ANNUAL REVIEWS

Annual Review of Control, Robotics, and Autonomous Systems

Recent Scalability Improvements for Semidefinite Programming with Applications in Machine Learning, Control, and Robotics

Anirudha Majumdar,[1] Georgina Hall,[2] and Amir Ali Ahmadi[3]

[1]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, New Jersey 08544, USA; email: ani.majumdar@princeton.edu

[2]Decision Sciences, INSEAD, 77300 Fontainebleau, France; email: georgina.hall@insead.edu

[3]Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, USA; email: aaa@princeton.edu

Annu. Rev. Control Robot. Auton. Syst. 2020. 3:331–60

First published as a Review in Advance on December 24, 2019

The *Annual Review of Control, Robotics, and Autonomous Systems* is online at control.annualreviews.org

https://doi.org/10.1146/annurev-control-091819-074326

Keywords

semidefinite programming, sum of squares programming, machine learning, control, robotics

Abstract

Historically, scalability has been a major challenge for the successful application of semidefinite programming in fields such as machine learning, control, and robotics. In this article, we survey recent approaches to this challenge, including those that exploit structure (e.g., sparsity and symmetry) in a problem, those that produce low-rank approximate solutions to semidefinite programs, those that use more scalable algorithms that rely on augmented Lagrangian techniques and the alternating-direction method of multipliers, and those that trade off scalability with conservatism (e.g., by approximating semidefinite programs with linear and second-order cone programs). For each class of approaches, we provide a high-level exposition, an entry point to the corresponding literature, and examples drawn from machine learning, control, or robotics. We also present a list of software packages that implement many of the techniques discussed in the review. Our hope is that this article will serve as a gateway to the rich and exciting literature on scalable semidefinite programming for both theorists and practitioners.

1. INTRODUCTION

A semidefinite program (SDP) is an optimization problem of the form

$$\begin{array}{ll} \min\limits_{X \in S_n} & \mathrm{Tr}(CX) \\ \text{subject to (s.t.)} & \mathrm{Tr}(A_{iX}) = b_i, \quad i = 1, \ldots, m, \\ & X \succeq 0, \end{array} \tag{1.}$$

where S_n denotes the set of $n \times n$ real symmetric matrices; Tr(.) stands for the trace of a matrix; $C, A_1, \ldots, A_m \in S_n$ and $b_1, \ldots, b_m \in \mathbb{R}$ are input data to the problem; and the notation $X \succeq 0$ denotes that the matrix X is constrained to be positive semidefinite (psd)—i.e., to have nonnegative eigenvalues. The dual to Problem 1 takes the form

$$\begin{array}{ll} \max\limits_{y \in \mathbb{R}^m} & \sum\limits_{i=1}^{m} y_i b_i \\ \text{s.t.} & C - \sum\limits_{i=1}^{m} y_i A_i \succeq 0. \end{array} \tag{2.}$$

SDPs have numerous fundamental applications throughout applied and computational mathematics. In combinatorics and discrete optimization, for instance, some of the most powerful relaxations for prominent problems of the field rely on semidefinite programming. For example, the current best approximation algorithm for the maximum-cut problem (1) and the only known polynomial-time algorithm for finding maximum stable sets in perfect graphs (2) make fundamental use of semidefinite programming. In machine learning, one is often faced with the task of learning a signal with a known structure from data (e.g., a low-rank matrix, a sparse eigenvector, or a monotone polynomial). The design of an appropriate convex optimization problem that induces such structure in a signal often leads to a semidefinite programming problem (see, e.g., 3; 4; 5, chapter 5). In polynomial optimization, SDP-based hierarchies based on the notion of sum of squares (sos) polynomials (see Section 1.4) are among the most promising approaches for obtaining global solutions in the absence of any convexity assumptions (6, 7). Finally, in control and robotics, various notions of the safety, performance, stability, and robustness of dynamical systems can be certified by searching for Lyapunov functions that satisfy appropriate nonnegativity constraints over the state space. When the candidate functions are parameterized as polynomials or other semialgebraic functions, this search can be automated by semidefinite programming via the SOS[1] approach (see 9 and Section 1.4 of this article). Overall, semidefinite programming has proven to be one of the most expressive classes of convex optimization problems, subsuming many other important classes, such as linear programming, convex quadratic programming, and second-order cone programming (10).

In addition to its expressiveness, another contributing factor to the popularity of semidefinite programming lies in its strong theoretical and computational properties, which to a certain extent parallel those of linear programming (10). For instance, strong duality between Problems 1 and 2 holds under mild assumptions. Moreover, many algorithms for linear programs, such as the ellipsoid method and interior-point methods, can and have been adapted to SDPs. While these

[1]We follow the convention introduced in Reference 8 of using lowercase abbreviations to refer to polynomials or matrices (when appropriate) and capitalized abbreviations to refer to cones of polynomials or matrices and optimization over said cones.

algorithms solve SDPs to arbitrary accuracy in polynomial time, in practice, they suffer from one serious impediment: scalability. When the problems considered have large n or m (as defined in Problem 1), solving-time and memory requirements tend to explode, making SDPs prohibitive to work with. Indeed, most solvers rely on interior-point methods as their algorithm of choice for SDPs, and interior-point methods produce iterates that are (as their name suggests) in the interior of the feasible set. Computing these large, dense matrix iterates and storing them leads to the issues described above. Devising methods to make SDPs more scalable is consequently an important and active area of research.

1.1. The Goal and Outline of This Survey

In this article, we review recent literature on scalability improvements for semidefinite programming, with a particular focus on methods that have proved useful in machine learning, control, and robotics. We consider these representative application areas since they are timely and because many problems that arise in them have straightforward semidefinite programming–based relaxations or reformulations but involve very large problem instances. This article is a follow-up to a workshop at the 2016 IEEE Conference on Decision and Control on the same topic, titled "Solving Large SDPs in Control, Machine Learning, and Robotics"; abstracts and slides for the 11 lectures given at this workshop are available at **http://aaa.princeton.edu/largesdps**.

The outline of this article is as follows. In Section 1.4, we provide a brief introduction to SOS proofs of nonnegativity of polynomials. This background material is relevant to many of the applications of semidefinite programming we discuss (e.g., in Sections 2.1.1 and 5.1). Section 2 presents approaches that enhance scalability by exploiting structure (e.g., symmetry or sparsity) in a problem. Section 3 presents methods for generating low-rank solutions to SDPs in order to reduce computation-time and storage-space requirements. Section 4 reviews more scalable alternatives to interior-point methods for solving SDPs based on augmented Lagrangian techniques and the alternating-direction method of multipliers. Section 5 presents approaches that trade off scalability with conservatism of solutions (e.g., by approximating SDPs with linear or second-order cone programs). In Section 6, we present a list of software packages that implement many of the techniques discussed. Section 7 concludes the article and highlights promising directions for future work.

1.2. Relationship Between Approaches Discussed in the Article

The approaches for improving scalability that we discuss in Sections 2–5 are largely complementary, and these sections can thus be read independently of each other. We provide the following rough guide to the different sections (which may be used, e.g., by a practitioner interested in a specific application where scalability is a challenge):

- If the SDP of interest (corresponding to Problem 1) has special structure (e.g., symmetry, sparsity, or degeneracy), refer to Section 2.
- If either the SDP has low-rank (optimal) solutions or one desires good-quality low-rank feasible points to the SDP, refer to Section 3.
- If approximately feasible solutions with good objective values are of interest (i.e., slight violation of the constraints can be tolerated), refer to Section 4.
- If conservative feasible solutions to the SDP (i.e., points that satisfy all constraints but are potentially suboptimal) are valuable, refer to Section 5.

The interested reader may also use this guide to explore the software packages listed in Section 6, which are similarly organized according to sections in the article.

In principle, the approaches presented in the different sections may be combined to further enhance scalability (e.g., in order to tackle a problem exhibiting sparsity for which slightly suboptimal feasible solutions are valuable). We highlight such combinations of approaches where possible in Sections 2–5.

1.3. Related Work Not Covered by This Survey

The number of algorithms that have been proposed as alternatives to interior-point methods for semidefinite programming is large and rapidly growing. While we have attempted to present some of the major themes, our survey is certainly not exhaustive. Some of the interesting contributions that this article does not cover due to space limitations include the recent first-order methods developed by Renegar (11), the multiplicative weights update algorithm of Arora et al. (12), the storage-optimal algorithm of Ding et al. (13), the iterative projection methods of Henrion & Malick (14), the conditional gradient–based augmented Lagrangian framework of Yurtsever et al. (15), the regularization methods of Nie & Wang (16) for SDP relaxations of polynomial optimization, the accelerated first-order method of Bertsimas et al. (17) for the SOS relaxation of unconstrained polynomial optimization, and the spectral algorithms of Hopkins et al. (18) for certain SOS problems arising in machine learning.

We also remark that there has been a relatively large recent literature at the intersection of optimization and machine learning where SDP relaxations of certain nonconvex problems are avoided altogether, and instead the nonconvex problem is tackled directly with local descent algorithms. These algorithms can often scale to larger problems compared with the SDP approach (at least when the SDPs are solved by interior-point methods). Moreover, under certain technical caveats and statistical assumptions on problem data, it is sometimes possible to prove guarantees for these algorithms similar to those of the SDP relaxation. We do not cover this literature in this article but point the interested reader to Reference 19 for a list of references.

1.4. A Brief Review of Sum of Squares Proofs of Nonnegativity

SOS constraints on multivariate polynomials are responsible for a substantial fraction of recent applications of semidefinite programming. For the convenience of the reader, we briefly review the basics of this concept and the context in which it arises. This will be relevant, e.g., for a better understanding of Sections 2.1.1 and 5.1.

Recall that a closed basic semialgebraic set is a subset of the Euclidean space of the form

$$\Omega := \{x \in \mathbb{R}^n \mid g_i(x) \geq 0, i = 1, \ldots, m\},$$

where $g_1, \ldots, g_m$ are polynomial functions. It is relatively well known by now that many fundamental problems of computational mathematics can be cast as optimization problems where the decision variables are the coefficients of a multivariate polynomial p, the objective function is a linear function of these coefficients, and the constraints are twofold: (*a*) affine constraints in the coefficients of p and (*b*) the constraint that

$$p(x) \geq 0 \ \forall x \in \Omega, \qquad 3.$$

where Ω is a given closed basic semialgebraic set. For example, in a polynomial optimization problem—i.e., the problem of minimizing a polynomial function f on Ω—the optimal value is equal to the largest constant γ for which $p(x) := f(x) - \gamma$ satisfies the constraint shown in Equation 3. (For numerous other applications, see References 20 and 21 and section 1.1 of

Reference 8.) Unfortunately, imposing the constraint in Equation 3 (or even asking a decision version of it for a fixed polynomial p) is already NP-hard when p is a quartic polynomial and $\Omega = \mathbb{R}^n$ or when p is quadratic and Ω is a polytope.

An idea pioneered to a large extent by Lasserre (6) and Parrilo (7) has been to write algebraic sufficient conditions for Equation 3 based on the concept of sos polynomials. We say that a polynomial h is a sos if it can be written as $h = \sum_i q_i^2$ for some polynomials q_i. Observe that if we succeed in finding sos polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$ such that the polynomial identity

$$p(x) = \sigma_0(x) + \sum_{i=1}^{m} \sigma_i(x) g_i(x) \qquad 4.$$

holds, then clearly Equation 3 must be satisfied. Conversely, a celebrated result in algebraic geometry (22) states that if $g_1, \ldots, g_m$ satisfy the so-called Archimedean property (a condition slightly stronger than compactness of the set Ω), then positivity of p on Ω guarantees the existence of sos polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$ that satisfy the algebraic identity in Equation 4.

The computational appeal of the SOS approach stems from the fact that the search for sos polynomials $\sigma_0, \sigma_1, \ldots, \sigma_m$ of a given degree that verify the polynomial identity in Equation 4 can be automated via semidefinite programming. This is true even when some coefficients of the polynomial p are left as decision variables. This claim is a straightforward consequence of the following well-known fact (see, e.g., 9): A polynomial h of degree $2d$ is a sos if and only if there exists a symmetric matrix Q that is psd and verifies the identity

$$h(x) = z(x)^T Q z(x),$$

where $z(x)$ here denotes the vector of all monomials in x of degree less than or equal to d. Note that the size of the semidefinite constraint in an SDP that imposes a SOS constraint on an n-variate polynomial of degree $2d$ is $\binom{n+d}{d} \times \binom{n+d}{d} \approx n^d \times n^d$. This is the reason why SOS optimization problems—i.e., SDPs arising from SOS constraints on polynomials—often quickly run into scalability limitations.

