Calculus Basics For Stochastic Gradient Descent

Amritesh Kumar, Neuraldemy (https://neuraldemy.com/)

January 2024

1 Introduction

Definition of a Derivative:

In the world of calculus, a derivative is like a magnifying glass for understanding change. Simply put, it tells us how a function is changing at any given point. Imagine you're tracking the position of a moving car - the derivative would reveal how its position changes with respect to time. Mathematically, the derivative of a function $f(x)$ at a specific point x is denoted as $f'(x)$ or $\frac{df}{dx}$. It represents the rate of change of $f(x)$ concerning x at that particular point.

Calculating Derivatives of Simple Functions:

Let's dive into the practical side. Calculating derivatives isn't as intimidating as it may sound. For a simple function like $f(x) = 2x^2$, finding its derivative involves applying a set of rules. For a power function $f(x) = x^n$, the derivative is $f'(x) = nx^{(n-1)}$. In our example, $f'(x) = 4x$, indicating that the rate at which the function $2x^2$ is changing at any point is four times its current value.

Chain Rule, Product Rule, and Quotient Rule:

Now, let's talk about the superpowers of derivatives - the Chain Rule, Product Rule, and Quotient Rule.

Chain Rule: The Chain Rule is your go-to when dealing with composite functions. If you have a function within a function, like $f(g(x))$, the Chain Rule helps you find the derivative. Mathematically, if $h(x) = f(g(x))$, then $h'(x) = f'(g(x)) \cdot g'(x)$. For example, if $f(x) = e^{2x}$ and $g(x) = 2x$, the derivative of $f(g(x))$ using the Chain Rule is $2e^{2x}$.

Product Rule: The Product Rule is handy when dealing with the product of two functions. If $h(x) = f(x) \cdot g(x)$, then $h'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$. Imagine $f(x) = 3x^2$ and $g(x) = \cos(x)$. Applying the Product Rule, the derivative of $h(x) = f(x) \cdot g(x)$ would be $6x \cdot \cos(x) - 3x^2 \cdot \sin(x)$.

Quotient Rule: When dealing with the ratio of two functions, the Quotient Rule comes to the rescue. If $h(x) = \frac{f(x)}{g(x)}$, then $h'(x) = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{(g(x))^2}$. Consider $f(x) = \sin(x)$ and $g(x) = x^2$. Applying the Quotient Rule, the derivative of $h(x) = \frac{\sin(x)}{x^2}$ becomes a bit involved but manageable: $\frac{x \cdot \cos(x) \cdot x^2 - \sin(x) \cdot 2x}{(x^2)^2}$.

Partial Derivatives:

Extending derivatives to functions with multiple variables.

Partial derivatives emerge as a powerful tool when dealing with functions that depend on multiple variables. Imagine a scenario where a function doesn't just rely on a single variable like x but involves multiple variables, say x and y. This is where partial derivatives step in to provide insights into how the function changes concerning each variable independently.

Mathematically, if we have a function $f(x, y)$ and we want to know how it changes concerning x while keeping y constant (or vice versa), we take the partial derivative. The partial derivative of f with respect to x is denoted as $\frac{\partial f}{\partial x}$, and similarly, the partial derivative with respect to y is $\frac{\partial f}{\partial y}$.

Partial Derivatives and Gradients:

Partial derivatives pave the way for another essential concept - gradients. The gradient of a multivariable function is like a compass that points in the direction of the steepest ascent. It's a vector composed of partial derivatives, giving us the rates of change in each direction.

For a function $f(x, y)$, the gradient is represented as $\nabla f = \langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \rangle$. The symbol ∇ is called "nabla," and it indicates the vector of partial derivatives.

Chain Rule:

The chain rule, becomes even more versatile when dealing with functions of multiple variables. Let's embark on a journey to unravel the intricacies of the chain rule in this multivariable landscape.

Consider a function $f(u, v)$ where both u and v are functions of another variable, say x. In mathematical terms, $u = u(x)$ and $v = v(x)$. Now, if we want to find the derivative of f concerning x, the chain rule steps in with its multivariable prowess.

Mathematically, the chain rule for multivariable functions is expressed as follows:

$$\frac{df}{dx} = \frac{\partial f}{\partial u} \cdot \frac{du}{dx} + \frac{\partial f}{\partial v} \cdot \frac{dv}{dx}$$

Breaking it down, the first term $\frac{\partial f}{\partial u} \cdot \frac{du}{dx}$ tells us how f changes concerning u (partial derivative of f with respect to u) multiplied by how u changes concerning x ($\frac{du}{dx}$). Similarly, the second term $\frac{\partial f}{\partial v} \cdot \frac{dv}{dx}$ gives us the contribution from the v side.

