High Confidence Off-Policy Evaluation (HCOPE)

Philip Thomas
CMU 15-899E, Real Life Reinforcement Learning, Fall 2015
Sequential Decision Problems

Agent

Environment

State, $s$
(Discrete / Continuous)
(Fully observable / Partially observable)

Action, $a$
(Discrete / Continuous)
Example: Digital Marketing
Example: Educational Games
Example: Decision Support Systems
Example: Gridworld

Grid:

- Initial state: (1,1)
- Reward $R_t = -10$ at (2,2)
- Reward $R_t = 1$ at (2,4)
- Terminal state: (4,4) with $R_t = 10$
Example: Mountain Car
Reinforcement Learning Algorithms

- Sarsa
- Q-learning
- LSPI
- Fitted Q Iteration
- REINFORCE
- Residual Gradient
- Continuous-Time Actor-Critic
- Value Gradient
- POWER
- PILCO
- LSPI
- PIPI
- Policy Gradient
- DQN
- Double Q-Learning
- Deterministic Policy Gradient
- NAC-LSTD
- INAC
- Average-Reward INAC
- Unbiased NAC
- Projected NAC
- Risk-sensitive policy gradient
- Natural Sarsa
- PGPE / PGPE-SyS
- True Online
- GTD/TDC
- ARP
- GPTD
- Auto-Actor Auto-Critic
- Approximate Value Iteration
If you apply an existing method, do you have confidence that it will work?
Notation

• $s$: State
• $a$: Action
• $S_t, A_t$: State, and action at time $t$
• $\pi(a|s) = \Pr(A_t = a|S_t = s)$
• $\tau = (S_0, A_0, S_1, ..., S_L, A_L)$
• $G(\tau) \in [0,1]$  
• $\rho(\pi) = \mathbb{E}[G(\tau)|\tau \sim \pi]$
Two Goals:

• High confidence off-policy evaluation (HCOPE)

  Historical Data, $\mathcal{D}$
  Proposed Policy, $\pi_e$
  Confidence Level, $\delta$

  $1 - \delta$ confidence lower bound on $\rho(\pi_e)$

• Safe Policy Improvement (SPI)

  Historical Data, $\mathcal{D}$
  Performance baseline, $\rho_-$
  Confidence Level, $\delta$

  An improved* policy, $\pi$

*The probability that $\pi$’s performance is below $\rho_-$ is at most $\delta$
High Confidence Off-Policy Evaluation

- Historical data: $\mathcal{D} = \{(\tau_i, \pi_i) : \tau_i \sim \pi_i\}_{i=1}^n$
- Evaluation policy, $\pi_e$
- Confidence level, $\delta$
- Compute $\text{HCOPE}(\pi_e | \mathcal{D}, \delta)$ such that

$$\Pr(\rho(\pi_e) \geq \text{HCOPE}(\pi_e | \mathcal{D}, \delta)) \geq 1 - \delta$$
Importance Sampling

• We would like to estimate

\[ \theta := \mathbb{E}[f(x)|x \sim p] \]

• Monte Carlo estimator:
  • Sample \( X_1, \ldots, X_n \) from \( p \) and set:

\[ \hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} f(X_i) \]

• Nice properties
  • The Monte Carlo estimator is strongly consistent:

\[ \hat{\theta}_n \xrightarrow{a.s.} \theta \]

  • The Monte Carlo estimator is unbiased for all \( n \geq 1 \):

\[ \mathbb{E}[\hat{\theta}_n] = \theta \]
Importance Sampling

• We would like to estimate
  \[ \theta := \mathbb{E}[f(x) | x \sim p] \]
• ... but we can only sample from a distribution, \( q \), not \( p \).
• Assume: if \( q(x) = 0 \) then \( f(x)p(x) = 0 \). Then:
Importance Sampling

• We would like to estimate

\[ \theta := \mathbb{E}[f(x) | x \sim p] \]

• Importance sampling estimator:
  • Sample \( X_1, \ldots, X_n \) from \( q \) and set:
  \[
  \hat{\theta}_n := \frac{1}{n} \sum_{i=1}^{n} \frac{p(X_i)}{q(X_i)} f(X_i)
  \]

