Tightness of Relaxations for Sparsity and Rank

Nati Srebro

TTI-Chicago

Sparse Prediction with the k-Support Nom, NIPS 2012

Andreas Argyriou (TTIC→'Ecole Centrale Paris), Rina Foygel (TTIC→Stanford), S

Concentration-Based Guarantees for Low-Rank Matrix Reconstruction, COLT 2011

Rinay Foygel, S

Rank, Trace Norm and Max Norm, COLT 2005

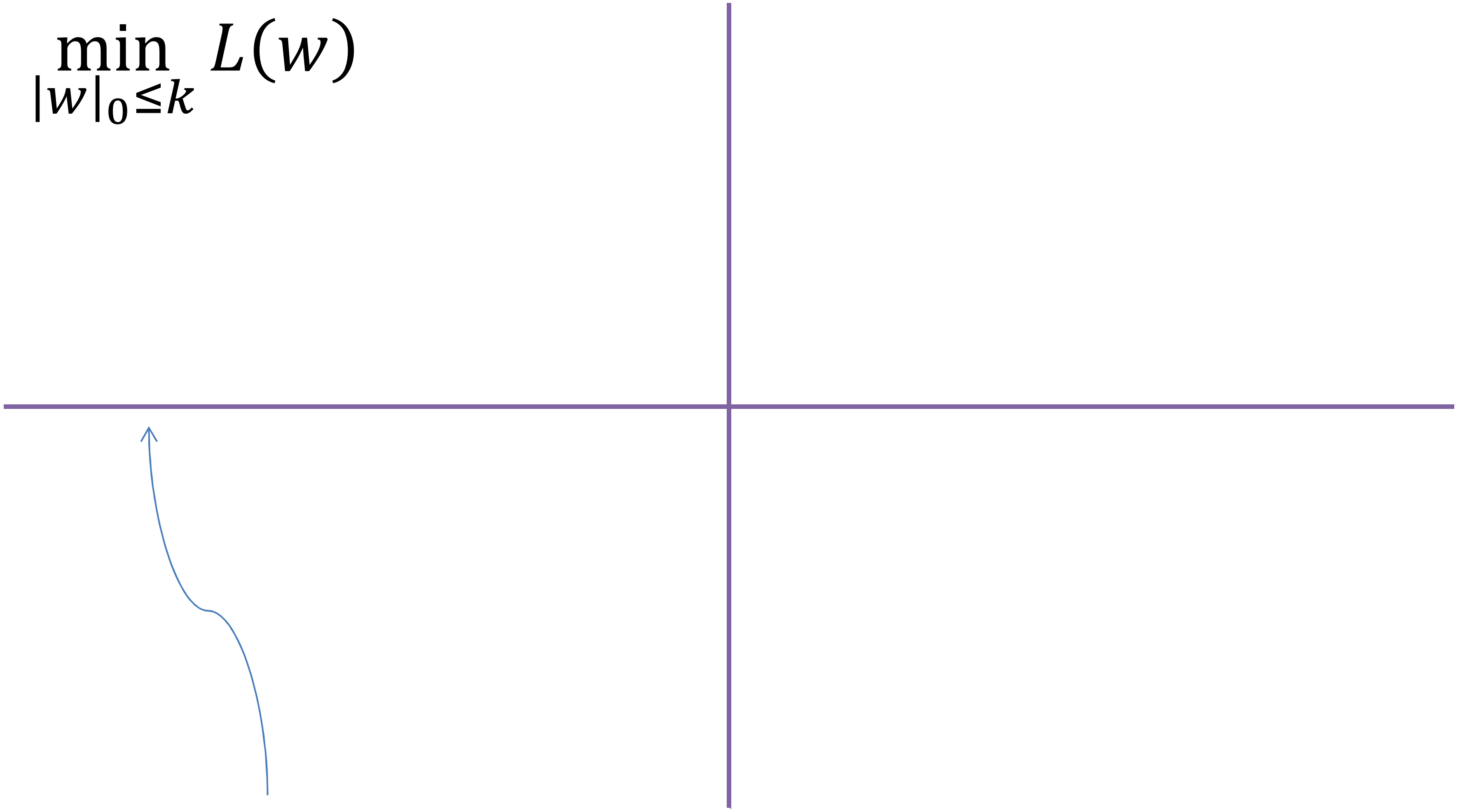
S, Adi Shraibman (Tel Aviv)

