CO 471 Semidefinite Optimization

Ilia Chtcherbakov

Spring 2016

0 Preamble

CO 471, Semidefinite Optimization. Offered by Professor Chris Godsil at the University of Waterloo in the Spring 2016 term. Composed, mlbaked by Ilia Chtcherbakov (`ischtche@uwaterloo.ca`).

Textbook: M. Laurent and F. Vallentin. *Semidefinite Optimization.* Mastermath notes, Spring 2012.

0.1 Introduction

2 May Semidefinite programming is a relatively simple but quite powerful generalization of linear programming. Recall that a linear program is a problem of the form

$$\max \ \mathbf{c}^\mathsf{T}\mathbf{x} \ \text{ s.t. } \ A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq 0 \ ,$$

where $\mathbf{c}^\mathsf{T}\mathbf{x}$ is an inner product and $A\mathbf{x} = \mathbf{b}$ is a matrix equation, which can be interpreted as a series of inner products $\mathbf{a}_i^\mathsf{T}\mathbf{x} = b_i$. We generalize this to a *semidefinite program* by instead taking

$$\max \ \langle C, X\rangle \ \text{ s.t. } \ \langle A_i, X\rangle = b_i,\ X \succcurlyeq 0 \ ,$$

where $\langle\cdot,\cdot\rangle$ is now a matrix inner product, A_i and C are matrices, and $X \succcurlyeq 0$ is the condition that X be a *positive semidefinite matrix*.

Semidefinite programs can be solved exactly, a property that is treasured among optimization problems, and there is a lot of curious theory behind precisely how much power they afford you. An awful lot more can be said about semidefinite optimization than for, say, convex or continuous optimization.

A healthy amount of linear algebra is required, and there are extensive applications in graph theory so familiarity with that subject is also assumed.

0.2 Preliminaries

Let V be an $\mathbb{R}$–vector space. A **bilinear form** $\beta : V^2 \to \mathbb{R}$ is a map which is linear in each coordinate—that is, for any $a \in V$, the maps $x \mapsto \beta(a, x)$ and $x \mapsto \beta(x, a)$ are linear. β is **symmetric** if $\beta(x, y) = \beta(y, x)$ for all $x, y \in V$. For example, if M is a square matrix on V, $\beta(x, y) = x^\mathsf{T}My$ is a bilinear form, and it is symmetric iff M is. β is **nondegenerate** if, whenever $\beta(a, x) = 0$ for all $x \in V$, then $a = 0$—in the above example, β is nondegenerate if M is invertible.

A bilinear form β is an **inner product** if it is symmetric, nondegenerate, and $\beta(x, x) \geq 0$ for all $x \in V$. We will denote inner products with angle brackets $\langle\cdot,\cdot\rangle$. By default, for vectors $x, y \in V$, we will use the dot product $\langle x, y\rangle := x^\mathsf{T}y$, which is an inner product.

Let $\mathbb{R}^{n\times n}$ is the space of $n \times n$ matrices, and recall the trace function $\operatorname{tr} : \mathbb{R}^{n\times n} \to \mathbb{R}$. First of all, $\operatorname{tr} A = \sum_i A_{ii}$, so it is linear. For matrices $A, B \in \mathbb{R}^{n\times n}$, we define $\langle A, B\rangle := \operatorname{tr} A^\mathsf{T}B$, and immediately we see it is a bilinear form. Also, $\operatorname{tr} A = \operatorname{tr} A^\mathsf{T}$, so $\langle\cdot,\cdot\rangle$ is symmetric.

Write $A = [\, a_1 \cdots a_n \,]$ in terms of its column vectors. Then $\langle A, A\rangle = \operatorname{tr} A^\mathsf{T}A = \sum_i (A^\mathsf{T}A)_{ii} = \sum_i \langle a_i, a_i\rangle \geq 0$, with equality iff each $a_i = 0$, that is, $A = 0$. So $\langle\cdot,\cdot\rangle$ is an inner product on $\mathbb{R}^{n\times n}$.

