Page 1

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N} y_n \lg\left[f(B^T A X_n)\right]$$

$$S = \sum_{k=1}^{K} (B^T A X_n)_k \quad \text{--- this is a scalar}$$

$y_n : 1 \times K$

$B : \dim \times K$

$A : \dim \times N_w$

$x_n : N_w \times 1$

First, write.

B_k is kth column of B.

$$\mathcal{L}_n = y_n \cdot \lg\left[f(B^T A X_n)\right] = \sum_{k=1}^{K} y_{nk} \cdot \underline{\lg\left[f(B_k^T A X_n)\right]}$$

Then take derivative.

$$\frac{\partial}{\partial B_k}\mathcal{L}_n = \sum_{k=1}^{K} y_{nk} \cdot \left(\lg\left[f(B_k^T A X_n)\right]\right)'$$

$$= \sum_{k=1}^{K} y_{nk} \cdot \frac{1}{f(B_k^T A X_n)} \cdot \left[f(B_k^T A X_n)\right]'$$

$$= \sum_{k=1}^{K} y_{nk} \cdot \frac{1}{f(B_k^T A X_n)} \cdot \frac{(B_k^T A X_n)' \cdot S - (B_k^T A X_n) \cdot S'}{S^2}$$

combine.

$$= \sum_{k=1}^{K} y_{nk} \cdot \frac{1}{f(B_k^T A X_n)} \cdot \frac{(A X_n) \cdot S - (B_k^T A X_n) \cdot (A X_n)}{S^2}$$

dim×1.

$$= \sum_{k=1}^{K} y_{nk} \cdot \left[f(B_k^T A X_n)\right]^{-1} \cdot \underline{\frac{1}{S^2} \cdot (A X_n)} \cdot \left(\sum_{\substack{k' \neq k \\ k'=1}}^{K} B_{k'}^T A X_n\right)$$

$$= \left[\sum_{k=1}^{K} y_{nk} \cdot \frac{\sum_{k=1, k' \neq k}^{K} B_{k'}^T A X_n}{f(B_k^T A X_n)}\right] \cdot \underbrace{\frac{1}{S^2} \cdot (A X_n)}_{\text{independent of } k}$$

⋮

$$f(B_k^T A X_n) = \frac{B_k^T A X_n}{S}$$

see left above.

this $k' \neq k$

$$S' = \left[\sum_{k'=1}^{K} (B^T A X_n)_{k'}\right]' = \sum_{k'=1}^{K} (B_{k'}^T A X_n)' = A X_n$$

(because only the term with B_k survives)

(cont)

~~$\sum y_{nk}$~~ **trick 1**: $\sum_{\substack{k'=1\\k'\neq k}}^{K} B_{k'}^T A X_n = \overset{\overset{K\times 1}{\cancel{\dim\times}}}{(B^T A\cdot X_n)^T} \cdot \mathbb{1}_k$.

where $\mathbb{1}_k = [1, 1, \dots, 1, \underset{\substack{\uparrow\\ \text{kth element}}}{0}, 1, \dots]$

trick2 $f(B_k^T A X_n) = \frac{1}{S} \cdot B_k^T A X_n$.

Plug both tricks into original formula, we have

$$= \sum_{k=1}^{K} y_{nk} \cdot \frac{(B^T A X_n)^T \cdot \mathbb{1}_k}{B_k^T A X_n \cdot \cancel{\frac{1}{S}}} \cdot \frac{1}{S^2} \cdot (A X_n)$$

cancel one.

$$= \cancel{\sum y_{nk} \cdots \frac{}{X_n^T A^T B_k}}$$

$$= \underline{\sum_{k=1}^{K} y_{nk} \cdot \frac{\mathbb{1}_k^T (B^T A X_n)}{B_k^T A X_n}} \cdot \frac{1}{S} \cdot (A X_n)$$

$$= [y_{n1}\mathbb{1}_1^T (B^T A X_n), \dots, y_{nK}\mathbb{1}_K^T (B^T A X_n)] \begin{bmatrix} \frac{1}{B_1^T A X_n} \\ \vdots \\ \frac{1}{B_K^T A X_n} \end{bmatrix} \cdot \frac{1}{S} \cdot (A X_n)$$

$$\underbrace{\frac{\partial}{\partial B_k} L_n}_{\uparrow} = \underbrace{[y_{n1}\mathbb{1}_1^T, \dots, y_{nk}\mathbb{1}_k^T]}_{\substack{\uparrow\\ \text{define a matrix}}} \cdot \underbrace{(B^T A X_n)}_{\substack{\uparrow\\ \text{easy to}\\ \text{compute}}} \cdot \underbrace{\begin{bmatrix} \frac{1}{B_1^T A X_n} \\ \vdots \\ \frac{1}{B_k^T A X_n} \end{bmatrix}}_{\substack{\uparrow\\ \text{nothing but}\\ B^T A X_n\\ \text{then take}\\ \text{row-wise inverse}}} \cdot \underbrace{\frac{1}{S}}_{\substack{\uparrow\\ \text{easy}}} \cdot \underbrace{(A X_n)}_{\substack{\uparrow\\ \dim\times 1}}$$

this is gradient at a ~~single~~ single instance (can be used in stochastic gradient descent).