Towards theoretical understanding of Domain Adaptation Learning

Shai Ben-David

University of Waterloo

ECML/PKDD 2009

LNIID **Workshop**

Success and limitations of Machine Learning Theory

- Some remarkable successes, like *Boosting* and *Support Vector Machines* are founded on theoretical insights.
- However, our mainstream theory is based on simplifying assumptions that many potential applications fail to meet.
- Major challenge –

  Extend ML theory to cover a wider variety of realistic settings.

A major common assumption-
The data-generating distribution is stationary

Most of the statistical learning guarantees are based on assuming that

The learning environment is unchanged throughout the learning process.

Formally, it is common to assume that

both the training and the test examples are generated i.i.d. by the **same fixed** *probability distribution.*

This is unrealistic for many ML applications
(and, therefore I think its great to have this workshop).

Outline of this talk

- Some general issues and principles.
- A example "success story" (or two).
- Some inherent limitations.
- Major open questions.

Learning when Training and Test distributions differ

- *Driving lessons* – train on one car, test on another.
- *Spam filters* – train on email arriving at one address, test on a different mailbox.
- *Natural Language Processing* tasks- train on some content domains, test on others.

Our Goal:
Figure out when and how will DA learning succeed

- While domain adaptation is being used in practical ML tasks, we do not have satisfactory theoretical justification for that.

- Why should one desire such theoretical backbone?

  - To provide success guarantees to common heuristics.
  - To help understand under what circumstances common DA techniques may fail.
  - To help choose the right leaning paradigm for a given task.
  - To guide the development of novel DA algorithms.
  - .....

Main issue-
MODELING TASK RELATEDNES

Preliminary convention – *a learning task is a joint distribution over points and labels,* P *over* $X \times \{0,1\}$.

- Most obvious desired relatedness – labels are the same.

  That is the conditionals $P(l|x)$ are the same for both the training and target tasks – this is the "*covariate shift*" assumption.

- ***Note*** – this may be meaningless unless we assume some further assumptions about unlabeled dist.
- Resulting needed assumption – relatedness of unlabeled dist.'s (How should we measure that?)
- Relaxing "covariate shift" – allow some label differences – how should we measure label similarity then?

Success example:
The POS tagging Inductive Transfer Problem

POS tagging is a common preprocessing step in NLP systems.

- *Can an automatic POS tagger be trained on one domain (say,* ***legal documents****) and then be used on a different domain (say,* ***biomedical abstracts****).*

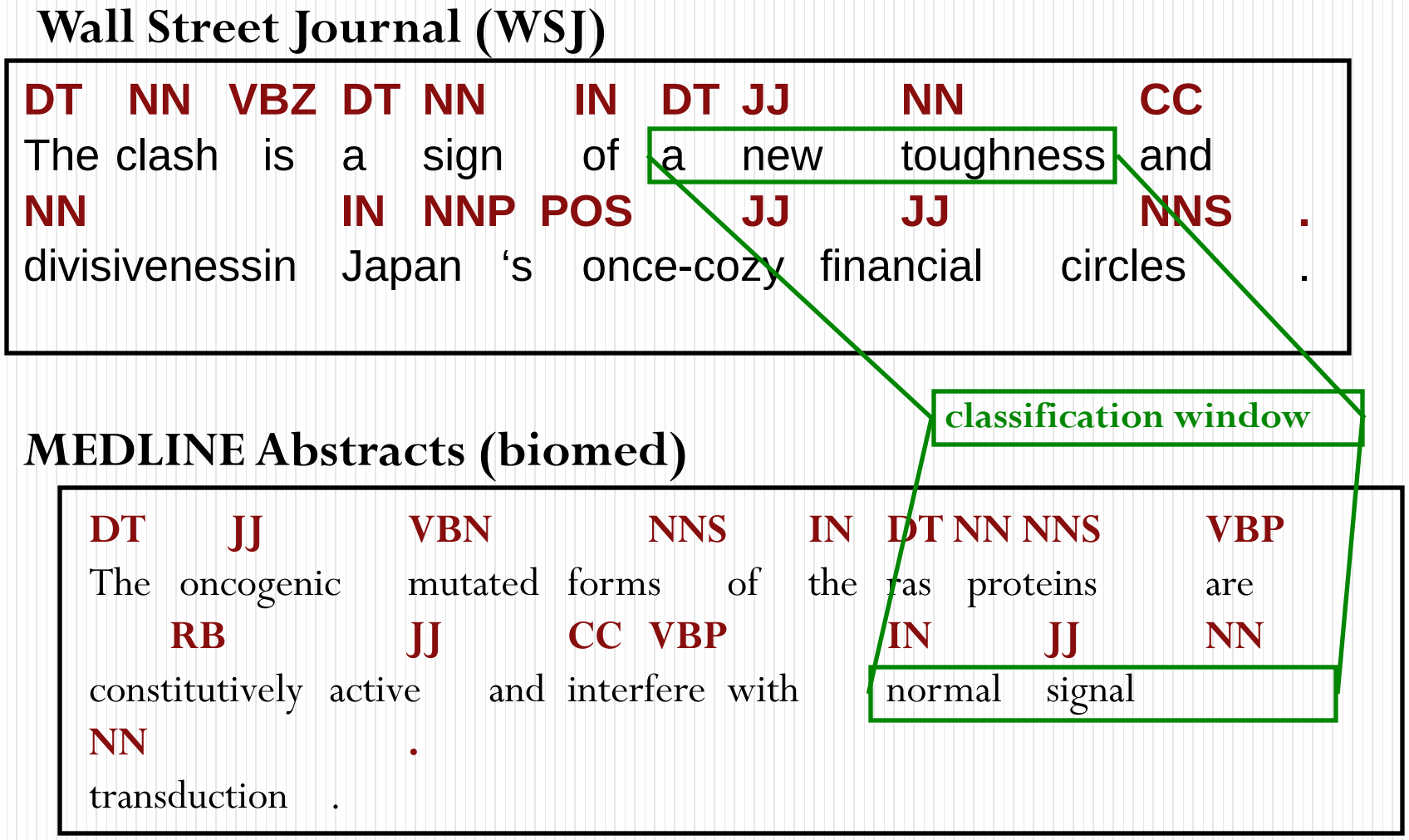
- The issue here is the discrepancy between the classifier's training and test data distributions.

