Generalization Error Bounds for Collaborative Prediction with Low-Rank Matrices

Nathan Srebro
Department of Computer Science, University of Toronto

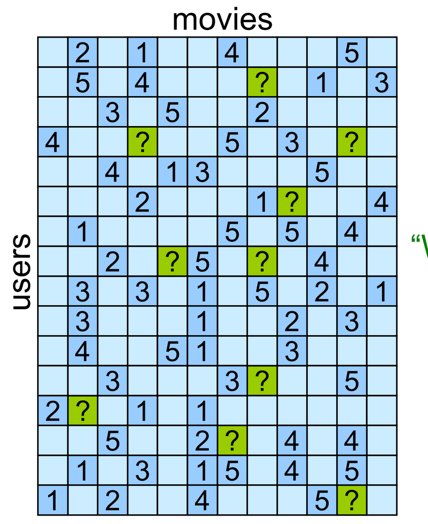
Noga Alon
School of Mathematical Sciences, Tel Aviv University

Tommi Jaakkola
Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Results

Collaborative Prediction

Based on partially observed matrix
⇒ Predict unobserved entries



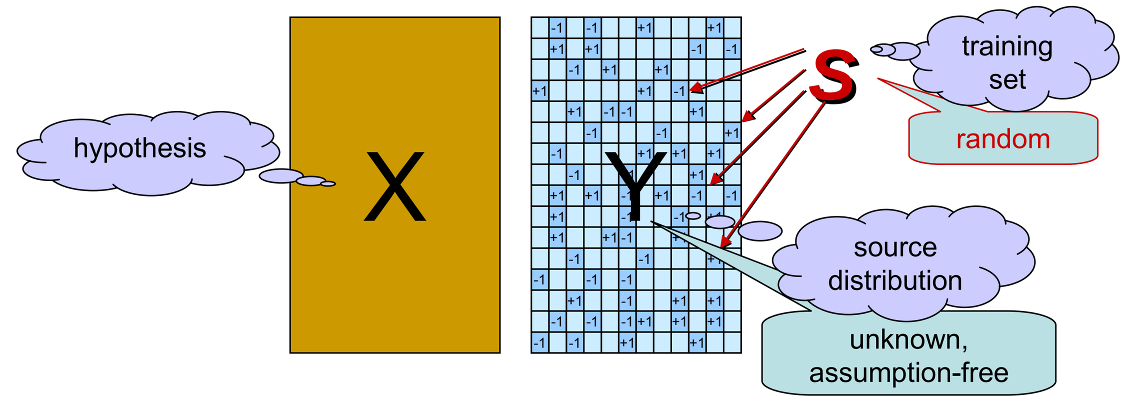
"Will user i like movie j?"

Generalization Error Bounds

$D(\mathbf{X};\mathbf{Y}) = \sum_{ij} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/nm$ *generalization error*

$D_S(\mathbf{X};\mathbf{Y}) = \sum_{ij\in S} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/|S|$ *empirical error*

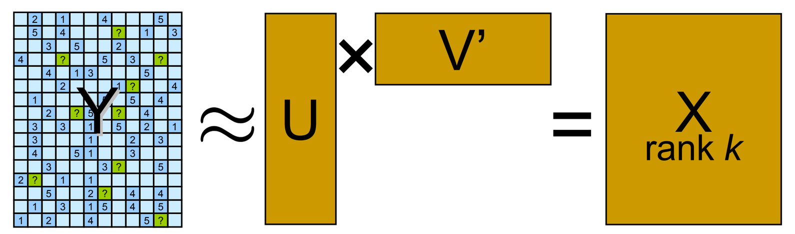
$$\forall_{\mathbf{Y}} \Pr_S \left(\forall_{\text{rank-}k\ \mathbf{X}}\ D(\mathbf{X};\mathbf{Y}) < D_S(\mathbf{X};\mathbf{Y}) + \varepsilon \right) > 1-\delta$$



$$loss(x,y) = \begin{cases} 0 & sign(x) = y \\ 1 & sign(x) \neq y \end{cases} \qquad \varepsilon = \sqrt{\frac{k(n+m)\log\frac{8em}{k} + \log\frac{1}{\delta}}{2|S|}}$$

monotone $loss(x,y) \le 1$: $\quad \varepsilon = 6\sqrt{\frac{k(n+m)\log\frac{8em}{k}\log\frac{|S|}{k(n+m)} + \log\frac{1}{\delta}}{|S|}}$

