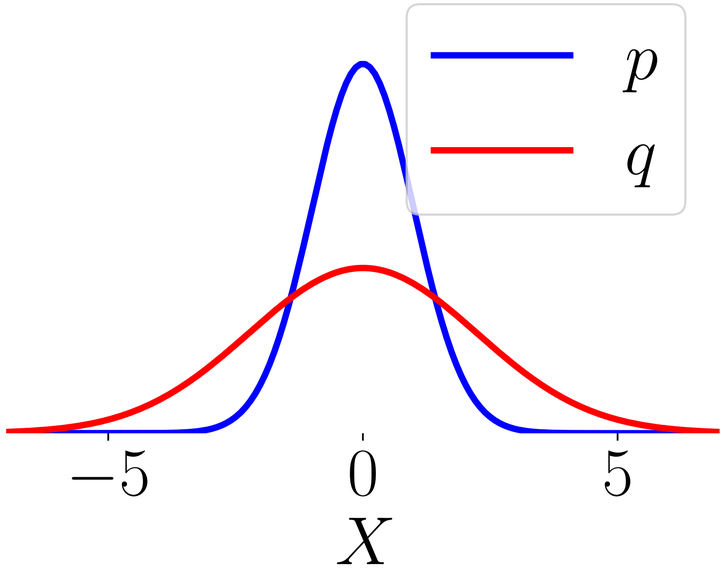
$$\text{(empirical) } \hat{D}_1(\mathsf{X}, \mathsf{Y}) = \left|\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{j=1}^{n} y_j\right|$$

Case 2: Same Mean, Different Variances



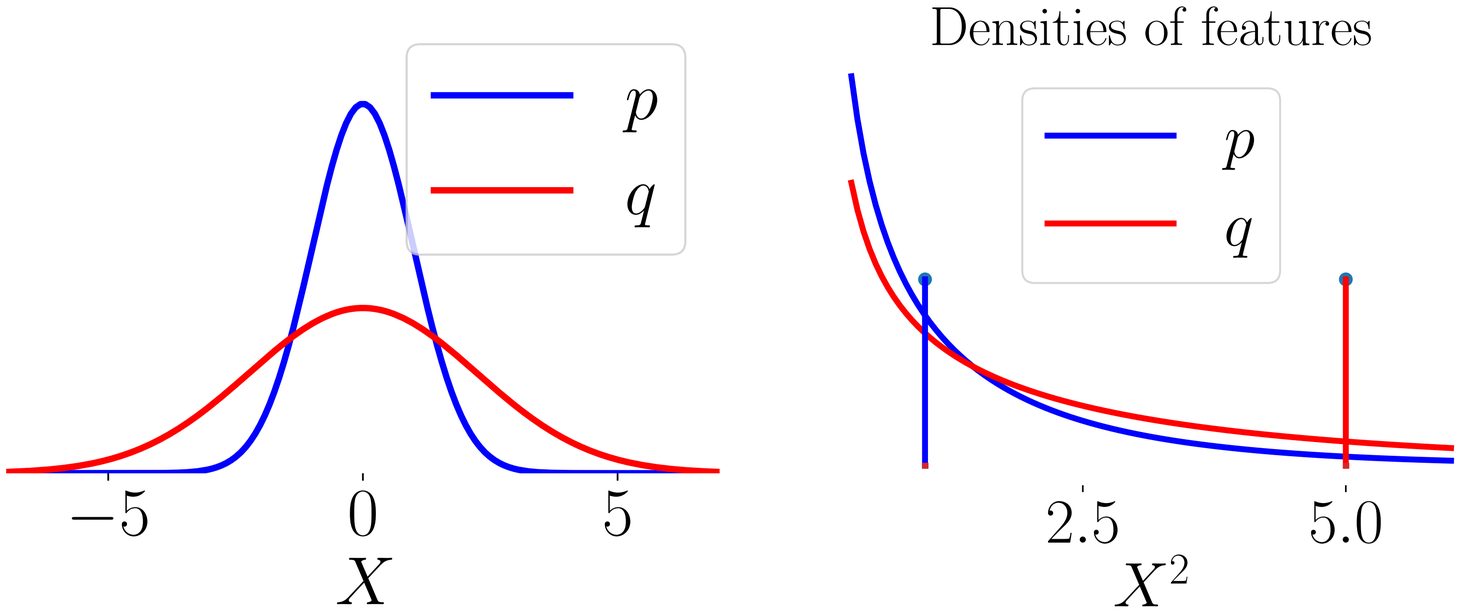
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

Case 2: Same Mean, Different Variances



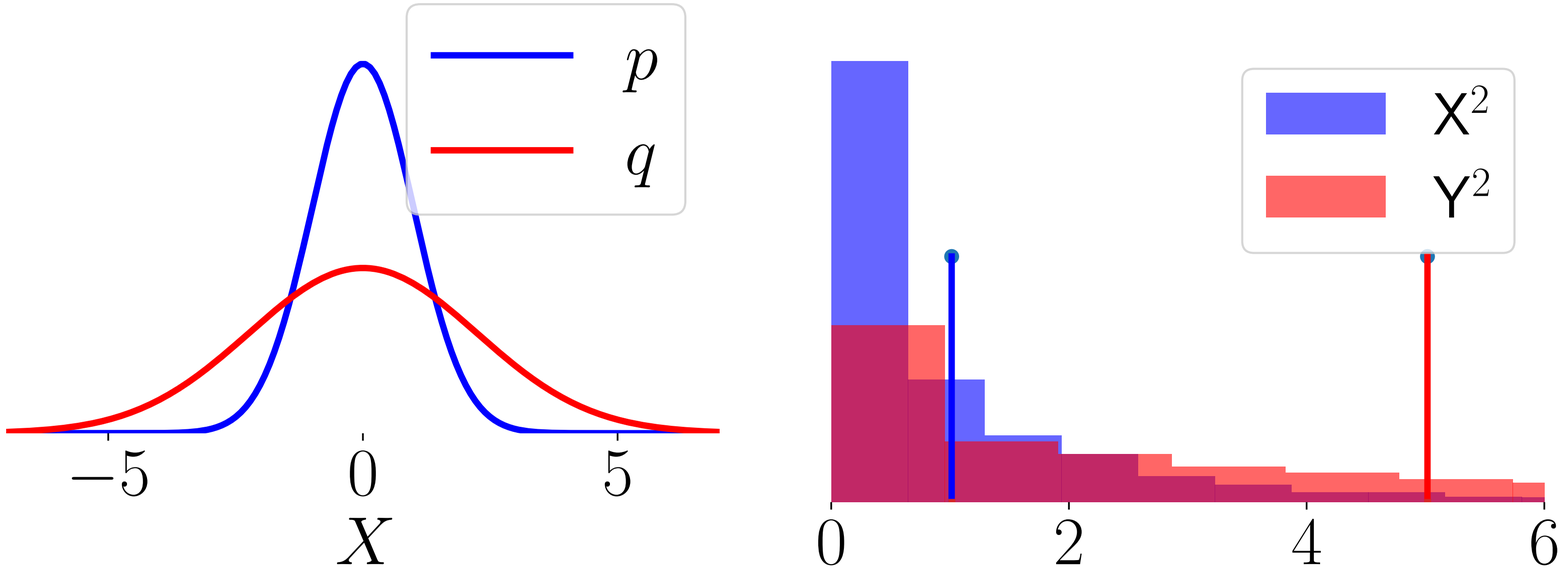
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

Case 2: Same Mean, Different Variances



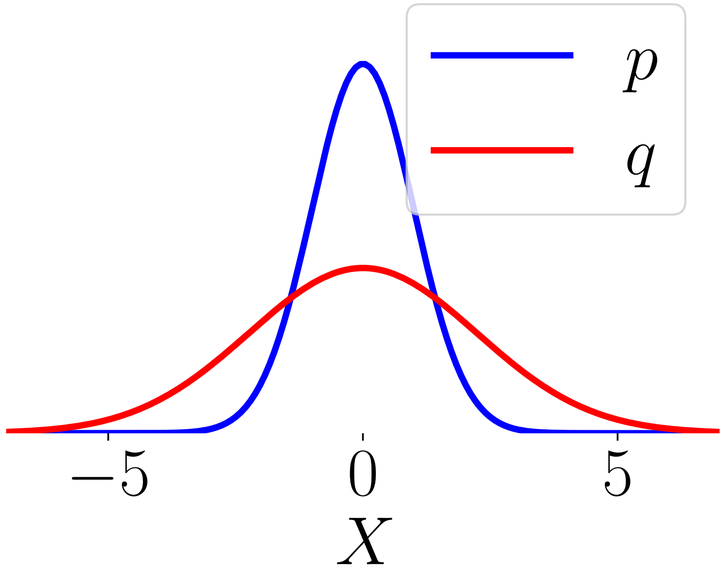
- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

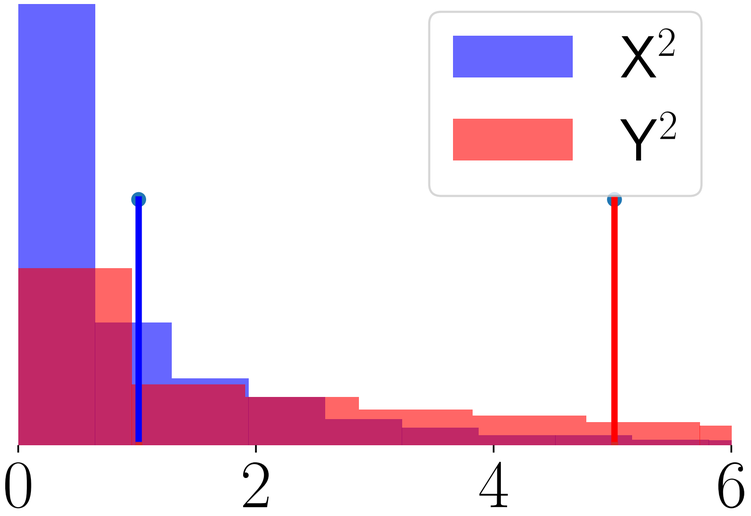
Case 2: Same Mean, Different Variances



- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

Case 2: Same Mean, Different Variances



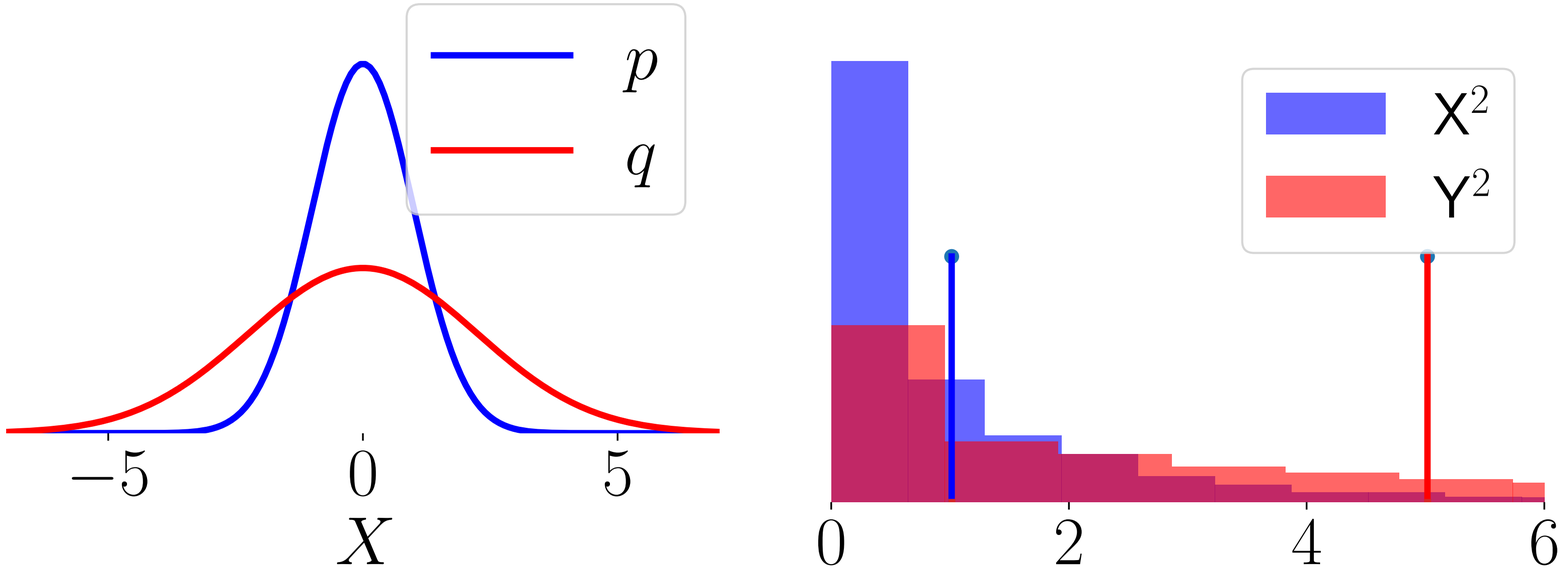


- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.
- New "distance":

$$D_2(p, q) = \left\|\mathbb{E}_{X\sim p}[\phi(X)] - \mathbb{E}_{Y\sim q}[\phi(X)]\right\|,$$

where $\phi(x) = (x, x^2)^\top$.

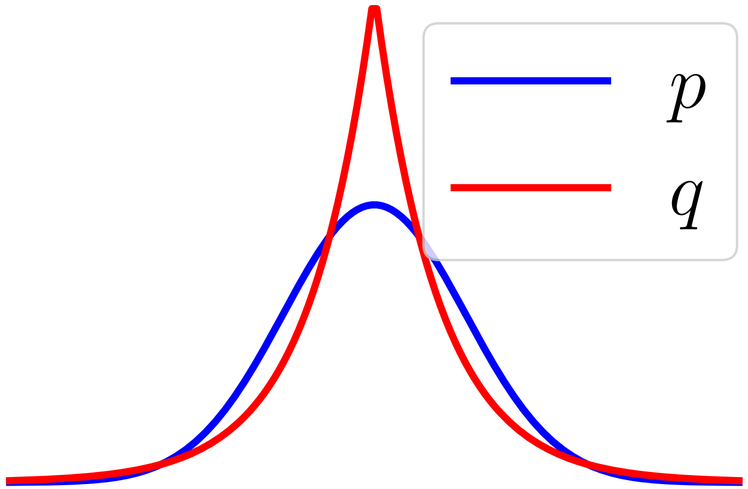
Case 2: Same Mean, Different Variances



- $D_1(p, q) := |\mathbb{E}_{X\sim p}[X] - \mathbb{E}_{Y\sim q}[Y]|$ cannot detect the difference. Why?
- **Idea**: look at difference in means of features $\phi(\cdot)$ of X and Y.

$$D_2(p, q) = \left\| \begin{pmatrix} \mathbb{E}_{X\sim p}[X] \\ \mathbb{E}_{X\sim p}[X^2] \end{pmatrix} - \begin{pmatrix} \mathbb{E}_{X\sim q}[X] \\ \mathbb{E}_{X\sim q}[X^2] \end{pmatrix} \right\|.$$

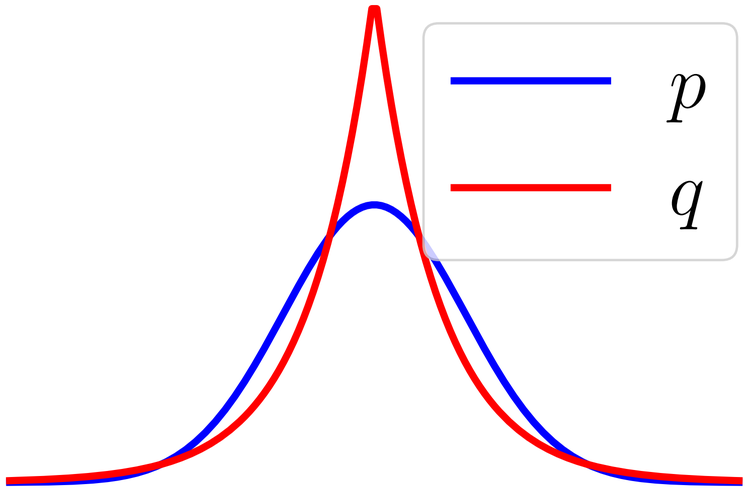
Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

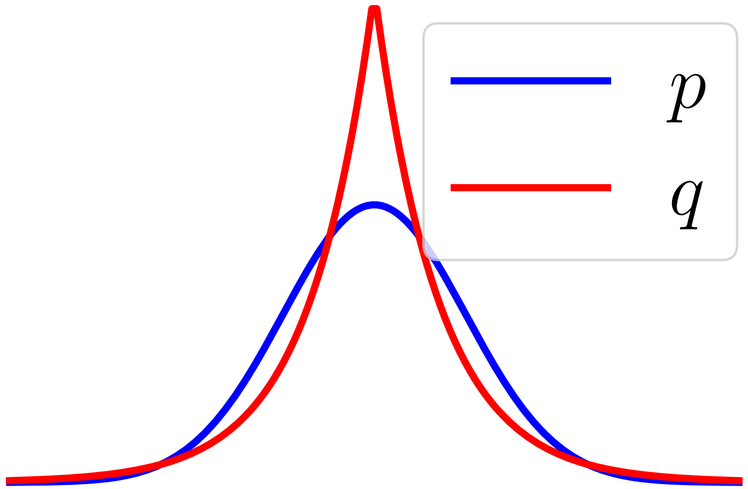
Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- **Solution**: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

Case 3: Difference in High-Order Moments



- p = Gaussian distribution,
  q = Laplace distribution.
- Same mean **and** variance.
- D_2 fails.

- $\phi(x) = (x, x^2, x^4)^\top$ works. Difference is in kurtosis (4^{th} moment).
- $\phi(x) = (x, x^2, x^4, \cos x, e^x, \ldots)^\top$. But, when to stop?
- Solution: Use an infinite-dimensional feature map $\phi(\cdot)$ with the kernel trick.

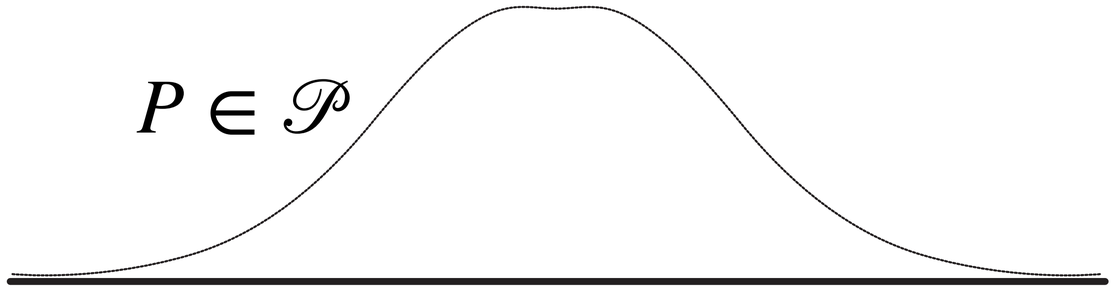
Working with probability measures

Working with probability measures

The probability “simplex” $\mathscr{P}$

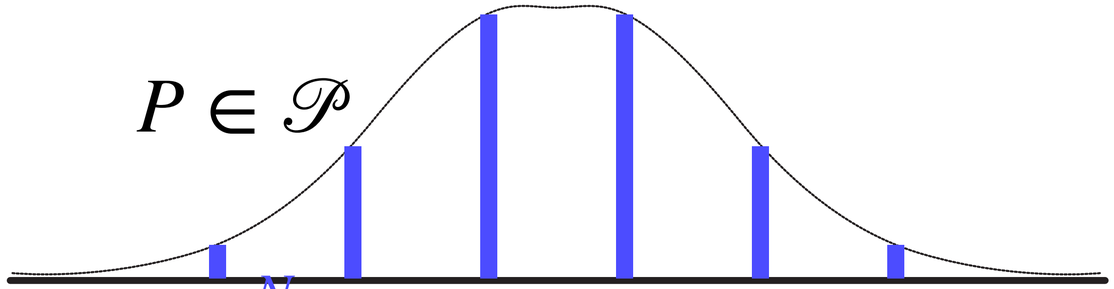
Working with probability measures

The probability “simplex” $\mathscr{P}$

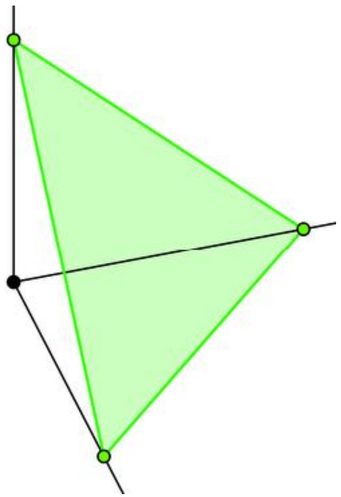


Working with probability measures

The probability “simplex” $\mathscr{P}$

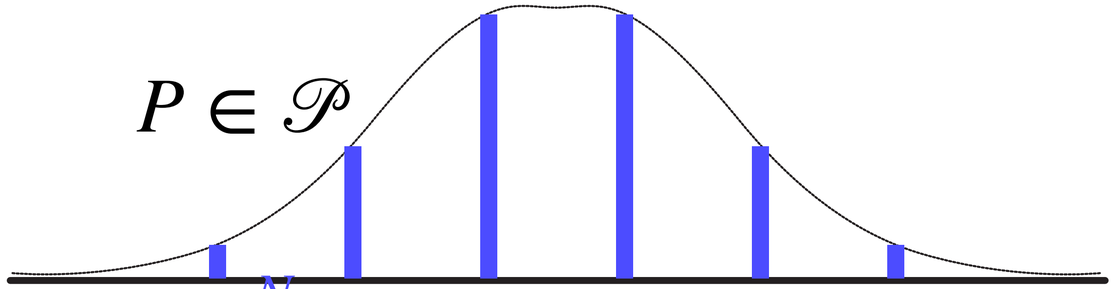


$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$

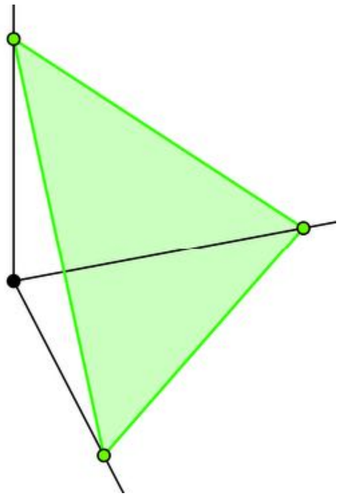


Working with probability measures

The probability "simplex" $\mathscr{P}$



$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$



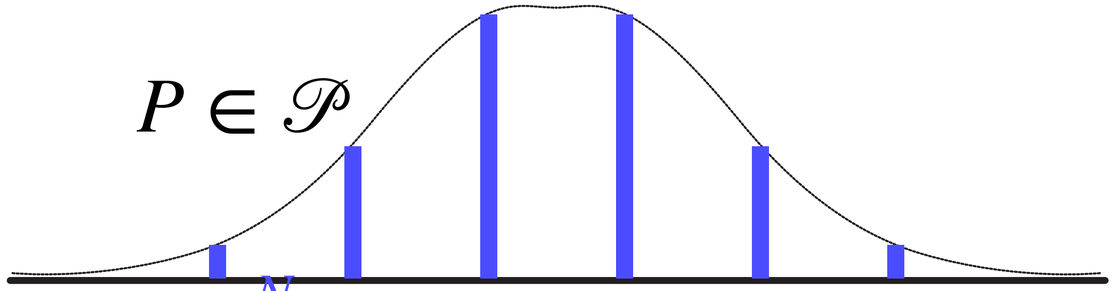
Empirical data distribution

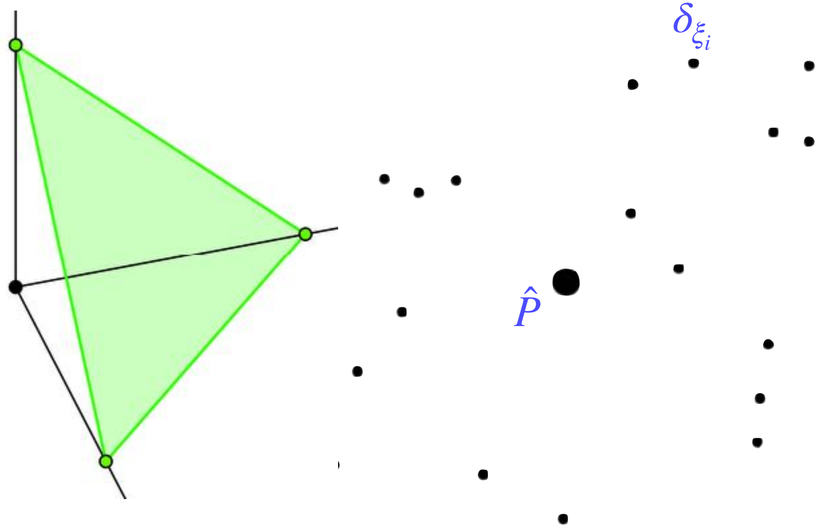
Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$

