△ SGD

$w_{t+1} \leftarrow w_t - \alpha \nabla f(w)$

△ SGD momentum

$v_t \leftarrow \beta v_{t-1} + (1-\beta)\nabla f(w)$

$w_{t+1} \leftarrow w_t - \alpha v_t$

△ Adagrad

$g_t = \nabla f(w)$

$G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^T$

$$\begin{bmatrix} w_{t+1,1} \\ w_{t+1,2} \\ \vdots \\ w_{t+1,n} \end{bmatrix} = \begin{bmatrix} w_{t,1} \\ w_{t,2} \\ \vdots \\ w_{t,n} \end{bmatrix} - \alpha \begin{bmatrix} \frac{1}{\sqrt{G_t^{(1,1)}+\epsilon}} & & & \\ & \frac{1}{\sqrt{G_t^{(2,2)}+\epsilon}} & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{G_t}} \end{bmatrix} \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(n)} \end{bmatrix}$$

△ Adam

$g_t \leftarrow \nabla f(w)$

$m_t \leftarrow \beta_1 (m_{t-1}) + (1-\beta_1) g_t$

$v_t \leftarrow \beta_2 (v_{t-1}) + (1-\beta_2) g_t g_t^T$

$\hat{m}_t \leftarrow \frac{m_t}{1-\beta_1}$

$\hat{v}_t \leftarrow \frac{v_t}{1-\beta_2}$

$w_{t+1} \leftarrow w_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t+\epsilon}} g_t$

Information gain :

$$h(S) = -\sum_{i=1}^{n} p(s) \log(P(S))$$

class

$$H(S) - \left(\frac{|S_1|}{|S|} H(S_1) + \frac{|S_2|}{|S|} H(S_2) \right)$$

$$P[\,|E_{in}(h) - E(h)| > \epsilon] \le 2\,|H|\, e^{-2\epsilon^2 N}$$

$$P[\,|E_{in}(h) - E(h)| > \epsilon] \le 4\, m_H(2N)\, e^{-\frac{1}{8}\epsilon^2 N}$$

dichotomies :

split of set
where two subset
are exclusive,
whose union is original set !

Adagrad

$g_t = \nabla f(w)$

$G_t = \sum_{\tau=1}^{t} g_\tau g_\tau^T$

$$\begin{bmatrix} w_{t+1}^1 \\ w_{t+1}^2 \\ \vdots \\ w_{t+1}^n \end{bmatrix} = \begin{bmatrix} w_t^1 \\ w_t^2 \\ \vdots \\ w_t^n \end{bmatrix} - \alpha \begin{bmatrix} \frac{1}{\sqrt{G_t^{(1,1)}+\epsilon}} & & 0 \\ & \frac{1}{\sqrt{G_t^{(2,2)}+\epsilon}} & \\ & & \ddots & \\ 0 & & \frac{1}{\sqrt{G_t^{(n,n)}+\epsilon}} \end{bmatrix} \begin{bmatrix} g_t^{(1)} \\ g_t^{(2)} \\ \vdots \\ g_t^{(n)} \end{bmatrix}$$

Adam

① $g_t = \nabla f(w)$

② $m_t = (1-\beta_1) m_{t-1} + \beta g_t$

$v_t = (1-\beta_2) v_{t-1} + \beta g_t^2$

③ $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$

$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$

④ $w_{t+1} = w_t - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t+\epsilon}} g_t$

$B(N,k)$

N \ k	1	2	3	4	5	6
1	1	2	2	2	2	2
2	1	3	4	4	4	4
3	1	4	7	8	8	8
4	1	5	11	15	16	16
5	1	6	16	26	31	32
6	1	7	22	42	57	63

$$m_H(N) \le B(N,k) \le \sum_{i=1}^{k} \binom{N}{i} \le N^k$$