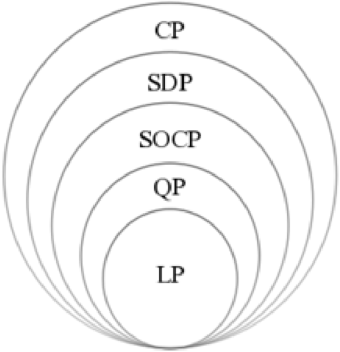
Optimization in Machine Learning

Constrained Optimization
Linear Programming



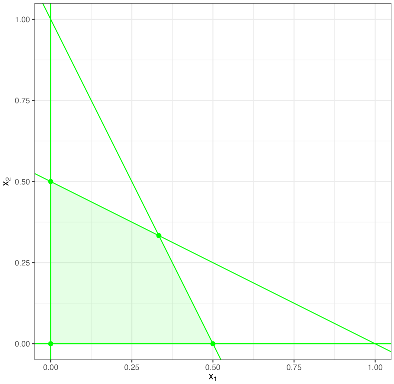
Learning goals

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

LINEAR PROGRAMMING

- **Linear problems** (LP): linear objective + linear constraints
- **Example:**

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



LP EXAMPLES: QUANTILE REGRESSION

- **(Sparse) Quantile regression**:

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

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

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

- **Case $\tau = 1/2$:** Median regression (a.k.a. least absolute errors (LAE), least absolute deviations (LAD))
- Parameter $t \geq 0$ determines regularization

LP EXAMPLES: DANTZIG SELECTOR

- **Dantzig selector**:

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

  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 $\|x\|_\infty = \max\{|x_1|, \dots, |x_n|\}$
- Similar to (and behaves similar to) the Lasso
- Introduced for variable selection by Tao and Candès
- Details about LPs in statistical estimation: see ▸ PhD thesis of Yonggong Gao

LP: STANDARD FORM

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

$$\max_{\boldsymbol{x} \in \mathbb{R}^n} \mathbf{c}^T \boldsymbol{x} \quad \text{s.t. } \mathbf{A}\boldsymbol{x} \leq \mathbf{b}, \; \boldsymbol{x} \geq 0$$

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

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

LP: CONVERTING TO STANDARD FORM

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

LP: STANDARD FORM EXAMPLE

- Example:

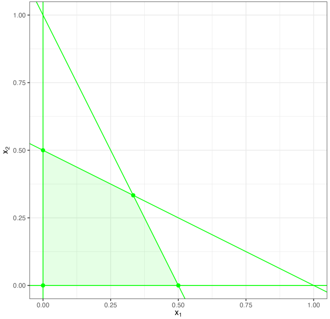
$$\min -x_1 - x_2 \quad \text{s.t. } x_1 + 2x_2 \leq 1,\; 2x_1 + x_2 \leq 1,\; x_1, x_2 \geq 0$$

- Can also be formulated as:

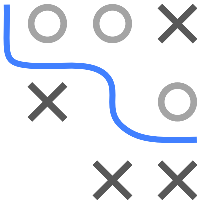
$$\max(1,1)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{s.t. } \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \boldsymbol{x} \leq \begin{pmatrix} 1 \\ 1 \end{pmatrix},\; \boldsymbol{x} \geq 0$$

GEOMETRIC INTERPRETATION: FEASIBLE SET

- *i*-th inequality constraint: $\mathbf{a}_i^T \boldsymbol{x} \leq b_i$
- Points $\{\boldsymbol{x} : \mathbf{a}_i^T \boldsymbol{x} = b_i\}$ form a hyperplane in $\mathbb{R}^n$
  ($\mathbf{a}_i$ is perpendicular to the hyperplane and called **normal vector**)
- Points $\{\boldsymbol{x} : \mathbf{a}_i^T \boldsymbol{x} \geq b_i\}$ lie on the side of the hyperplane into which the normal vector points ("half-space")
- Each inequality divides the space into two halves
- **Claim:** Points satisfying **all** inequalities form a **convex polytope**

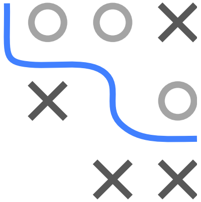


GEOMETRIC INTERPRETATION: POLYTOPES



- 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: CONVEXITY



- Points $\{\boldsymbol{x} : \mathbf{a}_i^T \boldsymbol{x} = b_i\}$ lie on the boundary of the polytope

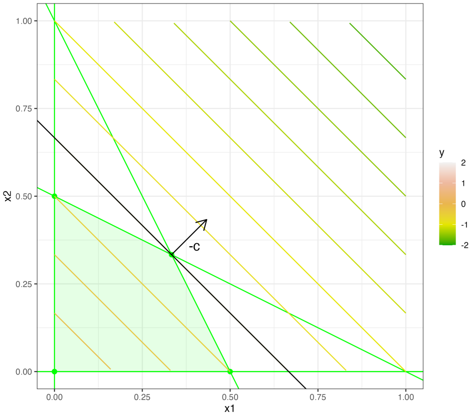
- Polytope $\{\boldsymbol{x} : \mathbf{A}\boldsymbol{x} \leq \mathbf{b}\}$ is convex: For $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{S}$ and $t \in [0, 1]$

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

- Polytope $\{\boldsymbol{x} : \mathbf{A}\boldsymbol{x} \leq \mathbf{b}\}$ is an n-**simplex**, i.e., convex hull of $n + 1$ *affinely independent* points

GEOMETRIC INTERPRETATION: OBJECTIVE FUNCTION

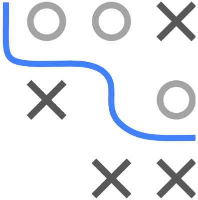
- **Linear case:** Contour lines form a hyperplane
- **Observe: c** is gradient and perpendicular to contour lines
- Solution "touches" the polygon

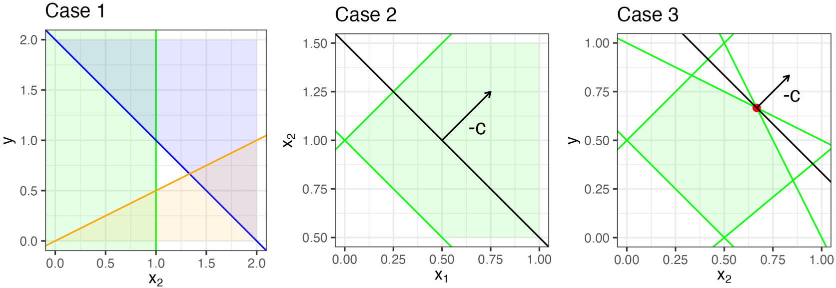


SOLUTIONS TO LP

There are 3 cases for 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