Formal Framework for Multi-Task Learning with Provable Generalization Bounds

Shai Ben-David

University of Waterloo, Canada

NIPS Workshop
December 2005

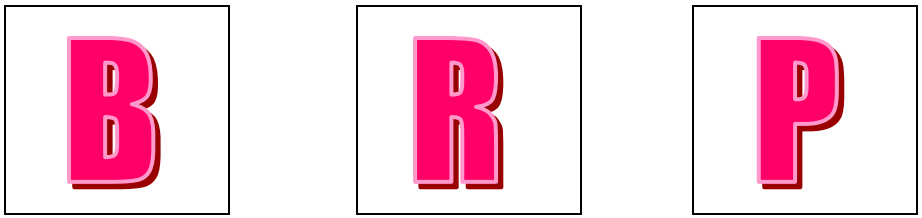
Multitask Learning

Intuition:

- Greater sample sizes allow for better classification predictions.
- We can compensate for small sample size by using additional samples from “related” learning problems.

Example:

In learning character recognition, some characters are somehow related.



Verifying This Intuition

Empirically

Empirical work has shown that data from extra "related" tasks does improve accuracy.

[Intrator and Adelman, '96], [Thrun, '96] [Caruana, '97], [Heskes, '00] [Bakker and Heskes '03]

Verifying This Intuition

Why theoretical analysis?

Some potential benefits of theoretical analysis for MTL

- Providing performance guarantees
- Suggesting new areas of applications
- Inspiring the development of new algorithmic paradigms.

A key component –
Modeling Task Relatedness

There is a wide variety of approaches.

They reflect

- Different potential target applications.
- Different frameworks for modeling user's prior knowledge.
- Personal and research community tastes …

Common Task-relatedness Approaches

- *Bayesian: Probability model for task generation* **[Baxter ’00], [Heskes ’00], [The, Seeger, Jordan ’05], [Zhang, Gharamani, Yang ’05]**
- *Between-task noise correlations* **[Greene ’02]**
- *Hidden common data structure*
  - *Implicit structure (common kernels)* **[Evgeniou, Micchelli, Pontil ’05]**
  - *Explicit structure (PCA)* **[Ando, Zhang ’04]**

A different approach- Transformation-Relatedness

While its difficult to pin down the *exact form* of the relationship between different learning tasks, it is often the case that we do have some *idea about the structure* of this relationship.

We would like to exploit this partial knowledge.

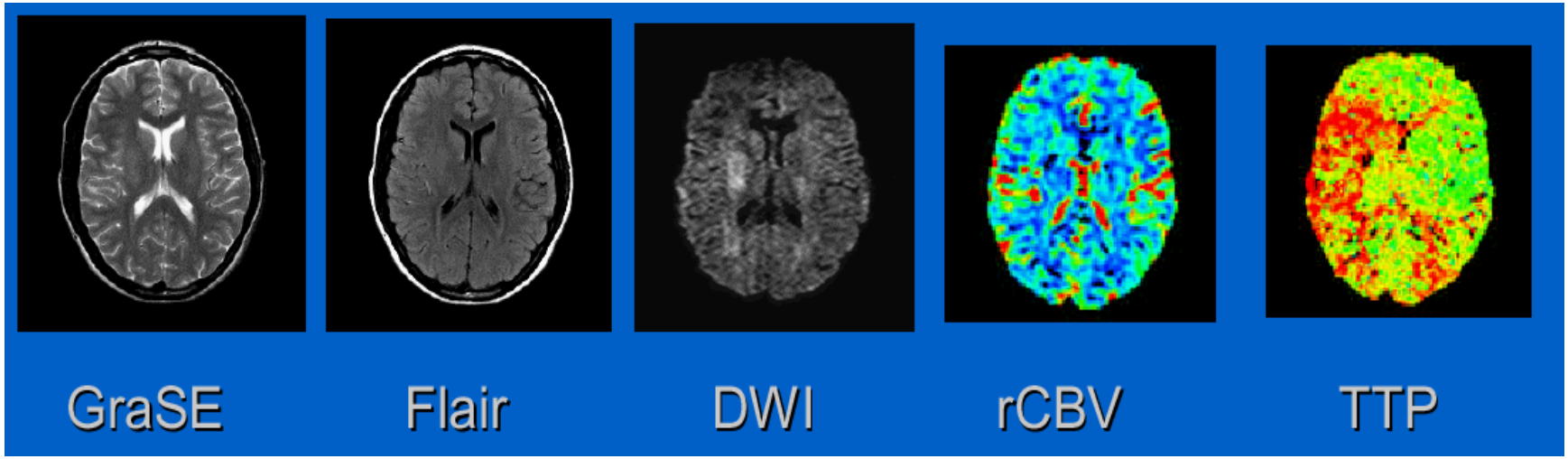
We consider situations in which the different tasks are all generated by the same data probability distribution.

