10 Appendix

10.1 Case Studies

As a common issue in MIS, the general estimators are usually difficult to optimize due to the minimax form. One solution is to choose the discriminator class ($\mathcal{Q}$ in our case) to be an RKHS, which often leads to a closed-form solution to the inner max and reduces the minimax optimization to a single minimization problem [16, 18, 26]. Below we show that this is also the case for our estimator, and provide the closed-form expression for the inner maximization when $\mathcal{Q}$ is an RKHS.

Lemma 10.1. *Let $\langle .,. \rangle_{\mathcal{H}_K}$ be the inner-product of $\mathcal{H}_K$ which satisfies the Reproducible Kernel Hilbert Space (RKHS) property. When the function space $\mathcal{Q} = \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}; \langle q, q \rangle_{\mathcal{H}_K} \leq 1\}$, the term $\max_{q \in \mathcal{Q}} L_w(w, \beta, q)^2$ has the following closed-form expression:*

$$
\begin{aligned}
&\mathbb{E}_{\substack{(s,a,s')\sim\mu \\ (\tilde{s},\tilde{a},\tilde{s}')\sim\mu}}[w(s,a) \cdot w(\tilde{s},\tilde{a}) \cdot \beta(s,a) \cdot \beta(\tilde{s},\tilde{a}) \cdot (K((s,a),(\tilde{s},\tilde{a})) - 2\gamma \mathbb{E}_{a'\sim\pi(\cdot|s')}[K((s',a'),(\tilde{s},\tilde{a}))] \\
&+ \gamma^2 \mathbb{E}_{\substack{a'\sim\pi(.s') \\ \tilde{a}'\sim\pi(.\tilde{s}')}}[K((s',a'),(\tilde{s}',\tilde{a}'))])] - 2(1-\gamma)\mathbb{E}_{\substack{(s,a,s')\sim\mu \\ \tilde{s}\sim d_0, \tilde{a}\sim\pi(\cdot|\tilde{s})}}[w(s,a) \cdot \beta(s,a) \cdot (K((s,a),(\tilde{s},\tilde{a})) \\
&- \gamma \mathbb{E}_{a'\sim\pi(.s')}[K((s',a'),(\tilde{s},\tilde{a}))]] + (1-\gamma)^2 \mathbb{E}_{\substack{s\sim d_0, a\sim\pi(\cdot|s) \\ \tilde{s}\sim d_0, \tilde{a}\sim\pi(\cdot|\tilde{s})}}[K((s,a),(\tilde{s},\tilde{a}))].
\end{aligned}
$$

Furthermore, when we use linear functions to approximate both w and q, the final estimator has a closed-form solution

Lemma 10.2. *Consider linear parameterization $w(s,a) = \phi(s,a)^T \alpha$, where $\phi \in \mathbb{R}^d$ is a feature map in $\mathbb{R}^d$ and α is the linear coefficients. Similarly let $q(s,a) = \Psi(s,a)^T \zeta$ where $\Psi \in \mathbb{R}^d$. Then, assuming that we have an estimate of $\frac{d^{\pi}_{P_{tr}}}{\mu}$ as $\hat{\beta}$, we can empirically estimate $\hat{w}$ using Equation 8, which has a closed-form expression $\hat{w}(s,a) = \phi(s,a)^T \hat{\alpha}$, where*

$$
\hat{\alpha} = (\mathbb{E}_{n,(s,a,s')\sim\mu}[(\Psi(s,a) - \gamma\Psi(s',\pi)) \cdot \phi(s,a)^T \cdot \hat{\beta}(s,a)])^{-1} (1-\gamma) \mathbb{E}_{n,s\sim d_0}[\Psi(s,\pi)] \quad (10)
$$

provided that the matrix being inverted is non-singular. Here, $\mathbb{E}_n$ is the empirical expectation using n-samples.

Detailed proof for these Lemma can be found in section 10.4 and 10.5 respectively.

10.2 Q-Function Estimator

In this section, we show an extension of our idea that can approximate the Q-function in the target environment. Similar to we did in the previous section, we now consider the OPE error of a candidate function q, that is, $|(1-\gamma)\mathbb{E}_{s\sim d_0}[q(s,\pi)] - J(\pi)|$, under the assumption that $w_{P_{te}/P_{tr}} \in conv(\mathcal{W})$:

$$
\begin{aligned}
&|(1-\gamma)\mathbb{E}_{s\sim d_0}[q(s,\pi)] - J_{P_{te}}(\pi)| = |\mathbb{E}_{\substack{(s,a)\sim d^{\pi}_{P_{te}}, \\ r\sim R(s,a), s'\sim P(s,a)}}[q(s,a) - \gamma q(s',\pi)] - \mathbb{E}_{\substack{(s,a)\sim d^{\pi}_{P_{tr}} \\ r\sim R(s,a)}}[W_{P_{te}/P_{tr}} \cdot r]| \\
&= |\mathbb{E}_{\substack{(s,a)\sim\mu, \\ r\sim R(s,a), s'\sim P(s,a)}}[W_{P_{te}/P_{tr}} \cdot \beta \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{\substack{(s,a)\sim d^{\pi}_{P_{tr}} \\ r\sim R(s,a)}}[W_{P_{te}/P_{tr}} \cdot r]| \\
&\leq \sup_{w\in\mathcal{W}} |\mathbb{E}_{\substack{(s,a)\sim\mu, \\ r\sim R(s,a), s'\sim P(s,a)}}[w \cdot \beta \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{\substack{(s,a)\sim d^{\pi}_{P_{tr}} \\ r\sim R(s,a)}}[w \cdot r]| \\
&=: \sup_{w\in\mathcal{W}} L_q(w, \beta, q).
\end{aligned}
\quad (11)
$$

The inequality step uses the assumption that $w_{P_{te}/P_{tr}} \in conv(\mathcal{W})$, and the final expression is a valid upper bound on the error of using q for estimating $J_{P_{te}}(\pi)$. It is also easy to see that the bound is tight because $q = Q^{\pi}_{P_{te}}$ satisfies the Bellman equation on all state-action pairs, and hence $L_q(w, \beta, Q^{\pi}_{P_{te}}) \equiv 0$.

Using this derivation, we propose the following estimator which will estimate $Q^{\pi}_{P_{te}}$.

$$Q^{\pi}_{P_{te}} \approx \hat{q} := \arg\min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} L_q(w,\beta,q). \tag{12}$$

Below we provide the results that parallel Lemmas 10.1 and 10.2 for the Q-function estimator.

Lemma 10.3. *Let $\langle .,. \rangle_{\mathcal{H}_K}$ be the inner-product of $\mathcal{H}_K$ which satisfies the Reproducible Kernel Hilbert Space (RKHS) property. When the function space $\mathcal{W} = \{w : \mathcal{S}\times\mathcal{A} \to \mathbb{R} | \langle w,w\rangle_{\mathcal{H}_K} \le 1\}$. The term $\max_{w\in\mathcal{W}} L_q(w,\beta,q)^2$ has a closed form expression.*

We defer the detailed expression and its proof to Appendix 10.6.

Lemma 10.4. *Let $w = \phi(s,a)^T\alpha$ where $\phi \in \mathbb{R}^d$ is some basis function. Let $q(s,a) = \Psi(s,a)^T\zeta$, where $\Psi(s,a) \in \mathbb{R}^d$. Then, assuming that we have an estimate of $\frac{d^{\pi}_{P_{tr}}}{\mu}$ as $\hat{\beta}$, we can empirically estimate $\hat{q}$ using uniqueness condition similar to Equation 12, which has a closed-form expression $\hat{w}(s,a) = \Psi(s,a)^T\hat{\zeta}$, where*

$$\hat{\zeta} = (\mathbb{E}^{\pi}_{n,\mu}[\hat{\beta}\cdot(\Phi(s,a)\Psi(s,a)^T - \gamma\Phi(s,a)\Psi(s',\pi)]))^{-1}\mathbb{E}_{n,(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a))}[\Phi(s,a)\cdot r] \tag{13}$$

where, $\mathbb{E}_n$ is the empirical expectation calculated over n-samples and assuming that the provided matrix is non-singular.