Let's delve into an example to make this more concrete. Consider the function $f(u, v) = u^2 + v^3$, where $u = x^2$ and $v = \sin(x)$. We want to find $\frac{df}{dx}$. Applying the chain rule:

$$\frac{df}{dx} = 2u \cdot 2x + 3v^2 \cdot \cos(x)$$

Substituting $u = x^2$ and $v = \sin(x)$:

$$\frac{df}{dx} = 2x^2 \cdot 2x + 3\sin^2(x) \cdot \cos(x)$$

Simplifying further gives us the final expression for $\frac{df}{dx}$. This example illustrates how the chain rule adapts seamlessly to the multivariable scenario, providing a systematic approach to finding derivatives when functions are entangled in a web of variables.

Gradient: Definition and interpretation of gradients.

The gradient emerges as a fundamental concept, akin to a compass guiding us through the topography of functions. The *gradient* of a scalar-valued function $f(x, y)$ is a vector that points in the direction of the steepest ascent at a given point. It's denoted by ∇f or $\vec{\nabla} f$, and mathematically, it is represented as:

$$\nabla f = \langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \rangle$$

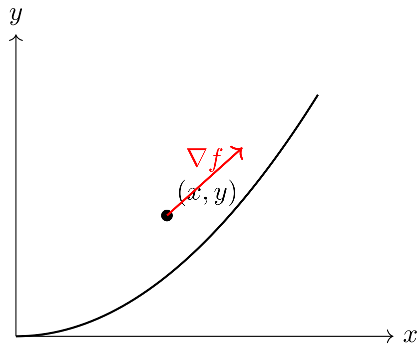
Here, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are the partial derivatives of f concerning x and y, respectively. The gradient vector ∇f encapsulates the rates of change in each direction.

Interpretation of Gradients:

Imagine standing on a hilly landscape, and $f(x, y)$ represents the height above sea level at each point (x, y). The gradient at a particular point would then be a vector pointing uphill, indicating the direction of the steepest ascent. The magnitude of the gradient vector represents the steepness of the ascent.

Visual Representation:

Let's visualize this concept with a diagram:



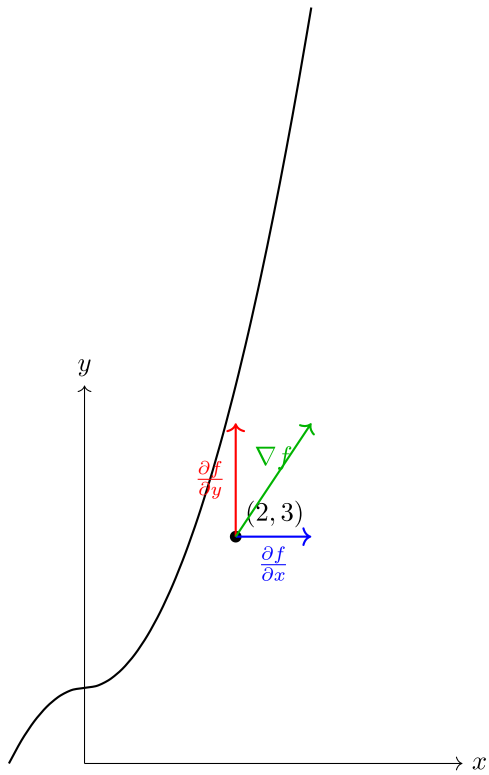
In this illustration, the function $f(x, y) = 0.2x^2$ represents a hill. At the point (x, y), the gradient vector ∇f points in the direction of the steepest ascent. The longer the vector, the steeper the ascent.

Example:

Let's take a concrete example to solidify our understanding. Consider the function $f(x, y) = x^2 + y^2$. The gradient of f is:

$$\nabla f = \langle 2x, 2y \rangle$$

At a specific point $(2, 3)$, the gradient vector is $\nabla f = \langle 4, 6 \rangle$. This means that at the point $(2, 3)$, the function f has the steepest ascent in the direction of $\langle 4, 6 \rangle$.

