Planning with Participation Constraints

Abstract

We pose and study the problem of planning in Markov decision processes (MDPs), subject to participation constraints as studied in mechanism design. In this problem, a planner must work with a self-interested agent on a given MDP. Each action in the MDP provides an immediate reward to the planner and a (possibly different) reward to the agent. The agent has no control in choosing the actions, but has the option to end the entire process at any time. The goal of the planner is to find a policy that maximizes her cumulative reward, taking into consideration the agent’s ability to terminate.

We give a fully polynomial-time approximation scheme for this problem. En route, we present polynomial-time algorithms for computing (exact) optimal policies for important special cases of this problem, including when the time horizon is constant, or when the MDP exhibits a “definitive decisions” property. We illustrate our algorithms with two different game-theoretic applications: the problem of assigning rides in ride-sharing and the problem of designing screening policies. Our results imply efficient algorithms for computing (approximately) optimal policies in both applications.

1 Introduction

In this paper, we consider the general problem of planning in a Markov decision process (MDP), when the actions are chosen by the planner, but must be taken by an agent that is separate from the planner. The agent cannot choose the actions, but can refuse to participate at any time. To illustrate the setting, consider the following scenario. A ride-hailing company is designing a system to assign tasks to drivers. The company monitors the location (or state) of each driver in real-time via a mobile app. The company then has to choose which ride to assign to a driver (the company’s choice of action), with each possible assignment resulting in some value to the company. If drivers had no incentives of their own (e.g., were self-driving vehicles belonging to the company), this could be modeled straightforwardly as an MDP. However, the driver also obtains rewards from the rides, which may be different. For example, drivers generally prefer not to take rides that will lead them to isolated areas, or that are likely to involve troublesome passengers (e.g., pick-up at a bar at closing time). The company still wants to assign these rides, but does not want the driver to become sufficiently unhappy that they sign off (stop participating) altogether.

It is therefore crucial to take into consideration drivers’ interests in planning task assignments, and provide enough incentives for drivers to remain in the system even when they receive less profitable or even unprofitable tasks. One intuitive solution is to promise drivers who receive undesirable tasks high priority for receiving more profitable tasks (or other perks) when they finish the current task. However, implementing this solution in a way that is most beneficial to the company can be challenging — incentivizing drivers and maximizing profit can be in conflict with each other, and it is not immediately clear how to balance them optimally.

In fact, the company faces a planning problem with participation constraints: the company as the planner (henceforth the principal) wants to compute and commit to a policy that maximizes her cumulative reward, subject to the constraint that each driver (henceforth the agent) must be motivated to participate in the policy, and not just at the beginning but also from any point onward.

In addition to ride-hailing, many other real-world scenarios can be modeled as planning with participation constraints. We provide a few such examples:

- Financial services, such as automated investment platforms (“robo-advisors”) or traditional mutual funds. In the case of robo-advisors, when a customer puts in money, the platform automatically creates and maintains a portfolio for the customer, and trades on behalf of the customer without asking for confirmation. The customer can see the trades and has the option to withdraw their money (i.e., to quit) at any time. The customer wants to make as much money as possible in a given amount of time, while the platform wants to keep the customer and optimize their own objective (which partially aligns with the customer’s goal, but may also include other considerations, such as charging an advisor fee).

2Restricting the focus to policies under which no agent ever drops out is without loss of generality by a revelation principle: if the optimal policy involves an agent quitting at some point, since the principal knows when the agent would quit, she can use an equivalent policy where the agent is told that he will not have to do anything else (e.g., take any further rides) at that point.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

有时，司机可以看见目的地；有时他们不能但仍然怀疑目的地基于出发点。
• Distributed computing platforms (e.g., MQL 5, Load-
Team, or Golem Network) where an agent can sell CPU
hours. Typically, such platforms assign tasks to machines
based on how much task providers are willing to pay and
how much computing power is required. Agents are paid
an amount of money that depends on the task when a task
is finished. The agents want to make more money, while
the platform may want to finish as many tasks as possi-
ble. If the agent is receiving too little payment, they may
decide to quit the platform and make use of their CPU
hours in other ways (e.g., mining cryptocurrencies).

• Crowdsourcing marketplaces (e.g., Amazon Mechanical
Turk). Workers are paid after they finish a task (e.g., com-
pleting surveys, labeling images). Workers can quit the
platform or switch to another platform if they are not
making enough money relatively. The platform wants to
assign tasks (or a list of possible tasks to choose from)
to workers and decide how much to pay to retain the work-
ers, while optimizing their own objectives at the same
time (e.g., charging a commission fee).

In this paper, we pose and study the computational prob-
lem of finding an (approximately) optimal policy in MDPs
subject to participation constraints. We focus on the setting
where agents have no private information, because in all the
aforementioned applications, we have a very good first-order
approximation of the agent’s reward function. For example,
ride-sharing drivers want to make more money in a given
amount of time, and applicants going through a screening
process would like to pass. (See Section 1.2 for more dis-
cussion on related work that considers private information.)

1.1 Our Results

Our main result is a fully polynomial-time approxima-
tion scheme (FPTAS) for planning with participation con-
straints.

• We can compute a policy satisfying participation
constraints and loses at most an additive \( \varepsilon > 0 \)
in the principal’s utility compared to the optimal
policy. Moreover, we can compute such a policy
in time \( \text{poly}(m, n, 1/\varepsilon) \), where \( m \) is the total
number of actions and \( n \) is the number of states.

Theorem 4 is a formal version of the above claim. We ob-
tain this FPTAS by first designing an exponential-time algo-
rithm (Algorithm 1 in Section 3.1) which computes an exact
optimal policy, and then carefully discretizing the algorithm
without violating participation constraints (Section 3.2).

When the MDP has some additional structures, our exact
algorithm (Algorithm 1) in fact runs in polynomial time.

• There is an exact algorithm (Algorithm 1) for
planning with participation constraints that runs
in time \( \text{poly}(m^* n) \), where \( m^* \) is the maximum
number of actions available at a state and \( n \) is the number of states.

• When the time horizon is \( H \), the exact algorithm
runs in time \( \text{poly}(m^* H, n^H) \).

• When the MDP has a “definitive decisions”
property (see Definition 5 in Section 3.3), the exact
algorithm runs in time \( \text{poly}(m, n) \), where \( m \) is total number of actions.

The formal statements of these results appear as Lemma 3
in Section 3.1 and Lemma 6 in Section 3.3.

Some applications, such as the ride-hailing model that we
gave above, are straightforward to fit into our framework.
Consequently, the above results immediately imply efficient
approximation algorithms for the ride assignment problem
in ride-hailing, as well as efficient exact algorithms when
the instance admits certain additional structures.

Our algorithmic framework in fact allows more and some-
what surprising applications. To illustrate this, we consider
the design of optimal screening policies (Section 4), where
the goal is to separate “good” candidates from “bad” ones
by observing their performance in a series of tests. The more
outcomes we observe, the more confident we are — but can-
didates may drop out if we ask them to take so many tests
that the cost of taking tests outweighs the expected utility
from passing the screening. By casting the problem as plan-
ning with participation constraints and applying our algo-
rithms, we obtain the following results (formal statements in
Theorem 8) for this problem.

