Outline

1. Topics that I expect you to already know

2. Introduction to nonlinear unconstrained optimization
   - Smooth problem example
   - Structured non-smooth example

3. Summary
## Calculus
- derivatives
- gradients
- Jacobians
- Hessians
- Taylor’s expansion

## Real Analysis
- sequences (subsequences, boundedness, accumulation points, etc.)
- continuity and limits

## Linear Algebra
- vectors and vector norms
- matrices and matrix norms
- matrix properties, e.g., symmetric, positive (semi) definite, (non)singular, etc.
- determinants, eigenvalues, and eigenvectors
- matrix factorizations

### The basic problem

<table>
<thead>
<tr>
<th>minimize $f(x)$</th>
<th>$\text{subject to}$ $x \in \mathbb{R}^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>objective function $f : \mathbb{R}^n \to \mathbb{R}$</td>
<td></td>
</tr>
<tr>
<td>may maximize $f$ by minimizing the function $\bar{f}(x) := -f(x)$</td>
<td></td>
</tr>
</tbody>
</table>

### Definition (global minimizer)

The vector $x^*$ is a global minimizer if $f(x^*) \leq f(x) \text{ for all } x \in \mathbb{R}^n$.

### Definition (local minimizer)

The vector $x^*$ is a local minimizer if $f(x^*) \leq f(x) \text{ for all } x \text{ satisfying } \|x - x^*\| \leq \varepsilon \text{ for some } \varepsilon > 0$.

### Definition (strict local minimizer)

The vector $x^*$ is a strict local minimizer if $f(x^*) < f(x) \text{ for all } x \neq x^* \text{ satisfying } \|x - x^*\|_2 \leq \varepsilon \text{ for some } \varepsilon > 0$.

Notes

---
The basic problem

\[
\text{minimize } f(x)
\]

Notation

- gradient \(g(x) := \nabla f(x) \in \mathbb{R}^n\)
- Hessian \(H(x) := \nabla^2 f(x) \in \mathbb{R}^{n \times n}\)

- we will always assume that \(f\) is a continuous function
- we will study when
  - \(f\) is once or twice continuously differentiable (smooth optimization)
  - \(f\) is non-differentiable but with structure (structured non-smooth optimization)
  - \(n\) is small, medium, and large
- we will not study
  - derivatives of \(f\) are too expensive or unavailable (derivative free optimization)
  - integer programming, i.e., optimization where (some) variables are required to take integer values (Combinatorial Optimization EN.553.766)
  - computing a global minimizer (Introduction to Convexity EN.553.465/665, Convex Optimization EN.553.765, Stochastic Search and Optimization EN.553.763)

Data fitting example

January 1801: asteroid Ceres is discovered, but in Autumn 1801 it “disappeared”. Gauss considers an elliptic orbit instead of a circular orbit

- circular orbit \(x^2 + y^2 = r^2\) for some \(r > 0\)
- elliptic (conic section) orbit \(\alpha x^2 + \beta y^2 + \gamma xy = 1\) for some \(\alpha, \beta,\) and \(\gamma\)

How did he do it?

- used a collection of \(N\) previous location measurements \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)
- found the “best” ellipse by computing
  \[
  (\alpha^*, \beta^*, \gamma^*) = \arg\min_{\alpha, \beta, \gamma} \sum_{i=1}^{N} (\alpha x_i^2 + \beta y_i^2 + \gamma x_i y_i - 1)^2
  \]
- looked for Ceres along the ellipse defined by
  \[
  \alpha^* x^2 + \beta^* y^2 + \gamma^* xy = 1
  \]
- the objective function \(f(\alpha, \beta, \gamma) = \sum_{i=1}^{N} (\alpha x_i^2 + \beta y_i^2 + \gamma x_i y_i - 1)^2\) is nonlinear and twice continuously differentiable
Speech recognition (multi-class regression)

- Number of classes $N_c \approx 100$ (basic units of sound)
- Number of features $N_f \approx 10$ thousand (coefficients in the mathematical representation of a digital sample of sound)
- Number of parameters $\approx 1$ million ($\#$ classes $\times$ $\#$ features)
- Number of data points $N_d \approx 10$ billion and growing (size of data)
- Compute $w^*$ as solution to
  \[
  \minimize_w f(w) + \lambda \|w\|_1 \quad (\lambda > 0 \text{ is a sparsity parameter})
  \]
  where
  \[
  f(w) := -\sum_{i=1}^{N_d} \log \left( \frac{\exp(w_i^T x_i)}{\sum_{j=1}^{N_c} \exp(w_j^T x_i)} \right)
  \]
- Predicted probability of new input $\hat{x}$ being in class $k$ is
  \[
  p(y = k | x = \hat{x}) = \frac{\exp(w_k^T \hat{x})}{\sum_{j=1}^{N_c} \exp(w_j^T \hat{x})}
  \]
- Major challenges
  - $f(w) + \lambda \|w\|_1$ is nonlinear
  - $\|w\|_1$ is non-smooth when $w_i = 0$ for some $i$ (structured non-smooth)
  - $\nabla f(w)$ is very expensive! Must sum up 10 billion gradients

Unconstrained optimization problems may

- be convex or nonconvex, but typically nonlinear.
- have an objective function that is twice continuously differentiable, once continuously differentiable, structurally non-smooth, non-smooth
- contain continuous and/or discrete variables
- vary in size
  - small scale $\approx 1 - 100$ variables
  - medium scale $\approx 10^3$ variables
  - large scale $\approx 10^4 - 10^5$ variables
  - very large scale $\geq 10^6$ variables
  - infinite dimensional

We may be interested in

- a local solution or global solution
- the minimum value of the objective function and/or the minimizer
- finding multiple distinct minimizers
- the lowest value of the objective given time constraints or limits on the number of allowed evaluations of the objective function