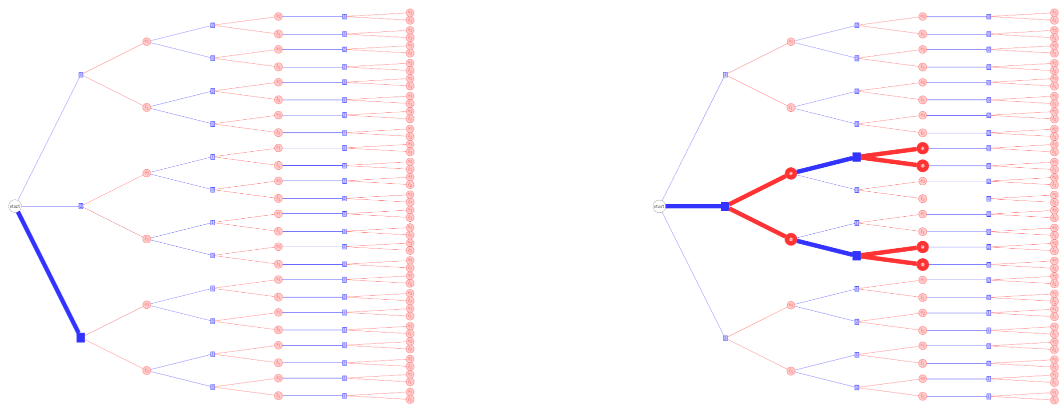
Supplementary Material

A Illustration of Policies



(a) A policy of just playing item 3. This policy has depth 1.

(b) A policy that plays item 2 first. If it is small, it plays item 1 whereas if it is large it plays item 3. After this, the final item is determined due to the fact that there are only 3 items in the problem. This policy has depth 2.

Figure 4: Examples of policies in the simple 3 item, 2 sizes stochastic knapsack problem. Each blue line represents choosing an item and the red lines represent the sizes of the previous items.

B Illustration of Bounds

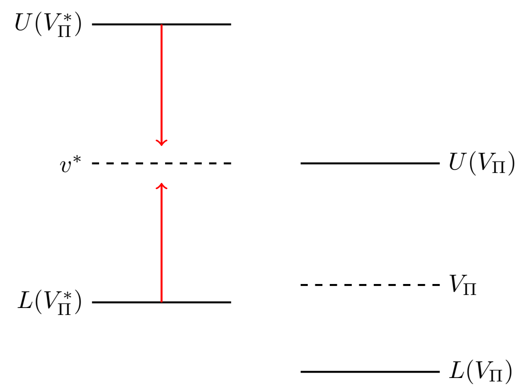


Figure 5: Example of where just looking at the optimistic policy might fail: If we always play the optimistic policy then, since $U(V^+_{\Pi^*}) \geq U(V^+_\Pi)$, we will always play Π^* and so the confidence bounds on Π will not shrink. This means that $L(V^+_{\Pi^*})$ will never be (epsilon) greater than the best alternative upper bound so there will not be enough confidence to conclude we have found the best policy.

C Algorithms

In these algorithms `Generate`(i) samples a reward and item size pair from the generative model of item i, whereas `sample`(A, k) samples from a set A with replacement to get k samples. The notation $i(d) = \Pi(d, b)$ indicates that item $i(d)$ was chosen by policy Π at depth d when the remaining capacity was b.

Algorithm 3: $\texttt{EstimateValue}(\Pi, m)$

Initialization: For all $i \in I$, $\mathcal{S}_i = \mathcal{S}_i^*$
for $j = 1, \ldots, m$ **do**
  $B_0 = B$;
  for $d = 1, \ldots, d(\Pi)$ **do**
    $i(d) = \Pi(d, B_{d-1})$;
    if $|\mathcal{S}_{i(d)}! \leq 0$ **then** $(r_{i(d)}, c_{i(d)}) = \texttt{Generate}(i(d))$, $\mathcal{S}_i^* = \mathcal{S}_i^* \cup \{r_{i(d)}, c_{i(d)})\}$;
      else $(r_{i(d)}, c_{i(d)}) = \texttt{sample}(\mathcal{S}_i, 1)$, and $\mathcal{S}_i = \mathcal{S}_i \setminus \{(r_{i(d)}, c_{i(d)})\}$;
    $B_d = B_{d-1} - c_{i(d)}$;
    if $B_d < 0$ **then** $r_{i(d)} = 0$;
  end
  $\overline{V_\Pi}^{(j)} = \sum_{d=1}^{d(\Pi)} r_{i(d)}$;
end
return $(\overline{V_\Pi}_m = \frac{1}{m} \sum_{j=1}^{m} \overline{V_\Pi}^{(j)}, \mathcal{S}^*)$

