A. Proof of Lemma 1

Using the Bellman evaluation equation, we have

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

\[ \Leftrightarrow Q^\pi_{\text{soft}}(s, a) + g(s) = r_2(s, a) + g(s) - \gamma \mathbb{E}_{a'} [g(s')] + \mathbb{E}_{s', a'} \left[ Q^\pi_{\text{soft}}(s', a') + g(s') - \alpha \ln \pi(a'|s') \right] \]

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

This proves the stated result.

B. Proof of Theorem 1

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

\[ \pi'(a|s) = \frac{\exp\{Q^\pi_{\text{soft}}(s, a)\}}{Z(s)} \]

\[ = \frac{\exp\{Q^\pi_{\text{soft}}(s, a) + g(s)\}}{Z(s) \exp g(s)} \]

\[ = \frac{\exp\{Q^\pi_{\text{soft}}(s, a)\}}{Z'(s)}. \tag{6} \]

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 \( \pi_1 \) and \( \pi_2 \) 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:

\[ \pi_2(a|s) = \frac{\exp\{Q^\pi_{\text{soft}}(s, a)\}}{Z_1(s)} \]

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

\[ \alpha \ln \pi_2(a|s) = Q^\pi_{\text{soft}}(s, a) - \ln Z_1(s) = Q^\pi_{\text{soft}}(s, a) + f(s). \]

According to Lemma 1, this means that \( \alpha \ln \pi_2(a|s) \) is the Q-function associated to the shaped reward function \( \bar{r}(s, a) = r(s, a) + f(s) - \gamma \mathbb{E}_{s', a'}[f(s')] \) for the policy \( \pi_1 \). Using the fact that this Q-function satisfies the Bellman equation, we have

\[ \alpha \ln \pi_2(a|s) = \bar{r}(s, a) + \gamma \mathbb{E}_{s', a'} [\alpha \ln \pi_2(a'|s') - \alpha \ln \pi_1(a'|s')] \]

\[ = \bar{r}(s, a) - \alpha \gamma \mathbb{E}_{a'} [\text{KL}(\pi_1(\cdot|s')||\pi_2(\cdot|s'))] \]

\[ \Leftrightarrow \bar{r}(s, a) = \alpha \ln \pi_2(a|s) + \alpha \gamma \mathbb{E}_{a'\sim \bar{\pi}_1(\cdot|a, s)} [\text{KL}(\pi_1(\cdot|s')||\pi_2(\cdot|s'))] . \]

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