Class Ratio Estimation using MMD

J. SAKETHA NATH (IIT BOMBAY)

COLLABORATORS: ARUN IYER (YAHOO!), SUNITA SARAWAGI (IIT B)

Motivation

Excellent middle eastern cuisine on historic Murphy avenue in Sunnyvale. We had a reservation for 8, and they were kind enough to seat us outdoors, which was wonderful on this beautiful day in...more

Came here for the first time a couple weeks ago on a week night - wait was not that bad. We were seated promptly and had time to look over menu. I ordered the Beriani Dajaj with Chicken (I saw...more

SO MAD! I have been driving past this place for months now. It always looked good, and the pictures online looked lovely. Sadly, not the case when you come in. I walked in and no one was at the...more

Yahoo! Local Restaurant Reviews

Motivation

Excellent middle eastern cuisine on historic Murphy avenue in Sunnyvale. We had a reservation for 8, and they were kind enough to seat us outdoors, which was wonderful on this beautiful day in...more

Came here for the first time a couple weeks ago on a week night - wait was not that bad. We were seated promptly and had time to look over menu. I ordered the Beriani Dajaj with Chicken (I saw...more

SO MAD! I have been driving past this place for months now. It always looked good, and the pictures online looked lovely. Sadly, not the case when you come in. I walked in and no one was at the...more

Yahoo! Local Restaurant Reviews

Motivation

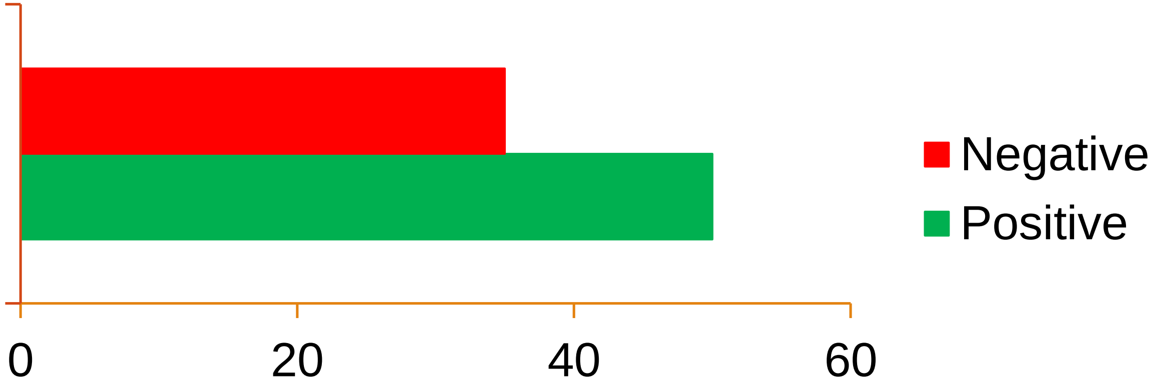
Excellent middle eastern cuisine on historic Murphy avenue in Sunnyvale. We had a reservation for 8, and they were kind enough to seat us outdoors, which was wonderful on this beautiful day in...more

Came here for the first time a couple weeks ago on a week night - wait was not that bad. We were seated promptly and had time to look over menu. I ordered the Beriani Dajaj with Chicken (I saw...more

SO MAD! I have been driving past this place for months now. It always looked good, and the pictures online looked lovely. Sadly, not the case when you come in. I walked in and no one was at the...more

Yahoo! Local Restaurant Reviews

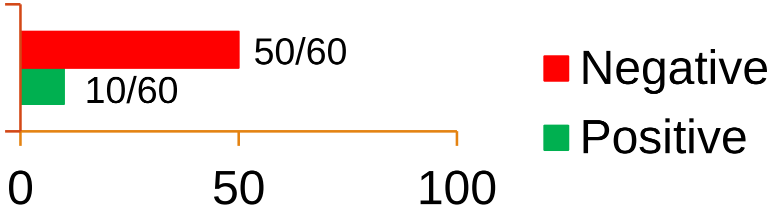
No. +ve , -ve is enough

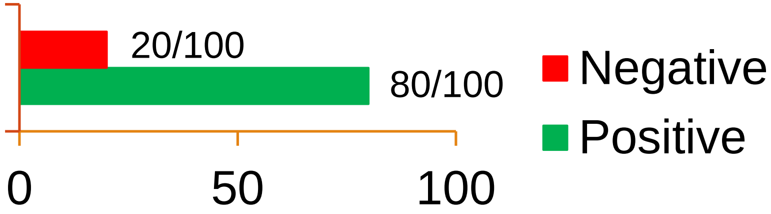


Definition: Class Ratio Estimation

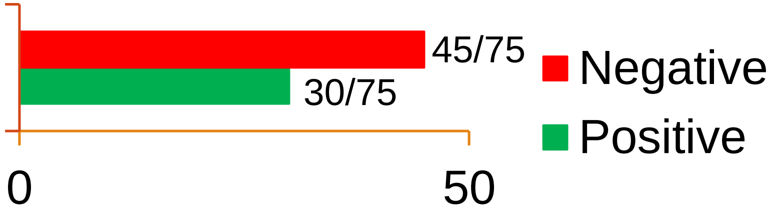
- Estimate fraction of instances belonging to each class in unlabelled set
  - Need **not** estimate per-instance labels
- Pose as supervised Learning problem
  - Labelled training instances

