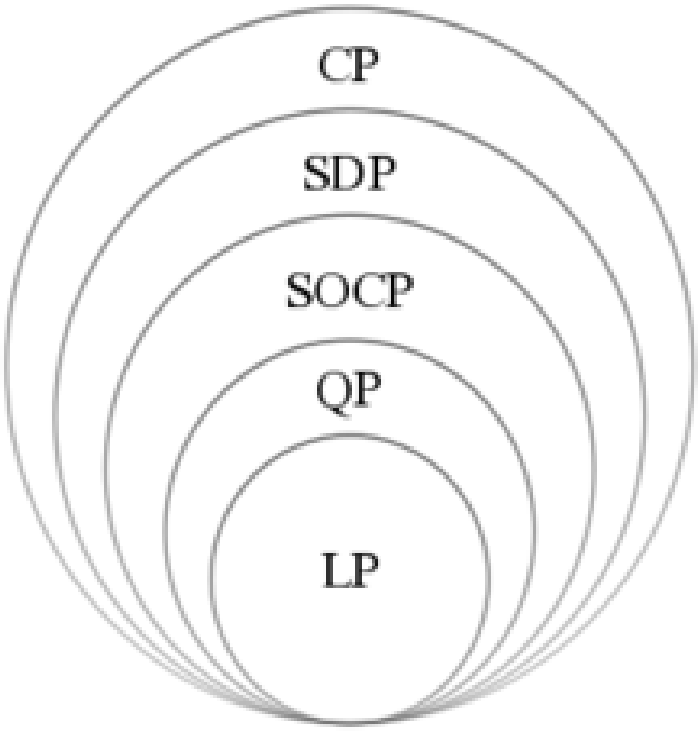
Optimization in Machine Learning

Linear Programming



Learning goals

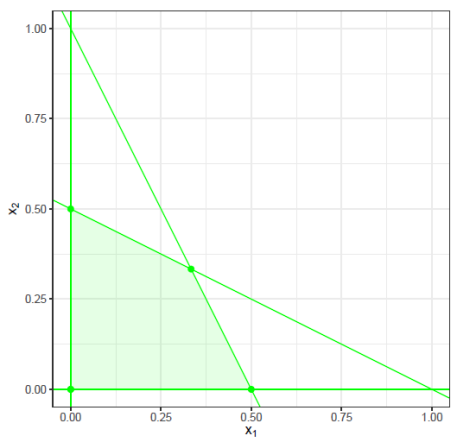
- Instances of LPs underlying statistical estimation
- Definition of an LP
- Geometric intuition of LPs

LINEAR PROGRAMMING

Special case: Linear programming (LP)
objective function and constraints are **linear functions**.

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**:

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

where scalar β_0 and $\boldsymbol{\beta} \in \mathbb{R}^p$ are the quantile regression coefficients for $\tau \in [0, 1]$, and $\rho_\tau(\cdot)$ is the check function defined as

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

When parameter $\tau = 1/2$, quantile regression amounts to median regression, least absolute error (LAE), or least absolute deviation (LAD). As in Lasso, $t \geq 0$ is a tuning parameter.

LINEAR PROGRAMMING

- **Dantzig selector**:

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

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, 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

W.l.o.g. Linear programming is specified using the so-called **standard form**.

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

The inequality constraints $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$ ($\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$) and $\boldsymbol{x} \geq 0$ are to be understood componentwise.

The condition $\boldsymbol{x} \geq 0$ is known as the **non-negativity constraint** and the vector $\boldsymbol{c}$ is known as the **cost vector**.

LINEAR PROGRAMMING

Linear optimization problems can be converted to the standard form by the following operations:

- Maximization instead of minimization: multiplication of the cost vector $\boldsymbol{c}$ by -1
- Less than or equal instead of greater than or equal: multiply the inequality by -1
- Equality instead of inequality: replace $\boldsymbol{a}_i\boldsymbol{x} = b_i$ with two conditions of inequality $\boldsymbol{a}_i\boldsymbol{x} \geq b_i$ and $\boldsymbol{a}_i\boldsymbol{x} \leq b_i$
- Variable without non-negativity constraint: replace x_i with $x_i^+ - x_i^-$ with $x_i^+, x_i^- \geq 0$ (positive or negative part).

In the following we assume that the LP is given in standard form.