Working with probability measures

The probability "simplex" $\mathscr{P}$





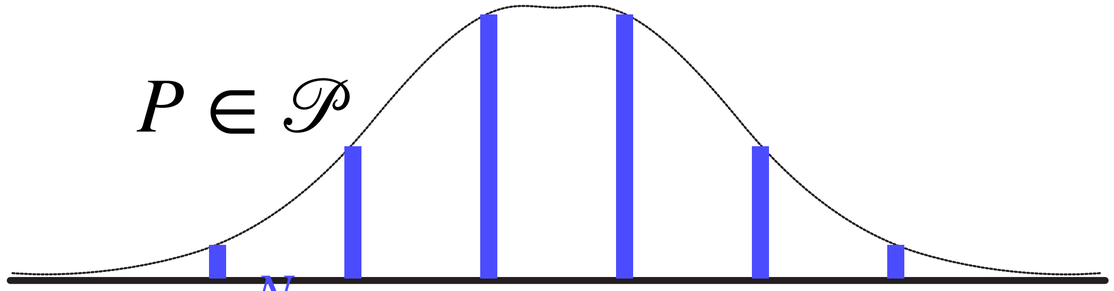
Empirical data distribution

Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

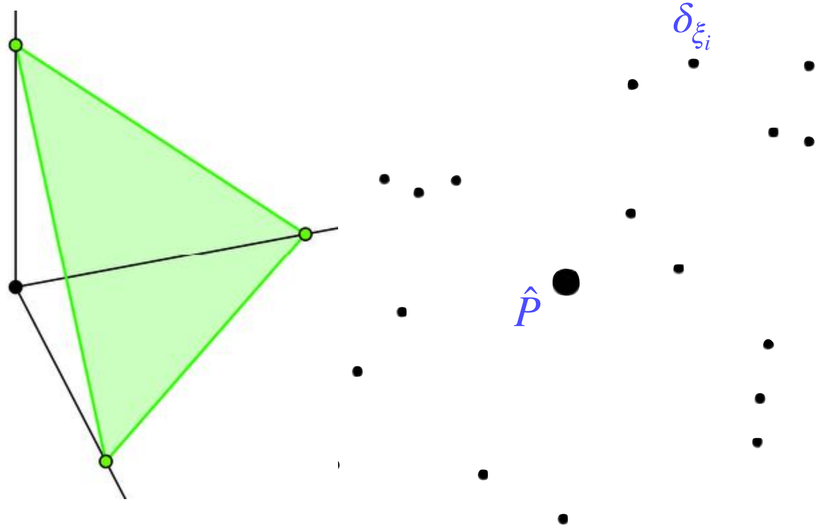
empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$

Working with probability measures

The probability "simplex" $\mathscr{P}$



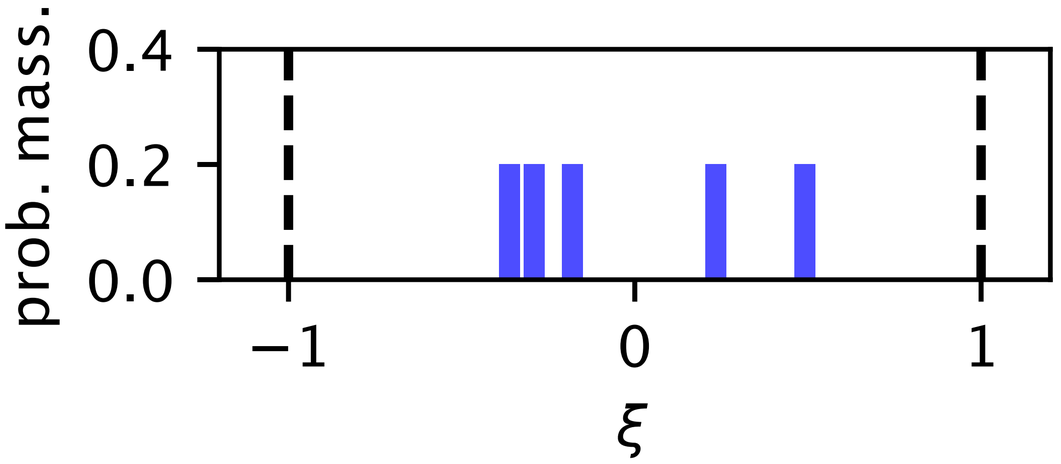
$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$



Empirical data distribution

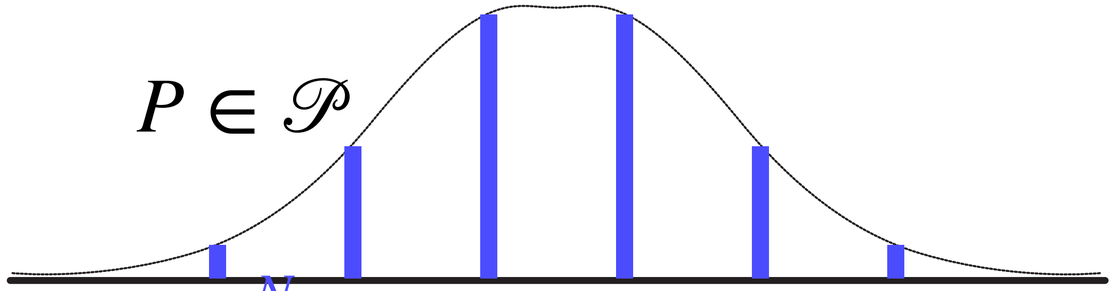
Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$

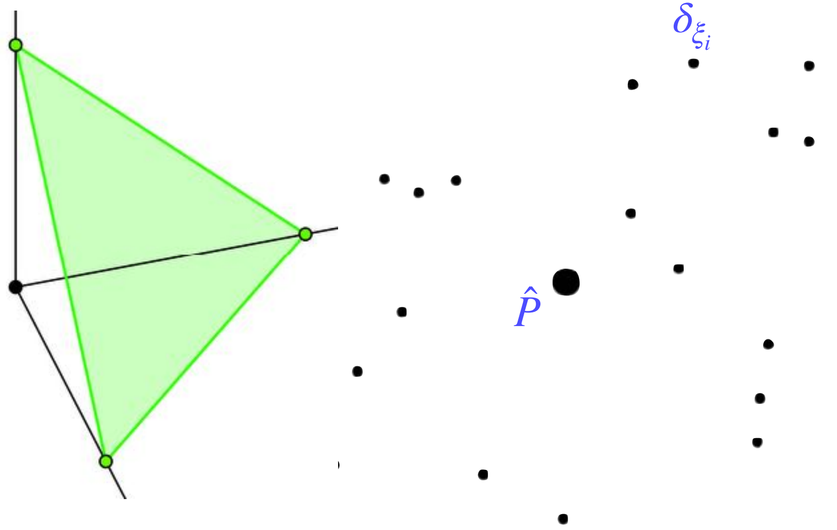


Working with probability measures

The probability "simplex" $\mathscr{P}$



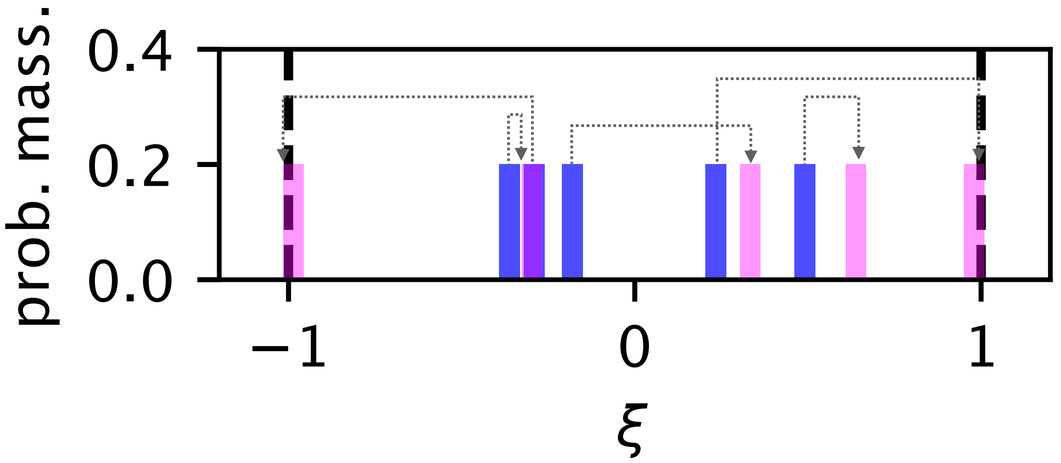
$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$



Empirical data distribution

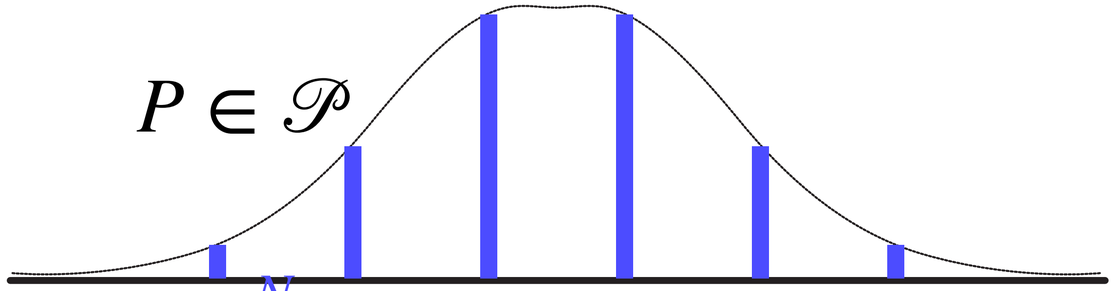
Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$

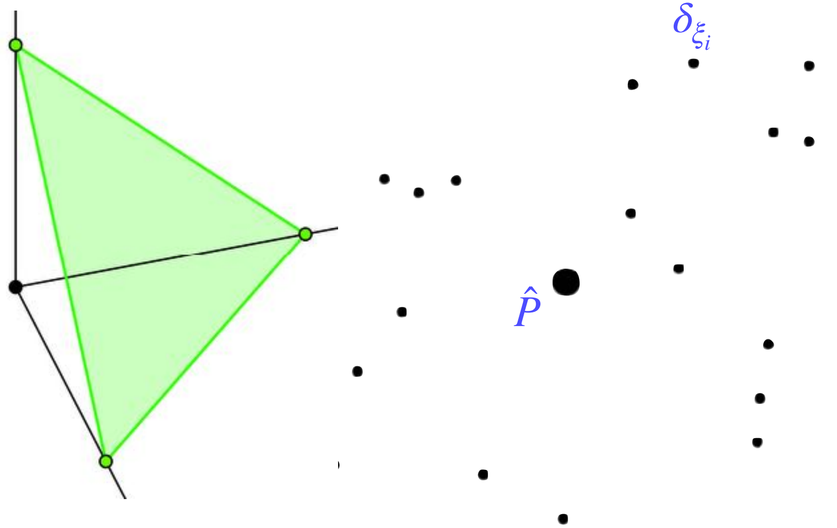


Working with probability measures

The probability "simplex" $\mathscr{P}$



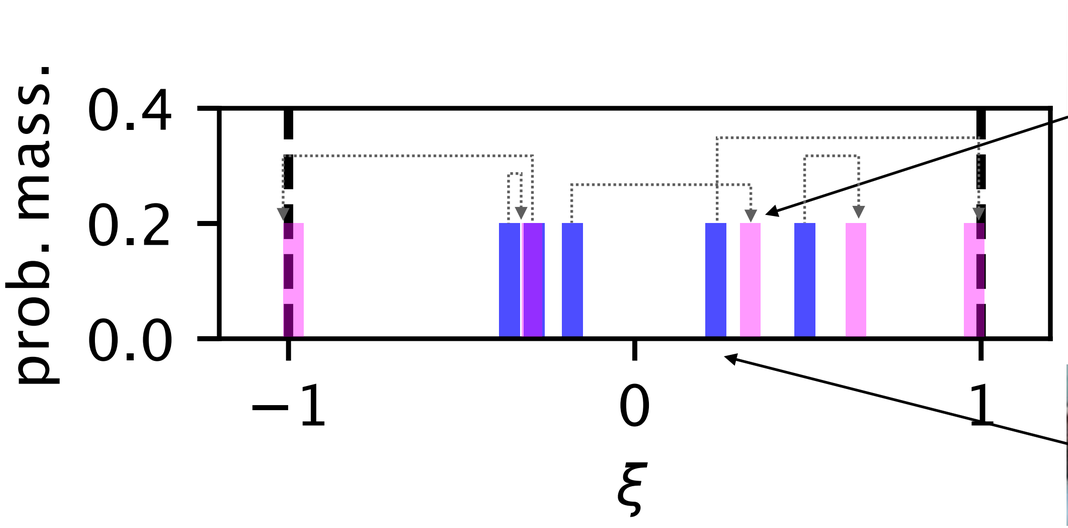
$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$



Empirical data distribution

Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

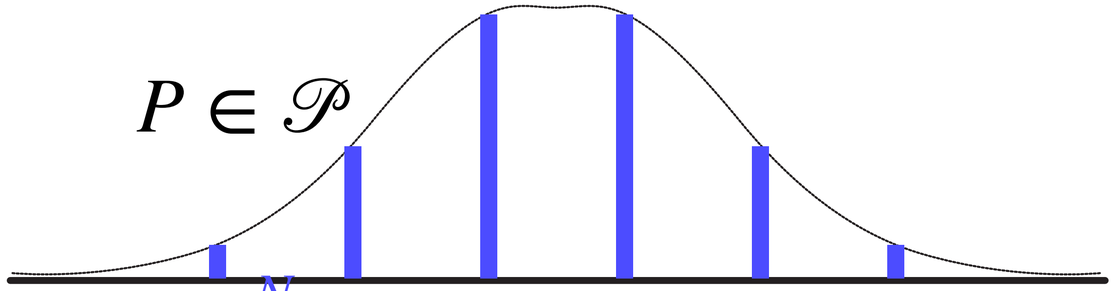
empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$



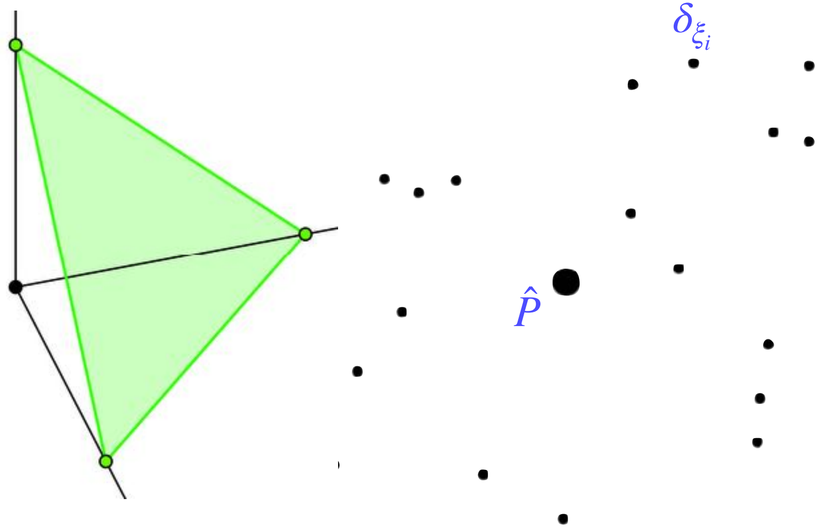
Adversarial attack: Goodfellow et al. 2015

Working with probability measures

The probability "simplex" $\mathscr{P}$



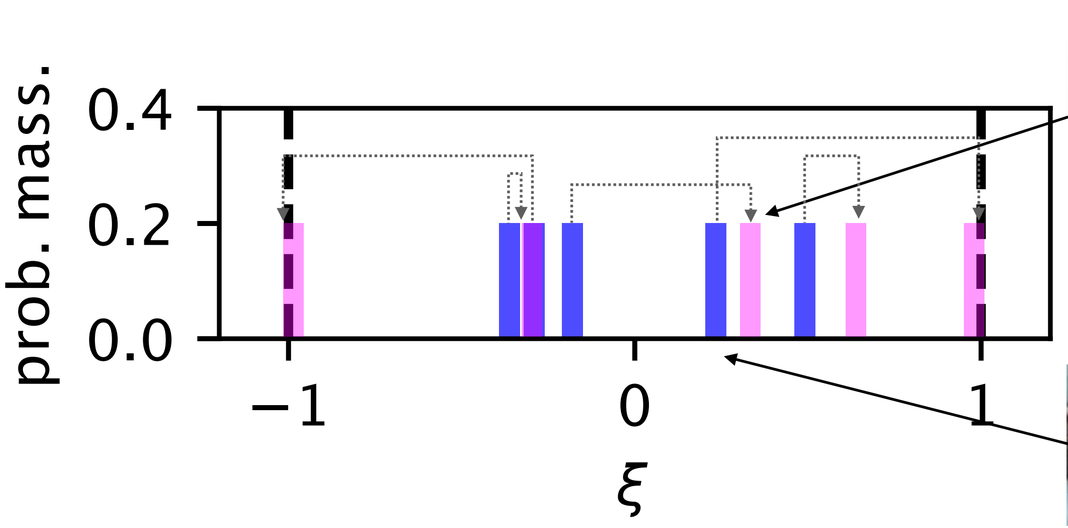
$$\sum_{i=1}^{N} \gamma_i \, \delta_{\xi_i} \qquad \gamma \in \Delta_d \subset \mathbb{R}^d$$



Empirical data distribution

Given data samples: $\xi_1, \xi_2, \ldots, \xi_N$

empirical data distribution $\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{\xi_i}$



Adversarial attack: Goodfellow et al. 2015

P might only differ slightly from $\hat{P}$, but can break the system.

Wasserstein distance

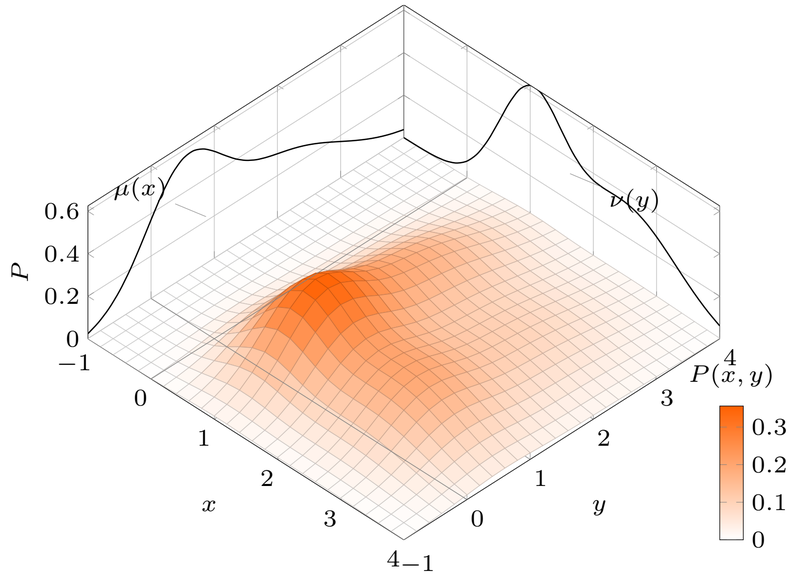
The optimal value of the Kantorovich problem

$$\mathcal{W}_p(P,Q) = \inf_{\pi \in \Pi(P,Q)} \left\{ \int \|u-v\|^p \, d\pi(u,v) \right\}^{1/p}$$

where $p \geq 1$, π is a *t*ransport plan (coupling) that belongs to the set of valid plans

$$\Pi(\mu,\nu) = \{\pi \in P(X \times X) : (\pi_0)_{\#}\gamma = \mu, (\pi_1)_{\#}\gamma = \nu, \}.$$

π_0 and π_1 are the two projections onto the marginal distributions.



Wasserstein distance

The optimal value of the Kantorovich problem

$$\mathcal{W}_p(P,Q) = \inf_{\pi \in \Pi(P,Q)} \left\{ \int \|u-v\|^p \, d\pi(u,v) \right\}^{1/p}$$

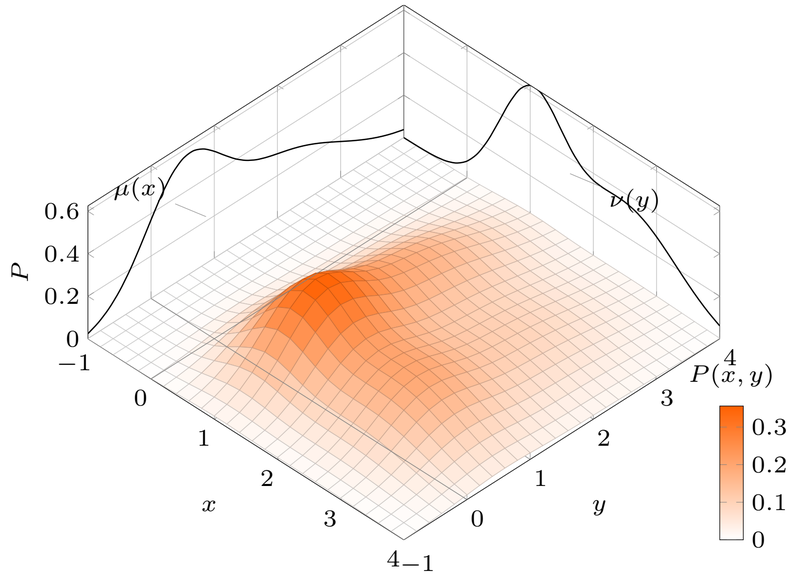
where $p \geq 1$, π is a *t*ransport plan (coupling) that belongs to the set of valid plans

$$\Pi(\mu,\nu) = \{\pi \in P(X \times X) : (\pi_0)_{\#}\gamma = \mu, \, (\pi_1)_{\#}\gamma = \nu, \}.$$

π_0 and π_1 are the two projections onto the marginal distributions.

If $p = 1$, this is equivalent to IPM with $1-$Lipschitz family

$$W_1(P,Q) = \sup_{\text{lip}(f) \leq 1} \int f \, d(P-Q)$$



Wasserstein distance

The optimal value of the Kantorovich problem

$$\mathcal{W}_p(P,Q) = \inf_{\pi \in \Pi(P,Q)} \left\{ \int \|u-v\|^p \, d\pi(u,v) \right\}^{1/p}$$

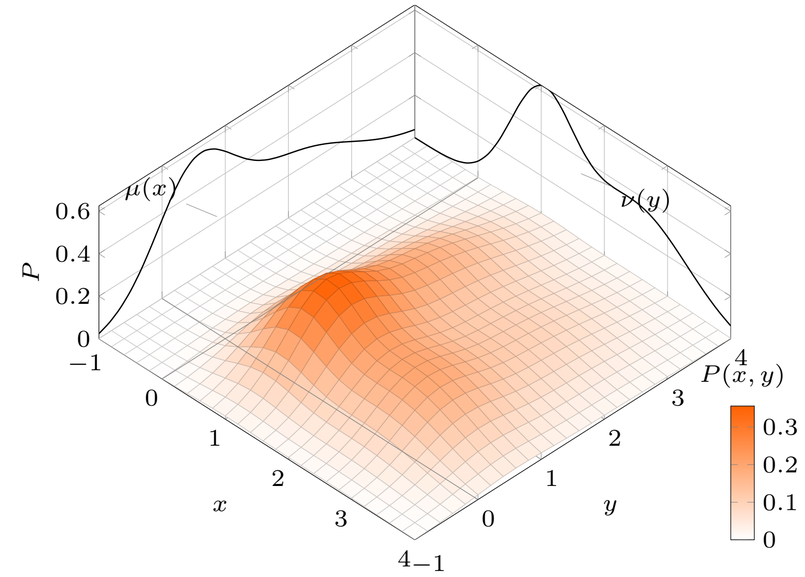
where $p \geq 1$, π is a *t*ransport plan (coupling) that belongs to the set of valid plans

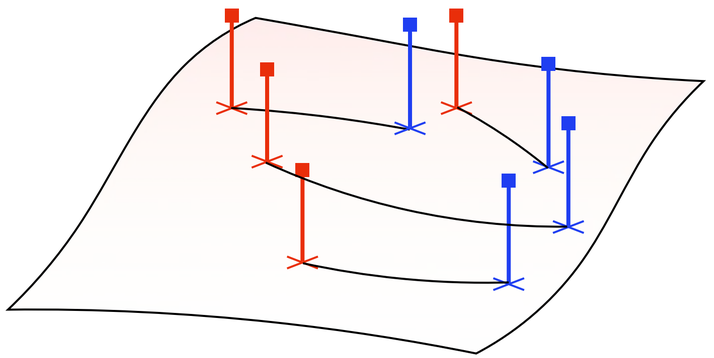
$$\Pi(\mu,\nu) = \{\pi \in P(X \times X) : (\pi_0)_{\#}\gamma = \mu, \, (\pi_1)_{\#}\gamma = \nu, \}.$$

π_0 and π_1 are the two projections onto the marginal distributions.

