Collaborative Learning under Strategic Behavior: Mechanisms for Eliciting Feedback in Principal-Agent Bandit Games

Ramakrishnan Krishnamurthy, Arpit Agarwal, Lakshmi Subramanian, Maximilian Nickel

Keywords: Non-cooperative Game Theory, Multi-Armed Bandits, Multi-Agent Learning, Equilibrium behaviour

Summary

Collaborative multi-armed bandits (MAB) has emerged as a promising framework that allows multiple agents to share the *burden of exploration* and find optimal actions for a shared task. While several recent results demonstrate the benefit of collaboration in minimizing per-agent regret, prior works on collaborative MABs implicitly assume that all participating agents behave truthfully. However, the case of *strategic* agent behavior—where an agent may *free-ride* on the information shared by others without performing exploration—has received limited attention. Such free-riding strategies can lead to a collapse in exploration, resulting in a high regret for all honest agents. Our paper addresses this problem of collaborative multi-armed bandits in the presence of strategic agent behavior.

Contribution(s)

1. Our main contribution is to design mechanisms for the principal and arm-playing algorithms for the agents so that truthful behavior, i.e., performing sufficient exploration and reporting feedback accurately, is a Nash equilibrium.
   Context: To the best of our knowledge, this is the first mechanism for collaborative multi-agent bandits that provably disincentivizes free-riding while preserving optimal collaborative regret guarantees under strategic behaviour of the agents.
2. Under this Nash equilibrium, the per-agent regret with collaboration is lower by a factor of $\sqrt{M}$ compared to the non-collaborative case, where M is the number of agents.
   Context: As $\sqrt{M}$ is the best possible factor of regret reduction even when agents are altruistic/non-strategic, our results establish that it is possible to achieve the benefit of altruistic collaboration even in the presence of strategic agents who may want to free-ride.
3. For the single communication-event setting, we exhaustively characterize the set of all Nash equilibria induced by our protocols and show that no equilibrium permits free-riding.
   Context: This characterization rules out unintended equilibria involving free-riding and ensures that dishonest behaviour cannot arise as part of any Nash equilibrium under our mechanism.
4. Our experiments empirically validate our theoretical findings. A free-riding agent enjoys lower regret, but suffers significantly higher penalty, thus disincentivizing him from free-riding.
   Context: Not applicable.

Collaborative Learning under Strategic Behavior: Mechanisms for Eliciting Feedback in Principal-Agent Bandit Games

Ramakrishnan Krishnamurthy[1], **Arpit Agarwal**[2], **Lakshmi Subramanian**[1], **Maximilian Nickel**[3]
rk4312@nyu.edu, aarpit@cse.iitb.ac.in, lakshmi@nyu.edu, maxn@meta.com

[1]**Courant Institute of Mathematical Sciences, New York University**
[2]**Indian Institute of Technology Bombay** [3]**FAIR, Meta**

Abstract

Collaborative multi-armed bandits (MAB) has emerged as a promising framework that allows multiple agents to share the *burden of exploration* and find optimal actions for a shared task. While several recent results demonstrate the benefit of collaboration in minimizing per-agent regret, prior works on collaborative MABs implicitly assume that all participating agents behave truthfully. However, the case of *strategic* agent behavior—where an agent may *free-ride* on the information shared by others without performing exploration—has received limited attention. Such free-riding strategies can lead to a collapse in exploration, resulting in a high regret for all honest agents.

This paper addresses the problem of collaborative multi-armed bandits in the presence of strategic agent behavior. We design a statistical penalty mechanism that incentivizes truthful behavior—namely, performing sufficient exploration and accurately reporting feedback—by making such behavior a Nash equilibrium. Furthermore, under this Nash equilibrium, the per-agent regret with collaboration is smaller by a factor of $\sqrt{M}$ compared to the non-collaborative case, where M is the number of agents. Our results show that near-optimal collaborative regret can be retained even in the presence of strategic agents who may otherwise be incentivized to free-ride. Our experiments empirically validate our theoretical findings.

1 Introduction

The multi-armed bandits (MAB) framework models scenarios where an agent repeatedly faces the *explore-exploit* dilemma—whether to gather new information or act optimally based on current knowledge (Lattimore & Szepesvári, 2020). This framework has broad applications, including clinical trials (Thompson, 1933; Villar et al., 2015), recommendation systems (Li et al., 2010), advertising (Buccapatnam et al., 2013), dynamic pricing (Trovò et al., 2018), portfolio selection (Huo & Fu, 2017), hedging (Cannelli et al., 2023), auctions (Sharma et al., 2012), and resource allocation (Krishnasamy et al., 2021).

Recent interest has surged in *collaborative* MABs (Dubey & Pentland, 2020; Wang et al., 2019; Li & Wang, 2022; He et al., 2022) driven by the potential of data sharing and interoperability to boost both individual and societal outcomes (Mancini, 2021). In this setting, multiple agents jointly solve a bandit problem by sharing information, allowing them to distribute the *burden of exploration* and act based on the collective information. Practical examples include hospitals collaborating in vaccine trials (Rieke et al., 2020; Flores et al., 2021), NGOs coordinating mobile interventions

for healthcare (Ou et al., 2022), cities tackling common issues like waste management (Mao & Perrault, 2024), autonomous vehicles jointly learning safe routes (Ferdowsi et al., 2019), and content recommendation platforms sharing aggregate statistics to enhance user personalization (Bouneffouf et al., 2020). Prior work shows such collaboration significantly reduces per-agent regret (Wang et al., 2019).

However, these benefits often rely on the unrealistic assumption that all agents truthfully share information and contribute to exploration. This assumption is arguably idealistic and can be easily violated in practice due to the presence of self-interested *strategic* agents who might be tempted to 'free-ride' on the information shared by others. For instance, Jung et al. (2020) showed that an agent observing the actions and rewards of another regret-minimizing agent can adopt a 'free-riding' strategy which suffers negligible regret. Hence, as many agents may be tempted to adopt such free-riding strategies, this behavior can ultimately lead to a collapse of exploration and the breakdown of collaborative regret minimization.

Despite its importance, the issue of strategic 'free-riding' remains underexplored in collaborative MABs. This paper addresses the question:

> **Research Question.** *Is it possible to incentivize strategic agents to truthfully participate—and avoid "free-riding"—in collaborative multi-armed bandits?*

To study this, we combine insights from two areas of research: mechanism design and multi-agent MABs. We propose mechanisms that a central principal can use to incentivize truthful behaviour in a linear stochastic MAB setting with M agents and time horizon T. Each agent's aim is to minimize their own regret on the bandit instance. We use monetary penalty as the main device by which the principal can incentivize truthful behavior. The principal acts as a central coordinator who collects (and shares) information from (to) agents and also enforces the mechanism to penalize the agents. Importantly, the principal does not directly observe the actions or rewards collected by agents and must instead infer deviations statistically from reported estimates. The principal's objective is to minimize the social regret—the sum of all agents' regrets. Communication occurs in synchronized rounds between principal and agents.

