Adversarially Robust Kernel Smoothing

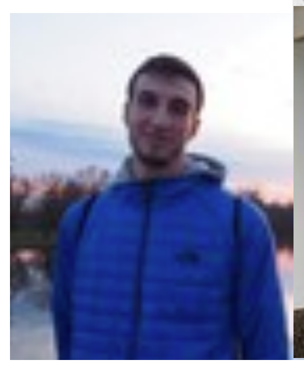
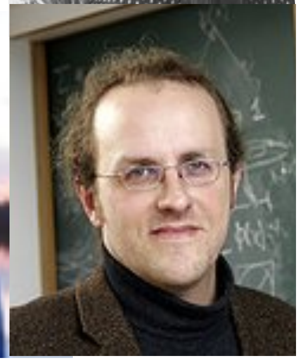
Jia-Jie Zhu[1,3], Christina Kouridi[2,3],
Yassine Nemmour[3], Bernhard Schölkopf[3]

[1]Weierstrass Institute for Applied Analysis and Stochastics
Berlin, Germany

[2]InstaDeep Ltd.
London, United Kingdom

[3]Max Planck Institute for Intelligent Systems
Tübingen, Germany



The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)
March, 2022

Learning under distribution shift

Learning under distribution shift

Empirical Risk Minimization

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

Learning under distribution shift

Empirical Risk Minimization

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

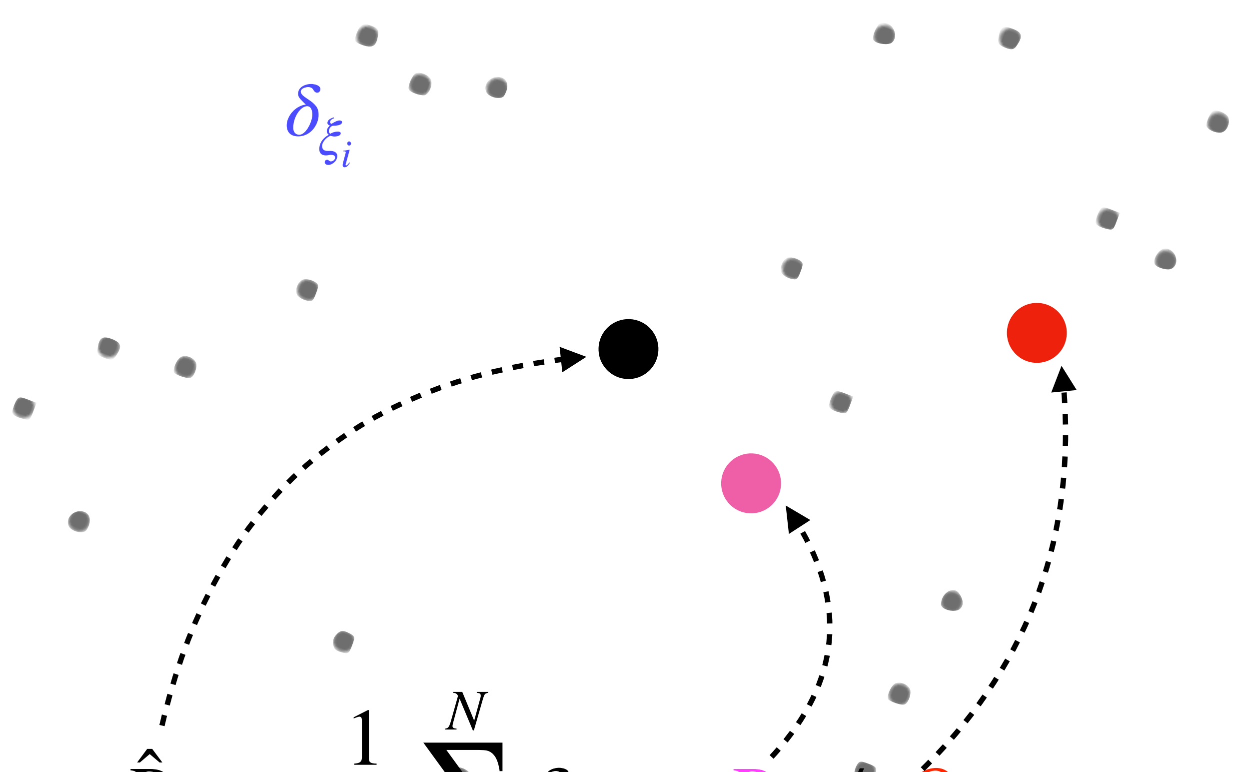
- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under <u>data distribution shifts</u>, when $Q \; (\neq P_0)$



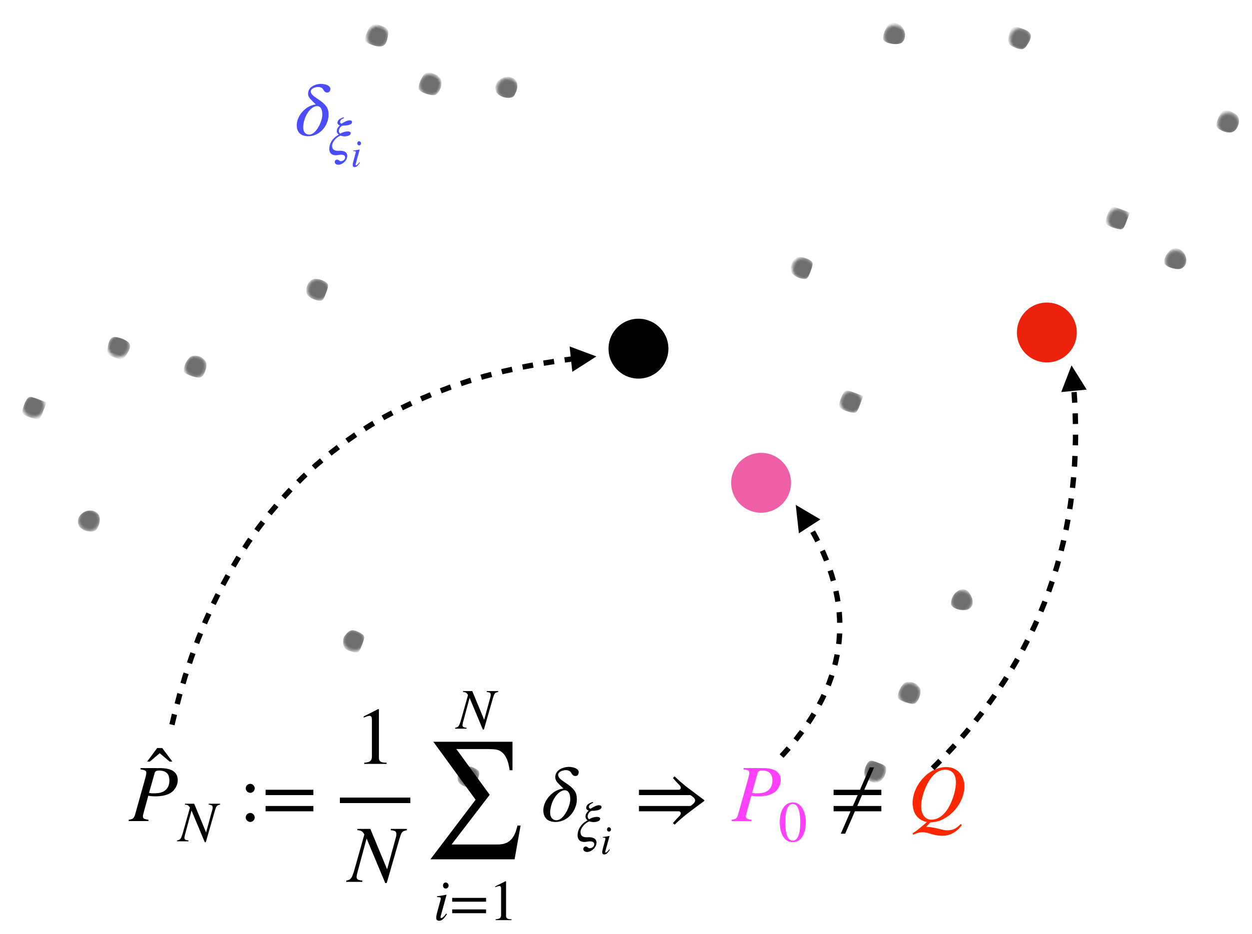
$$\hat{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under <u>data distribution shifts</u>, when Q $(\neq P_0)$



Distributionally Robust Learning

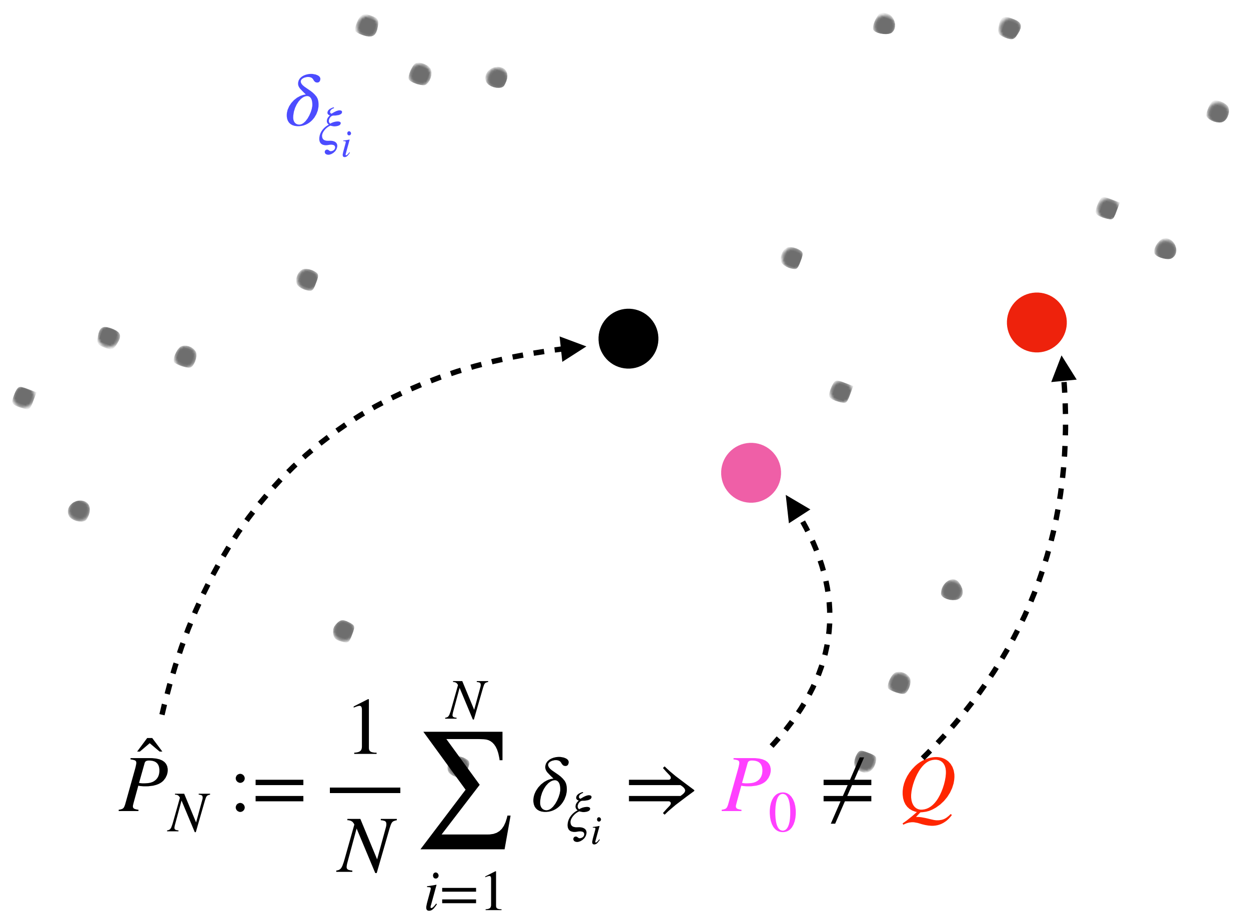
$$\min_{\theta} \sup_{Q \in \mathcal{M}} \mathbb{E}_Q L(\theta, \xi)$$