• We can compute an optimal screening policy that
uses at most \( N \) tests in time \( \text{poly}(N) \). Moreover,
an approximately optimal policy losing at most
an additive \( \varepsilon > 0 \) in the principal’s utility can be
computed in time \( \text{poly}(1/\varepsilon) \).

1.2 Related Work

Our results are along the line of research on planning in
Markov decision processes (MDPs). The basic problem of
computing optimal policies in MDPs has been extensively
studied, and numerous efficient methods have been pro-
posed, e.g., (Bellman 1957; Howard 1960; Puterman and
Shin 1978). The problem studied in this paper is closely re-
lated to constrained MDPs (CMDPs) (see (Altman 1999) for
a comprehensive survey), where the goal is to find a reward-
maximizing policy subject to overall constraints, such as
that the expected cumulative cost of the policy must not ex-
ceed a predefined threshold. It is known that CMDPs can be
solved using linear programming (Altman and Spieksma
1995; Altman 1996, 1998). The key difference between our
model and CMDPs is that we consider uniform participa-
tion constraints, which must hold in all states throughout
the execution of the policy, rather than just in the initial state.
In particular, such participation constraints may require any
(nearly) optimal policy to be history-dependent, whereas
in CMDPs, one can focus on Markovian (i.e., history-
independent) policies without loss of generality (Altman
1999). More recently, reinforcement learning in CMDPs has
received increasing attention (Achiam et al. 2017; Tessler,
Mankowitz, and Mannor 2018; Cheung 2019; Le, Voloshin,
and Yue 2019; Brantley et al. 2020; Efroni, Mannor, and
Pirotta 2020; Singh, Gupta, and Shroff 2020; Ding et al.
2021). These results are quite different from ours, because
they tend to focus on exploration rather than the planning
problem itself, and the model we study in this paper is differ-
ent from CMDPs. A related model is multi-objective MDPs
(MOMDPs) (see (Rojiers et al. 2013) for a comprehensive
survey), where the planner cares about multiple objectives
of actions $A_s$, where $A_{s_{\text{term}}} = \emptyset$, i.e., no actions can be played at the terminal state. The principal’s and the agent’s rewards are captured by two reward functions $r_P : S \times A \to \mathbb{R}$ and $r_A : S \times A \to \mathbb{R}$. That is, in state $s \in S$, taking action $a \in A_s$ gives the principal a reward of $r_P(s, a)$ and the agent a reward of $r_A(s, a)$.

When action $a \in A_s$ is played in state $s \in S$, the probabilistic transition to the next state is given by a distribution $P(s, a) \in \Delta(S)$ over the state space, where $P(s, a, s')$ is the probability that the next state is $s' \in S$. Throughout the paper, we assume the transition operator $P$ is acyclic, meaning that the probability of reaching the same state twice is 0 no matter what actions are played. In other words, if a state $s$ can reach another state $s'$ with positive probability, $s'$ cannot reach $s$. Consequently, we can label the states in topological order. That is, $S = [n]$ where $s_{\text{init}} = 1$ and $s_{\text{term}} = n$, such that all transitions go from a state with smaller index to one with larger index.

**Histories and policies.** A history of length $t$ is a sequence of interleaved states and actions $(s_1, a_1, s_2, a_2, \ldots, s_t, a_t)$, where $s_i \in S$ and $a_i \in A_i$ for each $i \in [t]$, and $s_i < s_{i+1}$ for each $i \in [t-1]$. Such a history corresponds to a (partial) run of the MDP, in which the initial state is $s_1$ and action $a_1$ is played, and the second state is $s_2$ and action $a_2$ is played, etc. Let $\mathcal{H}_t$ be the set of all histories of length $t$, and $\mathcal{H}_0 = \{\emptyset\}$ where $\emptyset$ denotes the empty history. Let $\mathcal{H} = \bigcup_{t \in [n-1]} \mathcal{H}_t \cup \mathcal{H}_0$ be the set of all histories, and let

$$\mathcal{H}_{<s} = \{(s_1, a_1, \ldots, s_t, a_t) \in \mathcal{H} \mid s_t < s\}$$

be the set of histories visiting only states before $s$. For a history $h = (s_1, a_1, \ldots, s_t, a_t)$ and a state-action pair $(s, a)$, let $h + (s, a)$ be the history given by appending $(s, a)$ to $h$:

$$h + (s, a) = (s_1, a_1, \ldots, s_t, a_t, s, a).$$

We say a history $h'$ extends another history $h$, if $h$ is a prefix of $h'$, with the notation $h' \supset h$.

A policy $\pi$ is a (probabilistic) mapping from history-state pairs to distributions over actions. For any $h \in \mathcal{H}, s \in S$, and $a \in A_s$, $\pi(h, s, a)$ is the probability that action $a$ will be chosen when state $s$ is reached with history $h$. Note that we allow policies to be randomized and history-dependent, which is necessary when we later consider participation constraints. We always require that $\pi(h, s, a) = 0$ if $a \notin A_s$.

Fixing a policy $\pi$, the utility (or value) function of the principal $u^*_P : \mathcal{H} \times S \to \mathbb{R}$ can be defined recursively.

$$u^*_P(h, s) = \sum_{a \in S_s} \pi(h, s, a) \left(r_P(s, a) + \sum_{s' \in S} P(s, a, s') \cdot u^*_P(h + (s, a), s')\right),$$

and $u^*_P(h, s) = 0$ if $s = s_{\text{term}}$. Note that $u^*_P(h, s)$ is well-defined given that transitions are acyclic. In particular, to having a distribution of initial states and multiple terminal states.

In our model, it is without loss of generality, i.e., equivalent
Algorithm 1: Algorithm for computing the Pareto frontier of the principal/agent’s utility for all states.

**Input:** State space $S$, sets of actions $\{A\}_s$, reward functions $r_P$ and $r_A$, transition operator $P$.

**Output:** Pareto frontier functions $\{f_s\}_s$ for all states.

1. let $f_n = \{(0,0)\}$;
2. for $s = n-1, \ldots, 1$ do
3.   initialize $f_s = \emptyset$;
4.   for $a \in A_s$ do
5.     call Algorithm 2 and let $g_{s,a} \leftarrow \text{subcurve}(s,a)$;
6.     /* $g_{s,a}$ is the Pareto frontier after taking action $a$ in state $s$ */
7.     let $g_{s,a} \leftarrow \{(x + r_A(s,a), y + r_P(s,a)) | (x, y) \in g_{s,a}\}$;
8.     /* shift $g_{s,a}$ by adding $(r_A(s,a), r_P(s,a))$ to all its turning points */
9.     return $\{f_s\}_s$;
10. /* Algorithm 3 requires additional bookkeeping information from Algorithms 1 and 2; for ease of presentation, we do not explicitly return it */

### 3.1 An Exact Algorithm

**Overview of Algorithm 1.** Algorithm 1 has the same high-level structure as backward induction in standard (finite-horizon) MDPs. Recall that states are labeled in topological order. Without participation constraints, we can simply process the states backward, and recursively compute the optimal onward utility and the corresponding sub-policy in every state. However, in our problem, the algorithm has to keep track of the cumulative reward of the principal and the agent, and enforce the participation constraints.

