MFDNN

Homework 8

24/05/01

1. $$\cancel{\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} Y_{ij}(J(X))_{ij} = \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} Y_{ij}\big((A(X.\text{reshape}(mn))).\text{reshape}(n/2,n/2)\big)_{ij}}$$

$$\cancel{= \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} Y_{ij}\Big(\sum_{k=1}^{mn}(A)_{\frac{n}{2}(i-1)+j,k}\,(X)_{\lfloor k/n\rfloor+1,\,k \bmod m}\Big)}$$

$$\cancel{(J^T(Y))_{ij} = \sum_{k=1}^{mn/4}(A^T)_{n(i-1)+j,k}\,(Y)_{\lfloor \frac{2k}{n}\rfloor+1,\,k-\frac{n}{2}\cdot\lfloor\frac{2k}{n}\rfloor}}$$

mean (2x2) stride 2

$$A = \begin{pmatrix} \frac14 & \frac14 & \frac14 & \frac14 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac14 & \frac14 & \frac14 & \frac14 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 & 0 & 0 & 0 & \frac14 & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \frac14 & \frac14 & \frac14 & \frac14 \end{pmatrix} \in \mathbb{R}^{(mn/4)\times(mn)}$$

the constant is $\frac14$ clearly

$$A^T = \begin{pmatrix} \frac14 & 0 & \cdots & 0 \\ \frac14 & 0 & \cdots & 0 \\ \frac14 & 0 & \cdots & 0 \\ \frac14 & 0 & \cdots & 0 \\ 0 & \frac14 & \cdots & 0 \\ 0 & \frac14 & \cdots & 0 \\ 0 & \frac14 & \cdots & 0 \\ 0 & \frac14 & \cdots & 0 \\ 0 & 0 & \frac14 & 0 \\ \vdots & \vdots & & \frac14 \\ & & \ddots & \frac14 \\ & & & \frac14 \\ 0 & 0 & \cdots & \frac14 \end{pmatrix} \in \mathbb{R}^{(mn)\times(mn/4)}$$

- J is a linear operator since $J(\alpha X) = \alpha J(X)$ since the average is also homogeneous $\dfrac{\sum_{i=1}^4 \alpha x_i}{4} = \alpha\dfrac{\sum_{i=1}^4 x_i}{4}$

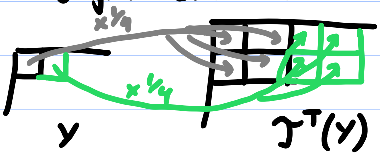
and $J(X_1+X_2) = J(X_1)+J(X_2)$ since,

$$\frac{(x_{11}+x_{21})+(x_{12}+x_{22})+(x_{13}+x_{23})+(x_{14}+x_{24})}{4} = \frac{x_{11}+x_{12}+x_{13}+x_{14}}{4} + \frac{x_{21}+x_{22}+x_{23}+x_{24}}{4}$$

So can be expressed as matrix A (see above)

- First four pixels of output of $J^T(Y)$ are all $\frac14$ of the first input pixel of Y.
  The next four " " second "
  ⋮

So every four pixels/elements in $J^T(Y)$ originate from one input pixel of Y and all have value $\frac14$ of that origin pixel's origin. i.e. can be seen as $\frac14\times$ nearest neighbour upsample



number of input channels

2. `layer = nn.ConvTranspose2d(in_channels=Cin, out_chanls=Cin, kernel_size=r, stride=r, bias=False)`
`layer.weight.data = torch.eye(Cin).unsqueeze(-1).unsqueeze(-1) * torch.ones(1,1,r,r)`

$C_{in}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ C_{in} broadcast

we don't want all 1s, only an $r \times r$ kernel of 1s once per channel; each in channel should only directly influence itself in the 'out channels' i.e. 1-1 so other channels' effect should be 0 (hence eye and not all ones)

is equivalent to:

```
... data.fill_(0)
for i in range(Cin):
    for h in range(r):
        for w in range(r):
            layer.weight.data[i, i, h, w] = 1
```

The kernel of ones for each channel copies each value in the input into a $k \times k$ grid which is the same as nearest neighbour upsampling.