Theorem 10.5. *Let $\hat{\beta}$ be our estimation of β using [20]. We utilize this $\hat{\beta}$ to further optimize for $\hat{w}_n$ (equation 8) using n samples. In both cases, $\mathbb{E}_{(s,a)\sim d^{\pi}_{P_{tr}}}[\cdot]$ is also approximated with n samples from the simulator P_{tr}. Then, under Assumptions 1 and 2 along with the additional assumption that $Q^{\pi}_{P_{te}} \in C(\mathcal{Q})$ with probability at least $1-\delta$, We can guarantee the OPE error for $\hat{q}_n$ which was optimized using equation 12 on n samples.*

$$\begin{aligned}
&|(1-\gamma)\mathbb{E}_{d_0}[\hat{q}_n(s,\pi)] - J_P(\pi)| \le \\
&\min_{q\in\mathcal{Q}}\max_{w\in\mathcal{W}} L_q(w,\beta,q) + 4\mathcal{R}_n(\mathcal{W},\mathcal{Q}) + 2C_{\mathcal{W}}\frac{R_{max}}{1-\gamma}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \\
&+ C_{\mathcal{W}}\frac{R_{max}}{1-\gamma}\cdot\tilde{O}\left(\sqrt{\|\frac{d^{\pi}_{P_{tr}}}{\mu}\|_{\infty}\left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_{\mathcal{F}}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}\right)}\right)
\end{aligned}$$

where $\mathcal{R}_n(\mathcal{F}), \mathcal{R}_n(\mathcal{W},\mathcal{Q})$ are the Radamacher complexities of function classes $\{(x,y)\to f(x) - \log(f(y)) : f\in\mathcal{F}\}$ and $\{(s,a,s')\to (w(s,a)\cdot\frac{d^{\pi}_{P'}(s,a)}{\mu(s,a)}\cdot(q(s,a)-\gamma q(s',\pi)) : w\in\mathcal{W}, q\in\mathcal{Q}\}$, respectively, $\|d^{\pi}_{P'}/\mu\|_{\infty} := \max_{s,a} d^{\pi}_{P'}(s,a)/\mu(s,a)$ measures the distribution shift between $d^{\pi}_{P'}$ and μ, and $\tilde{O}(\cdot)$ is the big-Oh notation suppressing logarithmic factors. Under the assumption $w^{\pi}_{P_{tr}/P_{te}} \in C(\mathcal{W})$,

10.3 Derivation for β-GradientDICE

We will show a demonstration on finite state-action space. The following identity holds true for $\tau_* = \frac{d^{\pi}_{P_{te}}}{d^{\pi}_{P_{tr}}}$. Let us assume that we have the diagonal matrix D with diagonal elements being $d^{\pi}_{P_{tr}}$. The following identity holds true.

$$D\tau_* = \mathcal{T}\tau_* \tag{14}$$

Where, $d_0(s,a) = d_0(s)\pi(a|s)$ and $\mathcal{T}$ is the reverse bellman operator

$$\mathcal{T}y = (1-\gamma)d_0(s,a) + \gamma P^T_{\pi} Dy$$

Where, $P_{\pi}((s,a),(s',a')) = P_{te}(s'|s,a)\pi(a'|s')$ To estimate τ, we can simply run the following optimization

$$\tau := \arg\min_{\tau:S\times A\to\mathbb{R}} |D\tau - \mathcal{T}\tau|^2_{D^{-1}} + \frac{\lambda}{2}((d^{\pi}_{P_{tr}})^T\tau - 1)$$

Here, $|y|_\Sigma^2 = y^T \Sigma y$. The optimization above can be simplified in form of expectation over $d^\pi_{P_{tr}}$.

$$\mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[(\frac{\delta(s,a)}{d^\pi_{tr}(s,a)})^2] + \frac{\lambda}{2}((d^\pi_{P_{tr}})^T\tau - 1)$$

With, $\delta(s,a) = D\tau - \mathcal{T}\tau$, We can now apply Fenchel Conjugate principle to get the following

$$\max_{f:S\times A\to\mathbb{R}} \mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[\frac{\delta(s,a)}{d^\pi_{P_{tr}}} f(s,a) - \frac{1}{2}f(s,a)^2] + \max_{\eta\in\mathbb{R}}(\mathbb{E}_{d^\pi_{P_{tr}}}[\eta\tau(s,a) - \eta] - \frac{\eta^2}{2})$$

If we simplify the above optimization, we get the following form

$$\begin{aligned}
&\frac{d^\pi_{P_{te}}}{d^\pi_{P_{te}}} := \arg\min_{\tau:S\times A\to\mathbb{R}} \max_{f:S\times A\to\mathbb{R},\eta\in\mathbb{R}} L(\tau,\eta,f) \\
&= (1-\gamma)\mathbb{E}_{s_0\sim d_0, a_0\sim\pi(\cdot|s_0)}[f(s_0,a_0)] + \gamma \mathbb{E}_{\substack{(s,a)\sim d^\pi_{P_{tr}} \\ s'\sim P_{te}(\cdot|s,a), a'\sim\pi(\cdot|s')}}[\tau(s,a)f(s',a')] \\
&- \mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[\tau(s,a)f(s,a)] - \frac{1}{2}\mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[f(s,a)^2] + \lambda\mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[\eta\tau(s,a) - \eta^2/2].
\end{aligned}$$

While we don't have samples from $(s,a,s') \sim d^\pi_{P_{tr}}$. We can simply re-weight the term $\mathbb{E}_{\substack{(s,a)\sim d^\pi_{P_{tr}} \\ s'\sim P_{te}(\cdot|s,a), a'\sim\pi(\cdot|s')}}[\tau(s,a)f(s',a')]$ with $\beta(s,a) = \frac{d^\pi_{P_{tr}}}{\mu}$. This completes the derivation of β-GradientDICE.

$$\begin{aligned}
&\frac{d^\pi_{P_{te}}}{d^\pi_{P_{te}}} := \arg\min_{\tau:S\times A\to\mathbb{R}} \max_{f:S\times A\to\mathbb{R},\eta\in\mathbb{R}} L(\tau,\eta,f) \\
&= (1-\gamma)\mathbb{E}_{s_0\sim d_0, a_0\sim\pi(\cdot|s_0)}[f(s_0,a_0)] + \gamma \mathbb{E}_{(s,a,s')\sim\mu, a'\sim\pi(\cdot|s')}[\beta(s,a)\tau(s,a)f(s',a')] \\
&- \mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[\tau(s,a)f(s,a)] - \frac{1}{2}\mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[f(s,a)^2] + \lambda\mathbb{E}_{(s,a)\sim d^\pi_{P_{tr}}}[\eta\tau(s,a) - \eta^2/2].
\end{aligned}$$

10.4 Proof of Lemma 10.1

Since $\mathcal{Q}$ belongs to the RKHS space. We can use the reproducible property of RKHS to re-write the optimization in the following form.

$$\begin{aligned}
&L_w(w,\beta,q)^2 = (\mathbb{E}_{(s,a)\sim\mu, s'\sim P_{te}(s,a)}[w(s,a)\cdot\beta(s,a)\cdot(q(s,a) - \gamma q(s',\pi))] - (1-\gamma)\mathbb{E}_{s\sim d_0}[q(s,\pi)])^2 \\
&= (\mathbb{E}_{(s,a)\sim\mu, s'\sim P_{te}(s,a)}[w(s,a)\cdot\beta(s,a)\cdot(\langle q, K((s,a),.),\cdot\rangle_{\mathcal{H}_K} - \gamma\mathbb{E}_{a'\sim\pi(.s')}[\langle q, K((s',a'),.),\cdot\rangle_{\mathcal{H}_K}]] \\
&\qquad - (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\pi(.|s)}[\langle q, K((s,a),.),\cdot\rangle_{\mathcal{H}_K}]))^2 \\
&= \max_{q\in\mathcal{Q}} \langle q, q^*\rangle^2_{\mathcal{H}_K}
\end{aligned} \tag{15}$$

