Khintchine-Type Inequalities and Their Applications in Optimization

Anthony Man-Cho So

Department of Systems Engineering & Engineering Management

The Chinese University of Hong Kong

ISDS-Kolloquium

Universitaet Wien

29 June 2009

Background

- A central question in probability theory is to understand the behavior of a sum of independent random variables.
  - Moment inequalities
  - Tail inequalities
  - ...
- Many applications
  - Packet routing
  - Approximate counting
  - Randomized rounding
  - ...

Background

- Consider the following very simple setting. Let
  - a be a vector of real numbers
  - $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs
- What can you say about $\sum \xi_i a_i$?
- In 1923, Khintchine proved that:

$$E\left|\sum \xi_i a_i\right|^p \le K_p \|a\|_2^p$$

- Considerable efforts have been focused on finding the best constant K_p. Haagerup (1982) showed that $K_p = \Theta\left(p^{p/2}\right)$.

Khintchine's Inequality

- The above moment inequality is known as Khintchine's inequality.
- Note that an application of Markov's inequality immediately gives a tail bound:

$$\Pr\left(\left|\sum \xi_i a_i\right| > t\right) = \Pr\left(\left|\sum \xi_i a_i\right|^p > t^p\right) \leq t^{-p} E\left|\sum \xi_i a_i\right|^p$$

- The inequality has since been extended in many directions, for example:
  - each a_i is a vector
  - each a_i is an element in some Banach space
- This gives rise to many Khintchine-type inequalities.

Khintchine's Inequality

- In this talk we are interested in the following setting:
  - $Q_1,\ldots,Q_h$ are $m \times n$ real matrices
  - $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs
- What can we say about $\sum \xi_i Q_i$?
- Specifically, can we bound the spectral norm (i.e. the largest singular value) of $\sum \xi_i Q_i$?
- Such a problem arises in the analyses of many optimization problems.

Khintchine's Inequality

- The form of Khintchine's inequality suggests that we may want to look for an inequality of the form:

$$E\left\|\sum \xi_i Q_i\right\|_{S_p}^p \le K_p \left(\sum \|Q_i\|_{S_p}^2\right)^{p/2}$$

  for some appropriate constant K_p.
- However, as we shall see, we may want some other normalizations on the RHS as well.

Application 1: QP with Norm Constraint

- Consider the following problem:

$$
\begin{aligned}
\max \quad & X \bullet \vec{A}X \\
\text{s.t.} \quad & X \bullet \vec{B}_i X \le 1 \quad \text{for } i = 1, \ldots, L \\
& \vec{C}X = 0 \\
& \|X\|_\infty \le 1 \\
& X \in M^{m,n}
\end{aligned}
\qquad \text{(P)}
$$

where:

- $M^{m,n}$ is the space of m-by-n matrices equipped with the Frobenius inner product $X \bullet Y = tr\left(XY^T\right) = tr\left(X^T Y\right)$
- $\vec{A}, \vec{B}_i$ are symmetric linear mappings, with $\vec{B}_i$ psd
- $\vec{C}$ is a linear mapping
- $\|X\|_\infty$ is the spectral norm (largest singular value) of X

Motivation

- Problem (P) arises in many applications.
- The Procrustes Problem:
  - Given: K collections $P_1,\ldots,P_K$ of points in R^n of the same cardinality, say m.
  - Goal: Find rotations $X_1,\ldots,X_K$ that make these collections as close to each other as possible.
  - Mathematically, we want to:

$$\min \sum_{1\le i<j\le K} \sum_{l=1}^{m} \left\| X_i A_{il} - X_j A_{jl} \right\|_2^2 \;\Leftrightarrow\; \max \sum_{1\le i<j\le K} tr\left(A_i^T X_i^T X_j A_j \right)$$

s.t. $X_i^T X_i = I$ for $i=1,\ldots,K$. This can be put into the form (P).

Procrustes: A Greek Legend

- In Greek mythology, Procrustes was a bandit from Attica who claimed that he had an iron bed that “fits everyone”. However,
  - if a guest was too short, he would stretch him by hamming or racking the body to fit;
  - if a guest was too tall, he would amputate the excess length.
- In either event, the guest died.
- Eventually, Procrustes was made to taste his own medicine by the Attic hero Theseus.

