Learning with Imperfect Data

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

Joint work with:
Yishay Mansour (Tel-Aviv & Google) and
Afshin Rostamizadeh (Courant Institute).

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)].

This Talk

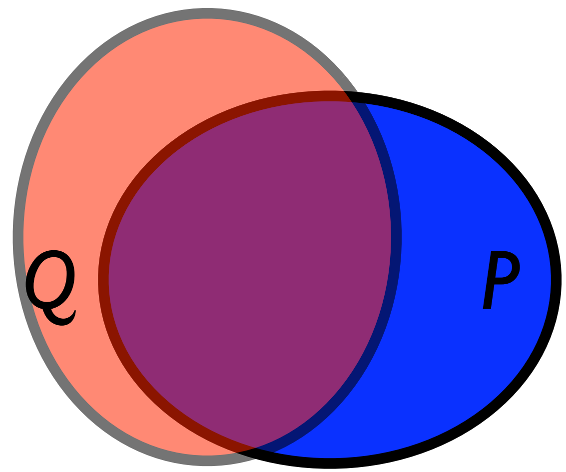
- Domain adaptation problem
- Discrepancy distance
- Theoretical guarantees
- Algorithm
- Experiments

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,

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

Distribution Mismatch



Which distance should we use
to compare these distributions?

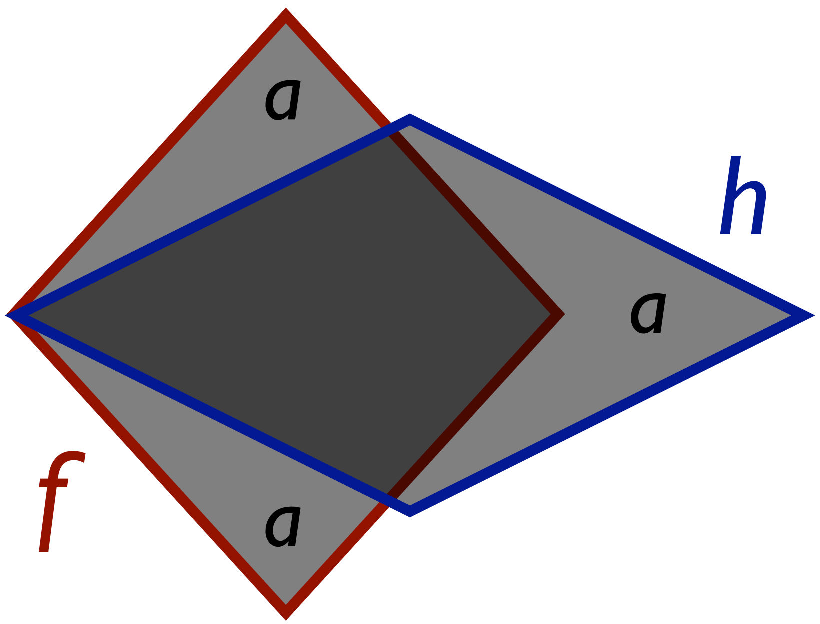
Simple Analysis

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

But, is this bound informative?

Example - 0/1 Loss



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

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:

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

Discrepancy Distance

- Definition:

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

- 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. For L_q loss, $L_q(y, y') = |y - y'|^q$, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathrm{disc}(P, Q) \leq \mathrm{disc}(\widehat{P}, \widehat{Q}) + 4q\Big(\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + \widehat{\mathfrak{R}}_{\mathcal{T}}(H)\Big) + 3M\left(\sqrt{\frac{\log\frac{4}{\delta}}{2m}} + \sqrt{\frac{\log\frac{4}{\delta}}{2n}}\right).$$

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:

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

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),

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

Kernel-Based Reg. Algorithms

- Algorithms minimizing objective function:

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

where K is a positive definite symmetric kernel,
$\lambda > 0$ is a trade-off parameter, and
$\widehat{R}_{\widehat{Q}}(h)$ 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 $x \in X, y \in Y$,

