OPTIMAL ROBUST SUBSIDY POLICIES FOR IRRATIONAL AGENT IN PRINCIPAL-AGENT MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate a principal-agent problem modeled within a Markov Decision Process, where the principal and the agent have their own rewards. The principal can provide subsidies to influence the agent's action choices, and the agent's resulting action policy determines the rewards accrued to the principal. Our focus is on designing a robust subsidy scheme that maximizes the principal's cumulative expected return, even when the agent displays bounded rationality and may deviate from the optimal action policy after receiving subsidies.

As a baseline, we first analyze the case of a perfectly rational agent and show that the principal's optimal subsidy coincides with the policy that maximizes social welfare, the sum of the utilities of both the principal and the agent. We then introduce a bounded-rationality model: the globally ϵ-incentive-compatible agent, who accepts any policy whose expected cumulative utility lies within ϵ of the personal optimum. In this setting, we prove that the optimal robust subsidy scheme problem simplifies to a one-dimensional concave optimization. This reduction not only yields a clean analytical solution but also highlights a key structural insight: optimal subsidies are concentrated along the social-welfare-maximizing trajectories. We further characterize the loss in social welfare—the degradation under the robust subsidy scheme compared to the maximum achievable—and provide an upper bound on this loss. Finally, we investigate a finer-grained, state-wise ϵ-incentive-compatible model. In this setting, we show that under two natural definitions of state-wise incentive-compatibility, the problem becomes intractable: one definition results in a non-Markovian agent action policy, while the other renders the search for an optimal subsidy scheme NP-hard.

1 INTRODUCTION

The principal–agent problem (often modeled as a Stackelberg game) has long been central to the study of strategic interactions where one party acts on behalf of another, yet with potentially misaligned incentives. This setting arises frequently in economics and governance: for example, governments design taxes, subsidies, and public investments to guide individual behavior toward socially beneficial outcomes. However, in decentralized markets, each participant ultimately pursues their own utility, and centralized guidance can only partially influence outcomes. A similar dynamic appears in machine learning, where reinforcement learning with human feedback (RLHF) is employed to align large language models (LLMs) with societal values such as ethics and legal compliance. In both cases, the principal faces the fundamental challenge of shaping an agent's behavior without direct control, while respecting both parties' interests.

In this paper, we investigate the principal–agent problem within the framework of a Markov Decision Process (MDP), where the principal can provide subsidies to influence the agent's action choices. More specifically, in our setting, each action under each state yields two distinct rewards: one for the principal and one for the agent. The principal may also assign non-negative subsidies to actions. The agent selects an action policy based on its own reward combined with subsidies offered by the principal. The principal, in turn, strategically designs these subsidies to influence the agent's choices, aiming to maximize the principal's overall payoff, which equals the total principal's reward associated with the agent's chosen action minus the subsidies provided.

A natural assumption in such models is that the agent always behaves rationally, selecting the trajectory that maximizes the sum of the agent's own reward and the subsidies provided by the principal. Yet in practice, this assumption is often violated: agents may deviate from perfect rationality due to bounded cognition, incomplete information, or limited computational power. For example, in economics, individuals may fail to optimize utility precisely because of uncertainty or behavioral biases. Similarly, in reinforcement learning, approximate training algorithms may yield suboptimal policies due to limited exploration or finite computation.

Motivated by these considerations, we ask:

How should the principal design subsidies when the agent may behave irrationally?

Our goal is to identify a **robust subsidy scheme** that guarantees the principal the best possible expected cumulative return in the worst-case scenario.

Our Contributions We introduce a theoretical framework based on Markov Decision Processes (MDPs) to model the principal-agent problem and formulate the design of an optimal robust subsidy scheme as a minimax optimization problem. Within this framework, we systematically analyze three agent models: the perfectly rational agent, the globally ϵ-incentive-compatible (IC) agent, and the state-wise ϵ-IC agent. For each model, we provide structural insights and algorithmic solutions.

We first study a *perfectly rational agent* as a baseline, who always selects actions that maximize its own reward. In Theorem 3.1, we characterize the optimal subsidy scheme and show in Proposition 3.2 that it suffices to subsidize only actions that maximize social welfare, defined as the sum of the principal's and agent's utilities. Under this scheme, the agent's best-response policy aligns with the social welfare-maximizing policy, establishing a clear benchmark for incentive alignment.

Next, we consider *globally ϵ-IC agents*, who tolerate at most an ϵ loss relative to their optimal reward under a given subsidy scheme. Unlike perfectly rational agents, these agents may adopt stochastic policies, making the principal's optimization a nontrivial bi-level problem. Theorem 4.1 shows that this problem can be equivalently reduced to maximizing a one-dimensional concave function over a bounded interval, allowing efficient solution via standard first-order methods. Structurally, in Proposition 4.2, we show the optimal subsidy mirrors the perfectly rational case by exclusively rewarding actions that align with maximizing social welfare; and, in the worst-case response, the agent's policy will assign positive probability to the socially optimal actions, though it may also mix with other actions. We further provide a quantitative analysis of the gap between the total payoff achieved under this robust scheme and the maximum possible social welfare, as shown in Proposition 4.3.