A Related Problem

- A closely related problem, namely that without the norm constraint, is pretty well understood.
- An $O(\log L)$-approximation (where L = no. of constraints) is possible using SDP relaxation (cf. Nemirovski et al. 1999).
- Many applications: clustering, signal processing, etc.

$$\begin{aligned} \max \quad & X \bullet \vec{A} X \\ \text{s.t.} \quad & X \bullet \vec{B}_i X \le 1 \\ & \vec{C} X = 0 \\ & \| X \|_\infty \le 1 \\ & X \in M^{m,n} \end{aligned}$$

A Related Problem

- A closely related problem, namely that without the norm constraint, is pretty well understood.
- An $O(\log L)$ approximation (where L = no. of constraints) is possible using SDP relaxation (cf. Nemirovski et al. 1999).
- Many applications: clustering, signal processing, etc.

$$\begin{aligned} \max \quad & X \bullet \vec{A} X \\ \text{s.t.} \quad & X \bullet \vec{B}_i X \le 1 \\ & \vec{c} X = 0 \\ & \| X \|_\infty \le 1 \\ & X \in M^{m,n} \end{aligned}$$

Question: Does a similar result hold for (P)?

An SDP Relaxation (Nemirovski '07)

- The linear mappings $\vec{A}, \vec{B}_i, \vec{C}$ can be identified with matrices of the appropriate dimension.

$$\begin{aligned}
\max \quad & X \bullet \vec{A}X \qquad A \bullet Y \\
\text{s.t.} \quad & X \bullet \vec{B}_i X \le 1 \qquad B_i \bullet Y \le 1 \\
& \vec{C}X = 0 \qquad C \bullet Y = 0 \\
& \|X\|_\infty \le 1 \\
& X \in M^{m,n} \\
& Y \text{ a Gram matrix} \\
& Y \in S^{mn}
\end{aligned}$$

An SDP Relaxation (Nemirovski '07)

- The linear mappings $\vec{A}, \vec{B}_i, \vec{C}$ can be identified with matrices of the appropriate dimension.
- The norm constraint $\|X\|_\infty \leq 1$ is equivalent to:

$$XX^T \leq I \Leftrightarrow X^T X \leq I$$

These can be expressed as LMIs using appropriate linear mappings.

$$\begin{aligned} \max \quad & A \bullet Y \\ \text{s.t.} \quad & B_i \bullet Y \leq 1 \\ & C \bullet Y = 0 \\ & \vec{S} Y \leq I \\ & \vec{T} Y \leq I \\ & Y \text{ a Gram matrix} \\ & Y \in S^{mn} \end{aligned}$$

An SDP Relaxation (Nemirovski '07)

- We now have the following problem:

$$(P') \qquad \begin{aligned} \max \quad & A \bullet Y \\ \text{s.t.} \quad & B_i \bullet Y \leq 1 \\ & C \bullet Y = 0 \\ & \vec{S} Y \leq I \\ & \vec{T} Y \leq I \\ & Y \text{ a Gram matrix} \end{aligned}$$

- The standard move now is to relax the Gram matrix constraint to an psd constraint.

An SDP Relaxation (Nemirovski '07)

- We now have the following problem:

(SDP)

$$\begin{aligned} \max \quad & A \bullet Y \\ \text{s.t.} \quad & B_i \bullet Y \le 1 \\ & C \bullet Y = 0 \\ & \vec{S} Y \le I \\ & \vec{T} Y \le I \\ & Y \text{ a Gram matrix} \end{aligned} \qquad Y \ge 0, Y \in S^{mn}$$

- The standard move now is to relax the Gram matrix constraint to an psd constraint.
- Note that while $\vec{S} Y \le I$ and $\vec{T} Y \le I$ are redundant in (P'), they are NOT redundant in (SDP).

