سوال اول قسمت الف: چون در اکثر مواقع وارون ماتریس X وجود ندارد و در این موقع از شبه وارون ماتریس X یعنی X^+ یا به بیان دقیق‌تر از وارون مور-پنروز استفاده می‌کنیم. ابتدا ماتریس X را به صورت ۳ ماتریس $U\Sigma V^T$ تجزیه می‌کنیم با روش SVD سپس شبه وارون را از رابطه زیر بدست می‌آوریم:

$$\underbrace{X = U\Sigma V^T}_{} \qquad X^+ = V\Sigma^+ U^T$$

همچنین برای محاسبه ماتریس Σ^+ ما، ماتریس Σ را به عنوان در نظر گرفته و تمام عناصر آن را با یک مقدار آستانه مقایسه می‌کنیم اگر از آستانه کمتر بود آن‌گاه صفر در نظر می‌گیریم سپس عناصر غیر صفر را معکوس می‌کنیم و در آخر ماتریس بدست آمده را استفاده می‌کنیم و Σ^+ را بدست می‌آوریم و به جای Σ می‌گذاریم و X^+ را محاسبه می‌کنیم و در نهایت $X^+ y = w$ می‌شود. نکته این است که شبه وارون همیشه تعریف شده است.

سوال اول قسمت ب: ابتدا ما باید تابع تخمین زننده را بدست آوریم که با این داده‌ها fit شوند. ما با استفاده از پکیج sklearn این تابع را به صورت زیر بدست آوردیم:

$$\hat{y} = 0.975838368x + 12.752268015847438 19$$

سپس فرمول Durbin-Watson را تقسیم:

$$e_T = (y - \hat{y}) \quad , \quad d = \frac{\sum_{T=2}^{T}(e_T - e_{T-1})^2}{\sum_{T=1}^{T} e_T^2}$$

$$\alpha = 0.05 \quad , \quad \sum_{T=1}^{T} e_T^2 = 47824.55449526$$

$e_{1T} = 70.428117302$

$e_{2T} = 62.58009813$

$e_{3T} = 64.32452622$

$e_{4T} = 66.27378265$

$e_{5T} = 68.13029559$

$e_{6T} = 70.58197219$

$e_{7T} = 72.623984$

$e_{8T} = 73.6925739$

$e_{9T} = 75.83941812$

$e_{10T} = 74.13219965$

$$\hat{\rho} = \frac{(e_2-e_1)^2 + (e_3-e_2)^2 + \cdots + (e_{10}-e_9)^2}{\sum_{T=1}^{T} e_T^2}$$

$$= \frac{30.67335709}{47824.55449526} = \boxed{0.00064135 \simeq 0}$$

$$d \simeq 2(1-\hat{\rho}) = 2(1-0.0006) \approx \boxed{2} \rightarrow d \simeq 2, \rho = 0$$

d_L, d_U ?

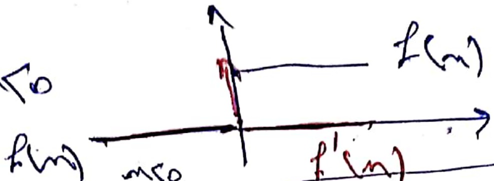
$d > d_U$? DO NOT REJECT H_0 ⟹ no first order serial correlation

$d_U > d > d_L$? INCONCLUSIVE

$d < d_L$? REJECT H_0

سوال دوم : ۲

Binary step : $\begin{cases} f(x) = 1 & x \geq 0 \\ f(x) = 0 & x < 0 \end{cases}$ (0 → threshold)



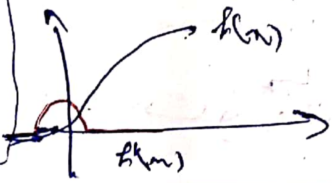
$f'(x) \begin{cases} 0 & x > 0 \\ \delta(0) & x = 0 \\ 0 & x < 0 \end{cases}$

