Optimization methods
Lecture 2: Convex functions properties. Automatic differentiation

Alexandr Katrutsa

Modern State of Artificial Intelligence Masters Program
Moscow Institute of Physics and Technology

Brief reminder of the last lecture

- Introduction and course details

Brief reminder of the last lecture

- Introduction and course details
- Convex sets and their properties

Brief reminder of the last lecture

- Introduction and course details
- Convex sets and their properties
- Convex functions, how to recognize and construct them

Plan for today

- L-smooth convex functions

Plan for today

- L-smooth convex functions
- More about strongly convex functions

Plan for today

- L-smooth convex functions
- More about strongly convex functions
- Matrix calculus: brief reminder

Plan for today

- L-smooth convex functions
- More about strongly convex functions
- Matrix calculus: brief reminder
- Automatic differentiation technique

L-smooth function

Definition

Let $L > 0$. A function f is called L-smooth if it is differentiable and satisfies

$$\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2.$$

Descent lemma

Let f be an L-smooth function. Then for any $\mathbf{x}, \mathbf{y}$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

Proof of descent lemma

- The fundamental theorem of calculus

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$

Proof of descent lemma

- The fundamental theorem of calculus

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$

- $f(\mathbf{y}) - f(\mathbf{x}) =$
  $\langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$

Proof of descent lemma

- The fundamental theorem of calculus

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$

- $f(\mathbf{y}) - f(\mathbf{x}) =$
$\langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$
- $|f(\mathbf{y}) - f(\mathbf{x}) - \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \int_0^1 |\langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| dt \leq \int_0^1 \| f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}) \| \|\mathbf{y} - \mathbf{x}\| dt$

Proof of descent lemma

- The fundamental theorem of calculus
$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$
- $f(\mathbf{y}) - f(\mathbf{x}) = \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$
- $|f(\mathbf{y}) - f(\mathbf{x}) - \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \int_0^1 |\langle f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| dt \leq \int_0^1 \| f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}) \| \| \mathbf{y} - \mathbf{x} \| dt$
- $\int_0^1 \| f'(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f'(\mathbf{x}) \| \| \mathbf{y} - \mathbf{x} \| dt \leq \int_0^1 tL \| \mathbf{y} - \mathbf{x} \|_2^2 dt = \frac{L}{2} \| \mathbf{y} - \mathbf{x} \|_2^2$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

- The function f is $\|\mathbf{A}\|_2$-smooth

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

- The function f is $\|\mathbf{A}\|_2$-smooth
- $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

- The function f is $\|\mathbf{A}\|_2$-smooth
- $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$
- Assume f is L-smooth, then consider the vector $\mathbf{z}$ such that $\|\mathbf{z}\|_2 = 1$ and $\|\mathbf{A}\mathbf{z}\|_2 = \|\mathbf{A}\|_2$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

- The function f is $\|\mathbf{A}\|_2$-smooth
- $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$
- Assume f is L-smooth, then consider the vector $\mathbf{z}$ such that $\|\mathbf{z}\|_2 = 1$ and $\|\mathbf{A}\mathbf{z}\|_2 = \|\mathbf{A}\|_2$
- Then
  $\|\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{z}\|_2 = \|f'(\mathbf{z}) - f'(0)\|_2 \leq L\|\mathbf{z} - 0\|_2 = L$

Example

- Consider $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$, where $\mathbf{A} \in \mathbf{S}^n$
- Its gradient $f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$
- Then by definition

$$\|\mathbf{A}\mathbf{x} - \mathbf{b} - \mathbf{A}\mathbf{y} + \mathbf{b}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{x} - \mathbf{y}\|_2$$