On final important property of the trace is that $\operatorname{tr} AB = \operatorname{tr} BA$ whenever both products are defined. In general, we may cyclically permute any product inside a trace, whenever the product is well-defined—for instance, $\operatorname{tr} ABC = \operatorname{tr} BCA$—but arbitrary permutations are not allowed:

$$\operatorname{tr}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix} = \operatorname{tr}\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix} = 1 \neq 2 = \operatorname{tr}\begin{bmatrix}2 & 0\\ 0 & 0\end{bmatrix} = \operatorname{tr}\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 1 & 0\end{bmatrix}.$$

A **norm** $\|\cdot\| : V \to \mathbb{R}$ is a map such that, for all $x \in V$: $\|x\| \geq 0$, with equality only if $x = 0$; $\|cx\| = |c|\|x\|$ for all $c \in \mathbb{R}$; and $\|x + y\| \leq \|x\| + \|y\|$ for all $y \in V$. If $\langle\cdot,\cdot\rangle$ is an inner product on V, then $\|x\| := \langle x, x\rangle^{\frac{1}{2}}$ is a norm. Some other useful norms are $\|x\|_1 := \sum_i |x_i|$ and $\|x\|_\infty := \max_i |x_i|$.

A matrix M is **positive semidefinite** or **psd** if it is symmetric and $x^\top M x \geq 0$ for all x. M is positive definite if equality holds only at $x = 0$, or equivalently if M is invertible. Write $M \succcurlyeq 0$ and $M \succ 0$ respectively. Also use $M \succcurlyeq N$ to mean $M - N \succcurlyeq 0$.

Proposition. *$\succcurlyeq$ is a partial order.*

4 May We are now ready to give an example of an SDP. Recall that a *coclique* in a graph $G = (V, E)$ is a pairwise nonadjacent subset of the vertices. Each subset $S \subseteq V$ may be represented by a characteristic vector $\chi \in \{0, 1\}^V$, where $\chi_v = 1$ if $v \in S$ and 0 otherwise. Note that $\chi\chi^\top \succcurlyeq 0$.

We will consider consider for each $S \subseteq V$ the matrix $X_S = \frac{1}{|S|}\chi\chi^\top$. Also, let $\mathbf{1}$ be the all-ones vector and $J = \mathbf{1}\mathbf{1}^\top$ the all-ones matrix. Notice that $\langle J, X_S\rangle = \operatorname{tr}\frac{1}{|S|}J\chi\chi^\top = \frac{1}{|S|}\operatorname{tr}\chi^\top J\chi = \frac{1}{|S|}|S|^2 = |S|$.

Let $\Omega = \{\, X \in \mathbb{R}^{n\times n} : \operatorname{tr} X = 1 \text{ and } X_{ij} = 0 \text{ if } ij \in E \,\}$. The X_S above are psd and are members of Ω. Hence, the value of the problem

$$\max\ \langle J, X\rangle \ \text{ s.t. }\ X \in \Omega,\ X \succcurlyeq 0$$

is an upper bound for $\alpha(G)$, the maximum size of a coclique in G. This is not the whole story, and there is plenty more that can be done, but as a first approximation this will do.

1 Positive Semidefinite Matrices

Recall that $M \succcurlyeq 0$ if $M = M^\top$ (or $M = M^*$ when working over $\mathbb{C}$) and $x^\top M x \geq 0$ for all x. We need to familiarize ourselves with psd matrices, because they have many nice properties we will be taking advantage of.

For example, observe that $A^\top A \succcurlyeq 0$, because $(A^\top A)^\top = A^\top A^{\top\top} = A^\top A$ and $x^\top A^\top A x = \|Ax\|^2 \geq 0$. Another example is, given vectors $x_1, ..., x_n$ in V an inner product space, their Gram matrix G is given by $G_{ij} = \langle x_i, x_j\rangle$. *Exercise.* Show that $G \succcurlyeq 0$.