- For this problem, *unlabeled* target examples are readily available.

Structural Correspondence Learning (Blitzer, McDonald, Pereira)

- Choose a set of "pivot words" (determiners, prepositions, connectors and frequently occurring verbs).
- Represent every word in a text as a vector of its correlations, within a small 'window", with each of the pivot words.
- Train a linear separator on the (images of) the training data coming from one domain and use it for tagging on the other.

Structural Correspondence Learning (Blitzer, McDonald, Pereira)



A concrete setting for our problem

A learner has access to:

1. *Labeled* training data, S, randomly sampled by some source (training) distribution P_S

and

2. An *unlabeled* sample T sampled from a target ('test') distribution P_T

The learner's task is to predict labels of points generated (and labeled) according to P_T

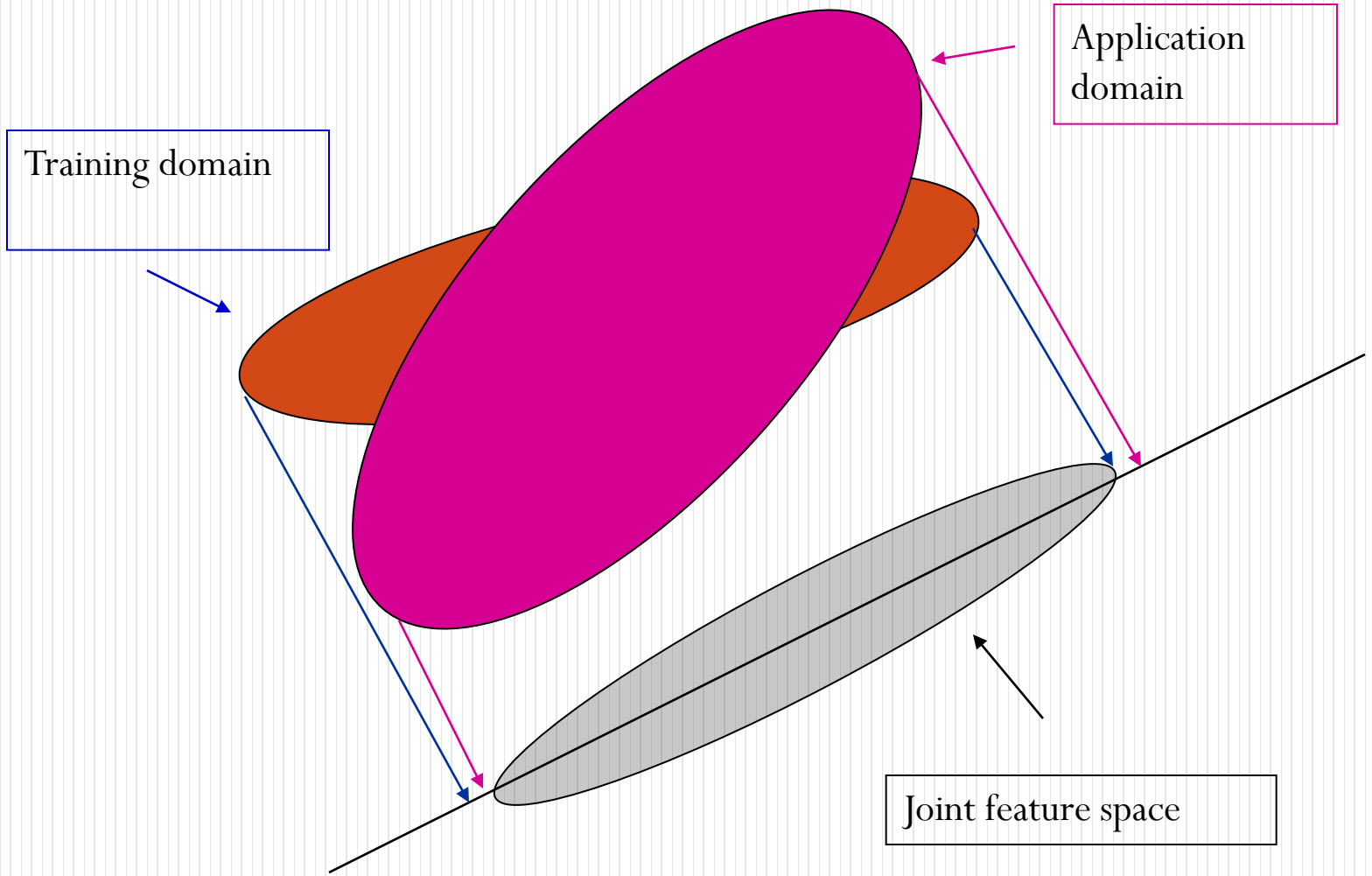
Our Inductive Transfer Paradigm

We propose to *embed* the original attribute space(s) into some *feature space* in which

1. *The two tasks look similar.*
2. *The source task can still be well classified.*

Then, treat the images of points from both distributions as if they are coming from a unique distribution.

The Common-Feature-Space Idea



How should one measure similarity of distributions?

Common measures of distributions similarity :

- **Total Variance**

$$TV(D, D') = Sup_{E\ measurable}(|D(E)-D'(E)|)$$

- **KL divergence**

These measures are too sensitive for our needs and cannot be reliably estimated from finite samples.

[Batu et al 2000]

A new distribution-similarity measure (Kifer, Ben-David and Gehrke '04)

For a class H of predictors, define:

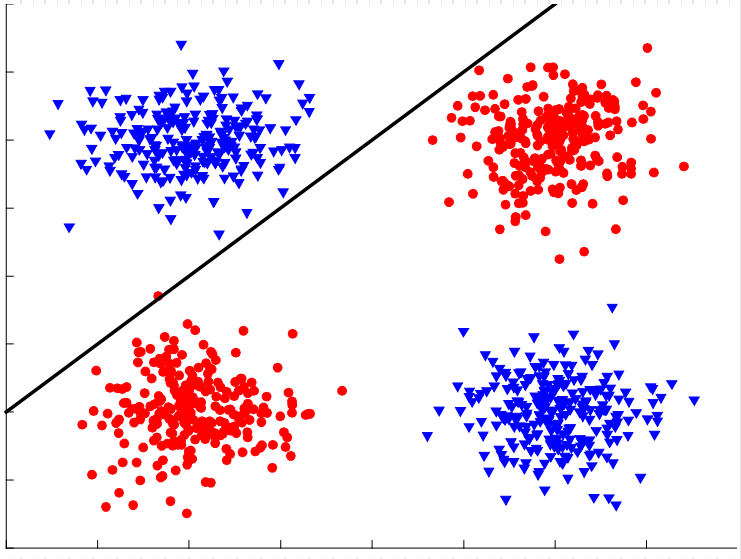
$$d_{\mathcal{H}}(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T) = 1 - 2 \min_{h' \in \mathcal{H}} \mathrm{err}(h')$$

