Confident Off-Policy Evaluation and Selection through Self-Normalized Importance Weighting: Supplementary Material

Ilja Kuzborskij **Claire Vernade** **András György** **Csaba Szepesvári**

DeepMind

A More on the stopping criteria for $\tilde{V}_t^{\text{SN}}$, $\tilde{U}_t^{\text{SN}}$, and $\tilde{B}_t$ in Algorithm 1

As discussed in Section 3, we control the simulation error introduced by the output of Algorithm 1 by applying a stopping criterion based on the empirical Bernstein's inequality (Theorem 4). In particular, for a user specified precision $\varepsilon > 0$, the estimation of V^{SN} is stopped when

$$\varepsilon \geq \sqrt{\frac{2\widehat{\text{Var}}(\tilde{V}_t^{\text{SN}})}{t}} + \frac{7}{3} \cdot \frac{2x}{t-1}$$

is satisfied. Suppose that the simulation has stopped after T_ε iterations. Then, the above guarantees w.p. at least $1 - e^{-x}, x > 0$ that $|V^{\text{SN}} - \tilde{V}_{T_\varepsilon}^{\text{SN}}| \leq \varepsilon$. We note that this comes by a direct application of Theorem 4 where the range $C = 2$, since $V^{\text{SN}} \leq 2$ a.s.

Similarly, we have a stopping criterion for U^{SN}, that is we stop when

$$\varepsilon \geq \sqrt{\frac{2\widehat{\text{Var}}(\tilde{U}_t^{\text{SN}})}{t}} + \frac{7}{3} \cdot \frac{2x}{t-1}$$

is satisfied. This gives w.h.p $|U^{\text{SN}} - \tilde{U}_{T_\varepsilon}^{\text{SN}}| \leq \varepsilon$.

In case of $\tilde{B}_T$, we control its simulation error indirectly through controlling an error $|Z_{T_\varepsilon}^{\text{inv}} - 1/Z| \leq \varepsilon$, i.e. stopping when

$$\varepsilon \geq \sqrt{\frac{2\widehat{\text{Var}}(Z_t^{\text{inv}})}{t}} + \frac{7}{3} \cdot \frac{Mx}{t-1} \,,$$

is satisfied, where $M = 1/\sum_i \min_{a \in [K]} \frac{\pi(a|X_i)}{\pi_b(a|X_i)}$ (note that $1/Z \leq M$ a.s. for fixed X_1^n). The reason for this becomes clear by observing a simple lower bound on B:

$$B = \min\left(1, \frac{1}{\mathbb{E}\left[\frac{n}{Z} \,\middle|\, X_1^n\right]}\right) \geq \min\left(1, \frac{1}{\mathbb{E}\left[n(Z_{T_\varepsilon}^{\text{inv}} + \varepsilon) \,\middle|\, X_1^n\right]}\right) \,.$$

Finally, we note that convergence of $\tilde{V}_t^{\text{SN}}$, $\tilde{U}_t^{\text{SN}}$, and $\tilde{B}_t$ might take different number of steps and in practice one would split Algorithm 1 into separate subroutines for estimation of respective quantities with different stopping criteria. As mentioned before the sample variance can be easily computed online, for instance by using Welford's method.

B Additional proofs

B.1 Proofs from Section 4

To prove Proposition 5 we will need the following statement:

Proposition 4. *Let $S = ((W_i, R_i))_{i=1}^n$ be independent random variables distributed according to some probability measure on $\mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$, let $f(S) = \frac{\sum_{i=1}^n W_i R_i}{\sum_{i=1}^n W_i}$, and $f_k(S^{(k)}) = \frac{\sum_{i\neq k} W_i R_i}{\sum_{i\neq k}^n W_i}$. Let $E_k = R_k - f_k(S^{\setminus k})$. Then for all $k \in [n]$,*

$$f(S) - f_k(S^{\setminus k}) = \frac{W_k E_k}{\sum_{i=1}^n W_i} \ .$$

Proposition 5. *Let $f(S) = \frac{\sum_{i=1}^n W_i R_i}{\sum_{i=1}^n W_i}$. Then,*

$$\sum_{k=1}^n \mathbb{E}\left[(f(S) - f(S^{(k)}))^2 \,\middle|\, W_1^k, X_1^n\right] \leq V^{\mathrm{SN}} = \sum_{k=1}^n \mathbb{E}\left[\left(\frac{W_k}{\sum_{i=1}^n W_i} + \frac{W_k'}{W_k' + \sum_{i\neq k} W_i}\right)^2 \,\middle|\, W_1^k, X_1^n\right] \ .$$

Proof. Denote

$$\tilde{W}_k = \frac{W_k}{\sum_{i=1}^n W_i} \ , \qquad \tilde{U}_k = \frac{W_k'}{W_k' + \sum_{i\neq k} W_i} \qquad k \in [n] \ .$$

By Proposition 4

$$\begin{aligned} f(S) - f(S^{(k)}) &= f(S) - f_k(S^{\setminus k}) + f_k(S^{\setminus k}) - f(S^{(k)}) \\ &= \frac{W_k E_k}{\sum_{i=1}^n W_i} - \frac{W_k' E_k'}{W_k' + \sum_{i\neq k} W_i} = \tilde{W}_k E_k - \tilde{U}_k E_k' \end{aligned}$$

where $E_k' = R_k' - f_k(S^{\setminus k})$. Taking square on both sides gives

$$\begin{aligned} \left(f(S) - f(S^{(k)})\right)^2 &= \tilde{W}_k^2 E_k^2 + \tilde{U}_k^2 (E_k')^2 - 2\tilde{W}_k \tilde{U}_k E_k E_k' \\ &\leq \tilde{W}_k^2 + \tilde{U}_k^2 + 2\tilde{W}_k \tilde{U}_k \qquad \text{(Since } E_k, E_k' \in [-1,1] \text{ a.s.)} \\ &= \left(\tilde{W}_k + \tilde{U}_k\right)^2 \ . \end{aligned}$$

□

Proof. From simple algebra (see Proposition 2 in (Kuzborskij and Szepesvári, 2019) discussion), we have

$$f(S) - f_k(S^{\setminus k}) = \frac{W_k(R_k - f_k(S^{\setminus k}))}{\sum_{i=1}^n W_i} \leq \frac{W_k}{\sum_{i=1}^n W_i} \qquad k \in [n] \ .$$

Then, the desired result follows from an application of Proposition 4, with $f = \hat{v}^{\mathrm{SN}}$ and $S = ((W_1, R_1), \ldots, (W_n, R_n))$, given the contexts. □

B.2 Polynomial Bounds for Weighted Importance Sampling

Since the exact calculation of V^{SN} could be prohibitive, we use a shortcut to lower bound the denominator $\sum_i W_i$. The promised lower bound is based on the following (more or less standard) result:

Lemma 1. *Assume that the non-negative random variables $W_1, W_2, \ldots, W_n$ are distributed independently from each other given $\mathcal{F}_0$. Then, for any $t \in [0, \sum_{k=1}^n \mathbb{E}[W_k \mid \mathcal{F}_0])$,*

$$\mathbb{P}\left(\sum_{i=1}^n W_i \leq t \,\middle|\, \mathcal{F}_0\right) \leq \exp\left(-\frac{\left(t - \sum_{k=1}^n \mathbb{E}\left[W_k \mid \mathcal{F}_0\right]\right)^2}{2\sum_{k=1}^n \mathbb{E}\left[W_k^2 \mid \mathcal{F}_0\right]}\right)$$

and in particular for any $x > 0$, with probability at least $1 - e^{-x}$,

$$\sum_{i=1}^{n} W_i > \sum_{k=1}^{n} \mathbb{E}[W_k \mid \mathcal{F}_0] - \sqrt{2x \sum_{k=1}^{n} \mathbb{E}\left[W_k^2 \mid \mathcal{F}_0\right]} \,. \tag{7}$$

Proof. We drop conditioning on $\mathcal{F}_0$ to simplify notation. Chernoff bound readily gives a bound on the lower tail

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq t\right) \leq \inf_{\lambda > 0} e^{\lambda t}\, \mathbb{E}\left[e^{-\lambda \sum_{i=1}^{n} X_i}\right] \,.$$

By independence of X_i

$$\begin{aligned}
\prod_{i=1}^{n} \mathbb{E}\left[e^{-\lambda X_i}\right] &\leq \prod_{i=1}^{n} \left(1 - \lambda\, \mathbb{E}\left[X_i\right] + \frac{\lambda^2}{2}\, \mathbb{E}\left[X_i^2\right]\right) && (e^{-x} \leq 1 - x + \tfrac{1}{2}x^2 \text{ for } x \geq 0) \\
&\leq e^{-\lambda \sum_{i=1}^{n} \mathbb{E}[X_1] + \frac{\lambda^2}{2} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]} && (1 + x \leq e^x \text{ for } x \in \mathbb{R} \text{ and i.i.d. assumption})
\end{aligned}$$

Getting back to the Chernoff bound gives,

$$\lambda = \max\left\{\frac{\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] - t}{\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]}, 0\right\} \,.$$

This proves the first result. The second result comes by inverting the bound and solving a quadratic equation. □

Proposition 3 (restated). *With probability at least $1 - 3e^{-x}$ for $x > 0$,*

$$v(\pi) \geq \frac{N_x}{n}\left(\hat{v}^{\mathrm{SN}}(\pi) - \sqrt{\frac{\sum_{k=1}^{n} \mathbb{E}[W_k^2 | X_k]}{N_x^2}\, e^x}\right) - \sqrt{\frac{x}{2n}} \,.$$

where

$$N_x = n - \sqrt{2x \sum_{k=1}^{n} \mathbb{E}\left[W_k^2 \mid X_1^n\right]} \,.$$

Proof of Proposition 3. The decomposition into the bias and the concentration is as in the proof of Theorem 1, where the concentration of contexts is handled once again through Hoeffding's inequality. Hence, we'll focus only on the concentration.

Let $Z = \hat{v}^{\mathrm{SN}}(\pi) - \mathbb{E}[\hat{v}^{\mathrm{SN}}(\pi)]$. Chebyshev's inequality gives us:

$$\mathbb{P}\left(|Z| \geq \sqrt{t \mathrm{Var}(\hat{v}^{\mathrm{SN}}(\pi) \mid X_1^n)}\right) \leq \frac{1}{t} \qquad t > 0 \,.$$

This implies

$$\begin{aligned}
|\hat{v}^{\mathrm{SN}}(\pi) - \mathbb{E}[\hat{v}^{\mathrm{SN}}(\pi)]| &\leq \sqrt{e^x\, \mathrm{Var}(\hat{v}^{\mathrm{SN}}(\pi) \mid X_1^n)} && (\text{w.p. at least } 1 - e^{-x},\ x > 0) \\
&\leq \sqrt{\frac{e^x}{N_x^2} \sum_{k=1}^{n} \mathbb{E}[W_k^2 | X_k]} && (\text{w.p. at least } 1 - 2e^{-x} \text{ (union bound)})
\end{aligned}$$

where by Efron-Stein's inequality and Proposition 2 of Kuzborskij and Szepesvári (2019):

$$\mathrm{Var}(\hat{v}^{\mathrm{SN}}(\pi) \mid X_1^n) \leq \mathbb{E}\left[\sum_{k=1}^{n} \left(\hat{v}_S^{\mathrm{SN}}(\pi) - \hat{v}_{S^{\setminus k}}^{\mathrm{SN}}(\pi)\right)^2 \,\middle|\, X_1^n\right] \leq \mathbb{E}\left[\frac{\sum_{k=1}^{n} W_k^2}{(\sum_{i=1}^{n} W_i)^2} \,\middle|\, X_1^n\right]$$

and a lower bound on the sum of weights comes from Lemma 1. □

B.3 Confidence Bound for λ-Corrected Importance Sampling Estimator

Recall the following empirical Bernstein bound given in Theorem 4.

Theorem 4 (Maurer and Pontil (2009)[5]). *Let $Z, Z_1, \dots, Z_n$ be i.i.d. random variables with values in $[0, C]$ and let $x > 0$. Then with probability at least $1 - 2e^{-x}$,*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z] \le \sqrt{\frac{2\mathrm{Var}(Z_1, \dots, Z_n)x}{n}} + \frac{7Cx}{3(n-1)}$$

where sample variance is defined as

$$\mathrm{Var}(Z_1, \dots, Z_n) = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} (Z_i - Z_j)^2 \ . \tag{8}$$

The following proposition states a concentration bound for the value when using the λ-IW estimator.

Proposition 1 (restated). *For the λ-IW estimator we have with probability at least $1 - 3e^{-x}$, for $x > 0$,*

$$v(\pi) \ge \hat{v}^{\text{IW-}\lambda}(\pi) - \sqrt{\frac{2x}{n}\mathrm{Var}(\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n)} - \frac{7x}{3\lambda(n-1)} - \frac{1}{n}\sum_{k=1}^{n}\sum_{a \in [K]} \pi(a|X_k)\left|\frac{\pi_b(a \mid X_k)}{\pi_b(a \mid X_k) + \lambda} - 1\right| - \sqrt{\frac{x}{2n}} \ .$$

and the variance of the estimator is defined as

$$\mathrm{Var}(\hat{v}^{\text{IW-}\lambda}(\pi)) = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \left(W_i^\lambda R_i - W_j^\lambda R_j\right)^2 \ . \tag{9}$$

Proof. We start with the decomposition

$$v(\pi) - \hat{v}^{\text{IW-}\lambda}(\pi) = \underbrace{v(\pi) - \mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n\right]}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n\right] - \hat{v}^{\text{IW-}\lambda}(\pi)}_{\text{Concentration}} \ .$$

