Learning in a Non-Ideal World

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

Joint work with Yishay Mansour and Afshin Rostami.

Standard Learning Assumptions

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

Modern Large Data Sets

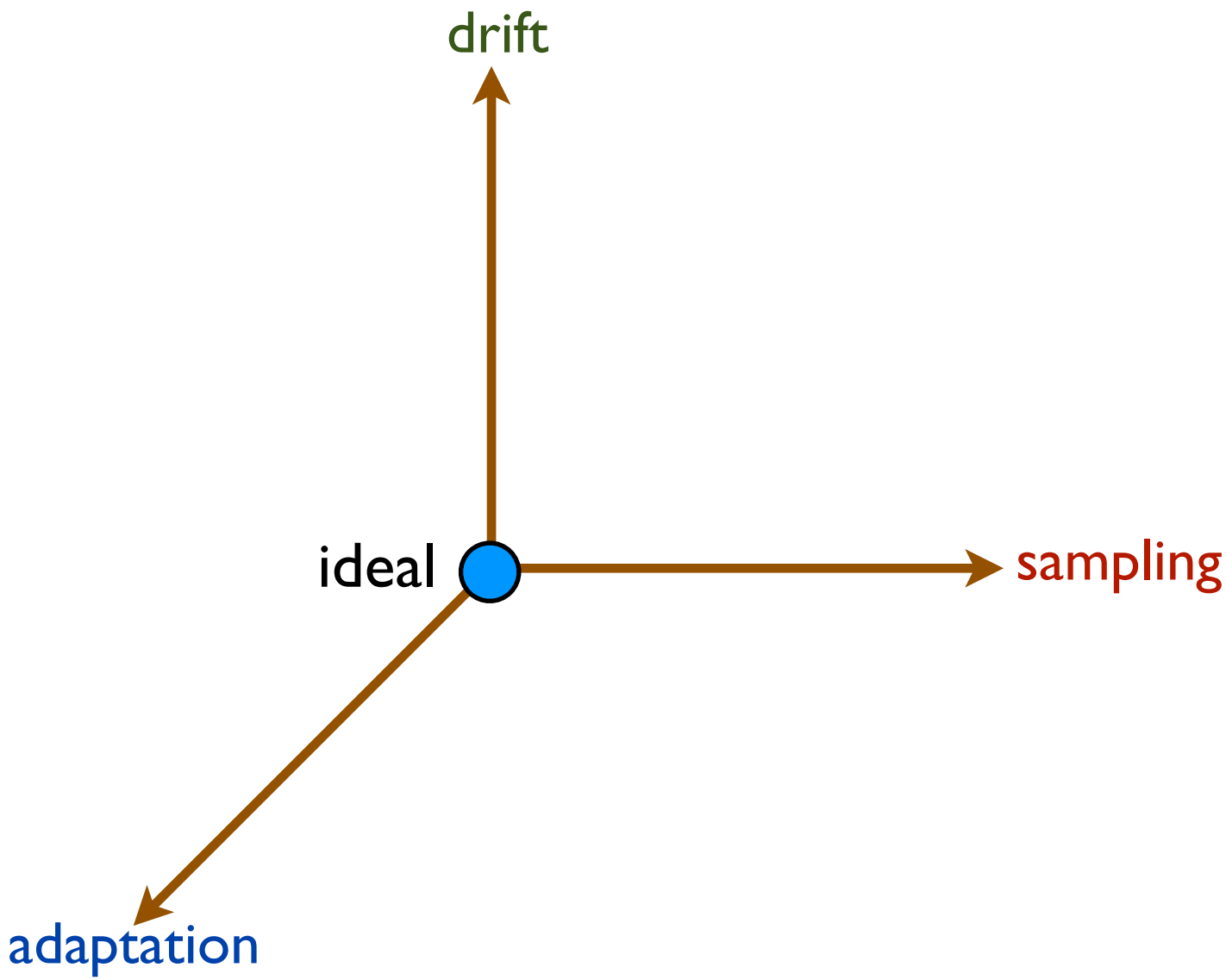
- Real-world applications:
  - Sample points are not drawn IID.
  - Training sample is biased.
  - Training points with uncertain labels.
  - Multiple training sources.
  - Distribution may drift with time.

