Discrepancy and Adaptation

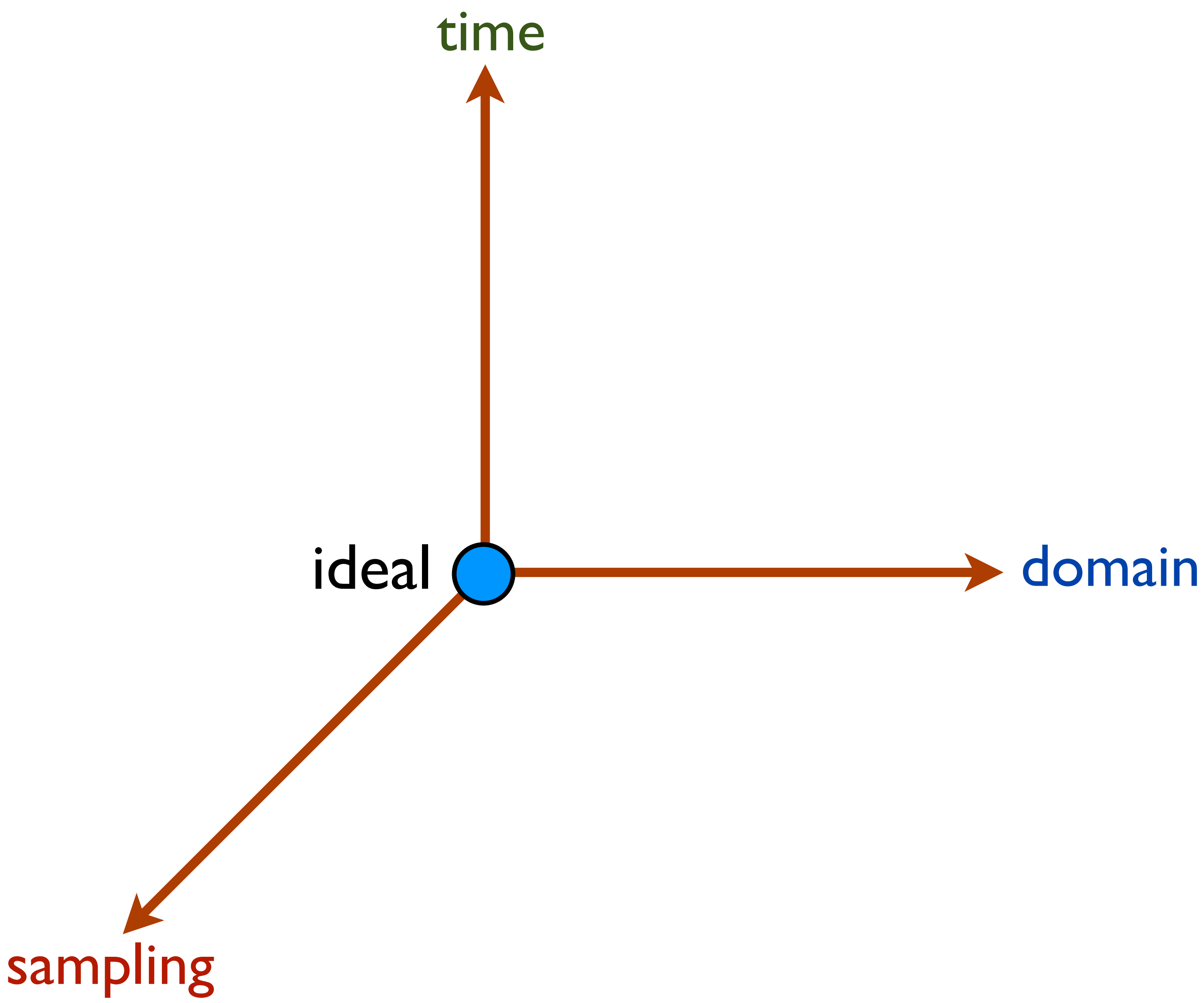
Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Includes joint work with Corinna Cortes,
Yishay Mansour, and Afshin Rostami.

