A Analysis of Existing Algorithms

Let $f^{\ast,\ast}$ denote a function that incorporates an attacker strategy. When $k = 0$, $f^{\text{CH, IS}}(D, w_y, g_y, k)$ is the result of applying the CH inequality to the IS weighted returns, obtained from $D$, which additionally includes $k$ copies of a trajectory with an IS weight of $w_y$ and return of $g_y$. Notice that $f^{\ast,\ast}$ is written in terms of IS weights. The following defines $f^{\text{CH, WIS}}$, written in terms of IS weights, when $k = 0$:

$$f^{\text{CH, WIS}}(D, w_y, g_y, 0) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

For the rest of the paper, we use the following notation. Let $I = \{I : \exists a \in A, \exists s \in S, I = \prod_{i=0}^{n-1} \pi_e(A_i = a, S_i = s)/\pi_b(A_i = a, S_i = s)\}$, i.e., the set of all IS weights that could be obtained from policies $\pi_e$ and $\pi_b$. The maximum and minimum IS weight is denoted by $\piE = \max(I)$ and $\piI = \min(I)$, respectively. For shorthand, let the sum of IS weights in $D$ be written as $\beta = \sum_{i=1}^{n} w_i$. Also, we assume that $\beta > 0$ to ensure that WIS is well-defined.

Next, we define a new term to describe how an attacker can increase the $1 - \delta$ confidence lower bound on the mean of a bounded and real-valued random variable. We say that $f^{\ast,\ast}$ is adversarially monotonic given its inputs, if an attacker can maximize $f^{\ast,\ast}$ by maximizing the value of the added samples. For brevity, we say that $f^{\ast,\ast}$ is adversarially monotonic.

**Definition 1.** $f^{\ast,\ast}$ is adversarially monotonic for $n > 1$, $k > 0$, $\pi_e$, $\pi_b$ and $D$ if both

1. There exists two constants $p \geq 0$ and $q \in [0, 1]$, with $pq \in [0, i^\ast]$, such that $f^{\ast,\ast}(D, p, q, k) \geq f^{\ast,\ast}(D, q, p, 0)$, i.e., adding $k$ copies of $pq$ does not decrease $f$;
2. $\frac{\partial}{\partial g_y} f^{\ast,\ast}(D, i^\ast, g_y, k) \geq 0$ and $\frac{\partial}{\partial w_y} f^{\ast,\ast}(D, w_y, 1, k) \geq 0$, with no local maximums, i.e., $f$ is a non-decreasing function w.r.t. the IS weight and return added by the attacker, respectively.

**Definition 1** means that $f^{\ast,\ast}$ is maximized when $w_y$ and $g_y$ is maximized. In other words, the optimal strategy is to add $k$ copies of the trajectory with the maximum IS weight and return. Notice that $f^{\ast,\ast}$ does not incorporate all possible attack functions, $M$: specifically, the set of attacks, where the attacker can choose to add $k$ different trajectories, is omitted. As described in Theorem 1 to perform a worst-case analysis, only the optimal attack must be incorporated as part of $f^{\ast,\ast}$.

In the following two lemmas, we show that a couple well-known Seldonian algorithms are adversarially monotonic.

**Lemma 1.** Under Assumptions 1, 2 and 3, $f^{\text{CH, IS}}$ is adversarially monotonic.

**Proof.** Let $w_y \geq \frac{1}{n} \sum_{i=1}^{n} w_i g_i + \frac{(n+k)}{k} \left( b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} - b \sqrt{\frac{\ln(1/\delta)}{2n}} \right)$ and $g_y = 1$. To show that $w_y g_y \in [0, i^\ast]$ as stated in [1] in Definition 1, it must be that $w_y \in [0, i^\ast]$. For all $i \in \{1, \ldots, n\}$, $w_i g_i \in [0, i^\ast]$. Thus, for any given dataset, $0 \leq 1/n \sum_{i=1}^{n} w_i g_i \leq i^\ast/n$. Using this fact, for any given $D$, the range of $w_y$ is

$$\frac{b(n+k)}{k} \left( \frac{\ln(1/\delta)}{2(n+k)} - \frac{\ln(1/\delta)}{2n} \right) \leq w_y \leq \frac{b(n+k)}{k} \left( \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} - \sqrt{\frac{\ln(1/\delta)}{2n}} \right).$$
Therefore, \( w_y \) can be selected such that \( w_y g_y \in [0, i^*] \). It follows that

