Subject Year:........ Month:........ Date:........

1) a) در گرادیان عادی فرمول به روزرسانی به شکل زیر است:

$$W_{k+1} = W_k - \eta \nabla J(W_k)$$

اما در momentum به شکل زیر عمل کرده:

$$W_{k+1} = W_k + (\Delta W_{k+1} \times -1)$$

تأثیر فیزیکی:

$$\underbrace{\Delta W_k}_{v_k} = \beta \underbrace{\Delta W_{k-1}}_{v_{k-1}} + \underbrace{\eta \nabla J(W_k)}_{a_k}$$

$$\Delta W_k = a_k + \beta a_{k-1} + \dots + \beta^{k-1} a_1$$

که اصولاً β را ۰٫۹ یا ۰٫۹۹ گرفته و باعث شده تغییرات براساس میانگین وزن‌دار گرادیان‌های قبلی در نظر گرفته شود و هرچقدر گرادیان مربوط به مراحل قبل‌تر باشد، تأثیرش کمتر است. مناطقی که گرادیان در چند مرحله جهتش در الگوریتم کاهش گرادیان یکسان است را

درخشان

در نظر بگیرید، یا به اصطلاح نواحی مسطح و دره‌ای، سرعت الگوریتم momentum بیشتر بوده و سریع‌تر همگرا شود. چرا که جمع یک سری بردار است که هم‌جهت بوده و همدیگر را تقویت کرده‌اند و ΔW بیشتری داشته‌ایم. همچنین در نواحی که انحنا منحنی زیاد است، به‌روزرسانی را هموارتر کرده و نوسانات را کاهش داده چرا که هنگامی که در این نواحی شیب‌ها الگوریتم به جای اینکه تیز شدید شود گرادیان‌های قبلی مانع این شده.

6) فرمول به‌روزرسانی در NAG:

$$W_{k+1} = W_k - \Delta W_{k+1}$$

$$\Delta W_k = \beta \Delta W_{k-1} + \eta \left(\nabla J(\underbrace{W_{k} \cancel{+} - \beta \Delta W_{k-1}}_{W_{k-1} \cancel{+} - \beta \Delta W_{k-1}}) \right)$$

درست نیست

در NAG به جای اینکه تغییرات وزن قبل را به گرادیان آن نقطه اضافه کنیم، ابتدا گرادیان را در یک حرکت جلوتر حساب کرده که همان $\Delta J(W_{k-1}+B\Delta W_{k-1})$ هست را حساب کرده و پس جمع می کنیم. یعنی از نقطه اول به نقطه $W_{k-1}-B\Delta W_{k-1}$ رفته و پس تصمیم گرفته. که به اصطلاح look ahead نام دارد. می توان گفت در NAG دقیق تر به روز رسانی انجام شده و خطر بیش از حد جهش کردن یا overshooting کمتر بوده و بسیار استیبل تر بوده.

c) عملا این کار را به کمک جمع مربعات مشتقات انجام داده. به این گونه که:

$$S_k = S_{k-1} + (\partial_W E_k)^2$$

$$W_k = w_{k-1} - \frac{\eta}{\sqrt{S_k + \epsilon}} (\partial_W E_k)$$

درخشان

از محدودیت هایی که ایجاد می‌کند کاهش شدید نرخ یادگیری به دلیل

جمع مربعات شیب‌هاست در طول زمان که کلا $\forall k: S^{(k+1)} \geq S^{(k)}$

و باعث شدن به روزرسانی کم شود و بسیار آهسته.

d) از میانگین وزن‌دار استفاده کرده و باعث کاهش اضافه

شدن مقادیر بزرگ نسبت به حالت قبلی شده: $S_k = \beta S_{k-1} + (1-\beta)(\nabla E(w_k))^2$

پس مانند AdaGrad اسکیل می‌کند اما S_k ملایم‌تر شدن و باعث

بسیار کند شدن الگوریتم شده.

e) در Momentum از میانگین گرادیان به جای خود گرادیان

استفاده کردیم و داریم:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla J(w_t)$$

در RMSProp سعی کردیم به گونه‌ای نرمال‌سازی را برای نرخ یادگیری

انجام دهیم و داشتیم:

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla J(w_t))^2$$

که عملاً نرخ یادگیری را کنترل می‌کرد برای ایمان.

حال اصلاح‌هایی انجام داده برای جلوگیری از صفر بودن مقادیر اولیه:

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t} \qquad \hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

ترکیب این ۲ برابر Adam است و داریم:

$$w_{k+1} = w_k - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t$$

از ۲ ویژگی First moment, Second moment استفاده کرده.

