12th Workshop on Planning, Perception and Navigation for Intelligent Vehicles (PPNIV)

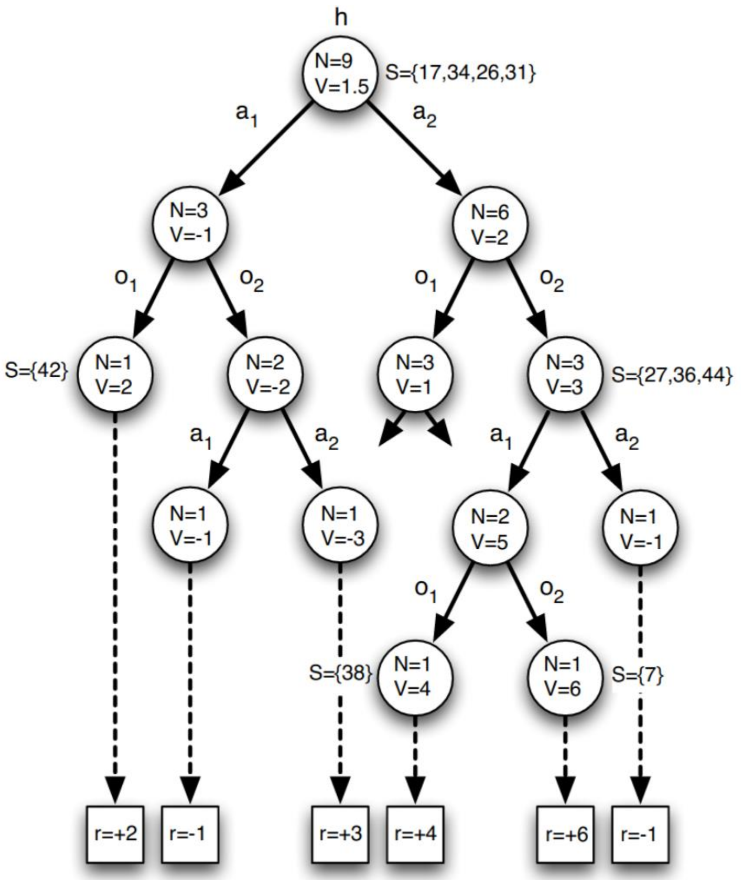
October 25th, 2020

Exploiting Continuity of Rewards: Efficient Sampling in POMDPs with Lipschitz Bandits

