COMP 761: Lecture 37 – Neural Networks II

David Rolnick

November 30, 2020

Problem

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

(Please don't post your ideas in the chat just yet, we'll discuss the problem soon in class.)

Course Announcements

Course Announcements

- Office hours today right after class.

Course Announcements

- Office hours today right after class.
- Reminder: Final two classes in the course are this Wed and **Thurs**.

Review: The gradient

Review: The gradient

- The *gradient* ∇f of a multivariable function $f(x) = f(x_1, \ldots, x_n)$ is the vector of partial derivatives with respect to the variables:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 & \cdots & \partial f / \partial x_n \end{bmatrix}.$$

Review: The gradient

- The *gradient* ∇f of a multivariable function $f(x) = f(x_1, \ldots, x_n)$ is the vector of partial derivatives with respect to the variables:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 & \cdots & \partial f / \partial x_n \end{bmatrix}.$$

- Can use gradient to estimate the amount that f changes:

$$f(x_1 + \epsilon_1, \ldots, x_n + \epsilon_n) \approx f(x_1, \ldots, x_n) + (\nabla f) \cdot \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}.$$

Review: The gradient

- The *gradient* ∇f of a multivariable function $f(x) = f(x_1, \ldots, x_n)$ is the vector of partial derivatives with respect to the variables:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 & \cdots & \partial f / \partial x_n \end{bmatrix}.$$

- Can use gradient to estimate the amount that f changes:

$$f(x_1 + \epsilon_1, \ldots, x_n + \epsilon_n) \approx f(x_1, \ldots, x_n) + (\nabla f) \cdot \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}.$$

- Dot product maximized when vectors aligned, so $\begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}$ should point along gradient (∇f).

Review: The gradient

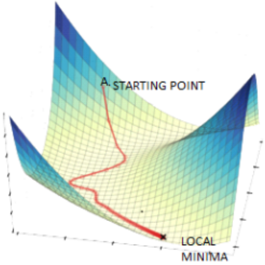
- The *gradient* ∇f of a multivariable function $f(x) = f(x_1, \ldots, x_n)$ is the vector of partial derivatives with respect to the variables:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_1 & \cdots & \partial f / \partial x_n \end{bmatrix}.$$

- Can use gradient to estimate the amount that f changes:

$$f(x_1 + \epsilon_1, \ldots, x_n + \epsilon_n) \approx f(x_1, \ldots, x_n) + (\nabla f) \cdot \begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}.$$

- Dot product maximized when vectors aligned, so $\begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}$ should point along gradient (∇f).
- Likewise, greatest *decrease* when $\begin{bmatrix} \epsilon_1 & \cdots & \epsilon_n \end{bmatrix}$ pointing along negative gradient ($-\nabla f$).



Gradient descent

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

  where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.
- Then, if γ is very small, we can use the approximation:

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\approx f(x^k) + (\nabla f(x^k)) \cdot (-\gamma \nabla f(x^k)) \\ &= f(x^k) - \gamma ||\nabla f(x^k)||^2. \end{aligned}$$

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

  $$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

  where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.
- Then, if γ is very small, we can use the approximation:

  $$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\approx f(x^k) + (\nabla f(x^k)) \cdot (-\gamma \nabla f(x^k)) \\ &= f(x^k) - \gamma ||\nabla f(x^k)||^2. \end{aligned}$$

- We are essentially taking a step in the direction that decreases the function the most.

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

  $$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

  where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.
- Then, if γ is very small, we can use the approximation:

  $$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\approx f(x^k) + (\nabla f(x^k)) \cdot (-\gamma \nabla f(x^k)) \\ &= f(x^k) - \gamma ||\nabla f(x^k)||^2. \end{aligned}$$

- We are essentially taking a step in the direction that decreases the function the most.
- We repeat until converge to a minimum (i.e. steps become really small).

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.
- Then, if γ is very small, we can use the approximation:

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\approx f(x^k) + (\nabla f(x^k)) \cdot (-\gamma \nabla f(x^k)) \\ &= f(x^k) - \gamma ||\nabla f(x^k)||^2. \end{aligned}$$

- We are essentially taking a step in the direction that decreases the function the most.
- We repeat until converge to a minimum (i.e. steps become really small).
- (The steps are small if $||\nabla f||^2 \approx 0$, which happens near the minimum.)