Where,

$$q^*(\cdot) = \mathbb{E}_\mu[w(s,a)\cdot\beta(s,a)\cdot(K((s,a),.) - \gamma\mathbb{E}_{a'\sim\pi(.s')}[K((s',a'),.)]] - (1-\gamma)\mathbb{E}_{s\sim d_0, a\sim\pi(.|s)}[K((s,a),.)]) \tag{16}$$

We go from first line to the second line by exploiting the linear properties of the RKHS function space. Given the constraint that $\mathcal{Q} = \{q : \mathcal{S}\times\mathcal{A}\to\mathbb{R};\ \langle q,q\rangle_{\mathcal{H}_K} \leq 1\}$ we can maximise $\max_q L(w,\beta,q)^2$ using Cauchy-Shwartz inequality

$$\begin{aligned}
&\max_q L_w(w,\beta,q)^2 = \langle q^*, q^*\rangle^2_{\mathcal{H}_K} \\
&= \mathbb{E}_{\substack{(s,a,s')\sim\mu \\ (\tilde{s},\tilde{a},\tilde{s}')\sim\mu}}[w(s,a)\cdot w(\tilde{s},\tilde{a})\cdot\beta(s,a)\cdot\beta(\tilde{s},\tilde{a})\cdot(K((s,a),(\tilde{s},\tilde{a})) - 2\gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}[K((s',a'),(\tilde{s},\tilde{a}))] \\
&+ \gamma^2\mathbb{E}_{\substack{a'\sim\pi(.s') \\ \tilde{a}'\sim\pi(.\tilde{s}')}}[K((s',a'),(\tilde{s}',\tilde{a}'))])] - 2(1-\gamma)\mathbb{E}_{\substack{(s,a,s')\sim\mu \\ \tilde{s}\sim d_0, \tilde{a}\sim\pi(\cdot|\tilde{s})}}[w(s,a)\cdot\beta(s,a)\cdot(K((s,a),(\tilde{s},\tilde{a})) \\
&- \gamma\mathbb{E}_{a'\sim\pi(.s')}[K((s',a'),(\tilde{s},\tilde{a}))]] + (1-\gamma)^2\mathbb{E}_{\substack{s\sim d_0, a\sim\pi(\cdot|s) \\ \tilde{s}\sim d_0, \tilde{a}\sim\pi(\cdot|\tilde{s})}}[K((s,a),(\tilde{s},\tilde{a}))]
\end{aligned}$$

This completes the proof.

10.5 Proof of Lemma 10.2

Substituting the functional forms for $q(s,a) = \Psi(s,a)^T\zeta$ and $w(s,a) = \phi(s,a)^T\alpha$ we get the following expression for $L_{n,w}(w,\hat{\beta},q)$. Where, $\hat{\beta}$ is an estimate of $\frac{d^{\pi}_{P_{tr}}}{\mu}$

$$L_{n,w}(w,\hat{\beta},q) = \mathbb{E}_{n,\mu}[\phi(s,a)^T\alpha \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a) - \gamma\Psi(s',\pi))^T\zeta)] - (1-\gamma)\mathbb{E}_{n,d_0}[\Psi(s,\pi)^T\zeta]$$

Using the uniqueness condition we derived in equation 6, we can go about finding the value of α by equating $L(w,\hat{\beta},q)$ to zero.

$$\mathbb{E}_{n,\mu}[\phi(s,a)^T\alpha \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a) - \gamma\Psi(s',\pi))^T\zeta)] - (1-\gamma)\mathbb{E}_{n,d_0}[\Psi(s,\pi)^T\zeta] = 0$$
$$\alpha^T\mathbb{E}_{n,\mu}[\phi(s,a) \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a) - \gamma\Psi(s',\pi))^T)]\zeta = (1-\gamma)\mathbb{E}_{n,d_0}[\Psi(s,\pi)^T]\zeta$$

Since the loss is linear in ζ, we can solve for α using the matrix inversion operation.

$$\hat{\alpha} = (\mathbb{E}_{n,\mu}[(\Psi(s,a) - \gamma\Psi(s',\pi)) \cdot \phi(s,a)^T \cdot \hat{\beta}])^{-1}(1-\gamma)\mathbb{E}_{n,d_0}[\Psi(s,\pi)]$$

This completes the proof.

10.6 Proof of Lemma 10.3

Consider the loss function $\sup_{w\in\mathcal{W}} L_q(w,\beta,q)^2$. Since $\mathcal{W}$ is in RKHS space. Using reproducible property of RKHS space we can re-write this maximization as follows,

$$\max_{w\in\mathcal{W}} L_q(w,\beta,q)^2 = \max_{w\in\mathcal{W}}(\mathbb{E}_{(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[w(s,a) \cdot \beta(s,a) \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[w(s,a)$$
$$\max_{w\in\mathcal{W}}(\mathbb{E}_{(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[\langle w, K(s,a),\cdot\rangle_{\mathcal{H}_K} \cdot \beta(s,a) \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[\langle w, K(s,a),\cdot\rangle_{\mathcal{H}_K}$$
$$\max_{w\in\mathcal{W}}\langle w, \mathbb{E}_{(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[K((s,a),\cdot) \cdot \beta(s,a) \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[K((s,a),\cdot) \cdot r]\rangle_{\mathcal{H}_K})^2$$
$$\max_{w\in\mathcal{W}}\langle w, w^*\rangle^2_{\mathcal{H}_K} = \langle w^*, w^*\rangle^2_{\mathcal{H}_K}$$

Where, we use the linear properties of RKHS spaces and then followed by using Cauchy-Shwartz inequality, to compute the maximization. Where, w^* has the following expression.

$$w^*(\cdot) = \mathbb{E}_{(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[K((s,a),\cdot) \cdot \beta(s,a) \cdot (q(s,a) - \gamma q(s',\pi))] - \mathbb{E}_{(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[K((s,a),\cdot) \cdot r]$$

The maximization expression thus takes the following form

$$\langle w^*, w^*\rangle^2_{\mathcal{H}_K} = \mathbb{E}_{\substack{(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)\\(\tilde{s},\tilde{a})\sim\mu,\tilde{s}'\sim P(s,a),r\sim R(s,a)}}[K((s,a),(\tilde{s},\tilde{a})) \cdot \beta(s,a) \cdot \beta(\tilde{s},\tilde{a}) \cdot \Delta(q,s,a,s') \cdot \Delta(q,\tilde{s},\tilde{a},\tilde{s}')]$$
$$-2\mathbb{E}_{\substack{(s,a)\sim\mu,s'\sim P(s,a)\\(\tilde{s},\tilde{a})\sim d^{\pi}_{P_{tr}},\tilde{r}\sim R(s,a),}}[K((s,a),(\tilde{s},\tilde{a})) \cdot \beta(s,a) \cdot \Delta(q,s,a,s') \cdot r] + \mathbb{E}_{\substack{(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)\\(\tilde{s},\tilde{a})\sim d^{\pi}_{P_{tr}},\tilde{r}\sim R(s,a)}}[K((s,a),(\tilde{s},\tilde{a})) \cdot r \cdot \tilde{r}]$$

Where, $\Delta(q,s,a,s') = q(s,a) - \gamma q(s',\pi)$.
This completes the proof.

