Supplementary Materials: State Relevance for Off-Policy Evaluation

A. Lemma

Given any mapping $\theta' : \mathcal{S} \to \{0, 1\}$,

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\rho^{\mathtt{C}}_{\theta'}(\tau^{(1)})] = 1 \tag{1}$$

Proof. From the definition of the OSIRIS weight,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\rho^{\mathtt{C}}_{\theta'}(\tau^{(1)})] &= \mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}\Big[\prod_{t=1}^{T}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]^{1-\theta'(s_t)}\Big] && (2)\\
&= \int\Big[\prod_{t=1}^{T}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]^{1-\theta'(s_t)}\pi_b(a_t\,|\,s_t)P(s_{t+1}\,|\,s_t,a_t)P(s_1)\Big]\,d\tau && (3)\\
&= \int\Big[\prod_{t=1}^{T}\pi_e(a_t\,|\,s_t)^{1-\theta'(s_t)}\pi_b(a_t\,|\,s_t)^{\theta'(s_t)}P(s_{t+1}\,|\,s_t,a_t)P(s_1)\Big]\,d\tau && (4)\\
&= 1 && (5)
\end{aligned}$$

where we use the joint distribution of trajectories generated from π_b (3) and cancel terms (4). Finally, $\theta'(s_t) \in \{0, 1\}$ selects either π_e or π_b as the integrand, both of which integrate to 1 by definition, so the full joint distribution also integrates to 1 (5). □

Notice that if we did not assume θ' is given, then we should assume that θ' is calculated using $\mathcal{D} \setminus \tau^{(1)}$. Otherwise, θ' would have statistical dependence with the transitions in trajectory $\tau^{(1)}$, which could not be resolved in (3).

B. Derivations of Sources of IS Variance

B.1. Variance due to Long Trajectory Length

Proposition 1. *For any subsets $\mathcal{T}_1 \subsetneq \mathcal{T}_2 \subseteq \{1, \ldots, T\}$, if $\pi_e \neq \pi_b$, and $\rho_1(\tau), \ldots, \rho_T(\tau)$ are mutually independent, then*

$$\mathop{\mathrm{Var}}_{\tau\sim\pi_b}\Big[\prod_{t\in\mathcal{T}_1}\rho_t(\tau)\,|\,s_1,\ldots,s_T\Big] < \mathop{\mathrm{Var}}_{\tau\sim\pi_b}\Big[\prod_{t\in\mathcal{T}_2}\rho_t(\tau)\,|\,s_1,\ldots,s_T\Big] \tag{6}$$

Proof. We first consider the variance of the product of general mutually independent random variables $X_1, \ldots, X_N$:

$$\begin{aligned}
\mathrm{Var}\Big[\prod_{n=1}^{N}X_n\Big] &= \mathbb{E}\Big[\prod_{n=1}^{N}X_n^2\Big] - \mathbb{E}\Big[\prod_{n=1}^{N}X_n\Big]^2 && (7)\\
&= \prod_{n=1}^{N}\mathbb{E}\left[X_n^2\right] - \prod_{n=1}^{N}\mathbb{E}\left[X_n\right]^2 && (8)\\
&= \prod_{n=1}^{N}\Big(\mathbb{E}\left[X_n^2\right] - \mathbb{E}\left[X_n\right]^2 + \mathbb{E}\left[X_n\right]^2\Big) - \prod_{n=1}^{N}\mathbb{E}\left[X_n\right]^2 && (9)\\
&= \prod_{n=1}^{N}\Big(\mathrm{Var}\left[X_n\right] + \mathbb{E}\left[X_n\right]^2\Big) - \prod_{n=1}^{N}\mathbb{E}\left[X_n\right]^2 && (10)
\end{aligned}$$

where we use the definition of variance (7), use the assumption that the variables are independent (8), introduce the same term (9), and reuse the definition of variance (10).

Using this fact, the inequality is equivalent to

$$\begin{aligned}&\prod_{t\in\mathcal{T}_1}\Big(\operatorname*{Var}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]+\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]^2\Big)-\prod_{t\in\mathcal{T}_1}\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]^2\\ &<\prod_{t\in\mathcal{T}_2}\Big(\operatorname*{Var}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]+\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]^2\Big)-\prod_{t\in\mathcal{T}_2}\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]^2 \qquad (11)\end{aligned}$$

We apply Equation 1:

$$\prod_{t\in\mathcal{T}_1}\Big(\operatorname*{Var}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]+1\Big)-1<\prod_{t\in\mathcal{T}_2}\Big(\operatorname*{Var}_{\tau\sim\pi_b}\left[\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\,|\,s_1,\ldots,s_T\right]+1\Big)-1 \qquad (12)$$

In general, variance is greater than or equal to zero; here the variance is strictly greater than zero because $\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}$is not constant because we assume $\pi_e \neq \pi_b$. Altogether, because LHS is the product of fewer variances > 0, we have that LHS $<$ RHS. □