Ömer Sahin Tas, Felix Hauser and Martin Lauer

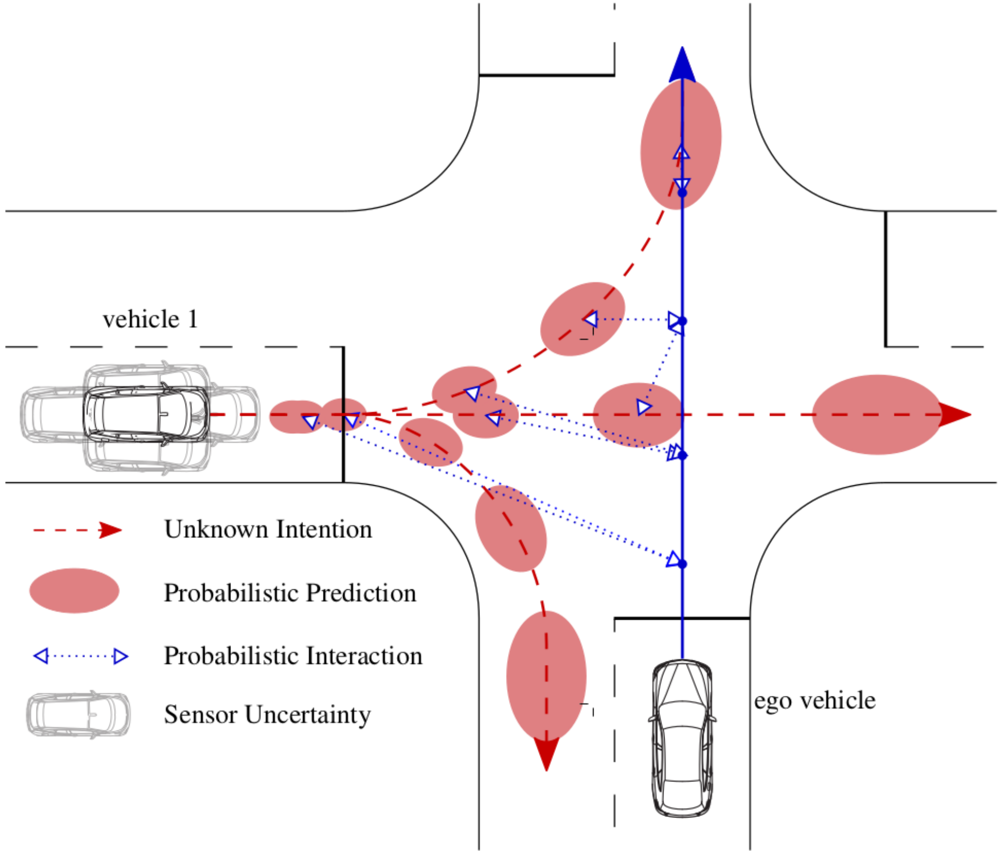
Karlsruhe Institute of Technology

POMDP Framework

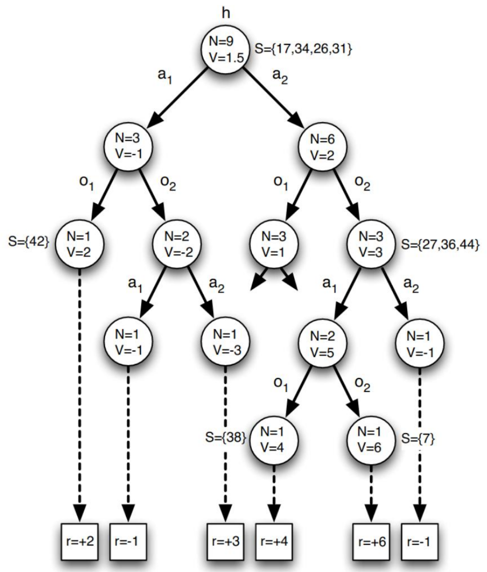


Silver & Veness, *Monte-Carlo Planning in Large POMDPs*, NIPS 2010.

Motion Planning in Automated Driving with the POMDP Framework



Hubmann et al., *Automated Driving in Uncertain Environments: Planning with Interaction and Uncertain Maneuver Prediction*, Transactions on Intelligent Vehicles 2018.



Silver & Veness, *Monte-Carlo Planning in Large POMDPs*, NIPS 2010.

Motion Planning in Automated Driving with the POMDP Framework

Map data

$$\rho = (p_i)_{i=1,\ldots,n} \qquad p_i = (x_i, y_i, l_i, \kappa_i, v_i)^\top$$

States, Observations, Actions

$$s = (s_0, s_1, s_2, \ldots, s_k) \qquad s_0 = (l_0, v_0) \quad s_k = (l_k, v_k, \rho_k)$$

$$o = (o_1, o_2, \ldots, o_k) \qquad o_k = (x_k, y_k, v_k)^\top$$

$$a \in [-3\,\mathrm{m\,s^{-2}}, 1\,\mathrm{m\,s^{-2}}]$$

Motion Planning in Automated Driving with the POMDP Framework

Transition Model

$$a_k = \max(a_{\mathrm{ref},k} + a_{\mathrm{int},k},\ a^{-}) + a_{\mathrm{noise},k}$$

Observation Model

$$(l, v, \rho) \rightarrow (x, y, v)$$

$$x_{\mathrm{noise}}, y_{\mathrm{noise}} \sim \mathcal{N}(0, \sigma^2_{o,\mathrm{pos}})$$

$$v_{\mathrm{noise}} \sim \mathcal{N}(0, \sigma^2_{o,\mathrm{vel}})$$

Motion Planning in Automated Driving with the POMDP Framework

Reward Model

$$r = r_{\mathrm{coll}} + r_v + r_{j,\mathrm{lon}} + r_{a,\mathrm{lat}}$$

$$r_{\mathrm{coll}} = \begin{cases} 0 & \text{no collision} \\ \zeta_{\mathrm{coll}} & \text{ego vehicle collides} \end{cases}$$

$$r_v = \begin{cases} \zeta_v \ (v_0 - v_{\mathrm{ref}})^2 & \text{if } v_0 \geq v_{\mathrm{ref}} \\ \zeta_v \ \log\left(1 + (v_0 - v_{\mathrm{ref}})^2\right) & \text{otherwise} \end{cases}$$

$$r_{j,\mathrm{lon}} = \zeta_{j,\mathrm{lon}} {j_0}^2$$

$$r_{a,\mathrm{lat}} = \zeta_{a,\mathrm{lat}} \left(\kappa \ v_0^2\right)^2$$

Multi-armed Bandits

$$\mathcal{A} = \{a_1, a_2, \ldots, a_K\}$$

Upper Confidence Bound (UCB)

$$b_t(a) = \hat{\mu}_t(a) + c\sqrt{\frac{2\log t}{n_t(a)}}$$

Algorithm 1: Upper Confidence Bound (UCB)

if $t \leq K$ **then**
  Choose arm from $\{a : n_t(a) = 0\}$ at random
else
  Choose arm $a_t = \arg\max_{a \in \mathcal{A}} b_t(a)$

UCB-V

$$b_t(a) = \hat{\mu}_t(a) + \sqrt{\frac{2\hat{\sigma}_t^2(a)\log t}{n_t(a)}} + \frac{3c\log t}{n_t(a)}$$

