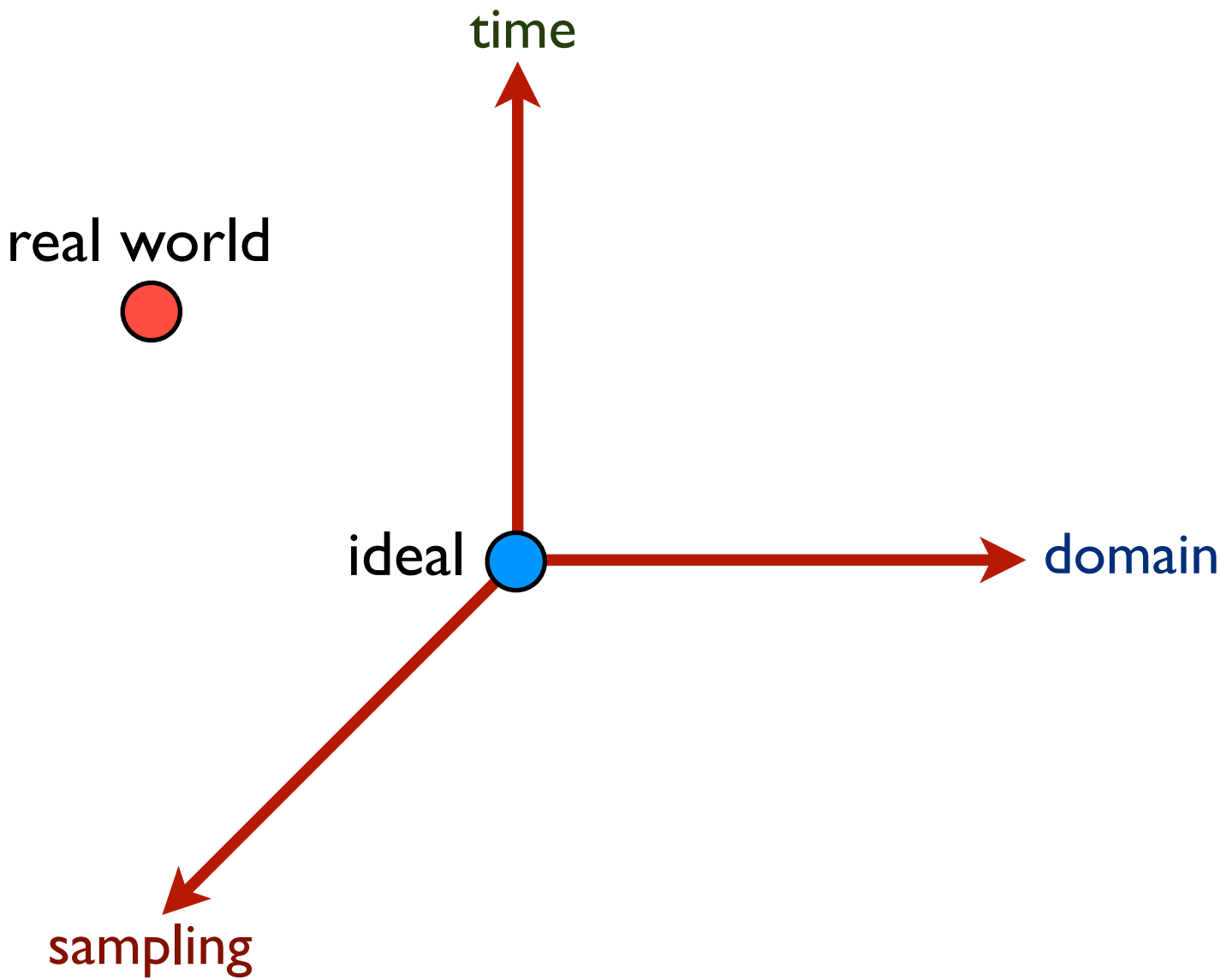
Recent Theoretical and Algorithmic Advances in Domain Adaptation

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 vs Real World



This Talk

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

Domain Adaptation

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

→ Solution critical for applications.

