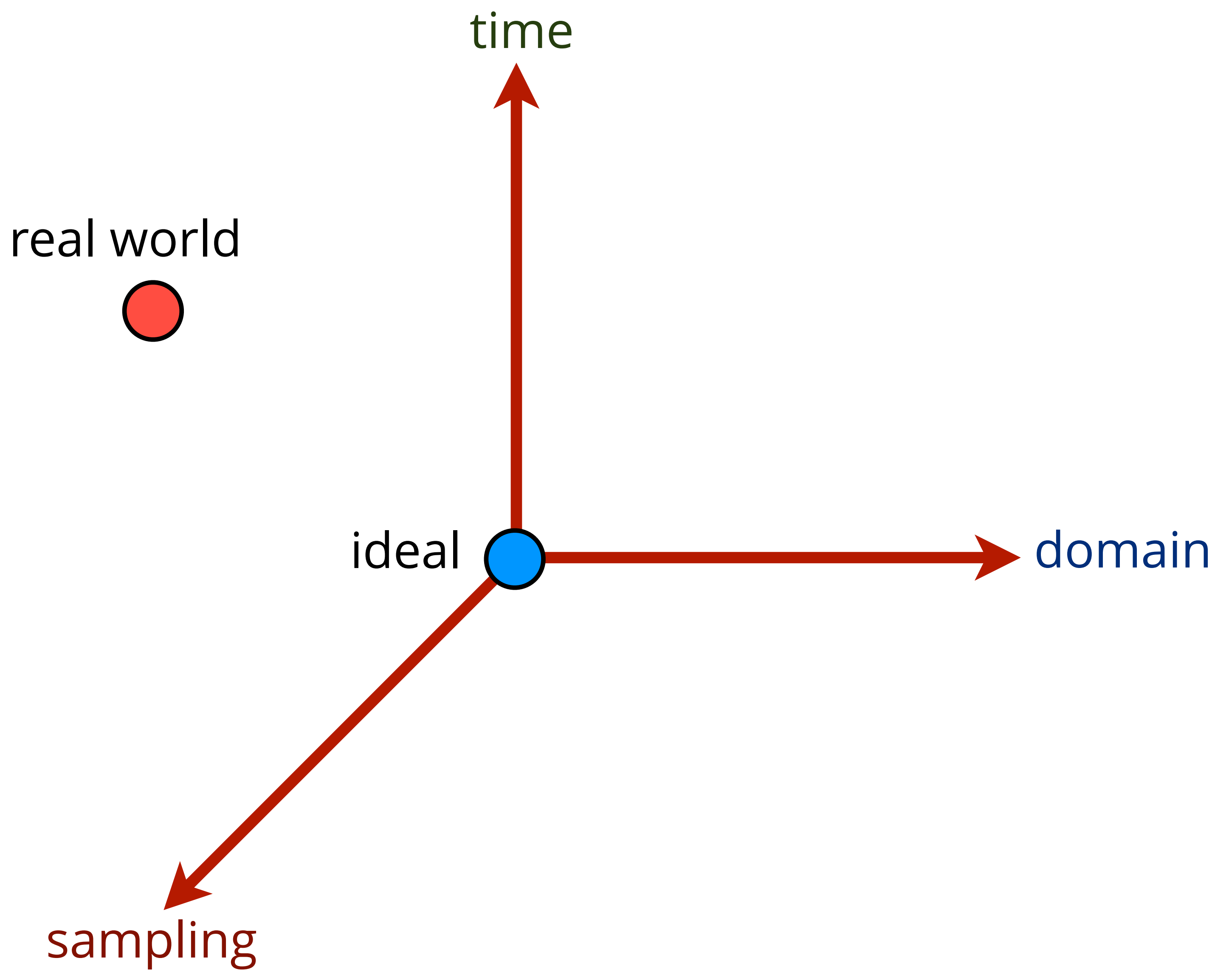
Advanced Machine Learning

Domain Adaptation

MEHRYAR MOHRI MOHRI@

COURANT INSTITUTE & GOOGLE RESEARCH

Non-Ideal World



Outline

- Domain adaptation.
- Multiple-source domain adaptation.

Domain Adaptation

- Sentiment analysis.
- Language modeling, part-of-speech tagging.
- Statistical parsing.
- Speech recognition.
- Computer vision.

→ Solution critical for applications.

This Talk

- Domain adaptation
  - Discrepancy
  - Theoretical guarantees
  - Algorithm
  - Enhancements

Domain Adaptation Problem

- Domains: source (Q, f_Q), target (P, f_P).
- Input:
  - labeled sample S drawn from source.
  - unlabeled sample T drawn from target.
- Problem: find hypothesis h in H with small expected loss with respect to target domain, that is

$$\mathcal{L}_P(h, f_P) = \mathop{\mathrm{E}}_{x \sim P}\Big[L\big(h(x), f_P(x)\big)\Big].$$

Sample Bias Correction Pb

- Problem: special case of domain adaptation with
  - $f_Q = f_P$.
  - $\text{supp}(Q) \subseteq \text{supp}(P)$.

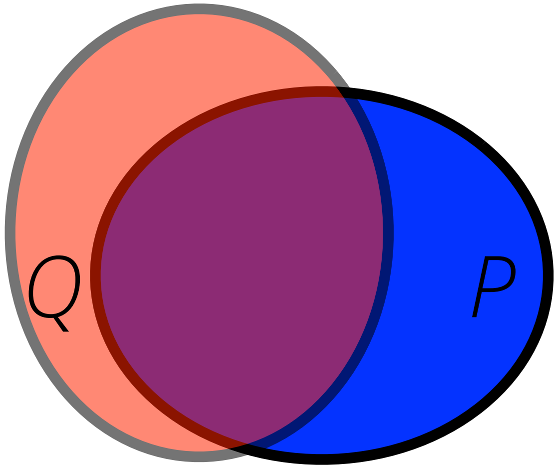
Related Work in Theory

- Single-source adaptation:
  - relation between adaptation and the d_A distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).
  - a few negative examples of adaptation (Ben-David et al. (AISTATS 2010)).
  - analysis and learning guarantees for importance weighting (Cortes, Mansour, and MM (NIPS 2010)).

Related Work in Theory

- Multiple-source:
  - same input distribution, but different labels (Crammer et al. , 2005, 2006).
  - theoretical analysis and method for multiple-source adaptation (Mansour, MM, Rostamizadeh, 2008).

Distribution Mismatch



Which distance should we use
to compare these distributions?

Simple Analysis

- Proposition: assume that the loss L is bounded by M, then

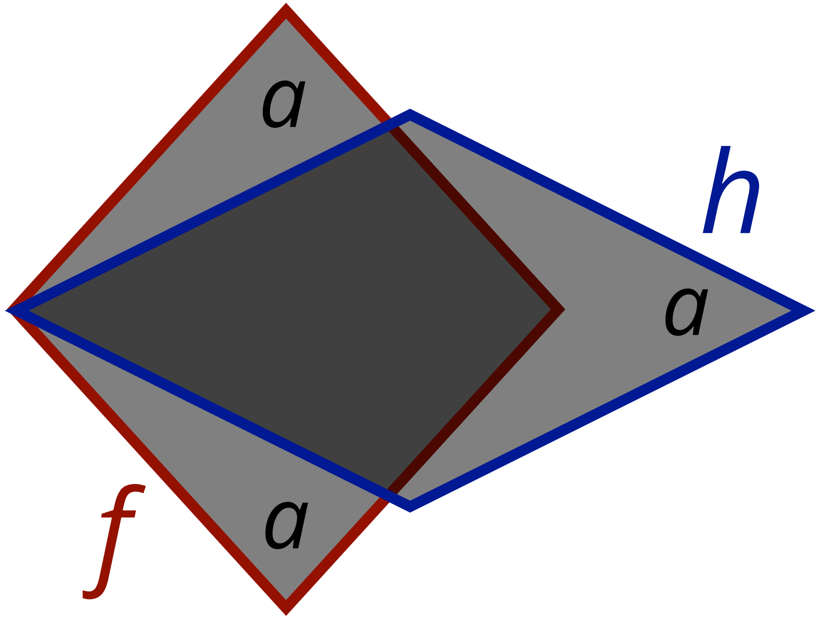
$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M\, L_1(Q, P).$$

- Proof:

$$\begin{aligned}
|\mathcal{L}_P(h, f) - \mathcal{L}_Q(h, f)| &= \Big| \mathop{\mathrm{E}}_{x \sim P}\big[L\big((h(x), f(x)\big)\big] - \mathop{\mathrm{E}}_{x \sim Q}\big[L\big((h(x), f(x)\big)\big] \Big| \\
&= \Big| \sum_x \big(P(x) - Q(x)\big) L\big((h(x), f(x)\big) \Big| \\
&\leq M \sum_x \big|P(x) - Q(x)\big|.
\end{aligned}$$

But, is this bound informative?

