Semidefinite Optimization

Chris Godsil

July 28, 2016

Preface

A course in semidefinite optimization for fourth year undergraduate students and graduate students. The aim is to introduce the theory and its applications. We offer a quite brief treatment of interior point methods, and make no attempt to discuss implementations. We provide applications to graph theory, coding theory, geometry, quantum imformation.

Some of the prerequisites are as follows.

1. Be able to formulate, and understand the formulation of linear programs.

2. Understand duality of linear programs—be able to write down the dual of a linear program and understand what it means.

3. Analysis: open sets, closed sets, compactness.

4. Linear algebra: all of it. Also norms on vector spaces and convex sets.

5. Graph theory: the basics, cliques, cocliques/independent sets, colouring.

Contents

Chapter 1

Basics

Roughly speaking, we're following Laurent and Vallentin for a while.

1.1 LPs and Inner Products

Discuss key properties of feasible regions in optimization problems.

Write LPs in terms of inner products: instead of

$$\max c^T x, \quad Ax = b, \quad x \geq 0,$$

rather, if A is $m \times n$ and A_i is the i-th row of A, then

$$\max \langle c, x \rangle, \quad \langle A_i, x \rangle = b_i, \ (i = 1, \dots m), \quad x \geq 0$$

Inner products on matrices. Change to upper case to get a semidefinite program:

$$\max \langle C, X \rangle, \quad \langle A_i, X \rangle = b_i, \ (i = 1, \dots m), \quad X \succcurlyeq 0.$$

Some explanations are in order. See the following sections in this chapter.

1.2 Inner Products & Norms

A *bilinear form* on a vector space V over $\mathbb{F}$ is a map $\beta : V \times V \to \mathbb{F}$ that is linear in each variable. It is symmetric if

$$\beta(u, v) = \beta(v, u)$$

for all u and v. It is *non-degenerate* if whenever $\beta(a,u) = 0$ for all u, then $a = 0$.

The simplest examples arise when $V = \mathbb{R}^d$ and

$$\beta(u,v) = u^T M v$$

for some $d \times d$ matrix M. This form is bilinear, and it is symmetric if and only if M is symmetric. Note that if $f_1, \ldots, f_d$ is a basis for V, the d^2 values $\beta(f_i, f_j)$ determine the form. In fact if we stack these entries into a $d \times d$ matrix M in the obvious way, and if $[u]$ denotes the coordinates of u relative to the given basis, then

$$\beta(u,v) = [u]^T M[v].$$

A bilinear form is non-degenerate if and only if the associated matrix M is invertible.

A bilinear form β on a real vector space is an *inner product* if it is symmetric and

(a) $\beta(x,x) \geq 0$ for all x in V.

(b) If $\beta(x,x) = 0$, then $x = 0$.

As you well know, if β is the form corresponding to the identity matrix, then it is an inner product.

From this point on we will usually denote the value of a form on the pair (u,v) by

$$\langle x, y \rangle.$$

This means that in the worst case you might need to consider the context to decide exactly what the form is. You may assume that if we're working over a finite-dimensional real vector space, we will use the traditional inner product.

The set $\mathrm{Mat}_{m\times n}(\mathbb{R})$ of $m \times n$ real matrices is a vector space and the map that takes the pair (A,B) to $\mathrm{tr}(A^T B)$ is easily seen to be bilinear. With slightly more effort, you may show that it is an inner product. We will make extensive use of this inner product in these notes. We will usually denote its value by $\langle A, B \rangle$.

A *norm* on a real vector space V is a function that sends a vector v to a real number $\|v\|$ such that:

(a) If v is a vector and c is a scalar, then $\|cv\| = |c|\|v\|$.

(b) For all vectors v we have $\|v\| \geq 0$, and if $\|v\| = 0$, then $v = 0$.

(c) $\|u + v\| \leq \|u\| + \|v\|$.

If $\langle \cdot, \cdot \rangle$ is an inner product on v, then the function

$$v \mapsto \sqrt{\langle v, v \rangle}$$

is a norm. (But there are many useful norms that do not arise in this way. Ask your favourite analyst for details.)

1.3 Positive Semidefinite Matrices

A real matrix M is *positive semidefinite* if:

(a) It is symmetric.

(b) For all vectors x we have $x^T Mx \geq 0$.

A matrix M is *positive definite* if it is positive semidefinite and invertible. Equivalently, it is positive definite if is positive semidefinite, and if $x^T Mx = 0$, then $x = 0$. (We will prove the equivalence of these two definitions later.) If A and B are matrices, we write $A \succcurlyeq B$ to denote that $A - B$ is positive semidefinite. Thus A is positive semidefinite if and only if $A \succcurlyeq 0$.

A matrix M is positive definite if and only if the associated form

$$\beta(x, y)) = x^T My$$

is an inner product.

We can now decode the definition of a semidefinite program given in the first section. We can view it as arising from a linear program by replacing vectors by matrices, and replacing the non-negativity condition by the assumption that feasible matrices must be positive semidefinite. It could be objected that, although we have shown this translation from LP to SDP is "feasible", the real question is how can we make use of it.

1.4 Cocliques in Graphs

We develop an example of a semidefinite program. A *coclique* (aka independent set) in a graph G is a set of vertices S (say), such that no two vertices in S are adjacent. We can represent S as a characteristic vector x on the vertices of G, such that $x_i = 0$ if the vertex $i \notin S$ and $x_i = 1$ if $i \in S$. If this was a course in linear programming we would happily work with these characteristic vectors, but we need matrices.

So, to the coclique S we associate the matrix $X = |S|^{-1}xx^T$. Observe that

$$\operatorname{tr}(X) = \operatorname{tr}(|S|^{-1}xx^T) = \frac{1}{|S|}\operatorname{tr}(xx^T) = \frac{1}{|S|}\operatorname{tr}(x^Tx) = \frac{1}{|S|}|S| = 1.$$

Further if J is the matrix with all entries equal to 1 and $\mathbf{1}$ is the vector with all entries 1, then $J = \mathbf{1}\mathbf{1}^T$

$$\langle J, X\rangle = \frac{1}{|S|}\operatorname{tr}(Jxx^T) = \frac{1}{|S|}\operatorname{tr}(\mathbf{1}\mathbf{1}^Txx^T) = \frac{|S|^2}{|S|} = |S|.$$

None of this makes any use of the fact that x is the characteristic vector of a coclique. We observe that if i and j are vertices in G, then $X_{i,j} = 1$ if and only if i and j are distinct vertices in S; in particular $X_{i,j} = 0$ if i and j are adjacent.

We conclude that the maximum value of $\langle J, X\rangle$, where X lies on the affine space

$$\Omega = \Big\{X : \operatorname{tr}(X) = 1,\ X_{i,j} = 0 \text{ if } \{i,j\} \in E(G)\Big\}$$

and $X \succcurlyeq 0$, is an upper bound on the maximum size $\alpha(G)$ of a coclique in G. The maximum is called the *theta number* of G, and denoted $\theta(G)$. We point out that it would be more useful to have a semidefinite program whose minimum value was an upper bound for $\alpha(G)$.

You should prove that if we restrict the feasible region to the diagonal matrices in Ω, the resulting problem is a linear program.

Chapter 2

Positive Semidefinite Matrices

2.1 Characterizations

We recall that a matrix A is positive semidefinite if it is symmetric and $x^T Ax = 0$ for all x, and it is positive definite if it is positive semidefinite and, if $x^T Ax = 0$ then $x = 0$. A diagonal matrix is positive semidefinite if and only if its diagonal entries are non-negative.

If $A = M^T M$, then

$$x^T AX = x^T M^T MX = \|Mx\|^2 \geq 0;$$

hence this gives us a large supply of positive semidefinite matrices. There is a generalization. If $x_1, \ldots, x_d$ are vectors in an inner product space, their *Gram matrix* G is the $d \times d$ matrix with

$$G_{i,j} = \langle x_i, x_j \rangle.$$

The exercises provide you with the opportunity to show that a Gram matrix is positive semidefinite, and it is positive definite if and only if the vectors $x_1, \ldots, x_d$ are linearly independent.

2.1.1 Theorem. *Let A be a symmetric matrix. The following are equivalent:*

(a) A is positive semidefinite.

(b) $x^T Ax \geq 0$ for all x.

(c) $A = B^T B$ *for some matrix* B.

(d) All eigenvalues of A *are non-negative.*

Proof. That (b) implies (a) is the definition, and we proved (c) implies (b) above. For the next step assume that the eigenvalues of A are non-negative. Then there is an orthogonal matrix L and a diagonal matrix D such that $A = LDL^T$. Since the eigenvalues of A are non-negative, the diagonal entries of D are non-negative and therefore there is a diagonal matrix with non-negative entries whose square is D. Originality is not required, and so we denote it by $D^{1/2}$. Now we have

$$A = LD^{1/2}D^{1/2}L^T = LD^{1/2}L^T \, LD^{1/2}L^T$$

and, and if $B = LD^{1/2}L$, then $A = B^2$. Since B is symmetric, this shows that (d) implies (c).

To complete the proof, we show (a) implies (d). If $Az = \lambda z$ and $z \neq 0$, then

$$0 \leq z^T Az = \lambda z^T z,$$

whence $\lambda \geq 0$. □

One consequence of our proof is that every positive semidefinite matrix has a square root (in fact 2^n square roots if A is $n \times n$).

If B is chosen so $A = B^T B$, then

$$\det(A) = \det(B^T)\det(B) = det(B)^2 \geq 0.$$

2.1.2 Theorem. *A matrix* A *is positive semidefinite if and only* $\langle A, X\rangle \geq 0$ *for all positive semidefinite matrices* X.

Proof. If $A \succcurlyeq 0$, then $A = B^T B$ for some matrix B and so

$$\langle A, X\rangle = \operatorname{tr}(AX) = \operatorname{tr}(B^T BX) = \operatorname{tr}(BXB^T).$$

Since $X \succcurlyeq 0$ we see that $BXB^T \succcurlyeq 0$ and hence $\operatorname{tr}(BXB^T) \geq 0$.

For the other direction, note that if x is a vector, then xx^T is positive semidefinite. If $\langle A, X\rangle \geq 0$ for all X, the $\langle A, xx^T\rangle \geq 0$ for all x. However

$$\langle A, xx^T\rangle = \operatorname{tr}(Axx^T) = \operatorname{tr}(x^T Ax) = x^T Ax$$

and therefore $x^T Ax \geq 0$ for all x. □

A real matrix P is a *projection* (more precisely, represents orthogonal projection onto some subspace) if $P = P^T$ and $P^2 = P$. And here "some subspace" is the column space of P. A matrix P such that $P^2 = P$ is said to be *idempotent* if $P^2 = P$, so a projection is a symmetric idempotent. If P is idempotent, so is $I - P$, and consequently if P is a projection, the so is $I - P$. Projections are necessarily positive semidefinite; you should be able to provide three proofs of this. For any vector x we have

$$xx^T\, xx^T = x^T x\, xx^T$$

and so if $\|x\| = 1$, then xx^T is a projection—onto the 1-dimensional subspace spanned by x. If P and Q are projections and $PQ = QP = 0$, we say that they are *orthogonal*; in this case $P + Q$ is also a projection (onto the direct sum of the column spaces of P and Q).

2.2 Sums of Projections

Suppose $A \succcurlyeq 0$. Then

$$\begin{aligned} 0 \le (x+ty)^T A(x+ty) &= x^T Ax + tx^T Ay + ty^T Ax + t^2 y^T AY \\ &= x^T Ax + 2tx^T Ay + t^2 y^T AY. \end{aligned}$$

The last term is a quadratic polynomial in t and, since it is non-negative for all t, its discriminant cannot be positive, i.e.,

$$(x^T Ay)^2 - x^T Ax\, y^T Ay \le 0.$$

Equivalently we have a form of the Cauchy-Schwarz inequality: if $A \succcurlyeq 0$, then

$$(x^T Ay)^2 \le x^T Ax\, y^T Ay.$$

(You will be asked to characterize the case where equality holds.)

Any symmetric matrix A with rank one can be expressed in the form $\pm xx^T$ for some x (and x is a unit vector if and only if $\mathrm{tr}(A) = 1$.) Thus symmetric matrices of rank one with non-nagative diagonal are positive semidefinite.

2.2.1 Theorem. *A positive semidefinite matrix with rank r can be written as the sum of at most r positive semidefinite matrices with rank one.*

Proof. We go by induction on $\operatorname{rk}(A)$. Choose a vector x such that $Ax \neq 0$—if no such vector exists then $A = 0$ and the result holds. Define

$$B = A - \frac{1}{x^T Ax} Axx^T A$$

and note that $B - A$ is a symmetric matrix with rank one. We claim that $B \succcurlyeq 0$. We have

$$y^T By = y^T Ay - \frac{1}{x^T Ax} y^T Axx^T Ay = y^T Ay - \frac{x^T Ay^2}{x^T Ax}$$

and, by our version of Cauchy-Schwarz we deduce that $B \succcurlyeq 0$. Since the column space of B is a subspace of the column space of A, and since

$$Bx = Ax - \frac{1}{x^T Ax} Axx^T Ax = Ax - Ax = 0,$$

we have $\operatorname{rk}(B) < \operatorname{rk}(A)$. By induction B is a sum of at most $\operatorname{rk}(A) - 1$ symmetric matrices of rank one, and the result follows. □

One consequence of this is that $\langle A, P \rangle \geq 0$ for all positive semidefinite matrices P if and only if $\langle A, P \rangle \geq 0$ for all positive semidefinite matrices of rank one. (We have already met a version of this.) We also note that it is not hard to that we can replace "at most r" in the statement of the theorem by "exactly r".

2.3 Principal Submatrices

The following result will get a lot of use.

2.3.1 Lemma. *If A is positive semidefinite, each principal submatrix of A os positive semidefinite.*

Proof. Suppose A is $n \times n$ and $K \subseteq \{1, \ldots, \}$. Let x be a vector in $\mathbb{R}^K$ and let $\hat{x}$ be the vector in $\mathbb{R}^n$ such that

$$\hat{x}_i = \begin{cases} x_i, & i \in K; \\ 0, & i \notin K. \end{cases}$$

If B is the principal submatrix of A with rows and columns indexed by K, then

$$\hat{x}^T A \hat{x} = x^T Bx$$

and so it follows that $B \succcurlyeq 0$ if $A \succcurlyeq 0$. □

Each diagonal entry of a matrix is 1×1 principal submatrix, so it follows that the diagonal entries of a positive semidefinite matrix are non-negative. (We could also prove this by observing that $A_{r,r} = e_r^T A e_r$.)

Since each principal submatrix of a positive semidefinite matrix is positive semidefinite, each principal minor of a positive semidefinite matrix is non-negative. The converse is true, but we will not need it.

2.3.2 Lemma. *If $A \succcurlyeq 0$ and $U^T AU = 0$ for some matrix U, then $AU = 0$.*

Proof. If $A \succcurlyeq 0$ then $A = B^T B$ for some matrix B. Hence

$$U^T AU = U^T B^T BU = (BU)^T BU = \langle BU, BU\rangle$$

and so $U^T AU = 0$ implies $BU = 0$. If $BU = 0$, then $Ax = B^T BU = 0$. □

The most commonly used case of this is when U is a vector. There is an important consequence of this: if $M \succcurlyeq 0$ and $M_{i,i} = 0$, then all entries in the i-th row and in the i-th column of M are zero. For if $M_{i,i} = 0$, then $e_i^T M e_i = 0$ and hence $Me_i = 0$ (and $e_i^T M = 0$.)

2.4 Cholesky

Suppose the symmetric matrix M has the 2×2 block form

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where A is invertible. We then have the factorization

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}$$

The matrix

$$S = D - CA^{-1}B$$

is called the *Schur complement* of A in M. Note that

$$\det(M) = \det(A)\det(D - CA^{-1}B).$$

2.4.1 Theorem. *If the symmetric matrix M has the 2×2 block form*

$$M = \begin{pmatrix} A & B^T \\ B & D \end{pmatrix}$$

where A is invertible, then M is positive semidefinite if and only if A and $D - BA^{-1}B^T$ are positive semidefinite.

Proof. We can rewrite our factorization above as

$$M = P^T \begin{pmatrix} A & 0 \\ 0 & D - B^T A^{-1} B \end{pmatrix} P$$

where

$$P = \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}.$$

It follows that $M \succcurlyeq 0$ if and only if

$$\begin{pmatrix} A & 0 \\ 0 & D - B^T A^{-1} B \end{pmatrix} \succcurlyeq 0.$$ □

We can use this to give a variant of the characterization of positive semidefinite matrices by factorization; this will also give us an algorithm for recognizing if a matrix is positive semidefinite.

2.4.2 Lemma. *If A is positive semidefinite, there is a lower triangular matrix L such that $A = LL^T$.*

Proof. Assume A is $n \times n$. The proof is by induction on n, and the result is trivial if $n = 1$. Assume $n \geq 2$. If $A_{1,1} = 0$, then the first row and column of A are zero. If we let A_1 be the matrix we get by deleting the first row and column of A, then the theorem holds for A_1 by induction.

So suppose $a = A_{1,1} \neq 0$. Then we may assume

$$A = \begin{pmatrix} a & b^T \\ b & A_1 \end{pmatrix}$$

and, by the previous theorem, there is a lower triangular matrix N such that

$$N^{-1} A N^{-T} = \begin{pmatrix} a & 0 \\ 0 & A_1 - a^{-1} b b^T \end{pmatrix}.$$

Again by the previous theorem $a > 0$ and $A_1 - a^{-1}bb^T \succcurlyeq 0$. So it follows by induction that there is a lower triangular matrix M such that $A = MDM^T$, where D is diagonal with non-negative diagonal entries. The matrix L we need is $MD^{1/2}$. □

2.5 Kronecker and Schur Products

If A and B are matrices, their *Kronecker product* $A \otimes B$ is the matrix we get by replacing the ij-entry of A with the matrix $A_{i,j}B$. If A is $m \times n$ and B is $r \times s$, then $A \otimes B$ is a $mr \times ns$ matrix. We can view its rows and columns as indexed by ordered pairs of integers, thus

$$(A \otimes B)_{(i,k),(j,\ell)} = A_{i,j}B_{k,\ell}.$$

If A is $m \times 1$ and B is $1 \times n$, then $A \otimes B$ is an $m \times n$ matrix.

The Kronecker product is bilinear but is not symmetric. One of the most important properties of the Kronecker product is that if the matrix products AC and BD are defined, then

$$(A \otimes B)(C \otimes D) = AC \otimes BD$$

and, in particular for vectors u and v of the right order

$$(A \otimes B)(u \otimes v) = Au \otimes Bv.$$

Hence if u and v are eigenvectors of A and B respectively, then $u \otimes v$ is an eigenvector for $A \otimes B$. We also see that

$$A \otimes B = (A \otimes I)(I \otimes B) = (I \otimes B)(A \otimes I).$$

Many graph products can be defined in terms of Kronecker product.

We also have

$$(A \otimes B)^T = A^T \otimes B^T;$$

note though that $A \otimes A^T$ is not symmetric. (You might show that if P represents the linear map sending $u \otimes v$ to $v \otimes u$, then $P(A \otimes A^T)$ is symmetric.)

It is not hard to show that $\operatorname{tr}(A \otimes B) = \operatorname{tr}(A)\operatorname{tr}(B)$. The behaviour of the determinant is more complex, and is left as an exercise.

