۱) نشان می‌دهم که Hessian تابع یک ماتریس PSD است پس این تابع convex و دارای minimum است و با الگوریتم گرادیان کاهشی می‌توانیم به نقطه minimum آن برسیم

می‌توان ω_0 را درون ω با extend کردن x در نظر گرفت.

$$z = \sigma(h)$$

$$h = \omega^T x + \omega_0 = \omega^T x$$

$$E(\omega) = -\sum_i y_i \ln(z_i) + (1-y_i)\ln(1-z_i)$$

$$\frac{\partial E(\omega)}{\partial \omega} = -\sum \frac{\partial(---)}{\partial \omega}$$

$$\frac{\partial(---)}{\partial \omega} = \frac{\partial\left(y_i \ln(z_i) + (1-y_i)\ln(1-z_i)\right)}{\partial \omega} = y_i \frac{\partial \ln(z_i)}{\partial z_i} * \frac{\partial(z_i)}{\partial h_i} * \frac{\partial h_i}{\partial \omega} + (1-y_i)\frac{\partial \ln(1-z_i)}{\partial z_i} * \frac{\partial(z_i)}{\partial h_i} * \frac{\partial h_i}{\partial \omega}$$

$$= \frac{\partial(z_i)}{\partial h} * \frac{\partial h}{\partial \omega}\left[y_i \frac{\partial \ln(z_i)}{\partial z_i} + (1-y_i)\frac{\partial \ln(1-z_i)}{\partial z_i}\right]$$

$$\frac{\partial(z_i)}{\partial h_i} = \sigma(h_i)(\sigma(h_i)-1) = z_i(z_i-1)$$

$$\frac{\partial h_i}{\partial \omega} = x_i$$

$$\Rightarrow \frac{\partial(---)}{\partial \omega} = x_i(z_i - y_i) \Rightarrow \frac{\partial E(\omega)}{\partial \omega} = -\sum_i x_i(\hat{y}_i - y_i) = \sum_i x_i(y_i - \hat{y}_i)$$

$$= \boxed{X^T(y-\hat{y})}$$

$$\begin{cases} \dfrac{\partial E}{\partial \omega} = \sum_i x_i(y_i - \hat{y}_i) \\ \dfrac{\partial E}{\partial b} = \sum_i (y_i - \hat{y}_i) \end{cases}$$

$$H = \frac{\partial}{\partial \omega}\left[\sum_i x_i(y_i - z_i)\right] = \sum_i x_i \frac{\partial}{\partial \omega}[y_i - z_i]$$

$$\frac{\partial}{\partial \omega}[y_i - z_i] = -x_i \frac{\partial z_i}{\partial h_i} \times \frac{\partial h_i}{\partial \omega} = -x_i(z_i)(z_i-1)x_i^T = \underbrace{(z_i)(1-z_i)}_{\alpha_i} x_i x_i^T$$

$$0 < z_i = \sigma(h_i) < 1$$

$$0 < 1 - z_i < 1$$

$$\Rightarrow 0 < \alpha_i < 1$$

Positive

$$\Rightarrow H = \sum_i \alpha_i x_i x_i^T = X^T \operatorname{diag}(\alpha_i) X = \underbrace{X^T \sqrt{\operatorname{diag}(\alpha_i)}}_{A^T}\,\underbrace{\sqrt{\operatorname{diag}(\alpha_i)}\, X}_{A}$$

$H = A^T A \Rightarrow H \succ 0 \Rightarrow$ یک نقطه مینیمم دارد.