A key issue

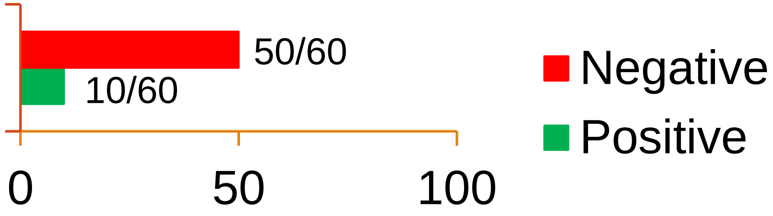

50/60
10/60
0
50
100
Negative
Positive

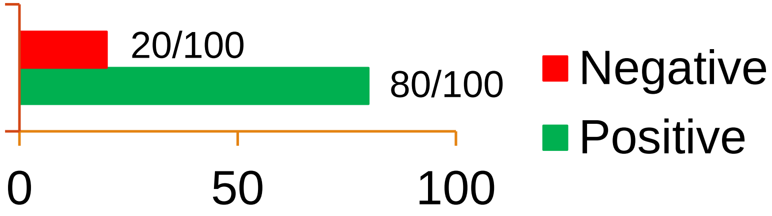

20/100
80/100
0
50
100
Negative
Positive


House of Pizza
HOUSE OF PIZZA

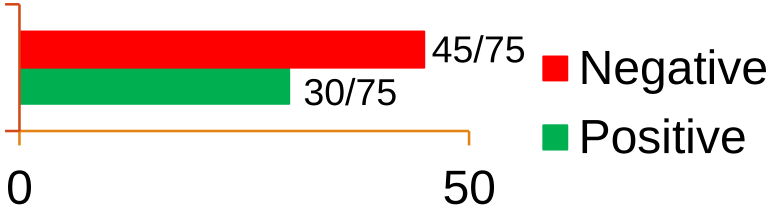

45/75
30/75
0
50
Negative
Positive

A key issue





- Training, test distr. may be different
  - Class ratios vary
  - Class-conditionals are same



Existing methods

- Multi-class classification (Baseline)
  - Optimized for instance level accuracy
  - Class shift is not well-studied
- Class ratio estimation (train, test class conditionals are same)
  - F-divergence based [PS12]
  - Maximum mean discrepancy [Zh13]
  - No theoretical analysis

Existing methods

- Multi-class classification (Baseline)
  - Optimized for instance level accuracy
  - Class shift is not well-studied
- Class ratio estimation (train, test class conditionals are same)
  - F-divergence based [PS12]
  - Maximum mean discrepancy [Zh13]
  - No theoretical analysis

Notation

- Given:
  - Labelled training set $L = \{(x_1, y_1), \dots, (x_l, y_l)\}, \quad y_i \in \{1, \dots, c\}$.
  - Unlabelled set $U = \{z_1, \dots, z_u\}$
  - Universal Kernel k, its feature map ϕ, and its RKHS H
- Goal: Find fraction of each class in U
  - i.e., find $\theta_1, \dots, \theta_c$

Notation

- Given:
  - Labelled training set $L = \{(x_1, y_1), \dots, (x_l, y_l)\},\ y_i \in \{1, \dots, c\}$.
  - Unlabelled set $U = \{z_1, \dots, z_u\}$
  - Universal Kernel k, its feature map ϕ, and its RKHS H
- Goal: Find fraction of each class in U
  - i.e., find $\theta_1, \dots, \theta_c$
- Key assumption: $P^L_{X/Y} = P^U_{X/Y}$
  - P^U_Y need not be P^L_Y
  - P^U_Y may be de-generate!

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} P_Y^U(i) P_{X/Y}^U(x/i)$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^U(x/i)$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^{L}(x/i)$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^{\boldsymbol{L}}(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

$$\min_{\theta \in \Delta_c} \left\| \mathrm{E}_{P_X^U}[\phi(X)] - \sum_{i=1}^{c} \theta_i \, \mathrm{E}_{P_{\frac{X}{Y}}^L}[\phi(X)/i] \right\|_H^2$$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

$$\approx \min_{\theta \in \Delta_c} \left\| \frac{1}{u} \sum_{j=1}^{u} \phi(z_j) - \sum_{i=1}^{c} \theta_i \left(\frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j) \right) \right\|_2^2$$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

$$\approx \min_{\theta \in \Delta_c} \left\| \frac{1}{u} \sum_{j=1}^{u} \phi(z_j) - \sum_{i=1}^{c} \theta_i \left(\frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j) \right) \right\|_2^2$$

Simple convex QP

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

Consistency?

$$\approx \min_{\theta \in \Delta_c} \left\| \frac{1}{u} \sum_{j=1}^{u} \phi(z_j) - \sum_{i=1}^{c} \theta_i \left(\frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j) \right) \right\|_2^2$$

MMD based method

- Idea:
  - $P_X^U(x) = \sum_{i=1}^{c} \theta_i P_{X/Y}^L(x/i)$
  - Find θ minimizes dist. between above
    - Use MMD as distance
- Maximum Mean Discrepancy (MMD) [FM53]
  - $MMD(P_1, P_2) \equiv \left\| \mathrm{E}_{P_1}[\phi(X)] - \mathrm{E}_{P_2}[\phi(X)] \right\|_H$, where k is universal

Learning bounds!

$$\approx \min_{\theta \in \Delta_c} \left\| \frac{1}{u} \sum_{j=1}^{u} \phi(z_j) - \sum_{i=1}^{c} \theta_i \left(\frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j) \right) \right\|_2^2$$

Key Contributions

- Theoretical Analysis
  - Derive learning bounds
  - Simple proof
  - Works with de-generate P_Y^U

Key Contributions

- Theoretical Analysis
  - Derive learning bounds
  - Simple proof
  - Works with de-generate P_Y^U
- Hints at right kernel
  - SDP formulation for kernel learning (convex!)
  - Improved generalization

Theorem

$$\hat{\theta} \equiv \operatorname*{argmin}_{\theta \in \Delta_c} \left\| \frac{1}{u} \sum_{j=1}^{u} \phi(z_j) - \sum_{i=1}^{c} \theta_i \left(\frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j) \right) \right\|_2^2$$

Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta \in \Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i} \phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c} \phi(x_j)$$

Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta \in \Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{R^2\left(\frac{c^2+1}{u} + \sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^TA)}$$

Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta \in \Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{R^2\left(\frac{c^2+1}{u} + \sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^TA)}$$

Towards Consistency

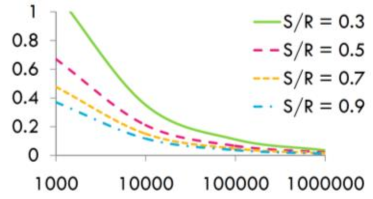
Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta\in\Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \frac{R^2\left(\frac{c^2+1}{u} + \sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^TA)}$$



Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta\in\Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\left\|\hat{\theta}-\theta^*\right\|_2^2 \le \frac{R^2\left(\frac{c^2+1}{u}+\sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^TA)}$$

Higher u is better!

Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta \in \Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{R^2\left(\frac{c^2+1}{u} + \sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^T A)}$$

k determines A, R

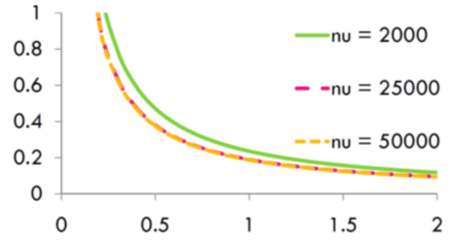
Theorem

$$\hat{\theta}_{1:c-1} \equiv \underset{\theta \in \Lambda_c}{\operatorname{argmin}}\left(h(\theta) \equiv \|A\theta - a\|_2^2\right),$$

$$A^i = \frac{1}{l_i}\sum_{j=1}^{l_i}\phi(x_j) - \frac{1}{l_c}\sum_{j=1}^{l_c}\phi(x_j)$$

If A has full column rank, then with probability atleast $1-\delta$, we have:

$$\|\hat{\theta} - \theta^*\|_2^2 \le \frac{R^2\left(\frac{c^2+1}{u} + \sum_{i=1}^{c}\frac{2}{l_i}\right)\left(1+\sqrt{\log\frac{2}{\delta}}\right)^2}{mineig(A^T A)}$$



Proof sketch

- TST: $\{h(\theta^*) - h(\hat{\theta})\} \xrightarrow{p} 0$, as $l, u \to \infty$
  - $h(\theta^*)$ satisfies bounded difference property
  - Follows from Mc Diarmid's inequality and upper bounding $\mathrm{E}[h(\theta^*)]$

Proof sketch

- TST: $\left\{h(\theta^*) - h(\hat{\theta})\right\} \xrightarrow{p} 0$, as $l, u \to \infty$
  - $h(\theta^*)$ satisfies bounded difference property
  - Follows from Mc Diarmid's inequality and upper bounding $\mathrm{E}[h(\theta^*)]$

- TST: $\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \frac{h(\theta^*) - h(\hat{\theta})}{mineig(A^T A)}$
  - Optimality conditions at $\hat{\theta}$
  - Elementary properties of quadratic

Kernel Learning

- Pre-processing step (otherwise also possible)
- Given: Universal $k_1, \dots, k_n$
- Goal: optimize $w \geq 0$ for "best" $k = \sum_{i=1}^{n} w_i k_i$

Kernel Learning

- Pre-processing step (otherwise also possible)
- Given: Universal $k_1, \ldots, k_n$
- Goal: optimize $w \geq 0$ for "best" $k = \sum_{i=1}^{n} w_i k_i$

- Two objectives:
  - w that minimizes terms in bound
  - w that minimizes an empirical term

Kernel Learning – bound terms

$$mineig(A^T A) = mineig\left(\sum_{i=1}^{n} w_i \, A_i^T A_i\right)$$

- Maximization of above term is convex
  - Infact, expressible as LMI

Kernel Learning – bound terms

$$mineig(A^T A) = mineig\left(\sum_{i=1}^{n} w_i A_i^T A_i\right)$$

- Maximization of above term is convex
  - Infact, expressible as LMI

$R^2 = \sum_{i=1}^{n} w_i^2 R_i^2 = \|w\|_2{}^2$ (normalized kernels)

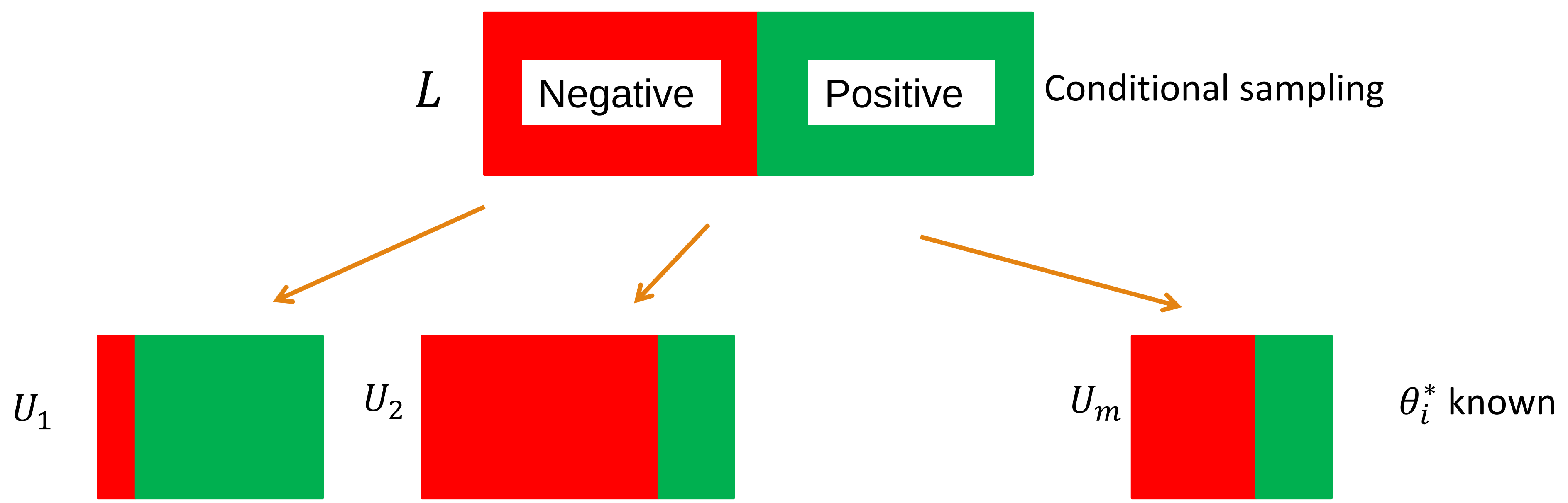
- Minimization of above term is convex

Kernel Learning – empirical term

- Empirical term: w is indeed good for several unlabelled sets
  - Unlabelled sets generated from L

