Homework 2

Yuanteng Chen

1. Why Learning Rates Cannot be Too Big.

(a) For what values of learning rate $\eta > 0$ is the recurrence (3) stable?

Solution:

$$W_{t+1} = (1-2\eta\sigma^2)W_t + 2\eta\sigma y$$

$$2\eta\sigma y = 2\eta\sigma^2 \cdot \frac{y}{\sigma}$$

$$\therefore \quad W_{t+1} = (1-2\eta\sigma^2)(W_t - \frac{y}{\sigma}) + \frac{y}{\sigma}$$

$$W_1 = (1-2\eta\sigma^2)(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma}$$

$$W_2 = (1-2\eta\sigma^2)(W_1 - \frac{y}{\sigma}) + \frac{y}{\sigma}$$

$$= (1-2\eta\sigma^2)[(1-2\eta\sigma^2)(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma} - \frac{y}{\sigma}] + \frac{y}{\sigma}$$

$$= (1-2\eta\sigma^2)^2(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma}$$

$$\vdots$$

$$W_{t+1} = (1-2\eta\sigma^2)^{t+1}(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma}$$

$|1-2\eta\sigma^2| < 1 \iff$ recurrence is stable

$$\therefore \quad 0 < \eta < \frac{1}{\sigma^2}$$

(b). get within a factor $(1-\varepsilon)$ of w^*

$$\Downarrow$$

$$|W_t - w^*| < \varepsilon |w^*|$$

$\because W_{t+1} = (1-2\eta\sigma^2)^{t+1}(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma}$

$W^* = \frac{y}{\sigma}$

$|W_t - \frac{y}{\sigma}| < \varepsilon |\frac{y}{\sigma}|$

$W_t - \frac{y}{\sigma} = (1-2\eta\sigma^2)^t (W_0 - \frac{y}{\sigma})$

$\because$ initial condition $W_0 = 0$

$\therefore |(1-2\eta\sigma^2)^t(-\frac{y}{\sigma})| < \varepsilon |\frac{y}{\sigma}|$

$\therefore (1-2\eta\sigma^2)^t < \varepsilon$

$\log (1-2\eta\sigma^2)^t < \log\varepsilon$

$\because \log(1-2\eta\sigma^2) < 0$

$\therefore \quad t > \frac{\log\varepsilon}{\log(1-2\eta\sigma^2)}$

(c) $\begin{bmatrix} \sigma_L & 0 \\ 0 & \sigma_S \end{bmatrix} \begin{bmatrix} W[1] \\ W[2] \end{bmatrix} = \begin{bmatrix} y[1] \\ y[2] \end{bmatrix}$

$\sigma_L >> \sigma_S$

initial condition $w=0$, a single learning rate η

Solution: assume $\Sigma = \begin{bmatrix} \sigma_L & 0 \\ 0 & \sigma_S \end{bmatrix}$

same as (d):

$W_{t+1} = (E - 2\eta\Sigma^2) W_t + 2\eta \Sigma y$

$\Sigma^2 = \begin{bmatrix} \sigma_L^2, & 0 \\ 0, & \sigma_S^2 \end{bmatrix}$ $\quad E - 2\eta\Sigma^2 = \begin{bmatrix} 1-2\eta\sigma_L^2 & 0 \\ 0 & 1-2\eta\sigma_S^2 \end{bmatrix}$

$\therefore \begin{cases} |1-2\eta\sigma_l^2| < 1 \\ |1-2\eta\sigma_s^2| < 1 \end{cases}$ $\because \sigma_l >> \sigma_s$ $\therefore \eta < \frac{1}{\sigma_l^2} < \frac{1}{\sigma_s^2}$

(d) depending on η, σ_l, σ_s, which of the two dimensions is converging faster and which one is converging slower?

Solution: in (c) $W_{t+1} = (1-2\eta\sigma^2)^{t+1}(W_0 - \frac{y}{\sigma}) + \frac{y}{\sigma}$

$$\therefore \begin{cases} W[1]_{t+1} = (1-2\eta\sigma_l^2)^{t+1}(W_0 - \frac{y[1]}{\sigma_l}) + \frac{y[1]}{\sigma_l} \\ W[2]_{t+1} = (1-2\eta\sigma_s^2)^{t+1}(W_0 - \frac{y[2]}{\sigma_s}) + \frac{y[2]}{\sigma_s} \end{cases}$$

if $|1-2\eta\sigma_l^2| < |1-2\eta\sigma_s^2| < 1$

then $W[1]$ is converging faster,

else $W[2]$ is converging faster.