Let $e_1, \dots, e_n$ denote the standard basis of $\mathbb{R}^n$. If A is an $m \times n$ matrix, we define $\operatorname{vec}(A)$ to be the column vector

$$\begin{pmatrix} Ae_1 \\ \vdots \\ Ae_n \end{pmatrix}.$$

You are invited to verify that for an matrix X we have

$$\operatorname{vec}(AXB^T) = (A \otimes B)\operatorname{vec}(X).$$

In other words, the linear map on the space of matrices that sends X to AXB^T can be represented by the matrix $A \otimes B$.

The *Schur product* $A \circ B$ of two $m \times n$ matrices is the $m \times n$ matrix defined by the condition

$$(A \circ B)_{i,j} = A_{i,j}B_{i,j}.$$

It may also be referred to as the Hadamard product (although Hadamard never used it), or as the ‘bad student’s product. If is bilinear and symmetric.

The Kronecker and Schur products play nicely together. Thus, assuming the orders line up,

$$(A \otimes B) \circ (C \otimes D) = (A \circ C) \otimes (B \circ D)$$

and

$$(A \circ B) \otimes (C \circ D) = (A \times C) \circ (B \otimes D).$$

(Perhaps you should check the last one.)

Naturally we are concerned with products of positive semidefinite matrices.

2.5.1 Lemma. *If A and B are positive semidefinite, so is $A \otimes B$. If A and B are positive semidefinite matrices of the same order, then $A \circ B$ is positive semidefinite.*

Proof. It is probably simplest to prove the Kronecker product of positive semidefinite matrices is positive semidefinite using the characterization in terms of factorizations, but it’s your choice. For the Schur product, one simple approach is to observe that $A \circ B$ is a principal submatrix of $A \otimes B$ and the former is positive semidefinite if the latter is. □

The Schur product allows us to provide another presentation of our standard inner product. Let $\operatorname{sum}(M)$ denote the sum of the entries of the matrix M. Then

$$\langle A, B \rangle = \operatorname{sum}(A \circ B).$$

2.6 Normal Linear Algebra

In this section we will work with complex inner product spaces. The inner product on $\mathbb{C}^d$ is

$$\langle x, y\rangle = \sum_r \bar{x}_r y_r$$

and on complex matrices it is

$$\langle A, B\rangle = \operatorname{tr} A^* B = \operatorname{sum} \overline{A} \circ B.$$

Our complex inner products are linear in the second variable and conjugate-linear in the first. (This is one of two conventions.) A matrix A is *Hermitian* if $A = \overline{A}^T$; we will usually denote the complex-conjugate of A by A^*.

2.6.1 Lemma. *The matrix A is Hermitian if and only if $\langle x, Ay\rangle = \langle Ax, y\rangle$, for all x and y.* □

If the subspace U is A-invariant, then its orthogonal complement $U^\perp$ is A^*-invariant, and therefore if A is Hermitian then the orthogonal complement of an A-invariant subspace is A-invariant. This the key fact we nee to prove that Hermitian matrices are diagonalizable, but we want more.

A matrix A is *normal* if $AA^* = A^*A$. Clearly Hermitian matrices are normal; unitary matrices $U^*U = I$ provide another large class. If A is normal, so are $A + A^*$ and $A - A^*$, whence

$$A = \frac{1}{2}(A + A^*) + i\frac{1}{2i}(A - A^*)$$

where $A + A^*$ and $\frac{1}{2i}(A - A^*)$ are Hermitian and commute. We could view these as the real and imaginary parts of A.

The following lemma is a fundamental tool.

2.6.2 Lemma. *The matrix A is normal if and only if $\langle Ax, Ax\rangle = \langle A^*x, A^*x\rangle$ for all x.*

Proof. We have $\langle x, Ay\rangle = \langle A^*x, y\rangle$ for any A. If A is normal,

$$\langle Ax, Ay\rangle = \langle A^*Ax, y\rangle = \langle AA^*x, y\rangle = \langle A^*x, A^*y\rangle.$$ □

2.6.3 Lemma. *If A is normal and $Ax = \lambda x$, then $A^*x = \bar{\lambda}x$.*

Proof. If $Ax = \lambda x$, then $(\lambda I - A)x = 0$. If A is normal then so is $\lambda I - A$ and

$$(\lambda I - A)^* = \bar{\lambda} I - A^*$$

As

$$\langle (\lambda I - A)x, (\lambda I - A)x \rangle = 0,$$

it follows that

$$\langle (\bar{\lambda} I - A^*)x, (\bar{\lambda} I - A^*)x \rangle = 0$$

whence $A^* x = \bar{\lambda} x$. □

2.7 Spectral Decomposition for Normal Matrices

Spectral decomposition is a reformulation of the result that a normal matrix can be unitarily diagonalized. It proves a powerful for working with functions of matrices.

2.7.1 Theorem. *A matrix A is normal if and only if it is unitarily similar to a diagonal matrix, i.e., $A = LDL^*$ where D is diagonal and L is unitary.*

Proof. First the easy part. If $A = LDL^*$ with $L^* L = I$, then

$$A^* = (LDL^*)^* = LD^* L^*$$

and since the diagonal matrices D and D^* commute, so do A and A^*.

For the converse, we sketch the proof that, if A is normal, there is an orthonormal basis that consists of eigenvectors of A. Then L is the matrix with these eigenvectors as columns and the diagonal entries of D are the eigenvalues of A.

The proof now goes by induction on the dimension d. The result is trivial if $d \leq 1$, so assume $d > 1$. The matrix A has an eigenvector, z say, with eigenvalue λ. From the previous section, this is also an eigenvector for A^* (with eigenvalue $\bar{\lambda}$). Hence the subspace spnned by z is A- and A^*-invariant, and therefore its orthogonal complement $z^\perp$ is A^*- and A-invariant. If B denotes the restriction of A to $z^\perp$, then B^* is equal to the restriction of A^* to $z^\perp$ and therefore B is normal. By induction there is an orthonormal basis for $z^\perp$ consisting of eigenvectors of B and this basis, together with $\|z\|^{-1} z$, forms the basis we need. □

Assume $A = LDL^*$ and that the eigenvalues of A are $\theta_1, \dots, \theta_m$. Then there are diagonal matrices $D_1, \dots, D_m$ with diagonal entries 0 and 1 such that

$$I = \sum_r D_r, \qquad D = \sum_r \theta_r D_r.$$

We define matrices $E_1, \dots, E_m$ by

$$E_r = LD^r L^*.$$

Then

$$\sum_r E_r = LDL^* = A$$

and

$$AE_r = LDL^* \, LD_r L^* = LDD_r L^* = \theta_r LD_r L^* = \theta_r E_r.$$

Moreover $E_r^2 = E_r$ and $E_r E_s = 0$ if $r \neq s$. We refer to $E_1, \dots, E_m$ as the *spectral idempotents* of A.

2.7.2 Lemma. *The spectral idempotents of a normal matrix A are polynomials in A.* □

2.8 Algebras

Matrix algebras, operators on spaces of matrices. Centralizers of permutation groups.

Projections onto subspaces of matrix algebras. Partial trace.

Chapter 3

Convex Sets

3.1 Open Sets, Closed Sets

Let V be a real vector space with a norm $\|\cdot\|$. (You may assume this is the usual Euclidean norm.) The *ball of radius* r about u in V is the set

$$\{x \in V : \|x - u\| \leq r\}.$$

A subset S of V is *open* (relative to our norm) if for each v in S, there is $\epsilon > 0$ such that the ball of radius ϵ about v is a subset of S. A subset of V is closed if its complement is open. Finite unions and intersections of open sets are open, ditto for closed sets. Further an arbitrary union of open sets is open, but an arbitrary intersection need not be.

A subset S of V is *bounded* if there is a scalar C such that $\|v\| \leq C$ for all v in S.

A collection of subsets $\mathcal{C}$ *covers* S if S is a subset of the union of the elements of $\mathcal{C}$. A subset S of V is *compact* if, for any collection of open sets that covers S, there is a finite subset of the collection that covers S. In our setting a subset is compact if and only it is closed and bounded. We will make free use of this fact.

3.1.1 Theorem. *If f is a continuous function on a compact set C, there is an element c of C such that $f(x) \leq f(c)$ for all x in C.* □

If C is a subset of $\mathbb{R}^n$, and point x in C is an *interior point* if, for some $\epsilon > 0$, the ball of radius ϵ about u is contained in C. The set of all interior points of A is the *interior* of A; a point of C that is not an interior point is a *boundary point*. The interior of a set is open, and might be empty. An open set has no boundary points. The set of boundary points of C is often denoted by ∂C.

3.2 Affine Space

Let V be a vector space over $\mathbb{F}$. We say a vector y is an affine linear combination of vectors $x_1, \ldots, x_d$ if there are scalars $a_1, \ldots, a_d$ such that

$$y = \sum_r a_r x_r, \qquad \sum_r a_r = 1.$$

The set of all affine linear combinations of two distinct vectors is the affine line through them. An *affine subspace* of V subset of V that is closed under taking affine linear combinations. You should prove that the affine subspaces are the cosets of the (usual) subspaces of V.

A set of vectors $x_0, \ldots, x_d$ is *affinely dependent* if there are scalars $a_0, \ldots, a_d$, not all zero, such that

$$\sum_r a_r x_r = 0, \qquad \sum_r a_r = 0.$$

We can relate affine dependence to linear dependence: if $x \in V = \mathbb{F}^d$, let $\hat{x}$ denote the vector in $\mathbb{F}^{d+1}$ we get by adjoining an extra coordinate, with entry equal to 1. Then vectors $x_1, \ldots, x_d$ are affinely dependent if and only if $\hat{x}_1, \ldots, \hat{x}_d$ is linearly dependent. A set that is not *affinely dependent* is *affinely independent*. The maximum cardinality of an independent subset of an affine space is its dimension.

Note that all this works over any field, of any characteristic. Further note that the study of a vector space V and its linear subspace is a lightly disguised version of projective geometry. Affine geoemtry arises when we allow "subspaces" that do not contain the zero vector. On the other hand, the map $x \mapsto \hat{x}$ used above shows that we can embed an affine space of dimension d into a vector space of dimension $d+1$ (which has projective dimension d, go figure).

An affine space has *affine dimension d* if the maximum size of an affinely independent subset is $d+1$. Such a subset may be called an *affine basis*.

3.2.1 Lemma. *Let C be a subset of $\mathbb{R}^d$ and let $x_0, \ldots, x_e$ be an affinely independent subset of C of maximum size. Then the affine subspace generated by $x_0, \ldots, x_e$ has affine dimension e and contains C.* □

3.3 Convex Sets

Optimization problems over compact sets are in general very difficult—a closed bounded set can be extraordinarily complicated. Convex sets are sufficiently general to be useful and not so complicated as to be unmanageable. In this section we begin our dealings with convex sets.

An element y in a real vector space is a *convex combination* of $x_1, \dots, x_d$ if there are scalars $a_1, \dots, a_d$ such that

$$y = \sum_r x_r, \qquad \sum_r a_r = 1, \qquad a_r \geq 0 \; (r = 1, \dots, d).$$

Thus y is an affine combination of $x_1, \dots, x_d$ and the coefficients in this combination are non-negative. An important special case is when $d = 2$; here the affine line through distinct points x_1 and x_2 is the set

$$sx_1 + (1 - s)x_2, \; s \in \mathbb{R},$$

and the convex combinations of the two points is the set

$$sx_1 + (1 - s)x_2, \; 0 \leq s \leq 1.$$

We will refer to this set as the *line segment* generated by x_1 and x_2, it consists of all points on the line through x_1 and x_2 that lie between x_1 and x_2.

A set C is *convex* if it is closed under taking convex combinations. The intersection of any collection of convex sets is convex. The *convex hull* of a set S is the intersection of all the convex sets that contain it. A *convex polytope* is the convex hull of a finite set of points. The unit ball relative to a given norm is convex.

One of the most important families of convex sets are half-spaces. A *half-space* in a real inner product space V is a set of the form

$$\{x \in V : \langle \alpha, x \rangle \leq b\}$$

where $\alpha \in V$ and $b \in \mathbb{R}$. Let us denote this by $H_{\alpha,b}$. We see that $H_{-\alpha,b}$ is a second half-space and $H_{\alpha,b} \cap H_{-\alpha,b}$ is the affine hyperplane with equation $\langle \alpha, x \rangle = b$. A *polyhedral set* is defined to be the intersection of a finite number of half-spaces. Thus the non-negative vectors in $\mathbb{R}^d$ form a polyhedral set.)

The set of $n \times n$ positive semidefinite matrices form an important example of a convex set. To see this. note that if $A \succcurlyeq 0$ and $c \geq 0$, then $cA \succcurlyeq$. Since the sum of two positive semidefinite matrices is positive semidefinite, it follows that any non-negative linear combination of positive semidefinite matrices is positive semidefinite, and hence the set of all positive semidefinite matrices is convex.

It is an important fact that a bounded polyhedral set is the convex hull of a finite set of points. One proof of this follows from the theory of linear programming. Much of what we do in the following sections is aimed at generalizing this result.

A point x in a convex set C is an *extreme point* if it is an endpoint of any line segment that is contained in C and contains x and a point in the relative interior of C. If C is the convex hull of a finite set of points S, the extreme points of C are a subset of S. It follows from the proof of Theorem 2.2.1 that any positive semidefinite matrix with rank greater than one is not an extreme point. To show that each rank-one positive semidefinite matrix is an extreme point, it suffices to show that if A and B are symmetric matrices of rank one and A is not a scalar multiple of B, then $\operatorname{rk}(A + B) > 1$. (See the exercises, if you do not want to do it now.)

The interior of a convex set may be empty—consider a line in $\mathbb{R}^2$. The problem arises because the line has dimension one and $\mathbb{R}^2$ has dimension two. A *simplex* is the convex hull of an affinely independent set of points. If β is an affine basis for an affine space of dimension d.

3.3.1 Lemma. *If $\mathcal{C}$ is a convex subset in a real affine space V and the affine span of $\mathcal{C}$ is V, then the interior of $\mathcal{C}$ is not empty.*

Proof. If $\mathcal{C}$ spans V, then $\mathcal{C}$ contains a basis $x_0, \ldots, x_d$ (say). The convex hull of this basis is a simplex and, in the exercises, you will have opportunity to prove that the interior of such a simplex is not empty. □

The *relative interior* of a subset $\mathcal{C}$ of a real affine space is the interior of $\mathcal{C}$, viewed as a subset of its affine span. We can use the term "relative interior" to avoid mentioning affine spans.

3.4 Nearest Points

We work over a real inner product space V. Suppose C is a closed subset of V and x is a point not in C. If $y \in C$, the set

$$\{z \in C : \|x - z\| \leq \|x - y\|\}$$

is bounded and, since it is the intersection of a closed ball with C, it is also closed. Therefore it is compact and hence there is a point z_0 in C which is a nearest point to x. Equivalently, the optimization problem of finding z in C such that $\|x - z\|$ is minimal has a solution.

Assume now that C is closed and convex and $x \notin C$, and suppose there are two distinct nearest points a and b. The line $a \vee b$ is a closed convex set and therefore its intersection with C is a closed line segment. It is easy to prove that there is a unique point on this line segment closest to x. This shows that there is a unique nearest point to x in C. (Thus closed implies existence, closed and convex implies unique existence.) One of the most important and useful properties of convex sets is, roughly speaking, that if C is a closed convex set and x is a point not in C, there is a hyperplane that separates x from C. We will need to be more precise about this, but that can wait. One proof of this claim follows from properties of the nearest-point map, which takes each point in V to the point in C nearest to it. This map is referred to as *metric projection*, and we denote metric projection onto a closed convex set C by π_C. Note that the image of π_C is contained in the boundary ∂C.

3.4.1 Lemma. *Let C be a non-empty closed convex set in the vector space V. If $x, y \in V$, then*

$$\|\pi_C(x) - \pi_C(y)\| \leq \|x - y\|.$$

Proof. We work using the four points x, y, $\pi_C(x)$ and $\pi_C(y)$; in practice this means we are working in affine space of dimension three. Let H_x and H_y denote the hyperplanes through $\pi_C(x)$ and $\pi_C(y)$ respectively with normal parallel to $h = \pi_C(y) - \pi_C(x)$. (Thus these hyperplanes are parallel.)

If $\langle x - \pi_C(x), h\rangle > 0$, then $\pi_C(x)$ is not the nearest point to x on the line segment $[\pi_C(y), \pi_C(x)]$, a contradiction to the definition of $\pi_C(x)$. The region between H_x and H_y divides V into two disjoint regions and, by what we have just seen, x lies in one of these regions and y in the other. Hence the distance between x and y is at least the distance between H_x and H_y. □

3.4.2 Corollary. *The metric projection map onto a non-empty closed convex set is a contraction, and therefore it is continuous.* □

Suppose C is closed and convex and $u \notin C$. If x lies on the half-line from $\pi_C(u)$ through u, then $\pi_C(x) = \pi_C(u)$. We leave this as an exercise.

3.4.3 Lemma. *Suppose C is a non-empty closed convex set in $\mathbb{R}^d$. Then for each point y in ∂C, there is a point x not in C such that $\pi_C(x) = y$.*

Proof. Assume $y \in \partial C$. Any closed ball of positive radius around y is convex, and so its intersection with C is a closed convex set. Accordingly it suffices to prove the theorem with this set in place of C; equivalently we may assume C is bounded and lies in the interior of some closed ball B.

We construct x by a limiting argument. Choose a sequence of points y_i in $\mathbb{R}^d \setminus C$ such that $d(y, y_i) < 1/i$, thus its limit is y. Because metric projection is continuous, the sequence of points $\pi_C(y_i)$ must converge to y.

The intersection of the line $y \vee \pi_C(y_i)$ with ∂B is a point which we denote by x_i, and is such that $\pi_C(x_i) = \pi_C(y_i)$. Since ∂B is compact, the sequence formed by the points x_i has a limit in ∂B, which we denote by x. Since limits and projection maps commute, we have $\pi_C(x) = y$. □

3.5 Separation

Two subsets A and B of $\mathbb{R}^d$ are *separated* if they lie on different sides of the hyperplane. (This is a slightly unusual use of the word separated, because $A \cap B$ need not be empty.) The hyperplane with equation $\langle h, x\rangle = c$ separates A and B if $\langle h, a\rangle \le c$ for all $a \in A$ and $\langle h, b\rangle \ge c$ for all b in B. Note that if we use the equation $\langle -h, x\rangle = -c$, the inequalities are reversed. If H separates A and B and

$$H \cap A = H \cap B = \emptyset,$$

we say that H *strictly separates* A and B. (These definitions are not entirely consistent with those in Laurent and Vallentin.)

A hyperplane H is a *supporting hyperplane* for a convex subset C of $\mathbb{R}^d$ if C lies in one of the half-spaces determined by H and $H \cap C \neq \emptyset$. If $y \in H \cap C$, we may say that H supports C at y.

3.5.1 Lemma. *Suppose C is a non-empty closed convex set in $\mathbb{R}^d$. If x is a point in $\mathbb{R}^d \setminus C$, the hyperplane through $\pi_C(x)$ with normal $x - \pi_C(x)$ supports C at $\pi_C(x)$.*

Proof. Set $y = \pi_C(x)$. Let H be the given hyperplane through y and suppose z is a point in C and in the same open half-space H^+ as x—thus $\langle z - y, x - y \rangle > 0$. We derive a contradiction.

