Supplemental Material for ICML 2020 Submission no. 2113 Information Particle Filter Tree: An Online Algorithm for POMDPs with Belief-Based Rewards on Continuous Domains

A. Multivariate Kernel Density Estimation

This section provides more details about kernel density estimation (KDE) and bandwidth selection based on (Silverman, 1986; Gisbert, 2003). We consider density estimation in a D-dimensional continuous state space. For a weighted particle set $\{(s_i, w_i)\}_{i=1}^m$ with normalized weights $\sum_{i=1}^m w_i = 1$ and particles $s_i \in \mathbb{R}^D$ the general KDE is given by

$$\hat{p}(s) = \sum_{i=1}^{m} \frac{w_i}{\sqrt{\det H}} K\left(H^{-\frac{1}{2}}(s - s_i)\right) \tag{1}$$

where K is the kernel function and $H \in \mathbb{R}^{D\times D}$ is the symmetric, positive definite bandwidth matrix. Since the resulting density estimate is not very sensitive to the choice of kernel function, we consider only the multivariate normal kernel

$$K(s) = \frac{1}{\sqrt{(2\pi)^D}} \exp\left(-\frac{1}{2} s^T s\right), \tag{2}$$

which results in the density estimate

$$\hat{p}(s) = \sum_{i=1}^{m} v_i \exp\left(-\frac{1}{2}(s - s_i)^T H^{-1}(s - s_i)\right) \tag{3}$$

with coefficients $v_i = \frac{w_i}{\sqrt{(2\pi)^D \det H}}$ for $i = 1, \dots, m$.

Bandwidth Selection More important is the choice of the bandwidth matrix H, as it defines the amount of smoothing applied to the particles. A small bandwidth will result in a tall, narrow peak at each particle, while increasing the bandwidth results in a smoother function. However, if the bandwidth is chosen too large, features are lost due to the smoothing effect of an increased bandwidth. Frequently, diagonal bandwidth matrices are employed, which allows to set different smoothing factors for each dimension. Moreover, the orientation of the smoothing can be chosen by using general symmetric, positive definite matrices.

A widely-used heuristic for the bandwidth is Silverman's rule of thumb which is defined by

$$\sqrt{H_{ii}} = \left(\frac{4}{D+2}\right)^{\frac{1}{D+4}} m^{-\frac{1}{D+4}} \hat{\sigma}_i, \quad i = 1, \dots, D, \tag{4}$$

where $\hat{\sigma}_i$ is the empirical standard deviation with respect to the ith dimension and $H_{ij} = 0$ for $i \neq j$. In the univariate case ($D = 1$) this reduces to

$$h = \left(\frac{4}{3}\right)^{\frac{1}{5}} m^{-\frac{1}{5}} \hat{\sigma} \approx 1.06 m^{-\frac{1}{5}} \hat{\sigma}. \tag{5}$$

It can be shown that this bandwidth is optimal if the underlying density is Gaussian (Silverman, 1986, p. 45). However, if the true density is not Gaussian, this choice is likely to over-smooth the result.

Computational Complexity In the general case of an arbitrary symmetric, positive definite bandwidth matrix, the matrix inversion and the determinant computation have a complexity of $\mathcal{O}(D^3)$. For diagonal bandwidth matrices, as we use in this work, the complexity of these operations reduces to $\mathcal{O}(D)$. Therefore, evaluating the density estimate in Equation (3) has a complexity of $\mathcal{O}(mD)$ and the KDE-based entropy estimation, as presented in Section 4.1, has a complexity of $\mathcal{O}(m^2 D)$.

B. Evaluation Hyperparameters

The hyperparameters used by the algorithms in the evaluation are listed in Table 1. Since the considered problems all have a small discrete action space, progressive widening is not used on the action space. Hence, the parameters k_a, α_a are not required.

Sunberg & Kochenderfer optimized the parameters of their algorithms POMCPOW and PFT-DPW for the Light Dark and Laser Tag problem with the cross entropy method (Mannor et al., 2003). In our work, we use the same parameters for these problems. For the Continuous Light Dark (CLD) problem, the parameters were chosen identical to the Light Dark problem, since the problems are very similar.

Since MCTS algorithms tend to be very sensitive with respect to the exploration constant c, we conducted additional experiments to inspect the influence of c on the benchmark. This is particularly important to ensure that the results for POMCPOW and PFT-DPW do not suffer from a suboptimal choice of the exploration constant. Figure 1 shows the results of 1000 simulations with varying parameter c

Table 1. Hyperparameters used in the Light Dark, Continuous Light Dark (CLD), and Laser Tag experiments

IPFT	Light Dark	CLD	Laser Tag
λ	50.0	60.0	4.0
m	20	20	20
c	100.0	100.0	26.0
k_o	5.0	5.0	4.0
α_o	$1/20$	$1/20$	$1/35$
POMCPOW	Light Dark	CLD	Laser Tag
c	90.0	90.0	26.0
k_o	5.0	5.0	4.0
α_o	$1/15$	$1/15$	$1/35$
PFT-DPW	Light Dark	CLD	Laser Tag
m	20	20	20
c	100.0	100.0	26.0
k_o	4.0	4.0	4.0
α_o	$1/10$	$1/10$	$1/35$

in the CLD problem with action spaces $\mathbb{A}_{10}$ and $\mathbb{A}_3$. The figure shows that the choice of exploration constant does not influence the results significantly.

