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

\[
\begin{align*}
U(V_\Pi) & \quad v^* \quad U(V_\Pi) \\
L(V_\Pi) & \quad V_\Pi \quad L(V_\Pi)
\end{align*}
\]

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 \(\Pi^*\) and so the confidence bounds on \(\Pi\) 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 \texttt{Generate}(i) samples a reward and item size pair from the generative model of item \(i\), whereas \texttt{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 \(\Pi\) at depth \(d\) when the remaining capacity was \(b\).
Algorithm 3: EstimateValue(Π, m)

Initialization: For all \( i \in I \), \( S_i = S_i^* \)
1. for \( j = 1, \ldots, m \) do
2. \( B_0 = B \);
3. for \( d = 1, \ldots, d(Π) \) do
4. \( i(d) = Π(d, B_{d-1}) \);
5. if \( |S_i(d)| \leq 0 \) then \((r_i(d), c_i(d)) = \text{Generate}(i(d)), S_i^* = S_i^* \cup \{r_i(d), c_i(d)\}\);
6. else \((r_i(d), c_i(d)) = \text{sample}(S_i, 1)\), and \( S_i = S_i \setminus \{(r_i(d), c_i(d))\}\);
7. \( B_d = B_{d-1} - c_i(d) \);
8. if \( B_d < 0 \) then \( r_i(d) = 0 \);
9. end
10. \( \overline{V}_Π^{(j)} = \sum_{d=1}^{d(Π)} r_i(d) \);
11. end
12. return \( \overline{V}_{Πm} = \frac{1}{m} \sum_{j=1}^{m} \overline{V}_Π^{(j)}, S^* \)

Algorithm 4: SampleBudget(Π, S)

Initialization: \( B_0 = B \) and for all \( i \in I \), \( S_i = S_i^* \)
1. for \( d = 1, \ldots, d(Π) \) do
2. \( i(d) = Π(d, B_{d-1}) \);
3. if \( |S_i(d)| \leq 0 \) then \((r_i(d), c_i(d)) = \text{Generate}(i(d)), S_i^* = S_i^* \cup \{r_i(d), c_i(d)\}\);
4. else \((r_i(d), c_i(d)) = \text{sample}(S_i, 1)\), and \( S_i = S_i \setminus \{(r_i(d), c_i(d))\}\);
5. \( B_d = B_{d-1} - c_i(d) \);
6. end
7. \( \overline{Ψ(BΠ)}^{(j)} = Ψ(\max\{B - \sum_{d=1}^{d(Π)} c_i(d), 0\}) \);
8. return \( \overline{Ψ(BΠ)}^{(j)}, S^* \)

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 \((Ω, A, P)\) be the probability space from Section 2, then for \( m_1 + m_2 \) independent samples of policy \( Π \), and \( δ_1, δ_2 > 0 \), with probability \( 1 - δ_1 - δ_2 \),

\[
\overline{V}_{Πm_1} - c_1 \leq V^+_Π \leq \overline{V}_{Πm_1} + \overline{Ψ(BΠ)}_{m_2} + c_1 + c_2.
\]

Where \( c_1 := \sqrt{Ψ(Π)^2 \log(2/δ_1)} / 2m_1 \) and \( c_2 := \sqrt{Ψ(Π)^2 \log(1/δ_2)} / 2m_2 \).

Proof: Consider the average value of policy \( Π \) over \( m_1 \) many trials. By Hoeffding’s Inequality,

\[
P\left(\overline{V}_{Πm_1} - E[Π] > c_1\right) \leq δ_1 \text{ and, similarly, } P\left(\overline{Ψ(BΠ)}_{m_2} - E[Ψ(BΠ)] > c_2\right) \leq δ_2.
\]

We are interested in the probability,

\[
P\left(\overline{V}_{Πm_1} - V^+_Π > t\right) \leq P\left(\overline{V}_{Πm_1} - E[Π] + |E[Π] - V^+_Π| > t\right) \leq P\left(\overline{V}_{Πm_1} - E[Π] + E[Ψ(BΠ)] > t\right).
\]

where the first line follows from the triangle inequality and the second from the definition of \( Ψ(BΠ) \). From the Hoeffding bounds and defining \( t = \overline{Ψ(BΠ)}_{m_2} + c_1 + c_2 \), we consider

\[
P\left(\overline{V}_{Πm_1} - E[Π] + E[Ψ(BΠ)] > \overline{Ψ(BΠ)}_{m_2} + c_1 + c_2\right).
\]