الگوریتم آموزشی $\Rightarrow \omega^{t+1} = \omega^t - \mu_t \dfrac{\partial E}{\partial \omega}$ ، $b^{t+1} = b^t - \mu_t \dfrac{\partial E}{\partial b}$

۲) (آ)

Covariate shift: مربوط به تغییر توزیع ورودی ها و node های یک شبکه به ازای نمونه های متفاوت است که می تواند باعث کاهش performance شبکه شود. در حالت کلی ممکن است پارامترهای شبکه برای داده های جدید تعمیم پذیری نداشته باشد و این باعث ایجاد مشکلاتی در عملکرد شبکه می شود. BN با نرمالیزه کردن داده های miniBatch به مقادیر مشخص این مشکل را مرتفع می نماید. در این حالت توزیع نمونه های ورودی شبکه به هم نزدیک اند و همچنین در جریان از یک لایه به لایه دیگر نیز تغییر زیادی نمی کند و در نتیجه این روش موجب کاهش اثر covariate shift و عملکرد بهتر و سریعتر مدل می شود.

ب) BN مانند اضافه کردن noise به داده های ورودی است که یک روش regularization مانند dropout است بنابراین به تعمیم پذیری مدل کمک می کند. همچنین با تغییر ورودی ها از overfit جلوگیری می کند و به جریان گرادیان نیز کمک می کند و از بزرگ شدن یا کوچک شدن آن جلوگیری می کند.

پ)

$$\frac{\partial L}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{x}_j} \times \frac{\partial \hat{x}_j}{\partial x_i}$$

$$\frac{\partial \hat{x}_j}{\partial x_i} = \begin{cases} 1 - 1/n & i = j \\ -1/n & i \neq j \end{cases} = \delta_{ij} - \frac{1}{n}$$

$$\Rightarrow \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i} - \frac{1}{n} \sum_{j=1}^{n} \frac{\partial L}{\partial \hat{x}_j}$$

$$\frac{\partial L}{\partial \hat{x}_i} = \sum_{j=1}^{n} \frac{\partial L}{\partial y_j} \cdot \frac{\partial y_j}{\partial \hat{x}_i}$$

$$\frac{\partial y_j}{\partial \hat{x}_i} = \gamma \delta_{ij}$$

$$\Rightarrow \frac{\partial L}{\partial \hat{x}_i} = \gamma \frac{\partial L}{\partial y_i}$$

$$\Longrightarrow \frac{\partial L}{\partial x_i} = \gamma \frac{\partial L}{\partial y_i} - \frac{\gamma}{n} \sum_{j=1}^{n} \frac{\partial L}{\partial y_j}$$

ت) $\frac{\partial L}{\partial x_i} = ?$

$n=1 \Rightarrow \frac{\partial L}{\partial x_i} = 0$

زمانی که یک داده داریم، تغییرات L نسبت به ورودی ها ثابت است چون $\hat{x} = 0$

$n \to \infty \Rightarrow \gamma \frac{\partial L}{\partial y_i}$

اما هرچه n بزرگتر می شود و به سمت ∞ می رود، تغییرات L نسبت به هر ورودی به صورت مستقل محاسبه می گردد و از سایر ورودی های Batch و ... اثری نمی پذیرد و در نتیجه اثر نرمالیزاسیون روی هر نقطه مستقل از سایر ورودی هاست.

۳) الف)

$$\hat{y}_k = \frac{e^{z_k^{(2)}}}{\sum_d^k e^{z_d^{(2)}}}$$

$$\frac{\partial \hat{y}_k}{\partial z_i^{(2)}} = \begin{cases} \hat{y}_k(1-\hat{y}_k) & k=i \\ \dfrac{-e^{z_k^{(2)}} e^{z_i^{(2)}}}{\underbrace{\sum_{d=1}^k e^{z_d^{(2)}}}_{\hat{y}_k} \underbrace{\sum_{d=1}^k e^{z_d}}_{\hat{y}_i}} = \hat{y}_k \hat{y}_i & k \neq i \end{cases}$$

ب)

$$\frac{\partial L}{\partial z_k^{(2)}} = \frac{\partial}{\partial z_k^{(2)}} -\sum y_i \log(\hat{y}_i) = \frac{\partial}{\partial z_k^{(2)}} -\sum_i y_i \left[z_i^{(2)} - \log \sum_d e^{z_d^{(2)}} \right]$$

$$= \frac{\partial}{\partial z_k^{(2)}} \sum_{i=1}^k y_i \log \sum_d^k e^{z_d^{(2)}} = \sum_{i=1}^k y_i \frac{e^{z_k}}{\sum_d^k e^{z_d^{(2)}}} = \sum_{i=1}^k y_i \hat{y}_k = \hat{y}_k$$

