CS 561, Gradient Descent

Jared Saia
University of New Mexico

The Problem

Given:

- Convex space $\mathcal{K}$
- Convex function f

Goal: Find $x \in \mathcal{K}$ that minimizes $f(x)$

Convexity

1. A convex *set* contains every point on every line segment drawn between any two points in the set.
2. A convex *function* ensures any line segment between two points on the function is above the function: $\forall x, y, \lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

   - Equivalently, a convex function has a convex *epigraph*: the set of points above the function is convex.
   - If the function is twice differentiable, it is convex iff its second derivative is always non-negative.
3. A function f is *concave* iff $-f$ is convex.

What is a gradient?

- The *gradient* of a function f (∇f) is just the derivatives of f written as a vector.
- Ex: The gradient of $f(x, y) = 2x + 3y$ is the vector $(2, 3)$
- Ex: The gradient of $f(x, y) = x^2 + y^2$ at the point $x = 2, y = 3$ is $(4, 6)$
- Ex: The gradient of $f(x, y) = xy$ at the point $x = 2, y = 3$ is $(3, 2)$

Gradient Descent Variables

- $D = \max_{x,y \in \mathcal{K}} |x - y|$
- G is an upperbound on $|\nabla f(x)|$ for any $x \in \mathcal{K}$

Note: all norms are 2-norms. D is known as the *diameter* of $\mathcal{K}$

Gradient Descent Algorithm

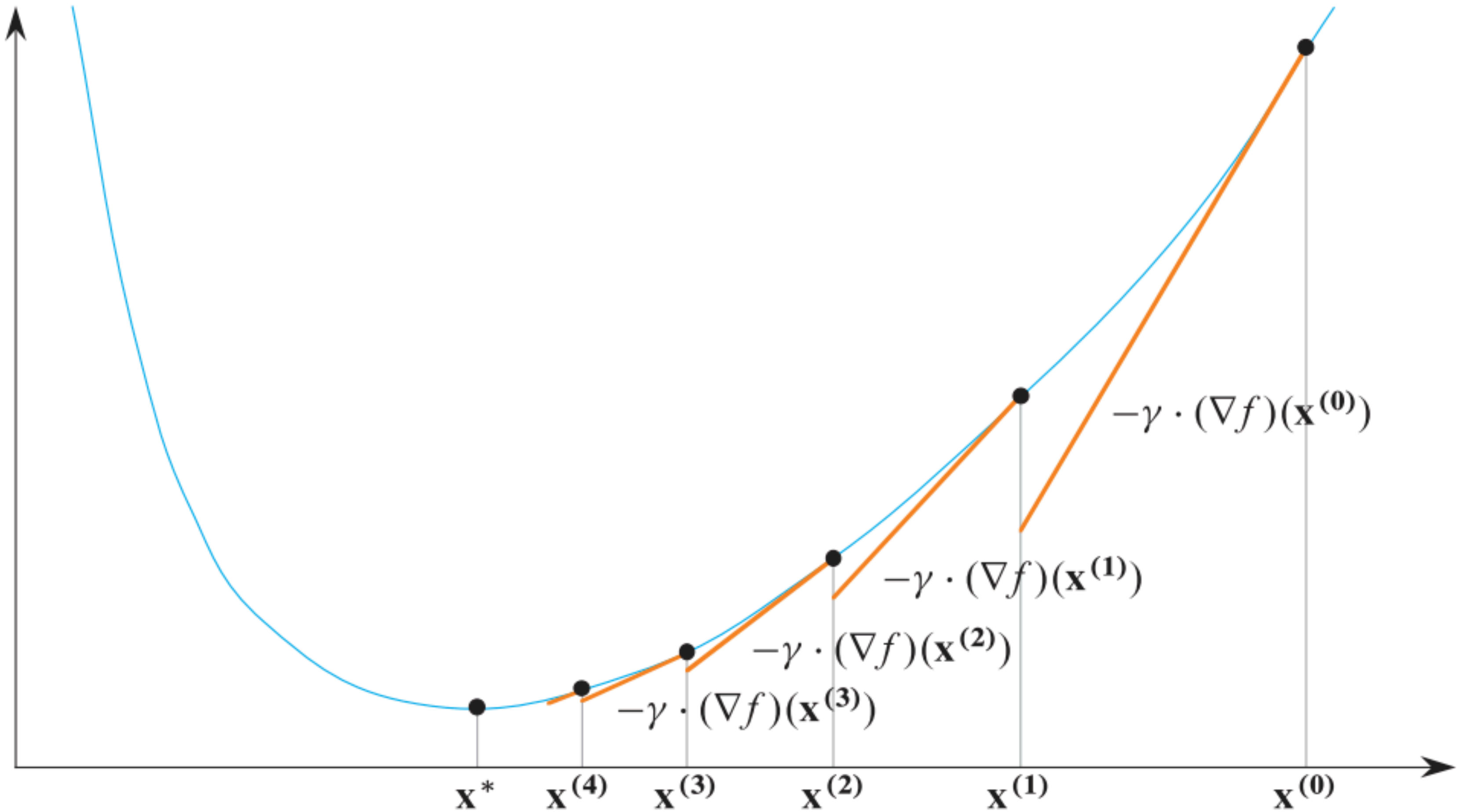
$\eta \leftarrow \frac{D}{G\sqrt{T}}$

Repeat for $i = 0$ to T:

1. $y_{i+1} \leftarrow x_i - \eta \nabla f(x_i)$
2. $x_{i+1} \leftarrow$ Projection of y_{i+1} onto $\mathcal{K}$

Output $z = \frac{1}{T} \sum_{i=1}^{T} x_i$

Example Run



$-\gamma \cdot (\nabla f)(\mathbf{x}^{(0)})$
$-\gamma \cdot (\nabla f)(\mathbf{x}^{(1)})$
$-\gamma \cdot (\nabla f)(\mathbf{x}^{(2)})$
$-\gamma \cdot (\nabla f)(\mathbf{x}^{(3)})$
$\mathbf{x}^*$
$\mathbf{x}^{(4)}$
$\mathbf{x}^{(3)}$
$\mathbf{x}^{(2)}$
$\mathbf{x}^{(1)}$
$\mathbf{x}^{(0)}$

Theorem 1 *Let $x^* \in \mathcal{K}$ be the value that minimizes f. Then, for any $\epsilon > 0$, if we set $T = \frac{D^2 G^2}{\epsilon^2}$, then:*

$$f(z) \leq f(x^*) + \epsilon$$

Fact 1: $f(x) - f(y) \leq \nabla f(x) \cdot (x - y)$

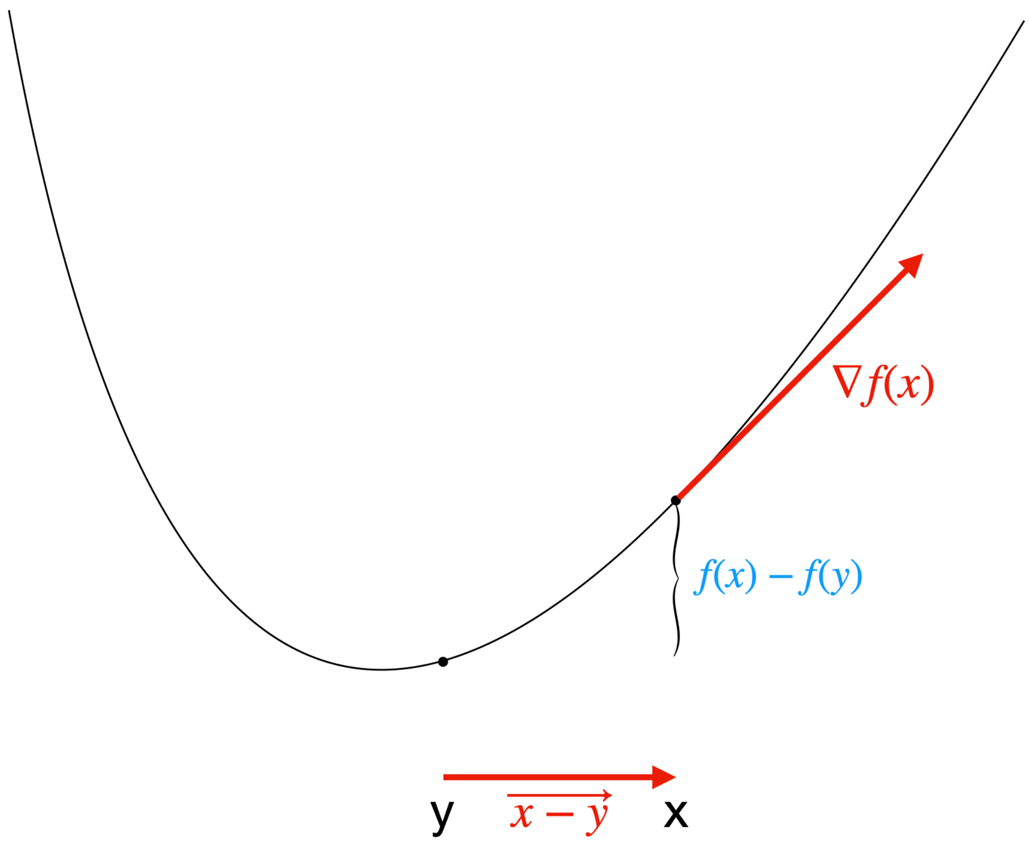
A convex function that is differentiable satisfies the following (basically, this says that the function is above the tangent plane at any point).