Theorem. *Let A be symmetric. TFAE:* (a) $A \succcurlyeq 0$. (b) *The eigenvalues of A are nonnegative.* (c) $A = B^\top B$.

Proof. Suppose $A \succcurlyeq 0$ and $Ax = \lambda x$ for $x \neq 0$. Then $0 \leq x^\top A x = \lambda x^\top x = \lambda\|x\|^2$, so $\lambda \geq 0$, so (a) implies (b).

Now suppose every eigenvalue of A is nonnegative. A is real symmetric, so it has a spectral decomposition $A = L^\top D L$ for L orthogonal—that is, $L^\top L = I$—and D a diagonal matrix whose entries are the eigenvalues of A. So there is a well-defined square root diagonal matrix $D^{\frac{1}{2}}$ such that $D^{\frac{1}{2}}D^{\frac{1}{2}} = D$. Then

$$A = L^\top D L = L^\top D^{\frac{1}{2}}D^{\frac{1}{2}}L = (D^{\frac{1}{2}}L)^\top(D^{\frac{1}{2}}L),$$

so (b) implies (c). Finally, we have seen already that (c) implies (a). □

Corollary. *If $A \succcurlyeq 0$ then* $\det A \geq 0$. *Proof.* $\det A$ is the product of algebraic eigenvalues of A. □

Theorem. *A symmetric matrix A is psd iff $\langle A, X\rangle \geq 0$ for all psd matrices X.*

Proof. If $X = Y^\top Y$, then $\langle A, X\rangle = \operatorname{tr} AX = \operatorname{tr} AY^\top Y = \operatorname{tr} YAY^\top$, which is nonnegative if $A \succcurlyeq 0$. Conversely, if $\langle A, X\rangle \geq 0$ for all $X \succcurlyeq 0$, then setting $X = xx^\top$ gives $0 \leq \langle A, xx^\top\rangle = \operatorname{tr} Axx^\top = x^\top Ax$. □

As a short tangent, let us consider the 2×2 psd matrices. Any positive rescaling of a psd matrix is psd, so up to rescaling we may suppose our matrices have unit trace. So we may assume the matrix is of the form

$$A = \begin{bmatrix} \frac{1}{2} + a & b \\ b & \frac{1}{2} - a \end{bmatrix}$$

for which we must have that $-\frac{1}{2} \le a \le \frac{1}{2}$ and $0 \le \det A = (\frac{1}{2} + a)(\frac{1}{2} - a) - b^2 = \frac{1}{4} - a^2 - b^2$, i.e. $a^2 + b^2 \le \frac{1}{4}$. This is the equation for a closed disk of radius $\frac{1}{2}$, meaning the 2×2 psd matrices have a circular cross-section.

A matrix P is a **projection** if it is symmetric and idempotent, that is, $P = P^\mathsf{T} = P^2$. It represents orthogonal projection on the column space $\mathrm{Col}(P) = \{\, Px : x \in V \,\}$. Each of its eigenvalues is either 0 or 1, so $P \succcurlyeq 0$. $I - P$ is also a projection, because $(I-P)^2 = I - P - P + P^2 = I - 2P + P = I - P = (I-P)^\mathsf{T}$; it represents projection onto the orthogonal complement of $\mathrm{Col}(P)$.

If $\|x\| = 1$ then we can compute that $(xx^\mathsf{T})^2 = (xx^\mathsf{T})(xx^\mathsf{T}) = x(x^\mathsf{T}x)x^\mathsf{T} = xx^\mathsf{T}$, so xx^T is a projection, onto the line $\mathbb{R}x = \{rx : r \in \mathbb{R}\}$. If M is a symmetric rank-1 matrix, then $M = \pm xx^\mathsf{T}$ for some vector x.