2. ENHANCING SCALABILITY BY EXPLOITING STRUCTURE

A major direction for improving the scalability of semidefinite programming has been the development of methods for exploiting problem-specific structure. Existing literature in this area has focused on two kinds of structure that appear frequently in applications in machine learning, control, and robotics: sparsity and symmetry.

2.1. Exploiting Sparsity

Many problems in control theory give rise to SDPs that exhibit structure in the form of chordal sparsity. We first give a brief introduction to the general notion of chordal sparsity and then highlight applications in control theory that exploit this structure to achieve significant gains in scalability. We begin by introducing a few relevant graph-theoretic notions.

Definition 1 (cycles and chords). A cycle of length $k \geq 3$ in a graph $\mathcal{G}$ is a sequence of distinct vertices $(v_1, v_2, \ldots, v_k)$ such that (v_k, v_1) is an edge and (v_i, v_{i+1}) is an edge for $i = 1, \ldots, k-1$. A chord (associated with a cycle) is an edge that is not part of the cycle but connects two vertices of the cycle.

Definition 2 (chordal graph). An undirected graph $\mathcal{G}$ is chordal if every cycle of length $k \geq 4$ has a chord.

Definition 3 (maximal clique). A clique of a graph $\mathcal{G}$ is a subset $\mathcal{C}$ of vertices such that for any distinct vertices $i, j \in \mathcal{C}$, (i, j) is an edge in $\mathcal{G}$. A clique is maximal if it is not a subset of another clique.

The graph-theoretic notions introduced above can be used to capture sparsity patterns of matrices. Let $\mathcal{E}$ and $\mathcal{V}$ denote the sets of edges and vertices, respectively, of an undirected graph $\mathcal{G}$. Let S_n denote the set of symmetric $n \times n$ matrices as above, and let $S_n(\mathcal{E})$ denote the set of matrices with a sparsity pattern defined by $\mathcal{G}$ as follows:

$$S_n(\mathcal{E}) := \{X \in S_n \mid X_{ij} = 0 \text{ if } (i, j) \notin \mathcal{E} \text{ for } i \neq j\}. \quad 5.$$

The following proposition is the primary result that allows one to exploit chordal sparsity (i.e., sparsity induced by a chordal graph) to improve the scalability of an SDP.

Proposition 1 (see Reference 23). Let $\mathcal{G}$ be a chordal graph with vertices $\mathcal{V}$, edges $\mathcal{E}$, and a set of maximal cliques $\Gamma = \{\mathcal{C}_1, \ldots, \mathcal{C}_p\}$. Then $X \in S_n(\mathcal{E})$ is psd if and only if there exists a set of matrices $X_k \in S(\mathcal{C}_k) := \{X \in S_n \mid X_{ij} = 0 \text{ if } (i, j) \notin \mathcal{C}_k \times \mathcal{C}_k\}$ for $k = 1, \ldots, p$, such that $X_k \succeq 0$ and $X = \sum_{k=1}^{p} X_k$.

The above statement allows one to equivalently express a large semidefinite constraint as a set of smaller semidefinite constraints (and additional equality constraints). This can lead to a significant increase in computational efficiency since the computational cost of solving an SDP is determined primarily by the size of the largest semidefinite constraint. We note that for a chordal graph, the problem of listing maximal cliques is amenable to efficient (i.e., polynomial time) algorithms (see 24 and references therein).

Proposition 1 [along with a related dual result given by Grone et al. (25)] is leveraged by several researchers in optimization to improve the efficiency of SDPs with chordal sparsity (26–28). These results have also been exploited by researchers in the context of control theory. In Reference 29, the authors applied these results to the problem of finding Lyapunov functions for sparse linear dynamical systems. In particular, by parameterizing Lyapunov functions with a chordal structure (i.e., a structure that allows one to apply Proposition 1), the resulting SDP finds a Lyapunov function significantly faster (up to a factor of ∼80 faster in terms of running time for instances where the largest maximal clique size is less than 15).

In References 24 and 30, the authors extend Proposition 1 to the case of matrices with block-chordal sparsity by demonstrating that such matrices can be decomposed into an equivalent set of smaller block-psd matrices (again with additional equality constraints). This result allows for applications to networked systems since it reasons about the sparsity of interconnections between subsystems. The authors apply these results to the problem of designing structured feedback controllers to stabilize large-scale networked systems.

In the work highlighted above, Proposition 1 is used to decompose a large semidefinite constraint into smaller ones with additional equality constraints. The resulting SDP can then be solved using standard techniques (e.g., standard interior-point methods). As noted in Reference 31, one challenge with this approach is that the additional equality constraints can sometimes nullify the computational benefits of dealing with smaller semidefinite constraints. A set of strategies that seek to overcome this issue involve exploiting chordal sparsity directly in the solver, e.g., directly in an interior-point method (26, 32, 33) or in a first-order method (31, 34–37). In the context of control theory, Reference 38 leverages sparse solvers (26, 33, 39) based on this idea to tackle robust stability analysis for large-scale sparsely interconnected systems.

2.1.1. Exploiting sparsity in polynomial optimization problems. The existence of structure in the form of sparsity can also be exploited for polynomial optimization problems (see Section 1.4). In Reference 40, the authors propose the Sparse-BSOS hierarchy, which is based on the BSOS (bounded degree SOS) hierarchy for polynomial optimization problems presented in Reference 41. The Sparse-BSOS hierarchy leverages the observation that, for many polynomial optimization problems that arise in practice, the polynomials that define the constraints of the problem exhibit sparsity in their coefficients. This observation is used to split the variables in the problem into smaller blocks of variables such that (*a*) each monomial of the polynomial objective function consists only of variables from one of the blocks and (*b*) each polynomial that defines the constraints also depends only on variables in only one of the blocks. Under the assumption that the sparsity in the objective and constraints satisfies the running intersection property (40), the Sparse-BSOS approach provides a hierarchy of SDPs whose optimal values converge to the global optimum of the original polynomial optimization problem. In addition, for the so-called class of SOS-convex polynomial optimization problems (for a definition, see 40) that satisfy the running intersection property, the hierarchy converges at the very first step. In contrast to the (standard) BSOS hierarchy, which also provides these guarantees, Sparse-BSOS results in semidefinite constraints of smaller block size. Numerical experiments in Reference 40 demonstrate that exploiting sparsity in the problem can result in the ability to solve large-scale polynomial optimization problems that are beyond the reach of the standard (i.e., nonsparse) BSOS approach.

In Reference 42, the authors utilize the Sparse-BSOS hierarchy to tackle the problem of simultaneous localization and mapping (SLAM) in robotics. SLAM (43) refers to the problem of simultaneously estimating the location (or trajectory of locations) of a robot and a model of the robot's environment. Two important versions of this problem are landmark SLAM and pose-graph SLAM. The landmark SLAM problem involves simultaneously estimating the position and orientation of the robot and the locations of landmarks in the environment. The pose-graph SLAM problem is a more restricted version that only requires estimating the position and orientation of the robot at each time step. Traditionally, these problems have been posed as maximum likelihood estimation problems and tackled using general nonlinear optimization methods (which do not come with guarantees on finding globally optimal solutions). In Reference 42, the authors formulate the planar landmark and pose-graph SLAM problems as polynomial optimization problems and demonstrate that they are SOS-convex. This allows them to apply the Sparse-BSOS hierarchy with the guarantee that the hierarchy converges to the global optimum at the very first step. The approach in this work contrasts with most prior works on landmark and pose-graph SLAM, which apply general-purpose nonlinear optimization algorithms to these problems. The numerical results in Reference 42 are used to compare the Sparse-BSOS approach to SLAM with nonlinear optimization techniques based on the Levenberg–Marquardt algorithm and demonstrate that the latter often yield suboptimal solutions, while the former finds the globally optimal solution. We also note that their approach contrasts with the SE-Sync approach (44) mentioned in Section 3.1.2, since the former relaxes the requirement of limited measurement noise imposed by the latter.

2.2. Exploiting Symmetry

Beyond sparsity, another general form of structure that arises in applications of semidefinite programming to machine learning, control, and robotics is symmetry. Mathematically, symmetry refers to a transformation of the variables in a problem that does not change the underlying problem. As a concrete example, consider a multiagent robotic system composed of a large number of identical agents. The agents in the system can be interchanged while leaving the overall capabilities of the multiagent system invariant.

Symmetry reduction (45–48) is a powerful set of approaches for exploiting symmetries in SDPs. We provide a brief overview of symmetry reduction here and refer the reader to Reference 48 for a thorough exposition. Let S_n^+ denote the cone of $n \times n$ symmetric psd matrices. Given an SDP of the form of Problem 1, symmetry reduction performs the following two steps:

1. Find an appropriate (see 48) affine subspace $\mathcal{S} \subset S_n$ containing feasible solutions to the problem.
2. Express $\mathcal{S} \cap S_n^+$ as a linear transformation of a simpler cone $\mathcal{C} \subset S_m$, where $m < n$:

$$\mathcal{S} \cap S_n^+ = \{\Phi Z \mid Z \in \mathcal{C}\}. \tag{6.}$$

Symmetry reduction techniques work by taking $\mathcal{S}$ equal to the fixed-point subspace of the group action corresponding to the underlying symmetry group (for details, see 45, 48).

Once steps 1 and 2 above have been accomplished, one can solve the following problem instead of the original SDP:

$$\begin{aligned} \min_{Z \in S_m} \quad & \mathrm{Tr}(C\Phi Z) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i \Phi Z) = b_i, \quad i = 1, \ldots, m, \\ & Z \in \mathcal{C}. \end{aligned} \tag{7.}$$

The key advantage here is that Problem 7 is formulated over a lower-dimensional space S_m and can thus be more efficient to solve.

Symmetry reduction techniques have been fruitfully applied in many different applications of semidefinite programming to control theory (49–51). Reference 50, for example, exploits symmetry in linear dynamical systems; more precisely, it exploits invariances of the state-space realizations of linear systems under unitary coordinate transformations to obtain significant computational gains for control synthesis problems. In Reference 51 (chapter 7), symmetry reduction techniques are used to reduce the computational burden of certifying the stability of interconnected subsystems. Specifically, it exploits permutation symmetries in a network of dynamical systems (i.e., invariances to exchanging subsystems) to reduce the complexity of the resulting SDPs.

2.3. Facial Reduction

In certain cases, there may be additional structure beyond sparsity and symmetry that one can exploit to reduce the sizes of SDPs arising in practice. As an example, consider the search for a Lyapunov function $V : \mathbb{R}^n \to \mathbb{R}$ that proves the stability of the equilibrium $x_0 \in \mathbb{R}^n$ of a polynomial dynamical system $\dot{x} = f(x)$. In this case, one would like to search for functions V that satisfy the following degeneracy conditions: $V(x_0) = 0, \dot{V}(x_0) = 0$. This imposes structure on the resulting optimization problem by restricting the space of possible functions V. One powerful set of approaches for leveraging this kind of degeneracy structure is referred to as facial reduction (48, 52–55). The general procedure behind facial reduction is identical to that of symmetry reduction (steps 1 and 2 in Section 2.2). The primary difference is that facial reduction techniques identify the subspace $\mathcal{S}$ in a different way than symmetry reduction techniques do. In particular, $\mathcal{S}$ is found by exploiting a certain kind of geometric degeneracy condition (for details, see 48).

Facial reduction techniques have been used to improve the scalability of SDPs arising in machine learning, control, and robotics applications (48, 56, 57). We highlight Reference 48, which presents methods for automating the process of performing facial (and symmetry) reduction. As an example, Reference 48 considers the problem of analyzing the stability of a rimless wheel, which is a simple hybrid dynamical walking robot model. The facial reduction techniques presented reduce

computation times by roughly a factor of 60 for this problem. Similar gains in terms of scalability are also obtained for many other problems. Another recent facial reduction algorithm that stands out in terms of its simplicity of implementation is Sieve-SDP (58). This technique can detect redundancies in the constraints of an SDP (or sometimes infeasibility) by checking positive definiteness of some sub-blocks of the SDP data matrices. In Reference 58, the authors demonstrate the performance of this preprocessing strategy on several benchmark examples, including many arising from SDP relaxations of polynomial optimization problems.

3. ENHANCING SCALABILITY BY PRODUCING LOW-RANK SOLUTIONS

In this section, we focus on methods for producing low-rank feasible solutions to Problem 1 at relatively large scales. One may think of the techniques we present here as leveraging a very specific kind of structure (i.e., low-rank structure), similar to the approaches considered in Section 2. However, the technical approaches for leveraging low-rank structure are quite different in nature from the approaches presented in Section 2. Moreover, the literature on low-rank methods is quite vast. We thus treat these methods distinctly from the ones in Section 2. At a high level, one utilizes such methods in two different settings, which we describe now.