\[
f_{\text{CH, WIS}}(D, w_y, 1, k) = \frac{1}{n + k} \sum_{i=1}^{n} w_i g_i + \frac{k}{n + k} (w_y)(1) - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}}
\]

\[
\geq \frac{1}{n + k} \sum_{i=1}^{n} w_i g_i + \frac{k}{n + k} \left( \frac{1}{n} \sum_{i=1}^{n} w_i g_i + \frac{n + k}{k} \left( b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} - b \sqrt{\frac{\ln(1/\delta)}{2n}} \right) \right) - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}
\]

\[= f_{\text{CH, WIS}}(D, w_y, g_y, 0).\]

Next, we show that (2) in Definition 1 holds.

\[
\frac{\partial}{\partial w_y} f_{\star, \star}(D, w_y, g_y, k) = \frac{\partial}{\partial w_y} \left( \sum_{i=1}^{n} w_i g_i \right) + \frac{k w_y g_y}{n + k} - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}}
\]

\[
= \frac{k}{n + k}
\]

\[
\frac{\partial}{\partial w_y} f_{\star, \star}(D, w_y, 1, k) = \frac{k}{n + k}
\]

\[
\frac{\partial}{\partial g_y} f_{\star, \star}(D, w_y, g_y, k) = \frac{\partial}{\partial g_y} \left( \sum_{i=1}^{n} w_i g_i \right) + \frac{k w_y g_y}{n + k} - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}}
\]

\[
= \frac{k w_y}{n + k}
\]

\[
\frac{\partial}{\partial g_y} f_{\star, \star}(D, i^*, g_y, k) = \frac{k i^*}{n + k}
\]

Notice that both partial derivatives are non-negative when \( g_y = 1 \) and \( w_y = i^* \), respectively. To find any critical points, the following equations are solved simultaneously: \( \partial/\partial g_y f_{\text{CH, WIS}}(D, w_y, g_y, k) = 0 \) and \( \partial/\partial w_y f_{\text{CH, WIS}}(D, w_y, g_y, k) = 0 \). Notice that points along the line \((w_y, 0)\) and \((0, g_y)\) are all critical points. The following partial derivatives are computed to classify these points:

\[
\frac{\partial}{\partial w_y} (D, w_y, g_y, k) = 0.
\]

\[
\frac{\partial}{\partial (g_y)^2} (D, w_y, g_y, k) = 0.
\]

\[
\frac{\partial}{\partial g_y w_y} (D, w_y, g_y, k) = \frac{k}{n + k}.
\]

Using the second partial derivative test, the critical points are substituted into the following equation:

\[
\frac{\partial}{\partial (w_y)^2} \cdot \frac{\partial}{\partial (g_y)^2} - \left( \frac{\partial}{\partial g_y w_y} \right)^2 = - \left( \frac{k}{n + k} \right)^2,
\]

which is less than zero. Therefore, points along the line \((w_y, 0)\) and \((0, g_y)\) are saddle points.

**Lemma 2.** Under Assumptions 1 and 2, \( f_{\text{CH, WIS}} \) is adversarially monotonic.
Proof. First, we show that (1) in Definition 1 holds with $g_y = 1$ and $w_y = 0$.

\[
\begin{align*}
\phi_{\text{CH, WIS}}(D, w_y, g_y, k) &= \frac{1}{k w_y + \beta} \left( k w_y g_y + \sum_{i=1}^{n} w_i g_i \right) - b \sqrt{\frac{\ln(1/\beta)}{2(n + k)}} \\
\phi_{\text{CH, WIS}}(D, 0, 1, k) &= \frac{1}{b} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\beta)}{2(n + k)}} \\
&= \phi_{\text{CH, WIS}}(D, w_y, g_y, 0),
\end{align*}
\]

where (1) follows from $b \sqrt{\frac{\ln(1/\beta)}{2n}} > b \sqrt{\frac{\ln(1/\beta)}{2(n + k)}}$. Second, we show that (2) in Definition 1 holds.