10.7 Proof of Lemma 10.4

Substituting the functional forms of $q(s,a) = \Psi(s,a)^T\zeta$, $w(s,a) = \phi(s,a)^T\alpha$. Also substituting the estimate for $\frac{d^{\pi}_{P_{tr}}}{\mu}$ as $\hat{\beta}$. We get the following expression

$$L_{q,n}(w,\hat{\beta},q) =$$
$$|\mathbb{E}_{n,(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[\phi(s,a)^T\alpha \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a)^T\zeta - \gamma\Psi(s',\pi)^T\zeta)] - \mathbb{E}_{n,(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[\phi(s,a)^T\alpha \cdot r]|$$
$$= 0$$

Where, the equality comes from the uniqueness condition similar to equation 6

$$\alpha^T\mathbb{E}_{n,(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[\phi(s,a) \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a) - \gamma\Psi(s',\pi))^T]\zeta = \alpha^T\mathbb{E}_{n,(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[\phi(s,a) \cdot r]$$

Since the equations above are linear in α. So it suffices to show that the optimal solution can be reached if β is approximated as follows,

$$\hat{\zeta} = (\mathbb{E}_{n,(s,a)\sim\mu,s'\sim P(s,a),r\sim R(s,a)}[\phi(s,a) \cdot \hat{\beta}(s,a) \cdot (\Psi(s,a) - \gamma\Psi(s',\pi))^T])^{-1} \cdot \mathbb{E}_{n,(s,a)\sim d^{\pi}_{P_{tr}},r\sim R(s,a)}[\phi(s,a) \cdot r] \quad (17)$$

Where, $\mathbb{E}_{n,\cdot}$ denotes the empirical approximation of the expectation. This completes the proof.

10.8 Proof of Theorem 5.1

To prove this theorem, we will first require a Lemma that we need to prove first. This is as follows,

Lemma 10.6. *Under Assumptions 1 and 2, suppose we use n samples each from distribution P and Q to empirically estimate the ratio of $\frac{P}{Q}$ using equation 2. The estimation error can be bounded with probability at least $1-\delta$ as follows:*

$$\left\|\hat{f}_n - \frac{P}{Q}\right\|_\infty^2 \leq \tilde{O}\left(\|\frac{P}{Q}\|_\infty\left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right)\right) \tag{18}$$

Proof. Since, equation 2 is optimized using empirical samples it is an Empirical Risk Minimization (ERM) algorithm. We denote the original loss with respect to a function $f \in \mathcal{F}$ as $L(f)$. Using familiar result from learning theory (Corollary 6.1 [28]) with probability at-least $1-\delta$

$$L(\hat{f}_n) - L(\frac{P}{Q}) \leq 4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_\mathcal{F}\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} \tag{19}$$

With probability at least $1-\delta$. Where, $\mathcal{R}_n(\mathcal{F})$ is the Radamacher complexity of the function class

$$\{(p,q) \to f(q) - \log(f(p) : f \in \mathcal{F}\} \tag{20}$$

Now, let's turn our attention to the left hand side. Before we end up doing that let's define the estimation error $\bar{e}_n(x) = \hat{f}_n(x) - \frac{P(x)}{Q(x)}$. Thus, we can re-write the left hand side in terms of $\bar{e}$

$$\begin{aligned} L(\hat{f}_n) - L(\frac{P}{Q}) &= L(\frac{P}{Q} + \bar{e}_n) - L(\frac{P}{Q}) \\ &= \sum_{x\in\Omega} Q(x)\bar{e}_n(x) - \sum_{x\in\Omega} P(x) log(\frac{\bar{e}_n(x) + \frac{P(x)}{Q(x)}}{\frac{P(x)}{Q(x)}}) \\ &= \sum_{x\in\Omega} Q(x)(\bar{e}_n(x) - \frac{P(x)}{Q(x)}\log(1 + \frac{\bar{e}_n(x)}{\frac{P(x)}{Q(x)}})) \end{aligned} \tag{21}$$

Assuming that n is sufficiently large such that $|\frac{\bar{e}_n}{g^*}| \leq 1$. We can now use second order Taylor approximation for $\log(1+x)$ for $|x| < 1$

$$\begin{aligned} L(\hat{f}_n) - L(\frac{P}{Q}) = &\sum_{x\in\Omega} Q(x)(\bar{e}_n(x) \\ &- \frac{P(x)}{Q(x)} \cdot \left(\frac{\bar{e}_n(x)}{\frac{P(x)}{Q(x)}} - \frac{1}{2}(\frac{\bar{e}_n(x)}{\frac{P(x)}{Q(x)}})^2)\right) \\ = &\sum_{x\in\Omega} Q(x)\frac{1}{2}(\frac{\bar{e}_n(x)^2}{\frac{P(x)}{Q(x)}}) \end{aligned} \tag{22}$$

Combining equations 19 with the simplified LHS above, we can bound the error with probability at least $1-\delta$ that,

$$\sum_{x\in\Omega} Q(x)\frac{1}{2}(\frac{\bar{e}_n(x)^2}{\frac{P(x)}{Q(x)}}) \leq 4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_\mathcal{F}\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} \tag{23}$$

Under assumption 1 and 2 $\exists \tilde{x} \in \Omega$ such that $|\bar{e}_n(\tilde{x})| = \|\hat{f}_n - \frac{P}{Q}\|_\infty$. Thus, the equation above can be re-written as

$$\begin{aligned}
\frac{1}{K}\|\bar{e}_n\|_\infty^2 &\leq 2\frac{P(\tilde{x})}{Q(\tilde{x})}\left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_{\mathcal{F}}\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right) \\
\|\bar{e}_n\|_\infty^2 &\leq 2K \cdot \|\frac{P}{Q}\|_\infty\left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_{\mathcal{F}}\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right)
\end{aligned} \tag{24}$$

Where $Q(\tilde{x}) = \frac{1}{K}$. The last inequality comes from the fact that $\frac{P(\tilde{x})}{Q(\tilde{x})} \leq \sup_{x\in\Omega}\frac{P(x)}{Q(x)} = \|\frac{P}{Q}\|_\infty$. This completes the proof. $\square$

Using equation 4 we can upper bound the performance of our estimator as follows,

$$\begin{aligned}
&|\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}}, r\sim R(s,a)}[\hat{w}_n \cdot r] - J_P(\pi)| \leq \max_{q\in\mathcal{Q}}|L_w(\hat{w}_n, \frac{d^\pi_P}{\mu}, q)| \\
&\hat{w}_n = \arg\min_{w\in\mathcal{W}}\max_{q\in\mathcal{Q}} L_{n,w}(w, \hat{\beta}, q)
\end{aligned} \tag{25}$$

We also approximate $\frac{d^\pi_{P_{tr}}}{\mu} \sim \hat{\beta}$. This can be written as follows,

$$\hat{\beta} = \arg\max_{f\in\mathcal{F}}\frac{1}{n}\sum_i \ln f(x_i) - \frac{1}{m}\sum_j f(\tilde{x}_j) + \frac{\lambda}{2}I(f)^2, \tag{26}$$

where $I(f)$ is some regularization function to improve the statistical and computational stability of learning. We can the simplify the RHS of this upper-bound using the following simplification.

$$\begin{aligned}
&|\mathbb{E}_{(s,a)\sim d^\pi_{\mathcal{P}'}, r\sim R(s,a)}[\hat{w}_n \cdot r] - J_P(\pi)| \leq \max_{q\in\mathcal{Q}}|L_w(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| \\
&\leq \max_{q\in\mathcal{Q}}|L_w(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_{n,w}(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| + \max_{q\in\mathcal{Q}}|L_{n,w}(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| + \\
&\max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \hat{\beta}, q)| + \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \hat{\beta}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| + \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| \\
&\leq \max_{q\in\mathcal{Q}}|L_w(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_{n,w}(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| + \max_{q\in\mathcal{Q}}|L_{n,w}(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P}}{\mu}, q)| \\
&+ \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \hat{\beta}, q)| + \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \hat{\beta}, q)| - \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| + \max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q)| \\
&\leq \underbrace{2\max_{q\in\mathcal{Q}, w\in\mathcal{W}}||L_w(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| - |L_n(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)||}_{T1} + \underbrace{2\max_{q\in\mathcal{Q}}|L_w(\hat{w}, \frac{d^\pi_{P'}}{\mu}, q) - L_w(\hat{w}, \hat{\beta}, q)|}_{T2} + \min_{w\in\mathcal{W}}\max_{q\in\mathcal{Q}}|L_w(w, \frac{d^\pi_{P_{tr}}}{\mu}, q)|
\end{aligned}$$

