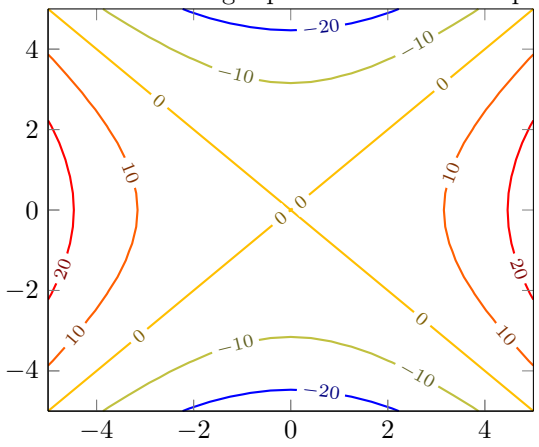
DEEP LEARNING WEEK 4

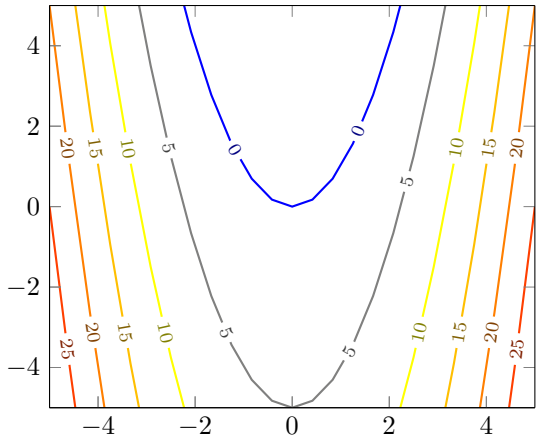
1. We have following functions x^3,$ln(x)$,e^x,x and 4. Which of the following functions has the steepest slope at x=1?
   a)x^3
   b)$ln(x)$
   c)e^x
   d)4
   Answer:a)
   Solution: Calculate the derivatives of following functions at x=1 and choose the function with highest absolute value

2. Which of the following represents the contour plot of the function f(x,y) = $x^2 - y^2$?

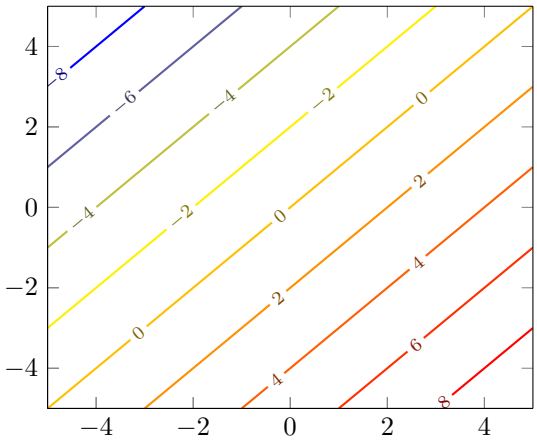
a)



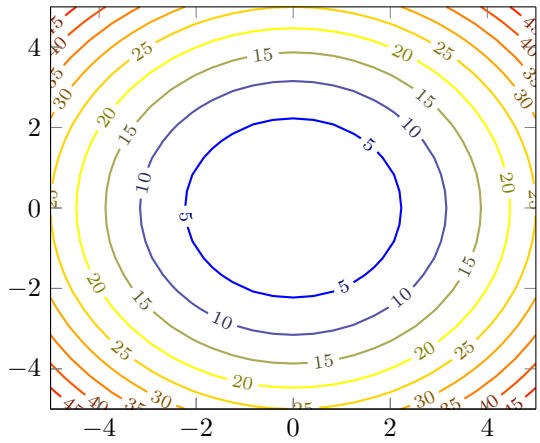
b)