Ideal World

- Standard learning assumptions:
  - same distribution for training and test.
  - distributions fixed over time.
  - IID sampling.

Ideal vs Real World



Domain Adaptation

- Sentiment analysis: appraisal information for some domains, e.g., movies, books, music, restaurants, but no labels for travel.
- Language modeling, part-of-speech tagging.
- Statistical parsing.
- Speech recognition.
- Computer vision.

→ Solution critical for applications.

Domain Adaptation Problem

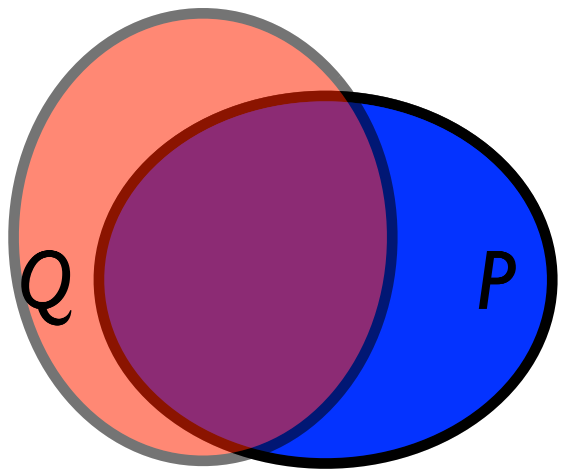
- Domains: source (Q, f_Q), target (P, f_P).
- Input: labeled sample S drawn from source, unlabeled sample T drawn from target.
- Problem: find hypothesis h in H with small expected loss with respect to target domain,

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

Previous Work - Adaptation Theory

- (Ben-David et al., NIPS 2006) & (Blitzer et al., NIPS 2007): bounds for binary classification based on d_A distance and λ_H, $3\times$ error issue.
- (Mansour, MM, Rostami, COLT 2009): learning bounds and analysis for general loss functions based on discrepancy and optimal hypotheses, favorable under plausible assumptions, pointwise loss guarantees for kernel algorithms.
- (Ben-David et al., AISTATS 2010): some negative examples for adaptation in binary classification.
- (Cortes, Mansour, and MM, NIPS 2010): analysis and learning guarantees for importance weighting.
- (Cortes and MM, ALT 2011): simpler and more general learning bounds, discrepancy minimization algorithm with kernels, efficient algorithm for solving SDP using smooth approximation technique.

Distribution Mismatch



Which distance should we use
to compare these distributions?

Discrepancy

(Mansour, MM, Rostami, 2009)

- Definition:

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

  - symmetric, verifies 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)).

Estimation from Finite Samples

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

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

Theoretical Guarantees

- Two types of questions:
  - difference between average loss of hypothesis h on Q versus P?
  - 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

- Notation:
  - $\mathcal{L}_Q(h_Q^*, f) = \min_{h \in H} \mathcal{L}_Q(h, f)$
  - $\mathcal{L}_P(h_P^*, f) = \min_{h \in H} \mathcal{L}_P(h, f)$
- Theorem: assume that L obeys the triangle inequality, then the following holds:

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

Some Special 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),

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

Kernel-Based Reg. (KBR) 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 .

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

Guarantees for KBR Algorithms

(Cortes, MM, 2011)

- Theorem: let K be a PDS kernel with $K(x,x) \leq R^2$ and L a loss function such that $L(\cdot, y)$ is μ-Lipschitz. Assume that $f_P \in H$, then, for all $(x,y) \in X \times Y$,

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

where $\eta = \max\{L(f_Q(x), f_P(x)) \colon x \in \mathrm{supp}(\widehat{Q})\}$.

