Appendix
Proof of Theorem 1

We now formulate and prove a series of claims and propositions towards proving Theorem 1. To simplify the proofs, we assume that all initial states of all the b-threads are not labeled as *must-finish*:

Assumption 1. $L(init) = 0$.

As demonstrated with our examples in Section II and Section VIII, this assumption is reasonable in practice. It does not restrict generality since adding an extra initial state that is not labeled as *must-finish* is always possible.

The first claim says that the accumulated rewards over a finite prefix of a run are either 0 or -1:

Claim 1. For every infinite b-program run $l = s^0 \xrightarrow{e^0} s^1 \xrightarrow{e^1} \cdots$ and time $t \geq 0$:

$$\sum_{k=0}^{t} R(s^{k-1}, e^{k-1}, s^k) = \begin{cases} 0 & \text{if } L(s^t) = 0; \\ -1 & \text{otherwise.} \end{cases}$$

Proof. By induction on t. The base case is given by Assumption 1. Assuming that the claim is true for $t-1$, If $L(s^{t-1}) = L(s^t)$ then $R(s^{t-1}, e^{t-1}, s^t) = 0$ and the claim follows. If $L(s^{t-1}) = 0$ and $L(s^t) = 1$ then $R(s^{t-1}, e^{t-1}, s^t) = -1$ and we get that

$$\sum_{k=0}^{t} R(s^{k-1}, e^{k-1}, s^k) = \sum_{k=0}^{t-1} R(s^{k-1}, e^{k-1}, s^k) - 1 = -1.$$

If $L(s_i^{t-1}) = 1$ and $L(s_i^t) = 0$ then $R(s^{t-1}, e^{t-1}, s^t) = 1$ and we get that

$$\sum_{k=0}^{t} R(s^{k-1}, e^{k-1}, s^k) = \sum_{k=0}^{t-1} R(s^{k-1}, e^{k-1}, s^k) + 1 = 0.$$

Hence, all cases are consistent with the claimed equation. □

We next show that the sequence of rewards is a repetition of the form: $0, \ldots, 0, -1, 0, \ldots, 0, 1$ where $0, \ldots, 0$ means, possibly empty, sequence of 0s. There may be an infinite tail of zeroes at the end, or the alternation can go forever.

Claim 2. For every infinite b-program run $l = s^0 \xrightarrow{e^0} s^1 \xrightarrow{e^1} \cdots$, let $(t_k)_{k=0}^{n}$ be the sequence of times where $R(s^{t_k}, e^{t_k}, s^{t_k+1}) \neq 0$. The length of the sequence can be finite, infinite, or empty, i.e., $n \in \mathbb{N} \cup \{\infty, -1\}$. Then for every $0 \leq k \leq n$: $R(s^{t_k}, e^{t_k}, s^{t_k+1}) = (-1)^{k+1}$.

Proof. Based on Definition 6 and Assumption 1, it is clear from that $R(s^{t_0}, e^{t_0}, s^{t_0+1}) = -1$. Assume towards contradiction that there is $k \geq 0$ such that $R(s^{t_k}, e^{t_k}, s^{t_k+1}) = R(s^{t_{k+1}}, e^{t_{k+1}}, s^{t_{k+1}+1})$. Since $R(s^t, e^t, s^{t+1}) = 0$ for every $t_k < t < t_{k+1}$, by the same definition, $L(s^{t_k+1}) = L(s^{t_{k+1}})$. This contradicts the definition of R where it is apparent that $L(s^{t_k+1}) = L(s^{t_{k+1}})$ implies $R(s^{t_k}, e^{t_k}, s^{t_k+1}) \neq R(s^{t_{k+1}}, e^{t_{k+1}}, s^{t_{k+1}+1})$. □

Using the above observation regarding the alternation of the sequence, we obtain a lower bound for the residual discounted accumulated reward of live runs based on the current state's label:

Claim 3. For every infinite live b-program run $l = s^0 \xrightarrow{e^0} s^1 \xrightarrow{e^1} \cdots$, time $t \geq 0$, and $\gamma < 1$:

$$\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) > \begin{cases} -1 & \text{if } L(s^t) = 0; \\ 0 & \text{if } L(s^t) = 1. \end{cases}$$

Proof. Similarly to the sequence used in Claim 2, let $(q_k)_{k=0}^{n_q}$ be the sequence of times after t where $R(s^{q_k}, e^{q_k}, s^{q_k+1}) = 1$, and $(r_k)_{k=0}^{n_r}$ be the sequence of times after t where $R(s^{r_k}, e^{r_k}, s^{r_k+1}) = -1$. Note that since run l is live, we have that $n_q \geq n_r$; otherwise, the run ends with infinitely many b-program's must-finish states. If both sequences are empty, it is clear from Definition 6 and Assumption 1 that $L(s_i^t) = 0$. In this case, all rewards are zero, and the claim holds trivially. Furthermore, if $L(s^t) = 1$ and $n_r < 0$, then $(q_k)_{k=0}^{n_q}$ is not empty, i.e., $n_q \geq 0$ or else the run ends with infinitely many must-finish states. In this case, we get that

$$\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) = \sum_{k=0}^{n_q} \gamma^{q_k} > 0.$$

If the sequences are not empty and $L(s_i^t) = 1$, from Claim 2 we get that $q_k < r_k$ for each $k \leq n_r$ and then

$$\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) \geq \sum_{k=0}^{n_r} (\gamma^{q_k} - \gamma^{r_k}) > 0.$$

If the sequences are not empty and $L(s_i^t) = 0$, from Claim 2 we get that $r_k < q_k < r_{k+1}$ for each $k < n_r - 1$ and

$$\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) \geq -\gamma^{r_0} + \sum_{k=0}^{n_r-1} (\gamma^{q_k} - \gamma^{r_{k+1}}) > -1.$$

Thus, the claim holds for all cases. □

In the opposite direction to the previous claim, we also show that if the optimal policy can achieve a positive residual discounted accumulated reward, it is possible to get to a state that is not labeled as *must-finish* (we will later use that to construct a live run).