Multi-armed Bandits

Pareto Optimal Sampling for Lipschitz Bandits (POSLB)

$$|\mu(a) - \mu(a')| \leq \mathscr{L}\,|a - a'|$$

Algorithm 1: POSLB [23, p. 22]

if $t \leq T$ **then**
  Choose arm from $\{a : n_t(a) = 0\}$ at random
else
  $a_t^* = \arg\max_{a \in \mathcal{A}} \hat{\mu}_t(a)$
  $f_t(a) =$
$$\begin{cases} \sum_{a' \in \mathcal{A}} n_t(a')\big(\hat{\mu}_t(a') - \lambda_t(a,a')\big)^2 (2\sigma^2)^{-1} & \text{if } a \neq a_t^* \\ n_t(a)\big(\hat{\mu}_t(a) - b_t(a_t^*)\big)^2 (2\sigma^2)^{-1} & \text{if } a = a_t^* \end{cases}$$
  with
  $\lambda_t(a,a') = \max\big(b_t(a_t^*) - \mathscr{L}\,|a - a'|\,,\ \hat{\mu}_t(a')\big)$
  Choose arm $a_t = \arg\max_{a \in \mathcal{A}} \log t - f_t(a)$

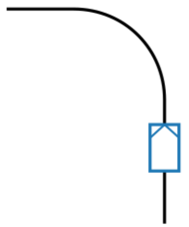
POSLB-V

$$\sigma_t^2(a) = \sigma^2 = \frac{n_t(a)}{2\log t}\left(\sqrt{\frac{2\hat{\sigma}_t^2(a)\log t}{n_t(a)}} + \frac{3c\log t}{n_t(a)}\right)^2$$

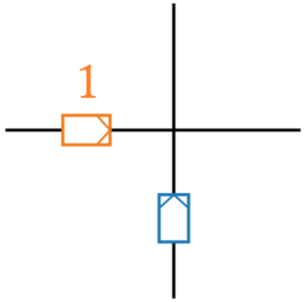
Evaluation

Scenarios

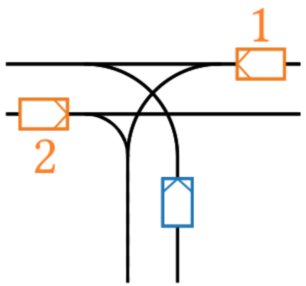
(a) Straight driving.



(b) Traversing curves.



(a) Collision scene.