Case 1 (there is a low-rank optimal solution to Problem 1). In this setting, we can establish a priori that among the optimal solutions to Problem 1, there must be one of low rank. An important case where this is known to hold is when Problem 1 does not have too many constraints. More specifically (see 59–61), if Problem 1 has m constraints and an optimal solution, then it also has a rank-r optimal solution with

$$\frac{r(r+1)}{2} \leq m. \qquad 8.$$

Note that we need $m < \frac{n(n+1)}{2}$ for this property to be useful.

In this setting, one may not particularly care about obtaining a low-rank (optimal) solution to the problem. Rather, searching for a low-rank solution (whose existence is guaranteed) can lead to significant gains in computation time and storage space. To see why, recall that an $n \times n$ psd matrix X has rank $r \leq n$ if and only if it can be written as $X = UU^T$, where $U \in \mathbb{R}^{n \times r}$. By searching for a low-rank solution, one can hope to work in the lower-dimensional space $\mathbb{R}^{n \times r}$ corresponding to U rather than the higher-dimensional space S_n corresponding to X. Doing so would lead to much cheaper algorithms because storage-space needs would drop from $O(n^2)$ to $O(nr)$, and, e.g., the important operation of matrix-vector multiplication would require $O(nr)$ flops rather than $O(n^2)$.

Case 2 (good-quality, low-rank feasible solutions to Problem 1 are of interest). In this setting, one considers SDPs whose optimal solutions may all have high rank. However, the goal is to find low-rank feasible solutions with good objective values. By making this compromise, one can again expect to work in the lower-dimensional space of low-rank feasible solutions and make gains in terms of computation time and storage space, as explained above.

Problems such as these occur in a variety of applications, most notably machine learning and combinatorial optimization. Consider, for instance, the problem of low-rank matrix completion (62) from machine learning, which is a cornerstone problem in user recommendation systems [e.g., as in the Netflix Prize (63)]. In this problem, we have access to a matrix $Z \in \mathbb{R}^{m \times n}$, which is partially complete—i.e., some of its entries are specified, but others are

unknown. The goal is to fill in the unknown entries. Without assuming some structure on Z, any completion of Z would be a valid guess, so the problem is ill defined. This can be remedied by assuming that Z should have low rank, which is a natural assumption in the context of user recommendation systems. Indeed, if one assumes that the matrix Z reflects the ratings a user i ($i = 1, \ldots, n$) gives a product j ($j = 1, \ldots, m$), then completing the matrix Z would involve inferring how a user would have rated a product that he or she has not yet rated based on existing ratings by both the user in question and other users. Under the assumption that a consumer base can be segmented into r categories (with $r \ll n$), it would then make sense to constrain the matrix Z to have rank less than or equal to r, which would reflect this segmentation. The matrix completion problem would then be written as

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}} \quad & \sum_{i,j} (X_{ij} - Z_{ij})^2 \\ \text{s.t.} \quad & \operatorname{rank}(X) \leq r, \end{aligned} \tag{9.}$$

where Z is the partially complete matrix corresponding to past data. The indices i, j in the objective function range over known entries of the matrix Z. Problem 9 can be relaxed to an SDP by replacing the constraint on the rank of X by a constraint on its nuclear norm (3), which is a semidefinite-representable surrogate of the rank. The problem then becomes

$$\begin{aligned} \min_{X \in \mathbb{R}^{m \times n}, W_1 \in S_m, W_2 \in S_n} \quad & \sum_{i,j} (X_{ij} - Z_{ij})^2 \\ \text{s.t.} \quad & \operatorname{Tr} \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \leq r, \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0. \end{aligned} \tag{10.}$$

The goal here is to obtain a feasible solution to the problem such that the matrix

$$\begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix}$$

has low rank and the objective value is small. This makes matrix completion a typical application of the algorithms we present next. Other machine learning problems that involve relaxing a rank-constrained SDP to an SDP via the use of the nuclear norm include certain formulations of clustering and maximum variance unfolding (64) and sparse principal component analysis (4).

The rest of this section is devoted to two of the most well-known approaches for computing low-rank feasible (and possibly optimal) solutions to Problem 1. The first is the Burer–Monteiro method and its variants (Section 3.1), and the second includes methods that use the Frank–Wolfe algorithm as their computational backbone (Section 3.2).

3.1. Burer–Monteiro-Based Methods

The key idea of the Burer–Monteiro algorithm (65) is to factor the decision variable X in Problem 1 as VV^T, where $V \in \mathbb{R}^{n \times r}$. Here, r can be chosen either such that a solution of rank r is guaranteed to exist (e.g., such that the inequality shown in Equation 8 holds—Case 1 above) or such that it corresponds to the rank that we would like our feasible solution to have (Case 2 above). Typically, the algorithm comes with some theoretical guarantees for the first case and is simply a

heuristic in the second. Using this factorization, we rewrite Problem 1 as

$$\begin{aligned} \min_{V \in \mathbb{R}^{n \times r}} \quad & \mathrm{Tr}(CVV^T) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i VV^T) = b_i, \quad i = 1, \ldots, m. \end{aligned} \tag{11.}$$

This reformulation has both advantages and disadvantages: On the upside, the problem to consider has smaller dimension than Problem 1 and has no conic constraint to contend with; on the downside, the problem has become nonconvex. Hence, local optimization methods may not always recover the global minimum of Problem 1 even if there is one of low rank. Research in this area has consequently sought to find algorithms to solve Problem 11 that provably recover global low-rank solutions to the problem if they exist, do so reasonably fast in terms of both theoretical and empirical convergence rates, and work on a large class of inputs A_i, b_i, and C. We review existing algorithms through the lens of these criteria.

3.1.1. Augmented Lagrangian approaches. The original method recommended by Burer & Monteiro (65) to solve Problem 11 is based on an augmented Lagrangian procedure (for another appearance of this general approach, see Section 4.2). The idea is to rewrite Problem 11 as an unconstrained optimization problem by adding terms to the objective that penalize infeasible points. This is done by constructing an augmented Lagrangian function, given here by

$$L_{y,\sigma}(V) = \mathrm{Tr}(CVV^T) - \sum_{i=1}^{m} y_i(\mathrm{Tr}(A_i VV^T) - b_i) + \frac{\sigma}{2} \sum_{i=1}^{m} (\mathrm{Tr}(A_i VV^T) - b_i)^2.$$

If y and $\sigma > 0$ are fixed appropriately, then minimizing $L_{y,\sigma}(\cdot)$ will yield an optimal solution to Problem 11. To obtain an appropriate pair (y, σ), the following iterative procedure is followed: Fix (y, σ) and minimize $L_{y,\sigma}(\cdot)$ to obtain V, then construct a new (y, σ) from V and repeat (for details, see 65). Note that this procedure involves the minimization of $L_{y,\sigma}$ with respect to V; choosing an appropriate algorithm for this subproblem is also an important step in the method. In Reference 65, the authors suggest using a limited-memory Broyden–Fletcher–Goldfarb–Shanno (LM-BFGS) algorithm (66) to do so. The LM-BFGS algorithm is a quasi-Newton algorithm for unconstrained optimization problems, where the objective is differentiable. Unlike Newton methods, it does not require computing the Hessian of the objective function, constructing instead an approximate representation of it from past gradients. Given that gradients of $L_{y,\sigma}$ can be computed quite quickly under some conditions (for details, see 65), this is a reasonably efficient algorithm to use. One can replace the LM-BFGS subroutine with a subgradient approach if Problem 11 has additional constraints that are nondifferentiable. Furthermore, the augmented Lagrangian approach has been tailored to directly handle inequality constraints of the type $\mathrm{Tr}(A_i X) \leq b_i$ in Problem 1 (see 64). Though there are no theoretical guarantees regarding convergence to global solutions, the authors in Reference 65 experimentally observe that the algorithm tends to return global minima of Problem 1, doing so in a much smaller amount of time than interior-point or bundle methods. In a follow-up paper (67), the authors slightly modify the augmented Lagrangian function and show convergence of the algorithm to a global minimum of Problem 1 (given that one exists of rank $\leq r$) if, among other things, each minimization of the modified $L_{y,\sigma}(\cdot)$ gives rise to a local minimum, a condition that is hard to test in practice.

3.1.2. Riemannian optimization approaches. Riemannian optimization concerns itself with the optimization of a smooth objective function over a smooth Riemannian manifold. We remind the reader that a Riemannian manifold $\mathcal{M}$ is a manifold that can be linearized at each point $x \in \mathcal{M}$ by a tangent space $T_x\mathcal{M}$, which is equipped with a Riemannian metric. We say that a Riemannian

manifold is smooth if the Riemannian metric varies smoothly as the point x varies (for a short introduction to such concepts, see, e.g., appendix A of Reference 68). In the Riemannian setting, one can define notions of both gradients and Hessians for manifolds (68, appendix A). Using these notions, one can extend the concepts of first-order and second-order necessary conditions of optimality to optimization over manifolds. These are almost identical to the ones for unconstrained optimization over $\mathbb{R}^n$, except that they involve the analogs of gradient and Hessians for manifolds. Similarly, one can generalize many classical algorithms for nonlinear unconstrained optimization (e.g., trust-region and gradient-descent methods) to the manifold setting. References 69 and 70 provide more information on the specifics of these algorithms, which we do not cover here.

Though it may not be immediately apparent why, Riemannian optimization methods have become a popular tool for tackling problems of the form of Problem 11 (71, 72). Indeed, the set

$$\mathcal{M} = \{V \in \mathbb{R}^{n\times r} \mid \mathrm{Tr}(A_i V V^T) = b_i, i = 1, \ldots, m\}$$

is a smooth Riemannian manifold under certain conditions on A_i (see 73, assumption 1), and the objective function $V \mapsto \mathrm{Tr}(CVV^T)$ is smooth. Hence, Problem 11 is exactly a Riemannian optimization problem. It can be shown that if $m < \frac{r(r+1)}{2}$, the conditions on $\mathcal{M}$ that make it smooth hold, and $\mathcal{M}$ is compact, then for almost all $C \in S_n$, any second-order critical point of Problem 11 is globally optimal to Problem 1 (72, theorem 2). (Note that an optimal solution of rank $\leq k$ is guaranteed to exist here from the assumptions and that we use the manifold definition of a second-order critical point.) If an algorithm that returns second-order critical points can be exhibited, then, in effect, we will have found a way of solving Problem 11. It turns out that Riemannian trust-region methods return such a point—under some conditions—regardless of initialization (74). In fact, it is shown in Reference 74 that an ϵ-accurate second-order critical point [i.e., a point $V \in \mathcal{M}$ with $||\mathrm{grad}(\mathrm{Tr}(CVV^T))|| \leq \epsilon$ and $\mathrm{Hessian}(\mathrm{Tr}(CVV^T)) \succeq -\epsilon I$] can be obtained in $O(1/\epsilon^2)$ iterations. Other Riemannian methods that come with some theoretical guarantees are the Riemannian steepest-descent algorithm and the Riemannian conjugate gradient algorithm, both of which are known to converge to a critical point of Problem 11 (74, 75). In practice, however, many more algorithms for nonlinear optimization can be and have been generalized to manifold settings (though their convergence properties have not necessarily been studied) and are used to solve problems such as Problem 11 [see, e.g., the list of solvers implemented in Manopt (76), one of the main toolboxes for optimization over manifolds].

We conclude this section by mentioning a few applications where Riemannian optimization methods have recently been used. These include not only machine learning problems such as synchronization (77), phase recovery (78), dictionary learning (79), and low-rank matrix completion (68), but also robotics problems such as the SLAM problem (see Section 2.1.1). In References 44, 80, and 81, the authors propose semidefinite programming relaxations for the maximum likelihood estimation problem arising from SLAM. These relaxations provide exact solutions to the maximum likelihood estimation problem, assuming that noise in the measurements made by the robot is below a certain threshold (for the exact conditions, see 44). In Reference 44, the authors tackle the challenge of scalability with an approach called SE-Sync, demonstrating that the resulting SDP admits low-rank solutions and using a Burer–Monteiro factorization combined with Riemannian optimization techniques (e.g., a Riemannian trust-region approach) to solve the SDP. Thus, assuming that the conditions on the measurement noise are satisfied, one can find globally optimal solutions to the maximum likelihood estimation problems arising from pose-graph SLAM (as well as other related maximum likelihood estimation problems arising from camera motion estimation and sensor network localization problems) via this method.

