# Prediction Games:

## From Maximum Likelihood Estimation to Active Learning, Fair Machine Learning, and Structured Prediction

Brian Ziebart

THE UNIVERSITY OF ILLINOIS AT CHICAGO

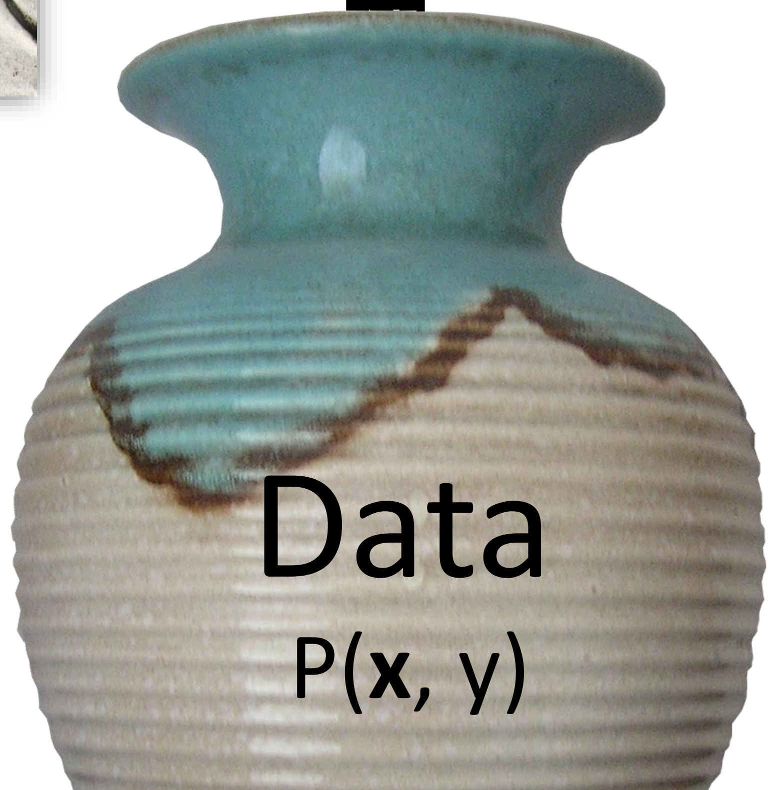
# Training

**Viagra**
**Cialis**
**cheap**
spam


dog

**x** y

**M samples**




**Sample dist.**
$\tilde{P}(\mathbf{x}, y)$

# Training

**Predictor** f: $X \rightarrow Y$

Viagra
Cialis
cheap

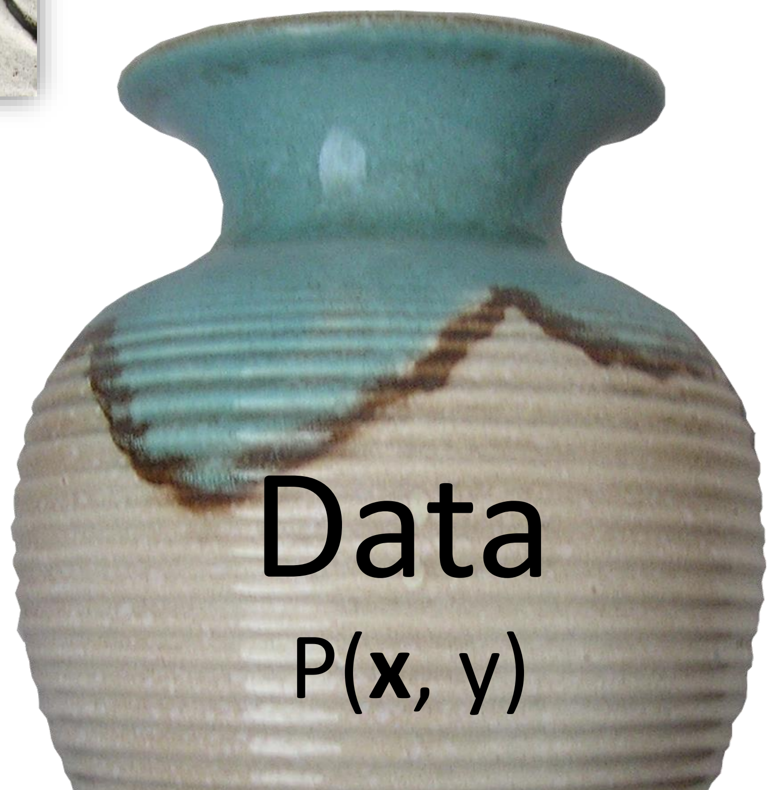
spam

dog

**x** y

**Sample dist.**

$\tilde{P}(\mathbf{x}, y)$



# Predictor f: $X \rightarrow Y$ **Testing**

$y$

| $\hat{y}$ | Dog | Cat | Car |
|---|---|---|---|
| Dog | 0 | 1 | 1 |
| Cat | 1 | 0 | 1 |
| Car | 1 | 1 | 0 |

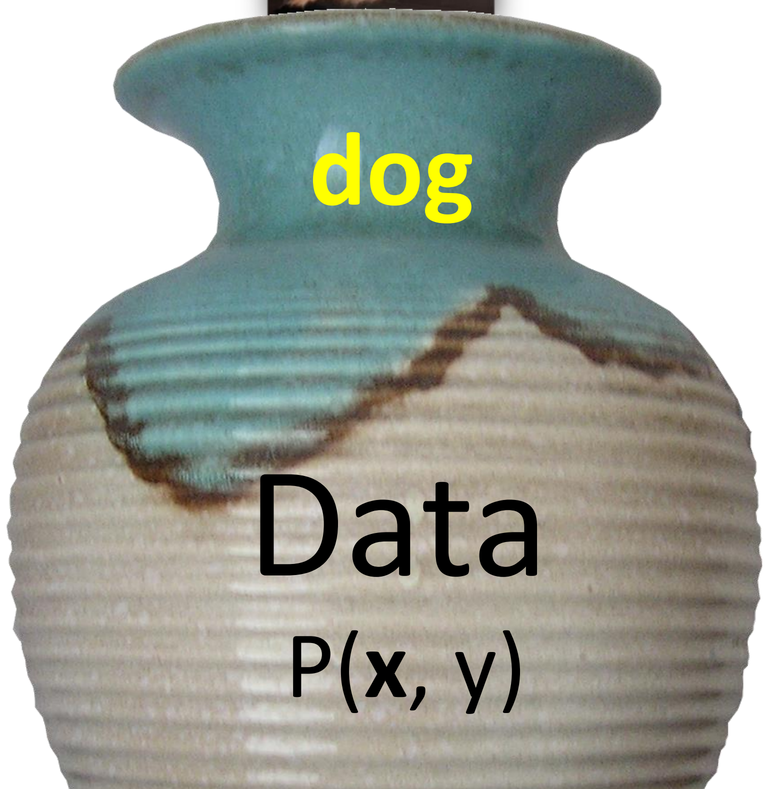
$y$

**Prediction:** $\hat{y} = f(\mathbf{x})$


**Loss:** $\mathrm{loss}(\hat{y}, y)$

**Expected Loss:**

$E_P[\mathrm{loss}(f(\mathbf{X}), Y)]$



# Predictor f: $\boldsymbol{X} \rightarrow Y$

## Standard Idea:

**Empirical Risk Minimization**

1. Restrict predictor to some set: f ∈ F (or regularize)
2. Minimize (surrogate) loss over predictor set

**Approximate loss**

**Exact training data**

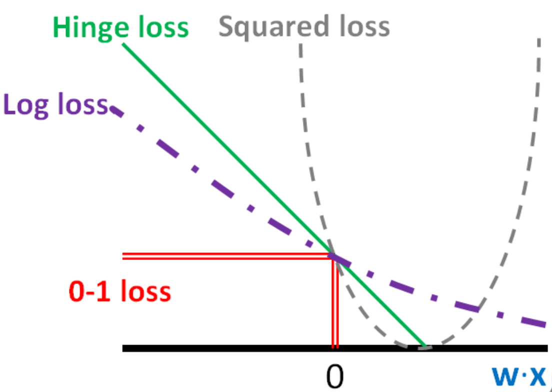
## Dist. Robust Idea:

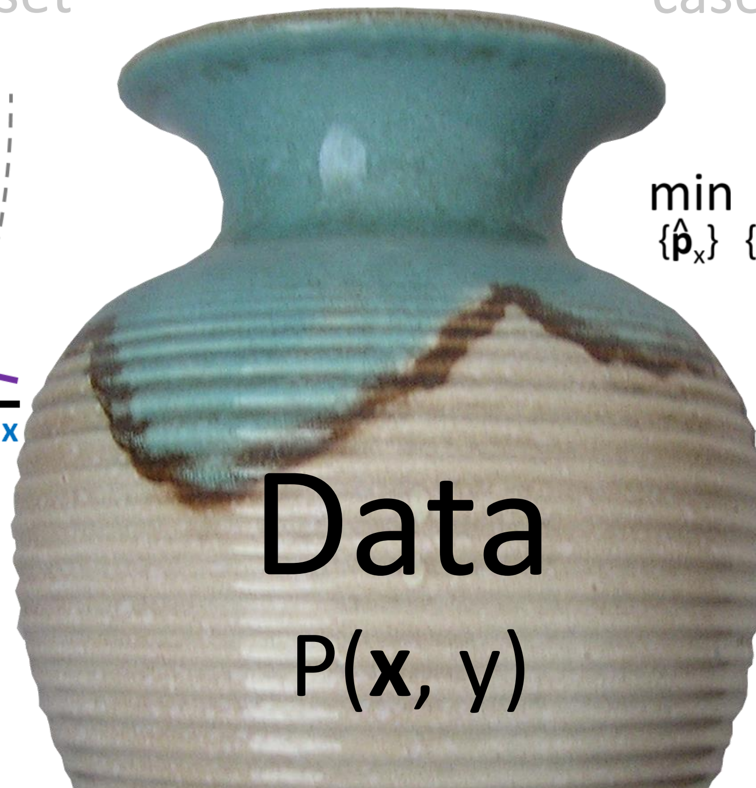
**Adversarial Risk Minimization**

1. Approximate true label dist. using sample dist. statistics
2. Minimize loss over worst-case approximation

**Exact loss**

**Approx. train labels**




$$\min_{\{\hat{\mathbf{p}}_x\}} \max_{\{\check{\mathbf{p}}_x\}\in\Xi} \sum_i \hat{\mathbf{p}}_{x_i}^{T} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \check{\mathbf{p}}_{x_i}$$

# Adversarial Supervised Learning Formulation

Construct predictor robust to worst label distribution:

$$\min_{\widehat{P}(\widehat{y}|\mathbf{x})\in\Delta} \max_{\check{P}(\check{y}|\mathbf{x})\in\Delta\cap\Xi} -\sum_{\mathbf{x},y} \tilde{P}(\mathbf{x})\check{P}(y|\mathbf{x}) \log \widehat{P}(y|\mathbf{x})$$

**Logarithmic loss** measures surprise from labels

**Predictor** P̂ minimizes loss (in probability simplex Δ)

**Adversary** P̌ maximizes loss, but must be similar to available data (set Ξ, e.g., $\mathbb{E}_{\substack{\mathbf{X}\sim\tilde{P} \\ Y|\mathbf{X}\sim\check{P}}}[\phi(\mathbf{X},Y)] = \mathbb{E}_{\mathbf{X},Y\sim\tilde{P}}[\phi(\mathbf{X},Y)]$)

Reduces to **maximizing entropy** (P̂ = P̌) (Topsøe 1979)

**Robust Bayesian Games** (Grünwald & Dawid 2004)

**DRO** with **expectations constraints** (Wieseman et al. 2014)

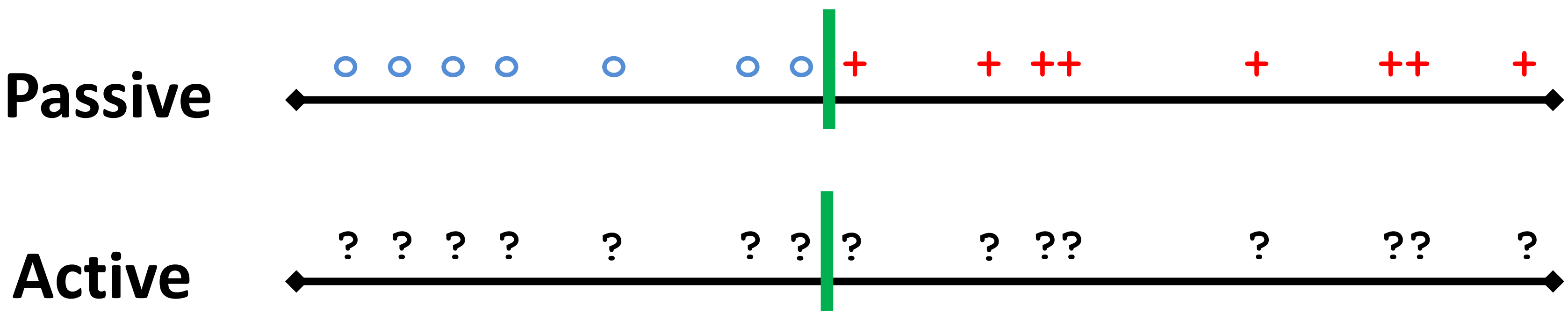
→ **Standard logistic regression and MLE**: $\hat{P}(y|\mathbf{x}) \propto e^{\theta\cdot\phi(\mathbf{x},y)}$

# Part 1: Covariate Shift & Active Learning

Joint work with Anqi Liu (NeurIPS 2014),
Anqi Liu & Lev Reyzin (AAAI 2015)

# Active Learning

Learn linear thresholds: x > c → + (o otherwise)




**4** vs. **15** labels

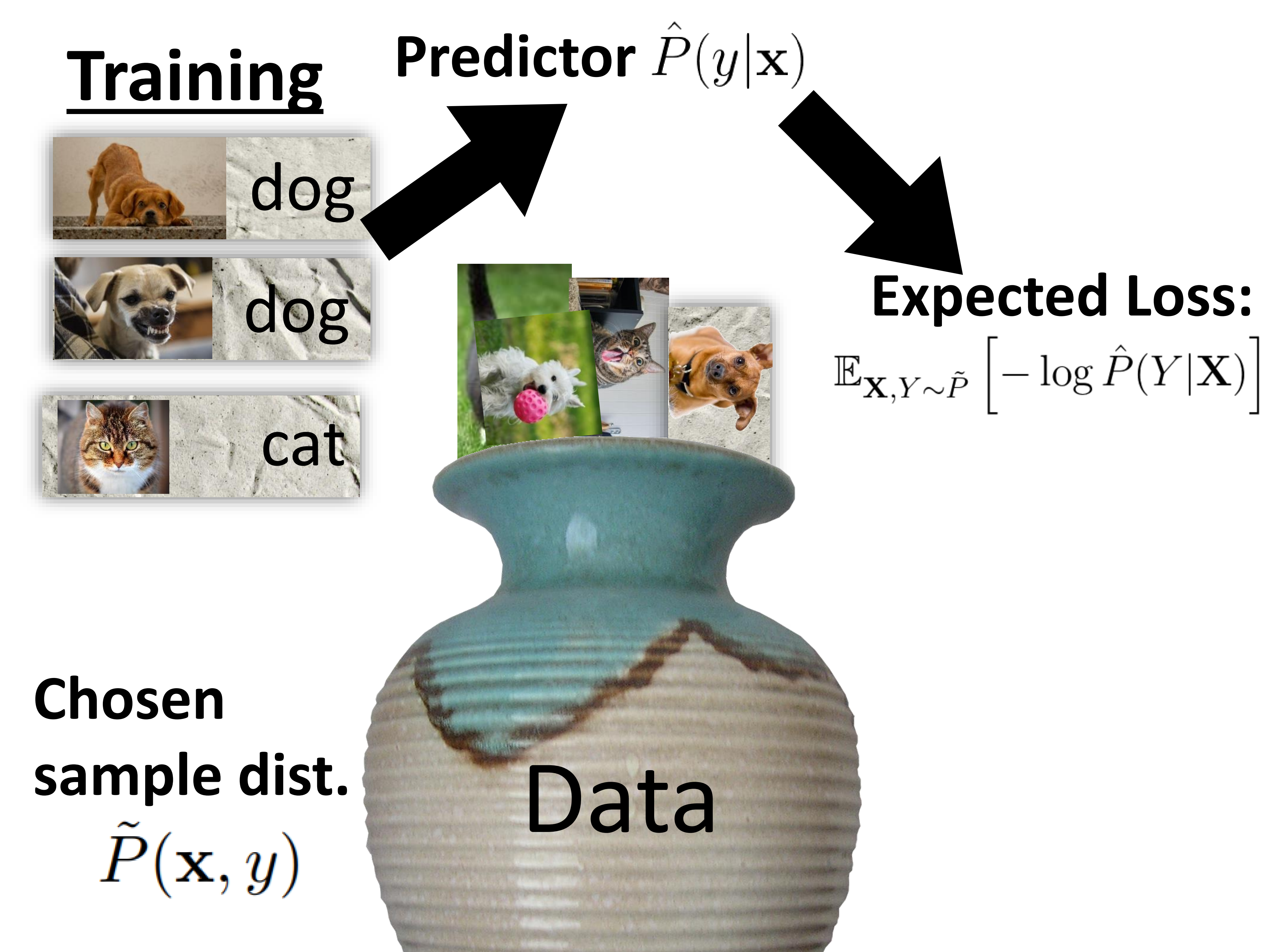
**O(log n)** vs. **O(n)**

# Training



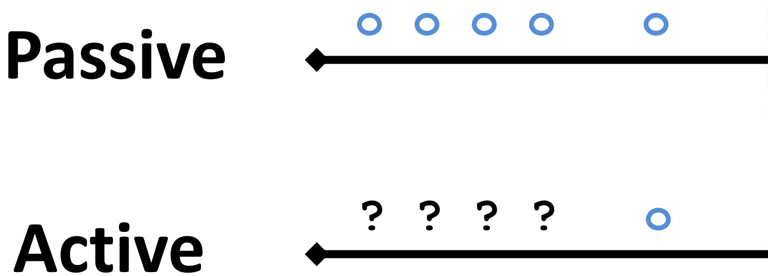
**Chosen sample dist.**

$$\tilde{P}(\mathbf{x}, y)$$


Training
Predictor $\hat{P}(y|\mathbf{x})$
dog
dog
cat
Expected Loss:
$\mathbb{E}_{\mathbf{X},Y\sim\tilde{P}}\left[-\log\hat{P}(Y|\mathbf{X})\right]$
Chosen
sample dist.
$\tilde{P}(\mathbf{x},y)$
Data