Finally, in Section 5, we examine *state-wise ϵ-IC agents*, for which the ϵ-tolerance must hold at each individual state. Two natural formalizations arise, each presenting distinct challenges. In the first formalization, the agent's worst-case response may necessitate a non-Markovian policy, thereby violating the foundational assumptions of the MDP framework and introducing history dependence that makes the problem computationally intractable. In the second formalization, while the agent's worst-case response remains polynomial-time computable, Theorem 5.1 demonstrates that the principal's problem becomes NP-hard. These findings illustrate that, although state-wise constraints are conceptually appealing, they introduce significant computational and modeling complexities that limit practical applicability.

Related work The principal–agent problem, a central concept in economics (Ross, 1973; Grossman & Hart, 1992), arises when a principal delegates tasks to an agent whose actions may be guided by self-interest. This framework underpins both contract theory (Laffont & Maskin, 1981; Guruganesh et al., 2021) and mechanism design (Myerson, 1982; Kadan et al., 2017).

Recent work has examined this problem in the setting of Markov Decision Processes (MDPs). Research in this area falls into two broad directions. The first, information design, seeks to influence the agent's beliefs, as in Bayesian persuasion (Gan et al., 2022; Wu et al., 2022; Bernasconi et al., 2023). The second, more closely aligned with our work, focuses on shaping the agent's incentives through policy teaching (Zhang & Parkes, 2008; Banihashem et al., 2022) or environment/model design (Thoma et al., 2024; Yu & Ho, 2022). A comprehensive survey is provided by Dütting et al. (2024). Among these, two approaches are most closely related to our study:

Contract-based models. This line of research integrates contract theory with MDPs, assuming the principal observes only states and offers state-dependent payments. Prior studies analyze subgame perfect equilibrium (Wu et al., 2024; Ivanov et al., 2024), showing that history-dependent contracts are necessary for farsighted agents (Bollini et al., 2024). These works typically assume perfectly rational agents and establish that the optimal contract design problem is NP-hard.

Reward shaping. In Reward shaping, the principal modifies the agent's incentives via additional rewards for specific state–action pairs, subject to a fixed budget (Ben-Porat et al., 2024), with the design problem remaining NP-hard. Extensions address behavioral uncertainty through robust reward design (Wu et al., 2025). In contrast, we incorporate incentive costs directly into the principal's objective, treating them as part of payoff optimization rather than an external constraint.

2 Problem Formulation

The Principal-Agent MDP Model We consider a principal-agent problem modeled as a time-inhomogeneous, finite-horizon Markov Decision Process (MDP). In this setting, the principal aims to achieve a goal by influencing an agent's actions. The principal can offer subsidies to incentivize the agent to follow a policy that benefits the principal.

Formally, we define the problem instance using the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{H}, \mathcal{P}, r_P, r_A, \hat{s}, \Pi \rangle$, where:

- $\mathcal{S}$ is the set of the finite states and $\mathcal{A}$ is the set of actions. We assume that both states and actions are *discrete*.
- $\mathcal{H} = \{0, 1, \cdots, H-1\}$ is the set of time steps, with H representing the time horizon.
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{H} \to \Delta(\mathcal{S})$ is the transition kernel , where $P(s'|s, a, h)$ indicates the probability of transferring to state $s' \in \mathcal{S}$ after executing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at timestep $h \in \mathcal{H}$.
- $r_P, r_A : \mathcal{S} \times \mathcal{A} \times \mathcal{H} \to \mathbb{R}$ are the reward functions of the principal and the agent, respectively, where $r_P(s, a, h)$ (resp. $r_A(s, a, h)$) denotes the reward obtained by the principal (resp. agent) when the agent executes action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$ at timestep $h \in \mathcal{H}$.
- Without loss of generality, $\hat{s}$ is the fixed starting state for the agent.

Subsidy Scheme and Action Policy The principal commits to a **subsidy scheme** $\Delta r : \mathcal{S} \times \mathcal{A} \times \mathcal{H} \to \mathbb{R}_{\geq 0}$. Here, $\Delta r(s, a, h)$ is a non-negative payment from the principal to the agent for taking action a in state s at timestep h. We denote the set of all feasible subsidy policies as $\mathcal{R}_\Delta$.

Given a subsidy Δr on action a in state s at timestep h, the effective rewards for the principal and agent become:

$$r_P^{\Delta r}(s, a, h) = r_P(s, a, h) - \Delta r(s, a, h) \quad \text{and} \quad r_A^{\Delta r}(s, a, h) = r_A(s, a, h) + \Delta r(s, a, h)$$

The agent observes the subsidy scheme and then chooses a Markovian **action policy** $\pi : \mathcal{S} \times \mathcal{H} \to \Delta(\mathcal{A})$. Based on the agent's (ir)rationality, for any given Δr, the agent will choose a policy from a specific set of feasible policies, which we denote by $\Pi(\Delta r)$.