Kernel Learning – empirical term

- Empirical term: w is indeed good for several unlabelled sets
  - Unlabelled sets generated from L



Kernel Learning – empirical term

- Won't work:
  - $\left\|\hat{\theta}_i^w - \theta_i^*\right\| \leq \epsilon \; \forall \; i$
  - $\left|h_i^w\left(\hat{\theta}_i^w\right) - h_i^w(\theta_i^*)\right| \leq \epsilon \; \forall \; i$
  - Both non-convex in w
  - Both do not avoid *extraneous* solutions

Kernel Learning – empirical term

- Won't work:
  - $\left\|\hat{\theta}_i^w - \theta_i^*\right\| \leq \epsilon \ \forall \ i$
  - $\left|h_i^w\left(\hat{\theta}_i^w\right) - h_i^w(\theta_i^*)\right| \leq \epsilon \ \forall \ i$
  - Both non-convex in w
  - Both do not avoid *extraneous* solutions

- Our idea:
  - $h_i^w(\theta) - h_i^w(\theta_i^*) \geq 1 \ \forall \ \|\theta - \theta_i^*\| > \epsilon$

Kernel Learning – empirical term

- Won't work:
  - $\left\|\hat{\theta}_i^w - \theta_i^*\right\| \leq \epsilon \ \forall \ i$
  - $\left|h_i^w\left(\hat{\theta}_i^w\right) - h_i^w(\theta_i^*)\right| \leq \epsilon \ \forall \ i$
  - Both non-convex in w
  - Both do not avoid *extraneous* solutions

- Our idea:
  - $h_i^w(\theta) - h_i^w(\theta_i^*) \geq \rho(\theta, \theta_i^*) - \xi_i \ \forall \ \|\theta - \theta_i^*\| > \epsilon, \xi_i \geq 0$

Kernel Learning – empirical term

- Won’t work:
  - $\left\|\hat{\theta}_i^w - \theta_i^*\right\| \leq \epsilon \ \forall \ i$
  - $\left|h_i^w\left(\hat{\theta}_i^w\right) - h_i^w(\theta_i^*)\right| \leq \epsilon \ \forall \ i$
  - Both non-convex in w
  - Both do not avoid *extraneous* solutions

- Our idea:
  - $w^T u_i \geq \rho(\theta, \theta_i^*) - \xi_i \ \forall \ \|\theta - \theta_i^*\| > \epsilon, \xi_i \geq 0$
  - Convex and avoids *extraneous* solutions

SDP formulation for Kernel Learning

$$\min_{w \in \mathrm{R}^n, \xi \in R^m} \quad \|w\|_2 + B\, maxeig\left(-\sum_{i=1}^{n} w_i\, A_i^T A_i\right) + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t.} \quad w^T u_i \geq \rho(\theta, \theta_i^*) - \xi_i \;\forall\, \|\theta - \theta_i^*\| > \epsilon, \xi_i \geq 0$$

SDP formulation for Kernel Learning

Sparsity

$$\min_{w \in \mathrm{R}^n, \xi \in R^m} \quad \|w\|_1 + B\ maxeig\left(-\sum_{i=1}^{n} w_i\, A_i^T A_i\right) + C \sum_{i=1}^{m} \xi_i$$

$$\mathrm{s.t.} \quad w^T u_i \geq \rho(\theta, \theta_i^*) - \xi_i\ \forall\ \|\theta - \theta_i^*\| > \epsilon, \xi_i \geq 0$$