B.2. Relationship between Trajectory Length and IS Weight

Proposition 2. *For any subsets $\mathcal{T}_1 \subsetneq \mathcal{T}_2 \subseteq \{1,\ldots,T\}$, if $\pi_e \neq \pi_b$, then*

$$\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\log\prod_{t\in\mathcal{T}_1}\rho_t(\tau)\Big]>\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\log\prod_{t\in\mathcal{T}_2}\rho_t(\tau)\Big] \qquad (13)$$

Proof. Using the identity for the log of a product and linearity of expectation, the inequality is equivalent to

$$\sum_{t\in\mathcal{T}_1}\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\log\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]>\sum_{t\in\mathcal{T}_2}\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\log\tfrac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big] \qquad (14)$$

By Jensen's inequality, for each individual expectation in the sum, $\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\log\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]\leq\log\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]=0$ where we also use Equation 1 ($\operatorname*{\mathbb{E}}_{\tau\sim\pi_b}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]=1$). Equality holds only if $\frac{\pi_e(a\,|\,s)}{\pi_b(a\,|\,s)}$ is constant, which is not true because we assume $\pi_e \neq \pi_b$. Thus, because the LHS is the sum of fewer expectations < 0, we have the strict inequality that LHS $>$ RHS.

□

C. Derivations for the General OSIRIS Estimator

C.1. Variance

Theorem 1. *Given any mapping $\theta' : \mathcal{S} \to \{0,1\}$, the variance of the OSIRIS estimator is:*

$$\begin{aligned}\operatorname*{Var}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta')]&=\operatorname*{Var}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{IS}}(\mathcal{D})]\\ &+\frac{1}{|\mathcal{D}|}\operatorname*{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2-\frac{1}{|\mathcal{D}|}\operatorname*{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')]^2 &(15a)\\ &-\frac{1}{|\mathcal{D}|}\operatorname*{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')^2]\operatorname*{Var}_{\mathcal{D}\sim\pi_b}[\rho^{\complement}_{\theta'}(\tau^{(1)})] &(15b)\\ &-\frac{1}{|\mathcal{D}|}\operatorname*{Cov}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')^2,\,\rho^{\complement}_{\theta'}(\tau^{(1)})^2] &(15c)\end{aligned}$$

Proof.

$$\operatorname*{Var}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D})]-\operatorname*{Var}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{IS}}(\mathcal{D})]$$

$$= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D})^2] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\mathcal{D})^2] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D})]^2 + \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\mathcal{D})]^2 \tag{16}$$

$$= \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2] - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})^2] - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2 \tag{17}$$

$$\begin{aligned}= &\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2] - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2\rho^{\complement}_{\theta'}(\tau^{(1)})^2] \\ &- \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2\end{aligned} \tag{18}$$

$$\begin{aligned}= &\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2] - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})^2] \\ &- \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2,\, \rho^{\complement}_{\theta'}(\tau^{(1)})^2]\end{aligned} \tag{19}$$

$$\begin{aligned}= &\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})^2] \\ &- \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2,\, \rho^{\complement}_{\theta'}(\tau^{(1)})^2]\end{aligned} \tag{20}$$

$$\begin{aligned}= &\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2]\Big(\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})]^2 - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})^2]\Big) \\ &- \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2,\, \rho^{\complement}_{\theta'}(\tau^{(1)})^2]\end{aligned} \tag{21}$$

$$\begin{aligned}= &-\frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2]\underset{\mathcal{D}\sim\pi_b}{\text{Var}}[\rho^{\complement}_{\theta'}(\tau^{(1)})] \\ &- \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})]^2 + \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})]^2 - \frac{1}{|\mathcal{D}|}\underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)})^2,\, \rho^{\complement}_{\theta'}(\tau^{(1)})^2]\end{aligned} \tag{22}$$

where we use the definition of variance (16), linearity of expectation (17), the relationship between IS and OSIRIS (18), the definition of covariance (19), Equation 1 (20), factor terms (21), and the definition of variance again (22). □

C.2. Bias

Theorem 2. *Given any mapping $\theta' : \mathcal{S} \to \{0, 1\}$, the mean of the OSIRIS estimator is:*

$$\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta')] = V^{\pi_e} - \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta'),\, \rho^{\complement}_{\theta'}(\tau^{(1)})] \tag{23}$$

Proof. From the fact that the IS estimator is unbiased:

$$\begin{aligned}V^{\pi_e} &= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{IS}}(\mathcal{D})] = \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{IS}}(\tau^{(1)})] &(24)\\ &= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')\cdot\rho^{\complement}_{\theta'}(\tau^{(1)})] &(25)\\ &= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\theta'}(\tau^{(1)})] + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta'),\, \rho^{\complement}_{\theta'}(\tau^{(1)})] &(26)\\ &= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta')] + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau^{(1)};\theta'),\, \rho^{\complement}_{\theta'}(\tau^{(1)})] &(27)\end{aligned}$$

where we use the assumption that trajectories are sampled i.i.d. (24), the relationship between IS and OSIRIS (25), the definition of covariance (26), and then Equation 1 (27). □