Define the events

\[
A_1 = \left\{\overline{V}_{Πm_1} - V^+_Π \leq E[Ψ(BΠ)] + c_1 \right\} \text{ and } A_2 = \left\{\overline{Ψ(BΠ)}_{m_2} - E[Ψ(BΠ)] \leq c_2 \right\}.
\]
Lemma 10

Let $\lambda > 0$, conditional expectations, it follows that for any $\lambda > 0$, $e^{\lambda Z_m}$ is a positive sub-martingale so we can apply Doob’s maximal inequality (Theorem 9) to get

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

Then, by Lemma 10, 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 $\lambda > 0$,

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

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

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

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

$$P \left( \sup_{m \leq n} Z_m \geq c \right) \leq \frac{E[e^{\lambda Z_m}]}{e^{\lambda c}} \leq \frac{e^{\lambda^2 D^2/2}}{e^{\lambda c}} = \exp \left\{ \frac{\lambda^2 D^2}{2} - \lambda c \right\}.$$
To model this dependence let us introduce a stopping time \( \tau \) based on past samples and we use this bound to decide if we need to continue sampling or if we can stop. We begin by noting that our samples of item size are dependent since in each iteration we construct a

Proof:

which hold with probability \( 1 - \delta_1 - \delta_2 \).

Proposition 11 (Proposition 2 in main text) The Algorithm \textbf{BoundValueShare} (Algorithm 2) returns confidence bounds,

\[
L(V_{II}^+ \pi) = V_{II m_1} - \sqrt{\frac{\Psi(B)^2 \log(2/\delta_1)}{2m_1}} \quad U(V_{II}^+ \pi) = V_{II m_1} + \Psi(B)_{m_1} + \frac{\sqrt{\Psi(B)^2 \log(2/\delta_1)}}{2m_1} + 2\Psi(B)\sqrt{\frac{1}{m_2} \log \left( \frac{8n}{\delta_2 m_2} \right)}
\]

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 \( \tau \) such that \( \tau(\omega) = n \) if our algorithm exits the loop at time \( n \). Consider the sequence

and define for \( m \geq 1 \)

\[
M_m = (m \land \tau)(\Psi(B) m_\land \tau - E[\Psi(B m)]) \quad \text{with} \quad M_0 = 0.
\]
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 $B_{\Pi,1}, \ldots, B_{\Pi,m}$

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

Combining this with the argument in Lemma 1 gives, for $m \geq 1$

$$\delta \Omega = \frac{\delta}{\sigma} N^{-1}_d$$

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

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

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 $1 - \delta_1 - \delta_2$.

\begin{lemma}
With probability $1 - \delta_{0.1} - \delta_{0.2}$, the bounds generated by BoundValueShare with parameters $\delta_{1,d} = \frac{\delta_{0.1}}{\sigma^2} N^{-1}_d$ and $\delta_{2,d} = \frac{\delta_{0.2}}{\sigma^2} N^{-1}_d$ hold for all policies $\Pi$ of depth $d = d(\Pi) \leq d^*$ simultaneously.
\end{lemma}

\begin{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 $\Pi$ 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,

$$P \left( \bigcap_{\Pi \in \mathcal{P}} \{L(V_{\Pi}^+ \leq V_{\Pi} \leq U(V_{\Pi}^+))\} \right) = 1 - P \left( \bigcup_{\Pi \in \mathcal{P}} \{L(V_{\Pi}^+ \leq V_{\Pi} \leq U(V_{\Pi}^+))\}^c \right) \geq 1 - \sum_{\Pi \in \mathcal{P}} P(\{L(V_{\Pi}^+ \leq V_{\Pi} \leq U(V_{\Pi}^+))\}^c) \geq 1 - \sum_{\Pi \in \mathcal{P}} (\delta_{d(\Pi),1} + \delta_{d(\Pi),2}) = 1 - \sum_{d=1}^{d^*} \delta_{d,1} + \delta_{d,2}$$

$\geq 1 - \sum_{d=1}^{d^*} N_d(\delta_{d,1} + \delta_{d,2}) \geq 1 - \sum_{d=1}^{d^*} N_d \left( \frac{\delta_{0.1}}{d^*} N^{-1}_d + \frac{\delta_{0.2}}{d^*} N^{-1}_d(\Pi_1) \right) = 1 - \sum_{d=1}^{d^*} \frac{1}{d^*} (\delta_{0.1} + \delta_{0.2}) = 1 - \delta_{0.1} - \delta_{0.2}$.
\end{proof}
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 $\Pi$, let $D(\Pi)$ be the set of all descendants of $\Pi$. Define the event $A = \bigcap_{\Pi \in D(\Pi)} \{ V_{\Pi} \in [L(\Pi^+), U(\Pi^+)] \}$. Then $P(A) \geq 1 - \delta_{0,1} - \delta_{0,2}$.

