Mathematical Foundation of DNN : HW 8

Jeong Min Lee

May 3, 2024

1

For $1 \leq i \leq m/2, 1 \leq j \leq n/2$

$$[\mathcal{T}(X)]_{ij} = \frac{1}{4}\left(X_{2i-1,2j-1} + X_{2i-1,2j} + X_{2i,2j-1} + X_{2i,2j}\right) \equiv \frac{1}{4}\text{sum}\left(X_{2i-1:2i,2j-1:2j}\right)$$

From the equation above, one can observe that $\mathcal{T}$ is linear operator.

$$\begin{aligned}
[\mathcal{T}(X+Y)]_{ij} &= \frac{1}{4}\text{sum}\left((X+Y)_{2i-1:2i,2j-1:2j}\right) \\
&= \frac{1}{4}\text{sum}\left(X_{2i-1:2i,2j-1:2j} + Y_{2i-1:2i,2j-1:2j}\right) \\
&= \frac{1}{4}\left(\text{sum}(X_{2i-1:2i,2j-1:2j}) + \text{sum}(Y_{2i-1:2i,2j-1:2j})\right) \\
&= [\mathcal{T}(X)]_{ij} + [\mathcal{T}(Y)]_{ij}
\end{aligned}$$

$$\begin{aligned}
\mathcal{T}(aX) &= \frac{1}{4}\text{sum}\left((aX)_{2i-1:2i,2j-1:2j}\right) \\
&= a \times \frac{1}{4}\text{sum}\left((X)_{2i-1:2i,2j-1:2j}\right) \\
&= a\mathcal{T}(X)
\end{aligned}$$

$$\begin{aligned}
&\sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}(\mathcal{T}(\mathcal{X}))_{ij} \\
&= \sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}\frac{1}{4}\text{sum}\left(X_{2i-1:2i,2j-1:2j}\right) \\
&= \frac{1}{4}\left[\sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}X_{2i-1,2j-1} + \sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}X_{2i-1,2j} + \sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}X_{2i,2j-1} + \sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}X_{2i,2j}\right] \\
&= \frac{1}{4}\left[\sum_{\substack{i=1,3,\cdots,m-1 \\ j=1,3,\cdots,n-1}} Y_{\frac{i+1}{2},\frac{j+1}{2}}X_{ij} + \sum_{\substack{i=1,3,\cdots,m-1 \\ j=2,4,\cdots,n}} Y_{\frac{i+1}{2},\frac{j}{2}}X_{ij} + \sum_{\substack{i=2,4,\cdots,m \\ j=1,3,\cdots,n-1}} Y_{\frac{i}{2},\frac{j+1}{2}}X_{ij} + \sum_{\substack{i=2,4,\cdots,m \\ j=2,4,\cdots,n}} Y_{\frac{i}{2},\frac{j}{2}}X_{i,j}\right] \\
&= \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1}^{n} Y_{\lfloor\frac{i+1}{2}\rfloor,\lfloor\frac{j+1}{2}\rfloor}X_{ij}
\end{aligned}$$

Consider the upsampling operator $\mathcal{U} : \mathbb{R}^{m/2\times n/2} \to \mathbb{R}^{m\times n}$. For $1 \leq i \leq m, 1 \leq j \leq n$,

$$[\mathcal{U}(Y)]_{ij} = \begin{cases} Y_{\frac{i+1}{2},\frac{j+1}{2}} \text{ if } i = odd, j = odd \\ Y_{\frac{i+1}{2},\frac{j}{2}} \text{ if } i = odd, j = even \\ Y_{\frac{i}{2},\frac{j+1}{2}} \text{ if } i = even, j = odd \\ Y_{\frac{i}{2},\frac{j}{2}} \text{ if } i = even, j = even \end{cases}$$

The mapping above can be written in short as follows: $[\mathcal{U}(Y)]_{ij} = Y_{\lfloor\frac{i+1}{2}\rfloor,\lfloor\frac{j+1}{2}\rfloor}$. Thus,

$$\sum_{i=1}^{m/2}\sum_{j=1}^{n/2} Y_{ij}(\mathcal{T}(\mathcal{X}))_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathcal{T}^T(Y)\right)_{ij}X_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{4}\left(\mathcal{U}(Y)\right)_{ij}X_{ij}$$

2

Suppose input $X \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$ and the output $Y \in \mathbb{R}^{C_{out} \times H_{out} \times W_{out}}$. Noting the **ConvTranspose2d()** operation, $H_{out} = (H_{in} - 1) \times \text{striding} + k$ without any consideration of padding. To make **ConvTranspose2d()** operation and **Upsample()** operation are equivalent, the filter size k and striding should satisfy the following equation.