We compute the *Pareto frontier* of the principal’s and the agent’s utility at every state. The Pareto frontier at state $s$ is a function $f_s$ such that $f_s(u)$ is the principal’s maximum achievable onward utility, subject to the agent’s onward utility is exactly $u$ (while satisfying onward participation constraints). We show that the functions $\{f_s\}_s$ can be computed efficiently in a recursive way, and consequently, the principal’s optimal overall utility can be read-off from the Pareto frontier in the initial state, and is equal to $\max_{u \geq 0} f_{\text{max}}(u)$.

Because we consider finite-horizon environments, no discount factor is required; if a discount factor $\gamma$ is nevertheless desired, we can add this into our model by introducing an additional probability of $1 - \gamma$ of immediately transitioning to the terminal state.

### 3. Computing Optimal Policies with Participation Constraints

In this section we present our main result: an algorithm that computes an $\varepsilon$-approximate optimal policy with participation constraints in time $\text{poly}(n, m, 1/\varepsilon)$, where $n = |S|$ is the number of states, and $m = \sum_s |A_s|$ is the total number of actions. We first present an exact algorithm, which runs in exponential time in the worst case, and then discretize the algorithm to allow for efficient approximation. Moreover, we show that the exact algorithm itself is strongly polynomial time if the input MDP satisfies a “definitive decisions” property (Definition 5).

$u^\pi_P(\emptyset, s_{\text{init}})$ is the principal’s expected cumulative reward under policy $\pi$.

Similarly, the agent’s utility function $u^\pi_A$ under policy $\pi$ can be defined as

$$u^\pi_A(h, s) = \sum_{a \in A_s} \pi(h, s, a) \left( r_A(s, a) + \sum_{s' \in S} P(s, a, s') \cdot u^\pi_A(h + (s, a), s') \right),$$

and $u^\pi_A(h, s) = 0$ if $s = s_{\text{term}}$.

**The planning problem.** We take the principal’s point of view and aim to find a policy $\pi$ that maximizes the principal’s expected total utility $u^\pi_P(\emptyset, s_{\text{init}})$. Unlike planning in standard MDPs, we have the additional constraint that the policy should always incentivize the agent to remain in the process. That is, we only consider policies $\pi$ that satisfy:

$$u^\pi_A(h, s) \geq 0, \quad \text{for all } h \in H \text{ and } s \in S.$$

This condition is commonly known as *individual-rationality* (IR) (or voluntary participation, or just participation) in economic theory, and capture the intuition that the agent should not be worse off by participating in the process. It is worth noting that, while our model as defined above only allows the agent to quit *before* an action is chosen (i.e., ex-ante IR), it is easy to extend the model to allow the agent to quit *after* seeing the randomly selected action (i.e., ex-post IR). This can be done by inserting a dummy state to represent “ready to take action $a$ in state $s$”.

We remark that it is sometimes better for the principal to let the agent quit, if the cost of keeping the agent in the system is too high. We capture this by adding a “quit” action for the principal in each non-terminal state, which gives both parties a reward of 0 and transitions the state of the world to $s_{\text{term}}$. Having added this action, imposing the IR constraint is without loss of generality: if our solution relied on the agent at some point refusing to participate further, we could instead just select the “quit” action at that point which the agent would be happy to do so.

**3.1 An Exact Algorithm**

**Overview of Algorithm 1.** Algorithm 1 has the same high-level structure as backward induction in standard (finite-horizon) MDPs. Recall that states are labeled in topological order. Without participation constraints, we can simply process the states backward, and recursively compute the optimal onward utility and the corresponding sub-policy in every state. However, in our problem, the algorithm has to keep track of the cumulative reward of the principal and the agent, and enforce the participation constraints.

We compute the *Pareto frontier* of the principal’s and the agent’s utility at every state. The Pareto frontier at state $s$ is a function $f_s$ such that $f_s(u)$ is the principal’s maximum achievable onward utility, subject to the agent’s onward utility is exactly $u$ (while satisfying onward participation constraints). We show that the functions $\{f_s\}_s$ can be computed efficiently in a recursive way, and consequently, the principal’s optimal overall utility can be read-off from the Pareto frontier in the initial state, and is equal to $\max_{u \geq 0} f_{\text{max}}(u)$.
Algorithmic insights. Algorithm 1 heavily exploits the fact that feasible policies (i.e., those ensuring participation) are closed under convex combination, and as a result, the Pareto frontier function at each state is concave. In each state $s$, we first compute, for each action $a$ that can be taken at that state, the Pareto frontier $g_{s,a}$ after taking that action, and then shift each $g_{s,a}$ by the principal/agent's reward for the state-action pair $(s,a)$. Then, since the policy can randomize among actions, we simply take the upper concave closure of all these shifted curves and then truncate it to ensure participation, which gives the Pareto frontier of this state.

A key subroutine of Algorithm 1 is Algorithm 2, which computes the Pareto frontier function $g(s,a)$ after taking action $a$ in state $s$ — we call this the subcurve for state-action pair $(s,a)$. Intuitively, to evaluate this subcurve at some $u$, one needs to determine the optimal way of distributing the agent’s utility $u$ into onward states, such that the resulting principal’s utility is maximized. Because the Pareto frontier functions of all onward states are concave, one can construct the entire subcurve using a water-filling procedure, where the construction always “follows” the Pareto frontier function that has the largest slope. In other words, if we want to increase the principal’s utility by an infinitesimal amount after $(s,a)$, we should increase the principal’s utility (and decrease the agent’s utility) in exactly one of the onward states, namely the onward state that provides the best tradeoff.

Implementing the optimal policy. Algorithm 1 solves only half of the problem — it computes the principal’s optimal utility, but not the corresponding policy. In fact, it could be the case that any optimal policy requires exponential space to explicitly describe (even approximately), so it is impossible to output an optimal policy in polynomial time. Nevertheless, given the Pareto frontier functions constructed in Algorithm 1, we present an algorithm (Algorithm 3) that can efficiently execute an optimal policy on the fly. At a high level, while there are exponentially many possible runs under the optimal policy, to execute the policy we need to follow only the one that actually happened.

More specifically, the optimal policy chooses a point $p_1$ on the Pareto frontier of $s_{\text{init}}$ to realize. We can backtrack how Algorithm 1 achieved $p_1$, which tells us to take an action (say $a_1$) and gives us a point $\{q_{s_2}\}$ on the Pareto frontier of each onward state $s$ reachable from $(s_{\text{init}},a_1)$. We then play action $a_1$ and transition to one of these states (say $s_2$) at random, and we can continue to execute the point $q_{s_2}$ on the Pareto frontier of $s_2$ using the same process as above.

Correctness of Algorithms 1 and 3. Now we are ready to prove the correctness of our algorithms.

