COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2021.

Lecture 24

LOGISTICS

- Problem Set 5 is posted. It is due 12/13. It is optional and can be used to replace your lowest problem set grade.
- Quiz due Monday, 8pm. Reminder that lowest quiz grade is dropped.
- The final will be on 12/16 from 10:30am-12:30pm. In the class.
- Final review sheet is posted under the 'Schedule Tab'. I may continue to add to this and we plan to post a practice exam(s).
- Several extra office hours will be held before the final. Times TBD.

Last Class:

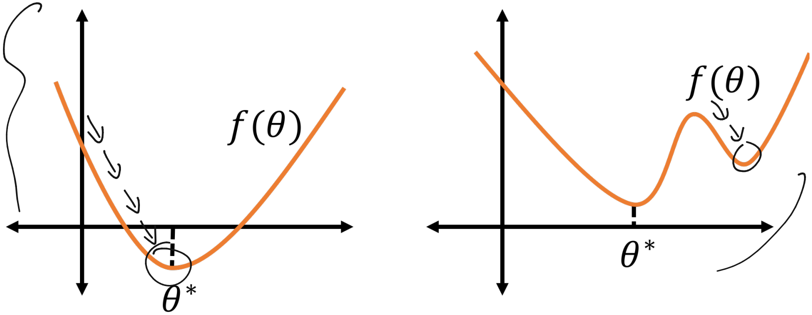
- Multivariable calculus review and gradient computation.
- Introduction to gradient descent. Motivation as a greedy algorithm.

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?

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

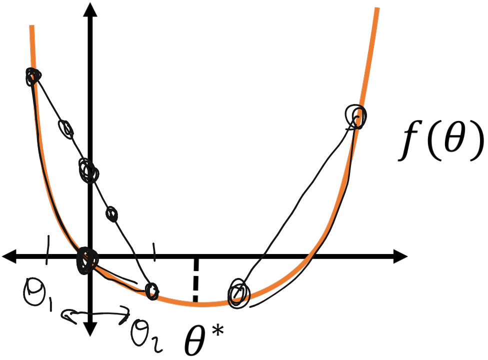


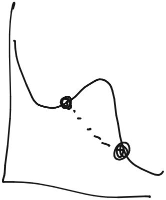
$f'(\theta_i)$

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 $\lambda \in [0, 1]$:

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



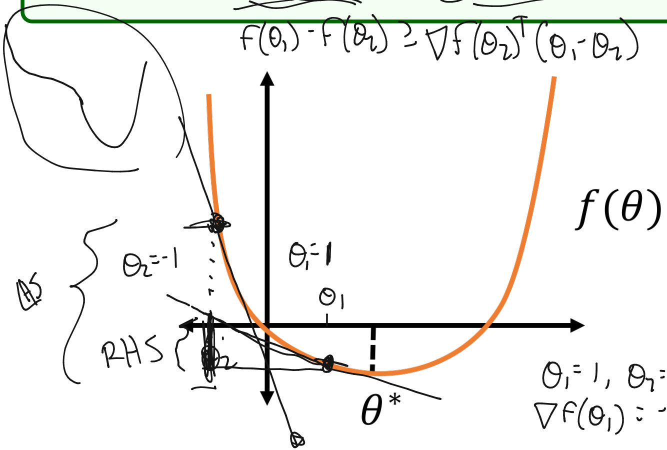


Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ ~~and $\lambda \in [0,1]$~~:

LHS RHS

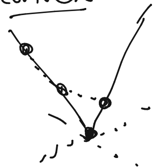
$$f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$$

$f(\theta_1) - f(\theta_2) \geq \nabla f(\theta_2)^T (\theta_1 - \theta_2)$



$f(x) = |x|$

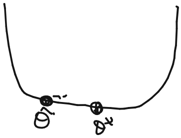
convex



$\theta_1 = 1$, $\theta_2 = -1$ $\quad$ $\theta_2 - \theta_1 = -2$

$\nabla f(\theta_1) = -1/2$ $\quad$ $-\frac{1}{2} \cdot -2 = 1$

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,...