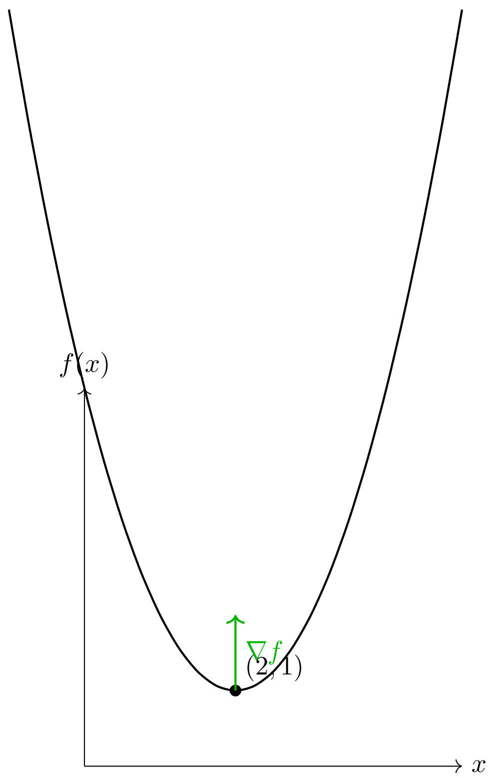


At the point $(2, 3)$, let's calculate the partial derivatives and the gradient:

$$\frac{\partial f}{\partial x} = 2x\Big|_{(2,3)} = 4$$
$$\frac{\partial f}{\partial y} = 2y\Big|_{(2,3)} = 6$$
$$\nabla f = \langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \rangle\Big|_{(2,3)} = \langle 4, 6 \rangle$$

In this example, the red vector represents the partial derivative with respect to y ($\frac{\partial f}{\partial y}$), the blue vector represents the partial derivative with respect to x ($\frac{\partial f}{\partial x}$), and the green vector represents the gradient vector (∇f) at the point $(2, 3)$. Intuitively, the partial derivatives indicate how the function changes concerning each variable individually, and the gradient provides the combined direction of the steepest ascent at that point. In the context of the Stochastic Gradient Descent (SGD) algorithm, understanding these derivatives and gradients helps in navigating the function's landscape to find the minimum, akin to finding the lowest point in a terrain. The gradient, with its direction and magnitude, serves as a valuable tool for understanding the landscape of SGD.

2D Visualization for Convex Function and Gradient:



Consider the convex function $f(x) = (x-2)^2 + 1$. This function has a single global minimum at $x = 2$, which is marked as the point $(2, 1)$. The green vector represents the gradient vector ∇f at this point. In convex functions, the gradient always points in the direction of the steepest ascent. At the global minimum, the gradient vector points directly away from the minimum, indicating the direction of the steepest ascent. Convex functions ensure that the global minimum is the lowest point, making them suitable for optimization algorithms.

1.1 Derivatives In N-Dimensions

Scalar Case:

Let's consider a scalar function $f(x)$ where x is a scalar variable. The derivative of f with respect to x is denoted as $f'(x)$ or $\frac{df}{dx}$. It represents the rate of change of the function with respect to x.

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Example:

$$f(x) = x^2$$
$$f'(x) = 2x$$

Gradient: Vector in, Scalar out:

Let's consider a vector-valued function $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x})]$ where $\mathbf{x}$ is a vector input. The gradient, denoted as $\nabla\mathbf{f}(\mathbf{x})$, is a vector of partial derivatives with respect to each component of $\mathbf{x}$.

$$\nabla\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \frac{\partial f_2}{\partial x_2} \\ \vdots \\ \frac{\partial f_n}{\partial x_n} \end{bmatrix}$$

Example:

$$\mathbf{f}(\mathbf{x}) = [x^2, 2x]$$

$$\nabla\mathbf{f}(\mathbf{x}) = \begin{bmatrix} 2x \\ 2 \end{bmatrix}$$

Jacobian: Vector in, Vector out:

Let's consider a vector-valued function $\mathbf{f}(\mathbf{x})$ where $\mathbf{x}$ is a vector input. The Jacobian matrix, denoted as $J(\mathbf{f})$ or $\frac{d\mathbf{f}}{d\mathbf{x}}$, is a matrix of partial derivatives. It represents the rate at which each component of the output vector changes concerning each component of the input vector.

$$J(\mathbf{f}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Example:

$$\mathbf{f}(\mathbf{x}) = [x^2, 2x]$$

$$J(\mathbf{f}) = \begin{bmatrix} 2x & 0 \\ 2 & 0 \end{bmatrix}$$

Generalized Jacobian: Tensor in, Tensor out:

Let's consider a tensor-valued function $\mathbf{F}(\mathbf{X})$ where $\mathbf{X}$ is a tensor input. The generalized Jacobian tensor, denoted as $J(\mathbf{F})$ or $\frac{d\mathbf{F}}{d\mathbf{X}}$, is a higher-order tensor of partial derivatives. It represents the rate at which each component of the output tensor changes concerning each component of the input tensor.

$$J(\mathbf{F}) = \frac{d\mathbf{F}}{d\mathbf{X}}$$

Example:

$$\mathbf{F}(\mathbf{X}) = \begin{bmatrix} x^2 & 2x \\ \sin(x) & e^x \end{bmatrix}$$

$$J(\mathbf{F}) = \begin{bmatrix} 2x & 2 \\ \cos(x) & e^x \end{bmatrix}$$

Optimization:

Basics of Optimization Problems:

Optimization involves finding the best solution among a set of feasible solutions. In the context of mathematical functions, optimization problems typically aim to find the minimum or maximum value of a function within a given domain.