3.(a) $p_x(x), p_y(x) \geq 0 \ \forall x$ since they are probability densities.

$$f \text{ convex} \Rightarrow f(\eta x_1 + (1-\eta)x_2) \leq \eta f(x_1) + (1-\eta) f(x_2) \quad \forall x_1, x_2 \in \mathbb{R}^+, \eta \in (0,1)$$

Let $z \geq 0$ and set $x_1 = x_2 = z$.

$$f(\eta z + (1-\eta)z) \leq \eta f(z) + (1-\eta) f(z)$$
$$f(1) \leq f(z)$$

Using given, $0 \leq f(z)$. Hence $f(z) \geq 0$ for any non-negative z.

Where-ever it is defined (i.e. where $p_y(x) \neq 0$), $\frac{p_x(x)}{p_y(x)} \geq 0$ hence $f\left(\frac{p_x(x)}{p_y(x)}\right) \geq 0$ by above.

If, for some x, $p_y(x) = 0$, then $f\left(\frac{p_x(x)}{p_y(x)}\right) p_y(x) = 0$.
Otherwise, if $p_y(x) \neq 0$, then $f\left(\frac{p_x(x)}{p_y(x)}\right) p_y(x) \geq 0$ since both terms of product are non-negative.

Hence the integrand is non-negative everywhere so the integral, $\int f\left(\frac{p_x(x)}{p_y(x)}\right) p_y(x)\,dx$, must always be non-negative also. That is $D_f(X \| Y) \geq 0$ as required.

3.(b) NB: $-\log$ is convex (HW2.4), as is $f(t) = t\log t$ (since $f''(t) = \frac{1}{t} > 0 \ \forall t \in \mathbb{R}$ ∵ in this case t is a division of two positive pdfs)

- $f(t) = -\log t$

$$D_f(X\|Y) = \int f\left(\frac{p_X(x)}{p_Y(x)}\right) p_Y(x)\,dx = \int -\log\left(\frac{p_X(x)}{p_Y(x)}\right) p_Y(x)\,dx$$

$$= \int \log\left(\frac{p_Y(x)}{p_X(x)}\right) p_Y(x)\,dx \quad (= D_{KL}(Y\|X))$$

$$\left(\begin{aligned} &= \underbrace{\int \log(p_Y(x))\,p_Y(x)\,dx}_{\substack{\text{Entropy of } Y: \\ -H(Y)}} - \underbrace{\int \log(p_X(x))\,p_Y(x)\,dx}_{\substack{\text{Cross-entropy of } X \text{ relative to } Y: \\ H(Y,X)}} \\ D_{KL}(X\|Y) = \int \log\left(\frac{p_X(x)}{p_Y(x)}\right) p_X(x) &= \underbrace{\int \log(p_X(x))\,p_X(x)\,dx}_{\substack{\text{Entropy of } X: \\ -H(X)}} - \underbrace{\int \log(p_Y(x))\,p_X(x)\,dx}_{\substack{\text{cross-entropy of } Y \text{ relative to } X: \\ H(X,Y)}} \end{aligned} \right)$$

- $f(t) = t\log t$

$$D_f(X\|Y) = \int \frac{p_X(x)}{p_Y(x)} \log\left(\frac{p_X(x)}{p_Y(x)}\right) p_Y(x)\,dx = \int p_X(x) \log\left(\frac{p_X(x)}{p_Y(x)}\right) dx = D_{KL}(X\|Y) \ \blacksquare$$

4. $G(u) = \inf\{x \in \mathbb{R} \mid u \leq F(x)\}$.

Considering $\mathbb{P}(G(U) \leq t)$, for any u s.t. $u \leq F(t)$, $G(u)$ is the smallest x s.t. $u \leq F(x)$.
So $x \leq t$, since for any $h > 0$, $F(x+h) > F(x)$ by right continuity of F (i.e. $t = x+h$).
This allows us to say $G(u) \leq t$

$$\Rightarrow u \leq F(\underbrace{G(u)}_{\text{'}x\text{'}}) \leq F(\overset{\text{'}x+h\text{'}}{t})$$

Therefore $\mathbb{P}(G(U) \leq t) = \mathbb{P}(U \leq F(t)) = F(t)$.
So $G(U)$ is an RV with CDF F as required.

