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:

$$x^2y = 62.5$$

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

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

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 $x = 5$:

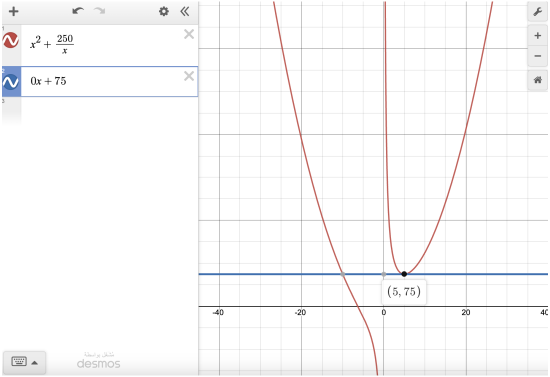

Figure: visualization $f(x)$ and the tangent line at $x = 5$. One can see that the minimum value for $f(x)$ is 75.

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

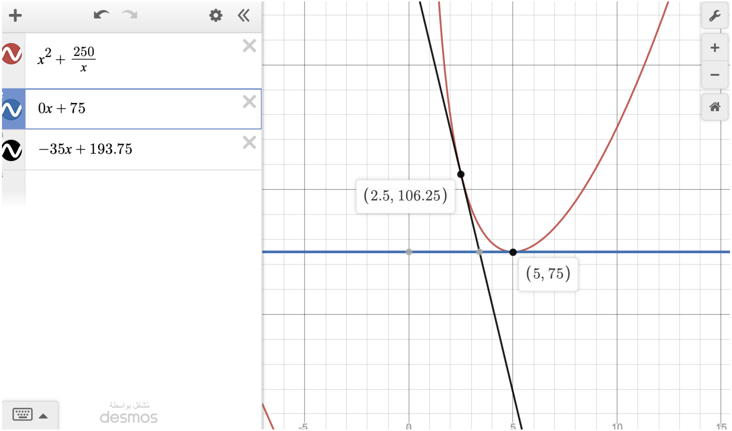

Figure: Example of negative derivative. Note that the derivative of $f(2.5)$ is -35 and it can be seen clearly that $f(2.6) < f(2.5)$.

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

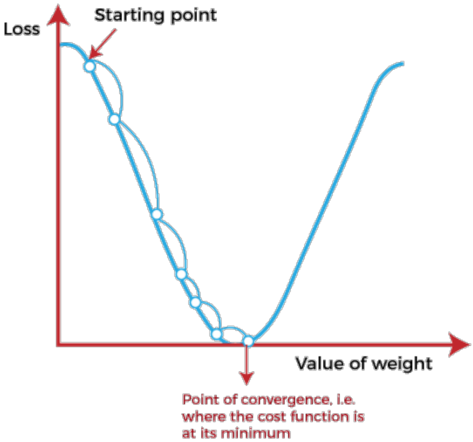

Figure: Example of gradient descent. Note that we start at point where the derivative is negative and we move to the right until we get to the minimum.

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:

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

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:

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

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:

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

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:

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

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`

The engineers thank you for your help in finding the best t to get the excellent ride.

2-D gradient Descent

The engineers thank you for your help in finding the best t to get the excellent ride. However, they believe they could do a little bit better. Therefore, they come up with the following route:

$$g(x, t_1, t_2) = t_1 \sin(x) + 3t_2 \cos(2x).$$

2-D gradient Descent

The engineers thank you for your help in finding the best t to get the excellent ride. However, they believe they could do a little bit better. Therefore, they come up with the following route:

$$g(x, t_1, t_2) = t_1 \sin(x) + 3t_2 \cos(2x).$$

Now, they want you to find the best (t_1, t_2) to get the best ride.

2-D gradient Descent

The engineers thank you for your help in finding the best t to get the excellent ride. However, they believe they could do a little bit better. Therefore, they come up with the following route:

$$g(x, t_1, t_2) = t_1 \sin(x) + 3t_2 \cos(2x).$$

Now, they want you to find the best (t_1, t_2) to get the best ride. Do you think this change would result in better or worse ride? Take five minutes to think about it and elaborate.