These problems must be addressed for learning to be effective.

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).

Non-Ideal World



This Talk

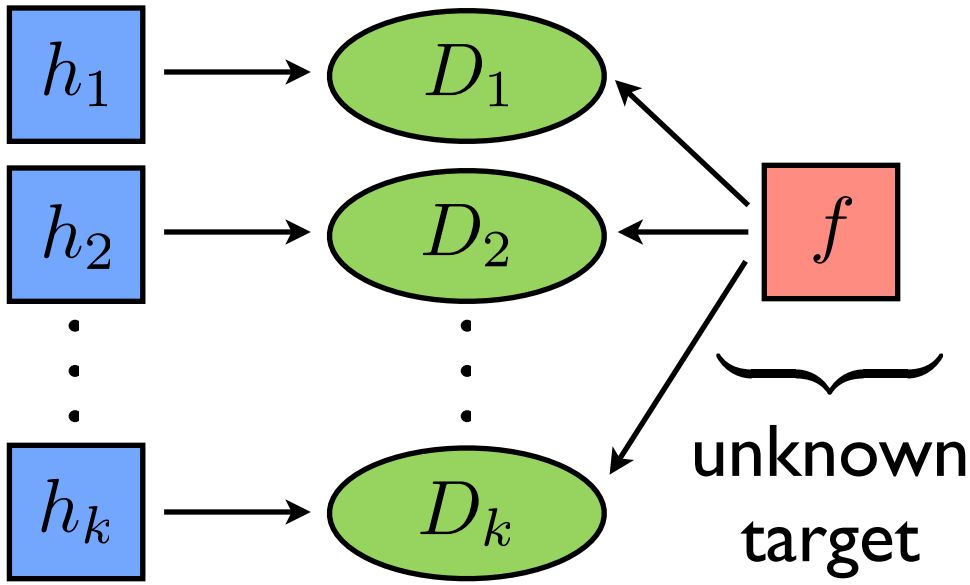
- Multiple source adaptation problem.
- Related work:
  - multiple sources but distinct problem: same input distribution but different labels, modulo disparity constraints (Crammer, Kearns, and Wortman, 2008).
  - single-source adaptation problem, e.g., (Blitzer, Crammer, Kulesza, Pereira, and Wortman, 2008), (Mansour, MM, and Rostami, 2009).

Motivation

- Adaptation: the problem of training and testing on differing but somewhat similar distributions.
- A natural ability for humans, can adapt to new tasks based on similar experiences.
- Examples:
  - Speech Recognition: adapt to different accents.
  - Sentiment Analysis: ratings available for TVs, laptops and CD players, but, how to rate general electronics?

Problem Formulation

- Given distributions and corresponding hypothesis:



$$\forall i, \underbrace{\mathcal{L}(D_i, h_i, f) \leq \epsilon}.$$

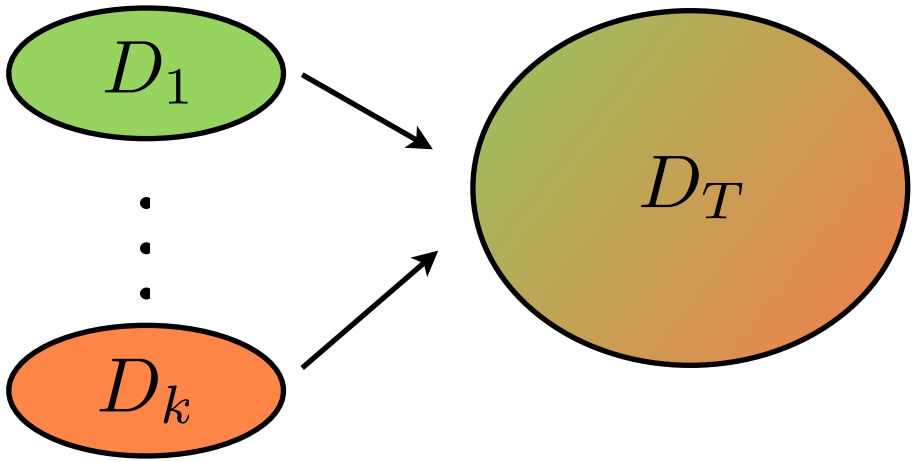
each hypothesis performs well in its domain.

