Introduction to Deep Learning

Tutorial



Arijit Mondal

Dept. of Computer Science & Engineering Indian Institute of Technology Patna arijit@iitp.ac.in

Deep Learning

1

Problem-1 • In typical gradient descent, we take steps of a constant size, so that:

 $w_{t+1} = w_t - \epsilon \nabla_w L(w_t)$

In the following, assume that
$$L$$
 is an arbitrary differentiable function.

• For very small ϵ what will generally be true? (a) $L(w_t) \geq L(\theta_{t+1})$, (b) $L(w_t) \leq L(w_{t+1})$,

(c) Cannot sav

• In typical gradient descent, we take steps of a constant size, so that:

$$w_{t+1} = w_t - \epsilon \nabla_w L(w_t)$$

In the following, assume that L is an arbitrary differentiable function.

In the following, assume that
$$L$$
 is an arbitrary differential

very small
$$\epsilon$$
 what will generally be true? (a)

• For very small ϵ what will generally be true? (a) $L(w_t) \geq L(\theta_{t+1})$, (b) $L(w_t) \leq L(w_{t+1})$,

• For a very big
$$\epsilon$$
 what will generally be true? (a) $L(w_t) \geq L(\theta_{t+1})$, (b) $L(w_t) \leq L(w_{t+1})$,

(c) Cannot sav

Problem-1 • In typical gradient descent, we take steps of a constant size, so that: $w_{t+1} = w_t - \epsilon \nabla_w L(w_t)$

In the following, assume that L is an arbitrary differentiable function.

• For very small ϵ what will generally be true? (a) $L(w_t) > L(\theta_{t+1})$, (b) $L(w_t) < L(w_{t+1})$,

• For a very big ϵ what will generally be true? (a) $L(w_t) > L(\theta_{t+1})$, (b) $L(w_t) < L(w_{t+1})$, (c) Cannot sav

the direction
$$\nabla_w L(w)$$
 and then uses ϵ' as the step size:
$$\epsilon' = \arg\min L(w_t - \nabla_w L(w_t)); \qquad w_{t+1} = w_t - \epsilon' \nabla_w L(w_t)$$

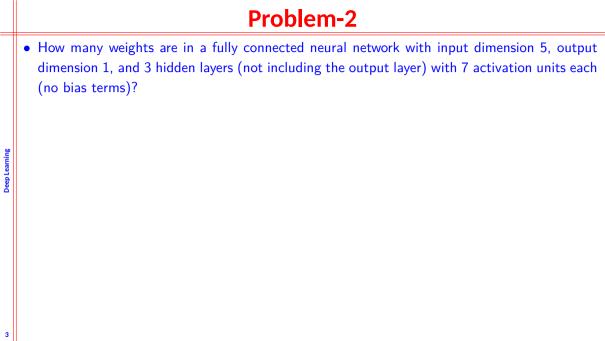
the direction $\nabla_w L(w)$ and then uses ϵ' as the step size:

 We would like to pick a perfect step size on every step and propose a new update rule that selects ϵ' to be the value step-size ϵ that decreases the objective as much as possible in

(a) $L(w_t) \ge L(\theta_{t+1})$, (b) $L(w_t) \le L(w_{t+1})$, (c) Cannot say

$$(w_t) \geq L(\theta_{t+1}),$$

$$)\geq L(heta_{t+1}), hinspace (b) L($$



	Problem-2
	• How many weights are in a fully connected neural network with input dimension 5, output dimension 1, and 3 hidden layers (not including the output layer) with 7 activation units each (no bias terms)?
	$ullet$ Find a relation between $ anh(x)$ and $\sigma(2x)$
Deep Learning	
3	

- For each of the following loss functions which activation functions in the last layer is appropriate?
 Negative Log-Likelihood Multiclass (NLLM) loss: a. Linear, b. Softmax, c. Sigmoid
 - Negative Log-Likelinood Multiclass (NLLM) loss: a. Linear, b. Softmax, c. Sigmoid
 - Squared loss: a. Linear, b. Softmax, c. Sigmoid
- For each of the following applications which activation function is appropriate?
 Map words in a news page to a predicted numerical change in a stock market mean: a.
- Linear, b. Softmax