① ✓
② ✗
③ ✓
④ ✓ نزدیک صفر

① : مشکل گرادیان صفر — مشتق ندارد
② متقارن است
③ هزینه محاسبه کم
④ مشتق مدنی است

sigmoid : $\sigma(x) = \frac{1}{1 - e^{x}}$

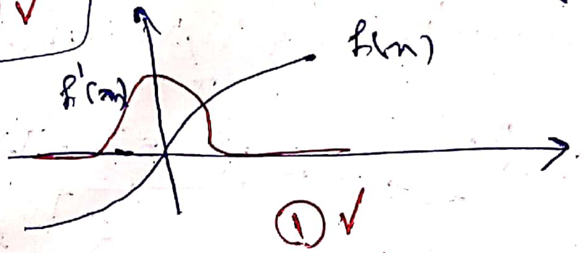
$\sigma'(x) = \sigma(x)(1 - \sigma(x))$



sigmoid
① ✓
② ✗
③ ✗
④ ✓

Tanh : $2\sigma(2x) - 1$

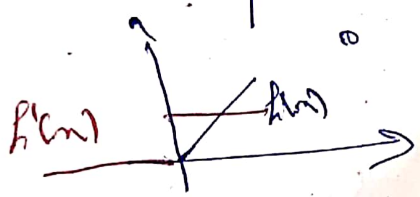
$\text{Tanh}'(x) : 1 - (\text{Tanh}(x)^2)$



① ✓
② ✓
③ ✗
④ ✓

$\text{ReLU}(x) = \max(0, x)$

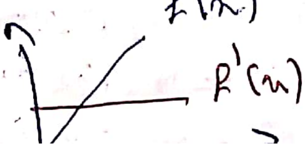
$f'(x) = \begin{cases} 1 & x > 0 \\ ? & x = 0 \\ 0 & x < 0 \end{cases}$



① ✓
② ✗
③ ✓
④ ✓ مرز افقی
در نقطه صفر

leaky ReLU : $\begin{cases} x & x > 0 \\ ax & x < 0 \end{cases}$

$f'(x) : \begin{cases} 1 & x > 0 \\ ? & x = 0 \\ -a & x < 0 \end{cases}$



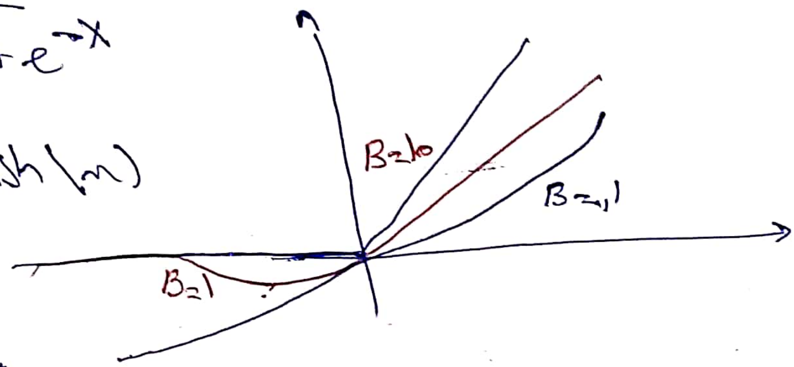
① ✗

سوال دوم قسمت ب:

$y = swish(x) = x\sigma(\beta x)$, β is constant

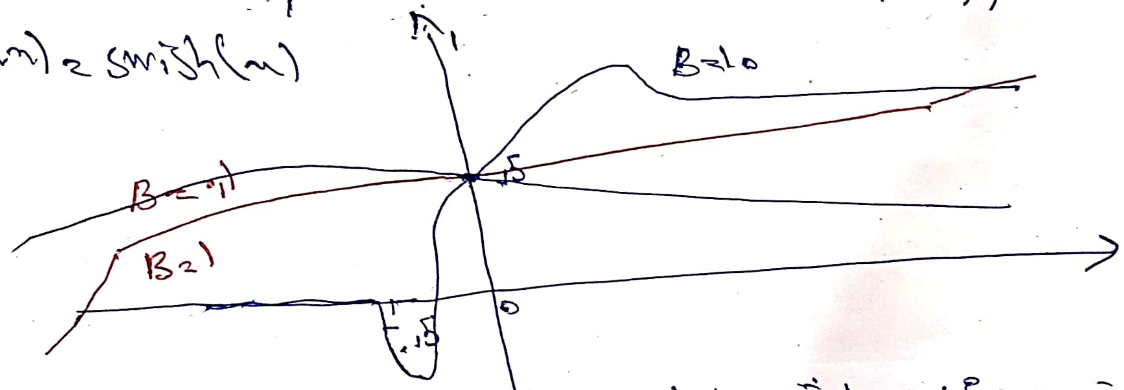
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$f(x) = swish(x)$



