APPENDIX: PROOF FOR EQ. (6)

In this section, we provide the step-by-step derivation for Eq. (6). We aim to compute the gradient of the loss function $\mathcal{L}_\pi(\theta)$ with respect to a specific logit parameter $\theta_{s,a}$ associated with action a. a denotes the specific action index associated with the logit parameter $\theta_{s,a}$ currently being updated (the target of the gradient), whereas a' is the random variable representing actions sampled from the policy $\pi_\theta(\cdot|s)$ used to calculate the expectation.

Recall the loss function defined as the negative expected soft value:

$$\mathcal{L}_\pi(\theta) = -\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}[g(s,a')] = -\sum_{a'\in\mathcal{A}} \pi_\theta(a'|s) g(s,a')$$

where $g(s,a') = Q_\phi(s,a') - \alpha \log \pi_\theta(a'|s)$.

Using the derivative of the softmax function (Lemma 2 from [14]), we have:

$$\frac{\partial \pi_\theta(a'|s)}{\partial \theta_{s,a}} = \pi_\theta(a'|s)\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right)$$

Now, we compute the gradient $\nabla_{\theta_{s,a}} \mathcal{L}_\pi(\theta)$:

$$\begin{aligned}
\nabla_{\theta_{s,a}} \mathcal{L}_\pi(\theta) &= -\sum_{a'\in\mathcal{A}} \frac{\partial \pi_\theta(a'|s)}{\partial \theta_{s,a}} g(s,a') \\
&= -\sum_{a'\in\mathcal{A}} \pi_\theta(a'|s)\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right) g(s,a') \\
&= -\mathbb{E}_{a'\sim\pi_\theta(\cdot|s)}\left[\left(\mathbf{1}\{a'=a\} - \pi_\theta(a|s)\right) g(s,a')\right] \quad \text{(Definition of Expectation)}
\end{aligned}$$

To simplify this expectation, we expand the summation by splitting it into the case where $a' = a$ and $a' \neq a$:

$$\begin{aligned}
\nabla_{\theta_{s,a}} \mathcal{L}_\pi(\theta) &= -\left[\pi_\theta(a|s)(1-\pi_\theta(a|s))g(s,a) + \sum_{a'\neq a} \pi_\theta(a'|s)(0-\pi_\theta(a|s))g(s,a')\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)^2 g(s,a) - \sum_{a'\neq a} \pi_\theta(a'|s)\pi_\theta(a|s)g(s,a')\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)\left(\pi_\theta(a|s)g(s,a) + \sum_{a'\neq a} \pi_\theta(a'|s)g(s,a')\right)\right] \\
&= -\left[\pi_\theta(a|s)g(s,a) - \pi_\theta(a|s)\underbrace{\sum_{a'\in\mathcal{A}} \pi_\theta(a'|s)g(s,a')}_{V_\pi(s)}\right] \\
&= -\pi_\theta(a|s)\left(g(s,a) - V_\pi(s)\right) \\
&= -\pi_\theta(a|s) A_{\text{soft}}(s,a)
\end{aligned}$$

This completes the derivation, confirming that the gradient descent update direction is proportional to the policy probability weighted by the soft advantage.