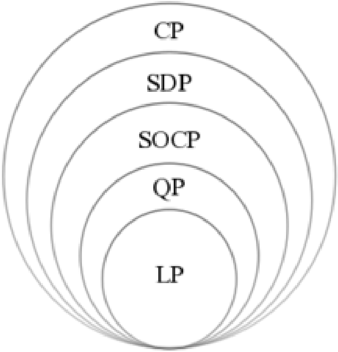
# Optimization in Machine Learning

## Linear Programming




### Learning goals

- Instances of LPs underlying statistical estimation
- Definition of an LP
- Geometric intuition of LPs

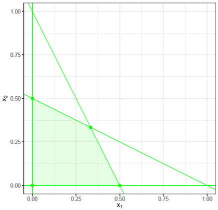
# LINEAR PROGRAMMING

**Linear problems** (LP):

**linear** objective function + **linear** constraints

**Example:**

$$
\begin{aligned}
\min \quad & -x_1 - x_2 \\
\text{s.t.} \quad & x_1 + 2x_2 \leq 1 \\
& 2x_1 + x_2 \leq 1 \\
& x_1, x_2 \geq 0
\end{aligned}
$$



# LINEAR PROGRAMMING

- **(Sparse) Quantile regression**:

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}} \quad & \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left( y^{(i)} - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}^{(i)} \right) \\ \text{s.t.} \quad & \|\boldsymbol{\beta}\|_1 \leq t \end{aligned}$$

where $\beta_0 \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are coefficients, and $\rho_\tau$, $\tau \in [0, 1]$, is the check function defined as

$$\rho_\tau(s) = \begin{cases} \tau \cdot s & \text{if } s > 0, \\ -1(1 - \tau) \cdot s & \text{if } s \leq 0. \end{cases}$$

**Case** $\tau = 1/2$**:** Median regression (a.k.a. least absolute errors (LAE), least absolute deviations (LAD))

Parameter $t \geq 0$ determines regularization.

# LINEAR PROGRAMMING

- **Dantzig selector**:

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \|\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{y})\|_\infty \leq \lambda \end{aligned}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, and $\lambda > 0$ is a tuning parameter. The infinity norm is defined as $\|x\|_\infty = \max\{|x_1|, \ldots, |x_i|, \ldots, |x_n|\}$ is

The Dantzig selector is similar (and behaves similar) to the Lasso and was introduced for variable selection in the seminal paper by Terence Tao and Emmanuel Candès (see moodle page for reference).

Details about LPs in statistical estimation can be found, e.g., in the PhD thesis of ▸ Yonggong Gao).

# LINEAR PROGRAMMING

LPs can be formulated in the **standard form**:

$$\begin{aligned} \max_{\mathbf{x}\in\mathbb{R}^n} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq 0 \end{aligned}$$

with $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$

- Constraints are to be understood **componentwise**
- $\mathbf{x} \geq 0$: "non-negativity constraint"
- $\mathbf{c}$: "cost vector"

# LINEAR PROGRAMMING

General LPs can be converted to standard form:

- $\min \longleftrightarrow \max$: multiply objective function by $-1$
- $\leq \longleftrightarrow \geq$: multiply inequality by $-1$
- $= \longleftrightarrow \leq, \geq$: replace $\mathbf{a}_i^\top \mathbf{x} = b_i$ by $\mathbf{a}_i^\top \mathbf{x} \geq b_i$ *and* $\mathbf{a}_i^\top \mathbf{x} \leq b_i$
- No non-negativity constraint: replace $x_i$ by $x_i^+ - x_i^-$ with $x_i^+, x_i^- \geq 0$ (positive and negative part)

# LINEAR PROGRAMMING

**Example:**

$$
\begin{aligned}
\text{min} \quad & -x_1 - x_2 \\
\text{s.t.} \quad & x_1 + 2x_2 \leq 1 \\
& 2x_1 + x_2 \leq 1 \\
& x_1, x_2 \geq 0
\end{aligned}
$$

can also be formulated as

$$
\begin{aligned}
\text{max} \quad & (1, 1) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
\text{s.t.} \quad & \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \mathbf{x} \leq \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
& \mathbf{x} \geq 0
\end{aligned}
$$

# GEOMETRIC INTERPRETATION

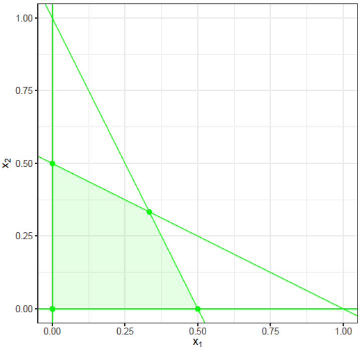
Linear programming can be interpreted geometrically.

