

Domain Adaptation Theory and Algorithms

Mehryar Mohri

Courant Institute and Google

mohri@cims.nyu.edu

Joint work with:

Corinna Cortes (Google)

Yishay Mansour (Tel-Aviv) and

Afshin Rostamizadeh (UC Berkeley).

Standard Learning Assumptions

- IID assumption.
- Same distribution for training and test.
- Distributions fixed over time.

Modern Large-Scale Data Sets

- Real-world applications:
 - Sample points are not drawn IID.
 - Training sample is biased.
 - Training points with uncertain labels.
 - Multiple training sources.
 - Distribution may drift with time.

These problems **must be addressed** for learning to be effective.

Domain Adaptation - Problem

■ Input:

- Labeled data from source domain.
- Unlabeled data from target domain.

■ Problem: use labeled and unlabeled data to derive hypothesis h with good performance on target domain.

- Thus, harder generalization problem than standard learning problem!

Domain Adaptation - Examples

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

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. (2006) and Blitzer et al. (2007)].

■ Multiple-source:

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

Existing Techniques?

- Importance weighting technique.
 - applications, e.g., sample bias correction.
 - experiments: method does not always work.
 - can lead to performance degradation!
 - recent new analysis and learning guarantees (Cortes, Mansour, and MM, NIPS 2010).

This Talk

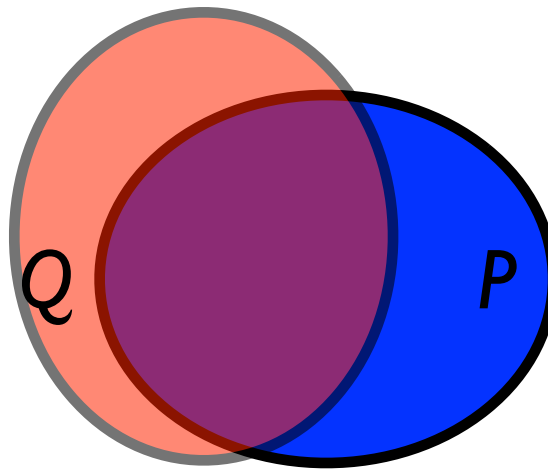
- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments

Learning Set-up

- **Distributions:** source Q , target P .
- **Target function(s):** f , or f_Q and f_P .
- **Input:** labeled sample drawn from Q , unlabeled sample drawn from P .
- **Problem:** find hypothesis h with small expected loss with respect to distribution P ,

$$\mathcal{L}_P(h, f) = \mathbb{E}_{x \sim P} \left[L(h(x), f(x)) \right].$$

Distribution Mismatch



Which distance should we use
to compare these distributions?

Simple Analysis

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

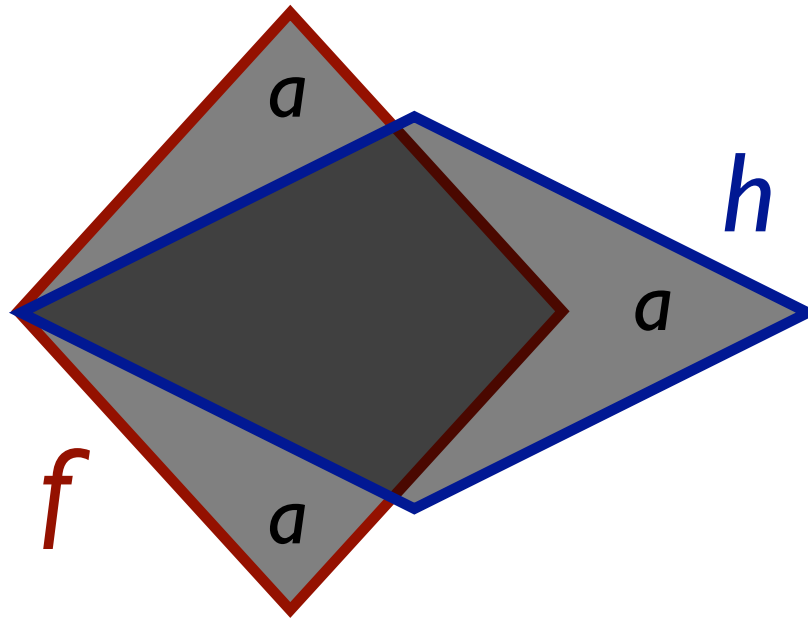
$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M l_1(Q, P).$$

■ **Proof:**

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

But, is this bound informative?

