Mathematical Programming manuscript No.
(will be inserted by the editor)

Moment Inequalities for Sums of Random Matrices and Their Applications in Optimization

Anthony Man–Cho So

Received: date / Accepted: date

Abstract In this paper, we consider various moment inequalities for sums of random matrices—which are well–studied in the functional analysis and probability theory literature—and demonstrate how they can be used to obtain the best known performance guarantees for several problems in optimization. First, we show that the validity of a recent conjecture of Nemirovski is actually a direct consequence of the so–called non–commutative Khintchine's inequality in functional analysis. Using this result, we show that an SDP–based algorithm of Nemirovski, which is developed for solving a class of quadratic optimization problems with orthogonality constraints, has a logarithmic approximation guarantee. This improves upon the polynomial approximation guarantee established earlier by Nemirovski. Furthermore, we obtain improved safe tractable approximations of a certain class of chance constrained linear matrix inequalities. Secondly, we consider a recent result of Delage and Ye on the so–called data–driven distributionally robust stochastic programming problem. One of the assumptions in the Delage–Ye result is that the underlying probability distribution has bounded support. However, using a suitable moment inequality, we show that the result in fact holds for a much larger class of probability distributions. Given the close connection between the behavior of sums of random matrices and the theoretical properties of various optimization problems, we expect that the moment inequalities discussed in this paper will find further applications in optimization.

Keywords Non–Commutative Khintchine's Inequality · Semidefinite Programming · Approximation Algorithms · Stochastic Programming

A preliminary version of this paper has appeared in the *Proceedings of the 20th Annual ACM–SIAM Symposium on Discrete Algorithms*, 2009.

Anthony Man–Cho So
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong,
Shatin, N. T., Hong Kong
E-mail: manchoso@se.cuhk.edu.hk

Mathematics Subject Classification (2000) 60B20 · 60F10 · 68W20 · 68W25 · 68W40 · 90C15 · 90C22

1 Introduction

In a recent groundbreaking work [25], Nemirovski showed that the theoretical properties of a host of optimization problems are closely related to the behavior of a sum of certain random matrices. Indeed, he showed that the construction of so–called safe tractable approximations of certain chance constrained linear matrix inequalities, as well as the analysis of a semidefinite relaxation of certain non–convex quadratic optimization problems, can be achieved by answering the following question:

Question (Q) *Let $\xi_1, \ldots, \xi_h$ be independent mean zero random variables, each of which is either (i) supported on $[-1, 1]$, or (ii) normally distributed with unit variance. Furthermore, let $Q_1, \ldots, Q_h$ be arbitrary $m \times n$ matrices. Under what conditions on $t > 0$ and $Q_1, \ldots, Q_h$ will we have an exponential decay of the tail probability*

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_\infty \geq t\right)?$$

Here, we use $\|A\|_\infty$ to denote the spectral norm (i.e., the largest singular value) of an $m \times n$ matrix A.

Motivated by such connection, Nemirovski proceeded to establish the following result, which constitutes one possible solution to Question (Q):

Theorem 1 *(Nemirovski [25]) Let $\xi_1, \ldots, \xi_h$ be independent random variables with zero first and third moments, and that each of them is either (i) supported on $[-1, 1]$, or (ii) normally distributed with unit variance. Furthermore, let $Q_1, \ldots, Q_h$ be arbitrary $m \times n$ matrices satisfying*

$$\sum_{i=1}^{h} Q_i Q_i^T \preceq I_m \quad \text{and} \quad \sum_{i=1}^{h} Q_i^T Q_i \preceq I_n, \tag{1}$$

where I_m (resp. I_n) is the $m \times m$ (resp. $n \times n$) identity matrix. Then, whenever $t \geq 7(m+n)^{1/6}$, one has

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_\infty \geq t\right) \leq 22 \exp\left(-t^2/32\right). \tag{2}$$

Note that in order for (2) to hold, some normalization of $Q_1, \ldots, Q_h$ is necessary. One such normalization is provided by condition (1), which is motivated by the requirements of the optimization problems considered in [25]. Now, even without knowing the details of those optimization problems, it is clearly desirable to have (2) holding (perhaps with different constants) for smaller values of t. Moreover, it would be nice to remove the assumption that the random

variables $\xi_1, \ldots, \xi_h$ have zero third moment. However, it is not easy, if not impossible, to extend Nemirovski's proof to obtain the desired improvements, as it involves some very tedious moment calculation. Nevertheless, Nemirovski made the following conjecture, the validity of which would represent a significant improvement over the result of Theorem 1:

Conjecture *(Nemirovski [25]) Let $\xi_1, \ldots, \xi_h$ be independent mean zero random variables, each of which is either (i) supported on $[-1, 1]$, or (ii) normally distributed with unit variance. Furthermore, let $Q_1, \ldots, Q_h$ be arbitrary $m \times n$ matrices satisfying (1). Then, whenever $t = \Omega(\sqrt{\ln(m+n)})$, one has*

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_\infty \geq t\right) \leq \Theta(1) \cdot \exp\left(-\Theta(1) \cdot t^2\right).$$

As argued in [25], the threshold $t = \Omega(\sqrt{\ln(m+n)})$ is in some sense the best one could hope for. Moreover, if the above conjecture is true, then it would immediately imply improved performance guarantees for various optimization problems. Therefore, there is great interest in determining the validity of this conjecture.

As it turns out, the behavior of the random variable $S_h \equiv \sum_{i=1}^{h} \xi_i Q_i$ has been extensively studied in the functional analysis and probability theory literature. One of the tools that is particularly useful for addressing Nemirovski's conjecture and other problems considered in this paper is the so–called Khintchine–type inequalities. Roughly speaking, such inequalities provide upper bounds on the p–norm of the random variable $\|S_h\|_\infty$ in terms of suitable normalizations of the matrices $Q_1, \ldots, Q_h$. Once these bounds are available, it is easy to derive tail bounds for $\|S_h\|_\infty$ using Markov's inequality. In this paper, we show that the validity of Nemirovski's conjecture is in fact a simple consequence of the so–called non–commutative Khintchine's inequality in functional analysis (see, e.g., [24,33,10]). As an immediate corollary, we obtain the best known performance guarantees for the two optimization problems considered in [25]—namely, the construction of safe tractable approximations of certain chance constrained linear matrix inequalities, and the analysis of a semidefinite relaxation of certain non–convex quadratic optimization problems with orthogonality constraints. To further demonstrate the power of Khintchine–type inequalities, we consider another such inequality, which is due to Tomczak–Jaegermann [42] and differs from the non–commutative Khintchine's inequality mainly on the normalization of $Q_1, \ldots, Q_h$, and show how it can be used to extend a recent result of Delage and Ye [11] on data–driven distributionally robust stochastic programming.

The rest of this paper is organized as follows. In Section 2, we discuss the various moment inequalities that will be used in our analyses. Then, in Section 3, we consider three problems in optimization and show how the moment inequalities introduced in Section 2 can be used to obtain the best known performance guarantees for them. Finally, we end with some concluding remarks in Section 4.

2 Moment Inequalities for Sums of Random Matrices

To motivate our discussion, let us first consider the case where $\xi_1, \ldots, \xi_h$ are independent Bernoulli random variables (i.e., each ξ_i takes on the values ± 1 with equal probability), and let $Q_1, \ldots, Q_h$ be arbitrary scalars (i.e., each Q_i is an 1×1 matrix). Recall that we are interested in determining the behavior of the random variable

$$\left\| \sum_{i=1}^{h} \xi_i Q_i \right\|_{\infty} = \left| \sum_{i=1}^{h} \xi_i Q_i \right|.$$

In 1923, in an effort to provide a sharp estimate on the rate of convergence in Borel's strong law of large numbers, Khintchine proved the following inequality that now bears his name [19]:

Khintchine's Inequality *Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables, and let $Q_1, Q_2, \ldots$ be an arbitrary sequence of scalars. Then, for any $h = 1, 2, \ldots$ and $p \in (0, \infty)$, there exists an absolute constant $c_p > 0$ such that*

$$\mathbb{E}\left[\left| \sum_{i=1}^{h} \xi_i Q_i \right|^p \right] \leq c_p \cdot \left(\sum_{i=1}^{h} |Q_i|^2 \right)^{p/2}. \tag{3}$$

(In fact, Khintchine only established the inequality for the case where $p \geq 2$ is an even integer. However, as shown in, e.g., [32], his proof can be extended to cover other values of p.) Since then, much effort has been spent on determining the optimal value of c_p in (3). In particular, it has been shown that for $p \geq 2$ (which will be the case of interest in our study), the value

$$c_p^* = \left(\frac{2^p}{\pi} \right)^{1/2} \Gamma\left(\frac{p+1}{2} \right)$$

is the best possible; see, e.g., [32] for a brief historical account of this result. Note that c_p^* is exactly equal to $\mathbb{E}\left[|Z|^p\right]$, where Z is a standard normal random variable. This should not be surprising, as the random variable $\sum_{i=1}^{h} \xi_i Q_i$ behaves (after suitable normalization) like a standard normal random variable by the Central Limit Theorem. Indeed, a quick calculation shows that (3) holds with equality and $c_p = c_p^*$ when $\xi_1, \xi_2, \ldots$ are i.i.d. standard normal random variables. Finally, using Stirling's formula (see, e.g., [34]), one can show that c_p^* is of order $p^{p/2}$ (in fact, $c_p^* < p^{p/2}$) for all $p \geq 2$.

Subsequent to the appearance of Khintchine's inequality, many extensions have been investigated. Of particular interest to us is the case where the elements $Q_1, Q_2, \ldots$ are arbitrary $m \times n$ matrices. In 1974, Tomczak–Jaegermann proved the following inequality, which can be viewed as a natural extension of (3):

Theorem 2 *(Tomczak–Jaegermann [42]) Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables, and let $Q_1, Q_2, \ldots$ be an arbitrary sequence of $m \times n$ matrices. Then, for any $h = 1, 2, \ldots$ and $p \geq 2$, we have*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_{S_p}^{p}\right] \leq p^{p/2} \cdot \left(\sum_{i=1}^{h} \|Q_i\|_{S_p}^2\right)^{p/2}.$$

Here, $\|A\|_{S_p}$ denotes the Schatten p–norm of an $m \times n$ matrix A, i.e., $\|A\|_{S_p} = \|\sigma(A)\|_p$, where $\sigma(A) \in \mathbb{R}^{\min\{m,n\}}$ is the vector of singular values of A, and $\|\cdot\|_p$ is the usual ℓ_p–norm.

