COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 23

Last Class:

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

Last Class:

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

This Class:

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

Goal: Find $\vec{\theta} \in \mathbb{R}^d$ that (nearly) minimizes convex function f.

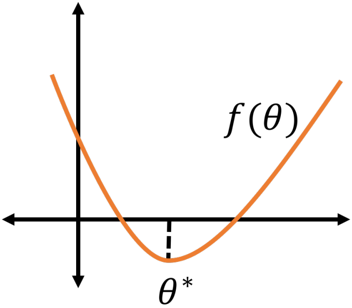
Gradient Descent Algorithm:

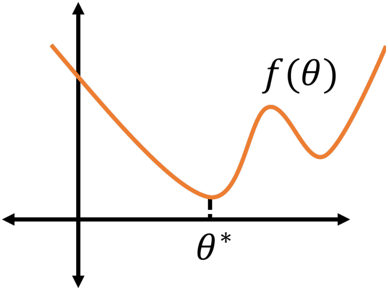
- Choose some initialization $\vec{\theta}^{(0)}$.
- For $i = 1, \ldots, t-1$
  - $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} - \eta \nabla f(\vec{\theta}^{(i-1)})$
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_1, \ldots \vec{\theta}_t} f(\vec{\theta}_i)$.

Step size η is chosen ahead of time or adapted during the algorithm. For now assume η stays the same in each iteration.

WHEN DOES GRADIENT DESCENT WORK?

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

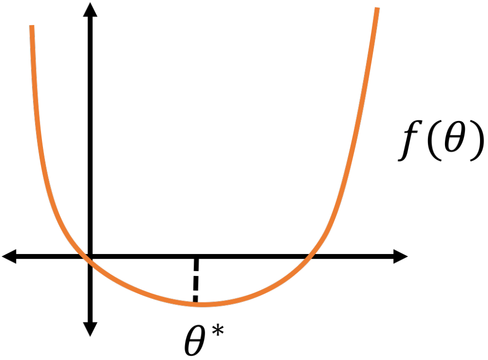




Gradient Descent Update in 1D: $\theta_{i+1} = \theta_i - \eta f'(\theta_i)$, i.e., increase θ if derivative is negative and decrease θ if derivative is positive.

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

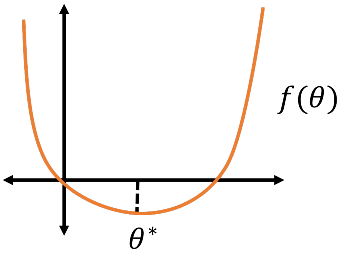
$$(1-\lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$



CONVEXITY

Corollary: A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is convex iff, for any $\theta_1, \theta_2 \in \mathbb{R}$:

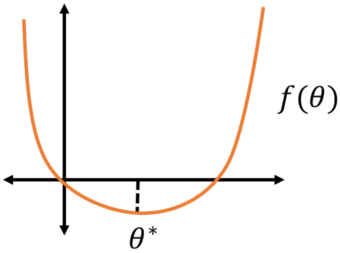
$$\text{“slope between } f(\theta_1) \text{ and } f(\theta_2)\text{”} = \frac{f(\theta_2) - f(\theta_1)}{\theta_2 - \theta_1} \geq f'(\theta_1)$$



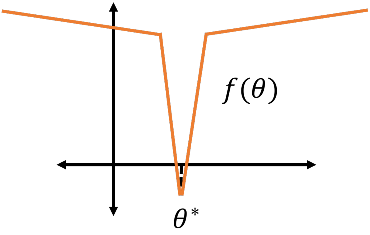
Corollary: A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is convex iff, for any $\theta_1, \theta_2 \in \mathbb{R}$:

$$\text{“slope between } f(\theta_1) \text{ and } f(\theta_2)\text{”} = \frac{f(\theta_2) - f(\theta_1)}{\theta_2 - \theta_1} \geq f'(\theta_1)$$

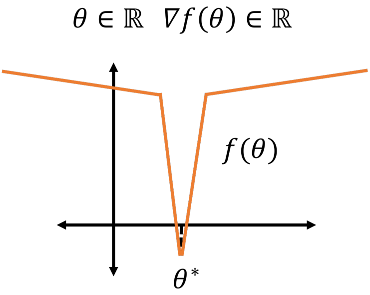
More generally, a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if and only if, for any $\vec{\theta}_1, \vec{\theta}_2 \in \mathbb{R}^d$: $f(\vec{\theta}_2) - f(\vec{\theta}_1) \geq \vec{\nabla} f(\vec{\theta}_1)^T \left(\vec{\theta}_2 - \vec{\theta}_1\right)$



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



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



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

For fast convergence, need to assume that the function is Lipschitz, i.e., size of gradient $\|\vec{\nabla} f(\vec{\theta})\|_2$ is bounded. We'll assume

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

Gradient Descent analysis for convex, Lipschitz functions.

