COMPSCL 514: ALGORITHMS FOR DATA SCIENCE

Cameron Musco
University of Massachusetts Amherst. Spring 2020.
Lecture 23
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.
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.
- Simple extension to projected gradient descent for constrained optimization.
  \[ f(\theta) \text{ s.t. } \theta \in S \]
**Definition – Convex Function:** A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if, for any $\tilde{\theta}_1, \tilde{\theta}_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$:

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

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

$$f(\tilde{\theta}_2) - f(\tilde{\theta}_1) \geq \nabla f(\tilde{\theta}_1)^T \left( \tilde{\theta}_2 - \tilde{\theta}_1 \right)$$
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.
- $\|\hat{\theta}_1 - \hat{\theta}_*\|_2 \leq R$ where $\hat{\theta}_1$ is the initialization point.

**Gradient Descent**

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

$$f(\hat{\theta}) \leq f(\tilde{\theta}_*) + \epsilon.$$
**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 $\bar{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

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

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

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

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

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

$$\leq \|\theta_i - \theta_*\|_2^2 + m^2 \sigma^2 - 2m \nabla f(\theta_i)^T (\theta_i - \theta_*)$$

$$\nabla f(\theta_i)^T (\theta_i - \theta_*) \leq \frac{\|\theta_i - \theta_*\|_2^2 - \|\theta_{i+1} - \theta_*\|_2^2 + m^2 \sigma^2}{2\eta}$$
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 $\tilde{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

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

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

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

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

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

Convexity

$$f(\theta_i) - f(\theta_*) \leq \nabla f(\theta_i)^T(\theta_i - \theta_*) \leq \text{RHS}$$
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 $\tilde{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

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

$$f(\hat{\theta}) \leq f(\tilde{\theta}^*) + \epsilon.$$

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

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

By Step 1:

$$\frac{1}{t} \sum_{i=1}^{t} f(\tilde{\theta}_i) - f(\tilde{\theta}^*) \leq \frac{mG^2}{2} + \frac{1}{t} \sum_{i=1}^{t} \frac{||\tilde{\theta}_i - \theta^*||^2 + ||\tilde{\theta}_{i+1} - \theta^*||^2}{2m}$$

$$\text{Want:} \min_{\theta} f(\theta) \leq \epsilon$$

$$f(\theta_k) - f(\theta_k) \leq \epsilon$$
**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 $\theta_*$, outputs $\hat{\theta}$ satisfying:

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

**Step 2:** 

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

$$n = \frac{R}{G\sqrt{t}}$$ 

$$\leq \frac{R\epsilon}{2\sqrt{t}} + \frac{R^2}{2
\sqrt{t}} = \frac{R\epsilon}{2\sqrt{t}} + \frac{R\epsilon}{2\sqrt{t}} = \frac{R\epsilon}{\sqrt{t}} \leq \epsilon.$$
Often want to perform convex optimization with convex constraints.

\[ \hat{\theta}^* = \arg \min_{\theta \in S} f(\theta), \]

where \( S \) is a convex set.
Often want to perform **convex optimization** with convex constraints.

\[
\hat{\theta}^* = \arg \min_{\hat{\theta} \in S} f(\hat{\theta}),
\]

where \( S \) is a **convex set**.

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

\[
(1 - \lambda)\hat{\theta}_1 + \lambda \cdot \hat{\theta}_2 \in S
\]
Often want to perform convex optimization with convex constraints.

\[ \tilde{\theta}^* = \arg \min_{\tilde{\theta} \in \mathcal{S}} f(\tilde{\theta}), \]

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 \( \tilde{\theta}_1, \tilde{\theta}_2 \in \mathcal{S} \) and \( \lambda \in [0, 1] \):

\[ (1 - \lambda)\tilde{\theta}_1 + \lambda \cdot \tilde{\theta}_2 \in \mathcal{S} \]

E.g. \( \mathcal{S} = \{ \tilde{\theta} \in \mathbb{R}^d : \| \tilde{\theta} \|_2 \leq 1 \} \).
For any convex set let $P_S(\cdot)$ denote the projection function onto $S$.

\begin{itemize}
  \item \( P_S(\vec{y}) = \arg \min_{\vec{\theta} \in S} \| \vec{\theta} - \vec{y} \|_2 \). 
\end{itemize}
For any convex set let $P_S(\cdot)$ denote the projection function onto $S$.

