COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco

University of Massachusetts Amherst. Fall 2020.

Lecture 23

- Problem Set 4 is due next Wednesday, 8pm.
- Week 12 Quiz is due Monday, 8pm.
- The final will be 12/3-12/4, in any two hour window.
- Final review sheet is posted under the 'Schedule Tab'. I will continue to add to this.
- Office hours will be held before the final. Times TBA.

Last Class:

- Multivariable calculus review and gradient computation.
- Introduction to gradient descent. Motivation as a greedy algorithm.
- Conditions under which we will analyze gradient descent: convexity and Lipschitzness.

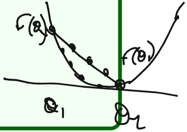
This Class:

- Analysis of gradient descent for Lipschitz, convex functions.
- Extension to projected gradient descent for constrained optimization.

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \Rightarrow \min_{\theta \in S} f(\theta)$$

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

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

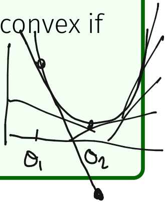


Corollary – Convex Function: A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0,1]$:

$\mathbb{R}^d$ $\mathbb{R}^d$

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

$\mathbb{R}$



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 $\vec{\theta}$.

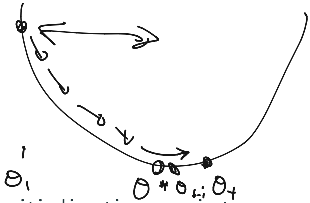
$f(\theta) = \theta^2$

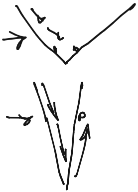
$f'(\theta) = 2\theta$

$f(\theta) = |\theta|$

$f'(\theta) \in \{-1, 1\}$

1-Lipschitz





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.

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

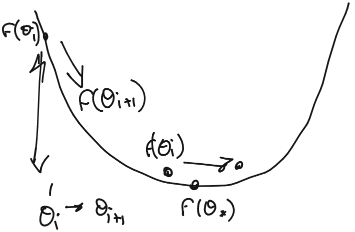
Theorem – GD on Convex Lipschitz Functions: For convex G-Lipschitz function f, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

current error   change in distance to opt

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

"noise" / "overshooting"



$\|a+b\|_2^2 = \|a\|_2^2 + 2a^T b + \|b\|_2^2$

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

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

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

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

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

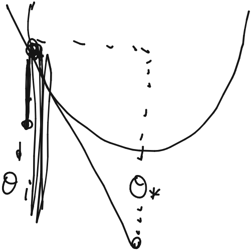
GD ANALYSIS PROOF

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.**



$f(\theta_*) - f(\theta_i) \geq \nabla f(\theta_i)^T(\theta_* - \theta_i)$

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

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:

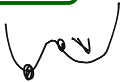
$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

R = radius

G = Lipschitzness

η = step size

$\|\theta_1 - \theta_*\|_2 \le R$ by assumption



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

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

$$\|\theta_1 - \theta_*\|^2 - \|\theta_2 - \theta_*\|^2 + \|\theta_2 - \theta_*\|^2 - \|\theta_3 - \theta_*\|_2^2 + \dots$$

$$\leq \frac{1}{t\cdot 2\eta}\left(\underbrace{\|\theta_1 - \theta_*\|_2^2 - \|\theta_{t+1} - \theta_*\|_2^2}_{\leq R^2}\right) + \frac{\eta G^2}{2} \leq \frac{R^2}{t\cdot 2\eta} + \frac{\eta G^2}{2}$$

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

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, $\boxed{\eta = \frac{R}{G\sqrt{t}}}$, and starting point within radius R of $\vec{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$A\theta \le b$

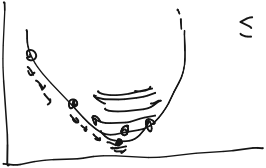
$$f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon.$$

$\eta \to 0$

$t \to \infty$

grand not covered

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



$$\le \frac{R^2}{2 \cdot \frac{R}{G\sqrt{t}} \cdot t} + \frac{RG^2}{G\sqrt{t}\cdot 2} = \frac{GR}{2\sqrt{t}} + \frac{GR}{2\sqrt{t}}$$

$$= \frac{GR}{\sqrt{t}}$$

$$\le \frac{GR}{\sqrt{\frac{R^2G^2}{\epsilon^2}}} = \epsilon$$

$\frac{1}{t}\sum_{i=1}^{t} f(\theta_i) - f(\theta_*) \le \epsilon$

$\Rightarrow \boxed{f(\hat{\theta}) - f(\theta^*) \le \epsilon}$

CONSTRAINED CONVEX OPTIMIZATION

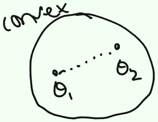
Often want to perform convex optimization with convex constraints.

