

## Deep Learning – HW1 Dry

Amit Rubinshtein – 324300722 – [amit.ru@campus.technion.ac.il](mailto:amit.ru@campus.technion.ac.il)

Nadav Offir – 213786197 - [Nadav.offir@campus.technion.ac.il](mailto:Nadav.offir@campus.technion.ac.il)

### Question 1

$$|f(x) - f(y) - \nabla f(y)^T(x - y)| = \left| \int_0^1 \nabla f(y + t(x - y))^T(x - y) dt - \nabla f(y)^T(x - y) \right| \leq$$

Cauchy-Schwarz

$$\int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \cdot \|x - y\| dt \leq$$

$$\int_0^1 \beta t \|x - y\|^2 dt = \frac{\beta}{2} \|x - y\|^2$$

## Question 2

1.  $\mathbb{E}[w(t)] = \mathbb{E}[w(t-1)] - \eta h \mathbb{E}[w(t-1)] = (1 - \eta h) \mathbb{E}[w(t-1)]$
2. We can show like the derivation in the class that:

$$\mathbb{E}[w(t)] = (1 - \eta h)^t \mathbb{E}[w(0)]$$

This converges to the minimum at the maximal rate when:

$$1 - \eta h = 0 \Rightarrow \eta_{optimal} = \frac{1}{h}$$

3. To converge we demand:

$$\lim_{t \rightarrow \infty} |(1 - \eta h)^t \mathbb{E}[w(0)]| < \infty$$

This demand is met when:

$$|1 - \eta h| < 1$$

$$0 < \eta < \frac{2}{h}$$

4.  $\mathbb{E}[w^2(t)] = (1 - 2\eta h + \eta^2(h^2 + \rho)) \mathbb{E}[w^2(t-1)] =$   
 $= ((1 - \eta h)^2 + \eta^2 \rho) \mathbb{E}[w^2(t-1)]$

That is because:  $\mathbb{E}[h_{n(t)}^2] = h^2 + \rho$

5. Once again we can show that:

$$\mathbb{E}[w^2(t)] = ((1 - \eta h)^2 + \eta^2 \rho)^t \mathbb{E}[w^2(0)]$$

We will write:

$$f(\eta) = 1 - 2\eta h + \eta^2(h^2 + \rho)$$

we will want to minimize  $f(\eta)$  in order to get the minimum at the maximum rate. So we will get:

$$\frac{df}{d\eta} = -2h + 2\eta(h^2 + \rho) = 0 \Rightarrow \eta_{optimal} = \frac{h}{h^2 + \rho}$$

and  $\frac{d^2f}{d\eta^2} = 2(h^2 + \rho) > 0$ , so  $\eta_{optimal}$  is indeed the minimum.

6. In order to converge we will demand:

$$|g(\eta)| < 1$$

Because  $f(\eta)$  is a sum of squares then it is always positive so:

$$|f(\eta)| = f(\eta) = 1 - 2\eta h + \eta^2(h^2 + \rho) < 1$$

$$\eta(h^2 + \rho) - 2h < 0$$

because  $\eta > 0$  we will get:

$$\eta(h^2 + \rho) - 2h < 0$$

$$\Rightarrow 0 < \eta < \frac{2h}{h^2 + \rho}$$

7. Because  $f(w) \propto w^2$  when  $\mathbb{E}[w^2(t)]$  converges so is  $\mathbb{E}[f(w(t))]$ . So we will get the same range as the previous sub-question:

$$0 < \eta < \frac{2h}{h^2 + \rho}$$

8. We will get the update equation:

$$\log(w(t)) = \log(w(t-1) - \eta h_{n(t)} w(t-1))$$

$$\begin{aligned}
&= \log\left((1 - \eta h_{n(t)})w(t-1)\right) = \log(w(t-1)) + \log(1 - \eta h_{n(t)}) \\
&= \log(w(0)) + \sum_{n=0}^t \log(1 - \eta h_n)
\end{aligned}$$

now:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log(w(t)) = \lim_{t \rightarrow \infty} \left[ \frac{1}{t} \log(w(0)) + \frac{1}{t} \sum_{n=0}^t \log(1 - \eta h_n) \right]$$

