10-708: Probabilistic Graphical Models, Spring 2020

19: RL as Inference 1

Lecturer: Maruan Al-Shedivat Scribe: Harshit Sikchi, Yufei Wang, Mengdi Xu, Tianwei Ni, Yash Oza

1 Intro to Reinforcement Learning

In supervised learning we have a collection of data $D = [x_i, y_i]_{i=1}^n$ where our aim is to learn a model that approximates $P(y|x)$. In unsupervised learning we have a collection of data $D = [(x_1, x_2, x_3..., x_d)_i]_{i=1}^n$ where we seek learn a model which approximates $P(x_1, x_2..., x_d)$. Reinforcement learning sorts of closes the loop where the agent can interact with the world, obtain samples and learn a policy where it can maximize a reward function in the given environment. Reinforcement learning is useful as ultimately we want to build autonomous intelligent machines that can perceive and interact with the world, exhibit purposeful goal directed behavior and learn from interactions. Recently, Reinforcement learning have seen a number of successes where an RL agent was able to beat the world-master in game of GO, also in robotics where it was demonstrated that an robot trained with RL was able to learn how to manipulate a Rubik's cube to solve it.

Reinforcement learning can be specified as a Markov Decision Process(MDP). A MDP is specified by a set of states(S), a set of possible actions(A), environment dynamics($P(s_{t+1}|s_t, a_t)$) and a reward function r(s,a). The environment dynamics specify the transition probability of an agent from state s_t to state s_{t+1} after it takes an action a. The reward function provides a scalar feedback specifying a utility of the action. A trajectory in this MDP is represented as

$$\tau = (s_1, a_1, r_1, s_2, a_2, r_2, ..., s_H) \tag{1}$$

Using this framework we can choose to solve two common problems. The first is to find a policy $\pi : S \to A$ that outputs actions for each given state such that the cumulative reward along the trajectory is maximized. Alternatively we might be interested to find out the underlying MDP given a set of optimal trajectories. The first problem is the standard RL objective whereas the second one is known as Inverse Reinforcement learning.

1.1 Definitions

The **cumulative return** from timestep t is defined as the rewards accumulated starting from timestep t

$$G_t = r_{t+1} + r_{t+2} + .. + r_T \tag{2}$$

If $t = \infty$ the sum can become diverge and we can use the notion of discount factor γ, where $0 < \gamma < 1$ to get a finite sum.

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... \tag{3}$$

$$= r_{t+1} + \gamma G_{t+1} \tag{4}$$

A **policy** is a mapping from state to action. It can be deterministic as well as stochastic. In most general form, at any state s,

$$a \sim \pi(a|s) \tag{5}$$

Value function of a state s is defined as the expected cumulative reward obtained when starting from state s and following policy π.

$$V_\pi(s) := E_\pi[G_t|s_t = s] = E_\pi[\sum_{k=0}^{T} \gamma^k r_{t+k+1}|s_t = s] \tag{6}$$

Value function of a state-action pair or more commonly known as **Q function** of a state-action pair (s,a) is defined as the expected cumulative reward obtained when starting from state s, taking an action a and following policy π thereafter.

$$Q_\pi(s,a) := E_\pi[G_t|s_t = s, a_t = a] = E_\pi[\sum_{k=0}^{T} \gamma^k r_{t+k+1}|s_t = s, a_t = a] \tag{7}$$

1.2 Bellman Equations for Value and Q functions

Given the definition of value and Q functions, it is natural to derive the following Bellman Equations:

$$\begin{aligned} V_\pi(s) &:= \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{T} \gamma^k r_{t+k+1}|s_t = s\right] \\ &= \mathbb{E}_\pi[r_{t+1} + \gamma G_{t+1}|s_t = s] \\ &= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a) + \gamma \mathbb{E}_\pi[G_{t+1}|s_{t+1} = s']] \\ &= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V_\pi(s')] \end{aligned} \tag{8}$$

$$\begin{aligned} Q_\pi(s,a) &:= \mathbb{E}_\pi[G_t|s_t = s, a_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{T} \gamma^k r_{t+k+1}|s_t = s, a_t = a\right] \\ &= r(s,a) + \gamma \mathbb{E}_\pi[G_{t+1}|s_t = s, a_t = a] \\ &= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s') \mathbb{E}_\pi[G_{t+1}|s_{t+1} = s', a_{t+1} = a'] \\ &= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi(a'|s') Q_\pi(s', a') \end{aligned} \tag{9}$$

