APPENDICES: Conditional Importance Sampling for Off-Policy Learning

A Proofs

Proposition 4.2. Assume the support condition (SC) holds. Given a trajectory functional Ψ and an associated SCF Φ, the estimator in Expression (11) is unbiased for $\mathbb{E}_{\eta^\pi}[\Psi(\tau_{0:n})]$. Further, its variance is no greater than that of the OIS estimator in Expression (10).

Proof. The proof of unbiasedness follows the logic of Proposition 3.1's proof and the proof for the variance upper bound follows the logic of Proposition 3.2's proof. Beginning with unbiasedness, we make the following calculation:

$$\begin{aligned}\mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right]\Psi(\tau_{0:n})\right] &\overset{(a)}{=} \mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\middle|\Phi(\tau_{0:n})\right]\right]\\ &\overset{(b)}{=} \mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\right]\\ &\overset{(c)}{=} \mathbb{E}_{\eta^\pi_{0:n}|_{(x,a)}}\left[\Psi(\tau_{0:n})\right]\,,\end{aligned}$$

where (a) follows since Φ is an SCF for Ψ (and hence $\Psi(\tau_{0:n})$ is fully determined by $\Phi(\tau_{0:n})$), (b) follows from the tower law of conditional expectations, and (c) follows from standard importance sampling theory.

For the variance result, we observe that

$$\mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right]\Psi(\tau_{0:n}) = \mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\middle|\Phi(\tau_{0:n})\right]\,,$$

which follows since Φ is an SCF for Ψ. Therefore, this estimator is a conditional expectation of the OIS estimator

$$\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\,,$$

and therefore the conclusion follows by direct application of Equation (9) which was used to establish Proposition 3.2, taking $Z_1 = \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})$ and $Z_2 = \Phi(\tau_{0:n})$. □

Proposition 4.3. For any given MDP, and pair of policies π and μ satisfying (SC), and target functional Ψ, the variance preorder *refines* the inclusion preorder. That is, for any two SCFs Φ_1, Φ_2 of Ψ, if $\Phi_1 \precsim \Phi_2$, then we have $\Phi_1 \precsim_{\mathbb{V}} \Phi_2$.

Proof. Assume we have $\Phi_1 \precsim \Phi_2$ for two sufficient conditioning functionals Φ_1, Φ_2 for Ψ. Since $\Phi_1(\tau_{0:n})$ is a function of $\Phi_2(\tau_{0:n})$, we have that $\mathbb{E}[\rho^{\pi,\mu}_{1:n-1}|\Phi_1(\tau_{0:n})] = \mathbb{E}[\mathbb{E}[\rho^{\pi,\mu}_{1:n-1}|\Phi_2(\tau_{0:n})]|\Phi_1(\tau_{0:n})]$ by the tower property for conditional expectations. The statement now follows from the conditional variance formula (9). □

Proposition 4.4. An SCF for Ψ for which the associated estimator in Expression (11) achieves minimal variance is Ψ itself.

Proof. This follows by first observing that $\Psi(\tau_{0:n})$ is a minimal sufficient conditioning functional for Ψ with respect to the ordering induced by $\precsim$; this is immediate from the definition. Next, since $\precsim_{\mathbb{V}}$ refines $\precsim$ (by Proposition 4.3), we have that $\Psi(\tau_{0:n})$ is also a minimal sufficient conditioning functional with respect to $\precsim_{\mathbb{V}}$, and the statement follows. □

Proposition 5.1. Assume the support condition (SC). For a given policy μ let $p^\mu|_{(x,a)}$ be the probability mass function of $\sum_{t=0}^{n-1}\gamma^t R_t$ under $\eta^\mu|_{(x,a)}$. Then we have

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\rho^{\pi,\mu}_{1:n-1}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] = \frac{p^\pi|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}{p^\mu|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}\,. \tag{12}$$

Proof. As in the discussion in Section 3, we have

$$\rho^{\pi,\mu}_{1:n-1} = \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})} \, .$$

