Constrained Optimization

Karl Stratos

All functions are assumed to be smooth in this note.

1 Setting

There are m **constraints** $c_1 \ldots c_m : \mathbb{R}^d \to \mathbb{R}$ whose indices are partitioned between $\mathcal{E}$ and $\mathcal{I}$, defining a closed **feasible set**

$$\Omega := \left\{ x \in \mathbb{R}^d : \;\; c_i(x) = 0 \;\; \forall i \in \mathcal{E}, \;\; c_i(x) \geq 0 \;\; \forall i \in \mathcal{I} \right\}$$

Given an objective function $f : \mathbb{R}^d \to \mathbb{R}$, our goal is to solve

$$\min_{x \in \Omega} \; f(x) \tag{1}$$

A point $x^* \in \mathbb{R}^d$ is a **local solution** to (1) if

1. $x^* \in \Omega$, and
2. There is a neighborhood $\mathcal{N}$ around x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{N} \cap \Omega$.

A useful special case of this setting is **convex programming** in which f and Ω are convex. In this case, all local solutions x^* are also global solutions (i.e., $f(x^*) \leq f(x)$ for all $x \in \Omega$) and themselves form a convex set.

Lemma 1.1. *If c_i is affine for all $i \in \mathcal{E}$ and concave for all $i \in \mathcal{I}$, then Ω is convex.*

Thus if f is convex, all inequality constraints are concave, and all equality constraints are affine, then to find a global solution it is sufficient to find any local solution.

2 Tangents

A vector $\delta \in \mathbb{R}^d$ is called a **tangent of Ω at** $x \in \Omega$ if there exists a *feasible* sequence $z_k \in \Omega$ approaching x and a *positive* sequence $t_k \in \mathbb{R}$ approaching 0 such that

$$\delta = \lim_{k \to \infty} \frac{z_k - x}{t_k}$$

The set of all tangents of Ω at $x^* \in \Omega$ is called the **tangent cone of Ω at** x^* and denoted by $T_\Omega(x^*)$. It is easy to check that $T_\Omega(x^*)$ is a cone rooted at 0.

$T_\Omega(x^*)$ characterizes feasible directions at $x^* \in \Omega$. To gain better insight, it is helpful to verify that if $\delta \in \mathbb{R}^d$ satisfies $x^* + \epsilon\delta \in \Omega$ for all sufficiently small $\epsilon > 0$, then $\delta \in T_\Omega(x^*)$.

Example 1. Let $\Omega = \left\{ x \in \mathbb{R}^2 : \; ||x|| = 1 \right\}$ and $x^* = (-1, 0)$. Then

$$T_\Omega(x^*) = \{(0, x_2) : \; x_2 \in \mathbb{R}\}$$

Example 2. Let $\Omega = \{x \in \mathbb{R}^2 : \ ||x|| \leq 1\}$ and $x^* = (-1, 0)$. Then

$$T_\Omega(x^*) = \{(x_1, x_2) : \ x_1 \geq 0, x_2 \in \mathbb{R}\}$$

See Appendix B for details of calculating tangents in the examples above. A critical result is that at local optimum, the gradient makes an accute angle with the tangent cone.

Lemma 2.1.

$$x^* \in \Omega \textit{ is a local solution of } (1) \qquad \Longrightarrow \qquad \langle \nabla f(x^*), \delta \rangle \geq 0 \quad \forall \delta \in T_\Omega(x^*)$$

Proof. Suppose $\langle \nabla f(x^*), \delta \rangle < 0$ for some $\delta \in T_\Omega(x^*)$. Then for some feasible sequence z_k approaching x^*, we eventually have

$$\langle \nabla f(x^*), z_k - x^* \rangle < 0$$

as $k \to \infty$. But this means $f(z_k) < f(x^*)$ for sufficiently large k. To see this, examine the Taylor approximation of $f(z_k)$ around x^*,

$$f(z_k) = f(x^*) + \underbrace{\langle \nabla f(x^*), z_k - x^* \rangle}_{<0} + \underbrace{o(||z_k - x^*||)}_{\to 0} \qquad \text{as } k \to \infty$$