$$f'(x) = swish'(x) = \beta f(\beta x) + \sigma(\beta x)(1-\beta f(\beta x))$$

$f'(x) = swish'(x)$



* این تابع پیوسته و در تمام نقطه‌ها مشتق‌پذیر است.
* برخلاف تابع ReLU از مشکل مردن نورون‌ها در حال حرکت رنج نمی‌برد.
* در بعضی از دیتاست‌ها مثل CIFAR و ImageNet از ReLU دقیق‌تر است
* به عنوان یک تابع فعال‌سازی حین اشباع از مشکل انفجار یا ناپدید شدن گرادیان رنج نمی‌برد.
* به‌علاوه یکی از تابع ReLU سریع‌تر عمل کرد لذا اما در مقایسه با بقیه توابع فعال‌سازی عملکرد مشابه دارد
* با توجه به پارامتر β که تأثیرگذار است که ترکیبی متقارن نیست و منجر به محو شدن گرادیان‌ها می‌شود.
* سرعت محاسبه این تابع نسبت به leaky ReLU و Parameterized ReLU کمتر است

$$hard\_swish(x) = x\frac{ReLU6(x+3)}{6}$$

هزینه محاسبه swish به علت استفاده از تابع sigmoid بسیار بیشتر از hard_swish است

سوال دوم قسمت ج: ایده یادگیری چندوظیفه‌ای (multi task learning) و یکی بودن بلوک‌های اولیه این است که قدرت تعمیم‌پذیری را با استفاده از اطلاعات دامنه موجود در سیگنال‌های آموزشی وظایف مربوط به عنوان یک سوگیری استقرایی بهبود می‌بخشد. این کار را با یادگیری کارها به صورت موازی در حین استفاده از یک نمایش مشترک انجام می‌دهد. آنچه برای هر کار آموخته می‌شود می‌تواند به یادگیری بهتر وظایف دیگر کمک کند.

* اشتراک‌گذاری پارامترها (Parameter sharing) در یادگیری چند دامنه‌ای (MDL) به دو روش انجام می‌گیرد که در آن پارامترها را به اشتراک بگذارند تا بتوانند تخصیص دادن منابع را کاهش دهند و در عین حال دقت پیش‌بینی را بهبود ببخشند. یکی از روش‌های اشتراک‌گذاری این است، اشتراک‌گذاری لایه‌های پایین شبکه عصبی عمیق بین دامنه‌ها و در عین حال استفاده از لایه‌های بالاتر جداگانه برای هر دامنه است.

سوال سوم قسمت آ: ① نمودار قرمز مربوط به momentum، نمودار سبز مربوط به Nesterov momentum، نمودار مشکی RMSProp است چون RMSProp در جهت محور عمودی مستقیم حرکت می‌کند و GD تقریباً همیشه سریع‌تر از momentum است. هم momentum هم Nag در ابتدا کند و هنگام ~~~~ نزدیک شدن به مینیمم سراسری سریع حرکت می‌کنند؛

② روش momentum در ابتدا بسیار آرام حرکت می‌کند اما هر چه نزدیک‌تر می‌شود شتاب می‌گیرد تا به نقاط مینیمم سریع‌تر برسد، برخلاف GD که همیشه با گام‌های آرام و منظم به سمت هدف حرکت می‌کند که این باعث کند بودن GD می‌شود و مشکل دیگر GD در نقاط فلات و تخت قرار گیری بسیار کند حرکت می‌کند یا اصلاً حرکت نمی‌کند اما در این نقاط momentum با سرعت بسیار زیاد در این نقاط حرکت می‌کند. و این ویژگی momentum باعث می‌شود که از نقاط تخت عبور کند، برخلاف GD که در آنها گیر می‌کند.

روش Nesterov momentum همان برتری‌های momentum بر GD را دارد. علاوه بر این روش، معمولاً سریع‌تر از momentum عمل می‌کند و به دلیل این قابلیت یعنی به جای محاسبه در موقعیت θ، گرادیان کمی جلوتر از ∇ در جهت momentum اندازه‌گیری می‌شود.

روش RMSProp برخلاف GD که به سمت شیب تند پایین می‌آید، در جهت شیب بُعد مستقیم حرکت می‌کند که مستقیماً به سمت مینیمم سراسری می‌رود و جهت حرکت اول خود را در جهت شیب محور سراسری تنظیم می‌کند.