Notation: $\mathcal{L}(D_i, h_i, f) = \mathop{\mathrm{E}}_{x \sim D_i}[L(h_i(x), f(x))]$.

Loss *L* assumed non-negative, bounded, convex and continuous.

Problem Formulation

- The unknown target distribution is a mixture of input distributions.



$$D_T(x) = \sum_{i=1}^{k} \lambda_i D_i(x)$$

- Task: choose a hypothesis mixture that performs well in target distribution.

$$h_z(x) = \sum_{i=1}^{k} z_i h_i(x) \qquad h_z(x) = \sum_{i=1}^{k} \frac{z_i D_i(x)}{\sum_{j=1}^{k} z_j D_j(x)} h_i(x)$$

convex combination rule distribution weighted combination

Main Results

- Although convex combination seems natural, we show that it can perform very poorly.
- Distribution weighted combination seems to be the "correct" combining rule.
- There exists a single "robust" distribution weighted hypothesis, that does well for any target mixture.

$$\forall f, \exists z, \forall \lambda, \ \mathcal{L}(D_\lambda, h_z, f) \leq \epsilon.$$

Known Target Distribution

- For some distributions, any convex combination performs poorly.

distribution weights

	D_T	D_0	D_1
a	0.5	1	0
b	0.5	0	1

hypothesis output

	f	h_0	h_1
a	1	1	0
b	0	1	0

- Base hypotheses have no error within domain.
- Any convex combination has error of 1/2.

Known Target Distribution

- If distribution is known, distribution weighted rule will always do well. Choose: $z = \lambda$.

$$h_\lambda(x) = \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{\sum_{j=1}^{k} \lambda_j D_j(x)} h_i(x).$$

- Proof:

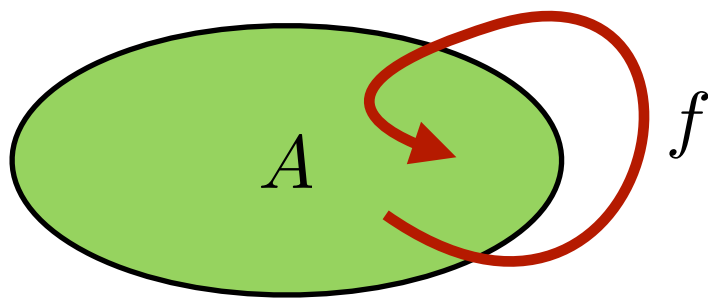
$$\begin{aligned}
\mathcal{L}(D_T, h_\lambda, f) &= \sum_{x \in X} L(h_\lambda(x), f(x)) D_T(x) \\
&\leq \sum_{x \in X} \sum_{i=1}^{k} \frac{\lambda_i D_i(x)}{D_T(x)} L(h_i(x), f(x)) D_T(x) \\
&= \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_i(x), f(x)) \leq \sum_{i=1}^{k} \lambda_i \epsilon = \epsilon.
\end{aligned}$$

Unknown Target Mixture

- Zero-sum game:
  - NATURE: select a target distribution D_i .
  - LEARNER: select a z , i.e. a distribution weighted hypothesis h_z.
  - Payoff: $\mathcal{L}(D_i, h_z, f)$.
  - Already shown: game value is at most ϵ .
- Minimax theorem (modulo discretization of z): there exists a mixture $\sum_j \alpha_j h_{z_j}$ of combination weighted hypothesis that does well for any distribution mixture.