Algorithm 4: $\texttt{SampleBudget}(\Pi, \mathcal{S})$

Initialization: $B_0 = B$ and for all $i \in I$, $\mathcal{S}_i = \mathcal{S}_i^*$
for $d = 1, \ldots, d(\Pi)$ **do**
  $i(d) = \Pi(d, B_{d-1})$;
  if $|\mathcal{S}_{i(d)}| \leq 0$ **then** $(r_{i(d)}, c_{i(d)}) = \texttt{Generate}(i(d))$, $\mathcal{S}_i^* = \mathcal{S}_i^* \cup \{r_{i(d)}, c_{i(d)})\}$;
    else $(r_{i(d)}, c_{i(d)}) = \texttt{sample}(\mathcal{S}_i, 1)$, and $\mathcal{S}_i = \mathcal{S}_i \setminus \{(r_{i(d)}, c_{i(d)})\}$;
  $B_d = B_{d-1} - c_{i(d)}$;
end
$\overline{\Psi(B_\Pi)}^{(j)} = \Psi(\max\{B - \sum_{d=1}^{d(\Pi)} c_{i(d)}, 0\})$;
return $\left(\overline{\Psi(B_\Pi)}^{(j)}, \mathcal{S}^*\right)$

D Proofs of Theoretical Results

For convenience we restate any results that appear in the main body of the paper before proving them.

D.1 Bounding the Value of a Policy

Lemma 7 (Lemma 1 in main text) *Let $(\Omega, \mathcal{A}, P)$ be the probability space from Section 2, then for $m_1 + m_2$ independent samples of policy Π, and $\delta_1, \delta_2 > 0$, with probability $1 - \delta_1 - \delta_2$,*

$$\overline{V_\Pi}_{m_1} - c_1 \leq V_\Pi^+ \leq \overline{V_\Pi}_{m_1} + \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2.$$

Where $c_1 := \sqrt{\frac{\Psi(B)^2 \log(2/\delta_1)}{2m_1}}$ and $c_2 := \sqrt{\frac{\Psi(B)^2 \log(1/\delta_2)}{2m_2}}$.

Proof: Consider the average value of policy Π over m_1 many trials. By Hoeffding's Inequality, $P\left(|\overline{V_\Pi}_{m_1} - E[V_\Pi]| > c_1\right) \leq \delta_1$ and, similarly, $P\left(|\overline{\Psi(B_\Pi)}_{m_2} - E[\Psi(B_\Pi)]| > c_2\right) \leq \delta_2$. We are interested in the probability,

$$\begin{aligned} P(|\overline{V_\Pi}_{m_1} - V_\Pi^+| > t) &\leq P(|\overline{V_\Pi}_{m_1} - E[V_\Pi]| + |E[V_\Pi] - V_\Pi^+| > t) \\ &\leq P(|\overline{V_\Pi}_{m_1} - E[V_\Pi]| + E[\Psi(B_\Pi)] > t). \end{aligned}$$

where the first line follows from the triangle inequality and the second from the definition of $\Psi(B_\Pi)$. From the Hoeffding bounds and defining $t = \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2$, we consider $P\left(|\overline{V_\Pi}_{m_1} - E[V_\Pi]| + E[\Psi(B_\Pi)] > \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2\right)$. Define the events

$$A_1 = \{|\overline{V_\Pi}_{m_1} - V_\Pi| + E[\Psi(B_\Pi)] \leq E[\Psi(B_\Pi)] + c_1\} \text{ and } A_2 = \left\{|\overline{\Psi(B_\Pi)}_{m_2} - E[\Psi(B_\Pi)]| \leq c_2\right\}.$$

Then,

$$\begin{aligned}P\left(|\overline{V_\Pi}_{m_1} - E[V_\Pi]| + E[\Psi(B_\Pi)] > \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2\right) &\leq P(\Omega\backslash(A_1 \cap A_2))\\ &\leq P(\Omega\backslash A_1) + P(\Omega\backslash A_2)\\ &\leq \delta_1 + \delta_2.\end{aligned}$$

Hence,

$$P\left(\overline{V_\Pi}_{m_1} - V_\Pi^+ > c_1\right) \leq P\left(\overline{V_\Pi}_{m_1} - V_\Pi > c_1\right) \leq \delta_1 < \delta_1 + \delta_2$$

which gives the left hand side of the result. For the right hand side,

$$\begin{aligned}&P\left(\overline{V_\Pi}_{m_1} - V_\Pi^+ < -\overline{\Psi(B_\Pi)}_{m_2} - c_1 - c_2\right)\\ &\qquad \leq P\left(\overline{V_\Pi}_{m_1} - E[V_\Pi] - E[\Psi(B_\Pi)] < -\overline{\Psi(B_\Pi)}_{m_2} - c_1 - c_2\right)\\ &\qquad \leq \delta_1 + \delta_2.\end{aligned}$$

□

Lemma 8 *Let $\{Z_m\}_{m=1}^\infty$ be a martingale with Z_m defined on the filtration $\mathcal{F}_m$, $E[Z_m] = 0$ and $|Z_m - Z_{m-1}| \leq d$ for all m where $Z_0 = 0$. Then,*

$$P\left(\exists m \leq n; \frac{Z_m}{m} \geq 2d^2\sqrt{\frac{2}{m}\log\left(\frac{n}{m}\frac{4}{\delta}\right)}\right) \leq \delta$$