Observing that $W_k^\lambda \le 1/\lambda$ The concentration term is bounded by Theorem 4 with $C = 1/\lambda$, that is:

$$\mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n\right] - \hat{v}^{\text{IW-}\lambda}(\pi) \ge \sqrt{\frac{2x}{n}\mathrm{Var}(\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n)} + \frac{7x}{3\lambda(n-1)} \ .$$

Now we focus on the bias term which is further decomposed as follows:

$$v(\pi) - \mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi)|X_1^n\right] = v(\pi) - \frac{1}{n}\sum_{k=1}^{n} v(\pi|X_k) + \frac{1}{n}\sum_{k=1}^{n} v(\pi|X_k) - \mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi)|X_1^n\right]$$

Since $(v(\pi|X_k))_{k\in[n]}$ are independent and they take values in the range $[0,1]$, by Hoeffding's inequality we have w.p. at least $1 - e^{-x}, x \ge 0$ that

$$\frac{1}{n}\sum_{k=1}^{n} v(\pi|X_k) - v(\pi) \le \sqrt{\frac{x}{2n}} \ .$$

[5] Maurer and Pontil (2009) stated inequality in another direction. However, we can show the one we stated by the symmetry of Bernstein's inequality.

Finally,

$$\begin{aligned}
\mathbb{E}\left[\hat{v}^{\text{IW-}\lambda}(\pi) \mid X_1^n\right] - \frac{1}{n}\sum_{k=1}^{n} v(\pi|X_k) &= \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\left(W_k^\lambda - W_k\right) R_k \mid X_k\right] \\
&= \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\left[\left(\frac{\pi(A_k \mid X_k)}{\pi_b(A_k \mid X_k) + \lambda} - \frac{\pi(A_k \mid X_k)}{\pi_b(A_k \mid X_k)}\right) R_k \,\middle|\, X_k\right] \\
&= \frac{1}{n}\sum_{k=1}^{n} \sum_{a\in[K]} \pi(a|X_k)\left(\frac{\pi_b(a \mid X_k)}{\pi_b(a \mid X_k) + \lambda} - 1\right) r(X_k, a) \\
&\leq \frac{1}{n}\sum_{k=1}^{n} \sum_{a\in[K]} \pi(a|X_k)\left|\frac{\pi_b(a \mid X_k)}{\pi_b(a \mid X_k) + \lambda} - 1\right| .
\end{aligned}$$

Putting all together and applying a union bound we get the statement w.p. at least $1 - 3e^{-x}$. □

B.4 Confidence Bound for λ-Corrected Doubly-Robust Estimator

Doubly-Robust (DR) estimators were introduced in the machine learning literature for off-policy evaluation by Dudík et al. (2011), and refined in Farajtabar et al. (2018); Su et al. (2019b). They combine a direct model estimator and IW, finding a compromise that should behave like IW with a reduced variance. To compute $\hat{v}^{\text{DR}}$, a reward estimator $\eta : \mathcal{X} \times [K] \to [0, 1]$ must be learned on a subset of the logged dataset. Then,

$$\hat{v}^{\text{DR}}(\pi) = \hat{V}_\eta(\pi) + \frac{1}{n}\sum_{i=1}^{n} W_i(R_i - \eta(X_i, A_i)),$$

where $\hat{V}_\eta(\pi) = (1/n)\sum_{i=1}^{n}\sum_{a\in[K]} \pi(a|X_i)\eta(X_i, a)$ is the expected reward of π given η.[6] Now we prove a very similar bound for the λ-Corrected Doubly-Robust estimator.

Proposition 2 (restated). *For the λ-DR estimator defined w.r.t. a fixed $\eta : \mathcal{X} \times [K] \to [0, 1]$ we have with probability at least $1 - 3e^{-x}$, for $x > 0$,*

$$\begin{aligned}
v(\pi) \geq \hat{v}^{\text{DR-}\lambda}(\pi) &- \sqrt{\frac{2x}{n}\operatorname{Var}(\hat{v}^{\text{DR-}\lambda}(\pi) \mid X_1^n)} - \frac{7}{3}\left(1 + \frac{1}{\lambda}\right)\frac{x}{n-1} \\
&- \frac{1}{n}\sum_{k=1}^{n}\sum_{a\in[K]} \pi(a|X_k)\left(\left|\frac{\pi_b(a \mid X_k)}{\pi_b(a \mid X_k) + \lambda} - 1\right| + \eta(a|X_k)\left(1 - \frac{\pi(a \mid X_k)}{\pi_b(a \mid X_k) + \lambda}\right)\right) - \sqrt{\frac{x}{2n}} .
\end{aligned}$$

and the variance of the estimator is defined as

$$\operatorname{Var}(\hat{v}^{\text{DR-}\lambda}(\pi)) = \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n} (Z_i - Z_j)^2 \tag{10}$$

where $Z_i = W_i^\lambda(R_i - \eta(X_i, A_i)) + \sum_{a\in[K]} \pi(a|X_i)\eta(a, X_i)$.

Proof. We follow the path in as in the proof of Proposition 1 with minor modifications. Once again, considering the decomposition

$$v(\pi) - \hat{v}^{\text{DR-}\lambda}(\pi) = \underbrace{v(\pi) - \mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi) \mid X_1^n\right]}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi) \mid X_1^n\right] - \hat{v}^{\text{DR-}\lambda}(\pi)}_{\text{Concentration}} .$$

and observing that $W_k^\lambda \leq 1/\lambda$, the concentration term is bounded by Theorem 4 with $C = 1 + 1/\lambda$ assuming that $\|\eta\|_\infty \leq 1$, that is:

$$\mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi) \mid X_1^n\right] - \hat{v}^{\text{DR-}\lambda}(\pi) \geq \sqrt{\frac{2x}{n}\operatorname{Var}(\hat{v}^{\text{DR-}\lambda}(\pi) \mid X_1^n)} + \frac{7}{3}\left(1 + \frac{1}{\lambda}\right)\frac{x}{n-1} .$$

[6] Note that since rewards can be negative, it is not clear how to incorporate DR estimation into our Theorem 1 due to use of Harris' inequality in Eq. (6).