Value Functions For any player $i \in \{P, A\}$, subsidy scheme Δr, and agent policy π, we define the standard state-value and action-value functions via the Bellman expectation equations:

$$V_i^{\pi, \Delta r}(s, h) = \sum_{a \in \mathcal{A}} \pi(a|s, h) Q_i^{\pi, \Delta r}(s, a, h)$$

$$Q_i^{\pi, \Delta r}(s, a, h) = r_i^{\Delta r}(s, a, h) + \sum_{s' \in \mathcal{S}} P(s'|s, a, h) V_i^{\pi, \Delta r}(s', h+1)$$

with the terminal condition $V_i^{\pi, \Delta r}(s, H) = 0$. Furthermore, we use $\overline{V}_A^{\Delta r}(s, h)$ and $\overline{Q}_A^{\Delta r}(s, a, h)$ to denote the optimal state-value and action-value functions attainable by the agent,

$$\overline{V}_A^{\Delta r}(s, h) = \max_a \overline{Q}_A^{\Delta r}(s, a, h)$$

$$\overline{Q}_A^{\Delta r}(s, a, h) = r_A(s, a, h) + \sum_{s' \in \mathcal{S}} P(s'|s, a, h) \overline{V}_A^{\Delta r}(s', h+1)$$

Additionally, $V_i^{\pi,\Delta r=0}(s,h)$, $Q_i^{\pi,\Delta r=0}(s,a,h)$, $\overline{V}_A^{\Delta r=0}(s,h)$ and $\overline{Q}_A^{\Delta r=0}(s,a,h)$ denote the corresponding value in the absence of subsidies.

Social Welfare We define social welfare as the aggregate reward of both the principal and the agent: $r_{\text{sw}}(s,a,h) \triangleq r_P(s,a,h) + r_A(s,a,h)$, which remains unaffected by the subsidy term Δr.

The *social welfare value functions*, V_{sw}^{π} and Q_{sw}^{π}, characterize the expected social welfare under an agent policy π:

$$V_{\text{sw}}^{\pi}(s,h) = \sum_{a\in\mathcal{A}} \pi(a|s,h)\, Q_{\text{sw}}^{\pi}(s,a,h),$$
$$Q_{\text{sw}}^{\pi}(s,a,h) = r_{\text{sw}}(s,a,h) + \sum_{s'\in\mathcal{S}} P(s'|s,a,h)\, V_{\text{sw}}^{\pi}(s',h+1).$$

Analogously, the optimal social welfare value functions, V_{sw}^{*} and Q_{sw}^{*}, are defined as:

$$V_{\text{sw}}^{*}(s,h) = \max_{a\in\mathcal{A}} Q_{\text{sw}}^{*}(s,a,h),$$
$$Q_{\text{sw}}^{*}(s,a,h) = r_{\text{sw}}(s,a,h) + \sum_{s'\in\mathcal{S}} P(s'|s,a,h)\, V_{\text{sw}}^{*}(s',h+1).$$

An action a is said to be **social-welfare-maximizing** in state s at timestep h if it is greedy with respect to the optimal Q-value, i.e., $a \in \arg\max_{a'\in\mathcal{A}} Q_{\text{sw}}^{*}(s,a',h)$.

Optimization Objective We consider a robust formulation where the principal seeks a subsidy scheme that performs best against the agent's worst-case response. The agent's adversarial action policy to a subsidy Δr is an agent policy $\pi_{\Delta r}$ that minimizes the principal's expected return within the feasible set $\Pi(\Delta r)$:

$$\pi_{\Delta r} \in \arg\min_{\pi\in\Pi(\Delta r)} V_P^{\pi,\Delta r}(\hat{s},h=0)$$

The principal's objective is to find the optimal subsidy scheme Δr^* that maximizes this worst-case outcome. The optimal value for the principal is therefore:

$$\text{OPT} \triangleq \max_{\Delta r\in\mathcal{R}_\Delta} \min_{\pi\in\Pi(\Delta r)} V_P^{\pi,\Delta r}(\hat{s},h=0) \tag{2.1}$$

3 Warm-up: The perfectly rational Agent

We begin with the simplest setting of a perfectly rational agent, defined as an agent that seeks to maximize its cumulative reward. Although this scenario is conceptually straightforward, it provides a crucial foundation for the subsequent analysis of more complex, irrational agents. We formalize this concept as follows.

Definition 3.1 (Perfectly Rational Agent)**.** *Given a subsidy scheme Δr, the action policy $\pi \in \Pi_0(\Delta r)$ of a perfectly rational agent satisfies the constraint*

$$V_A^{\pi,\Delta r}(\hat{s},h=0) \geq \overline{V}_A^{\Delta r}(\hat{s},h=0).$$

Tie-breaking Rule A tie-breaking rule dictates the agent's choice when multiple actions yield identical rewards. In this setting with a perfectly rational agent, we assume that when two options provide the same personal reward, the agent selects the more cooperative action—that is, the one that benefits the principal more. For example, consider a single state with two actions. Both give the agent a reward of 0, but the principal receives 2 for the first action and 0 for the second. Even a negligible subsidy on the first action makes it strictly preferred. As the subsidy approaches zero, the agent's choice remains the action with a higher principal value. Thus, tie-breaking systematically favors actions that increase the principal's payoff. This assumption allows for a tractable proof of optimality in this section, but it is important to note that we will not rely on this rule in the more general frameworks developed later in the paper.