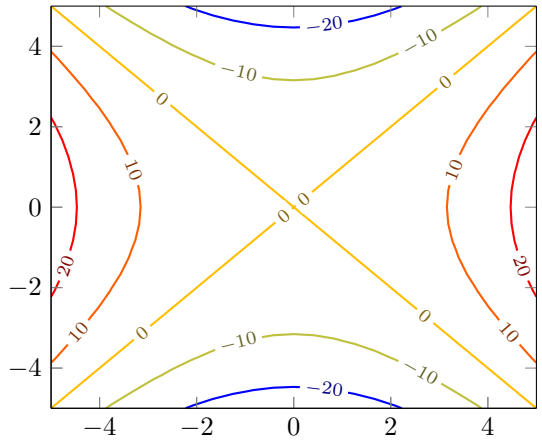


c)



d)





Answer: a)

3. Choose the correct options for the given gradient descent update rule $w_{t+1} = w_t - \eta \nabla w$ (η is the learning rate) (**MSQ**)
   a)The weight update is tiny at a gentle loss surface
   b)The weight update is tiny at a steep loss surface
   c)The weight update is large at a steep loss surface

d)The weight update is large at a gentle loss surface
Answer:a),c)
Solution: Gradient is small at a gentle loss surface and large at a steep loss surface. Gradient determines the size of updates.

4. Which of the following algorithms will result in more oscillations of the parameter during the training process of the neural network?

   a)Stochastic gradient descent
   b)Mini batch gradient descent
   c)Batch gradient descent
   d)Batch NAG

   Answer: a)
   Solution: Since in stochastic gradient descent we update weights based on one training example it is a poor approximation of true gradient compared to say Mini batch more batch gradient hence results in more oscillations.

5. Which of the following are among the disadvantages of Adagrad?

   a)It doesn't work well for the Sparse matrix.
   b)It usually goes past the minima.
   c)It gets stuck before reaching the minima.
   d)Weight updates are very small at the initial stages of the algorithm.

   Answer: c)
   Solution: It gets stuck before reaching the local minima since the learning rate of weight which is dense gets reduced exponentially.

6. Which of the following is a variant of gradient descent that uses an estimate of the next gradient to update the current position of the parameters?
   a) Momentum optimization
   b) Stochastic gradient descent
   c) Nesterov accelerated gradient descent
   d) Adagrad

   Answer: c) Nesterov accelerated gradient descent

   Solution: Nesterov gradient descent estimates the next position of the parameter and calculates the gradient of parameters at that position. The new position is determined using this gradient and the gradient at the original step.

7. Consider a gradient profile ∇W=[1,0.9,0.6,0.01,0.1,0.2,0.5,0.55,0.56]. Assume $v_{-1} = 0, \epsilon = 0, \beta = 0.9$ and the learning rate is η_{-1}. Suppose that we use the Adagrad algorithm then what is the value of $\eta_6 = \eta/sqrt(v_t + \epsilon)$?
   a)0.03
   b)0.06
   c)0.08
   d)0.006
   Answer:b)

Solution: Use the expression $v_t = v_{t-1} + (\nabla W)^2$ to get v_6 using the gradients list given in the question. Calculate $\eta_6 = \eta / sqrt(v_t + \epsilon)$

8. Which of the following can help avoid getting stuck in a poor local minimum while training a deep neural network?

   (a) Using a smaller learning rate.

   (b) Using a smaller batch size.

   (c) Using a shallow neural network instead.

   (d) None of the above.

   Answer: (d) **Solution:** None of the above methods can prevent the neural network from getting stuck in poor local minima.

9. What are the two main components of the ADAM optimizer?
   a) Momentum and learning rate.
   b)Gradient magnitude and previous gradient.
   c)Exponential weighted moving average and gradient variance.
   d)Learning rate and a regularization term.

   Answer: c) The two main components of the ADAM optimizer are exponential moving average and gradient variance.

10. What is the role of activation functions in deep learning?

    (a) Activation functions transform the output of a neuron into a non-linear function, allowing the network to learn complex patterns.

    (b) Activation functions make the network faster by reducing the number of iterations needed for training.

    (c) Activation functions are used to normalize the input data.

    (d) Activation functions are used to compute the loss function.

    Answer: a)

    Solution: Activation functions transform the output of a neuron into a non-linear function, which is important for learning complex patterns. Without activation functions, neural networks would be limited to linear transformations of the input data.