Machine Learning
Optimizations for Machine Learning

Dr. Pratyay Kuila

Dept. of Computer Science & Engineering
NIT Sikkim, Ravangla-737139

Introduction

- Many situations arise in machine learning where we would like to optimize the value of some function.

- That is, given a function $f : \mathbb{R}^n \to \mathbb{R}$, we want to find $x \in \mathbb{R}^n$ that minimizes (or maximizes) $f(x)$.

Classifications of Optimization Problems

- **Based on Variables**
  - Single-variable optimization
  - Multi-variable optimization
- **Based on Constraints**
  - Constrained
  - Unconstrained
- **Based on Nature of the Equations Involved**
  - Linear
  - Nonlinear
  - Geometric
  - Quadratic programming problems
- **Based on the Number of Objective Functions**
  - Single objective
  - Multi-objective

Unconstrained Optimization

- Suppose *f*(*x*) is a univariate function with continuous first-order and second order derivatives.
- In an *unconstrained optimization problem*, the task is to locate the solution x^* that maximizes or minimizes *f*(*x*) without imposing any constraints on x^*.
- The solution x^*, which is known as a *stationary point*, can be found by taking the first derivative of *f* and setting it to zero:

$$\frac{df}{dx}\Big|_{x=x^*} = 0$$

- *f*(x^*) can take a maximum or minimum value depending on the second-order derivative of the function:
  - x^* is a maximum stationary point if $\frac{d^2 f}{dx^2} < 0$ at $x = x^*$
  - x^* is a minimum stationary point if $\frac{d^2 f}{dx^2} > 0$ at $x = x^*$
  - x^* is a point of inflection when $\frac{d^2 f}{dx^2} = 0$ at $x = x^*$

Unconstrained Optimization

- This concept can be extended to a multivariate function, $f(x_1, x_2, ..., x_d)$, where the condition for finding a stationary point $\mathbf{x}^* = [x_1^*, x_2^*, ..., x_d^*]^T$ is

$$\frac{\partial f}{\partial x_i}\Big|_{x_i = x_i^*} = 0, \forall i = 1, 2, ..., d.$$

- We need to consider the partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ for all possible pairs of i and j.
- The complete set of second-order partial derivatives is given by the Hessian matrix.

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

Unconstrained Optimization

- A Hessian matrix **H** is *positive definite* if and only if $\mathbf{x}^T\mathbf{H}\mathbf{x} > 0$ for any non-zero vector **x**.
  - If $\mathbf{H}(\mathbf{x}^*)$ is *positive definite*, then $\mathbf{x}^*$ is a minimum stationary point.

- A Hessian is *negative definite* if and only if $\mathbf{x}^T\mathbf{H}\mathbf{x} < 0$ for any non-zero vector **x**.
  - If $\mathbf{H}(\mathbf{x}^*)$ is *negative definite*, then $\mathbf{x}^*$ is a maximum stationary point.

- A Hessian is *indefinite* if $\mathbf{x}^T\mathbf{H}\mathbf{x}$ is positive for some value of **x** and negative for others.
  - A stationary point with *indefinite* Hessian is a **saddle point**, which can have a minimum value in one direction, and a maximum value in another.

Unconstrained Optimization

Example. Suppose $f(x, y) = 3x^2+2y^3-2xy$. The conditions for finding the stationary points of this function are

$$\frac{\partial f}{\partial x} = 6x - 2y = 0$$
$$\frac{\partial f}{\partial y} = 6y^2 - 2x = 0$$

whose solutions are $x^* = y^* = 0$ or $x^* = \frac{1}{27}, y^* = \frac{1}{9}$.

The Hessian of f is $\mathbf{H}(x, y) = \begin{bmatrix} 6 & -2 \\ -2 & 12y \end{bmatrix}$

At $x = y = 0$, $\mathbf{H}(0, 0) = \begin{bmatrix} 6 & -2 \\ -2 & 0 \end{bmatrix}$

- Therefore, $\begin{bmatrix} x & y \end{bmatrix} \mathbf{H}(x, y) \begin{bmatrix} x & y \end{bmatrix}^T = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 6x - 2y \\ -2x \end{bmatrix} = 6x^2 - 4xy$

Unconstrained Optimization

- Since, $\begin{bmatrix} x & y \end{bmatrix} \mathbf{H}(x,y) \begin{bmatrix} x & y \end{bmatrix}^T = 6x^2 - 4xy = 2x(3x - 2y)$ which can be either positive or negative.
- The Hessian is indefinite and (0, 0) is a saddle point.

