Lecture 10

Math Foundations Team

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction

- We will look at nonlinear optimization concepts in this lecture.
- We already know how to compute gradient, but there are some minutiae of gradient descent that we need to address.
- Machine learning algorithms depend heavily on the correctness of the gradient since if the gradient is computed erroneously, the algorithms might fail to find the local or global optimum.
- We will also look into some challenges in non-linear optimization.

Optimization in machine learning

- A common loss function is $L(\boldsymbol{\theta}) = \sum_{i=1}^{i=n} \|\boldsymbol{\theta}^T \boldsymbol{x}_i - y_i\|^2$.
- Here $\boldsymbol{x}_i^T$ is the ith row vector in a matrix $\mathbf{X}$ consisting of n row vectors where each row vector represents a training point
- y_i contains the real-valued observation of the i^{th} training point.
- The above loss function occurs in least squared regression
- It represents the sum of squared differences between the observed values y_i in the data and the predicted values $\hat{y}_i = \boldsymbol{\theta}^T \boldsymbol{x}_i$.

Optimization of Additionally separable sum

- We can write the total objective function as

$$L(\theta) = \sum_{i=1}^{n} L_i(\theta)$$

- This type of linear separability is useful since it enables the use of techniques like stochastic gradient descent and mini-batch stochastic gradient descent.
- The idea here is that we can replace the gradient of the entire objective function with a sampled approximation.

Overfitting in machine learning

- In traditional optimization, we focus on updating parameters so that the objective function is minimized as much as possible.
- In machine learning, minimization of the objective function is performed over training data but the model is applied on test data which is unseen.
- We need to avoid the problem of overfitting the training data.

Overfitting in machine learning

- Suppose we have 4-dimensional data and one dependent variable, i,e output that is a function of the four inputs.
- Let the input parameters be x_1, x_2, x_3, x_4 and the output be y.
- We seek to learn parameters w_1, w_2, w_3, w_4, w_5 such that our prediction expression $\hat{y} = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5$ gives a good prediction.
- We would like to minimize the squared error $\sum_{i=1}^{n} \|y - \hat{y}\|_2^2$

Example of overfitting

- Consider the following data on 4 variables x_1, x_2, x_3, x_4 and associated output variable y.
- Let us say that this is a sample of real-life data where the output $y \approx x_1$.

x_1	x_2	x_3	x_4	y
61	2	3	0.1	49
40	0	4	0.5	40
68	0	10	1.0	70

Example of overfitting

- Minimizing squared error, we notice that one good solution is $w_1 = 1, w_2 = w_3 = w_4 = 0, w_5 = 0$.
- This solution does not give zero squared error with respect to the actual observations but gives an error close to zero.
- The solution $w_1 = 1, w_2 = w_3 = w_4 = 0, w_5 = 0$ is a very good one since it captures the real-life relationship between the output variable y and x_1.

Example of overfitting

- Consider $w_1 = 0, w_2 = 7, w_3 = 5, w_4 = 0, w_5 = 20$. This solution gives zero training error.
- It is a very poor solution since there is no dependence of the output variable y on x_1 while we know that there is actually a strong dependence between y and x_1.
- Therefore it will incur a high error on test-data.
- This example illustrates the idea that minimizing the loss function to the greatest extent may not be a good thing since the model may then perform poorly on real-life data.

Feature processing

- The loss function can have vastly different sensitivities to different model parameters and this can have an impact in the learning process.
- Consider a model where a person's wealth y is modeled in terms of his age x_1 and number of years of college education x_2.The formula for wealth y is $y = w_1 x_1^2 + w_2 x_2^2$.
- Age $[0, 100]$, number of years in college education $[0, 10]$.
- We have $\frac{\partial y}{\partial w_1} = x_1^2$ and $\frac{\partial y}{\partial w_2} = x_2^2$.
- Since x_1 and x_2 are generally very different in magnitude, we take small steps in respect of w_2 and large steps in respect of w_1.

Feature processing

- Taking small steps in w_2 and large steps in w_1 will make us go steadily towards the optimal value for w_2 but oscillate with respect to the optimal value of w_1, overshooting the target each time.
- This makes convergence very slow.
- It is therefore helpful to have features with similar variance.

