CONTRACTING WITH A REINFORCEMENT LEARNING AGENT BY PLAYING TRICK OR TREAT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study *principal-agent* problems where a *farsighted* agent takes costly actions in an MDP. The core challenge in these settings is that agent's actions are *hidden* to the principal, who can only observe their outcomes, namely state transitions and their associated rewards. Thus, the principal's goal is to devise a policy that incentives the agent to take actions leading to desirable outcomes, by committing to a payment scheme (a.k.a. *contract*) at each step. Interestingly, we show that Markovian policies are unfit in these settings, as they do not allow to achieve the optimal principal's utility and are constitutionally intractable. Thus, accounting for history in unavoidable, and this begets considerable additional challenges compared to standard MDPs. Nevertheless, we design an efficient algorithm to compute an approximately optimal policy, leveraging a compact way of representing histories for this purpose. Unfortunately, this policy incentivizes the agent to take the desired actions only approximately. To fix this, we design an efficient method to make such a policy incentive compatible, by only introducing a negligible loss in principal's utility, generalizing an existing technique for classical principal-agent problems.

1 INTRODUCTION

Over the last years, *principal-agent problems* have received a growing attention from the algorithmic game theory community (see, *e.g.*, (Dütting et al., 2019; Alon et al., 2021; Guruganesh et al., 2021; Dutting et al., 2021; Dütting et al., 2022; Castiglioni et al., 2022; Alon et al., 2023; Castiglioni et al., 2023a;b; Guruganesh et al., 2023)). At their core, these problems capture scenarios where a principal tries to steer the behavior of a self-interested agent towards favorable outcomes. In the simplest version of the problem, the agent has to take a costly action resulting in a stochastic outcome that determines a reward for the principal. The key challenge stems from the fact that the agent's action is *hidden* to the principal, who can only observe its resulting outcome. In order to incentivize the agent to take actions leading to favorable outcomes, the principal commits to a *contract*, which is a payment scheme specifying a monetary transfer from the principal to the agent for every possible outcome.

Thanks to their generality and flexibility, principal-agent problems are able to capture a terrific number of real-world applications, such as, *e.g.*, crowdsourcing (Ho et al., 2014), online labor platforms (Kaynar & Siddiq, 2023), blockchain (Cong & He, 2019), delegation of machine learning tasks (Cai et al., 2015), and pay-for-performance healthcare (Bastani et al., 2016; 2019). Moreover, algorithmic contract design in principal-agent problems is expected to play a crucial role in a world that increasingly relies on AI agents to perform complex tasks; see, *e.g.*, (Hadfield-Menell & Hadfield, 2019; Wang et al., 2024; Saig et al., 2024).

Despite the recent surge in popularity of principal-agent problems, most of the works on the topic focus on "one-shot" problems in which the principal and the agent interact only one time. Nevertheless, in most of the real-world problems of interest, it is natural to assume that the principal and the agent can interact multiple times in a *sequential* manner. Recently, some works (Ivanov et al., 2024; Wu et al., 2024; Gan et al., 2026) initiated the study of settings where a principal and an agent sequentially interact in a dynamic environment, which is modeled as a *Markov decision process* (MDP). In this paper, we study the problem of computing an optimal policy when the agent is *farsighted*. Differently from previous works, we allow the principal to commit to any policy, as long as it satisfies the classical incentive-compatibility and limited liability constraints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 CONTRIBUTIONS

As a preliminary and motivating step (see Section 3), we show Markovian policies are the wrong tool to deal with a farsighted agent. We provide two strong evidences supporting such a sentence: (i) *Markovian* policies are *not* sufficient to achieve the optimal principal's utility and (ii) they are also computationally intractable. Therefore, we need to deal with history-dependent policies, which in general may not even admit a compact (*i.e.*, polynomially-sized) representation, as the number of possible histories is exponential in the time horizon of the MDP. In Section 4 of this paper, we show that it is possible to focus w.l.o.g. on history-dependent policies that admit a compact representation, which we denote as *promise-form* policies. The rest of the paper (Sections 5 and 6) exploits the power of promise-form policies to design an algorithm that allows to implement an approximately-optimal history-dependent policy in polynomial time. The first building block of our algorithm (see Section 5) is an efficient approximation scheme to compute a promise-form policy that is optimal for the principal, but only *approximately* IC. The second building block (see Section 6) is an efficient procedure to transform an ϵ-IC promise-form policy into one that is *exactly* IC, by only incurring in a bounded loss in principal's expected utility. This gives the desired approximately-optimal history-dependent policy.

Let us remark that the idea of condensing all the information contained in a history into a single "promised value" has already been exploited to deal with *Bayesian* persuasion problems in MDPs by Bernasconi et al. (2023). In this paper, we apply such an idea to solve principal-agent problems in MDPs. This raises considerable additional challenges compared to the *Bayesian* persuasion case, which we discuss in more details in the following Section 1.2.

1.2 RELATED WORKS

In the following, we discuss the previous works that are the most related to this paper.

Principal-agent problems in MDPs The problem studied in this paper is closely related to those recently addressed by Wu et al. (2024) and Ivanov et al. (2024). Similarly to this paper, (Wu et al., 2024; Ivanov et al., 2024) analyze principal-agent settings in which a *farsighted* agent takes costly hidden actions in an MDP, with the principal committing to a contract at each step of the MDP. However, while we study the problem of computing an *optimal policy to commit to* for the principal., (Wu et al., 2024; Ivanov et al., 2024) consider the problem of computing a *subgame perfect equilibrium* of the principal-agent game, *i.e.*, they require the policy to be optimal in every possible principal-agent subgame. Our solution concept allows the principal to achieve a much better utility than the one attainable in a subgame perfect equilibrium. Namely, it allows the principal to implement "threats" that discourage the agent from overriding the course of action desired by the principal.

Finally, (Gan et al., 2026) considers a setting where two players execute actions in a partially observable MDP, with the first player committing to to policy beforehand. Notably, (Gan et al., 2026) assumes that the first player can automatically detect when the second player deviates from the recommended actions or misreport their observation. At high level, its approach is somehow orthogonal to the one presented in this paper. We discretize the agent's utility and build an ϵ-IC policy in a single backward pass, then we perform another pass to compute an IC policy. In contrast, (Gan et al., 2026) discretizes both players' utilities and, in a first pass, computes approximately optimal values under an IC policy, while in a second pass it recovers the actual policy.

Making contracts IC Our work extends a known result Dutting et al. (2021) for "one-shot" principal-agent problems, namely that an almost incentive-compatible contract can be converted into an incentive-compatible one. This result has also been extend to principal-agent problems involving, *e.g.*, menu of contracts Castiglioni et al. (2023a) or costs uncertainty Bernasconi et al. (2024).

Promises We devise a subclass of efficiently representable policies based on the concept of promises introduced by Bernasconi et al. (2023) in the context of Bayesian persuasion in an MDP. Their policies promise a minimum expected utility to the agent at each step. The main challenge in adapting their idea to contract design is that in our framework the principal *cannot* observe the agent's action. This prompted us to use a stricter notion of promises and to change the relative incentive-compatibility constraints in a more complex formulation.

2 PRELIMINARIES

2.1 MODEL

We study principal-agent problems where a *farsighted* agent takes actions in a *time-inhomogeneous finite-horizon* MDP. The actions performed by the agent are *hidden* to the principal, who can only observe their outcomes and the associated rewards. Formally, a problem instance is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{H}, \{P_h\}_{h\in\mathcal{H}}, \mu, \{r_h\}_{h\in\mathcal{H}}, \{c_h\}_{h\in\mathcal{H}})$. $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ a finite set of agent's actions, and $\mathcal{H} \coloneqq \{1, \dots, H\}$ a finite set of time steps, with H being the time horizon. Transition are governed by the function $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, with $P_h(s'|s, a)$ denoting the probability of reaching state $s' \in \mathcal{S}$ at step $h + 1$ when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at step $h \in \mathcal{H}$.[1] The initial state is sampled from the distribution $\mu \in \Delta(\mathcal{S})$. $r_h : \mathcal{S} \times \mathcal{S} \to [0, 1]$ and $c_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are the principal's reward and the agent's cost functions, respectively. When the the agent takes action a at step h inducing a transition from sate s to s', the principal obtains reward $r_h(s, s')$ and the agent suffers cost $c_h(s, a)$.[2] The interaction protocol is detailed in Algorithm 1.

Algorithm 1 Principal-Agent Interaction

Principal commits to $\pi : \mathcal{T} \to \Delta(\mathcal{X})$
$s_1 \sim \mu$, $\tau \leftarrow (s_1)$
for all $h = 1, \dots, H$ **do**
  Principal commits to $(p_h, a_h) \sim \pi(\tau)$
  Agent plays an action $\widehat{a}_h \in \mathcal{A}$ maximizing future cumulative expected utility
  $s_{h+1} \sim P_h(s_h, \widehat{a}_h)$
  Principal gets $r_h^{\mathrm{P}}(s_h, p_h, s_{h+1}) \coloneqq r_h(s_h, s_{h+1}) - p_h(s_{h+1})$
  Agent gets $r_h^{\mathrm{A}}(s_h, p_h, \widehat{a}_h, s_{h+1}) \coloneqq p_h(s_{h+1}) - c_h(s_h, \widehat{a}_h)$
  $\tau \leftarrow \tau \oplus (p_h, a_h, s_{h+1})$

At each step of the MDP, the principal selects a contract and an associated action recommendation for the agent. A *contract* is a payment scheme that prescribes a monetary transfer from the principal to the agent for every possible future state (*i.e.*, every possible outcome of agent's action). Formally, a contract is defined as a function $p : \mathcal{S} \to [0, B]$, with $p(s')$ encoding how much the principal pays the agent in the event that state $s' \in \mathcal{S}$ is reached at the next step of the MDP.[3] In the following, we denote by $\mathcal{P}$ the set of all the possible contracts, while we let $\mathcal{X} \coloneqq \mathcal{P} \times \mathcal{A}$ be the set of all the possible pairs of contracts and action recommendations the principal can choose from.

In the most general case, the principal can commit beforehand to a *non-stationary* and *non-Markovian* policy (henceforth called *history-dependent* policy for short). This defines a randomization over contract-recommendation pairs for every possible *history* realized while interacting with the agent in the MDP. Formally, a history-dependent policy is a function $\pi : \mathcal{T} \to \Delta(\mathcal{X})$, where $\mathcal{T}$ is the set of all the possible histories (up to any step of the MDP). Specifically, $\mathcal{T} \coloneqq \bigcup_{h\in\mathcal{H}} \mathcal{T}_h$, where $\mathcal{T}_h$ is the set of all histories up to step $h \in \mathcal{H}$, defined as follows:

$$\mathcal{T}_h \coloneqq \{\tau \mid \tau = (s_1, p_1, a_1, \dots, s_{h-1}, p_{h-1}, a_{h-1}, s_h) : s_i \in \mathcal{S}, p_i \in \mathcal{P}, a_i \in \mathcal{A}\},$$

with s_i denoting the state reached at the i-th step and (p_i, a_i) being the contract-recommendation pair selected by the principal at that step. Notice that histories do *not* include the actions actually played by the agent, but only those recommended by the principal, as the former are *not* observable. At step h, the agent plays action $\widehat{a}_h$ to maximize their future cumulative utility, taking into account both the pair $(p_h, a_h) \sim \pi(\tau_h)$ and the policy π that determines future expected payments. The agent sticks to the recommendation if such an action maximizes their future cumulative utility.

[1] In this paper, we denote by $\Delta(\mathcal{Y})$ the set of all the probability distributions defined over a set $\mathcal{Y}$.

[2] In this paper, for ease of presentation, we assume w.l.o.g. that there exists an action $a \in \mathcal{A}$ whose cost is $c_h(s, a) = 0$ for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$.

[3] As it is customary in the literature, in this paper we assume that the agent has *limited liability*, *i.e.*, payments are only from the principal to the agent and *not viceversa* (see, *e.g.*, (Dütting et al., 2019)), and, additionally, that payments are bounded above by some "sufficiently-big" constant $B \in \mathbb{R}_{>0}$ (see, *e.g.*, (Zhu et al., 2023)).

2.2 VALUE FUNCTIONS

Next, we introduce value functions for the principal and the agent. The principal's value function $V_h^{\mathrm{P},\pi} : \mathcal{T}_h \to \mathbb{R}$ encodes the principal's cumulative expected utility from step $h \in \mathcal{H}$ onwards under policy $\pi : \mathcal{T} \to \Delta(\mathcal{X})$, when assuming that the *agent sticks to recommendations*. Formally, for every history $\tau = (s_1, p_1, a_1, \ldots, s_{h-1}, p_{h-1}, a_{h-1}, s_h) \in \mathcal{T}_h$, the following holds:[4]

$$V_h^{\mathrm{P/A},\pi}(\tau) \coloneqq \sum_{(p,a)\in\mathcal{X}_\pi} \pi(p,a|\tau)\, Q_h^{\mathrm{P/A},\pi}(\tau,p,a),$$

where $Q_h^{\mathrm{P},\pi} : \mathcal{T}_h \times \mathcal{X}_\pi \to \mathbb{R}$ and $Q_h^{\mathrm{A},\pi} : \mathcal{T}_h \times \mathcal{X}_\pi \to \mathbb{R}$ are defined as:[5]

$$Q_h^{\mathrm{P/A},\pi}(\tau,p,a) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s_h,a)\left(r_h^{\mathrm{P/A}}(s_h,p,a,s') + V_{h+1}^{\mathrm{P/A},\pi}(\tau\oplus(p,a,s'))\right).$$

We also introduce the agent's deviation-value function $\widehat{Q}_h^{\mathrm{A},\pi} : \mathcal{T}_h \times \mathcal{X}_\pi \to \mathbb{R}$ to encode the agent's cumulative expected utility from step $h \in \mathcal{H}$ onwards under policy $\pi : \mathcal{T} \to \Delta(\mathcal{X})$, when the *agent deviates from recommended actions*. For every $\tau = (s_1, p_1, a_1, \ldots, s_{h-1}, p_{h-1}, a_{h-1}, s_h) \in \mathcal{T}_h$, contract $p \in \mathcal{P}$, and recommendation $a \in \mathcal{A}$, this is formally defined as:

$$\widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a) \coloneqq \max_{\widehat{a}\in A} \sum_{s'\in\mathcal{S}} P_h(s'|s_h,\widehat{a})\left(r_h^{\mathrm{A}}(s_h,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau\oplus(p,a,s'))\right),$$

where $\widehat{V}_h^{\mathrm{A},\pi} : \mathcal{T}_h \to \mathbb{R}$ is defined as follows for every $h \in \mathcal{H}$ and $\tau \in \mathcal{T}_h$:

$$\widehat{V}_h^{\mathrm{A},\pi}(\tau) \coloneqq \sum_{(p,a)\in\mathcal{X}_\pi} \pi(p,a|\tau)\, \widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a).$$

Intuitively, $\widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a)$ is the maximum cumulative expected utility that the agent can achieve by deviating from recommendations, once the principal has selected to the contract-recommendation pair (p,a) at step h given history τ. Notice that the cumulative expected utility achieved by deviating from step $h+1$ onwards is computed by using $\tau\oplus(p,a,s')$, *i.e.*, accounting for the recommendation a rather than the action $\widehat{a}$ actually played by the agent, as the principal cannot observe the latter.

Finally,we let $V^{\mathrm{P/A},\pi} \coloneqq \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P/A},\pi}(s)$ be the principal's (agent's) cumulative expected utility, when the agent sticks to recommended actions. Similarly, we let $\widehat{V}^{\mathrm{A},\pi} \coloneqq \sum_{s\in\mathcal{S}} \mu(s)\widehat{V}_1^{\mathrm{A},\pi}(s)$ be the expected agent's cumulative utility when they play in order to maximize it.

2.3 OPTIMAL POLICIES

Now, we are ready to formally introduce the optimization problem faced by the principal. First, we observe that, by well-known revelation-principle-style arguments (see, *e.g.*, (Gan et al., 2022)), it is possible to focus w.l.o.g. on policies that are *incentive compatible* and *direct*. The first property intuitively characterizes the policies under which the agent is "correctly" incentivized to follow recommendations, while the second one identifies policies that specify at most one contract for every possible action recommendation. These two properties are formally defined in the following.

Definition 1 (ϵ-IC policies)**.** *Let $\epsilon \geq 0$. A history-dependent policy $\pi : \mathcal{T} \to \Delta(\mathcal{X})$ is said to be* ϵ-incentive compatible (ϵ-IC) *if, for every step $h \in \mathcal{H}$ and history $\tau \in \mathcal{T}_h$, the following condition holds for every contract-recommendation pair $(p,a) \in \mathcal{X}$ such that $\pi(p,a|\tau) > 0$:*

$$Q_h^{\mathrm{A},\pi}(\tau,p,a) \geq \widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a) - \epsilon.$$

Moreover, we say that π is incentive compatible (IC) *if the condition holds for $\epsilon = 0$.*

[4] In this paper, we always work with policies $\pi : \mathcal{T} \to \Delta(\mathcal{X})$ that specify finitely-supported distributions. Thus, we can write expectations with respect to π as sums over the pairs $(p,a) \in \mathcal{X}_\pi \coloneqq \bigcup_{\tau\in\mathcal{T}} \mathrm{supp}(\pi(\tau))$, where $\mathrm{supp}(\pi(\tau))$ denotes the (finite) support of the probability distribution $\pi(\tau)$.

[5] With an abuse of notation we let $r_h^{\mathrm{P}}(s_h,p,a,s') = r_h^{\mathrm{P}}(s_h,p,s')$, albeit the principal's utility depends only on the transition and not on the agent's action.

We denote by Π the set of all the IC policies. Let us also notice that, if a policy $\pi : \mathcal{T} \to \Delta(\mathcal{X})$ is IC, then it is the case that $V_h^{\mathrm{A},\pi}(\tau) = \widehat{V}_h^{\mathrm{A},\pi}(\tau)$ for every step $h \in \mathcal{H}$ and history $\tau \in \mathcal{T}_h$.

Definition 2 (Direct policies)**.** *A policy $\pi : \mathcal{T} \to \Delta(\mathcal{X})$ is said to be* direct *if, for every history $\tau \in \mathcal{T}$ and action $a \in \mathcal{A}$, there exists at most one contract $p \in \mathcal{P}$ such that $\pi(p, a|\tau) > 0$.*

Thus, the principal's goal can be formulated as the problem of finding an *optimal* IC policy $\pi^\star \in \Pi$, which is one that achieves value OPT. This is formally defined as $\mathrm{OPT} \coloneqq \max_{\pi \in \Pi} V^{\mathrm{P},\pi}$.

3 Why do we need history-dependent policies?

As a preliminary step, we discuss why history-dependent policies are necessary. First, we show that history-dependent policies allow the principal to achieve a cumulative expected utility strictly larger than what is attainable by means of (non-stationary) Markovian policies. Intuitively, whenever there are multiple trajectories reaching a particular state, a history-dependent policy can induce the agent to follow one providing high utility to the principal. This is possible by "threatening" to do *not* pay the agent in the event that they deviate from the desired trajectory. In contrast, a Markovian policy cannot implement such a "threat", since it is *not* able to distinguish how a state has been reached. In Appendix A, we provide an extensive example in which the principal can benefit from implementing the "threats" described above, thus showing that the following proposition holds.

Proposition 1. *There exists a problem instance in which a history-dependent policy provides a principal's cumulative expected utility larger than that of any Markovian policy.*

Furthermore, we also show that computing an approximately-optimal non-stationary *Markovian* policy is NP-hard, while, as we show in the rest of this paper, it is possible to implement an approximately-optimal history-dependent policy by means of a polynomial-time algorithm.

Theorem 1. *There exists a constant $\delta > 0$ such that computing a non-stationary Markovian policy with value at least $(1-\delta)\mathrm{OPT}_{\mathrm{Mkv}}$ is NP-hard, where $\mathrm{OPT}_{\mathrm{Mkv}}$ is the principal's cumulative expected utility of an optimal non-stationary Markovian policy.*

4 Promise-form policies

As observed in Section 3, *Markovian* policies are both suboptimal and computationally intractable. Therefore, in this work, we focus on computing history-dependent policies. However, such policies pose considerable computational challenges, as representing them requires an exponential number of bits with respect to the time horizon H. To address this issue, we employ a class of policies, called *promise-form* policies, which can be efficiently represented and are guaranteed to perform as well as history-dependent ones. In this section, we first formally define promise-form policies, and then we show that there always exists an optimal promise-form policy.

A *promise-form* policy is defined as a tuple $\sigma \coloneqq \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$, where:

- $I_h : \mathcal{S} \to 2^{[0,H]}$ is a function specifying, for every state $s \in \mathcal{S}$, a finite set $I(s) \subseteq [0, H]$ of promises for step $h \in \mathcal{H}$. We assume that $|I_1(s)| = 1$ for every $s \in \mathcal{S}$ and we denote with $i(s)$ the single element belonging to $|I_1(s)|$. For ease of notation, we set $I_{H+1}(s) \coloneqq \{0\}$ for every $s \in \mathcal{S}$, and we let $\mathcal{I} \coloneqq \bigcup_{h \in \mathcal{H}} \bigcup_{s \in \mathcal{S}} I_h(s)$ be the set of all the possible promises.
- $J_h : \mathcal{S} \times \mathcal{I} \times \mathcal{A} \to 2^{\mathcal{P}}$ is a function defining, for every state $s \in \mathcal{S}$, promise $\iota \in I_h(s)$, and action $a \in \mathcal{A}$, a finite set $J_h(s, \iota, a) \subseteq \mathcal{P}$ of contracts that the principal can select at step $h \in \mathcal{H}$ when recommending action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ with promise $\iota \in \mathcal{I}$. Furthermore, we let $\mathcal{J} \coloneqq \bigcup_{h \in \mathcal{H}} \bigcup_{s \in \mathcal{S}} \bigcup_{\iota \in I_h(s)} \bigcup_{a \in \mathcal{A}} J_h(s, \iota, a)$.
- $\varphi_h : \mathcal{S} \times \mathcal{I} \to \Delta(\mathcal{J} \times \mathcal{A})$ encodes the policy at step $h \in \mathcal{H}$, with $\varphi_h(p, a|s, \iota)$ denoting the probability with which the principal selects contract $p \in J_h(s, \iota, a)$ and recommends action $a \in \mathcal{A}$, when in a state $s \in \mathcal{S}$ with promise $\iota \in \mathcal{I}_h(s)$.
- $g_h : \mathcal{S} \times \mathcal{I} \times \mathcal{J} \times \mathcal{A} \times \mathcal{S} \to \mathcal{I}$ is a promise function for step $h \in \mathcal{H}$, with $g_h(s, \iota, p, a, s') \in \mathcal{I}_{h+1}(s')$ being the promise for step $h+1$ when, at step $h \in \mathcal{H}$, the state is $s \in \mathcal{S}$, the promise is $\iota \in \mathcal{I}_h(s)$, the principal selects contract $p \in \mathcal{J}_h(s, \iota, a)$ and recommends action $a \in \mathcal{A}$, and the environment transitions to state $s' \in \mathcal{S}$.

Intuitively, under a promise-form policy, when reaching a state $s \in \mathcal{S}$ at step $h \in \mathcal{H}$, the principal promises an expected utility value $\iota \in I_h(s)$ to the agent. This represents how much the agent will gain by following the principal's recommendations from that point onwards.

Notice that, at each step $h \in \mathcal{H}$, the contract-recommendation pair selected by the principal only depends on the current state $s \in \mathcal{S}$ and the current promise $\iota \in I_h(s)$, through the function φ_h. This removes the dependency on the histories, whose number is exponential in time horizon H. However, both the cardinality of the set of promises $\mathcal{I}$ and that of the set of contracts $\mathcal{J}$ may be large in principle, and, thus, representing a promise-form policy may potentially require a prohibitive number of bits. In the rest of the paper, we show how to sidestep this issues by working with "sufficiently small" sets of promises and contracts.

Finally, let us observe that formally a promise-form $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ corresponds to a history-depend policy $\pi^\sigma : \mathcal{T} \to \Delta(\mathcal{X})$ that implements it. For every step $h \in \mathcal{H}$ and history $\tau \in \mathcal{T}_h$, such a policy is defined as $\pi^\sigma(\cdot|\tau) := \varphi_h(\cdot|s_h, \iota^\sigma_\tau)$, where $\iota^\sigma_\tau \in I_h(s_h)$ is computed by iteratively employing the function g_h over each step h in the history τ. Notice that the value and action-value functions of π^σ can be easily computed from the components that define the promise-form policy σ. Indeed, when the principal commits to a promise-form policy, their behavior at some step $h \in \mathcal{H}$ does *not* depend on the entire realized history $\tau \in \mathcal{T}_h$, but only on the current state $s \in \mathcal{S}$ and the current promise $\iota \in I_h(s_h)$. Thus, when dealing with promise-form policies, we can employ value and action-value functions that take as input only a state $s \in \mathcal{S}$ and a promise $\iota \in I_h(s)$, in place of an entire history $\tau \in \mathcal{T}_h$. The formal definitions of these promise-form value and action-value functions are deferred to Appendix B, while their equivalence to their history-dependent versions is stated by the following lemma:

Lemma 1. *Given a promise-form policy $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ and its corresponding history-dependent policy $\pi := \pi^\sigma$, for every step $h \in \mathcal{H}$, history $\tau \in \mathcal{T}_h$ ending in state $s := s_h$, action $a \in \mathcal{A}$, and contract $p \in \mathcal{P}$, the following hold:*

$$V^{\mathrm{P},\sigma}_h(s, \iota^\sigma_\tau) = V^{\mathrm{P},\pi}_h(\tau), \quad Q^{\mathrm{P},\sigma}_h(s, \iota^\sigma_\tau, p, a) = Q^{\mathrm{P},\pi}_h(\tau, p, a),$$
$$V^{\mathrm{A},\sigma}_h(s, \iota^\sigma_\tau) = V^{\mathrm{A},\pi}_h(\tau), \quad Q^{\mathrm{A},\sigma}_h(s, \iota^\sigma_\tau, p, a) = Q^{\mathrm{A},\pi}_h(\tau, p, a),$$
$$\widehat{V}^{\mathrm{A},\sigma}_h(s, \iota^\sigma_\tau) = \widehat{V}^{\mathrm{A},\pi}_h(\tau), \quad \widehat{Q}^{\mathrm{A},\sigma}_h(s, \iota^\sigma_\tau, p, a) = \widehat{Q}^{\mathrm{A},\pi}_h(\tau, p, a).$$

4.1 Honesty

We now introduce a subclass of promise-form policies in which the principal *honestly approximately fulfills the promises* made to the agent, which we call η-honest policies for some $\eta \geq 0$.