Domain Adaptation Problem

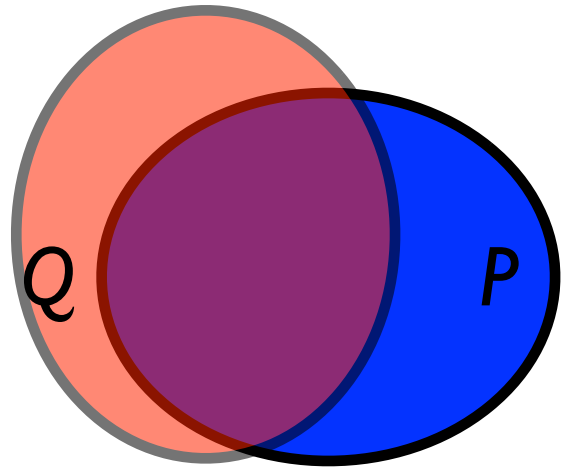
- Domains: source (Q, f_Q), target (P, f_P).
- Input:
  - labeled sample S drawn from source.
  - unlabeled sample T drawn from target.
- Problem: find hypothesis h in H with small expected loss with respect to target domain, that is

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

Some 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)].

Distribution Mismatch



Which distance should we use to compare these distributions?

Simple Analysis

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

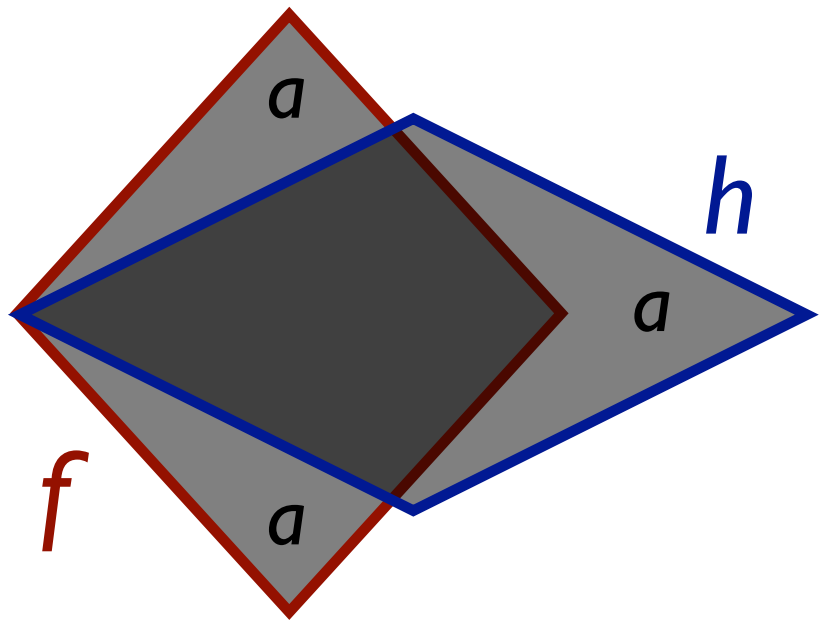
$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M\, L_1(Q, P).$$

- **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 - 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 $1-\delta$,

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

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**: $\Psi : h \mapsto \mathrm{E}_{x\sim P}[h^2(x)] - \mathrm{E}_{x\sim Q}[h^2(x)]$ 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 $P = Q$.

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) 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 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 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),

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

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

- 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.

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?