C.3. Decomposition of the Covariance Term

We can decompose the covariance involving the *product* of likelihood ratios into a sum of covariances involving *individual* likelihood ratios. We use the notation $\rho^{\complement}_{1:\ell}$ to indicate the product of the 1st to ℓth irrelevant likelihood ratios.

$$\underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta'),\, \rho^{\complement}_{\text{OSIRIS}}(\tau^{(1)};\theta')]$$

$$
\begin{aligned}
&= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{\text{OSIRIS}}(\tau^{(1)};\theta')] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\text{OSIRIS}}(\tau^{(1)};\theta')] && (28)\\
&= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{\text{OSIRIS}}(\tau^{(1)};\theta')] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')] && (29)\\
&= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-1}(\tau^{(1)};\theta')]\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\rho^{\complement}_{\ell}(\tau^{(1)};\theta')] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')]\\
&\quad + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-1}(\tau^{(1)};\theta'),\ \rho^{\complement}_{\ell}(\tau^{(1)};\theta')] && (30)\\
&= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-1}(\tau^{(1)};\theta')] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')]\\
&\quad + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-1}(\tau^{(1)};\theta'),\ \rho^{\complement}_{\ell}(\tau^{(1)};\theta')] && (31)\\
&= \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')] - \underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')]\\
&\quad + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-1}(\tau^{(1)};\theta'),\ \rho^{\complement}_{\ell}(\tau^{(1)};\theta')]\\
&\quad + \underset{\mathcal{D}\sim\pi_b}{\text{Cov}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')\rho^{\complement}_{1:\ell-2}(\tau^{(1)};\theta'),\ \rho^{\complement}_{\ell-1}(\tau^{(1)};\theta')]\\
&\quad + \cdots && (32)
\end{aligned}
$$

where we use the definition of covariance (28), Equation 1 (29), the definition of covariance again (30), Equation 1 again (31), and repeat (32). Eventually, the $\underset{\mathcal{D}\sim\pi_b}{\mathbb{E}}[\hat{V}_{\text{OSIRIS}}(\tau^{(1)};\theta')]$ will cancel, leaving us with the sum of individual covariances.

D. Derivations for the OSIRIS Estimator using State Relevance

D.1. Composite Policy π'_e

Lemma 1. *Let π'_e be a composite policy:*

$$
\pi'_e(a\,|\,s;\theta) \equiv \begin{cases} \pi_e(a\,|\,s) & \textit{if } \theta(s) = 1 \textit{ (relevant)} \\ \pi_b(a\,|\,s) & \textit{if } \theta(s) = 0 \textit{ (irrelevant)} \end{cases}. \tag{33}
$$

Then the policy values of π_e and π'_e are equal:

$$
V^{\pi_e} = V^{\pi'_e}. \tag{34}
$$

Proof. First, we will recall some notation. The state-action value function under policy π_e is

$$
Q^{\pi_e}(s,a) \equiv \underset{\tau\sim\pi_e}{\mathbb{E}}[g_{t:T}(\tau)\,|\,s_t = s,\ a_t = a], \tag{35}
$$

and the state value function under policy π_e is then

$$
V^{\pi_e}(s) \equiv \underset{a\sim\pi_e(\cdot\,|\,s)}{\mathbb{E}}\big[Q^{\pi_e}(s,a)\big]. \tag{36}
$$

Now, for any state s,

$$
V^{\pi_e}(s) = \underset{a\sim\pi_e(\cdot\,|\,s)}{\mathbb{E}}\big[Q^{\pi_e}(s,a)\big] = \underset{a\sim\pi'_e(\cdot\,|\,s)}{\mathbb{E}}\big[Q^{\pi_e}(s,a)\big] \tag{37}
$$

because if $\theta(s) = 0$, then we can replace π_e in the expectation with anything because, by definition of state irrelevance, $Q^{\pi_e}(s,a)$ is constant with respect to a. Otherwise, if $\theta(s) = 1$, then π'_e is simply equivalent to π_e by definition of the composite policy. By applying this logic with the recursive Bellman equation $V^{\pi_e}(s) = \mathbb{E}_{a\sim\pi_e(\cdot\,|\,s)}\Big[\mathbb{E}_{s'\sim P(\cdot\,|\,s,a)}\big[R(s,a) + \gamma V^{\pi_e}(s')\big]\Big]$, we conclude that for any state s,

$$
V^{\pi_e}(s) = V^{\pi'_e}(s). \tag{38}
$$

We take an expectation over the initial state distribution,

$$
\underset{s_1\sim P(s_1)}{\mathbb{E}}\big[V^{\pi_e}(s_1)\big] = \underset{s_1\sim P(s_1)}{\mathbb{E}}\big[V^{\pi'_e}(s_1)\big], \tag{39}
$$

which is equivalent to

$$V^{\pi_e} = V^{\pi'_e}. \tag{40}$$

□

D.2. Bias using θ

Theorem 3. *Given the true relevance mapping θ, the mean of the OSIRIS estimator is*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta)] = V^{\pi_e} \tag{41}$$