However, each learning task is the result of applying a different transformation to the training sample

While the identity of the specific transformations in not known to the learner, we do assume that he knows a set of functions that contains all these transformations.

Transformation Relatedness Example

Consider the task of learning to detect images of tumors through different brain-imaging techniques.



Merits of the Transformation-Relatedness Approach

- No prior assumption on way any specific task is chosen ("*worst case scenario*").
- Allow utilizing advantages from a *small number of tasks*.
- Allows the derivation of *generalization error bounds* that are provably superior to the corresponding single task learning bounds.
- Applicable to *multi-task clustering*.
- Applies to a different family scenarios than previous approaches.

Transformation Relatedness formal definition

- Fix a domain set X and let F be a family of 1-1 mappings from X to itself.

- Let P_i be probability distributions over $X \times \{0,1\}$.

  (Note that such a P completely determines a learning task)

- P_1 is F-related to P_2 if, for some $f \in F$,

  for any measurable $T \subseteq X \times \{0,1\}$,

  $$P_1(T) = P_2(f[T])$$

  (where $f[T] = \{(f(x), b) : (x, b) \in T\}$)

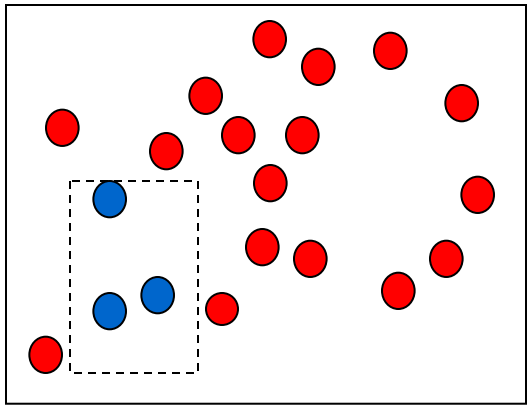
Example

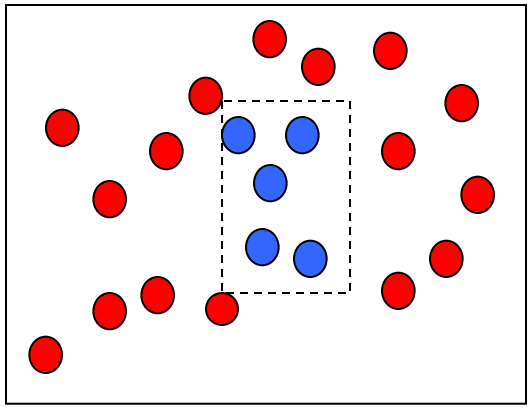
$X = \mathbf{R}^2$

F is the set of shifts:

Namely, $F=\{f_{a,b}(x, y)=(x+a, y+b): a,b \in \mathbf{R}\}$

Task 1





Task 2

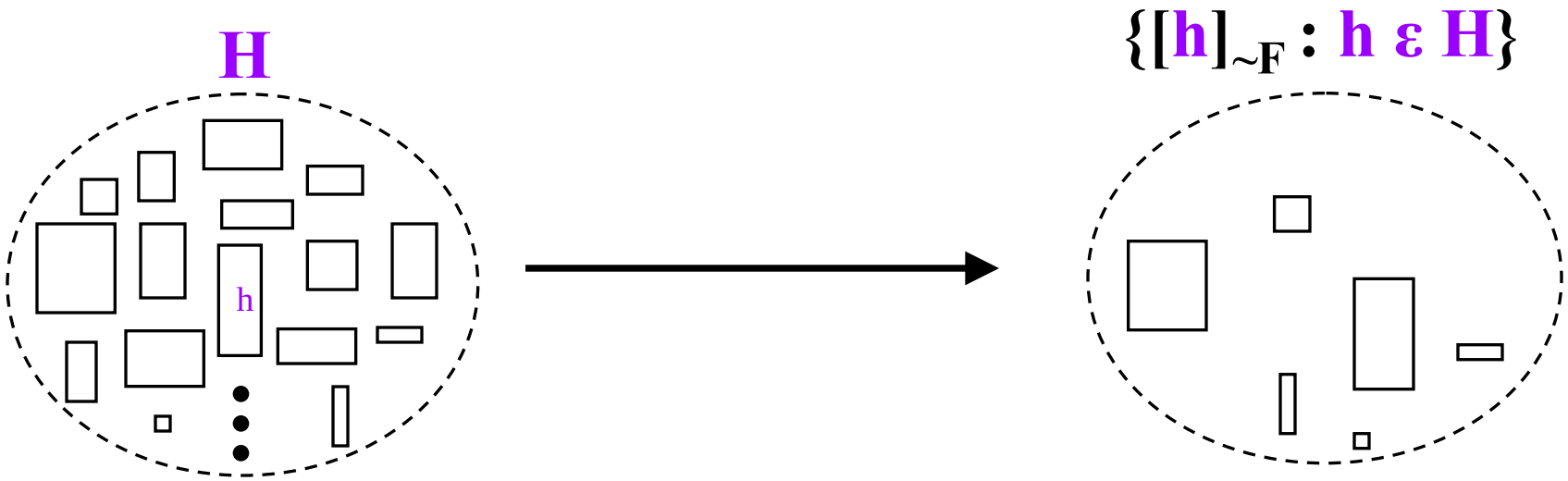
Applicability of Transformation Relatedness

F-related tasks arise in a variety of applications.

- **Learning from disparate databases**
  - ***F*** **accounts for the mappings between the different database schemas**
- **Speech recognition**
  - ***F*** **accounts for differences in tone, phrasing, accent, etc. between different speakers**
- **Image recognition**
  - ***F*** **accounts for differences in cameras, lighting conditions, etc.**

Exploiting Transformation Relatedness

- Use extra tasks to reduce the size of the hypothesis space, *H*.



- Then use standard learning methods on the primary task to select a classifier from the reduced hypothesis space.

Sample-Complexity Gains

- Our main result states that, given a set of *F*-related learning tasks, $P_1, \ldots P_n$, and samples of each, a learning algorithm can utilize the different samples to compute an equivalence class of a good hypothesis for P_1, based on knowing the transformation class *F* (but not the concrete transformation functions f_i).
- Roughly speaking, for the purpose of learning the *F*-invariants of a target hypothesis, samples labeled by different tasks are almost as useful as samples from any fixed specific task.

Formal Statements of Bounds some notation

- H is a class of predictors $h:X \rightarrow \{0,1\}$
- We assume that F acts as a group over H (e.g., if $h \in H$ and $f \in F$ then $h \circ f \in H$, and $f^{1} \in F$).
- $[h]_F$ denotes the equivalence class of h w.r.t F, namely, $[h]_F = \{h \circ f: f \in F\}$.
- $d_H(n)$ is a VC-type parameter measuring the complexity of determining the invariants of a predictor h modulo F- transformations on n tasks.

Generalization guarantee

- ***Theorem:***

  There is a transformation-relatedness version of Empirical Risk Minimization, ERM^{TR} that, for every H, F, F-related $P_1,\ldots P_n$, $\varepsilon>0$, and independent samples $S_1,..S_n$ of $P_1, \ldots P_n$ (respectively), finds $h \varepsilon H$ so that,

$$\mathrm{Er}^{P_1}(h) \le \inf\left\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{Er}^{S_i}(h_i) : h_1 \sim h_2 \sim \ldots . h_n\right\} + \mathrm{Max}\{\sqrt{d_H(n)}, \sqrt{VC\dim([h]_F)}\}/\sqrt{m}$$

  whenever, for all i $|S_i|>m$ (we ignore constants etc.)

