A. Proof of Theorem 1

Proof. For the base case $t = H+1$, since $V_{\text{DR}}^0 = V(s_{H+1}) = 0$, it is obvious that at the $(H+1)$-th step the estimator is unbiased with 0 variance, and the theorem holds. For the inductive step, suppose the theorem holds for step $t+1$. At time step t, we have:

$$
\begin{aligned}
&\mathbb{V}_t\big[V_{\text{DR}}^{H+1-t}\big] \\
&= \mathbb{E}_t\Big[\big(V_{\text{DR}}^{H+1-t}\big)^2\Big] - \Big(\mathbb{E}_t\big[V(s_t)\big]\Big)^2 \\
&= \mathbb{E}_t\Big[\Big(\widehat{V}(s_t) + \rho_t\big(r_t + \gamma V_{\text{DR}}^{H-t} - \widehat{Q}(s_t,a_t)\big)\Big)^2 \\
&\quad - V(s_t)^2\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\Big(\rho_t Q(s_t,a_t) - \rho_t\widehat{Q}(s_t,a_t) + \widehat{V}(s_t) \\
&\quad + \rho_t\big(r_t + \gamma V_{\text{DR}}^{H-t} - Q(s_t,a_t)\big)\Big)^2 - V(s_t)^2\Big] \\
&\quad + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\Big(-\rho_t\Delta(s_t,a_t) + \widehat{V}(s_t) + \rho_t(r_t - R(s_t,a_t)) \\
&\quad + \rho_t\gamma\big(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]\big)\Big)^2 \\
&\quad - V(s_t)^2\Big] + \mathbb{V}_t\big[V(s_t)\big] \qquad (15) \\
&= \mathbb{E}_t\Big[\mathbb{E}_t\Big[\big(-\rho_t\Delta(s_t,a_t) + \widehat{V}(s_t)\big)^2 - V(s_t)^2 \;\Big|\; s_t\Big]\Big] \\
&\quad + \mathbb{E}_t\Big[\mathbb{E}_{t+1}\big[\rho_t^2(r_t - R(s_t,a_t))^2\big]\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&\quad + \mathbb{E}_t\Big[\mathbb{E}_{t+1}\Big[\Big(\rho_t\gamma\big(V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]\big)\Big)^2\Big]\Big] \\
&= \mathbb{E}_t\Big[\mathbb{V}_t\big[-\rho_t\Delta(s_t,a_t) + \widehat{V}(s_t) \mid s_t\big]\Big] + \mathbb{E}_t\Big[\rho_t^2\,\mathbb{V}_{t+1}\big[r_t\big]\Big] \\
&\quad + \mathbb{E}_t\Big[\rho_t^2\gamma^2\,\mathbb{V}\big[V_{\text{DR}}^{H-t} \mid s_t, a_t\big]\Big] + \mathbb{V}_t\big[V(s_t)\big] \\
&= \mathbb{E}_t\Big[\mathbb{V}_t\big[\rho_t\Delta(s_t,a_t) \mid s_t\big]\Big] + \mathbb{E}_t\Big[\rho_t^2\,\mathbb{V}_{t+1}\big[r_t\big]\Big] \\
&\quad + \mathbb{E}_t\Big[\rho_t^2\gamma^2\,\mathbb{V}_{t+1}\big[V_{\text{DR}}^{H-t}\big]\Big] + \mathbb{V}_t\big[V(s_t)\big].
\end{aligned}
$$

This completes the proof. Note that from Eqn.(15) to the next step, we have used the fact that conditioned on s_t and a_t, $r_t - R(s_t,a_t)$ and $V_{\text{DR}}^{H-t} - \mathbb{E}_{t+1}\big[V(s_{t+1})\big]$ are independent and have zero means, and all the other terms are constants. Therefore, the square of the sum equals the sum of squares in expectation. □