(a (۲)

$$E\left[z_i^{(h)}\right] = E\left[\sum_{j=1}^{d_h} w_{ij}^{(h)} x_j\right] = \sum_{j=1}^{d_h} E\left[w_{ij}^{(h)} x_j\right]$$

$$\xlongequal{W^{(h)} \perp X^{(h)}} \sum_{j=1}^{d_h} E\left[w_{ij}^{(h)}\right] E\left[x_j^{(h)}\right] \xlongequal{E\left[w_{ij}^{(h)}\right]=0} \sum_{j=1}^{d_h} 0 \times E\left[x_j^{(h)}\right]$$

$$= \boxed{0}$$

(b

$$Var(z_i^{(h)}) = E\left[(z_i^{(h)})^2\right] - \underbrace{E^2\left[(z_i^{(h)})\right]}_{0} =$$

$$E\left[(z_i^{(h)})^2\right] = E\left[\left(\sum_{j=1}^{d_h} w_{ij}^{(h)} x_j^{(h)}\right)^2\right] \overset{1}{=}$$

$$E\left[\sum_{j=1}^{d_h} (w_{ij}^{(h)} x_j^{(h)})^2\right] = \sum_{j=1}^{d_h} E\left[(w_{ij}^{(h)} x_j^{(h)})^2\right] = \sum_{j=1}^{d_h} E\left[(w_{ij}^{(h)})^2 (x_j^{(h)})^2\right]$$

$$= \sum_{j=1}^{d_h} E\left[(w_{ij}^{(h)})^2\right] E\left[(x_j^{(h)})^2\right] \overset{2}{=} \sum_{j=1}^{d_h} Var(w_{ij}^{(h)})\, E\left[(x_j^{(h)})^2\right] =$$

$$\sum_{j=1}^{d_h} \sigma_h^2\, E\left[(x_j^{(h)})^2\right] = \boxed{d_h \sigma_h^2\, E\left[(x_j^{(h)})^2\right]}$$

توضیح رابطه 1 : به ازای k و k' هایی که $k \neq k'$ هست داریم :

$$E\left(w_{ik}^{(h)} w_{ik'}^{(h)} x_k^{(h)} x_{k'}^{(h)}\right) = \underbrace{E(w_{ik}^{(h)})}_{0} \times \dots \times E(x_{k'}^{(h)}) = 0$$

درخشان

توضیح رابطه ۲ : برای محاسبه واریانس داریم :

$$Var(w_{ij}^{(h)}) = E[(w_{ij}^{(h)})^2] - \underbrace{E^2[(w_{ij}^{(h)})]}_{0} = E[(w_{ij}^{(h)})^2]$$

c) ابتدا سمت راست را محاسبه کرده :

$$Var(z_j^{(h-1)}) = E(z_j^{2(h-1)}) - \underbrace{E^2(z_j^{(h-1)})}_{0} =$$

$$E[(z_j^{(h-1)})^2] = \int_{-\infty}^{+\infty} z^2 p(z)\,dz = 2\int_{0}^{+\infty} z^2 p(z)\,dz$$

$$\rightarrow \frac{1}{2} Var(z_j^{(h-1)}) = \boxed{\int_{0}^{+\infty} z^2 p(z)\,dz}$$

حال سمت چپ را محاسبه کرده و داریم :

$$E[(x_j^{(h)})^2] = E[(\sigma(z_j^{(h-1)}))^2] = \int_{-\infty}^{+\infty} \sigma^2(z)\, p(z)\,dz =$$

$$\int_{-\infty}^{0} 0 \times p(z)\,dz + \int_{0}^{+\infty} z^2 p(z)\,dz = \boxed{\int_{0}^{+\infty} z^2 p(z)\,dz}$$

d) برای اینکه واریانس ٤ حفظ شود داریم طبق روابط قبلی:

$$\frac{Var(z_i^{(h)})}{Var(z_i^{(h-1)})} = 1 \xrightarrow{b,c} \frac{d_h \sigma_h^2}{2} = 1 \longrightarrow$$

$$\boxed{\sigma_h^2 = \frac{2}{d_h}}$$

یعنی باید مقداردهی وزن ٤ به شکلی باشد که میانگین صفر بوده و

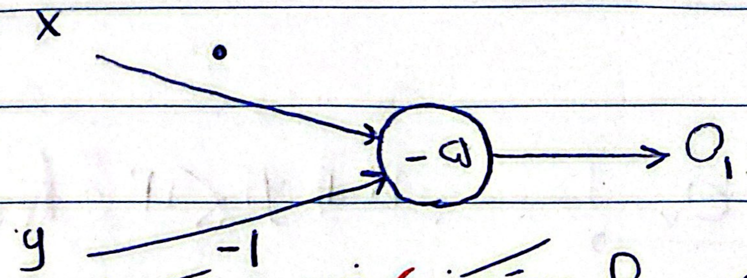
انحراف معیار آن‌ها برابر $\sqrt{\frac{2}{d_h}}$.

