Unsupervised Domain Adaptation Based on Source-guided Discrepancy

23th Sep.

Han Bao (The University of Tokyo / RIKEN AIP)

Research interests

supervised learning + real-world constraints

- Learning theory: how to handle performance metrics for **class-imbalance**

  [Bao & Sugiyama 19] (in submission)

- Reinforcement learning **with low-cost data**

  [WCBTS19] (ICML2019) Imitation Learning from Imperfect Demonstration

- Domain adaptation: how to learn when **training ≠ test** today's topic

  [KCBHSS19] (AAAI2019)
  Unsupervised Domain Adaptation Based on Source-guided Discrepancy

- Weak supervision: how to learn **without labels**

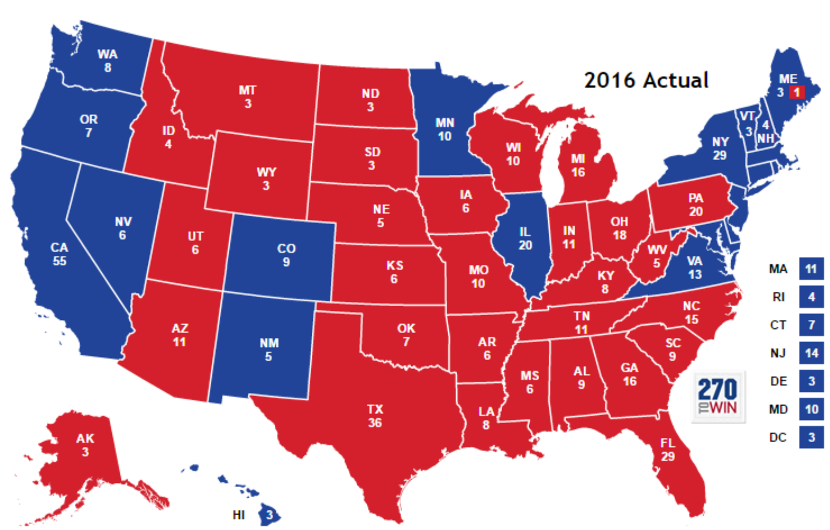
  [BNS18] (ICML2018)
  Classification from Pairwise Similarity and Unlabeled Data

Inference in Real-world

[Brownback & Novotny 2018]

- Prediction of President Election
  - cf. social desirability bias
  - tend to answer in the ways “what others desire”
  - unexpected results in 2016 US president election

Hard to obtain real answers!



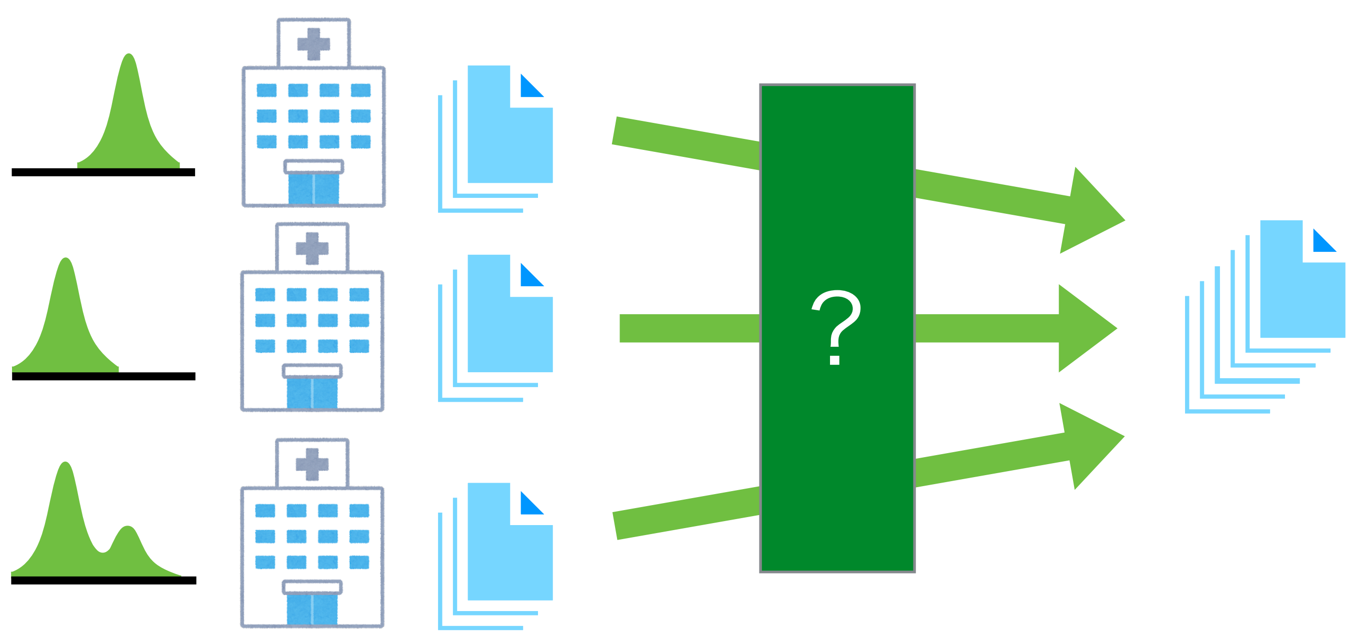
https://www.270towin.com/2016_Election/

Brownback, A., & Novotny, A. (2018). Social desirability bias and polling errors in the 2016 presidential election. *Journal of Behavioral and Experimental Economics*, *74*, 38-56.

Inference in Real-world

- Integration of hospital databases [Wachinger & Reuter 2016]
  - CAD (Computer-Aided Diagnosis) prevailing
  - each hospital has limited amount of data
  - want to unify among hospitals as much as possible

Data distribution may differ!

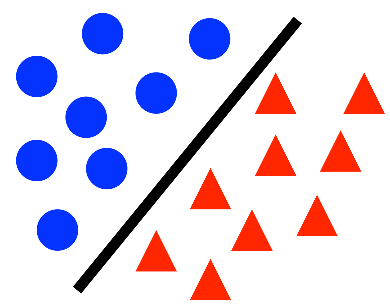


Wachinger, C., & Reuter, M. Alzheimer's Disease Neuroimaging Initiative. (2016). Domain adaptation for Alzheimer's disease diagnostics. *Neuroimage*, *139*, 470-479.

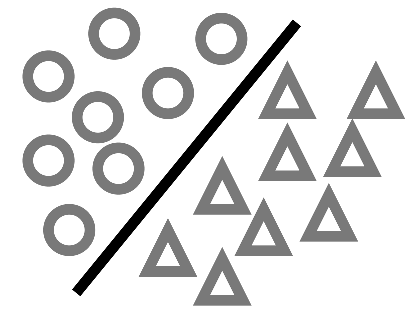
What's transfer learning?

