Optimization: math toolkit

A collection of definitions, theorems, and formulae

March 1, 2026

Contents

1 Optimization fundamentals

1.1 General Identity for Pointwise Minimization

Given a functional of the form

$$\mathcal{E}(f) = \int_X L(f(x),\, x)\, \mathrm{d}x,$$

where

- $f\colon \mathcal{X} \to \mathbb{R}$ is the function to be optimized.
- $L(f(x), x)$ depends on $f(x)$ and on x, but not on derivatives or integrals of f.

Then the minimization of $\mathcal{E}(f)$ over all functions f reduces to pointwise minimization:

$$\min_f \mathcal{E}(f) \;=\; \int_{\mathcal{X}} \Big[\min_{f(x)} L(f(x), x)\Big]\, \mathrm{d}x.$$

Hence we can compute the minimizer $f(x)$ pointwise for all $x \in \mathcal{X}$.

1.2 The predictor $f_p(x) = \mathbb{E}[Y \mid x]$ minimizes the L^2 Loss.

Among all (measurable) functions $g(x)$, the conditional expectation $f_p(x) = \mathbb{E}[Y \mid x]$ uniquely minimizes the expected squared error:

$$\arg\min_g \; \mathbb{E}\big[(Y - g(x))^2\big] \;=\; \mathbb{E}[Y \mid x].$$

As a proof, for any candidate $g(x)$,

$$\mathbb{E}\big[(Y - g(x))^2 \mid x\big] = \mathbb{E}\big[(Y - f_p(x) + f_p(x) - g(x))^2 \mid x\big].$$

Expanding the square and using $\mathbb{E}[Y - f_p(x) \mid x] = 0$ gives

$$\mathbb{E}\big[(Y - g(x))^2 \mid x\big] = \underbrace{\mathbb{E}\big[(Y - f_p(x))^2 \mid x\big]}_{\text{irreducible error}} + \big(f_p(x) - g(x)\big)^2.$$

Since the second term is nonnegative and vanishes exactly when $g(x) = f_p(x)$, it follows that

$$\mathbb{E}\big[(Y - g(x))^2\big] \;=\; \mathbb{E}\Big[\mathbb{E}\big[(Y - g(x))^2 \mid x\big]\Big] \;\geq\; \mathbb{E}\Big[\mathbb{E}\big[(Y - f_p(x))^2 \mid x\big]\Big],$$

with equality if and only if $g(x) = \mathbb{E}[Y \mid x]$ almost surely. Thus $f_p(x) = \mathbb{E}[Y \mid x]$ is the L^2-optimal predictor.

1.3 Derivative of the Squared Norm

For a matrix $A \in \mathbb{R}^{n \times n}$ and vector $x \in \mathbb{R}^n$,

$$\frac{\partial}{\partial x} \|Ax\|_2^2 \;=\; 2\, A^\top A\, x.$$

Intermediate Steps.

(1) Expand the squared norm:

$$\|Ax\|_2^2 = (Ax)^\top (Ax) = x^\top A^\top A x.$$

(2) Differentiate using the fact that $\frac{\partial}{\partial x}(x^\top M x) = (M + M^\top)x$ and here $M = A^\top A$ is symmetric:

$$\frac{\partial}{\partial x}(x^\top A^\top A x) = 2\, A^\top A\, x.$$

1.4 Softmin Function

For $x \in \mathbb{R}^n$ and inverse-temperature parameter $\beta > 0$, the *softmin* is defined componentwise by

$$\text{softmin}(x)_i \;=\; \frac{e^{-\beta\, x_i}}{\sum_{j=1}^n e^{-\beta\, x_j}}, \qquad i = 1, \ldots, n.$$

In the limit as $\beta \to \infty$, the softmin concentrates all mass on the index i minimizing x_i, thereby providing a smooth approximation to the min operator.

1.5 Weierstraus theorem

Let $f : \Omega \to \mathbb{R}$ be a continuous function on a nonempty, closed, bounded set Ω. Then by the *Weierstraus theorem* there always exists a $x^* \in \Omega$ s.t. $f(x^*) \leq f(x) \quad \forall x \in \Omega$.

- If Ω is closed but not bounded, then use a ball and the fact that the intersection of closed sets is closed
- If Ω is bounded but not closed, consider the closure, show $f(x)$ continuous on the closure, use Weierstraß, then prove the minimum is not on the boundary.

1.6 Open, closed, clopen sets

- A set is *open* if every point has a neighborhood entirely contained within the set.
- A set is *closed* if its complement is open.
- It is also possible to be neither: $[0, 1)$ is neither *open* or *closed*

Therefore $\mathbb{R}$ is open, but $\mathbb{R}$ is also closed since its complement $\emptyset$ is open $\implies$ $\mathbb{R}$ is "*clopen*".

1.7 Unconstrained optimization over convex $\Omega = \mathbb{R}^n$

The condition that $\nabla f(x) = 0$ is *necessary,* but not sufficient, in unconstrained optimization ($\Omega = \mathbb{R}^n$) since the directional derivative at x along some directions $d \in \mathbb{R}^n$ must satisfy $\langle \nabla f(x), d \rangle \geq 0 \quad \forall d \in \mathbb{R}^n \iff \nabla f(x) = 0$.