درسنامه

۳) a) ابتدا هر ناحیه محدب (مربع و مثلث) را شبکه اش

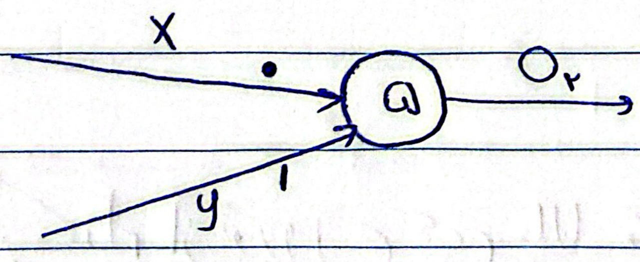
را جداگانه حساب کرده و داریم:

ابتدا ۷ ناحیه را پیاده سازی می کنیم:

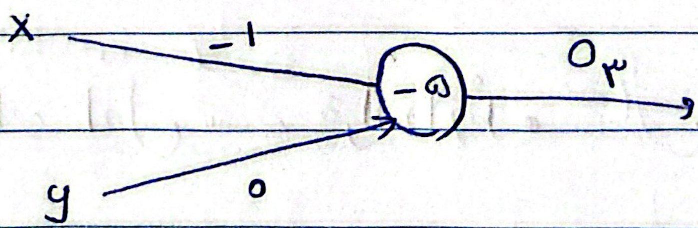


$y \leq 5 : R_1$

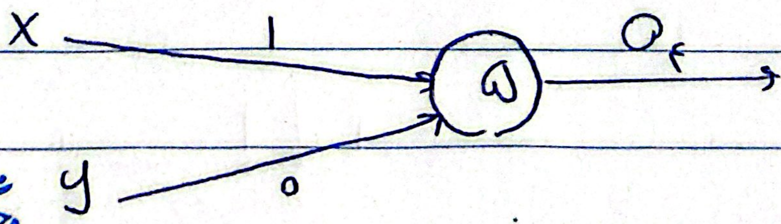
زمانی که ورودی داده شده در ناحیه R_1 صدق کند خروجی یک است.



