Deep Learning Assignment 1

Arya Jalali

February 19, 2023

1 Linear Algebra Recap

1.1

First we define a vector-valued function $\mathbf{f}$ below

$$\mathbf{f}([u,v,z]) = [\frac{\partial y}{\partial u}, \frac{\partial y}{\partial v}, \frac{\partial y}{\partial z}] \tag{1}$$

Using the definition we'll derive the Jacobian Matrix of $\mathbf{f}$

$$\mathbf{J} = [\frac{\partial \mathbf{f}}{\partial u}, \frac{\partial \mathbf{f}}{\partial v}, \frac{\partial \mathbf{f}}{\partial z}] = \begin{pmatrix} \frac{\partial y}{\partial u \partial u} & \frac{\partial y}{\partial u \partial v} & \frac{\partial y}{\partial u \partial z} \\ \frac{\partial y}{\partial v \partial u} & \frac{\partial y}{\partial v \partial v} & \frac{\partial y}{\partial v \partial z} \\ \frac{\partial y}{\partial z \partial u} & \frac{\partial y}{\partial z \partial v} & \frac{\partial y}{\partial z \partial z} \end{pmatrix} \tag{2}$$

Based on the definition, the derived matrix is equivalent to the Hessian matrix of our function $\mathbf{f}$. Note that for simplicity we omitted $\mathbf{f} = \nabla\psi$

1.2

(i) We'll first calculate the ij'th element , and generalize it from there.

$$\frac{\partial y_i}{\partial x_j} = \frac{1}{\partial x_j} \sum_{k=1}^{n} a_{ik} x_k = a_{ij} \rightarrow \frac{\partial y_i}{\partial x_j} = a_{ij} \rightarrow \frac{\partial y}{\partial x} = A \tag{3}$$

(ii) We'll use the same method

$$\frac{\partial y_i}{\partial z} = \frac{1}{\partial z} \sum_{k=1}^{n} a_{ik} x_k = \sum_{k=1}^{n} a_{ik} \frac{\partial x_k}{\partial z} = A \frac{\partial x}{\partial z}_k \rightarrow \frac{\partial y}{\partial z} = A \frac{\partial x}{\partial z} \tag{4}$$

(iii) Using the definition of derivative we can write

$$\frac{\partial \alpha}{\partial x} y^T A x = \lim_{h \to 0} \frac{y^T A(x+h) - y^T A(x)}{h} = \lim_{h \to 0} \frac{y^T A h}{h} = y^T A \tag{5}$$

$$\frac{\partial \alpha}{\partial y} = \lim_{h \to 0} \frac{(y+h)^T A x - y^T A x}{h} = \lim_{h \to 0} \frac{h^T A x}{h} = \lim_{h \to 0} \frac{x^T A^T h}{h} = x^T A^T \tag{6}$$

We can transpose the nominator in (6) since it's a scalar.

(iv)

$$\alpha = y_1 x_1 + \cdots + y_n x_n \rightarrow \frac{\partial \alpha}{\partial z} = \frac{\partial y_1}{\partial z} x_1 + y_1 \frac{\partial x_1}{\partial z} + \cdots + \frac{\partial y_n}{\partial z} x_n + y_n \frac{\partial x_n}{\partial z} \tag{7}$$

$$= \frac{\partial y_1}{\partial z} x_1 + \frac{\partial y_2}{\partial z} x_2 + \cdots + \frac{\partial y_n}{\partial z} x_n + \frac{\partial x_1}{\partial z} y_1 + \frac{\partial x_2}{\partial z} y_2 + \cdots + \frac{\partial x_n}{\partial z} y_n \tag{8}$$

$$\rightarrow \frac{\partial \alpha}{\partial z} = y^T \frac{\partial x}{\partial z}^T + x^T \frac{\partial y}{\partial z} \tag{9}$$

(v) Using $A^{-1}A = I$ we can write

$$\frac{\partial A^{-1}A}{\partial \alpha} = \frac{\partial I}{\partial \alpha} = 0 \rightarrow \frac{\partial A^{-1}A}{\partial \alpha}_{ij} = \frac{1}{\partial \alpha}\sum_{k}^{n} = a_{ik}^{-1}a_{kj} = 0 \tag{10}$$

$$\sum_{k}^{n} \frac{\partial a_{ik}^{-1}}{\partial \alpha}a_{kj} + a_{ik}^{-1}\frac{\partial a_{kj}}{\partial \alpha} = 0 \tag{11}$$

Let's write the equation as a matrix multiplication

$$\frac{\partial A^{-1}}{\partial \alpha}A = -A^{-1}\frac{\partial A}{\partial \alpha} \rightarrow \frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1} \tag{12}$$

1.3

$$A\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} = \begin{pmatrix}||a_1^T||_1^2\\||a_2^T||_1^2\\\vdots\\||a_n^T||_1^2\end{pmatrix} = \begin{pmatrix}m\\m\\\vdots\\m\end{pmatrix} \tag{13}$$

Therefore the vector $\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}$ is an eigenvector corresponding to the eigenvalue m.

