CME 305 Project: Learning matrices

Jason Lee, Carlos Sing-Long, Yuekai Sun · February 25, 2012

1 Theory

1.1 Abstract framework

A foundational concept in modern machine learning is to construct models from data by balancing the complexity of the model with the discrepancy between the model and the data. Therefore, one of the main concerns is to define meaningful ways to quantify complexity and discrepancy. To formulate an abstract framework, consider a class $\mathcal{F}$ of models, endowed with a set of operations on its elements. We define a function $\mathcal{C} : \mathcal{F} \mapsto \mathbb{R}_+$ that quantifies the complexity of these models. Any model with low values of $\mathcal{C}$ is deemed *simple* in this class. Furthermore, we assume that $\mathcal{F}$ is spanned by the simple models induced by $\mathcal{C}$. In addition, consider a function $\mathcal{D}_Y : \mathcal{F} \mapsto \mathbb{R}_+$ that measures the discrepancy between the data Y and any given model. In other words, any model with high values of $\mathcal{D}_Y$ cannot explain properly the data. Thus the problem reduces to determine reasonable choices of $\mathcal{C}$ and $\mathcal{D}_Y$ for a given application. Once these are made, one common scheme to construct the aforementioned model is to solve an optimization problem. Usually, this problem has the form

$$\underset{X \in \mathcal{F}}{\text{minimize}} \quad \mathcal{D}_Y(X) \quad \text{subject to} \quad \mathcal{C}(X) \in S \tag{1}$$

where $S \subset \mathbb{R}_+$ indicates the allowed complexity of the elements used to explain Y or

$$\underset{X \in \mathcal{F}}{\text{minimize}} \quad \mathcal{D}_Y(X) + \lambda \mathcal{C}(X) \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter that balances the trade-off between the discrepancy and complexity. By using an optimization problem, we have implicitly imposed a new constraint on $\mathcal{C}$ and $\mathcal{D}_Y$. Specifically, they must be chosen in such a way that (1) or (2) becomes a tractable problem.

1.2 rank and trace-norm

We now make more concrete assumptions on our setting. In what follows, we assume the data Y is known and we denote $f = \mathcal{D}_Y$. In several applications, both the models and the data can be represented in a meaningful way as a matrix with real entries, and therefore $\mathcal{F}$ becomes a class of matrices. Thus the question is how to define a measure of complexity for these models. A theoretically sound measure of complexity is the rank. This is consistent with our abstract formulation as classical results in linear algebra show that any matrix can be decomposed as a linear combination of rank-1 matrices. Therefore, low-rank matrices are deemed simple and (1) becomes

$$\underset{X}{\text{minimize}} \quad f(X) \quad \text{subject to} \quad \mathbf{rank}\, X \leq k \tag{3}$$

Unfortunately, this leads to an untractable optimization problem. For this reason, several research has been done to find a suitable regularization of the rank. One method that has received a huge amount of attention is the trace-norm $\|\cdot\|_{\text{tr}}$, defined as the sum of the singular values of X. The trace-norm promotes

low-rank by minimizing the ℓ_1-norm of the vector containing the singular values of X, thus encouraging its sparsity. It is of particular interest to consider (2) is this case

$$\underset{X \in \mathcal{F}}{\text{minimize}} \quad f(X) + \lambda \|X\|_{\text{tr}} \tag{4}$$

which becomes tractable. This relaxation proves to be extremely useful in a wide array where rank constraints are needed [1].

Although the trace-norm is a successful regularizer, it does not seem to be widely known that there might be additional ways to regularize the rank.

1.3 The max-norm

A different approach is to further explore the consequences of considering a matrix X of size $m \times n$ as an operator from ℓ_p to ℓ_q seen as finite-dimensional Banach spaces. A natural way to quantify the magnitude of X is to consider its corresponding operator norm

$$\|X\|_{\ell_p \to \ell_q} = \sup_{z \neq 0} \frac{\|Xz\|_{\ell_q}}{\|z\|_{\ell_p}}$$

where $\|\cdot\|_{\ell_p}$ is the standard ℓ_p-norm. Unfortunately the aforementioned norms are computationally intractable in several cases [2]. Furthermore, its is known that it is computationally hard to approximate them in these cases. In particular, its is NP-hard to approximate it to any constant factor for $1 < q < p < 2$, $1 < q < 2 < p$ and $2 \leq q < p$ whereas it is $O(2^{(\log n)^{1-\varepsilon}})$ hard in $1 < q = p < 2$ and $2 < q = p$.