- Usual machine learning

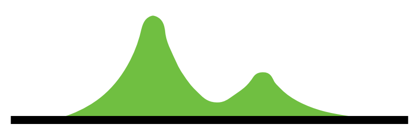
training data



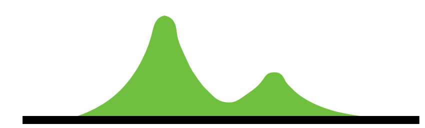
test data



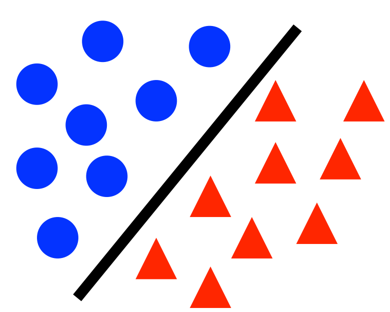
training distribution

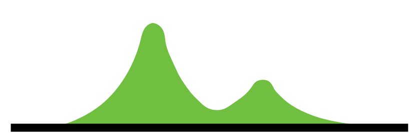


test distribution

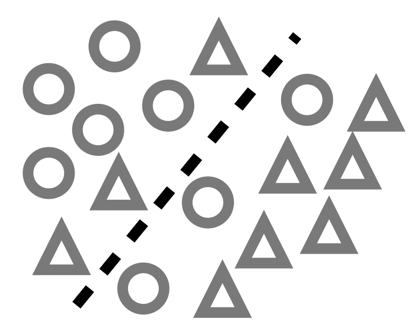


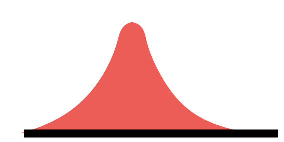
- Transfer learning





$\neq$

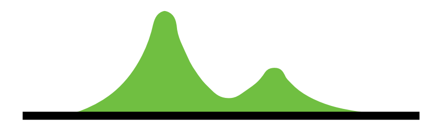




Many terminologies: transfer learning, covariate shift adaptation, domain adaptation, multi-task learning, etc.

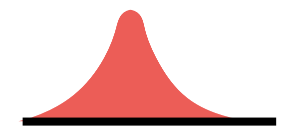
Unsupervised Domain Adaptation

- Input
  - training **labeled** data: $\{x_i, y_i\} \sim p_S$
    (source)



abundant

  - test **unlabeled** data: $\{x'_j\} \sim p_T$
    (target)



scarce

- Goal
  - obtain a predictor that performs well on test data

$$\underset{g}{\operatorname{argmin}} \ \mathrm{Err}_T(g) = \underbrace{\mathbb{E}_T[\ell(Y, g(X))]}_{\text{no access}}$$

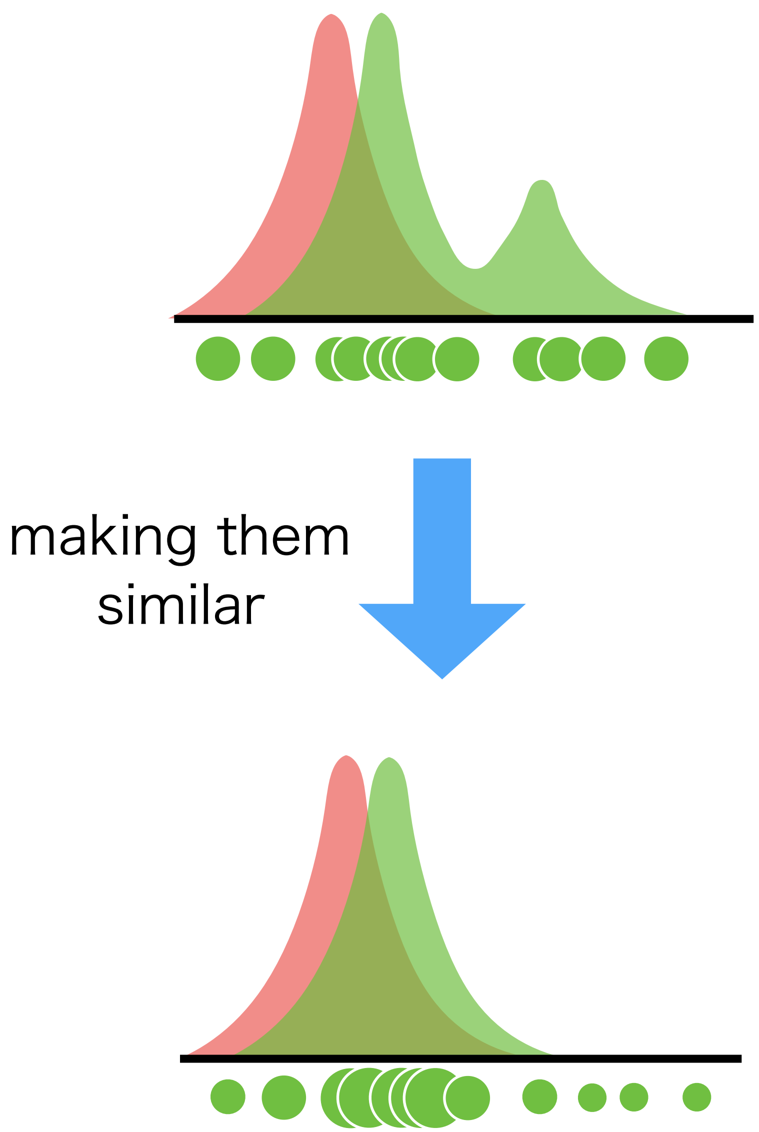
  - Q. How to estimate the target risk?