**Lemma 1.** Algorithm 1 computes the Pareto frontier functions for all states, i.e., for all $s \in S$, $h \in H_{c,s}$, and $u \geq 0$:
\[
f_s(u) = \max \{ u_{\pi}^A(h,s) \mid \pi : u_{\pi}^A(h,s) = u \text{ and } \\
\forall a \in A_s, \forall s' > s, \forall h' \in H_{c,s'} \text{ with } \\
h' \geq h + (s,a), u_{\pi}^A(h',s') \geq 0 \}.
\]

**Algorithm 2:** subcurve$(s,a)$: algorithm for computing the subcurve for a state-action pair.

**Input:** State $s \in S$, action $a \in A_s$, transition operator $P$, Pareto frontier functions \{ $f_{s'}$ $s' > s$, for downstream states $s'$.

**Output:** Subcurve for state-action pair $(s,a)$.

/* we represent each $f_{s'}$ and the subcurves by their turning points; let $f_{s'} = (f_{s',0}, \ldots, f_{s',k_{s'}})$ where $k_{s'}$ is the number of pieces of $f_{s'}$ and each $f_{s',i} = (x_{s',i}, y_{s',i})$ is a point in $R^2$ */

1. let $p_0 \leftarrow \sum_{s' > s} P(s,a,s') \cdot f_{s',0}$;

2. let the source point of $p_0$ from each onward state $s'$ be $f_{s',0}$;

/* the source points are for bookkeeping and are used in Algorithm 3 to execute the policy */

3. let $k \leftarrow \sum_{s' > s} k_{s'}$, and $c_{s'} = 0$ for all $s' > s$;

/* $c_{\ell}$ is the number of pieces of $f_{s'}$ that we have already used */

4. for $i = 1, \ldots, k$ do

5. let $L_i \leftarrow \{s' \mid s < s' \leq n, c_{s'} < k_{s'}\}$;

6. let $s_i \leftarrow \arg\max_{s' \in L_i} \{ y_{s',c_{s'} + 1} - y_{s',c_{s'}}, x_{s',c_{s'} + 1} - x_{s',c_{s'}} \}$;

7. let $p_i \leftarrow p_{i-1} + P(s,a,s_i) \cdot (x_{s',c_{s'} + 1} - x_{s',c_{s'}}, y_{s',c_{s'} + 1} - y_{s',c_{s'}})$;

8. let $c_{s_i} \leftarrow c_{s_i} + 1$;

9. let the source point of $p_i$ from each onward state $s'$ be $f_{s',c_{s'}}$;

10. return $(p_0, p_1, \ldots, p_k)$.

**Lemma 2.** Let $(s_1 = s_{\text{init}}, a_1, s_2, a_2, \ldots, a_{t-1}, s_t = s_{\text{term}})$ be any (random) output sequence of Algorithm 3. Then for each $i \in [1..t]$, $a_i \sim \pi^*(h_i, s_i)$, where $\pi^*$ is an optimal policy subject to participation constraints, and $h_i = (s_1, a_1, \ldots, s_{i-1}, a_{i-1})$. In particular, $\pi^*$ guarantees that
\[
u_{A_i}^*(h_i, s_i) = u_i \quad \text{and} \quad u_{\pi}^* (h_i, s_i) = f_{s_i} (u_i),
\]
where $u_i$ is computed in line 1 or line 12 in Algorithm 12.

**Runtime of Algorithm 1.** We remark that Algorithm 1 may take exponential time in the worst case. In fact, let $|f_s|$ be the number of pieces in $f_s$ (recall that all $f_s$ are piecewise-linear). Then for each $s \in S$, $|f_s|$ can be as large as $\Theta (|A_s| \cdot \sum_{s'} |f_{s'}|)$, so $|f_s|$ may grow exponentially as $s$ grows. Formally, we have the following result regarding the time complexity of Algorithm 1.

**Lemma 3.** Algorithm 1 runs in time $n \cdot (2m_0)^n$, where $n = |S|$ is the number of states and $m_0 = \max_{s \neq s_{\text{term}}} |A_s|$ is the maximum number of actions available in a single state. Moreover, suppose the state space can be partitioned into $H$ sets $\{S_1, \ldots, S_H\}$, such that for any two states $s \in S_i$ and $s' \in S_j$, for all $a \in A_s$, $P(s,a,s') > 0$ only if $j = i + 1$. Then Algorithm 1 runs in time $O((m_0 H)^d \log n)$. In particular, Algorithm 1 is polynomial-time if $H$ is constant.
Algorithm 3: Algorithm for executing the optimal policy on the fly (i.e., choosing actions on the fly)

**Input:** Problem instance \((S, \{A\}, r, \pi, r_A, P)\), Pareto frontier functions \(\{f_s\}\) and other auxiliary variables computed by Algorithm 1.

**Output:** A random trajectory \(\langle s_1 = s_{\text{init}}, a_1, \ldots, a_{t-1}, s_t = s_{\text{term}} \rangle\) of the optimal policy subject to participation constraints.

/* we require the following bookkeeping information: the source points (lines 2 and 9 of Algorithm 2), the Pareto frontier functions before truncation \(\{f_s\}\) (line 7 of Algorithm 1), and “which turning points come from which subcurve” (lines 7 and 8 of Algorithm 1) */

1. let \(i \leftarrow 1, u_1 \leftarrow \arg\max_{u} f_{s_{\text{init}}}^i(u), s_1 \leftarrow s_{\text{init}}\);
2. while \(s_i \neq s_{\text{term}}\) do
3.     let \(p_i^- : p_i^+ \in f_{s_i}\) and \(\alpha_i \in [0, 1]\) be such that
4.         \((u_i, f_{s_i}(u_i)) = \alpha_i \cdot p_i^- + (1 - \alpha_i) \cdot p_i^+;\)
5.     let \(a_i^-, a_i^+ \in A\) be such that \(p_i^-\) and \(p_i^+\) come from the (shifted) subcurves for \((s_i, a_i^-)\) and \((s_i, a_i^+)\) respectively;
6.     if \(\text{rand}_i \leq \alpha_i\) then
7.         let \(a_i \leftarrow a_i^-, p_i \leftarrow p_i^- - (r_A(s_i, a_i), r_P(s_i, a_i));\)
8.     else
9.         let \(a_i \leftarrow a_i^+, p_i \leftarrow p_i^+ - (r_A(s_i, a_i), r_P(s_i, a_i));\)
10.    play action \(a_i\) and transition to next state \(s_{i+1} \sim P(s_i, a_i)\);
11.    let \(q_i = (x_i, y_i)\) be the source point of \(p_i\) from onward state \(s_{i+1}\) in the subcurve (before shifting) for \((s_i, a_i)\);
12.    let \(u_{i+1} \leftarrow x_i, i \leftarrow i + 1;\)
13. return \(\langle s_1 = s_{\text{init}}, a_1, \ldots, a_{t-1}, a_{t-1}, s_t = s_{\text{term}} \rangle;\)

3.2 Achieving Efficiency by Approximation

Below we show how to make Algorithm 1 polynomial-time by sacrificing an arbitrarily small additive error \(\varepsilon\). The idea is simple: since each Pareto frontier function is concave, when the number of pieces in each function is large, there must be many pieces that are essentially the same and therefore can be merged at little cost.