This gives us a list of "candidate minimums" to investigate further

1.8 Extreme points

Let $C \subseteq \mathbb{R}^n$ be convex. A point $x \in C$ is an *extreme point* of C if it cannot be written as a nontrivial convex combination of two distinct points in C, i.e.,

$$x = \lambda y + (1 - \lambda) z, \quad y, z \in C,\ \lambda \in (0, 1) \implies y = z = x.$$

We write $\mathrm{ext}(C)$ for the set of extreme points of C (when C is a polytope, these are exactly its vertices).

1.9 Geometric balls

For $p \in [1, \infty]$, the ℓ_p unit ball in $\mathbb{R}^n$ is

$$B_p = \{x \in \mathbb{R}^n : \ \|x\|_p \leq 1\}.$$

A recurring theme in convex geometry/optimization is a tradeoff between (i) how many inequalities describe a convex set and (ii) how many extreme points it has.

1. **The ℓ_2-ball (Euclidean ball).**

   $$B_2 = \{x \in \mathbb{R}^n : \ \|x\|_2 \leq 1\}.$$

   In $\mathbb{R}^2$ the B_2 ball is a disk, and in higher dimensions it is a hypersphere. It is not a polytope, so it does not have a finite extreme-point description.

2. **The ℓ_∞-ball (hypercube).**

   $$B_\infty = [-1, 1]^n = \{x \in \mathbb{R}^n : \ -1 \leq x_i \leq 1 \text{ for all } i = 1, \ldots, n\}.$$

   This polytope can be described by $2n$ linear inequalities. However it has exponentially many extreme points:

   $$\mathrm{ext}(B_\infty) = \{(\pm 1, \ldots, \pm 1)\}, \qquad |\,\mathrm{ext}(B_\infty)| = 2^n.$$

3. **The ℓ_1-ball (crosspolytope).**

   $$B_1 = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\}.$$

   In $\mathbb{R}^2$ this is a diamond (a rotated square). As a polytope, it has only $2n$ extreme points:

   $$\mathrm{ext}(B_1) = \{\pm e_1, \ldots, \pm e_n\}, \qquad |\,\mathrm{ext}(B_1)| = 2n.$$

   But a direct linear-inequality description in the *original* variables x uses exponentially many constraints:

$$\|x\|_1 \le 1 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} s_i x_i \le 1 \ \text{ for all } s \in \{\pm 1\}^n,$$

i.e. 2^n inequalities (one per sign pattern).

Extended formulation (trade constraints for dimension). Introduce auxiliary variables $z \in \mathbb{R}^n$ to encode $|x_i|$:

$$\|x\|_1 \le 1 \quad \Longleftrightarrow \quad \exists z \in \mathbb{R}^n \text{ such that } \sum_{i=1}^{n} z_i \le 1, \quad z_i \ge 0, \quad -z_i \le x_i \le z_i \quad (i = 1, \ldots, n).$$

This replaces 2^n inequalities in x by $O(n)$ inequalities in (x, z), at the cost of adding n new variables (higher dimension).

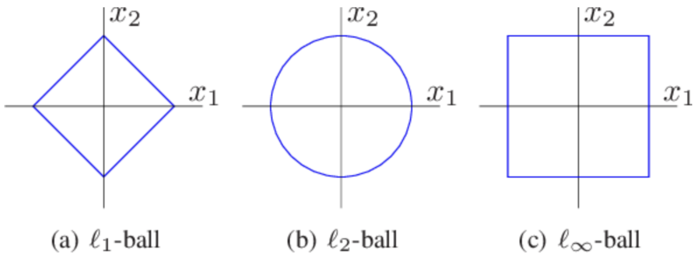

Figure 1: ℓ_p norms for $p = 1,\ 2,\ \infty$

2 Convexity

2.1 Convexity of a set

The set Ω is *convex* if

$$t\,x + (1-t)\,y \in \Omega \qquad \forall\, x, y \in \Omega, \quad t \in [0,1].$$

Note that convexity of Ω says nothing about being closed or bounded: For example,

- $\mathbb{R}$ is convex, not bounded, and clopen.
- $(0, 1)$ is convex, bounded, and open.
- $[0, 1)$ is convex, bounded, and neither open or closed in $\mathbb{R}$.

2.2 Convex functions

Let $\Omega \subseteq \mathbb{R}^n$ be convex. A function $f : \Omega \to \mathbb{R}$ is convex if

$$f(tx + (1-t)y) \ \le \ t\,f(x) + (1-t)\,f(y) \qquad \forall\, x, y \in \Omega,\ t \in [0,1].$$

The following definitions are equivalent:

1. $f(tx + (1-t)y) \le t\,f(x) + (1-t)\,f(y) \quad \forall\, x, y \in \Omega,\ t \in [0,1]$
2. $f(y) \ge f(x) + \langle \nabla f(x),\, y - x \rangle \quad \forall\, x, y \in \Omega$
3. $\langle \nabla f(y) - \nabla f(x),\, y - x \rangle \ge 0 \quad \forall\, x, y \in \Omega$