Adaptation Algorithm

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

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

- Solve modified KBR problem:

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

Discrepancy Min. - Input space

- For L_2 loss and $H = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$:

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

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

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

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

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

with $\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{\mathfrak{m}} z_i \mathbf{M}_i, \quad \mathbf{M}_0 = \sum_{j=\mathfrak{m}+1}^{\mathfrak{q}} \widehat{P}(\mathbf{x}_j) \mathbf{x}_j \mathbf{x}_j^\top, \quad \mathbf{M}_i = \mathbf{x}_i \mathbf{x}_i^\top, i \in [1, \mathfrak{m}].$

Discrepancy Min. - Input space

- For L_2 loss and $H = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$, can be cast as an SDP:

$$
\begin{aligned}
\text{minimize} \quad & \|\mathbf{M}(\mathbf{z})\|_2 \\
\text{subject to} \quad & \mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{\mathfrak{m}} z_i \mathbf{M}_i \\
& \mathbf{M}_0 = \sum_{j=\mathfrak{m}+1}^{\mathfrak{q}} \widehat{P}(\mathbf{x}_j) \mathbf{x}_j \mathbf{x}_j^\top \\
& \mathbf{M}_i = \mathbf{x}_i \mathbf{x}_i^\top, i \in [1, \mathfrak{m}] \\
& \mathbf{z}^\top \mathbf{1} = 1 \wedge \mathbf{z} \geq 0.
\end{aligned}
$$

→ what about if we want to use kernels?

Discrepancy Min. with Kernels

- For L_2 loss and $H = \{h \in \mathbb{H} \colon \|h\|_K \leq \Lambda\}$, proof that it can be cast as a similar SDP:

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{M}'(\mathbf{z})\|_2 \\ \text{subject to} \quad & \mathbf{M}'(\mathbf{z}) = \mathbf{M}'_0 - \sum_{i=1}^{m} z_i \mathbf{M}'_i \\ & \mathbf{M}'_0 = \mathbf{K}^{1/2} \mathbf{D}_0 \mathbf{K}^{1/2} \\ & \mathbf{M}'_i = \mathbf{K}^{1/2} \mathbf{D}_i \mathbf{K}^{1/2} \\ & \mathbf{z}^\top \mathbf{1} = 1 \wedge \mathbf{z} \geq 0. \end{aligned}$$

→ but, cannot be solved practically even for a few hundred points, even with best public SDP solvers.

Disc. Min. SDP Algorithm

- Smooth approximation:
  - $F\colon \mathbf{z} \mapsto \|\mathbf{M}(\mathbf{z})\|_2$ not differentiable.
  - $G_p\colon \mathbf{z} \mapsto \frac{1}{2}\operatorname{Tr}[\mathbf{M}(\mathbf{z})^{2p}]^{\frac{1}{p}}$: smooth unif. approximation.
- Algorithm: $\mathbf{J} = (\langle \mathbf{M}_i, \mathbf{M}_j \rangle_F)_{1 \leq i,j \leq \mathfrak{m}}$.

Algorithm 2

$\mathbf{u}_0 \leftarrow \operatorname{argmin}_{\mathbf{u} \in C} \mathbf{u}^\top \mathbf{J} \mathbf{u}$
for $k \geq 0$ **do**
  $\mathbf{v}_k \leftarrow \operatorname{argmin}_{\mathbf{u} \in C} \frac{2p-1}{2}(\mathbf{u} - \mathbf{u}_k)^\top \mathbf{J}(\mathbf{u} - \mathbf{u}_k) + \nabla G_p(\mathbf{M}(\mathbf{u}_k))^\top \mathbf{u}$
  $\mathbf{w}_k \leftarrow \operatorname{argmin}_{\mathbf{u} \in C} \frac{2p-1}{2}(\mathbf{u} - \mathbf{u}_0)^\top \mathbf{J}(\mathbf{u} - \mathbf{u}_0) + \sum_{i=0}^{k} \frac{i+1}{2} \nabla G_p(\mathbf{M}(\mathbf{u}_i))^\top \mathbf{u}$
  $\mathbf{u}_{k+1} \leftarrow \frac{2}{k+3}\mathbf{w}_k + \frac{k+1}{k+3}\mathbf{v}_k$
end for

Convergence Guarantee

- Let $r = \max_{\mathbf{z} \in C} \operatorname{rank}(\mathbf{M}(\mathbf{z})) \leq \max\{N, \sum_{i=0}^{n} \operatorname{rank}(\mathbf{M}_i)\}$.
- Theorem: for any $\epsilon > 0$, the algorithm solves the discrepancy minimization SDP with relative accuracy ϵ in $O(\sqrt{r \log r}/\epsilon)$ iterations.