Now we focus on the bias term which is further decomposed as follows:

$$v(\pi) - \mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi)|X_1^n\right] = v(\pi) - \frac{1}{n}\sum_{k=1}^n v(\pi|X_k) + \frac{1}{n}\sum_{k=1}^n v(\pi|X_k) - \mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi)|X_1^n\right]$$

As in the proof of Proposition 1 w.p. at least $1-e^{-x}, x \geq 0$ we have

$$\frac{1}{n}\sum_{k=1}^n v(\pi|X_k) - v(\pi) \leq \sqrt{\frac{x}{2n}}\,.$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\hat{v}^{\text{DR-}\lambda}(\pi) \;\middle|\; X_1^n\right] - \frac{1}{n}\sum_{k=1}^n v(\pi|X_k) \\
&= \frac{1}{n}\sum_{k=1}^n \left(\mathbb{E}\left[W_k^\lambda(R_k - \eta(X_k,A_k)) - W_k R_k \;\middle|\; X_k\right] + \sum_{a\in[K]}\pi(a|X_k)\eta(a,X_k)\right) \\
&= \frac{1}{n}\sum_{k=1}^n \left(\mathbb{E}\left[(W_k^\lambda - W_k)R_k - W_k^\lambda \eta(X_k,A_k) \;\middle|\; X_k\right] + \sum_{a\in[K]}\pi(a|X_k)\eta(a,X_k)\right) \\
&= \frac{1}{n}\sum_{k=1}^n \sum_{a\in[K]} \pi(a|X_k)\left(\frac{\pi_b(a\mid X_k)}{\pi_b(a\mid X_k)+\lambda} - 1\right) r(X_k,a) \\
&+ \frac{1}{n}\sum_{k=1}^n \sum_{a\in[K]} \pi(a|X_k)\eta(a|X_k)\left(1 - \frac{\pi(a\mid X_k)}{\pi_b(a\mid X_k)+\lambda}\right) \\
&\leq \frac{1}{n}\sum_{k=1}^n \sum_{a\in[K]} \pi(a|X_k)\left(\left|\frac{\pi_b(a\mid X_k)}{\pi_b(a\mid X_k)+\lambda} - 1\right| + \eta(a|X_k)\left(1 - \frac{\pi(a\mid X_k)}{\pi_b(a\mid X_k)+\lambda}\right)\right)
\end{aligned}$$

Putting all together and applying a union bound we get the statement w.p. at least $1-3e^{-x}$. □

C Additional Experimental Details

C.1 Policies.

Parametrized oracle-based policies. For a given dataset $((\boldsymbol{x}_i, y_i))_{i=1}^n \subset (\mathcal{X}\times\mathcal{Y})^n$, we assume we have access to an oracle $\rho : \mathcal{X} \to \mathcal{Y}$ that maps contexts to their true label[7]. We define an ideal Gibbs policy as $\pi^{\text{ideal}}(y \mid \boldsymbol{x}) \propto e^{\frac{1}{\tau}\mathbb{I}\{y=\rho(\boldsymbol{x})\}}$ and $\tau > 0$ is a temperature parameter. The smaller τ is, the more peaky is the distribution on the predicted label. To create mismatching policies, we consider a *faulty* policy type for which the peak is shifted to another, wrong action for a set of faulty actions $F \subset [K]$ (i.e., if $\rho(\boldsymbol{x}) \in F$, the peak is shifted by 1 cyclically), that is, a faulty policy $\pi^{\text{faulty}(F)}$ is the same as the ideal policy when $\rho(\boldsymbol{x}) \notin F$, and it has distribution $\pi^{\text{faulty}(F)}(y \mid \boldsymbol{x}) \propto e^{\frac{1}{\tau}\mathbb{I}\{y-1=\rho(\boldsymbol{x}) \bmod K\}}$.

In the following we consider faulty behavior policies, while one among the target policies is *ideal*.

Learnt policies. There is an important literature on off-policy learning (Swaminathan and Joachims, 2015a,c; Joachims et al., 2018) that considers the problem of directly learning a policy from logged bandit feedback. These algorithms minimize a loss defined by either IW or SN on a parametrized family of policy. We implements those two type of parametrized policies as follows: we introduce $\pi^{\boldsymbol{\Theta}}(y = k \mid \boldsymbol{x}) \propto e^{\frac{1}{\tau}\boldsymbol{x}^\top \boldsymbol{\theta}_k}$ with two choices of parameters given by the optimization problems: $\hat{\boldsymbol{\Theta}}_{\text{IW}} \in \arg\max_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times K}} \hat{v}^{\text{IW}}(\pi^{\boldsymbol{\Theta}})$, $\hat{\boldsymbol{\Theta}}_{\text{SN}} \in \arg\max_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times K}} \hat{v}^{\text{SN}}(\pi^{\boldsymbol{\Theta}})$. In practice we obtain these by running gradient descent with step size 0.01 for 10^5 steps. In all cases the temperature is set to $\tau = 0.1$.

[7] In general, this oracle has to be learnt, see discussions on datasets below.

C.2 Datasets and oracles

Synthetic dataset. To allow for a precise control of the distribution of the contexts, as well as of the sample size, we generate an underlying multiclass classification problem through the scikit-learn function `make_classification()` [8]. Then we obtain a ground truth oracle by training a classifier $\hat{r}$ with a regularized logistic regression (with hyperparameter tuned on the validation set).

Real Datasets. The chosen 8 datasets (see Table 3 in Appendix C) are loaded from OpenML (Dua and Graff, 2017), using scikit-learn (Pedregosa et al., 2011). To simplify and stabilize the Gibbs policy construction process, we use the true labels as the peaks of the Gibbs oracle. In the literature on off-policy evaluation, some experimental settings rely on a ground truth function, which is a multi-class classifier learned on a held-out full-information dataset. This ground truth then replaces the true labels in the policies. Depending on the accuracy of the learnt function, this might naturally induce noise in the policies by having them make mistakes due to a relatively bad oracle. Note that in the case of synthetic datasets, it is easy and costless to generate a large train set, get a highly accurate classifier, and discard this data. However, for real datasets, the more data is used for training the oracle, the less is available to generate a logged dataset and perform the actual off-policy evaluation experiments.

While this moves the process away from practice, it has the advantage of allowing a precise control of the values of the policies we create. This is a key point to design stable and reproducible experiments. Learning perfectly interpolating classifiers would lead to the same results, except for the time spent and the data used to do so.

Baselines. In addition to the confidence bound discussed in Section 5 we consider the standard DR estimator and the recent estimation algorithm of Karampatziakis et al. (2019) based on Empirical Likelihood (EL). For DR (and λ-DR), rewards are modeled by a ridge regressor (one per class) where a hyperparameter is tuned by a 10-fold cross-validation (leave-one-out cross-validation for sample size ≤ 100). For both λ-IW and λ-DR, λ is set to $1/\sqrt{n}$.

name	Yeast	PageBlok	OptDigits	SatImage	isolet	PenDigits	Letter	kropt
OpenML ID	181	30	28	182	300	32	6	184
Size	1484	5473	5620	6435	7797	10,992	20,000	28,056

Table 3: Real Datasets used in experiments

Empirical coverage analysis: the case of the Empirical Likelihood estimator. We run the same experiment as that presented in Figure 2 to study the tightness of the returned lower bound for each estimator: the Gibbs temperature is $\tau = 0.3$ and the sample size is $N = 1000$ (new dataset for each run), so that the Effective Sample Size is on average 650 ± 10. Results are shown on Figure 3. These simulations highlight two interesting facts that make EL a slightly different solution to our problem than all other state-of-the-art estimators. First, the returned lower bound is always very close to the true value, and on average just slightly under it. But while this should be a perfect property for our task, the returned value also suffers from quite a large variance such that in many runs the lower bound is larger than the true value (the confidence interval is violated). This seems to indicate that our setting has not yet reached the asymptotic regime in which the confidence interval should have a coverage probability close to $1 - \delta$. We conjecture that this may explain the bad performance of EL in our experiments on data (see Section 6).