If $p = 1$, this is equivalent to IPM with $1-$Lipschitz family

$$W_1(P,Q) = \sup_{\text{lip}(f) \leq 1} \int f \, d(P-Q)$$





Application to generative modeling: Wasserstein auto-encoder (WAE)

Application to generative modeling: Wasserstein auto-encoder (WAE)

Minimize the regularized reconstruction loss

$$\inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\left[c\left(X, G(Z)\right)\right] + \lambda \cdot \mathscr{D}_Z(Q_Z, P_Z)\,.$$

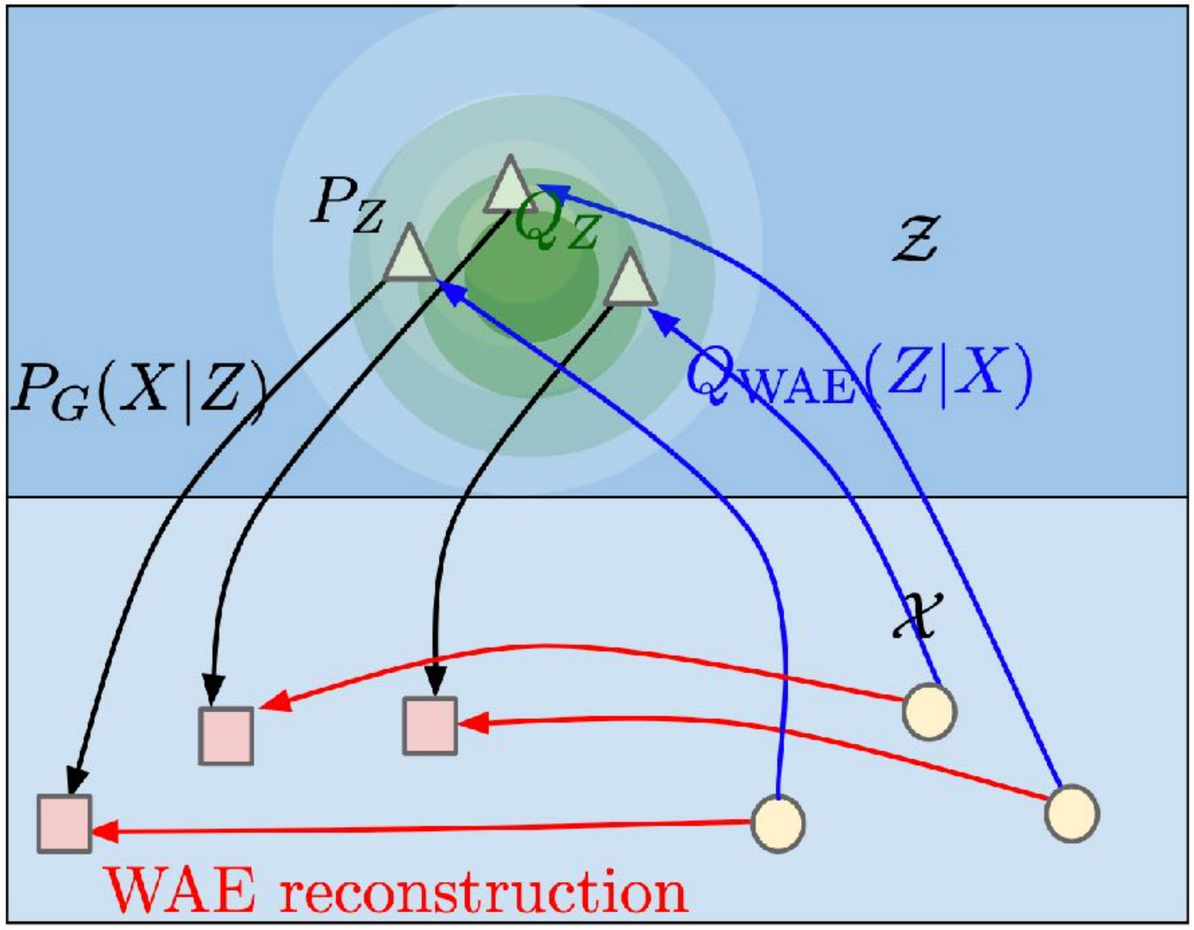
- the posterior distribution $Q(Z|X)$ is the encoder, X is data
- $\mathscr{D}_Z$ is some discrepancy measure in the latent space, e.g., MMD
- P_Z is the fixed target latent distribution, a.k.a. prior, such as a Gaussian or Uniform.
- In practice, we take the marginal distribution P_X to be the empirical data distribution $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$.

Application to generative modeling: Wasserstein auto-encoder (WAE)

Minimize the regularized reconstruction loss

$$\inf_{Q(Z|X)\in\mathcal{Q}} \mathbb{E}_{P_X}\mathbb{E}_{Q(Z|X)}\big[c\big(X, G(Z)\big)\big] + \lambda \cdot \mathscr{D}_Z(Q_Z, P_Z)\,.$$

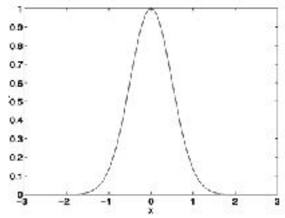
- the posterior distribution $Q(Z|X)$ is the encoder, X is data
- $\mathscr{D}_Z$ is some discrepancy measure in the latent space, e.g., MMD
- P_Z is the fixed target latent distribution, a.k.a. prior, such as a Gaussian or Uniform.
- In practice, we take the marginal distribution P_X to be the empirical data distribution $\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$.



Learning with kernels (the modern way)

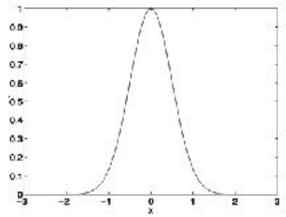
Learning with kernels (the modern way)

- A kernel is a symmetric function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.



Learning with kernels (the modern way)

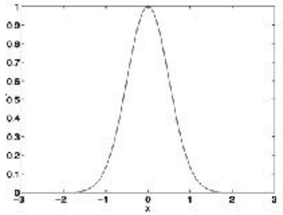
- A kernel is a symmetric function
  $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel
  $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.



- 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 $\mathcal{H}$.

Learning with kernels (the modern way)

- 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 $\mathcal{H}$.

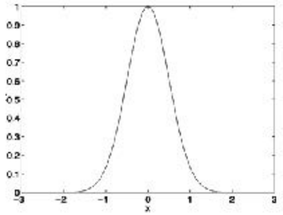
$\sim P$

$\sim Q$

$\mathcal{P}$

Learning with kernels (the modern way)

- 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 $\mathcal{H}$.

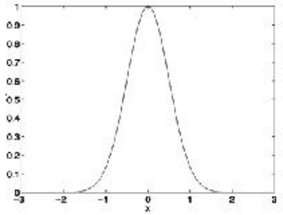
$\sim P$

$\sim Q$

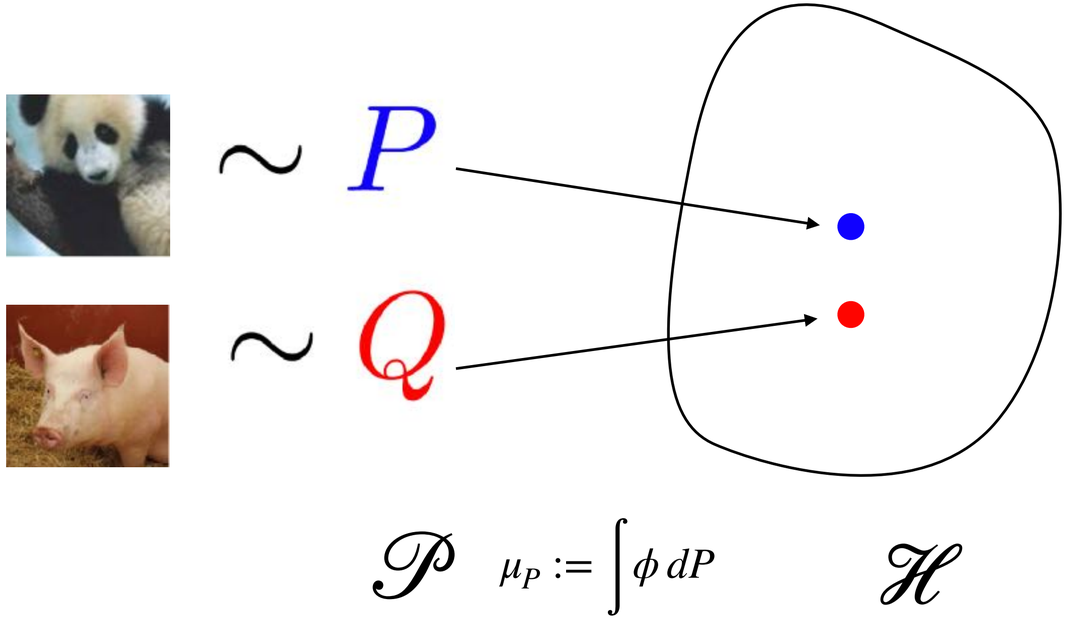
$\mathcal{P} \quad \mu_P := \int \phi \, dP$

Learning with kernels (the modern way)

- A kernel is a symmetric function
  $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel
  $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.

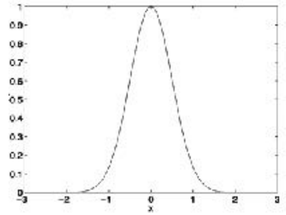


- A p.d. k corresponds to a Hilbert space $\mathcal{H}$ (RKHS), which satisfies the **reproducing property**
  $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$,
  $\phi(x) := k(x, \cdot\,)$ is the **canonical feature** of $\mathcal{H}$.

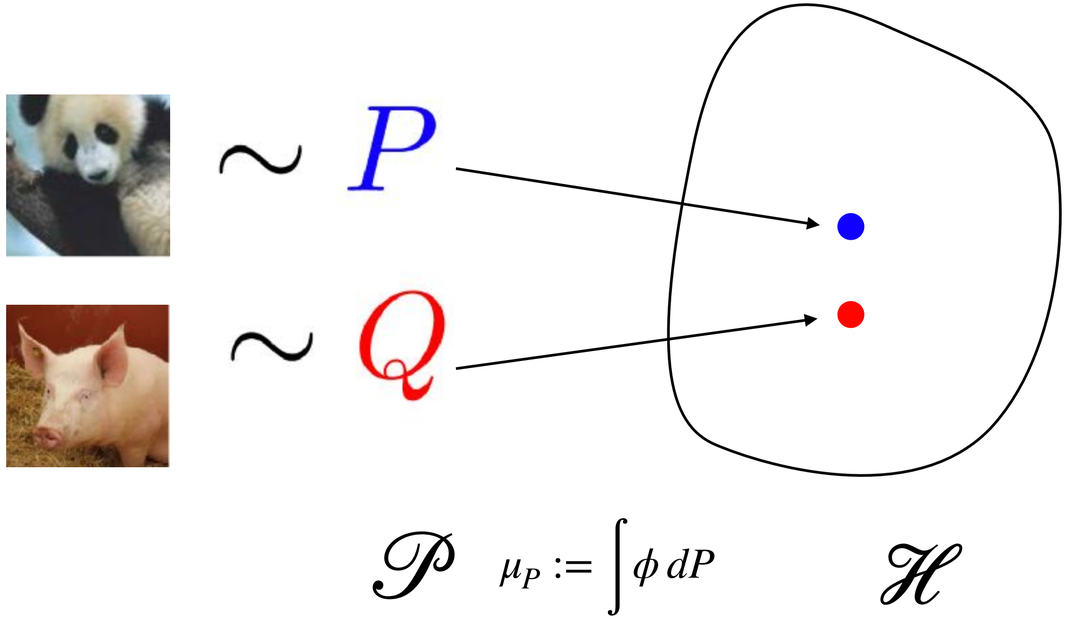


Learning with kernels (the modern way)

- 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 $\mathcal{H}$.

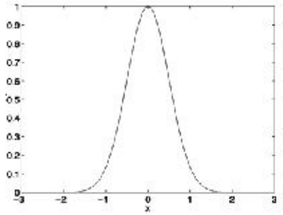


$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

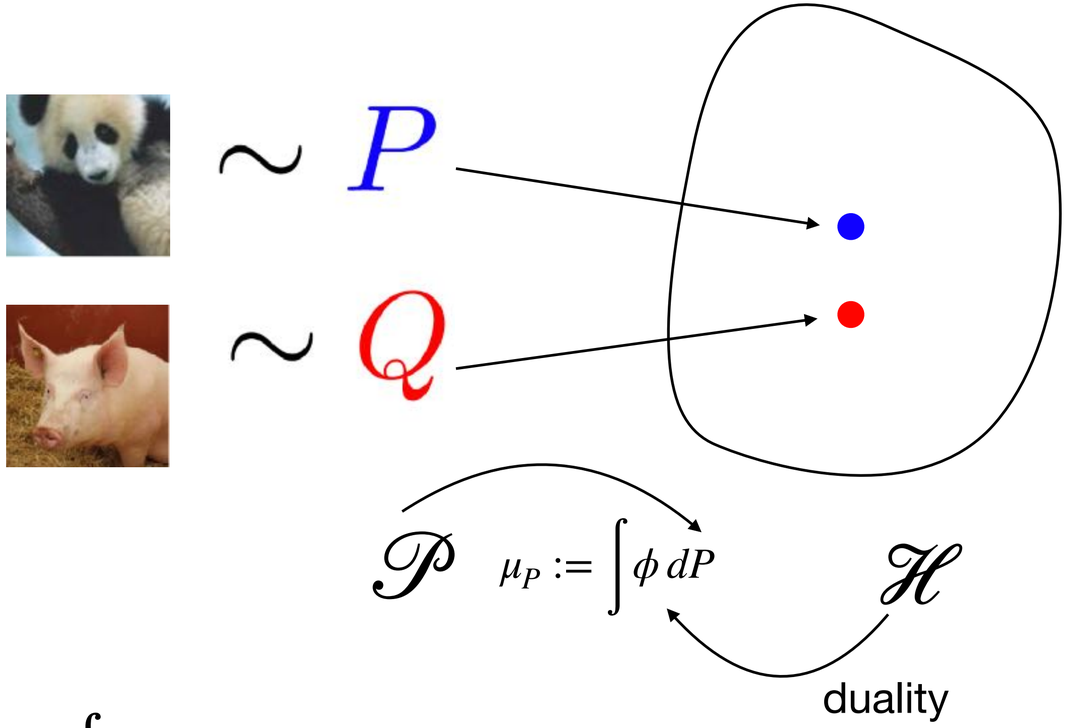
μ can be viewed as a generalized moment vector
e.g., let $\phi(x) = [x, x^2]^\top$

Learning with kernels (the modern way)

- 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 $\mathcal{H}$.

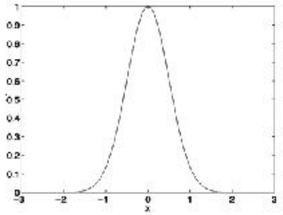


$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

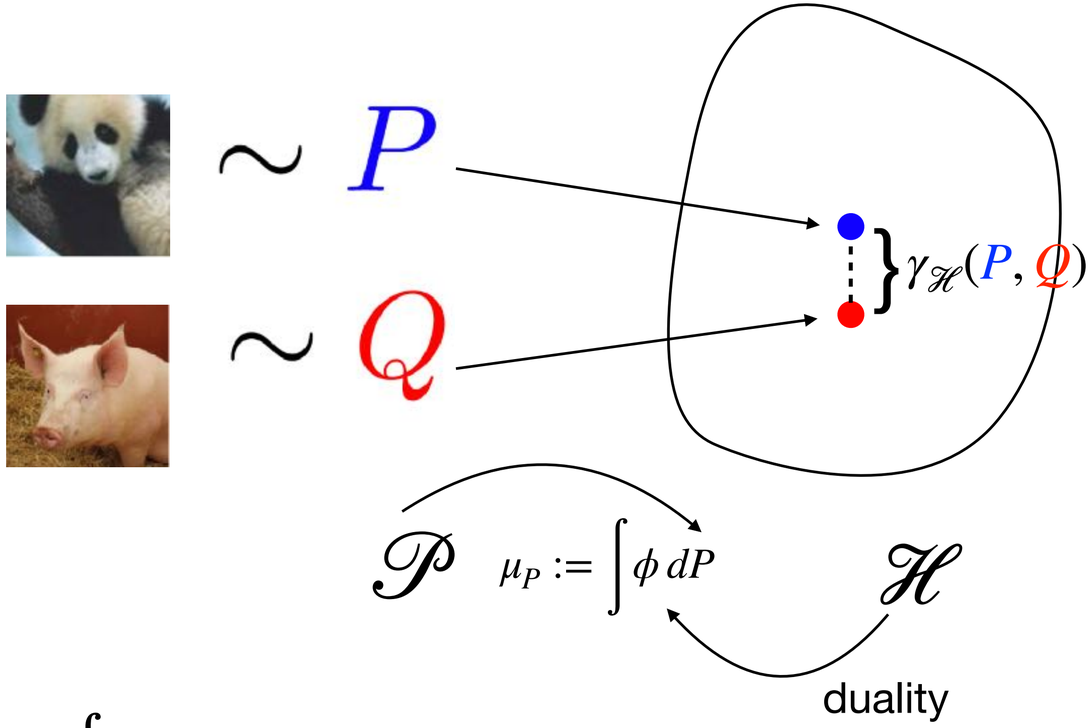
μ can be viewed as a generalized moment vector
e.g., let $\phi(x) = [x, x^2]^\top$

Learning with kernels (the modern way)

- 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 $\mathcal{H}$.

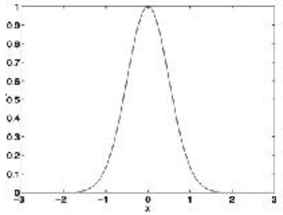


$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

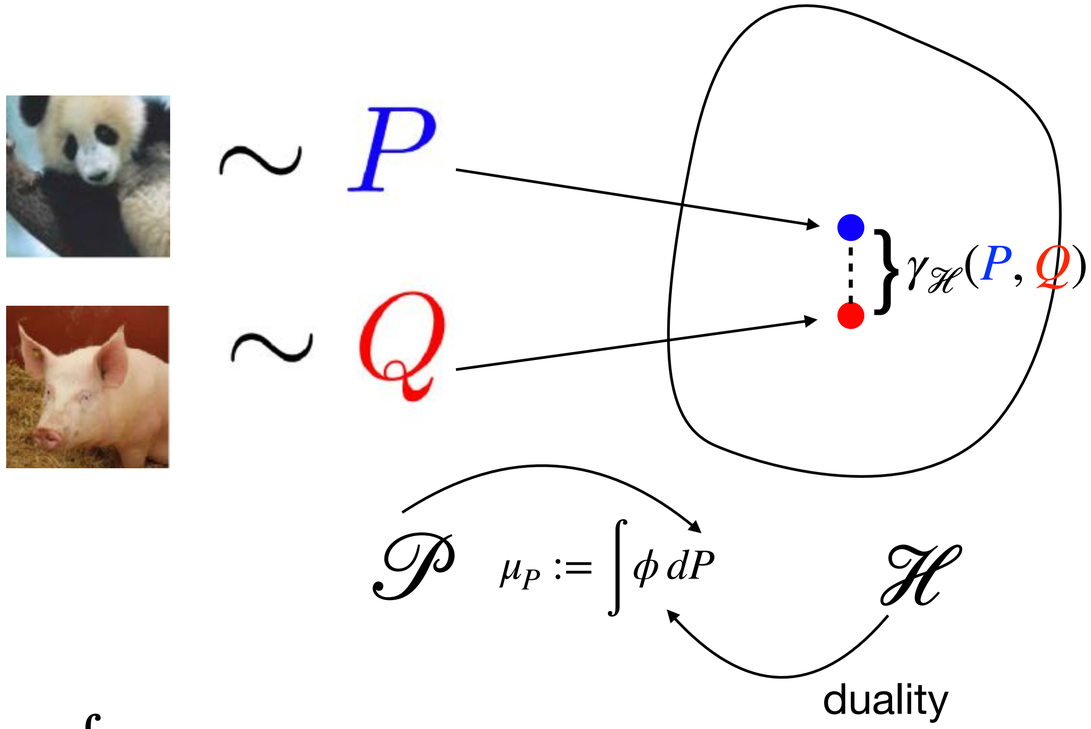
μ can be viewed as a generalized moment vector
e.g., let $\phi(x) = [x, x^2]^\top$

Learning with kernels (the modern way)

- 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), $\gamma_{\mathcal{H}}$ is a metric on $\mathcal{P}$.

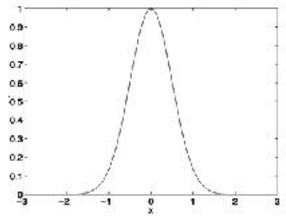


$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

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

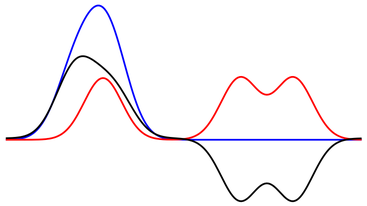
Learning with kernels (the modern way)

- A kernel is a symmetric function
  $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, e.g., Gaussian kernel
  $k(x, x') = \exp\left(-\|x - x'\|_2^2 \,/\, 2\sigma^2\right)$.



