Outline

Mathematical Background
- Lecture 4: Intro to Optimization
- Lecture 5: Gradient Descent

Lecture 4: Intro to Optimization
- Unconstrained Optimization
  - First Order Optimality
  - Second Order Optimality
- Convexity
  - What is convexity?
  - Convex optimization
- Constrained Optimization
  - Lagrangian
  - Examples
Unconstrained Optimization

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{s.t.} & \quad x \in \mathcal{X}
\end{align*}
\]

- \( x^* \in \mathcal{X} \) is a **global minimizer** if
  - \( f(x^*) \leq f(x) \) for any \( x \in \mathcal{X} \)
- \( x^* \in \mathcal{X} \) is a **local minimizer** if
  - \( f(x^*) \leq f(x) \), for any \( x \) in a neighborhood \( \mathcal{B}_\delta(x^*) \)
  - \( \mathcal{B}_\delta(x^*) = \{ x \mid \| x - x^* \|_2 \leq \delta \} \)
Uniqueness of Global Minimizer

If $\mathbf{x}^*$ is global minimizer, then

- Objective value $f(\mathbf{x}^*)$ is unique
- Solution $\mathbf{x}^*$ is not necessarily unique

Therefore:

- Suppose $f(\mathbf{x}) = g(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$ for some convex $g$.
- “minimize $f(\mathbf{x})$” has a global optimal $f(\mathbf{x}^*)$.
- But there could be multiple $\mathbf{x}^*$’s.
- Some $\mathbf{x}^*$ maybe better, but not in the sense of $f(\mathbf{x})$. 

[Diagram showing multiple minimizers $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$]
First and Second Order Optimality

\[ \nabla f(x^*) = 0 \quad \text{and} \quad \nabla^2 f(x^*) \succeq 0 \]

First order condition

Second order condition

**Necessary Condition:**
If \( x^* \) is a global (or local) minimizer, then
- \( \nabla f(x^*) = 0 \).
- \( \nabla^2 f(x^*) \succeq 0 \).

**Sufficient Condition:**
If \( x^* \) satisfies
- \( \nabla f(x^*) = 0 \).
- \( \nabla^2 f(x^*) \succ 0 \).
then \( x^* \) is a global (or local) minimizer.
Why? First Order

- Why is $\nabla f(x^*) = 0$ necessary?
- Suppose $x^*$ is the minimizer.
- Pick any direction $d$, and any step size $\epsilon$. Then

$$f(x^* + \epsilon d) = f(x^*) + \epsilon \nabla f(x^*)^T d + \mathcal{O}(\epsilon^2).$$

- Rearranging the terms yields

$$\lim_{\epsilon \to 0} \left\{ \frac{f(x^* + \epsilon d) - f(x^*)}{\epsilon} \right\} = \nabla f(x^*)^T d.$$

- So $\nabla f(x^*)^T d \geq 0$ for all $d$. True only when $\nabla f(x^*) = 0$. 
First Order Condition Illustrated

If $x^*$ is not optimal, then
\[ \nabla f(x^*)^T d \leq 0, \text{ for some } d \]

If $x^*$ is optimal, then
\[ \nabla f(x^*)^T d = 0, \text{ for all } d \]
Why? Second Order

Do third order approximation:

\[
f(x^* + \epsilon d) = f(x^*) + \epsilon \nabla f(x^*)^T d + \frac{\epsilon^2}{2} d^T \nabla^2 f(x^*) d + \frac{\epsilon^3}{6} O(\|d\|^3)
\]

Therefore,

\[
\frac{1}{\epsilon^2} \left[ f(x^* + \epsilon d) - f(x^*) \right] = \frac{1}{2} d^T \nabla^2 f(x^*) d + \frac{\epsilon}{6} O(\|d\|^3)
\]

\[
\lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \left[ f(x^* + \epsilon d) - f(x^*) \right] = \frac{1}{2} d^T \nabla^2 f(x^*) d + \lim_{\epsilon \to 0} \frac{\epsilon}{6} O(\|d\|^3)
\]

Hence,

\[
\frac{1}{2} d^T \nabla^2 f(x^*) d \geq 0, \quad \forall d.
\]

\Rightarrow \text{positive semi-definite!}
Second Order Condition Illustrated

If positive definite, then minimum point.

