A. Proof of Lemma 1

Using the Bellman evaluation equation, we have

$$Q_{\text{soft}}^{\pi,r_2}(s,a) = r_2(s,a) + \gamma \mathbb{E}_{s',a'}\left[Q_{\text{soft}}^{\pi,r_2}(s',a') - \alpha \ln \pi(a'|s')\right]. \tag{1}$$

$$\Leftrightarrow \underbrace{Q_{\text{soft}}^{\pi,r_2}(s,a) + g(s)}_{=Q_{\text{soft}}^{\pi,r_1}(s,a)} = \underbrace{r_2(s,a) + g(s) - \gamma \mathbb{E}_{s'}[g(s')]}_{r_1(s,a)} + \mathbb{E}_{s',a'}\left[\underbrace{Q_{\text{soft}}^{\pi,r_2}(s',a') + g(s')}_{Q_{\text{soft}}^{\pi,r_1}(s',a')} - \alpha \ln \pi(a'|s')\right] \tag{2}$$

$$\Leftrightarrow Q_{\text{soft}}^{\pi,r_1}(s,a) = r_1(s,a) + \gamma \mathbb{E}_{s',a'}\left[Q_{\text{soft}}^{\pi,r_1}(s',a') - \alpha \ln \pi(a'|s')\right]. \tag{3}$$

This proves the stated result.

B. Proof of Theorem 1

Let $\pi' = \text{SPI}_{r_1}\{\pi\}$. We have, for any state-action couple,

$$\pi'(a|s) = \frac{\exp\{Q_{\text{soft}}^{\pi,r_1}(s,a)\}}{Z(s)} \tag{4}$$

$$= \frac{\exp\{Q_{\text{soft}}^{\pi,r_1}(s,a) + g(s)\}}{Z(s)\exp g(s)} \tag{5}$$

$$= \frac{\exp\{Q_{\text{soft}}^{\pi,r_2}(s,a)\}}{Z'(s)}. \tag{6}$$

The last equations means that $\pi' = \text{SPI}_{r_2}\{\pi\}$, and so $\text{SPI}_{r_1}\{\pi\} = \text{SPI}_{r_2}\{\pi\}$. To see that both rewards provide the same optimal policy, it is sufficient to notice that an optimal policy is the unique policy being greedy respectively to itself, that is $\pi_* = \text{SPI}_r\{\pi_*\}$. So, $\text{SPI}_{r_1}\{\pi\}$ and $\text{SPI}_{r_2}\{\pi\}$ have necessarily the same fixed point.

C. Proof of Theorem 2

Let π_1 and π_2 be two successive policies such that $\pi_2 = \text{SPI}_r\{\pi_1\}$. This means that, for any state s and action a, we have:

$$\pi_2(a|s) = \frac{\exp\{Q_{\text{soft}}^{\pi_1}(s,a)\}}{Z_1(s)}$$

where $Z_1(s)$ is a normalization factor. Taking the logarithm of this expression, we get:

$$\alpha \ln \pi_2(a|s) = Q_{\text{soft}}^{\pi_1}(s,a) - \ln Z_1(s) = Q_{\text{soft}}^{\pi_1}(s,a) + f(s).$$

According to Lemma 1, this means that $\alpha \ln \pi_2(a|s)$ is the Q-function associated to the shaped reward function $\bar{r}(s,a) = r(s,a) + f(s) - \gamma \mathbb{E}_{s'}\left[f(s')\right]$ for the policy π_1. Using the fact that this Q-function satisfies the Bellman equation, we have

$$\begin{aligned}\alpha \ln \pi_2(a|s) &= \bar{r}(s,a) + \gamma \mathbb{E}_{s',a'}\left[\alpha \ln \pi_2(a'|s') - \alpha \ln \pi_1(a'|s')\right] \\ &= \bar{r}(s,a) - \alpha\gamma \mathbb{E}_{s'}\left[\text{KL}(\pi_1(.|s') \| \pi_2(.|s'))\right] \\ \Leftrightarrow \bar{r}(s,a) &= \alpha \ln \pi_2(a|s) + \alpha\gamma \mathbb{E}_{s' \sim \mathcal{P}(.|a,s)}\left[\text{KL}(\pi_1(.|s') \| \pi_2(.|s'))\right].\end{aligned}$$

The fact that both r and $\bar{r}$ have the same optimal policy is due to theorem 1. This proves the stated result.