Example - Zero-One Loss



$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| = |Q(a) - P(a)|$$

Discrepancy

(Mansour, MM, Rostami, COLT 2009)

- Definition:

$$\mathrm{disc}(P, Q) = \max_{h,h' \in H} \Big| \mathcal{L}_P(h, h') - \mathcal{L}_Q(h, h') \Big|.$$

  - symmetric, triangle inequality, in general not a distance.
  - helps compare distributions for arbitrary losses, e.g. hinge loss, or L_p loss.
  - generalization of d_A distance (Devroye et al. (1996); Kifer et al. (2004); Ben-David et al. (2007)).

Discrepancy - Properties

- Theorem: for L_q loss bounded by M, for any $\delta > 0$, with probability at least $1-\delta$,

$$\operatorname{disc}(P, Q) \leq \operatorname{disc}(\widehat{P}, \widehat{Q}) + 4q\Big(\widehat{\mathfrak{R}}_S(H) + \widehat{\mathfrak{R}}_T(H)\Big) + 3M\left(\sqrt{\frac{\log\frac{4}{\delta}}{2m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2n}}\right).$$

- Proof: Application of McDiarmid's inequality.

Discrepancy = Distance

(Cortes & MM (TCS 2013))

- Theorem: let K be a universal kernel (e.g., Gaussian kernel) and $H = \{h \in \mathbb{H}_K : \|h\|_K \leq \Lambda\}$. Then, for the L_2 loss, discrepancy is a distance over a compact set X.

- Proof: $\Psi : h \mapsto \mathrm{E}_{x \sim P}[h^2(x)] - \mathrm{E}_{x \sim Q}[h^2(x)]$ is continuous for norm $\|\cdot\|_\infty$, thus continuous on $C(X)$.
  - $\mathrm{disc}(P, Q) = 0$ implies $\Psi(h) = 0$ for all $h \in \mathbb{H}$ since:
  $$\forall h, h' \in H, \quad \left| \mathop{\mathbb{E}}_{x \sim P}[(h'(x) - h(x))]^2 - \mathop{\mathbb{E}}_{x \sim Q}[(h'(x) - h(x))]^2 \right| = 0.$$
  - since $\mathbb{H}$ is dense in $C(X)$, $\Psi = 0$ over $C(X)$.
  - thus, $\mathrm{E}_P[f] - \mathrm{E}_Q[f] = 0$ for all $f \geq 0$ in $C(X)$.
  - this implies $P = Q$.

Theoretical Guarantees

- Two types of questions:
  - difference between average loss of hypothesis h on P versus Q?
  - difference of loss (measured on P) between hypothesis h obtained when training on $(\widehat{Q}, f_Q)$ versus hypothesis h' obtained when training on $(\widehat{P}, f_P)$?

Generalization Bound

(Mansour, MM, Rostamizadeh (COLT 2009) + MM addition)

- Notation:
  - $\mathcal{L}_Q(h_Q^*, f_Q) = \min_{h \in H} \mathcal{L}_Q(h, f_Q)$.
  - $\mathcal{L}_P(h_P^*, f_P) = \min_{h \in H} \mathcal{L}_P(h, f_P)$.

- Theorem: assume that L obeys the triangle inequality, then the following holds:

$$\mathcal{L}_P(h, f_P) \leq \min_{h_Q, h_P \in H} \Big\{ \mathcal{L}_Q(h, h_Q) + \mathrm{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \\ + \min\big\{\mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P)\big\} \Big\}.$$

Proof

$$\mathcal{L}_P(h, f_P) \leq \min_{h_P \in H} \Big\{ \mathcal{L}_P(h, h_P) + \mathcal{L}_P(h_P, f_P) \Big\} \qquad \text{(triangle ineq.)}$$

$$\leq \min_{h_P \in H} \Big\{ \mathcal{L}_Q(h, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \Big\}$$

(def. of discrepancy)

$$\leq \min_{h_Q, h_P \in H} \Big\{ \mathcal{L}_Q(h, h_Q) + \mathcal{L}_Q(h_Q, h_P) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) \Big\}.$$

(triangle ineq.)

$$\mathcal{L}_P(h, f_P)$$

$$\leq \min_{h_Q, h_P \in H} \Big\{ \mathcal{L}_Q(h, h_Q) + \text{dis}(P, Q) + \mathcal{L}_P(h_P, f_P) + \min\{\mathcal{L}_Q(h_Q, h_P), \mathcal{L}_P(h_Q, h_P)\} \Big\}.$$

(rerun with the opposite order of min)

Some Natural Cases

- When $h^* = h^*_Q = h^*_P$,

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_Q(h, h^*) + \mathcal{L}_P(h^*, f_P) + \mathrm{disc}(P, Q).$$

- When $f_P \in H$ (consistent case),

$$|\mathcal{L}_P(h, f_P) - \mathcal{L}_Q(h, f_P)| \leq \mathrm{disc}(Q, P).$$

  - Bound of (Ben-David et al., NIPS 2006) or (Blitzer et al., NIPS 2007): always worse in these cases.

Regularized ERM Algorithms

- Objective function:

$$F_{\widehat{Q}}(h) = \lambda \|h\|_K^2 + \widehat{R}_{\widehat{Q}}(h),$$

where K is a PDS kernel;
$\lambda > 0$ is a trade-off parameter; and
$\widehat{R}_{\widehat{Q}}(h)$ is the empirical error of h.

  - broad family of algorithms including SVM, SVR, kernel ridge regression, etc.

Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

- Theorem: let K be a PDS kernel with $K(x,x) \leq R^2$ and L a convex loss function such that $L(\cdot, y)$ is μ-Lipschitz. Let h' be the minimizer of $F_{\widehat{P}}$ and h that of that $F_{\widehat{Q}}$, then, for all $(x,y) \in X \times Y$,

$$\left| L(h'(x), y) - L(h(x), y) \right| \leq \mu R \sqrt{\frac{\operatorname{disc}(\widehat{P}, \widehat{Q}) + \mu\, \eta_H(f_P, f_Q)}{\lambda}},$$

where

$$\eta_H(f_P, f_Q) = \inf_{h \in H} \Big\{ \max_{x \in \operatorname{supp}(\widehat{P})} |f_P(x) - h(x)| + \max_{x \in \operatorname{supp}(\widehat{Q})} |f_Q(x) - h(x)| \Big\}.$$

Proof

- By the property of the minimizers, there exist subgradients such that

$$2\lambda h' = -\delta R_{\widehat{P}}(h')$$

$$2\lambda h = -\delta R_{\widehat{Q}}(h).$$

- Thus,

$$\begin{aligned}
2\lambda \|h' - h\|^2 &= -\langle h' - h, \delta R_{\widehat{P}}(h') - \delta R_{\widehat{Q}}(h) \rangle \\
&= -\langle h' - h, \delta R_{\widehat{P}}(h') \rangle + \langle h' - h, \delta R_{\widehat{Q}}(h) \rangle \\
&\leq R_{\widehat{P}}(h) - R_{\widehat{P}}(h') + R_{\widehat{Q}}(h') - R_{\widehat{Q}}(h) \\
&\leq 2\mathrm{disc}(\widehat{P}, \widehat{Q}) + 2\mu\, \eta_H(f_P, f_Q).
\end{aligned}$$

Proof

- For any hypothesis h_0, we can write:

$$\begin{aligned}2\lambda\|h'-h\|_K^2 \leq\; & \big(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)\big)\\ & + \big(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)\big)\\ & + \big(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)\big) - \big(\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)\big).\end{aligned}$$

- Next, by the Lipschitzness, the following holds:

$$\big(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)\big) \leq 2\mu \operatorname*{E}_{x\sim\widehat{P}}[|f_P(x) - h_0(x)|]$$

$$\big(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)\big) - \big(\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)\big) \leq 2\mu \operatorname*{E}_{x\sim\widehat{Q}}[|f_Q(x) - h_0(x)|].$$

- Since h_0 is in H, we have

$$\big(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)\big) \leq 2\,\mathrm{disc}(\widehat{P}, \widehat{Q}).$$

Guarantees for Reg. ERM

(Cortes & MM (TCS 2013))

- Theorem: let K be a PDS kernel with $K(x,x) \leq R^2$ and L the L_2 loss bounded by M. Then, for all (x,y),

$$\big|L(h'(x),y) - L(h(x),y)\big| \leq \frac{R\sqrt{M}}{\lambda}\Big(\delta + \sqrt{\delta^2 + 4\lambda \mathrm{disc}(\widehat{P},\widehat{Q})}\Big),$$

where $\delta = \min_{h \in H} \Big\| \mathop{\mathrm{E}}_{x \sim \widehat{Q}} \Big[\big(h(x) - f_Q(x)\big)\Phi_K(x)\Big] - \mathop{\mathrm{E}}_{x \sim \widehat{P}} \Big[\big(h(x) - f_P(x)\big)\Phi_K(x)\Big]\Big\|_K.$

- For $f_P = f_Q = f$,
  - $\delta \leq R\epsilon$ if f is ε-close to H on samples.
  - $\delta = 0$ for a Gaussian kernel and f continuous.