\[
\frac{\partial}{\partial w_y} \phi_{\text{CH, WIS}}(D, w_y, g_y, k) = -\frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i g_i - \frac{k^2 w_y g_y}{(k w_y + \beta)^2} + \frac{k g_y}{(k w_y + \beta)}
\]

\[
= -\frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i g_i - \frac{k^2 w_y g_y}{(k w_y + \beta)^2} + \frac{k g_y (k w_y + \beta)}{(k w_y + \beta)^2}
\]

\[
= -\frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i g_i + \frac{k g_y}{(k w_y + \beta)^2}
\]

\[
= -\frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i g_i + \frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i g_y
\]

\[
= \frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i (g_y - g_i) = \frac{k}{(k w_y + \beta)^2} \sum_{i=1}^{n} w_i (1 - g_i).
\]

(2)

Notice that (2) is non-negative: 1) When $g_y = 1$, (2) is positive as long as there exists at least one $g_i < 1$ for $i \in \{1, \ldots, n\}$; 2) If all $g_i = 1$ in $D$, then (2) is zero. The following is the derivative of $\phi_{\text{CH, WIS}}(D, w_y, g_y, k)$ w.r.t. $g_y$:

\[
\frac{\partial}{\partial g_y} \phi_{\text{CH, WIS}}(D, w_y, g_y, k) = \frac{k w_y}{(k w_y + \beta)}
\]

(3)

\[
\frac{\partial}{\partial g_y} \phi_{\text{CH, WIS}}(D, i^*, g_y, k) = \frac{k i^*}{(k i^* + \beta)}
\]

which is also non-negative. To find any critical points, the following equations are solved simultaneously: $\partial/\partial g_y \phi_{\text{CH, WIS}}(D, w_y, g_y, k) = 0$ and $\partial/\partial w_y \phi_{\text{CH, WIS}}(D, w_y, g_y, k) = 0$. Notice that (3) is zero when $w_y = 0$. Plugging $w_y = 0$ into $\partial/\partial g_y \phi_{\text{CH, WIS}}(D, w_y, g_y, k) = 0$, and then solving for $g_y$, yields the $x$ coordinate of a critical point.

\[
\frac{k}{(k + (0))^2} \sum_{i=1}^{n} w_i (g_y - g_i) = 0
\]

\[
\frac{k}{(k + (0))^2} \sum_{i=1}^{n} w_i (g_y - g_i) = 0
\]

\[
g_y \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i g_i = 0
\]

\[
g_y = \frac{\sum_{i=1}^{n} w_i g_i}{\beta}.
\]
The following partial derivatives are computed to classify whether \((0, \sum_{i=1}^{n} w_{i} r_{i}/\beta)\) is a minimum, maximum or saddle point:
\[
\frac{\partial}{\partial (w_y)} (D, w_y, g_y, k) = \frac{-2k^2}{(\beta + kw_y)^3} \sum_{i=1}^{n} w_i (g_y - g_i).
\]
\[
\frac{\partial}{\partial (g_y)} (D, w_y, g_y, k) = 0.
\]
\[
\frac{\partial}{\partial g_y w_y} (D, w_y, g_y, k) = \frac{\partial}{\partial w_y} \left( \frac{kw_y}{\beta + kw_y} \right) = \frac{k}{(\beta + kw_y)^2}.
\]
Using the second partial derivative test, the critical point is substituted into the following equation:
\[
\frac{\partial}{\partial (w_y)^2} \cdot \frac{\partial}{\partial (g_y)^2} - \left( \frac{\partial}{\partial g_y w_y} \right)^2 = 0 - \left( \frac{k\beta}{(\beta + k(0))^2} \right)^2
\]
which is less than zero. Therefore, \((w_y = 0, g_y = \sum_{i=1}^{n} w_{i} r_{i}/\beta)\) is a saddle point.

Next, we describe the trajectory that must be added to \(D\) to execute the optimal attack.

**Definition 2 (Optimal Attack).** An optimal attack strategy for \(k > 0\) is to select
\[
\arg \max_{H \in H_{\pi_{b}}} f^{*,*} (D, w_y = w(H, \pi_e, \pi_b), g_y = g(H), k).
\]

**Definition 3 (Optimal Trajectory).** Given that a maximum exists, let \((a', s') \in \arg \max_{a \in A, s \in S} \frac{\pi_{b}(a, s)}{\pi_{e}(a, s)}\). If \(\frac{\pi_{b}(a, s)}{\pi_{e}(a, s)} > 1\), let \(H^* = \{S_0 = s', A_0 = a', R_0 = 1, \ldots, S_{T-1} = s', A_{T-1} = a', R_{T-1} = 1\}\). Otherwise, let \(H^* = \{S_0 = s', A_0 = a', R_0 = 1\}\).