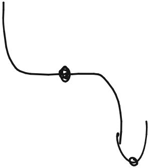
CONDITIONS FOR GRADIENT DESCENT CONVERGENCE

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:



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

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

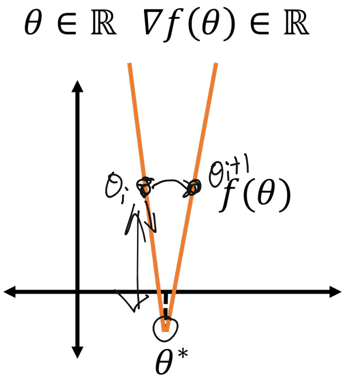
$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon = \min_{\vec{\theta}} f(\vec{\theta}) + \epsilon.$$

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:

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

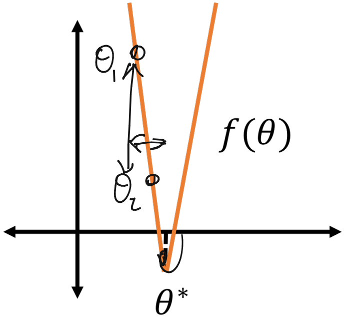
Examples: neural networks, clustering, mixture models.

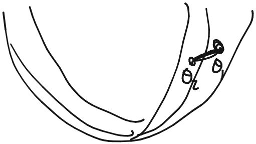


Gradient Descent Update:

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

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





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.:

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

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

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.

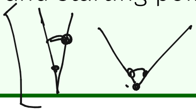
0 $\|\theta_*\|_2$

Gradient Descent

- 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 $\hat{\theta} = \arg\min_{\vec{\theta}_1, \ldots, \vec{\theta}_t} f(\vec{\theta}_i)$.

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:



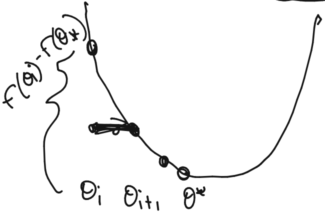
$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

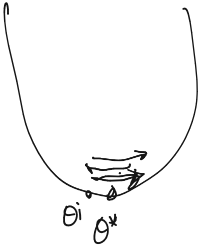
$\|\theta_1 - \theta_*\| \leq R$

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

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

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:

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

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

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

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

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

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

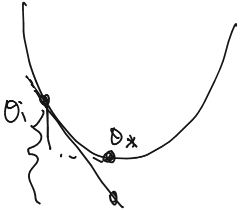
GD ANALYSIS PROOF

Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

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.



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:

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

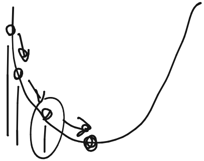
Step 1: For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

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

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



$$\underline{f(\hat{\theta}) - f(\theta_*)} \leq \frac{R^2}{2\eta t} + \frac{\eta G^2}{2}$$

$\hat{\theta} = \underline{\arg\min_{i=1\ldots t} f(\theta_i)}$

$\hat{\theta}$ is an output $\Rightarrow$ approx minimizer.

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:

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

not making it to opt

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

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

$$= \left(\sum_{i=1}^{t} \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2}{2\eta t}\right) + \frac{\eta G^2}{2} \leq \frac{R^2}{2\eta t} + \frac{\eta G^2}{2}$$

$$\frac{R^2}{2\eta \frac{R^2 G^2}{\epsilon^2}} + \frac{\eta G^2}{2} = \frac{\epsilon^2}{2\eta G^2} + \frac{\eta G^2}{2}$$

overshoot

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

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

Often want to perform convex optimization with convex constraints.

$$\vec{\theta}^* = \arg\min_{\vec{\theta}\in\mathcal{S}} f(\vec{\theta}),$$

where $\mathcal{S}$ is a convex set.

Previous slide: $t = \frac{R^2 G^2}{\epsilon^2}$ $\quad \eta = \frac{R}{G\sqrt{t}} = \frac{\epsilon}{G^2}$

$$f(\hat{\theta}) - f(\theta^*) \le \frac{\epsilon^2}{2G^2\eta} + \frac{\eta G^2}{2}$$

$$\le \frac{\epsilon^2}{2G^2\frac{\epsilon}{G^2}} + \frac{\frac{\epsilon}{G^2}G^2}{2} = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$