D Implementation of Algorithm 1

In this section we provide a code listing in Python for computing the bound of Theorem 1. In particular, the function `eslb(...)` implements computation of the bound through the Monte-Carlo simulation described in Algorithm 1. The function `eslb(t_probs, b_probs, weights, rewards, delta, n_iterations, n_batch_size)` takes 7 arguments: `t_prob` and `b_prob` are $[0,1]^{n \times K}$ matrices where the i-th row corresponds to $\pi(\cdot|X_i)$ and $\pi_b(\cdot|X_i)$ respectively. Next, `weights` is an vector of importance weights belonging to $\mathbb{R}_+^n$, similarly `rewards` is a reward vector in $[0,1]^n$, and $\delta \in (0,1)$ is an error probability (recall that the lower bound holds with probability at least

[8] See scikit-learn documentation

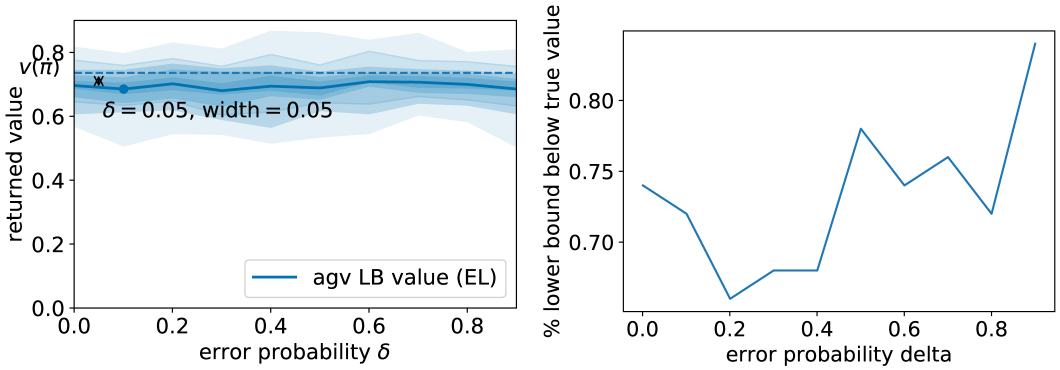

Figure 4: **Left:**Empirical tightness of the EL lower bound estimator. The returned value is on average very close to the true value. **Right:** Average rate of violated confidence interval: the empirical coverage is much worse than the true one (δ). Results averaged over 50 runs

$1-\delta$). Finally, `n_iterations` and `n_batch_size` are Monte-Carlo iterations and the sample size (batch size) used in the simulation (larger `n_batch_size` requires more memory but ensures faster convergence of the simulation). `eslb(...)` returns a Python dictionary holding 5 enries: entry `lower_bound` corresponds to the actual lower bound computed according to Theorem 1; `est_value` is $\hat{v}(\pi)$, `concentration` is a concentration term denoted by ϵ in Theorem 1, `mult_bias` is a multiplicative bias denoted by B, and `concentration_of_contexts` is a $\sqrt{x/(2n)}$ term.

Listing 1: Computation of the bound of Theorem 1: "eslb(...)" function.

```
# Copyright 2020 DeepMind Technologies Limited.
#
#
# Licensed under the Apache License, Version 2.0 (the "License");
# you may not use this file except in compliance with the License.
# You may obtain a copy of the License at
#
# https://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

from __future__ import division
from math import sqrt, log as ln
import numpy as np

def sample_from_simplices_m_times(p, m):
    """ Sample from n probability simplices m times.

    p -- n times K (matrix where earch row describes a probability simplex)
    m -- number of times to sample

    Returns n-times-m matrix of indices of simplex corners.
    """
    axis = 1
    r = np.expand_dims(np.random.rand(p.shape[1-axis], m), axis=axis)
    p_ = np.expand_dims(p.cumsum(axis=axis), axis=2)
    return (np.repeat(p_, m, axis=2) > r).argmax(axis=1)

def eslb(t_probs, b_probs, weights, rewards, delta, n_iterations, n_batch_size):
    """ Computes Efron-Stein lower bound of Theorem 1 as described in Algorithm 1.
    Here n is a sample size, while K is a number actions.

    t_probs -- n-times-K matrix, where $i$-th row correponds to $\pi(\cdot | X_i)$
    b_probs -- n-times-K matrix, where $i$-th row correponds to $\pi_b(\cdot | X_i)$
    weights -- n-sized vector of importance weights
    rewards -- n-sized reward vector
    delta -- error probability in (0,1)
    n_iterations -- Monte-Carlo simulation iterations
    n_batch_size -- Monte-Carlo simulation batch size

    Returns dictionary with 5 enries: lower_bound corresponds to the actual lower bound;
    est_value is an empirical value, concentration is a concentration term, mult_bias
    is a multiplicative bias, and while concentration_of_contexts is a term responsible
    for concentration of contexts.
    """
    conf = ln(2.0/delta)
    n = len(weights)
    ix_1_n = np.arange(n)
    W_cumsum = weights.cumsum()
    W_cumsum = np.repeat(np.expand_dims(W_cumsum, axis=1), n_batch_size, axis=1)
    W = np.repeat(np.expand_dims(weights, axis=1), n_batch_size, axis=1)

    weight_table = t_probs / b_probs

    V_unsumed = np.zeros((n,))
    E_V_unsumed = np.zeros((n,))
    Ehat_recip_W = 0.0

    for i in range(n_iterations):
        A_sampled = sample_from_simplices_m_times(b_probs, n_batch_size)
        W_sampled = weight_table[ix_1_n, A_sampled.T].T
        W_sampled_cumsum = W_sampled[::-1, :].cumsum(axis=0)[::-1, :]
        Z = np.copy(W_cumsum)
        Z[:-1, :] += W_sampled_cumsum[1:, :]

        A_sampled_for_U = sample_from_simplices_m_times(b_probs, n_batch_size)
        W_sampled_for_U = weight_table[ix_1_n, A_sampled_for_U.T].T
        A_sampled_for_B = sample_from_simplices_m_times(b_probs, n_batch_size)
        W_sampled_for_B = weight_table[ix_1_n, A_sampled_for_B.T].T

        Z_repk = Z - W + W_sampled
        W_tilde = W / Z
        U_tilde = W_sampled_for_U / Z_repk
        V_t = (W_tilde + U_tilde)**2

        E_V_new_item = ((W_sampled / W_sampled.sum(axis=0))**2).mean(axis=1)
        V_new_item = V_t.mean(axis=1)
```

```
        E_V_unsumed += (E_V_new_item - E_V_unsumed) / (i+1)
        V_unsumed += (V_new_item - V_unsumed) / (i+1)

        B_t = (1.0/W_sampled_for_B.sum(axis=0)).mean()
        Ehat_recip_W += (B_t - Ehat_recip_W) / (i+1)

    V = V_unsumed.sum()
    E_V = E_V_unsumed.sum()

    eff_N = 1.0 / Ehat_recip_W

    mult_bias = min(1.0, eff_N / n)
    concentration = sqrt(2.0 * (V + E_V) * (conf + 0.5 * ln(1 + V/E_V)))
    concentration_of_contexts = sqrt(conf / (2*n))
    est_value = weights.dot(rewards) / weights.sum()
    lower_bound = mult_bias * (est_value - concentration ) - concentration_of_contexts

    return dict(lower_bound=max(0, lower_bound), est_value=est_value, concentration=concentration,
                mult_bias=mult_bias, concentration_of_contexts=concentration_of_contexts)
```