B. Bias of DR-v2

Proof of Proposition 1. Let $V_{\text{DR-v2'}}$ denote Eqn.(12) with approximation $\widehat{P} = P$. Since $V_{\text{DR-v2}}$ is unbiased, the bias of $V_{\text{DR-v2'}}$ is then the expectation of $V_{\text{DR-v2'}} - V_{\text{DR-v2}}$. Define

$$\beta_t = \mathbb{E}_t\Big[V_{\text{DR-v2'}}^{H+1-t} - V_{\text{DR-v2}}^{H+1-t}\Big].$$

Then, β_1 is the bias we try to quantify, and is a constant. In general, β_t is a random variable that depends on $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$. Now we have

$$
\begin{aligned}
\beta_t &= \mathbb{E}_t\Big[\rho_t\gamma\,\big(V_{\text{DR-v2'}}^{H-t} - V_{\text{DR-v2}}^{H-t}\big) \\
&\qquad - \rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)} - 1\right)\Big] \\
&= \mathbb{E}_t\Big[\rho_t\gamma\beta_{t+1}\Big] - \mathbb{E}_t\Big[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)} - 1\right)\Big].
\end{aligned}
$$

In the second term of the last expression, the expectation is taken over the randomness of a_t and s_{t+1}; we keep a_t as a random variable and integrate out s_{t+1}, and get

$$
\begin{aligned}
&\mathbb{E}_t\Big[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)} - 1\right)\Big] \\
&= \mathbb{E}_t\Big[\mathbb{E}_{t+1}\Big[\rho_t\gamma\widehat{V}(s_{t+1})\left(\frac{\widehat{P}(s_{t+1}|s_t,a_t)}{P(s_{t+1}|s_t,a_t)} - 1\right)\Big]\Big] \\
&= \mathbb{E}_t\Big[\rho_t\gamma\sum_{s'}P(s'|s_t,a_t)\widehat{V}(s')\left(\frac{\widehat{P}(s'|s_t,a_t)}{P(s'|s_t,a_t)} - 1\right)\Big] \\
&= \mathbb{E}_t\Big[\rho_t\gamma\sum_{s'}\widehat{V}(s')\Big(\widehat{P}(s'|s_t,a_t) - P(s'|s_t,a_t)\Big)\Big].
\end{aligned}
$$

Recall that the expectation of the importance ratio is always 1, hence

$$
\begin{aligned}
\beta_t &\le \mathbb{E}_t\Big[\rho_t\gamma\,(\beta_{t+1} + \epsilon V_{\max})\Big] \\
&= \mathbb{E}_t\Big[\rho_t\gamma\beta_{t+1}\Big] + \gamma\epsilon V_{\max}.
\end{aligned}
$$

With an abuse of notation, we reuse β_t as its maximal absolute magnitude over all sample paths $s_1, a_1, \ldots, s_{t-1}, a_{t-1}$. Clearly we have $\beta_{H+1} = 0$, and

$$\beta_t \le \gamma(\beta_{t+1} + \epsilon V_{\max}).$$

Hence, $\beta_1 \le \epsilon V_{\max}\sum_{t=1}^{H}\gamma^t$. □

C. Cramer-Rao bound for discrete DAG MDPs

Here, we prove a lower bound for the relaxed setting where the MDP is a layered Directed Acyclic Graph instead of a tree. In such MDPs, the regions of the state space reachable in different time steps are disjoint (just as tree MDPs), but trajectories that separate in early steps can reunion at a same state later.

Definition 2 (Discrete DAG MDP)**.** An MDP is a *discrete Directed Acyclic Graph (DAG) MDP* if:

- The state space and the action space are finite.
- For any $s \in S$, there exists a unique $t \in \mathbb{N}$ such that, $\max_{\pi: S\to A} P(s_t = s \mid \pi) > 0$. In other words, a state only occurs at a particular time step.
- As a simplification, we assume $\gamma = 1$, and non-zero rewards only occur at the end of each H-step long tra-

jectory. We use an additional state s_{H+1} to encode the reward randomness so that reward function $R(s_{H+1})$ is deterministic and the domain can be solely parameterized by transition probabilities.

