Tightness of Relaxations for Sparsity and Rank

Nati Srebro
TTI-Chicago

Rank, Trace Norm and Max Norm, COLT 2005
S, Adi Shraibman (Tel Aviv)

Concentration-Based Guarantees for Low-Rank Matrix Reconstruction, COLT 2011
Rinay Foygel (TTIC→UChicago), S

Sparse Prediction with the k-Support Nom, NIPS 2012
Andreas Argyriou (TTIC→Ecole Centrale Paris), Rina Foygel, S

Clustering, Hamming Embedding, Generalized LSH and the Max Norm, ALT 2014
Behnam Neyshabur (TTIC), Yury Makarychev (TTIC), S
Relaxing Non-Convex Constraints

\[ \min_{\mathbf{w} \in \mathcal{C}} L(\mathbf{w}) \]

If \( \mathcal{C} \) is non-convex (e.g. “sparse”, “low rank”, “obeys certain dependency structure”, ”conjunction of three variables”).
\[ \text{e.g. learn non-convex hypothesis class } \mathcal{C} \]

Often relax to \( \mathcal{C} \subset \mathcal{W} \) convex:
\[ \min_{\mathbf{w} \in \mathcal{W}} L(\mathbf{w}) \]
\[ \text{i.e. learn predictor from larger hypothesis class } \mathcal{W} \]

- How tight are the standard relaxations we use?
  - How much larger is the estimation error?
  - By how much does the sample complexity increase
- Are there better relaxations?
Outline

• Part I: Relaxing Sparsity
  – $k$-support norm
  *Sparse Prediction with the $k$-Support Norm*, NIPS 2012, Argyriou, Foygel, S

• Part II: Relaxing Rank
  – Matrix max-norm
  *Concentration-Based Guarantees for Low-Rank Matrix Recon.*, COLT 2011, Foygel, S
  *Rank, Trace Norm and Max Norm*, COLT 2005, S, Shraibman

• Bonus: Relaxing Clustering
  *Clustering, Hamming Embedding, Generalized LSH and the Max Norm*, ALT 2014, Neyshabur, Makarychev, S
Sparsity

- Sparsity of \( w \in \mathbb{R}^d \):
  \[
  |w|_0 = \text{number of non-zero entries in } w
  \]

- "find best \textbf{sparse} \( w \in \mathbb{R}^d \)" ubiquitous
  - Feature selection / Sparse prediction
  - Compressed sensing
  - Sparse inverse covariance estimation
  - Minimizing number of errors, if \( w \) is error vector

- Common relaxation: \[
  |w|_1 = \sum_i |w_i|
  \]
Relaxing Sparsity Constraints

\[
\min_{|w|_0 \leq k} L(w)
\]

\[
\{ |w|_0 \leq k \}
\]
Relaxing Sparsity Constraints

\[
\min_{|w|_0 \leq k} L(w)
\]

\[
\{ |w|_0 \leq k, |w|_\infty \leq 1 \}
\]
Relaxing Sparsity Constraints

\[
\min_{|w|_0 \leq k} L(w)
\]

\[
\{ |w|_0 \leq k, |w|_\infty \leq 1 \} 
\subseteq \{ |w|_1 \leq k \}
\]
Relaxing Sparsity Constraints

\[
\min_{|w|_0 \leq k} L(w)
\]

\[
\{ |w|_0 \leq k, |w|_\infty \leq 1 \} \subseteq \{ |w|_1 \leq k \}
\]
Relaxing Sparsity Constraints

\[
\min_{|w|_0 \leq k} L(w)
\]

\[
\text{conv}(\{ |w|_0 \leq k, |w|_\infty \leq 1 \}) = \{ |w|_1 \leq k, |w|_\infty \leq 1 \}
\]
Sample Complexity

Want to minimize:

\[ L(w) = \mathbb{E}_{x,y} [\ell(w, x, y)] \]

Based in \( m \) iid samples \((x_i, y_i)\):

\[ \hat{w} = \arg \min_{w \in \mathcal{W}} \sum_{i=1}^{m} \ell(w, x_i, y_i) \]