Where, $\hat{w} = \arg\min_{w\in\mathcal{W}}\max_{q\in\mathcal{Q}}|L_w(w, \hat{\beta}, q)|$. Let's analyse each of the terms above one by one. Starting with T1 we get the following,

$$\begin{aligned}
T1 &= 2\max_{q\in\mathcal{Q}, w\in\mathcal{W}}|L_w(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| - |L_{n,w}(\hat{w}_n, \frac{d^\pi_{P'}}{\mu}, q)| \\
&\leq 2\mathcal{R}_n(\mathcal{W}, \mathcal{Q}) + C_{\mathcal{W}} \cdot C_{\mathcal{Q}}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \quad \text{w.p at-least } 1 - \frac{\delta}{2}
\end{aligned} \tag{27}$$

Where, the upper bound follows from [29]. Note that $\mathcal{R}_n(\mathcal{W}, \mathcal{Q})$ is the Radamacher Complexity for the following function class

$$\{(s, a, s') \to w(s, a)\frac{d^\pi_{P_{tr}}(s, a)}{\mu(s, a)}(q(s, a) - \gamma q(s', \pi)) : w \in \mathcal{W}, q \in \mathcal{Q}\} \tag{28}$$

For the term T2 we can simplify the expression as follows,

$$\begin{aligned}
T2 &= 2\max_{q\in\mathcal{Q}} |L_w(\hat{w}, \frac{d^\pi_{P_{te}}}{\mu}, q) - L_w(\hat{w}, \hat{\beta}, q)| \\
&= \max_{q\in\mathcal{Q}} |\mathbb{E}_{(s,a,s')\sim\mu}[(\hat{\beta} - \frac{d^\pi_{P'}}{\mu}) \cdot \hat{w}(s,a) \cdot (q(s,a) - \gamma q(s',\pi))]| \\
&= \max_{q\in\mathcal{Q}} |\mathbb{E}_{(s,a,s')\sim\mu}[\varepsilon(s,a) \cdot \hat{w}(s,a) \cdot (q(s,a) - \gamma q(s',\pi))]|. \leq 2C_{\mathcal{Q}} \cdot C_{\mathcal{W}} \|\varepsilon\|_\infty
\end{aligned} \tag{29}$$

Here, we assume that $\varepsilon(s,a) = \hat{\beta} - \frac{d^\pi_{P'}}{\mu}$.Combining equations 27, 29 along with equation 24 we get the following upper-bound with at-least $1-\delta$

$$|\mathbb{E}_{(s,a)\sim d^\pi_P, r\sim R(s,a)}[\hat{w}_n \cdot r] - J_P(\pi)| \leq \max_{q\in\mathcal{Q}} |L_w(\hat{w}, \frac{d^\pi_{P_{tr}}}{\mu}, q)| + 4\gamma C_{\mathcal{W}} \cdot C_{\mathcal{Q}} \cdot \|\varepsilon\|_\infty + 2\mathcal{R}_n(\mathcal{W},\mathcal{Q}) + C_{\mathcal{W}} \cdot C_{\mathcal{Q}} \sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \tag{30}$$

Using Lemma 10.6 we can bound $\|\varepsilon\|_\infty$ with probability $1 - \frac{\delta}{2}$ as follows,

$$\begin{aligned}
&|\mathbb{E}_{(s,a)\sim d^\pi_P, r\sim R(s,a)}[\hat{w}_n \cdot r] - J_P(\pi)| \leq \min_{w\in\mathcal{W}} \max_{q\in\mathcal{Q}} |L_w(w, \frac{d^\pi_{P_{tr}}}{\mu}, q)| \\
&+ 4C_{\mathcal{W}} \cdot C_{\mathcal{Q}} \cdot \sqrt{2K \cdot \|\frac{d^\pi_{P_{tr}}}{\mu}\|_\infty \left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_{\mathcal{F}}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}\right)} + 4\mathcal{R}_n(\mathcal{W},\mathcal{Q}) + 2C_{\mathcal{W}} \cdot C_{\mathcal{Q}} \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}
\end{aligned} \tag{31}$$

This completes the proof.

10.9 Proof of Theorem 10.5

Using equation 11, we can bound the performance of the q estimator as follows,

$$\begin{aligned}
&|(1-\gamma)\mathbb{E}_{d_0}[\hat{q}_n(s,\pi)] - J_P(\pi)| \leq \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| \\
&\leq \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| - \max_{w\in\mathcal{W}} |L_{n,q}(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| + \max_{w\in\mathcal{W}} |L_{n,q}(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| - \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| \\
&+ \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| - \max_{w\in\mathcal{W}} |L_q(w, \hat{\beta}, \hat{q})| + \max_{w\in\mathcal{W}} |L_q(w, \hat{\beta}, \hat{q})| - \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| + \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| \\
&\leq \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| - \max_{w\in\mathcal{W}} |L_{n,q}(w, \frac{d^\pi_{P_{te}}}{\mu}, \hat{q}_n)| + \max_{w\in\mathcal{W}} |L_{n,q}(w, \frac{d^\pi_{P}}{\mu}, \hat{q}_n)| - \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q}_n)| \\
&+ \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| - \max_{w\in\mathcal{W}} |L_q(w, \hat{\beta}, \hat{q})| + \max_{w\in\mathcal{W}} |L_q(w, \hat{\beta}, \hat{q})| - \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| + \max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q})| \\
&\leq 2 \underbrace{\max_{q\in\mathcal{Q}, w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{te}}}{\mu}, q) - L_{n,q}(w, \frac{d^\pi_{P_{te}}}{\mu}, q)|}_{T1} + 2 \underbrace{\max_{w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, \hat{q}) - L_q(w, \hat{\beta}, \hat{q})|}_{T2} + \min_{q\in\mathcal{Q}} \max_{w\in\mathcal{W}} L_q(w, \frac{d^\pi_{P_{tr}}}{\mu}, q)
\end{aligned} \tag{32}$$

Where, $\hat{q} = \arg\min_{q\in\mathcal{Q}} \max_{q\in\mathcal{W}} |L_q(w, \hat{\beta}, q)|$. Lets analyse each of these terms $T1, T2$ separately. For T1 we get the following,

$$\begin{aligned}
T1 &= 2 \max_{q\in\mathcal{Q}, w\in\mathcal{W}} |L_q(w, \frac{d^\pi_{P_{te}}}{\mu}, q) - L_{n,q}(w, \frac{d^\pi_{P_{te}}}{\mu}, q)| \\
&\leq 2\mathcal{R}_n(\mathcal{W},\mathcal{Q}) + C_{\mathcal{W}} \cdot \frac{R_{\max}}{1-\gamma} \sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \quad \text{w.p at-least } 1 - \frac{\delta}{2}
\end{aligned} \tag{33}$$

Where, the upper bound follows from [29]. Note that $\mathcal{R}_n(\mathcal{W}, \mathcal{Q})$ is the Radamacher Complexity for the following function class

$$\{(s,a,s') \to w(s,a)\frac{d^{\pi}_{P_{tr}}(s,a)}{\mu(s,a)}(q(s,a) - \gamma q(s',\pi)) : w \in \mathcal{W}, q \in \mathcal{Q}\} \tag{34}$$