using the law of large numbers because  $h_{n(t)}$  is i.i.d random sample, the number of times each  $h_n$  appears can be written as  $H_n(t) = 1^{\delta(h_{n(t)} - h_n)}$  we will get with probability 1:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^t H_n(t') = \mathbb{E}[h_{n(t)}] = \frac{1}{N}$$

using this going back to the main equation we will get:

$$= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=0}^t H_n(t') \sum_{n=1}^N \log(1 - \eta h_n) = \frac{1}{N} \sum_{n=1}^N \log(1 - \eta h_n)$$

with probability 1.

9. The condition that must be hold is  $q(\eta) < 0$  because than:

$$\frac{1}{t} \log w(t) \rightarrow q < 0 \Rightarrow \log w(t) \sim qt \rightarrow -\infty$$

then we will get:

$$w(t) = \exp(\log w(t)) \rightarrow 0$$

### Question 3

1. Chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}(w(t-1))}{\partial \eta_{t-1}} &= \frac{\partial \mathcal{L}(w(t-1))}{\partial w(t-1)} \frac{\partial w(t-1)}{\partial \eta_{t-1}} \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \frac{\partial}{\partial \eta_{t-1}} (w(t-2) - \eta_{t-1} \nabla \mathcal{L}(w(t-2))) \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \nabla \mathcal{L}(w(t-2))\end{aligned}$$

2. Chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}(w(t-1))}{\partial \alpha_{t-1}} &= \frac{\partial \mathcal{L}(w(t-1))}{\partial \eta_{t-1}} \frac{\partial \eta_{t-1}}{\partial \alpha_{t-1}} \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \nabla \mathcal{L}(w(t-2)) \frac{\partial}{\partial \alpha_{t-1}} \left( \eta_{t-2} - \alpha_{t-1} \frac{\nabla \mathcal{L}(w(t-2))}{\partial \eta_{t-2}} \right) \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \nabla \mathcal{L}(w(t-2)) \frac{\nabla \mathcal{L}(w(t-2))}{\partial \eta_{t-2}} \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \nabla \mathcal{L}(w(t-2)) \nabla \mathcal{L}(w(t-2)) \cdot \nabla \mathcal{L}(w(t-3))\end{aligned}$$

3. Chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}(w(t-1))}{\partial \eta_{t-2}} &= \frac{\partial \mathcal{L}(w(t-1))}{\partial w(t-1)} \frac{\partial w(t-1)}{\partial w(t-2)} \frac{\partial w(t-2)}{\partial \eta_{t-2}} \\ &= -\nabla \mathcal{L}(w(t-1)) \frac{\partial (w(t-2) - \eta_{t-1} \nabla \mathcal{L}(w(t-2)))}{\partial w(t-2)} \frac{\partial (w(t-3) - \eta_{t-2} \nabla \mathcal{L}(w(t-3)))}{\partial \eta_{t-2}} \\ &= -\nabla \mathcal{L}(w(t-1)) (I - \eta_{t-1} \nabla^2 \mathcal{L}(w(t-2))) \nabla \mathcal{L}(w(t-3))\end{aligned}$$

4. Chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}(w(t-1))}{\partial \eta_{t-\tau}} &= \frac{\partial \mathcal{L}(w(t-1))}{\partial w(t-1)} \frac{\partial w(t-1)}{\partial w(t-2)} \frac{\partial w(t-2)}{\partial w(t-3)} \dots \frac{\partial w(t-\tau)}{\partial \eta_{t-\tau}} \\ &= -\nabla \mathcal{L}(w(t-1)) \cdot \left[ \prod_{t'=1}^{\tau-1} I - \eta_{t-t'} \nabla^2 \mathcal{L}(w(t-t'-1)) \right] \nabla \mathcal{L}(w(t-\tau-1))\end{aligned}$$

### 5. Advantages of the first

- It updates the step size more frequently, which is beneficial because delayed updates can negatively affect performance, as discussed in class.
- It is computationally cheaper, since it avoids explicit Hessian computations. While Hessian-vector products are feasible in practice, they are still more expensive than standard gradient calculations.
- The second approach requires backpropagation through a long optimization trajectory, which is both memory-intensive (because the entire process must be stored) and potentially unstable due to vanishing or exploding gradients.