We then decompose

$$\begin{aligned}\mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] =& \mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{p_\pi|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n}|\sum_{t=0}^{n-1}\gamma^t R_t)}{p_\mu|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n}|\sum_{t=0}^{n-1}\gamma^t R_t)}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] \\ =& \frac{p_\pi|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}{p_\mu|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}\mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n}|\sum_{t=0}^{n-1}\gamma^t R_t)}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n}|\sum_{t=0}^{n-1}\gamma^t R_t)}\middle|\sum_{t=0}^{n-1}\gamma^t R_t\right] \\ =& \frac{p_\pi|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)}{p_\mu|_{(x,a)}(\sum_{t=0}^{n-1}\gamma^t R_t)} \, ,\end{aligned}$$

as required. □

Proposition 5.2. A global minimum for each of the objectives in Expressions (15) and (16) is given by

$$f_\theta(\Phi(\tau_{0:n})) = \mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right] \, .$$

Proof. We begin by restating Expression (15), and use the tower law of conditional expectation as follows:

$$\begin{aligned}&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)^2\right] \\ =&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)^2\middle|\Phi(\tau_{0:n})\right]\right] \, .\end{aligned}$$

The inner conditional expectation is of the form $\mathbb{E}_Y[(z-Y)^2]$; viewed as a function of z, it is well known that the minimiser of such an expression is $z = \mathbb{E}[Y]$. Thus, for a fixed value of $\Phi(\tau_{0:n})$, the optimal value of $f_\theta(\Phi(\tau_{0:n}))$ is given by

$$\mathbb{E}_{\eta^\mu}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right] \, .$$

Therefore, the global optimiser of Expression (15) is given precisely by the function

$$f_\theta(\Phi(\tau_{0:n})) = \mathbb{E}_{\eta^\mu}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right] \, ,$$

as required. For Expression (16), in a similar manner we can write the following:

$$\begin{aligned}&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)^2 \Psi(\tau_{0:n})^2\right] \\ =&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\left(f_\theta(\Phi(\tau_{0:n})) - \frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\right)^2\middle|\Phi(\tau_{0:n})\right]\Psi(\tau_{0:n})^2\right] \, ,\end{aligned}$$

with the equality following from the fact that Φ is a sufficient conditioning functional for Ψ. Now we may proceed in an identical manner to that for Expression (15), and the claim follows. □

We also record a precise result on the form of the SCIS weights described in Section 5 below.

Proposition A.1. As described in Section 5, assuming the support condition, we have

$$\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|X_t, A_t, R_t\right] = \frac{p^\pi_t|_{(x,a)}(X_t)}{p^\mu_t|_{(x,a)}(X_t)} \times \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} \, .$$

Proof. The proof follows by factorising the trajectory probabilities $\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})$, $\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})$ in the following manner, using the Markov property of the environment:

$$\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n}) = p^\pi_t|_{(x,a)}(X_t)\pi(A_t|X_t)\eta^\pi_{t:n}|_{(X_t,A_t)}(\tau_{t:n})\eta^\pi_{0:t-1}|_{(x,a)}(\tau_{0:t-1}|X_t)\,,$$

where we write $\eta^\pi_{0:t-1}|_{(x,a)}(\tau_{0:t-1}|X_t)$ for probability mass associated with the trajectory $\tau_{0:t-1}$ under $\eta^\pi_{0:t}$, conditional on the trajectory visiting the state X_t at time t. Using conditional independence, we therefore have