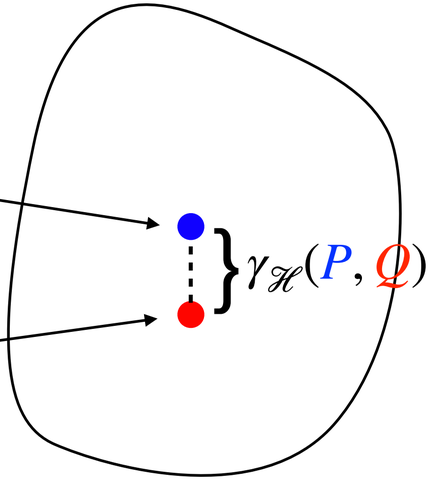
- A p.d. k corresponds to a Hilbert space $\mathcal{H}$ (RKHS), which satisfies the **reproducing property**
  $f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}, x \in \mathcal{X}$,
  $\phi(x) := k(x, \cdot\,)$ is the **canonical feature** of $\mathcal{H}$.
- If $\mathcal{H}$ is a large (dense in C), $\gamma_{\mathcal{H}}$ is a metric on $\mathcal{P}$.
- We can generalize to the more general **integral probability metric** (IPM)

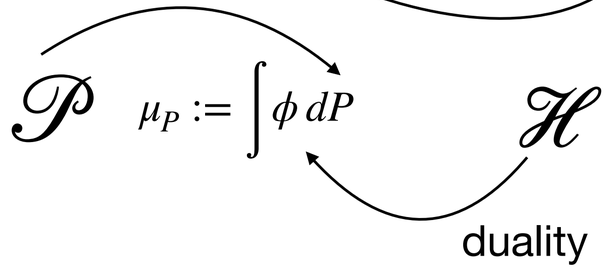
$$\text{IPM}(\mathcal{F}; P, Q) := \sup_{f \in \mathcal{F}} \int f \, d(P - Q).$$



$\sim P$

$\sim Q$



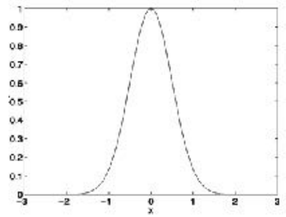


$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

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

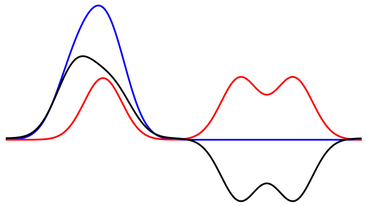
Learning with kernels (the modern way)

- 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), $\gamma_{\mathcal{H}}$ is a metric on $\mathcal{P}$.
- We can generalize to the more general **integral probability metric** (IPM)

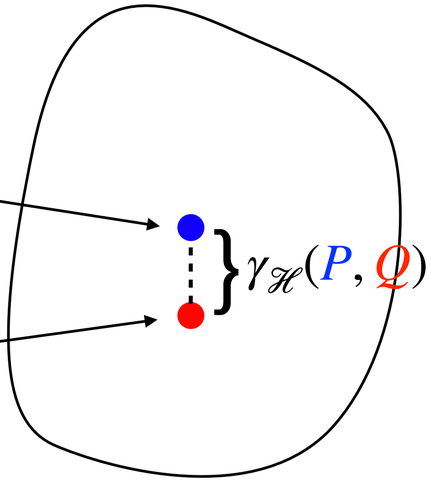
$$\text{IPM}(\mathcal{F}; P, Q) := \sup_{f \in \mathcal{F}} \int f \, d(P - Q).$$

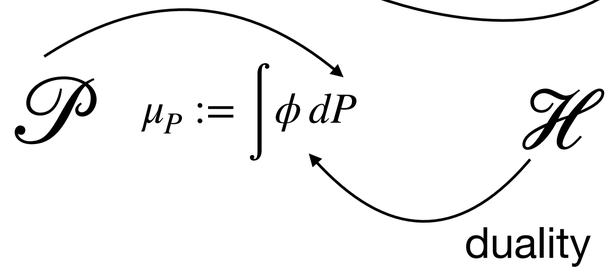


  - Special cases:
    $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ –> Maximum Mean Discrepancy (MMD)
    $\mathcal{F} = \{f : \|f\|_{\text{lip}} \leq 1\}$ –> Wasserstein (type-1)

$\sim P$

$\sim Q$

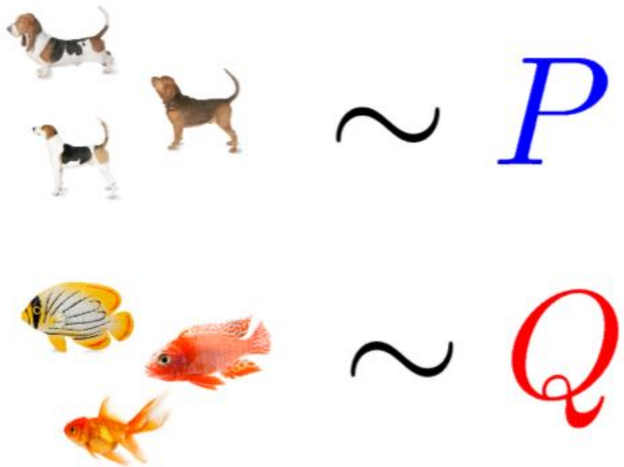




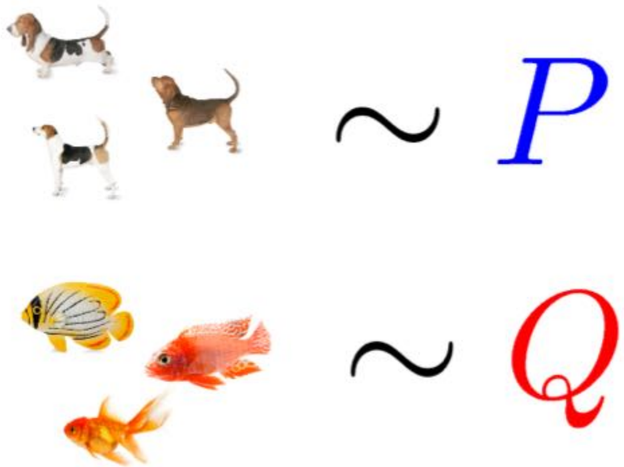
$\mu := \int \phi \, dP$ is the *(kernel)* ***mean embedding*** of P in $\mathcal{H}$.

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

The MMD



The MMD



MMD slides courtesy of Wittawat Jitkrittum

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2 \langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2\langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2\langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2\langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

$$\mathrm{MMD}(p, q) := \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}},$$

where $\phi(x)$ is in the Hilbert space $\mathcal{H}$.

- Recall $\|a - b\|^2 = \langle a, a \rangle - 2\langle a, b \rangle + \langle b, b \rangle$.

$$\begin{aligned}
\mathrm{MMD}^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\
&= \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{x' \sim p}[\phi(x')] \rangle - 2 \langle \mathbb{E}_{x \sim p}[\phi(x)], \mathbb{E}_{y \sim q}[\phi(y)] \rangle \\
&\quad + \langle \mathbb{E}_{y \sim q}[\phi(y)], \mathbb{E}_{y' \sim q}[\phi(y')] \rangle \\
&= \mathbb{E}_{x \sim p} \mathbb{E}_{x' \sim p} \langle \phi(x), \phi(x') \rangle - 2 \mathbb{E}_{x \sim p} \mathbb{E}_{y \sim q} \langle \phi(x), \phi(y) \rangle \\
&\quad + \mathbb{E}_{y \sim q} \mathbb{E}_{y' \sim q} \langle \phi(y), \phi(y') \rangle
\end{aligned}$$

- Depend on only the inner product $\langle \phi(x), \phi(y) \rangle$.
- Don't need $\phi(x)$ explicitly (could be ∞-dimensional!).

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

- Define $k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ (kernel).

$$\begin{aligned}\mathrm{MMD}_k^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x \sim p}\mathbb{E}_{x' \sim p} k(x, x') - 2\mathbb{E}_{x \sim p}\mathbb{E}_{y \sim q} k(x, y') \\ &\quad + \mathbb{E}_{y \sim q}\mathbb{E}_{y' \sim q} k(y, y').\end{aligned}$$

- Unbiased estimator:

$$\widehat{\mathrm{MMD}}_k^2(\mathsf{X}, \mathsf{Y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, y_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(y_i, y_j).$$

- $k(x, x') \approx$ similarity between x and x'.

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

- Define $k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ (kernel).

$$\begin{aligned}\mathrm{MMD}_k^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x \sim p}\mathbb{E}_{x' \sim p} k(x, x') - 2\mathbb{E}_{x \sim p}\mathbb{E}_{y \sim q} k(x, y') \\ &\quad + \mathbb{E}_{y \sim q}\mathbb{E}_{y' \sim q} k(y, y').\end{aligned}$$

- Unbiased estimator:

$$\begin{aligned}\widehat{\mathrm{MMD}}_k^2(\mathrm{X}, \mathrm{Y}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, y_j) \\ &\quad + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(y_i, y_j).\end{aligned}$$

- $k(x, x') \approx$ similarity between x and x'.

Maximum Mean Discrepancy (MMD) [Gretton et al., 2012a]

- Define $k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$ (kernel).

$$\begin{aligned}\mathrm{MMD}_k^2(p, q) &= \left\| \mathbb{E}_{x \sim p}[\phi(x)] - \mathbb{E}_{y \sim q}[\phi(y)] \right\|_{\mathcal{H}}^2 \\ &= \mathbb{E}_{x \sim p}\mathbb{E}_{x' \sim p} k(x, x') - 2\mathbb{E}_{x \sim p}\mathbb{E}_{y \sim q} k(x, y') \\ &\quad + \mathbb{E}_{y \sim q}\mathbb{E}_{y' \sim q} k(y, y').\end{aligned}$$

- Unbiased estimator:

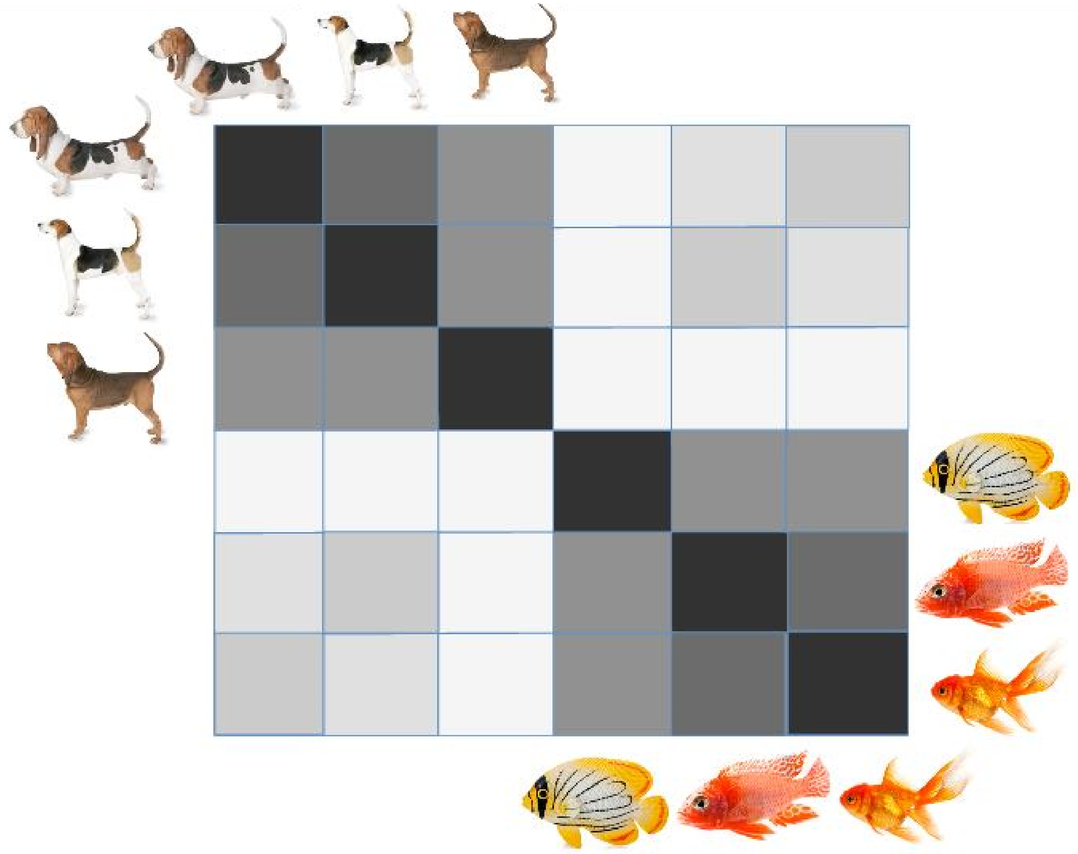
$$\begin{aligned}\widehat{\mathrm{MMD}}_k^2(\mathrm{X}, \mathrm{Y}) = &\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(x_i, x_j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, y_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} k(y_i, y_j).\end{aligned}$$

- $k(x, x') \approx$ similarity between x and x'.

See Gretton et al. 2012 for the biased estimator

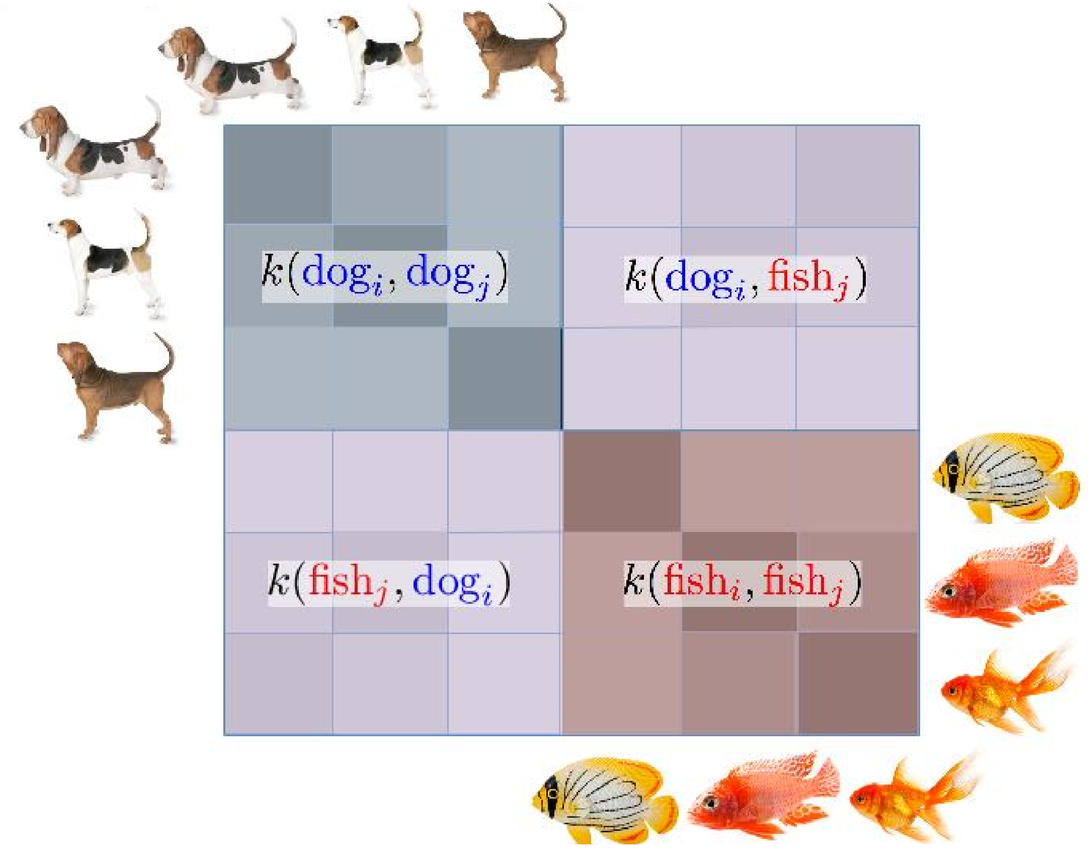
Intuition for the MMD

- Dogs $\sim p$ and fish $\sim q$.
- Each entry is one of $k(\text{dog}_i, \text{dog}_j)$, $k(\text{dog}_i, \text{fish}_j)$, or $k(\text{fish}_i, \text{fish}_j)$

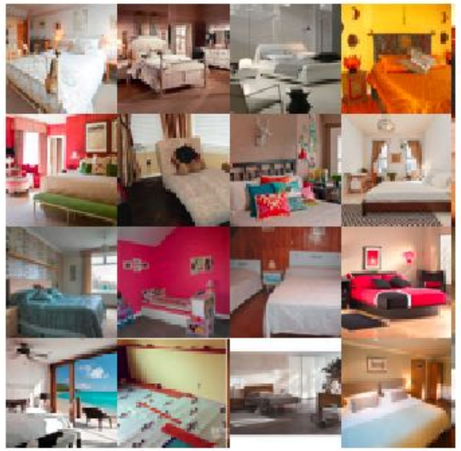


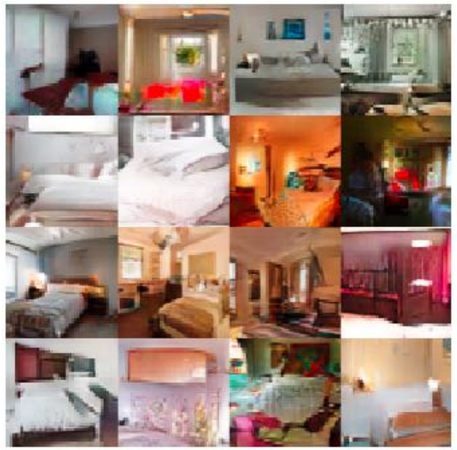
Intuition for the MMD

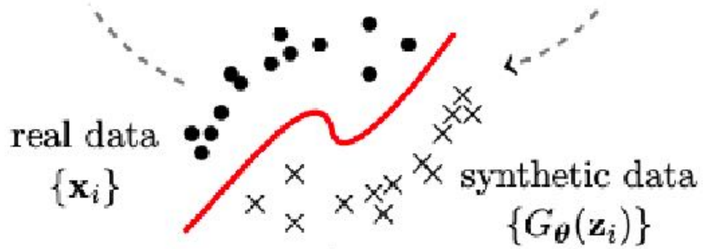
$$\widehat{\mathrm{MMD}}_k^2 = \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{dog}_i, \mathrm{dog}_j) - \frac{2}{n^2} \sum_{i,j} k(\mathrm{dog}_i, \mathrm{fish}_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(\mathrm{fish}_i, \mathrm{fish}_j)$$



Application to generative modeling: MMD for generative adversarial nets (GAN)







MMD Test

$\left\| \hat{\mu}_X - \hat{\mu}_{G_{\boldsymbol{\theta}}(Z)} \right\|_{\mathscr{H}}$ is zero?

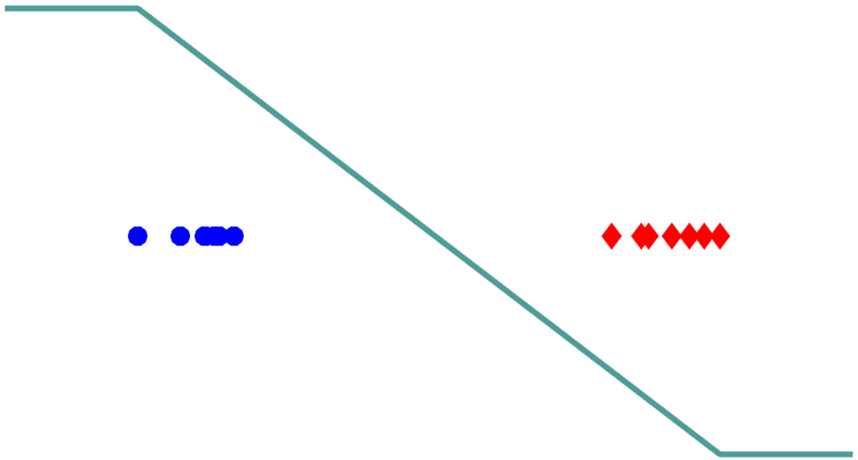
Comparison: Wasserstein-1 vs. MMD-k

Comparison: Wasserstein-1 vs. MMD-k

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

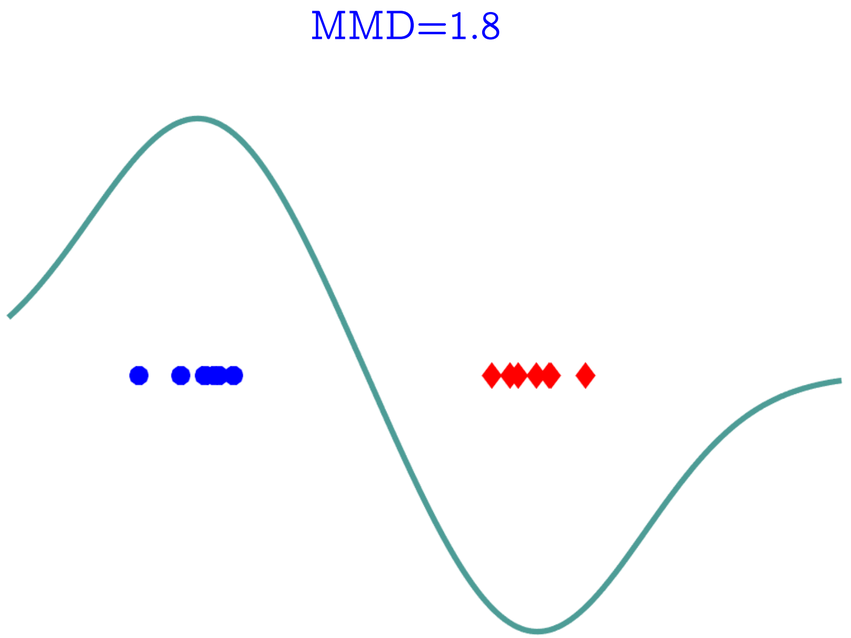
$||f||_L := \sup_{x \neq y} |f(x) - f(y)| / ||x - y||$

W_1=0.88



Comparison: Wasserstein-1 vs. MMD-k

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

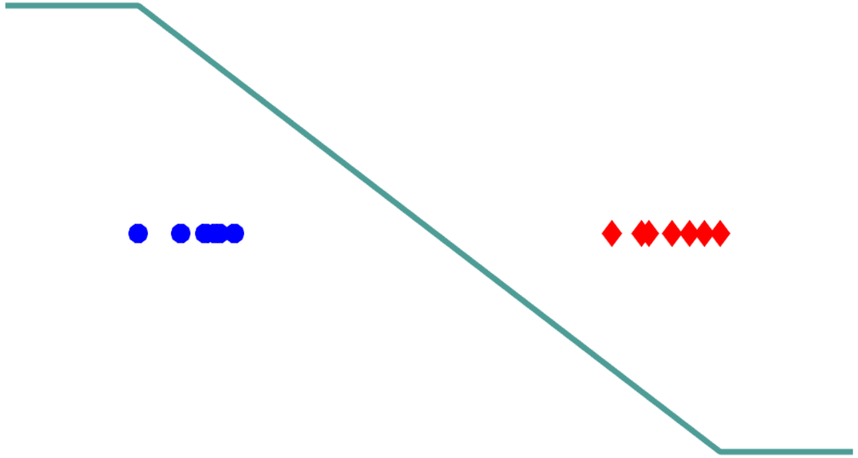


Comparison: Wasserstein-1 vs. MMD-k

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

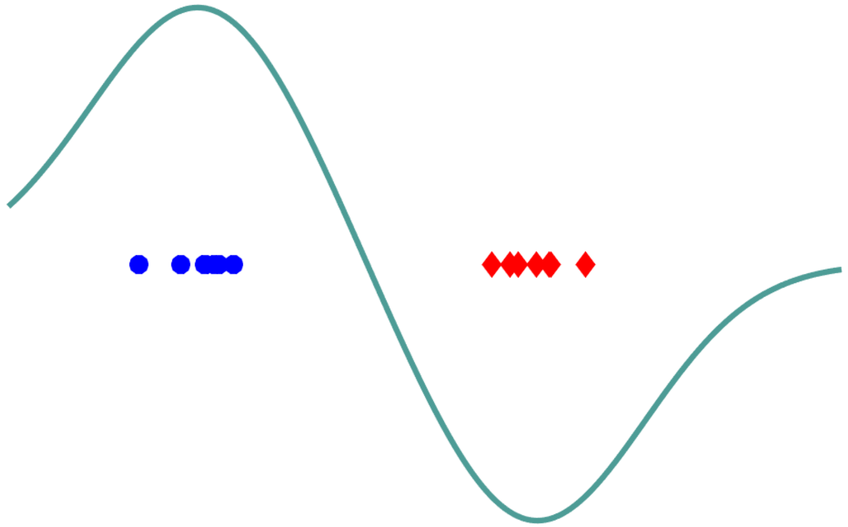
$||f||_L := \sup_{x \neq y} |f(x) - f(y)| / ||x - y||$

W_1=0.88



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

MMD=1.8



Comparison: Wasserstein-1 vs. MMD-k

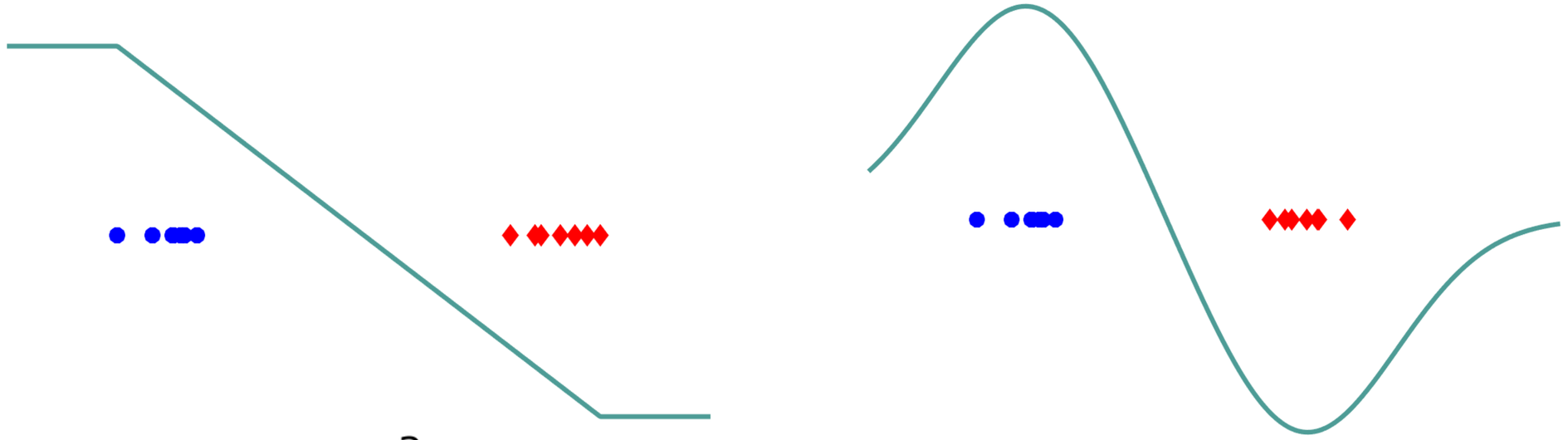
$W_1(P, Q) = \sup_{||f||_L \leq 1} E_P f(X) - E_Q f(Y).$

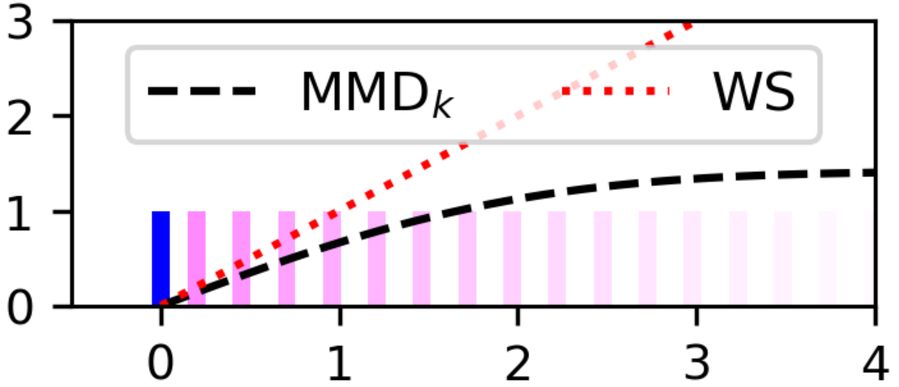
$||f||_L := \sup_{x \neq y} |f(x) - f(y)| / ||x - y||$

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

W_1=0.88

MMD=1.8





Distributional Robustness

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

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

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

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

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

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

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

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (DRO)

[Delage and Ye 2010, Scarf 1958]

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

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

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

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (DRO)

[Delage and Ye 2010, Scarf 1958]

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

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

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

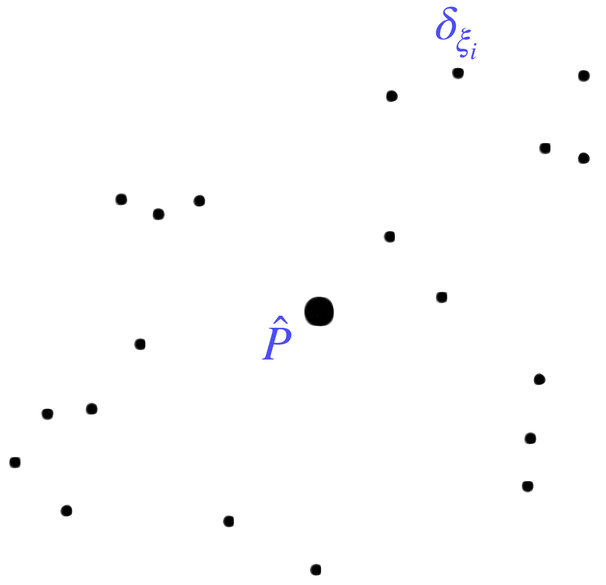
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (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 $\mathcal{K}$ (***ambiguity set***), e.g.,



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

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

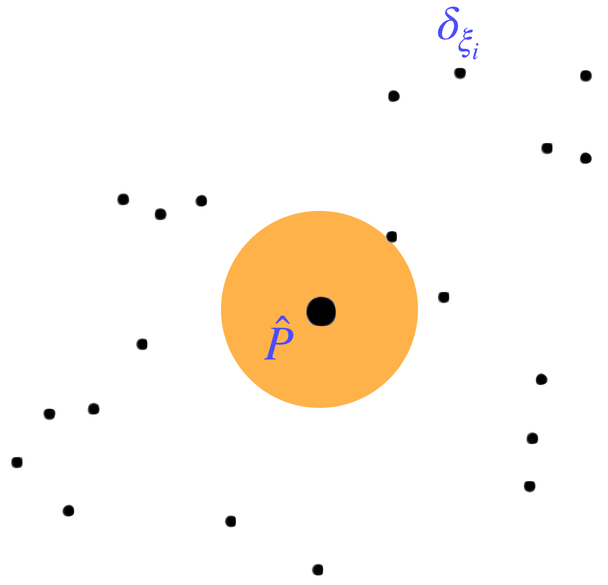
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (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 $\mathcal{K}$ (***ambiguity set***), e.g.,
  - $\mathcal{K}$ can be a metric-ball centered at $\hat{P}$, e.g., using relative entropy, optimal transport, and kernel methods .