• Nice properties (under mild assumptions)
  • The importance sampling estimator is strongly consistent:
    \[
    \hat{\theta}_n \xrightarrow{a.s.} \theta
    \]
  • The importance sampling estimator is unbiased for all \( n \geq 1 \):
    \[
    \mathbb{E} \left[ \hat{\theta}_n \right] = \theta
    \]
Importance Sampling

$$\rho(\pi_e) = \mathbb{E}_{\tau \sim \pi_e}[G(\tau)] = \mathbb{E}_{\tau \sim \pi_b}\left[\frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)}G(\tau)\right]$$

- **Evaluation Policy,** $\pi_e$
- **Behavior Policy,** $\pi_b$

Probability of trajectory
Importance Sampling for Reinforcement Learning  
(D. Precup, R. S. Sutton, and S. Dasgupta, 2001)

\[
\rho(\pi_e) = \mathbb{E}_{\tau \sim \pi_e} [G(\tau)] = \mathbb{E}_{\tau \sim \pi_b} \left[ \frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)} G(\tau) \right]
\]

\[
\frac{\Pr(\tau|\pi_e)}{\Pr(\tau|\pi_b)} G(\tau) = \frac{\prod_{t=0}^{L} \Pr(S_t|\text{past}) \Pr(A_t|\text{past,}\pi_e)}{\prod_{t=0}^{L} \Pr(S_t|\text{past}) \Pr(A_t|\text{past,}\pi_b)} G(\tau)
\]

\[
= \frac{\prod_{t=0}^{L} \Pr(A_t|\text{past,}\pi_e)}{\prod_{t=0}^{L} \Pr(A_t|\text{past,}\pi_b)} G(\tau)
\]

\[
= \frac{\prod_{t=0}^{L} \pi_e(A_t|S_t)}{\prod_{t=0}^{L} \pi_b(A_t|S_t)} G(\tau)
\]

\[
\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)
\]
Per-Decision Importance Sampling

• Use importance sampling to estimate each $R_t$.
  • Still and unbiased and strongly consistent estimator of $\rho(\pi_e)$.
  • Often has lower variance than ordinary importance sampling.
Historical Data, $\mathcal{D}$
Proposed Policy, $\pi_e$
Confidence Level, $\delta$

$1 - \delta$ confidence lower bound on $\rho(\pi_e)$

$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t | S_t)}{\pi_b(A_t | S_t)} G(\tau)$
Chernoff-Hoeffding Inequality

- Let $X_1, ..., X_n$ be $n$ independent identically distributed random variables such that:
  - $X_i \in [0, b]$
- Then with probability at least $1 - \delta$:

$$E[X_i] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}}$$
\[ \rho(\pi_e) = E[\hat{\rho}(\pi_e, \tau, \pi)] \geq \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}(\pi_e, \tau, \pi_i) - b \sqrt{\frac{\ln(1/\delta)}{2n}} \]

With probability at least \(1 - \delta\):

\[ E[X_i] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\ln(1/\delta)}{2n}} \]
Historical Data, $\mathcal{D}$

Proposed Policy, $\pi_e$

Confidence Level, $\delta$

$1 - \delta$ confidence lower bound on $\rho(\pi_e)$

$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t | S_t)}{\pi_b(A_t | S_t)} G(\tau)$
Historical Data, $\mathcal{D}$

Proposed Policy, $\pi_e$

Confidence Level, $\delta$

\[ \left\{ \hat{\rho}(\pi_e, \tau_i, \pi_i) \right\}_{i=1}^{n} \]

\[ \hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t | S_t)}{\pi_b(A_t | S_t)} G(\tau) \]

\[ \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}(\pi_e, \tau_i, \pi_i) - b \sqrt{\frac{\ln(1/\delta)}{2n}} \]

\[ 1 - \delta \text{ confidence lower bound on } \rho(\pi_e) \]
Example: Mountain Car

Figure 3: Mountain Car (Sarsa(\(\lambda\)))
Example: Mountain Car

- Using 100,000 trajectories
- Evaluation policy’s true performance is 0.19 ∈ [0,1].
- We get a 95% confidence lower bound of:

$$-5,831,000$$
What went wrong?

\[ \hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t | S_t)}{\pi_b(A_t | S_t)} G(\tau) \]
What went wrong?