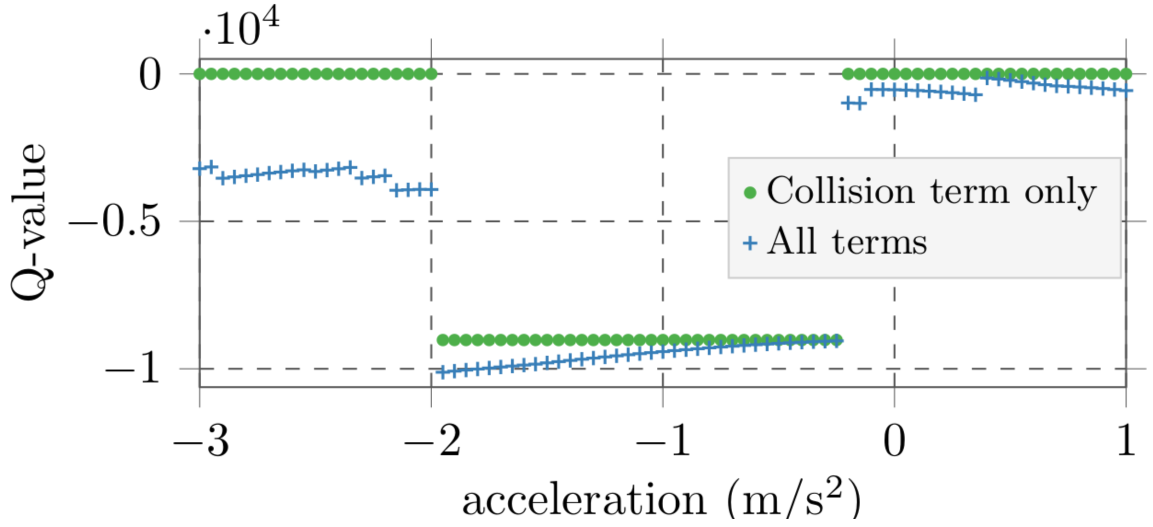


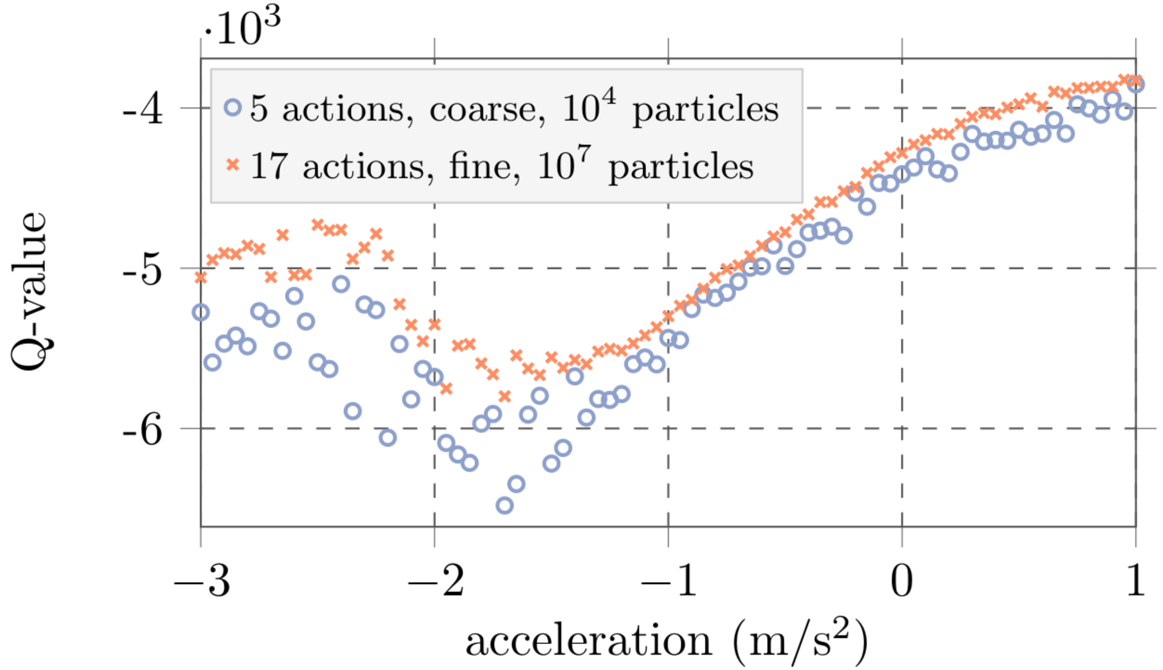
(b) Intersection scene.

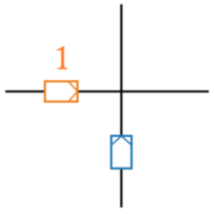
Scenario	Vehicle	Time-to-Intersection (s)			
$S_{\texttt{Coll}}$	ego	2.11			
	vehicle2	2.71			
$S_{\texttt{I-Lo}}$	ego	5.33	5.14	6.81	
	vehicle1	3.99	4.20	4.78	
	vehicle2	6.35	6.14	7.89	7.73
$S_{\texttt{I-Hi}}$	ego	2.66	2.28	5.63	
	vehicle1	2.78	3.23	4.52	
	vehicle2	3.42	3.05	6.21	5.92

Evaluation

Q-value



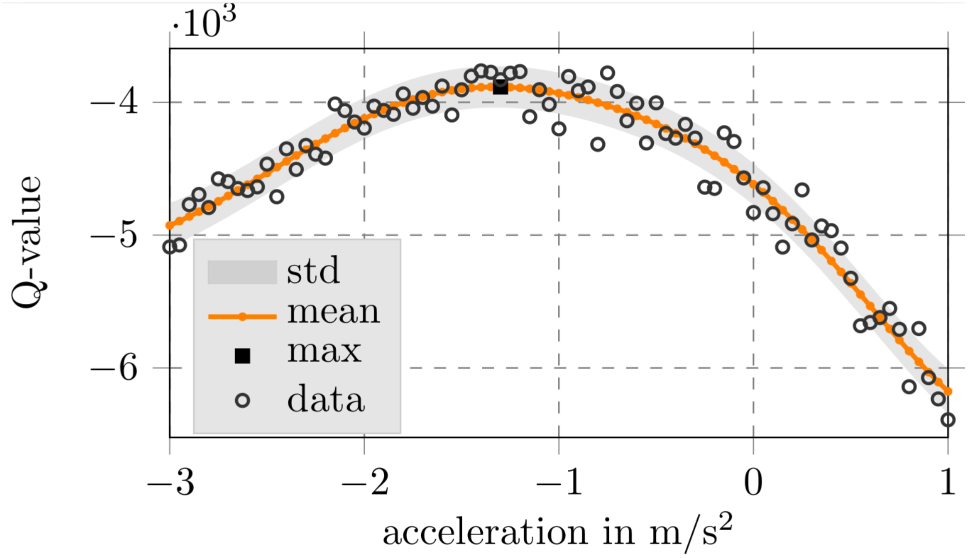




(a) Collision scene.