Proof. The key idea here is that the likelihood ratio omission procedure in OSIRIS is equivalent to pretending π'_e is the evaluation policy.

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}\left[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\theta)\right] \tag{42}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi_b}\left[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\tau;\theta)\right] \tag{43}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi_b}\left[g(\tau)\prod_{t=1}^{T}\left[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\right]^{\theta(s_t)}\right] \tag{44}$$

$$= \int g(\tau)\prod_{t=1}^{T}\left[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\right]^{\theta(s_t)}\pi_b(a_t\,|\,s_t)P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau \tag{45}$$

$$= \int g(\tau)\prod_{t=1}^{T}\left[\pi_e(a_t\,|\,s_t)\right]^{\theta(s_t)}\left[\pi_b(a_t\,|\,s_t)\right]^{1-\theta(s_t)}P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau \tag{46}$$

$$= \int g(\tau)\prod_{t=1}^{T}\pi'_e(a_t\,|\,s_t)P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau \tag{47}$$

$$= \mathop{\mathbb{E}}_{\tau\sim\pi'_e}[g(\tau)] = V^{\pi'_e} = V^{\pi_e} \tag{48}$$

where we use the assumption that trajectories are sampled i.i.d. (43), the definition of the OSIRIS estimator (44), the definition of the expectation (45), cancellation of terms (46), and the definition of the composite policy π'_e (47). Finally, we clean up by reintroducing the expectation, using Lemma 1, and recovering the definition of the true policy value (48). □

Note that if state s is truly relevant but was omitted by OSIRIS, then bias would be introduced from Equation 37:

$$\left|\mathop{\mathbb{E}}_{a\sim\pi_b}[Q^{\pi_e}(s,a)] - \mathop{\mathbb{E}}_{a\sim\pi_e}[Q^{\pi_e}(s,a)]\right| \neq 0 \tag{49}$$

However, if state s is truly irrelevant but was kept by OSIRIS, then there would be no effect on bias because Equation 37 still holds because π_e in the expectation can be replaced with anything.

D.3. Consistency using $\hat{\theta}$

Theorem 4. *If $|\mathcal{A}| = 2$ and $\alpha > 0$, then as $|\mathcal{D}| \to \infty$*

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}[\hat{V}^{\pi_e}_{\text{OSIRIS}}(\mathcal{D};\hat{\theta}(\,\cdot\,;\mathcal{D}))] = V^{\pi_e} \tag{50}$$

Proof. Welch's two-sample t-test is consistent, which means that for all $s \in \mathcal{S}$ where $\theta(s) = 1$, the test will give $\hat{\theta}(s) = 1$ as $|\mathcal{D}| \to \infty$. The binary classification of actions when calculating $\hat{\theta}$ has no effect on this fact because the action space is already assumed to be binary. Thus, as $|\mathcal{D}| \to \infty$, all relevant likelihood ratios will be kept by OSIRIS, so in this limit, the estimator will be unbiased (Appendix D.2). □

D.4. Step-Wise OSIRIS

Theorem 5. *Let state $s \in \mathcal{S}$ be irrelevant to the reward Δt-steps away if*

$$\mathop{\mathbb{E}}_{\tau\sim\pi_e}\left[r_{t+\Delta t}\,|\,s_t = s, a_t = a\right] = \text{constant}, \quad \forall a \in \mathcal{A}. \tag{51}$$

Otherwise, s is relevant to the reward Δt-steps away. Using this condition, we define $\theta_{\Delta t} : \mathcal{S} \to \{0,1\}$ where $\theta_{\Delta t}(s) = 0$ if s is irrelevant to the reward Δt-steps away, and otherwise $\theta_{\Delta t}(s) = 1$. Then

$$\hat{V}^{\pi_e}_{\substack{\text{step-wise}\\\text{OSIRIS}}}(\mathcal{D};\theta_{\Delta t}) \equiv \frac{1}{|\mathcal{D}|}\sum_{\tau\in\mathcal{D}}\sum_{t'=1}^{T}\left(\gamma^{t'-1}r_{t'}\prod_{t=1}^{T}\left[\rho_t(\tau)\right]^{\theta_{t'-t}(s_t)}\right) \tag{52}$$

is an unbiased estimator of V^{π_e}.