To make this concrete, instead of the exact Pareto frontier curves, we maintain their intersections with equally spaced horizontal lines, and use the upper concave closures of these intersections as proxy curves to approximate the actual Pareto frontier functions. Since the Pareto frontier curves are concave, each curve can intersect each horizontal line at most twice — once in the increasing phase, and once in the decreasing phase. Moreover, the error from this approximation is uniformly upper bounded by the distance between two neighboring horizontal lines. Formally, for any desired error \(\varepsilon > 0\), we insert the following operation after line 9 of Algorithm 1:

\[
\text{let } f_s \leftarrow \text{conv}\{f_s \cap \{(x, y) \mid x \in \mathbb{R}, y \in \{-n, -n + \frac{\varepsilon}{n}, -n + \frac{2\varepsilon}{n}, \ldots, n - \frac{2\varepsilon}{n}, n - \frac{\varepsilon}{n}, n\}\};
\]

The approximation guarantee and the runtime of the modified algorithms are stated in the following theorem.

**Theorem 4.** Suppose the principal’s reward function is bounded, i.e., for all \(s \in S\) and \(a \in A_s\),

\[-1 \leq r_P(s, a) \leq 1.\]

Then, the modified Algorithm 1 with parameter \(\varepsilon > 0\) together with Algorithm 3 implicitly computes a policy \(\pi^*\) guaranteeing participation, which satisfies

\[u_P^*(\emptyset, s_{\text{init}}) \geq u_P^*(\emptyset, s_{\text{init}}) - \varepsilon,\]

where \(\pi^*\) is the optimal policy guaranteeing participation. Moreover, the modified Algorithm 1 with parameter \(\varepsilon > 0\) takes time \(O((mn^3 \log n)/\varepsilon)\).

3.3 Achieving Efficiency with Definitive Decisions

Although Algorithm 1 can take exponential time in general, we show that it is in fact strongly polynomial-time for a special case of the problem, where all decisions are definitive.

Formally, we consider the model with the following restriction.

**Definition 5** (Definitive decisions). We say an MDP has the “definitive decisions” property if, for each state \(s \in S\), the set of available actions \(A_s\) at \(s\) includes at most one “decide-later” action, for which there are no restrictions on its transition probabilities; all other actions in \(A_s\) are “decisions” actions, which immediately terminate the process by taking the state deterministically to \(s_{\text{term}}\).

Notice that there are no restrictions on the rewards, so each action can induce arbitrary rewards for the principal and the agent. MDPs with definitive decisions property can capture many real-world scenarios, such as candidate screening and training programs (see Section 4).

We can show that our exact algorithm (Algorithm 1) runs in polynomial time if the input MDP satisfies the definitive decisions property.

**Lemma 6.** In the case of definitive decisions, Algorithm 1 computes the (exact) Pareto frontier functions for all states in time \(O(m^2 \log n)\).
4 Application: Screening Policy Design

In this section, we illustrate the wide applicability of our results by examining a natural application scenario: the design of screening policies. The setting we consider is the following. Suppose we run a crowdsourcing task that requires very specific qualifications, so we would like all crowd workers to be screened before entering the pool. Whether a worker actually qualifies can only be observed during the job (even workers themselves do not know beforehand), but we know that workers who qualify tend to perform better in certain simple online tests. In order to estimate workers’ qualifications, we provide these tests to all prospective workers, so they can potentially take as many as they wish.

Ideally, we would like to admit only workers who take a large number of tests and perform well on average, because in that case we are very confident that they qualify. However, if the time cost of repeatedly taking tests outweighs the benefit of getting the job, then workers will simply not try at all. Even if a worker starts trying, he may quickly give up after seeing a few negative results, deciding that his chance of eventually qualifying does not justify the time cost of keep trying. Taking this into consideration, our goal is to design a policy of admission that in expectation maximizes our gain (i.e., the amount of useful job done minus the total payment) by encouraging workers to take more tests.

The setup. Formally, there are two types of workers: “good” and “bad”. Before taking tests, neither we nor the worker know whether the worker is good or bad, but there are commonly known prior probabilities for both types, \( p_g \) and \( p_b \). A test has two possible results: pass (P) and fail (F). A good worker passes each test independently with probability \( q_g \), and a bad worker passes with probability \( q_b < q_g \). Without loss of generality, suppose our utility for accepting a good agent is \( u_g > 0 \), and that for accepting a bad one is \( u_b < 0 \); each worker receives utility 1 for being accepted, and \(-c\) for the time cost of taking a test.

A policy maps each sequence of test outcomes — e.g., “PFPPFPF” — to a distribution over three possible actions: accept, reject, and decide later. Our goal is to find an optimal policy, taking into consideration that workers may drop out at any moment if their onward expected utility is negative. We remark that our formulation and algorithms can in fact handle much more general settings, but we choose to present this specific application it admits some nice structures and it can demonstrate several interesting aspects of our results.

4.1 Screening Policy Design as Planning with Participation Constraints

We show how to cast the screening problem as planning with participation constraints. First observe that at any moment in the above process, our belief on the worker’s type (i.e., the posterior probability that a worker is good) only depends on the sequence of test outcomes so far. In fact, this probability depends only on the numbers of Ps and Fs that a worker gets. Therefore, without loss of generality, we can consider an MDP where the states are \( S = \{ (p, f) \mid p, f \in \mathbb{N}\} \cup \{s_{\text{term}}\} = \mathbb{N}^2 \cup \{s_{\text{term}}\}, \) where \( s_{\text{init}} = (0, 0) \). The posterior probability of being good in a state \((p, f)\) is given by:

\[
Pr[\text{good} \mid (p, f)] = \frac{p_g \cdot q_g^p (1 - q_g)^f}{p_g \cdot q_g^p (1 - q_g)^f + p_b \cdot q_b^p (1 - q_b)^f}.
\]

The set of actions associated with each state \( s \) has 3 elements: \( A_s = \{ \text{accept}, \text{reject}, \text{decide-later} \} \). Among these actions, accept and reject are decision actions, in the sense that they terminate the process immediately. The rewards they induce for the agent and the principal are: \( (1, Pr[\text{good} \mid (p, f)] \cdot (u_g - u_b) + u_b) \) for accept, and \((0, 0)\) for reject. The decide-later action induces rewards \((-c, 0)\) and transition probabilities:

\[
Pr[(p + 1, f) \mid (p, f)] = Pr[\text{good} \mid (p, f)] \cdot q_g + (1 - Pr[\text{good} \mid (p, f)]) \cdot q_b, \\
Pr[(p, f + 1) \mid (p, f)] = 1 - Pr[(p + 1, f) \mid (p, f)].
\]

Computing optimal policies. With the above formulation, we can apply our algorithms to find optimal policies. First consider the case where there is an upper limit \( N \) on the number of tests each worker can take. In such cases, the state space becomes \( S = \{ (p, b) \mid p + b \leq N \} \cup \{s_{\text{term}}\}, \) where \( |S| = O(N^2) \). Since the formulation has definitive constraints, we can directly apply Algorithm 1 to compute the optimal policy in time \( O(N^4 \log N) \).