- At x = 1/27, y = 1/9, $\mathbf{H}(x,y) = \begin{bmatrix} 6 & -2 \\ -2 & \frac{12}{19} \end{bmatrix}$

Since, $\begin{bmatrix} x & y \end{bmatrix} \mathbf{H}(x,y) \begin{bmatrix} x & y \end{bmatrix}^T = 4x^2 - 2xy + \frac{4y^2}{3} = 4(x - \frac{y}{4})^2 + \frac{13y^2}{4} > 0$ for non-zero x and y, the Hessian is positive definite.

- Therefore, (1/27, 1/9) is a minimum stationary point.

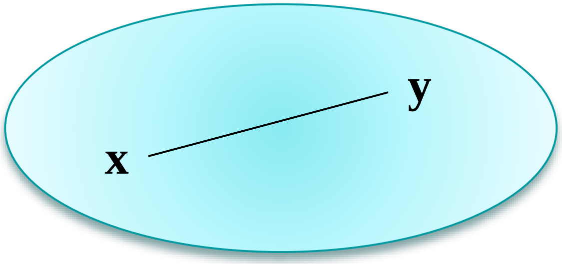
- The minimum value of f is -0.0014.

Convex Optimization

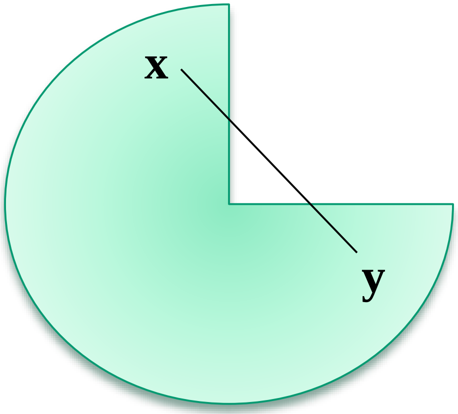
Definition (Convex set): *A set $X \subseteq \mathbb{R}^N$ is said to be convex if for any two points* $\mathbf{x}, \mathbf{y} \in X$ *the segment* $[\mathbf{x}, \mathbf{y}]$ *lies in* X*, that is*

$$\{\alpha\mathbf{x} + (1-\alpha)\mathbf{y} : \forall\alpha, 0 \le \alpha \le 1\} \subseteq X.$$

- In other words, it is impossible to find a pair of points in the set such that any of the points on the straight line joining them do not lie in the set.
- A circle, an ellipse, a square, or a half-moon are all convex sets.
- However, a three-quarter circle is not a convex set because one can draw a line between the two points inside the set, so that a portion of the line lies outside the set.