Proposition (Cauchy-Schwarz). *If $A \succcurlyeq 0$ is a real matrix, then $(x^\mathsf{T}Ay)^2 \le (x^\mathsf{T}Ax)(y^\mathsf{T}Ay)$.*

Proof. Let t be an indeterminate, and consider $(x+ty)^\mathsf{T}A(x+ty) = x^\mathsf{T}Ax + 2(x^\mathsf{T}Ay)t + (y^\mathsf{T}Ay)t^2$. If A is psd, this is nonnegative. Since this is a quadratic in t, this means it has at most one root. Hence its discriminant $(x^\mathsf{T}Ay)^2 - (x^\mathsf{T}Ax)(y^\mathsf{T}Ay)$ is at most 0. □

Exercise. Prove the Cauchy–Schwarz inequality for complex vector spaces. *Hint.* You may need to fiddle with the phase θ of a complex number $re^{i\theta}$, if proceeding as above.

Theorem. *If $A \succcurlyeq 0$ has rank r, then A is the sum of at most r rank-1 psd matrices.*

Proof. We may assume A is nonzero, so choose x such that $Ax \ne 0$. Let $B = A - \frac{Axx^\mathsf{T}A}{x^\mathsf{T}Ax} = A\big(I - \frac{xx^\mathsf{T}A}{x^\mathsf{T}Ax}\big)$. $\mathrm{Col}(B) \le \mathrm{Col}(A)$ and $Bx = 0$, so B has rank at most $r - 1$. Then by Cauchy–Schwarz, we can compute that

$$y^\mathsf{T}By = y^\mathsf{T}Ay - \frac{y^\mathsf{T}Axx^\mathsf{T}Ay}{x^\mathsf{T}Ax} = \frac{(x^\mathsf{T}Ax)(y^\mathsf{T}Ay) - (x^\mathsf{T}Ay)^2}{x^\mathsf{T}Ax} \ge 0.$$

The result follows by induction on r. □

9 May 1.1 The Cholesky Decomposition

Recall that A is psd iff $A = BB^\mathsf{T}$ for some B. We showed one direction explicitly by using the eigendecomposition $A = LDL^{-1}$. Theoretically, this is fine, but in practice this is numerically unstable and little tough to compute.

Instead, there exists an algorithm due to Cholesky which determines rather quickly if A is psd, and if it is, produces a suitable B. We can even arrange for B to be lower triangular. In the interests of laziness, we will merely present a theoretical sketch from which the algorithm can be recovered.

Consider the matrix $M = [\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}]$ where A and D are square and A is invertible.

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

so $\det M = \det A \cdot \det(D - CA^{-1}B)$. The matrix $D - CA^{-1}B$ is called the *Schur complement* to A in M.

Theorem. *If $M = M^\mathsf{T} = [\begin{smallmatrix} A & B^\mathsf{T} \\ B & D \end{smallmatrix}]$ and A is invertible, then $M \succcurlyeq 0$ iff $A \succcurlyeq 0$ and $D - BA^{-1}B^\mathsf{T} \succcurlyeq 0$.*

Proof. $P = [\begin{smallmatrix} I & A^{-1}B^\mathsf{T} \\ 0 & I \end{smallmatrix}]$ is invertible, so by the above, $M = P^\mathsf{T} \left[\begin{smallmatrix} A & \\ & D-BAB^\mathsf{T} \end{smallmatrix}\right] P$ is psd iff $\left[\begin{smallmatrix} A & \\ & D-BAB^\mathsf{T} \end{smallmatrix}\right]$ is. □

Exercise. If $M \succcurlyeq 0$ and $x^\mathsf{T}Mx = 0$, then $Mx = 0$.