Proof. This proof is analogous to Sections D.1 and D.2 where we use a composite policy

$$\pi'_e(a\,|\,s;\theta_{\Delta t}) \equiv \begin{cases}\pi_e(a\,|\,s) & \text{if } \theta_{\Delta t}(s) = 1 \text{ (relevant)}\\ \pi_b(a\,|\,s) & \text{if } \theta_{\Delta t}(s) = 0 \text{ (irrelevant)}\end{cases}. \tag{53}$$

We will first prove the lemma in Equation 56 and next prove that the step-wise OSIRIS estimator is equivalent to pretending π'_e is the evaluation policy.

First, the expected reward at each individual time step t' is

$$\begin{aligned}\mathop{\mathbb{E}}_{\tau\sim\pi_e}\left[R(s_{t'},a_{t'})\right] &= \mathop{\mathbb{E}}_{s_t\sim P(s_t)}\Big[\mathop{\mathbb{E}}_{a_t\sim\pi_e(a_t\,|\,s_t)}\Big[\mathop{\mathbb{E}}_{\tau\sim\pi_e}\left[R(s_{t'},a_{t'})\,|\,s_t,a_t\right]\Big]\Big] &(54)\\ &= \mathop{\mathbb{E}}_{s_t\sim P(s_t)}\Big[\mathop{\mathbb{E}}_{a_t\sim\pi'_e(a_t\,|\,s_t)}\Big[\mathop{\mathbb{E}}_{\tau\sim\pi_e}\left[R(s_{t'},a_{t'})\,|\,s_t,a_t\right]\Big]\Big] &(55)\end{aligned}$$

where we use the law of total expectation by introducing $P(s_t)$ as the stationary state distribution. Then we introduce the composite policy defined in Equation 53 because π'_e is equivalent to π_e except in states that are irrelevant (to the reward $(t'-t)$-steps away) where, by definition, we can replace π_e in the expectation with anything. Recursively applying this logic, we find that

$$\mathop{\mathbb{E}}_{\tau\sim\pi_e}\left[R(s_{t'},a_{t'})\right] = \mathop{\mathbb{E}}_{\tau\sim\pi'_e}\left[R(s_{t'},a_{t'})\right] \tag{56}$$

Now, we will use this fact to show the step-wise OSIRIS estimator is unbiased:

$$\begin{aligned}&\mathop{\mathbb{E}}_{\mathcal{D}\sim\pi_b}\Big[\hat{V}^{\pi_e}_{\substack{\text{step-wise}\\\text{OSIRIS}}}(\mathcal{D};\theta_{\Delta t})\Big] &(57)\\ &= \mathop{\mathbb{E}}_{\tau\sim\pi_b}\Big[\hat{V}^{\pi_e}_{\substack{\text{step-wise}\\\text{OSIRIS}}}(\tau;\theta_{\Delta t})\Big] &(58)\\ &= \mathop{\mathbb{E}}_{\tau\sim\pi_b}\Big[\sum_{t'=1}^{T}\gamma^{t'-1}r_{t'}\prod_{t=1}^{T}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]^{\theta_{t'-t}(s_t)}\Big] &(59)\\ &= \sum_{t'=1}^{T}\gamma^{t'-1}\mathop{\mathbb{E}}_{\tau\sim\pi_b}\Big[r_{t'}\prod_{t=1}^{T}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]^{\theta_{t'-t}(s_t)}\Big] &(60)\\ &= \sum_{t'=1}^{T}\gamma^{t'-1}\int r_{t'}\prod_{t=1}^{T}\Big[\frac{\pi_e(a_t\,|\,s_t)}{\pi_b(a_t\,|\,s_t)}\Big]^{\theta_{t'-t}(s_t)}\pi_b(a_t\,|\,s_t)P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau &(61)\\ &= \sum_{t'=1}^{T}\gamma^{t'-1}\int r_{t'}\prod_{t=1}^{T}\left[\pi_e(a_t\,|\,s_t)\right]^{\theta_{t'-t}(s_t)}\left[\pi_b(a_t\,|\,s_t)\right]^{1-\theta_{t'-t}(s_t)}P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau &(62)\\ &= \sum_{t'=1}^{T}\gamma^{t'-1}\int r_{t'}\prod_{t=1}^{T}\pi'_e(a_t\,|\,s_t)P(s_{t+1}\,|\,s_t,a_t)P(s_1)\,d\tau &(63)\end{aligned}$$

$$= \sum_{t'=1}^{T} \gamma^{t'-1} \mathop{\mathbb{E}}_{\tau \sim \pi'_e} [r_{t'}] = \sum_{t'=1}^{T} \gamma^{t'-1} \mathop{\mathbb{E}}_{\tau \sim \pi_e} [r_{t'}] = V^{\pi_e} \tag{64}$$

where we use the assumption that trajectories are sampled i.i.d. (58), the definition of the step-wise OSIRIS estimator (59), linearity of expectation (60), the definition of the expectation (61), cancellation of terms (62), and the definition of the composite policy π'_e (63). Finally, we clean up by reintroducing the expectation, using Equation 56, and recovering the definition of the true policy value (64). □

E. Environment Descriptions

All reported results are aggregated over 200 independent trials with discount factor $\gamma = 1$.