- The function f is $\|\mathbf{A}\|_2$-smooth
- $\|\mathbf{A}\|_2 = \lambda_{\max}(\mathbf{A})$
- Assume f is L-smooth, then consider the vector $\mathbf{z}$ such that $\|\mathbf{z}\|_2 = 1$ and $\|\mathbf{A}\mathbf{z}\|_2 = \|\mathbf{A}\|_2$
- Then
  $\|\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{z}\|_2 = \|f'(\mathbf{z}) - f'(0)\|_2 \leq L\|\mathbf{z} - 0\|_2 = L$
- $\|\mathbf{A}\|_2$ is indeed the smallest smooth parameter

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \le L$ for any $\mathbf{x}$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha \mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha\mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$
   - $\lim\limits_{\alpha \to +0} \frac{\|f'(\mathbf{x}+\alpha\mathbf{d}) - f'(\mathbf{x})\|_2}{\alpha} = \|f''(\mathbf{x})\mathbf{d}\|_2 \leq L\|\mathbf{d}\|_2$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha\mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$
   - $\lim\limits_{\alpha \to +0} \frac{\|f'(\mathbf{x}+\alpha\mathbf{d}) - f'(\mathbf{x})\|_2}{\alpha} = \|f''(\mathbf{x})\mathbf{d}\|_2 \leq L\|\mathbf{d}\|_2$
   - Since this inequality holds for any $\mathbf{d}$, then $\|f''(\mathbf{x})\|_2 \leq L$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha\mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$
   - $\lim\limits_{\alpha \to +0} \frac{\|f'(\mathbf{x}+\alpha\mathbf{d}) - f'(\mathbf{x})\|_2}{\alpha} = \|f''(\mathbf{x})\mathbf{d}\|_2 \leq L\|\mathbf{d}\|_2$
   - Since this inequality holds for any $\mathbf{d}$, then $\|f''(\mathbf{x})\|_2 \leq L$
2. Assume $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha\mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$
   - $\lim\limits_{\alpha \to +0} \frac{\|f'(\mathbf{x}+\alpha\mathbf{d}) - f'(\mathbf{x})\|_2}{\alpha} = \|f''(\mathbf{x})\mathbf{d}\|_2 \leq L\|\mathbf{d}\|_2$
   - Since this inequality holds for any $\mathbf{d}$, then $\|f''(\mathbf{x})\|_2 \leq L$
2. Assume $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$
   - According to fundamental calculus theorem $f'(\mathbf{y}) - f'(\mathbf{x}) = \int_0^1 f''(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})dt$

Second order characteristic

Claim

Let $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, then for given $L > 0$ the following claims are equivalent

- f is L-smooth
- $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$

Proof

1. Assume f is L-smooth
   - Then for any $\mathbf{d}$ and $\alpha > 0$: $\|f'(\mathbf{x} + \alpha \mathbf{d}) - f'(\mathbf{x})\|_2 \leq \alpha L \|\mathbf{d}\|_2$
   - $\lim\limits_{\alpha \to +0} \frac{\|f'(\mathbf{x}+\alpha\mathbf{d}) - f'(\mathbf{x})\|_2}{\alpha} = \|f''(\mathbf{x})\mathbf{d}\|_2 \leq L\|\mathbf{d}\|_2$
   - Since this inequality holds for any $\mathbf{d}$, then $\|f''(\mathbf{x})\|_2 \leq L$
2. Assume $\|f''(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x}$
   - According to fundamental calculus theorem $f'(\mathbf{y}) - f'(\mathbf{x}) = \int_0^1 f''(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) dt$
   - $\|f'(\mathbf{y}) - f'(\mathbf{x})\|_2 \leq \left(\int_0^1 \|f''(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\|_2 dt \right) \|\mathbf{y} - \mathbf{x}\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2$

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

- f is L-smooth

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

- f is L-smooth
- $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

- f is L-smooth
- $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$
- $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

- f is L-smooth
- $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$
- $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|f'(\mathbf{x}) - f'(\mathbf{y})\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$
- $\langle f'(\mathbf{x}) - f'(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|f'(\mathbf{x}) - f'(\mathbf{y})\|_2^2$

L-smooth convex function

Claim

Let f be differentiable and convex function and $L > 0$. Then the following claims are equivalent:

- f is L-smooth
- $f(\mathbf{y}) \le f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$
- $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|f'(\mathbf{x}) - f'(\mathbf{y})\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$
- $\langle f'(\mathbf{x}) - f'(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ge \frac{1}{L} \|f'(\mathbf{x}) - f'(\mathbf{y})\|_2^2$
- $f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \ge \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) - \frac{L}{2} \lambda (1 - \lambda) \|\mathbf{x} - \mathbf{y}\|_2^2$ for any pair $\mathbf{x}, \mathbf{y}$

Strongly convex function

Definition: reminder

Function $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is called **strongly** convex with constant $m > 0$, if $\mathcal{X}$ is convex set and $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ и $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1 - \alpha) f(\mathbf{x}_2) - \frac{m}{2}\alpha(1 - \alpha)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

Uniqueness of minimizer

If function f is strictly convex, then its minimizer is unique.

Strongly convex function

Definition: reminder

Function $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is called **strongly** convex with constant $m > 0$, if $\mathcal{X}$ is convex set and $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ и $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) - \frac{m}{2}\alpha(1-\alpha)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

Uniqueness of minimizer

If function f is strictly convex, then its minimizer is unique.

Proof

Strongly convex function

Definition: reminder

Function $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is called **strongly** convex with constant $m > 0$, if $\mathcal{X}$ is convex set and $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ и $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) - \frac{m}{2}\alpha(1-\alpha)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

Uniqueness of minimizer

If function f is strictly convex, then its minimizer is unique.

Proof

- Assume, that there are two points $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $f(\mathbf{x}_1) = f(\mathbf{x}_2)$ and they are minimizers

Strongly convex function

Definition: reminder

Function $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is called **strongly** convex with constant $m > 0$, if $\mathcal{X}$ is convex set and $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ и $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) - \frac{m}{2}\alpha(1-\alpha)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2$$

Uniqueness of minimizer

If function f is strictly convex, then its minimizer is unique.

Proof

- Assume, that there are two points $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $f(\mathbf{x}_1) = f(\mathbf{x}_2)$ and they are minimizers
- Consider a point $\mathbf{z}$ from a segment $[\mathbf{x}_1, \mathbf{x}_2]$ and $f(\mathbf{z}) < \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2) = f(\mathbf{x}_2)$

Strongly convex function

Definition: reminder

Function $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ is called **strongly** convex with constant $m > 0$, if $\mathcal{X}$ is convex set and $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ и $\alpha \in [0, 1]$ we have:

$$f(\alpha\mathbf{x}_1+(1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1)+(1-\alpha)f(\mathbf{x}_2)-\frac{m}{2}\alpha(1-\alpha)\|\mathbf{x}_1-\mathbf{x}_2\|_2^2$$

Uniqueness of minimizer

If function f is strictly convex, then its minimizer is unique.

Proof

- Assume, that there are two points $\mathbf{x}_1 \neq \mathbf{x}_2$ such that $f(\mathbf{x}_1) = f(\mathbf{x}_2)$ and they are minimizers
- Consider a point $\mathbf{z}$ from a segment $[\mathbf{x}_1, \mathbf{x}_2]$ and $f(\mathbf{z}) < \lambda f(\mathbf{x}_1) + (1 - \lambda)f(\mathbf{x}_2) = f(\mathbf{x}_2)$
- Then if we take λ sufficiently close to 1, we get a contradiction with the assumption that $\mathbf{x}_2$ is a minimizer

Facts about strongly convex functions

Claim

The following conditions are all equivalent to the strong convexity of a differentiable function f

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{m}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$
- $\langle f'(\mathbf{x}) - f'(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq m \|\mathbf{x} - \mathbf{y}\|_2^2$

Facts about strongly convex functions

Claim

The following conditions are all equivalent to the strong convexity of a differentiable function f

- $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle f'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$
- $\langle f'(\mathbf{x}) - f'(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq m\|\mathbf{x} - \mathbf{y}\|_2^2$