$$H_{out} = (H_{in} - 1) \times \text{striding} + k = rH_{in}$$

Selecting $k = r$, striding $= r$ is the simplest way to hold the condition above. Also, to do nearest upsampling, the weight of the filter should be initialized to be 1 with dimension $r \times r$. Thus,

```
layer = nn.ConvTranspose2d(C_in,C_out, kernel_size = r, stride = r)
layer.weight.data = torch.ones(r,r)
```

3

a

$$\begin{aligned}
D_f(X||Y) &= \int f\left(\frac{P_X(x)}{P_Y(x)}\right) P_Y(x)dx \\
&= \mathbb{E}\left[f\left(\frac{P_X}{P_Y}\right)\right] \\
&\geq f\left(\mathbb{E}\left[\frac{P_X}{P_Y}\right]\right) \\
&= f\left(\int \frac{P_X(x)}{P_Y(x)} P_Y(x)dx\right) \\
&= f(1) = 0
\end{aligned}$$

b

To show that $y = -\log x$ and $y = x\log x$ are convex, I used the following theorem : The differentiable function f is strictly convex *if and only if* its derivate f' is strictly increasing (Refer to the appendix to see the proof of the following theorem). Noting that logarithm function is defined on $\{x \in \mathbb{R} | x > 0\}$,

$$\begin{aligned}
\frac{d^2}{dx^2}(-\log x) &= \frac{d}{dx}\left(-\frac{1}{x}\right) = \frac{1}{x^2} \geq 0 \\
\frac{d^2}{dx^2}(x\log x) &= \frac{d}{dx}\log x + 1 = \frac{1}{x} \geq 0
\end{aligned}$$

Also, it is obvious that $\log 1 = 0, 1 \times \log 1 = 0$. Thus, the f-divergence of both $-\log x, x\log x$ are well defined. Then, one can find that f-divergence of both functions are KL-divergence.

$$\begin{aligned}
D_f(X||Y) &= \int -\log\left(\frac{P_X(x)}{P_Y(x)}\right) P_Y(x)dx = \int P_Y(x)\log\left(\frac{P_Y(x)}{P_X(x)}\right) dx = D_{KL}\left(P_Y||P_X\right) \\
D_f(X||Y) &= \int P_Y(x)\frac{P_X(x)}{P_Y(x)}\log\left(\frac{P_X(x)}{P_Y(x)}\right) dx = \int P_X(x)\log\left(\frac{P_X(x)}{P_Y(x)}\right) dx = D_{KL}(P_X||P_Y)
\end{aligned}$$

4

$$\mathbb{P}(G(U) \leq t) = \mathbb{P}\left(\inf\{x \in \mathbb{R} | U \leq F(x)\} \leq t\right)$$

Define two event A, B as follows :

$$\begin{aligned}
A &= \{\inf\{x \in \mathbb{R} | U \leq F(x)\} \leq t\} \\
B &= \{U \leq F(t)\}
\end{aligned}$$

Since the CDF of U, $F_U(u)$, is $F_U(u) = u$ for $u \in [0, 1]$, $\mathbb{P}(B) = \mathbb{P}(U \leq F(t)) = F_U(F(t)) = F(t)$. If $A = B$, then $\mathbb{P}(A) = \mathbb{P}(B) = F(x)$ and the proof is done. To show that $A = B$, consider the both cases : $B \subset A, A \subset B$.

(i) : $A \subset B$

Suppose A is true. Then, $\inf\{x \in \mathbb{R} | U \leq F(x)\} \leq t$. This implies $\exists x_0 \in \mathbb{R}, F(x_0) \geq U$ for $x_0 \leq t$. Since the CDF F is nondecreasing, $F(t) \geq F(x_0)$. Thus, $U \leq F(x_0) \leq F(t)$, which concludes that B is true.