Saddle point (positive semi-definite)
Mathematical Background

- Lecture 4: Intro to Optimization
- Lecture 5: Gradient Descent

Lecture 4: Intro to Optimization

- Unconstrained Optimization
  - First Order Optimality
  - Second Order Optimality
- Convexity
  - What is convexity?
  - Convex optimization
- Constrained Optimization
  - Lagrangian
  - Examples
Most Optimization Problems are Not Easy

Minimize the log-sum-exp function:

\[ f(x) = \log \left( \sum_{i=1}^{m} \exp(a_i^T x + b_i) \right) \]

- Gradient is (exercise)

\[ \nabla f(x^*) = \frac{1}{\sum_{j=1}^{m} \exp(a_j^T x^* + b_j)} \sum_{i=1}^{m} \exp(a_i^T x^* + b_i) a_i. \]

- Non-linear equation. No closed-form solution.
- Need iterative algorithms, e.g., gradient descent.
- Or off-the-shelf optimization solver, e.g., CVX.
Disciplined optimization: It translates the problem for you.
Developed by S. Boyd and colleagues (Stanford).
E.g., Minimize \( f(x) = \log \left( \sum_{i=1}^{n} \exp(a_i^T x + b_i) \right) + \lambda \|x\|^2 \).

```python
import cvxpy as cp
import numpy as np

n = 100
d = 3
A = np.random.randn(n, d)
b = np.random.randn(n)
lambda_ = 0.1

x = cp.Variable(d)
objective = cp.Minimize(cp.log_sum_exp(A*x - b) + lambda_*cp.sum_squares(x))
constraints = []
prob = cp.Problem(objective, constraints)

optimal_objective_value = prob.solve()
print(optimal_objective_value)
print(x.value)
```
**Convex Function**

**Definition**

Let $x \in \mathcal{X}$ and $y \in \mathcal{X}$. Let $0 \leq \lambda \leq 1$. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is **convex** over $\mathcal{X}$ if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

The function is called strictly convex if “$\leq$” is replaced by “$<$”.

![Convex Function Diagram](image)
Example: Which one is convex?
Verifying Convexity

Any of the following conditions is **necessary** and **sufficient** for convexity:

1. By definition:
   \[
   f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}).
   \]
   - Function value is lower than the line.

2. First Order Convexity:
   \[
   f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.
   \]
   - Tangent line is always lower than the function

3. Second Order Convexity: \( f \) is convex over \( \mathcal{X} \) if and only if
   \[
   \nabla^2 f(\mathbf{x}) \succeq 0 \quad \forall \mathbf{x} \in \mathcal{X}.
   \]
   - Curvature is positive.
Tangent Line Condition Illustrated

\[ f(x) + \nabla f(x)^T (y-x) \]
Outline

Mathematical Background

- Lecture 4: Intro to Optimization
- Lecture 5: Gradient Descent

Lecture 4: Intro to Optimization

- Unconstrained Optimization
  - First Order Optimality
  - Second Order Optimality
- Convexity
  - What is convexity?
  - Convex optimization
- Constrained Optimization
  - Lagrangian
  - Examples
Constrained Optimization

**Equality** Constrained Optimization:

\[
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad h_j(x) = 0, \quad j = 1, \ldots, k.
\end{align*}
\]

Requires a function: Lagrangian function

\[
\mathcal{L}(x, \nu) \overset{\text{def}}{=} f(x) - \sum_{j=1}^{k} \nu_j h_j(x).
\]

\( \nu = [\nu_1, \ldots, \nu_k] \): **Lagrange multipliers** or the **dual variables**.