1.3 Optimal Policies and Value Functions

Goal of RL: find the optimal policy that achieves the highest expected returns. A policy π is better or equal to π' ($\pi \geq \pi'$) if its expected return is greater than that of π' in all states:

$$\pi \geq \pi' \Leftrightarrow V_\pi(s) \geq V_{\pi'}(s) \forall s \in S \tag{10}$$

Given this, we can define the optimal value and Q functions, and the Bellman Optimality Equations:

$$V_*(s) := \max_\pi V_\pi(s) = \max_a \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V_*(s')] \tag{11}$$

$$Q_*(s,a) := \max_\pi Q_\pi(s,a) = \sum_{s'} p(s'|s,a)[r(s,a) + \gamma \max_{a'} Q_*(s',a')] \tag{12}$$

If we can compute the optimal Q values $Q_*(s,a)$, then we can recover the optimal policy $\pi_*(a|s)$ as:

$$\pi_*(a|s) = \delta\left(a = \arg\max_a Q_*(s,a)\right) \tag{13}$$

To recover a set of optimal trajectories, we just need to execute the optimal policy:

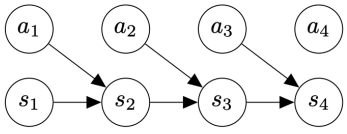
$$\begin{aligned} \tau_* &= (s_1^*, a_1^*, r_1^*, s_2^*, a_2^*, r_2^*, ...) \\ s_{t+1}^* &\sim p(s_{t+1}|s_t, a_t^* = \arg\max_a Q_*(s,a)) \end{aligned} \tag{14}$$

2 RL and Control as Inference: The GM framework

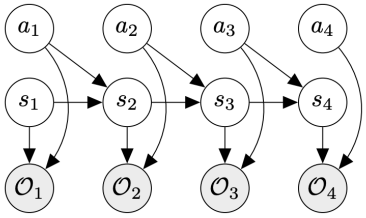
2.1 MDP as a Graphical Model

The graphical model for a standard MDP is shown on the left of Fig.1. The state is a Markov Chain and the states and actions are both random variables.

In MDP some transitions are rewarded with high rewards, and we hope to up weight the trajectories with high rewards and down weight the suboptimal ones. Therefore we augment the graphical model with an optimality variable $\mathcal{O}_t$ which is observable and makes it a Hidden Markov Process. The conditional distribution of the optimality variable is $p(\mathcal{O}_t = 1|s_t, a_t) = \exp(r(s_t, a_t))$. High rewards means the high probability of being optimal at time point t. Note that here we assume the reward are adjusted to make sure $p(\mathcal{O}_t = 1|s_t, a_t)$ is a probability distribution.



Initial state	$s_0 \sim p_0(s)$
Transition	$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$
Policy	$a_t \sim \pi(a_t \mid s_t)$
Reward	$r_t = r(s_t, a_t)$



Initial state	$s_0 \sim p_0(s)$
Transition	$s_{t+1} \sim p(s_{t+1} \mid s_t, a_t)$
Policy	$a_t \sim \pi(a_t \mid s_t)$
Reward	$r_t = r(s_t, a_t)$
Optimality	$p(\mathcal{O}_t = 1 \mid s_t, a_t) = \exp(r(s_t, a_t))$

Figure 1: MDP as a Graphical Model

Why the optimality variable $\mathcal{O}_t$ is important?

- The auxiliary variable $\mathcal{O}_t$ allows us to incorporate the reward information into a probabilistic generative process for sampling the trajectories. We can solve control and planning problems using probabilistic inference algorithms in this Hidden Markov Model.
- It allows us to probabilistically specify a model of optimal behavior, is importance for inverse RL.

- It also provides an explanation for why stochastic behavior might be preferred (for the explanation and transfer learning point of view).

Given the graphical model, we can

- Given a reward, determine how likely a trajectory to be optimal. Mathematically, we can compute $p(\tau, \mathcal{O}_{1:T})$, the probability of a trajectory τ given acting optimally throughout the trajectory.

$$\begin{aligned} p(\tau, \mathcal{O}_{1:T}) &\propto p(s_1) \prod_{t=1}^{T} p(a_t|s_t) p(s_{t+1}|s_t, a_t) p(\mathcal{O}_t|s_t, a_t) \\ &= p(s_1) \prod_{t=1}^{T} p(s_{t+1}|s_t, a_t) \exp(r(s_t, a_t) + \log p(a_t|s_t)) \\ &= \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1}|s_t, a_t) \right] \exp\left(\sum_{t=1}^{T} r(s_t, a_t) + \log p(a_t|s_t) \right) \end{aligned}$$

