A. Derivation of Shrinkage Estimators for Non-combinatorial Setting

In this section we provide detailed derivations for the two estimators in the non-combinatorial setting.

We first derive the pessimistic version. Recall that the optimization problem decouples across (x, a), so we focus on a single (x, a) pair such that $\mu(a \mid x) > 0$ since only such pairs can appear in the data. For conciseness, we omit the dependence on (x, a) and simply write $w = w(x, a)$, $\hat{w} = \hat{w}(x, a)$ and $\mu = \mu(a \mid x)$. Fixing $\lambda \geq 0$, we must solve

$$\operatorname*{Minimize}_{\hat{w} \in \mathbb{R}} \Bigl[\mu \hat{w}^2 + 2\lambda \left|\mu(\hat{w} - w)\right|\Bigr].$$

(Note that we allow any $\hat{w} \in \mathbb{R}$, but we will see that the solution will actually satisfy $0 \leq \hat{w} \leq w$.) Since $\mu > 0$, this minimization problem is strongly convex and therefore has a unique minimizer. By first-order optimality, $\hat{w}$ is a minimizer if and only if

$$2\mu\hat{w} + 2\lambda\mu v = 0 \qquad \text{and} \qquad v \in \partial\left|\hat{w} - w\right| = \begin{cases} 1 \text{ if} & \hat{w} > w, \\ [-1, 1] \text{ if} & \hat{w} = w, \\ -1 \text{ if} & \hat{w} < w. \end{cases} \tag{11}$$

Since $\mu > 0$, the first equation can be rewritten as

$$\hat{w} = -\lambda v.$$

Now a simple case analysis shows that if $w > \lambda$ then the choice $\hat{w} = \lambda$, $v = -1$ satisfies Eq. (11), and if $0 \leq w \leq \lambda$ then the choice $\hat{w} = w$, $v = -w/\lambda$ satisfies Eq. (11), yielding

$$\hat{w}_{\mathrm{p},\lambda}(x, a) = \min\{\lambda, w(x, a)\},$$

which is the clipped estimator.

For the optimistic version, the optimization problem is

$$\operatorname*{Minimize}_{\hat{w} \in \mathbb{R}} \Bigl[\mu \hat{w}^2 w^2 / z + \lambda\mu(\hat{w} - w)^2 / z\Bigr],$$

where $z = z(x, a)$. The optimality conditions are

$$2\mu w^2 \hat{w}/z + 2\lambda\mu(\hat{w} - w)/z = 0.$$

This gives the optimistic estimator

$$\hat{w}_{\mathrm{o},\lambda}(x, a) = \frac{\lambda}{w(x, a)^2 + \lambda} w(x, a).$$

Notice that this estimator does not depend on the weighting function z, so it does not depend on how we train the regression model.

B. Derivation of the Shrinkage Estimator for Combinatorial Setting

We provide a complete derivation of the shrinkage estimator for combinatorial actions. We use the notation $w(x, \mathbf{a}) = \mathbf{w}(x)^\top \mathbf{a}$. We assume that the regression model takes form $\hat{\eta}(x, \mathbf{a}) = \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a}$, satisfies $\hat{\eta}(x, \mathbf{a}) \in [0, 1]$, and is trained to minimize

$$L(\hat{\boldsymbol{\eta}}) \coloneqq \mathop{\mathbb{E}}_{\mu}\left[\, z(x, \mathbf{a})(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a})^2 \,\right]$$

for some $z(x, \mathbf{a}) > 0$. We assume that the linearity assumption holds, so we can write $\eta(x, \mathbf{a}) = \boldsymbol{\eta}(x)^\top \mathbf{a}$. And we also assume that $\operatorname{span}\bigl(\operatorname{supp} \pi(\cdot \mid x)\bigr) \subseteq \operatorname{span}\bigl(\operatorname{supp} \mu(\cdot \mid x)\bigr)$, so, as shown by Swaminathan et al. (2017), the pseudo-inverse estimator is unbiased:

$$\mathop{\mathbb{E}}_{(x,\mathbf{a},r)\sim\mu}\Bigl[\hat{\boldsymbol{\eta}}(x)^\top \mathbf{q}_{\pi,x} + w(x, \mathbf{a})\bigl(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a}\bigr)\Bigr] = \mathop{\mathbb{E}}_{(x,\mathbf{a},r)\sim\pi}[r]. \tag{12}$$

Therefore, if we replace w by an arbitrary function $\hat{w}$ (not necessarily linear in $\mathbf{a}$), we obtain the expression for the bias

$$\mathrm{Bias}(\hat{w}) = \mathop{\mathbb{E}}_{\mu}\left[\,\big(\hat{w}(x,\mathbf{a}) - w(x,\mathbf{a})\big)(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a})\,\right]. \tag{13}$$

Using Cauchy–Schwarz inequality, we can bound the bias in terms of $L(\hat{\boldsymbol{\eta}})$:

$$\mathrm{Bias}(\hat{w}) \le \sqrt{\mathop{\mathbb{E}}_{\mu}\left[\,\big(\hat{w}(x,\mathbf{a}) - w(x,\mathbf{a})\big)^2 / z(x,\mathbf{a})\,\right]} \cdot \sqrt{L(\hat{\boldsymbol{\eta}})}.$$

For the variance bound, we begin with a proxy based on Proposition 2, and then bound it using Cauchy–Schwarz inequality, the fact that $\hat{\boldsymbol{\eta}}(x)^\top \mathbf{a}$ and r are bounded in $[0,1]$, and an additional assumption that $|\hat{w}(x,\mathbf{a})| \le |w(x,\mathbf{a})|$ (which we will show is true for the specific optimistic estimator that we derive below):

$$\begin{aligned}\mathrm{Var}(\hat{w}) &\approx \frac{1}{n}\mathop{\mathbb{E}}_{\mu}\left[\,\hat{w}(x,\mathbf{a})^2\big(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a}\big)^2\,\right] \le \frac{1}{n}\sqrt{\mathop{\mathbb{E}}_{\mu}\Big[\hat{w}(x,\mathbf{a})^4 / z(x,\mathbf{a})\Big]} \cdot \sqrt{\mathop{\mathbb{E}}_{\mu}\left[\,z(x,\mathbf{a})\big(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a}\big)^4\,\right]} \\ &\le \frac{1}{n}\sqrt{\mathop{\mathbb{E}}_{\mu}\Big[\hat{w}(x,\mathbf{a})^2 w(x,\mathbf{a})^2 / z(x,\mathbf{a})\Big]} \cdot \sqrt{L(\hat{\boldsymbol{\eta}})}.\end{aligned}$$

Similar to non-combinatorial setting, the solutions of the resulting MSE bound must lie on the Pareto front parameterized by a single scalar $\lambda \in [0,\infty]$:

$$\mathop{\mathrm{Minimize}}_{\hat{w}} \;\; \lambda \mathop{\mathbb{E}}_{\mu}\left[\,\tfrac{1}{z(x,\mathbf{a})}\big(\hat{w}(x,\mathbf{a}) - w(x,\mathbf{a})\big)^2\,\right] + \mathop{\mathbb{E}}_{\mu}\left[\,\tfrac{w(x,\mathbf{a})^2}{z(x,\mathbf{a})}\hat{w}(x,\mathbf{a})^2\,\right].$$

This decomposes across $(x,\mathbf{a})$ and by first-order optimality, we obtain the same solution as in non-combinatorial setting:

$$\hat{w}(x,\mathbf{a}) = \frac{\lambda}{w(x,\mathbf{a})^2 + \lambda} w(x,\mathbf{a}).$$

Note that these weights satisfy $|\hat{w}(x,\mathbf{a})| \le |w(x,\mathbf{a})|$, and $\hat{w}$ matches the sign of w, so in fact a stronger shrinkage property holds: $\hat{w}(x,\mathbf{a}) = c(x,\mathbf{a})w(x,\mathbf{a})$ where $c(x,\mathbf{a}) \in [0,1]$. Also note that when μ is supported on a linearly independent set of actions for any given x, then we can pick $\hat{\mathbf{w}}(x) \in \mathbb{R}^d$ to satisfy $\hat{w}(x,\mathbf{a}) = \hat{\mathbf{w}}(x)^\top \mathbf{a}$ across all actions in the support of $\mu(\cdot \mid x)$, thus satisfying the assumptions of Proposition 2. Plugging the expression for $\hat{w}$ back into the pseudo-inverse estimator yields

$$\hat{V}_{\text{DRos-PI}}(\pi; \hat{\boldsymbol{\eta}}, \lambda) := \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\eta}}(x_i)^\top \mathbf{q}_{\pi,x_i} + \left(\frac{\lambda}{\lambda + (\mathbf{w}(x_i)^\top \mathbf{a}_i)^2}\right)\mathbf{w}(x_i)^\top \mathbf{a}_i (r_i - \hat{\boldsymbol{\eta}}(x_i)^\top \mathbf{a}_i).$$

C. Proofs

C.1. Proof of Proposition 1

The law of total variance gives

$$\begin{aligned}\mathrm{Var}(\hat{w}) &= \frac{1}{n}\mathop{\mathrm{Var}}_{x,a,r\sim\mu}\left(\sum_{a'\in\mathcal{A}} \pi(a' \mid x)\hat{\eta}(x,a') + \hat{w}(x,a)(r - \hat{\eta}(x,a))\right) \\ &= \underbrace{\frac{1}{n}\mathop{\mathbb{E}}_{x}\mathop{\mathrm{Var}}_{a,r\sim\mu}\left(\sum_{a'\in\mathcal{A}} \pi(a' \mid x)\hat{\eta}(x,a') + \hat{w}(x,a)(r - \hat{\eta}(x,a))\right)}_{=:T_1} \\ &\quad + \underbrace{\frac{1}{n}\mathop{\mathrm{Var}}_{x}\mathop{\mathbb{E}}_{a,r\sim\mu}\left(\sum_{a'\in\mathcal{A}} \pi(a' \mid x)\hat{\eta}(x,a') + \hat{w}(x,a)(r - \hat{\eta}(x,a))\right)}_{=:T_2}.\end{aligned}$$

For T_1, since $\sum_{a' \in \mathcal{A}} \pi(a' \mid x)\hat{\eta}(x, a')$ does not depend on a, r, it does not contribute to the conditional variance, and we get

$$T_1 = \mathop{\mathbb{E}}_{x} \mathop{\mathrm{Var}}_{a,r} \left(\hat{w}(x,a)(r - \hat{\eta}(x,a)) \right) = \mathop{\mathbb{E}}_{x,a,r} \left[\hat{w}(x,a)^2 (r - \hat{\eta}(x,a))^2 \right] - \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{a,r} \left[\hat{w}(x,a)(r - \hat{\eta}(x,a)) \right]^2 \right].$$

The first term is our variance proxy. To bound the second term, write $\hat{w}(x,a) = c(x,a)w(x,a)$ for some $c(x,a) \in [0,1]$, which is possible since $0 \le w \le \hat{w}$ by assumption. The second term can then be rewritten and bounded as

$$\begin{aligned} 0 \le \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{a,r \sim \mu} \left[\hat{w}(x,a)\big(r - \hat{\eta}(x,a)\big) \right]^2 \right] &= \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{a,r \sim \mu} \left[w(x,a)c(x,a)\big(r - \hat{\eta}(x,a)\big) \right]^2 \right] \\ &= \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{a,r \sim \pi} \left[c(x,a)\big(r - \hat{\eta}(x,a)\big) \right]^2 \right] \le 1, \end{aligned}$$

where the second equality follows by the unbiasedness of inverse-propensity scoring, and the final bound follows because $c(x,a), r, \hat{\eta}(x,a) \in [0,1]$.