Proof: The proof is similar to that of Lemma B.1 in Perchet, Rigollet, Chassang, and Snowberg (2016) and will make use of the following standard results:

Theorem 9 Doob's maximal inequality: *Let Z be a non-negative submartingale. Then for $c > 0$,*

$$P\left(\sup_{k\leq n} Z_k \geq c\right) \leq \frac{E[Z_n]}{c}.$$

Proof: See, for example, Williams (1991), Theorem 14.6, page 137. □

Lemma 10 *Let Z_n be a martingale such that $|Z_i - Z_{i-1}| \leq d_i$ for all i with probability 1. Then, for $\lambda > 0$,*

$$E[e^{\lambda Z_n}] \leq e^{\frac{\lambda^2 D^2}{2}},$$

where $D^2 = \sum_{i=1}^n d_i^2$.

Proof: See the proof of the Azuma-Hoeffding inequality in Azuma (1967). □

Then, for the proof of Lemma 8, we first notice that since $\{Z_m\}_{m=1}^\infty$ is a martingale, by Jensen's inequality for conditional expectations, it follows that for any $\lambda > 0$,

$$E[e^{\lambda Z_m}|\mathcal{F}_{m-1}] \geq e^{\lambda E[Z_m|\mathcal{F}_{m-1}]} = e^{\lambda Z_{m-1}}.$$

Hence, for any $\lambda > 0$, $\{e^{\lambda Z_m}\}_{m=1}^\infty$ is a positive sub-martingale so we can apply Doob's maximal inequality (Theorem 9) to get

$$P\left(\sup_{m\leq n} Z_m \geq c\right) = P\left(\sup_{m\leq n} e^{\lambda Z_m} \geq e^{\lambda c}\right) \leq \frac{E[e^{\lambda Z_n}]}{e^{\lambda c}}.$$

Then, by Lemma 10, since $|Z_i - Z_{i-1}| \leq d$ for all i, it follows that

$$P\left(\sup_{m\leq n} Z_m \geq c\right) \leq \frac{E[e^{\lambda Z_n}]}{e^{\lambda c}} \leq \frac{e^{\lambda^2 D^2/2}}{e^{\lambda c}} = \exp\left\{\frac{\lambda^2 D^2}{2} - \lambda c\right\}. \tag{5}$$

Minimizing the right hand side with respect to λ gives $\hat{\lambda} = \frac{c}{D^2}$ and substituting this back into (5) gives,

$$P\left(\sup_{m\leq n} Z_m \geq c\right) \leq \exp\left\{-\frac{c^2}{2D^2}\right\}.$$

Then, since we are considering the case where $d_i = d$ for all i, $D^2 = nd^2$ and so,

$$P\left(\sup_{m\leq n} Z_m \geq c\right) \leq \exp\left\{-\frac{c^2}{2nd^2}\right\}.$$

Further, if we are interested in $P(\sup_{k\leq m\leq n} Z_m \geq c)$, we can redefine the indices to get

$$P\left(\sup_{k\leq m\leq n} Z_m \geq c\right) = P\left(\sup_{m'\leq n-k+1} Z_m \geq c\right) \leq \exp\left\{-\frac{c^2}{2(n-k+1)d^2}\right\}. \tag{6}$$

We then define $\varepsilon_m = 2d\sqrt{\frac{1}{m}\log\left(\frac{n}{m}\frac{8}{\delta}\right)}$ and use a peeling argument similar to that in Lemma B.1 of Perchet et al. (2016) to get

$$\begin{aligned}
P\left(\exists m \leq n; \frac{Z_m}{m} \geq \varepsilon_m\right) &\leq \sum_{t=0}^{\lfloor \log_2(n)\rfloor+1} P\left(\bigcup_{m=2^t}^{2^{t+1}-1}\left\{\frac{Z_m}{m} \geq \varepsilon_m\right\}\right) && \text{(by union bound)}\\
&\leq \sum_{t=0}^{\lfloor \log_2(n)\rfloor+1} P\left(\bigcup_{m=2^t}^{2^{t+1}-1}\left\{\frac{Z_m}{m} \geq \varepsilon_{2^{t+1}}\right\}\right) && \text{(since } \varepsilon_m \text{ decreasing in } m)\\
&\leq \sum_{t=0}^{\lfloor \log_2(n)\rfloor+1} P\left(\bigcup_{m=2^t}^{2^{t+1}-1}\left\{Z_m \geq 2^t\varepsilon_{2^{t+1}}\right\}\right) && \text{(as } m \geq 2^t)\\
&\leq \sum_{t=0}^{\lfloor \log_2(n)\rfloor+1} \exp\left\{-\frac{(2^t\varepsilon_{2^{t+1}})^2}{2^{t+1}d^2}\right\} && \text{(from (6))}\\
&\leq \sum_{t=0}^{\lfloor \log_2(n)\rfloor+1} \frac{2^{t+1}\delta}{8n} && \text{(substituting } \varepsilon_{2^{t+1}})\\
&\leq \frac{2^{\log_2(n)+3}\delta}{8n} = \delta. && \left(\text{since } \sum_{i=1}^{k} 2^i = 2^{k+1}-1\right)
\end{aligned}$$