Example - 0/1 Loss



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

d_A distance

■ Definition:

$$d_A(Q_1, Q_2) = \sup_{a \in A} |Q_1(a) - Q_2(a)|.$$

where A is a set of regions or subsets of X
[Devroye et al. (1996), Kifer et al. (2004)], Ben-David et al. (2007),
Blitzer et al. (2007)].

■ For 0/1 loss, the natural choice is the set of all possible disagreement regions:

$$A = H \Delta H = \{|h' - h| : h, h' \in H\}.$$

Discrepancy Distance

■ Definition:

$$\text{disc}(Q_1, Q_2) = \max_{h, h' \in H} \left| \mathcal{L}_{Q_1}(h', h) - \mathcal{L}_{Q_2}(h', h) \right|.$$

- Relationship with discrepancy in combinatorial contexts [Chazelle (2000)].
- d_A is a special case, 0-1 loss.
- helps compare distributions for other losses, e.g. hinge loss, L_p loss.
- symmetric, verifies triangle inequality, in general not a distance.

Discrepancy - Properties

- **Theorem:** the discrepancy distance can be estimated from finite samples for H with finite VC dimension d . For L_q loss, $L_q(y, y') = |y - y'|^q$, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned} \text{disc}(P, Q) \\ \leq \text{disc}(\hat{P}, \hat{Q}) + O\left(q\sqrt{\frac{d}{m}} + q\sqrt{\frac{d}{n}}\right) + 3M\left(\sqrt{\frac{\log \frac{4}{\delta}}{2m}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n}}\right). \end{aligned}$$

This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments

Theoretical Guarantees

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

$$\begin{aligned} \mathcal{L}_P(h, f_P) &\leq \mathcal{L}_Q(h, h_Q^*) + \mathcal{L}_P(h_P^*, f_P) + \text{disc}(P, Q) \\ &\quad + \mathcal{L}_Q(h_Q^*, h_P^*). \end{aligned}$$

Some Special Cases

■ When $h^* = h_Q^* = h_P^*$,

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

■ When $f_P \in H$ (consistent case),

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

Kernel-Based Reg. Algorithms

- Algorithms minimizing objective function:

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

where K is a positive definite symmetric kernel,
 $\lambda > 0$ is a trade-off parameter, and
 $\hat{R}_{\hat{Q}}(h)$ the empirical error of h .

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

Guarantees for KBR Algorithms

- **Theorem:** let K be a positive definite symmetric kernel with $\forall x, K(x, x) \leq \kappa$ and the loss s.t. $L(\cdot, y)$ is σ -Lipschitz. Assume that $f_P \in H$ and that f_P and f_Q coincide on the training sample. Then, for all $x \in X, y \in Y$,

$$|L(h'(x), y) - L(h(x), y)| \leq \kappa \sigma \sqrt{\frac{\text{disc}(\hat{P}, \hat{Q})}{\lambda}}.$$

Guarantees for KBR Algorithms

- **Theorem:** same assumptions but f_P and f_Q potentially different on the training sample, H bounded by M , and L the square loss; then, for all $x \in X, y \in Y$,

$$\begin{aligned} & |L(h'(x), y) - L(h(x), y)| \leq \\ & \frac{2\kappa M}{\lambda} \left(\kappa\delta + \sqrt{\kappa^2\delta^2 + 4\lambda \text{disc}_L(\hat{P}, \hat{Q})} \right), \end{aligned}$$

with $\delta^2 = L_{\hat{Q}}(f_Q(x), f_P(x)) \ll 1$.

Empirical Discrepancy

- Discrepancy distance $\text{disc}(\hat{P}, \hat{Q})$ 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 problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments

Algorithm - Idea

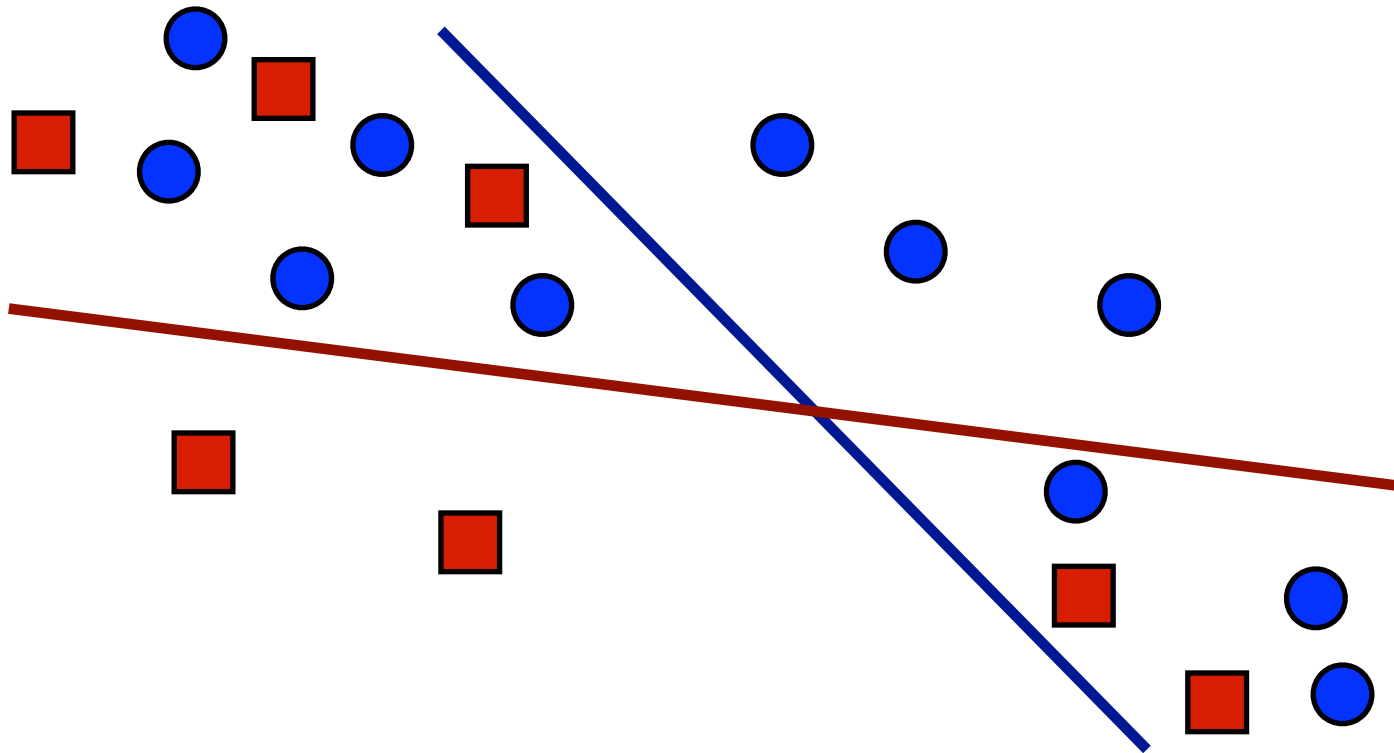
- The training sample is given, but we can search for a new empirical distribution \hat{Q}' such that

$$\hat{Q}' = \operatorname{argmin}_{\hat{Q}' \in \mathcal{Q}} \operatorname{disc}(\hat{P}, \hat{Q}'),$$

where \mathcal{Q} is the set of distributions with support $\operatorname{supp}(\hat{Q})$.

- can be interpreted as *reweighting* training points.

Case of Halfspaces



Min-Max Problem

■ Reformulation:

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

- game theoretical interpretation.
- gives lower bound:

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

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

Classification - 0/1 Loss

■ Problem:

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

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

Classification - 0/1 Loss

■ Linear program (LP):