\# samples \( m \) so that \( L(\hat{w}) \leq \inf_{w \in \mathcal{W}} L(w) + \epsilon \):

- For \( \mathcal{C} = \{w \in \mathbb{R}^d, |w|_0 \leq k\} \):
  \[ m = O\left( k \log(d)/\epsilon^2 \right) \]
- For \( \mathcal{W} = \{w \in \mathbb{R}^d, |w|_1 \leq k, |w|_\infty \leq 1\} \):
  \[ m = O\left( |w|_1^2 \log(d)/\epsilon^2 \right) = O(K^2 \log(d)/\epsilon^2) \]

Can be reduced to \( \frac{1}{\epsilon} \cdot \frac{L^* + \epsilon}{\epsilon} \).
Measuring Scale by $|w|_2$

- Replace $|w|_\infty \leq 1$ by $|w|_2 \leq 1$ (or $\leq B$)
  - Robustness
  - Scale of $E[\langle w, x \rangle^2]$
  - Generalization
The Elastic Net

\[ \{ |w|_0 \leq k, |w|_2 \leq 1 \} \subset \{ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1 \} \]

\[ |w|_{\text{en}}^k = \max \left( |w|_2, \frac{|w|_1}{\sqrt{k}} \right) \]

- Sample Complexity (# samples m so that \( L(\hat{w}) \leq \inf_{w \in \mathcal{W}} L(w) + \epsilon \)):
  \[ O \left( |w|_1^2 \log(d) / \epsilon^2 \right) = O \left( k \log(d) / \epsilon^2 \right) \]
The Elastic Net

\[ \text{conv}( \{ |w|_0 \leq k, |w|_2 \leq 1 \}) \]
\[ \subset \{ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1 \} \]

\[ |w|_k^{\text{en}} = \max \left( |w|_2, \frac{|w|_1}{\sqrt{k}} \right) \]

- Sample Complexity (# samples \( m \) so that \( L(\hat{w}) \leq \inf_{w \in \mathcal{W}} L(w) + \epsilon \)):
  \[ O\left( |w|_1^2 \log(d) / \epsilon^2 \right) = O\left( k \log(d) / \epsilon^2 \right) \]
The Elastic Net

\[ \text{conv}( \{ |w|_0 \leq k, |w|_2 \leq 1 \} ) \subset \{ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1 \} = \{ |w|_k^{\text{en}} \leq 1 \} \]

\[ |w|_k^{\text{en}} = \max\left( |w|_2, \frac{|w|_1}{\sqrt{k}} \right) \]
The \( k \)-Support Norm

\[
\text{conv}\left( \{ |w|_0 \leq k, |w|_2 \leq 1 \} \right) = \{ |w|_k^{sp} \leq 1 \} \\
\subset \{ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1 \} = \{ |w|_k^{en} \leq 1 \}
\]
How to Make Paper Lanterns

Looking for instructions on how to make paper lanterns? My husband designed an easy template for making paper lanterns in a cute round shape. They look a bit oriental, don't you think?
**The $k$-Support Norm**

\[
\{ |w|_{k}^{sp} \leq 1 \} = \text{conv}( \{ |w|_0 \leq k, |w|_2 \leq 1 \} )
\]

- Can be viewed as Overlap Group Lasso where the “groups” are all $k$-subsets:
  \[
  |w|_{k}^{sp} = \inf_{v_I} \{ \sum_{I \subset [d], |I| = k} |v_I|_2 \mid \text{supp}(v_I) = I, \sum v_I = w \}
  \]

\[
|w|_1^{sp} = |w|_1 \\
|w|_d^{sp} = |w|_2
\]