3.1 Optimal Subsidy Scheme

With the definition of perfect rationality, we now address the problem of determining the optimal subsidy scheme Δr^*. The following theorem characterizes the principal's optimal payoff and the optimal subsidy scheme. Detailed proof is deferred to Appendix A.3.

Theorem 3.1 (Optimal Subsidy Scheme)**.** *For a perfectly rational agent, the principal's optimal payoff is given by*

$$V_{sw}^*(\hat{s}, h=0) \;-\; \overline{V}_A^{\Delta r=0}(\hat{s}, h=0),$$

that is, the maximum attainable social welfare (over all action policies) minus the maximum reward the agent can obtain in the absence of subsidies. Furthermore, there exists an optimal subsidy scheme Δr^* such that, for every state–action–timestep triple (s, a, h),

$$\Delta r^*(s,a,h) \;=\; \overline{V}_A^{\Delta r=0}(s,h) - \overline{Q}_A^{\Delta r=0}(s,a,h). \tag{3.1}$$

Proof Sketch. The principal's optimal payoff is bounded above by $V_{\text{sw}}^*(\hat{s}, h=0) - \overline{V}_A^{\Delta r=0}(\hat{s}, h=0)$, since the total value of the principal and agent cannot exceed the maximum possible social welfare, and the agent will not accept less than their stand-alone value without subsidies. This upper bound is achieved under the subsidy scheme Δr^* defined in equation (3.1). Under this scheme, the agent's adjusted Q-values are equalized across all actions: $Q_A^{\Delta r^*}(s,a,h) = \overline{V}_A^{\Delta r=0}(s,h)$ for all (s,a,h). Thus, the agent is indifferent among all actions. Our provisional tie-breaking rule then ensures the agent selects actions that maximize the principal's reward, allowing the principal's payoff to exactly reach the upper bound. □

Although Theorem 3.1 identifies an optimal subsidy scheme that provides transfers on nearly all actions, the following proposition shows that, to achieve optimal rewards, the principal needs to subsidize only the social-welfare-maximizing actions. The detailed proof is deferred to Appendix A.4.

Proposition 3.2 (Social Welfare)**.** *There exists an optimal subsidy scheme Δr_{sw} that assigns positive transfers exclusively to social-welfare-maximizing actions. Under Δr_{sw}, the agent implements social-welfare-maximizing agent policy π_{sw}, allowing the principal to attain the maximum achievable social welfare.*

4 Optimal Policies for Globally ϵ-IC Agents

When an agent is no longer perfectly rational, the optimality of its response ceases to be the sole factor guiding its decisions. To model such bounded rationality, a natural approach is to assume that the agent can tolerate a maximum reward loss of ϵ, in line with the classical notion of ϵ-incentive compatibility (IC). However, since we are dealing with sequential decision-making, several interpretations of ϵ-IC are possible. Here, we focus on the so-called *globally ϵ-IC agent*, which constrains only the cumulative reward loss over the entire decision horizon.

Definition 4.1. *An agent is a globally ϵ-IC agent if and only if, given a subsidy scheme Δr, the action policy $\pi \in \Pi_\epsilon^g(\Delta r)$ satisfies*

$$V_A^{\pi,\Delta r}(\hat{s}, h=0) \geq \overline{V}_A^{\Delta r}(\hat{s}, h=0) - \epsilon.$$

4.1 Optimal Subsidy Scheme

We now consider the problem of determining the optimal subsidy scheme Δr^*. Unlike the perfectly rational case, the agent's best-response policy may be stochastic.

To handle this, we reformulate the objective (2.1) using occupancy measures. Specifically, let $\mu(s,a,h)$ denote the probability that the agent takes action a in state s at timestep h. Replacing the policy π with its corresponding occupancy measure μ, the optimization problem becomes

$$\max_{\Delta r \in \mathcal{R}_\Delta} \min_{\mu \in M(\Delta r)} \sum_{s,a,h} \mu(s,a,h)\big(r_P(s,a,h) - \Delta r(s,a,h)\big), \tag{4.1}$$

where $M(\Delta r)$ is the set of occupancy measures satisfying the following constraints:

$$\text{Initial state:} \quad \sum_a \mu(\hat{s}, a, h=0) = 1, \quad \sum_a \mu(s, a, h=0) = 0 \quad \forall s \neq \hat{s}, \tag{4.2a}$$

$$\text{Transition:} \quad \sum_a \mu(s, a, h) = \sum_{s', a'} \mu(s', a', h-1) P(s|s', a', h-1), \tag{4.2b}$$

$$\text{Non-negativity:} \quad \mu(s, a, h) \geq 0, \tag{4.2c}$$

$$\text{Global } \epsilon\text{-IC:} \quad \sum_{s,a,h} \mu(s, a, h)\big(r_A(s, a, h) + \Delta r(s, a, h)\big) \geq \overline{V}_A^{\Delta r}(\hat{s}, h=0) - \epsilon. \tag{4.2d}$$

Directly solving this program is challenging for two main reasons. First, the feasible set of μ is not fixed but depends on the choice of Δr, creating a coupling between the inner and outer variables that distinguishes our setting from standard minimax formulations. Second, defining $f(\Delta r) = \min_{\mu \in M(\Delta r)} \sum_{s,a,h} \mu(s, a, h)\big(r_P(s, a, h) - \Delta r(s, a, h)\big)$ shows that $f(\Delta r)$ is not concave in Δr (see Appendix A.2.1 for example). Consequently, the outer problem $\max_{\Delta r} f(\Delta r)$ is not a concave maximization , which rules out standard convex optimization methods.

