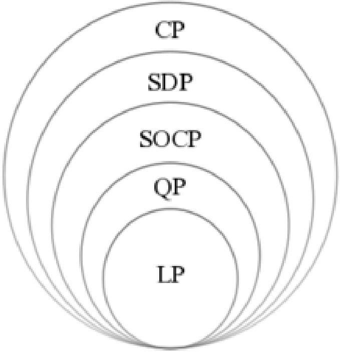
Optimization in Machine Learning

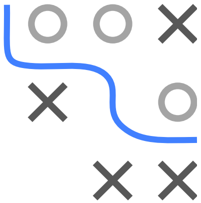
Constrained Optimization
Introduction



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 $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$, we want to find the MLE

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

- In some cases, θ can only take **certain values**
- If f is a Poisson distribution, we require $\lambda \geq 0$
- If f is a multinomial distribution

$$f(x_1, \ldots, x_p; n; \theta_1, \ldots, \theta_p) = \begin{cases} \frac{n!}{x_1! \cdots x_p!} \theta_1^{x_1} \cdots \theta_p^{x_p} & \text{if } x_1 + \ldots + x_p = n \\ 0 & \text{else} \end{cases}$$

- The probabilities θ_i must lie between 0 and 1 and sum to 1:

$$0 \leq \theta_i \leq 1 \text{ for all } i \quad \text{and} \quad \theta_1 + \ldots + \theta_p = 1$$

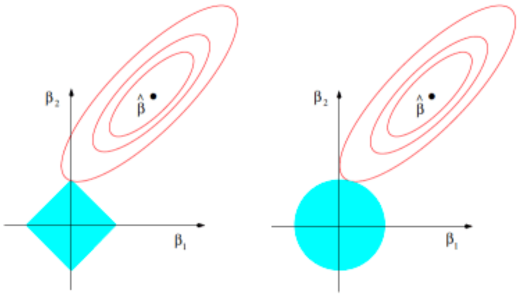
CONSTRAINED OPTIMIZATION IN ML

- **Lasso regression**:

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

- **Ridge regression**:

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



CONSTRAINED OPTIMIZATION IN ML

- **Constrained Lasso regression**:

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

- 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. ▸ Gaines, Kim, and Zhou 2018

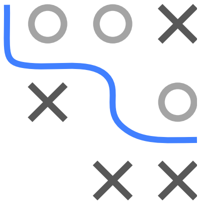
CONSTRAINED OPTIMIZATION IN ML

- Dual formulation of the SVM: convex QP with box constraints plus one linear constraint

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y^{(i)} y^{(j)} \left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

CONSTRAINED OPTIMIZATION



- **Constrained Optimization problem**:

$$\min f(\boldsymbol{x}) \quad \text{s.t. } g_i(\boldsymbol{x}) \leq 0 \ (i = 1, \dots, k), \quad h_j(\boldsymbol{x}) = 0 \ (j = 1, \dots, l)$$

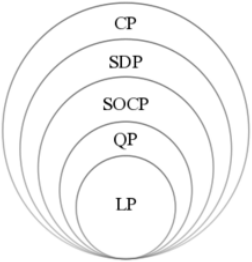
- $g_i : \mathbb{R}^d \to \mathbb{R}$ are **inequality constraints**
- $h_j : \mathbb{R}^d \to \mathbb{R}$ are **equality constraints**

- The **feasible set** is the set of inputs $\boldsymbol{x}$ that fulfill the constraints:

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

CONSTRAINED CONVEX OPTIMIZATION

- **Convex programs**: convex objective f, convex inequality constraints g_i, and affine equality constraints h_j (i.e., $h_j(\boldsymbol{x}) = \mathbf{A}_j^T \boldsymbol{x} - \mathbf{b}_j$)

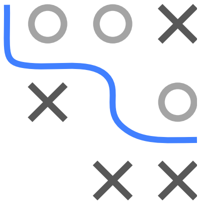


- **Linear program (LP)**: f and all constraints g_i, h_j are linear
- **Quadratic program (QP)**: f is a quadratic form, constraints are linear

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \mathbf{Q}\boldsymbol{x} + \mathbf{c}^T \boldsymbol{x} + d \quad \text{for } \mathbf{Q} \in \mathbb{R}^{d \times d}, \mathbf{c} \in \mathbb{R}^d, d \in \mathbb{R}$$

- Also: second-order cone programs (SOCP), semidefinite programs (SDP), cone programs (CP)

CONSTRAINED CONVEX OPTIMIZATION



- SOCPs play a pivotal role in statistics and engineering
  ▸ Lobo et al. 1998
- In ML, SDPs are at the heart of, e.g., learning kernels from data
  ▸ Lanckriet et al. 2004
- This categorization of convex optimization problem classes helps design specialized *optimization methods* tailored to specific problem types
- Keyword: disciplined convex programming ▸ Grant, Boyd, and Ye 2006