ج)

$$\frac{\partial L}{\partial w^{(1)}} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial z^{(2)}} * \frac{\partial z^{(2)}}{\partial a^{(1)}} * \frac{\partial a^{(1)}}{\partial \hat{a}^{(1)}} * \frac{\partial \hat{a}^{(1)}}{\partial z^{(1)}} * \underbrace{\frac{\partial z^{(1)}}{\partial w^{(1)}}}_{x}$$

$$\frac{\partial \hat{a}^{(1)}}{\partial z_i^{(1)}} = \begin{cases} 1 & z_i^{(1)} \geq 0 \\ 0.01 & z_i^{(1)} < 0 \end{cases}$$ vector

$$\frac{\partial a^{(1)}}{\partial \hat{a}^{(1)}} = \begin{cases} 1 & \text{with Probability of } 0.8 \\ 0 & \text{with Probability of } 0.2 \end{cases}$$ vector

$$\frac{\partial z^{(2)}}{\partial a^{(1)}} = w^{(2)}$$

$$\frac{\partial L}{\partial z^{(2)}} = \hat{y}_k$$

با توجه به قسمت ۲

Jacobian of f $R^m \longrightarrow R^n$ is $J \in R^{n \times m}$

$$H(y) = \begin{pmatrix} \frac{\partial^2 y}{\partial u^2} & \frac{\partial^2 y}{\partial u \partial v} & \frac{\partial^2 y}{\partial u \partial z} \\ \frac{\partial^2 y}{\partial u \partial v} & \frac{\partial^2 y}{\partial v^2} & \frac{\partial^2 u}{\partial z \partial v} \\ \frac{\partial^2 y}{\partial u \partial z} & \frac{\partial^2 y}{\partial v \partial z} & \frac{\partial^2 u}{\partial z^2} \end{pmatrix}$$

$$\nabla(y) = \begin{pmatrix} \frac{\partial y}{\partial u} \\ \frac{\partial y}{\partial v} \\ \frac{\partial y}{\partial z} \end{pmatrix} \qquad F = \begin{pmatrix} f_1(u,v,z) \\ f_2(u,v,z) \\ f_3(u,v,z) \end{pmatrix}$$

$$J(F) = \begin{pmatrix} \frac{\partial f_1}{\partial u} & \frac{\partial f_1}{\partial v} & \frac{\partial f_1}{\partial z} \\ \frac{\partial f_2}{\partial u} & \frac{\partial f_2}{\partial v} & \frac{\partial f_2}{\partial z} \\ \frac{\partial f_3}{\partial u} & \frac{\partial f_3}{\partial v} & \frac{\partial f_3}{\partial z} \end{pmatrix}$$

$$\boxed{J(\nabla_y) = H(y)}$$

۵)

$$J_1 = 0.5\left(y_d - \sum \delta_k W_k x_k\right)^2$$

$$\frac{\partial J_1}{\partial w_i} = -\left(y_d - \sum_{k=1}^{n} \delta_k w_k x_k\right)\delta_i x_i = \sum_{k=1}^{n} \delta_i \delta_k x_i x_k w_k - y_d \delta_i x_i$$

$$E\left[\frac{\partial J_1}{\partial w_i}\right] = E\left[\sum_{\substack{k=1\\k\neq i}}^{n} \delta_i \delta_k x_i x_k w_k\right] + E\left[\delta_i^2 x_i^2 w_i\right] - E\left[y_d \delta_i x_i\right]$$

$$E[\delta_i \delta_k] = E[\delta_i]E[\delta_k] = 1 \qquad (1+\sigma^2)\, x_i^2 w_i \qquad -x_i y_d$$

