Optimization For Machine Learning

Majid Almarhoumi

Gradient Descent

Introduction

In this model, our goal is to:

Introduction

In this model, our goal is to:

- Calculate approximate value of functions using derivatives.

Introduction

In this model, our goal is to:

- Calculate approximate value of functions using derivatives.
- Calculate approximate value of derivatives using functions.

Introduction

In this model, our goal is to:

- Calculate approximate value of functions using derivatives.
- Calculate approximate value of derivatives using functions.
- Understand the meaning of derivatives in higher dimensions.

Introduction

In this model, our goal is to:

- Calculate approximate value of functions using derivatives.
- Calculate approximate value of derivatives using functions.
- Understand the meaning of derivatives in higher dimensions.
- calculate the gradient descent algorithm in one dimension.

Introduction

In this model, our goal is to:

- Calculate approximate value of functions using derivatives.
- Calculate approximate value of derivatives using functions.
- Understand the meaning of derivatives in higher dimensions.
- calculate the gradient descent algorithm in one dimension.
- Understand the gradient descent algorithm in multiple dimensions.

Optimization and gradient descent may sound like big concepts. We can think of them in this way:

Optimization and Gradient Descent

Optimization and gradient descent may sound like big concepts. We can think of them in this way:
Optimization: Getting the best possible value from a function (maximum or minimum).

Optimization and Gradient Descent

Optimization and gradient descent may sound like big concepts. We can think of them in this way:
Optimization: Getting the best possible value from a function (maximum or minimum).
This can be seen from the fish tank example yesterday where the fish owner wanted to minimize the cost of glass she had to pay for to get a tank with a volume of 62.5 in^3.

Optimization and Gradient Descent

Optimization and gradient descent may sound like big concepts. We can think of them in this way:
Optimization: Getting the best possible value from a function (maximum or minimum).
This can be seen from the fish tank example yesterday where the fish owner wanted to minimize the cost of glass she had to pay for to get a tank with a volume of 62.5 in^3.
Gradient Descent: Getting the computer to find that best value for us. Thus, making our lives much easier!

Fish Tank I

Fish Tank I

In our fish tank problem, we wanted the volume to be $62,5$. Thus we had:

Fish Tank I

In our fish tank problem, we wanted the volume to be $62,5$. Thus we had:

$$x^2 y = 62.5$$

Fish Tank I

In our fish tank problem, we wanted the volume to be $62,5$. Thus we had:

$$x^2y = 62.5$$

However, we also wanted to minimize the surface area of the tank (one base and four sides). So we want to minimize:

Fish Tank I

In our fish tank problem, we wanted the volume to be $62, 5$. Thus we had:

$$x^2 y = 62.5$$

However, we also wanted to minimize the surface area of the tank (one base and four sides). So we want to minimize:

$$f(x, y) = x^2 + 4xy$$

Fish Tank I

In our fish tank problem, we wanted the volume to be $62,5$. Thus we had:

$$x^2 y = 62.5$$

However, we also wanted to minimize the surface area of the tank (one base and four sides). So we want to minimize:

$$f(x, y) = x^2 + 4xy$$

Or we can substitute $y = 62.5/x^2$ to get:

$$f(x) = x^2 + \frac{4 \times 62.5}{x}$$

Fish Tank I

From our calculus knowledge, we need to equate the derivative of $f(x)$ to zero to get the best possible value of x. Therefore:

Fish Tank I

From our calculus knowledge, we need to equate the derivative of $f(x)$ to zero to get the best possible value of x. Therefore:

$$\frac{\partial f(x)}{\partial x} = 2x - \frac{250}{x^2} = 0$$

Fish Tank I

From our calculus knowledge, we need to equate the derivative of $f(x)$ to zero to get the best possible value of x. Therefore:

$$\frac{\partial f(x)}{\partial x} = 2x - \frac{250}{x^2} = 0$$

Multiply by x^2 and divide by 2 to get:

$$x^3 - 125 = 0$$

Fish Tank I

From our calculus knowledge, we need to equate the derivative of $f(x)$ to zero to get the best possible value of x. Therefore:

$$\frac{\partial f(x)}{\partial x} = 2x - \frac{250}{x^2} = 0$$

Multiply by x^2 and divide by 2 to get:

$$x^3 - 125 = 0$$

Simple calculation here shows that the best value for x is 5.