- Dual norm: $2$-$k$ symmetric gauge norm

\[
|u|_{k}^{sp*} = \sqrt{\sum_{i=1}^{k} (|u|_i^a)^2} = |\text{top } k \text{ elements in } u|_2
\]

\[
|w|_1^{sp*} = |w|_{\infty} \\
|w|_d^{sp*} = |w|_2
\]
**Computation and Optimization**

\[
|w|_k^{sp} = \sqrt{\sum_{i=1}^{k-r-1} (|w|_i^\downarrow)^2 + \frac{1}{r+1} \left( \sum_{i=k-r}^{d} |w|_i^\downarrow \right)^2}
\]

where:

\[
|w|_{k-r-1} > \frac{1}{r+1} \sum_{i=k-r}^{d} |w|_i^\downarrow \geq |w|_{k-r}^{sp}
\]

- Can compute \( |w|_k^{sp} \) in time \( O(d \log(d)) \)
- Can compute \( \nabla |w|_k^{sp} \) in time \( O(d \log(d)) \)
- Can compute prox map in time \( O(d (\log(d)+k)) \):

\[
\text{prox}_\lambda(w) = \arg \min_u \frac{1}{2} |u - w|^2_2 + \lambda (|u|_k^{sp})^2
\]

\( \Rightarrow \) can optimize \( \min_{|w|_k^{sp} \leq B} L(w) \) or \( \min L(w) + \lambda |w|_k^{sp} \) using e.g. FISTA
**k-Support vs Elastic Net**

\[
\{ |w|_k^{\text{sp}} \leq 1 \} = \text{conv}( \{ |w|_0 \leq k, |w|_2 \leq 1 \} ) \\
\subset \{ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1 \} = \{ |w|_k^{\text{en}} \leq 1 \}
\]

- \(|w|_k^{\text{en}} \leq |w|_k^{\text{sp}}|
- \(|w|_1^{\text{en}} = |w|_1^{\text{sp}} = |w|_1\) and \(|w|_d^{\text{en}} = |w|_d^{\text{sp}} = |w|_2\)
- For \(w = (k^{1.5}, 1, 1, ..., 1) \in \mathbb{R}^d, d = k^2 + 1\):

\[
k^{1.5} \left(1 + \frac{1}{\sqrt{k}}\right) = |w|_k^{\text{en}} < |w|_k^{\text{sp}} = \sqrt{2} \cdot k^{1.5}
\]

\(\Rightarrow\) Gap could be as much as \(\sqrt{2}\)

**Theorem:** \(|w|_k^{\text{en}} \leq |w|_k^{\text{sp}} \leq \sqrt{2} \cdot |w|_k^{\text{en}}\)
## Experiments

<table>
<thead>
<tr>
<th></th>
<th>Zou+Hastie Synthetic (d=40, k=15, strong correlations)</th>
<th>South African Heart</th>
<th>20 Newsgroups</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lasso</strong></td>
<td>0.27</td>
<td>0.18</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Elastic Net</strong></td>
<td>0.23</td>
<td>0.18</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>k-Support</strong></td>
<td><strong>0.21</strong></td>
<td>0.18</td>
<td><strong>0.69</strong></td>
</tr>
</tbody>
</table>

Mean Squared Error on test data. Parameters $\lambda$, $k$ selected on validation set.
Summary: $k$-Support Norm

• When discussing “tightness” of convex relaxation, scale constraint is important!

• $k$-support norm is tightest convex relaxation of sparsity with an $l_2$ constraint

• efficiently to computable and optimizable

• strictly tighter then elastic net (relaxing $|w|_0$ to $|w|_1$)

• ... but only up to a factor of $\sqrt{2}$

⇒ elastic net is tight up to $\sqrt{2}$
Part II: Matrices

• Relax \{ \text{rank}(X) \leq k \}

• With what scale constraint?