Low-Rank Matrix Factorization



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

$$\text{minimize} \sum_{ij\in S} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})$$

observed entries — prediction — observation

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

[Sarwar, Karypis, Konstan, Riedl **Applications of dimesionality reduction in recommender systems—a case study** *WebKDD* 2000]
[Hoffman **Latent semantic models for collaborative filtering** *ACM Trns. Inf. Syst.* 2004]
[Marlin, Zemel **Modeling user rating profiles for collaborative filtering** *NIPS* 2003]
[Canney **GAP: A Factor Model for Discrete Data** *SIGIR* 2004]
many others…

Prior work

- Assuming a low-rank structure (eigengap) in **Y**, predict entries:
  - Asymptotic behavior
    [Azar, Fiat, Karlin, McSherry Saia **Spectral analysis of data** *STOC* 2001]
  - Sample complexity, query strategy
    [Drineas, Kerenidis, Raghavan **Competitive recommendation systems** *STOC 2002*]
- Bounds on residual errors, no assumptions on **Y**:
  [Shaw-Taylor, Cristianini, Kandola **On the concentration of spectral properties** *NIPS* 2002]
  Subset of rows fully observed, bound is on distance of new rows to learned *subspace*

In this work: collaborative prediction analysis (entry prediction), no assumptions on **Y**.

Major Assumption: Random Observations

Although we did not make any assumptions about the true preferences **Y**, we made a very strong assumption about the set ***S*** of observed entries: we assumed entries as selected uniformly at random. Although the uniformity requirement can be relaxed:

$D(\mathbf{X};\mathbf{Y}) = \mathbf{E}_{ij}[\text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})]$ $\quad D_S(\mathbf{X};\mathbf{Y}) = \sum_{ij\in S} \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij})/|S|$

same observation distribution

$$\forall_{\mathbf{Y}} \Pr_S \left(\forall_{\mathbf{X}}\ D(\mathbf{X};\mathbf{Y}) < D_S(\mathbf{X};\mathbf{Y}) + \varepsilon \right) > 1-\delta$$

This is not very satisfying: we are guaranteed good generalization only on items the user is likely to observe on its own—not on items we might recommend.

Different low-rank methods differ in how they relate real-valued entries in X to the observations (preferences) Y, possibly through a probabilistic model, and in the associated contrast (loss) functions.

Low-rank models of co-occurrence or frequency data

	Multinomial	Independent Binomials	Independent Bernoulli
Mean parameterization $0 \le X_{ij} \le 1$ $E[Y_{ij}\mid X_{ij}]=X_{ij}$	Aspect Model (pLSA) [Hoffman+99] ≡ NMF if $\sum X_{ij}=1$	$Y_{ij}\mid X_{ij} \sim Bin(N,X_{ij})$ ≈ NMF [Lee+01]	$P(Y_{ij}=1) = X_{ij}$
Natural parameterization unconstrained X_{ij}	SDR [Globerson+02]	$Y_{ij}\mid X_{ij} \sim Bin(N,g(X_{ij}))$	Logistic Low Rank Approximation [Schein+03]

≈ hinge loss

row features most informative about columns

Exponential PCA: [Collins+01]
$p(Y_{ij}\mid X_{ij}) \propto \exp(Y_{ij}X_{ij}+F(Y_{ij}))$

$g(x)=1/(1+e^x)$

Proofs

Binary Labels, Zero-One Loss

For particular **X**,**Y**: $\quad \text{loss}(\mathbf{X}_{ij};\mathbf{Y}_{ij}) \sim \text{Bernoulli}(D(\mathbf{X};\mathbf{Y}))$

random — random

$$\Pr\left(D_S(\mathbf{X};\mathbf{Y}) < D(\mathbf{X};\mathbf{Y}) - \varepsilon \right) < e^{-2|S|\varepsilon^2}$$

Union bound over all possible **X**s: $\quad \forall_{\mathbf{Y}} \Pr_S \left(\forall_{\mathbf{X}}\ D(\mathbf{X},\mathbf{Y}) < D_S(\mathbf{X},\mathbf{Y}) + \varepsilon \right) > 1-\delta$

$$\varepsilon = \sqrt{\frac{\log(\#\text{ possible } \mathbf{X}\text{s}) + \log\frac{1}{\delta}}{2|S|}}$$