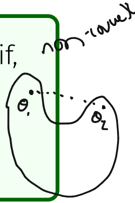
$$\vec{\theta}^* = \arg\min_{\vec{\theta}\in\mathcal{S} \subseteq \mathbb{R}^d} f(\vec{\theta}),$$

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

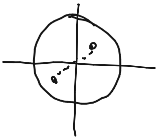
Definition – Convex Set: A set $\mathcal{S} \subseteq \mathbb{R}^d$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathcal{S}$ and $\lambda \in [0,1]$:

$$(1-\lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$$





E.g. $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$.



$\theta_1, \theta_2 \in S$

$\theta' = (1-\lambda)\theta_1 + \lambda\theta_2$

$\boxed{\theta' \in S}$

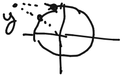
$\|\theta'\|_2 = \|(1-\lambda)\theta_1 + \lambda\theta_2\|_2$

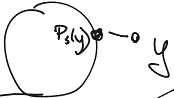
$\leq \|(1-\lambda)\theta_1\|_2 + \|\lambda\theta_2\|_2$

$\leq 1-\lambda+\lambda \leq 1$

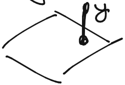
$y \in S, \quad P_S(y) = y$

For any convex set let $P_{\mathcal{S}}(\cdot)$ denote the projection function onto $\mathcal{S}$.





- $P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$.
- For $\mathcal{S} = \{\vec{\theta} \in \mathbb{R}^d : \|\vec{\theta}\|_2 \leq 1\}$ what is $P_{\mathcal{S}}(\vec{y})$? $= \dfrac{\vec{y}}{\max(1, \|y\|_2)}$

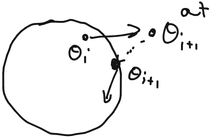


- For $\mathcal{S}$ being a k dimensional subspace of $\mathbb{R}^d$, what is $P_{\mathcal{S}}(\vec{y})$?

$P_S(y) = VV^T y$ where V is orthonormal basis for S.

Projected Gradient Descent

- Choose some initialization $\vec{\theta}_1$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t-1$
  - $\vec{\theta}_{i+1}^{(out)} = \vec{\theta}_i - \eta \cdot \vec{\nabla} f(\vec{\theta}_i)$
  - $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

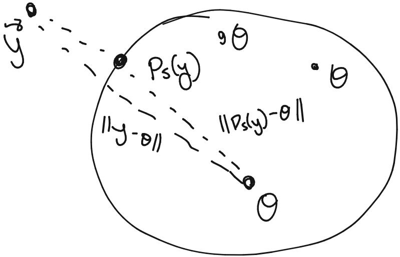


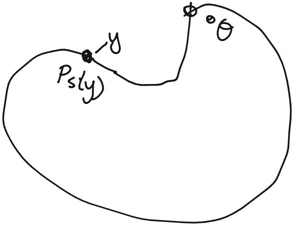
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_i} f(\vec{\theta}_i)$.

Projected gradient descent can be analyzed identically to gradient descent!

Theorem – Projection to a convex set: For any convex set $\mathcal{S} \subseteq \mathbb{R}^d$, $\vec{y} \in \mathbb{R}^d$, and $\vec{\theta} \in \mathcal{S}$,

$$\|P_{\mathcal{S}}(\vec{y}) - \vec{\theta}\|_2 \leq \|\vec{y} - \vec{\theta}\|_2.$$





Theorem – Projected GD: For convex G-Lipschitz function f, and convex set $\mathcal{S}$, Projected 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:

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

Recall: $\vec{\theta}_{i+1}^{(out)} = \vec{\theta}_i - \eta \cdot \vec{\nabla} f(\vec{\theta}_i)$ and $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

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

- Follow from GD analysis

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

$\theta_* \in S$

$\|\theta_{i+1} - \theta_*\| =$

$\|P_S(\theta_{i+1}^{out}) - \theta_*\|$

$\leq \|\theta_{i+1}^{out} - \theta_*\|$

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

$\leq \epsilon$

1. Gradient descent analysis
   - convex
   - Lipschitz

2. Convex sets & constrained optimization
   "convex optimization w/ convex constraints"
3. Solved w/ Projected Gradient Descent.

$$S : \{x : \|x\| \leq 1\}$$

$$P_S(y) = \frac{y}{\max(1, \|y\|)}$$