• Trace-norm (aka nuclear norm, \|\text{spectrum}\|_1) is tightest relaxation subject to spectral norm (\|\text{spectrum}\|_{\infty}):
  \[
  \{ |X|_{tr} \leq k, |X|_{sp} \leq 1 \} = \text{conv}(\{ \text{rank}(X) \leq k, |X|_{sp} \leq 1 \})
  \]
Constraining Avg Entry Magnitude

- Relax \( \{ \text{rank}(X) \leq k, \frac{1}{nm} |X|^2_F \leq 1 \} \)

- \( |X|^2_F = |\text{spectrum}|_2 \), vector case carries over:
  - \( \left\{ \frac{1}{nm} |X|^2_{\text{tr}} \leq k, \frac{1}{nm} |X|^2_F \leq nm \right\} \) tight up to a factor of \( \sqrt{2} \)
  - Convex hull (tight relaxation) give by \( k \)-support norm applied to spectrum
    - Can calculate and optimize, just like vector case

- But often \( |X|_\infty \) more natural
  - Required for (noisy) matrix completion guarantees
The Matrix Max-Norm

• Recall: $|X|_{tr} = \min_{X=UV} |U|_F |V|_F$

• The Max-Norm: $|X|_{\max} = \min_{X=UV} |U|_{2,\infty} |V|_{2,\infty}$
  
  – Not a spectral function!
  – SDP representable
  – Super-fast non-convex opt [Lee et al 2010]
  – Fast 1st order optimization [PRISMA: Argyriou, Orabona, S 2012]

• $\frac{1}{nm} |X|_2^2 \leq |X|_{\max}^2 \leq \text{rank}(X) \cdot |X|_\infty^2$
  
  – Contrast with: $\frac{1}{nm} |X|_2^2 \leq \text{rank}(X) \cdot \frac{1}{nm} |X|_F^2$
Trace-Norm vs Max-Norm

\{ \text{rank}(X) \leq k, |X|_\infty \leq 1 \} 
\subset \{ |X|_{\text{max}}^2 \leq k, |X|_\infty \leq 1 \} 
\subset \left\{ \frac{1}{nm} |X|_{tr}^2 \leq k, |X|_\infty \leq 1 \right\}

- Gap between relaxations as large as \((n/k)^{2/3}\):

\[
X = \begin{bmatrix}
H_{\sqrt{n^2k}} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0
\end{bmatrix}
\]

\[
\frac{1}{nm} |X|_{tr}^2 = k \\
|X|_{\text{max}}^2 = \frac{3}{\sqrt{n^2k}}
\]
Sample Complexity for Low-Rank Matrix Reconstruction

- $Y \approx \text{low rank } M$, observe random subset $S$ of entries

- \#sample to get $\frac{1}{nm} \| \hat{Y} - Y \|_1 \leq \frac{1}{nm} \| M - Y \|_1 + \varepsilon$
  
  (or, if $Y=M+iid$ noise, to get $\frac{1}{nm} \| \hat{Y} - M \|_F^2 \leq \varepsilon$)

  - Using trace-norm: $O( \text{rank}(M) \log(n) (n+m) / \varepsilon^2 )$
  - Using max-norm: $O( \text{rank}(M) (n+m) / \varepsilon^2 )$

- If entries sampled non-uniformly:
  
  - Using trace-norm: $O( \text{rank}(M) \sqrt{n} (n+m) / \varepsilon^2)$
  - Using max-norm: $O( \text{rank}(M) (n+m) / \varepsilon^2 )$
The Trace-Norm with Non-Uniform Sampling

- Both A, B of rank 2
- Sampling:
  - uniform in A w.p. $\frac{1}{2}$
  - uniform w.p. $\frac{1}{2}$

- Regularizing with the rank or with the max-norm:
  sample complexity $\propto n$, i.e. $O(1)$ per row

- Regularizing with the trace-norm:
  number $\propto n^{4/3}$, i.e. $O(n^{1/3})$ per row!!!