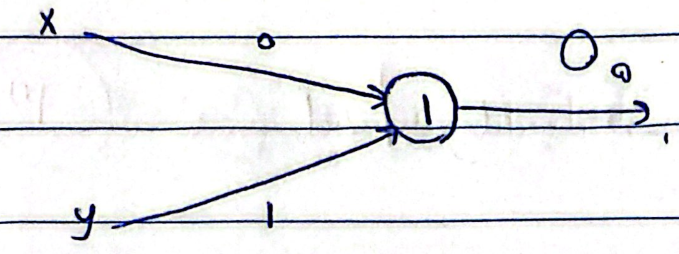
$y \geq 5 : R_2$



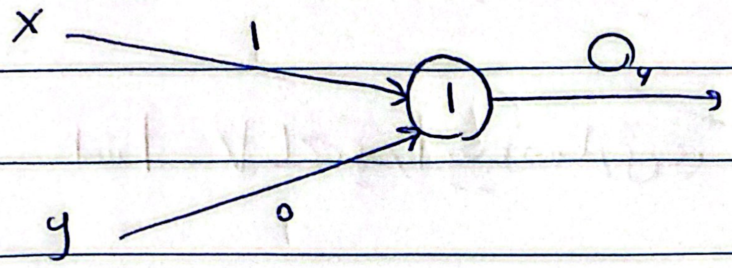
$x \leq 5 : R_3$



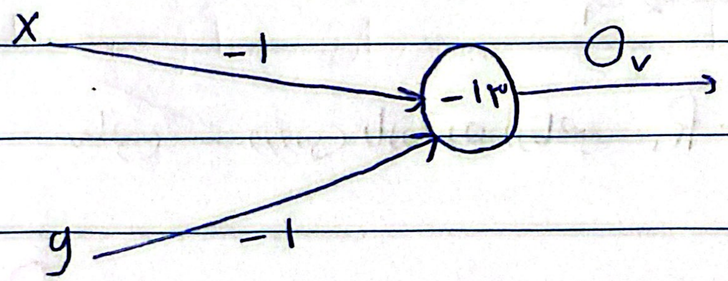
$x \geq 5 : R_4$



$y \geqslant 1 : R_5$



$x \geqslant 1 : R_6$

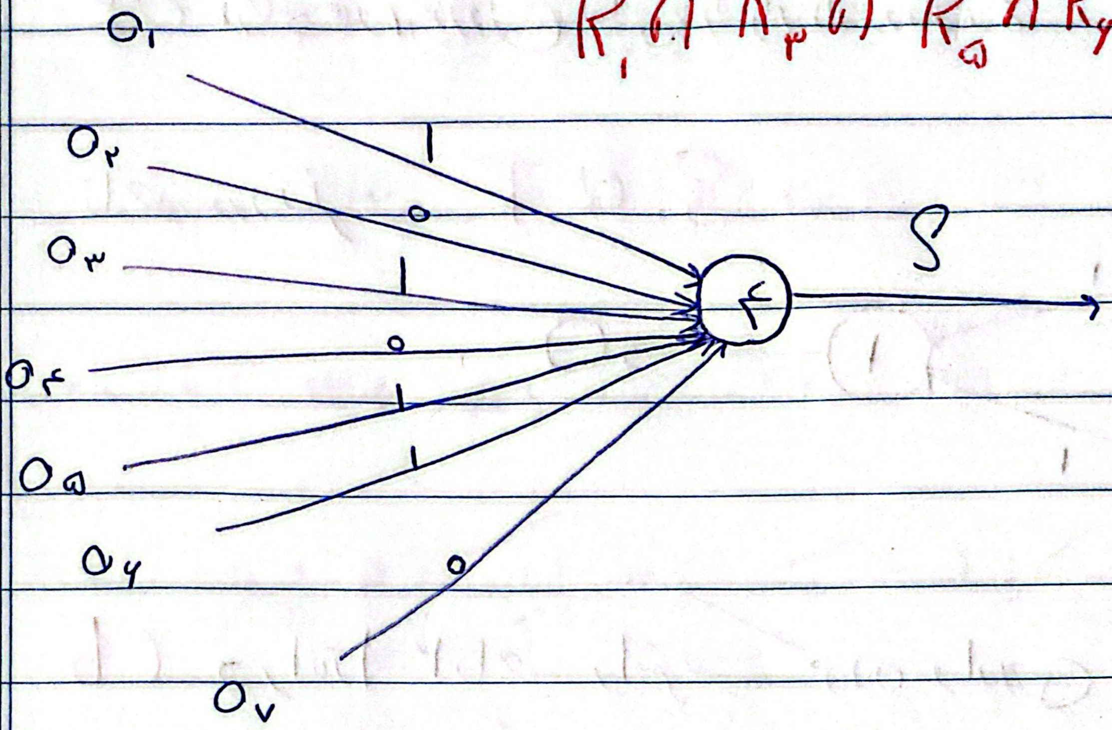


$x + y \leqslant 13 : R_7$

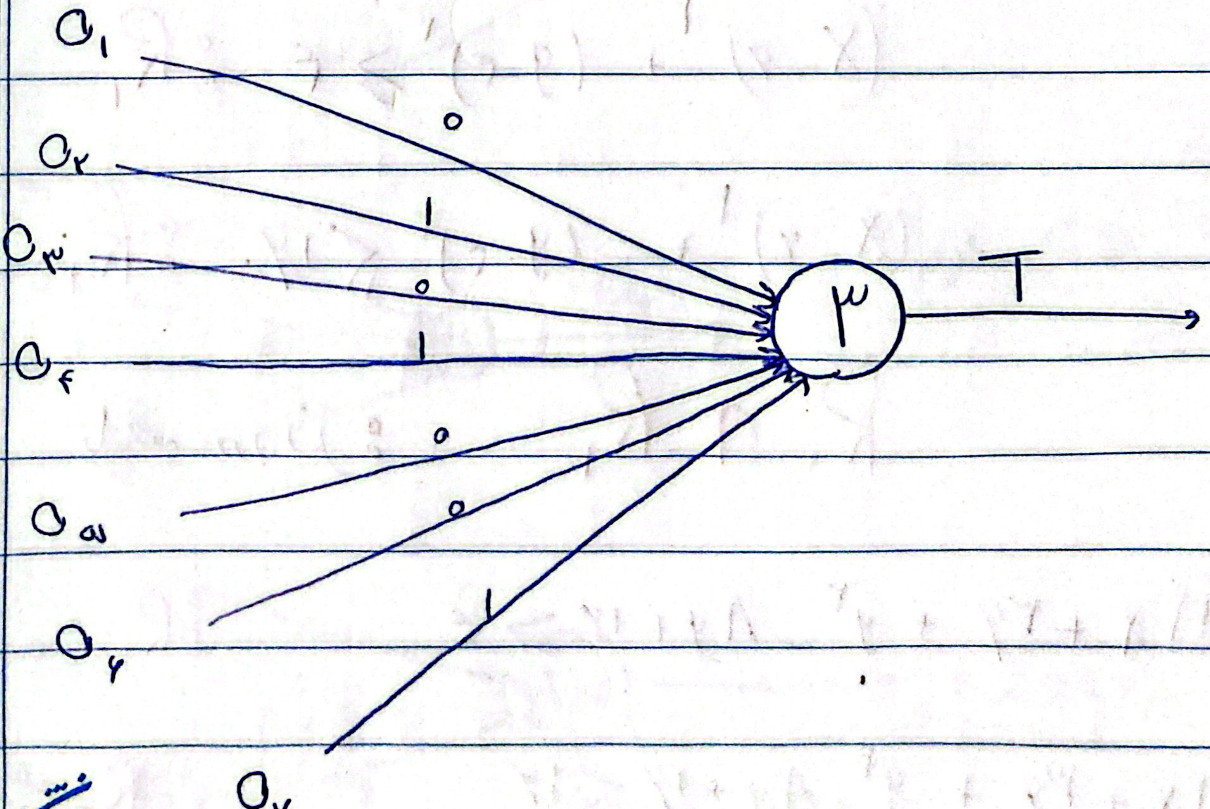
هر کدام از نورون های بالا که ورودی آن ها x و y است در لایه اول شبکه قرار دارند. حال هر کدام از ناحیه های محدب را در لایه دوم شبکه پیاده کرده و داریم:

Subject Year: Month: Date:

