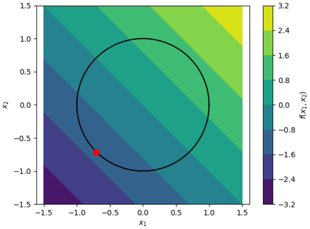
Optimization

Constrained problems



Learning goals

- TODO
- TODO

GENERAL DEFINITION

Consider the **optimization problem**

$$\min_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x})$$

with objective function

$$f : \mathcal{S} \to \mathbb{R}.$$

The problem is called **constrained**, if the domain $\mathcal{S}$ is restricted:

$$\mathcal{S} \subsetneq \mathbb{R}^d.$$

$\mathcal{S}$ is typically defined via functions called **constraints**:

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

where

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

GENERAL DEFINITION

We also write the general constrained optimization problem as:

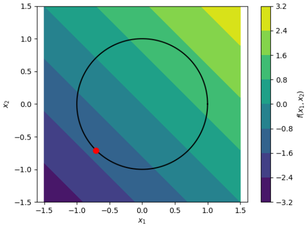
$$
\begin{aligned}
\min \quad & f(\mathbf{x}) & \\
\text{such that} \quad & g_i(\mathbf{x}) \leq 0 & \text{for } i = 1, \ldots, k \\
& h_j(\mathbf{x}) = 0 & \text{for } j = 1, \ldots, l.
\end{aligned}
$$

Types of constraints.... If f and all constraints are **smooth**, the problem is **smooth**.

EXAMPLE 1: UNIT CIRCLE

Example for a constrained optimization problem: minimization on the unit circle

$$\begin{array}{ll} \min & f(x_1, x_2) = x_1 + x_2 \\ \text{s.t.} & g(x_1, x_2) = x_1^2 + x_2^2 - 1 = 0 \end{array}$$



EXAMPLE 2: MAXIMUM LIKELIHOOD ESTIMATION

Example: Maximum Likelihood Estimation

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

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

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. $\lambda \geq 0$

EXAMPLE 2: MAXIMUM LIKELIHOOD ESTIMATION

- 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

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

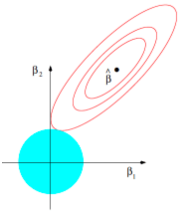
EXAMPLE 3: RIDGE REGRESSION

In Ridge regression, we add an L_2 penalty on $\boldsymbol{\theta}$:

$$\hat{\theta}_{\text{Ridge}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta} \right) \right)^2 + \lambda ||\boldsymbol{\theta}||_2^2 \right\}$$

To get a better understanding of the geometry, we reformulate the optimization as a constrained problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_2 \leq t \end{aligned}$$



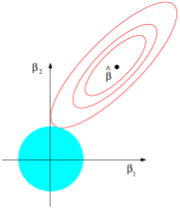
EXAMPLE 3: RIDGE REGRESSION

In Ridge regression, we add an L_2 penalty on $\boldsymbol{\theta}$:

$$\hat{\theta}_{\text{Ridge}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 + \lambda ||\boldsymbol{\theta}||_2^2 \right\}$$

To get a better understanding of the geometry, we reformulate the optimization as a constrained problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_2 \leq t \end{aligned}$$



These are smooth, (strongly) convex optimization problems in quadratic form. Usually, the unconstrained formulation is used. Again, we can either compute the closed form solution given by $\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X} + \lambda \boldsymbol{I})^{-1} \mathbf{X}^\top \mathbf{y}$ or use a gradient based optimization method.

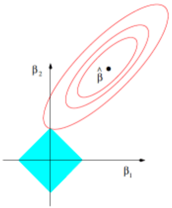
EXAMPLE 4: LASSO REGRESSION

In LASSO regression, we add an L_1 penalty on $\boldsymbol{\theta}$:

$$\hat{\theta}_{\text{Lasso}} \quad = \quad \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 + \lambda ||\boldsymbol{\theta}||_1 \right\}$$

Analogously, the problem can be reformulated as a constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_1 \leq t \end{aligned}$$



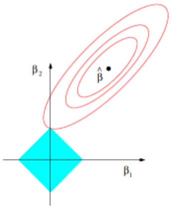
EXAMPLE 4: LASSO REGRESSION

In LASSO regression, we add an L_1 penalty on $\boldsymbol{\theta}$:

$$\hat{\theta}_{\text{Lasso}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} \left(y^{(i)} - f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right)^2 + \lambda ||\boldsymbol{\theta}||_1 \right\}$$