E.1. Gridworlds

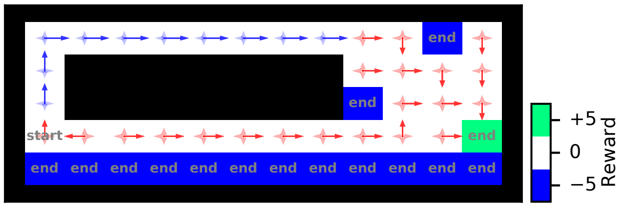


Figure 1: Environment and policies for Gridworld experiments.

The Gridworld environment and policies are shown in Figure 1. From each state, the agent takes the action indicated by the large arrow with probability $1 - \epsilon$ and takes a random action with probability ϵ. The evaluation policy has $\epsilon = 0.1$, and the behavior policy has $\epsilon = 0.5$. Rewards are given for entering the indicated states. The data size is $|\mathcal{D}| = 25$. These parameters describe the *Dilly-Dallying Gridworld*. The *Express Gridworld* variant is identical except that the behavior policy has $\epsilon = 0.2$ only in the corridor states (blue arrows in Figure 1).

E.2. Lunar Lander

Lunar Lander is a popular deterministic benchmark environment in OpenAI gym with an 8-dimensional continuous state space and 4 discrete actions (Brockman et al., 2016). The goal is to safely land on the ground by firing engines that push the agent left, right, or up. The reward from each transition is related to the agent's distance from the target landing site with small penalties for firing engines; around 100-140 is accumulated for a typical landing. A successful landing earns an additional $+100$ whereas a crash earns -100, and both events end the trajectory. Or the trajectory is automatically ended after 1000 steps.

We used ϵ-greedy behavior and evaluation policies based on an optimal policy learned by DQN (Mnih et al., 2013). The behavior policy takes a random action with probability $\epsilon = 0.1$, and the evaluation policy does so with probability $\epsilon = 0.05$. The data size is $|\mathcal{D}| = 50$.

Calculating $\hat{\theta}(s; \mathcal{D})$ requires a discrete state space, so only for this step of the OSIRIS algorithm, we discretized the state space by creating three linearly spaced bins per state dimension. In each trial $\mathcal{D}$, this procedure resulted in about 200 discrete states that were represented at least once.

E.3. Cart Pole

Cart Pole is another deterministic benchmark environment in OpenAI gym. It has a 4-dimensional continuous state space and 2 discrete actions (Barto et al., 1983; Brockman et al., 2016). The goal is to apply leftward or rightward force to the bottom (the cart) of an inverted pendulum (the pole) to keep it balanced upright. The agent gets $+1$ reward for each transition until the the pole falls, which ends the trajectory.

The behavior and evaluation policies and the discrete state space were obtained as described for Lunar Lander, but the

Table 1: Comparison of mean squared errors for alternative implementations of OSIRIS state relevance implementation.

		Algorithm 1		g_τ-Binary $\mathcal{A}$		Smirnov		NN as $\hat{Q}^{\pi_e}$		On-
		OSIRIS	OSIRWIS	OSIRIS	OSIRWIS	OSIRIS	OSIRWIS	OSIRIS	OSIRWIS	Policy
Dilly-Dallying Gridworld	Mean	**1.3**	1.1	0.5	0.4	0.5	0.5	0.9	0.0	4.3
	Std	2.0	1.8	**1.7**	1.8	1.7	1.8	5.3	2.4	0.6
	RMSE	**3.6**	3.7	4.2	4.3	4.1	4.2	6.3	4.9	0.6
Express Gridworld	Mean	0.7	0.8	0.7	**0.9**	0.5	0.6	0.7	0.3	4.3
	Std	**1.2**	1.3	1.3	1.5	1.3	1.4	2.9	2.1	0.6
	RMSE	3.8	**3.7**	3.8	3.7	4.0	4.0	4.6	4.5	0.6
Cart Pole	Mean	**1068.9**	759.7	970.0	622.3	582.0	640.7	608.2	608.2	1503.6
	Std	3961.8	318.5	5337.1	381.8	619.0	308.3	**97.6**	**97.6**	244.8
	RMSE	3985.5	**809.2**	5363.7	960.4	1110.2	916.3	900.7	900.7	244.8
Lunar Lander	Mean	**244.6**	234.8	241.2	259.7	286.9	248.9	222.9	249.4	245.3
	Std	55.3	23.5	230.3	**13.5**	171.7	16.3	62.0	14.4	6.8
	RMSE	55.3	25.7	230.3	19.7	176.6	16.7	65.9	**15.0**	6.8

policies here used $\epsilon = 0.25$ and $\epsilon = 0.2$, respectively. The data size is $|\mathcal{D}| = 50$. In each trial $\mathcal{D}$, about 16 discrete states were represented at least once.