Definition 3 (η-Honesty)**.** *A promise-form policy $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ is* η-honest*, with $\eta \geq 0$, if the following holds for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, and promise $\iota \in I_h(s)$:*

$$\left| \sum_{(p,a) \in \mathcal{X}_\sigma} \varphi_h(p, a|s, \iota) \sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^{\mathrm{A}}_h(s, p, a, s') + g_h(s, \iota, p, a, s') \right) - \iota \right| \leq \eta.$$

Intuitively, an η-honest policy ensures that, at the last time step H, the difference between the promise and the actual agent's expected utility is at most η. Thus, by employing an inductive argument, it is possible to prove the following lemma.

Lemma 2. *Let $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ be an η-honest promise-form policy. Then, for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$ it holds that $\left| V^{\mathrm{A},\sigma}_h(s, \iota) - \iota \right| \leq \eta(H - h + 1)$ for every $\iota \in I_h(s)$.*

We also introduce a crucial lemma showing that if an η-honest promise-form policy satisfies a suitable a set of "local"constraints, then such a policy is also approximately IC.

Lemma 3. *Let $\zeta > 0$ and $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ be an η-honest promise-form policy such that the following constraint holds for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, promise $\iota \in I_h(s)$, action $a, \widehat{a} \in \mathcal{A}$, and contract $p \in \mathcal{P}$ such that $\varphi_h(p, a|s, \iota) > 0$:*

$$\begin{aligned} &\sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^{\mathrm{A}}_h(s, p, a, s') + g_h(s, \iota, p, a, s') \right) \geq \\ &\qquad \sum_{s' \in \mathcal{S}} P_h(s'|s, \widehat{a}) \left(r^{\mathrm{A}}_h(s, p, \widehat{a}, s') + g_h(s, \iota, p, a, s') \right) - \zeta. \end{aligned} \tag{1}$$

Then, the promise-form policy σ is $(2\eta\zeta H^2)$-IC.

Thanks to Equation 1, we can require a policy to be approximately incentive compatible by only imposing a "local" constraint. Indeed, for every state $s \in \mathcal{S}$ and time step $h \in \mathcal{H}$, Equation 1 depends only on the current promise $\iota \in I_h(s)$ and the components that define the promise-form policy σ. Finally, given the previous results, we can prove that promise-form policies represent a subclass of history-dependent policies that is powerful enough to include the optimum.

Theorem 2. *There always exists an honest IC promising-form policy that achieves a principal's expected cumulative reward equal to* OPT.

5 Approximation Policy

In this section, we develop a polynomial-time algorithm to compute a promise-form policy σ that achieves a value $V^{\mathrm{P},\sigma} \geq \mathrm{OPT}$, while being ϵ-IC for a given ϵ. Let us recall that the value of the policy is computed assuming that the agent sticks to recommendations, even if the policy is not exactly IC.

5.1 Discretization method

We employ a finite set of promises $\mathcal{I} \subseteq \mathcal{D}_\delta := \{k\delta \mid 0 \leq k \leq \lfloor HB/\delta \rfloor\}$, where δ is a suitable discretization step. This ensures that σ will be polynomially sized so long as δ is polynomial itself.

Our approach is based on dynamic programming (Algorithm 2). Starting from $h = H$, we build the components $\{\varphi_h, J_h, g_h\}$ defining our our policy on step at a time. The algorithm keeps track of $H+1$ tables $M_h^\delta : \mathcal{S} \times \mathcal{D}_\delta \to \mathbb{R} \cup \{-\infty\}$, where $M_h^\delta(s,\iota)$ encodes an approximation of $V_h^{\mathrm{P},\sigma}(s,\iota)$. A value of $-\infty$ indicates that the principal cannot promise ι honestly, therefore $\iota \notin I_h(s)$.

To fill the table M_h and compute $\{\varphi_h, J_h, g_h\}$ at step h, Algorithm 2 computes a tuple $(\alpha, \{p_a\}_{a\in\mathcal{A}}, z, v)$ for every promise $\iota \in \mathcal{D}_\delta$ and state $s \in \mathcal{S}$. Specifically, $\alpha \in \Delta(\mathcal{A})$ is a probability distribution over the agent's actions, where α_a is the probability of recommending action $a \in \mathcal{A}$, while p_a is the contract to incentivize it. The function $z : \mathcal{A}\times\mathcal{S} \to \mathcal{D}_\delta$ specifies the future promises $g_h(s,\iota,p_a,a,s') = z(a,s')$ for every $a \in \mathcal{A}$ and $s' \in \mathcal{S}$. Finally, v is the value to be stored in $M_h^\delta(s,\iota)$.

Algorithm 2 `Approximation policy`

Require: $\delta \in (0,1)$.
1: $M_{H+1}^\delta(s,0) \leftarrow 0 \quad \forall s \in \mathcal{S}$
2: $M_{H+1}^\delta(s,\iota) \leftarrow -\infty \quad \forall s \in \mathcal{S}, \iota \in \mathcal{D}_\delta \setminus \{0\}$
3: **for all** $h \in \mathcal{H}, s \in \mathcal{S}$ **do**
4: $\quad I_h(s) \leftarrow 0$
5: $\quad$ **for all** $\iota \in \mathcal{D}_\delta$ **do**
6: $\quad\quad (\alpha, \{p_a\}_{a\in\mathcal{A}}, z, v) \leftarrow O_{\iota,\delta}^{h,s}(M_{h+1}^\delta)$
7: $\quad\quad M_h^\delta(s,\iota) \leftarrow v$
8: $\quad\quad$ **for all** $a \in \mathcal{A} : \alpha_a > 0$ **do**
9: $\quad\quad\quad \varphi_h(p_a, a|s,\iota) \leftarrow \alpha_a$
10: $\quad\quad\quad J_h(s,\iota,p_a,a) \leftarrow \{p_a\}$
11: $\quad\quad\quad g_h(s,\iota,p_a,a,s') \leftarrow z(a,s') \; \forall s' \in \mathcal{S}$
12: $\quad\quad$ **if** $v > -\infty$ **then**
13: $\quad\quad\quad I_h(s) \leftarrow I_h(s) \cup \{\iota\}$
14: **for all** $s \in \mathcal{S}$ **do**
15: $\quad I_1(s) \leftarrow \arg\max_{\iota \in I_1(s)} M_1^\delta(s,\iota)$
16: **Return** $\{(I_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$

Crucially, the oracle employed at Line 6 is responsible for ensuring that σ is approximately IC, and that its value when the agent sticks to the recommendations is at least OPT. We now describe the properties of the solution (α, p, z, v) computed by this oracle, and leave its implementation to Section 5.2. The tuples (α, p, z) computed by the approximate oracle belongs to suitably search space $\Psi_{h,s}^\iota(\lambda,\zeta)$ defined by the following constraints:

$$\left| \sum_{a\in A} \alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^A(s,p_a,a,s') + z(a,s') \right) - \iota \right| \leq \lambda \tag{2}$$

$$\alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^A(s,p_a,a,s') + z(a,s') \right) \geq \alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^A(s,p_a,\widehat{a},s') + z(a,s') \right) - \alpha_a\zeta \quad \forall a,\widehat{a} \in \mathcal{A} \tag{3}$$

Intuitively, Equation 2 corresponds to honesty with a tolerance $\lambda > 0$. Equation 3 corresponds instead to Equation 1, and is crucial to ensure that σ is approximately IC by Lemma 3. Observe that since

every term in Equation 3 multiplies α_a, the equation is satisfied when $\alpha_a = 0$, *i.e.* when action a is never recommended. We also introduce the following function for every state $s \in \mathcal{S}$, step $h \in \mathcal{H}$:

$$F_{h,s,M_{h+1}}(\alpha, p, z) \coloneqq \sum_{a \in \mathcal{A}} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^P(s, p_a, s') + M_{h+1}(s, z(a,s')) \right),$$

which represents an approximation of the principal's utility $V_h^{\mathrm{P},\sigma}(s,\iota)$ if the tuple (α, p, z) is used to define σ at state $s \in \mathcal{S}$ and step $h \in \mathcal{H}$. We are now ready to provide the definition of *approximation oracle* that $O_{\iota,\delta}^{h,s}(M_{h+1}^{\delta})$ must satisfy whenever it is executed.

Definition 4 (Approximation oracle)**.** *An algorithm $O_{\iota,\delta}^{h,s}(M_{h+1}^{\delta})$ is an approximation oracle if it returns a tuple (α, p, z, v), with $\alpha \in \Delta(\mathcal{A}), p \in \mathcal{P}^{|\mathcal{A}|}, z : \mathcal{A} \times \mathcal{S} \to \mathcal{D}_\delta, v \in \mathbb{R} \cup \{-\infty\}$, such that:*

1. *for every $(\bar{\alpha}, \bar{p}, \bar{z}) \in \Psi_{h,s}^{\iota}(\lambda = \delta, \zeta = 2\delta)$, it holds $v \geq F_{h,s,M}(\bar{\alpha}, \bar{p}, \bar{z})$.*
2. *if $v > -\infty$, then $F_{h,s,M}(\alpha, p, z) \geq v$.*
3. *if $v > -\infty$, then $(\alpha, p, z) \in \Psi_{h,s}^{\iota}(\lambda = 2\delta, \zeta = 4\delta)$ and $F_{h,s,M}(\alpha, p, z) > -\infty$.*

We now give an overview on how this definition allows us to prove that Algorithm 2 computes an almost IC policy with value at least OPT (the full proofs are deferred to Appendix C).

We first establish the following relationship between the tables $M_h(s,\iota)$ and the optimal policy.[6]

Lemma 4. *Let $\sigma^\star = \{(I_h^\star, J_h^\star, \varphi_h^\star, g_h^\star)\}_{h \in \mathcal{H}}$ be an honest, direct, IC and optimal promise-form policy, and M_h^δ, I_h be computed accordingly to Algorithm 2. For every $h \in \mathcal{H}$, $s \in \mathcal{S}$, and $\iota \in I_h^\star(s)$, it holds that $M_h^\delta(s, \lceil \iota \rceil_\delta) \geq V_h^{P,\sigma^\star}(s,\iota)$ and $M_h^\delta(s, \lfloor \iota \rfloor_\delta) \geq V_h^{P,\sigma^\star}(s,\iota)$.*

The lemma above holds because for any $h \in \mathcal{H}, s \in \mathcal{S}, \iota \in I_h^{\sigma^\star}(s)$, we can build $(\bar{\alpha}, \bar{p})$ by using the contracts and recommendations prescribed by $\varphi_h^\star(s,\iota)$, and $\bar{z}$ by discretizing the future promises prescribed by $g^\star$. The resulting tuple $(\bar{\alpha}, \bar{p}, \bar{z})$ belongs to $\Psi_{h,s}^{\lfloor \iota \rfloor_\delta}$. Since $M(s, \lfloor \iota \rfloor_\delta)$ is filled with the value v provided by the oracle, Property 1 proves the statement. A similar argument holds for $\lceil \iota \rceil_\delta$.

We now link the values of the tables M to those of the policy σ computed by Algorithm 2.

Lemma 5. *Let $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ be the promise-form policy computed by Algorithm 2. For every $h \in \mathcal{H}$, $s \in \mathcal{S}$, and $\iota \in I_h(s)$, it holds that $V_h^{\mathrm{P},\sigma}(s,\iota) \geq M_h^\delta(s,\iota)$.*

The proof of Lemma 5 relies on Property 2. At the last step $h = H$, we have $V_h^{\mathrm{P},\sigma}(s,\iota) = F_{h,s,M_{h+1}}(\alpha, p, z)$, where the tuple (α, p, z, v) is that computed by the oracle. Therefore the value $v \leq F_{h,s,M_{h+1}}(\alpha, p, z)$ stored in $M(s,\iota)$ is smaller than the actual value of σ. Proceeding by induction over the steps $h < H$ concludes the proof.

Finally, we observe that if in state $s \in \mathcal{S}$ the optimal policy $\sigma^\star$ employs promise ι, while our algorithm selects $\bar{\iota}$, we have:

$$V_1^{\mathrm{P},\sigma}(s,\bar{\iota}) \geq M_1(s,\bar{\iota}) \geq M_1(s, \lfloor \iota \rfloor_\delta) \geq V_h^{\mathrm{P},\sigma^\star}(s,\iota),$$

where the first inequality holds by Lemma 5, the second by construction (Line 15), and the last one by Lemma 4. We conclude that $V^{\mathrm{P},\sigma} \geq V^{\mathrm{P},\sigma^\star} = \mathrm{OPT}$. Thanks to Property 3 we can also show that σ is also approximately IC, proving the following theorem.

Theorem 3. *For any $\epsilon > 0$, Algorithm 2 instantiated with $\delta = \epsilon/16H^2$ computes an ϵ-IC promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ in time polynomial in the instance size and $1/\epsilon$. Furthermore, σ is such that $V^{P,\pi^\sigma} \geq \mathrm{OPT}$, while $\mathcal{I} = \mathcal{O}(H^4 B/\epsilon)$ and $\mathcal{J} = \mathcal{O}(H^4 B|\mathcal{S}||\mathcal{A}|/\epsilon)$.*

5.2 Approximation oracle

We now describe Algorithm 3, which provides an approximation oracle according to Definition 4 by solving a new optimization problem. For the ease of notation, we introduce the set of promises $\mathcal{D}_\delta^{M,s} \coloneqq \{\iota \in \mathcal{D}_\delta \mid M(s,\iota) > -\infty\}$ that can be kept in state $s \in \mathcal{S}$.

[6] We let $\lfloor \widetilde{\iota} \rfloor_\delta \coloneqq \arg\min_{\iota \in \mathcal{D}_\delta, \iota \leq \widetilde{\iota}} |\iota - \widetilde{\iota}|$ and $\lceil \widetilde{\iota} \rceil_\delta \coloneqq \arg\min_{\iota \in \mathcal{D}_\delta, \iota \geq \widetilde{\iota}} |\iota - \widetilde{\iota}|$ be the lower (upper) discretization of $\widetilde{\iota}$.

We define the optimization problem $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ over variables $\alpha \in \Delta(\mathcal{A})$, $p \in \mathcal{P}^{|\mathcal{A}|}$, and $\widetilde{q} : \mathcal{A} \times \mathcal{S} \to \Delta(\mathcal{D}_\delta)$. Intuitively, instead of discrete promises $z(a, s')$, we implicitly use continuous promises $q(a, s') \coloneqq \sum_{\iota' \in \mathcal{D}^{M,s'}_{\delta}} \widetilde{q}(\iota'|a, s')\iota'$, defined as weighted sums of the discrete ones. We further define $y(a, s') \coloneqq \sum_{\iota' \in \mathcal{D}^{M,s'}_{\delta}} \widetilde{q}(\iota'|a, s')M(s', \iota')$, extending the table M to these continuous promises. Given these definitions, we define the optimization problem $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ and its value $G^{\delta}_{h,s,M}(\alpha, p, \widetilde{q})$ as:

$$\widehat{G}^{\delta}_{h,s,\iota}(M) \coloneqq \max_{\substack{\alpha \in \Delta(\mathcal{A}),\ p \in \mathcal{P}^{|\mathcal{A}|} \\ \widetilde{q}:\mathcal{A}\times\mathcal{S}\to\Delta(\mathcal{D}_\delta)}} G^{\delta}_{h,s,M}(\alpha, p, \widetilde{q}) \coloneqq \sum_{a\in\mathcal{A}} \alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r^P_h(s, p^a, s') + y(a, s')\right),$$

$$\text{s.t. } (\alpha, p, q) \in \Psi^{\iota}_{h,s}(\lambda = \delta, \zeta = 2\delta) \text{ and } \widetilde{q}(a, s') \in \Delta(\mathcal{D}^{M,s'}_{\delta}) \ \forall a \in \mathcal{A}, s \in \mathcal{S}.$$

Algorithm 3 `Approximation oracle`

Require: h, s, δ, ι, M.
1: **if** $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ is feasible **then**
2: $(\alpha, p, \widetilde{q}) \leftarrow$ Solution to $\mathcal{R}^{\delta}_{h,s,\iota}(M)$
3: $v \leftarrow G^{\delta}_{h,s,M}(\alpha, p, \widetilde{q})$
4: $z(a, s') \leftarrow$ Equation 4
5: **else**
6: $(\alpha, p, z) \leftarrow$ anything, $v \leftarrow -\infty$
7: **Return** (α, p, z, v)

While $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ is a quadratic problem, it is equivalent to an LP thanks to its peculiar structure. We refer the reader to Appendix D for details. Therefore, it can be solved in polynomial time. Algorithm 3 solves such an optimization problem and recovers discrete promises as follows:

$$z(a, s') = d^{M,s'}_{\delta}(q(a, s')) \coloneqq \operatorname*{arg\,max}_{\iota' \in \mathcal{D}^{M,s'}_{\delta} : |\iota' - q(a,s')| < \delta} M(s', \iota') \quad \forall a \in \mathcal{A}, s' \in \mathcal{S}. \tag{4}$$

Finally, Algorithm 3 returns the tuple (α, p, z, v), where $v = \widehat{G}^{\delta}_{h,s,\iota}(M)$.

Lemma 6. *If Algorithm 3 is used as an approximate oracle $O^{h,s}_{\iota,\delta}$ during the entire execution of Algorithm 2, then it satisfies Definition 4.*

To prove the lemma, we observe that v is larger than the value of any feasible solution of $\mathcal{R}^{\delta}_{h,s,\iota}(M)$, hence Algorithm 3 satisfies Property 1. Furthermore, we prove Property 2 by some concavity properties of the function $G^{\delta}_{h,s,M}$ (see Appendix D). Finally, we have $(\alpha, p, q) \in \Psi^{\iota}_{h,s}(\delta, 2\delta)$ by construction, while Equation 4 changes the promises by at most δ, leading to $(\alpha, p, z) \in \Psi^{\iota}_{h,s}(2\delta, 4\delta)$.

6 From approximately-IC policies to an approximate optimum

Algorithm 2 computes an ϵ-IC promise-form policy σ that achieves a principal's cumulative expected utility of at least OPT. In this section we show how to derive an IC policy with value almost OPT. In the classical one-shot problem, (Dütting et al., 2019) propose a method to transform ϵ-IC contract-recommendation pair (a, p) into an IC one. It is sufficient to compute new payments as a linear combination of the original ones and the principal reward for each outcome of the game. The pair (p', a') composed of this new contract and the action it incentivizes, possibly different from a, is IC and provides almost the same utility as of (a, p).

We generalize this approach to promise-form policies in MDPs. Given an ϵ-IC promise-form policy σ, for every $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota \in I_h(s)$, our procedure substitutes every contract p proposed by σ with a contract p' defined as:

$$p'(s') = (1 - \sqrt{\epsilon})p(s') + \sqrt{\epsilon}r_h(s, s') \forall s' \in \mathcal{S}.$$

Then, the procedure performs two backward passes to recompute the best-responses of the agent to this new contracts, which become the new recommendations, and the agent's expected utilities that substitute the promises. We provide here the main result, and leave the details to Appendix E.

Theorem 4. *Given a direct and ϵ-IC promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ that achieves at least $V^{\mathrm{P},\sigma} \geq$ OPT, Algorithm 4 (detailed in Appendix E) computes an IC and honest promise-form policy σ^3 in time polynomial in the size of σ and the instance. Furthermore, the value of σ^3 is at least $V^{\mathrm{P},\sigma^3} \geq \mathrm{OPT} - (H + 1)\sqrt{\epsilon}$, while its size is such that $|\mathcal{I}^3| \leq |\mathcal{I}|$ and $|\mathcal{J}^3| \leq |\mathcal{J}|$.*

While the technique is similar to the one proposed by Dütting et al. (2019), the analysis requires to consider the long term interaction between the principal and the agent. Specifically, in Appendix E we observe that for every possible trajectory induced by policy σ when the agent sticks to recommendations, there is another trajectory induced by σ^3 that provides a similar utility.

AI USE STATEMENT

In this work we have not used generative AI tools.

REFERENCES

Paola Alimonti and Viggo Kann. Some apx-completeness results for cubic graphs. *Theoretical Computer Science*, 237(1-2):123–134, 2000.

Tal Alon, Paul Dütting, and Inbal Talgam-Cohen. Contracts with private cost per unit-of-effort. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 52–69, 2021.

Tal Alon, Paul Duetting, Yingkai Li, and Inbal Talgam-Cohen. Bayesian analysis of linear contracts. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, pp. 66, 2023.

Hamsa Bastani, Mohsen Bayati, Mark Braverman, Ramki Gummadi, and Ramesh Johari. Analysis of medicare pay-for-performance contracts. *Available at SSRN 2839143*, 2016.

Hamsa Bastani, Joel Goh, and Mohsen Bayati. Evidence of upcoding in pay-for-performance programs. *Management Science*, 65(3):1042–1060, 2019.

Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, and Mirco Mutti. Persuading farsighted receivers in mdps: the power of honesty. *Advances in Neural Information Processing Systems*, 36, 2023.

Martino Bernasconi, Matteo Castiglioni, and Alberto Marchesi. Regret-minimizing contracts: Agency under uncertainty. *arXiv preprint arXiv:2402.13156*, 2024.

Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997. ISBN 1886529191.

Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*, pp. 280–296. PMLR, 2015.

Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Bayesian agency: Linear versus tractable contracts. *Artificial Intelligence*, 307:103684, 2022.

Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Designing menus of contracts efficiently: The power of randomization. *Artificial Intelligence*, 318:103881, 2023a.

Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Multi-agent contract design: How to commission multiple agents with individual outcomes. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 412–448, 2023b.

Lin William Cong and Zhiguo He. Blockchain disruption and smart contracts. *The Review of Financial Studies*, 32(5):1754–1797, 2019.

Paul Dütting, Tim Roughgarden, and Inbal Talgam-Cohen. Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pp. 369–387, 2019.

Paul Dutting, Tim Roughgarden, and Inbal Talgam-Cohen. The complexity of contracts. *SIAM Journal on Computing*, 50(1):211–254, 2021.

Paul Dütting, Tomer Ezra, Michal Feldman, and Thomas Kesselheim. Combinatorial contracts. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 815–826. IEEE, 2022.

Jiarui Gan, Minbiao Han, Jibang Wu, and Haifeng Xu. Optimal coordination in generalized principal-agent problems: A revisit and extensions. *arXiv preprint arXiv:2209.01146*, 2022.

Jiarui Gan, Rupak Majumdar, Debmalya Mandal, and Goran Radanovic. Stochastic principal-agent problems: Computing and learning optimal history-dependent policies. *Advances in Neural Information Processing Systems*, 38:156800–156825, 2026.

Guru Guruganesh, Jon Schneider, and Joshua R Wang. Contracts under moral hazard and adverse selection. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 563–582, 2021.

Guru Guruganesh, Jon Schneider, Joshua Wang, and Junyao Zhao. The power of menus in contract design. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, pp. 818–848, 2023.

Dylan Hadfield-Menell and Gillian K Hadfield. Incomplete contracting and ai alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 417–422, 2019.

Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pp. 359–376, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450325653.

Dima Ivanov, Paul Dütting, Inbal Talgam-Cohen, Tonghan Wang, and David C Parkes. Principal-agent reinforcement learning. *arXiv preprint arXiv:2407.18074*, 2024.

Nur Kaynar and Auyon Siddiq. Estimating effects of incentive contracts in online labor platforms. *Management Science*, 69(4):2106–2126, 2023.

Eden Saig, Ohad Einav, and Inbal Talgam-Cohen. Incentivizing quality text generation via statistical contracts. *arXiv preprint arXiv:2406.11118*, 2024.

Tonghan Wang, Paul Duetting, Dmitry Ivanov, Inbal Talgam-Cohen, and David C Parkes. Deep contract design via discontinuous networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Jibang Wu, Siyu Chen, Mengdi Wang, Huazheng Wang, and Haifeng Xu. Contractual reinforcement learning: Pulling arms with invisible hands. *arXiv preprint arXiv:2407.01458*, 2024.

Banghua Zhu, Stephen Bates, Zhuoran Yang, Yixin Wang, Jiantao Jiao, and Michael I Jordan. The sample complexity of online contract design. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 1188–1188, 2023.

APPENDIX

A MARKOVIAN POLICIES

In this section we introduce Markovian policies and show wh they are not the "right" type of policies to tackle our problem. First, we introduce the notation regarding Markovian policies. Then, we propose an instance where no Markovian policy is optimal, and thus a history-dependent policy must be considered. Finally, we prove that approximating the optimal Markovian policy is NP-hard.