□

Proposition 11 (Proposition 2 in main text) *The Algorithm* `BoundValueShare` *(Algorithm 2) returns confidence bounds,*

$$L(V_\Pi^+) = \overline{V_\Pi}_{m_1} - \sqrt{\frac{\Psi(B)^2\log(2/\delta_1)}{2m_1}} \quad U(V_\Pi^+) = \overline{V_\Pi}_{m_1} + \overline{\Psi(B_\Pi)}_{m_2} + \sqrt{\frac{\Psi(B)^2\log(2/\delta_1)}{2m_1}} + 2\Psi(B)\sqrt{\frac{1}{m_2}\log\left(\frac{8n}{\delta_2 m_2}\right)}$$

which hold with probability $1-\delta_1-\delta_2$.

Proof: We begin by noting that our samples of item size are dependent since in each iteration we construct a bound based on past samples and we use this bound to decide if we need to continue sampling or if we can stop. To model this dependence let us introduce a stopping time τ such that $\tau(\omega) = n$ if our algorithm exits the loop at time n. Consider the sequence

$$\overline{\Psi(B_\Pi)}_{1\wedge\tau}, \overline{\Psi(B_\Pi)}_{2\wedge\tau}, \dots$$

and define for $m \geq 1$

$$M_m = (m\wedge\tau)(\overline{\Psi(B_\Pi)}_{m\wedge\tau} - E[\Psi(B_\Pi)]) \quad \text{with} \quad M_0 = 0.$$

Furthermore, define the filtration $\mathcal{F}_m = \sigma(B_{\Pi,1}, \ldots, B_{\Pi,m})$ then for $m \geq 1$

$$E[M_m|\mathcal{F}_{m-1}] = E[M_m|\mathcal{F}_{m-1}, \tau \leq m-1] + E[M_m|\mathcal{F}_{m-1}, \tau > m-1].$$

Now

$$E[M_m|\mathcal{F}_{m-1}, \tau \leq m-1] = E[M_{m-1}|\tau \leq m-1].$$

and due to independence of the samples $B_{\Pi,1}, \ldots, B_{\Pi,m}$

$$\begin{aligned}
&E[M_m|\mathcal{F}_{m-1}, \tau > m-1] \\
&\quad = E[m(\overline{\Psi(B_\Pi)}_m - E[\Psi(B_\Pi)])|\mathcal{F}_{m-1}, \tau > m-1] \\
&\quad = E\left[\sum_{j=1}^{m-1} \Psi(B_{\Pi,j}) + \Psi(B_{\Pi,m}) - mE[\Psi(B_\Pi)] \middle| \mathcal{F}_{m-1}, \tau > m-1\right] \\
&\quad = (m-1)E[\overline{\Psi(B_\Pi)}_{m-1} - E[\Psi(B_\Pi)]|\mathcal{F}_{m-1}, \tau > m-1] \\
&\qquad\qquad + E[\Psi(B_{\Pi,m}) - E[\Psi(B_\Pi)]|\mathcal{F}_{m-1}, \tau > m-1] \\
&\quad = E[M_{m-1}|\tau > m-1] + E[\Psi(B_{\Pi,m})] - E[\Psi(B_\Pi)] = E[M_{m-1}|\tau > m-1].
\end{aligned}$$

Hence, $E[M_m|\mathcal{F}_{m-1}] = M_{m-1}$ and M_m is a martingale with increments $|M_m - M_{m-1}| \leq |\Psi(B_{\Pi,m}) - E[\Psi(B_\Pi)]| \leq \Psi(B)$. We could apply the Azuma-Hoeffding inequality to gain guarantees for individual m-values. Alternatively, we can use Lemma 8 to get,

$$P\left(\sup_{m \leq n} \frac{M_m}{m} \geq 2\Psi(B)\sqrt{\frac{1}{m}\log\left(\frac{8n}{\delta m}\right)}\right) \leq \delta_2.$$

Combining this with the argument in Lemma 1 gives

$$\overline{V_\Pi}_{m_1} - c_1 \leq V_\Pi^+ \leq \overline{V_\Pi}_{m_1} + \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2,$$

where $c_1 := \sqrt{\frac{\Psi(B)^2 \log(2/\delta_1)}{2m_1}}$ and $c_2 := 2\Psi(B)\sqrt{\frac{1}{m_2}\log\left(\frac{8n}{\delta_2 m_2}\right)}$ and these bounds hold with probability $1 - \delta_1 - \delta_2$.