For the term T2 we can simplify the expression as follows,

$$\begin{aligned} T2 &= 2\max_{w\in\mathcal{W}} |L_q(w, \frac{d^{\pi}_{P_{tr}}}{\mu}, \hat{q}) - L_q(w, \hat{\beta}, \hat{q})| \\ &= \max_{w\in\mathcal{W}} |\mathbb{E}_{(s,a,s')\sim\mu}[(\hat{\beta} - \frac{d^{\pi}_{P'}}{\mu}) \cdot w(s,a) \cdot (\hat{q}(s,a) - \gamma\hat{q}(s',\pi)]| \\ &\leq 2C_{\mathcal{W}}\frac{R_{max}}{1-\gamma}\|\varepsilon\|_{\infty} \end{aligned} \tag{35}$$

Combining equation 33 and 35 along with equation 24 we can bound the error in evaluation as follows,

$$\begin{aligned} &|(1-\gamma)\mathbb{E}_{d_0}[\hat{q}_n(s,\pi)] - J_P(\pi)| \leq \min_{q\in\mathcal{Q}}\max_{w\in\mathcal{W}} L_q(w, \frac{d^{\pi}}{\mu}, q) + 2\mathcal{R}_n(\mathcal{W},\mathcal{Q}) + 4C_{\mathcal{W}} \cdot \frac{R_{\max}}{1-\gamma}\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \\ &+ 2C_{\mathcal{W}}\frac{R_{max}}{1-\gamma}\sqrt{2K \cdot \|\frac{d^{\pi}_{P_{tr}}}{\mu}\|_{\infty}\left(4\mathbb{E}\mathcal{R}_n(\mathcal{F}) + C_{\mathcal{F}}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}}\right)} \\ &\text{w.p at-least } 1-\delta \end{aligned} \tag{36}$$

This completes the proof

10.10 Additional Experimental Details and Additional Results

Experiment Setup We conduct experiments on both Sim2Sim and Sim2Real environments. For Sim2Sim experiments we demonstrate our results over a range of different types of environments like Tabular (Taxi), Discrete-control (cartpole) and continuous control (Reacher and Halfcheetah). For the Sim2Sim experiments over a diverse set of simulation and the real world environments like gravity, arm-length, friction and maximum torque. For all the experiments we mention here, we will first generate an offline data which was collected using known behavior policy μ. For the sake of these experiments, behavior policy are parameterized by a factor δ which basically dictates the amount of noise added to a pre-trained model. We similarly parameterise the target policy target policy by α. We experiment over different pairs of training and test environments. We typically keep the simulation environment fixed and vary the test environment. We call the key parametric difference between the training and the test environment as the Sim2Real gap. Detailed information for each set of experiments is provided below.
Learning β: We parameterize β as two-layered neural network with ReLU activation layers for intermediate layers. We experimented with two different kinds of final activation layer, squared and tanh. We observed that tanh layer scaled to go from 0 to 10 worked best for these set of experiments.
Learning w: For most of our experiments on β-DICE we use the framework of GradientDICE. GradientDICE algorithms are typically two layered neural networks which use orthogonal initialisation. Inner activation is ReLU and the final activation layer is linear.
Baselines We compare with the following baselines:

- Simulator: This is the baseline of trusting the simulator's evaluation and not using data from the test environment.
- Model-free MIS: We include DualDICE, GradientDICE, GenDICE [21, 23, 22] as state-of-the-art baselines for model-free MIS, which only uses data from the test environment and does not use simulator information.

- Residual dynamics: We fit a model for OPE from test-environment data with the simulator as the "base" prediction and only learn a correction term.

- DR-DICE [24]: the previous baselines ignore some of the available information (e.g., model-free MIS does not use simulator information) or use them in a naïve manner. Therefore, we additionally include a doubly-robust (DR) MIS estimator [24] that can organically blend the model information with the test-environment data.

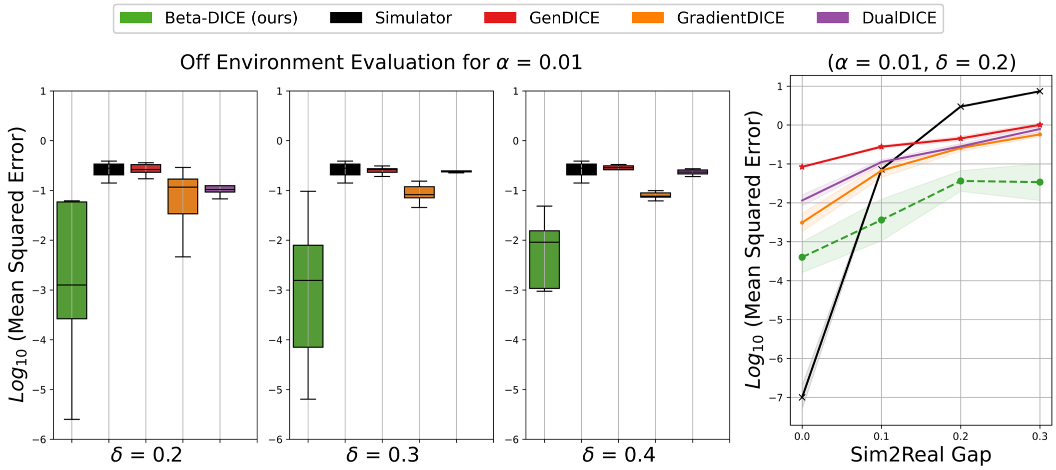
Taxi Environment: Taxi environment has 500 states and 6 discrete actions. For these set of experiments the simulator environment involves deterministic transition between two states. For the real world environment, we experiment with environments where the transition is deterministic with probability $(1-\tau)$ and random with probability τ. With τ being the Sim2Real gap. To collect data, we use a behavior policy that chooses optimal action (which was learnt using Q-learning) with probability $1-\delta$ and a random action with probability δ. Target policies are similarly parameterised but with α. In figure 3 we demonstrate the performance of β-DICE for $\alpha = 0.1$. In these set of experiments, we evaluate performance of β-DICE over 3 different types of behavior policies $\delta = \{0.2, 0.3, .0.4\}$ and three different sets of target policies $\alpha = \{0.01, 0.1, 0.2\}$. For two sets of behavior and target policy, we also show the effect of sim gap on evaluation error. We observe that evaluation error increases with increasing sim2sim gap. For these set of experiments we used a discounting factor $\gamma = 0.9$ and limited our offline trajectory collection to 150 timesteps. Learning rate for β is 1e-4, the learning rate for w is 1e-4. We observe that β-DICE is able to outperform the state-of-the-art MIS baseline comfortably.

Cartpole Environment: For discrete control problems, we choose the Cartpole environment [30]. For the simulator we choose cartpole environment with gravity equals to $10m/s^2$. For the test environment, we choose gravity to be $(\tau)m/s^2$. With τ being the Sim2Sim gap. Our behavior policy is chosen to be a mixture of optimal policy (which was trained using Cross Entropy method) π_* and a uniformly random policy U such that $\mu = (1-\delta)\pi_* + (\delta)U$. Our target policy is similarly parameterised by α. We demonstrate results over different sets of behavior policies $\delta = \{0.4, 0.5, 0.6\}$ and evaluate performance over a set of $\alpha = \{0.2, 0.5, 0.8\}$ and simreal gap $\tau = \{5, 10, 20\}m/s^2$. In figures 2a and 4 (with additional baselines), we demonstrate our experiments over different sets of behavior policies and target policies and observe that our method is more than capable of improving upon state-of-the-art baseline with information from simulation. Our discounting factor $\gamma = 0.99$ and timesteps is limited to 200. Learning rate for β is 1e-4 and learning rate for w is 1e-2.

