Adapting to a Non-Ideal World

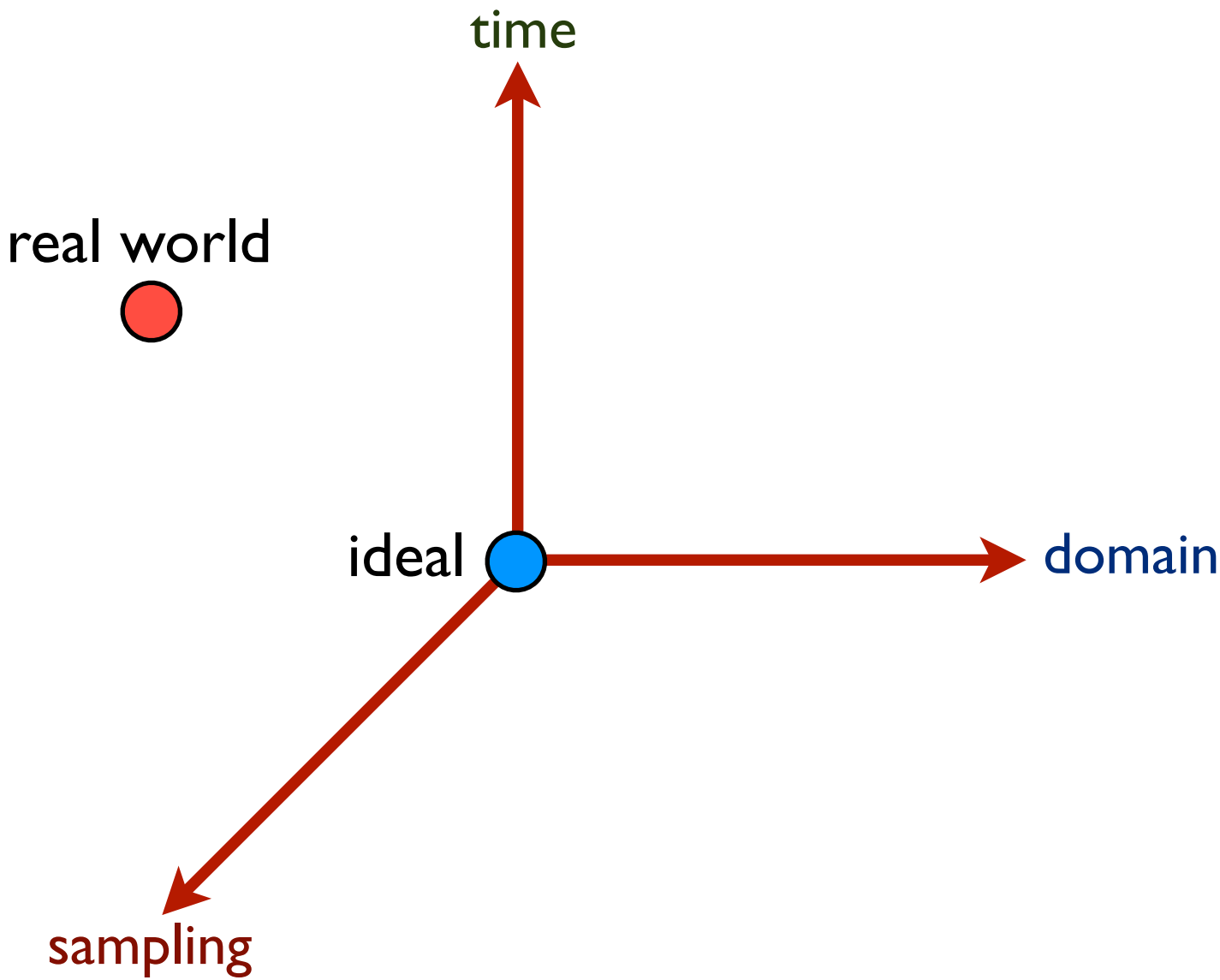
Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Includes joint work with Corinna Cortes,
Yishay Mansour, Andres Muñoz, and Afshin Rostamizadeh.

Ideal World

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

Ideal vs Real World



This Talk

- Domain adaptation
  - Discrepancy distance
  - Theoretical guarantees
  - Algorithm
  - Experiments
- Drifting scenario

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, that is

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

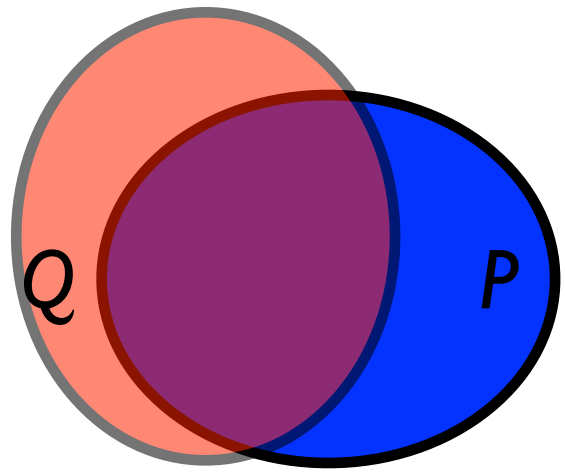
Related Work

- Single-source adaptation:
  - language modeling, probabilistic parsers, maxent models: source domain used to define a prior.
  - relation between adaptation and the d_A distance [Ben-David et al. (NIPS 2006) and Blitzer et al. (NIPS 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