3.1.3. Coordinate-descent approaches. One of the more recent trends for solving problems of the form of Problem 11 has been to use coordinate-descent (or block-coordinate-descent) methods (82, 83). These methods have been proposed to solve diagonally constrained problems, i.e., problems of the type

$$\begin{aligned} \min_{V \in \mathbb{R}^{n \times r}} \quad & \mathrm{Tr}(CVV^T) \\ \text{s.t.} \quad & ||v_i^T|| = b_i, \ i = 1, \ldots, n, \end{aligned} \tag{12.}$$

where $||\cdot||$ is the 2-norm and v_i^T is the ith row of V. This subset of SDPs covers machine learning applications such as certain formulations of graphical model inference and community detection (82). Coordinate-descent methods are easy to explain for this case. For clarity of exposition, let us assume that $b_i = 1$ for $i = 1, \ldots, m$. Initialization consists of randomly choosing $v_i, i = 1, \ldots, n$ on the sphere. Then, at each iteration, an index i is picked and fixed, and the goal is to find a feasible vector v_i such that $\mathrm{Tr}(CVV^T)$ is minimized. We update the objective by replacing the previous v_i with this new v_i and proceed to the next iteration. This algorithm can be slightly modified to consider descent on blocks of coordinates rather than one single coordinate at each time step (82). In that setting, it can be shown that if the indices are picked appropriately, then coordinate block-descent methods converge to an ϵ-accurate critical point of Problem 12 in time $O(\frac{1}{\epsilon})$. Each iteration of coordinate-descent methods is much less costly (by an order of n, approximately) than its Riemannian trust-region counterpart, however. As a consequence, numerical results tend to show that as n increases, using coordinate-descent methods can be preferable (82).

3.2. Frank–Wolfe-Based Methods

We start this subsection with a brief presentation of the Frank–Wolfe algorithm (84), also known as the conditional gradient algorithm, a version of which will be used in the approaches we present. This is an algorithm for constrained optimization problems of the type $\min_{x \in S} f(x)$, where S is a compact and convex set and f is a convex and differentiable function. At iteration k, a linear approximation of f (obtained using the first-order Taylor approximation of f around the current iterate x_k) is minimized over S, and a minimizer s_k is obtained. The next iterate x_{k+1} is then constructed by taking a convex combination of the previous iterate x_k and the minimizer s_k before repeating the process. Intuitively, the reason why Frank–Wolfe algorithms are popular for obtaining low-rank feasible solutions for SDPs is that, if the algorithm is initialized with a rank-1 matrix and each minimizer is also a rank-1 matrix, then the kth iterate is of rank at most k. This enables us to control the rank of the solution that is given as output. We first present an algorithm developed by Hazan (85) in Section 3.2.1 and then follow up with some extensions of Hazan's algorithm in Section 3.2.2.

3.2.1. Hazan's algorithm. Hazan's algorithm is designed to produce low-rank solutions to SDPs of the type

$$\begin{aligned} \min_{X \in S_n} \quad & \mathrm{Tr}(CX) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i X) = b_i, \ i = 1, \ldots, m, \\ & X \succeq 0 \ \text{ and } \ \mathrm{Tr}(X) = 1. \end{aligned} \tag{13.}$$

The additional constraint $\mathrm{Tr}(X) = 1$ is required by the algorithm but can be slightly relaxed to the constraint $\mathrm{Tr}(X) \leq t$, where t is fixed (86, chapter 5). We consider the version with the $\mathrm{Tr}(X) = 1$ constraint here. A key subroutine of Hazan's algorithm is solving the following optimization

problem:

$$\begin{aligned} \min_{X \in S_n} \quad & f(X) \\ \text{s.t.} \quad & \mathrm{Tr}(X) = 1 \text{ and } X \succeq 0, \end{aligned} \tag{14.}$$

where f is a convex function. This is of interest in itself because some problems can be cast in the form of Problem 14, an example being given in Reference 85. Let f^* be the optimal value of this problem. It is for this subroutine that a Frank–Wolfe-type iterative algorithm is used. This algorithm can be described as follows: Initialize X_1 to be a rank-1 matrix with trace 1. Then, at iteration k, compute the eigenvector v_k corresponding to the maximum eigenvalue of $\nabla f(X_k)$ written in matrix form. Finally, for a step size α_k, set $X_{k+1} = (1 - \alpha_k)X_k + \alpha_k(-v_k v_k^T)$ and iterate. This algorithm returns a $\frac{1}{k}$-approximate solution to Problem 14 [i.e., a matrix X such that $\mathrm{Tr}(X) \leq 1 + \frac{1}{k}$ and $f(X) \leq f^* + \Omega(\frac{1}{k})$] with rank at most k in k iterations.

To see how this subroutine can be used to solve Problem 13, first note that one can reduce this problem to a sequence of SDP feasibility problems via bisection. It is therefore sufficient to explain how one can leverage the Frank–Wolfe-type algorithm described above to solve an SDP feasibility problem of the type

$$\mathrm{Tr}(A_i X) \leq b_i, i = 1, \ldots, m, X \succeq 0, \mathrm{Tr}(X) = 1.$$

This can be done by letting $f(X) = \frac{1}{M} \log\left(\sum_{i=1}^{m} e^{M \cdot (\mathrm{Tr}(\mathrm{A_i X}) - \mathrm{b_i})}\right)$, where $M = k \log(m)$, $\frac{1}{k}$ being the desired accuracy (see 85). The algorithm is then guaranteed to return a $\frac{1}{k}$-approximate solution of rank at most k in $O(k^2)$ iterations.

It is interesting to note that this latter algorithm is nearly identical to the multiplicative weights update algorithm for SDPs, which also produces low-rank solutions (12) and has exactly the same guarantees. However, it is derived in a completely different fashion.

3.2.2. Other Frank–Wolfe-based methods. A caveat of Hazan's algorithm is that the rank of the solution returned is linked to its accuracy. One may think that, if the solution is known to be low rank and the Frank–Wolfe algorithm is initialized with a low-rank matrix, then all iterates of the algorithm produce a low-rank matrix. Unfortunately, this is not always the case. In fact, what can be observed in practice is that the Frank–Wolfe algorithm typically returns iterates with increasing rank up to a certain point, where the rank decreases again, with the final solution returned being low rank (see, e.g., the numerical experiments in Reference 87). As a consequence, if one requires an accurate solution, one will likely have to deal with high-rank intermediate iterates, with the storage and computational problems that this entails.

Here, we present two different methods designed to avoid this issue for optimization problems of the type

$$\begin{aligned} \min_{X \in \mathbb{R}^{n \times d}} \quad & f(\mathcal{A}X) \\ \text{s.t.} \quad & ||X||_* \leq 1, \end{aligned} \tag{15.}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is a convex and differentiable function, $|| \cdot ||_*$ denotes the nuclear norm, and $\mathcal{A} : \mathbb{R}^{n \times d} \to \mathbb{R}^m$ is a linear operator that maps X to $(\mathrm{Tr}(A_1 X), \ldots, \mathrm{Tr}(A_m X))^T$. The relaxation of the low-rank matrix completion problem given in Equation 10, for instance, fits into this format. Similarly to Problem 14, the Frank–Wolfe algorithm can be applied to Problem 15 quite readily. Initialization is done as before by setting X_0 to be some rank-1 matrix. At iteration k, one computes an update direction by finding a pair (u_k, v_k) of singular vectors corresponding to the maximum

singular value of $\mathcal{A}^*(\nabla f(\mathcal{A}X_k))$, where $\mathcal{A}^* : \mathbb{R}^m \mapsto \mathbb{R}^{n\times d}$ is the adjoint operator to $\mathcal{A}$ given by $\mathcal{A}^* z = \sum_{i=1}^m z_i A_i$. We then construct a new iterate $X_{k+1} = (1-\alpha)X_k - \alpha u_k v_k^T$ and repeat.

The first method, described in Reference 88, is a modification of the Frank–Wolfe algorithm as described above. The key idea is that, at every iteration, the algorithm chooses between two types of steps. One type is the step described above, which is derived from the singular vectors of the matrix $\mathcal{A}^*(\nabla f(\mathcal{A}X_k))$, referred to in Reference 88 as a "regular step"; the other is what they call an "in-face step." One way to take such a step is to minimize $\mathrm{Tr}((\mathcal{A}^*\nabla f(AX_k))^T X)$ over X in the minimal face of the nuclear norm unit ball to which X_k belongs. The matrix Z_k obtained from this minimization is then used to produce the next iterate: $X_{k+1} = (1-\alpha)X_k + \alpha Z_k$. (Reference 88 explains how this can be done efficiently.) The advantage of such a step is that the matrix X_{k+1} constructed has rank no larger than X_k. If the choice between a regular step and an in-face step is appropriately made, it can still be the case that one has a $\frac{1}{k}$-approximate solution after k iterations, though the rank of the kth iterate could be much less than k depending on the number of in-face steps used.

The second method, given in Reference 87, provides an alternative way of dealing with the possibly high rank of the intermediate iterates X_k produced by the Frank–Wolfe algorithm applied to Problem 15. To get around this problem, the authors devise a dual formulation of the Frank–Wolfe algorithm described above, where the primal iterates $X_k \in \mathbb{R}^{n\times d}$ are replaced by dual iterates $z_k \in \mathbb{R}^m$, which are of smaller size. There is a need, however, to keep track of the iterates X_k in some sense, as the ultimate goal is to recover a solution to Problem 15 at the end of the algorithm. To do this, the authors suggest a procedure based on sketching: Two random matrices $\Omega \in \mathbb{R}^{d\times j}$ and $\Psi \in \mathbb{R}^{l\times n}$ are generated such that j and l are of the same order as the rank r of the solution to Problem 15, and two new matrices Y and W are obtained as $Y = X\Omega$ and $W = \Psi X$. Note that the matrices (Y, W) typically have smaller rank than the iterates X_k that they represent. Any update that is made in the algorithm is then reflected on (Y, W) rather than on the matrix X. At the end of the algorithm, one can reconstruct a rank r solution to the problem from the updated pair (Y, W) (for details, see 87). Note that in this approach, one never needs to store or utilize high-rank intermediate iterates.

4. SCALABILITY VIA THE ALTERNATING-DIRECTION METHOD OF MULTIPLIERS AND THE AUGMENTED LAGRANGIAN METHOD

An attractive alternative to using interior-point methods for solving SDPs is to use first-order methods. Compared with interior-point methods, first-order methods can scale to significantly larger problem sizes while trading off the accuracy of the resulting output (more precisely, first-order methods can require more time to achieve accuracy similar to that of interior-point methods). First-order methods are thus particularly attractive for applications that demand scalability but do not require extremely accurate feasible solutions (e.g., some machine learning applications highlighted below). Here, we describe two recent approaches that involve first-order methods: one based on the alternating-direction method of multipliers (ADMM) (89) and one based on the augmented Lagrangian method (90, 91).

4.1. Solving Semidefinite Programs Using the Alternating-Direction Method of Multipliers

In Reference 89, the authors present a first-order method for solving general cone programs, with SDPs as a special case. Here, we first describe the basic idea behind ADMM and then describe how it is applied to SDPs; Reference 92 provides a more thorough exposition of ADMM.

Consider a convex optimization problem of the form

$$\begin{aligned} \min_{x,z} \quad & f(x) + g(z) \\ \text{s.t.} \quad & x = z. \end{aligned} \tag{16.}$$

Here, the convex functions f and g can be nonsmooth and take on infinite values (e.g., to encode additional constraints). ADMM is a method for solving problems of the form of Problem 16 (more generally, ADMM can solve problems where x and z have an affine relationship). ADMM operates by iteratively updating solutions x^k and z^k and a dual variable λ^k (associated with the constraint $x = z$) via the following steps:

1. $x^{k+1} = \underset{x}{\text{argmin}} \ (f(x) + \frac{\rho}{2}\|x - z^k - \lambda^k\|_2^2)$,
2. $z^{k+1} = \underset{z}{\text{argmin}} \, (g(z) + \frac{\rho}{2}\|x^{k+1} - z - \lambda^k\|_2^2)$,
3. $\lambda^{k+1} = \lambda^k - x^{k+1} + z^{k+1}$,

where $\rho > 0$ is a parameter. The initializations z^0 and λ^0 are usually taken to be zero but can be arbitrary. Under mild technical conditions (see 92), the steps above converge. In particular, the cost $f(x^k) + g(z^k)$ converges to the optimal value of Problem 16, the residual $x^k - z^k$ converges to zero, and λ^k converges to an optimal dual variable.