A generic optimization problem can be formulated as follows:

$$\text{Minimize} \quad f(\mathbf{x})$$

$$\text{Subject to} \quad \mathbf{x} \in \mathcal{D}$$

Here, $f(\mathbf{x})$ is the objective function to be minimized, and $\mathcal{D}$ is the feasible set representing the constraints on the input variables $\mathbf{x}$.

Types of Optimization Problems:

Optimization problems come in various forms, broadly categorized into two types:
1. Unconstrained Optimization: In unconstrained optimization, the objective is to find the minimum or maximum of a function without any constraints on the input variables. The general form is:

$$\text{Minimize (or Maximize)} \quad f(\mathbf{x})$$

2. Constrained Optimization: Constrained optimization involves finding the minimum or maximum of a function while considering constraints on the input variables. The general form is:

$$\text{Minimize (or Maximize)} \quad f(\mathbf{x})$$

$$\text{Subject to} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m$$

Here, $g_i(\mathbf{x})$ are inequality constraints.

Minimizing a Function using Derivatives:

Derivatives are instrumental in optimization problems. To find the minimum or maximum of a function, we look for critical points where the derivative is zero:

$$\frac{df}{d\mathbf{x}} = \mathbf{0}$$

After finding critical points, we can use the second derivative test to determine their nature: 1. If $f''(\mathbf{x}) > 0$, the point is a local minimum. 2. If $f''(\mathbf{x}) < 0$, the point is a local maximum. 3. If $f''(\mathbf{x}) = 0$, the test is inconclusive.

Second Derivative Test Examples:

Case 1: Local Minimum
Consider the function $f(x) = x^2 - 4x + 5$. Find the critical points and use the second derivative test.

$$\frac{df}{dx} = 2x - 4$$

Setting $\frac{df}{dx} = 0$, we get $x = 2$. Now, calculate the second derivative:

$$\frac{d^2f}{dx^2} = 2$$

Since $f''(x) > 0$, the critical point $x = 2$ corresponds to a local minimum.

Case 2: Local Maximum
Consider the function $f(x) = -x^2 + 4x - 5$. Find the critical points and use the second derivative test.

$$\frac{df}{dx} = -2x + 4$$

Setting $\frac{df}{dx} = 0$, we get $x = 2$. Now, calculate the second derivative:

$$\frac{d^2f}{dx^2} = -2$$

Since $f''(x) < 0$, the critical point $x = 2$ corresponds to a local maximum.

Case 3: Inconclusive Test
Consider the function $f(x) = x^3$. Find the critical points and use the second derivative test.

$$\frac{df}{dx} = 3x^2$$

Setting $\frac{df}{dx} = 0$, we get $x = 0$. Now, calculate the second derivative:

$$\frac{d^2 f}{dx^2} = 6x$$

At $x = 0$, $f''(x) = 0$. The second derivative test is inconclusive.

Hessian Matrix: Understanding the Hessian Matrix and its Role in Optimization

In optimization, the Hessian matrix emerges as a key player, wielding its influence on the trajectory of algorithms seeking to find optimal solutions. At its core, the Hessian matrix is a square matrix of second-order partial derivatives. For a function $f(x_1, x_2, \ldots, x_n)$, the Hessian matrix $\mathbf{H}$ is composed of the second partial derivatives of f with respect to its variables:

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

Each element $\mathbf{H}_{ij}$ in the Hessian matrix represents the second mixed partial derivative of f with respect to x_i and x_j.

Geometric Interpretation:

To grasp the essence of the Hessian matrix, let's consider its geometric interpretation. Imagine you're traversing mountainous terrain, and f represents the elevation at each point. The first-order partial derivatives of f (gradient vector) guide you in the steepest uphill direction. Now, the Hessian matrix plays a crucial role in describing the curvature of the terrain. For a point where the second partial derivatives are large, indicating a concave upward shape, the Hessian matrix captures this information. Conversely, if the second partial derivatives are small or negative, suggesting a concave downward shape, the Hessian matrix reflects this curvature.

Positive Definiteness:

A significant property of the Hessian matrix is its positive definiteness. If the Hessian matrix is positive definite at a critical point, it signifies that the point is a local minimum. In optimization problems, finding such minima is often the goal.