ادامه سوال ۳ قسمت الف (۳-۱): از آنجا که الگوریتم ~~[illegible]~~ یک الگوریتم با آهنگ یادگیری

ADAM

~~[illegible]~~ تطبیقی است و بیان میکند که به دست آوردن هر ابرپارامتر مثل آهنگ یادگیری η توسط کاربر داده شود؟

نیازی ندارد. نسبت به الگوریتم های قاعده گرادیان است. این الگوریتم میانگین متحرک نمایی

گرادیان های قبلی را محاسبه می‌کند و ~~[illegible]~~ میانگین متحرک نمایی مجذور گرادیان های

گذشته را هم محاسبه خواهد کرد.

سوال سوم قسمت ج :

(هم چنین نورون‌های لایه مخفی)

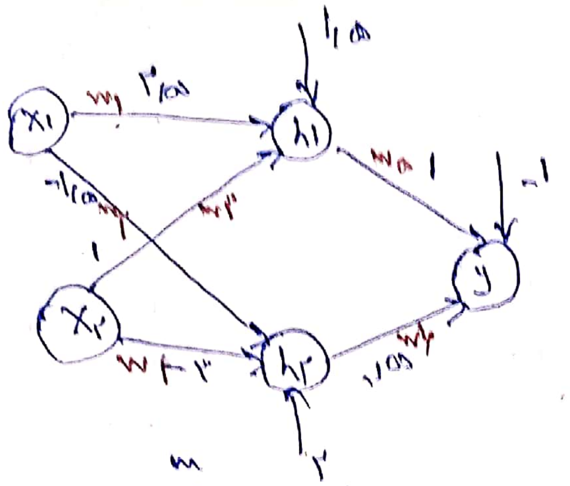
① چون هنگام استفاده از dropout برخی از نورون‌ها در هر epoch به طور تصادفی با احتمال p غیر فعال می‌شوند پس نقش آن‌ها به طور کامل در این epoch در شبکه کم می‌شود. و از طرفی در تمام ایپاک‌ها غیر فعال نمی‌شوند. نورون‌های دیگر که به روش صحیح آموزش داده می‌شوند همچنین نورون‌های مجاور خود وابسته نمی‌شوند و قادر هستند مستقل و جدا هستند. آن‌ها مخصوص به قسمت خاصی نمی‌شوند ورودی خاص وابسته نیستند و مجبور می‌شوند به تمام نورون‌های ورودی توجه کنند که نتیجه آن حساسیت کمتر به تغییرات ورودی است. در نهایت مشابه این است که به این صورت آموزش داده می‌شود و مدل تعمیم پذیری بیشتری خواهیم داشت که ملاک هم نقاط ضعف تکیه که باعث افزایش بیش قدرت تعمیم نشود همان منظم سازی است

② ما با استفاده از dropout در هر بار (ایپاک) آموزش برخی نورون‌ها را به صورت تصادفی با احتمال p غیر فعال می‌کنیم که این باعث به وجود آمدن شبکه جدیدی در هر بار آموزش می‌شود. در مجموع 2^n شبکه جدید داریم با فرض اینکه n تعداد نورون‌ها باشد که حذف می‌شوند و این مانند تکنیک یادگیری جمعی است که ما k مدل متفاوت داریم و در نهایت پیش‌بینی dropout میانگین این شبکه‌های مختلف گزارش می‌شود که می‌توان مطمئن‌ترین مدل را در نهایت انتخاب کرد که این شبیه یادگیری جمعی با تکنیک رای‌گیری است.

③ ما فقط از تکنیک حذف نورون‌ها (غیر فعال سازی) در مرحله آموزش استفاده می‌کنیم و در مرحله تست نورون‌ها را به این صورت حذف نمی‌کنیم. و در مرحله پیش‌بینی و شناسایی نورون را حذف نمی‌کنیم و در مرحله Refrence از یک mask استفاده می‌کنیم.
Refrence

①

سوال ۴ قسمت ۲ ب:



Forward:

$$\hat{y} = \sigma(\underbrace{h_1 \times w_5 + h_2 \times w_6 - 1}_{z}) = \sigma(1.05) = 0.74$$

$$h_1 = x_1 w_1 + x_2 w_3 + 1.5 = 0 + 2.5 = \sigma(2.5) = 0.92$$

$$h_2 = x_1 w_2 + x_2 w_4 + 2 = 0 - 1 = \sigma(-1) = 0.27$$

$$f(z)_{ACT} = \sigma(z) \quad , \quad f'(z) = \sigma(z)(1-\sigma(z))$$

$\alpha = 0.01$

$x_1 = 0$

$x_2 = 1$

$y = 1$

$$J(w) = \sum_{i=1}^{m} \left[-y^i \ln h(x^i) - (1-y^i) \ln(1-h(x^i)) \right] =$$

