Computational Dynamic Mechanism Design

Sadie Zhao
Harvard University
sadie_zhao@g.harvard.edu

Amy Greenwald
Brown University
amy_greenwald@brown.edu

Yanchen Jiang
Harvard University
yanchen_jiang@g.harvard.edu

Denizalp Goktas
Cornell Tech
goktasdenizalp@gmail.com

Yiling Chen
Harvard University
yiling@seas.harvard.edu

Abstract

Dynamic mechanism design (DMD) provides a principled framework for coordinating self-interested agents whose private information evolves over time, but history dependence and dynamic incentive constraints make general analytical solutions difficult. Using a dynamic revelation principle, we formulate optimal dynamic mechanism design as a bilevel optimization problem over parameterized direct mechanisms: the principal optimizes the mechanism, while agent-side unilateral deviation problems verify incentive compatibility. We model each agent's problem as a parameterized POMDP, reduce it to a sufficient-statistics MDP, and run stochastic natural policy gradient to compute approximate best-response policies with finite-sample error bounds. These policies serve as oracle outputs for the principal's problem, yielding first-order optimization procedures with convergence guarantees that approximate stationary solutions under bounded oracle error. This formulation separates two representation choices, principal-side history compression and agent-side sufficient statistics, and guides the design of our experiments, which explore how representations can affect optimization in DMD.

1 Introduction

Dynamic mechanism design (DMD) is a framework in which to study how to design rules for settings in which agents' private information evolves over time and current decisions affect future incentives, such as dynamic auctions [6, 5, 1, 50, 39], dynamic screening and contracting [19, 3, 9], dynamic matching [31], and dynamic public goods provision [16, 15]. Their broad applicability gives rise to their main difficulty: a mechanism must condition on long histories of behavior and anticipate how outcomes shape agents' future information, incentives, and strategies. In a *bandit auction*, for example, a seller repeatedly sells perishable goods to buyers whose values evolve with past allocations. Agents face intertemporal trade-offs: winning today can reveal private information that affects future prices, while shading reports can reduce current allocations and alter future opportunities. As a result, DMD is richer than static mechanism design, and substantially more challenging.

Mechanism design is difficult because the principal chooses rules while anticipating induced equilibrium behavior among self-interested agents. In static environments, Myerson's revelation principle [46] provides a powerful simplification: the designer can fix truthful reporting as the desired equilibrium and search for direct mechanisms that rationalize this outcome. Sugaya and Wolitzky [53]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

extend this logic to multistage games. In the dynamic environments we study, their results allow us to restrict attention to direct dynamic mechanisms that induce truthful reporting as a Bayesian Nash equilibrium. Although this reduction is very powerful, analytical DMD remains difficult because dynamic incentive constraints are history dependent. Past approaches have imposed structure in order to reduce these constraints to local or otherwise manageable conditions: Pavan et al. [50] develop a dynamic envelope theorem framework, while Kakade et al. [39] identify restrictions such as separability of information across time. These results provide deep economic insights, but they also motivate computational approaches that are applicable in a wider range of environments.

This paper develops a computational approach to DMD in the spirit of differentiable economics [27]. We formulate optimal DMD as a bilevel optimization (eq. (2)) over a parameterized class of direct mechanisms. The upper level optimizes the principal's payoff, while the lower level verifies incentive compatibility by finding, for each agent, an optimal unilateral deviation assuming all other agents are truthful. We integrate these lower-level deviation problems into the upper-level optimization through a *deviation oracle*: fixing a mechanism, for each agent, the oracle returns an approximate best-response deviation policy. This bilevel formulation exposes two representation choices. The first is *principal-side history compression*: what information from past reports and outcomes the learned mechanism conditions on. The second is the *agent-side sufficient statistic*: how a deviating agent represents private information in the unilateral deviation problem. By enriching the state space, we formulate each agent's unilateral deviation problem as a partially observable Markov decision process (POMDP) [43], which we reduce to a sufficient-statistics Markov decision process (SSMDP), e.g., a belief-state MDP. The sufficient statistic may be the (usually intractable) full history in the general construction, while tractable representations use compact statistics such as Bayesian belief states in finite type spaces or recurrent embeddings in continuous type spaces.

To solve this bilevel optimization, we combine first-order optimization over mechanism parameters with policy-gradient-based methods for the agent-side deviation oracles. We study two outer procedures: a direct penalty method and a binary-search-based acceptance method; for the inner problems, we assume agents solve their unilateral deviation problems using natural policy gradient [40]. Our convergence analysis does not claim global optimality of the learned mechanism. Instead, it quantifies how approximate deviation policies affect stochastic first-order optimization and establishes convergence to approximate stationary solutions of the corresponding penalty or acceptance objectives under the stated assumptions. This is the level of guarantee one should expect in a nonconvex, nonsmooth bilevel problem with learned inner oracles.

We conduct two sets of experiments. In budget-constrained dynamic auctions (the first set), we study the interplay of principal-side history compression and optimization by comparing compact state representations with richer history summaries. Richer representations lead gradient-based learning to improved mechanisms in both settings, including an independent-type setting, where a compact state representation is theoretically sufficient for optimal design [24]. The revenue gains are substantially larger in a correlated-type setting, where this sufficiency guarantee no longer applies. In dynamic bandit auctions (the second set), we study agent-side representations. Specifically, we conduct incentive audits by varying deviation oracles' information, architecture, and optimization method, and find that our learned mechanisms resist exploitability by strong deviation oracles.

In summary, the key contribution of this work is a computational framework for learning approximately incentive-compatible dynamic mechanisms with convergence guarantees, which, to our knowledge, is the most powerful such framework to date.

2 Optimal Dynamic Mechanism Design

We start by defining dynamic mechanism design environments and the games dynamic mechanisms induce. Using the dynamic revelation principle, we then formulate optimal dynamic mechanism design as a bilevel optimization problem.

2.1 Dynamic Mechanism Design Model

Dynamic Mechanism Design Environment. A *dynamic mechanism design environment (DMD)* $\mathcal{D} \doteq (n, T, \mathcal{T}, \mathcal{X}, (\omega_i)_{i\in[n]}, (F_i)_{i\in[n]}, (u_i)_{i\in[n]\cup\{0\}}, \gamma)$ specifies the number of agents n, horizon $T \in \mathbb{N} \cup \{\infty\}$, joint private type space $\mathcal{T} = \bigtimes_{i\in[n]} \mathcal{T}_i$, joint public outcome space $\mathcal{X} = \bigtimes_{i\in[n]} \mathcal{X}_i$,

commonly-known initial type distributions $\omega_i \in \Delta(\mathcal{T}_i)$, independent type transition kernels F_i, agent immediate rewards $(u_i)_{i\in[n]}$, principal immediate reward u_0, and discount factor $\gamma \in (0, 1]$, with $\gamma < 1$ when $T = \infty$. Outcomes can encode allocations, payments, or other payoff-relevant decisions. For each agent i, F_i governs the next type conditional on the current type and realized outcome history, i.e., $t_{i,\tau+1} \sim F_i(\cdot \mid t_{i,\tau}, \boldsymbol{x}^{1:\tau})$.[1] Immediate rewards are also history dependent: agent i's period-τ reward is $u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau})$, with the principal's reward defined analogously.

Dynamic mechanisms and induced games. Given a DMD $\mathcal{D}$, a *dynamic mechanism* is a tuple $M = (\mathcal{R}, g)$ with joint report space $\mathcal{R} = \bigtimes_{i\in[n]} \mathcal{R}_i$. Writing $\boldsymbol{x}^{1:\tau-1}$ and $\hat{\boldsymbol{t}}^{1:\tau}$ for outcome and report histories, the rule $g : \bigcup_{\tau=1}^{T} \left(\mathcal{X}^{1:\tau-1} \times \mathcal{R}^{1:\tau}\right) \to \mathcal{X}$ selects $\boldsymbol{x}_\tau = g(\boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau})$. A mechanism is *direct* if $\mathcal{R} = \mathcal{T}$, and we denote the spaces of dynamic and direct mechanisms by $\mathcal{M}$ and $\mathcal{M}^{\text{direct}}$.

Together, $\mathcal{D}$ and M induce an *agent dynamic game* $(\mathcal{D}, M)$. At time τ, each agent i observes private type $t_{i,\tau}$ and reports $\hat{t}_{i,\tau}$. The mechanism observes report profile $\hat{\boldsymbol{t}}_\tau$ and outputs a public outcome $\boldsymbol{x}_\tau$. Immediate rewards $(u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau}))_{i\in[n]}$ and $u_0(\boldsymbol{t}, \boldsymbol{x}^{1:\tau})$ are then disbursed, and each next type is drawn independently from $F_i(\cdot \mid t_{i,\tau}, \boldsymbol{x}^{1:\tau})$.

Reports affect outcomes, and outcomes affect future types and rewards, so that reporting strategies have dynamic consequences. For all agents $i \in [n]$, let $\boldsymbol{t}_i^{1:\tau}$, $\hat{\boldsymbol{t}}_i^{1:\tau-1}$, and $\boldsymbol{x}^{1:\tau-1}$ denote agent i's private type history, her past report history, and the public outcome history available before reporting at time τ. A reporting strategy for agent i is a collection $\boldsymbol{\pi}_i = \{\pi_{i,\tau} : \mathcal{T}_i^{1:\tau} \times \mathcal{R}_i^{1:\tau-1} \times \mathcal{X}^{1:\tau-1} \to \mathcal{R}_i\}_{\tau=1}^{T}$. We denote agent i's strategy space by Π_i and the joint strategy space by $\Pi = \bigtimes_{i\in[n]} \Pi_i$. When M is direct, the *truthful reporting strategy* reports the current true type at every history, i.e., $\pi_{i,\tau}(\boldsymbol{t}_i^{1:\tau}, \hat{\boldsymbol{t}}_i^{1:\tau-1}, \boldsymbol{x}^{1:\tau-1}) = t_{i,\tau}$. We denote the truthful profile by $\boldsymbol{\pi}^\dagger$.

Agent i's long-run payoff under strategy profile $\boldsymbol{\pi}$ in mechanism M is given by $U_i(\boldsymbol{\pi}; M) \doteq \mathbb{E}^{\boldsymbol{\pi}, M}\left[\sum_{\tau=1}^{T} \gamma^{\tau-1} u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau})\right]$, and the principal's payoff $U_0(\boldsymbol{\pi}; M)$ is defined analogously using u_0. A *payoff-relevant history*, or *result*, is a full trajectory $(\boldsymbol{t}^{1:T}, \boldsymbol{x}^{1:T})$. A mechanism *implements* a result distribution if the distribution is induced by an equilibrium strategy profile of $(\mathcal{D}, M)$.

Equilibrium foundation. Given a DMD $\mathcal{D}$ and a fixed dynamic mechanism M, a reporting strategy profile $\boldsymbol{\pi}$ is a *Bayesian Nash equilibrium (BNE)* of $(\mathcal{D}, M)$ if no agent can improve her payoff by a unilateral reporting deviation: $U_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}; M) \geq U_i(\boldsymbol{\pi}_i', \boldsymbol{\pi}_{-i}; M)$ for all $i \in [n]$ and all $\boldsymbol{\pi}_i' \in \Pi_i$. We denote the set of BNE by $\text{BNE}(\mathcal{D}, M)$.

We rely on two equilibrium-foundation results from Sugaya and Wolitzky [53], stated formally in Section B.1. First, in our setting, the choice of solution concept (BNE, perfect Bayesian equilibrium, or sequential equilibrium) does not affect the set of implementable result distributions. Second, any equilibrium-implementable result distribution can also be induced by a direct mechanism with truthful reporting as the equilibrium. Thus we formulate ODMD over direct mechanisms with truthful reporting as the target BNE, without loss at the level of implementable result distributions.

2.2 Optimal Dynamic Mechanism Design

The goal in ODMD is to find a dynamic mechanism that maximizes the principal's equilibrium payoff. By the aforementioned equilibrium-equivalence and dynamic revelation principle results, we formulate this problem over direct dynamic mechanisms and use truthful reporting as the target BNE. The remaining requirements are that truthful reporting is indeed an equilibrium and that agents receive nonnegative expected payoff from participating.

Fix a DMD $\mathcal{D}$. A direct dynamic mechanism M is *dynamic incentive compatible (DIC)* if $\boldsymbol{\pi}^\dagger \in \text{BNE}(\mathcal{D}, M)$. It is *(ex-ante) dynamically individually rational (DIR)*[2] if truthful reporting gives every

[1] The environment primitives are known to both the principal and the agents; however, realized types remain private. The principal observes reports and public outcomes; each agent observes her own type and report histories and public outcomes. For learning, closed-form transition kernels are unnecessary if online sampling under candidate mechanisms and reporting policies supports the gradient estimators described in Section 4.1. Our pathwise implementation additionally requires a differentiable simulator.

[2] We impose ex-ante DIR because participation occurs at the beginning of the dynamic relationship, before the agent observes the full future private-information path, and the mechanism is analyzed under commitment. Ex-ante DIR is therefore the appropriate participation benchmark.

agent nonnegative expected discounted payoff, i.e., $U_i(\boldsymbol{\pi}^\dagger; M) \geq 0$ for all $i \in [n]$. The ODMD problem is thus defined as the following bilevel optimization problem:

$$\max_{M \in \mathcal{M}^{\text{direct}}} U_0(\boldsymbol{\pi}^\dagger; M) \tag{1a}$$

$$\text{s.t. } \boldsymbol{\pi}^\dagger \in \text{BNE}(\mathcal{D}, M) \tag{1b}$$

$$U_i(\boldsymbol{\pi}^\dagger; M) \geq 0 \ \ \forall i \in [n] \tag{1c}$$

Equation (1a) maximizes the principal's payoff under truthful reporting; Equation (1b) enforces DIC; and Equation (1c) enforces DIR.

Because DIC fixes truthful reporting by construction, Equation 1 does not require computing or selecting among equilibria of the induced agent game. Instead, DIC reduces to n unilateral deviation inequalities: for each agent $i \in [n]$, $\max_{\boldsymbol{\pi}_i \in \Pi_i} U_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}^\dagger_{-i}; M) - U_i(\boldsymbol{\pi}^\dagger; M) \leq 0$, where the left-hand side is agent i's maximal gain by unilaterally deviating from truthful reporting. In other words, DIC holds iff truthful reporting maximizes each agent's payoff when all other agents report truthfully.

3 Parameterized Bilevel Formulation and Representations

Section 2 formulates ODMD as a bilevel optimization problem over direct mechanisms with DIC and DIR constraints. In this section, we work towards making this formulation computationally tractable. On the principal's side, we parameterize the mechanism class and compress the mechanism's history. On the agents' side, each unilateral deviation constraint gives rise to a unilateral deviation problem, which we formulate first as a POMDP, and then, as a sufficient statistic MDP.

3.1 Principal-Side Representation: Mechanism Parameterization and History Compression

To manage the vast mechanism space, we restrict attention to a parameterized family of direct dynamic mechanisms. For $d \in \mathbb{N}$, let $\Theta \subseteq \mathbb{R}^d$ be a parameter space and let $\mathcal{M}^\Theta \doteq \{M^{\boldsymbol{\theta}} = (\mathcal{T}, g^{\boldsymbol{\theta}}) \mid \boldsymbol{\theta} \in \Theta\}$ denote the corresponding family. The resulting optimization is over the chosen class $\mathcal{M}^\Theta$, not over all direct mechanisms.

A second principal-side choice is how the mechanism represents history. A fully general dynamic mechanism would condition on a history of all past reports and outcomes. For tractability, especially in infinite-horizon settings, we represent the mechanism's history by a finite-dimensional summary $c_\tau \doteq \Phi(\boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau}) \in \mathbb{R}^d$, where Φ is a *compression function*. We assume this summary can be updated recursively as $c_{\tau+1} = \Phi'(c_\tau, \boldsymbol{x}_\tau, \hat{\boldsymbol{t}}_{\tau+1})$, using the current summary, the new outcome, and the next report profile without reconstructing the entire history.

Assuming a compressed history, the mechanism rule is now evaluated as $g^{\boldsymbol{\theta}}(c_\tau)$. Compression can be a modeling restriction or a computational convenience. In Pavan et al. [50] and Kakade et al. [39], DMD becomes analytically tractable under structural assumptions that render finite summaries of the histories sufficient for optimality. In contrast, we learn compressed history representations and corresponding mechanisms, and study how different summaries affect performance.

3.2 Agent-Side Representation: POMDP Formulation and Sufficient Statistic MDPs

POMDP Formulation Fix a mechanism parameter $\boldsymbol{\theta} \in \Theta$ and an agent $i \in [n]$, and suppose all agents other than i report truthfully. The adversarial DIC constraint requires computing agent i's best reporting deviation against $\boldsymbol{\pi}^\dagger_{-i}$. We model this unilateral deviation problem as a parameterized *agent partially observable Markov decision process (POMDP)* $\mathcal{P}_i^{\boldsymbol{\theta}}$. The environment is only partially, not fully, observable because agent i does not observe the other agents' private types. This POMDP is parameterized by the mechanism parameter $\boldsymbol{\theta}$ because the mechanism rule $g^{\boldsymbol{\theta}}$, and therefore the reward function and transition kernels depend on $\boldsymbol{\theta}$.

The actions in $\mathcal{P}_i^{\boldsymbol{\theta}}$ are agent i's possible reports, i.e., $\mathcal{A}_i = \mathcal{T}_i$. The latent state s_τ at time τ is chosen to ensure the induced dynamics are Markov. In the general full-history construction, $s_\tau = (\boldsymbol{t}_\tau, \boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau-1})$. The current joint types determine rewards and future type evolution, while outcome and report histories capture payoff, transition, and mechanism history dependence. Given s_τ and action $a_{i,\tau}$, assuming other agents report truthfully, the mechanism generates $\boldsymbol{x}_\tau =$

$g^{\boldsymbol{\theta}}(\boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau-1}, (a_{i,\tau}, t_{-i,\tau}))$, agent i receives $r_i(s_\tau, a_{i,\tau}; \boldsymbol{\theta}) = u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau})$, and the next state appends the current reports and outcome after sampling next types from F. In the observation model, $o_{i,\tau} = (\boldsymbol{x}_{\tau-1}, t_{i,\tau})$, so the induced *history space* is $\mathcal{I}_{i,\tau} = (\mathcal{O}_i \times \mathcal{A}_i)^{1:\tau-1} \times \mathcal{O}_i$ with $(\boldsymbol{t}_i^{1:\tau}, \hat{\boldsymbol{t}}^{1:\tau-1}, \boldsymbol{x}^{1:\tau-1}) \in \mathcal{I}_{i,\tau}$. Thus, policies in $\mathcal{P}_i^{\boldsymbol{\theta}}$ correspond exactly to reporting strategies for agent i in the induced agent dynamic game.

Let $R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})$ denote agent i's expected discounted reward in $\mathcal{P}_i^{\boldsymbol{\theta}}$ under policy $\boldsymbol{\pi}_i$. By construction, $R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) = U_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^\dagger; M^{\boldsymbol{\theta}})$. Therefore, solving $\mathcal{P}_i^{\boldsymbol{\theta}}$ optimally computes the maximal unilateral deviation value defined by the DIC constraint, namely $\varphi_i(\boldsymbol{\theta}) \doteq \max_{\boldsymbol{\pi}_i \in \Pi_i} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) - R_i(\boldsymbol{\pi}_i^\dagger; \boldsymbol{\theta})$. Note that $\varphi_i(\boldsymbol{\theta}) \geq 0$, because truthful reporting is a feasible policy with zero deviation gain.. We call this problem the *unilateral deviation problem*. Full POMDP details are provided in Section D.4.

Sufficient Statistic MDPs An agent's history within its POMDP can be large; indeed, it can grow indefinitely in an infinite-horizon setting. A *sufficient statistic* $\Phi_{i,\tau} : \mathcal{I}_{i,\tau} \to \mathcal{Y}_{i,\tau}$ summarizes this history, while preserving the requisite Markovian transition and reward information needed to solve the unilateral deviation problem. The full history itself is always sufficient, but usually too large for computation. Sufficient-statistic (SS) MDP reductions are useful when they yield smaller statistics than the full history: e.g., exact Bayesian beliefs in finite-state POMDPs, Kalman filters in continuous-state POMDPs, and recurrent neural embeddings as approximately sufficient statistics.

For convergence analyses, finite-horizon SSMDPs can be made stationary by augmenting the statistic with time and an absorbing terminal state. Infinite-horizon SSMDPs require a finite-dimensional payoff-relevant state, stationary dynamics and rewards, and a time-invariant statistic with time-homogeneous induced transitions and rewards; see Sections D.2 and D.3. We denote the resulting unified infinite-horizon stationary SSMDP by $\mathcal{S}_i^{\boldsymbol{\theta}}$. Since the statistic is sufficient, for any $\boldsymbol{\pi}_i \in \Pi_i$ of the POMDP $\mathcal{P}_i^{\boldsymbol{\theta}}$, there exists a corresponding $\hat{\boldsymbol{\pi}}_i$ of the SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}}$ s.t. $\hat{\boldsymbol{\pi}}_i$'s expected discounted reward in $\mathcal{S}_i^{\boldsymbol{\theta}}$, namely $J_i(\hat{\boldsymbol{\pi}}_i; \boldsymbol{\theta})$ equals $R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})$. This agent-side statistic is only used to solve the agent's unilateral deviation problem, and need not match the mechanism's compressed history.

3.3 Parameterized Bilevel ODMD

Combining the mechanism's parameterization with the agent-side unilateral deviation problem yields the following parameterized bilevel ODMD problem:

$$\max_{\boldsymbol{\theta} \in \Theta} \; U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}}) \tag{2a}$$

$$\text{s.t.} \quad \varphi_i(\boldsymbol{\theta}) \leq 0 \;\; \forall i \in [n] \tag{2b}$$

$$U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}}) \geq 0 \;\; \forall i \in [n] \tag{2c}$$

Equation (2a) optimizes the principal's payoff under truthful reporting; Equation (2b) enforces DIC through solutions to the agents' unilateral deviation problems; and Equation (2c) enforces DIR.

4 Computational Framework and Convergence

Section 3 reframes our bilevel formulation of ODMD as an outer optimization over mechanism parameters with the inner adversarial DIC constraints expressed as unilateral deviations. This section presents a three-step computational approach to solving this bilevel optimization problem: 1. We reformulate the outer objective so that it is amenable to projected stochastic subgradient descent (PSSD) [47, 23]; 2. We provide an outer convergence guarantee (Thm 4.1) assuming approximate agent deviation oracles. 3. We describe how to build these oracles (i.e., solve the inner optimization) by running natural policy gradient [40] on the SSMDPs (Thm. 4.2).

4.1 Principal-Side Optimization with Deviation Oracles

An exact solution to the unilateral deviation problem is usually unavailable, as in general, it requires solving an infinite-horizon SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}}$ optimally. But we can still approximate a solution. Since $\mathcal{S}_i^{\boldsymbol{\theta}}$ is stationary, it suffices to consider the class of stationary, Markov policies $\hat{\Pi}_i$. For tractability and differentiability, we approximate this space by a parameterized class of randomized stationary, Markov policies $\hat{\Pi}_i^{\mathcal{W}} = \{\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}} \mid \boldsymbol{w} \in \mathcal{W}\}$. We then assume a *deviation oracle*, which returns a

policy $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}} \in \hat{\Pi}_i^{\mathcal{W}}$ that approximately solves agent i's unilateral deviation problem. We use these deviation oracles to reformulate our parameterized bilevel formulation of ODMD (Eq. 2) in terms of two alternative population outer[3] objectives.

Outer objectives The first is a *penalty reformulation*:

$$G_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \doteq -U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}}) + \beta_1 \sum_{i\in[n]} \varphi_i(\boldsymbol{\theta}) + \beta_2 \sum_{i\in[n]} \max\{0, -U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})\}. \tag{3}$$

Minimizing $G_{\boldsymbol{\beta}}$ ensures high payoff for the principal while penalizing DIC and DIR violations for the agents. We optimize this objective using PSSD and deviation-oracle estimates of the subgradients of the DIC terms. (See Algorithm 1, Appendix C).

The second is an *acceptance objective*, given a candidate principal payoff level δ:

$$H_{\boldsymbol{\beta}}(\boldsymbol{\theta}; \delta) \doteq \max\{0, \delta - U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})\} + \beta_1 \sum_{i\in[n]} \varphi_i(\boldsymbol{\theta}) + \beta_2 \sum_{i\in[n]} \max\{0, -U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})\}. \tag{4}$$

For a fixed δ, minimizing $H_{\boldsymbol{\beta}}(\cdot; \delta)$ tests whether the target payoff δ can be achieved while keeping DIC and DIR violations small: in the exact zero-violation case, any minimizer implements a DIC and DIR mechanism whose principal payoff under truthful reporting reaches the target (see Proposition 3, Appendix E). The acceptance framework wraps this inner optimization in a binary search over δ: for each fixed δ, we run PSSD on $H_{\boldsymbol{\beta}}(\cdot; \delta)$ and declare the target acceptable if the resulting violation is below a prescribed tolerance; binary search over δ then identifies the largest acceptable target up to accuracy α_1. (See Algorithm 2, Appendix C.)