Proof

- For any hypothesis h_0, we can write as for previous result:

$$\begin{aligned}2\lambda\|h'-h\|_K^2 \leq \big(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0)\big)\\ + \big(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)\big)\\ + \big(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)\big) - \big(\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)\big).\end{aligned}$$

- Next, for the squared loss, we have:

$$\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0) = \mathop{\mathrm{E}}_{x\sim\widehat{P}}\big[(h_0(x) - f_P(x))(2h(x) - f_P(x) - h_0(x))\big]$$

$$\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{P}}(h', h_0) = \mathop{\mathrm{E}}_{x\sim\widehat{P}}\big[(h_0(x) - f_P(x))(2h'(x) - f_P(x) - h_0(x))\big].$$

- Thus,

$$\begin{aligned}&\big(\mathcal{L}_{\widehat{Q}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, f_Q)\big) - \big(\mathcal{L}_{\widehat{Q}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', f_Q)\big)\\ &\qquad = -2 \mathop{\mathrm{E}}_{x\sim\widehat{Q}}\big[(h_0(x) - f_Q(x))(h(x) - h'(x))\big].\end{aligned}$$

Proof

- As for previous theorem, we have

$$\big(\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{Q}}(h, h_0)\big) - \big(\mathcal{L}_{\widehat{P}}(h', h_0) - \mathcal{L}_{\widehat{Q}}(h', h_0)\big) \leq 2\,\mathrm{disc}(\widehat{P}, \widehat{Q}).$$

- Thus, $2\lambda \|h' - h\|_K^2 \leq 2\,\mathrm{disc}(\widehat{P}, \widehat{Q}) + 2\Delta$ with:

$$\begin{aligned}\Delta &= \Big\langle h - h', \operatorname*{E}_{x \sim \widehat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \operatorname*{E}_{x \sim \widehat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)]\Big\rangle \\ &\leq \|h - h'\|_K \Big\| \operatorname*{E}_{x \sim \widehat{P}}[(h_0(x) - f_P(x))K(x, \cdot)] - \operatorname*{E}_{x \sim \widehat{Q}}[(h_0(x) - f_Q(x))K(x, \cdot)]\Big\|_K.\end{aligned}$$

- The result follows by solving second-degree inequality.

Empirical Discrepancy

- Discrepancy measure $\mathrm{disc}(\widehat{P}, \widehat{Q})$ critical term in bounds.
- Smaller empirical discrepancy guarantees closeness of pointwise losses of *h'* and *h*.
- But, can we further reduce the discrepancy?

Algorithm - Idea

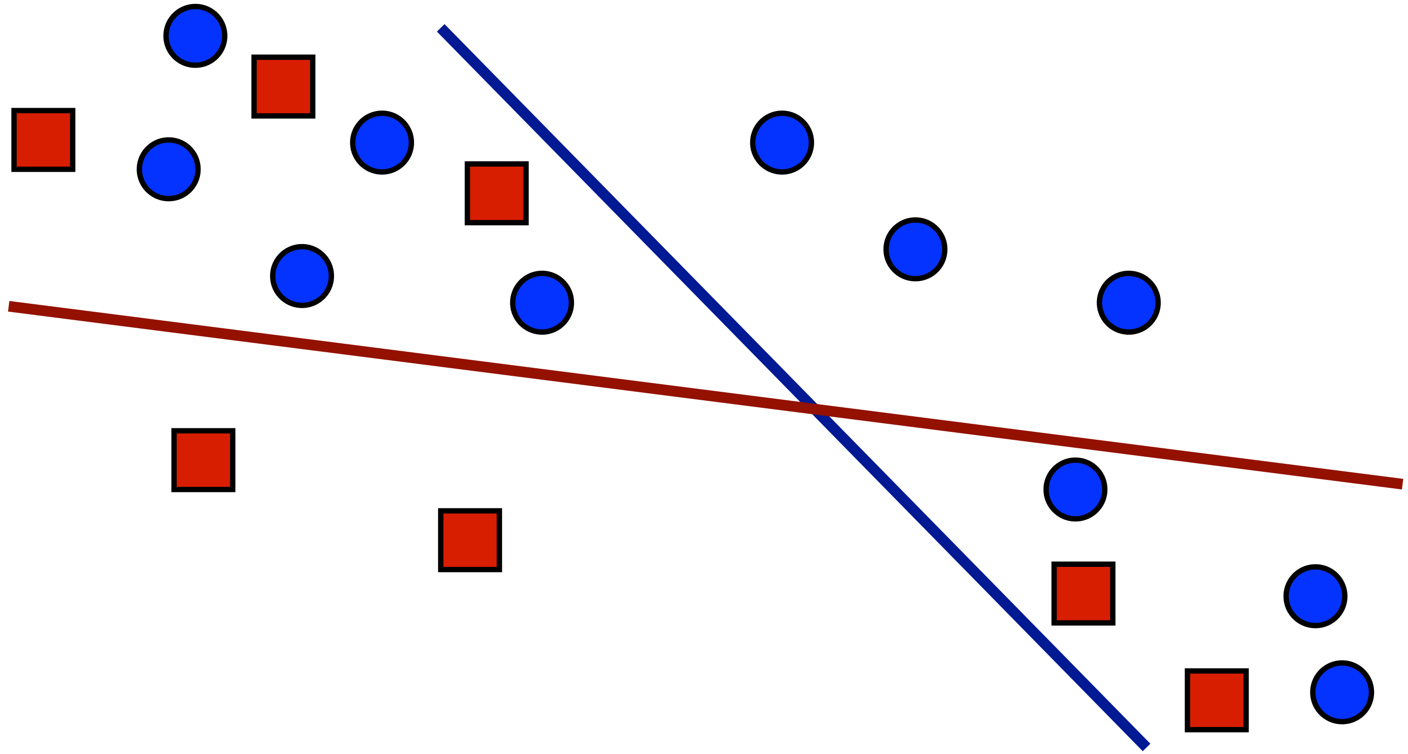
- Search for a new empirical distribution q^* with same support:

$$q^* = \underset{\text{supp}(q) \subseteq \text{supp}(\widehat{Q})}{\text{argmin}} \ \text{disc}(\widehat{P}, q).$$

- Solve modified optimization problem:

$$\min_h F_{q^*}(h) = \sum_{i=1}^{m} q^*(x_i) L(h(x_i), y_i) + \lambda \|h\|_K^2.$$

Case of Halfspaces



Min-Max Problem

- Reformulation:

$$\widehat{Q}' = \operatorname*{argmin}_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} \left| \mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h) \right|.$$

  - game theoretical interpretation.
  - gives lower bound:

$$\max_{h, h' \in H} \min_{\widehat{Q}' \in \mathcal{Q}} \left| \mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h) \right| \leq$$

$$\min_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} \left| \mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h) \right|.$$

Classification - 0/1 Loss

- Problem:

$$\min_{Q'} \max_{a \in H \Delta H} \left| \widehat{Q}'(a) - \widehat{P}(a) \right|$$

$$\text{subject to} \quad \forall x \in S_Q, \widehat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \widehat{Q}'(x) = 1.$$

