# **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  $(\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)})$ , 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 λ to be non-negative, i.e. λ ≥ 0 0 0 X X 0 X X

### **CONSTRAINED OPTIMIZATION IN STATISTICS / 2**

• If *f* is a multinomial distribution

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

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

$$0 \le heta_i \le 1$$
 for all  $i$   
 $heta_1 + \dots + heta_p = 1.$ 

X

 $\times \times$ 

#### **CONSTRAINED OPTIMIZATION IN ML**

• Lasso regression:

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

• Ridge regression:

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

#### **CONSTRAINED OPTIMIZATION IN ML / 2**





#### **CONSTRAINED OPTIMIZATION IN ML / 3**

• Constrained Lasso regression:

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

$$\mathbf{C}\boldsymbol{\beta} \leq \mathbf{d}$$

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{b},$$

× 0 0 × ×

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 / 4**

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

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

× × 0 × × ×

## **CONSTRAINED OPTIMIZATION**

General definition of a Constrained Optimization problem:

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

where

- $g_i : \mathbb{R}^d \to \mathbb{R}, i = 1, ..., k$  are inequality constraints,
- $h_j : \mathbb{R}^d \to \mathbb{R}, j = 1, ..., I$  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  $h_i$  (i.e.  $h_i(\mathbf{x}) = \mathbf{A}_i^{\top}\mathbf{x} - \mathbf{b}_i$ ).

Convex programs can be categorized into





# CONSTRAINED CONVEX OPTIMIZATION / 2

- Linear program (LP): objective function *f* and all constraints *g<sub>i</sub>*, *h<sub>j</sub>* 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).

× 0 0 × × ×

# **CONSTRAINED CONVEX OPTIMIZATION / 3**

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).

× × 0 × × ×