Discriminate between S and T

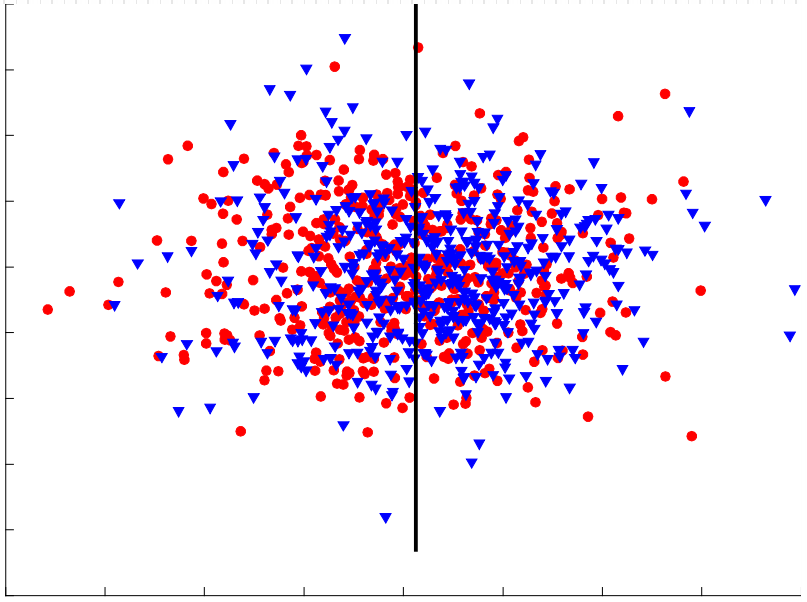
Namely, use classifiers in H to try to separate the two distributions.

Failure to separate means that the distributions are close.

Demonstration of the new distance



H-distance: 0.5



H-distance: 0.134

Estimating d_A from samples (Ben-David, Kifer, Gehrke)

Theorem 3.1. Let P_1, P_2 be any probability distributions over some domain X and let $\mathcal{A}$ be a family of subsets of X and $\epsilon \in (0,1)$. If S_1, S_2 are i.i.d m samples drawn by P_1, P_2 respectively, then,

$$P\left[\exists A \in \mathcal{A} \left| |P_1(A) - P_2(A)| - |S_1(A) - S_2(A)| \right| \geq \epsilon\right]$$

$$< \Pi_{\mathcal{A}}(2m) 4e^{-m\epsilon^2/4}$$

It follows that

$$P\left[|d_{\mathcal{A}}(P_1, P_2) - d_{\mathcal{A}}(S_1, S_2)| \geq \epsilon\right] < \Pi_{\mathcal{A}}(2m) 4e^{-m\epsilon^2/4}$$

A Bound on Target Domain Error

Theorem Assume we have m be a random labeled sample of the source domain, S, and let $\tilde{\mathcal{U}}_S$ and $\tilde{\mathcal{U}}_T$ be random unlabeled samples of size m' from $\tilde{\mathcal{D}}_S$ and $\tilde{\mathcal{D}}_T$ respectively. Then with probability $1-\delta$, for every $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \hat{\epsilon}_S(h) + \sqrt{\frac{4}{m}\left(d \log \frac{2em}{d} + \log \frac{4}{\delta}\right)}$$

$$+\lambda + d_{\Delta\mathcal{H}}(\tilde{\mathcal{U}}_S, \tilde{\mathcal{U}}_T) + \sqrt{\frac{16}{m'}\left(d \log(2m') + \log \frac{4}{\delta}\right)}$$

