COMPSCI 514: Algorithms for Data Science

Cameron Musco

University of Massachusetts Amherst. Fall 2022.

Lecture 24

Logistics

- Problem Set 5 is posted. It is due 12/12 (last day of classes). It is optional and can give up to a 5% boost on your final grade.
- The final will be on 12/14 in this room, 10:30am-12:30pm.
- It is not cumulative and will follow a similar format to the midterm.
- Final review sheet is posted under the 'Schedule Tab' and practice exams posted on Moodle.
- My office hours today will end early at 3pm. I'll hold additional office hours in LGRC A215 on Friday 2:30-4:30pm and Monday 10am-12pm.

Summary

Last Class:

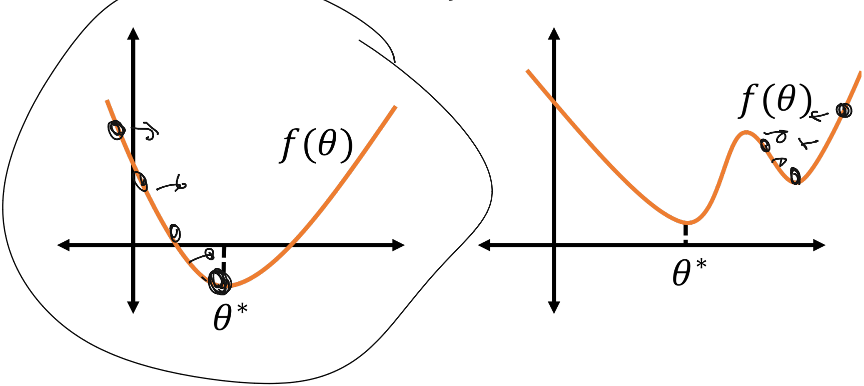
- Multivariable calculus review and gradient computation.
- Introduction to gradient descent. Motivation as a greedy algorithm.

This Class:

- Conditions under which we will analyze gradient descent: convexity and Lipschitzness.
- Analysis of gradient descent for Lipschitz, convex functions.
- Extension to projected gradient descent for constrained optimization.

When Does Gradient Descent Work?

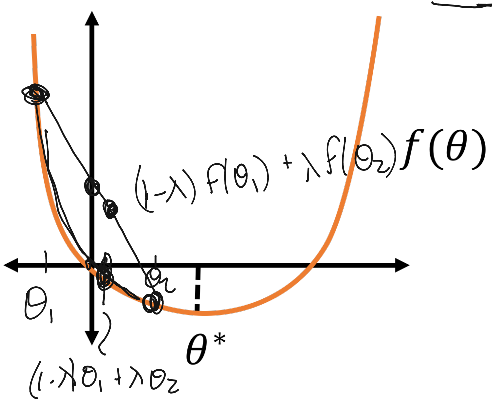
$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$

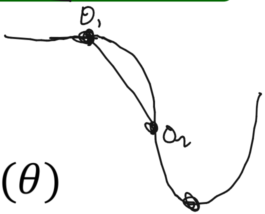


Gradient Descent Update: $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$

Definition – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1-\lambda)\cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda)\cdot\vec{\theta}_1 + \lambda\cdot\vec{\theta}_2\right)$$





Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

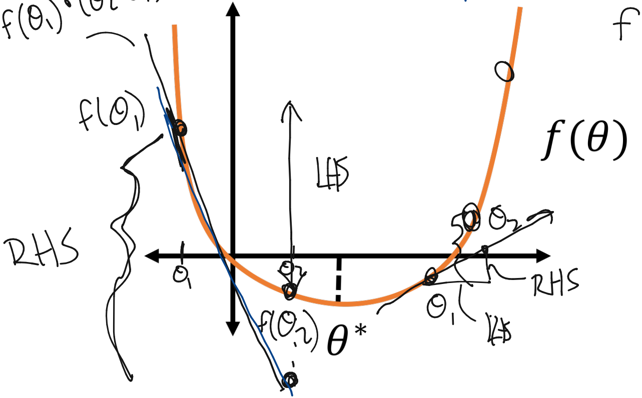
$$\underbrace{f(\vec{\theta}_2) - f(\vec{\theta}_1)}_{\mathbb{R}} \geq \underbrace{\vec{\nabla} f(\vec{\theta}_1)^T}_{\mathbb{R}} \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$

$\langle x, y \rangle = x^T y$

$\langle \nabla f(\theta_1), \theta_2 - \theta_1 \rangle$

$f'(\theta_1) \cdot (\theta_2 - \theta_1)$

$\nabla f(\theta_1) \cdot (\theta_2 - \theta_1)$



Conditions for Gradient Descent Convergence

Convex Functions: After sufficient iterations, if the step size η is chosen appropriately, gradient descent will converge to a approximate minimizer $\hat{\theta}$ with:

$$\underline{f(\hat{\theta})} \leq \underline{f(\vec{\theta}_*) + \epsilon} = \min_{\vec{\theta}} f(\vec{\theta}) + \epsilon.$$

Examples: least squares regression, logistic regression, sparse regression (lasso), regularized regression, SVMs,...