- Multiple-source:
  - related problem: same input distribution, but different labels modulo disparity constraints [Crammer, Kearns, and Wortman (NIPS 2005, NIPS 2006)].
  - theoretical analysis and method for multiple-source adaptation using the notion of Rényi divergence [Mansour, MM, and Rostami (NIPS 2008, UAI 2009)].

Distribution Mismatch



Which distance should we use to compare these distributions?

Simple Analysis

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

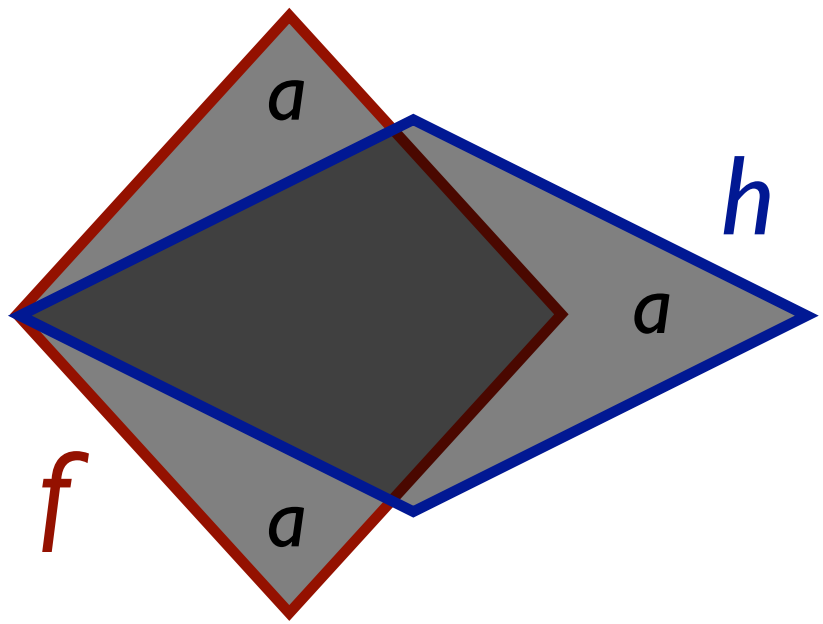
$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M\, L_1(Q, P).$$

- Proof:

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

But, is this bound informative?

Example - 0/1 Loss



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

Discrepancy

(Mansour, MM, Rostami, 2009)

- Definition:

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

- 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 $1-\delta$,

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

This Talk

- Domain adaptation
  - Discrepancy distance
  - Theoretical guarantees
  - Algorithm
  - Experiments
- Drifting scenario

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

[Mansour, MM, Rostami (COLT 2009)]

- 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 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) & (Blitzer et al., NIPS 2007): always worse in these cases.

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 .

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

Guarantees for KBR Algorithms

[Cortes & MM (ALT 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 $\eta = \max\{L(f_Q(x), f_P(x)) \colon x \in \mathrm{supp}(\widehat{Q})\}$.

Guarantees for KBR Algorithms

[Cortes & MM (2012)]

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

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

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

- $\delta = 0$ for Gaussian kernels and $f_P = f_Q$ continuous.

Empirical Discrepancy

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

This Talk

- Domain adaptation
  - Discrepancy distance
  - Theoretical guarantees
  - Algorithm
  - Experiments
- Drifting scenario