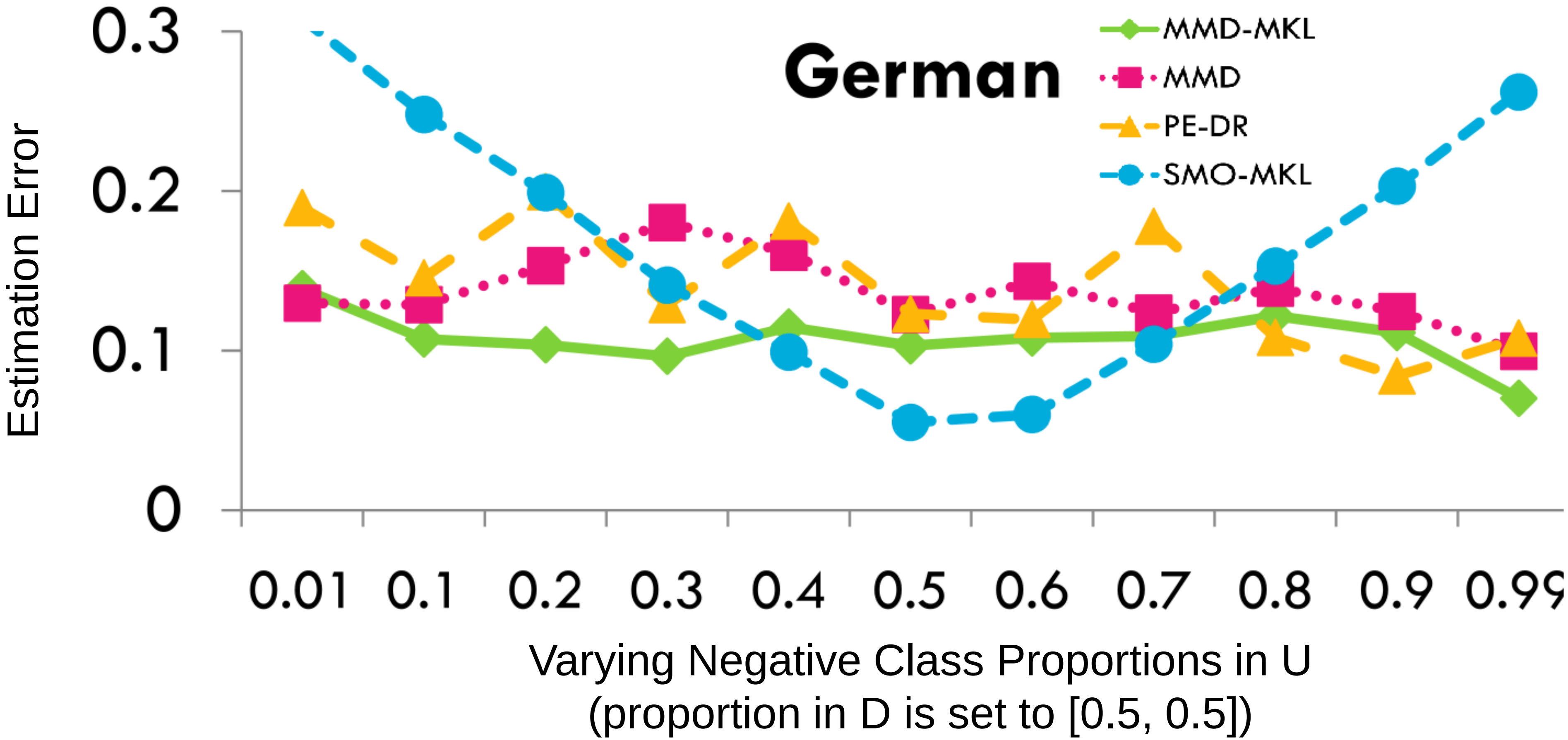
SDP formulation for Kernel Learning

$$\min_{w \in \mathrm{R}^n, \xi \in R^m} \quad \|w\|_1 + B\, maxeig\left(-\sum_{i=1}^{n} w_i\, A_i^T A_i\right) + C \sum_{i=1}^{m} \xi_i$$

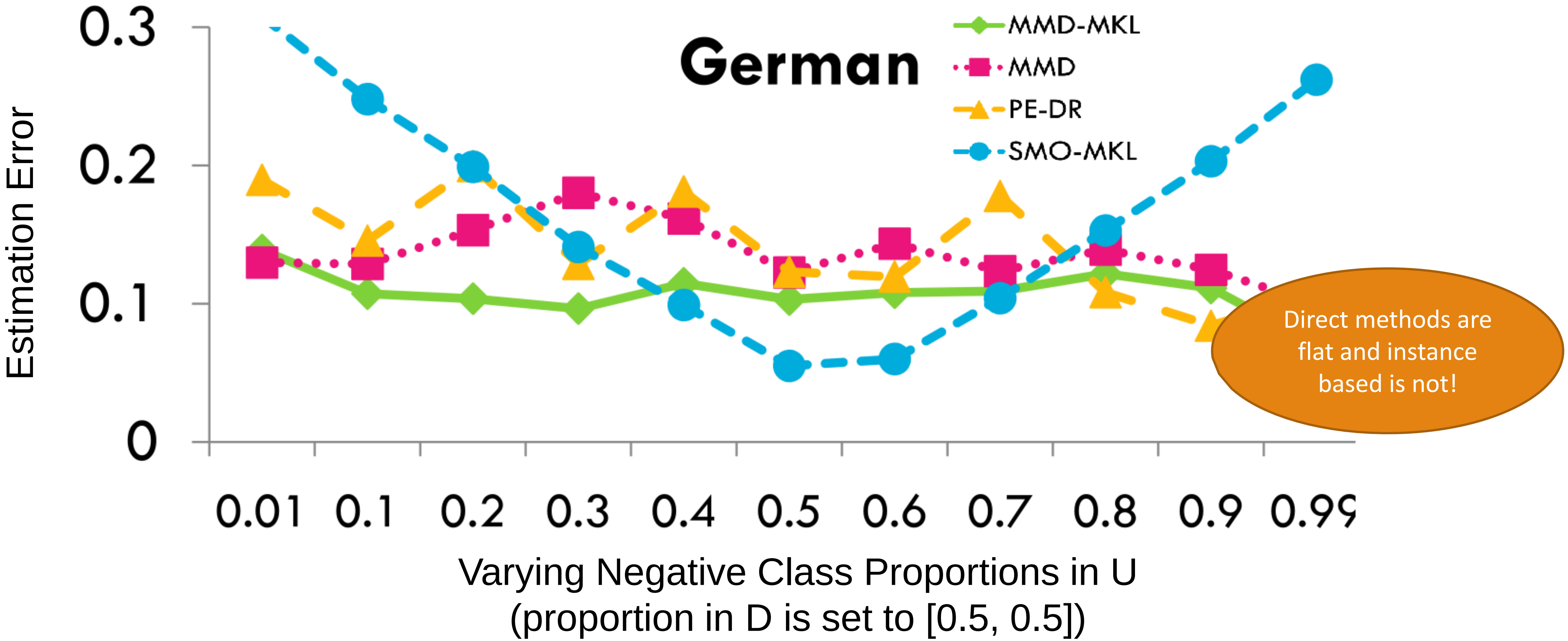
$$\text{s.t.} \quad w^T u_i \geq \rho(\theta, \theta_i^*) - \xi_i\ \forall\ \|\theta - \theta_i^*\| > \epsilon, \xi_i \geq 0$$

Solved using cutting planes algorithm [Ar14]