In our main theorem, we show the problem can be reformulated to a one-dimensional concave optimization (Theorem 4.1). The approach leverages the dual of the inner optimization problem and swaps the order of optimization between the subsidy scheme Δr and the dual variables (α, V). The optimal subsidy scheme can then be expressed as the difference between the V-function and Q-function, analogous to the perfectly rational case.

Theorem 4.1. *The optimization problem (4.1) is equivalent to maximizing a concave function $F(x)$, formulated as*

$$\max_{x \in [0,1)} F(x) = x V_{sw}^*(\hat{s}, h=0) - V_x^*(\hat{s}, h=0) - \frac{x}{1-x}\epsilon,$$

where, for each state s and timestep h, $V_x^*(s, h) \triangleq \max_\pi \Big\{ x V_{sw}^\pi(s, h) - V_P^{\pi, \Delta r=0}(s, h) \Big\}$.

Furthermore, for an optimal x^*, there exists an optimal subsidy scheme Δr^* such that

$$\Delta r^*(s, a, h) = V_{x^*}^*(s, h) - Q_{x^*}^*(s, a, h) \tag{4.3}$$

where $Q_{x^*}^*(s, a, h) \triangleq x^* r_{sw}(s, a, h) - r_P(s, a, h) + \sum_{s' \in \mathcal{S}} P(s'|s, a, h) V_{x^*}^*(s', h+1)$.

Proof. We begin by considering the inner program over the state-action occupancy measure μ for a fixed subsidy scheme Δr. This program is a linear program. By introducing dual variables $\alpha \in \mathbb{R}_+$ for the globally ϵ-IC constraint (4.2d) and $V \in \mathbb{R}^{|\mathcal{S}|(H+1)}$ for the transition (4.2a) and initial state (4.2b) constraints, we can express the problem in its dual form. Combining this with the outer maximization over Δr, α, and V yields the following optimization problem:

$$\max_{\alpha \geq 0, V} \; V(\hat{s}, h=0) - \alpha\epsilon + \alpha \max_{\Delta r} \overline{V}_A^{\Delta r}(\hat{s}, h=0)$$

such that $V(s, h) \leq r_P(s, a, h) - \alpha r_A(s, a, h) - (1+\alpha)\Delta r(s, a, h) + \sum_{s' \in \mathcal{S}} P(s'|s, a, h) V(s', h+1)$ for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $h \in \mathcal{H}$; and with the terminal condition $V(s, H) = 0$ for any state $s \in \mathcal{S}$.

Next, we exchange $\max_{\Delta r}$ and $\max_{\alpha \geq 0, V}$ and analyze maximization over Δr for a fixed V and α. Notice that the objective is non-decreasing with respect to Δr, since $\overline{V}_A^{\Delta r}(\hat{s}, h=0)$ represents the maximum value attainable by the agent under the subsidy Δr. Additionally, the constraints impose an upper bound on each $\Delta r(s, a, h)$:

$$\Delta r(s, a, h) \leq \frac{1}{1+\alpha}\Big(-V(s, h) + \sum_{s' \in \mathcal{S}} P(s'|s, a, h) V(s', h+1) + r_P(s, a, h) - \alpha r_A(s, a, h) \Big).$$

Thus, the optimal choice for Δr is to take this upper bound, making the inequality hold with equality. Given α and V, substituting the optimal choice of Δr, the RHS of the above inequality, into

$\overline{V}_A^{\Delta r}(\hat{s}, h=0) = \max_\pi \mathbb{E}_\pi\Big[\sum_{t=0}^{H-1} r_A(s_t,a_t,t) + \Delta r(s_t,a_t,t)\Big]$ gives

$$\overline{V}_A^{\Delta r}(\hat{s}, h=0) = \max_\pi \frac{1}{1+\alpha}\,\mathbb{E}_\pi\Bigg[\sum_{t=0}^{H-1}\big(r_P(s_t,a_t,t) + r_A(s_t,a_t,t)\big) + \sum_{s_{t+1}\in\mathcal{S}} P(s_{t+1}|s_t,a_t,t)V(s_{t+1},t+1) - V(s_t,t)\Bigg]$$
$$= \frac{1}{1+\alpha}\Big(V_{\text{sw}}^*(\hat{s}, h=0) - V(\hat{s}, h=0)\Big).$$

Substituting this back, the problem reduces to

$$\max_{\alpha\geq 0}\max_V \frac{1}{1+\alpha}V(\hat{s},h=0) + \frac{\alpha}{1+\alpha}V_{\text{sw}}^*(\hat{s},h=0) - \alpha\epsilon$$
$$\text{s.t. } V(s,h) \leq r_P(s,a,h) - \alpha r_A(s,a,h) + \sum_{s'\in\mathcal{S}} P(s'|s,a,h)V(s',h+1),$$
$$V(s,H)\leq 0.$$