**Proof:** When BoundValueShare is called for a policy $\Pi$ with $d(\Pi) = d$, it is done so with parameters $\delta_{d,1} = \frac{\delta}{d^2} N_d^{-1}$ and $\delta_{d,2} = \frac{\delta}{d^2} N_d^{-1}$, where $\delta_{d,1}$ and $\delta_{d,2}$ are used to control the accuracy of the estimated value of $V_{\Pi}$ and $E\Psi(B_{\Pi})$ respectively. It follows from Proposition 2 that for any active policy $\Pi$, the probability that the interval $[V_{\Pi,m_1} - c_1, V_{\Pi,m_1} + \Psi(B_{\Pi})]_{m_2} + c_1 + c_2$ generated by BoundValueShare does not contain $V_{\Pi}^+$ is less than $\delta_{d,1} + \delta_{d,2}$. Furthermore, from standard Hoeffding bounds, the probability that $V_{\Pi}$ is outside the interval $[V_{\Pi} - c_1, V_{\Pi} + c_1]$ is less than $\delta_{d,1}$. Since any descendant policy $\Pi'$ of $\Pi$ consists of adding at least one item to the knapsack and item rewards are all $\geq 0$, 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 $P(A) > 1 - \sum_{d=1}^{d_{\Pi}}(\delta_{d,1} + \delta_{d,2})N_d = 1 - \delta_{0,1} - \delta_{0,2}$. □

The result of the proposition follows by noting that the true optimal policy $\Pi^{OPT}$ will be a descendant of $\Pi_i$ for some $i \in I$. Let $\Pi^*$ be the policy outputted by the algorithm. By the stopping criterion, $L(V_{\Pi_i}^+) + \epsilon \geq \max_{\Pi \in ACTIVE \setminus \{\Pi_i\}} U(V_{\Pi}^+)$ 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 $\Pi'$ of $\Pi^{OPT}$ 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 $\Pi$ 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, $U(V_{\Pi}^+) - L(V_{\Pi}^+) \leq \Psi(B_{\Pi})_{m_1} + c_2 + 2c_1 = U(\Psi(B_{\Pi})), 2c_1$. For a complete policy, $U(\Psi(B_{\Pi})) \leq \frac{\epsilon}{4}$ and according to BoundValueShare, $m_1$ is chosen such that $2c_1 \leq \frac{\epsilon}{4}$ which implies $U(V_{\Pi}^+) - L(V_{\Pi}^+) \leq \epsilon$.

If $\Pi$ 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

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

Hence,

$$U(V_{\Pi}^+) - L(V_{\Pi}^+) \leq U(\Psi(B_{\Pi})) + 2c_1$$

$$\leq 3U(\Psi(B_{\Pi})) \quad (8)$$

$$\leq 3(E\Psi(B_{\Pi}) + 2c_2)$$

$$\leq 3 \left( E\Psi(B_{\Pi}) + E\Psi(B_{\Pi}) - \frac{\epsilon}{4} \right) \quad (9)$$

$$\leq 6E\Psi(B_{\Pi}) - \frac{3}{4}\epsilon.$$

Where (8) follows since, when $L(\Psi(B_{\Pi})) \geq \frac{\epsilon}{4}$, we sample the value of policy $\Pi$ 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, $\Pi_t \in Q^*$ 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 \( \Pi^* \) in the active set, denote this policy by \( \Pi^*_t \). If \( \Pi_t = \Pi^*_t \), then by Lemma 14 \( V_{\Pi_t} \in [L(V^+_{\Pi_t}), U(V^+_{\Pi_t})] \). Hence,

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

Where the last inequality will hold for any incomplete policy (since for an incomplete policy \( L(B_{\Pi_t}) \geq \frac{\epsilon}{4} \)) and so, \( \Pi_t \in Q^\epsilon \). For \( \Pi_t = \Pi^*_t \), since \( \frac{3}{4}\epsilon \geq \epsilon \), \( \Pi_t \in Q^\epsilon \).