□

Lemma 12 *With probability* $1 - \delta_{0,1} - \delta_{0,2}$*, the bounds generated by* `BoundValueShare` *with parameters* $\delta_{1,d} = \frac{\delta_{0,1}}{d^*} N_d^{-1}$ *and* $\delta_{2,d} = \frac{\delta_{0,2}}{d^*} N_d^{-1}$ *hold for all policies* Π *of depth* $d = d(\Pi) \leq d^*$ *simultaneously.*

Proof: The probability that all bounds hold simultaneously is $P(\bigcap_{\Pi \in \mathcal{P}} \{L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)\})$ where $\mathcal{P}$ is the set of all policies. From Proposition 2, for any policy Π of depth $d = d(\Pi)$, $P(L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)) \geq 1 - \delta_{d,1} - \delta_{d,2}$. Then,

$$\begin{aligned}
P\left(\bigcap_{\Pi \in \mathcal{P}} \{L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)\}\right) &= 1 - P\left(\bigcup_{\Pi \in \mathcal{P}} \{L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)\}^c\right) \\
&\geq 1 - \sum_{\Pi \in \mathcal{P}} P(\{L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)\}^c) \\
&\geq 1 - \sum_{\Pi \in \mathcal{P}} (\delta_{d(\Pi),1} + \delta_{d(\Pi),2}) \\
&= 1 - \sum_{d=1}^{d^*} N_d(\delta_{d,1} + \delta_{d,2}) \\
&\geq 1 - \sum_{d=1}^{d^*} N_d\left(\frac{\delta_{0,1}}{d^*} N_d^{-1} + \frac{\delta_{0,2}}{d^*} N_{d(\Pi_t)}^{-1}\right) \\
&= 1 - \sum_{d=1}^{d^*} \frac{1}{d^*}(\delta_{0,1} + \delta_{0,2}) = 1 - \delta_{0,1} - \delta_{0,2}
\end{aligned}$$

□

D.2 Theoretical Results for Optimistic Stochastic Knapsacks (OpStoK)

Proposition 13 (Proposition 4 in main text) *With probability at least* $(1 - \delta_{0,1} - \delta_{0,2})$*, the algorithm* `OpStoK` *returns a policy with value at least* $v^* - \epsilon$.

Proof: The proof follows from the following lemma.

Lemma 14 *For every round of the algorithm and incomplete policy* Π*, let* $D(\Pi)$ *be the set of all descendants of* Π*. Define the event* $A = \bigcap_{\Pi' \in D(\Pi)} \{V_{\Pi'} \in [L(V_\Pi^+), U(V_\Pi^+)]\}$*. Then* $P(A) \geq 1 - \delta_{0,1} - \delta_{0,2}$.

Proof: When `BoundValueShare` is called for a policy Π with $d(\Pi) = d$, it is done so with parameters $\delta_{d,1} = \frac{\delta_{0,1}}{d^*} N_d^{-1}$ and $\delta_{d,2} = \frac{\delta_{0,2}}{d^*} N_d^{-1}$, where $\delta_{d,1}$ and $\delta_{d,2}$ are used to control the accuracy of the estimated value of V_Π and $E\Psi(B_\Pi)$ respectively. It follows from Proposition 2, that for any active policy Π, the probability that the interval $\left[\overline{V_\Pi}_{m_1} - c_1, \overline{V_\Pi}_{m_1} + \overline{\Psi(B_\Pi)}_{m_2} + c_1 + c_2\right]$ generated by `BoundValueShare` does not contain V_Π^+ is less than $\delta_{d,1} + \delta_{d,2}$. Furthermore, from standard Hoeffding bounds, the probability that V_Π is outside the interval $[V_\Pi - c_1, V_\Pi + c_1]$ is less than $\delta_{d,1}$. Since any descendant policy Π' of Π consists of adding at least one item to the knapsack and item rewards are all ≥ 0, it follows that $V_\Pi \leq V_{\Pi'} \leq V_\Pi^+$. Hence, the probability of the value of a descendant policy being outside the interval $[L(V_\Pi^+), U(V_\Pi^+)]$ is less than $\delta_{d,1} + \delta_{d,2}$. By the same argument as in Lemma 12, it can be shown that $P(A) > 1 - \sum_{d=1}^{d^*} (\delta_{d,1} + \delta_{d,2}) N_d = 1 - \delta_{0,1} - \delta_{0,2}$. □

The result of the proposition follows by noting that the true optimal policy Π^{OPT} will be a descendant of Π_i for some $i \in I$. Let Π^* be the policy outputted by the algorithm. By the stopping criterion, $L(V_{\Pi^*}^+) + \epsilon \geq \max_{\Pi \in \text{ACTIVE} \setminus \{\Pi^*\}} \geq U(V_\Pi^+)$ for any $\Pi \in$ ACTIVE. From the expansion rule of `OpStoK`, it follows that either $\Pi^{OPT} \in$ ACTIVE or there exists some ancestor policy Π' of Π^{OPT} in ACTIVE. In the first case, $V_{\Pi^{OPT}} = v^* \leq U(V_{\Pi^{OPT}}^+)$ whereas in the latter $V_{\Pi^{OPT}} = v^* \leq U(V_{\Pi'}^+)$ with high probability from Lemma 14. In either case, it follows that $L(V_{\Pi^*}^+) + \epsilon \geq v^*$ and so $V_{\Pi^*} + \epsilon \geq v^*$.

□

Lemma 15 *If* Π *is a complete policy then,* $U(V_\Pi^+) - L(V_\Pi^+) \leq \epsilon$*, otherwise* $U(V_\Pi^+) - L(V_\Pi^+) \leq 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon$.