$R_1 \cap R_3 \cap R_5 \cap R_6 =$ مربع

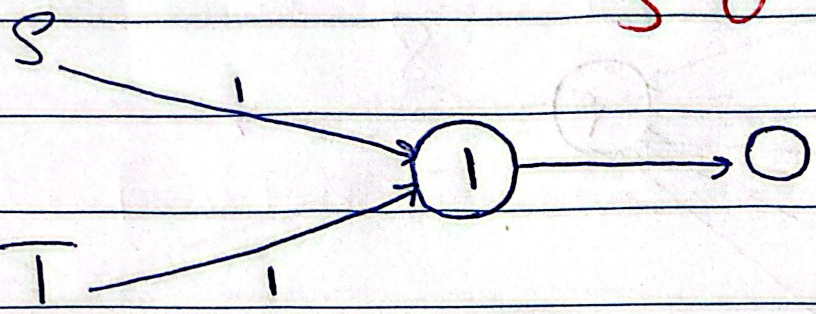


$R_2 \cap R_4 \cap R_7 =$ مثلث



حال در لایه سوم نورون نتایج را اجتماع داده و داریم :

ناحیه مورد نظر = $S \cup T$



b) در ابتدا ۲ ناحیه را مشخص کرده و از روی آن ٤ ناحیه مورد نظر را پیدا کنیم :

$R_1 : (x-6)^2 + (y-4)^2 \geq 4$

$R_2 : (x-6)^2 + (y-4)^2 \leq 16$

ناحیه مورد نظر : $R_1 \cap R_2$

$R_1 : x^2 - 12x + 36 + y^2 - 8y + 16 \geq 4$

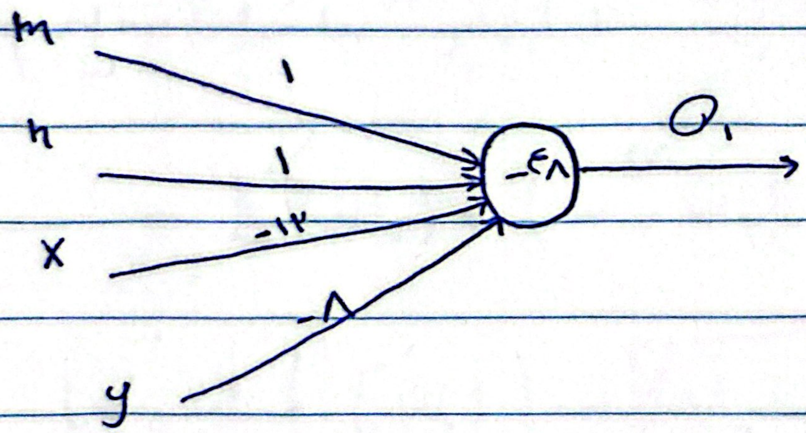
$R_2 : x^2 - 12x + 36 + y^2 - 8y + 16 \leq 16$

