Lecture 1: Introduction and foundations
Large-scale optimization

Jens Sjölund   Sebastian Mair

August 31, 2023

Presentations

- Welcome!
- Jens Sjölund
  - Assistant professor at Systems and Control.
  - Background as industry researcher working in radiotherapy.
  - Leading a research group focusing on machine learning and optimization.
- Sebastian.
  - Postdoc at Systems and Control.
  - Research on data-efficient machine learning
- You? Raise of hands:
  - physics,
  - geophysics,
  - chemistry,
  - biology,
  - machine learning,
  - computer science,
  - MSc students,
  - other?

This Lecture

Recommended reading: Chapter 1 and Sections 2.1-2.3 in Wright and Recht 2022.

- Course contents
- The anatomy of an optimization problem (variables, objective, constraints, feasible set, etc.)
- Examples from data science/machine learning (logistic regression)
- Taylor's theorem and L-smoothness

Course contents

In this course we only consider *continuous* optimization problem, which have the following anatomy:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \le b_i,\ i = 1, \dots, m. \end{aligned} \tag{1}$$

- *Optimization variables* $x \in \mathbb{R}^n$ (also decision variables)
- *Objective function* $f_0(x) : \mathbb{R}^n \to \mathbb{R}$
- *Constraint functions* $f_i(x) : \mathbb{R}^n \to \mathbb{R}$
- An optimization problem without constraints is called *unconstrained.*
- Both the objective and constraint functions have *scalar* outputs!
- By convention, always minimize (maximize by flipping sign of objective function)
- The constraints define a *feasible set* $\Omega \subset \mathbb{R}^n$. An equivalent form of the problem is thus:

  $$\underset{x \in \Omega}{\text{minimize}} \quad f_0(x) \tag{2}$$

  where $\Omega = \{x \mid f_i(x) \le b_i,\ i = 1, \dots, m\}$.
- The first step in solving an optimization problem is to state it precisely!
- In applications, a problem can often be formulated in different ways.
- Can have a dramatic impact on the computational effort required to solve it!

Example: quadratic function

$$\underset{x}{\text{minimize}} \quad (x-1)(x-3). \tag{3}$$

- $x \in \mathbb{R}$
- $f_0(x) = (x-1)(x-3)$
- Unconstrained, $\Omega = \mathbb{R}$
- Draw
- How solved? Graphically or by setting gradient to zero.

$$\nabla f_0(x) = (x-3) + (x-1) = 2x - 4. \tag{4}$$

$$\nabla f_0(x) = 0 \implies x = 2. \tag{5}$$

- Is this all we need?

Example: logistic regression

- Classic method in machine learning and statistics.
- Setting: dataset $\mathcal{D} = \{(a_1, y_1), \dots, (a_m, y_m)\}$, where a_i are n-dimensional features and $y \in \{-1, +1\}$ are labels.
- A logistic regression models is trained by finding the parameters x that minimize

$$f_0(x) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + e^{-y_i \cdot a_i^\top x}\right). \tag{6}$$

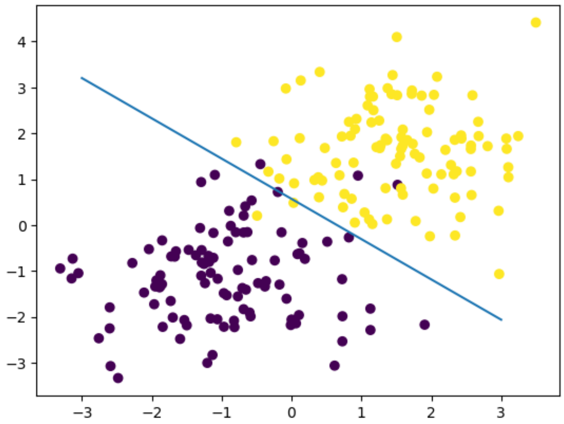

Figure 1: Binary classification data using $m = 200$ and $n = 2$.

- Partial derivatives

$$\frac{\partial f_0}{\partial x_j} = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 + e^{-y_i \cdot a_i^\top x}} \cdot (-y_i a_{ij}) e^{-y_i \cdot a_i^\top x} \tag{7}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{-y_i a_{ij}}{1 + e^{y_i \cdot a_i^\top x}}. \tag{8}$$

- This is a *nonlinear* function in x, so $\nabla f_0(x)$ is a system of nonlinear equations. Almost never has closed-form solution.
- Have to use a numerical, iterative, method.
- Suggestions from the audience?
- Numerical comparison: how many iterations of Adam vs. Newton?
- Conclusion: Newton's method is fantastic (also in theory)!

Example: large-scale problem

- DOROTHEA is a drug discovery dataset. Chemical compounds represented by structural molecular features must be classified as active (binding to thrombin) or inactive.
- Training data set consists of $m = 1909$ compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting. Each compound is described by a binary label and $n = 139,351$ binary features, which describe three-dimensional properties of the molecule.
- What happens when we try to use Newton's method to optimize a logistic regression classifier?
- Since the Hessian $\nabla^2 f_0(x) \in \mathbb{R}^{n \times n}$, evaluating and storing it takes up a lot of compute/memory (155 GB). Computing its inverse is even worse. Intractable!
- In comparison, computing and storing the gradient $\nabla f_0(x) \in \mathbb{R}^n$ is nothing (1.1 MB).
- Conclusion: we can at most use first-order methods (gradient information) to solve *large-scale* problems.
- Unless the problem has special *structure*, e.g., sparse, banded, circulant, low-rank.

