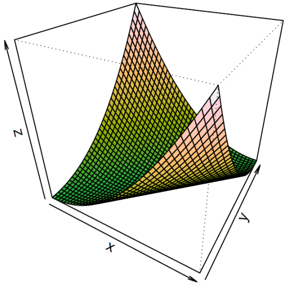
Optimization

Convex optimization problems



Learning goals

- TODO
- TODO

GENERAL DEFINITION

Consider the **optimization problem**

$$\min_{\mathbf{x} \in \mathcal{S} \subseteq \mathbb{R}^d} f(\mathbf{x})$$

with objective function

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

The problem is called **convex**

- f is a convex function
- $\mathcal{S}$ is a convex set.

How do constraints need to look like such that $\mathcal{S}$ is convex? Linear constraints are okay; ...

EXAMPLE 1: QUADRATIC FORMS

Discuss when a quadratic form corresponds to a convex optimization problem

EXAMPLE 2: SVM DUAL

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

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

EXAMPLE 2: SVM DUAL

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

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

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

EXAMPLE 3: RISK MIN. IN MACHINE LEARNING

- $\mathcal{D} = \left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)}\right)\right)$ denotes a dataset where $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ is a model, parameterized by $\boldsymbol{\theta}$ (e.g. linear model).
- Let $L\left(y, f(\mathbf{x})\right)$ be the point-wise loss function which measures the error of a prediction $f(\mathbf{x})$ compared to the true output y.
- We want to find the model which minimizes the **empirical risk**

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right).$$

Formulate without θ and then explain why we usually parameterize the hypothesis space.

EXAMPLE 3: RISK MIN. MACHINE LEARNING

Machine learning consists of three components:

$$\textbf{Machine Learning} = \underbrace{\textbf{Hypothesis Space + Risk}}_{\text{Formulating the optimization problem}} + \underbrace{\textbf{Optimization}}_{\text{Solving it}}$$

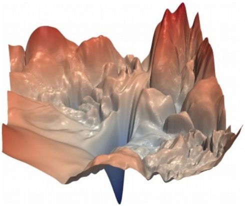
- **Hypothesis Space:** Define (and restrict!) what kind of model f can be learned from the data.
- **Risk:** Define the risk function $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ that quantifies how well a specific model performs on a given data set via a suitable loss function L.
- **Optimization:** Solve the resulting optimization problem through optimizing the risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ over the hypothesis space.

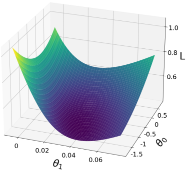
EXAMPLE 3: RISK MIN. MACHINE LEARNING

The (computational) complexity of the optimization problem

$$\arg\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

and hence the choice of the numerical optimization algorithm is influenced by the model structure and the choice of the loss function:, i.e., smoothness, convexity.





Loss landscapes of ML problems.

Left: ResNet-56, right: Logistic regression with cross-entropy loss

Source: `https://arxiv.org/pdf/1712.09913.pdf`

EXAMPLE 3A: NORMAL REGRESSION

EXAMPLE 3B: LOGISTIC REGRESSION

EXAMPLE 3C: NEURAL NETWORK

WHY IS CONVEXITY DESIRABLE?