# Active Learning

Learn linear thr

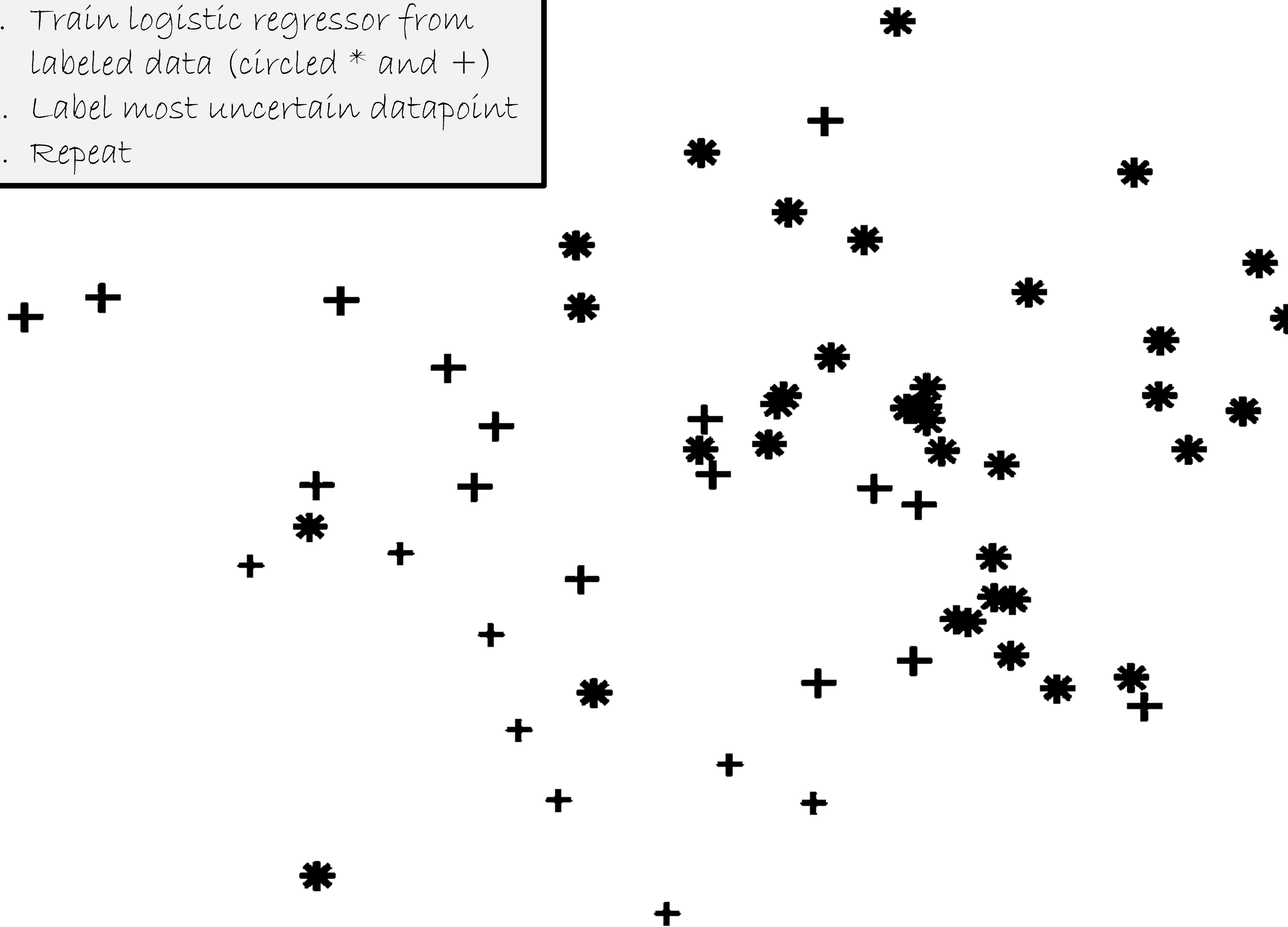



# Results frequently **worse** than passive learning!

"random sampling ... may be more advisable than taking one's chances on active learning with an inappropriate learning model"

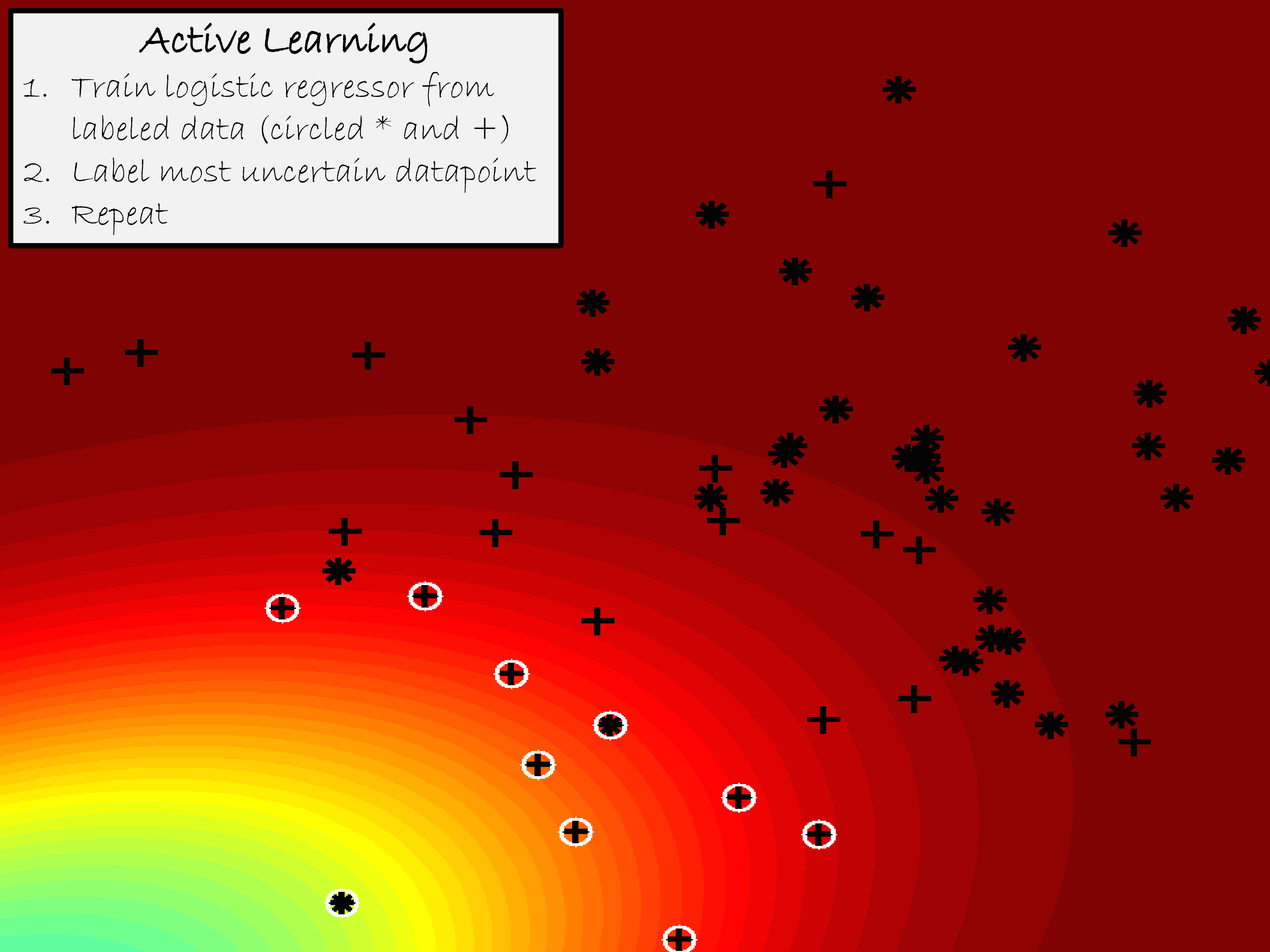
# Active Learning

1. Train logistic regressor from labeled data (circled * and +)
2. Label most uncertain datapoint
3. Repeat



## Active Learning

1. Train logistic regressor from labeled data (circled * and +)
2. Label most uncertain datapoint
3. Repeat



# What's wrong with this recipe?

**Active Learning**

1. Train logistic regressor from labeled data
2. Label most uncertain datapoint
3. Repeat

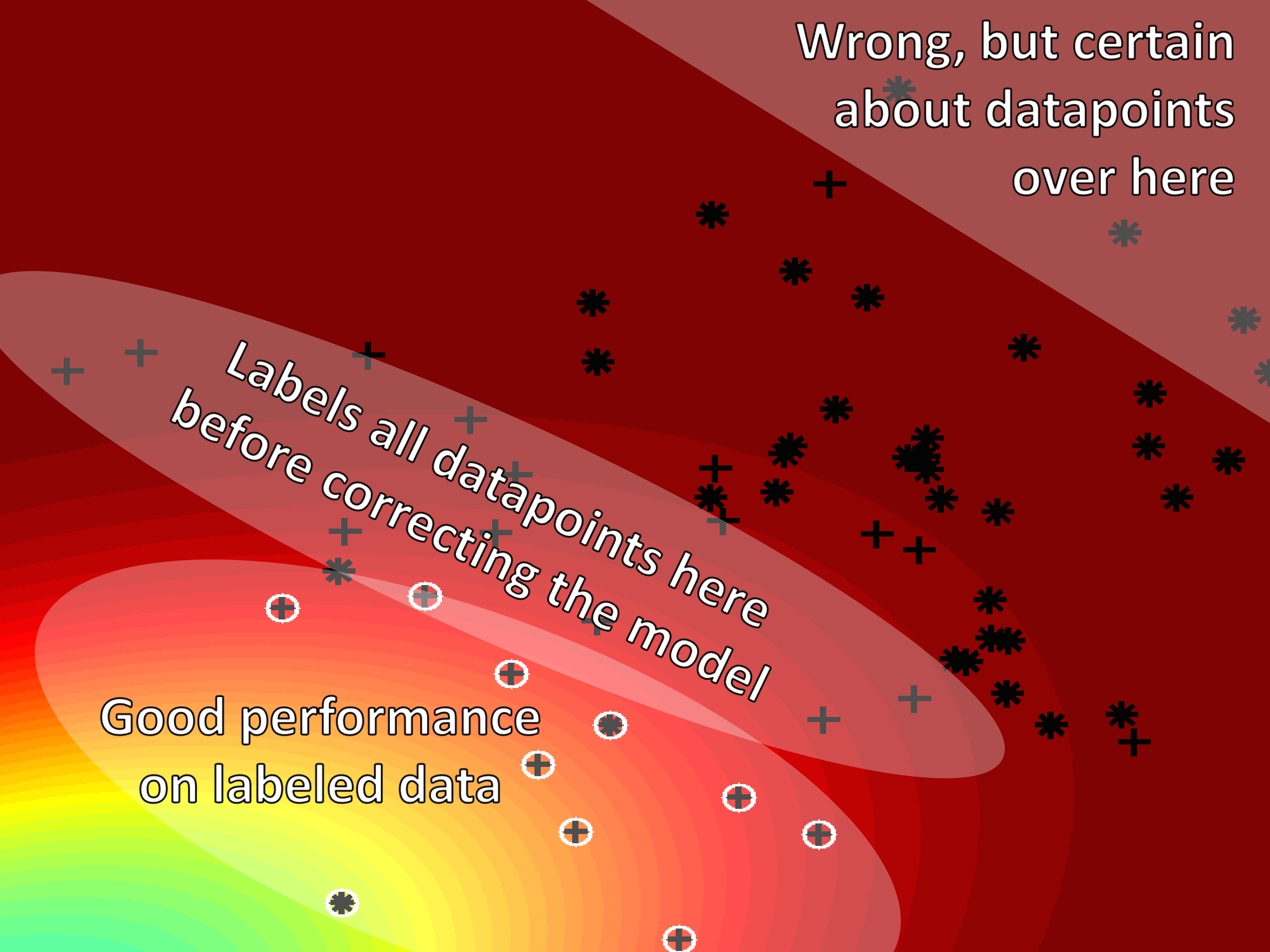
# What's wrong with this recipe?

**Active Learning**

1. Train logistic regressor from labeled data
2. Label most uncertain datapoint
3. Repeat

Assumes IID data

Produces non-IID data


Wrong, but certain
about datapoints
over here
Labels all datapoints here
before correcting the model
Good performance
on labeled data

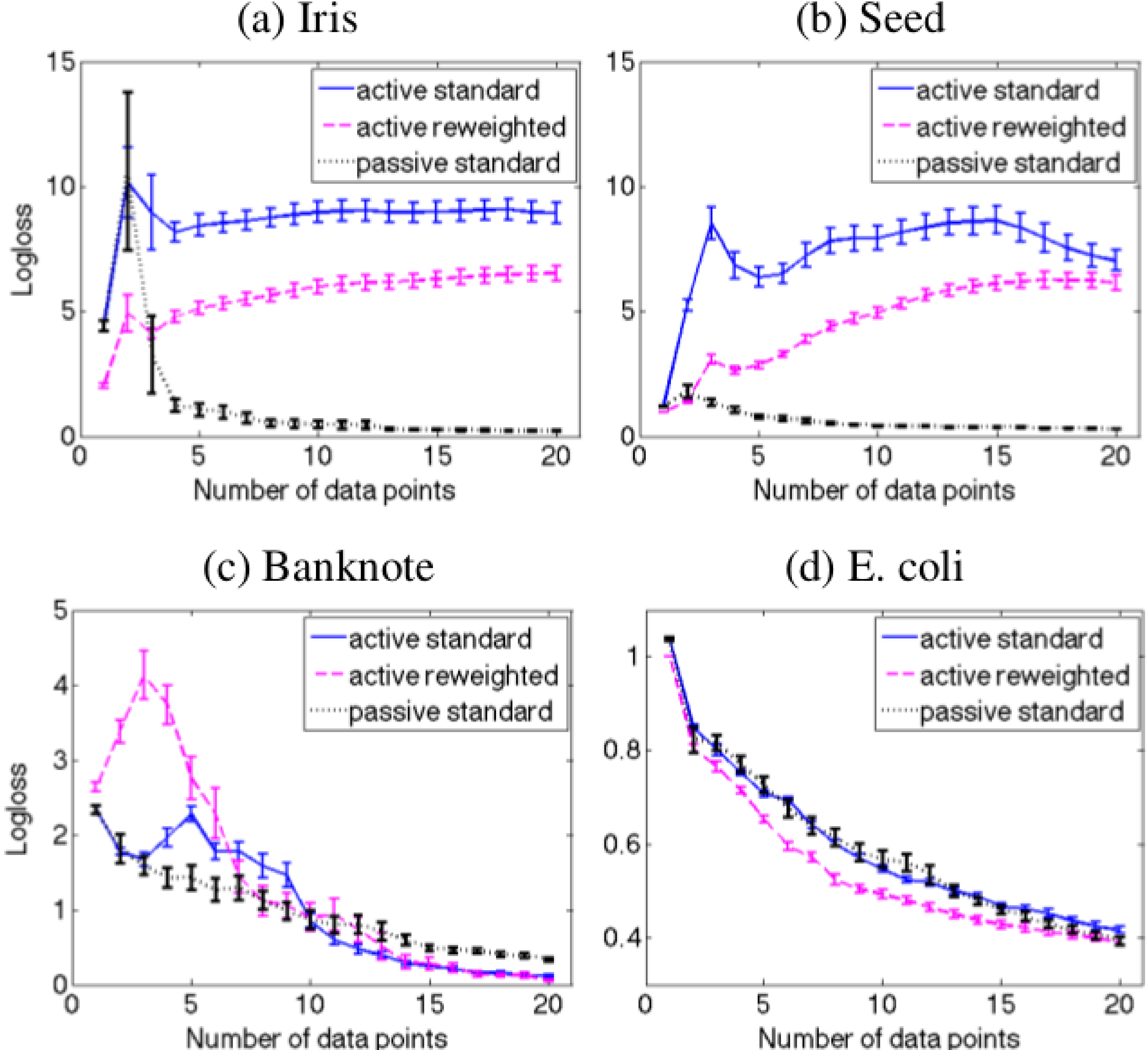
# Re-Weighted Empirical Risk Minimization