4. $\nabla^2 f(x) \succeq 0 \quad \forall\, x \in \Omega$ (if Ω is open)

2.3 Convex functions lies above their linearization

Let f be convex on Ω. Then

$$f(y) \;\geq\; f(x) + \langle \nabla f(x),\, y - x \rangle \qquad \forall\, x, y \in \Omega.$$

"f lies above its linearization at all points $x \in \Omega$"

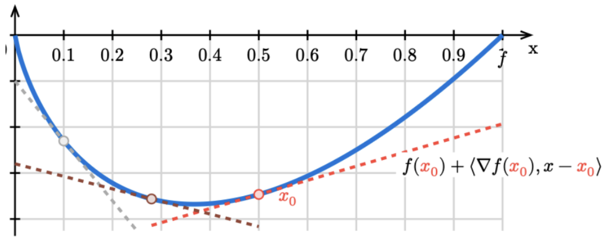

Figure 2: Convex functions lie entirely above their linearization at any point

2.4 Basic Convex Functions (Convex Atoms)

We can build f from known convex "atoms" using operations that preserve convexity (sums, nonnegative scaling, composition with affine maps, pointwise supremum, perspective, etc.). Below is a non-exhaustive list of standard convex functions (on their natural domains):

- Power functions: x^p for $p \geq 1$ or $p \leq 0$; and $-x^p$ for $0 \leq p \leq 1$.
- Exponential and logarithm: e^x, $-\log x$ (for $x > 0$), and $x \log x$ (for $x > 0$).
- Affine functions: $a^\top x + b$.
- Quadratics and norms:

$$x^\top x, \qquad \frac{x^\top x}{y}\ (y > 0), \qquad \sqrt{x^\top x}, \qquad \|x\|\ \text{(any norm)}.$$

- Pointwise operations: $\max(x_1, \ldots, x_n)$, $\log(e^{x_1} + \cdots + e^{x_n})$.
- Matrix functions: $\log\det(X^{-1})$ for $X \succ 0$.

Remark: Each of these is convex on its "natural" domain (e.g. $x > 0$ for $\log x$, $X \succ 0$ for $\log\det X^{-1}$)

2.5 Rules for preserving convexity

If f, g are convex functions on a common convex domain and h is convex and nondecreasing, then the following operations yield convex functions:

- *Nonnegative scaling:* If $\alpha \geq 0$, then αf is convex.
- *Sum:* $f + g$ is convex.

- *Affine pre-composition:* For any matrix A and vector b, $x \mapsto f(Ax+b)$ is convex.

- *Pointwise maximum:* If f_i $(i = 1, \dots, m)$ are convex, then $x \mapsto \max_i f_i(x)$ is convex.

- *Pointwise supremum:* If $f(x, y)$ is convex in x for each $y \in S$, then $g(x) = \sup_{y \in S} f(x, y)$ is convex.

- *Partial minimization (infimum):* If $f(x, y)$ is jointly convex in (x, y) and C is a convex set, then $g(x) = \inf_{y \in C} f(x, y)$ is convex.

- *Monotone composition:* If h is convex and nondecreasing and f is convex, then $g(x) = h(f(x))$ is convex.

2.6 Constructing Complex Convex Functions

Examples. By starting from basic convex atoms and applying the above rules, one can build more elaborate objectives and constraints:

- *Piecewise-linear (max-of-affines):*

$$f(x) = \max_{i=1,\dots,k} (a_i^\top x + b_i).$$

- ℓ_1-regularized least squares:

$$f(x) = \|Ax - b\|_2^2 \; + \; \lambda \, \|x\|_1, \quad \lambda \geq 0.$$

- *Distance to a convex set C:*

$$f(x) = \inf_{y \in C} \|x - y\|_2.$$

2.7 FOOC are *necessary and sufficient* with convex functions f over convex sets Ω

Let Ω be convex and f be convex & differentiable. Then

$$-\nabla f(x) \in \mathcal{N}_\Omega(x) \iff x \text{ is a minimizer of } f \text{ over } \Omega$$

- We know that over a convex set Ω, FOOC are necessary
- When we also have that $f(x)$ is convex too, then the FOOC become *necessary* ***and*** *sufficient*!

2.8 Strict and strong convexity

- f is *strictly convex* if

$$f(tx + (1-t)y) < t\, f(x) + (1-t)\, f(y) \qquad \forall\, x \neq y \in \Omega, \quad t \in (0, 1).$$

- f is *strongly convex* with modulus μ if

$$f(x) - \tfrac{\mu}{2}\|x\|_2^2 \quad \text{is convex.}$$

2.9 Strictly convex functions have at most 1 minimizer

Let Ω be convex, and f be strictly convex, then f has at most one minimizer over Ω.

2.10 Strong $\implies$ Strict $\implies$ Convex

Strongly convex functions must "beat" a quadratic term, so clearly since $\|x\|_2^2 > 0$ for all $x \neq 0$ then any strongly convex function will be strictly convex.

2.11 Projections of points onto convex sets are unique

Since $\frac{\mu}{2}\|x-y\|_2^2$ is strongly convex $\implies$ strictly convex, then projections of points onto convex sets are unique.