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

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

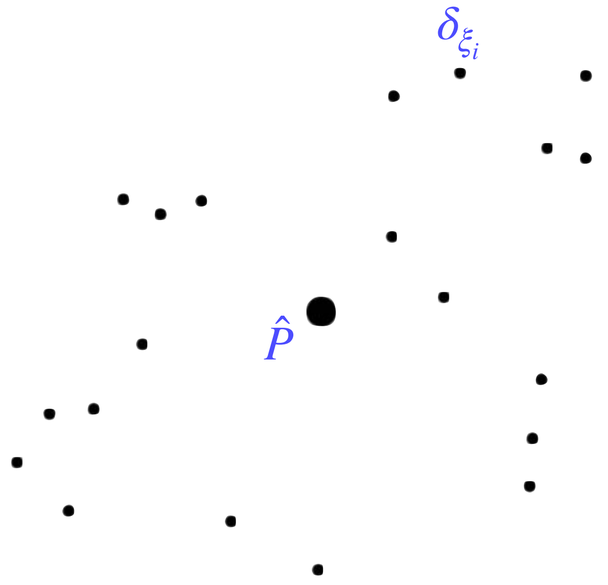
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (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 $\mathcal{K}$ (***ambiguity set***), e.g.,
  - $\mathcal{K}$ can be a metric-ball centered at $\hat{P}$, e.g., using relative entropy, optimal transport, and kernel methods .



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

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

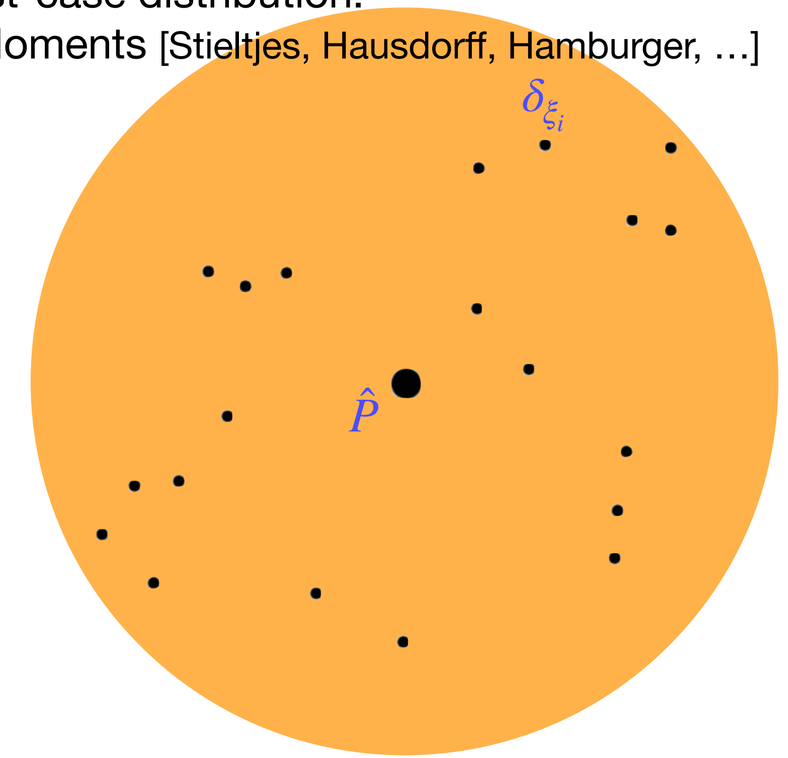
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (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 $\mathcal{K}$ (***ambiguity set***), e.g.,
  - $\mathcal{K}$ can be a metric-ball centered at $\hat{P}$, e.g., using relative entropy, optimal transport, and kernel methods .



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

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

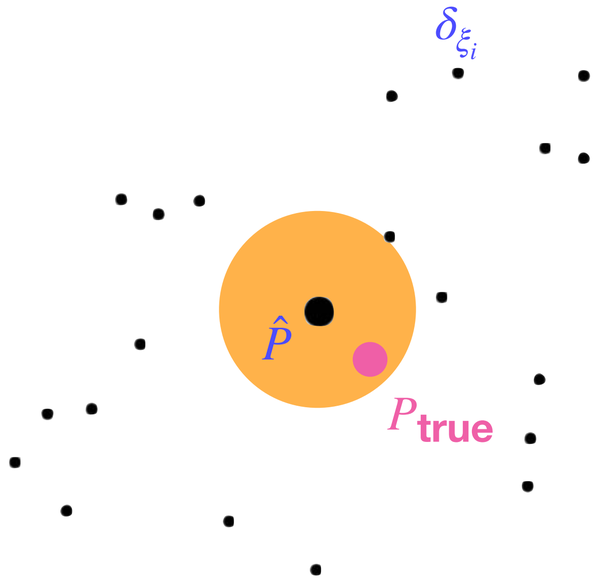
(RO) $\min_{\theta} \sup_{\xi \in \mathcal{U}} l(\theta, \xi)$

$\min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$ (DRO)

[Delage and Ye 2010, Scarf 1958]

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

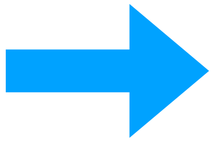
- Robustifies against a set of probability measures $\mathcal{K}$ (***ambiguity set***), e.g.,
  - $\mathcal{K}$ can be a metric-ball centered at $\hat{P}$, e.g., using relative entropy, optimal transport, and kernel methods .
    - One way of constructing ambiguity region: one can quantify the empirical mean convergence rate $\gamma(\hat{P}, P_{\text{true}}) \leq \epsilon$.



What distributions does the ambiguity set $\mathscr{K}$ contain?

$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

Uncertainty set



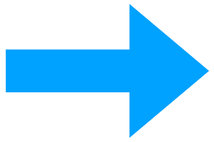
$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

What distributions does the ambiguity set $\mathscr{K}$ contain?

$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

Uncertainty set



$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$

What distributions does the ambiguity set $\mathscr{K}$ contain?

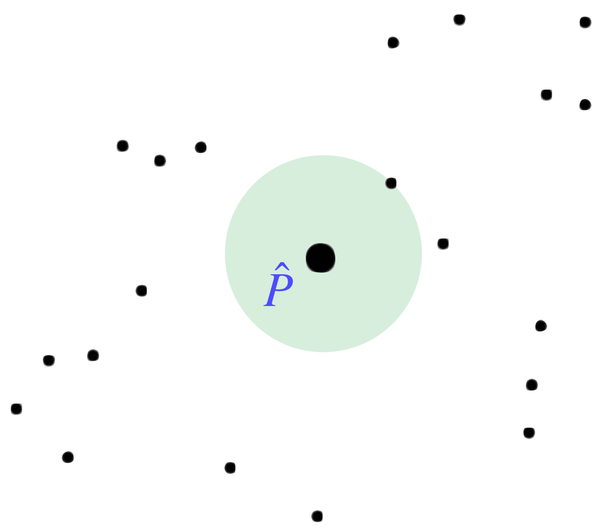
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$



What distributions does the ambiguity set $\mathscr{K}$ contain?

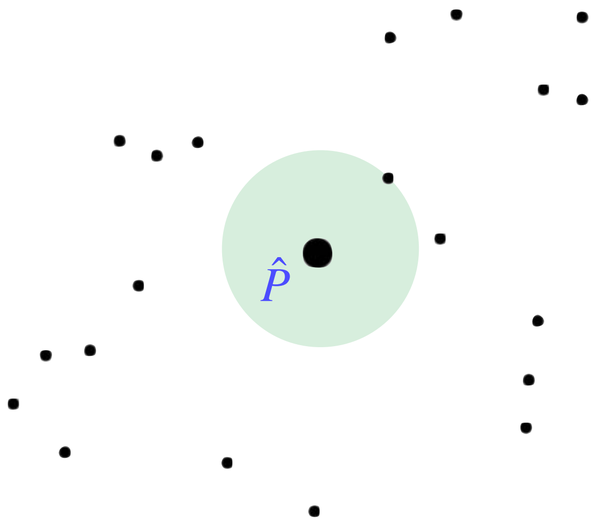
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

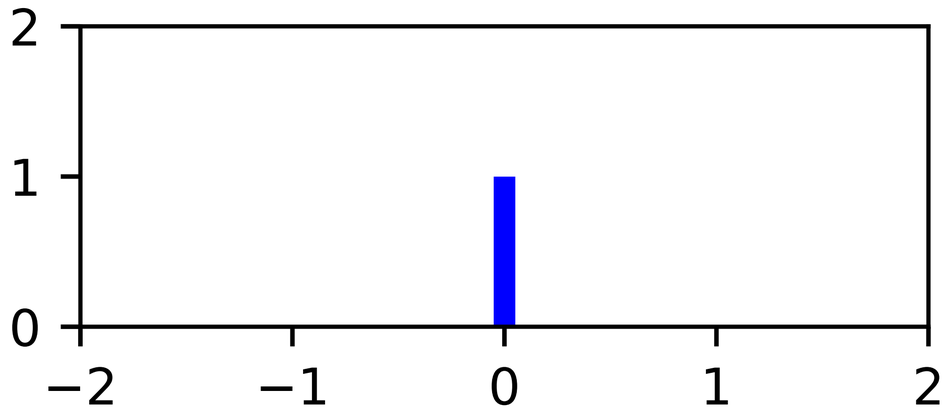
Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$





What distributions does the ambiguity set $\mathscr{K}$ contain?

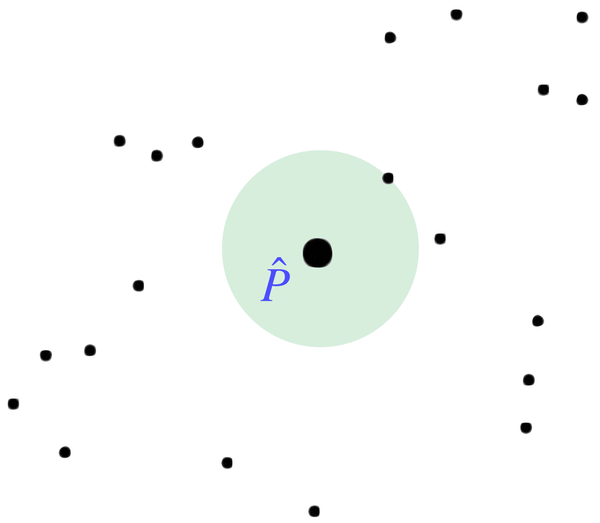
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

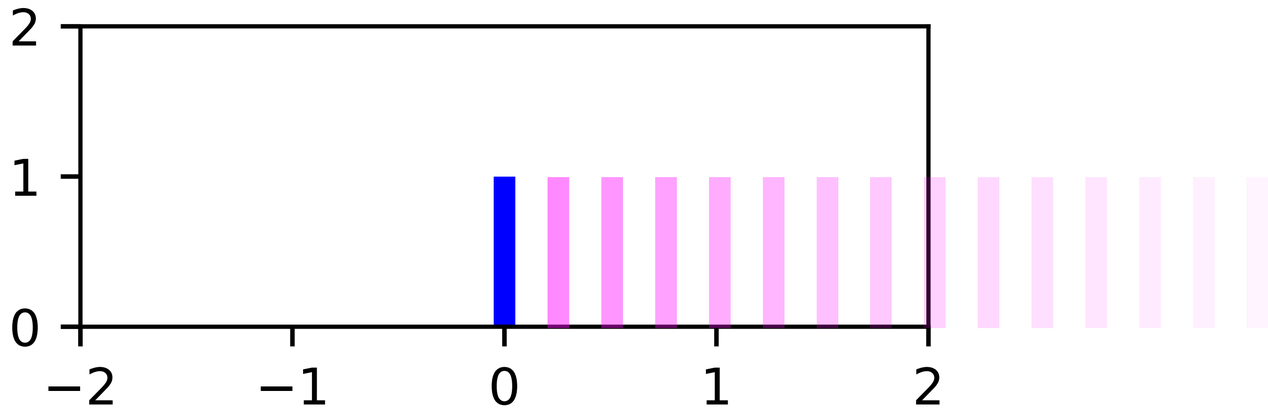
Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$





What distributions does the ambiguity set $\mathscr{K}$ contain?

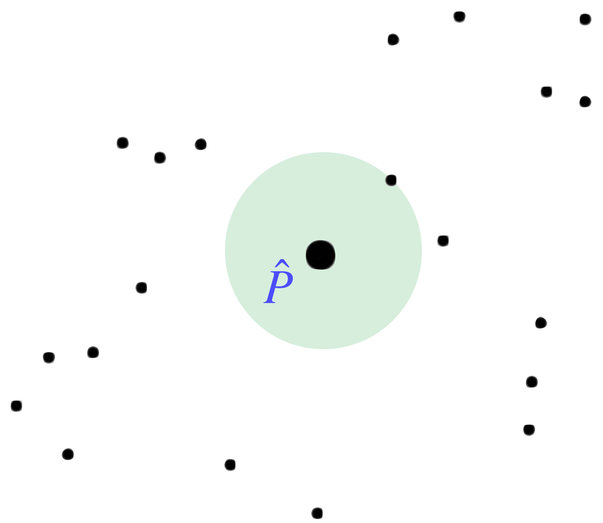
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

$\mathscr{X}$: Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

$\mathscr{K}$: Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$



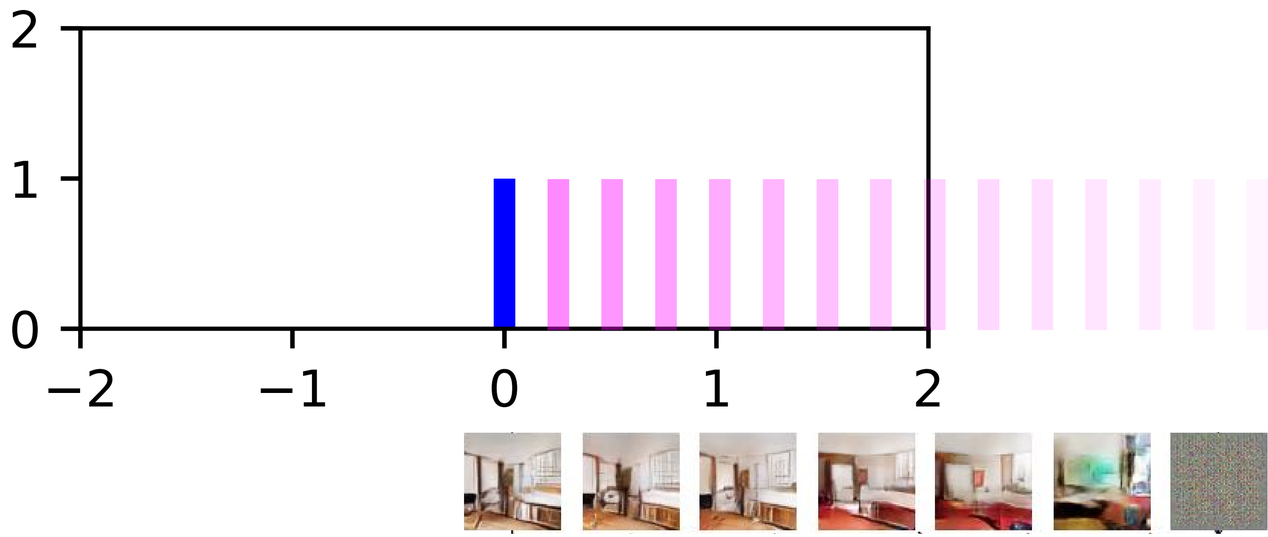


Image credit: Arjovsky'17

What distributions does the ambiguity set $\mathscr{K}$ contain?

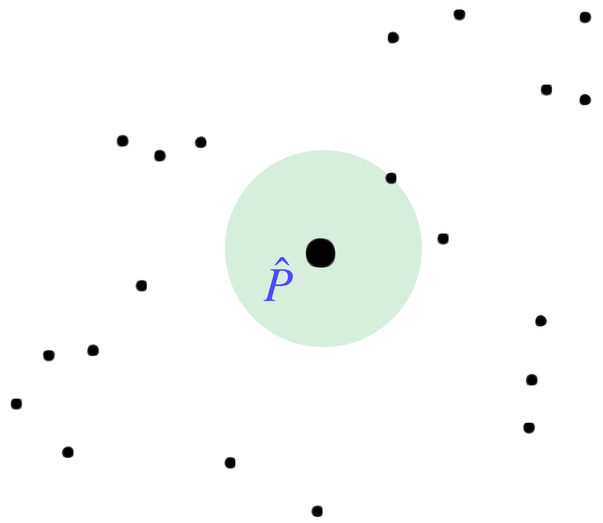
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$



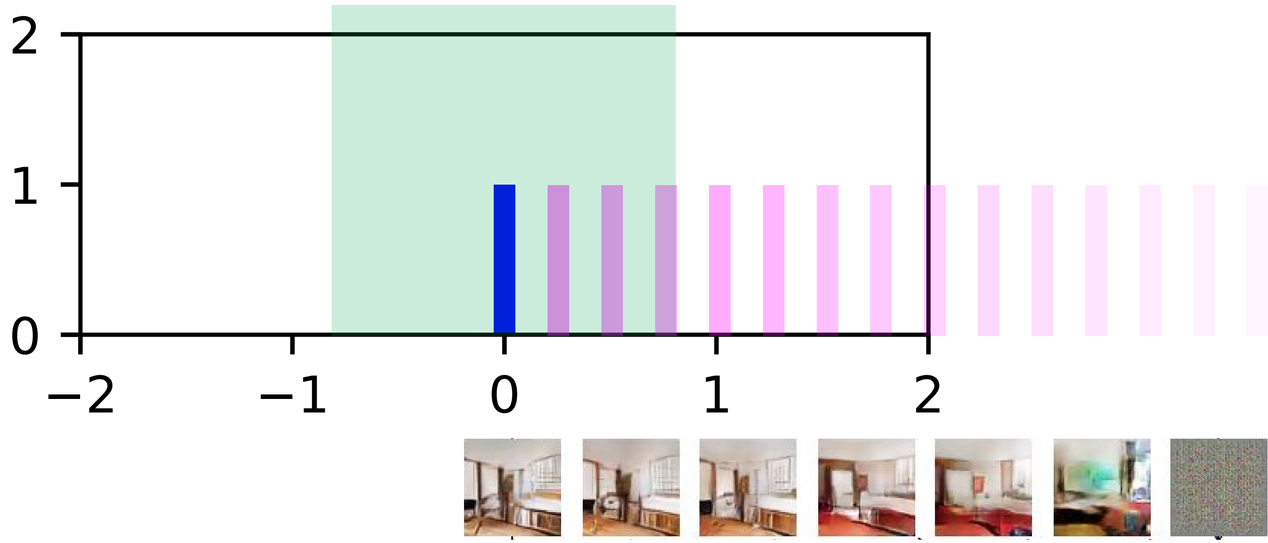


Image credit: Arjovsky'17

What distributions does the ambiguity set $\mathscr{K}$ contain?

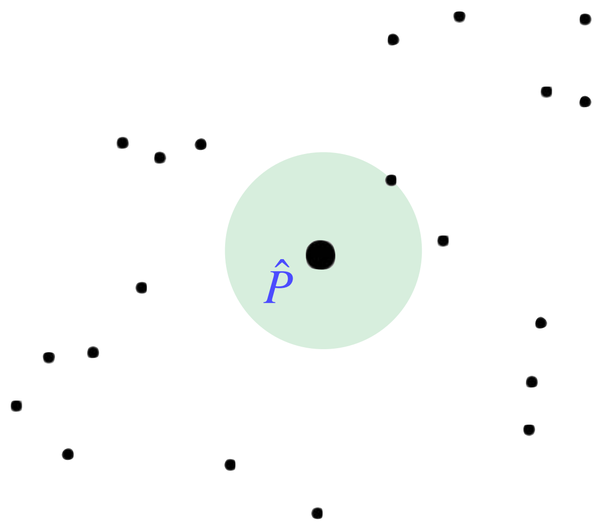
$$\min_{\theta} \sup_{\xi \in \mathscr{X}} l(\theta, \xi) \quad \text{(RO)}$$

Uncertainty set

$$\min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_P l(\theta, \xi) \quad \text{(DRO)}$$

Ambiguity set

$$\mathscr{K} = \left\{ P : \gamma(P, \hat{P}) \leq \epsilon \right\}$$



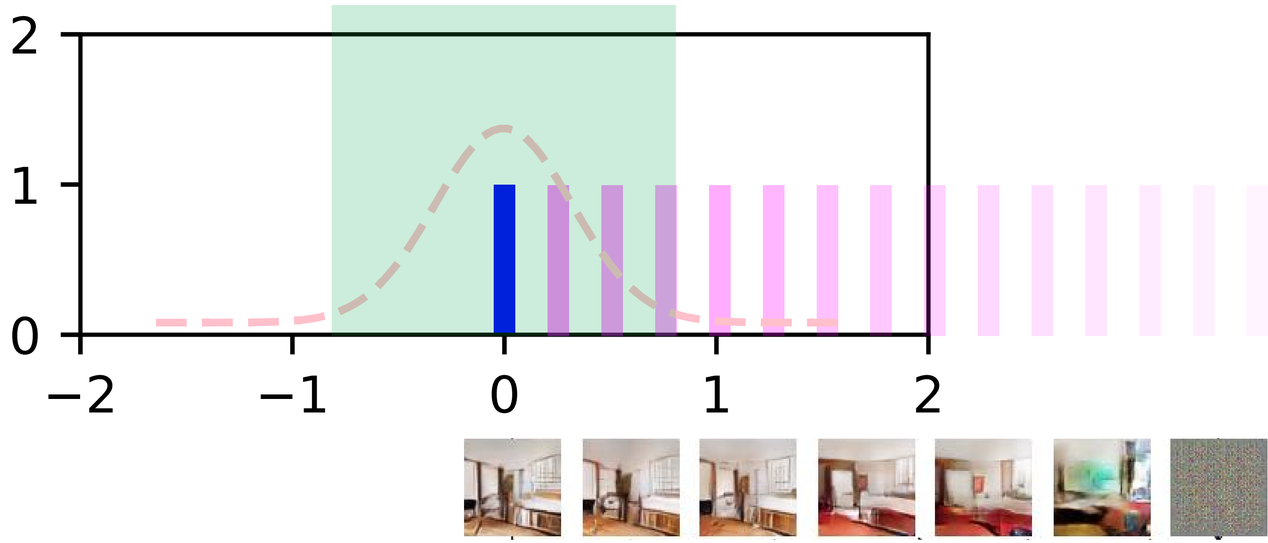


Image credit: Arjovsky'17

Choosing different ambiguity sets

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

- singleton $\mathcal{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

- singleton $\mathcal{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM
- entire space $\mathcal{K} = \mathcal{P}$: RO

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

- singleton $\mathcal{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM
- entire space $\mathcal{K} = \mathcal{P}$: RO
- polytope $\mathcal{K} = \text{conv}\{\delta_{\xi_1}, \ldots, \delta_{\xi_1}\}$: SVMs, scenario opt.

Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

- singleton $\mathcal{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM
- entire space $\mathcal{K} = \mathcal{P}$: RO
- polytope $\mathcal{K} = \text{conv}\{\delta_{\xi_1}, \ldots, \delta_{\xi_1}\}$: SVMs, scenario opt.
- metric-ball $\mathcal{K} = \{P: \ \text{MMD}(P, \hat{P}; \mathcal{H}) \leq \epsilon\}$

  $\mathcal{K} = \{P: \ \text{W}_\alpha(P, \hat{P}) \leq \epsilon\}$

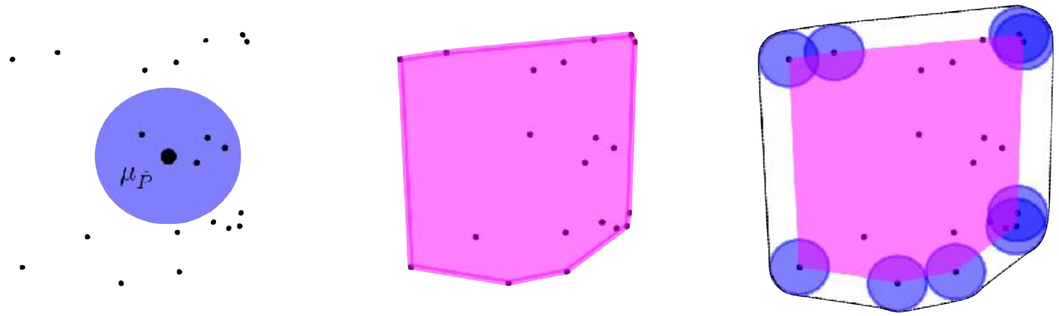
Choosing different ambiguity sets

$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathscr{K}} \mathbb{E}_{P} l(\theta, \xi)$$

Examples of ambiguity set $\mathscr{K} \subseteq \mathscr{P}$

- singleton $\mathscr{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM
- entire space $\mathscr{K} = \mathscr{P}$: RO
- polytope $\mathscr{K} = \text{conv}\{\delta_{\xi_1}, \ldots, \delta_{\xi_1}\}$: SVMs, scenario opt.
- metric-ball $\mathscr{K} = \{P: \ \text{MMD}(P, \hat{P}; \mathscr{H}) \leq \epsilon\}$

  $\mathscr{K} = \{P: \ \text{W}_{\alpha}(P, \hat{P}) \leq \epsilon\}$

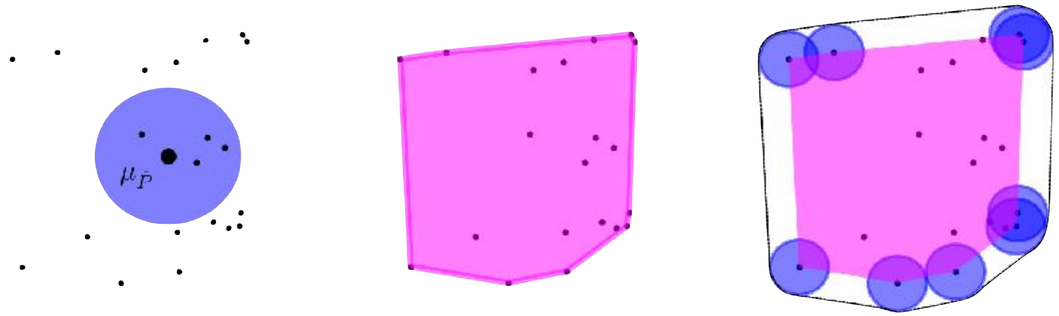


Choosing different ambiguity sets

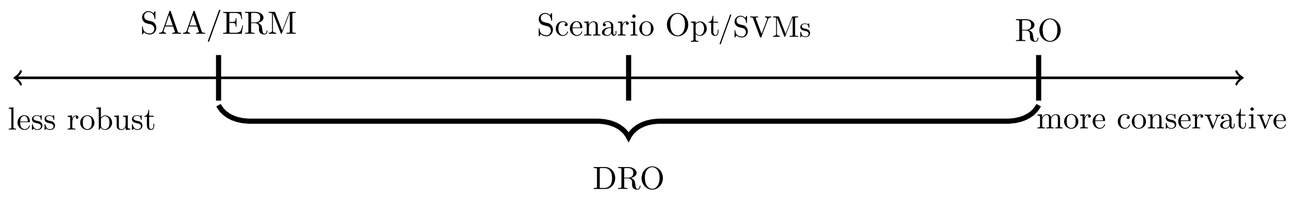
$$\text{(DRO)} \quad \min_{\theta} \sup_{P \in \mathcal{K}} \mathbb{E}_P l(\theta, \xi)$$

Examples of ambiguity set $\mathcal{K} \subseteq \mathcal{P}$

- singleton $\mathcal{K} = \{\frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i}\}$: ERM
- entire space $\mathcal{K} = \mathcal{P}$: RO



- polytope $\mathcal{K} = \text{conv}\{\delta_{\xi_1}, \ldots, \delta_{\xi_1}\}$: SVMs, scenario opt.
- metric-ball $\mathcal{K} = \{P: \ \text{MMD}(P, \hat{P}; \mathcal{H}) \leq \epsilon\}$
  $\mathcal{K} = \{P: \ \text{W}_\alpha(P, \hat{P}) \leq \epsilon\}$



The lifting trick of DRO

The lifting trick of DRO

- How do we solve metric-ball constrained DRO?

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

The lifting trick of DRO

- How do we solve metric-ball constrained DRO?

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

- Remember the RO we solved earlier?

$$\min_{\theta} \quad \sup_{\|p-\hat{p}\|_\alpha\leq\epsilon} \quad p^T\theta$$

The lifting trick of DRO

- How do we solve metric-ball constrained DRO?

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

- Remember the RO we solved earlier?

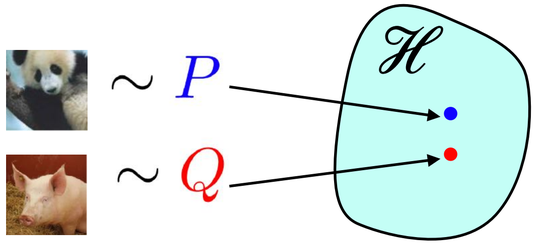
$$\min_{\theta} \quad \sup_{\|p-\hat{p}\|_\alpha\leq\epsilon} \quad p^T\theta$$

- By comparing them, we find that DRO has the same structure as RO! This motivates us to solve DRO with convex duality (but DRO is ∞ -dimensional!).

Dual program of DRO

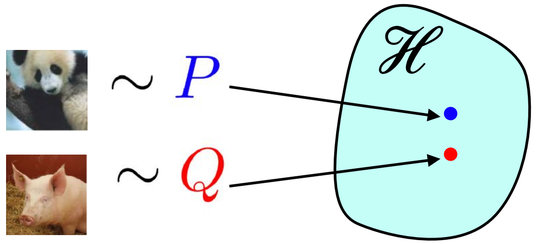
Dual program of DRO

$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\leq\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$



Dual program of DRO

$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\leq\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$

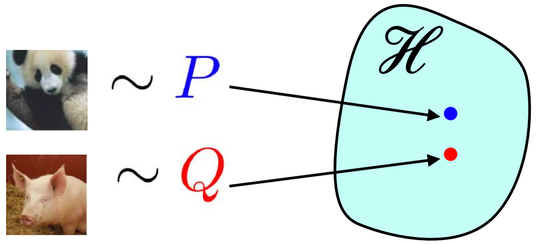


Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \quad \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta, \cdot) \leq f,$$

Dual program of DRO

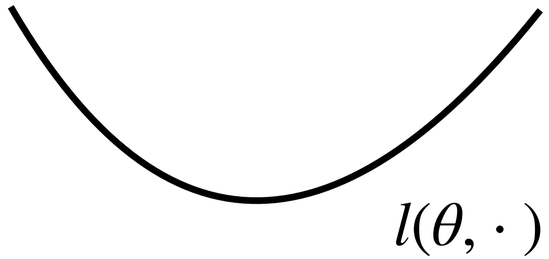
$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\leq\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

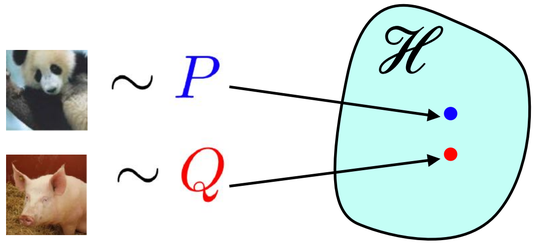
$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \quad \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta, \cdot) \leq f,$$

Geometric intuition



Dual program of DRO

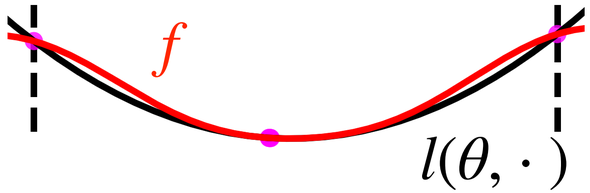
$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\leq\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

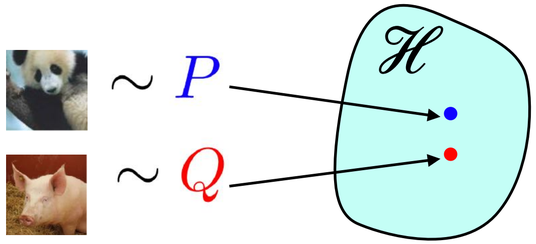
$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta, \cdot) \leq f,$$

Geometric intuition



Dual program of DRO

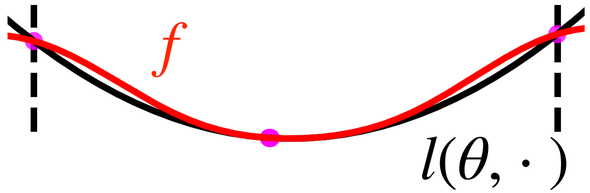
$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\le\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \quad \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta, \cdot) \le f,$$

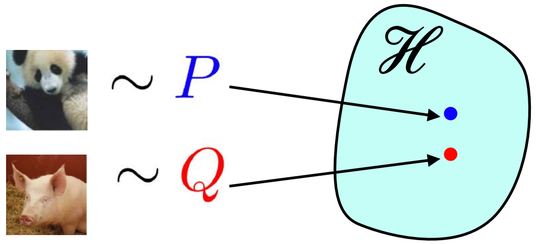
Geometric intuition



Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

Dual program of DRO

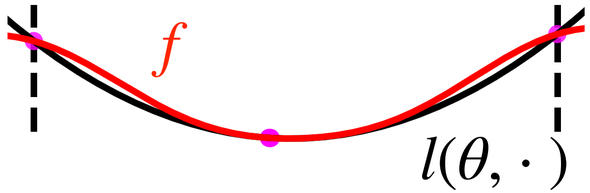
$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\le\epsilon} \quad \mathbb{E}_{P} l(\theta, \xi)$$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \quad \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta, \cdot) \le f,$$

Geometric intuition

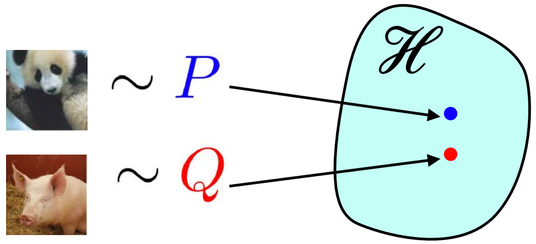


Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

Intuition: **flatten the curve, smooth is robust**

Dual program of DRO

$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\le\epsilon} \quad \mathbb{E}_P l(\theta,\xi)$$

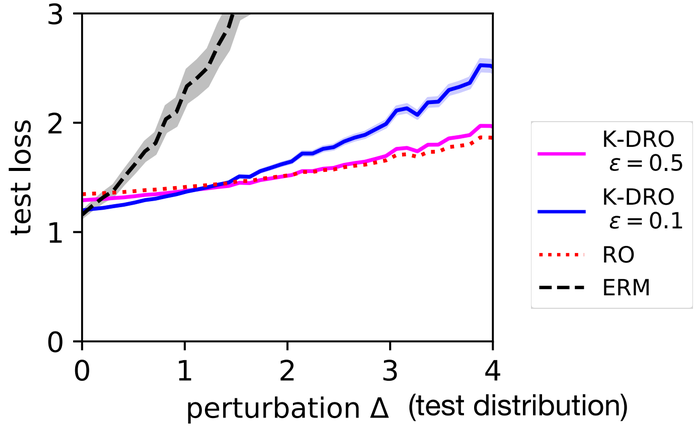


Example. Robust least squares

[El Ghaoui Lebret '97]

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

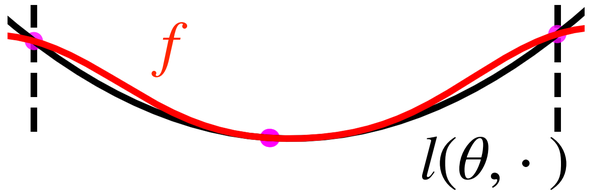
Given historical samples $\xi_1, \xi_2, \ldots, \xi_N$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \quad \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t. } l(\theta,\cdot) \le f,$$

Geometric intuition

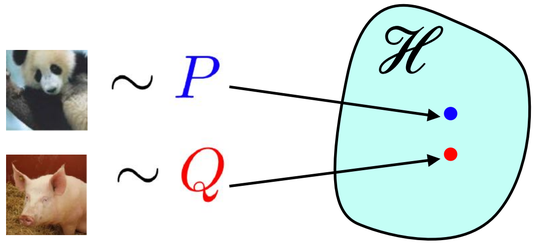


Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

Intuition: **flatten the curve, smooth is robust**

Dual program of DRO

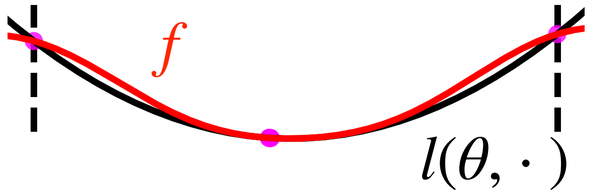
$$\text{(P)} \quad \min_{\theta} \quad \sup_{\gamma(P,\hat{P})\le\epsilon} \quad \mathbb{E}_{P} l(\theta,\xi)$$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

$$\text{(D)} \quad \min_{\theta, f\in\mathscr{H}} \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.} \; l(\theta,\cdot) \le f,$$

Geometric intuition



Smoothness of f ↭ Distributional robustness (↭ Size of $\mathscr{H}$)

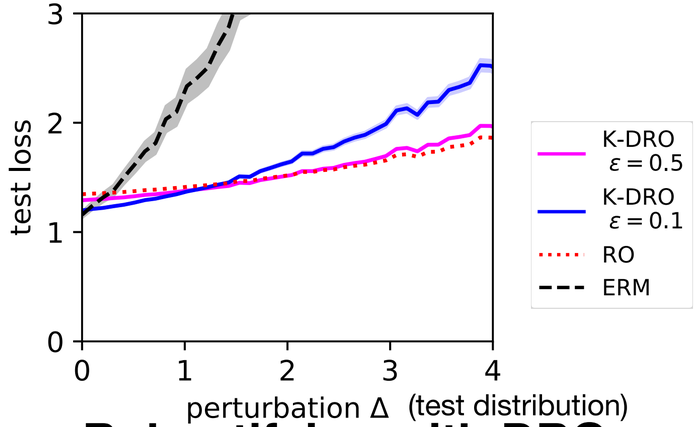
Intuition: **flatten the curve, smooth is robust**

Example. Robust least squares

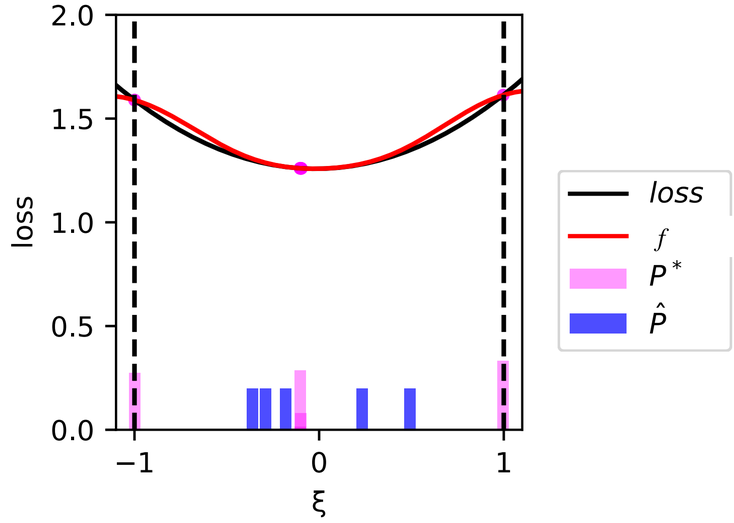
[El Ghaoui Lebret '97]

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

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

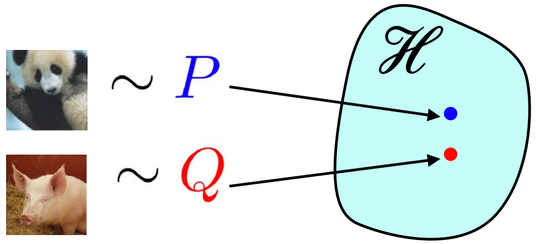


Robustifying with DRO



Dual program of DRO

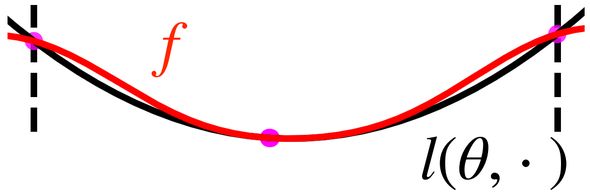
(P) $\min_{\theta} \sup_{\gamma(P,\hat{P})\le\epsilon} \mathbb{E}_P l(\theta,\xi)$



Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

(D) $\min_{\theta, f\in\mathscr{H}} \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot) \le f,$

Geometric intuition



Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

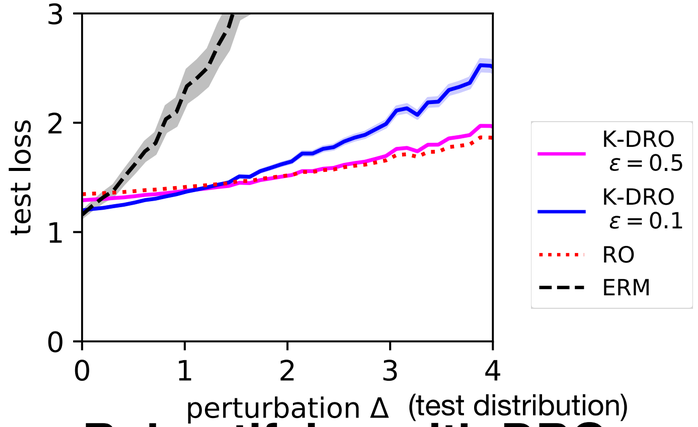
Intuition: **flatten the curve, smooth is robust**

Example. Robust least squares

[El Ghaoui Lebret '97]

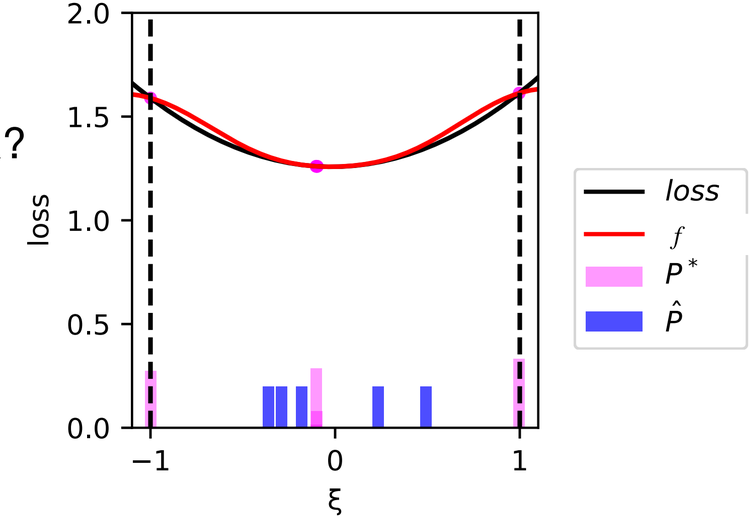
$\text{minimize}\ l(\theta,\xi) := \|A(\xi)\cdot\theta - b\|_2^2$

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



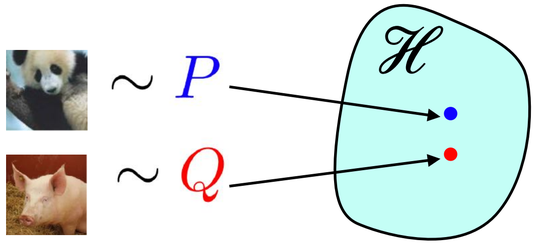
Robustifying with DRO

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



Dual program of DRO

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

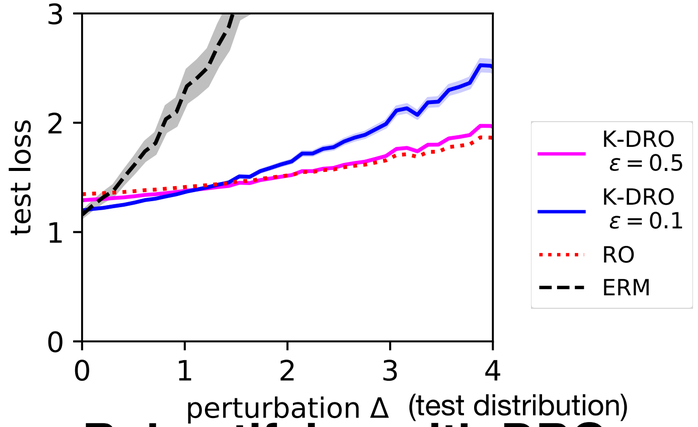


Example. Robust least squares

[El Ghaoui Lebret '97]

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

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

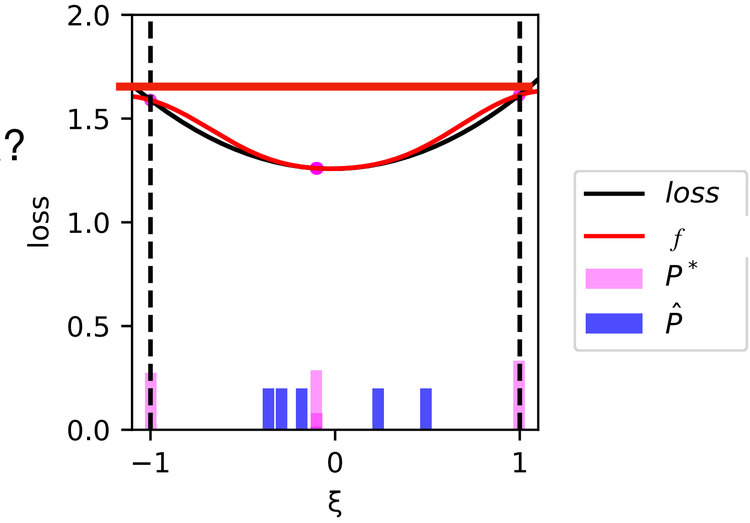


Theorem (Simplified: **Generalized DRO duality, Zhu et al. '20**). DRO (P) is equivalent to solving its dual problem