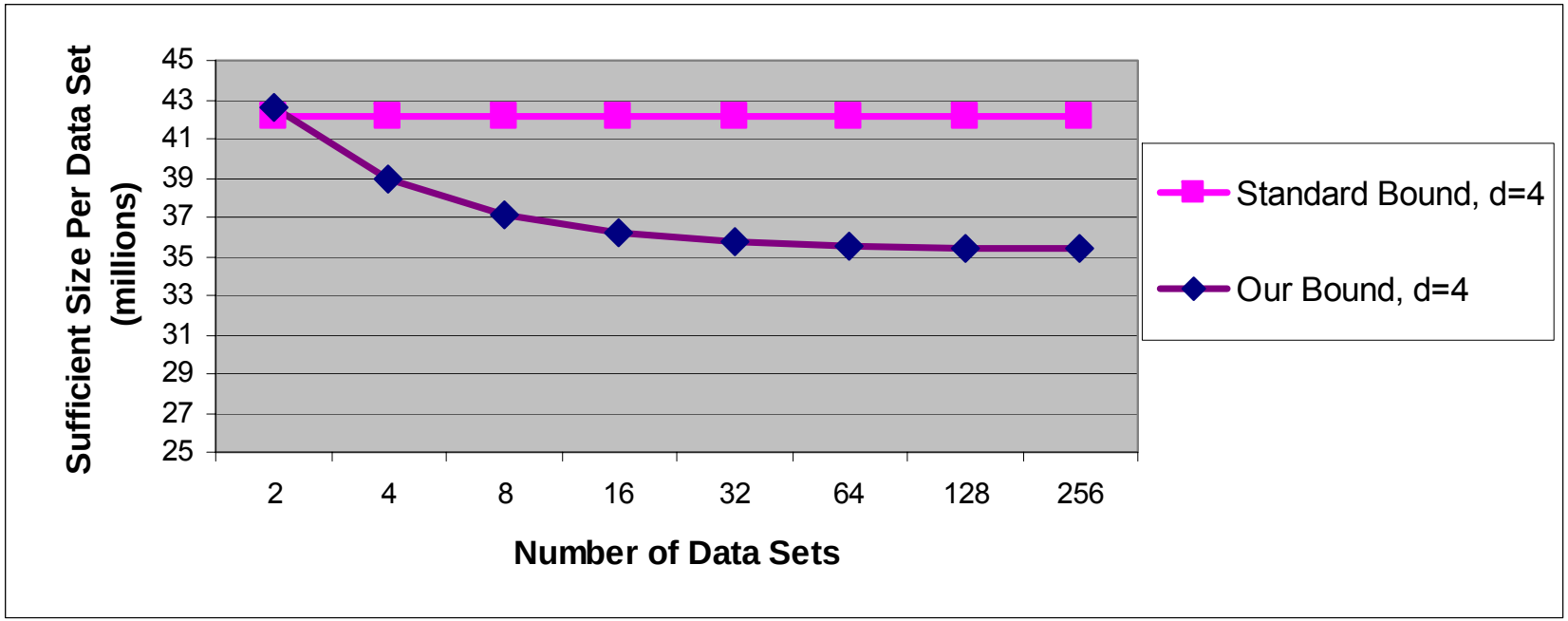
What have we gained?

- *Recall that, if one disregards the extra tasks P_i, then the usual VC bounds give a similar bound with $VCdim(H)$ replacing $d_H(n)$ and $VCdim([h])$.*

- *$d_H(n)$ is roughly the number of parameters that has to be set to determine any particular h from $[h]_F$*

- *In the example of hyper-rectangles in R^d transformed under (d-dimensional) shifts, the VCdim goes from $2d$ to d.*

How do the bounds compare?



For the case of *H* being the class of hyper-rectangles and *F* the class of Euclidean shifts, ε=0.001

Transformation-Relation learning algorithm

- We developed a version of *Decision Tree* learning for multi-task learning in the transformation-relation setting.