CONVEX SET



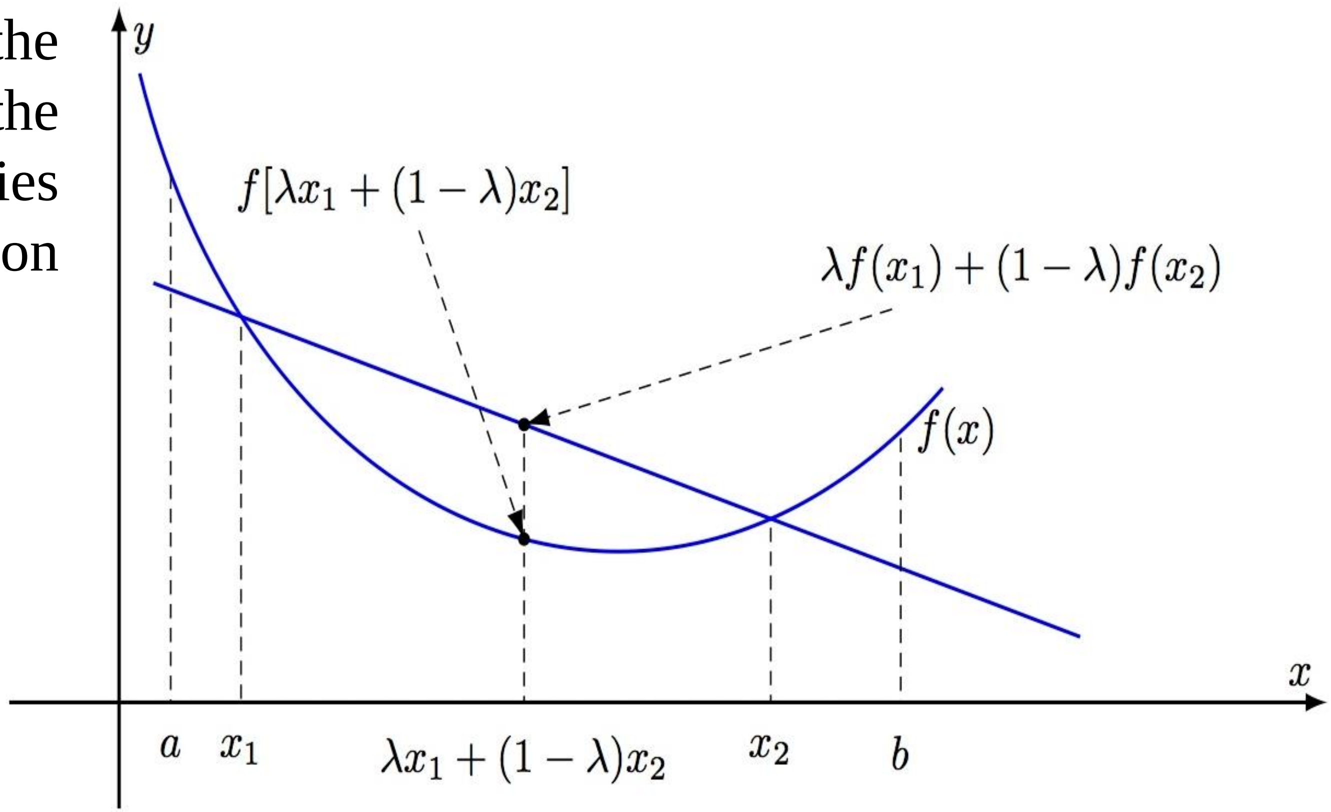
NON-CONVEX SET

Convex Optimization

Definition (Convex function): *Let X be a convex set. A function $f : X \to \mathbb{R}$ is said to be convex if for all $\mathbf{x}, \mathbf{y} \in X$ and $\lambda \in [0, 1]$,*

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

That is, if the segment joining the two points lies entirely above or on the graph of $f(\mathbf{x})$.



Convex Optimization

Definition (Convex function): *Let X be a convex set. A function $f : X \rightarrow \mathbb{R}$ is said to be convex if for all* $\mathbf{x}, \mathbf{y} \in X$ *and* $\lambda \in [0, 1]$,

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

- A function $f(x)$ is *strictly convex* if,

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- The convexity of a function is tested by checking the Hessian matrix of the function.

- *If the Hessian matrix is positive-definite or positive-semidefinite for all values of x in the search space, the function is a convex function.*

Convex Optimization

Definition (Convex function): *Let X be a convex set. A function $f : X \to \mathbb{R}$ is said to be convex if for all* $\mathbf{x}, \mathbf{y} \in X$ *and* $\lambda \in [0, 1]$,

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

- A function $f(x)$ is defined as a *concave function* if the function $-f(x)$ is a convex function. Therefore,

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- A function $f(x)$ is *strictly concave* if,

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) > \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

- An optimization problem is called a *convex optimization problem* if the objective function and the constraint functions (if any) are convex.