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

\[ b \approx 10^{9.4} \]

Largest observed importance weighted return: 316.
Another problem:

\[ \hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t | S_t)}{\pi_b(A_t | S_t)} G(\tau) \]

- One behavior policy
  - Independent and identically distributed
- More than one behavior policy
  - Independent
Conservative Policy Iteration

(S. Kakade and J. Langford, 2002)

- \( \approx 1,000,000 \) trajectories for a single policy improvement.
PAC-RL (T. Lattimore and M. Hutter, 2012)

- $\approx 10^{17}$ time steps to guarantee convergence to a near-optimal policy.
Thesis

High Confidence Off-Policy Evaluation (HCOPE) and Safe Policy Improvement (SPI) are tractable using a practical amount of data.
Historical Data, $\mathcal{D}$

Proposed Policy, $\pi_e$

Confidence Level, $\delta$

$1 - \delta$ confidence lower bound on $\rho(\pi_e)$

$\hat{\rho}(\pi_e, \tau, \pi_b) = \prod_{t=0}^{L} \frac{\pi_e(A_t|S_t)}{\pi_b(A_t|S_t)} G(\tau)$

$\frac{1}{n} \sum_{i=1}^{n} \hat{\rho}(\pi_e, \tau_i, \pi_i) - b \sqrt{\frac{\ln(1/\delta)}{2n}}$

$1 - \delta$ confidence lower bound on $\rho(\pi_e)$
<table>
<thead>
<tr>
<th>Name</th>
<th>Direct Dependence on $b$</th>
<th>Identically Distributed Only</th>
<th>Exact or Approximate</th>
<th>Reference</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH</td>
<td>$\Theta\left(\frac{b}{\sqrt{n}}\right)$</td>
<td>No</td>
<td>Exact</td>
<td>(Massart, 2007)</td>
<td>None</td>
</tr>
<tr>
<td>MPeB</td>
<td>$\Theta\left(\frac{b}{n}\right)$</td>
<td>No</td>
<td>Exact</td>
<td>(Maurer and Pontil, 2009, Theorem 11)</td>
<td>Requires all random variables to have the same range.</td>
</tr>
<tr>
<td>BM</td>
<td>$\Theta\left(\frac{b}{\sqrt{n}}\right)$</td>
<td>Yes</td>
<td>Exact</td>
<td>(Bubeck et al., 2012)</td>
<td>None.</td>
</tr>
</tbody>
</table>

| CUT  | None                     | No                            | Exact               | Theorem 23 | None. |
Theorem 17 (Chernoff-Hoeffding (CH) Inequality). Let \( \{X_i\}_{i=1}^n \) be \( n \) independent random variables such that \( \Pr(X_i \in [a_i, b_i]) = 1 \), for all \( i \in \{1, \ldots, n\} \), where all \( a_i \in \mathbb{R} \) and \( b_i \in \mathbb{R} \). Then

\[
\Pr \left( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \sqrt{\frac{\ln \left( \frac{1}{\delta} \right) \sum_{i=1}^{n} (b_i - a_i)^2}{2n^2}} \right) \geq 1 - \delta. \tag{4.6}
\]
Theorem 18 (Maurer and Pontil’s Empirical Bernstein (MPeB) Inequality). Let \( \{X_i\}_{i=1}^n \) be \( n \) independent random variables such that \( \Pr(X_i \in [a, b]) = 1 \), for all \( i \in \{1, \ldots, n\} \), where \( a \in \mathbb{R} \) and \( b \in \mathbb{R} \). Then