Proof: By the bounds in Proposition 2, $U(V_\Pi^+) - L(V_\Pi^+) \leq \overline{\Psi(B_\Pi)}_{m_2} + c_2 + 2c_1 = U(\Psi(B_\Pi)) + 2c_1$. For a complete policy, $U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}$ and according to `BoundValueShare`, m_1 is chosen such that $2c_1 \leq \frac{\epsilon}{2}$ which implies $U(V_\Pi^+) - L(V_\Pi^+) \leq \epsilon$.
If Π is not complete, by the sampling strategy in `BoundValueShare`, we continue sampling the remaining budget until $L(\Psi(B_\Pi)) \geq \frac{\epsilon}{4}$. In this setting, the maximal width of the confidence interval of $E\Psi(B_\Pi)$ will satisfy

$$2c_2 \leq E\Psi(B_\Pi) - \frac{\epsilon}{4}. \tag{7}$$

Hence,

$$\begin{aligned}
U(V_\Pi^+) - L(V_\Pi^+) &\leq U(\Psi(B_\Pi)) + 2c_1 \\
&\leq 3U(\Psi(B_\Pi)) \qquad (8) \\
&\leq 3(E\Psi(B_\Pi) + 2c_2) \\
&\leq 3\left(E\Psi(B_\Pi) + E\Psi(B_\Pi) - \frac{\epsilon}{4}\right) \qquad (9) \\
&\leq 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon.
\end{aligned}$$

Where (8) follows since, when $L(\Psi(B_\Pi)) \geq \frac{\epsilon}{4}$, we sample the value of policy Π until $c_1 \leq U(\Psi(B_\Pi))$, and (9) by substituting in (7). □

Lemma 16 (Lemma 3 in main text) *Assume that* $L(V_\Pi^+) \leq V_\Pi \leq U(V_\Pi^+)$ *holds simultaneously for all policies* $\Pi \in$ ACTIVE *with* $U(V_\Pi^+)$ *and* $L(V_\Pi^+)$ *as defined in Proposition 2. Then,* $\Pi_t \in \mathcal{Q}^\epsilon$ *for every policy selected by* ***OpStoK*** *at every time point* t*, except for possibly the last one.*

Proof: Since, when we expand a policy, we replace it in ACTIVE by all its child policies, at any time point $t \geq 1$ there will be one ancestor of Π^* in the active set, denote this policy by Π^*_t. If $\Pi_t = \Pi^*_t$, then by Lemma 14, $V_{\Pi^*} \in [L(V^+_{\Pi_t}), U(V^+_{\Pi_t})]$. Hence,

$$V_\Pi + 6E\Psi(B_\Pi) + \frac{3}{4}\epsilon \geq U(V^+_\Pi) \geq v^* \geq v^* - 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon + \epsilon.$$

Where the last inequality will hold for any incomplete policy (since for an incomplete policy $L(B_\Pi) \geq \frac{\epsilon}{4}$) and so, $\Pi_t \in \mathcal{Q}^\epsilon$. For $\Pi_t = \Pi^*$, since $\frac{6}{4}\epsilon \geq \epsilon$, $\Pi_t \in \mathcal{Q}^\epsilon$.

Assume $\Pi_t \neq \Pi^*_t$. If Π_t is a complete policy, $U(V^+_{\Pi_t}) - L(V^+_{\Pi_t}) \leq \epsilon$. For a complete policy Π to be selected, it must have the largest $U(V^+_\Pi)$, since most alternative policies will have larger $U(\Psi(B_\Pi))$. Hence $\Pi^{(1)}_t = \Pi_t$ and

$$L(V^+_{\Pi^{(1)}_t}) + \epsilon \geq U(V^+_{\Pi^{(1)}_t}) \geq \max_{\Pi \in \text{ACTIVE} \setminus \{\Pi^{(1)}_t\}} U(V^+_\Pi),$$

so the algorithm stops.

Assume $\Pi_t = \Pi^{(1)}_t \neq \Pi^*_t$ is an incomplete policy. By Lemma 15, for an incomplete policy,

$$U(V^+_\Pi) - L(V^+_\Pi) \leq 3U(\Psi(B_\Pi)) \leq 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon. \tag{10}$$

Then, if the termination criteria is not met,

$$\begin{aligned} V_{\Pi_t} \geq L(V^+_{\Pi_t}) \quad \implies \quad V_{\Pi_t} + 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon - \epsilon &\geq L(V^+_{\Pi_t}) + 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon - \epsilon \\ &\geq U(V^+_{\Pi_t}) - \epsilon \\ &\geq \max_{\Pi \in \text{ACTIVE} \setminus \{\Pi_t\}} U(V^+_\Pi) - \epsilon \\ &\geq L(V^+_{\Pi_t}) \\ &\geq U(V^+_{\Pi_t}) - 6E\Psi(B_\Pi) + \frac{3}{4}\epsilon \\ &\geq U(V^+_{\Pi^*_t}) - 6E\Psi(B_\Pi) + \frac{3}{4}\epsilon \\ &\geq v^* - 6E\Psi(B_\Pi) + \frac{3}{4}\epsilon \end{aligned}$$

which follows since $\Pi^{(1)}_t$ is chosen to be the policy with largest upper bound. Therefore, $\Pi_t \in \mathcal{Q}^\epsilon$.