For T_2, of course we have $T_2 \ge 0$, and further

$$\begin{aligned} T_2 &= \mathop{\mathrm{Var}}_{x} \left[\mathop{\mathbb{E}}_{a \sim \pi} \big[\hat{\eta}(x,a)\big] + \mathop{\mathbb{E}}_{a \sim \mu} \big[\hat{w}(x,a)\big(\eta(x,a) - \hat{\eta}(x,a)\big)\big] \right] \\ &\le \mathbb{E}_x \left[\left(\mathop{\mathbb{E}}_{a \sim \pi} [\hat{\eta}(x,a)] + \mathop{\mathbb{E}}_{a \sim \mu} \big[\hat{w}(x,a)\big(\eta(x,a) - \hat{\eta}(x,a)\big)\big] \right)^2 \right] \\ &= \mathbb{E}_x \left[\left(\mathop{\mathbb{E}}_{a \sim \pi} [\hat{\eta}(x,a)] + \mathop{\mathbb{E}}_{a \sim \mu} \big[w(x,a)c(x,a)\big(\eta(x,a) - \hat{\eta}(x,a)\big)\big] \right)^2 \right] \\ &= \mathbb{E}_x \left[\mathop{\mathbb{E}}_{a \sim \pi} \Big[\hat{\eta}(x,a) + c(x,a)\big(\eta(x,a) - \hat{\eta}(x,a)\big)\Big]^2 \right] \\ &= \mathbb{E}_x \left[\mathop{\mathbb{E}}_{a \sim \pi} \Big[\big(1 - c(x,a)\big)\hat{\eta}(x,a) + c(x,a)\eta(x,a)\Big]^2 \right] \le 1, \end{aligned}$$

where we again write $\hat{w}(x,a) = c(x,a)w(x,a)$ for some $c(x,a) \in [0,1]$, then appeal to the unbiasedness of the inverse-propensity scoring, and finally use the bounds $c(x,a), \eta(x,a), \hat{\eta}(x,a) \in [0,1]$.

Therefore, we can write

$$\mathrm{Var}(\hat{w}) - \frac{1}{n} \mathop{\mathbb{E}}_{x,a,r} \left[\hat{w}(x,a)^2 (r - \hat{\eta}(x,a))^2 \right] = -\frac{1}{n} \mathop{\mathbb{E}}_{x} \left[\mathop{\mathbb{E}}_{a,r} \left[\hat{w}(x,a)(r - \hat{\eta}(x,a)) \right]^2 \right] + \frac{1}{n} T_2,$$

and we have just shown that the right hand side is in $\left[-\frac{1}{n}, \frac{1}{n}\right]$. This proves the proposition.

C.2. Proof of Proposition 2

We begin by deriving a simple expression for the pseudo-inverse $\Gamma_{\mu,x}^{\dagger}$. Consider a fixed x and let $s = |\mathcal{B}_x|$ be the size of the basis $\mathcal{B}_x$ (note that s might be a function of x). Let $D_{\mu,x} \in \mathbb{R}^{s \times s}$ denote the diagonal matrix $D_{\mu,x} = \mathrm{diag}\{\mu(\mathbf{a} \mid x)\}_{\mathbf{a} \in \mathcal{B}_x}$. Recall that B_x is the matrix with $\mathbf{a} \in \mathcal{B}_x$ in its columns. The matrix $\Gamma_{\mu,x}$ can then be written as

$$\Gamma_{\mu,x} = B_x D_{\mu,x} B_x^{\top}.$$

To obtain its psedo-inverse, we use tho following fact:

Fact 4. *Let $B \in \mathbb{R}^{d \times s}$ be a matrix with linearly independent columns and let $K = B^{\top} B$. Then for any invertible diagonal matrix $D \in \mathbb{R}^{s \times s}$, we have*

$$(BDB^{\top})^{\dagger} = BK^{-1}D^{-1}K^{-1}B^{\top},$$

where K^{-1} is well defined thanks to the linear independence of columns of B.

Proof. Let $G := BDB^\top$ and $G' := BK^{-1}D^{-1}K^{-1}B^\top$. To show that $G^\dagger = G'$, it suffices to argue that $GG'G = G$ and $G'GG' = G'$:

$$\begin{aligned} GG'G &= BDB^\top BK^{-1}D^{-1}K^{-1}B^\top BDB^\top = BDB^\top = G, \\ G'GG' &= BK^{-1}D^{-1}K^{-1}B^\top BDB^\top BK^{-1}D^{-1}K^{-1}B^\top = BK^{-1}D^{-1}K^{-1}B^\top = G'. \end{aligned}$$ □

Using this fact, we thus have

$$\Gamma^\dagger_{\mu,x} = B_x K_x^{-1} D^{-1}_{\mu,x} K_x^{-1} B_x^\top, \tag{14}$$

where $K_x = B_x^\top B_x$.

We are now ready to start the proof of Proposition 2. Similarly to the proof of Proposition 1, we first apply the law of total variance

$$\begin{aligned} n\,\mathrm{Var}(\hat{w}) &= \operatorname*{Var}_{x,\mathbf{a},r\sim\mu}\left(\,\hat{\boldsymbol{\eta}}(x)^\top \mathbf{q}_{\pi,x} + \hat{\mathbf{w}}(x)^\top \mathbf{a}(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a})\,\right) \\ &= \underbrace{\mathop{\mathbb{E}}_{x}\operatorname*{Var}_{\mathbf{a},r\sim\mu}\left(\,\hat{\boldsymbol{\eta}}(x)^\top \mathbf{q}_{\pi,x} + \hat{\mathbf{w}}(x)^\top \mathbf{a}(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a})\,\right)}_{=:T_1} + \underbrace{\operatorname*{Var}_{x}\mathop{\mathbb{E}}_{\mathbf{a},r\sim\mu}\left(\,\hat{\boldsymbol{\eta}}(x)^\top \mathbf{q}_{\pi,x} + \hat{\mathbf{w}}(x)^\top \mathbf{a}(r - \hat{\boldsymbol{\eta}}(x)^\top \mathbf{a})\,\right)}_{=:T_2}. \end{aligned}$$

To analyze T_2, we first rewrite and bound the inner expectation for a fixed x. We drop the dependence on x from the notation and write $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}(x)$, $\mathbf{q}_\pi = \mathbf{q}_{\pi,x}$, $\hat{\mathbf{w}} = \hat{\mathbf{w}}(x)$, $\boldsymbol{\eta} = \boldsymbol{\eta}(x)$, and $\Gamma_\mu = \Gamma_{\mu,x}$:

$$\begin{aligned} &\mathop{\mathbb{E}}_{\mathbf{a},r\sim\mu}\Big[\hat{\boldsymbol{\eta}}^\top \mathbf{q}_\pi + \hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\Big] \\ &\qquad = \hat{\boldsymbol{\eta}}^\top \mathbf{q}_\pi + \mathop{\mathbb{E}}_{\mathbf{a}\sim\mu}\Big[\hat{\mathbf{w}}^\top \mathbf{a}(\mathbf{a}^\top \boldsymbol{\eta} - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\Big] = \hat{\boldsymbol{\eta}}^\top \mathbf{q}_\pi + \hat{\mathbf{w}}^\top \left(\mathop{\mathbb{E}}_{\mathbf{a}\sim\mu}[\mathbf{a}\mathbf{a}^\top]\right)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) = \hat{\boldsymbol{\eta}}^\top \mathbf{q}_\pi + \hat{\mathbf{w}}^\top \Gamma_\mu(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) \\ &\qquad = \hat{\boldsymbol{\eta}}^\top B\mathbf{v}_\pi + \hat{\mathbf{w}}^\top (BD_\mu B^\top)(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}) \end{aligned} \tag{15}$$

where in the last step we introduced shorthands $B = B_x$, $D_\mu = D_{\mu,x}$ and $\mathbf{v}_\pi = \mathbf{v}_{\pi,x}$.

To continue with the derivation, observe that by assumption, we have $\hat{\mathbf{w}}^\top \mathbf{a} = c(x,\mathbf{a})\mathbf{w}^\top \mathbf{a}$ for all $\mathbf{a} \in \mathcal{B}_x$, and so we can write $\hat{\mathbf{w}}^\top B = \mathbf{w}^\top BC$ where C is a diagonal matrix with entries $c(x,\mathbf{a})$ across $\mathbf{a} \in \mathcal{B}_x$. Next, using the fact that $\mathbf{w} = \Gamma_\mu^\dagger \mathbf{q}_\pi$ and then plugging in Eq. (14), we obtain

$$\hat{\mathbf{w}}^\top B = \mathbf{w}^\top BC = \mathbf{q}_\pi^\top \Gamma_\mu^\dagger BC = \mathbf{v}_\pi^\top B^\top \Big(BK^{-1}D_\mu^{-1}K^{-1}B^\top\Big)BC = \mathbf{v}_\pi^\top D_\mu^{-1} C, \tag{16}$$

where we introduced the shorthand $K = K_x$. Now combining Eqs. (15) and (16), we obtain

$$\begin{aligned} \left|\mathop{\mathbb{E}}_{\mathbf{a},r\sim\mu}\Big[\hat{\boldsymbol{\eta}}^\top \mathbf{q}_\pi + \hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\Big]\right| &= \left|\mathbf{v}_\pi^\top B^\top \hat{\boldsymbol{\eta}} + \mathbf{v}_\pi^\top D_\mu^{-1} C D_\mu B^\top (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})\right| \\ &= \left|\mathbf{v}_\pi^\top \Big(B^\top \hat{\boldsymbol{\eta}} + CB^\top(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})\Big)\right| \\ &\le \|\mathbf{v}_\pi\|_1 \cdot \big\|(I - C)B^\top \hat{\boldsymbol{\eta}} + CB^\top \boldsymbol{\eta}\big\|_\infty, \end{aligned} \tag{17}$$

where in the last step we used Holder's inequality. For the ℓ_∞ norm, we get

$$\big\|(I - C)B^\top \hat{\boldsymbol{\eta}} + CB^\top \boldsymbol{\eta}\big\|_\infty^2 = \max_{\mathbf{a}\in\mathcal{B}_x}\Big|\big(1 - c(x,\mathbf{a})\big)\hat{\eta}(x,\mathbf{a}) + c(x,\mathbf{a})\eta(x,\mathbf{a})\Big| \le 1,$$

where the last step follows because $\eta(x,\mathbf{a}), \hat{\eta}(x,\mathbf{a}), c(x,\mathbf{a}) \in [0,1]$. Therefore, we have that $0 \le T_2 \le \mathbb{E}_x\big[\|\mathbf{v}_{\pi,x}\|_1^2\big]$.