2.12 Strict convexity $\not\Rightarrow$ existence of a minimizer over a convex Ω

Recall that Weirstrauss, which applies for for any function f, only works when Ω is bounded and closed.

Thrm: If Ω not both closed and open, then f might not have a minimum, even if we add that Ω is convex *and* f is strictly convex.

Example: Let $C = (0, 1)$ (bounded, not closed, and convex) and $f(x) = x^2$ (strictly convex, continuous). Then $\inf_{x \in C} f(x) = 0$, but $0 \notin C$, so $\min_{x \in C} f$ does not exist.

2.13 Projections onto convex sets

- **(Thrm)** Let Ω be nonempty, closed, convex. If $y \notin \Omega$ then there exist $u \in \mathbb{R}^n$ and $v \in \mathbb{R}$ such that $\langle u, y\rangle < v$ and $\langle u, x\rangle \geq v \quad \forall x \in \Omega$.
- By construction u is perpendicular to the direction from y to x^* (its projection on Ω), $u = x^* - y, \quad v = \langle u, x^*\rangle$, where x^* is found by solving $\min_{x \in \Omega} \frac{1}{2}\|x - y\|_2^2$. This problem has $\nabla f(x) = x - y$, and the FOOC $\langle \nabla f(x), x - x^*\rangle \geq 0$ for all $x \in \Omega$ (since we know this is necessary).

3 Normal cones and

3.1 Normal cone of a convex set

Let $\Omega \subseteq \mathbb{R}^n$ be a convex set and $x \in \Omega$. The *normal cone* to Ω at x is

$$\mathcal{N}_\Omega(x) := \{ d \in \mathbb{R}^n : \langle d, y - x\rangle \leq 0 \quad \forall y \in \Omega \}.$$

The interpretation is that a vector $d \in \mathbb{R}^n$ lies in the normal cone at x if and only if it makes a non-acute angle (inner product ≤ 0) with every direction $y - x$ pointing into Ω.

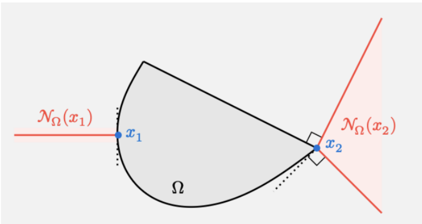

Figure 3: Normal cone to a convex set

3.2 Constrained optimization over convex $\Omega \subseteq \mathbb{R}^n$

Let Ω be convex and f differentiable. For a point $x \in \Omega$ to be a minimizer , it is necessary that $\langle \nabla f(x),\, y - x \rangle \geq 0 \quad \forall\, x, y \in \Omega$. Since the direction y-x keeps us in Ω when it's convex
(Key) When Ω is convex and f is differentiable, the first-order optimality condition is

$$-\nabla f(x) \;\in\; \mathcal{N}_\Omega(x).$$

(Example) Given a hyperplane

$$\Omega = \{\, y \in \mathbb{R}^n : \langle a, y \rangle = 0 \}, \quad \mathcal{N}_\Omega(x) := \mathrm{span}(a) = \{\lambda\, a : \lambda \in \mathbb{R}\},$$

to generalize, if

$$\Omega := \{\, y \in \mathbb{R}^n : A\, y = b \}, \quad \text{then} \quad \mathcal{N}_\Omega(x) := \mathrm{colspan}(A^T) = \mathrm{rowspan}(A).$$

(Key) Whenever we have $\min f(x)$ over $\Omega = \{x \in \mathbb{R}^n \mid Ax = b\}$ then at optimality it must hold that $-\nabla f(x) = A^T \lambda \quad$ for some $\lambda \in \mathbb{R}^d$. since a linear combination of the columns of A^T is the definition of $\mathrm{colspan}(A^T)$.

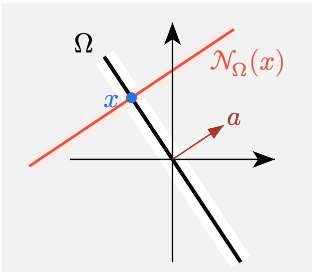

Figure 4: Normal cone to a hyperplane is the perpendicular plane (row span)

3.3 Normal cone at an intersection of halfspaces

Let $\Omega \subseteq \mathbb{R}^n$ be given as the intersection of m halfspaces $\langle a_j,\, x \rangle \leq b_j$. Then

$$\mathcal{N}_\Omega(x) := \Big\{ \sum_{j \in I(x)} \lambda_j\, a_j : \lambda_j \in \mathbb{R}_{\geq 0} \Big\}, \quad \text{for} \quad I(x) = \{\, j \in [m] : \langle a_j, x \rangle = b_j \},$$

where $I(x)$ is the set of active constraints at $x \in \Omega$.