1.1 Our contributions

We design two provable mechanisms: (i) LsHon (**L**ow-**S**witching **Hon**est protocol) for the agents, and (ii) MonPen (**Mon**etary **Pen**alty protocol) for the principal. Together, these mechanisms enforce truthful behaviour—namely, diligently performing the prescribed exploration and accurately reporting the information required by LsHon. Our main contributions are as follows.

A. Incentive-compatible collaboration. When the principal follows MonPen under common knowledge, it is a Nash equilibrium for all agents to follow LsHon (Theorem 1). To the best of our knowledge, this is the first mechanism that provably disincentivizes 'free-riding' in collaborative multi-armed bandit settings. Because LsHon prescribes truthful behaviour and because honesty constitutes a Nash equilibrium under the maintained assumptions, the induced protocol pairing is Bayesian-Nash incentive compatible (BNIC).

This guarantee relies on a key design principle in MonPen: the principal performs carefully constructed statistical manipulations—despite not knowing the true model parameter—and applies a penalty function whose expected value converges for honest agents but diverges for dishonest or free-riding agents.

B. Optimal regret under strategic behaviour. Under the equilibrium where all agents follow LsHon, every agent achieves an order-wise optimal regret of $\tilde{O}(d\sqrt{T/M})$ (Theorem 2) while incurring an arbitrarily small (non-negative) expected penalty. Here, d denotes the ambient dimension of the action space. Compared to the per-agent regret of $\tilde{O}(d\sqrt{T})$ achievable without collaboration (Abbasi-Yadkori et al., 2011), this represents a $\sqrt{M}$-fold reduction in per-agent regret—even

when agents behave strategically. Achieving this guarantee requires LsHon to employ a *low-switching* bandit policy (Abbasi-Yadkori et al., 2011), which enables synchronized arm selections across agents and facilitates efficient information sharing.

C. Characterization of all Nash equilibria in the single communication-event case. Beyond establishing the truthful equilibrium, we characterize the set of all possible Nash equilibria induced by MonPen for a single communication event (Theorem 3). We show that a strategy profile forms a Nash equilibrium only if all agents apply the same/identical bias to their reported estimates while not free-riding on exploration. This characterization rules out equilibria involving free-riding, although coordinated equilibria involving a shared bias remain possible.

D. Empirical validation. We complement our theoretical results with experiments in a semi-synthetic environment derived from the MovieLens dataset (Harper & Konstan, 2015). Our experiments show that although free-riding agents may temporarily achieve lower regret (higher utility) than truthful agents, the penalties imposed by our mechanism ultimately reduce their overall utility below that of honest agents, thereby discouraging dishonest behaviour.

1.2 Related Work

Stochastic linear bandits. The stochastic multi-armed bandits (MAB) framework models the exploration-exploitation trade-off in sequential decision-making. In linear bandits, actions are represented as vectors in a d-dimensional space, and expected rewards are a linear function of these embeddings. This setting is well-suited to problems with contextual or side information (Chu et al., 2011; Wang et al., 2021; Moradipari et al., 2021). A standard solution approach is the 'Optimism in the Face Of Uncertainty' paradigm (Auer et al., 2002). For linear bandits, the OFUL algorithm (a.k.a. 'LinUCB') (Abbasi-Yadkori et al., 2011) achieves a regret of $\tilde{O}(d\sqrt{T})$ even when actions sets changes over time horizon T. It is optimal (up to log factors) as it matches the $\Omega(d\sqrt{T})$ lower bound (Dani et al., 2008).

More recently, linear bandits have been studied in multi-agent settings where agents communicate via a central principal (Hillel et al., 2013; Tao et al., 2019). Wang et al. (2019) introduced DisLinUCB, a multi-agent extension of LinUCB, which achieves a $\tilde{O}(d\sqrt{MT})$ social regret—i.e., a per-agent improvement by a factor of $1/\sqrt{M}$ over single-agent settings. However, these works assume all agents are truthful, and do not account for strategic behaviour such as free-riding. We juxtapose our work against these in Table 1.

Table 1: Contrast of our work with the literature.

Work	Regret-optimal?	Collaborative?	Strategic agents?
Abbasi-Yadkori et al. (2011)	✓	✗	✗
Wang et al. (2019)	✓	✓	✗
Our Work	✓	✓	✓

Strategic agents. A growing body of work addresses strategic behavior in collaborative MABs. Free-riding arises when an agent benefits from another agent's exploration or shared information without contributing back. Jung et al. (2020) show that an agent observing other no-regret learners can attain $O(1)$ regret (in T, but with an $1/\Delta$ factor otherwise) by estimating rewards solely from observations shared by other agents. This challenge intensifies when data sharing incurs costs—as studied in linear bandits (Wei et al., 2024) and supervised learning (Karimireddy et al., 2022; Murhekar et al., 2024). In such cases, agents may opt out unless incentivized. A common approach is to design mechanisms offering payments that offset agents' costs and rewards their contributions; for example, Krishnamurthy et al. (2026) study how learning algorithms should behave to offer fair payouts to the agents.

Our work differs from these approaches in two important ways. First, we study incentives for truthful exploration rather than merely participation or communication. Second, we establish regret guarantees at equilibrium and characterize strategic behavior induced by the mechanism.

For the interested reader, we also provide a more detailed literature survey in Appendix B.

2 Setting & Preliminaries

Communication setup. We consider M agents and a single trusted principal. The communication network between these agents and principal is set up according to a star topology. More specifically, every agent can communicate with the principal by sending and receiving packets, but the agents cannot communicate amongst themselves directly. The communication happens synchronously over rounds, i.e., the principal will coordinate upon a time-step at which all agents will send packets to the principal, and after receiving all packets, the principal will then respond to each agent. We denote the sequence of time-steps when communication events happen by $\mathcal{T} = (\tau_1, \tau_2, \ldots, \tau_q)$. The total communication cost is the total number of real numbers sent/received by the principal across all the communication events.

Multi-agent linear bandit setting. Each agent $a \in \mathcal{A}$ is faced with the same linear bandit instance with time-horizon T. At each trial/time-step $t \in [T]$, the agent needs to choose an action $x_{a,t}$ from the decision set $\mathcal{X}_t \subseteq \mathbb{R}^d$ that spans $\mathbb{R}^d$. Note that $\mathcal{X}_t$ is identical across all agents, but can vary over time. Upon playing action $x_{a,t}$, the agent observes its reward $y_{a,t} = \langle \theta^*, x_{a,t} \rangle + \eta_{a,t}$, where $\theta^* \in \mathbb{R}^d$ is an unknown parameter (and no agent has auxiliary information, such as a prior, about θ^*) and $\eta_{a,t} \sim \mathcal{N}(0, 1)$ is zero-mean i.i.d. random noise. Given knowledge of θ^*, the optimal action at trial t is $x_t^* = \arg\max_{x \in \mathcal{X}_t} \langle \theta^*, x \rangle$ and the expected cumulative reward of such optimal actions is given by $\sum_{t=1}^{T} \langle \theta^*, x_t^* \rangle$. We will measure the *regret* of the agent against this optimal strategy, i.e.,[1]

$$R_a(T) = \sum_{t=1}^{T} \langle \theta^*, x_t^* \rangle - \langle \theta^*, x_{a,t} \rangle . \tag{1}$$

To analyze regret, we make standard assumptions: $\|\theta^*\|_2 \leq 1$ and $\|x\|_2 \leq 1$ for all $x \in \mathcal{X}_t$ for all t.