2-D Gradient Descent

This change usually leads to better results. That is because in worst case scenario, we can pick $t_1 = t_2$ and that would give us the initial model.

2-D Gradient Descent

This change usually leads to better results. That is because in worst case scenario, we can pick $t_1 = t_2$ and that would give us the initial model. So having this extra degree of freedom, we know that we cannot do worse than the initial model.

2-D Gradient Descent

This change usually leads to better results. That is because in worst case scenario, we can pick $t_1 = t_2$ and that would give us the initial model. So having this extra degree of freedom, we know that we cannot do worse than the initial model.
The question now is. How do we find the new loss function? And how do we take its derivative?

Multivariable Functions I

In the calculus section. We took derivatives of functions that depend on one variable x.

Multivariable Functions I

In the calculus section. We took derivatives of functions that depend on one variable x. One example may look like:

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

And

$$\frac{\partial f(x)}{\partial x} = 2x.$$

Now, we would like to think of derivatives of functions that depend on two variables, take this function for example:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Multivariable Functions I

In the calculus section. We took derivatives of functions that depend on one variable x. One example may look like:

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

And

$$\frac{\partial f(x)}{\partial x} = 2x.$$

Now, we would like to think of derivatives of functions that depend on two variables, take this function for example:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Here is a visualization of this function in the 3-dimensional space.
`https://www.geogebra.org/3d/hv4e7ety`

Multivariable functions II

Given:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Multivariable functions II

Given:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Calculate $f(3, 2)$, $f(3, 3)$ and $f(-2, 3)$.

Multivariable functions II

Given:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Calculate $f(3, 2)$, $f(3, 3)$ and $f(-2, 3)$. To calculate the minimum and maximum of this function, we need to define partial derivatives.

Partial derivatives I

Partial derivative: the Partial derivative of multivariable function is the derivative with respect to one of its variables.

Partial derivatives I

Partial derivative: the Partial derivative of multivariable function is the derivative with respect to one of its variables. Thus, if a function depends on two variables, it has two partial derivatives. for our example:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

Partial derivatives I

Partial derivative: the Partial derivative of multivariable function is the derivative with respect to one of its variables. Thus, if a function depends on two variables, it has two partial derivatives. for our example:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

has the partial derivative with respect to x:

$$\frac{\partial f(x, y)}{\partial x} = 2(x - y)$$

Partial derivatives I

Partial derivative: the Partial derivative of multivariable function is the derivative with respect to one of its variables. Thus, if a function depends on two variables, it has two partial derivatives. for our example:

$$f(x, y) = (x - y)^2 + (y - 1)^2 - 1.$$

has the partial derivative with respect to x:

$$\frac{\partial f(x, y)}{\partial x} = 2(x - y)$$

and the partial derivative with respect to y:

$$\frac{\partial f(x, y)}{\partial y} = 2(x - y)(0 - 1) + 2(y - 1)$$

Based on this example, can you think of the similarities of partial derivatives to regular derivatives?

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply.

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) = a$
- $\frac{\partial}{\partial x}(x^n + y^n) =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) = a$
- $\frac{\partial}{\partial x}(x^n + y^n) = nx^{n-1}$
- $\frac{\partial}{\partial x} xy =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) = a$
- $\frac{\partial}{\partial x}(x^n + y^n) = nx^{n-1}$
- $\frac{\partial}{\partial x} xy = y$
- $\frac{\partial}{\partial x} e^x + e^y =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) = a$
- $\frac{\partial}{\partial x}(x^n + y^n) = nx^{n-1}$
- $\frac{\partial}{\partial x} xy = y$
- $\frac{\partial}{\partial x} e^x + e^y = e^x$
- $\frac{\partial}{\partial x} x^n y^m =$

Partial Derivatives II

When taking the partial derivative with respect to one variable, we treat all other variables as constants. and the same rules of single variable derivatives apply. Some of these rules are (a, b, c are constants, x, y are variables):

- $\frac{\partial}{\partial x} c = 0$
- $\frac{\partial}{\partial x} y = 0$
- $\frac{\partial}{\partial x}(ax + by + c) = a$
- $\frac{\partial}{\partial x}(x^n + y^n) = nx^{n-1}$
- $\frac{\partial}{\partial x} xy = y$
- $\frac{\partial}{\partial x} e^x + e^y = e^x$
- $\frac{\partial}{\partial x} x^n y^m = nx^{n-1} y^m$

Try to calculate these derivatives with respect to y on your own time.