To bound T_1, we first note that $\hat{\boldsymbol{\eta}}(x)^\top \mathbf{q}_{\pi,x}$ is independent of $\mathbf{a}$ and r, and so it does not contribute to the variance, and so

$$T_1 = \mathop{\mathbb{E}}_{x,\mathbf{a},r\sim\mu}\Big[\big(\,\hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\,\big)^2\Big] - \mathop{\mathbb{E}}_{x}\Big[\mathop{\mathbb{E}}_{\mathbf{a},r\sim\mu}\big[\,\hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\,\big]^2\Big].$$

The second term satisfies

$$0 \le \mathop{\mathbb{E}}_{x}\Big[\mathop{\mathbb{E}}_{\mathbf{a},r\sim\mu}\big[\,\hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\,\big]^2\Big] \le \mathop{\mathbb{E}}_{x}\Big[\|\mathbf{v}_{\pi,x}\|_1^2 \cdot \big\|CB^\top(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})\big\|_\infty^2\Big] \le \mathop{\mathbb{E}}_{x}\big[\|\mathbf{v}_{\pi,x}\|_1^2\big],$$

where we applied similar reasoning as in Eq. (17). Combining this bound with the bound on T_2 completes the proof:

$$\left|n\,\mathrm{Var}(\hat{w}) - \mathop{\mathbb{E}}_{x,\mathbf{a},r\sim\mu}\Big[\big(\,\hat{\mathbf{w}}^\top \mathbf{a}(r - \mathbf{a}^\top \hat{\boldsymbol{\eta}})\,\big)^2\Big]\right| \le \mathop{\mathbb{E}}_{x}\big[\|\mathbf{v}_{\pi,x}\|_1^2\big].$$

C.3. Proof of Theorem 3

The main technical part of the proof is a deviation inequality for the sample variance. For this, let us fix θ, which we drop from notation, and focus on estimating the variance

$$\operatorname{Var}(Z) = \mathbb{E}[(Z - \mathbb{E}(z))^2] \text{ with } \widehat{\operatorname{Var}} = \frac{1}{2n(n-1)} \sum_{\substack{1 \le i,j \le n \\ i \ne j}} (Z_i - Z_j)^2.$$

We have the following lemma

Lemma 5 (Variance estimation). *Let $Z_1, \dots, Z_n$ be iid random variables, and assume that $|Z_i| \le R$ almost surely. Then there exists a constant $C > 0$ such that for any $\delta \in (0,1)$, with probability at least $1 - \delta$*

$$\left| \operatorname{Var}(Z) - \widehat{\operatorname{Var}} \right| \le (C+3) \left(\sqrt{\frac{2R^2 \operatorname{Var}(Z) \log(6C/\delta)}{n}} + \frac{2R^2 \log(6C/\delta)}{3n} \right).$$

Proof. For this lemma only, define $\mu = \mathbb{E}[Z]$. By direct calculation

$$\operatorname{Var}(Z) = \mathbb{E}\left[\, Z^2 \,\right] - \mu^2, \qquad \widehat{\operatorname{Var}} = \frac{1}{n} \sum_{i=1}^n Z_i^2 - \frac{1}{n(n-1)} \sum_{i \ne j} Z_i Z_j.$$

We work with the second term first. Let $Z_1', \dots, Z_n'$ be an iid sample, independent of $Z_1, \dots, Z_n$. Now, by Theorem 3.4.1 of De la Pena & Giné (2012), we have

$$\mathbb{P}\left[\left| \frac{1}{n(n-1)} \sum_{i \ne j} (Z_i - \mu)(Z_j - \mu) \right| > t \right] \le C\mathbb{P}\left[\left| \frac{1}{n(n-1)} \sum_{i \ne j} (Z_i - \mu)(Z_j' - \mu) \right| > t/C \right]$$

for a universal constant $C > 0$. Thus, we have decoupled the U-statistic. Now let us condition on $Z_1, \dots, Z_n$ and write $X_j = \frac{1}{n-1} \sum_{i \ne j} (Z_i - \mu)$, which conditional on $Z_1, \dots, Z_n$ is non-random. We will apply Bernstein's inequality on $\frac{1}{n} \sum_{j=1}^n X_j (Z_j' - \mu)$, which is a centered random variable, conditional on $Z_{1:n}$. This gives that with probability at least $1 - \delta$

$$\begin{aligned} \left| \frac{1}{n} \sum_{j=1}^n X_j (Z_j' - \mu) \right| &\le \sqrt{\frac{2 \frac{1}{n} \sum_{j=1}^n \operatorname{Var}(X_j Z_j') \log(2/\delta)}{n}} + \frac{2 \max_j \sup |\, X_j (Z_j - \mu) \,| \log(2/\delta)}{3n}. \\ &\le \max_j |X_j| \left(\sqrt{\frac{2 \operatorname{Var}(Z) \log(2/\delta)}{n}} + \frac{2R \log(2/\delta)}{3n} \right). \end{aligned}$$

This bound holds with high probability for any $\{X_j\}_{j=1}^n$. In particular, since $|X_j| \le R$ almost surely, we get that with probability $1 - \delta$

$$\left| \frac{1}{n(n-1)} \sum_{i \ne j} (Z_i - \mu)(Z_j - \mu) \right| \le C \sqrt{\frac{2R^2 \operatorname{Var}(Z) \log(2C/\delta)}{n}} + \frac{2CR^2 \log(2C/\delta)}{3n}.$$

The factors of C arise from working through the decoupling inequality.

Next, by a standard application of Bernstein's inequality, with probability at least $1 - \delta$, we have

$$\left| \frac{1}{n} \sum_{i=1}^n Z_i - \mu \right| \le \sqrt{\frac{2 \operatorname{Var}(Z) \log(2/\delta)}{n}} + \frac{2R \log(2/\delta)}{3n}.$$

Therefore, with probability $1-2\delta$ we have

$$\begin{aligned}\left|\frac{1}{n(n-1)}\sum_{i\neq j} Z_iZ_j - \mu^2\right| &\leq \left|\frac{1}{n(n-1)}\sum_{i\neq j}(Z_i-\mu)(Z_j-\mu)\right| + 2\left|\frac{1}{n}\sum_{i=1}^n Z_i\mu - \mu^2\right| \\ &\leq \left|\frac{1}{n(n-1)}\sum_{i\neq j}(Z_i-\mu)(Z_j-\mu)\right| + 2R\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mu\right| \\ &\leq (C+2)\sqrt{\frac{R^2\operatorname{Var}(Z)\log(2C/\delta)}{n}} + \frac{2(C+2)R^2\log(2C/\delta)}{3n}\end{aligned}$$

Let us now address the first term, a simple application of Bernstein's inequality gives that with probability at least $1-\delta$

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - \mathbb{E}[Z^2]\right| &\leq \sqrt{\frac{2\operatorname{Var}(Z^2)\log(2/\delta)}{n}} + \frac{2R^2\log(2/\delta)}{3n} \\ &\leq \sqrt{\frac{2R^2\operatorname{Var}(Z)\log(2/\delta)}{n}} + \frac{2R^2\log(2/\delta)}{3n}.\end{aligned}$$

Combining the two inequalities, we obtain the result. □

Since we are estimating the variance of the sample average estimator, we divide by another factor of n. Meanwhile the range and the variance terms themselves are certainly $O(1)$, so the error terms in Lemma 5 are $O(n^{-3/2})$ and $O(n^{-2})$ respectively. Formally, there exists a universal constants $C_1, C_2 > 0$ such that for any $\delta \in (0,1)$ with probability at least $1-\delta$ we have

$$\left|\operatorname{Var}(\theta) - \widehat{\operatorname{Var}}(\theta)\right| \leq C_1\sqrt{\frac{\log(1/\delta)}{n^3}} + C_2\frac{\log(1/\delta)}{n^2}.$$

By adjusting the constant, we can simplify the expression by removing the n^{-2} term. In other words, there exists a different universal constant $C > 0$ such that

$$\left|\operatorname{Var}(\theta) - \widehat{\operatorname{Var}}(\theta)\right| \leq \frac{C\log(1/\delta)}{n^{3/2}}$$

holds with probability at least $1-\delta$.

For the model selection result, first apply Lemma 5 for all $\theta \in \Theta$, taking a union bound. Further take a union bound over the event that $\text{Bias}(\theta) \leq \text{BiasUB}(\theta)$ for all $\theta \in \Theta$, if it is needed. Then, observe that for any $\theta_0 \in \Theta_0$ we have

$$\begin{aligned}\text{MSE}(\hat\theta) &= \text{Bias}(\hat\theta)^2 + \operatorname{Var}(\hat\theta) \leq \text{BiasUB}(\hat\theta)^2 + \widehat{\operatorname{Var}}(\hat\theta) + \frac{C\log(|\Theta|/\delta)}{n^{3/2}} \\ &\leq \text{BiasUB}(\theta_0)^2 + \widehat{\operatorname{Var}}(\theta_0) + \frac{C\log(|\Theta|/\delta)}{n^{3/2}} \leq 0 + \operatorname{Var}(\theta_0) + \frac{2C\log(|\Theta|/\delta)}{n^{3/2}} = \text{MSE}(\theta_0) + \frac{2C\log(|\Theta|/\delta)}{n^{3/2}}.\end{aligned}$$

The first inequality uses Lemma 5 and the fact that Bias $\leq$ BiasUB. The second uses that $\hat\theta$ optimizes this quantity, and the third uses the property that $\text{BiasUB}(\theta_0) = 0$ by assumption. Note that the universal constant here is slightly different from the one in the variance bound, since we have also taken a union bound for the bias term.

C.4. Construction of Upper Bounds on Bias

In this section we give detailed construction of bias upper bounds that we use in the model selection procedure. Recall that this is for the analysis only. Empirically we found that using the estimators alone — not the upper bounds — leads to better performance.

Throughout, we fix a set of hyperparameters θ, which we suppress from the notation.

Direct bias estimation. The most straightforward bias estimator is to simply approximate the expectation with a sample average.

$$\widetilde{\text{Bias}} = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\hat{w}(x_i, a_i) - w(x_i, a_i) \right) \left(r_i - \hat{\eta}(x_i, a_i) \right) \right|.$$

This estimator has finite-sum structure, and naively, each term is bounded in $[-w_\infty, w_\infty]$ where $w_\infty = \max_{x,a} w(x, a)$. The variance is at most $\mathbb{E}_\mu[w(x, a)^2]$. Hence Bernstein's inequality gives that with probability at least $1 - \delta$

$$\left| \widetilde{\text{Bias}} - \text{Bias} \right| \le \sqrt{\frac{2\,\mathbb{E}_\mu[w(x, a)^2] \log(2/\delta)}{n}} + \frac{2 w_\infty \log(2/\delta)}{3n}.$$

Inflating the estimate by the right hand side gives BiasUB, which is a high probability upper bound on Bias.

Pessimistic estimation. The bias bound used in the pessimistic estimator and its natural sample estimator are

$$\mathop{\mathbb{E}}_{\mu}\left[\left|\hat{w}(x, a) - w(x, a)\right|\right], \qquad \widetilde{\text{Bias}} = \frac{1}{n} \sum_{i} \sum_{a} \mu(a \mid x_i) \left|\hat{w}(x_i, a) - w(x_i, a)\right| = \frac{1}{n} \sum_{i} \sum_{a} \pi(a \mid x_i) \left| \frac{\hat{w}(x_i, a)}{w(x_i, a)} - 1 \right|.$$

Note that since we have already eliminated the dependence on the reward, we can analytically evaluate the expectation over actions, which will lead to lower variance in the estimate.