We next introduce some quantities that are typical to the LinUCB algorithmic approach (see, e.g., Lattimore & Szepesvári (2020)). Each agent a maintains statistics of arms played and rewards observed by him in two quantities: the 'design' matrix $V_{a,t}^s := \sum_{q=1}^{t} x_{a,q} x_{a,q}^\top$, and the 'results' matrix $B_{a,t}^s := \sum_{q=1}^{t} x_{a,q} y_{a,q}$, superscript s denotes 'self-play'. Additionally, denote by $V_{a,t}^c := \sum_{b \in \mathcal{A} \setminus \{a\}} V_{b,t}^s$ and $B_{a,t}^c := \sum_{b \in \mathcal{A} \setminus \{a\}} B_{b,t}^s$ the design and result matrices from statistics of all other agents put together, the superscript c denotes 'collaboration'. Further, let $\theta_{a,t}$ be the estimate (to be elaborated in Section 3.1) of the unknown θ^* that agent a computes at time t.

Strategic Agent. A rational agent might not truthfully follow the arm play algorithm prescribed by the principal and instead devise their own strategy to 'free-ride' (see Definition 1) on the other agents' statistics provided to him by the principal. Write $V_{a,\tau}^s$ to be the design matrix that agent a will have had at time τ if he completely adheres to arm playing algorithm of LsHon. But, by deviating from this algorithm to free-ride, he would end up with another design matrix $V_{a,\tau}^{s,\varepsilon} = V_{a,\tau}^s + \varepsilon_{a,\tau}^V$, where $\varepsilon_{a,\tau}^V \in \mathbb{R}^{d \times d}$ quantifies this strategic deviation.

Definition 1 (Free-rider)**.** *A rational agent $a \in \mathcal{A}$ is said to* free-ride *if, by some communication event $\tau \in [T]$, he explores not as much as* LsHon *requires him to, i.e., $V_{a,\tau}^{s,\varepsilon} \prec V_{a,\tau}^s$. Equivalently, an agent with a deviation $\varepsilon_{a,\tau}^V \prec 0$ that is negative definite is a free-rider.*

While this work primarily studies the aspects of agents free-riding (deviating design matrices), we also consider it possible for agents to misreport (bias) their parameter estimates, albeit with some restrictions captured in Assumption 2. So, the agent a reports his estimate as $\theta_{a,\tau}^\varepsilon = \theta_{a,\tau} + \varepsilon_{a,\tau}^\theta$, with a bias $\varepsilon_{a,\tau}^\theta \in \mathbb{R}^d$.

Assumption 2 (Agent knowledge and behaviour)**.** *At any communication step τ, the knowledge and behaviours of all agents $a \in \mathcal{A}$ are constrained by the following:*

1. *The agent does not have any prior/auxiliary information about the linear bandit model parameter $\theta^* \in \mathbb{R}^d$.*

[1] This notion of regret is sometimes referred to as *pseudo-regret* in the literature.

2. *The agent shall compute his estimate $\theta_{a,\tau}$ using all samples from self-play $B^s_{a,\tau}$ until the current time-step and samples from collaboration $B^c_{a,\tau'}$ until the preceding communication event $\tau' < \tau$, and (as a consequence of Item 1) not from any other information .*
3. *To the estimate $\theta_{a,\tau}$, the bias added by an agent $\varepsilon^\theta_{a,\tau} \in \mathbb{R}^d$ is some constant value independent of quantities used in Item 2.*

Assumption 2 restricts the strategic behavior available to agents. In particular, we consider deviations through under-exploration and additive reporting bias independent of locally observed data. Studying richer classes of strategic manipulation remains an important direction for future work.

While an honest protocol-adhering agent a implicitly has $\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0$ [2] for all communication rounds τ, a rational agent a shall pick $(\varepsilon^\theta_{a,\tau}, \varepsilon^V_{a,\tau})$ strategically. Denote by short-hands $\varepsilon_{a,\tau} = 0$ the honest strategy of $\varepsilon^\theta_{a,\tau} = 0$ and $\varepsilon^V_{a,\tau} = 0$, and by $\varepsilon_{a,\tau} \neq 0$ any free-riding dishonest strategy where $\varepsilon^\theta_{a,\tau} \neq 0$ or $\varepsilon^V_{a,\tau} \prec 0$.

Agent-principal contract. The goal of the principal—despite the agents being strategic—is to minimize the social regret, i.e., the cumulative regret $\sum_{a\in\mathcal{A}} R_a(T)$ of all agents. However, to guard against such free-riding strategies and ensure agents act truthfully, we consider a setting where a contract exists between the agents and the principal. Under this contract, the principal can levy enforceable monetary penalties at communication events based on a statistical assessment of the agent's adherence to the protocol. In this case, the goal of every agent a is to minimize the sum of his regret and penalty levied on him: $R_a(T) + P_a(T)$, where $P_a(T)$ is the overall penalty levied on agent a.

2.1 Framework

Algorithm 1 Framework for multi-agent linear bandits with rational agents

1: **for** time-step $t = 1, 2, \ldots, T$ .. **do**
2: **for all** agents $a \in \mathcal{A}$ **do**
3: **[Agent]** Chooses and plays arm $x_{a,t}$ and observes reward $y_{a,t}$.
4: **if** Communication event occurs at $\tau \leftarrow t$ **then**
5: **[Agent]** Reports estimate $\theta^\varepsilon_{a,\tau}$ to principal.
6: **[Principal]** Shares back aggregated statistics $V^c_{a,\tau}, B^{c,\varepsilon}_{a,\tau}$ to agent a.
7: **[Principal]** Levies penalty $P_{a,\tau}$ on agent a.
8: **end if**
9: **end for**
10: **end for**

We summarize the interactions and data flow in Framework 1. Each agent individually plays the bandit instance using its locally available statistics (line 3). The chosen arms and observed reward remain private and are not shared by default with other agents or the principal. At certain time-steps τ, the principal initiates communication events (line 4). During these events, agents report $\theta^\varepsilon_{a,\tau}$s as their current parameter estimates (line 5) to the principal, and then, the principal sends the aggregated statistics (design and results matrices) back to the agents and imposes a penalty $P_{a,\tau}$ (lines 6-7). This setup resembles that of Wang et al. (2019), with some differences in the nature of statistics transferred and how communication is triggered. These communication events naturally induce a game among agents, which we formalize in normal form as follows:

Definition 3 (Principal-agent bandit game)**.** *Let $\Gamma = [\mathcal{A}, (S_a)_{a\in\mathcal{A}}, (u_a)_{a\in\mathcal{A}}]$ denote the normal form* principal-agent bandit game *, where*

1. *The set of players is the finite set of agents $\mathcal{A}$.*
2. *Each player $a \in \mathcal{A}$ chooses strategy $s_a = ((\varepsilon^\theta_{a,\tau}, \varepsilon^V_{a,\tau}) : \tau \in \mathcal{T}) \in S_a \subseteq \left(\mathbb{R}^d \times \mathbb{R}^{d\times d}\right)^q$.*

[2] We simplify notation and overload 0 to denote different quantities (vector, matrix, tuples of quantities) of all zeros, disambiguated by the l.h.s. quantity compared to.