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under <u>data distribution shifts</u>, when Q $(\neq P_0)$



Distributionally Robust Learning

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

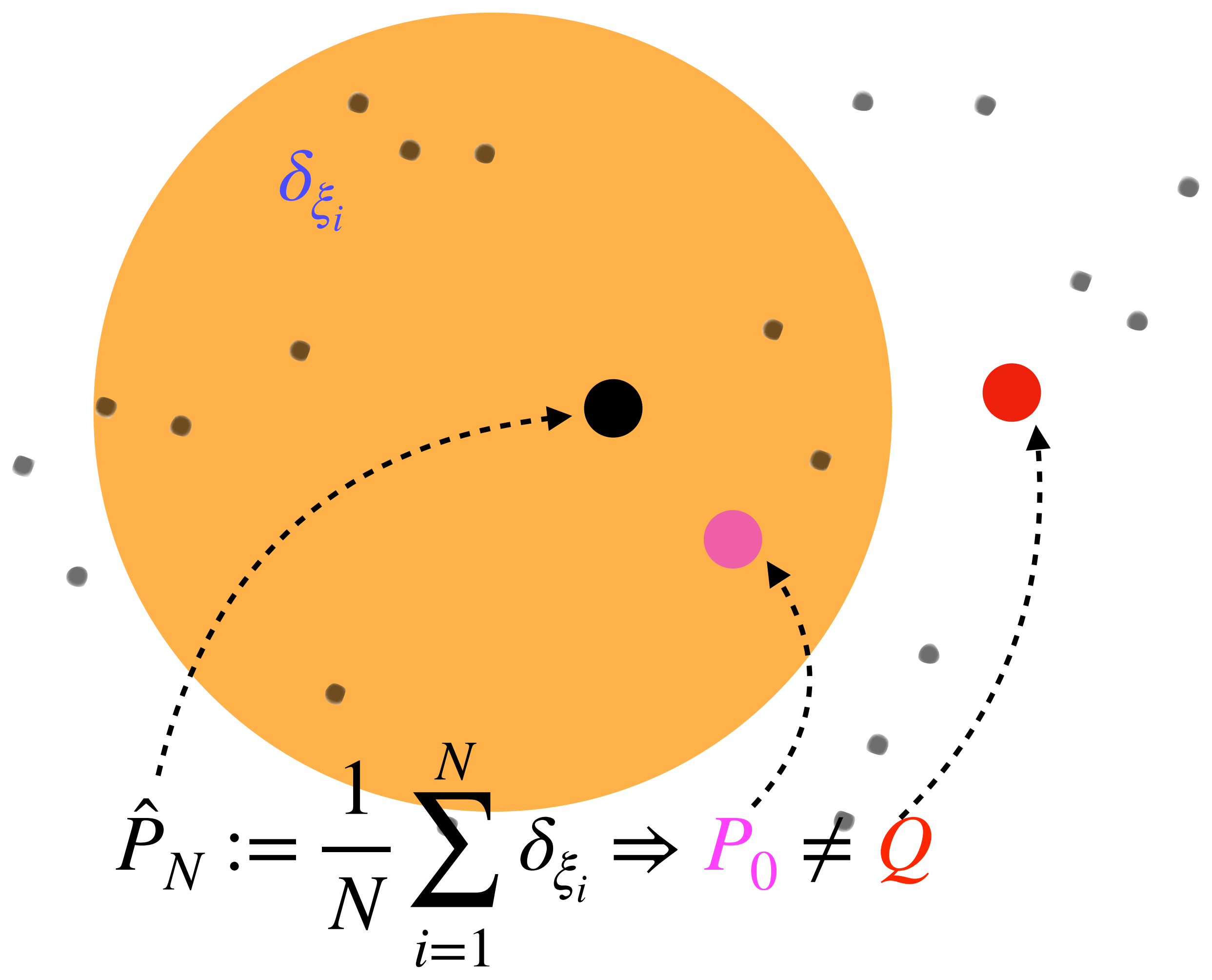
- Minimize risk under a local worst-case distribution Q

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$



Distributionally Robust Learning

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

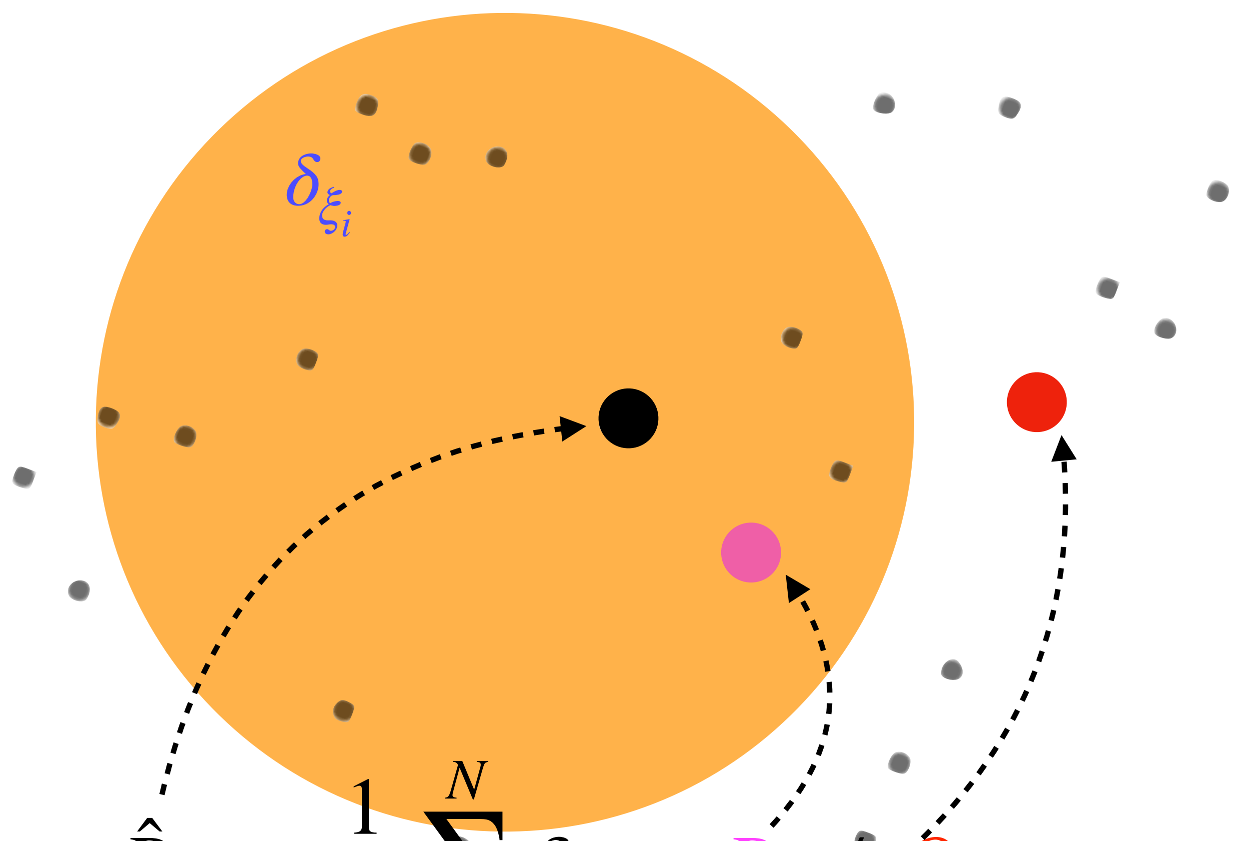
- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathscr{M}$. Example: maximum mean discrepancy-ball $\{Q : \mathrm{MMD}(Q, \hat{P}_N) \leq \rho\}$ or Wasserstein-ball

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$



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

Distributionally Robust Learning

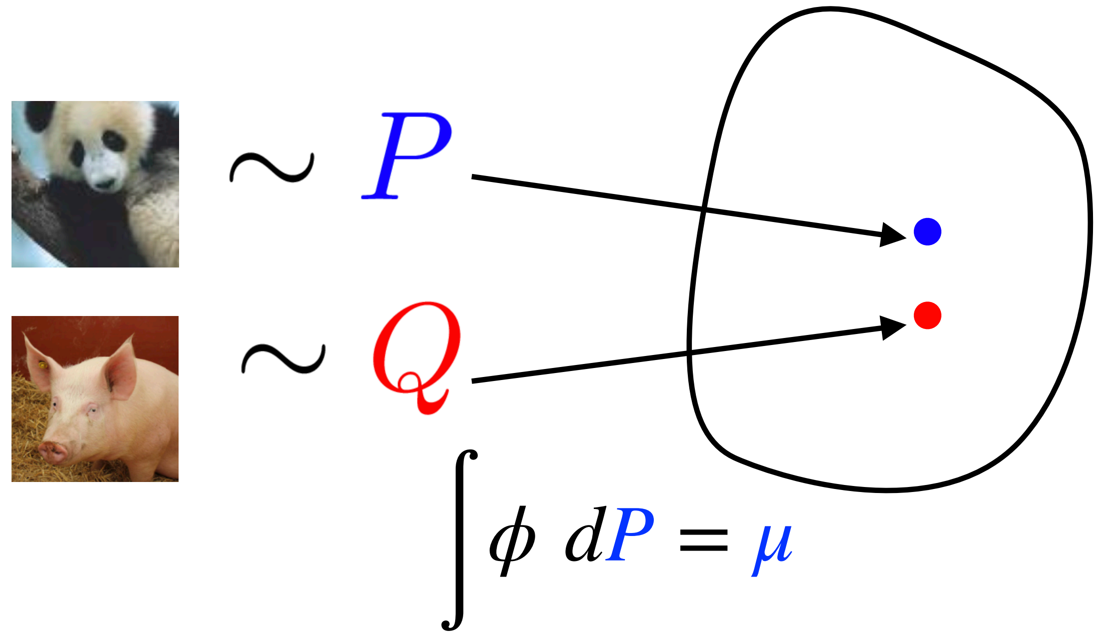
$$\min_{\theta} \sup_{Q \in \mathscr{M}} \mathbb{E}_Q L(\theta, \xi)$$