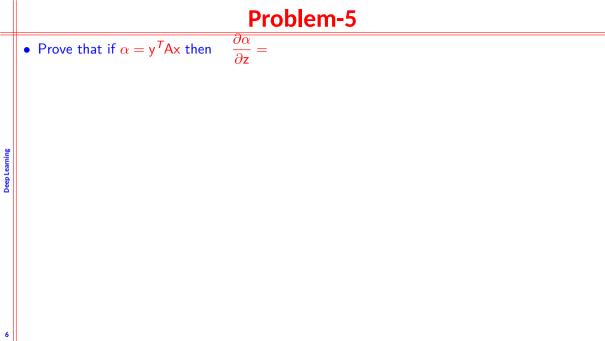
- For each of the following loss functions which activation functions in the last layer is appropriate?
- Negative Log-Likelihood Multiclass (NLLM) loss: a. Linear, b. Softmax, c. Sigmoid
- Squared loss: a. Linear, b. Softmax, c. Sigmoid
- For each of the following applications which activation function is appropriate?
 - Map words in a news page to a predicted numerical change in a stock market mean: a. Linear, b. Softmax
 - Map a satellite image to the probability it will rain at that location during the next day:
 a. Linear, b. Softmax

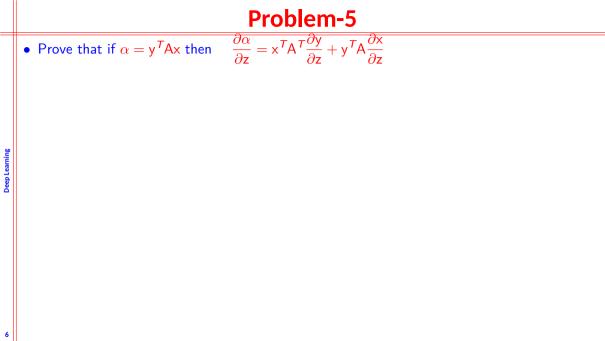
- For each of the following loss functions which activation functions in the last layer is appropriate?
 - Negative Log-Likelihood Multiclass (NLLM) loss: a. Linear, b. Softmax, c. Sigmoid
 - Squared loss: a. Linear, b. Softmax, c. Sigmoid
- For each of the following applications which activation function is appropriate?

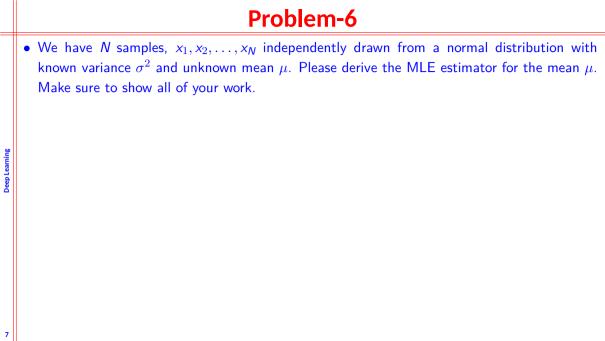
 Man words in a news page to a predicted numerical change in a stock market mean: a
 - Map words in a news page to a predicted numerical change in a stock market mean: a. Linear, b. Softmax
 - Map a satellite image to the probability it will rain at that location during the next day:
 a. Linear, b. Softmax
 - Map words in an email to which one of a fixed set of folders it should be filed in: a. Linear,
 b. Softmax

- What is the gradient of the function at (1,1)
- If we initialize gradient descent to (1,1) with $\epsilon = 0.0001$, what are the values of (x,y)

- after the first iteration of gradient descent?







• We have N samples, x_1, x_2, \dots, x_N independently drawn from a normal distribution with

- known variance σ^2 and unknown mean μ . Please derive the MLE estimator for the mean μ . Make sure to show all of your work.
- Consider the following recurrence: $(x_{t+1}, y_{t+1}) = (f(x_t, y_t), g(x_t, y_t))$. Here, f() and g() are multivariate functions. Derive an expression for $\frac{\partial x_{t+2}}{\partial x_t}$ in terms of only x_t and y_t .
- Consider a two-input neuron that multiplies its two inputs x_1 and x_2 to obtain the output o. Let L be the loss function that is computed at o. Suppose that you know that $\frac{\partial L}{\partial o} = 5, x_1 = 2$
- and $x_2 = 3$. Compute the values of $\frac{\partial L}{\partial x_1}$, $\frac{\partial L}{\partial x_2}$.