Outline

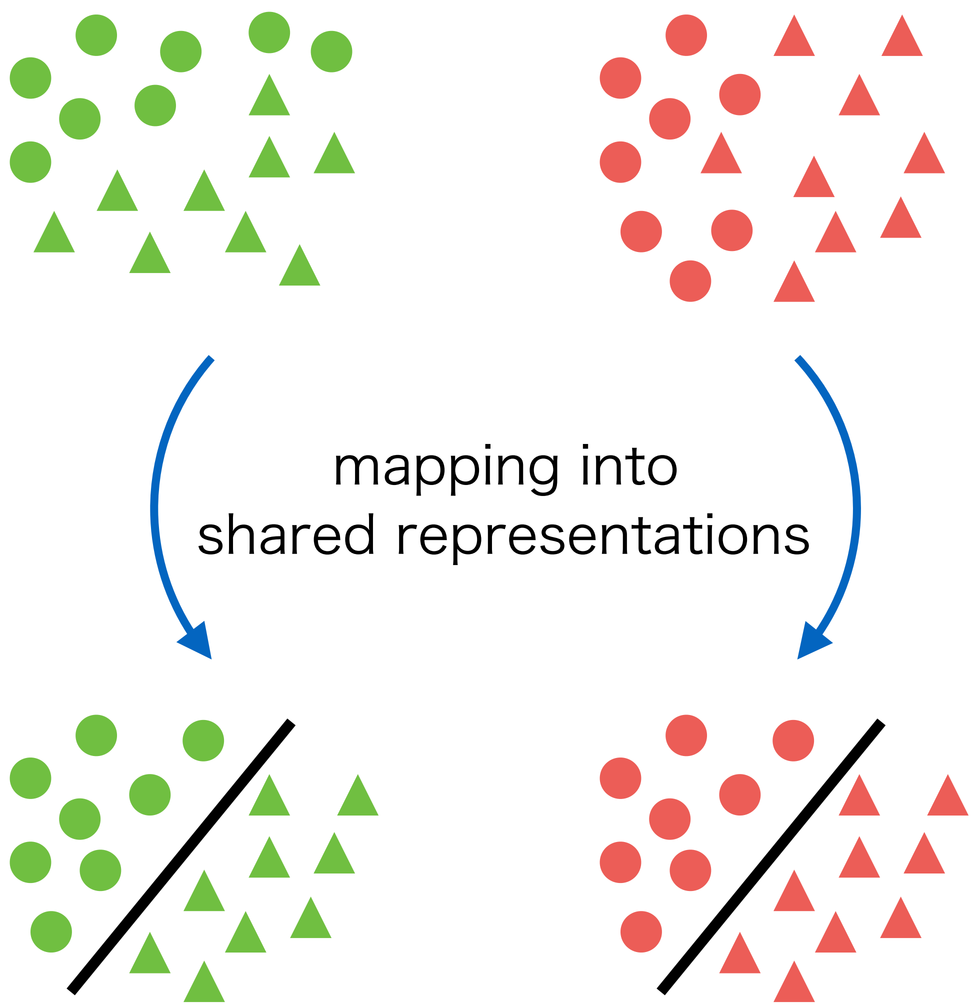
- Introduction — Transfer Learning
- **History/Comparison of Existing Approaches**
- Proposed Method
- Experiments and Future Work