Claim 4. If $Q^*(s^t, e^t) > 0$ then there is a path $s^t \xrightarrow{e^t} s^{t+1} \xrightarrow{e^{t+1}} \cdots \xrightarrow{e^{t+m_t-1}} s^{t+m_t}$ such that $L(s^{t+m_t}) = 0$.

Proof. Using the optimal policy π^*, we construct a path by defining $e^{t'} = \pi^*(s^{t'})$ for every $t' > t$, and choosing $s^{t'+1}$ to be the only state such that $P(s^{t'}, e^{t'}, s^{t'+1}) = 1$. There is only one such state since the b-program transitions (as defined in Definition 3) are deterministic. Assume, towards contradiction, that $L(s^{t'}) = 1$ for every $t' \geq t$. Then $Q^*(s^t, e^t) = \sum_{t'=t}^{\infty} \gamma^{t'} R(s^{t'}, e^{t'}, s^{t'+1}) = 0$, which contradicts the assumption. This gives us that the path that we have constructed is as required. □

We are now ready to state and prove the two propositions that establish the correctness of our approach, starting with showing that an execution mechanism that generates all Q^* compatible runs is complete in the sense that it generates all possible live runs:

Proposition 1. A live b-program run is Q^*-compatible.

Proof. Let $l = s^0 \xrightarrow{e^0} s^1 \xrightarrow{e^1} \cdots$ be a live run. To prove that l is Q^*-compatible we now show that the term in Definition 7 holds for every time t.

If $L(s^t) = 0$, from Claim 1 we get that $\sum_{k=0}^{t} R(s^k, e^k, s^{k+1}) = 0$ and, as shown in Claim 3 $\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) > -1$.

If $L(s^t) = 1$, from Claim 1 we get that $\sum_{k=0}^{t} R(s^k, e^k, s^{k+1}) = -1$ and, as shown in Claim 3 $\sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) > 0$.

In both cases, when adding the terms together, we get that for every time t $\sum_{k=0}^{t} R(s^k, e^k, s^{k+1}) + \sum_{k=t}^{\infty} \gamma^k R(s^k, e^k, s^{k+1}) > -1$.

By the definition of Q^*, the optimal-value function, $\sum_{k=0}^{t} R(s^k, e^k, s^{k+1}) + Q^*(s^t, e^t) > -1$.

We get that run l is Q^*-compatible by Definition 7. $\square$

Second, we show that an execution that generates Q^* compatible runs according to the distribution defined in Definition 8 is sound in the sense that it generates live runs with probability one.

Proposition 2. A Q^*-compatible b-program run is almost surely live.

Proof. Let π be the policy defined in Definition 8 using the optimal value function Q^*. We will show that π generates a live run with probability one. Since, by definition, π always draws a Q^* compatible runs, we have $\sum_{k=0}^{t-1} R(s^k, e^k, s^{k+1}) + Q^*(s^t, e^t) > -1$ for all $t \geq 0$. To generate a non-live run, π needs from some point of time, t_0, to always visit states that are labeled as *must-finish*, i.e., $L(s^t) = 1$ for all $t > t_0$. By Claim 1, for all $t > t_0$, $\sum_{k=0}^{t-1} R(s^k, e^k, s^{k+1}) = -1$ and we get that $Q^*(s^t, e^t) > 0$.

Therefore, by Claim 4 for every $t > t_0$ there is a path $s^t \xrightarrow{\hat{e}^t} \hat{s}^{t+1} \xrightarrow{\hat{e}^{t+1}} \cdots \xrightarrow{\hat{e}^{t+m_{t-1}}} \hat{s}^{t+m_t}$ such that $L(\hat{s}^{t+m_t}) = 0$. Assume that this is the first such index, i.e., $L(\hat{s}^{t+m_t-1}) = 1$ and we get that $\sum_{k=t}^{t'-1} R(\hat{s}^k, \hat{e}^k, \hat{s}^{k+1}) = 0$ for every $t \leq t' < t + m_t - 1$. This means that all the states along this path satisfy

$$\sum_{k=0}^{t-1} R(s^k, e^k, s^{k+1}) + \sum_{k=t}^{t'-1} R(\hat{s}^k, \hat{e}^k, \hat{s}^{k+1}) + Q^*(\hat{s}^{t'}, \hat{e}^{t'}) > -1$$

and that π could have chosen this path. The probability that it will not choose any of these paths is zero. $\square$

Finally, we get that Theorem 1 holds, a live b-program run is Q^*-compatible, and a Q^*-compatible b-program run is almost surely live:

Proof of Theorem 1. Follows by Proposition 1 and Proposition 2 above. $\square$