Student Name: Mingqian Liao
Student ID: 3038745426

Deep Neural Networks(EECS182)

Homework2

---

1. Why Learning Rates Cannot be Too Big

(a) From the original formula, we can derive that

$$w_{t+1} - \frac{y}{\sigma} = (1 - 2\eta\sigma^2)(w_t - \frac{y}{\sigma})$$

If the recurrence is stable, then the $|1 - 2\eta\sigma^2|$ is less than 1. So that we have $-1 < 1 - 2\eta\sigma^2 < 1$, therefore, $0 < \eta < \frac{1}{\sigma^2}$.

(b) Assume that we get within a factor $1-\epsilon$ of w^*, we have $(1-2\eta\sigma^2)^t|w_0 - \frac{y}{\sigma}| = \epsilon|\frac{y}{\sigma}|$, when $w_0 = 0$, then we have $t \log(1 - 2\eta\sigma^2) = \log \epsilon$, therefore, $t = \frac{\log|\epsilon|}{\log|(1-2\eta\sigma^2)|}$

(c) Similar to (a), we can derive two equations,

$$|1 - 2\eta\sigma_l^2| < 1 \Longrightarrow \eta < \frac{1}{\sigma_l^2}$$

$$|1 - 2\eta\sigma_s^2| < 1 \Longrightarrow \eta < \frac{1}{\sigma_s^2}$$

Since $\sigma_l \gg \sigma_s$, we have $0 < \eta < \frac{1}{\sigma_l^2}$. The σ_l will limit our learning rate.

(d) If the value of η will change, then we can plot the figure of the η, σ_l, σ_s, so the converging rate will depend on the relative magnitude of the two function $\min\{|1 - 2\eta\sigma_l^2|, |1 - 2\eta\sigma_s^2|\}$, the result of the function is the slower one, and the other is the faster one.

(e) From the figure plotted in (d), we can discover that the fastest overall convergence to the solution is the point that when the two lines intercept. The corresponding η is $\frac{1}{\sigma_s^2+\sigma_l^2}$

(f) No, because the edge variable to determine the convergence rate is the σ_l and σ_s.

(g) Since $X = U\Sigma V^T$, and from the OLS solution we have derived that $w = V(\Sigma^T\Sigma)^{-1}\Sigma^T U^T y$, and we could denote that $\hat{w} = (\Sigma^T\Sigma)^{-1}\Sigma^T\hat{y}$, so that $\hat{w}_i = \frac{1}{\sigma_i} y_i \sigma_i * \hat{w}_i = y_i$, which is obviously relevant to the analysis above.

2. Accelerating Gradient Descent with Momentum

(a)

$$\begin{aligned}
w_{t+1} - w^* &= w_t - \eta z_{t+1} - w^* \\
&= w_t - \eta(1-\beta)z_t - \eta\beta(2X^TXw_t - 2X^Ty) - (X^TX)^{-1}X^Ty \\
&= (I - 2\eta\beta X^TX)w_t - \eta(1-\beta)z_t + 2\eta\beta X^Ty - (X^TX)^{-1}X^Ty \\
&= (I - 2\eta\beta X^TX)(w_t - w_0) - \eta(1-\beta)z_t \\
&= V(I - 2\eta\beta\Sigma^T\Sigma)V^T(w_t - w_0) - \eta(1-\beta)V^Tz_t \\
\Longrightarrow x_{t+1} &= (I - 2\eta\beta\Sigma^T\Sigma)x_t - \eta(1-\beta)a_t
\end{aligned}$$

$$
\begin{aligned}
z_{t+1} &= (1-\beta)z_t + 2\beta(X^TX)(w_t - w^*) \\
&= (1-\beta)z_t + 2\beta V(\Sigma^T\Sigma)V^T(w_t - w^*) \\
\implies a_{t+1} &= (1-\beta)a_t + 2\beta(\Sigma^T\Sigma)x_t
\end{aligned}
$$

Therefore,

$$x_{t+1}[i] = (I - 2\eta\beta\Sigma^T\Sigma)x_t[i] - \eta(1-\beta)a_t[i]$$

$$a_{t+1}[i] = (1-\beta)a_t[i] + 2\beta(\Sigma^T\Sigma)x_t[i]$$

(b) Assume that

$$R = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

From the equation above, we can derive that

$$
\begin{aligned}
a_{t+1}[i] &= aa_t[i] + bx_t[i] \\
x_{t+1}[i] &= ca_t[i] + dx^t[i]
\end{aligned}
$$

Therefore,

$$R = \begin{bmatrix} 1-\beta & 2\beta\Sigma^T\Sigma \\ -\eta(1-\beta) & I - 2\eta\beta\Sigma^T\Sigma \end{bmatrix}$$

(c) Eigenvalue of $R[i]$ is

$$\frac{2 - \beta - 2\eta\beta\sigma_i^2 \pm \sqrt{\beta^2(2\eta\sigma_i^2 - 1)^2 - 8\eta(1-\beta)\beta\sigma_i^2}}{2}$$