Both of these objectives are nonconvex and nonsmooth in general: the dependence of the principal's payoff on $\boldsymbol{\theta}$ is nonconvex, hinge terms introduce nonsmoothness, and the DIC terms $\varphi_i(\boldsymbol{\theta})$ are computed adversarially: i.e., they are defined by maximization problems over deviation policies. Hence, the outer convergence theorem is a first-order stationarity guarantee for the corresponding objective, not a global optimality or exact feasibility guarantee.

Stochastic subgradient estimators

We optimize the outer objectives in Eqs. 3 and 4 using subgradients. The subdifferential $\partial[z]_+$ of the hinge function $[z]_+ = \max\{z, 0\}$ is given by $\partial[z]_+ = \{\{1\}$ if $z > 0$; $\{0\}$ if $z < 0$; and $[0, 1]$ if $z = 0\}$. In our analysis, we evaluate the hinge selector $s_+(z) \in \partial[z]_+$ at population payoffs.

At iteration k, given oracle policies $\boldsymbol{\pi}_i^{(k)}$ for all agents i, we estimate each ***DIC subgradient*** by $\widehat{g}_{\varphi,i}^{(k)} \doteq \widehat{g}_{R,i}^{(k)} - \widehat{g}_{U,i}^{(k)}$, where $\widehat{g}_{R,i}^{(k)}$ estimates $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)})$, while $\widehat{g}_{U,i}^{(k)}$ is an unbiased estimate of $\nabla_{\boldsymbol{\theta}} U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}})$. For the two objectives, respectively,

$$\widehat{g}_G^{(k)} \doteq -\widehat{g}_{U,0}^{(k)} + \beta_1 \sum_{i\in[n]} \widehat{g}_{\varphi,i}^{(k)} - \beta_2 \sum_{i\in[n]} s_+\Big(-U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}})\Big)\, \widehat{g}_{U,i}^{(k)} \tag{5}$$

$$\widehat{g}_H^{(k)} \doteq -s_+\Big(\delta - U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}})\Big)\, \widehat{g}_{U,0}^{(k)} + \beta_1 \sum_{i\in[n]} \widehat{g}_{\varphi,i}^{(k)} - \beta_2 \sum_{i\in[n]} s_+\Big(-U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}})\Big)\, \widehat{g}_{U,i}^{(k)} \tag{6}$$

We assume the payoff-gradient estimators are conditionally unbiased with bounded second moments (Assumption 2, Appendix E). In particular, $\widehat{g}_{R,i}^{(k)}$ is conditionally unbiased for $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)})$, with the returned policy held fixed. However, because this policy is only an approximate best response, the resulting $\widehat{g}_{\varphi,i}^{(k)}$ may be biased relative to the exact DIC subdifferential; Lemma 1 bounds this *deviation-oracle bias*.[4]

Convergence with approximate deviation oracles

[3]With deviation oracles, these objectives are essentially single-level reformulations of the problem. Nonetheless, we continue to refer to them as "outer" objectives, because in practice we run an inner optimization to compute the DIC terms.

[4]The analysis evaluates hinge selectors at population payoffs. Using estimated payoffs can introduce additional *hinge-selection bias*, because sampling noise can change whether a hinge penalty is positive. This contribution is distinct from the *deviation-oracle bias* bounded by Δ_k in Eq. 15. A bound on hinge-selection bias can be added to Δ_k in the third term of Eq. 7; see Remark 1. We do not quantify this additional contribution here.

Let F denote either $G_{\boldsymbol{\beta}}(\cdot)$ or $H_{\boldsymbol{\beta}}(\cdot;\delta)$, and set $F'(\boldsymbol{\theta}) \doteq F(\boldsymbol{\theta}) + \iota_\Theta(\boldsymbol{\theta})$. For F weakly convex and $\alpha > 0$, we define the *Moreau envelope* $F'_\alpha(\boldsymbol{\theta}) \doteq \min_{\boldsymbol{\theta}' \in \Theta}\{F(\boldsymbol{\theta}') + \frac{1}{2\alpha}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2\}$. Then $\|\nabla F'_\alpha(\boldsymbol{\theta})\|$ is a smooth stationarity measure for α below the weak-convexity threshold. Moreover, when $\|\nabla F'_\alpha(\boldsymbol{\theta})\|$ is small, $\boldsymbol{\theta}$ is close to an approximately first-order stationary point of the nonsmooth constrained objective F.

The only nonstandard component of the subgradients is the DIC term. Suppose the oracle returns $\boldsymbol{\pi}_i^{(k)}$ at iteration k with conditional gap $\mathbb{E}\left[V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)};\boldsymbol{\theta}^{(k)}) \mid \mathcal{F}_k\right] \le \bar{\varepsilon}_i^{(k)}$ for all i, where $V_i(\boldsymbol{\theta}) \doteq \max_{\boldsymbol{\pi}_i \in \Pi_i} R_i(\boldsymbol{\pi}_i;\boldsymbol{\theta})$, $\mathcal{F}_k$ records the information available before the oracle call at outer iteration k, and $\bar{\varepsilon}_i^{(k)}$ bounds the expected oracle suboptimality conditional on this information, with the expectation over the randomness in the oracle call. Then Lemma 1, Appendix E, gives $\mathbb{E}[\widehat{g}_{\varphi,i}^{(k)} \mid \mathcal{F}_k] \in \partial_C \varphi_i(\boldsymbol{\theta}^{(k)}) + \ell_{\nabla R_i,\boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}\, \mathbb{B}$.[5] This is the inexact-oracle contribution: with exact oracles, the update matches the stochastic linear one-sided model of Davis and Drusvyatskiy [23], while approximate oracles add an $O(\sum_i \sqrt{\bar{\varepsilon}_i^{(k)}})\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\|$ model error; see Section D.8.

Assumptions 1–2 in Appendix E give compactness, bounded payoffs, fixed-policy smoothness, well-posed deviation values, oracle stability, and unbiased bounded-variance payoff-gradient estimators; under them, both objectives are weakly convex.

Theorem 4.1. *[Moreau-envelope convergence for $G_{\boldsymbol{\beta}}$ and $H_{\boldsymbol{\beta}}$] Fix $F \in \{G_{\boldsymbol{\beta}}(\cdot), H_{\boldsymbol{\beta}}(\cdot;\delta)\}$ and run projected stochastic subgradient descent with learning rate schedule $\{\eta_{\boldsymbol{\theta}}^{(k)}\}$: i.e., $\boldsymbol{\theta}^{(k+1)} = \Pi_\Theta\Big(\boldsymbol{\theta}^{(k)} - \eta_{\boldsymbol{\theta}}^{(k)}\, \widehat{g}_F^{(k)}\Big)$, where $\widehat{g}_F^{(k)}$ is $\widehat{g}_G^{(k)}$ from Eq. 5 or $\widehat{g}_H^{(k)}$ from Eq. 6, depending on the objective. Suppose the deviation oracle satisfies $\mathbb{E}\left[V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)};\boldsymbol{\theta}^{(k)}) \mid \mathcal{F}_k\right] \le \bar{\varepsilon}_i^{(k)}$ for all agents i, and let Δ_k be the deviation oracle bias term defined by Eq. 15. Fix any $\alpha \in (0, 1/\rho)$, and let R_K be sampled from $\{0,\dots,K-1\}$ with $\mathbb{P}(R_K = k) = \eta_{\boldsymbol{\theta}}^{(k)} / \sum_{t=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(t)}$. Then there exist constants $C_1, C_2, C_3 > 0$ s.t.*

$$\mathbb{E}\left[\|\nabla F'_\alpha(\boldsymbol{\theta}^{(R)})\|^2\right] \le \underbrace{C_1 \frac{F'_\alpha(\boldsymbol{\theta}^{(0)}) - \inf_{\boldsymbol{\theta}} F'(\boldsymbol{\theta})}{\sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{finite-time optimization error}} + \underbrace{C_2 \frac{\varsigma_F^2 \sum_{k=0}^{K-1} {\eta_{\boldsymbol{\theta}}^{(k)}}^2}{\sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{stochastic-gradient error}} + \underbrace{C_3 \frac{\sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)} \Delta_k}{\sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{deviation-oracle bias}}. \tag{7}$$

Consequently, if $\sum_k \eta_{\boldsymbol{\theta}}^{(k)} = \infty$, $\sum_k {\eta_{\boldsymbol{\theta}}^{(k)}}^2 < \infty$, and $\sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)} \Delta_k / \sum_{k=0}^{K-1} \eta_{\boldsymbol{\theta}}^{(k)} \to 0$, then the randomly selected iterate R_K with $\mathbb{P}(R_K = k) \propto \eta_{\boldsymbol{\theta}}^{(k)}$ satisfies $\mathbb{E}\left[\|\nabla F'_\alpha(\boldsymbol{\theta}^{(R_K)})\|^2\right] \to 0$.

The three terms in Eq. 7 have distinct interpretations: finite-time optimization progress, stochastic-subgradient noise, and deviation-oracle error, respectively. As long as the oracle's accuracy improves sufficiently, PSSD converges to Moreau stationarity, while with fixed oracle accuracy, it converges to an oracle-error-dependent neighborhood. The theorem is thus a local stationarity guarantee for the chosen mechanism class and objective.

4.2 Agent-Side Deviation Oracles via Natural Policy Gradient

Natural policy gradient for the deviation oracle Next, we analyze stochastic NPG as one principled implementation of an approximate deviation oracle.[6] Fix a mechanism parameter $\boldsymbol{\theta}$ and an agent i with SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}}$. The oracle targets the infinite-horizon unilateral deviation problem $\max_{\boldsymbol{w} \in \mathcal{W}} J_i(\boldsymbol{w};\boldsymbol{\theta})$, where $J_i(\boldsymbol{w};\boldsymbol{\theta})$ is the expected discounted reward of the policy $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}} \in \hat{\Pi}_i^{\boldsymbol{w}}$.

The Fisher information matrix $F_i(\boldsymbol{w};\boldsymbol{\theta}) \doteq \mathbb{E}_{(y_i,a_i)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta}}}\left[\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i) \nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)^\top\right]$, where $\delta_i^{\boldsymbol{w},\boldsymbol{\theta}}$ denotes the normalized discounted state-action occupancy measure in $\mathcal{S}_i^{\boldsymbol{\theta}}$ induced by $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}$.

[5] Here ∂_C denotes the Clarke subdifferential and $\mathbb{B}$, the closed unit Euclidean ball; see Appendix B. The constants $\ell_{\nabla R_i,\boldsymbol{\pi}}$ and $\varepsilon_{i,\mathrm{opt}}$ are defined in the oracle-stability condition of Assumption 1. The former controls how the mechanism's gradient varies with the deviating policy, while the latter converts payoff suboptimality into distance from an optimal policy.

[6] Other first-order oracle solvers can be used provided they return policies satisfying the approximate optimality condition required by Theorem 4.1 In our experiments, we use differentiable rollouts and first-order gradient ascent for the oracle.

Natural policy gradient (NPG) [40] preconditions the policy gradient with the Fisher information. The ideal NPG direction is $F_i(\boldsymbol{w};\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{w}} J_i(\boldsymbol{w};\boldsymbol{\theta})$. In practice, however, the Fisher information and this gradient are estimated from finite rollouts of length H. The resulting estimators are unbiased for the gradient of the truncated objective and for its associated truncated Fisher information:

$$J_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) \doteq \frac{1-\gamma^H}{1-\gamma} \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}} \left[\hat{r}_i(y_i,a_i)\right],$$
$$F_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) \doteq \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}} \left[\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i) \nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)^\top\right],$$

where $\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}$ denotes the normalized length-H discounted occupancy measure. At inner iteration t, length-H simulated trajectories under the current SSMDP policy $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}^{(t)}}$ produce an unbiased truncated policy-gradient estimator $\widehat{\nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w}^{(t)};\boldsymbol{\theta})}$, while the empirical Fisher estimator targets the truncated Fisher matrix. The matrix $\widetilde{F}_{i,H}^{(t)}$ denotes the regularized or spectrally clipped empirical truncated Fisher matrix estimator, which is what is actually inverted in the update. Thus, the stochastic truncated NPG update is $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \eta_{\boldsymbol{w}} \left(\widetilde{F}_{i,H}^{(t)}\right)^{-1} \widehat{\nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w}^{(t)};\boldsymbol{\theta})}$. For the analysis, the oracle returns a uniformly random inner iterate $\boldsymbol{w}^{(T)}$, $T \sim \mathrm{Unif}\{0,\cdots,K-1\}$.

Deviation-oracle accuracy The outer convergence theorem only requires a conditional expected bound on the oracle optimality gap. The following theorem shows that stochastic NPG supplies such an oracle up to policy-class approximation, compatible-function approximation, optimization, direction-estimation, and truncation errors. Assumption 4 (Appendix E) imposes SSMDP and policy-class regularity, including bounded rewards, smooth scores, compatible function approximation, and nondegenerate Fisher information. Assumption 5 (Appendix E) gives conditional bounds on the stochastic NPG directions generated from rollouts.

Theorem 4.2. *[Deviation-oracle accuracy via stochastic NPG] Fix a mechanism parameter $\boldsymbol{\theta} \in \Theta$ and an agent $i \in [n]$. Let $J_i^*(\boldsymbol{\theta}) \doteq \sup_{\hat{\boldsymbol{\pi}}_i \in \hat{\Pi}_i} J_i(\hat{\boldsymbol{\pi}}_i;\boldsymbol{\theta})$. Let $\boldsymbol{w}^{(0)},\ldots,\boldsymbol{w}^{(K-1)}$ be generated by stochastic truncated NPG on the H-truncated SSMDP, and let $T \sim \mathrm{Unif}\{0,\ldots,K-1\}$ be independent of the rollout randomness. Under Assumptions 4–5, the random-iterate $\boldsymbol{w}^{(T)}$ satisfies*

$$\mathbb{E}\left[J_i^*(\boldsymbol{\theta}) - J_i(\boldsymbol{w}^{(T)};\boldsymbol{\theta})\right] \le \underbrace{\varepsilon_{\mathrm{approx}}(i,\boldsymbol{\theta})}_{\textit{policy-class approximation}} + \underbrace{\frac{1-\gamma^H}{1-\gamma}\sqrt{\varepsilon_{\mathrm{cfa},H}(i,\boldsymbol{\theta})}}_{\textit{compatible approximation}} + \underbrace{\frac{D_{\mathrm{KL},H}}{\eta_{\boldsymbol{w}} K}}_{\textit{optimization}} + \underbrace{\frac{M\eta_{\boldsymbol{w}}}{2} C_{d,H}^2}_{\textit{direction size}} + \underbrace{G C_{\mathrm{dir},H}}_{\textit{direction error}} + \underbrace{2\frac{\gamma^H}{1-\gamma} R_{\max}}_{\textit{truncation}}. \tag{8}$$

Moreover, the same bound holds for the true best-so-far iterate.

Corollary 1 (Appendix E) provides the appropriate hyperparameter settings (e.g., $\eta_{\boldsymbol{w}}$, K, etc.) to ensure that stochastic truncated NPG supplies the conditional expected oracle accuracy required by Theorem 4.1, thereby obtaining the truncation, optimization, stochastic-estimation, and approximation terms given by Theorem 4.2. As a result, the DIC certificate produced by the outer algorithms (Algorithms 1 and 2) is only as strong as the deviation oracle used to optimize the deviation policy.

5 Experiments

Our experiments study how representation choices affect both the mechanisms learned by our gradient-based optimization framework and their incentive properties. We simulate budget-constrained auctions to study how principal-side history representations affect revenue, and bandit auctions to assess how agent-side history representations affect the extent to which a deviation oracle can certify that a mechanism is truly DIC. Our environments, network architectures, and training and evaluation protocols are described in detail in Appendices F and G.[7]

For each learned mechanism, we evaluate the principal's payoff and then we reoptimize the agents' deviation policies while holding the mechanism fixed. We report estimated unilateral payoff gains as *evaluation DIC loss*, with all other agents reporting truthfully. Budget experiments sum positive per-agent gains; the bandit aggregation conventions are specified in Appendix G. The *evaluation*

[7] Our code is available here.

Type Setting	A	A_wide	B	C
e1: independent	2.132 ± 0.008	2.152 ± 0.015 (+0.96%)	2.184 ± 0.014 (+2.43%)	2.156 ± 0.026 (+1.13%)
e2: correlated	3.116 ± 0.024	3.158 ± 0.039 (+1.35%)	3.398 ± 0.060 (+9.06%)	3.426 ± 0.015 (+9.95%)

Table 1: Sample mean $\pm$ standard deviation revenue evaluated on fresh trajectories after training. Parentheses show percentage improvement in mean revenue over A within each setting. All mechanisms achieve evaluation DIC ≤ 0.04 and 0 DIR by construction.

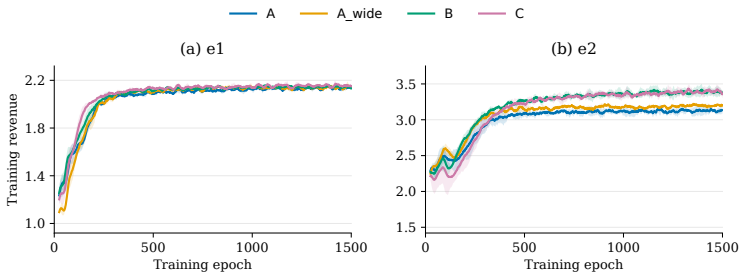


Figure 1: Training revenue in e1 and e2. Curves depict means, with each run smoothed over a trailing 25-epoch moving average. Shaded bands show $\pm$ one sample standard deviation across seeds.

DIR loss is given by $\sum_{i\in[n]} \max\{0, -U_i(\widehat{\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}}})\}$. These evaluation losses are computed on fresh trajectories.

Budget Auctions: Principal-Side History Compression and Optimization We designed budget-auction experiments to study how the principal's history representation can impact revenue. Following our model of a DMD environment, we allow the buyers' private values and public budgets to evolve. The mechanism then chooses allocations and payments subject to budget feasibility.

We consider two settings, each with two buyers who have an initial budget of 3. In setting e1, there is one good, and values are drawn independently from $\mathrm{Unif}[0, 1]$ across buyers each period. In setting e2, there are two goods, and each buyer's next value depends on both her current private value and her allocation. Budget feasibility and ex-post DIR are enforced by construction.

Our theoretical reference point is the lossless history-compression framework of Deng et al. [24]. In their budget auctions, they use *generalized bank account mechanisms* (GBAMs) to summarize each buyer's history through a cumulative utility balance α_i and remaining budget β_i. Under their assumptions, including independence of types across buyers and periods, such a GBAM-enhanced auction can achieve the profit of any DIC, ex-post DIR, budget-feasible dynamic mechanism. We use this result as an information-compression benchmark for e1: it says that a GBAM representation can be without loss, but it does not provide a closed-form characterization of the optimal mechanism or of the optimal profit, nor does it imply that gradient descent will easily find an optimal GBAM.

We compare our gradient-based approach across four representations. A uses a compact GBAM-style state: current reports, time, cumulative utility α_i, and remaining budgets β_i. A_wide uses the same inputs with a larger network. B augments A with cumulative allocation and payment summaries, while C adds a masked, time-indexed attention summary of past reports, allocations, and payments. A_wide, B, and C have equal parameter counts. We train a mechanism using the acceptance method with equal per-trial budgets and three prescribed seeds. Then, we evaluate the mechanism's revenue and the DIC constraints on fresh trajectories. Our complete protocol is described in Appendix F.

Table 1 shows that richer representations led to higher mean revenue in both settings. The gains are modest in e1: B improves over A by 2.43%, while C improves over A by 1.13%. In e2, the gains over A are much larger: 9.06% for B and 9.95% for C, compared with only 1.35% for A_wide, which increases network capacity without adding information. Perhaps unsurprisingly, the richer-history advantage is substantially larger in e2, the setting outside the sufficiency guarantee.

We also ran distillation experiments to examine the extent to which a compact mechanism can recover a richer representation's advantage. Specifically, we trained a student with A's architecture (a compact GBAM) to imitate a richer teacher's output, and then we initialized our acceptance method with this trained student. In e1, this produced an A student with revenue 2.184, which differs by only 0.01 from its B teacher's, namely 2.195. In e2, however, although the corresponding B-to-A and C-to-A procedures improved compact mechanisms, they did so with much larger teacher–student gaps (0.122

		Discrete, two goods M1-trained				Discrete, two goods M2-trained				Continuous, one good M1-trained				Continuous, two goods M1-trained			
Info	Search	Base	Wide	Attn	GRU	Base	Wide	Attn	GRU	MLP	Wide	Attn	GRU	MLP	Wide	Attn	GRU
M1	PW	−0.014	0.022	0.000	0.001	−0.012	−0.003	−0.011	−0.010	0.002	0.008	0.002	0.002	0.010	0.010	0.009	0.010
M1	ES	0.044	−0.086	0.049	0.045	0.036	−0.036	0.036	0.032	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
M1	NPG	0.071	0.023	0.074	0.068	0.050	0.058	0.049	0.047	0.000	0.000	0.000	0.000	0.003	0.001	0.003	0.002
M1	NPG+	0.097	0.014	0.077	0.070	0.072	0.054	0.048	0.049	0.000	0.000	0.000	0.000	0.002	0.001	0.002	0.002
M2	PW	0.032	0.063	0.044	0.046	0.005	0.023	0.006	0.020	0.002	0.008	0.002	0.002	0.010	0.010	0.009	0.010
M2	ES	0.053	−0.142	0.038	0.071	0.017	−0.103	0.028	0.040	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000
M2	NPG	0.041	0.049	**0.102**	**0.101**	0.020	0.060	0.063	0.061	0.000	0.000	0.000	0.000	0.003	0.002	0.002	0.003
M2	NPG+	**0.104**	0.035	**0.108**	**0.104**	0.064	0.060	0.063	0.063	0.000	0.000	0.000	0.000	0.003	0.000	0.002	0.002

Table 2: Evaluation DIC loss of fixed baseline mechanisms, described by the column headers, computed with various deviation oracles (e.g., M1, NPG, Attn).

and 0.164, respectively). Finally, in the e2 setting, we continued training the selected A mechanism for an additional 500 epochs. A's revenue increased, but without closing the gap to B or C. In sum, stronger compact mechanisms can indeed be learned through distillation and further optimization.

Bandit Auctions: Agent-Side Representation and Deviation-Oracle Reliability In a bandit auction, a seller repeatedly allocates perishable goods to buyers whose private valuations evolve as a function of past allocations. Current reports, therefore, affect both immediate rewards and future allocation opportunities. We consider both a discrete- and a continuous-type setting, with one or two goods. In all settings, there are two buyers, and the seller, who aims to maximize her expected long-term profits, incurs a cost of 0.1 per allocated good.

We designed experiments that assess the exploitability of our learned mechanisms: i.e., the extent to which they fail to enforce the agents' unilateral deviation constraints. More specifically, we fix a mechanism and then audit it using a *deviation oracle*, which is characterized along three dimensions:

- *Information.* For discrete types, M1 uses the buyer's current type, a Bayesian belief over opponents' types, and the public allocation history, while M2 adds own-report history. For continuous types, M1 uses encoder-specific features of her current type, her own reports, and the public allocation history; additional recurrent features encode past types. M2 adds own-payment history.
- *Architecture.* We compare a base MLP on flattened history features (Base/MLP), a wider network (Wide), an attention network (Attn), and recurrent features (GRU).
- *Optimization.* We compare pathwise gradients (PW), natural policy gradients (NPG), larger-budget NPG (NPG+), and gradient-free evolution strategies (ES), using multiple restarts.

Our first baseline mechanism uses M1 information, pathwise gradients, and a base flattened MLP. For this (and all) baseline mechanisms, evaluation DIC loss (against the deviation oracle used during training) and estimated ex-ante DIR loss were both below 0.1. We freeze this mechanism to audit it: i.e., to evaluate it against alternative deviation oracle configurations.

Assuming one good, the learned mechanism is 0.1 approximately DIC, as verified by exact dynamic programming. For an M1-trained two-good mechanism, however, DIC loss exceeds 0.1 under M2 with NPG and NPG+. Training with M2 information instead, while retaining the base architecture and pathwise optimization, yielded DIC loss below 0.1 everywhere, with 0.072 as the largest loss. (See Table 2.) The corresponding revenue suffers only slightly, dropping from 15.570 to 15.528.

In the continuous types cases, both for one and two goods, the largest recorded DIC losses discovered during our audit were 0.008 and 0.010, respectively. Changing the information, the architecture, nor the learning method increased these values beyond the baseline. These experiments do not provide a guarantee that the mechanisms learned by our framework are DIC, but they do provide strong evidence that they are approximately DIC, as they are robust to strong deviation oracles.

