Supplementary Material for Monte Carlo Bayesian Reinforcement Learning

1. Proof of Theorem 1

To prove the theorem, we need two lemmas.

Lemma 1. *For any* $\tau_1 \in (0,1)$,

$$\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\left((|\hat{\pi}||O|+2)\ln|\hat{\pi}| + |\hat{\pi}|\ln|A| + \ln(2/\tau_1)\right)}{K}}$$

with probability at most τ_1.

Proof. To prove the lemma, we consider all policies and bound the probability that any policy π has estimate $\hat{V}_\pi$ with error larger the its specified bound.

Consider an arbitrary policy π_i with size i. By definition, V_{π_i} is the value of π_i for the hybrid POMDP $\mathcal{P}$, which is a constant calculated with respect to the initial belief $b^0_{\mathcal{P}}(\theta)$ over all possible parameter values. Let $V(\pi_i, \theta)$ be the value of policy π_i when parameter θ is used. Therefore, we have $V_{\pi_i} = E\left(V(\pi_i, \theta)\right)$.

On the other hand, $\hat{V}_{\pi_i}$ is the value of π_i for the discrete POMDP $\hat{\mathcal{P}}$, which is formulated with a uniform prior over K hypotheses $\left(\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^K\right)$. We can write $\hat{V}_{\pi_i} = \frac{1}{K}\sum_{k=1}^{K} V(\pi_i, \hat{\theta}^k)$.

As the K hypotheses $\left(\hat{\theta}^1, \hat{\theta}^2, \ldots, \hat{\theta}^K\right)$ are independently sampled from $b^0_{\mathcal{P}}(\theta)$, Hoeffding's inequality gives

$$\begin{aligned} & p\left(\hat{V}_{\pi_i} - V_{\pi_i} \geq \epsilon_i\right) \\ = \;& p\left(\frac{1}{K}\sum_{k=1}^{K} V(\pi_i, \hat{\theta}^k) - E\left(V(\pi_i,\theta)\right) \geq \epsilon_i\right) \\ \leq \;& \exp\left(-\frac{K\epsilon_i^2}{2C^2}\right), \end{aligned} \tag{1}$$

where $C = \frac{R_{\max}}{1-\gamma}$.

Let E_{π_i} denote the event that $\hat{V}_{\pi_i} - V_{\pi_i} \geq \epsilon_i$ and δ_i denote the RHS of inequality (1). Applying the union bound and inequality (1), the probability that at least one policy π_i has error greater than ϵ_i, for all i, is bounded as follows

$$\begin{aligned} p\left(\bigcup_{\forall \pi_i, i} E_{\pi_i}\right) \;&\leq\; \sum_{\forall \pi_i, i} p\left(E_{\pi_i}\right) \\ &\leq\; \sum_{i=1}^{\infty} |\Pi_i|\,\delta_i. \end{aligned} \tag{2}$$

Here, $|\Pi_i|$ denotes the number of policies with size i. In a policy π_i, each node has $|A|$ possible labels and $|O|$ outgoing edges. Each edge has i possible ending nodes. Therefore, $|\Pi_i| = \left(|A| \cdot i^{|O|}\right)^i$.

Set

$$\delta_i = \frac{\tau_1}{2i^2|\Pi_i|} \tag{3}$$

and plug it into inequality (2), we have

$$\begin{aligned} p\left(\bigcup_{\forall \pi_i, i} E_{\pi_i}\right) \;&\leq\; \frac{\tau_1}{2}\sum_{i=1}^{\infty}\frac{1}{i^2} \\ &\leq\; \frac{6\tau_1}{\boldsymbol{\pi}^2}\sum_{i=1}^{\infty}\frac{1}{i^2} \\ &=\; \tau_1, \end{aligned}$$

where $\boldsymbol{\pi}$ denotes the constant $3.1415\ldots$.

Note that the event $\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \epsilon_{|\hat{\pi}|}$ is a subset of $\bigcup_{\forall \pi_i, i} E_{\pi_i}$. Therefore,

$$p\left(\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \epsilon_{|\hat{\pi}|}\right) \leq p\left(\bigcup_{\forall \pi_i, i} E_{\pi_i}\right) \leq \tau_1.$$

Since δ_i is defined to be the RHS of inequality (1), plugging into Equation (3), we have

$$\exp\left(-\frac{K\epsilon_i^2}{2C^2}\right) = \frac{\tau_1}{2i^2|\Pi_i|}.$$

Solving for $\epsilon_{|\hat{\pi}|}$, we obtain

$$\epsilon_{|\hat{\pi}|} = \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\left((|\hat{\pi}||O|+2)\ln|\hat{\pi}| + |\hat{\pi}|\ln|A| + \ln(2/\tau_1)\right)}{K}}.$$

□

Lemma 2. *Assume that* $\hat{V}_{\hat{\pi}^\star} - \hat{V}_{\hat{\pi}} \leq \delta$. *For any* $\tau_2 \in (0,1)$,

$$V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\ln(1/\tau_2)}{K}} + \delta$$

with probability at most τ_2.