(Shimodaira 2000, Kanamori and Shimodaira 2003)

$$\max_{\theta} \mathbb{E}_{\tilde{P}_{\text{train}}(\mathbf{x},y)} \left[ \log \hat{P}_{\theta}(Y|\mathbf{X}) \right] + \lambda||\theta||$$

$$\frac{P_{\text{test}}(\mathbf{x})}{P_{\text{train}}(\mathbf{x})}$$

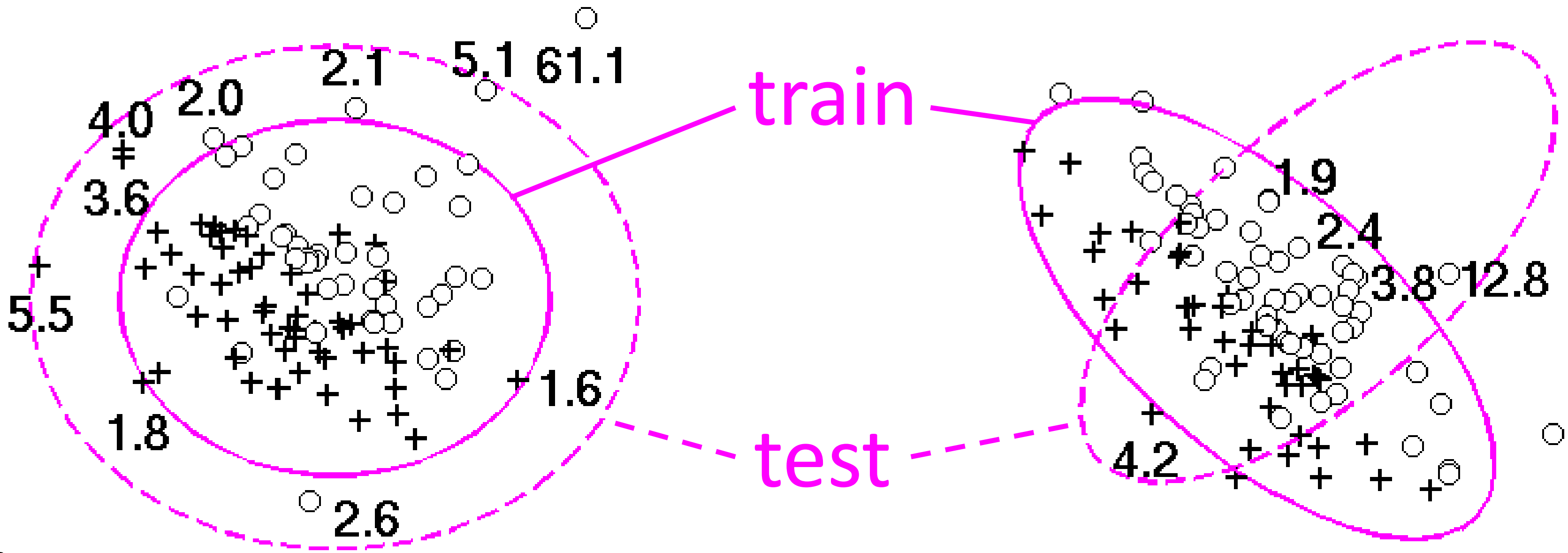
If $\frac{P_{\text{test}}(\mathbf{x})}{P_{\text{train}}(\mathbf{x})} > 0$, asymptotically unbiased estimator!



(a) Iris
(b) Seed
(c) Banknote
(d) E. coli
Logloss
Number of data points
active standard
active reweighted
passive standard
0
5
10
15
20
0.4
0.6
0.8
1

# Re-Weighted Empirical Risk Minimization

(Shimodaira 2000, Kanamori and Shimodaira 2003)




**Issues:**

- <u>High variance</u> estimates
- Slow (or <u>no</u>) convergence (Cortes et al. 2010)
- Especially bad for small sample sizes

# Adversarial Prediction for Sample Selection Bias

$$\min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{\tilde{P}(\mathbf{x})\check{P}(\check{y}|\mathbf{x})} \left[ -\log \hat{P}(\check{Y}|\mathbf{X}) \right]$$

$$\text{such that: } \mathbb{E}_{\tilde{P}(\mathbf{x})\check{P}(\check{y}|\mathbf{x})} \left[ \phi(\mathbf{X}, \check{Y}) \right] = \tilde{\phi}$$

IID: $\hat{P}_\theta(y|\mathbf{x}) \propto e^{\theta \cdot \phi(\mathbf{x}, y)}$

# Adversarial Prediction for Sample Selection Bias

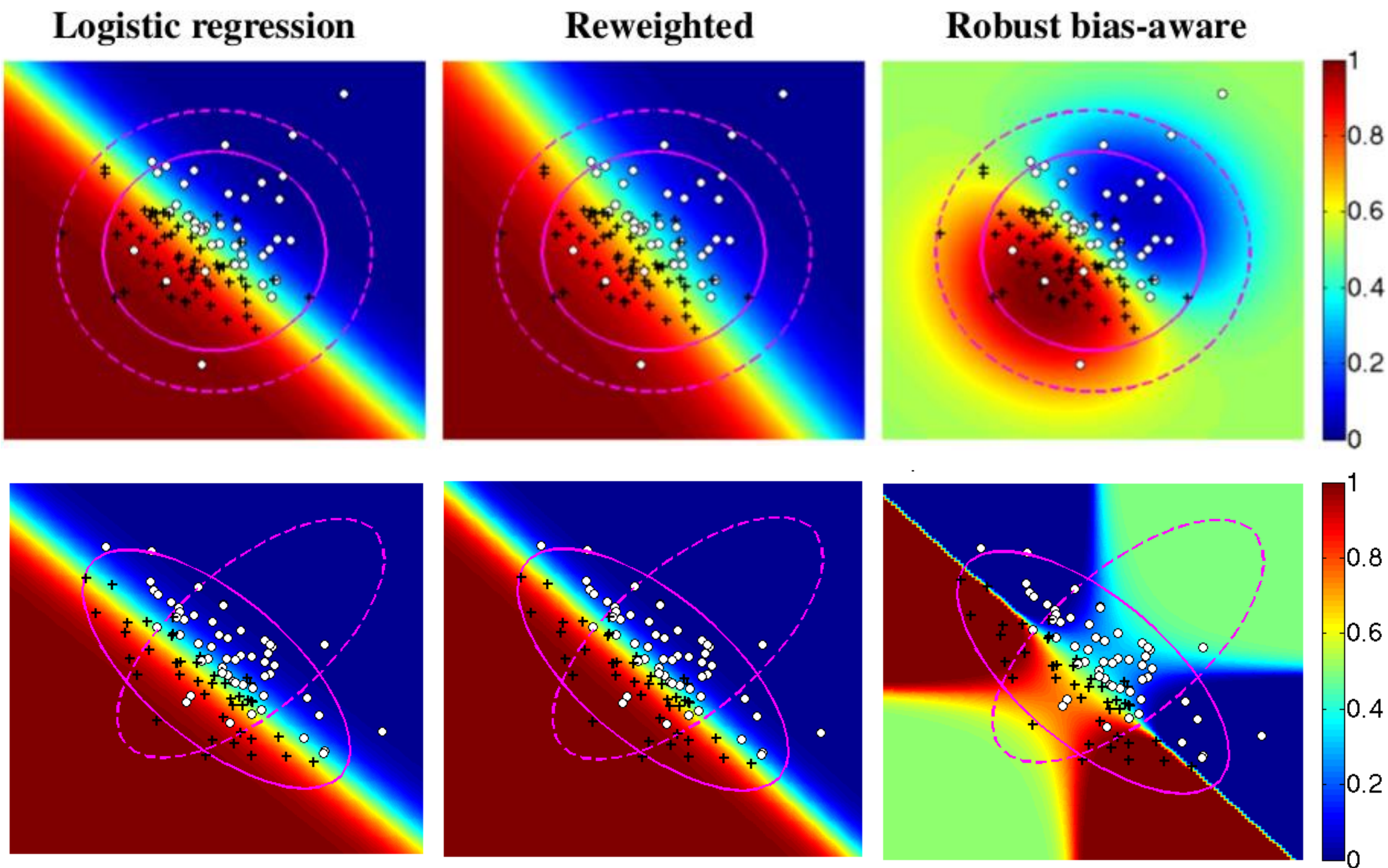
$$\min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{P_{\text{test}}(\mathbf{x})\check{P}(\check{y}|\mathbf{x})} \left[ -\log \hat{P}(\check{Y}|\mathbf{X}) \right]$$

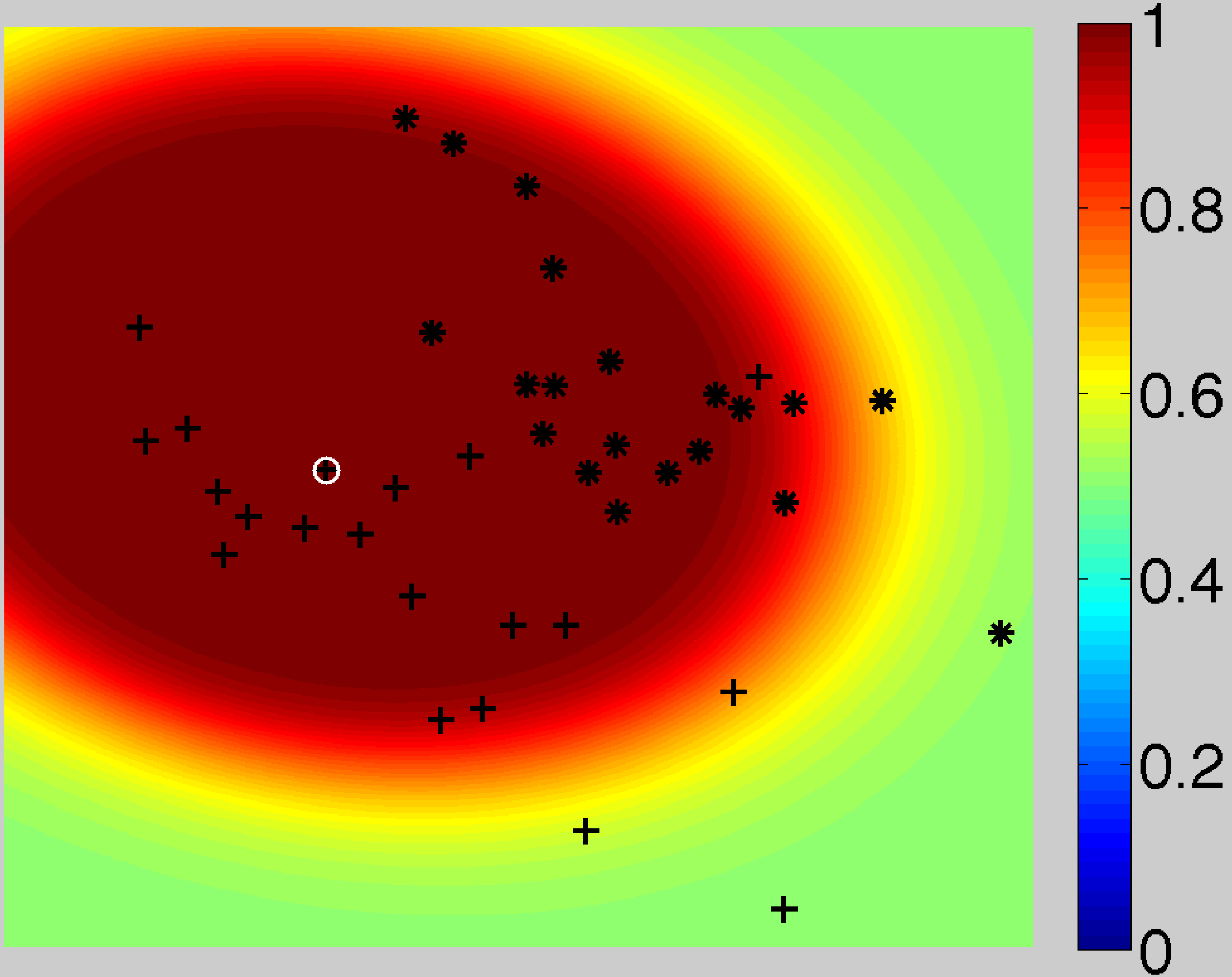
$$\text{such that: } \mathbb{E}_{P_{\text{train}}(\mathbf{x})\check{P}(\check{y}|\mathbf{x})} \left[ \phi(\mathbf{X}, \check{Y}) \right] = \tilde{\phi}$$

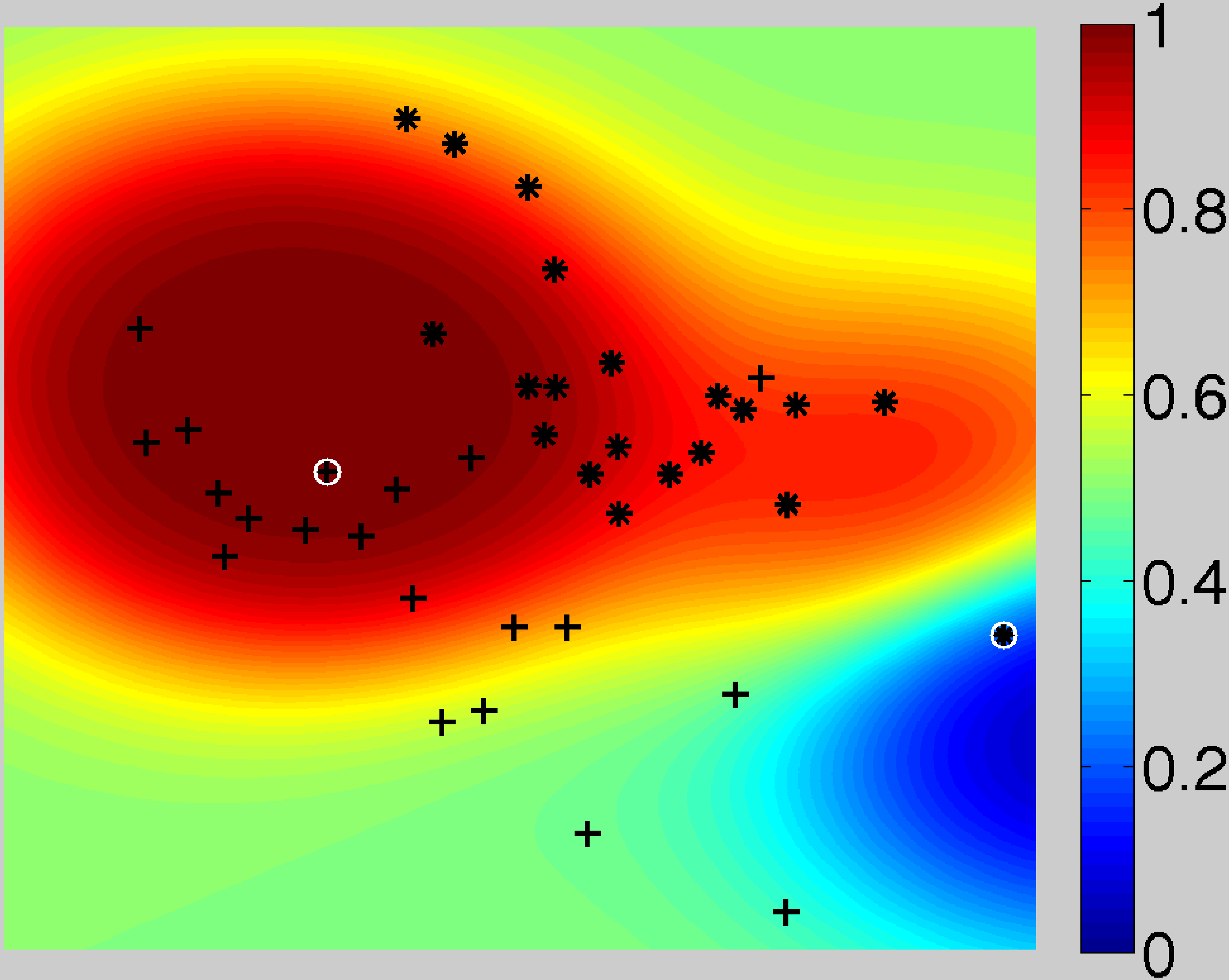
IID: $\hat{P}_\theta(y|\mathbf{x}) \propto e^{\theta \cdot \phi(\mathbf{x},y)}$