We have

$$x - (sy + (1 - s)z) = (x - y) - (1 - s)(z - y)$$

and therefore

$$\|x - (sy + (1 - s)z)\|^2 = \|x - y\|^2 - 2(1 - s)\langle x - y, z - y \rangle + (1 - s)^2 \|z - y\|^2$$

It follows that for small positive values of s, the point $sy + (1 - s)z$ is closer than x to $y = \pi_C(x)$. Since C is convex, $[y, z] \subseteq C$, and so we have our contradiction.

We conclude that $C \cap H^+ = \emptyset$, and therefore H is a supporting hyperplane for C that separates x from C. □

3.6 Two More Proofs of Separation

We follow Barvinok **?**.

3.6.1 Theorem. *Let C be an open convex set in $\mathbb{R}^d$. If x is a point in $\mathbb{R}^d \setminus C$, there is a hyperplane through x such that one of the open half-spaces determined by H contains C.*

Proof. We may assume without loss that $x = 0$.

We first treat the case $d = 2$. Let S be the unit circle in $\mathbb{R}^2$ and let α be the image of C under the map $x \mapsto u/\|u\|$. Since C is compact, α is a connected set and therefore it is an arc in S. If $u \in S$, then $u = v/\|v\|$ for some v in C. If ℓ is the line through v parallel to the tangent to S at u, then $\ell \cap C$ is an open interval that contains v and the projection onto S of this interval is open. Therefore α is open.

If the length of α is not less than π, then α contains a pair of antipodal points, and therefore C contains points v and w such that

$$\frac{1}{\|v\|} v = -\frac{1}{\|w\|} w,$$

but by convexity it follows that $0 \in C$.

If p is an "endpoint" of α, the line through 0 and p is the hyperplane we need.

Now assume $d \geq 3$. If P is a plane on 0, then $P \cap C$ is an open convex set and, by what we have just proved, there is a line ℓ in P that contains 0 and is disjoint from C. Finally let H be a maximal affine subspace such that $0 \in H$ and $H \cap C = \emptyset$. We claim that H is a hyperplane. To prove this we work in the quotient space V/H. If H is not a hyperplane, then $\dim(V/H) \geq 2$. The image of C in V/H is open (prove it) and consequently there is a line ℓ in V/H disjoint from the image of C such that $0 \in \ell$. The preimage of ℓ in V is a hyperplane that contains 0 and is disjoint from C. □

Our second proof follows Tunçel **?**.

3.6.2 Theorem. *Let C be a non-empty closed convex set in $\mathbb{R}^d$ that does not contain* 0*. Then there is a non-zero vector h and a positive scalar a such that*

$$C \subseteq \{x \in \mathbb{R}^d : h^T x \geq a\}.$$

Proof. Let u be a point in C such that $\|u\|$ is a minimum and define a to be $u^T u$. For each vector x in C, the line segment $[x, c]$ is contained in C. If $0 < s \leq 1$, then $sx + (1-s)u \in C$. Hence

$$u^T u \leq \|sx + (1-s)u\|^2 = \|s(x-u) + u\|^2 = s^2\|x-u\|^2 + 2su^T(x-u) + u^T u$$

and so

$$s^2\|x-u\|^2 + 2su^T(x-u) \geq 0.$$

Therefore

$$u^T(x-u) \geq -\frac{s}{2}\|x-u\|^2$$

and, taking limits as s decreases to 0, we see that $u^T(x-u) \geq 0$. As this holds for all x, we conclude that the theorem holds with $h = u$ and $a = u^T u$.

3.7 Extreme Points

Recall from Section 3.3 that a point x in a convex set C is an *extreme point* if it is an endpoint of any line segment that is contained in C and contains both x and a point in the relative interior of C. Equivalently, x is extreme in C if whenever a and b are points of C and $x = sa + (1 - s)b$, where $0 \leq s \leq 1$, then $s = 0$ or $s = 1$.

If C is a compact convex set and H is a supporting hyperplane for C, then each extreme point of $H \cap C$ is an extreme point of C. (Another exercise.)

3.7.1 Theorem. *A compact convex set if the convex hull of its extreme points.*

Proof. We proceed by induction on the dimension d of the affine space spanned by C. If $d = 0$ then C is a point and we're done. Assume $d \geq 1$. By hypothesis, the relative interior of C is not empty.

Suppose first that x is a boundary point of C. By Lemma 3.5.1, there is a supporting hyperplane H at x. We consider the set $D = H \cap C$. This is a compact convex set of dimension at most $d - 1$ that contains x; by induction it follows that x is a convex combination of extreme points of D. By our remark above it follows that x is a convex combination of extreme points of D.

So we may assume that x lies in the interior of C. Hence there is a line segment contained in C with x in its relative interior. The affine line that contains this line segment meets C in a closed line segment that joins two points on the boundary of C. Since these points are convex combinations of extreme points of C, so is x. □

If H is a supporting hyperplane for a convex set C, the intersection $H \cap C$ is a *face* of C. A face is convex and is closed if C is. You might prove that each extreme point of a convex polytope is a face.

3.8 Unit Balls and Norms

Unit balls are closed, centrally symmetric convex sets. Any such subset gives a norm.

3.9 Cones

A subset $\mathcal{C}$ of a real vector space V is a *convex cone* if it is closed under taking non-negative linear combinations. Equivalently

(a) If $x \in \mathcal{C}$ and $a \geq 0$, then $ax \in \mathcal{C}$.

(b) If $x, y \in \mathcal{C}$, then $x + y \in \mathcal{C}$.

We could also define a convex cone to be a convex set that is closed under multiplication by non-negative scalars. In these notes, cone will mean convex cone.

Two important examples of cones are the set of non-negative vectors in $\mathbb{R}^d$, and the positive semidefinite matrices in the space of $n \times n$ real matrices. For a third example, we have the *Lorentz cone* or *hyperbolic cone*, which is the subset of pairs (x, t) in $\mathbb{R}^d \times \mathbb{R}$ such that $\|x\| \leq t$. As yet another example, the set of symmetric matrices M such that $x^T M x \geq 0$ for all non-negative vectors x is a convex cone, known as the *copositive cone*.

According to the definition, linear subspaces are cones, but this case is of very little interest to us. We say a cone is *pointed* if it does not contain a line through the origin. In the exercises you may meet with the *Minkowski sum* $A + B$ of convex sets A and B, defined by

$$A + B := \{a + b : a \in A,\ b \in B\}.$$

You might prove that any cone that is not pointed is the Minkowski sum of a subspace with a pointed cone.

We can use cones in V to create partial orders on V. For if K is a cone in V and we declare $x \geq y$ if $x - y \in K$, then $\geq$ is a partial order on V.

If K is a convex cone and M is a linear map, then

$$MK := \{Mx : x \in K\}$$

is a convex cone. If M is invertible and $MK = K$ we say M is an *automorphism* of K. A cone K is *homogeneous* if, for each pair of points x and y in the interior of K, there is an automorphism of K that maps x to y. We note that A and B are invertible matrices, then the map $M \mapsto AMB^T$ is an automorphism of the set of positive semidefinite matrices and it can be shown that all automorphisms of this cone have this form.

The convex hull of a single point in a convex cone is called a *ray* of the cone. In convex cones, extreme rays take on the role of extreme points: a ray R of a convex cone C is an *extreme ray* if it is a face.

3.9.1 Lemma. *If K is a closed convex cone and H with equation $\langle h, x\rangle = c$ supports K at a non-zero vector u, then $c = 0$.*

If $x \in K$ then

$$\langle h, x \rangle \leq c$$

and, if $\lambda > 0$, then $\langle h, \lambda x \rangle \leq c$ and thus

$$\langle h, x \rangle \leq \lambda^{-1} c$$

for all positive λ. If $c < 0$, we take λ small and deduce that $\langle h, x \rangle$ is less than any negative number, which is impossible. If $c > 0$, then we take λ large and deduce that $\langle h, x \rangle \leq 0$. □

One important consequence of this result is that a closed convex cone is the intersection of closed half-spaces given by hyperplanes through the origin.

If K is a convex cone in an inner product space, its *dual* K^* is given by

$$K^* = \{y \in V : \langle y, x \rangle \geq 0, \ \forall x \in K\}.$$

Since K^* is defined as the intersection of closed half-spaces, it is closed (whether or not K is). A cone K is *self dual* if $K = K^*$. It is easy to see that if K is the set of non-negative vectors in $\mathbb{R}^d$, or the set of all positive semidefinite matrices, then it is self dual. We observe that if K and L are convex cones and $K \subseteq L$, then $L^* \subseteq K^*$. Similarly it is easy to check that for any convex cone K we have $K \subseteq K^{**}$. In fact:

3.9.2 Theorem. *If K is a closed convex cone, then $K = K^{**}$.* □

3.10 Bases for Convex Cones

Let K be a convex cone. A subset B of K is a *base* for K if $0 \notin B$ and, for each vector x in $K \setminus 0$, there is a unique vector b in B and a unique positive scalar λ such that $x = \lambda b$. Thus we see that the rays generated by the elements of B partition the non-zero vectors in K, and each ray meets B in exactly one point.

To give one example, if K is the cone of positive semidefinite matrices, then the set

$$\{M \in K : \mathrm{tr}(M) = 1\}$$

is a compact base for K. More generally, if K is closed and H is an affine hyperplane that does not contain 0 and $H \cap K$ is bounded, then $H \cap K$ is a base for K that is compact and convex. (We do not require bases to be convex, but they will be usually.)

3.10.1 Lemma. *If a cone in* $\mathbb{R}^m$ *has a compact base, it is closed.*

Proof. Let B be a compact base for K and let u be a point not in K. We show that there is an open neighbourhood of u disjoint from K. Define δ by

$$\delta = \min\{\|x\| : x \in B\},$$

thus δ is the minimum distance of a point in B from 0. Let λ_1 denote $(\|u\| + 1)\delta$ and let U_1 be the open ball of radius λ_1 centred on u. If $\lambda > \lambda_1$, then $\lambda B \cap U_1 = \emptyset$.

Assume $X = [0, \lambda_1] \times B$ and consider the map $\phi : X \to \mathbb{R}^m$ given by

$$\phi(\lambda, x) = \lambda x.$$

Since B is compact, so is X. The image $\phi(X)$ of X is compact and so closed in $\mathbb{R}^m$. As $u \notin K$, we see that $u \notin \phi(X)$, and therefore there is a neighbourhood U_2 of u such that $U \cap \phi(X) = \emptyset$. If $\lambda \geq 0$, then

$$(U_1 \cap U_2) \cap \lambda B = \emptyset.$$

□

Chapter 4

Lovász Theta

Assume we have a fixed graph G and define a *word* to be a sequence of vertices of G. Two words α and β are *confoundable* if for each i either $\alpha_i = \beta_i$ of α_i and β_i are adjacent in G. We are concerned with the maximum size of a set of pairwise non-confoundable words of length k. In this chapter we work through Lovász's paper on the Shannon capacity of a graph.

4.1 The Strong Product

Let G and H be graphs. The *strong product* $G \boxtimes H$ of G and H is the graph with vertex set $V(G) \times V(H)$, where pairs (u, x) and (v, y) are adjacent if

(a) $u = v$ and $x \sim y$, or

(b) $u \sim v$ and $x = y$, or

(c) $u \sim v$ and $x \sim y$.

On special case: if $G = K_m$ and $H = K_n$, then $G \boxtimes H = K_{mn}$. If $\alpha(G)$ is the maximum size of a coclique in G and $\omega(G)$ is the maximum size of a clique, you may prove that

$$\alpha(G \boxtimes H) \le \alpha(H)\alpha(H), \qquad \omega(G \boxtimes H) = \omega(G)\omega(H).$$

The strong product is associative, i.e.,

$$(G \boxtimes H) \boxtimes K \cong G \boxtimes (H \boxtimes K).$$

we use $G^{\boxtimes k}$ to denote the strong product of k copies of G.

4.1.1 Lemma. *If G and H are graphs with respective adjacency matrics $A(G) and A(H)$, then*

$$A(G \boxtimes H) = I \otimes A(H) + A(G) \otimes I + A(G) \otimes A(H). \qquad \square$$

4.1.2 Lemma. *A set of words of length k over G is non-confoundable if and only if it forms a coclique in $G^{\boxtimes k}$.*

The *complement* of $\overline{G}$ of the graph G is the graph with vertex set $V(G)$, wehre two vertices are adjacent in $\overline{G}$ if they are distinct and not adjacent in G.

4.1.3 Lemma. *For any graph G we have $\alpha(G \boxtimes \overline{G}) \geq |V(G)|$.*

Proof. The diagonal vertices (i, i) form a coclique of size $|V(G)|$. $\square$

Since C_5 is self-complementary, if follows that $\alpha(C_5 \otimes C_5) \geq 5$, which is strictly greater than $\alpha(C_5)^2$.

The *Shannon capacity* of $\Theta(G)$ the graph G is defined by

$$\Theta(G) = \sup_k \alpha(G^{\boxtimes k})^{1/k}.$$

It can be shown that

$$\sup_k \alpha(G^{\boxtimes k})^{1/k} = \lim_k \alpha(G^{\boxtimes k})^{1/k}.$$

Since $\alpha(G \boxtimes H) \geq \alpha(G)\alpha(H)$ we see that $\alpha(G) \leq \Theta(G)$. Our aim in this chapter is to derive upper bounds on $\Theta(G)$. We observe that if $\mu(G)$ is a graph parameter such that

(a) $\alpha(G) \leq \mu(G)$,

1. $\mu(G \boxtimes H) \leq \mu(G)\mu(H)$,

then

$$\alpha(G^{\boxtimes k}) \leq \mu(G^{\boxtimes k}) \leq \mu(G)^k$$

whence $\Theta(G) \leq \mu(G)$. If (b) holds we say the parameter $\mu(G)$ is submultiplicative, and thus any submultiplicative parameter that is an upper bound for $\alpha(G)$ is an upper bound for $\Theta(G)$.

4.2 Graph Homomorphisms

If G and H are graphs, a map $\psi : V(G) \to V(H)$ is a *graph homomorphism* if, for each pair of adjacent vertices i and j in G, their images $\psi(i)$ and $\psi(j)$ are adjacent in H. Examples:

(a) If G is a subgraph of H, the identity map is a homomorphism.

(b) A proper m-colouring of G is a homomorphism from G to K_m.

(c) An automorphism of G is a homomorphism.

(d) If G is bipartite and has at least one edge, then we have homomorphisms $K_2 \to G$ and $G \to K_2$. (We say that G is *homomorphically equivalent* to K_2.)

(e) Let $\Omega(d)$ be the graph whose vertices are the unit vectors in $\mathbb{R}^d$, where two vectors are adjacent if they are orthogonal. A homomorphism from $\overline{G}$ to $\Omega(d)$ is called an *orthonormal representation* of G.

The cliques of maximum size in $\Omega(d)$ are the orthonormal bases of $\mathbb{R}^d$. Thus each clique in $\Omega(d)$ is contained in a clique of size d.

If $\psi : G \to H$ is a graph homomorphism and $v \in V(H)$, the set

$$\psi^{-1}(v) = \{u \in V(G) : \psi(u) = v\}$$

is called a *fibre* of ψ. Each fibre is necessarily a coclique in G.

4.3 Fractional Colourings

Let G be a graph and let N be the matrix whose columns are the characteristic vectors of the cocliques of G. Note that G has an m-colouring if and only if there are m cocliques in G whose union is $V(G)$. The union of a set of cocliques with characteristic vectors $u_1, \ldots, u_m$ is equal to $V(G)$ if and only if all entries of $\sum_i u_1$ are positive, and therefore the 01-vectors x such that $\mathbf{1}^T x = m$ and $Nx \geq 1$ correspond to the m-colourings of G. Consequently the value of the linear program

$$\min \mathbf{1}^T x, \qquad Nx \geq \mathbf{1}, \qquad x \geq 0$$

is a lower bound in the chromatic number $\chi(G)$ of G. We denote this lower bound by $\chi_f(G)$ and refer to it as the *fractional chromatic number* of G. An optimal solution to the LP is a *fractional colouring*. The dual to this LP is

$$\max y^T\mathbf{1}, \qquad y^T N \le \mathbf{1}^T, \qquad y \ge 0.$$

The 01-vectors yv such that $y^T N \le \mathbf{1}$ must be characterisitic vectors of cliques in G. Therefore the value of the dual is an upper bound on $\omega(G)$; we denote this value by $\omega_f(G)$ and we refer to an optimal solution as a *fractional clique*. We have

$$\omega(G) \le \omega_f(G) = \chi_f(G) \le \chi(G).$$

One consequence of this is that

$$\alpha(G) \le \chi_f(\overline{G}).$$

We now prove that

$$\chi_f(G \boxtimes H) \le \chi_f(G)\chi_f(H),$$

from which it follows that $\Theta(G) \le \chi_f(G)$.

Let N' be the matrix we get from N by deleting the columns corresponding to cocliques that are not maximal (by inclusion). If $Nx \ge 1$, there is a non-negative vector z such that $Nz \ge 1$ and $\mathbf{1}^T z = \mathbf{1}^T x$—just replace each coclique in the support of x by a maximal coclique.

For any graph G, let M_G denote the matrix whose columns are the maximal (by inclusion) cliques in G. Then

$$M_{G\boxtimes H} = M_G \otimes M_H.$$

Now $\chi_f(G \boxtimes H)$, the minimum value of $\mathbf{1}^T z$ for non-negative vectors z such that

$$\mathbf{1}^T z \le M_{G\boxtimes H} z = (M_G \otimes M_H)z,$$

is bounded above by the minimum value for non-negative vectors x and y such that

$$\mathbf{1}^T(x \otimes y) \le (M_G \otimes M_H)(x \otimes y) = M_G x \otimes M_H y,$$

and this is equal to $\chi_f(\overline{G})\chi_f(\overline{H})$. Thus we have Shannon's result:

4.3.1 Theorem. *For any graph G, we have $\Theta(G) \le \chi_f(\overline{G})$.* □

4.4 Using Symmetry

An automorphism of a graph G is a permutation γ of its vertices such $u^\gamma \sim v^\gamma$ if and only if $u \sim v$. Each permutation can be represented by a permutation matrix $P(\gamma)$ such that $Pe_i = e_{i\gamma}$; the permutation is an automorphism if and only if P commutes with the adjacency matrix of the graph. Two vertices u and v lie in the same orbit of $\mathrm{Aut}(G)$ if there is an automorphism γ such that $u^\gamma = v$. A graph is *vertex transitive* if there is only one orbit.

Let N be our vertex-coclique incidence matrix. If P represents an automorphism of G, then $PN = NQ$ for some permutation matrix Q. So if $y^T N \leq \mathbf{1}^T$,

$$y^T PN = y^T NQ \leq \mathbf{1}^T Q = \mathbf{1}^T$$

and hence $y^T P$ is feasible if and only if y is. Further, since $P\mathbf{1} = \mathbf{1}$,

$$y^T \mathbf{1} = y^T P \mathbf{1}.$$

Therefore

$$\hat{y} = \frac{1}{|\mathrm{Aut}(G)|} \sum_{P \in \mathrm{Aut}(G)} y^T P$$

is a convex combination of feasible solutions and so is itself feasible, with the same value as y. Further it is constant on the orbits of $\mathrm{Aut}(G)$. (You should prove this.) Since our choice of y was arbitrary, we can assume that $\hat{y}$ is optimal. If G is vertex transitive, we conclude that there is an optimal solution to

$$\max y^T \mathbf{1}, \qquad y^T N \leq \mathbf{1}^T, \qquad y \geq 0$$

that is a constant vector. This constant vector must be

$$\frac{1}{\alpha(G)} \mathbf{1}$$

and its value is $v/\alpha(G)$.