$$f(x + z) \geq f(x) + \nabla f(x) \cdot z, \text{ for all } x, z$$

Seting $z = y - x$, we get:

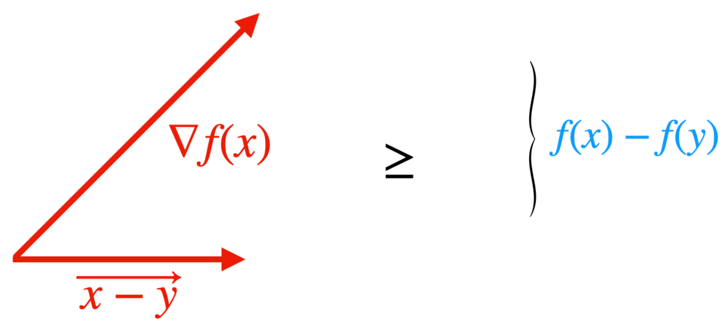
$$f(x) - f(y) \leq \nabla f(x) \cdot (x - y) \text{ for all } x, y$$

Fact 1: Picture



Fact:

$$\overrightarrow{x-y} \cdot \nabla f(x) \;\geq\; f(x) - f(y)$$



Proof of Theorem 1 (I)

$$
\begin{aligned}
|x_{i+1} - x^*|^2 &\leq |y_{i+1} - x^*|^2 \\
&= |x_i - x^* - \eta \nabla f(x_i)|^2 \\
&= |x_i - x^*|^2 + \eta^2 |\nabla f(x_i)|^2 - 2\eta \nabla f(x_i) \cdot (x_i - x^*)
\end{aligned}
$$

First step holds since x_{i+1} projects y_{i+1} onto a space that contains x^*. Second step holds by definition of y_{i+1}. Last step holds since $|v|^2 = v \cdot v$.

Proof of Theorem 1 (II)

From last slide:

$$|x_{i+1} - x^*|^2 \;\leq\; |x_i - x^*|^2 + \eta^2 |\nabla f(x_i)|^2 - 2\eta \nabla f(x_i) \cdot (x_i - x^*)$$

Reorganizing, and using definition of G:

$$\nabla f(x_i) \cdot (x_i - x^*) \;\leq\; \frac{1}{2\eta}\left(|x_i - x^*|^2 - |x_{i+1} - x^*|^2\right) + \frac{\eta}{2} G^2$$

Using Fact 1:

$$f(x_i) - f(x^*) \;\leq\; \frac{1}{2\eta}\left(|x_i - x^*|^2 - |x_{i+1} - x^*|^2\right) + \frac{\eta}{2} G^2 \qquad (1)$$

Proof of Theorem 1 (III)

Sum last inequality for $i = 1$ to T. After cancellations:

$$\sum_{i=1}^{T} (f(x_i) - f(x^*)) \leq \frac{1}{2\eta}\left(|x_1 - x^*|^2 - |x_{T+1} - x^*|^2\right) + \frac{T\eta}{2}G^2$$

Divide the above by T. By convexity, $f\left(\frac{1}{T}(\sum_i x_i)\right) \leq \frac{1}{T}\sum_i f(x_i)$. Since $z = \frac{1}{T}\sum_i x_i$, we get

$$f(z) - f(x^*) \leq \frac{D^2}{2\eta T} + \frac{\eta}{2}G^2.$$

Since $\eta = \frac{D}{G\sqrt{T}}$, the right hand side is at most $\frac{DG}{\sqrt{T}}$. Since $T = \frac{D^2G^2}{\epsilon^2}$, we have $f(z) \leq f(x^*) + \epsilon$

Online Gradient Descent

- Surprisingly, the gradient descent algorithm can work even when the function to minimize changes in every round!
- Even if these functions are chosen by an adversary! - So long as they are always convex.
- We just need to make a slight tweak in the algorithm (next slide - can you spot the differences?)