Classification - 0/1 Loss

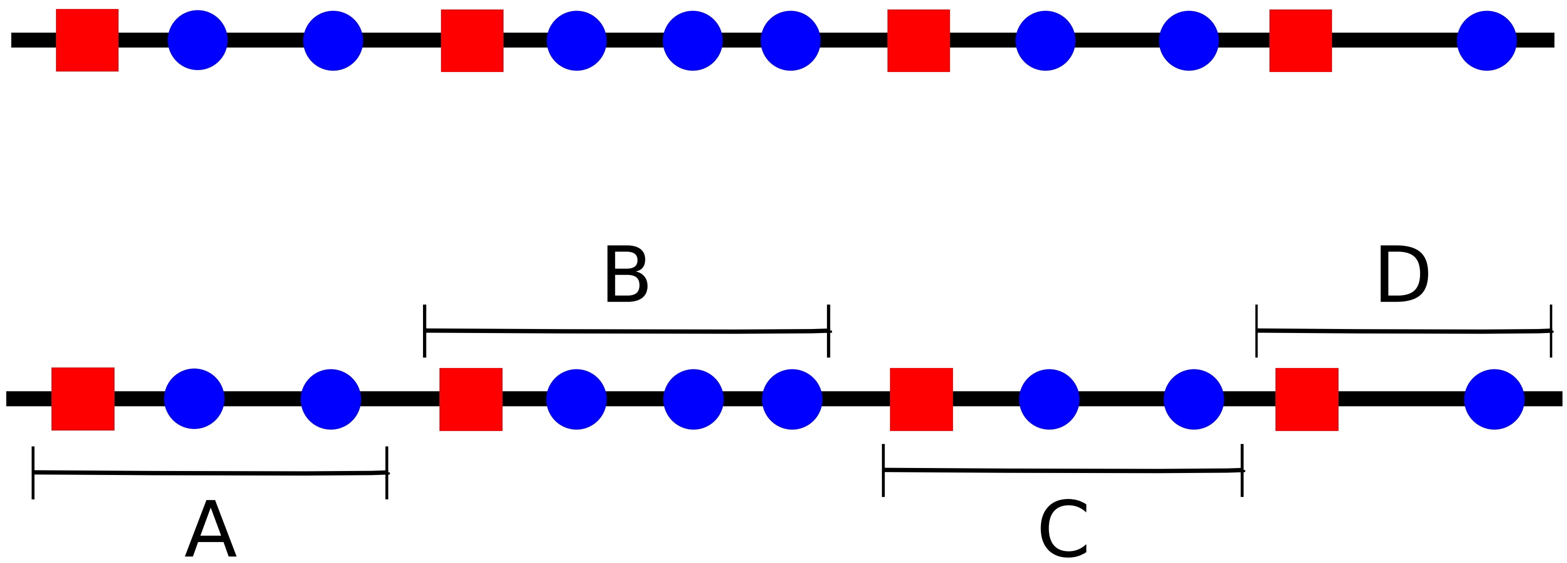
- Linear program (LP):

$$\begin{aligned} \min_{Q'} \quad & \delta \\ \text{subject to} \quad & \forall a \in H\Delta H, \widehat{Q}'(a) - \widehat{P}(a) \leq \delta \\ & \forall a \in H\Delta H, \widehat{P}(a) - \widehat{Q}'(a) \leq \delta \\ & \forall x \in S_Q, \widehat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \widehat{Q}'(x) = 1. \end{aligned}$$

  - No. of constraints bounded by shattering coefficient.

$$\Pi_{H\Delta H}(m_0 + n_0)$$

Algorithm - 1D



Regression - L2 Loss

- Problem:

$$\min_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} \left| \mathop{\mathrm{E}}_{\widehat{P}}[(h'(x) - h(x))^2] - \mathop{\mathrm{E}}_{\widehat{Q}'}[(h'(x) - h(x))^2] \right|.$$

$$\min_{\widehat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \left| \mathop{\mathbb{E}}_{\widehat{P}}[((\mathbf{w}' - \mathbf{w})^\top \mathbf{x})^2] - \mathop{\mathbb{E}}_{\widehat{Q}'}[((\mathbf{w}' - \mathbf{w})^\top \mathbf{x})^2] \right|$$

$$= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \left| \sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x}))[(\mathbf{w}' - \mathbf{w})^\top \mathbf{x}]^2 \right|$$

$$= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \left| \sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x}))[\mathbf{u}^\top \mathbf{x}]^2 \right|$$

$$= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \left| \mathbf{u}^\top \Big(\sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x})) \mathbf{x}\mathbf{x}^\top \Big) \mathbf{u} \right|.$$

Regression - L2 Loss

- Problem equivalent to

$$\min_{\substack{\|\mathbf{z}\|_1=1\\ \mathbf{z}\geq 0}} \max_{\|\mathbf{u}\|=1} |\mathbf{u}^\top \mathbf{M}(\mathbf{z})\mathbf{u}|,$$

with: $\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i,$

$$\mathbf{M}_0 = \sum_{\mathbf{x}\in S} P(\mathbf{x})\mathbf{x}\mathbf{x}^\top$$

$$\mathbf{M}_i = \mathbf{s}_i\mathbf{s}_i^\top$$

elements of $\mathrm{supp}(\widehat{Q})$

Regression - L2 Loss

- Semi-definite program (SDP): linear hypotheses.

$$
\begin{aligned}
\min_{\mathbf{z},\lambda} \quad & \lambda \\
\text{subject to} \quad & \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0 \\
& \lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0 \\
& \mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0,
\end{aligned}
$$

where the matrix $\mathbf{M}(\mathbf{z})$ is defined by:

$$
\mathbf{M}(\mathbf{z}) = \sum_{\mathbf{x} \in S} \widehat{P}(\mathbf{x}) \mathbf{x} \mathbf{x}^\top - \sum_{i=1}^{m_0} z_i \mathbf{s}_i \mathbf{s}_i^\top .
$$

Regression - L2 Loss

- SDP: generalization to H RKHS for some kernel K.

$$\begin{aligned} \min_{\mathbf{z},\lambda} \quad & \lambda \\ \text{subject to} \quad & \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0 \\ & \lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0 \\ & \mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0, \end{aligned}$$

with:
$$\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i$$
$$\mathbf{M}_0 = \mathbf{K}^{1/2} \operatorname{diag}(P(s_1), \ldots, P(s_{p_0}))\mathbf{K}^{1/2}$$
$$\mathbf{M}_i = \mathbf{K}^{1/2} \mathbf{I}_i \mathbf{K}^{1/2}.$$

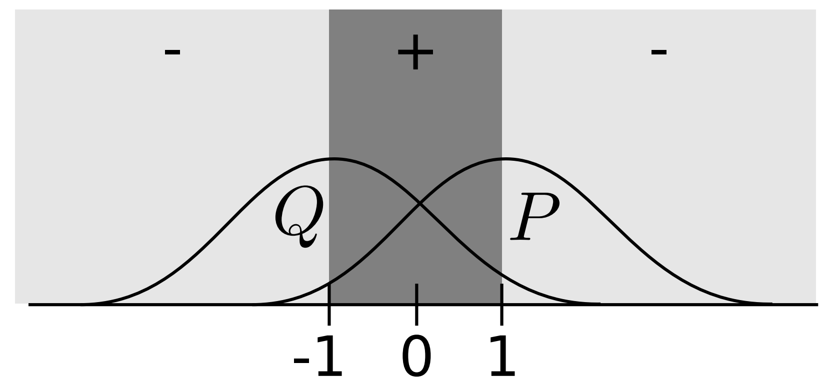
Discrepancy Min. Algorithm

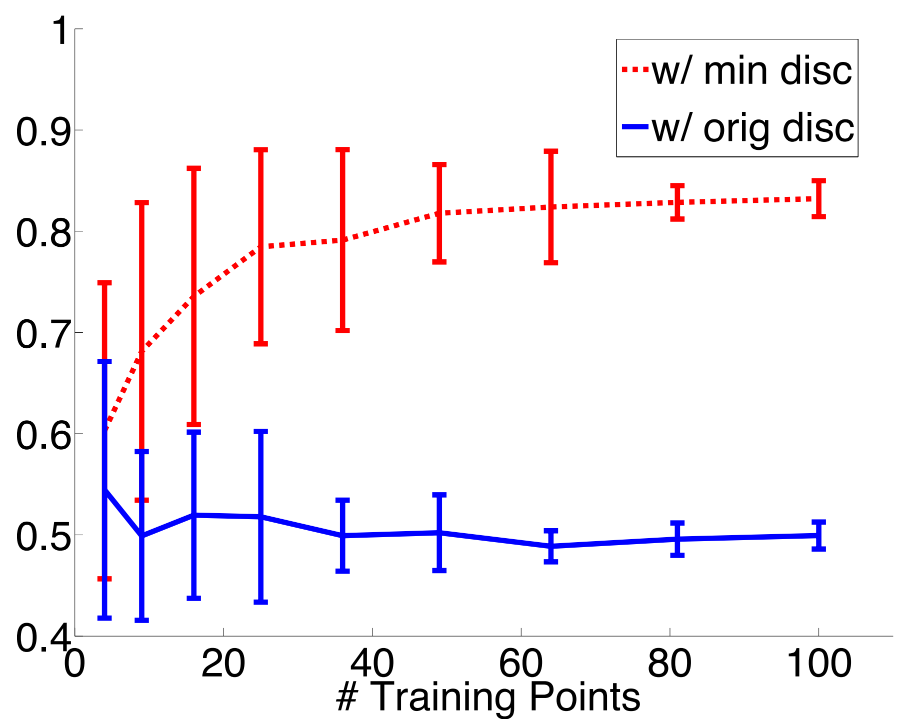
(Cortes & MM (TCS 2013))

- Convex optimization:
  - cast as semi-definite programming (SDP) prob.
  - efficient solution using smooth optimization.
- Algorithm and solution for arbitrary kernels.
- Outperforms other algorithms in experiments.