Potential Solutions

- Importance Weighting

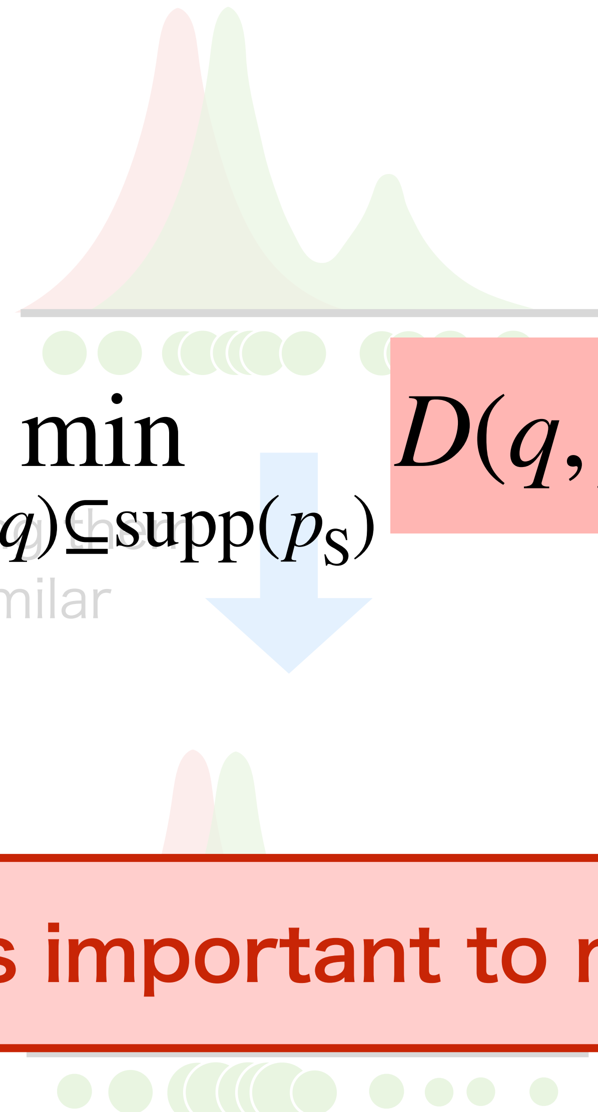


- Representation Learning

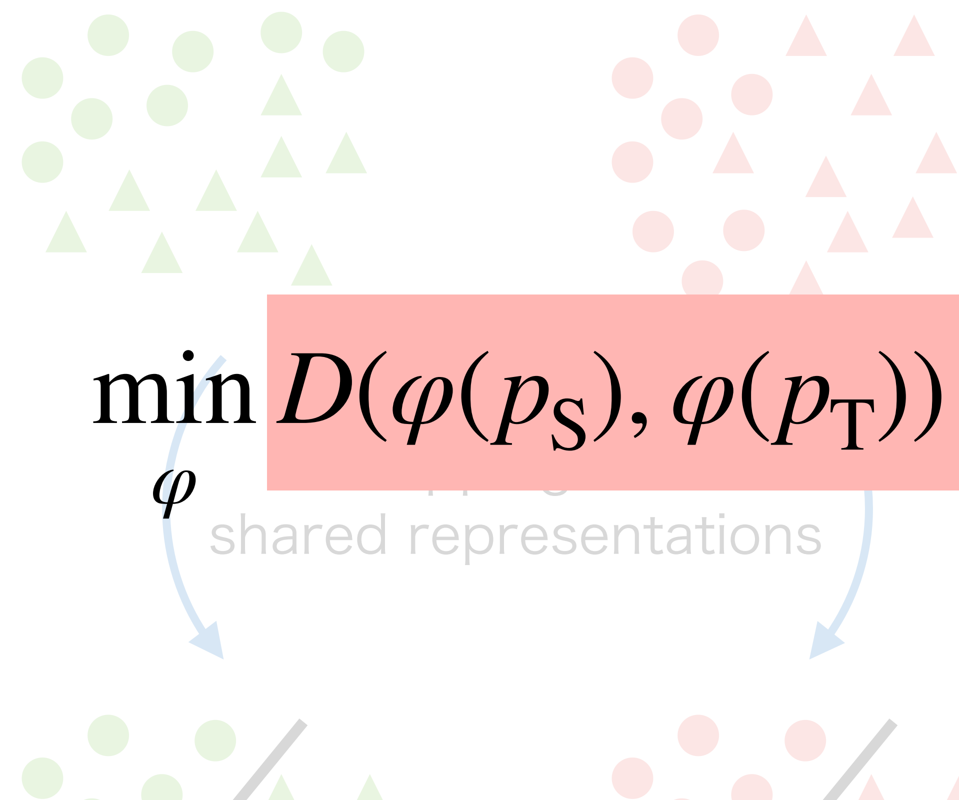


Potential Solutions

- Importance Weighting
- Representation Learning



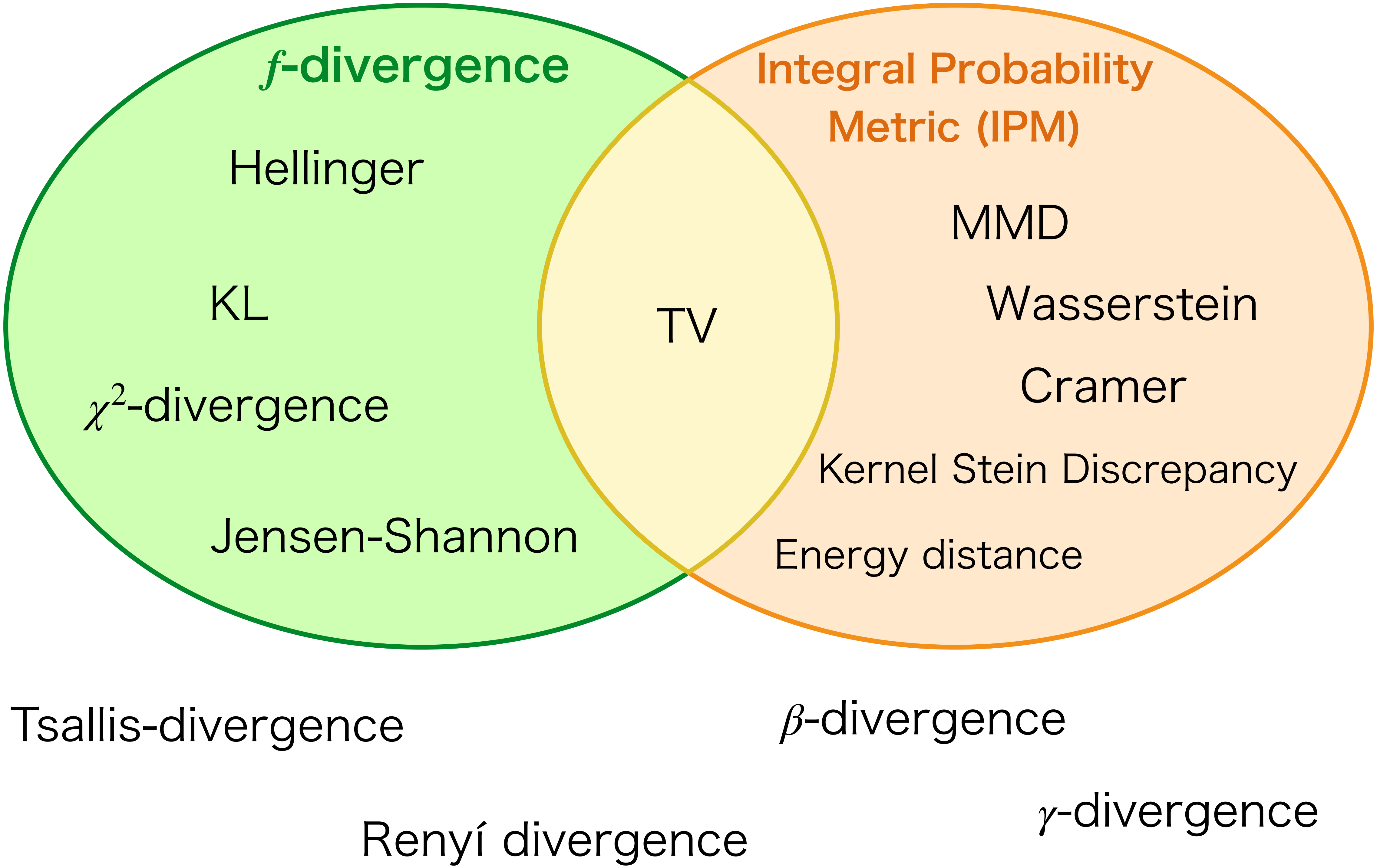
$$\min_{\mathrm{supp}(q)\subseteq \mathrm{supp}(p_\mathrm{S})} D(q, p_\mathrm{T})$$



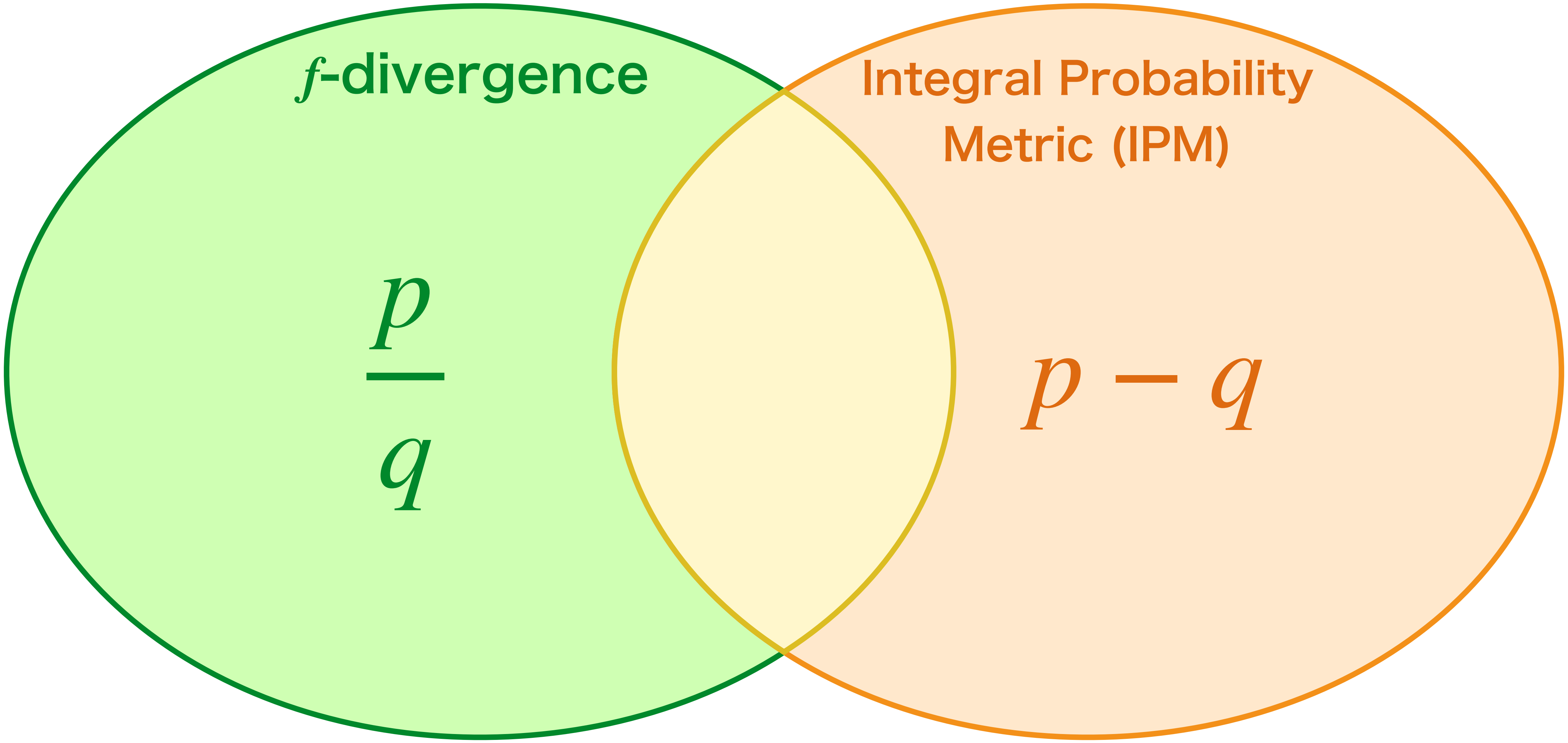
$$\min_{\varphi} D(\varphi(p_\mathrm{S}), \varphi(p_\mathrm{T}))$$

It's important to measure closeness of distributions!