As it turns out, the normalization $\sum_{i=1}^{h} \|Q_i\|_{S_p}^2$ is not the only one possible in order for a Khintchine–type inequality to hold. In 1986, Lust–Piquard showed that with a different normalization, an inequality similar to the one in Theorem 2 is also valid:

Theorem 3 *(Non–Commutative Khintchine's Inequality; Lust–Piquard [24]) Let $\xi_1, \xi_2, \ldots$ be a sequence of independent Bernoulli random variables, and let $Q_1, Q_2, \ldots$ be an arbitrary sequence of $m \times n$ matrices. Then, for any $h = 1, 2, \ldots$ and $p \geq 2$, there exists an absolute constant $\gamma_p > 0$ such that*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_{S_p}^{p}\right] \leq \gamma_p \cdot \max\left\{\left\|\left(\sum_{i=1}^{h} Q_i Q_i^T\right)^{1/2}\right\|_{S_p}^{p}, \left\|\left(\sum_{i=1}^{h} Q_i^T Q_i\right)^{1/2}\right\|_{S_p}^{p}\right\}.$$

Unfortunately, the proof of Lust–Piquard does not provide an estimate for γ_p. In [33], Pisier showed that $\gamma_p \leq \alpha p^{p/2}$ for some absolute constant $\alpha > 0$. Using a result of Buchholz [10], it can be shown that $\alpha \leq 2^{-p/4}(\pi/e)^{p/2} < 1$ for all $p \geq 2$ (see, e.g., [43]). We note that Theorem 3 is also valid (with $\gamma_p \leq \alpha p^{p/2} < p^{p/2}$) when $\xi_1, \xi_2, \ldots$ are i.i.d. standard normal random variables [10].

As a first illustration of the power of the aforementioned inequalities, let us see how Theorem 3 can be used to resolve Nemirovski's conjecture in the affirmative:

Theorem 4 *Let $\xi_1, \ldots, \xi_h$ be independent mean zero random variables, each of which is either (i) supported on $[-1, 1]$, or (ii) normally distributed with unit variance. Furthermore, let $Q_1, \ldots, Q_h$ be arbitrary $m \times n$ matrices satisfying $\max\{m, n\} \geq 2$ and condition (1). Then, for any $\beta \geq 1/2$, we have*

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_{\infty} \geq \sqrt{2e(1+\beta)\ln\max\{m,n\}}\right) \leq (\max\{m,n\})^{-\beta}$$

if $\xi_1, \ldots, \xi_h$ are i.i.d. Bernoulli or standard normal random variables; and

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i Q_i\right\|_{\infty} \geq \sqrt{8e(1+\beta)\ln\max\{m,n\}}\right) \leq (\max\{m,n\})^{-\beta}$$

if $\xi_1, \ldots, \xi_h$ are independent mean zero random variables supported on $[-1, 1]$.

Proof Since $Q_1, \ldots, Q_h$ satisfy condition (1), all the eigenvalues of $\sum_{i=1}^h Q_i Q_i^T$ and $\sum_{i=1}^h Q_i^T Q_i$ lie in $[0,1]$. It follows that

$$\left\| \left(\sum_{i=1}^h Q_i Q_i^T \right)^{1/2} \right\|_{S_p} \leq m^{1/p} \quad \text{and} \quad \left\| \left(\sum_{i=1}^h Q_i^T Q_i \right)^{1/2} \right\|_{S_p} \leq n^{1/p}.$$

Now, let us first consider the case where $\xi_1, \ldots, \xi_h$ are i.i.d. Bernoulli or standard normal random variables. By Theorem 3 and the remarks following it, we have

$$\mathbb{E}\left[\left\| \sum_{i=1}^h \xi_i Q_i \right\|_\infty^p \right] \leq \mathbb{E}\left[\left\| \sum_{i=1}^h \xi_i Q_i \right\|_{S_p}^p \right] \leq p^{p/2} \cdot \max\{m,n\}$$

for any $p \geq 2$. Thus, by Markov's inequality, for any $t > 0$ and $p \geq 2$, we have

$$\Pr\left(\left\| \sum_{i=1}^h \xi_i Q_i \right\|_\infty \geq t \right) \leq t^{-p} \cdot \mathbb{E}\left[\left\| \sum_{i=1}^h \xi_i Q_i \right\|_\infty^p \right] \leq \frac{p^{p/2} \cdot \max\{m,n\}}{t^p}.$$

Upon setting $t = \sqrt{2e(1+\beta)\ln \max\{m,n\}}$ and $p = t^2/e > 2$ (since $\beta \geq 1/2$ and $\max\{m,n\} \geq 2$ by assumption), we obtain

$$\Pr\left(\left\| \sum_{i=1}^h \xi_i Q_i \right\|_\infty \geq \sqrt{2e(1+\beta)\ln \max\{m,n\}} \right) \leq (\max\{m,n\})^{-\beta}$$

as desired.

Next, we consider the case where $\xi_1, \ldots, \xi_h$ are independent mean zero random variables supported on $[-1,1]$. Let $\epsilon_1, \ldots, \epsilon_h$ be i.i.d. Bernoulli random variables that are independent of the ξ_i's. A standard symmetrization argument (see, e.g., [21, Lemma 6.3]), together with Fubini's theorem and Theorem 3, implies that

$$\begin{aligned}
\mathbb{E}\left[\left\| \sum_{i=1}^h \xi_i Q_i \right\|_{S_p}^p \right] &\leq 2^p \cdot \mathbb{E}_\xi \mathbb{E}_\epsilon \left[\left\| \sum_{i=1}^h \epsilon_i \xi_i Q_i \right\|_{S_p}^p \right] \\
&\leq 2^p \cdot p^{p/2} \cdot \mathbb{E}_\xi \left[\max\left\{ \left\| \left(\sum_{i=1}^h \xi_i^2 Q_i Q_i^T \right)^{1/2} \right\|_{S_p}^p, \right. \right. \\
&\qquad \left. \left. \left\| \left(\sum_{i=1}^h \xi_i^2 Q_i^T Q_i \right)^{1/2} \right\|_{S_p}^p \right\} \right] \\
&\leq 2^p \cdot p^{p/2} \cdot \max\{m,n\}.
\end{aligned}$$

Here, $\mathbb{E}_\xi$ (resp. $\mathbb{E}_\epsilon$) denotes the mathematical expectation with respect to the random variables $\xi_1, \ldots, \xi_h$ (resp. $\epsilon_1, \ldots, \epsilon_h$). The desired result then follows from an application of Markov's inequality. □

In the sequel, we will demonstrate further how the inequalities introduced in this section can be used to tackle various problems in optimization.

3 Applications

3.1 Non–Convex Quadratic Optimization with Orthogonality Constraints

Consider the following class of quadratic optimization problems:

$$\text{(QP–OC)} \qquad \begin{array}{lll} \text{maximize} & X \bullet \mathcal{A}X & \\ \text{subject to} & X \bullet \mathcal{B}_i X \leq 1 \quad \text{for } i = 1, \ldots, L, & (a) \\ & \mathcal{C}X = \mathbf{0}, & (b) \\ & \|X\|_\infty \leq 1, & (c) \\ & X \in \mathscr{M}^{m,n}, & (d) \end{array}$$

where

- $\mathscr{M}^{m,n}$ is the space of $m \times n$ real matrices equipped with the trace inner product $X \bullet Y = \operatorname{tr}\left(XY^T\right) = \operatorname{tr}\left(X^T Y\right)$;
- $\mathcal{A}, \mathcal{B}_1, \ldots, \mathcal{B}_L : \mathscr{M}^{m,n} \to \mathscr{M}^{m,n}$ are self–adjoint linear operators (in particular, they can be represented as symmetric $mn \times mn$ matrices);
- $\mathcal{B}_1, \ldots, \mathcal{B}_L$ are positive semidefinite;
- $\mathcal{C} : \mathscr{M}^{m,n} \to \mathbb{R}^u$ is a linear mapping (in particular, it can be represented as an $u \times mn$ matrix);
- $\|X\|_\infty$ is the spectral norm of X (recall that

$$\|X\|_\infty = \max\left\{\|Xv\|_2 : v \in \mathbb{R}^n,\ \|v\|_2 = 1\right\}$$

by the Courant–Fischer theorem; see, e.g., [17, Theorem 7.3.10]).

As pointed out by Nemirovski [25], Problem (QP–OC) is quite general and captures several well–studied problems in the literature as special cases. Before we proceed to study the algorithmic aspects of (QP–OC), let us consider two such problems, namely, the Procrustes Problem and the so–called orthogonal relaxation of the Quadratic Assignment Problem. A common feature of these two problems is that they both contain the orthogonality constraint $X^T X = I$, which at first sight does not seem to fit into the form (QP–OC). However, by exploiting the structure of these problems, we may relax the orthogonality constraint to the norm constraint $\|X\|_\infty \leq 1$ with no loss of generality.

The Procrustes Problem

In the Procrustes Problem, one is given K collections $\mathcal{P}_1, \ldots, \mathcal{P}_K$ of points in $\mathbb{R}^n$ with the same cardinality $|\mathcal{P}_1| = \cdots = |\mathcal{P}_k| = m$, and the goal is to find rotations that make these collections as close to each other as possible. More

precisely, let A_i be an $n \times m$ matrix whose l–th column represents the l–th point in the i–th collection, where $i = 1, \ldots, K$ and $l = 1, \ldots, m$. The goal is to find K $n \times n$ orthogonal matrices $X_1, \ldots, X_K$ such that the quantity

$$\sum_{1 \le i < j \le K} \sum_{l=1}^{m} \|X_i A_{il} - X_j A_{jl}\|_2^2$$

is minimized. Here, A_{il} is the l–th column of the matrix A_i, where $i = 1, \ldots, K$ and $l = 1, \ldots, m$. Note that the quantity $\|X_i A_{il} - X_j A_{jl}\|_2^2$ represents the squared Euclidean distance between the l–th transformed point in the i–th collection and the l–th transformed point in the j–th collection. The Procrustes Problem is first studied in psychometrics and has now found applications in shape and image analyses, market research and biometric identification, just to name a few (see [15] for details). It is not hard to show that the Procrustes Problem as defined above is equivalent to

$$\text{maximize} \quad \sum_{1 \le i < j \le K} \operatorname{tr}\left(A_i^T X_i^T X_j A_j\right) \quad \text{subject to} \quad X_i^T X_i = I \quad \text{for } i = 1, \ldots, K. \tag{4}$$

Now, notice that the objective function is linear in each of the X_i's. Thus, we may relax the orthogonality constraint $X_i^T X_i = I$ to the norm constraint $\|X_i\|_\infty \le 1$ without affecting the optimal value of the problem (we refer the reader to [25] for details). In other words, Problem (4) has the same optimal value as the problem

$$\text{maximize} \quad \sum_{1 \le i < j \le K} \operatorname{tr}\left(A_i^T X_i^T X_j A_j\right) \quad \text{subject to} \quad \|X_i\|_\infty \le 1 \quad \text{for } i = 1, \ldots, K,$$

which, after some elementary manipulations, can be cast into the form (QP–OC). □

Orthogonal Relaxation of the Quadratic Assignment Problem

In the Quadratic Assignment Problem (QAP), one is given a set $\mathcal{N} = \{1, \ldots, n\}$, two $n \times n$ symmetric matrices A and B, and an $n \times n$ matrix C, and the goal is to find a permutation π on $\mathcal{N}$ such that the quantity $\sum_{i=1}^n \sum_{j=1}^n A_{\pi(i)\pi(j)} B_{ij} - 2\sum_{i=1}^n C_{i\pi(i)}$ is maximized. Equivalently, one can formulate the QAP as follows (see, e.g., [20,46]):

$$\begin{aligned} \text{maximize} \quad & \operatorname{tr}\left(AXBX^T - 2CX^T\right) \\ \text{subject to} \quad & XX^T = I, \\ & X_{ij} \in \{0, 1\} \qquad \text{for } i = 1, \ldots, n;\ j = 1, \ldots, n. \end{aligned} \tag{5}$$