Online GD Algorithm

$\eta \leftarrow \frac{D}{G\sqrt{T}}$

Repeat for $i = 0$ to T:

1. $y_{i+1} \leftarrow x_i - \eta \nabla f_i(x_i)$
2. $x_{i+1} \leftarrow$ Projection of y_{i+1} onto $\mathcal{K}$

Online Gradient Theorem

Theorem 2 (Zinkevich's Theorem) *Let* $x^* \in \mathcal{K}$ *be the value that minimizes* $\sum_{i=1}^{T} f_i(x^*)$*. Then, for all* $T > 0$,

$$\frac{1}{T}\sum_{i=1}^{T}\left(f_i(x_i) - f_i(x^*)\right) \leq \frac{DG}{\sqrt{T}}.$$

Left hand side of this inequality is called the *regret* per step.

Proof

- Equation 1 from Slide 9 bounds the regret for step i
- Sum regrets over all i and divide by T to get the theorem!

Applctn: Portfolio Management

- From Section 16.6 in Arora notes

Portfolio Management

- Imagine you are investing in n stocks
- For i, $1 \leq i \leq n$, and $t > 1$, define

$$r_t[i] = \frac{\text{Price of stock } i \text{ on day } t}{\text{Price of stock } i \text{ on day } t-1}$$

- Let x^* be an optimal allocation of your money among the n stocks in hindsight.
- Q: Can we design an algorithm that is competitive with x^*?

Portfolio Management

- Our goal: Choose an allocation, x_t for each day t, that maximizes

$$\prod_t r_t \cdot x_t$$

- Taking logs, we get that we want to maximize:

$$\sum_t \log(r_t \cdot x_t)$$

- Same as minimizing

$$-\sum_t \log(r_t \cdot x_t)$$

- This last function is convex and so by Zinkevich's theorem, online gradient descent tracks

$$-\sum_t \log(r_t \cdot x^*)$$

Stochastic Gradient Descent

The final major trick of GD enables significant speed up. Assume we want to minimize over just one function, f, again.

- In each step, i, we estimate the gradient of f at x_i based on *one* random data item
- Call this random gradient g_i, where $E(g_i) = \nabla f(x_i)$
- Then, using the g_i's we get essentially same results as if we had the true gradient

Stochastic GD Algorithm

$\eta \leftarrow \frac{D}{G\sqrt{T}}$

Repeat for $i = 0$ to T:

1. $g_i \leftarrow$ a random vector, such that $E(g_i) = \nabla f(x_i)$
2. $y_{i+1} \leftarrow x_i - \eta g_i$
3. $x_{i+1} \leftarrow$ Projection of y_{i+1} onto $\mathcal{K}$

Output $z = \frac{1}{T}\sum_{i=1}^{T} x_i$

Stochastic GD Theorem

Theorem 3 $E(f(z)) \leq f(x^*) + \frac{DG}{\sqrt{T}}$.

Proof (1/2)

$$\begin{aligned} E(f(z)) &= E\left(f\left(\frac{1}{T}\sum_{i=1}^{T} x_i\right)\right) \\ &\leq E\left(\frac{1}{T}\ \sum_{i=1}^{T} f(x_i)\right) \quad \text{By convexity of f} \\ &\leq \frac{1}{T}E\left(\sum_{i=1}^{T} f(x_i)\right) \quad \text{Since E(cX) = cE(X) for constant c} \end{aligned}$$

Proof (2/2)

$$
\begin{aligned}
E(f(z) - f(x^*)) &\leq \frac{1}{T} E(\sum_{i=1}^{T} (f(x_i) - f(x^*))) && \text{By previous slide} \\
&\leq \frac{1}{T} \sum_i E(\nabla f(x_i) \cdot (x_i - x^*)) && \text{Using Fact 1} \\
&= \frac{1}{T} \sum_i E(g_i \cdot (x_i - x^*)) && \text{Cuz } E(g_i \cdot x) = \nabla f(x_i) \cdot x \\
&= \frac{1}{T} \sum_i E(f_i(x_i) - f_i(x^*)) && \text{Letting } f_i(x) = g_i \cdot x \\
&= E\left(\frac{1}{T} \sum_{i=1}^{T} (f_i(x_i) - f_i(x^*))\right) && \text{Linearity of Exp.} \\
&\leq \frac{DG}{\sqrt{T}} && \text{Regret bound using Zinkevich's Thm}
\end{aligned}
$$

Two Notes on Proof

- Requirement in Step 3: $E(g_i \cdot x) = \nabla f(x_i) \cdot x$, for all x
- Holds since dot product is linear, and $E(g_i) = \nabla f(x_i)$

- Requirement in Last Step: $f_i(x)$ is convex. Needed to use Zinkevich
- Holds since $f_i(x) = g_i \cdot x$ is *linear*

Take Away

Gradient Descent comes in 3 flavors:

- Standard Gradient Descent
- Online Gradient Descent
  Works even when function is changing
- Stochastic Gradient Descent
  Just need the correct gradient in **expectation**