6 Conclusion

In this paper, we developed a computational framework for learning approximately optimal incentive-compatible dynamic mechanisms with convergence guarantees. We also investigated how principal-side and agent-side representations affect the learned mechanisms and their incentive properties. In future work, we plan to investigate how to design learning mechanisms from offline trajectories, and scale the framework to larger, more realistic dynamic mechanism design settings.

Acknowledgments

This research was supported by the Office of Naval Research under grant N00014-24-1-2657 and the National Science Foundation under grant IIS2147187. The authors also wish to thank anonymous reviewers for their helpful feedback.

References

[1] Susan Athey and Ilya Segal. An Efficient Dynamic Mechanism. *Econometrica*, 81(6):2463–2485, 2013. ISSN 1468-0262. doi: 10.3982/ECTA6995. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA6995. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA6995.

[2] Moshe Babaioff, Yogeshwer Sharma, and Aleksandrs Slivkins. Characterizing truthful multi-armed bandit mechanisms. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 79–88, 2009.

[3] Marco Battaglini. Long-Term Contracting with Markovian Consumers. *American Economic Review*, 95(3):637–658, June 2005. ISSN 0002-8282. doi: 10.1257/0002828054201369. URL https://www.aeaweb.org/articles?id=10.1257/0002828054201369.

[4] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[5] Dirk Bergemann and Maher Said. Dynamic Auctions: A Survey, March 2010. URL https://papers.ssrn.com/abstract=1576613.

[6] Dirk Bergemann and Juuso Välimäki. The Dynamic Pivot Mechanism. *Econometrica*, 78(2):771–789, 2010. ISSN 0012-9682. URL https://www.jstor.org/stable/40664492.

[7] Dirk Bergemann and Juuso Välimäki. Dynamic Mechanism Design: An Introduction. *Journal of Economic Literature*, 57(2):235–274, 2019. ISSN 0022-0515. URL https://www.jstor.org/stable/26673239.

[8] Dimitri P. Bertsekas and Steven E. Shreve. *Stochastic Optimal Control: The Discrete-Time Case | Guide books | ACM Digital Library*. URL https://dl.acm.org/doi/book/10.5555/1512940.

[9] David Besanko. Multi-period contracts between principal and agent with adverse selection. *Economics Letters*, 17(1):33–37, January 1985. ISSN 0165-1765. doi: 10.1016/0165-1765(85)90122-3. URL https://www.sciencedirect.com/science/article/pii/0165176585901223.

[10] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. *arXiv preprint arXiv:2105.15183*, 2021.

[11] Simon Board. Durable-Goods Monopoly with Varying Demand. *The Review of Economic Studies*, 75(2):391–413, 2008. ISSN 0034-6527. URL https://www.jstor.org/stable/20185037.

[12] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

[13] Gianluca Brero, Alon Eden, Matthias Gerstgrasser, David C. Parkes, and Duncan Rheingans-Yoo. Reinforcement Learning of Sequential Price Mechanisms, May 2021. URL http://arxiv.org/abs/2010.01180. arXiv:2010.01180 [cs].

[14] Gianluca Brero, Alon Eden, Darshan Chakrabarti, Matthias Gerstgrasser, Amy Greenwald, Vincent Li, and David C. Parkes. Stackelberg POMDP: A Reinforcement Learning Approach for Economic Design, November 2023. URL http://arxiv.org/abs/2210.03852. arXiv:2210.03852 [cs].

[15] Evan M. Calford and Timothy N. Cason. Contingent Reasoning and Dynamic Public Goods Provision. *American Economic Journal: Microeconomics*, 16(2):236–266, May 2024. ISSN 1945-7669. doi: 10.1257/mic.20220111. URL https://www.aeaweb.org/articles?id=10.1257/mic.20220111.

[16] Bhaskar Chakravorti. Dynamic public goods provision with coalitional manipulation. *Journal of Public Economics*, 56(1):143–161, January 1995. ISSN 0047-2727. doi: 10.1016/0047-2727(93)01410-C. URL https://www.sciencedirect.com/science/article/pii/004727279301410C.

[17] Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, UAI'02, pages 103–110, San Francisco, CA, USA, August 2002. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-897-9.

[18] Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM conference on Electronic commerce*, EC '04, pages 132–141, New York, NY, USA, May 2004. Association for Computing Machinery. ISBN 978-1-58113-771-2. doi: 10.1145/988772.988793. URL https://doi.org/10.1145/988772.988793.

[19] Pascal Courty and Li Hao. Sequential Screening. *The Review of Economic Studies*, 67(4):697–717, October 2000. ISSN 0034-6527. doi: 10.1111/1467-937X.00150. URL https://doi.org/10.1111/1467-937X.00150.

[20] Michael Curry, Tuomas Sandholm, and John Dickerson. Differentiable economics for randomized affine maximizer auctions. *arXiv preprint arXiv:2202.02872*, 2022.

[21] Michael Curry, Vinzenz Thoma, Darshan Chakrabarti, Stephen McAleer, Christian Kroer, Tuomas Sandholm, Niao He, and Sven Seuken. Automated Design of Affine Maximizer Mechanisms in Dynamic Settings, February 2024. URL http://arxiv.org/abs/2402.08129. arXiv:2402.08129 [cs].

[22] Michael J. Curry, Zhou Fan, Yanchen Jiang, Sai Srivatsa Ravindranath, Tonghan Wang, and David C. Parkes. Automated mechanism design: A survey. *ACM SIGecom Exchanges*, 22(2):102–120, 2025.

[23] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[24] Yuan Deng, Vahab Mirrokni, and Song Zuo. Non-clairvoyant dynamic mechanism design with budget constraints and beyond. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 369, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 978-1-4503-8554-1. doi: 10.1145/3465456.3467545. URL https://dl.acm.org/doi/10.1145/3465456.3467545.

[25] Zhijian Duan, Haoran Sun, Yurong Chen, and Xiaotie Deng. A scalable neural network for dsic affine maximizer auction design. *Advances in Neural Information Processing Systems*, 36:56169–56185, 2023.

[26] Paul Duetting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism Design for Large Language Models, October 2023. URL http://arxiv.org/abs/2310.10826. arXiv:2310.10826 [cs, econ].

[27] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C. Parkes, and Sai Srivatsa Ravindranath. Optimal Auctions through Deep Learning: Advances in Differentiable Economics. *Journal of the ACM*, page 3630749, November 2023. ISSN 0004-5411, 1557-735X. doi: 10.1145/3630749. URL https://dl.acm.org/doi/10.1145/3630749.

[28] Péter Eso and Balázs Szentes. Optimal Information Disclosure in Auctions and the Handicap Auction. *The Review of Economic Studies*, 74(3):705–731, 2007. ISSN 0034-6527. URL https://www.jstor.org/stable/4626158.

[29] Alireza Fallah, Michael Jordan, and Annie Ulichney. Fair allocation in dynamic mechanism design. *Advances in Neural Information Processing Systems*, 37:125935–125966, 2024.

[30] Zhe Feng, Harikrishna Narasimhan, and David C Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 354–362, 2018.

[31] Daniel Fershtman and Alessandro Pavan. Pandora's Auctions: Dynamic Matching with Unknown Preferences. *American Economic Review*, 107(5):186–190, May 2017. ISSN 0002-8282. doi: 10.1257/aer.p20171043. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20171043.

[32] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning, May 2016. URL http://arxiv.org/abs/1605.06676. arXiv:1605.06676 [cs].

[33] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 5–22, 2014.

[34] Noah Golowich, Harikrishna Narasimhan, and David C Parkes. Deep learning for multi-facility location mechanism design. In *Proceedings of the Twenty-seventh International Joint Conference on Artificial Intelligence and the Twenty-third European Conference on Artificial Intelligence*, pages 261–267, 2018.

[35] Amy Greenwald, Jiacui Li, and Eric Sodomka. Approximating Equilibria in Sequential Auctions with Incomplete Information and Multi-Unit Demand. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/hash/801c14f07f9724229175b8ef8b4585a8-Abstract.html.

[36] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernandez del Rio, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[37] Matthew Hausknecht and Peter Stone. Deep Recurrent Q-Learning for Partially Observable MDPs, January 2017. URL http://arxiv.org/abs/1507.06527. arXiv:1507.06527 [cs].

[38] Tom Hennigan, Trevor Cai, Tamara Norman, Lena Martens, and Igor Babuschkin. Haiku: Sonnet for JAX, 2020. URL http://github.com/deepmind/dm-haiku.

[39] Sham Kakade, Ilan Lobel, and Hamid Nazerzadeh. Optimal Dynamic Mechanism Design and the Virtual Pivot Mechanism, March 2013. URL https://papers.ssrn.com/abstract=1782211.

[40] Sham M Kakade. A Natural Policy Gradient. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://papers.nips.cc/paper_files/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

[41] Kevin Kuo, Anthony Ostuni, Elizabeth Horishny, Michael J Curry, Samuel Dooley, Ping-yeh Chiang, Tom Goldstein, and John P Dickerson. Proportionnet: Balancing fairness and revenue for auction design with deep learning. *arXiv preprint arXiv:2010.06398*, 2020.

[42] Haoming Li, Yumou Liu, Zhenzhe Zheng, Zhilin Zhang, Jian Xu, and Fan Wu. Truthful bandit mechanisms for repeated two-stage ad auctions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1565–1575, 2024.

[43] Michael L. Littman. A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125, June 2009. ISSN 0022-2496. doi: 10.1016/j.jmp.2009.01.005. URL https://www.sciencedirect.com/science/article/pii/S0022249609000042.

[44] Yanli Liu, Kaiqing Zhang, Tamer Ba{\c{s}}ar, and Wotao Yin. An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods, November 2022. URL http://arxiv.org/abs/2211.07937. arXiv:2211.07937 [cs].

[45] Vahab Mirrokni, Renato Paes Leme, Pingzhong Tang, and Song Zuo. Optimal Dynamic Auctions are Virtual Welfare Maximizers, December 2018. URL http://arxiv.org/abs/1812.02993. arXiv:1812.02993 [cs, econ].

[46] Roger B. Myerson. Optimal Auction Design. *Mathematics of Operations Research*, 6(1):58–73, February 1981. ISSN 0364-765X. doi: 10.1287/moor.6.1.58. URL https://pubsonline.informs.org/doi/10.1287/moor.6.1.58.

[47] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

[48] Mallesh M. Pai and Rakesh Vohra. Optimal Dynamic Auctions and Simple Index Rules. *Mathematics of Operations Research*, 38(4):682–697, 2013. ISSN 0364-765X. URL https://www.jstor.org/stable/24540877.

[49] Alessandro Pavan. Dynamic Mechanism Design: Robustness and Endogenous Types. In Bo Honoré, Ariel Pakes, Monika Piazzesi, and Larry Samuelson, editors, *Advances in Economics and Econometrics*, pages 1–62. Cambridge University Press, 1 edition, November 2017. ISBN 978-1-316-51052-0 978-1-108-22716-2 978-1-108-40000-8. doi: 10.1017/9781108227162.001. URL https://www.cambridge.org/core/product/identifier/CBO9781108227162A010/type/book_part.

[50] Alessandro Pavan, Ilya Segal, and Juuso Toikka. Dynamic Mechanism Design: A Myersonian Approach. *Econometrica*, 82(2):601–653, 2014. ISSN 0012-9682. URL https://www.jstor.org/stable/24029270.

[51] Sai Srivatsa Ravindranath, Yanchen Jiang, and David C Parkes. Data market design through deep learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 6662–6689. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1577ea3eaf8dacb99f64e4496c3ecddf-Paper-Conference.pdf.

[52] Maher Said. Sequential auctions with randomly arriving buyers. *Games and Economic Behavior*, 73(1):236–243, September 2011. ISSN 0899-8256. doi: 10.1016/j.geb.2010.12.010. URL https://www.sciencedirect.com/science/article/pii/S0899825611000054.

[53] Takuo Sugaya and Alexander Wolitzky. The Revelation Principle in Multistage Games. *The Review of Economic Studies*, 88(3 (320)):1503–1540, 2021. ISSN 0034-6527. URL https://www.jstor.org/stable/27031806.

[54] Andrea Tacchetti, DJ Strouse, Marta Garnelo, Thore Graepel, and Yoram Bachrach. A neural architecture for designing truthful and efficient auctions. *arXiv preprint arXiv:1907.05181*, 2019.

[55] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

[56] Arun Venkatraman, Nicholas Rhinehart, Wen Sun, Lerrel Pinto, Martial Hebert, Byron Boots, Kris M. Kitani, and J. Andrew Bagnell. Predictive-State Decoders: Encoding the Future into Recurrent Networks, September 2017. URL http://arxiv.org/abs/1709.08520. arXiv:1709.08520 [stat].

[57] Gustavo Vulcano, Garrett van Ryzin, and Costis Maglaras. Optimal Dynamic Auctions for Revenue Management. *Management Science*, 48(11):1388–1407, 2002. ISSN 0025-1909. URL https://www.jstor.org/stable/822614.

[58] Tonghan Wang, Yanchen Jiang, and David C. Parkes. Gemnet: Menu-based, strategy-proof multi-bidder auctions through deep learning. In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 1100, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400707049. doi: 10.1145/3670865.3673454. URL https://doi.org/10.1145/3670865.3673454.

[59] Tonghan Wang, Yanchen Jiang, and David C Parkes. Bundleflow: Deep menus for combinatorial auctions by diffusion-based optimization. *arXiv preprint arXiv:2502.15283*, 2025.

[60] Noah Williams. Persistent private information. *Econometrica*, 79(4):1233–1275, 2011.

[61] 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, February 2022. URL http://arxiv.org/abs/2202.10678. arXiv:2202.10678 [cs, econ].

[62] Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Computing Optimal Equilibria and Mechanisms via Learning in Zero-Sum Extensive-Form Games, May 2024. URL http://arxiv.org/abs/2306.05216. arXiv:2306.05216 [cs].

[63] Hanrui Zhang and Vincent Conitzer. Automated Dynamic Mechanism Design, May 2021. URL http://arxiv.org/abs/2105.06008. arXiv:2105.06008 [cs].

A Additional Related Work

Foundations and efficiency in DMD. For broad surveys of dynamic mechanism design (DMD), see Bergemann and Välimäki [7] and Pavan [49]. A large efficiency-oriented thread establishes dynamic analogues of Groves/VCG and studies institutional constraints. The dynamic pivot mechanism implements efficient allocations under evolving private information [6]. Related work formulates allocation and pricing with stochastic arrivals/departures and timing via MDP/online models and sequential auctions [e.g., 35, 52, 57]. Beyond dominant-strategy templates, Athey and Segal [1] characterize efficient dynamic mechanisms in environments with intertemporal private information, and Pavan et al. [50] provide a general revelation and envelope approach for dynamic screening. Other work studies dynamic revenue maximization in models with time-varying/evolving types [48, 45]. Complementary strands analyze dynamic pricing/screening and institutional constraints–including durable-goods monopoly with varying demand over time [11], sequential screening [19], and a revelation principle for multistage games that clarifies the implementability of extensive-form mechanisms [53]. Continuous-time dynamic contracting addresses persistent private information and commitment [60]. Fairness and design constraints have also been embedded into dynamic settings [29].

Information design in dynamic environments. A complementary literature studies what the designer/principal should *reveal* over time. Canonical treatments of (static) information design include Bergemann and Morris [4]. Dynamic counterparts analyze sequential persuasion and optimal disclosure when actions and beliefs evolve, e.g., Wu et al. [61], Eso and Szentes [28]. These tools interact with DMD through belief dynamics, experimentation, and commitment to disclosure policies.

Automated Mechanism Design (AMD) Automated mechanism design originated with Conitzer and Sandholm [17, 18], who cast mechanism search as a constrained optimization problem and studied computational complexity and constraint generation. More recent approaches utilize deep learning to scale to larger and more general settings. Duetting et al. [26] introduced deep neural networks for auction design, improving representational flexibility. This approach, termed *differentiable economics*, has since been extended to budget-constrained bidders [30], payment minimization [54], multi-facility location [34], fairness-revenue trade-offs [41], affine maximizer auctions [20, 25],

and data markets [51]. Structural advancements, such as Wang et al. [58], achieve exact strategy-proofness rather than approximate incentive compatibility. The framework has also been extended to the single-bidder combinatorial auction setting [59]. For a comprehensive survey, see [22].

Dynamic automated mechanism design (DAMD). Recent work pushes the AMD viewpoint into dynamic settings. Zhang and Conitzer [63] introduce Automated Dynamic Mechanism Design with linear-program formulations (and hardness for long horizons). RL-based approaches learn policies for sequential pricing/auctions: Brero et al. [13] learn sequential posted-price mechanisms; Brero et al. [14] study Stackelberg-POMDP formulations and sample-based computation of Stackelberg equilibria relevant to dynamic market design. Parallel efforts connect equilibrium computation with learning in auction games, e.g., Zhang et al. [62]. There is also growing interest in automated design for structured mechanism classes such as affine maximizers in dynamic settings [21]. These works collectively demonstrate that learning-based search over mechanism spaces (with explicit incentive constraints or ex-post compliance) can scale beyond hand-crafted analytic templates.

Truthful bandits and incentivized exploration. A complementary line studies *truthful bandit* and incentivized-exploration mechanisms in repeated/dynamic markets, clarifying when exploration must be separated from exploitation and how to maintain (approximate) BIC while learning [2, 33, 42]. These results inform the limits and design of learn-as-you-go dynamic platforms.

Positioning relative to this paper. Relative to these literatures, our approach searches directly over parameterized *dynamic* mechanisms and enforces IC/IR as equilibrium constraints in the induced dynamic game. This provides a complementary route to dynamic AMD that is not tied to a particular analytic template (e.g., dynamic VCG) and is amenable to multidimensional types and unknown dynamics, while also accommodating belief updates and exploration when relevant.

B Preliminaries

Notation We denote by $[n]$ the set of integers $\{1, \dots, n\}$ and $[n^*]$ by $\{0\} \cup [n]$. Let $\mathcal{X}$ be any set, and let $(\mathcal{X}, \mathscr{B}_\mathcal{X})$ denote the associated measurable space, where $\mathscr{B}_\mathcal{X}$ is the Borel σ-algebra on $\mathcal{X}$. Unless otherwise noted, we will take $\mathscr{B}_\mathcal{X}$ as the default σ-algebra on $\mathcal{X}$. Additionally, we denote the orthogonal projection operator onto a set $\mathcal{X}$ by $\Pi_\mathcal{X}(\boldsymbol{x}) \doteq \arg\min_{\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$. For any $\mathcal{A} \subseteq \mathcal{X}$, we define the indicator function $\mathbb{1}_\mathcal{A} : \mathcal{X} \to \{0, 1\}$ by $\mathbb{1}_\mathcal{A}(x) = 1$ if $x \in \mathcal{A}$ and $\mathbb{1}_\mathcal{A}(x) = 0$ otherwise. For a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\partial_C f(x)$ its Clarke subdifferential at x. When f is continuously differentiable, $\partial_C f(x) = \{\nabla f(x)\}$; when f is convex, $\partial_C f(x)$ coincides with the usual convex subdifferential. We write $\mathbb{B}$ for the closed unit Euclidean ball. For a closed convex set $\Theta \subseteq \mathbb{R}^d$, we denote its penalty indicator function by

$$\iota_\Theta(\theta) = \begin{cases} 0, & \theta \in \Theta, \\ +\infty, & \theta \notin \Theta. \end{cases}$$

Terminology. Fix any norm $\|\cdot\|$. Given $\mathcal{A} \subset \mathbb{R}^d$, the function $f : \mathcal{A} \to \mathbb{R}$ is said to be ℓ_f-*Lipschitz-continuous* for some $\ell_f \in \mathbb{R}_+$ iff $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}, \|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\| \leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. If the gradient of f is $\ell_{\nabla f}$-Lipschitz-continuous for some $\ell_{\nabla f} \in \mathbb{R}_+$, f is called $\ell_{\nabla f}$-*Lipschitz-smooth*. The function f is said to be a ℓ_f-contraction (resp. non-expansion) if it is Lipschitz-continuous with coefficient $\ell_f < 1$ (resp. $\ell_f = 1$).

B.1 Equilibrium Foundations for ODMD

This appendix records the implementability notation and equilibrium-foundation results used in Section 2. Fix a DMD $\mathcal{D}$. We define the space of *payoff-relevant histories*, or *results*, as

$$\mathcal{H} \doteq \{(\boldsymbol{t}^{1:T}, \boldsymbol{x}^{1:T}) \in \mathcal{T}^{1:T} \times \mathcal{X}^{1:T} \mid F_i(t_{i,\tau+1} \mid t_{i,\tau}, \boldsymbol{x}^{1:\tau}) > 0 \; \forall i \in [n], \; \forall \tau \leq T-1\}.$$

Any dynamic mechanism M together with a reporting strategy profile $\boldsymbol{\pi}$ induces a distribution over results, denoted $\nu^{M,\boldsymbol{\pi}} \in \Delta(\mathcal{H})$. A *result distribution* $\nu \in \Delta(\mathcal{H})$ is *implementable* by a dynamic mechanism M under a specified equilibrium concept if there exists an equilibrium reporting strategy profile $\boldsymbol{\pi}$ of the agent dynamic game $(\mathcal{D}, M)$, with respect to that equilibrium concept, such that $\nu = \nu^{M,\boldsymbol{\pi}}$.

Proposition 1. *[Equilibrium Equivalence (Proposition 4, Claim 4 [53])] Consider a dynamic mechanism design problem $\mathcal{D}$. The set of result distributions implementable by dynamic mechanisms is identical under Bayesian Nash equilibrium, perfect Bayesian equilibrium, and sequential equilibrium.*

Proposition 2. *[Dynamic Revelation Principle under BNE (Proposition 2, [53])] Consider a dynamic mechanism design environment $\mathcal{D}$. If a result distribution $\nu \in \Delta(\mathcal{H})$ is implementable by some possibly indirect dynamic mechanism under a BNE, then the same result distribution is implementable by a direct dynamic mechanism in which truthful reporting by all agents constitutes a BNE.*

Taken together, Propositions 1 and 2 imply that, for the dynamic mechanism design problems considered in the main text, implementability can be characterized in terms of direct mechanisms under truthful Bayesian Nash equilibrium.