Again we perform a fairly naive analysis. Since $0 \le \hat{w}(x, a) \le w(x, a)$, the random variables, equal to the inner sum over a, take values in $[0, 1]$. Therefore, Hoeffding's inequality gives that with probability $1 - \delta$

$$\text{Bias} \le \widetilde{\text{Bias}} + \sqrt{\frac{\log(1/\delta)}{2n}},$$

and we use the right hand side for our high probability upper bound.

Optimistic estimation. For the optimistic bound, we must estimate two terms, one involving the regressor and one involving the importance weights. We use

$$T_1 \coloneqq \frac{1}{n} \sum_{i=1}^{n} z(x_i, a_i)(r_i - \hat{\eta}(x_i, a_i))^2, \qquad T_2 \coloneqq \frac{1}{n} \sum_{i=1}^{n} \sum_{a} \mu(a \mid x_i) \frac{\left|\hat{w}(x_i, a) - w(x_i, a)\right|^2}{z(x_i, a)}.$$

Note here that the former uses sampled actions from μ, but does not involve the importance weight, while the latter involves the importance weight but analytically evaluates the expectation over μ. Thus we can expect that both are fairly low variance.

For both, we use Bernstein's inequality. For T_1, each term is bounded in $[-z_\infty, z_\infty]$ where $z_\infty = \max_{x,a} z(x, a)$ and its variance is bounded by $\mathbb{E}_\mu[z(x, a)^2]$. Thus we get that with probability at least $1 - \delta/2$

$$\mathbb{E}\left[\, z(x, a)(r - \hat{\eta}(x, a))^2 \,\right] \le T_1 + \sqrt{\frac{2\,\mathbb{E}_\mu[z(x, a)^2] \log(2/\delta)}{n}} + \frac{2 z_\infty \log(2/\delta)}{3n}.$$

For T_2, we similarly to the pessimistic case convert the inner expectation w.r.t. $\mu(a \mid x_i)$ to an expectation w.r.t. $\pi(a \mid x_i)$, obtaining a random variable bounded between 0 and $\max_{x,a} w(x, a)/z(x, a)$. Using Hoeffding's inequality, we obtain that with probability $1 - 2\delta$

$$\mathop{\mathbb{E}}_{\mu}\left[\left|\hat{w}(x, a) - w(x, a)\right|^2 / z(x, a)\right] \le T_2 + \sqrt{\frac{\max_{x,a} \frac{w(x,a)}{z(x,a)} \log(2/\delta)}{2n}}.$$

The high probability upper bound follows by multiplying the two right hand sides together and taking square root.

D. Experimental Details and Additional Results

D.1. Experimental Details and Results for Atomic Actions

Dataset statistics. We use datasets from the UCI Machine Learning Repository (Dua & Graff, 2017). Dataset statistics are displayed in Table 4.

Hyperparameter grid. For our shrinkage estimators and SWITCH, we choose the shrinkage coefficients from a grid of 30 geometrically spaced values. For the pessimistic estimator and SWITCH, the largest and smallest values in the grid are the 0.05 quantile and 0.95 quantile of the importance weights. For the optimistic estimator, the largest and smallest values are $0.01 \times (w_{0.05})^2$ and $100 \times (w_{0.95})^2$ where $w_{0.05}$ and $w_{0.95}$ are the 0.05 and 0.95 quantile of the importance weights.

For the off-policy learning experiments, we only consider the shrinkage coefficients in $\{0.0, 0.1, 1, 10, 100, 1000, \infty\}$ during training, while for model selection, we use the same grid as in the evaluation experiments.

MRDR. Farajtabar et al. (2018) propose training the regression model with a specific choice of weighting z, which we also use in our experiments. When the evaluation policy π is deterministic, they set $z(x, a) = \mathbf{1}\{\pi(x) = a\} \cdot \frac{1-\mu(a|x)}{\mu(a|x)^2}$. For stochastic policies, following the implementation of Farajtabar et al., we sample $a_i \sim \pi(\cdot \mid x_i)$ for each example in the dataset used to train the reward predictor. Then we proceed as if the evaluation policy deterministically chooses a_i on example x_i.

Ablation study for deterministic target policy. Since MRDR is more suited to deterministic policies, we also report the results of our regressor and shrinkage ablations for a deterministic target policy $\pi_{1,\text{det}}$ in Table 5. As with the stochastic policies, the estimator influences the choice of reward predictor, but note that $z = 1$ and $z = w$ are more favorable here. This is likely due to high variance suffered from training with $z = w^2$, because the importance weights are larger with a deterministic policy. Our shrinkage ablation reveals that both estimator types are important also when the target policy is deterministic.

Ablation study for model selection. In Table 6, we show the comparison of different model selection methods under different reward predictors and different shrinkage types. In most cases, Dir-all (DRs-direct where the bias bound is estimated as the pointwise minimum of (1) the bias, (2) the optimistic bound and (3) the pessimistic bound) and Up-all (all bias estimates are adjusted by adding twice standard error before taking pointwise minimum) are most frequently statistically indistinguishable from the best, which suggests that our proposed bias estimate (by taking pointwise minimum of the three) is robust and adaptive.

Comparisons across additional experimental conditions. In Figure 4 and Figure 5, we compare our new estimators, DRs-direct and DRs-upper, with baselines across various conditions (apart from deterministic versus stochastic rewards from the main paper). We first investigate the performance under *friendly logging* (logging and evaluation policies are derived from the same deterministic policy, $\pi_{1,det}$), *adversarial logging* (logging and evaluation policies are derived from different policies $\pi_{1,det}$, $\pi_{2,det}$), and *uniform logging* (logging policy is uniform over all actions). Then we plot the performance in the small sample regime, where we aggregate the 108 conditions (6 logging policies, 9 datasets, deterministic/stochastic reward) at just 200 bandit samples.

Comparisons across all reward predictors. In Table 7–Table 11, we compare the performance of DRs-direct and DRs-upper against baselines across various choices of reward predictors. We begin with using the best reward predictor type for each method (matching the setting of the main paper), and then consider each reward predictor in turn, across all estimators. We report the number of conditions where each estimator is statistically indistinguishable from the best, and the number of conditions where each estimator statistically dominates all others. DRs-upper is most often in the top group and most often the unique winner. DRs-direct is also better than snIPS, snDR, and SWITCH. These results suggest that our shrinkage estimators are robust to different choices of reward predictors, and not just limited to the recommended set $\{\hat{\eta} \equiv 0, z = w^2\}$.

Robustness of DRs-direct **and** DRs-upper **(w.r.t. inclusion of more reward predictors)** In Figure 6, we test the robustness of our proposed methods as we incorporate more reward predictors. Our practical suggestions is to use $\{\hat{\eta} \equiv 0, z = w^2\}$ (shown as DRs-direct and DRs-upper in the figure). Here we also evaluate these methods when selecting from all reward predictors in the set $\{\hat{\eta} \equiv 0, z \equiv 1, z = w, z = w^2, \text{MRDR}\}$ (shown as DRs-direct (all) and DRs-upper (all) in the figure). For DRs-direct, the curves almost match, suggesting that it is quite robust. However, DRs-upper is less robust to including additional reward predictors.

Learning curves. At the end of appendix, we provide learning curves across all conditions. Dataset *glass* is excluded since we only ran it for a single sample size $n = 214$.

Dataset	Glass	Ecoli	Vehicle	Yeast	PageBlok	OptDigits	SatImage	PenDigits	Letter
Actions	6	8	4	10	5	10	6	10	26
Examples	214	336	846	1484	5473	5620	6435	10992	20000

Table 4. Dataset statistics.

	$\hat{\eta} \equiv 0$	$z \equiv 1$	$z = w$	$z = w^2$	MRDR
DM	0 (0)	54 (30)	59 (23)	35 (4)	24 (6)
DR	28 (1)	94 (11)	85 (0)	85 (1)	85 (0)
snDR	65 (7)	86 (7)	79 (0)	72 (0)	71 (0)
DRs	14 (9)	51 (17)	65 (14)	54 (6)	47 (4)

	DRps	DRos
$\hat{\eta} \equiv 0$	13	59
$z \equiv 1$	29	55
$z = w$	26	66
$z = w^2$	30	67
MRDR	29	63

Table 5. Ablation analysis for *deterministic* target policy $\pi_{1,\text{det}}$ across experimental conditions. Left: we compare reward predictors using a fixed estimator (with oracle tuning if applicable). We report the number of conditions where a regressor is statistically indistinguishable from the best and, in parenthesis, the number of conditions where it statistically dominates all others. Right: we compare different shrinkage types using a fixed reward predictor (with oracle tuning) reporting the number of conditions where one statistically dominates the other.

D.2. Experimental Details for Combinatorial Actions

Hyperparameter grid. We select the hyperparameter λ from the grid of 15 geometrically spaced values, with the smallest value $0.01 \times (w_{0.05})^2$ and the largest value $100 \times (w_{0.95})^2$, where $w_{0.05}$ and $w_{0.95}$ are the 0.05 and 0.95 quantiles of the weights $w(x_i, \mathbf{a}_i)$ on the logged data. We also add two boundary values $\lambda = 10^{-50}$ and $\lambda = 10^{30}$ to include DM and DR-PI as special cases.

Basis construction. We use the logging distribution supported on a linearly independent set of actions, i.e., a basis, constructed following Algorithm 1. The number of elements of the basis for the action space of lists of length ℓ out of m items is $1 + \ell(m - 1)$.

Algorithm 1 Constructing basis for the action space of lists of length ℓ out of m items.

```
*Actions* $\mathbf{a}$ *are represented as tuples of size* $\ell$ *with entries* $a[i] \in \{0, \dots, m-1\}$*, indexed by* $i \in \{0, \dots, \ell-1\}$.
*Actions* $\mathbf{a}$ *correspond to vectors in* $\mathbb{R}^{\ell m}$*, obtained by representing each* $a[i]$ *as a vector of standard basis in* $\mathbb{R}^m$*,*
*and concatenating these vectors.*
**Assume:** Greedy action is the tuple $\mathbf{g} = [1, 2, \dots, \ell]$
Initialize $\mathcal{B} = \{\mathbf{g}\}$
**for** $i = 1$ **to** $\ell - 1$ **do**
    Set $\mathbf{a} = \mathbf{g}$
    Set $a[0] = i$ and $a[i] = 0$, $\mathcal{B} = \mathcal{B} \cup \{\mathbf{a}\}$
**end for**
**for** $j = \ell$ **to** $m - 1$ **do**
    Set $\mathbf{a} = \mathbf{g}$
    Set $a[0] = j$ and $\mathcal{B} = \mathcal{B} \cup \{\mathbf{a}\}$
**end for**
**for** $i = 1$ **to** $\ell - 1$ **do**
    **for** $i' = 1$ **to** $\ell - 1$ such that $i \neq i'$ **do**
        Set $\mathbf{a} = \mathbf{g}$
        Set $a[0] = i$, $a[i] = i'$ and $a[i'] = 0$, $\mathcal{B} = \mathcal{B} \cup \{\mathbf{a}\}$
    **end for**
    **for** $j = \ell$ **to** $m - 1$ **do**
        Set $\mathbf{a} = \mathbf{g}$
        Set $a[0] = i$, $a[i] = j$ and $\mathcal{B} = \mathcal{B} \cup \{\mathbf{a}\}$
    **end for**
**end for**
**for** $i = 1$ **to** $\ell - 1$ **do**
    Set $\mathbf{a} = \mathbf{g}$
    Set $a[0] = \ell$, $a[i] = 0$ and $\mathcal{B} = \mathcal{B} \cup \{\mathbf{a}\}$
**end for**
Return $\mathcal{B}$
```