Divergences



Divergences



What is a good measure?

- Postulate: classification risks should be closer if distances between distributions are small

$$\mathrm{Err}_{\mathrm{T}}(g) - \mathrm{Err}_{\mathrm{S}}(g) \leq D(p_{\mathrm{T}}, p_{\mathrm{S}}) + C$$

$$\mathbb{E}_{\mathrm{T}}[\ell(g)] - \mathbb{E}_{\mathrm{S}}[\ell(g)]$$

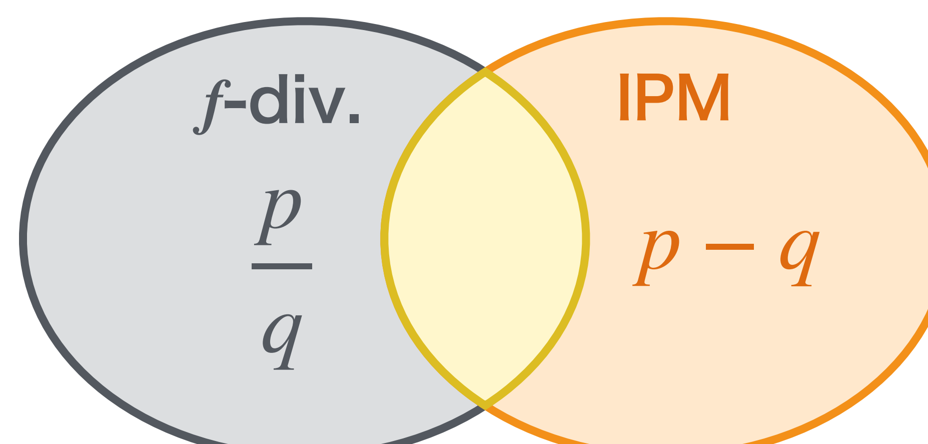
- IPM could be a more suitable family!
  - IPM: $D_\Gamma(p, q) = \sup_{\gamma \in \Gamma} \left| \mathbb{E}_p[\gamma] - \mathbb{E}_q[\gamma] \right|$
  - represented in **difference of expectations**

Γ : real-valued function class
(e.g. 1-Lipschitz for Wasserstein)

expectation over marginal of source dist.

$$\mathrm{Err}_{\mathrm{S}}[g] = \mathbb{E}_{p_{\mathrm{S}}}[\underbrace{\ell}_{\text{loss func.}}(g(X), \underbrace{f_{\mathrm{S}}}_{\text{labeling func.}}(X))]$$

(parallel notation for target domain as well)



Simple Approach: Total Variation

[Kifer+ VLDB2004]

- Total Variation $D_{\mathrm{TV}}(p,q) = 2 \sup_{A:\mathrm{mes'ble}} \left| p(A) - q(A) \right|$

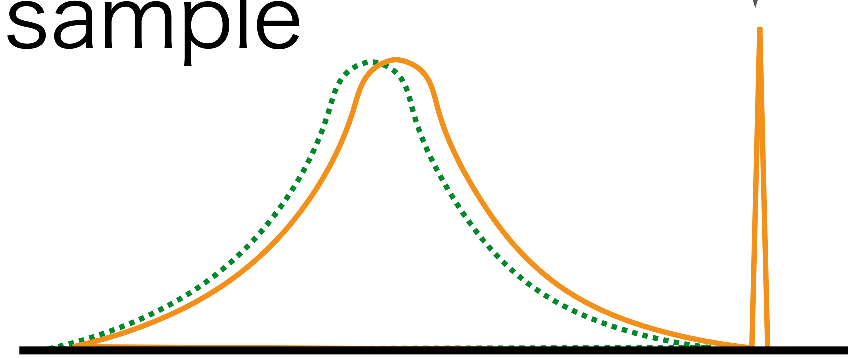
p, q are distributions over $\mathcal{X}$

- classification risk bound

$$\mathrm{Err}_{\mathrm{T}}(g) - \mathrm{Err}_{\mathrm{S}}(g) \leq$$
$$D_{\mathrm{TV}}(p_{\mathrm{S}}, p_{\mathrm{T}}) + \min\{\mathbb{E}_{\mathrm{S}}[\,|f_{\mathrm{S}} - f_{\mathrm{T}}|\,], \mathbb{E}_{\mathrm{T}}[\,|f_{\mathrm{S}} - f_{\mathrm{T}}|\,]\}$$

- Problems
  - TV is overly pessimistic
  - TV is hard estimate within finite sample

we can make a distribution with arbitrarily large TV



Kifer, D., Ben-David, S., & Gehrke, J. (2004, August). Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 180-191). VLDB Endowment.

First Attempt: $\mathcal{H}\Delta\mathcal{H}$-divergence

[Kifer+ VLDB2004; Blitzer+ NeurIPS2008]

Definition ($\mathcal{H}$-divergence)

$$D_{\mathcal{H}}(p,q) = 2\sup_{g\in\mathcal{H}} \left| p(g(X)=1) - q(g(X)=1) \right| \quad ; \mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$$

- $D_{\mathcal{H}}(p,q) \le D_{\mathrm{TV}}(p,q)$ by def. $\Rightarrow$ could be less pessimistic
- estimator $\hat{D}_{\mathcal{H}}(p,q)$ can be computed by ERM in $\mathcal{H}$ (omitted)

Lemma (finite-sample convergence)

Let $d = \mathrm{VCdim}(\mathcal{H})$. Then, with prob. at least $1-\delta$,

$$D_{\mathcal{H}}(p_{\mathrm{S}}, p_{\mathrm{T}}) \le \underbrace{\hat{D}_{\mathcal{H}}(p_{\mathrm{S}}, p_{\mathrm{T}})} + \tilde{O}_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right)$$

empirical estimator

$\hat{D}_{\mathcal{H}}(p_{\mathrm{S}}, p_{\mathrm{T}}) = 2\sup \left| \frac{1}{n_{\mathrm{S}}} \sum_{x\in S} \mathbf{1}_{\{g(x)=1\}} - \frac{1}{n_{\mathrm{T}}} \sum_{x\in T} \mathbf{1}_{\{g(x)=1\}} \right|$

Kifer, D., Ben-David, S., & Gehrke, J. (2004, August). Detecting change in data streams. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 180-191). VLDB Endowment.
Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2008). Learning bounds for domain adaptation. In Advances in neural information processing systems (pp. 129-136).