$$\begin{aligned}
&\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^\mu_{0:n}|_{(x,a)}(\tau_{0:n})}\middle| X_t, A_t, R_t\right] \\
=&\frac{p^\pi_t|_{(x,a)}(X_t)\pi(A_t|X_t)}{p^\mu_t|_{(x,a)}(X_t)\mu(A_t|X_t)}\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{t:n}|_{(X_t,A_t)}(\tau_{t:n})\eta^\pi_{0:t}|_{(x,a)}(\tau_{0:t-1}|X_t)}{\eta^\mu_{t:n}|_{(X_t,A_t)}(\tau_{t:n})\eta^\mu_{0:t}|_{(x,a)}(\tau_{0:t-1}|X_t)}\middle| X_t, A_t, R_t\right] \\
=&\frac{p^\pi_t|_{(x,a)}(X_t)\pi(A_t|X_t)}{p^\mu_t|_{(x,a)}(X_t)\mu(A_t|X_t)}\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{t:n}|_{(X_t,A_t)}(\tau_{t:n})}{\eta^\mu_{t:n}|_{(X_t,A_t)}(\tau_{t:n})}\middle| X_t, A_t\right]\mathbb{E}_{\eta^\mu|_{(x,a)}}\left[\frac{\eta^\pi_{0:t}|_{(x,a)}(\tau_{0:t-1}|X_t)}{\eta^\mu_{0:t}|_{(x,a)}(\tau_{0:t-1}|X_t)}\middle| X_t\right] \\
=&\frac{p^\pi_t|_{(x,a)}(X_t)\pi(A_t|X_t)}{p^\mu_t|_{(x,a)}(X_t)\mu(A_t|X_t)}\,,
\end{aligned}$$

as required. The final equality follows since both of the conditional expectations are in fact expectations of Radon-Nikodym derivatives under the measure in the "denominator" of the derivative, and hence evaluate to 1 almost surely. □

B Experimental details

B.1 Environment

Chain. We use a 6-state chain environment, with absorbing states at each end of the chain. Two actions, `left` and `right`, are available at each state of the chain. Transitions corrupted with $p\%$ noise means that with probability p, a transition to a uniformly-random adjacent state (independent of the action taken) occurs. Each non-terminal step incurs a reward of $+1$, whilst reaching an absorbing state incurs a one-off reward of $+10$, and the episode then terminates. The initial state of the environment is taken to be the third state from the left. Figure 4 provides an illustration.

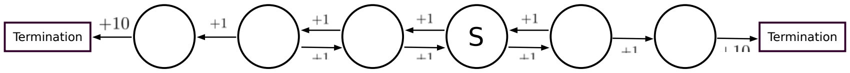

Figure 4: Illustration of the chain environment.

B.2 Other experimental details: operator estimation

Throughout, the discount factor is taken to be $\gamma = 0.99$, and the Q-function used to form the target $(T^\pi)^n Q$ has its entries sampled independently from the $N(0, 0.1)$ distribution. The policies π and μ are drawn independently, with each $\pi(\cdot|x)$ and $\mu(\cdot|x)$ drawn independently from a Dirichlet$(1, \ldots, 1)$ distribution. Default values of parameters are taken as $n = 5$, the transition noise level is set to 10%, and the learning rate is set to 0.1, and 100 repetitions of each experiments are performed to compute the bootstrapped confidence intervals.

B.3 Other experimental details: policy evaluation

The environment and default parameters are exactly the same as in the operator estimation experiments, with the exception that the Q-function is initialised so that all coordinates are 0, and $n = 3$. We estimate bootstrap confidence intervals using 500 repetitions of each experiment. In the linear function approximation experiments, we use a version of tile-coding [Sutton and Barto, 2018]; the specification parametrisation we use is as follows. For a chain of length K, we take a weight vector $\mathbf{w} = (w_{k,a}|k \in [K-1], a \in \mathcal{A}) \in \mathbb{R}^{(K-1)\times|\mathcal{A}|}$. Labelling the states of the chain $x_1, \ldots, x_K$, we parametrise $Q(x_1, a)$ by $w_{1,a}$, $Q(x_K, a)$ by $w_{K-1,a}$, and $Q(x_k, a)$ by $\frac{1}{2}w_{k-1,a} + \frac{1}{2}w_{k,a}$, for each $a \in \mathcal{A}$ and each $1 < k < K$; this is illustrated in Figure 5. The weight vector is initialised with all coordinates equal to 0 in all experiments.