3. Then, each player $a \in \mathcal{A}$ receives utility

$$u_a((s_a)_{a\in\mathcal{A}}) = -R_a(T) - P_a(T), \tag{2}$$

where $R_a(T)$ is the regret as in Equation (1), and $P_a(T) = \sum_{\tau\in\mathcal{T}} P_{a,\tau}$ is the penalty as in line 7 of Alg. 1.

The goal of this work is to jointly design mechanisms (and therein the penalty functions) for the principal to use and arm-playing algorithms for the agents to use, that dovetail and delicately balance the following aspects: the principal mechanism should provably induce equilibrium with honest behaviour that prevent free-riding; arm play algorithms should enable/permit the possibility of designing such mechanisms; the arm play algorithms should have provably good regret guarantees; and also have low penalty guarantees in conjunction with the principal mechanism.

3 Our Mechanism and Algorithm

We introduce protocols LsHon for the agent behaviour (Algorithm 2) and MonPen for principal behaviour (Algorithm 3) that fit into our setting Framework 1.

3.1 Agent Protocol

As part of the contract, the agent is required to adhere to LsHon which prescribes to choose and play actions in the following way. The agent shall make use of the total statistics $V_{a,\tau} = V^{s,\varepsilon}_{a,\tau} + V^{c}_{a,\tau}$ and $B_{a,\tau} = B^{s,\varepsilon}_{a,\tau} + B^{c,\varepsilon}_{a,\tau}$ from both self-play and from collaboration with other agents, as of the last communication event at time τ. The choice of actions is based on the low-switching algorithm of Abbasi-Yadkori et al. (2011) which is similar to the LinUCB algorithm except that it changes/updates its parameter estimate very infrequently. This algorithm balances exploration and exploitation by constructing a confidence set around the current parameter estimate and selecting an *optimistic* action based on this confidence set. Specifically, the agent computes the ridge regression (RR) estimator

$$\theta^{\mathrm{RR}}_{a,\tau} = (\lambda I + V_{a,\tau})^{-1}(B_{a,\tau})\,, \tag{3}$$

where $\lambda > 0$ is the ridge regression parameter. Then, the agent constructs the confidence set

$$C_{a,\tau} = \left\{\theta' \in \mathbb{R}^d : \left\|\theta' - \theta^{\mathrm{RR}}_{a,\tau}\right\|_{\lambda I + V_{a,\tau}} \le \beta_t(\delta)\right\}\,, \tag{4}$$

where $\beta_t(\delta)$ is the radius for a confidence parameter δ (see Equation (36) in Appendix). The agent uses this confidence set $C_{a,\tau}$ to choose the arm (breaking ties in some deterministic pre-agreed way) to play, at all time-steps until the next communication event (line 4).

Algorithm 2 LsHon - Agent's Truthful Protocol with low-switching arm play.

1: Most recent communication time $\tau \leftarrow 0$.
2: **for** time-step $t = 1, 2, \ldots, T$ **do**
3: Compute $\theta_{a,\tau}$ and $C_{a,\tau}$ using Equations (3) and (4) with statistics until last communication event τ.
4: Choose and play arm

$$x_{a,t} \in \arg\max_{x\in\mathcal{X}_t} \max_{\theta'\in C_{a,\tau}} \langle x, \theta' \rangle\,.$$

5: **if** Communication round is initiated **then**
6: Report true OLS parameter estimate $\theta^{\varepsilon}_{a,t} \leftarrow \theta_{a,t}$ as in Equation (5) to the principal.
7: Recent communication event time $\tau \leftarrow t$.
8: **end if**
9: **end for**

Since the confidence set is not updated every single time-step, the algorithm is said to follow a 'low-switching' or 'deferred update' paradigm where feedback from playing arms is processed in

batches only during communication events. As a result of this, the choices of arms played are independent of the reward noise from the ongoing bandit play (since the last comm. event), and thus is a deterministic function of the set $C_{a,\tau}$. This property enables the principal to compute the sequence of actions an agent will play through the ongoing bandit play, and thus compute the anticipated design $V^s_{a,\tau}$ every agent a that obeys LsHon shall have at communication event τ.

Finally, whenever a communication event is initiated, the agent shall compute Ordinary Least Squares (OLS) parameter estimate using all locally available statistics (from self-play until current time t and from collaboration as of preceding communication event τ) as

$$\theta_{a,t} = \left(V^{s,\varepsilon}_{a,t} + V^{c}_{a,\tau}\right)^{\dagger} \left(B^{s,\varepsilon}_{a,t} + B^{c,\varepsilon}_{a,\tau}\right) \tag{5}$$

and report it to the principal (line 6). Here, while the free-rider's design $V^{s,\varepsilon}_{a,\tau}$ can be non-invertible, we assume the honest design matrices $V^s_{a,\tau}$ are full-rank and invertible by every communication event τ. Note this can also be explicitly achieved by having a dedicated phase of d time-steps, wherein every agent plays a sequence of actions to span $\mathbb{R}^d$ which make the design matrices full-rank and invertible, at the beginning of arm-play after a communication event.

3.2 Principal Protocol

As per MonPen, the principal initiates communication events whenever sufficient new statistics have been generated (line 4), i.e., when the determinant value of the cumulative global design matrix $V^{\mathcal{A}}_t := \sum_{a\in\mathcal{A}} V^s_{a,t}$ has increased by a multiplicative factor of $(1+c_1)$ since the last communication event τ. Note this initiates at most $O(d \log T)$ communication events. During a communication event τ, the principal receives parameter estimates from all agents (line 6). As the parameter estimates $\theta^{\varepsilon}_{a,\tau}$s are from statistics from both self-play and other agents' statistics provided to agent a (as per Equation (5)), the principal attempts to recover the agent a's self-play statistics as follows:

$$B^{s,\varepsilon}_{a,\tau} = \left(V^s_{a,\tau} + V^c_{a,\tau'}\right)\theta^{\varepsilon}_{a,\tau} - B^{c,\varepsilon}_{a,\tau'}, \tag{6}$$

where the principal is aware of $V^c_{a,\tau'}, B^{c,\varepsilon}_{a,\tau'}$ as the statistics it sent to a in the last (preceding) communication event $\tau' < \tau$, and the principal can compute $V^s_{a,\tau}$ due to the determinism of the choice of arm played by the low-switching algorithm in LsHon. Since the principal observes only reported parameter estimates and not the underlying actions or rewards, reconstructing self-play statistics is non-trivial and relies crucially on the low-switching determinism of LsHon. The principal, then, shall return to every agent a the complete aggregate statistics

$$V^c_{a,\tau} = \sum_{b\in\mathcal{A}\setminus\{a\}} V^s_{b,\tau} \qquad \text{and} \qquad B^{c,\varepsilon}_{a,\tau} = \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s,\varepsilon}_{b,\tau} \tag{7}$$

of all other agents (line 8). Note in each communication event $O(d^2)$ floating-point values are exchanged between each agent and the principal, thereby bringing the overall communication complexity of the protocol to $O(d^3 M \log T)$.