The constraints in Problem (5) force the matrix X to be a permutation matrix. Indeed, it is well–known that X satisfies the constraints in (5) iff X is a permutation matrix. The QAP is a classical problem in combinatorial optimization and has found many applications (see, e.g., [30]). However, it is also a notoriously hard computational problem. Therefore, various relaxations

have been proposed. One such relaxation, called the orthogonal relaxation, is obtained by dropping the binary constraints in (5). In other words, consider the following problem:

$$\text{maximize } \operatorname{tr}\left(AXBX^T - 2CX^T\right) \text{ subject to } XX^T = I. \tag{6}$$

Note that we can also add the redundant constraint $X^TX = I$ to Problem (6), and indeed this can tighten the relaxation (see, e.g., [46,3,4]). Now, suppose that we have $A, B \succ \mathbf{0}$. Let $A^{1/2}$ and $B^{1/2}$ be the $n \times n$ symmetric positive definite matrices such that $A = A^{1/2}A^{1/2}$ and $B = B^{1/2}B^{1/2}$. Then, the objective function in Problem (6) can be written as

$$\operatorname{tr}\left((A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T - 2CX^T\right),$$

which is a convex quadratic form in X. Consequently, we may relax the orthogonality constraint $XX^T = I$ to the norm constraint $\|X\|_\infty \leq 1$ without affecting the optimal value of the problem (again, we refer the reader to [25] for details). In particular, Problem (6) is equivalent to

$$\text{maximize } \operatorname{tr}\left(AXBX^T - 2CX^T\right) \text{ subject to } \|X\|_\infty \leq 1,$$

which can be cast into the form (QP–OC) after a standard homogenization argument (see, e.g., [25]).

The above reformulation is based on the assumption that $A, B \succ \mathbf{0}$. However, observe that the set of optimal solutions to Problem (6) does not change if we make the substitution $A \leftarrow A+\alpha I$ and $B \leftarrow B+\beta I$ for any fixed $\alpha, \beta \in \mathbb{R}$. Thus, we may always assume without loss that $A, B \succ \mathbf{0}$. We should emphasize, however, that such a substitution could significantly weaken subsequent relaxations of Problem (6). □

Given the generality of Problem (QP–OC), it should come as no surprise that it is NP–hard. Indeed, let A be an $m \times m$ symmetric positive semidefinite matrix, and consider the following binary quadratic optimization problem:

$$\text{(BQP)} \quad \text{maximize } x^TAx \text{ subject to } x_i^2 = 1 \text{ for } i = 1, \ldots, L.$$

It is well–known that (BQP) includes the MAX–CUT problem as a special case and hence it is NP–hard (see, e.g., [29]). Now, by the convexity of the objective function $x \mapsto x^TAx$, we see that (BQP) is equivalent to

$$\text{(BQP}') \quad \text{maximize } x^TAx \text{ subject to } \|x\|_\infty \leq 1 \text{ for } i = 1, \ldots, L,$$

which is an instance of (QP–OC) with $n = 1$, $\mathcal{B}_1 = \cdots = \mathcal{B}_L = \mathbf{0}$ and $\mathcal{C} = \mathbf{0}$. It follows that (QP–OC) is also NP–hard, as claimed.

The above hardness result motivates us to search for efficient algorithms that can solve Problem (QP–OC) approximately. In [25], Nemirovski proposed to use semidefinite programming (SDP) relaxation to tackle (QP–OC). In recent years, SDP has become an invaluable tool in the design of approximation algorithms. Beginning with the seminal work of Goemans and Williamson

[14], who showed that SDP can be used to obtain good approximation algorithms for MAX–CUT and various satisfiability problems, researchers have successfully employed the SDP approach to design approximation algorithms for problems in combinatorial optimization (see, e.g., [13,18,2,6,5]), telecommunications (see, e.g., [23,39,38]) and quadratic optimization (see, e.g., [29, 44,41,40]). In fact, for many of those problems, the SDP approach yields the best known approximation to date, and the situation is no different for the case of (QP–OC). Before we delve into the details of Nemirovski's approach, let us describe some of its high–level ideas.

As it turns out, the main feature that distinguishes (QP–OC) from the quadratic optimization problems considered in the approximation algorithms literature is the norm constraint (QP–OC(c)). Indeed, if we drop the norm constraint (QP–OC(c)), then (QP–OC) becomes a usual quadratic program, and an $O(\ln L)$ approximation algorithm for it is known [26][1]. In [25], Nemirovski showed that a natural SDP relaxation of (QP–OC) together with a simple rounding scheme yields an $O\left(\max\left\{(m+n)^{1/3}, \ln L\right\}\right)$ approximation algorithm for (QP–OC). The rounding scheme proposed in [25] resembles that of Nemirovski et al. [26]. Roughly speaking, it consists of the following steps:

1. extract from the optimal SDP solution a set $S = \{v_1, \ldots, v_{mn}\}$ of vectors and apply a suitable orthogonal transformation to S to obtain vectors $v'_1, \ldots, v'_{mn}$
2. generate a random vector $\xi = (\xi_1, \ldots, \xi_{mn})$, where the entries are i.i.d. Bernoulli random variables
3. form the (random) vector $\zeta = \sum_{i=1}^{mn} \xi_i v'_i$ and extract from ζ a candidate solution matrix $\widehat{X}$

In order to analyze the performance of such a procedure, one needs to determine the behavior of $\widehat{X}$ with respect to both the objective function and the constraints in (QP–OC). Intuitively, the objective function and the constraints (QP–OC(a)) and (QP–OC(b)) pose no difficulty, as one should be able to analyze the behavior of $\widehat{X}$ with respect to those in a manner similar to that in [26]. However, it is more challenging to analyze the behavior of $\widehat{X}$ with respect to the norm constraint (QP–OC(c)). Indeed, as it was shown in [25], the problem boils down to that of answering Question (Q), i.e., we need to estimate the typical spectral norm of a sum of certain random matrices. Fortunately, this can be easily done using the moment inequalities introduced in Section 2. In particular, using Theorem 4, we show that the SDP–based algorithm described in [25] actually yields an $O\left(\ln \max\{m, n, L\}\right)$ approximation for (QP–OC). This significantly improves upon the $O\left(\max\left\{(m+n)^{1/3}, \ln L\right\}\right)$ bound established in [25] and provides the first logarithmic approximation guarantee for (QP–OC). Before we prove this result, let us review the derivation of Nemirovski's SDP relaxation of (QP–OC).

[1] In fact, for the case where the norm constraint (QP–OC(c)) is absent and $L \leq 2$, Problem (QP–OC) can be efficiently solved using SDP. This follows from the results of Shapiro [36], Barvinok [7] and Pataki [31]; see also [45,8,1] for related results.

3.1.1 A Semidefinite Relaxation of (QP–OC)

The ideas used in the derivation of the SDP relaxation are fairly standard: first linearize the quadratic terms and then tighten the relaxation with positive semidefinite constraints. To begin, let us identify the mapping $\mathcal{A}$ with an $mn \times mn$ symmetric matrix A whose rows and columns are indexed by pairs (i, j), where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Specifically, let e_i be the i–th standard basis vector whose dimension depends on the context. The (k, l)–th column of A is given by $\operatorname{Vec} \mathcal{A}E_{kl}$, where $E_{kl} = e_k e_l^T \in \mathscr{M}^{m,n}$ and $\operatorname{Vec} X$ denotes the mn–dimensional vector obtained by stacking the columns of the $m \times n$ matrix X into a single column. Finally, the entries of A are given by

$$A_{(i,j)(k,l)} = E_{ij} \bullet \mathcal{A}E_{kl} \qquad \text{for } i, k = 1, \ldots, m;\ j, l = 1, \ldots, n.$$

In a similar fashion, we identify the mappings $\mathcal{B}_i$ with $mn \times mn$ symmetric positive semidefinite matrices B_i, where $i = 1, \ldots, L$. For the mapping $\mathcal{C}$, we identify it with an $u \times mn$ matrix C whose entries are given by

$$C_{i,(k,l)} = e_i^T \mathcal{C} E_{kl} \qquad \text{for } i = 1, \ldots, u;\ k = 1, \ldots, m;\ l = 1, \ldots, n.$$

Now, for $X \in \mathscr{M}^{m,n}$, let $\operatorname{Gram} X$ be the $mn \times mn$ positive semidefinite matrix $(\operatorname{Vec} X)(\operatorname{Vec} X)^T$. It is then clear that

$$X \bullet \mathcal{A}X = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} (\mathcal{A}X)_{ij} = \sum_{i,k=1}^{m} \sum_{j,l=1}^{n} A_{(i,j)(k,l)} X_{ij} X_{kl} = A \bullet \operatorname{Gram} X.$$

Similarly, we have $X \bullet \mathcal{B}_i X = B_i \bullet \operatorname{Gram} X$ for $i = 1, \ldots, L$. To express the constraints (QP–OC(b)) and (QP–OC(c)) in terms of $\operatorname{Gram} X$, we follow the approach of Zhao et al. [46]. First, observe that

$$\mathcal{C}X = \mathbf{0} \Leftrightarrow C \operatorname{Vec} X = \mathbf{0} \Leftrightarrow \|C \operatorname{Vec} X\|_2^2 = 0 \Leftrightarrow (\operatorname{Vec} X)^T C^T C (\operatorname{Vec} X) = 0.$$

Thus, we have

$$\mathcal{C}X = \mathbf{0} \Leftrightarrow C^T C (\operatorname{Vec} X)(\operatorname{Vec} X)^T = 0 \Leftrightarrow C^T C \bullet \operatorname{Gram} X = 0.$$

Next, observe that $\|X\|_\infty \leq 1$ iff $XX^T \preceq I_m$. Now, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$, the (i, j)–th entry of XX^T is $\sum_{k=1}^{n} X_{ik} X_{jk}$. It follows that the entries of XX^T are linear combinations of the entries in $\operatorname{Gram} X$, which in turn implies the existence of a linear mapping $\mathcal{S} : \mathscr{S}^{mn} \to \mathscr{S}^m$ such that $XX^T \preceq I_m$ iff $\mathcal{S} \operatorname{Gram} X \preceq I_m$ (here, $\mathscr{S}^m$ is the space of $m \times m$ real symmetric matrices). In a similar fashion, we have $\|X\|_\infty \leq 1$ iff $X^T X \preceq I_n$, and there exists a linear mapping $\mathcal{T} : \mathscr{S}^{mn} \to \mathscr{S}^n$ such that $X^T X \preceq I_n$ iff $\mathcal{T} \operatorname{Gram} X \preceq I_n$. Note that both the linear mappings $\mathcal{S}$ and $\mathcal{T}$ can be specified explicitly as matrices of appropriate dimensions in polynomial time;

see, e.g., [46]. Now, upon putting the pieces together and using the fact that $\operatorname{Gram} X \succeq \mathbf{0}$, we obtain the following SDP relaxation of (QP–OC):

$$\text{(QP–OC–SDR)} \qquad \begin{array}{lll} \text{maximize} & A \bullet Y & \\ \text{subject to} & B_i \bullet Y \leq 1 & \text{for } i = 1, \ldots, L, \\ & C^T C \bullet Y = 0, & \\ & \mathcal{S}Y \preceq I_m,\ \mathcal{T}Y \preceq I_n, & \\ & Y \in \mathscr{S}^{mn},\ Y \succeq \mathbf{0}. & \end{array}$$

Although the constraints $\mathcal{S} \operatorname{Gram} X \preceq I_m$ and $\mathcal{T} \operatorname{Gram} X \preceq I_n$ imply each other and are thus redundant in (QP–OC), the corresponding relaxed constraints $\mathcal{S}Y \preceq I_m$ and $\mathcal{T}Y \preceq I_n$ are *not* redundant in (QP–OC–SDR). In fact, the inclusion of these constraints can significantly strengthen various SDP relaxations of the Quadratic Assignment Problem (see, e.g., [46,3,4]). Moreover, as we shall see, they play a crucial role in the quality analysis of (QP–OC–SDR).