LINEAR PROGRAMMING

Example:

The example above

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

can also be formulated as

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

GEOMETRIC INTERPRETATION

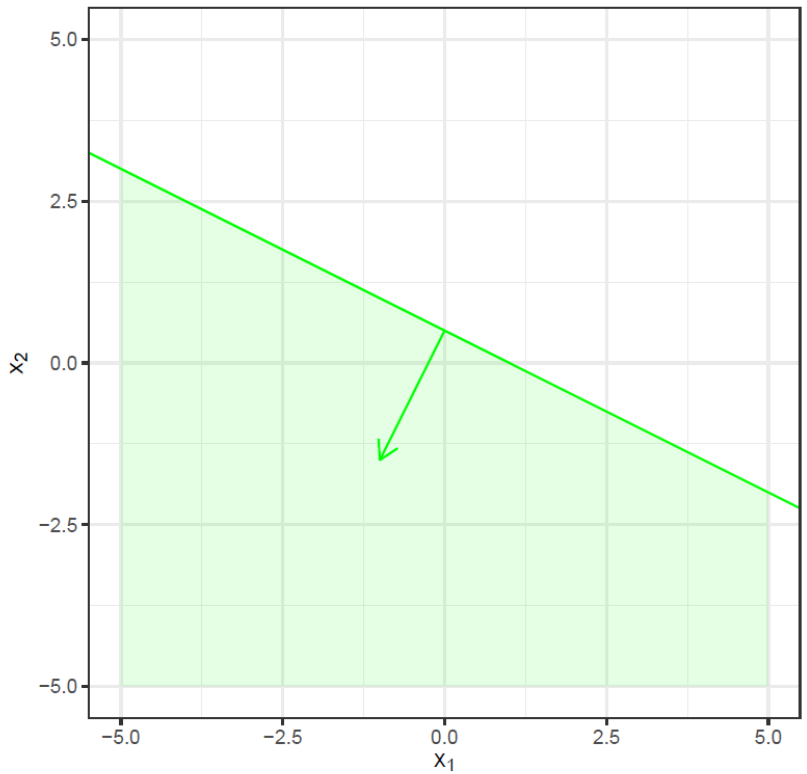
Linear programming can be interpreted geometrically.

Feasible set:

- Let $\boldsymbol{a}_i\boldsymbol{x} \geq b_i$ be the i-th line of the inequality conditions.
- The points that satisfy the linear system $\boldsymbol{a}_i\boldsymbol{x} = b_i$ form a hyperplane in n-dimensional space.
- The vector $\boldsymbol{a}_i$ is perpendicular to the plane and is called the **normal vector**.

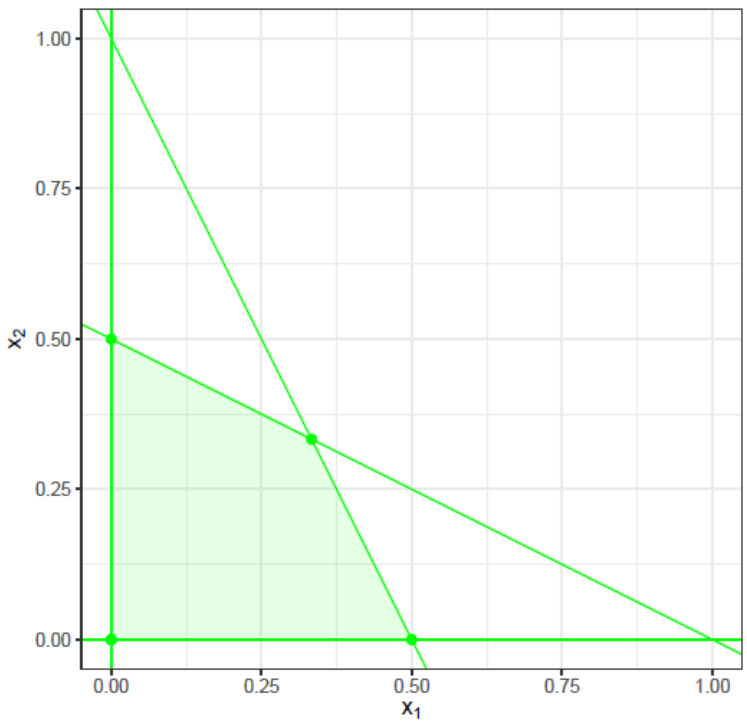
GEOMETRIC INTERPRETATION

- The set of points $\{\boldsymbol{x} : \boldsymbol{a}_i\boldsymbol{x} \geq b_i\}$ consists of points on the side of the hyperplane into which the normal vector points (**half-space**).