Now, observe that if some diagonal entry M_{ii} of a psd matrix M is zero. then $\mathbf{e}_i^\mathsf{T}M\mathbf{e}_i = M_{ii} = 0$, so $M\mathbf{e}_i = 0$, meaning that entire row (and by symmetry, column) is zero. So either we can find a strictly positive diagonal entry, or M is not psd, or M is zero. Take this entry to be A, so that it is invertible, and by induction we have an iterative algorithm for testing whether M is psd, and if so, constructing and N such that $M = NN^\mathsf{T}$.

1.2 Matrix products

Besides the usual matrix product, we shall introduce two more useful products of matrices. These behave well with respect to semidefiniteness, and are handy for expressing matrix manipulations.

If A and B are matrices, $m \times n$ and $p \times q$ respectively, then their **Kronecker product** is the $mp \times nq$ matrix $A \otimes B := [A_{ij}B]_{i,j}$. To wit, every entry of $A \otimes B$ is of the form $A_{ij}B_{i'j'}$. The Kronecker product is linear in each factor,[1] but generally not commutative. $(A \otimes B)^\mathsf{T} = A^\mathsf{T} \otimes B^\mathsf{T}$ and $\operatorname{tr} A \otimes B = \operatorname{tr} A \cdot \operatorname{tr} B$.

Exercise. $\det A \otimes B = \det A \cdot \det B$.

If the multiplications AC and BD are defined, then $(A \otimes B)(C \otimes D) = AC \otimes BD$. Consequently, if x is a λ-eigenvector for A and y is a μ-eigenvector for B, then $(A \otimes B)(x \otimes y) = Ax \otimes By = \lambda x \otimes \mu y = \lambda\mu(x \otimes y)$. So the eigenvalues of $A \otimes B$ are precisely the products of the eigenvalues of A and B.

If A is $m \times n$, then let $\operatorname{vec}(A) := \sum_{i,j} A_{ij}(\mathbf{e}_i \otimes \mathbf{e}_j) \in \mathbb{R}^{mn}$ be the listing of its columns from left to right. Then we have the useful equation $\operatorname{vec}(AXB^\mathsf{T}) = (B \otimes A)\operatorname{vec}(X)$. (*Exercise.*) So we may rewrite matrix equations $AX = XB$ as $(I \otimes A - B^\mathsf{T} \otimes I)\operatorname{vec}(X) = 0$. This is interesting because $(I \otimes A)(B \otimes I) = B \otimes A = (B \otimes I)(I \otimes A)$, so there are many matrices that commute, which is very useful from a linear algebraic perspective.

If A and B are both $m \times n$ matrices, then the **Schur product** of A and B is the $m \times n$ matrix $A \circ B$ such that $(A \circ B)_{ij} := A_{ij}B_{ij}$. This product is also linear in each variable, and commutative if the ring of coefficients is. If A and B are psd, then so is $A \circ B$. One way to see this is to note that it is a principal submatrix of $A \otimes B$, which is itself positive semidefinite. (*Exercise.*) Finally, note that $\operatorname{tr} A^\mathsf{T} B = \operatorname{sum}(A \circ B)$, where $\operatorname{sum}(X) := \sum_{i,j} X_{ij} = \mathbf{1}^\mathsf{T} X \mathbf{1}$.

2 Convex Geometry

We will now develop some basic theory of convexity and convex optimization, so that we may apply it and specialize it to our case of positive semidefinite programming. There must unfortunately be some minor quibbling about affine linear spaces and topology of convex sets, but shortly thereafter we will find a reasonably firm grasp on convex geometry and proceed with optimization.

A vector $y \in \mathbb{R}^d$ is an **affine linear combination** of vectors $x_1, ..., x_m \in \mathbb{R}^d$ if $y = \sum_i a_i x_i$ for coefficients $a_i \in \mathbb{R}$ satisfying $\sum_i a_i = 1$. Likewise, $x_1, ..., x_m$ are **affinely dependent** if there exist $a_i \in \mathbb{R}$ not all zero such that $\sum_i a_i x_i = 0$ and $\sum_i a_i = 0$.