Gradient Descent

Definition (Gradient): *Let $f : X \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$ be a differentiable function. Then, the* gradient *of f at $\mathbf{x} \in X$ is the vector in $\mathbb{R}^N$ denoted by $\nabla f(\mathbf{x})$ and defined by*

$$\nabla f(\vec{x}) = \left[\frac{\partial f(\vec{x})}{\partial x_1} \quad \frac{\partial f(\vec{x})}{\partial x_2} \quad \cdots \quad \frac{\partial f(\vec{x})}{\partial x_N} \right]$$

- Notice $\nabla f(x)$ is itself a vector, whose components are the partial derivatives of f with respect to each of the x_i.

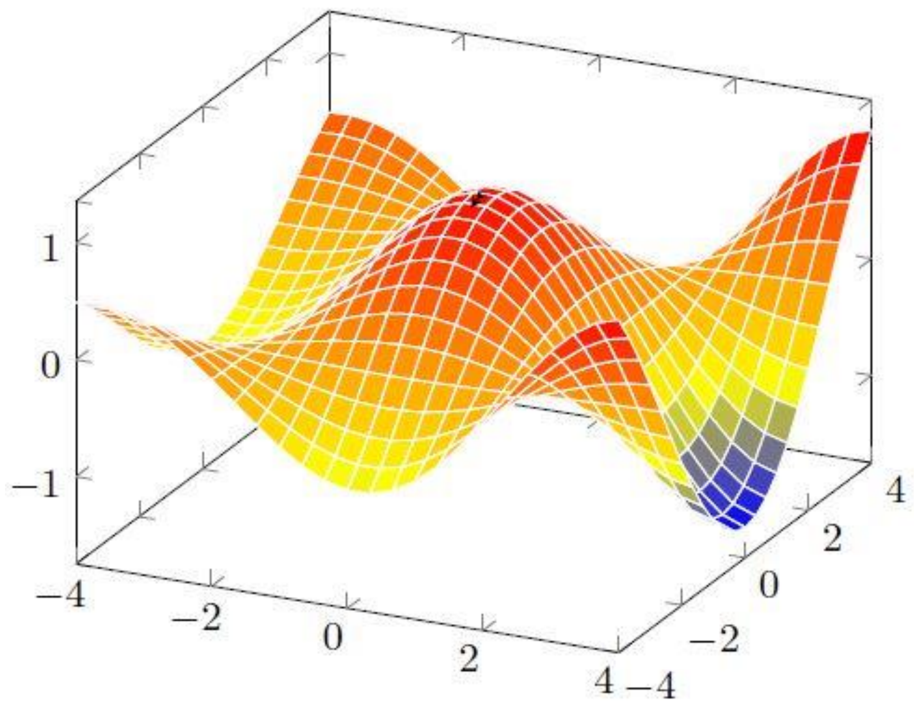
- When interpreted as a vector in weight space, *the gradient specifies the direction that produces the steepest increase in f.*

- The *negative of this vector therefore gives the direction of steepest decrease.*

Gradient Descent

Definition (Gradient): *Let $f : X \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$ be a differentiable function. Then, the* gradient *of f at* **x**∈*X is the vector in $\mathbb{R}^N$ denoted by* ∇*f*(**x**) *and defined by*