Experiments

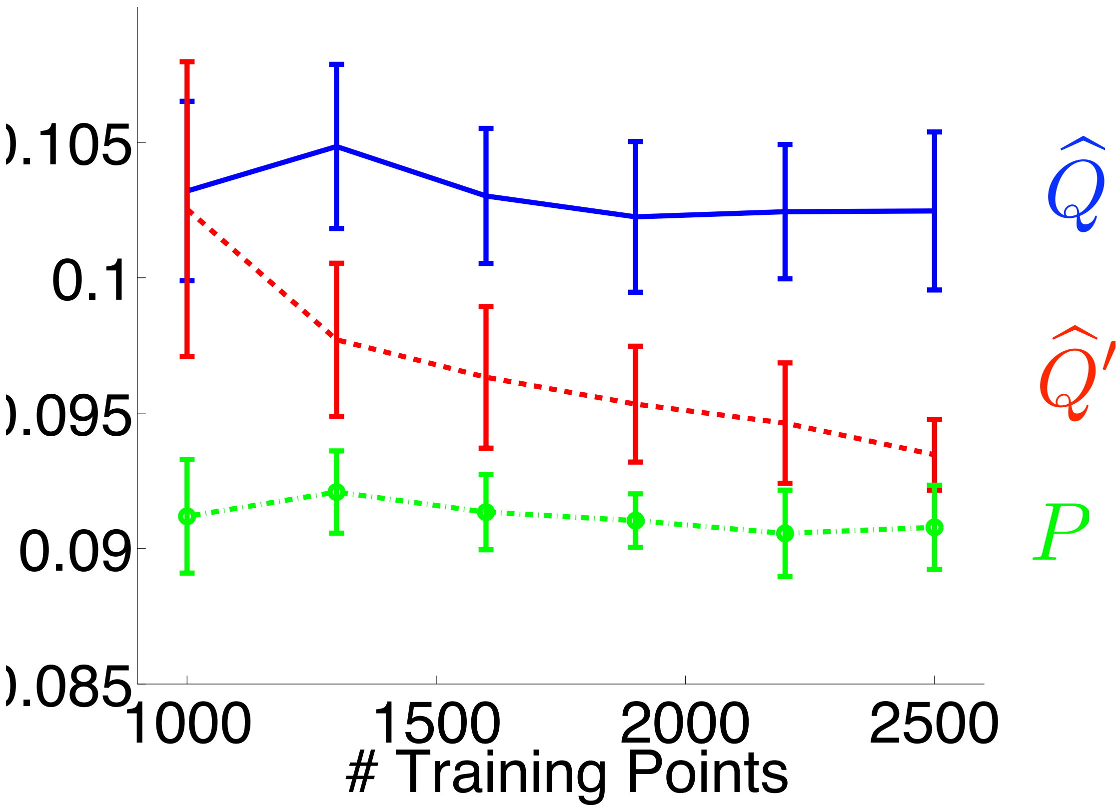
- Classification:
  - *Q* and *P* Gaussians.
  - *H*: halfspaces.
  - *f*: interval [-1, +1].





Experiments

- Regression:



SDP solved in about 15s using SeDuMi on 3GHz CPU with 2GB memory.

Experiments

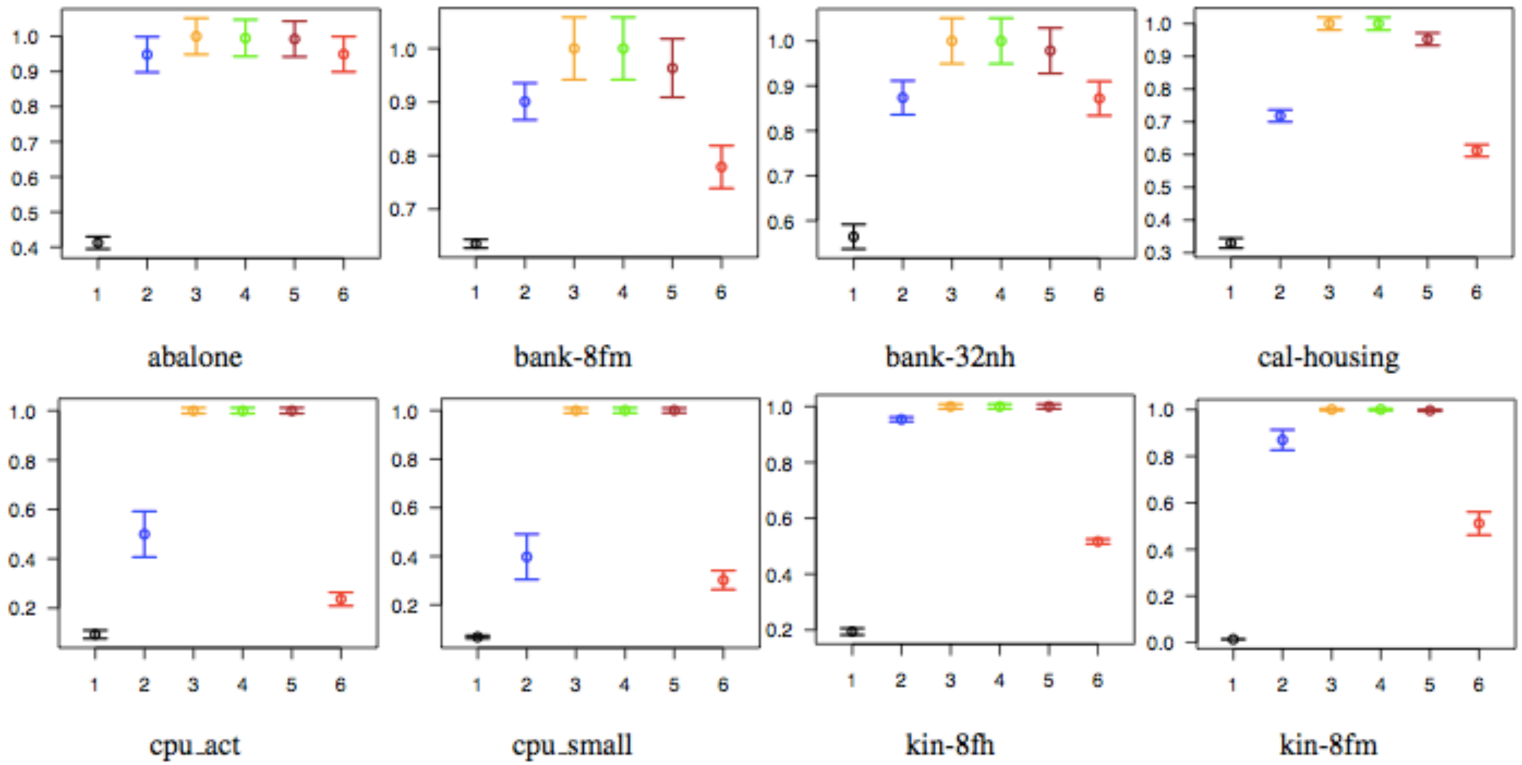


Fig. 11. Results with "easy-to-learn" biasing scheme: Relative MSE performance of (1): Optimal (in black); (2): KMM (in blue); (3): KLIEP (in orange); (4): Uniform (in green); (5): Two-Stage (in brown); and (6): DM (in red). Errors are normalized so that the average MSE of Uniform is 1.

Enhancement

(Cortes, MM, and Muñoz (2014))

- Shortcomings:
  - discrepancy depends on maximizing pair of hypotheses.
  - $\longrightarrow$ DM algorithm too conservative.
- Ideas:
  - finer quantity: *generalized discrepancy*, hypothesis-dependent.
  - reweighting depending on hypothesis.

Algorithm

(Cortes, MM, and Muñoz (2014))

- Choose Q_h such that objectives are unif. close:

$$\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{Q}_h}(h, f_Q)$$

$$\lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P).$$

- Ideally:

$$\mathsf{Q}_h = \operatorname*{argmin}_{\mathsf{q}} \left|\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)\right|.$$

- Using convex surrogate H'':

$$\mathsf{Q}_h = \operatorname*{argmin}_{\mathsf{q}} \max_{h'' \in H''} \left|\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}(h, h'')\right|.$$

Optimization

(Cortes, MM, and Muñoz (2014))

$$\begin{aligned}
\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) &= \operatorname*{argmin}_{l \in \{\mathcal{L}_{\mathsf{q}}(h, f_Q) : \mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})\}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')| \\
&= \operatorname*{argmin}_{l \in \mathbb{R}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')| \\
&= \frac{1}{2}\Big(\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big).
\end{aligned}$$

→ Convex optimization problem (loss jointly convex):

$$\min_h \lambda \|h\|_K^2 + \frac{1}{2}\Big(\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big).$$

Convex Surrogate Hyp. Set

(Cortes, MM, and Muñoz (2014))

- Choice of H'' among balls

$$B(r) = \{h'' \in H | \mathcal{L}_\mathsf{q}(h'', f_Q) \leq r^p\}.$$

- Generalization bound proven to be more favorable than DM for some choices of radius r.