Implications of strong conveity

If function f is strongly convex, then

- $\frac{1}{2}\|f'(\mathbf{x})\|_2^2 \geq m(f(\mathbf{x}) - f(\mathbf{x}^*))$
- $\|f'(\mathbf{x}) - f'(\mathbf{y})\|_2 \geq m\|\mathbf{x} - \mathbf{y}\|_2$

Summary on the convex functions properties

- Criterion of L-smoothness of convex function

Summary on the convex functions properties

- Criterion of L-smoothness of convex function
- Properties and facts about strongly convex functions

Summary on the convex functions properties

- Criterion of L-smoothness of convex function
- Properties and facts about strongly convex functions
- Why these characteristics of convex functions are important?

What are the next parts?

- Matrix calculus
- Automatic differentiation

Main definitions

Let $f : D \to E$, derivative related entity $\frac{\partial f}{\partial x} \in G$:

D	E	G	Name
$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	Derivative, $f'(x)$
$\mathbb{R}^n$	$\mathbb{R}$	$\mathbb{R}^n$	Gradient, $\frac{\partial f}{\partial x_i}$
$\mathbb{R}^n$	$\mathbb{R}^m$	$\mathbb{R}^{m \times n}$	Jacobi matrix, $\frac{\partial f_i}{\partial x_j}$
$\mathbb{R}^{m \times n}$	$\mathbb{R}$	$\mathbb{R}^{m \times n}$	$\frac{\partial f}{\partial x_{ij}}$

A square $n \times n$ matrix of second derivatives $\mathbf{H} = [h_{ij}]$ in the case of $f : \mathbb{R}^n \to \mathbb{R}$ is called hessian and has the following elements $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Main technique

Matrix representation of function: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x},\ \mathbf{A} \in \mathbf{S}^n$

↓

Scalar representation: $f = \sum_{ij} a_{ij} x_i x_j$

$\xrightarrow{f'_k = \frac{\partial f}{\partial x_k}}$

Scalar representation of gradient: $f'_k(\mathbf{x}) = 2\sum_j a_{kj} x_j$

↑

Matrix representation of gradient: $f'(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

Main technique

Matrix representation of function: $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x},\ \mathbf{A} \in \mathbf{S}^n$

Scalar representation: $f = \sum_{ij} a_{ij} x_i x_j$

$f'_k = \frac{\partial f}{\partial x_k}$

Scalar representation of gradient: $f'_k(\mathbf{x}) = 2\sum_j a_{kj} x_j$

Matrix representation of gradient: $f'(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$

- There is another approach to compute gradients based on a set of rules
- You can use it if you are sure

Composition of functions: scalar case

- Let $f(\mathbf{x}) = g(u(\mathbf{x}))$, then $f'(\mathbf{x}) = \frac{\partial g}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$

Composition of functions: scalar case

- Let $f(\mathbf{x}) = g(u(\mathbf{x}))$, then $f'(\mathbf{x}) = \frac{\partial g}{\partial u} \frac{\partial u}{\partial \mathbf{x}}$
- Check the dimensions consistency and verify the correctness of the form $\frac{\partial g}{\partial u}$.

Composition of functions: scalar case

- Let $f(\mathbf{x}) = g(u(\mathbf{x}))$, then $f'(\mathbf{x}) = \frac{\partial g}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$
- Check the dimensions consistency and verify the correctness of the form $\frac{\partial g}{\partial u}$.

Examples

Composition of functions: scalar case

- Let $f(\mathbf{x}) = g(u(\mathbf{x}))$, then $f'(\mathbf{x}) = \frac{\partial g}{\partial u}\frac{\partial u}{\partial \mathbf{x}}$
- Check the dimensions consistency and verify the correctness of the form $\frac{\partial g}{\partial u}$.