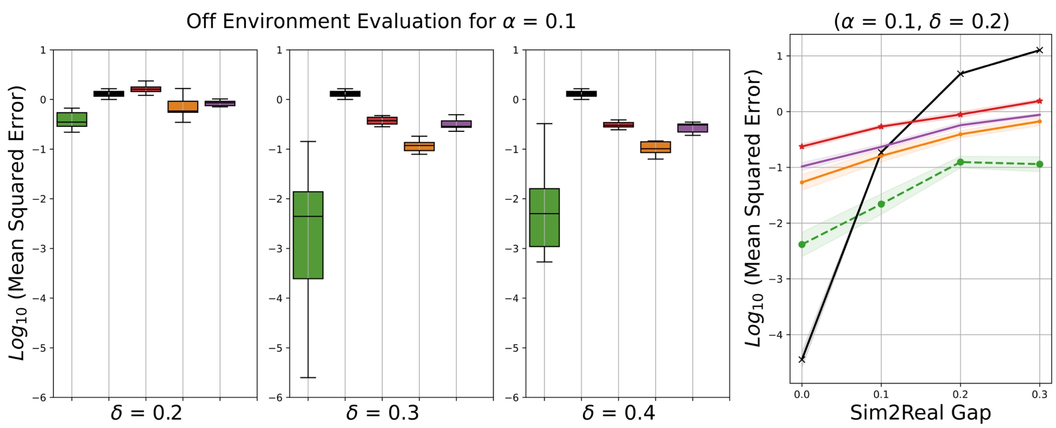
Reacher Environment: For continuous control, we experiment with RoboschoolReacher environment. For these set of environments, we choose training environment as the one where the length of both links are 0.1 m. The test environment is chosen to be one, where the length of both the links is $(0.1+\tau)m$. We choose behavior policy as the addition of an optimal policy plus a zero mean normal policy whose standard deviation is δ. For our experiments, $\delta = \{0.4, 0.5, 0.6\}$, $\alpha = \{0.0, 0.1, 0.2\}$ and $\tau = \{-0.5, -0.25, 0.0, 0.25\}m$. In figures 2b and 5, we demonstrate our experiments over different sets of behavior policies and target policies and observe that our method is more than capable of improving upon state-of-the-art baseline with information from simulation. In figure 2b and 5a, we also demonstrate the effect of β-DICE with sim2sim gap over two sets of (δ, α). Our discounting factor $\gamma = 0.99$ and timesteps is limited to 150. Learning rate for β is 1e-4 and learning rate for w is 3e-3.

HalfCheetah Environment: For continuous control, we experiment with RoboschoolCheetah environment. For these set of environments, we choose training environment as the one where the maximum torque to the joints is 0.9. The test environment is chosen to be one, where the length of both the links is $0.9+\tau$ N.m We choose behavior policy as the addition of an optimal policy with zero mean normal policy whose standard deviation is δ. For behavior policy the delta is taken to be $\delta = \{0.4, 0.5, 0.6\}$ and the target policy is taken to be $\alpha = \{0.0, 0.1, 0.2\}$. Due to limited computation, we experimented only with $\tau = 0.4Nm$. In figures 6, we demonstrate our experiments over different sets of behavior policies and target policies and observe that our method is more than capable of improving upon state-of-the-art baseline with information from simulation.

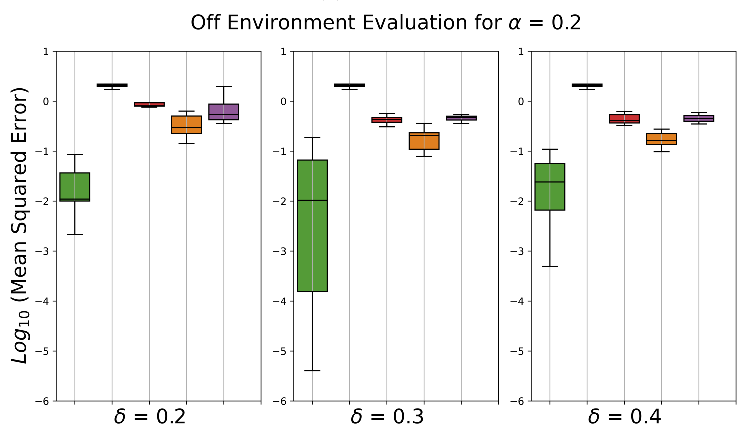
Our discounting factor $\gamma = 0.99$ and timesteps is limited to 150. Learning rate for β, w is 1e-4.



(a) $\alpha = 0.01$

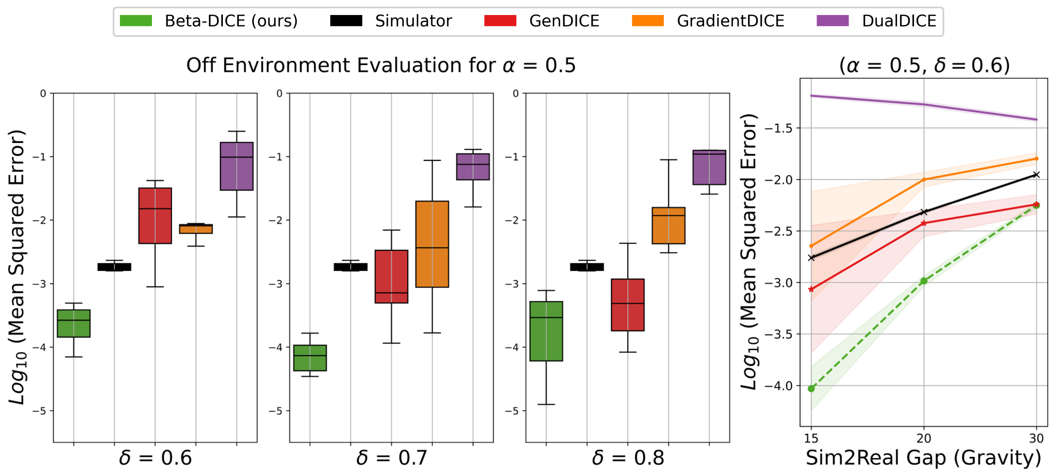


(b) $\alpha = 0.1$

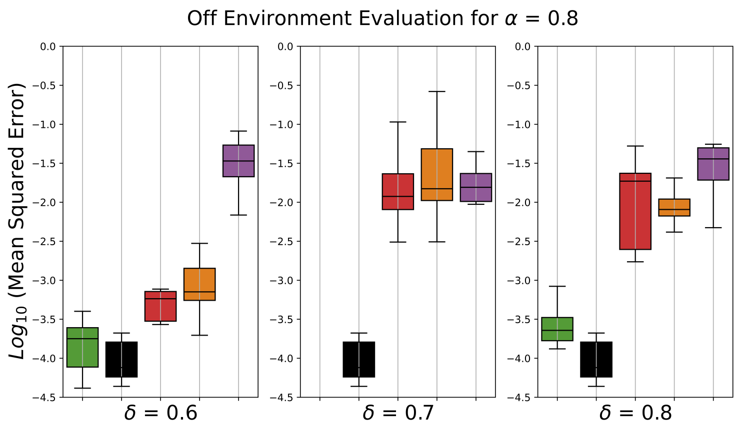


(c) $\alpha = 0.2$

Figure 3: Each of the above figure demonstrates the effect of evaluation over varying behavior policies $\delta = \{0.2, 0.3, 0.4\}$ on a fixed target policy using β-DICE for the taxi environment. For these set of experiments the training environment is the default transition parameters, while the test environment has $\tau = 0.1$. In (a), (b), (c) the target policies that we use are $\alpha = \{0.01, 0.1, 0.2\}$. Additionally for (a), (b) (RHS) we also show the effect of varying sim2real gap on target policy evaluation using β-DICE (while keeping δ, α fixed).