Thus given any neighborhood $\mathcal{N}$ of x^*, we can produce $z_k \in \mathcal{N} \cap \Omega$ such that $f(z_k) < f(x^*)$ by choosing large enough k. This contradicts the assumption that x^* is a local solution. □

3 Linearized Feasible Directions

Tangents are defined geometrically and difficult to manipulate. Instead, we work with an algebraic approximation of feasible directions. Given $x^* \in \Omega$, define a set of **active constraints**

$$\mathcal{A}(x^*) := \mathcal{E} \cup \{i \in \mathcal{I} : \ c_i(x^*) = 0\}$$

and a set of **linearized feasible directions**

$$\mathcal{F}(x^*) := \{\delta \in \mathbb{R}^d : \ \langle \nabla c_i(x^*), \delta \rangle = 0 \quad \forall i \in \mathcal{E}, \quad \langle \nabla c_i(x^*), \delta \rangle \geq 0 \quad \forall i \in \mathcal{A}(x^*) \cap \mathcal{I}\}$$

Here is some motivation for the definition of $\mathcal{F}(x^*)$. Suppose the constraints are all *linear*. Then for any $x \in \Omega$ and $\delta \in \mathbb{R}^d$,

$$\begin{aligned} c_i(x + \delta) = 0 \qquad &\Longleftrightarrow \qquad \langle \nabla c_i(x), \delta \rangle = 0 \qquad \forall i \in \mathcal{E} \\ c_i(x + \delta) \geq 0 \qquad &\Longleftrightarrow \qquad \langle \nabla c_i(x), \delta \rangle \geq 0 \qquad \forall i \in \mathcal{I} \end{aligned}$$

So the set of feasible directions are fully characterized by vectors that form a right angle to $\nabla c_i(x)$ for $i \in \mathcal{E}$ and an acute angle to $\nabla c_i(x)$ for $i \in \mathcal{I}$ such that $c_i(x) = 0$. The reason we do not require an acute angle for $i \in \mathcal{I}$ such that $c_i(x) > 0$ is that

$$c_i(x + \epsilon\delta) = \underbrace{c_i(x)}_{>0} + \epsilon \langle \nabla c_i(x), \delta \rangle$$

remains positive for sufficiently small $\epsilon > 0$ no matter what δ is.

It is again easy to verify that $\mathcal{F}(x^*)$ is a cone rooted at 0. We will also use a matrix form. Let $B(x^*)$ be a matrix with $\{\nabla c_i(x^*)\}_{i \in \mathcal{A}(x^*) \cap \mathcal{I}}$ as columns. Let $C(x^*)$ be a matrix with $\{\nabla c_i(x^*)\}_{i \in \mathcal{E}}$ as columns. Then we can write

$$\mathcal{F}(x^*) = \left\{\delta \in \mathbb{R}^d : \ B(x^*)^\top \delta \geq 0, \quad C(x^*)^\top \delta = 0\right\}$$

Linearized feasible directions *overestimate* tangents. That is, a tangent is always a linearized feasible direction, but not every linearized feasible direction is a tangent. To ensure that the linearization is faithful, we require a non-degeneracy condition on the constraint gradients (Appendix A).

Lemma 3.1. *For all $x \in \Omega$, we have $\mathcal{T}_\Omega(x) \subseteq \mathcal{F}(x)$ with equality iff $\{\nabla c_i(x)\}_{i \in \mathcal{A}(x)}$ are linearly independent.*

A full proof of Lemma 3.1 is rather complicated; we refer to Nocedal and Wright (2006) for one.