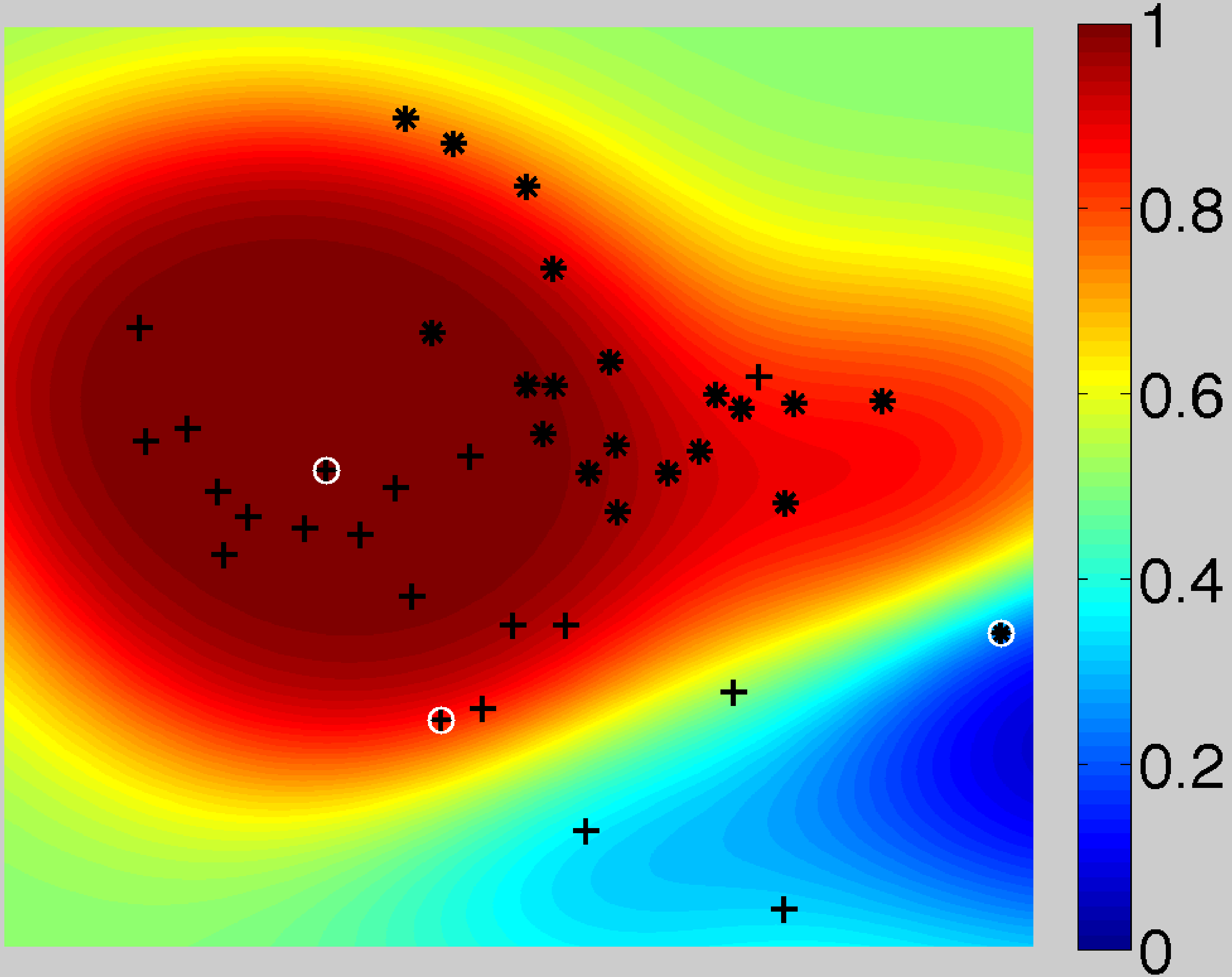
Covariate Shift: $\hat{P}_\theta(y|\mathbf{x}) \propto e^{\frac{P_{\text{train}}(\mathbf{x})}{P_{\text{test}}(\mathbf{x})}\theta \cdot \phi(\mathbf{x},y)}$

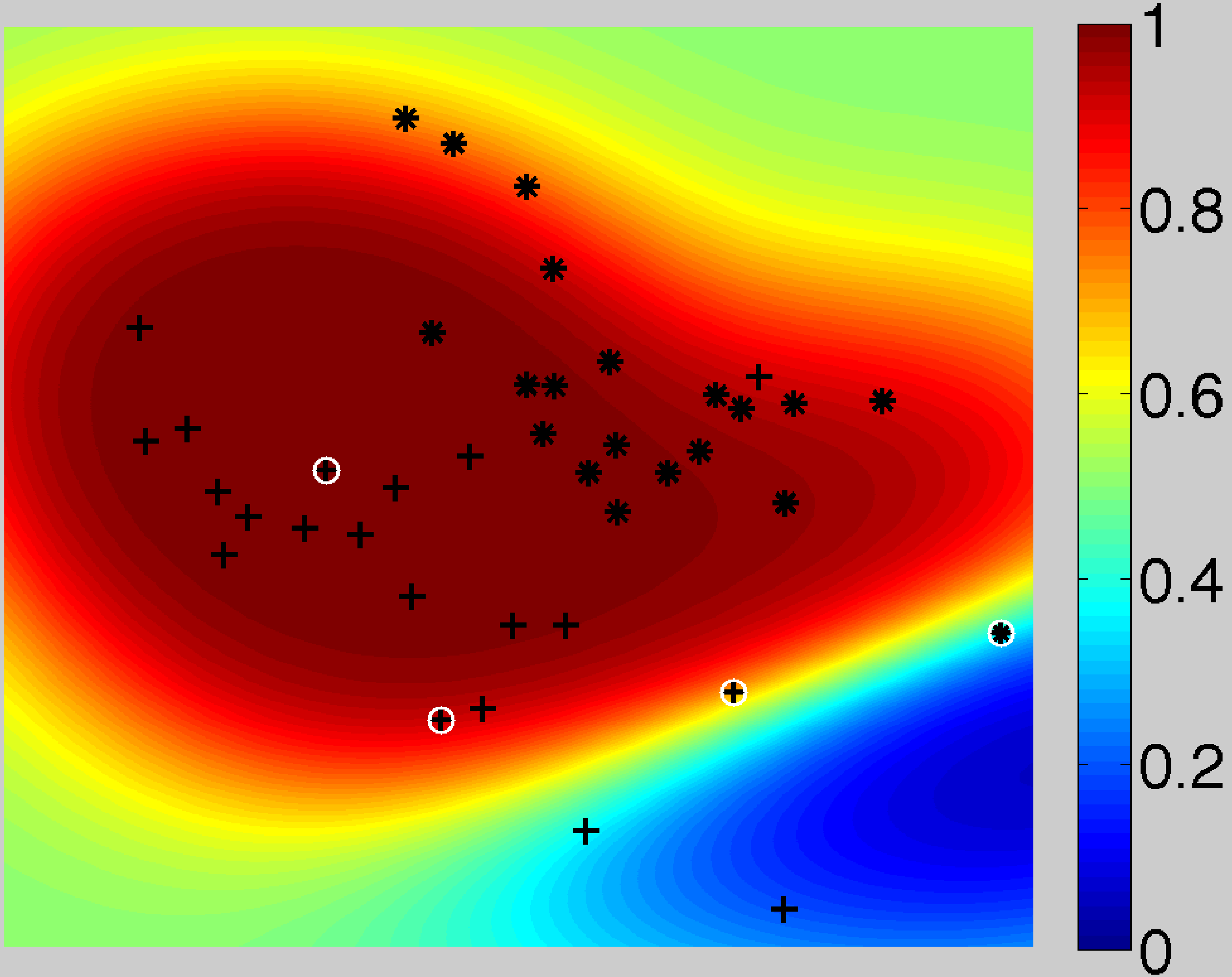
# Adversarial Prediction for Sample Selection Bias