Experimental Results for MTL Decision Trees

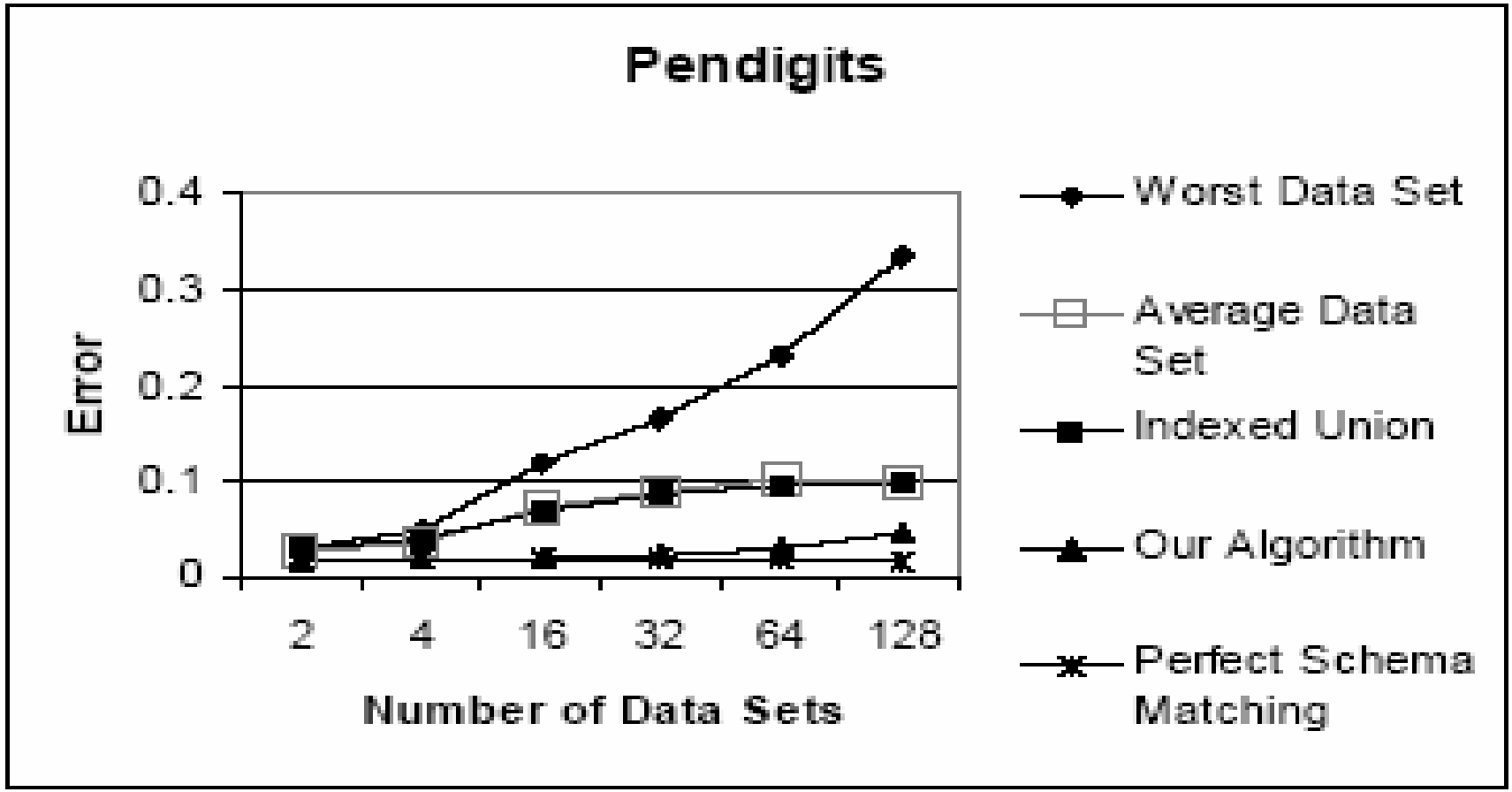


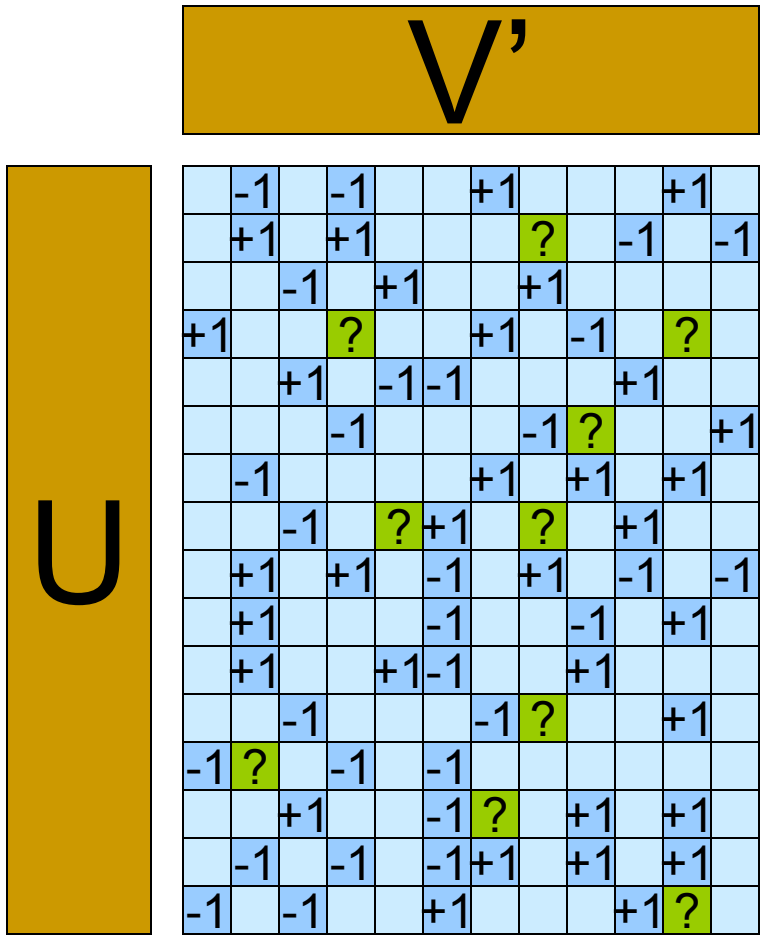
Fig. 7. Results for the Pendigits data set

Inductive Transfer and Collaborative Prediction

- *The task of Collaborative Prediction (building a movie recommendation system based on preferences of past customers on past movies), can be viewed as a particular case of MTL.*

- *By assuming the existence of hidden linear structure, of the labels matrix, Srebro Jaakkola and Alon have developed generalization bounds for that framework.*

Collaborative Prediction with Matrix Factorization



Fit factorizable (low-rank) matrix **X=UV'** to observed entries.

minimize Σloss(X_{ij};Y_{ij})