A.1 MARKOVIAN POLICIES

A Markovian (time-inhomogeneous) policy prescribes the current contract p and action a depending only the current state s and time step h. W.l.o.g., we restrict our attention to deterministic Markovian policies, as the optimal one is deterministic. Formally, a Markovian policy $\rho \coloneqq \{p_h^s, a_h^s\}_{s\in\mathcal{S},h\in\mathcal{H}}$ prescribes contract p_h^s and action a_h^s at step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$.

As for history-dependent policies, we defined the value-functions for the principal and the agent. The principal's value function $V^{\mathrm{P},\rho} : \mathcal{S} \times \mathcal{H}$ returns, given a state $s \in \mathcal{S}$ and step $h \in \mathcal{H}$, the principal's expected cumulative utility from state s and step h, formally:

$$V_h^{\mathrm{P},\rho}(s) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s, a_h^s) \left(r_h^{\mathrm{P}}(s, p_h^s, s') + V_{h+1}^{\mathrm{P},\rho}(s') \right).$$

Similarly, the agent's value function $V^{\mathrm{P},\rho} : \mathcal{S} \times \mathcal{H}$ returns, given a state $s \in \mathcal{S}$ and step $h \in \mathcal{H}$, the agent's expected cumulative utility from state s and step h, formally:

$$V_h^{\mathrm{A},\rho}(s) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s, a_h^s) \left(r_h^{\mathrm{A}}(s, p_h^s, a_h^s, s') + V_{h+1}^{\mathrm{A},\rho}(s') \right).$$

We will only consider IC Markovian policies, *i.e.*, policies $\rho \coloneqq \{p_h^s, a_h^s\}_{s\in\mathcal{S},h\in\mathcal{H}}$ such that for every state $s \in \mathcal{S}$ and step $h \in \mathcal{H}$, the action a_h^s maximizes the agent's expected cumulative utility from state s and step h.[7] As a result, a IC Markovian policy must satisfy the following:

$$\sum_{s'\in\mathcal{S}} P_h(s'|s, a_h^s) \left(r_h^{\mathrm{A}}(s, p_h^s, a_h^s, s') + V_{h+1}^{\mathrm{A},\rho}(s') \right) \geq \sum_{s'\in\mathcal{S}} P_h(s'|s, \widehat{a}) \left(r_h^{\mathrm{P}}(s, p_h^s, \widehat{a}, s') + V_{h+1}^{\mathrm{A},\rho}(s') \right),$$

for every state $s \in \mathcal{S}$, step $h \in \mathcal{H}$, and action $\widehat{a} \in \mathcal{A}$. We denote with ϱ the set of IC Markovian policies.

Finally, the expected principal's cumulative utility of a Markovian policy $\rho \coloneqq \{p_h^s, a_h^s\}_{s\in\mathcal{S},h\in\mathcal{H}}$ is defined as:

$$V^{\mathrm{P},\rho} \coloneqq \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\rho}(s),$$

while the optimal Markovian policy $\rho^\star$ maximizes the expected principal's cumulative utility, formally:

$$\rho^\star \coloneqq \arg\max_{\rho\in X} V^{\mathrm{P},\rho}$$

[7] Observe that, as long as the agent break ties in favor of the principal when indifferent between multiple actions, a persuasive Markovian policy does not need to recommend an action and can prescribes the contract only.

A.2 MARKOVIAN POLICIES ARE SUBOPTIMAL

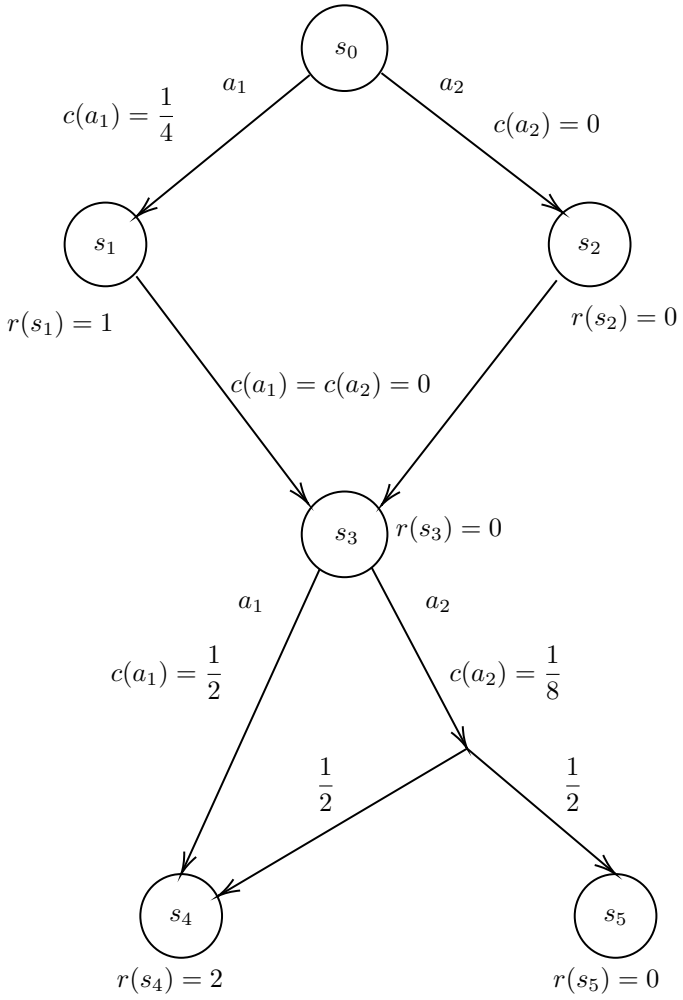

Figure 1: An instance where no Markovian policy is optimal.

Proposition 1. *There exists a problem instance in which a history-dependent policy provides a principal's cumulative expected utility larger than that of any Markovian policy.*

Proof. Consider the instance depicted in Figure 1. The initial state is s_0 and the time horizon is $H = 3$. In this example every state can only be reached at a single time step, thus we will drop the subscript h when denoting functions that depend on them. Furthermore, rewards are not normalized in the set $[0, 1]$ to ease the exposition.

From state s_0, the agent can play action a_1 with cost $1/4$ to reach state s_2, or action a_2 with cost zero to reach state s_2. Furthermore, the principal collects a reward of one is state s_1 is reached and of zero if, instead, the state transitions to s_2. From state s_1 and s_2 the agent can only reach state s_3 with no cost. For the sake of simplicity, we assume that the set of actions depends on the state, and that a single action a_1 is available in this states. From state s_3 the agent can play either a_1 with cost $1/2$ or a_2 with cost $1/8$. We further assume the agent can play an action a_3 without cost and providing zero reward to the principal. This action is not depicted in Figure 1 and it is only required to guarantee a cost 0 action and individual Rationality. States s_4 and s_5 are "terminal states", in the sense that they are reached at the end of the time horizon and no action is taken in them.

We observe that the agent chooses an action only in states s_0 and s_3, while in states s_1 ad s_2 there is a single available action. Furthermore, given a policy that in state s_1 proposes a contract with payment

$p^{s_1}(s_3) = k > 0$, we can build an equivalent policy by increasing the previous payment $p^{s_0}(s_1)$ by k and setting $p^{s_1}(s_3)$ to zero. A similar argument holds for state s_2. As a result, we can consider policies that do not pay in states s_1 and s_2, *i.e.* they propose a contract with $p^{s_1}(s_3) = p^{s_2}(s_3) = 0$.

First, we upperbound the principal's utility achievable with Markovian policies. Against a Markovian policy in this instance, the agent acts myopically maximizing their immediate utility. Indeed, the first choice of the agent is in state s_0 and impacts the utilities at step $h = 1$. Then, regardless of this choice, both the principal and the agent collects rewards of zero at step $h = 2$, and the state reaches s_3 at step $h = 3$. Finally, the Markovian policy ignores how state s_3 has been reached. As a result, the agent can act in order to maximize their immediate utility in state s_1, as this action does not impact the next choice at state s_3, and in state s_3, which is reached at the end of the time horizon. Hence, the optimal Markovian policy $\rho \coloneqq \{p^s, a^s\}_{s\in\mathcal{S}}$ maximizes the principal's immediate utility in each state. In particular, such a policy prescribes action $a^{s_0} = a_1$ in state s_0, with payment $p^{s_0}(s_1) = 1/4$ if the future state is s_1 and $p^{s_0}(s_2) = 0$ if the future state s_2. In state s_3, the action a_1 is incentivized by a contract p^{s_3} such that:

$$p^{s_3}(s_4) - \frac{1}{2} \geq \frac{1}{2}p^{s_3}(s_5) + \frac{1}{2}p^{s_3}(s_5) - \frac{1}{8}.$$

The inequality above implies:

$$p^{s_3}(s_4) \geq \frac{3}{4} + p^{s_3}(s_5).$$

Thus, the contract to incentivize action a_1 with the minimum payment is the one that pays $p^{s_3}(s_4) = 3/4$ for the terminal state s_4 and zero for any other. With such a contract, the principal collects an immediate utility of $V^{\mathrm{P},\rho}(s_3) = 5/4$. This is optimal, as any other agent's action would provide an immediate utility smaller than 1. Overall, the expected cumulative of the principal is $V^{\mathrm{P},\rho}(s_0) = 2$.

Now, we design an history-dependent policy with strictly larger utility, concluding the proof. We observe that while s_3 is always reached, the contract and hence the agent's action in s_3 depend on the agent's action in s_0. Consequently, the agent does not necessarily play in order to maximize the immediate utility in state s_0, since the action in this state influences the following choice. This can be exploited by a history-dependent policy in order to reduce the payment in state s_0 with respect to the optimal Markovian policy.

Let p_0 be the contract that does not provide any payment, *i.e.*, $p_0(s') = 0$ for every $s' \in \mathcal{S}$. We define the following two histories:

$$\tau_1 \coloneqq (s_0, p_0, a_1, s_1, p_0, a_1, s_3)$$
$$\tau_2 \coloneqq (s_0, p_0, a_1, s_2, p_0, a_1, s_3).$$

Intuitively, in both histories the principal recommends action a_1 in state s_0. History τ_1 corresponds to the case in which the agent plays action a_1 in state s_0, while history τ_2 encodes the case in which the agent ignores the recommendation and plays a_2. Thus, these two histories represent the two paths through which the agent can reach state s_3.

We design a deterministic history-dependent policy π that, in state s_0, prescribes action a_1 and does not provide any payment. If the state s_3 is reached through history τ_1, the policy prescribes the same contract proposed by the Markovian policy in s_3 and recommend action a_1. Instead, it does not provide any payment for history τ_2, while it recommends action a_3.

We can show that the policy π is IC. Indeed, as shown before, the agent is incentivized to play action a_1 in state s_1 when the history is τ_1. This provides an immediate utility:

$$V^{\mathrm{A},\pi}(\tau_1) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}.$$

Instead, if the history is τ_2, then the agent plays action a_3 with no cost and no payment, as any other action incurs in a cost with no payment. Thus:

$$V^{\mathrm{A},\pi}(\tau_2) = 0.$$

Furthermore, if the agent plays a_1 in s_0, then it collects a cumulative utility of:

$$V^{\mathrm{A},\pi}(\tau_0) = V^{\mathrm{A},\pi}(\tau_1) - \frac{1}{4} = 0,$$

where $\tau_0 \coloneqq (s_0)$ is the initial history. Instead, by playing action a_2 in s_0, the agent collects a cumulative utility of zero. As a result, by recommending action a_1 in s_0 the policy π is IC.

Finally, the policy π provides the principal with a cumulative utility of:

$$V^{\mathrm{P},\pi}(\tau_0) = 1 + V^{\mathrm{P},\pi}(\tau_1) = 1 + 2 - \frac{3}{4} = \frac{9}{4} > 2 = V^{\mathrm{P},\rho}(s_0),$$

concluding the proof. □

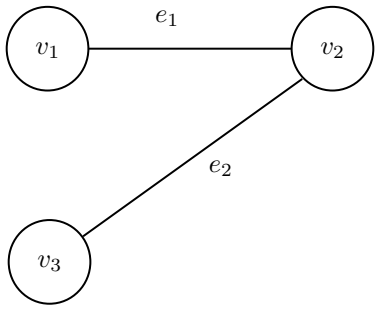

Figure 2: Illustrative instance of a cubic graph with $V = \{v_1, v_2, v_3\}$ and $E = \{e_1, e_2\}$.

Theorem 1. *There exists a constant $\delta > 0$ such that computing a non-stationary Markovian policy with value at least $(1-\delta)\mathrm{OPT}_{\mathrm{Mkv}}$ is NP-hard, where $\mathrm{OPT}_{\mathrm{Mkv}}$ is the principal's cumulative expected utility of an optimal non-stationary Markovian policy.*

Proof. We reduce from an NP-Hard problem related to find vertex covers in cubic graphs. In particular, Alimonti & Kann (2000) show that there exists a constant $\epsilon > 0$ such that the following promise problem is NP-Hard: Given a graph (V, E) in which each nodes has degree at most 3, distinguish whether there is a vertex cover of size k or all the vertex covers have size at least $k(1+\epsilon)$.

Construction. Given an graph (V, E), we build an instance as follows. In the following, for the ease of exposition we assume that the set of actions depends on the state. There is a state s_e for each edge $e \in E$. The agent has three actions in each state s_e, $e = (v_1, v_2)$:

- $a_\varnothing$ that leads to state $s_\varnothing$ that has reward $r(s_e, s_\varnothing) = 0$ and cost $c(s_e, a_1) = 0$,
- a_{v_1} that has cost $c(s_e, a_1) = 1/4$ and leads deterministically to node s_{v_1},
- a_{v_2} that has cost $c(s_e, a_1) = 1/4$ and leads deterministically to node s_{v_2}.

In state $s_\varnothing$, there is a single action a_1 with cost $c(s_\varnothing, a_1) = 0$ that leads to a state $\hat{s}_\varnothing$.

There exists a state $\hat{s}$. In this state, the agent has only an action a_1 available with cost $c(\hat{s}, a_1) = 0$ that leads to an uniform distribution over s_v, $v \in V$. Finally, for each $v \in V$, there is a state s_v with two available actions:

- a_1 that leads deterministically to s_1 and has cost $c(s_v, a_1) = \frac{1}{2}$,
- a_2 that lead to s_1 with probability $1/2$ and to s_2 with probability $1/2$ and has cost $c(s_v, a_2) = 1/8$,
- a_3 that leads deterministically to s_2 and has cost $c(s_v, a_3) = 0$.

The time horizon is $H = 3$, and hence the interaction ends when the agent enters in state s_1 or s_2. Hence, it is not required to specified the action in these states. Moreover, since each state can be reach only in a specific time step, we remove this dependency.

Then, we define the rewards. Since they depend only on the reached state, we drop the dependency on the initial state, *i.e.*, $r(s') = r(s, s')$ for all $s, s' \in \mathcal{S}$. The reward reaching a state s_v, $v \in V$, is 1,

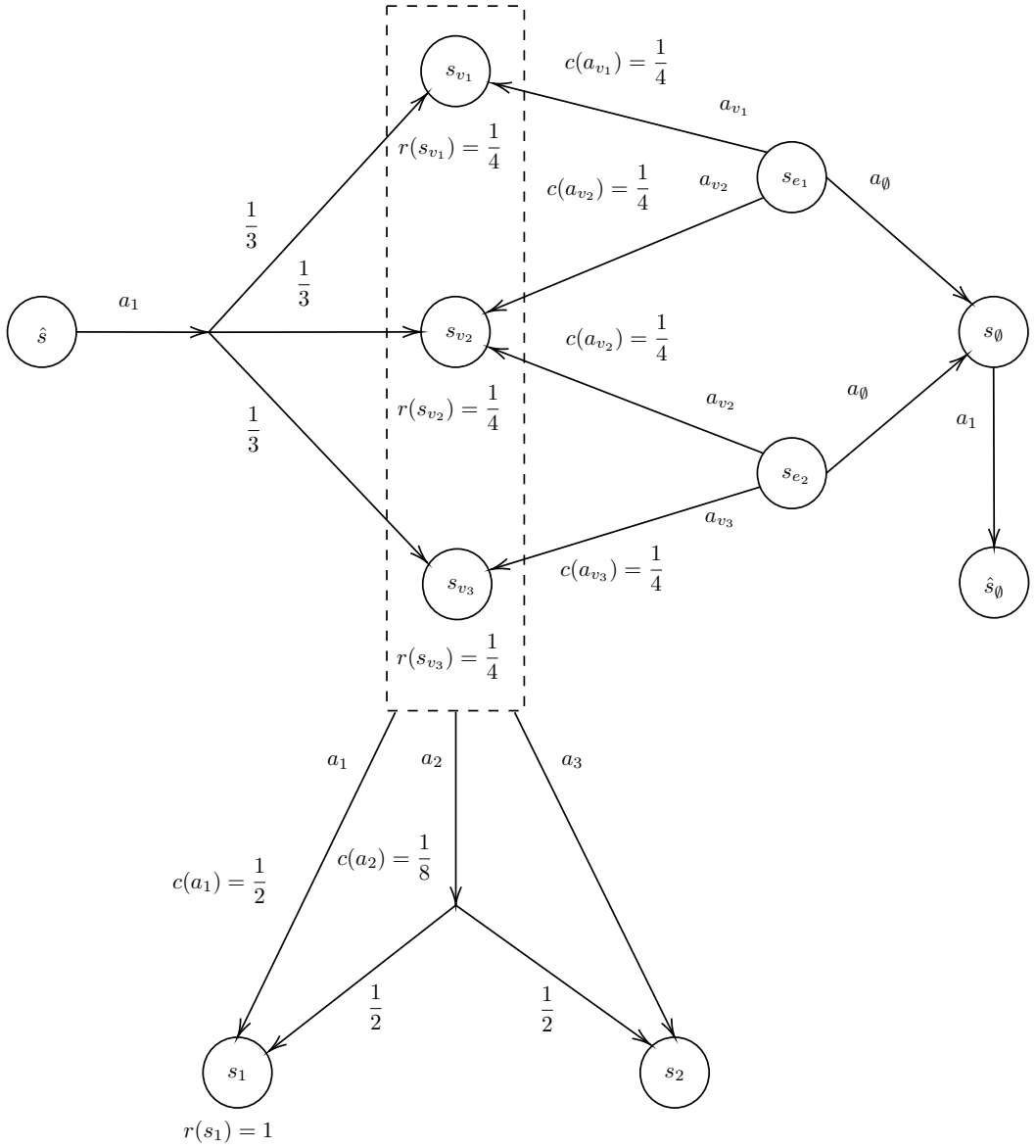

Figure 3: Instance of the Contract MDP problem constructed from the Cubic Graph depict in Figure 2. Costs and rewards are reported only when different from zero. The states s_{v_1}, s_{v_2} and s_{v_3} share the same set of actions $\{a_1, a_2, a_3\}$.

i.e., $r(s_v) = \frac{1}{4}$ for all v. Moreover, the reward reaching a state s_1 is $r(s_1) = 1$. All the other rewards are set to 0.

We conclude the construction specifing the initial state. The initial state is $\hat{s}$ with probability $\frac{1}{10}$, and for each $e \in E$ the initial state is s_e with probability $\frac{9}{10|E|}$.

If. Suppose that there exists a vertex cover V^* of size k. Consider a Markovian policy that sets payment on all nodes s_v, $v \in V^*$ to $p^{s_v}(s_1) = \frac{3}{4}$. On all nodes s_v, $v \notin V^*$, the payment is set to $p^{s_v}(s_1) = \frac{1}{4}$. All the other payments are set to 0.

It is easy to see that when the initial node is s_e, $e = (v_1, v_2) \in E$, the agent will take action a_v for a $v \in \{v_1, v_2\} \cap V^*$, and then a_1 in the reached state s_v with $v \in \{v_1, v_2\}$. Indeed, the utility of

the agent playing this sequence of actions is $3/4 - 1/4 - 1/2 = 0$. Playing sequence a_1 and a_2, the utility of the agent is $3/8 - 1/4 - 1/8 = 0$. Finally, playing any other sequence of actions the payment is 0, and hence the agent's utility is at most 0. Then, simple calculations show that the cumulative principal's utility is $5/4 - 3/4 = 1/2$ when the initial state is s_e, $e \in E$.

If the initial state is a $\hat{s}$, the agent plays the only action, i.e., a_1, and the transaction leads to a node s_v, $v \in V^*$, the principal's utility is $\frac{1}{4}$. Indeed, the agent is incentivized to play action s_1, with reward 1 and payment $3/4$.

If the initial state is a $\hat{s}$ and the transaction leads to a node s_v, $v \in V \setminus V^*$ the principal's utility is $\frac{1}{2} - \frac{1}{8} = \frac{3}{8}$. Indeed, the agent is incentivized to play action a_2, with expected reward 1 and expected payment $1/8$.

Hence, the expected utility is:

$$\frac{9}{10}\frac{1}{2} + \frac{1}{10|V|}\left[\frac{1}{4}k + \frac{3}{8}(|V| - k)\right] = \frac{9}{20} + \frac{3}{80} - \frac{1}{80}\frac{k}{|V|} = \frac{39}{80} - \frac{1}{80}\frac{k}{|V|}.$$

Only If. Suppose that all the vertex covers have size at least $(1+\epsilon)k$.

Consider a policy $\rho := \{p^s, a^s\}_{s \in \mathcal{S}}$ of the agent. We upperbound the utility of any policy such that the set of actions $\{a^s\}_{s \in \mathcal{S}}$ is IC. To do so, we show that it is sufficient to consider a subset of the IC contraints for the agent.

In particular, we consider the following "sub-components" of the graph. For each node s_v with $a^{s_v} = a_1$, we let $E(v)$ be the set of e such that $a^{s_e} = a_v$. Given a $v \in V$, we analyze the component $v \cup E(v)$ with $E(v) \neq \emptyset$. Notice that the agent plays action a_1 in node s_v by definition of $E(v)$ $E(v) \neq \emptyset$. Then, it must be the case that $p^{s_v}(s_1) \geq \frac{3}{4}$. This implies that if the initial node is an $e \in E(v)$, the principal's utility is at most $\frac{1}{2}$ since the reward is at most $\frac{5}{4}$ and the payment at least $\frac{3}{4}$, while if the initial node is $\hat{s}$ and the transaction leads to state v, the utility is at most $1/4$ by a similar argument.

Then, we upperbound the principal's utility if the initial node is s_e, $e \notin \cup_{v \in V} E(v)$. If the played action is $a_\varnothing$, the principal's utility is clearly 0. If the played action is a_v, $v \in V$, and the following action is a_2 or a_3 (recall that by definition of $E(v)$ the agent will not play a_1), the utility is upperbounded by the social welfare of the sequence of actions that is $3/8$.

Finally, consider the case in which the initial node is $\hat{s}$, the agent plays action a_1, and the transaction leads to state v, $E(v) = \emptyset$. Also in case, indipendently from the played action the principal's utility is upperbounded by the social welfare $\frac{3}{8}$.

Now, we count the number of nodes belonging to each component. In particular, we let $n = \cup_{v \in V} E(v)$, and $m = |\{v : E(v) \neq \emptyset\}|$. Then, by the previous analysis of all the cases we can conclude that the expected principal's utility is at most:

$$\frac{9}{10}\frac{1}{2}\frac{n}{|E|} + \frac{9}{10}\frac{3}{8}\left(1 - \frac{n}{|E|}\right) + \frac{1}{10}\frac{1}{4}\frac{m}{|V|} + \frac{1}{10}\frac{3}{8}\left(1 - \frac{m}{|V|}\right) = \frac{30}{80} + \frac{9}{80}\frac{n}{|E|} - \frac{1}{80}\frac{m}{|V|}.$$

Now, we we proceed providing a lowerbound on m. Intuitively, we know that if $n = |E|$, then $m \geq (1+\epsilon)k$ since it is an vertex cover. We can generalize this argument noticing that adding $|E| - n$ nodes to $\{v : E(v) \neq \emptyset\}$ it is possible to build a vertex cover. To do that, it is sufficient to take all the edge $e(v_1, v_2)$ not in $\cup_v E(v)$ and add v_1 to the vertex cover. Hence, by the assumption on the size of vertex cover, we get

$$m + |E| - n \geq (1+\epsilon)k.$$

This implies:

$$m/|V| \geq \frac{(1+\epsilon)k + n - |E|}{|V|} = \frac{(1+\epsilon)k}{|V|} - \frac{|E| - n}{|V|} \geq \frac{(1+\epsilon)k}{|V|} - \frac{2}{3}\frac{|E| - n}{|E|},$$

where the last inequality follows since each nodes has degree at most 3. Hence,

$$\frac{30}{80} + \frac{9}{80}\frac{n}{|E|} - \frac{1}{80}\frac{m}{|V|} \leq \frac{30}{80} + \frac{9}{80}\frac{n}{|E|} - \frac{1}{80}\left[\frac{(1+\epsilon)k}{|V|} - \frac{2}{3}\frac{|E| - n}{|E|}\right]$$

$$
\begin{aligned}
&= \frac{92}{240} + \frac{25}{240}\frac{n}{|E|} - \frac{1}{80}\frac{(1+\epsilon)k}{|V|} \\
&\leq \frac{92}{240} + \frac{25}{240} - \frac{1}{80}\frac{(1+\epsilon)k}{|V|} \\
&\leq \frac{39}{80} - \frac{1}{80}\frac{(1+\epsilon)k}{|V|}.
\end{aligned}
$$

The proof is concluded observing that $\frac{k}{|V|}$ is constant.[8]. Indeed, it implies that

$$
\frac{\frac{39}{80} - \frac{1}{80}\frac{(1+\epsilon)k}{|V|}}{\frac{39}{80} - \frac{1}{80}\frac{k}{|V|}}
$$

is a constant strictly smaller than 1. □

B Omitted details and proofs from Section 4

B.1 Value functions of promise-form policies

As we discussed in Section 4, given a promise-form policy we can define equivalent (According to Lemma 1) value and action-value functions that exploit the particular structure of this class of policies. Formally, given a promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$, we define the principal/agent value function ${V_h}^{\mathrm{P/A},\sigma}$ as the function that, for each step $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota \in I_h(s)$, recursively satisfies:

$$
{V_h}^{\mathrm{P/A},\sigma}(s,\iota) \coloneqq \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_h(p,a|s,\iota){Q_h}^{\mathrm{P/A},\sigma}(s,\iota,p,a),
$$

where the principal and the agent action-value functions are recursively defined as follows:

$$
{Q_h}^{\mathrm{P},\sigma}(s,\iota,p,a) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\left(r_h^{\mathrm{P}}(s,p,s') + Q_{h+1}^{\mathrm{P},\sigma}(s', g_h(s,\iota,p,a,s'))\right),
$$

and:

$$
Q_h^{\mathrm{A},\sigma}(s,\iota,p,a) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\left(r_h^{\mathrm{A}}(s,p,a,s') + V_{h+1}^{\mathrm{A},\sigma}(s', g_h(s,\iota,p,a,s'))\right),
$$

for each step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, contract $p \in \mathcal{J}$, action $a \in \mathcal{A}$ and promise $\iota \in I_h(s)$. In the equations above, the support $\mathcal{X}_\sigma$ of the policy σ is defined as $\mathcal{X}_\sigma \coloneqq \mathcal{X}_{\pi^\sigma}$.