The bound rests on bounding the number of possible **X**s. The behavior of $loss(\mathbf{X}_{ij},\mathbf{Y}_{ij})$ only depends on sign(**X**), and so it is enough to bound the number of sign configurations:

$$F(n,m,k) = \{ \text{sign}(\mathbf{X}) \in \{-,+\}^{n\times m} \mid \mathbf{X} \in \mathbb{R}^{n\times m}, \text{rank } \mathbf{X} \le k\}$$
$$f(n,m,k) = \#F(n,m,k)$$

Sign Configurations of Low-Rank Matrices

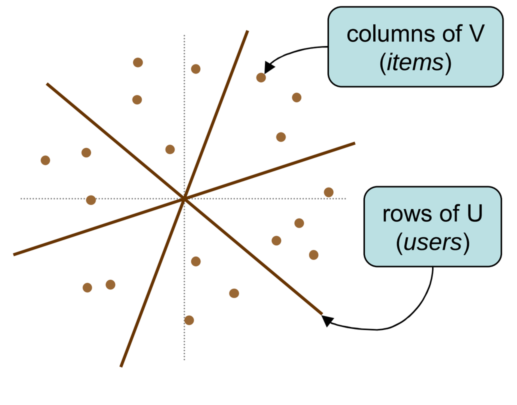
Following [Alon **Tools from higher algebra** *Handbook of Combinatorics* 1995], similar to [Alon, Frankl, Rödel **Geometric realization of set systems and probabilistic communication complexity** *FOCS* 1985]

V ($k \times m$):

1.11	-0.81	0.[illegible]7	2.24	0.57
0.22	-0.58	[illegible]70	-0.52	0.00

U ($n \times k$) and **UV** ($n \times m$):

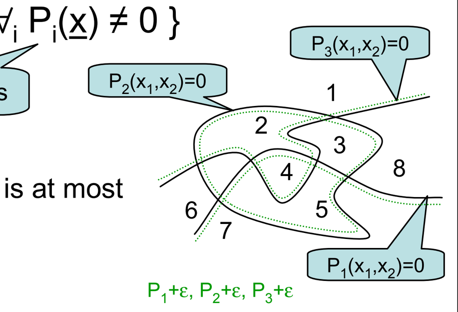
U		UV				
-0.46	0.65	-0.36	-0.00	-0.98	-1.36	-0.26
-0.20	0.70	-0.07	-0.24	-1.14	-0.82	-0.12
0.02	0.74	0.19	-0.44	-1.27	-0.34	0.01
-0.61	-0.59	-0.81	0.84	1.17	-1.07	-0.35
-0.50	0.33	-0.48	0.21	-0.43	-1.28	-0.28
-0.38	0.31	-0.35	0.13	-0.43	-1.02	-0.22
-0.35	-0.69	-0.54	0.68	1.26	-0.43	-0.20
-0.[illegible]9	0.56	-0.22	0.40	0.98	0.10	-0.05
0.[illegible]0	0.57	1.13	-1.06	-1.21	1.72	0.51
1.27	-0.52	1.30	-0.73	0.55	3.12	0.72
0.44	0.17	0.53	-0.45	-0.40	0.90	0.25
-0.06	0.09	-0.04	-0.00	-0.14	-0.17	-0.03
-0.66	-0.29	-0.80	0.70	0.67	-1.32	-0.37
-1.65	0.09	-1.81	1.28	0.28	-3.73	-0.93
0.58	0.16	0.68	-0.56	-0.43	1.21	0.33



Warren (1968): The number of connected components of $\{ \underline{x} \mid \forall_i\ P_i(\underline{x}) \neq 0 \}$ is at most $\left(\frac{4e\ (\text{degree})\ (\#\text{polys})}{(\#\text{ variables})}\right)^{(\#\text{ variables})}$

polynomials

Corollary: The number of sign configurations of the polynomials is at most $\left(\frac{4e\ (\text{degree})\ (\#\text{polys})}{(\#\text{ variables})}\right)^{(\#\text{ variables})}$



nk variables — *mk* variables

$$\mathbf{X}_{ij} = \sum_r \mathbf{U}_{ir} \mathbf{V}_{rj}$$

nm polynomials of degree 2

$$f(n,m,k) \le \left(\frac{4e \cdot 2 \cdot nm}{k(n+m)}\right)^{k(n+m)} = 2^{k(n+m)\log(8em/k)}$$

General Bounded Loss Functions

For 0/1 loss, behavior of entries of **X** around zero enough.
More generally, need to bound complexity of behavior everywhere.