(e) when $|1-2\eta\sigma_l^2| = |1-2\eta\sigma_s^2|$, we get the fastest overall convergence to the solution.

$\because \sigma_l >> \sigma_s$ $\therefore$ $2 = 2\eta(\sigma_l^2+\sigma_s^2)$

$$\eta = \frac{1}{\sigma_l^2+\sigma_s^2}$$

(f): the speed of converging of σ_i $(i \neq l, s)$

is between 6_l and 6_s,

so they will not influence the choice of possible learning rates.

(g)

I have no idea ......

2. Accelerating Gradient Descent with Momentum.

$$L(w) = \|y - Xw\|_2^2$$

$$w_{t+1} = w_t - \eta z_{t+1}$$

$$z_{t+1} = (1-\beta) z_t + \beta g_t$$

the gradient descent update:

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

$$w^* = (X^T X)^{-1} X^T y$$

(18) $w_{t+1} = w_t - \eta z_{t+1}$

$$z_{t+1} = (1-\beta) z_t + \beta (2X^T X w_t - 2X^T y)$$

$$x_t = V^T (w_t - w^*)$$

$$a_t = V^T z_t$$

(a) ① $w_{t+1} = w_t - \eta z_{t+1}$

$$V^T w_{t+1} = V^T w_t - \eta V^T z_{t+1}$$

$$V^T w_{t+1}[i] = V^T w_t[i] - \eta V^T z_{t+1}[i]$$

$\because x_t = V^T(w_t - w^*)$

$\therefore V^T w_{t+1} = x_{t+1} + V^T \cdot w^*$

$V^T w_t = x_t + V^T \cdot w^*$

$\therefore x_{t+1} + V^T \cdot w^* = x_t + V^T \cdot w^* - \eta V^T z_{t+1}$

$x_{t+1} = x_i - \eta V^T z_{t+1}$

$\because a_t = V^T z_t$

$\therefore x_{t+1} = x_t - \eta a_{t+1}$

$x_{t+1}[i] = x_t[i] - \eta a_{t+1}[i]$

②. $z_{t+1} = (1-\beta) z_t + \beta (2X^T X w_t - 2X^T y)$

$V^T z_{t+1} = (1-\beta) V^T z_t + \beta V^T (2X^T X w_t - 2X^T y)$

$a_{t+1} = (1-\beta) a_t + \beta V^T (2V\Sigma U^T U \Sigma V^T w_t - 2X^T y)$

$a_{t+1} = (1-\beta) a_t + \beta V^T (2V\Sigma^2 V^T w_t - 2X^T y)$

$a_{t+1} = (1-\beta) a_t + \beta (2V^T V \Sigma^2 V^T w_t - 2V^T X^T y)$

$a_{t+1} = (1-\beta) a_t + \beta (2\Sigma^2 V^T w_t - 2V^T X^T y)$

$\because x_t = V^T(w_t - w^*)$

$V x_t = w_t - w^* \quad \therefore w_t = V x_t + w^*$

$\therefore a_{t+1} = (1-\beta) a_t + \beta (2\Sigma^2 V^T (V x_t + w^*) - 2V^T X^T y)$

$a_{t+1} = (1-\beta) a_t + \beta (2\Sigma^2 x_t + 2\Sigma^2 V^T w^* - 2V^T X^T y)$

$$\because\ w^* = (X^TX)^{-1}X^Ty$$
$$= (V\Sigma^2V^T)^{-1}X^Ty$$
$$= V\Sigma^{-2}V^TX^Ty$$

plug w^* into:

$$a_{t+1} = (1-\beta)a_t + \beta(2\Sigma^2x_t + 2\Sigma^2V^TV\Sigma^{-2}V^TX^Ty - 2V^TX^Ty)$$

$$a_{t+1} = (1-\beta)a_t + \beta(2\Sigma^2x_t + 2V^TX^Ty - 2V^TX^Ty)$$
$$a_{t+1} = (1-\beta)a_t + 2\beta\Sigma^2x_t$$
$$\Downarrow$$
$$a_{t+1}[i] = (1-\beta)a_t[i] + 2\beta\sigma_i^2x_t[i]$$

$$\therefore \begin{cases} w_{t+1} = w_t - \eta z_{t+1} \\ z_{t+1} = (1-\beta)z_t + \beta(2X^TXw_t - 2X^Ty) \end{cases}$$
$$\Downarrow$$
$$\begin{cases} a_{t+1}[i] = (1-\beta)a_t[i] + 2\beta\sigma_i^2x_t[i] \\ x_{t+1}[i] = x_t[i] - \eta a_{t+1}[i] \end{cases}$$