Theorem 3. *For discrete DAG MDPs, the variance of any unbiased estimator is lower bounded by*

$$\sum_{t=1}^{H+1} \mathbb{E}\Big[\frac{P_1(s_{t-1},a_{t-1})^2}{P_0(s_{t-1},a_{t-1})^2}\mathbb{V}_t\big[V(s_t)\big]\Big],$$

where for trajectory τ,
$P_0(\tau) = \mu(s_1)\pi_0(a_1|s_1)P(s_2|s_1,a_1)\dots P(s_{H+1}|s_H,a_H)$, *and $P_0(s_t,a_t)$ is its marginal probability; $P_1(\cdot)$ is similarly defined for π_1.*

Remark Compared to Theorem 2, the cumulative importance ratio $\rho_{1:t-1}$ is replaced by the state-action occupancy ratio $P_1(s_{t-1},a_{t-1})/P_0(s_{t-1},a_{t-1})$ in Theorem 3. The two ratios are equal when each state can only be reached by a unique sample path. In general, however, $\mathbb{E}\big[P_1(s_{t-1},a_{t-1})^2/P_0(s_{t-1},a_{t-1})^2\mathbb{V}_t\big[V(s_t)\big]\big] \le \mathbb{E}\big[\rho_{1:t-1}^2\mathbb{V}_t\big[V(s_t)\big]\big]$, hence DAG MDPs are easier than tree MDPs for off-policy value evaluation.

Below we give the proof of Theorem 3, which is almost identical to the proof of Theorem 2.

Proof of Theorem 3. We parameterize the MDP by $\mu(s_1)$ and $P(s_{t+1}|s_t,a_t)$ for $t = 1,\dots,H$. For convenience we will treat $\mu(s_1)$ as $P(s_1|\emptyset)$, so all the parameters can be represented as $P(s_{t+1}|s_t,a_t)$ (for $t=0$ there is a single s_0 and a). These parameters are subject to the normalization constraints that have to be taken into consideration in the Cramer-Rao bound, namely $\forall t,s_t,a_t,\sum_{s_{t+1}} P(s_{t+1}|s_t,a_t)=1$.

$$\begin{bmatrix} 1\cdots1 & & & \\ & 1\cdots1 & & \\ & & \ddots & \\ & & & 1\cdots1 \end{bmatrix}\theta = \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} \tag{16}$$

where $\theta_{s_t,a_t,s_{t+1}} = P(o|s_t,a_t)$. The matrix on the left is effectively the Jacobian of the constraints, which we denote as F. We index its rows by (s_t,a_t), so $F_{(s_t,a_t),(s_t,a_t,s_{t+1})} = 1$ and other entries are 0. Let U be a matrix whose column vectors consist an orthonormal basis for the null space of F. From Moore Jr (2010, Eqn. (3.3) and Corollary 3.10), we have the Constrained Cramer-Rao Bound (CCRB) being[2] (the dependence on θ in all terms are omitted):

$$KU(U^\top IU)^{-1}U^\top K^\top, \tag{17}$$

where I is the Fisher Information Matrix (FIM), and K is the Jacobian of the quantity we want to estimate; they are computed below. We start with I, which is

$$I = \mathbb{E}\Big[\Big(\frac{\partial \log P_0(\tau)}{\partial\theta}\Big)\Big(\frac{\partial \log P_0(\tau)}{\partial\theta}\Big)^\top\Big]. \tag{18}$$

To calculate I, we define a new notation $g(\tau)$, which is a vector of indicator functions and $g(\tau)_{s_t,a_t,s_{t+1}} = 1$ when (s_t,a_t,s_{t+1}) appears in trajectory τ. Using this notation, we have

$$\frac{\partial \log P_0(\tau)}{\partial\theta} = \theta^{\circ-1}\circ g(\tau), \tag{19}$$

where $\circ$ denotes element-wise power/multiplication. Then we can rewrite the FIM as

$$\begin{aligned} I &= \mathbb{E}\Big[[\theta_i^{-1}\theta_j^{-1}]_{ij}\circ(g(\tau)g(\tau)^\top)\Big] \\ &= [\theta_i^{-1}\theta_j^{-1}]_{ij}\circ\mathbb{E}\big[(g(\tau)g(\tau)^\top)\big], \end{aligned} \tag{20}$$