Evaluation

Convergence



$\|\mathcal{A}\|$	Straight	Curve	$S_{\texttt{Coll}}$	$S_{\texttt{I-Lo}}$	$S_{\texttt{I-Hi}}$
5	0.0	−1.0	1.0	−1.0	−1.0
9	0.0	−1.5	1.0	−1.0	−1.5
17	0.0	−1.5	1.0	−1.0	−1.25
33	0.0	−1.0	0.875	−0.875	−1.125

$\|\mathcal{A}\|$	Straight	Curve	$S_{\texttt{Coll}}$	$S_{\texttt{I-Lo}}$	$S_{\texttt{I-Hi}}$
5	1247	1847	1003	1241	573
9	1271	2157	1981	1345	728
17	1336	2280	2742	1453	787
33	1370	2246	1260	1432	1033

Evaluation

Convergence

$$\text{MAE}_n = \frac{1}{m} \sum_{i=0}^{m-1} \left| a^*_{i,n} - a^* \right|$$

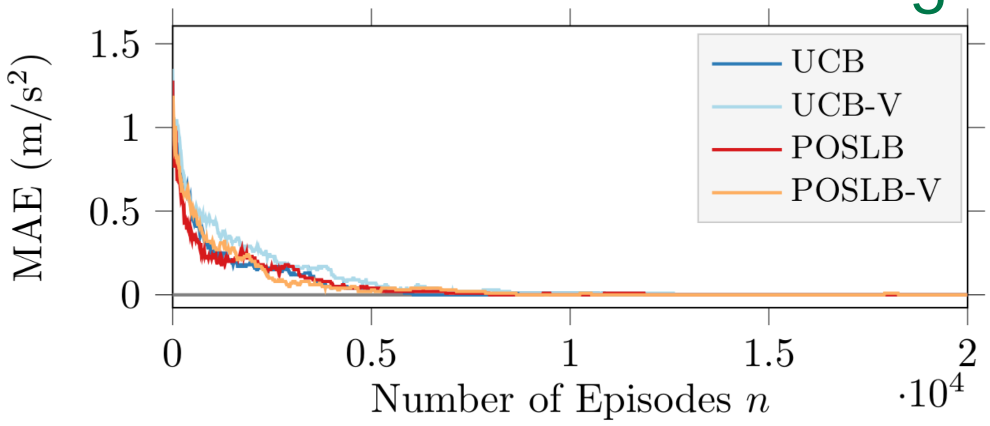
$$a^*_n = \arg\max_{\mathrm{a} \in \mathcal{A}} Q_n(h_0, a)$$

$\lvert\mathcal{A}\rvert$	Straight	Curve	$\mathtt{S}_{\mathtt{Coll}}$	$\mathtt{S}_{\mathtt{I-Lo}}$	$\mathtt{S}_{\mathtt{I-Hi}}$
5	0.0	−1.0	1.0	−1.0	−1.0
9	0.0	−1.5	1.0	−1.0	−1.5
17	0.0	−1.5	1.0	−1.0	−1.25
33	0.0	−1.0	0.875	−0.875	−1.125