In Reference 89, the general ADMM procedure described above is applied to SDPs. A key idea is to apply ADMM to the homogeneous self-dual embedding of the primal SDP (Problem 1) and its dual (Problem 2). The homogeneous self-dual embedding converts the primal–dual pair of optimization problems into a single convex feasibility problem. This is done by embedding the Karush–Kuhn–Tucker (KKT) optimality conditions associated with the primal and dual SDPs into a single system of equations and semidefinite constraints. If the original primal–dual pair has a feasible solution, then it can be recovered from a solution to the embedding. If, on the other hand, the original primal–dual pair is not solvable, then the embedding allows one to produce a certificate of infeasibility. The homogeneous self-dual embedding is commonly used in interior-point methods (93, 94). Reference 89 provides more details on this technique.

To summarize, the approach presented in Reference 89 consists of the following three steps:

1. Form the homogeneous self-dual embedding associated with the primal–dual SDP pair.
2. Express this embedding in the form of Problem 16 required by ADMM.
3. Apply the ADMM steps to this problem.

In Reference 89, the authors also present techniques for efficiently implementing step 3, including methods for efficiently performing the projections required to implement the ADMM steps, scaling or preconditioning the problem data to improve convergence, and choosing stopping criteria. The resulting approach achieves significant speedups as compared with interior-point methods for a number of large-scale conic optimization problems (with some loss in the accuracy of the solution). As an example, Reference 89 considers the problem of robust principal component analysis (95) in machine learning. This problem corresponds to recovering the principal components of a data matrix even when a fraction of its entries are arbitrarily corrupted by noise. More specifically, suppose one is given a (large) data matrix M that is known to be decomposable as

$$M = L + S,$$

where L is a low-rank matrix and S is sparse (but nothing further is known, e.g., the locations or number of nonzero entries of S). Interestingly, this problem can be solved using semidefinite programming under relatively mild conditions (see 95). For this problem, the ADMM approach

presented in Reference 89 is able to tackle large-scale instances where standard interior-point solvers [such as SeDuMi (93) and SDPT3 (96)] run out of memory. Moreover, the approach provides significant (order of magnitude) speedups on smaller instances while resulting in only a small loss in accuracy (less than 3×10^{-4} reconstruction error of the low-rank matrix L).

4.2. Augmented Lagrangian Methods

Next, we discuss two closely related approaches for solving SDPs using augmented Lagrangian methods: SDPNAL (a Newton-CG augmented Lagrangian method for solving semidefinite programming; see 90) and SDPNAL+ (see 91). SDPNAL operates on Problem 2 by defining an augmented Lagrangian function $L_\sigma : \mathbb{R}^m \times S_n \rightarrow \mathbb{R}$ as

$$L_\sigma(y, X) = -b^T y + \frac{1}{2\sigma}\left(\|\Pi\left(X - \sigma\left(C - \sum_{i=1}^{m} y_i A_i\right)\right)\|^2 + \|X\|^2\right), \qquad 17.$$

where $\sigma > 0$ is a penalty parameter, $\|\cdot\|$ corresponds to the Frobenius norm, and $\Pi(\cdot)$ is the metric projection operator onto the set S_n^+ of $n \times n$ symmetric psd matrices. More precisely, $\Pi(Y)$ is the (unique) optimal solution to the following convex problem:

$$\min_{Z \in S_n^+} \frac{1}{2}\|Z - Y\|. \qquad 18.$$

The augmented Lagrangian function is continuously differentiable [since $\|\Pi(\cdot)\|^2$ is continuously differentiable] and may be viewed as the (usual) Lagrangian function associated with a regularized version of Problem 2 (for a more thorough exposition, see 92, 97).

The augmented Lagrangian method iteratively updates solutions to the dual SDP shown in Problem 2 and the primal SDP shown in Problem 1 using the following steps:

1. $y^{k+1} = \underset{y}{\operatorname{argmin}}\, L_{\sigma_k}(y, X^k)$,
2. $X^{k+1} = \Pi(X^k - \sigma_k(C - \sum_{i=1}^{m} y_i A_i))$,
3. $\sigma_{k+1} = \rho\sigma_k$ for $\rho \geq 1$.

To implement this method, we thus require the ability to solve the inner optimization problems in steps 1 and 2 above. The objective functions in these inner optimization problems are convex and continuously differentiable but not twice continuously differentiable (see 90). The SDPNAL method presented in Reference 90 works by applying a semismooth Newton method combined with the conjugate gradient method to approximately solve the inner problems. Conditions under which the resulting iterates converge to the optimal solution were established in Reference 90. In Reference 91, the authors present SDPNAL+, which builds on the approach in Reference 90. However, for certain kinds of SDPs (degenerate SDPs; for a formal definition, see 91), SDPNAL can encounter numerical difficulties. SDPNAL+ tackles this issue by directly working with the primal SDP and using a different technique to solve the inner optimization problems that arise from the augmented Lagrangian method (specifically, a majorized semismooth Newton method).

In Reference 91, the authors compare the performance of SDPNAL+ with that of two other first-order methods on various numerical examples. In particular, they consider SDP relaxations for the machine learning problem of k-means clustering. The goal here is to partition a given set of data points into k clusters such that the total sum-of-squared Euclidean distances from each data point to its assigned cluster center is minimized. In Reference 98, the authors present semidefinite relaxations for finding approximate solutions to this NP-hard problem. The numerical experiments presented in Reference 91 demonstrate that SDPNAL+ is able to solve the resulting SDPs

with higher accuracy than other first-order methods while also resulting in faster running times (e.g., order-of-magnitude gains in running time on large instances of the clustering problem).

5. TRADING OFF SCALABILITY WITH CONSERVATISM

In this section, we discuss approaches that afford gains in scalability but are potentially conservative. We use the word conservative to refer to guaranteed feasible solutions that may be suboptimal (in contrast to approaches that may produce points that slightly violate some constraints). In particular, here we discuss approaches that provide the user with a tuning parameter for trading off scalability with conservatism. Conceptually, we classify such approaches into two groups: special-purpose approaches that leverage domain knowledge associated with the target application, and general-purpose approaches that apply broadly across application domains.

As an example of an approach that falls under the first category, we highlight recent work on neural network verification using semidefinite programming. Over the last few years, a large body of work has explored the susceptibility of neural networks to adversarial inputs (or, more generally, uncertainty in the inputs) (99–102). For example, in the context of image classification, neural networks can be made to switch their classification labels by adding very small perturbations (imperceptible to humans) to the input image. Motivated by the potentially serious consequences of such fragilities in safety-critical applications (e.g., autonomous cars), there has been a recent explosion of work on verifying the robustness of neural networks to perturbations to the input (103–108; for a more comprehensive literature review, see 105). Recently, approaches based on semidefinite programming have been proposed to tackle this challenge (109–111). The challenge of scalability for this application is particularly pronounced since the number of decision variables in the resulting SDPs grows with the number of neurons in the network. In Reference 111, the authors propose an approach for verifying neural networks by capturing their input–output relationships using quadratic constraints. This technique allows one to certify that, for a given set of inputs (e.g., a set of images obtained by making small perturbations to a training image), the output is contained within a specified set (e.g., outputs that assign the same label). The authors consider neural networks with different types of activation functions [rectified linear activation unit (ReLU), sigmoid, etc.] and propose different sets of quadratic constraints for each one. Importantly, by including or excluding different kinds of quadratic constraints, the approach allows one to trade off scalability with conservatism. Moreover, by exploiting the modular structure of neural networks, the approach scales to neural networks with approximately 5,000 neurons. The scalability afforded by the modular approach, however, comes at the cost of conservatism.

Next, we highlight approaches that fall under the second category mentioned above: general-purpose approaches for trading off scalability with conservatism.

5.1. DSOS and SDSOS Optimization

In Reference 8, the authors introduce diagonally dominant sum of squares (DSOS) and scaled diagonally dominant sum of squares (SDSOS) optimization as linear-programming- and second-order-cone-programming-based alternatives to SOS and semidefinite optimization that allow for a trade-off between computation time and solution quality. The following definitions are central to this framework.

Definition 4 (dd and sdd matrices). A symmetric matrix $A = (a_{ij})$ is diagonally dominant (dd) if

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|$$

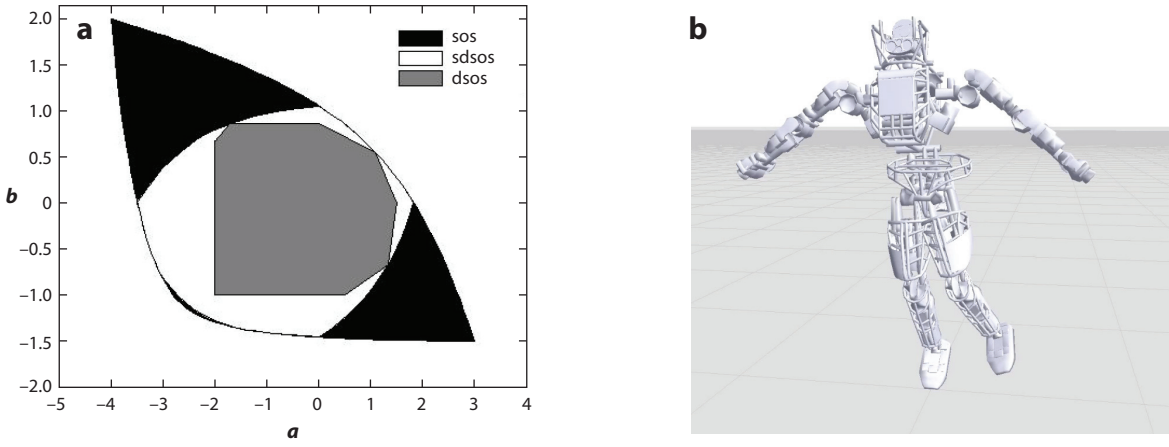

Figure 1

Diagonally dominant sum of squares (dsos) and scaled diagonally dominant sum of squares (sdsos) polynomials form structured subsets of sum of squares (sos) polynomials, which can prove useful when optimization over sos polynomials is too expensive. (*a*) The set of parameters a and b for which the polynomial $p(x_1, x_2) = x_1^4 + x_2^4 + ax_1^3x_2 + (1 - \frac{1}{2}a - \frac{1}{2}b)x_1^2x_2^2 + 2bx_1x_2^3$ is dsos, sdsos, or sos. Panel adapted from Reference 8. Copyright © 2019 Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved. (*b*) Use of sdsos optimization to balance a model of a humanoid robot with 30 states and 14 control inputs on one foot. Atlas® robot image provided courtesy of Boston Dynamics, Inc.

for all i. A symmetric matrix A is scaled diagonally dominant (sdd) if there exists a diagonal matrix D, with positive diagonal entries, such that DAD is dd.

Definition 5 (dsos and sdsos polynomials). A polynomial $p := p(x)$ is said to be diagonally dominant sum of squares (dsos) if it can be written as $p(x) = \sum_i \alpha_i q_i^2(x)$ for some scalars $\alpha_i \geq 0$ and some polynomials $q_i(x)$ that each involve at most two monomials with a coefficient of ± 1. We say that a polynomial p is scaled diagonally dominant sum of squares (sdsos) if it can be written as $p(x) = \sum_i q_i^2(x)$ for some polynomials q_i that involve at most two monomials with an arbitrary coefficient.

The implications dsos $\Rightarrow$ sdsos $\Rightarrow$ sos readily follow; **Figure 1*a*** shows a comparison of the three notions on a parametric family of bivariate quartic polynomials. Similarly, in view of Gershgorin's circle theorem (112), the implications dd $\Rightarrow$ sdd $\Rightarrow$ psd are straightforward to establish. In Reference 8, the authors connect the above definitions via the following statement.

Theorem 1. A polynomial p of degree $2d$ is dsos or sdsos if and only if it admits a representation as $p(x) = z^T(x)Qz(x)$, where $z(x)$ is the standard monomial vector of degree $\leq d$ and Q is a dd or sdd matrix, respectively.

By combining this theorem with some linear algebraic observations, the authors show in Reference 8 that optimization of a linear objective function over the intersection of the cone of dsos or sdsos polynomials of a given degree with an affine subspace can be carried out via linear programming or second-order cone programming, respectively. These are two mature classes of convex optimization problems that can be solved significantly faster than SDPs. The linear programs (LPs) and second-order cone programs (SOCPs) that arise from dsos and sdsos constraints on polynomials are termed DSOS and SDSOS optimization problems, respectively; they are used to quickly produce feasible, but possibly suboptimal, solutions to SOS optimization problems. In the special case where the polynomials involved have degree two, DSOS and SDSOS optimization

problems can be used to produce approximate solutions to SDPs. We also remark that, in applications where SOS programming is used as a relaxation—i.e., to provide an outer approximation of a (typically intractable) feasible set—this approach replaces the semidefinite constraint with a membership constraint in the dual of the cone of dsos or sdsos polynomials (for more details, see, e.g., section 3.4 of Reference 113).