The notation above

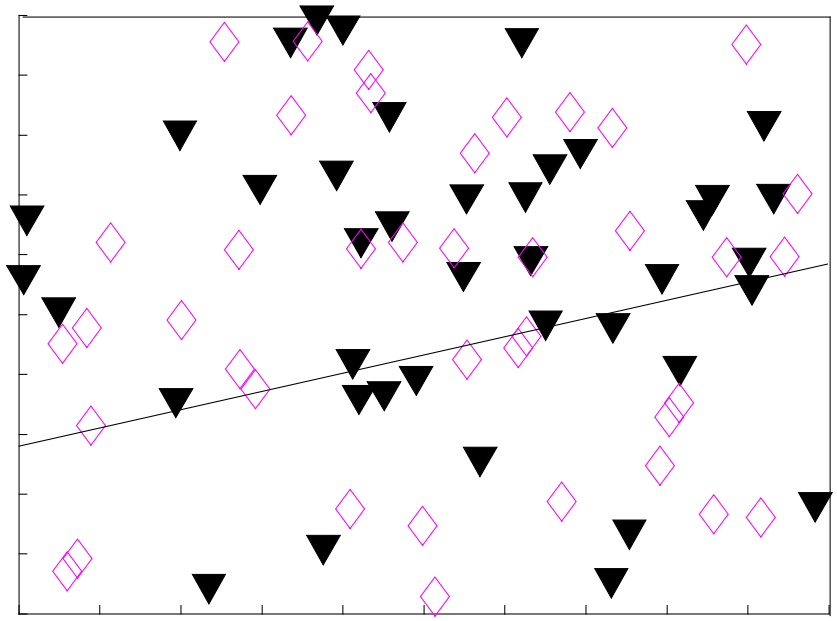
Where d is the VC-dim of H, and

$$\lambda = \inf_{h \in H} (\mathrm{Er}_T(h) + \mathrm{Er}_S(h))$$

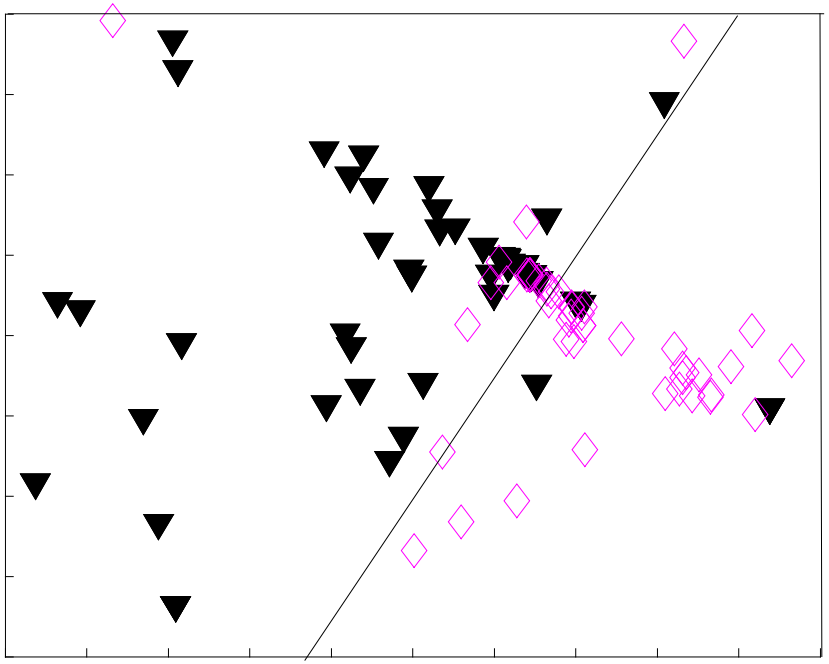
Note that this is a measure of the relatedness of labelling functions of the two tasks.