Algorithm - Idea

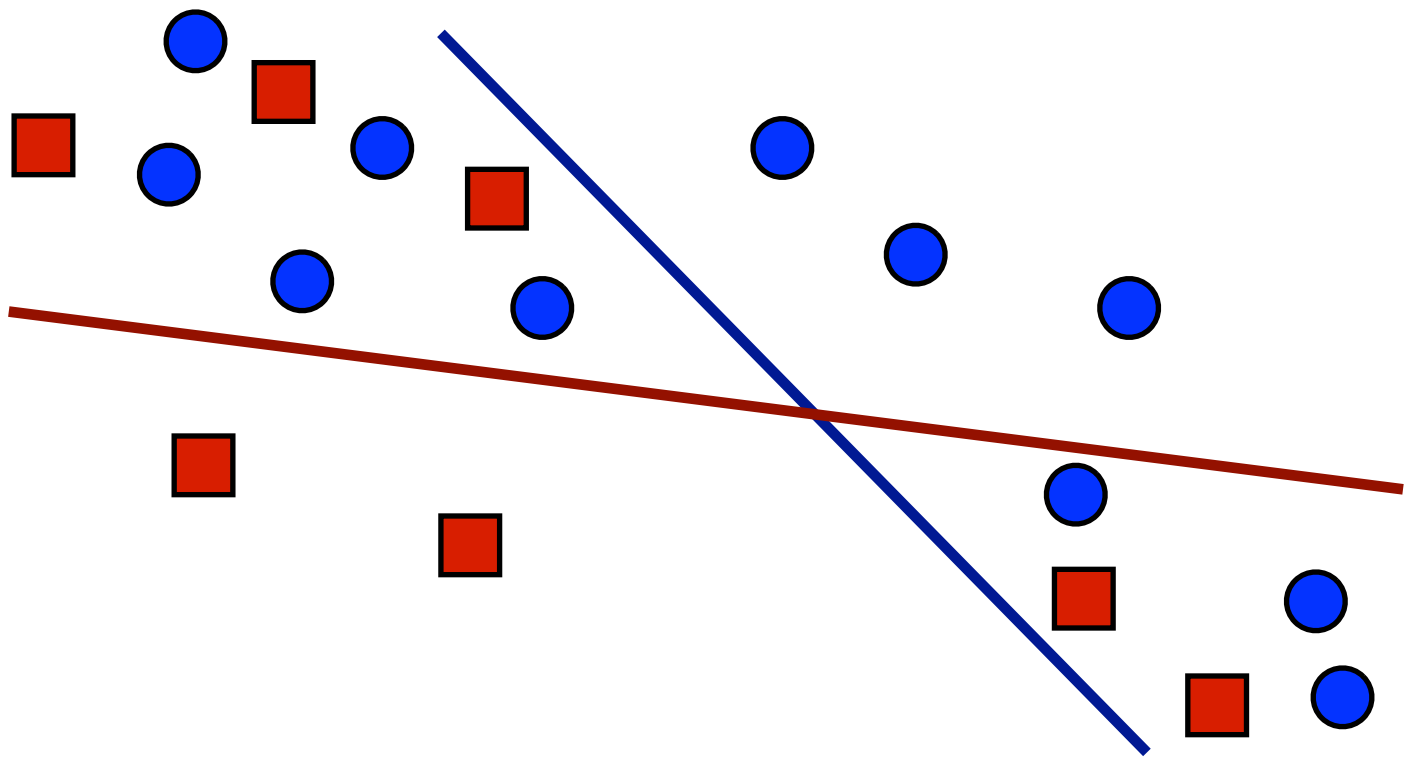
- Search for a new empirical distribution q^* with same support:

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

- Solve modified 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



Discrepancy Minimization 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

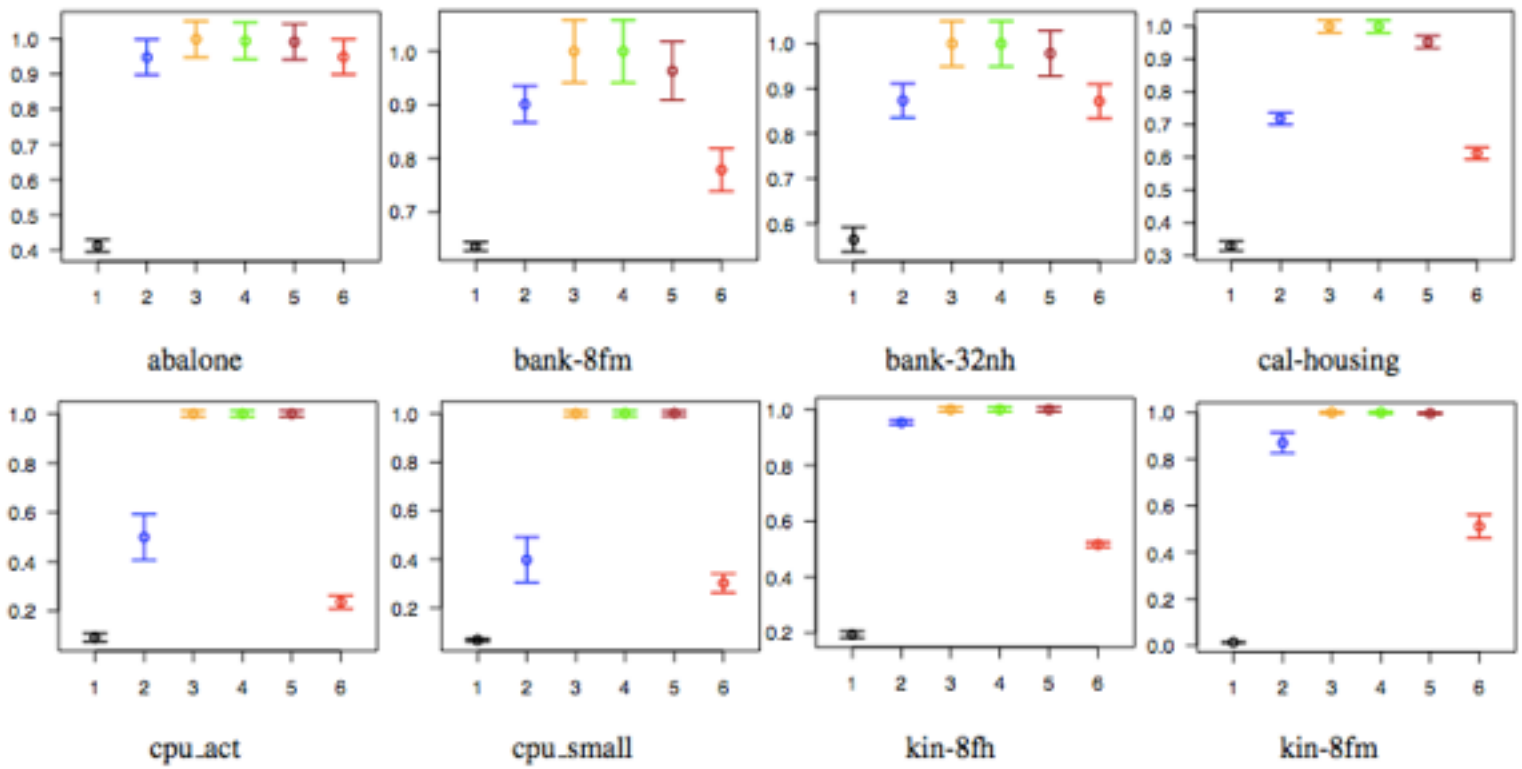


Fig. 11. Results with "easy-to-learn" biasing scheme: Relative MSE performance of (1): Optimal (in black); (2): KMM (in blue); (3): KLIEP (in orange); (4): Uniform (in green); (5): Two-Stage (in brown); and (6): DM (in red). Errors are normalized so that the average MSE of Uniform is 1.

Enhancement

[Cortes, MM, and Munoz (2014)]

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

Algorithm

[Cortes, MM, and Munoz (2014)]

- Choose Q_h such that objectives are unif. close:

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

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

- Ideally:

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

- Using convex surrogate H'' :

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

Optimization

[Cortes, MM, and Munoz (2014)]

$$\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \operatorname*{argmin}_{l \in \{\mathcal{L}_{\mathbf{q}}(h, f_Q) : \mathbf{q} \in \mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})\}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')|$$

$$= \operatorname*{argmin}_{l \in \mathbb{R}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')|$$

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

→ Convex optimization problem:

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

Convex Surrogate Hypothesis Set

[Cortes, MM, and Munoz (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

- Theorie 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.