# CS224 - Assignment 3

O.L.

September 2019

# 1

## 1.a

### 1.a.1

Momentum update is another approach that almost always enjoys better converge rates on deep networks. This update can be motivated from a physical perspective of the optimization problem. In particular, the loss can be interpreted as the height of a hilly terrain (and therefore also to the potential energy since $U = mgh$ and therefore $U \sim h$ ). Initializing the parameters with random numbers is equivalent to setting a particle with zero initial velocity at some location. The optimization process can then be seen as equivalent to the process of simulating the parameter vector (i.e. a particle) as rolling on the landscape.

Since the force on the particle is related to the gradient of potential energy (i.e. $F = \nabla U$ ), the force felt by the particle is precisely the (negative) gradient of the loss function. Moreover, $F = ma$ so the (negative) gradient is in this view proportional to the acceleration of the particle.

As an unfortunate misnomer, this variable is in optimization referred to as momentum (its typical value is about 0.9), but its physical meaning is more consistent with the coefficient of friction. Effectively, this variable damps the velocity and reduces the kinetic energy of the system, or otherwise the particle would never come to a stop at the bottom of a hill.

### 1.a.2

The nominator: $\alpha \circ m \sim \nabla J$.
The denominator: $\alpha \circ m \sim (\nabla J)^2$.
So the second term: $\frac{1}{\nabla J}$

we multiply the exponential average computed till the last update with a hyperparameter, represented by the greek symbol nu. We then multiply the square of the current gradient with (1 - nu). We then add them together to get the exponential average till the current time step.

The reason why we use exponential average is because as we saw, in the momentum example, it helps us weigh the more recent gradient updates more

than the less recent ones. In fact, the name "exponential" comes from the fact that the weightage of previous terms falls exponentially.

## 1.b

### 1.b.1

Writing the expectation as follows:

$$E_{p_{drop}}[h_{drop}]_i = E_{p_{drop}}[\gamma \mathbf{d} \circ \mathbf{h}]_i = \gamma E_{p_{drop}}[d_i h_i] = \gamma(0 \times p_{drop} + 1 \times (1 - p_{drop}))h_i = \gamma h_i(1 - p_{drop})$$

Therefore:

$$\gamma = \frac{1}{1 - p_{drop}}$$

### 1.b.2

Question:Why should we apply dropout during training but not during evaluation?
Answer: Because in training every node has a probability of $p_{drop}$ of dropping, it's mean that $p_{drop}$ of the total nodes that connected into the following node are dropped. In testing there $\frac{1}{1-p_{drop}}$ times the nodes as in the training, so in needed to be factorized - this can be done during training as well(The $\gamma$ multiplication).

# 2

## 2.a

| Stack | Buffer | New dependency | Transition |
|---|---|---|---|
| [ROOT] | [I, parsed, this, sentence, correctly] | | Initial Config |
| [ROOT,I] | [parsed, this, sentence, correctly] | | SHIFT |
| [ROOT,I,parsed] | [this, sentence, correctly] | | SHIFT |
| [ROOT,parsed] | [this, sentence, correctly] | parsed → I | LEFT-ARC |
| [ROOT,parsed,this] | [sentence, correctly] | | SHIFT |
| [ROOT,parsed,this,sentence] | [correctly] | | SHIFT |
| [ROOT,parsed,sentence] | [correctly] | sentence→ this | LEFT-ARC |
| [ROOT,parsed] | [correctly] | parsed→ sentence | RIGHT-ARC |
| [ROOT,parsed,correctly] | [] | | SHIFT |
| [ROOT,parsed] | [] | parsed→ correctly | RIGHT-ARC |
| [ROOT] | [] | ROOT→ parsed | RIGHT-ARC |

The buffer is empty and the stack size is one.

## 2.b

$2n + 1$ Steps:
$n$ - Steps to insert all the words from the buffer into the stack.
$n$ - Steps to make dependencies between 2 words(and so one word leaves the stack).
1 - The final step.

## 2.e

TESTING:
Restoring the best model weights found on the dev set
Final evaluation on test set
2919736it [00:00, 26957575.37it/s]
- test UAS: 89.08

## 2.f

i.
- Error type: Prepositional Phrase Attachment Error
- Incorrect dependency: heading → fearing
- Correct dependency: was → fearing

ii.
- Error type: Coordination Attachment Error
- Incorrect dependency: makes → rescue
- Correct dependency: rush → rescue

iii.
- Error type: Prepositional Phrase Attachment Error
- Incorrect dependency: named → Midland
- Correct dependency: guy → Midland

iv.
- Error type: Modifier Attachment Error
- Incorrect dependency: one → crucial
- Correct dependency: elements → crucial