Conditions for Gradient Descent Convergence

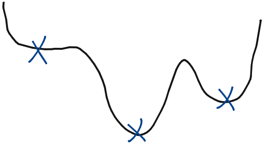
Convex Functions: After sufficient iterations, if the step size η is chosen appropriately, gradient descent will converge to a approximate minimizer $\hat{\theta}$ with:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon = \min_{\vec{\theta}} f(\vec{\theta}) + \epsilon.$$

Examples: least squares regression, logistic regression, sparse regression (lasso), regularized regression, SVMs,...

Non-Convex Functions: After sufficient iterations, gradient descent will converge to a approximate stationary point $\hat{\theta}$ with:

$$\|\nabla f(\hat{\theta})\|_2 \leq \epsilon.$$



Conditions for Gradient Descent Convergence

Convex Functions: After sufficient iterations, if the step size η is chosen appropriately, gradient descent will converge to a approximate minimizer $\hat{\theta}$ with:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon = \min_{\vec{\theta}} f(\vec{\theta}) + \epsilon.$$

Examples: least squares regression, logistic regression, sparse regression (lasso), regularized regression, SVMs,...

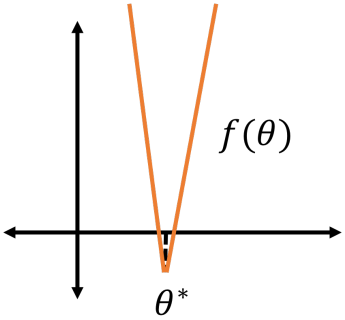
Non-Convex Functions: After sufficient iterations, gradient descent will converge to a approximate stationary point $\hat{\theta}$ with:

$$\|\nabla f(\hat{\theta})\|_2 \leq \epsilon.$$

Examples: neural networks, clustering, mixture models.

Lipschitz Functions

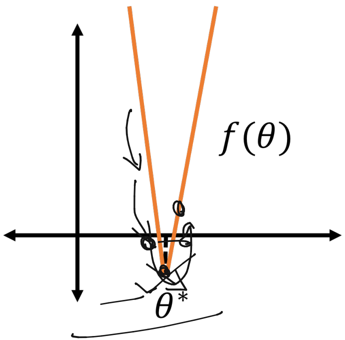
$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$



Gradient Descent Update:
$\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$

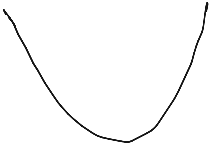
Lipschitz Functions

$\theta \in \mathbb{R} \quad \nabla f(\theta) \in \mathbb{R}$



Gradient Descent Update:

$\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$



Need to assume that the function is Lipschitz (size of gradient is bounded): There is some G s.t.:

"mean value theorem"

$$\forall \vec{\theta}: \quad \|\vec{\nabla} f(\vec{\theta})\|_2 \leq G \Leftrightarrow \forall \vec{\theta}_1, \vec{\theta}_2: \quad \underbrace{|f(\vec{\theta}_1) - f(\vec{\theta}_2)| \leq G \cdot \|\vec{\theta}_1 - \vec{\theta}_2\|_2}$$

$|f'(\theta)| \leq G$

Well-Behaved Functions

Definition – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$(1-\lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$

Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

$$f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$

Definition – Lipschitz Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is G-Lipschitz if $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq G$ for all $\vec{\theta}$.

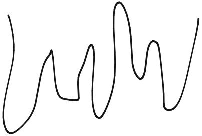
GD Analysis – Convex Functions

$\vec{\theta}_* = \arg\min_{\vec{\theta} \in \mathbb{R}^d} f(\vec{\theta})$

Assume that:

- f is convex.
- f is G-Lipschitz.
- $\|\vec{\theta}_1 - \vec{\theta}_*\|_2 \leq R$ where $\vec{\theta}_1$ is the initialization point.

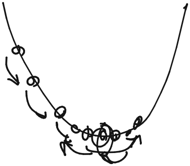
$\vec{\theta}_1 = 0$



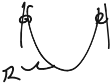
Gradient Descent

$+\varepsilon$

- Choose some initialization $\vec{\theta}_1$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t-1$
  - $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \vec{\nabla} f(\vec{\theta}_i)$
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_1, \ldots, \vec{\theta}_t} f(\vec{\theta}_i)$.