However, the operator interpretation allows us to use factorizations as a meaningful way of quantifying the complexity of X. Namely, we can fix a Banach space Z and a pair of operators $U : Z \mapsto \ell_q$ and $V : \ell_p \mapsto Z$ such that $X = UV$. In this case, we would like to find the simplest among all such decompositions. The complexity is quantified as

$$\gamma_Z(X) = \inf\{\|U\|_{Z \to \ell_q} \|V\|_{\ell_p \to Z} : X = UV\}$$

A classical result from Banach spaces theory shows that for fixed Z, γ_Z defines a norm over the space of operators from ℓ_p to ℓ_q. Of particular interest for us will be the case $Z = \ell_2$, $q = \infty$ and $p = 1$. We obtain γ_2

$$\gamma_2(X) = \inf\{\|U\|_{\ell_2 \to \ell_\infty} \|V\|_{\ell_2 \to \ell_\infty} : X = UV\}$$

which is formulated in terms of norms that are computationally tractable. To make the connection with the singular value decomposition clearer, it is custom to define $X = UV^T$. In this case, a standard result from analysis shows that $\|V^T\|_{\ell_1 \to \ell_2} = \|V\|_{\ell_2 \to \ell_\infty}$, so that

$$\gamma_2(X) = \inf\{\|U\|_{\ell_2 \to \ell_\infty} \|V\|_{\ell_2 \to \ell_\infty} : X = UV^T\}$$

Thus we have our first definition

Definition 1.1. *The **max-norm** of a matrix X is*

$$\|X\|_{\max} = \gamma_2(X) = \inf\{\|U\|_{\ell_2 \to \ell_\infty} \|V\|_{\ell_2 \to \ell_\infty} : X = UV^T\}$$

To explain why max-norm is a suitable name for γ_2, denote x_i^T the rows of X and consider the norm $\ell_2 \to \ell_\infty$

$$\begin{aligned}
\|X\|_{\ell_2\to\ell_\infty} &= \sup_{\|z\|_{\ell_2}=1} \|Xz\|_{\ell_\infty} \\
&= \sup_{\|z\|_{\ell^2}=1} \max_i |\langle x_i,\, z\rangle| \\
&\leq \sup_{\|z\|_{\ell_2}=1} \max_i \|x_i\|_{\ell_2}\|z\|_{\ell_2} \\
&= \max_i \|x_i\|_{\ell_2}
\end{aligned}$$

Since the inequality becomes equality for $z_i = x^\star/\|x^\star\|_{\ell_2}$ with $x^\star \in \arg\max_i \|x\|_{\ell_2}$, we see that $\|X\|_{\ell_2\to\ell_\infty}$ is the maximum ℓ_2-norm of the *rows* of X. This clearly implies

$$\|X\|_{\max} = \inf_{X=UV^T} \max_{i,j} \|u_i\|_{\ell_2}\|v_j\|_{\ell_2} \tag{5}$$

where u_i are the *rows* of U and v_j are the *rows* of V. This expression suggests that the max-norm measures complexity by looking at the maximum norms of the rows of the factors U and V for $X = UV^T$. But more can be said aside from these facts. We will show that the max-norm has a handful of interesting properties. For example, the max-norm is invariant under the adjoint operator, i.e. $\|X^T\|_{\max} = \|X\|_{\max}$, and so is its dual. This is clear from the definition, as

$$\begin{aligned}
\|X^T\|_{\max} &= \inf\{\|U\|_{\ell_2\mapsto\ell_\infty^m}\|V\|_{\ell_2\mapsto\ell_\infty^n} :\, X^T = UV^T\} \\
&= \inf\{\|U\|_{\ell_2\mapsto\ell_\infty^m}\|V\|_{\ell_2\mapsto\ell_\infty^n} :\, X = VU^T\} \\
&= \inf\{\|U\|_{\ell_2\mapsto\ell_\infty^m}\|V\|_{\ell_2\mapsto\ell_\infty^n} :\, X = VU^T\} \\
&= \|X\|_{\max}
\end{aligned}$$

and

$$\begin{aligned}
\|X^T\|_{\max}^* &= \sup_{\|Y\|_{\max}\leq 1} \operatorname{trace}(X^T)^T Y \\
&= \sup_{\|Y\|_{\max}\leq 1} \operatorname{trace} YX \\
&= \sup_{\|Y^T\|_{\max}\leq 1} \operatorname{trace} Y^T X \\
&= \sup_{\|Y\|_{\max}\leq 1} \operatorname{trace} Y^T X \\
&= \|X\|_{\max}^*
\end{aligned}$$