Now, using the ellipsoid method [16], the semidefinite program (QP–OC–SDR) can be solved to within an additive error of $\epsilon > 0$ in polynomial time. Specifically, let θ^* be the optimal value of (QP–OC–SDR). Then, for any $\epsilon > 0$, we can compute in polynomial time a $Y' \succeq \mathbf{0}$ that is feasible for (QP–OC–SDR) and satisfies $\theta' \equiv A \bullet Y' \geq \theta^* - \epsilon$.

3.1.2 Analysis of the SDP Relaxation

Let us now analyze the quality of the SDP relaxation (QP–OC–SDR). Our goal in this section is to prove the following theorem:

Theorem 5 *There exists an efficient randomized algorithm that, given a feasible solution of (QP–OC–SDR) with objective value θ', produces an $m \times n$ matrix $\overline{X}$ such that*

(a) $\overline{X}$ is feasible for (QP–OC); and
(b) $\overline{X} \bullet \mathcal{A}\overline{X} \geq \Omega\left(1/\ln \max\{m, n, L\}\right) \cdot \theta'$.

To begin, let us consider the following rounding scheme of Nemirovski [25] that converts a feasible solution Y' of (QP–OC–SDR) to a random $m \times n$ matrix $\widehat{X}$. Since $Y' \succeq \mathbf{0}$, there exists a positive semidefinite matrix $Y'^{1/2} \in \mathscr{S}^{mn}$ such that $Y' = Y'^{1/2} Y'^{1/2}$. Moreover, the matrix $Y'^{1/2} A Y'^{1/2}$ is symmetric, and hence it admits a spectral decomposition $Y'^{1/2} A Y'^{1/2} = U^T \Lambda U$, where Λ is an $mn \times mn$ diagonal matrix and U is an $mn \times mn$ orthogonal matrix. Now, we generate a random mn–dimensional vector $\xi = (\xi_{ij})_{1 \leq i \leq m,\, 1 \leq j \leq n}$, where the entries are i.i.d. Bernoulli random variables. Then, we define the random $m \times n$ matrix $\widehat{X}$ via $\operatorname{Vec} \widehat{X} = Y'^{1/2} U^T \xi$.

Clearly, the above rounding scheme can be implemented in polynomial time. We are now interested in the quality of the solution $\widehat{X}$. In the sequel, we assume that $\max\{m, n\} \geq 2$. The following proposition is established in [25]. For completeness' sake, we include the proof here.

Proposition 1 *The solution $\widehat{X}$ returned by the rounding procedure satisfies*

(a) $\widehat{X} \bullet \mathcal{A}\widehat{X} \equiv \theta'$;
(b) $\mathbb{E}[\widehat{X} \bullet \mathcal{B}_i\widehat{X}] \leq 1$ *for* $i = 1, \ldots, L$;
(c) $\mathcal{C}\widehat{X} \equiv \mathbf{0}$; *and*
(d) $\mathbb{E}[\widehat{X}\widehat{X}^T] \preceq I_m$ *and* $\mathbb{E}[\widehat{X}^T\widehat{X}] \preceq I_n$.

Proof Observe that

$$\widehat{X} \bullet \mathcal{A}\widehat{X} = A \bullet \operatorname{Gram} \widehat{X} = \xi^T U Y'^{1/2} A Y'^{1/2} U^T \xi = \xi^T U U^T \Lambda U U^T \xi = \operatorname{tr}(\Lambda).$$

On the other hand, we have

$$\operatorname{tr}(\Lambda) = \operatorname{tr}\left(U^T \Lambda U\right) = \operatorname{tr}\left(Y'^{1/2} A Y'^{1/2}\right) = A \bullet Y' = \theta'.$$

This establishes (a). Next, we compute

$$\mathbb{E}[\widehat{X} \bullet \mathcal{B}_i\widehat{X}] = \mathbb{E}[B_i \bullet \operatorname{Gram} \widehat{X}] = B_i \bullet \mathbb{E}\left[Y'^{1/2} U^T \xi\xi^T U Y'^{1/2}\right] = B_i \bullet Y' \leq 1,$$

where the third equality follows from the fact that $\mathbb{E}\left[\xi\xi^T\right] = I_{mn}$. This proves (b). To prove (c), we first note that $\mathcal{C}\widehat{X} = C\operatorname{Vec} \widehat{X} = CY'^{1/2}U^T\xi$. Now, observe that

$$\begin{aligned}
\mathbb{E}[\|\mathcal{C}\widehat{X}\|_2^2] &= \mathbb{E}\left[\xi^T U Y'^{1/2} C^T C Y'^{1/2} U^T \xi\right] \\
&= U Y'^{1/2} C^T C Y'^{1/2} U^T \bullet \mathbb{E}\left[\xi\xi^T\right] \\
&= C^T C \bullet Y' \\
&= 0.
\end{aligned}$$

Since $\|\mathcal{C}\widehat{X}\|_2^2 \geq 0$ for any realization of $\xi \in \{-1, +1\}^{mn}$, it follows that $\mathcal{C}\widehat{X} \equiv \mathbf{0}$, as desired. Finally, to prove (d), it suffices to observe that

$$\mathbb{E}[\widehat{X}\widehat{X}^T] = \mathbb{E}[\mathcal{S}\operatorname{Gram} \widehat{X}] = \mathcal{S}\left(Y'^{1/2} U^T \mathbb{E}\left[\xi\xi^T\right] U Y'^{1/2}\right) = \mathcal{S}Y' \preceq I_m,$$

where the second equality follows from the linearity of $\mathcal{S}$. In a similar fashion, we obtain

$$\mathbb{E}[\widehat{X}^T\widehat{X}] = \mathbb{E}[\mathcal{T}\operatorname{Gram} \widehat{X}] = \mathcal{T}\left(Y'^{1/2} U^T \mathbb{E}\left[\xi\xi^T\right] U Y'^{1/2}\right) = \mathcal{T}Y' \preceq I_n.$$

This completes the proof. □

To obtain the results claimed in Theorem 5, we need to analyze the behavior of $\widehat{X}$ with respect to the constraints (QP–OC(a)) and (QP–OC(c)). Specifically, we would like to show that the event

$$\left\{\widehat{X} \bullet \mathcal{B}_i\widehat{X} \leq \Gamma^2 \text{ for all } i = 1, \ldots, L\right\} \bigcap \left\{\|\widehat{X}\|_\infty \leq \Gamma\right\}$$

occurs with constant probability, where $\Gamma = O\left(\sqrt{\ln \max\{m,n,L\}}\right)$. This can be done using the results in Section 2. First, let us tackle the norm constraint (QP–OC(c)). Observe that the matrix $\widehat{X}$ has the form $\sum_{k=1}^{m}\sum_{l=1}^{n}\xi_{kl}Q_{kl}$, where each Q_{kl} is an $m \times n$ matrix. Indeed, the (i,j)–th entry of Q_{kl} is simply $\left(Y'^{1/2}U^T\right)_{(i,j)(k,l)}$, where $i = 1,\ldots,m$ and $j = 1,\ldots,n$. Now, we compute

$$\mathbb{E}[\widehat{X}\widehat{X}^T] = \mathbb{E}\left[\sum_{k,k'=1}^{m}\sum_{l,l'=1}^{n}\xi_{kl}\xi_{k'l'}Q_{kl}Q_{k'l'}^T\right] = \sum_{k=1}^{m}\sum_{l=1}^{n}Q_{kl}Q_{kl}^T$$

and

$$\mathbb{E}[\widehat{X}^T\widehat{X}] = \mathbb{E}\left[\sum_{k,k'=1}^{m}\sum_{l,l'=1}^{n}\xi_{kl}\xi_{k'l'}Q_{kl}^TQ_{k'l'}\right] = \sum_{k=1}^{m}\sum_{l=1}^{n}Q_{kl}^TQ_{kl}.$$

By Proposition 1, we have

$$\sum_{k=1}^{m}\sum_{l=1}^{n}Q_{kl}Q_{kl}^T \preceq I_m \quad \text{and} \quad \sum_{k=1}^{m}\sum_{l=1}^{n}Q_{kl}^TQ_{kl} \preceq I_n.$$

Thus, by Theorem 4, we obtain the following result:

Proposition 2 *For any $\beta \geq 1/2$, we have*

$$\Pr\left(\|\widehat{X}\|_\infty \geq \sqrt{2e(1+\beta)\ln\max\{m,n\}}\right) \leq (\max\{m,n\})^{-\beta}.$$

Now, let us analyze the behavior of $\widehat{X}$ with respect to the constraint (QP–OC(a)). We have the following proposition:

Proposition 3 *For any $\beta \geq 1/2$, we have*

$$\Pr\left(\widehat{X} \bullet \mathcal{B}_i\widehat{X} \geq 2e(1+\beta)\ln\max\{m,n,L\}\right) \leq (\max\{m,n,L\})^{-(1+\beta)}$$

for $i = 1,\ldots,L$.