	Dir-all	Dir-naive	Dir-opt	Dir-pes	Up-all	Up-naive	Up-opt	Up-pes
0-pes	71	67	74	78	63	60	79	79
0-opt	75	68	71	81	64	62	66	81
0-best	63	59	67	74	56	54	64	76
w^2-pes	47	41	5	3	72	71	5	3
w^2-opt	47	40	3	2	73	70	3	2
w^2-best	47	42	4	3	75	72	4	3
best-pes	51	46	7	5	69	70	6	5
best-opt	49	46	7	3	69	70	5	3
best-best	50	46	8	4	71	70	6	4
all-best	49	46	7	6	65	67	6	6

Table 6. Comparison of model selection methods when paired with different reward predictor sets and shrinkage types. As in other tables, we record the number of conditions in which this model selection method is statistically indistinguishable from the best, for fixed reward predictor set and shrinkage types. Columns are indexed by model selection methods, "Dir" denotes taking sample average and "Up" denotes inflating sample averages with twice the standard error. "Naive" denotes directly estimating bias, "opt" denotes estimating optimistic bias bound, "pes" denotes pessimistic bias bound, and "all" denotes taking the pointwise minimum of all three. Rows are indexed by reward predictors: $\hat{\eta} \equiv 0$, $z = w^2$, "best" denotes selecting over both, and "all" denotes selecting over these and additionally $z = 1$, $z = w$, and MRDR. Rows are also indexed by shrinkage type, optimistic, pessimistic, and best, which denotes model selection over both.

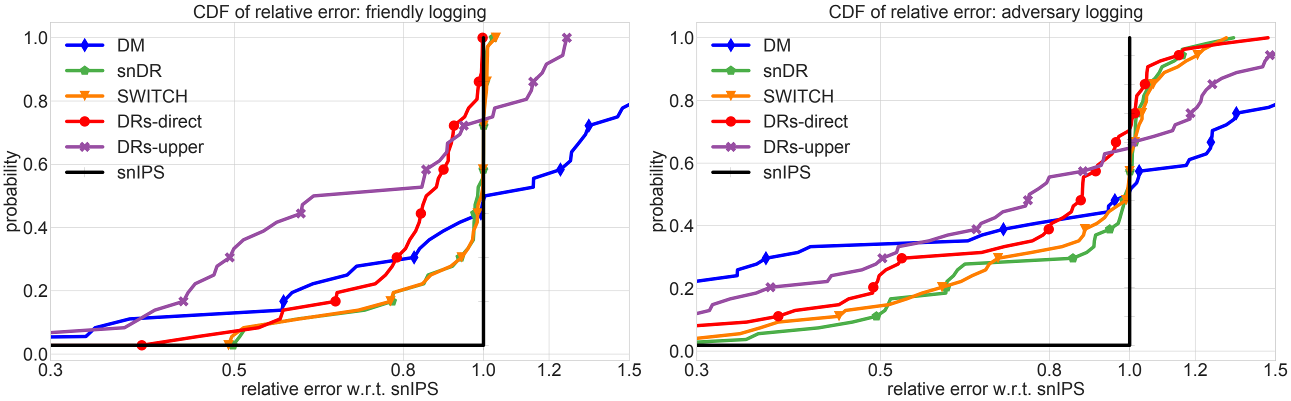


Figure 4. CDF plots of normalized MSE aggregated across all conditions with friendly scenario (left) and adversary scenario (right).

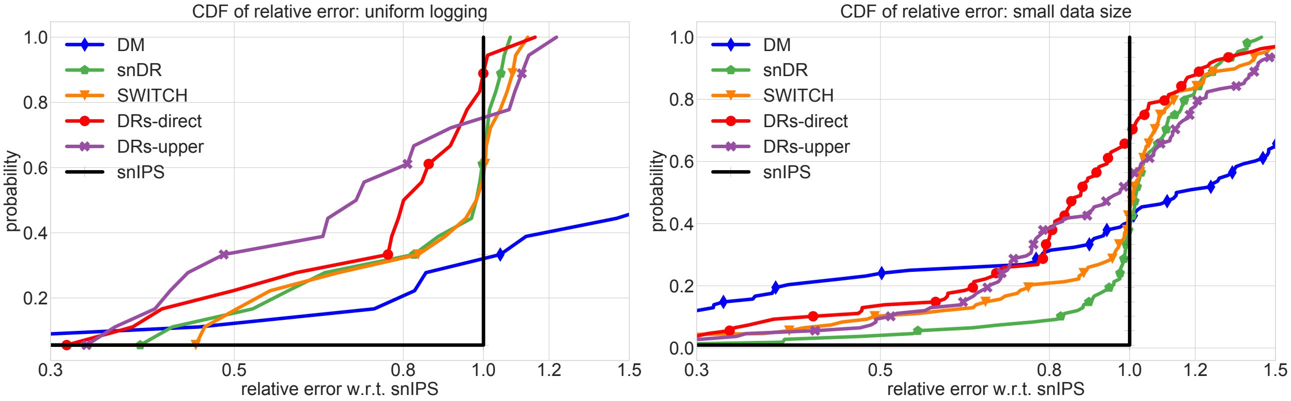


Figure 5. CDF plots of normalized MSE aggregated across all conditions with uniform logging policy scenario (left) and small data regime (right).

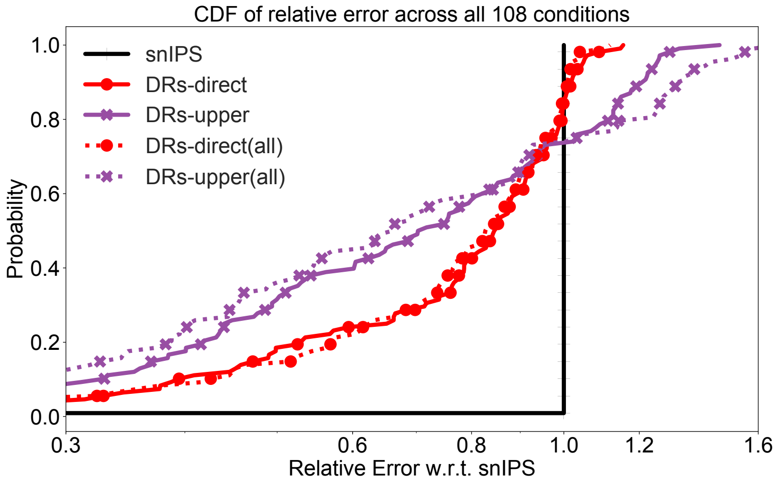


Figure 6. Robustness test for DRs-direct and DRs-upper. DRs-direct (all) and DRs-upper (all) means the corresponding method with reward predictor select from all possible cases $\{\hat{\eta} = 0, z = 1, z = w, z = w^2\}$ and MRDR.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	1	39	1	1	27	56
Unique Best	0	28	0	0	14	52

Table 7. Significance testing for different estimators across all conditions, with each estimator using its best reward predictor (DM uses $z = 1$, snDR uses $z = w$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = w^2\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	10	39	19	15	25	49
Unique Best	0	23	3	1	5	43

Table 8. Significance testing for different estimators across all conditions (DM, snDR use reward $z = 1$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = 1\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	3	44	5	5	30	58
Unique Best	0	23	0	0	6	51

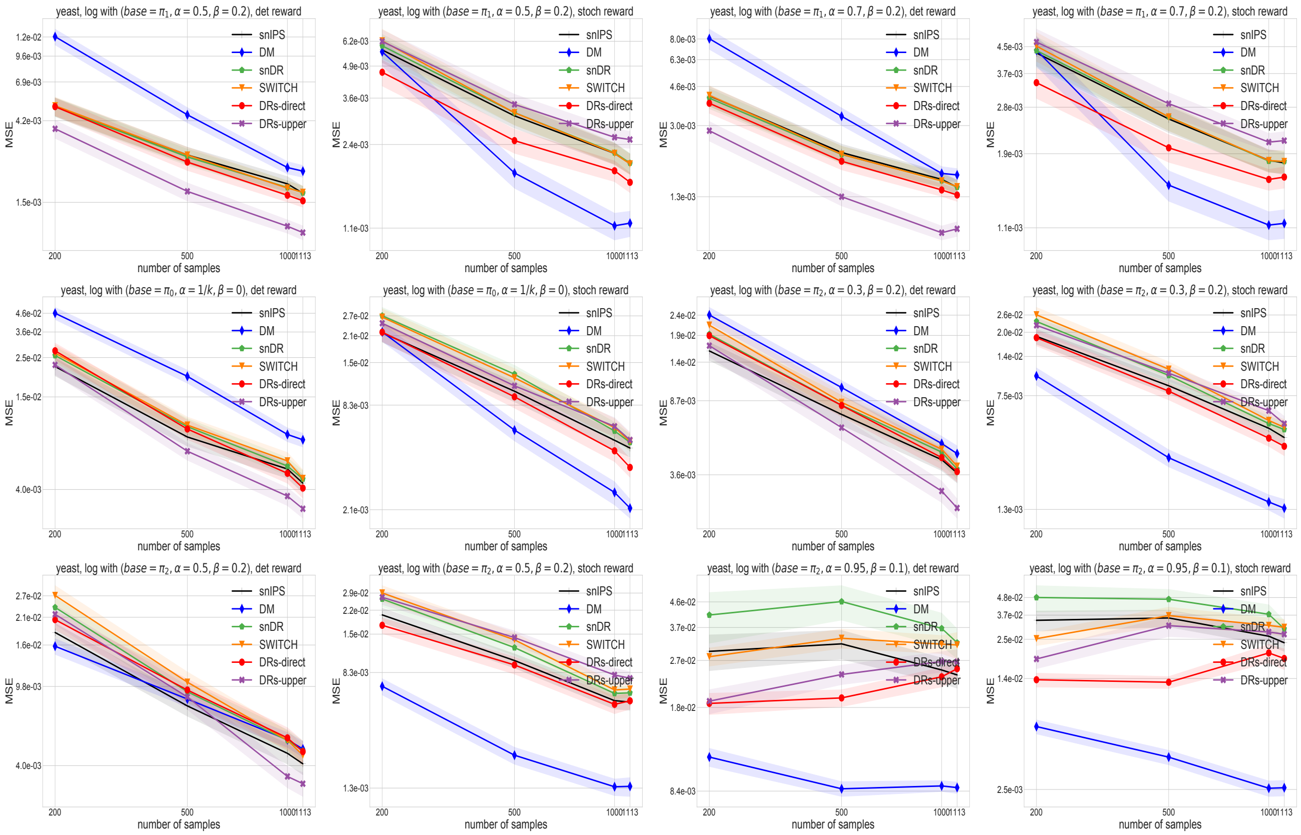
Table 9. Significance testing for different estimators across all conditions (DM, snDR use reward $z = w$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = w\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