4.4.1 Lemma. *If G is a vertex-transitive graph on v vertices, then*

$$\chi_f(G) = \frac{v}{\alpha(G)}. \qquad \square$$

4.4.2 Corollary. *If G is a vertex-transitive graph on v vertices, then*

$$\theta(G)\omega(G) \leq v. \qquad \square$$

This implies that, for vertex-transitive graphs, $\alpha(G)\omega(G) \leq v$.

4.5 Lovász $\theta(G)$

Lovász showed how to use the geometry of orthonormal representations of graphs to derive improved bounds on the Shannon capacity. We first define a parameter. Suppose the vectors $u_1, \dots, u_n$ provide an orthonormal representation of G. Then

$$\theta(G) := \min_{\|c\|=1} \max_{1 \le i \le n} \frac{1}{\langle c, u_i \rangle^2}.$$

A vector c which realizes the value $\theta(G)$ is called the *handle* of the representation.

4.5.1 Lemma. *For any graph G we have $\theta(G) \le \alpha(G)$.*

Proof. Let S be a coclique in G and let $u_1, \dots, u_n$ be an orthonormal representation of G. Then

$$1 = \langle c, c \rangle \ge \sum_{i \in S} \langle c, u_i \rangle^2 \ge \frac{\alpha(G)}{\theta(G)},$$

and the lemma follows. □

As a (very relevant) example, we show that $\theta(C_5) \le \sqrt{5}$. Let the unit vectors $u_0, \dots, u_4$ be the vertices of a regular pentagon centred at the origin. Then

$$\langle u_0, u_2 \rangle = \cos(4\pi/5) \approx -0.8090$$

and so we do not have an orthonormal representation yet. View $u_0, \dots, u_4$ as vectors in the xy-plane in $\mathbb{R}^3$. If y is a scalar multiple of the vector $(0, 0, 1)$, then $\langle u_i, y \rangle = 0$ and

$$\langle u_i + y, u_j + y \rangle = \langle u_i, u_j \rangle + \langle y, y \rangle.$$

Therefore if we choose y so its squared length is $-\cos(4\pi/5)$, then the vectors

$$\frac{1}{\|u_i + y\|}(u_i + y), \qquad (i = 1, \dots, n)$$

are an orthonormal representation of C_5 and the vector $(0, 0, 1)$ is a handle for this representation. The square of the inner product of the handle with any of the above vectors is

$$\frac{-\cos(4\pi/5)}{1 - \cos(4\pi/5)} = \frac{1}{\sqrt{5}}.$$

We now prove that $\theta(G)$ is submultiplicative:

4.5.2 Lemma. *For any graphs G and H we have $\theta(G \boxtimes H) \leq \theta(G)\theta(H)$.*

Proof. Let $u_1, \ldots, u_m$ and $v_1, \ldots, v_n$ be optimal orthonormal representations with handles c and d respectively. Then the vectors $u_i \otimes v_j$ form an orthonormal representation for $G \boxtimes H$ and

$$\begin{aligned} \theta(G \boxtimes H) &\leq \max_{i,j} \frac{1}{\langle c \otimes d, u_i \otimes v_j \rangle^2} \\ &= \max_i \frac{1}{\langle c, u_i \rangle^2} \max_j \frac{1}{\langle d, v_j \rangle^2} \\ &= \theta(G)\theta(H). \end{aligned}$$ □

We have shown that $\theta(G)$ is a submultiplicative upper bound for $\alpha(G)$, hence:

4.5.3 Theorem. *For any graph G we have $\theta(G) \leq \Theta(G)$.* □

We note that we follow the usual convention that a vertex is not adjacent to itself. In his paper Lovász redefines adjacency so each vertex is adjacent to itself.

4.6 A Semidefinite Program?

We present an optimization problem whose value is $\theta(G)$. We use $\lambda_1(M)$ to denote the larges eigenvalue of the matrix M (which will invariably be symmetric).

4.6.1 Theorem. *Let G be a graph with vertex set $\{1, \ldots, n\}$ and let $\Phi(G)$ be the set of symmetric $n \times n$ matrices such $M_{i,j} = 1$ if $i = j$ or i and j are not adjacent. Then*

$$\theta(G) = \min_{M \in \Phi} \lambda_1(M).$$

Proof. Let $u_1, \ldots, u_n$ be an optimal orthonormal representation for G with handle c. Define vectors v_i by

$$v_i := c - \langle c, u_i \rangle^{-1} u_i;$$

then

$$\langle v_i, v_j \rangle = -1 + \frac{\langle u_i, u_j \rangle}{\langle c, u_i \rangle \langle c, u_j \rangle}.$$

Let D be the $n \times n$ diagonal matrix with ii-entry equal to $1/\langle c, u_i \rangle^2$. If we use C to denote the Gram matrix of the vectors v_i and set $M = D - C$, then $M \in \Phi$ and $D - M \succcurlyeq 0$. As $\theta(G)I \succcurlyeq D$, we see that $\theta(G)I - M \succcurlyeq 0$ and therefore $\theta(G) \geq \lambda_1(M)$. It follows that

$$\theta(G) \geq \min_{M \in \Phi} \lambda_1(M).$$

To get equality, suppose $M \in \Phi$ and $\lambda = \lambda_1(M)$. Then $\lambda I - M \succcurlyeq$ and accordingly $\lambda I - M$ is the Gram matrix of vectors $x_1, \ldots, x_n$. Let c be a unit vector orthogonal to each of these vectors and define

$$u_i =:= \frac{1}{\sqrt{\lambda}}(c + x_i).$$

Then u_i is a unit vector and, if i and j are distinct and

$$\langle u_i, u_j \rangle = \frac{1}{\lambda}(1 + \langle u_i, u_j \rangle) = 0;$$

hence these vectors form an orthonormal representation of G. As $\langle c, u_i \rangle^2 = 1/\lambda$, it follows that $\theta(G) \leq \lambda$, that is

$$\theta(G) \leq \min_{M \in \Phi} \lambda_1(M). \qquad \square$$

In the first part of this proof, if $D \neq \theta(G)I$ then $\theta(G)$ does not meet the upper bound. We conclude that if $u_1, \ldots, u_n$ is an optimal orthonormal representation for G with handle c, then $\langle c, u_i \rangle$ is independent of the vertex i.

If G is a regular graph, we can derive an explicit upper bound for $\theta(G)$. Suppose G is k-regular with adjacency matrix A. Then for any real t, we have $J - tA \in \Phi$; we compute $\lambda_1(J - tI)$. First, $\mathbf{1}$ is an eigenvector for $J - tI$ with eigenvalue $n - tk$. Suppose z is an eigenvector for A with eigenvalue λ and $\langle \mathbf{1}, z \rangle = 0$. Then

$$(J - tA)z = \mathbf{1}\mathbf{1}^T z - tAz = -t\lambda z$$

and, if τ is the least eigenvalue of A, the largest eigenvalue of $J - tA$ is minimal when $n - tk = -t\tau$. In this case $t = n/(k - \tau)$ and $\lambda_1(J - tA) = n(-\tau)/(k - \tau)$. So we have the bounds

$$\alpha(G) \leq \theta(G) \leq \frac{n(-\tau)}{k - \tau}.$$

This bound is tight for complete graphs, regular bipartite graphs and the Petersen graph (for which $\tau = -2$).

4.7 A Semidefinite Program

The program for $\theta(G)$ in the previous section is not obviously a semidefinite program. In this section we present a semidefinite program for $\theta(G)$. We make extensive use of the fact that for symmetric matrices A and B of the same order,

$$\langle A, B\rangle = \operatorname{tr} AB = \operatorname{sum}(A \circ B).$$

4.7.1 Theorem. *Let G be a graph with vertex set $\{1, \dots, n\}$ and let Ω be the set of $n \times n$ positive semidefinite matrices N such that* $\operatorname{tr}(N) = 1$ *and $N_{i,j} = 0$ if i and j are adjacent. Then*

$$\theta(G) = \max_{N \in \Omega} \langle J, N\rangle.$$

Proof. Let M any matrix in Φ (defined in the statement of Theorem 4.6.1) with largest eigenvalue $\theta(G)$ and suppose $N \in \Omega$. Then $\langle J, N\rangle = \langle M, N\rangle$ and consequently

$$\theta(G) - \langle J, N\rangle = \theta(G)I - \langle M, N\rangle = \langle \theta(G)I - M, N\rangle \succcurlyeq 0.$$

This shows that $\theta(G) \geq \max_{N\in\Omega}\langle J, N\rangle$, we need to show that equality holds.

Let A be the adjacency matrix of G. If h is a unit vector in $\mathbb{R}^n$, define $\hat{h}$ to be the ordered pair in $\mathbb{R}^{n^2} \times \mathbb{R}$ given by

$$\hat{h} = (A \circ hh^T, \langle \mathbf{1}, h\rangle^2)$$

and define z in $\mathbb{R}^{n^2} \times \mathbb{R}$ by

$$z = (0, \theta(G)).$$

We claim that z lies in the convex hull of the vectors $\hat{h}$ (as h runs over all unit vectors in $\mathbb{R}^n$). If our claim were false, there would be a separating hyperplane, that is, a matrix B and a scalar y and a scalar β such that

$$\langle B, A \circ hh^T \rangle + y\langle \mathbf{1}, h\rangle^2 \leq \beta \tag{4.7.1}$$

but $y\theta(G) > \beta$. Taking $h = e_1$, we also see that

$$y = y\langle \mathbf{1}, e_1\rangle^2 \leq \beta$$

and so $\beta > 0$. Hence $y\theta(G) > \beta \geq y$ and, as $\theta(G) \geq 1$ we see that $y > 0$. By rescaling if needed, we may assume $y = 1$ and then $\beta < \theta(G)$. Since

$$\langle B, A \circ hh^T \rangle = \operatorname{sum}(B \circ A \circ hh^T),$$

we can further assume that $B_{i,j} = 0$ if $i = j$ or $i \sim j$ (or, equivalently, take $B = B \circ A$.)

Now we define M by

$$M = J + B$$

and observe that

$$\begin{aligned}\langle B, A \circ hh^T \rangle + \langle \mathbf{1}, h\rangle^2 &= \operatorname{sum}(B \circ A \circ hh^T) + \operatorname{tr}(Jhh^T)\\ &= \operatorname{sum}(B \circ hh^T) + \operatorname{sum}(hh^T)\\ &= \operatorname{sum}((B + J) \circ hh^T)\\ &= \operatorname{sum}(M \circ hh^T)\\ &= \operatorname{tr}(Mhh^T)\\ &= h^T M h.\end{aligned}$$

By Equation (4.7.1) (with $y = 1$) we see that $h^T M h \leq \beta$ for all unit vectors h, and therefore $\lambda_1(M) \leq \beta < \theta(G)$. On the other hand, $M \in \Phi$ and therefore $\lambda_1(M) \geq \theta(G)$. This contradiction forces us to conclude that z lies in the convex hull of the vectors $\hat{h}$.

We now construct our optimal solution. We know there are non-negative scalars $a_1, \ldots, a_m$ summing to 1 and unit vectors $h_1, \ldots, h_m$ such that

$$\sum_{i=1}^m a_i A \circ h_i h_i^T = 0, \qquad \sum_{i=1}^m a_i \langle \mathbf{1}, h\rangle^2 = \theta(G).$$

Define

$$N = \sum_{i=1}^{m} a_i h_i h_i^T;$$

clearly $N \in \Omega$ and

$$\langle J, N\rangle = \sum_{i=1}^{m} a_i \langle \mathbf{1}, h_i\rangle^2 = \theta(G).$$

Hence N is an optimal solution with value $\theta(G)$. □

4.8 Using Complements

4.8.1 Theorem. *If $v_1, \ldots, v_n$ ranges over all orthonormal representations of $\overline{G}$ and d ranges over all unit vectors, then*

$$\theta(G) = \max \sum_i \langle d, v_i\rangle^2.$$

Proof. Let $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ be orthonormal representations (of G and $\overline{G}$ respectively, and let c and d be any vectors. Then

$$\langle u_i \otimes v_i, u_j \otimes v_j\rangle = \delta_{i,j},$$

hence the vectors $u_i \otimes v_i$ are orthogonal and therefore

$$\sum_{i=1}^{n} \langle c \otimes d, u_i \otimes v_i\rangle^2 \le \langle c \otimes d, c \otimes d\rangle = \langle c, c\rangle \langle d, d\rangle.$$

Accordingly

$$\sum_{i=1}^{n} \langle c, u_i\rangle^2 \langle d, v_i\rangle^2 \le \langle c, c\rangle \langle d, d\rangle$$

and if we assume that $u_1, \ldots, u_n$ is optimal and c is a handle for it, we deduce that

$$\sum_{i=1}^{n} \langle d, v_i\rangle^2 \le \theta(G).$$

To complete the proof we must show that equality holds. Let N a matrix in the set Ω of Theorem 4.7.1. Since $N \succcurlyeq 0$, it is the Gram matrix of a set of vectors $w_1, \ldots, w_n$ where

$$1 = \operatorname{tr}(N) = \sum_i \langle w_i, w_i\rangle$$

and

$$\theta(G) = \langle J, N\rangle = \mathbf{1}^T N \mathbf{1} = \langle \sum_i w_i, \sum_i w_i\rangle.$$

If we define

$$v_i = \frac{1}{\|w_i\|}, \qquad d = \frac{1}{\|\sum_i w_i\|} \sum_i w_i,$$

then $v_1, \ldots, v_n$ is an orthonormal representation of $\overline{G}$ and, using Cauchy-Schwarz, we have

$$\begin{aligned}\sum_i \langle d, v_i\rangle^2 &= \left(\sum_i \langle w_i, w_i\rangle\right)\left(\sum_i \langle d, v_i\rangle^2\right)\\ &\geq \left(\sum_i \|w_i\| \langle d, v_i\rangle\right)^2\\ &= \left(\sum_i \langle d, w_i\rangle\right)^2\\ &= \langle d, \sum_i w_i\rangle^2\\ &= \langle \sum_i w_i, \sum_i w_i\rangle\\ &= \theta(G). \qquad \square\end{aligned}$$

Suppose $v_1, \ldots, v_n$ is an orthonormal representation for $\overline{G}$ and let B denote the matrix with $v_1, \ldots, v_n$ as its columns. Then $B^T B$ is the Gram matrix of $v_1, \ldots, v_n$, and it is a positive semidefinite matrix with $B_{i,i} = 1$ for all vertices i abd $B_{i,j} = 0$ if i and j are not adjacent. We also note that for any vector d

$$\sum_i \langle d, v_i\rangle = d^T B B^T d,$$

and therefore the maximum value of the sum over unit vectors d is $\lambda_1(BB^T)$.

4.8.2 Lemma. *If B is an $m \times n$ matrix and C is an $n \times m$ matrix, then BC and CB have the same non-zero eigenvalues with the same multiplicities.* □

From this lemma we see that $\lambda_1(BB^T) = \lambda_1(B^T B)$, from which we get:

4.8.3 Corollary. *Let Ψ be the set of the Gram matrices of the orthonormal representations of $\overline{G}$. Then*

$$\theta(G) = \max_{Q \in \Psi} \lambda_1(Q). \qquad \square$$

4.9 Other Bounds on the Shannon Capacity

4.9.1 Theorem. *For any graph G we have $\theta(G) \leq \alpha_f(G)$.*

Proof. Let $u_1, \ldots, u_n$ be an orthonormal representation of $\overline{G}$ and let c be a unit vector such that

$$\theta(G) = \sum_i \langle c, u_i \rangle^2.$$

(The representation and the vector c exist by Theorem **??**.) If C is a clique in G then the vectors

$$\{u_i : i \in C\}$$

are orthogonal and therefore

$$\sum_{i \in C} \langle c, u_i \rangle^2 \leq \langle c, c \rangle = 1.$$

So if N is the vertex-clique incidence matrix of G and z is the vector with $z_i = \langle c, u_i \rangle^2$, then

$$z^T N \leq \mathbf{1}^T$$

and $\langle \mathbf{1}, z \rangle$ is bounded above by $\alpha_f(G)$. □

Note that $\alpha_f(G) = \chi_f(\overline{G}) = \chi_f(\overline{G})$.

4.9.2 Theorem. *If G admits an orthonormal representation in $\mathbb{R}^d$, then $\theta(G) \leq d$.*

Proof. Let $u_1, \ldots, u_n$ be an orthonormal representation of G in $\mathbb{R}^d$. Then the vectors $u_i \otimes u_i$ also provide an orthonormal representation of G. Let $e_1, \ldots, e_d$ be an orthonormal basis for $\mathbb{R}^d$ and set

$$b = \frac{1}{\sqrt{d}} \sum_i u_i \otimes u_i.$$

Then $\langle b, b \rangle = 1$ and

$$\langle u_i \otimes u_i, b \rangle = \frac{1}{\sqrt{d}} \sum_k \langle e_k \otimes e_k, u_k \otimes u_k \rangle = \frac{1}{\sqrt{d}} \sum_k \langle e_k, u_i \rangle^2 = \frac{1}{\sqrt{d}}.$$

We conclude that $\theta(G) \leq d$. □

Chapter 5

Duality and Semidefinite Programs

We study duality in conic optimization problems: this is a class of optimization problems just general enough to include linear and semidefinite optimization as special cases.

5.1 Conic Programs

If K is a closed convex cone, the optimization problem

$$\sup\langle C, X\rangle, \qquad \langle A_i, X\rangle = b_i \quad (i = 1, \ldots, m), \qquad X \in K$$

is a *conic programming problem*. There are three important special cases:

(a) $K = (\mathbb{R}^n)_{\geq 0}$ (non-negative vectors in $\mathbb{R}^n$).

(b) $K = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \|x\| \leq t\}$, the Lorentz cone.

(c) K is the set of positive semidefinite matrices.

Each of these cones in homogeneous and self-dual. There is a second optimization problem related to the above conic programming problem:

$$\inf y^T b, \qquad \sum_i y^T A_i - C \in K^*.$$

(Here $b = (b_1, \ldots, b_m)^T$.) We declare this to be the dual to previous primal problem.

We have

$$\begin{aligned} y^T b - \langle C, X\rangle &= \sum_i y_i b_i - \langle \sum_i y_i A_i, X\rangle + \langle \sum_i y_i A_i, X\rangle - \langle C, X\rangle \\ &= \sum_i y_i(b_i - \langle A_i, X\rangle) + \langle \sum_i y_i A_i - C, X\rangle \end{aligned}$$

If X is primal feasible and y is dual feasible, then the first term in this sum is zero and, since $\sum_i y_i A_i - C \in K^*$ and $C \in K$, the second term is non-negative. This has brought us to the weak duality theorem for conic programs:

5.1.1 Theorem. *If y is dual feasible and X is primal feasible, then*

$$y^T b \geq \langle C, X\rangle$$

and equality holds if and only if

$$\langle \sum_i y_i A_i - C, X\rangle = 0. \qquad \square$$

We note that if $M, N \succcurlyeq 0$, then $\langle M, N\rangle = 0$ if and only if $MN = 0$.