where $[\theta_i^{-1}\theta_j^{-1}]_{ij}$ is a matrix expressed by its (i,j)-th element. Now we compute $\mathbb{E}\big[g(\tau)g(\tau)^\top\big]$. On the diagonal, it is $P_0(s_t,a_t,s_{t+1})$, so the diagonal of I is $\frac{P_0(s_t,a_t)}{P(s_{t+1}|s_t,a_t)}$; for non-diagonal entries whose row indexing and column indexing tuples are at the same time step, the value is 0; in other cases, suppose row is (s_t,a_t,s_{t+1}) and column is $s_{t'},a_{t'},s_{t'+1}$, and without loss of generality assume $t'<t$, then the entry is $P_0(s_{t'},a_{t'},s_{t'+1},s_t,a_t,s_{t+1})$, with the corresponding entries in I being $\frac{P_0(s_{t'},a_{t'},s_{t'+1},s_t,a_t,s_{t+1})}{P(s_{t'+1}|s_{t'},a_{t'})P(s_{t+1}|s_t,a_t)} = P_0(s_{t'},a_{t'})P_0(s_t,a_t|s_{t'+1})$.

Then, we calculate $(U^\top IU)^{-1}$. To avoid the difficulty of taking inverse of this non-diagonal matrix, we apply the following trick to diagonalize I: note that for any matrix X with matching dimensions,

$$U^\top IU = U^\top(F^\top X^\top + I + XF)U, \tag{21}$$

because by definition U is orthogonal to F. We can design X so that $D = F^\top X^\top + I + XF$ is a diagonal matrix, and $D_{(s_t,a_t,s_{t+1}),(s_t,a_t,s_{t+1})} = I_{(s_t,a_t,s_{t+1}),(s_t,a_t,s_{t+1})} = \frac{P_0(s_t,a_t)}{P(s_{t+1}|s_t,a_t)}$. This is achieved by having XF eliminate all the non-diagonal entries of I in the upper triangle without touching anything on the diagonal or below, and by symmetry $F^\top X^\top$ will deal with the lower triangle. The particular X we take is $X_{(s_{t'},a_{t'},s_{t'+1}),(s_t,a_t)} = -P_0(s_{t'},a_{t'})P_0(s_t,a_t|s_{t'+1})\mathbb{I}(t'<t)$, and it is not hard to verify that this construction diagonalizes I.

With the diagonalization trick, we have $(U^\top IU)^{-1} = (U^\top DU)^{-1}$. Since CCRB is invariant to the choice of U, and we observe that the rows of F are orthogonal, we choose U as follows: let $n_{(s_t,a_t)}$ be the number of 1's in $F_{(s_t,a_t),(\cdot)}$, and $U_{(s_t,a_t)}$ be the $n_{(s_t,a_t)}\times(n_{(s_t,a_t)}-1)$ matrix

[2] In fact, existing literature on Contrained Cramer-Rao Bound does not deal with the situation where the unconstrained parameters break the normalization constraints (which we are facing). However, this can be easily tackled by changing the model slightly to $P(o|h,a) = \theta_{hao}/\sum_{o'}\theta_{hao'}$, which resolves the issue and gives the same result.

with orthonormal columns in the null space of $\begin{bmatrix}1 & \dots & 1\end{bmatrix}$ ($n_{(s_t,a_t)}$ 1's); finally, we choose U to be a block diagonal matrix $U = \mathrm{diag}(\{U_{(s_t,a_t)}\})$, where $U_{(s_t,a_t)}$'s are the diagonal blocks, and it is easy to verify that U is column orthonormal and $FU = 0$. Similarly, we write $D = \mathrm{diag}(\{D_{(s_t,a_t)}\})$ where $D_{(s_t,a_t)}$ is a diagonal matrix with $(D_{(s_t,a_t)})_{s_{t+1},s_{t+1}} = P_0(s_t,a_t)/P(s_{t+1}|s_t,a_t)$, and