Proof For $i = 1,\ldots,L$, we have

$$\widehat{X} \bullet \mathcal{B}_i\widehat{X} = B_i \bullet Y'^{1/2}U^T\xi\xi^TUY'^{1/2} = \xi^TB_i'\xi,$$

where $B_i' = UY'^{1/2}B_iY'^{1/2}U^T \succeq \mathbf{0}$ because $B_i \succeq \mathbf{0}$. In particular, we may write $\widehat{X} \bullet \mathcal{B}_i\widehat{X} = \left\|(B_i')^{1/2}\xi\right\|_2^2$. Now, by Proposition 1 and a straightforward calculation, we have

$$1 \geq \mathbb{E}[\widehat{X} \bullet \mathcal{B}_i\widehat{X}] = \sum_{k=1}^{m}\sum_{l=1}^{n}(B_i')_{(k,l)(k,l)} = \sum_{k=1}^{m}\sum_{l=1}^{n}\left\|(B_i')^{1/2}_{(\cdot,\cdot)(k,l)}\right\|_2^2, \qquad (7)$$

where $(B_i')^{1/2}_{(\cdot,\cdot)(k,l)}$ is the (k,l)–th column of $(B_i')^{1/2}$. Since for any vector v and scalar $p \geq 2$, we have $\|v\|_{S_p} = \|v\|_2$, it follows from (7) and Theorem 2 that

$$\mathbb{E}\left[\left\|\sum_{k=1}^{m}\sum_{l=1}^{n}\xi_{kl}(B_i')^{1/2}_{(\cdot,\cdot)(k,l)}\right\|_2^p\right] \leq p^{p/2}\cdot\left(\sum_{k=1}^{m}\sum_{l=1}^{n}\left\|(B_i')^{1/2}_{(\cdot,\cdot)(k,l)}\right\|_2^2\right)^{p/2} \leq p^{p/2}. \tag{8}$$

In particular, by combining (8) with Markov's inequality, we conclude that for any $\beta \geq 1/2$,

$$\begin{aligned}
&\Pr\left(\widehat{X}\bullet\mathcal{B}_i\widehat{X} \geq 2e(1+\beta)\ln\max\{m,n,L\}\right)\\
&= \Pr\left(\left\|\sum_{k=1}^{m}\sum_{l=1}^{n}\xi_{kl}(B_i')^{1/2}_{(\cdot,\cdot)(k,l)}\right\|_2 \geq \sqrt{2e(1+\beta)\ln\max\{m,n,L\}}\right)\\
&\leq (\max\{m,n,L\})^{-(1+\beta)},
\end{aligned}$$

as desired. □

Remarks One of the referees has suggested a more elementary argument for proving Proposition 3. It relies on the following theorem, whose proof can be found in [21] (see inequality (1.9)):

Theorem 6 *Let $f:\mathbb{R}^h \to \mathbb{R}$ be a convex function satisfying*

$$|f(x)-f(y)| \leq L\cdot\|x-y\|_2 \qquad \textit{for all } x,y\in\mathbb{R}^h.$$

Let $\xi \in \{-1,+1\}^h$ be a vector of i.i.d. Bernoulli random variables, and let M_f be the median[2] *of $f(\xi)$. Then, for any $t>0$, we have*

$$\Pr\left(|f(\xi)-M_f| > t\right) \leq 4\exp\left(-t^2/8L^2\right).$$

To prove Proposition 3 (with slightly different constants) using Theorem 6, consider a fixed $i = 1,\ldots,L$, and let $f:\mathbb{R}^h \to \mathbb{R}$ be given by $f(x) = \left\|(B_i')^{1/2}x\right\|_2$. It is clear that f is convex, and for any $x,y\in\mathbb{R}^h$, we have

$$\begin{aligned}
|f(x)-f(y)| &\leq \left\|(B_i')^{1/2}(x-y)\right\|_2\\
&\leq \left\|(B_i')^{1/2}\right\|_\infty\cdot\|x-y\|_2\\
&\leq \left((B_i')^{1/2}\bullet(B_i')^{1/2}\right)\cdot\|x-y\|_2\\
&= (B_i\bullet Y')\cdot\|x-y\|_2\\
&\leq \|x-y\|_2.
\end{aligned}$$

[2] Recall that the median M_X of a real–valued random variable X is defined as $M_X = \sup\{t\in\mathbb{R} : \Pr(X\leq t)\leq 1/2\}$.

Now, since $f(\xi)$ is a non–negative random variable, we have $M_f \leq 2\mathbb{E}\left[f(\xi)\right]$ by Markov's inequality. Using (7) and Hölder's inequality, we compute

$$\mathbb{E}\left[f(\xi)\right] \leq \left(\mathbb{E}\left[f(\xi)^2\right]\right)^{1/2} \leq 1.$$

Hence, it follows that

$$\Pr\left(\left\|(B_i')^{1/2}\xi\right\|_2 \geq 2+t\right) \leq \Pr\left(\left\|(B_i')^{1/2}\xi\right\|_2 \geq M_f+t\right) \leq 4\exp\left(-t^2/8\right).$$

Upon setting $t = \sqrt{8(1+\beta)\ln\left(4\max\{m,n,L\}\right)}$, we conclude that for any $\beta \geq 1/2$,

$$\Pr\left(\widehat{X} \bullet \mathcal{B}_i\widehat{X} \geq 16(1+\beta)\ln\left(4\max\{m,n,L\}\right)\right) \leq \left(\max\{m,n,L\}\right)^{-(1+\beta)}.$$

Modulo the different constants, this establishes Proposition 3. □

We are now ready to finish the proof of Theorem 5.

Proof of Theorem 5 Let $\beta = 2$ in Propositions 2 and 3. Since $\max\{m,n\} \geq 2$ by assumption, we see from Propositions 1, 2 and 3 that with probability at least $1-(1/4+1/4) = 1/2$, the solution $\widehat{X}$ returned by the rounding scheme will satisfy

(a) $\widehat{X} \bullet \mathcal{A}\widehat{X} = \theta'$;
(b) $\widehat{X} \bullet \mathcal{B}_i\widehat{X} \leq 6e\ln\max\{m,n,L\}$ for $i = 1,\ldots,L$;
(c) $\mathcal{C}\widehat{X} = \mathbf{0}$; and
(d) $\|\widehat{X}\|_\infty \leq \sqrt{6e\ln\max\{m,n,L\}}$.

It follows that the matrix $\overline{X} = \widehat{X}/\sqrt{6e\ln\max\{m,n,L\}}$ has the required properties. □

3.2 Safe Tractable Approximations of Chance Constrained Linear Matrix Inequalities

Another of Nemirovski's original motivation for studying Question (Q) is to develop a so–called safe tractable approximation of the following chance constrained optimization problem:

$$\text{(CCP)} \qquad \begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & F(x) \leq \mathbf{0}, \\ & \Pr\left(\mathcal{A}_0(x) - \sum_{i=1}^{h}\xi_i\mathcal{A}_i(x) \succeq \mathbf{0}\right) \geq 1-\epsilon, \quad (\dagger) \\ & x \in \mathbb{R}^n. \end{array}$$

Here, $c \in \mathbb{R}^n$ is a given objective vector; $F : \mathbb{R}^n \to \mathbb{R}^l$ is an efficiently computable vector–valued function with convex components; $\mathcal{A}_0, \mathcal{A}_1, \ldots, \mathcal{A}_h : \mathbb{R}^n \to \mathscr{S}^m$ are affine functions in x with $\mathcal{A}_0(x) \succ \mathbf{0}$ for all $x \in \mathbb{R}^n$; $\xi_1, \ldots, \xi_h$

are independent (but not necessarily identical) mean zero random variables; and $\epsilon \in (0,1)$ is the error tolerance parameter. We assume that $m \geq 2$, so that (†) is indeed a chance constrained linear matrix inequality. The chance constrained problem (CCP) arises in many engineering applications, such as truss topology design, communications system design, and problems in control theory, and has received much attention lately (see, e.g., [27,28,25,9,37,22]). Unfortunately, the constraint (†) in (CCP) is generally intractable. In an attempt to circumvent this problem, Ben–Tal and Nemirovski [25,9] considered a safe tractable approximation of (†)—that is, a system of constraints $\mathcal{H}$ such that (i) x is feasible for (†) whenever it is feasible for $\mathcal{H}$, and (ii) the constraints in $\mathcal{H}$ are efficiently computable. Specifically, their strategy is as follows. First, observe that

$$\Pr\left(\mathcal{A}_0(x) - \sum_{i=1}^{h} \xi_i \mathcal{A}_i(x) \succeq \mathbf{0}\right) = \Pr\left(\sum_{i=1}^{h} \xi_i \mathcal{A}'_i(x) \preceq I\right), \tag{9}$$

where $\mathcal{A}'_i(x) = \mathcal{A}_0(x)^{-1/2}\mathcal{A}_i(x)\mathcal{A}_0(x)^{-1/2}$. Now, suppose that one can choose $\gamma = \gamma(\epsilon) > 0$ such that whenever

$$\sum_{i=1}^{h} \left(\mathcal{A}'_i(x)\right)^2 \preceq \gamma^2 I \tag{10}$$

holds, the constraint (†) is satisfied. Then, (10) will be a sufficient condition for (†) to hold. The upshot of (10) is that it can be expressed as a linear matrix inequality using the Schur complement (see [9]):

$$\begin{bmatrix} \gamma\mathcal{A}_0(x) & \mathcal{A}_1(x) & \cdots & \mathcal{A}_h(x) \\ \mathcal{A}_1(x) & \gamma\mathcal{A}_0(x) & & \\ \vdots & & \ddots & \\ \mathcal{A}_h(x) & & & \gamma\mathcal{A}_0(x) \end{bmatrix} \succeq \mathbf{0}. \tag{11}$$

Thus, by replacing (†) with (11), Problem (CCP) becomes tractable. Moreover, any solution $x \in \mathbb{R}^n$ that satisfies $F(x) \leq \mathbf{0}$ and (11) will be feasible for the original chance constrained problem (CCP).

Now, it is not hard to see that if the random variables $\xi_1, \ldots, \xi_h$ satisfy the conditions of Theorem 4 and if (10) holds for $\gamma \geq \gamma(\epsilon) \equiv \left(\sqrt{8e\ln(m/\epsilon)}\right)^{-1}$, then for any $\epsilon \in (0, 1/2]$, we have

$$\Pr\left(\sum_{i=1}^{h} \xi_i \mathcal{A}'_i(x) \preceq I\right) = \Pr\left(\left\|\sum_{i=1}^{h} \xi_i \mathcal{A}'_i(x)\right\|_\infty \leq 1\right) > 1 - \epsilon. \tag{12}$$

Indeed, observe that

$$\Pr\left(\left\|\sum_{i=1}^{h} \xi_i \mathcal{A}'_i(x)\right\|_\infty > 1\right) = \Pr\left(\left\|\sum_{i=1}^{h} \xi_i \left(\frac{1}{\gamma}\mathcal{A}'_i(x)\right)\right\|_\infty > \frac{1}{\gamma}\right).$$

Since $\sum_{i=1}^{h}(\gamma^{-1}\mathcal{A}_i'(x))^2 \preceq I$, we conclude by Theorem 3 and Markov's inequality that (cf. the proof of Theorem 4)

$$\Pr\left(\left\|\sum_{i=1}^{h}\xi_i\left(\frac{1}{\gamma}\mathcal{A}_i'(x)\right)\right\|_\infty > \frac{1}{\gamma}\right) \leq m\cdot\exp\left(-1/(8e\gamma^2)\right) \leq \epsilon.$$

This establishes (12), as desired. Now, upon recalling (9), we obtain the following theorem:

Theorem 7 *Let $\xi_1,\ldots,\xi_h$ be independent mean zero random variables, each of which is either (i) supported on $[-1,1]$, or (ii) normally distributed with unit variance. Consider the chance constrained problem (*Ccp*). Then, for any $\epsilon \in (0,1/2]$, the positive semidefinite constraint (11) with $\gamma \geq \gamma(\epsilon) \equiv \left(\sqrt{8e\ln(m/\epsilon)}\right)^{-1}$ is a safe tractable approximation of (†).*

Theorem 7 improves upon Nemirovski's result in [25], which requires $\gamma = \Omega\left(m^{1/6}+\sqrt{\ln(1/\epsilon)}\right)$ before one could assert that the constraint (11) is a safe tractable approximation of (†).