Even in the idealized setting where the number of tests a worker can take is arbitrarily large (which is possible because workers may not consider sunk cost when making decisions), we can still approximate the optimal policy to arbitrary precision. More specifically, the following lemma states that there exists a policy that uses at most \( N \) tests and achieves a principal’s utility at most \( O(1/N) \) worse than the optimal policy.

Lemma 7. Let \( OPT \) be the principal’s optimal utility in screening policy design with no limit on the number of tests. For any \( N \), there is a policy guaranteeing participation which plays reject for all \( (p, f) \) where \( p + f > N \), and achieves expected utility \( OPT - O(1/N) \) for the principal.

Combining Lemma 6 with Lemma 7, we can compute an \( \varepsilon \)-approximately optimal policy in time \( O((\log(1/\varepsilon)/\varepsilon)^8) \) for any \( \varepsilon > 0 \). To summarize, we have the following theorem.

Theorem 8. We can compute an optimal screening policy in time \( O(N^4 \log N) \) when the maximum number of tests a worker is willing to take is \( N \). Moreover, when workers may take arbitrarily many tests, we can compute an approximately optimal policy up to any additive error \( \varepsilon > 0 \) in time \( O((\log(1/\varepsilon)/\varepsilon)^8) \).

5 Future Directions

We leave a number of open questions as interesting avenues for future work: Is there a poly-time algorithm (or a hardness result) for exactly solving MDP with participation constraints? What if the policy must be deterministic? What additional applications can fit in the MDP model in this paper? Another future direction is the case with incomplete information, e.g., when the agent’s reward function is only approximately known. One potential solution is to run a robust version of our algorithm that allows some slack in the IR constraints, where the slack corresponds to a confidence interval for our estimate of the agent’s reward function.
References


A Omitted Proofs

Proof of Lemma 1. We use mathematical induction. We know \( f_n \) is the Pareto frontier at \( s_{\text{term}} = n \) since the only possible onward utility for both the principal and the agent is 0. For any \( s \in [n-1] \), assume that \( f_{s'} \) is the Pareto frontier for every state \( s' > s \). We only need to show that for all \( h \in \mathcal{H}_{<s} \) and \( u \in \mathbb{R}^+ \),

\[
f_s(u) = \max\{u_p(h, s) \mid \pi : u_A^\pi(h, s) = u \text{ and } \forall a \in A_s, s' > s, \mathcal{H}_{<s'} \ni h' \geq h + (s, a), u_A^\pi(h', s') \geq 0\}.
\]

First observe that \( f_s \) is achievable by construction. That is, for any possible value \( u \in \mathbb{R}^+ \) of the agent’s utility (and history \( h \in \mathcal{H}_{<s} \) which is essentially irrelevant), there exists a policy \( \pi \) ensuring onward participation (i.e., \( u_A^\pi(h', s') \geq 0 \) for all \( a \in A_s, s' > s \), and \( \mathcal{H}_{<s'} \ni h' \geq h + (s, a) \)) such that \( f_s(u) = u_p(h, s) \). This is because \( f_{s'} \) is achievable for any \( s' > s \), so the subcurve for \((s, a)\) for every \( a \in A_s \) (which is constructed explicitly based on \( \{f_{s'}\}_{s' > s} \)) is achievable. Since \( f_s \) is constructed by randomizing among actions corresponding to these subcurves, \( f_s \) itself is also achievable. See the proof of Lemma 2 for a more detailed argument.

We now show the other direction that \( f_s \) uniformly upper bounds the principal’s optimal utility. That is, for any policy \( \pi \) ensuring onward participation,

\[
f_s(u_A^\pi(h, s)) \geq u_p^\pi(h, s).
\]

At a high level, this is because \( f_s \) is constructed to achieve an optimal way.

By construction, \( f_s \) takes the upper concave closure of the shifted subcurves, and consequently \( f_s \) randomizes among these actions optimally. We only need to show that for any \( a \in A_s \), the subcurve for \((s, a)\) computed by Algorithm 2 uniformly upper bounds the principal’s utility after taking action \( a \) in state \( s \). Given the induction hypothesis, the above is equivalent to: letting \( g_{s,a} \) be the subcurve (before shifting) for \((s, a)\), for any \( u \in \mathbb{R} \), and \( u_{s'} \in \mathbb{R} \) for all \( s' > s \) satisfying

\[
u = \sum_{s' > s} P(s, a, s') \cdot u_{s'},
\]

we always have

\[
g_{s,a}(u) \geq \sum_{s' > s} P(s, a, s') \cdot f_{s'}(u_{s'}).
\]

Suppose otherwise, i.e., there exists \( u \) and \( \{u_{s'}\}_{s'} \) satisfying

\[
u = \sum_{s' > s} P(s, a, s') \cdot u_{s'} \quad \text{and} \quad g_{s,a}(u) < \sum_{s' > s} P(s, a, s') \cdot f_{s'}(u_{s'}).
\]

Let \( \pi \) be the policy achieving \( g_{s,a}(u) \), and \( u_A^* = u_A^\pi(s') \) for each \( s' > s \). Note that \( \{u_A^* \}_{s'} \) satisfy

\[
u = \sum_{s' > s} P(s, a, s') \cdot u_A^* \quad \text{and} \quad g_{s,a}(u) = \sum_{s' > s} P(s, a, s') \cdot f_{s'}(u_A^*).
\]

For any \( \alpha \in [0, 1] \), let \( u_{s'}^\alpha = \alpha \cdot u_{s'} + (1 - \alpha) \cdot u_A^* \) for each \( s' > s \), and let

\[
g_{s,a}^\alpha(u) = \sum_{s' > s} P(s, a, s') \cdot f_{s'}(u_{s'}^\alpha).
\]

Recall that by construction, each \( f_{s'} \) is concave, so we have

\[
g_{s,a}^\alpha(u) \geq \sum_{s' > s} P(s, a, s') \cdot \left( \alpha \cdot f_{s'}(u_{s'}) + (1 - \alpha) \cdot f_{s'}(u_A^*) \right).
\]

In particular, for any \( \alpha > 0 \),

\[
g_{s,a}^\alpha(u) > g_{s,a}^0(u) = g_{s,a}(u).
\]

Moreover, since each \( f_{s'} \) is piecewise linear, there exists some \( \varepsilon > 0 \), such that \( g_{s,a}^\alpha(u) \) is linear in \( \alpha \) for \( \alpha \in [0, \varepsilon] \).