By the stopping criteria of `OpStoK`, if the algorithm does not stop and select $\Pi^{(1)}_t$ as the optimal policy, then $\Pi_t = \Pi^{(2)}_t$ and

$$L(V^+_{\Pi^{(1)}_t}) + \epsilon < \max_{\Pi \in \text{ACTIVE} \setminus \{\Pi^{(1)}_t\}} U(V^+_\Pi) = U(V^+_{\Pi^{(2)}_t}).$$

By equation (10),

$$L(V^+_{\Pi^{(1)}_t}) + 6E\Psi(B_\Pi) - \frac{3}{4}\epsilon \geq U(V^+_{\Pi^{(1)}_t}).$$

and by the selection criterion $U(\Psi(B_{\Pi_t^{(2)}})) \geq U(\Psi(B_{\Pi_t^{(1)}}))$. Therefore, for $\Pi_t = \Pi_t^{(2)} \neq \Pi_t^*$,

$$\begin{aligned} V_{\Pi_t} + 12E\Psi(B_\Pi) - \frac{6}{4}\epsilon - \epsilon &\geq L(V_{\Pi_t}^+) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon - \epsilon \\ &\geq U(V_{\Pi_t}^+) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon - \epsilon \\ &\geq U(V_{\Pi_t}^+) + 3U(\Psi(B_{\Pi_t})) - \epsilon \\ &\geq U(V_{\Pi_t}^+) + 3U(\Psi(B_{\Pi_t^{(1)}})) - \epsilon \\ &\geq L(V_{\Pi_t^{(1)}}^+) + 3U(\Psi(B_{\Pi_t^{(1)}})) \\ &\geq U(V_{\Pi_t^{(1)}}^+) \\ &\geq U(V_{\Pi_t^*}^+) \\ &\geq v^*. \end{aligned}$$

Hence $\Pi_t \in \mathcal{Q}^\epsilon$. □

Theorem 17 (Theorem 5 in main text) *The total number of samples required by `OpStoK` is bounded from above by,*

$$\sum_{\Pi \in \mathcal{Q}^\epsilon} (m_1(\Pi) + m_2(\Pi))\, d(\Pi),$$

with probability $1 - \delta_{0,2}$.

Proof: The result follows from the following three lemmas.

Lemma 18 *For* $\Pi \in \mathcal{A}^\epsilon$ *of depth* $d = d(\Pi)$*, then, with probability* $1 - \delta_{d,2}$*, the minimum number of samples of the value and remaining budget of the policy* Π *are bounded by*

$$m_1(\Pi) = \left\lceil \frac{8\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{\epsilon^2} \right\rceil \qquad \text{and} \qquad m_2(\Pi) = m^*,$$

where m^* *is the smallest integer satisfying* $\frac{16\Psi(B)^2}{(E\Psi(B_\Pi) - \epsilon/2)^2} \leq \frac{m}{\log(8n/m\delta_2)}$ *with* n *defined as in* (2).

Proof: When $E\Psi(B_\Pi) \leq \frac{\epsilon}{4}$, the event $\{U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}\}$ will eventually occur with enough samples of the remaining budget of the policy. With probability greater than $1-\delta_{d,2}$, this will happen when $2c_2 \leq \frac{\epsilon}{2} - E\Psi(B_\Pi)$, since by Hoeffding's Inequality we know $\overline{\Psi(B_\Pi)}_{m_2} \in [E\Psi(B_\Pi) - c_2, E\Psi(B_\Pi) + c_2]$ where c_2 is as defined in Lemma 1. From this, it follows that $U(\Psi(B_\Pi)) \in [E\Psi(B_\Pi), E\Psi(B_\Pi) + 2c_2]$. We want to make sure that $U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}$ will eventually happen so we need to construct a confidence interval such that c_2 satisfies $E\Psi(B_\Pi) + 2c_2 \leq \frac{\epsilon}{2}$. Therefore we select m_2 such that,

$$\begin{aligned} & 2c_2 \leq \frac{\epsilon}{2} - E\Psi(B_\Pi) \\ \Longrightarrow\ & 4\Psi(B)\sqrt{\frac{2\log(\frac{8n}{m_2\delta_{d,2}})}{m_2}} \leq \frac{\epsilon}{2} - E\Psi(B_\Pi) \\ \Longrightarrow\ & \frac{16\Psi(B)^2}{(E\Psi(B_\Pi) - \epsilon/2)^2} \leq \frac{m_2}{\log(4n/m_2\delta_2)}. \end{aligned}$$

Defining, $m_2(\Pi) = m^*$, where m^* is the smallest integer satisfying the above, is therefore an upper bound on the minimum number of samples necessary to ensure that $U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}$ with probability greater than $1 - \delta_{d,2}$. When $U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}$, `BoundValueShare` requires $m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{\epsilon^2} \right\rceil$ samples of the value of the policy to ensure $2c_1 \leq \frac{\epsilon}{2}$. □