(ii) : $B \subset A$

Suppose B is true. Then, $U \leq F(t)$. Consider a set $S = \{x | F(x) \geq U\}$. Then, $t \in S$. This implies $\inf S = \inf\{x | F(x) \geq U\} \leq t$. Thus, the event A is true. As a result, $A = B$ and this ends up with $\mathbb{P}(G(U) \leq t) = F(t)$

5

From $X = AY + b$, $\varphi(X) = Y = A^{-1}(X - b)$. According to the prperty of $p_X(x)$ produced by the condition of the problem,

$$\begin{aligned}
P_X(x) &= P_Y(\varphi(x)) \left| \det \frac{\partial \varphi}{\partial x}(x) \right| \\
&= P_Y(\varphi(x)) \left| \det A^{-1} \right| \\
&= \frac{1}{\det A} \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\|A^{-1}(x-b)\|^2} \\
&= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-b)^T (A^{-1})^T A^{-1}(x-b)} (\because \det \Sigma = \det A \det A^T = (\det A)^2) \\
&= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(x-b)^T \Sigma^{-1}(x-b)} (\because \Sigma^{-1} = (AA^T)^{-1} = (A^{-1})^T A^{-1})
\end{aligned}$$

6

Algorithm 1 Inverse Permutation Pseudo Code

Let n be the length of the list σ
Initialize an empty list called *result*
for $i = 1$ **to** n **do**
    for $j = 1$ **to** n **do**
        if $\sigma[i]$ equals $j + 1$ **then**
            Append $j + 1$ to the *result* list
        end if
    end for
end for

7

a

$$[P_\sigma x]_i = e_{\sigma(i)}^T x = x_{\sigma(i)}$$

b

$$\begin{aligned}
\left(P_\sigma P_\sigma^T\right)_{ij} &= \sum_k P_\sigma^{(ik)} P_\sigma^{(jk)} \\
&= \sum_k e_{\sigma(i)}^{(k)} e_{\sigma(j)}^{(k)} \\
&= \sum_k \delta_{\sigma(i),k} \delta_{\sigma(j),k} \\
&= \delta_{\sigma(i),\sigma(j)} \\
&= \delta_{ij} \ (\because \sigma \text{ is bijection})
\end{aligned}$$

$$
\begin{aligned}
(P_{\sigma^{-1}}P_{\sigma})_{ij} &= \sum_k P_{\sigma^{-1}}^{(ik)} P_{\sigma}^{(kj)} \\
&= \sum_k e_{\sigma^{-1}(i)}^{(k)} e_{\sigma(k)}^{(j)} \\
&= \sum_k \delta_{\sigma^{-1}(i),k}\delta_{\sigma(k),j} \\
&= \delta_{ij} \ (\because \delta_{\sigma^{-1}(i),k}\delta_{\sigma(k),j} = 1 \text{ if } k = \sigma^{-1}(i), \sigma(k) = j)
\end{aligned}
$$

c

$$
\begin{aligned}
&\det\left(P_\sigma P_\sigma^T\right) = (\det P_\sigma)^2 = 1 \\
&\therefore |\det P_\sigma| = 1
\end{aligned}
$$

A Proof of Theorem in problem 3-b

proof

$\Longrightarrow$: $a, b \in \mathbb{R}(a < b)$. Let$x_1, x_2, x_3 \in \mathbb{R}$ be chosen s.t. $a < x_1 < x_2 < x_3 < b$.

From Chordal Slope Lemma, which is famous lemma applicable to convex function, the following inequalities are held.

$$
\frac{f(x_1)-f(a)}{x_1-a} < \frac{f(x_2)-f(x_1)}{x_2-x_1} < \frac{f(x_3)-f(x_2)}{x_3-x_2} < \frac{f(b)-f(x_3)}{b-x_3} \tag{1}
$$

$$
\frac{f(x_2)-f(a)}{x_2-a} < \frac{f(b)-f(x_2)}{b-x_2} \tag{2}
$$

Taking limit $x_1 \to a, x_3 \to b$ on equation 1,

$$
f'(a) \leq \frac{f(x_2)-f(a)}{x_2-a} < \frac{f(b)-f(x_2)}{b-x_2} \leq f'(b) \tag{3}
$$

and, therefore, $f'(a) < f'(b)$ for arbitrary $a, b \in \mathbb{R}$.

$\Longleftarrow$: Let $x_1, x_2, x_3 \in \mathbb{R}$ s.t. $x_1 < x_2 < x_3$. From Mean Value Theorem,

$$
\exists a \text{ s.t } \frac{f(x_2)-f(x_1)}{x_2-x_1} = f'(a) \tag{4}
$$

$$
\exists b \text{ s.t } \frac{f(x_3)-f(x_2)}{x_3-x_2} = f'(b) \tag{5}
$$

Note that $x_1 < a < x_2 < b < x_3$. Since f' is strictly increasing, $f'(a) < f'(b)$, which implies

$$
\frac{f(x_2)-f(x_1)}{x_2-x_1} < \frac{f(x_3)-f(x_2)}{x_3-x_2} \tag{6}
$$

The result of equation 6 is equivalent that f is strictly convex according to Chordal Slope Lemma.$\blacksquare$