- Given a collection of optimal trajectories, infer the reward and priors, which is basically an inverse RL question.

$$\begin{aligned} p(\tau, \mathcal{O}_{1:T}, \theta, \phi) &\propto \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1}|s_t, a_t) \right] \exp\left(\sum_{t=1}^{T} r_\phi(s_t, a_t) + \log p_\theta(a_t|s_t) \right) \\ &= \left[p(s_1) \prod_{t=1}^{T} p(s_{t+1}|s_t, a_t) \right] \exp\left(\sum_{t=1}^{T} \phi^T f_r(s_t, a_t) + \log \theta^T f_p(a_t|s_t) \right) \end{aligned}$$

  The problem is a featurized CRF. By recovering the parametric potential functions f_r and f_p, we can learn the reward recovered from the trajectories. Note that CRF is undirected and does not preserve the casual structure; this model is more restrictive and known as MEMM.

- Given a reward, infer the optimal policy by calculating $p(a_t|s_t, \mathcal{O}_{t:T})$. Instead of solving the optimization problem, we now can solve the inference problem.

2.2 Optimal Policy via Inference

Now we aim to infer the optimal policy $p(\mathbf{s}_t|\mathbf{a}_t, \mathcal{O}_{t:T})$ by standard message passing algorithm. It is sufficient to compute the backward message $\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)$, which denotes the probability of a trajectory to be optimal from t to T starting from the state and action at time t. Also we introduce the message $\beta_t(\mathbf{s}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t)$. Then the messages can be computed recursively:

$$\begin{aligned} \beta_t(\mathbf{s}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t) &= \int_{\mathcal{A}} p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t) p(\mathbf{a}_t|\mathbf{s}_t) d\mathbf{a}_t = \int_{\mathcal{A}} \beta_t(\mathbf{s}_t, \mathbf{a}_t) p(\mathbf{a}_t|\mathbf{s}_t) d\mathbf{a}_t \\ \beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t) &= \int_{\mathcal{A}} \beta_{t+1}(\mathbf{s}_{t+1}) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) p(\mathcal{O}_t|\mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1} \end{aligned} \tag{15}$$

Then the optimal action distribution can be derived by two backward messages:

$$\begin{aligned} \pi(\mathbf{a}_t|\mathbf{s}_t) := p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{t:T}) &= \frac{p(\mathbf{s}_t, \mathbf{a}_t|\mathcal{O}_{t:T})}{p(\mathbf{s}_t|\mathcal{O}_{t:T})} = \frac{p(\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{t:T})}{p(\mathbf{s}_t, \mathcal{O}_{t:T})} = \frac{p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t) p(\mathbf{s}_t) p(\mathbf{a}_t|\mathbf{s}_t)}{p(\mathcal{O}_{t:T}|\mathbf{s}_t) p(\mathbf{s}_t)} \\ &\propto \frac{p(\mathcal{O}_{t:T}|\mathbf{s}_t, \mathbf{a}_t)}{p(\mathcal{O}_{t:T}|\mathbf{s}_t)} = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)} \end{aligned} \tag{16}$$

Here we assume the action prior is a uniform distribution $p(\mathbf{a}_t|\mathbf{s}_t) = 1/|\mathcal{A}|$.

Then in order to get more intuition in RL, we introduce the message in log-space:

$$\begin{aligned} Q(\mathbf{s}_t, \mathbf{a}_t) &= \log \beta_t(\mathbf{s}_t, \mathbf{a}_t) \\ V(\mathbf{s}_t) &= \log \beta_t(\mathbf{s}_t) \\ \pi(\mathbf{a}_t|\mathbf{s}_t) &\propto \exp(Q(\mathbf{s}_t, \mathbf{a}_t) - V(\mathbf{s}_t)) = \exp(A_t(\mathbf{s}_t, \mathbf{a}_t)) \end{aligned} \tag{17}$$

Actually, the log-messages Q, V correspond to the state-action and state value function in a soft version. The action distribution is proportional to advantage value. Moreover, by the relationship in 15, we can derive the following relationship for Q, V:

$$\begin{aligned} V(\mathbf{s}_t) &= \log \int_{\mathcal{A}} \exp(Q(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t \approx \max_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t) \\ Q(\mathbf{s}_t, \mathbf{a}_t) &= \log p(\mathcal{O}_t|\mathbf{a}_t, \mathbf{s}_t) + \log \int \beta_{t+1}(\mathbf{s}_{t+1}) p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1} = r(\mathbf{s}_t, \mathbf{a}_t) + \log \mathbb{E}_{\mathbf{s}_{t+1}}[\exp(V(\mathbf{s}_{t+1})] \end{aligned} \tag{18}$$

Thus V can be seen as the soft-max of Q. When the dynamic is deterministic, the second relationship is exactly Bellman equation backup:

$$Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V(\mathbf{s}_{t+1}) \tag{19}$$

However, when the dynamic is stochastic, the update is optimistic, because it will be largely determined by the max of next state value, which creates risk-seeking behavior. This issue will be mitigated by variational inference in the next section.