Guarantees for KBR Algorithms

- 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 $(x,y) \in X \times Y$,

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

where

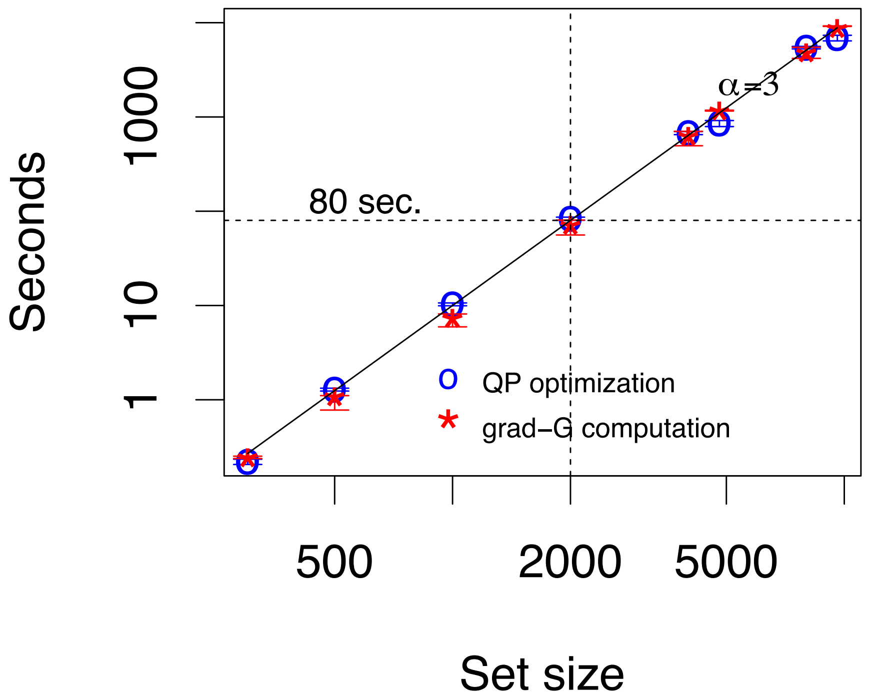
$$\delta = \min_{h \in H} \Big\| \mathop{\mathrm{E}}_{x \sim \widehat{Q}}\Big[\big(h(x) - f_Q(x)\big)\Phi_K(x)\Big] - \mathop{\mathrm{E}}_{x \sim \widehat{P}}\Big[\big(h(x) - f_P(x)\big)\Phi_K(x)\Big]\Big\|_K.$$