Observing the inner optimization over $V(s,h)$ coincides with form of minimizing cumulative reward in an MDP with modified reward $r_P - \alpha r_A$. By letting $x = \frac{\alpha}{1+\alpha}$ and introducing $V_x^*(s,h)$ equals $= -\frac{1}{1+\alpha}$ times the optimal value of $V(s,h)$, the formulation equals

$$\max_{x\in(0,1]} \quad x\cdot V_{\text{sw}}^*(\hat{s},h=0) - V_x^*(\hat{s},h=0) - \frac{x}{1-x}\epsilon$$

where $$V_x^*(\hat{s},h=0) \triangleq -(1-x)\cdot\min_\pi\left\{V_P^{\pi,\Delta r=0}(\hat{s},h=0) - \frac{x}{1-x}V_A^{\pi,\Delta r=0}(\hat{s},h=0)\right\}$$
$$= \max_\pi\big\{xV_{\text{sw}}^\pi(\hat{s},h=0) - V_P^{\pi,\Delta r=0}(\hat{s},h=0)\big\}.$$

Restricting π to deterministic action policies does not change the value of $V_x^*(\hat{s},h=0)$, and under this restriction, $V_x^*(\hat{s},h=0)$ is the maximum of finitely many linear functions in x, so the objective function is concave over the interval $[0,1)$. □

Markovian vs. Non-Markovian A process is called **Markovian** if it depends solely on its current state, independent of its past trajectory. Conversely, a process is **non-Markovian** if it can depend on historical states, i.e., it possesses "memory."

In our framework, both the principal and the agent may adopt non-Markovian strategies. For example, the principal might determine subsidies based not only on the agent's current action but also on past actions. Similarly, in equation (4.1), the agent could adopt a non-Markovian globally ϵ-IC policy to reduce the principal's reward. Nevertheless, the following two key observations establish that it suffices to restrict attention to Markovian strategies.

First observation: Given a Markovian subsidy scheme of the principal, there always exists a Markovian globally ϵ-IC policy for the agent that minimizes the principal's reward. This follows from the fact that the inner optimization problem in equation (4.1) is a linear program. Any non-Markovian ϵ-IC policy can be represented by an occupancy measure $\mu(s,a,h)$, which specifies the probability of taking action a in state s at timestep h. Such an occupancy measure can always be replicated by a Markovian policy, ensuring identical rewards for both the principal and the agent.

Second observation: Among all possible subsidy schemes—Markovian or non-Markovian—the Markovian scheme specified in equation (4.3) is optimal. A non-Markovian scheme can be transformed into a Markovian one by augmenting the state space to encode the relevant history. By Theorem 4.1, for each state–action pair in this augmented representation, the scheme in equation (4.3) coincides exactly with its Markovian counterpart.

Remark We briefly examine the boundary cases of x^* and ϵ in Theorem 4.1. When $\epsilon = 0$, as $x^* \to 1$, the principal's value approaches $V_{\text{sw}}^*(\hat{s},h=0) - V_A^{\Delta r=0}(\hat{s},h=0)$, consistent with the tie-breaking rule in the perfectly rational case. This shows that the globally ϵ-IC agent naturally generalizes the perfectly rational agent.

4.2 Action Policy

According to Theorem 4.1, the optimal subsidy scheme Δr^* takes a form similar to that in the perfectly rational case. The following proposition shows that the principal can still allocate positive transfers exclusively to the social-welfare-maximizing actions. Furthermore, the agent is still willing to cooperate with the principal to a certain extent by choosing one social-welfare-maximizing agent policy π_{sw} with probability x^*, the optimal solution in Theorem 4.1. The detailed proof of the following proposition is deferred to Appendix A.5.

Proposition 4.2 (Optimal subsidy scheme and action policy)**.** *There exists an optimal subsidy scheme Δr_{sw} that assigns positive reward transfers solely to social-welfare-maximizing actions. Meanwhile, there exists a globally ϵ-IC action policy $\pi_{\Delta r_{sw}}$ minimizing the principal's reward, which is the mixture of a social-welfare-maximizing agent policy π_{sw} and one other action policy, placing a weight of at least x^* on π_{sw}.*

Proof Sketch. The proof relies on two key insights. First, under the optimal subsidy scheme Δr^*, the policy π_{sw} achieves the maximum agent expected cumulative reward, $\overline{V}_A^{\Delta r^*}(\hat{s}, h = 0)$. This implies that it is sufficient to provide subsidies only along the trajectories induced by π_{sw}, without affecting the optimal value for the principal. Second, there exists an action policy $\hat{\pi}$ whose agent value falls below $\overline{V}_A^{\Delta r^*}(\hat{s}, h = 0) - \epsilon$, which can be combined with π_{sw} to form the globally ϵ-IC policy $\pi_{\Delta r_{\text{sw}}}$, such that the dual of the global ϵ-incentive compatibility constraint is tight. □

