#McGill

$J(\omega)$: expected return of policy π.

$d^\pi(s)$: stationary distribution.

Note that policy gradient theorem says

$$\nabla_z J(\pi) = \mathbb{E}_{s\sim d^\pi,\, a\sim\pi(\cdot|s)}\left[\nabla_z \log \pi(a|s)\cdot A^\pi(s|a)\right] \quad (*)$$

Since $\pi(a|s) = \dfrac{\exp(z(s,a_i))}{\sum_{a\in A}\exp(z(s,a))}$

$$\Rightarrow \underbrace{\log(\pi(a|s))} = z(s,a_i) - \log\left(\sum_{a\in A}\exp(z(s,a))\right)$$

$$\nabla \log(\pi(a|s)) = \frac{\nabla_z \pi(a|s)}{\pi(a|s)} \qquad (**)$$

substitute $(**)$ into $(*)$, get

$$\nabla_z J(\pi) = \mathbb{E}_{s\sim d^\pi,\, a\sim\pi(\cdot|s)}\left[\frac{\nabla_z \pi(a|s)}{\pi(a|s)}\cdot A^\pi(s|a)\right]$$

$$= \sum_{s\in S} d^\pi(s)\sum_{a\in A}\cancel{\pi(a|s)}\,\frac{\nabla_z \pi(a|s)}{\cancel{\pi(a|s)}}\,A^\pi(s|a)$$

$$= \sum_{s\in S} d^\pi(s)\underbrace{\left(\sum_{a\in A}\nabla_z \pi(a|s)\right)}_{\text{the integration of a gradient returns the original function : } \pi(a|s).} A^\pi(s|a)$$

$$= \sum_{s\in S} d^\pi(s)\left(\pi(a|s)\, A^\pi(s|a)\right)$$

since $d^\pi(s) := \sum_{s\in S} d^\pi(s)\, P(s'|s,\pi(s))$

$$= d^\pi(s)\,\pi(a|s)\,A^\pi(s|a) \quad \text{as desired}$$ ☒