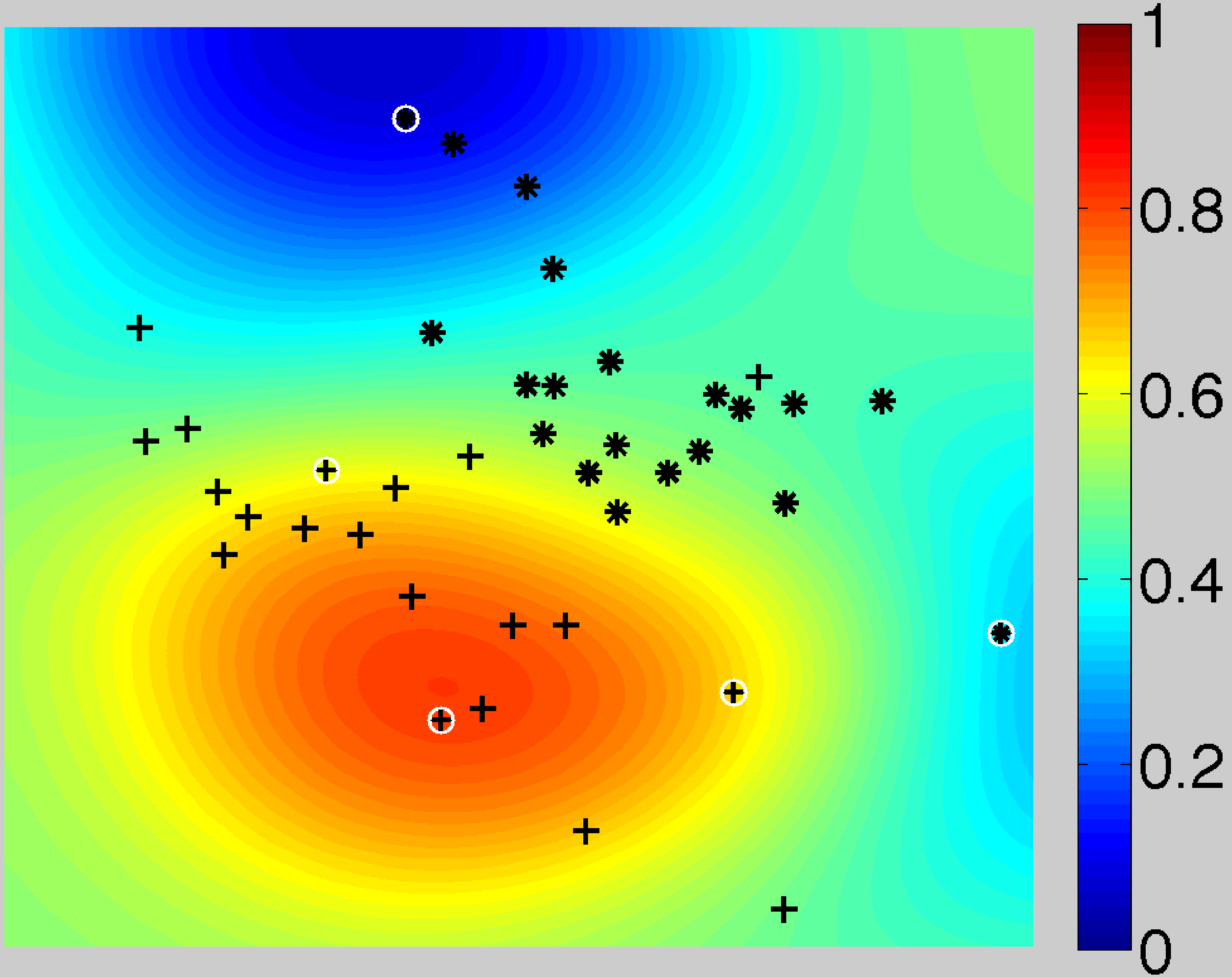


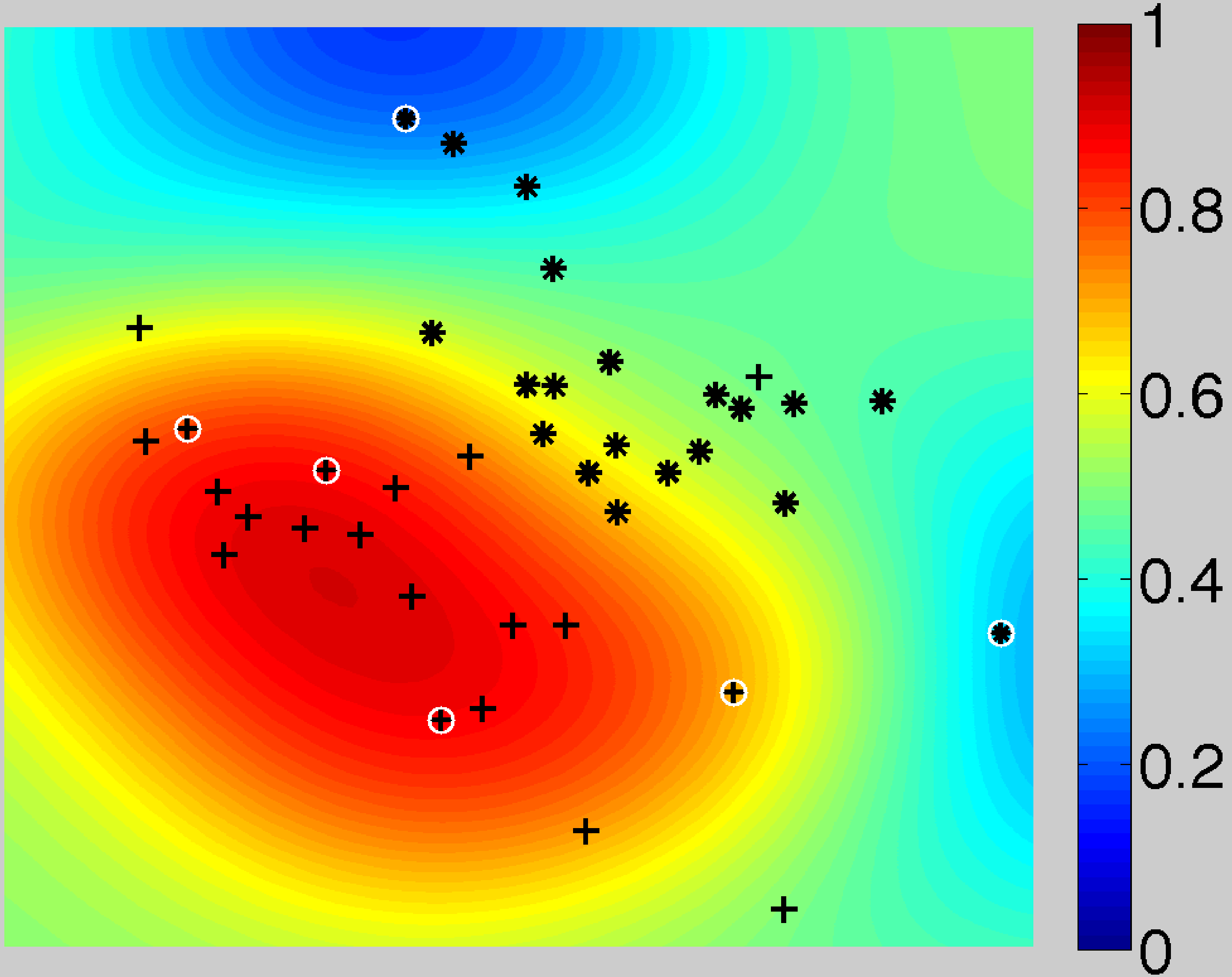

1
0.8
0.6
0.4
0.2
0

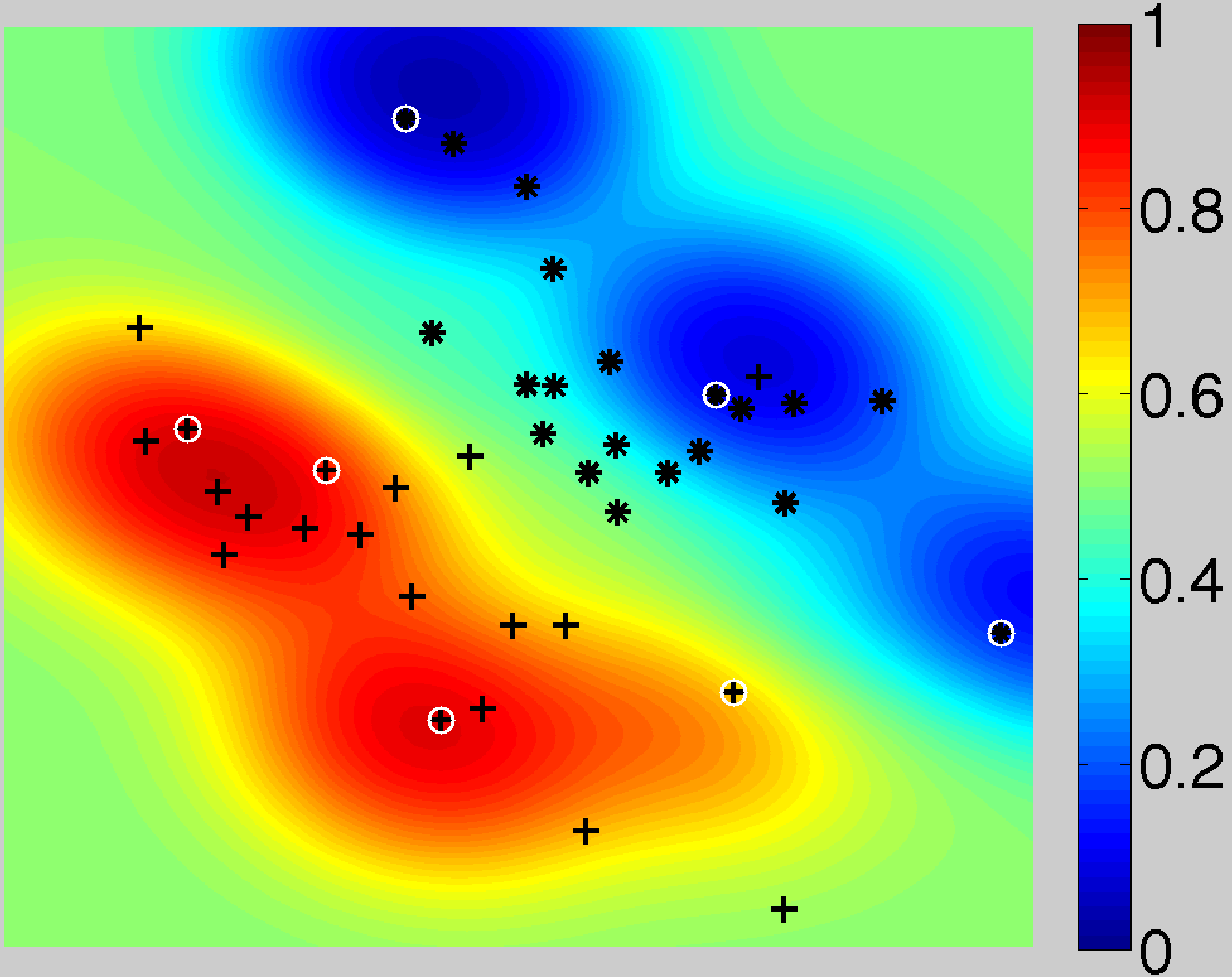

1
0.8
0.6
0.4
0.2
0

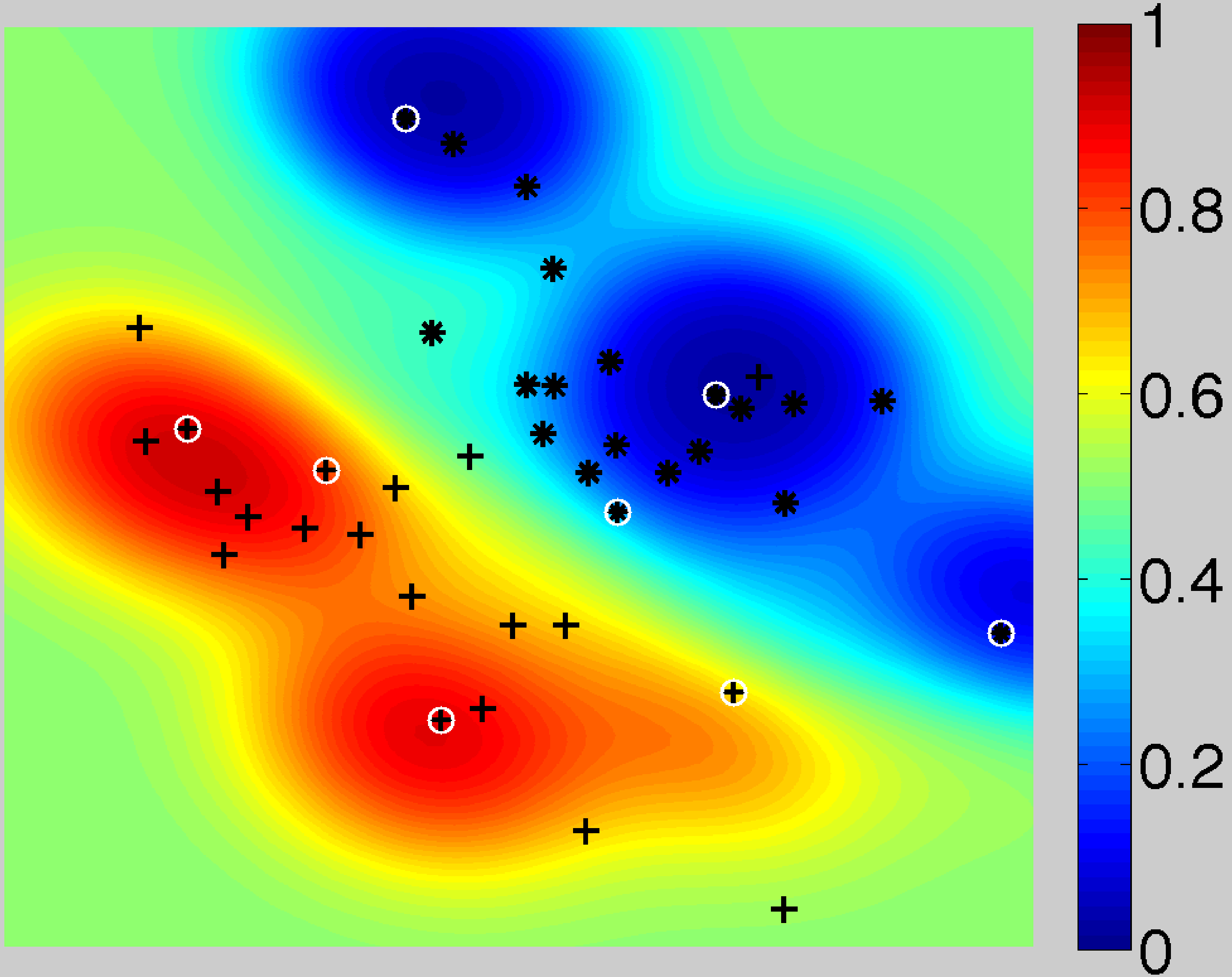

1
0.8
0.6
0.4
0.2
0


1
0.8
0.6
0.4
0.2
0

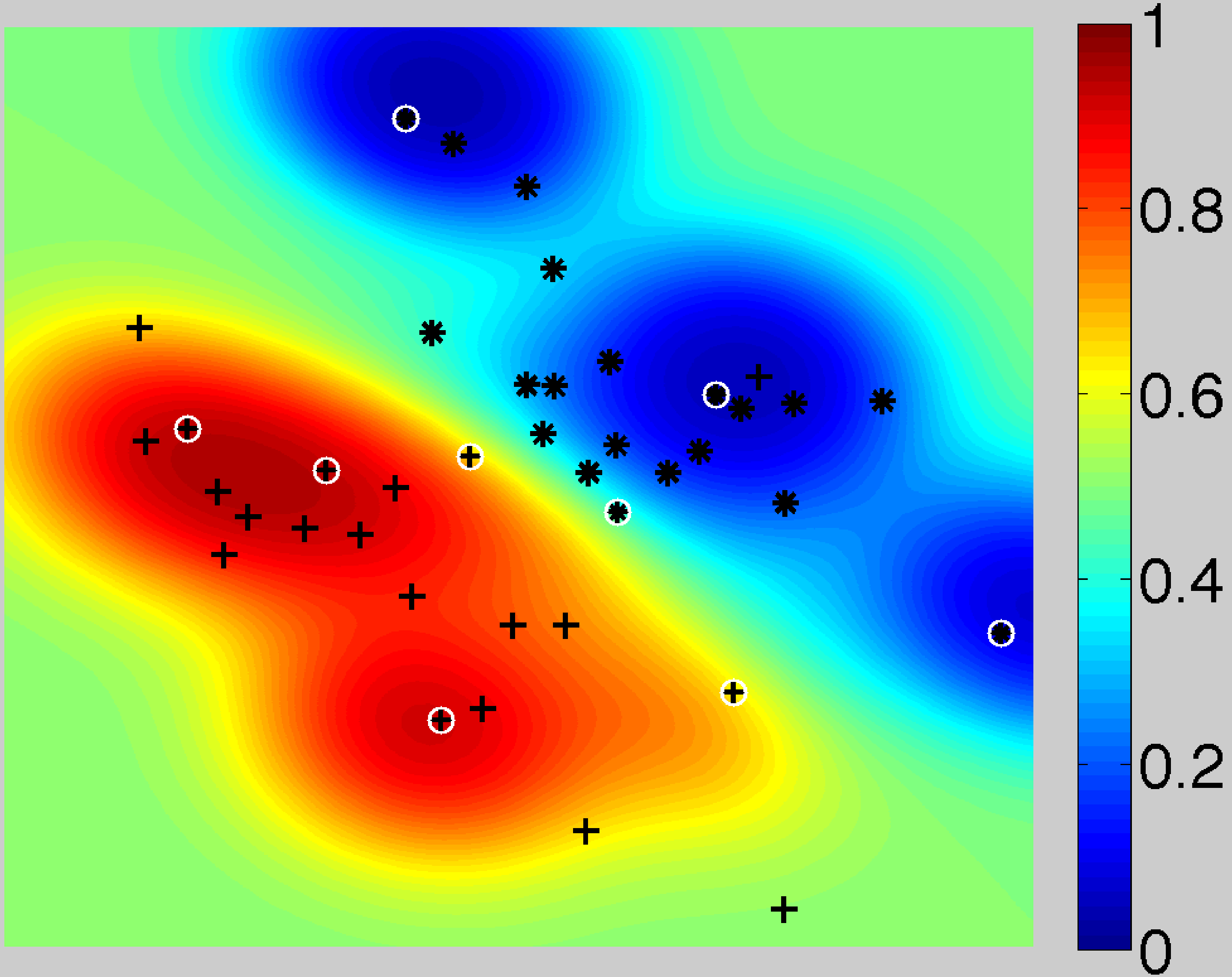

1
0.8
0.6
0.4
0.2
0


1
0.8
0.6
0.4
0.2
0

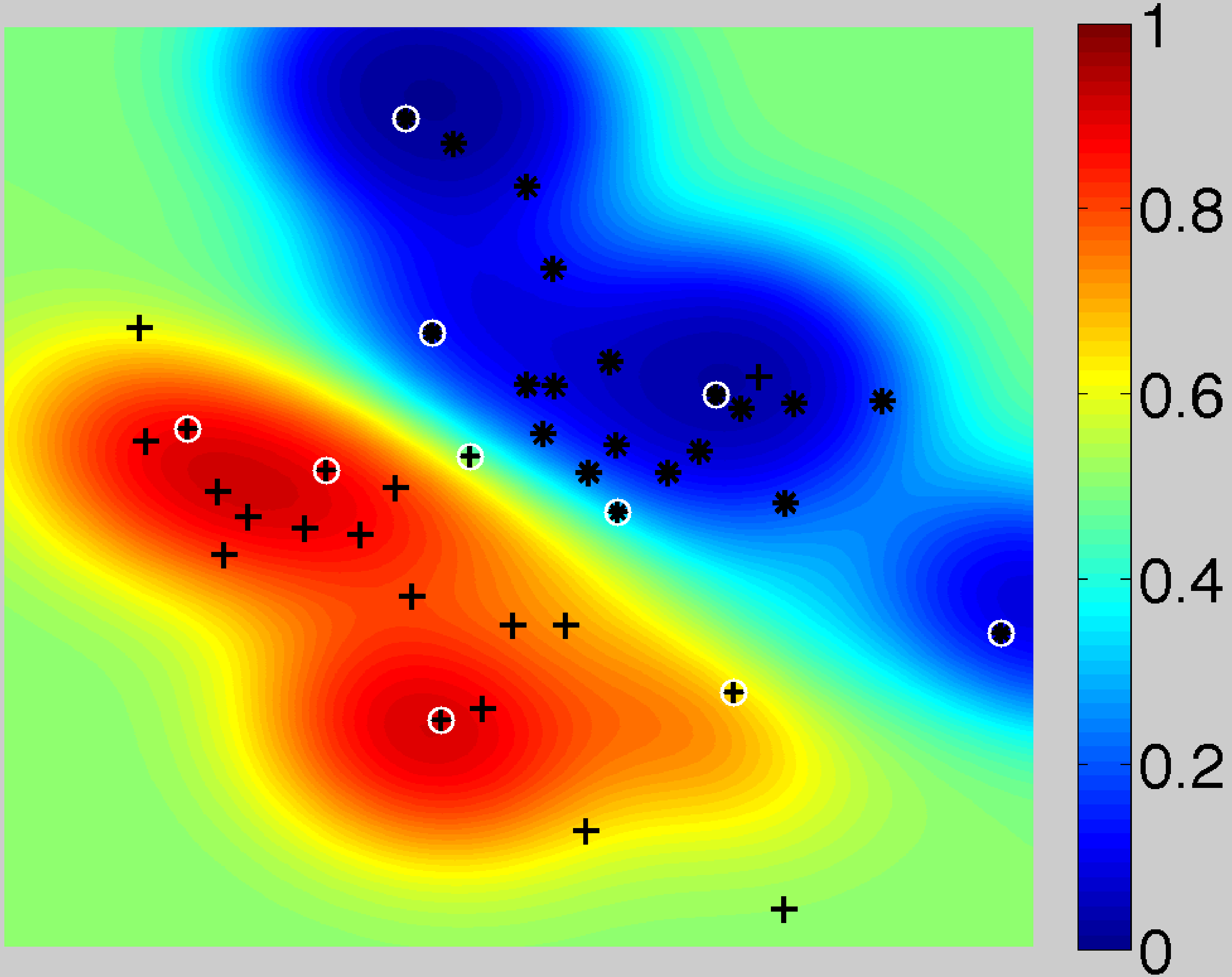

1
0.8
0.6
0.4
0.2
0

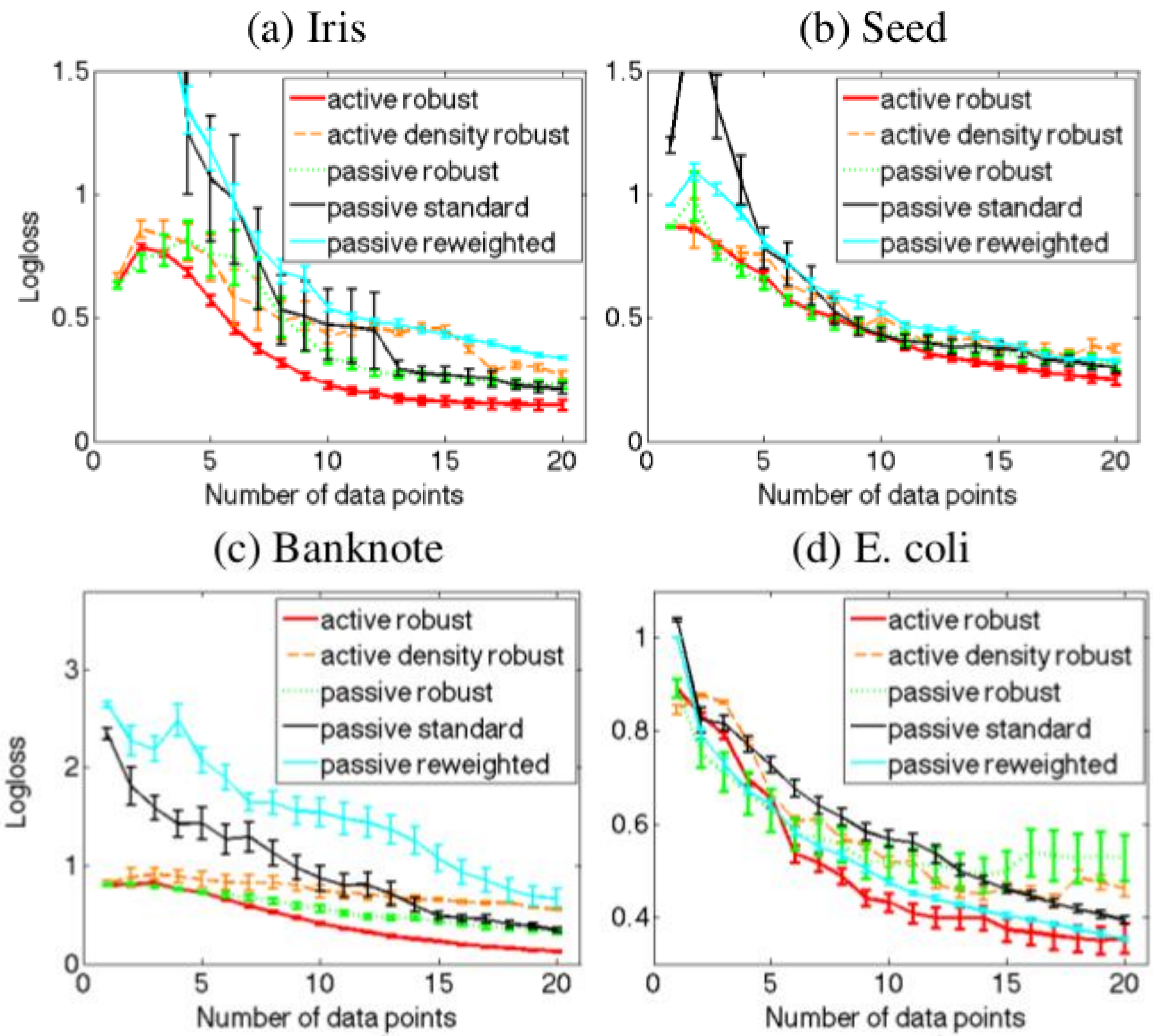


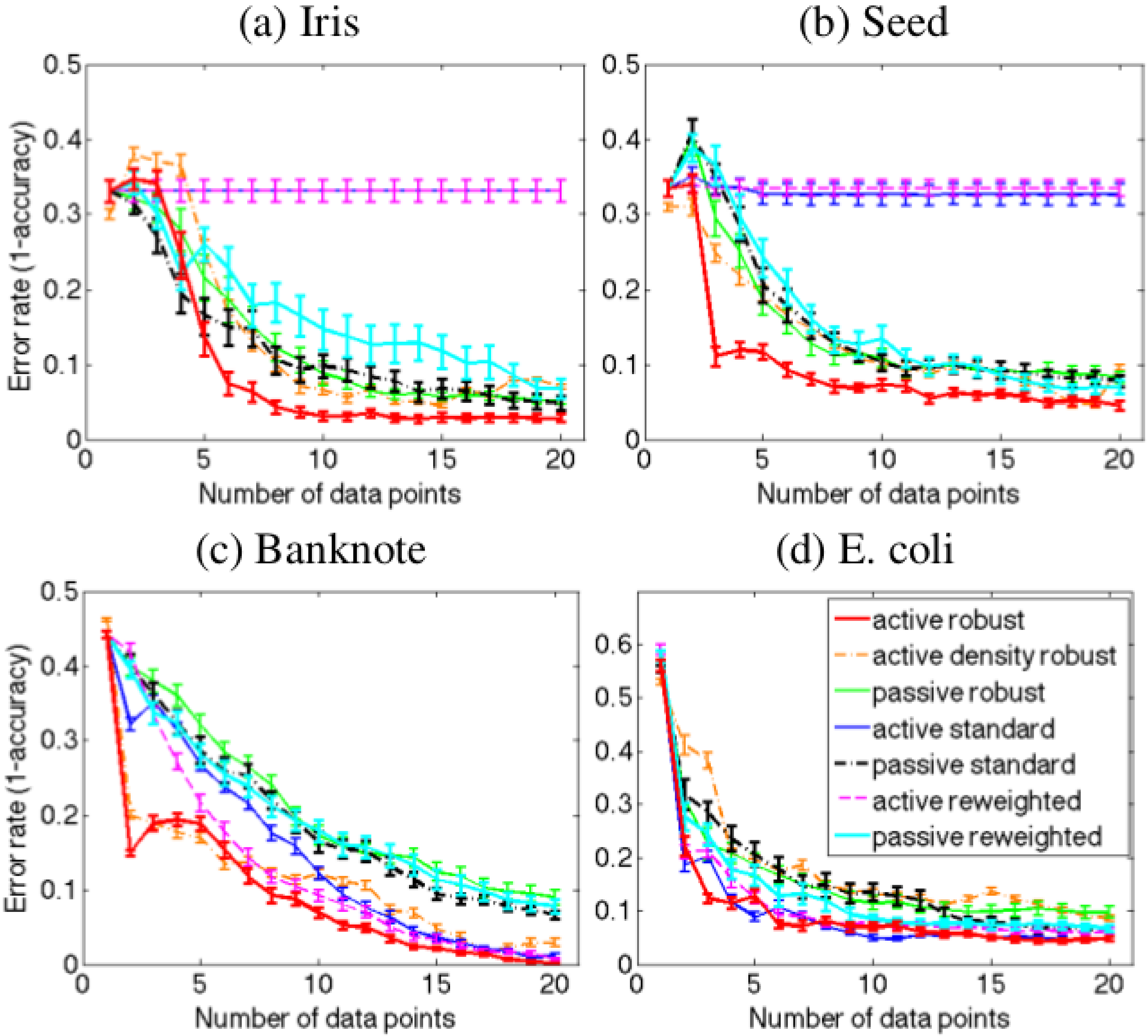
1
0.8
0.6
0.4
0.2
0



1
0.8
0.6
0.4
0.2
0


1
0.8
0.6
0.4
0.2
0


(a) Iris
(b) Seed
(c) Banknote
(d) E. coli
active robust
active density robust
passive robust
passive standard
passive reweighted
Logloss
Number of data points
0
5
10
15
20
0.5
1
1.5
2
3
0.4
0.6
0.8



(a) Iris
(b) Seed
(c) Banknote
(d) E. coli
Error rate (1-accuracy)
Number of data points
0
0.1
0.2
0.3
0.4
0.5
0.6
5
10
15
20
active robust
active density robust
passive robust
active standard
passive standard
active reweighted
passive reweighted

# Part 2: Group Fairness

Joint work with: Ashkan Rezaei, Rizal Fathony, Omid Memarrast (AAAI 2020)

# Fairness for data-driven decision making

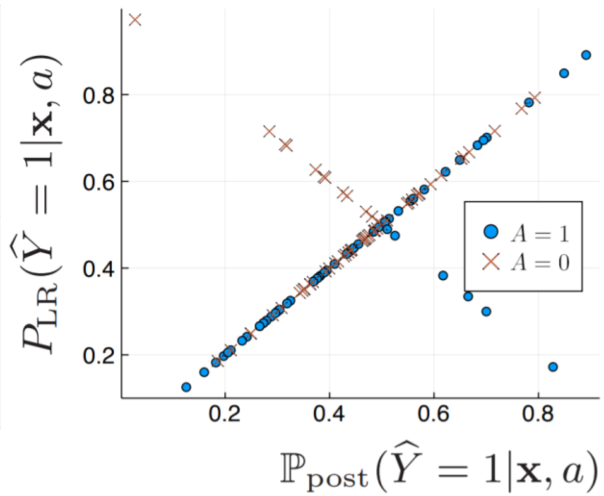
**Group 1**

| | Qualified | Unqualified |
|---|---|---|
| **Accept** | True Positive ($TP_1$) | False Positive ($FP_1$) |
| **Reject** | False Negative ($FN_1$) | True Negative ($TN_1$) |

**Group 2**

| | Qualified | Unqualified |
|---|---|---|
| **Accept** | True Positive ($TP_2$) | False Positive ($FP_2$) |
| **Reject** | False Negative ($FN_2$) | True Negative ($TN_2$) |

(Hardt et al. 2016)




**Demographic Parity:** Decision $\perp\!\!\!\perp$ Group

$$(TP_1+FP_1)/N_1 = (TP_2+FP_2)/N_2$$

**Equalized Opportunity:** Decision $\perp\!\!\!\perp$ Group | Qualified=True

$$TP_1/(TP_1+FN_1) = TP_2/(TP_2+FN_2)$$