(b) $$\begin{bmatrix} a_{t+1}[i] \\ x_{t+1}[i] \end{bmatrix} = R_i \begin{bmatrix} a_t[i] \\ x_t[i] \end{bmatrix}$$ Derive R_i

$$\because \begin{cases} a_{t+1}[i] = (1-\beta)a_t[i] + 2\beta\sigma_i^2x_t[i] \\ x_{t+1}[i] = x_t[i] - \eta a_{t+1}[i] \end{cases}$$

$$x_{t+1}[i] = x_t[i] - \eta\, a_{t+1}[i]$$

$$= x_t[i] - \eta(1-\beta)a_t[i] - 2\eta\beta\sigma_i^2 x_t[i]$$

$$= (1-2\eta\beta\sigma_i^2)x_t[i] - \eta(1-\beta)a_t[i]$$

$$\therefore \begin{cases} a_{t+1}[i] = (1-\beta)a_t[i] + 2\beta\sigma_i^2 x_t[i] \\ x_{t+1}[i] = (1-2\eta\beta\sigma_i^2)x_t[i] - \eta(1-\beta)a_t[i] \end{cases}$$

$$\therefore R_i = \begin{bmatrix} 1-\beta, & 2\beta\sigma_i^2 \\ \eta(\beta-1), & 1-2\eta\beta\sigma_i^2 \end{bmatrix}$$

c()

$$R_i \vec{x} = \lambda \vec{x}$$

$$(R_i - \lambda E)\vec{x} = 0$$

$$|R_i - \lambda E| = 0$$

$$\begin{vmatrix} 1-\beta-\lambda & 2\beta\sigma_i^2 \\ \eta(\beta-1) & 1-2\eta\beta\sigma_i^2-\lambda \end{vmatrix} = 0$$

$$1-\beta-\lambda-2\eta\beta\sigma_i^2+2\eta\beta^2\sigma_i^2+2\eta\beta\sigma_i^2\lambda-\lambda+\beta\lambda+\lambda^2$$
$$+2\eta\sigma_i^2\beta(1-\beta) = 0$$

$$\lambda^2-(2-\beta-2\eta\beta\sigma_i^2)\lambda+(1-\beta) = 0$$

$$\Delta = (2-\beta-2\eta\beta\sigma_i^2)^2-4(1-\beta)$$

$$\begin{cases} \Delta \geq 0 \iff \text{real eigenvalues} \\ \Delta < 0 \iff \text{complex eigenvalues.} \end{cases}$$

(d). when λ is repeated: $\Delta = 0$

$$(2-\beta-2\eta\beta\sigma_i^2)^2 = 4(1-\beta)$$

$$2-\beta-2\eta\beta\sigma_i^2 = \pm 2\sqrt{1-\beta}$$

$$\eta = \frac{2-\beta \mp 2\sqrt{1-\beta}}{2\beta\sigma_i^2}$$

$\therefore$ highest $\eta = \dfrac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2}$

(e) when λ is real

$$\eta > \frac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2} \quad \text{or} \quad \eta < \frac{2-\beta-2\sqrt{1-\beta}}{2\beta\sigma_i^2}$$

(f) when λ is complex:

$$\frac{2-\beta+2\sqrt{1-\beta}}{2\beta\sigma_i^2} < \eta < \frac{2-\beta-2\sqrt{1-\beta}}{2\beta\sigma_i^2}$$

(g) optimal rate $= \dfrac{(\sigma_{max}/\sigma_{min})^2-1}{(\sigma_{max}/\sigma_{min})^2+1}$

$\sigma_{max}^2 = 5 \quad \sigma_{min}^2 = 0.05$

$\therefore$ rate $= \frac{100-1}{100+1} = \frac{99}{101}$

using ordinary gradient descent:

$$\left(\frac{99}{101}\right)^{T_1} \le 99.5\%$$

using this learning rate with momentum:

in (c):

we got $\lambda_1 \cdot \lambda_2 = \sqrt{1-\beta} = \sqrt{0.9} < 1$

$\therefore$ the higher one of λ_1, λ_2 is $\ge \sqrt{0.9}$

$\therefore$ the convergence rate $r \ge \sqrt{0.9}$

$$(r)^{T_2} \le 99.5\%$$

$$\log (r)^{T_2} \le \log 99.5\%$$

$$T_2 \log r \le \log 99.5\%$$

$$T_2 \ge \frac{\log 99.5\%}{\log r}$$

$T_1 > T_2$ $\left(\sqrt{0.9} < \frac{99}{101}\right)$

3- Regularization and Instance Noise

$\tilde{X}_i = X_i + N_i \quad N_i \sim N(0, \sigma^2 I_n)$

$$\tilde{X} = \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \cdots \\ \tilde{x}_m^T \end{bmatrix} \quad \tilde{x}_i \in R^n \text{ and } y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in R^m$$