Backward:

$$\frac{\partial J(w)}{\partial w_5} = \frac{\partial J(w)}{\partial f(z) = h(x^i)} \times \frac{\partial h(x^i) = f(z)}{\partial x^i = z} \times \frac{\partial x^i = z}{\partial w_5} = \left(\frac{-y}{a} + \frac{(1-y)}{1-a}\right) \times \sigma(z) \times (1-\sigma(z))$$

$$h(x^i) = f(z) = a = \hat{y} \quad \times w_5 = \left(\frac{-1}{0.62} + \frac{2}{0.38}\right) \times 0.62 \times 0.38 \times 1 = \frac{-38}{100} + \frac{124}{100} = \frac{86}{100}$$

$$\frac{\partial f(z)}{\partial z} = \sigma(z) \times (1-\sigma(z)) = 0.62 \times 0.38 = 0.2356$$

$$\frac{\partial z}{\partial w_5} = h_1 \times w_5 + h_2 \times w_6 - 1 = w_5 = 1$$

$$w_5 \leftarrow w_5 - \alpha \frac{\partial J_w}{\partial w_5} = 1 - 0.01 \times \frac{86}{100} = 1 - \frac{86}{1000} = \frac{14}{1000} \checkmark$$

$$\frac{\partial J(w)}{\partial w_6} = \frac{\partial J(w)}{\partial f(z)} \times \frac{\partial f(z)}{\partial z} \times \frac{\partial z}{\partial w_6} = \left(\frac{-38}{100} + \frac{124}{100}\right) \times -\frac{1}{2} = \frac{-7}{1000}$$

$$w_6 \leftarrow w_6 - \alpha \frac{\partial J_w}{\partial w_6} = \frac{1}{2} - 0.01 \times \frac{-7}{1000} = \frac{1}{2} + \frac{7}{10000} = 0.50007 \checkmark$$

$$\frac{\partial J(w)}{\partial w_1} = \frac{\partial J(w)}{\partial f(z)} \times \frac{\partial f(z)}{\partial z} \times \frac{\partial z}{\partial h_1} \times \frac{\partial h_1}{\partial w_1} = \left(\frac{-38}{100} + \frac{124}{100}\right) \times \frac{14}{1000} \times x_1 = 0 = 0$$

$$w_1 \leftarrow w_1 - 0 = w_1 \checkmark$$

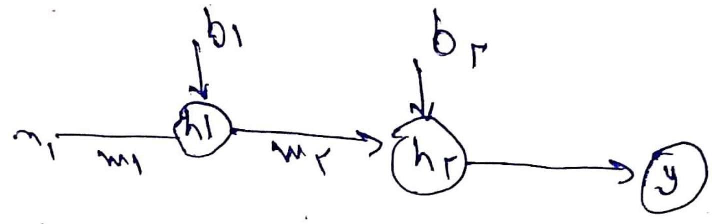
①

سوال چهارم قسمت ب

ابتدا شکل گراف را رسم می‌کنیم:

$y_i = 1$

$X = [x_1]$



$$\frac{\partial J}{\partial w_1}$$

$$J = -\sum y_i \log(\hat{y}_i)$$

$$\frac{\partial J}{\partial w_1} = \frac{\partial J}{\partial \text{Softmax}(h_2)} \times \frac{\partial \text{Softmax}(h_2)}{\partial h_2} \times \frac{\partial h_2}{\partial \text{ReLU}(z_2)} \times \frac{\partial \text{ReLU}(z_2)}{\partial z_2} \times \frac{\partial z_2}{\partial h_1} \times \frac{\partial h_1}{\partial \text{ReLU}(z_1)} \times \frac{\partial \text{ReLU}(z_1)}{\partial z_1} \times \frac{\partial z_1}{\partial w_1}$$

$$\frac{\partial J}{\partial \text{Softmax}(h_2)} = \frac{-y}{\hat{y}} + \frac{1-y}{1-\hat{y}}$$

$$\frac{\partial \text{Softmax}(h_2)}{\partial h_2} = \frac{\partial \frac{e^{a_i}}{\sum_{k=1}^{N} e^{a_k}}}{\partial a_j} = \frac{\partial S_i}{\partial a_j} = S_i(\delta_{ij} - S_j)$$