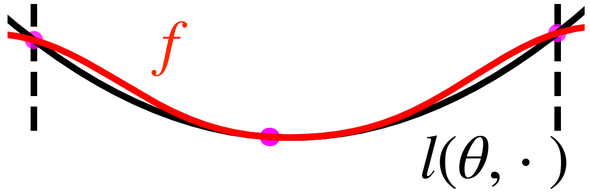
(D) $\min_{\theta, f\in\mathscr{H}} \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t. } l(\theta,\cdot) \le f,$

Robustifying with DRO

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



Geometric intuition



Smoothness of f ↔ Distributional robustness (↔ Size of $\mathscr{H}$)

Intuition: **flatten the curve, smooth is robust**

Special cases of DRO dual program

$$\text{(P)}\ \min_{\theta}\ \sup_{\gamma(P,\hat{P})\le\epsilon}\ \mathbb{E}_{P} l(\theta,\xi) \quad \Longleftrightarrow \quad \text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon \|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot) \le f$$

Special cases of DRO dual program

$$\text{(P)}\ \min_{\theta}\ \sup_{\gamma(P,\hat{P})\le\epsilon}\ \mathbb{E}_{P} l(\theta,\xi) \quad\Longleftrightarrow\quad \text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot)\le f$$

- Type-1 Wasserstein DRO. Suppose the loss function $l(\theta,)$ is Lipschitz continuous, and we choose γ to be the W-1 distance. (D) is then equivalent to Lipschitz regularized ERM

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta,\xi) + \epsilon\mathrm{Lip}(l).$$

Special cases of DRO dual program

$$\text{(P)}\ \min_{\theta}\ \sup_{\gamma(P,\hat{P})\le\epsilon}\ \mathbb{E}_{P} l(\theta,\xi) \quad\Longleftrightarrow\quad \text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot)\le f$$

- Type-1 Wasserstein DRO. Suppose the loss function $l(\theta,)$ is Lipschitz continuous, and we choose γ to be the W-1 distance. (D) is then equivalent to Lipschitz regularized ERM

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta,\xi) + \epsilon\mathrm{Lip}(l).$$

- Distributionally robust logistic regression (with exact regularization coefficient)

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta^{\top}x_i, y_i) + \epsilon\cdot\|\theta\|_2.$$

Special cases of DRO dual program

$$\text{(P)}\ \min_{\theta}\ \sup_{\gamma(P,\hat{P})\le\epsilon}\ \mathbb{E}_{P} l(\theta,\xi) \quad\Longleftrightarrow\quad \text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot)\le f$$

- Type-1 Wasserstein DRO. Suppose the loss function $l(\theta,)$ is Lipschitz continuous, and we choose γ to be the W-1 distance. (D) is then equivalent to Lipschitz regularized ERM

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta,\xi) + \epsilon\mathrm{Lip}(l).$$

- Distributionally robust logistic regression (with exact regularization coefficient)

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta^{\top}x_i, y_i) + \epsilon\cdot\|\theta\|_2.$$

- Empirical gradient penalty for DNN

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(f_\theta(x_i), y_i) + \lambda\cdot\max_{i}\|\nabla f_\theta(x_i)\|_2.$$

Special cases of DRO dual program

$$\text{(P)}\ \min_{\theta}\ \sup_{\gamma(P,\hat{P})\le\epsilon}\ \mathbb{E}_{P} l(\theta,\xi) \quad\Longleftrightarrow\quad \text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot)\le f$$

- Type-1 Wasserstein DRO. Suppose the loss function $l(\theta,)$ is Lipschitz continuous, and we choose γ to be the W-1 distance. (D) is then equivalent to Lipschitz regularized ERM

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta,\xi) + \epsilon\mathrm{Lip}(l).$$

- Distributionally robust logistic regression (with exact regularization coefficient)

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(\theta^{\top}x_i, y_i) + \epsilon\cdot\|\theta\|_2.$$

- Empirical gradient penalty for DNN

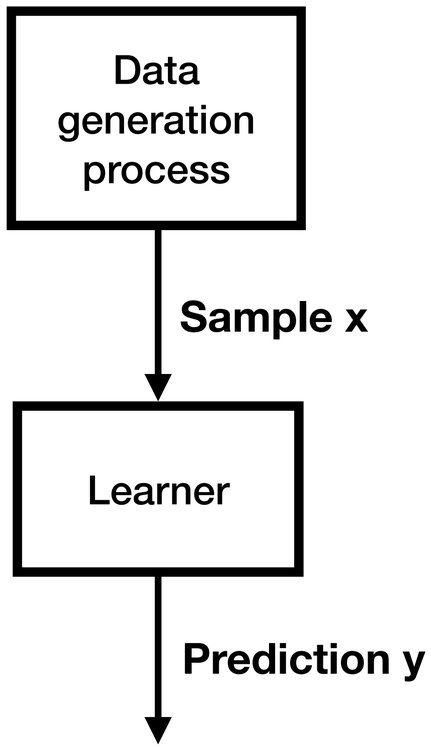
$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N} l(f_\theta(x_i), y_i) + \lambda\cdot\max_{i}\|\nabla f_\theta(x_i)\|_2.$$

- Kernel DRO. If we choose to be the MMD, then solving DRO becomes a kernel-based learning

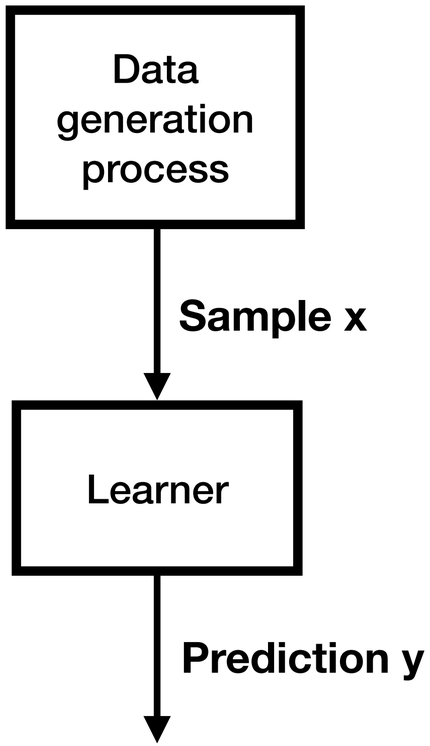
$$\text{(D)}\ \min_{\theta, f\in\mathscr{H}}\ \mathbb{E}_{\hat{P}} f + \epsilon\|f\|_{\mathscr{H}} \quad \text{s.t.}\ l(\theta,\cdot)\le f.$$

Multi-stage robust decision-making

SL

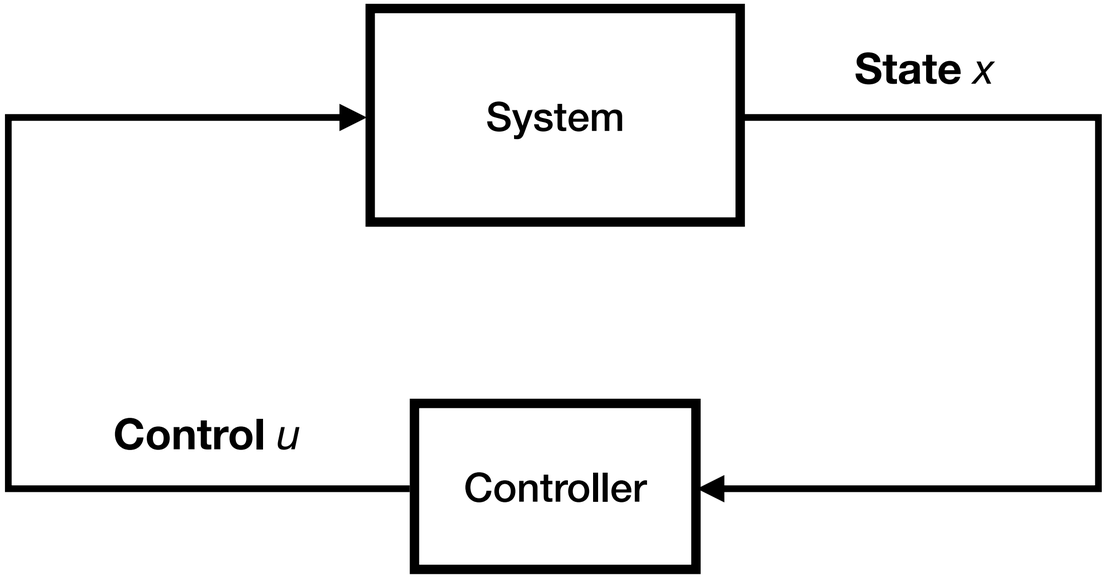


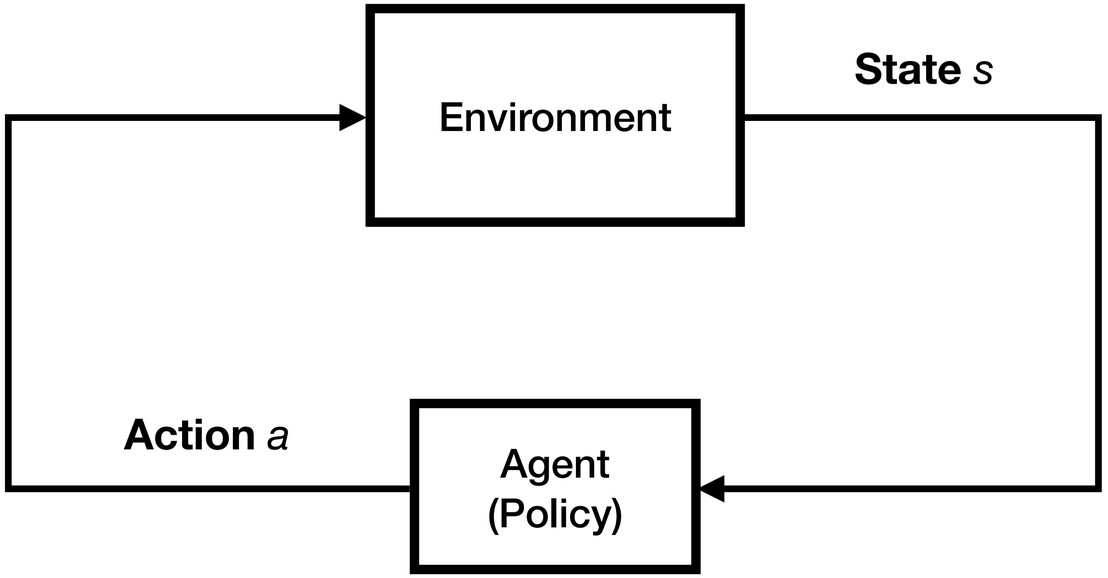
SL



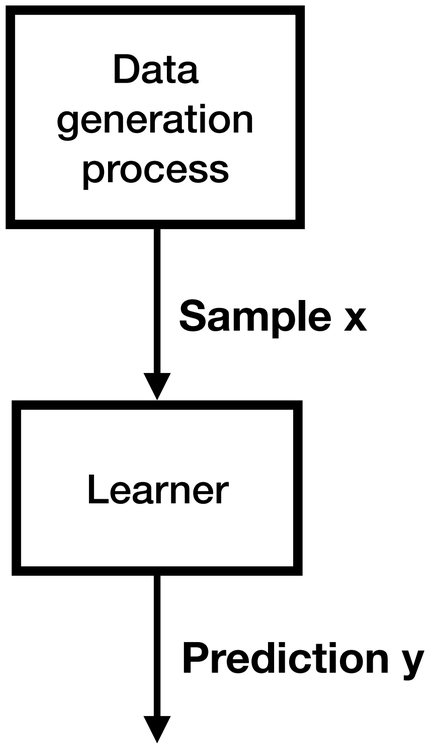
Control

RL

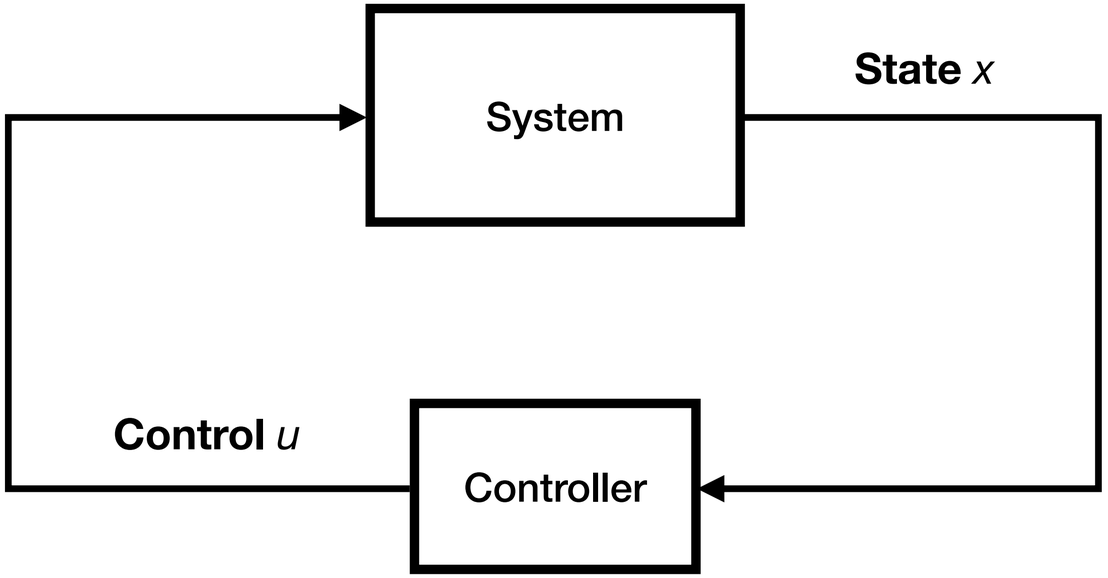




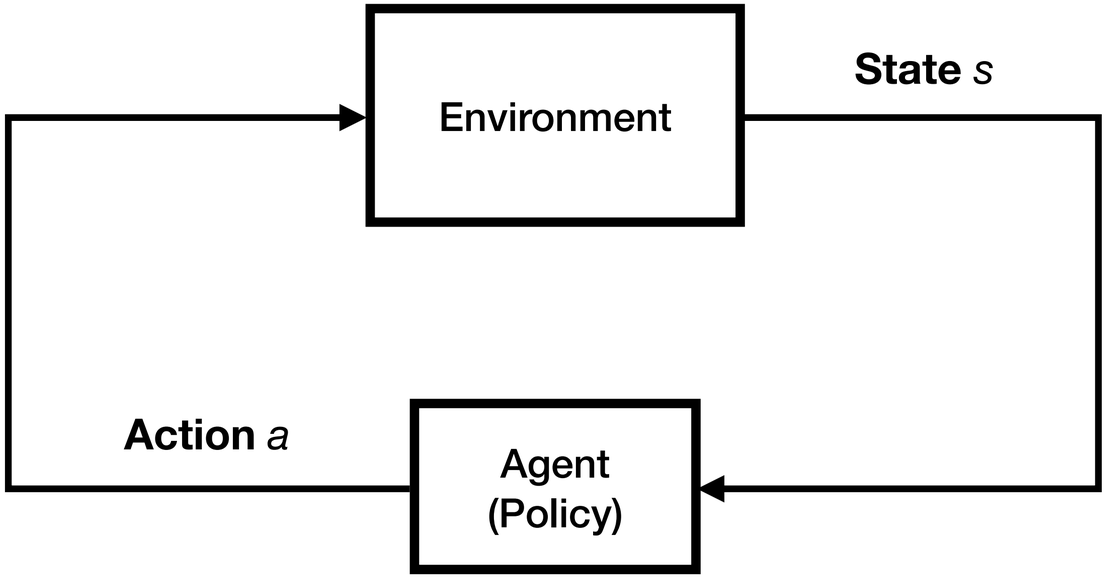
SL



Control



RL



Distinction: feedback in systems causes **distribution shift** in the data-generation processes

Reinforcement learning problem

$$\max_{\mu} \quad \mathbb{E} \sum_{t=1}^{T} \textbf{reward}(x_t, \mu(x_t))$$

Reinforcement learning problem

$$\max_{\mu} \quad \mathbb{E} \sum_{t=1}^{T} \textbf{reward}(x_t, \mu(x_t))$$

$$\min_{u_{1:T}} \quad \sum_{t=1}^{T} \textbf{cost}(x_t, u_t)$$

Optimal control problem

Reinforcement learning problem

$$\max_{\mu} \quad \mathbb{E} \sum_{t=1}^{T} \textbf{reward}(x_t, \mu(x_t))$$

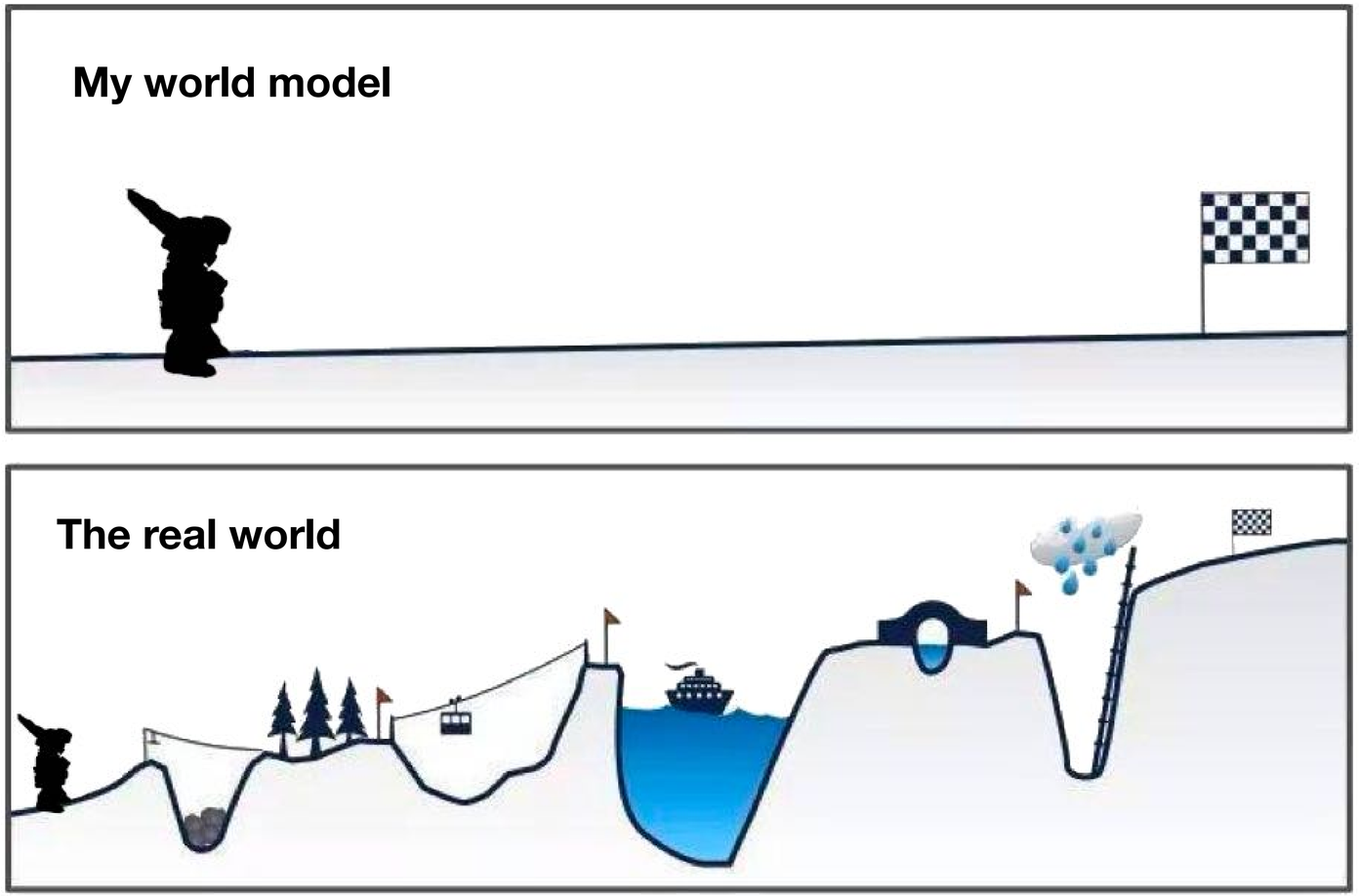
$$\min_{u_{1:T}} \quad \sum_{t=1}^{T} \textbf{cost}(x_t, u_t)$$

Optimal control problem

Why do we need to be robust in RL/control?

Why do we need to be robust in RL/control?

Why do we need to be robust in RL/control?



Model Predictive Control

Model Predictive Control

Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

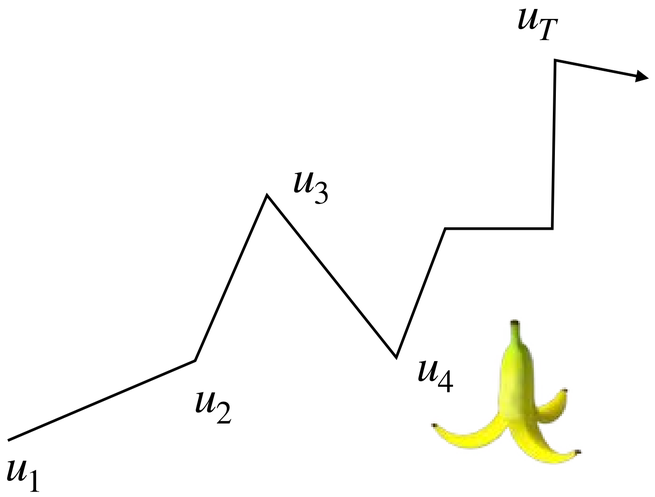
$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0,\ldots,N-1$$

Model Predictive Control



Optimal control problem

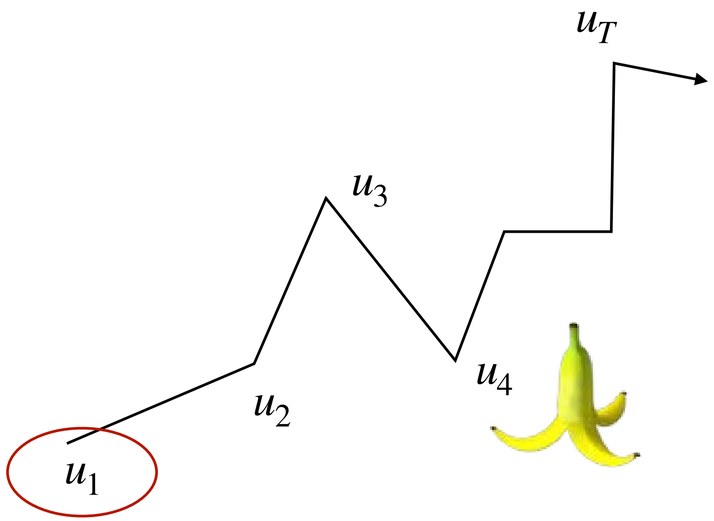
(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

subject to

$$x_{t+1} = f(x_t, u_t)$$
$$c(x_t, u_t) \leq 0$$
$$x_0 = s$$
$$t = 0,\ldots,N-1$$

Model Predictive Control



Optimal control problem

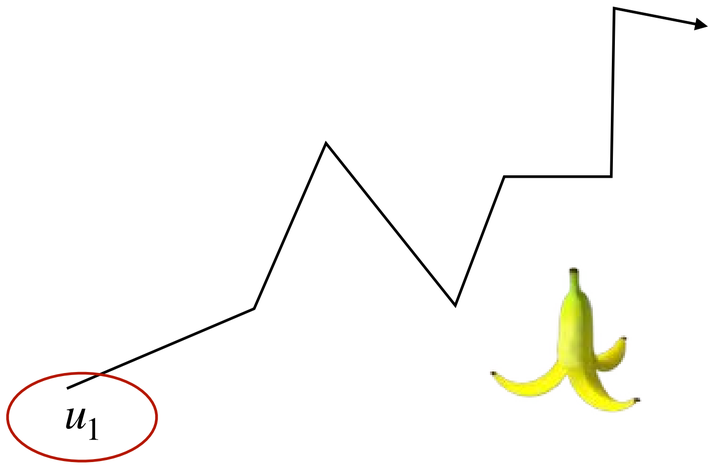
(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

subject to

$$x_{t+1} = f(x_t, u_t)$$
$$c(x_t, u_t) \leq 0$$
$$x_0 = s$$
$$t = 0,\ldots,N-1$$

Model Predictive Control



Optimal control problem

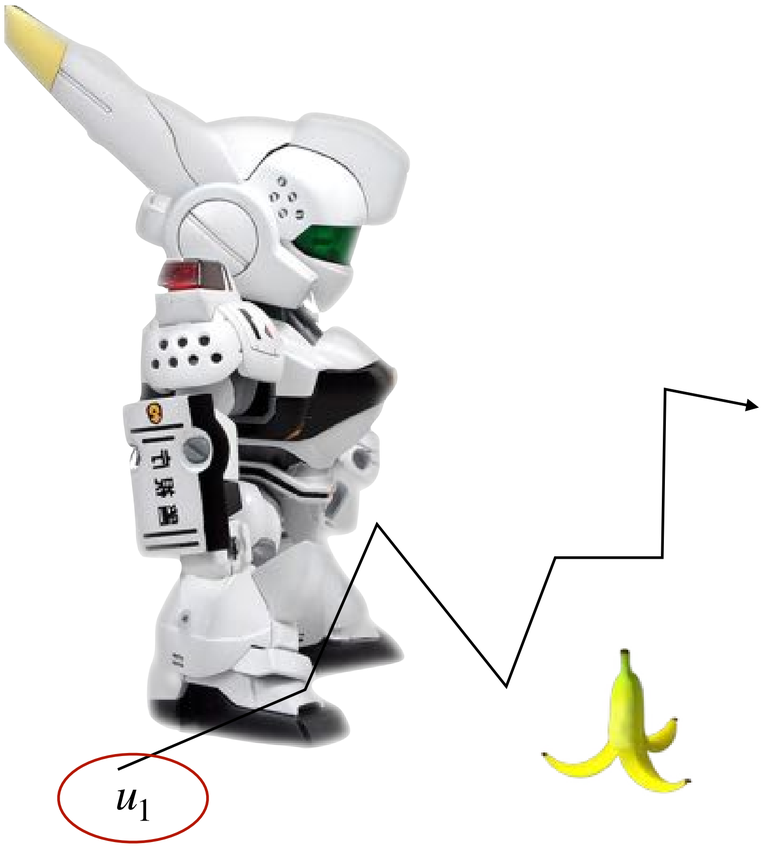
(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

subject to

$$x_{t+1} = f(x_t, u_t)$$
$$c(x_t, u_t) \leq 0$$
$$x_0 = s$$
$$t = 0,\ldots,N-1$$

Model Predictive Control



Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0,\ldots,N-1$$

Model Predictive Control

Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0,\ldots,N-1$$

Model Predictive Control

Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

subject to

$$x_{t+1} = f(x_t, u_t)$$
$$c(x_t, u_t) \leq 0$$
$$x_0 = s$$
$$t = 0,\ldots,N-1$$

Model Predictive Control

Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0,\ldots,N-1$$

You are using MPC when you drive or catch a train.