**Equalized Odds:** Decision $\perp\!\!\!\perp$ Group | Qualified

$$TP_1/(TP_1+FN_1) = TP_2/(TP_2+FN_2);\ FP_1/(FP_1+TN_1) = FP_1/(FP_1+TN_1)$$

# Fair and Robust Log Loss Predictor

$$\min_{\mathbb{P}\in\Delta\cap\Gamma}\max_{\mathbb{Q}\in\Delta\cap\Xi}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \mathbb{Q}(\widehat{y}|\mathbf{x},a,y)}}\left[-\log\mathbb{P}(\widehat{Y}|\mathbf{X},A,Y)\right].$$

$$\Xi:\left\{\mathbb{Q}\mid\mathbb{E}_{\widetilde{P}(\mathbf{x});\mathbb{Q}(\widehat{y}|\mathbf{x})}[\phi(\mathbf{X},\widehat{Y})]=\mathbb{E}_{\widetilde{P}(\mathbf{x},y)}\left[\phi(\mathbf{X},Y)\right]\right\},$$

$\Gamma$ : $\mathbb{P}$ is fair

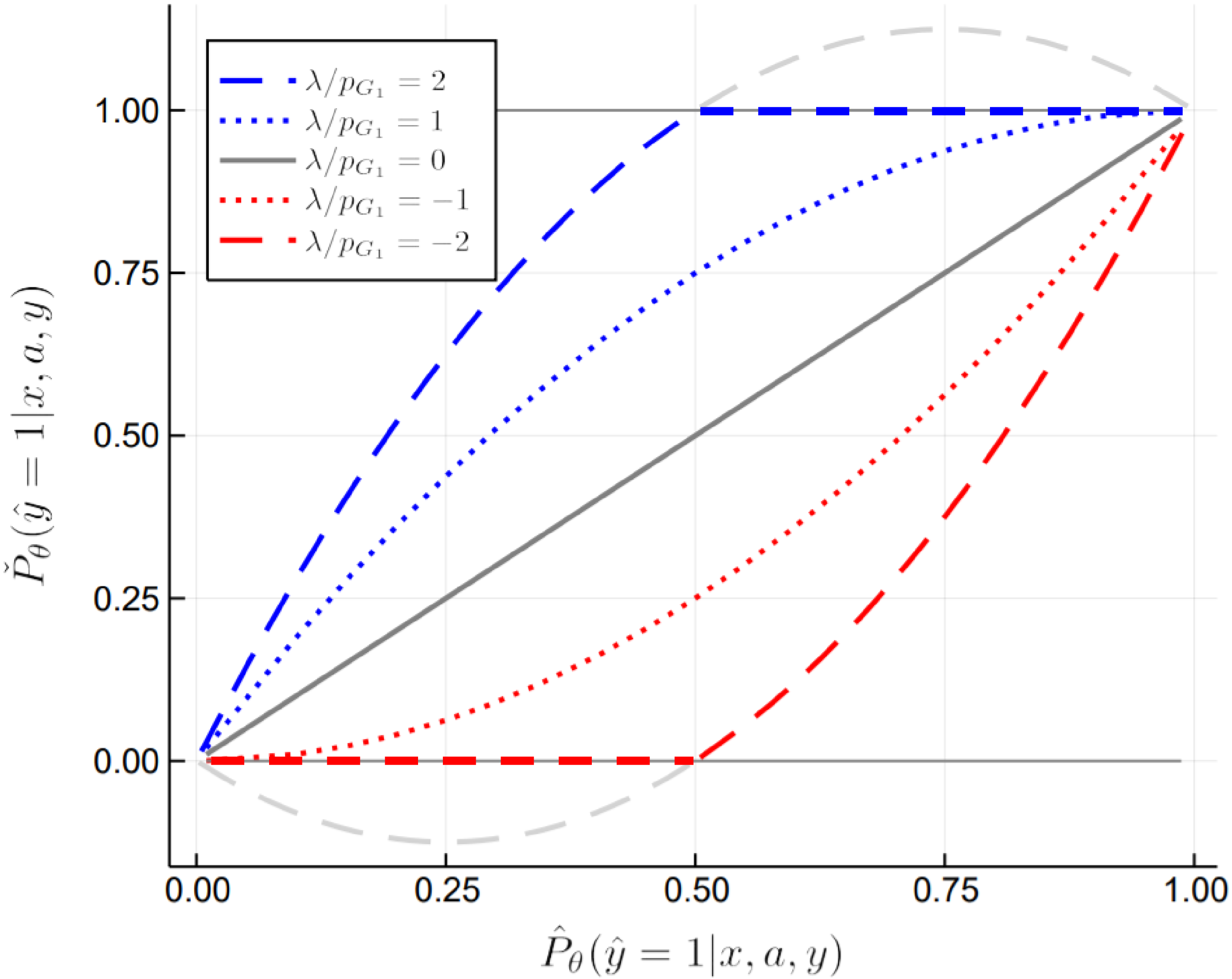
$$\Gamma:\left\{\mathbb{P}\mid\frac{1}{p_{\gamma_1}}\mathbb{E}_{\substack{\widetilde{P}(,a,y)\\ \mathbb{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_1(A,Y))]=\frac{1}{p_{\gamma_0}}\mathbb{E}_{\substack{\widetilde{P}(\mathbf{x},a,y)\\ \mathbb{P}(\widehat{y}|\mathbf{x},a,y)}}[\mathbb{I}(\widehat{Y}=1\wedge\gamma_0(A,Y))]\right\}$$

$$\begin{aligned}
\Gamma_{\mathsf{dp}} &\iff \gamma_j(A,Y)=\mathbb{I}(A=j);\\
\Gamma_{\mathsf{e.opp}} &\iff \gamma_j(A,Y)=\mathbb{I}(A=j\wedge Y=1);\\
\Gamma_{\mathsf{e.odd}} &\iff \gamma_j(A,Y)=\begin{bmatrix}\mathbb{I}(A=j\wedge Y=1)\\ \mathbb{I}(A=j\wedge Y=0)\end{bmatrix}.
\end{aligned}$$

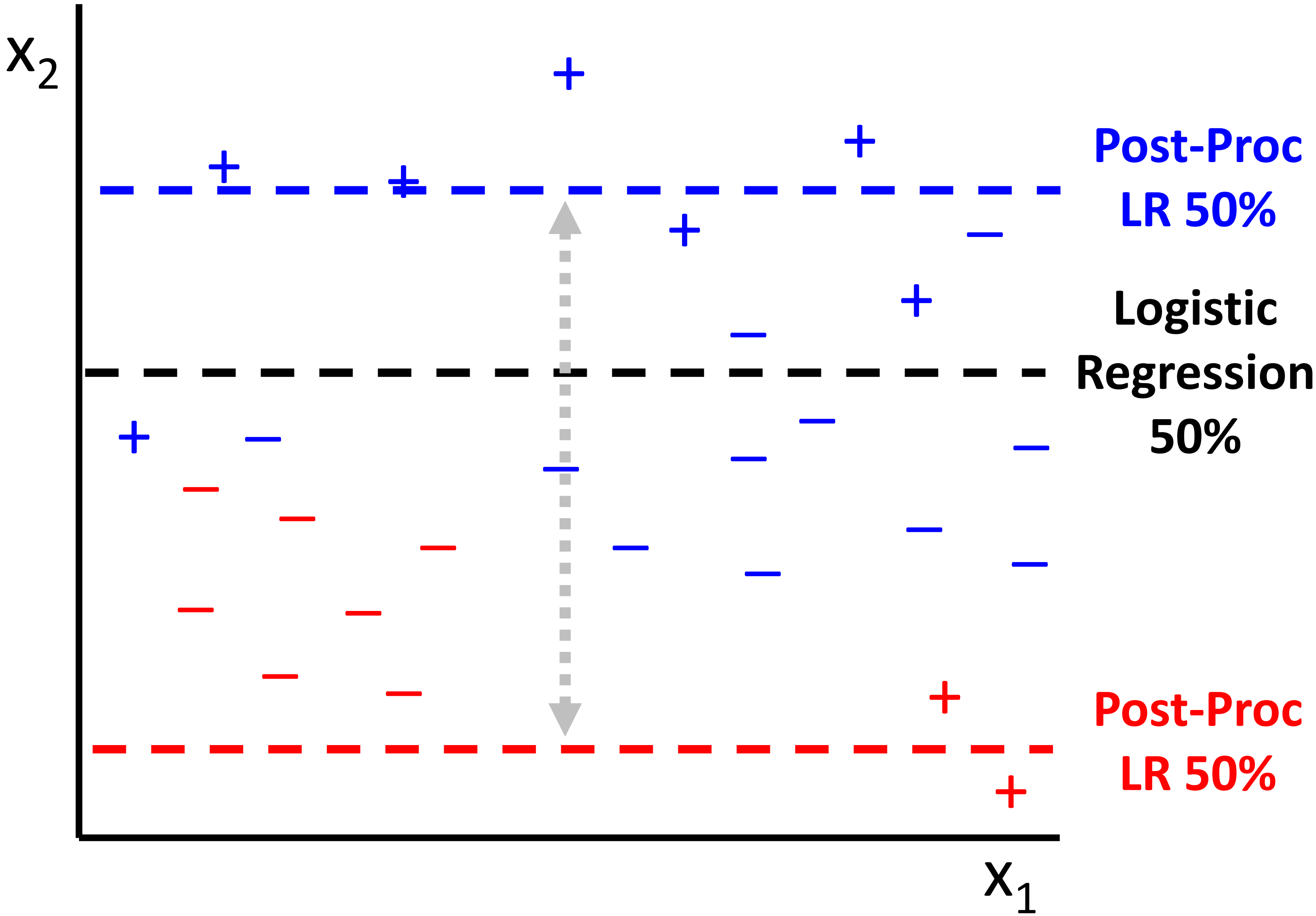
(Agarwal et al. 2018)

$$\widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \begin{cases} \min\left\{\frac{\exp(\theta^\top\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, \frac{p_{\gamma_1}}{\lambda}\right\} & \textit{if } \gamma_1(a,y) \\ \max\left\{\frac{\exp(\theta^\top\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})}, 1-\frac{p_{\gamma_0}}{\lambda}\right\} & \textit{if } \gamma_0(a,y) \\ \frac{\exp(\theta^\top\phi(\mathbf{x},1))}{Z_\theta(\mathbf{x})} & \textit{otherwise}; \end{cases}$$