We focus on this small interval where \( g_{s,a}^\alpha(u) \) is linear in \( \alpha \) and in fact, every \( f_{s'}(\alpha \cdot (u_{s'} - u_A^*) + u_A^*) \) is linear in \( \alpha \). So we have

\[
\frac{\partial g_{s,a}^\alpha(u)}{\partial \alpha} \bigg|_{\alpha = \varepsilon/2} = \sum_{s'} P(s, a, s') \cdot f_{s'}(\alpha \cdot (u_{s'} - u_A^*) + u_A^*)
\]

\[
= \sum_{s'} P(s, a, s') \cdot (u_{s'} - u_A^*) \cdot \frac{\partial f_{s'}(t)}{\partial t} \bigg|_{t=u_A^*}.
\]
Here,

\[
\frac{\partial f_x(t)}{\partial t} \bigg|_{t = u_{s'}} = \begin{cases}
\lim_{t \to u_{s'}} \frac{f_x(u_{s'}) - f_x(t)}{u_{s'} - t}, & \text{if } u_{s'} < u_{s''} \\
\lim_{t \to u_{s'}} \frac{f_x(u_{s'}) - f_x(t)}{u_{s'} - t}, & \text{if } u_{s'} > u_{s''} \\
0, & \text{otherwise.}
\end{cases}
\]

Let

\[ S^+ = \{ s' \mid s' > s, u_{s'} > u_{s''} \} \quad \text{and} \quad S^- = \{ s' \mid s' > s, u_{s'} < u_{s''} \}. \]

Then we have

\[
\sum_{s' \in S^+} P(s, a, s') \cdot (u_{s'} - u_{s''}) \cdot \frac{\partial f_x(t)}{\partial t} \bigg|_{t = u_{s'}} - \sum_{s' \in S^-} P(s, a, s') \cdot (u_{s'} - u_{s''}) \cdot \frac{\partial f_x(t)}{\partial t} \bigg|_{t = u_{s''}} = \frac{\partial g_{s, a, \pi}(u)}{\partial \alpha} \bigg|_{\alpha = \epsilon/2} > 0.
\]

Since

\[
\sum_{s' \in S^+} P(s, a, s') \cdot (u_{s'} - u_{s''}) = \sum_{s' \in S^-} P(s, a, s') \cdot (u_{s'} - u_{s''}),
\]

it must be the case that for some \( s^+ \in S^+ \) and \( s^- \in S^- \),

\[
\frac{\partial f_x(t)}{\partial t} \bigg|_{t = u_{s^+}} > \frac{\partial f_x(t)}{\partial t} \bigg|_{t = u_{s^-}}.
\]

This contradicts the construction of \( g_{s, a, \pi} \), since the algorithm must exhaust the current piece (containing \( u_{s''} \)) before following the current piece (containing \( u_{s''} \)) of \( f_{s''} \), but by the definition of \( S^+ \) and \( S^- \), the algorithm has not finished the current piece of \( f_{s''} \), but has already started following the current piece of \( f_{s''} \). This concludes the proof.

**Proof of Lemma 2.** We only need to show the second half of the statement, i.e., for any \( (s_1, \ldots, s_i) \),

\[ u_{A}^+(h_i, s_i) = u_i \quad \text{and} \quad u_{P}^+(h_i, s_i) = f_{s_i}(u_i), \]

which implies the first half, since \( u_1 = \arg\max_{u'} f_{s_{\text{min}}}(u') \), and by Lemma 1, \( f_s \) uniformly upper bounds the maximum utility achievable by the principal.

We apply induction on \( i \). When \( i = n - 1 \), we have

\[ p_i^- = (r_A(s_i, a_i^-), r_P(s_i, a_i^-)) \quad \text{and} \quad p_i^+ = (r_A(s_i, a_i^+), r_P(s_i, a_i^+)), \]

where by construction,

\[ \alpha_i \cdot p_i^- + (1 - \alpha_i) \cdot p_i^+ = (u_i, \hat{f}_s(u_i)). \]

So,

\[ u_{A}^+(h_i, s_i) = \alpha_i \cdot r_A(s_i, a_i^-) + (1 - \alpha_i) \cdot r_A(s_i, a_i^+) = u_i, \]

and similarly,

\[ u_{P}^+(h_i, s_i) = \alpha_i \cdot r_P(s_i, a_i^-) + (1 - \alpha_i) \cdot r_P(s_i, a_i^+) = \hat{f}_s(u_i) = f_{s_i}(u_i). \]

Now for \( i < n - 1 \), suppose the statement holds for all \( i' > i \), and consider \( u_{A}^+(h_i, s_i) \) and \( u_{P}^+(h_i, s_i) \). We have

\[
u_{A}^+(h_i, s_i) = \alpha_i \cdot \left( r_A(s_i, a_i^-) + \sum_{s' > s_i} P(s_i, a_i^-, s') \cdot u_{A}^+(h_i + (s_i, a_i^-), s') \right)
\]

\[ + (1 - \alpha_i) \cdot \left( r_A(s_i, a_i^+) + \sum_{s' > s_i} P(s_i, a_i^+, s') \cdot u_{A}^+(h_i + (s_i, a_i^+), s') \right).\]

For each \( s' > s_i \), let \( u_{i+1,s'} \) (resp. \( u_{i+1,s'}^{+} \)) be the value of variable \( u_{i+1} \) computed by Algorithm 3 on input sequence \( s_1, \ldots, s_i, s' \) conditioning on the first \( i \) output actions being \( a_1, \ldots, a_i^- \) (resp. \( a_1, \ldots, a_i^+ \)). By the induction hypothesis, we have

\[ u_{i+1,s'}^- = u_{A}^+(h_i + (s_i, a_i^-), s') \quad \text{and} \quad u_{i+1,s'}^+ = u_{A}^+(h_i + (s_i, a_i^+), s'). \]
By the construction of source points (lines 2 and 9 in Algorithm 2), we have
\[
u_i = \alpha_i \cdot \left( r_A(s_i, a^+_i) + \sum_{s' > s} P(s_i, a^+_i, s') \cdot u_{i+1, s'} \right)
+ (1 - \alpha_i) \cdot \left( r_A(s_i, a^-_i) + \sum_{s' > s} P(s_i, a^-_i, s') \cdot u_{i+1, s'} \right)
= \alpha_i \cdot \left( r_A(s_i, a^+_i) + \sum_{s' > s_i} P(s_i, a^+_i, s') \cdot u_A^*(h_i + (s_i, a^+_i), s') \right)
+ (1 - \alpha_i) \cdot \left( r_A(s_i, a^-_i) + \sum_{s' > s_i} P(s_i, a^-_i, s') \cdot u_A^*(h_i + (s_i, a^-_i), s') \right)
= u_A^*(h_i, s_i).
\]
Similarly one can show that
\[
f_{s_i}(u_i) = u_P^*(h_i, s_i).
\]
This concludes the proof.

