where $P \left( s \rightarrow s^* \text{ transition} \middle| \text{Goal was reached from } s \text{ via optimal policy} \right) = T(s,a^*,s')P^*(s')/P^*(s)$.

These equations essentially say that if $a^*$ was executed in $s$ and, as a result of following an optimal policy $\pi^*$ the goal was reached, then with probability $T(s,a^*,s')P^*(s')/P^*(s)$ action $a^*$ must have caused a transition from $s$ to $s_1$, with probability $T(s,a^*,s_2')P^*(s_2')$ it must have caused a transition to $s_2$, and so on. This means that if we want to find the vector $[J^*|P^*]$ of expected costs of goal-reaching trajectories under $\pi^*$, then it is enough to find the optimal cost function of MDP $M_{P^*} = \langle S_{P^*}, A_{P^*}, T_{P^*}, \mathcal{C}_{P^*}, \mathcal{G}_{P^*}, s_0^{P^*} \rangle$, where $\mathcal{G}_{P^*}$ and $s_0^{P^*}$ (if known) are the same as $\mathcal{G}$ and $s_0$ for the iSSP MDP that we are trying to solve; $S_{P^*}$ is the same as $S$ for $M$ but does not include dead ends, i.e., states $s$ for which $P^*(s) = 0$; $A_{P^*} = \cup_{s \in S} A_{P^*}(s)$, i.e., the set of actions consists of all $P^*$-greedy actions in each state; for each $a^* \in A_{P^*}$, $T_{P^*}(s,a^*,s') = T(s,a^*,s')P^*(s')$, as above, and $a^*$ is “applicable” only in $s$; and $\mathcal{G}_{P^*}(s,a)$ is the same as $\mathcal{C}$ for $M$, except it is defined only for $a \in A_{P^*}$.

As it turns out, we already know how to solve MDPs such as $M_{P^*}$:

**Theorem 4.** For an iSSP MDP $M$ with $P^*(s_0) > 0$, MDP $M_{P^*}$ constructed from $M$ as above is an SSP MDP.

**Proof sketch.** Indeed, $M_{P^*}$ is “almost” like the original iSSP MDP, but has at least one proper policy because, by construction, it has no dead ends.

Now, as we know [Bertsekas, 1995], $J^*$ for the SSP $M_{P^*}$ satisfies $J^*(s) = \min_{a \in A_{P^*}} \mathcal{C}_{P^*}(s,a) + \sum_{s' \in S} T_{P^*}(s,a,s')J^*(s')$. Therefore, by plugging in $T_{P^*}(s,a,s')P^*(s')$ in place of $T_{P^*}(s,a,s')$ and $[J^*|P^*]$ in place of $J^*$, we can state the following theorem for the original iSSP MDP $M$:

**Theorem 5.** For an iSSP MDP with the optimal goal-probability function $P^*$, the optimal cost function $[J^*|P^*]$ characterizing the minimum expected cost of trajectories that reach the goal satisfies

$$[J^*|P^*](s) = 0 \; \forall s \text{ s.t. } P^*(s) = 0$$

$$[J^*|P^*](s) = \min_{a \in A_{P^*}} \left\{ \mathcal{C}(s,a) + \sum_{s' \in S} \frac{T(s,a,s')P^*(s')}{P^*(s)} [J^*|P^*](s') \right\}$$

**Putting It All Together.** Our construction not only let us derive the optimality equation for $[J^*|P^*]$, but also implies that $[J^*|P^*]$ can be found via VI, as in the case of SSP MDPs [Bertsekas, 1995], over $P^*$-optimal actions and non-dead-end states. Moreover, since the optimal policy for an SSP MDP is greedy w.r.t. the optimal cost function and solving an iSSP MDP ultimately reduces to solving an SSP, the following important result holds:

**Theorem 6.** For every iSSP MDP, there exists a Markovian deterministic policy $\pi^*$ that can be derived from $P^*$ and $[J^*|P^*]$ for non-dead-end states using

$$\pi^*(s) = \arg \min_{a \in A_{P^*}} \left\{ \mathcal{C}(s,a) + \sum_{s' \in S} \frac{T(s,a,s')P^*(s')}{P^*(s)} [J^*|P^*](s') \right\}$$

Combining optimality equations 11 and 12 for $P^*$ and $[J^*|P^*]$ respectively with Equation 13, we present a VI-based algorithm for solving iSSP MDPs, called IVI (Infinite-penalty Value Iteration) in Algorithm 1.

**Algorithm 1: IVI**

**Input:** iSSP MDP $M$

**Output:** Optimal policy $\pi^*$ for non-dead-end states of $M$

1. Find $P^*$ using arbitrarily initialized VI$_{MP}$.
2. Find $[J^*|P^*]$ using arbitrarily initialized VI$_{SSP}$ over $M_{P^*}$ with update equations 12

Return $\pi^*$ derived from $P^*$ and $[J^*|P^*]$ via Equation 13

**Algorithm 2: SHS**

**Input:** iSSP MDP $M$

**Output:** Optimal policy $\pi^*_{ss}$ for non-dead-end states of $M$ rooted at $s_0$

1. Find $P^*_{s_0}$ using FRET initialized with an admissible heuristic $\hat{P} \geq P^*$
2. Find $[J^*|P^*]_{s_0}$ using FIND-AND-REVISE over $M_{P^*}$ with optimality equations 12, initialized with an admissible heuristic $\hat{J} \leq [J^*|P^*]$

Return $\pi^*_{s_0}$ derived from $P^*_{s_0}$ and $[J^*|P^*]_{s_0}$ via Equation 13

As we established, solving an iSSP MDP with VI is a two-stage process, whose first stage solves a MAXPROB MDP and whose second stage solves an SSP MDP. In the Background section we mentioned that both of these kinds of MDPs can be solved with heuristic search; MAXPROB — with the FRET framework, and SSP — with the FIND-AND-REVISE framework. This allows us to construct a heuristic search schema called SHS (Staged Heuristic Search) for iSSP MDPs, presented in Algorithm 2.

There are two major differences between Algorithms 1 and 2. The first one is that SHS produces functions $P^*_{s_0}$ and $[J^*|P^*]_{s_0}$ that are guaranteed to be optimal only over the states visited by some optimal policy $\pi_{s_0}^*$ starting from the initial state $s_0$. Accordingly, the SHS-produced policy $\pi_{s_0}^*$ specifies actions only for these states and does not prescribe any for other states. Second, SHS requires two admissible heuristics to find an optimal (partial) policy, one ($\hat{P}$) being an upper bound on $P^*$ and the other ($\hat{J}$) being a lower bound on $[J^*|P^*]$. 