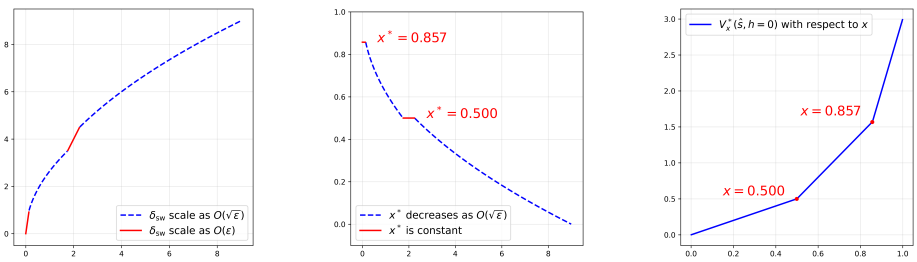
4.3 Social Welfare

We define the social welfare gap δ_{sw} as the difference between the maximum attainable welfare and the welfare achieved under the optimal subsidy scheme Δr^*. When $\epsilon \to +\infty$, the agent can effectively bypass the global ϵ-IC constraint and freely select any action policy. In this limit, the welfare gap becomes $\delta_{\text{sw}} = V_{\text{sw}}^*(\hat{s}, h = 0) - \min_\pi V_{\text{sw}}^\pi(\hat{s}, h = 0)$. Our objective is to characterize the upper bound on δ_{sw} and the rate at which social welfare declines as a function of ϵ, particularly in the regime where ϵ remains small. We first establish the following upper bound on δ_{sw}.

Proposition 4.3. *Given ϵ and the corresponding optimal solution $x^* \in (0, 1)$, the social welfare gap is $\delta_{sw} = \frac{\epsilon}{1-x^*}$ and it is upper bounded by $O(\sqrt{\epsilon})$.*

This $O(\sqrt{\epsilon})$ bound can be achieved in certain specific cases (see Appendix A.2.2 for an example). However, in most cases, the social welfare gap δ_{sw} exhibits two different growth rates—$O(\sqrt{\epsilon})$ or $O(\epsilon)$—depending on whether V_x^* is differentiable at x^*. A concrete example is provided below, while detailed discussions are deferred to Appendix A.6.1.

Example Consider a single-period scenario with three actions and $\epsilon = 1$. For the first action, the principal's reward is 7 and the agent's reward is 3. For the second action, the principal's reward is 1 and the agent's reward is 2. For the third action, the principal's reward is 1 and the agent's reward is 0. Figure 1a shows that x can grow at rates of $O(\epsilon)$ and $O(\sqrt{\epsilon})$, corresponding to the cases in Figure 1b where x remains constant or grows at $O(\sqrt{\epsilon})$. Figure 1c depicts the piecewise-linear relationship between $V_x^*(\hat{s}, h = 0)$ and x, where the constant-x value in Figure 1b coincides with the break point of $V_x^*(\hat{s}, h = 0)$, a non-differentiable point of the objective function.



(a) δ_{sw} versus ϵ (b) x^* versus ϵ (c) $V_x^*(\hat{s}, h = 0)$ versus x

Figure 1: Curves of δ_{sw} and x^* versus ϵ when $V_x^*(\hat{s}, h = 0)$ is non-differentiable.

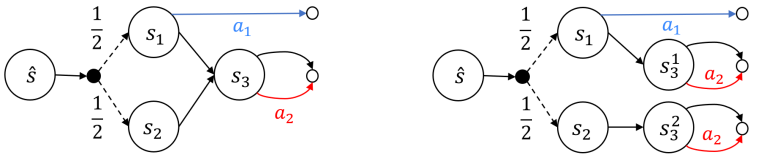
5 State-Wise ϵ-IC Agent

In this section, we examine the *state-wise ϵ-IC agent*, which differs from the globally ϵ-IC agent in that incentive compatibility is enforced locally at each state and decision step. Intuitively, such an agent ensures that its chosen action remains within ϵ of the best immediate value available at that decision point. While the idea is simple, constructing a mathematically consistent and tractable formalization is more subtle. We provide two definitions below.

Value-Consistent State-Wise ϵ-IC Agent We first define the *value-consistent state-wise ϵ-IC agent*, where the agent's action at each state must approximate the optimal reward within ϵ.

Definition 5.1. *An agent is a value-consistent state-wise ϵ-IC agent if, under a subsidy scheme Δr, the induced policy $\pi \in \Pi_\epsilon^v(\Delta r)$ satisfies $V_A^{\pi,\Delta r}(s,h) \geq \overline{V}_A^{\Delta r}(s,h) - \epsilon$ for all $s \in \mathcal{S}$ and $h \in \mathcal{H}$.*

A key challenge with this formulation is that the agent's policy minimizing the principal's reward under a given subsidy scheme may be **non-Markovian**. In such cases, the agent's policy cannot be represented within polynomial size.



(a) Original MDP instance (b) History-dependent MDP expansion

Figure 2: Illustration of value-consistent state-wise ϵ-IC agents.