Statistical test. Next, the principal evaluates probabilistically the extent of dishonesty/deviation of every agent a and assigns (line 13) a commensurate monetary penalty to the agent by performing a statistical test as follows.

We introduce some notations to capture the design and result matrices corresponding to intervals $[\tau', t]$ of time: $V^c_{a,\tau':\tau} := V^c_{a,\tau} - V^c_{a,\tau'}$, $B^{c,\varepsilon}_{a,\tau':\tau} := B^{c,\varepsilon}_{a,\tau} - B^{c,\varepsilon}_{a,\tau'}$. Analogously define $V^s_{a,\tau':\tau}$ and $B^{s,\varepsilon}_{a,\tau':\tau}$. Observe that $V^c_{a,\tau':\tau} = \sum_{b\in\mathcal{A}\setminus\{a\}} V^s_{b,\tau':\tau}$ and $B^{c,\varepsilon}_{a,\tau':\tau} = \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s,\varepsilon}_{b,\tau':\tau}$. We also summarize all our notations in Appendix A.

The principal computes the statistics of all other agents $b \in \mathcal{A}\setminus\{a\}$ in the time period since the preceding communication event τ' until the current communication event τ as $V^c_{a,\tau':\tau}$ and $B^{c,\varepsilon}_{a,\tau':\tau}$, and attempts to compute its/their corresponding OLS estimate

$$\theta^{c,\varepsilon}_{a,\tau':\tau} = (V^c_{a,\tau':\tau})^{-1} B^{c,\varepsilon}_{a,\tau':\tau}. \tag{8}$$

Algorithm 3 MONPEN - Principal's protocol with monetary penalty

Parameters: $c_1 > 0, c_2 > 0, \lambda > 0$.
Input: Estimate $\theta^{\varepsilon}_{a,\tau}$ from every agent $a \in \mathcal{A}$ at every communication event τ.
Output: Statistics sent back and penalty $P_{a,\tau}$ for every agent a at every communication event τ.

1: Most recent communication event time $\tau \leftarrow 0$.
2: **for** time-step $t = 1, 2, \ldots, T$ **do**
3: Compute agent designs $V^s_{a,t}$s as per LSHON, compute global design $V^{\mathcal{A}}_t$.
4: **if** $\det\left(\lambda I + V^{\mathcal{A}}_t\right) \geq (1 + c_1) \cdot \det\left(\lambda I + V^{\mathcal{A}}_\tau\right)$ **then**
5: Initiate communication event, $\tau \leftarrow t$.
6: Get parameter estimates $\theta^{\varepsilon}_{a,\tau}$ from all agents a.
7: **for all** agent $a \in \mathcal{A}$ **do**
8: Send to agent a the statistics $V^c_{a,\tau}$, $B^{c,\varepsilon}_{a,\tau}$ of all other agents as in Equations (6) and (7).
9: Compute intermediary vector z_a as in Equation (9).
10: **for** dimension $i \in [d]$ **do**
11: With $x \leftarrow (z^i_a)^2$, compute component-wise penalty $P^i = \dfrac{\sqrt{2\pi x}}{(x+1)^{1+c_2}} \cdot \exp\left(\dfrac{x}{2}\right) - \frac{1}{c_2}$.
12: **end for**
13: Levy monetary penalty $P_{a,\tau} = \frac{1}{d}\sum_i P^i$ on agent a.
14: **end for**
15: **end if**
16: **end for**

Next, the principal compares this derived estimate to agent a's reported estimate $\theta^{\varepsilon}_{a,\tau}$ to get an intermediary d-dimensional vector z_a (line 9) as

$$z_a = \left(\left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} + \left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2} \left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau}\right). \tag{9}$$

Finally, the components of z_a are fed to a hand-crafted penalty function (lines 10-11) to get the penalty for agent a at this communication event τ. We shall later show that the specific exponential form of the penalty function is not arbitrary—it is carefully designed to amplify statistical inconsistencies between an agent's reported estimates and what would be expected under truthful exploration. By making the penalty depend exponentially on the squared magnitude of the reshaped estimate components, the mechanism creates a sharp separation between honest and dishonest behavior: the expected penalty for an honest agent remains bounded, whereas that of a free-rider diverges. Notice that this penalty assigned to an agent, $P_{a,\tau}$, depends not only on the agent's own actions and estimates but also on those of the other agents that agent a can't observe beforehand.

4 Results and Discussion

We establish in Theorem 1 that the principal's protocol MONPEN is designed in a way such that it induces an equilibrium of honesty (adherence to protocol) among all agents when prescribed to follow protocol LSHON.

Theorem 1. *Let the principal follow* MONPEN *in common knowledge and agents adhere to Assumption 2. Then, the strategy profile of all agents being honest, $s^* = ((\varepsilon^{\theta}_{a,\tau} = 0, \varepsilon^{V}_{a,\tau} = 0) : \forall \tau \in \mathcal{T})_{a \in \mathcal{A}}$, is a Nash Equilibrium of the principal-agents bandit game Γ.*

With this explicit specification of the equilibrium strategy profile, we note it is *this* equilibrium behaviour that is incorporated in LSHON for the agents. Thus, the above establishes that the protocol pairing MONPEN-LSHON, our principal protocol and agent algorithm, is Bayesian-Nash Incentive Compatible wherein it is beneficial for each agent to reveal their true parameter estimates as long as the other agents are also doing so. Next, we consider the trajectory where this equilibrium occurs and analyze the bandit regret for an agent from playing the stochastic linear bandit instance.

Theorem 2. *[Regret at Equilibrium] In the collaborative linear bandits problem, let the principal follow* MONPEN *and all agents follow* LSHON*. Then, the expected regret of every agent $a \in \mathcal{A}$ is*

$$R_a(T) = O\left(d\sqrt{\frac{T\log(TM)}{M}}\right).$$

And, the expected penalty $P_a(T)$ of every agent is $O(1)$.

Our result shows that we can attain the same regret bound as that of altruistic collaboration even when all agents are rational and self-interested. Wang et al. (2019) showed that if the agents are implicitly altruistic, then $\tilde{O}(d\sqrt{MT})$ social regret is achievable. Here, Theorem 2 shows that the same social regret of $\tilde{O}(d\sqrt{MT})$ is attained when agents are strategic.