11 May Predictably, a set of vectors is affinely independent if it is not affinely dependent, and the affine dimension of a set is one less than the size of a largest affinely independent subset. An affine subspace is a set of vectors which is closed under taking affine linear combinations. Verifying this is well-defined is routine and tedious.

Claim. *The vectors $x_0, ..., x_m$ are affinely independent iff $x_1 - x_0, ..., x_m - x_0$ are linearly independent.* □

Recall a bit of topology. Fixing a norm $\|\cdot\|$ on $\mathbb{R}^d$, the *open ball* of radius $r > 0$ about some $x \in \mathbb{R}^d$ is the set $B(r; x) = \{y \in \mathbb{R}^d : \|x - y\| < r\}$. *Open sets* are defined to be arbitrary unions of open balls, and these are closed under arbitrary union and finite intersection. *Closed sets*, on the other hand, are defined as complements of open sets. The *interior* of a set is the union of all the open balls it contains. The *boundary* of a closed set is the collection of points not in its interior. Finally, a set K is *compact* if it is closed and bounded—that is, contained in some open ball.

Note that usually we will want to consider the *relative interior* of a set S, with respect to the minimal affine space X containing S. Here, a *relatively open ball* in X of radius $r > 0$ about $x \in X$ is $B_X(r; x) = \{y \in X : \|x - y\| < r\}$. We will frequently consider sets which have a lower dimensionality than the ambient space, and these sets will always have empty interior, but frequently their relative interior will be nonempty and of great interest.

Theorem. *Let $S \subseteq \mathbb{R}^d$ and $\{x_0, ..., x_m\} \subseteq S$ a maximal affinely independent subset. Then the affine space generated by $\{x_0, ..., x_m\}$ has affine dimension m and contains S. Proof. Exercise.* □

[1] For those familiar with tensor products, the Kronecker product is precisely the tensor product of matrices. If we consider A and B, representing linear maps $V \to V'$ and $W \to W'$ respectively, then $A \otimes B : V \otimes W \to V' \otimes W'$ is such that $(A \otimes B)(v \otimes w) = Av \otimes Bw$.

An element y is a **convex combination** of $x_1, ..., x_m$ if $y = \sum_i a_i x_i$ where $\sum_i a_i = 1$ and each $a_i \geq 0$. A set is **convex** if it is closed under taking convex combinations. Equivalently, C is convex if for all $x, y \in C$, the line segment $[x, y] = \{tx + (1-t)y : 0 \leq t \leq 1\} \subseteq C$. A point x in a convex set C is **extreme** if it does not lie in the relative interior of any line segment in C.

The intersection of any family of convex sets is itelf convex. The **convex hull** of a set S is the smallest convex set containing S, i.e. the intersection of every convex superset of S. The convex hull of a finite set of points is called a (convex) **polytope**.

A (closed) **half-space** is a set of the form $H = \{x \in \mathbb{R}^d : \langle h, x \rangle \leq c\}$ for some $h \in \mathbb{R}^d$ and $c \in \mathbb{R}$. An **open half-space** is the interior of a closed half-space, that is, a set of the form $\{x \in \mathbb{R}^d : \langle h, x \rangle < c\}$. Both sort of half-space are convex. So the intersection of any collection of closed half-spaces is a closed convex set. The intersection of a finite number of half-spaces is a (convex) **polyhedron**.

Proposition. *A bounded polyhedron is a polytope.*

While we will not need the proposition above, and hence omit the proof, it is worth noting that it is nontrivial.

2.1 Closed convex sets

While convex sets are tame and unsurprising objects, it is not quite a simple matter to show they behave the way we expect, and it is important for us to make sure we prove things about convex sets in general, because the spaces in which they live do not always look like familiar Euclidean space. We focus our attention on proving two important results about convexity: that a closed convex set is an intersection of half-spaces, and that it is the convex hull of its extreme points.