3.3 Data–Driven Distributionally Robust Stochastic Programming

In this section, we consider another class of optimization problems that involve data uncertainty and show how the moment inequalities introduced in Section 2 can be used to establish performance guarantees for them. To begin, let $\mathcal{X}\subset\mathbb{R}^n$ be a set of admissible actions, and let $(\Omega,\mathscr{B},\mathbb{P})$ be a probability space. Consider the following generic stochastic programming problem:

$$\text{(Sp)}\quad \text{minimize}\ \ \mathbb{E}_{\mathbb{P}}\left[f(x,\omega)\right]\ \ \text{subject to}\ \ x\in\mathcal{X}.$$

Here, the uncertain parameter $\omega\in\Omega$ is distributed according to $\mathbb{P}$, and $f:\mathcal{X}\times\Omega\to\mathbb{R}$ is the objective function. An assumption that is implicit in the formulation of (Sp) (and also (Ccp)) is that the distribution $\mathbb{P}$ of the uncertain parameter is fixed a priori. However, in practice, there could be uncertainty about the distribution as well. For instance, in the newsvendor problem, which is one of the simplest and most fundamental stochastic optimization problems, the newsvendor may only have information about the mean and variance of the customers' demand [35]. In order to guard against unpleasant surprises that may arise due to distributional uncertainty, one could consider the following so–called distributionally robust stochastic programming problem:

$$\text{(Drsp)}\quad \text{minimize}\ \ \max_{\mathbb{P}\in\mathcal{D}}\mathbb{E}_{\mathbb{P}}[f(x,\omega)]\ \ \text{subject to}\ \ x\in\mathcal{X}.$$

Here, $\mathcal{D}$ is the set of admissible distributions of the uncertain parameter, which is usually assumed to contain the true distribution. Many candidate distribution sets $\mathcal{D}$ have been proposed and studied in the literature; see, e.g., [12] and

the references therein. In particular, in a recent work, Delage and Ye [11] considered the case where $\mathcal{D}$ is the set of distributions on $\mathbb{R}^h$ whose mean vectors and covariance matrices are within a certain "distance" of the corresponding empirical estimates. Specifically, let $\mu_0 \in \mathbb{R}^h$ (resp. $\Sigma_0 \in \mathbb{R}^{h\times h}$) be an empirical estimate of the mean vector (resp. covariance matrix) of the underlying distribution, where it is assumed that $\Sigma_0 \succ \mathbf{0}$. Let $\gamma_m > 0$ and $\gamma_c \geq 1$ be parameters to be chosen, and let $S \subset \mathbb{R}^h$ be a closed convex set that is known to contain the support of the underlying distribution. Now, consider the set

$$\begin{aligned}\mathcal{D} &\equiv \mathcal{D}(S, \mu_0, \Sigma_0, \gamma_m, \gamma_c) \\ &= \left\{ \mathbb{P} \in \mathcal{M} : \begin{bmatrix} \mathbb{P}(\omega \in S) = 1 \\ (\mathbb{E}_{\mathbb{P}}[\omega] - \mu_0)^T \Sigma_0^{-1} (\mathbb{E}_{\mathbb{P}}[\omega] - \mu_0) \leq \gamma_m \\ \mathbb{E}_{\mathbb{P}}\left[(\omega - \mu_0)(\omega - \mu_0)^T\right] \preceq \gamma_c \Sigma_0 \end{bmatrix} \right\}, \end{aligned} \tag{13}$$

where $\mathcal{M}$ is the set of probability measures on the measurable space $(\mathbb{R}^h, \mathscr{B})$. An important advantage of using the set $\mathcal{D}$ is that under some mild assumptions, Problem (DRSP) can be formulated as a convex optimization problem. However, there is an undesirable feature in the definition of $\mathcal{D}$. Specifically, if the parameters $\gamma_m > 0$ and $\gamma_c \geq 1$ are not chosen carefully, then there is no guarantee that $\mathcal{D}$ will contain the true distribution. In particular, the optimization problem (DRSP) may have nothing to do with reality! Thus, it is natural to ask whether there is a rule for choosing $\gamma_m > 0$ and $\gamma_c \geq 1$ so that the set $\mathcal{D}$ defined in (13) will contain the true distribution with high probability. In [11], Delage and Ye showed that if the support of the true distribution $\mathbb{P}$ is bounded, and if one is able to generate independent samples according to $\mathbb{P}$, then the answer to this question is affirmative. Specifically, they proved the following theorem:

Theorem 8 *(Delage and Ye [11]) Let $\mathbb{P}$ be the true distribution of the uncertain parameter $\omega \in \mathbb{R}^h$. Let $\mu = \mathbb{E}_{\mathbb{P}}[\omega] \in \mathbb{R}^h$ and $\Sigma = \mathbb{E}_{\mathbb{P}}\left[(\omega - \mu)(\omega - \mu)^T\right] \in \mathbb{R}^{h\times h}$ be the true mean vector and covariance matrix of $\omega \in \mathbb{R}^h$, respectively. Suppose that $\Sigma \succ \mathbf{0}$, and that there exists an $R > 0$ such that*

$$\mathbb{P}\left((\omega - \mu)^T \Sigma^{-1} (\omega - \mu) \leq R^2\right) = 1. \tag{14}$$

Now, let $\delta \in (0, 1)$ be a confidence parameter, and let $\omega^1, \ldots, \omega^M \in \mathbb{R}^h$ be M independent samples generated according to $\mathbb{P}$, where

$$M > R^4 \left(\sqrt{1 - h/R^4} + \sqrt{\ln(2/\delta)}\right)^2.$$

Furthermore, let

$$\mu_0 = \frac{1}{M} \sum_{i=1}^{M} \omega_i \quad \text{and} \quad \Sigma_0 = \frac{1}{M} \sum_{i=1}^{M} (\omega_i - \mu_0)(\omega_i - \mu_0)^T$$

be the empirical estimates of the mean vector $\mu \in \mathbb{R}^h$ and covariance matrix $\Sigma \in \mathbb{R}^{h\times h}$, respectively. Then, with probability at least $1-\delta$ (over the choices of $\omega^1, \ldots, \omega^M$), the following constraints will be satisfied:

$$(\mu_0 - \mu)\Sigma^{-1}(\mu_0 - \mu) \leq \frac{R^2}{M}\left(2 + \sqrt{2\ln(2/\delta)}\right)^2,$$

$$\Sigma \preceq \left[1 - \frac{R^2}{\sqrt{M}}\left(\sqrt{1 - h/R^4} + \sqrt{\ln(4/\delta)}\right) - \frac{R^2}{M}\left(2 + \sqrt{2\ln(2/\delta)}\right)^2\right]^{-1}\Sigma_0,$$

$$\Sigma_0 \preceq \left[1 + \frac{R^2}{\sqrt{M}}\left(\sqrt{1 - h/R^4} + \sqrt{\ln(4/\delta)}\right)\right]\Sigma.$$

Moreover, there exist $\gamma_m > 0$ and $\gamma_c \geq 1$ such that with probability at least $1-\delta$, the true distribution $\mathbb{P}$ will belong to the set $\mathcal{D}$ defined in (13), where $S = \{\omega \in \mathbb{R}^h : (\omega - \mu)^T\Sigma^{-1}(\omega - \mu) \leq R^2\}$.

As it turns out, Theorem 8 can be extended to the case where the true distribution does not have bounded support, but rather satisfies a (weaker) moment growth condition. One way of proving such an extension is to use the moment inequalities introduced in Section 2. Before we present the proof, let us be more precise about the assumption we make on the true distribution $\mathbb{P}$:

Condition (G) *The true distribution $\mathbb{P}$ of the uncertain parameter $\omega \in \mathbb{R}^h$ satisfies the following moment growth condition: There exists an absolute constant $c > 0$ such that for any $p \geq 1$, the following holds:*

$$\mathbb{E}_{\mathbb{P}}\left[\left\|\Sigma^{-1/2}(\omega - \mu)\right\|_2^p\right] \leq (cp)^{p/2}.$$

Clearly, if $\mathbb{P}$ satisfies the bounded–support condition (14), then it also satisfies Condition (G) with $c \leq R$. However, the converse is not true. For instance, suppose that $\mathbb{P}$ is the standard one–dimensional normal distribution. Then, we have $\mu = 0$ and $\Sigma = 1$. It is clear that $\mathbb{P}(\omega^2 \leq R^2) < 1$ for any $R > 0$. On the other hand, for any $p \geq 1$, we have

$$\mathbb{E}_{\mathbb{P}}\left[|\omega|^p\right] < p^{p/2}.$$

Thus, Condition (G) is strictly weaker than the bounded–support condition (14).

Now, let us prove the following result concerning the empirical estimate μ_0 of the true mean vector μ. It extends [11, Corollary 1] to the case where the true distribution $\mathbb{P}$ may have unbounded support but satisfies the moment growth condition (G).

Proposition 4 *Suppose that the true distribution $\mathbb{P}$ of the uncertain parameter $\omega \in \mathbb{R}^h$ satisfies Condition (G). Let $M \geq 1$ be an integer, and let $\delta \in (0, e^{-2})$ be a confidence parameter. Consider M independent samples*

$\omega^1, \ldots, \omega^M \in \mathbb{R}^h$ *generated according to* $\mathbb{P}$. *Then, with probability at least* $1-\delta$, *we will have*

$$(\mu_0 - \mu)^T \Sigma^{-1} (\mu_0 - \mu) \leq \frac{4ce^2 \ln^2(1/\delta)}{M},$$

where $\mu_0 = M^{-1} \sum_{i=1}^M \omega_i$.

Proof Define $\zeta_i = \Sigma^{-1/2}(\omega_i - \mu)$ for $i = 1, \ldots, M$, and let $\epsilon_1, \ldots, \epsilon_M$ be i.i.d. Bernoulli random variables. Since $\mathbb{E}_{\mathbb{P}}[\zeta_i] = \mathbf{0}$ for $i = 1, \ldots, M$, by the convexity of $x \mapsto |x|^p$ on $\mathbb{R}_+$ for any $p \geq 1$ and a standard symmetrization argument (see, e.g., [21, Lemma 6.3]), we have

$$\mathbb{E}_{\mathbb{P}} \left[\left\| \sum_{i=1}^M \zeta_i \right\|_2^p \right] \leq 2^p \cdot \mathbb{E} \left[\left\| \sum_{i=1}^M \epsilon_i \zeta_i \right\|_2^p \right]$$

for any $p \geq 1$. Now, by Theorem 2 and Jensen's inequality, conditioned on the ζ_i's, we have

$$\mathbb{E}_{\epsilon} \left[\left\| \sum_{i=1}^M \epsilon_i \zeta_i \right\|_2^p \right] < p^{p/2} \cdot \left(\sum_{i=1}^M \|\zeta_i\|_2^2 \right)^{p/2} \leq M^{p/2-1} \cdot p^{p/2} \cdot \left(\sum_{i=1}^M \|\zeta_i\|_2^p \right)$$

for any $p \geq 2$. Hence, it follows from Fubini's theorem and Condition (G) that for any $p \geq 2$,

$$\mathbb{E}_{\mathbb{P}} \left[\left\| \sum_{i=1}^M \zeta_i \right\|_2^p \right] \leq 2^p \cdot M^{p/2-1} \cdot p^{p/2} \cdot \left(\sum_{i=1}^M \mathbb{E}_{\mathbb{P}} \left[\|\zeta_i\|_2^p \right] \right) \leq 2^p \cdot (cM)^{p/2} \cdot p^p.$$

Now, by Markov's inequality, for any $t > 0$ and $p \geq 2$, we compute

$$\mathbb{P} \left(\left\| \frac{1}{M} \sum_{i=1}^M \zeta_i \right\|_2 > t \right) \leq t^{-p} \cdot M^{-p} \cdot \mathbb{E}_{\mathbb{P}} \left[\left\| \sum_{i=1}^M \zeta_i \right\|_2^p \right] \leq \frac{2^p \cdot c^{p/2} \cdot p^p}{t^p \cdot M^{p/2}}.$$

Upon setting $t = e\sqrt{4c/M} \ln(1/\delta)$ and $p = t/e\sqrt{4c/M} \geq 2$ (since we have $\delta \leq e^{-2}$ by assumption), we conclude that

$$\begin{aligned} &\mathbb{P} \left((\mu_0 - \mu)^T \Sigma^{-1} (\mu_0 - \mu) > \frac{4ce^2 \ln^2(1/\delta)}{M} \right) \\ &= \mathbb{P} \left(\left\| \frac{1}{M} \sum_{i=1}^M \zeta_i \right\|_2^2 > \frac{4ce^2 \ln^2(1/\delta)}{M} \right) \\ &\leq \delta, \end{aligned}$$

as desired. □

Next, we establish a relationship between the matrix $\tilde{\Sigma} = M^{-1} \sum_{i=1}^M (\omega_i - \mu)(\omega_i - \mu)^T$ and the true covariance matrix Σ. Note that $\tilde{\Sigma}$ is not the empirical

estimate of Σ, as it is defined in terms of the true mean vector μ. Nevertheless, as we shall see in Theorem 9, such a relationship will enable us to derive bounds on the empirical estimate Σ_0 of the true covariance matrix.