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

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

Then, if the termination criteria is not met,

\[
V_{\Pi_t} \geq L(V^+_{\Pi_t}) \implies V_{\Pi_t} + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon - \epsilon \geq L(V^+_{\Pi_t}) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon - \epsilon \\
\geq U(V^+_{\Pi_t}) - \epsilon \\
\geq \max_{\Pi \in \text{Active} \setminus \{\Pi_t\}} U(V^+_{\Pi_t}) - \epsilon \\
\geq L(V^+_{\Pi_t}) \\
\geq U(V^+_{\Pi_t}) - 6E\Psi(B_{\Pi_t}) + \frac{3}{4}\epsilon \\
\geq U(V^+_{\Pi_t}) - 6E\Psi(B_{\Pi_t}) + \frac{3}{4}\epsilon \\
\geq v^* - 6E\Psi(B_{\Pi_t}) + \frac{3}{4}\epsilon
\]

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

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

\[
L(V^+_{\Pi_t^{(1)}}) + \epsilon < \max_{\Pi \in \text{Active} \setminus \{\Pi_t^{(1)}\}} U(V^+_{\Pi_t}) = U(V^+_{\Pi_t^{(2)}}).
\]

By equation (10),

\[
L(V^+_{\Pi_t^{(1)}}) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4}\epsilon \geq U(V^+_{\Pi_t^{(1)}}).
\]
and by the selection criterion \( U(\Psi(B_{\Pi_1^{(2)}})) \geq U(\Psi(B_{\Pi_1^{(1)}})) \). Therefore, for \( \Pi_t = \Pi_1^{(2)} \neq \Pi_1^{(1)} \),
\[
V_{\Pi_t} + 12E\Psi(B_{\Pi_t}) - \frac{6}{4} \epsilon - \epsilon \geq L(V_{\Pi_t}^+) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4} \epsilon + 6E\Psi(B_{\Pi_t}) - \frac{3}{4} \epsilon - \epsilon
\]
\[
\geq U(V_{\Pi_t}^+) + 6E\Psi(B_{\Pi_t}) - \frac{3}{4} \epsilon - \epsilon
\]
\[
\geq U(V_{\Pi_t}^+) + 3U(\Psi(B_{\Pi_t})) - \epsilon
\]
\[
\geq U(V_{\Pi_t}^+) + 3U(\Psi(B_{\Pi_t^{(1)}})) - \epsilon
\]
\[
\geq L(V_{\Pi_t^{(1)}}^+) + 3U(\Psi(B_{\Pi_t^{(1)}}))
\]
\[
\geq U(V_{\Pi_t^{(1)}}^+)
\]
\[
\geq U(V_{\Pi_t}^+)
\]
\[
\geq v^*.
\]

Hence \( \Pi_t \in \mathcal{Q}^c \). \( \square \)

**Theorem 17** (Theorem 5 in main text) The total number of samples required by OpStoK is bounded from above by,
\[
\sum_{\Pi \in \mathcal{Q}^c} (m_1(\Pi) + m_2(\Pi)) d(\Pi),
\]
with probability \( 1 - \delta_{0.2} \).

**Proof:** The result follows from the following three lemmas.

**Lemma 18** For \( \Pi \in \mathcal{A}^c \) 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 \( \Pi \) are bounded by
\[
m_1(\Pi) = \left\lceil \frac{8\Psi(B)^2 \log(\frac{2}{\delta_{d,2}})}{\epsilon^2} \right\rceil \quad \text{and} \quad m_2(\Pi) = m^*,
\]
where \( m^* \) is the smallest integer satisfying
\[
\frac{16\Psi(B)^2}{(E\Psi(B))^{2/\gamma}} \leq \frac{m}{\log(\frac{2n}{m_2})} \quad \text{with} \quad n \text{ defined as in (2)}.
\]

**Proof:** When \( E\Psi(B_{\Pi_1}) \leq \frac{\epsilon}{2} \), the event \( \{U(\Psi(B_{\Pi_1})) \leq \frac{\epsilon}{2}\} \) 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 \( 2c_2 \leq \frac{\epsilon}{2} - E\Psi(B_{\Pi_1}) \), since by Hoeffding’s Inequality we know \( \Psi(B_{\Pi_1}) \leq m_2 \in [E\Psi(B_{\Pi_1}) - c_2, E\Psi(B_{\Pi_1}) + c_2] \) where \( c_2 \) is as defined in Lemma \( \square \). From this, it follows that \( U(\Psi(B_{\Pi_1})) \in [E\Psi(B_{\Pi_1}) - E\Psi(B_{\Pi_1}) + 2c_2] \). We want to make sure that \( U(\Psi(B_{\Pi_1})) \leq \frac{\epsilon}{2} \) will eventually happen so we need to construct a confidence interval such that \( c_2 \) satisfies \( E\Psi(B_{\Pi_1}) + 2c_2 \). Therefore we select \( m_2 \) such that,
\[
2c_2 \leq \frac{\epsilon}{2} - E\Psi(B_{\Pi_1})
\]
\[
\Rightarrow 4\Psi(B) \sqrt{\frac{2\log(\frac{8n}{m_2})}{m_2}} \leq \frac{\epsilon}{2} - E\Psi(B_{\Pi_1})
\]
\[
\Rightarrow \frac{16\Psi(B)^2}{(E\Psi(B_{\Pi_1}) - \gamma/2)^2} \leq \frac{m_2}{\log(4n/(m_2\gamma))}.
\]