The Theorem 2 immediately leads to the following:

Corollary 1 (Utility at Equilibrium)**.** *In the principal-agent bandit game Γ, when the principal follows* MONPEN *and all agents follow* LSHON*, the negative utility of every agent $a \in \mathcal{A}$ is upper bounded as $-u_a(s^*) = O\left(d\sqrt{\frac{T\log(TM)}{M}}\right)$ at NE strategy profile s^* (mentioned in Theorem 1).*

If each agent operates individually without the collaboration, they can achieve an $\tilde{O}(d\sqrt{T})$ expected regret using the LinUCB algorithm (Abbasi-Yadkori et al., 2011). However, our Theorem 2 (sim. Corollary 1) shows an improvement of the expected regret (sim. negative utility) bound to $\tilde{O}(d\sqrt{T/M})$ when all agents participate in the collaboration adhering to the mentioned equilibrium behaviour. Upfront, the agents are better off participating in this collaboration than abstaining, despite the threat of a monetary penalty. Thus, our protocol pairing is ex ante Individually Rational (IR) in equilibrium. Further, from the single agent lower bound of Dani et al. (2008), it follows that even with communication event at every time-step and complete sharing of statistics, superfluously set up as a single agent playing for MT time sequentially, the per-agent regret bound is lower bounded as $\Omega(d\sqrt{T/M})$. This shows that our algorithm attains minimax optimal regret guarantee.

While we designed the MONPEN, LSHON tailored to show that honesty is a Nash Equilibrium in the game Γ, it is imperative to ask if these protocols also induce other unintended equilibriums to which agents might converge. We answer this question in Theorem 3 for the single communication event case, where we explicitly enumerate all possible NEs.

Theorem 3. *[All Nash Equilibria] Let the principal follow* MONPEN *in common knowledge. If there is one communication event in the principal-agent bandit game Γ, then the set of all Nash equilibria is*

$$S \subseteq \left\{ \left(\varepsilon_a^\theta, \varepsilon_a^V\right)_{a\in\mathcal{A}} \;\middle|\; \varepsilon_a^V \not\prec 0 \text{ and } \exists\, \beta \in \mathbb{R}^d \text{ s.t. } \forall a \in \mathcal{A},\ \varepsilon_a^\theta = \beta \right\}$$

In words, a strategy profile is a Nash equilibrium only if all agents apply the same identical bias $\beta \in \mathbb{R}^d$ to their reported estimates, and do not free-ride in it ($\varepsilon_a^V \not\prec 0$).

This crucially establishes that any free-riding strategy can *not* be a part of any Nash equilibrium. Further, although the strategy profiles where all agents uniformly apply a certain non-zero bias to their reported values constitutes an NE, these strategy profiles fare poorly in regret minimization as all agents together mislead the principal.

However, note that our Nash Equilibrium (discussed in Theorem 1) leads to a per-agent optimal minimax regret as mentioned above, and this leads to a socially optimal regret bound of $\widetilde{O}\left(d\sqrt{TM}\right)$. Owing to this optimality, even when other equilibria exist, they cannot have a better/lesser minimax social regret guarantee. In fact, this shows that no hand-crafted strategy profile for the set of agents (even if they do not constitute an equilibrium) can lead to a better minimax social regret, showing that our Nash equilibrium of honesty is socially optimal.

We discuss some limitations of our setting and results in Appendix H. Deferring the formal proofs to Appendices D to F, we provide brief proof sketches next.

4.1 Proof sketches

Proof Sketch of Theorem 1 [Nash Equilibrium of Honesty.] We want to show that when all other agents are honest, agent a enjoys the highest utility when he is honest in every single communication round. Towards this, we shall show that the expected penalty of an agent is $O(1)$ if he is honest (Lemmas 10 and 17) and that it doesn't converge (tends to $+\infty$) otherwise (Lemmas 13 and 19).

This result primarily hinges on the principal's ability to distinguish between an honest and dishonest agent, probabilistically, from their reported parameter estimates $\theta^{\varepsilon}_{a,\tau}$s without actually knowing the true parameter value θ^*. Building on the intuition that a free-rider's parameter estimate $\theta^{\varepsilon}_{a,\tau}$ will be generally 'poorer' (or less statistically reliable) than that of honest agents' as they don't perform the requisite exploration, the principal computes a specific statistical quantity $z_a \in \mathbb{R}^d$ (in line 9 of MONPEN) by comparing the estimate of agent a with the aggregate estimate of all other agents' new statistics that is yet to be given to agent a. We shall show that this z_a doesn't depend on the unknown θ^*, and conditioned on the honesty of all other agents, z_a follows a 0-mean Gaussian distribution with all components having a marginal variance 1 if a was honest, but if a was dishonest and free-riding, some component's marginal variance goes strictly above 1 (Claim 14).

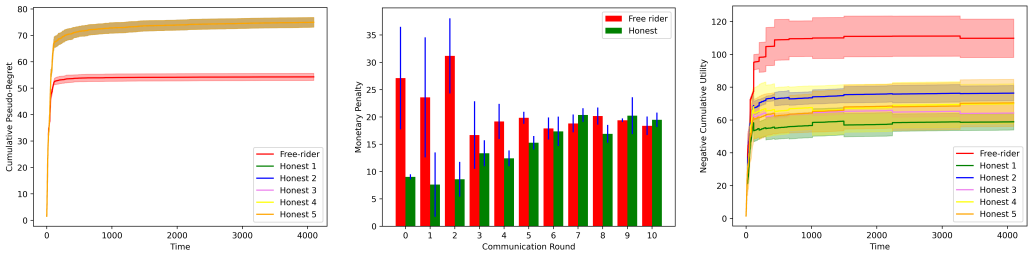
Taking advantage of this distinction, the principal computes the penalty of agent a as function of these components of z_a (in line 11) specifically designed such that its expected value converges to a constant for variance 1 (for an honest agent) and doesn't converge for variance more than 1 (for a dishonest agent). □

Proof Sketch of Theorem 2 [Regret bound at honest behaviour.] The key idea is that all (honest) agents shall explore and contribute *equally* to the collaboration. This holds due to the deterministic nature, across all agents, of the arm playing algorithm in LSHON between two communication events. As the learning only occurs in batches during the times of communication, all agents use the same confidence set $C_{a,\tau}$ for the entire period until the next communication event. Thus, we shall show (in Lemma 22) that choice of arm played (in line 4) is the same across all agents leading to all agents maintain an identical design matrix $V_{a,t}$. This shall establish that all agents share the burden of exploration equally.

We shall then quantify the benefit of collaboration among M agents as an improvement (reduction) in individual regret by a factor of $\sqrt{M}$ for all agents, by using this fact that all agents contribute equally to the collaboration. The remainder of the proof shall then broadly use the line of argument in Abbasi-Yadkori et al. (2011) as follows: First, we shall upper bound the instantaneous regret $r_{a,t}$ of an agent at time t, and shall then use it to bound the total regret of agent a to get the desired bound of $R_a(T)$. □

Proof sketch of Theorem 3 [Characterizing all Nash Equilibriums.] The principal's mechanism MONPEN is constructed so that the expected penalty of an agent remains finite only when the statistic z_a computed by the principal follows the reference distribution $z_a \sim \mathcal{N}(0, I_d)$; otherwise the expected penalty diverges to $+\infty$. Hence, given the strategies of other agents, any best response of agent a must ensure that the induced distribution of z_a satisfies this condition. We therefore characterize best responses by enforcing the two requirements $\mathbb{E}[z_a] = 0$ and $\mathrm{Var}[z_a] \preceq I_d$.