These properties allows us to express the dual of the max-norm as

$$\begin{aligned}
\|X\|_{\max}^* &= \|X^T\|_{\max}^* \\
&= \sup_{\|Y\|_{\max}\leq 1} \operatorname{trace} XY \\
&= \sup_{\|Y^T\|_{\max}\leq 1} \operatorname{trace} XY^T \\
&= \sup_{\|Y\|_{\max}\leq 1} \sum_i \langle x_i,\, y_i\rangle \\
&= \sup_{\|Y\|_{\max}\leq 1} \sum_{i,j} X_{i,j}Y_{i,j}
\end{aligned}$$

where x_i and y_i represent the *rows* of X and Y respectively. By using the decomposition $Y = UV^T$ we obtain

$$\|X\|_{\max}^* = \sup_{\|Y\|_{\max}\leq 1} \sum_{i,j} X_{i,j}\langle u_i,\, v_j\rangle$$

where u_i and v_j where defined previously. This expression hides a very deep relationship with the theory of absolutely summing operators in Banach space theory. The connection is made through Grothendieck's inequality. We state the theorem in a slightly different form as in the standard reference [3], with the notation changed to fit ours [4]

Theorem 1.1. *There is a universal constant K_G for which, given any Hilbert space H, any $n \in \mathbb{N}$, any $n \times n$ scalar matrix $(X_{i,j})$ and any vectors $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ in $B_H = \{x : \|x\|_H \leq 1\}$, such that*

$$\left|\sum_{i,j} X_{i,j}\langle u_i,\, v_j\rangle\right| \leq K_G \max\left\{\left|\sum_{i,j} X_{i,j}s_it_j\right| :\, |s_i| \leq 1,\, |t_j| \leq 1\right\}$$

The best possible value for the constant K_G is called Grothendieck's constant and its unknown so far. However, bounds are known and $1.5 < K_G < 1.8$. This clearly implies

$$\|X\|_{\max}^* \leq K_G \max\left\{\left|\sum_{i,j} X_{i,j}s_it_j\right| :\, |s_i| \leq 1,\, |t_j| \leq 1\right\}$$

This inequality raises the question: is there any object defined as the right-hand side? The answer is yes and, surprisingly, it is the $\ell_\infty \to \ell_1$-norm. Indeed, let v be any vector such that $\|v\|_{\ell_\infty} \leq 1$. Then $|v_j| \leq 1$ and

$$\|Xv\|_{\ell_1} = \sum_i \left|\sum_j X_{i,j}v_j\right|$$

It is clear that among all u_i such that $|u_i| \leq 1$ we have

$$\|Xv\|_{\ell_1} \geq \sum_i u_i \sum_j X_{i,j}v_j = \sum_{i,j} X_{i,j}u_iv_j$$

Since the equality becomes an equality for $u_i = \text{sign}\left|\sum_j X_{i,j} v_j\right|$ we have

$$\|X\|_{\ell_\infty \to \ell_1} = \sup_{\|v\|_{\ell_\infty} \leq 1} \|Xv\|_{\ell_1} = \max\left\{ \sum_{i,j} X_{i,j} u_i v_j : |u_i| \leq 1,\, |v_j| \leq 1 \right\}$$

from which the inequality

$$\|X\|^*_{\max} \leq K_G \|X\|_{\ell_\infty \to \ell_1}$$

follows. This suggests that the dual of the max-norm is equivalent to the $\|X\|_{\ell_\infty \to \ell_1}$ norm. In fact, to prove the remaining inequality, we note that, for a matrix decomposition $Y = UV^T$ we have

$$\left\{ \sum_{i,j} X_{i,j} s_i t_j : |s_i| \leq 1,\, |t_j| \leq 1 \right\} \subset \left\{ \sum_{i,j} X_{i,j} \langle u_i,\, v_j \rangle : Y = UV^T,\, \|Y\|_{\max} \leq \sqrt{mn} \right\}$$

Indeed, for a given set of values $|s_i|, |t_j| \leq 1$ we can construct $u_i = s_i \sum_i e_i$ and $v_j = t_j \sum_i e_i$, where $\{e_i\}$ are the elements of the canonical basis. In this case clearly $\|u_1\|_{\ell_2} \|v_j\|_{\ell_2} \leq |s_i||t_j|\sqrt{mn}$ and the inclusion follows. Remark that there might be better constants to obtain the inclusion. Therefore, by taking supremum over the two sides

$$\|X\|_{\ell_\infty \to \ell_1} \leq \sup_{\|Y\|_{\max} \leq \sqrt{mn}} \sum_{i,j} X_{i,j} Y_{i,j}$$

We may rescale the set $\{\|Y\|_{\max} \leq \sqrt{mn}\,\}$ so that

$$\|X\|_{\ell_\infty \to \ell_1} \leq \sqrt{mn} \sup_{\|Y\|_{\max} \leq 1} \sum_{i,j} X_{i,j} Y_{i,j} = \sqrt{mn}\, \|X\|^*_{\max}$$

This inequality implies

$$\|X\|_{\ell_\infty \to \ell_1} \leq \|X\|^*_{\max} \leq K_G \|X\|_{\ell_\infty \mapsto \ell_1}$$

or, by taking duals

$$\frac{1}{K_G} \|X\|^*_{\ell_\infty \mapsto \ell_1} \leq \|X\|_{\max} \leq \|X\|^*_{\ell_\infty \mapsto \ell_1}$$