E Value Bound Decomposition

In this section we present the decomposition of each confidence bound on the value for various estimators evaluated in Section 6. In particular, the decomposition is done w.r.t. the respective lower bounds on the concentration, bias, and concentration of contexts terms:

$$v(\pi) - \hat{v}(\pi) = \underbrace{v(\pi) - \mathbb{E}\left[v(\pi) \,|\, X_1^n\right]}_{\text{Concentration of contexts}} + \underbrace{\mathbb{E}\left[v(\pi) \,|\, X_1^n\right] - \mathbb{E}\left[\hat{v}(\pi) \mid X_1^n\right]}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\hat{v}(\pi) \mid X_1^n\right] - \hat{v}(\pi)}_{\text{Concentration}} .$$

In the following tables each term is presented w.r.t. three target policies discussed in Section 6.1: That is `Ideal` is π^{ideal}, while `Gibbs-fitted-IW` is $\pi^{\hat{\Theta}_{\text{IW}}}$, and `Gibbs-fitted-SN` is $\pi^{\hat{\Theta}_{\text{SN}}}$.

E.1 Synthetic Dataset

Table 4: Concentration term ϵ for different confidence intervals and target policies.

Concentration	5000	10000	20000
ESLB: Ideal	0.680 ± 0.061	0.497 ± 0.019	0.346 ± 0.013
ESLB: Gibbs-fitted-IW	0.650 ± 0.079	0.498 ± 0.018	0.363 ± 0.011
ESLB: Gibbs-fitted-SN	0.770 ± 0.068	0.561 ± 0.029	0.378 ± 0.017
λ-IW: Ideal	0.346 ± 0.023	0.271 ± 0.008	0.206 ± 0.007
λ-IW: Gibbs-fitted-IW	0.350 ± 0.024	0.274 ± 0.008	0.208 ± 0.007
λ-IW: Gibbs-fitted-SN	0.348 ± 0.024	0.273 ± 0.008	0.207 ± 0.007
Cheb-SN: Ideal	5.437 ± 0.000	3.242 ± 0.000	2.030 ± 0.000
Cheb-SN: Gibbs-fitted-IW	4.006 ± 0.438	2.969 ± 0.086	2.034 ± 0.026
Cheb-SN: Gibbs-fitted-SN	6.991 ± 0.227	3.982 ± 0.122	2.344 ± 0.027
DR: Ideal	-	-	-
DR: Gibbs-fitted-IW	-	-	-
DR: Gibbs-fitted-SN	-	-	-
λ-DR: Ideal	0.412 ± 0.018	0.305 ± 0.018	0.218 ± 0.009
λ-DR: Gibbs-fitted-IW	0.435 ± 0.017	0.310 ± 0.023	0.228 ± 0.006
λ-DR: Gibbs-fitted-SN	0.416 ± 0.030	0.310 ± 0.013	0.210 ± 0.010
Emp.Lik. Ideal	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-

Table 5: Bias term B for different confidence intervals and target policies.

Bias (multiplicative for SN-based CIs)	5000	10000	20000
ESLB: Ideal	0.988 ± 0.000	0.994 ± 0.000	0.997 ± 0.000
ESLB: Gibbs-fitted-IW	0.992 ± 0.001	0.995 ± 0.000	0.997 ± 0.000
ESLB: Gibbs-fitted-SN	0.984 ± 0.001	0.992 ± 0.001	0.996 ± 0.000
λ-IW: Ideal	0.293 ± 0.000	0.261 ± 0.000	0.219 ± 0.000
λ-IW: Gibbs-fitted-IW	0.212 ± 0.025	0.232 ± 0.008	0.217 ± 0.004
λ-IW: Gibbs-fitted-SN	0.393 ± 0.011	0.347 ± 0.014	0.272 ± 0.005
Cheb-SN: Ideal	0.599 ± 0.000	0.715 ± 0.000	0.800 ± 0.000
Cheb-SN: Gibbs-fitted-IW	0.671 ± 0.024	0.733 ± 0.006	0.800 ± 0.002
Cheb-SN: Gibbs-fitted-SN	0.538 ± 0.008	0.671 ± 0.007	0.776 ± 0.002
DR: Ideal	-	-	-
DR: Gibbs-fitted-IW	-	-	-
DR: Gibbs-fitted-SN	-	-	-
λ-DR: Ideal	0.515 ± 0.059	0.540 ± 0.064	0.590 ± 0.046
λ-DR: Gibbs-fitted-IW	0.430 ± 0.095	0.570 ± 0.063	0.679 ± 0.045
λ-DR: Gibbs-fitted-SN	0.767 ± 0.113	0.744 ± 0.086	0.790 ± 0.038
Emp.Lik. Ideal	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-

Table 6: Concentration of contexts term for different confidence intervals and target policies.

Concentration of contexts	5000	10000	20000
ESLB: Ideal	0.025 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
ESLB: Gibbs-fitted-IW	0.025 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
ESLB: Gibbs-fitted-SN	0.025 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
λ-IW: Ideal	0.026 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
λ-IW: Gibbs-fitted-IW	0.026 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
λ-IW: Gibbs-fitted-SN	0.026 ± 0.000	0.018 ± 0.000	0.013 ± 0.000
Cheb-SN: Ideal	0.300 ± 0.000	0.212 ± 0.000	0.150 ± 0.000
Cheb-SN: Gibbs-fitted-IW	0.300 ± 0.000	0.212 ± 0.000	0.150 ± 0.000
Cheb-SN: Gibbs-fitted-SN	0.300 ± 0.000	0.212 ± 0.000	0.150 ± 0.000
DR: Ideal	-	-	-
DR: Gibbs-fitted-IW	-	-	-
DR: Gibbs-fitted-SN	-	-	-
λ-DR: Ideal	0.037 ± 0.000	0.026 ± 0.000	0.018 ± 0.000
λ-DR: Gibbs-fitted-IW	0.037 ± 0.000	0.026 ± 0.000	0.018 ± 0.000
λ-DR: Gibbs-fitted-SN	0.037 ± 0.000	0.026 ± 0.000	0.018 ± 0.000
Emp.Lik. Ideal	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-

Table 7: Empirical value $\hat{v}$ for different estimators and target policies.