5.1.1. Impact on applications. A key practical benefit of the DSOS/SDSOS approach is that it can be used as a plug-in in any application area of SOS programming. The software package (see Section 6) that accompanies Reference 8 facilitates this procedure. Section 4 of Reference 8 shows via numerical experiments from diverse application areas—polynomial optimization, statistics and machine learning, derivative pricing, and control theory—that with reasonable trade-offs in accuracy, one can achieve noticeable speedups and handle problems at scales that are currently beyond the reach of traditional SOS approaches. (Reference 8 also includes comparisons with SDP solvers such as MOSEK, SeDuMi, and SDPNAL+.) For the problem of sparse principal component analysis in machine learning, for example, the experiments in Reference 8 show that the SDSOS approach is more than 1,000 times faster than the standard SDP approach on 100 × 100 input instances, while sacrificing only approximately 2–3% in optimality (see section 4.5 of Reference 8). As another example, **Figure 1*b*** illustrates a humanoid robot with 30 state variables and 14 control inputs that SDSOS optimization is able to stabilize on one foot (114). A nonlinear control problem of this scale is beyond the reach of standard solvers for SOS programs at the moment. In a different paper (115), the authors show the potential of DSOS and SDSOS optimization for real-time applications. More specifically, they use these techniques to compute, every few milliseconds, certificates of collision avoidance for a simple model of an unmanned vehicle that navigates through a cluttered environment.

5.1.2. Some guarantees of the DSOS/SDSOS approach. On the theoretical front, DSOS and SDSOS optimization enjoys some of the same guarantees as SOS optimization. For example, classical theorems in algebraic geometry can be utilized to conclude that any even positive definite homogeneous polynomial is the ratio of two dsos polynomials (section 3.2 of Reference 8). From this observation alone, one can design a hierarchy of linear programming relaxations that can solve any copositive program (116) to arbitrary accuracy (section 4.2 of Reference 8). This idea can be extended to achieve the same result for any polynomial optimization problem with a compact feasible set (see section 4.2 of Reference 117).

5.2. Adaptive Improvements to DSOS and SDSOS Optimization

While DSOS and SDSOS techniques result in significant gains in terms of solving times and scalability, they inevitably lead to some loss in solution accuracy when compared with the SOS approach. In this subsection, we briefly outline two possible strategies to mitigate this loss. These strategies solve a sequence of adaptive LPs or SOCPs that inner approximate the feasible set of a SOS program in a direction of interest. For brevity of exposition, we explain how the strategies can be applied to approximate the generic SDP given in Problem 1. A treatment tailored to the case of SOS programs can be found in the references we provide.

5.2.1. Iterative change of basis. In Reference 113, the authors build on the notions of diagonal and scaled diagonal dominance to provide a sequence of improving inner approximations to the cone P_n of psd matrices in the direction of the objective function of an SDP at hand. The idea is simple: define a family of cones

$$DD(U) := \{M \in S_n \mid M = U^T Q U \text{ for some dd matrix } Q\},$$

parameterized by an $n \times n$ matrix U. [One can think of $DD(U)$ as the set of matrices that are dd after a change of coordinates via the matrix U.] Optimizing over the set $DD(U)$ is an LP since U is fixed, and the defining constraints are linear in the coefficients of the two unknown matrices M and Q. Furthermore, the matrices in $DD(U)$ are all psd—i.e., $\forall\, U$, $DD(U) \subseteq P_n$.

The proposal in Reference 113 is to solve a sequence of LPs, indexed by k, by replacing the condition $X \succeq 0$ with $X \in DD(U_k)$:

$$\begin{aligned} DSOS_k := \min_{X \in S_n} \quad & \mathrm{Tr}(CX) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X \in DD(U_k). \end{aligned} \tag{19.}$$

The sequence of matrices $\{U_k\}$ is defined as

$$\begin{aligned} U_0 &= I, \\ U_{k+1} &= \mathrm{Chol}(X_k), \end{aligned} \tag{20.}$$

where X_k is an optimal solution to the LP in Equation 19, and $\mathrm{Chol}(\cdot)$ denotes the Cholesky decomposition of a matrix (this can also be replaced with the matrix square root operation).

Note that the first LP in the sequence optimizes over the set of dd matrices. Defining U_{k+1} as a Cholesky factor of X_k guarantees improvement of the optimal value in each iteration. Indeed, since $X_k = U_{k+1}^T I U_{k+1}$ and the identity matrix I is dd, we see that $X_k \in DD(U_{k+1})$ and hence is feasible for iteration $k+1$. This implies that the optimal value at iteration $k+1$ is at least as good as the optimal value at the previous iteration—i.e., $DSOS_{k+1} \leq DSOS_k$ (in fact, the inequality is strict under mild assumptions; see Reference 113).

In an analogous fashion, one can construct a sequence of SOCPs that inner approximate P_n with increasing quality. This time, the authors define a family of cones

$$SDD(U) := \{M \in S_n \mid M = U^T Q U, \text{ for some sdd matrix } Q\},$$

parameterized again by an $n \times n$ matrix U. For any U, optimizing over the set $SDD(U)$ is an SOCP, and we have $SDD(U) \subseteq P_n$. This leads us to the following iterative SOCP sequence:

$$\begin{aligned} SDSOS_k := \min_{X \in S_n} \quad & \mathrm{Tr}(CX) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X \in SDD(U_k). \end{aligned} \tag{21.}$$

Assuming the existence of an optimal solution X_k at each iteration, we can define the sequence $\{U_k\}$ iteratively in the same way as was done in Equation 20. Using similar reasoning, we have $SDSOS_{k+1} \leq SDSOS_k$. In practice, the sequence of upper bounds $\{SDSOS_k\}$ approaches the SDP optimal value faster than the sequence of the LP upper bounds $\{DSOS_k\}$.

Figure 2 shows the improvement (in every direction) obtained by just a single iteration of this approach. Note that the SOCP in particular fills up almost the entire spectrahedron in a single iteration.

5.2.2. Column generation. In Reference 118, the authors design another iterative method for inner approximating the set of psd matrices using linear and second-order cone programming. Their approach combines DSOS and SDSOS techniques with ideas from the theory of column generation in large-scale linear and integer programming. The high-level idea is to approximate

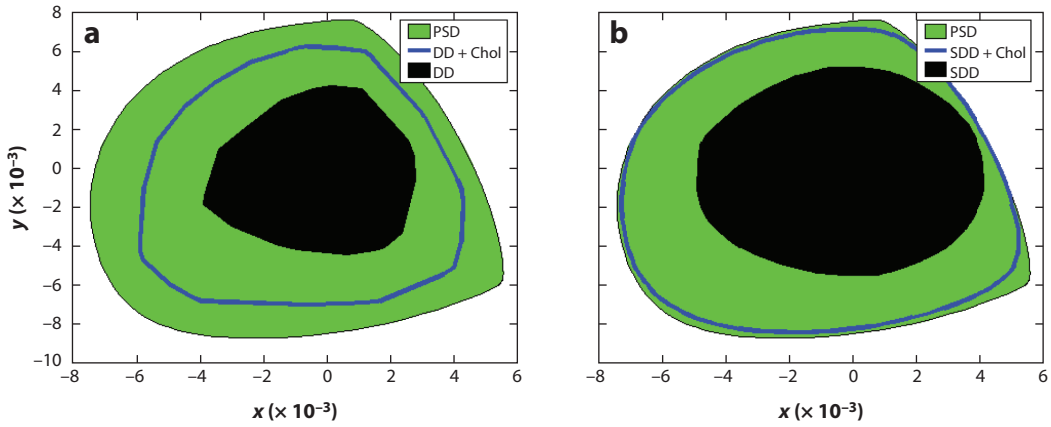

Figure 2

Improvements (in all directions) obtained after one iteration of the change-of-basis algorithm on (*a*) the diagonally dominant (DD) approximation and (*b*) the scaled diagonally dominant (SDD) approximation of a randomly generated semidefinite program (SDP). In each panel, the outer set in green shows the feasible set of the randomly generated SDP. The sets in black are the DD and SDD inner approximations, and the blue line shows the boundary of the improved inner approximation after one iteration. Additional abbreviations: Chol, Cholesky decomposition; PSD, positive semidefinite. Figure adapted from Reference 113, published by the American Mathematical Society.

the SDP in Problem 1 by a sequence of LPs (parameterized by t):

$$\begin{aligned} \min_{X \in S_n, \alpha_i} \quad & \mathrm{Tr}(CX) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X = \sum_{i=1}^{t} \alpha_i B_i, \\ & \alpha_i \geq 0, \quad i = 1, \ldots, t, \end{aligned} \tag{22.}$$

where $B_1, \ldots, B_t$ are fixed psd matrices. These matrices are initialized to be the extreme rays of $n \times n$ dd matrices, which turn out to be all rank-1 matrices $v_i v_i^T$, where the vector v_i has at most two nonzero components, each equal to ± 1 (119). Once this initial LP is solved, one adds a new psd matrix B_j (or sometimes several such matrices) to Problem 22 and resolves. This process then continues. In each step, the new matrices B_j are picked carefully to bring the optimal value of the LP closer to that of the SDP. Usually, the construction of B_j involves solving a pricing subproblem (in the terminology of the column generation literature), which adds appropriate cutting planes to the dual of Problem 22 (for details, see 118).

The SOCP analog of this process is similar. The SDP in Problem 1 is inner approximated by a sequence of SOCPs (parameterized by t):

$$\begin{aligned} \min_{X \in S_n, \Lambda_i \in S_2} \quad & \mathrm{Tr}(CX) \\ \text{s.t.} \quad & \mathrm{Tr}(A_i X) = b_i, \quad i = 1, \ldots, m, \\ & X = \sum_{i=1}^{t} V_i \Lambda_i V_i^T, \\ & \Lambda_i \succeq 0, \quad i = 1, \ldots, t, \end{aligned} \tag{23.}$$

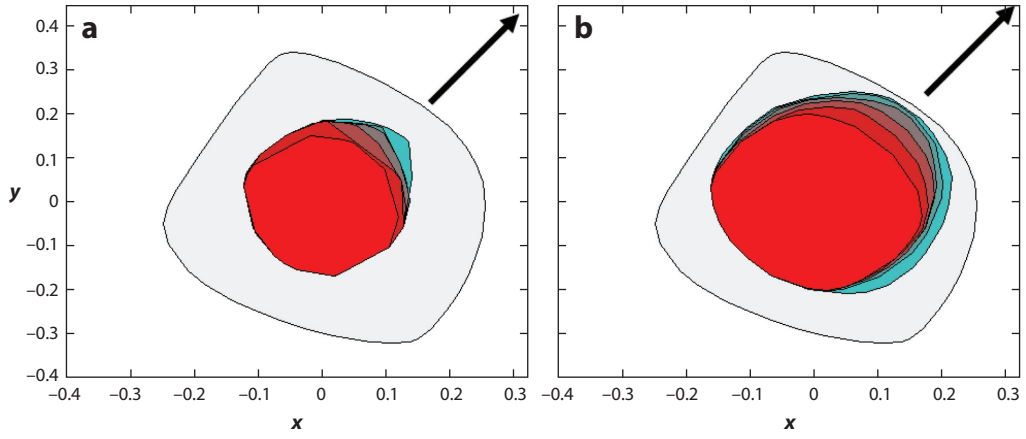

Figure 3

Successive improvements (in the northeast direction) obtained by applying five iterations of the column generation method to (*a*) the diagonally dominant (DD) inner approximation and (*b*) the scaled diagonally dominant (SDD) inner approximation of a randomly generated semidefinite program (SDP). In each panel, the outer gray set is the feasible set of the randomly generated SDP, and the colored sets show the inner approximations at each iteration. Figure adapted from Reference 118 with permission from Elsevier.

where $V_1, \ldots, V_t$ are fixed $n \times 2$ matrices. They are initialized as the set of matrices that have zeros everywhere except for a 1 in the first column in position j and a 1 in the second column in position $k \neq j$. This gives exactly the set of $n \times n$ sdd matrices (8, lemma 3.8]. In subsequent steps, one appropriately chosen matrix V_i (or sometimes several matrices) are added to Problem 23. These matrices are obtained by solving pricing subproblems and help bring the optimal value of the SOCP closer to that of the SDP in each iteration.

Figure 3 shows the improvement obtained by five iterations of this process on a randomly generated SDP, where the objective is to maximize a linear function in the northeast direction.