Fish Tank I

We can see that this is in fact the case by noting that the derivative of $f(x)$ is a flat line at $x = 5$:

Fish Tank I

We can see that this is in fact the case by noting that the derivative of $f(x)$ is a flat line at $x = 5$:

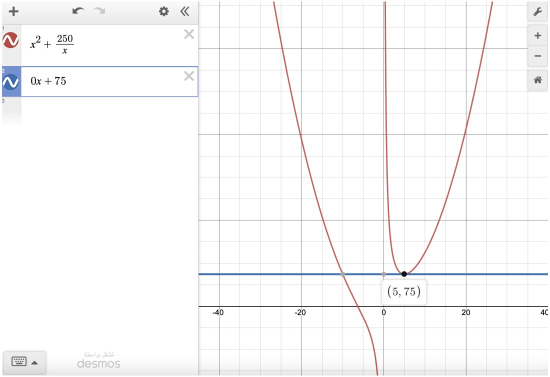

Figure: visualization $f(x)$ and the tangent line at $x = 5$. One can see that the minimum value for $f(x)$ is 75.

Fish Tank II

The route we took before to solve the optimization problem was purely mathematical. Now, how can we get the computer to do all that stuff for us? Here is a simple gradient descent code implemented that shows us the best value of x:

Fish Tank II

```
x = 3
for i in range(30):
  x=x-0.02*(2*x-250/x/x)
  print(x)
```

```
3.4355555555555557
3.7217529400804534
3.933855993083802
4.099598875017223
4.233115128060058
4.342819845080659
4.434217214369401
4.511142504623798
4.576392139890364
4.632075548785568
4.679826229739367
4.720935618851382
4.756441878828203
4.787190993197505
4.813880116881221
4.837089149885119
4.857304254702394
4.874935716560974
4.890331739199433
4.903789260810716
4.915562545678371
4.925870088497689
4.934900220036111
4.942815700083474
4.949757511249499
4.9558480152995825
4.961193596001637
4.965886884647884
```

Figure: Note how the values of x get closer and closer to $x = 5$.

Fish Tank II

Now that we have seen that the computer is able to solve these kind of problems for us. The question becomes how does it do that? What are the tools we need to understand to implement these kind of codes?

Derivatives and Approximations

The first tool we need to understand is how derivatives work inside the computer. Granted, the computer cannot compute derivatives like we do. Then, how can we make the computer calculate a derivative of a function?

Derivatives and Approximations

The first tool we need to understand is how derivatives work inside the computer. Granted, the computer cannot compute derivatives like we do. Then, how can we make the computer calculate a derivative of a function? From the definition of the derivative, we know that:

$$\frac{\partial f(x)}{\partial x} \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

From this law, try to answer the following questions:

- Given $f(4) = 20$ and $f(4.1) = 20.91$ compute the derivative of $f(4)$.

Derivatives and Approximations

The first tool we need to understand is how derivatives work inside the computer. Granted, the computer cannot compute derivatives like we do. Then, how can we make the computer calculate a derivative of a function? From the definition of the derivative, we know that:

$$\frac{\partial f(x)}{\partial x} \approx \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

From this law, try to answer the following questions:

- Given $f(4) = 20$ and $f(4.1) = 20.91$ compute the derivative of $f(4)$.
- Given $f(x) = x^2$ approximate $f(4.1)$ using $f(4)$ and the derivative of $f(4)$.