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathscr{M}$. Example: maximum mean discrepancy-ball $\{Q : \text{MMD}(Q, \hat{P}_N) \leq \rho\}$ or Wasserstein-ball

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

$$= \mathbb{E}_{x, x' \sim Q} k(x, x') + \mathbb{E}_{y, y' \sim P} k(y, y')$$

$$-2\mathbb{E}_{x \sim Q, y \sim P} k(x, y) \, .$$

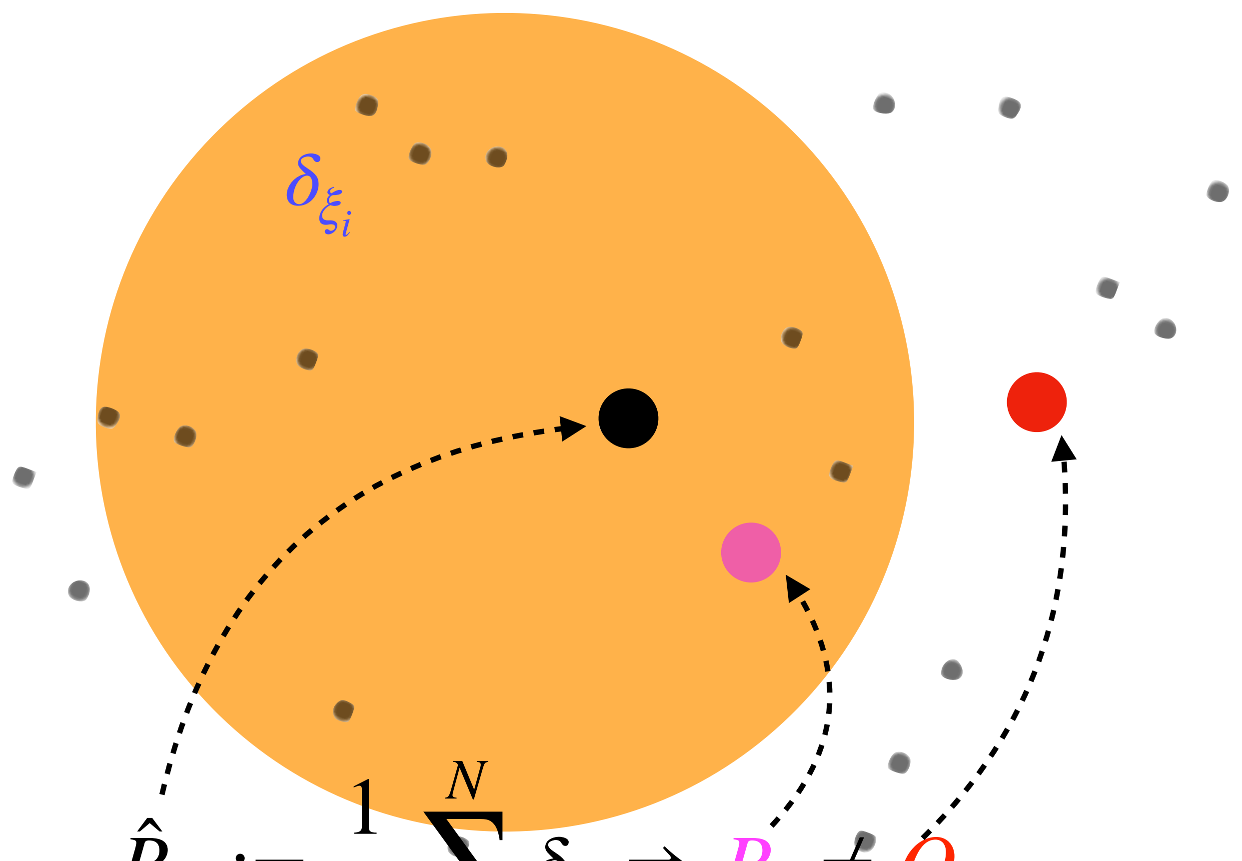


Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under data distribution shifts, when $Q\ (\neq P_0)$

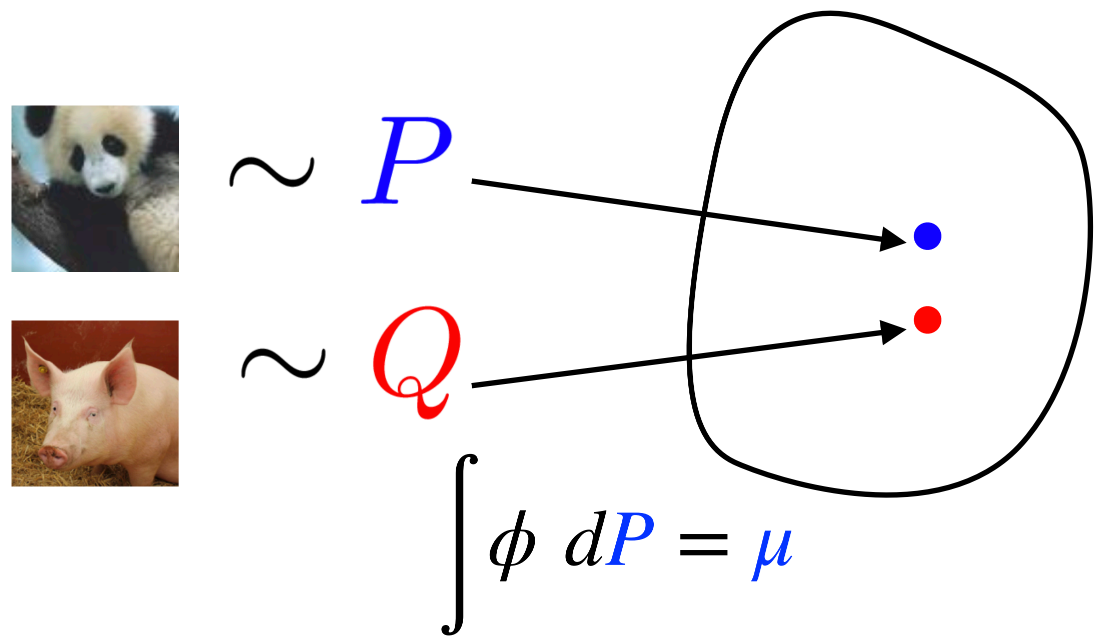


$$\hat{P}_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

Distributionally Robust Learning

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

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathcal{M}$. Example: maximum mean discrepancy-ball $\{Q : \mathrm{MMD}(Q, \hat{P}_N) \leq \rho\}$ or Wasserstein-ball
- $$\mathrm{MMD}_{\mathcal{H}}(Q,P) := \sup_{\|f\|_{\mathcal{H}}\leq 1} \int f\, d(Q-P)$$
$$= \mathbb{E}_{x,x'\sim Q} k(x,x') + \mathbb{E}_{y,y'\sim P} k(y,y')$$
$$-2\mathbb{E}_{x\sim Q, y\sim P} k(x,y)\,.$$



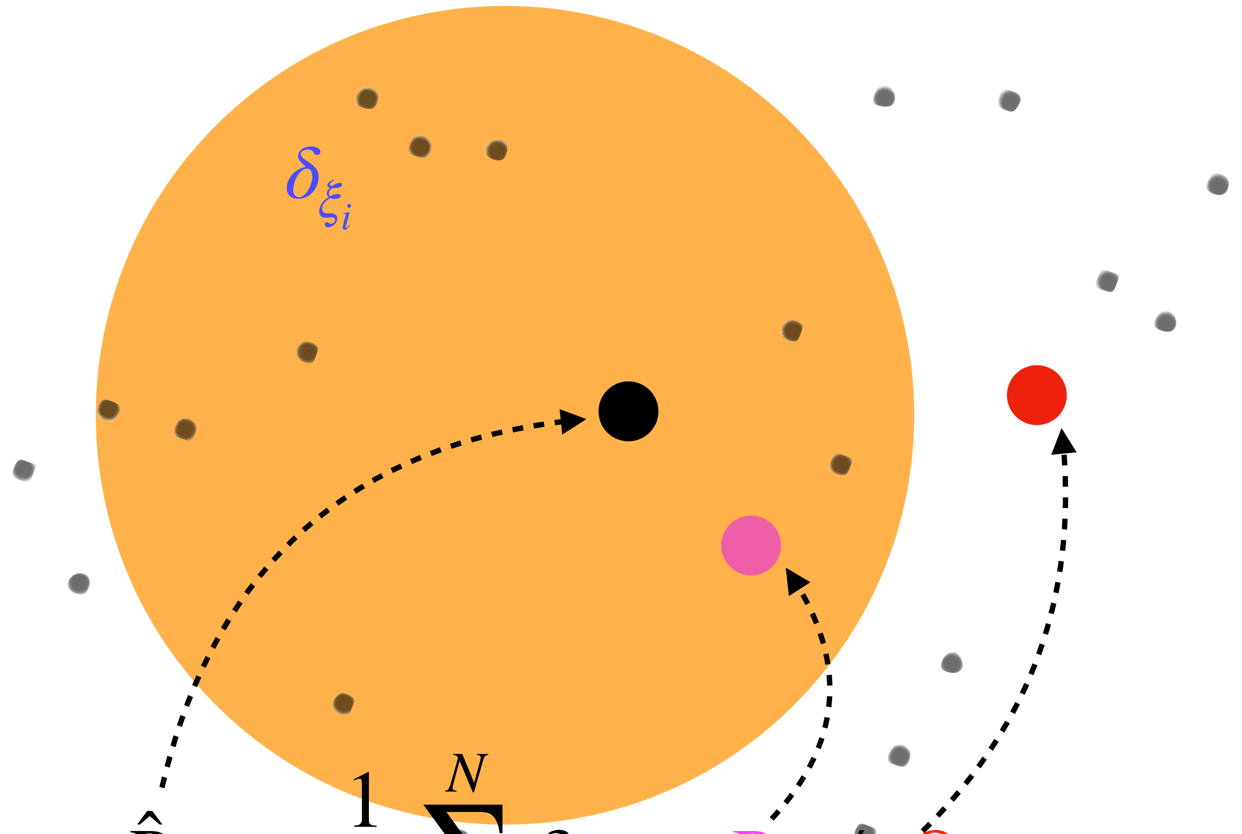
- We can bound performance under $Q\ (\neq P_0)$ beyond statistical fluctuation (classical learning theory)