6. SOFTWARE: SOLVERS AND MODELING LANGUAGES

In this section, we list software tools that allow a user to specify and solve SDPs. This list is not meant to be exhaustive; rather, the goal is to provide the interested practitioner with a starting point in terms of software for implementing some of the approaches reviewed in this article:

- Software for specifying SDPs: YALMIP (120), CVXPY (121), PICOS (122), SPOT (123), and JuMP (124). These software packages allow one to easily specify SDPs and provide interfaces to various solvers mentioned above. YALMIP and SPOT are MATLAB packages, CVXPY and PICOS are Python packages, and JuMP is a Julia package.
- General-purpose SDP solvers that implement interior-point methods: MOSEK (94), SeDuMi (93), and SDPT^3 (96). The MOSEK solver provides an interface with MATLAB, C, Python, and Java. SeDuMi and SDPT^3 interface with MATLAB.
- Solvers based on ADMM and augmented Lagrangian methods (see Section 4): SDPNAL/SDPNAL+ (91) and SCS (125). SDPNAL and SDPNAL+ provide a MATLAB interface, while SCS provides interfaces to C, C++, Python, MATLAB, R, and Julia.
- Software for Riemannian optimization (see Section 3.1): Manopt (76) and Pymanopt (126). Manopt is a MATLAB-based toolbox for optimization on manifolds, while Pymanopt is Python based.

- Software for exploiting chordal sparsity in SDPs (see Section 2.1): SparseCoLO (28) and CDCS (31). SparseCoLO is a MATLAB toolbox for converting SDPs with chordal sparsity to equivalent SDPs with smaller-sized SDP constraints. CDCS is a solver that exploits chordal sparsity and provides a MATLAB interface.
- Software for exploiting structure via facial reduction (see Section 2.3): frlib (127). This is a MATLAB toolbox and interfaces with the YALMIP and SPOT packages mentioned above.
- Parsers for SOS programs (see Section 1.4): YALMIP (120), SPOT (123), SOSTOOLS (128), and SumofSquares.jl (129). These packages allow one to specify SOS programs and convert them to a form that can be solved using SDP solvers. YALMIP, SPOT, and SOSTOOLS are MATLAB packages, while SumofSquares.jl is a Julia package.
- Software for parsing DSOS and SDSOS programs (see Section 5.1): SPOT (123). This MATLAB toolbox allows one to parse DSOS and SDSOS programs. SPOT provides interfaces to LP and SOCP solvers, including MOSEK. The appendix of Reference 8 provides a tutorial on solving DSOS and SDSOS programs using SPOT.

7. CONCLUSIONS

Semidefinite programming is an exciting and active area of research with a vast number of applications in fields including machine learning, control, and robotics. Historically, the lack of scalability of semidefinite programming has been a major impediment to fulfilling the potential that it brings to these application domains. In this article, we have reviewed recent approaches for addressing this challenge, including techniques that exploit structure such as sparsity and symmetry in a problem, approaches that produce low-rank feasible solutions to SDPs, methods for solving SDPs via augmented Lagrangian and ADMM techniques, and approaches that trade off scalability with conservatism, such as techniques for approximating SDPs with LPs or SOCPs. We have also provided a list of software packages associated with these approaches. Our hope is that this article will serve as an entry point for both practitioners working in application domains that demand scalability and researchers seeking to contribute to theory and algorithms for scalable semidefinite programming.

Significant work remains to be done to build on the advancements reviewed here. Semidefinite programming is still far from being a mature technology like linear or quadratic programming. Potential directions for future work include (*a*) gaining a better theoretical understanding of the convergence properties of algorithms for low-rank SDPs (see Section 3); (*b*) identifying structure beyond chordal sparsity, symmetry, and degeneracy (see Section 2) that arises in practice and methods for exploiting such structure; (*c*) exploring different first-order methods for solving SDPs and understanding their relative merits (see Section 4); (*d*) understanding the power of LPs and SOCPs for approximating SDPs in an adaptive (as opposed to one-shot) fashion (see Section 5); (*e*) finding ways to combine the different approaches for improving scalability reviewed in this article; and (*f*) further developing mature software that allows practitioners to deploy semidefinite programming technology on future applications, including ones that involve solving SDPs in real time.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Cemil Dibek for his constructive feedback on the first draft of this article. This work was partially supported by a Defense Advanced Research Projects Agency (DARPA) Young Faculty

Award, a CAREER Award from the National Science Foundation (NSF), a Google Faculty Award, an Innovation Award from the School of Engineering and Applied Sciences at Princeton University, a Sloan Fellowship, a Multidisciplinary Research Program of the University Research Initiative (MURI) Award from the Air Force Office of Scientific Research, an Office of Naval Research award (N00014-18-1-2873), an NSF Computer and Information Science and Engineering (CISE) Research Initiation Initiative (CRII) award (IIS-1755038), and an Amazon Research Award.

LITERATURE CITED

1. Goemans MX, Williamson DP. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM* 42:1115–45
2. Lovász L. 1979. On the Shannon capacity of a graph. *IEEE Trans. Inf. Theory* 25:1–7
3. Recht B, Fazel M, Parrilo PA. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* 52:471–501
4. d'Aspremont A, Ghaoui LE, Jordan MI, Lanckriet GR. 2007. A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* 49:434–48
5. Hall G. 2018. *Optimization over nonnegative and convex polynomials with and without semidefinite programming*. PhD Thesis, Princeton Univ., Princeton, NJ
6. Lasserre JB. 2001. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.* 11:796–817
7. Parrilo PA. 2003. Semidefinite programming relaxations for semialgebraic problems. *Math. Progr.* 96:293–320
8. Ahmadi AA, Majumdar A. 2019. DSOS and SDSOS optimization: more tractable alternatives to sum of squares and semidefinite optimization. *SIAM J. Appl. Algebr. Geom.* 3:192–230
9. Parrilo PA. 2000. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD Thesis, Calif. Inst. Technol., Pasadena
10. Vandenberghe L, Boyd S. 1996. Semidefinite programming. *SIAM Rev.* 38:49–95
11. Renegar J. 2019. Accelerated first-order methods for hyperbolic programming. *Math. Progr.* 173:1–35
12. Arora S, Hazan E, Kale S. 2005. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *46th Annual IEEE Symposium on Foundations of Computer Science*, pp. 339–48. Piscataway, NJ: IEEE
13. Ding L, Yurtsever A, Cevher V, Tropp JA, Udell M. 2019. An optimal-storage approach to semidefinite programming using approximate complementarity. arXiv:1902.03373 [math.OC]
14. Henrion D, Malick J. 2011. Projection methods for conic feasibility problems: applications to polynomial sum-of-squares decompositions. *Optim. Methods Softw.* 26:23–46
15. Yurtsever A, Fercoq O, Cevher V. 2019. A conditional gradient-based augmented Lagrangian framework. arXiv:1901.04013 [math.OC]
16. Nie J, Wang L. 2012. Regularization methods for SDP relaxations in large-scale polynomial optimization. *SIAM J. Optim.* 22:408–28
17. Bertsimas D, Freund RM, Sun XA. 2013. An accelerated first-order method for solving SOS relaxations of unconstrained polynomial optimization problems. *Optim. Methods Softw.* 28:424–41
18. Hopkins SB, Schramm T, Shi J, Steurer D. 2016. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, pp. 178–91. New York: ACM
19. Sun J. 2019. Provable nonconvex methods/algorithms. **https://sunju.org/research/nonconvex**
20. Blekherman G, Parrilo PA, Thomas R, eds. 2013. *Semidefinite Optimization and Convex Algebraic Geometry*. Philadelphia: Soc. Ind. Appl. Math.
21. Hall G. 2019. Engineering and business applications of sum of squares polynomials. arXiv:1906.07961 [math.OC]
22. Putinar M. 1993. Positive polynomials on compact semi-algebraic sets. *Indiana Univ. Math. J.* 42:969–84
23. Agler J, Helton W, McCullough S, Rodman L. 1988. Positive semidefinite matrices with a given sparsity pattern. *Linear Algebra Appl.* 107:101–49

24. Zheng Y, Mason RP, Papachristodoulou A. 2017. Scalable design of structured controllers using chordal decomposition. *IEEE Trans. Autom. Control* 63:752–67
25. Grone R, Johnson CR, Sá EM, Wolkowicz H. 1984. Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* 58:109–24
26. Fukuda M, Kojima M, Murota K, Nakata K. 2001. Exploiting sparsity in semidefinite programming via matrix completion I: general framework. *SIAM J. Optim.* 11:647–74
27. Kim S, Kojima M, Mevissen M, Yamashita M. 2011. Exploiting sparsity in linear and nonlinear matrix inequalities via positive semidefinite matrix completion. *Math. Progr.* 129:33–68
28. Fujisawa K, Kim S, Kojima M, Okamoto Y, Yamashita M. 2009. *User's manual for SparseCoLO: conversion methods for sparse conic-form linear optimization problems*. Res. Rep. B-453, Dep. Math. Comput. Sci., Tokyo Inst. Technol., Tokyo
29. Mason RP, Papachristodoulou A. 2014. Chordal sparsity, decomposing SDPs and the Lyapunov equation. In *2014 American Control Conference*, pp. 531–37. Piscataway, NJ: IEEE
30. Zheng Y, Mason RP, Papachristodoulou A. 2016. A chordal decomposition approach to scalable design of structured feedback gains over directed graphs. In *2016 IEEE 55th IEEE Conference on Decision and Control*, pp. 6909–14. Piscataway, NJ: IEEE
31. Zheng Y, Fantuzzi G, Papachristodoulou A, Goulart P, Wynn A. 2020. Chordal decomposition in operator-splitting methods for sparse semidefinite programs. *Math. Progr.* 180:489–532
32. Burer S. 2003. Semidefinite programming in the space of partial positive semidefinite matrices. *SIAM J. Optim.* 14:139–72
33. Andersen MS, Dahl J, Vandenberghe L. 2010. Implementation of nonsymmetric interior-point methods for linear optimization over sparse matrix cones. *Math. Progr. Comput.* 2:167–201
34. Sun Y, Andersen MS, Vandenberghe L. 2014. Decomposition in conic optimization with partially separable structure. *SIAM J. Optim.* 24:873–97
35. Kalbat A, Lavaei J. 2015. A fast distributed algorithm for decomposable semidefinite programs. In *2015 54th IEEE Conference on Decision and Control*, pp. 1742–49. Piscataway, NJ: IEEE
36. Madani R, Kalbat A, Lavaei J. 2015. ADMM for sparse semidefinite programming with applications to optimal power flow problem. In *2015 54th IEEE Conference on Decision and Control*, pp. 5932–39. Piscataway, NJ: IEEE
37. Sun Y, Vandenberghe L. 2015. Decomposition methods for sparse matrix nearness problems. *SIAM J. Matrix Anal. Appl.* 36:1691–717
38. Andersen MS, Pakazad SK, Hansson A, Rantzer A. 2014. Robust stability analysis of sparsely interconnected uncertain systems. *IEEE Trans. Autom. Control* 59:2151–56
39. Benson SJ, Ye Y. 2006. *DSDP5 user guide—software for semidefinite programming*. Tech. Rep. ANL/MCS-TM-277, Argonne Natl. Lab., Lemont, IL
40. Weisser T, Lasserre JB, Toh KC. 2018. Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity. *Math. Progr. Comput.* 10:1–32
41. Lasserre JB, Toh KC, Yang S. 2017. A bounded degree SOS hierarchy for polynomial optimization. *EURO J. Comput. Optim.* 5:87–117
42. Mangelson JG, Liu J, Eustice RM, Vasudevan R. 2018. Guaranteed globally optimal planar pose graph and landmark SLAM via sparse-bounded sums-of-squares programming. arXiv:1809.07744 [cs.RO]
43. Thrun S, Burgard W, Fox D. 2005. *Probabilistic Robotics*. Cambridge, MA: MIT Press
44. Rosen DM, Carlone L, Bandeira AS, Leonard JJ. 2019. SE-Sync: a certifiably correct algorithm for synchronization over the special Euclidean group. *Int. J. Robot. Res.* 38:95–125
45. Gatermann K, Parrilo PA. 2004. Symmetry groups, semidefinite programs, and sums of squares. *J. Pure Appl. Algebra* 192:95–128
46. Vallentin F. 2009. Symmetry in semidefinite programs. *Linear Algebra Appl.* 430:360–69
47. De Klerk E. 2010. Exploiting special structure in semidefinite programming: a survey of theory and applications. *Eur. J. Oper. Res.* 201:1–10
48. Permenter FN. 2017. *Reduction methods in semidefinite and conic optimization*. PhD Thesis, Mass. Inst. Technol., Cambridge
49. Henrion D, Garulli A, eds. 2005. *Positive Polynomials in Control*. Berlin: Springer