$\hat{v}(\pi)$	5000	10000	20000
ESLB: Ideal	0.973 ± 0.005	0.974 ± 0.002	0.974 ± 0.002
ESLB: Gibbs-fitted-IW	0.822 ± 0.059	0.901 ± 0.033	0.901 ± 0.016
ESLB: Gibbs-fitted-SN	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000
λ-IW: Ideal	0.691 ± 0.044	0.727 ± 0.021	0.760 ± 0.026
λ-IW: Gibbs-fitted-IW	0.689 ± 0.045	0.732 ± 0.022	0.745 ± 0.027
λ-IW: Gibbs-fitted-SN	0.546 ± 0.040	0.587 ± 0.020	0.659 ± 0.024
Cheb-SN: Ideal	0.973 ± 0.005	0.974 ± 0.002	0.974 ± 0.002
Cheb-SN: Gibbs-fitted-IW	0.822 ± 0.059	0.901 ± 0.033	0.901 ± 0.016
Cheb-SN: Gibbs-fitted-SN	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000
DR: Ideal	-	-	-
DR: Gibbs-fitted-IW	-	-	-
DR: Gibbs-fitted-SN	-	-	-
λ-DR: Ideal	0.771 ± 0.052	0.828 ± 0.039	0.879 ± 0.030
λ-DR: Gibbs-fitted-IW	0.769 ± 0.049	0.837 ± 0.038	0.867 ± 0.032
λ-DR: Gibbs-fitted-SN	0.780 ± 0.066	0.791 ± 0.042	0.882 ± 0.025
Emp.Lik. Ideal	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-

E.2 UCI Datasets

Table 8: Concentration term for different confidence intervals and target policies.

Concentration / Name Size	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	isolet 7797	PenDigits 10992	Letter 20000	kropt 28056
ESLB: Ideal	0.424 ± 0.003	0.384 ± 0.038	0.672 ± 0.052	0.778 ± 0.020	0.402 ± 0.041	0.494 ± 0.018	0.241 ± 0.018	0.279 ± 0.006
ESLB: Gibbs-fitted-IW	2.056 ± 0.082	0.839 ± 0.054	0.858 ± 0.025	0.810 ± 0.088	0.827 ± 0.006	0.676 ± 0.047	0.665 ± 0.013	0.595 ± 0.009
ESLB: Gibbs-fitted-SN	3.060 ± 0.496	0.993 ± 0.154	1.640 ± 0.204	1.242 ± 0.144	1.229 ± 0.033	0.996 ± 0.148	0.833 ± 0.035	0.709 ± 0.023
λ-IW: Ideal	0.633 ± 0.001	0.352 ± 0.020	0.385 ± 0.029	0.400 ± 0.016	0.321 ± 0.021	0.301 ± 0.015	0.208 ± 0.012	0.196 ± 0.006
λ-IW: Gibbs-fitted-IW	0.821 ± 0.071	0.418 ± 0.022	0.501 ± 0.015	0.435 ± 0.022	0.570 ± 0.015	0.366 ± 0.009	0.347 ± 0.003	0.263 ± 0.004
λ-IW: Gibbs-fitted-SN	0.782 ± 0.056	0.393 ± 0.017	0.466 ± 0.015	0.428 ± 0.024	0.483 ± 0.023	0.337 ± 0.014	0.303 ± 0.006	0.243 ± 0.005
λ-DR: Ideal	0.927 ± 0.004	0.484 ± 0.025	0.524 ± 0.031	0.486 ± 0.017	0.430 ± 0.023	0.370 ± 0.017	0.263 ± 0.015	0.242 ± 0.013
λ-DR: Gibbs-fitted-IW	1.179 ± 0.051	0.550 ± 0.034	0.665 ± 0.027	0.555 ± 0.035	0.755 ± 0.015	0.451 ± 0.019	0.427 ± 0.017	0.344 ± 0.009
λ-DR: Gibbs-fitted-SN	0.985 ± 0.048	0.516 ± 0.024	0.544 ± 0.035	0.487 ± 0.023	0.527 ± 0.037	0.378 ± 0.024	0.304 ± 0.016	0.265 ± 0.010
Cheb-SN: Ideal	3.133 ± 0	2.464 ± 0	5.450 ± 0	6.852 ± 0	2.405 ± 0	3.258 ± 0	1.268 ± 0	1.541 ± 0
Cheb-SN: Gibbs-fitted-IW	$-\infty$	6.275 ± 0.994	3.720 ± 0.561	5.665 ± 2.055	2.629 ± 0.305	4.375 ± 0.727	5.652 ± 0.242	4.668 ± 0.126
Cheb-SN: Gibbs-fitted-SN	$-\infty$	11.510 ± 3.473	33.712 ± 4.930	17.818 ± 5.089	21.090 ± 0.994	8.806 ± 1.524	6.555 ± 0.139	4.779 ± 0.085
DR: Ideal	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-IW	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-SN	-	-	-	-	-	-	-	-
Emp.Lik. Ideal	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-	-	-	-	-	-

Table 9: Bias term for different confidence intervals and target policies.

Bias (multiplicative for SN-based CIs) / Name Size	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	isolet 7797	PenDigits 10992	Letter 20000	kropt 28056
ESLB: Ideal	0.996 ± 0.001	0.997 ± 0	0.989 ± 0.001	0.985 ± 0.001	0.997 ± 0.001	0.994 ± 0.001	0.999 ± 0	0.998 ± 0
ESLB: Gibbs-fitted-IW	0.892 ± 0.018	0.987 ± 0.003	0.993 ± 0.002	0.989 ± 0.005	0.996 ± 0.001	0.992 ± 0.002	0.988 ± 0.001	0.990 ± 0.001
ESLB: Gibbs-fitted-SN	0.789 ± 0.030	0.976 ± 0.005	0.951 ± 0.004	0.965 ± 0.008	0.962 ± 0.003	0.979 ± 0.004	0.985 ± 0.001	0.990 ± 0.001
λ-IW: Ideal	0.054 ± 0	0.067 ± 0	0.170 ± 0	0.244 ± 0	0.076 ± 0	0.151 ± 0	0.056 ± 0	0.097 ± 0
λ-IW: Gibbs-fitted-IW	0.487 ± 0.067	0.218 ± 0.033	0.118 ± 0.021	0.210 ± 0.079	0.091 ± 0.015	0.238 ± 0.047	0.509 ± 0.024	0.516 ± 0.016
λ-IW: Gibbs-fitted-SN	0.789 ± 0.058	0.361 ± 0.076	0.718 ± 0.052	0.554 ± 0.095	0.725 ± 0.007	0.506 ± 0.079	0.602 ± 0.015	0.528 ± 0.012
λ-DR: Ideal	0.123 ± 0.013	0.115 ± 0.011	0.319 ± 0.035	0.478 ± 0.040	0.313 ± 0.009	0.343 ± 0.033	0.306 ± 0.007	0.294 ± 0.006
λ-DR: Gibbs-fitted-IW	1.127 ± 0.174	0.439 ± 0.074	0.488 ± 0.099	0.557 ± 0.165	1.220 ± 0.077	0.706 ± 0.132	1.762 ± 0.101	1.533 ± 0.070
λ-DR: Gibbs-fitted-SN	1.443 ± 0.182	0.582 ± 0.119	1.305 ± 0.103	1.154 ± 0.205	1.439 ± 0.070	1.097 ± 0.157	1.395 ± 0.049	1.256 ± 0.032
Cheb-SN: Ideal	0.722 ± 0	0.767 ± 0	0.599 ± 0	0.543 ± 0	0.772 ± 0	0.714 ± 0	0.865 ± 0	0.841 ± 0
Cheb-SN: Gibbs-fitted-IW	0 ± 0	0.567 ± 0.043	0.688 ± 0.031	0.601 ± 0.080	0.756 ± 0.021	0.652 ± 0.038	0.590 ± 0.010	0.635 ± 0.006
Cheb-SN: Gibbs-fitted-SN	0 ± 0	0.426 ± 0.066	0.198 ± 0.033	0.325 ± 0.062	0.279 ± 0.010	0.484 ± 0.043	0.554 ± 0.005	0.630 ± 0.004
DR: Ideal	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-IW	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-SN	-	-	-	-	-	-	-	-
Emp.Lik. Ideal	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-	-	-	-	-	-