**Theorem 1.** For any adversarially monotonic off-policy estimator, the optimal attack strategy is to add \(k\) repetitions of \(H^*\) to \(D\).

**Proof.** An optimal attack strategy is equivalent to
\[
\arg \max_{H \in H_{\pi_{b}}} f^{*,*} (D, w(H, \pi_e, \pi_b), g(H), k) = \arg \max_{i^* \in I, g^* \in [0, 1]} f^{*,*} (D, i^*, g^*, k).
\]
For any off-policy estimator that is adversarially monotonic, by \([1]\) of Definition \([1]\) there exists a \(pq\) such that
\[
f^{*,*} (D, p, q, k) \geq f^{*,*} (D, p, q, 0).
\]
A return that maximizes \(f^{*,*} (D, w_y, g_y, k)\) implies that
\[
\max_{g^* \in [0, 1]} f^{*,*} (D, p, g^*, k) \geq f^{*,*} (D, p, q, k).
\]

\(f^{CH, IS}\) and \(f^{CH, WIS}\) are non-decreasing w.r.t. the return. Therefore,
\[
\arg \max_{g^* \in [0, 1]} f^{*,*} (D, p, g^*, k) = \max_{g^* \in [0, 1]} g^*.
\]
Setting \(g^* = 1\), an importance weight that maximizes \(f^{*,*} (D, w_y, 1, k)\) implies that
\[
\max_{i^* \in I} f^{*,*} (D, i^*, 1, k) \geq f^{*,*} (D, p, 1, k).
\]
\(f^{CH, IS}\) and \(f^{CH, WIS}\) are also non-decreasing w.r.t. the importance weight. So,
\[
\arg \max_{i^* \in I} f^{*,*} (D, i^*, 1, k) = \max_{i^* \in I} i^*.
\]
Since the IS weight is a product of ratios over the length of a trajectory, the ratio at a single time step is maximized.

\[
\max_{i^* \in I} = \max_{\alpha \in A, s \in S} \prod_{t=0}^{\tau-1} \frac{\pi_e(A_t = a, S_t = s)}{\pi_b(A_t = a, S_t = s)} = \begin{cases} 
\left( \max_{\alpha \in A, s \in S} \frac{\pi_e(a,s)}{\pi_b(a,s)} \right)^\tau & \text{if } \max_{\alpha \in A, s \in S} \frac{\pi_e(a,s)}{\pi_b(a,s)} > 1, \\
\max_{\alpha \in A, s \in S} \frac{\pi_e(a,s)}{\pi_b(a,s)} & \text{otherwise}.
\end{cases}
\]

To create \( H^* \), if the ratio at a single time step is greater than 1, \( \alpha' \) and \( s' \) is repeated for the maximum length of the trajectory, \( \tau \); otherwise, \( \alpha' \) and \( s' \) is repeated only for a single time step. Thus, \( H^* \) represents the trajectory with the largest return and importance weight.

Next, we show how Equations 2 and 1, that define quasi-\( \alpha \)-security and \( \alpha \)-security, respectively, apply to \( L^* \). Specifically, we show that a safety test using \( L^* \) as a metric is a valid safety test that first predicts the performance of \( \pi_e \) using \( D \), and then bounds the predicted performance with high probability. If \( L^* (\pi_e, D) > J(b) \), the safety test returns True; otherwise it returns False.

**Lemma 3.** A safety test using \( L^* \) is quasi-\( \alpha \)-secure if \( \forall m \in M \), \( \Pr \left( L^* (\pi_e, m(D, k)) > J(b) + \alpha \right) \leq \Pr \left( L^* (\pi_e, D) > J(b) \right) \).