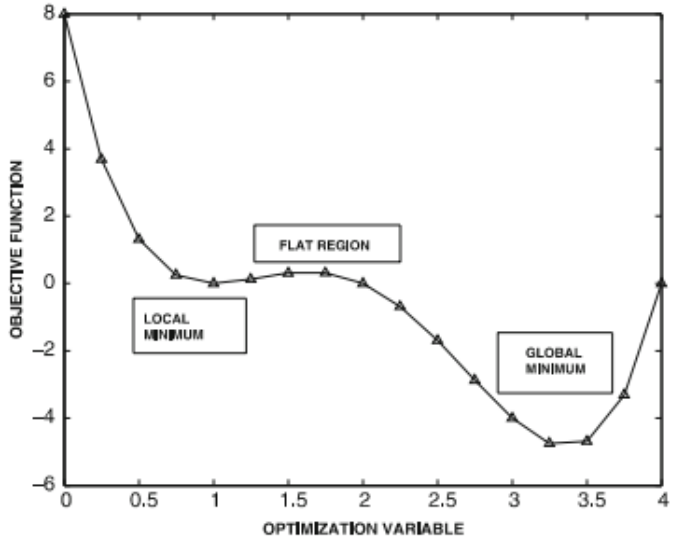
Feature processing

- Two techniques used for achieving similar variance are mean-centering and feature normalization.
- In case of mean-centering a vector of column-wise means is subtracted from each data point.
- In case of feature normalization, each feature value is divided by its standard deviation.
- In case of min-max normalization we scale the jth feature of the ith data point as follows: $x_{ij} = \frac{x_{ij} - \min_j}{\max_j - \min_j}$.

Challenges in Gradient-Based Optimization

innovate achieve lead

1. Flat regions and local optima
2. Different levels of curvature in different directions



(a) Local optima with flat regions

Local optima and flat regions

Consider a function defined as follows

- $F(x) = (x-1)^2((x-3)^2 - 1)$.
- $F'(x) = 2(x-1)((x-1)(x-3) + (x-3)^2 - 1) = 0$.
- The solutions to this equation are
  $x = 1, x = \frac{5-\sqrt{3}}{2} = 1.634, x = \frac{5+\sqrt{3}}{2} = 3.366$.

Local optima and flat regions

- We can show that the first and third roots are minima since $F''(x) > 0$ at these points while the second point is a maximum since $F''(x) < 0$.
- $F(1) = 0$, $F(\frac{5-\sqrt{3}}{2}) = 0.348$, $F(\frac{5+\sqrt{3}}{2}) = -4.848$.
- If we start gradient descent from any point less than 1.634, we will arrive only at a local minimum.
- We might never arrive at a global minimum if we keep choosing wrong starting points.
- The problem becomes worse with high-dimensionality.

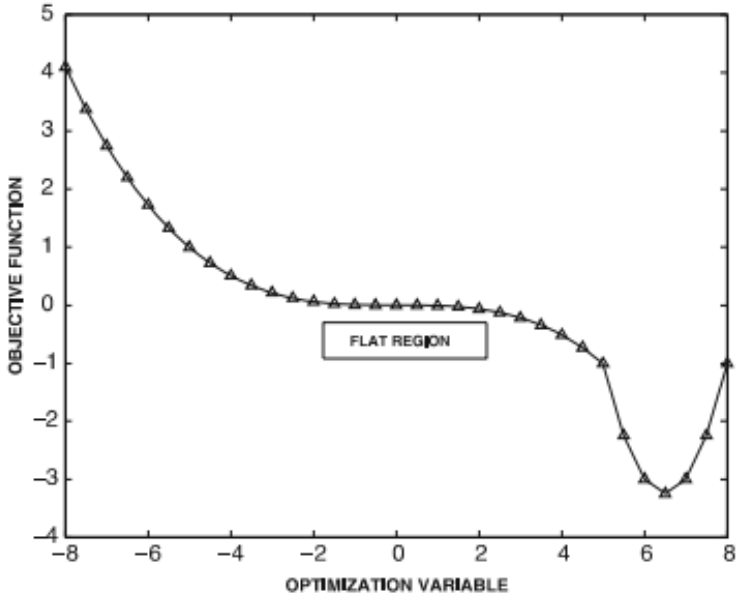
Local optima and flat regions

- Consider $F(x_1, x_2, \ldots x_d) = A_1(x_1) + A_2(x_2) + \ldots + A_n(x_n)$.
- Let $A_i(x_i)$ have k_i local minima.
- Setting $\frac{\partial F}{\partial x_i} = 0 \ \ \forall i$, we note that any point $(x_1^*, x_2^*, \ldots, x_d^*)$, where x_i^* is a local minima of the function $A_i(x_i)$, is a solution to $\frac{\partial F}{\partial x_i} = 0$.