Balancing Losses

- Brouwer's Fixed Point theorem: for any compact, convex, non-empty set A and any continuous function $f\colon A \to A$, there exists x such that: $f(x) = x$.



Notation:
$$\mathcal{L}_i^z := \mathcal{L}(D_i, h_z, f).$$

- Define mapping ϕ by: $[\phi(z)]_i = \dfrac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z}$.
- By fixed point theorem (modulo continuity):

$$\exists z\colon \forall i, z_i = \frac{z_i \mathcal{L}_i^z}{\sum_j z_j \mathcal{L}_j^z} \implies \forall i, \mathcal{L}_i^z = \sum_j z_j \mathcal{L}_j^z =: \gamma.$$

Bounding Loss

- For fixed point z,

$$\mathcal{L}(D_z, h_z, f) = \sum_{x \in X} L(h_z(x), f(x)) \Big(\sum_{i=1}^{k} z_i D_i(x) \Big)$$
$$= \sum_{i=1}^{k} z_i \sum_{x \in X} D_i(x) L(h_z(x), f(x))$$
$$= \sum_{i=1}^{k} z_i \mathcal{L}_i^z = \sum_{i=1}^{k} z_i \gamma = \gamma.$$

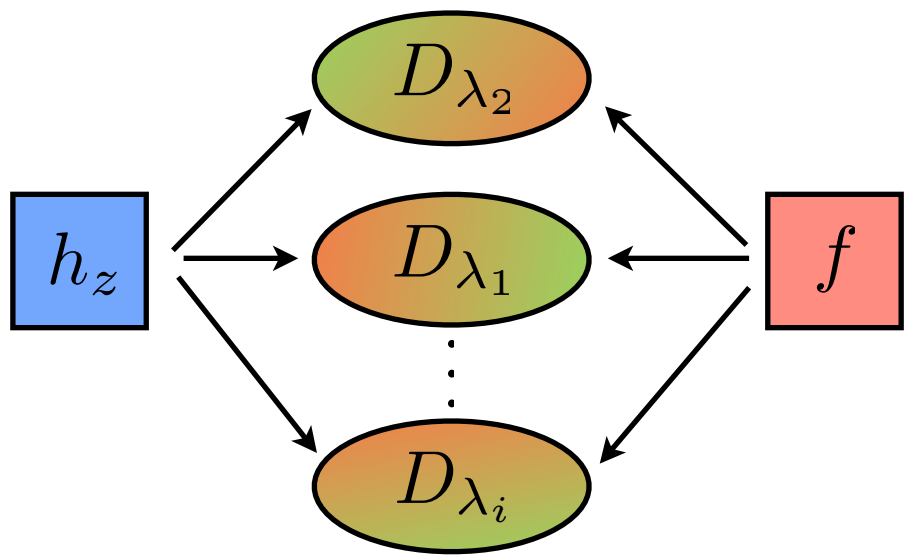
- Also, by convexity,

$$\gamma = \mathcal{L}(D_z, h_z, f) \leq \sum_{x \in X} \sum_{i=1}^{k} \frac{z_i D_i(x)}{D_z(x)} L(h_i(x), f(x)) D_z(x) = \sum_{i=1}^{k} z_i \mathcal{L}(D_i, h_i, f) \leq \epsilon.$$

Bounding Loss

- Thus, $\gamma \leq \epsilon$ and for any mixture λ,

$$\mathcal{L}(D_\lambda, h_z, f) = \sum_{i=1}^{k} \lambda_i \mathcal{L}(D_i, h_z, f) \leq \sum_{i=1}^{k} \lambda_i \gamma = \gamma \leq \epsilon.$$



Details

- To deal with non-continuity refine hypotheses:

$$h_z^\eta(x) = \sum_{i=1}^{k} \frac{z_i D_i(x) + \eta/k}{\sum_{j=1}^{k} z_j D_j(x) + \eta} h_i(x).$$

- Theorem: for any target function f and any $\delta > 0$,

$$\exists \eta > 0, z: \ \ \forall \lambda, \mathcal{L}(D_\lambda, h_z^\eta, f) \leq \epsilon + \delta.$$

- If loss obeys triangle inequality:

$$\forall \delta > 0, \exists z, \eta > 0, \forall \lambda, f \in \mathcal{F}, \ \ \mathcal{L}(D_\lambda, h_z^\eta, f) \leq 3\epsilon + \delta.$$

holds for all admissible target functions.