Learning under distribution shift

Empirical Risk Minimization

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

- Robust under statistical fluctuation, e.g., we can bound $\mathbb{E}_{P_0} l(\hat{\theta}, \xi)$
- Not robust under data distribution shifts, when Q $(\neq P_0)$



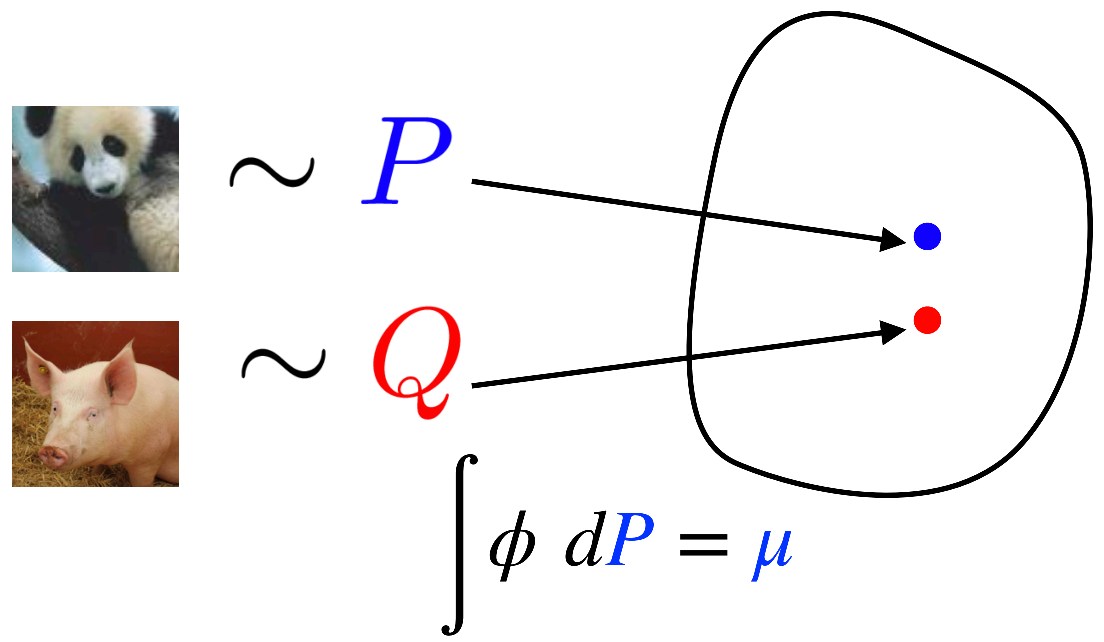
$$\hat{P}_N := \frac{1}{N} \sum_{i=1}^{N} \delta_{\xi_i} \Rightarrow P_0 \neq Q$$

Distributionally Robust Learning

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

- Minimize risk under a local worst-case distribution Q
- Distribution shift described by an ambiguity set $\mathcal{M}$. Example: maximum mean discrepancy-ball $\{Q : \mathrm{MMD}(Q, \hat{P}_N) \leq \rho\}$ or Wasserstein-ball

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



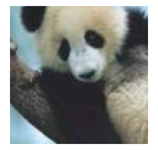
- We can bound performance under Q $(\neq P_0)$ beyond statistical fluctuation (classical learning theory)
- **Question**: how do we actually solve an MMD-constrained optimization problem? (**Non-trivial**!)

Kernel distributionally robust optimization

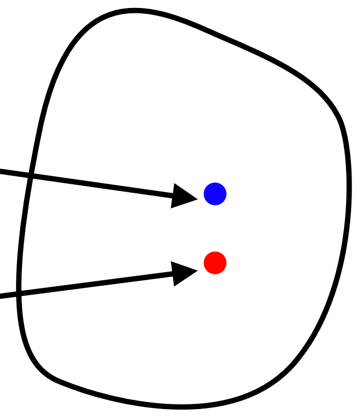
Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

 $\sim P$

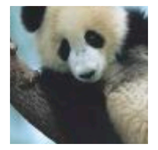
 $\sim Q$



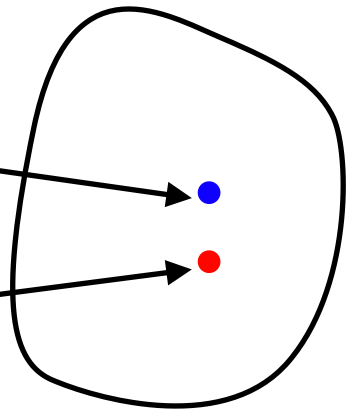
Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

 $\sim P$

 $\sim Q$



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

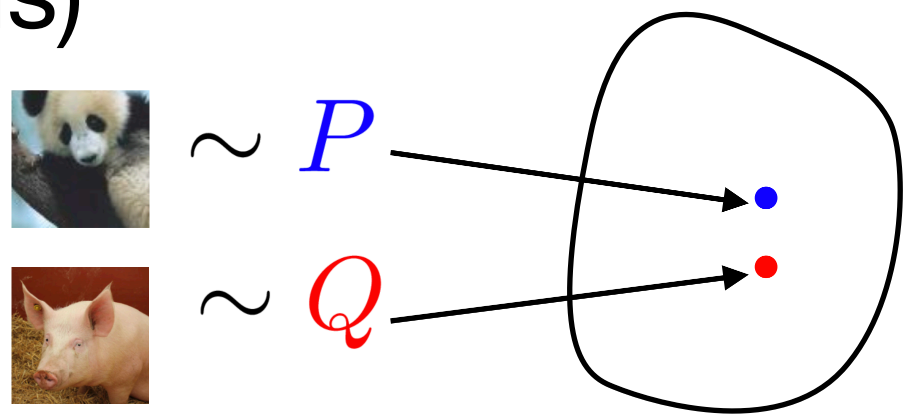
$$\text{(K)}\ \min_{\theta, f\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} f(\xi_i) + \epsilon\|f\|_{\mathcal{H}} \quad \text{subject to } l(\theta, \cdot\,) \leq f$$

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

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

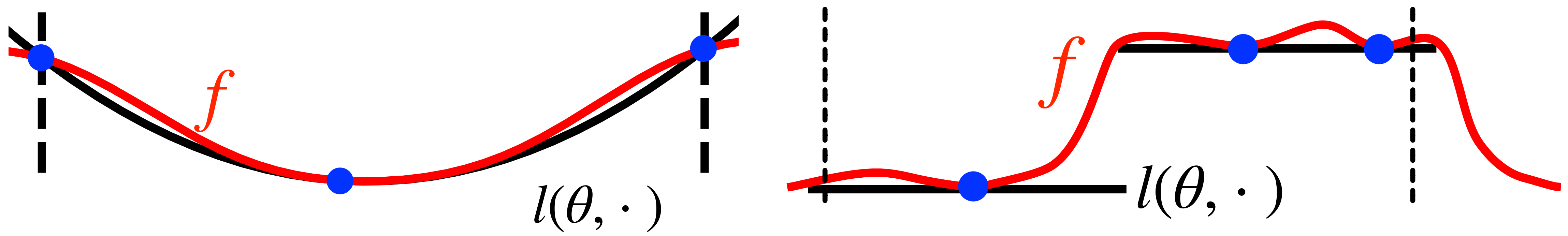


Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

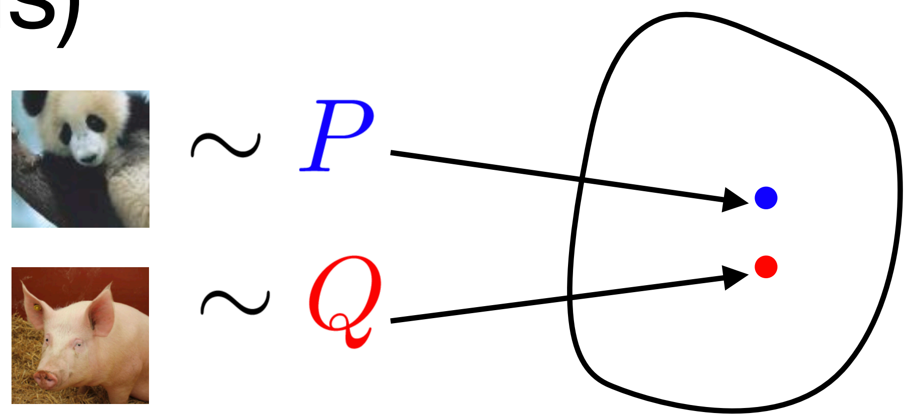
Geometric intuition: using kernel approximations as robust surrogate losses



Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

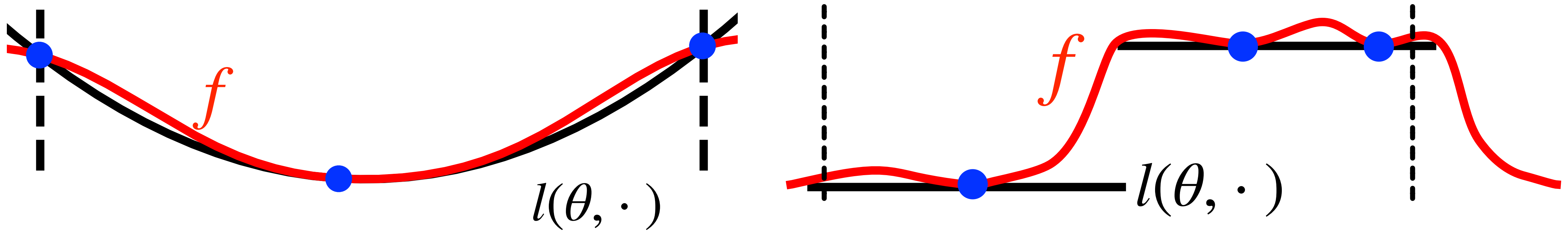


Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

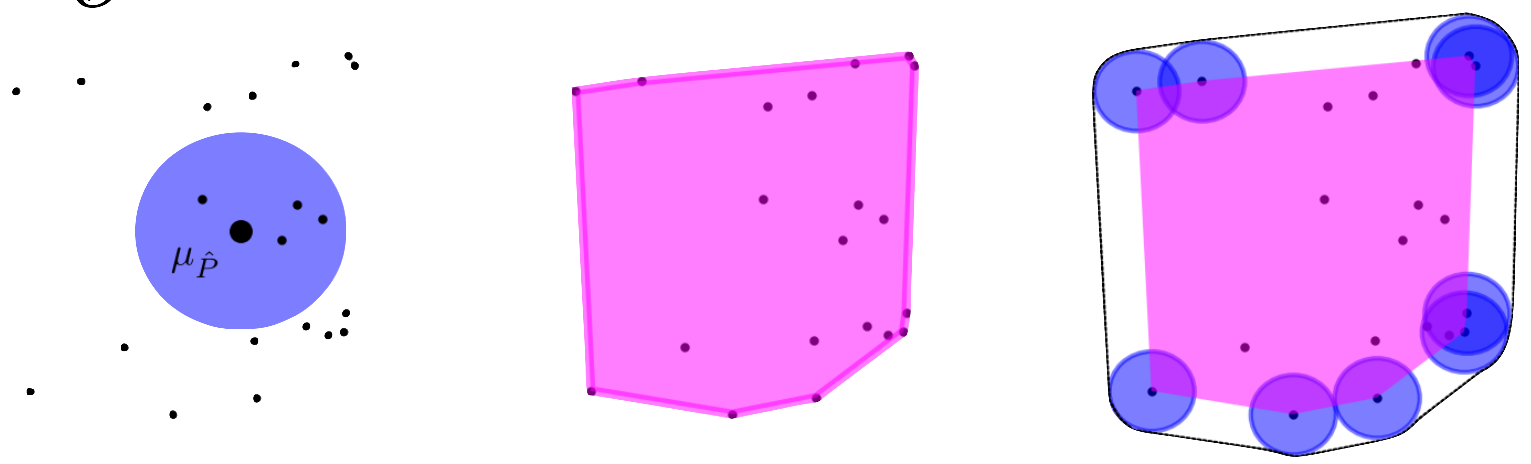
Geometric intuition: using kernel approximations as robust surrogate losses



- Extension to more general ambiguity geometry

$\min_{\theta, f\in\mathcal{H}} \delta^*_{\mu(\mathcal{M})}(f) \quad \text{subject to } l(\theta,\cdot) \le f.$

$\delta^*_{\mathscr{C}}$ denotes the support function of the set $\mathscr{C}$

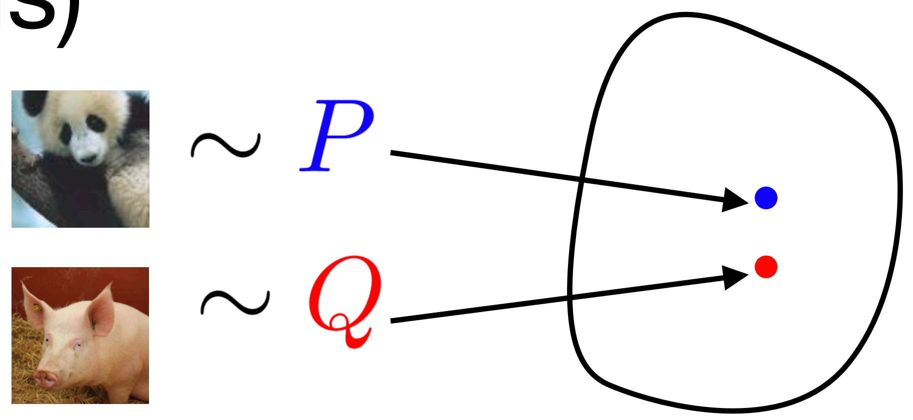


Many alg. as special cases, e.g., SVM, multi-kernel…

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

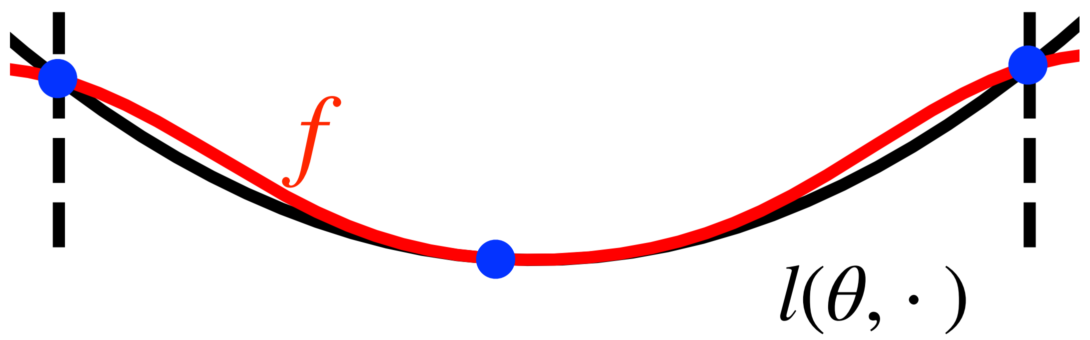


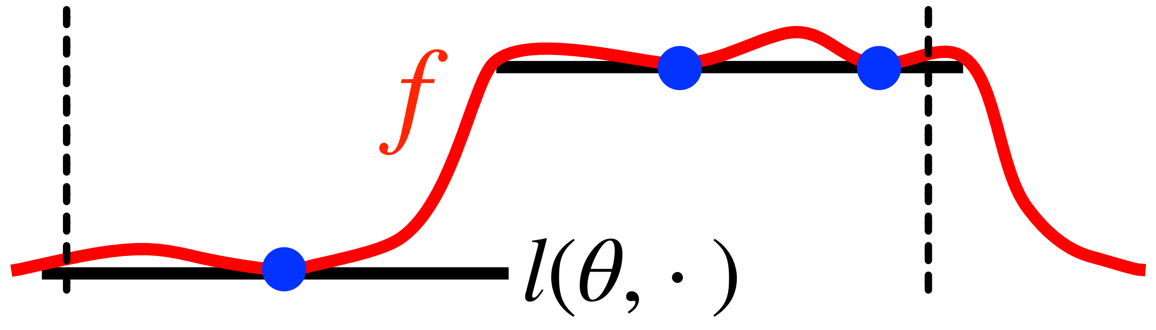
Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

Geometric intuition: using kernel approximations as robust surrogate losses

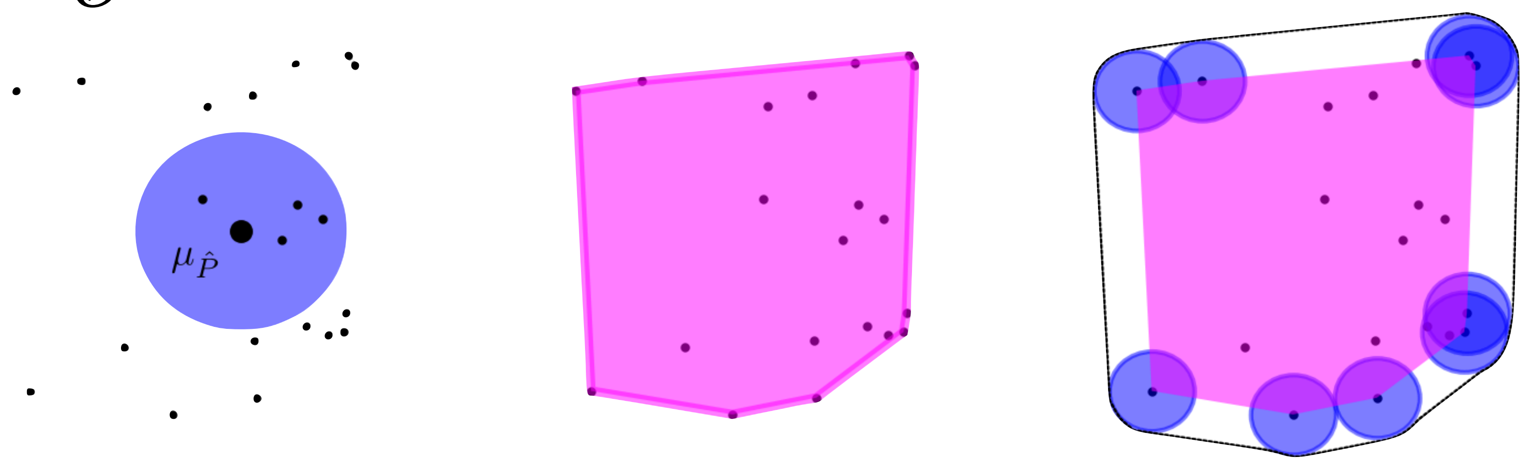




- Extension to more general ambiguity geometry

$$\min_{\theta, f\in\mathcal{H}} \delta^*_{\mu(\mathcal{M})}(f) \quad \text{subject to } l(\theta,\cdot\,) \le f.$$

$\delta^*_{\mathcal{C}}$ denotes the support function of the set $\mathcal{C}$



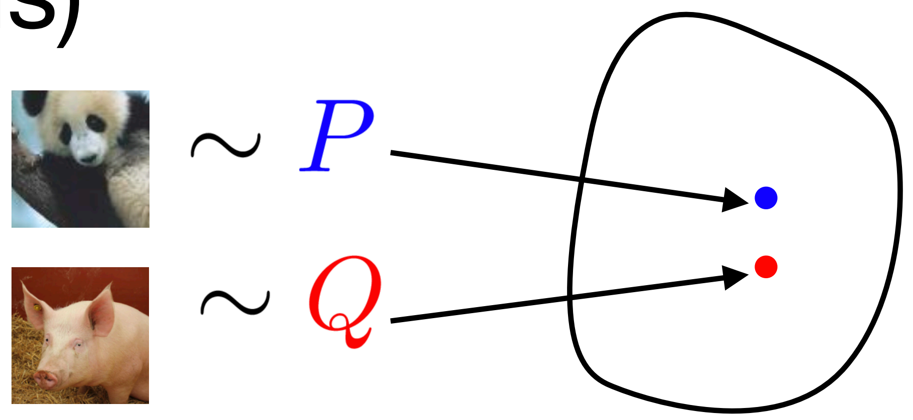
Many alg. as special cases, e.g., SVM, multi-kernel…

- Decision variable f can be interpreted as the test function in the kernel two-sample test