We have a reformulation when $\Omega = \{\, x \in \mathbb{R}^n : A\, x \leq b \}$. Then

$$\mathcal{N}_\Omega(x) = \{ A^T \lambda : \lambda \geq 0,\ \lambda^T (b - A\, x) = 0 \}.$$

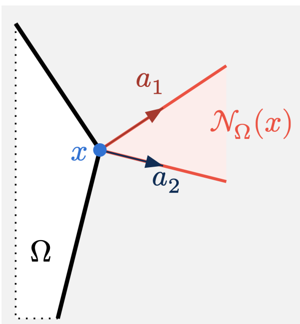

Figure 5: Normal cone at an intersection of halfspaces

3.4 Cones to be aware of

Conic optimization takes the form

$$\begin{aligned} \min \quad & f(x) \\ st. \quad & Ax = b \\ & x \in \mathcal{K} \quad \longleftarrow \text{ non-empty closed convex cone} \end{aligned}$$

- Lorentz cone → Second order conic programming

$$\mathcal{K} = \mathcal{L}^n := \{(x, z) \in \mathbb{R}^n \times \mathbb{R} : z \geq \|x\|_2\}.$$

- Semidefinite cone → Semidefinite programming

$$\mathcal{K} = \mathcal{S}^n := \{ X \in S^n : X \succeq 0\} = \{X \in S^n : a^T X a \geq 0 \quad \forall\, a \in \mathbb{R}^n\}.$$

  where S^n is the set of $n \times n$ symmetric matrices. (Semidefinite programming subsumes both LP's and SOCP's since both cases can be expressed as a particular subset of the semidefinite cone)

3.5 Normal cone at the intersection of an affine subspace and a closed convex set

(Thrm) Define an affine subspace as

$$H := \{ x \in \mathbb{R}^n : Ax = b\}$$

and let S be a closed convex set (not necessarily a cone) with $S^0 \cap H \neq \emptyset$, where S^0 is the interior of S. Then for any $x \in H \cap S$, the normal cone

$$\mathcal{N}_{H \cap S}(x) = \mathcal{N}_H(x) + \mathcal{N}_S(x).$$

Furthermore, since we know that $\mathcal{N}_H(x) = \text{colspan}(A^T)$, then $\mathcal{N}_{H\cap S}(x) = \{A^T \mu + z : \mu \in \mathbb{R}^m,\ z \in \mathcal{N}_S(x)\}$.

3.6 Polar cones, dual cones

(Def) The *polar cone* of $\mathcal{K}$, denoted $\mathcal{K}^\circ := \mathcal{N}_\mathcal{K}(0)$, is $\mathcal{K}^\circ = \{ d : \langle d, y\rangle \leq 0\, \forall\, y \in \mathcal{K}\}$.

(Def) The *dual cone* of $\mathcal{K}$, denoted $\mathcal{K}^* := -\mathcal{K}^\circ = -\mathcal{N}_\mathcal{K}(0)$, is $\mathcal{K}^* = \{ d : \langle d, y\rangle \geq 0\, \forall\, y \in \mathcal{K}\}$.

3.7 Self dual cones, normal cones at cones

(Thrm) The Lorentz, non-negative, semidefinite cones are self-dual: $\mathcal{K}^* = \mathcal{K}$.

(Thrm) Let $\mathcal{K}$ be a non-empty closed convex cone. Then $\mathcal{N}_\mathcal{K}(x) = \{ d \in \mathcal{K}^* : \langle d, x\rangle \geq 0\}$. where the proof is based on two separate inclusions, using that if $x \in \mathcal{K}$, then $\frac{1}{2}x, 2x \in \mathcal{K}$ from which we use $\langle d, x\rangle = 0$

3.8 Cone of non-negative polynomials

The cone of non-negative polynomials is the set of all polynomials $p \in \mathbb{R}[x]$ such that $p(x) \geq 0 \quad \forall\, x \in \mathbb{R}^n$, denoted $\mathbb{R}[x]_0$.

(Def) The cone of nonnegative polynomials (in one variable) of degree $2d$ is

$$\mathbb{R}[x]_{1,2d} = \left\{ (p_0, \ldots, p_{2d}) \in \mathbb{R}^{2d} : \sum_{k=0}^{2d} p_k x^k \geq 0 \quad \forall x \in \mathbb{R} \right\}.$$

The generalization of this to n variables of maximum degree $2d$ is $\mathbb{R}[x]_{n,2d}$

4 Duality

4.1 Duality

Consider the primal for $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$:

$$\begin{aligned} \max_{x\in\mathbb{R}^n} \quad & f(x) = c^\top x \\ \text{s.t.} \quad & A\,x \leq b \end{aligned} \tag{P}$$

And consider the dual for $\lambda \in \mathbb{R}^m_{\geq 0}$:

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^m_{\geq 0}} \quad & g(\lambda) = b^\top \lambda \\ \text{s.t.} \quad & A^\top \lambda = c \end{aligned} \tag{D}$$

Strong duality: If (P) admits an optimal x^*, then (D) admits an optimal λ^* such that:

- $c^T x^* = b^T \lambda^*$
- Additionally, λ^* satisfies complementary slackness: $(\lambda^*)^T\,(b - A\,x^*) = 0$.

4.2 Definition of a cone

(Def) A set S is a *cone* if for any $x \in S$ and $\lambda \in \mathbb{R}_{\geq 0}$, we also have $\lambda x \in S$.