First Attempt: $\mathscr{H}\Delta\mathscr{H}$-divergence

[Kifer+ VLDB2004; Blitzer+ NeurIPS2008]

Definition (symmetric difference hypothesis $\mathscr{H}\Delta\mathscr{H}$)

$g \in \mathscr{H}\Delta\mathscr{H} \quad \Longleftrightarrow \quad g = h \oplus h'$ for some $h, h' \in \mathscr{H}$ ($\oplus$: XOR)

Theorem (domain adaptation bound)

Let $d = \mathrm{VCdim}(\mathscr{H})$. Then, with prob. at least $1 - \delta$, for any g,

$$\mathrm{Err}_{\mathrm{T}}(g) \leq \mathrm{Err}_{\mathrm{S}}(g) + \frac{1}{2}\hat{D}_{\mathscr{H}\Delta\mathscr{H}}(p_{\mathrm{S}}, p_{\mathrm{T}}) + \tilde{O}_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right) + \lambda$$

where $\lambda = \min_{h\in\mathscr{H}} \mathrm{Err}_{\mathrm{S}}(h) + \mathrm{Err}_{\mathrm{T}}(h)$ (joint minimizer)

Issues

- $\hat{D}_{\mathscr{H}\Delta\mathscr{H}}$ is **intractable**; though $\hat{D}_{\mathscr{H}}$ is tractable
- λ is intrinsically **impossible to estimate**; assume to be small

($\because$ $\mathrm{Err}_{\mathrm{T}}$ cannot be accessed)

Extension: discrepancy measure

[Mansour+ COLT2009]

Definition (discrepancy)

$$D_{\mathrm{disc},\ell}(p,q) = \sup_{g,g'\in\mathcal{H}} \left|\mathrm{Err}_p(g,g') - \mathrm{Err}_q(g,g')\right| \;;\; \mathrm{Err}(g,g') = \int \underbrace{\ell(g(X),g'(X))}_{\text{loss is generalized}} dp$$

- intuition: seeking for potential labelings maximizing diff. of losses
- $\hat{D}_{\mathrm{disc},\ell}$: empirical estimator of $D_{\mathrm{disc},\ell}$; $\hat{D}_{\mathrm{disc},\ell}(p,q) = \sup_{g,g'\in\mathcal{H}} \left|\widehat{\mathrm{Err}}_p(g,g') - \widehat{\mathrm{Err}}_q(g,g')\right|$

Lemma (finite-sample convergence)

Let Rademacher averages of $\mathcal{H}$ on the distribution p_S (p_T resp.) are bounded by $O_p(n_\mathrm{S}^{-1/2})$ ($O_p(n_\mathrm{T}^{-1/2})$ resp.). Assume ℓ is Lipschitz cont. Then, with prob. at least $1-\delta$,

$$D_{\mathrm{disc},\ell}(p_\mathrm{S},p_\mathrm{T}) \le \hat{D}_{\mathrm{disc},\ell}(p_\mathrm{S},p_\mathrm{T}) + O_p\left(\frac{1}{\sqrt{\min\{n_\mathrm{S},n_\mathrm{T}\}}}\right)$$

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of Computational Learning Theory.*

Extension: discrepancy measure

[Mansour+ COLT2009]

Theorem (domain adaptation bound)

Let Rademacher averages of $\mathscr{H}$ on the distribution p_{S} (p_{T} resp.) are bounded by $O_p(n_{\mathrm{S}}^{-1/2})$ ($O_p(n_{\mathrm{T}}^{-1/2})$ resp.). Assume $\mathscr{H}$ is symmetric. Then, with prob. at least $1-\delta$, for any g,

$$\mathrm{Err}_{\mathrm{T}}(g, f_{\mathrm{T}}) - \underbrace{\mathrm{Err}^*_{\mathrm{T}}}_{=\mathrm{Err}_{\mathrm{T}}(g^*_{\mathrm{T}}, f_{\mathrm{T}})} \leq \widehat{\mathrm{Err}}_{\mathrm{S}}(g, g^*_{\mathrm{S}}) + \hat{D}_{\mathrm{disc},01}(p_{\mathrm{S}}, p_{\mathrm{T}}) + O_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right) + \lambda$$

where $\lambda = \mathrm{Err}_{\mathrm{T}}(g^*_{\mathrm{S}}, g^*_{\mathrm{T}})$ (joint minimizer)

Issues

- $\hat{D}_{\mathrm{disc},\ell}$ is generally **intractable**; needs joint sup of g and g'

  (tractable in simple cases)
- λ is intrinsically **impossible to estimate**; assume to be small

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of Computational Learning Theory.*

Comparison of Existing Measures

Q. Can we construct a tractable/tighter measure?

- hard to estimate
- pessimistic

Total Variation

$D_{\mathscr{H}\Delta\mathscr{H}}$-divergence
[KBG04][BBCP06]

- DA bound
- intractable

discrepancy
[MMR09]

? ? ?

Outline

- Introduction — Transfer Learning
- History/Comparison of Existing Approaches
- **Proposed Method**
- Experiments and Future Work

Proposed: Source-guided Discrepancy

Idea: supremum with one variable should be tractable

Definition (Source-guided Discrepancy)

$$D_{\mathrm{sd},\ell}(p,q) = \sup_{g\in\mathscr{H}} \left| \mathrm{Err}_p(g, g_{\mathrm{S}}^*) - \mathrm{Err}_q(g, g_{\mathrm{S}}^*) \right| \; ; \; \mathrm{Err}(g,g') = \int \ell(g(X), g'(X)) dp$$

fix one function

where $g_{\mathrm{S}}^* = \operatorname{argmin}_{g\in\mathscr{H}} \mathrm{Err}_{\mathrm{S}}(g)$ (source risk minimizer)

cf. (discrepancy)

$$D_{\mathrm{disc},\ell}(p,q) = \sup_{g, g'\in\mathscr{H}} \left| \mathrm{Err}_p(g, g') - \mathrm{Err}_q(g, g') \right|$$

- $D_{\mathrm{sd},\ell}(p,q) \leq D_{\mathrm{disc},\ell}(p,q)$ by definition (S-disc is finer)

S-disc Estimator = ERM

- Consider binary classification (loss function: ℓ_{01})
  - assume $\mathscr{H}$ is symmetric: $g \in \mathscr{H} \implies -g \in \mathscr{H}$

Theorem $\hat{D}_{\mathrm{sd},01}(p_{\mathrm{S}}, p_{\mathrm{T}}) = 1 - \min_{g \in \mathscr{H}} J_{\ell_{01}}(g)$

where $J_\ell(g) = \frac{1}{n_{\mathrm{S}}} \sum_{i=1}^{n_{\mathrm{S}}} \ell(g(x_i^{\mathrm{S}}), \underbrace{g_{\mathrm{S}}^*(x_i^{\mathrm{S}})}_{\text{source: labeled by } g_{\mathrm{S}}^*}) + \frac{1}{n_{\mathrm{T}}} \sum_{j=1}^{n_{\mathrm{T}}} \ell(g(x_j^{\mathrm{T}}), \underbrace{-g_{\mathrm{S}}^*(x_j^{\mathrm{T}})}_{\text{target: labeled by } -g_{\mathrm{S}}^*})$ (cost-sensitive risk)

