MAT 222 Linear Algebra
Week 11
Lecture Notes 1

Murat Karaçayır

Akdeniz University
Department of Mathematics

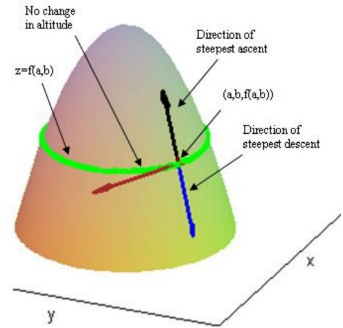
13rd May 2025

Optimization

- Given $z = f(x_1, x_2 \ldots, x_n)$, in many applications we need the maximum or minimum value of f. This type of problems are called **optimization**.
- A few examples are: Minimizing cost or maximizing profit, maximizing the distance traveled or minimizing the fuel consumption for a constant distance, adjusting the time for traffic lights at an intersection, fitting a curve to a data, etc.
- Sometimes we consider the problem in the whole domain of f (**unconstrained optimization**), and sometimes we consider it in a subset of the domain satisfying a certain condition (**unconstrained optimization**).
- In each case, the first derivative test plays an important role.
- We will mainly consider examples with $n = 2$ or $n = 3$, but the methods are also valid for other dimensions.

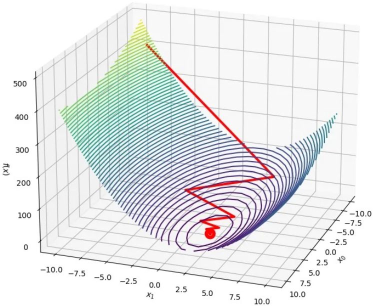
Gradient Descent Method

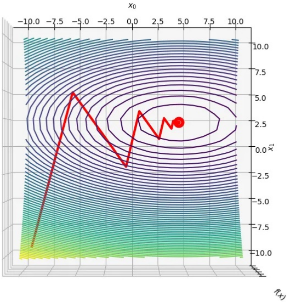
- **Gradient Descent Method** is one of the simplest methods used to compute the absolute minimum of a function of several variables.
- Let us demonstrate it for $n = 2$ variables. So we have $z = f(x, y)$.
- Problem: Minimize $f(x, y)$ over all (x, y) in the domain of f.
- Note that we can also use it to maximize $f(x, y)$. Because maximizing $f(x, y)$ is the same as minimizing $-f(x, y)$.



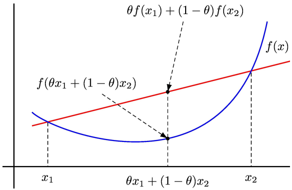
- The method is based on the following observation: The gradient of f at a certain point is the direction of the steepest increase.
- But this implies the following: The opposite direction of ∇f should be the direction of steepest descent.
- This shows that it is feasible to move in the direction of $-\nabla f$ to reach a minimum of f.

Gradient Descent Method





- Above you see a few steps of the gradient descent method. On the left is the graph of the function while you see the level curves on the right.



- The intuition is that starting from any point on the surface, if you let (x, y) in the direction of the steepest descent, which is $-\nabla f$, then eventually you will get arbitrarily close to the pit in the middle. This pit is the minimum we are looking for.
- If f is a **convex function**, gradient descent method always works.
- Such functions constitute a generalization of convex functions in one variable. (See the figure)

Gradient Descent Method: Algorithm

- Suppose we have a convex function $z = f(x, y)$. Our aim is to find a global (absolute) minimum of f. The steps are as follows:

1. Choose an initial point $\mathbf{x}_0 = (x_0, y_0)$.
2. Determine a tolerance $\varepsilon > 0$.
3. Choose a coefficient $\alpha > 0$ (known as the **learning rate**).
4. For $k = 0, 1, 2, \ldots$, update the point as $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k)$.
5. Update k as $k + 1$.
6. Stop when $\|\nabla(\mathbf{x}_k)\| < \varepsilon$.
7. $\mathbf{x}_k$ (called a **minimizer**) is the approximate location of the minimum of the function f and $f(\mathbf{x}_k)$ is approximately the minimum value of f.

Gradient Descent Method: Example

Example: Let us consider the function $f(x, y) = 1 + x^2 + 2y^2$. We know that there is a global minimum at $(0, 0)$ and the minimum value is $f(0, 0) = 1$. But let us try to find it using gradient descent method.

- **Initialization:** Choose $\mathbf{x}_0 = (4, 3)$ (It can be anything else). Set $\alpha = 0.1, \varepsilon = 0.0001$. Calculate the gradient of f. $\nabla f = (2x, 4y)$.
- **Step 1:** $\nabla f(\mathbf{x}_0) = \nabla f(4, 3) = (8, 12)$.
  $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (4, 3) - 0.1(8, 12) = (3.2, 1.8)$.
- **Step 2:** $\nabla f(\mathbf{x}_1) = \nabla f(3.2, 1.8) = (6.4, 7.2)$.
  $\mathbf{x}_2 = \mathbf{x}_1 - \alpha \nabla f(\mathbf{x}_1) = (3.2, 1.8) - 0.1(6.4, 7.2) = (2.56, 1.08)$.
- ...
- After 51 steps, we have $\mathbf{x}_{51} = (0.00004567192616, 0.00000000001454906299)$. The norm of its gradient is $\|\nabla f(\mathbf{x}_{51})\| = 0.0000456719 < \varepsilon = 0.0001$, so we stop here. $\mathbf{x}_{51}$ is the approximate value for the global minimum.

Exercise: Try to solve this example using $\alpha = 0.2$. Then use $\alpha = 0.5$. What do you observe?

Gradient Descent and Linear Regression

- Suppose we have experimental data represented by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we want to fit a regression line $\hat{y} = \beta_0 + \beta_1 x$ to this line.
- Consider the sum of the squares of the residuals given by

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- The above expression is a function of β_0 and β_1 so let us denote it by $f(\beta_0, \beta_1)$.
- Thus, we have the following unconstrained optimization problem:

  Minimize $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ over all $(\beta_0, \beta_1) \in \mathbb{R}^2$.
- f can be shown to be a convex function so gradient descent method is suitable for this problem.
- Note that fitting any other curve (a parabola, exponential function etc.) also gives rise to a convex function f, only the number of variables would be different in that case.

Linear Regression with Gradient Descent: Example

x_i	1	2	2.5	3	3.5	4.5	4.7	5.2	6.1	6.1	6.8
y_i	1.5	1	2	2	3.7	3	5	4	5.8	5	5.7

- Suppose we want to fit a least squares regression line $y = \beta_0 + \beta_1 x$ to the above $n = 11$ data points.
- We can do it by solving the 11×11 system determined by the normal equations, but let us do it using gradient descent method.
- The function to minimize is the squares of the residuals

$$f(\beta_0, \beta_1) = \sum_{i=1}^{11} (y_i - \beta_0 - \beta_1 x_i)^2$$

- For this example we set the learning rate $\alpha = 0.001$, the tolerance $\varepsilon = 10^{-6}$ and the initial point $\mathbf{x}_0 = (-1, 1)$.
- After 4492 steps the algorithm stops and outputs the solution $\hat{\beta} = (-0.022514383, 0.85787792)^T$.
- Thus the least squares regression line that best fits this data is

$$y = -0.022514383 + 0.85787792x$$

Exercise: Observe what happens using the values $\alpha = 10^{-4}$ and $\alpha = 0.01$.