Quality of SDP Relaxation

- So how well does (SDP) do?
- Nemirovski (2007) proved that an

$$O\left(\max\left\{(m+n)^{1/3}, \log L\right\}\right)$$

  approximation is possible.
  - He also conjectured that an $O\left(\log \max\left\{m, n, L\right\}\right)$ approximation should be achievable.
- Observation (S. 2008): Nemirovski's conjecture is true.
- The proof relies on certain Khintchine-type inequalities.

Rounding the SDP Solution

- A standard way of generating a solution $\hat{X}$ to (P) from a solution Y^* to (SDP) is via randomization.
- Specifically:
  - based on $\vec{A}$ extract from Y^* a set of vectors $\{v_1, \ldots, v_{mn}\}$
  - generate a Bernoulli random vector $\{\xi_1, \ldots, \xi_{mn}\}$
  - form the (random) vector

$$\zeta = \sum_{i=1}^{mn} \xi_i v_i$$

and "de-vectorize" it to obtain a candidate solution matrix $\hat{X}$

Quality of $\hat{X}$

follows from construction

- It is not hard to show that:

$$\hat{X} \bullet \vec{A}\hat{X} = v^*_{sdp} \quad \text{and} \quad \vec{C}\hat{X} = 0$$

(P)

$$\max \quad X \bullet \vec{A}X$$

$$\text{s.t.} \quad X \bullet \vec{B}_i X \le 1$$

$$\vec{C}X = 0$$

$$\|X\|_\infty \le 1$$

$$X \in M^{m,n}$$

Quality of $\hat{X}$

- It is not hard to show that:

  $\hat{X} \bullet \vec{A}\hat{X} = v^*_{sdp}$ and $\vec{C}\hat{X} = 0$

- Thus, to analyze the quality of $\hat{x}$, it remains to bound the following quantities:

  $\Pr\left(\hat{X} \bullet \vec{B}_i \hat{X} > t^2\right)$ and $\Pr\left(\left\|\hat{X}\right\|_\infty > t\right)$

- If these are small, then we can assert that $\hat{X} / t$ is a feasible solution to (P) of value $\geq v^*_{sdp} / t^2 \geq v^* / t^2$ with constant probability.

(P)

$$\max \quad X \bullet \vec{A}X$$

$$\text{s.t.} \quad X \bullet \vec{B}_i X \leq 1$$

$$\vec{C}X = 0$$

$$\|X\|_\infty \leq 1$$

$$X \in M^{m,n}$$

Outline of the Approach

- Those two tail probabilities can be estimated using Khintchine-type inequalities.
- First, the problem of bounding $\Pr\left(\hat{X} \bullet \vec{B}_i \hat{X} > t^2\right)$ can be shown to be equivalent to the following:
  - Let $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs. Let $\{w_i\}$ be vectors satisfying $\sum \|w_i\|_2^2 \leq 1$. Determine an upper bound on $\Pr\left(\left\|\sum \xi_i w_i\right\|_2 \geq t\right)$.

a normalization condition

Outline of the Approach

- On the other hand, the problem of bounding $\Pr\left(\left\|\hat{X}\right\|_\infty > t\right)$ is equivalent to the following:
  - Let $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs. Let $\{Q_i\}$ be m by n matrices satisfying $\sum Q_i Q_i^T \le I$ and $\sum Q_i^T Q_i \le I$. Determine an upper bound on $\Pr\left(\left\|\sum \xi_i Q_i\right\|_\infty \ge t\right)$.

from the constraints $\vec{S}Y \le I, \vec{T}Y \le I$

Tool: Khintchine-type Inequalities

- For the first problem...
  - Let $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs. Let $\{w_i\}$ be arbitrary vectors.
  - Theorem (Tomczak-Jaegermann 1974):

$$E\left\|\sum \xi_i w_i\right\|_2^p \le p^{p/2}\left(\sum \|w_i\|_2^2\right)^{p/2}$$

  - Note that the bound is independent of the number of vectors in the collection!
  - Corollary: Let $T=max\{m,n,L\}$. Then,

$$\begin{aligned}&\Pr\left(\hat{X} \bullet \vec{B}_i \hat{X} > \Omega\left(\beta \log T\right)\right)\\ &= \Pr\left(\left\|\sum \xi_i w_i\right\|_2 > \Omega\left(\sqrt{\beta \log T}\right)\right)\\ &\le O\left(T^{-\beta}\right)\end{aligned}$$

