COMPSCI 514: ALGORITHMS FOR DATA SCIENCE

Andrew McGregor

Lecture 22

SUMMARY

Last Class:

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

SUMMARY

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.

Let $\vec{e}_i \in \mathbb{R}^d$ denote the i^{th} standard basis vector,

$$\vec{e}_i = \underbrace{[0, 0, 1, 0, 0, \ldots, 0]}_{\text{1 at position } i} .$$

Let $\vec{e}_i \in \mathbb{R}^d$ denote the i^{th} standard basis vector,

$$\vec{e}_i = \underbrace{[0, 0, 1, 0, 0, \ldots, 0]}_{\text{1 at position } i}.$$

Partial Derivative:

$$\frac{\partial f}{\partial \vec{\theta}(i)} = \lim_{\epsilon \to 0} \frac{f(\vec{\theta} + \epsilon \cdot \vec{e}_i) - f(\vec{\theta})}{\epsilon}.$$

Let $\vec{e}_i \in \mathbb{R}^d$ denote the i^{th} standard basis vector,

$$\vec{e}_i = \underbrace{[0, 0, 1, 0, 0, \ldots, 0]}_{\text{1 at position } i}.$$

Partial Derivative:

$$\frac{\partial f}{\partial \vec{\theta}(i)} = \lim_{\epsilon \to 0} \frac{f(\vec{\theta} + \epsilon \cdot \vec{e}_i) - f(\vec{\theta})}{\epsilon}.$$

Directional Derivative: For unit vector $\vec{v}$,

$$D_{\vec{v}}\, f(\vec{\theta}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta} + \epsilon \vec{v}) - f(\vec{\theta})}{\epsilon}.$$

Gradient: Just a 'list' of the partial derivatives.

$$\vec{\nabla} f(\vec{\theta}) = \begin{bmatrix} \frac{\partial f}{\partial \vec{\theta}(1)} \\ \frac{\partial f}{\partial \vec{\theta}(2)} \\ \vdots \\ \frac{\partial f}{\partial \vec{\theta}(d)} \end{bmatrix}$$

Gradient: Just a 'list' of the partial derivatives.

$$\vec{\nabla} f(\vec{\theta}) = \begin{bmatrix} \frac{\partial f}{\partial \vec{\theta}(1)} \\ \frac{\partial f}{\partial \vec{\theta}(2)} \\ \vdots \\ \frac{\partial f}{\partial \vec{\theta}(d)} \end{bmatrix}$$

Directional Derivative in Terms of the Gradient:

$$D_{\vec{v}}\, f(\vec{\theta}) = \langle \vec{v}, \vec{\nabla} f(\vec{\theta}) \rangle.$$

Often the functions we are trying to optimize are very complex (e.g., a neural network). We will assume access to:

Function Evaluation: Can compute $f(\vec{\theta})$ for any $\vec{\theta}$.

Gradient Evaluation: Can compute $\vec{\nabla} f(\vec{\theta})$ for any $\vec{\theta}$.

FUNCTION ACCESS

Often the functions we are trying to optimize are very complex (e.g., a neural network). We will assume access to:

Function Evaluation: Can compute $f(\vec{\theta})$ for any $\vec{\theta}$.

Gradient Evaluation: Can compute $\vec{\nabla} f(\vec{\theta})$ for any $\vec{\theta}$.

In neural networks:

- Function evaluation is called a forward pass (propogate an input through the network).
- Gradient evaluation is called a backward pass (compute the gradient via chain rule, using backpropagation).

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta \vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta \vec{v})$.

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta \vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta \vec{v})$.

$$D_{\vec{v}}\ f(\vec{\theta}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta} + \epsilon \vec{v}) - f(\vec{\theta})}{\epsilon}.$$

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta\vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta\vec{v})$.

$$D_{\vec{v}}\, f(\vec{\theta}^{(i-1)}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta}^{(i-1)} + \epsilon\vec{v}) - f(\vec{\theta}^{(i-1)})}{\epsilon}.$$

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta\vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta\vec{v})$.

$$D_{\vec{v}}\ f(\vec{\theta}^{(i-1)}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta}^{(i-1)} + \epsilon\vec{v}) - f(\vec{\theta}^{(i-1)})}{\epsilon}.$$