Analogously, the problem can be reformulated as a constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \sum_{i=1}^{n} \left(\boldsymbol{\theta}^\top \mathbf{x}^{(i)} - y^{(i)} \right)^2 \\ \text{s.t.} \quad & \|\boldsymbol{\theta}\|_1 \leq t \end{aligned}$$



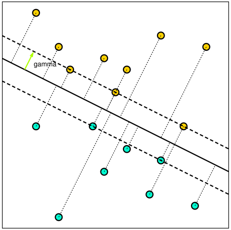
These are non-smooth, convex optimization problems. There is no closed form solution and optimization is harder due to the non-differentiability of the constraint. Here, we could use derivative-free optimization methods, e.g. coordinate descent.

EXAMPLE 4: LASSO REGRESSION

Add geometric interpretation for L1 (clipping operator).

EXAMPLE 6: SUPPORT VECTOR MACHINES

- In a linear support vector machine problem, we want to find a linear decision boundary which separates the classes with a **maximum** safety distance.
- This means, the distance to the points that are closest to the hyperplane ("safety margin γ") should be **maximized**.
- We allow violations of the margin constraints via slack variables $\zeta^{(i)} \geq 0$



The safety margin γ is indicated by a green arrow.

A more thorough introduction to SVMs is given in "Supervised learning".

EXAMPLE 5: SUPPORT VECTOR MACHINES

This soft-margin SVM optimization problem can be reformulated:

$$
\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\theta}_0, \zeta^{(i)}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^{n} \zeta^{(i)} \\
\text{s.t.} \quad & y^{(i)} \left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \boldsymbol{\theta}_0 \right) \geq 1 - \zeta^{(i)} \quad \forall\, i \in \{1, \dots, n\}, \\
\text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, n\}.
\end{aligned}
$$

The parameter C controls the trade-off between the two conflicting objectives of maximizing the size of the margin and minimizing the frequency and size of margin violations.

EXAMPLE 5: SUPPORT VECTOR MACHINES

This soft-margin SVM optimization problem can be reformulated:

$$
\begin{aligned}
\min_{\boldsymbol{\theta}, \boldsymbol{\theta}_0, \zeta^{(i)}} \quad & \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^{n} \zeta^{(i)} \\
\text{s.t.} \quad & y^{(i)} \left(\left\langle \boldsymbol{\theta}, \mathbf{x}^{(i)} \right\rangle + \boldsymbol{\theta}_0 \right) \geq 1 - \zeta^{(i)} \quad \forall\, i \in \{1, \dots, n\}, \\
\text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall\, i \in \{1, \dots, n\}.
\end{aligned}
$$

The parameter C controls the trade-off between the two conflicting objectives of maximizing the size of the margin and minimizing the frequency and size of margin violations. This is a convex optimization problem – particularly, a quadratic program with linear constraints and is known as the **primal** problem.

EXAMPLE 5: SUPPORT VECTOR MACHINES

We could directly solve the primal problem, but usually the SVM is solved in the **dual**:

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

EXAMPLE 5: SUPPORT VECTOR MACHINES

We could directly solve the primal problem, but usually the SVM is solved in the **dual**:

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

This is a convex quadratic program with box constraints plus one linear constraint.

EXAMPLE 5: SUPPORT VECTOR MACHINES

When applying the kernel trick to the dual (soft-margin) SVM problem by replacing $\left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle$ by kernels $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, we get the non-linear SVM:

$$
\begin{aligned}
\max_{\alpha \in \mathbb{R}^n} \quad & \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(\mathbf{y}) \boldsymbol{K} \operatorname{diag}(\mathbf{y}) \alpha \\
\text{s.t.} \quad & \alpha^\top \mathbf{y} = 0, \\
& 0 \leq \alpha \leq C,
\end{aligned}
$$

where $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$.

EXAMPLE 5: SUPPORT VECTOR MACHINES

When applying the kernel trick to the dual (soft-margin) SVM problem by replacing $\left\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \right\rangle$ by kernels $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, we get the non-linear SVM:

$$
\begin{aligned}
\max_{\alpha \in \mathbb{R}^n} \quad & \mathbf{1}^\top \alpha - \frac{1}{2} \alpha^\top \operatorname{diag}(\mathbf{y}) \boldsymbol{K} \operatorname{diag}(\mathbf{y}) \alpha \\
\text{s.t.} \quad & \alpha^\top \mathbf{y} = 0, \\
& 0 \leq \alpha \leq C,
\end{aligned}
$$

where $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. This is still a constrained convex quadratic problem, because $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is positive semi-definite.