4.1-c-

$$\ell_{DPO}(\phi) = -\mathbb{E}_{(x,y_w,y_t)\sim D}\left[\log\sigma(\delta(\phi))\right] \text{ where}$$

where $\delta(\phi) = \beta\log\frac{\pi_\phi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\phi(y_l|x)}{\pi_{ref}(y_l|x)}$

$$\rightarrow \ell_{DPO}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma(\delta(\phi))\right] \text{ ——— } \textcircled{1}$$

Gradient on both sides $\nabla_\phi \ell_{DPO}(\phi) = \nabla_\phi(-\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma(\delta(\phi))\right]$

Properties of sigmoid → $\sigma(z) = \frac{1}{1+e^{-z}} \rightarrow \sigma'(z) = \sigma(z)(1-\sigma(z))$

Derivative of $\log(x) = \frac{1}{x}$

$$\nabla_\phi \ell_{DPO}(\phi) = -\mathbb{E}_{(x,y_w,y_t)\sim D}\left[\frac{1}{\sigma(\delta(\phi))}\cdot\sigma(\delta(\phi))\left[1-\sigma(\delta(\phi))\right]\cdot\nabla_\phi\delta(\phi)\right]$$

$$= -\mathbb{E}_{(x,y_w,y_t)\sim D}\left[\left[1-\sigma(\delta(\phi))\right]\cdot\nabla_\phi\delta(\phi)\right]$$

(Property of sigmoid $= \sigma(-\delta(\phi)) = 1-\sigma(\delta(\phi))$)

$$\nabla_\phi \ell_{DPO}(\phi) = -\mathbb{E}_{(x,y_w,y_t)\sim D}\left[\sigma(-\delta(\phi))\cdot\underbrace{\nabla_\phi\delta(\phi)}_{\textcircled{a}}\right] \text{ ——— } \textcircled{2}$$

Solving $\textcircled{a}$ $\nabla_\phi\delta(\phi) = \nabla_\phi\left[\beta\log\frac{\pi_\phi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\phi(y_l|w)}{\pi_{ref}(y_l|w)}\right]$

$$= \nabla_\phi\left[\beta\log\pi_\phi(y_w|x) - \beta\log\pi_{ref}(y_w|x) - \beta\log\pi_\phi(y_l|x) + \beta\log\pi_{ref}(y_l|x)\right]$$

$$\nabla_\phi\delta(\phi) = \beta\frac{\nabla_\phi\pi_\phi(y_w|x)}{\pi_\phi(y_w|x)} - \frac{\beta\nabla_\phi\pi_\phi(y_l|x)}{\pi_\phi(y_l|x)} \text{ ——— } \textcircled{b}$$

Putting $\textcircled{b}$ back in $\textcircled{2}$

$$\nabla_\phi \ell_{DPO}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\,\sigma(-\delta(\phi))\left[\beta\frac{\nabla_\phi\pi_\phi(y_w|x)}{\pi_\phi(y_w|x)} - \frac{\beta\nabla_\phi\pi_\phi(y_l|x)}{\pi_\phi(y_l|x)}\right]$$

Hence, Proved

$$\boxed{\nabla_\phi \ell_{DPO}(\phi) = \mathbb{E}_{(x,y_w,y_l)\sim D}\,\sigma(-\delta(\phi))\left[-\beta\frac{\nabla_\phi\pi_\phi(y_w|x)}{\pi_\phi(y_w|x)} + \beta\frac{\nabla_\phi\pi_\phi(y_l|x)}{\pi_\phi(y_l|x)}\right]}$$