The significance of this inequality lies on the fact that the dual $\ell_\infty \to \ell_1$-norm belongs to a class called nuclear norms. In fact, we can write

$$\|X\|^*_{\ell_\infty \to \ell_1} = \inf\left\{ \|\sigma\|_1 : X = \sum_i \sigma_i u_i v_i^T,\, \|u_i\|_{\ell_\infty} = 1,\, \|v_i\|_{\ell_\infty} = 1 \right\}$$

This result has its origins on trace duality and it is beyond the scope of this work. The interested reader may check the corresponding references [3, 4]. The equivalence with the matrix norm implies

$$\|X\|_{\max} \approx \inf\left\{ \|\sigma\|_{\ell_1} : X = \sum_i \sigma_i u_i v_i^T,\, \|u_i\|_{\ell_\infty} = 1,\, \|v_i\|_{\ell_\infty} = 1 \right\}$$

The trace-norm belongs to the same class and, in fact, we have the equivalent expression

$$\|X\|_{\text{tr}} = \inf\left\{ \|\sigma\|_{\ell_1} : X = \sum_i \sigma_i u_i v_i^T,\, \|u_i\|_{\ell_2} = 1,\, \|v_i\|_{\ell_2} = 1 \right\}$$

This shows that our new approach to measuring complexity led us to a new method of regularizing the rank of a matrix. Indeed, the trace-norm promotes low-rank by using factors with unit ℓ_2-norm whereas the max-norm promotes low-rank by using factors with unit ℓ_∞-norm. This suggests that the max-norm ought to be able to outperform the trace-norm in cases when the data is uniformly bounded.

1.4 Computation of the max-norm

The theoretical analysis performed in the previous section suggest a tractable method to compute the max-norm. To do this, we try to solve

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & \max_{i,j} \|u_i\|_{\ell_2}^2/2 + \|v_j\|_{\ell_2}^2/2 \\ \text{subject to} \quad & X = UV^T \end{aligned}$$

The objective is to find an equivalent convex formulation to this problem. Using standard optimization techniques, this program can be formulated equivalently as

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & t \\ \text{subject to} \quad & X = UV^T \\ & \|u_i\|_{\ell_2}^2 \leq t \\ & \|v_j\|_{\ell_2}^2 \leq t \end{aligned}$$

By using the fact that $\|u_i\|_{\ell_2}^2 = (UU^T)_{i,i}$ and $\|v_j\|_{\ell_2}^2 = (VV^T)_{j,j}$ we obtain

$$\begin{aligned} \underset{U,V}{\text{minimize}} \quad & t \\ \text{subject to} \quad & X = UV^T \\ & \mathbf{diag}\, UU^T \leq t \\ & \mathbf{diag}\, VV^T \leq t \end{aligned}$$

This problem is non-convex, as $X = UV^T$ is not a convex constraint. But we can make the following observation

$$\begin{bmatrix} U \\ V \end{bmatrix} \begin{bmatrix} U^T & V^T \end{bmatrix} = \begin{bmatrix} UU^T & X \\ X^T & VV^T \end{bmatrix}$$

This matrix is clearly symmetric and positive definite. From this, we may use the change of variables $UU^T = W_1$ and $VV^T = W_2$ to write the equivalent program

$$\begin{aligned} \underset{W_1,W_2}{\text{minimize}} \quad & t \\ \text{subject to} \quad & \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0 \\ & \mathbf{diag}\, W_1 \leq t \\ & \mathbf{diag}\, W_2 \leq t \end{aligned}$$

This is an SDP and there are several available techniques to solve it. The problem of computing the max-norm is therefore tractable.

1.5 Connections to Quadratic Integer Programming

An interesting connection can be made between the max-norm and some Quadratic Integer Programs (QIP). For example, consider a generic QIP

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & x^T A x \\ \text{subject to} \quad & x_i \in \{-1, 1\} \end{aligned}$$

A standard way to produce a relaxation to this problem is to introduce vectors s, t and define

$$\begin{aligned} \underset{s,t}{\text{maximize}} \quad & s^T A t \\ \text{subject to} \quad & s_i \in \{-1, 1\}, \quad t_j \in \{-1, 1\} \end{aligned}$$

This is the so-called Grothendieck's problem [5]. We can further relax the integer constraints by imposing

$$\begin{aligned} \underset{s,t}{\text{maximize}} \quad & s^T A t \\ \text{subject to} \quad & |s_i| \leq 1, \quad |t_j| \leq 1 \end{aligned}$$

which is precisely $\|A\|_{\max}^*$. As a particular example, consider the QIP formulation of the max-cut problem for an undirected graph $G = (V, E)$

$$\begin{aligned} \underset{x}{\text{maximize}} \quad & \frac{1}{2} \sum_{(i,j) \in E} (1 - x_i x_j) \\ \text{subject to} \quad & x_i \in \{-1, 1\} \end{aligned}$$