Visualizing the Classifiers

Scatter plots of nouns **(diamonds)** and verbs **(triangles)**



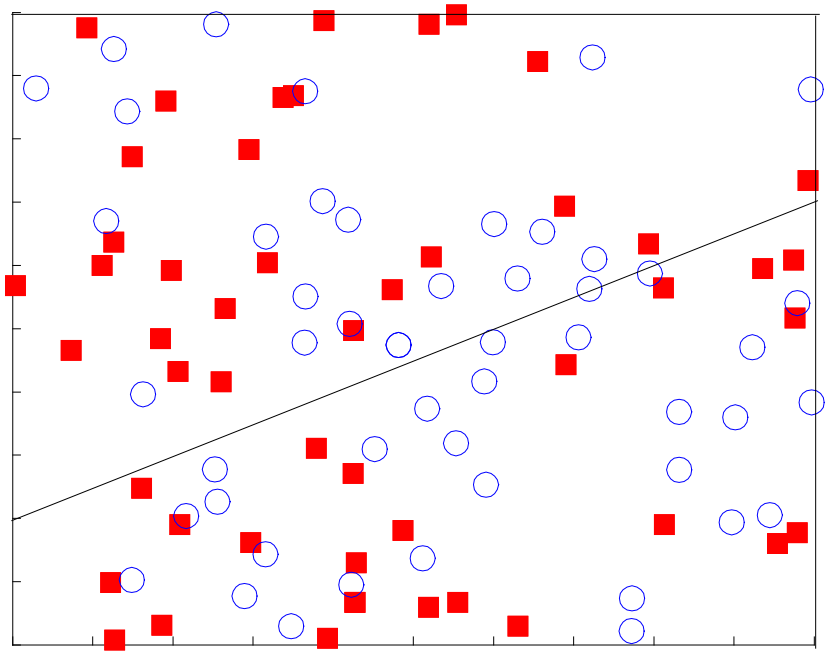
Random Projections



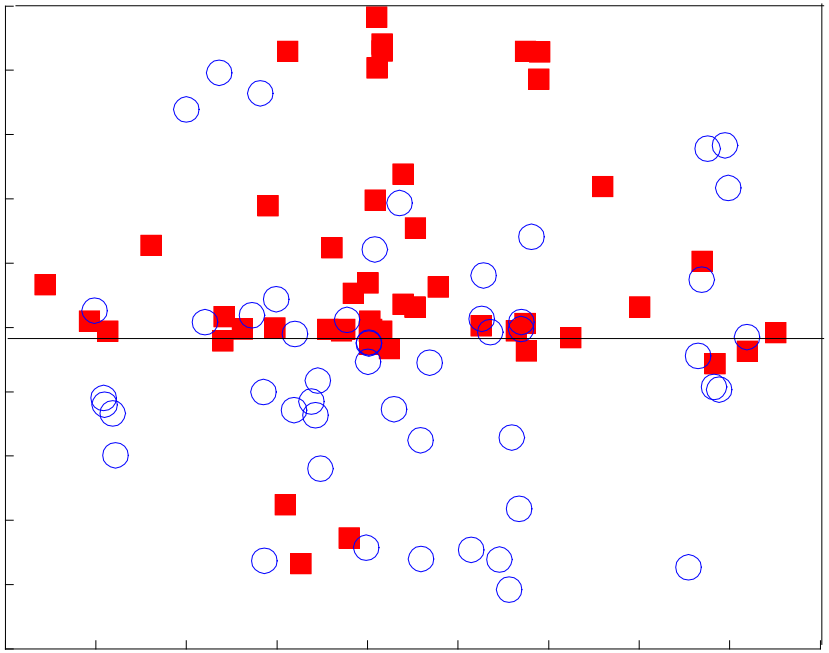
Structural Correspondence Learning

Visualizing the d_H -distance

Scatter plots of financial (squares) and biomedical (circles) words



Random Projections



Structural Correspondence Learning

The algorithmic conclusion

Find a feature space representation, R, such that:

1. The (unlabeled) distributions induced by the *Source* and *Target* distributions under a representation R are similar.