حال $x^2 = m$ و $y^2 = n$ گرفته و داریم

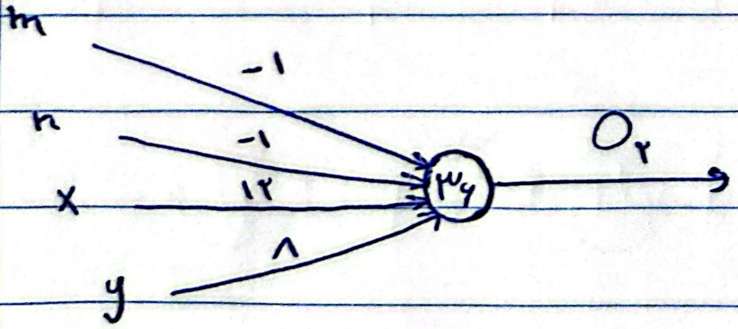
R_1: $m + n - 12x - 8y \geq -48$

R_2: $m + n - 12x - 8y \leq -36$

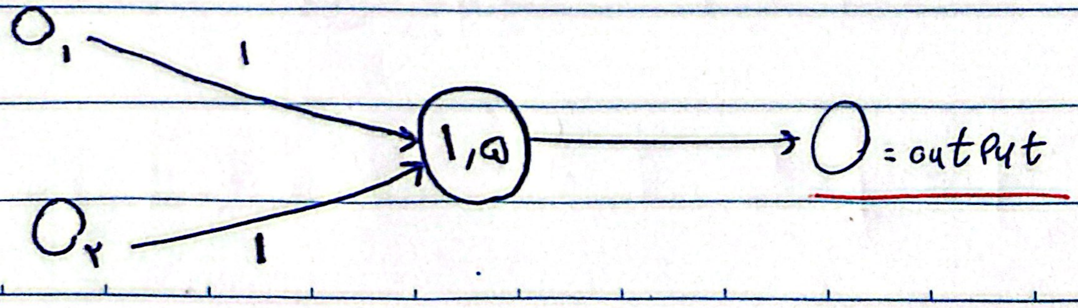
R_1: حال کافی است یک ابر صفحه را پیاده کنیم مانند قبل:



R_2: مانند قبل داریم:



$R_1 \cap R_2$:



Subject Year:........ Month:........ Date:........

(4) (a ابتدا بسط تیلور $E(w)$ را حول w_0 نوشته و داریم:

$$E(w) = E(w_0) + (w-w_0)E'(w_0) + \frac{(w-w_0)^2}{2}E''(w_0)$$

$\xrightarrow{\text{مشتق}}$ $$\frac{\partial E(w)}{\partial w} = 0 + E'(w_0) + (w-w_0)E''(w_0)$$

حال فرض می‌کنیم که w_0 برابر w^* که $w^* = \arg\min_w E(w)$: $\rightarrow E'(w) = 0$

$$E'(w^*) = 0 = E'(w_0) + (w^* - w_0)E''(w_0) \longrightarrow$$

$$w^* = w_0 + \left(-\frac{E'(w_0)}{E''(w_0)}\right) = \boxed{w_0 - \frac{1}{E''(w_0)} \frac{\partial E}{\partial w}\Big|_{w=w_0}}$$

این به این معناست که برای رسیدن به w^* به ازای هر نقطه‌ای (w_0)

$h = \frac{1}{a}$ که یک عدد ثابت است باید تنظیم شود.

a^{-1}

Subject Year: Month: Date:

b) تابع $\cosh x = \frac{e^{x}+e^{-x}}{2}$ به این شکل است. نقطه‌ای

که $E(\omega)=\cosh\omega$ را کمینه کند $\omega=0$ است.

فرض کنید در نقطه‌ی ω_0 هستیم و داریم:

$$\omega_0 - \frac{1}{E''(\omega_0)}E'(\omega_0) = \omega_0 - \frac{e^{\omega_0}-e^{-\omega_0}}{e^{\omega_0}+e^{-\omega_0}}$$

$$= \frac{e^{\omega_0}(\omega_0-1)+e^{-\omega_0}(\omega_0+1)}{e^{\omega_0}+e^{-\omega_0}} \neq 0 \quad \forall \omega_0 \in \mathbb{R}$$

c) کافی است اثبات کنیم که رابطه زیر برقرار است:

$$\left|w_{t+1}+\frac{b}{a}\right| > \left|w_t+\frac{b}{a}\right| \quad \forall t \in \mathbb{N}$$

$$w_{t+1} = w_t + \Delta w = w_t - \frac{k}{a}(aw_t+b)$$

$$\left|w_{t+1}+\frac{b}{a}\right| = \left|w_t - kw_t - \frac{kb}{a} + \frac{b}{a}\right| =$$