**Proof of Lemma 3.** For the first part of the theorem, we only need to show inductively that
\[
|f_s| \leq m_0 \cdot \sum_{s' > s} |f_{s'}|.
\]
This is true because for each \(a \in A_s\),
\[
|g_{s,a}| \leq \sum_{s' > s} |f_{s'}|,
\]
and
\[
|f_s| \leq \sum_{a \in A_s} |g_{s,a}| \leq m_0 \cdot \sum_{s' > s} |f_{s'}|.
\]
For the second half, we show that for the \(i\)-th layer where \(i \in [H - 1]\),
\[
\sum_{s \in S_i} |f_s| \leq m_0 n \cdot \sum_{s' \in S_{i+1}} |f_{s'}|.
\]
In fact, for each \(s \in S_i\), by the same reasoning as in the first part,
\[
|f_s| \leq \sum_{a \in A_s} |g_{s,a}| \leq m_0 \cdot \sum_{s' \in S_{i+1}} |f_{s'}|.
\]
So
\[
\sum_{s \in S_i} |f_s| \leq |S_i| \cdot \sum_{s' \in S_{i+1}} |f_{s'}| \leq n \cdot \sum_{s' \in S_{i+1}} |f_{s'}|.
\]
As a result, for each \(i \in [H]\),
\[
\sum_{s \in S_i} |f_s| \leq (m_0 n)^{H-i}.
\]
So for any \(s \in S\) (where the worst case is when \(s \in S_1\)),
\[
|f_s| \leq m_0^H n^{H-1},
\]
and because of the use of heaps, computing \(f_s\) takes time \(O(|f_s| \log n)\). So the overall time complexity is
\[
n \cdot O(m_0^H n^{H-1} \log n) = O((m_0 n)^H \log n).
\]

**Proof of Theorem 4.** First consider the approximation guarantee. Let \(f^*_s\) be the actual Pareto frontier functions (i.e., the ones computed by the unmodified version of Algorithm 1). We inductively show that for each \(s \in S = [n]\), for all values \(u\) of the agent’s utility achievable by policies ensuring onward participation,
\[
f^*_s(u) - \frac{(n - s)}{n} \cdot \varepsilon \leq f_s(u) \leq f^*_s(u).
\]
When \( s = n \), it is clear that \( f_s = f_s^* \). For \( s < n \), by the induction hypothesis and the construction of \( f_s^* \), for all \( u \) achievable by policies ensuring onward participation,

\[
f^*_s(u) - \frac{(n - s - 1)}{n} \cdot \varepsilon \leq f_s(u) \leq f^*_s(u).
\]

Moreover, by the construction of \( f_s \), since \( \hat{f}_s \) is concave and bounded between \(-n\) and \( n \), for all \( u \) achievable,

\[
\hat{f}_s(u) - \frac{\varepsilon}{n} \leq f_s(u) \leq \hat{f}_s(u),
\]

which implies

\[
f^*_s(u) - \frac{\varepsilon}{n} \leq f_s(u) \leq f^*_s(u).
\]

This finishes the proof of the approximation guarantee.

For the time complexity, first observe that the number of pieces in each \( f_s \) is \( O(n^2/\varepsilon) \). This is because \( \hat{f}_s \) is concave, so \( \hat{f}_s \) can intersect each horizontal line at most twice. Observe that the bottleneck of Algorithm 1 is the subcurve \((s, a)\) procedure. In each call to subcurve \((s, a)\), the total number of pieces \( k \) of \( g_{s,a} \) can be as large as

\[
k = \sum_{s' > s} |f_{s'}| = \sum_{s' > s} O(n^2/\varepsilon) = O(n^3/\varepsilon).
\]

In constructing each of these pieces, the algorithm needs to retrieve a piece from some downstream state with the largest slope, which can be done in time \( O(\log n) \) by maintaining a heap of pieces. So the total time required by a single call to subcurve \((s, a)\) is \( O((n^3 \log n)/\varepsilon) \). Now observe that for each state \( s \), subcurve \((s, a)\) must be run \(|A_s|\) times, so in total, the algorithm makes \( \sum_s |A_s| = m \) calls to subcurve \((s, a)\). This takes time \( O((mn^3 \log n)/\varepsilon) \). One can check that all other parts of the algorithm can be done in the same amount of time.

**Proof of Lemma 6.** We only need to show that for each \( s \in \mathcal{S} \), the number of pieces in \( f_s \) is \( O(m) \), i.e., \( |f_s| = O(m) \). Let \( \text{slope}(f_s) \) denote the set of all distinct slopes of pieces in \( f_s \). Then equivalently, we only need to show that \(|\text{slope}(f_s)| = O(m)\).

Below we show inductively a stronger claim, i.e., for each \( s \),

\[
\left| \bigcup_{s' \geq s} \text{slope}(f_{s'}) \right| \leq 2 \sum_{s' > s} |A_{s'}|.
\]

For \( s = s_{\text{term}} = n \), this is trivially true. Now consider some \( s \neq s_{\text{term}} \). Let \( a_{s,0} \in A_s \) be the “decide later” action for \( s \). Then \( \hat{f}_s \) is the upper concave closure of \( g_{s,a_{s,0}} \), the shifted subcurve for \((s, a_{s,0})\) and \(|A_s| - 1\) points, each corresponding to the rewards of a “decision” action. First note that

\[
\text{slope}(g_{s,a_{s,0}}) \subseteq \bigcup_{s' > s} \text{slope}(f_{s'}).\n\]

This is because the subcurve procedure preserves the slopes of every piece constructed. Then, by taking the concave closure of \( g_{s,a_{s,0}} \) and \(|A_s| - 1\) points, there are at most \( 2(|A_s| - 1) \) new slopes introduced, because each point can introduce at most 2 new slopes. So we have

\[
\left| \bigcup_{s' \geq s} \text{slope}(f_{s'}) \right| = \left| \bigcup_{s' > s} \text{slope}(f_{s'}) \right| + \left| \text{slope}(f_{s}) \setminus \left( \bigcup_{s' > s} \text{slope}(f_{s'}) \right) \right|
\]

\[
\leq 2 \sum_{s' > s} |A_{s'}| + \left| \text{slope}(f_{s}) \setminus \left( \bigcup_{s' > s} \text{slope}(f_{s'}) \right) \right| \quad \text{(induction hypothesis)}
\]

\[
\leq 2 \sum_{s' > s} |A_{s'}| + 2(|A_s| - 1) \quad \text{(bound on number of new slopes)}
\]

\[
\leq 2 \sum_{s' \geq s} |A_{s'}|.
\]

This concludes the proof.
Proof of Lemma 7. Consider an optimal policy $\pi^*$ subject to participation constraints. Since $u_A^\pi(\emptyset, s_{\text{init}}) \geq 0$, we have

$$c \cdot E[T] \leq 1,$$

where $T$ is the (random) number of tests an worker takes before being accepted or rejected. By the Markov bound, we have

$$\Pr[T \geq N] \leq 1/(cN).$$

Given this, we modify $\pi^*$ into a policy $\pi_N$ which always plays reject after the first $N$ tests, while approximately preserving the principal’s utility. We first truncate $\pi^*$ at $N$, and let $\pi'$ be such that

$$\pi'(h, s) = \pi^*(h, s)$$

whenever $|h| < N$, and $\pi'$ deterministically plays reject otherwise. Observe that if participation constraints were satisfied by $\pi'$, then $\pi'$ would satisfy the conditions required by the lemma. This is because $\pi'$ behaves differently than $\pi^*$ with probability at most $1/(cN)$, and as a result, the utility loss is at most $2/(cN)$. However, since $\pi'$ may not satisfy participation constraints, we need further modifications. That is, we increase the probability of playing accept at any state $(p, f)$ where $p + f = N$ (and let the resulting policy be $\pi_N$), to the extent that $u_A^\pi_N(h, (p, f)) = u_A^\pi(h, (p, f))$ for all $h \in H_{N-1}$. This modification again can hurt the principal’s utility by at most $2/(cN)$ for exactly the same reason. Overall, $\pi_N$ is a policy satisfying the conditions of the lemma, where the utility gap is at most $4/(cN)$. \qed