Solution \((x^*, \nu^*)\) satisfies

\[
\begin{align*}
\nabla_x \mathcal{L}(x^*, \nu^*) &= 0, \\
\nabla_\nu \mathcal{L}(x^*, \nu^*) &= 0.
\end{align*}
\]
Example: Illustrating Lagrangian

- Consider the problem

\[
\begin{align*}
\text{minimize} & \quad x_1 + x_2 \\
\text{subject to} & \quad x_1^2 + x_2^2 = 2.
\end{align*}
\]

- Minimizer is \( x = (-1, -1) \).
Example: Illustrating Lagrangian
Example: Illustrating Lagrangian

\[ \nabla f(x^*) = \lambda \nabla h(x^*) \]

\[ \Rightarrow \nabla L(x^*, \lambda) = \nabla f(x^*) - \lambda \nabla h(x^*) = 0 \]
Example: $\ell_2$-minimization with constraint

\[
\text{minimize } \frac{1}{2}\|x - x_0\|^2, \quad \text{subject to } Ax = y.
\]

The Lagrangian function of the problem is

\[
\mathcal{L}(x, \nu) = \frac{1}{2}\|x - x_0\|^2 - \nu^T(Ax - y).
\]

The first order optimality condition requires

\[
\nabla_x \mathcal{L}(x, \nu) = (x - x_0) - A^T\nu = 0
\]
\[
\nabla_{\nu} \mathcal{L}(x, \nu) = Ax - y = 0.
\]

Multiply the first equation by $A$ on both sides:

\[
A(x - x_0) - AA^T\nu = 0
\]
\[
\Rightarrow \quad \underbrace{Ax - Ax_0}_{y} = AA^T\nu
\]
\[
\Rightarrow \quad y - Ax_0 = AA^T\nu
\]
\[
\Rightarrow \quad (AA^T)^{-1}(y - Ax_0) = \nu
\]
Example: $\ell_2$-minimization with constraint

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|x - x_0\|^2, \\
\text{subject to} & \quad Ax = y.
\end{align*}
\]

The first order optimality condition requires

\[
\begin{align*}
\nabla_x \mathcal{L}(x, \nu) &= (x - x_0) - A^T \nu = 0 \\
\nabla_\nu \mathcal{L}(x, \nu) &= Ax - y = 0.
\end{align*}
\]

We just showed: $\nu = (AA^T)^{-1} (y - A x_0)$. Substituting this result into the first order optimality yields

\[
x = x_0 + A^T \nu \\
= x_0 + A^T (AA^T)^{-1} (y - A x_0)
\]

Therefore, the solution is $x = x_0 + A^T (AA^T)^{-1} (y - A x_0)$. 
Special Case

\[ \begin{align*}
\text{minimize} & \quad \frac{1}{2} ||x - x_0||^2, \\
\text{subject to} & \quad Ax = y.
\end{align*} \]

Special case: When \( Ax = y \) is simplified to \( w^T x = 0 \).

- \( w^T x = 0 \) is a line.
- Find a point \( x \) on the line that is closest to \( x_0 \).
- Solution is

\[ x = x_0 + w(w^T w)^{-1}(0 - w^T x_0) \]

\[ = x_0 - \left( \frac{w^T x_0}{||w||^2} \right)^T w. \]
In practice ...

- Use CVX to solve problem
- Here is a MATLAB code
- Exercise: Turn it into Python.

```matlab
% MATLAB code: Use CVX to solve min ||x - x0||, s.t. Ax = y
m = 3; n = 2*m;
A = randn(m,n); xstar = randn(n,1);
y = A*xstar;
x0 = randn(n,1);

cvx_begin
    variable x(n)
    minimize( norm(x-x0) )
    subject to
        A*x == y;
    cvx_end

% you may compare with the solution x0 + A'*inv(A*A')*(y-A*x0).
```
Reading List

Unconstrained Optimality Conditions

- Nocedal-Wright, Numerical Optimization. (Chapter 2.1)
- Boyd-Vandenberghe, Convex Optimization. (Chapter 9.1)

Convexity

- Nocedal-Wright, Numerical Optimization. (Chapter 1)
- Boyd-Vandenberghe, Convex Optimization. (Chapter 2 and 3)
- CMU, Convex Optimization (Lecture 2 and 4)
  
  https://www.stat.cmu.edu/~ryantibs/convexopt-F18/
- Stanford CS 229 (Tutorial)
  
  http://cs229.stanford.edu/section/cs229-cvxopt.pdf
- UCSD ECE 273 (Tutorial)
  

Constrained Optimization

- Nocedal-Wright, Numerical Optimization. (Chapter 12.1)