[Salakhutdinov § 10]

improved to $O(n^{3/2})$ by [Hazan Kale Shalev-Shwartz 12]
# Experiments on Netflix

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>%Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>NetFlix Cinematch: (baseline)</td>
<td>0.9525</td>
<td>0</td>
</tr>
<tr>
<td>TraceNorm:</td>
<td>0.9235</td>
<td>3.04</td>
</tr>
<tr>
<td>MaxNorm:</td>
<td>0.9138</td>
<td>4.06</td>
</tr>
<tr>
<td>Weighted TraceNorm:</td>
<td>0.9078</td>
<td>4.69</td>
</tr>
<tr>
<td>Smoothed Wghtd TrNorm:</td>
<td>0.9068</td>
<td>4.80</td>
</tr>
<tr>
<td>Local MaxNorm</td>
<td>0.9063</td>
<td>4.85</td>
</tr>
<tr>
<td>Winning team:</td>
<td>0.8553</td>
<td>10.20</td>
</tr>
</tbody>
</table>
Tightness of Max-Norm Relaxation

• Grothendik’s inequality:
  \[
  \text{conv}\{ \text{rank}(X) \leq 1, |X|_\infty \leq 1 \} \\
  \subset \{ |X|_{\text{max}}^2 \leq 1 \} \\
  \subset 1.79 \cdot \text{conv}\{ \text{rank}(X) \leq 1, |X|_\infty \leq 1 \}
  \]

• What about larger k?
  \[
  \text{conv}\{ \text{rank}(X) \leq k, |X|_\infty \leq 1 \} \\
  \subset \{ |X|_{\text{max}}^2 \leq k, |X|_\infty \leq 1 \} \\
  \subset G(k) \cdot \text{conv}\{ \text{rank}(X) \leq k, |X|_\infty \leq 1 \}
  \]

• How does G(k) grow?
  \[
  1.4 \leq G(k) \leq \sqrt{k} \cdot 1.79
  \]
Cluster Matrices

• 2-way clustering matrices (cut matrices):

\[
\begin{bmatrix}
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
-1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & +1 & +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
+1 & -1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

\[
\text{conv}\left(\{X = uu' \mid u \in \pm 1^n\}\right) = \text{cut-polytope} = \text{conv}\left(\{\text{rank}(X) \leq 1, |X|_\infty \leq 1, X \succeq 0\}\right)
\subset \{|X|_{\max} \leq 1, X \succeq 0\}
\subset \{|X|_{\tr} \leq n, X \succeq 0\}
\]
Multi-Way Cut Matrices

\[
X_{ij} = \delta_{u_i, u_j} \bigg| u \in [k]^n \bigg\}
\subset \left\{ |X|_{\text{max}} \leq 3 - \frac{4}{k} \right\}
\subset \text{conv}(\{ \text{rank}(X) \leq k, |X|_{\infty} \leq 1 \})
\subset \left\{ |X|_{\text{max}} \leq \sqrt{k}, |X|_{\infty} \leq 1 \right\}
\]

- \text{conv} \left( \left\{ X_{ij} = \delta_{u_i, u_j} \bigg| u \in \mathbb{N}^n \right\} \right) \subset \left\{ |X|_{\text{max}} \leq 3 \right\}
Tightness of Relaxation: Bi-Clustering

\[ X_{ij} = \delta_{u_i,v_j} \mid u \in [k]^n, v \in [k]^m \]
\[ \subset \{ |X|_{\text{max}} \leq 3 \} \]
\[ \subset 5.37 \text{conv} \left( \{ X_{ij} = \delta_{u_i,v_j} \mid u \in [k]^n, v \in [k]^m \} \right) \]

- For symmetric clustering: no tractable constant ratio (or dependent only on \( k \)) convex relaxation
Summary

- When discussing “tightness” of convex relaxation, scale constraint is important!

- Relaxing sparsity with bounded $l_2$ scale:
  - $k$-support norm is tightest convex relaxation
  - Elastic net is tight up to $\sqrt{2}$

- Relaxing rank constraint for bounded entry matrices:
  - Max-norm much tighter than trace-norm
  - Better reconstruction guarantees; often better empirical performance

Rank, Trace Norm and Max Norm, COLT 2005
Concentration-Based Guarantees for Low-Rank Matrix Reconstruction, COLT 2011
Sparse Prediction with the k-Support Nom, NIPS 2012
Clustering, Hamming Embedding, Generalized LSH and the Max Norm, ALT 2014