A Simple Algorithm

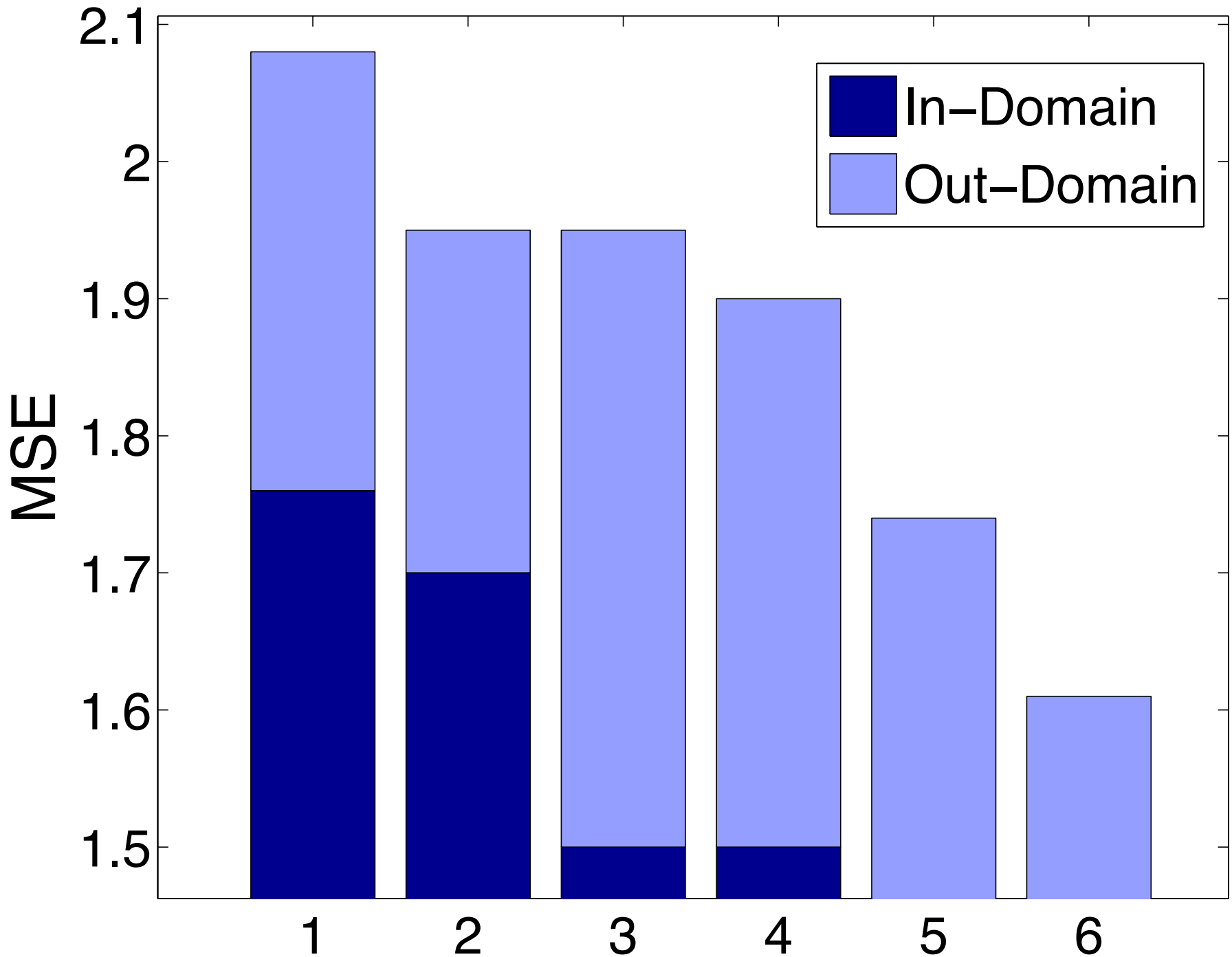
- A simple constructive algorithm, choose z with uniform weights:

$$\begin{aligned}
\mathcal{L}(D_\lambda, h_u, f) &= \sum_x D_\lambda(x) L\left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x)\right) \\
&= \sum_x \left(\sum_{m=1}^k \lambda_m D_m(x)\right) L\left(\sum_{i=1}^k \frac{D_i(x)}{\sum_{j=1}^k D_j(x)} h_i(x), f(x)\right) \\
&\leq \sum_x \underbrace{\frac{\sum_{m=1}^k \lambda_m D_m(x)}{\sum_{j=1}^k D_j(x)}}_{\leq 1} \sum_{i=1}^k D_i(x) L\left(h_i(x), f(x)\right) \\
&\leq \sum_{i=1}^k \sum_x D_i(x) L\left(h_i(x), f(x)\right) = \sum_{i=1}^k \mathcal{L}(D_i, h_i, f) = \sum_{i=1}^k \epsilon_i \boxed{\leq k\epsilon.}
\end{aligned}$$

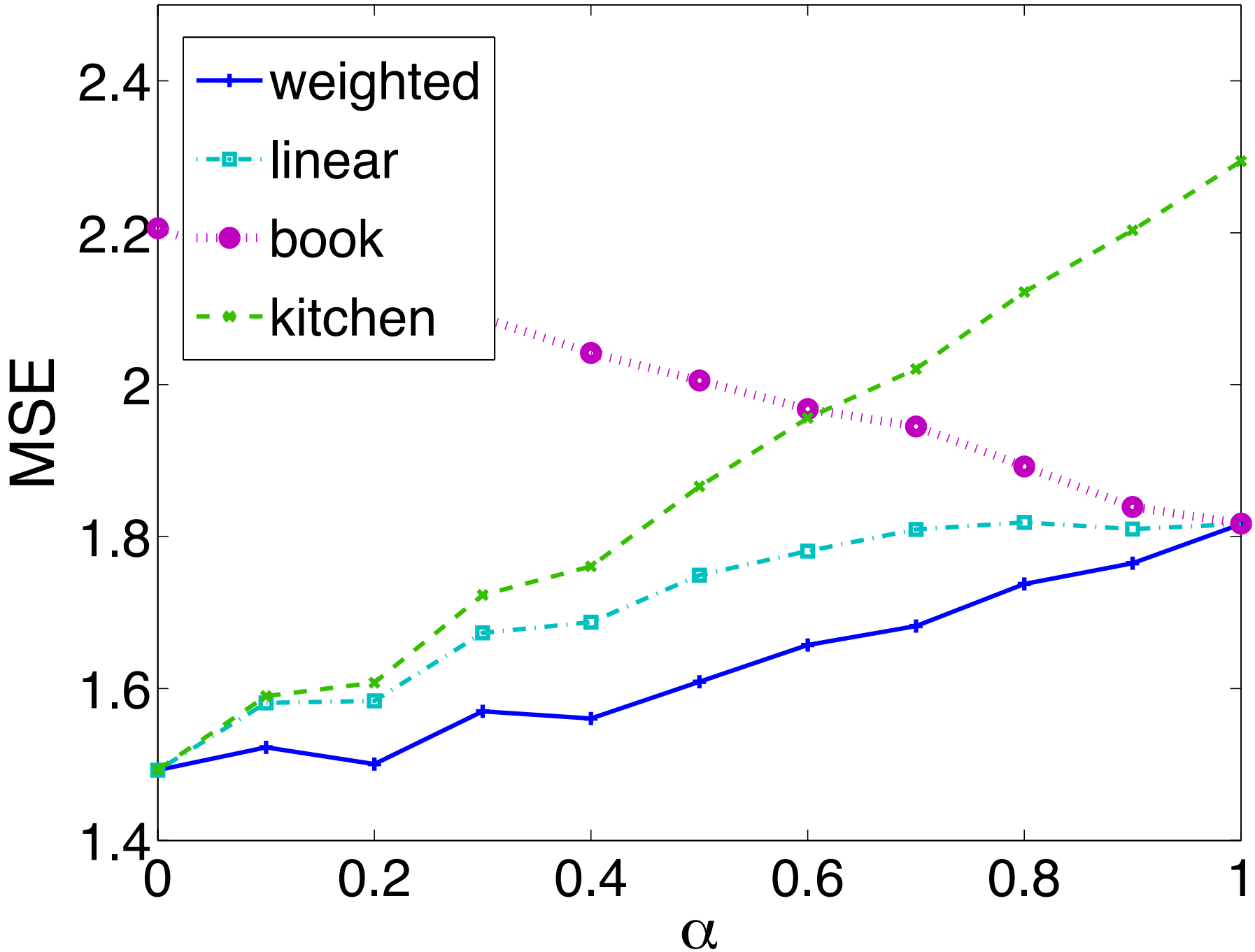
Preliminary Empirical Results

- Sentiment Analysis - given a product review (text string), predict a rating (between 1.0 and 5.0).
- 4 Domains: Books, DVDs, Electronics and Kitchen Appliances.
- Base hypotheses are trained within each domain (Support Vector Regression).
- We are not given the distributions. We model each distribution using a bag of words model.
- We then test the distribution combination rule on known target mixture domains.