5.2 Difficulties with Duality

In dealing with linear programming, we learnt that if an LP and its dual were both feasible, then both the primal and dual had optimal solutions, and the value of these solutions were equal. For conic programs, life is not so kind, as we illustrate by examples.

We assume our primal program is $\sup\langle C, X\rangle$ subject to conditions

$$\langle A_i, X\rangle = b_i, \qquad X \in \mathcal{K}.$$

(Here $\mathcal{K}$ is a pointed, closed, convex cone.) Thus we can specify both problems by giving C, the matrices $A_1, \ldots, A_m$, the vector b, and $\mathcal{K}$. The dual program requires us to find $\inf y^T b$ subject to the condition

$$\sum_i y_i A_i - C \in \mathcal{K}^*.$$

In understanding these examples, it will be useful to recall that for symmetric matrices,

$$\langle M, N\rangle = \operatorname{tr}(MN) = \operatorname{sum} M \circ N.$$

For our first example

$$C = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}; \quad A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}; \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The constraint $\langle A_1, X\rangle = 1$ implies that $X_{1,1} = 1$; the constraint $\langle A_2, X\rangle = 0$ implies that $X_{2,2} = 0$ and, since $X \succcurlyeq 0$, this implies that $X_{1,2} = X_{2,1} = 0$. Consequently

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

is the only feasible solution, hence it is optimal with value 1.

The dual program to find $\inf y_1$ subject to

$$y_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + y_2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \succcurlyeq 0.$$

The left side here is

$$\begin{pmatrix} y_1 & 1 \\ 1 & y_2 \end{pmatrix}$$

which you may verify is positive semidefinite if and only $y_1, y_2 \geq 0$ and $y_1 y_2 \geq 0$. From this it is clear that value of the dual is 0, but it is not attained.

For the second example

$$C = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} A_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

These constraints on the primal imply that

$$X_{1,1} = 0, \quad 2X_{1,3} + X_{2,2} = 1,$$

whence the first row and column of X must be zero and $X_{2,2} = 1$. Therefore all feasible solutions have $X_{2,2} = 1$ and their value is -1. In particular the matrix

$$E_{2,2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

is optimal with value -1. Turning to the dual we find that

$$y_1A_1 + y_2A_2 - C = \begin{pmatrix} y_1+1 & 0 & y_2 \\ 0 & y_2+1 & 0 \\ y_2 & 0 & 0 \end{pmatrix}$$

and for this to be positive semidefinite, we must have $y_2 = 0$. The value of a feasible solution is $-y_2$. Hence the optimal value is 0 and this is realized when $y = 0$. In this example both the primal and dual programs attain their optimal solutions, by we have a *duality gap*, that is $\langle C, X\rangle < y^Tb$ for the optimal X and y.

To summarize, we have seen two ways in which duality for conic programs might fail:

(a) The value of the dual (or primal) might not be attained, or

(b) There is a duality gap.

5.3 Duals of Duals

Our formulation of the primal and dual conic programming problems appears to lack symmetry. Here we show that, despite this, the dual of the dual is the primal. To this end, we rewrite the primal and dual programs.

We assume $\mathcal{K}$ is a closed pointed convex cone; our primal program is

$$\sup\langle C, X\rangle, \quad \langle A_i, X\rangle = b_i \ (i = 1, \ldots, m), \quad X \in \mathcal{K}$$

and its dual is

$$\inf y^Tb, \quad \sum_i A_i - C \in \mathcal{K}^*.$$

Start with the primal. Assuming it is feasible, choose X_0 such that

$$\langle A_i, X_0\rangle = b_i$$

and define $\mathcal{L}$ to be the space of symmetric matrices X such that

$$\langle A_i, X\rangle = 0$$

for each i. Its orthogonal complement is

$$\mathcal{L}^\perp = \{\sum_i y_iA_i : y_i \in \mathbb{R}\}.$$

We see that $\langle A_i, X\rangle = b_i$ for all i if and only if $X \in X_0 + L$. (Here $X_0 + L$ is a coset of L.) So our primal becomes

$$\sup\langle C, X\rangle, \quad X \in (X_0 + L) \cap \mathcal{K}.$$

For the dual

$$y^T b = \sum_i y_i \langle A_i, X_0\rangle = \langle \sum_i y_i A_i, X_0\rangle = \langle C, X_0\rangle + \langle \sum_i y_i A_i - C, X_0\rangle$$

and so we can write as

$$\langle C, X_0\rangle + \inf\langle X_0, Z\rangle, \quad Z \in (-C + \mathcal{L}^\perp) \cap \mathcal{K}^*.$$

5.4 Strong Duality

We continue to work with primal and dual conic programs in canonical form:

$$\sup\langle C, X\rangle, \quad \langle A_i, X\rangle = b_i \ (i = 1, \ldots, m), \quad X \in \mathcal{K}$$

and

$$\inf y^T b, \quad \sum_i y_i A_i - C \in \mathcal{K}^*$$

respectively.

5.4.1 Theorem. *If there is a vector y such that $\sum_i y_i A_i - C \in \operatorname{int}(\mathcal{K}^*)$ and the value of the dual is bounded below, the primal problem has an optimal solution and there is no duality gap.*

Proof. Let d^* denote the optimal value of the dual. We define

$$\mathcal{M} = \Big\{\sum_i y_i A_i - C, \ y^T b \le d^*\Big\}.$$

(Note that $\mathcal{M} \cap \mathcal{K}^*$ is the set of dual-optimal solutions and, since the dual is feasible, $\mathcal{M} \neq \emptyset$.)

There is one trivial case: if $b = 0$ then $X = 0$ is optimal and both primal and dual have the value 0. The rest of the proof breaks into three parts (and some window dressing.)

We claim first that $\mathcal{M} \cap \operatorname{int}(\mathcal{K}^*) = \emptyset$. Otherwise there is a vector y such that

$$\sum_i y_i A_i - C \in \operatorname{int}(\mathcal{K}^*), \quad y^T b \leq d^*.$$

Since $b \neq 0$, there is an index i such that $b_i \neq 0$, say $i = 1$. If $b_1 < 0$, then

$$(y_1 + \epsilon) A_1 + \sum_{i>1} y_i A_i - C \in \mathcal{K}^*$$

for small enough positive values of ϵ and therefore,

$$(y_1 + \epsilon) b_1 + \sum_{i>1} y_i b_i < y^T b \leq d^*.$$

This contradicts our assumption that d^* is the value of the dual. If $b_1 > 0$, repeat the above with ϵ small and negative.

Since $\mathcal{M}$ and $\mathcal{K}$ are convex sets whose relative interiors are disjoint, they can be separated by an affine hyperplane. So there is X such that

$$\sup\{\langle X, Z\rangle : Z \in \mathcal{M}\} \leq \inf\{\langle X, Z\rangle : Z \in \mathcal{K}^*\}. \tag{5.4.1}$$

We use this X to construct an optimal solution to the primal.

First we show that $X \in \mathcal{K}$. For this it is enough to show that

$$\inf_{Z \in \mathcal{K}^*} \langle X, Z\rangle \geq 0,$$

because this implies $X \in (\mathcal{K}^*)^* = \mathcal{K}$. If $Z_1 \in \mathcal{K}^*$ and $\langle X, Z_1\rangle < 0$ then

$$\langle X, \lambda Z_1\rangle \to -\infty$$

as λ increases, and this contradicts our assumption that the value of the dual is finite.

Next we claim that there is a positive real number μ such that $\langle A_i, X\rangle = \mu b_i$ for all i and $\langle C, X\rangle \geq \mu d^*$. Since $0 \in \mathcal{K}^*$, we have that

$$\inf_{Z \in \mathcal{K}^*} \langle X, Z\rangle = 0$$

and by Equation (5.4.1) we find that

$$\sup_{X \in \mathcal{M}} \langle X, Z\rangle \leq 0.$$

Therefore if y is such that $y^T b \leq d^*$, then

$$\langle X, \sum_i y_i A_i - C \rangle \leq 0$$

or, equivalently that if $y^T b \leq d^*$, then

$$\sum_i y_i \langle A_i, X \rangle \leq \langle C, X \rangle.$$

Consequently we have shown that that the half-space

$$\{y : y^T b \leq d^*\}$$

is contained in the half-space

$$\{y : \sum_i y_i \langle A_i, X \rangle \leq \langle C, X \rangle\};$$

this implies that the normal vectors for these half-spaces must be parallel.

In other words there is a non-negative real number μ such that

$$\langle A_i, X \rangle = \mu b_i \ (i = 1, \dots, m)$$

and

$$\mu d^* \leq \langle C, X \rangle.$$

To complete the proof of the claim, we show $\mu > 0$. If $\mu = 0$, the previous inequality implies that $\langle C, X \rangle \geq 0$. By hypothesis there is a vector $\bar{y}$ such that $\sum_i \bar{y}_i A_i - C \in \text{int}(\mathcal{K}^*)$. If $\mu = 0$ then for this vector, since $X \neq 0$,

$$0 < \langle \sum_i \bar{y}_i A_i - C, X \rangle = -\langle C, X \rangle.$$

and $\langle C, X \rangle < 0$. This is a contradiction.

To complete the proof of the theorem, we observe that

$$\langle C, \mu^{-1} X \rangle \geq d^*$$

and therefore by complementary slackness $\mu^{-1} X$ is an optimal solution to the primal with value d^*. □

Despite our efforts, we have not proved that, under the hypotheses of the theorem, the optimum value of the dual is obtained.

A dual feasible point in $\text{int}\, \mathcal{K}^*$ is sometimes called a *Slater point*.

5.5 Farkas's Lemma

Farkas's lemma tells us that if the system

$$Ax = b, \quad x \geq 0$$

has no solution, there is a vector y such that

$$y^T A \geq 0, \quad y^T b < 0.$$

Thus y can be viewed as certifying that our system is infeasible.

Nothing so strong holds for conic programs, but there is a good approximation (which ought to be enough for an optimization course).

5.5.1 Theorem. *Let $\mathcal{K}$ be a pointed closed convex cone with non-empty interior and assume that the equations*

$$\langle A_i, X \rangle = b_i, \quad (i = 1, \ldots, m)$$

have a solution. Then exactly one of the following statements is true:

(a) There is X in $\operatorname{int}(\mathcal{K})$ such that $\langle A_i, X \rangle = b_i$ for all i.

(b) There is y such that $\sum_i y_i A_i \in \mathcal{K}^* \setminus \{0\}$ and $y^T b \leq 0$.

Proof. If X satisfies (a) and y satisfies (b), then

$$0 \leq \langle \sum_i y_i A_i, X \rangle = \sum_i y_i \langle A_i, X \rangle = y^T b \leq 0$$

but, as $\sum_i y_i A_i \in \operatorname{int}(\mathcal{K}^*)$ and $X \in \operatorname{int}(\mathcal{K})$, we have

$$\langle \sum_i y_i A_i, X \rangle > 0.$$

Hence (a) and (b) cannot both hold.

Assume now that our system of linear equations does not admit a solution in $\operatorname{int}(\mathcal{K})$. Define the subspace $\mathcal{L}$ by

$$\mathcal{L} = \{X : \langle A_i, X \rangle = 0, \ i = 1, \ldots, m\}.$$

Let X_0 be a feasible solution to the system; then the set of feasible solutions is the affine space $X_0 + \mathcal{L}$ and by assumption $(X_0 + \mathcal{L}) \cap \operatorname{int}(\mathcal{K}) = \emptyset$.

There there is a hyperplane separating $X_0 + \mathcal{L}$ and $\mathrm{int}(\mathcal{K})$, equivalently there is a 'vector' C and a scalar β such that for all X in $\mathcal{K}$ we have $\langle C, X\rangle \geq \beta$ and, for all X in $X_0 + \mathcal{L}$ we have $\langle C, X\rangle \leq \beta$. Since $0 \in \mathcal{K}$ it follows that $\beta \leq 0$ and, since $\langle C, tX\rangle \geq \beta$ for all X in $\mathcal{K}$ and $t > 0$, it follows that $C \in \mathcal{K}^*$.

Further, if $X \in \mathcal{L}$ and $t \in \mathbb{R}$, we have

$$\langle C, tX + X_0\rangle \leq \beta,$$

whence $\langle C, X\rangle = 0$. This implies that $C \in \mathcal{L}^\perp$ and consequently C is a linear combination of $A_1, \dots, A_m$, say

$$C = \sum_i y_i A_i$$

for some y. We know that $C \in \mathcal{K}^* \setminus 0$ and as $X_0 \in X_0 + \mathcal{L}$, we have

$$y^T b = \sum_i y_i \langle A_i, X_0\rangle = \langle C, X_0\rangle \leq \beta \leq 0. \qquad \Box$$

5.6 A Second Approach to Duality

[We're following Barvinok.] We take our primal program to be

$$\inf \langle C, X\rangle, \quad \langle A, X_i\rangle = b_i \ (i = 1, \dots m), \quad X \in K$$

with dual

$$\sup y^T b, \quad \sum_j y_j A_j - C \in K^*.$$

Note that we have swapped our inf and sup and so, as you should verify, complementary slackness now implies that $y^T b \leq \langle C, X\rangle$ (with the right weasel words in place).

Given $n \times n$ matrices $A_1, \dots, A_m$, we define a linear map $\widehat{A} : \mathrm{Mat}_{n\times n}(\mathbb{R}) \to \mathbb{R}^{m+1}$ by

$$\widehat{A}(X) = (\langle A_1, X\rangle, \dots, \langle A_m, X\rangle, \langle C, X\rangle).$$

5.6.1 Theorem. *Let K be a convex cone and suppose that the cone $\widehat{A}(K)$ is closed. If there is a primal feasible solution, then the duality gap is zero and, if the primal is bounded below, there is a primal optimal solution.*

Proof. If the primal is not bounded below, there are no primal feaasible solutions and the duality gap is zero.

We assume that that the primal is bounded below, with value γ. Fix b in $\mathbb{R}^m$ and consider the line $L = (b, \tau)$. The intersection $L \cap \widehat{A}(K)$ is a closed set of points

$$\{(b, \langle C, X\rangle)$$

where X is primal feasible. Since there is a primal feasible solution and the primal is bounded below, this intersection is a closed bounded interval or a closed ray that is bounded below. Hence there is a primal feasible solution X such that $\langle C, X\rangle = \gamma$, and this solution is optimal.

Let β denote the optimal value of the dual. By complementary slackness, $\beta \le \gamma$. We claim that if $\epsilon > 0$, there is a dual feasible solution y such that $y^T b \ge \gamma - \epsilon$. This would imply that $\beta = \gamma$.

If our claim was false, then

$$(b, \gamma - \epsilon) \notin \widehat{A}(K)$$

and, since $\widehat{A}(K)$ is closed, this point can be strictly separated from $\widehat{A}(K)$ by a hyperplane. Hence there is a point (y, σ) in $\mathbb{R}^m \oplus \mathbb{R}$ and a number α such that for all dual-feasible y

$$y^T b + \sigma(\gamma - \epsilon) > \alpha$$

and, for all primal feasible X,

$$\sum_j y_j \langle A_j, X\rangle + \sigma\langle C, X\rangle < \alpha.$$

As we may take $X = 0$, we may assume $\alpha > 0$.

Next we show $\sigma < 0$. If c is positive we have

$$\sum_j y_j \langle A_j, cX\rangle + \sigma\langle C, cX\rangle = c\sum_j y_j \langle A_j, X\rangle + c\sigma\langle C, X\rangle < \alpha.$$

Since K is a cone, this implies that

$$\sum_j y_j \langle A_j, X\rangle + \sigma\langle C, X\rangle \le 0$$

for all X in K.

So, for all X in K we must have

$$y^T b + \sigma(\gamma - \epsilon) > 0, \qquad \sum_j y_j^T \langle A_j, X\rangle + \sigma\langle C, X\rangle \leq 0.$$

If X_0 is primal optimal, then $\langle C, X_0\rangle = \gamma$ and

$$ipA_jX = b_j, \quad (j = 1, \ldots, m)$$

and consequently

$$y^T b - (\gamma - \epsilon) > 0$$

and, for all x in K,

$$\sum_j y_j\langle A_j, X\rangle - \langle C, X\rangle = \langle X, \sum_j y_j A_j\rangle - \langle C, X\rangle = \langle X, \sum_j y_j A_j - C\rangle \leq 0.$$

Therefore $C - \sum_j y_j A_j \in K^*$. We conclude that y is dual feasible with $y^T b \geq \gamma - \epsilon$. □

5.7 Strong Duality, Again

We use the theorem from the previous section to derive strong duality. The following lemma will be useful.

5.7.1 Lemma. *Let $T : V \to W$ be linear and let K be a cone in V with a compact convex base. If $\ker(T) \cap K = \{0\}$, then $T(K)$ is a closed convex cone in W.*

Proof. Let B be a compact base for K and set $C = T(B)$. Then C is compact and convex and $0 \notin C$ and, further, $T(K)$ is generated (as a cone) by $T(C)$. By Lemma 3.10.1, we conclude that $T(K)$ is a closed convex cone. □

As in the previous section, our primal problem is

$$\inf\langle C, X\rangle, \quad \langle A_i, X\rangle = b_i \ (i = 1, \ldots, m), \quad X \succcurlyeq 0.$$

5.7.2 Lemma. *If there exist real numbers $y_1, \ldots, y_m$ and a real number ρ such that $\sum_i y_i A_i + \rho C \succ 0$, and the primal is feasible, then the duality gap is zero. Moreover, if the primal is bounded below, there is an optimal solution for the primal.*

Proof. We apply our map $\widehat{A}$ from the previous section. If $X \in \ker(\widehat{A})$, then

$$\langle A_1, X\rangle = \cdots = \langle A_m, X\rangle = \langle C, X\rangle = 0$$

and therefore $\langle B, X\rangle = 0$. Since $B \succ 0$, it follows that $X = 0$. Since the cone of positive semidefinite matrices has a compact base, the lemma follows from the above lemma and Theorem 5.6.1.

Chapter 6

Algorithms

We discuss the standard approach to solving semidefinite programs.

6.1 Convex Functions

The *epigraph* of a real f function on a real vector space V is the subset of $V \times \mathbb{R}$:

$$\{(x,y) : y \geq f(x),\ x \in V\}.$$

A function is *convex* if its domain is convex and its epigraph is a convex subset of $V \times \mathbb{R}$. We call f *concave* if $-f$ is convex.

Any norm on a vector space is a convex function.

Note that if f is convex according to the above definition and $0 \leq a \leq 1$ and $x_1 \leq x_2$, then

$$a(x_1, f(x_1)) + (1-a)(x_2, f(x_2)) = (ax_1 + (1-a)x_2, af(x_1) + (1-a)f(x_2))$$

belongs to the epigraph of f, whence we have

$$af(x_1) + (1-a)f(x_2) \geq af(x_1) + (1-a)f(x_2).$$

Conversely if this holds for all x_1 and x_2, then f is convex. This inequality is often used as the definition of a convex function. If the above inequality is tight except when $t = 0$ or $t = 1$, we say f is *strictly convex*.

If x and y belong to a convex set D and f is a function on D. Then

$$f((1-a)x + ay) = f(x + a(y-x)) =: g(a)$$

and thus f is convex on the line segment joining x to y if and only if g is convex on $[0,1]$. (This is an easy exercise.) Hence we can reduce the problem of deciding if f is convex to a collection of one-variable problems.