4 KKT Conditions

We finally present the **KKT conditions**: necessary (but not sufficient) conditions for local optimium in (1). The conditions mandate that if $x^* \in \Omega$ is a local solution, then the gradient of the objective $\nabla f(x^*)$ is a particular conical combination of the constraint gradients $\nabla c_1(x^*) \dots \nabla c_m(x^*)$. The main ingredients behind the result are:

- Tangents yield a necessary geometric condition on $\nabla f(x^*)$ (Lemma 2.1).
- Linearized feasible directions can substitute tangents to yield a necessary algebraic condition on $\nabla f(x^*)$, provided that $\{\nabla c_i(x)\}_{i \in \mathcal{A}(x)}$ are linearly independent (Lemma 3.1).
- This algebraic condition can be used in Farkas' lemma to show that $\nabla f(x^*)$ lies in the cone generated by the active constraint gradients at x^*.

Theorem 4.1. *Suppose $x^* \in \Omega$ is a local solution of* (1) *and $\{\nabla c_i(x^*)\}_{i \in \mathcal{A}(x^*)}$ are linearly independent. Then there exists a unique $\lambda^* \in \mathbb{R}^m$ such that*

$$\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* \nabla c_i(x^*) \tag{2}$$

$$c_i(x^*) = 0 \qquad \forall i \in \mathcal{E} \tag{3}$$

$$c_i(x^*) \geq 0 \qquad \forall i \in \mathcal{I} \tag{4}$$

$$\lambda_i^* \geq 0 \qquad \forall i \in \mathcal{I} \tag{5}$$

$$\lambda_i^* c_i(x^*) = 0 \qquad \forall i = 1 \dots m \tag{6}$$

Proof. If $\delta \in \mathbb{R}^d$ satisfies $B(x^*)^\top \delta \geq 0$ and $C(x^*)^\top \delta = 0$, then we must have $\langle \nabla f(x^*), \delta \rangle \geq 0$ by Lemma 3.1 and 2.1 since $\delta \in \mathcal{F}(x^*)$. Thus there is no hyperplane separating $\nabla f(x^*)$ from the cone

$$K_{B(x^*), C(x^*)} = \left\{B(x^*)\lambda^{(1)} + C(x^*)\lambda^{(2)} : \ \ \lambda^{(1)} \geq 0, \ \lambda^{(2)} \in \mathbb{R}^{|\mathcal{E}|}\right\}$$

and consequently $\nabla f(x^*) \in K_{B(x^*),C(x^*)}$ by Farkas' lemma (Appendix D). The coefficients $\lambda^{(1)} \geq 0$ and $\lambda^{(2)} \in \mathbb{R}^{|\mathcal{E}|}$ associated with $\nabla f(x^*)$ are unique since the columns of $B(x^*)$ and $C(x^*)$ are linearly independent.

Now we set $\lambda^* \in \mathbb{R}^m$ so that λ_i^* is the corresponding coefficient in $\lambda^{(1)}$ or $\lambda^{(2)}$ if $i \in \mathcal{A}(x^*)$ and zero if $i \notin \mathcal{A}(x^*)$. It is easy to check the conditions (2–6) hold under this λ^*. The uniqueness follows by observing that $\lambda_i^* = 0$ is the only possible value for $i \notin \mathcal{A}(x^*)$ that does not break the condition (6).

□

The **Lagrangian** $L(x, \lambda_1 \dots \lambda_m)$ associated with (1) is defined as

$$L(x, \lambda_1 \dots \lambda_m) := f(x) - \sum_{i=1}^{m} \lambda_i c_i(x)$$

Note that we can re-write the KKT conditions as stationary conditions on the Lagrangian.

Theorem 4.2. *Suppose $x^* \in \Omega$ is a local solution of (1) and $\{\nabla c_i(x^*)\}_{i \in \mathcal{A}(x^*)}$ are linearly independent. Then there exists a unique $\lambda^* \in \mathbb{R}^m$ such that*

$$\nabla_x L(x^*, \lambda_1^* \dots \lambda_m^*) = 0 \tag{7}$$

$$\nabla_{\lambda_i} L(x^*, \lambda_1^* \dots \lambda_m^*) = 0 \qquad \forall i \in \mathcal{E} \tag{8}$$

$$\nabla_{\lambda_i} L(x^*, \lambda_1^* \dots \lambda_m^*) \leq 0 \qquad \forall i \in \mathcal{I} \tag{9}$$

$$\lambda_i^* \geq 0 \qquad \forall i \in \mathcal{I} \tag{10}$$

$$\lambda_i^* c_i(x^*) = 0 \qquad \forall i = 1 \dots m \tag{11}$$

The value λ_i^* in Theorem 4.1 and 4.2 is called a **Lagrange multiplier** and represents the "sensitivity" of the optimization problem to constraint c_i.