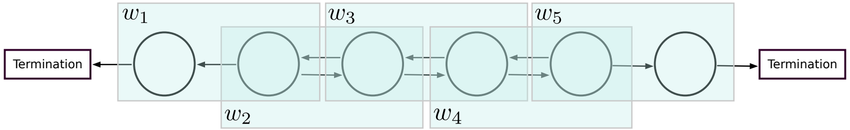

Figure 5: An illustration of the tile-coding scheme used in the linear function approximation scheme; the figure shows how feature weights (for each action) are allocated states. The value prediction at each state is given by averaging the weights allocated to the state.

B.4 Further experimental results

In this section, we give in Figure 6 the results described in Section 6.2, including also results for oracle versions of the CIS algorithms in question. We observe that the performance of the online versions of CIS algorithms generally closely track that of their oracle counterparts.

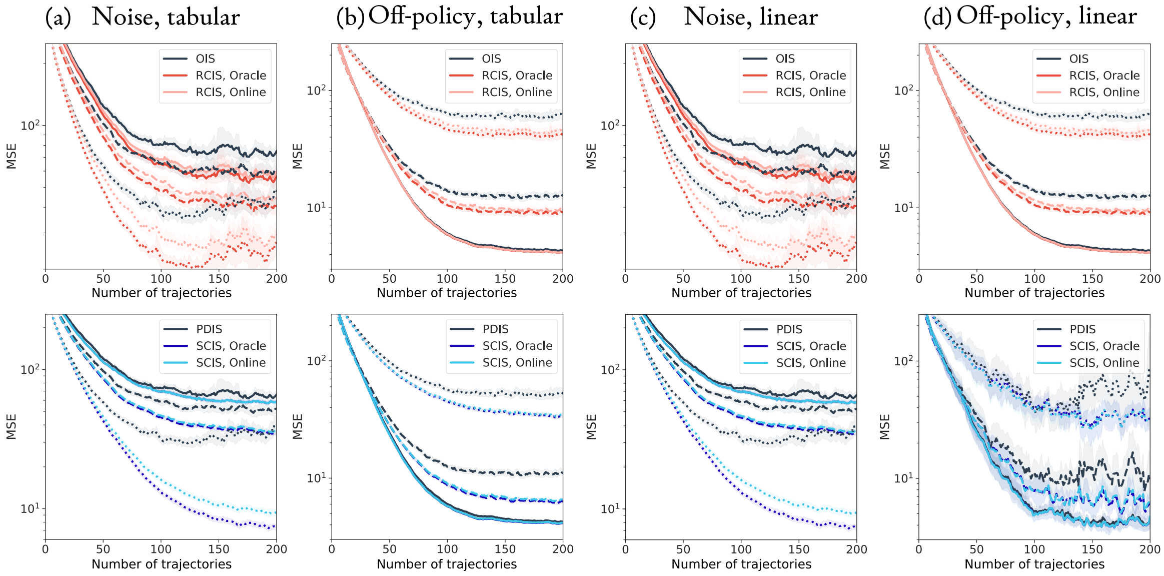

Figure 6: Policy evaluation MSE as a function of number of trajectories for OIS, RCIS, PDIS, and SCIS, with both tabular and function approximation variants. Shaded regions indicate bootstrapped 95% confidence intervals.

C Extending the CIS framework

C.1 A measure-theoretic perspective on conditional importance sampling

In this section, we give a measure-theoretic treatment of the conditional importance sampling framework introduced in Section 4 of the main paper. We do not provide any fundamentally new results relative to the main paper, but we believe the measure-theoretic exposition gives a useful perspective, and may be useful for future work.

We begin by returning to the trajectory importance-weighted estimator given in Expression (10) in the main paper:

$$\frac{\eta^{\pi}_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^{\mu}_{0:n}|_{(x,a)}(\tau_{0:n})}\Psi(\tau_{0:n})\,.$$

This expression weights the target quantity $\Psi(\tau_{0:n})$ by the importance weight associated with the proposal distribution $\eta^{\mu}_{0:n}$ and the target distribution $\eta^{\pi}_{0:n}$. A conditional importance sampling estimator is formed by taking a function Φ that in the language of the main paper, is a sufficient conditioning functional for Ψ, and forming the new estimator

$$\mathbb{E}\left[\frac{\eta^{\pi}_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^{\mu}_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\Phi(\tau_{0:n})\right]\Psi(\tau_{0:n})\,.$$

Proposition 4.2 then shows that the variance of the conditioned estimator is no greater than that of the trajectory-weighted estimator, and, roughly speaking, in many cases it is strictly lower.