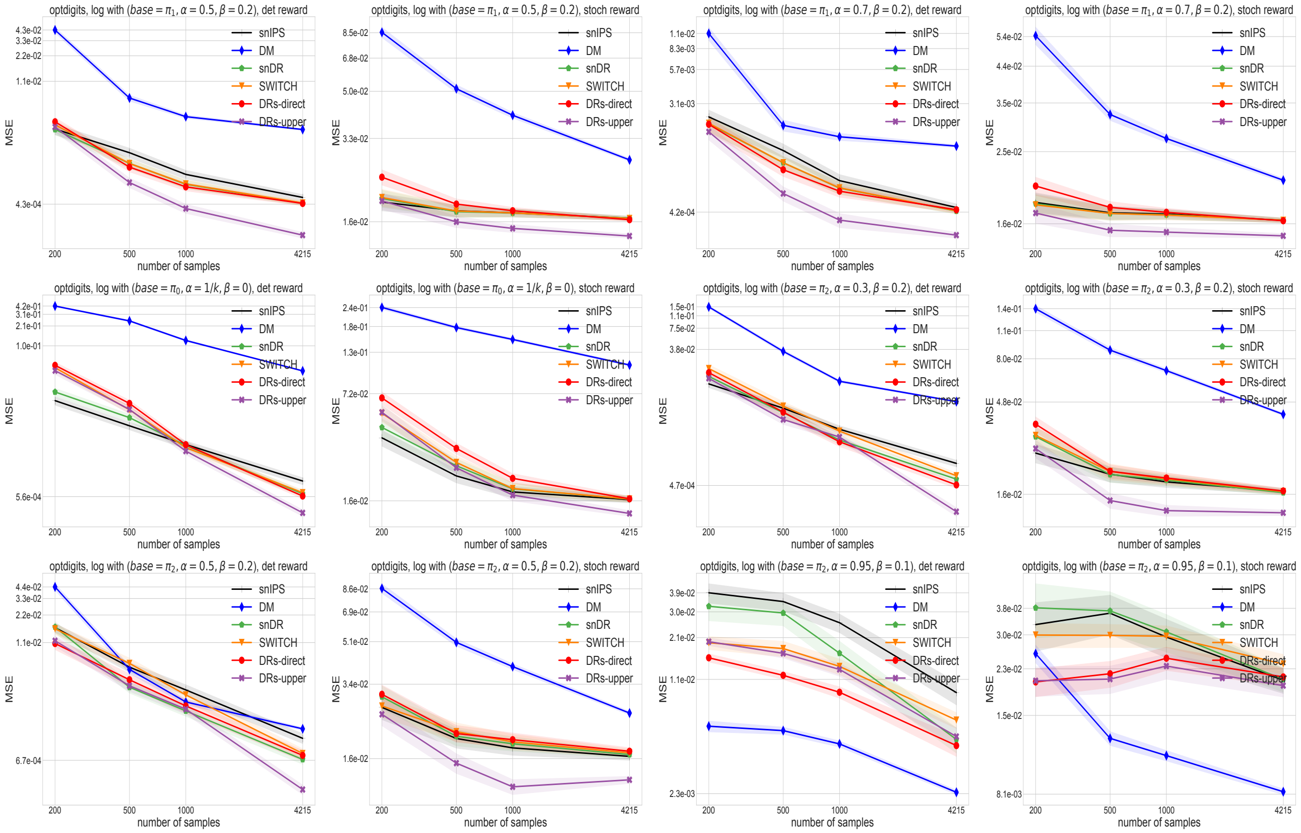
	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	5	36	5	4	43	63
Unique Best	0	15	0	0	13	46

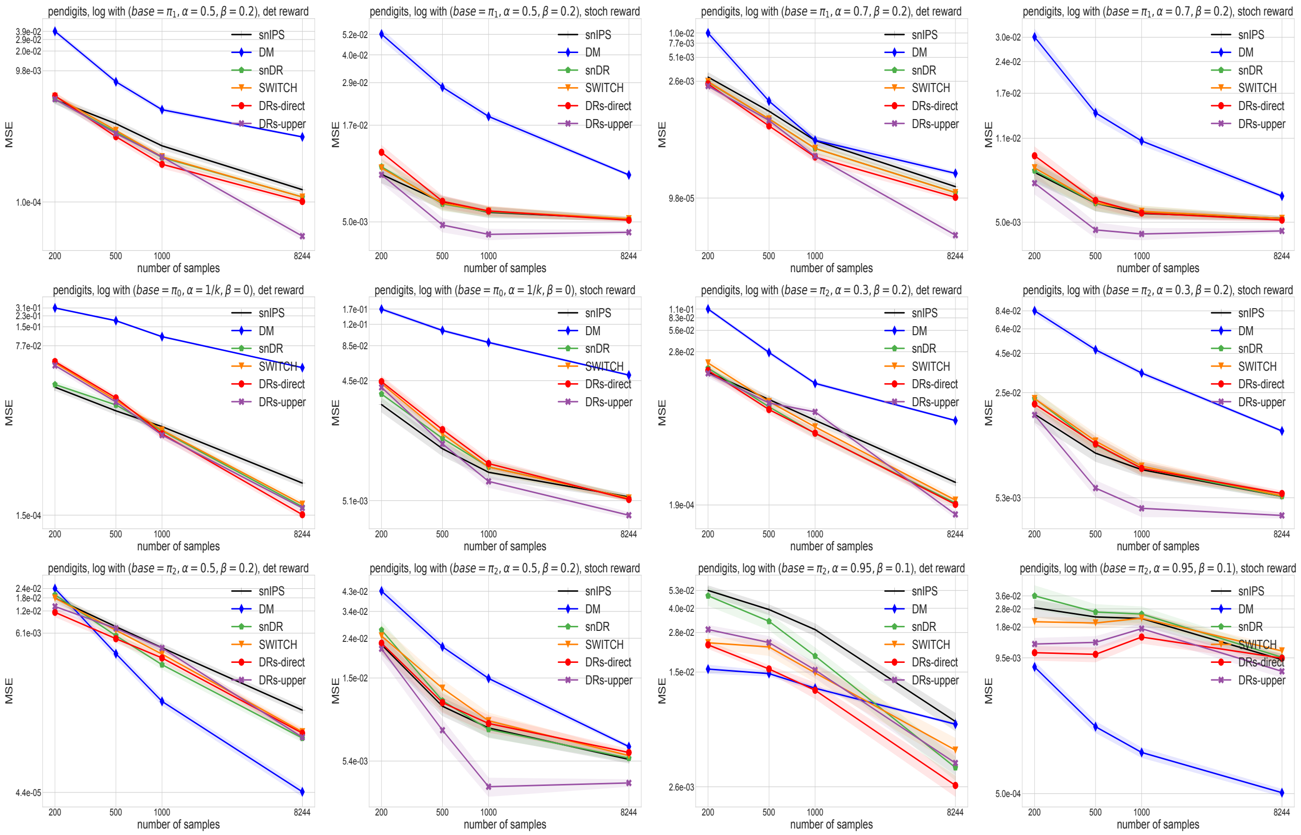
Table 10. Significance testing for different estimators across all conditions (DM, snDR use reward $z = w^2$, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, z = w^2\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

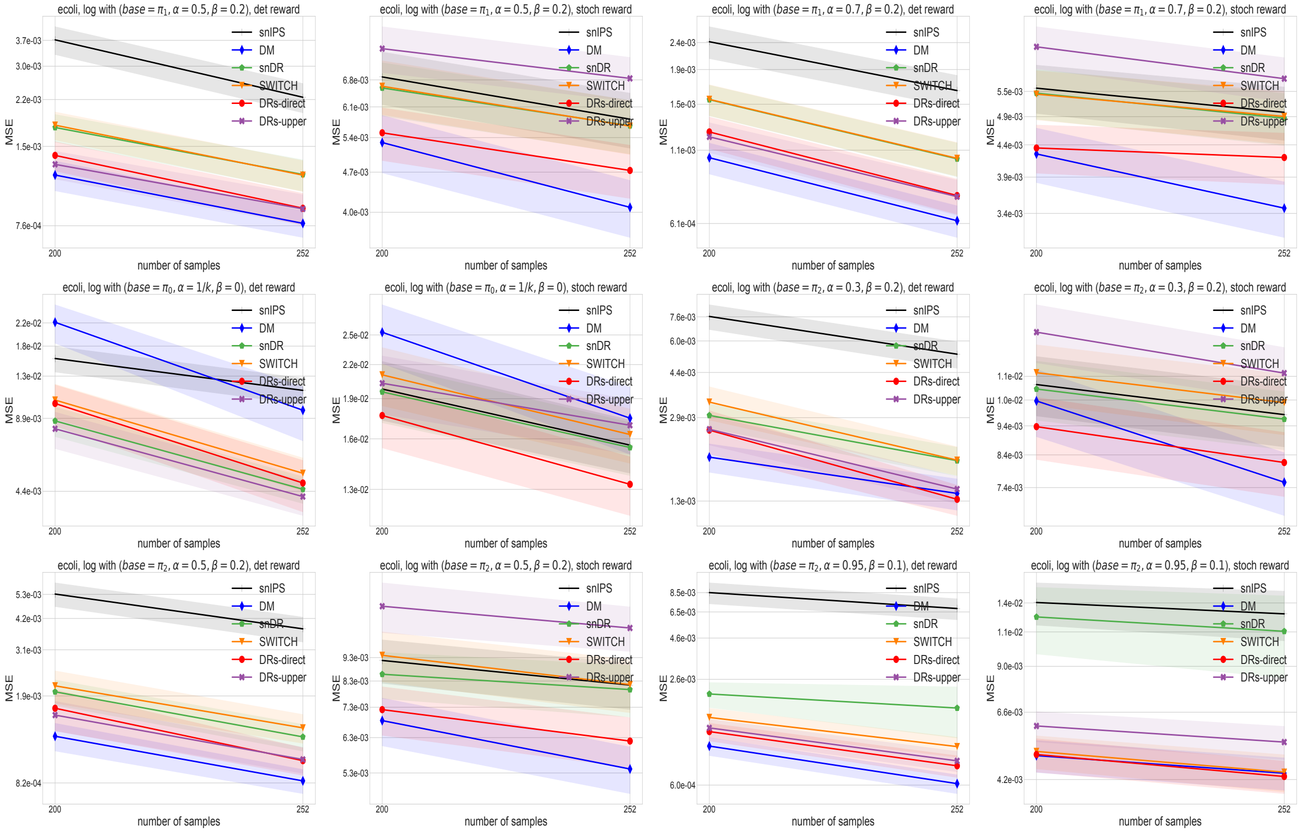
	snIPS	DM	snDR	SWITCH	DRs-direct	DRs-upper
Best or Tied	10	17	8	10	37	73
Unique Best	0	7	0	0	17	57