Mathematically, a symmetric matrix $\mathbf{A}$ is positive definite if, for any nonzero vector $\mathbf{v}$, the quadratic form $\mathbf{v}^T \mathbf{A} \mathbf{v}$ is positive:

$$\mathbf{v}^T \mathbf{A} \mathbf{v} > 0$$

The positive definiteness of the Hessian matrix at a critical point provides a reliable indication that the point is a minimum.

Negative Definiteness:

Conversely, the negative definiteness of the Hessian matrix at a critical point implies a local maximum. The quadratic form $\mathbf{v}^T \mathbf{A} \mathbf{v}$ would now be negative for any nonzero vector $\mathbf{v}$.

Understanding the signs of the eigenvalues of the Hessian matrix is a practical way to determine positive or negative definiteness. Positive eigenvalues indicate positive definiteness, while negative eigenvalues signify negative definiteness.

Indefiniteness:

If the Hessian matrix has both positive and negative eigenvalues, the critical point is saddle-shaped, indicating neither a minimum nor a maximum. This scenario brings forth the concept of indefiniteness. The Hessian matrix, with its eigenvalues acting as messengers of curvature, thus becomes a valuable tool for optimization algorithms seeking to navigate the landscape of a multi-dimensional function.

Role in Optimization:

Optimization algorithms, ranging from classic gradient descent to more sophisticated quasi-Newton methods, leverage the information encapsulated in the Hessian matrix to guide their steps towards optimal solutions.

Quadratic Approximation:

One notable application of the Hessian matrix is in quadratic approximation. Around a critical point, the function f can often be well-approximated by a quadratic form. The Hessian matrix plays a central role in constructing this quadratic approximation.

For a function f and a point $\mathbf{x}_0$, the quadratic approximation is given by:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0)^T \nabla f(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

This quadratic form, with the Hessian matrix at its core, provides a local model of the function around the critical point, aiding optimization algorithms in making informed decisions about the direction and step size.

Example:

Let's delve into an example to illustrate the role of the Hessian matrix in optimization. Consider the function $f(x, y) = x^2 + 4y^2 + 4xy$. We aim to find the critical points and analyze the definiteness of the Hessian matrix at these points.

To find the critical points, we need to solve for $\nabla f = \mathbf{0}$:

$$\frac{\partial f}{\partial x} = 2x + 4y = 0$$
$$\frac{\partial f}{\partial y} = 8y + 4x = 0$$

Solving this system of equations, we find the critical point $(x, y) = (0, 0)$.

Now, let's calculate the Hessian matrix:

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

Evaluating the second partial derivatives and substituting the critical point $(0, 0)$, we get:

$$\mathbf{H} = \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}$$

Now, let's find the eigenvalues of the Hessian matrix to determine its definiteness. The characteristic equation is given by:

$$\det(\mathbf{H} - \lambda \mathbf{I}) = 0$$

Solving this equation, we find the eigenvalues to be $\lambda_1 = 0$ and $\lambda_2 = 10$. Since we have one positive and one zero eigenvalue, the Hessian matrix is indefinite, indicating that the critical point $(0, 0)$ is a saddle point.

This example showcases how the Hessian matrix aids in characterizing critical points and provides insights into the nature of the optimization landscape. Hessian matrix stands as a powerful tool, offering insights into the curvature of functions and guiding algorithms towards optimal solutions. Its role in determining the definiteness of critical points shapes the foundation of optimization strategies, enabling efficient exploration of complex, multi-dimensional spaces.

Convex Functions and Convex Optimization

Convex optimization is a fascinating field that finds applications in various domains, including machine learning, statistics, and operations research. At the core of this field lies the concept of convex functions. Let's explore what convex functions are and how they play a crucial role in optimization.

Convex Functions

A function is considered convex if, roughly speaking, the line segment between any two points on its graph lies above the graph itself. Mathematically, a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if, for all x, y in its domain and for all λ in the interval $[0, 1]$, the following inequality holds:

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

This inequality essentially means that the function lies below the line segment connecting any two points on its graph. Convex functions have several important properties that make them significant in optimization.

Properties of Convex Functions

1. Global Minima: Convex functions have a desirable property - any local minimum is also a global minimum. This makes optimization problems involving convex functions more manageable.

2. Optimality Conditions: A critical point of a convex function is also a global minimum. This simplifies the process of finding optimal solutions in convex optimization.

Examples of Convex Functions

Let's explore a few examples of convex functions:

1. Linear Functions: Any linear function $f(x) = ax + b$ is convex.

2. Quadratic Functions: Quadratic functions $f(x) = ax^2 + bx + c$ are convex if $a \geq 0$.

3. Exponential Functions: The exponential function $f(x) = e^{ax}$ is convex for any a.

4. Norms: The p-norms (L^p norms) are convex. For example, the L^2 norm $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}$ is convex.