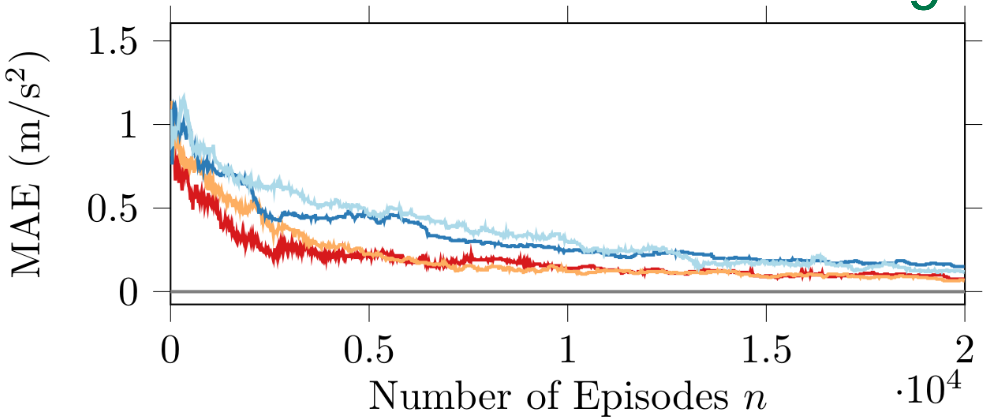
Evaluation

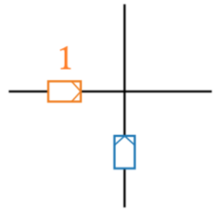
Convergence

5



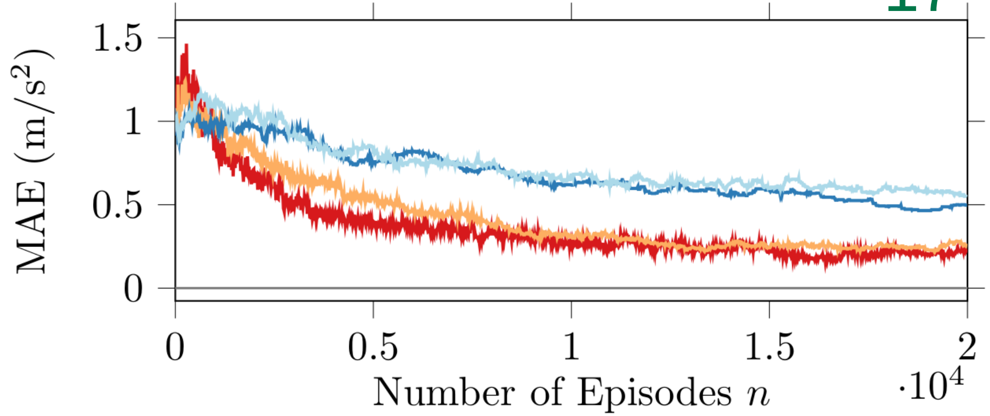
9



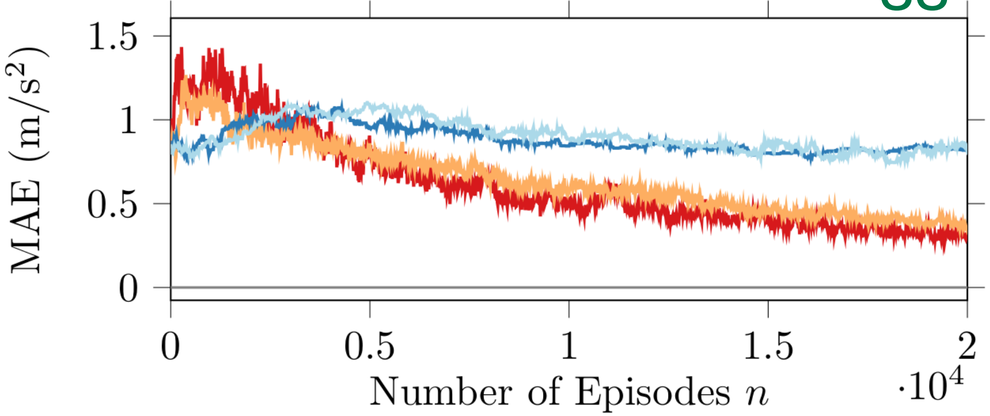


(a) Collision scene.

17

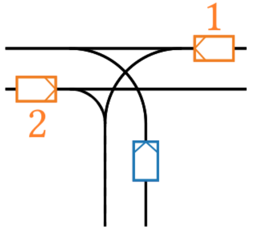


33



Evaluation

Convergence

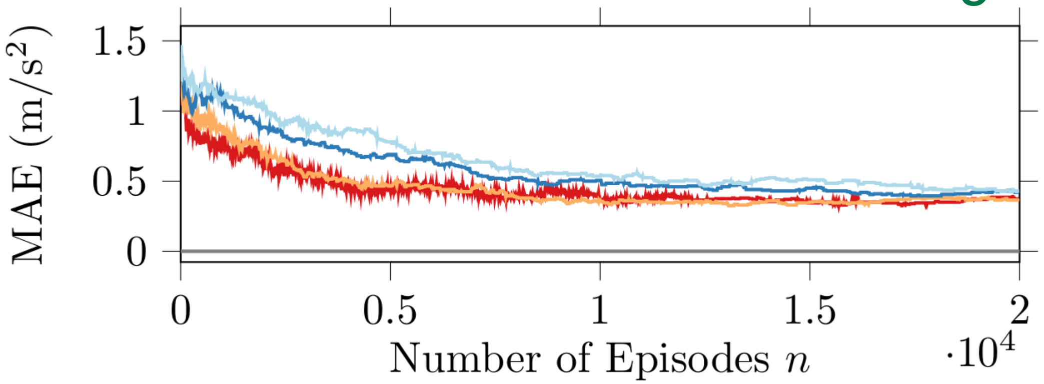


(b) Intersection scene.