$$\begin{aligned}
&U(U^\top IU)^{-1}U^\top = U(U^\top DU)^{-1}U^\top \\
&= U(\mathrm{diag}(\{U_{(s_t,a_t)}^\top\})\mathrm{diag}(\{D_{(s_t,a_t)}\})\mathrm{diag}(\{U_{(s_t,a_t)}\}))^{-1}U \\
&= U\mathrm{diag}(\{\big(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)}\big)^{-1}\})U \\
&= \mathrm{diag}(\{U_{(s_t,a_t)}\big(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)}\big)^{-1}U_{(s_t,a_t)}^\top\}).
\end{aligned} \tag{22}$$

Notice that each block in Eqn.(22) is simply $1/P_0(s_t,a_t)$ times the CCRB of a multinomial distribution $P(\cdot|s_t,a_t)$. The CCRB of a multinomial distribution p can be easily computed by an alternative formula (Moore Jr, 2010, Eqn. (3.12))), which gives $\mathrm{diag}(p) - pp^\top$, so we have,

$$\begin{aligned}
&U_{(s_t,a_t)}\big(U_{(s_t,a_t)}^\top D_{(s_t,a_t)}U_{(s_t,a_t)}\big)^{-1}U_{(s_t,a_t)}^\top \\
&= \frac{\mathrm{diag}(P(\cdot|s_t,a_t)) - P(\cdot|s_t,a_t)P(\cdot|s_t,a_t)^\top}{P_0(s_t,a_t)}.
\end{aligned} \tag{23}$$

We then calculate K. Recall that we want to estimate

$$\begin{aligned}
v = v^{\pi_1,H} = \sum_{s_1}\mu(s_1)\sum_{a_1}\pi_1(a_1|s_1)\dots \\
\sum_{s_{H+1}}P(s_{H+1}|s_H,a_H)R(s_{H+1}),
\end{aligned} \tag{24}$$

and its Jacobian is $K = (\partial v/\partial\theta_t)^\top$, with $K_{(s_t,a_t,s_{t+1})} = P_1(s_t,a_t)V(s_{t+1})$, where $P_1(\tau) = \mu(s_1)\pi_1(a_1)\dots P(s_{H+1}|s_H,a_H)$ and $P_1(s_t,a_t)$ is the marginal probability.

Finally, putting all the pieces together, we have Eqn.(17) equal to

$$\begin{aligned}
&\sum_{s_t,a}\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)}\Big(\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)V(s_{t+1})^2 \\
&\qquad -\big(\sum_{s_{t+1}}P(s_{t+1}|s_t,a_t)V(s_{t+1})\big)^2\Big) \\
&= \sum_{t=0}^{H}\sum_{s_t}P_0(s_t,a_t)\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)^2}\mathbb{V}\big[V(s_{t+1})\;\big|\;s_t,a\big] \\
&= \sum_{t=0}^{H}\mathbb{E}\Big[\frac{P_1(s_t,a_t)^2}{P_0(s_t,a_t)^2}\mathbb{V}_{t+1}\big[V(s_{t+1})\big]\Big] \\
&= \sum_{t=1}^{H+1}\mathbb{E}\Big[\frac{P_1(s_{t-1},a_{t-1})^2}{P_0(s_{t-1},a_{t-1})^2}\mathbb{V}_t\big[V(s_t)\big]\Big].
\end{aligned}$$

□

D. Experiment Details

Here, we provide full details on the experiments that are omitted in the main paper due to space limit.

D.1. Mountain Car

Domain Description Mountain car is a widely used benchmark problem for RL with a 2-dimensional continuous state space (position and velocity) and deterministic dynamics (Singh & Sutton, 1996). The state space is $[-1.2, 0.6] \times [-0.07, 0.07]$, and there are 3 discrete actions. The agent receives -1 reward every time step with a discount factor 0.99, and an episode terminates when the first dimension of state reaches the right boundary. The initial state distribution is set to uniformly random, and behavior policy is uniformly random over the 3 actions. The typical horizon for this problem is 400, which can be too large for IS and its variants, therefore we accelerate the dynamics such that given (s,a), the next state s' is obtained by calling the original transition function 4 times holding a fixed, and we set the horizon to 100. A similar modification was taken by Thomas (2015), where every 20 steps are compressed as one step.