**Proof.** For \( x \in \mathbb{N}^+ \), let \( P : \Pi \times D_n^{+} \to \mathbb{R}^x \) denote any function to predict the performance of some \( \pi_e \in \Pi \), using data \( D \) collected from \( \pi_b \). Also, let \( \mathcal{B} : \mathbb{R}^x \times [0, 1] \to \mathbb{R} \) denote any function that bounds performance with high probability, \( 1 - \delta \), where \( \delta \in [0, 1] \). Starting with the definition of quasi-\( \alpha \)-security, we have that \( \forall m \in M \),

\[
\Pr \left( \varphi(\pi_e, m(D, k)), J(b) + \alpha \right) = \text{True} \leq \Pr \left( \varphi(\pi_e, D), J(b) \right) = \text{True}
\]

\[
\iff \Pr \left( \mathcal{B} \left( P(\pi_e, m(D, k)), \delta \right) > J(b) + \alpha \right) \leq \Pr \left( \mathcal{B} \left( P(\pi_e, D), \delta \right) > J(b) \right)
\]

\[
\iff \Pr \left( L^* (\pi_e, m(D, k)) > J(b) + \alpha \right) \leq \Pr \left( L^* (\pi_e, D) > J(b) \right).
\]

**Lemma 4.** A safety test using \( L^* \) is \( \alpha \)-secure if \( \forall m \in M \), \( \Pr \left( L^* (\pi_e, m(D, k)) > J(b) + \alpha \right) < \delta \).

**Proof.** For \( x \in \mathbb{N}^+ \), let \( P : \Pi \times D_n^{+} \to \mathbb{R}^x \) denote any function to predict the performance of some \( \pi_e \in \Pi \), using data \( D \) collected from \( \pi_b \). Also, let \( \mathcal{B} : \mathbb{R}^x \times [0, 1] \to \mathbb{R} \) denote any function that bounds performance with high probability, \( 1 - \delta \), where \( \delta \in [0, 1] \). Starting with the definition of \( \alpha \)-security, we have that \( \forall m \in M \),

\[
\Pr \left( \varphi(\pi_e, m(D, k)), J(b) + \alpha \right) = \text{True} < \delta
\]

\[
\iff \Pr \left( \mathcal{B} \left( P(\pi_e, m(D, k)), \delta \right) > J(b) + \alpha \right) < \delta
\]

\[
\iff \Pr \left( L^* (\pi_e, m(D, k)) > J(b) + \alpha \right) < \delta.
\]

In Lemma 4, we describe a condition that must hold in order to compute a valid \( \alpha \). The condition states that a valid \( \alpha \) must be equal to or greater than the largest increase in the 1 − \( \delta \) confidence lower bound on \( J(\pi_e) \) across all datasets \( D \in D_n^{+} \) and all attack strategies (i.e., the optimal attack).

**Lemma 5.** A safety test using \( L^* \) is quasi-\( \alpha \)-secure or \( \alpha \)-secure if \( \forall D \in D_n^{+} \) and \( \forall m \in M \), \( L^* (\pi_e, m(D, k)) \leq L^* (\pi_e, D) + \alpha \).
Proof. If \( L^{*,*}(\pi_e, m(D, k)) \leq L^{*,*}(\pi_e, D) + \alpha \), then
\[
L^{*,*}(\pi_e, D) \geq L^{*,*}(\pi_e, m(D, k)) - \alpha.
\] (4)
A safety test checks whether \( L^{*,*}(\pi_e, D) > J(\pi_b) \). When (4) holds \( \forall D \in D_n^\alpha \) and \( \forall m \in M \),
\[
\Pr(L^{*,*}(\pi_e, D) > J(\pi_b)) \geq \Pr(L^{*,*}(\pi_e, m(D, k)) - \alpha > J(\pi_b)),
\] (5)
and hence via algebra that
\[
\Pr(L^{*,*}(\pi_e, m(D, k)) > J(\pi_b) + \alpha) \leq \Pr(L^{*,*}(\pi_e, D) > J(\pi_b)),
\]
which, by Lemma (4), implies that a safety test using \( L^{*,*} \) is quasi-\( \alpha \)-secure. In the case of \( \alpha \)-security, by Assumption (3) we require a “safe” safety test. That is,
\[
\Pr(L^{*,*}(\pi_e, D) > J(\pi_b)) < \delta.
\] (6)
From the transitive property of \( \geq \), we can conclude from (5) and (6) that
\[
\Pr(L^{*,*}(\pi_e, m(D, k)) - \alpha > J(\pi_b)) < \delta,
\]
and hence via algebra that
\[
\Pr(L^{*,*}(\pi_e, m(D, k)) > J(\pi_b) + \alpha) < \delta,
\]
which, by Lemma (4), implies that a safety test using \( L^{*,*} \) is \( \alpha \)-secure.