Proof. First, as an instance of inequality (1), we have

$$p\left(V_{\pi^\star} - \hat{V}_{\pi^\star} \geq \epsilon\right) \leq \exp\left(-\frac{K\epsilon^2}{2C^2}\right),$$

where $C = \frac{R_{\max}}{1-\gamma}$.

Next, we show that the event $V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon + \delta$ is a subset of the event $V_{\pi^\star} - \hat{V}_{\pi^\star} \geq \epsilon$. This is true because if $V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon + \delta$ holds, then

$$\begin{aligned} V_{\pi^\star} - \hat{V}_{\pi^\star} &= \left(V_{\pi^\star} - \hat{V}_{\hat{\pi}}\right) - \left(\hat{V}_{\pi^\star} - \hat{V}_{\hat{\pi}}\right) \\ &\geq \left(V_{\pi^\star} - \hat{V}_{\hat{\pi}}\right) - \left(\hat{V}_{\hat{\pi}^\star} - \hat{V}_{\hat{\pi}}\right) \\ &\geq \epsilon + \delta - \delta \\ &= \epsilon, \end{aligned}$$

where $\hat{\pi}^\star$ denotes the optimal policy to the discrete POMDP $\hat{\mathcal{P}}$ and $\hat{V}_{\hat{\pi}^\star} \geq \hat{V}_{\pi^\star}$.

Finally, we have

$$\begin{aligned} p(V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon + \delta) &\leq p(V_{\pi^\star} - \hat{V}_{\pi^\star} \geq \epsilon) \\ &\leq \exp\left(-\frac{K\epsilon^2}{2C^2}\right). \end{aligned}$$

Set the RHS equal to τ_2, we conclude that

$$V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \frac{R_{\max}}{1-\gamma}\sqrt{\frac{2\ln(1/\tau_2)}{K}} + \delta$$

with probability at most τ_2. □

Lemmas 1 and 2 identify two sources of the approximation error in Theorem 1, and provide PAC bounds on them. Now we can prove the theorem by combining the two error bounds.

Proof. It is clear that $V_{\pi^\star} - V_{\hat{\pi}} \geq \epsilon_1 + \epsilon_2$ implies that either $\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \epsilon_1$ or $V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon_2$ should hold. Therefore, the event $V_{\pi^\star} - V_{\hat{\pi}} \geq \epsilon_1 + \epsilon_2$ is a subset of the event $\left(\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \epsilon_1\right) \cup \left(V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon_2\right)$ and

$$p\left(V_{\pi^\star} - V_{\hat{\pi}} \geq \epsilon_1 + \epsilon_2\right) \leq p\left(\left(\hat{V}_{\hat{\pi}} - V_{\hat{\pi}} \geq \epsilon_1\right) \cup \left(V_{\pi^\star} - \hat{V}_{\hat{\pi}} \geq \epsilon_2\right)\right).$$

Combining this with Lemmas 1 and 2 and setting $\tau_1 = \tau_2 = \frac{\tau}{2}$, we have

$$\begin{aligned} V_{\pi^\star} - V_{\hat{\pi}} \leq\ & \delta + \frac{R_{\max}}{1-\gamma}\left(\sqrt{\frac{2\ln(2/\tau)}{K}} + \right. \\ & \left. \sqrt{\frac{2\left((|\hat{\pi}||O|+2)\ln|\hat{\pi}| + |\hat{\pi}|\ln|A| + \ln(4/\tau)\right)}{K}}\right) \end{aligned}$$

with probability at least $1-\tau$.

Note that

$$\sqrt{\frac{2\ln(2/\tau)}{K}} < \sqrt{\frac{2\left((|\hat{\pi}||O|+2)\ln|\hat{\pi}| + |\hat{\pi}|\ln|A| + \ln(4/\tau)\right)}{K}}.$$

Therefore,

$$V_{\pi^\star} - V_{\hat{\pi}} \leq \frac{2R_{\max}}{1-\gamma}\sqrt{\frac{2\left((|\hat{\pi}||O|+2)\ln|\hat{\pi}| + |\hat{\pi}|\ln|A| + \ln(4/\tau)\right)}{K}} + \delta$$

with probability at least $1-\tau$. □

2. Detailed Settings of Intersection Navigation Problem

In this section we provide the detailed settings of the Intersection Navigation problem.

As discussed in Section 5.4, for a given driver A, the underlying decision problem for agent R is modeled as a POMDP. The state consists of the positions and velocities of R and A. For simplicity, the positions are discretized into a uniform grid with cell size $5\,\mathrm{m} \times 5\,\mathrm{m}$, as shown in Figure 1(a). The velocities are uniformly discretized into 5 levels, ranging from $0\,\mathrm{m/s}$ to $4\,\mathrm{m/s}$. Each time step has duration $\Delta t = 0.5\,\mathrm{s}$.

