Optimization for Machine Learning

Introduction

Clément W. Royer

M2 IASD - 2025/2026

September 18, 2025

Dauphine | PSL
UNIVERSITÉ PARIS

About this course

Online resources

- Course webpage:
  https://www.lamsade.dauphine.fr/~croyer/teachOML.html
- Teams channel (you should be registered).

Human resources (me)

- Faculty at Dauphine and teaching this course in IASD since 2019.
- Research area: Optimization.
- Since September: Deputy director of the CS track.

My email: clement.royer@lamsade.dauphine.fr

Course roadmap

New format in 2025-2026

- 8 sessions on Thursday afternoons, 2.15pm-5.30pm (16 before!)
- A typical session:
  - First half (∼90min): Lecture.
  - Hands-on part (∼60min): Exercises or Lab.
    No need to turn them after the class!
  - Last part (∼30min): Research-oriented topic.

Course roadmap

New format in 2025-2026

- 8 sessions on Thursday afternoons, 2.15pm-5.30pm (16 before!)
- A typical session:
  - First half ($\sim$90min): Lecture.
  - Hands-on part ($\sim$60min): Exercises or Lab.
    No need to turn them after the class!
  - Last part ($\sim$30min): Research-oriented topic.

Please...

- Make it to the class on time.
- Respect break duration.

Grading

Your grade

- 50% exam (on December 11, 2 hours).
- 50% homework (individual, due ??)
  $\rightarrow$ Should consist in an extended lab session.

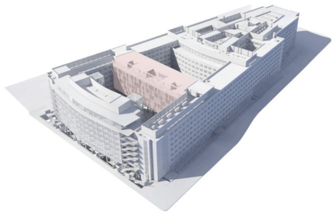
Previous exams (2019-2024)

- Partly irrelevant (48$\rightarrow$24 hours)!
- I'll give the relevant parts as additional exercises.

Optimization: *A field of study concerned with choosing the best decision out of a set of alternatives*

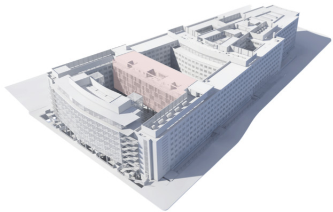
Optimization: *A field of study concerned with choosing the best decision out of a set of alternatives*

- 1980-2000 Rise as a computational mathematics field, successful model-driven applications
  Ex) Power systems, vaccine dispatch, allocating office space in Dauphine during renovation



Optimization: *A field of study concerned with choosing the best decision out of a set of alternatives*

- 1980-2000 Rise as a computational mathematics field, successful model-driven applications
  Ex) Power systems, vaccine dispatch, allocating office space in Dauphine during renovation



- 2000-2020 Shift to machine learning applications and data-driven problems.

Typical optimization problem for ML

- **Data**, e.g. $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$ for supervised learning.
- **Model class** $\mathcal{H} = \{\boldsymbol{h}(\cdot; \boldsymbol{w}), \boldsymbol{w} \in \mathbb{R}^d\}$
- **Loss function** ℓ.

Empirical risk minimization

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{w}), \boldsymbol{y}_i)}_{f(\boldsymbol{w})} + \lambda \Omega(\boldsymbol{w})$$

- f: Data-fitting term.
- Ω: Regularization term.

Example 1: Linear regression

$$\operatorname*{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 = \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i)^2.$$

- Simplest data analysis task possible.
- $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$.
- Nontrivial to solve when $n, d \gg 1$.

Example 1: Linear regression

$$\operatorname*{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2n}\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_2^2 = \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i)^2.$$

- Simplest data analysis task possible.
- $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$.
- Nontrivial to solve when $n, d \gg 1$.

Alternate losses for linear regression

- ℓ_1 loss: $\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_1 = \sum_{i=1}^n |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i|$
- Chebyshev loss: $\|\boldsymbol{X}\boldsymbol{w}-\boldsymbol{y}\|_\infty = \max_{1\le i\le n} |\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}-y_i|$.
- And more!