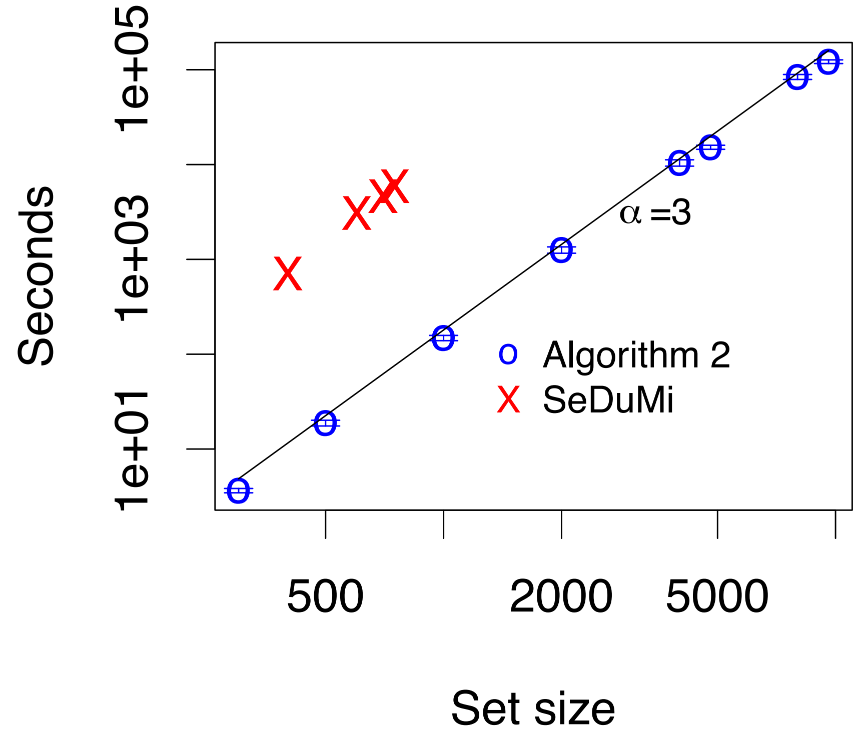
Discrepancy = Distance

- **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.

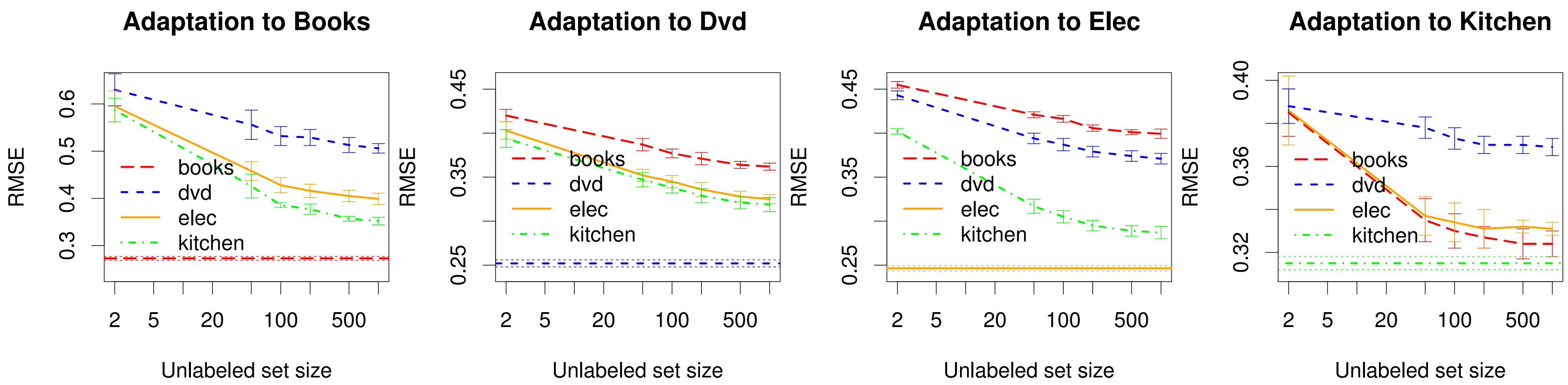
- **Proof**: $\Psi\colon h \mapsto \mathrm{E}_{x \sim P}[h^2(x)] - \mathrm{E}_{x \sim Q}[h^2(x)]$ is Lipschitz for norm $\| \cdot \|_\infty$, thus continuous on $C(X)$.
  - $\mathrm{disc}(P, Q) = 0$ implies $\Psi(h) = 0$ for all $h \in \mathbb{H}$.
  - since $\mathbb{H}$ is dense in $C(X)$, $\Psi = 0$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

Experiments - Time





Experiments - Performance



- Multi-domain sentiment analysis data set (Blitzer et al. 2007): `books`, `dvd`, `elec`, `kitchen`.
- Treated as regression task.

Conclusion

- Recent and upcoming:
  - theoretical properties of discrepancy minimization.
  - explicit learning guarantees in terms of size of unlabeled data.
  - adaptation with small amount of labeled data: analysis, theory, and algorithms.