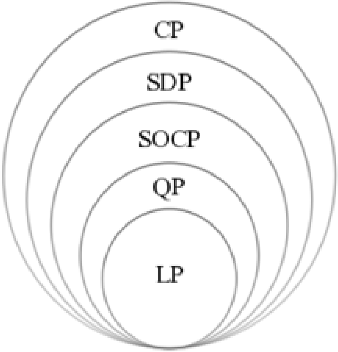
Optimization in Machine Learning

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, θ can only take **certain values**.

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

CONSTRAINED OPTIMIZATION IN STATISTICS

- If f is a multinomial distribution

$$f(x_1, ..., x_p; n; \theta_1, ..., \theta_p) = \begin{cases} \binom{n!}{x_1! \cdot x_2! \dots x_p!} \theta_1^{x_1} \cdot ... \cdot \theta_p^{x_p} & \text{if } x_1 + ... + x_p = n \\ 0 & \text{else} \end{cases}$$

The probabilities θ_i must lie between 0 and 1 and add up to 1, i.e. we require

$$\begin{aligned} 0 \leq \theta_i \leq 1 \qquad & \text{for all } i \\ \theta_1 + ... + \theta_p = 1. & \end{aligned}$$

CONSTRAINED OPTIMIZATION IN ML

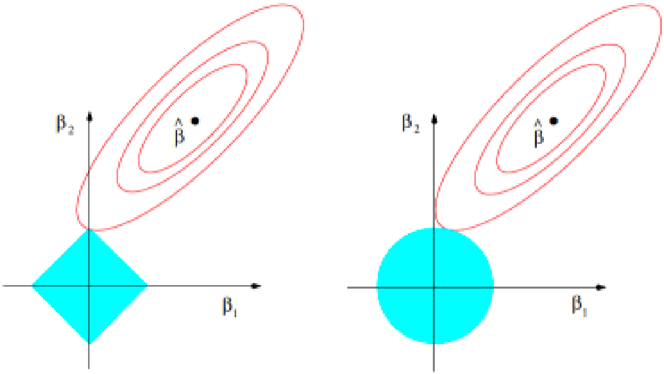
- **Lasso regression**:

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

- **Ridge regression**:

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

CONSTRAINED OPTIMIZATION IN ML



CONSTRAINED OPTIMIZATION IN ML

- **Constrained Lasso regression**:

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

where the matrices $\mathbf{A} \in \mathbb{R}^{l \times p}$ and $\mathbf{C} \in \mathbb{R}^{k \times p}$ have full row rank.

This model includes many Lasso variants as special cases, e.g., the Generalized Lasso, (sparse) isotonic regression, log-contrast regression for compositional data, etc. (see, e.g., Gaines et al., 2018).

CONSTRAINED OPTIMIZATION IN ML

Remember the dual formulation of the SVM, which is a convex quadratic program with box constraints plus one linear constraint:

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y^{(i)} y^{(j)} \left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle \\
\text{s.t.} \quad & 0 \leq \alpha_i \leq C, \\
& \sum_{i=1}^{n} \alpha_i y^{(i)} = 0,
\end{aligned}
$$

CONSTRAINED OPTIMIZATION

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

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

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.,

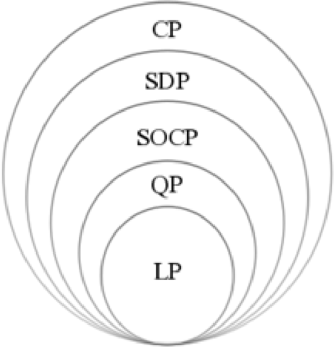
$$\mathcal{S} := \{\mathbf{x} \in \mathbb{R}^d \mid g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0 \; \forall \; i, j\},$$

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



CONSTRAINED CONVEX OPTIMIZATION

- Linear program (LP): objective function f and all constraints g_i, h_j are linear functions
- Quadratic program (QP): objective function f is a quadratic form, i.e.

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{c}^\top \mathbf{x} + \mathbf{d}$$

  for $\mathbf{Q} \in \mathbb{R}^{d \times d}, \mathbf{c} \in \mathbb{R}^d, d \in \mathbb{R}$, and constraints are linear.

as well as second-order cone programs (SOCP), semidefinite programs (SDP), and cone programs (CP).

CONSTRAINED CONVEX OPTIMIZATION

SOCPs play a pivotal role in statistics and engineering and have been popularized in the seminal article by ▸ Lobo et al., 1998.

In ML, SDPs are at the heart of, e.g., learning kernels from data (see, e.g., ▸ Lanckriet et al., 2004).

In general, this categorization of convex optimization problem classes helps in the design of specialized *optimization methods* that are tailored toward the specific type of convex optimization problem (keyword: disciplined convex programming ▸ Grant et al., 2006).