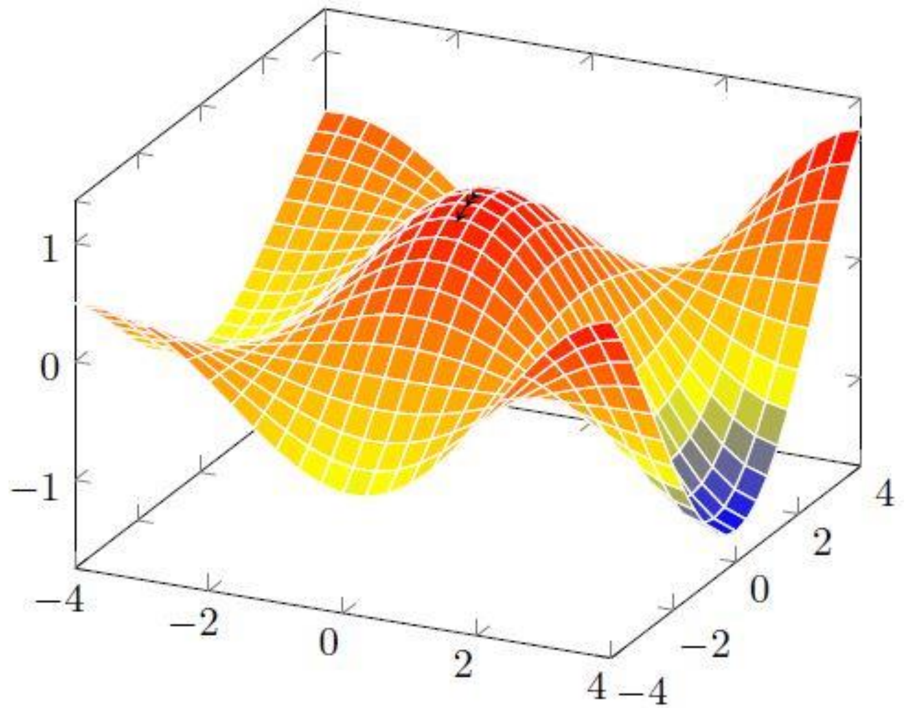
$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \frac{\partial f(\vec{x})}{\partial x_2} & \cdots & \frac{\partial f(\vec{x})}{\partial x_N} \end{bmatrix}$$



Gradient Descent

Definition (Gradient): *Let $f : X \subseteq \mathbb{R}^N \rightarrow \mathbb{R}$ be a differentiable function. Then, the* gradient *of f at* $\mathbf{x} \in X$ *is the vector in $\mathbb{R}^N$ denoted by* $\nabla f(\mathbf{x})$ *and defined by*

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} & \frac{\partial f(\vec{x})}{\partial x_2} & \cdots & \frac{\partial f(\vec{x})}{\partial x_N} \end{bmatrix}$$



Gradient Descent

- The gradient descent method assumes that the function $f(\mathbf{x})$ is differentiable and computes the stationary point as follows:

$$\vec{\mathrm{x}}_{[t+1]} = \vec{\mathrm{x}}_{[t]} - \eta \nabla f\left(\vec{\mathrm{x}}_{[t]}\right)$$

- Here η is a positive constant called the *learning rate*, which determines the step size in the gradient descent search.

- The *negative sign* is present because we want to move the weight vector in the direction that ***decreases*** f.

- In this method, the location of **x** is updated in the direction of the steepest descent, which means that **x** is moved towards the decreasing value of f.

Gradient Descent

- The gradient descent method assumes that the function $f(\mathbf{x})$ is differentiable and computes the stationary point as follows:

$$\vec{x}_{[t+1]} = \vec{x}_{[t]} - \eta \nabla f(\vec{x}_{[t]})$$

Algorithm 1 : Gradient Descent

Input:Initial weights $\vec{x}_{[0]}$, iterations T, learning rate η.
Output:Final weights $\vec{x}_{[T]}$.

1: **for** $t = 0$ to $T - 1$
2: Compute $\nabla f(\vec{x}_{[t]})$.
3: $\vec{x}_{[t+1]} = \vec{x}_{[t]} - \eta \nabla f(\vec{x}_{[t]})$.
4: **end for**
5: Return $\vec{x}_{[T]}$.

Gradient Descent

Example: Find the minimum of the following function with two variables: $f(x,y) = x^2 + 2y^2$.

The general form of the gradient vector is given by: $\nabla f(\vec{x}) = \begin{bmatrix} 2x & 4y \end{bmatrix}$.

