DSCI 572: Supervised Learning II
Optimization

Muhammad Abdul-Mageed

muhammad.mageed@ubc.ca

Deep Learning & NLP Lab

The University of British Columbia

Table of Contents

Optimization

Optimization

- The task of maximizing or minimizing a function, called **objective function** or **criterion**.
- When we are minimizing a function, we call it **cost function**, **loss function**, or **error function**.
- Usually denoted with a superscript $*$: $x* = \arg\min f(x)$

Calculus Refresher: Functions

Functions

- A function describes a relationship between an **input** and an **output**:

$$f(x) = 2x + 3$$

- A 'function' can **take another function** as input:

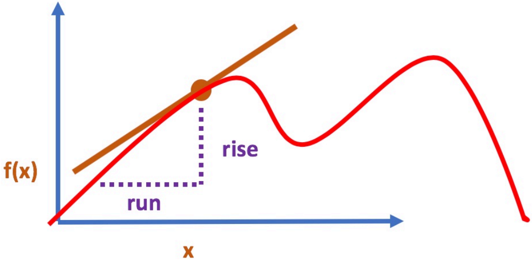
$$f(x) = g(x)$$

$$f(x) = o(g(x))$$

Derivative of a Function

Derivative of f(x)

- The **derivative** of a function $y = f(x)$, where x and y are real numbers is denoted as $y = f'(x)$ or $\frac{dy}{dx}$.
- The derivative $f'(x)$ gives the **slope** of $f(x)$ at the point x.
- The derivative tells us how a **small change in the input** results in a corresponding **change in the output**.



Derivative of a Linear Function

Linear Function

- A linear function has the **same derivative everywhere**

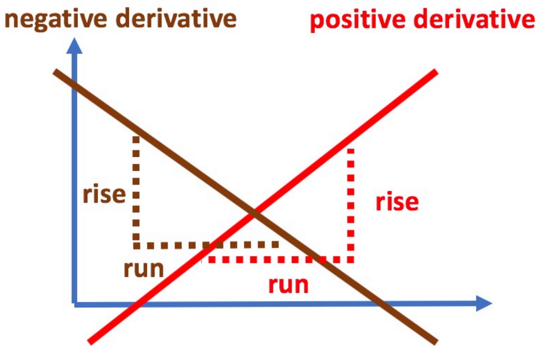


Figure: Positive and negative derivatives

Changing x with amount Δx

- With a small change in x, we get a new point Δx

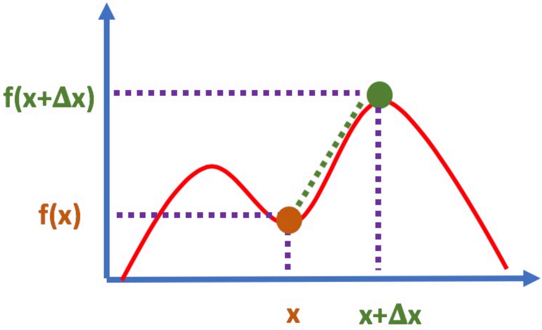


Figure: We have a new function $f(x + \Delta x)$

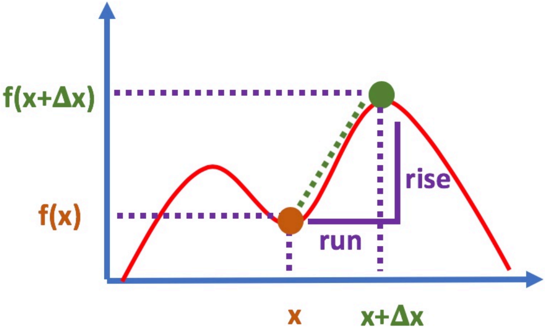


Figure: **Rise** $= f(x + \Delta x) - f(x)$; **run** $= \Delta x$

Derivative at x

- Derivative at x:

$$\approx \frac{rise}{run} = \frac{f(x+\Delta x) - f(x)}{(x+\Delta x) - x} = \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

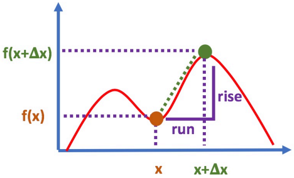


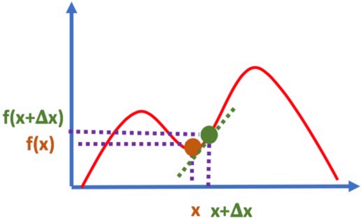
Figure: **Rise** $= f(x+\Delta x) - f(x)$; **run** $= \Delta x$

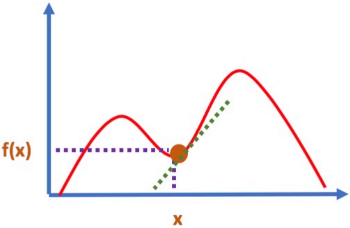
Limit

Derivative at x

- As Δx gets smaller, the line connecting the two functions becomes better and better approximation of the actual derivative at x. (**limit**)

$$\frac{df}{dx} = f^{'}(x) = \lim_{\Delta x \to 0} \left(\frac{f(x + \Delta x) - f(x)}{\Delta x} \right)$$





Derivative

Derivative & Gradient Descent

- We use the derivative to introduce changes in x to make **small improvements in** y.
- By moving x in small steps with the opposite side of the derivative, we can reduce $f(x)$. This is **gradient descent**.

