Note on Maximum Entropy RL and Soft Value Iteration

Wen Sun[1]

[1]Department of Computer Science, Cornell University

April 22, 2021

1 Problem Formulation

Consider a finite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, c, P, H, \mu\}$, where we work on the minimizing cost setting, i.e., $c(s,a) \in [0,1]$ is a cost function rather than a reward function.

Recall the for finite horizon setting, the optimal policy will be a sequence of time-dependent policies. For notation simplicity, we denote $\pi = \{\pi_0, \ldots, \pi_{H-1}\}$.

Different from the traditional setting, we are interested in optimizing the following objective:

$$\min_{\pi} \sum_{h=0}^{H-1} \mathbb{E}_{s\sim \mathbb{P}_h^{\pi}(\cdot;\mu)} \left[\mathbb{E}_{a\sim\pi_h(\cdot|s)} c(s,a) - \text{entropy}(\pi_h(\cdot|s))\right]$$

Namely, we want to find a policy that minimizes the expected cost, while at the same make sure that the policy has large entropy (i.e., being more stochastic in choosing actions).

Using the definition of entropy, we can rewrite the above objective as follows:

$$\min_{\pi} \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim \mathbb{P}_h^{\pi}(\cdot,\cdot;\mu)} \left[c(s,a) + \ln \pi_h(a|s)\right]$$

This looks like a regular RL objective, except that we have a new cost term that is $c(s,a) + \ln \pi_h(a|s)$ which is dependent on our policy. Note that if the new cost were independent of our policy, then we could use classic approaches such as value iteration or policy iteration to compute the optimal policy. But what we could do if the cost term contains $\ln \pi_h(a|s)$? We will see that we can still solve this problem using Dynamic programming again.

2 Soft Value Iteration

In the process of Dynamic programming, we will define $V_h^{\star}(s)$ as follows:

$$V_h^{\star}(s) = \min_{\pi_h \ldots \pi_{H-1}} \mathbb{E}\left[\sum_{t=h}^{H-1} c(s_t,a_t) + \ln \pi_t(a_t|s_t) | s_h = s, a_t \sim \pi_t(\cdot|s_t)\right],$$

i.e., the minimum expected total cost-to-go (including the $\ln \pi_t(a_t|s_t)$ term) starting at state s at time step h.

For the base case of DP, we will just start from the ghost step H (recall in H-step finite horizon MDP, we end at $H-1$), i.e., we set:

$$V_H^{\star}(s) = 0, \forall s.$$

Namely, this is an additional time step, we do not need to do anything there and there is no cost there either.

Now let us start the inductive hypothesis step. Assume that we have $V^\star_{h+1}(s)$ for all s available. We show that how we can compute $\pi^\star_h$ and $V^\star_h$ here.

As usual, let us denote $Q^\star_h(s,a)$ as follows:

$$Q^\star_h(s,a) = c(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\star_{h+1}(s').$$

Note that $Q^\star_h$ contains the entropy-related terms from $h+1$ (since $V^\star_{h+1}$ includes the entropy-related terms from $h+1$ to H) but it does not include the entropy-related term at h yet.

Let us focus on s, and try to compute the optimal strategy at s at time step h. The optimal strategy at s,h should be the following minimizer:

$$\min_{\rho\in\Delta(A)} \left[\mathbb{E}_{a\sim\rho}\left[c(s,a) + \ln\rho(a) + \mathbb{E}_{s'\sim P(s,a)} V^\star_{h+1}(s')\right]\right]$$

which is equivalent to:

$$\min_{\rho\in\Delta(A)} \left[\mathbb{E}_{a\sim\rho}\left[Q^\star_h(s,a) + \ln\rho(a)\right]\right]$$

We can compute the exact minimizer of the above formulation using Lagrange formulation (note that here we have an implicit constraint $\sum_{a\in A}\rho(a) = 1$ as ρ is a distribution over actions). We skip the Lagrange formulation step, but it is not hard to verify the the optimal strategy at s can be computed as follows:

$$\pi^\star_h(\cdot|s) = \frac{\exp(-Q^\star_h(s,a))}{\sum_{a'\in A}\exp(-Q^\star_h(s,a))}.$$

This means that we assign high probabilities to actions (exponential with respect to $-Q^\star(s,a)$) that have low cost-to-go $Q^\star(s,a)$, and assign non-zero low probability to actions that have high cost-to-go $Q^\star_h(s,a)$. [1]

With $\pi^\star_h(\cdot|s)$ available, we can compute $V^\star_h(s)$ as follows:

$$\begin{aligned}
V^\star_h(s) &= \min_{\rho\in\Delta(A)} \left[\mathbb{E}_{a\sim\rho}(c(s,a) + \ln\rho(a) + \mathbb{E}_{s'\sim P(s,a)} V^\star_{h+1}(s'))\right] \\
&= \mathbb{E}_{a\sim\pi^\star_h(\cdot|s)}(c(s,a) + \ln\rho(a) + \mathbb{E}_{s'\sim P(s,a)} V^\star_{h+1}(s')) \\
&= -\ln\left(\sum_{a\in A}\exp\left(-Q^\star_h(s,a)\right)\right).
\end{aligned}$$

This is a soft-min operator, i.e., instead of $\min_a Q^\star(s,a)$, we smooth the value out here using $-\ln(\sum_a \exp(-Q^\star(s,a)))$.

Now we can continue the above process until we are done at time step $h=0$. We output $\pi^\star_h$ for all $h = 0,\ldots,H-1$.

Remark Let us consider an extreme case where $Q^\star(s,a) = 0$ while $Q^\star(s,a') = +\infty$ for any $a' \neq a$. Namely a has zero cost-to-go while other actions all have extremely large cost-to-go. In this case, we will have $\pi^\star(a|s) = 1$ and $\pi^\star(a'|s) = 0$ for $a' \neq a$, and $V^\star_h(s) = \min_a Q^\star(s,a)$. Namely, in the extreme case, the policy becomes deterministic again, and the soft-min operator simply picks the minimum value. Another extreme is that all actions have the same cost-to-go, i.e., $Q^\star_h(s,a) = Q^\star_h(s,a')$ for all $a,a' \in A$. In this case, $\pi^\star_h(\cdot|s)$ will be a uniform distribution over A. Namely, in this case, $\pi^\star_h(\cdot|s)$ will not commit to a unique action (all actions are equally good) but instead will try to be as random as possible (recall uniform distribution has the maximum entropy).

References

[1] This step should be contrast to the classic DP step where we would have $\pi^\star_h(s) = \arg\min_a Q^\star_h(s,a)$.