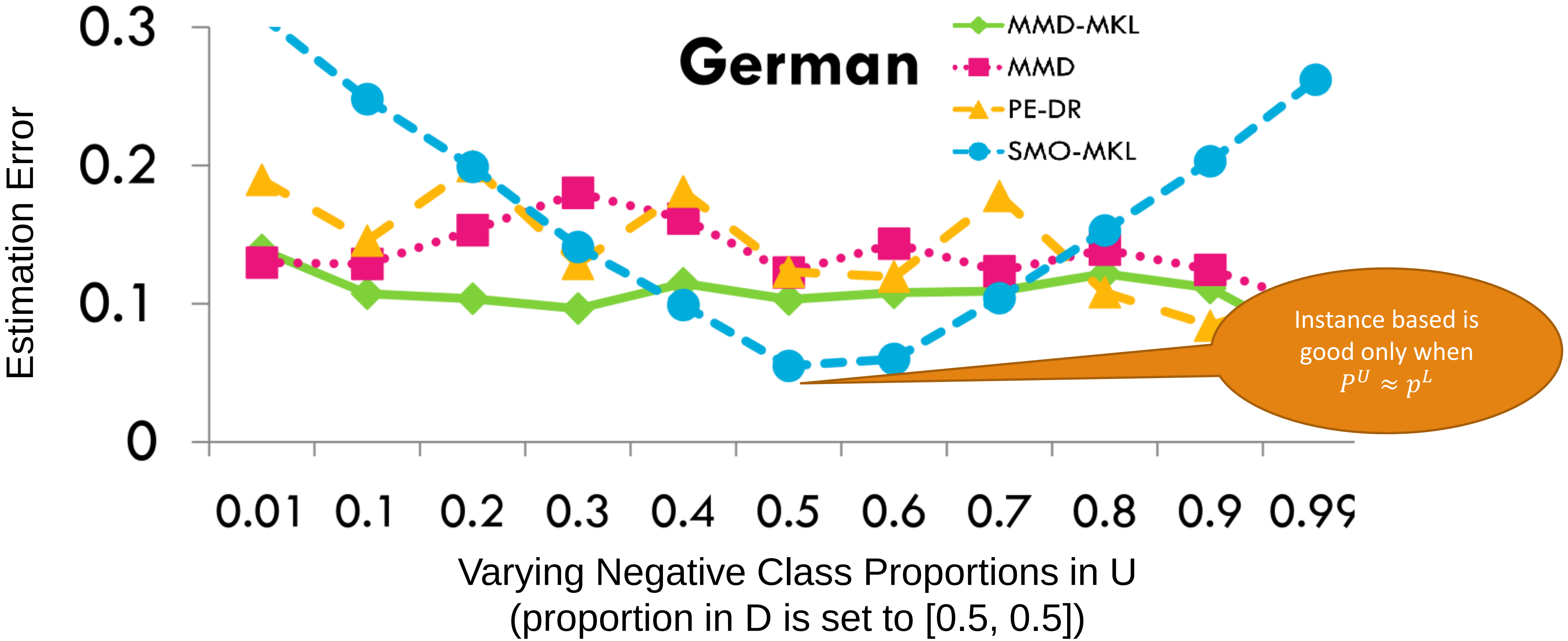
Results: Binary Class Dataset (UCI)



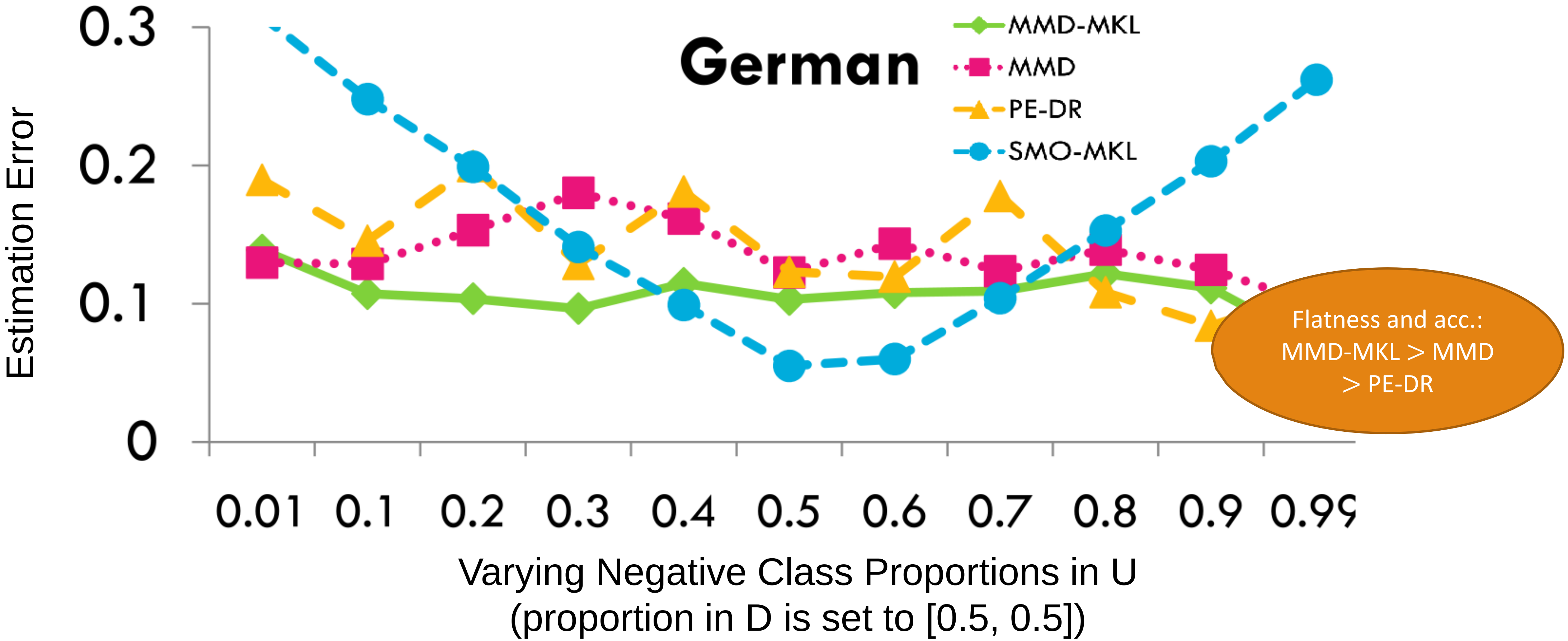
Results: Binary Class Dataset (UCI)