- known variance σ^2 and unknown mean μ . Please derive the MLE estimator for the mean μ . Make sure to show all of your work. • Consider the following recurrence: $(x_{t+1}, y_{t+1}) = (f(x_t, y_t), g(x_t, y_t))$. Here, f() and g() are
- multivariate functions. Derive an expression for $\frac{\partial x_{t+2}}{\partial x_t}$ in terms of only x_t and y_t . • Consider a two-input neuron that multiplies its two inputs x_1 and x_2 to obtain the output o.
- Let L be the loss function that is computed at o. Suppose that you know that $\frac{\partial L}{\partial a} = 5, x_1 = 2$ and $x_2 = 3$. Compute the values of $\frac{\partial L}{\partial x_1}$, $\frac{\partial L}{\partial x_2}$.
- Consider the softmax as output function ie. $o_i = softmax(v) = \frac{exp(v_i)}{\sum_{k} exp(v_k)}$. Show that $\frac{\partial o_i}{\partial v_i}$ is $o_i(1 - o_i)$ when i = j. Find when $i \neq j$.

function of the dot product $r_1^T r_2$.

Problem-7

SGD for optimization. Let us assume that we sample the subfunction f_2 and we start from $x_0 = 0$. Find the new value of x ie. x_1 (say). Find the relation between $f(x_0)$ and $f(x_1)$. • Suppose each word is represented as unit vector having dimension d. Consider two words are represented as r_1 and r_2 . Show that Euclidean distance $||r_1 - r_2||$ is a monotonically decreasing

x₀ = 0. Find the new value of x ie. x₁ (say). Find the relation between f(x₀) and f(x₁).
Suppose each word is represented as unit vector having dimension d. Consider two words are represented as r₁ and r₂. Show that Euclidean distance ||r₁-r₂|| is a monotonically decreasing function of the dot product r₁^Tr₂.

• Consider a binary classification problem. To avoid overlay confident prediction, we transform the prediction y to lie in the interval [0.1,0.9]. In other words, we take $y=0.8\sigma(z)+0.1$ where σ denotes the logistic function. We still use the cross entropy loss. For a positive training example, sketch the cross entropy loss as a function of z.

• Consider the function $f(x,y) = \frac{1}{2}(x^2 + by^2)$ where $0 < b \le 1$. We apply gradient descent with

exact line search method. Here the step size (α) is computed as follows $\alpha = \arg\min_{\alpha} f(x - 1)$

 $\alpha \nabla_x f(x)$). Let us assume that we start from $(x_0, y_0) = (b, 1)$. Find the value of (x_k, y_k) . Can

you find any interesting property of two consecutive gradients?

• Let $\theta^* \in \mathbb{R}^d$ and let $f(\theta) = \frac{1}{2} \|\theta - \theta^*\|^2$. Show that the Hessian of f is identity matrix.

 X^TX

• Consider the function $f(x,y) = \frac{1}{2}(x^2 + by^2)$ where $0 < b \le 1$. We apply gradient descent with

you find any interesting property of two consecutive gradients?

exact line search method. Here the step size (α) is computed as follows $\alpha = \arg\min_{\alpha} f(x - 1)$

• Let $\theta^* \in \mathbb{R}^d$ and let $f(\theta) = \frac{1}{2} \|\theta - \theta^*\|^2$. Show that the Hessian of f is identity matrix.

 $\alpha \nabla_x f(x)$). Let us assume that we start from $(x_0, y_0) = (b, 1)$. Find the value of (x_k, y_k) . Can

• Let $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$. For $\theta \in \mathbb{R}^d$ let $g(\theta) = \frac{1}{2} ||X\theta - y||^2$. Show that the Hessian of g is

Problem-8 • Consider the function $f(x,y) = \frac{1}{2}(x^2 + by^2)$ where $0 < b \le 1$. We apply gradient descent with

exact line search method. Here the step size (α) is computed as follows $\alpha = \arg\min_{\alpha} f(x - \alpha)$ $\alpha \nabla_x f(x)$). Let us assume that we start from $(x_0, y_0) = (b, 1)$. Find the value of (x_k, y_k) . Can you find any interesting property of two consecutive gradients?

• Let $\theta^* \in \mathbb{R}^d$ and let $f(\theta) = \frac{1}{2} \|\theta - \theta^*\|^2$. Show that the Hessian of f is identity matrix.

• Let
$$X \in R^{n \times d}$$
 and $y \in R^n$. For $\theta \in R^d$ let $g(\theta) = \frac{1}{2} ||X\theta - y||^2$. Show that the Hessian of g is $X^T X$.

• A random variable follows an exponential distribution with parameter λ ($\lambda > 0$) if it has the following density: $p(t) = \lambda e^{-\lambda t}$, $t \in [0, \infty)$. This distribution is often used to model waiting times between events. Imagine you are given i.i.d. data $T = (t_1, \dots, t_n)$ where each t_i is modeled as being drawn from an exponential distribution with parameter λ . (a) Compute the log-probability of T given λ . (b) Solve for $\hat{\lambda}_{\text{MLE}}$