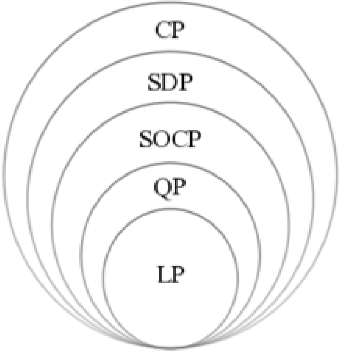
# Optimization in Machine Learning

## Constrained Optimization




**Learning goals**

- Examples of constrained optimization in statistics and ML
- General definition
- Hierarchy of convex constrained problems

# CONSTRAINED OPTIMIZATION IN STATISTICS

**Example**: Maximum Likelihood Estimation

For data $\left(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\right)$, we want to find the maximum likelihood estimate

$$\max_{\theta} L(\theta) = \prod_{i=1}^{n} f(\mathbf{x}^{(i)}, \theta)$$

In some cases, $\theta$ can only take **certain values**.

- If $f$ is a Poisson distribution, we require the rate $\lambda$ to be non-negative, i.e. $\lambda \geq 0$

# CONSTRAINED OPTIMIZATION IN ML

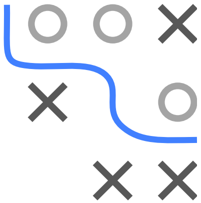
- **Lasso regression**:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \left( y^{(i)} - \beta^T \mathbf{x}^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\beta\|_1 \leq t \end{aligned}$$

- **Ridge regression**:

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \frac{1}{n} \sum_{i=1}^{n} \left( y^{(i)} - \beta^T \mathbf{x}^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\beta\|_2 \leq t \end{aligned}$$

# CONSTRAINED OPTIMIZATION



**General definition** of a **Constrained Optimization problem**:

$$\begin{aligned} \min \quad & f(\mathbf{x}) & \\ \text{such that} \quad & g_i(\mathbf{x}) \leq 0 & \text{for } i = 1, \dots, k \\ & h_j(\mathbf{x}) = 0 & \text{for } j = 1, \dots, l, \end{aligned}$$

where

- $g_i : \mathbb{R}^d \to \mathbb{R}, i = 1, ..., k$ are inequality constraints,
- $h_j : \mathbb{R}^d \to \mathbb{R}, j = 1, ..., l$ are equality constraints.

The set of inputs **x** that fulfill the constraints, i.e.,

$$\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^d \mid g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0 \;\forall\; i, j\},$$

is known as the **feasible set**.

# CONSTRAINED CONVEX OPTIMIZATION

Special cases of constrained optimization problems are **convex programs**, with convex objective function $f$, convex inequality constraints $g_i$, and affine equality constraints $h_j$ (i.e. $h_j(\mathbf{x}) = \mathbf{A}_j^\top \mathbf{x} - \mathbf{b}_j$).

Convex programs can be categorized into