If the state s_t is perfectly observable, the "expert's policy" is defined as the Markovian policy that maximizes the likelihood of observed actions given observed states, i.e.:

$$\pi^e = \arg\max_{\pi} \sum_{n=1}^{N} \sum_{t=0}^{T-1} \log \pi(a_{t,n}|s_{t,n}) \tag{1}$$

However, in the POMDP model, the expert's policy only makes sense in the context of a *specific* model of perception because the states are not directly observable. Thus with beliefs $b_{\theta,t}$ the expert's policy is the solution to the maximum likelihood problem:

$$\pi^e(\cdot|b_\theta) = \arg\max_{\pi(\cdot|b_\theta)} \sum_{n=1}^{N} \sum_{t=0}^{T-1} \log \pi(a_{t,n}|b_{\theta,t,n}) \tag{2}$$

A regularization with respect to the "expert's policy" may be implemented through the specification of the information processing costs as follows:

$$c(\pi(\cdot|b_t)) = \alpha \mathcal{D}_{KL}(\pi(\cdot|b_{\theta,t})||\pi^e(\cdot|b_{\theta,t}))$$

where π^e is the "expert's policy" defined in (2). With this specification, the inner problem of IRL becomes:

$$\max_{\pi} E[\sum_{t\geq 0} \gamma^t (r_\phi(s_t, a_t) - c(\pi(\cdot|b_t))]$$

with solution

$$\pi_{\theta,\phi}(a|b_{\theta,t}) = \frac{\pi^e(a|b_{\theta,t}) \exp Q_{\theta,\phi,t}(b_{\theta,t}, a)}{\sum_{\tilde{a}\in\mathcal{A}} \pi^e(\tilde{a}|b_{\theta,t}) \exp Q_{\theta,\phi,t}(b_{\theta,t}, \tilde{a})} \tag{3}$$

The regularized estimation problem takes the form:

$$\begin{aligned} \max_{\theta,\phi} & \sum_{n=1}^{N} \sum_{t=0}^{T-1} \left[\log \sigma_\theta(o_{t+1,n}|b_{\theta,t,n}, a_{t,n}) + \log \pi_{\theta,\phi}(a_{t,n}|b_{\theta,t,n}) \right] \\ \text{s.t.} \quad & (3) \end{aligned}$$

where σ_θ is the observation probability map.