4.1 Sufficient Condition

For completeness we include the second-order sufficient optimality condition. This statement is a weaker version of Theorem 12.6 of Nocedal and Wright (2006). Note that the constraint qualification (i.e., that $\{\nabla c_i(x^*)\}_{i \in \mathcal{A}(x^*)}$ are linearly independent) is not required.

Theorem 4.3. *Suppose for $x^* \in \mathbb{R}^d$ we are able to find Lagrange multipliers $\lambda^* \in \mathbb{R}^m$ that satisfy the KKT conditions. If*

$$\nabla_{xx} L(x^*, \lambda_1^* \dots \lambda_m^*) \succ 0 \tag{12}$$

then x^* is a (strict) local solution of (1).

The following corollary is a useful shortcut in many convex programming problems.

Corollary 4.4 (Shortcut to convex programming)**.** *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is convex. Suppose $c_i : \mathbb{R}^d \to \mathbb{R}$ is concave for all $i \in \mathcal{I}$ and affine for all $i \in \mathcal{E}$. Suppose f is strictly convex or c_i is strictly concave for some $i \in \mathcal{I}$. Then x^* is the unique global solution of (1) iff it satisfies the KKT conditions.*

Proof. The Hesian of the Lagrangian is positive definite,

$$\nabla_{xx} L(x, \lambda_1 \dots \lambda_m) = \nabla_{xx} f(x) - \sum_{i=1}^{m} \lambda_i \nabla_{xx} c_i(x) \succ 0$$

so x^* is a strict local solution by Theorem 4.3 and therefore the unique global solution by Lemma 1.1. □

4.2 Duality

The KKT conditions can be seen as finding a *saddle point* of the Lagrangian: we minimize $L(x, \lambda_1 \dots \lambda_m)$ over x and maximize $L(x, \lambda_1 \dots \lambda_m)$ over $\lambda_1 \dots \lambda_m$. More specifically, the **dual objective function** $q : \mathbb{R}^m \to \mathbb{R}$ is defined as

$$q(\lambda_1 \dots \lambda_m) := \inf_{x \in \mathbb{R}^d} L(x, \lambda_1 \dots \lambda_m)$$

It can be shown that q is always concave. The **dual problem** corresponding to (1) is

$$\max_{\substack{\lambda_1 \dots \lambda_m \in \mathbb{R}: \\ \lambda_i \geq 0 \;\; \forall i \in \mathcal{I}}} q(\lambda_1 \dots \lambda_m) \tag{13}$$

The dual is a lower bound of (1): if $x \in \Omega$, then for all $\lambda_1 \dots \lambda_m$

$$q(\lambda_1 \dots \lambda_m) \leq f(x)$$

In certain cases, the inequality is tight and we can solve the original problem (1) by solving its dual (13). A full treatment of the theory of duality is beyond the scope of this note.

5 Examples

5.1 Norm Regularization vs. Norm Constraint

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and consider the l_2-regularized minimization with $\lambda > 0$:

$$x^* = \arg\min_{x \in \mathbb{R}^d} f(x) + \lambda \, ||x||^2 \tag{14}$$

Since this is strictly convex, x^* is unique and must satisfy

$$\nabla_x f(x^*) + 2\lambda x^* = 0$$

Now consider the constrained problem

$$x^+ = \arg\min_{x \in \mathbb{R}^d : ||x|| \leq ||x^*||} f(x) \tag{15}$$

The corresponding Lagrangian is

$$L(x, \lambda) = f(x) - \lambda(||x^*||^2 - ||x||^2)$$

By Corollary 4.4, $x^+ \in \mathbb{R}^d$ is the unique global minimum of (15) iff $||x^+|| = ||x^*||$ and for some $\lambda' > 0$

$$\nabla_x f(x^+) + 2\lambda' x^+ = 0$$

Thus $x^* = x^+$.