GEOMETRIC INTERPRETATION

- Each of the m inequalities divides the n-dimensional space into two halves.
- **Claim:** The points that satisfy **all** inequalities form a **convex polytope**.



GEOMETRIC INTERPRETATION

In geometry, a **polytope** denotes a generalized polygon in an arbitrary dimension. A polytope consists of several subpolytopes.

- A 0 polytope is a single point,
- A 1 polytope is a line,
- A 2 polytope is a polygon, ...

In general, a d polytope is formed from several $(d-1)$ polytopes (so-called facets) which in turn can have a $(d-2)$ polytope in common. Thus a 3 polytope (e.g. a cube) has several sides / facets, some of which have common edges, etc.

GEOMETRIC INTERPRETATION

The points for which $\boldsymbol{a}_i \boldsymbol{x} = b_i$ applies lie on the **facet** of the polytope.

The polytope which is defined by the inequalities $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$ is convex. For two points $\boldsymbol{x}_1, \boldsymbol{x}_2$, which lie in the polytope, any point that results from the convex combination of the two points, is again inside the polytope:

$$\begin{aligned} \boldsymbol{A}(\boldsymbol{x}_1 + t(\boldsymbol{x}_2 - \boldsymbol{x}_1)) &= \boldsymbol{A}\boldsymbol{x}_1 + t(\boldsymbol{A}\boldsymbol{x}_2 - \boldsymbol{A}\boldsymbol{x}_1) \\ &= (1-t)\underbrace{\boldsymbol{A}\boldsymbol{x}_1}_{\geq \boldsymbol{b}} + t\underbrace{\boldsymbol{A}\boldsymbol{x}_2}_{\geq \boldsymbol{b}} \\ &\geq (1-t)\boldsymbol{b} + t\boldsymbol{b} = \boldsymbol{b} \qquad \text{for } t \in [0,1]) \end{aligned}$$

A polytope formed by the convex hull of $(n+1)$ affin independent points in $\mathbb{R}^n$ is also called n-**simplex**.

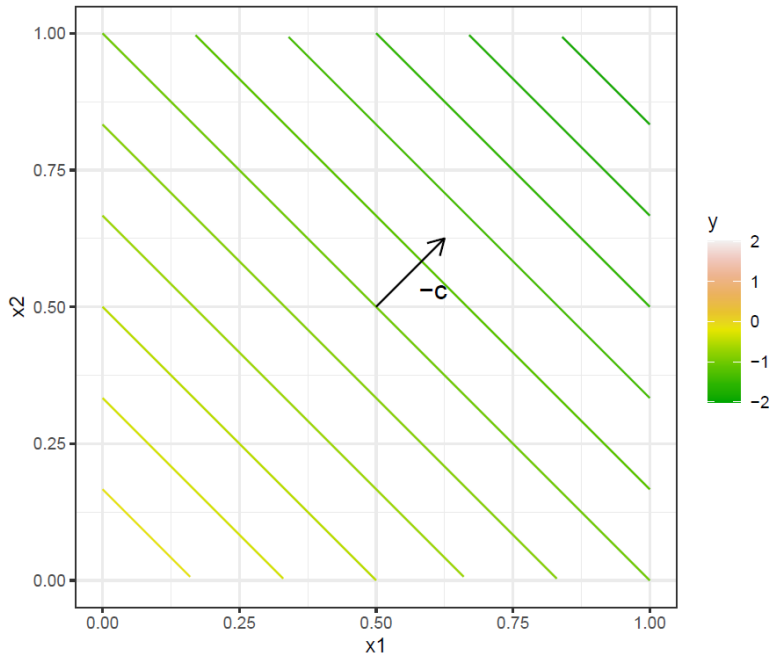
GEOMETRIC INTERPRETATION

Objective function:

- For the objective function, we consider the **contour lines**. Points that lie on the same contour line have the same objective function value.
- In the linear case, the contours $y = \boldsymbol{c}^T \boldsymbol{x}$ are also a hyperplane for fixed y.
- The vector $\boldsymbol{c}$ is again perpendicular to the respective contour lines.
- The vector $\boldsymbol{c}$ can also be interpreted as a gradient. The **negative** gradient $-\boldsymbol{c}$ points in the direction of the “steepest” descent.