Algorithm - Idea

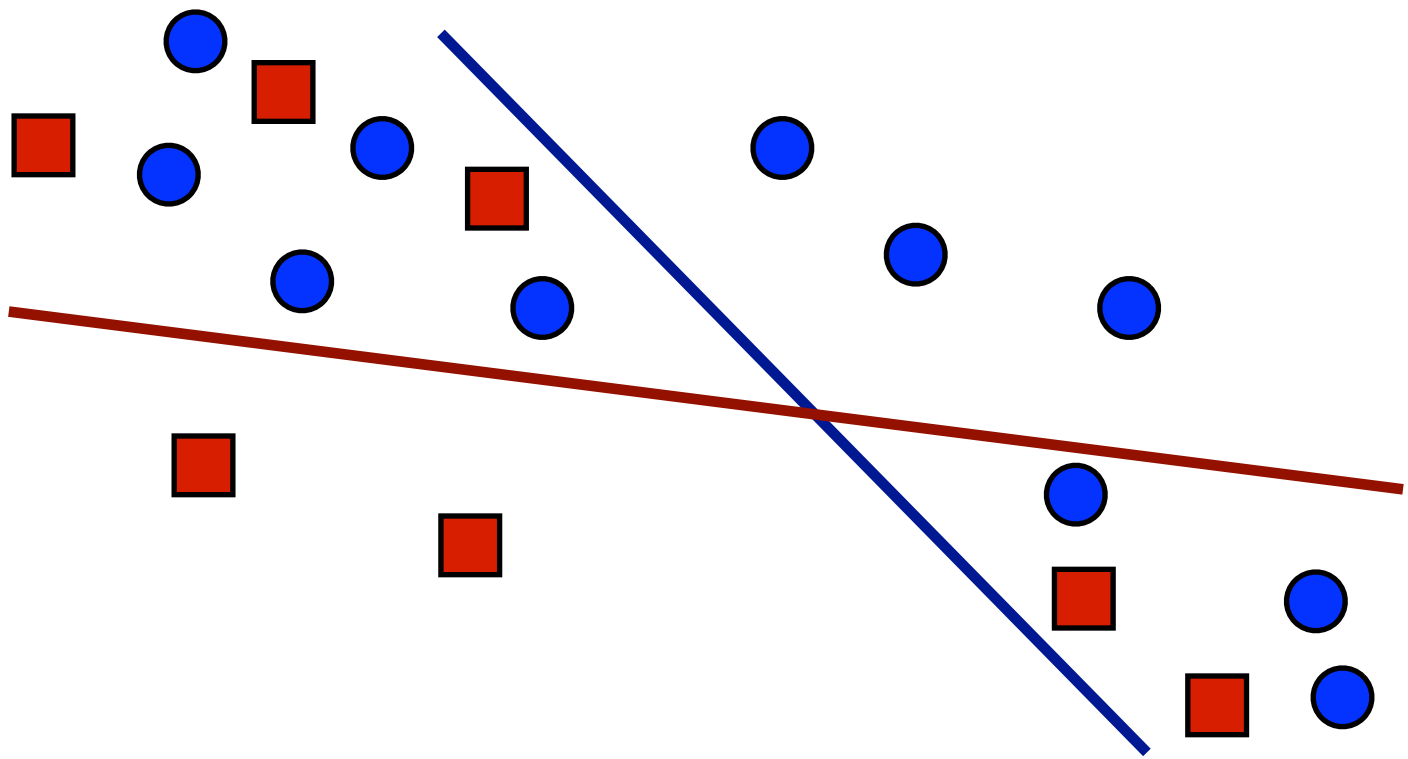
- Search for a new empirical distribution q^* with same support:

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

- Solve modified KBR problem:

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

Case of Halfspaces



Regression - L2 Loss

- Problem:

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

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:

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

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

→ what about if we want to use kernels?

Discrepancy Min. with Kernels

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

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

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

Discrepancy Min. SDP Algorithm

[Cortes & MM (ALT 2011)]

- Smooth approximation [Nesterov (1983, 2005)]:
  - $F\colon \mathbf{z} \mapsto \|\mathbf{M}(\mathbf{z})\|_2$ 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: $\mathbf{J} = (\langle \mathbf{M}_i, \mathbf{M}_j \rangle_F)_{1 \leq i,j \leq \mathrm{m}}$.