Reference Numerical Optimization (Nocedal and Write, 2006)

A Non-Degenerate Constraint Gradients

We give some additional results that shed light on why we need a non-degeneracy condition on the constraint gradients to ensure the faithfulness of linearization.

A.1 Single equality constraint.

Suppose we have a single equality constraint $c(x) = 0$. The feasible set is given by

$$\Omega = \left\{ x \in \mathbb{R}^d : \ c(x) = 0 \right\}$$

A fact from multivariable calculus states that at any $x \in \mathbb{R}^d$, the gradient $\nabla c(x)$ orthogonal to the level set at x,

$$\left\{ y \in \mathbb{R}^d : c(y) = c(x) \right\}$$

provided that $\nabla c(x) \neq 0$.

Lemma A.1. *At any point x, either $\nabla c(x) = 0$ or for all $\delta \in \mathbb{R}^d$,*

$$c(x+\delta) = c(x) \qquad \Longleftrightarrow \qquad \langle \nabla c(x), \delta \rangle = 0$$

A.2 Multiple equality constraints.

Suppose we have m equality constraints $c_1 \ldots c_m$. The feasible set is given by

$$\Omega = \left\{ x \in \mathbb{R}^d : \ c_1(x) = \cdots = c_m(x) = 0 \right\}$$

An extension of Lemma A.1 is the following.

Lemma A.2. *At any point x, either $\nabla c_1(x) \ldots \nabla c_m(x) \in \mathbb{R}^d$ are linearly dependent or for all $\delta \in \mathbb{R}^d$,*

$$c_i(x+\delta) = c_i(x) \ \ \forall i = 1 \ldots m \qquad \Longleftrightarrow \qquad \delta \perp \mathit{span}\{\nabla c_1(x) \ldots \nabla c_m(x)\}$$

Proof. ($\Rightarrow$) We have $\nabla c_i(x) \neq 0$ for each i, so we can apply Lemma A.1 and obtain $\langle \nabla c_i(x), \delta \rangle = 0$. ($\Leftarrow$) Since $\langle \sum_i \lambda_i \nabla c_i(x), \delta \rangle = 0$ for any choice of $\lambda_1 \ldots \lambda_m \in \mathbb{R}$, we set

$$\lambda_i = \begin{cases} 1 & \text{if } \langle \nabla c_i(x), \delta \rangle \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Thus we have $\sum_i |\langle \nabla c_i(x), \delta \rangle| = 0$. This means that $\langle \nabla c_i(x), \delta \rangle = 0$ and thus $c_i(x+\delta) = c_i(x)$ for each i again by Lemma A.1. $\square$

B Calculating Tangents

Example 1. Let $\Omega = \left\{ x \in \mathbb{R}^2 : \ ||x|| = 1 \right\}$ and $x^* = (-1, 0)$.

- One tangent is given by $(0, -1)$. Define for all $k \in \mathbb{N}$

$$z_k = \left(-\sqrt{1 - \frac{1}{k^2}},\ -\frac{1}{k}\right) \qquad\qquad t_k = ||z_k - x^*||$$

  Note that $z_k \in \Omega$ approaches $(-1, 0)$ from $(0, -1)$. It can be verified that $\lim_{k\to\infty}(z_k - x^*)/t_k = (0, -1)$.

- Another tangent $(0, 1)$ is given by flipping the sign of $[z_k]_2$ so that z_k approaches $(-1, 0)$ from $(0, 1)$.