B Proof of Theorem 1

The result of (5) for the estimator that uses CH and IS is the following:
\[
\alpha' = \max_{D \in D_n^\alpha} f^{CH,IS}(D, i^*, 1, k) - L^{CH,IS}(\pi_e, D)
\]
\[
= \max_{D \in D_n^\alpha} \frac{1}{n+k} \sum_{i=1}^{n} w_i g_i + \frac{k}{n+k} (i^*)(1) - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} - \left( \frac{1}{n} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}} \right)
\]
\[
= \max_{D \in D_n^\alpha} b \sqrt{\frac{\ln(1/\delta)}{2n}} - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)} (i^* - \frac{\sum_{i=1}^{n} w_i g_i}{n}).
\]
Recall that \( b \) represents the upper bound of all IS weighted returns. Let \( b = i^* \), and \( g_i = 0 \) for all \( i \in \{1, \ldots, n\} \).
\[
\alpha' = i^* \sqrt{\frac{\ln(1/\delta)}{2n}} - i^* \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)} (i^* - 0)
\]
\[
= i^* \left( \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)} \right).
\]
The result of (5) for the estimator that uses CH and WIS is the following:
\[
\alpha' = \max_{D \in D_n^\alpha} f^{CH,WIS}(D, i^*, 1, k) - L^{CH,WIS}(\pi_e, D)
\]
\[
= \max_{D \in D_n^\alpha} \frac{1}{k i^* + \sum_{i=1}^{n} w_i} \left( \sum_{i=1}^{n} w_i g_i + k(i^*)(1) \right) - \frac{1}{k i^* + \sum_{i=1}^{n} w_i} \left( \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}} \right)
\]
\[
= \max_{D \in D_n^\alpha} b \sqrt{\frac{\ln(1/\delta)}{2n}} - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k i^*}{(k i^* + \beta)} \left( 1 - \frac{\sum_{i=1}^{n} w_i g_i}{\beta} \right).
\]
Let \( g_i = 0 \) for all \( i \in \{1, \ldots, n\} \). Also, notice that \( b = 1 \) because importance weighted returns are in range \([0, 1]\) for WIS (since only IS weights are clipped).
\[
\alpha' = \max_{D \in D_n^\alpha} \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k i^*}{(k i^* + \beta)}.
\]
Recall that \( \beta \neq 0 \). So, let \( w_i = 0 \) for all \( i \in \{1, \ldots, n-1\} \) and \( w_n = i^\min \).
\[
\alpha' = \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k i^*}{(i^\min + k i^*)}.
\]
C Panacea: An Algorithm for Safe and Secure Policy Improvement

Table 1: $\alpha$-security of Panacea.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH, IS</td>
<td>$c\left(\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)}\right)$</td>
</tr>
<tr>
<td>CH, WIS</td>
<td>$\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{kc}{(n+k)^2}$</td>
</tr>
</tbody>
</table>

Algorithm 1 Panacea($D, \pi_c, \alpha, k$)

1: Compute $c$, using $\alpha$ and $k$, given estimator
2: for $H \in D$ do
3: if IS weight computed using $H$ is greater than $c$ then
4: Set IS weight to $c$
5: return clipped $D$

C.1 Proof of Corollary 1

Let $\alpha'$ and $k'$ denote the user-specified inputs to Panacea. Based on Table 1, $c_{\text{CH, IS}} = \alpha'/(\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)})$ if $k' = k$. Recall that $b$ is the upper bound on all IS weighted returns. Due to clipping, $b = c_{\text{CH, IS}}$, and let $g_i = 0$ for all $i \in \{1, \ldots, n\}$. The result of (6) for the estimator that uses CH and IS is the following:

$$
\max_{D \in \mathcal{D}_n^k} f_{\text{CH, IS}}(\text{Panacea}(D, c_{\text{CH, IS}}, c_{\text{CH, IS}}, 1, k)) - L_{\text{CH, IS}}(\pi_c, \text{Panacea}(D, c_{\text{CH, IS}}))
$$

$$
= \max_{D \in \mathcal{D}_n^k} \left(\frac{1}{n+k} \sum_{i=1}^{n} w_i g_i + \frac{k}{n+k} (c_{\text{CH, IS}}) - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} - \frac{1}{n} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}\right)
$$

$$
= \max_{D \in \mathcal{D}_n^k} \left(b \sqrt{\frac{\ln(1/\beta)}{2n}} - b \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)} (c_{\text{CH, IS}} - \frac{1}{n} \sum_{i=1}^{n} w_i g_i)\right)
$$