(Thrm) If Ω is a nonempty, closed, convex *cone* and $y \notin \Omega$, then there exists $u \in \mathbb{R}^n$ such that $\langle u, y\rangle < 0 \quad \text{and} \quad \langle u, x\rangle \geq 0 \quad \forall\, x \in \Omega$. → e.g. we just looked at $f(x) = \frac{1}{2}\|x - y\|^2$ and set $v = 0$. This is the theorem above with $v = 0$.

4.3 Separation oracle

(Def) A *strong separation oracle* for a closed convex set $\Omega \subseteq \mathbb{R}^n$ is an algorithm which, given any test point $y \in \mathbb{R}^n$, correctly outputs exactly one of:

1. $y \in \Omega$,
2. $(y \notin \Omega,\, u)$ where $u \in \mathbb{R}^n$ satisfies $\langle u, y\rangle < \langle u, x\rangle$ for all $x \in \Omega$.

→ If $\Omega = \{\, x \in \mathbb{R}^n : A\,x \leq b\}$, where $A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}$, then if $y \notin \Omega$ there is some j with $a_j^T y > b_j$, so $-a_j$ defines a valid separating hyperplane.

4.4 Farkas' Lemma

For any A, b in $Ax \leq b$, exactly one of the following holds:

- $Ax \leq b$ has a solution,
- there exists a vector $y \geq 0$ such that $A^T y = 0 \quad \text{and} \quad b^T y < 0$.

4.5 KKT conditions for functional constraints

Motivation: $-\nabla f(x)$ should belong to the normal cone to the linearization of the binding constraints at x

- A functional constraint problem takes the form of

$$\begin{aligned} \min \quad & f(x) \\ st. \quad & h_i(x) = 0 \quad \text{for } i \in [r] \\ & g_j(x) \leq 0 \quad \text{for } j \in [s] \text{non-empty closed convex cone} \end{aligned}$$

(Def) The KKT conditions at x of an optimization problem with functional constraints of the form above are: "There exist μ_i and λ_j s.t. the following holds"

$$-\nabla f(x) = \sum_{i=1}^{r} \mu_i \nabla h_i(x) \ + \ \sum_{j=1}^{s} \lambda_j \nabla g_j(x) \quad \text{"Stationarity"}$$

$$\mu_i \in \mathbb{R} \quad \forall i \in [r], \quad \lambda_j \geq 0 \quad \forall j \in [s] \quad \text{"Dual feasibility"}$$

$$\lambda_j g_j(x) = 0 \quad \forall j \in [s] \quad \text{"Complementary slackness"}$$

- However, the KKT conditions might not hold at x (aka there might not exist μ_i, λ_j) even when x is optimal. We
- For example, KKT fails at x when we have collapsing linearizations
- To describe the instances of functional constraints where the KKT conditions *do* hold, we use *constraint qualification*

4.6 Optimality qualifications using the KKT conditions

The following give conditions *at a given* x which are either necessary or necessary and sufficient for x to to optimal:

- KKT are necessary for optimality at a given x when the binding inequalities $\{g_j\}_{j \in I(x)}$ are concave differentiable and the equality constraints are affine.
- KKT are necessary for optimality at a given x if the set of gradients $\{\nabla h_i(x) : i \in [r]\} \cup \{\nabla g_j(x) : j \in I(x)\}$ is linearly independent.
- **(Slater's Condition)** The binding inequality constraints $\{g_j\}_{j \in I(x)}$ are convex differentiable, the equality constraints $\{h_i\}_{i \in [r]}$ are affine, and there exists a strictly feasible point x_0 for the binding inequalities, i.e.

  $$g_j(x_0) < 0 \quad \forall j \in I(x).$$

  $\rightarrow$ Then if $f(x)$ is convex and $\{h_i\}$ & $\{g_j\}$ satisfy Slater's conditions, then the KKT conditions at x are *both* necessary & sufficient for x to be optimum!

4.7 Relationship between optimality and the Lagrangian function

When Slater's conditions hold, the KKT conditions are necessary and sufficient for x to be optimal.

When we are in this setting, the following are equivilant (imply eachother)

1. A) the point x^* is optimal

2. B) the point $x^* \in \Omega$ admits $\mu \in \mathbb{R}^r$ and $\lambda \in \mathbb{R}^s_{\geq 0}$ such that

$$-\nabla f(x^*) = \sum_{i=1}^{r} \mu_i \nabla h_i(x^*) \; + \; \sum_{j=1}^{s} \lambda_j \nabla g_j(x^*), \qquad \lambda_j \, g_j(x^*) = 0 \;\; \forall \, j \in [s].$$

   the multipliers μ, λ are nothing but a reflection of the expression for the normal cone to the linearization of the feasible set: it just so happens that the normal cone to linear constraints can be written as a linear combination of vectors, and μ, λ are simply these combination coefficients.

3. C) there exists $\mu \in \mathbb{R}^r$ and $\lambda \in \mathbb{R}^s_{\geq 0}$ such that x^* is a minimizer of the Lagrangian

$$L(x; \lambda, \mu) := f(x) + \sum_{i=1}^{r} \mu_i \, h_i(x) + \sum_{j=1}^{s} \lambda_j \, g_j(x) \quad \text{over } x \in \mathbb{R}^n, \; x^* \in \Omega,$$

   and $\lambda_j \, g_j(x^*) = 0$ for all $j \in [s]$.