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

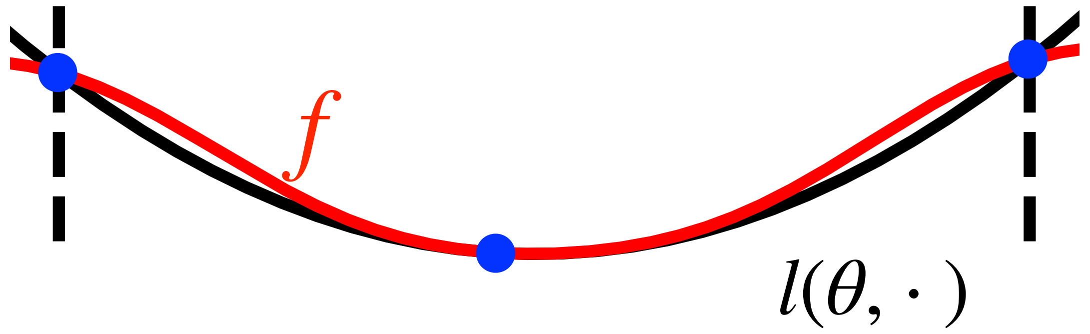


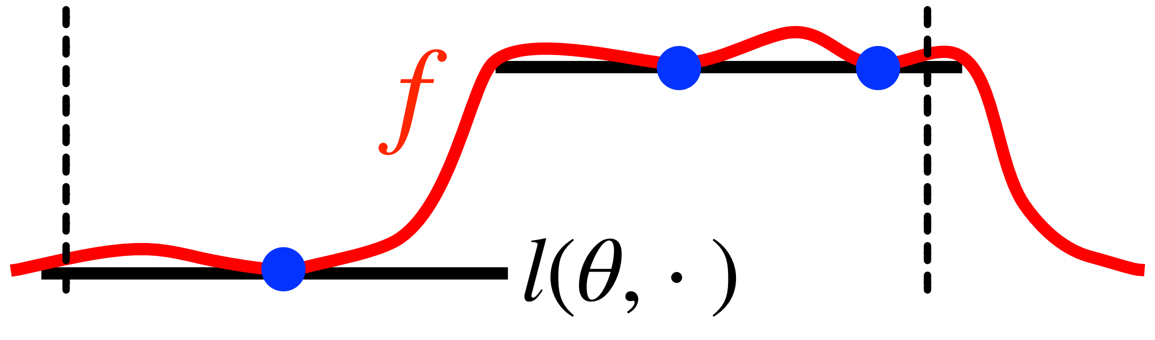
Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

Geometric intuition: using kernel approximations as robust surrogate losses

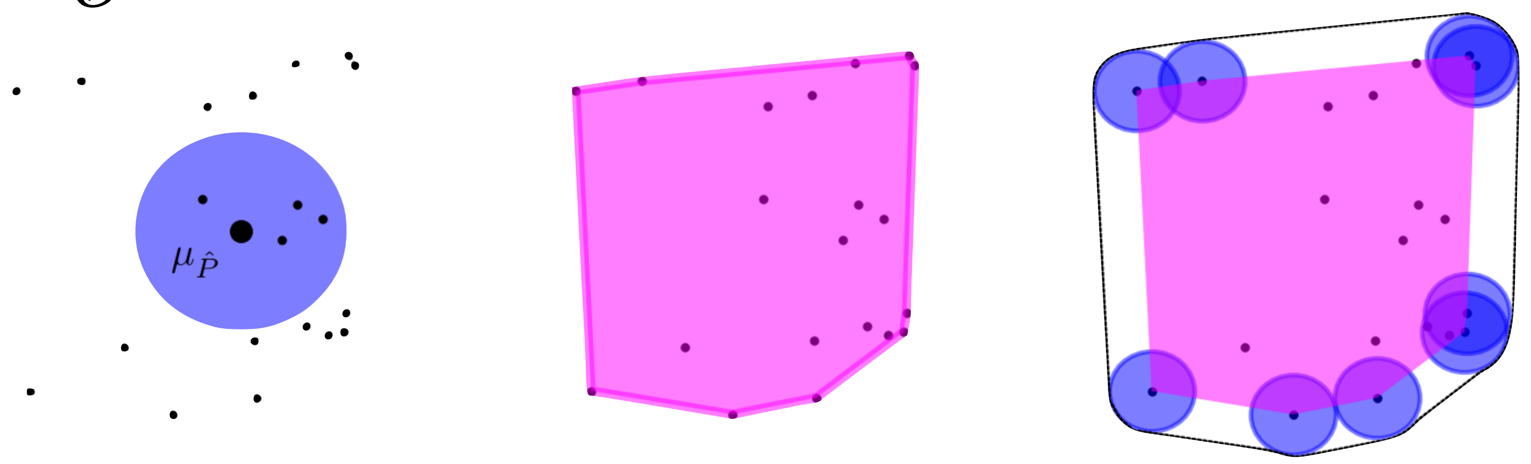




- Extension to more general ambiguity geometry

$$\min_{\theta, f\in\mathcal{H}} \delta^*_{\mu(\mathcal{M})}(f) \quad \text{subject to } l(\theta,\cdot\,) \le f.$$

$\delta^*_{\mathscr{C}}$ denotes the support function of the set $\mathscr{C}$



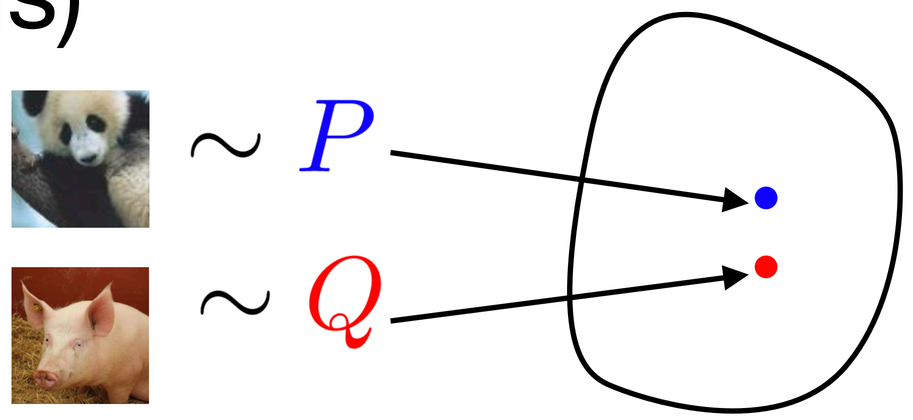
Many alg. as special cases, e.g., SVM, multi-kernel…

- Decision variable f can be interpreted as the test function in the kernel two-sample test
- Comparison with Wasserstein DRO:

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

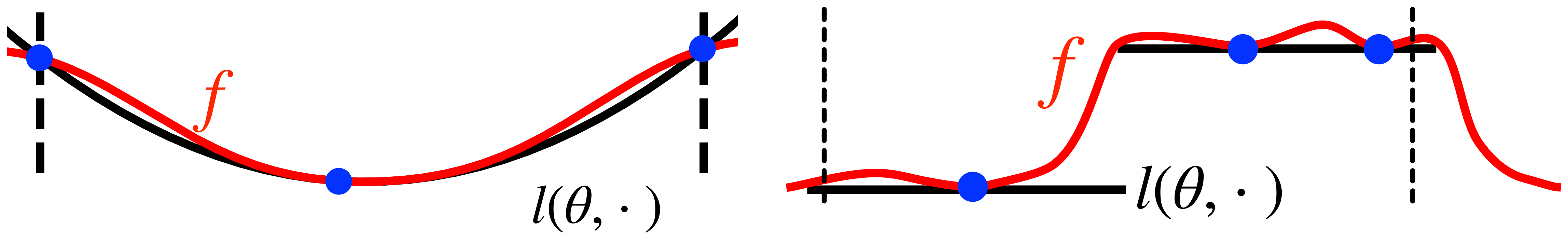


Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

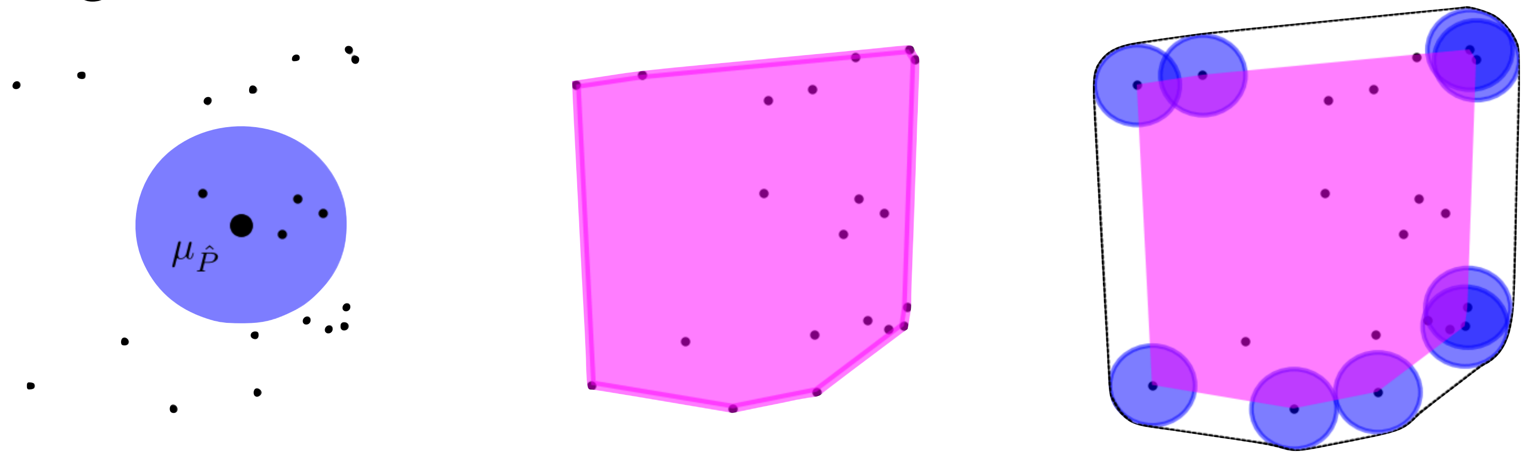
Geometric intuition: using kernel approximations as robust surrogate losses



- Extension to more general ambiguity geometry

$$\min_{\theta, f\in\mathcal{H}} \delta^*_{\mu(\mathcal{M})}(f) \quad \text{subject to } l(\theta,\cdot) \le f.$$