Derivatives and Approximations

The first tool we need to understand is how derivatives work inside the computer. Granted, the computer cannot compute derivatives like we do. Then, how can we make the computer calculate a derivative of a function? From the definition of the derivative, we know that:

$$\frac{\partial f(x)}{\partial x} \approx \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

From this law, try to answer the following questions:

- Given $f(4) = 20$ and $f(4.1) = 20.91$ compute the derivative of $f(4)$.
- Given $f(x) = x^2$ approximate $f(4.1)$ using $f(4)$ and the derivative of $f(4)$.

Now, can you think of a way of how to calculate derivatives on the computer?

About Derivatives

There are key things to keep in mind when thinking about derivatives:

- When the derivative of $f(x_0)$ is negative, then $f(x_0 + \Delta x) < f(x_0)$.

About Derivatives

There are key things to keep in mind when thinking about derivatives:

- When the derivative of $f(x_0)$ is negative, then $f(x_0 + \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is positive, then $f(x_0 - \Delta x) < f(x_0)$.

About Derivatives

There are key things to keep in mind when thinking about derivatives:

- When the derivative of $f(x_0)$ is negative, then $f(x_0 + \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is positive, then $f(x_0 - \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is zero, then $f(x_0)$ is either minimum or maximum.

About Derivatives

There are key things to keep in mind when thinking about derivatives:

- When the derivative of $f(x_0)$ is negative, then $f(x_0 + \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is positive, then $f(x_0 - \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is zero, then $f(x_0)$ is either minimum or maximum.
- We can estimate the derivative of $f(x_0)$ by the following identity:

$$\left.\frac{\partial f(x)}{\partial x}\right|_{x_0} \approx \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

About Derivatives

There are key things to keep in mind when thinking about derivatives:

- When the derivative of $f(x_0)$ is negative, then $f(x_0 + \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is positive, then $f(x_0 - \Delta x) < f(x_0)$.
- When the derivative of $f(x_0)$ is zero, then $f(x_0)$ is either minimum or maximum.
- We can estimate the derivative of $f(x_0)$ by the following identity:

$$\left.\frac{\partial f(x)}{\partial x}\right|_{x_0} \approx \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

- This identity can be fed directly to the computer to evaluate the derivative at any point x_0 we want. This could be useful some cases where the derivative is hard to compute.

About Derivatives

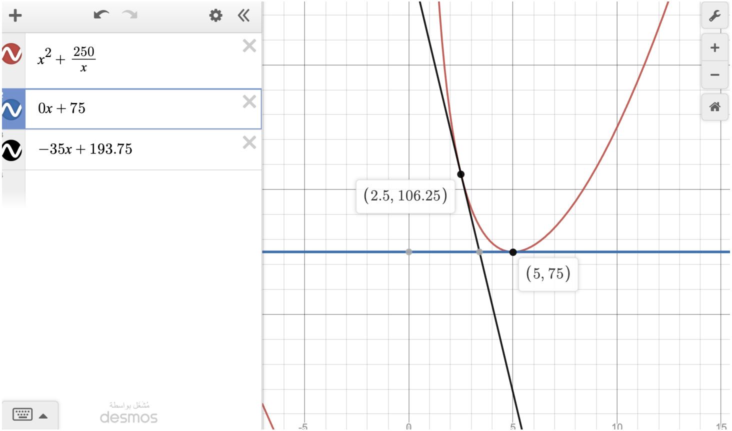

Figure: Example of negative derivative. Note that the derivative of $f(2.5)$ is -35 and it can be seen clearly that $f(2.6) < f(2.5)$.

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

1. Define the function $f(x)$ that you want to optimize.

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

1. Define the function $f(x)$ that you want to optimize.
2. Choose a starting point x_0 (can be your favourite number).

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

1. Define the function $f(x)$ that you want to optimize.
2. Choose a starting point x_0 (can be your favourite number).
3. Calculate the gradient of $f(x)$ at this point, by any of the methods we took.