F. Extended Results

F.1. Alternative Implementations

To demonstrate the extent to which our analysis is robust to specific implementation choices, here we present empirical results for other possible implementations (Table 1).

The *Algorithm 1* implementation refers to the procedure presented in the main text with Welch's two-sample t-test. We also tried using *Smirnov*'s non-parametric test (Hodges, 1958) where we increased the significance level to $\alpha = 0.2$.

These two-sample statistical tests require that we binarize the action space and discretize the state space. First, in the main text, we proposed binarizing the actions based on their likelihood ratios. Here, we also tried assigning action classes based on whether they led to above- or below-average trajectory returns, which is the $g(\tau)$*-Binary* $\mathcal{A}$ implementation. Finally, we also tried directly regressing $\hat{Q}^{\pi_e}$ with a neural network model. This *NN as* $\hat{Q}^{\pi_e}$ implementation can certainly handle continuous states and it can handle action spaces $|\mathcal{A}| > 2$. The model had one 32-dimensional hidden layer. For each trial (i.e. sample of $\mathcal{D}$), it was trained to minimize the Huber loss for 500 epochs, each looking at a minibatch of 128 transitions. We use the trained $\hat{Q}^{\pi_e}$ to calculate $\hat{\theta}(s)$ for a given s. First, we calculate the standard deviation of $\hat{Q}^{\pi_e}(s, a)$ for all $a \in \mathcal{A}$. We normalize this by dividing by the mean of $\hat{Q}^{\pi_e}(s, a)$ for all $a \in \mathcal{A}$. This value gives us a sense of how much the Q^{π_e} value function varies for different actions taken from the same state. If the value is greater than $\alpha = 0.4$, then we output $\hat{\theta}(s) = 1$ (relevant), and otherwise, $\hat{\theta}(s) = 0$ (irrelevant).

In Table 2, we also provide empirical results for the OSIRIS estimator using an "oracle" state relevance mapping where $\hat{\theta}(s) = 0$ (irrelevant) for all s in the Gridworld corridor, and otherwise $\hat{\theta}(s) = 1$ (relevant). These results represent a practical bound on the accuracy improvement from likelihood ratio omission. As such, they also give a rough picture of the amount of variance contributed from estimating $\hat{\theta}$ vs noise inherent to the data or variance contributed by OSIRIS's manipulating trajectory lengths (which is expected to be small).

F.2. Express Gridworld

See Figures 2, 3, and 4a for empirical results in the Express Gridworld environment. This figures show the same qualitative trends as reported in the main text for the Dilly-Dallying Gridworld. Figure 4a shows OSIRIS was more likely to label corridor states are relevant in Express Gridworld than Dilly-Dallying Gridworld, which is expected because there are fewer visits to the corridor states, so there is likely more noise being picked up by the statistical test.

Table 2: Mean squared errors for OSIRIS using an "oracle" state relevance mapping.

		ORACLE	
		OSIRIS	OSIRWIS
DILLY-DALLYING GRIDWORLD	MEAN	3.1	3.4
	STD	1.9	1.2
	RMSE	2.3	1.5
EXPRESS GRIDWORLD	MEAN	2.9	2.9
	STD	3.1	1.8
	RMSE	3.4	2.3

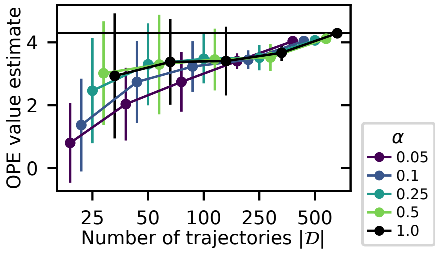


Figure 2: Express Gridworld. Distributions of OSIRWIS estimates showing estimator consistency. Dots represent means, error bars represent standard deviations, and horizontal line represents the mean of the on-policy MC estimator. Colors indicate α values, where $\alpha = 1$ is equivalent to ordinary WIS.

F.3. Estimated State Relevance

We show the estimated state relevance over all state dimensions in Cart Pole (Figure 4b) and Lunar Lander (Figure 4c).

References

Barto, A. G., Sutton, R. S., and Anderson, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13(5):834–846, 1983.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym, 2016. arXiv:1606.01540.

Hodges, J. L. The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486, 1958.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with Deep Reinforcement Learning, 2013. arXiv:1312.5602.