5. $\underline{X} = A\underline{Y} + \underline{b}$

$A^{-1}(\underline{X}-\underline{b}) = \underline{Y}$ (A invertible)

i.e. $\varphi(\underline{z}) = A^{-1}(\underline{z}-\underline{b})$

$$\left(\frac{\partial \varphi(\underline{z})}{\partial \underline{z}}\right)_{ij} = \frac{\partial \varphi_i(\underline{z})}{\partial z_j} = \frac{\partial}{\partial z_j}\left(\sum_k^n (A^{-1})_{ik}(z_k - b_k)\right)$$

$$= (A^{-1})_{ij} \quad \Rightarrow \frac{\partial \varphi(\underline{z})}{\partial \underline{z}} = A^{-1}$$

$(\Sigma := AA^T)$

$$\left|\det\left(\frac{\partial \varphi(\underline{z})}{\partial \underline{z}}\right)\right| = |\det(A^{-1})| = \left|\frac{1}{\det(A)}\right| = \frac{1}{\sqrt{\det(A)^2}} = \frac{1}{\sqrt{\det(A)\det(A^T)}} = \frac{1}{\sqrt{\det(\Sigma)}}$$

by properties of det

$$p_X(\underline{x}) = p_Y(\varphi(\underline{x}))\left|\det \frac{\partial \varphi(\underline{x})}{\partial \underline{x}}\right|$$

$$= \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\|\varphi(\underline{x})\|^2} \cdot \frac{1}{\sqrt{\det \Sigma}}$$

$$\|\varphi(\underline{x})\|^2 = (A^{-1}(\underline{x}-\underline{b}))^T(A^{-1}(\underline{x}-\underline{b}))$$
$$= (\underline{x}-\underline{b})^T(A^{-1})^T A^{-1}(\underline{x}-\underline{b})$$
$$= (\underline{x}-\underline{b})^T(A^T)^{-1}A^{-1}(\underline{x}-\underline{b})$$
$$= (\underline{x}-\underline{b})^T(\underbrace{AA^T}_{\Sigma})^{-1}(\underline{x}-\underline{b})$$

$$= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-\frac{1}{2}(\underline{x}-\underline{b})^T \Sigma^{-1}(\underline{x}-\underline{b})}$$

as desired ∎

6. $1, 2, \ldots, n$

$\downarrow \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$

Permuted: 1 in pos σ_1

2 in pos σ_2

⋮

n in pos σ_n

$\downarrow$ move element in pos σ_1 to 1

" σ_2 to 2

⋮

```
σ_inverse = [0,...,0] //length n
σ_pos = 1
for each element σ_el of σ: //iterate over given σ list
    σ_inverse[σ_el] = σ_pos //1-based indexing
    σ_pos += 1
output σ_inverse //σ^-1
```

$\sigma^{-1}(j) =$ "σ.index(j)"

$$= \sum_{k=1}^{n} \delta_{\sigma(k),j}\, k$$

7. (a) $(P_\sigma x)_i = \sum_k^n (P_\sigma)_{ik} x_k = \sum_k^n (\underline{e}_{\sigma(i)})_k x_k = x_{\sigma(i)}$, since $(e_{\sigma(i)})_j = 1$ if $\sigma(i) = j$ else 0

(b) P_σ is orthonormal since the standard unit vectors are orthonormal and each $\sigma(1), \dots, \sigma(n)$ is unique
So $P_\sigma P_\sigma^T = P_\sigma^T P_\sigma = I$, i.e. the transpose of P_σ is its inverse.
So directly we have $P_\sigma^{-1} = P_\sigma^T$ (the first equality)

$$P_\sigma^T = \left[\underline{e}_{\sigma(1)} \ \ \underline{e}_{\sigma(2)} \ \cdots \ \underline{e}_{\sigma(n)}\right] = \begin{bmatrix} \underline{e}^T_{\sigma^{-1}(1)} \\ \underline{e}^T_{\sigma^{-1}(2)} \\ \vdots \\ \underline{e}^T_{\sigma^{-1}(n)} \end{bmatrix} = P_{\sigma^{-1}}$$ so $P_\sigma^T = P_{\sigma^{-1}}$ (outer equality)

a 1 in kth column on first row
implies $\sigma(k) = 1$ (i.e. $\underline{e}_1$ has leading 1)
The vector $\underline{e}_{\sigma^{-1}(1)}$ also has a 1 as it's
kth entry therefore. So transpose is first row, etc.

(c) $\det P_\sigma = \det P_\sigma^T$ by properties of determinant function

$\det P_\sigma^{-1} = \dfrac{1}{\det P_\sigma}$ also by properties "

But $P_\sigma^{-1} = P_\sigma^T$, so $\dfrac{1}{\det P_\sigma} = \det P_\sigma \iff 1 = (\det P_\sigma)^2 \iff |\det P_\sigma| = 1$ ∎