- Radius r chosen via cross-validation using small amount of labeled data from target.

- Further improvement of empirical results.

Conclusion

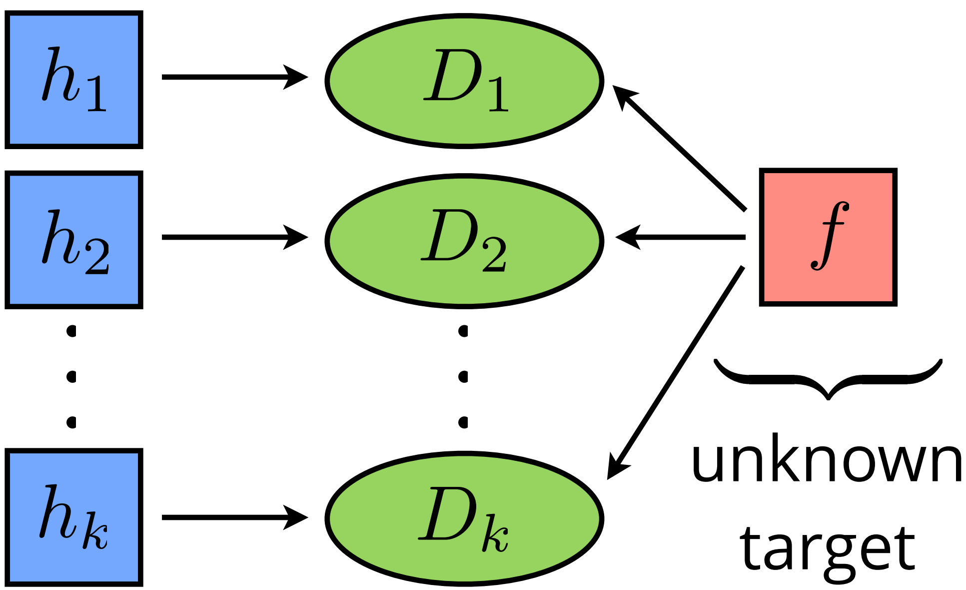
- Theory of adaptation based on discrepancy:
  - key term in analysis of adaptation and drifting.
  - discrepancy minimization algorithm DM.
  - compares favorably to other adaptation algorithms in experiments.
- Generalized discrepancy:
  - extension to hypothesis-dependent reweighting.
  - convex optimization problem.
  - further empirical improvements.
- Further generalization: (Awasthi, Cortes, MM, 2024).

Outline

- Domain adaptation.
- Multiple-source domain adaptation.

Problem Formulation

- Given distributions and corresponding hypotheses:



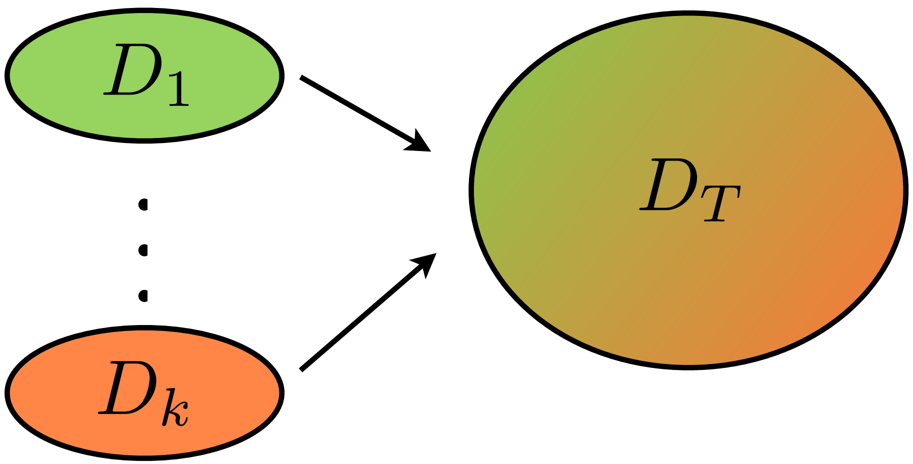
$$\underbrace{\forall i, \mathcal{L}(D_i, h_i, f) \leq \epsilon.}_{\text{each hypothesis performs well in its domain.}}$$

Notation: $\mathcal{L}(D_i, h_i, f) = \underset{x \sim D_i}{\mathrm{E}}[L(h_i(x), f(x))].$

Loss L assumed non-negative, bounded, convex and continuous.

Problem Formulation

- The unknown target distribution is a mixture of input distributions.



$$D_T(x) = \sum_{i=1}^{k} \lambda_i D_i(x)$$

- Task: choose a hypothesis mixture that performs well in target distribution.

$$h_z(x) = \sum_{i=1}^{k} z_i h_i(x)$$

convex combination rule

$$h_z(x) = \sum_{i=1}^{k} \frac{z_i D_i(x)}{\sum_{j=1}^{k} z_j D_j(x)} h_i(x)$$

distribution weighted combination

Known Target Distribution

- For some distributions, any convex combination performs poorly.

distribution weights

	D_T	D_0	D_1
a	0.5	1	0
b	0.5	0	1

hypothesis output

	f	h_0	h_1
a	1	1	0
b	0	1	0

- base hypotheses have no error within domain.
- any convex combination has error of 1/2!

Main Results

- Thus, although convex combinations seem natural, they can perform very poorly.
- We will show that distribution weighted combinations seem to define the "right" combination rule.
- There exists a single "robust" distribution weighted hypothesis, that does well for any target mixture.

$$\boxed{\forall f, \exists z, \forall \lambda,\ \mathcal{L}(D_\lambda, h_z, f) \leq \epsilon.}$$

Known Target Distribution

- If distribution is known, distribution weighted rule will always do well. Choose: $z = \lambda$.

$$h_\lambda(x) = \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{\sum_{j=1}^{k} \lambda_j D_j(x)} h_i(x).$$

- Proof:

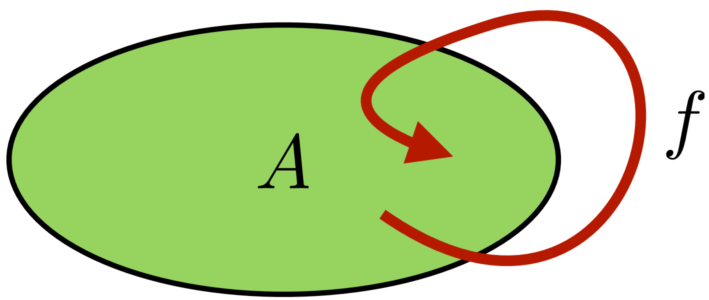
$$\begin{aligned}
\mathcal{L}(D_T, h_\lambda, f) &= \sum_{x \in X} L(h_\lambda(x), f(x)) D_T(x) \\
&\leq \sum_{x \in X} \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{D_T(x)} L(h_i(x), f(x)) D_T(x) \\
&= \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_i(x), f(x)) \leq \sum_{i=1}^{k} \lambda_i \epsilon = \epsilon.
\end{aligned}$$

Unknown Target Mixture

- Zero-sum game:
  - NATURE: select a target distribution D_i.
  - LEARNER: select a z, i.e. a distribution weighted hypothesis h_z.
  - Payoff: $\mathcal{L}(D_i, h_z, f)$.
  - Already shown: game value is at most ϵ.
- Minimax theorem (modulo discretization of z): there exists a mixture $\sum_j \alpha_j h_{z_j}$ of distribution weighted hypothesis that does well for any distribution mixture.

Balancing Losses

- Brouwer's Fixed Point theorem: for any compact, convex, non-empty set A and any continuous function $f\colon A \to A$, there exists x such that: $f(x) = x$.



