Domain Adaptation for Regression

Corinna Cortes
Google Research
corinna@google.com

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

Motivation

- Applications: distinct training and test distributions.
  - 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.

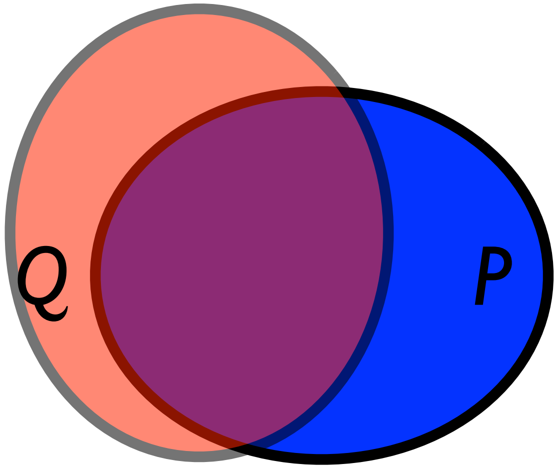
→ Solution critical for applications.
This talk: regression problems.

Domain Adaptation Problem

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

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

Distribution Mismatch



Which distance should we use
to compare these distributions?

Discrepancy Distance

(Mansour, MM, Rostami, 2009)

- Definition:

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

  - symmetric, verifies triangle inequality, in general not a distance.
  - helps compare distributions for arbitrary losses, e.g. hinge loss, or L_p loss.
  - can be estimated from finite samples, Rademacher complexity bounds.

Previous Work

- (Ben-David et al., NIPS 2006) & (Blitzer et al., NIPS 2007): bounds for binary classification based on d_A distance and λ_H term (cannot be estimated).
- (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): series of negative results for adaptation in binary classification.

Theoretical Guarantees

- Two types of questions:
  - difference between average loss of hypothesis h on Q versus P?
  - difference of loss between hypothesis h obtained when training on $(\widehat{Q}, f_Q)$ versus hypothesis h' obtained when training on $(\widehat{P}, f_P)$.

Kernel-Based Reg. (KBR) Algorithms

- Algorithms minimizing 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

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

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

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

Adaptation Algorithm

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

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

- 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 L2 loss and $H = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \colon \|\mathbf{w}\| \leq \Lambda\}$, can be cast as an SDP (Mansour, MM, Rostami, COLT 2009):

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

→ what about if we want to use kernels?

Discrepancy Min. with Kernels

- For L2 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.

Smooth Approximation

(Nesterov, 1983, 2005)

- Convex optimization problem: $\mathrm{minimize}_{\mathbf{z} \in C} F(\mathbf{z})$.
- Smooth:
  - C closed convex, F Lipschitz continuous gradient.
  - algorithm: $O(1/\sqrt{\epsilon})$, optimal for problem class.
- Non-smooth:
  - F Lipschitz continuous.
  - find G uniform ϵ-approximation of F.
  - algorithm: $O(1/\epsilon)$.