The celebrated Goemans-Williamson SDP relaxation of this problem "lifts" each node by a vector u_i and reformulates the problem as

$$\begin{aligned} \underset{u_i}{\text{maximize}} \quad & \frac{1}{2} \sum_{(i,j) \in E} (1 - \langle u_i, u_j \rangle) \\ \text{subject to} \quad & \langle u_i, u_j \rangle \in \{-1, 1\} \end{aligned}$$

We can interpret $\langle u_i, u_j \rangle$ as the component of a matrix X so that the objective becomes $\frac{1}{2} \sum_{(i,j) \in E} (1 - X_{i,j})$ and $X = UU^T$, where u_i are the rows of U. This further imposes the constraint $X \succeq 0$. The constraint $\langle u_i, u_j \rangle \in \{-1, 1\}$ can be relaxed as usual by $|\langle u_i, u_j \rangle| \leq 1$, or equivalently, $|X_{i,j}| \leq 1$. This constraint can be reformulated as $\|X\|_{\max} \leq 1$ by considering $X = UV^T$ and consequently we can solve

$$\begin{aligned} \underset{X}{\text{maximize}} \quad & \frac{1}{2} \sum_{(i,j) \in E} (1 - X_{i,j}) \\ \text{subject to} \quad & \|X\|_{\max} \leq 1, \quad |X_{i,j}| \leq 1 \end{aligned}$$

which also shows a connection between the max-norm and different problems related to discrete optimization.

1.6 Collaborative filtering interpretation

The canonical example of a collaborative filtering problem is the Netflix problem. Let $X \in \mathbb{R}^{m \times n}, m > n$ be an matrix with observed entries $X_{ij}, (i, j) \in \Omega$. Let every row of X represent a user and every column, a

movie. The entries of X are user ratings of movies. A rating is positive if a user likes a movie and negative if the a user dislikes a movie.

Consider the problem

$$\begin{aligned} \underset{Y}{\text{minimize}} \quad & \|Y\|_{\max} \\ \text{subject to} \quad & \operatorname{sgn}(X_{ij})\, Y_{ij} \geq 1,\ (i,j) \in \Omega \end{aligned}$$

We are fitting a regularized matrix, Y to our data, the observed entries of X. Y has a low rank factorization $Y = UV^T$ hence $Y_{ij} = u_i v_j^T$, where u_i is the i-th row of U and v_i is the i-th row of V. We can interpret u_i as a feature vector for user i and the v_i as a linear classifier that classifies the users into users that like and dislike movie j. The constraint $\operatorname{sgn}(X_{ij})\, Y_{ij} \geq 1,\ (i,j) \in \Omega$ ensures that the classifiers for movie j correctly classifies users who have rated the movie. The max-norm lends itself very well to this application because user preferences are uniformly bounded data.

Regularizing using the max-norm can be interpreted as finding a linear classifier for every movie that separates the users who like and dislike the movie with a large margin [6]. Given a set of linearly separable data with feature vectors $u_i \in \mathbb{R}^n$ and labels $l_i \in \{-1, 1\}, i = 1, \ldots, n$, the linear classifier, v that maximizes the margin between the two classes is the solution to

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|v\|^2 \\ \text{subject to} \quad & l_i(v^T u_i - b) \geq 1, i = 1, \ldots, n \end{aligned}$$

Recall the definition of the max-norm as

$$\|X\|_{\max} := \inf\left\{\|U\|_{2,\infty}\|V\|_{2,\infty} : X = UV^T\right\}$$

where $\|X\|_{2,\infty}$ is the maximum 2-norm of the rows [7]. Hence, by minimizing the max-norm of Y, we are maximizing the minimum margin with which the linear classifiers (rows of V) separate users.

2 Experiments

2.1 Semi-supervised learning

We are given a set of data $x_i, i = 1, \ldots, n$ that we would like to cluster. Some of the data $x_i, i \in \Omega$ are labeled but most are unlabeled. We first construct an k-nearest neighbor graph, G, with weights $w_{ij} = \max(d_i(j), d_j(i))$ where $d_i(j) = e^{-\frac{\|x_i - x_j\|}{2\sigma_i^2}}$ and σ_i is the distance from x_i to its kth nearest neighbor. We then build the weighted affinity matrix A of G and sort the labeled data into their respective clusters. Hence, we edit the entries of A corresponding to labeled data.

$$A_{ij} = \begin{cases} 1 & x_i \text{ and } x_j \text{ are in the same class.} \\ -1 & x_i \text{ and } x_j \text{ are not in the same class.} \end{cases} \quad \forall i, j \in \Omega$$

We then learn B, a regularized version of our raw affinity matrix A by solving

$$\underset{B}{\text{minimize}} \quad \|A - B\|_{Fro} + \lambda \|B\|_{\max}$$

and use spectral clustering on B to cluster our data. Ideally, we would like B to be block diagonal, with each block corresponding to a cluster.