Outline

- Part I: Relaxing Sparsity
  - *k*-support norm
- Part II: Relaxing Rank
  - Matrix max-norm

Relaxing Sparsity Constraints

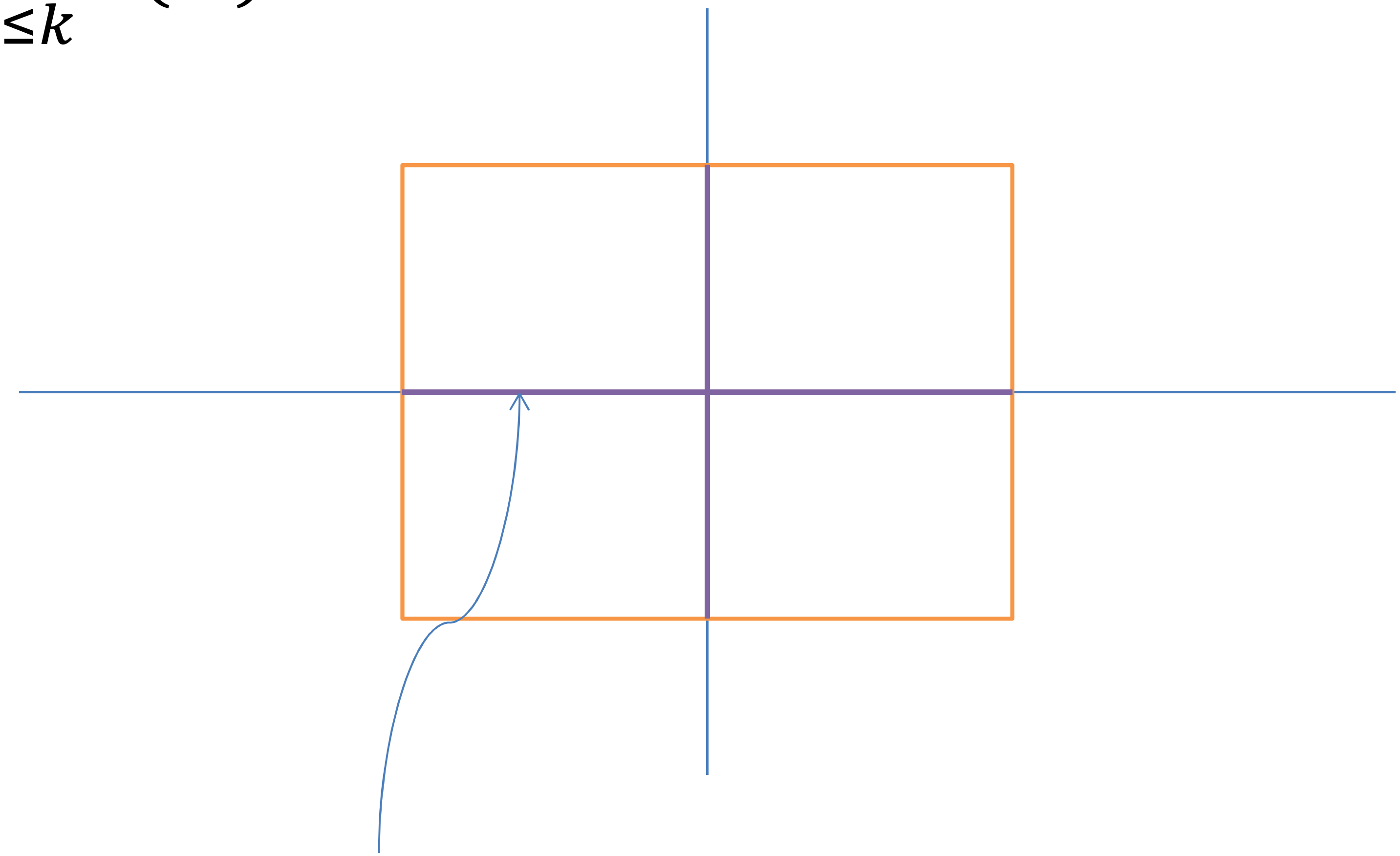
$$\min_{|w|_0 \leq k} L(w)$$



$\{ \, |w|_0 \leq k \qquad\qquad \}$

Relaxing Sparsity Constraints

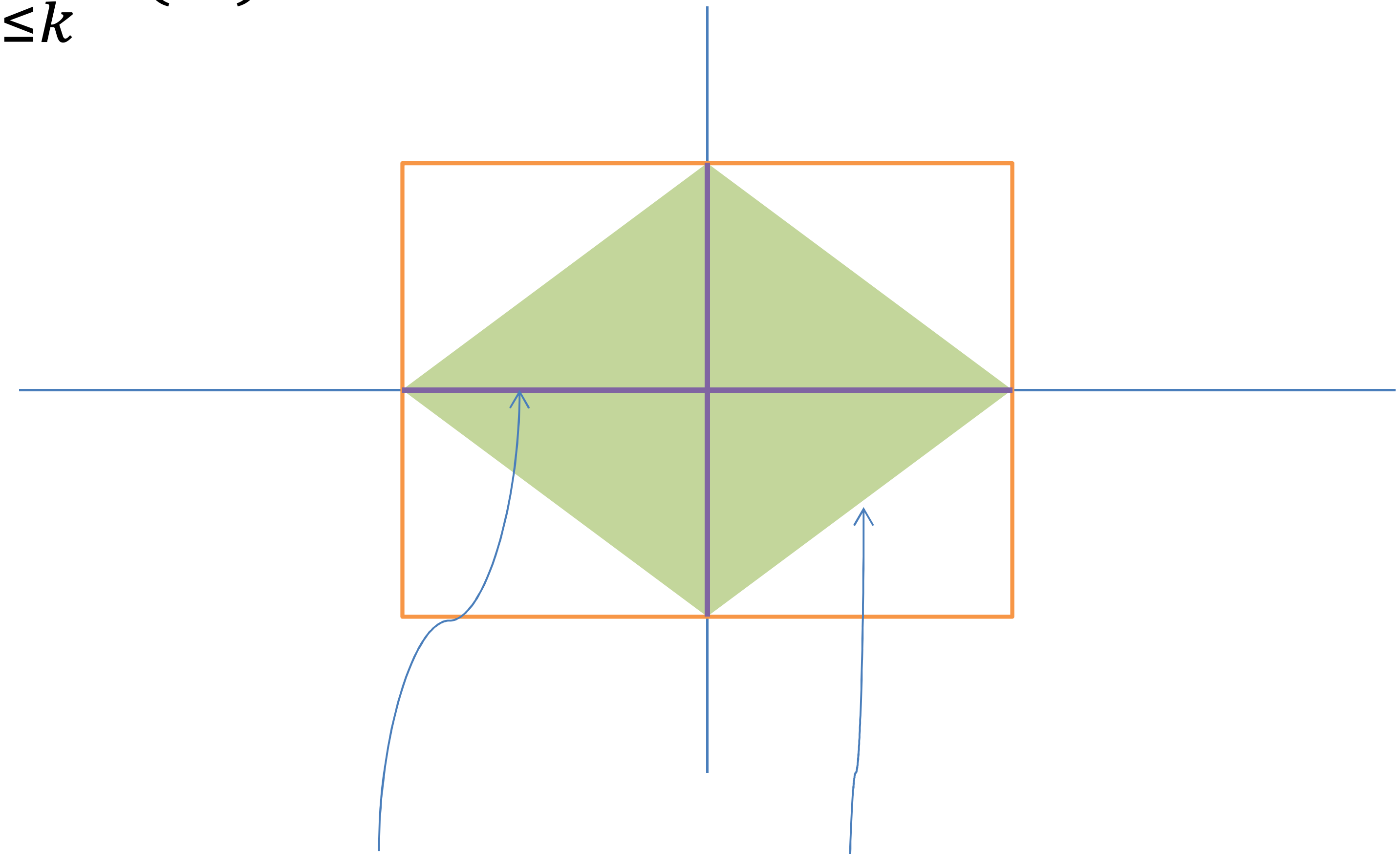
$$\min_{|w|_0 \le k} L(w)$$



$$\{ \, |w|_0 \le k \, , \, |w|_\infty \le 1 \, \}$$

Relaxing Sparsity Constraints

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

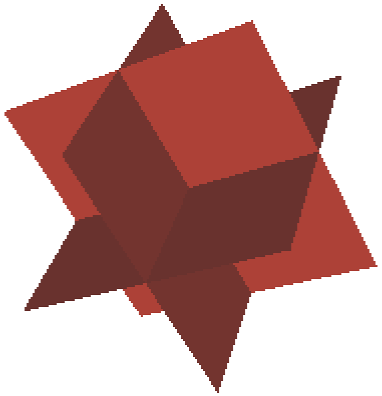


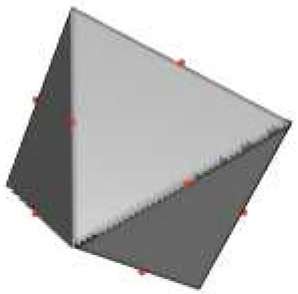
$\{ \ |w|_0 \leq k \ , \ |w|_\infty \leq 1 \ \}$