Proposition 5 *Suppose that the true distribution $\mathbb{P}$ of the uncertain parameter $\omega \in \mathbb{R}^h$ satisfies Condition (G). Let $M \geq 1$ be an integer, and let $\delta \in (0, 2he^{-3})$ be a confidence parameter. Consider M independent samples $\omega^1, \ldots, \omega^M \in \mathbb{R}^h$ generated according to $\mathbb{P}$, where*

$$M > 16c'^2(2e/3)^3 \ln^3(2h/\delta)$$

and $c' = \max\{c, 1\}$. Then, with probability at least $1 - \delta$, we will have

$$(1-t)\Sigma \preceq \tilde{\Sigma} \preceq (1+t)\Sigma,$$

where $\tilde{\Sigma} = M^{-1}\sum_{i=1}^{M}(\omega_i - \mu)(\omega_i - \mu)^T$ and

$$t = \frac{4c'(2e/3)^{3/2}\ln^{3/2}(2h/\delta)}{\sqrt{M}}.$$

Proof As in the proof of Proposition 4, define $\zeta_i = \Sigma^{-1/2}(\omega_i - \mu)$ for $i = 1, \ldots, M$, and let $\epsilon_1, \ldots, \epsilon_M$ be i.i.d. Bernoulli random variables. Consider the matrices $Q_i = \zeta_i\zeta_i^T - I \in \mathbb{R}^{h\times h}$, where $i = 1, \ldots, M$. Since

$$\mathbb{E}_{\mathbb{P}}\left[\zeta_i\zeta_i^T\right] = \Sigma^{-1/2}\mathbb{E}_{\mathbb{P}}\left[(\omega_i - \mu)(\omega_i - \mu)^T\right]\Sigma^{-1/2} = I,$$

we have $\mathbb{E}_{\mathbb{P}}[Q_i] = \mathbf{0}$, where $i = 1, \ldots, M$. Thus, a standard symmetrization argument yields

$$\mathbb{E}_{\mathbb{P}}\left[\left\|\sum_{i=1}^{M} Q_i\right\|_{S_p}^p\right] \leq 2^p \cdot \mathbb{E}\left[\left\|\sum_{i=1}^{M}\epsilon_i Q_i\right\|_{S_p}^p\right] \tag{15}$$

for any $p \geq 1$. Now, by Theorem 2 and Jensen's inequality, conditioned on the ζ_i's, we have

$$\mathbb{E}_{\epsilon}\left[\left\|\sum_{i=1}^{M} Q_i\right\|_{S_p}^p\right] < p^{p/2}\cdot\left(\sum_{i=1}^{M}\|Q_i\|_{S_p}^2\right)^{p/2} \leq M^{p/2-1}\cdot p^{p/2}\cdot\left(\sum_{i=1}^{M}\|Q_i\|_{S_p}^p\right) \tag{16}$$

for any $p \geq 2$. Note that for $i = 1, \ldots, M$, the eigenvalues of Q_i are -1 (with multiplicity $h-1$) and $\|\zeta_i\|_2^2 - 1$ (with multiplicity 1). Hence, it follows that

$$\|Q_i\|_{S_p}^p = h - 1 + \left|\|\zeta_i\|_2^2 - 1\right|^p. \tag{17}$$

Upon substituting (16), (17) into (15) and noting that the ζ_i's are identically distributed, we have

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}\left[\left\|\sum_{i=1}^{M} Q_i\right\|_{S_p}^p\right] &\leq 2^p \cdot M^{p/2}\cdot p^{p/2}\cdot\mathbb{E}_{\mathbb{P}}\left[h - 1 + \left|\|\zeta_1\|_2^2 - 1\right|^p\right] \\
&\leq 2^p \cdot M^{p/2}\cdot p^{p/2}\cdot\left(h + \mathbb{E}_{\mathbb{P}}\left[\|\zeta_1\|_2^{2p}\right]\right) \qquad (18)\\
&\leq 2^p \cdot M^{p/2}\cdot p^{p/2}\cdot\left(h + (2cp)^p\right), \qquad (19)
\end{aligned}$$

where (18) follows from the fact that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}\left[\left|\|\zeta_1\|_2^2-1\right|^p\right] &= \int_{\{\|\zeta_1\|_2^2<1/2\}}\left|\|\zeta_1\|_2^2-1\right|^p\,d\mathbb{P}\\ &\quad+\int_{\{\|\zeta_1\|_2^2\geq 1/2\}}\left|\|\zeta_1\|_2^2-1\right|^p\,d\mathbb{P}\\ &\leq 1+\int_{\mathbb{R}^h}\|\zeta_1\|_2^{2p}\,d\mathbb{P}\\ &= 1+\mathbb{E}_{\mathbb{P}}\left[\|\zeta_1\|_2^{2p}\right],\end{aligned}$$

and (19) follows from Condition (G). Thus, by Markov's inequality, for any $t>0$ and $p\geq 2$, we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^M Q_i\right\|_\infty > t\right) &\leq t^{-p}\cdot M^{-p}\cdot\mathbb{E}_{\mathbb{P}}\left[\left\|\sum_{i=1}^M Q_i\right\|_{S_p}^p\right]\\ &\leq \frac{2^p\cdot p^{p/2}\cdot(h+(2cp)^p)}{t^p\cdot M^{p/2}}.\end{aligned}$$

Now, define $c'=\max\{c,1\}$, and set

$$t=\frac{4c'(2e/3)^{3/2}\ln^{3/2}(2h/\delta)}{\sqrt{M}}\quad\text{and}\quad p=\left(\frac{t}{4c'e^{3/2}/\sqrt{M}}\right)^{2/3}. \tag{20}$$

Note that $p=2\ln(2h/\delta)/3\geq 2$, since $\delta\leq 2he^{-3}$ by assumption. Moreover, we have

$$\frac{2^p\cdot p^{p/2}\cdot h}{t^p\cdot M^{p/2}}=\frac{2^p\cdot p^{p/2}\cdot h}{(4c')^p\cdot e^{3p/2}\cdot p^{3p/2}}\leq\frac{\delta}{2}$$

and

$$\frac{2^p\cdot p^{p/2}\cdot(2cp)^p}{t^p\cdot M^{p/2}}=\frac{c^p}{c'^p\cdot e^{3p/2}}\leq\frac{\delta}{2h}\leq\frac{\delta}{2}.$$

It follows that

$$\mathbb{P}\left(\left\|\frac{1}{M}\sum_{i=1}^M Q_i\right\|_\infty > \frac{4c'(2e/3)^{3/2}\ln^{3/2}(2h/\delta)}{\sqrt{M}}\right)\leq\delta. \tag{21}$$

Now, let $\lambda_1,\ldots,\lambda_h$ be the eigenvalues of $\tilde{\Sigma}'=M^{-1}\sum_{i=1}^M\zeta_i\zeta_i^T$. Then, (21) implies that with probability at least $1-\delta$, we will have $|1-\lambda_i|\leq t$ for $i=1,\ldots,h$, where t is given by (20). In particular, whenever

$$M>16c'^2(2e/3)^3\ln^3(2h/\delta)$$

so that $t < 1$, all the eigenvalues of $\tilde{\Sigma}'$ will lie between $[1-t, 1+t]$ with probability at least $1-\delta$. In other words, the linear matrix inequalities

$$(1-t)I \preceq \tilde{\Sigma}' \preceq (1+t)I$$

will hold with probability at least $1-\delta$. Upon noting that $\tilde{\Sigma} = \Sigma^{1/2}\tilde{\Sigma}'\Sigma^{1/2}$, the proof is completed. □

Recall that our original goal is to prove that with high probability, the true distribution $\mathbb{P}$ will belong to the distribution set $\mathcal{D}$ defined in (13). This can now be achieved by combining the results of Propositions 4, 5 and the arguments in [11]. Specifically, we prove the following theorem, which extends the corresponding result in [11] (see Theorem 8 of this paper) to the case where the true distribution $\mathbb{P}$ only satisifes the moment growth condition (G):

Theorem 9 *Suppose that the true distribution $\mathbb{P}$ of the uncertain parameter $\omega \in \mathbb{R}^h$ satisfies Condition (G). Let*

$$\mu = \mathbb{E}_{\mathbb{P}}[\omega] \in \mathbb{R}^h \quad \textit{and} \quad \Sigma = \mathbb{E}_{\mathbb{P}}\left[(\omega-\mu)(\omega-\mu)^T\right] \in \mathbb{R}^{h\times h}$$

be the true mean vector and covariance matrix of $\omega \in \mathbb{R}^h$, respectively, with $\Sigma \succ \mathbf{0}$. Let $\delta \in (0, 2e^{-3})$ be a confidence parameter, and let $\omega^1, \ldots, \omega^M \in \mathbb{R}^h$ be M independent samples generated according to $\mathbb{P}$, where

$$M > 32c'^2(2e/3)^3 \ln^3(4h/\delta)$$

and $c' = \max\{c, 1\}$. Let

$$\mu_0 = \frac{1}{M}\sum_{i=1}^{M}\omega_i \quad \textit{and} \quad \Sigma_0 = \frac{1}{M}\sum_{i=1}^{M}(\omega_i-\mu_0)(\omega_i-\mu_0)^T$$

be the empirical estimates of the mean vector $\mu \in \mathbb{R}^h$ and covariance matrix $\Sigma \in \mathbb{R}^{h\times h}$, respectively, and define

$$t_m = \frac{4ce^2\ln^2(2/\delta)}{M} \quad \textit{and} \quad t_c = \frac{4c'(2e/3)^{3/2}\ln^{3/2}(4h/\delta)}{\sqrt{M}}.$$

Then, with probability at least $1-\delta$ (over the choices of $\omega^1, \ldots, \omega^M$), the following constraints will be satisfied:

$$(\mu_0-\mu)^T\Sigma_0^{-1}(\mu_0-\mu) \le \frac{t_m}{1-t_c-t_m} \equiv \gamma_m,$$

$$\mathbb{E}_{\mathbb{P}}\left[(\omega-\mu_0)(\omega-\mu_0)^T\right] \preceq \frac{1+t_m}{1-t_c-t_m}\Sigma_0 \equiv \gamma_c\Sigma_0.$$

In particular, we have $\mathbb{P} \in \mathcal{D}(\mathbb{R}^h, \mu_0, \Sigma_0, \gamma_m, \gamma_c)$.