Gradient descent

- In *gradient descent*, we use this property of the gradient to our benefit to find the minimum of a function.
- Starting at some point $x^0 \in \mathbb{R}^n$, we progressively find points $x^1, x^2, \ldots \in \mathbb{R}^n$ by setting:

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

  where $\gamma > 0$ is a fixed *learning rate* and $\nabla f(x^k)$ means the gradient of f evaluated at x^k.
- Then, if γ is very small, we can use the approximation:

$$\begin{aligned} f(x^{k+1}) &= f(x^k - \gamma \nabla f(x^k)) \\ &\approx f(x^k) + (\nabla f(x^k)) \cdot (-\gamma \nabla f(x^k)) \\ &= f(x^k) - \gamma ||\nabla f(x^k)||^2. \end{aligned}$$

- We are essentially taking a step in the direction that decreases the function the most.
- We repeat until converge to a minimum (i.e. steps become really small).
- (The steps are small if $||\nabla f||^2 \approx 0$, which happens near the minimum.)
- Can similarly do gradient *ascent* to find maximum: $x^{k+1} = x^k + \gamma \nabla f(x^k)$.

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- What are ways that gradient descent can fail?

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- What are ways that gradient descent can fail?
- There are essentially three ways that gradient descent can “fail”:

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- What are ways that gradient descent can fail?
- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point.

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- What are ways that gradient descent can fail?
- There are essentially three ways that gradient descent can “fail”:
  - The iterative algorithm converges to the wrong point.
  - The algorithm doesn’t converge.

Gradient descent failure modes

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- What are ways that gradient descent can fail?
- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point.
  - The algorithm doesn't converge.
  - The algorithm converges, but very slowly.

Gradient descent – failure mode 1

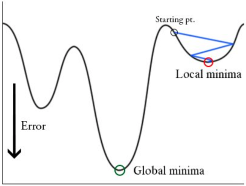
$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge.
  - The algorithm converges, but very slowly.

Gradient descent – failure mode 1

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge.
  - The algorithm converges, but very slowly.
- The first situation happens if there is a local minimum that isn't a global minimum – the algorithm then essentially gets stuck.



Gradient descent – failure mode 2

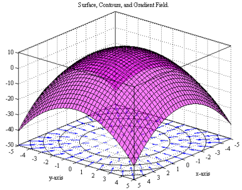
$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can “fail”:
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn’t converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly.

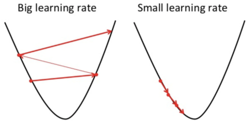
Gradient descent – failure mode 2

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly.
- The second situation happens if (i) the function isn't bounded and can descend forever, or (ii) the learning rate is too large and gradient descent bounces around without settling into a minimum.



(i)



(ii)

Gradient descent – failure mode 3

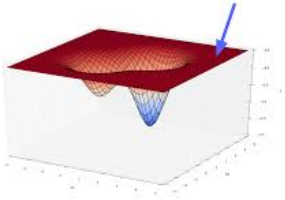
$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can “fail”:
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn’t converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.

Gradient descent – failure mode 3

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.
- The third situation happens if the gradient moves into a "flat region" where the gradient is very small.



Gradient descent – preventing failures

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.

Gradient descent – preventing failures

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.
- We are generally working with bounded functions.

Gradient descent – preventing failures

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.
- We are generally working with bounded functions.
- We can pick a very low learning rate (or decrease it as we go).

Gradient descent – preventing failures

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can “fail”:
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn’t converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.
- We are generally working with bounded functions.
- We can pick a very low learning rate (or decrease it as we go).
- Flat regions can be a big problem (as we will see later).