- $P_S(\bar{y}) = \arg \min_{\tilde{\theta} \in S} \| \tilde{\theta} - \bar{y} \|_2$.
- For $S = \{ \tilde{\theta} \in \mathbb{R}^d : \| \tilde{\theta} \|_2 \leq 1 \}$ what is $P_S(\bar{y})$?

$$P_S(y) = \frac{y}{\| y \|_2}$$
For any convex set let $P_S(\cdot)$ denote the projection function onto $S$.

- $P_S(\bar{y}) = \arg \min_{\tilde{\theta} \in S} \|\tilde{\theta} - \bar{y}\|_2$.
- For $S = \{\tilde{\theta} \in \mathbb{R}^d : \|\tilde{\theta}\|_2 \leq 1\}$ what is $P_S(\bar{y})$?
- For $S$ being a $k$ dimensional subspace of $\mathbb{R}^d$, what is $P_S(\bar{y})$?

\[
S = \{\tilde{\theta} : \tilde{\theta} = Vc \tilde{\xi} \}
\]

\[
P_S(y) = VV^T y
\]
For any convex set let $P_S(\cdot)$ denote the projection function onto $S$.

- $P_S(\bar{y}) = \arg\min_{\theta \in S} \|\theta - \bar{y}\|_2$.
- For $S = \{\bar{\theta} \in \mathbb{R}^d : \|\bar{\theta}\|_2 \leq 1\}$ what is $P_S(\bar{y})$?
- For $S$ being a $k$ dimensional subspace of $\mathbb{R}^d$, what is $P_S(\bar{y})$?

Projected Gradient Descent

- Choose some initialization $\hat{\theta}_1$ and set $\eta = \frac{R}{G\sqrt{t}}$.
- For $i = 1, \ldots, t - 1$
  - $\hat{\theta}_{i+1}^{(\text{out})} = \hat{\theta}_i - \eta \cdot \nabla f(\hat{\theta}_i)$
  - $\hat{\theta}_{i+1} = P_S(\hat{\theta}_{i+1}^{(\text{out})})$.
- Return $\hat{\theta} = \arg\min_{\hat{\theta}_i} f(\hat{\theta}_i)$. 
Projected gradient descent can be analyzed identically to gradient descent!
Projected gradient descent can be analyzed identically to gradient descent!

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

$$\|P_S(\vec{y}) - \vec{\theta}\|_2 \leq \|\vec{y} - \vec{\theta}\|_2.$$
**Theorem – Projected GD:** For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\tilde{\theta}_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\tilde{\theta}_*) + \epsilon = \min_{\tilde{\theta} \in S} f(\tilde{\theta}) + \epsilon$$
Theorem – Projected GD: For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\theta_*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta_*) + \epsilon = \min_{\tilde{\theta} \in S} f(\tilde{\theta}) + \epsilon$$

Recall: $\theta_{i+1}^{(out)} = \hat{\theta}_i - \eta \cdot \nabla f(\hat{\theta}_i)$ and $\hat{\theta}_{i+1} = P_S(\theta_{i+1}^{(out)})$. 
Theorem – Projected GD: For convex $G$-Lipschitz function $f$, and convex set $S$, Projected 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 $\theta^*$, outputs $\hat{\theta}$ satisfying:

$$f(\hat{\theta}) \leq f(\theta^*) + \epsilon = \min_{\bar{\theta} \in S} f(\bar{\theta}) + \epsilon$$

Recall: $\theta^{(out)}_{i+1} = \theta^t - \eta \cdot \nabla f(\theta^t)$ and $\theta^t_{i+1} = P_S(\theta^{(out)}_{i+1})$.

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


**Theorem – Projected GD:** For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\tilde{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

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

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

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

$$\|\tilde{\theta}_{i+1} - \tilde{\theta}_*\| \leq \|\tilde{\theta}_{i+1}^{(out)} - \tilde{\theta}_*\|$$
Theorem – Projected GD: For convex $G$-Lipschitz function $f$, and convex set $S$, 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 $\tilde{\theta}_*$, outputs $\hat{\theta}$ satisfying:

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

Recall: $\tilde{\theta}_{i+1}^{(out)} = \tilde{\theta}_i - \eta \cdot \nabla f(\tilde{\theta}_i)$ and $\tilde{\theta}_{i+1} = \text{Proj}_S(\tilde{\theta}_{i+1}^{(out)})$.

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

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

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