So for small η:

$$f(\vec{\theta}^{(i)}) - f(\vec{\theta}^{(i-1)}) = f(\vec{\theta}^{(i-1)} + \eta\vec{v}) - f(\vec{\theta}^{(i-1)})$$

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta\vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta\vec{v})$.

$$D_{\vec{v}}\ f(\vec{\theta}^{(i-1)}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta}^{(i-1)} + \epsilon\vec{v}) - f(\vec{\theta}^{(i-1)})}{\epsilon}.$$

So for small η:

$$f(\vec{\theta}^{(i)}) - f(\vec{\theta}^{(i-1)}) = f(\vec{\theta}^{(i-1)} + \eta\vec{v}) - f(\vec{\theta}^{(i-1)}) \approx \eta \cdot D_{\vec{v}} f(\vec{\theta}^{(i-1)})$$

Gradient Descent Greedy Approach

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta \vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta \vec{v})$.

$$D_{\vec{v}}\ f(\vec{\theta}^{(i-1)}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta}^{(i-1)} + \epsilon \vec{v}) - f(\vec{\theta}^{(i-1)})}{\epsilon}.$$

So for small η:

$$\begin{aligned} f(\vec{\theta}^{(i)}) - f(\vec{\theta}^{(i-1)}) = f(\vec{\theta}^{(i-1)} + \eta \vec{v}) - f(\vec{\theta}^{(i-1)}) &\approx \eta \cdot D_{\vec{v}} f(\vec{\theta}^{(i-1)}) \\ &= \eta \cdot \langle \vec{v}, \vec{\nabla} f(\vec{\theta}^{(i-1)}) \rangle. \end{aligned}$$

GRADIENT DESCENT GREEDY APPROACH

Gradient descent is a greedy iterative optimization algorithm: Starting at $\vec{\theta}^{(0)}$, in each iteration let $\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \eta \vec{v}$, where η is a (small) 'step size' and $\vec{v}$ is a direction chosen to minimize $f(\vec{\theta}^{(i-1)} + \eta \vec{v})$.

$$D_{\vec{v}}\ f(\vec{\theta}^{(i-1)}) = \lim_{\epsilon \to 0} \frac{f(\vec{\theta}^{(i-1)} + \epsilon \vec{v}) - f(\vec{\theta}^{(i-1)})}{\epsilon}.$$

So for small η:

$$\begin{aligned} f(\vec{\theta}^{(i)}) - f(\vec{\theta}^{(i-1)}) = f(\vec{\theta}^{(i-1)} + \eta \vec{v}) - f(\vec{\theta}^{(i-1)}) &\approx \eta \cdot D_{\vec{v}} f(\vec{\theta}^{(i-1)}) \\ &= \eta \cdot \langle \vec{v}, \vec{\nabla} f(\vec{\theta}^{(i-1)}) \rangle. \end{aligned}$$

We want to choose $\vec{v}$ minimizing $\langle \vec{v}, \vec{\nabla} f(\vec{\theta}^{(i-1)}) \rangle$ – i.e., pointing in the direction of $\vec{\nabla} f(\vec{\theta}^{(i-1)})$ but with the opposite sign.

FUNCTION MINIMIZATION VIA GRADIENT DESCENT

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

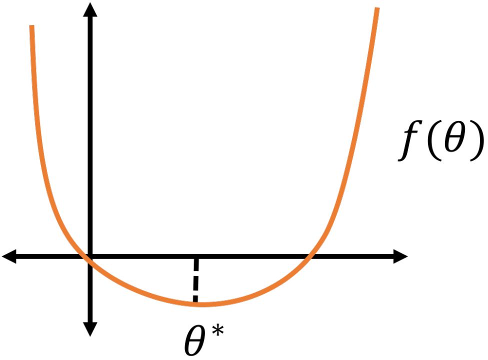
Gradient Descent Algorithm:

- 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.