C Omitted Algorithms

Algorithm 1: Direct Penalty Framework

In: parameter space Θ, number of iterations K, penalty weights $\beta_1, \beta_2 > 0$, learning rates $\{\eta^{(k)}\}_{k=0}^{K-1}$
Out: $\boldsymbol{\theta}^{(K)}$
Initialize $\boldsymbol{\theta}^{(0)} \in \Theta$
for $k = 0, 1, \ldots, K-1$ **do**
  IC oracle:
  for *each agent* $i \in [n]$ **do**
    Approximately solve $\boldsymbol{\pi}_i^{(k)} \approx \arg\max_{\pi_i \in \Pi_i} R_i(\pi_i; \boldsymbol{\theta}^{(k)})$
    $\widehat{\varphi}_i^{(k)} \leftarrow R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{\dagger}; \boldsymbol{\theta}^{(k)})$
  end
  Form penalized objective function

$$\widehat{G}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)}) = -U_0(\boldsymbol{\pi}^{\dagger}; M^{\boldsymbol{\theta}^{(k)}}) + \beta_1 \sum_i \widehat{\varphi}_i^{(k)} + \beta_2 \sum_i \max\left\{0, -U_i(\boldsymbol{\pi}^{\dagger}; M^{\boldsymbol{\theta}^{(k)}})\right\}$$

  Compute subgradient estimate $\widehat{g}_G^{(k)} \in \partial_{\boldsymbol{\theta}} \widehat{G}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)})$
  Update $\boldsymbol{\theta}^{(k+1)} \leftarrow \Pi_{\Theta}\left(\boldsymbol{\theta}^{(k)} - \eta^{(k)} \widehat{g}_G^{(k)}\right)$
end
return $\boldsymbol{\theta}^{(K)}$

Algorithm 2: Binary Search-Based Feasibility Framework

In: Θ, bounds $\delta_{\min} < \delta_{\max}$, accuracy α_1, tolerance α_2, penalty weights $\beta_1, \beta_2 > 0$, learning rates $\{\eta^{(k)}\}_{k=0}^{K-1}$

Out: Feasible mechanism $\widehat{\boldsymbol{\theta}}$ and payoff level $\widehat{\delta}$

Initialize $\widehat{\delta} \leftarrow \delta_{\min}$

while $\delta_{\max} - \delta_{\min} > \alpha_1$ **do**

  $\delta \leftarrow (\delta_{\min} + \delta_{\max})/2$, Initialize $\boldsymbol{\theta}^{(0)} \in \Theta$

  Initialize $\widehat{H}_{\min} \leftarrow +\infty$ and $\boldsymbol{\theta}_{\text{best}} \leftarrow \boldsymbol{\theta}^{(0)}$

  for $k = 0, 1, \ldots, K-1$ **do**

    IC oracle:

    for *each agent* $i \in [n]$ **do**

      Approximately solve $\boldsymbol{\pi}_i^{(k)} \approx \arg\max_{\pi_i \in \Pi_i} R_i(\pi_i; \boldsymbol{\theta}^{(k)})$

      $\widehat{\varphi}_i^{(k)} \leftarrow R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{\dagger}; \boldsymbol{\theta}^{(k)})$

    end

    Form feasibility objective

$$\widehat{H}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)}; \delta) = \max\{0, \delta - U_0(\boldsymbol{\pi}^{\dagger}; M^{\boldsymbol{\theta}^{(k)}})\} + \beta_1 \sum_{i \in [n]} \widehat{\varphi}_i^{(k)} \quad + \beta_2 \sum_{i \in [n]} \max\left\{0, -U_i(\boldsymbol{\pi}^{\dagger}; M^{\boldsymbol{\theta}^{(k)}})\right\}.$$

    if $\widehat{H}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)}; \delta) < \widehat{H}_{\min}$ **then**

      $\widehat{H}_{\min} \leftarrow \widehat{H}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)}; \delta)$, $\boldsymbol{\theta}_{\text{best}} \leftarrow \boldsymbol{\theta}^{(k)}$

    end

    Obtain subgradient estimate $\widehat{g}_H^{(k)} \in \partial_{\boldsymbol{\theta}} \widehat{H}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^{(k)}; \delta)$, Update $\boldsymbol{\theta}^{(k+1)} \leftarrow \Pi_{\Theta}\left(\boldsymbol{\theta}^{(k)} - \eta^{(k)} \widehat{g}_H^{(k)}\right)$

  end

  if $\widehat{H}_{\min} \le \alpha_2$ **then**

    `// δ feasible`

    $\delta_{\min} \leftarrow \delta$, $\widehat{\delta} \leftarrow \delta$, $\widehat{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}_{\text{best}}$

  else

    $\delta_{\max} \leftarrow \delta$

  end

end

return $(\widehat{\boldsymbol{\theta}}, \widehat{\delta})$

D Omitted Discussion and Technical Details

D.1 Comparison Between Two Computational Frameworks

The acceptance-based framework fundamentally differs from penalty methods in how it relaxes the original problem. Rather than trading off the objective and constraint violations via a fixed penalty weight, it separates value targeting from constraint satisfaction: for each candidate δ, the algorithm asks whether δ is attainable under (approximate) IC and IR constraints. This separation offers several advantages. *First*, it directly targets the mechanism design objective of maximizing the principal's payoff, rather than optimizing a surrogate objective that encapsulates tradeoffs, and thus avoids the need to tune penalty parameters. Although the acceptance objective also includes penalty terms, any sufficiently large penalty yields the same feasibility decision. In contrast, in penalty methods, the penalty weight directly determines the objective-constraint tradeoff and must be carefully tuned. *Second*, the acceptance framework is more robust to oracle approximation error. For a fixed target payoff level, oracle error affects only the feasibility test and does not alter the relative weighting between objective maximization and constraint enforcement. In contrast, in penalty-based optimization, oracle errors are scaled by the penalty weight, which can distort optimization dynamics and make performance highly sensitive to tuning.

While it does have some advantages, we emphasize that the acceptance framework does not eliminate nonconvexity or oracle approximation error. Rather, it provides a complementary outer-loop structure that is often more robust to these challenges. In practice, penalty-based methods can be used as inner solvers within the feasibility procedure or as warm-start heuristics.

D.2 Time-Augmentation for Finite-Horizon SSMDP

Consider the finite-horizon SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}} = \Big(T, (\mathcal{Y}_{i,\tau})_{\tau\in[T]}, \mathcal{A}_i, (\hat{P}_{i,\tau})_{\tau\in[T]}, \gamma, \hat{\mu}, (\hat{r}_{i,\tau})_{\tau\in[T]}\Big)$. Essentially, one can alternatively consider an infinite-horizon MDP where the state space is defined as $\mathcal{Y}_\tau = \bigcup_{\tau\in[T]}(\mathcal{Y}_{i,\tau}\times\{\tau\})\cup\{y_{\text{abs}}\}$ where y_{abs} is a costless absorbing state and always transitions to itself. Thus, the stationary transition kernel $\hat{P}$ can be easily defined as $\hat{P}_i((y_i', \tau+1) \mid (y_i,\tau), a_{i,\tau}) = \hat{P}_{i,\tau}(y_i' \mid y_i, a_{i,\tau})$ for all $\tau < T$, and $\hat{P}_i(y_{i,\text{abs}} \mid (y_i,T), a) = \hat{P}_i(y_{i,\text{abs}} \mid y_{i,\text{abs}}, a) = 1$; and the stationary reward function is similarly adopted: $\hat{r}_i((y,\tau), a_{i,\tau}) = \hat{r}_{i,\tau}(y, a_{i,\tau})$ while $\hat{r}_i((y_{i,\text{abs}},\tau), a_{i,\tau}) = 0$. Moreover, any nonstationary policy $\hat{\boldsymbol{\pi}}_i = (\hat{\pi}_{i,1}, \cdots, \hat{\pi}_{i,T})$ in the original nonstationary SSMDP now corresponds to a stationary policy where $\hat{\boldsymbol{\pi}}_i(y_i,\tau) = \hat{\pi}_{i,\tau}(y_i)$.

D.3 Stationarity for Infinite-Horizon SSMDPs

The infinite-horizon case is different from the finite-horizon construction above: stationarity cannot be obtained merely by relabeling time-dependent objects. For a computationally tractable stationary infinite-horizon deviation problem, the growing history must be summarized by a fixed payoff-relevant state, or by an information state built from such a summary, with time-homogeneous controlled dynamics and rewards. Fix an infinite-horizon agent POMDP $\mathcal{P}_i^{\boldsymbol{\theta}}$ and a statistic $\Phi_i : \mathcal{I}_i \to \mathcal{Y}_i$. The statistic yields a stationary SSMDP when the state space $\mathcal{Y}_i$ and action space $\mathcal{A}_i$ are fixed over time and there exist a stationary kernel $\hat{P}_i(\cdot \mid y_i, a_i)$ and a stationary reward function $\hat{r}_i(y_i, a_i)$ such that, for every private history $\iota_{i,\tau}$ with $\Phi_i(\iota_{i,\tau}) = y_i$ and every report a_i,

$$\mathbb{P}\big(\Phi_i(\iota_{i,\tau+1}) \in \underline{\mathcal{Y}_i} \mid \iota_{i,\tau}, a_i\big) = \hat{P}_i(\underline{\mathcal{Y}_i} \mid y_i, a_i)$$

for all measurable $\underline{\mathcal{Y}_i} \subseteq \mathcal{Y}_i$, and

$$\mathbb{E}\big[r_i(S_\tau, a_i; \boldsymbol{\theta}) \mid \iota_{i,\tau}, a_i\big] = \hat{r}_i(y_i, a_i).$$

Both right-hand sides must be independent of τ. Under these conditions, the reduced control problem is the stationary infinite-horizon SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}} = (\infty, \mathcal{Y}_i, \mathcal{A}_i, \hat{P}_i, \gamma, \hat{\mu}, \hat{r}_i)$.

A common way to verify these conditions is to first identify a finite-dimensional payoff-relevant state compression with stationary dynamics and rewards. That is, the primitive rewards, observation law, and type transition law of the agent POMDP depend on the growing history only through a fixed compressed state, and this compressed state has a time-homogeneous transition law given the

agent's report. The agent-side statistic can then be a belief over this compressed state, or another information state that is sufficient for predicting the next statistic and current expected reward. This is the infinite-horizon modeling condition that keeps the deviation oracle finite-dimensional and stationary. The principal-side compressed state is related but conceptually separate: it determines what the learned mechanism can condition on, while the payoff-relevant state for the deviation problem must also carry the information needed to predict utilities, observations, and type transitions. In applications the two summaries may coincide, but they need not be identical.

D.4 Agent POMDP Details

This appendix spells out the POMDP construction used in Section 3. Fix a DMD $\mathcal{D}$, a parameterized direct mechanism $M^{\boldsymbol{\theta}} = (\mathcal{T}, g^{\boldsymbol{\theta}})$, and an agent $i \in [n]$. Other agents report truthfully, so agent i's unilateral deviation problem can be represented as $\mathcal{P}_i^{\boldsymbol{\theta}} \doteq (T, \mathcal{S}, \mathcal{A}_i, P_i, \gamma, \mu, r_i, \mathcal{O}_i, O_i)$.

The horizon T and discount factor γ are inherited from $\mathcal{D}$. The action space is $\mathcal{A}_i = \mathcal{T}_i$. In the full-history construction, the latent state space is

$$\mathcal{S} = \bigcup_{\tau=1}^{T} \mathcal{T} \times \mathcal{X}^{1:\tau-1} \times \mathcal{T}^{1:\tau-1},$$

with $s_\tau = (\boldsymbol{t}_\tau, \boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau-1})$. The initial distribution is induced by the priors over initial types, with empty outcome and report histories.

Given state s_τ and action $a_{i,\tau} \in \mathcal{A}_i$, the other agents' truthful reports are $t_{-i,\tau}$, the current joint report is $\hat{\boldsymbol{t}}_\tau = (a_{i,\tau}, t_{-i,\tau})$, and the mechanism chooses

$$\boldsymbol{x}_\tau = g^{\boldsymbol{\theta}}(\boldsymbol{x}^{1:\tau-1}, \hat{\boldsymbol{t}}^{1:\tau-1}, \hat{\boldsymbol{t}}_\tau).$$

The transition kernel updates the outcome and report histories deterministically and draws next-period types according to the independent per-agent kernels:

$$\prod_{i \in [n]} F_i(t_{i,\tau+1} \mid t_{i,\tau}, \boldsymbol{x}^{1:\tau}).$$

The reward is the deviating agent's period payoff under the induced outcome,

$$r_i(s_\tau, a_{i,\tau}; \boldsymbol{\theta}) = u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau}).$$

The observation space is $\mathcal{O}_i = \mathcal{X} \times \mathcal{T}_i$. The baseline observation at time τ is $o_{i,\tau} = (\boldsymbol{x}_{\tau-1}, t_{i,\tau})$, and the observation kernel O_i is the deterministic projection from s_τ to these observed components. Hence the private history space is

$$\mathcal{I}_{i,\tau} = (\mathcal{O}_i \times \mathcal{A}_i)^{1:\tau-1} \times \mathcal{O}_i = \mathcal{T}_i^{1:\tau} \times \mathcal{T}_i^{1:\tau-1} \times \mathcal{X}^{1:\tau-1}.$$

Policies in this POMDP are therefore equivalent to reporting strategies for agent i in the agent dynamic game $(\mathcal{D}, M^{\boldsymbol{\theta}})$ when all other agents report truthfully. For any such policy,

$$R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) \doteq \mathbb{E}^{\boldsymbol{\pi}_i}\left[\sum_{\tau=1}^{T} \gamma^{\tau-1} r_i(S_\tau, A_{i,\tau}; \boldsymbol{\theta})\right] = U_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^{\dagger}; M^{\boldsymbol{\theta}}).$$

D.5 Examples of Sufficient Statistics

We briefly discuss several standard choices of sufficient statistics for reducing agent POMDPs to SSMDPs here.

Bayesian belief states. A canonical sufficient statistic is the agent's Bayesian belief over the underlying payoff-relevant state. Let $b_{i,\tau} \in \Delta(S)$ denote agent i's posterior belief conditional on her information history $\iota_{i,\tau}$. The belief evolves deterministically via Bayes' rule as a function of the previous belief, the agent's action, and the newly observed outcome, inducing a fully observable belief-state MDP with belief-averaged rewards. While belief-state reductions are exact in finite-state environments, they remain well defined but computationally challenging in continuous state or type spaces, where the belief lies in an infinite-dimensional space. In such settings, practical implementations rely on

approximations such as particle filtering or parametric function approximation, trading off precision for tractability.

Recurrent neural embeddings. To avoid explicit belief tracking, we also consider learned recurrent embeddings as implicit sufficient statistics. The statistic at time τ consists of a public embedding c_τ summarizing publicly observed signals and an agent-specific private embedding $h_{i,\tau}$ summarizing private observations. This design follows common practice in POMDPs and multi-agent RL [37, 56, 32], where recurrent architectures are employed to capture and update information over time.

Other classical examples. Beyond belief states, classical POMDP reductions employ a variety of sufficient statistics that summarize observation histories. Predictive state representations (PSRs) replace beliefs with vectors of predictions of future observable events, yielding Markovian dynamics in prediction space. Finite-memory controllers and information-state abstractions restrict attention to a finite set of internal states updated deterministically from observations and actions. In continuous-state settings, parametric filters such as Kalman filters and extended Kalman filters provide finite-dimensional sufficient statistics under linear-Gaussian or locally linear assumptions. All of these constructions induce SSMDPs by ensuring that the statistic admits Markovian transitions and reward sufficiency.

D.6 Deterministic vs. Randomized Policies

From an economic perspective, we are ultimately interested in optimizing over deterministic reporting strategies, or equivalently, deterministic stationary policies on the SSMDP. For an SSMDP with Borel state and action spaces and bounded rewards, it is well known that randomization does not improve the optimal value: for any $\varepsilon > 0$, there exists a deterministic stationary policy whose expected discounted reward is within ε of the optimal value [8].

Nevertheless, our analysis is conducted over a class of smoothly parameterized randomized stationary policies. This relaxation is purely technical and is required to apply policy gradient and natural policy gradient methods, which rely on differentiability of the policy and non-degeneracy of the Fisher information matrix. Importantly, allowing randomized policies does not alter the optimal value benchmark: the supremum of the expected return over randomized policies coincides with that over deterministic policies. Consequently, convergence guarantees stated for randomized policies imply convergence in value to an ε-optimal deterministic policy for arbitrarily small ε.

D.7 Sufficient Conditions for Assumption 4 and Assumption 5

Note that most conditions in Assumption 4 and Assumption 5 can be readily satisfied by imposing standard regularity conditions on the DMD $\mathcal{D}$. For example, assume that for any agent $i \in [n]$, the type space $\mathcal{T}_i$ is a non-empty, compact Borel space, and each agent has bounded immediate reward, i.e., there exists $U_{\max} < \infty$ such that, for any agent $i \in [n]$, $\tau \in [T]$, $t_{i,\tau} \in \mathcal{T}_i$, and $\boldsymbol{x}^{1:\tau} \in \mathcal{X}^{1:\tau}$, $|u_i(t_{i,\tau}, \boldsymbol{x}^{1:\tau})| \leq U_{\max}$.

The key condition that guarantees the convergence of natural policy gradient is the non-degeneracy of the Fisher information matrix, which is a regularity condition on the interaction between the policy parameterization and the state-action visitation distribution induced by the policy. Formally, the Fisher matrix $F_i(\boldsymbol{w})$ is positive definite if and only if no nonzero parameter direction is almost orthogonal to the policy score under the induced occupancy measure. This requires that (i) the policy parameterization be identifiable so that different parameter directions correspond to genuinely different action distributions, and (ii) the induced state-action distribution place sufficient mass on regions where these directions are "excited."

This condition can be ensured by standard choices of policy classes. When $\mathcal{A}_i$ is discrete, a softmax policy with bounded logits and explicit exploration (e.g., an ε-mixture with the uniform distribution) guarantees that every action has uniformly positive probability, and that the covariance of the score vectors spans the full parameter space $\mathcal{W}$ whenever the feature representation is non-degenerate. For continuous action space $\mathcal{A}_i$, a common choice is a (possibly truncated) Gaussian policy with mean parameterized by $\boldsymbol{w}$ and covariance uniformly bounded below by $\sigma_{\min^2} I$. In this case, the score with respect to the mean parameters has nonzero variance in every direction, and the Fisher matrix reduces to the expected Jacobian covariance of the mean map under the discounted state visitation measure.

Uniform positive definiteness then follows provided the Jacobian of the mean map has full column rank on a set of states visited with positive probability.

In both cases, Fisher non-degeneracy is achieved by design through a combination of (i) a non-redundant policy parameterization and (ii) enforced exploration, and is standard in the analysis of natural policy gradient methods. Similar sufficient conditions are discussed in, e.g., Appendix B.2 of [44].

D.8 Connection to Davis–Drusvyatskiy Moreau-Envelope Descent

Remark 1. *Let $\mathcal{F}_k$ denote the filtration generated by the algorithm before constructing $\widehat{g}_F^{(k)}$. By Lemma 1 and Assumption 2, for either objective $F \in \{G_{\boldsymbol{\beta}}(\cdot), H_{\boldsymbol{\beta}}(\cdot;\delta)\}$, there exist $v_F^{(k)} \in \partial_C F(\boldsymbol{\theta}^{(k)})$ and $b_F^{(k)} \in \mathbb{R}^d$ such that*

$$\mathbb{E}\Big[\widehat{g}_F^{(k)} \mid \mathcal{F}_k\Big] = v_F^{(k)} + b_F^{(k)}, \qquad \|b_F^{(k)}\| \leq \Delta_k, \tag{9}$$

where

$$\Delta_k \doteq \beta_1 \sum_{i\in[n]} \ell_{\nabla R_i,\boldsymbol{\pi}}\, \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}. \tag{10}$$

The term Δ_k captures the bias induced by using approximate deviation oracles to estimate the incentive-compatibility terms φ_i; the smooth payoff-gradient estimators are unbiased. If empirical payoff estimates are used to determine hinge activity, the same relation holds with $\Delta_k^{\mathrm{emp}} \doteq \Delta_k + b_{\mathrm{hinge}}^{(k)}$ in place of Δ_k, where $b_{\mathrm{hinge}}^{(k)}$ bounds the additional compositional bias from empirical hinge selectors. Moreover, under Assumptions 1 and 2, after increasing ς_F^2 if necessary, the estimator has uniformly bounded conditional second moment:

$$\mathbb{E}\Big[\|\widehat{g}_F^{(k)}\|^2 \mid \mathcal{F}_k\Big] \leq \varsigma_F^2. \tag{11}$$

For either objective F, define the random linear model

$$m_k(\boldsymbol{\theta}) = F(\boldsymbol{\theta}^{(k)}) + \langle \widehat{g}_F^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)} \rangle. \tag{12}$$

Since F is ρ-weakly convex, (9) implies that for all $\boldsymbol{\theta} \in \Theta$,

$$\mathbb{E}[m_k(\boldsymbol{\theta}) \mid \mathcal{F}_k] - F(\boldsymbol{\theta}) \leq \frac{\rho}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\|^2 + \Delta_k \|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\|. \tag{13}$$

Thus, when the deviation oracle is exact, $\Delta_k = 0$ and (13) reduces to the stochastic one-sided linear model condition of Davis and Drusvyatskiy [23]. For $\Delta_k > 0$, (13) is an inexact one-sided model with an additional linear error term. With empirical hinge selectors, the same statement holds with Δ_k^{emp} in place of Δ_k. The projected update

$$\boldsymbol{\theta}^{(k+1)} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\{ \langle \widehat{g}_F^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)} \rangle + \frac{1}{2\eta_{\boldsymbol{\theta}}^{(k)}} \|\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\|^2 \right\} \tag{14}$$

is therefore a prox-regularized minimization step on this inexact linear model.

E Omitted Assumptions, Results, and Proofs

Proposition 3. *Given a DMD $\mathcal{D}$ and a class of parameterized direct dynamic mechanisms $\mathcal{M}^{\Theta}$. For any $\delta^* \in \mathbb{R}$, if there exists $\beta_1, \beta_2 > 0$ s.t. $\min_{\boldsymbol{\theta}\in\Theta} H_{\boldsymbol{\beta}}(\boldsymbol{\theta}; \delta^*) = 0$, with $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Theta} H_{\boldsymbol{\beta}}(\boldsymbol{\theta}; \delta^*)$, then $M^{\boldsymbol{\theta}^*}$ is IC and IR, and the expected principal payoff under truthful reporting satisfies $U_0(\boldsymbol{\pi}^\dagger; \boldsymbol{\theta}^*) \geq \delta^*$.*

Proof. Consider any $\delta^* \in \mathbb{R}$ and $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\Theta} H_{\boldsymbol{\beta}}(\boldsymbol{\theta}; \delta^*)$, and assume that $H_{\boldsymbol{\beta}}(\boldsymbol{\theta}^*; \delta^*) = 0$. Each term in $H_{\boldsymbol{\beta}}$ is nonnegative: the principal target violation and IR violations are hinge losses, and

$$\varphi_i(\boldsymbol{\theta}) = \max_{\boldsymbol{\pi}_i \in \Pi_i} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) - R_i(\pi_i^\dagger; \boldsymbol{\theta}) \geq 0$$

because truthful reporting is one feasible reporting policy for agent i. Since $\beta_1, \beta_2 > 0$ and the sum of nonnegative terms is zero, each term must be zero at $\boldsymbol{\theta}^*$. Thus $\varphi_i(\boldsymbol{\theta}^*) = 0$ for all $i \in [n]$, so $M^{\boldsymbol{\theta}^*}$ is IC; $U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^*}) \geq 0$ for all $i \in [n]$, so it is IR; and $(\delta^* - U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^*}))_+ = 0$, so the expected principal payoff under truthful reporting satisfies $U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^*}) \geq \delta^*$. □

Assumption 1 (Outer-loop DMD regularity)**.** *Given a DMD $\mathcal{D}$ and a class of parameterized direct dynamic mechanisms $\mathcal{M}^{\Theta}$, the following hold.*

1. ***(Parameter domain and bounded payoffs)** The parameter space Θ is nonempty, closed, convex, and compact, and all principal and agent discounted payoffs are uniformly bounded.*
2. ***(Uniform fixed-policy smoothness)** For every agent $i \in [n]$ and every reporting policy $\boldsymbol{\pi}_i \in \Pi_i$, the map $\boldsymbol{\theta} \mapsto R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})$ is continuously differentiable on an open neighborhood of Θ. Moreover, there exist constants $L_{R,i}, L_{\nabla R,i} < \infty$ such that, for all $\boldsymbol{\pi}_i \in \Pi_i$ and all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,*

$$\|\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})\| \leq L_{R,i},$$
$$\|\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}')\| \leq L_{\nabla R,i} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

   The truthful payoff maps $\boldsymbol{\theta} \mapsto U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})$ and $\boldsymbol{\theta} \mapsto U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})$ have uniformly bounded and Lipschitz gradients on Θ. Under these conditions, the functions $G_{\boldsymbol{\beta}}$ and $H_{\boldsymbol{\beta}}(\cdot; \delta)$ are ρ-weakly convex on Θ for some finite ρ.
3. ***(Well-posed deviation values)** For every $\boldsymbol{\theta} \in \Theta$, $V_i(\boldsymbol{\theta}) \doteq \max_{\boldsymbol{\pi}_i \in \Pi_i} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})$ is finite and the maximizer set $\Pi_i^*(\boldsymbol{\theta})$ is nonempty.*
4. ***(Clarke-Danskin regularity)** For every $i \in [n]$ and $\boldsymbol{\theta} \in \Theta$, the deviation value satisfies the Clarke-Danskin formula*

$$\partial_C V_i(\boldsymbol{\theta}) = \operatorname{conv}\left\{\nabla_{\boldsymbol{\theta}} R_i(\pi_{i,*}; \boldsymbol{\theta}) : \pi_{i,*} \in \Pi_i^*(\boldsymbol{\theta})\right\}.$$

   Consequently, for $\varphi_i(\boldsymbol{\theta}) \doteq V_i(\boldsymbol{\theta}) - R_i(\pi_i^\dagger; \boldsymbol{\theta})$, we have

$$\partial_C \varphi_i(\boldsymbol{\theta}) = \partial_C V_i(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} R_i(\pi_i^\dagger; \boldsymbol{\theta}).$$

5. ***(Oracle stability)** There exists $\varepsilon_{i,\mathrm{opt}} > 0$ such that, for any $\boldsymbol{\theta} \in \Theta$ and any $\boldsymbol{\pi}_i \in \Pi_i$, if $V_i(\boldsymbol{\theta}) - R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) \leq \varepsilon$, then*

$$\operatorname{dist}\left(\boldsymbol{\pi}_i, \Pi_i^*(\boldsymbol{\theta})\right) \leq \varepsilon_{i,\mathrm{opt}} \sqrt{\varepsilon}.$$

   Moreover, $\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta})$ is Lipschitz in $\boldsymbol{\pi}_i$ with constant $\ell_{\nabla R_i, \boldsymbol{\pi}}$, uniformly over $\boldsymbol{\theta} \in \Theta$:

$$\|\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i; \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i'; \boldsymbol{\theta})\| \leq \ell_{\nabla R_i, \boldsymbol{\pi}} \operatorname{dist}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_i').$$

The oracle-stability condition is a quadratic-growth-type condition for the inner deviation problem; it is used only to convert value suboptimality of the returned deviation policy into bias of the outer IC subgradient estimator.