Whilst this perspective of conditioning on functionals Φ of the trajectory is conceptually straightforward and clearly hints at how such techniques can be implemented in practice, as described in Section 5.3, there are some subtleties introduced by this perspective that make the analysis of the method less straightforward. One such case is illustrated by the following example: consider two sufficient conditioning functionals Φ_1 and Φ_2 for a target Ψ, which happen to be related according to the identity $\Phi_1(\tau_{0:n}) = 2\Phi_2(\tau_{0:n})$ for all $\tau_{0:n}$. Intuitively, Φ_1 and Φ_2 encode the same information about $\tau_{0:n}$, and thus the estimators they produce are identical. We might therefore like to be able to treat Φ_1 and Φ_2 as "identical" in our analysis, and yet this is made difficult by the focus of the analysis on *functionals* of the trajectory. This is related to the need to work with *preorders* in Section 4.1, rather than the perhaps more familiar notion of *partial orders*. One route around this difficulty is to define an equivalence relation over functions of the trajectory, rigorously encoding the notion of "captures the same information about $\tau_{0:n}$", and then to work instead with equivalence classes of trajectory functionals under this relation. However, this has the potential to be very unwieldy, and further, it turns out this is essentially equivalent to a much more familiar collection of objects from measure theory, known as sigma-algebras. For formal definitions and background on sigma-algebras, see for example Billingsley [1995]. We note that technically speaking, it is necessary to constrain functionals of the trajectory to be *measurable*; we do not mention this condition further in this section, but return to it in Appendix C.2 when describing the application of the conditional importance sampling framework to more general classes of MDPs. For a general random variable Z, we write $\mathscr{F}_Z$ for the sigma-algebra generated by Z; in the discussion that follows, all random variables will be defined over the same probability space, which we therefore suppress from the notation in what follows.

The counterpart to a sufficient conditioning functional Φ is a *sufficient conditioning sigma-algebra* (SCSA) $\mathscr{F}$, which is defined as being a sigma-algebra over the same measurable space as $\mathscr{F}_{\tau_{0:n}}$, with the property that $\mathscr{F}_{\Psi(\tau_{0:n})} \subseteq \mathscr{F}$. With this definition, a functional Φ is an SCF if and only if $\mathscr{F}_{\Phi(\tau_{0:n})}$ is an SCSA. The corresponding importance sampling estimator is then given by

$$\mathbb{E}_{\eta^{\mu}|_{(x,a)}}\left[\frac{\eta^{\pi}_{0:n}|_{(x,a)}(\tau_{0:n})}{\eta^{\mu}_{0:n}|_{(x,a)}(\tau_{0:n})}\middle|\mathscr{F}\right]\Psi(\tau_{0:n})\,.$$

The analogue of the preorder $\precsim$ over conditioning functionals is the *inclusion partial order* $\subseteq$ over sigma-algebras; we have $\Phi_1 \precsim \Phi_2$ if and only if $\mathscr{F}_{\Phi_1(\tau_{0:n})} \subseteq \mathscr{F}_{\Phi_2(\tau_{0:n})}$. Further, if for two conditioning functionals Φ_1 and Φ_2 we have $\Phi_1 \precsim \Phi_2$ and $\Phi_2 \precsim \Phi_1$ (that is, roughly speaking, Φ_1 and Φ_2 encode the same information about the trajectory), then we have $\mathscr{F}_{\Phi_1(\tau_{0:n})} = \mathscr{F}_{\Phi_2(\tau_{0:n})}$. Thus, working with sigma-algebras eliminates the issue of several conditioning objects representing exactly the same information about the trajectory.