Disc. Min. SDP Problem

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

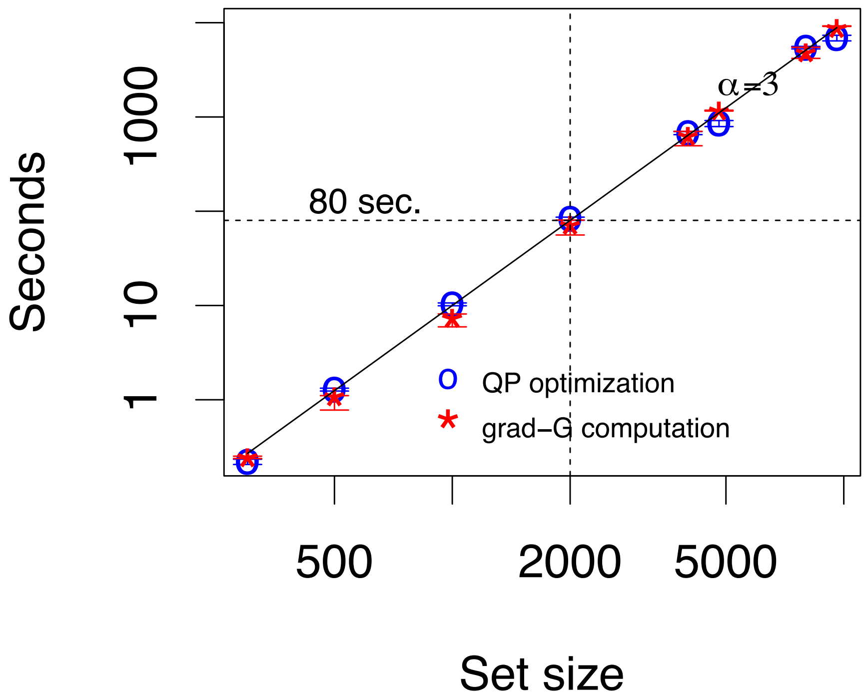
Algorithm 2

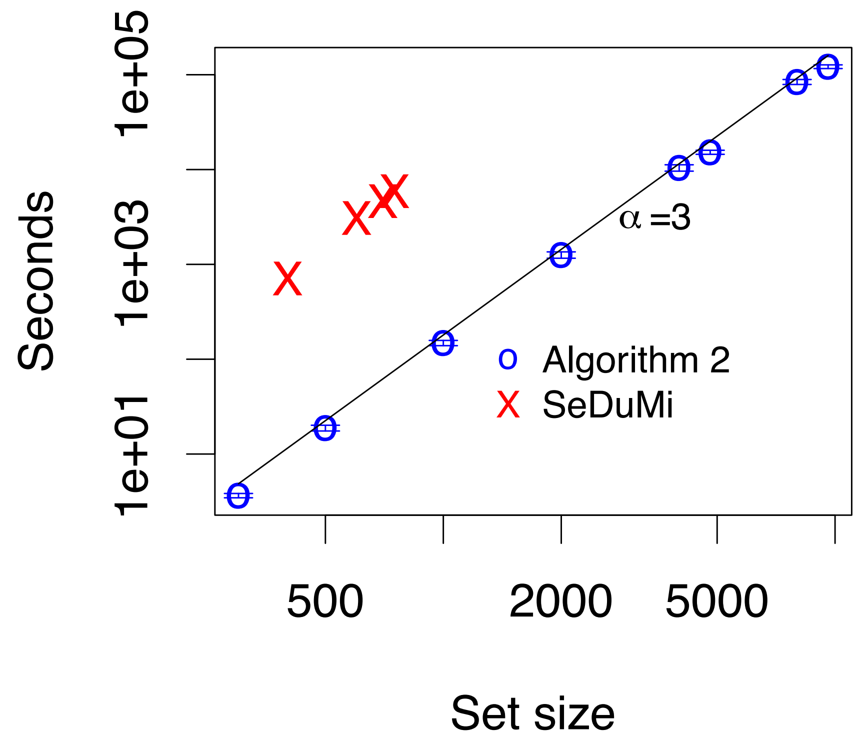
$\mathbf{u}_0 \leftarrow \mathrm{argmin}_{\mathbf{u} \in C}\, \mathbf{u}^\top \mathbf{J} \mathbf{u}$
for $k \geq 0$ **do**
  $\mathbf{v}_k \leftarrow \mathrm{argmin}_{\mathbf{u} \in C}\, \frac{2p-1}{2}(\mathbf{u} - \mathbf{u}_k)^\top \mathbf{J}(\mathbf{u} - \mathbf{u}_k) + \nabla G_p(\mathbf{M}(\mathbf{u}_k))^\top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

Convergence Guarantee

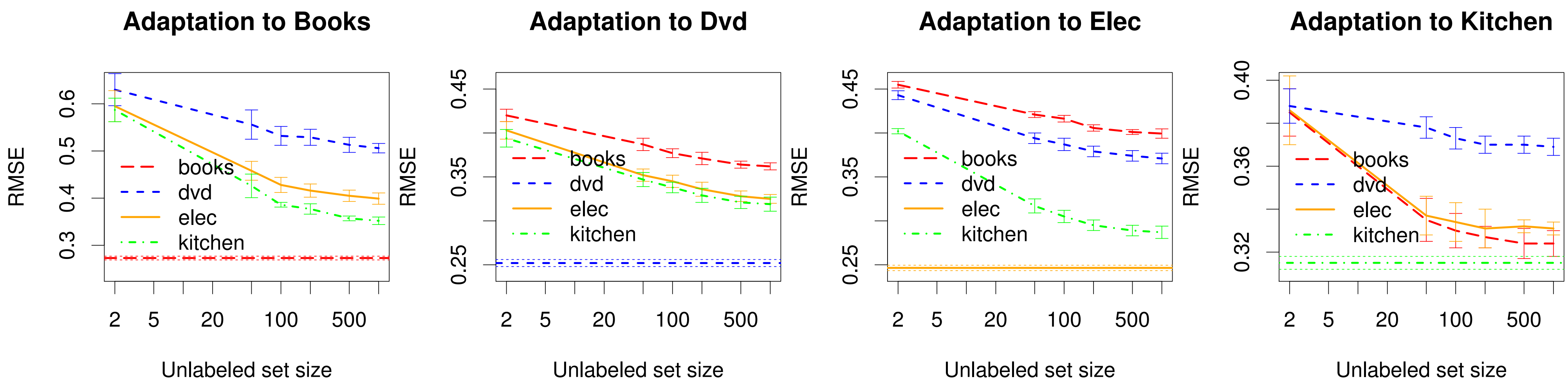
- Let $r = \max_{\mathbf{z} \in C} \operatorname{rank}(\mathbf{M}(\mathbf{z})) \leq \max\{N, \sum_{i=0}^{n} \operatorname{rank}(\mathbf{M}_i)\}$.
- Theorem: for any $\epsilon > 0$, the algorithm solves the discrepancy minimization SDP with relative accuracy ϵ in $O(\sqrt{r \log r}/\epsilon)$ iterations.

Experiments - Time





Experiments - Performance



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

Conclusion

- Theoretical results for DA in regression.
  - new pointwise loss guarantees for general class of loss functions.
  - disc. min. adaptation extended to kernels.
- Efficient algorithm for solving discrepancy minimization.
  - shown to scale to relatively large data sets.
  - empirically shown to be effective.
- Still many adaptation questions left to address!