Assumption 2 (Unbiased payoff-gradient estimators)**.** *At outer iteration k, let $\widehat{g}_{U,0}^{(k)}$ estimate the truthful principal payoff gradient, let $\widehat{g}_{U,i}^{(k)}$ estimate the truthful payoff gradient for agent i, and let $\widehat{g}_{R,i}^{(k)}$ estimate the fixed-deviation payoff gradient used to form $\widehat{g}_{\varphi,i}^{(k)} \doteq \widehat{g}_{R,i}^{(k)} - \widehat{g}_{U,i}^{(k)}$. Conditional on $\mathcal{F}_k$ and the oracle output $\boldsymbol{\pi}_i^{(k)}$, these estimators satisfy*

$$\begin{aligned}
\mathbb{E}[\widehat{g}_{U,0}^{(k)} \mid \mathcal{F}_k] &= \nabla_{\boldsymbol{\theta}} U_0(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}}),\\
\mathbb{E}[\widehat{g}_{U,i}^{(k)} \mid \mathcal{F}_k] &= \nabla_{\boldsymbol{\theta}} U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}^{(k)}}),\\
\mathbb{E}[g_{R,i}^{(k)} \mid \mathcal{F}_k, \boldsymbol{\pi}_i^{(k)}] &= \nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}), \qquad \forall i \in [n].
\end{aligned}$$

Moreover, their conditional second moments are uniformly bounded: there exists $\varsigma_{\mathrm{pay}} < \infty$ such that, for all $i \in [n]$,

$$\mathbb{E}[\|\widehat{g}_{U,0}^{(k)}\|^2 |\, \mathcal{F}_k], \quad \mathbb{E}[\|\widehat{g}_{U,i}^{(k)}\|^2 |\, \mathcal{F}_k], \quad \mathbb{E}[\|\widehat{g}_{R,i}^{(k)}\|^2 |\, \mathcal{F}_k, \boldsymbol{\pi}_i^{(k)}] \le \varsigma_{\mathrm{pay}}^2.$$

Lemma 1 (Approximate IC subgradient estimator)**.** *Suppose Assumptions 1–2 hold. Fix an outer iteration k and an agent $i \in [n]$. Let $\mathcal{F}_k$ be the filtration generated before the deviation oracle is called. Suppose the oracle returns a possibly randomized policy $\boldsymbol{\pi}_i^{(k)}$ satisfying $\mathbb{E}[V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}) \mid \mathcal{F}_k] \le \bar{\varepsilon}_i^{(k)}$. Define $\widehat{g}_{\varphi,i}^{(k)} \doteq \widehat{g}_{R,i}^{(k)} - \widehat{g}_{U,i}^{(k)}$. Then*

$$\operatorname{dist}\left(\mathbb{E}[\widehat{g}_{\varphi,i}^{(k)} \mid \mathcal{F}_k], \partial_C \varphi_i(\boldsymbol{\theta}^{(k)})\right) \le \ell_{\nabla R_i, \boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}.$$

Equivalently, $\mathbb{E}[\widehat{g}_{\varphi,i}^{(k)} \mid \mathcal{F}_k] \in \partial_C \varphi_i(\boldsymbol{\theta}^{(k)}) + \ell_{\nabla R_i, \boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}\, \mathbb{B}$.

Proof. For each realized oracle output $\boldsymbol{\pi}_i^{(k)}$, choose $\boldsymbol{\pi}_i^*(\boldsymbol{\pi}_i^{(k)}) \in \Pi_i^*(\boldsymbol{\theta}^{(k)})$ satisfying

$$\operatorname{dist}\left(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i^*(\boldsymbol{\pi}_i^{(k)})\right) \le \varepsilon_{i,\mathrm{opt}} \sqrt{V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)})},$$

which is possible by Assumption 1. Define

$$v_i^{(k)} \doteq \nabla_{\boldsymbol{\theta}} R_i\left(\boldsymbol{\pi}_i^*(\boldsymbol{\pi}_i^{(k)}); \boldsymbol{\theta}^{(k)}\right) - \nabla_{\boldsymbol{\theta}} R_i(\pi_i^\dagger; \boldsymbol{\theta}^{(k)}).$$

By Clarke-Danskin regularity, $v_i^{(k)} \in \partial_C \varphi_i(\boldsymbol{\theta}^{(k)})$ for every realization. Since $\partial_C \varphi_i(\boldsymbol{\theta}^{(k)})$ is convex and closed, $\mathbb{E}[v_i^{(k)} \mid \mathcal{F}_k] \in \partial_C \varphi_i(\boldsymbol{\theta}^{(k)})$. Using Assumption 2, Lipschitzness in the policy argument, and Jensen's inequality,

$$\begin{aligned}
&\operatorname{dist}\left(\mathbb{E}[\widehat{g}_{\varphi,i}^{(k)} \mid \mathcal{F}_k], \partial_C \varphi_i(\boldsymbol{\theta}^{(k)})\right)\\
&\le \left\|\mathbb{E}[\widehat{g}_{\varphi,i}^{(k)} \mid \mathcal{F}_k] - \mathbb{E}[v_i^{(k)} \mid \mathcal{F}_k]\right\|\\
&= \left\|\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}) - \nabla_{\boldsymbol{\theta}} R_i\left(\boldsymbol{\pi}_i^*(\boldsymbol{\pi}_i^{(k)}); \boldsymbol{\theta}^{(k)}\right) \mid \mathcal{F}_k\right]\right\|\\
&\le \ell_{\nabla R_i, \boldsymbol{\pi}}\, \mathbb{E}\left[\operatorname{dist}\left(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i^*(\boldsymbol{\pi}_i^{(k)})\right) \mid \mathcal{F}_k\right]\\
&\le \ell_{\nabla R_i, \boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}}\, \mathbb{E}\left[\sqrt{V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)})} \mid \mathcal{F}_k\right]\\
&\le \ell_{\nabla R_i, \boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}} \sqrt{\mathbb{E}\left[V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)}; \boldsymbol{\theta}^{(k)}) \mid \mathcal{F}_k\right]}\\
&\le \ell_{\nabla R_i, \boldsymbol{\pi}} \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}.
\end{aligned}$$

This proves the distance bound. The inclusion follows from the definition of distance to a closed set. □

Theorem 4.1. *[Moreau-envelope convergence for $G_{\boldsymbol{\beta}}$ and $H_{\boldsymbol{\beta}}$] Fix $F \in \{G_{\boldsymbol{\beta}}(\cdot), H_{\boldsymbol{\beta}}(\cdot;\delta)\}$ and run projected stochastic subgradient descent with learning rate schedule $\{\eta_{\boldsymbol{\theta}}^{(k)}\}$: i.e., $\boldsymbol{\theta}^{(k+1)} = \Pi_\Theta\Big(\boldsymbol{\theta}^{(k)} - \eta_{\boldsymbol{\theta}}^{(k)}\,\widehat{g}_F^{(k)}\Big)$, where $\widehat{g}_F^{(k)}$ is $\widehat{g}_G^{(k)}$ from Eq. 5 or $\widehat{g}_H^{(k)}$ from Eq. 6, depending on the objective. Suppose the deviation oracle satisfies $\mathbb{E}\left[V_i(\boldsymbol{\theta}^{(k)}) - R_i(\boldsymbol{\pi}_i^{(k)};\boldsymbol{\theta}^{(k)}) \mid \mathcal{F}_k\right] \le \bar{\varepsilon}_i^{(k)}$ for all agents i, and let Δ_k be the deviation oracle bias term defined by Eq. 15. Fix any $\alpha \in (0, 1/\rho)$, and let R_K be sampled from $\{0,\dots,K-1\}$ with $\mathbb{P}(R_K = k) = \eta_{\boldsymbol{\theta}}^{(k)}/\sum_{t=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(t)}$. Then there exist constants $C_1, C_2, C_3 > 0$ s.t.*

$$\mathbb{E}\left[\|\nabla F_\alpha'(\boldsymbol{\theta}^{(R)})\|^2\right] \le \underbrace{C_1 \frac{F_\alpha'(\boldsymbol{\theta}^{(0)}) - \inf_{\boldsymbol{\theta}} F'(\boldsymbol{\theta})}{\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{finite-time optimization error}} + \underbrace{C_2 \frac{\varsigma_F^2 \sum_{k=0}^{K-1}{\eta_{\boldsymbol{\theta}}^{(k)}}^2}{\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{stochastic-gradient error}} + \underbrace{C_3 \frac{\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}\Delta_k}{\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}}}_{\textit{deviation-oracle bias}}. \tag{7}$$

Consequently, if $\sum_k \eta_{\boldsymbol{\theta}}^{(k)} = \infty$, $\sum_k {\eta_{\boldsymbol{\theta}}^{(k)}}^2 < \infty$, and $\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}\Delta_k / \sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)} \to 0$, then the randomly selected iterate R_K with $\mathbb{P}(R_K = k) \propto \eta_{\boldsymbol{\theta}}^{(k)}$ satisfies $\mathbb{E}\left[\|\nabla F_\alpha'(\boldsymbol{\theta}^{(R_K)})\|^2\right] \to 0$.

Proof. We first verify weak convexity. By Assumption 1, for each agent i and policy $\boldsymbol{\pi}_i \in \Pi_i$, the map $\boldsymbol{\theta} \mapsto R_i(\boldsymbol{\pi}_i;\boldsymbol{\theta})$ has Lipschitz gradient on Θ, uniformly over $\boldsymbol{\pi}_i$. Hence each such map is weakly convex on Θ. Since $V_i(\boldsymbol{\theta}) = \max_{\boldsymbol{\pi}_i \in \Pi_i} R_i(\boldsymbol{\pi}_i;\boldsymbol{\theta})$, and pointwise maxima preserve weak convexity, the function V_i is weakly convex. Moreover, since $R_i(\boldsymbol{\pi}_i^\dagger;\cdot)$ has Lipschitz gradient, $-R_i(\boldsymbol{\pi}_i^\dagger;\cdot)$ is weakly convex. Therefore $\varphi_i(\boldsymbol{\theta}) = V_i(\boldsymbol{\theta}) - R_i(\boldsymbol{\pi}_i^\dagger;\boldsymbol{\theta})$ is weakly convex after increasing the weak-convexity parameter.

The truthful payoff maps have Lipschitz gradients by Assumption 1. Hence $-U_0(\boldsymbol{\pi}^\dagger; M^\cdot)$ is weakly convex, and the hinge terms $\max\{0, -U_i(\boldsymbol{\pi}^\dagger; M^\cdot)\}$ and $\max\{0, \delta - U_0(\boldsymbol{\pi}^\dagger; M^\cdot)\}$ are weakly convex because each is the pointwise maximum of 0 and a smooth weakly convex function. Nonnegative weighted sums preserve weak convexity, so F is ρ-weakly convex for some finite ρ. Since Θ is closed and convex, ι_Θ is convex, so $F' = F + \iota_\Theta$ is also ρ-weakly convex as an extended-valued function.

Define

$$\Delta_k \doteq \beta_1 \sum_{i\in[n]} \ell_{\nabla R_i, \boldsymbol{\pi}}\, \varepsilon_{i,\mathrm{opt}} \sqrt{\bar{\varepsilon}_i^{(k)}}. \tag{15}$$

Next, Lemma 1 and Assumption 2 imply the bias relation

$$\mathbb{E}\Big[\widehat{g}_F^{(k)} \mid \mathcal{F}_k\Big] = v_F^{(k)} + b_F^{(k)}, \qquad v_F^{(k)} \in \partial_C F(\boldsymbol{\theta}^{(k)}), \qquad \|b_F^{(k)}\| \le \Delta_k. \tag{16}$$

The term Δ_k captures the bias induced by using approximate deviation oracles to estimate the incentive-compatibility terms φ_i; the smooth payoff-gradient estimators are unbiased. If empirical payoff estimates are used to determine hinge activity, the same proof applies with $\Delta_k^{\mathrm{emp}} \doteq \Delta_k + b_{\mathrm{hinge}}^{(k)}$ in place of Δ_k, where $b_{\mathrm{hinge}}^{(k)}$ bounds the additional compositional bias from empirical hinge selectors. Assumptions 1 and 2 also imply, after increasing ς_F^2 if necessary, that

$$\mathbb{E}\Big[\|\widehat{g}_F^{(k)}\|^2 \mid \mathcal{F}_k\Big] \le \varsigma_F^2. \tag{17}$$

Let $\bar{\boldsymbol{\theta}}^{(k)} = \mathrm{prox}_{\alpha F'}(\boldsymbol{\theta}^{(k)})$. Then

$$\nabla F_\alpha'(\boldsymbol{\theta}^{(k)}) = \alpha^{-1}(\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}).$$

Because $\bar{\boldsymbol{\theta}}^{(k)} \in \Theta$ and the update is a projection onto Θ,

$$\begin{aligned}\|\boldsymbol{\theta}^{((k+1))} - \bar{\boldsymbol{\theta}}^{(k)}\|^2 &\le \|\boldsymbol{\theta}^{(k)} - \eta_{\boldsymbol{\theta}}^{(k)}\widehat{g}_F^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\|^2 \\ &= \|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\|^2 - 2\eta_{\boldsymbol{\theta}}^{(k)}\langle \widehat{g}_F^{(k)}, \boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\rangle + {\eta_{\boldsymbol{\theta}}^{(k)}}^2 \|\widehat{g}_F^{(k)}\|^2. \end{aligned} \tag{18}$$

By the definition of the Moreau envelope,

$$F'_\alpha(\boldsymbol{\theta}^{((k+1))}) \le F'(\bar{\boldsymbol{\theta}}^{(k)}) + \frac{1}{2\alpha}\|\boldsymbol{\theta}^{((k+1))} - \bar{\boldsymbol{\theta}}^{(k)}\|^2$$
$$\le F'_\alpha(\boldsymbol{\theta}^{(k)}) - \frac{\eta_{\boldsymbol{\theta}}^{(k)}}{\alpha}\langle \widehat{g}_F^{(k)}, \boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\rangle + \frac{{\eta_{\boldsymbol{\theta}}^{(k)}}^2}{2\alpha}\|\widehat{g}_F^{(k)}\|^2. \tag{19}$$

Taking conditional expectation and using (16)–(17),

$$\mathbb{E}\Big[F'_\alpha(\boldsymbol{\theta}^{((k+1))}) \mid \mathcal{F}_k\Big] \le F'_\alpha(\boldsymbol{\theta}^{(k)}) - \frac{\eta_{\boldsymbol{\theta}}^{(k)}}{\alpha}\langle v_F^{(k)}, \boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\rangle$$
$$+ \eta_{\boldsymbol{\theta}}^{(k)}\|b_F^{(k)}\|\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\| + \frac{{\eta_{\boldsymbol{\theta}}^{(k)}}^2}{2\alpha}\varsigma_F^2. \tag{20}$$

Since F is ρ-weakly convex and $v_F^{(k)} \in \partial_C F(\boldsymbol{\theta}^{(k)})$, the weak-convexity subgradient inequality gives

$$F(\bar{\boldsymbol{\theta}}^{(k)}) \ge F(\boldsymbol{\theta}^{(k)}) + \langle v_F^{(k)}, \bar{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)}\rangle - \frac{\rho}{2}\|\bar{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^{(k)}\|^2.$$

Also, by the definition of $\bar{\boldsymbol{\theta}}^{(k)}$ and the fact that $\boldsymbol{\theta}^{(k)} \in \Theta$,

$$F(\boldsymbol{\theta}^{(k)}) - F'(\bar{\boldsymbol{\theta}}^{(k)}) \ge \frac{1}{2\alpha}\|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\|^2.$$

Combining the last two displays yields

$$\langle v_F^{(k)}, \boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\rangle \ge \frac{1-\rho\alpha}{2\alpha}\|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\|^2 = \frac{\alpha(1-\rho\alpha)}{2}\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\|^2. \tag{21}$$

Substituting (21) into (20) and using $\|b_F^{(k)}\| \le \Delta_k$, we obtain

$$\mathbb{E}\Big[F'_\alpha(\boldsymbol{\theta}^{((k+1))}) \mid \mathcal{F}_k\Big] \le F'_\alpha(\boldsymbol{\theta}^{(k)}) - \frac{1-\rho\alpha}{2}\eta_{\boldsymbol{\theta}}^{(k)}\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\|^2 + \eta_{\boldsymbol{\theta}}^{(k)}\Delta_k\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\| + \frac{{\eta_{\boldsymbol{\theta}}^{(k)}}^2}{2\alpha}\varsigma_F^2. \tag{22}$$

Because Θ is compact and both $\boldsymbol{\theta}^{(k)}$ and $\bar{\boldsymbol{\theta}}^{(k)}$ lie in Θ,

$$\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\| = \alpha^{-1}\|\boldsymbol{\theta}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\| \le D_\Theta/\alpha, \qquad D_\Theta \doteq \sup_{\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|.$$

Therefore (22) implies

$$\mathbb{E}\Big[F'_\alpha(\boldsymbol{\theta}^{((k+1))}) \mid \mathcal{F}_k\Big] \le F'_\alpha(\boldsymbol{\theta}^{(k)}) - \frac{1-\rho\alpha}{2}\eta_{\boldsymbol{\theta}}^{(k)}\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\|^2 + \frac{D_\Theta}{\alpha}\eta_{\boldsymbol{\theta}}^{(k)}\Delta_k + \frac{{\eta_{\boldsymbol{\theta}}^{(k)}}^2}{2\alpha}\varsigma_F^2. \tag{23}$$

Taking total expectation, summing from $k = 0$ to $K-1$, and using $F'_\alpha(\boldsymbol{\theta}^{(K)}) \ge \inf_{\boldsymbol{\theta}} F'(\boldsymbol{\theta})$ gives

$$\frac{1-\rho\alpha}{2}\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}\,\mathbb{E}\Big[\|\nabla F'_\alpha(\boldsymbol{\theta}^{(k)})\|^2\Big] \le F'_\alpha(\boldsymbol{\theta}^{(0)}) - \inf_{\boldsymbol{\theta}} F'(\boldsymbol{\theta}) + \frac{\varsigma_F^2}{2\alpha}\sum_{k=0}^{K-1}{\eta_{\boldsymbol{\theta}}^{(k)}}^2$$
$$+ \frac{D_\Theta}{\alpha}\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}\Delta_k.$$

Dividing by $\sum_{k=0}^{K-1}\eta_{\boldsymbol{\theta}}^{(k)}$ and using the definition of the random index R yields the claimed bound. With empirical hinge selectors, the same argument gives the same bound with Δ_k^{emp} in place of Δ_k. □

Assumption 3 (Sufficient Statistics)**.** *A statistic $\Phi_{i,\tau}$ is* sufficient (for control) *if it satisfies*

- ***(Markovian transition)*** *There exists a Borel-measurable stochastic kernel $\hat{P}_{i,\tau}(\cdot \mid y_{i,\tau}, a_{i,\tau})$ on $\mathcal{Y}_{i,\tau+1}$ s.t. for any $\underline{\mathcal{Y}_{i,\tau+1}} \in \mathscr{B}_{\mathcal{Y}_{i,\tau+1}}$ and any $\iota_{i,\tau}$ with $\Phi_{i,\tau}(\iota_{i,\tau}) = y_{i,\tau}$, $\mathbb{P}^{\boldsymbol{\pi}}\Big[\Phi_{i,\tau+1}(\iota_{i,\tau+1}) \in \underline{\mathcal{Y}_{i,\tau+1}} \mid \iota_{i,\tau}, a_{i,\tau}\Big] = \hat{P}_{i,\tau}\Big(\underline{\mathcal{Y}_{i,\tau+1}} \mid y_{i,\tau}, a_{i,\tau}\Big)$.*
- ***(Reward Sufficiency)*** *There exists a measurable function $\hat{r}_{i,\tau} : \mathcal{Y}_{i,\tau} \times \mathcal{A} \to \mathbb{R}$ s.t. for any $\iota_{i,\tau}$ with $\Phi_{i,\tau}(\iota_{i,\tau}) = y_{i,\tau}$, $\mathbb{E}^{\boldsymbol{\pi}}\left[r_i(s_\tau, a_{i,\tau}) \mid \iota_{i,\tau}, a_{i,\tau}\right] = \hat{r}_{i,\tau}(y_{i,\tau}, a_{i,\tau})$.*

Truncation convention. Our goal is to solve the infinite-horizon deviation problem. In computation, however, policy-gradient and Fisher estimates are obtained from finite rollouts of length H. Accordingly, the stochastic NPG analysis is carried out for the truncated value $J_{i,H}$, for which the length-H trajectory estimators are unbiased. The resulting guarantee is then lifted back to the infinite-horizon objective J_i using the uniform truncation bound. For finite-horizon deviation problems, we set $H = T$. In this case, the finite-horizon SSMDP may be written as a stationary MDP by augmenting the sufficient statistic with the time index and an absorbing terminal state. For infinite-horizon deviation problems, H is only a rollout truncation parameter; when the infinite-horizon policy class is stationary, the rollout time index is used only in the analysis of truncated value and advantage functions, not as an additional policy input.

For $\boldsymbol{w} \in \mathcal{W}$, define the normalized discounted state-action occupancy measure induced by $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}$ in $\mathcal{S}_i^{\boldsymbol{\theta}}$ by

$$\delta_i^{\boldsymbol{w},\boldsymbol{\theta}}(B) \doteq (1-\gamma) \mathop{\mathbb{E}}_{\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}} \left[\sum_{h=0}^{\infty} \gamma^h \mathbf{1}\{(y_{i,h}, a_{i,h}) \in B\} \right]. \tag{24}$$

The corresponding infinite-horizon value and Fisher matrix are

$$J_i(\boldsymbol{w};\boldsymbol{\theta}) \doteq \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{(y_i,a_i)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta}}} \left[\hat{r}_i(y_i,a_i)\right], \tag{25}$$

$$F_i(\boldsymbol{w};\boldsymbol{\theta}) \doteq \mathop{\mathbb{E}}_{(y_i,a_i)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta}}} \left[\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i) \nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)^\top\right]. \tag{26}$$

For the truncated analysis, define the normalized length-H discounted state-action-time occupancy measure by

$$\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}(B) \doteq \frac{1-\gamma}{1-\gamma^H} \mathop{\mathbb{E}}_{\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}} \left[\sum_{h=0}^{H-1} \gamma^h \mathbf{1}\{(y_{i,h}, a_{i,h}, h) \in B\} \right]. \tag{27}$$

The corresponding truncated value and Fisher matrix are

$$J_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) \doteq \frac{1-\gamma^H}{1-\gamma} \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}} \left[\hat{r}_i(y_i,a_i)\right], \tag{28}$$

$$F_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) \doteq \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}} \left[\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i) \nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)^\top\right]. \tag{29}$$

The rollout expectations in (24) and (27) are over rollouts in $\mathcal{S}_i^{\boldsymbol{\theta}}$ from the oracle initial distribution, with actions drawn from $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}$. Let $\upsilon_i^{\boldsymbol{w},\boldsymbol{\theta},H}$ denote the state marginal of $\delta_i^{\boldsymbol{w},\boldsymbol{\theta},H}$.

Assumption 4 (SSMDP and Policy Regularity)**.** *For each agent $i \in [n]$, mechanism parameter $\boldsymbol{\theta} \in \Theta$, and rollout horizon $H \in \mathbb{N}$, the SSMDP $\mathcal{S}_i^{\boldsymbol{\theta}}$ and policy class $\{\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{W}\}$ satisfy:*

1. ***(Measurable spaces)*** *$(\mathcal{Y}_i, \mathscr{B}_{\mathcal{Y}_i})$ is a Borel space, and $\mathcal{A}_i$ is a nonempty compact Borel space.*

2. ***(Markov kernel regularity)*** *$\hat{P}_i(\cdot \mid y_i, a_i)$ is a Borel-measurable stochastic kernel on $\mathcal{Y}_i$. Moreover, for each bounded measurable $f : \mathcal{Y}_i \to \mathbb{R}$, the map*

$$(y_i, a_i) \mapsto \int f(y_i')\, \hat{P}_i(dy_i' \mid y_i, a_i)$$

is measurable.

3. ***(Bounded rewards)*** *The SSMDP reward is uniformly bounded:*

$$\sup_{y_i\in\mathcal{Y}_i,\ a_i\in\mathcal{A}_i} |\hat{r}_i(y_i,a_i)| \le R_{\max} < \infty.$$

4. ***(Policy regularity)*** *There exist constants $G, M < \infty$ such that, for all y_i, a_i, and $\boldsymbol{w}, \boldsymbol{w}' \in \mathcal{W}$,*

$$\|\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)\| \le G,$$

and

$$\|\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i) - \nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}'}(a_i \mid y_i)\| \le M\|\boldsymbol{w} - \boldsymbol{w}'\|.$$

5. ***(Policy-class expressiveness)*** *Define the inherent infinite-horizon policy-class approximation gap*

$$\varepsilon_{\mathrm{approx}}(i,\boldsymbol{\theta}) \doteq \sup_{\hat{\boldsymbol{\pi}}_i \in \hat{\Pi}_i} J_i(\hat{\boldsymbol{\pi}}_i;\boldsymbol{\theta}) - \sup_{\boldsymbol{w}\in\mathcal{W}} J_i(\boldsymbol{w};\boldsymbol{\theta}).$$

We assume $\varepsilon_{\mathrm{approx}}(i,\boldsymbol{\theta}) < \infty$.