Critical Points & Minima

- **Critical Points** (*aka* **stationary points**): Points where $f'(x) = 0$, and the derivative provides no information which direction to move
- **Local minimum**: a point where $f(x)$ is lower than at all neighboring points, making it not possible to decrease the function by making infinitesimal steps
- **Local maximum**: a point where $f(x)$ is higher than at all neighboring points, so it is not possible to increase the function by making infinitesimal steps
- **Saddle points:** Critical points that are neither maxima nor minima
- **Global minimum:** A point that obtains the absolute lowest value of $f(x)$

Gradient-Based Optimization

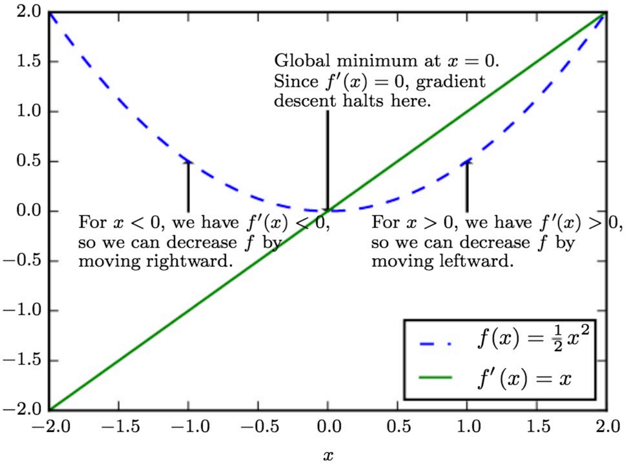


Figure: Gradient Descent. [From Goodfellow et al., 2016]

Critical Points

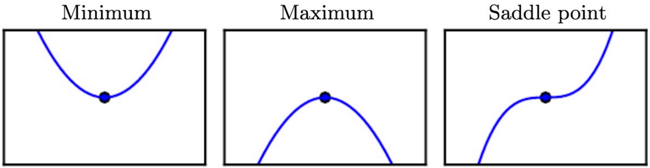


Figure: Different types of critical points. [From Goodfellow et al., 2016]

The Gradient: A Vector of Partial Derivatives

- **Gradient** generalizes the notion of derivative to the case of a function where the derivative is with respect to a **vector** ($f : \mathbb{R}^n \rightarrow \mathbb{R}$)
- As such, the gradient of f is the **vector containing all the partial derivatives**, denoted $\nabla_x f(x)$.
- $\frac{\partial}{\partial_{xi}} f(x)$: measures how f changes as only the variable x_i increases at point x
- Element i of the gradient is the **partial derivative of f with respect to x_i**
- **Critical points in multiple dimensions**: Points where every element of the gradient is equal to zero.

Directional Derivative and Gradient Descent

Directional Derivative

- **Directional derivatives** tell us how a multivariable function changes as we move along some vector in its input space.
- The **directional derivative** in direction $\mathbf{u}$: The slope of the function f in direction u. ($\mathbf{u}$ is a unit vector)
- To minimize f, we would like to find the direction in which f decreases the fastest
- This is minimized when $\mathbf{u}$ points in the opposite direction as the gradient: the gradient points directly uphill, and the negative gradient points directly downhill.
- So, we can **decrease f by moving in the direction of the negative gradient**
- This is known as the method of **steepest descent**, or **gradient descent**.

Directional Derivative and Gradient Descent

Directional Derivative

- **Gradient descent**: proposes a new point:

$$x' = x - \eta \nabla_x f(x)$$

- η: Known as the **learning rate**: a positive scalar determining the size of the step
- We can set η to a **small constant**, or **use line search**, among other methods.
- **Line search**: evaluate the function $f(x - \eta \nabla_x f(x))$ for several values of η and chose the one resulting in the smallest objective function value. See Wikipedia on "line search" [link].
- Steepest descent **converges** when every element of the gradient is zero, or very close to zero
- **Note:** Book uses ϵ instead of η

Beyond the Gradient: Jacobian

Jacobian

- Sometimes we need to find all the partial derivatives of a function whose input and output are both vectors
- **Jacobian matrix**: the matrix containing all these partial derivatives
- For a function **f**: $\mathbb{R}^{\mathrm{m}} \rightarrow \mathbb{R}^{\mathrm{n}}$, then the Jacobian matrix **J** $\in\mathbb{R}^{nxm}$ of $\mathbf{f}$ is defined such that:

$$J_{i,j} = \frac{\partial}{\partial_{xj}} f(\mathbf{x})_i$$

Beyond the Gradient: Hessian

Hessian

- We also make use of the derivative of the derivative, *aka* **second derivative** or the **Hessian matrix**.
- In a single dimension, we can denote $\frac{d^2}{d_{x^2}}$ by $f^{''}(x)$.
- The **second derivative** tells us how the first derivative will change as we vary the input
- This tells us `whether a gradient step will cause as much of an improvement as we would expect based on the gradient alone.`
- For more on the **Hessian**, see ch04 of Goodfellow et al. (2016).