- There cannot be a tangent δ such that $\delta_1 \neq 0$, since this requires $z_k \notin \Omega$ no matter how large k is. Thus the tangent cone is the vertical axis

$$T_\Omega(x^*) = \{(0, x_2) :\ x_2 \in \mathbb{R}\}$$

Example 2. Let $\Omega = \left\{x \in \mathbb{R}^2 :\ ||x|| \leq 1\right\}$ and $x^* = (-1, 0)$.

- We claim that any $x_1 \geq 0$ and $x_2 \in \mathbb{R}$ defines a tangent $(x_1, x_2)/\,||x||$. Define for all $k \geq (x_1^2 + x_2^2)/(2x_1)$

$$z_k = \left(\frac{x_1}{k} - 1,\ \frac{x_2}{k}\right) \qquad\qquad t_k = ||z_k - x||$$

  Note that $z_k \in \Omega$ approaches $(-1, 0)$ from the right side. Since $(z_k - x)/t_k = (x_1, x_2)/\,||x||$, the claim follows.

- There cannot be a tangent δ such that $\delta_1 < 0$, since this requires $z_k \notin \Omega$ no matter how large k is. Thus the tangent cone is the half plane

$$T_\Omega(x^*) = \{(x_1, x_2) :\ x_1 \geq 0, x_2 \in \mathbb{R}\}$$

C Asymptotic Notations

For $g(x) \neq 0$, we write $f(x) = o(g(x))$ as $x \to 0$ if

$$\lim_{x\to 0} \frac{f(x)}{g(x)} = 0$$

(i.e., g grows much faster than f). The asymptotic notation gives an insightful form of Taylor's theorem[1],

$$f(x) = f(y) + \langle \nabla f(y), x - y\rangle + o(||x - y||) \qquad\qquad \text{as } y \to x$$

which states that the error of the first-order Taylor approximation of $f(x)$ around y is dominated by $||x - y||$:

$$\lim_{y\to x} \frac{f(x) - (f(y) + \langle \nabla f(y), x - y\rangle)}{||x - y||} = 0$$

We also use of the notation to re-characterize the limit definition of a tangent δ of Ω at x^*. Let $z_k \in \mathbb{R}^d$ and $t_k > 0$ be the sequences associated with δ. Then

$$\begin{aligned} \lim_{k\to\infty} \frac{z_k - x^*}{t_k} = \delta \qquad &\Longleftrightarrow \qquad \frac{z_k - x^*}{t_k} - \delta = o(1) \qquad \text{as } k \to \infty \\ &\Longleftrightarrow \qquad z_k - t_k\delta - x^* = o(t_k) \qquad \text{as } k \to \infty \end{aligned}$$

which states that the error of $z_k - t_k\delta$ in approximating x^* is dominated by t_k.

[1] Theorem 3 at `http://eml.berkeley.edu/~anderson/Econ204/TaylorsTheoremTimeless.pdf`

D Convex Cones and Farkas' Lemma

Following Nocedal and Wright (2006), we use a formulation of a convex cone that accommodates unconstrained variables. Given any $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{n\times p}$, we define an associated cone

$$K_{B,C} := \{By + Ct : \; y \geq 0, \; t \in \mathbb{R}^p\}$$

This can be re-written in a more conventional form $K_D := \{Dz : \; z \geq 0\}$ where $D := [B \;\; C \;\; -C] \in \mathbb{R}^{n\times(m+2p)}$ and $z := y \oplus t_1 \oplus t_2 \in \mathbb{R}^{m+2p}$ since

$$K_{B,C} = \{By + Ct_1 - Ct_2 : \; y, t_1, t_2 \geq 0\}$$

Lemma D.1 (Farkas). *Pick any $g \in \mathbb{R}^n$. Then strictly one of the following is true:*

1. $g \in K_{B,C}$
2. *There exists $\delta \in \mathbb{R}^n$ such that $g^\top \delta < 0$, $B^\top \delta \geq 0$, and $C^\top \delta = 0$.*