We test our method on the Iris dataset and the MNIST dataset. The Iris dataset consists of three classes. The first two are linearly separable but the third class is not linearly separable from the first two. We plot our learned affinity matrix versus raw affinity matrix in Figure 1. We also use our method to cluster handwritten digits in the MNIST dataset. The MNIST dataset consists of images of handwritten digist 0 thru 9. The digits 3 and 5 and 4 and 9 are known to be hard to cluster. We plot our learned affinity matrix versus raw affinity matrix in Figures 2 and 3.

2.2 Matrix Completion

In matrix completion, there is an underlying matrix Y that is only partially observed, and the goal is to recover the entire matrix. This problem arises in the setting of the collaborative filtering. In the Netflix problem, Y corresponds to a matrix of user-movie ratings; this matrix is extremely sparse since most users have only watched a small subset of the movies. Thus the recovery problem corresponds to predicting unobserved user-movie ratings.

This is a highly under-determined problem, but under the prior that Y is low rank (i.e. $\mathbf{rank}\,(Y) < min(n, m)$) the trace-norm heuristic succeeds, under certain assumptions on the distribution of singular vectors(CItations needed). Due to the intimate connection between the max-norm, rank, and trace-norm, we study the performance of the max-norm heuristic for matrix completion. Previous work on real-world datasets (Movielens and Netflix) demonstrate that the max-norm outperforms the trace-norm for collaborative filtering tasks. We focus our experiments on the problem of exact matrix completion, instead of noisy real-world data to better understand the theoretical properties of both norms.

The two optimization problems we compare are:

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{\text{tr}} \\ \text{subject to} \quad & X_{ij} = M_{ij},\ (i,j) \in \Omega, \\ & X \succeq 0. \end{aligned}$$

and

$$\begin{aligned} \underset{X}{\text{minimize}} \quad & \|X\|_{2,\infty} \\ \text{subject to} \quad & X_{ij} = M_{ij},\ (i,j) \in \Omega, \\ & X \succeq 0. \end{aligned}$$

In our experimental setup, $Y \in \mathbb{R}^{n \times n}$ is generated as $Y = VV^T$, where each entry of $V \in \mathbb{R}^{n \times r}$ is random: $V_{ij} = \pm 1$ with probability half. Over the experiments, the rank of Y is varied by varying the number of columns of V and $\|\Omega\|$ is chosen as multiples of the degrees of freedom of a rank r symmetric matrix. We declare the recovery succesful if $\frac{\|X-Y\|_{Fro}}{\|Y\|_{Fro}} < 1e-3$. The results are summarized in Figure 2.2. We observe that the max-norm is more succesful than the trace-norm at recovering low-rank matrices generated from sign vectors. At $2 \times d.o.f.$ samples, the max-norm frequently succeeds,yet the trace-norm fails. This is expected since the max-norm measures the complexity of a matrix in terms of the size of its decomposition in the space of sign vectors. These experimental results indicate that succesful recovery for the max-norm requires only $O(nr)$ observed entries. In our future work, we will study the max-norm recovery guarantees similar to guarantees for the trace-norm given by CANDES, RECHT ET AL.

3 k-Planes Problem

Given a set of k-planes, the goal is determine a segmentation/clustering of the points according to the plane they belong to. Recent work [8] proposes to solve the following program:

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{Z}\|_{\text{tr}} \\ \text{subject to} \quad & \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}. \end{aligned}$$

The recovered matrix $\boldsymbol{Z}$ is then used to form an affinity matrix $\tilde{\boldsymbol{Z}} = Z + Z^T$ and spectral clustering is applied. In the ideal case, $\boldsymbol{Z}$ will be block-diagonal because any point can be expressed as a linear combination of points within its plane. Thus $\boldsymbol{Z} = diag(\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_k})$. Since the ideal $\boldsymbol{Z}$ is block-diagonal, it has small max-norm since the max-norm measures the complexity of a matrix in terms of a decomposition into sign-vectors. Thus it is natural to instead solve the following program:

$$\begin{aligned} \text{minimize} \quad & \|\boldsymbol{Z}\|_{\max} \\ \text{subject to} \quad & \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z}. \end{aligned}$$

We compare the performance of the trace-norm and max-norm on several synthetic datasets and the Hopkins 155 dataset [9], the standard motion segmentation benchmark dataset for the k-planes problem. The max-norm slightly outperforms the trace-norm in the experiments we attempted. We lea