We also define the value and action-value functions of the agent in the case in which the agent does not follow the principal's recommendation as follows:

$$
\widehat{V}_h^{\mathrm{A},\sigma}(s,\iota) \coloneqq \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi(p,a|s,\iota)\widehat{Q}_h^{\mathrm{A},\sigma}(s,\iota,p,a),
$$

$$
\widehat{Q}_h^{\mathrm{A},\sigma}(s,\iota,p,a) \coloneqq \max_{\widehat{a}\in A} \sum_{s'\in\mathcal{S}} P(s'|s,\widehat{a})\left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\sigma}(s', g_h(s,\iota,p,a,s'))\right).
$$

Finally, we observe that the value of a promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ can be written as:

$$
V^{\mathrm{P},\pi^\sigma} = \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\pi}(s) = V^{\mathrm{P},\sigma} = \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\sigma}(s, i(s))
$$

thanks to Lemma 1.

[8] Each non-trivial node is connected to at least another node. Hence $|E| \geq |V|/2$. Hence $k \geq |E|/3 \geq |V|/6$.

B.2 OMITTED PROOFS

Lemma 1. *Given a promise-form policy $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ and its corresponding history-dependent policy $\pi := \pi^\sigma$, for every step $h \in \mathcal{H}$, history $\tau \in \mathcal{T}_h$ ending in state $s := s_h$, action $a \in \mathcal{A}$, and contract $p \in \mathcal{P}$, the following hold:*

$$\begin{aligned}
V_h^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma) &= V_h^{\mathrm{P},\pi}(\tau), \quad Q_h^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a) = Q_h^{\mathrm{P},\pi}(\tau,p,a),\\
V_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma) &= V_h^{\mathrm{A},\pi}(\tau), \quad Q_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = Q_h^{\mathrm{A},\pi}(\tau,p,a),\\
\widehat{V}_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma) &= \widehat{V}_h^{\mathrm{A},\pi}(\tau), \quad \widehat{Q}_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = \widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a).
\end{aligned}$$

Proof. As a first step, we observe that $\mathcal{X}_\pi = \mathcal{X}_\sigma$ by construction. We prove the statement by induction over the steps $\mathcal{H}$ starting from $h = H$. For $h = H$, by the definitions of $Q_H^{\mathrm{P},\pi}(\tau,p,a)$ and $Q_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a)$, it holds that:

$$Q_H^{\mathrm{P},\pi}(\tau,p,a) = \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^P(s,p,s') = Q_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a)$$

for every history $\tau \in \mathcal{T}_H$ ending in state $s \in \mathcal{S}$, and pair $(p,a) \in \mathcal{X}$. Furthermore, since by construction $\varphi_H(p,a|s,\iota_\tau^\sigma) = \pi(p,a|\tau)$, it holds that:

$$V_H^{\mathrm{P},\pi}(\tau) = \sum_{(p,a)\in\mathcal{X}_\sigma} \pi(p,a|\tau) Q_H^{\mathrm{P},\pi}(\tau,p,a) = \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_H(p,a|s,\iota_\tau^\sigma) Q_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a) = V_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma).$$

With a similar argument, it is possible to show that $V_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma) = V_H^{\mathrm{A},\pi}(\tau)$ and $Q_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = Q_H^{\mathrm{A},\pi}(\tau,p,a)$ for every state $s \in \mathcal{S}$, history $\tau \in \mathcal{T}_H$ ending in state s, and pair $(p,a) \in \mathcal{X}$. Furthermore, we observe that:

$$\widehat{Q}_H^{\mathrm{A},\pi}(\tau,p,a) = \max_{\widehat{a}\in A} \sum_{s'\in\mathcal{S}} P_H(s'|s,\widehat{a}) r^{\mathrm{A}}(s,p,a,s') = \widehat{Q}_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a).$$

Consequently:

$$\widehat{V}_H^{\mathrm{A},\pi}(\tau) = \sum_{(p,a)\in\mathcal{X}_\sigma} \pi(p,a|\tau) \widehat{Q}_H^{\mathrm{A},\pi}(\tau,p,a) = \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_H(p,a|s,\iota_\tau^\sigma) \widehat{Q}_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = \widehat{V}_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma).$$

Now we consider a generic step $0 < h < H$ and assume that the statement holds for $h+1$. For every history $\tau \in \mathcal{T}_h$, contract $p \in \mathcal{P}$ and action $a \in \mathcal{A}$, the following holds:

$$Q_h^{\mathrm{P},\pi}(\tau,p,a) = \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\pi}(\tau\oplus(p,a,s')) \right).$$

We observe that $V_{h+1}^{\mathrm{P},\pi}(\tau\oplus(p,a,s')) = V_{h+1}^{\mathrm{P},\sigma}(s',\iota_{\tau\oplus(p,a,s')}^\sigma)$ by inductive hypothesis. Furthermore, $\iota_{\tau\oplus(p,a,s')}^\sigma = g_h(s,\iota_\tau^\sigma,p,a,s')$ by construction. Consequently, the following holds:[9]

$$\begin{aligned}
Q_h^{\mathrm{P},\pi}(\tau,p,a) &= \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\pi}(\tau\oplus(p,a,s')) \right)\\
&= \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma}(s',g_h(s,\iota_\tau^\sigma,p,a,s')) \right)\\
&= Q_h^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a)
\end{aligned}$$

As a result:

$$V_h^{\mathrm{P},\pi}(\tau) = \sum_{(p,a)\in\mathcal{X}_\sigma} \pi(p,a|\tau) Q_h^{\mathrm{P},\pi}(\tau,p,a) = \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_H(p,a|s,\iota_\tau^\sigma) Q_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma,p,a) = V_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma).$$

[9] We let $g_h(s,\iota,p,a,s') := \min(I_{h+1}(s'))$ for any contract $p \notin J_h(s,\iota,a)$. This way, the policy π is well defined for every history, even those that happen with probability zero (as they include contracts that are never prescribed).

A similar argument shows that $V_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma) = V_H^{\mathrm{A},\pi}(\tau)$ and $Q_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = Q_H^{\mathrm{A},\pi}(\tau,p,a)$.

Finally, we have that:

$$\begin{aligned}\widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a) &= \max_{\widehat{a}\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau\oplus(p,a,s'))\right)\\ &= \max_{\widehat{a}\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\sigma}(s', g_h(s,\iota_\tau^\sigma,p,a,s'))\right)\\ &= \widehat{Q}_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a)\end{aligned}$$

and

$$\widehat{V}_h^{\mathrm{A},\pi}(\tau) = \sum_{(p,a)\in\mathcal{X}_\sigma} \pi(p,a|\tau)\widehat{Q}_h^{\mathrm{A},\pi}(\tau,p,a) = \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_H(p,a|s,\iota_\tau^\sigma)\widehat{Q}_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) = \widehat{V}_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma),$$

concluding the proof. □

Lemma 2. *Let $\sigma := \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ be an η-honest promise-form policy. Then, for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$ it holds that $\left|V_h^{\mathrm{A},\sigma}(s,\iota) - \iota\right| \leq \eta(H-h+1)$ for every $\iota \in I_h(s)$.*

Proof. We prove the statement by induction over the time steps $h \in \mathcal{H}$, starting from $h = H$. When $h = H$, for every state $s \in \mathcal{S}$, promise $\iota \in I_H(s)$ the following holds:

$$\begin{aligned}\left|V_H^{\mathrm{A},\sigma}(s,\iota) - \iota\right| &= \left|\sum_{(p,a)\in\mathcal{X}} \varphi_H(p,a|s,\iota) Q_H^{\mathrm{A},\sigma}(s,\iota,p,a) - \iota\right|\\ &= \left|\sum_{(p,a)\in\mathcal{X}} \varphi_H(p,a|s,\iota) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{A}}(s,p,a,s') - \iota\right|\\ &= \left|\sum_{(p,a)\in\mathcal{X}} \varphi_H(p,a|s,\iota) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) \left(r_H^{\mathrm{A}}(s,p,a,s') + g_H(s,\iota,p,a,s')\right) - \iota\right|\\ &\leq \eta,\end{aligned}$$

where the last equality holds because $I_{H+1}(s) = \{0\}$ by definition, while the inequality holds because σ is η-honest at step H.

Now we consider a time step $h < H$ and assume that the statement holds for $h' = h+1$. Then, for every state $s \in \mathcal{S}$, promise $\iota \in I_h(s)$ and pair $(p,a) \in \mathcal{X}$ such that $\varphi_H(p,a|s,\iota) > 0$ the following holds:

$$\begin{aligned}Q_h^{\mathrm{A},\sigma}(s,\iota,p,a) &= \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + V_{h+1}^{\mathrm{A},\sigma}(s', g_h(s,\iota,p,a,s'))\right)\\ &\geq \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s,\iota,p,a,s') - \eta(H-h)\right)\\ &= \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s,\iota,p,a,s')\right) - \eta(H-h),\end{aligned}$$

where the first inequality holds by inductive hypothesis, while the last one holds because we have $\sum_{s'\in\mathcal{S}} P_h(s'|s,a) = 1$. Thanks to this result and since σ is η-honest, we have that:

$$\begin{aligned}V_h^{\mathrm{A},\sigma}(s,\iota) &= \sum_{(p,a)\in\mathcal{X}_\sigma} \varphi_h(p,a|s,\iota) Q_h^{\mathrm{A},\sigma}(s,\iota,p,a)\\ &\geq \iota - \eta - \eta(H-h)\\ &= \iota - \eta(H-h+1).\end{aligned}$$

With a similar argument, we can show that $V_h^{\mathrm{A},\sigma}(s,\iota) \leq \iota + \eta(H-h+1)$, concluding the proof. □

Lemma 3. *Let $\zeta > 0$ and $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ be an η-honest promise-form policy such that the following constraint holds for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, promise $\iota \in I_h(s)$, action $a, \widehat{a} \in \mathcal{A}$, and contract $p \in \mathcal{P}$ such that $\varphi_h(p, a|s, \iota) > 0$:*

$$\sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s,\iota,p,a,s')\right) \geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + g_h(s,\iota,p,a,s')\right) - \zeta. \tag{1}$$

Then, the promise-form policy σ is $(2\eta\zeta H^2)$-IC.

Proof. Let π be the history-dependent policy that implements σ. In order to prove that σ (equivalently π) is ϵ-IC, we show that for every step $h \in \mathcal{H}$, history $\tau \in \mathcal{T}_h$ contract $p \in \mathcal{P}$, and pair of actions $a, \widehat{a} \in \mathcal{A}$ such that $\pi(p, a|\tau) > 0$, the following holds:

$$Q_h^{\mathrm{A},\pi}(\tau, p, a) \geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau \oplus (p,a,s'))\right) - \epsilon_h, \tag{5}$$

where:

$$\epsilon_h = \begin{cases} \zeta, & h = H, \\ 2\eta(H-h+1) + \zeta + \epsilon_{h+1}, & h = 1, \dots, H-1. \end{cases}$$

Let us observe that $\epsilon_1 \leq 2\eta\zeta H^2$. Therefore, Equation 5 implies that the policy is $(2\eta\zeta H^2)$-IC.

In order to prove that Equation 5 holds for every time step $h \in \mathcal{H}$, we reason by induction, starting from $h = H$. Take any state $s \in \mathcal{S}$, history $\tau \in \mathcal{T}$ ending in state s, contract $p \in \mathcal{P}$ and pair of actions $a, \widehat{a} \in \mathcal{A}$ such that $\pi(p, a|\tau) > 0$. Then, thanks to Lemma 1, we have that $Q_h^{\mathrm{A},\pi}(\tau, p, a) = Q_h^{\mathrm{A},\sigma}(s, \iota_\tau^\sigma, p, a)$. Moreover, observing that $I_{H+1}(s) \coloneqq \{0\}$, the action-value of the agent can be lower bounded as follows:

$$\begin{aligned} Q_H^{\mathrm{A},\pi}(\tau,p,a) &= Q_H^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) \\ &= \sum_{s' \in \mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{A}}(s,p,a,s') \\ &= \sum_{s' \in \mathcal{S}} P_H(s'|s,a) \left(r_H^{\mathrm{A}}(s,p,a,s') + g_H(s,\iota,p,a,s')\right) \\ &\geq \sum_{s' \in \mathcal{S}} P_H(s'|s,\widehat{a}) \left(r_H^{\mathrm{A}}(s,p,\widehat{a},s') + g_H(s,\iota,p,a,s')\right) - \zeta \\ &= \sum_{s' \in \mathcal{S}} P_H(s'|s,\widehat{a}) r_H^{\mathrm{A}}(s,p,\widehat{a},s') - \zeta, \end{aligned}$$

where the inequality holds by hypothesis thanks to Equation 1. Since $\epsilon_H = 0$, we conclude that Equation 5 holds for $h = H$.

Now we let $h < H$ and we assume that Equation 5 holds for $h' = h + 1$. Thus, given the inductive hypothesis, the following holds:

$$Q_{h+1}^{\mathrm{A},\pi}(\tau,p,a) + \epsilon_{h+1} \geq \widehat{Q}_{h+1}^{\mathrm{A},\pi}(\tau,p,a),$$

for every history $\tau \in \mathcal{T}_{h+1}$ and every contract $p \in \mathcal{P}$ and action $a, \in \mathcal{A}$ such that $\pi(p, a|\tau) > 0$. Furthermore, observing that $\sum_{(p,a) \in \mathcal{X}} \pi(p, a|\tau) = 1$, we can prove that $V_{h+1}^{\mathrm{A},\pi}(\tau) + \epsilon_{h+1} \geq \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau)$, which implies that:

$$V_{h+1}^{\mathrm{A},\pi}(\tau) \geq \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau) - \epsilon_{h+1}. \tag{6}$$

Consider any other action $\widehat{a} \neq a$. Since $\sum_{s' \in \mathcal{S}} P_h(s'|s,a) = 1$, the action-value function of the agent can be lower bounded as follows:

$$\begin{aligned} &Q_h^{\mathrm{A},\pi}(\tau,p,a) \\ &= Q_h^{\mathrm{A},\sigma}(s,\iota_\tau^\sigma,p,a) \end{aligned} \tag{7}$$

$$= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + V_{h+1}^{\mathrm{A},\sigma}(s', g_h(s, \iota_\tau^\sigma, p, a, s')) \right)$$

$$\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s, \iota_\tau^\sigma, p, a, s') - \eta(H-h+1) \right) \tag{8}$$

$$= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s, \iota_\tau^\sigma, p, a, s') \right) - \eta(H-h+1)$$

$$\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + g_h(s, \iota_\tau^\sigma, p, a, s') \right) - \eta(H-h+1) - \zeta \tag{9}$$

$$\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + V_{h+1}^{\mathrm{A},\pi}(\tau \oplus (p,a,s')) \right) - 2\eta(H-h+1) - \zeta \tag{10}$$

$$\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau \oplus (p,a,s')) \right) - 2\eta(H-h+1) - \zeta - \epsilon_{h+1} \tag{11}$$

where Equation 8 holds thanks to Lemma 2 ($V_{h+1}^{\mathrm{A},\sigma}(s', g_h(s, \iota_\tau^\sigma, p, a, s')) \geq g_h(s, \iota_\tau^\sigma, p, a, s') - \eta(H-h+1)$), Equation 9 employs Equation 1, Equation 10 holds due to Lemma 2 and Lemma 1, while Equation 11 holds due to Equation 6. As a result, by considering that $Q_h^{\mathrm{A},\sigma}(s, \iota_\tau^\sigma, p, a) = Q_h^{\mathrm{A},\pi}(\tau, p, a)$, Equation 5 is satisfied. This concludes the proof. □

Theorem 2. *There always exists an honest IC promising-form policy that achieves a principal's expected cumulative reward equal to* OPT.

Proof. We show that, given an optimal IC policy $\pi^\star$, one can always build a promise-form $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ that achieves the same expected principal's cumulative utility. Let us define, for every step $h \in \mathcal{H}$, the set $\mathcal{T}_h^\pi$ as:

$$\mathcal{T}_h^\pi := \{\tau \in \mathcal{T}_h | \tau = (s_1, p_1, a_1, \ldots, s_{h-1}, p_{h-1}, a_{h-1}, s), \pi(p_i, a_i | \tau(s_i)) > 0, i \in \mathcal{H} \cap [1, H-1]\},$$

where $\tau(s_i) = (s_1, p_1, a_1, \ldots, s_{i-1}, p_{i-1}, a_{i-1}, s) > 0$. Given a history-dependent policy π with finite support $\mathcal{X}_\pi$, we define $\sigma = \sigma_\pi$ as follows:

- $I_h(s) := \{V_h^{\mathrm{A},\pi}(\tau) | \tau \in \mathcal{T}_h^\pi\}$ for all $s \in \mathcal{S}$ and $h \in \mathcal{H}$. Observe that $|I_1(s)| = 1$ for very $s \in \mathcal{S}$, since $\mathcal{T}_1^\pi = \{(s_i) \mid s_i \in \mathcal{S}\}$.
- $\varphi_h(p, a | s, \iota) = \pi(p, a | \tau_{h,s,\iota})$ for all $(p,a) \in \mathcal{X}$, $h \in \mathcal{H}$, $s \in \mathcal{S}$ and $\iota \in I_h(s)$, where:

$$\tau_{h,s,\iota} \in \underset{\substack{\tau \in \mathcal{T}_h^\pi: \\ \tau = (s_1, \ldots, s_h), \\ s_h = s \,\wedge\, V_h^{\mathrm{A},\pi}(\tau) = \iota}}{\arg\max} \{V_h^{\mathrm{P},\pi}(\tau)\}$$

- $J_h(s, \iota, a) = \{p \in \mathcal{P} \mid \pi(p, a | \tau_{h,s,\iota}) > 0\}$ for every $h \in \mathcal{H}$, $\iota \in I_h(s)$ and $a \in \mathcal{A}$.
- $g_h(s, \iota, p, a, s') = V_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p, a, s'))$ for all $h \in \mathcal{H}$, $s, s' \in \mathcal{S}$, $\iota \in I_h(s)$, $a \in \mathcal{A}$ and $p \in J_h(s, \iota, a)$.

We observe that given this definition, it holds that $\mathcal{X}_\sigma \subseteq \mathcal{X}_\pi$. Furthermore, for every $h \in \mathcal{H}$ the set $\mathcal{T}_h^\pi$ is finite. Consequently, both the sets $\mathcal{I}$ and $\mathcal{J}$ are finite.

As a first step, we show that, σ is η-honest with $\eta = 0$. For every $h \in \mathcal{H}$, $s \in \mathcal{S}$, $\iota \in I_h(s)$ and $(p,a) \in \mathcal{X}$ such that $\varphi_h(p, a | s, \iota) = \pi(p, a | \tau_{h,s,\iota}) > 0$ the following holds:

$$\sum_{(p,a) \in \mathcal{X}_\sigma} \varphi(p, a | s, \iota) \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s, \iota, p, a, s') \right)$$

$$= \sum_{\substack{(p,a) \in \mathcal{X}: \\ \pi(p, a | \tau_{h,s,\iota}) > 0}} \pi(p, a | \tau_{h,s,\iota}) \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + V_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p,a,s')) \right)$$

$$= V_h^{\mathrm{A},\pi}(\tau_{h,s,\iota}) = \iota,$$

where the first equality holds thanks to the definitions of g_h and φ_h, the second one due to the definition of $V_h^{\mathrm{A},\pi}(\tau_{h,s,\iota})$, and the last equality thanks to the definition of $\tau_{h,s,\iota}$. Thus, σ is η-honest, with $\eta = 0$.

Now we prove that, given that π is IC and σ is honest, σ is IC as well. In order to do that, we show that Equation 1 is satisfied and apply Lemma 3. First, we observe that thanks to the definitions of $Q_h^{\mathrm{P},\pi}(\tau_{h,s,\iota}, p, a)$ and $g_h(s, \iota, a, s')$:

$$\begin{aligned} Q_h^{\mathrm{A},\pi}(\tau_{h,s,\iota}, p, a) &= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + V_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p,a,s')) \right) \\ &= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s,p,a,s') + g_h(s,\iota,p,a,s') \right), \end{aligned} \tag{12}$$

for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, promise $\iota \in I_h(s)$, and action $a \in \mathcal{A}$ with corresponding contract $p \in \mathcal{P}$ such that $\varphi_h(p, a|s, \iota) > 0$. Furthermore, for every action $\widehat{a} \in \mathcal{A}$:

$$\begin{aligned} Q_h^{\mathrm{A},\pi}(\tau_{h,s,\iota}, p, a) &\geq \sum_{s' \in \mathcal{S}} P(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + \widehat{V}_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p,a,s')) \right) \\ &= \sum_{s' \in \mathcal{S}} P(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + V_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p,a,s')) \right) \\ &= \sum_{s' \in \mathcal{S}} P(s'|s,\widehat{a}) \left(r_h^{\mathrm{A}}(s,p,\widehat{a},s') + g_h(s,\iota,p,a,s') \right) \end{aligned} \tag{13}$$

since π is IC and $V_{h+1}^{\mathrm{A},\pi}(\tau_{h,s,\iota} \oplus (p,a,s')) = g_h(s,\iota,p,a,s')$. As a result, by combining Equation 12 and Equation 13, we have that Equation 1 is satisfied. Consequently, by applying Lemma 3, the policy σ is incentive compatible.

In order to conclude the proof, we show that the promise-form policy σ achieves a principal's expected cumulative utility at least as good as that obtained by π. Let π^σ be the history-dependent policy that implements σ (notice that it may be different from π). Formally, we prove by induction over $h \in \mathcal{H}$ that $V_h^{\mathrm{P},\pi^\sigma}(\tau) \geq V_h^{\mathrm{P},\pi}(\tau)$ for every step $h \in \mathcal{H}$ and history $\tau \in \mathcal{T}_h^\pi$.

As a base case, we consider $h = H$. For every state $s \in \mathcal{S}$ and history $\tau \in \mathcal{T}_H^\pi$ ending in state s it holds that:

$$\begin{aligned} V_H^{\mathrm{P},\pi}(\tau) &\leq V_H^{\mathrm{P},\pi}(\tau_{H,s,\iota_\tau^\sigma}) \\ &= \sum_{(p,a) \in \mathcal{X}_\pi} \pi(p,a|\tau_{H,s,\iota_\tau^\sigma}) \sum_{s' \in \mathcal{S}} P_H(s'|s,a) r_H^P(s,p,s') \\ &= \sum_{(p,a) \in \mathcal{X}_\sigma} \varphi_H(p,a|s,\iota_\tau^\sigma) \sum_{s' \in \mathcal{S}} P_H(s'|s,a) r_H^P(s,p,s') \\ &= V_H^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma) \\ &= V_h^{\mathrm{P},\pi^\sigma}(\tau), \end{aligned}$$

where the inequality holds due to the definition of $\tau_{H,s,\iota_\tau^\sigma}$ and the last equality one thanks to Lemma 1.