2. There *exist* a predictor in H that works reasonably well for the training data (in the feature space).

To predict:

Represent *your test point in the feature space, and use a good training classifier to predict its label.*

Evaluating the generalization bound

Recall the generalization bound we have:

$$\varepsilon_T(\mathrm{h}) \le \varepsilon_S(\mathrm{h}) + \mathrm{d}_{\Delta \mathrm{H}}(U_S, U_T) + \inf_{h \in \mathrm{H}} (\mathrm{Er}_T(h) + \mathrm{Er}_S(h))$$

Can this bound be improved?

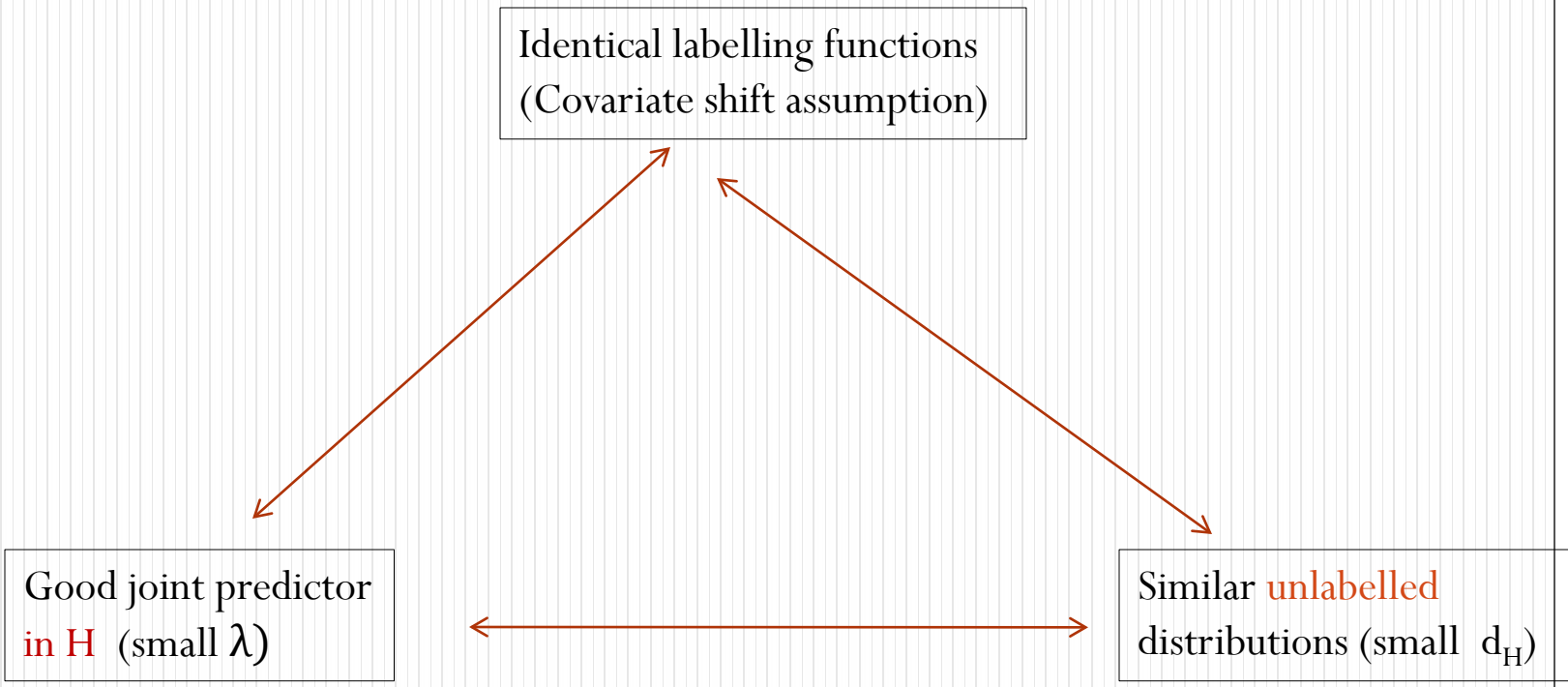
Other candidate relatedness measures: The covariate shift assumption