6. ***(Truncated comparator)*** *Let* $A_{i,H}^{\boldsymbol{w},\boldsymbol{\theta}}$ *denote the advantage function of* $\hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}$ *for the length-*H *truncated value in* (28)*. Thus*

$$J_{i,H}(\boldsymbol{w}';\boldsymbol{\theta}) - J_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) = \frac{1-\gamma^H}{1-\gamma} \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\boldsymbol{w}',\boldsymbol{\theta},H}} \left[A_{i,H}^{\boldsymbol{w},\boldsymbol{\theta}}(y_i,a_i,h)\right].$$

We assume that the restricted truncated optimum is attained: there exists

$$w_{i,H}^*(\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{w}\in\mathcal{W}} J_{i,H}(\boldsymbol{w};\boldsymbol{\theta}).$$

7. ***(Truncated compatible function approximation)*** *Define the exact truncated natural policy-gradient direction*

$$d_{i,H}^{\natural}(\boldsymbol{w};\boldsymbol{\theta}) \doteq F_{i,H}(\boldsymbol{w};\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w};\boldsymbol{\theta}).$$

Let $\delta_i^{\star,\boldsymbol{\theta},H} \doteq \delta_i^{w_{i,H}^*(\boldsymbol{\theta}),\boldsymbol{\theta},H}$ *be the normalized discounted state-action occupancy measure of the best-in-class truncated deviation policy* $\hat{\boldsymbol{\pi}}_i^{w_{i,H}^*(\boldsymbol{\theta})}$*. We assume*

$$\sup_{\boldsymbol{w}\in\mathcal{W}} \mathop{\mathbb{E}}_{(y_i,a_i,h)\sim\delta_i^{\star,\boldsymbol{\theta},H}} \left[A_{i,H}^{\boldsymbol{w},\boldsymbol{\theta}}(y_i,a_i,h) - \frac{1-\gamma}{1-\gamma^H} d_{i,H}^{\natural}(\boldsymbol{w};\boldsymbol{\theta})^{\top}\nabla_{\boldsymbol{w}} \log \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}}(a_i \mid y_i)\right]^2 \le \varepsilon_{\mathrm{cfa},H}(i,\boldsymbol{\theta}) < \infty.$$

8. ***(Nondegeneracy of truncated Fisher information)*** *There exists* $\lambda_F > 0$ *such that, for all* $\boldsymbol{w} \in \mathcal{W}$*,*

$$F_{i,H}(\boldsymbol{w};\boldsymbol{\theta}) \succeq \lambda_F I.$$

9. ***(Bounded comparator KL)*** *There exists* $D_{\mathrm{KL},H} < \infty$ *such that*

$$\mathop{\mathbb{E}}_{y_i\sim\upsilon_i^{w_{i,H}^*(\boldsymbol{\theta}),\boldsymbol{\theta},H}} \left[\mathrm{KL}\Big(\hat{\boldsymbol{\pi}}_i^{w_{i,H}^*(\boldsymbol{\theta})}(\cdot \mid y_i) \,\Big\|\, \hat{\boldsymbol{\pi}}_i^{\boldsymbol{w}^{(0)}}(\cdot \mid y_i)\Big)\right] \le D_{\mathrm{KL},H}.$$

10. ***(Bounded truncated policy gradient)*** *There exists* $G_{J,H} < \infty$ *such that, for all* $\boldsymbol{w} \in \mathcal{W}$*,*

$$\|\nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w};\boldsymbol{\theta})\| \le G_{J,H}.$$

Assumption 5 (Conditional Stochastic Estimation for Truncated NPG)**.** *Fix an agent* $i \in [n]$*, mechanism parameter* $\boldsymbol{\theta} \in \Theta$*, rollout horizon* $H \in \mathbb{N}$*, and inner NPG iteration* t*. Let* $\mathcal{F}_t$ *denote the filtration generated by all randomness before the length-*H *trajectories at inner iteration* t *are sampled. For the current parameter* $\boldsymbol{w}^{(t)}$*, let*

$$\widehat{g}_{i,H}^{(t)} \equiv \nabla_{\boldsymbol{w}} \widehat{J_{i,H}(\boldsymbol{w}^{(t)};\boldsymbol{\theta})}$$

be the empirical truncated policy-gradient estimator computed from κ *i.i.d. length-*H *trajectories. Let* $\widehat{F}_{i,H}^{(t)}$ *be the corresponding empirical truncated Fisher matrix, and let* $\widetilde{F}_{i,H}^{(t)}$ *denote the regularized Fisher matrix actually inverted in the NPG update, for example after spectral clipping or ridge regularization. The stochastic NPG direction is*

$$\widehat{d}_{i,H}^{(t)} \doteq \left(\widetilde{F}_{i,H}^{(t)}\right)^{-1} \widehat{g}_{i,H}^{(t)}.$$

The following conditions hold.

1. ***(Conditional unbiasedness and variance of truncated gradient)***

$$\mathbb{E}[\widehat{g}_{i,H}^{(t)} \mid \mathcal{F}_t] = \nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta}),$$

and

$$\mathbb{E}\Big[\|\widehat{g}_{i,H}^{(t)} - \nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta})\|^2 |\, \mathcal{F}_t\Big] \le \frac{\varsigma_J^2}{\kappa}.$$

2. ***(Regularized Fisher inverse control)*** *Conditional on $\mathcal{F}_t$,*

$$\widetilde{F}_{i,H}^{(t)} \succeq \frac{\lambda_F}{2} I \qquad \textit{almost surely},$$

and

$$\mathbb{E}\Big[\|\widetilde{F}_{i,H}^{(t)} - F_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta})\|_{\mathrm{op}} |\, \mathcal{F}_t\Big] \le \varepsilon_F(\kappa, H),$$

where $\varepsilon_F(\kappa, H) \to 0$ as $\kappa, H \to \infty$.

Theorem 4.2. *[Deviation-oracle accuracy via stochastic NPG] Fix a mechanism parameter $\boldsymbol{\theta} \in \Theta$ and an agent $i \in [n]$. Let $J_i^*(\boldsymbol{\theta}) \doteq \sup_{\hat{\boldsymbol{\pi}}_i \in \hat{\Pi}_i} J_i(\hat{\boldsymbol{\pi}}_i; \boldsymbol{\theta})$. Let $\boldsymbol{w}^{(0)}, \dots, \boldsymbol{w}^{(K-1)}$ be generated by stochastic truncated NPG on the H-truncated SSMDP, and let $T \sim \mathrm{Unif}\{0, \dots, K-1\}$ be independent of the rollout randomness. Under Assumptions 4–5, the random-iterate $\boldsymbol{w}^{(T)}$ satisfies*

$$\mathbb{E}\Big[J_i^*(\boldsymbol{\theta}) - J_i(\boldsymbol{w}^{(T)}; \boldsymbol{\theta})\Big] \le \underbrace{\varepsilon_{\mathrm{approx}}(i, \boldsymbol{\theta})}_{\textit{policy-class approximation}} + \underbrace{\frac{1-\gamma^H}{1-\gamma}\sqrt{\varepsilon_{\mathrm{cfa},H}(i, \boldsymbol{\theta})}}_{\textit{compatible approximation}} + \underbrace{\frac{D_{\mathrm{KL},H}}{\eta_{\boldsymbol{w}} K}}_{\textit{optimization}} + \underbrace{\frac{M\eta_{\boldsymbol{w}}}{2} C_{d,H}^2}_{\textit{direction size}} + \underbrace{G C_{\mathrm{dir},H}}_{\textit{direction error}} + \underbrace{2\frac{\gamma^H}{1-\gamma} R_{\max}}_{\textit{truncation}}. \tag{8}$$

Moreover, the same bound holds for the true best-so-far iterate.

The constants in (8) are

$$C_{d,H}^2 \doteq \frac{4}{\lambda_F^2}\left(G_{J,H}^2 + \frac{\varsigma_J^2}{\kappa}\right), \qquad C_{\mathrm{dir},H} \doteq \frac{2G_{J,H}}{\lambda_F^2}\varepsilon_F(\kappa, H) + \frac{2\varsigma_J}{\lambda_F\sqrt{\kappa}}.$$

Since $(1-\gamma^H)/(1-\gamma) \le (1-\gamma)^{-1}$, the compatible approximation term is at most $\sqrt{\varepsilon_{\mathrm{cfa},H}(i, \boldsymbol{\theta})}/(1-\gamma)$. A validation-best iterate selected using independent rollout estimates incurs the usual additional validation error.

Proof. We first analyze the H-truncated SSMDP and then lift the result to the infinite-horizon objective. Define

$$J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) \doteq \sup_{\boldsymbol{w} \in \mathcal{W}} J_{i,H}(\boldsymbol{w}; \boldsymbol{\theta}) = J_{i,H}(w_{i,H}^*(\boldsymbol{\theta}); \boldsymbol{\theta}).$$

For notational compactness, write

$$g_H^{(t)} \doteq \nabla_{\boldsymbol{w}} J_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta}), \qquad F_H^{(t)} \doteq F_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta}),$$
$$d_H^{\natural,(t)} \doteq \left(F_H^{(t)}\right)^{-1} g_H^{(t)}, \qquad \widehat{d}_H^{(t)} \doteq \left(\widetilde{F}_{i,H}^{(t)}\right)^{-1} \widehat{g}_{i,H}^{(t)}.$$

Step 1: Liu-style average-iterate inequality for the truncated objective. The proof of Liu et al. [44, Proposition 4.5] applies to the time-augmented H-truncated SSMDP. Using the score-smoothness constant M, the score bound G, the finite-horizon performance-difference identity in Assumption 4, and the truncated compatible-function-approximation condition, we obtain the pathwise bound

$$J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) - \frac{1}{K}\sum_{t=0}^{K-1} J_{i,H}(\boldsymbol{w}^{(t)}; \boldsymbol{\theta}) \le \frac{1-\gamma^H}{1-\gamma}\sqrt{\varepsilon_{\mathrm{cfa},H}(i, \boldsymbol{\theta})} + \frac{D_{\mathrm{KL},H}}{\eta_{\boldsymbol{w}} K} + \frac{M\eta_{\boldsymbol{w}}}{2K}\sum_{t=0}^{K-1}\|\widehat{d}_H^{(t)}\|^2 + \frac{G}{K}\sum_{t=0}^{K-1}\|\widehat{d}_H^{(t)} - d_H^{\natural,(t)}\|.$$

The factor $(1-\gamma^H)/(1-\gamma)$ appears because the normalized truncated occupancy satisfies

$$\mathop{\mathbb{E}}_{\delta_i^\star,\boldsymbol{\theta},H}[A_{i,H}^{\boldsymbol{w},\boldsymbol{\theta}}] = \frac{1-\gamma}{1-\gamma^H}\left(J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) - J_{i,H}(\boldsymbol{w};\boldsymbol{\theta})\right).$$

Step 2: Bound the squared update direction. By Assumption 5,

$$\left\|\left(\widetilde{F}_{i,H}^{(t)}\right)^{-1}\right\|_{\mathrm{op}} \le \frac{2}{\lambda_F} \qquad \text{almost surely.}$$

Therefore, conditional on $\mathcal{F}_t$,

$$\begin{aligned}
\mathbb{E}[\|\widehat{d}_H^{(t)}\|^2 |\, \mathcal{F}_t] &\le \frac{4}{\lambda_F^2}\,\mathbb{E}[\|\widehat{g}_{i,H}^{(t)}\|^2 |\, \mathcal{F}_t] \\
&\le \frac{4}{\lambda_F^2}\left(\|g_H^{(t)}\|^2 + \frac{\varsigma_J^2}{\kappa}\right) \\
&\le \frac{4}{\lambda_F^2}\left(G_{J,H}^2 + \frac{\varsigma_J^2}{\kappa}\right) \doteq C_{d,H}^2.
\end{aligned}$$

Step 3: Bound the direction error. Using the identity

$$\begin{aligned}
\widehat{d}_H^{(t)} - d_H^{\natural,(t)} = &\left[\left(\widetilde{F}_{i,H}^{(t)}\right)^{-1} - \left(F_H^{(t)}\right)^{-1}\right] g_H^{(t)} \\
&+ \left(\widetilde{F}_{i,H}^{(t)}\right)^{-1}\left(\widehat{g}_{i,H}^{(t)} - g_H^{(t)}\right),
\end{aligned}$$

Assumptions 4–5 give

$$\left\|\left(\widetilde{F}_{i,H}^{(t)}\right)^{-1} - \left(F_H^{(t)}\right)^{-1}\right\|_{\mathrm{op}} \le \frac{2}{\lambda_F^2}\|\widetilde{F}_{i,H}^{(t)} - F_H^{(t)}\|_{\mathrm{op}}.$$

Hence, conditional on $\mathcal{F}_t$,

$$\begin{aligned}
\mathbb{E}[\|\widehat{d}_H^{(t)} - d_H^{\natural,(t)}\| |\, \mathcal{F}_t] &\le \frac{2G_{J,H}}{\lambda_F^2}\varepsilon_F(\kappa,H) + \frac{2}{\lambda_F}\,\mathbb{E}[\|\widehat{g}_{i,H}^{(t)} - g_H^{(t)}\| |\, \mathcal{F}_t] \\
&\le \frac{2G_{J,H}}{\lambda_F^2}\varepsilon_F(\kappa,H) + \frac{2\varsigma_J}{\lambda_F\sqrt{\kappa}} \doteq C_{\mathrm{dir},H},
\end{aligned}$$

where the last inequality uses Jensen's inequality and the conditional variance bound in Assumption 5.

Step 4: Take expectations and pass to a uniform random iterate. Taking total expectation in Step 1 and using Steps 2–3 yields

$$\begin{aligned}
J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) - \frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[J_{i,H}(\boldsymbol{w}^{(t)};\boldsymbol{\theta})] \le &\frac{1-\gamma^H}{1-\gamma}\sqrt{\varepsilon_{\mathrm{cfa},H}(i,\boldsymbol{\theta})} + \frac{D_{\mathrm{KL},H}}{\eta_{\boldsymbol{w}}K} \\
&+ \frac{M\eta_{\boldsymbol{w}}}{2}C_{d,H}^2 + GC_{\mathrm{dir},H}.
\end{aligned}$$

Since $T \sim \mathrm{Unif}\{0,\dots,K-1\}$ is independent of the NPG randomness,

$$\mathbb{E}[J_{i,H}(\boldsymbol{w}^{(T)};\boldsymbol{\theta})] = \frac{1}{K}\sum_{t=0}^{K-1}\mathbb{E}[J_{i,H}(\boldsymbol{w}^{(t)};\boldsymbol{\theta})].$$

Thus

$$\begin{aligned}
J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) - \mathbb{E}[J_{i,H}(\boldsymbol{w}^{(T)};\boldsymbol{\theta})] \le &\frac{1-\gamma^H}{1-\gamma}\sqrt{\varepsilon_{\mathrm{cfa},H}(i,\boldsymbol{\theta})} + \frac{D_{\mathrm{KL},H}}{\eta_{\boldsymbol{w}}K} \\
&+ \frac{M\eta_{\boldsymbol{w}}}{2}C_{d,H}^2 + GC_{\mathrm{dir},H}.
\end{aligned}$$

Step 5: Lift from the truncated objective to the infinite-horizon objective. By bounded rewards,

$$\sup_{\boldsymbol{w}\in\mathcal{W}} |J_i(\boldsymbol{w};\boldsymbol{\theta}) - J_{i,H}(\boldsymbol{w};\boldsymbol{\theta})| \le e_H, \qquad e_H \doteq \frac{\gamma^H}{1-\gamma} R_{\max}.$$

Therefore,

$$\sup_{\boldsymbol{w}\in\mathcal{W}} J_i(\boldsymbol{w};\boldsymbol{\theta}) - \mathbb{E}[J_i(\boldsymbol{w}^{(T)};\boldsymbol{\theta})] \le J_{i,H}^{\star,\mathcal{W}}(\boldsymbol{\theta}) - \mathbb{E}[J_{i,H}(\boldsymbol{w}^{(T)};\boldsymbol{\theta})] + 2e_H.$$

Finally, by the definition of the infinite-horizon policy-class approximation gap,

$$J_i^*(\boldsymbol{\theta}) - \sup_{\boldsymbol{w}\in\mathcal{W}} J_i(\boldsymbol{w};\boldsymbol{\theta}) = \varepsilon_{\text{approx}}(i,\boldsymbol{\theta}).$$

Combining the last three displays proves (8).

The ideal best-so-far claim follows because, for $t^\star \in \arg\max_{0\le t<K} J_i(\boldsymbol{w}^{(t)};\boldsymbol{\theta})$,

$$J_i(\boldsymbol{w}^{(t^\star)};\boldsymbol{\theta}) \ge \frac{1}{K}\sum_{t=0}^{K-1} J_i(\boldsymbol{w}^{(t)};\boldsymbol{\theta})$$

for every realized trajectory of the algorithm. If a validation maximizer $\widehat{t}$ is used instead and

$$\max_{0\le t<K} |\widehat{J}_{\text{val}}(\boldsymbol{w}^{(t)};\boldsymbol{\theta}) - J_i(\boldsymbol{w}^{(t)};\boldsymbol{\theta})| \le \varepsilon_{\text{val}},$$

then the standard validation-selection argument gives an additional $2\varepsilon_{\text{val}}$ term. □

Corollary 1. *[Inner oracle error via stochastic NPG] Fix an outer iteration r, agent $i \in [n]$, mechanism parameter $\boldsymbol{\theta}^{(r)} \in \Theta$, and inner NPG parameters $(K_r, H_r, \kappa_r, \eta_{\boldsymbol{w}}^{(r)})$. Define*

$$\bar{\varepsilon}_{i,\text{NPG}}^{(r)} \doteq \varepsilon_{\text{approx}}(i,\boldsymbol{\theta}^{(r)}) + \frac{1-\gamma^{H_r}}{1-\gamma}\sqrt{\varepsilon_{\text{cfa},H_r}(i,\boldsymbol{\theta}^{(r)})} + \frac{D_{\text{KL},H_r}}{\eta_{\boldsymbol{w}}^{(r)} K_r} + \frac{M\eta_{\boldsymbol{w}}^{(r)}}{2} C_{d,H_r}^2 + GC_{\text{dir},H_r} + 2\frac{\gamma^{H_r}}{1-\gamma} R_{\max}.$$

By the SSMDP reduction, optimizing over $\widehat{\mathcal{P}}_i$ is value-equivalent to optimizing over admissible reporting deviations in P_i, so $J_i^\star(\theta) = V_i(\theta)$. Under Assumptions 4–5, the stochastic NPG oracle output $\boldsymbol{\pi}_i^{(r)}$ satisfies

$$\mathbb{E}\Big[V_i(\boldsymbol{\theta}^{(r)}) - R_i(\boldsymbol{\pi}_i^{(r)};\boldsymbol{\theta}^{(r)}) \mid \mathcal{F}_r\Big] \le \bar{\varepsilon}_{i,\text{NPG}}^{(r)}.$$

F Budget Auction Experiment Details

The budget-auction experiments study how principal-side history compression affects representation and optimization. We first specify the environments, network classes, and common training protocol, then give the frozen three-seed results and the additional refinement and distillation studies.

F.1 Environments and Feasibility

There are two buyers, $T = 5$ periods, discount factor $\gamma = 0.9$, and initial budget $B_i = 3$ for each buyer. In this appendix, $v_{i,t}$ and $r_{i,t}$ denote a buyer's value and report, $x_{i,t}$ her allocation vector, and $p_{i,t}$ her payment. Stage utility is $v_{i,t} \cdot x_{i,t} - p_{i,t}$. The seller maximizes expected discounted revenue; allocation costs are zero. All coordinates of values and reports lie in $[0, 1]$.

In `e2`, write $s_{i,g,t} = \text{logit}(v_{i,g,t})$, with values clipped to $[10^{-6}, 1-10^{-6}]$ for this transformation. The transition is $s_{i,g,t+1} = s_{i,g,t} + [0.01(1-x_{i,g,t}) + x_{i,g,t}]\epsilon_{i,g,t}$ and $v_{i,g,t+1} = \text{sigmoid}(s_{i,g,t+1})$, where the $\epsilon_{i,g,t}$ are independent standard Gaussian variables. This transition uses the current allocation, without an additional accumulated-allocation factor. Thus, `e2` changes both the number of goods and the type dynamics relative to `e1`.

Setting	Goods	Initial values	Subsequent values
`e1`	1	Independent $\mathrm{Unif}[0,1]$	Independent $\mathrm{Unif}[0,1]$
`e2`	2	Independent $\mathrm{Unif}[0,1]^2$	Allocation-dependent logit dynamics

Table 3: Budget-auction environments. Buyers are independent initially; in `e2`, each buyer's future values depend on her private current values and allocations.

History-compression reference. The theoretical reference point is the generalized bank account mechanism (GBAM) framework of Deng et al. [24]. In its budget application, the state tracks each buyer's cumulative utility balance $\alpha_{i,t} = \sum_{s<t}(r_{i,s} \cdot x_{i,s} - p_{i,s})$ and remaining budget $\beta_{i,t} = B_i - \sum_{s<t} p_{i,s}$. Under their assumptions, including independence of types across buyers and periods, this compressed state suffices to match the revenue of any DIC, ex-post DIR, budget-feasible dynamic mechanism. We use this informational benchmark for `e1`; its guarantee does not extend to the privately correlated dynamics of `e2`. This is an existence and compression result: it neither characterizes the optimal mechanism or revenue in closed form nor guarantees that a particular finite network and gradient-based training procedure can recover them.

Hard feasibility transform. Independently of the history representation, all four mechanism architectures use the same output transform to enforce allocation feasibility, budget feasibility, and ex-post DIR in both settings. This construction uses the balances defined above and does not rely on the GBAM sufficiency guarantee. For each good, a softmax over buyer logits and a learned no-allocation logit ensures $\sum_i x_{i,g,t} \le 1$. If $z_{i,t}$ is the payment head's output, the payment is

$$p_{i,t} = \max\{0, \min\{\beta_{i,t}, \alpha_{i,t} + r_{i,t} \cdot x_{i,t}\}\} \,\mathrm{sigmoid}(z_{i,t}). \tag{30}$$

The balances update by $\alpha_{i,t+1} = \alpha_{i,t} + r_{i,t} \cdot x_{i,t} - p_{i,t}$ and $\beta_{i,t+1} = \beta_{i,t} - p_{i,t}$. Starting from $(0, B_i)$, these updates keep both balances nonnegative. Hence cumulative payments respect the budget, and truthful reporting gives nonnegative cumulative utility at every prefix. Summation by parts then also gives nonnegative discounted truthful utility. The mechanism receives reports and its own recorded outcomes, not the simulator's hidden private values.

F.2 Mechanism Representation Variants

We compare four representations of each buyer's history. All use the current report $r_{i,t}$, the balances $\alpha_{i,t}$ and $\beta_{i,t}$, and normalized time $\tau_t = (t-1)/(T-1)$.

Variant `A`. The compact GBAM-style representation uses only $(r_{i,t}, \alpha_{i,t}, \beta_{i,t}, \tau_t)$.

Variant `A_wide`. This variant uses exactly the same inputs and feature-encoder architecture as `A`, but a larger subsequent MLP and more trainable parameters. It introduces no additional information.

Variant `B`. This variant adds $(\sum_{s<t} x_{i,s}, \sum_{s<t} p_{i,s})$ to `A`. Since $\sum_{s<t} p_{i,s} = B_i - \beta_{i,t}$, cumulative payments are informationally redundant, but exposing them separately can change optimization.

Variant `C`. This variant retains the cumulative summaries in `B` and additionally encodes the ordered histories $(r_{i,s})_{s<t}$, $(x_{i,s})_{s<t}$, and $(p_{i,s})_{s<t}$ through learned, time-indexed attention pools.

All variants use the common feasibility transform in Section F.1. Within each setting, `A_wide`, `B`, and `C` have equal numbers of trainable parameters; `A` has fewer parameters. Network dimensions are specified next.

F.3 Network Architecture

F.3.1 Mechanism Networks

Write $\mathrm{MLP}_{d_0 \to d_1 \to \cdots \to d_L}$ for a fully connected network with input dimension d_0, layer widths $d_1, \ldots, d_L$, ReLU activations between affine layers, and an affine output layer. A superscript $+$ additionally applies ReLU to the output. Brackets with $\|$ denote concatenation. Network weights are shared across buyers within each mechanism; the four variants are trained separately.