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

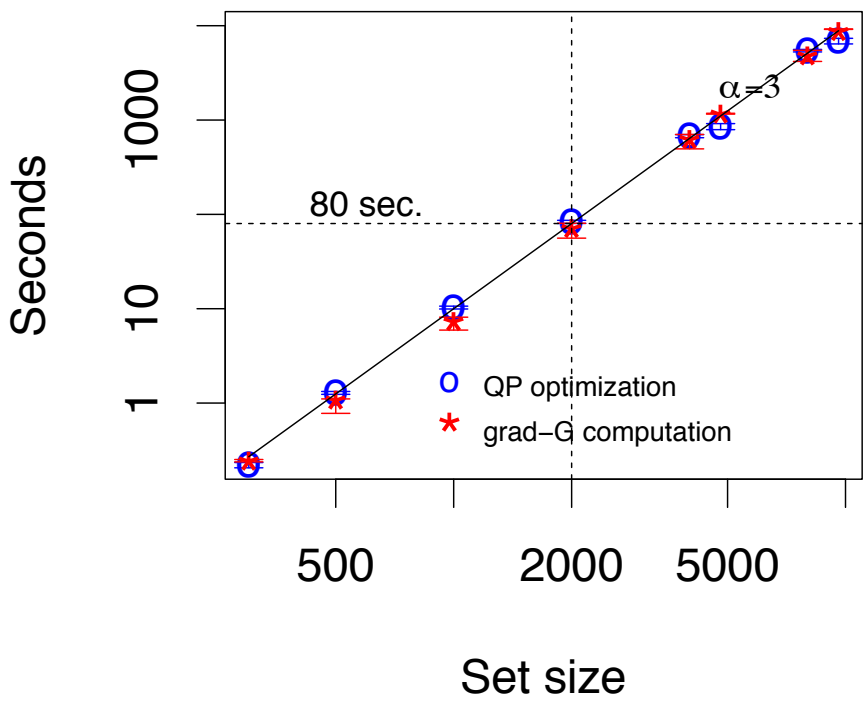
Convergence Guarantee

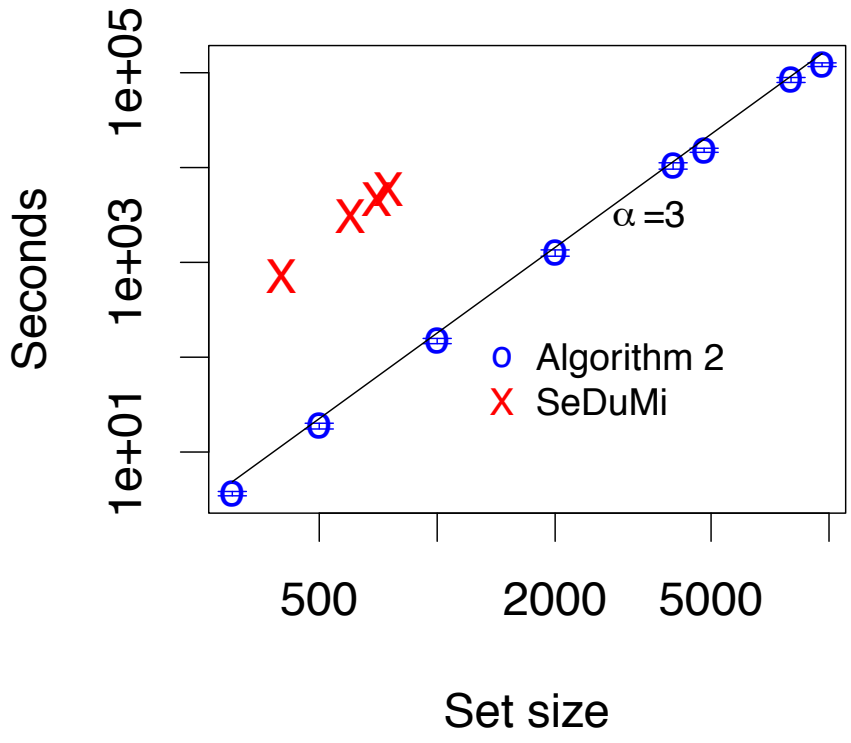
- Let $r = \max_{\mathbf{z} \in C} \operatorname{rank}(\mathbf{M}(\mathbf{z})) \leq \max\{N, \sum_{i=0}^{n} \operatorname{rank}(\mathbf{M}_i)\}$.
- Theorem: for any $\epsilon > 0$, the algorithm solves the discrepancy minimization SDP with relative accuracy ϵ in $O(\sqrt{r \log r}/\epsilon)$ iterations.

This Talk

- Domain adaptation
  - Discrepancy distance
  - Theoretical guarantees
  - Algorithm
  - Experiments
- Drifting scenario