6.2 Differentiable Functions

If f is twice-differentiable, there is a vector-valued function ∇f and a matrix-valued function $\nabla^2 f$ such that, ignoring terms of degree greater than two in $\|x\|$,

$$f(a+x) = f(a) + \langle \nabla f(a), x\rangle + \frac{1}{2}\langle \nabla^2 f(a)x, x\rangle$$

The function ∇f is the *gradient* of f and $\nabla^2 f$ is its *Hessian*. Of course

$$(\nabla f(a))_i = \frac{\partial f}{\partial x_i}, \qquad (\nabla^2 f(a))_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

6.2.1 Lemma. *A twice-differentiable function on an open convex subset D of a real vector space V is convex if and only $\nabla^2 f$ is positive semidefinite at all points in D.*

Proof. We define the function g on $\mathbb{R}$ by

$$f((1-s)a + sx) = f(a + s(x-a))$$

and if we define $g(s) = f(a + s(x-a))$, then

$$g''(0) = \langle \nabla^2 f(a)(x-a), x-a\rangle$$

and g is convex at zero if and only if $\nabla^2 f(a) \succcurlyeq 0$. □

6.2.2 Theorem. *The function $f(X) = -\log(\det(X))$ is convex on the set of positive semidefinite matrices.*

Proof. Assume $X \succ 0$ and H is symmetric. For any t,

$$X + tH = X^{1/2}(I + tX^{-1/2}HX^{-1/2})X^{1/2}$$

and accordingly

$$\det(X + tY) = \det(X)\det(I + tX^{-1/2}HX^{-1/2}).$$

Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $X^{-1/2}HX^{-1/2}$; then

$$\det(I + X^{-1/2}HX^{-1/2}) = \prod_i (1 + \lambda_i).$$

Hence

$$\begin{aligned}\log(\det(X + tH)) &= \log(\det(X)) + \log(I + tX^{-1/2}HX^{-1/2}) \\ &= \log(\det(X)) + \log\left(\prod_i (1 + t\lambda_i)\right) \\ &= \log(\det(X)) + \sum_i \log(1 + t\lambda_i).\end{aligned}$$

We have

$$\log(1 + t\lambda)'' = -\frac{\lambda^2}{(1 + t\lambda)^2} \leq 0$$

and therefore $-\log(1 + t\lambda_i)$ is a convex function of t. Since a convex combination of convex functions is convex, we conclude that $-\log(\det(X))$ is convex. □

We compute the gradient of $\log(\det(X))$. We use the notation of the theorem and define

$$\hat{H} = X^{-1/2}HX^{-1/2}.$$

If we ignore quadratic and higher order terms in t,

$$\det(I + t\hat{H}) = 1 + t\,\mathrm{tr}(\hat{H}).$$

and it follows that

$$\langle \nabla f(X), H \rangle = -\,\mathrm{tr}(X^{-1/2}HX^{-1/2}) = -\langle X^{-1}, H \rangle.$$

It can also be shown that

$$\langle \nabla^2 f(X)H, H \rangle = -\,\mathrm{tr}(X^{-1}HX^{-1}H) = -\,\mathrm{tr}((X^{-1/2}HX^{-1/2})^2).$$

6.3 An Interior Point Method

We sketch an algorithm for solving semidefinite programs. The basic idea is to solve a sequence of problems of the form

$$\max\langle C, X \rangle + \mu \det(\log(X)), \quad \langle A_i, X \rangle = b_i \ (i = 1, \ldots, m), \quad X \succ 0.$$

(Note that X is constrained to be positive definite, not just positive semidefinite.) Here $\mu > 0$. We claim that for each positive value of μ, this problem has a unique optimum solution $X^*(\mu)$ and that, as $\mu \to 0$, the solutions $X^*(\mu)$ converge to a solution of our standard semidefinite program, i.e, the above problem with $\mu = 0$ and $X \succcurlyeq 0$.

The key point is that because of the *barrier function* $\mu \log \det(X)$, in solving this problems we can ignore the constraint $X \succ 0$. Thus the actual problem is to maximize a convex function over an affine space; this is a completely standard problem and can be solved by Newton's method. To apply Newton's method, we need the gradient and Hessian of the barrier function, but we have already determined these.

Chapter 7

Codes, Colourings, Packings

We present applications of semidefinite optimization to coding theory, and related topics.

7.1 Geometry of Code Words

We fix an integer n and define a word to be an element of $\mathbb{Z}_2^n$. The *Hamming distance* $\mathrm{h}(\alpha, \beta)$ between words α and β is defined by

$$\mathrm{h}(\alpha, \beta) = |\{i : \alpha_i \neq \beta_i\}|.$$

This is a metric on the set of all words. A code C is simply a set of words, we may refer to its elements as *code words*. If C is a subspace of $\mathbb{Z}_2^n$, it is a *linear code*.

The set of words at distance at most r from a code word x is the *ball of radius* r about x. The *packing radius* of a code C is the maximum integer e such that the balls of radius e about the code words in C are pairwise disjoint. The *covering radius* of C is the maximum distance r such that each binary vector of length n lies in the ball of radius r about some code word. The *minimum distance* of a code C is the minimum distance between two distinct words in C.

If $B_k(x)$ denote the ball of radius k about the word x, then

$$|B_k(x)| = \sum_{i=0}^{k} \binom{n}{i}$$

This leads us to the the so-called sphere-packing bound: if the packing radius of C is e, then

$$|C| \leq \frac{2^n}{\sum_{i=0}^{e}}\binom{n}{i}.$$

A code is *perfect* if we have equality in this bound. In general though, $\sum_{i=0}^{e}\binom{n}{i}$ is not a power of two, and so most codes are not perfect. The question we will be concerned with is to derive good upper bounds on the size of a code with packing radius e.

We can express this as a problem in graph theory. The *Hamming graph* $H(n,2)$ has vertex set $\mathbb{Z}_2^n$, where two words are adjacent if they differ in exactly one position, ie., their Hamming distance is 1. This is an n-regular graph on 2^n vertices, probably better known as the d-cube. If X denotes the Hamming graph, its *i-th distance graph* X_i is the graph with $V(X_i) = V(X)$, and with words α and β adjacent in X_i if and only if they are at distance i in X. (So $X_1 = X$.) If Y_e denote the edge-disjoint union of the graphs $X_1, \dots, X_e$, then our coding theory problem is equivalent to determining the maximum size of a coclique in Y_e.

7.2 A Matrix Algebra

We denote the adjacency matrix of the i-th distance graph X_i by A_i and set $A_0 = I$.

7.2.1 Lemma. *Let $A)i$ denote the adjacency matrix of the i-th distance graph of $H(n,2)$. There are constants b_i and c_i such that*

$$A_1 A_i = c_i A_{i-1} + b_i A_{i-1}. \qquad \square$$

7.2.2 Corollary. *There are polynomials $p_1, \dots, p_n$ such that* $\deg(p_i) = i$ *and $A_i = p_i(A_1)$.* □

These results have many consequences. Let $\mathcal{B}$ denote the vector space spanned by the matrices $A_0, \dots, A_n$. This space is closed under multiplication by A_1 and, consequently, it is closed under multiplication. A vector space of matrices that contains I and is closed under matrix multiplication is a *matrix algebra*. Since $A_i \circ A_j = 0$ if $i \neq j$, the set

$$\{0, A_0, \dots, A_n\}$$

is closed under Schur multiplication, and it follows that $\mathcal{B}$ is also Schur-closed. As $\sum_i A_0 = J$, it contains the Schur identity J. Finally

$$\operatorname{tr} A_i A_j = \operatorname{sum} A_i \circ A_j = \begin{cases} \operatorname{sum}(A_i), & i = j; \\ 0, & i \neq j. \end{cases}$$

This shows that $A_0, \dots, A_n$ is an orthogonal basis for $\mathcal{B}$.

There is a second useful orthogonal basis for $\mathcal{B}$. Assume $A = A_1$ and let

$$A = \sum_i \theta_i E_i$$

be the spectral decomposition of A. Then the spectral idempotents E_i span $\mathcal{B}$. As $\operatorname{tr}(E_i E_j) = 0$ if $i \neq j$, we see that these idempotents form an orthogonal basis for $\mathcal{B}$, and it follows that A has exactly $n+1$ distinct eigenvalues. As E_j represents orthogonal projection onto the θ_j eigenspace, $\operatorname{tr}(E_j)$ is equal to the dimension of this eigenspace or, equivalently, to the multiplicity of θ_j as an eigenvalue of A.

7.3 A Projection

7.3.1 Theorem. *Let M be a matrix with rows and columns indexed by $\mathbb{Z}_2^n$ and let $\widehat{M}$ denote the orthogonal projection of M onto the subspace $\mathcal{B}$ of* $\mathrm{Mat}_{2\times 2^n}(\mathbb{R})$. *Then*

$$\widehat{M} = \sum_{i=0}^{n} \frac{\langle A_i, M\rangle}{\langle A_i, A_i\rangle} A_i = \sum_{j=0}^{n} \frac{\langle E_j, M\rangle}{\langle E_j, E_j\rangle} E_j.$$

If $M \geq 0$, then $\widehat{M} \geq 0$; if $M \succcurlyeq 0$, then $\widehat{M} \succcurlyeq 0$.

Proof. The two expressions for $\widehat{M}$ are just the usual formula for the image of a projection onto a subspace, given an orthogonal basis for the subspace. For the final two claims note first that

$$\langle A_i, M\rangle = \operatorname{sum}(A_i \circ M)$$

and this is non-negative if M is non-negative. Next, the idempotents E_j are positive semidefinite, and so if $M \succcurlyeq 0$, then $\langle E_j, M\rangle \geq 0$. □

The coefficient

$$\frac{\langle E_j, M\rangle}{\langle E_j, E_j\rangle}$$

in the expansion of $\widehat{M}$ relative to the E_j's is the eigenvalue of $\widehat{M}$ on the θ_j-eigenspace of A_1.

We will develop formulas for the entries of the matrices E_j later, but we can determine one of these almost for free. Since $J \in \mathcal{B}$, there are scalars $\mu 0n$ such that

$$J = \sum_j \mu_j E_j.$$

As $J \succcurlyeq 0$, each eigenvalue μ_j is non-negative and accordingly

$$1 = \operatorname{rk}(J) = \sum_{j:\mu_j \neq 0} \operatorname{rk}(E_j).$$

It follows that exactly one of the eigenvalues μ_j is not zero, and therefore J is a scalar multiple of E_j for some j. It is traditional to assume $\mu_0 \neq 0$, and then $J = 2^n E_0$, or

$$E_0 = \frac{1}{2^n} J.$$

This implies that $E_j J = 0$ if $j \neq 0$.

7.3.2 Lemma. *Let M be a matrix with rows and columns indexed by $\mathbb{Z}_2^n$ and let $\widehat{M}$ denote the orthogonal projection of M onto the subspace $\mathcal{B}$ of* $\operatorname{Mat}_{2\times 2^n}(\mathbb{R})$. *Then* $\operatorname{tr}(\widehat{M}) = \operatorname{tr}(M)$ *and* $\operatorname{sum}(\widehat{M}) = \operatorname{sum}(M)$.

Proof. As $A_0 = I$ and $\sum_i A_i = J$, we have $\operatorname{tr}(A_j) = 0$ if $j \neq 0$. It follows immediately that

$$\operatorname{tr}(\widehat{M}) = \frac{\langle I, M\rangle}{\langle I, I\rangle} \operatorname{tr}(I) = \langle I, M\rangle = \operatorname{tr}(M).$$

If $j \neq 0$, then

$$\operatorname{sum}(E_j) = \operatorname{sum}(J \circ E_j) = n \operatorname{sum}(E_0 \circ E_j) = n \operatorname{tr}(E_0 E_j) = 0$$

and consequently

$$\operatorname{sum}(\widehat{M}) = \langle E_0, M\rangle \operatorname{sum}(E_0) = \operatorname{sum}(J \circ M) = \operatorname{sum}(M). \qquad \square$$

7.4 A Bound on Codes

Suppose C is a code with packing radius as least e, so the minimum distance between two distinct words of C is at least $2e+1$. We determine a linear program which optimum value is an upper bound on C.

Let x be the characteristic vector of C and set $M = xx^T$. From Theorem 7.3.1, we have

$$\widehat{M} = \sum_{i=0}^{n} \frac{\langle A_i, M\rangle}{\langle A_i, A_i\rangle} A_i.$$

Further

$$\langle A_i, M\rangle = \operatorname{sum}(A_i \circ M),$$

which shows that $\langle A_i, M\rangle$ is equal to the number of ordered pairs of elements (u,v) of C, such that $\operatorname{h}(u,v) = i$. We conclude that $\langle A_i, M\rangle = 0$ when $1 \le i \le 2e$.

From Lemma 7.3.2 we have

$$\operatorname{tr}(\widehat{M}) = \operatorname{tr}(M) = |C|, \qquad \operatorname{sum}(\widehat{M}) = \operatorname{sum}(M) = |C|^2$$

and therefore

$$|C| = \frac{\operatorname{sum}(\widehat{M})}{\operatorname{tr}(\widehat{M})}.$$

7.4.1 Lemma. *The maximum size of a code of length n and packing radius e is equal to the maximum value of $\operatorname{sum}(N)/\operatorname{tr}(N)$, where N runs over the positive semidefinite matrices in the algebra $\mathcal{B}$ such that $N \circ A_i = 0$ if $1 \le i \le 2i$.* □

Thus we see that we can compute an upper bound on the size of a code by solving a semidefinite optimization problem. In fact, we only need to deal with a linear program. To see this we first note that each matrix A_i is a linear combination of the spectral idempotents $E_0, \dots, E_n$ and therefore there are scalars $p_i(j)$ such that

$$A_i = \sum_j p_i(j) E_j.$$

The scalars $p_i(j)$ for $j = 0, \dots, n$ are eigenvalues of A_i. If $N \in \mathcal{B}$, we also have

$$N = \sum_i \nu_i A_i$$

and therefore

$$N = \sum_{i,j} \nu_i p_i(j) E_j = \sum_j \left(\sum_i \nu_i p_i(j) \right) E_j.$$

Here the coefficient of E_j is an eigenvalue of N, and therefore $N \succcurlyeq 0$ if and only if

$$\sum_i \nu_i p_i(j) \geq 0$$

for $j = 0, \ldots, n$.

Accordingly our upper bound is the value of the linear program

$$\max \sum_i \nu_i \operatorname{sum}(A_i)$$

subject to

$$\nu_0 = 2^{-n}, \quad \nu_i \geq 0, \ (i = 1, \ldots, n)$$

and

$$\sum_i \nu_i p_i(j) \geq 0.$$

These bounds were first derived by Delsarte in his 1973 Ph. D. thesis. The bounds were good and no major improvements were found until work of Schrijver in 2005. We discuss this in the following sections.

7.5 Schur-Closed Algebras

We discuss work of Schrijver's which lead to significant improvements in the upper bounds on the size of codes. The key step is to work with a superalgebra of the algebra $\mathcal{B}$. This larger algebra is not commutative, and so we have to work a little harder.

We introduce diagonal 01-matrices, with rows and columns index by $V(H(n,2))$. These are defined by the constraint that $(D_i)_{u,u} = 1$ if and only if $\mathrm{h}(0,u) = r$. Clearly $\sum_i D_i = I$. We define $\mathcal{T}$ to be the matrix algebra generated by the matrices $D_0, \ldots, D_n$ together with the matrices in $\mathcal{B}$. Thus

$$\mathcal{T} = \langle A_0, \ldots, A_n, \ D_0, \ldots, D_n \rangle.$$

7.5.1 Lemma. *The algebra $\mathcal{T}$ is closed under transposes.* □

We leave the proof of this lemma as an exercise.

We also need to use the groups of automorphism of the graph $H(n,2)$, which we denote by G. In fact we are only concerned with the subgroup G_0 of G, consisting of the elements of G that fix 0. We view automorphisms of $H(n,2)$ as permutation matrices that commute with $A_1 = A(H(n,2))$. It can be shown that $G_0 \cong \mathrm{Sym}(n)$, but we will not need this. We will make use of the following:

7.5.2 Lemma. *The algebra $\mathcal{T}$ is equal to the set of matrices that commute with each permutation matrix in G_0.* □

In other terms, $\mathcal{T}$ is the *commutant* of G_0. One direction of this is immediate—the matrices $A_0, \ldots, A_n$ and $D_0, \ldots, D_n$ lie in the commutant of G_0, and so the commutant contains $\mathcal{T}$—the more difficult step to prove the reverse inclusion.

7.5.3 Lemma. *The commutant of a set of permutation matrices is a Schur-closed matrix algebra that contains J.*

Proof. If A and B are matrices that commutes with permutation matrix P, you may easily verify that $A \circ B$ commutes with P. □

7.5.4 Corollary. *A Schur-closed matrix algebra that contains J has a basis of* 01*-matrices that sums to J (and hence is orthogonal).*

7.5.5 Corollary. *If M is a non-negative matrix, the orthogonal projection of M onto $\mathcal{T}$ is non-negative.* □

7.6 Properties of Projections

We aim to prove that orthogonal projection on to $\mathcal{T}$ sends positive semidefinite matrices to positive semidefinite matrices. It will be convenient to denote the orthogonal projection map by $\mathcal{E}$.

To begin we establish some important properties of $\mathcal{E}$. Note that we have a direct sum decomposition

$$\mathrm{Mat}_{n\times n}(\mathbb{R}) = \mathcal{T} \oplus \mathcal{T}^{\perp}.$$

7.6.1 Lemma. *For all matrices M we have $\mathrm{tr}(\mathcal{E}(M)) = \mathrm{tr}(M)$.*

Proof. As

$$\mathcal{E}(M - \mathcal{E}(M)) = \mathcal{E}(M) - \mathcal{E}(M) = 0,$$

we see that $M - \mathcal{E}(M) \in \mathcal{T}^\perp$. Consequently $M - \mathcal{E}(M)$ is orthogonal to I and hence

$$0 = \langle I, M - \mathcal{E}(M) \rangle = \operatorname{tr}(M - \mathcal{E}(M)).$$ □

7.6.2 Lemma. *For any matrix M and for any matrix N in $\mathcal{T}$, we have*

$$\mathcal{E}(MN) = \mathcal{E}(M)N, \qquad \mathcal{E}(NM) = N\mathcal{E}(M).$$

Proof. We see that

$$MN - \mathcal{E}(MN) \in \mathcal{T}^\perp, \quad M - \mathcal{E}(M) \in \mathcal{T}^\perp.$$

Since $\mathcal{T}$ is transpose-closed, it is N^T-invariant and hence $\mathcal{T}^\perp$ is N-invariant. It follows that

$$MN - \mathcal{E}(M)N = (M - \mathcal{E}(M))N \in \mathcal{T}^\perp$$

and therefore

$$\mathcal{E}(M)N - \mathcal{E}(MN) \in \mathcal{T}^\perp.$$

As $\mathcal{E}(M)N - \mathcal{E}(MN) \in \mathcal{T}$, it follows that $\mathcal{E}(M)N - \mathcal{E}(MN) = 0$.

The second claim follows similarly. □

7.7 Projections are Positive

We aim to rpove that the image $\mathcal{E}(M)$ in $\mathcal{T}$ of a positive semidefinite matrix M is positive semidefinite.