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 \( U(\Psi(B_{\Pi_1})) \leq \frac{\epsilon}{2} \) with probability greater than \( 1 - \delta_{d,2} \).

When \( U(\Psi(B_{\Pi_1})) \leq \frac{\epsilon}{2} \), **BoundValueShare** requires \( m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log(\frac{2}{\delta_{d,2}})}{\epsilon^2} \right\rceil \) samples of the value of the policy to ensure \( 2c_1 \leq \frac{\epsilon}{2} \). \( \square \)
Lemma 19 For \( \Pi \in B^d \) 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 \( \Pi \) are bounded by

\[
m_1(\Pi) \leq \left\lceil \frac{\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{2E\Psi(B)\epsilon^2} \right\rceil \quad \text{and} \quad m_2(\Pi) = m^*,
\]

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

Proof: When \( E\Psi(B) \leq \frac{2}{7} \), by noting that the event \( \{E(\Psi(B)) \geq \frac{2}{7}\} \) will eventually happen and using a very similar argument to Lemma 18, it follows that \( m_2(\Pi) \) is the smallest integer solution to

\[
\frac{16\Psi(B)^2}{(E\Psi(B) - \epsilon/4)^2} \leq \frac{m}{\log(n/m \delta_2)},
\]

with probability greater than \( 1 - \delta_{d,2} \). Whenever \( L(\Psi(B)) > \frac{2}{7} \), \( \text{BoundValueShare} \) requires \( m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{(U(\Psi(B)))^2} \right\rceil \) samples of the value of policy \( \Pi \). Since \( U(\Psi(B)) \in [E\Psi(B) + 2c_2] \) with probability \( 1 - \delta_{d,2} \), \( U(\Psi(B)) \geq E\Psi(B) \), and so,

\[
m_1(\Pi) = \left\lceil \frac{2\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{(U(\Psi(B)))^2} \right\rceil \leq \left\lceil \frac{2\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{E\Psi(B)\epsilon^2} \right\rceil
\]

and the result holds.

\( \square \)

Lemma 20 For \( \Pi \in C^d \) 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 \( \Pi \) are bounded by

\[
m_1(\Pi) \leq \max \left\{ \left\lceil \frac{8\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{\epsilon^2} \right\rceil, \frac{\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{2E\Psi(B)\epsilon^2} \right\}
\]

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

Proof: When \( \frac{2}{7} < E\Psi(B) < \frac{2}{5} \), 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 \( \frac{16\Psi(B)^2}{(\epsilon/4)^2} \leq \frac{m}{\log(n/m \delta_2)} \).

In order to determine the number of samples of the value required by \( \text{BoundValueShare} \), we need to know which of \( \{U(\Psi(B)) \leq \frac{2}{7}\} \) or \( \{L(\Psi(B)) > \frac{2}{7}\} \) occurs first. However, when \( \Pi \in C^d \), we do not know this so the best we can do is bound \( m_1(\Pi) \) by the maximum of the two alternatives,

\[
m_1(\Pi) \leq \max \left\{ \left\lceil \frac{2\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{\epsilon^2} \right\rceil, \frac{2\Psi(B)^2 \log\left(\frac{2}{8d \epsilon_1}\right)}{E\Psi(B)\epsilon^2} \right\}
\]

\( \square \)

The result of the theorem then follows by noting that for any policy \( \Pi \) 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 \( \text{OpStoK} \) will be in \( Q^s \) so it suffices to sum over all policies in \( Q^s \). This result assumes that all confidence bounds hold, whereas we know that for any policy \( \Pi \) 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 \( |Q^s| \leq |P| \), the probability should be considerably greater than \( 1 - \delta_{0,2} \). \( \square \)