**Advantages of the second approach:**

- The step size is updated based on the loss at a later point in the optimization, which is more closely related to the final objective (the loss at the end of training). In contrast, the first approach is greedy and focuses only on minimizing the loss after a single update step.

#### Question 4

1. We will do each part separately and then join them:

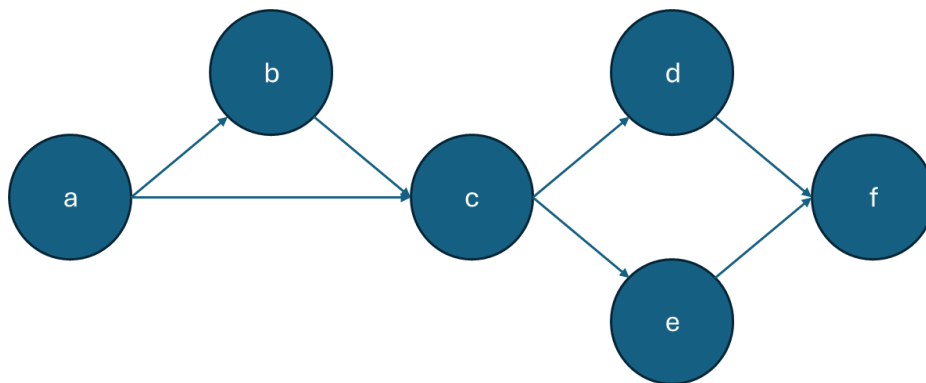
$$g(x) = \exp(\exp(x) + \exp(x)^2); q(x) = \sin(\exp(x) + \exp(x)^2)$$

$$\frac{dg}{dx} = \exp(\exp(x) + \exp(x)^2) \cdot (\exp(x) + 2 \exp(x)^2)$$

$$\frac{dq}{dx} = \cos(\exp(x) + \exp(x)^2) \cdot (\exp(x) + 2 \exp(x)^2)$$

$$\frac{df}{dx} = (\exp(x) + 2 \exp(x)^2) \cdot (\cos(\exp(x) + \exp(x)^2) + \exp(\exp(x) + \exp(x)^2))$$

- 2.



- 3.

$$\frac{df}{dd} = 1$$

$$\frac{df}{de} = 1$$

$$\frac{df}{dc} = \frac{df}{dd} \cdot \frac{dd}{dc} + \frac{df}{de} \cdot \frac{de}{dc} = \exp(c) + \cos(c)$$

$$\frac{df}{db} = \frac{df}{dc} \cdot \frac{dc}{db} = \exp(c) + \cos(c)$$

$$\frac{df}{da} = \frac{df}{dc} \cdot \frac{dc}{da} = (\exp(c) + \cos(c))(1 + 2a)$$

$$= (\exp(a + a^2) + \cos(a + a^2))(1 + 2a)$$

$$\frac{df}{dx} = \frac{df}{da} \cdot \frac{da}{dx} = (\exp(a + a^2) + \cos(a + a^2))(1 + 2a) \exp(x)$$

$$= (\exp(\exp(x) + \exp(x)^2) + \cos(\exp(x) + \exp(x)^2))(1 + 2 \exp(x)) \exp(x)$$

### Question 5

$$\begin{aligned}f(g(h(x + \epsilon x')) &= f(g(h(x) + \epsilon h'(x)x')) \\&= f(g(h(x)) + \epsilon g'(h(x))h'(x)x') \\&= f(g(h(x))) + \epsilon f'(g(h(x)))g'(h(x))h'(x)x'\end{aligned}$$

Therefore:

$$\begin{aligned}\left. \frac{df(x)}{dx} \right|_x &= \text{coefficient of epsilon}(\text{dual version}(f)(x + 1 \cdot \epsilon)) \\&= f'(g(h(x)))g'(h(x))h'(x)\end{aligned}$$