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

1. Define the function $f(x)$ that you want to optimize.
2. Choose a starting point x_0 (can be your favourite number).
3. Calculate the gradient of $f(x)$ at this point, by any of the methods we took.
4. If we want to minimize $f(x)$, then make a scaled move to the opposite direction of the derivative and pick a new point there.

Gradient Descent Algorithm

As mentioned before, gradient descent is a method in which we ask the computer to find the best possible value for us, which means we have to break the method into simpler steps which can be done by the computer:

1. Define the function $f(x)$ that you want to optimize.
2. Choose a starting point x_0 (can be your favourite number).
3. Calculate the gradient of $f(x)$ at this point, by any of the methods we took.
4. If we want to minimize $f(x)$, then make a scaled move to the opposite direction of the derivative and pick a new point there.
5. repeat steps 3,4 until we get a gradient close to 0.

Gradient Descent Algorithm

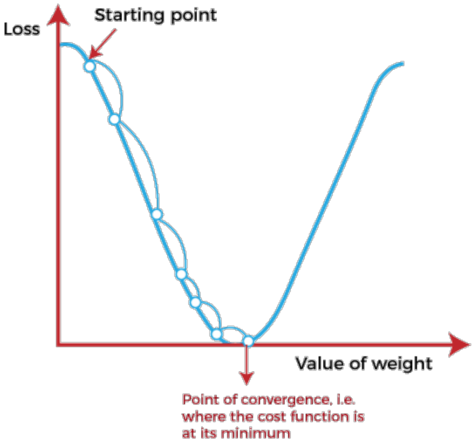

Figure: Example of gradient descent. Note that we start at point where the derivative is negative and we move to the right until we get to the minimum.

Gradient descent example

Now that we understand the steps for gradient descent, let us do an example:

Gradient descent example

Now that we understand the steps for gradient descent, let us do an example:
A new amusement park is about to open in Jeddah. The engineers would like to build a fun wiggly ride that shows you all the cool places in the park. The ride they came up with had the following route:

$$g(x) = t \sin x + 3t \cos(2x).$$

Gradient descent example

Now that we understand the steps for gradient descent, let us do an example:
A new amusement park is about to open in Jeddah. The engineers would like to build a fun wiggly ride that shows you all the cool places in the park. The ride they came up with had the following route:

$$g(x) = t\sin x + 3t\cos(2x).$$

They also know that the cool places are located at the points $(x, y) = \{(1, -3), (3, 6), (5, -7), (7, 2), (9, 5), (11, -8)\}$.

Gradient descent example

Now that we understand the steps for gradient descent, let us do an example:
A new amusement park is about to open in Jeddah. The engineers would like to build a fun wiggly ride that shows you all the cool places in the park. The ride they came up with had the following route:

$$g(x) = t \sin x + 3t \cos(2x).$$

They also know that the cool places are located at the points $(x, y) = \{(1, -3), (3, 6), (5, -7), (7, 2), (9, 5), (11, -8)\}$. Unfortunately, they do not know what is the best t that would pass along most of the cool places. Can you help them?

Gradient descent example

Now that we understand the steps for gradient descent, let us do an example:
A new amusement park is about to open in Jeddah. The engineers would like to build a fun wiggly ride that shows you all the cool places in the park. The ride they came up with had the following route:

$$g(x) = t \sin x + 3t \cos(2x).$$

They also know that the cool places are located at the points $(x, y) = \{(1, -3), (3, 6), (5, -7), (7, 2), (9, 5), (11, -8)\}$. Unfortunately, they do not know what is the best t that would pass along most of the cool places. Can you help them? play with this graph to better understand the problem:
`https://www.desmos.com/calculator/jyp7hitvh2`
`https://colab.research.google.com/drive/1r2DWXCumlL_gfWj1Ks-cnNBB1R6tplT_#scrollTo=iJeGCgYxx883`