For any threshold matrix $\mathbf{T} \in \mathbb{R}^{n\times m}$, bound number of relative sign configurations: #{ sign(**X-T**) | rank(**X**)=k }

nk variables — *mk* variables

$$(\mathbf{X}-\mathbf{T})_{ij} = \sum_r \mathbf{U}_{ir}\mathbf{V}_{rj} - \mathbf{T}_{ij}$$

nm polynomials of degree 2

$$\#\{ \text{sign}(\mathbf{X}-\mathbf{T}) \mid \text{rank}(\mathbf{X})=k \} \le \left(\frac{4e\cdot 2\cdot nm}{k(n+m)}\right)^{k(n+m)} = 2^{k(n+m)\log(8em/k)}$$

Viewing matrices as a mappings from index pairs to values: $(i,j) \mapsto \mathbf{X}_{ij}$, this gives us a bound of $k(n+m)\log(8em/k)$ on the pseudo-dimension of rank-k matrices. We can now invoke standard results bounding the generalization error in terms of the pseudo-dimension.

A class $\mathcal{F}$ of real-valued functions **pseudo-shatters** the points $x_1,\dots,x_n$ with thresholds $t_1,\dots,t_n$ if for every binary labeling of the points $(s_1,\dots,s_n) \in \{+,-\}^n$ there exists $f \in \mathcal{F}$ s.t. $f(x_i) \le t_i$ iff $s_i = -$.

The **pseudo-dimension** of a class $\mathcal{F}$ is the supremum over n for which there exist n points and thresholds that can be pseudo-shattered.

A Weaker Bound Using Realizable Oriented Matriods

For a fixed **V**, each row is linear classification of columns in **V**, and there are $\le 2(k+1)m^{k-1}$ such classifications. Overall, for each fixed **V**, the number of possible sign matrices is bounded by:

$$\#\{ \text{sign}(\mathbf{UV}) \mid \mathbf{U} \in \mathbb{R}^{n\times k}\} \le (2(k+1)m^{k-1})^n$$

This should be multiplied by the number of **V**s, or rather the number of **V**s yielding different sets of possible classification vectors:

$$M(\mathbf{V}) = \{ \text{sign}(u'\mathbf{V}) \mid u \in R^k \}$$

set of covectors of a realizable oriented matriod (where sign ∈ {-,0,+})

$$\#\{ M(\mathbf{V}) \mid \mathbf{V} \in R^{k\times m} \} \le m^{k(k+1)m}$$

[Goodman Pollack 1986], [Alon 1986] bound on number of realizable oriented matriods

$$\Downarrow$$

$$f(n,m,k) \le (2(k+1)m^{k-1})^n\, m^{k(k+1)m} \le 2^{k(n+m)\log(2m) + k^2 m \log(2m)}$$

extraneous k^2 term!

Why not treat as combined classifiers?

For MMMF (max-margin/low-norm matrix factorizations), generalization error bounds obtained by viewing MMMF as a "combined" classifier, a convex combination of unit-norm rank-1 matrices.
Rank-*K* matrices can be viewed as "combined", or "voting" classifiers, each combining *k* rank-1 matrices.
Can a similar approach be taken for low-rank matrices?

- Scale-sensitive complexity (log covering numbers, Rademacher complexity) carries over to convex hull. (scale-invariant complexity certainly not conserved for convex hull)
- VC-dimension scales gracefully with *k* for combinations of *k* classifiers
  ⇒ generalization error bounds for linear combinations of signs of low-rank matrices
- Pseudo-dimension of a linear combinations of *k* functions from a low-pseudo-dimension class?

Counter Example: A family $\mathcal{F}$ of functions closed under scalar multiplication, with pseudo-dimension 3, such that $\{ f_1+f_2 \mid f_1,f_2 \in \mathcal{F} \}$ has infinite pseudo-dimension:

$$\mathcal{F} = \{ \alpha\cdot f_A,\ \alpha\cdot g_A \mid \alpha \in \mathbb{R},\ A \in \mathbb{N} \}$$
$$f_A(x) = 2^{xA} + 1_{x\in A}$$
$$g_A(x) = 2^{xA}$$

Consider a 1:1 mapping $\mathbb{N} \leftrightarrow 2^{\mathbb{N}}$. 'A' denotes both a number, and the corresponding subset.