Gradient descent – preventing failures

$$x^{k+1} = x^k - \gamma \nabla f(x^k).$$

- There are essentially three ways that gradient descent can "fail":
  - The iterative algorithm converges to the wrong point – **local minma**.
  - The algorithm doesn't converge – **unbounded function** or **large learning rate**
  - The algorithm converges, but very slowly – **flat regions**.
- We are generally working with bounded functions.
- We can pick a very low learning rate (or decrease it as we go).
- Flat regions can be a big problem (as we will see later).
- We can ignore local minima if the function is *convex*.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- One way to define it is to use the graph of $f(x)$.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- One way to define it is to use the graph of $f(x)$.
- f is convex if for any two points $x, y \in \mathbb{R}^n$, the segment between $(x, f(x))$ and $(y, f(y))$ never goes below the graph:

$$\text{for any } 0 \leq t \leq 1, \qquad tf(x) + (1-t)f(y) \geq f(tx + (1-t)y).$$

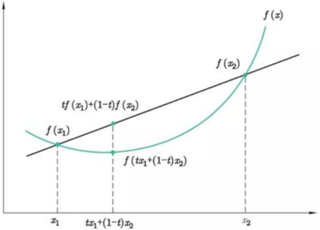
Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- One way to define it is to use the graph of $f(x)$.
- f is convex if for any two points $x, y \in \mathbb{R}^n$, the segment between $(x, f(x))$ and $(y, f(y))$ never goes below the graph:

$$\text{for any } 0 \leq t \leq 1, \qquad tf(x) + (1-t)f(y) \geq f(tx + (1-t)y).$$

- This is a generalization of the definition of *convexity* we saw for univariate functions.



Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- Another way to define convexity is to use the Hessian of $f(x)$.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- Another way to define convexity is to use the Hessian of $f(x)$.
- Since $x = [x_1, x_2, \ldots, x_n]$ is a vector, we can't just write $f'' > 0$, we have to consider derivatives with respect to all the x_i.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- Another way to define convexity is to use the Hessian of $f(x)$.
- Since $x = [x_1, x_2, \dots, x_n]$ is a vector, we can't just write $f'' > 0$, we have to consider derivatives with respect to all the x_i.
- The *Hessian* of f is the matrix that captures all the possible second derivatives:

$$\mathbf{H}f = \begin{bmatrix} \partial^2 f / \partial x_1 \partial x_1 & \partial^2 f / \partial x_1 \partial x_2 & \cdots & \partial f / \partial x_1 \partial x_n \\ \partial^2 f / \partial x_2 \partial x_1 & \partial^2 f / \partial x_2 \partial x_2 & \cdots & \partial f / \partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f / \partial x_n \partial x_1 & \partial^2 f / \partial x_n \partial x_2 & \cdots & \partial f / \partial x_n \partial x_n \end{bmatrix}.$$

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- Another way to define convexity is to use the Hessian of $f(x)$.
- Since $x = [x_1, x_2, \ldots, x_n]$ is a vector, we can't just write $f'' > 0$, we have to consider derivatives with respect to all the x_i.
- The *Hessian* of f is the matrix that captures all the possible second derivatives:

$$\mathbf{H}f = \begin{bmatrix} \partial^2 f/\partial x_1 \partial x_1 & \partial^2 f/\partial x_1 \partial x_2 & \cdots & \partial f/\partial x_1 \partial x_n \\ \partial^2 f/\partial x_2 \partial x_1 & \partial^2 f/\partial x_2 \partial x_2 & \cdots & \partial f/\partial x_2 \partial x_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f/\partial x_n \partial x_1 & \partial^2 f/\partial x_n \partial x_2 & \cdots & \partial f/\partial x_n \partial x_n \end{bmatrix}.$$

- Let's see how we can use that.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x + \epsilon) \approx f(x) + \epsilon f'(x)$.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

- What is the second-order approximation?

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

- What is the second-order approximation?
- With a bit more algebra, we can prove that it is:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x) + (1/2)\epsilon^T (\mathbf{H} f(x))\epsilon.$$

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x + \epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x + \epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2)f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x + \epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

- What is the second-order approximation?
- With a bit more algebra, we can prove that it is:

$$f(x + \epsilon) \approx f(x) + \epsilon \cdot \nabla f(x) + (1/2)\epsilon^T(\mathbf{H}f(x))\epsilon.$$

- So what should the analogous statement be to $f''(x) \geq 0$?