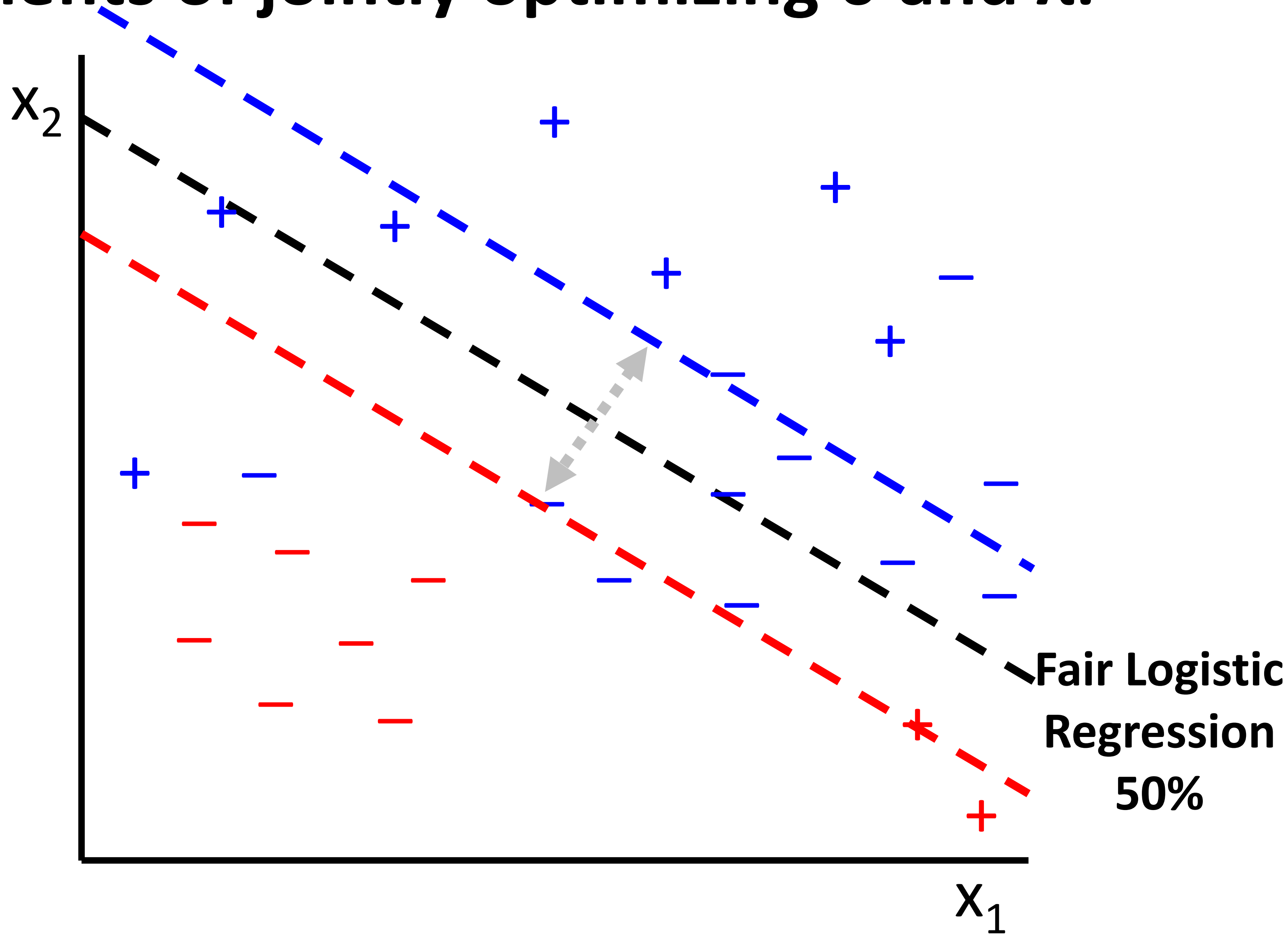
$$\breve{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y) = \widehat{P}_{\theta,\lambda}(\widehat{y}=1|\mathbf{x},a,y)\times \begin{cases} \left(1+\frac{\lambda}{p_{\gamma_1}}\widehat{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\right) & \textit{if } \gamma_1(a,y) \\ \left(1-\frac{\lambda}{p_{\gamma_0}}\widehat{P}_{\theta,\lambda}(\widehat{y}=0|\mathbf{x},a,y)\right) & \textit{if } \gamma_0(a,y) \\ 1 & \textit{otherwise}. \end{cases}$$



# Benefits of jointly optimizing θ and λ:



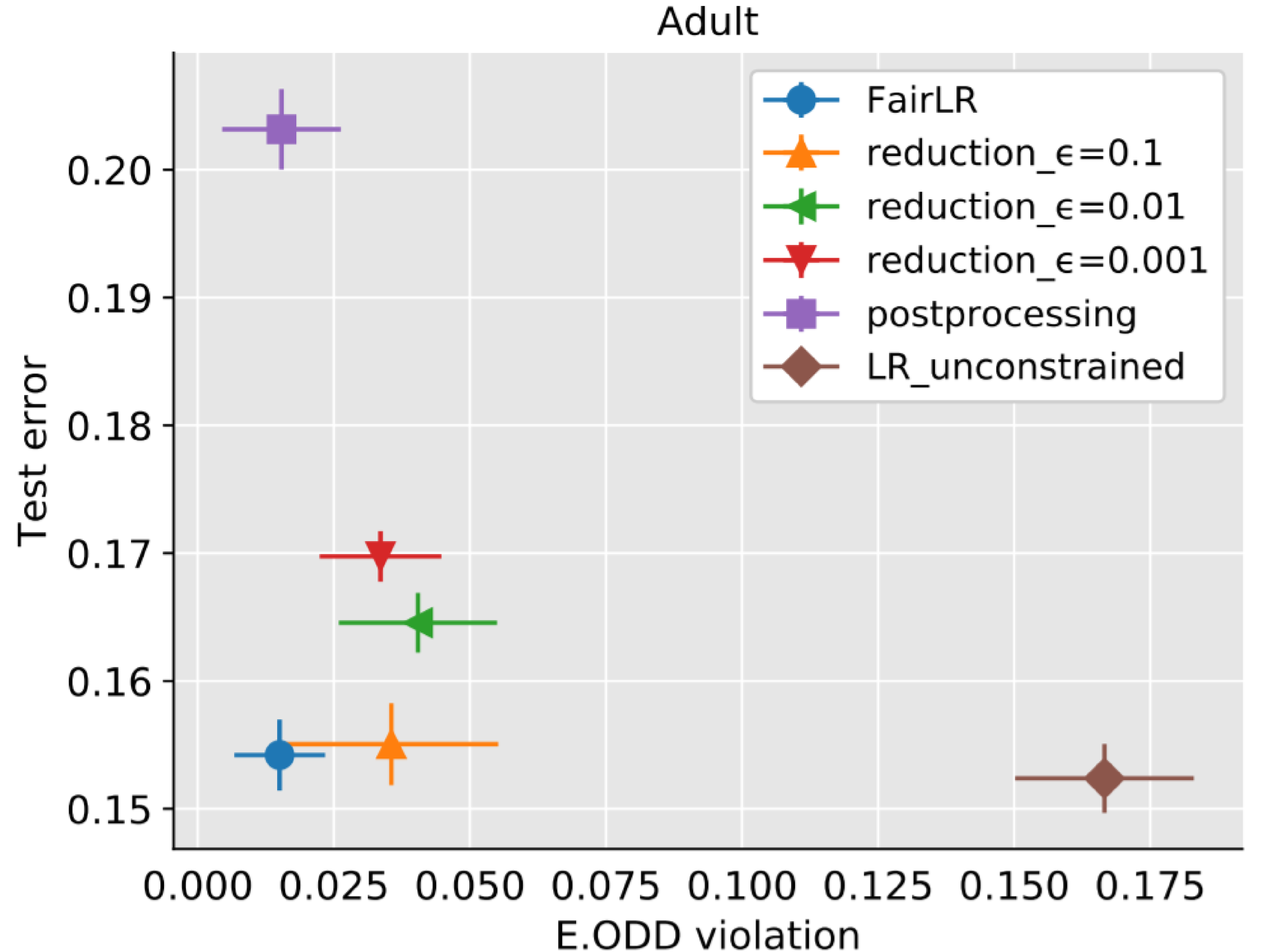
# Benefits of jointly optimizing θ and λ:



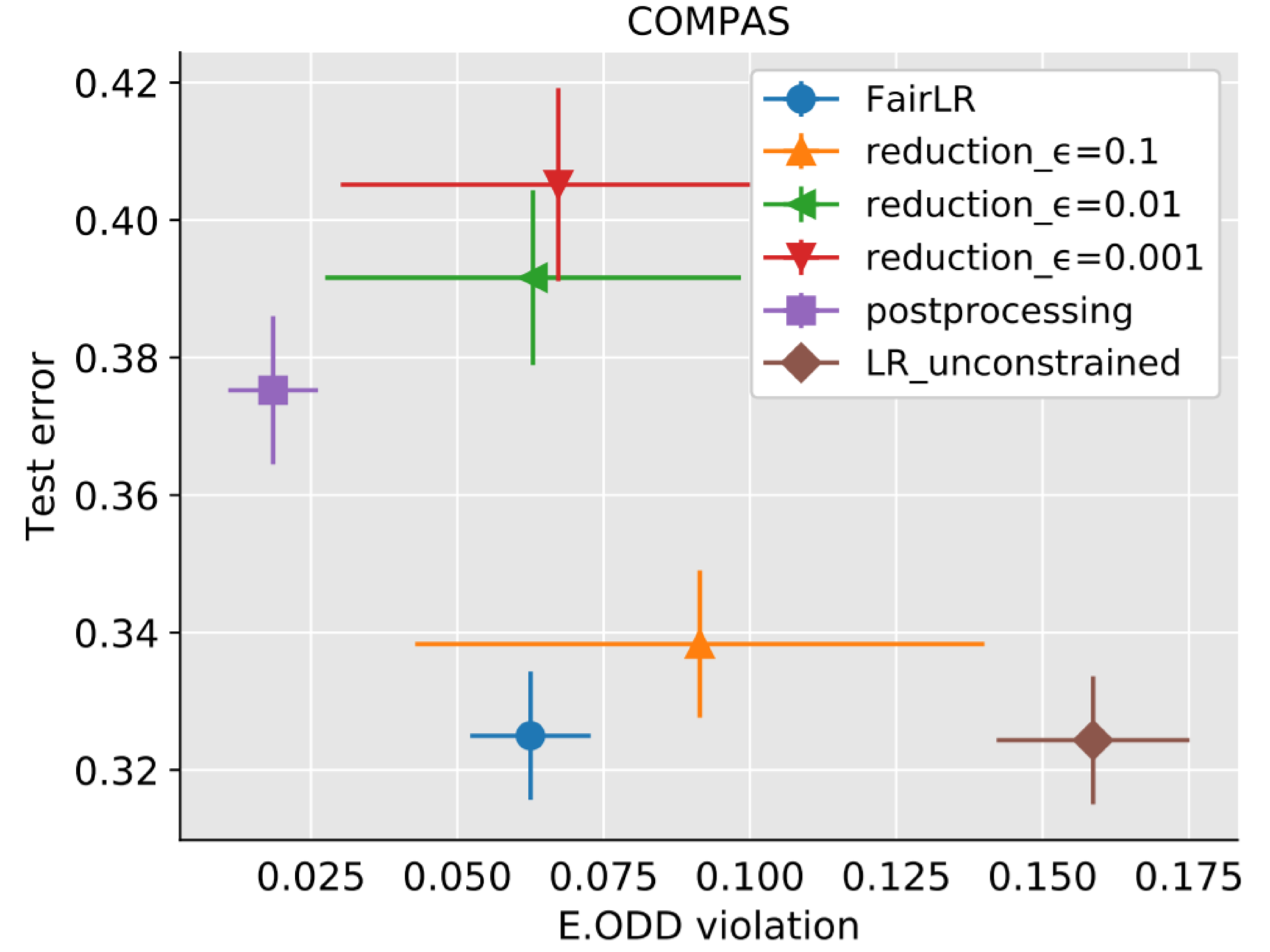
# Experiments

| | Label | Protected Attribute | Features | Examples |
|---|---|---|---|---|
| **UCI Adult** | Income > $50k | Gender | 12 | 45,222 |
| **COMPAS** | recidivism | Race | 10 | 6,167 |

- Evaluation on 20 random splits (70%/30% train/test)
- Baselines
  - Unconstrained (unfair) logistic regression
  - Reweighting approach (Kamiran & Calders 2012)
  - Cost-sensitive reduction approach (Agarwal et al. 2018)
  - Post-processing (Hardt et al. 2016)



Adult
FairLR
reduction_ϵ=0.1
reduction_ϵ=0.01
reduction_ϵ=0.001
postprocessing
LR_unconstrained
Test error
0.20
0.19
0.18
0.17
0.16
0.15
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
E.ODD violation

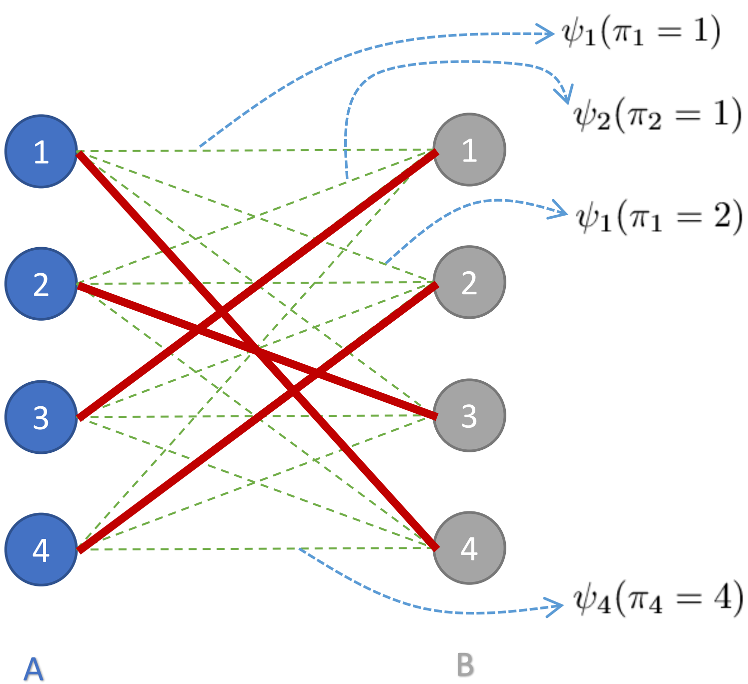


COMPAS
Test error
E.ODD violation
0.42
0.40
0.38
0.36
0.34
0.32
0.025
0.050
0.075
0.100
0.125
0.150
0.175
FairLR
reduction_ϵ=0.1
reduction_ϵ=0.01
reduction_ϵ=0.001
postprocessing
LR_unconstrained

# Part 3: Structured Prediction

Joint work with: Rizal Fathony, Sima Behpour, Xinhua Zhang (ICML 2018)

# Bipartite Matching Task




$\pi = [4, 3, 1, 2]$

Maximum weighted bipartite matching:

$$\max_{\pi \in \Pi} \psi(\pi) = \max_{\pi \in \Pi} \sum_i \psi_i(\pi_i)$$

Machine learning task:

Learn appropriate weights $\psi_i(\cdot)$

Objective:

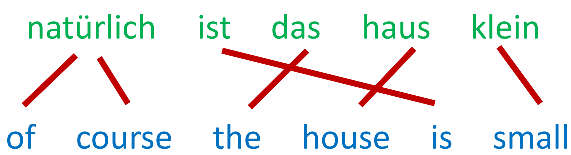
Minimize a loss metric, e.g., the Hamming loss

$$\mathrm{loss}_{\mathrm{Ham}}(\pi, \pi') = \sum_{i=1}^{n} 1(\pi'_i \neq \pi_i)$$