Model Predictive Control

Optimal control problem

(discrete-time)

$$\underset{u_0,\dots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0, \dots, N-1$$

You are using MPC when you drive or catch a train.

Learning-based MPC is an active research area!

Robust and stochastic model predictive control

Robust and stochastic model predictive control

Optimal control problem

(discrete-time)

$$\underset{u_0,\ldots,u_{N-1}}{\text{minimize}} \quad \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N)$$

$$\text{subject to} \quad x_{t+1} = f(x_t, u_t)$$

$$c(x_t, u_t) \leq 0$$

$$x_0 = s$$

$$t = 0,\ldots,N-1$$

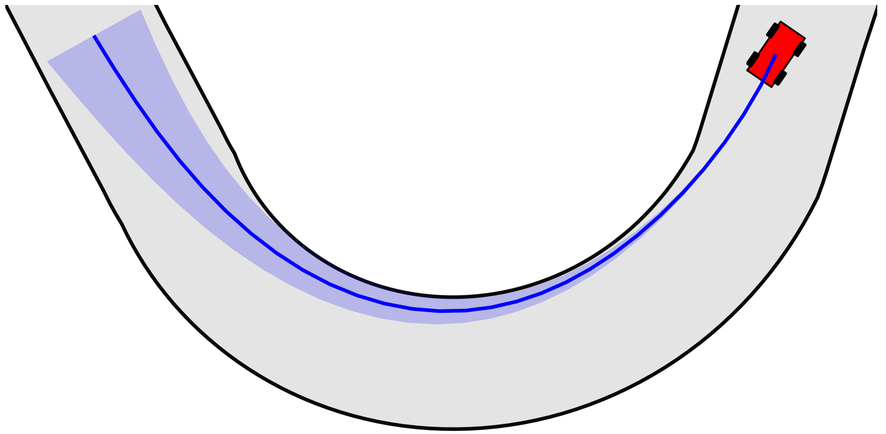


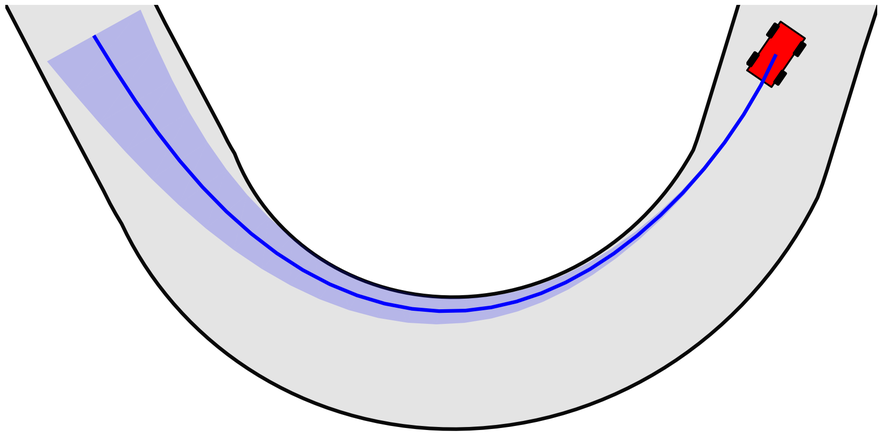
Figure credit: Hewing et al.'18, Z. et al.'20

Robust and stochastic model predictive control

Optimal control problem

(discrete-time)

$$\begin{aligned} \underset{u_0,\dots,u_{N-1}}{\text{minimize}} \quad & \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N) \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t) \\ & c(x_t, u_t) \leq 0 \\ & x_0 = s \\ & t = 0,\dots,N-1 \end{aligned}$$



Guaranteed Margins for LQG Regulators

JOHN C. DOYLE

***Abstract*—There are none.**

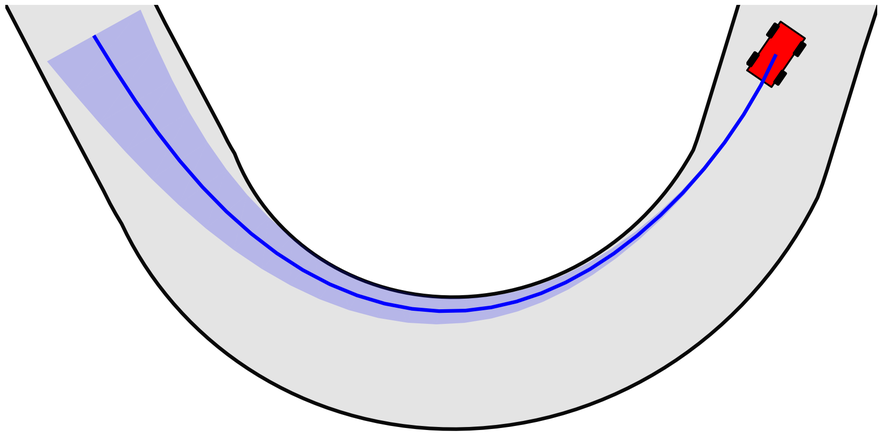
Figure credit: Hewing et al.'18, Z. et al.'20

Robust and stochastic model predictive control

Optimal control problem

(discrete-time)

$$\begin{aligned} \underset{u_0,\dots,u_{N-1}}{\text{minimize}} \quad & \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N) \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t) \\ & c(x_t, u_t) \leq 0 \\ & x_0 = s \\ & t = 0,\dots,N-1 \end{aligned}$$



Guaranteed Margins for LQG Regulators

JOHN C. DOYLE

***Abstract*—There are none.**

Optimization (alone) is not robust.

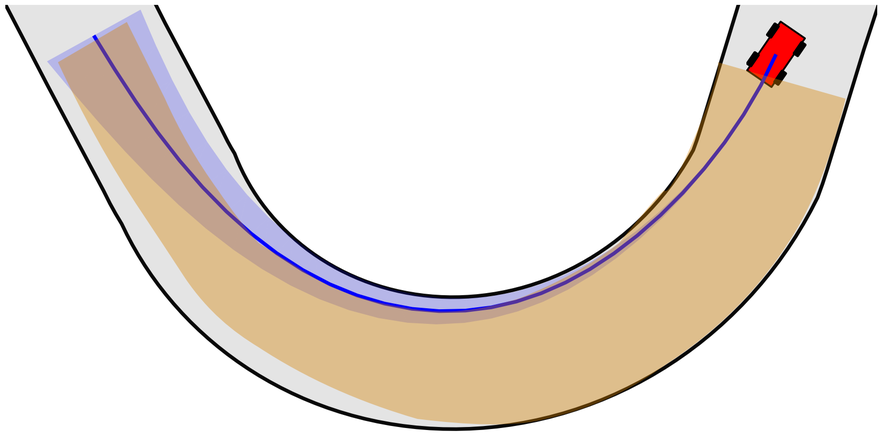
Figure credit: Hewing et al.'18, Z. et al.'20

Robust and stochastic model predictive control

Optimal control problem

(discrete-time)

$$\begin{aligned} \underset{u_0,\dots,u_{N-1}}{\text{minimize}} \quad & \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N) \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t) \\ & c(x_t, u_t) \leq 0 \\ & x_0 = s \\ & t = 0,\dots,N-1 \end{aligned}$$



- **Constraint Tightening strategy**: Drive within a narrower "*tube*" than the original track!

Guaranteed Margins for LQG Regulators

JOHN C. DOYLE

***Abstract*—There are none.**

Optimization (alone) is not robust.

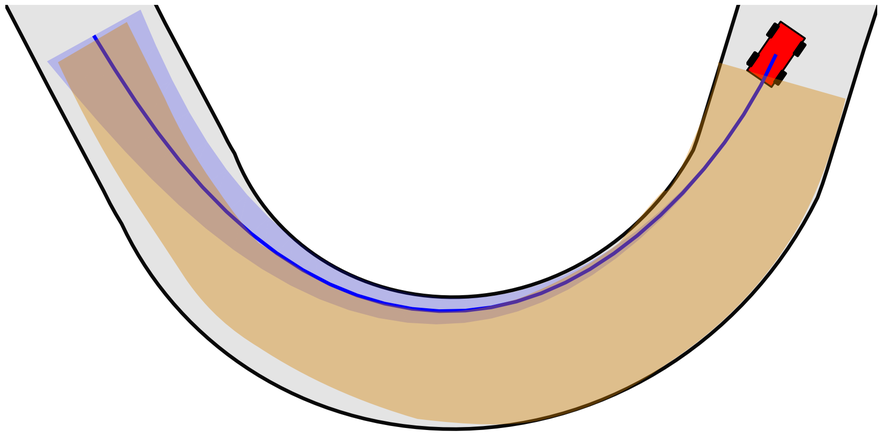
Figure credit: Hewing et al.'18, Z. et al.'20

Robust and stochastic model predictive control

Optimal control problem

(discrete-time)

$$\begin{aligned} \underset{u_0,\dots,u_{N-1}}{\text{minimize}} \quad & \sum_{t=1}^{N-1} l_t(x_t, u_t) + l_N(x_N) \\ \text{subject to} \quad & x_{t+1} = f(x_t, u_t) \\ & c(x_t, u_t) \leq 0 \\ & x_0 = s \\ & t = 0,\dots,N-1 \end{aligned}$$



Guaranteed Margins for LQG Regulators

JOHN C. DOYLE

***Abstract*—There are none.**

Optimization (alone) is not robust.

- **Constraint Tightening strategy**: Drive within a narrower "*tube*" than the original track!
- In practice, this often boils down to replacing the constraint $c(x_t, u_t) \leq 0$ by

  $c(x_t, u_t) + b \leq 0$ for some $b > 0$.

Figure credit: Hewing et al.'18, Z. et al.'20

Learning for control:
Three main classes of methods

Learning for control: Three main classes of methods

- Learning-based control, a.k.a., model-based reinforcement learning
  - Approximate the system model, then solve OCP

Learning for control: Three main classes of methods

- Learning-based control, a.k.a., model-based reinforcement learning
  - Approximate the system model, then solve OCP
- Imitation learning
  - Directly learning a control policy from demonstration

Learning for control: Three main classes of methods

- Learning-based control, a.k.a., model-based reinforcement learning
  - Approximate the system model, then solve OCP
- Imitation learning
  - Directly learning a control policy from demonstration
- Model-free reinforcement learning
  - Policy/value iteration using Bellman, policy search, Q-learning, …

Learning for control:
Three main classes of methods

- Learning-based control, a.k.a., model-based reinforcement learning
  - Approximate the system model, then solve OCP
- Imitation learning
  - Directly learning a control policy from demonstration
- Model-free reinforcement learning
  - Policy/value iteration using Bellman, policy search, Q-learning, …

They all suffer from distribution shift!

Covariate shift in imitation learning

Covariate shift in imitation learning

Train

Formula1Net π_θ

Covariate shift in imitation learning

Train

Formula1Net π_θ

(IL): $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_e}(dx),$

Covariate shift in imitation learning

Train

Formula1Net π_θ

(IL): $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_e}(dx),$

What can go wrong?

Covariate shift in imitation learning

Train

Formula1Net π_θ

(IL): $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_e}(dx),$

What can go wrong?

- **Shift in the induced state distributions μ^π.**

Covariate shift in imitation learning

Train

Formula1Net π_θ

(IL): $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_e}(dx),$

What can go wrong?

- **Shift in the induced state distributions μ^π.**
- **Since we care about the performance**

  $\int l(x, \pi_\theta(x)) \mu^{\pi_\theta}(dx)$

  not $\int l(x, \pi_\theta(x)) \mu^{\pi_e}(dx)$.

  But π_θ, π_e are only similar under the μ^{π_e}.

Covariate shift in imitation learning

Train

Formula1Net π_θ

(IL): $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_e}(dx),$

What can go wrong?

- **Shift in the induced state distributions μ^π.**
- **Since we care about the performance**

  $$\int l(x, \pi_\theta(x)) \mu^{\pi_\theta}(dx)$$

  not $\int l(x, \pi_\theta(x)) \mu^{\pi_e}(dx)$.

  But π_θ, π_e are only similar under the μ^{π_e}.
- **Same goes for learning dynamics for control**

Fix: covariate shift in imitation learning

Fix: covariate shift in imitation learning

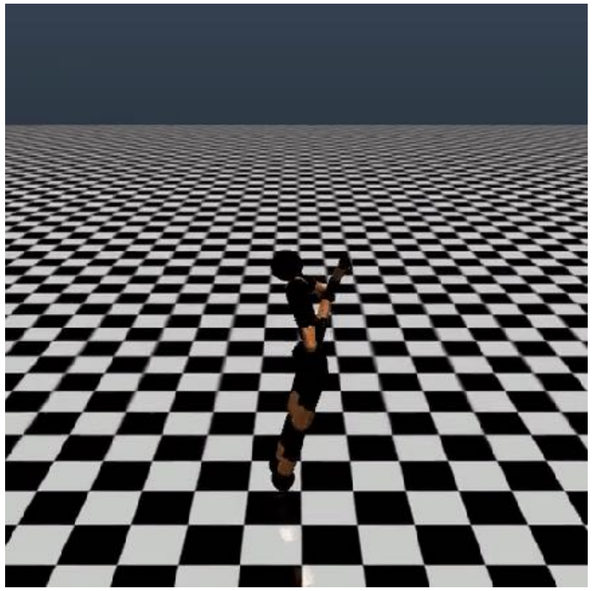
DAgger (Data Aggregation)
1. Train a policy π_θ from expert data D
2. Run π_θ (on the robot) to obtain new states
3. Ask the expert to label the states with actions and obtain new data D'
4. Add the labeled data to the training set $D = D \cup D'$ (aggregation)

Fix: covariate shift in imitation learning

DAgger (Data Aggregation)
1. Train a policy π_θ from expert data D
2. Run π_θ (on the robot) to obtain new states
3. Ask the expert to label the states with actions and obtain new data D'
4. Add the labeled data to the training set $D = D \cup D'$ (aggregation)

Want: $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_\theta}(dx)$

Fix: covariate shift in imitation learning



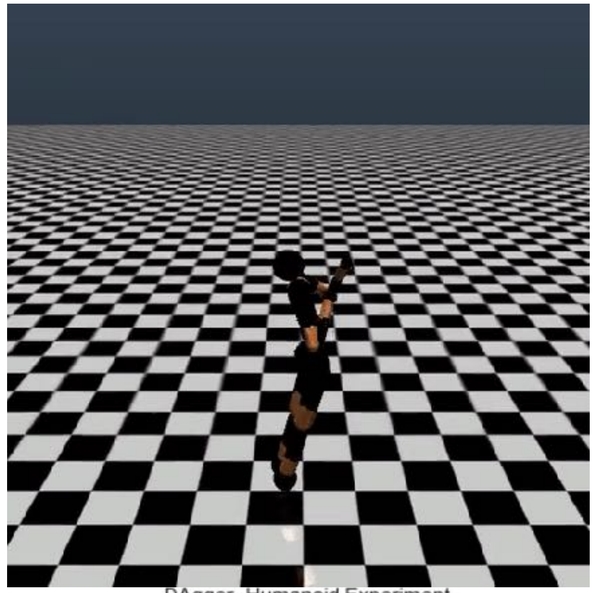
DAgger (Data Aggregation)
1. Train a policy π_θ from expert data D
2. Run π_θ (on the robot) to obtain new states
3. Ask the expert to label the states with actions and obtain new data D'
4. Add the labeled data to the training set $D = D \cup D'$ (aggregation)

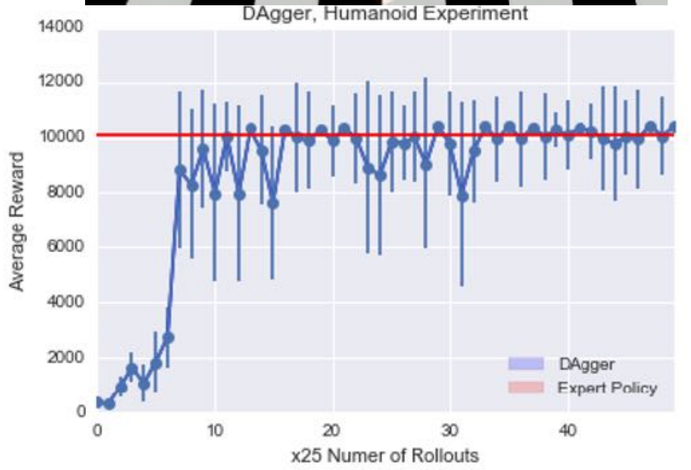
Want: $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_\theta}(dx)$

Fix: covariate shift in imitation learning

DAgger (Data Aggregation)
1. Train a policy π_θ from expert data D
2. Run π_θ (on the robot) to obtain new states
3. Ask the expert to label the states with actions and obtain new data D'
4. Add the labeled data to the training set $D = D \cup D'$ (aggregation)

Want: $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_\theta}(dx)$

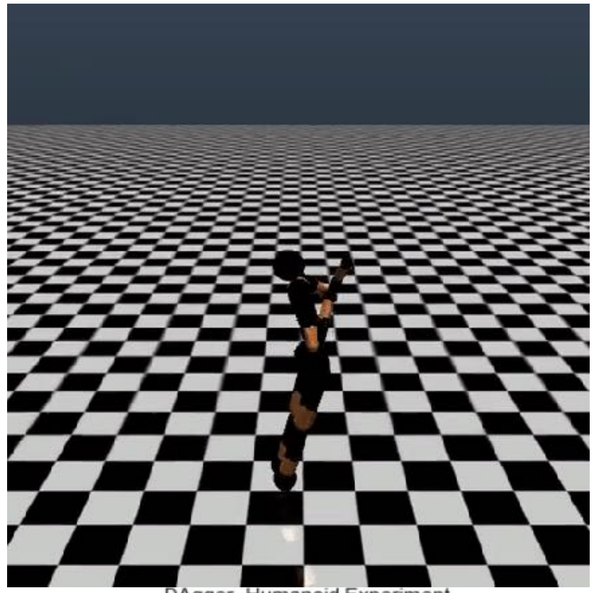


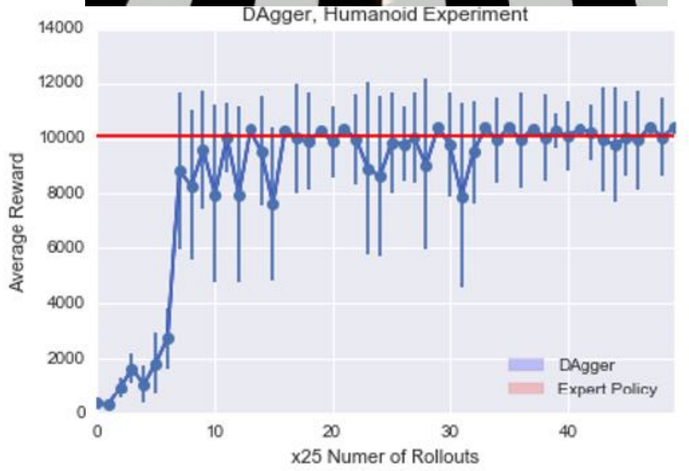


Fix: covariate shift in imitation learning

DAgger (Data Aggregation)
1. Train a policy π_θ from expert data D
2. Run π_θ (on the robot) to obtain new states
3. Ask the expert to label the states with actions and obtain new data D'
4. Add the labeled data to the training set $D = D \cup D'$ (aggregation)

Want: $\min_{\theta \in \Theta} \int \|\pi_\theta(x) - \pi_e(x)\|^2 \mu^{\pi_\theta}(dx)$





Other related topics: off-policy/offline RL

Conclusions

Conclusions

- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.

Conclusions

- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.

Conclusions

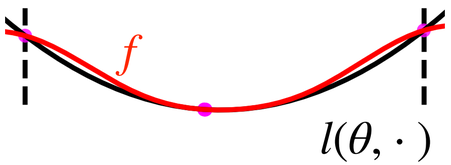
- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.
- We looked at the robustness and distribution shift of multi-stage decision-making and control.

Conclusions

- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.
- We looked at the robustness and distribution shift of multi-stage decision-making and control.
- We have applied the robustness principle for decision-making :)

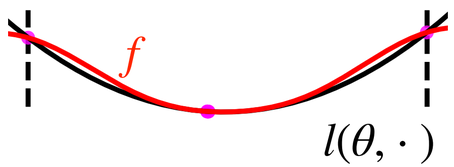
Conclusions

- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.
- We looked at the robustness and distribution shift of multi-stage decision-making and control.
- We have applied the robustness principle for decision-making :)
- Takeaway
  - **Flatten the curve, smooth is robust**



Conclusions

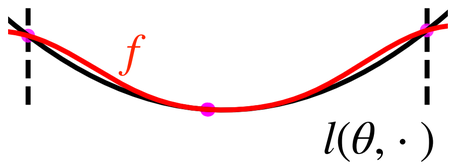
- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.
- We looked at the robustness and distribution shift of multi-stage decision-making and control.
- We have applied the robustness principle for decision-making :)
- Takeaway
  - **Flatten the curve, smooth is robust**



Other topics

Conclusions

- We have introduced distributional metrics such as the MMD, Wasserstein distance as tools for comparing distributions.
- We learned about RO and DRO, and the principles of robustness against distribution shift in optimization and machine learning.
- We looked at the robustness and distribution shift of multi-stage decision-making and control.
- We have applied the robustness principle for decision-making :)
- Takeaway
  - **Flatten the curve, smooth is robust**



Other topics

- Generalizaiton and statistical bounds of DRO, minimax rate of non-parametric statistics
- Distributionally robust MPC
- Distributionally robust dynamics learning
- Off-policy RL, off-line RL
- Multi-stage DRO
- Adversarial attacks
- Causal inference
- Double descent
- Robust generative modeling
- You tell me!

Thank you!

J.J. (Jia-Jie) Zhu
jj-zhu.github.io
zhu@wias-berlin.de

Weierstrass-Institute, Berlin &
Max-Planck-Institute, Tübingen
Germany