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

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 $x \in X, y \in Y$,

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

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

Empirical Discrepancy

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

Algorithm - Idea

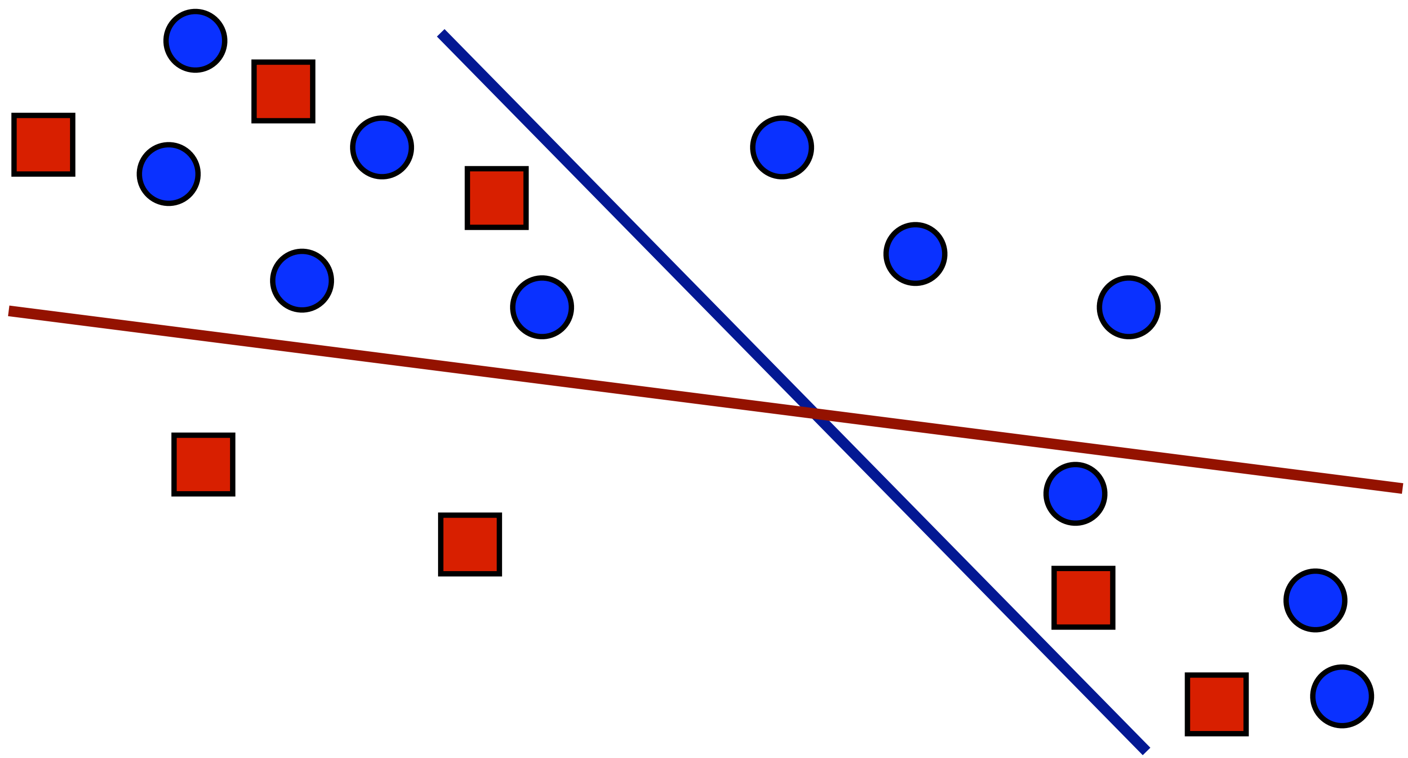
- The training sample is given, but we can search for a new empirical distribution $\widehat{Q}'$ such that

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

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

  - can be interpreted as *reweighting* training points.

Case of Halfspaces



Min-Max Problem

- Reformulation:

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

  - game theoretical interpretation.
  - gives lower bound:

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

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

Classification - 0/1 Loss

- Problem:

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

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

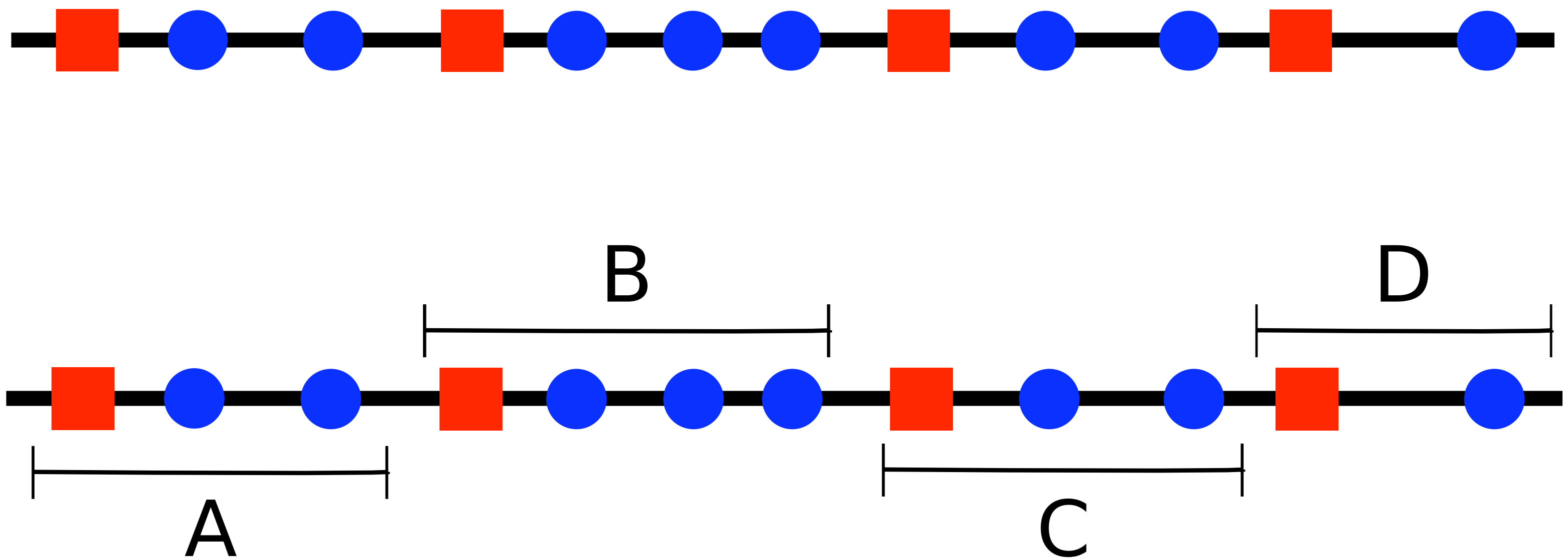
Classification - 0/1 Loss

- 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:

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

Regression - L2 Loss

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

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

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

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

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

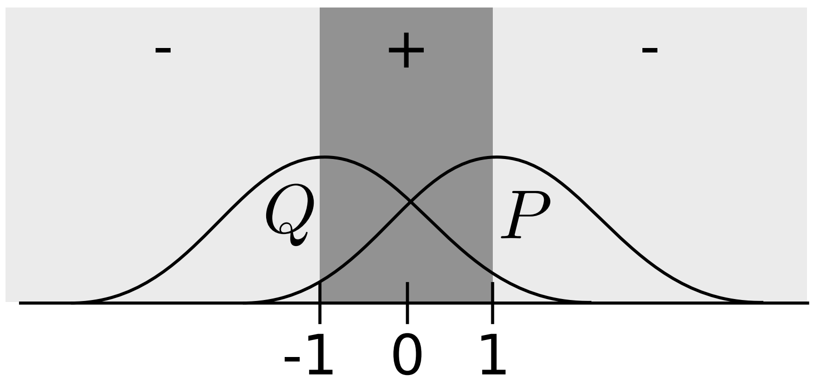
$$\mathbf{M}_i = \mathbf{K}^{1/2}\mathbf{I}_i\mathbf{K}^{1/2}.$$