In conclusion, with the PGM augmented by optimality variables, we reduced the optimal control to inference in a HMM-like model, and make its connections with dynamic programming, value iteration in RL field.

3 Connections to Variational Inference

3.1 Which objective does the inference optimize?

Recall that the optimal trajectory distribution:

$$p(\tau) \propto \left[p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right) \tag{20}$$

Now we aim to optimize an approximate policy to be optimized to be closed to this trajectory distribution. Let the policy be $\pi(\mathbf{a}_t|\mathbf{s}_t)$, then its trajectory distribution under **deterministic** dynamics (where $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{t:T}) = p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$):

$$\begin{aligned} \hat{p}(\tau) &= p(\mathbf{s}_1|\mathcal{O}_{1:T}) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{t:T}) \pi(\mathbf{a}_t|\mathbf{s}_t) \\ &= p(\mathbf{s}_1|\mathcal{O}_{1:T}) \prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \pi(\mathbf{a}_t|\mathbf{s}_t) \end{aligned} \tag{21}$$

In case of exact inference derived in the last section, the match is exact, i.e. $D_{KL}(\hat{p}(\tau)||p(\tau)) = 0$. Therefore we can view the optimization objective as maximizing the negative KL divergence:

$$\begin{aligned}
&\max_{\pi} - D_{KL}(\hat{p}(\tau)||p(\tau)) \\
&\quad = \mathbb{E}_{\tau\sim\hat{p}}[\log p(\tau) - \log \hat{p}(\tau)] \\
&\quad = \mathbb{E}_{\tau\sim\hat{p}}\left[\log \frac{p(\mathbf{s}_1)}{p(\mathbf{s}_1)} + \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t) + \log \frac{p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}{p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}\right] \\
&\quad = \mathbb{E}_{\tau\sim\hat{p}}\left[\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)\right] \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\hat{p}}[r(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)] \\
&\quad = \sum_{t=1}^{T} \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\hat{p}}[r(\mathbf{s}_t, \mathbf{a}_t)] + \mathbb{E}_{\mathbf{s}\sim\hat{p}}[\mathcal{H}(\pi(\mathbf{a}_t|\mathbf{s}_t))]
\end{aligned} \tag{22}$$

Now, the dynamics under deterministic conditions are given as -
$Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V(\mathbf{s}_{t+1})$
and under stochastic conditions are given as -
$Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log E_{\mathbf{s}_{t+1}\sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)}[\exp(V(\mathbf{s}_{t+1}))]$

Here, rather than having the optimistic term (which assumes that if any of the future states have a high reward regardless of the intermediary states that lead to there, the exponential term will favor that high-reward-state only), we would like to ask the question - given that a high reward was obtained in the past, what is the action probability, given that the transition probability did not change ?

3.2 Control via variational inference

To address the above mentioned issues, we would use Variational Inference, where the goal is to find $q(s_{1:T}, a_{1:T})$ such that it approximates $p(s_{1:T}, a_{1:T}|\mathcal{O}_{1:T})$ while the dynamics stays fixed to $p(s_{t+1}|s_t, a_t)$.

The distribution over optimal trajectories is given as $p(\tau) = \left[p(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\right] \exp\left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t)\right)$ and the policy induced distribution is given as

$q(\tau) = q(\mathbf{s}_1)\prod_{t=1}^{T} p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\, q(\mathbf{a}_t|\mathbf{s}_t)$

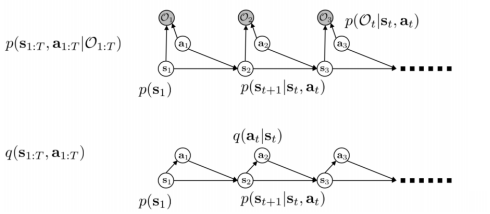


Figure 2: Graphical models with and without the optimality variables. Using variational inference, we try to find a variational distribution(bottom) that approximates the original distribution(above) well.