Two iterations of the algorithm, T=2 and η = 0.1 are shown below:

- Initial $t = 0$: $\vec{x}_{[0]} = (4, 3)$ # This is just a randomly chosen point
- At $t = 1$ # At $t = 0$: $f(x,y) = 4^2 + 2\times3^2 = 34$

$$\vec{x}_{[1]} = \vec{x}_{[0]} - \eta\nabla f(\vec{x}_{[0]})$$
$$\vec{x}_{[1]} = (4, 3) - 0.1 \times (8, 12)$$
$$\vec{x}_{[1]} = (3.2, 1.8)$$

At $t = 1$: $f(x,y) = (3.2)^2+2\times(1.8)^2 = 16.72$

- At $t = 2$

$$\vec{x}_{[2]} = \vec{x}_{[1]} - \eta\nabla f(\vec{x}_{[1]})$$
$$\vec{x}_{[2]} = (3.2, 1.8) - 0.1 \times (6.4, 7.2)$$
$$\vec{x}_{[2]} = (2.56, 1.08)$$

At $t = 2$: $f(x,y) = (2.56)^2+2\times(1.08)^2 = 8.8864$

- If we keep running the above iterations, the procedure will eventually end up at the point where the function is minimum, i.e., (0,0).

Constrained Optimization

A *constrained optimization problem* comprises an objective function together with a number of equality and inequality constraints.

- **Equality Constraints:**

Consider the problem of finding the minimum value of $f(x_1, x_2,..., x_d)$ subjected to equality constraints of the form:

$$
\begin{aligned}
g_1(\mathbf{x}) &= 0 \\
g_2(\mathbf{x}) &= 0 \\
&\cdots \\
g_p(\mathbf{x}) &= 0
\end{aligned}
$$

- A method known as Lagrange multipliers can be used to solve the constrained optimization problem.
- Various other methods are also in the literature.

Constrained Optimization

- **Equality Constraints:**

Lagrange multipliers method involves the following steps:

1. Define the Lagrangian, $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{p} \lambda_i g_i(\mathbf{x})$ where λ_i is a dummy variable called the *Lagrange multiplier*.

2. Set the first-order derivatives of the Lagrangian with respect to $\mathbf{x}$ and the Lagrange multipliers to zero,

$$\frac{\partial L}{\partial x_1} = 0 \qquad \frac{\partial L}{\partial \lambda_1} = 0$$
$$\frac{\partial L}{\partial x_2} = 0 \qquad \frac{\partial L}{\partial \lambda_2} = 0$$
$$\cdots \qquad \cdots$$
$$\frac{\partial L}{\partial x_d} = 0 \qquad \frac{\partial L}{\partial \lambda_p} = 0$$

3. Solve the $(d + p)$ equations in step 2 to obtain the stationary point $\mathbf{x}^*$ and the corresponding values for λ_i's.

Constrained Optimization

➢ Equality Constraints:

Example (Lagrange multipliers): Let $f(x, y) = x + 2y$. Suppose we want to minimize the function $f(x, y)$ subject to the constraint $x^2 + y^2 - 4 = 0$.

1. First, we introduce the Lagrangian

$$L(x, y, \lambda) = (x + 2y) + \lambda(x^2 + y^2 - 4)$$

Only one Lagrange multiplier is used, as we have only one constraint.

2. Set the first-order derivatives:

$$\frac{\partial L}{\partial x} = 1 + 2\lambda x = 0 \qquad \frac{\partial L}{\partial y} = 2 + 2\lambda y = 0$$

$$\frac{\partial L}{\partial \lambda} = x^2 + y^2 - 4 = 0$$

3. Solving these equations yields, $\lambda = \pm\frac{\sqrt{5}}{4}, x = \mp\frac{2}{\sqrt{5}}$ and $y = \mp\frac{4}{\sqrt{5}}$

When, $\lambda = \frac{\sqrt{5}}{4}, f(-\frac{2}{\sqrt{5}}, -\frac{4}{\sqrt{5}}) = -\frac{10}{\sqrt{5}}$

When, $\lambda = -\frac{\sqrt{5}}{4}, f(\frac{2}{\sqrt{5}}, \frac{4}{\sqrt{5}}) = \frac{10}{\sqrt{5}}$

Thus, the function $f(x, y)$ has its minimum value at $x = -\frac{2}{\sqrt{5}}, y = -\frac{4}{\sqrt{5}}$

Constrained Optimization

Inequality Constraints:

Consider the problem of finding the minimum value of $f(x_1, x_2,..., x_d)$ subjected to inequality constraints of the form:

$$
\begin{aligned}
h_1(\mathbf{x}) &\leq 0 \\
h_2(\mathbf{x}) &\leq 0 \\
&\cdot\ \cdot\ \cdot \\
h_q(\mathbf{x}) &\leq 0
\end{aligned}
$$

- The method for solving this problem is quite similar to the Lagrange method.
- However, the inequality constraints impose additional conditions to the optimization problem. The optimization problem stated above leads to the following Lagrangian.