Examples

1. ℓ_2 vector norm: $f(\mathbf{x}) = \|\mathbf{x}\|_2$

Composition of functions: scalar case

- Let $f(\mathbf{x}) = g(u(\mathbf{x}))$, then $f'(\mathbf{x}) = \frac{\partial g}{\partial u} \frac{\partial u}{\partial \mathbf{x}}$
- Check the dimensions consistency and verify the correctness of the form $\frac{\partial g}{\partial u}$.

Examples

1. ℓ_2 vector norm: $f(\mathbf{x}) = \|\mathbf{x}\|_2$
2. Trace of the matrix product: $f(\mathbf{X}) = \mathrm{trace}(\mathbf{X}^\top \mathbf{X})$

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ is the k-th element of the gradient

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ is the k-th element of the gradient
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, where $\mathbf{J}$ — Jacobi matrix of h

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ is the k-th element of the gradient
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, where $\mathbf{J}$ — Jacobi matrix of h

Examples

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ is the k-th element of the gradient
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, where $\mathbf{J}$ — Jacobi matrix of h

Examples

- $h(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $g(\mathbf{u}) = \|\mathbf{u}\|_2^2$. Find $f'(\mathbf{x})$

Composition of functions: vector case

- $f(\mathbf{x}) = g(h(\mathbf{x}))$, where $h : \mathbb{R}^n \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}$
- $\frac{\partial f}{\partial x_k} = \sum_j \frac{\partial g}{\partial h_j} \frac{\partial h_j}{\partial x_k} = \sum_j J_{jk} \frac{\partial g}{\partial h_j}$ is the k-th element of the gradient
- $\frac{\partial f}{\partial \mathbf{x}} = \mathbf{J}^\top \frac{\partial g}{\partial h}$, where $\mathbf{J}$ — Jacobi matrix of h

Examples

- $h(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, $g(\mathbf{u}) = \|\mathbf{u}\|_2^2$. Find $f'(\mathbf{x})$
- $h(\mathbf{x}) = \cos(\mathbf{x})$ elementwise, $g(\mathbf{u}) = \sum_i u_i$. Find $\frac{\partial f}{\partial \mathbf{x}}$

Examples of non-differentiable convex function

- We already know that $\max$ operation preserves convexity

Examples of non-differentiable convex function

- We already know that $\max$ operation preserves convexity
- However, this operation produces non-differentiable convex function

Examples of non-differentiable convex function

- We already know that $\max$ operation preserves convexity
- However, this operation produces non-differentiable convex function
- Typical examples are $|x|$, $\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_1$

Examples of non-differentiable convex function

- We already know that $\max$ operation preserves convexity
- However, this operation produces non-differentiable convex function
- Typical examples are $|x|$, $\|\mathbf{x}\|_\infty$ and $\|\mathbf{x}\|_1$

Compressed sensing

$$\begin{aligned} &\min_{\mathbf{x}} \|\mathbf{x}\|_1 \\ \text{s.t. } &\mathbf{A}\mathbf{x} = \mathbf{b} \end{aligned}$$

Summary on matrix calculus

- Gradient, hessian and Jacobi matrix

Summary on matrix calculus

- Gradient, hessian and Jacobi matrix
- Compositions of functions

Summary on matrix calculus

- Gradient, hessian and Jacobi matrix
- Compositions of functions
- Chain rule helps a lot

From chain rule to `autodiff`[1]

Motivating example

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^{k} \frac{\partial h_i}{\partial g_k} \frac{\partial g_k}{\partial x_j}$

[1]Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^{k} \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^k \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

- $f = f_L \circ \ldots \circ f_1$ can be represented as a computational graph

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^k \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

- $f = f_L \circ \ldots \circ f_1$ can be represented as a computational graph
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^{k} \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

- $f = f_L \circ \ldots \circ f_1$ can be represented as a computational graph
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

How to compute $\mathbf{J}_f$

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to autodiff[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^k \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

- $f = f_L \circ \ldots \circ f_1$ can be represented as a computational graph
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

How to compute $\mathbf{J}_f$

- From right to left — forward mode

[1]Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

From chain rule to `autodiff`[1]