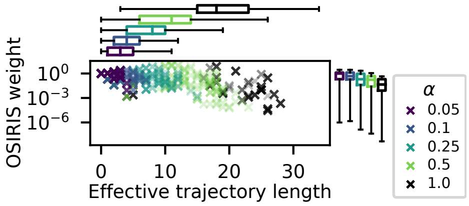
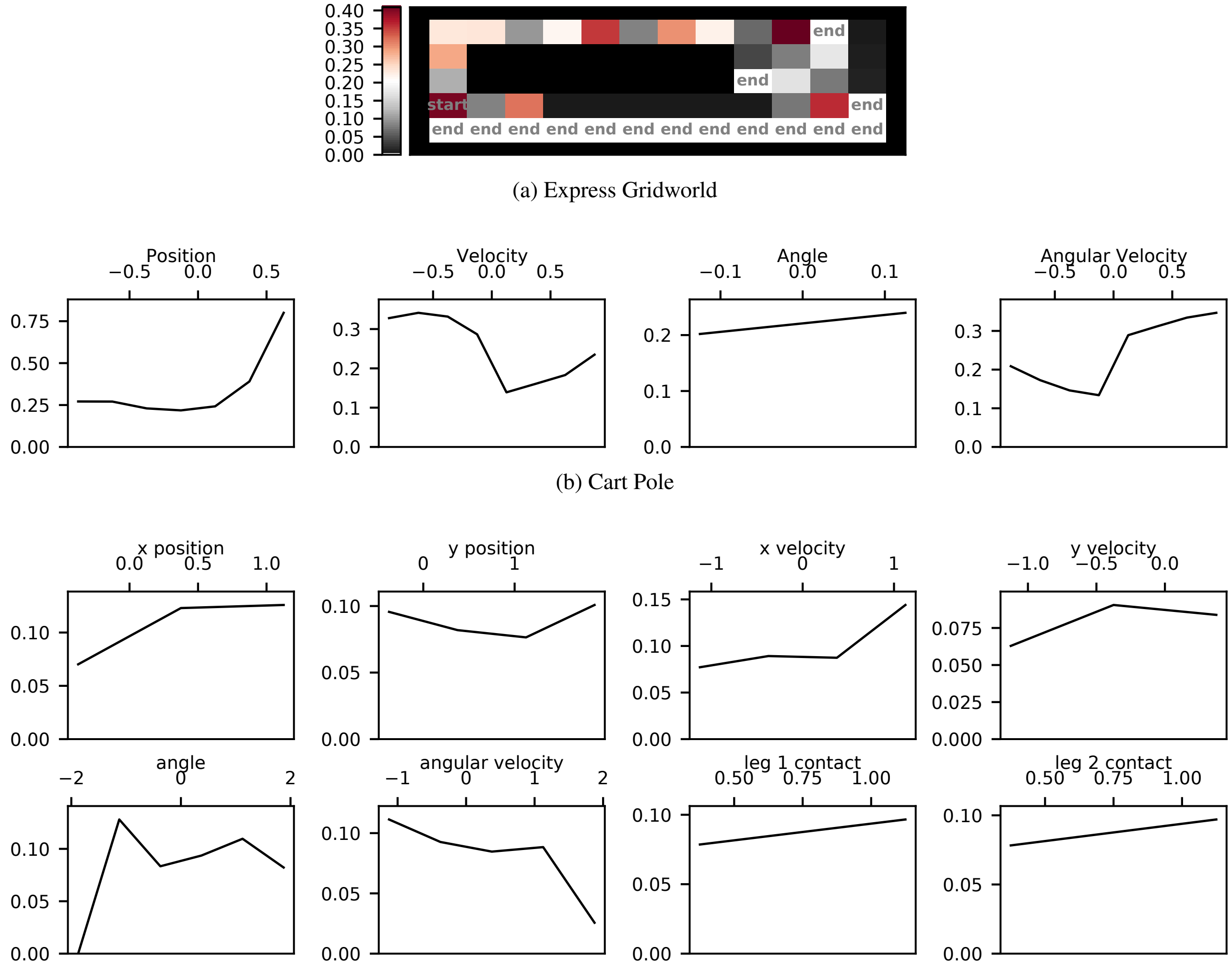


Figure 3: Express Gridworld. Scatter plots show correlation between OSIRIS weights and effective trajectory lengths $\sum_{t=1}^{T} \hat{\theta}(s_t)$. Boxplots show variance reduction of OSIRIS weights by shortening and evening of the effective trajectory lengths. Colors indicate α values, where $\alpha = 1$ is equivalent to ordinary IS.



(a) Express Gridworld

(b) Cart Pole

(c) Lunar Lander

Figure 4: Mean of estimated state relevance $\hat{\theta}(s)$ from visits to the indicated states is represented by color (a) or on the y axis (b, c). States identified as relevant (i.e. $\hat{\theta}(s) = 1$) are key decision points where trajectory outcome is sensitive to action taken.