- Estimation Algorithm
  - train a classifier only using source (g_{S}^*)
  - minimize cost-sensitive risk J_ℓ

Similar idea to $\mathscr{H}$-divergence, but we don't need to use $\mathscr{H}\Delta\mathscr{H}$

Finite-Sample Consistency

Theorem

Let Rademacher averages of $\mathcal{H} \otimes \mathcal{H}$ on the distribution p_{S} (p_{T} resp.) are bounded by $O_p(n_{\mathrm{S}}^{-1/2})$ ($O_p(n_{\mathrm{T}}^{-1/2})$ resp.). Then, with prob. at least $1 - \delta$,

$$D_{\mathrm{sd},\ell}(p_{\mathrm{S}}, p_{\mathrm{T}}) \leq \hat{D}_{\mathrm{sd},\ell}(p_{\mathrm{S}}, p_{\mathrm{T}}) + O_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right)$$

- empirical S-disc is tractable
- consistent (as well as $D_{\mathcal{H}}$, D_{disc})

- $\mathcal{H} \otimes \mathcal{H} = \{g \cdot g' \mid g, g' \in \mathcal{H}\}$
- $\mathrm{Rad}(\mathcal{H}) = O_p(n^{-1/2}) \implies \mathrm{Rad}(\mathcal{H} \otimes \mathcal{H}) = O_p(n^{-1/2})$

Domain Adaptation Bound

Theorem (domain adaptation bound)

Let Rademacher averages of $\mathscr{H} \otimes \mathscr{H}$ on the distribution p_{S} (p_{T} resp.) are bounded by $O_p(n_{\mathrm{S}}^{-1/2})$ ($O_p(n_{\mathrm{T}}^{-1/2})$ resp.).
Assume the loss ℓ satisfies the triangle inequality.
Then, with prob. at least $1-\delta$, for any g,

$$\mathrm{Err}_{\mathrm{T}}(g, f_{\mathrm{T}}) - \mathrm{Err}_{\mathrm{T}}^{*} \leq \widehat{\mathrm{Err}}_{\mathrm{S}}(g, g_{\mathrm{S}}^{*}) + \hat{D}_{\mathrm{sd},\ell}(p_{\mathrm{S}}, p_{\mathrm{T}}) + O_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right) + \lambda$$

where $\lambda = \mathrm{Err}_{\mathrm{T}}(g_{\mathrm{S}}^{*}, g_{\mathrm{T}}^{*})$ (joint minimizer)

- $\hat{D}_{\mathrm{sd},\ell}$ is tractable
- $D_{\mathrm{sd},\ell} \leq D_{\mathrm{disc},\ell}$ (always tighter bound)
- λ is impossible to estimate

Summary

Source-guided Discrepancy

$$D_{\mathrm{sd},\ell}(p,q) = \sup_{g \in \mathscr{H}} \left| \mathrm{Err}_p(g, g^*_{\mathrm{S}}) - \mathrm{Err}_q(g, g^*_{\mathrm{S}}) \right|$$

fix one function

DA bound

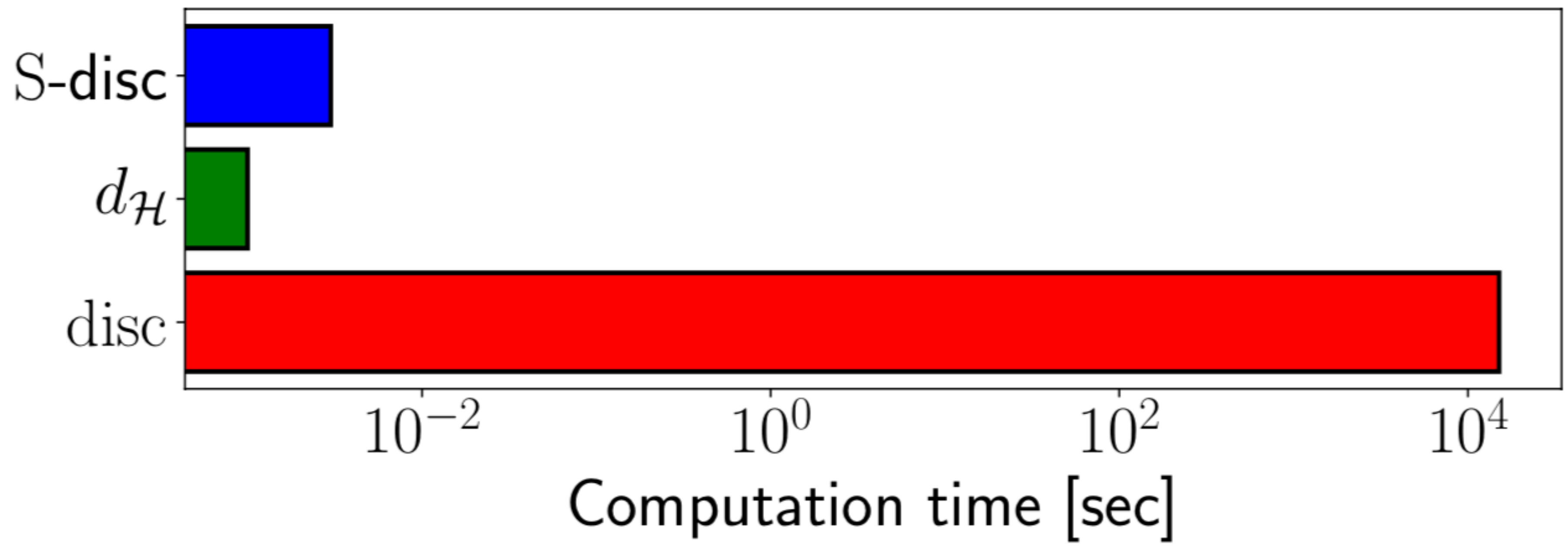
$$\mathrm{Err}_{\mathrm{T}}(g, f_{\mathrm{T}}) - \mathrm{Err}^*_{\mathrm{T}} \leq \widehat{\mathrm{Err}}_{\mathrm{S}}(g, g^*_{\mathrm{S}}) + \hat{D}_{\mathrm{sd},\ell}(p_{\mathrm{S}}, p_{\mathrm{T}}) + O_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right) + \lambda$$

- Tractable estimator: can be computed by ERM
- Tighter measure
- DA bound, but still λ (impossible term) exists

Outline

- Introduction — Transfer Learning
- History/Comparison of Existing Approaches
- Proposed Method
- **Experiments and Future Work**

Computational Time



- $d = 2$, 200 synthetic examples for both source and target
- $d_{\mathscr{H}}$ is an approximator of $D_{\mathscr{H}\Delta\mathscr{H}}$
  - faster, but does not entail DA bound
- discrepancy is computed via approximation
  - resorted to semi-definite relaxation