Model Construction The model we construct for this domain uses a simple discretization (state aggregation): the two state variables are multiplied by 2^6 and 2^8 respectively and the rounded integers are treated as the abstract state. We then estimate the model parameters from data using a tabular approach. Unseen aggregated state-action pairs are assumed to have reward $R_{\min} = -1$ and a self-loop transition. Both the models that produces π_{train} and that used for off-policy evaluation are constructed in the same way.

Data sizes & other details The dataset sizes are $|D_{\text{train}}| = 2000$ and $|D_{\text{eval}}| = 5000$. We split D_{eval} such that $D_{\text{test}} \in \{10, 100, 1000, 2000, 3000, 4000, 4900, 4990\}$. DR-bsl uses the step-dependent constant function

$$\widehat{Q}(s_t,a_t) = \frac{R_{\min}(1-\gamma^{H-t+1})}{1-\gamma}.$$

Since the estimators in the IS family typically has a highly skewed distribution, the estimates can occasionally go largely out of range, and we crop such outliers in $[V_{\min}, V_{\max}]$ to ensure that we can get statistically significant experiment results within a reasonable number of simulations. The same treatment is also applied to the experiment on Sailing.

D.2. Sailing

Domain Description The sailing domain (Kocsis & Szepesvári, 2006) is a stochastic shortest-path problem, where the agent sails on a grid (in our experiment, a map of size 10×10) with wind blowing in random directions,

aiming at the terminal location on the top-right corner. The state is represented by 4 integer variables, representing either location or direction. At each step, the agent chooses to move in one of the 8 directions, (moving against the wind or running off the grid is prohibited), and receives a negative reward that depends on moving direction, wind direction, and other factors, ranging from $R_{\min} = -3-4\sqrt{2}$ to $R_{\max} = 0$ (absorbing). The problem is non-discounting, and we use $\gamma = 0.99$ for easy convergence when computing π_{train}.

Model Construction We apply Kernel-based Reinforcement Learning (Ormoneit & Sen, 2002) and supply a smoothing kernel in the joint space of states and actions. The kernel we use takes the form $\exp(-\|\cdot\|/b)$, where $\|\cdot\|$ is the ℓ_2-distance in $S \times A$,[3] and b is the kernel bandwidth, set to 0.25.

Data sizes & other details The data sizes are $|D_{\text{train}}| = 1000$ and $|D_{\text{eval}}| = 2500$, and we split D_{eval} such that $D_{\text{test}} \in \{5, 50, 500, 1000, 1500, 2000, 2450, 2495\}$. DR-bsl uses the step-dependent constant function

$$\widehat{Q}(s_t, a_t) = \frac{R_{\min}}{2} \frac{1-\gamma^{H-t+1}}{1-\gamma},$$

for the reason that in Sail $R_{\min}$ is rarely reached hence too pessimistic as a rough estimate of the magnitude of reward obtained per step.

D.3. KDD Cup 1998 Donation Dataset

Here are further details for experiments with the KDD donation dataset:

1. The size of dataset generated from the simulator for off-policy evaluation is equal to that of the true dataset (the one we use to fit the simulator at the very beginning; there are 3754 trajectories in that dataset).

2. The policy π_{train} is generated by training a recurrent neural network on the original data to fit a Q-value function (Li et al., 2015b).

3. Since there are many possible next-states for each state-action pair, for computational efficiency we use a sparse-sample approach when estimating $\widehat{Q}$ using the fitted model $\widehat{M}$: for each (s, a), we randomly sample several next-states from $\widehat{P}(\cdot|s, a)$, and cache them as a particle representation for the next-state distribution. The number of particles is set to 5 which is enough to ensure high accuracy.

[3] The difference of two directions is defined as the angle between them (in degrees) divided by 45°. For computational efficiency, the kernel function is cropped to 0 whenever two state-action pairs deviate more than 1 in any of the dimensions.