Local optima and flat regions

- This is because $\frac{\partial F}{\partial x_i}|_{(x_1^*, x_2^*, \ldots x_d^*)} = A_i'(x_i^*) = 0$ since x_i^* is a local minimum of $A_i(x_i)$.
- There are $\prod_{i=1}^{i=d} k_i$ local minima for the function $F(x_1, x_2, \ldots, x_d)$, which is very large number of points.
- Gradient descent could be stuck at any one of these points which might be far from the global optimum.

Local optima and flat regions



(b) Only global optimum with flat region

- Another problem to contend with is the presence of flat regions where the gradient is close to zero.

Local optima and flat regions

- Flat regions are problematic because the speed of descent depends on the magnitude of the gradient, given a fixed learning rate.
- The optimization process will take a long time to cross a flat region of space which will make convergence slow.

Different levels of curvature

- In multi-dimensional settings, the components of the gradient with respect to different parameters can vary widely. This will cause convergence problems since there is oscillation in the update step with respect to some components and a steady movement with respect to other components.
- Consider the simplest possible case of a bowl-like convex, quadratic objective function with a single global minimum - $L = x^2 + y^2$ represents a perfectly circular bowl, and the function $L = x^2 + 4y^2$.
- We shall show contour plots of both functions and how gradient descent performs on finding the minimum of the two functions.

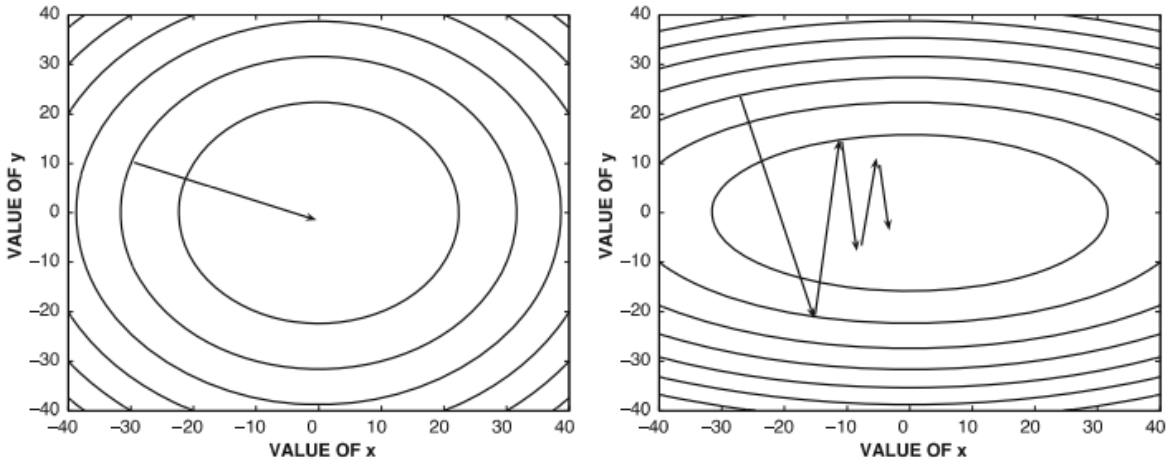
Different levels of curvature

- What is the qualitative difference between $L = x^2 + y^2$ and $L = x^2 + 4y^2$. Intuitively one looks symmetric in x and y, and the other is not.
- The second loss function is more sensitive to changes in y as compared to x - it looks like an elliptical bowl. The specific sensitivity depends on the position of x, y.
- Looking at the second-order derivatives we can see that for the second function $\frac{\partial^2 L}{\partial^2 x^2}$, and $\frac{\partial^2 L}{\partial^2 y^2}$ are very different.