$\subseteq \{ \ |w|_1 \leq k \qquad\qquad\qquad \}$

Relaxing Sparsity Constraints

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



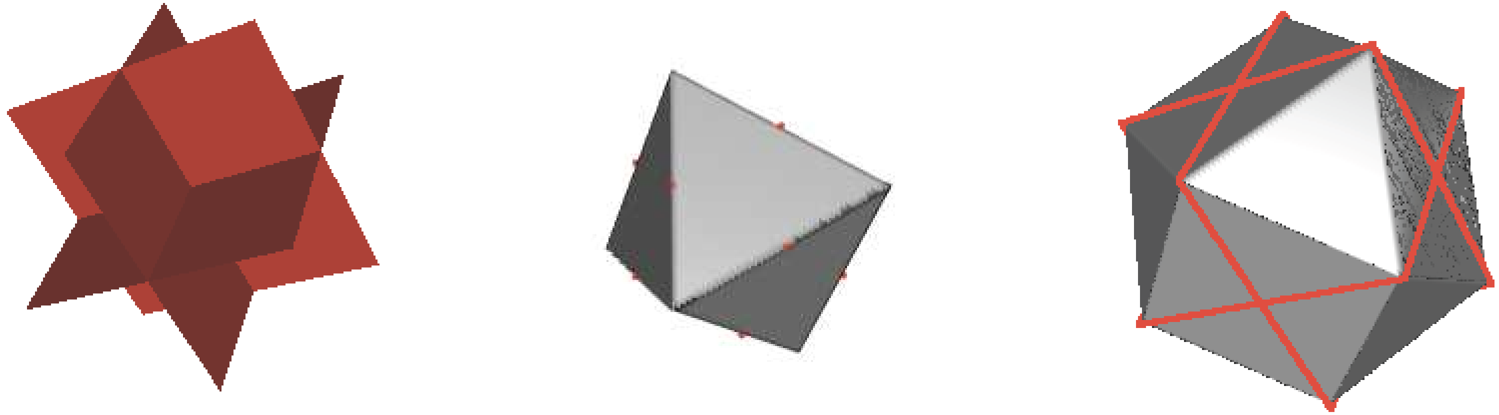


$$\{\ |w|_0 \leq k\ ,\ |w|_\infty \leq 1\ \}$$

$$\subseteq \{\ |w|_1 \leq k \qquad\qquad\qquad\qquad \}$$

Relaxing Sparsity Constraints

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



$\text{conv}\left(\{\, |w|_0 \leq k\,,\ |w|_\infty \leq 1 \,\}\right)$

$= \{\, |w|_1 \leq k\,,\ |w|_\infty \leq 1 \,\}$

Sample Complexity

Want to minimize:

$$L(w) = E_{x,y}[l(w,x,y)]$$

Based in m iid samples (x_i,y_i):

$$\hat{w} = \arg\min_{w\in\mathcal{W}} \sum_{i=1..m} l(w,x_i,y_i)$$

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

- For $\mathcal{W} = \{ w\in R^d,\ |w|_0 \leq k \}$:

  $$m = O(\ k \log(d) / \epsilon^2\ )$$

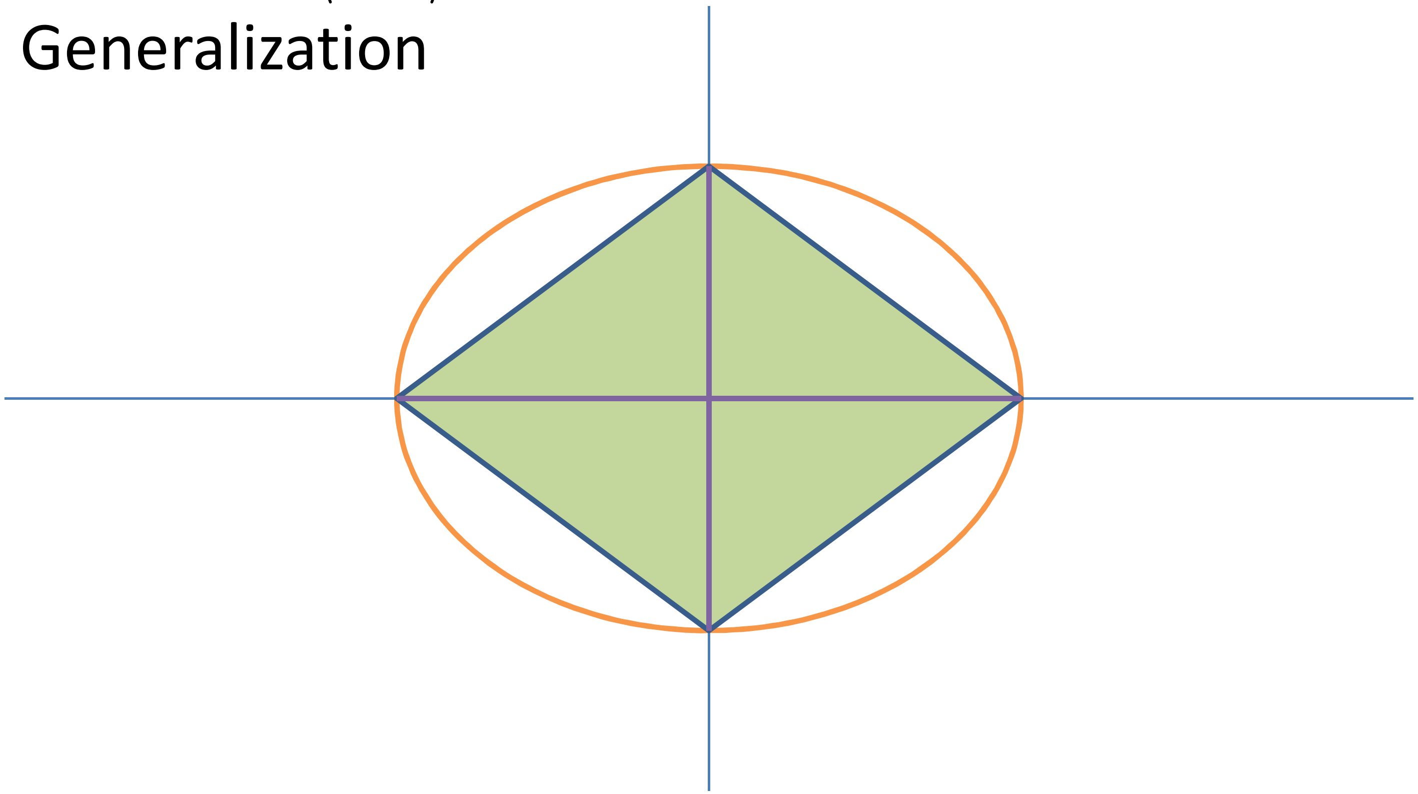
- For $\mathcal{W} = \{ w\in R^d,\ |w|_1 \leq k,\ |w|_\infty \leq k \}$:

  $$m = O(\ |w|_1^2 \log(d) / \epsilon^2\ ) = O(\ k^2 \log(d) / \epsilon^2\ )$$

Can be reduced to $1/\epsilon\cdot((L^*+\epsilon)/\epsilon)$

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

$$\{\ |w|_0 \leq k, |w|_2 \leq 1\}$$

$$\subset \left\{\ |w|_1 \leq \sqrt{k}, |w|_2 \leq 1\right\} = \{\ |w|_k^{\mathbf{en}} \leq 1\}$$

$$|w|_k^{\mathbf{en}} = \max\left(|\boldsymbol{w}|_{\mathbf{2}}, \frac{|\boldsymbol{w}|_{\mathbf{1}}}{\sqrt{\boldsymbol{k}}}\right)$$

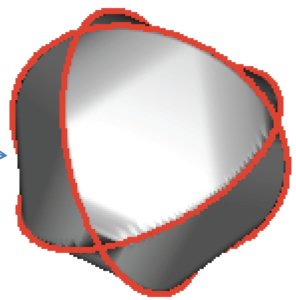
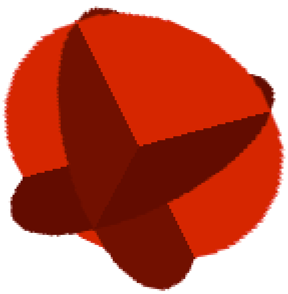
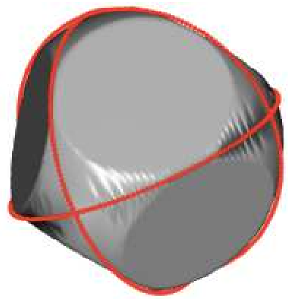
- Sample Complexity (# samples m so that L($\hat{w}$)$\leq$inf$_{w\in\mathcal{W}}$ L(w)+ϵ):