Course structure

- 10 lectures, Wednesdays 13:15-15:00 in 101142.
- See recommended reading and problems for each lecture.
- Three blocks:
  - Unconstrained optimization (3 lectures)
  - Scalability (3 lectures)
  - Constrained optimization, additional theory (4 lectures)
- Examination: 3 homeworks, one for each block. Peer-review due one week after deadline.
- Optional project (3 credits): 2-page proposals due on lecture 9. Project report and presentation just before Christmas.

Foundations

What does it mean to solve an optimization problem?

Consider the unconstrained optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x). \tag{9}$$

See the example in figure 2.

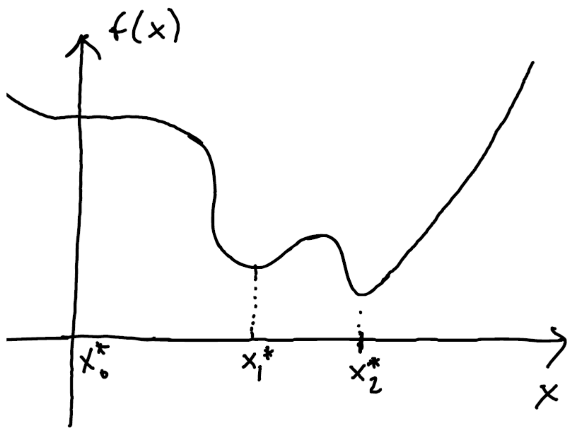

Figure 2: x_0^* and x_1^* are local minimizers while x_2^* is a global minimizer.

- Let $\mathcal{D} = \text{dom}(f)$ be the domain of f, i.e. where f is defined.
- $x^* \in \mathcal{D}$ *local minimizer* if $f(x) \geq f(x^*)$ for all x in a neighborhood of x^*.
- $x^* \in \mathcal{D}$ *global minimizer* if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.

In the case where $\mathcal{D} = \mathbb{R}^n$, we have:

- Necessary conditions for x^* to be a local minimum:
  - If f continuously differentiable, $\nabla f(x^*) = 0$ (first-order necessary condition).
  - If f twice continuously differentiable, also $\nabla^2 f(x^*) \succeq 0$ (second-order necessary condition).

- Second-order sufficient condition: $\nabla^2 f(x^*) \succ 0$.
- Conditions for global minimum require *convexity*—next lecture!
- Proofs require two tools: Taylor's theorem and Lipschitz smoothness.

Taylor's theorem

- In the univariate case, the fundamental theorem of calculus gives

$$f(x+p) - f(x) = \int_x^{x+p} f'(t)\,dt = \{t = x + \gamma p\} = \int_0^1 f'(x+\gamma p)p\,d\gamma \tag{10}$$

$$\iff f(x+p) = f(x) + \int_0^1 f'(x+\gamma p)p\,d\gamma \tag{11}$$

- Current value plus (mean value of the gradient times p).
- Multivariate case:

$$f(x+p) = f(x) + \int_0^1 \nabla f(x+\gamma p)^\top p\,d\gamma \tag{12}$$

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p)p\,d\gamma \tag{13}$$

  Integral forms of Taylor's theorem.

- If f is continuously differentiable, it follows from the mean-value theorem that there exists a $\gamma^* \in (0,1)$ such that

$$\int_0^1 \nabla f(x+\gamma p)^\top p\,d\gamma = \nabla f(x+\gamma^* p)^\top p \tag{14}$$

$$\implies f(x+p) = \nabla f(x) + \nabla f(x+\gamma^* p)^\top p. \tag{15}$$

  Mean-value form of Taylor's theorem.

L-smoothness

- A function f is L_0-*Lipschitz* if

$$|f(x) - f(y)| \leq L_0\|x-y\|. \tag{16}$$

  Click here for a visualization.

- A continuously differentiable f is L-*smooth* if its gradient is L-Lipschitz,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|. \tag{17}$$

- If f is twice differentiable then

$$-LI \preceq \nabla^2 f(x) \preceq LI \tag{18}$$

  for all x.

- L-smoothness is can be used to upper-bound a function f by a quadratic function,

  Lemma. *For an L-smooth function f, we have for any $x, y \in dom(f)$ that*

$$f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2. \tag{19}$$

Proof. First use integral form of Taylor's theorem with $p = y - x$, then Cauchy-Schwarz inequality, and finally L-smoothness,

$$
\begin{aligned}
& f(y) - f(x) - \nabla f(x)^\top (y - x) && (20)\\
&= \int_0^1 \nabla f(x + \gamma(y - x))^\top (y - x)\, d\gamma - \nabla f(x)^\top (y - x) && (21)\\
&= \int_0^1 \left(\nabla f(x + \gamma(y - x))^\top - \nabla f(x)\right)^\top (y - x)\, d\gamma && (22)\\
&\leq \int_0^1 \|\nabla f(x + \gamma(y - x))^\top - \nabla f(x)\| \|y - x\| d\gamma && (23)\\
&\leq \int_0^1 L\|x + \gamma(y - x) - x\| \|y - x\| d\gamma && (24)\\
&= L\|y - x\|^2 \int_0^1 \gamma\, d\gamma = \frac{L}{2}\|y - x\|^2. && (25)
\end{aligned}
$$

Rearranging completes the proof. □

Recap of inequalities

Cauchy-Schwarz inequality

$$|\langle x, y\rangle| \leq \|x\|\|y\|. \tag{26}$$

Triangle inequality

$$\|x + y\| \leq \|x\| + \|y\|. \tag{27}$$

Reverse triangle inequality

$$\|\|x\| - \|y\|\| \leq \|x - y\|. \tag{28}$$

References

Wright, Stephen J and Benjamin Recht (2022). *Optimization for data analysis.* Cambridge University Press.