In each time step, the agent R can take 3 actions to change its velocity: ACCELERATE ($1\,\mathrm{m/s^2}$), MAINTAIN ($0\,\mathrm{m/s^2}$), and DECELERATE ($-1\,\mathrm{m/s^2}$). The actions are imperfect with a failure rate 5%. For ACCELERATE and DECELERATE, a failure causes no change to the speed. For MAINTAIN, a failure causes the speed to be increased or decreased by $1\,\mathrm{m/s^2}$ at random.

After taking an action, the POMDP transits to a new state, and the agent R receives an observation on that state. The observation on its own position and speed is accurate. In contrast, the observation on the state of A is noisy. With probability 10%, R will wrongly observe that A locates at the grid cells adjacent to its actual position. The observation model for the speed of A is defined in a similar way.

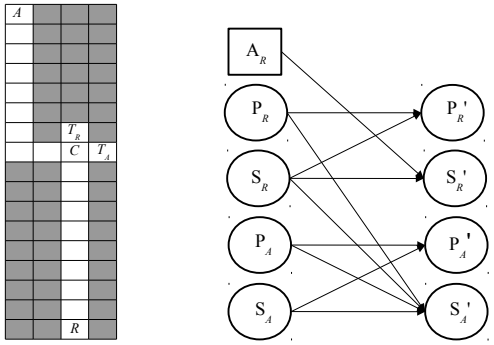
For the reward function, R receives a reward of 500 for safely crossing the intersection, and a large penalty -2500 for collision with A. To expedite R crossing the intersection faster, we give it a penalty of -5 in each time step.

The transition function is defined based on the action taken by R and the driving strategy of A. Figure 1(b) shows the dynamic Bayesian network that encodes the structure of the transition function. The current speed S'_R of R depends on the action A_R and the speed S_R in the previous step. Its transition function is naturally defined according to the effect of the noisy actions.

The current position P'_R of R depends on its position P_R and speed S_R in the previous step. The transition function is defined as

$$\Pr(P'_R|P_R, S_R) = \begin{cases} 1 - \frac{1}{t}, & P'_R = P_R \\ \frac{1}{t}, & P'_R \text{ is the cell next to } P_R \\ 0, & \text{otherwise} \end{cases}$$

where t is the expected number of steps to move from P_R to P'_R under speed S_R. Intuitively, the longer it takes to



(a) (b)

Figure 1. (a) The discretized environment. The two vehicles start in cells R and A, and move towards the terminal cells T_R and T_A, respectively. C is the point of potential collision. Shaded cells are not passable. (b) The DBN for the transition function of the POMDP.

travel from P_R to P'_R, the smaller the transition probability. The transition function for the position P'_A of A is defined in a similar way.

The current speed S'_A of A depends on the previous states of both A and R. This allows us to model reactive drivers that take the state of R into consideration when driving. The transition function for S'_A is defined using the Gipps car following model (Gipps, 1981), which is a parametric model commonly used in transportation research for estimating car velocity. It estimates the car speed by the following equations:

$$\begin{aligned} v_{\text{safe}} &= v(t) + \frac{g(t) - v(t)\tau}{\frac{v}{b} + \tau}, \\ v_{\text{des}} &\leftarrow \min\left(v_{\max}, v + a, v_{\text{safe}}\right), \\ v &\leftarrow \max\left(0, rand\left(v_{\text{des}} - \sigma a, v_{\text{des}}\right)\right), \end{aligned}$$

where a and b are the acceleration and deceleration of the car respectively, τ is the reaction time of the driver, and σ is the driver's imperfection in control. $g(t)$ is the distance between the car in consideration and the car it is interacting with. $v_{\max}$ is the maximum speed of the car and has been set as a constant $5\,\text{m/s}$. $rand(x_1, x_2)$ denotes a random number between x_1 and x_2. The reader is referred to (Gipps, 1981; Krauss et al., 1997) for details.

The transition function has 4 parameters that depends on the driving strategy of A : (1) the driver's imperfection in control σ, (2) the reaction time of the driver τ, (3) the acceleration of the car a, and (4) the deceleration of the car d. In practice, the agent R does not know what type of driver A it is facing. Therefore, it needs to learn the 4 parameters of the transition function and at the same time cross the intersection safely and efficiently.

Finally, we used SUMO to evaluate POMDP and handcrafted policies. SUMO is an well established open-source package for microscopic road traffic simulation (Behrisch et al., 2011). It also adopts the Gipps model as the car velocity updating rule.

References

Behrisch, M., Bieker, L., Erdmann, J., and Krajzewicz., D. Sumo – simulation of urban mobility: An overview. In *Int. Conf. on Advances in System Simulation*, pp. 63–68, 2011.

Gipps, P. G. A behavioural car following model for computer simulation. *Transportation Research B*, 15:105–111, 1981.

Krauss, S., Wagner, P., and Gawron, C. Metastable states in a microscopic model of traffic flow. *Physical Review E*, 55(304): 55–97, 1997.