  O($|w|_1^2$log(d) / ϵ^2) = O(k log(d) / ϵ^2)

The Elastic Net

$$\mathrm{conv}(\,\{\,|w|_0 \le k, |w|_2 \le 1\}\,)$$
$$\subset \{\,|w|_1 \le \sqrt{k}, |w|_2 \le 1\} = \{\,|w|_k^{\mathbf{en}} \le 1\}$$

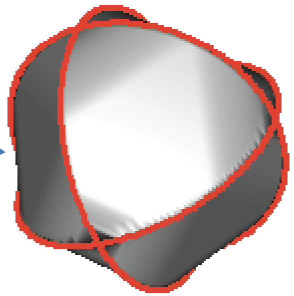
$$|w|_k^{\mathbf{en}} = \max\left(|\boldsymbol{w}|_{\mathbf{2}}, \frac{|\boldsymbol{w}|_{\mathbf{1}}}{\sqrt{\boldsymbol{k}}}\right)$$

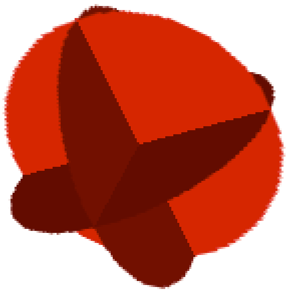




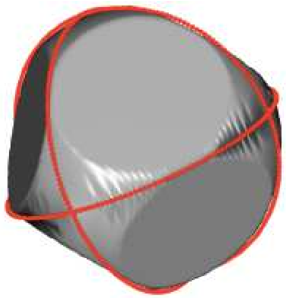
The *k*-Support Norm

$$\mathrm{conv}(\ \{\ |w|_0 \le k, |w|_2 \le 1\} \quad = \{\ |w|_k^{\mathbf{sp}} \le 1\ \}$$

$$\subset \{\ |w|_1 \le \sqrt{k}, |w|_2 \le 1\} = \{\ |w|_k^{\mathbf{en}} \le 1\}$$







The *k*-Support Norm

$$\{\ |w|_k^{\mathbf{sp}} \leq 1\ \} = \mathrm{conv}(\ \{\ |w|_0 \leq k, |w|_2 \leq 1\}\ )$$

- Can be viewed as Overlap Group Lasso where the "groups" are all *k*-subsets:

$$|w|_k^{\mathbf{sp}} = \inf_{v_I} \left\{ \sum_{I \subset [d], |I|=k} |v_I|_2 \ \middle|\ \mathrm{supp}(v_I) = I, \sum v_I = w \right\}$$

$$|w|_1^{\mathbf{sp}} = |w|_1 \qquad\qquad |w|_d^{\mathbf{sp}} = |w|_2$$

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

$$|u|_k^{\mathbf{sp}*} = \sqrt{\textstyle\sum_{i=1}^{k} (|u|_i^{a})^2} = |\text{top } k \text{ elements in u}|_2$$

$$|w|_1^{\mathbf{sp}*} = |w|_\infty \qquad\qquad |w|_d^{\mathbf{sp}*} = |w|_2$$

Computation and Optimization

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

where:

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

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

$$\mathrm{prox}_\lambda(w) = \arg\min_u \frac{1}{2}|u-w|_2^2 + \lambda\left(|u|_k^{\mathbf{sp}}\right)^2$$

$\Rightarrow$ can optimize $\min_{|w|_k^{\mathbf{sp}} \leq B} L(w)$ or $\min L(w) + \lambda\, |w|_k^{\mathbf{sp}}$ using e.g. FISTA

k-Support vs Elastic Net

$$\{\, |w|_k^{\mathbf{sp}} \le 1 \,\} = \mathrm{conv}(\, \{\, |w|_0 \le k, |w|_2 \le 1\}\,)$$
$$\subset \left\{\, |w|_1 \le \sqrt{k}, |w|_2 \le 1\right\} = \{\, |w|_k^{\mathbf{en}} \le 1\}$$