Table 1: k-planes Segmentation Results

$(d1, \ldots, dk; D)$	Max-Norm	Trace-Norm	(Max,Tie,Trace)
(1 1 3 3; 6)	**4.66%**	8.15%	(67,12,21)
(1 2 3 4; 8)	**1.81%**	2.22%	(39,30,31)
(2 2 3; 5)	**7.80%**	12.67%	(64,11,25)
Kanatanaki 1	**0%**	**0%**	NA
Kanatanaki 2	**0%**	**0%**	NA
Kanatanaki 3	**0%**	**0%**	NA

Table 2: Table of results for k-planes. $(d1, \ldots, dk; D)$ indicates the k different dimensions of each plane and D is the ambient dimension. The error rates for the max-norm and trace-norm are reported in columns 2 and 3. The 4th column indicates the number of times that the max-norm strictly outperformed the trace-norm, the number of ties, and the number of times trace-norm strictly outperformed the max-norm. The first 3 experiments are from synthetically generated data and averaged over 100 trials. The last 3 are from the Hopkins 155 motion segmentation dataset.

4 Future Work

We proposed a relatively unstudied norm that promotes low rank and block-diagonality. Several experiments demonstrate that using this norm as a regularizer leads to block-diagonal similarity matrices, which are easily clustered/classified. This indicates that max-norm regularization is extremely promising in supervised and unsupervised learning, due to the block-diagonal promoting properties. Future work includes formalizing the notion of learning block-diagonal matrices, scaling up the optimization algorithm, and studying other regularizers that promote similarity matrices with desirable properties.

References

[1] Benjamin Recht, Maryam Fazel, and Pablo Parrilo. Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 2007. To appear.

[2] Aditya Bhaskara and Aravindan Vijayaraghavan. Approximating matrix p-norms. In *SODA '11: Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 2011.

[3] G. J. O. Jameson. *Summing and Nuclear Norms in Banach Space Theory*. Number 8 in London Mathematical Society Student Texts. Cambridge University Press, Cambridge, UK, 1987.

[4] Joe Diestel, Hans Jarchow, and Andrew Tonge. *Absolutely Summing Operators*. Number 43 in Cambridge Studies in Advanced Mathematics. Cambrige University Press, April 1995.

[5] Noga Alon and Assaf Naor. Approximating the cut-norm via Grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006.

[6] Nathan Srebro, Jason Rennie, and Tommi Jaakkola. Maximum margin matrix factorization. In *Advances in Neural Information Processing Systems*, 2004.

[7] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *18th Annual Conference on Learning Theory (COLT)*, 2005.

[8] Liu et al. Robust subspace segmentation by low-rank representation. *International Conference on Machine Learning 2010*, pages 1–10.

[9] Roberto Tron and Rene Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. *Proceedings of the CVPR 2007*, 15:1–8, 2007.

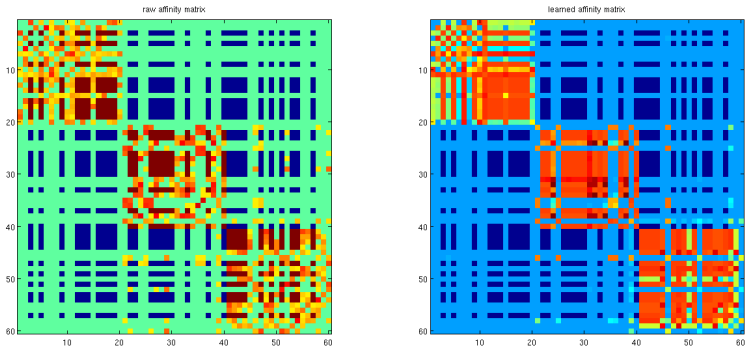


Figure 1: Raw (left) and learned (right) affinity matrix for the Iris dataset

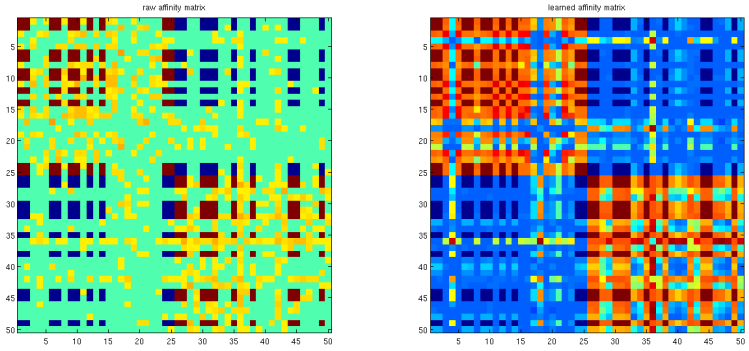


Figure 2: Raw (left) and learned (right) affinity matrix for the digits 3 and 5.