4.8 Slater's Condition for Convex Optimization

Consider the convex program

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \; i = 1, \ldots, m, \\ & Ax = b, \end{aligned}$$

where $f_0, \ldots, f_m$ are convex functions on a domain $D \subseteq \mathbb{R}^n$, and $A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$.

Slater's Condition: The functions $\{f_i\}_{i=1}^m$ satisfy Slater's condition if there exists an x^* in the *relative interior* of D such that

$$f_i(x^*) < 0 \quad \text{for all } i = 1, \ldots, m.$$

Relaxed Slater Condition. Suppose that among the constraints some k functions $f_1, \ldots, f_k$ are affine. A relaxed Slater point $x^* \in \operatorname{relint} D$ then satisfies

$$f_i(x^*) \leq 0, \; i = 1, \ldots, k, \quad f_i(x^*) < 0, \; i = k+1, \ldots, m.$$

Consequence (Strong Duality). If the convex program is bounded below and (relaxed) Slater's condition holds, then strong duality applies: the optimal value of the primal problem equals that of its Lagrange dual, and a dual optimum is attained.

Remark. Slater's condition guarantees zero duality gap and the existence of Lagrange multipliers under mild regularity (convexity and constraint qualifications).

4.9 Relationship between duality and the Lagrangian

Weak Duality: for any choice of λ, μ with $\lambda \in \mathbb{R}^s_{\geq 0}$, $\mu \in \mathbb{R}^r$,

$$\inf_{x \in \mathbb{R}^n} L(x; \lambda, \mu) \; \leq \; \text{value}(P),$$

which is true even when Slater's conditions don't hold.

Strong Duality: if (P) admits a min, then the dual (D)

$$\max_{\mu,\lambda} \min_{x\in\mathbb{R}^n} L(x;\lambda,\mu) \quad \text{s.t.} \quad \lambda \in \mathbb{R}^s_{\geq 0},\ \mu \in \mathbb{R}^r$$

admits an optimal value λ, μ such that value(D) = value(P).

5 Sum of squares

5.1 Sum of squares polynomials

(Def) A sum of squares polynomial is a polynomial $p \in \mathbb{R}[x]$ such that it can be written as

$$p(x) = \sum_k P_k(x)^2 \quad \forall\, x \in \mathbb{R}^n,$$

for appropriate polynomials $P_k(x) \in \mathbb{R}[x]$. Denote the cone of sums of squares by $\Sigma[x]$.

5.2 When is a polynomial $p(x)$ SOS?

(Thrm) A polynomial over the reals $p(x) \in \mathbb{R}[x]$ satisfies $p(x) \in \Sigma[x]$ ("$p(x)$ is sos") if and only if there exists a positive-definite matrix $Q \in \mathbb{S}^{d\times d}$ such that

$$p(x) = v_d(x)^T\, Q\, v_d(x),$$

where if p has degree $2d$ and v_d is the vector of all monomials up to degree d.

(Example) If $p(x_1, x_2)$ has degree 4, then $v_d(x) = [\, x_1^2,\ x_2^2,\ x_1x_2,\ x_1,\ x_2,\ 1\,]^T$.

5.3 Using SOS formulations to prove minima of polynomials

To prove that $p(x) \geq t$ over $x \in \mathbb{R}^n$, try to find a Q that proves $p(x) - t \in \Sigma[x]$. $\rightarrow$

$$\begin{cases} \text{if feasible, then } p(x) \geq t\ \forall\, x \in \mathbb{R}^n, \\ \text{if infeasible, then } \exists\, x \in \mathbb{R}^n \text{ s.t. } p(x) < t. \end{cases}$$

5.4 When is $\mathbb{R}[x]_0^{n,d} = \Sigma[x]$?

This answers the question of what polynomials can be written as a sum of squares of polynomials.

(Thrm) Let $n = \#$ variables, $d =$ degree. Then $\mathbb{R}[x]_0^{n,d} = \Sigma[x]$ if and only if $n = 1$ for any degree d, $d = 2$ for any number of variables n, $n = 2$ and $d = 4$.

$\rightarrow$ Otherwise, the cone of sums of squares is a strict subset of the cone of non-negative polynomials: $\Sigma[x] \subset \mathbb{R}[x]_0$. E.g. the Motzkin polynomial is non-negative but cannot be written as sos.