Now we consider $h < H$ and assume that the statement holds for $h' = h + 1$. Then for every history $\tau \in \mathcal{T}_h^\pi$ ending in some state $s \in \mathcal{S}$ the following holds:

$$\begin{aligned} V_h^{\mathrm{P},\pi}(\tau) &\leq V_h^{\mathrm{P},\pi}(\tau_{h,s,\iota_\tau^\sigma}) \\ &= \sum_{(p,a) \in \mathcal{X}_\pi} \pi(p,a|\tau_{h,s,\iota_\tau^\sigma}) Q_h^{\mathrm{P},\pi}(\tau_{h,s,\iota_\tau^\sigma}, p, a) \\ &= \sum_{(p,a) \in \mathcal{X}_\pi} \varphi_h(p,a|s,\iota_\tau^\sigma) Q_h^{\mathrm{P},\pi}(\tau_{h,s,\iota_\tau^\sigma}, p, a), \end{aligned} \tag{14}$$

where the inequality holds due to the definition of $\tau_{H,s,\iota_\tau^\sigma}$. Furthermore, for every contract $p \in \mathcal{P}$ and action $a \in \mathcal{A}$ such that $\pi(p, a|\tau) > 0$, it holds that:

$$Q_h^{\mathrm{P},\pi}(\tau_{h,s,\iota_\tau^\sigma}, p, a) = \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\pi}(\tau_{h,s,\iota_\tau^\sigma} \oplus (p,a,s')) \right)$$

$$= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + g_h(s, \iota_\tau^\sigma, p, a, s') \right) \tag{15}$$

$$\leq \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\pi^\sigma}(\tau_{h+1,s',g_h(s,\iota_\tau^\sigma,p,a,s')}) \right) \tag{16}$$

$$= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma}(s', g_h(s, \iota_\tau^\sigma, p, a, s')) \right) \tag{17}$$

$$= Q_h^{\mathrm{P},\sigma}(s, \iota_\tau^\sigma)$$

where Equation 15 holds due to the definition of $g_h(s, \iota_\tau^\sigma, p, a, s')$, Equation 16 holds by induction, while Equation 17 follows from Lemma 1. Thus, by combining this result with Equation 14 we get:

$$V_h^{\mathrm{P},\pi}(\tau) \leq \sum_{(p,a) \in \mathcal{X}_\pi} \varphi_h(p,a|s,\iota_\tau^\sigma) Q_h^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma) = V_h^{\mathrm{P},\sigma}(s,\iota_\tau^\sigma) = V_h^{\mathrm{P},\pi^\sigma}(\tau),$$

where the last equality holds thanks to Lemma 1 and the definition of ι_τ^σ, concluding the proof.

Since $V_h^{\mathrm{P},\pi^\sigma}(\tau) \geq V_h^{\mathrm{P},\pi}(\tau)$ for every step $h \in \mathcal{H}$ and history $\tau \in \mathcal{T}_h^\pi$, it follows that:

$$V^{\mathrm{P},\pi} = \sum_{s \in \mathcal{S}} \mu(s) V_1^{\mathrm{P},\pi}(s) = \sum_{s \in \mathcal{S}} \mu(s) V_1^{\mathrm{P},\pi^\sigma}(s) = V^{\mathrm{P},\pi^\sigma},$$

as $\mathcal{T}_1^\pi = \mathcal{T}_1 = \{(s) \mid s \in \mathcal{S}\}$.

Finally, we observe that the promise-form policy $\sigma^\star = \sigma_{\pi^\star}^\star$, where $\pi^\star$ is an optimal history-dependent policy, is optimal, as $V^{\mathrm{P},\pi^{\sigma^\star}} \geq V^{\mathrm{P},\pi^\star}$. □

C Proofs omitted from Section 5

Lemma 5. *Let $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ be the promise-form policy computed by Algorithm 2. For every $h \in \mathcal{H}$, $s \in \mathcal{S}$, and $\iota \in I_h(s)$, it holds that $V_h^{\mathrm{P},\sigma}(s,\iota) \geq M_h^\delta(s,\iota)$.*

Proof. Let $\sigma = \{\{(I_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}\}$ be the promise-form policy returned by Algorithm 2. We prove the statement by induction. For $h = H$, for any state $s \in \mathcal{S}$ and promise $\iota \in I_H(s)$, we let $(\alpha, p, z, v) = O_{\iota,\delta}^{h,s}(M_{H+1})$. Then, the following holds:

$$\begin{aligned} V_H^{\mathrm{P},\sigma}(s,\iota) &= \sum_{(p,a) \in \mathcal{X}} \varphi_H(p,a|s,\iota) \sum_{s' \in \mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s') \\ &= \sum_{a \in \mathcal{A}} \alpha_a \sum_{s' \in \mathcal{S}} P_H(s'|,s,a) r_H^{\mathrm{P}}(s,p^a,s') \\ &= F_{H,s,M_{H+1}^\delta}(\alpha,p,z) \\ &\geq v, \end{aligned}$$

where the second equality holds by construction (see Algorithm 2), while the inequality follows from Definition 4. Since $v = M_H^\delta(s,\iota)$, the statement holds for the base case $h = H$.

Now we consider a step $h < H$ and we assume that the statement holds for $h' = h + 1$. Furthermore, we let $(\alpha, p, z, v) = O_{\iota,\delta}^{h,s}(M_{h+1})$. Then, the following holds:

$$\begin{aligned} V_h^{\mathrm{P},\sigma}(s,\iota) &= \sum_{(p,a) \in \mathcal{X}_\sigma} \varphi_h(p,a|s,\iota) Q_h^{\mathrm{P},\sigma}(s,\iota,p,a) \\ &= \sum_{a \in \mathcal{A}} \alpha_a Q_h^{\mathrm{P},\sigma}(s,\iota,p^a,a) \\ &= \sum_{a \in \mathcal{A}} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p^a,s') + V_{h+1}^{\mathrm{P},\sigma}(s', g_h(s,\iota,p^a,a,s')) \right) \\ &\geq \sum_{a \in \mathcal{A}} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p^a,s') + M_{h+1}^\delta(s', g_h(s,\iota,p^a,a,s')) \right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{a\in\mathcal{A}} \alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\left(r_h^{\mathrm{P}}(s,p^a,s') + M_{h+1}^{\delta}(s',z(a,s'))\right)\\
&= F_{h,s,M_{h+1}^{\delta}}(\alpha,p,z)\\
&\geq v,
\end{aligned}
$$

where the first inequality holds employing the inductive hypothesis and the last inequality follows from Definition 4. The proof is concluded observing that $v = M_h^{\delta}(s,\iota)$ by construction. □

Lemma 4. *Let $\sigma^\star = \{(I_h^\star, J_h^\star, \varphi_h^\star, g_h^\star)\}_{h\in\mathcal{H}}$ be an honest, direct, IC and optimal promise-form policy, and M_h^δ, I_h be computed accordingly to Algorithm 2. For every $h \in \mathcal{H}$, $s \in \mathcal{S}$, and $\iota \in I_h^\star(s)$, it holds that $M_h^\delta(s, \lceil\iota\rceil_\delta) \geq V_h^{P,\sigma^\star}(s,\iota)$ and $M_h^\delta(s,\lfloor\iota\rfloor_\delta) \geq V_h^{P,\sigma^\star}(s,\iota)$.*

Proof. We recall that, w.l.o.g., we can consider an optimal direct, honest and IC promise-form policy $\sigma^\star$. Thus, for every action $a \in \mathcal{A}$ there is at most a single contract $p_h^{s,\iota,a}$ such that $\varphi_h^\star(p_h^{s,\iota,a}, a|s,\iota) > 0$. For the sake of notation, if $\varphi_h^\star(p,a|s,\iota) = 0$ for every $p \in \mathcal{P}$, we let $p_h^{s,\iota,a}$ be any contract in $\mathcal{P}$. We further recall (see the proof of Theorem 2) that $\sigma^\star$ satisfies Equation 1 with tolerance $\zeta = 0$.

We prove the statement by induction, starting from $h = H$. We consider a state $s \in \mathcal{S}$ and a promise $\widetilde{\iota} \in I_H^\star(s)$. Let (α,p,z) be the solution defined as follows:

$$
\begin{aligned}
\alpha_a &= \varphi_H^\star(p_H^{s,\iota,a}, a|s,\widetilde{\iota}) && \forall a\in\mathcal{A},\\
p_a &= p_H^{s,\iota,a} && \forall a\in\mathcal{A},\\
z(a,s') &= 0 && \forall a\in\mathcal{A}, s'\in\mathcal{S}.
\end{aligned}
$$

We can easily verify that $(\alpha,p,z) \in \Psi_{h,s}^{\widetilde{\iota}}(\lambda = 0, \zeta = 0)$ (the argument is essentially the same used for $h < H$ in the following, after noticing that future promises and utilities are zero by definition). Let now $\iota \in \{\lfloor\widetilde{\iota}\rfloor_\delta, \lceil\widetilde{\iota}\rceil_\delta\}$. As ι is at most at distance δ from $\widetilde{\delta}$, therefore $(\alpha,p,z) \in \Psi_{h,s}^{\widetilde{\iota}}(\lambda = \delta, \zeta = 0)$. Now let v be the solution return by the oracle. By Property 2 Definition 4 we have:

$$
\begin{aligned}
M_h^\delta(s,\iota) &= v\\
&\geq F_{h,s,M_{h+1}}(\alpha,p,z)\\
&= \sum_{a\in\mathcal{A}} \alpha_a \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^P(s,p^a,s')\\
&= \sum_{(p',a)\in\mathcal{X}_\sigma^\star} \varphi_H^\star(p',a|s,\widetilde{\iota}) \sum_{s'\in\mathcal{S}} P_H(s',s,a) r_H^P(s,p',s')\\
&= \mathrm{V}_H^{P,\sigma^\star}(s,\widetilde{\iota}),
\end{aligned}
$$

where we exploit the fact that all entries of M_{H+1}^δ are zero by construction.

Now consider any $h = 1,\dots,H-1$. We will prove that $M_h(s,\lfloor\widetilde{\iota}\rfloor_\delta) \geq V_h^{\mathrm{P},\sigma^\star}(s,\widetilde{\iota})$, but a similar argument (by setting $z(a,s') = \lceil g_h^\star(s,\iota,a,s')\rceil_\delta$ in the following definition) proves $M_h(s,\lceil\widetilde{\iota}\rceil_\delta) \geq V_h^{\mathrm{P},\sigma^\star}(s,\widetilde{\iota})$. Let $(\bar\alpha,\bar p,\bar z)$ be defined as follows:

$$
\begin{aligned}
\bar\alpha_a &= \varphi_h^\star(p_h^{s,\iota,a}, a|s,\widetilde{\iota}) && \forall a\in\mathcal{A}\\
\bar p_a &= p_h^{s,\iota,a} && \forall a\in\mathcal{A}\\
\bar z(a,s') &= \lfloor g_h^\star(s,\iota,a,\bar p_a,s')\rfloor_\delta && \forall a\in\mathcal{A}, s'\in\mathcal{S}.
\end{aligned}
$$

We argue that $(\bar\alpha,\bar p,\bar z) \in \Psi_{h,s}^{\iota}(\lambda = \delta, \zeta = 2\delta)$. In order to do so, we first define:

$$
d \coloneqq \sum_{a\in A} \bar\alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\left(z(a,s') - g_h^\star(s,\iota,a,\bar p_a,s')\right),
$$

and we observe that $|d| \leq \delta$. First, consider Constraint 2:

$$
\left|\sum_{a\in A} \bar\alpha_a \sum_{s'\in\mathcal{S}} P_h(s'|s,a)\left(r_h^A(s,\bar p_a,a,s') + \bar z(a,s')\right) - \iota\right|
$$

$$
\begin{aligned}
&= \left| \sum_{a \in A} \bar{\alpha}_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^A(s, \bar{p}_a, a, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) + d - \iota \right| \\
&\leq \left| \sum_{a \in A} \bar{\alpha}_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^A(s, \bar{p}_a, a, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) - \iota \right| + |d| \\
&= \left| \sum_{(p,a) \in \mathcal{X}_\sigma} \varphi_h(p, a|s, \iota) \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{A}}(s, p, a, s') + g_h(s, \widetilde{\iota}, p, a, s') \right) - \iota \right| + |d| \\
&= |d| \leq \delta,
\end{aligned}
$$

where the first two equalities hold by the definitions of d and z, the inequality by applying the triangle inequality, and the last equality since $\sigma^\star$ is honest. Therefore, $(\bar{\alpha}, \bar{p}, \bar{z})$ satisfies Constraint 2 with $\lambda = \delta$. Now consider Constraint 3 for any two actions $a, \widehat{a} \in \mathcal{A}$ with $\alpha_a, \alpha_{\widehat{a}} > 0$. Similarly to what we did for Constraint 2, let us define for the ease of notation:

$$
\begin{aligned}
k(a) &\coloneqq \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(z(a, s') - g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right), \\
k(\widehat{a}) &\coloneqq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(z(a, s') - g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right).
\end{aligned}
$$

We then observe:

$$
\begin{aligned}
&\sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^A(s, \bar{p}_a, a, s') + \bar{z}(a, s') \right) \\
&= \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^A(s, \bar{p}_a, a, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) + k(a) \\
&\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^A(s, \bar{p}_a, \widehat{a}, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) + k(a) \\
&= \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^A(s, \bar{p}_a, \widehat{a}, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) + k(a) - k(\widehat{a}) \\
&\geq \sum_{s' \in \mathcal{S}} P_h(s'|s,\widehat{a}) \left(r_h^A(s, \bar{p}_a, \widehat{a}, s') + g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \right) - 2\delta,
\end{aligned}
$$

where the equalities hold by the definition of the quantities involved. The first inequality above holds since $\sigma^\star$ is IC and honest, thus it satisfies Equation 1 with $\zeta = 0$, and the last inequality holds since $|k(a) - k(\widehat{a})| \leq 2\delta$. As result Constraint 3 is satisfied with $\zeta = 2\delta$, and we conclude that $(\bar{\alpha}, \bar{p}, \bar{z}) \in \Psi^\iota_{h,s}(\lambda = \delta, \zeta = 2\delta)$.

We conclude the proof by observing the following:

$$
\begin{aligned}
M_h(s, \iota) &= v \\
&\geq F_{h,s,M_{h+1}}(\bar{\alpha}, \bar{p}, \bar{z}) \\
&= \sum_{a \in \mathcal{A}} \bar{\alpha}_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^P(s, \bar{p}_a, s') + M_{h+1}(s, \bar{z}(a, s')) \right) \\
&= \sum_{a \in \mathcal{A}} \bar{\alpha}_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^P(s, \bar{p}_a, s') + M_{h+1}(s, \lfloor g_h^\star(s, \widetilde{\iota}, a, \bar{p}_a, s') \rfloor_\delta) \right) \\
&\geq \sum_{a \in \mathcal{A}} \bar{\alpha}_a \sum_{s' \in \mathcal{S}} P_h(s'|s,a) \left(r_h^P(s, \bar{p}_a, s') + V_{h+1}^{\sigma^\star}(s', \widetilde{\iota}) \right) = V_{h+1}^{\sigma^\star}(s, \widetilde{\iota}),
\end{aligned}
$$

where the first inequality holds by Property 1 Definition 4, and the second by the inductive hypothesis. □

Theorem 3. *For any $\epsilon > 0$, Algorithm 2 instantiated with $\delta = \epsilon/16H^2$ computes an ϵ-IC promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h \in \mathcal{H}}$ in time polynomial in the instance size and $1/\epsilon$. Furthermore, σ is such that $V^{P,\pi^\sigma} \geq$ OPT, while $\mathcal{I} = \mathcal{O}(H^4 B/\epsilon)$ and $\mathcal{J} = \mathcal{O}(H^4 B|\mathcal{S}||\mathcal{A}|/\epsilon)$.*

Proof. First, we observe that σ returned by Algorithm 2 is a promise-form policy. Indeed, we have that $|I_1(s)| = 1$ for every state $s \in \mathcal{S}$ by construction (see Line 15 in Algorithm 2). By satisfying Definition 4, the approximating oracle guarantees the remaining properties of promise-form policies.

As a further step, we show that σ is η-honest, with $\eta = \lambda = 2\delta$. We observe that the policy σ is constructed from the solutions provided by an approximation oracle that satisfies Definition 4. Furthermore, for every state $s \in \mathcal{S}$ and step $h \in \mathcal{H}$, a promise $\iota \in \mathcal{D}_\delta$ is added to the set $I_h(s)$ only if the value v is larger than $-\infty$. Thus, according to Definition 4, the solutions (α, p, q) used to construct σ belong to the corresponding set $\Psi^{h,s}_{\iota,h}(\lambda = 2\delta, \zeta = 4\delta)$ and satisfy Equation 2 for $\lambda = 2\delta$. Consequently, one can easily verify that σ is 2δ-honest.

Since σ is 2δ-honest, we can employ Lemma 3 to prove that it is also approximately IC. In order to do that, we observe that σ satisfies Equation 1 with $\zeta = 4\delta$, as the solutions returned by the approximation oracle satisfy Equation 3 with such a tolerance ζ. Therefore, the policy is $(16\delta H^2)$-IC. Thus, by observing that $\delta = \epsilon/16H^2$, it follows that σ is ϵ-IC.

To conclude the proof, we prove that the principal's expected cumulative utility achieved by of sigma is at least OPT. To show that, we consider an optimal and honest promise-form policy $\sigma^\star = \{I^\star_h, \varphi^\star_h, g^\star_h\}_{h \in \mathcal{H}}$, which always exists according to Theorem 2. By combining Lemma 5 and Lemma 4, we have that $V^{\mathrm{P},\sigma}_h(s, \lfloor \iota \rfloor_\delta) \geq \mathrm{V}^{\mathrm{P},\sigma^\star}_h(s, \iota)$ for every state $s \in \mathcal{S}$, step $h \in \mathcal{H} \setminus \{1\}$, and promise $\iota \in I^\star_h(s)$. Furthermore, at the time step $h = 1$, thanks to Lemma 5, Lemma 4 and Line 15 in Algorithm 2, it holds that $V^{\mathrm{P},\sigma}_1(s, i(s)) \geq V^{\mathrm{P},\sigma^\star}_1(s, i^\star(s))$. Then, we have:

$$
\begin{aligned}
\mathrm{OPT} &= V^{\mathrm{P},\pi^{\sigma^\star}} \\
&= \sum_{s \in \mathcal{S}} \mu(s) V^{\mathrm{P},\pi^{\sigma^\star}}_1(s) \\
&= \sum_{s \in \mathcal{S}} \mu(s) V^{\mathrm{P},\sigma^\star}_1(s, i^\star(s)) \\
&\leq \sum_{s \in \mathcal{S}} \mu(s) V^{\mathrm{P},\sigma}_1(s, i(s)) \\
&= V^{\mathrm{P},\pi^\sigma},
\end{aligned}
$$

which is the expected cumulative principal's utility of the policy π^σ that implements σ.

Finally, we observe that the size of σ is polynomial in $1/\epsilon$ and the instance size. Indeed, the number of possible promises is given by:

$$
|\mathcal{I}| \leq |\mathcal{D}_\delta| H = \mathcal{O}\left(\frac{HB}{\delta} H\right) = \mathcal{O}\left(\frac{H^4 B}{\epsilon}\right)
$$

and, since σ prescribes at most one contract for each action and promise in every state and time step, the number of contracts is bounded by:

$$
|\mathcal{J}| \leq |\mathcal{I}||\mathcal{S}||\mathcal{A}| = \mathcal{O}\left(|\mathcal{S}||\mathcal{A}| \frac{H^4 B}{\epsilon}\right).
$$

Finally, given that the size of $\mathcal{D}_\delta$ is polynomial in $1/\epsilon$ and the instance size and the oracle can be executed in polynomial time, it follows that Algorithm 2 runs in polynomial time. □

D Building an Approximation Oracle

In this section we provide the details of Algorithm 3. In particular, we show how to compute an optimal solution for Problem $\mathcal{R}^\delta_{h,s,\iota}(M)$ by means of an equivalent LP. Then, we prove that Algorithm 3 satisfies Definition 4.

D.1 AN EQUIVALENT LP

We rewrite the complete problem $\mathcal{R}^{\delta}_{h,s,\iota}(M)$, splitting Equation 2 in two different sets of constraints for clarity:

$$\max_{\substack{\alpha\in\Delta(\mathcal{A})\\ p\in\mathcal{P}^{|\mathcal{A}|}\\ \widetilde{q}:\mathcal{A}\times\mathcal{S}\to\Delta(\mathcal{D}_\delta)}} \quad \sum_{a\in\mathcal{A}}\alpha_a\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(r^P_h(s,p^a,s')+y(a,s')\right) \tag{18a}$$

$$\text{s.t.} \quad \sum_{a\in A}\alpha_a\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(r^A_h(s,p^a,a,s')+q(a,s')\right)\geq \iota-\lambda \tag{18b}$$

$$\sum_{a\in A}\alpha_a\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(r^A_h(s,p^a,a,s')+q(a,s')\right)\leq \iota+\lambda \tag{18c}$$

$$\alpha_a\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(r^A_h(s,p^a,a,s')+q(a,s')\right)\geq \alpha_a\sum_{s'\in\mathcal{S}}P_h(s'|s,\widehat{a})\left(r^A_h(s,p^a,\widehat{a},s')+q(a,s')\right)-\alpha_a\zeta\ \ \forall a,\widehat{a}\in\mathcal{A}, \tag{18d}$$

$$\sum_{\iota'\in\mathcal{D}^{M,s'}_\delta}\widetilde{q}(\iota'|a,'s)=1\quad\forall a\in A,s'\in\mathcal{S} \tag{18e}$$

We observe that the Program 18a includes some quadratic terms, specifically the products $\alpha_a\widetilde{q}(\iota'|a,s')$ and $\alpha_a p^a(s')$ in the objective function and Constraint 18b, Constraint 18c and Constraint 18d. In order to manage them, we introduce the variables $\gamma^a_{s'}$ (corresponding to $\alpha_a p^a(s')$), $\xi^{\iota'}_{a,s'}$ (corresponding to $\alpha_a\widetilde{q}(\iota'|a,s')$), and ν_a (corresponding to α_a). Then, we define the linear program $\mathcal{L}^{\delta}_{h,s,\iota}(M)$ defined as follows:

$$\max_{\substack{\gamma^a_{s'}\in\mathbb{R}_+\\ \xi^{\iota'}_{a,s'}\in\mathbb{R}_+\\ \nu_a\in\mathbb{R}_+:|\nu|_1=1}} \quad \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(\nu_a r_h(s,s')-\gamma^a_{s'}+\sum_{\iota'\in\mathcal{D}^{M,s}_\delta}\xi^{\iota'}_{a,s'}M(s',\iota')\right) \tag{19a}$$

$$\text{s.t.} \quad \sum_{a\in A}\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(\gamma^a_{s'}-\nu_a c_h(s,a)+\sum_{\iota'\in\mathcal{D}^{M,s}_\delta}\xi^{\iota'}_{a,s'}\iota'\right)\geq\iota-\lambda \tag{19b}$$

$$\sum_{a\in A}\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(\gamma^a_{s'}-\nu_a c_h(s,a)+\sum_{\iota'\in\mathcal{D}^{M,s}_\delta}\xi^{\iota'}_{a,s'}\iota'\right)\leq\iota+\lambda \tag{19c}$$

$$\sum_{s'\in\mathcal{S}}P_h(s'|s,a)\left(\gamma^a_{s'}-\nu_a c_h(s,a)+\sum_{\iota'\in\mathcal{D}^{M,s}_\delta}\xi^{\iota'}_{a,s'}\iota'\right)\geq\sum_{s'\in\mathcal{S}}P_h(s'|s,\widehat{a})\left(\gamma^a_{s'}-\nu_a c_h(s,\widehat{a})+\sum_{\iota'\in\mathcal{D}^{M,s}_\delta}\xi^{\iota'}_{a,s'}\iota'\right)-\nu_a\zeta\quad\forall a,\widehat{a}\in\mathcal{A} \tag{19d}$$

$$\gamma^a_{s'}\leq B\nu_a\quad\forall a\in\mathcal{A},s'\in\mathcal{S} \tag{19e}$$

$$\sum_{\iota'\in\mathcal{D}^{M,s'}_\delta}\xi^{\iota'}_{a,s'}=\nu_a\quad\forall a\in A,s'\in\mathcal{S} \tag{19f}$$

Now we show that the problems $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ and $\mathcal{L}^{\delta}_{h,s,\iota}(M)$ are equivalent, *i.e.,* given a solution of $\mathcal{R}^{\delta}_{h,s,\iota}(M)$, one can compute a solution of $\mathcal{L}^{\delta}_{h,s,\iota}(M)$, and viceversa.