- $|w|_k^{\mathbf{el}} \le |w|_k^{\mathbf{sp}}$
- $|w|_1^{\mathbf{sp}}$=$|w|_1^{\mathbf{sp}}$=$|w|_1$ and $|w|_d^{\mathbf{sp}}$=$|w|_d^{\mathbf{sp}}$=$|w|_2$
- For $w = (k^{1.5}, 1,1, \ldots, 1) \in R^d$, $d = k^2 + 1$:

$$k^{1.5}\left(1 + \frac{1}{\sqrt{k}}\right) = |w|_k^{\mathbf{el}} < |w|_k^{\mathbf{sp}} = \sqrt{2} \cdot k^{1.5}$$

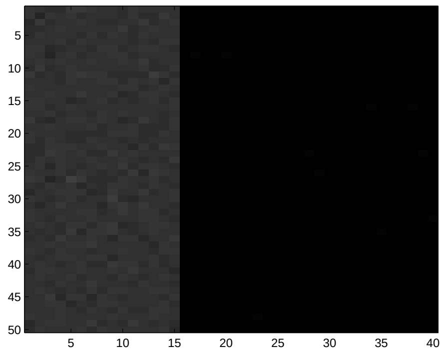
$\Rightarrow$ Gap could be as much as $\sqrt{2}$

Theorem: $\quad |w|_k^{\mathbf{el}} \le |w|_k^{\mathbf{sp}} \le \sqrt{2} \cdot |w|_k^{\mathbf{el}}$

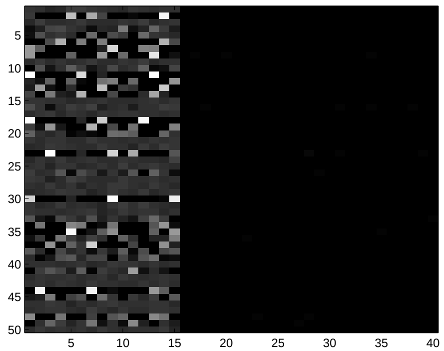
Experiments

	Zou+Hastie Synthetic (d=40,k=15, strong correlations)	South African Heart	20 Newsgroups
Lasso	0.27	0.18	0.70
Elastic Net	0.23	0.18	0.70
k-Support	**0.21**	0.18	**0.69**

Mean Squared Error on test data.
Parameters λ, k selected on validation set.



k-Support



Elastic Net

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 $\sqrt{2}$

$\Rightarrow$ elastic net is tight up to $\sqrt{2}$

Part II: Rank

- Relax { rank(X) $\leq$ k }
- With what scale constraint?

- Trace-norm (aka nuclear norm, $|\text{spectrum}|_1$) is tightest relaxation subject to spectral norm ($|\text{spectrum}|_\infty$):

$$\left\{ |X|_{\text{tr}} \leq k, |X|_{\text{sp}} \leq 1 \right\}$$
$$= \text{conv}\left(\left\{\text{rank}(X) \leq k, |X|_{\text{sp}} \leq 1\right\}\right)$$

Constraining Avg Entry Magnitude

- Relax $\{\mathrm{rank}(X) \le k, \frac{1}{nm}|X|_F^2 \le 1\}$
- $|X|_F^2 = |\mathrm{spectrum}|_2$, vector case carries over:
  - $\left\{\frac{1}{nm}|X|_{\mathrm{tr}}^2 \le k, \frac{1}{nm}|X|_F^2 \le nm\right\}$ 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: $|X|_{\text{tr}} = \min_{X=UV'} |U|_F |V|_F$
- 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]

$|U|_{2,\infty} = \max_i |U_i|_2$

- $\frac{1}{nm} |X|_{\text{tr}}^2 \leq |X|_{\max}^2 \leq \text{rank}(X) \cdot |X|_\infty^2$
  - Contrast with: $\frac{1}{nm} |X|_{\text{tr}}^2 \leq \text{rank}(X) \cdot \frac{1}{nm} |X|_F^2$

Trace-Norm vs Max-Norm

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

- Gap between relaxations as large as $\sqrt[3]{n}$:

$$X = \begin{bmatrix} H_{\sqrt[3]{n^2k}} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$|X|_{\max} = \sqrt[3]{n/k} \qquad\qquad \frac{1}{\sqrt{nn}}|X|_{\mathrm{tr}} = 1$$

(Gap between max-norm and trace-norm as large as n)