7.7.1 Lemma. *For all matrices M we have $\mathcal{E}(M^T) = \mathcal{E}(M)^T$.*

Proof. If $N \in \mathcal{T}$, then

$$\begin{aligned} \operatorname{tr}(\mathcal{E}(M^T)N) = \operatorname{tr}(\mathcal{E}(M^T N)) = \operatorname{tr}(M^T N) &= \operatorname{tr}(N^T M) \\ &= \operatorname{tr}(N^T \mathcal{E}(M)) \\ &= \operatorname{tr}(\mathcal{E}(M)^T N). \end{aligned}$$

Therefore

$$\operatorname{tr}((\mathcal{E}(M^T) - \mathcal{E}(M)^T)N) = 0$$

for all N in $\mathcal{T}$ and hence $\mathcal{E}(M^T) = \mathcal{E}(M)^T$. □

7.7.2 Theorem. *If $M \succcurlyeq 0$ then $\mathcal{E}(M) \succcurlyeq 0$.*

Proof. By the lemma, $N = \mathcal{E}(M)$ is symmetric and so using the spectral decomposition of N we have

$$N = E - F$$

where $E, F \succcurlyeq 0$ and $EF = FE = 0$. We assume $F \neq 0$ and derive a contradiction. If $F \neq 0$, there is a spectral idempotent F_1 of N associated to a negative eigenvalue λ of N and $FF_1 = \lambda F_1$. Now

$$0 > \operatorname{tr}(NF_1) = \operatorname{tr}(\mathcal{E}(M)F_1) = \operatorname{tr}(\mathcal{E}(MF_1)) = \operatorname{tr}(MF_1)$$

but M and F_1 are both positive semidefinite and therefore $\operatorname{tr}(MF_1) \geq 0$. □

The strategey for deriving an upper bound on the size of code now runs as follows. If x were the the characteristic vector of a code and $M = xx^T$, then $\mathcal{E}(M)$ is a non-negative and positive semidefinite matrix in $\mathcal{T}$. We can derive an upper bound by computing the maximum value of $\operatorname{sum}(N)/\operatorname{tr}(N)$ where N runs over the non-negative positive semidefinite matrices in $\mathcal{T}$.

7.8 Vector Colourings

Suppose $-1 \leq \alpha \leq 1$. The vertices of the graph $S(d, \alpha)$ are the unit vectors in $\mathbb{R}^d$, with unit vectors x and y adjacent if $\langle x, y \rangle \leq \alpha$. (In practice α will be negative.) A graph G has a *vector β-colouring* if there is a homomorphism

$$G \to S\left(d, -\frac{1}{\beta - 1}\right)$$

for some dimension d. We have $\beta = 1 - 1/\alpha$. The least value of β such that G has a vector β-colouring is the *vector chromatic number* of G, denoted $\chi_{\text{vec}}(G)$.

To take care of one trivial case, we note that $S(d, -1)$ is a disjoint union of copies of K_2, and so a vector 2-colourable graph is bipartite. We also note that if there is a homomorphism from H to G and G is vector β-colourable, so is H.

7.8.1 Lemma. *The complete graph K_n has a vector n-colouring.*

Proof. Let $e_1, \dots, e_n$ be the standard basis for $\mathbb{R}^n$ and define vectors $v_1, \dots, v_n$ by

$$v_i = \frac{1}{\sqrt{1 - \frac{1}{n}}} \left(e_i - \frac{1}{n} \mathbf{1} \right).$$

Then $\|v_i\| = 1$ and, if $i \neq j$, we find that

$$\langle v_i, v_j \rangle = -\frac{1}{n-1}$$

Therefore the map $i \mapsto v_i$ is a vector $(n-1)$-colouring of K_n. □

Not too surprisingly, $\chi_{\text{vec}}(K_n) = n - 1$; this is one consequwnce of the following.

7.8.2 Lemma. *We have* $\omega(G) \le \chi_{\text{vec}}(G)$.

Proof. Suppose C is a clique in G and the map $i \mapsto v_i$ is a vector β-colouring. Define

$$v_C = \sum_{i \in C} v_i.$$

Then

$$\begin{aligned} 0 \le \langle v_C, v_C \rangle &= |C| + \sum_{i \neq j} \langle v_i, v_j \rangle \\ &= |C| + (|C|^2 - |C|) \left(-\frac{1}{\beta - 1} \right) \\ &= |C| \left(1 - \frac{|C| - 1}{\beta - 1} \right) \end{aligned}$$

and consequently $\beta \ge |C|$. □

7.9 Semidefinite Programs for $\chi_{\text{vec}}(G)$

Assume G is a graph with adjacency matrix A. The value of the following problem is $\chi_{\text{vec}}(G)$:

$$\min \alpha$$

subject to

$$M \circ I = I, \quad M \circ A \le \alpha A, \quad M \succcurlyeq 0.$$

Any matrix that satisfies these constraints is the Gram matrix of a vector β-colouring, with $\beta = 1 - \frac{1}{\alpha}$. We assert that the dual to this problem can be written as

$$\min \operatorname{tr}(N)$$

given that

$$\operatorname{sum}(N) = 1, \quad N \circ (J - I - A) = 0, \quad N \geq 0, \quad N \succcurlyeq 0.$$

To see the relation between these problems, note that if M is primal feasible and N is dual feasible, then

$$\begin{aligned} 0 \leq \operatorname{tr}(MN) = \operatorname{sum}(M \circ N) &= \operatorname{tr}(D_N) + \operatorname{sum}(M \circ (N - (N \circ I))) \\ &\leq \operatorname{tr}(N) + \alpha(\operatorname{sum}(N) - \operatorname{tr}(N)) \\ &= \alpha + (1 - \alpha)\operatorname{tr}(N). \end{aligned}$$

It follows that

$$1 - \frac{1}{\alpha} \leq \frac{1}{\operatorname{tr}(N)}.$$

From this we conclude that if N is dual feasible, then $1/\operatorname{tr}(N)$ is an upper bound on the vector-chromatic number of G.

Referring back to Theorem 4.7.1, we have that $\theta(\overline{G})$ is equal to

$$\max \operatorname{sum}(N)$$

given that

$$\operatorname{tr}(N) = 1, \qquad N \circ (J - I - A) = 0, \qquad N \succcurlyeq 0.$$

You may prove that this is equal to

$$\max \frac{1}{\operatorname{tr}(N)}$$

subject to

$$\operatorname{sum}(N) = 1, \qquad N \circ (J - I - A) = 0, \qquad N \succcurlyeq 0.$$

An immediate consequence of this is that:

7.9.1 Corollary. *For any graph G we have $\chi_{\mathrm{vec}}(G) \leq \theta(\overline{G})$.* □

Let $S_=(d, \alpha)$ denote the graph with the unit vectors in $\mathbb{R}^d$ as its vertices, with two unit vectors u and v adjacent if and only if $\langle u, v \rangle = \alpha$. If we set $\beta = 1 - \frac{1}{\alpha}$, then we say a graph G has a *strict vector β-colouring* if there is a homomorphism from G to $S_=(d, \alpha)$.

The strict vector-chromatic number of G is equal to $\theta(\overline{G})$.

7.10 Bounds for $\chi_{\mathrm{vec}}(G)$

We work with the dual version of the optimization problem for $\chi_{\mathrm{vec}}(G)$ (actually $1/\chi_{\mathrm{vec}}(G)$):

$$\min \operatorname{tr}(N)$$

given that

$$\operatorname{sum}(N) = 1, \quad N \circ (J - I - A) = 0, \quad N \geq 0, \quad N \succcurlyeq 0.$$

Suppose $A = A(G)$ and τ is the least eigenvalue of A. We assume G has v vertices and e edges and that $e > 0$, whence $\tau < 0$. Therefore $A - \tau I$ is positive semidefinite and non-negative and

$$\frac{\operatorname{tr}(A - \tau I)}{\operatorname{sum}(A - \tau I)} = -\frac{v\tau}{2e - v\tau} = \frac{1}{1 - \frac{2e/v}{\tau}}.$$

This is an upper bound on the value of problem, and implies that

$$\chi_{\mathrm{vec}}(G) \geq 1 - \frac{2e/v}{\tau}.$$

We point out that $2e/v$ is the average valency of a vertex in G.

7.10.1 Lemma. *Let G be a k-regular graph with least eigenvalue τ and let E_τ denote the corresponding spectral idempotent. If there are constants x and y such that*

$$E_\tau \circ I = xI, \qquad E_\tau \circ A = yA$$

then $\chi_{\mathrm{vec}}(G) = 1 - \frac{k}{\tau}$.

Proof. If M and N are primal and dual optimal solutions, then $\operatorname{tr}(MN) = 0$ and therefore $MN = NM = 0$. Let us be optimistic and suppose that N is scalar multiple of $A - \tau I$. Then the columns of M lie in $\ker(A - \tau I)$ and therefore they are eigenvectors for A with eigenvalue τ. Accordingly M must be a positive semidefinite matrix whose columns are eigenvectors for A with eigenvalue τ. This suggests we should take M to be a scalar multiple of the spectral idempotent E_τ.

Assume $v = |V(G)|$ and let m denote the multiplicity of τ. Suppose that E_τ satisfies the hypotheses of the lemma. Then

$$vx = \operatorname{sum}(xI) = \operatorname{sum}(E_\tau \circ I) = \operatorname{tr}(E_\tau) = m$$

and, since $2e = vk$,

$$vky = \operatorname{sum}(yA) = \operatorname{sum}(E_\tau \circ A) = \operatorname{tr}(E_\tau A) = \tau \operatorname{tr}(E_\tau) = m\tau.$$

This shows that $x = m/v$ and $y = m\tau/(vk)$. We define

$$M = \frac{v}{m} E_\tau.$$

Then $M \circ I = I$ and $M \succcurlyeq 0$ and

$$M \circ A = \frac{v}{m} E_\tau \circ A = \frac{v}{m} \frac{m\tau}{vk} A = \frac{\tau}{k} A.$$

Therefore τ/k is an upper bound on the value α of the primal, and so G is vector β-colourable with

$$\beta = 1 - \frac{k}{\tau}.$$

We have shown that $\chi_{\text{vec}}(G) \leq 1 - \frac{k}{\tau}$. Comparing this with the upper bound derived previously, we conclude that equality holds. □

A careful reading of this proof will show that, in this case, $\chi_{\text{vec}}(G) = \theta(\overline{G})$, since the vector colouring constructed is strict.

An *arc* in a graph is an ordered pair of adjacent vertices, and a graph G is arc-transitive if its automorphism group acts transitively on the arcs of G. (An arc-transitive graph is necessarily vertex and edge transitive.) The hypotheses of the lemma hold for all arc-transitive graphs, and so:

7.10.2 Corollary. *If G is arc transitive with valency k and least eigenvalue τ, then $\chi_{\text{vec}}(G) = 1 - \frac{k}{\tau}$.* □

7.11 The Kissing Number

The *kissing number* τ_d is the maximum number of pairwise disjoint copies of the unit sphere in $\mathbb{R}^d$ that we can arrange so that so that each of the τ_d spheres touches some given sphere. (We regard spheres as disjoint if their interiors are disjoint.) It is not hard to convince yourself that $\tau_2 = 6$. Newton and Gregory had a famous dispute over whether τ_3 was 12 or 13. Newton argued for 12, whereas Gregory thought that 13 might be possible. Newton was proved right, eventually. We are going to discuss a proof that $\tau_8 = 240$.

We define a *spherical cap* with centre x to be the set

$$\{y : \|y\| = 1,\ x^T y \geq \cos\gamma\}.$$

(In general $\gamma \leq \pi/2$.) Suppose have N pairwise disjoint unit spheres touching the unit sphere centred at the origin. Then the intersections of these N spheres with the sphere S of radius two centred at the origin form a set of N pairwise disjoint caps on S. So the problem of determining the kissing number can be solved if we determine the maximum size of a set of pairwise disjoint caps on S (with the right size.) To determine the size of the caps, note that if three spheres of the same size are touching then their centres are the vertices of an equilateral triangle.

7.11.1 Lemma. *The kissing number τ_d is the maximum size of a set C of unit vectors such that $x^T y \leq 1/2$ for all distinct x and y in C.* □

We will shortly write down a semidefinite program which gives an upper bound on τ_d, but we need an additional concept. We will be working with symmetric bivariate functions on the unit sphere, such function may be referred to as a *kernel*. By way of (a very pertinent) example, if $p(t)$ is a real polynomial, then the function $p(x^T y)$ is a kernel. We say that a kernel is *positive semidefinite* if, for each set of unit vectors $x_1, \ldots, x_m$, the matrix

$$(K(x_r, x_s))_{r,s}$$

is positive semidefinite. A real function f defined on $[-1, 1]$ is positive semidefinite if the kernel $f(x^T y)$ is positive semidefinite; if f is positive semidefinite, we write $f \succcurlyeq 0$.

7.11.2 Lemma. *The value of the semidefinite program*

$$\inf \lambda, \quad p(1) = \lambda, \quad p(t) \leq -1 \text{ (if } t \leq 1/2), \quad p \succcurlyeq 0$$

is an upper bound on τ_d.

Proof. Suppose $C = \{x_1, \ldots, x_m\}$. Then, since $p \geq 0$, we have

$$0 \leq \sum_{r,s} p(x_r^T x_s) \leq |C|(\lambda - 1) + |C|(|C| - 1)(-1) = |C|(\lambda - |C|)$$

and so $|C| \leq \lambda$. □

7.12 Gegenbauer Polynomials

If f and g are continuous functions on the unit sphere Ω, then we have an inner product

$$\langle f, g\rangle = \int_\Omega f(z)g(z)\,dz.$$

A function f on the unit sphere is *zonal* relatiove to the unit vector a if $f(z)$ is determined by the value a^Tz. For zonal functions it can be shown that

$$\int_\Omega f(a^Tz)g(a^Tz)\,dz = \frac{1}{\gamma_d}\int_{-1}^{1} f(t)g(t)(1-t^2)^{(d-3)/2},$$

where γ_d is a scalar whose value will not play a role. We can view the right side as an inner product on the space of continuous functions on $[-1,1]$. Hence we can apply Gram-Schmidt to the sequence of polynomial

$$1, t, t^2, \dots$$

to form an orthogonal set of polynomials g_i, where $\deg(g_i) = i$. These are known as *Gegenbauer polynomials*. These polynomials are only determined up to multiplication by a nonzero scalar; we say they are *normalized* if $g_i(1) = 1$. (There could be a problem if $g_i(1) = 0$ for some i, you might show that this cannot happen.) We define functions $g_{a,i}$ on Ω by

$$g_{a,i}(z) = g_i(a^Tz).$$

One very important property of the functions $g_{a,i}$ is the *addition rule*:

7.12.1 Theorem. *For any two points a and b on the unit sphere,*

$$\langle g_{a,i}, g_{b,j}\rangle = \delta_{i,j}\, g_{a,i}(b). \qquad \square$$

We do not prove this. We note that $g_{a,i}(b) = g_i(a^Tb)$. Our chief application of this result is the following:

7.12.2 Lemma. *The kernel associated to the Gegenbauer polynomial g_i is positive semidefinite.*

Proof. By the addition rule

$$g_i(x^Ty) = \int_\Omega g_i(x^Tz)g_i(z^Ty)\,dz$$

and therefore

$$\begin{aligned}\sum_{r,s} a_r g_i(x^T y) a_s &= \sum_{r,s} \int_\Omega a_r g_i(x^T z) g_i(z^T y) a_s \, dz \\ &= \int_\Omega \Big(g_i(x_r^T z)\Big)^2 \\ &\geq 0.\end{aligned}$$

It follows that g_i is positive semidefinite. □

Any polynomial $f(t)$ can be exressed as a linear combination of Gegenbauer polynomials and, if

$$f(t) = f_0 g_0 + \cdots f_d g_d$$

then f is positve semidefinite if the Gegenbauer coefficients f_k are non-negative. (The converse is true, but we will not need it.) Therefore we can rewrite our semidefinite program bounding τ_d as follows:

$$\inf \lambda, \quad f_0 + \cdots + f_d = \lambda - 1, \quad \sum_r f_r g_r(t) \leq -1 \ (t \leq 1/2), \quad f_1, \ldots, f_d \geq 0.$$

There is one significant difficulty with this formulation, in that the constraint that $f(t) \leq -1$ when $t \in [-1, 1/2]$ is actually an infinite set of constraints. We will see how to deal with this later.

7.13 The Optimal Solution

We define

$$f(t) = -1 + \frac{320}{3}(t+1)\left(t + \frac{1}{2}\right)^2 t^2 \left(t - \frac{1}{2}\right).$$

You may show that $f(1) = 239$ and

$$f(-1_= f(-1/2) = f(0) f(1/2) = -1$$

We claim that $f(t) \leq -1$ if $t \in [-1, 1/2]$ (which can be verified by plotting software) and that the Gegenbauer coefficients of f are non-negative. In consequence

$$\tau_8 \leq 240.$$

Let C be the set consisting of the 112 vectors

$$\pm e_i \omega e_j, \quad (1 \leq i, j \leq 8)$$

along with the 128 vectors

$$(\pm 1)$$

of length eight and with even number of minus signs. In the exercises you are invited to use these vectors to prove that $\tau_8 \geq 240$ (these vectors can serve as the centres of the kissing spheres).

Chapter 8

Quantum Channels

8.1 Systems

A *quantum system* is a complex inner product space, and a state of the system is a 1-dimensional subspace. We could represent a state by a unit vector z that spans it, but then if a is complex number of norm 1, the unit vectors az and z determine the same state. In this context, physicists refer to a as a *phase factor*. The matrix $P = zz^T$ represents projection onto the line spanned by z, but does not depend on the choice unit vector. In our terms, each state is specified by positive semidefinite matrix with rank 1 and trace 1.

Physicists find it convenient to work with a more general class of objects. For them, a *density matrix* is a positive semidefinite matrix with trace 1 and density matrices specify so-called *mixed states*. In this context they refer to the states specified by density matrices with rank 1 as *pure states*. A mixed state can be viewed as a convex combination of pure states. To see this, recall that a positive semidefinite matrix P is Hermitian, hence has a spectral decomposition

$$P = \sum_r \theta_r E_r$$

where the idempotents E_r are positive semidefinite. If $P \succcurlyeq 0$, then $\theta_r \geq$ for all r and if $\operatorname{tr}(P) = 1$, then $\sum_r \theta_r = 1$. However in ??? we saw that any positive semidefinite matrix can be expressed as a sum of positive semidefinite matrices of rank 1, and these expressions are not unique.

If a quantum system is a complex inner product space, a *composite* quantum system is an inner product space which can be expressed as a tensor product of smaller systems. This simple fact is the cause of much of the weirdness of quantum physics.

8.2 Complex and Hermitian Matrices

If M is a complex matrix, M^* will denote its congugate-transpose. (Physicists normally use $M^\dagger$, which doesn't work so well on the blackboard.) We have an inner product

$$\langle M, N\rangle = \operatorname{tr}(M^* N) = \operatorname{sum}(\overline{M} \circ N)$$

for which

$$\langle N, M\rangle = \overline{\langle M, N\rangle}.$$

A matrix M is *Hermitian* if $M^* = M$. If M and N are Hermitian then $\langle M, N\rangle$ is real. A real matrix is Hermitian if and only if it is symmetric. If M is Hermitian then, in general, the matrix iM is not Hermitian, hence the Hermitian matrices do not form a vector space over $\mathbb{C}$. They do form a real vector space. The matrices

$$\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}, \quad \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}, \quad \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \quad \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}$$

form an orthogonal basis for the space of 2×2 Hermitian matrices, and it is not hard to prove that the space of $d \times d$ Hermitian matrices has dimension d^2.