Lemma 7. *Given a feasible solution $(\alpha, p, \widetilde{q})$ of $\mathcal{R}^{\delta}_{h,s,\iota}$, there exists a solution (ν, γ, ξ) of $\mathcal{L}^{\delta}_{h,s,\iota}$ with the same value.*

Proof. Consider the solution (ν, γ, ξ) defined as follows:

$$\begin{aligned} \nu_a &= \alpha_a && \forall a \in \mathcal{A}, \\ \gamma^a_{s'} &= \alpha_a p^a(s') && \forall a \in \mathcal{A}, s' \in \mathcal{S}, \\ \xi^{\iota'}_{a,s'} &= \widetilde{q}(\iota'|, a, s')\alpha_a && \forall a \in A, s' \in \mathcal{S}, \iota' \in D^{M,s}_{\delta}. \end{aligned}$$

Then one can show by direct calculation that (ν, γ, ξ) is a feasible solution of $\mathcal{L}^{\delta}_{h,s,\iota}$ and that its value is $G^{\delta}_{h,s,M}(\alpha, p, \widetilde{q})$. □

Lemma 8. *Given a feasible solution (ν, γ, ξ) of $\mathcal{L}^{\delta}_{h,s,\iota}$, there exists a feasible solution $(\alpha, p, \widetilde{q})$ of $\mathcal{R}^{\delta}_{h,s,\iota}$ with at least the same value.*

Proof. Consider the solution $(\alpha, p, \widetilde{q})$ defined as follows:

$$\begin{aligned} \alpha_a &= \nu_a && \forall a \in \mathcal{A}, \\ p^a(s') &= \begin{cases} \frac{\gamma^a_{s'}}{\alpha_a} & \alpha_a > 0 \\ 0 & \alpha_a = 0 \end{cases} && \forall a \in \mathcal{A}, s' \in \mathcal{S}, \\ \widetilde{q}(\iota'|a, s') &= \begin{cases} \frac{\xi^{\iota'}_{a,s'}}{\alpha_a} & \alpha_a > 0 \\ 0 & \alpha_a = 0 \end{cases} && \forall a \in \mathcal{A}, s' \in \mathcal{S}, \iota' \in D^{M,s}_{\delta}. \end{aligned}$$

One can easily observe that the following holds:

$$\begin{aligned} \nu_a &= \alpha_a && \forall a \in \mathcal{A}, \\ \gamma^a_{s'} &= \alpha_a p^a(s') && \forall a \in \mathcal{A}, s' \in \mathcal{S}, \\ \xi^{\iota'}_{a,s'} &= \widetilde{q}(\iota'|, a, s')\alpha_a && \forall a \in A, s' \in \mathcal{S}, \iota' \in D^{M,s}_{\delta}. \end{aligned}$$

Thus, by substituting the values of (ν, γ, ξ) in Constraint 18b, Constraint 18c, and Constraint 18d of Program 18a, one can observe that they are equivalent to Constraint 19b, Constraint 19c, and Constraint 19d of Program 19a respectively. Furthermore, the remaining constraints of Program 19a make sure that the payments are correctly bounded by B, and that α and $\widetilde{q}$ represent distributions. As a result, $(\alpha, p, \widetilde{q})$ is a feasible solution for Program 18a. Furthermore, direct calculations show that it achieves the same value of (ν, γ, ξ). □

By employing Lemma 8 and Lemma 7, we can show that the problem $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ can be solved in polynomial time, as stated by the following Lemma:

Lemma 9. *The problem $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ can be solved in time polynomial in $1/\delta$ and the instance size.*

Proof. Consider an optimal feasible solution (ν, γ, ξ) of $\mathcal{L}^{\delta}_{h,s,\iota}(M)$. Such a solution can be computed in polynomial time, since $\mathcal{L}^{\delta}_{h,s,\iota}(M)$ is a linear program with a number of variables and constraints polynomial in $1/\delta$ and the instance size. Now apply the transformation described in the proof of Lemma 8 to obtain a feasible solution $(\alpha, p, \tilde{q})$ of $\mathcal{R}^{\delta}_{h,s,\iota}(M)$ with the same value.

We show that such a solution is optimal. In order to do this, suppose by contradiction that there exists a different feasible solution $(\alpha', p', \tilde{q}')$ such that $G^{\delta}_{h,s,\iota}(\alpha', p', \tilde{q}') > G^{\delta}_{h,s,\iota}(\alpha, p, \tilde{q})$. Then apply the transformation provided in the proof of Lemma 7 to obtain a feasible solution (ν', γ', ξ') of $\mathcal{L}^{\delta}_{h,s,\iota}(M)$ with the same value $G^{\delta}_{h,s,\iota}(\alpha', p', \tilde{q}')$. As such, according to Lemma 7 the solution (ν', γ', ξ') achieves a value strictly larger than (ν, γ, ξ), which contradicts the hypothesis that (ν, γ, ξ) is an optimal solution. □

D.2 Approximation oracle and concavity properties

Lemma 9 implies that Algorithm 3 can be executed in polynomial time. Now we show that Algorithm 3 is an approximation oracle accordingly to Definition 4. First, we need to introduce two additional technical lemmas. In particular, we show that for any step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$, there exists a promise $\bar{\iota} \in \mathcal{D}_\delta$ such that $\mathcal{D}_\delta^{M_h^\delta, s} = [0, \bar{\iota}] \cap \mathcal{D}_\delta$. Then, we prove that the function $x : [0, \bar{\iota}] \to \mathbb{R}$ defined as $x(\iota) \coloneqq \widehat{G}_{h,s,\iota}(M)$ is concave.

Lemma 10. *Suppose that Algorithm 3 is used as an approximate oracle $O_{\iota,\delta}^{h,s}$ during the entire execution of Algorithm 2. Then for every $h \in \mathcal{H}$ and $s \in \mathcal{S}$, there exists some $\bar{\iota} \in \mathcal{D}_\delta$ such that $\mathcal{D}_\delta^{M_h^\delta, s} = [0, \bar{\iota}] \cap \mathcal{D}_\delta$; specifically $0 \in \mathcal{D}_\delta^{M_h^\delta, s}$.*

Proof. In order to prove that $0 \in \mathcal{D}_\delta^{M_h^\delta, s}$, we reason by induction. We observe that for every state $s' \in \mathcal{S}$, it holds that $M_{H+1}^\delta(s', 0) = 0$, while $M_{H+1}^\delta(s', \iota') = -\infty$ for every $\iota' > 0$. Thus, the base case $h = H + 1$ trivially satisfy the statement. Consider $h \leq H$ and any state $s \in \mathcal{S}$ and the promise $\iota = 0$. We prove that the problem $\mathcal{R}_{h,s,0}^\delta(M)$ admits a solution. We recall that there is an action $\bar{a} \in \mathcal{A}$ such that $c_H(s, \bar{a}) = 0$. Thus, we consider the solution $(\alpha, p, \widetilde{q})$ defined as follows:

$$\alpha_a = \begin{cases} 1, & a = \bar{a} \\ 0, & \text{otherwise} \end{cases} \qquad \forall a \in \mathcal{A}$$

$$p^a(s') = 0 \qquad \forall a \in \mathcal{A}, s' \in \mathcal{S}$$

$$\widetilde{q}(\iota' | a, s') = \begin{cases} 1, & \iota' = 0 \\ 0, & \text{otherwise} \end{cases} \qquad \forall a \in \mathcal{A}, s' \in \mathcal{S}, \iota' \in \mathcal{D}_\delta.$$

One can easily observe that this is a feasible solution to $\mathcal{R}_{h,s,0}^\delta(M)$. Consequently, $0 \in \mathcal{D}_\delta^{M_h^\delta, s}$.

As a further step, we prove that if a given $\bar{\iota} \in \mathcal{D}_\delta$ does not belong to $\mathcal{D}_\delta^{M_h^\delta, s}$, then no $\widehat{\iota} > \bar{\iota}$ belongs to $\mathcal{D}_\delta^{M_h^\delta, s}$ (*i.e.*, $\mathcal{R}_{h,s,\widehat{\iota}}^\delta(M)$ is unfeasible). Suppose, by contradiction, that there exists a $\widehat{\iota} > \bar{\iota}$ such that $\widehat{\iota} \in \mathcal{D}_\delta^{M_h^\delta, s}$. We build an optimization problem $\mathcal{V}_{h,s,\widehat{\iota}}^\delta(M)$ as follows: take Program 19a (*i.e.*,$\mathcal{L}_{h,s,\iota}^\delta(M)$) and let ι be a variable instead of a constant. Furthermore, add the constraint that $0 \leq \iota \leq \widehat{\iota}$. One can easily observe that $(\nu, \gamma, \xi, \widehat{\iota})$ is a feasible solution for $\mathcal{V}_{h,s,\widehat{\iota}}^\delta(M)$ if and only if (ν, γ, ξ) is a feasible solution for $\mathcal{L}_{h,s,\iota}^\delta(M)$. Since both $\mathcal{R}_{h,s,0}^\delta(M)$ and $\mathcal{R}_{h,s,\widehat{\iota}}^\delta(M)$ are feasible, by Lemma 7 $\mathcal{L}_{h,s,0}^\delta(M)$ and $\mathcal{L}_{h,s,\widehat{\iota}}^\delta(M)$ admit feasible solutions too. Consequently, since the feasibility space of $\mathcal{V}_{h,s,\widehat{\iota}}^\delta(M)$ is a polytope, it follows that there exist some α, p and $\widetilde{q}$ such that $(\alpha, p, \widetilde{q}, \bar{\iota})$ is a feasible solution for $\mathcal{V}_{h,s,\widehat{\iota}}^\delta(M)$, as $0 \leq \bar{\iota} \leq \widehat{\iota}$. As a result, $\mathcal{L}_{h,s,\bar{\iota}}^\delta(M)$ and $\mathcal{R}_{h,s,\bar{\iota}}^\delta(M)$ are feasible, thus $\bar{\iota} \in \mathcal{D}_\delta^{M_h^\delta, s}$, which is a contradiction. □

Lemma 11. *Consider a fixed $h \in \mathcal{H}$, $s \in \mathcal{S}$ and let $M = M_{h+1}^\delta$. If $\mathcal{D}_\delta^{M_h^\delta, s} = \{\iota \in \mathcal{D}_\delta \mid \iota \leq \bar{\iota}\}$ for some $\bar{\iota} \in \mathcal{D}_\delta$, then the function $x : [0, \bar{\iota}] \to \mathbb{R}$ defined as $x(\iota) \coloneqq \widehat{G}_{h,s,\iota}(M)$ is concave.*

Proof. Consider Program 19a and the value $\widehat{L}_{h,s,\iota}(M)$ of its optimal solution. Thanks to Theorem 5.1 by Bertsimas & Tsitsiklis (1997), we have that the function $x'(\iota) = \widehat{L}_{h,s,\iota}^\delta(M)$ for any fixed step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$ is concave. Furthermore, the optimal solution of Program 19a has the same value of the optimal solution of Problem 18a (see the proof of Lemma 9). Consequently:

$$x(\iota) = \widehat{G}_{h,s,\iota}(M) = \widehat{L}_{h,s,\iota}(M) = x'(\iota),$$

thus x is concave. □

D.3 Proof of Lemma 6

We split the proof of Lemma 6 into three lemmas, one for each property of Definition 4.

Lemma 12. *If Algorithm 3 is used as an approximate oracle during the entire execution of Algorithm 2, then Property 2 is satisfied.*

Proof. Consider a fixed state $s \in \mathcal{S}$, time step $h \in \mathcal{H}$ and promise $\iota \in \mathcal{D}_\delta$. In the following we let $M \coloneqq M^\delta_{h+1}$, and $(\alpha, p, z, v) = O^{h,s}_{\iota,\delta}(M)$ be computed accordingly to Algorithm 3. We argue that:

$$M(s', z(a, s')) \geq y(a, s') = \sum_{\iota' \in \mathcal{D}^{M,s'}_\delta} \widetilde{q}(\iota'|a, s') M(s', \iota'), \tag{20}$$

for every $a \in \mathcal{A}$, and $s' \in \mathcal{S}$. Such a result would imply that $v = G_{h,s,M}(\alpha, p, \widetilde{q}) \leq F_{h,s,M}(\alpha, p, z)$, by simply taking the definitions of these two functions. Consequently, if Equation 20 holds, then Property 2 is satisfied. We prove Equation 20 by induction, starting from $h = H$ (where it is trivially satisfied as $M(s', z(a, s')) = y(a, s') = 0$), up to $h = 1$.

Let us consider a fixed action $a \in \mathcal{A}$ and future state $s' \in \mathcal{S}$. Thanks to Lemma 10, there exists some $\bar{\iota} \in \mathcal{D}_\delta$ such that $\mathcal{D}^{M,s'}_\delta = \{\iota' \in \mathcal{D}_\delta \mid \iota' \leq \bar{\iota}\}$, while $z(a, s') \in \mathcal{D}^{M,s'}_\delta$ by definition (Equation 4). Furthermore, let us define:

$$\mathcal{M} \coloneqq \{(\iota', m) \mid \iota' \in \mathcal{D}^{M,s'}_\delta, m = M(s', \iota')\}.$$

If we define the concave (Lemma 11) function $x(\iota') \coloneqq \widehat{G}_{h+1,s',\iota'}(M^\delta_{h+2})$, then we have that $M(s', \iota') = x(\iota')$ for every $\iota' \in \mathcal{D}^{M,s'}_\delta$. As a result, the polygon $\text{co}(\mathcal{M})$ is the convex hull of the points $(\iota', x(\iota'))$, with $\iota' \in \mathcal{D}^{M,s'}_\delta$ and x being a concave function. We further observe that both the points $(q(a, s'), y(a, s'))$ and $(z(a, s'), M(s', z(a, s'))) = (z(a, s'), x(z(a, s'))) \in \mathcal{M}$ belong to $\text{co}(\mathcal{M})$.

Suppose that $q(a, s') \in \mathcal{D}^{M,s'}_\delta$. Then $q(a, s') = z(a, s')$, as there are no other values in $\mathcal{D}^{M,s'}_\delta$ at distance less than δ from $q(a, s')$. As the points $(z(a, s'), y(a, s'))$ and $(z(a, s'), x(z(a, s')))$ both belong to $\text{co}(\mathcal{M})$ and x is concave, it holds that $y(a, s') \leq x(z(a, s')) = M(s', z(a, s'))$.

On the other hand, if $q(a, s') \notin \mathcal{D}^{M,s'}_\delta$, then $0 < q(a, s') < \bar{\iota}$ and the discrete values $\lfloor q(a, s') \rfloor_\delta$ and $\lceil q(a, s') \rceil_\delta$ belong to $\mathcal{D}^{M,s'}_\delta$. As the polygon $\text{co}(\mathcal{M})$ is the convex hull of the points $(\iota', x(\iota'))$, with $\iota' \in \mathcal{D}^{M,s'}_\delta$ and x is a concave function, it follows that either $q(a, s') \leq x(\lfloor q(a, s') \rfloor_\delta)$ or $q(a, s') \leq x(\lceil q(a, s') \rceil_\delta)$. Consequently, by construction Equation 4 selects $z(a, s') \in \{\lfloor q(a, s') \rfloor_\delta, \lceil q(a, s') \rceil_\delta\}$ such that $M(s', z(a, s')) = x(z(a, s')) \geq y(a, s')$. □

Lemma 13. *If Algorithm 3 is used as an approximate oracle during the entire execution of Algorithm 2, then Property 1 is satisfied.*

Proof. Consider a fixed state $s \in \mathcal{S}$, time step $h \in \mathcal{H}$ and promise $\iota \in \mathcal{D}_\delta$. In the following we let $M \coloneqq M^\delta_{h+1}$, and $(\alpha, p, z, v) = O^{h,s}_{\iota,\delta}(M)$ be computed accordingly to Algorithm 3.

If $\Psi^\iota_{h,s}(\lambda = \delta, \zeta = 2\delta)$ is empty or any any solution within it has value $-\infty$, the the property is trivially satisfied (since it does not impose any constraint). We thus exclude these cases in the following.

Let now $(\bar{\alpha}, \bar{p}, \bar{z}) \in \Psi^\iota_{h,s}(\lambda = \delta, \zeta = 2\delta)$, with $F_{h,s,M}(\bar{\alpha}, \bar{p}, \bar{z}) > -\infty$. We show how to build $\widetilde{q} : \mathcal{A} \times \mathcal{S} \to \Delta(\mathcal{D}_\delta)$ such that $(\bar{\alpha}, \bar{p}, \widetilde{q})$ is a feasible solution for problem $\mathcal{R}^\delta_{h,s,\iota}(M)$, and $G_{h,s,M}(\alpha, p, \widetilde{q}) = F_{h,s,M}(\alpha, p, q)$. Specifically, let $\widetilde{q}(\cdot|a, s')$ be the distribution that put all the probability mass on $\bar{z}(a, s')$, for every $a \in \mathcal{A}$ and $s' \in \mathcal{S}$. The corresponding function q defined in Section 5.2 is thus equal to $\bar{z}$, and as a result $G_{h,s,M}(\alpha, p, \widetilde{q}) = F_{h,s,M}(\alpha, p, q)$ and $(\alpha, p, q) \in \Psi^\iota_{h,s}(\lambda = \delta, \zeta = 2\delta)$. Furthermore, we have $M(s', z(a, s')) > -\infty$ for every $s' \in \mathcal{S}$ and $a \in \mathcal{A}$, as otherwise we would have $F_{h,s,M}(\alpha, p, q) = -\infty$. As a result, $\widetilde{q} \in \mathcal{Q}^M_\delta$, and $(\bar{\alpha}, \bar{p}, \widetilde{q})$ is a feasible solution for problem $\mathcal{R}^\delta_{h,s,\iota}(M)$. We conclude that by soling problem $\mathcal{R}^\delta_{h,s,\iota}(M)$, the oracle must return:

$$v \geq G_{h,s,M}(\alpha, p, \widetilde{q}) = F_{h,s,M}(\alpha, p, q),$$

proving the statement. □

Lemma 14. *If Algorithm 3 is used as an approximate oracle during the entire execution of Algorithm 2, then Property 3 is satisfied.*

Proof. Consider a fixed state $s \in \mathcal{S}$, time step $h \in \mathcal{H}$ and promise $\iota \in \mathcal{D}_\delta$. In the following we let $M \coloneqq M^\delta_{h+1}$, and $(\alpha, p, z, v) = O^{h,s}_{\iota,\delta}(M)$ be computed accordingly to Algorithm 3. We also let $(\alpha, p, \widetilde{q})$ be the optimal solution to problem $\mathcal{R}_{h,s,\iota}(M)$ (if the problem is unfeasible, then $v = -\infty$ and Property 3 is satisfied).

We now show that starting from $(\alpha, p, \widetilde{q})$, Algorithm 3 builds (α, p, z) such that $F_{h,s,M}(\alpha, p, z) > -\infty$. Such a condition is equivalent to imposing that $M(s', z(a, s')) > -\infty$ for all possible future states $s' \in \mathcal{S}$ and actions $a \in \mathcal{A}$. Fix some $s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Recall that the set $\mathcal{D}^{M,s'}_\delta$ of promises such that $M(s', z(a, s')) > -\infty$ is a set of consecutive discrete promises, which starts from 0 and ends in some $\bar{\iota} \in \mathcal{D}_\delta$ (Lemma 10). Since $\widetilde{q}(\cdot|a, s')$ randomizes only over promises in $\bar{\iota} \in \mathcal{D}_\delta$ (see the constraints of problem $\mathcal{R}_{h,s,\iota}(M)$), it must be the case that $q(a, s) \in [0, \bar{\iota}]$. By Equation 4, $z(a, s') \in \mathcal{D}_\delta$ is either $\lfloor q(a, s') \rfloor_\delta$ or $\lceil q(a, s') \rceil_\delta$, and both of them belong to $[0, \bar{\iota}]$, which implies that $M(s', z(a, s')) > -\infty$.

In order to conclude the proof, we show that from $(\alpha, p, \widetilde{q})$, Algorithm 3 builds $(\alpha, p, z) \in \Psi^\iota_{h,s}(\lambda = 2\delta, \zeta = 4\delta)$. We have that $(\alpha, p, q) \in \Psi^\iota_{h,s}(\lambda = \delta, \zeta = 2\delta)$. Define $d = \sum_{a \in A} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s, a)(q(a, s') - z(a, s'))$. We have $|d| \leq \delta$ by Equation 4 and:

$$
\begin{aligned}
&\left| \sum_{a \in A} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^A_h(s, p_a, a, s') + z(a, s') \right) - \iota \right| \\
&= \left| \sum_{a \in A} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^A_h(s, p_a, a, s') + q(a, s') \right) - d - \iota \right| \\
&\leq \left| \sum_{a \in A} \alpha_a \sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^A_h(s, p_a, a, s') + q(a, s') \right) - \iota \right| + |d| \\
&\leq \delta + |d| \leq 2\delta,
\end{aligned}
$$

where we employed the triangle inequality and the fact that (α, p, q) satisfies Equation 2 with $\lambda = \delta$. As a result, (α, p, z) satisfies Equation 2 with $\lambda = 2\delta$. Finally, we consider Equation 3 for any two actions $a, \widehat{a}$ with $\alpha_a > 0$:

$$
\begin{aligned}
\sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^A_h(s, p_a, a, s') + z(a, s') \right) &\geq \sum_{s' \in \mathcal{S}} P_h(s'|s, a) \left(r^A_h(s, p_a, a, s') + q(a, s') \right) - \delta \\
&\geq \sum_{s' \in \mathcal{S}} P_h(s'|s, \widehat{a}) \left(r^A_h(s, p_a, \widehat{a}, s') + q(a, s') \right) - 3\delta \\
&\geq \sum_{s' \in \mathcal{S}} P_h(s'|s, \widehat{a}) \left(r^A_h(s, p_a, \widehat{a}, s') + q(a, s') \right) - 4\delta,
\end{aligned}
$$

where the second equality holds because (α, p, q) satisfies Equation 3 with $\zeta = 2\delta$, and the others by the fact that $|z(a, s') - q(a, s')| \leq \delta$. We conclude that $(\alpha, p, z) \in \Psi^\iota_{h,s}(\lambda = 2\delta, \zeta = 4\delta)$. □

Lemma 6. *If Algorithm 3 is used as an approximate oracle $O^{h,s}_{\iota,\delta}$ during the entire execution of Algorithm 2, then it satisfies Definition 4.*

Proof. The statement is proved by combining Lemma 12, Lemma 13, and Lemma 14. □

E From approximately IC to IC

In this section, we describe an algorithm to convert an ϵ-IC, direct, promise-form policy σ with value $V^{\mathrm{P},\sigma} \geq \mathrm{OPT}$ into an IC-promise form policy σ^3 that achieves at least $\mathrm{OPT} - (H+1)\sqrt{\epsilon}$. The pseudocode is provided in Algorithm 4.

Algorithm 4 `From ϵ−IC to IC`

Require: direct ϵ-IC $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ that achieves $V^{\mathrm{P},\sigma} \geq OPT$.
1: $\sigma^1 \leftarrow$ `Change-Contracts`(σ)
2: $\sigma^2 \leftarrow$ `Realign-Actions`(σ^1)
3: $\sigma^3 \leftarrow$ `Realign-Promises`(σ^2)
4: **Return** σ^3

Algorithm 4 is composed of three subprocedures. The core of the algorithm is the first subprocedure, `Change-Contracts`, which takes in input the policy σ and computes a new policy $\sigma^1 = \{I_h^1, J_h^1, \varphi_h^1, g_h^1\}_{h\in\mathcal{H}}$ by changing its contracts according to the mapping $f_{h,s} : p \mapsto (1-\sqrt{\epsilon})p + \sqrt{\epsilon}r_h(s,\cdot)$. In other words, the new contracts are a linear combination of the original contracts and the reward collected by the principal in the next states. Formally, for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$, the policy σ^1 is defined as:

$$\begin{cases} I_h^1(s) = I_h(s) & \\ J_h^1(s,\iota,a) = \{f_{h,s}(p) \mid p \in I_h(s,\iota,a)\} & \iota \in I_h(s), a \in \mathcal{A}, \\ \varphi_h^1(s, f_{h,s}(p)|s,\iota) = \varphi_h(s,p|s,\iota) & \iota \in I_h(s), a \in \mathcal{A}, p \in I_h(s,\iota,a), \\ g_h^1(s,\iota,a,f_{h,s}(p),s') = g_h(s,\iota,a,p,s') & \iota \in I_h(s), a \in \mathcal{A}, p \in I_h(s.\iota,a), s' \in \mathcal{S}. \end{cases} \tag{21}$$

The promise-form policy σ^1 ensures that the principal attains expected utility at least $\mathrm{OPT}-(H+1)\sqrt{\epsilon}$ if the agent deviates from the recommendations and plays a best response at each step, maximizing their own cumulative utility. Therefore, algorithm 4 employs its second subprocedure, `Realign-Actions`, to recompute the recommendations in order to match the best responses played by the agent. This produces a new policy σ^2 satisfying $V^{\mathrm{P},\sigma^2} \geq \mathrm{OPT} - (H+1)\sqrt{\epsilon}$. As a last step, the third subprocedure, `Realign-Promises`, recomputes the promises so that the final policy is also honest.

E.1 Changing contracts

In order to analyze the subprocedure `Change-Contracts` of Algorithm 4, we need to analyze the principal's and agent's utilities both when the agent best responds to a policy and when it follows the recommendations. In order to do so, we need to introduce two additional concepts. First, we introduce the following class of *pseudo-histories*. At step $h \in \mathcal{H}$, the set of pseudo-histories $\mathcal{T}_h^A$ is defined as:

$$\mathcal{T}_h^A \coloneqq \{t \mid t = (s_1, a_1, \dots, s_h, a_h) : s_i \in \mathcal{S}, a_i \in \mathcal{A}\},$$

while the set $\mathcal{T}^A \coloneqq \mathcal{T}_1^A \cap \mathcal{T}_2^A \cap \dots \cap \mathcal{T}_H^A$ includes all possible pseudo-histories. A pseudo-history $t \in \mathcal{T}_h^A$ ends at a decision point of the agent: after receiving the recommendation $a_h \in \mathcal{A}$ in state $s_h \in \mathcal{S}$ at step $h \in \mathcal{H}$, the agent decides wether to follow the recommendation or deviate from it. The actual action $\widehat{a} \in \mathcal{A}$ played by the agent determines the subsequent state $s_{h+1} \sim P_h(s_h, \widehat{a})$. Crucially, the pseudo-histories do not include the contracts proposed by the principal, which can be uniquely determined from the recommendations only when working with direct policies. This facilitates the analysis, as we do not have to map the histories induced by σ to those induced by σ^1 by changing the corresponding contracts. With an abuse of notation, given a history $\tau = (s_1, p_1, a_1, \dots, s_h)$ and an action $a \in \mathcal{A}$, we let $\tau \oplus a$ be the pseudo-history constructed by removing the contracts from τ and adding the recommendation a at the end.