Motivating example

- $f = h(g(\mathbf{x}))$, where $h : \mathbb{R}^k \to \mathbb{R}^m$, $g : \mathbb{R}^n \to \mathbb{R}^k$
- $\mathbf{J}_f = \mathbf{J}_h(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$ or $J_f^{(i,j)} = \frac{\partial f_i}{\partial x_j} = \sum_{l=1}^k \frac{\partial h_i}{\partial g_k}\frac{\partial g_k}{\partial x_j}$

Generalization

- $f = f_L \circ \ldots \circ f_1$ can be represented as a computational graph
- $\mathbf{J}_f = \mathbf{J}_L \cdot \ldots \cdot \mathbf{J}_1$

How to compute $\mathbf{J}_f$

- From right to left — forward mode
- From left to right — backward mode

[1] Griewank A., Walther A. Evaluating derivatives: principles and techniques of algorithmic differentiation. – Society for Industrial and Applied Mathematics, 2008.

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

- Choose element x_j

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

- Choose element x_j
- Assign vector $\mathbf{u} = \mathbf{e}_j$, where e_j is the j-th orth

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

- Choose element x_j
- Assign vector $\mathbf{u} = \mathbf{e}_j$, where e_j is the j-th orth
- Multiply recursively $\mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ from right to left

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

- Choose element x_j
- Assign vector $\mathbf{u} = \mathbf{e}_j$, where e_j is the j-th orth
- Multiply recursively $\mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ from right to left
- Multiplication and computation of $f = f_L \circ \dots \circ f_1$ are performed simultaneously

Forward mode

Main idea

Compute $\frac{\partial f_i}{\partial x_k}$ for all i and fixed k, i.e. compute the j-th column of matrix $\mathbf{J}_f$

Implementation

- Choose element x_j
- Assign vector $\mathbf{u} = \mathbf{e}_j$, where e_j is the j-th orth
- Multiply recursively $\mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1 \mathbf{u}$ from right to left
- Multiplication and computation of $f = f_L \circ \ldots \circ f_1$ are performed simultaneously
- Every function f_i has to compute not only its own value but also the result of product $\mathbf{J}_i$ by given vector

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k
- Assign $\mathbf{u} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th orth

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k
- Assign $\mathbf{u} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th orth
- Multiply recursively $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ from left to right

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k
- Assign $\mathbf{u} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th orth
- Multiply recursively $\mathbf{u}^\top \mathbf{J}_L \dots \mathbf{J}_2 \mathbf{J}_1$ from left to right
- Compute f firstly, and after product from above. Therefore, we have to make two sweeps over computational graph

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k
- Assign $\mathbf{u} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th orth
- Multiply recursively $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ from left to right
- Compute f firstly, and after product from above. Therefore, we have to make two sweeps over computational graph
- Every f_i has to compute not only its own value but also multiplication of $\mathbf{J}_i^\top$ by vector

Backward mode or backpropagation

Main idea

Compute $\frac{\partial f_k}{\partial x_i}$ for all i and for fix k, i.e. compute the j-th row of the matrix $\mathbf{J}_f$

Implementation

- Choose element f_k
- Assign $\mathbf{u} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the k-th orth
- Multiply recursively $\mathbf{u}^\top \mathbf{J}_L \ldots \mathbf{J}_2 \mathbf{J}_1$ from left to right
- Compute f firstly, and after product from above. Therefore, we have to make two sweeps over computational graph
- Every f_i has to compute not only its own value but also multiplication of $\mathbf{J}_i^\top$ by vector

If $m = 1$, then $\mathbf{u} = 1$ and the result of backward mode differentiation equals to gradient!