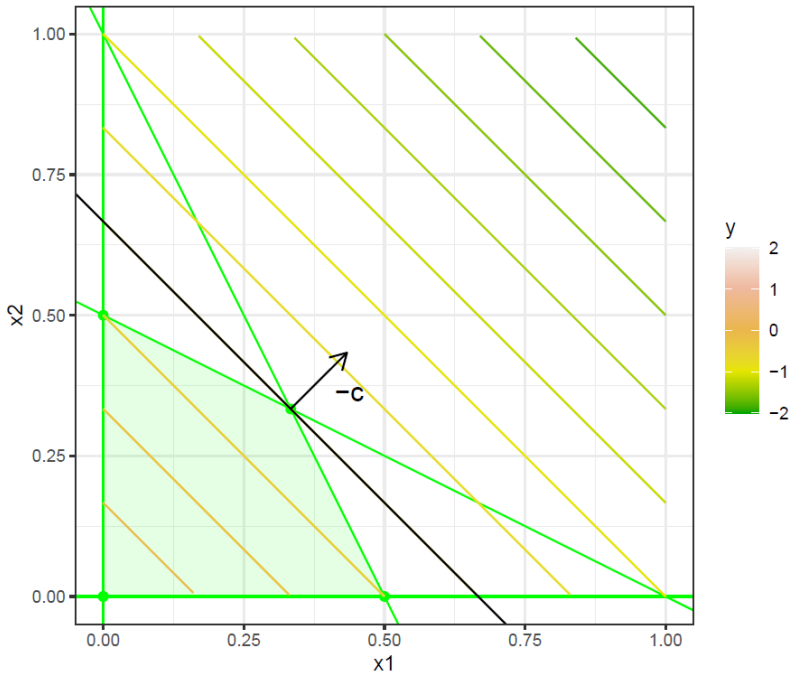
GEOMETRIC INTERPRETATION

- The value of the function becomes smaller when we go in the direction of the **negative gradient** $-\boldsymbol{c}$.



GEOMETRIC INTERPRETATION

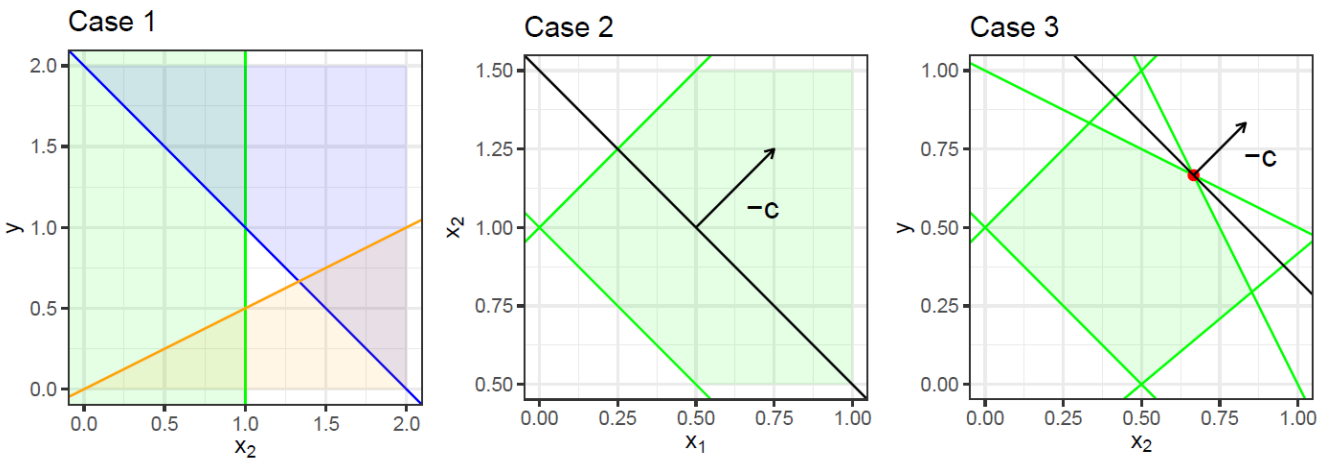
- So we move the objective function **in the opposite direction** of the normal vector until the line just touches the polygon.



SOLUTIONS TO LP

There are three ways to solve Linear programming:

1. LP is **infeasible**, the feasible set is empty ($\mathcal{S} = \emptyset$)
2. LP is unconstrained
3. LP has at least one optimal 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.