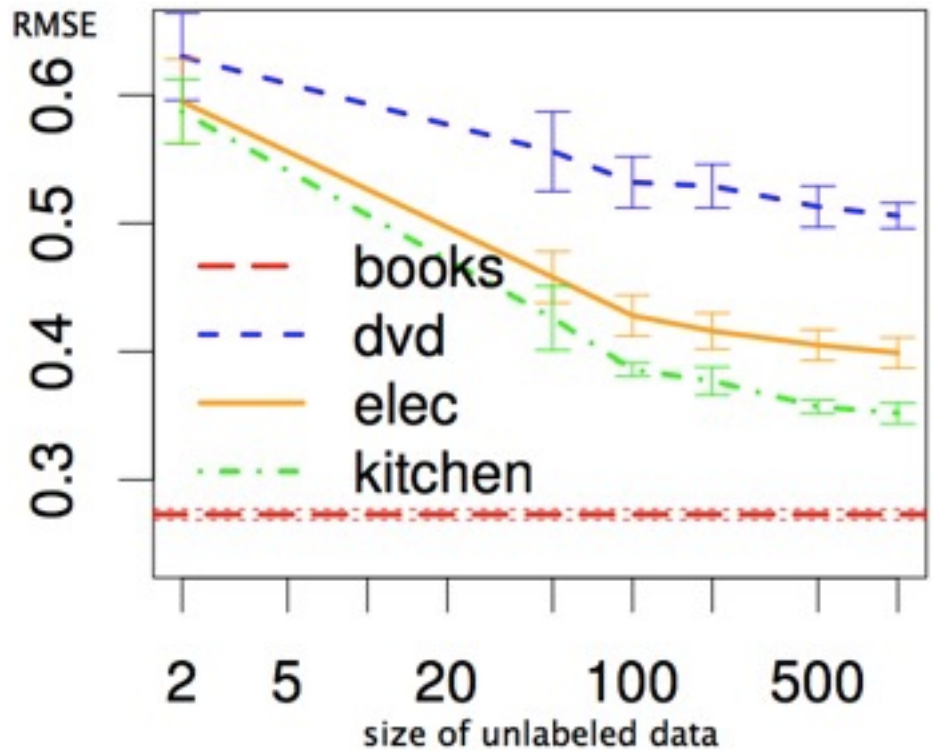
Experiments - Time





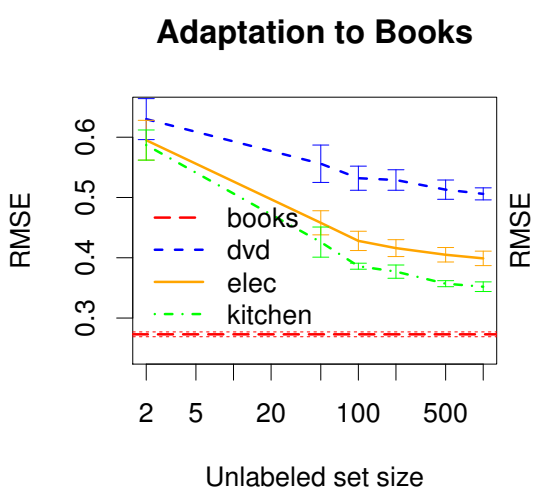
Experiments - Performance

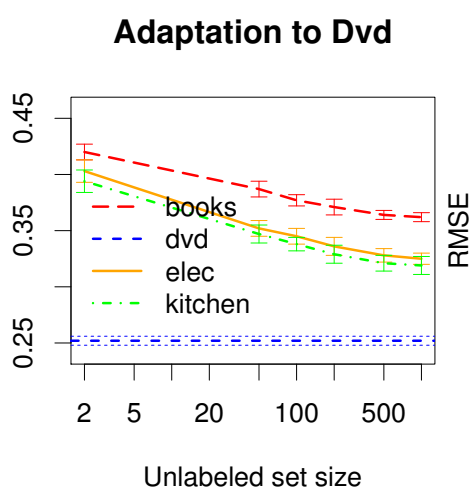
adaptation to books

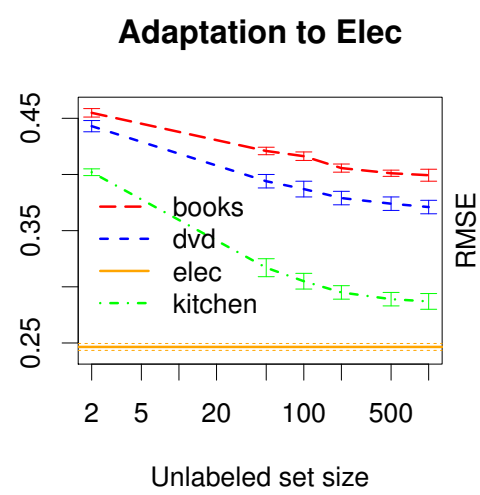


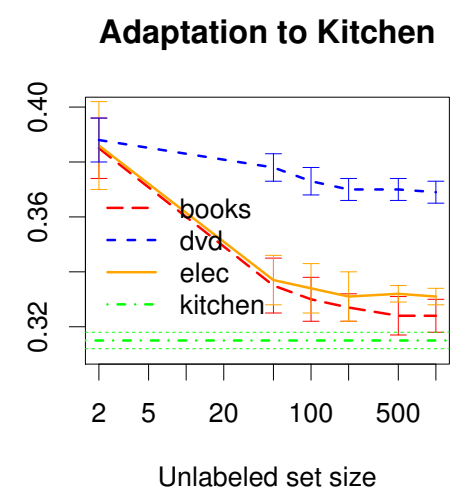
- Multi-domain sentiment analysis data set (Blitzer et al. 2007): `books`, `dvd`, `elec`, `kitchen`.
- Treated as regression task.

Experiments - Performance









This Talk

- Domain adaptation
  - Discrepancy distance
  - Theoretical guarantees
  - Algorithm
  - Experiments
- Drifting scenario

Generalized Discrepancy

- Definition: given two distributions D and D' over $\mathcal{X} \times \mathcal{Y}$, the $\mathcal{Y}$-discrepancy is defined as follows:

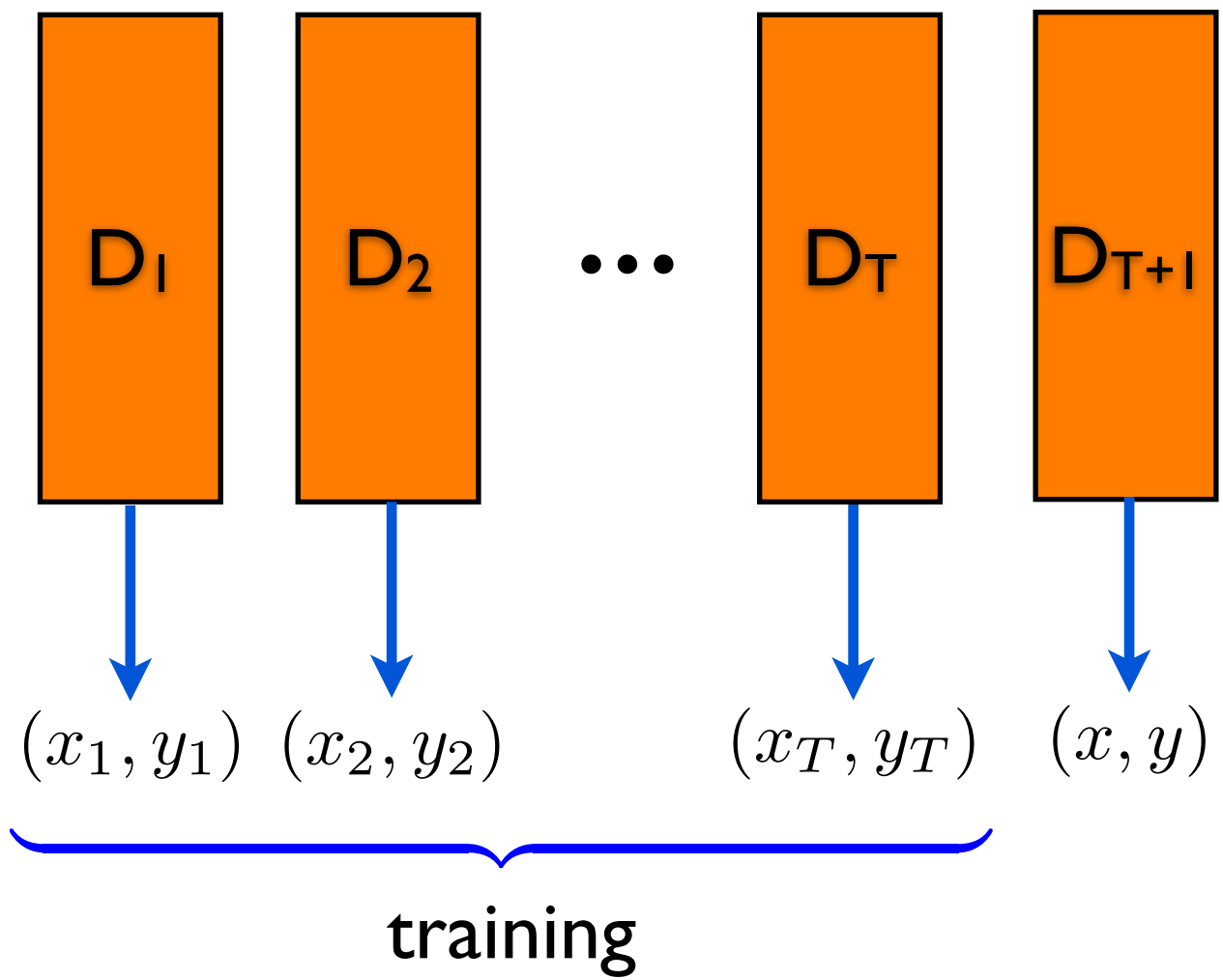
$$\mathrm{disc}_{\mathcal{Y}}(D, D') = \sup_{h \in H} \left| \mathcal{L}_{D'}(h) - \mathcal{L}_{D}(h) \right|,$$

where $\mathcal{L}_D(h) = \mathop{\mathrm{E}}_{(x,y)\sim D}[L(h(x), y)]$.

Drifting PAC Scenario

[MM & Muñoz, 2012]

- Model:



Drifting PAC - Guarantee

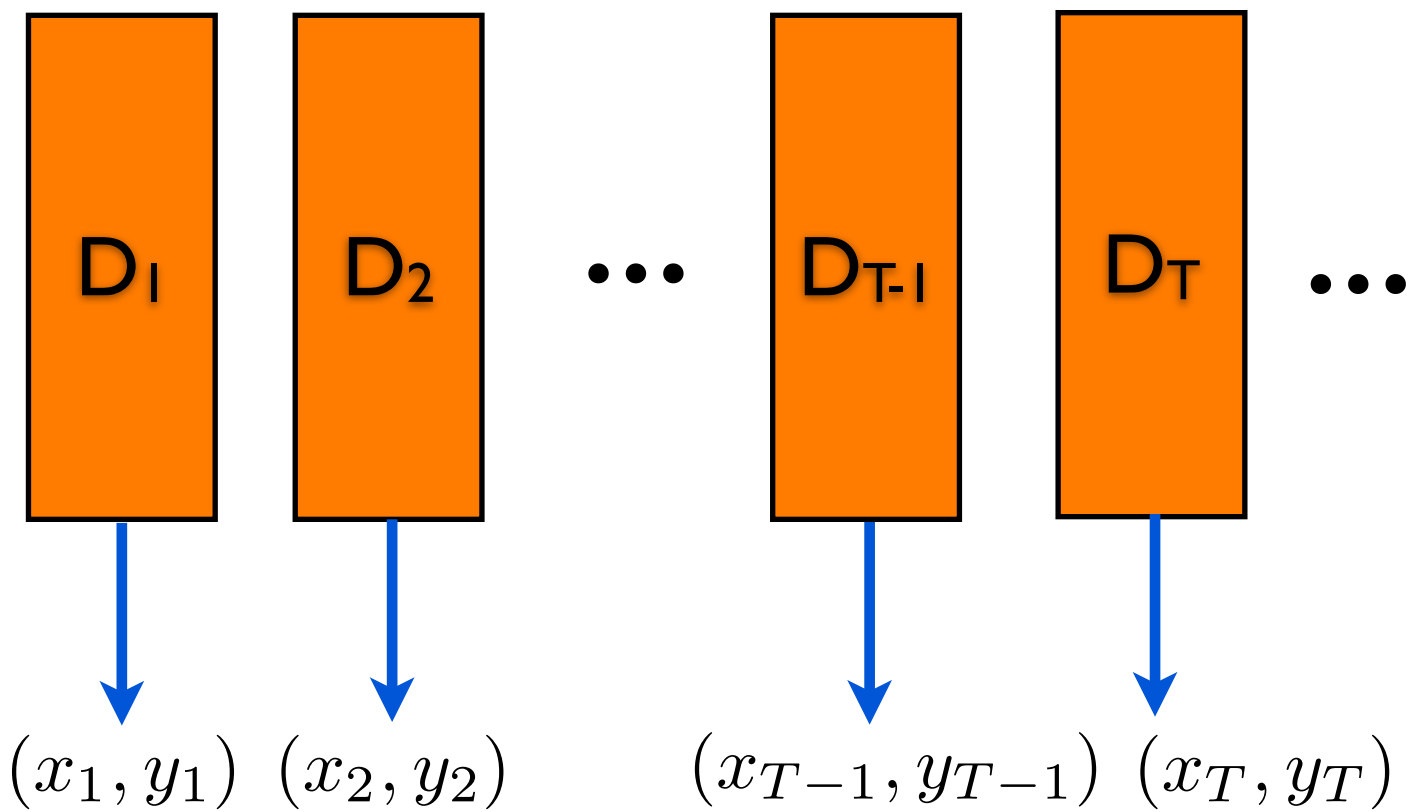
- **Theorem**: assume that the loss L is bounded by M. Let $D_1, \ldots, D_{T+1}$ be a sequence of distributions and let $H_L = \{(x, y) \mapsto L(h(x), y) \colon h \in H\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathcal{L}_{D_{T+1}}(h) \leq \frac{1}{T} \sum_{t=1}^{T} L(h(x_t), y_t) + 2\mathfrak{R}_T(H_L) + \frac{1}{T} \sum_{t=1}^{T} \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + M \sqrt{\frac{\log \frac{2}{\delta}}{2T}}.$$

Drifting Tracking Scenario

[Bartlett, COLT 1992]

- Model: perpetual training and testing.



- tracking: $\sup_{(D_t)} \{ \mathop{\mathrm{E}}_{\substack{(x_t, y_T) \\ \sim D_T}} [L(h_{t-1}(x_T, y_T)] : \forall t, \mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta \} < \inf_{h \in H} \mathcal{L}_{D_T}(h) + \epsilon.$

Drifting Tracking Scenario

- Expected mistake:

$$M_T(D) = \mathrm{E}_{S\sim D}[\widehat{M}_T] = \mathrm{E}_{S\sim D}[L(h_{T-1}(x_T), y_T)].$$

  - supremum over all distributions $D = \bigotimes_{t=1}^{T} D_t$:

$$\widetilde{M}_T = \sup_{\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1})<\Delta} M_T(D).$$

- Algorithm $\mathcal{A}$ (Δ, ϵ) - tracks H if for T sufficiently large,

$$\widetilde{M}_T < \inf_{h\in H} \mathcal{L}_{D_T}(h) + \epsilon.$$

Drifting Tracking - Guarantee

- Theorem: H hypothesis set with VC-dimension d and $\mathfrak{R}_T(H_L) \leq O(\sqrt{d/T})$; sequence of distributions such that $\mathrm{disc}_{\mathcal{Y}}(D_t, D_{t+1}) \leq \Delta$ for all t. If $\Delta = O(d\epsilon^3)$, there exists an algorithm that (Δ, ϵ)-tracks H.

  - improvement over previous L_1 result of (Long, (1999).