Sample Complexity for Low-Rank Matrix Reconstruction

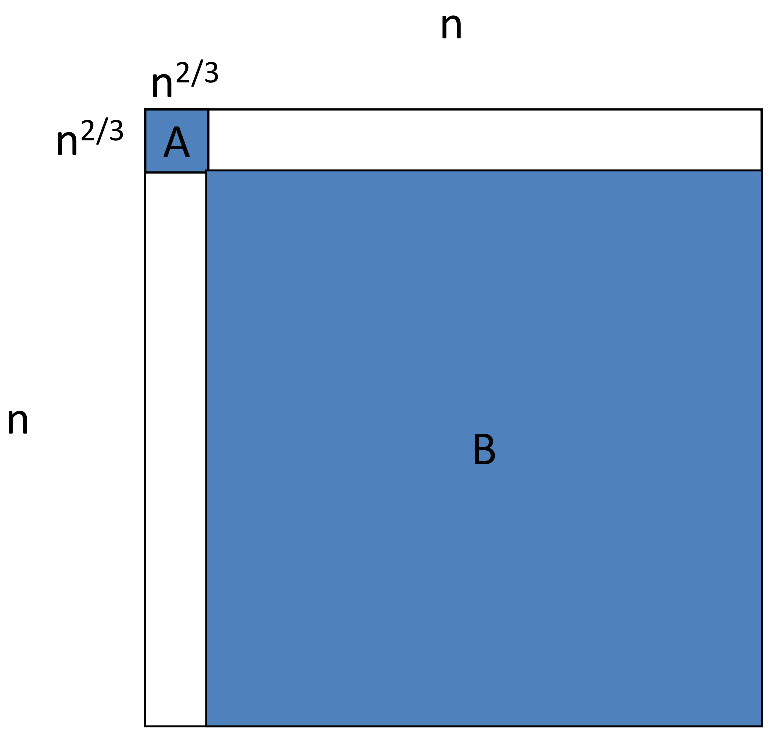
- Y $\approx$ low rank M, observe random subset S of entries
- #sample to get $\frac{1}{nm}|X-Y|_1 \leq \frac{1}{nm}|Y-M|_1 + \epsilon$

  (or, if Y=M+iid noise, to get $\frac{1}{nm}|X-M|_F^2 \leq \epsilon$)
  - Using trace-norm: O(rank(M) (n+m) **log(n)** / ϵ^2)
  - Using max-norm: O(rank(M) (n+m) / ϵ^2)

- If entries sampled non-uniformly:
  - Using trace-norm:

    Ω(rank(M) (n+m) $\sqrt[3]{\boldsymbol{n}}$ / ϵ^2) ~ O(rank(M) (n+m) $\sqrt{\boldsymbol{n}}$ / ϵ^2)
  - Using max-norm: O(rank(M) (n+m) / ϵ^2)

The Trace-Norm with Non-Uniform Sampling



- Both A,B of rank 2
- Sampling:
  - uniform in A w.p. ½
  - uniform w.p. ½

- 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 **S** 10]

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

Experiments on Netflix

	RMSE	%improvement
NetFlix Cinematch: (baseline)	0.9525	0
TraceNorm:	0.9235	3.04
MaxNorm:	0.9138	4.06
Weighted TraceNorm:	0.9078	4.69
Smoothed Wghtd TrNorm:	0.9068	4.80
Local MaxNorm	*0.9063*	*4.85*
Winning team:	0.8553	10.20

Tightness of Max-Norm Relaxation

- Grothendik's inequality:

$$\begin{aligned}&\mathrm{conv}(\{\ \mathrm{rank}(X) \le 1, |X|_\infty \le 1\})\\ &\qquad \subset \{|X|_{\max}^2 \le 1\}\\ &\qquad\qquad \subset 1.79 \cdot \mathrm{conv}(\{\mathrm{rank}(X) \le 1, |X|_\infty \le 1\})\end{aligned}$$

- What about larger k?

$$\begin{aligned}&\mathrm{conv}(\{\ \mathrm{rank}(X) \le k, |X|_\infty \le 1\})\\ &\qquad \subset \{|X|_{\max}^2 \le k, |X|_\infty \le 1\}\\ &\qquad\qquad \subset \mathrm{G(k)} \cdot \mathrm{conv}(\{\mathrm{rank}(X) \le k, |X|_\infty \le 1\})\end{aligned}$$

- How does G(k) grow?

  $1.4 \le \mathrm{G(k)} \le \sqrt{k} \cdot 1.79$

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:

  (bounded entries required for reconstruction gurantees)
  - Max-norm much tighter then trace-norm
  - Better reconstruction guarantees; often better empirical performance