$\delta^*_{\mathcal{C}}$ denotes the support function of the set $\mathcal{C}$



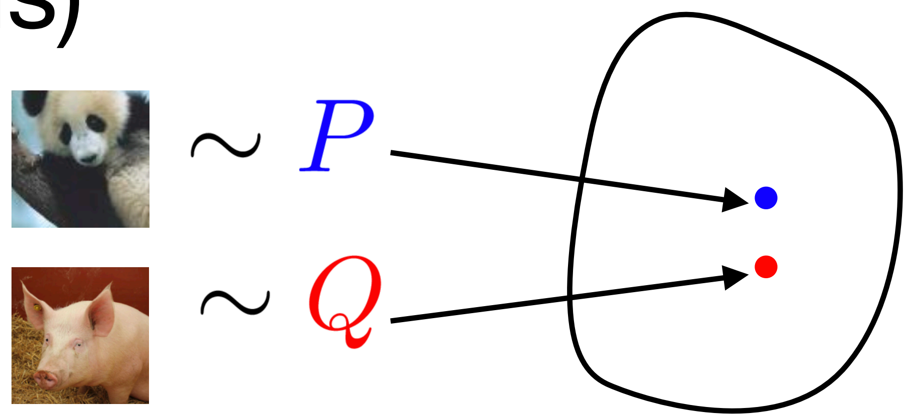
Many alg. as special cases, e.g., SVM, multi-kernel…

- Decision variable f can be interpreted as the test function in the kernel two-sample test
- Comparison with Wasserstein DRO:
  - MMD enjoys closed-form estimator for fast computation and favorable convergence rate

Kernel distributionally robust optimization

Primal DRO (not solvable as it is)

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

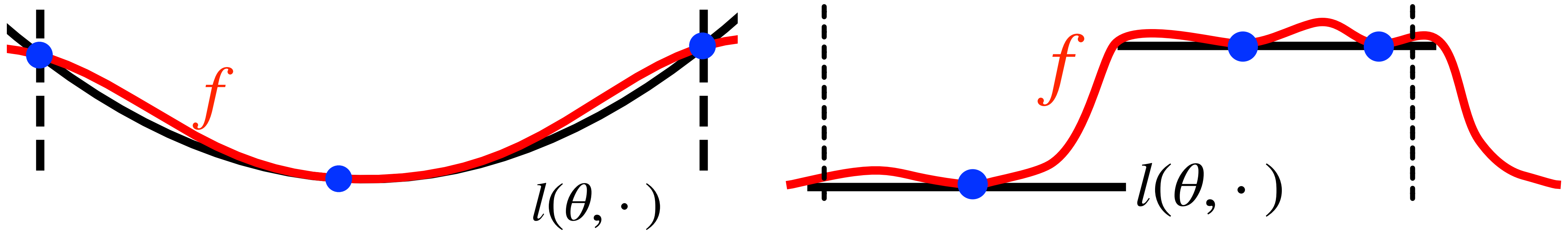


Kernel DRO Theorem (simplified). [Z. et al. AISTATS 2021] *DRO problem is equivalent to the a dual kernel learning problem, i.e., (DRO)=(K).*

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

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

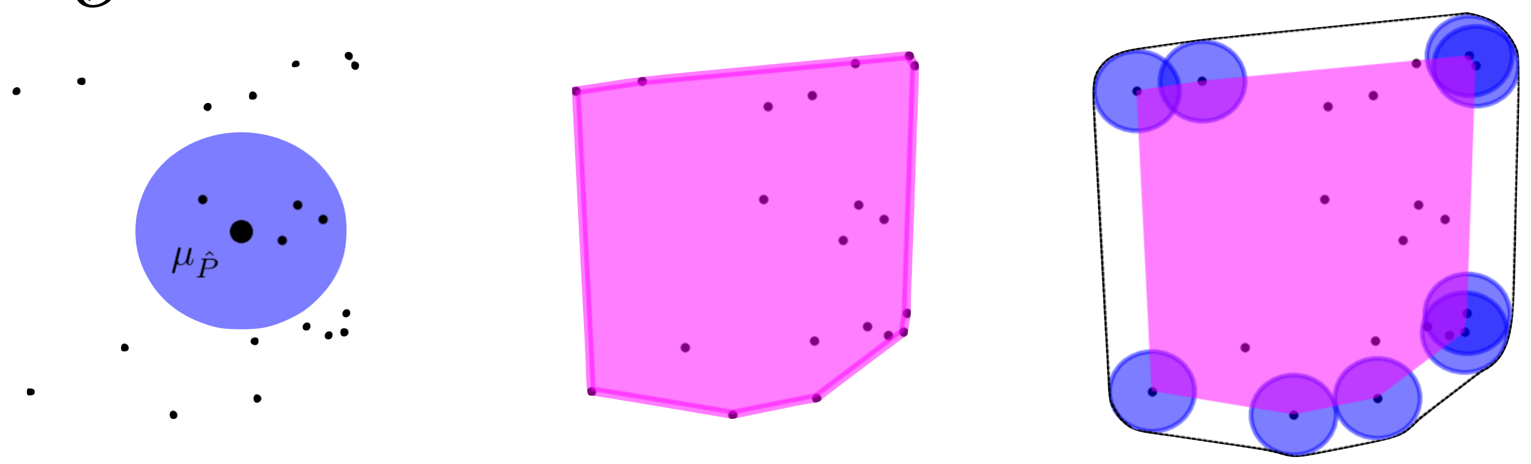
Geometric intuition: using kernel approximations as robust surrogate losses



- Extension to more general ambiguity geometry

$$\min_{\theta, f\in\mathcal{H}} \delta^*_{\mu(\mathcal{M})}(f) \quad \text{subject to } l(\theta,\cdot) \le f.$$

$\delta^*_{\mathscr{C}}$ denotes the support function of the set $\mathscr{C}$



Many alg. as special cases, e.g., SVM, multi-kernel…

- Decision variable f can be interpreted as the test function in the kernel two-sample test
- Comparison with Wasserstein DRO:
  - MMD enjoys closed-form estimator for fast computation and favorable convergence rate
  - For general ML loss $l(\theta,\cdot)$ with nonlinear models, there exists no exact reformulation of Wasserstein DRO. Kernel DRO can be applied in such cases thanks to the universality of RKHSs.

Adversarially Robust Kernel Smoothing (ARKS)

Adversarially Robust Kernel Smoothing (ARKS)

Example. Certified adversarially robust learning
(Classify the presence of glasses)

Adversarially Robust Kernel Smoothing (ARKS)

Example. Certified adversarially robust learning
(Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

Adversarially Robust Kernel Smoothing (ARKS)

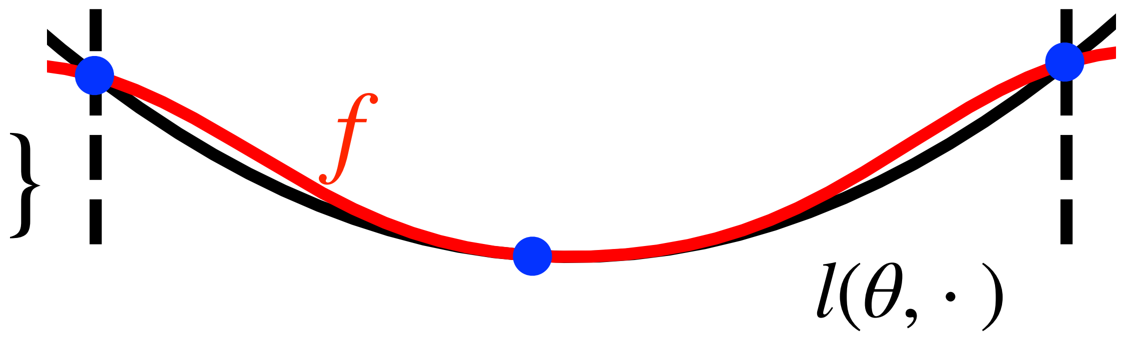
Example. Certified adversarially robust learning (Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

We construct a solution:

$f(x) = \sup_{z} \{ l(\theta, z) k(z, x) \}$



•kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$ (OT in log-scale)

c: transport cost in OT, $\sigma > 0$: bandwidth

Adversarially Robust Kernel Smoothing (ARKS)

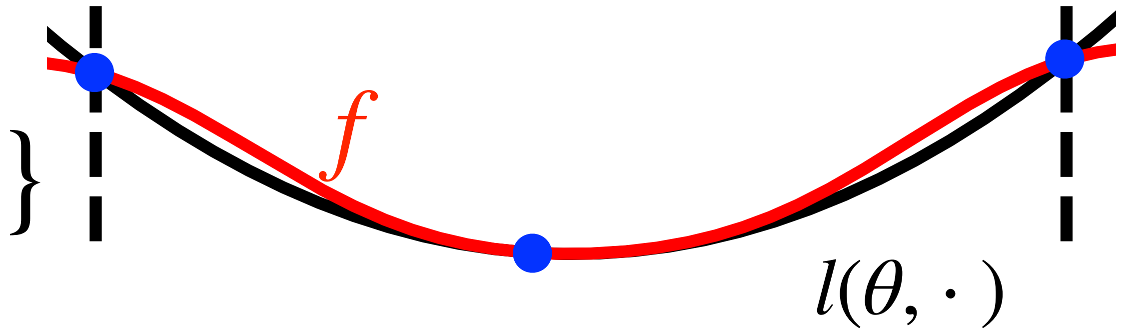
Example. Certified adversarially robust learning (Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

We construct a solution:

$$f(x) = \sup_{z} \{ l(\theta, z) k(z, x) \}$$



•kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$ (OT in log-scale)

c: transport cost in OT, $\sigma > 0$: bandwidth

✓ infinite constraint satisfied: $l(\theta, x) \leq f(x), \forall x$

Adversarially Robust Kernel Smoothing (ARKS)

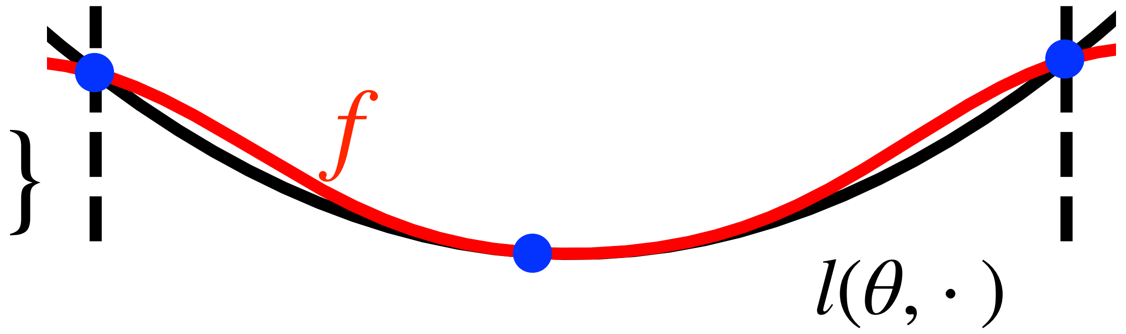
Example. Certified adversarially robust learning
(Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