General On-Line to Batch

- **Theorem**: assume that the loss L is bounded by M and that it is convex with respect to its first argument. Let $h_1, \ldots, h_T$ be the hypotheses returned by algorithm $\mathcal{A}$ and $h = \sum_{t=1}^{T} w_t h_t$. Then, with probability at least $1 - \delta$,

$$\mathcal{L}_{D_{T+1}}(h) \leq \sum_{t=1}^{T} w_t L(h_t(x_t), y_t) + \sum_{t=1}^{T} w_t \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{1}{\delta}}$$

$$\mathcal{L}_{D_{T+1}}(h) \leq \inf_{h \in H} \mathcal{L}(h) + \frac{R_T}{T} + \sum_{t=1}^{T} w_t \mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + M\|\mathbf{w} - \mathbf{u}_0\|_1 + 2M\|\mathbf{w}\|_2 \sqrt{2 \log \frac{2}{\delta}}.$$

Drifting Algorithm

- Idea: find best mixture weight.

$$\min_{\mathbf{w}} \quad \lambda \|\mathbf{w}\|_2^2 + \sum_{t=1}^{T} w_t \left(\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) + L(h_t(x_t), y_t)\right)$$

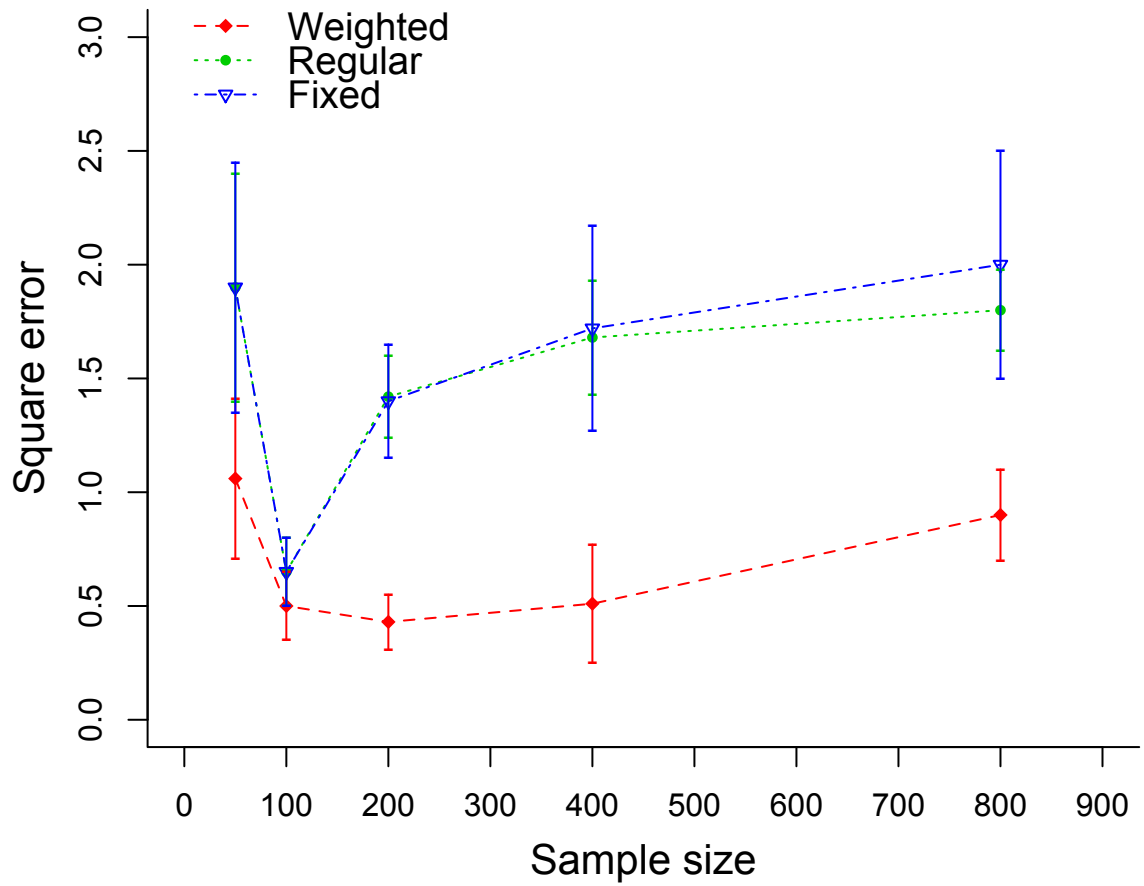
$$\text{subject to:} \quad \left(\sum_{t=1}^{T} w_t = 1\right) \wedge (\forall t \in [1, T], w_t \geq 0).$$

  - deterministic setting:

$$\mathrm{disc}_{\mathcal{Y}}(D_t, D_{T+1}) \leq \mathrm{disc}(D_t, D_{T+1}).$$

  - estimation of discrepancies using independent samples.

Experiments



- base on-line algorithm: Widrow-Hoff.
- uniform weights, uniform over last 100, and weighted algorithms.

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.