Notation:
$$\mathcal{L}_i^z := \mathcal{L}(D_i, h_z, f).$$

- Define mapping ϕ by: $[\phi(z)]_i = \dfrac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z}$.
- By fixed point theorem (modulo continuity):

$$\exists z \colon \forall i, z_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z} \implies \forall i, \mathcal{L}_i^z = \sum_j z_j \mathcal{L}_j^z =: \gamma.$$

Bounding Loss

- For fixed point z,

$$\mathcal{L}(D_z, h_z, f) = \sum_{x \in X} L(h_z(x), f(x)) \Big(\sum_{i=1}^{k} z_i D_i(x) \Big)$$
$$= \sum_{i=1}^{k} z_i \sum_{x \in X} D_i(x) L(h_z(x), f(x))$$
$$= \sum_{i=1}^{k} z_i \mathcal{L}_i^z = \sum_{i=1}^{k} z_i \gamma = \gamma.$$

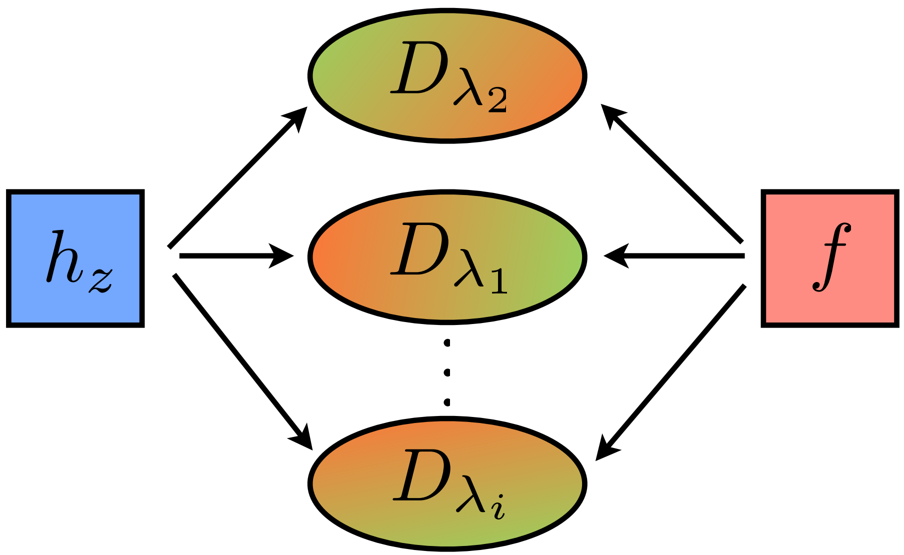
- Also, by convexity,

$$\gamma = \mathcal{L}(D_z, h_z, f) \leq \sum_{x \in X} \sum_{i=1}^{k} \frac{z_i D_i(x)}{D_z(x)} L(h_i(x), f(x)) D_z(x) = \sum_{i=1}^{k} z_i \mathcal{L}(D_i, h_i, f) \leq \epsilon.$$

Bounding Loss

- Thus, $\gamma \leq \epsilon$ and for any mixture λ,

$$\mathcal{L}(D_\lambda, h_z, f) = \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_z, f) \leq \sum_{i=1}^{k} \lambda_i \gamma = \gamma \leq \epsilon.$$



Details

- To deal with non-continuity refine hypotheses:

$$h_z^\eta(x) = \sum_{i=1}^k \frac{z_i D_i(x) + \eta/k}{\sum_{j=1}^k z_j D_j(x) + \eta} h_i(x).$$

- Theorem: for any target function f and any $\delta > 0$,

$$\exists \eta > 0, z\colon\ \forall \lambda, \mathcal{L}(D_\lambda, h_z^\eta, f) \leq \epsilon + \delta.$$

- If loss obeys triangle inequality:

$$\forall \delta > 0, \exists z, \eta > 0, \forall \lambda, f \in \mathcal{F},\ \mathcal{L}(D_\lambda, h_z^\eta, f) \leq 3\epsilon + \delta.$$

holds for all admissible target functions.

A Simple Algorithm

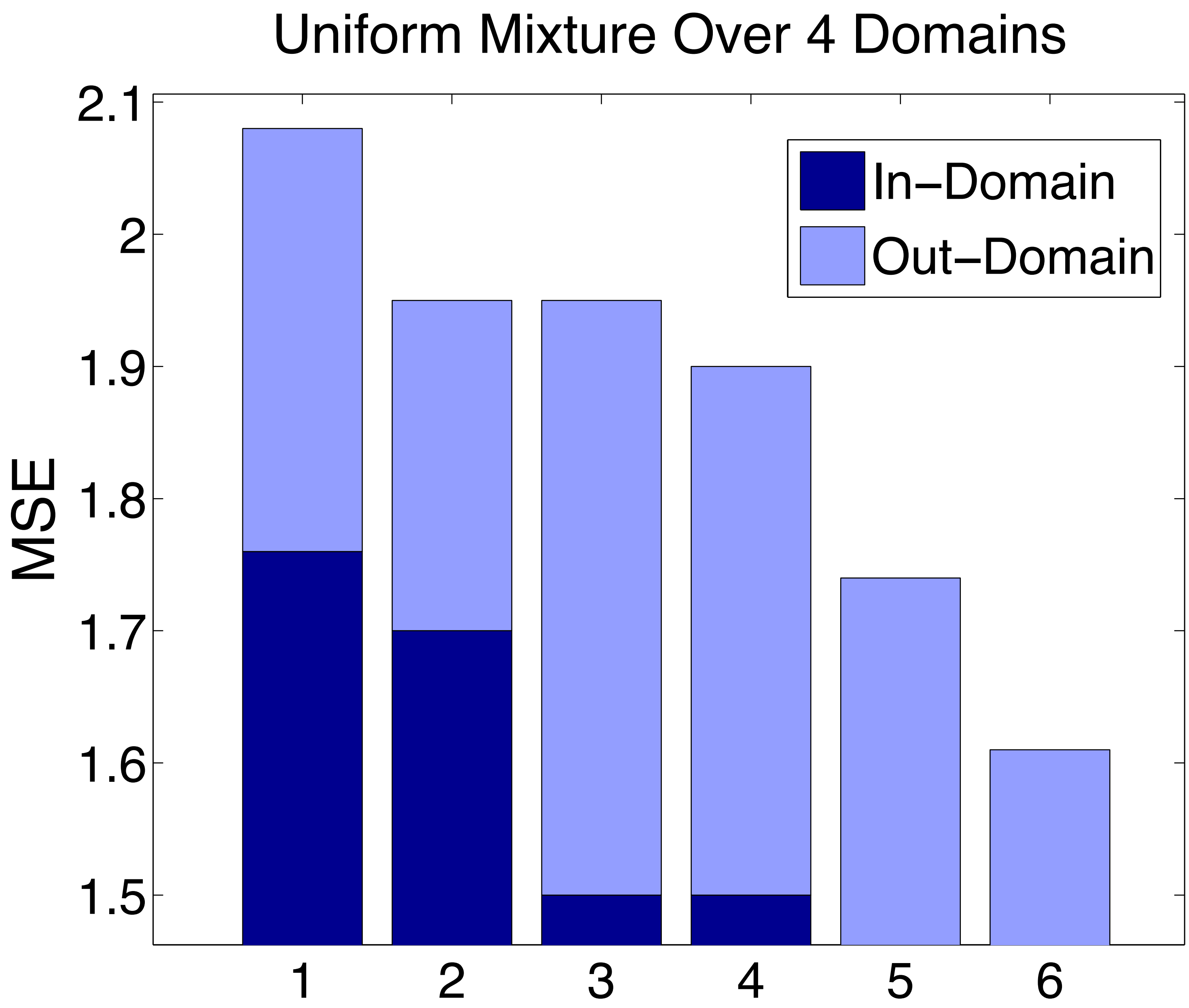
- A simple constructive algorithm, choose *z* with uniform weights:

$$
\begin{aligned}
\mathcal{L}(D_\lambda, h_u, f) &= \sum_x D_\lambda(x) L\left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x)\right) \\
&= \sum_x \left(\sum_{m=1}^k \lambda_m D_m(x)\right) L\left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x)\right) \\
&\leq \sum_x \underbrace{\frac{\sum_{m=1}^k \lambda_m D_m(x)}{\sum_{j=1}^k D_j(x)}}_{\leq 1} \sum_{i=1}^k D_i(x) L\left(h_i(x), f(x)\right) \\
&\leq \sum_{i=1}^k \sum_x D_i(x) L\left(h_i(x), f(x)\right) = \sum_{i=1}^k \mathcal{L}(D_i, h_i, f) = \sum_{i=1}^k \epsilon_i \boxed{\leq k\epsilon.}
\end{aligned}
$$