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{Ju}) \leq 2.5 C(f(\mathbf{x}))$

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \le 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top \mathbf{u}) \le 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously
- Backward mode: additional memory is needed since the intermediate values of f_{i-1} is necessary to store for computing $\mathbf{J}_i^\top \mathbf{u}$

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top \mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously
- Backward mode: additional memory is needed since the intermediate values of f_{i-1} is necessary to store for computing $\mathbf{J}_i^\top \mathbf{u}$

Take home message

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously
- Backward mode: additional memory is needed since the intermediate values of f_{i-1} is necessary to store for computing $\mathbf{J}_i^\top\mathbf{u}$

Take home message

- If $m \ll n$, use backward mode

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously
- Backward mode: additional memory is needed since the intermediate values of f_{i-1} is necessary to store for computing $\mathbf{J}_i^\top\mathbf{u}$

Take home message

- If $m \ll n$, use backward mode
- If $m \geq n$, use forward mode

Forward vs backward modes

Computational complexity

- Forward mode: $C(f(\mathbf{x}), \mathbf{J}\mathbf{u}) \leq 2.5C(f(\mathbf{x}))$
- Backward mode: $C(f(\mathbf{x}), \mathbf{J}^\top\mathbf{u}) \leq 4C(f(\mathbf{x}))$

Memory consumption

- Forward mode: no extra memory is needed, result column of $\mathbf{J}$ and f are computed simultaneously
- Backward mode: additional memory is needed since the intermediate values of f_{i-1} is necessary to store for computing $\mathbf{J}_i^\top\mathbf{u}$

Take home message

- If $m \ll n$, use backward mode
- If $m \geq n$, use forward mode

Different implementations can significantly optimize all computations!

Example

Given function $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Find $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$

$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Example

Given function $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Find $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$

$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Forward mode

- Compute $\frac{\partial f}{\partial x_2}$
- $w_1 = x_1$, $w_2 = x_2$
- $\frac{\partial w_1}{\partial x_1} = 0$, $\frac{\partial w_2}{\partial x_2} = 1$
- $w_3 = 3w_2^2 \frac{\partial w_2}{\partial x_2}$
- $w_4 = \frac{\partial w_1}{\partial x_1} + w_3$
- $w_5 = -\sin(w_1 + w_2^3) w_4$
- $w_6 = 2\cos(w_1 + w_2^3) w_5$
- $w_6 = \frac{\partial f}{\partial x_2}$

Example

Given function $f(x_1, x_2) = \cos^2(x_1 + x_2^3)$. Find $\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}$

$$f(x_1, x_2) = f_1(f_2(f_3(x_1, f_4(x_2))))$$

Forward mode

- Compute $\frac{\partial f}{\partial x_2}$
- $w_1 = x_1$, $w_2 = x_2$
- $\frac{\partial w_1}{\partial x_1} = 0$, $\frac{\partial w_2}{\partial x_2} = 1$
- $w_3 = 3w_2^2 \frac{\partial w_2}{\partial x_2}$
- $w_4 = \frac{\partial w_1}{\partial x_1} + w_3$
- $w_5 = -\sin(w_1 + w_2^3) w_4$
- $w_6 = 2\cos(w_1 + w_2^3) w_5$
- $w_6 = \frac{\partial f}{\partial x_2}$

Backward mode

- $w_0 = 1$
- $w_1 = \frac{\partial f_1}{\partial f_2} w_0 = 2 f_2 w_0$
- $w_2 = \frac{\partial f_2}{\partial f_3} w_1 = -\sin(f_3) w_1$
- $w_3 = \frac{\partial f}{\partial x_1} = \frac{\partial f_3}{\partial x_1} w_2 = w_2$
- $w_4 = \frac{\partial f_3}{\partial f_4} w_2$
- $w_5 = \frac{\partial f}{\partial x_2} = \frac{\partial f_4}{\partial x_2} w_4 = 3x_2^2 w_4$

Summary on autodiff

- Chain rule leads to autodiff technique

Summary on autodiff

- Chain rule leads to autodiff technique
- Forward mode vs backward mode

Summary on autodiff

- Chain rule leads to autodiff technique
- Forward mode vs backward mode
- Typical use cases and issues