Hence, we can compute the Evidence Lower Bound as -

$$\begin{aligned}
\log \mathrm{p}\left(\mathcal{O}_{1:\mathrm{T}}\right) &= \log \iint \mathrm{p}\left(\mathcal{O}_{1:\mathrm{T}}, \mathbf{s}_{1:\mathrm{T}}, \mathbf{a}_{1:\mathrm{T}}\right) \mathrm{d}\mathbf{s}_{1:\mathrm{T}} \mathrm{d}\mathbf{a}_{1:\mathrm{T}} \\
&= \log \iint \mathrm{p}\left(\mathcal{O}_{1:\mathrm{T}}, \mathbf{s}_{1:\mathrm{T}}, \mathbf{a}_{1:\mathrm{T}}\right) \frac{\mathrm{q}\left(\mathbf{s}_{1:\mathrm{T}}, \mathbf{a}_{1:\mathrm{T}}\right)}{\mathrm{q}\left(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)} \mathrm{d}\mathbf{s}_{1:\mathrm{T}} \mathrm{d}\mathbf{a}_{1:T} \\
&= \log \mathbb{E}_{\left(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right) \sim \mathbf{q}\left(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)}\left[\frac{p\left(\mathcal{O}_{1:T}, \mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)}{q\left(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right)}\right] \\
&\geq \mathbb{E}_{\left(\mathrm{s}_1\right)}\left[\log \mathrm{p}\left(\mathcal{O}_{1:\mathrm{T}}, \mathbf{s}_{1:\mathrm{T}}, \mathbf{a}_{1:\mathrm{T}}\right) - \log \mathrm{q}\left(\mathbf{s}_{1:\mathrm{T}}, \mathbf{a}_{1:\mathrm{T}}\right)\right] \\
&= \mathbb{E}_{\tau \sim q}\left[\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right) - \log q\left(\mathbf{a}_t | \mathbf{s}_t\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\left(s_t, \mathbf{a}_t\right) \sim q}\left[\mathbf{r}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] + H\left(\mathbf{q}\left(\mathbf{a}_t | \mathbf{s}_t\right)\right)
\end{aligned}$$

using Jensen's inequality to lower bound the log-probability of the observable variables. Hence, now the objective is composed of two components just like the deterministic case, but in terms of the variational distribution. The first term is the expected return induced by the variational policy, and the second term is the entropy of the variational policy. Now, to obtain the optimal policy, we have

$$q\left(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}\right) = p\left(\mathbf{s}_1\right) \Pi_t p\left(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t\right) q\left(\mathbf{a}_t | \mathbf{s}_t\right) \tag{23}$$

$$\log p\left(\mathcal{O}_{1:T}\right) \geq \sum_t E_{\left(\mathbf{s}_t, \mathbf{a}_t\right) \sim q}\left[r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \mathcal{H}\left(q\left(\mathbf{a}_t | \mathbf{s}_t\right)\right)\right] \tag{24}$$

Solving further, we have

$$q\left(\mathbf{a}_T | \mathbf{s}_T\right) = \arg\max E_{\mathbf{s}_T \sim q\left(\mathbf{s}_T\right)}\left[E_{\mathbf{a}_T \sim q\left(\mathbf{a}_T | \mathbf{s}_T\right)}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right] + \mathcal{H}\left(q\left(\mathbf{a}_T | \mathbf{s}_T\right)\right)\right] \tag{25}$$

$$\arg\max E_{\mathbf{s}_T \sim q\left(\mathbf{s}_T\right)}\left[E_{\mathbf{a}_T \sim q\left(\mathbf{a}_T | \mathbf{s}_T\right)}\left[r\left(\mathbf{s}_T, \mathbf{a}_T\right) - \log q\left(\mathbf{a}_T | \mathbf{s}_T\right)\right]\right] \tag{26}$$

This is minimized when $q\left(\mathbf{a}_T | \mathbf{s}_T\right) \propto \exp\left(r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right)$

$$q\left(\mathbf{a}_T | \mathbf{s}_T\right) = \frac{\exp\left(r\left(\mathbf{s}_T, \mathbf{a}_T\right)\right)}{\int \exp\left(r\left(\mathbf{s}_T, \mathbf{a}\right)\right) d\mathbf{a}} = \exp\left(Q\left(\mathbf{s}_T, \mathbf{a}_T\right) - V\left(\mathbf{s}_T\right)\right) \tag{27}$$

And the value function is given as -

$$V\left(\mathbf{s}_T\right) = \log \int \exp\left(Q\left(\mathbf{s}_T, \mathbf{a}_T\right)\right) d\mathbf{a}_T \tag{28}$$