We now introduce the concept of *agent's function*, which determines the action played by the agent at each decision point (*i.e.*, pseudo-history). Formally, an agent's function $\alpha : \mathcal{T}^A \to \mathcal{A}$ maps each pseudo-history to the action selected by the agent in that decision point. We can now define the principal's and agent's value functions given any direct policy $\pi \in \Pi$ and agent's function $\alpha : \mathcal{T}^A \to \mathcal{A}$. For every history $\tau = (s_1, p_1, a_1, \dots, s_h, p_h, a_h, s_h) \in \mathcal{T}$ ending in step $h \in \mathcal{H}$, let

$$V_h^{\mathrm{P/A},\pi,\alpha}(\tau) \coloneqq \sum_{(p,a)\in\mathcal{X}_\pi} \pi(p,a|\tau) Q_h^{\mathrm{P/A},\pi,\alpha}(\tau,p,a),$$

where:

$$Q_h^{\mathrm{P/A},\pi,\alpha}(\tau,p,a) \coloneqq \sum_{s'\in\mathcal{S}} P_h(s'|s,\alpha(\tau\oplus a))\Big(r_h^{\mathrm{P/A}}(s,p,\alpha(\tau\oplus a),s') + V_{h+1}^{\mathrm{P},\pi,\alpha}(\tau\oplus(p,a,s'))\Big).$$

We also let $V^{\text{P/A},\pi,\alpha} \coloneqq \sum_{s\in\mathcal{S}} \mu(s_1) V_1^{\text{P/A},\pi,\alpha}(s)$. With an abuse of notation, if σ is a promise-form policy implemented by π, we let $V^{\text{P/A},\sigma,\alpha} \coloneqq V^{\text{P/A},\pi,\alpha}$.

Given a policy π, we are interested in two agent's functions. The first is the *recommended agent's function* α is formally defined as:

$$\alpha(t) \coloneqq a_h \quad \forall t = (s_1, a_1, \dots, s_h, a_h) \in \mathcal{T}^A,$$

which corresponds to an agent that sticks to the recommendations. We also consider the *incentivized agent's functions*, which correspond to agents that always play rationally. Formally, an agent's function $\widehat{\alpha}$ is incentivized by π if:

$$\widehat{\alpha}(\tau \oplus a) \in \arg\max_{\widehat{a}\in\mathcal{A}} \sum_{s'\in\mathcal{S}} P_h(s'|s_h, \widehat{a}) \left(p_h(s') - c_h(s_h, \widehat{a}) + \widehat{V}_{h+1}^{\text{A},\pi}(\tau \oplus (p, a, s')), \right)$$

for every history $\tau \in \mathcal{T}$ and pair (p, a) such that $\pi(p, a|\tau) > 0$. We observe that there may be multiple incentivized policies that break ties differently. Furthermore, by definition we have $V^{\text{P/A},\pi,\alpha} = V^{\text{P/A},\pi}$ when α is the recommended agent's function, and $V^{\text{A},\pi,\widehat{\alpha}} = \widehat{V}^{\text{A},\pi}$ when $\widehat{\alpha}$ is an incentivized agent's function.

Given the definition above, our goal is to prove that given a direct promise-form policy σ, the procedure `Change-Contracts` computes a policy σ^1 such that $V^{\text{P},\sigma^1,\widehat{\alpha}} \geq V^{\text{P},\sigma} - (H+1)\sqrt{\epsilon}$. In order to do so, we first introduce this additional lemma, which allows us to rewrite the the principal's and agent's utilities without employing recursion. Given a direct policy π and a history $\tau \in \mathcal{T}$, let $\pi(a|\tau)$ the probability of recommending action $a \in \mathcal{A}$, and $p^\pi_{\tau,a}$ be the corresponding contract recommended with positive probability (or any contract if $\pi(a|\tau) = 0$). Similarly, given a pseudo-history $t \in \mathcal{T}^A$, we let p^π_t be the unique contract proposed to the agent at the decision point t. Then the following lemma holds.

Lemma 15. *Let π be a direct policy and α be an agent's function. The corresponding principal's and agent values can be written as:*

$$V^{P/A,\pi,\alpha} = \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1})|\pi, \alpha) r_h^{P/A}\left(s_h, p^\pi_t, \alpha(t), s')\right),$$

where for any pseudo-history $t = (s_1, a_1, \dots, s_h, a_h)$ and subsequent state $s \in \mathcal{S}$, the probability of $t \oplus s$ is:

$$\mathbb{P}(t \oplus (s_{h+1})|\pi, \alpha) = \mu(s_1) \prod_{k=1}^{h} \pi(a_k|(s_1, p_1, a_1, \dots, s_k)) P_k(s_{k+1}|s_k, \alpha(s_1, a_1, \dots, s_k, a_k)),$$

with $p_k = p^\pi_{(s_1,a_1,\dots,s_k,a_k)}$, $1 \leq k \leq h$.

Proof. We will provide the proof for $V^{\text{P},\pi,\alpha}$, but a similar argument proves the result for the agent's value function. Let us define, for every pseudo-history $t = (s_1, a_1, \dots, s_h, a_h)$, we define the history $\tau^t = (s_1, p_1, a_1, \dots, s_h)$ whose contracts are those proposed by σ^1, namely:

$$p_k = p^{\pi^1}_{t_k} \quad \forall 1 \leq k < h, t_k = (s_1, a_1, \dots, s_k, a_k).$$

Notice that τ^t is constructed by removing a_h, and then adding the missing contracts. With an abuse of notation, we will let $\pi(t) = \pi(\tau^t)$; in other words, when we evaluate π over a sequence of states and recommendations, the missing contracts are those that the direct π recommends at each step of the trajectory. Let us also define the set $\mathcal{T}_{\pi,h} \subseteq \mathcal{T}_h$ of histories that π can induce with positive probabilities, namely those histories $\tau = (s_1, p_1, a_1, \dots, s_h)$ such that $\pi(p_k, a_k|s_1, \dots, s_k) > 0$ for every $1 \leq k < h$. Finally, define:

$$\mathbb{P}(t \oplus s_{h+1}|\pi, \alpha, \tau) = \prod_{k=\bar{h}}^{h} \pi(a_k|s_1, \dots, s_k) P_h(s_{k+1}|s_k, \alpha(s_1, \dots, s_k, a_k)).$$

for every pseudo-history $t = (s_1, a_1, \dots, s_k) \in \mathcal{T}$ and subsequent state $s_{h+1} \in \mathcal{S}$. Finally, for the purpose of this proof, when we denote a pseudo-history with the symbol t, we denote with s_k (a_k) the k-th state (recommendation) in it, and with $h \in \mathcal{H}$ the last step, so that $t \in \mathcal{T}_h^A$.

We prove by induction that, for every $\tau = (s_1, p_1, a_1, \dots, s_{\bar{h}}) \in \mathcal{T}_{\pi,\bar{h}}$:

$$V_h^{\mathrm{P},\pi,\alpha}(\tau) = \sum_{\substack{t>\tau,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t \oplus s_{h+1}|\pi,\alpha,\tau) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}),$$

where, with an abuse of notation, the condition $t > \tau$ indicates all the pseudo-histories that begin with $(s_1, a_1, \dots, s_{\bar{h}})$.

The base step is $\bar{h} = H$. Consider a history $\tau = (s_1, p_1, a_1, \dots, s_H) \in \mathcal{T}_{\pi,H}$ ending at step $h \coloneqq \bar{h} \coloneqq H$. Then:

$$\begin{aligned}
V_H^{\mathrm{P},\pi,\alpha}(\tau) &= \sum_{(p,a)\in\mathcal{X}_\pi} \pi(p,a|\tau) \sum_{s_{h+1}\in\mathcal{S}} P_h(s_{h+1}|s_h, \alpha(\tau\oplus a)) r_h^{\mathrm{P}}(s_h, p, s_{h+1}) \\
&= \sum_{\substack{a\in\mathcal{A},\\ t=\tau\oplus a}} \pi(a|\tau) \sum_{s'\in\mathcal{S}} P_h(s_{h+1}|s_h, \alpha(\tau\oplus a)) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}) \\
&= \sum_{t>\tau} \pi(a|\tau^t) \sum_{s_{h+1}\in\mathcal{S}} P_h(s_{h+1}|s_h, \alpha(t)) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}) \\
&= \sum_{\substack{t>\tau,\\ s_{h+1}\in\mathcal{S}}} \pi(a|\tau^t) P_h(s_{h+1}|s_h, \alpha(t)) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}) \\
&= \sum_{\substack{t>\tau,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus s|\pi,\alpha,\tau) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}).
\end{aligned}$$

Now consider a step $\bar{h} < H$. For any history $\tau = (s_1, p_1, a_1, \dots, s_{\bar{h}}) \in \mathcal{T}_{\pi,\bar{h}}$, we have:

$$\begin{aligned}
&V_{\bar{h}}^{\mathrm{P},\pi,\alpha}(\tau) \\
&= \sum_{\substack{a\in\mathcal{A},\\ p:\pi(p,a|\tau)>0}} \pi(p,a|\tau) \sum_{s_{\bar{h}+1}\in\mathcal{S}} P_h(s'|s_{\bar{h}}, \alpha(\tau\oplus a)) \left(r_h^{\mathrm{P}}(s_{\bar{h}}, p, s_{\bar{h}+1}) + V_{\bar{h}+1}^{\mathrm{P},\pi,\alpha}(\tau\oplus(p,a,s_{\bar{h}+1})) \right) \\
&= \sum_{\substack{a_{\bar{h}}\in\mathcal{A},\\ p_{\bar{h}}:\pi(p_{\bar{h}},a_{\bar{h}}|\tau)>0\\ s_{\bar{h}+1}\in\mathcal{S}}} \pi(a_{\bar{h}}|\tau) P_h(s'|s_{\bar{h}}, \alpha(\tau\oplus a_{\bar{h}})) \left(r_h^{\mathrm{P}}(s_{\bar{h}}, p, s_{\bar{h}+1}) + V_{\bar{h}+1}^{\mathrm{P},\pi,\alpha}(\tau\oplus(p_{\bar{h}},a_{\bar{h}},s_{\bar{h}+1})) \right) \\
&= \sum_{\substack{t>\tau,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus s_{h+1}|\pi,\alpha,\tau) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}),
\end{aligned}$$

where the last equality holds by induction.

As a result:

$$\begin{aligned}
V^{\mathrm{P},\pi,\alpha} &= \sum_{s_1\in\mathcal{S}} \mu(s_1) V_1^{\mathrm{P},\pi,\alpha}(s_1) \\
&= \sum_{s_1\in\mathcal{S}} \mu(s_1) \sum_{\substack{t>s_1\in\mathcal{T}_\pi,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus s_{h+1}|\pi,\alpha,s_1) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}) \\
&= \sum_{\substack{t\in\mathcal{T},\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus s_{h+1}|\pi,\alpha,) r_h^{\mathrm{P}}(s_h, p_t^\pi, s_{h+1}),
\end{aligned}$$

concluding the proof. □

Thanks to Lemma 15, we can prove our main result concerning the subprocedure `Change-Contracts`. The proof leverages Lemma 15 to recover a structure similar to that found in one-shot games, then adapts the classical proof on this structure.

Lemma 16. *Given an ϵ-IC direct promise-form policy σ, let σ^1 be the policy defined according to Equation 21. Then $V^{\mathrm{P},\sigma^1,\alpha'} \geq V^{\mathrm{P},\sigma} - (H+1)\sqrt{\epsilon}$ for any agent's function α' incentivized by σ^1.*

Proof. In the following, we let $\pi = \pi^\sigma$ and $\pi^1 = \pi^{\sigma^1}$ be the history-dependent policies that implement σ and σ^1 respectively. We also let α be the recommended agent's function, and $\widehat{\alpha}$ the agent's function incentivized by π^1. Furthermore, when we denote a pseudo-history with the symbol t, we denote with s_k (a_k) the k-th state (recommendation) in it, and with $h \in \mathcal{H}$ the last step, so that $t \in \mathcal{T}_h^A$.

Observe that, for any pseudo-history $t \in \mathcal{T}^A$, subsequent state $s \in \mathcal{S}$, and agent's function $\widetilde{\alpha}$, we have:

$$\mathbb{P}(t \oplus (s_{h+1}) | \pi, \widetilde{\alpha}) = \mathbb{P}(t \oplus (s_{h+1}) | \pi^1, \widetilde{\alpha}) =: \mathbb{P}(t \oplus (s_{h+1}) | \widetilde{\alpha}),$$

as the two policies differ only in the contracts that they propose, but not in the actions that they recommend. Therefore, by employing Lemma 15, we have:

$$\begin{aligned} V^{\mathrm{P},\widetilde{\pi},\widetilde{\alpha}} &= \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widetilde{\alpha}) \left(r_h(s_h, s') - p_t^{\widetilde{\pi}}(s') \right), \\ V^{\mathrm{A},\widetilde{\pi},\widetilde{\alpha}} &= \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widetilde{\alpha}) \left(p_t^{\widetilde{\pi}}(s') - c_h(s_h, \widetilde{\alpha}(t)) \right) \end{aligned} \tag{22}$$

for any $\widetilde{\pi} \in \{\pi, \pi^1\}$ and $\widetilde{\alpha} \in \{\alpha, \widehat{\alpha}\}$. These utilities have a structure similar to those of a one-shot principal-agent game. For instance, the principal's utility in a one-shot game is computed as the sum of the probability of each outcome (which depends only on the agent's action), multiplied by the difference of the associated reward and payment. Similarly, in our sequential game $V^{\mathrm{P},\widetilde{\pi},\widetilde{\alpha}}$ is the sum of the probabilities $\mathbb{P}(t \oplus (s_{h+1}) | \widetilde{\alpha})$ (which do not depend on $\widetilde{\pi}$) multiplied by the associated reward minus the payment $(r_h(s_h, s') - p_{t,a}^{\widetilde{\pi}}(s'))$. We exploit this similarity to adapt recover a result similar to that of one-shot games Dütting et al. (2019). Specifically, we follow an argument similar to the one proposed in Castiglioni et al. (2023a) for Bayesian one-shot games.

Since σ^1 is ϵ-IC, we have $V^{\mathrm{A},\pi,\alpha} \geq V^{\mathrm{A},\pi,\widehat{\alpha}} - \epsilon$, which thanks to Equation 22 we rewrite as:

$$\begin{aligned} &\sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \alpha) \left(p_t^\pi(s_{h+1}) - c_h(s_h, \alpha(t)) \right) \\ &\qquad \geq \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widehat{\alpha}) \left(p_t^\pi(s_{h+1}) - c_h(s_h, \widehat{\alpha}(t)) \right) - \epsilon. \end{aligned} \tag{23}$$

We also have that $V^{\mathrm{A},\pi^1,\widehat{\alpha}} \geq V^{\mathrm{A},\pi^1,\alpha}$, since $\widehat{\alpha}$ is incentived by π^1. Therefore:

$$\begin{aligned} &\sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widehat{\alpha}) \left(p_t^{\pi^1}(s_{h+1}) - c_h(s_h, \widehat{\alpha}(t)) \right) \\ &\qquad \geq \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \alpha) \left(p_t^{\pi^1}(s_{h+1}) - c_h(s_h, \alpha(t)) \right). \end{aligned}$$

Noticing that $p_t^{\pi^1}(s_{h+1}) = (1 - \sqrt{\epsilon}) p_t^\pi(s_{h+1}) + \sqrt{\epsilon} r_h(s_h, s_{h+1})$ by Equation 21, we rewrite the inequality above as:

$$\begin{aligned} &\sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widehat{\alpha}) \left((1 - \sqrt{\epsilon}) p_t^\pi(s_{h+1}) + \sqrt{\epsilon} r_h(s_h, s_{h+1}) - c_h(s_h, \widehat{\alpha}(t)) \right) \\ &\qquad \geq \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \alpha) \left((1 - \sqrt{\epsilon}) p_t^\pi(s_{h+1}) + \sqrt{\epsilon} r_h(s_h, s_{h+1}) - c_h(s_h, \alpha(t)) \right). \end{aligned} \tag{24}$$

We now employ Equation 23 and Equation 23 to prove the lemma, *i.e.* to upper bound the difference $V^{\mathrm{P},\pi,\alpha} - V^{\mathrm{P},\pi^1,\widehat{\alpha}}$ with $(H+1)\sqrt{\epsilon}$.

$$
\begin{aligned}
&V^{\mathrm{P},\pi,\alpha} - V^{\mathrm{P},\pi^1,\widehat{\alpha}} \\
&= \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\alpha)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \\
&\qquad\qquad - \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|,\widehat{\alpha})\left(r_h(s_h,s_{h+1}) - p_t^{\pi^1}(s_{h+1})\right) \\
&= \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\alpha)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \\
&\qquad\qquad - \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left(r_h(s_h,s_{h+1}) - (1-\sqrt{\epsilon})p_t^{\pi}(s_{h+1}) - \sqrt{\epsilon}r_h(s_h,s_{h+1})\right) \\
&= \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\alpha)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \\
&\qquad\qquad - \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left((1-\sqrt{\epsilon})r_h(s_h,s_{h+1}) - (1-\sqrt{\epsilon})p_t^{\pi}(s_{h+1})\right) \\
&= \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\alpha)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \\
&\qquad\qquad - (1-\sqrt{\epsilon}) \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \\
&\leq \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \left(\mathbb{P}(t\oplus(s_{h+1})|\alpha) - \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\right)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) + H\sqrt{\epsilon}, \qquad (25)
\end{aligned}
$$

where the second equality holds according to Equation 21, and the last inequality holds because:

$$
\sqrt{\epsilon} \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) = \sqrt{\epsilon}V^{\mathrm{P},\pi,\widehat{\alpha}} \leq H\sqrt{\epsilon}.
$$

By Equation 25, in order to conclude the proof we must prove that:

$$
\sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \left(\mathbb{P}(t\oplus(s_{h+1})|\alpha) - \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\right)\left(r_h(s_h,s_{h+1}) - p_t^{\pi}(s_{h+1})\right) \leq \sqrt{\epsilon} \qquad (26)
$$

In order to so, we observe that Equation 24 implies:

$$
\begin{aligned}
&\sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left(p_t^{\pi}(s_{h+1}) - c_h(s_h,\widehat{\alpha}(t))\right) \\
&\qquad\qquad - \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\alpha)\left(p_t^{\pi}(s_{h+1}) - c_h(s_h,\alpha(t))\right) \\
&\geq \sqrt{\epsilon} \sum_{\substack{t\in\mathcal{T}^A,\\ s_{h+1}\in\mathcal{S}}} \mathbb{P}(t\oplus(s_{h+1})|\widehat{\alpha})\left(p_t^{\pi}(s_{h+1}) - r_h(s_h,s_{h+1})\right)
\end{aligned}
$$

$$- \sqrt{\epsilon} \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \alpha) \left(p_t^\pi(s_{h+1}) - r_h(s_h, s_{h+1})\right)$$

$$= \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \Big(\mathbb{P}(t \oplus (s_{h+1}) | \widehat{\alpha}) - \mathbb{P}(t \oplus (s_{h+1}) | \alpha)\Big) \Big(p_t^\pi(s_{h+1}) - r_h(s_h, s_{h+1})\Big).$$

By combining this result with Equation 23, we get:

$$\epsilon \geq \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \widehat{\alpha}) \left(p_t^\pi(s_{h+1}) - c_h(s_h, \widehat{\alpha}(t))\right)$$

$$- \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \mathbb{P}(t \oplus (s_{h+1}) | \alpha) \left(p_t^\pi(s_{h+1}) - c_h(s_h, \alpha(t))\right)$$

$$\geq \sqrt{\epsilon} \sum_{\substack{t \in \mathcal{T}^A, \\ s_{h+1} \in \mathcal{S}}} \Big(\mathbb{P}(t \oplus (s_{h+1}) \widehat{\alpha}) - \mathbb{P}(t \oplus (s_{h+1}) | \alpha)\Big) \Big(p_t^\pi(s_{h+1}) - r_h(s_h, s_{h+1})\Big).$$

Diving by $\sqrt{\epsilon}$, we have that Equation 26 is verified, concluding the proof. □

E.2 Realigning recommendations and promises

In this section we describe the `Realign-Actions` and `Realign-Promises` employed by Algorithm 4. At the end of subsection, we will also provide the proof of Theorem 4.

The pseudocode of `Realign-Actions` is provided in Algorithm 5. It computes the actual best response a' for every contract proposed by the policy σ^1 (Line 7), and builds σ^2 by prescribing the same contract and recommending a'. We remark that while the best way to break ties at Line 7 Algorithm 5 is to choose the action that maximizes the principal's cumulative utility, the theoretical guarantees of our approach hold regardless of how these ties are broken. In order to compute the values of the function $\widehat{V}^{\mathrm{A},\sigma^1}$, the algorithm employs dynamic programming. In particular, at Line 5 it computes and stores the value of $\widehat{V}_h^{\mathrm{A},\sigma^1}(s, \iota)$ by using the values of $\widehat{V}_{h+1}^{\mathrm{A},\sigma^1}$ computed at the previous time step. Furthermore, the values of $\widehat{V}_{h+1}^{\mathrm{A},\sigma^1}$ are employed at Line 7.

Algorithm 5 `Realign-actions`

Require: promise-form policy $\sigma^1 = \{(I_h^1, J_h^1, \varphi_h^1, g_h^1)\}_{h \in \mathcal{H}}$.

1: **for** $h = H, \dots, 1$ **do**
2: $\quad I_h^2 \leftarrow I_h^1$
3: $\quad$ **for all** $s \in \mathcal{S}, \iota \in I_h^1(s)$ **do**
4: $\quad\quad J_h^2(s, \iota, a) \leftarrow \emptyset \quad \forall a \in \mathcal{A}$
5: $\quad\quad$ Compute and store $\widehat{V}_h^{\mathrm{A},\sigma^1}(s, \iota)$
6: $\quad\quad$ **for all** $a \in \mathcal{A}, p \in J_h^1(s, \iota, a)$ **do**
7: $\quad\quad\quad a' \leftarrow \arg\max_{\widehat{a} \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P_h(s' | s, \widehat{a}) \left(r_h^{\mathrm{A}}(s, p, \widehat{a}, s') + \widehat{V}_{h+1}^{\mathrm{A},\sigma^1}(s', g_h^1(s, \iota, p, a, s'))\right)$
8: $\quad\quad\quad \varphi_h^2(p, a' | s, \iota) \leftarrow \varphi_h^1(p, a | s, \iota)$
9: $\quad\quad\quad J_h^2(s, \iota, a') \leftarrow J_h^2(s, \iota, a') \cup \{p\}$
10: $\quad\quad\quad g_h^2(s, \iota, p, a', s') \leftarrow g_h^1(s, \iota, p, a, s') \quad \forall s' \in \mathcal{S}$
11: **Return** $\sigma^2 = \{I_h^2, J_h^2, \varphi_h^2, g_h^2\}_{h \in \mathcal{H}}$

Algorithm 5 provides two main guarantees. The first is that the value of the policy σ^2 is at least $V^{\mathrm{P},\pi^{\sigma^2}} = V^{\mathrm{P},\pi^{\sigma^1},\alpha}$ for some agent's function α incentivized by σ^2. The actual incentivized function depends on how ties are broken at Line 7 Algorithm 5. The second important guarantee is that the policy σ^2 is IC. These properties are formalized in the following lemma:

Lemma 17. *Let σ^1 be a direct promise-form policy. Then Algorithm 5 returns an IC promise-form policy σ^2 with value $V^{P,\sigma^2} = V^{P,\sigma^1,\alpha}$ for some agent's function α incentivized by σ^1. Furthermore, Algorithm 5 runs in polynomial time, while $|\mathcal{I}^2| = |\mathcal{I}^1|$ and $|\mathcal{J}^2| = |\mathcal{J}^1|$.*

Proof. In the following we let π^1 and π^2 the history-dependent policies that implement σ^1 and σ^2, respectively.

Let us define, for every pseudo-history $t = (s_1, a_1, \dots, s_h, a_h)$, the history $\tau^t = (s_1, p_1, a_1, \dots, s_h)$ whose contracts are those proposed by σ^1, namely:

$$p_k = p_{t_k}^{\pi^1} \quad \forall 1 \leq k < h, t_k = (s_1, a_1, \dots, s_k, a_k).$$

Let us also $\iota_t^{\sigma^1} \coloneqq \iota_{\tau^t}^{\sigma^1}$ be the promise associated with this history. We observe that the action $a' \coloneqq f_h(s, \iota, a)$ computed at Line 7 depends only on the state $s \in \mathcal{S}$, the promise $\iota \in I_h^1(s)$, and the original recommendation $a \in \mathcal{A}$. Since σ^1 is direct, the corresponding contract $p \in J_h(s, \iota, a)$ is unique. Furthermore, by construction the action $f_h(s, \iota, a)$ maximizes the agent's utility when the principal plays the policy σ^1. Thus, there exists an agent's function α incentivized by π^1 such that $\alpha(t) = f_h(s_h, \iota_t^{\sigma^1}, a_h)$ for every $t \in \mathcal{T}$.

Let us also define, the function $V_h^{\mathrm{P},\sigma^1,\alpha} : \mathcal{S} \times \mathcal{I}^1 \to \mathbb{R}$ as:

$$V_h^{\mathrm{P/A},\sigma^1,\alpha}(s, \iota) \coloneqq \sum_{(p,a) \in \mathcal{X}_{\sigma^1}} Q_h^{\mathrm{P/A},\sigma^1,\alpha}(s, \iota, p, a),$$

where:

$$Q_h^{\mathrm{P/A},\sigma^1,\alpha}(s, \iota, p, a) \coloneqq \sum_{s' \in \mathcal{S}} P_h(s'|s, \widehat{a}) \left(r_h^{\mathrm{P/A}}(s, p, \widehat{a}, s') + V_{h+1}^{\mathrm{P/A},\sigma^1}(s', g_h^1(s, \iota, p, a, s')) \right),$$

and $\widehat{a} = f(s, \iota, a)$. Intuitively, $V_h^{\mathrm{P},\sigma^1,\alpha}$ is "equivalent" to the function $V_h^{\mathrm{P},\pi^1,\alpha}$, but it is written leveraging the promise-form representation. With an argument similar to the one employed by Lemma 1, one can show that for every step $\bar{h} \in \mathcal{H}$ and history $\bar{\tau} = (s_1, p_1, a_1, \dots, s_{\bar{h}}) \in \mathcal{T}_{\bar{h}}$, it holds that:

$$V_{\bar{h}}^{\mathrm{P/A},\pi^1,\alpha}(\bar{\tau}) = V_{\bar{h}}^{\mathrm{P},\sigma^1,\alpha}(s_{\bar{h}}, \iota_{\bar{\tau}}^{\sigma^1}), \quad \text{and } Q_{\bar{h},p,a}^{\mathrm{P/A},\pi^1,\alpha}(\bar{\tau}) = Q_{\bar{h}}^{\mathrm{P/A},\sigma^1,\alpha}(s_{\bar{h}}, \iota_{\bar{\tau}}^{\sigma^1}, p, a) \tag{27}$$

for every $(p, a) \in \mathcal{X}_{\pi^1}$.