50. Cogill R, Lall S, Parrilo PA. 2008. Structured semidefinite programs for the control of symmetric systems. *Automatica* 44:1411–17
51. Arcak M, Meissen C, Packard A. 2016. *Networks of Dissipative Systems: Compositional Certification of Stability, Performance, and Safety*. Cham, Switz.: Springer
52. Drusvyatskiy D, Wolkowicz H. 2017. The many faces of degeneracy in conic optimization. *Found. Trends Optim.* 3:77–170
53. Borwein J, Wolkowicz H. 1981. Regularizing the abstract convex program. *J. Math. Anal. Appl.* 83:495–530
54. Pataki G. 2013. Strong duality in conic linear programming: facial reduction and extended duals. In *Computational and Analytical Mathematics*, ed. DH Bailey, HH Bauschke, P Borwein, F Garvan, M Théra, et al., pp. 613–34. New York: Springer
55. Kungurtsev V, Marecek J. 2018. A two-step pre-processing for semidefinite programming. arXiv:1806.10868 [math.OC]
56. Waki H, Ebihara Y, Sebe N. 2016. Reduction of SDPs in H∞ control of SISO systems and performance limitations analysis. In *2016 55th IEEE Conference on Decision and Control*, pp. 646–51. Piscataway, NJ: IEEE
57. Krislock N, Wolkowicz H. 2010. Explicit sensor network localization using semidefinite representations and facial reductions. *SIAM J. Optim.* 20:2679–708
58. Zhu Y, Pataki G, Tran-Dinh Q. 2019. Sieve-SDP: a simple facial reduction algorithm to preprocess semidefinite programs. *Math. Program. Comput.* 11:503–86
59. Barvinok AI. 1995. Problems of distance geometry and convex properties of quadratic maps. *Discrete Comput. Geom.* 13:189–202
60. Pataki G. 1998. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Math. Oper. Res.* 23:339–58
61. Lemon A, So AMC, Ye Y. 2016. Low-rank semidefinite programming: theory and applications. *Found. Trends Optim.* 2:1–156
62. Candès EJ, Plan Y. 2010. Matrix completion with noise. *Proc. IEEE* 98:925–36
63. Bennett J, Lanning S. 2007. The Netflix Prize. In *Proceedings of KDD Cup and Workshop 2007*, pp. 3–6. New York: ACM
64. Kulis B, Surendran AC, Platt JC. 2007. Fast low-rank semidefinite programming for embedding and clustering. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pp. 235–42. Proc. Mach. Learn. Res. Vol. 2. N.p.: PMLR
65. Burer S, Monteiro RD. 2003. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Math. Progr.* 95:329–57
66. Liu DC, Nocedal J. 1989. On the limited memory BFGS method for large scale optimization. *Math. Progr.* 45:503–28
67. Burer S, Monteiro RD. 2005. Local minima and convergence in low-rank semidefinite programming. *Math. Progr.* 103:427–44
68. Boumal N, Absil PA. 2011. RTRMC: a Riemannian trust-region method for low-rank matrix completion. In *Advances in Neural Information Processing Systems 24*, ed. J Shawe-Taylor, RS Zemel, PL Bartlett, F Pereira, KQ Weinberger, pp. 406–14. Red Hook, NY: Curran
69. Absil PA, Baker CG, Gallivan KA. 2007. Trust-region methods on Riemannian manifolds. *Found. Comput. Math.* 7:303–30
70. Absil PA, Mahony R, Sepulchre R. 2009. *Optimization Algorithms on Matrix Manifolds*. Princeton, NJ: Princeton Univ. Press
71. Journée M, Bach F, Absil PA, Sepulchre R. 2010. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM J. Optim.* 20:2327–51
72. Boumal N, Voroninski V, Bandeira A. 2016. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems 29*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 2757–65. Red Hook, NY: Curran
73. Boumal N, Voroninski V, Bandeira AS. 2018. Deterministic guarantees for Burer-Monteiro factorizations of smooth semidefinite programs. arXiv:1804.02008 [math.OC]

74. Boumal N, Absil PA, Cartis C. 2018. Global rates of convergence for nonconvex optimization on manifolds. *IMA J. Numer. Anal.* 39:1–33
75. Sato H, Iwai T. 2015. A new, globally convergent Riemannian conjugate gradient method. *Optimization* 64:1011–31
76. Boumal N, Mishra B, Absil PA, Sepulchre R. 2014. Manopt, a MATLAB toolbox for optimization on manifolds. *J. Mach. Learn. Res.* 15:1455–59
77. Mei S, Misiakiewicz T, Montanari A, Oliveira RI. 2017. Solving SDPs for synchronization and maxcut problems via the Grothendieck inequality. arXiv:1703.08729 [math.OC]
78. Sun J, Qu Q, Wright J. 2018. A geometric analysis of phase retrieval. *Found. Comput. Math.* 18:1131–98
79. Sun J, Qu Q, Wright J. 2016. Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inf. Theory* 63:885–914
80. Carlone L, Rosen DM, Calafiore G, Leonard JJ, Dellaert F. 2015. Lagrangian duality in 3D SLAM: verification techniques and optimal solutions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 125–32. Piscataway, NJ: IEEE
81. Carlone L, Calafiore GC, Tommolillo C, Dellaert F. 2016. Planar pose graph optimization: duality, optimal solutions, and verification. *IEEE Trans. Robot.* 32:545–65
82. Erdogdu MA, Ozdaglar A, Parrilo PA, Vanli ND. 2018. Convergence rate of block-coordinate maximization Burer-Monteiro method for solving large SDPs. arXiv:1807.04428 [math.OC]
83. Wang PW, Chang WC, Kolter JZ. 2017. The mixing method: coordinate descent for low-rank semidefinite programming. arXiv:1706.00476 [math.OC]
84. Frank M, Wolfe P. 1956. An algorithm for quadratic programming. *Naval Res. Logist. Q.* 3:95–110
85. Hazan E. 2008. Sparse approximate solutions to semidefinite programs. In *Latin American Symposium on Theoretical Informatics*, ed. ES Laber, C Bornstein, LT Nogueira, L Faria, pp. 301–16. Berlin: Springer
86. Gärtner B, Matousek J. 2012. *Approximation Algorithms and Semidefinite Programming*. Berlin: Springer
87. Yurtsever A, Udell M, Tropp JA, Cevher V. 2017. Sketchy decisions: convex low-rank matrix optimization with optimal storage. arXiv:1702.06838 [math.OC]
88. Freund RM, Grigas P, Mazumder R. 2017. An extended Frank–Wolfe method with in-face directions, and its application to low-rank matrix completion. *SIAM J. Optim.* 27:319–46
89. O'Donoghue B, Chu E, Parikh N, Boyd S. 2016. Conic optimization via operator splitting and homogeneous self-dual embedding. *J. Optim. Theory Appl.* 169:1042–68
90. Zhao XY, Sun D, Toh KC. 2010. A Newton-CG augmented Lagrangian method for semidefinite programming. *SIAM J. Optim.* 20:1737–65
91. Yang L, Sun D, Toh KC. 2015. SDPNAL+: a majorized semismooth Newton-CG augmented Lagrangian method for semidefinite programming with nonnegative constraints. *Math. Progr. Comput.* 7:331–66
92. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3:1–122
93. Sturm J. 2001. SeDuMi. *Lehigh University*. **http://sedumi.ie.lehigh.edu**
94. Mosek. 2019. Introducing the MOSEK Optimization Suite 9.0.105. *Mosek*. **https://docs.mosek.com/9.0/intro/index.html**
95. Candès EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? *J. ACM* 58:11
96. Toh KC, Tütüncü RH, Todd MJ. SDPT^3 - a MATLAB software package for semidefinite-quadratic-linear programming. *Carnegie Mellon University*. **http://www.math.cmu.edu/~reha/sdpt3.html**
97. Rockafellar RT. 1976. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.* 1:97–116
98. Peng J, Wei Y. 2007. Approximating K-means-type clustering via semidefinite programming. *SIAM J. Optim.* 18:186–205
99. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, et al. 2013. Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]
100. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy*, pp. 372–87. Piscataway, NJ: IEEE

101. Zheng S, Song Y, Leung T, Goodfellow I. 2016. Improving the robustness of deep neural networks via stability training. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4480–88. Piscataway, NJ: IEEE
102. Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. 2017. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 86–94. Piscataway, NJ: IEEE
103. Tjeng V, Xiao K, Tedrake R. 2017. Evaluating robustness of neural networks with mixed integer programming. arXiv:1711.07356 [cs.LG]
104. Wong E, Kolter JZ. 2017. Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv:1711.00851 [cs.LG]
105. Liu C, Arnon T, Lazarus C, Barrett C, Kochenderfer MJ. 2019. Algorithms for verifying deep neural networks. arXiv:1903.06758 [cs.LG]
106. Pulina L, Tacchella A. 2012. Challenging SMT solvers to verify neural networks. *AI Commun.* 25:117–35
107. Xiang W, Musau P, Wild AA, Lopez DM, Hamilton N, et al. 2018. Verification for machine learning, autonomy, and neural networks survey. arXiv:1810.01989 [cs.AI]
108. Wang S, Pei K, Whitehouse J, Yang J, Jana S. 2018. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 6367–77. Red Hook, NY: Curran
109. Raghunathan A, Steinhardt J, Liang PS. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31*, ed. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, pp. 10877–87. Red Hook, NY: Curran
110. Qin C, O'Donoghue B, Bunel R, Stanforth R, Gowal S, et al. 2019. Verification of non-linear specifications for neural networks. arXiv:1902.09592 [cs.LG]
111. Fazlyab M, Morari M, Pappas GJ. 2019. Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. arXiv:1903.01287 [math.OC]
112. Geršgorin SA. 1931. Über die Abgrenzung der Eigenwerte einer Matrix. *Bull. Acad. Sci. URSS Classe Sci. Math.* 6:749–54
113. Ahmadi AA, Hall G. 2017. Sum of squares basis pursuit with linear and second order cone programming. In *Algebraic and Geometric Methods in Discrete Mathematics*, ed. HA Harrington, M Omar, M Wright, pp. 27–53. Providence, RI: Am. Math. Soc.
114. Majumdar A, Ahmadi AA, Tedrake R. 2014. Control and verification of high-dimensional systems with DSOS and SDSOS programming. In *53rd IEEE Conference on Decision and Control*, pp. 394–401. Piscataway, NJ: IEEE
115. Ahmadi AA, Majumdar A. 2016. Some applications of polynomial optimization in operations research and real-time decision making. *Optim. Lett.* 10:709–29
116. Dür M. 2010. Copositive programming – a survey. In *Recent Advances in Optimization and Its Applications in Engineering*, ed. M Diehl, F Glineur, E Jarlebring, W Michiels, pp. 3–20. Berlin: Springer
117. Ahmadi AA, Hall G. 2019. On the construction of converging hierarchies for polynomial optimization based on certificates of global positivity. *Math. Oper. Res.* 44:1148–509
118. Ahmadi AA, Dash S, Hall G. 2017. Optimization over structured subsets of positive semidefinite matrices via column generation. *Discrete Optim.* 24:129–51
119. Barker G, Carlson D. 1975. Cones of diagonally dominant matrices. *Pac. J. Math.* 57:15–32
120. Löfberg J. 2004. YALMIP: a toolbox for modeling and optimization in MATLAB. In *2004 IEEE International Symposium on Computer Aided Control Systems Design*, pp. 284–89. Piscataway, NJ: IEEE
121. Diamond S, Boyd S. 2016. CVXPY: a Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* 17:83
122. Stahlberg M. 2018. PICOS. *GitLab*. **https://picos-api.gitlab.io/picos**
123. Posa M. 2013. Systems Polynomial Optimization Toolbox. *GitHub*. **https://github.com/mposa/spot**
124. Dunning I, Huchette J, Lubin M. 2017. JuMP: a modeling language for mathematical optimization. *SIAM Rev.* 59:295–320
125. O'Donoghue B, Chu E, Parikh N, Boyd S. 2017. SCS. *GitHub*. **https://github.com/cvxgrp/scs**

126. Townsend J, Koep N, Weichwald S. 2016. Pymanopt: a Python toolbox for optimization on manifolds using automatic differentiation. *J. Mach. Learn. Res.* 17:4755–59
127. Permenter F, Parrilo P. 2018. Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone. *Math. Progr.* 171:1–54
128. Papachristodoulou A, Anderson J, Valmorbida G, Prajna S, Seiler P, Parrilo PA. 2013. SOSTOOLS: sum of squares optimization toolbox for MATLAB. *Caltech*. **http://www.cds.caltech.edu/sostools**
129. JuliaOpt. 2019. Sum of squares programming for Julia. *GitHub*. **https://github.com/JuliaOpt/SumOfSquares.jl**