The parameters for IPFT were selected by running 1000 simulations with different parameter sets and choosing the best parameters. Since IPFT is based on PFT-DPW, the PFT-DPW parameters served as a starting point for this procedure. An initial guess for the information weight λ was determined by scaling the information gathering term such that it has the same order of magnitude as the expected reward.

In general, the values α_o for the observation widening are quite small. This essentially results in a limited number of child nodes and thereby allows the tree to grow deeper. Sunberg & Kochenderfer note that it might be sufficient to simply limit the number of child nodes to a fixed number (2017).

References

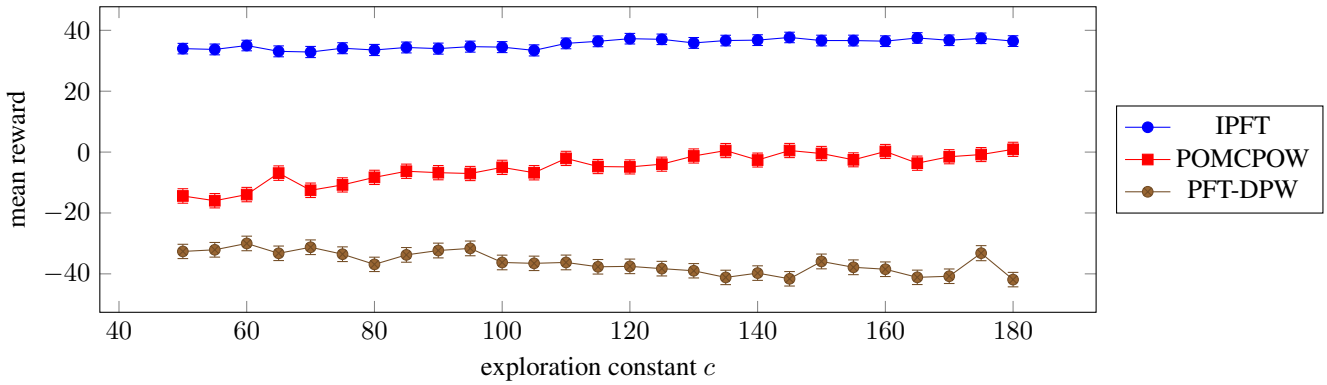
Gisbert, F. J. G. Weighted samples, kernel density estimators and convergence. *Empirical Economics*, 28(2):335–351, 2003.

Mannor, S., Rubinstein, R., and Gat, Y. The cross entropy method for fast policy search. In *Proceedings of the International Conference on International Conference on Machine Learning*, pp. 512–519. AAAI Press, 2003.

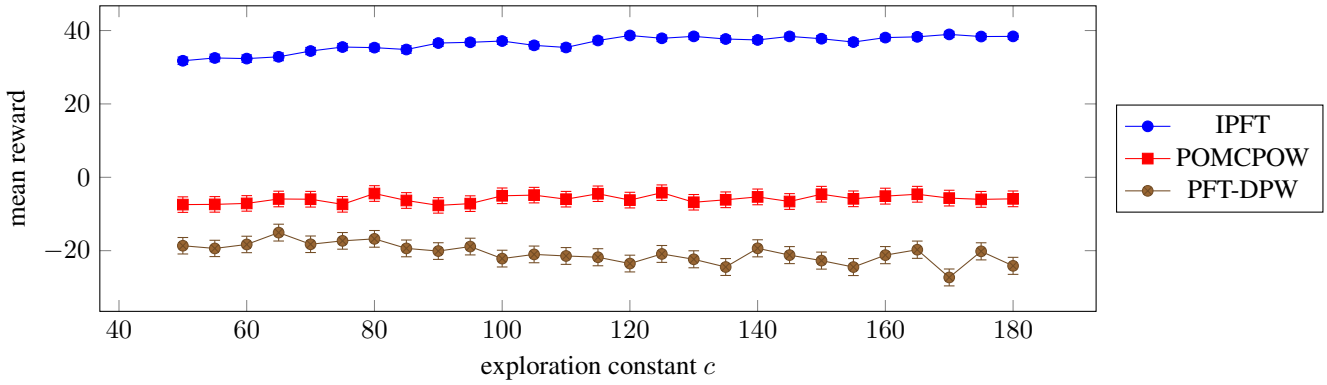
Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

Sunberg, Z. and Kochenderfer, M. Online algorithms for POMDPs with continuous state, action, and observation spaces (extended version). *CoRR*, abs/1709.06196, 2017.

Sunberg, Z. and Kochenderfer, M. Online Algorithms for POMDPs with Continuous State, Action, and Observation Spaces. In *Proceedings of the International Conference on Automated Planning and Scheduling*, pp. 259–263, 2018.



a) Action space $\mathbb{A}_{10}$.



b) Action space $\mathbb{A}_3$.

Figure 1. Sensitivity of the reward with respect to the exploration constant c in the Continuous Light Dark problem. The mean reward and its standard deviation over 1000 simulations are depicted.