Source Selection

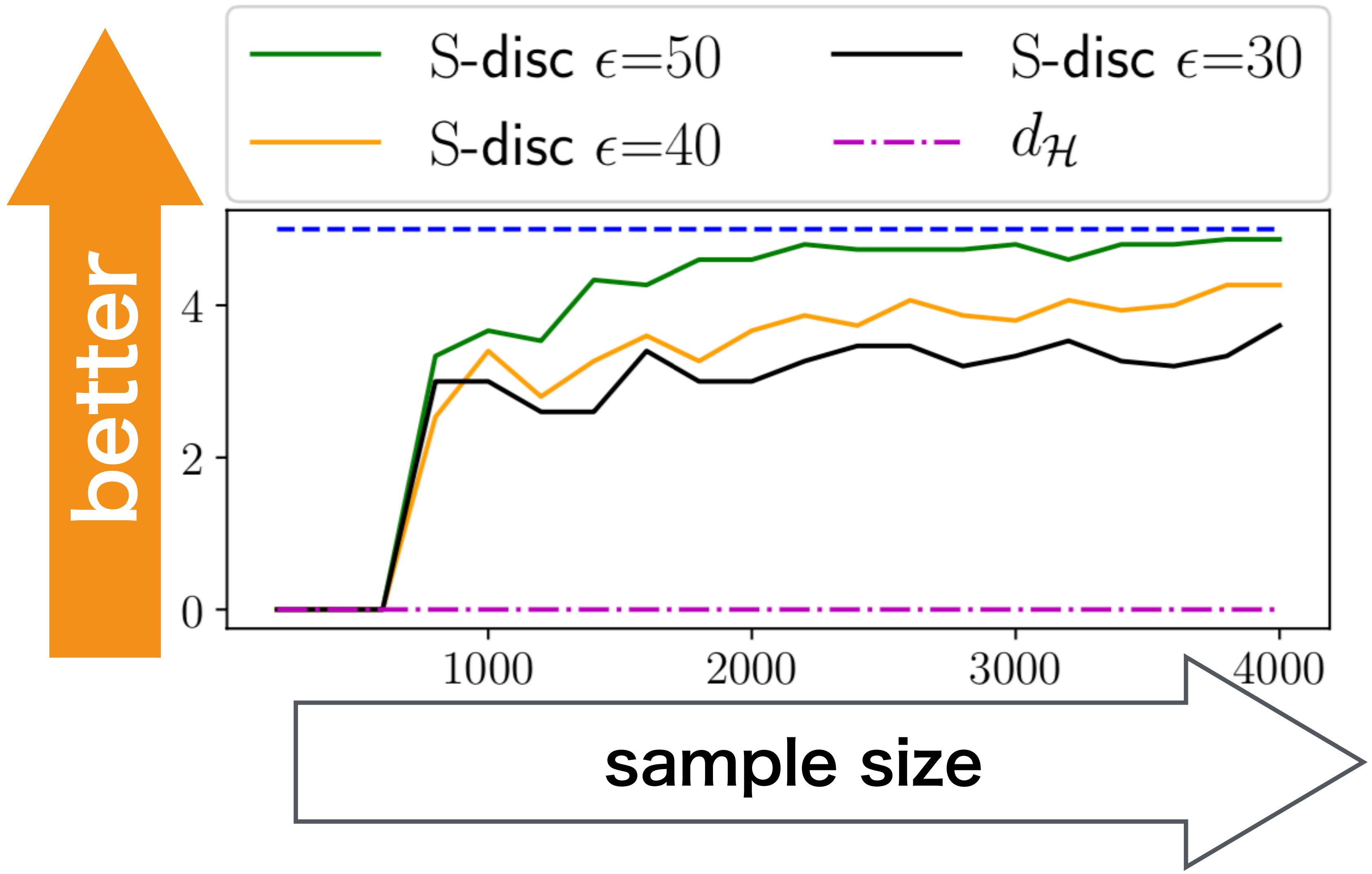
- Domains
  - source: 5 clean MNIST-M, 5 noisy MNIST-M
  - target: MNIST
    (clean MNIST-M is known to be useful for MNIST)
- Setup
  - measure the distance between target and each source
  - sort in ascending order

5 clean MNIST-M should admit smaller distance than noisy ones

MNIST

MNIST-M

Source Selection



Vertical-axis: # of clean MNIST-M domains in top 5

- S-disc successfully capture the difference between clean and noisy MNIST-M

Following Work

[Zhang+ ICML2019]

Source-guided Discrepancy

$$D_{\mathrm{sd},\ell}(p,q) = \sup_{g \in \mathscr{H}} \left| \mathrm{Err}_p(g, g_{\mathrm{S}}^*) - \mathrm{Err}_q(g, g_{\mathrm{S}}^*) \right|$$

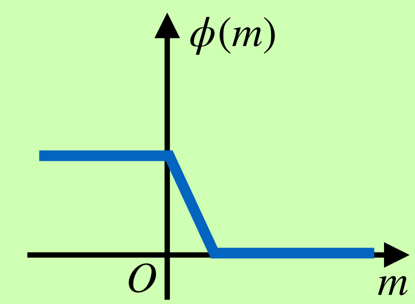
fix source-risk minimizer

Definition: Margin Disparity Discrepancy

②limited to margin loss

$$D_{\mathrm{MDD},f,\ell}(p,q) = \sup_{g \in \mathscr{H}} \left| \mathrm{Err}_p(g, f) - \mathrm{Err}_q(g, f) \right|$$

①fix an arbitrary



DA bound based on MDD

$$\mathrm{Err}_{\mathrm{T}}(g, f_{\mathrm{T}}) \le \widehat{\mathrm{Err}}_{\mathrm{S}}(g, g_{\mathrm{S}}^*) + \hat{D}_{\mathrm{MDD},g,\ell}(p_{\mathrm{S}}, p_{\mathrm{T}}) + O_p\left(\frac{1}{\sqrt{\min\{n_{\mathrm{S}}, n_{\mathrm{T}}\}}}\right) + \lambda$$

⇒ extended to **multi-class (one-vs-all)** case

Zhang, Y., Liu, T., Long, M., & Jordan, M. I. (2019). Bridging Theory and Algorithm for Domain Adaptation. In *ICML,* 2019.

Conclusion

- Discrepancy measure is important in domain adaptation
  - IPM is a nice family; can be connected to DA bound
  - "**fixing one function**" would be a good idea

$$D_{\mathrm{sd},\ell}(p,q) = \sup_{g \in \mathcal{H}} \left| \mathrm{Err}_p(g, g^*_{\mathrm{S}}) - \mathrm{Err}_q(g, g^*_{\mathrm{S}}) \right|$$

fix source-risk minimizer

- Potential directions
  - remove the unestimable term in DA bound (λ)
  - any "optimality" in DA bound?

rethinking DA framework (adaptation algorithms, available supervision) might be needed…