Different levels of curvature

- The second-order derivative measures the rate of change of the gradient - a high second-order derivative means high curvature.
- From the point of view of gradient descent we want moderate curvature in all dimensions as it would means that the gradient does not change too much in some dimensions compared to others.
- We can then make gradient-descent steps of large sizes.

Different levels of curvature



(a) Loss function is circular bowl
$L = x^2 + y^2$

(b) Loss function is elliptical bowl
$L = x^2 + 4y^2$

Figure 5.2: The effect of the shape of the loss function on steepest-gradient descent

Different levels of curvature

- In case of the perfect bowl, a sufficiently large step-size from any point can take us directly to the optimum of the function in one-step, since the gradient at any point points towards the optimum of the function.
- This is not true for the elliptical bowl, the gradient at any point does not point to the optimum of the function.

Contour plots

- Note that the gradient at any point is orthogonal to the contour line at that point.
- This because the dot product of the gradient ∇F and a small displacement $\delta \boldsymbol{x}$ along the contour line gives the change in the value of the function along the displacement $\boldsymbol{x}$.
- Since the function remains constant along the contour line, $\nabla F.\boldsymbol{x} = 0$

Differential curvature

- A closer look at the contour plot for the elliptical bowl case shows that in the y-direction, we see oscillatory movement as in each step we correct the mistake of overshooting made in the previous step. The gradient component along the y-direction is more than the component along the x-direction.
- Along the x-direction, we make small movements towards the optimum x-value. Overall, after many training steps we find that we have made little progress to the optimum.
- It needs to be kept in mind that the path of steepest descent in most objective functions is only an instantaneous direction of best improvement, and is not the correct direction of descent in the longer term.

Revisiting feature normalization

- We show how to address in some measure the differential curvature problem by feature normalization.
- Consider the following toy dataset, where the two input attributes are x_1 and x_2 , and the output attribute is y.
- We intend to find a relationship of the form $y = w_1x_1 + w_2x_2$ from the data. The coefficients w_1 and w_2 are found using gradient descent on the loss function computed from the data.

x_1	x_2	y
0.1	25	7
0.8	10	1
0.4	10	4

Revisiting feature normalization

- Loss function:

$$J(\boldsymbol{w}) = (0.1w_1+25w_2-7)^2+(0.8w_1+10w_2-1)^2+(0.4w_1+10w_2-4)^2$$

- Objective function is much more sensitive to w_2 than w_1
- One way to get around this issue is to standardize each column to zero mean and unit variance
- The coefficients for w_1 and w_2 will become much more similar, and differential curvature will be reduced.

Constrained Optimization

- Optimization problems are of 2 types
  1. Unconstrained Optimization
  2. Constrained Optimization
- We discussed algorithms to solve unconstrained optimization
- How do we find the solution to an optimization problem with constraints?
- Example of a constrained optimization problem

$$\begin{aligned} \textit{maximize} \quad & f(x,y) = x^2 y \\ \text{subject to} \quad & g(x,y) : x^2 + y^2 = 1 \end{aligned}$$

Constrained Optimization : Lagrange Multiplier Method

- Constrained maximization (minimization) problem is rewritten as a Lagrange function whose optimal point is a saddle point, i.e. a global maximum (minimum)
- Lagrange function use Lagrange multipliers as a strategy for finding the local maxima and minima of a function subject to constraints
- Maximum of $f(x, y)$ under constraint $g(x, y)$ is obtained when their gradients point to same direction
- Introduce a Lagrange multiplier λ for the equality constraint Mathematically, $\nabla f(x, y) = \lambda \nabla g(x, y)$

Constrained Optimization : Lagrange Multiplier Method

Consider the following optimization problem

$$\begin{aligned} &\text{maximize } xy \\ &\text{subject to} \quad x + y = 6 \end{aligned}$$

- The Lagrangian is $L(x, y) = xy - \lambda(x + y - 6)$

$$\frac{\partial L(x,y)}{\partial \lambda} = x + y - 6 = 0$$

$$\left.\begin{aligned} \frac{\partial L(x,y)}{\partial x} &= y - \lambda = 0 \\ \frac{\partial L(x,y)}{\partial y} &= x - \lambda = 0 \end{aligned}\right\} \Rightarrow x = y = \lambda$$

$$\Rightarrow x = y = 3$$

- x and y values remain same even if you take $+\lambda$ or $-\lambda$ for equality constraint