Related Work

Warren's Theorem and Configuration Counting

Warren's Theorem, and a weaker result of Milnor, have a long history in combinatorics and learning theory:

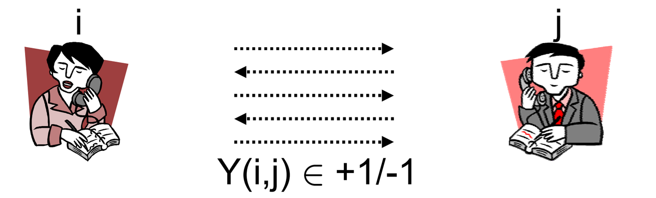
[Goodman Pollack **Upper bounds for configurations and polytopes in R^d** *Disc Comp Geom* 1986]
[Alon 1986 **The number of polytopes, configurations and real matroids** *Mathematika* 1986]
Bound on the number of non-equivalent point configurations (realizable oriented matriods). Can be used to obtain weak bound on number of sign configurations of low-rank matrices **(green panel)**.

[Ben-David, Lindenbaum **Localization vs. identification of semi-algebraic sets** *COLT* 1993]
VC-dimension of set of transformations of an image, used to analyze sample complexity of determining location.

[Goldberg, Jerrum **Bounding the VC dimension of concept classes parameterized by real numbers** *COLT* 1993]
VC-dimension of any concept classes parameterized by real numbers, where each concept can be written as logical formula over polynomial inequalities
≤ 2 (# of params describing each concept) log(8e (degree of polys used) (# of polys used))
Can be applied to collaborative prediction with low-rank matrices, where:
$X(i,j) = \vee_{i',j'} (i=i' \wedge j=j' \wedge \sum_r U_{i'r}V_{rj'}>0)$
yielding: VC-dim(rank-k matrices) $\le 2\cdot k(n+m)\cdot\log(8e\cdot 2\cdot 3nm) \le 2k(n+m)\log(48enm)$
By directly applying Warren's Theorem we:
- avoided symmetrization (for 0/1 error)
- avoided Sauer's lemma, and a log|S| term in the generalization error bound
- bounded the pseudo-dimension and obtained generalization error bounds for general loss functions

More on Sign Configurations of Low-Rank Matrices

Unbounded Error Communication Complexity



$Y(i,j) \in +1/-1$

Unbounded error communication complexity C
= Randomized protocol, always $\le C$ bits, P(correct answer)> ½

[Paturi, Simon **Probabilistic communication complexity** *FOCS* 84]
log rank **X**⌉ ≤ C ≤ ⌈log rank **X**⌉ , sign(**X**)=**Y**

[Alon, Frankl, Rödel **Geometric realization of set systems and probabilistic communication complexity** *FOCS* 1985]
Bound # sign configurations
counting arguments ⇒ ∃ **Y** with rank(**X**)>n/32 ⇒ ∃ Y:$\{0,1\}^r \times \{0,1\}^r \to \{0,1\}$ with C>r-5

Embedability as Linear Classification

Can all concept classes be embedded as linear classifications in a low dimensional space?
$\mathcal{C}=\{c_1,\dots,c_n\}$ can be embedded as k-dimensional linear classification ⇔ Rank-k **X**, s.t. $c_i(j)=\text{sign}(\mathbf{X}_{ij})$
counting arguments ⇒ ∃ small concept class, not embeddable as low dimensional linear classification

Explicit Examples

These counting arguments provide only existence proofs, not explicit constructions of sign configurations with no low-rank realization (i.e. functions with high unbounded error communication complexity, or concept classes that cannot be embedded as low-dimensional linear classification).

[Forester **A linear bound on the unbounded error communication complexity** *CCC* 2001]
rank(X) ≥ n / |sign(**X**)|$_2$ (spectral norm of sign(**X**))
In particular, the $2^r\times 2^r$ Hadamard matrix cannot be realized with rank(**X**)$<2^{r/2}$

In this example, rank(**X**)≥√n. No known explicit example with rank(**X**)=Ω(n).

http://www.cs.toronto.edu/~nati/genlowrank/