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

- What is the second-order approximation?
- With a bit more algebra, we can prove that it is:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x) + (1/2)\epsilon^T (\mathbf{H} f(x))\epsilon.$$

- The analogous statement to $f''(x) \geq 0$ is that for any vector $\epsilon \in \mathbb{R}^n$, we have $\epsilon^T (\mathbf{H} f(x))\epsilon \geq 0$, which is the same as saying that $\mathbf{H} f(x)$ is a *positive semi-definite matrix*.

Convex functions

For a univariate function $f(x)$ (where $x \in \mathbb{R}$), you can test if it's convex by checking if $f''(x) \geq 0$. What should it mean for a multivariate function $f(x)$ to be convex, where $x \in \mathbb{R}^n$ is a vector and $f(x)$ is a scalar?

- We saw that for $x \in \mathbb{R}$, we had the first-order (linear) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x)$.
- And the second-order (quadratic) approximation $f(x+\epsilon) \approx f(x) + \epsilon f'(x) + (\epsilon/2) f''(x)$.
- Now, if $x \in \mathbb{R}^n$ and ϵ is also $\in \mathbb{R}^n$, we had the first-order approximation:

$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x).$$

- What is the second-order approximation?
- With a bit more algebra, we can prove that it is:

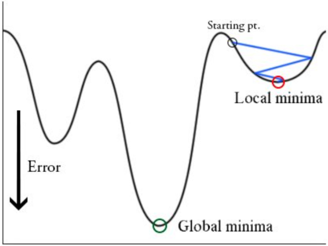
$$f(x+\epsilon) \approx f(x) + \epsilon \cdot \nabla f(x) + (1/2)\epsilon^T (\mathbf{H} f(x))\epsilon.$$

- The analogous statement to $f''(x) \geq 0$ is that for any vector $\epsilon \in \mathbb{R}^n$, we have $\epsilon^T (\mathbf{H} f(x))\epsilon \geq 0$, which is the same as saying that $\mathbf{H} f(x)$ is a *positive semi-definite matrix*.
- Equivalently (though we won't prove this): $\mathbf{H} f(x)$ has all eigenvalues ≥ 0.

Gradient descent in convex functions

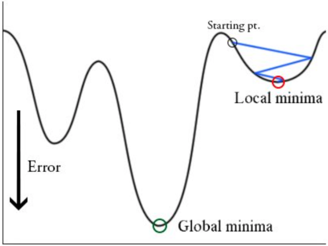
Gradient descent in convex functions

- For a convex function f, any local minimum of f is a global minimum of f.



Gradient descent in convex functions

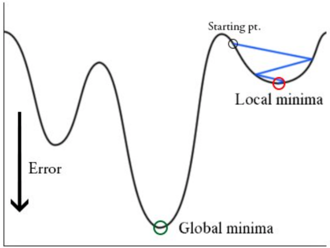
- For a convex function f, any local minimum of f is a global minimum of f.



- Informal way of seeing this: Suppose towards contradiction we have a local minimum x and a global minimum y with $f(y) < f(x)$.

Gradient descent in convex functions

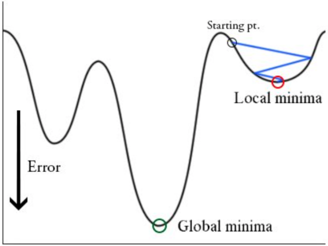
- For a convex function f, any local minimum of f is a global minimum of f.



- Informal way of seeing this: Suppose towards contradiction we have a local minimum x and a global minimum y with $f(y) < f(x)$.
- Then, the segment between x and y has to go below the graph of f, since x is a local minimum.

Gradient descent in convex functions

- For a convex function f, any local minimum of f is a global minimum of f.



- Informal way of seeing this: Suppose towards contradiction we have a local minimum x and a global minimum y with $f(y) < f(x)$.
- Then, the segment between x and y has to go below the graph of f, since x is a local minimum.
- That means f can't be convex!