# Bipartite Matching Applications

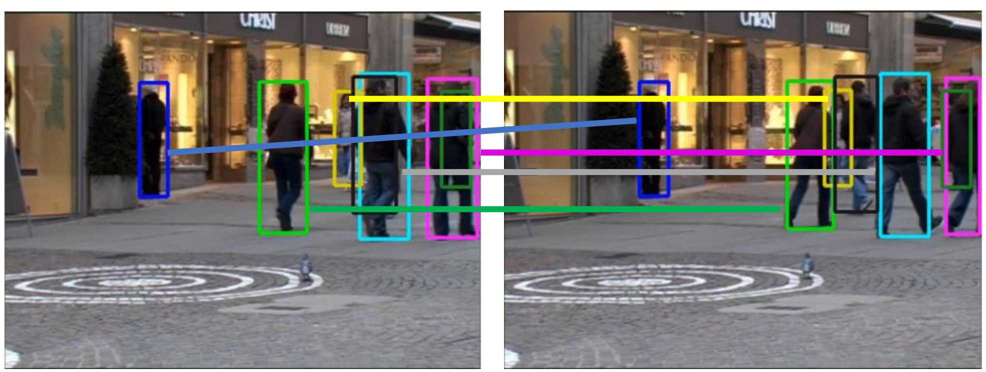
## Word alignment

(Taskar et. al., 2005; Pado & Lapta, 2006; Mac-Cartney et. al., 2008)



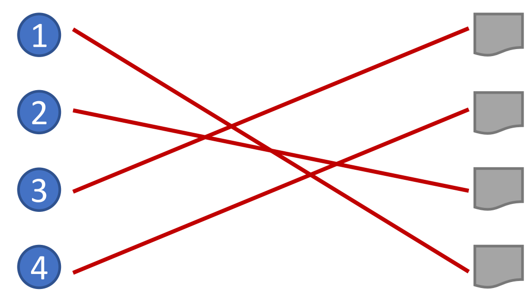

## Correspondence between images

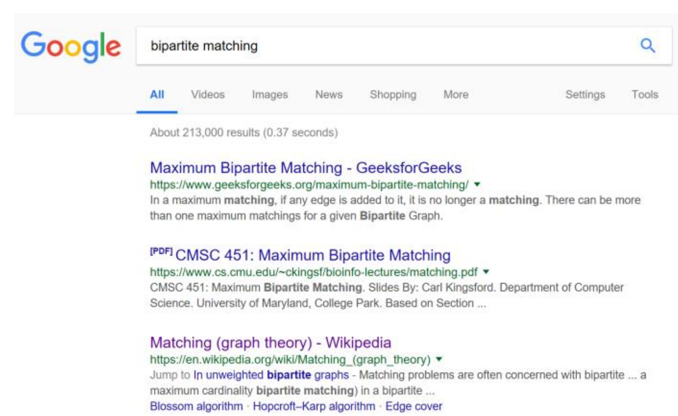
(Belongie et. al., 2002; Dellaert et. al., 2003)



## Learning to rank documents

(Dwork et. al., 2001; Le & Smola, 2007)



# Learning Bipartite Matchings

## 1 Conditional Random Field

(Petterson et. al., 2009; Volkovs & Zemel, 2012)

$$P_\psi(\pi) = \frac{1}{Z_\psi} \exp\left(\sum_{i=1}^{n} \psi_i(\pi_i)\right)$$

$$Z_\psi = \sum_{\pi} \prod_{i=1}^{n} \exp\left(\psi_i(\pi_i)\right) = \text{perm}(\mathbf{M})$$

where $M_{i,j} = \exp\left(\psi_i(j)\right)$

**Fisher consistent**
Produces Bayes optimal prediction in ideal case

**Computationally intractable**
Normalization term requires matrix permanent computation (a #P-hard problem). Approximation is needed.

## 2 Structured SVM

(Tsochantaridis et. al., 2005)

Based on CS hinge loss
solved using constraint generation

$$\min_{\psi} \mathbb{E}_{\pi\sim\tilde{P}}\left[\max_{\pi'} \{\text{loss}(\pi,\pi') + \psi(\pi')\} - \psi(\pi)\right]$$

$\tilde{P}$ is the empirical distribution

**Computationally efficient**
Hungarian algorithm for computing the maximum violated constraints

**No Fisher consistency guarantee**
Not consistent for distribution with no majority label

# Adversarial Bipartite Matchings

[Fathony et al., ICML 2018]

**Fisher consistency guaranteed**

Primal:

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} [\text{loss}(\hat{\pi},\check{\pi})]$$

$$\text{s.t. } \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x,\check{\pi}_i)\right] = \mathbb{E}_{(x,\pi)\sim\tilde{P}} \left[\sum_{i=1}^{n} \phi_i(x,\pi_i)\right]$$

Dual:

$$\min_{\theta} \mathbb{E}_{x,\pi\sim\tilde{P}} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\substack{\hat{\pi}|x\sim\hat{P} \\ \check{\pi}|x\sim\check{P}}} \left[\text{loss}(\hat{\pi},\check{\pi}) + \theta\cdot\sum_{i=1}^{n} (\phi_i(x,\check{\pi}_i) - \phi_i(x,\pi_i))\right]$$

Augmented Hamming loss matrix for $n = 3$ permutations

| | $\check{\pi}=123$ | $\check{\pi}=132$ | $\check{\pi}=213$ | $\check{\pi}=231$ | $\check{\pi}=312$ | $\check{\pi}=321$ |
|---|---|---|---|---|---|---|
| $\hat{\pi}=123$ | $0+\delta_{123}$ | $2+\delta_{132}$ | $2+\delta_{213}$ | $3+\delta_{231}$ | $3+\delta_{312}$ | $2+\delta_{321}$ |
| $\hat{\pi}=132$ | $2+\delta_{123}$ | $0+\delta_{132}$ | $3+\delta_{213}$ | $2+\delta_{231}$ | $2+\delta_{312}$ | $3+\delta_{321}$ |
| $\hat{\pi}=213$ | $2+\delta_{123}$ | $3+\delta_{132}$ | $0+\delta_{213}$ | $2+\delta_{231}$ | $2+\delta_{312}$ | $3+\delta_{321}$ |
| $\hat{\pi}=231$ | $3+\delta_{123}$ | $2+\delta_{132}$ | $2+\delta_{213}$ | $0+\delta_{231}$ | $3+\delta_{312}$ | $2+\delta_{321}$ |
| $\hat{\pi}=312$ | $3+\delta_{123}$ | $2+\delta_{132}$ | $2+\delta_{213}$ | $3+\delta_{231}$ | $0+\delta_{312}$ | $2+\delta_{321}$ |
| $\hat{\pi}=321$ | $2+\delta_{123}$ | $3+\delta_{132}$ | $3+\delta_{213}$ | $2+\delta_{231}$ | $2+\delta_{312}$ | $0+\delta_{321}$ |

size: $n! \times n!$

Intractable for modestly-sized $n$

# Adversarial Bipartite Matchings

[Fathony et al., ICML 2018]

**Dual:** $\min_{\theta} \mathbb{E}_{(x,\pi)\sim\tilde{P}} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} \left[ \sum_{i=1}^{n} I(\pi'_i \neq \pi_i) + \theta \cdot \sum_{i=1}^{n} (\phi_i(x,\check{\pi}_i) - \phi_i(x,\pi_i)) \right]$

## Marginal Distribution Matrices:

Predictor $\mathbf{P} =$

| | 1 | 2 | 3 |
|---|---|---|---|
| $\hat{\pi}_1$ | $p_{1,1}$ | $p_{1,2}$ | $p_{1,3}$ |
| $\hat{\pi}_2$ | $p_{2,1}$ | $p_{2,2}$ | $p_{2,3}$ |
| $\hat{\pi}_3$ | $p_{3,1}$ | $p_{3,2}$ | $p_{3,3}$ |

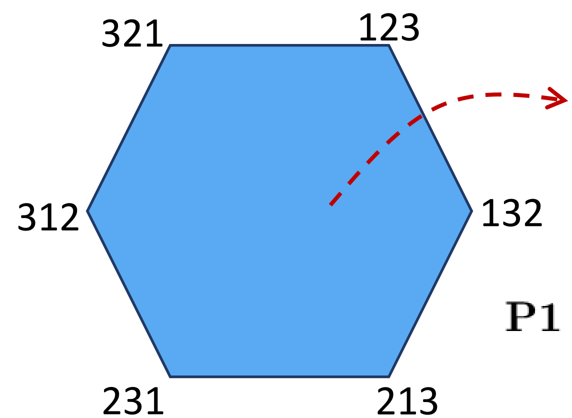
$$p_{i,j} = \hat{P}(\hat{\pi}_i = j)$$

Adversary $\mathbf{Q} =$

| | 1 | 2 | 3 |
|---|---|---|---|
| $\check{\pi}_1$ | $q_{1,1}$ | $q_{1,2}$ | $q_{1,3}$ |
| $\check{\pi}_2$ | $q_{2,1}$ | $q_{2,2}$ | $q_{2,3}$ |
| $\check{\pi}_3$ | $q_{3,1}$ | $q_{3,2}$ | $q_{3,3}$ |

$$q_{i,j} = \check{P}(\check{\pi}_i = j)$$

## Birkhoff – Von Neumann theorem:




convex polytope whose points are doubly stochastic matrices

$$\mathbf{P1} = \mathbf{P}^\top \mathbf{1} = \mathbf{Q1} = \mathbf{Q}^\top \mathbf{1} = \mathbf{1}$$

reduce variables from $O(n!)$ to $O(n^2)$

## Marginal Formulation:

Rearrange the optimization order and add regularization and smoothing penalties

$$\max_{\mathbf{Q}\geq\mathbf{0}} \min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \min_{\mathbf{P}_i\geq\mathbf{0}} \left[ \langle \mathbf{Q}_i - \mathbf{Y}_i, \textstyle\sum_k \theta_k \mathbf{X}_{i,k} \rangle - \langle \mathbf{P}_i, \mathbf{Q}_i \rangle + \frac{\mu}{2}\|\mathbf{P}_i\|_F^2 - \frac{\mu}{2}\|\mathbf{Q}_i\|_F^2 \right] + \frac{\lambda}{2}\|\theta\|_2^2$$

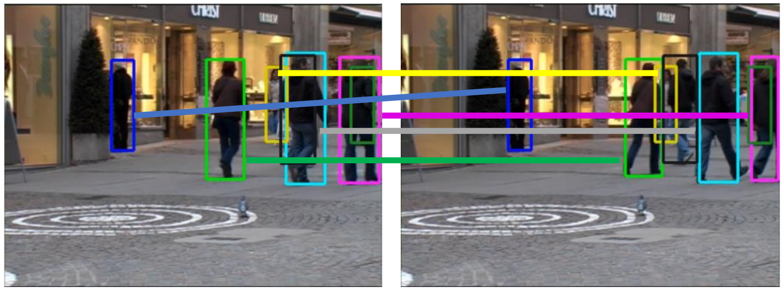
$$\text{s.t. : } \mathbf{P}_i\mathbf{1} = \mathbf{P}_i^\top\mathbf{1} = \mathbf{Q}_i\mathbf{1} = \mathbf{Q}_i^\top\mathbf{1} = \mathbf{1}, \quad \forall i$$

projected Quasi-Newton (Schmidt, et.al., 2009) for Q; closed-form for $\theta$;

projection to doubly-stochastic matrix for P using ADMM

# Adversarial Bipartite Matchings
[Fathony et al., ICML 2018]

## Application: Video Tracking



## Datasets

*Table 3.* Dataset properties

| DATASET | # ELEMENTS | # EXAMPLES |
|---|---|---|
| TUD-CAMPUS | 12 | 70 |
| TUD-STADTMITTE | 16 | 178 |
| ETH-SUNNYDAY | 18 | 353 |
| ETH-BAHNHOF | 34 | 999 |
| ETH-PEDCROSS2 | 30 | 836 |

## Empirical runtime (until convergence)

*Table 5.* Running time (in seconds) of the model for various number of elements $n$ with fixed number of samples ($m = 50$)

| DATASET | # ELEMENTS | ADV MARG. | SSVM |
|---|---|---|---|
| CAMPUS | 12 | 1.96 | 0.22 |
| STADTMITTE | 16 | 2.46 | 0.25 |
| SUNNYDAY | 18 | 2.75 | 0.15 |
| PEDCROSS2 | 30 | 8.18 | 0.26 |
| BAHNHOF | 34 | 9.79 | 0.31 |

Adversarial. Marginal Formulation: grows (roughly) quadratically in $n$

CRF: impractical even for $n = 20$
(Petterson et. al., 2009)

# Adversarial Bipartite Matchings

[Fathony et al., ICML 2018]

Table 1: The mean and standard deviation (in parenthesis) of the average accuracy (1 - the average Hamming loss) for the adversarial bipartite matching model compared with Structured-SVM.

| Training/ Testing | Adv. Bipartite Matching | Structured SVM |
|---|---|---|
| Campus/ Stadtmitte | 0.662 (0.08) | 0.662 (0.08) |
| Stadtmitte/ Campus | 0.667 (0.11) | 0.660 (0.12) |
| Bahnhof/ Sunnyday | **0.754** (0.10) | 0.729 (0.15) |
| Pedcross2/ Sunnyday | **0.750** (0.10) | 0.736 (0.13) |
| Sunnyday/ Bahnhof | **0.751** (0.18) | 0.739 (0.20) |
| Pedcross2/ Bahnhof | **0.763** (0.16) | 0.731 (0.21) |
| Bahnhof/ Pedcross2 | **0.714** (0.16) | 0.701 (0.18) |
| Sunnyday/ Pedcross2 | **0.712** (0.17) | 0.700 (0.18) |

6 pairs of dataset significantly outperforms SSVM

2 pairs of dataset competitive with SSVM

# Adversarial Bipartite Matchings

[Fathony et al., ICML 2018]

| | **Efficient?** | **Consistent?** | **Performs well?** |
|---|---|---|---|
| Conditional Random Field (CRF)<br>(Petterson et. al., 2009; Volkovs & Zemel, 2012) | ✗ | ✓ | ? |
| Structured SVM<br>(Tsochantaridis et. al., 2005) | ✓ | ✗ | – |
| Adversarial Bipartite Matching<br>(our approach) | ✓ | ✓ | ✓ |

# Summary & Conclusions

$$\min_{\hat{P}(\hat{y}|\mathbf{x}) \in \Delta \cap \Gamma} \max_{\check{P}(\check{y}|\mathbf{x}) \in \Delta \cap \Xi} \mathbb{E}_{\substack{\mathbf{x} \sim \tilde{P} \\ \hat{\mathbf{y}}|\mathbf{x} \sim \hat{P} \\ \check{\mathbf{y}}|\mathbf{x} \sim \check{P}}} \left[ \text{loss}(\hat{\mathbf{Y}}, \check{\mathbf{Y}}) \right]$$

**Covariate Shift/Active Learning:** $P_{train}(\mathbf{x}) \neq P_{test}(\mathbf{x})$

→ **Avoids harmful extrapolations**

**Fairness:** Minimizer also satisfies fairness requirements (Γ)

→ **Robust/smooth group fairness**

**Structured Prediction:** Structured objects **y**, bilinear loss

→ **Consistency and Computational Tractability**

**Foundational framework for parametric predictors**

**Versatile for a wide range of settings**

# Questions?


- Liu, Ziebart. "*Robust Classification Under Sample Selection Bias*." NeurIPS 2014.
- Liu, Reyzin, Ziebart. "*Shift-Pessimistic Active Learning Using Robust Bias-Aware Prediction*." AAAI 2015.
- Rezaei, Fathony, Memmarest, Ziebart. "*Fairness for robust log loss classification*." AAAI 2020.
- Fathony, Behpour, Zhang, Ziebart. "*Efficient and Consistent Adversarial Bipartite Matching*." ICML 2018.
- Fathony, Liu, Asif, Ziebart. "*Adversarial Multiclass Classification: A Risk Minimization Perspective*." NeurIPS 2016.
- Fathony, Bashiri, Ziebart. "*Adversarial Surrogate Losses for Ordinal Regression*." NeurIPS 2017.
- Wang, Xing, Asif, Ziebart. "*Adversarial Prediction Games for Multivariate Losses*." NeurIPS 2015.
- Tirinzoni, Chen, Petrik, Ziebart. "*Policy-conditioned uncertainty sets for robust Markov Decision Processes*." NeurIPS 2018.
- Bashiri, Ziebart, Zhang. "*Distributionally Robust Imitation Learning*." NeurIPS 2021.



NRI Award #1227495 CAREER Award #1652530

IIS Award #1526379 FAI Award #1939743