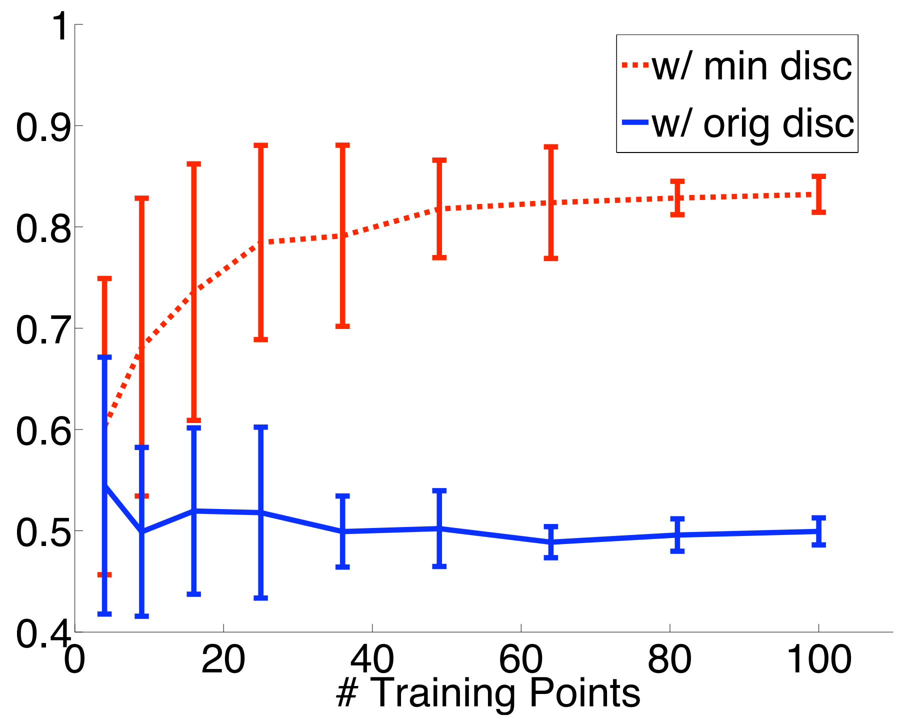
This Talk

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

Experiments

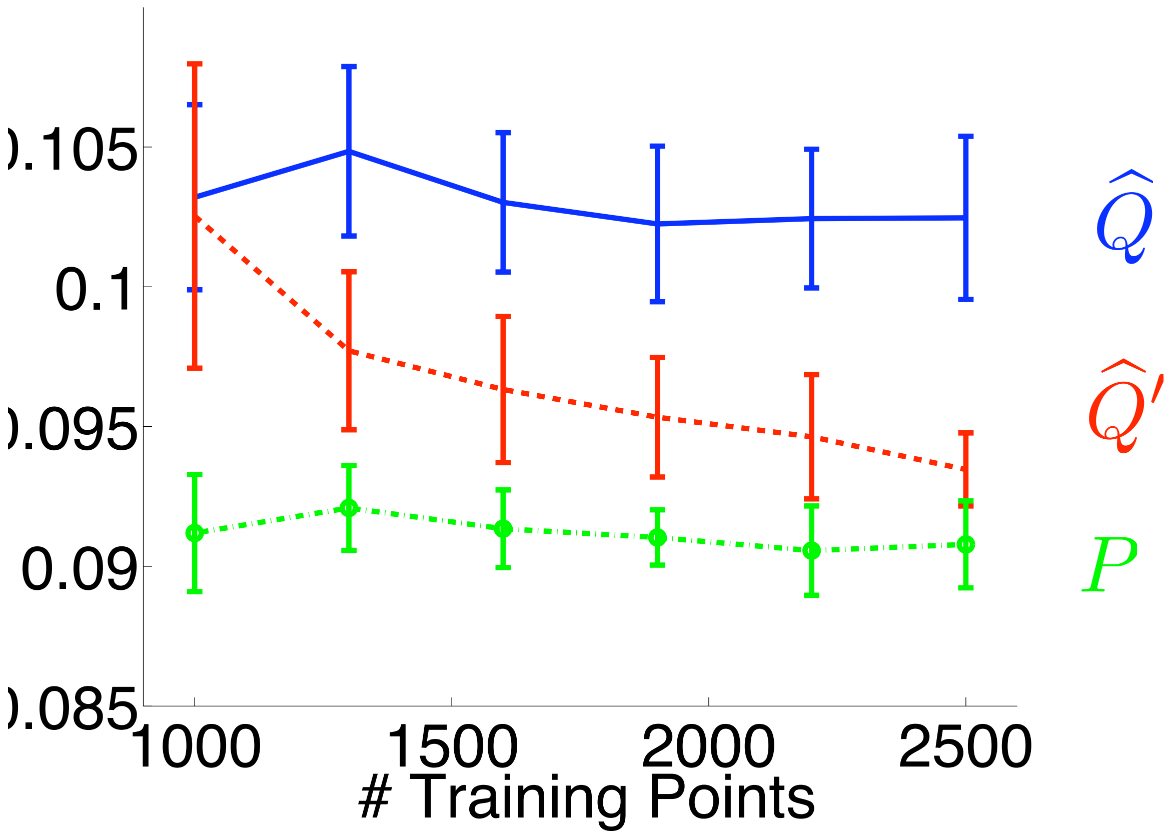
- Classification:
  - *Q* and *P* Gaussians.
  - *H*: halfspaces.
  - *f*: interval [-1, +1].





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.