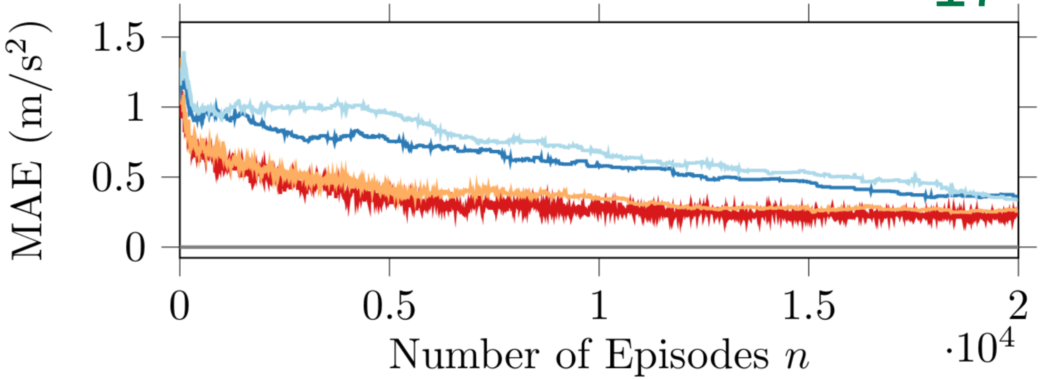
$S_{\text{I-Hi}}$

UCB
UCB-V
POSLB
POSLB-V

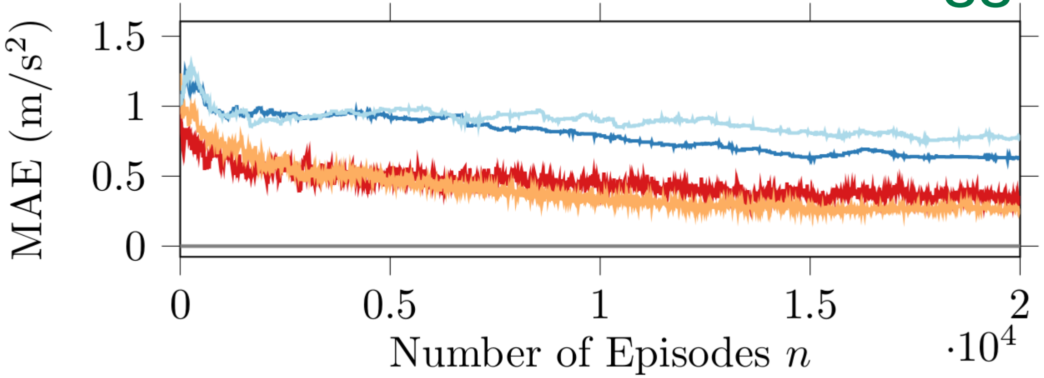
9



17

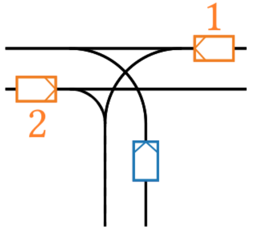


33



Evaluation

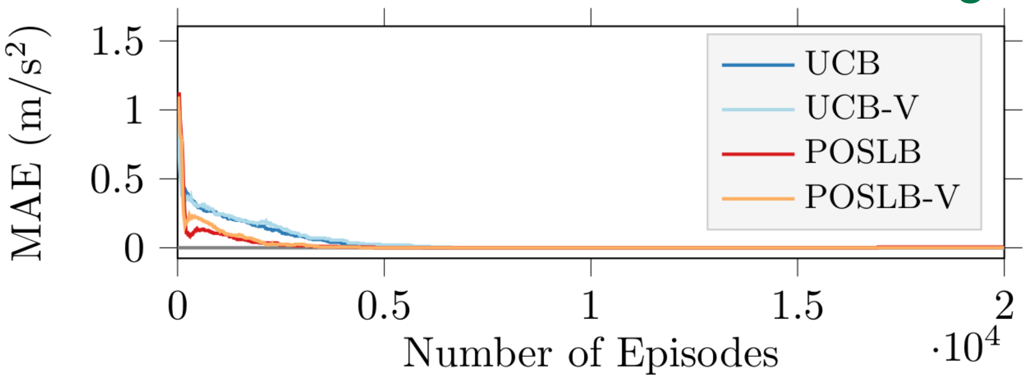
Convergence



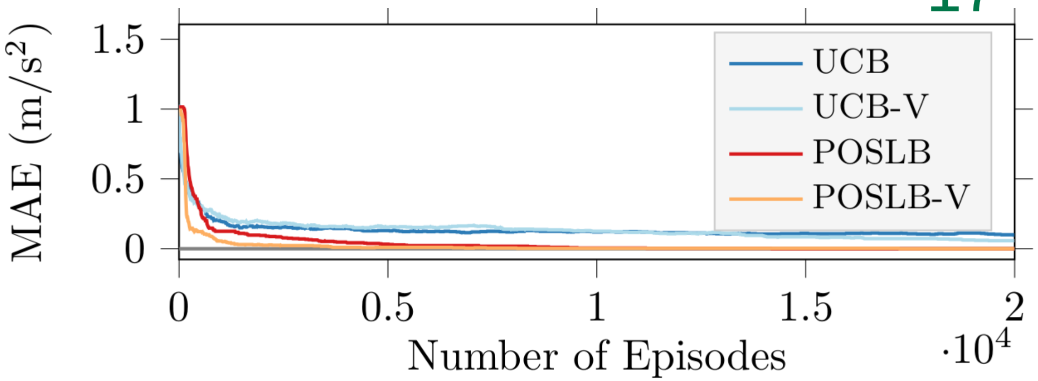
(b) Intersection scene.