**Feasible set:**

- $i$-th inequality constraint: $\mathbf{a}_i^\top \mathbf{x} \leq b_i$
- Points $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} = b_i\}$ form a hyperplane in $\mathbb{R}^n$
  ($\mathbf{a}_i$ is perpendicular to the hyperplane and called **normal vector**)
- Points $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} \geq b_i\}$ lie on the side of the hyperplane into which the normal vector points ("half-space")

# GEOMETRIC INTERPRETATION

- Each inequality divides the space into two halves.
- **Claim:** Points satisfying **all** inequalities form a **convex polytope**.



# GEOMETRIC INTERPRETATION

Geometry: A **polytope** is a generalized polygon in arbitrary dimensions.

A polytope consists of several sub-polytopes:

- 0-polytope: point
- 1-polytope: line
- 2-polytope: polygon, ...

**General:**

- $d$-polytope is formed from several $(d-1)$-polytopes ("facets")
- $(d-1)$-polytope is formed from several $(d-2)$-polytopes

# GEOMETRIC INTERPRETATION

**Observe:** Points $\{\mathbf{x} : \mathbf{a}_i^\top \mathbf{x} = b_i\}$ lie on the boundary of the polytope.

- Polytope $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is convex: For $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{S}$ and $t \in [0, 1]$

$$
\begin{aligned}
\mathbf{A}(\mathbf{x}_1 + t(\mathbf{x}_2 - \mathbf{x}_1)) &= \mathbf{A}\mathbf{x}_1 + t(\mathbf{A}\mathbf{x}_2 - \mathbf{A}\mathbf{x}_1) \\
&= (1 - t)\underbrace{\mathbf{A}\mathbf{x}_1}_{\leq \mathbf{b}} + t\underbrace{\mathbf{A}\mathbf{x}_2}_{\leq \mathbf{b}} \\
&\leq (1 - t)\mathbf{b} + t\mathbf{b} = \mathbf{b}
\end{aligned}
$$

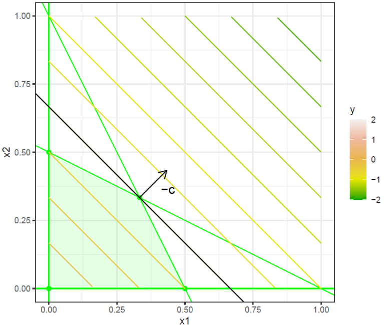
- Polytope $\{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ is an $n$-**simplex**, i.e.,

  convex hull of $n + 1$ *affinely independent* points

# GEOMETRIC INTERPRETATION

## Objective function:

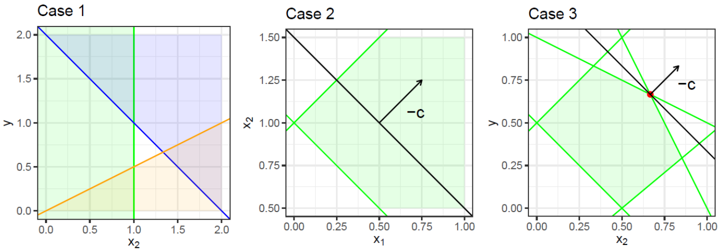
- **Linear case:** Contour lines form a hyperplane
- **Observe:** **c** is gradient and perpendicular to contour lines
- Solution "touches" the polygon



# SOLUTIONS TO LP

There are 3 ways to solve linear programming:

1. Feasible set is **empty** $\Rightarrow$ LP is infeasible
2. Feasible set is **"unbounded"**
3. Feasible set is **"bounded"** $\Rightarrow$ LP has at least one solution



# SOLUTIONS TO LP

- If LP is solvable and constrained (neither case 1 nor case 2), there is always an optimal point that can **not** be convexly combined from other points in the polytope.
- The optimal solution is then a corner, edge or side of the polytope.