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$. Do you remember how did we do it for one variable functions $f(x)$?

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.
Do you remember how did we do it for one variable functions $f(x)$?
For multivariable functions, it is a similar process.

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.
Do you remember how did we do it for one variable functions $f(x)$?
For multivariable functions, it is a similar process.

- Take the partial derivative with respect to x and equate it to zero.

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.
Do you remember how did we do it for one variable functions $f(x)$?
For multivariable functions, it is a similar process.

- Take the partial derivative with respect to x and equate it to zero.
- Take the partial derivative with respect to y and equate it to zero.

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.
Do you remember how did we do it for one variable functions $f(x)$?
For multivariable functions, it is a similar process.

- Take the partial derivative with respect to x and equate it to zero.
- Take the partial derivative with respect to y and equate it to zero.
- Solve the system of equations by any method we know. (We could use linear algebra methods from before).

So our example becomes:

Optimization of Multivariable Functions

Now, we would like to find (x, y) that would minimize or maximize $f(x, y)$.
Do you remember how did we do it for one variable functions $f(x)$?
For multivariable functions, it is a similar process.

- Take the partial derivative with respect to x and equate it to zero.
- Take the partial derivative with respect to y and equate it to zero.
- Solve the system of equations by any method we know. (We could use linear algebra methods from before).

So our example becomes:

$$\begin{cases} \frac{\partial f(x,y)}{\partial x} = 2(x-y) = 0; \\ \frac{\partial f(x,y)}{\partial y} = 2(x-y)(0-1) + 2(y-1) = 0. \end{cases} \tag{1}$$

Can you solve this system of equations and find (x, y) that minimize the value of $f(x, y)$?

For a multivariable function $f(x, y)$, we define the gradient ∇ of f as the vector (from Linear Algebra) formed by the partial derivatives:

Partial Derivatives, Gradients and Multivariable Functions

For a multivariable function $f(x, y)$, we define the gradient ∇ of f as the vector (from Linear Algebra) formed by the partial derivatives:

$$\nabla f(x, y) = (\frac{\partial}{\partial x} f(x, y), \frac{\partial}{\partial y} f(x, y)) = (2x - 2y, 4y - 2x - 2).$$

Partial Derivatives, Gradients and Multivariable Functions

For a multivariable function $f(x, y)$, we define the gradient ∇ of f as the vector (from Linear Algebra) formed by the partial derivatives:

$$\nabla f(x, y) = (\frac{\partial}{\partial x} f(x, y), \frac{\partial}{\partial y} f(x, y)) = (2x - 2y, 4y - 2x - 2).$$

This gradient is defined at every point of the space (for differentiable functions) and always points to the direction of the highest increase.

Partial Derivatives, Gradients and Multivariable Functions

For a multivariable function $f(x, y)$, we define the gradient ∇ of f as the vector (from Linear Algebra) formed by the partial derivatives:

$$\nabla f(x, y) = (\frac{\partial}{\partial x} f(x, y), \frac{\partial}{\partial y} f(x, y)) = (2x - 2y, 4y - 2x - 2).$$

This gradient is defined at every point of the space (for differentiable functions) and always points to the direction of the highest increase. Calculate the gradient of $f(2, 3), f(1, 1), f(3, 3)$.

Partial Derivatives, Gradients and Multivariable Functions

For a multivariable function $f(x, y)$, we define the gradient ∇ of f as the vector (from Linear Algebra) formed by the partial derivatives:

$$\nabla f(x, y) = (\frac{\partial}{\partial x} f(x, y), \frac{\partial}{\partial y} f(x, y)) = (2x - 2y, 4y - 2x - 2).$$

This gradient is defined at every point of the space (for differentiable functions) and always points to the direction of the highest increase.
Calculate the gradient of $f(2, 3), f(1, 1), f(3, 3)$.
Can you think of how we could use the gradient in the amusement park problem?