Now we prove by induction that $V_h^{\mathrm{P/A},\sigma^2}(s, \iota) = V_h^{\mathrm{P},\sigma^1,\alpha}(s, \iota)$ for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota \in I_h^1(s)$ (observe that $I_h^1(s) = I_h^2(s)$ by construction). In other words, we prove that the value of policy σ^2 when the agent follows the recommendations of such a policy is equal to the value of σ^1 when the agent plays according to α. For the sake of exposition, we prove it for the principal's value function, but the same argument holds for the agent's one. The base step is $h = H$:

$$\begin{aligned}
V_H^{\mathrm{P},\sigma^2}(s, \iota) &= \sum_{\substack{(p,a'): \\ a' \in \mathcal{A}, \\ p \in J_H^2(s,\iota,a')}} \varphi_H^2(p, a'|s, \iota) \sum_{s' \in \mathcal{S}} P_H(s'|s, a') \left(r_H(s, s') - p(s') \right) \\
&= \sum_{\substack{(p,a): \\ a \in \mathcal{A}, \\ p \in J_H^1(s,\iota,a)}} \varphi_H^2(p, a'(s, \iota, p, a)|s, \iota) \sum_{s' \in \mathcal{S}} P_H(s'|s, a'(s, \iota, a)) \left(r_H(s, s') - p(s') \right) \\
&= \sum_{\substack{(p,a): \\ a \in \mathcal{A}, \\ p \in J_H^1(s,\iota,a)}} \varphi_H^1(p, a|s, \iota) \sum_{s' \in \mathcal{S}} P_H(s'|s, a'(s, \iota, a)) \left(r_H(s, s') - p(s') \right) = V_H^{\mathrm{P},\sigma^1,\alpha}(s, \iota),
\end{aligned}$$

where the second equality holds by construction, as for every action $a' \in \mathcal{A}$ there is another action $a \in \mathcal{A}$ such that $a' = a'(s, \iota, a)$ and $J_H^2(s, \iota, a') = J_H^1(s, \iota, a)$, while the third equality holds thanks to Line 8 Algorithm 5.

Now we consider a step $h < H$ and assume that $V_{h+1}^{\mathrm{P},\sigma^2}(s', \iota') = V_{h+1}^{\mathrm{P},\sigma^1,\alpha}(s', \iota')$ for every state $s' \in \mathcal{S}$ and promise $\iota' \in I_{h+1}^1(s') = I_{h+1}^2(s')$. Then, for every state $s \in \mathcal{S}$ and promise $\iota \in I_h^1(s)$ the following holds:

$$V_h^{\mathrm{P},\sigma^2}(s, \iota)$$

$$
\begin{aligned}
&= \sum_{\substack{(p,a'):\\ a'\in\mathcal{A},\\ p\in J_h^2(s,\iota,a')}} \varphi_h^2(p,a'|s,\iota) \sum_{s'\in\mathcal{S}} P_h(s'|s,a') \left(r_h(s,s') - p(s') + V_{h+1}^{\mathrm{P},\sigma^2}(s', g^2(s,\iota,p,a',s'))\right) \\
&= \sum_{\substack{(p,a):\\ a\in\mathcal{A},\\ p\in J_h^1(s,\iota,a)}} \varphi_h^1(p,a|s,\iota) \sum_{s'\in\mathcal{S}} P_h(s'|s,a') \left(r_h(s,s') - p(s') + V_{h+1}^{\mathrm{P},\sigma^2}(s', g^2(s,\iota,p,a',s'))\right) \\
&= \sum_{\substack{(p,a):\\ a\in\mathcal{A},\\ p\in J_h^1(s,\iota,a)}} \varphi_h^1(p,a|s,\iota) \sum_{s'\in\mathcal{S}} P_h(s'|s,a') \left(r_h(s,s') - p(s') + V_{h+1}^{\mathrm{P},\sigma^2}(s', g^1(s,\iota,p,a,s'))\right) \\
&= \sum_{\substack{(p,a):\\ a\in\mathcal{A},\\ p\in J_h^1(s,\iota,a)}} \varphi_h^1(p,a|s,\iota) \sum_{s'\in\mathcal{S}} P_h(s'|s,a') \left(r_h(s,s') - p(s') + V_{h+1}^{\mathrm{P},\sigma^1,\alpha}(s', g^1(s,\iota,p,a,s'))\right) \\
&= V_h^{\mathrm{P},\sigma^1,\alpha}(s,\iota).
\end{aligned}
$$

where the second equality holds, with $a' = a'(s,\iota,p,a)$, by means of the same argument employed in the base step $h = H$, while the third one holds since $g^1(s,\iota,p,a,s') = g^2(s,\iota,p,a',s')$ by construction (see Line 10 Algorithm 5), and the second to last one thanks to the inductive hypothesis. Consequently, it holds that $V_h^{\mathrm{P},\sigma^2}(s,\iota) = V_h^{\mathrm{P},\sigma^1,\alpha}(s,\iota)$ for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota \in I_h^1(s) = I_h^2(s)$.

As a result, we can show that $V^{\mathrm{P},\pi^2} \geq V^{\mathrm{P},\pi^1,\alpha}$, as:

$$
\begin{aligned}
V^{\mathrm{P},\pi^2} &= \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\pi^2}(s) \\
&= \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\sigma^2}(s, i^2(s)) \\
&= \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\sigma^1,\alpha}(s, i^1(s)) \\
&= \sum_{s\in\mathcal{S}} \mu(s) V_1^{\mathrm{P},\pi^1,\alpha}(s) = V^{\mathrm{P},\pi^1,\alpha},
\end{aligned}
$$

where the first equality holds by definition, the second one thanks to Lemma 1, the third one thanks to the result proved by induction above and because σ^1 and σ^2 share the initial promises, and the second to last one holds according to Equation 27.

As a further step, we observe that the same argument provided above shows that $Q_h^{\mathrm{A},\sigma^2}(s,\iota,p,a) = Q_h^{\mathrm{A},\sigma^1,\alpha}(s,\iota,p,a)$. Furthermore, Algorithm 5 ensures that $\widehat{V}_h^{\sigma^2}(s,\iota) = V_h^{\mathrm{A},\sigma^1,\alpha}(s,\iota)$. This implies that the policy σ^2 is IC.

To conclude the proof, we observe that σ^2 uses the same set of promises of σ^1 and contracts (up to the mapping $a \mapsto f(a,s,\iota)$). It is also easy to verify that Algorithm 5 runs in time polynomial in the sizes of the instance and σ^1. □

The last subprocedure of Algorithm 4 is `Realing-Promises`, provided in Algorithm 6. While σ^2 is IC, it is not honest. Indeed, it keeps the promises of the original policy σ, which work correctly as "internal states" of the policy, but have no semantic meaning. For completeness, Algorithm 6 changes the promises of policy σ^2 and builds a new policy σ^3 that is honest. Intuitively it maps each promise ι_2 of σ^2 to a new honest promise ι_3. In particular, for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota_2 \in I_h^2(s)$, the new honest promise is $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s,\iota_2)$ (we recall that since σ^2 is IC, it holds that $V_h^{\mathrm{A},\sigma^2}(s,\iota_2) = \widehat{V}_h^{\mathrm{A},\sigma^2}(s,\iota_2)$). The other functions of the policy σ^3 are defined in order to preserve this mapping, *i.e.*, $\varphi_h^3(s,\iota_3) \coloneqq \varphi_h^2(s,\iota_2)$ and $g_h^3(s,\iota_3,p,a,s') \coloneqq V_h^{\mathrm{A},\sigma^2}(s', g^2(s,\iota_2,p,a,s'))$.

Algorithm 6 `Realign-promises`

Require: IC promise-form policy $\sigma^2 = \{(I_h^2, J_h^2, \varphi_h^2, g_h^2)\}_{h\in\mathcal{H}}$.
Initialize dictionaries $C_h(s, \iota_3)$ and $G_h(s, \iota_3, \iota_2)$ to empty.
for $h = H, \dots, 1$ **do**
  for all $s \in \mathcal{S}$ **do**
    for all $\iota_2 \in I_h^2(s)$ **do**
      Compute and store $V_h^{\mathrm{A},\sigma^2}(s, \iota_2)$
      $\iota_3 \leftarrow V_h^{\mathrm{A},\sigma^2}(s, \iota_2)$
      $C_h(s, \iota_3) \leftarrow C_h(s, \iota_3) \cup \{\iota_2\}$
      $G_h(s, \iota_3, \iota_2) \leftarrow$the value of Equation 28.
      $I_h^3(s) \leftarrow I_h^3(s) \cup \{\iota_3\}$
    for all $\iota_3 \in I_h^3(s)$ **do**
      $\iota_2 \leftarrow \arg\max_{\widetilde{\iota}_2 \in C_h(s,\iota_3)} G_h(s, \iota_3, \widetilde{\iota}_2)$
      $\varphi_h^3(p, a|s, \iota_3) \leftarrow \varphi_h^2(p, a|s, \iota_2) \quad \forall a \in \mathcal{A}, p \in J_h^2(s, \iota_2, a)$
      $J_h^3(s, \iota_3, a) \leftarrow J_h^2(s, \iota_2, a) \quad \forall a \in \mathcal{A}$
      $g_h^3(s, \iota_3, p, a, s') \leftarrow V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s, \iota_2, p, a, s')) \quad \forall a \in \mathcal{A}, p \in J_h^2(s, \iota_2, a), s' \in \mathcal{S}$
      Compute and store $V_h^{\mathrm{P},\sigma^3}(s, \iota_3)$
Return $\sigma^3 = \{I_h^3, J_h^3, \varphi_h^3, g_h^3\}_{h\in\mathcal{H}}$

It may happen that two different promises $\iota_2, \widetilde{\iota}_2 \in I_h^2(s)$ are mapped to the same new promise $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s, \iota_2) = V_h^{\mathrm{A},\sigma^2}(s, \widetilde{\iota}_2)$. This produces a "conflict", as generally we cannot have $\varphi_h^3(s, \iota_3) = \varphi_h^2(s, \iota_2)$ and $\varphi_h^3(s, \iota_3) = \varphi_h^2(s, \widetilde{\iota}_2)$ at the same time. To solve this problem, we define the functions $\varphi_h^3(s, \iota_3)$ and $g_h^3(s, \iota_3, p, a, s')$ considering the old promise that achieves the largest principal value, *i.e.*, $\varphi_h^3(s, \iota_3) \coloneqq \varphi_h^2(s, \iota_2)$ and $g_h^3(s, \iota_3, p, a, s') \coloneqq V_h^{\mathrm{A},\sigma^2}(s', g_h^2(s, \iota_2, p, a, s'))$ if they provide the largest principal value, or the same using $\widetilde{\iota}_2$ in place of ι_2 otherwise. The value that the principal achieves by using an old promise $\iota_2 \in I_h^2(s)$ to solve a conflict in $\iota_3 \in I_h^3(s)$ is:

$$G_h(s, \iota_3, \iota_2) \coloneqq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p, a|s, \iota_2) \sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r_h^{\mathrm{P}}(s, p, s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', \iota') \right), \tag{28}$$

where $\iota' \coloneqq V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s, \iota_2, p, a, s'))$. Formally, Algorithm 6 makes use of two dictionaries for each time step, $C_h : \mathcal{S} \times \mathcal{I}^3 \to \mathcal{I}^2$ and $G_h : \mathcal{S} \times \mathcal{I}^3 \times \mathcal{I}^2 \to \mathbb{R}$. C_h stores, for every state $s \in \mathcal{S}$ and promise $\iota_3 \in \mathcal{I}_h^3(s)$, the set of old promises $C_h(s, \iota_3) \subseteq I_h^2(s)$ that have been mapped into ι_3. When $|C_h(s, \iota_3)| > 1$, a conflict arises, and Algorithm 6 employs the dictionary G_h, which is filled according to Equation 28. Let us also observe that Algorithm 6 employs dynamic programming to compute the values of the functions $V_h^{\mathrm{A},\sigma^2}$ and $V_h^{\mathrm{P},\sigma^3}$.

The policy σ^3 computed by Algorithm 6 is guaranteed to provide at least the same utility of σ^2, while being both IC and honest. This result is formalized in the following lemma.

Lemma 18. *Let σ^2 be an IC promise-form policy. Then Algorithm 6 returns an IC and honest promise-form policy σ^3 with value $V^{P,\sigma^3} \geq V^{P,\sigma^2}$. Furthermore, Algorithm 6 runs in time polynomial in $|\mathcal{J}^2|$, $|\mathcal{I}^2|$ and the instance size, while $|\mathcal{J}^3| \leq |\mathcal{J}^2|$ and $|\mathcal{I}^3| \leq |\mathcal{I}^2|$.*

Proof. As a first step, we observe that for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$, by construction:

$$I_h^3(s) = \{\iota_3 \in \mathbb{R} \mid \iota_3 = V_h^{\mathrm{A},\sigma^2}(s, \iota_2), \iota \in I_h^2(s)\}.$$

and:

$$C_h(s, \iota_3) = \{\iota_2 \in I_h^2(s) \mid \iota_3 = V_h^{\mathrm{A},\sigma^2}(s, \iota_2)\} \quad \forall \iota_3 \in I_h^3(s),$$

where $V_h^{\mathrm{A},\sigma^2} = \widehat{V}_h^{\mathrm{A},\sigma^2}$ as σ^2 is IC. Furthermore, $\mathcal{X}_{\sigma^3} \subseteq \mathcal{X}_{\sigma^2}$, as σ^3 prescribes only pairs $(p, a) \in \mathcal{X}$ that are prescribed also by σ^2.

Now we will first show that $V^{\mathrm{P},\sigma^3} \geq V^{\mathrm{P},\sigma^2}$, and then that the policy σ^3 is both honest and IC.

We prove by recursion that for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, and promise $\widetilde{\iota}_2 \in I_h^2(s)$, it holds that $V_h^{\mathrm{P},\sigma^2}(s,\widetilde{\iota}_2) \leq V_h^{\mathrm{P},\sigma^3}(s,\iota_3)$, where $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s,\widetilde{\iota}_2)$. Let:

$$\iota_2 = \underset{\bar{\iota}_2 \in C_h(s,\iota_3)}{\arg\max}\ G_h(s,\iota_3,\bar{\iota}_2).$$

be the value selected at Line 11 Algorithm 6. We also observe that $\iota_2, \widetilde{\iota}_2 \in C_h(s,\iota_3)$.

The base step is $h = H$. Then, according to Equation 28 and the definition of ι_2:

$$\sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_H^2(p,a|s,\iota_2) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s') \geq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_H^2(p,a|s,\widetilde{\iota}_2) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s').$$

Consequently:

$$\begin{aligned}
V_H^{\mathrm{P},\sigma^2}(s,\widetilde{\iota}_2) &= \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_H^2(p,a|s,\widetilde{\iota}_2) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s') \\
&\leq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_H^2(p,a|s,\iota_2) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s') \\
&= \sum_{(p,a)\in\mathcal{X}_{\sigma^3}} \varphi_H^3(p,a|s,\iota_3) \sum_{s'\in\mathcal{S}} P_H(s'|s,a) r_H^{\mathrm{P}}(s,p,s') = V_H^{\mathrm{P},\sigma^3}(s,\iota_3),
\end{aligned}$$

where the second to last equality holds because $\mathcal{X}_{\sigma^3} \subseteq \mathcal{X}_{\sigma^2}$ and, as of Line 12 Algorithm 6, $\varphi_H^2(p,a|s,\iota_2) = \varphi_H^3(p,a|s,\iota_3)$.

Now consider $h < H$ and assume that $V_{h+1}^{\mathrm{P},\sigma^2}(s,\widetilde{\iota}_2) \leq V_{h+1}^{\mathrm{P},\sigma^3}(s,\iota_3)$ for every state $s \in \mathcal{S}$, and promises $\widetilde{\iota}_2 \in I_{h+1}^2(s)$ and $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s,\widetilde{\iota}_2)$. Let $h \in \mathcal{H}$, $s \in \mathcal{S}$, and consider two promises $\widetilde{\iota}_2 \in I_h^2(s)$ and $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s,\widetilde{\iota})$. Let also ι_2 be the value selected at Line 11 Algorithm 6. Then, according to Equation 28 and the definition of ι_2:

$$\begin{aligned}
&\sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\iota_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \Big(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s,\iota_2,p,a,s'))) \Big) \\
&\quad \geq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\widetilde{\iota}_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \Big(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s,\widetilde{\iota}_2,p,a,s'))) \Big).
\end{aligned} \tag{29}$$

Consequently:

$$\begin{aligned}
&V_h^{\mathrm{P},\sigma^2}(s,\widetilde{\iota}_2) \\
&= \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\widetilde{\iota}_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^2}(s', g_h(s,\widetilde{\iota}_2,p,a,s')) \right) \\
&\leq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\widetilde{\iota}_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s,\widetilde{\iota}_2,p,a,s'))) \right) \\
&\leq \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\iota_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s,\iota_2,p,a,s'))) \right) \\
&= \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p,a|s,\iota_2) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', g_h^3(s,\iota_3,p,a,s')) \right) \\
&= \sum_{(p,a)\in\mathcal{X}_{\sigma^3}} \varphi_h^3(p,a|s,\iota_3) \sum_{s'\in\mathcal{S}} P_h(s'|s,a) \left(r_h^{\mathrm{P}}(s,p,s') + V_{h+1}^{\mathrm{P},\sigma^3}(s', g_h^3(s,\iota_3,p,a,s')) \right) \\
&= V_h^{\mathrm{P},\sigma^3}(s,\iota_3)
\end{aligned}$$

where the first inequality holds thanks to the inductive hypothesis and the second one according to Equation 29. Furthermore, by construction, $\varphi_h^2(p,a|s,\iota_2) = \varphi_h^3(p,a|s,\iota_3)$ and $g_h^2(s,\iota_2,p,a,s') = g_h^3(s,\iota_3,p,a,s')$, which proves the last three equalities.

Let $h \in \mathcal{H}$ and $s \in \mathcal{S}$. Given the structure of the set $I_h^3(s)$, for every $\iota_3 \in I_h^3(s)$ there is a promise $\iota_2 \in I_h^2(s)$ such that $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s, \widetilde{\iota}_2)$. Since $V_{h+1}^{\mathrm{P},\sigma^3}(s, \iota_3) \geq V_{h+1}^{\mathrm{P},\sigma^2}(s, \widetilde{\iota}_2)$, it follows that $V^{\mathrm{P},\pi^3} \geq V^{\mathrm{P},\pi^2}$, where π^3 and π^2 are the history-dependent policies that implement σ^3 and σ^2 respectively.

As a further step, we show that the policy σ^3 is honest. We prove that for every step $h \in \mathcal{H}$, state $s \in \mathcal{S}$ and promise $\iota_3 \in I_h^3(s)$, the following holds:

$$\sum_{(p,a)\in\mathcal{X}_{\sigma^3}} \varphi_h^3(p, a|s, \iota_3) \sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r_h^{\mathrm{A}}(s, p, a, s') + g_h^3(s, \iota_3, p, a, s')\right) = \iota_3,$$

where, by construction, $\iota_3 = V_h^{\mathrm{A},\sigma^2}(s, \iota_2)$ for some promise $\iota_2 \in I_h^2(s)$. We observe that:

$$\begin{aligned}
&\sum_{(p,a)\in\mathcal{X}_{\sigma^3}} \varphi_h^3(p, a|s, \iota_3) \sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r_h^{\mathrm{A}}(s, p, a, s') + g_h^3(s, \iota_3, p, a, s')\right) \\
&= \sum_{(p,a)\in\mathcal{X}_{\sigma^2}} \varphi_h^2(p, a|s, \iota_2) \sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r_h^{\mathrm{A}}(s, p, a, s') + V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s, \iota_2, p, a, s'))\right) \\
&= V_h^{\mathrm{A},\sigma^2}(s, \iota_2),
\end{aligned}$$

where the first equality holds because $\varphi_h^3(p, a|s, \iota_3) = \varphi_h^2(p, a|s, \iota_2)$ (see Line 12) and $g_h^3(s, \iota_3, p, a, s') = V_{h+1}^{\mathrm{A},\sigma^2}(s', g_h^2(s, \iota_2, p, a, s'))$ given the definition of $I_{h+1}^3(s')$ and Line 14 Algorithm 6. As a result, σ^3 is honest.

As a further step, we show that σ^3 is IC by means of Lemma 3. Thus, we have to prove that for any step $h \in \mathcal{H}$, state $s \in \mathcal{S}$, promise $\iota_3 \in I_h^3(s)$, actions $a, \widehat{a} \in \mathcal{A}$ and contract $p \in J_h^3(s, \iota_3, a)$, the following holds:

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} P_h(s'|s, a) \left(r_h^{\mathrm{A}}(s, p, a, s') + g_h^3(s, \iota_3, p, a, s')\right) \geq \\
&\qquad \sum_{s'\in\mathcal{S}} P_h(s'|s, \widehat{a}) \left(r_h^{\mathrm{A}}(s, p, \widehat{a}, s') + g_h^3(s, \iota_3, p, a, s')\right).
\end{aligned}$$

Let $\iota_2 \in \mathcal{I}_h^2(s)$ be such that $\iota_2 = V_h^{\mathrm{A},\sigma^2}(s, \iota_2)$. Then $g_h^3(s, \iota_3, p, a, s') = V_h^{\mathrm{A},\sigma^2}(s, g_h^2(s, \iota_2, p, a, s'))$. Since the policy σ^2 is IC, it follows that the inequality above is satisfied. As a result, σ^3 is IC according to Lemma 3.

In order to conclude the proof, we observe that Algorithm 6 runs in time polynomial in the sizes of σ^2 and the instance. Furthermore, since for every step $h \in \mathcal{H}$ and state $s \in \mathcal{S}$ we have that $|I_h^3(s)| \leq |I_h^2(s)|$ by construction, it follows that $|\mathcal{I}^3| \leq |\mathcal{I}^2|$. A similar result holds for the contracts: fix a step $h \in \mathcal{H}$, a state $s \in \mathcal{S}$ and a promise $\iota^3 \in I_h^3(s)$. For every promise $\iota_3 \in I_h^3(s)$ there is a promise $\iota_2 \in I_h(s)$ such that $J_h^3(s, \iota_3, a) = J_h^2(s, \iota_2, a)$ for every $a \in \mathcal{A}$. Such a promise $\iota_2 = \iota_2(\iota_3)$ is different for every $\iota_3 \in I_h^3(s)$. As $|I_h^3(s)| \leq |I_h^2(s)|$, it follows that $|\mathcal{J}^3| \leq |\mathcal{J}^2|$. □

We conclude by providing the proof of Theorem 4.

Theorem 4. *Given a direct and ϵ-IC promise-form policy $\sigma = \{(I_h, J_h, \varphi_h, g_h)\}_{h\in\mathcal{H}}$ that achieves at least $V^{\mathrm{P},\sigma} \geq$ OPT, Algorithm 4 (detailed in Appendix E) computes an IC and honest promise-form policy σ^3 in time polynomial in the size of σ and the instance. Furthermore, the value of σ^3 is at least $V^{\mathrm{P},\sigma^3} \geq$ OPT $- (H+1)\sqrt{\epsilon}$, while its size is such that $|\mathcal{I}^3| \leq |\mathcal{I}|$ and $|\mathcal{J}^3| \leq |\mathcal{J}|$.*

Proof. By Lemma 16, the policy σ^1 computed by Algorithm 4 is such that $V^{\mathrm{P},\sigma^1,\alpha} \geq V^{\mathrm{P},\sigma} - (H+1)\sqrt{\epsilon}$ for some agent's function α. Since $V^{\mathrm{P},\sigma} \geq$ OPT, it follows that:

$$V^{\mathrm{P},\sigma^1,\alpha} \geq \text{OPT} - (H+1)\sqrt{\epsilon}.$$

As a further step we observe that, thanks to Lemma 17, the policy σ^2 is IC and achieves a value of:

$$V^{\mathrm{P},\sigma^2} = V^{\mathrm{P},\sigma^1,\alpha} \geq \text{OPT} - (H+1)\sqrt{\epsilon}.$$

Finally, according to Lemma 18 the policy σ^2 is both IC and honest, and achieves a value of:

$$V^{\mathrm{P},\sigma^3} \geq V^{\mathrm{P},\sigma^2} \geq \mathrm{OPT} - (H+1)\sqrt{\epsilon}.$$

To conclude the proof, we observe that every subprocedure of Algorithm 4 can be executed in polynomial time, and, according to Lemma 17 and Lemma 18, the size of σ^3 is given by $|\mathcal{I}^3| \leq |\mathcal{I}|$ and $|\mathcal{J}^3| \leq |\mathcal{J}|$. □