Representation	Concatenated feature dimension (`e1`; `e2`)	Trainable parameters	
		`e1`	`e2`
`A`	96; 96	37,859	38,053
`A_wide`	96; 96	57,366	57,688
`B`	143; 146	57,366	57,688
`C`	224; 224	57,366	57,688

Table 4: Feature dimensions and model sizes of the budget mechanism networks. Parameter counts include feature encoders, the shared MLP, and both output heads. Matching counts controls network size, not functional equivalence of the resulting classes.

Common feature encoders. Let G denote the number of goods. Each variant encodes the current report and normalized balances separately:

$$h_{\mathrm{rep},i,t} = \mathrm{MLP}_{G\to 128\to 64}(r_{i,t}) \in \mathbb{R}^{64},$$
$$h_{\mathrm{st},i,t} = \mathrm{MLP}_{3\to 64\to 32}([\alpha_{i,t}/B_i \| \beta_{i,t}/B_i \| \tau_t]) \in \mathbb{R}^{32}.$$

Variant `A`. Concatenate these features and apply the shared MLP:

$$q^A_{i,t} = [h_{\mathrm{rep},i,t} \| h_{\mathrm{st},i,t}] \in \mathbb{R}^{96}, \qquad h^A_{i,t} = \mathrm{MLP}^+_{96\to 128\to 64}(q^A_{i,t}) \in \mathbb{R}^{64}.$$

Variant `A_wide`. The concatenated feature vector remains $q^A_{i,t} \in \mathbb{R}^{96}$. The subsequent MLP is

$$h^{A_\mathrm{wide}}_{i,t} = \begin{cases} \mathrm{MLP}^+_{96\to 187\to 86\to 66}(q^A_{i,t}), & \texttt{e1}, \\ \mathrm{MLP}^+_{96\to 193\to 87\to 58}(q^A_{i,t}), & \texttt{e2}. \end{cases}$$

Thus, the increase in model size comes from additional network parameters, with identical input information and input-feature dimension.

Variant `B`. Encode the cumulative summaries as

$$h_{\mathrm{alloc},i,t} = \mathrm{MLP}_{G\to 64\to 32}\left(\sum_{s<t} x_{i,s}\right) \in \mathbb{R}^{32},$$
$$h^B_{\mathrm{pay},i,t} = \mathrm{MLP}_{1\to 32\to d_p}\left(\sum_{s<t} p_{i,s}\right) \in \mathbb{R}^{d_p},$$
$$q^B_{i,t} = [h_{\mathrm{rep},i,t} \| h_{\mathrm{st},i,t} \| h_{\mathrm{alloc},i,t} \| h^B_{\mathrm{pay},i,t}] \in \mathbb{R}^{128+d_p},$$

where $d_p = 15$ in `e1` and $d_p = 18$ in `e2`. The shared MLP is

$$h^B_{i,t} = \begin{cases} \mathrm{MLP}^+_{143\to 160\to 82}(q^B_{i,t}), & \texttt{e1}, \\ \mathrm{MLP}^+_{146\to 143\to 98}(q^B_{i,t}), & \texttt{e2}. \end{cases}$$

Variant `C`. This variant uses the same allocation encoder as `B`, a cumulative-payment encoder $h^C_{\mathrm{pay},i,t} = \mathrm{MLP}_{1\to 32\to 16}\big(\sum_{s<t} p_{i,s}\big)$, and three history encoders:

$$h^{\mathrm{rh}}_{i,t} = \mathrm{AttPool}_{32}((r_{i,s})_{s<t}) \in \mathbb{R}^{32},$$
$$h^{\mathrm{ah}}_{i,t} = \mathrm{AttPool}_{32}((x_{i,s})_{s<t}) \in \mathbb{R}^{32},$$
$$h^{\mathrm{ph}}_{i,t} = \mathrm{AttPool}_{16}((p_{i,s})_{s<t}) \in \mathbb{R}^{16},$$
$$q^C_{i,t} = [h_{\mathrm{rep},i,t} \| h_{\mathrm{st},i,t} \| h_{\mathrm{alloc},i,t} \| h^C_{\mathrm{pay},i,t} \| h^{\mathrm{rh}}_{i,t} \| h^{\mathrm{ah}}_{i,t} \| h^{\mathrm{ph}}_{i,t}] \in \mathbb{R}^{224},$$
$$h^C_{i,t} = \mathrm{MLP}^+_{224\to 128\to 64}(q^C_{i,t}) \in \mathbb{R}^{64}.$$

For an input sequence $(y_s)_{s<t}$, $\mathrm{AttPool}_d$ forms $e_s = \mathrm{ReLU}(W[y_s \| \tau_s] + b) \in \mathbb{R}^d$, assigns weights proportional to $\exp(w^\top e_s + b_0)$ over $s < t$, and returns their weighted sum. Unobserved periods are masked, and an empty history produces zero. Each of the three pools has its own parameters.

Allocation and payment heads. For each variant V, let d_V be the dimension of $h_{i,t}^V$. The allocation logits are $\ell_{i,t}^V = W_x h_{i,t}^V + b_x \in \mathbb{R}^G$. For each good, softmax normalization over the buyers' logits and a learned no-allocation logit produces the allocation. The payment score is $z_{i,t}^V = \mathrm{MLP}_{d_V \to 64 \to 32 \to 1}(h_{i,t}^V)$, followed by the common cap in Equation (30).

F.3.2 Agent Deviation-Policy Network

The deviation policy uses an MLP on ordered observation histories, followed by a logit-residual reporting head. Its components are as follows. Histories below are flattened in temporal order and zero-padded: private types have $T+1$ slots, while reports and public outcomes have T slots.

- **Public-history features.** The common feature vector f_t^{pub} contains all buyers' past allocations and payments, initial budgets, and the past-period mask $m_t = (\mathbf{1}\{s < t\})_{s=1}^T$:

$$f_t^{\mathrm{pub}} = [\,\mathrm{vec}((x_{j,s})_{j\in[n],s<t}) \| \mathrm{vec}((p_{j,s})_{j\in[n],s<t}) \| (B_j)_{j\in[n]} \| m_t] \in \mathbb{R}^{TnG+Tn+n+T}.$$

- **Private-history features.** For buyer i, the private feature vector contains her types through the current period and her past reports:

$$f_{i,t}^{\mathrm{priv}} = [\,\mathrm{vec}((v_{i,s})_{s\le t}) \| \mathrm{vec}((r_{i,s})_{s<t})] \in \mathbb{R}^{(T+1)G+TG}.$$

  Opponents' private types and reports are excluded.

- **History MLP.** Concatenate these features with the one-hot buyer identifier $e_i \in \mathbb{R}^n$ and normalized time τ_t, then apply two fully connected ReLU layers:

$$f_{i,t} = [f_{i,t}^{\mathrm{priv}} \| f_t^{\mathrm{pub}} \| e_i \| \tau_t] \in \mathbb{R}^{d_H}, \qquad h_{i,t} = \mathrm{MLP}^+_{d_H \to 128 \to 128}(f_{i,t}) \in \mathbb{R}^{128}.$$

  The input dimension is $d_H = (T+1)G + TG + TnG + Tn + 2n + T + 1$, giving 41 in `e1` and 62 in `e2`. Network weights are shared across buyers.

- **Reporting head.** An affine layer produces a clipped logit residual and transforms the current type into a report:

$$\delta_{i,t} = \mathrm{clip}(W_{\mathrm{out}} h_{i,t} + b_{\mathrm{out}}, -20, 20) \in \mathbb{R}^G,$$
$$r_{i,g,t} = \frac{v_{i,g,t}\exp(\delta_{i,g,t})}{1 - v_{i,g,t} + v_{i,g,t}\exp(\delta_{i,g,t})} \in [0,1].$$

  The output layer is initialized to zero, so the initial policy reports truthfully. The implementation uses a numerically stable form defined at the boundaries, and gradients propagate through the current private type, whose evolution depends on earlier deviations.

The same input construction and reporting head are used with larger hidden widths in the evaluation searches described below. All other buyers report truthfully when one buyer's deviation is evaluated.

F.4 Training, Evaluation, and Frozen Selection

Acceptance training. Each main trial runs 1,500 outer epochs, with 200 inner updates and one mechanism update per epoch, using batches of 64 trajectories. Both Adam learning rates are 3×10^{-4}. The DIC weight is 5 in `e1` and 10 in `e2`; the DIR penalty is zero because feasibility is enforced structurally. Inner optimizer states persist across ordinary outer iterations, and the policy is reset every 50 epochs. Training also compares simple truthful, low, high, and half-value reports. Early stopping is disabled.

Targets and candidate checkpoints. The initial grid has three prescribed seeds for every representation and targets $2.1, 2.2, 2.3$ in `e1` and $3.2, 3.3, 3.4$ in `e2`, giving 72 trials. Additional shared target extensions and two lower-target `e2/A` trials expand the completed candidate pool. These extensions are recorded in the released configuration index. Thus, the per-trial training budget is common, while the total number of selection opportunities is not identical across all representation–seed pairs.

The final selection considers checkpoints at epochs 500, 750, 1,000, 1,250, and 1,500. A regular evaluation ranks eligible checkpoints by revenue, breaking ties in favor of the earlier epoch. Stronger

evaluation is applied to the highest-ranked checkpoint and, if it fails the DIC or feasibility criterion, to the next ranked checkpoint. At most two checkpoints receive this stronger evaluation; missing the revenue target alone does not trigger a second attempt. The selected mechanism for a representation and seed is the highest-revenue valid result across its completed targets. All main selections are frozen before the subsequent exploratory extensions.

Deviation search and evaluation DIC loss. For a fixed mechanism, each buyer's policy is optimized with all other buyers truthful. Let $\widehat{d}_i$ be the estimated discounted payoff difference from truthful reporting for the selected deviation. The budget evaluation DIC loss is $\sum_i \max\{0, \widehat{d}_i\}$. Regular search uses 600 pathwise steps with three restarts and 200 NPG steps with three restarts. Strong search uses 1,200 pathwise steps and 400 NPG steps, each with six restarts, over policy widths 128, 256, and 512. Search batches have 128 trajectories; policy selection uses 2,048 validation trajectories. NPG uses report-noise scale 0.3. Additional evolution-strategy searches use width 128, 200 steps, 16 antithetic pairs, batches of 64 trajectories, perturbation scale 0.01, learning rate 0.001, gradient clipping at 0.5, and three restarts per buyer. The resulting candidates are compared on a further 2,048-trajectory validation sample.

Revenue and payoff differences are then evaluated in 32 batches of 256 trajectories. If $\widehat{\mathrm{se}}_i$ is the standard error of buyer i's batch-mean difference, the reported sampling upper bound is $\sum_i \max\{0, \widehat{d}_i + 1.96\widehat{\mathrm{se}}_i\}$. The factor 1.96 gives a Bonferroni-adjusted one-sided normal approximation for two buyers. A candidate is valid when this bound is at most 0.04 and the observed allocation, payment, budget, and truthful-utility feasibility checks pass. This bound accounts for sampling error for the tested policies; it is not an upper bound on unsearched deviations. Since these evaluations determine the chosen checkpoint and target, the reported main revenues are selected validation results, not held-out test results.

F.5 Frozen Three-Seed Results

Table 5 reports the selected revenues for each seed, and Table 6 summarizes their relative improvements.

Setting	Representation	Seed 1	Seed 2	Seed 3	Mean $\pm$ sample SD
`e1`	`A`	2.14004	2.13137	2.12412	2.13184 ± 0.00797
	`A_wide`	2.16744	2.15106	2.13832	2.15227 ± 0.01460
	`B`	2.19494	2.18723	2.16848	2.18355 ± 0.01361
	`C`	2.18590	2.14588	2.13624	2.15601 ± 0.02634
`e2`	`A`	3.14351	3.09621	3.10888	3.11620 ± 0.02449
	`A_wide`	3.12671	3.14630	3.20181	3.15827 ± 0.03895
	`B`	3.32997	3.44202	3.42340	3.39846 ± 0.06005
	`C`	3.42008	3.41515	3.44352	3.42625 ± 0.01515

Table 5: Frozen budget-auction selected validation revenues. The standard deviation is across three seeds. All 24 selected mechanisms meet the 0.04 sampling-upper-bound criterion and observed feasibility checks. Exact trial identifiers, selected epochs, targets, and checkpoint hashes accompany the released results.

Setting	`A_wide` over `A`	`B` over `A`	`C` over `A`
`e1`	0.96%	2.43%	1.13%
`e2`	1.35%	9.06%	9.95%

Table 6: Relative improvement in mean selected validation revenue. These are descriptive comparisons of the frozen selections.

The modest gains in `e1`, together with the distillation results below, support an approximation/optimization contribution even where compact states have an informational sufficiency guarantee. In `e2`, the richer histories yield larger gains than widening the compact network. This comparison identifies the larger empirical representation gap in the correlated environment; it does not by itself prove that the compact state is insufficient or isolate correlation from the change in the number of goods.

Figure 2 shows the individual training trajectories for these frozen selections.

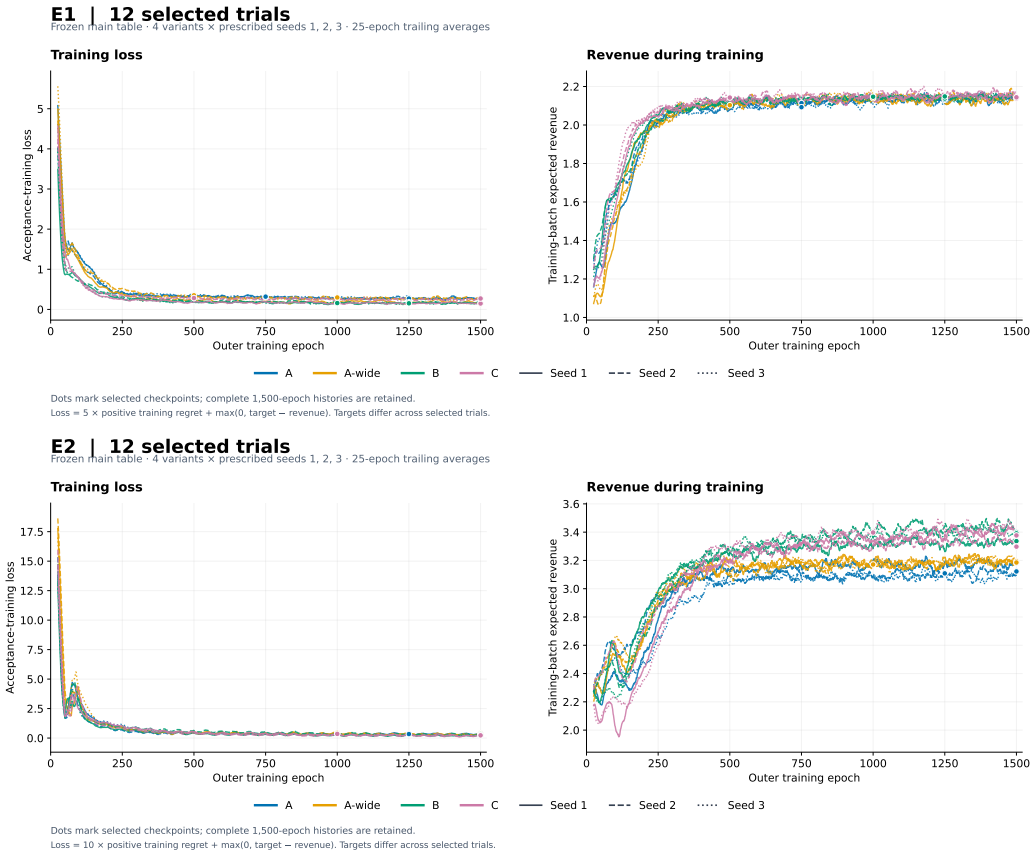


Figure 2: Training trajectories for the frozen budget selections in `e1` (top) and `e2` (bottom). Each curve corresponds to one selected trial and uses a trailing 25-epoch average. Loss and revenue are measured on training batches. Acceptance targets differ across trials, so loss levels are not a common objective-value comparison. A selected checkpoint may precede the end of its displayed trajectory. Figure 1 summarizes these selected runs using mean revenue and $\pm$ one sample standard deviation across seeds.

F.6 Refinement and Distillation

These studies examine how much of the richer representations' advantage can be recovered by additional optimization of a compact mechanism. They retain the main selections above and report separate validation results.

Additional acceptance training. We continue one `e2/A` mechanism and one `e2/C` mechanism for 500 outer epochs with DIC weight 10, retaining the common inner training schedule. Candidate checkpoints are at 250 and 500 additional epochs. The `A` revenue increases from 3.14351 to 3.20789 (selected at 250 epochs; sampling DIC upper bound 0.03986). The `C` revenue increases from 3.44352 to 3.44684 (500 epochs; upper bound 0.03149). Their targets are 3.22230 and 3.52045, respectively. Additional training improves `A` but leaves a substantial gap in this finite refinement study.

Distillation from teacher and acceptance training. Distillation samples 2,048 truthful teacher trajectories for fitting and 512 separate teacher trajectories for validation. At each visited history, an `A` student predicts the teacher's allocation vector, including the no-allocation component, and its payment fraction before the feasibility cap. The supervised objective is the sum of the two mean squared errors, with equal weights. We use Adam with learning rate 10^{-3} and batch size 256, evaluating the fit every 10 epochs. The best validation fit initializes acceptance training. Imitating the teacher on truthful trajectories alone does not impose DIC; the subsequent acceptance stage trains the student against deviation policies.

Setting	Teacher	Teacher revenue	Student revenue	Remaining gap	DIC upper
e1	B	2.19494	2.18399	0.01095	0.03729
e1	C	2.18590	2.15763	0.02828	0.01806
e2	B	3.44171	3.31965	0.12206	0.03546
e2	C	3.44352	3.27944	0.16408	0.03390

Table 7: Distillation into A with 1,000 supervised-fit epochs followed by 1,000 acceptance-training epochs. Revenues are selected validation values; the remaining gap is teacher minus student revenue, and the DIC upper bound is for the selected student. Teacher identities are fixed within each row and need not be a main-table seed selection.

We fit each student for 1,000 supervised epochs, select the best validation fit over epochs 0–1,000, and initialize a fresh 1,000-epoch acceptance run, evaluating checkpoints every 250 epochs. In e1, the acceptance targets are 2.18686 and 2.17633 for the B and C teachers, respectively, with DIC weight 5. In e2, the corresponding targets are 3.36989 and 3.37045, with DIC weight 20.

Distilling from B nearly closes the e1 teacher–student gap, demonstrating that a stronger compact mechanism can be reached from a richer teacher. The tested e2 procedures also improve compact mechanisms but leave larger gaps. These studies do not include matched fresh-A controls, so they establish the performance of the combined procedures rather than a separate causal effect of imitation. The remaining finite-training gaps leave both optimization difficulty and representation limitations open.

Figure 3 shows the acceptance-training trajectories of the distilled students.

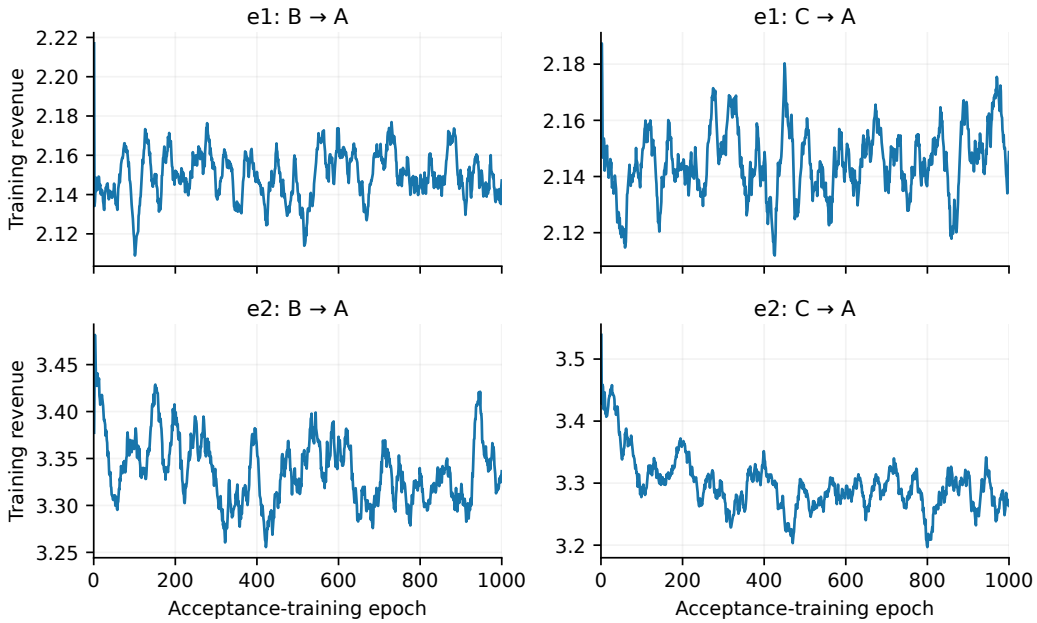


Figure 3: Acceptance-training revenue after 1,000 epochs of teacher distillation, with trailing 25-epoch averages. Each panel shows 1,000 acceptance-training epochs initialized from the selected supervised fit. These are training-batch trajectories; selected validation revenues appear in Table 7.

G Bandit Auction Experiment Details

The bandit-auction experiments assess the reliability of deviation oracles. We train a mechanism, hold it fixed, and vary the information, architecture, and optimization of the deviation policy. This separates the mechanism's response to a stronger training oracle from the strength of the subsequent evaluation searches.

G.1 Auction Environments

All four environments have two buyers, horizon $T = 5$, discount factor $\gamma = 0.9$, and allocation cost $c = 0.1$ per good. At period t, buyer i has private value $v_{i,t}$, submits report $r_{i,t}$, receives allocation $x_{i,t}$, and pays $p_{i,t}$. Her stage utility is $v_{i,t} \cdot x_{i,t} - p_{i,t}$; the principal's is $\sum_i p_{i,t} - c \sum_{i,g} x_{i,g,t}$. Values evolve with allocations, so a report affects both current payoffs and future opportunities. These experiments have no budget constraints.

Types	Goods	Report domain	Initial value distribution
Discrete	1	$\{1, 2, 3\}$	Uniform
Discrete	2	$\{1, 2, 3\}^2$	Uniform
Continuous	1	$[0, 5]$	$\mathcal{N}(2.5, 0.1^2)$
Continuous	2	$[0, 5]^2$	Independent $\mathcal{N}(2.5, 0.1^2)$ coordinates

Table 8: Bandit-auction environments. Initial draws are independent across buyers. The continuous initial sampler is Gaussian without truncation; subsequent logit transitions and the reporting policy use the stated bounds.

Discrete dynamics. The transition on the finite grid $\mathcal{V} = \{1, 2, 3\}^G$ is

$$P(v' \mid v, x) = \frac{\exp\{-\frac{1}{2}(\sum_g x_g)\|v' - v\|_1\}}{\sum_{\widetilde{v} \in \mathcal{V}} \exp\{-\frac{1}{2}(\sum_g x_g)\|\widetilde{v} - v\|_1\}}. \tag{31}$$

The simulator updates both buyers after each period. In particular, zero allocation produces a uniform next-type draw under this kernel; it does not freeze the current type.

Continuous dynamics. Write $s_{i,g,t} = \mathrm{logit}(v_{i,g,t}/5)$, clipping the normalized value to $[10^{-6}, 1 - 10^{-6}]$. Set $a_{i,g,t} = \frac{1}{2}\sum_{s \le t} x_{i,g,s}$, including the current allocation. Then $s_{i,g,t+1} = s_{i,g,t} + \sigma_{i,g,t}\epsilon_{i,g,t}$ and $v_{i,g,t+1} = 5\,\mathrm{sigmoid}(s_{i,g,t+1})$, with independent standard Gaussian $\epsilon_{i,g,t}$ and $\sigma_{i,g,t} = (1 - x_{i,g,t})\sigma_{\mathrm{no}} + x_{i,g,t}/[\lambda(1 + a_{i,g,t})]$. Both continuous configurations use $\sigma_{\mathrm{no}} = 0$ and $\lambda = 3$.

G.2 Mechanism Network

All four environments use the same mechanism architecture, with weights shared across buyers. We use the network notation of Section F.3: $\mathrm{MLP}_{d_0 \to \cdots \to d_L}$ has ReLU hidden layers and an affine output; a superscript $+$ adds a final ReLU. Brackets with $\|$ denote concatenation. Let G be the number of goods and $\tau_t = (t-1)/(T-1)$. The mechanism receives reports and recorded outcomes, without hidden private types.