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

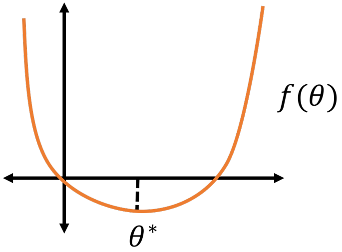
$$(1-\lambda) \cdot f(\vec{\theta}_1) + \lambda \cdot f(\vec{\theta}_2) \geq f\left((1-\lambda) \cdot \vec{\theta}_1 + \lambda \cdot \vec{\theta}_2\right)$$



CONVEXITY

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

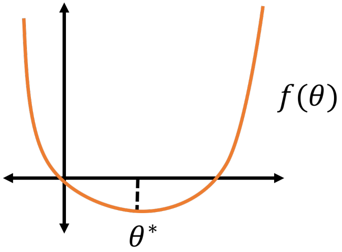
$$\text{“slope between } f(\theta_1) \text{ and } f(\theta_2)\text{”} = \frac{f(\theta_2) - f(\theta_1)}{\theta_2 - \theta_1} \geq f'(\theta_1)$$



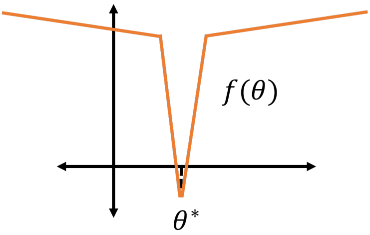
Corollary: A function $f : \mathbb{R} \to \mathbb{R}$ is convex iff, for any $\theta_1, \theta_2 \in \mathbb{R}$:

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

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

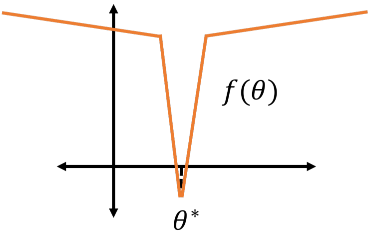


$\theta \in \mathbb{R}$ $\nabla f(\theta) \in \mathbb{R}$



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

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



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

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 $f : \mathbb{R}^d \to \mathbb{R}$, 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:** $\vec{\nabla} f(\vec{\theta}_i)^T \vec{a}_i \leq \frac{\|\vec{a}_i\|_2^2 - \|\vec{a}_{i+1}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$ where $\vec{a}_i = \vec{\theta}_i - \vec{\theta}_*$.

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

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

using$\|a - b\|_2^2 = \|a\|_2^2 - 2a^T b + \|b\|_2^2$. Then rearrange.

GD ANALYSIS PROOF

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

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

  using $\|a - b\|_2^2 = \|a\|_2^2 - 2a^Tb + \|b\|_2^2$. Then rearrange.

- **Step 2:** By convexity, for all i,

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

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

  using$\|a - b\|_2^2 = \|a\|_2^2 - 2a^T b + \|b\|_2^2$. Then rearrange.

- **Step 2:** By convexity, for all i,

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

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

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

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 2:** For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{a}_i\|_2^2 - \|\vec{a}_{i+1}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$
- **Step 3:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$.

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 2:** For all i, $f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{\|\vec{a}_i\|_2^2 - \|\vec{a}_{i+1}\|_2^2}{2\eta} + \frac{\eta G^2}{2}$
- **Step 3:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2}$.

Proof of Step 3:

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

$$\leq \frac{t\eta G^2}{2} + \frac{1}{2\eta}\|\vec{\theta}_0 - \vec{\theta}_*\|_2^2 \leq \frac{t\eta G^2}{2} + \frac{R^2}{2\eta}$$

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

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 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 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 $f : \mathbb{R}^d \to \mathbb{R}$, 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:** $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}_i) - f(\vec{\theta}_*) \leq \frac{R^2}{2\eta \cdot t} + \frac{\eta G^2}{2} \leq \epsilon$.
- 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 $\mathcal{S}$:

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

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 $P_{\mathcal{S}}(\vec{y})$?

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 $\vec{\theta} \in \mathcal{S}$,

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

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

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 $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

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 $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

Proof from earlier establishes that 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 $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

Proof from earlier establishes that for all i,

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

But Projection Lemma then ensures that 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 $\vec{\theta}_{i+1} = P_{\mathcal{S}}(\vec{\theta}_{i+1}^{(out)})$.

Proof from earlier establishes that for all i,

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

But Projection Lemma then ensures that for all i,

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

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