First, analyzing the mean condition shows that the best response bias of agent a must equal the average bias of the other agents. Consequently, any strategy profile that is a fixed point of best responses must have a common bias $\varepsilon^{\theta}_a = \beta$ for all a. Second, we analyze the variance condition by introducing a scalar slack variable f_a that captures deviations in the design matrix and corresponds to the degree of free-riding. The variance constraint implies an inequality relating f_a to the values chosen by other agents. This inequality forces f_a to lie no farther from 1 than the deviations of the other agents. Iterating this best-response relation shows that any self-consistent profile must satisfy $f_a = 1$ for all a, which corresponds to the absence of free-riding (equivalently, $\varepsilon^{V}_a \not\prec 0$). □



(a) (b) (c)

Figure 1: **MovieLens problem instance** (a) Regret of agents as a function of time t. Note that regrets are identical for all honest agents. (b) Penalty of agents as a function of communication rounds. (c) Negative utility of agents.

5 Experiments

We demonstrate the effectiveness of our approach in disincentivizing 'free-riding' under two different linear bandit environments: (1) synthetically generated, (2) based on the MovieLens dataset (Harper & Konstan, 2015). We give details of bandit environment setup in Appendix G.1.

Presence of free-rider. The purpose of this experiment is to demonstrate the usefulness of our Protocol for multi-agent stochastic linear bandits problem in the following sense. Existing algorithms are not immune to the presence of strategic agents, i.e., a strategic agent will be able to 'free-ride' leading to a bad social outcome over time due to a collective lack of exploration. On the other hand our algorithm will be able to impose a monetary penalty to the agents, the penalty serves as a deterrent to such 'free-riding' and creates an incentive to be truthful. To elucidate this, in our collaborative learning setup, we let one of the agents 'free-ride' and analyse how the regret and penalty of the agents behave. Specifically, the free-rider agent, say a, plays the best arm according to the $\theta_{a,\tau}$ based on total statistics from self-play and from communication by the principal during the preceding communication event, and does not perform exploration of his own. Specifically, he plays $x_{a,t} = \arg\max_{x \in \mathcal{X}_t} \langle x, \theta_{a,\tau} \rangle$, while all other agents fulfill the exploration as dictated in (line 4 of) agent protocol LsHon. All agents, including the free-rider, report their current estimate $\theta_{a,t}$ honestly.

Disincentivization of free-riding. We observe that the regret of the free-rider is lower than that of the honest agents (Fig. 1a). The exploration information that the free-rider obtains during the communication events more than sufficiently compensates for his own lack of exploration, thereby letting him enjoy a better regret. At this juncture, we note that the algorithms in the literature (whose goals were different to that of handling strategic agents) such as Wang et al. (2019) are not immune to this phenomenon of agents free-riding. However, in our work, we observe that the principal's MonPen penalizes the free-rider significantly more than it penalizes the honest agents (in Fig. 1b). Hence, this sum of regret and penalty, i.e., the negative utility, is higher for the free-rider than for the honest agent (in Fig. 1c), which makes free-riding undesirable.

It is also interesting to note the variation of penalty across the different communication events (in plots Fig. 1b). The free-rider incurs a lot more penalty than the honest agent in the initial few rounds, whereas their penalties appear to become similar towards the later stages. This can be attributed to the inherent nature of the arm play algorithms employed by the two. The free-rider always exploits the current estimate. The honest agent balances his exploration and exploitation, inherently leaning towards exploring more in the early stages and towards exploiting more in the later part. Thus, while the exploitative arm-play of the free-rider starkly differs from the highly exploratory arm-play of the honest agent in the initial stages and thus results in heavier penalties for the free-rider, it

should be noted their arm-plays get 'closer' to that of each other towards the later stages when both predominantly (and rightly) exploit, resulting in similar penalties.

Nevertheless, albeit lesser than a free-rider, we observe that even an honest agent incurs some positive penalty (Fig. 1b). To mitigate this empirical phenomenon, we can consider redistribution of the accrued penalty equally back among the agents. That is, we subtract from the penalty (add to the utility) of every agent a uniform value of $\frac{1}{M}\sum_{a\in\mathcal{A}} P_a(T)$. This can be interpreted as the agents transferring the monetary penalty to one another, as opposed to the principal collecting the penalties (and keeping for himself). Note that as the total penalties levied over all agents is still non-negative, the MONPEN remains budget-feasible. As this is a fixed quantity across agents, it doesn't affect the separation of curves of different agents witnessed in Fig. 1c.

Robustness of MONPEN. Fig. 2 illustrates the robustness of our penalty mechanism, MONPEN, along two key dimensions. First, we evaluate performance under misspecified reward noise, i.e., when the environment has different levels of stochasticity (reward variance) than the noise level (1.0) for which MONPEN was designed. Second, we assess robustness against varying degrees of free-riding, controlled by the *play fraction*—the fraction of time-steps the free-rider follows LSHON in a round-robin manner. During the remaining time-steps, the free-rider plays no action, observes no reward, and incurs no regret.

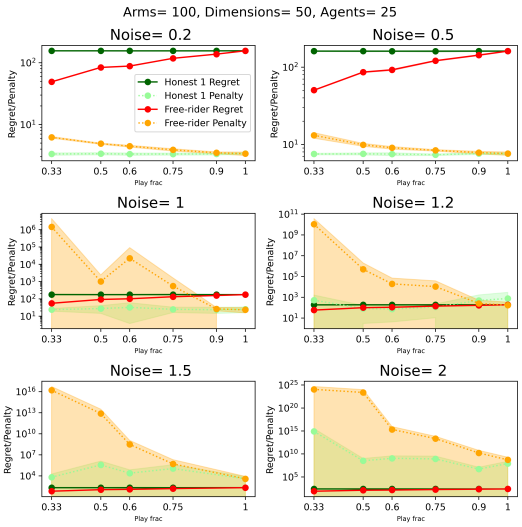


Figure 2: Variation of penalty and regret (y-axis) w.r.t. different free-riding extents and different environment reward noise (x-axis).

As expected, the free-rider incurs lower regret when their play fraction is small, and this regret increases steadily as play fraction increases. At the design noise level (1.0), the honest agent consistently receives low penalties (both in average and variance), while the free-rider receives significantly higher penalties when free-riding is extensive (i.e., low play fraction). MONPEN penalizes the free-rider heavily up to a play fraction of 0.75. However, at higher play fractions (≥ 0.9), the distinction between the honest agent and free-rider becomes less reliable. These trends also hold under moderate noise misspecification (noise=1.2).