Convex Optimization

Convex optimization deals with the optimization of convex functions subject to convex constraints. The general form of a convex optimization problem is:

$$\text{Minimize } f_0(x)$$

$$\text{Subject to } f_i(x) \leq 0, \quad i = 1, 2, \ldots, m$$

$$h_i(x) = 0, \quad i = 1, 2, \ldots, p$$

Where f_0, f_i, h_i are convex functions. The convexity of the objective function and constraints simplifies the optimization process, making it possible to find global optima efficiently.

Convex optimization problems have widespread applications, such as linear programming, quadratic programming, and support vector machines in machine learning.

Hessian in Convex Optimization

In convex optimization, the Hessian matrix plays a crucial role, especially when dealing with multivariable convex functions. The Hessian provides information about the local behavior of the function, helping us determine whether a critical point is a minimum, maximum, or a saddle point.

Hessian Matrix

For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian matrix, denoted by $\nabla^2 f$ or H, is the matrix of second partial derivatives. If $f(x)$ has continuous second partial derivatives, the Hessian matrix is given by:

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

The Hessian is symmetric if the mixed partial derivatives are continuous.

Positive Definiteness

One of the key properties of the Hessian is positive definiteness. A symmetric matrix A is positive definite if, for any nonzero vector v, the following inequality holds:

$$v^T A v > 0$$

For convex functions, a critical point x^* is a local minimum if and only if the Hessian at that point, $\nabla^2 f(x^*)$, is positive definite.

Graphical Representation

Let's visualize the concept of convexity with a graphical representation. Consider a simple convex function:

$$f(x) = x^2$$

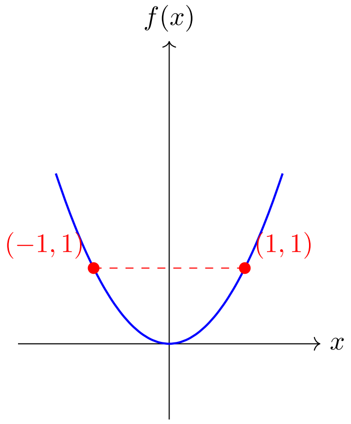


Figure 1: Graph of the convex function $f(x) = x^2$

In the graph, you can observe that any line segment between two points on the curve lies above the curve itself, confirming the convexity of the function.

Hessian in Convex Optimization

The Hessian matrix plays a crucial role in convex optimization, particularly in determining the convexity of a function. The Hessian matrix, denoted as H_f, is the matrix of second-order partial derivatives of a function f with respect to its variables. For a function $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian matrix is given by:

$$
H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}
$$

Convex Optimization Techniques

Convex optimization encompasses a variety of techniques tailored for efficiently finding optimal solutions. These techniques can be broadly categorized into constrained and unconstrained optimization.

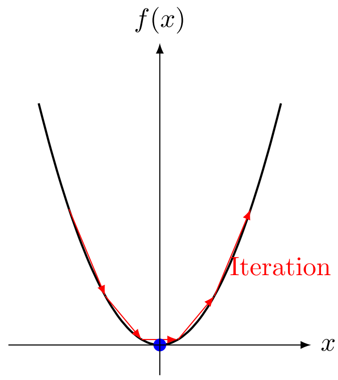
Unconstrained Optimization

In unconstrained optimization, the goal is to find the minimum (or maximum) of a convex function without any constraints on the variables. Two well-known algorithms for unconstrained optimization are:

1. Gradient Descent: This iterative method utilizes the gradient of the objective function to iteratively update the solution. The update rule is given by:

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

where α is the step size. Gradient descent is a simple yet powerful algorithm for unconstrained optimization. It iteratively moves towards the minimum of the convex function using the negative gradient. Let's visualize the steps of gradient descent.

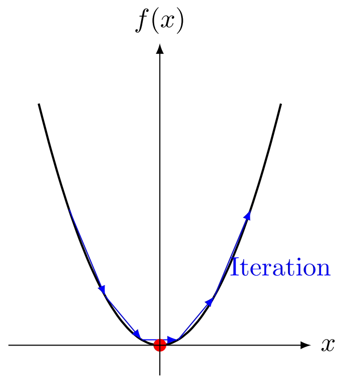


Gradient Descent for Unconstrained Optimization. In this diagram, each red arrow represents a step taken by the gradient descent algorithm towards the optimal point (marked in blue). More in the next note.