Proof Note that the condition on M ensures that $t_c + t_m < 1$. Now, by Propositions 4 and 5, with probability at least $1 - \delta$, the following constraints will be satisfied:

$$(\mu_0 - \mu)^T \Sigma^{-1} (\mu_0 - \mu) \leq t_m, \tag{22}$$

$$(1 - t_c)\Sigma \preceq \tilde{\Sigma} \preceq (1 + t_c)\Sigma. \tag{23}$$

In particular, using the argument in the proof of [11, Theorem 2], one can derive from (22) and (23) that

$$(1 - t_c)\Sigma \preceq \tilde{\Sigma} \preceq \Sigma_0 + t_m \Sigma.$$

This implies that

$$(\mu_0 - \mu)^T \Sigma_0^{-1} (\mu_0 - \mu) \leq \frac{1}{1 - t_c - t_m} (\mu_0 - \mu)^T \Sigma^{-1} (\mu_0 - \mu) \leq \frac{t_m}{1 - t_c - t_m} = \gamma_m. \tag{24}$$

Now, using (24) and the argument in the proof of [11, Corollary 4], we have

$$\mathbb{E}_{\mathbb{P}} \left[(\omega - \mu_0)(\omega - \mu_0)^T \right] - \gamma_m \Sigma_0 \preceq \frac{1}{1 - t_c - t_m} \Sigma_0,$$

which yields

$$\mathbb{E}_{\mathbb{P}} \left[(\omega - \mu_0)(\omega - \mu_0)^T \right] \preceq \frac{1 + t_m}{1 - t_c - t_m} \Sigma_0 = \gamma_c \Sigma_0.$$

This completes the proof of Theorem 9. □

4 Conclusion

In this paper, we explored the close connection between moment inequalities for sums of certain random matrices and the performance analyses of several problems in optimization. As a result, we obtained the best known performance guarantees for several optimization problems. Given the power and wide applicability of the moment inequalities considered in this paper, it would be interesting to find other problems for which they apply.

Acknowledgements This research is supported by CUHK Direct Grant No. 2050401. The author would like to express his gratitude to the referees, whose detailed comments greatly improve the presentation of the paper.

References

1. Ai, W., Zhang, S.: Strong Duality for the CDT Subproblem: A Necessary and Sufficient Condition. SIAM Journal on Optimization **19**(4), 1735–1756 (2009)
2. Alon, N., Makarychev, K., Makarychev, Y., Naor, A.: Quadratic Forms on Graphs. Inventiones Mathematicae **163**(3), 499–522 (2006)
3. Anstreicher, K., Chen, X., Wolkowicz, H., Yuan, Y.X.: Strong Duality for a Trust–Region Type Relaxation of the Quadratic Assignment Problem. Linear Algebra and Its Applications **301**(1–3), 121–136 (1999)
4. Anstreicher, K., Wolkowicz, H.: On Lagrangian Relaxation of Quadratic Matrix Constraints. SIAM Journal on Matrix Analysis and Applications **22**(1), 41–55 (2000)
5. Arora, S., Lee, J.R., Naor, A.: Euclidean Distortion and the Sparsest Cut. Journal of the American Mathematical Society **21**(1), 1–21 (2008)
6. Arora, S., Rao, S., Vazirani, U.: Expander Flows, Geometric Embeddings and Graph Partitioning. Journal of the ACM **56**(2), Article 5 (2009)
7. Barvinok, A.I.: Problems of Distance Geometry and Convex Properties of Quadratic Maps. Discrete and Computational Geometry **13**, 189–202 (1995)
8. Beck, A., Eldar, Y.C.: Strong Duality in Nonconvex Quadratic Optimization with Two Quadratic Constraints. SIAM Journal on Optimization **17**(3), 844–860 (2006)
9. Ben-Tal, A., Nemirovski, A.: On Safe Tractable Approximations of Chance–Constrained Linear Matrix Inequalities. Mathematics of Operations Research **34**(1), 1–25 (2009)
10. Buchholz, A.: Operator Khintchine Inequality in Non–Commutative Probability. Mathematische Annalen **319**, 1–16 (2001)
11. Delage, E., Ye, Y.: Distributionally Robust Optimization under Moment Uncertainty with Application to Data–Driven Problems (2009). To appear in *Operations Research*
12. Dupačová, J.: Stochastic Programming: Minimax Approach. In: C.A. Floudas, P.M. Pardalos (eds.) Encyclopedia of Optimization, second edn., pp. 3778–3782. Springer Science+Business Media, LLC, New York (2009)
13. Goemans, M.X.: Semidefinite Programming in Combinatorial Optimization. Mathematical Programming **79**, 143–161 (1997)
14. Goemans, M.X., Williamson, D.P.: Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. Journal of the ACM **42**(6), 1115–1145 (1995)
15. Gower, J.C., Dijksterhuis, G.B.: Procrustes Problems, *Oxford Statistical Science Series*, vol. 30. Oxford University Press, New York (2004)
16. Grötschel, M., Lovász, L., Schrijver, A.: Geometric Algorithms and Combinatorial Optimization, *Algorithms and Combinatorics*, vol. 2, second corrected edn. Springer–Verlag, Berlin Heidelberg (1993)
17. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
18. Karger, D., Motwani, R., Sudan, M.: Approximate Graph Coloring by Semidefinite Programming. Journal of the ACM **45**(2), 246–265 (1998)
19. Khintchine, A.: Über Dyadische Brüche. Mathematische Zeitschrift **23**, 109–116 (1923)
20. Koopmans, T.C., Beckmann, M.: Assignment Problems and the Location of Economic Activities. Econometrica **25**(1), 53–76 (1957)
21. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes, *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*, vol. 23. Springer–Verlag, Berlin Heidelberg (1991)
22. Li, W.L., Zhang, Y.J., So, A.M.C., Win, M.Z.: Slow Adaptive OFDMA through Chance Constrained Programming (2009). Preprint
23. Luo, Z.Q., Sidiropoulos, N.D., Tseng, P., Zhang, S.: Approximation Bounds for Quadratic Optimization with Homogeneous Quadratic Constraints. SIAM Journal on Optimization **18**(1), 1–28 (2007)
24. Lust-Piquard, F.: Inégalités de Khintchine dans C_p $(1 < p < \infty)$. Comptes Rendus de l'Académie des Sciences de Paris, Série I **303**(7), 289–292 (1986)
25. Nemirovski, A.: Sums of Random Symmetric Matrices and Quadratic Optimization under Orthogonality Constraints. Mathematical Programming, Series B **109**(2–3), 283–317 (2007)

26. Nemirovski, A., Roos, C., Terlaky, T.: On Maximization of Quadratic Form over Intersection of Ellipsoids with Common Center. Mathematical Programming, Series A **86**, 463–473 (1999)
27. Nemirovski, A., Shapiro, A.: Convex Approximations of Chance Constrained Programs. SIAM Journal on Optimization **17**(4), 969–996 (2006)
28. Nemirovski, A., Shapiro, A.: Scenario Approximations of Chance Constraints. In: G. Calafiore, F. Dabbene (eds.) Probabilistic and Randomized Methods for Design under Uncertainty, pp. 3–47. Springer–Verlag, London (2006)
29. Nesterov, Yu.: Quality of Semidefinite Relaxation for Nonconvex Quadratic Optimization. CORE Discussion Paper 9719, Université Catholique de Louvain, Belgium (1997)
30. Pardalos, P.M., Wolkowicz, H. (eds.): Quadratic Assignment and Related Problems, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 16. American Mathematical Society, Providence, Rhode Island (1994)
31. Pataki, G.: On the Rank of Extreme Matrices in Semidefinite Programs and the Multiplicity of Optimal Eigenvalues. Mathematics of Operations Research **23**(2), 339–358 (1998)
32. Peshkir, G., Shiryaev, A.N.: The Khintchine Inequalities and Martingale Expanding Sphere of Their Action. Russian Mathematical Surveys **50**(5), 849–904 (1995)
33. Pisier, G.: Non–Commutative Vector Valued L_p–Spaces and Completely p–Summing Maps. Astérisque **247** (1998)
34. Quine, M.P.: A Calculus–Based Proof of a Stirling Formula for the Gamma Function. International Journal of Mathematical Education in Science and Technology **28**(6), 914–917 (1997)
35. Scarf, H.: A Min–Max Solution of an Inventory Problem. In: K.J. Arrow, S. Karlin, H. Scarf (eds.) Studies in the Mathematical Theory of Inventory and Production, pp. 201–209. Stanford University Press, Stanford (1958)
36. Shapiro, A.: Rank–Reducibility of a Symmetric Matrix and Sampling Theory of Minimum Trace Factor Analysis. Psychometrika **47**(2), 187–199 (1982)
37. Shenouda, M.B., Davidson, T.N.: Outage–Based Designs for Multi–User Transceivers. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), pp. 2389–2392 (2009)
38. So, A.M.C.: On the Performance of Semidefinite Relaxation MIMO Detectors for QAM Constellations. In: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009), pp. 2449–2452 (2009)
39. So, A.M.C.: Probabilistic Analysis of the Semidefinite Relaxation Detector in Digital Communications (2010). To appear in *Proceedings of the 21st Annual ACM–SIAM Symposium on Discrete Algorithms (SODA 2010)*
40. So, A.M.C., Ye, Y., Zhang, J.: A Unified Theorem on SDP Rank Reduction. Mathematics of Operations Research **33**(4), 910–920 (2008)
41. So, A.M.C., Zhang, J., Ye, Y.: On Approximating Complex Quadratic Optimization Problems via Semidefinite Programming Relaxations. Mathematical Programming, Series B **110**(1), 93–110 (2007)
42. Tomczak-Jaegermann, N.: The Moduli of Smoothness and Convexity and the Rademacher Averages of Trace Classes S_p ($1 \leq p < \infty$). Studia Mathematica **50**, 163–182 (1974)
43. Tropp, J.A.: The Random Paving Property for Uniformly Bounded Matrices. Studia Mathematica **185**(1), 67–82 (2008)
44. Ye, Y.: Approximating Global Quadratic Optimization with Convex Quadratic Constraints. Journal of Global Optimization **15**(1), 1–17 (1999)
45. Ye, Y., Zhang, S.: New Results on Quadratic Minimization. SIAM Journal on Optimization **14**(1), 245–267 (2003)
46. Zhao, Q., Karisch, S.E., Rendl, F., Wolkowicz, H.: Semidefinite Programming Relaxations for the Quadratic Assignment Problem. Journal of Combinatorial Optimization **2**(1), 71–109 (1998)