$$\min_{Q'} \delta$$

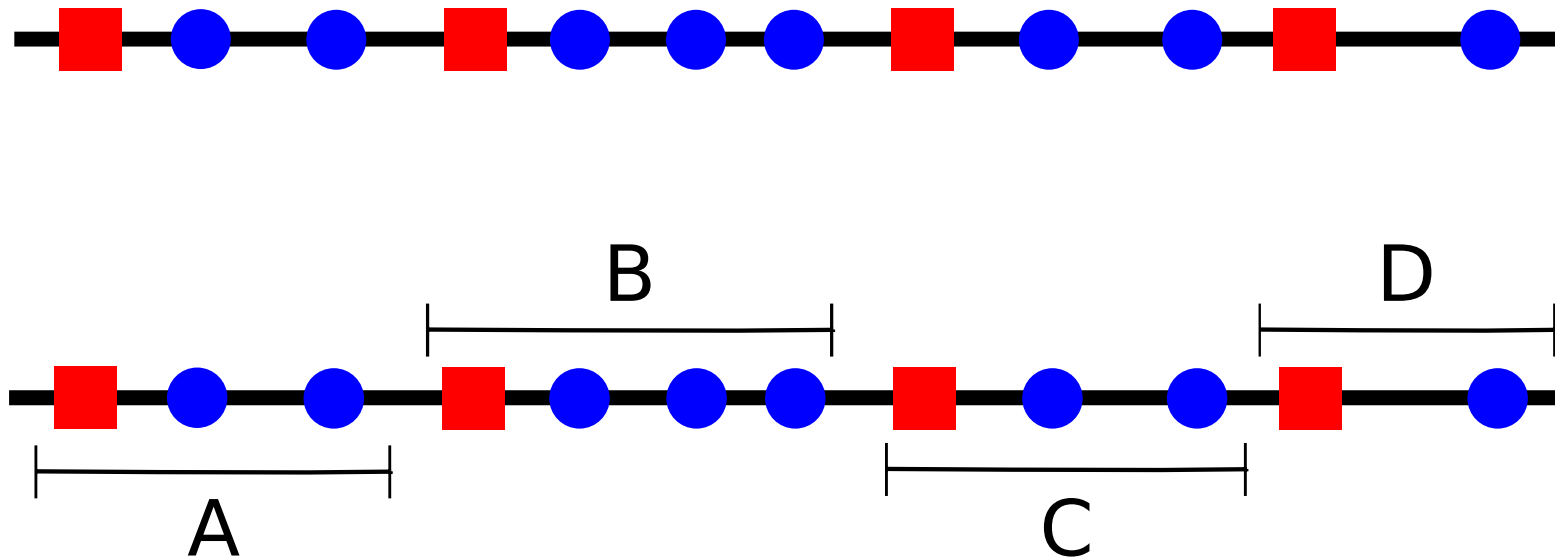
$$\text{subject to } \forall a \in H \Delta H, \hat{Q}'(a) - \hat{P}(a) \leq \delta$$

$$\forall a \in H \Delta H, \hat{P}(a) - \hat{Q}'(a) \leq \delta$$

$$\forall x \in S_Q, \hat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \hat{Q}'(x) = 1.$$

- No. of constraints bounded by shattering coefficient $\Pi_{H \Delta H}(m_0 + n_0)$.

Algorithm - ID



Regression - L2 Loss

■ Problem:

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

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:

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

elements of $\text{supp}(\hat{Q})$



Regression - L2 Loss

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

$$\min_{\mathbf{z}, \lambda} \quad \lambda$$

$$\text{subject to} \quad \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0$$

$$\mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0,$$

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

$$\mathbf{M}_0 = \mathbf{K}^{1/2} \text{diag}(P(s_1), \dots, P(s_{p_0})) \mathbf{K}^{1/2}$$

$$\mathbf{M}_i = \mathbf{K}^{1/2} \mathbf{I}_i \mathbf{K}^{1/2}.$$

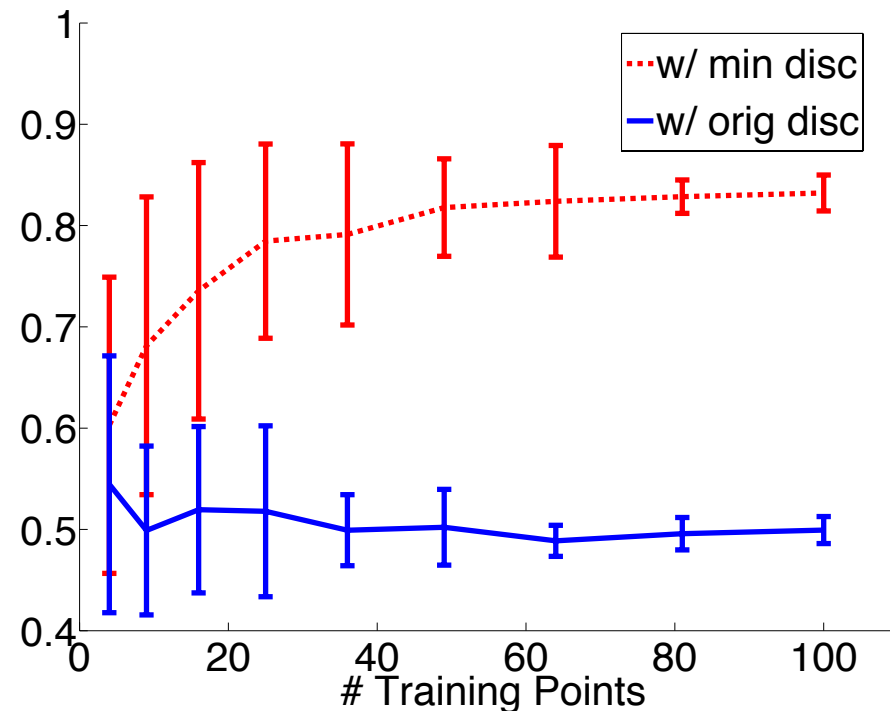
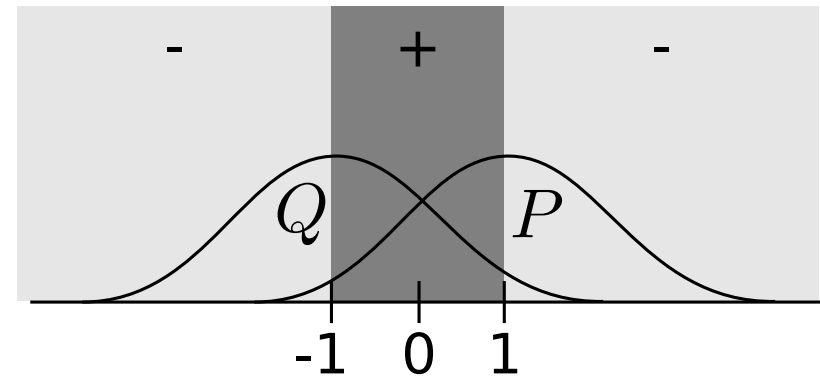
This Talk

- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments

Experiments

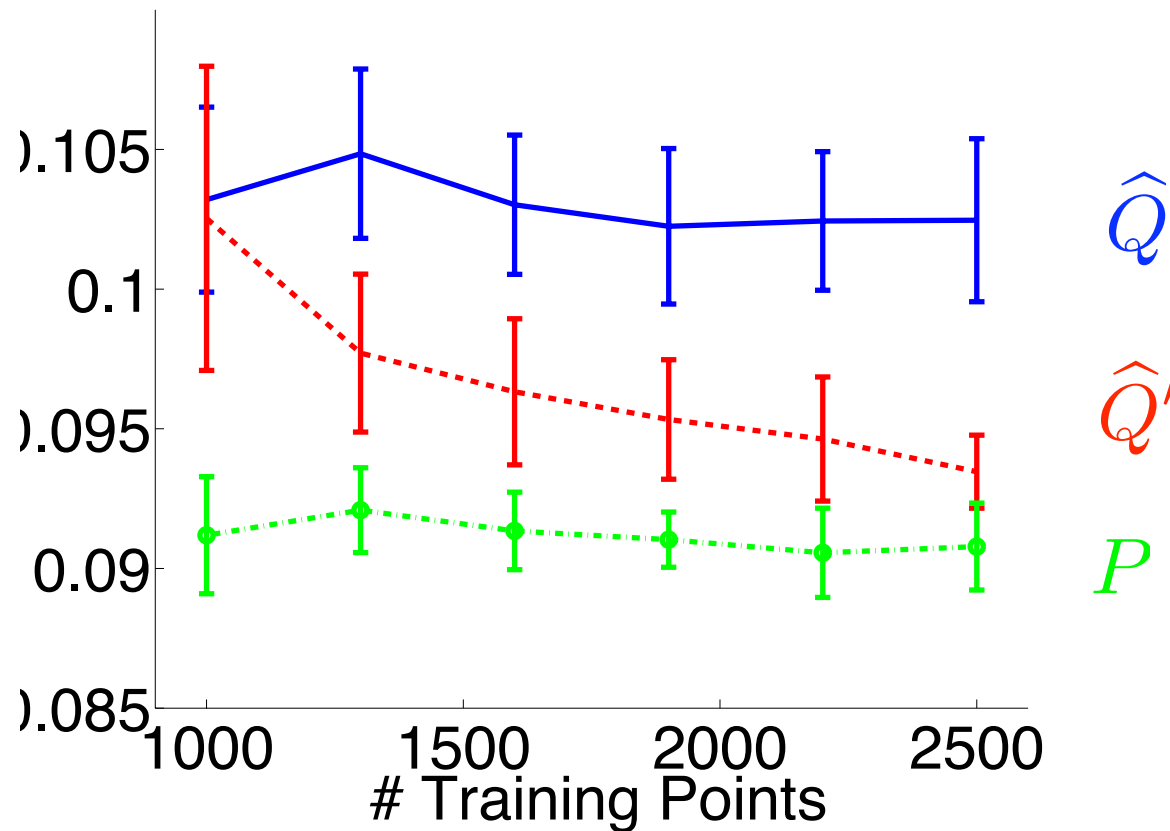
■ Classification:

- 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.

Conclusion

- **Discrepancy distance:** appears as the ‘right’ measure of difference of distributions for adaptation.
- **Theoretical analysis:** generalization bounds and strong guarantees for a large class of algorithms.
- **Algorithm:** discrepancy minimization algorithms for other loss functions, more efficient large-scale algorithms.