\[
\Pr \left( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7(b-a) \ln \left( \frac{2}{\delta} \right)}{3(n-1)} - \sqrt{\frac{2 \ln \left( \frac{2}{\delta} \right)}{n} \frac{1}{n(n-1)} \sum_{i,j=1}^{n} \frac{(X_i - X_j)^2}{2}} \right) \geq 1 - \delta.
\]
Theorem 19 (Anderson and Massart’s (AM) Inequality). Let $\{X_i\}_{i=1}^n$ be $n$ independent and identically distributed random variables such that $X_i \geq a$, for all $i \in \{1, \ldots, n\}$, where $a \in \mathbb{R}$. Then

$$\Pr \left( \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] \geq Z_n - \sum_{i=0}^{n-1} (Z_{i+1} - Z_i) \min \left\{ 1, \frac{i}{n} + \sqrt{\frac{\ln(2/\delta)}{2n}} \right\} \right) \geq 1 - \delta,$$

where $Z_0 = a$ and $\{Z_i\}_{i=1}^n$ are $\{X_i\}_{i=1}^n$, sorted such that $Z_1 \leq Z_2 \leq \ldots \leq Z_n$. 
Extending Maurer’s Inequality

• First Key Idea:
  • Generalize: random variables with different ranges.
  • Specialize: random variables with the same mean.
Extending Maurer’s Inequality

• Second Key Idea:
  • Removing the upper tail only decreases the expected value.
Theorem 1. Let $X_1, \ldots, X_n$ be $n$ independent real-valued random variables such that for each $i \in \{1, \ldots, n\}$, we have $\Pr[0 \leq X_i] = 1$, $\mathbb{E}[X_i] \leq \mu$, and some threshold value $c_i > 0$. Let $\delta > 0$ and $Y_i := \min\{X_i, c_i\}$. Then with probability at least $1 - \delta$, we have

$$\mu \geq \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2}.$$ (3)
Tradeoff
Threshold Optimization

• Use 20% of the data to optimize $c$.
• Use 80% to compute lower bound with optimized $c$. 
Given \( n \) samples, \( \mathcal{X} := \{X_i\}_{i=1}^n \), predict what the lower bound would be if computed from \( m \) samples, rather than \( n \).

\[
\widehat{\text{CUT}}(\mathcal{X}, \delta, c, m) := \left[ \frac{1}{n} \sum_{i=1}^{n} \min\{X_i, c\} - \frac{7c \ln(2/\delta)}{3(m - 1)} \right] - \left[ \frac{\ln(2/\delta)}{m} \frac{2}{n(n - 1)} \left( n \sum_{i=1}^{n} (\min\{X_i, c\})^2 - \left( \sum_{i=1}^{n} \min\{X_i, c\} \right)^2 \right) \right],
\]

sample mean of \( \mathcal{X} \) (after being collapsed)

sample variance of \( \mathcal{X} \) (after being collapsed)
Algorithm 4.11: CUT($X_1, \ldots, X_n, \delta$): Uses the CUT inequality to return a $1 - \delta$ confidence lower bound on $\mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} X_i]$.

**Constants:** This algorithm has a real-valued hyperparameter, $c_{\text{min}} \geq 0$, which is the smallest allowed threshold. It should be chosen based on the application. For HCOPE we use $c_{\text{min}} = 1$.

** Assumes:** The $X_i$ are independent random variables such that $\Pr(X_i \geq 0) = 1$ for all $i \in \{1, \ldots, n\}$.

1. Randomly select $1/5$ of the $X_i$ and place them in a set $\mathcal{X}_{\text{pre}}$ and the remainder in $\mathcal{X}_{\text{post}}$;
   
   // Optimize threshold using $\mathcal{X}_{\text{pre}}$

2. $c^* \in \arg \max_{c \in [1, \infty]} \text{CUT}(\mathcal{X}_{\text{pre}}, \delta, c, |\mathcal{X}_{\text{post}}|)$; // \text{CUT} is defined in (4.15)

3. $c^* = \max\{c_{\text{min}}, c^*\}$; // Do not let $c^*$ become too small
   
   // Compute lower bound using optimized threshold, $c^*$ and $\mathcal{X}_{\text{post}}$