When $\beta^2(2\eta\sigma_i^2 - 1)^2 - 8\eta(1-\beta)\beta\sigma_i^2 < 0$, then they are complex.
When $\beta^2(2\eta\sigma_i^2 - 1)^2 - 8\eta(1-\beta)\beta\sigma_i^2 = 0$, then they are repeated and purely real.
When $\beta^2(2\eta\sigma_i^2 - 1)^2 - 8\eta(1-\beta)\beta\sigma_i^2 > 0$, then they are purely real.
(h) From equation(10), we know that $\eta^* = \frac{1}{\sigma_{min}^2 + \sigma_{max}^2}$, so that the largest eigenvalue and the smallest eigenvalue will impact the learning rate, and then influence the gradients and parameters updates.

(i) The gradient descent with Momentum is faster, because it can accumulate the past gradient and push the cost further even if the current gradient is very small even zero.

3. Regulation and Instance Noise

(a)

$$\arg\min_w E[\|\hat{X}w - y\|^2] = \arg\min_w E[\sum_{i=1}^m (\hat{X}_i w - y)^2]$$

$$= \arg\min_w E[\sum_{i=1}^m (X_i w - y)^2 + \sum_{i=1}^m (N_i w)^2 + \sum_{i=1}^m 2(X_i w - y)(N_i w)]$$

$$= \arg\min_w E[\|X_i w - y\|^2] + \sum_{i=1}^{m} E[(N_i w)]^2 + \sum_{i=1}^{m} 2E[\sum_{i=1}^{m}(X_i w - y)(N_i w)]$$

$$= \arg\min_w \frac{1}{m}\|Xw - y\|^2 + \sum_{i=1}^{m} Var[N_i w] = \arg\min_w \frac{1}{m}\|Xw - y\|^2 + \sigma^2 w$$

Therefore, $\lambda = \sigma^2$.

(b) $\frac{\partial L}{\partial w} = \hat{X}^T(\hat{X}w - y)$

$$w_{t+1} = w_t - \eta \hat{X}_t^T(\hat{X}_t w - y)$$

Therefore,

$$E[w_{t+1}] = E[w_t] - E[\eta \hat{X}_t^T(X_t \hat{w_t} - y)]$$

$$E[w_{t+1}] = E[w_t] - \eta E[(x + N_t^T)[(x + N_t)w_t - y]]$$

$$E[w_{t+1}] = E[w_t] - \eta(x^2 E[w_t] - xy + E[\sum_{i=1}^{m} N_{t,i}^2]E[w_t])$$

$$E[w_{t+1}] = E[w_t] - \eta(x^2 E[w_t] - xy + Var[N_t]E[w_t])$$

$$E[w_{t+1}] = E[w_t] - \eta(x^2 E[w_t] - xy + \sigma^2 E[w_t])$$

Therefore,

$$E[w_{t+1}] = (1 - (\eta x^2 + \eta\sigma^2))E[w_t] + \eta xy$$

(c) We can construct that

$$E[w_{t+1}] - \frac{xy}{x^2 + \sigma^2} = (1 - (\eta x^2 + \eta\sigma^2))(E[w_t] - \frac{xy}{x^2 + \sigma^2})$$

Therefore, to make the expectation of the learned weight to converge, $|(1 - \eta x^2 + \eta\sigma^2)| < 1$, so that, $-1 < 1 - (\eta x^2 + \eta\sigma^2) < 1$, we can obtain that $\eta < \frac{2}{x^2+\sigma^2}$.

(d) Converge to $\frac{xy}{x^2+\sigma^2}$, so that $w^* = (x^2 + \sigma^2)^{-1}xy$

4. An Alternate MAP Interpretation of Ridge Regression

   MAP $= \arg\max_w P(w|Y = y)$, since w, Y are Gaussian distributed, the $P(w|Y = y)$ is also Gaussian distributed.

   Therefore,

$$\arg\max_w P(w|Y = y) = E(w|Y = y) = \sum_{wY} \sum_{YY}^{-1} y$$

$$\sum_{wY} = E(wY^T) = E(w(w^T X^T + \sqrt{\lambda}N^T)) = E(ww^T X^T + \sqrt{\lambda}wN^T) = IX^T$$

$$\sum_{YY} = E(YY^T) = E((Xw + \sqrt{\lambda N})(Xw + \sqrt{\lambda}N)^T)$$

$$= XE[(ww^T)x^T] + \lambda E(NN^T) = XIX^T + \lambda I$$

   Therefore, $w^* = \sum_{wY} \sum_{YY}^{-1} = X^T(XX^T + \lambda)^{-1}y$

5. Coding Question: Initialization and Optimizers

   (a) the gradient norm of the "he" initialization is high at first, and it decreases very fast during the iteration. Maybe it's because the variance is dependent to the fan-in size, and it can fit the size well.

6. Homework Process and Study Group
    (a) The tutorial of numpy
    (b) Name: Chuanchen SID: 3038743333
    (c) Approximately 16 hours