5.5 If the polynomial has roots, must come with even multiplicity

6 Polarity and oracle equivalence

6.1 Definition of a polar set

(Def) Let Ω° be a compact and convex set such that

$$B_r(0) \subseteq \Omega \subseteq B_R(0) \quad \text{for some radii } 0 < r < R$$

then the polar set is

$$\Omega^\circ := \{ y \in \mathbb{R}^n : \langle x, y \rangle \leq 1 \quad \forall x \in \Omega \}.$$

6.2 Properties of the polar set

It has the following properties:

- Ω° is convex and compact
- $B_{\frac{1}{R}}(0) \subseteq \Omega^\circ \subseteq B_{\frac{1}{r}}(0)$
- $(\Omega^\circ)^\circ = \Omega$

6.3 Equivilance between methods on Ω and Ω°

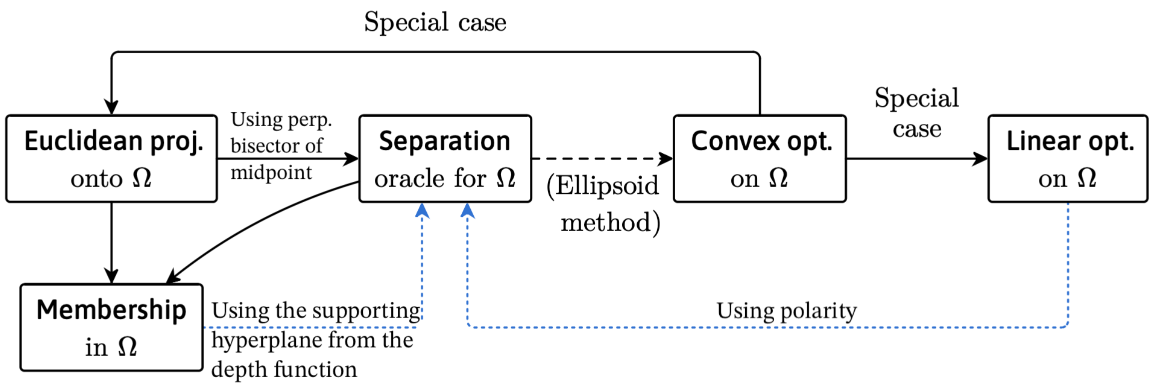

Figure 6: Equivalence between methods

Part I SDPs

1 Conic optimization

1.1 Detecting shared roots using the resultant ↔ Sylvester matrix

Define two polynomials with $a_n \neq 0$ and $b_m \neq 0$

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 \in \mathbb{R}[x], \qquad g(x) = b_m x^m + b_{m-1} x^{m-1} + \cdots + b_0 \in \mathbb{R}[x].$$

Then define the Sylvester matrix $\mathrm{Syl}(f,g)$ (size $(n+m)\times(n+m)$) by stacking the shifted coefficient rows:

$$\mathrm{Syl}(f,g) = \begin{pmatrix} a_n & a_{n-1} & \cdots & a_0 & & & \\ & a_n & a_{n-1} & \cdots & a_0 & & \\ & & \ddots & & & \ddots & \\ & & & a_n & a_{n-1} & \cdots & a_0 \\ b_m & b_{m-1} & \cdots & b_0 & & & \\ & b_m & b_{m-1} & \cdots & b_0 & & \\ & & \ddots & & & \ddots & \\ & & & b_m & b_{m-1} & \cdots & b_0 \end{pmatrix},$$

where the top block consists of m shifted rows of $(a_n,\ldots,a_0)$ and the bottom block consists of n shifted rows of $(b_m,\ldots,b_0)$, with blanks interpreted as 0.

$$\mathrm{Res}(f,g) \;:=\; \det\big(\mathrm{Syl}(f,g)\big).$$

Key identity (root factorization form). If $\alpha_1,\ldots,\alpha_n$ are the (complex) roots of f counted with multiplicity, then

$$\mathrm{Res}(f,g) = a_n^m \prod_{i=1}^{n} g(\alpha_i).$$

Equivalently, if $\beta_1,\ldots,\beta_m$ are the roots of g,

$$\mathrm{Res}(f,g) = (-1)^{nm} b_m^n \prod_{j=1}^{m} f(\beta_j).$$

Vanishing criterion.

$$\mathrm{Res}(f,g) = 0 \quad\Longleftrightarrow\quad f \text{ and } g \text{ have a common root in } \mathbb{C}$$

Over $\mathbb{R}$, this still detects common complex roots; a common real root is a special case. For clarity

$$\mathrm{Res}(f,g) = 0$$

means that there exists α such that $f(\alpha) = g(\alpha) = 0$, so the elimination of x from the system has a solution

1.2 Deteting multiplicity ≥ 2 using the discriminant

Define the discriminant of f by

$$\mathrm{Disc}(f) \;:=\; (-1)^{\frac{n(n-1)}{2}} \frac{1}{a_n} \mathrm{Res}(f,f').$$

where $f'(x)$ is the derivative of f. One nice property is that a real/complex number α is a root of f of multiplicity ≥ 2 of f if and only if

$$f(\alpha) = 0 \quad\text{and}\quad f'(\alpha) = 0.$$

We then have the root multiplicity criterion result, which says that

$$\mathrm{Disc}(f) = 0 \quad\Longleftrightarrow\quad f \text{ has a multiple root (i.e. multiplicity } \geq 2)$$

If f takes the special form $f(x) = a_n \prod_{i=1}^{n}(x - \alpha_i)$ over $\mathbb{C}$, then

$$\mathrm{Disc}(f) = a_n^{2n-2} \prod_{1 \le i < j \le n} (\alpha_i - \alpha_j)^2.$$