2. Newton's Method: Building upon gradient descent, Newton's method incorporates second-order information through the Hessian matrix. The update rule is given by:

$$x_{k+1} = x_k - \alpha(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

Newton's method converges faster than gradient descent. Let's visualize Newton's method:



In this diagram, each blue arrow represents a step taken by Newton's method towards the optimal point (marked in red).

Constrained Optimization

Constrained optimization deals with finding the optimal solution within a set of constraints. Linear programming and the Kuhn-Tucker conditions are integral components of this category.

Linear Programming (LP)

Linear programming involves optimizing a linear objective function subject to linear constraints. Consider the following LP problem:

$$\text{Minimize } c^T x$$

$$\text{Subject to } Ax \leq b, \quad x \geq 0$$

Here, c is the cost vector, A is the coefficient matrix, b is the right-hand side vector, and x is the vector of decision variables. The feasible region formed by the constraints is a convex polytope, and the optimal solution lies at a vertex of this polytope.

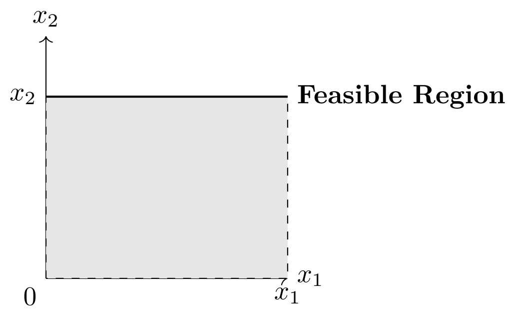


Figure 2: Feasible region for a linear programming problem

In this example, the feasible region is the convex polytope defined by the constraints, and the optimization goal is to find the point within this region that minimizes the linear objective function.

Kuhn-Tucker Conditions

For more general constrained optimization problems, the Kuhn-Tucker conditions come into play. These conditions extend the concept of Lagrange multipliers to handle inequality constraints.

Consider the optimization problem:

$$\text{Minimize } f(x)$$

$$\text{Subject to } g_i(x) \leq 0, \quad i = 1, 2, \ldots, m$$

$$h_j(x) = 0, \quad j = 1, 2, \ldots, p$$

The Kuhn-Tucker conditions are a set of conditions that characterize the optimal solutions to this problem. They involve a combination of the gradient of the Lagrangian and complementary slackness conditions. Consider the problem:

$$\text{Minimize } f(x)$$

$$\text{Subject to } g_i(x) \leq 0, \quad h_j(x) = 0$$

The Kuhn-Tucker conditions state that at an optimal solution x^*, there exist Lagrange multipliers λ_i and ν_j such that:

1. Stationarity: $\nabla f(x^*) + \sum_i \lambda_i \nabla g_i(x^*) + \sum_j \nu_j \nabla h_j(x^*) = 0$
2. Primal Feasibility: $g_i(x^*) \leq 0$ for all i
3. Dual Feasibility: $\lambda_i \geq 0$ for all i
4. Complementary Slackness: $\lambda_i g_i(x^*) = 0$ for all i

These conditions ensure that the optimal solution satisfies both the primal and dual problem requirements.

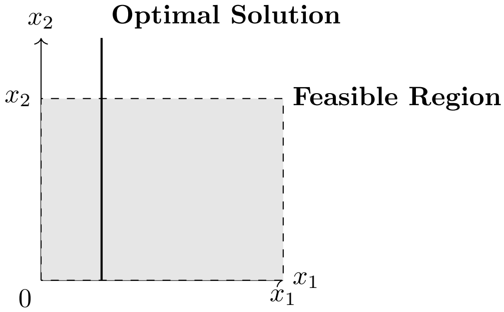


Figure 3: Optimal solution satisfying Kuhn-Tucker conditions

This diagram illustrates an optimal solution within the feasible region, satisfying the Kuhn-Tucker conditions.

Properties of Convex and Smooth Functions

Lipschitz Continuity

A Lipschitz continuous function is one for which the rate at which the function can change is bounded. Specifically, a function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L if, for all x, y in its domain, the following inequality holds:

$$|f(x) - f(y)| \leq L\|x - y\|$$

This condition provides a measure of the smoothness of the function, indicating how quickly the function values can change with respect to changes in the input.

Consider the function $f(x) = |x|$. This function is Lipschitz continuous with Lipschitz constant $L = 1$ since $|f(x) - f(y)| \leq \|x - y\|$.