$$\sum_{\substack{k=1\\k\neq i}}^{n} x_i w_k x_k$$

$$= \sum_{\substack{k=1\\k\neq i}}^{n} x_i w_k x_k + (1+\sigma^2) w_i x_i^2 - x_i y_d = x_i\left[\sum_{k=1}^{n} w_k x_k + \sigma^2 w_i - y_d\right]$$

$$= -x_i\left[y_d - \sum_{k=1}^{n} w_k x_k\right] + w_i x_i^2 \sigma_i^2$$

non-regularized: $J_c = 0.5\left(y_d - \sum w_k x_k\right)^2$

$$\boxed{\frac{\partial J_c}{\partial w_i} = -x_i\left[y_d - \sum_{k=1}^{n} w_k x_k\right]}$$

$$\Rightarrow E\left[\frac{\partial J_1}{\partial w_i}\right] = \frac{\partial J_c}{\partial w_i} + w_i x_i^2 \, Var(\sigma_i)$$

مشاهده می‌شود که ضرایب w_k با یک متغیر تصادفی δ_k نویزی می‌شوند و در ادامه تابع loss نسبت به ضرایب مانند regularization عمل می‌کند و مقداری نویز به ضرایب اضافه می‌کند

ج ۶) اثبات همگرایی روش نیوتن:

$f = g(x)$

$f(x^*) = 0$ ، $f'(x^*) \neq 0$ ، $f''(x)$ پیوسته

* فرض می‌شود که x_k به x^* همگرا می‌شود:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \;;\; k = 1, 2, \dots$$

$$|x_{k+1} - x^*| \leq M\,|x_k - x^*|^2 \quad \text{if} \quad M > \frac{|f''(x^*)|}{|f'(x^*)|}$$

اثبات: $e_k = x_k - x^* \Rightarrow x^* = x_k - e_k$

بسط تیلور: $f(x) = f(x_k) + (x - x_k)f'(x_k) + R_1$

$$R_1 = f''(\xi)\frac{(x - x_k)^2}{2!} \,,\; x_k < \xi < x^*$$

$$\Rightarrow \underbrace{f(x_k - e_k)}_{f(x^*) = 0} = f(x_k) - e_k f'(x_k) + \frac{(e_k)^2}{2!} f''(\xi)$$

$$\Rightarrow \quad 0 = f(x_k) - e_k f'(x_k) + \frac{e_k^2}{2!} f''(\xi)$$

تا زمانی که x_k به اندازه‌ی کافی به x^* نزدیک باشد، چون $0 \neq f'(x^*)$ پس $f'(x_k) \neq 0$

$$\Rightarrow \quad 0 = \frac{f(x_k)}{f'(x_k)} - \underbrace{e_k}_{(x_k - x^*)} + \frac{e_k^2}{2!}\,\frac{f''(\xi)}{f'(x_k)}$$

$$\Rightarrow \quad 0 = \frac{f(x_k)}{f'(x_k)} - (x_k - x^*) + \frac{e_k^2 f''(\xi)}{2f'(x_k)}$$

$-x_{k+1}$

$$\Rightarrow \quad x_{k+1} - x^* = \frac{\underbrace{e_k}_{(x_k - x^*)}{}^2\, f''(\xi)}{2f'(x_k)}$$

$$\Rightarrow \quad x_{k+1} - x^* = (x_k - x^*)^2 \frac{f''(\xi)}{2f'(x_k)}$$

$$\Rightarrow \quad |x_{k+1} - x^*| \leq \frac{|f''(\xi)|}{2|f'(x_k)|} |x_k - x^*|^2$$

با توجه به پیوستگی f' و f'' ، $f'(x_k)$ به $f'(x^*)$ همگرا می‌شود و ξ که بین x_k و x^* است نیز به x^* همگرا می‌شود و $f''(\xi)$ به $f''(x^*)$ همگرا می‌شود

$$\Rightarrow \quad |x_{k+1} - x^*| \leq M |x_k - x^*|^2 \quad , \quad M > \frac{|f''(x^*)|}{2|f'(x^*)|}$$

x^* نقطه بهینه $g(x)$ است چرا که $g'(x^*) = 0$

با توجه به نتیجه قبل:

$$|x_{k+1} - x^*| \leq M |x_k - x^*|^2 \quad , \quad M > \frac{|g^{(3)}(x^*)|}{2|g''(x^*)|}$$