Assume that:

- f is convex.
- f is G Lipschitz, i.e., $\|\vec{\nabla} f(\vec{\theta})\|_2 \leq G$ for all $\vec{\theta}$.
- $\|\vec{\theta}_1 - \vec{\theta}_*\|_2 \leq R$ 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 $i = 1, \ldots, t-1$
  - $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta \nabla f(\vec{\theta}_i)$
- Return $\hat{\theta} = \arg\min_{\vec{\theta}_1, \ldots \vec{\theta}_t} f(\vec{\theta}_i)$.

Theorem: For convex G-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within R of θ_*, outputs $\hat{\theta}$ satisfying $f(\hat{\theta}) \leq f(\theta_*) + \epsilon$.

Theorem: For convex G-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within R of θ_*, outputs $\hat{\theta}$ satisfying $f(\hat{\theta}) \leq f(\theta_*) + \epsilon$.

- Substituting $\theta_{i+1} = \theta_i - \eta f'(\theta_i)$ and letting $a_i = \theta_i - \theta_*$ gives:

$$a_{i+1}^2 = (\theta_{i+1} - \theta_*)^2 = (a_i - \eta f'(\theta_i))^2 = a_i^2 - 2\eta f'(\theta_i) a_i + (\eta f'(\theta_i))^2$$

Theorem: For convex G-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, GD run with $t \geq \frac{R^2G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within R of θ_*, outputs $\hat{\theta}$ satisfying $f(\hat{\theta}) \leq f(\theta_*) + \epsilon$.

- Substituting $\theta_{i+1} = \theta_i - \eta f'(\theta_i)$ and letting $a_i = \theta_i - \theta_*$ gives:

$$a_{i+1}^2 = (\theta_{i+1} - \theta_*)^2 = (a_i - \eta f'(\theta_i))^2 = a_i^2 - 2\eta f'(\theta_i)a_i + (\eta f'(\theta_i))^2$$

- Rearrange and use convexity to show:

$$f(\theta_i) - f(\theta_*) \leq f'(\theta_i)a_i = \frac{1}{2\eta}\left(a_i^2 - a_{i+1}^2\right) + \eta(f'(\theta_i))^2/2$$

GD ANALYSIS PROOF

Theorem: For convex G-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within R of θ_*, outputs $\hat{\theta}$ satisfying $f(\hat{\theta}) \leq f(\theta_*) + \epsilon$.

- Substituting $\theta_{i+1} = \theta_i - \eta f'(\theta_i)$ and letting $a_i = \theta_i - \theta_*$ gives:

$$a_{i+1}^2 = (\theta_{i+1} - \theta_*)^2 = (a_i - \eta f'(\theta_i))^2 = a_i^2 - 2\eta f'(\theta_i) a_i + (\eta f'(\theta_i))^2$$

- Rearrange and use convexity to show:

$$f(\theta_i) - f(\theta_*) \leq f'(\theta_i) a_i = \frac{1}{2\eta}\left(a_i^2 - a_{i+1}^2\right) + \eta (f'(\theta_i))^2/2$$

- Summing over i and using the fact $|f'(\theta_i)| \leq G$,

$$\frac{1}{t}\sum_{i=1}^t \left(f(\theta_i) - f(\theta_*)\right) \leq \left(\frac{1}{2t\eta}\sum_{i=1}^t (a_i^2 - a_{i+1}^2)\right) + \frac{\eta G^2}{2} \leq \frac{a_1^2}{2t\eta} + \frac{\eta G^2}{2}$$

Theorem: For convex G-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$, GD run with $t \geq \frac{R^2 G^2}{\epsilon^2}$ iterations, $\eta = \frac{R}{G\sqrt{t}}$, and starting point within R of θ_*, outputs $\hat{\theta}$ satisfying $f(\hat{\theta}) \leq f(\theta_*) + \epsilon$.