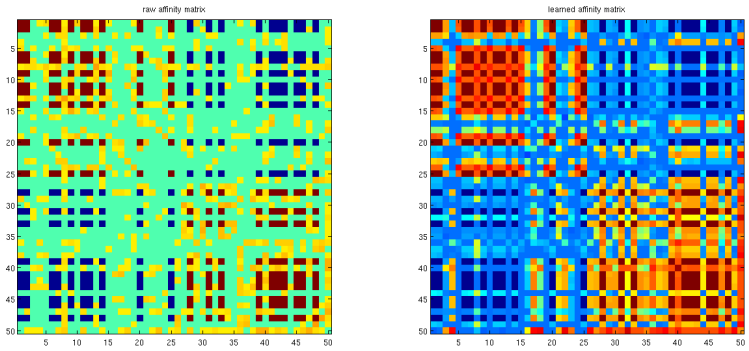
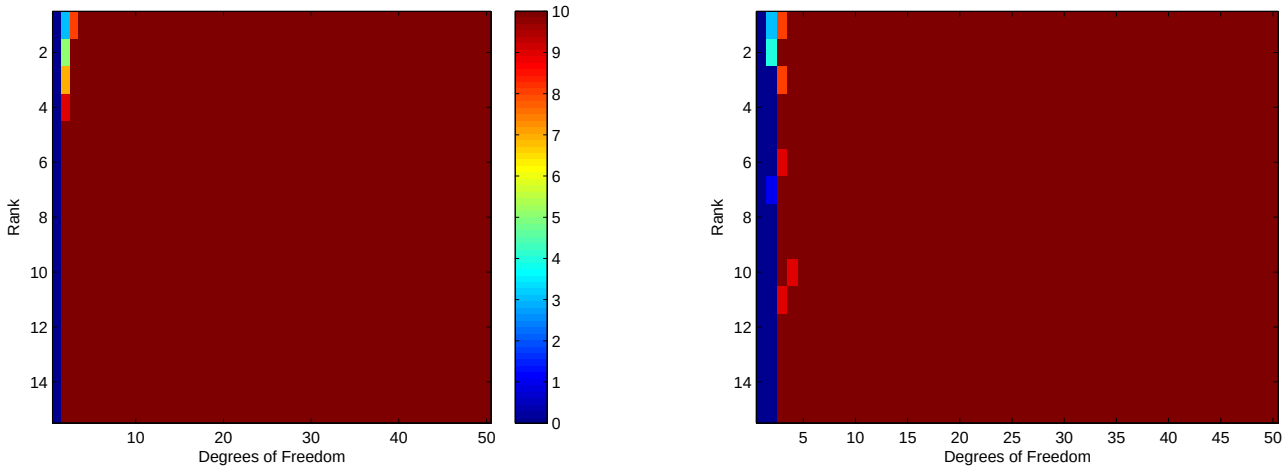


Figure 3: Raw (left) and learned (right) affinity matrix for the digits 4 and 9.



(a) Max-Norm

(b) Trace-Norm

Figure 4: Color represents the probability of succesful recovery over 10 trials. x-axis represents the number of observed entries and y-axis represents the rank. Y is 50×50 symmetric, positive-definite.

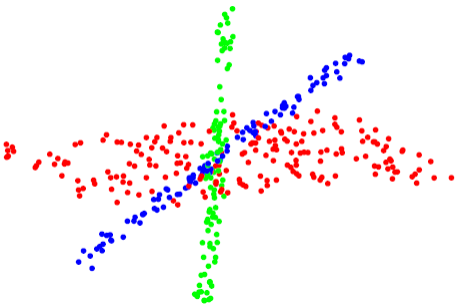


Figure 5: k-plane segmentation problem.

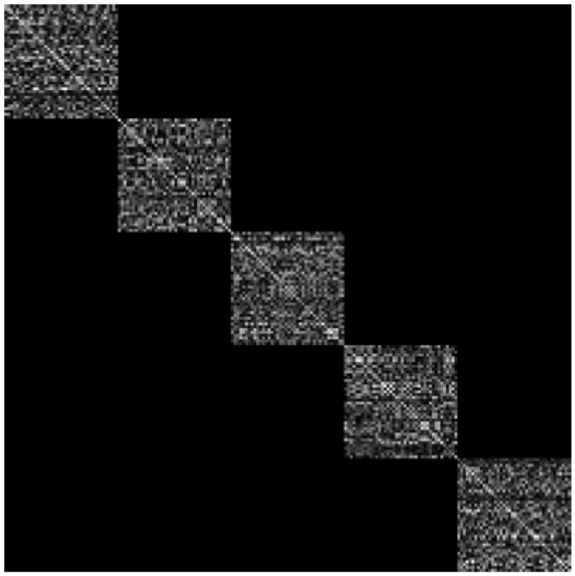


Figure 6: Ideal $\boldsymbol{Z}$ from solving trace/max-norm problem for k-planes.