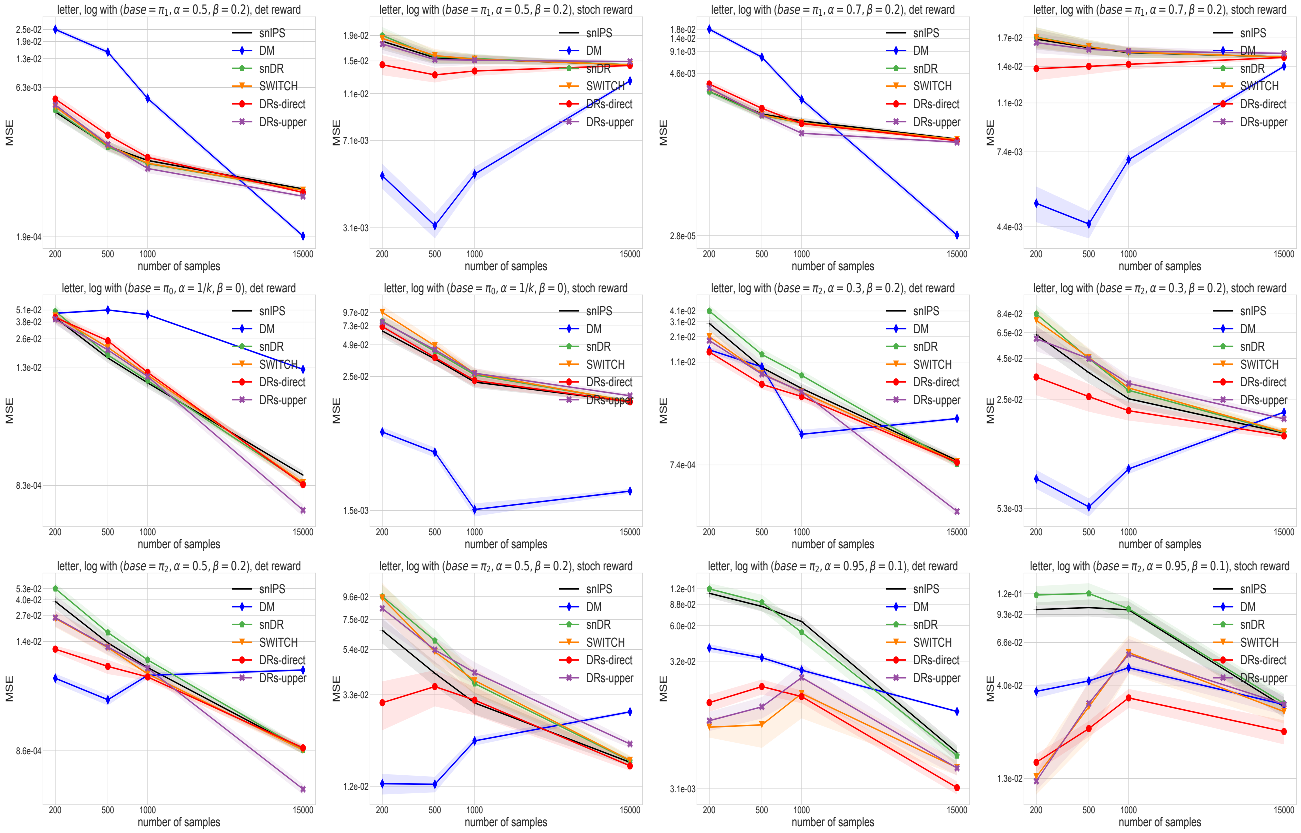
Table 11. Significance testing for different estimators across all conditions (DM, snDR use reward estimated from MRDR, while SWITCH and DRs use reward from $\{\hat{\eta} = 0, \text{MRDR}\}$). In the top row, we report the number of conditions where each estimator is statistically indistinguishable from the best, and in the bottom row we report the number of conditions where each estimator is the uniquely best.

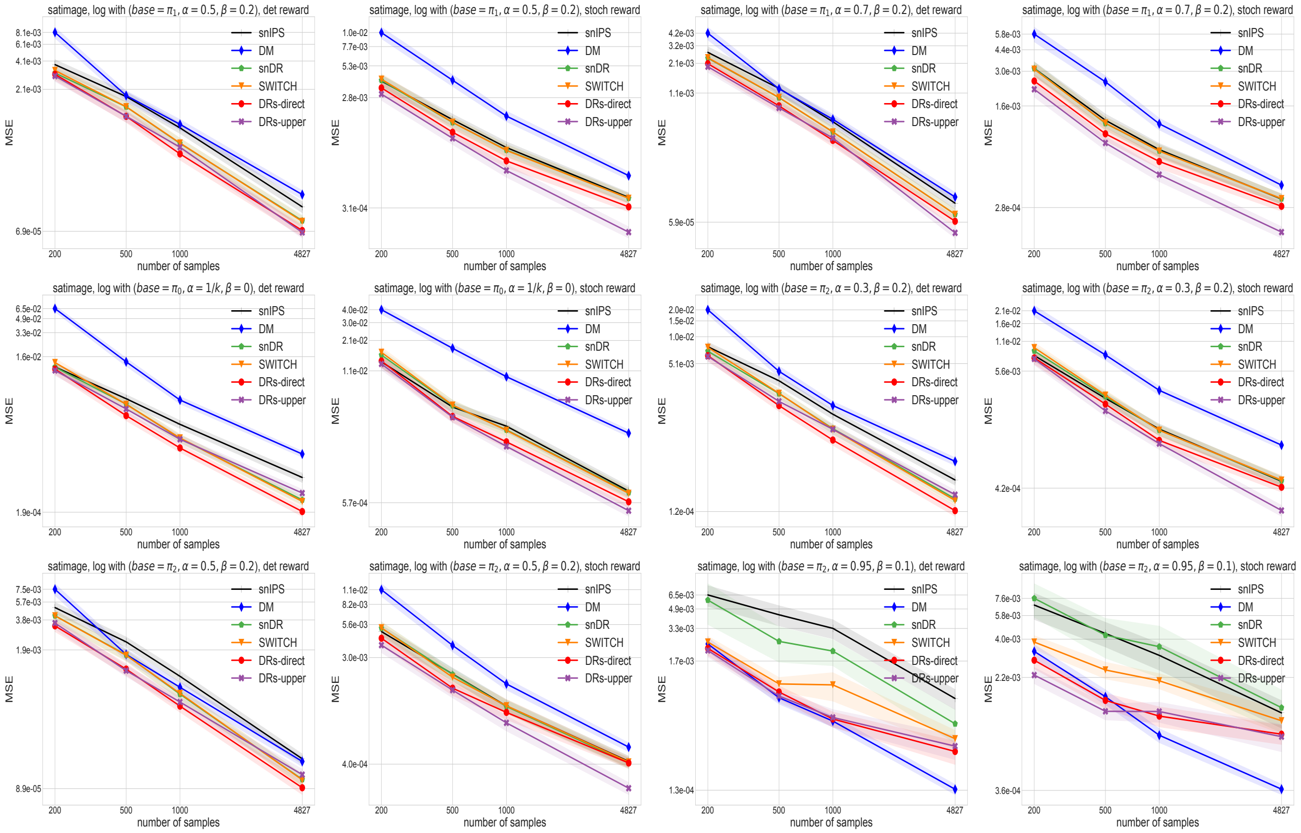

yeast, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
1.2e-02
9.6e-03
6.9e-03
4.2e-03
1.5e-03
yeast, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
6.2e-03
4.9e-03
3.6e-03
2.4e-03
1.1e-03
yeast, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
8.0e-03
6.3e-03
4.6e-03
3.0e-03
1.3e-03
yeast, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
4.5e-03
3.7e-03
2.8e-03
1.9e-03
1.1e-03
yeast, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
4.6e-02
3.6e-02
2.5e-02
1.5e-02
4.0e-03
yeast, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
2.7e-02
2.1e-02
1.5e-02
8.3e-03
2.1e-03
yeast, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
2.4e-02
1.9e-02
1.4e-02
8.7e-03
3.6e-03
yeast, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
2.6e-02
2.0e-02
1.4e-02
7.5e-03
1.3e-03
yeast, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
2.7e-02
2.1e-02
1.6e-02
9.8e-03
4.0e-03
yeast, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
2.9e-02
2.2e-02
1.5e-02
8.3e-03
1.3e-03
yeast, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
4.6e-02
3.7e-02
2.7e-02
1.8e-02
8.4e-03
yeast, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
4.8e-02
3.7e-02
2.5e-02
1.4e-02
2.5e-03
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
200
500
10001113
number of samples

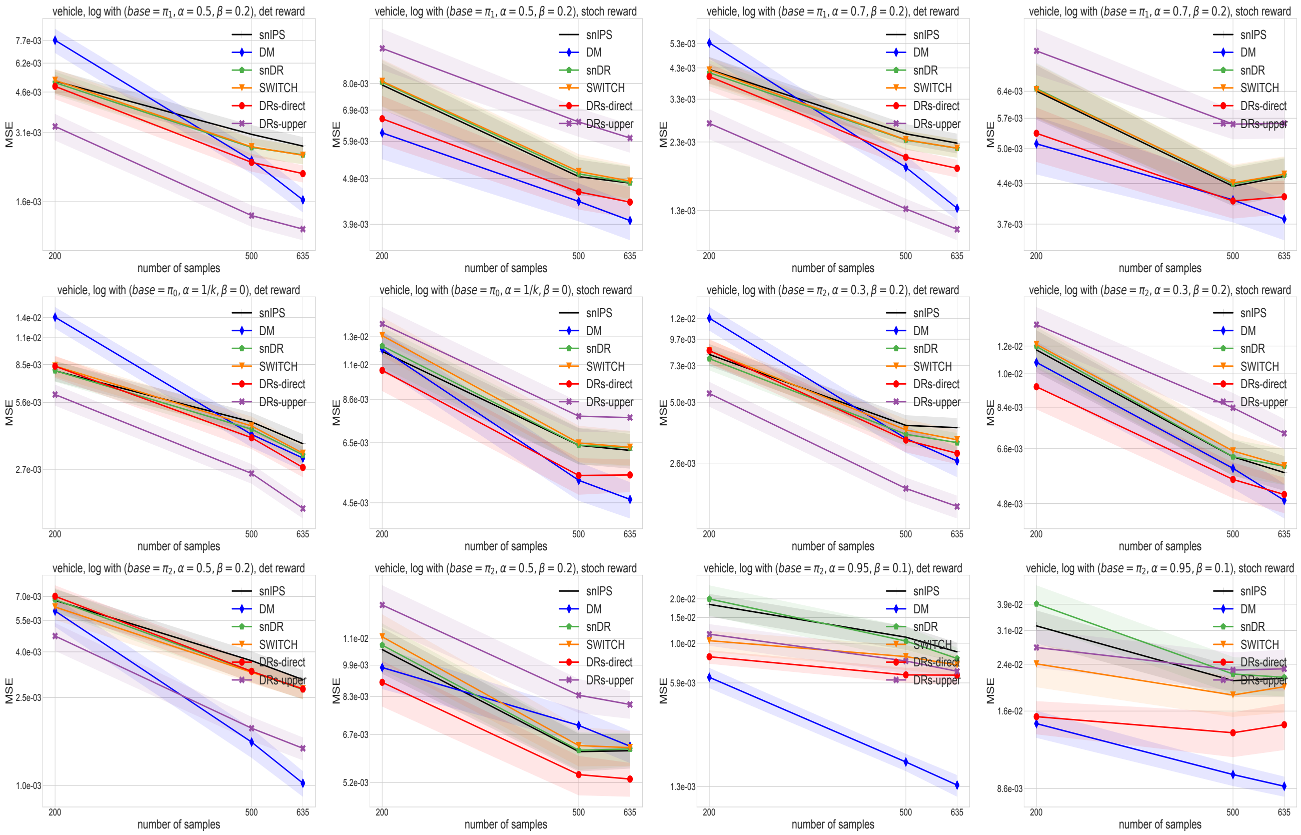

optdigits, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
4.3e-02
3.3e-02
2.2e-02
1.1e-02
4.3e-04
optdigits, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
8.5e-02
6.8e-02
5.0e-02
3.3e-02
1.6e-02
optdigits, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
1.1e-02
8.3e-03
5.7e-03
3.1e-03
4.2e-04
optdigits, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
5.4e-02
4.4e-02
3.5e-02
2.5e-02
1.6e-02
optdigits, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
4.2e-01
3.1e-01
2.1e-01
1.0e-01
5.6e-04
optdigits, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
2.4e-01
1.8e-01
1.3e-01
7.2e-02
1.6e-02
optdigits, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
1.5e-01
1.1e-01
7.5e-02
3.8e-02
4.7e-04
optdigits, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
1.4e-01
1.1e-01
8.0e-02
4.8e-02
1.6e-02
optdigits, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
4.4e-02
3.3e-02
2.2e-02
1.1e-02
6.7e-04
optdigits, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
8.6e-02
6.9e-02
5.1e-02
3.4e-02
1.6e-02
optdigits, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
3.9e-02
3.0e-02
2.1e-02
1.1e-02
2.3e-03
optdigits, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
3.8e-02
3.0e-02
2.3e-02
1.5e-02
8.1e-03
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
200
500
1000
4215
number of samples