Subject Year: Month: Date:

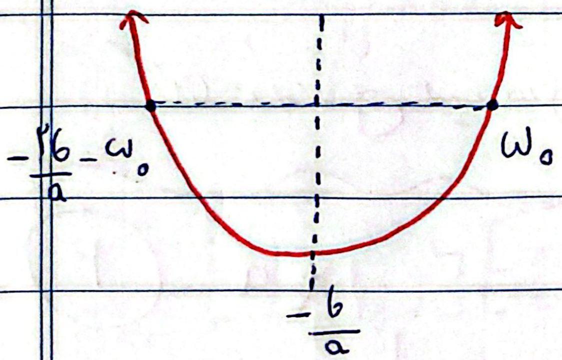
$$|w_t(1-k) + \frac{b}{a}(1-k)| = |1-k||w_t + \frac{b}{a}|$$

طبق فرض مسئله باید $k>2$ باشد:

$$k>2 \rightarrow k-1>1 \rightarrow |k-1| = |1-k| > 1 \rightarrow$$

$$|1-k||w_t+\frac{b}{a}| > |w_t + \frac{b}{a}| \checkmark$$

d) به شکل روبه رو توجه کنید:



اگر h برابر h بهینه باشد، به $-\frac{b}{a}$ می‌رسیم. اگر $h=2h_{opt}$

باشد دقیقاً نسبت به خط $x=-\frac{b}{a}$ قرینه شده وضعیت جدید. پس وقتی

$h_{opt} < h < 2h_{opt}$ باشد بین نقاط $-\frac{b}{a}$ و w_0 ، $-\frac{2b}{a}-w_0$ خواهیم بود که نسبت به

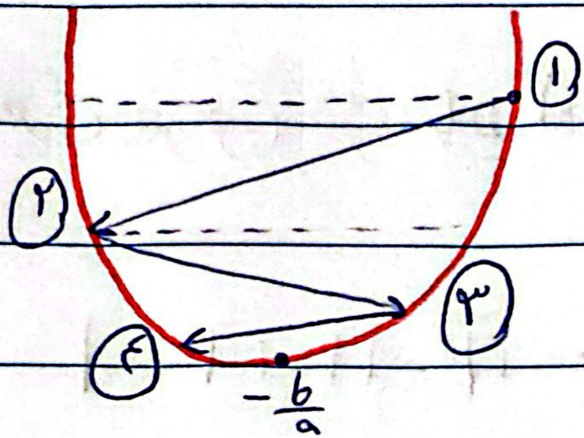
وضعیت قبلی بهتر بوده ولی قرینه شده است نسبت به محور $-\frac{b}{a}$ و پایین‌تر آن.

درخشان

Subject Year:........ Month:........ Date:........

قرارداریم تا زمانی فاصله ما کم و کمتر شده و حول نقطه بهینه نوسان کرده :

مثال :



در این حالت از هر نقطه ای به جز نقطه بهینه شروع کنیم هیچ

گاه به نقطه بهینه نرسیدیم اما فاصله کم تر شده در هر تکرار الگوریتم .

الف) ابتدا $\hat{y}$ را از طریق feed_forward حساب کرده و داریم:

$$(x_1, x_2, x_3, x_4) = (-0.7, 1.2, 1.1, -2)$$

$$S_1 = w_1 x_1 + w_2 x_2 = -1.7 \times -0.7 + 0.1 \times 1.2 = 1.19 + 0.12 = 1.31$$

$$S_2 = w_3 x_3 + w_4 x_4 = -0.9 \times 1.1 + -1.8 \times -2 = -0.99 + 3.6 = 2.61$$

$$h_1 = \frac{1}{1+e^{-s_1}} = 0.79 = \sigma(s_1)$$

$$h_2 = \frac{1}{1+e^{-s_2}} = 0.95 = \sigma(s_2)$$

$$S_3 = w_5 h_1 + w_6 h_2 = -0.2 \times 0.79 + 0.5 \times 0.95 = 0.32$$

$$\hat{y} = \frac{1}{1+e^{-s_3}} = 0.58 = \sigma(S_3)$$

حال $\frac{\partial L}{\partial w_1}$ را محاسبه کرده و داریم:

chain rull:

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial s_3} \times \frac{\partial s_3}{\partial h_1} \times \frac{\partial h_1}{\partial s_1} \times \frac{\partial s_1}{\partial w_1} =$$

$$-2(y-\hat{y}) \times \sigma(s_3)(1-\sigma(s_3)) \times w_5 \times \sigma(s_1)(1-\sigma(s_1)) \times x_1 =$$

$$\underbrace{-2(0.5-0.58)}_{0.16} \times 0.58 \times 0.42 \times -0.2 \times 0.79 \times 0.21 \times (-0.7) = \boxed{9\times 10^{-4}}$$

۶) a توزیع برنولی دارای pmf زیر است:

$$f(r) = \begin{cases} p & r=1 \\ 1-p & r=0 \end{cases}$$

حال داریم:

$$E[R_{ni}] = \sum_r r\,f(r) = 1\times p + 0\times(1-p)$$

$$= \boxed{p}$$

Subject Year:........ Month:........ Date:........

$$E[R_{ni} R_{nj}] :$$

$$\text{if } i=j \longrightarrow E[R_{ni}^2] = \sum_{r} r^2 f(r) = 1 \times \rho = \rho$$

$$\text{if } i \neq j \xrightarrow{\text{مستقل اند}} E[R_{ni}]\, E[R_{nj}] = \rho \times \rho = \rho^2$$

$$\Longrightarrow \boxed{\delta_{ij}\, \rho + (1-\delta_{ij})\, \rho^2} \qquad \delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$$

6) می دانیم که : $\sum_{X_i} E(X) = E \sum_{X_i} (X)$:

$$E[E(W)] = \sum_{n=1}^{N} \sum_{k=1}^{K} E\left[\left\{ y_{nk} - \sum_{i=1}^{D} w_{ki} R_{ni} x_{ni} \right\}^2\right]$$

فرض کنید که داریم : $O_{nk} = \sum_{i=1}^{D} w_{ki} R_{ni} x_{ni} \longrightarrow$

$$E[(y_{nk} - O_{nk})^2] = E[y_{nk}^2 - 2 y_{nk} O_{nk} + O_{nk}^2] =$$

$$y_{nk}^2 - 2 y_{nk} E[O_{nk}] + E[O_{nk}^2]$$

از طرفی می دانیم که : $Var(X) = E[X^2] - E^2[X]$

$$E[O^2_{nk}] = Var(O_{nk}) + E^2[O_{nk}] \longrightarrow$$ پس داریم :

$$E[(y_{nk} - O_{nk})^2] = \underbrace{y^2_{nk} - 2y_{nk}E[O_{nk}] + E^2[O_{nk}]} + Var(O_{nk})$$

$$= (y_{nk} - E[O_{nk}])^2 + Var(O_{nk})$$

از طرفی داریم :

$$E[O_{nk}] = E\left[\sum_{i=1}^{D} w_{ki} R_{ni} x_{ni}\right] =$$

$$\sum_{i=1}^{D} E[w_{ki} R_{ni} x_{ni}] = \sum_{i=1}^{D} w_{ki} x_{ni} E[R] = \rho \sum_{i=1}^{D} w_{ki} x_{ni}$$

$$Var(O_{nk}) = Var\left(\sum_{i=1}^{D} w_{ki} R_{ni} x_{ni}\right) \xrightarrow[i \neq j]{R_{ni} \perp R_{nj}}$$

$$\sum_{i=1}^{D} var(w_{ki} R_{ni} x_{ni}) = \sum_{i=1}^{D} w^2_{ki} x^2_{ni} \, var(R_{ni})$$

$$= \sum_{i=1}^{D} w^2_{ki} x^2_{ni} (\underbrace{E[R^2_{ni}] - E^2_{ni}[R]}_{\rho - \rho^2}) =$$

$$\rho(1-\rho) \sum_{i=1}^{D} w^2_{ki} x^2_{ni}$$

طبق روابط بالا داریم:

$$E\left[E(w)\right] =$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\left(y_{nk}-\rho\sum_{i=1}^{D}w_{ki}x_{ni}\right)^2+\rho(1-\rho)\sum_{i=1}^{D}w_{ki}^2x_{ni}^2\right]=$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\left(y_{nk}-\rho\sum_{i=1}^{D}w_{ki}x_{ni}\right)^2+\rho(1-\rho)\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{i=1}^{D}w_{ki}^2x_{ni}^2$$

C) همانطور که در روابط بالا دیدیم، شبیه واریانس متغیر تصادفی برنولی است و داریم:

$$\boxed{Var(O_{nk})=\rho(1-\rho)\sum_{i=1}^{D}w_{ki}^2x_{ni}^2}$$