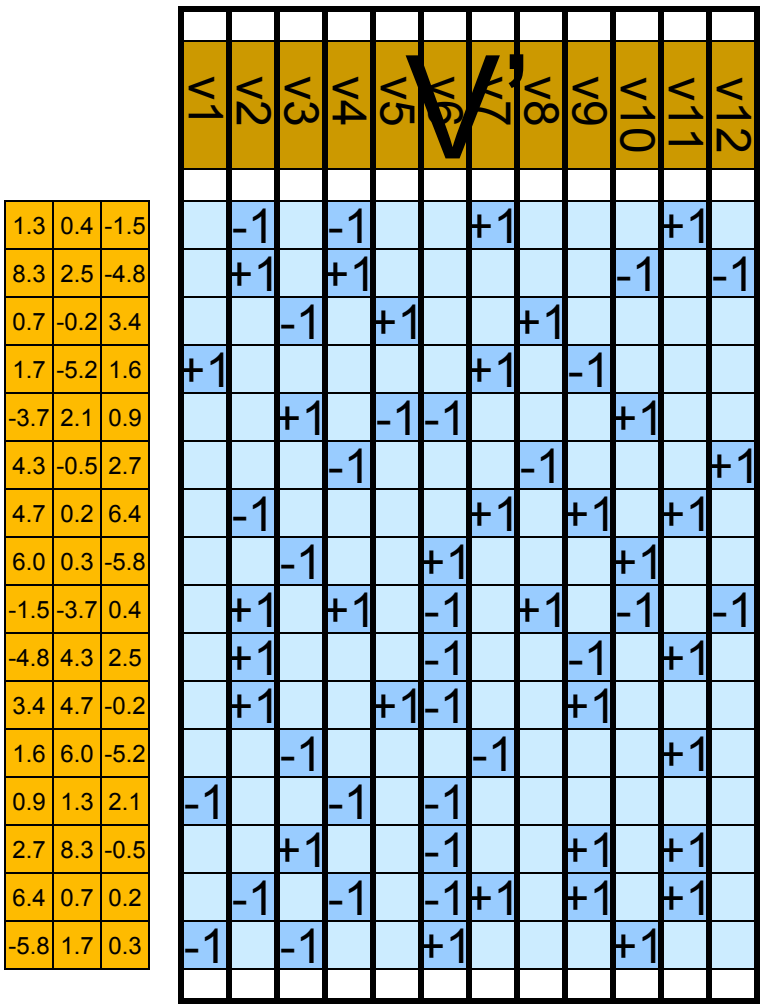
prediction

observation

Use matrix **X** to predict unobserved entries.

[Sarwar+00] [Azar+01] [Hoffman04] [Marlin+04]

Collaborative Prediction with Matrix Factorization



When U is fixed,

each row is a linear classification problem:

- rows of U are feature vectors
- columns of V are linear classifiers

Fitting U **and** V:

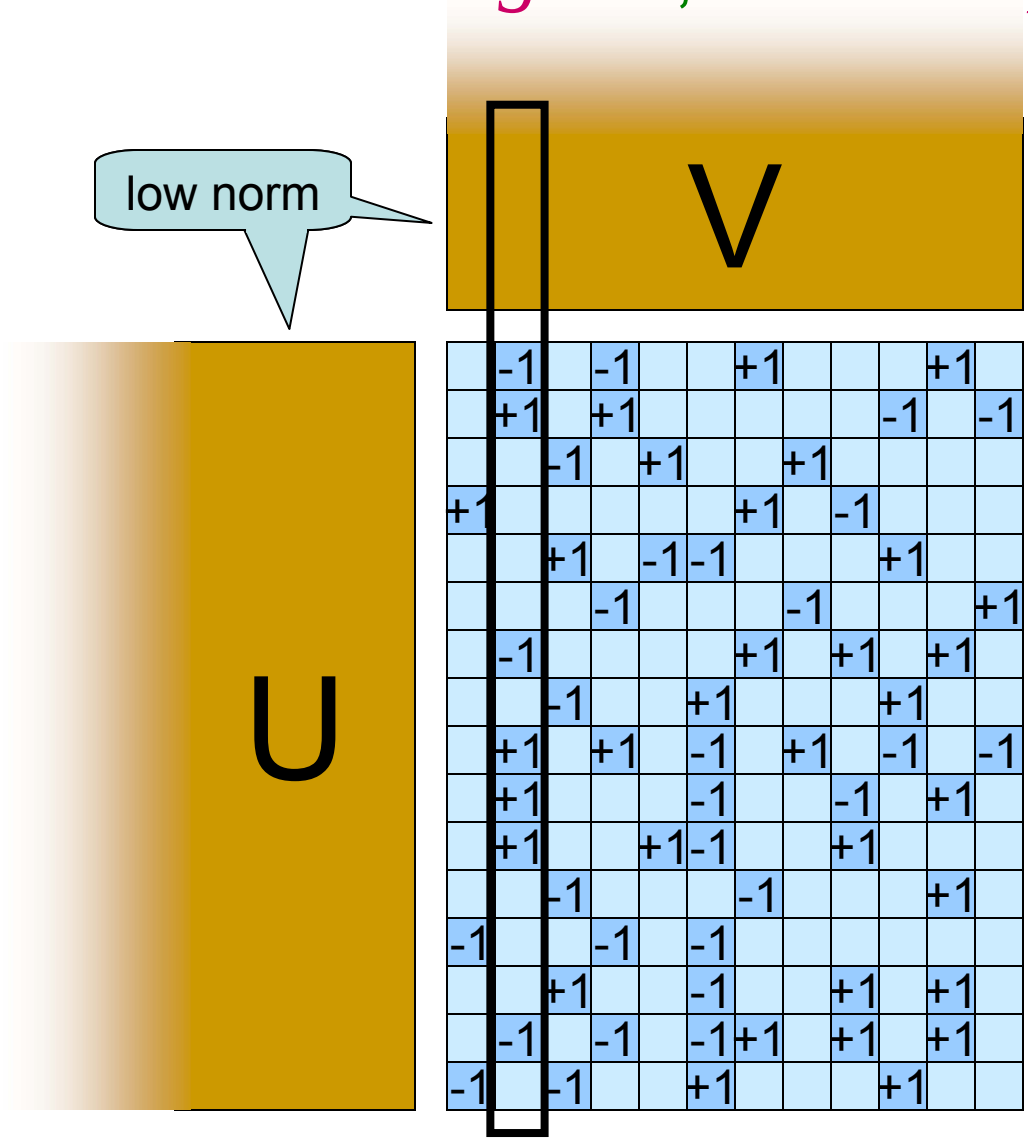
Learning features that work well across all classification problems.