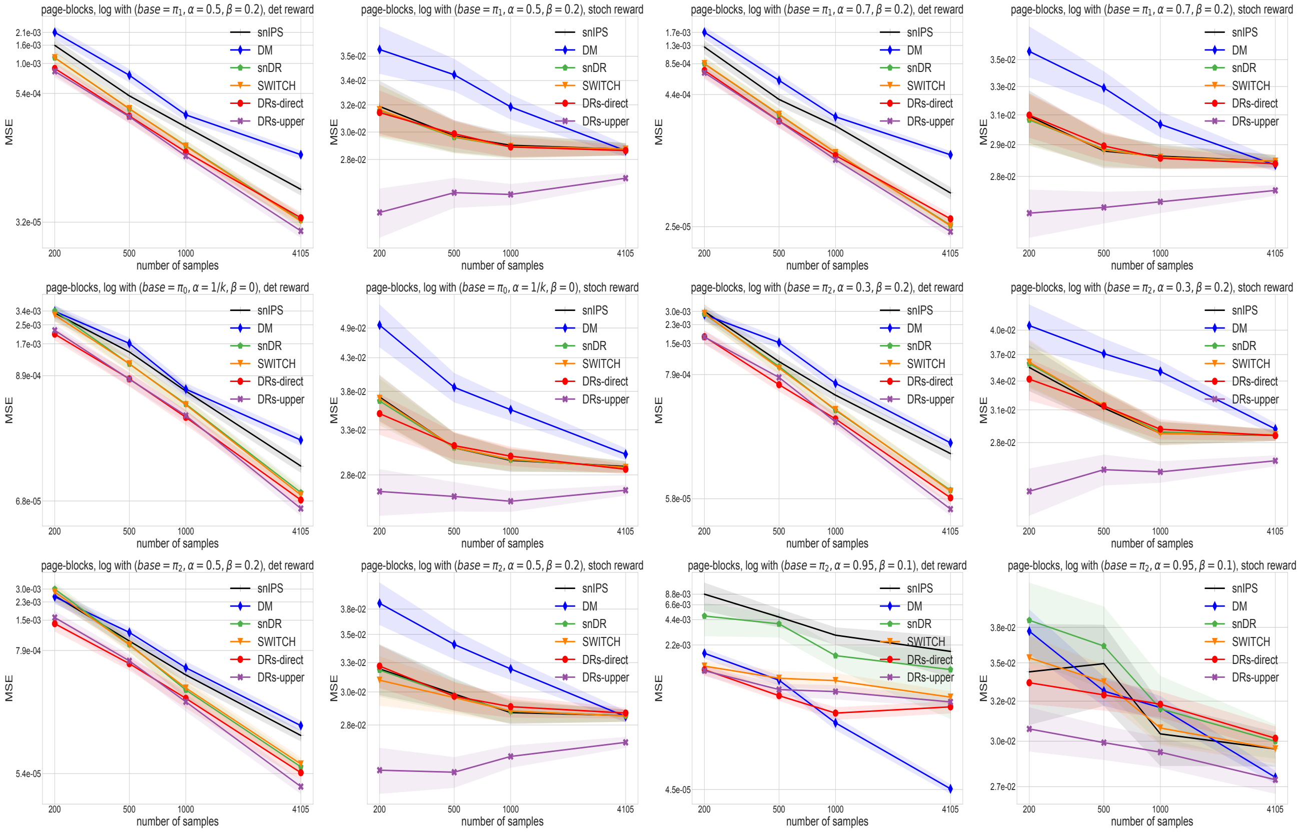

pendigits, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
3.9e-02
2.9e-02
2.0e-02
9.8e-03
1.0e-04
200
500
1000
8244
number of samples
pendigits, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
5.2e-02
4.0e-02
2.9e-02
1.7e-02
5.0e-03
pendigits, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
1.0e-02
7.7e-03
5.1e-03
2.6e-03
9.8e-05
pendigits, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
3.0e-02
2.4e-02
1.7e-02
1.1e-02
5.0e-03
pendigits, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
3.1e-01
2.3e-01
1.5e-01
7.7e-02
1.5e-04
pendigits, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
1.7e-01
1.2e-01
8.5e-02
4.5e-02
5.1e-03
pendigits, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
1.1e-01
8.3e-02
5.6e-02
2.8e-02
1.9e-04
pendigits, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
8.4e-02
6.4e-02
4.5e-02
2.5e-02
5.3e-03
pendigits, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
2.4e-02
1.8e-02
1.2e-02
6.1e-03
4.4e-05
pendigits, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
4.3e-02
3.4e-02
2.4e-02
1.5e-02
5.4e-03
pendigits, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
5.3e-02
4.0e-02
2.8e-02
1.5e-02
2.6e-03
pendigits, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
3.6e-02
2.8e-02
1.8e-02
9.5e-03
5.0e-04


ecoli, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
3.7e-03
3.0e-03
2.2e-03
1.5e-03
7.6e-04
200
252
number of samples
ecoli, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
6.8e-03
6.1e-03
5.4e-03
4.7e-03
4.0e-03
ecoli, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
2.4e-03
1.9e-03
1.5e-03
1.1e-03
6.1e-04
ecoli, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
5.5e-03
4.9e-03
4.4e-03
3.9e-03
3.4e-03
ecoli, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
2.2e-02
1.8e-02
1.3e-02
8.9e-03
4.4e-03
ecoli, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
2.5e-02
2.2e-02
1.9e-02
1.6e-02
1.3e-02
ecoli, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
7.6e-03
6.0e-03
4.4e-03
2.9e-03
1.3e-03
ecoli, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
1.1e-02
1.0e-02
9.4e-03
8.4e-03
7.4e-03
ecoli, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
5.3e-03
4.2e-03
3.1e-03
1.9e-03
8.2e-04
ecoli, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
9.3e-03
8.3e-03
7.3e-03
6.3e-03
5.3e-03
ecoli, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
8.5e-03
6.5e-03
4.6e-03
2.6e-03
6.0e-04
ecoli, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
1.4e-02
1.1e-02
9.0e-03
6.6e-03
4.2e-03


letter, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
2.5e-02
1.9e-02
1.3e-02
6.3e-03
1.9e-04
letter, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
1.9e-02
1.5e-02
1.1e-02
7.1e-03
3.1e-03
letter, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
1.8e-02
1.4e-02
9.1e-03
4.6e-03
2.8e-05
letter, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
1.7e-02
1.4e-02
1.1e-02
7.4e-03
4.4e-03
letter, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
5.1e-02
3.8e-02
2.6e-02
1.3e-02
8.3e-04
letter, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
9.7e-02
7.3e-02
4.9e-02
2.5e-02
1.5e-03
letter, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
4.1e-02
3.1e-02
2.1e-02
1.1e-02
7.4e-04
letter, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
8.4e-02
6.5e-02
4.5e-02
2.5e-02
5.3e-03
letter, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
5.3e-02
4.0e-02
2.7e-02
1.4e-02
8.6e-04
letter, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
9.6e-02
7.5e-02
5.4e-02
3.3e-02
1.2e-02
letter, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
1.2e-01
8.8e-02
6.0e-02
3.2e-02
3.1e-03
letter, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
1.2e-01
9.3e-02
6.6e-02
4.0e-02
1.3e-02
MSE
200
500
1000
15000
number of samples
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper


satimage, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
8.1e-03
6.1e-03
4.1e-03
2.1e-03
6.9e-05
satimage, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
1.0e-02
7.7e-03
5.3e-03
2.8e-03
3.1e-04
satimage, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
4.2e-03
3.2e-03
2.1e-03
1.1e-03
5.9e-05
satimage, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
5.8e-03
4.4e-03
3.0e-03
1.6e-03
2.8e-04
satimage, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
6.5e-02
4.9e-02
3.3e-02
1.6e-02
1.9e-04
satimage, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
4.0e-02
3.0e-02
2.1e-02
1.1e-02
5.7e-04
satimage, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
2.0e-02
1.5e-02
1.0e-02
5.1e-03
1.2e-04
satimage, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
2.1e-02
1.6e-02
1.1e-02
5.6e-03
4.2e-04
satimage, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
7.5e-03
5.7e-03
3.8e-03
1.9e-03
8.9e-05
satimage, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
1.1e-02
8.2e-03
5.6e-03
3.0e-03
4.0e-04
satimage, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
6.5e-03
4.9e-03
3.3e-03
1.7e-03
1.3e-04
satimage, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
7.6e-03
5.8e-03
4.0e-03
2.2e-03
3.6e-04
MSE
200
500
1000
4827
number of samples
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper


vehicle, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
vehicle, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
vehicle, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
vehicle, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
vehicle, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
vehicle, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
vehicle, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
vehicle, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
vehicle, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
vehicle, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
vehicle, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
vehicle, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
number of samples
200
500
635


page-blocks, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), det reward
page-blocks, log with ($base = \pi_1, \alpha = 0.5, \beta = 0.2$), stoch reward
page-blocks, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), det reward
page-blocks, log with ($base = \pi_1, \alpha = 0.7, \beta = 0.2$), stoch reward
page-blocks, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), det reward
page-blocks, log with ($base = \pi_0, \alpha = 1/k, \beta = 0$), stoch reward
page-blocks, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), det reward
page-blocks, log with ($base = \pi_2, \alpha = 0.3, \beta = 0.2$), stoch reward
page-blocks, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), det reward
page-blocks, log with ($base = \pi_2, \alpha = 0.5, \beta = 0.2$), stoch reward
page-blocks, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), det reward
page-blocks, log with ($base = \pi_2, \alpha = 0.95, \beta = 0.1$), stoch reward
snIPS
DM
snDR
SWITCH
DRs-direct
DRs-upper
MSE
number of samples
200
500
1000
4105