- **Current-report encoder.** Append normalized time to each buyer's report:

$$h_{\mathrm{rep},i,t} = \mathrm{MLP}^{+}_{G+1 \to 128 \to 64}([r_{i,t} \| \tau_t]) \in \mathbb{R}^{64}.$$

- **Allocation-history encoder.** Encode cumulative allocations by good:

$$h_{\mathrm{alloc},i,t} = \mathrm{MLP}^{+}_{G \to 96 \to 48}\left(\sum_{s<t} x_{i,s}\right) \in \mathbb{R}^{48}.$$

- **Report-history encoder.** Let $\widetilde{r}_{i,s} = r_{i,s}$ for $s < t$ and zero otherwise, for $s = 1, \ldots, T$. The attention pool is

$$e_{i,s} = \mathrm{ReLU}(W_r \widetilde{r}_{i,s}) \in \mathbb{R}^{32}, \quad w_{i,s} = \frac{\exp(a^\top e_{i,s})}{\sum_{u=1}^{T} \exp(a^\top e_{i,u})}, \quad h_{\mathrm{hist},i,t} = \sum_{s=1}^{T} w_{i,s} e_{i,s} \in \mathbb{R}^{32}.$$

  The projection and scoring map have no biases. The normalization includes the zero-padded slots, and tokens have no added time coordinate.

- **Combined feature network.** Concatenate the three encodings:

$$q_{i,t} = [h_{\mathrm{rep},i,t} \| h_{\mathrm{alloc},i,t} \| h_{\mathrm{hist},i,t}] \in \mathbb{R}^{144}, \qquad h_{i,t} = \mathrm{MLP}^{+}_{144\to128\to64}(q_{i,t}) \in \mathbb{R}^{64}.$$

- **Allocation and payment heads.** Let $\ell_{i,t} = W_x h_{i,t} + b_x \in \mathbb{R}^G$ and let $\ell_{0,g}$ be a learned no-allocation logit. Then

$$x_{i,g,t} = \frac{\exp(\ell_{i,g,t})}{\exp(\ell_{0,g}) + \sum_{j\in[n]} \exp(\ell_{j,g,t})}, \qquad p_{i,t} = \mathrm{ReLU}(\mathrm{MLP}_{64\to64\to32\to1}(h_{i,t}))\,.$$

  Thus allocations are feasible and payments are nonnegative. Individual rationality is handled by the training objective.

G.3 Deviation-Oracle Axes and Network Architectures

We vary the oracle's information, neural architecture, and optimization method. The information specifies the available observations, the architecture maps them to reports, and the optimization method searches over the policy parameters.

G.3.1 Information Representations

For discrete types, let $b_{i,t} \in \Delta((\{1,2,3\}^G)^n)$ denote buyer i's belief over the joint type profile, with her own type fixed at its observed value. Set $c^{\mathrm{disc}}_{i,t} = [v_{i,t} \| b_{i,t}] \in \mathbb{R}^{G+K^n}$, where $K = 3^G$. The history tokens are

$$y^{\mathrm{disc,M1}}_{i,s} = \mathrm{vec}((x_{j,s})_{j\in[n]}) \in \mathbb{R}^{nG},$$
$$y^{\mathrm{disc,M2}}_{i,s} = [r_{i,s} \| \mathrm{vec}((x_{j,s})_{j\in[n]})] \in \mathbb{R}^{G+nG}.$$

Only $s < t$ is observed; all other slots are zero-padded to length T. The filter propagates the opponent's distribution using the known transition kernel and observed allocation, conditioning on reconstructed opponent reports. The information-access verification is described in Section G.6. The implementation names `m15` and `infostate` correspond to paper `M1` and `M2`, respectively.

For continuous types, the raw-history encoders use

$$y^{\mathrm{cont,M1}}_{i,s} = [r_{i,s} \| \mathrm{vec}((x_{j,s})_{j\in[n]})] \in \mathbb{R}^{G+nG},$$
$$y^{\mathrm{cont,M2}}_{i,s} = [r_{i,s} \| \mathrm{vec}((x_{j,s})_{j\in[n]}) \| p_{i,s}] \in \mathbb{R}^{G+nG+1},$$

with the same padding convention. Recurrent features additionally encode past private types. Thus, `M1`/`M2` specify the information change within an encoder family: own reports for discrete types and own payments for continuous types.

G.3.2 Deviation-Policy Networks

We first define the feature maps, then their common MLP and reporting heads. All maps below are shared across buyers. Let $Y_{i,t}$ denote the length-T token sequence from the chosen information representation.

Base/MLP and Wide. For discrete types, both variants flatten the available histories, with input $f_{i,t} = [v_{i,t} \| b_{i,t} \| \mathrm{vec}(Y_{i,t})]$. In the flattened implementation, own-report and public-allocation histories are concatenated in separate temporal blocks. For continuous `MLP`, $f_{i,t} = [v_{i,t} \| \mathrm{vec}(Y_{i,t})]$; continuous `Wide` instead uses the recurrent features defined below. If $d_f = \dim(f_{i,t})$, the reporting-head features are

$$z_{i,t} = \mathrm{MLP}^{+}_{d_f\to h\to h}(f_{i,t}) \qquad \text{for Base/MLP, Attn, and GRU,}$$
$$z_{i,t} = \mathrm{MLP}^{+}_{d_f\to 4h\to 4h\to 4h}(f_{i,t}) \qquad \text{for Wide.}$$

Discrete PW and ES use $h = 128$, while discrete NPG uses $h = 32$. Continuous searches use $h = 128$. The discrete Base NPG implementation uses a separate categorical head: the current type is one-hot encoded and concatenated with the joint-profile belief and the specified flattened histories, giving $f^{\mathrm{NPG}}_{i,t}$. Its logits are $L_{i,t} = W_2 \tanh(W_1 f^{\mathrm{NPG}}_{i,t} + b_1) + b_2$, with a single hidden layer of width 32 and no truthful-prototype term. The other discrete NPG architectures use the encoders and prototype head defined here.

Attention encoder. For discrete types the query context is $c_{i,t}^{\mathrm{disc}}$; for continuous types it is $c_{i,t}^{\mathrm{cont}} = [v_{i,t} \| \mathrm{vec}((\sum_{s<t} x_{j,s})_{j\in[n]})]$. Define

$$q_{i,t} = W_q c_{i,t} + b_q, \qquad k_{i,s} = W_k y_{i,s} + b_k, \qquad u_{i,s} = W_u y_{i,s} + b_u,$$
$$w_{i,s} = \frac{\exp(q_{i,t}^\top k_{i,s}/\sqrt{h})}{\sum_{u=1}^{T} \exp(q_{i,t}^\top k_{i,u}/\sqrt{h})}, \qquad a_{i,t} = \sum_{s=1}^{T} w_{i,s} u_{i,s} \in \mathbb{R}^h.$$

The query, key, and value vectors have dimension h. Attention is normalized over all padded slots. The MLP input is $[v_{i,t} \| b_{i,t} \| a_{i,t}]$ for discrete types and $[v_{i,t} \| a_{i,t}]$ for continuous types.

Discrete recurrent encoder. A trainable width-h GRU processes $Y_{i,t}$ from an all-zero hidden state, including its padded slots. Its final state $a_{i,t} \in \mathbb{R}^h$ gives the MLP input $f_{i,t} = [v_{i,t} \| b_{i,t} \| a_{i,t}]$. The GRU and reporting head are optimized jointly.

Continuous recurrent features. The public and private feature generators have dimensions 64 and 128. Their weights are randomly initialized and held fixed during mechanism training and oracle search. On completing period t, they update using

$$\begin{aligned}
o_t^{\mathrm{pub}} &= [\tau_t \| \mathrm{vec}((x_{j,t})_{j\in[n]})], \\
c_{t+1} &= \mathrm{GRU}_{64}(\mathrm{MLP}^+_{1+nG\to 64}(o_t^{\mathrm{pub}}), c_t), \\
o_{i,t}^{\mathrm{priv}} &= [\tau_t \| v_{i,t+1} \| r_{i,t}], \qquad \widetilde{c}_{t+1} = \mathrm{MLP}^+_{64\to 128}(c_{t+1}), \\
H_{i,t+1} &= \mathrm{GRU}_{128}\Big(\mathrm{MLP}^+_{1+2G+128\to 128}([o_{i,t}^{\mathrm{priv}} \| \widetilde{c}_{t+1}]), H_{i,t}\Big).
\end{aligned}$$

The first G coordinates of $H_{i,t+1}$ are then set to $v_{i,t+1}$. Initially $c_1 = 0$; the private encoder is initialized from the current type, zero previous report, and zero public state. A separately trainable affine projection $P : \mathbb{R}^{64} \to \mathbb{R}^{128}$, initialized to zero, gives the `M1` input $f_{i,t} = [H_{i,t} \| P(c_t)] \in \mathbb{R}^{256}$. Under `M2`, append $\mathrm{MLP}^+_{T\to h}((p_{i,s})_{s<t})$, with own payments zero-padded to the horizon. `GRU` and `Wide` use these same features and differ in the subsequent MLP widths.

Discrete reporting head. For report-grid points $v^{(1)}, \ldots, v^{(K)}$, the logits are

$$L_{i,t,k} = \kappa \mathbf{1}\{v_{i,t} = v^{(k)}\} + (W_{\mathrm{out}} z_{i,t} + b_{\mathrm{out}})_k, \qquad \kappa = \mathrm{softplus}(\kappa_{\mathrm{raw}}).$$

The truthful-prototype strength is initialized at 3. Pathwise training samples reports with unit-temperature straight-through Gumbel-softmax: the forward report is a grid point, while the backward pass uses soft probabilities. Deterministic evaluation takes the largest logit; NPG optimizes the corresponding categorical policy.

Continuous reporting head. An affine output produces $\delta_{i,t} = W_{\mathrm{out}} z_{i,t} + b_{\mathrm{out}} \in \mathbb{R}^G$. For each coordinate,

$$p^0_{i,g,t} = \mathrm{clip}(v_{i,g,t}/5, 10^{-6}, 1 - 10^{-6}), \qquad r_{i,g,t} = 5\,\mathrm{sigmoid}\big(\mathrm{logit}(p^0_{i,g,t}) + \delta_{i,g,t}\big).$$

This keeps reports in $[0, 5]$. The original pathwise implementation treats the truthful component p^0 as constant when differentiating. NPG adds Gaussian noise to the logit residual for score-function optimization.

G.3.3 Optimization Methods

All four methods optimize unilateral payoff gains with the mechanism fixed and other buyers truthful. We use multiple initializations and retain the policy with the largest evaluated gain. Their update rules differ as follows.

- **Pathwise gradients (PW).** Differentiate sampled discounted payoffs through the implemented reporting and rollout maps, and apply Adam ascent to the policy parameters. Discrete reports use the straight-through Gumbel-softmax rule described above; continuous reports use the bounded logit-residual map. After optimization, the discrete policy is evaluated using its largest-logit report, and the continuous policy using its deterministic report.

- **Natural policy gradients (NPG).** Sample reports from a categorical policy for discrete types and from a Gaussian policy in logit space for continuous types. Let w denote the policy parameters. A score-function gradient estimate $\widehat{g}$ and empirical Fisher matrix $\widehat{F}$ give the ascent direction $d = (\widehat{F} + \lambda I)^{-1}\widehat{g}$ and update $w \leftarrow w + \eta d$. Discrete searches weight per-period action scores by discounted return-to-go, centered across trajectories at each period; their Fisher estimate averages outer products of per-period scores. Continuous searches weight trajectory scores by centered total discounted gains and form the Fisher matrix from those trajectory scores. Sampled reports are held fixed in score differentiation.

  The discrete Base policy uses a direct regularized Fisher solve, damping $\lambda = 0.001$, and step size 0.3. The other discrete architectures and all continuous searches use the Woodbury identity to solve the regularized system through the score matrix, with damping 0.1. Their step size is $\min\{1, \sqrt{2\delta/\max\{\widehat{g}^\top d, 10^{-12}\}}\}$, with $\delta = 0.02$. Continuous logit noise has standard deviation 0.3. Final evaluation uses the deterministic largest-logit or Gaussian-mean policy.

- **Larger-budget NPG (NPG+).** Use the same policy-gradient construction and update rules as NPG, but increase the optimization steps, restarts, and/or trajectories per step. This tests whether additional search computation reveals deviations missed by the standard budget. The allocation of this additional computation depends on the setting and architecture; Table 10 specifies it.

- **Evolution strategies (ES).** Draw parameter perturbations $\epsilon_k \sim \mathcal{N}(0, I)$ and evaluate the paired policies $w \pm \sigma\epsilon_k$, using common trajectory randomness across the population. Rank the resulting payoff gains and rescale the ranks to $[-1/2, 1/2]$. The weighted perturbation average gives an ascent direction, which is applied with Adam. This search uses payoff evaluations without differentiating the reporting rule or simulator. We periodically evaluate the unperturbed policy and retain the best evaluated parameters across iterations and restarts.

Optimization steps, restart counts, and trajectory budgets are given in Section G.4.2.

G.4 Training and Evaluation Configuration

G.4.1 Training and Evaluation of Baseline Mechanisms

We use acceptance training with `M1` deviation policies and pathwise updates. For discrete types, the policy is the base flattened-history MLP; for continuous types, it is an MLP reporting head on fixed GRU features. The continuous residual head has initialization standard deviation 1, as specified in the saved configurations. The discrete head uses truthful-prototype logits of initial strength 3 and a learned residual; sampled reports use straight-through Gumbel-softmax. Thus, discrete pathwise training differentiates a surrogate for the categorical sampling operation. The retrained discrete two-good mechanism changes the training oracle from `M1` to `M2`, retaining the base architecture, pathwise optimization, and the other training controls.

Training parameter	Discrete one good	Discrete two goods	Continuous one good	Continuous two goods
Outer epochs	2,000	2,000	7,500	5,000
Inner steps	500	500	400	400
Batch trajectories	64	64	64	64
Mechanism learning rate	10^{-4}	10^{-4}	5×10^{-4}	5×10^{-4}
Policy learning rate	0.005	0.005	5×10^{-4}	5×10^{-4}
DIC weight	5	5	5	5
DIR weight	5	5	1	1
Profit target	7.617	15.25	9.2	12.0
Policy hidden width	128	128	128	128

Table 9: Saved bandit training configurations. All use seed 42, Adam updates, one outer step per epoch, and checkpoint intervals of 50 epochs. Discrete training uses three inner restarts. The `M2`-trained two-good mechanism uses the same column as its `M1`-trained counterpart.

Table 9 summarizes the training settings. The selected discrete checkpoints are at epochs 1,600 (one good), 1,950 (two goods, `M1`), and 1,800 (two goods, `M2`).

Baseline evaluation. The discrete checkpoint-selection procedure reoptimizes the base pathwise oracle for 600 steps with three restarts, learning rate 5×10^{-4}, and batches of 128 trajectories; evaluation uses 32 batches of 256 trajectories. Its selection gate uses the average per-buyer DIC estimate and a trajectory-level negative-utility diagnostic, each with tolerance 0.1. The continuous comparisons retain the saved best checkpoint from the corresponding training run.

For the ex-ante DIR results reported here, we separately evaluate each frozen mechanism on 64 independent batches of 1,024 truthful trajectories. We estimate each buyer's expected discounted utility and then compute $\sum_{i\in[n]} \max\{0, -\widehat{U_i(\boldsymbol{\pi}^\dagger; M^{\boldsymbol{\theta}})}\}$. The hinge is applied after averaging. Standard errors are computed from the 64 batch means. This forward evaluation leaves both mechanism and policy parameters fixed.

G.4.2 Training and Evaluation against Alternative Deviation Oracles

Oracle optimization on a fixed mechanism. For each frozen mechanism, we optimize deviation policies across the information–architecture–optimization grid, holding other buyers truthful. Multiple initializations produce candidate policies, which are compared using the search driver's evaluation samples. Table 10 gives the principal search controls; the released per-cell configurations specify the evaluation sample sizes and initialization settings. The grid contains different amounts of search computation.

Search	Discrete	Continuous
PW	1,500 steps, 8 restarts	1,500 steps, 8 restarts
NPG	300 steps, 6 restarts, 512 trajectories/step; Base step size 0.3, damping 0.001	300 steps, 6 restarts, 512 trajectories/step; step size 1, damping 0.1
NPG+	Base: 500 steps, 1,024 trajectories/step; 24 restarts for original two-good mechanism, 12 for retrained mechanism. Other encoders: 400 steps, 6 restarts, 512 trajectories/step	600 steps, 12 restarts, 1,024 trajectories/step
ES	300 generations, 4 restarts, 64 antithetic pairs, perturbation scale 0.05, step size 0.02, 128 trajectories per fitness estimate	

Table 10: Bandit evaluation search budgets. NPG+ base evaluations for the original discrete two-good mechanism use 100,000 trajectories; other NPG evaluations use 20,000. Non-Base discrete NPG uses damping 0.1 and step-size control 0.02 as described in Section G.3.3. Continuous NPG uses report-noise scale 0.3 and the same step-size control. The grid deliberately includes different amounts of search computation.

Mechanism retraining against `M2`. For discrete two-good auctions, we additionally train the mechanism against `M2`, retaining the base MLP, pathwise optimization, and training controls in Table 9. We then freeze this mechanism and repeat the full 32-cell grid: two information representations, four architectures, and four optimization methods. This is distinct from optimizing a stronger oracle against the original fixed mechanism.

Evaluation DIC loss. All bandit tables report *evaluation DIC loss*. Let $\widehat{d}_i$ denote buyer i's estimated unilateral payoff gain relative to truthful reporting, with the other buyer truthful. Discrete NPG and NPG+ entries use $\widehat{d}_1$; discrete PW/ES and continuous entries use $(\widehat{d}_1 + \widehat{d}_2)/2$. Negative entries indicate that the evaluated deviation policy performs worse than truthful reporting. The exact finite calculation below evaluates each buyer separately. These are the bandit aggregation conventions referenced in Section 5; budget experiments instead sum positive per-buyer gains. An average below the reference tolerance alone does not certify that each buyer's gain is below it.

G.5 Experimental Results

Table 2 reports the complete information–architecture–optimization grid for the original and retrained discrete two-good mechanisms and the two continuous mechanisms. The largest saved entries are summarized in Table 11. The full 32-cell grid is available for the retrained discrete mechanism, including every `M1` and `M2` search/architecture combination. Its maximum 0.07221 occurs with `M1`,

base NPG+; the base M2 NPG+ entry is 0.06387. Thus, its improvement is supported by the entire tested grid, not just reevaluation with its training oracle. Principal payoff changes from 15.57045 to 15.52846.

Types	Goods / training information	Largest neural entry	Exact finite DP
Discrete	One / M1	0.0826	0.08702
Discrete	Two / M1	0.1082	—
Discrete	Two / M2	0.0722	—
Continuous	One / M1	0.0083	—
Continuous	Two / M1	0.0103	—

Table 11: Bandit deviation-oracle results. The continuous and two-good columns summarize saved neural searches under the reporting convention above; the single-good discrete row gives the largest gain from the two NPG comparisons described in Section G.6 and the saved exact maximum per-buyer gain in the finite model. Dashes indicate that no corresponding exact computation is reported.

Ex-ante individual rationality. The estimated ex-ante evaluation DIR loss is 0.0 for all baseline mechanisms and the M2-retrained mechanism in the discrete two-good setting. Both buyers' estimated truthful utilities have positive lower endpoints in their approximate 95% confidence intervals.

G.6 Exact Finite-Setting Calibration

For the frozen single-good mechanism, dynamic programming enumerates the finite type and report histories, takes the best current report at each legal information state, and integrates the known transition probabilities and initial prior. The comparison is between the optimal unilateral reporting policy and truthful reporting, with the other buyer truthful. The saved calculation records an ex-ante gain of 0.08701646 for each buyer, below the reference tolerance 0.1. For comparison, NPG gives gains 0.07917 with full public allocation history and 0.08259 with cumulative public allocations; both policies also receive current type, the joint-profile belief, and own-report history. These two point estimates recover approximately 91–95% of the recorded exact gain.

The information check verifies that the opponent's report can be reconstructed from the public outcomes and the buyer's own report for this saved mechanism. Exhaustive verification covers all five periods and 66,429 reachable decision cases, within the recorded numerical precision. This connects the report-based finite DP calculation to information actually available to a buyer. For the two-good mechanisms, the saved reconstruction checks cover depth three and sampled full trajectories; they do not provide the same exhaustive five-period verification. We therefore do not attach an exact finite-model certificate to their neural comparisons.

Relation to the budget study. The budget study varies principal-side history representations and studies the resulting optimization gaps. The bandit study holds learned mechanisms fixed to examine agent-side representations and deviation-oracle strength, with a separate retraining comparison showing the mechanism's response to a richer training oracle.

H Computing Resources, Software, and Code Availability

Computing resources. The bandit training and deviation-oracle experiments used a cloud GPU server equipped with eight NVIDIA L40S GPUs. The budget experiments were distributed across local and remote GPU workers, including local execution of additional distillation runs. The separate ex-ante DIR evaluation of the frozen bandit mechanisms used CPU execution. Hardware and software environments therefore differed across studies. Experiment configurations, random seeds, saved results, and reproduction scripts are provided in the accompanying code repository.

Software environment. The implementation is written in Python [55]. The numerical and differentiable-programming components use standard scientific-computing libraries, including NumPy [36], JAX [12], Haiku [38], and JAXopt [10]. Data handling, visualization, and auxiliary numerical routines use standard Python packages such as Pandas, Matplotlib, and CVXPY where appropriate. Exact package versions are documented in the code repository.

Licenses. The experiments rely on widely used open-source software. Python is distributed under the Python Software Foundation License. NumPy, Pandas, and Matplotlib are distributed under permissive BSD-style or PSF-compatible licenses. JAX, Haiku, JAXopt, and CVXPY are distributed under Apache 2.0 licenses. We use these packages as external dependencies and do not redistribute modified versions of them as part of the paper.

Code availability. Code for the experiments, including implementation files, configuration files, and reproduction scripts, is available at the code repository. The repository contains the experiment-specific settings used for the results reported in Appendices G and F.

I LLM Usage Declaration

Large language models were used as auxiliary tools during the preparation of this work. Their use included partial assistance with code implementation, as well as writing, editing, formatting, and proofreading portions of the manuscript, including parts of the theoretical exposition. All code produced or modified with LLM assistance was reviewed, tested, and checked carefully by the authors. All mathematical statements, proofs, experimental designs, results, and conclusions were verified by the authors, who take full responsibility for the content of the paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope: a parameterized bilevel framework for computational dynamic mechanism design, convergence guarantees under approximate deviation oracles, and experiments in dynamic auction environments. The claims are supported by the formulation in Sections 2–4 and the experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the computational guarantees, including local stationarity rather than global optimality, dependence on oracle accuracy, and approximation error from the chosen mechanism and policy classes; see Section 4 and Appendix D. The experiments also discuss finite-sample, optimization, and representation effects in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical assumptions are stated in Section 4 and Appendix E, including regularity, stochastic-gradient, sufficient-statistic, and stochastic NPG assumptions. Full proofs of the main convergence and oracle-accuracy results are provided in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the experimental environments, architectures, training protocols, evaluation procedures, and hyperparameters in Section 5 and Appendices G–F. Code and reproduction scripts are provided in the code repository described in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use simulated dynamic auction environments rather than external datasets, and the code is available at the code repository linked in Section 5 and Appendix H. The repository contains implementation files, configuration files, and scripts for reproducing the reported results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main experimental setup is summarized in Section 5, with full environment specifications, mechanism and policy architectures, training configurations, hyperparameters, and evaluation protocols given in Appendices G and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The experimental tables report mean performance with error bars for profit, evaluation IC loss, and evaluation IR loss where applicable. The evaluation protocol uses fresh random keys and samples, as described in Appendices G.4.1 and F.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix H describes the study-specific compute environments: a cloud server equipped with eight NVIDIA L40S GPUs for bandit training and deviation searches, local and remote GPU workers for budget experiments, and CPU execution for the separate ex-ante DIR evaluation. Experiment configurations and reproduction scripts are provided in the code repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and computational, uses simulated auction environments, and does not involve human subjects, private data, scraped datasets, or deployed decision systems. The authors have reviewed the NeurIPS Code of Ethics and are not aware of any deviation from it.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper studies foundational methods for designing dynamic mechanisms, with potential positive impact through better tools for analyzing incentive systems and potential risks if such tools are applied without appropriate domain-specific safeguards. The scope and limitations of the learned mechanisms are discussed in Sections 4 and 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, pretrained generative models, or systems intended for direct deployment. The released assets are code and simulated experiment configurations for reproducing the paper's computational results.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix H lists the main software dependencies and their licenses, including Python, NumPy, Pandas, Matplotlib, JAX, Haiku, JAXopt, and CVXPY. The paper uses these packages as external dependencies and does not redistribute modified versions of them.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases code and simulated experiment configurations as new research assets through the code repository described in Appendix H. The repository documents the implementation and reproduction scripts for the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, user studies, surveys, or experiments with human subjects. All experiments are conducted in simulated dynamic auction environments.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, crowdsourcing, or collection of personal data, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Appendix I declares that LLMs were used as auxiliary tools for partial code implementation and for writing, editing, formatting, and proofreading. The authors reviewed and verified all code, mathematical statements, proofs, experiments, results, and conclusions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.