Max-Margin Matrix Factorization:

Geometric Interpretation:

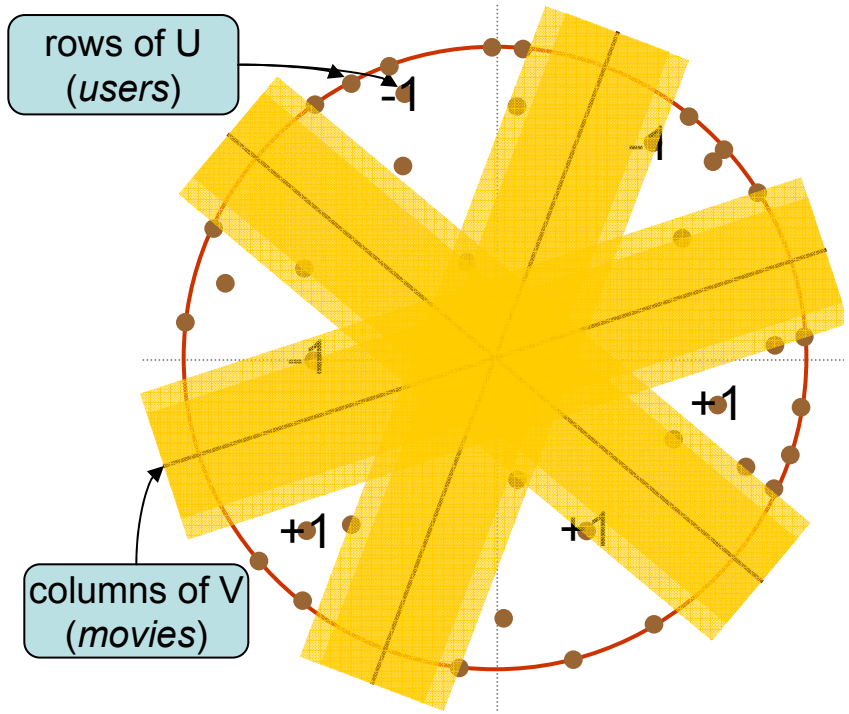
Bounding Norms of U, V instead of Inner-Products

Embedding Points and Separating Hyperplanes



bound norms uniformly:

$(\max_i |U_i|^2)\,(\max_j |V_j|^2) \le R^2$



For observed $Y_{ij} \in \pm 1$:

$Y_{ij}\, X_{ij} \ge 1$ *Margin*

$\langle U_i, V_j \rangle$

Conclusions and future work
Some concrete directions

- *Any tools that assume relevance of structure of the unlabeled data for class prediction, can be applied in the context of MTL.*
  In particular
  - *Semi-supervised learning*
  - *Clustering*
  - *Dimensionality reduction*

- *Tools assuming structure of the labels, like label-matrix low rank, can be readily applied to MTL.*

- *The Transformation-relatedness framework is an 'orthogonal' component. Can be applied on top of the above methods but can provide multi-tasking benefits also in cases where no semi-supervised structure exists.*

Conclusions and future work - basic principles

- A notion of task relatedness is key to any LTL theory.
- Its Inconceivable to hope for a unique general notion of class relatedness.
- The importance of practical experimental evidence of LTL.
  - To indicate new types of task-relatedness
  - To verify the validity of our prior assumptions
  - To motivate and justify research