2 Optimization

2.1

Taking the gradient of the function gives us

$$\nabla f = \begin{pmatrix}6x_1 - 3x_2 + 12x_1^2 + 4x_1^3\\4x_2 - 3x_1\end{pmatrix} \tag{14}$$

We can find the saddle points by solving the linear system $\nabla f = 0$

$$6x_1 - 3x_2 + 12x_1^2 + 4x_1^3 = 0 \tag{15}$$

$$4x_2 - 3x_1 = 0 \Rightarrow x_2 = \frac{3}{4}x_1 \tag{16}$$

$$6x_1 - \frac{9}{4}x_1 + 12x_1^2 + 4x_1^3 = 0 \equiv x_1(6 - \frac{9}{4} + 12x_1 + 4x_1^2) = 0 \tag{17}$$

$$f(x_1) = \begin{cases}x_1 &= 0,\\x_1 &= \frac{1}{4}(\sqrt{21} - 6),\\x_1 &= \frac{1}{4}(-\sqrt{21} - 6)\end{cases}$$

Using $x_2 = \frac{3}{4}x_1$ we get $(0,0), (\frac{1}{4}(\sqrt{21} - 6), \frac{3}{16}(\sqrt{21} - 6)), \frac{3}{16}(-\sqrt{21} - 6)$ as our saddle points.

$$\begin{cases}f_{x_1x_1} &= 6 + 24x_1 + 12x_1^2,\\f_{x_2x_2} &= 4,\\f_{x_1x_2} &= -3\end{cases}$$

$$\begin{cases}D(0) > 0, \frac{\partial^2 f}{\partial x_1^2} > 0 & \Rightarrow (0,0) \text{ is a local minimum}\\D(\frac{1}{4}(\sqrt{21} - 6)) < 0 & \Rightarrow (\frac{1}{4}(\sqrt{21} - 6), \frac{3}{16}(\sqrt{21} - 6)) \text{ is a saddle point}\\D(\frac{1}{4}(-\sqrt{21} - 6)) > 0, \frac{\partial^2 f}{\partial x_1^2} > 0 & \Rightarrow (\frac{1}{4}(-\sqrt{21} - 6), \frac{3}{16}(-\sqrt{21} - 6)) \text{ is a local minimum}\end{cases}$$

2.2

2.2.1

To perform one step of the Newton-Raphson method, we need to calculate the gradient and Hessian of $f(x)$ at the point $(0, 0)$, which are given by:

$$\nabla f(x) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}) = (4x + 3.4\pi sin(0.2\pi x) - y, 4y - 17cos(0.2\pi x) - x) \rightarrow \nabla f(x)|_{(x,y)=(0,0)} = (0, -17)$$

$$H_f = \begin{pmatrix} 4 + 0.68\pi^2 ycos(0.2\pi x) & 3.4\pi sin(0.2\pi x) - 1 \\ 3.4\pi sin(0.2\pi x) - 1 & 4 \end{pmatrix}_{(x,y)=(0,0)} = \begin{pmatrix} 4 & -1 \\ -1 & 4 \end{pmatrix}$$

$$H_f^{-1} = \begin{pmatrix} \frac{4}{15} & \frac{1}{15} \\ \frac{1}{15} & \frac{4}{15} \end{pmatrix} \tag{18}$$

Letting α be 1 and Substituting the values we calculated above, we get:

$$x_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{4}{15} & \frac{1}{15} \\ \frac{1}{15} & \frac{4}{15} \end{pmatrix} \begin{pmatrix} 0 \\ -17 \end{pmatrix} = \begin{pmatrix} \frac{17}{15} \\ \frac{68}{15} \end{pmatrix} \tag{19}$$

2.2.2

(i) We are simply calculating the gradient of our loss function

(ii) Maintaining a running average of all past steps helps us have a smoother descent towards the minimum

(iii) Calculating the mean squared gives us an estimate of the second moment. We will use this to scale the gradient in each dimension (the higher derivative suggests avoiding large changes in the variable)

(iv) We are performing bias correction so we can have an unbiased estimator

(v) Same as before

(vi) The final update rule combines all previous equations and depends on both current gradient and past values of the second and first moment.

We will prove this for m_t, the proof is analogous for v_t

$$m_1 = (1-\beta_1)g_1, \;\; m_2 = \beta_1(1-\beta_1)g_1 + (1-\beta_1)g_2, m_3 = \beta_1^2(1-\beta_1)g_1 + \beta_1(1-\beta_1)g_2 + (1-\beta_1)g_3 \tag{20}$$

We can generalize a bit, and write a closed form for m_t

$$m_t = (1-\beta_1)\sum_{i=0}^{t} \beta_1^{t-i} g_i \rightarrow E[m] = (1-\beta_1)\sum_{i=0}^{t} E[\beta_1^{t-i} g_i] = E[g](1-\beta_1^t) \tag{21}$$

The $(1-\beta_1^t)$ is what we have to get rid of so we can get an unbiased estimator. Dividing our moving average by $(1-\beta_1^t)$ corrects this bias.

$$E[\tilde{m}] = \frac{1}{1-\beta_1^t} E[m] = E[g] \tag{22}$$

3 Backpropagation

3.1

$$\frac{\partial \mathcal{L}_{reg}}{\partial w} = \frac{\partial \mathcal{L}}{\partial w} + \lambda \frac{\partial R}{\partial w}$$
$$\frac{\partial R}{\partial w} = w$$
$$\frac{\partial R}{\partial b} = 0$$
$$\frac{\partial y}{\partial w} = \sigma(z)(1-\sigma(z))x$$
$$\frac{\partial L}{\partial w} = (y-t)\sigma(z)(1-\sigma(z))x$$
$$\frac{\partial L}{\partial b} = (y-t)\sigma(z)(1-\sigma(z))$$

Using chain rule we get

$$\frac{\mathcal{L}_{reg}}{\partial b} = (y-t)\sigma(z)(1-\sigma(z))$$
$$\frac{\mathcal{L}_{reg}}{\partial w} = (y-t)\sigma(z)(1-\sigma(z))x + \lambda w$$

3.2

Small weights can avoid problems such as vanishing gradient, because large values would make $\sigma(z)$ get close to 1. This combined with the process power bigger weights require compels us to use smaller weights. Randomness helps SGD converge, and drastically reduces the probability of getting stuck in a local minima.