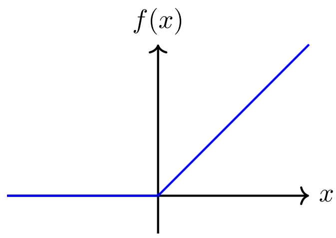


Figure 4: Lipschitz Continuous Function $f(x) = |x|$

Strong Convexity

A strongly convex function is more than just convex; it has a curvature that is strictly positive. Specifically, a function $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with strong convexity parameter μ if, for all x, y in its domain and for all t in the interval $[0, 1]$, the following inequality holds:

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\mu}{2}t(1-t)\|x - y\|^2$$

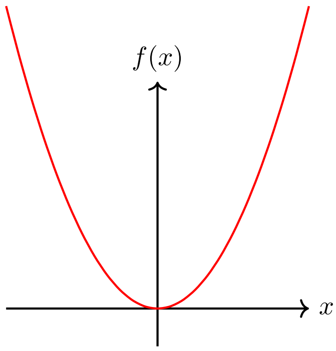


Figure 5: Strongly Convex Function $f(x) = x^2$

Strong convexity provides a stronger control over the growth of the function compared to convexity alone.

Polyak-Lojasiewicz Condition

The Polyak-Lojasiewicz (PL) condition is a property that certain nonconvex functions can possess. A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the PL condition if there exist constants α and c such that, for all x in its domain, the following inequality holds:

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \alpha(f(x) - c)$$

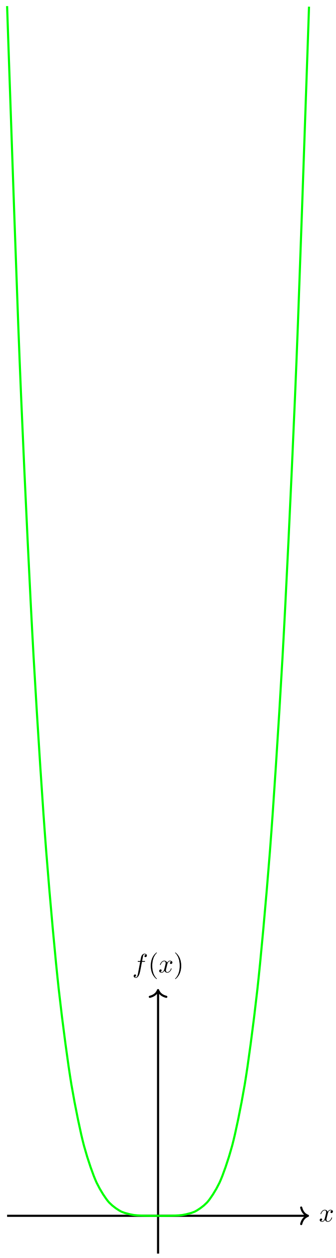


Figure 6: PL Condition Satisfying Function $f(x) = x^4$

The PL condition implies a form of smoothness and plays a crucial role in the convergence analysis of optimization algorithms.

Smoothness

Smoothness of a function refers to the behavior of its gradient. A function $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth if its gradient is L-Lipschitz continuous. Mathematically, for all x, y in its domain, the following inequality holds:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

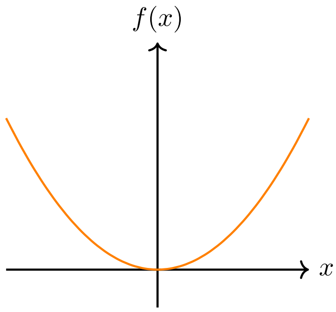


Figure 7: Smooth Function $f(x) = \frac{1}{2}x^2$

Smoothness is a desirable property in optimization as it ensures a controlled rate of change for the gradient.

Smoothness and Nonconvexity

Smoothness alone does not guarantee convexity. Nonconvex functions can still be smooth. Consider the example of the nonconvex function $f(x) = x^4$. It is smooth, but not convex.

Smoothness and Convexity

When a function is both smooth and convex, optimization becomes particularly well-behaved. Convexity ensures a unique global minimum, and smoothness ensures that optimization algorithms converge efficiently.

Convergence for Convex and Smooth Functions

For convex and smooth functions, optimization algorithms exhibit strong convergence properties. Gradient-based algorithms, such as gradient descent, have convergence guarantees, reaching the global minimum in a finite number of steps.

Convergence for Strongly Convex and Smooth Functions

If a function is both strongly convex and smooth, the convergence is even more pronounced. Algorithms converge at an accelerated rate, often quadratically, towards the global minimum.

Convergence for Polyak-Lojasiewicz and Smooth Functions

Functions satisfying the Polyak-Lojasiewicz condition and being smooth allow for convergence guarantees in nonconvex settings. Although not as strong as convexity, the PL condition ensures convergence to a critical point.

For in-depth mathematical details on these advanced topics, I highly recommend reading this paper: https://arxiv.org/abs/2301.11235. Now, let's jump to the tutorial and talk about SGD.