Final Notes on Gradient Descent algorithm

Final Notes on Gradient Descent algorithm

Quick quiz: Why did we name the algorithm Gradient Descent?
Since the gradient points to the direction of the highest increase. What we do is compute the gradient at some point, and slide (descend) in the opposite direction of the gradient to get to a smaller value.
See the animations on this website for clearer visualization of Gradient Descent.
`https://towardsdatascience.com/a-visual-explanation-of-gradient-descent-methods-momentum-ada`

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. https://www.desmos.com/calculator/hdjqa0hrys //

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss?

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss? What happens when t_3 is very large?

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss? What happens when t_3 is very large? Do you think it is feasible in real life?

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss? What happens when t_3 is very large? Do you think it is feasible in real life? Can you say something about the role of $g_1(x)$ and $g_2(x)$ that are being graphed?

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss? What happens when t_3 is very large? Do you think it is feasible in real life? Can you say something about the role of $g_1(x)$ and $g_2(x)$ that are being graphed? Can you think why would the engineers add t_3^2 in $g_2(x)$?

More Problems!

Going back to the amusement park problem. The engineers ask you if it would be feasible to have a route that looks like the following:

$$g(x, t_1, t_2, t_3) = t_1 \sin(x) + 3t_2 \cos(2t_3 x).$$

What do you think of this? take 3 minutes and share your thoughts. Look at the following graph of the function g, play with it for a bit. `https://www.desmos.com/calculator/hdjqa0hrys` // Do you think that this new route g is going to have smaller loss? What happens when t_3 is very large? Do you think it is feasible in real life? Can you say something about the role of $g_1(x)$ and $g_2(x)$ that are being graphed? Can you think why would the engineers add t_3^2 in $g_2(x)$? Let us go back to the amusement park problem and continue solving by picking the optimal (t_1, t_2, t_3) to minimize the loss.
`https://colab.research.google.com/drive/1r2DWXCumlL_gfWj1Ks-cnNBB1R6tplT_#scrollTo=uokIjv5py8sg`

Does Gradient Descent Always Work?

Now that we have studied Gradient Descent, do you think that it would work in all optimization problems? Can you reason why or give an example of why it should not?

Does Gradient Descent Always Work?

Now that we have studied Gradient Descent, do you think that it would work in all optimization problems? Can you reason why or give an example of why it should not?
In fact, gradient descent does not always work, and usually fails to work to optimize functions with more than one minimum.

Does Gradient Descent Always Work?

Now that we have studied Gradient Descent, do you think that it would work in all optimization problems? Can you reason why or give an example of why it should not?

In fact, gradient descent does not always work, and usually fails to work to optimize functions with more than one minimum. Look at the plot of the following function:

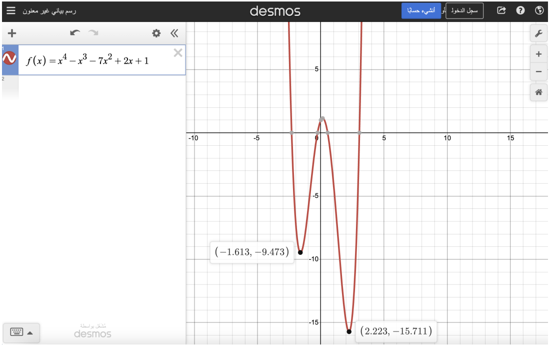

Figure: Example of function f with two minimum points. The gradient descent

Your Turn!

For the rest of the class, try to come up with a real life situation where someone would need to compute a minimum of a function. describe the situation, come up with the loss function, compute its derivative and find the minimizing values on paper.

Your Turn!

For the rest of the class, try to come up with a real life situation where someone would need to compute a minimum of a function. describe the situation, come up with the loss function, compute its derivative and find the minimizing values on paper. Finally, code the problem in python and make sure that your answers for the loss and the optimal values match.

Your Turn!

For the rest of the class, try to come up with a real life situation where someone would need to compute a minimum of a function. describe the situation, come up with the loss function, compute its derivative and find the minimizing values on paper. Finally, code the problem in python and make sure that your answers for the loss and the optimal values match. Take this problem home with you and do not hesitate to ask any questions!