$f(x) = x^2$



GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

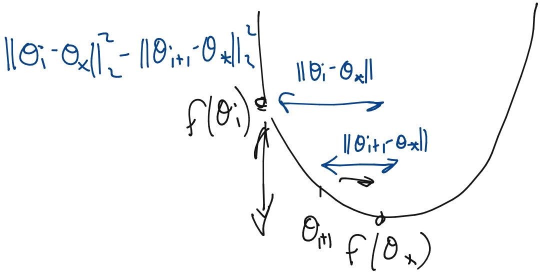
GD Analysis Proof

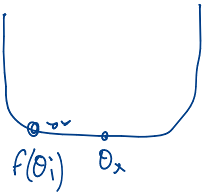
Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

could be negative

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Visually:





Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$\|a-b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 - 2\langle a,b\rangle$

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \theta_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$. Formally:

$\theta_i - \theta_*$

$\eta \nabla f(\theta_i)$

$$\|\theta_{i+1} - \theta_*\|_2^2 = \|\theta_i - \eta \nabla f(\theta_i) - \theta_*\|_2^2$$

$$= \|\theta_i - \theta_*\|_2^2 + \|\eta \nabla f(\theta_i)\|_2^2 - 2\eta \nabla f(\theta_i)^T(\theta_i - \theta_*)$$

$$\nabla f(\theta_i)^T(\theta_i - \theta_*) = \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2 + \underbrace{\|\eta \nabla f(\theta_i)\|_2^2}_{\leq \eta^2 G^2}}{2\eta}$$

$+ \frac{\eta G^2}{2}$

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\vec{\nabla} f(\vec{\theta}_i)^T (\vec{\theta}_i - \vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

GD Analysis Proof

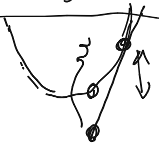
Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Step 1.1: $\vec{\nabla} f(\vec{\theta}_i)^T(\vec{\theta}_i - \vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$ $\implies$ Step 1 by convexity.

$$\boxed{f(\theta_i) - f(\theta_*) \leq \nabla f(\theta_i)^T (\theta_i - \theta_*)}$$



Step 1.1 → Step 1 by convexity

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \implies$

Step 2: $\underline{\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*)} \leq \underline{\frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}}$.

$\hat{\theta} = \arg\min f(\theta_i)$

$f(\hat{\theta}) \leq \frac{1}{t}\sum_{i=1}^{t} f(\theta_i)$

if $\frac{1}{t}\sum f(\theta_i) - f(\theta_*) \leq \epsilon$

$\Rightarrow f(\hat{\theta}) - f(\theta_*) \leq \epsilon$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \underbrace{\frac{R^2}{2\eta \cdot t}}^{\varepsilon/2} + \underbrace{\frac{\eta G^2}{2}}^{\varepsilon/2}$. $\leq \varepsilon$

Step 1

$$\frac{1}{t}\sum f(\theta_i) - f(\theta_*) \leq \frac{1}{t}\sum\left[\frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}\right]$$

$$\frac{R^2}{2\eta t}$$

$$\leq \frac{1}{t}\underbrace{\sum\left[\frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta}\right]} + \frac{\eta G^2}{2}$$

$$\|\theta_1 - \theta_*\|_2^2 - \cancel{\|\theta_2 - \theta_*\|_2^2} + \cancel{\|\theta_2 - \theta_*\|_2^2} - \|\theta_3 - \theta_*\|_2^2 + \ldots$$

$$\|\theta_1 - \theta_*\|_2^2 - \|\theta_{t+1} - \theta_*\|_2^2 \leq \|\theta_1 - \theta_*\|_2^2 \leq R^2$$

GD Analysis Proof

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

distance from opt

Step 2: $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \underbrace{\frac{R^2}{2\eta \cdot t}}_{\frac{\varepsilon}{2}} + \underbrace{\frac{\eta G^2}{2}}_{\frac{\varepsilon}{2}}$. — overshoot

$t = \frac{R^2G^2}{\varepsilon^2}$

$\eta = \frac{R}{G\sqrt{t}} = \frac{R}{G\frac{RG}{\varepsilon}} = \frac{\varepsilon}{G^2}$

$$\frac{R^2}{2\eta t} + \frac{\eta G^2}{2}$$

$$\frac{R^2}{2\frac{R}{G\sqrt{t}} \cdot t} + \cancel{\frac{\varepsilon}{G^2} \cdot \frac{G^2}{2}}^{\,\varepsilon/2}$$

$$\frac{GR}{2 \cdot \sqrt{t}} = \frac{GR}{2 \cdot \frac{RG}{\varepsilon}} = \frac{\varepsilon}{2}$$