$$\underset{w}{\operatorname{argmin}} E[\|\tilde{X}w - y\|^2]$$

(a)

$$E[\|\tilde{X}w - y\|^2]$$

$$= E\left[\sum_{i=1}^{m} (\tilde{x}_i^T w - y_i)^2\right]$$

$$= \sum_{i=1}^{m} E[((x_i + N_i)^T w - y_i)^2]$$

$$= \sum_{i=1}^{m} E[(x_i^T w + N_i^T w - y_i)^2]$$

$$= \sum_{i=1}^{m} E[((x_i^T w - y_i) + N_i^T w)^2]$$

$$= \sum_{i=1}^{m} E[(x_i^T w - y_i)^2 - 2(N_i^T w)(x_i^T w - y_i) + (N_i^T w)^2]$$

$$= \sum_{i=1}^{m} E[(x_i^T w - y_i)^2] - 2E((N_i^T w)(x_i^T w - y_i)) + E(N_i^T w)^2$$

$$= \sum_{i=1}^{m} (x_i^T w - y_i)^2 - 2E[(N_i^T w)(x_i^T w - y_i)] + E(w^T N_i N_i^T w)$$

$\because N_i \in (0, \sigma^2 I_n)$

$\therefore 2E[(N_i^T w)(x_i^T w - y_i)] = 0$

$E[N_i N_i^T] = \sigma^2 I_n$

$$\therefore = \sum_{i=1}^{m} (x_i^T w - y_i)^2 + w^T \sigma^2 I_n w$$

$$= (Xw - y)^2 + \sigma^2 \|w\|^2 \cdot m$$

$$\Downarrow$$

is equivalent to a regularized least squares problem:

$$\underset{w}{\operatorname{argmin}} \frac{1}{m} \|Xw - y\|^2 + \lambda \|w\|^2$$

(b). $\tilde{X_i} = X + N_t$, $N_t \in (0, \sigma^2)$

$L(w) = \frac{1}{2} (\tilde{X} w - y)^2$ $\quad w_0 = 0$

$$\frac{\partial L}{\partial w} = (\tilde{X} w - y) \tilde{X}$$

$$= (w(X + N_t) - y)(X + N_t)$$

$$= w(X + N_t)^2 - y(X + N_t)$$

$$= w(X^2 + 2XN_t + N_t^2) - y(X + N_t)$$

$$\therefore W_{t+1} = W_t - \eta \frac{\partial L}{\partial w}$$

$$= W_t - \eta [W_t (X^2 + 2XN_t + N_t^2) - y(X + N_t)]$$

$$= W_t (1 - \eta (X^2 + 2XN_t + N_t^2)) - \eta y (X + N_t)$$

as N_i is i.i.d. $E(N_i) = 0$ $\quad E(N_i^2) = \sigma^2$

$$E(W_{t+1}) = E(W_t) \cdot E(1 - \eta(X^2 + 2XN_t + N_t^2)) - E(\eta y (X + N_t))$$

$$E(W_{t+1}) = E(W_t) \cdot E(1 - \eta(X^2 + \sigma^2)) - E(\eta y X)$$

$$= E(W_t)(1 - \eta(X^2 + \sigma^2)) - \eta y X$$

(c). For what values of learning rate η do we expect the expectation of the learned weight to converge using gradient descent?

Solution:

gradient descent to converge

$\Leftrightarrow \quad -1 < 1-\eta(x^2+\sigma^2) < 1$

$$0 < \eta < \frac{2}{x^2+\sigma^2}$$

(d) what would we expect $E(W_t)$ to converge as $t \to \infty$? How does this differ from the situation without noise?

Solution:

when $E(W_t)$ to converge:

$$\frac{\partial L}{\partial w} = w(x^2+2xN_t+N_t^2)-y(x-N_t)$$

$$E\left(\frac{\partial L}{\partial w}\right) = w(x^2+\sigma^2)-yx=0$$

$$w = \frac{yx}{x^2+\sigma^2}$$

$$w = \frac{y}{x}\,\frac{1}{1+\frac{\sigma^2}{x^2}}$$

without noise: $w = \frac{y}{x}$

there is a scalar value $\frac{1}{1+\frac{\sigma^2}{x^2}}$ when noise is added to x.

7. (a) CSDN, ChatGPT

(b)

(c) 16 hours.