C.2 Generalising the conditional importance sampling framework to other classes of MDPs

We have restricted the presentation in the main paper to MDPs with finite state and action spaces and reward distributions with finite support for ease of exposition, and to avoid having to introduce measure-theoretic terminology such as Radon-Nikodym derivatives to deal with more general classes of MDPs. Nevertheless, the framework described in the main paper applies much more generally, such as for certain classes of MDPs with continuous state and/or action spaces. In this section, we briefly describe how the framework generalises to these settings. The aim is not to be exhaustive, but rather to indicate how key concepts change when moving away from the assumptions of the main paper; for a rigorous treatment of the measure-theoretic issues that arise in MDPs with more general state and action spaces, see Bertsekas and Shreve [2007].

Consider now an MDP with a general state space $\mathcal{X}$ and action space $\mathcal{A}$, each equipped with a sigma-algebra, and consider $\mathbb{R}$, the domain of rewards in the MDP, to be equipped with its usual Borel sigma-algebra. Given measurable transition kernel $P : \mathcal{X} \times \mathcal{A} \rightarrow \mathscr{P}(\mathcal{X})$, reward kernel $\mathcal{R} : \mathcal{X} \times \mathcal{A} \rightarrow \mathscr{P}(\mathbb{R})$, initial state distribution $\nu \in \mathscr{P}(\mathcal{X})$, and two Markov policies $\pi, \mu : \mathcal{X} \rightarrow \mathscr{P}(\mathcal{A})$, we can straightforwardly define trajectory measures $\eta^{\mu}_{0:n}$, $\eta^{\pi}_{0:n}$, and conditional trajectory measures $\eta^{\mu}_{0:n}|_{(x,a)}$, $\eta^{\pi}_{0:n}|_{(x,a)}$ over the relevant product space. The key requirement in order to be able to carry out importance sampling in this more general case is that $\eta^\pi_{0:n}|_{(x,a)}$ is absolutely continuous with respect to $\eta^\mu_{0:n}|_{(x,a)}$. When this is the case, the Radon-Nikodym derivative

$$\frac{\mathrm{d}\eta^\pi_{0:n}|_{(x,a)}}{\mathrm{d}\eta^\mu_{0:n}|_{(x,a)}}(\tau_{0:n})$$

exists, and has the property that for a measurable functional Ψ of the trajectory, under mild integrability conditions, we have

$$\mathbb{E}_{\eta^\mu_{0:n}|_{(x,a)}}\left[\frac{\mathrm{d}\eta^\pi_{0:n}|_{(x,a)}}{\mathrm{d}\eta^\mu_{0:n}|_{(x,a)}}(\tau_{0:n})\Psi(\tau_{0:n})\right] = \mathbb{E}_{\eta^\pi_{0:n}|_{(x,a)}}[\Psi(\tau_{0:n})] \ ,$$

the fundamental property we require an importance weight to satisfy. The CIS framework of the main paper can thus be extended to these more general settings by computing conditional expectations of the Radon-Nikodym derivative of the two trajectory measures. We conclude by noting that in several practical applications of interest, $\mathcal{X}$ and $\mathcal{A}$ are themselves subsets of Euclidean spaces, with $\pi(\cdot|x)$ and $\mu(\cdot|x)$ taken to have densities with respect to Lebesgue measure for each $x \in \mathcal{X}$; in such circumstances, under mild assumptions, the Radon-Nikodym derivative can be expressed in the familiar form of a product of action density ratios; that is

$$\frac{\mathrm{d}\eta^\pi_{0:n}|_{(x,a)}}{\mathrm{d}\eta^\mu_{0:n}|_{(x,a)}}(\tau_{0:n}) = \prod_{t=1}^{n-1} \frac{\pi(A_t|X_t)}{\mu(A_t|X_t)} \, .$$

However, in cases where the action distribution $\pi(\cdot|x)$ is *not* absolutely continuous with respect to $\mu(\cdot|x)$, such as in deterministic policy gradient algorithms [Silver et al., 2014, Lillicrap et al., 2016], the measure $\eta^\pi_{0:n}$ is *not* absolutely continuous with respect to $\eta^\mu_{0:n}$, meaning that the Radon-Nikodym derivative does not exist, and so importance sampling, and in particular the CIS framework, cannot straightforwardly be applied.