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\cancel{\frac{\partial h_2}{\partial \text{ReLU}(z_2)}} \quad \frac{\partial \text{ReLU}(z_2)}{\partial z_2} = \begin{cases} 1 & z_2 > 0 \\ ? & z_2 = 0 \\ 0 & z_2 < 0 \end{cases}$$

$$\frac{\partial h_2}{\partial \text{ReLU}(z_2)} = 1 \text{ و } \frac{\partial z_2}{\partial h_1} = w_2 \text{ , } \frac{\partial h_1}{\text{ReLU}(z_1)} = 1 \text{ و } \frac{\partial z_1}{w_1} = x \text{ و } \frac{\partial z_2}{w_2} = h_1$$

$$\frac{\partial J}{\partial w_2} = \frac{\partial J}{\partial \text{Softmax}(h_2)} \times \frac{\partial \text{Softmax}(h_2)}{\partial h_2} \times \frac{\partial h_2}{\partial \text{ReLU}(z_2)} \times \frac{\partial \text{ReLU}(z_2)}{\partial z_2} \times \frac{\partial z_2}{w_2}$$

۱

ادامه سوال چهارم، قسمت ج:

$$* = \frac{\partial J}{\partial b_1} = \frac{\partial J}{\partial softmax(h_2)} \times \frac{\partial softmax(h_2)}{\partial h_2} \times \frac{\partial h_2}{\partial ReLU(z_2)} \times \frac{\partial ReLU(z_2)}{\partial z_2}$$

$$\times \frac{\partial z_2}{\partial h_1} \times \frac{\partial h_1}{\partial ReLU(z_1)} \times \frac{\partial ReLU(z_1)}{z_1} \times \frac{\partial z_1}{b_1}$$

$$\frac{\partial z_1}{\partial b_1} = 1$$

$$\frac{\partial J}{\partial b_2} = \frac{\partial J}{\partial softmax(h_2)} \times \frac{\partial softmax(h_2)}{\partial h_2} \times \frac{\partial h_2}{\partial ReLU(z_2)} \times \frac{\partial ReLU(z_2)}{\partial z_2} \times \frac{\partial z_2}{b_2}$$

$$\frac{\partial z_2}{b_2} = 1 \quad , \quad \frac{\partial J}{\partial b_1} \times \frac{\partial J}{\partial b_2} = \left(\frac{\partial J}{\partial softmax} \times \frac{\partial softmax}{\partial h_2} \times \frac{\partial h_2}{\partial ReLU(z_2)} \times \frac{\partial ReLU(z_2)}{\partial z_2} \right)^2$$

$$\times \frac{\partial z_2}{\partial h_1} \times \frac{\partial h_1}{\partial ReLU(z_1)} \times \frac{\partial ReLU(z_1)}{z_1} \times \frac{\partial z_1}{b_1} \times \frac{\partial z_2}{b_2}$$

(۲)

ادامه سوال ۴ قسمت ج ه

$$\frac{\nabla J(w)}{w_3} = \frac{\nabla J(w)}{f(z)} \times \frac{f(z)}{z} \times \frac{z}{h_1} \times \frac{h_1}{w_3} = \left(\frac{-38}{1000} + \frac{124}{100}\right) \times \frac{14}{1000} \times 1 = 0.001$$

$$w_3 \Leftarrow w_3 - \frac{1}{10} \times 0.001 = 1 - .9999$$

$$\frac{\nabla J(w)}{w_2} = \frac{\nabla J(w)}{f(z)} \times \frac{f(z)}{z} \times \frac{z}{h_2} \times \frac{h_2}{w_2} = \left(\frac{-38}{1000} + \frac{124}{100}\right) \times \frac{1}{2} \times 0 = 0$$

$$w_2 \Leftarrow w_2 - 0 = w_2$$

$$\frac{\nabla J(w)}{w_4} = \frac{\nabla J(w)}{f(z)} \times \frac{f(z)}{z} \times \frac{z}{h_2} \times \frac{h_2}{w_4} = \left(\frac{-38}{1000} + \frac{124}{100}\right) \times \frac{1}{2} \times \frac{\cancel{6}^{1}}{\cancel{4}} = 0.601$$

$$w_4 \Leftarrow w_4 - \frac{1}{10} \times 0.601 = -3 = -3.06$$