4. return $\text{CUT}(\mathcal{X}_{\text{post}}, \delta, c^*, |\mathcal{X}_{\text{post}}|)$;
<table>
<thead>
<tr>
<th></th>
<th>CUT</th>
<th>Chernoff-Hoeffding</th>
<th>Maurer</th>
<th>Anderson</th>
<th>Bubeck et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>95% Confidence lower</td>
<td>0.145</td>
<td>-5,831,000</td>
<td>-129,703</td>
<td>0.055</td>
<td>-.046</td>
</tr>
<tr>
<td>bound on the mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Digital Marketing Example

• 10 real-valued features
• Two groups of campaigns to choose between
• User interactions limited to $L = 10$
• Data collected from a Fortune 20 company
• Data was not used directly.
Example: Digital Marketing
Example: Digital Marketing
Example: Digital Marketing
Example: Digital Marketing
Example: Digital Marketing
Example: Digital Marketing
We can now evaluate policies proposed by RL algorithms
Two Goals:

• High confidence off-policy evaluation (HCOPE)
  - Historical Data, $\mathcal{D}$
  - Proposed Policy, $\pi_e$
  - Confidence Level, $\delta$
  - $1 - \delta$ confidence lower bound on $\rho(\pi_e)$

• Safe Policy Improvement (SPI)
  - Historical Data, $\mathcal{D}$
  - Performance baseline, $\rho_-$
  - Confidence Level, $\delta$
  - An improved* policy, $\pi$

*The probability that $\pi$’s performance is below $\rho_-$ is at most $\delta$
Safe Policy Improvement (SPI)

$$\text{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\text{No Solution Found}\} \cup \Pi$$

$$\Pr (\text{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\pi: \rho(\pi) < \rho_-\}) < \delta$$

<table>
<thead>
<tr>
<th>Exact</th>
<th>Approximate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; \delta$</td>
<td>$\approx \delta$</td>
</tr>
</tbody>
</table>
SPI(\mathcal{D}, \rho_-, \delta)

1. Return **NO SOLUTION FOUND.**
SPI(\(\mathcal{D}, \rho_-, \delta\))
1. Return NO SOLUTION FOUND.

\[
\Pr \left( \text{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\pi : \rho(\pi) < \rho_-\} \right) < \delta
\]
$\text{SPI}(\mathcal{D}, \rho_-, \delta)$

1. Return NO SOLUTION FOUND.

\[
\Pr (\text{SPI}(\mathcal{D}, \rho_-, \delta) \in \{\pi: \rho(\pi) < \rho_\text{\textdollar}\}) < \delta
\]

• Want a batch RL algorithm that
  • Satisfies this inequality
  • Often returns a policy
Thoughts?
Historical Data

- Training Set (20%)
- Testing Set (80%)

Candidate Policy

“Safety” Test
Algorithm 5.3: \( \text{SPI}^{\dagger,\ast}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, \delta, \rho_-) \): Use the historical data, partitioned into \( \mathcal{D}_{\text{train}} \) and \( \mathcal{D}_{\text{test}} \), to search for a safe policy (with \( 1 - \delta \) confidence lower bound at least \( \rho_- \)). If none is found, then return \text{NO SOLUTION FOUND}. Although other \( \dagger \) and \( \ddagger \) could be used, we have only provided complete pseudocode for \( (\dagger, \ddagger) \in \{(\text{NPDIS, CUT}), (\text{CWPDIS, BCa})\} \). We allow for \( \ast \in \{\text{None, } k\text{-fold}\} \). Assumption 1 is not required.