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{q} \lambda_i h_i(\mathbf{x})$$

Constrained Optimization

Inequality Constraints:

- The optimization problem stated above leads to the following Lagrangian.

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{q} \lambda_i h_i(\mathbf{x})$$

and constraints known as the Karush-Kuhn-Tucker (KKT) conditions:

$$\begin{aligned} \frac{\partial L}{\partial x_i} &= 0, \forall i = 1, 2, ..., d \\ h_i(\mathbf{x}) &\leq 0, \forall i = 1, 2, ..., q \\ \lambda_i &\geq 0, \forall i = 1, 2, ..., q \\ \lambda_i h_i(\mathbf{x}) &= 0, \forall i = 1, 2, ..., q \end{aligned}$$

Notice that the Lagrange multipliers are no longer unbounded in the presence of inequality constraints.

Constrained Optimization

Inequality Constraints:

Example (KKT Conditions): Suppose we want to minimize the function $f(x, y) = (x - 1)^2 + (y - 3)^2$ subject to the following constraints:

$$x + y \leq 2, \text{ and } y \geq x.$$

The Lagrangian for this problem is

$$L = (x-1)^2 + (y-3)^2 + \lambda_1(x+y-2) + \lambda_2(x-y)$$

subjected to the following KKT constraints:

$$\frac{\partial L}{\partial x} = 2(x-1) + \lambda_1 + \lambda_2 = 0$$

$$\frac{\partial L}{\partial y} = 2(y-3) + \lambda_1 - \lambda_2 = 0$$

$$\lambda_1(x+y-2) = 0$$

$$\lambda_2(x-y) = 0$$

$$\lambda_1 \geq 0, \lambda_2 \geq 0, x+y \leq 2, y \geq x$$

Constrained Optimization

Inequality Constraints:

Case 1: $\lambda_1 = 0, \lambda_2 = 0$. In this case, we obtain the following equations:

$$2(x - 1) = 0 \text{ and } 2(y - 3) = 0,$$

whose solution is given by $x = 1$ and $y = 3$. Since $x + y = 4$, this is not a feasible solution because it violates the constraint $x + y \leq 2$.

Case 2: $\lambda_1 = 0, \lambda_2 \neq 0$. In this case, we obtain the following equations:

$$x - y = 0, \qquad 2(x - 1) + \lambda_2 = 0, \qquad 2(y - 3) - \lambda_2 = 0,$$

whose solution is given by $x = 2$, $y = 2$, and $\lambda_2 = -2$, which is not a feasible solution because it violates the conditions $\lambda_2 \geq 0$ and $x+y \leq 2$.

Case 3: $\lambda_1 \neq 0, \lambda_2 = 0$. In this case, we obtain the following equations:

$$x + y - 2 = 0, \qquad 2(x - 1) + \lambda_1 = 0, \qquad -2(x + 1) + \lambda_1 = 0,$$

whose solution is given by $x = 0$, $y = 2$, and $\lambda_1 = 2$, which is a feasible solution.

Summary

- For convex optimization, the local optima is the global optima.
- In the case of differentiable unconstrained convex optimization problems, setting the gradient to "zero" provides a simple means for identifying candidate local optima.
- The Karush-Kuhn-Tucker (KKT) conditions are applicable for constraints optimization with inequality constraints.
- The Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient to reach to global optima for *convex optimization problems*.
- Gradient descent approach is applicable for *unconstrained convex optimization problems.*

Books:

1. Linear Algebra and Optimization for Machine Learning by Charu C. Aggarwal, Springer.
2. Introduction to Data Mining by PN Tang, M Steinbach, V Kumar, Pearson.
3. Convex Optimization by Stephen Boyd and Lieven Vandenberghe. Cambridge, 2004.