We construct a solution:

$$f(x) = \sup_{z} \{ l(\theta, z) k(z, x) \}$$



- kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$ (OT in log-scale)

  c: transport cost in OT, $\sigma > 0$: bandwidth

✓ infinite constraint satisfied: $l(\theta, x) \leq f(x), \forall x$

✓ applies to loss with practical models, e.g., DNN

Adversarially Robust Kernel Smoothing (ARKS)

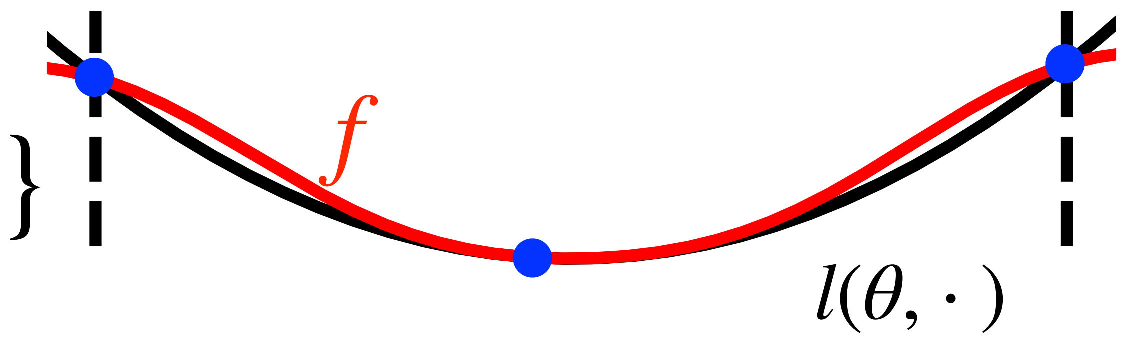
Example. Certified adversarially robust learning (Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

We construct a solution:

$f(x) = \sup_{z} \{ l(\theta, z) k(z, x) \}$



•kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$ (OT in log-scale)

c: transport cost in OT, $\sigma > 0$: bandwidth

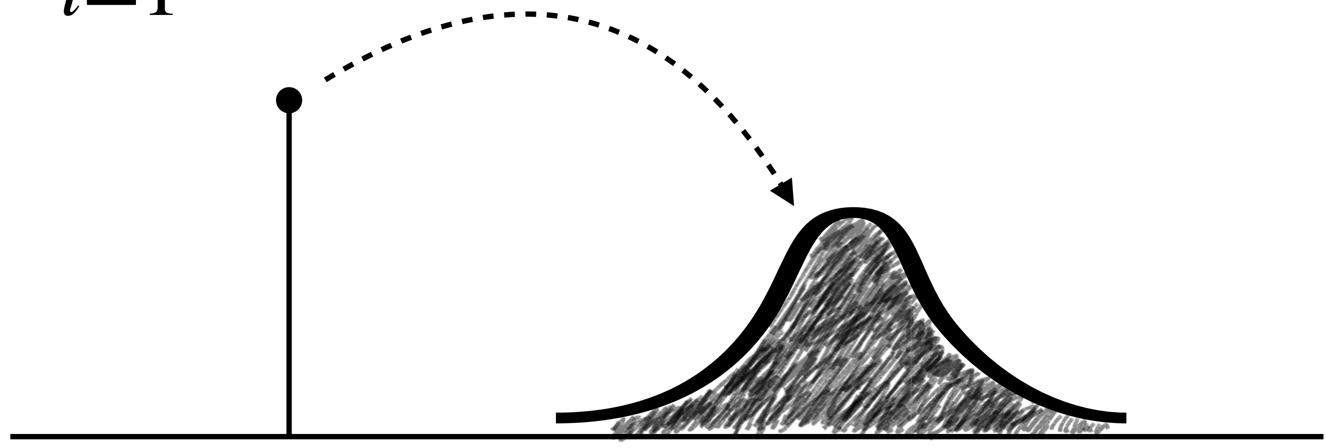
✓ infinite constraint satisfied: $l(\theta, x) \leq f(x), \forall x$

✓ applies to loss with practical models, e.g., DNN

Distributionally robust learning with ARKS

(ARKS) $\quad \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_{z} \{ l(\theta, z) k(z, \xi_i) \}$

Intuition: modeling adversarial perturbation using density $k(z, \xi_i)$.



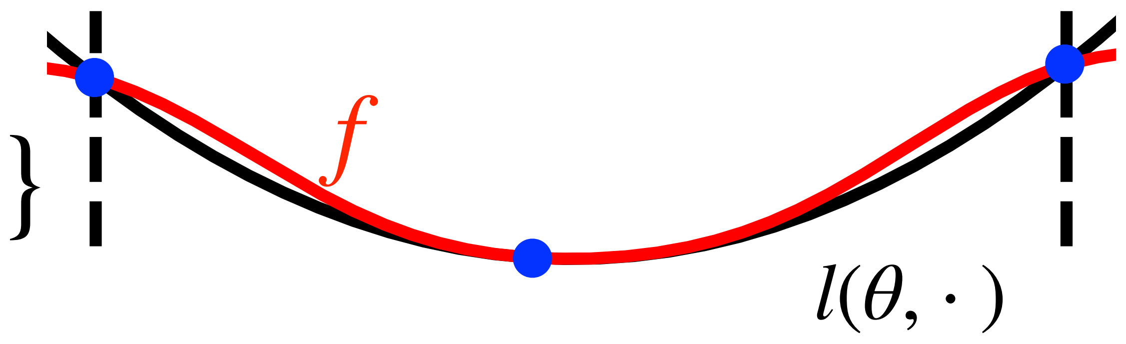
Adversarially Robust Kernel Smoothing (ARKS)

Example. Certified adversarially robust learning
(Classify the presence of glasses)

Recall: **Kernel DRO Theorem**:

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

We construct a solution:
$f(x) = \sup_z \{ l(\theta, z) k(z, x) \}$



•kernel choice: $k(x, x') := e^{-c(x,x')/\sigma}$ (OT in log-scale)
c: transport cost in OT, $\sigma > 0$: bandwidth

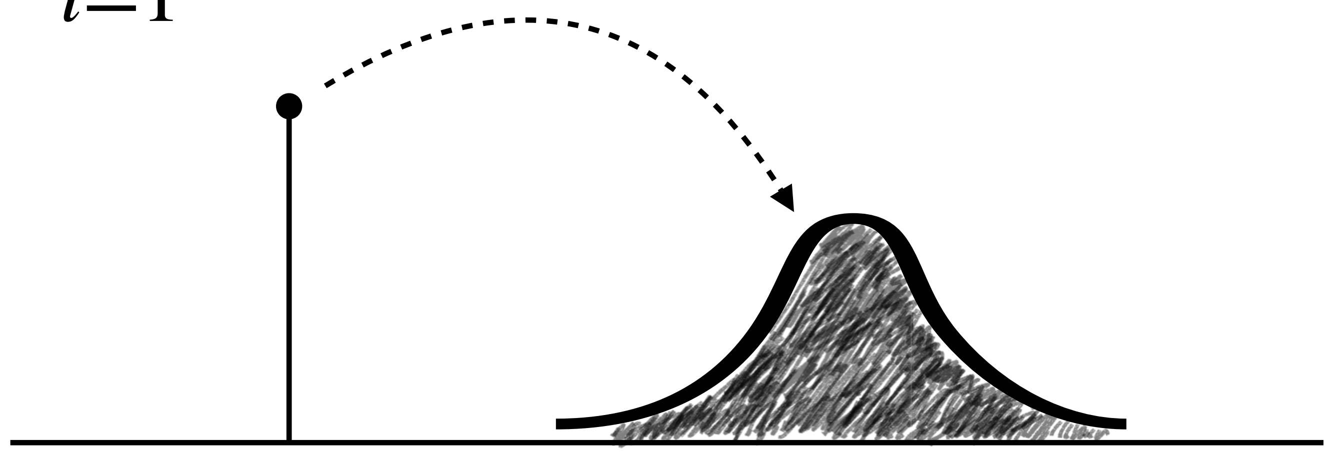
✓ infinite constraint satisfied: $l(\theta, x) \leq f(x), \forall x$

✓ applies to loss with practical models, e.g., DNN

Distributionally robust learning with ARKS

$$\text{(ARKS)} \quad \min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sup_z \{ l(\theta, z) k(z, \xi_i) \}$$

Intuition: modeling adversarial perturbation using density $k(z, \xi_i)$.



Distributional robustness certificate.

$\mathcal{W}_c(\,\cdot\,, \cdot\,)$: OT metric with transport cost c

$\epsilon_N \to 0$, computable robustness certificate:

$$\sup_{\mathcal{W}_c(Q, P_0) \leq \rho} \mathbb{E}_Q \ln l(\hat{\theta}, \xi) \leq \ln \underbrace{\left\{ \frac{1}{N} \sum_{i=1}^{N} \sup_z \{ l(\hat{\theta}, z) k(z, \xi_i) \} \right\}}_{\text{ARKS objective}} + \frac{\rho}{\sigma} + \epsilon_N$$

Conclusion

Conclusion

- We proposed a distributionally robust learning algorithm by constructing solutions using the duality theorem of Kernel DRO.

Conclusion

- We proposed a distributionally robust learning algorithm by constructing solutions using the duality theorem of Kernel DRO.
- Exploiting the connection between OT and kernel methods, we provide a distributional robustness certificate.

Conclusion

- We proposed a distributionally robust learning algorithm by constructing solutions using the duality theorem of Kernel DRO.
- Exploiting the connection between OT and kernel methods, we provide a <u>distributional robustness certificate</u>.
- In contrast with many DRO algorithms, our distributionally robust learning algorithm applies to large-scale learning with DNNs. This is enabled by the kernel DRO theorem that allows us to use <u>regularized kernel approximation for general functions outside RKHSs</u>.

Conclusion

- We proposed a distributionally robust learning algorithm by constructing solutions using the duality theorem of Kernel DRO.
- Exploiting the connection between OT and kernel methods, we provide a distributional robustness certificate.
- In contrast with many DRO algorithms, our distributionally robust learning algorithm applies to large-scale learning with DNNs. This is enabled by the kernel DRO theorem that allows us to use regularized kernel approximation for general functions outside RKHSs.

Interesting future directions

- Design specific kernels for robustness beyond norm-ball perturbation
- Physics, information geometry, and general dynamic OT
- Causal inference via distributional robustness

References

- **Zhu**, J.-J., Jitkrittum, W., Diehl, M. & Schölkopf, B. Kernel Distributionally Robust Optimization. AISTATS 2021
- **Zhu**, J.-J., Kouridi, C., Nemmour, Y. & Schölkopf, B. Adversarially Robust Kernel Smoothing. AISTATS 2022

Code

- KDRO: https://github.com/jj-zhu/kdro
- ARKS: https://github.com/christinakouridi/arks

Website: jj-zhu.github.io