Lecture 29

Understanding

Gradient Descent.

* Midterm already graded

* Drop date is 11/18 (Wednesday)
**CONVEX FUNCTIONS**

A function \( f \) is convex if

\[
    f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y)
\]

"Function below the chord"

**Characterizations** (all equivalent, under suitable conditions)

- "First-order": for all \( x, y \)

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

Linear approximation at \( x \)

- "Second-order": for all \( x \)

\[
    \frac{\partial^2 f(x)}{\partial x^2}
\]

Hessian

\( \text{is a PSD matrix} \)

"Quadratic approximation has positive curvature"
WHY IS CONVEXITY SO AWESOME?

• LOCAL MINIMA ARE ALSO GLOBAL

• SIMPLE CHARACTERIZATION OF OPTIMALITY

\[ \nabla f(x^*) = 0 \]

"Gradient must vanish"
"Zero gradient at the optimal point"

• A SIMPLE ALGORITHM ("gradient method")

\[ x_{k+1} = x_k - \gamma \nabla f(x_k) \]
**Example:** Square roots.

**How to compute the square root of a number?**

E.g., \( \sqrt{\beta} \)? \( (\beta > 0) \)

→ Here's an approach, using gradient descent

(*Disclaimer: Better methods exist!*)

**Step 1:** Given \( \beta \), find a convex function with \( \sqrt{\beta} \) as the minimum (or minimizer)

E.g.: \[ F(x) = x + \frac{\beta}{x} \quad (x > 0) \]

Why? \[ \frac{dF(x)}{dx} = 1 - \frac{\beta}{x^2} = 0 \quad \Rightarrow \quad x = \pm \sqrt{\beta} \]
**Step 2:** Let's write gradient descent!

\[ x_{k+1} = x_k - \gamma \nabla f(x_k) = x_k - \gamma \left( 1 - \frac{\beta}{x_k^2} \right) \]

Take, e.g. \( \beta = 23 \) and \( \gamma = 0.1 \) (from Julia)

\[ \sqrt{23} \]

![Graph showing convergence to \( \sqrt{23} \approx 4.7958 \)]
Does every stepsize work?

E.g., in the univariate case

\[ F(x) = \frac{1}{2} x^2 \]
\[ \nabla F(x) = x \]

\[ x_{k+1} = x_k - \gamma \nabla F(x_k) \]
\[ x_{k+1} = x_k - \gamma x_k = (1 - \gamma) x_k \]

\[ |1 - \gamma| < 1 \]
What happens for quadratic functions?

\[ F(x) = \frac{1}{2} x^T A x \]

(without loss of generality)

\[ \nabla F(x) = A x \]

\[ A \text{ is pd} \]

(so \( f \) is convex).

Eigenvalues

\[ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0 \]

\[ x_{k+1} = x_k - \gamma \nabla F(x_k) = x_k - \gamma A x_k \]

\[ x_{k+1} = (I - \gamma A) x_k \]

Q: When does this converge?

\[ x_{k+1} = M x_k \]

\[ M = I - \gamma A \]

\[ \text{convergence} \]

\[ |\lambda_i(M)| < 1 \]
Q: What is the "best" stepsize?

\[ A = \begin{pmatrix} \lambda_2 & 0 \\ 0 & \lambda_n \end{pmatrix} \quad I - \sigma A = \begin{pmatrix} 1 - \sigma \lambda_1 & 0 \\ 0 & 1 - \sigma \lambda_n \end{pmatrix} \]

\[ |\text{eig}(I - \sigma A)| = |1 - \sigma \lambda_1| \]

\[ \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} \]

\[ \eta \]

\[ \gamma \]

\[ \frac{2}{\lambda_1} \]

\[ \Rightarrow \text{ if } 0 < \sigma < \frac{2}{\lambda_1} \]

The gradient method converges.
If we measure convergence rate by

\[ R = \| I - \sigma A \| = \max_i \left| 1 - \sigma \lambda_i \right|, \]

"best" stepsize is \( \frac{2}{\lambda_1 + \lambda_n} \).

This choice gives a rate

\[ R \leq \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n} = \frac{Q - 1}{Q + 1} \]

where \( Q = \frac{\lambda_1}{\lambda_n} \) is the condition number of the matrix \( A \).
Behavior in 2D

"Zigzag behavior"

(similar in higher dims)

\[ A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \]

\[ Q = 1 \]

\[ f(x,y) = x^2 + y^2 \]

\[ A = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix} \]

\[ Q = 10 \]

\[ f(x,y) = x^2 + 10y^2 \]

\[ Q = 100 \]
What happens in the general convex case?

- Does it converge?
- For what values of stepsize?

\[ x_{k+1} = x_k - \gamma \nabla f(x_k) \]

For simplicity

- Assume minimum value is zero:

\[ f(x^*) = 0 \]

(easy, by shifting the function)

\[ f(x) \rightarrow f(x) - f(x^*) \]
Q: **How to quantify convergence?**

Diagram:

- Function value $F(x(k))$ decreasing as iteration number $k$ increases.
- Upper bound.

Q: **What properties of $F(x)$ should it depend on?**

Recall:

- $F(x)$ convex $\iff$ Hessian PSD (non-negative eigenvalues).

Eigenvalues of Hessian $\frac{\partial^2 F}{\partial x^2}$ may depend on the point $x$. 
CASE I: ("Smooth") or ("Lipschitz gradient")

$\lambda_i$ eigenvalues of Hessian

$0 \leq \lambda_i \leq L$

(at all points)

Choose Step size

$\gamma = \frac{1}{4L}$

Then:

$f(x_k) \leq \frac{L}{2k} \|x_0 - x^*\|^2$

$\frac{4k}{L} \rightarrow 0$ as $k \rightarrow \infty$

$\Rightarrow$ GRADIENT DESCENT CONVERGES.
CASE II: ("smooth and strongly convex")

- Stronger assumption:

\[ m \leq \lambda_i \leq L \] (at all points)

Upper and lower bounds on curvature

Choose stepsize \( \gamma = \frac{2}{L + m} \)

Then:

\[ f(x_k) \leq \frac{L}{2} \left( \frac{Q - 1}{Q + 1} \right)^{2k} ||x_0 - x^*||^2 \]

where \( Q = \frac{L}{m} \) "condition number" \( (Q > 1) \)

\[ \Rightarrow \text{CONVERGES MUCH FASTER!} \]
**Intuition / Summary**

- **Stepsizes**
  - Too small: slow (but converges)
  - Too large: oscillate, or diverge

**Function Properties**

1. **Smooth**
   - \( \lambda \geq \frac{1}{3} \)
   - "Slow"
   - \( O\left( \frac{1}{n} \right) \)

2. **Smooth + Strongly Convex**
   - \( m \leq \lambda \leq \frac{1}{\kappa} \)
   - "Fast" or "Linear"
   - \( O\left( \kappa^k \right) \)
   - (for some \( \kappa < 1 \))

- No explicit dependence on dimension

- **Condition number**
  - \( Q = \frac{\lambda_1}{\lambda_n} \)
  - \( \frac{1}{m} \)

Plays a big role (want \( Q \) small)