\begin{align*}
1 & \pi_c \leftarrow \text{GETCANDIDATEPOLICY}^{\dagger,\ast}(\mathcal{D}_{\text{train}}, \delta, \rho_-, |\mathcal{D}_{\text{test}}|) ; \\
2 & \text{if } HCOPE^{\dagger}_{\pi_c}(\mathcal{D}_{\text{test}}, \delta) \geq \rho_- \text{ then} \\
3 & \quad \text{return } \pi_c ; \\
4 & \text{return } \text{NO SOLUTION FOUND}
\end{align*}
\[ f_\dagger^\dagger (\pi, \mathcal{D}, \delta, \rho_-, m) := \begin{cases} 
\hat{\rho} (\pi|\mathcal{D}) & \text{if } \text{HCOPE}_\dagger^\dagger (\pi, \mathcal{D}, \delta, m) \geq \rho_-, \\
\text{HCOPE}_\dagger^\dagger (\pi, \mathcal{D}, \delta, m) & \text{otherwise}. 
\end{cases} \]

**Algorithm 5.4:** \text{GETCANDIDATEPOLICY}_\dagger^\dagger,\text{None} (\mathcal{D}_\text{train}, \delta, \rho_-, m): Use the historical data, partitioned into \( \mathcal{D}_\text{train} \) to search for the candidate policy that is predicted to be safe and perform the best (or to be closest to safe if none are predicted to be safe). Although other \( \dagger \) and \( \dagger \) could be used, we have only provided complete pseudocode for \( (\dagger, \dagger) \in \{(\text{NPDIS, CUT}), (\text{CWPDIS, BCa})\} \). Assumption 1 is not required.

1. \text{return} \( \arg \max_\pi f_\dagger^\dagger (\pi, \mathcal{D}_\text{train}, \delta, \rho_-, m) \);
Algorithm 5.5: GET\textsc{CandidatePolicy}^{†,k-fold}(D_{\text{train}}, \delta, \rho_-, m): Use the historical data, $D_{\text{train}}$, to search for the candidate policy that is predicted to be safe and perform the best (or to be closest to safe if none are predicted to be safe). Although other $\dagger$ and $\ddagger$ could be used, we have only provided complete pseudocode for $(\dagger, \ddagger) \in \{(\text{NPDIS}, \text{CUT}), (\text{CWPDIS}, \text{BCa})\}$. Assumption 1 is not required.

\begin{itemize}
  \item[1] $\lambda^* \leftarrow \arg\max_{\lambda \in [0,1]} \text{CROSSVALIDATE}^{\dagger}_\ddagger(\lambda, D_{\text{train}}, \delta, \rho_-, m)$;
  \item[2] $\pi^* \leftarrow \arg\max_{\pi} f_{\dagger}^{\dagger}(\mu_{\lambda^*, \pi_0, \pi}, D_{\text{train}}, \delta, \rho_-, m)$;
  \item[3] return $\mu_{\lambda^*, \pi_0, \pi^*}$;
\end{itemize}
Approximate Confidence Intervals

• What if we knew that the importance weighted returns were normally distributed?
  • One-sided Student’s t-test.
**Algorithm 4.5:** $TT(X_1, \ldots, X_n, \delta)$: Uses the TT to return an approximate $1 - \delta$ confidence lower bound on $E[\frac{1}{n} \sum_{i=1}^{n} X_i]$.

**Assumes:** The $X_i$ are independent random variables with finite variance.

<table>
<thead>
<tr>
<th>Step</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\frac{1}{n} \sum_{i=1}^{n} X_i - \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}<em>n)^2}}{\sqrt{n}} t</em>{1-\delta,n-1}$</td>
</tr>
</tbody>
</table>
Approximate Confidence Intervals

- Bootstrap (BCa)
  - Estimate CDF with sample CDF:
    \[ F_n(x) := \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leq x} \]

- What if we assume that the samples come from a distribution like \( F_n \)?
  - Sort \( X_1, \ldots, X_n \) and return \( X_{\delta n} \)
ksdensity(gamrnd(2, 50, 10000000, 1))
Example: Gridworld
Example: Mountain Car
Example: Digital Marketing
DAEDALUS

• Apply SPI algorithm repeatedly.
• Per-iteration guarantee.
• DAEDALUS2: Exact HCPI until the first change to the policy, then approximate.
Example: Gridworld
Example: Mountain Car
Example: Digital Marketing
Safe policy improvement is tractable

A policy improvement algorithm that has a low probability of returning a bad policy.