Proof. Suppose $g \notin K_{B,C}$. Since $K_{B,C}$ is closed (Lemma D.2), we may define $\hat{s} \in K_{B,C}$ by

$$\hat{s} := \operatorname*{arg\,min}_{s \in K_{B,C}} ||s - g||_2^2 \tag{16}$$

Let $\delta := \hat{s} - g$. It has two geometric properties:

1. $\hat{s}^\top \delta = 0$: Since $\alpha \hat{s} \in K_{B,C}$ for $\alpha \geq 0$, $\alpha^* = 1$ is a local minimizer of $||\alpha \hat{s} - g||_2^2$. The claim follows from the gradient condition at α^*.
2. $s^\top \delta \geq 0$ for all $s \in K_{B,C}$: By the convexity of $K_{B,C}$, for $\theta \in (0, 1]$,

$$\begin{aligned} ||\hat{s} + \theta (s - \hat{s}) - g||_2^2 \geq ||\hat{s} - g||_2^2 \quad &\Rightarrow \quad \theta \, ||s - \hat{s}||_2^2 + 2\langle s - \hat{s}, \hat{s} - g\rangle \geq 0 \\ &\Rightarrow \quad \langle s - \hat{s}, \hat{s} - g\rangle \geq 0 \\ &\Rightarrow \quad \langle s, \hat{s} - g\rangle \geq 0 \end{aligned}$$

where the second step takes limit $\theta \downarrow 0$ and the third step uses $\hat{s}^\top \delta = 0$.

From the first property, we have $g^\top \delta = (\hat{s} - \delta)^\top \delta = -\,||\delta||_2^2 < 0$. From the second property, we have $y^\top B^\top \delta + t^\top C^\top \delta \geq 0$ for all $y \geq 0$ and $t \in \mathbb{R}^p$. Choose $y = 0$ to obtain $t^\top C^\top \delta \geq 0$ for all $t \in \mathbb{R}^p$: it follows that $C^\top \delta = 0$. Choose $t = 0$ to obtain $y^\top B^\top \delta \geq 0$ for all $y \geq 0$: it follows that $B^\top \delta \geq 0$.

Finally, it is easy to verify that we cannot have both $g \in K_{B,C}$ and $\delta \in \mathbb{R}^n$ such that $g^\top \delta < 0$, $B^\top \delta \geq 0$, and $C^\top \delta = 0$. □

Lemma D.2. *Let $\{s^k\}$ be any sequence inside $K_D := \{Dz : \; z \geq 0\}$ such that $s = \lim_{k\to\infty} s^k$. Then $s \in K_D$.*

Proof. For each k, we can find $I_k \subseteq [m]$ such that $s^k = D_{I_k} z^k$ for some $z^k > 0$ and D_{I_k} has linearly independent columns (Lemma D.3). Since $\mathcal{P}([m])$ is finite, there is an index set I that appears infinitely many times. Taking a subsequence corresponding to I and using the fact that every subsequnece converges to the same limit s, without loss of generality we may assume that $s^k = D_I z^k$ for all k. Taking limit $k \to \infty$ on both sides, we have $s = D_I z$ where $z := \lim_{k\to\infty} z^k \geq 0$. Since D_I has linearly independent columns, we have $z = D_I^+ s$, thus z is well-defined. The result $s \in K_D$ follows by observing that we can construct $z' \geq 0$ such that $s = Dz'$ by setting $z'_i = z_i$ if $i \in I$ and $z'_i = 0$ otherwise. $\square$

Lemma D.3. *Pick $y \in K_D := \{Dz : \ z \geq 0\}$. If $I \subseteq [m]$ is an index set of minimal size such that $y = D_I z$ for some $z > 0$, then D_I has linearly independent columns.*

Proof. Suppose false. Then $D_I w = 0$ for some $w \neq 0$, so $y = D_I(z + \tau w)$ for all $\tau \in \mathbb{R}$. We can choose τ so that $\bar{z} := z + \tau w$ has a 0 entry. The fact that $y = D_I \bar{z}$ contradicts the assumption that I is minimal. $\square$