Tool: Khintchine-type Inequalities

- On the other hand...
  - Let $\{\xi_i\}$ be i.i.d. $\{\pm 1\}$ RVs. Let $\{Q_i\}$ be m by n matrices satisfying $\sum Q_i Q_i^T \le I$ and $\sum Q_i^T Q_i \le I$.
  - Theorem (Lust-Piquard 1986, Pisier 1998, Buchholz 2001):

$$E\left\|\sum \xi_i Q_i\right\|_{S_p}^p \le p^{p/2} \max\{m,n\}$$

  - Using the fact that $\|S\|_\infty \le \|S\|_{S_p}$, we obtain:

    Corollary: Let $T=max\{m,n,L\}$. Then,

$$\begin{aligned}
&\Pr\left(\left\|\hat{X}\right\|_\infty > \Omega\left(\sqrt{\beta \log T}\right)\right) \\
&= \Pr\left(\left\|\sum \xi_i Q_i\right\|_\infty > \Omega\left(\sqrt{\beta \log T}\right)\right) \\
&\le O\left(T^{-\beta}\right)
\end{aligned}$$

Putting the Pieces Together

- By picking β appropriately, the above result shows that the rounding scheme of Nemirovski (2007) actually produces a feasible solution $\hat{x}$ to (P) that is within a logarithmic factor from the optimum.

Application 2: Chance-Constrained LMIs

- Consider the following chance-constrained optimization problem:

$$\text{(P)} \qquad \begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & F(x) \le 0 \\ & \Pr\left[A_0(x) - \sum \xi_i A_i(x) \ge 0\right] \ge 1 - \varepsilon \\ & x \in R^n \end{aligned}$$

  - F is an efficiently computable vector-valued function with convex components;
  - each A_i maps x into a symmetric matrix;
  - we assume that $A_0(x) > 0$ for all x.
- Such a problem arises, e.g., in control theory and is in general intractable.

A Safe Tractable Approximation

- One approach for processing (P) is the so-called safe tractable approximation, i.e. a system $\mathcal{H}$ of constraints such that:
  - x is feasible for (P) whenever it is feasible for $\mathcal{H}$
  - the constraints in $\mathcal{H}$ are efficiently computable
- To develop $\mathcal{H}$, observe that:

$$\Pr\left[A_0(x) - \sum \xi_i A_i(x) \geq 0\right] \geq \Pr\left[-I \leq \sum \xi_i A_i'(x) \leq I\right]$$

where:

$$A_i'(x) = A_0^{-1/2}(x) A_i(x) A_0^{-1/2}(x)$$

A Safe Tractable Approximation

- Now, using a matrix version of Khintchine's inequality, one can show that for "nice" ξ:

$$\Pr\left(\left\|\sum \xi_i A_i'(x)\right\|_\infty \le 1\right) \ge 1-\varepsilon$$

whenever:

$$(*) \qquad \sum \left(A_i'(x)\right)^2 \le O\left(\frac{1}{\ln(1/\varepsilon)}\right)\cdot I$$

- The upshot of (*) is that it can be written as an LMI:

$$Z(x) \equiv \begin{bmatrix} \gamma A_0(x) & A_1(x) & \cdots & A_h(x) \\ A_1(x) & \gamma A_0(x) & & \\ \vdots & & \ddots & \\ A_h(x) & & & \gamma A_0(x) \end{bmatrix} \ge 0, \quad \gamma \equiv O\left(\frac{1}{\ln(1/\varepsilon)}\right)$$

A Safe Tractable Approximation

- Thus, we obtain the following safe tractable approximation of (P):

$$\begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & F(x) \leq 0 \\ & Z(x) \geq 0 \\ & x \in R^n \end{aligned}$$

Conclusion

- Moment inequalities are very useful in analyzing randomized algorithms.
  - SDP rank reduction algorithm
  - bounds on stochastic optimization problems
  - analysis of SDP-based detector for MIMO channels
- Find more applications!

Thank You!