Preliminary Empirical Results

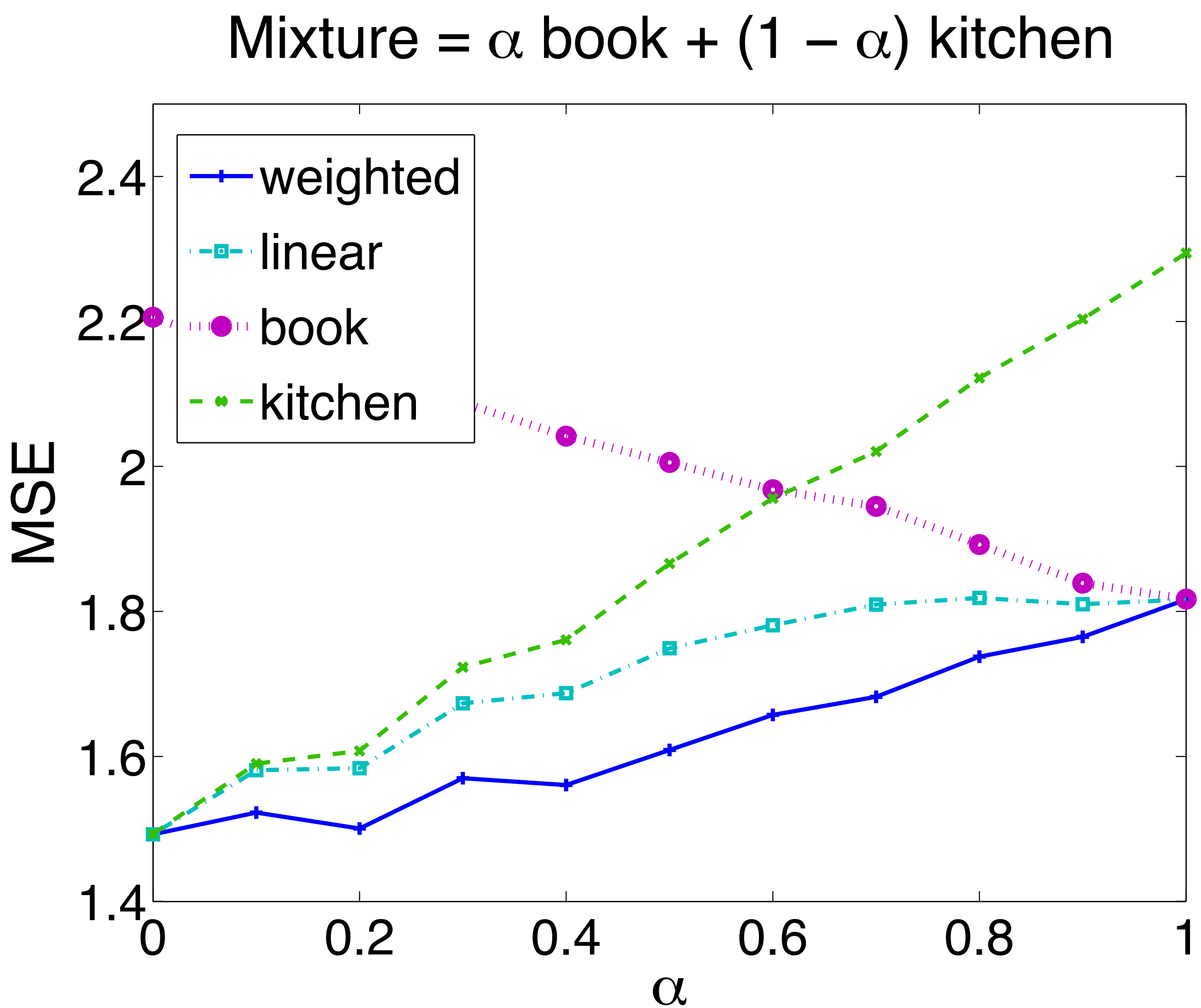
- Sentiment Analysis - given a product review (text string), predict a rating (between 1.0 and 5.0).
- 4 Domains: Books, DVDs, Electronics and Kitchen Appliances.
- Base hypotheses are trained within each domain (Support Vector Regression).
- We are not given the distributions. We model each distribution using a bag of words model.
- We then test the distribution combination rule on known target mixture domains.

Empirical Results



Empirical Results

- 2 class



Conclusion

- Formulation of the multiple source adaptation problem.
- Theoretical analysis for mixture distributions.
- Efficient algorithm for finding distribution weighted combination hypothesis?
- Beyond mixture distributions?

Rényi Divergences

- Definition: for $\alpha \geq 0$,

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_x P(x) \left[\frac{P(x)}{Q(x)}\right]^{\alpha-1}.$$

  - $\alpha = 1$: coincides with relative entropy.
  - $\alpha = 2$: logarithm of expected probability ratio;

$$D_\alpha(P\|Q) = \log \mathop{\mathrm{E}}_{x \sim P} \left[\frac{P(x)}{Q(x)}\right].$$

  - $\alpha = +\infty$: logarithm of maximum probability ratio;

$$D_\alpha(P\|Q) = \log \sup_{x \sim P} \left[\frac{P(x)}{Q(x)}\right].$$

Extensions - Arbitrary Target

(Mansour, MM, and Rostami, 2009)

- Theorem: for any $\delta > 0$ and $\alpha > 1$,

$$\exists \eta, z : \forall P, \mathcal{L}(P, h_z^\eta, f) \leq \Big[d_\alpha(P \| \mathcal{Q})(\epsilon + \delta) \Big]^{\frac{\alpha - 1}{\alpha}} M^{\frac{1}{\alpha}}.$$

$$\mathcal{Q} = \Big\{ \sum_{i=1}^{k} \lambda_i D_i : \lambda \in \Delta_k \Big\}$$

P

measured in terms of Rényi divergence,

$$d_\alpha(P, Q) = \Big[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \Big]^{\frac{1}{\alpha - 1}}.$$

Proof

- By Hölder's inequality, for any hypothesis h,

$$
\begin{aligned}
\mathcal{L}(P,h,f) &= \sum_x \frac{P(x)}{Q^{\frac{\alpha-1}{\alpha}}(x)} Q^{\frac{\alpha-1}{\alpha}}(x) L(h(x), f(x)) \\
&\leq \left[\sum_x \frac{P^\alpha x)}{Q^{\alpha-1}(x)}\right]^{\frac{1}{\alpha}} \left[\sum_x Q(x) L^{\frac{\alpha}{\alpha-1}}(h(x), f(x))\right]^{\frac{\alpha-1}{\alpha}} \\
&= (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\mathop{\mathrm{E}}_{x\sim Q}[L^{\frac{\alpha}{\alpha-1}}(h(x), f(x))]\right]^{\frac{\alpha-1}{\alpha}} \\
&= (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\mathop{\mathrm{E}}_{x\sim Q}[L(h(x), f(x)) L^{\frac{1}{\alpha-1}}(h(x), f(x))]\right]^{\frac{\alpha-1}{\alpha}} \\
&\leq (d_\alpha(P\|Q))^{\frac{\alpha-1}{\alpha}} \left[\mathcal{L}(Q,h,f) M^{\frac{1}{\alpha-1}}\right]^{\frac{\alpha-1}{\alpha}}.
\end{aligned}
$$

Other Extensions

(Mansour, MM, and Rostami, 2009)

- Approximate distributions (estimated):
  - similar results shown depending on divergence between true and estimated distributions.
- Different source target functions f_i:
  - similar results when target functions close to f on target distribution.

References

- S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *NIPS*. 2007.
- S. Ben-David, T. Lu, T. Luu, and D. Pal. Impossibility theorems for domain adaptation. *Journal of Machine Learning Research-Proceedings Track*, 9:129–136, 2010.
- S. Bickel, M. Bruckner, and T. Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10:2137–2155, 2009.
- J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *NIPS.* 2008.
- Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *NIPS*. 2010.
- Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 2014.

References

- Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In Proceedings of *ALT*. 2008.
- Koby Crammer, Michael Kearns, and Jennifer Wortman. Learning from Data of Variable Quality. In Proceedings of *NIPS*, 2006.
- Koby Crammer, Michael Kearns, and Jennifer Wortman.Learning from multiple sources.In Proceedings of *NIPS*, 2007.
- Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.
- M. Dredze, J. Blitzer, P. P. Talukdar, K. Ganchev, J. Graca, and F. Pereira. Frustratingly Hard Domain Adaptation for Parsing. In *CoNLL*, 2007.
- J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, volume 19, pages 601–608. 2006.

References

- Kifer D., Ben-David S., Gehrke J. Detecting change in data streams. In: Proceedings of *VLDB*, pp 180–191. 2004.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *COLT*. 2009.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In Advances in *NIPS*, pages 1041-1048. 2009.
- Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In Proceedings of *UAI*. 2009.
- Mehryar Mohri and Andres Muñoz Medina. New analysis and algorithm for learning with drifting distributions. In Proceedings of *ALT*, volume 7568, pages 124-138. 2012.

References

- H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227– 244, 2000.

- M. Sugiyama, S. Nakajima, H. Kashima, P. von Bunau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*. 2008.

- M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bunau, and M.Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.