Example 2: SVM Classification

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \frac{1}{n}\sum_{i=1}^{n} \max\left\{1 - y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w}), 0\right\} + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2.$$

- $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$.
- Hinge loss $(h, y) \mapsto \max\{1 - y\,h, 0\}$.
- Regularization term with $\lambda \geq 0$.

Example 3: Binary classification using CNNs

$$\operatorname*{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i \mathrm{CNN}(\boldsymbol{x}_i;\boldsymbol{w})) + \lambda\|\boldsymbol{w}\|_1.$$

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i \mathrm{CNN}(\boldsymbol{x}_i;\boldsymbol{w})) + \lambda\|\boldsymbol{w}\|_1.$$

- Cross-entropy/Logistic loss.
- $\boldsymbol{x}_i \in \mathbb{R}^{d_0 \times d_0 \times c_0}$ (image), $y_i \in \{-1,1\}$ (class).

$$\operatorname*{minimize}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i \mathrm{CNN}(\boldsymbol{x}_i;\boldsymbol{w})) + \lambda\|\boldsymbol{w}\|_1.$$

- Cross-entropy/Logistic loss.
- $\boldsymbol{x}_i \in \mathbb{R}^{d_0\times d_0\times c_0}$ (image), $y_i \in \{-1,1\}$ (class).
- $\mathrm{CNN} : \boldsymbol{x}_i = \boldsymbol{z}^{(0)} \mapsto \boldsymbol{z}^{(1)} \mapsto \cdots \mapsto \boldsymbol{z}^{(L)}$, where

$$\boldsymbol{z}^{(l)}_{ijk} = \phi\left(\sum_{m,n,p} \boldsymbol{W}^{(l-1)}_{m,n,p,k}\boldsymbol{z}^{(l-1)}_{i-m,j-n,p} + \boldsymbol{b}^{(l-1)}_k\right).$$

$\phi(\boldsymbol{z}) = [\max(\boldsymbol{z}_i,0)]_i$ (ReLU activation).

Example 3: Binary classification using CNNs

$$\operatorname*{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \mathrm{CNN}(\boldsymbol{x}_i; \boldsymbol{w})) + \lambda \|\boldsymbol{w}\|_1.$$

- Cross-entropy/Logistic loss.
- $\boldsymbol{x}_i \in \mathbb{R}^{d_0 \times d_0 \times c_0}$ (image), $y_i \in \{-1, 1\}$ (class).
- $\mathrm{CNN} : \boldsymbol{x}_i = \boldsymbol{z}^{(0)} \mapsto \boldsymbol{z}^{(1)} \mapsto \cdots \mapsto \boldsymbol{z}^{(L)}$, where

$$\boldsymbol{z}_{ijk}^{(l)} = \phi\left(\sum_{m,n,p} \boldsymbol{W}_{m,n,p,k}^{(l-1)} \boldsymbol{z}_{i-m,j-n,p}^{(l-1)} + \boldsymbol{b}_k^{(l-1)} \right).$$

  $\phi(\boldsymbol{z}) = [\max(\boldsymbol{z}_i, 0)]_i$ (ReLU activation).
- $\boldsymbol{w}$ concatenates all $(\boldsymbol{W}^l, \boldsymbol{b}^l)_{l=0\dots(L-1)}$.

Takeaways

Generic form: $\text{minimize}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$.

Common traits

- Defined based on data.
- Use continuous functions (linear, ReLU, log/exp).

Distinctive aspects

- Model complexity/Number of parameters.
- Nonlinearity of operations.
- Regularization/Lack thereof.

Goals of the course

Understand...

- the **structure** of optimization problems arising in machine learning,
- the main optimization **algorithms** used in this setting,
- the challenges in **implementing** these methods at scale.

Goals of the course

Understand...

- the **structure** of optimization problems arising in machine learning,
- the main optimization **algorithms** used in this setting,
- the challenges in **implementing** these methods at scale.

Tentative outline

1. Basics of optimization
2. Automatic differentiation
3. Gradient descent
4. Beyond gradient descent
5. Stochastic gradient
6. Regularization
7. Second-order methods.