A common assumption in domain adaptation literature is
The Covariate Shift assumption –

Both the source and the target tasks are labelled by the same function (assigning labels to domain point deterministically)

Does it help to improve the domain adaptation performance?

Sufficient conditions for DA learning

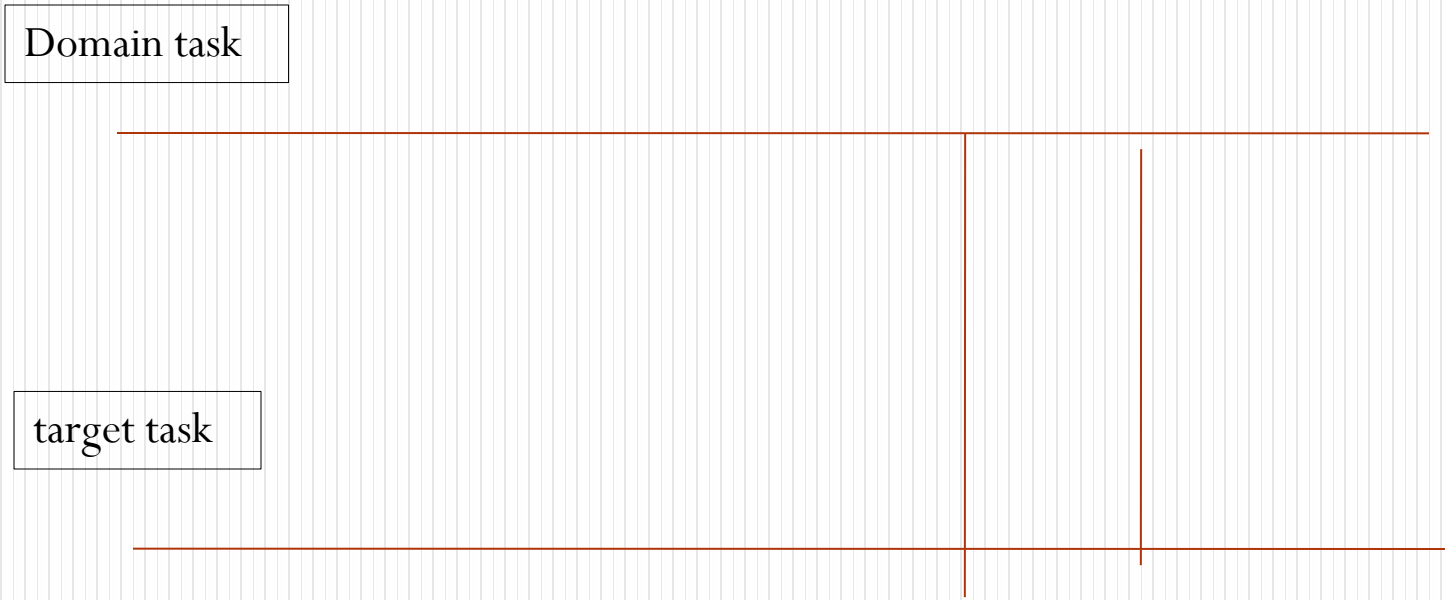


Demonstrating uselessness of the Covariate Shift assumption

Note that here we have the covariate shift assumption + small d_H

Yet, the ***All-1*** h has small error on the domain task and large error on the target