$$
= c_{\text{CH, IS}} \left(\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)} (c_{\text{CH, IS}} - 0)\right)
$$

$$
= c_{\text{CH, IS}} \left(\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)}\right)
$$

$$
= \alpha' \left(\sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n+k)}} + \frac{k}{(n+k)}\right)
$$

$$
= \alpha'.
$$

For WIS, recall that no matter how the clipping weight is set, $b \leq 1$ because importance weighted returns are in range $[0, 1]$, and $\beta \neq 0$. So, let $w_i = 0$ for all $i \in \{1, \ldots, n-1\}$ and $w_n = i_{\text{min}}$. Also, let $g_i = 0$ for all $i \in \{1, \ldots, n\}$. Based on Table 1, $c_{\text{CH, WIS}} = i_{\text{min}} (\alpha' - \sqrt{\frac{\ln(1/\beta)}{2n}} + \sqrt{\frac{\ln(1/\beta)}{2(n+k)}})/k (1 - \alpha' + \sqrt{\frac{\ln(1/\beta)}{2n}} - \sqrt{\frac{\ln(1/\beta)}{2(n+k)}})$ if $k' = k$. The result of (6) for the estimator that uses
CH and WIS is the following:
\[
\max_{D \in \mathcal{D}_n} f_{\text{CH, WIS}}(\text{Panacea}(D, c_{\text{CH, WIS}}), c_{\text{CH, WIS}}, 1, k) - L_{\text{CH, WIS}}(\pi_c, \text{Panacea}(D, c_{\text{CH, WIS}}))
\]
\[
= \max_{D \in \mathcal{D}_n} \frac{1}{kc_{\text{CH, WIS}} + \sum_{i=1}^{n} w_i} \left( \sum_{i=1}^{n} w_i g_i + k(c_{\text{CH, WIS}}) \right) - b \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} - \left( \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i g_i - b \sqrt{\frac{\ln(1/\delta)}{2n}} \right)
\]
\[
= \max_{D \in \mathcal{D}_n} b \sqrt{\frac{\ln(1/\delta)}{2n}} - b \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{kc_{\text{CH, WIS}}}{(kc_{\text{CH, WIS}} + \beta)} \left( 1 - \sum_{i=1}^{n} w_i g_i \right)
\]
\[
\leq \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{kc_{\text{CH, WIS}}}{(kc_{\text{CH, WIS}} + \delta)}
\]
\[
= \left( \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} \right) + \left( \alpha - \sqrt{\frac{\ln(1/\delta)}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} \right)
\]
\[
= \alpha - \sqrt{\frac{\ln(1/\delta)}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}}
\]
\[= \alpha'.
\]

C.2 Proof of Corollary 2

Let \( \alpha \) and \( k' \) denote the user-specified inputs to Panacea. If \( k' = k \), i.e., the user inputs the correct number of trajectories added by the attacker, the result of (6) for the estimator that uses CH and IS is the following:
\[
\alpha = \max_{D \in \mathcal{D}_n} f^{\text{CH, IS}}(\text{Panacea}(D, c, 1, k)) - L^{\text{CH, IS}}(\pi_c, \text{Panacea}(D, c))
\]
\[
\alpha = c \left( \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{k}{n + k} \right)
\]
\[
c = \frac{\beta}{\left( \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{k}{n + k} \right)}.
\]

If \( k' = k \), the result of (6) for the estimator that uses CH and WIS is the following:
\[
\max_{D \in \mathcal{D}_n} f^{\text{CH, WIS}}(\text{Panacea}(D, c, 1, k)) - L^{\text{CH, WIS}}(\pi_c, \text{Panacea}(D, c)) \leq \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{kc}{(kc + \delta)}.
\]

Setting the right-hand side of (7) to \( \alpha \), and solving for \( c \) equals:
\[
\alpha = \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \frac{kc}{(\delta + kc)}
\]
\[
k \alpha = \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}}
\]
\[
k - k \alpha = \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}}
\]
\[
k (1 - \alpha) = \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}}
\]
\[
c = \frac{\min(\alpha - \sqrt{\frac{\ln(1/\delta)}{2n}} + \sqrt{\frac{\ln(1/\delta)}{2(n + k)}})}{k (1 - \alpha + \sqrt{\frac{\ln(1/\delta)}{2n}} - \sqrt{\frac{\ln(1/\delta)}{2(n + k)}})}.
\]