Lemma 19 *For $\Pi \in \mathcal{B}^\epsilon$ of depth $d = d(\Pi)$, then, with probability $1 - \delta_{d,2}$, the minimum number of samples of the value and remaining budget of the policy Π are bounded by*

$$m_1(\Pi) \leq \left\lceil \frac{\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{2E\Psi(B_\Pi)^2} \right\rceil \quad \text{and} \quad m_2(\Pi) = m^*,$$

where m^* is the smallest integer satisfying $\frac{16\Psi(B)^2}{(E\Psi(B_\Pi) - \epsilon/4)^2} \leq \frac{m}{\log(8n/m\delta_2)}$ with n defined as in (2).

Proof: When $E\Psi(B_\Pi) \geq \frac{\epsilon}{2}$, by noting that the event $\{L(\Psi(B_\Pi)) \geq \frac{\epsilon}{4}\}$ will eventually happen and using a very similar argument to Lemma 18, it follows that $m_2(\Pi)$ is the smallest integer solution to

$$\frac{16\Psi(B)^2}{(E\Psi(B_\Pi) - \epsilon/4)^2} \leq \frac{m}{\log(8n/m\delta_2)},$$

with probability greater than $1 - \delta_{d,2}$. Whenever $L(\Psi(B_\Pi)) \geq \frac{\epsilon}{4}$, `BoundValueShare` requires $m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{(U(\Psi(B_\Pi))^2} \right\rceil$ samples of the value of policy Π. Since $U(\Psi(B_\Pi)) \in [E\Psi(B_\Pi), E\Psi(B_\Pi) + 2c_2]$ with probability $1 - \delta_{0,2}$, $U(\Psi(B_\Pi)) \geq E\Psi(B_\Pi)$, and so,

$$m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{(U(\Psi(B_\Pi))^2} \right\rceil \leq \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{E\Psi(B_\Pi)^2} \right\rceil$$

and the result holds. □

Lemma 20 *For $\Pi \in \mathcal{C}^\epsilon$ of depth $d = d(\Pi)$, then, with probability $1 - \delta_{d,2}$, the minimum number of samples of the value and remaining budget of the policy Π are bounded by*

$$m_1(\Pi) \leq \max \left\{ \left\lceil \frac{8\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{\epsilon^2} \right\rceil, \left\lceil \frac{\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{2E\Psi(B_\Pi)^2} \right\rceil \right\}$$

and $m_2(\Pi) = m^*$,where m^* is the smallest integer satisfying $\frac{16\Psi(B)^2}{(\epsilon/4)^2} \leq \frac{m}{\log(8n/m\delta_2)}$ with n defined as in (2).

Proof: When $\frac{\epsilon}{4} < E\Psi(B_\Pi) < \frac{\epsilon}{2}$, then the minimum width we will need a confidence interval to be is $\epsilon/4$. By an argument similar to Lemma 18, we can deduce that $m_2(\Pi)$ will be the smallest integer satisfying $\frac{16\Psi(B)^2}{(\epsilon/4)^2} \leq \frac{m}{\log(8n/m\delta_2)}$.

In order to determine the number of samples of the value required by `BoundValueShare`, we need to know which of $\{U(\Psi(B_\Pi)) \leq \frac{\epsilon}{2}\}$ or $\{L(\Psi(B_\Pi)) \geq \frac{\epsilon}{4}\}$ occurs first. However, when $\Pi \in \mathcal{C}^\epsilon$, we do not know this so the best we can do is bound $m_1(\Pi)$ by the maximum of the two alternatives,

$$m_1(\Pi) \leq \max \left\{ \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{\epsilon^2} \right\rceil, \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,1}})}{E\Psi(B_\Pi)^2} \right\rceil \right\}.$$

□

The result of the theorem then follows by noting that for any policy Π of depth $d(\Pi)$, it will be necessary to have $m_1(\Pi)$ samples of the value of the policy and $m_2(\Pi)$ samples of the value of the policy. This requires $m_1(\Pi)d(\Pi)$ samples of item rewards, $m_1(\Pi)d(\Pi)$ samples of item sizes (to calculate the rewards) and $m_2(\Pi)d(\Pi)$ samples of item sizes (to calculate remaining budget), thus a total of $(m_1(\Pi)+m_2(\Pi))d(\Pi)$ calls to the generative model. From Lemma 3, any policy expanded by `OpStoK` will be in $\mathcal{Q}^\epsilon$ so it suffices to sum over all policies in $\mathcal{Q}^\epsilon$. This result assumes that all confidence bounds hold, whereas we know that for any policy Π of depth $d(\Pi)$, the probability of the confidence bound holding is greater than $1 - \delta_{d,2}$. By an argument similar to Lemma 12, the probability that all bounds hold is greater than $1 - \delta_{0,2}$. Note that, since $|\mathcal{Q}^\epsilon| \leq |\mathcal{P}|$, the probability should be considerably greater than $1 - \delta_{0,2}$. □