Table 10: Concentration of contexts term for different confidence intervals and target policies.

Concentration of contexts / Name Size	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	isolet 7797	PenDigits 10992	Letter 20000	kropt 28056
ESLB: Ideal	0.066 ± 0	0.034 ± 0	0.034 ± 0	0.032 ± 0	0.029 ± 0	0.024 ± 0	0.018 ± 0	0.015 ± 0
ESLB: Gibbs-fitted-IW	0.066 ± 0	0.034 ± 0	0.034 ± 0	0.032 ± 0	0.029 ± 0	0.024 ± 0	0.018 ± 0	0.015 ± 0
ESLB: Gibbs-fitted-SN	0.066 ± 0	0.034 ± 0	0.034 ± 0	0.032 ± 0	0.029 ± 0	0.024 ± 0	0.018 ± 0	0.015 ± 0
λ-IW: Ideal	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
λ-IW: Gibbs-fitted-IW	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
λ-IW: Gibbs-fitted-SN	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
λ-DR: Ideal	0.096 ± 0	0.050 ± 0	0.049 ± 0	0.046 ± 0	0.042 ± 0	0.035 ± 0	0.026 ± 0	0.022 ± 0
λ-DR: Gibbs-fitted-IW	0.096 ± 0	0.050 ± 0	0.049 ± 0	0.046 ± 0	0.042 ± 0	0.035 ± 0	0.026 ± 0	0.022 ± 0
λ-DR: Gibbs-fitted-SN	0.096 ± 0	0.050 ± 0	0.049 ± 0	0.046 ± 0	0.042 ± 0	0.035 ± 0	0.026 ± 0	0.022 ± 0
Cheb-SN: Ideal	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
Cheb-SN: Gibbs-fitted-IW	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
Cheb-SN: Gibbs-fitted-SN	0.068 ± 0	0.035 ± 0	0.035 ± 0	0.033 ± 0	0.030 ± 0	0.025 ± 0	0.018 ± 0	0.016 ± 0
DR: Ideal	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-IW	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-SN	-	-	-	-	-	-	-	-
Emp.Lik. Ideal	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-	-	-	-	-	-

Table 11: Empirical value $\hat{v}$ for different estimators and target policies.

$\hat{v}(\pi)$ / Name Size	Yeast 1484	PageBlok 5473	OptDigits 5620	SatImage 6435	isolet 7797	PenDigits 10992	Letter 20000	kropt 28056
ESLB: Gibbs-fitted-IW	0.367 ± 0.068	0.612 ± 0.075	0.510 ± 0.057	0.593 ± 0.070	0.349 ± 0.019	0.604 ± 0.066	0.425 ± 0.035	0.411 ± 0.023
ESLB: Gibbs-fitted-SN	0.997 ± 0.004	0.919 ± 0.028	0.999 ± 0.002	0.964 ± 0.035	1.000 ± 0	0.969 ± 0.017	0.984 ± 0.013	0.840 ± 0.053
ESLB: Ideal	0.903 ± 0.013	0.916 ± 0.007	0.907 ± 0.023	0.921 ± 0.024	0.918 ± 0.005	0.908 ± 0.021	0.912 ± 0.007	0.908 ± 0.010
λ-IW: Gibbs-fitted-IW	0.585 ± 0.041	0.725 ± 0.040	0.963 ± 0.023	0.700 ± 0.038	1.680 ± 0.068	0.772 ± 0.049	0.679 ± 0.027	0.400 ± 0.019
λ-IW: Gibbs-fitted-SN	0.175 ± 0.055	0.514 ± 0.103	0.192 ± 0.070	0.352 ± 0.100	0.381 ± 0.058	0.361 ± 0.101	0.297 ± 0.032	0.267 ± 0.019
λ-IW: Ideal	0.851 ± 0.013	0.848 ± 0.013	0.750 ± 0.033	0.691 ± 0.024	0.849 ± 0.014	0.772 ± 0.024	0.862 ± 0.016	0.820 ± 0.010
λ-DR: Ideal	0.877 ± 0.019	0.870 ± 0.013	0.820 ± 0.046	0.800 ± 0.038	0.875 ± 0.016	0.835 ± 0.029	0.883 ± 0.013	0.834 ± 0.016
λ-DR: Gibbs-fitted-IW	0.682 ± 0.098	0.770 ± 0.052	0.730 ± 0.082	0.710 ± 0.069	0.528 ± 0.125	0.734 ± 0.054	0.477 ± 0.064	0.530 ± 0.044
λ-DR: Gibbs-fitted-SN	0.683 ± 0.086	0.671 ± 0.102	0.583 ± 0.061	0.752 ± 0.069	0.722 ± 0.071	0.743 ± 0.073	0.718 ± 0.048	0.687 ± 0.039
Cheb-SN: Ideal	0.903 ± 0.013	0.916 ± 0.007	0.907 ± 0.023	0.921 ± 0.024	0.918 ± 0.005	0.908 ± 0.021	0.912 ± 0.007	0.908 ± 0.010
Cheb-SN: Gibbs-fitted-IW	0.367 ± 0.068	0.612 ± 0.075	0.510 ± 0.057	0.593 ± 0.070	0.349 ± 0.019	0.604 ± 0.066	0.425 ± 0.035	0.411 ± 0.023
Cheb-SN: Gibbs-fitted-SN	0.997 ± 0.004	0.919 ± 0.028	0.999 ± 0.002	0.964 ± 0.035	1.000 ± 0	0.969 ± 0.017	0.984 ± 0.013	0.840 ± 0.053
DR: Ideal	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-IW	-	-	-	-	-	-	-	-
DR: Gibbs-fitted-SN	-	-	-	-	-	-	-	-
Emp.Lik. Ideal	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-IW	-	-	-	-	-	-	-	-
Emp.Lik. Gibbs-fitted-SN	-	-	-	-	-	-	-	-