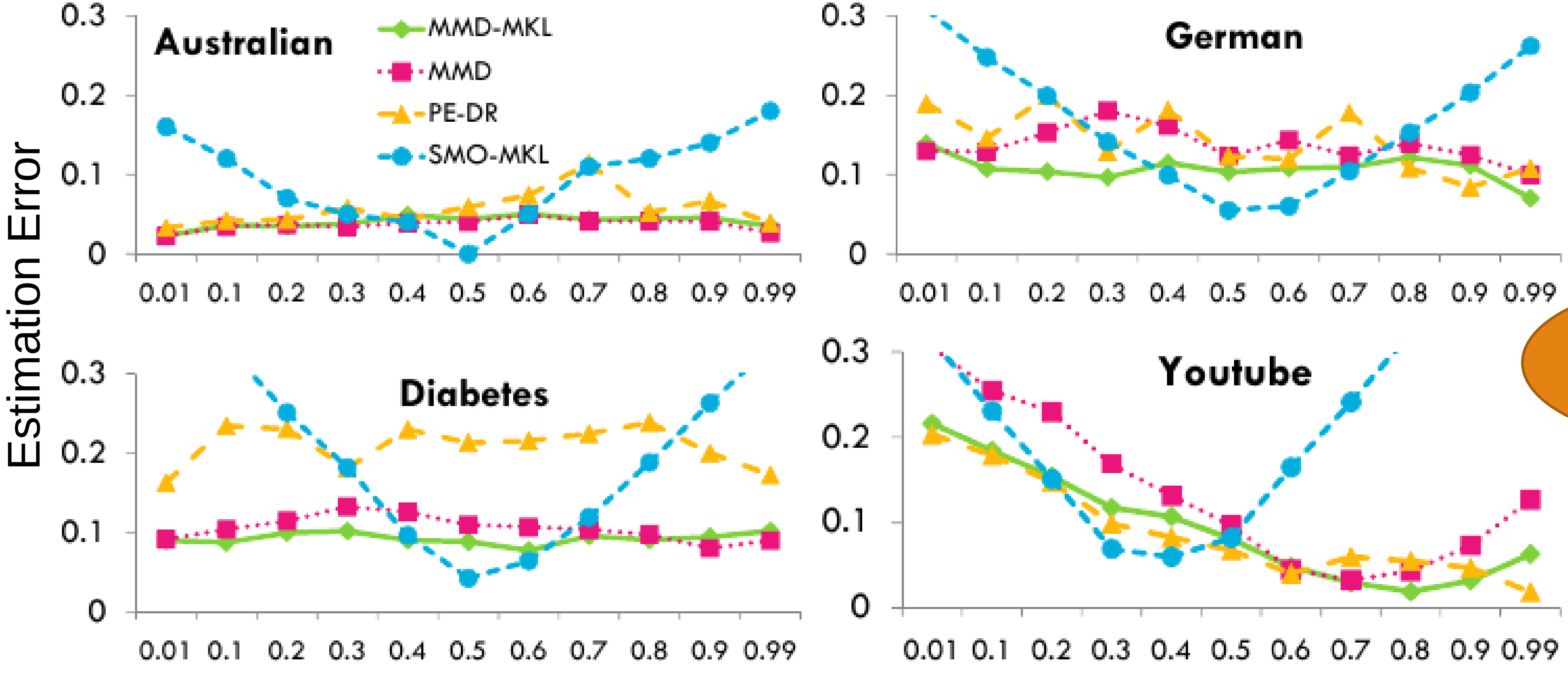
Results: Binary Class Dataset (UCI)



Results: Binary Class Dataset (UCI)