A real matrix S is *skew symmetric* if $S^T = -S$, and a real matrix is skew-symmetric if and only if iS is Hermitian.

A complex matrix M is positive semidefinite if it Hermitian and

$$\langle z, Mz\rangle \geq 0$$

for all vectors z. (Note that $\langle z, Mz\rangle = z^* Mz$ and so is guaranteed to be real.) The 2×2 positive semidefinite matrices over $\mathbb{C}$ with trace 1 are of the form

$$\begin{pmatrix}\frac{1}{2} + a & c + di\\ c - di & \frac{1}{2} - a\end{pmatrix}$$

where a, c, d are real and

$$a^2 + c^2 + d^2 \leq \frac{1}{4}.$$

Thus there is a bijection from the set of 2×2 positive semidefinite matrices with trace 1 and the points of a sphere in $\mathbb{R}^3$. Physicists refer to the image of this bijection as the *Bloch sphere*. Note that the points on its surface correspond to pure states. Note also that the 2×2 positive semidefinite matrices with trace 1 are a subset of $\mathrm{Mat}_{2\times 2}(\mathbb{C})$, and this subset is not a sphere.

8.3 Measurement and Channels

Suppose the state space of a quantum system has dimension d. Then the result of a measurement of the system will take one of d values. For our purposes a measurement $\mathcal{M}$ of this system is a sequence of d matrices $M_1, \ldots, M_d$ such that $M_i \succcurlyeq 0$ and $\sum M_i = I$. The spectral idempotents of a Hermitian matrix H provide an example, but we do not require that the components M_i commute in general (and in particular, they are not orthogonal). If the state of the system is given by a density matrix D, then the result of a measurement modelled by $\mathcal{M}$ is i, with probability

$$\mathrm{tr}(DM_i).$$

We see that $\mathrm{tr}(DM_i) \geq 0$ because D and M_i are positive semidefinite, and also

$$\sum_i \mathrm{tr}(DM_i) = \mathrm{tr}\Big(D \sum_i M_i\Big) = \mathrm{tr}(D) = 1.$$

One key point is that the result of the measurement is a random variable on $\{1, \ldots, d\}$.

One way of describing general operations on quantum systems is by means of *channels*. Suppose we are given two quantum systems X and Y of respective dimensions d and e. A channel is a linear map $\mathcal{L}$ from $d \times d$ Hermitian matrices to $e \times e$ hermitian matrices such that

(a) $\mathrm{tr}(\mathcal{L}(M)) = \mathrm{tr}(M)$.

(b) If $M \succcurlyeq 0$, then $\mathcal{L}(M) \succcurlyeq 0$.

If (a) holds, we say that cŁ is *trace preserving*. Condition (b) is not all we need, the technical term is that $\mathcal{L}$ must be completely positive. We point out that $\mathcal{L}$ is necessarily of the form

$$\mathcal{L}(M) = \sum_i A_i M B_i^*$$

for suitable matrices A_i, B_i.

8.4 Hermitian SDPs

We consider optimization problems of the form

$$\sup\langle A, X\rangle, \qquad \Phi(X) = B, \quad X \succcurlyeq 0.$$

Here A and B are Hermitian matrices and Φ is a linear map from the space of $d \times d$ Hermitian matrices to the space of $e \times e$ Hermitian matrices. Note that if X is Hermitian and C is arbitrary, then

$$\langle C, X\rangle = \operatorname{tr}(C^* X) = \operatorname{tr}(X^* C) = \operatorname{tr}(XC) = \operatorname{tr}(CX) = \langle C^*, X\rangle$$

and therefore

$$\langle C, X\rangle = \frac{1}{2}(\langle C, X\rangle + \langle C^*, X\rangle).$$

Thus there is no loss of generality in assuming A is Hermitian. Previously the linear constraints for our semidefinite programs were of the form $\langle M_i, X\rangle = b_i$; we can convert this to the above form by taking B to be diagonal with $B_{i,i} = b_i$. (We then need to rewrite the left sides, but we leave this as an exercise.) The program above is in primal form, its dual is

$$\inf\langle Y, B\rangle, \qquad \Phi^*(Y) \succcurlyeq A, \qquad Y = Y^*.$$

Observe that

$$\langle A, X\rangle \le \langle \Phi^*(Y), X\rangle = \langle Y, \Phi(X)\rangle = \langle Y, B\rangle.$$

If X and Y are both optimal then

$$\langle A, X\rangle = \langle \Phi^*(Y), X\rangle$$

and so

$$\langle \Phi^*(Y) - A, X\rangle = 0.$$

Since $\Phi^*(Y) - A \succcurlyeq 0$ and $X \succcurlyeq 0$, this implies that

$$\Phi^*(Y)X = AX.$$

We also have (trivially)

$$Y\Phi(X) = YB.$$

8.5 Optimal Measurements

We consider an optimal measurement problem. A quantum system is prepared in a state with density matrix D_a with probability p_a, for $a \in \{1, \dots, d\}$. Supposed we have a measurement

$$\mathcal{M} = (M_1, \dots, M_d).$$

If we carry out this measurement on the system and it is in state i, we will conclude after our measurement that it is in state j with probability $\langle D_i, M\rangle_j$. The probability that we correctly identify the state is $\langle D_i, M_i\rangle$. Our problem is then to choose $\mathcal{M}$ so that

$$\sum_a p_a \langle D_a, M_a\rangle$$

is as large as possible. Here we produce a characterization of the optimal measurements.

8.5.1 Lemma. *A measurement $\mathcal{M} = (M_1, \dots, M_d)$ maximizes $\sum_i p_i \langle D_a i, M_i\rangle$ if and only if, for each j in $\{1, \dots, d\}$, we have*

$$\sum_i p_i M_i D_i \succcurlyeq p_j D_j.$$

Proof. We want to find

$$\max \sum_i p_i \langle M_i, D_i\rangle, \qquad \sum_i M_i = I, \qquad M_i \succcurlyeq 0.$$

The dual problem is

$$\min \operatorname{tr}(X), \qquad X \succcurlyeq p_i D_i \ (i \in \{1, \dots, d\}), \qquad X = X^*.$$

We have

$$\begin{aligned}
\operatorname{tr}(X) - \sum_i p_i \langle D_i, M_i\rangle &= \operatorname{tr}(X) - \sum_i \langle p_i D_i, M_i\rangle \\
&= \operatorname{tr}(X) - \sum_i \langle X, M_i\rangle + \sum_i \langle X, M_i\rangle - \sum_i \langle p_i D_i, M_i\rangle \\
&= \langle X, I - \sum_i M_i\rangle + \sum_i \langle X - p_i D_i, M_i\rangle \\
&= \sum_i \langle X - p_i D_i, M_i\rangle
\end{aligned}$$

and since $X - p_i D_i$ and M_i are positive semidefinite, the last term is nonnegative. Both the primal and dual have Slater points and it follows that if X is primal optimal,

$$\sum_i \langle X - p_i D_i, M_i \rangle = 0$$

and this holds if and only if

$$(X - p_i D_i) M_i = 0$$

for all i. Summing this over i yields

$$X = \sum_i p_a D_i M_i.$$

As X is Hermitian we also have $X = \sum_i p_a M_i D_i$. □

Chapter 9

Copositive, Completely Positive

9.1 Motkin-Straus

We start by presenting an influential result due to Motzkin and Straus.

9.1.1 Theorem. *For any graph G we have*

$$\frac{1}{\alpha(G)} = \min\{x^T(A+I)x : \mathbf{1}^T x = 1,\ x \geq 0\}.$$

Proof. Let $f(G)$ denote the value of this program. If S is a coclique with characteristic vector z and $x := |S|^{-1}z$ then x is feasible and

$$x^T(A+I)x = x^TAx + x^Tx = 0 = \frac{1}{|S|}.$$

Therefore $f(G) \leq \alpha(G)^{-1}$.

We proceed by induction on $n = |V(G)|$ to show that $f(G) \geq \alpha(G)^{-1}$. If $n = 1$, the result holds trivially. We assume inductively that $f(H) \geq \alpha(H)^{-1}$ for all proper induced subgraphs H of G. Let y be an optimal solution to the program. We distinguish two cases.

First, suppose there is a vertex i in G such that $y_i = 0$, and set $H = G\backslash i$. Then

$$\alpha(G)^{-1} \leq \alpha(H)^{-1} \leq f(H) \leq 2\sum_{ij\in E(H)} y_iy_j + \sum_{j\neq i} y_j^2 = f(G),$$

which proves the inequality we need.

So now we assume there is an optimum solution y such that all entries of y are positive. If $\mathbf{1}^T h = 0$, then

$$(y+h)^T(A+I)(y+h) = y^T(A+I)y + 2y^T(A+I)h + h^T(A+I)h$$

and, if h is small enough we may ignore the final term. Hence we deduce that if y is optimal, then $\mathbf{1}^T h = 0$ implies that $y^T(A+I)h = 0$. It follows that the subspace $\mathbf{1}^\perp$ is contained in the subspace $((A+I)y)^\perp$, and therefore for some scalar λ we have

$$(A+I)y = \lambda\mathbf{1}.$$

(You may reach the same conclusion using Lagrange multipliers, if you prefer.)

If G has no edges, $y = \lambda\mathbf{1}$ and since $\mathbf{1}^T y = 1$, we have $\lambda = 1/n$ and

$$y^T(A+I)y = \frac{1}{n}.$$

Hence $f(x) = 1/n$ as expected. Otherwise, if ij is an edge in G we set $h = \epsilon(e_i - e_j)$ and calculate

$$(y+h)^T(A+I)(y+h) = y^TAy + 2y^T(A+I)h + h^T(A+I)h.$$

Here $h^T(A+I)h = 0$ and, since $(A+I)y = \lambda\mathbf{1}$, we also see that $y^T(A+I)h = 0$. By choosing ϵ appropriately, we can arrange that $y+h \geq 0$ and some entry of $y+h$ is zero. The theorem now follows by induction. □

9.2 Copositive and Completely Positive Matrices

A matrix M is *copositive* if it is symmetric and $x^TMx \geq 0$ for all non-negative vectors x. Clearly positive semidefinite matrices are copositive, as are symmetric non-negative matrices. The set of copositive matrices is evidently a closed convex cone which contains both the cone of positive semidefinite matrices and the cone of symmetric non-negative matrices. We denote the cone of symmetric copositive matrices of order $n \times n$ by $\mathcal{C}_n$.

If we refer to a copositive matrix, you may assume we mean a *symmetric* copositive matrix.

A matrix is *completely positive* is it can be expressed as a sum of matrices

$$xx^T, \quad x \geq 0.$$

Note that if $xx^T \geq 0$, we may assume $x \geq 0$. It is immediate that the copositve matrices form a convex cone.

9.2.1 Lemma. *The completely positive matrices form a closed convex cone.*

Proof. The problem is to show that this cone is closed. The first step, which we leave as an exercise, is to show that there is an integer N such that any $n \times n$ completely positive matrix is a sum of at most N matrices of the form xx^T, where $x \geq 0$. Now prove that the set $\{xx^T : x \geq 0\}$ is closed, and that the Minkowski sum of a finite number of closed sets is closed. □

9.2.2 Lemma. *The cone of completely positive matrices is the dual of the cone of copositive matrices.*

Proof. If $M = \sum_i x_i x_i^T$ for non-negative vectors $x_1, \ldots, x_m$, then

$$\langle M, N \rangle = \operatorname{tr}\left(\sum_i x_i x_i^T N\right) = \sum_i \operatorname{tr}(x_i x_i^T N) = \sum_i x_i^T N x_i.$$

Hence if N is copositive, $\langle M, N \rangle \geq 0$ and so N lies in the dual to the cone of completely positive matrices.

If N is not copositive, there is a vector y such that $y \geq 0$ and $\langle N, yy^T \rangle < 0$ Therefore N does not lie in the dual to the cone of completely positive matrices.

If $\mathcal{P}$ denotes the cone of completely positive matrices, we have shown that $\mathcal{C}_n = \mathcal{P}^*$ and therefore

$$\mathcal{C}_n^* = (\mathcal{P}^*)^*.$$

Since $\mathcal{P}$ is closed, we have $(\mathcal{P}^*)^* = \mathcal{P}$. □

Because of this result, we can use $\mathcal{C}_n^*$ to denote the cone of completely positive matrices.

9.3 Motzkin-Straus as a Conic Program

We aim to show that the problem

$$\min\{x^T(A+I)x : x \geq 0,\ \mathbf{1}^T x = 1\}$$

(with value $\alpha(G)^{-1}$) is equivalent to

$$\min\{\langle A+I, X\rangle : \langle J, X\rangle = 1,\ X \in \mathcal{C}_n^*\}.$$

We present part of the argument as separate lemma.

9.3.1 Lemma. *A matrix Y is an extreme point of the set*

$$\{X \in \mathcal{C}_n^*,\ \langle J, X\rangle = 1\}$$

if and only if $Y = yy^T$, where $y \geq 0$ and $\mathbf{1}^T y = 1$.

Proof. In the exercises you will have the opportunity to prove that the extreme points of the feasible region of the conic program in the theorem are the matrices yy^T where $y \geq 0$ and $\mathbf{1}^T y = 1$.

Now suppose Y is an extreme point of M. Then for some k,

$$Y = \sum_{i=1}^k y_i y_i^T$$

where $y_i > 0$ for all i. Then $\mathbf{1}^T y_i > 0$ for all i and we may define matrices Z_i by

$$Z_i = (\mathbf{1}^T e_i)^{-2} y_i y_i^T.$$

Then Z_i is completely positive and $\langle J, Z_i\rangle = 1$, hence $Z_i \in \mathcal{M}$. Now

$$1 = \mathbf{1}^T Y \mathbf{1} = \sum_i (\mathbf{1}^T y_i)^2$$

and

$$Y = \sum_i (\mathbf{1}^T y_i)^2 Z_i,$$

we see that Y is a convex combination of matrices in M. Since it is an extreme point we must have $Y = Z_i$ for some i. □

9.3.2 Theorem. *For any graph G, we have*

$$\alpha(G)^{-1} = \min\{\langle A+I, X\rangle : \langle J, X\rangle = 1,\ X \in \mathcal{C}_n^*\}.$$

Proof. Note that

$$x^T(A+I)x = \langle A+I, xx^T\rangle$$

and so the difference btween the Motzkin-Straus problem and the one just stated is that in the Motzkin-Straus problem our feasible set is

$$\{xx^T : x \geq 0,\ \mathbf{1}^T x = 1\}$$

whereas in the conic program above, it is

$$\{X : \langle J, X\rangle = 1,\ X \in \mathcal{C}_n^*\}.$$

Since $\mathbf{1}^T x = 1$ if and only if $\langle J, xx^T\rangle$, we see that in the conic program we are minimizing $\langle A+I, X\rangle$ over the set $\mathcal{M}$ of completely positive matrices X with $\operatorname{sum}(X) = 1$, and in the Motzkin-Straus problem we are minimizing $\langle A+I, X\rangle$ over the extreme points $X = xx^T$ of $\mathcal{M}$. Since our objective function is convex on the set of non-negative vectors in $\mathbb{R}^n$, as you should prove, it assumes its minimum on an extreme point. Hence our result follows. □

The dual of the program in this theorem is

$$\max\{\lambda : I + A - \lambda J \in \mathcal{C}_n\}.$$

9.4 Copositive Programs for $\alpha(G)$

Recall that

$$\theta(G) = \max\{\langle J, X\rangle : A \circ X = 0,\ \operatorname{tr}(X) = 1,\ X \succcurlyeq 0\}.$$

If S is a coclique in G with characteristic vector x and we set $X = |S|^{-1}xx^T$, then X is feasible in the above program for $\theta(G)$, and its value is $|S|$. Hence $\alpha(G) \leq \theta(G)$, as we have seen before. By replacing the semidefinite cone by the copositive cone, we get the following:

9.4.1 Theorem. *If G is a graph on n vertices, then*

$$\alpha(G) = \max\{\langle J, X\rangle : A \circ X = 0,\ \operatorname{tr}(X) = 1,\ X \in \mathcal{C}_n^*\}.$$

Proof. The extreme rays of the convex cone

$$\{X \in \mathcal{C}_n^* : A \circ X = 0\}$$

are generated by the matrices xx^T with $x \geq 0$, and therefore the extreme points of the feasible region in our program above are the matrices xx^T with $x \geq 0$ and $\operatorname{tr}(xx^T) = 1$.

Since the optimum value must be attained at an extreme point, there is an optimal solution $Y = yy^T$ with $y \geq 0$ and $\|y\| = 1$. Since $A \circ Y = 0$, we see that $\operatorname{supp}(y)$ must be a coclique, S say. If we denote the optimum value by λ, then

$$\lambda = \max\{(\mathbf{1}^T x)^2 : \|x\| = 1,\ x \geq 0,\ \operatorname{supp}(x) = \operatorname{supp}(y)\}.$$

it is not hard to show that this maximum is realized when x is constant on its support, and hence that $\lambda = \alpha(G)$. □

Beacuse the matrices X in $\mathcal{C}_n^*$ are non-negative, $A \circ X = 0$ if and only if $\operatorname{tr}(AX) = 0$, and so we may rewrite the program in the theorem as

$$\max\{\langle J, X\rangle : \langle A, X\rangle = 0,\ \operatorname{tr}(X) = 1,\ X \in \mathcal{C}_n^*\},$$

and this dual of this is

$$\inf_{\lambda, y \in \mathbb{R}}\{\lambda I + yA - J \in \mathcal{C}_n\}.$$

The dual is strictly feasible—the matrix $(n+1)I - J$ is feasible, but the primal is not, because some entries of X must be zero. To deal with this, we will make use of the following lemma.

9.4.2 Lemma. *If $\epsilon \geq 0$, then the matrix $(1+\epsilon)\alpha(G)(I+A) - J$ is copositive.*

Proof. Let Q_ϵ denote the matrix $(1+\epsilon)\alpha(G)(I+A) - J$ and let Δ denote the simplex formed by the vectors x with $x \geq 0$ and $\mathbf{1}^T x = 1$. If $\mathbf{1}^T x = 1$, then

$$\begin{aligned} x^T Q_\epsilon x &= (1+\epsilon)\alpha(G)(x^T x + x^T A x) - x^T J x \\ &= (1+\epsilon)\alpha(G)(x^T x + x^T A x) - 1. \end{aligned}$$

We sketch the remaining steps. Choose x feasible. If $\operatorname{supp}(x)$ contains a pair of adjacenct vertices i and j, we can adjust the values of x_i and x_j to get a new feasible vector x with x_i or x_j zero. Thus we reduce to the case where $\operatorname{supp}(x)$ is a coclique, and then we show that x must constant on its support. □

Since the cone of copositive matrices is closed, we conclude that

$$Q_0 = \lambda(G)(I + A) - J$$

is copositive. As an immediate corollary, we have:

9.4.3 Corollary. *For any graph G,*

$$\alpha(G) = \min\{\lambda : \lambda(I + A) - J \in \mathcal{C}_n\}. \qquad \square$$

Index