Let C be a nonempty closed convex set, and let x be any point. We will show that there is a unique point $y \in C$ that is nearest to x. Existence follows because, if we fix any $c \in C \neq 0$, then the nearest point y is the minimizer of the continuous function $\|x - y\|$ on the compact set $C \cap \bar{B}(\|x - c\|; x)$ and this minimizer must be optimal since any nearest point y in C to x must be a member of C and $\|x - y\| \leq \|x - c\|$.

So it remains to see this point is unique. Suppose $y, z \in C$ are two distinct nearest points to x. Then we would like to say that $\frac{1}{2}y + \frac{1}{2}z \in C$ is strictly nearer, because that's how distance works. This reasoning is valid, so long as we make explicit that we are supposing that our norm $\|\cdot\|$ is *strictly convex*: that closed balls in this norm are strictly convex, i.e. no line segment is contained in the boundary. We will content ourselves with showing this for $\|\cdot\|$, which follows from considering when Cauchy–Schwarz holds at equality. (*Exercise.*)

16 May So the map π_C sending a point to its nearest neighbour in C is a well-defined function. This is often called the *metric projection* function.

Theorem. *For all points x and y, $\|\pi_C(x) - \pi_C(y)\| \leq \|x - y\|$.*

Before we prove this, note that since we are working with four points, we can mentally assume that we are working in affine dimension at most 3. Mental reductions like this are useful when visualizing problems in convex geometry, since convexity itself is in some sense a 1-dimensional property: given two points, we require only that the line segment they bound is contained in any convex set containing those two points. This should not ultimately factor into the proof we write, but is a useful heuristic when faced with something requiring proof.

Proof. If $\pi_C(x) = \pi_C(y)$, then the result is trivial. So we may assume they are distinct. Consider the hyperplanes H_x and H_y, both with normal $h = \pi_C(x) - \pi_C(y)$, intersecting $\pi_C(x)$ and $\pi_C(y)$ respectively. x must be on the same side of H_y as $\pi_C(x)$, or else $\pi_C(y)$ is nearer to x then $\pi_C(x)$.

So suppose for a contradiction that x is between H_x and H_y. Then the hyperplane H' with normal h and intersecting x is between H_x and H_y, so it intersects $z \in [\pi_C(x), \pi_C(y)]$. Since $\|\cdot\|$ is strictly convex, the only way for $\pi_C(x)$ to be nearer to x than z, is if $x = z$, so that $H' = H_x$.

So by symmetry, we have that each of x or y is in the correct closed half-space bounded by H_x and H_y, respectively. Altogether, $\|\pi_C(x) - \pi_C(y)\| \leq \|x - y\|$. □

Thus, π_C is a *contractive* map, and consequently it is continuous.

Exercise. Let $u \notin C$ and suppose x lies on the half-line from $\pi_C(u)$ through u. Show that $\pi_C(x) = \pi_C(u)$.

We are quickly approaching the proof that a closed convex set is the intersection of closed half-spaces.

Theorem. *For each y in the boundary ∂C of C, there exists $x \notin C$ such that $\pi_C(x) = y$.*

Proof (Sketch). Any closed ball of positive radius is closed and convex, and the intersection of C with a closed ball about y is convex, so it suffices to prove the result for compact convex sets. Consider some closed ball $B \supsetneq C$. Let $(y_n)_{n \geq 1} \subseteq B \setminus C$ be a sequence of points converging to y. Then $(\pi_C(y_n))_{n \geq 1} \subseteq \partial C$ converges to y as well.

So construct $(x_n)_{n \geq 1} \subseteq \partial B$ by taking x_n to be the point on the boundary of B that intersects the half-line from $\pi_C(y_n)$ through y_n. Passing to a convergent subsequence, we find $x_n \to x$ and $\pi_C(x_n) = \pi_C(y_n) \to y$, so by the continuity of projection, we are done. □