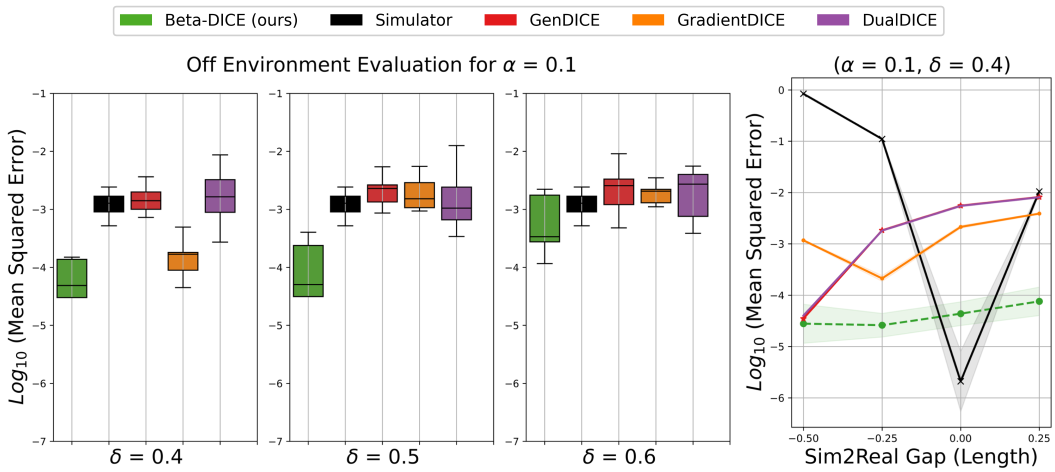


(a) $\alpha = 0.5$

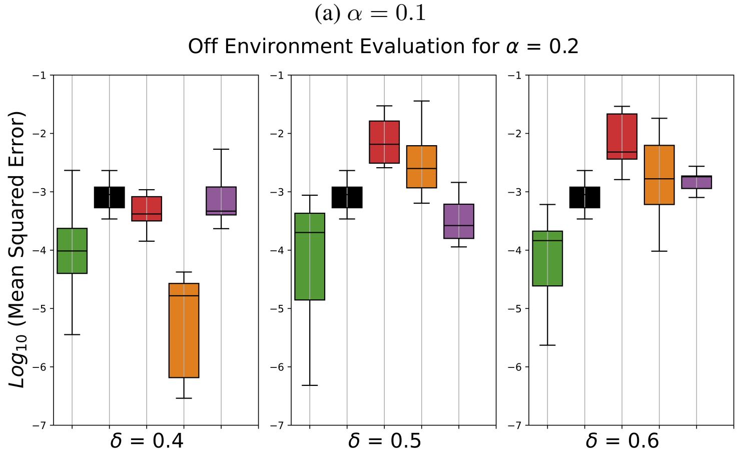


(b) $\alpha = 0.8$

Figure 4: Each of the above figure demonstrates the effect of evaluation over varying behavior policies $\delta = \{0.6, 0.7, 0.8\}$ on a fixed target policy using β-DICE for the cartpole environment. For these set of experiments the training environment has gravity = $10m/s^2$, while the test environment has gravity = $15.0m/s^2$. In (a), (b) the target policies that we use are $\alpha = \{0.5, 0.8\}$. Additionally for (a), (RHS) we also show the effect of varying sim2real gap on target policy evaluation using β-DICE (while keeping δ, α fixed).

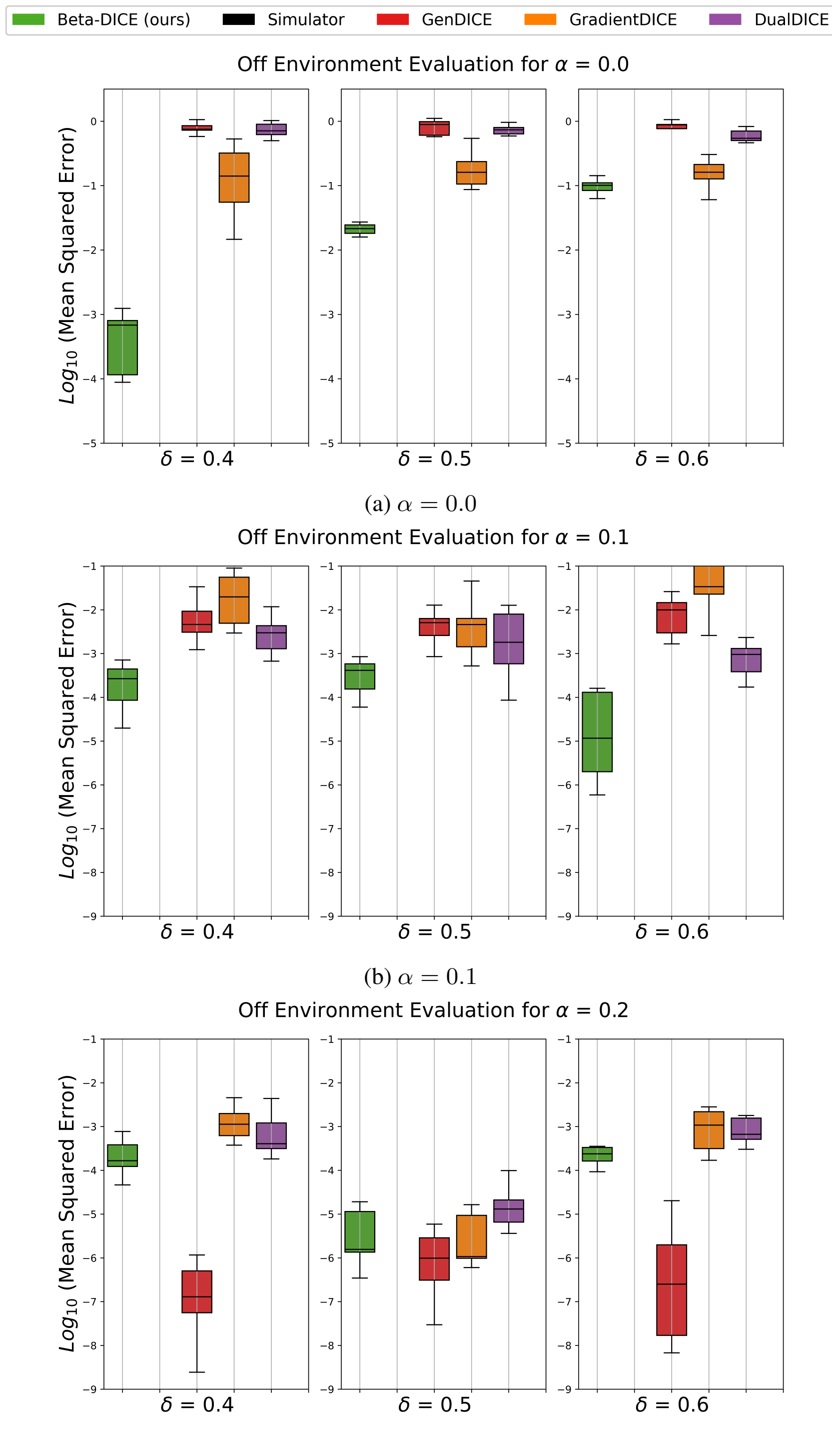


(a) $\alpha = 0.1$



(b) $\alpha = 0.2$

Figure 5: Each of the above figure demonstrates the effect of evaluation over varying behavior policies $\delta = \{0.4, 0.5, 0.6\}$ on a fixed target policy using β-DICE for the reacher environment. For these set of experiments the training environment has length = $0.1m$, while the test environment has length = $0.075m$. In (a), (b) the target policies that we use are $\alpha = \{0.1, 0.2\}$. Additionally for (a), (RHS) we also show the effect of varying sim2real gap on target policy evaluation using β-DICE (while keeping δ, α fixed).



(a) $\alpha = 0.0$

(b) $\alpha = 0.1$

(c) $\alpha = 0.2$

Figure 6: Each of the above figure demonstrates the effect of evaluation over varying behavior policies $\delta = \{0.4, 0.5, 0.6\}$ on a fixed target policy using β-DICE for the half cheetah environment. For these set of experiments the training environment has length = $0.9Nm$, while the test environment has length = $1.3Nm$. In (a), (b), (c) the target policies that we use are $\alpha = \{0.0, 0.1, 0.2\}$