To illustrate, consider the post-subsidy MDP in Figure 2a, where (i) for action a_1 at s_1: principal reward 100, agent reward 3; (ii) for action a_2 at s_3: principal reward 2, agent reward 2; and (iii) for all other actions: reward 0. Under a **Markovian policy**, the value-consistent state-wise ϵ-IC agent minimizes the principal's reward by selecting a_2 at s_3, and steering toward s_3 from s_1. This yields a principal reward of 2. However, under a **non-Markovian policy**, we can duplicate s_3 into two history-dependent states, s_3^1 and s_3^2. At s_3^1, the agent always selects a_2, while at s_3^2, the agent mixes between two actions with equal probability. This reduces the principal's expected reward to 1.5.

Greedy State-Wise ϵ-IC Agent To avoid non-Markovian behavior, we introduce the *greedy state-wise ϵ-IC agent*, which replaces recursive value computations with greedy look-ahead. Once the subsidy scheme is fixed, $\overline{V}_A^{\Delta r}$ becomes deterministic, and the agent greedily minimizes the principal's value through local decisions.

Definition 5.2. *An agent is a greedy state-wise ϵ-IC agent if, under subsidy scheme Δr, the induced policy $\pi \in \Pi_\epsilon^s(\Delta r)$ satisfies, for all $s \in S$, $h \in \mathcal{H}$:*

$$\sum_{a \in \mathcal{A}} \pi(a|s,h) \Big(r_A^{\Delta r}(s,a,h) + \sum_{s' \in \mathcal{S}} P(s'|s,a,h)\, \overline{V}_A^{\Delta r}(s',h+1) \Big) \;\geq\; \overline{V}_A^{\Delta r}(s,h) - \epsilon.$$

However, even in this simplified greedy setting, designing the principal's optimal subsidy scheme remains computationally intractable. The complete proof is deferred to Appendix A.7.

Theorem 5.1. *Given a greedy state-wise ϵ-IC agent, computing the principal's optimal subsidy scheme is NP-hard.*

6 Conclusion

In this paper, we study a principal-agent problem with the aim of designing a robust subsidy scheme that maximizes the cumulative expected return in the presence of an irrational agent. We demonstrate that, under the globally ϵ-IC assumption, the optimal subsidy scheme can be effectively determined, representing a natural extension of the perfectly rational case. We further show that formulating the state-wise ϵ-IC follower is computationally challenging. As future work, it would be interesting to consider scenarios in which the principal does not have prior knowledge of the agent's reward function or the value of ϵ, such as in a learning-based setting.

REFERENCES

Kiarash Banihashem, Adish Singla, Jiarui Gan, and Goran Radanovic. Admissible policy teaching through reward design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6037–6045, 2022.

Omer Ben-Porat, Yishay Mansour, Michal Moshkovitz, and Boaz Taitler. Principal-agent reward shaping in mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 9502–9510, 2024.

Martino Bernasconi, Matteo Castiglioni, Alberto Marchesi, and Mirco Mutti. Persuading farsighted receivers in mdps: the power of honesty. *Advances in Neural Information Processing Systems*, 36:14987–15014, 2023.

Matteo Bollini, Francesco Bacchiocchi, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. Contracting with a reinforcement learning agent by playing trick or treat. *arXiv preprint arXiv:2410.13520*, 2024.

Paul Dütting, Michal Feldman, and Inbal Talgam-Cohen. Algorithmic contract theory: A survey. *Foundations and Trends® in Theoretical Computer Science*, 16(3-4):211–412, 2024.

Jiarui Gan, Rupak Majumdar, Goran Radanovic, and Adish Singla. Bayesian persuasion in sequential decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5025–5033, 2022.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of insurance economics: Readings in economics and finance*, pp. 302–340. Springer, 1992.

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.

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

Ohad Kadan, Philip J Reny, and Jeroen M Swinkels. Existence of optimal mechanisms in principal-agent problems. *Econometrica*, 85(3):769–823, 2017.

Jean-Jacques Laffont and Eric Maskin. *The theory of incentives: An overview*. Université des sciences sociales, Faculté des sciences économiques, 1981.

Roger B Myerson. Optimal coordination mechanisms in generalized principal–agent problems. *Journal of mathematical economics*, 10(1):67–81, 1982.

Stephen A Ross. The economic theory of agency: The principal's problem. *The American economic review*, 63(2):134–139, 1973.

Vinzenz Thoma, Barna Pásztor, Andreas Krause, Giorgia Ramponi, and Yifan Hu. Contextual bilevel reinforcement learning for incentive alignment. *Advances in Neural Information Processing Systems*, 37:127369–127435, 2024.

Jibang Wu, Zixuan Zhang, Zhe Feng, Zhaoran Wang, Zhuoran Yang, Michael I Jordan, and Haifeng Xu. Sequential information design: Markov persuasion process and its efficient reinforcement learning. *arXiv preprint arXiv:2202.10678*, 2022.

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.

Shuo Wu, Haoxiang Ma, Jie Fu, and Shuo Han. Robust reward design for markov decision processes. *Journal of Artificial Intelligence Research*, 84, 2025.

Guanghui Yu and Chien-Ju Ho. Environment design for biased decision makers. In *IJCAI*, pp. 592–598, 2022.

Haoqi Zhang and David C Parkes. Value-based policy teaching with active indirect elicitation. In *AAAI*, volume 8, pp. 208–214, 2008.