$S_{\text{I-Lo}}$

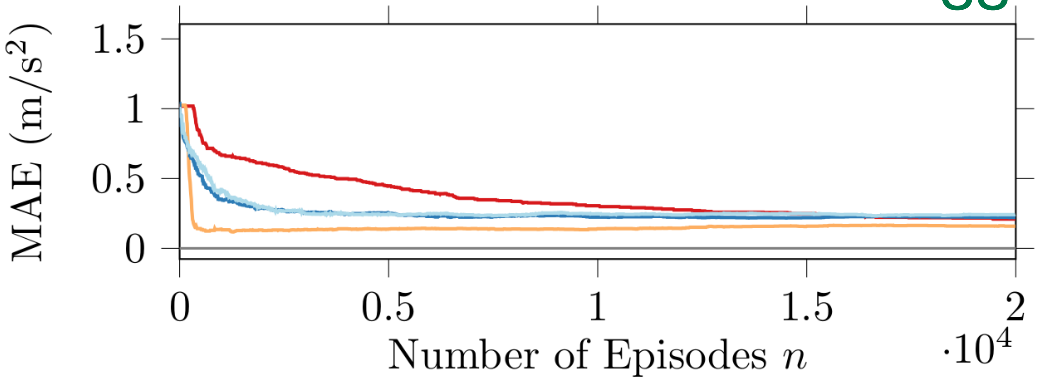
9



17

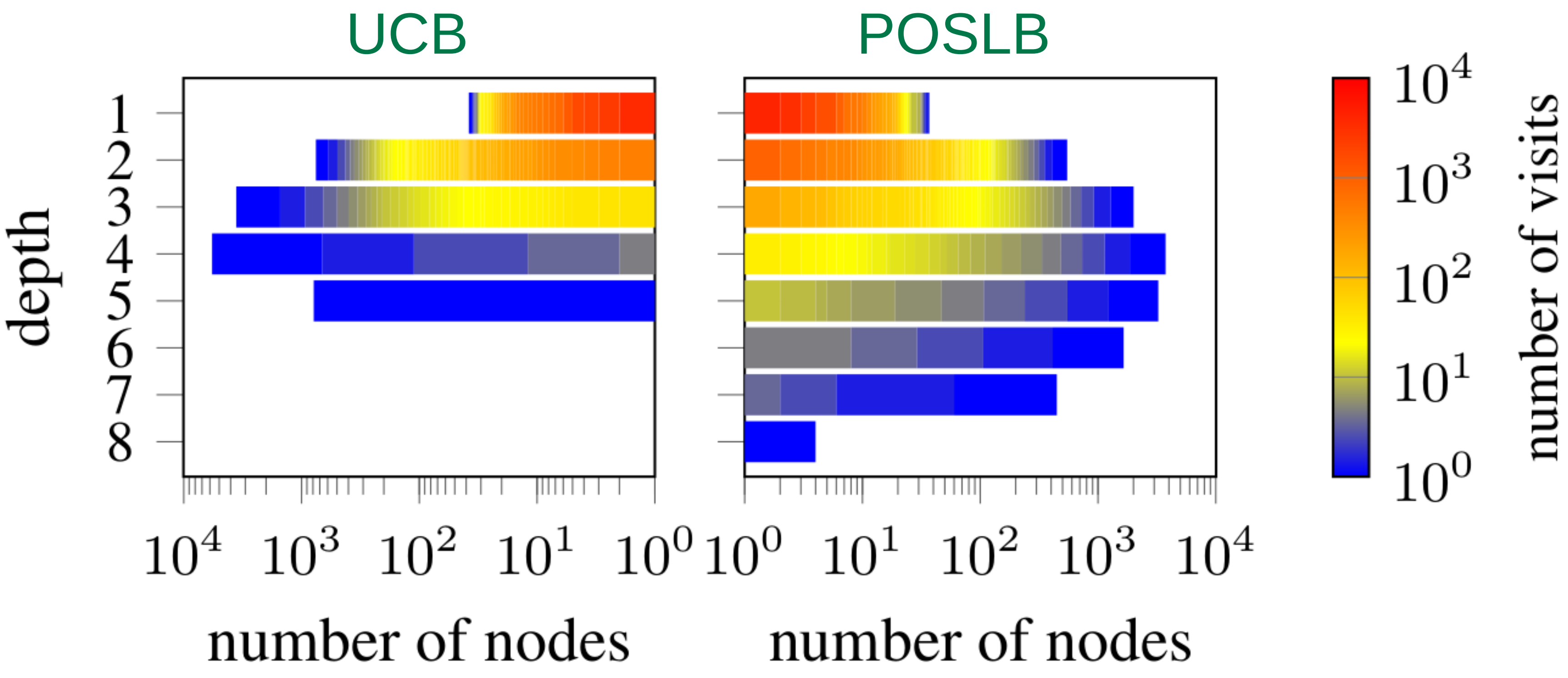


33



Evaluation

Tree Depth



Conclusion

- Uncertainties in the transition and observation model have a smoothing effect on the discontinuities
- Utilizing the continuity of Q-values allows significant performance improvements
- POSLB-V bandit algorithm

➡ This work enables the use of POMDPs for problems where multiple actions need to be considered, such as in motion planning.

Thanks!

FZI Research Center for Information Technology

Ömer Sahin Tas
Department on Mobile Perception Systems

Haid-und-Neu-Str. 10-14
76131 Karlsruhe

E-Mail: tas@fzi.de