1.

For counter $= 1$ upto Niter

- Choose an integer k uniformly at random from $\{1, 2, 3, \ldots, N\}$
- $x^{\{k\}}$ is current training data point
- $a^{[1]} = x^{\{k\}}$
- For $l = 2$ upto L
  - $z^{[l]} = W^{[l]} a^{[l-1]} + b^{[l]}$
  - $a^{[l]} = \sigma\left(z^{[l]}\right)$
  - $D^{[l]} = \operatorname{diag}\left(\sigma'(z^{[l]})\right)$
- end
- $\delta^{[L]} = D^{[L]}\left(a^{[L]} - y(x^{\{k\}})\right)$
- For $l = L - 1$ downto 2
  - $\delta^{[l]} = D^{[l]}(W^{[l+1]})^T \delta^{[l+1]}$
- end
- For $l = L$ downto 2
  - $W^{[l]} \to W^{[l]} - \eta\, \delta^{[l]} {a^{[l-1]}}^T$
  - $b^{[l]} \to b^{[l]} - \eta\, \delta^{[l]}$
- end

end

The picture[1] above is the pseudo code of SGD.

Now, we just follow the steps to calculate θ^1.

Take $a^{[1]} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.

For $l = 2$ to 2, so we only need to deal with the case $l = 2$.

$z[2] = W^{[2]} a^{[1]} + b^{[2]} = \begin{pmatrix} 5 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + 4 = 21.$

$a^{[2]} = \sigma(z^{[2]}) = \sigma(21)$, where $\sigma(x) = \dfrac{1}{1 + e^{-x}}$, the sigmoid function.

$D = \operatorname{diag}(\sigma'(z^{[2]})) = \sigma(21)(1 - \sigma(21))$

$\delta^{[2]} = \sigma(21)(1 - \sigma(21))(\sigma(21) - 3).$

The next for loop start from $l = 1 < 2$, so we can ignore it.

For $l = 2$ to 2, again, we only need to deal with the case $l = 2$.

Substitute $W^{[2]} - \eta \delta^{[2]} (a^{[1]})^T = \begin{pmatrix} 5 & 6 \end{pmatrix} - \eta \cdot \sigma(21)(1 - \sigma(21))(\sigma(21) - 3) \begin{pmatrix} 1 & 2 \end{pmatrix}$ for $W^{[2]}$.

On the other hand, substitute $b^{[2]} - \eta \delta^{[2]} = 4 - \eta \cdot \sigma(21)(1 - \sigma(21))(\sigma(21) - 3)$ for $b^{[2]}$.

Hence, we have $\theta^1 = \begin{pmatrix} 4 - \eta \cdot k & 5 - \eta \cdot k & 6 - 2\eta \cdot k \end{pmatrix}$, where $k = \sigma(21)(1 - \sigma(21))(\sigma(21) - 3)$.

[1] The picture is captured from

Deep Learning: An Introduction for Applied Mathematicians/ Authors: Catherine F. Higham and Desmond J. Higham

2. (a) Recall that $\sigma(x) = \dfrac{1}{1+e^{-x}} = \dfrac{e^x}{e^x+1}$.

$$\begin{aligned}
\sigma'(x) &= -\frac{1}{(1+e^{-x})^2} \cdot (-e^{-x}) \\
&= \frac{e^{-x}}{1+2e^{-x}+e^{-2x}} \\
&= \frac{e^x}{e^{2x}+2e^x+1} \\
&= \frac{1}{e^x+1} \cdot \frac{e^x}{e^x+1} \\
&= \sigma(x)(1-\sigma(x)) \\
&= -(\sigma(x))^2 + \sigma(x) \\
&= -\sigma^2 + \sigma \\
\sigma''(x) &= -2\sigma\sigma' + \sigma' \\
&= (-2\sigma+1)(-\sigma^2+\sigma) \\
&= \sigma(1-\sigma)(1-2\sigma) \\
\sigma'''(x) &= (6\sigma^2 - 6\sigma + 1) \cdot \sigma' \\
&= (6\sigma^2 - 6\sigma + 1)\sigma(1-\sigma)
\end{aligned}$$

(b) Recall that $\tanh x = \dfrac{e^{2x}-1}{e^{2x}+1}$.

Observe that the denominator is similar.

$$\begin{cases}
\sigma(2x) &= \dfrac{e^{2x}}{e^{2x}+1} \\
1-\sigma(2x) &= \dfrac{1}{e^{2x}+1}
\end{cases}$$

Hence, we have

$$\tanh x = 2\sigma(2x) - 1$$

3. I guess that the next questions may be

   (a) Is the SGD convergent?

   (b) If so, the point it converges to must be the global minima?

   (c) If it is not the global minima, how do we fix the "stop"?

   Actually, I have already heard something about these questions and I just leave a brief reference answer here.

   For (a), yes, but I have never proved it.

   For (b), no, it might be saddle point, or in general case, most of them are just saddle points [2].

   However, the graph of the Coss function may be rugged, so the problem we met more frequently is we can't even touch the saddle point, unless we do the for loop with plenty of times.

   Thus, also answer (c) simultaneously, we can use the concept of momentum [3] in Physics, like substitute the gradient we minus now with the "general" one which considering the former or the sum of all gradients we have now or change η in each l, for example, divide it by the norm of the gradient [4].

[2] See What Happened When Gradient is almost zero?
[3] See 【機器學習2021】類神經網路訓練不起來怎麼辦(二)：批次(batch) 與動量(momentum) 22:37
[4] See 【機器學習2021】類神經網路訓練不起來怎麼辦(三)：自動調整學習速率(Learning Rate) 00:00