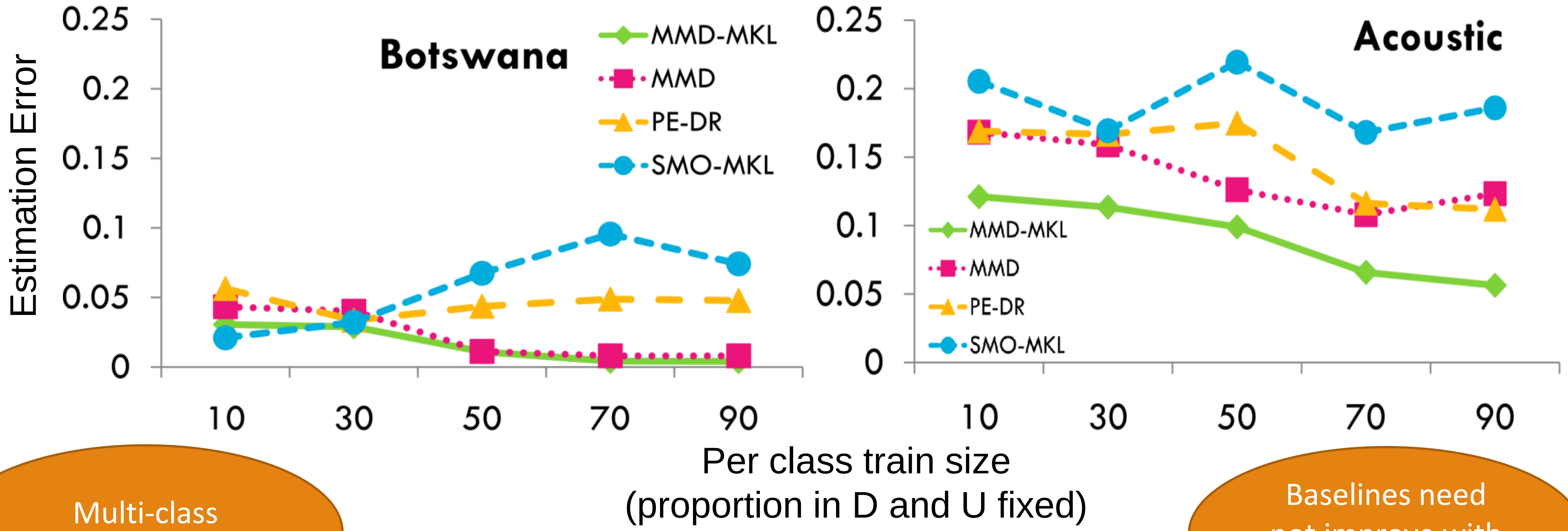
Other binary classification results



Same trend!

Varying Negative Class Proportions in U
(proportion in D is set to [0.5, 0.5])

Variation with data size



Multi-class datasets

Baselines need not improve with data size!

Summary

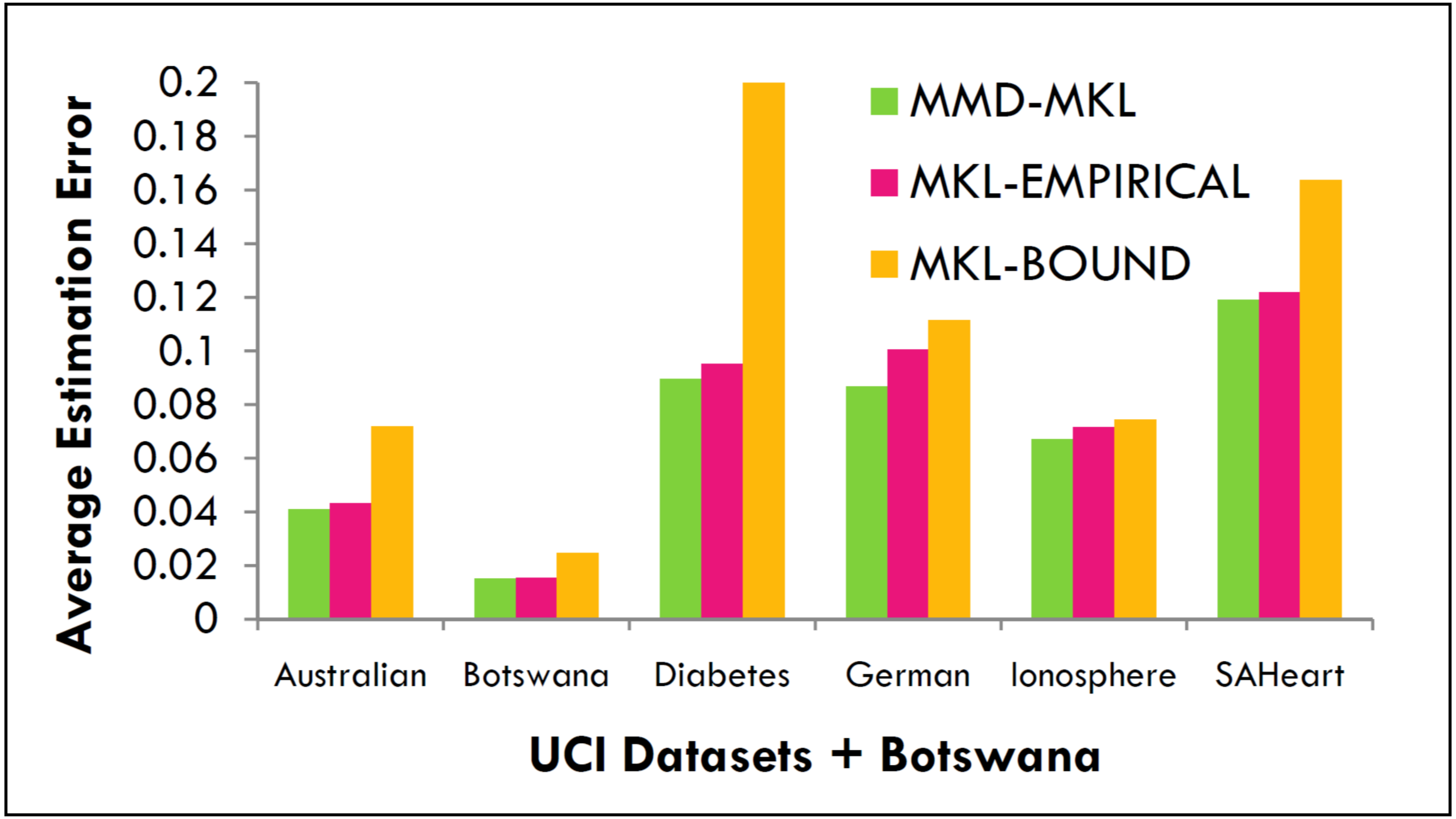
- MMD based estimator for class ratio estimation
- Learning bounds for it
- Bounds provide insight for kernel learning
- SDP formulation for kernel learning
- MMD+MKL improves state-of-the-art
  - Upto 60% overall
  - Upto 40% because of kernel learning

Thanks for listening.
Questions?

References

- [PS12] Plessis, Marthinus D. and Sugiyama, Masashi. *Semisupervised learning of class balance under class-prior change by distribution matching*. In ICML, 2012.
- [Zh13] Zhang, Kun, Scholkopf, Bernhard, Muandet, Krikamol, and Wang, Zhikun. *Domain adaptation under target and conditional shift*. In ICML, 2013.
- [Ar14] Arun Iyer, J. Saketha Nath, Sunita Sarawagi. *Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection*. In ICML, 2014.

Effect of bound



Kernels