- Substituting $\theta_{i+1} = \theta_i - \eta f'(\theta_i)$ and letting $a_i = \theta_i - \theta_*$ gives:

$$a_{i+1}^2 = (\theta_{i+1} - \theta_*)^2 = (a_i - \eta f'(\theta_i))^2 = a_i^2 - 2\eta f'(\theta_i) a_i + (\eta f'(\theta_i))^2$$

- Rearrange and use convexity to show:

$$f(\theta_i) - f(\theta_*) \leq f'(\theta_i) a_i = \frac{1}{2\eta}\left(a_i^2 - a_{i+1}^2\right) + \eta (f'(\theta_i))^2/2$$

- Summing over i and using the fact $|f'(\theta_i)| \leq G$,

$$\frac{1}{t}\sum_{i=1}^{t}\left(f(\theta_i) - f(\theta_*)\right) \leq \left(\frac{1}{2t\eta}\sum_{i=1}^{t}(a_i^2 - a_{i+1}^2)\right) + \frac{\eta G^2}{2} \leq \frac{a_1^2}{2t\eta} + \frac{\eta G^2}{2}$$

- Using $a_1^2 \leq R^2$ and $f(\hat{\theta}) - f(\theta^*) \leq \frac{1}{t}\sum_{i=1}^{t}\left(f(\theta_i) - f(\theta_*)\right)$

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

GD ANALYSIS PROOF

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \theta_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$.

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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 via Convexity.

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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 via Convexity. Proof of Step 1.1:

$$\begin{aligned}
\|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2 &= \|\vec{\theta}_i - \eta \vec{\nabla} f(\vec{\theta}_i) - \vec{\theta}_*\|_2^2 \\
&= \|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - 2\eta \vec{\nabla} f(\vec{\theta}_i)^T(\vec{\theta}_i - \vec{\theta}_*) + \|\eta \vec{\nabla} f(\theta_i)\|_2^2
\end{aligned}$$

using fact $\|a + b\|_2^2 = \|a\|_2^2 + 2a^T b + \|b\|_2^2$.

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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 via Convexity. Proof of Step 1.1:

$$\begin{aligned}
\|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2 &= \|\vec{\theta}_i - \eta\vec{\nabla} f(\vec{\theta}_i) - \vec{\theta}_*\|_2^2 \\
&= \|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - 2\eta\vec{\nabla} f(\vec{\theta}_i)^T(\vec{\theta}_i - \vec{\theta}_*) + \|\eta\vec{\nabla} f(\theta_i)\|_2^2
\end{aligned}$$

using fact $\|a+b\|_2^2 = \|a\|_2^2 + 2a^Tb + \|b\|_2^2$. Since $\|\eta\vec{\nabla} f(\vec{\theta}_i)\|_2^2 \leq \eta^2 G^2$,

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

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2}$

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \Longrightarrow$

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

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2}{2\eta} + \frac{\eta G^2}{2} \Longrightarrow$

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

Proof of Step 2:

$$
\begin{aligned}
\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) &\leq \frac{t\eta G^2}{2} + \frac{1}{2\eta}\sum_{i=0}^{t-1}\left(\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2\right) \\
&= \frac{t\eta G^2}{2} + \frac{1}{2\eta}\|\vec{\theta}_0 - \vec{\theta}_*\|_2^2 \leq \frac{t\eta G^2}{2} + \frac{R^2}{2\eta}
\end{aligned}
$$

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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 $f(\hat{\theta}) \leq f(\vec{\theta}_*) + \epsilon$.

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, 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:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, 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:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2} \leq \epsilon$.

GD ANALYSIS PROOF

Theorem: For convex G-Lipschitz function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, 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:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2} \leq \epsilon$.
- Result follows since $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) \geq f(\hat{\theta})$.

CONSTRAINED CONVEX OPTIMIZATION

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.

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.

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 $\lambda \in [0, 1]$: $(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$

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.

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 $\lambda \in [0, 1]$: $(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$

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

$$P_{\mathcal{S}}(\vec{y}) = \arg\min_{\vec{\theta} \in \mathcal{S}} \|\vec{\theta} - \vec{y}\|_2$$

CONSTRAINED CONVEX OPTIMIZATION

Often want to perform convex optimization with convex constraints.

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

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 $\lambda \in [0, 1]$: $(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$

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

$$P_{\mathcal{S}}(\vec{y}) = \underset{\vec{\theta} \in \mathcal{S}}{\arg\min} \|\vec{\theta} - \vec{y}\|_2$$

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

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.

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 $\lambda \in [0, 1]$: $(1 - \lambda)\vec{\theta}_1 + \lambda \cdot \vec{\theta}_2 \in \mathcal{S}$

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

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

CONVEX PROJECTIONS

Analysis of projected gradient descent is almost identifcal to gradient descent analysis!

Analysis of projected gradient descent is almost identifcal to gradient descent analysis! Just need to appeal to following geometric result:

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

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

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

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

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

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

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

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

PROJECTED GRADIENT DESCENT ANALYSIS

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

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

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

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