۷) آ)

$L(z,y) = -\sum_{i=1}^{k} y_i \ln \hat{y}_i$

$$\frac{\partial L}{\partial z_j} = \frac{\partial}{\partial z_j}\left[-\sum_{i=1}^{k} y_i\left(z_i - \ln\left(\sum_{d=1}^{k} e^{z_d}\right)\right)\right] = \frac{\partial}{\partial z_j}\left[-\sum_{i=1}^{k} y_i z_i\right] + \frac{\partial}{\partial z_j}\left[\sum y_i \ln\left(\sum_{d=1}^{k} e^{z_d}\right)\right]$$

$$-\frac{\partial}{\partial z_j}\sum_{i=1}^{k} y_i z_i = -y_j \quad ، \quad \frac{\partial}{\partial z_j}\left[\sum_{i=1}^{k} y_i \ln\left(\sum_{d=1}^{k} e^{z_d}\right)\right] = \sum_{i=1}^{k} y_i \underbrace{\frac{e^{z_j}}{\sum_{d=1}^{k} e^{z_d}}}_{\hat{y}_j} = \sum_{i=1}^{k} y_i \hat{y}_j$$

$$\Rightarrow \frac{\partial L}{\partial z_j} = \underbrace{\cancel{\sum_{i=1}^{k} y_i \hat{y}_j}}_{\hat{y}_j} - y_j = \hat{y}_j - y_j \Rightarrow \boxed{\nabla_z L = \hat{Y} - Y}$$

چون y به صورت one hot است.

ب)

$$H = \nabla_z^2 L = \begin{bmatrix} \frac{\partial^2 L}{\partial z_1 \partial z_1} & \cdots & \frac{\partial^2 L}{\partial z_1 \partial z_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 L}{\partial z_1 \partial z_k} & \cdots & \frac{\partial^2 L}{\partial z_k \partial z_k} \end{bmatrix} , \quad \frac{\partial^2 L}{\partial z_i \partial z_j} = \frac{\partial}{\partial z_j}(\hat{y}_i - y_i) = \begin{cases} \hat{y}_j - \hat{y}_j^2 & i=j \\ -\hat{y}_i \hat{y}_j & i \neq j \end{cases}$$

$$\Rightarrow \boxed{H = diag(\hat{y}_i) - \hat{y}\hat{y}^T} \quad ، \quad x^T H x = x^T diag(\hat{y}_i) x - x^T \hat{y}\hat{y}^T x$$

ب)

$$= \sum_{i=1}^{k} \hat{y}_i x_i^2 - \left(\sum x_i \hat{y}_i\right)^2$$

با توجه به اینکه $\hat{y}_i$ احتمال است و یک عدد مثبت است من به صورت $(\sqrt{\hat{y}_i})^2$ در نظر می‌گیرم و همچنین $\sum_{i=1}^{k} \hat{y}_i = 1$ چون احتمال است:

$\Rightarrow$ طبق کوشی شوارتز داریم

$$\left(\sum x_i \sqrt{\hat{y}_i}\sqrt{\hat{y}_i}\right)^2 \leq \sum \left(x_i \sqrt{\hat{y}_i}\right)^2 \sum \left(\sqrt{\hat{y}_i}\right)^2 = \sum x_i^2 \hat{y}_i \underbrace{\sum \hat{y}_i}_{=1}$$

چون احتمال است

$$\Rightarrow \boxed{\sum x_i^2 \hat{y}_i \geq \left(\sum x_i \hat{y}_i\right)^2}$$

$$\Rightarrow x^T H x \geq 0 \quad \forall x \in R^k \Rightarrow H \succeq 0$$ H is PSD $\Rightarrow$ مثبت نیمه معین است

تابع زیان cross entropy نسبت به $\underline{Z}$ یک تابع محدب است.