#4.

Consider a function $\varphi : [0,\infty] \to \mathbb{R}$ where $\varphi(x) = -\log x$.

First $[0,\infty]$ is a convex set since $0 < \eta x_1 + (1-\eta) x_2$ for $\forall x_1, x_2 \in [0,\infty]$, $\eta \in (0,1)$.

$\forall x_1, x_2 \in [0,\infty]$, $\eta \in (0,1)$

① $x_1 = x_2 \Rightarrow \varphi(\eta x_1 + (1-\eta) x_2) \le \eta \varphi(x_1) + (1-\eta)\varphi(x_2)$ (trivial) — Ⓐ

② $x_1 \ne x_2 \Rightarrow$ wlog) $x_1 < x_2$

$t = \eta x_1 + (1-\eta) x_2 \Rightarrow x_1 < t < x_2$.

By MVT) $\exists c_1 \in (x_1, t)$ s.t. $\dfrac{\varphi(t) - \varphi(x_1)}{t - x_1} = \varphi'(c_1)$

$\exists c_2 \in (t, x_2)$ s.t. $\dfrac{\varphi(x_2) - \varphi(t)}{x_2 - t} = \varphi'(c_2)$.

$\varphi''(x) = \frac{1}{x^2} > 0 \Rightarrow \varphi'(x)$ is a strictly increasing function.

$c_1 < c_2 \Rightarrow \varphi'(c_1) < \varphi'(c_2) \Rightarrow \dfrac{\varphi(t) - \varphi(x_1)}{t - x_1} < \dfrac{\varphi(x_2) - \varphi(t)}{x_2 - t}$

$\Rightarrow (x_2 - t)(\varphi(t) - \varphi(x_1)) < (t - x_1)(\varphi(x_2) - \varphi(t))$

$\Rightarrow (x_2 - x_1)\varphi(t) < (x_2 - t)\varphi(x_1) + (t - x_1)\varphi(x_2) = \eta(x_2 - x_1)\varphi(x_1) + (1-\eta)(x_2 - x_1)\varphi(x_2)$

$\Rightarrow \varphi(\eta x_1 + (1-\eta) x_2) < \eta\varphi(x_1) + (1-\eta)\varphi(x_2)$ — Ⓑ

$\Rightarrow$ By Ⓐ, Ⓑ, φ is a convex function.

for any pmf $p, q \in \mathbb{R}^n$

$$D_{KL}(p \| q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum_{i=1}^{n} -\log \frac{q_i}{p_i} \cdot p_i$$

$$= E_{x \sim P}\left[-\log \frac{q(x)}{p(x)}\right]$$ where P is the probability distribution with pmf p.

$$\ge -\log\left(E_{x\sim P}\left[\frac{q(x)}{p(x)}\right]\right) = -\log\left(\sum_{i=1}^{n} \frac{q_i}{p_i} \cdot p_i\right) = 0.$$

Holds by the Jensen's Inequality since $\varphi(x) = -\log x$ is convex.

■

#5.

By the analysis of ② in #4 $\Rightarrow$ Ⓑ shows that $\varphi : [0,\infty] \to \mathbb{R}$ is strictly convex.

Therefore, by the Jensen's inequality, $D_{KL}(p\|q) = E_{x\sim P}\left[-\log \frac{q(x)}{p(x)}\right] > -\log E_{x\sim P}\left[\frac{q(x)}{p(x)}\right] = 0$.

$p \ne q$ shows that the equality cannot hold.

■

#6

$$\frac{\partial}{\partial u_i} f_\theta(x) = \frac{\partial}{\partial u_i} \sum_{j=1}^{p} u_j \,\sigma(a_j x + b_j) = \sigma(a_i x + b_i)$$

$$\Rightarrow \nabla_u f_\theta(x) = \sigma(ax+b)$$

$$\frac{\partial}{\partial b_i} f_\theta(x) = \frac{\partial}{\partial b_i} \sum_{j=1}^{p} u_j \,\sigma(a_j x + b_j) = \sum_{j=1}^{p} u_j \cdot \frac{\partial}{\partial b_i} \left(\sigma(a_j x + b_j)\right)$$

$$= \sum_{j=1}^{p} u_j \cdot \delta_{ij} \cdot \sigma'(a_j x + b_j) = u_i \cdot \sigma'(a_i x + b_i)$$

$$\Rightarrow \nabla_b f_\theta(x) = \left(u_1 \sigma'(a_1 x + b_1), \cdots, u_p \sigma'(a_p x + b_p)\right)^T.$$

$$= \mathrm{diag}\left(\sigma'(ax+b)\right) u.$$

$$\frac{\partial}{\partial a_i} f_\theta(x) = \frac{\partial}{\partial a_i} \sum_{j=1}^{p} u_j \,\sigma(a_j x + b_j) = u_i \sigma'(a_i x + b_i) \cdot x. = \frac{\partial}{\partial b_i} f_\theta(x) \cdot x$$

$$\Rightarrow \nabla_a f_\theta(x) = \nabla_b f_\theta(x) \cdot x = \mathrm{diag}\left(\sigma'(ax+b)\right) \cdot ux.$$

$\blacksquare$