Demonstrating uselessness of the Covariate Shift assumption



Note that here we have the covariate shift assumption + small λ

Yet, the al-1 h has small error on the domain task and large error on the target

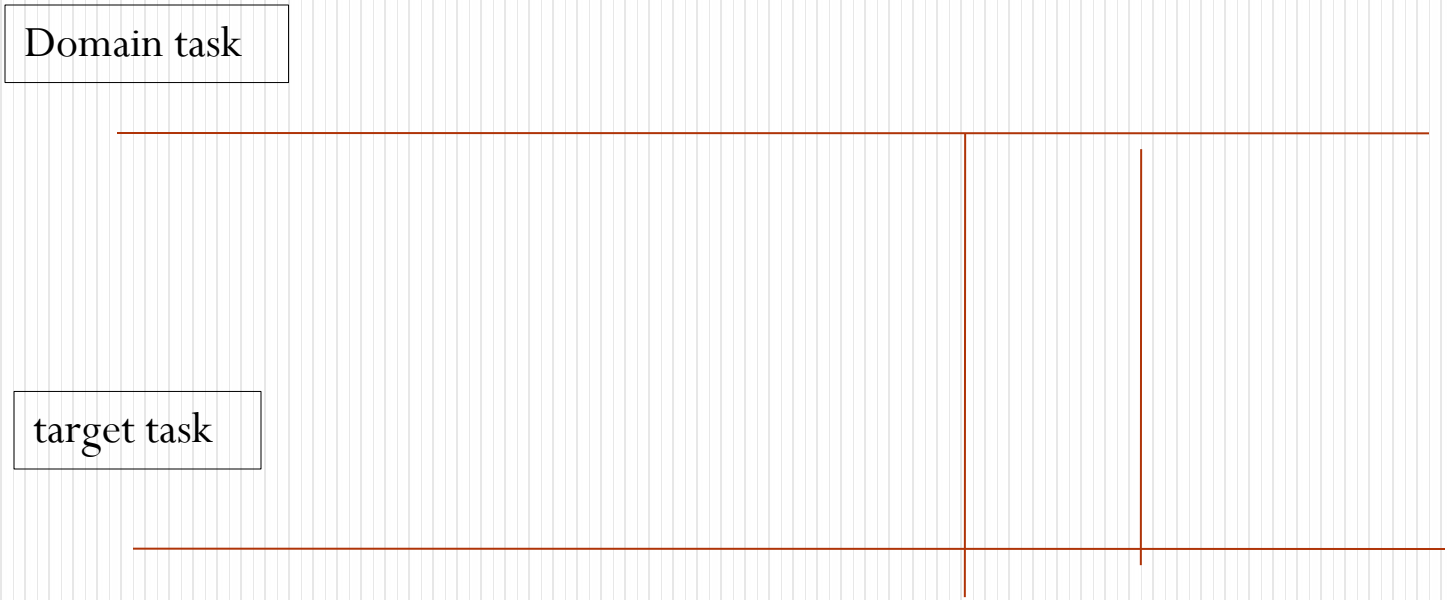
Conservative vs. Adaptive algorithms

- Can we do better than learning on the source domain and applying the SAME hypothesis to the target domain?
- We should! But how?
- The Afshani-Mohari-Mansour idea:

  Use the target unlabeled sample to reweigh the training sample. Choose h that minimizes the training error w.r.t. This reweighted training sample.

- It may fail badly!
- Q: Under what assumptions will it work?
- Q: What other types of adaptive DA may work?

Demonstrating the failure of the AMM reweighting paradigm



Note that here , although we have small λ,
and without reweighting the learner will have zero target error,
after reweighting the learner will choose the ***All-0*** hypothesis
and fail badly

Major remaining open questions

- Improve our basic generalization-error bound.
- Find relatedness parameters under which different paradigms work (e.g., ERM with respect to task-reweighted training sample).
- Come up with different adaptive (rather than conservative) learning algorithms (or even just conditions under which the reweighting trick is guaranteed to succeed)
- Come up with more user-friendly useful relatedness notions.