Uniform Mixture Over 4 Domains



Mixture = α book + (1 – α) kitchen



Conclusion

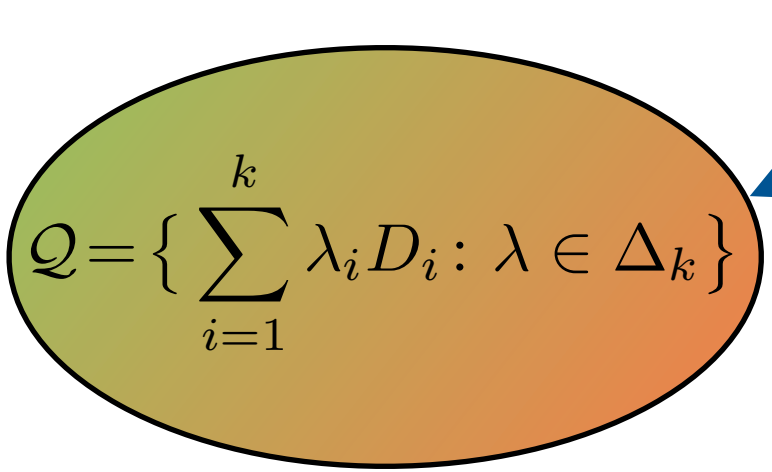
- Formulation of the multiple source adaptation problem.
- Theoretical analysis for mixture distributions.
- Efficient algorithm for finding distribution weighted combination hypothesis?
- Beyond mixture distributions?

Extensions - Arbitrary Target Distrib.

(Mansour, MM, and Rostami, 2010)

- Theorem: for any $\delta > 0$,

$$\exists \eta, z : \forall P, \mathcal{L}(P, h_z^\eta, f) \leq \left[d_\alpha(P \| \mathcal{Q})(\epsilon + \delta) \right]^{\frac{\alpha-1}{\alpha}}.$$



measured in terms of Rényi divergence,

$$d_\alpha(P, Q) = \left[\sum_x \frac{P^\alpha(x)}{Q^{\alpha-1}(x)} \right]^{\frac{1}{\alpha-1}}.$$

Other Extensions

- Approximate distributions (estimated):
  - similar results shown depending on divergence between true and estimated distributions.
- Different source target functions f_i :
  - similar results when target functions close to f on target distribution.