On the flip side, the honest agent continues to receive low penalties—comparable to order of regret—so long as the reward noise remains ≤ 1.2. This suggests that MONPEN is robust to underestimation of noise up to 20%. Beyond that (e.g., noise $= 1.5, 2$), both the honest agent and free-rider begin to incur large penalties, indicating the limits of MONPEN's tolerance to noise misspecification.

6 Conclusion

This paper studies collaborative linear bandits in the presence of strategic agents and proposes new protocols for the principal that induce truthful behaviour as a Nash equilibrium. By establishing minimax optimal regret guarantees at this equilibrium, we show that one can obtain almost the full benefit of altruistic collaboration despite strategic behaviour from agents.

As future work, we aim to expand our work to consider an asymmetric set of agents with different action sets, different prior knowledge/beliefs about the model parameter, and alternative communication frameworks to study the dichotomy between the benefits of collaboration and ill-effects of strategic agent behaviour.

Acknowledgements

Rama Krishnamurthy was supported by the Meta-AI mentorship program from Meta Platforms Inc.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Baruch Awerbuch and Robert Kleinberg. Competitive collaborative learning. *Journal of Computer and System Sciences*, 74(8):1271–1288, 2008.

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pp. 1–8. IEEE, 2020.

Etienne Boursier and Vianney Perchet. Selfish robustness and equilibria in multi-player bandits. In *Conference on Learning Theory*, pp. 530–581. PMLR, 2020.

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, pp. 961–987. PMLR, 2020.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Multi-armed bandits in the presence of side observations in social networks. In *52nd IEEE conference on decision and control*, pp. 7309–7314. IEEE, 2013.

Loris Cannelli, Giuseppe Nuti, Marzio Sala, and Oleg Szehr. Hedging using reinforcement learning: Contextual k-armed bandit versus q-learning. *The Journal of Finance and Data Science*, 9: 100101, 2023.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, volume 2, pp. 3, 2008.

Abhimanyu Dubey and AlexSandy' Pentland. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

Aidin Ferdowsi, Samad Ali, Walid Saad, and Narayan B Mandayam. Cyber-physical security and safety of autonomous connected vehicles: Optimal control meets multi-armed bandit learning. *IEEE Transactions on Communications*, 67(10):7228–7244, 2019.

Mona Flores, Ittai Dayan, Holger Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Abidin, Andrew Liu, Anthony Costa, Bradford Wood, et al. Federated learning used for predicting outcomes in sars-cov-2 patients. *Research Square*, 2021.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Multi-agent heterogeneous stochastic linear bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 300–316. Springer, 2022.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. DOI: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Jiafan He, Tianhao Wang, Yifei Min, and Quanquan Gu. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. *Advances in neural information processing systems*, 35:4762–4775, 2022.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

Christopher Jung, Sampath Kannan, and Neil Lutz. Quantifying the burden of exploration and the unfairness of free riding. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1892–1904. SIAM, 2020.

Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

Sai Praneeth Karimireddy, Wenshuo Guo, and Michael Jordan. Mechanisms that incentivize data sharing in federated learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

Nikolai Karpov and Qin Zhang. Collaborative best arm identification with limited communication on non-iid data. *arXiv preprint arXiv:2207.08015*, 2022.

Nathan Korda, Balazs Szorenyi, and Shuai Li. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pp. 1301–1309. PMLR, 2016.

Ramakrishnan Krishnamurthy, Arpit Agarwal, Lakshmi Subramanian, and Maximilian Nickel. Creator incentives in recommender systems: A cooperative game-theoretic approach for stable and fair collaboration in multi-agent bandits. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations research*, 69(1):315–330, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Chuanhao Li and Hongning Wang. Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 6529–6553. PMLR, 2022.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

James Mancini. Data portability, interoperability and digital platform competition: Oecd background paper. *Interoperability and Digital Platform Competition: OECD Background Paper (June 8, 2021)*, 2021.

Yi Mao and Andrew Perrault. Time-constrained restless multi-armed bandits with applications to city service scheduling. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, pp. 2375–2377, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

Aritra Mitra, Arman Adibi, George J Pappas, and Hamed Hassani. Collaborative linear bandits with adversarial agents: Near-optimal regret bounds. *Advances in Neural Information Processing Systems*, 35:22602–22616, 2022.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767, 2021.

Ahmadreza Moradipari, Mohammad Ghavamzadeh, and Mahnoosh Alizadeh. Collaborative multi-agent stochastic linear bandits. In *2022 American Control Conference (ACC)*, pp. 2761–2766. IEEE, 2022.

Aniket Murhekar, Zhuowen Yuan, Bhaskar Ray Chaudhury, Bo Li, and Ruta Mehta. Incentives in federated learning: Equilibria, dynamics, and mechanisms for welfare maximization. *Advances in Neural Information Processing Systems*, 36, 2024.

Han-Ching Ou, Christoph Siebenbrunner, Jackson A. Killian, Meredith B. Brooks, David Kempe, Yevgeniy Vorobeychik, and Milind Tambe. Networked restless multi-armed bandits for mobile interventions. In Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (eds.), *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, pp. 1001–1009. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

Akash Das Sharma, Sujit Gujar, and Y Narahari. Truthful multi-armed bandit mechanisms for multi-slot sponsored search auctions. *Current Science*, pp. 1064–1077, 2012.

Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *International conference on artificial intelligence and statistics*, pp. 2917–2925. PMLR, 2021.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International conference on machine learning*, pp. 19–27. PMLR, 2013.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 126–146. IEEE, 2019.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Improving multi-armed bandit algorithms in online pricing settings. *International Journal of Approximate Reasoning*, 98: 196–235, 2018.

Daniel Vial, Sanjay Shakkottai, and R Srikant. Robust multi-agent multi-armed bandits. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 161–170, 2021.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Lingda Wang, Bingcong Li, Huozhi Zhou, Georgios B Giannakis, Lav R Varshney, and Zhizhen Zhao. Adversarial linear contextual bandits with graph-structured side observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10156–10164, 2021.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2019.

Zhepei Wei, Chuanhao Li, Haifeng Xu, and Hongning Wang. Incentivized communication for federated bandits. *Advances in Neural Information Processing Systems*, 36:54399–54420, 2023.

Zhepei Wei, Chuanhao Li, Tianze Ren, Haifeng Xu, and Hongning Wang. Incentivized truthful communication for federated bandits. *arXiv preprint arXiv:2402.04485*, 2024.

Renzhe Xu, Haotian Wang, Xingxuan Zhang, Bo Li, and Peng Cui. Competing for shareable arms in multi-player multi-armed bandits. *arXiv preprint arXiv:2305.19158*, 2023.

Zheqing Zhu, Rodrigo de Salvo Braz, Jalaj Bhandari, Daniel Jiang, Yi Wan, Yonathan Efroni, Liyuan Wang, Ruiyang Xu, Hongbo Guo, Alex Nikulkov, Dmytro Korenkevych, Ürün Dogan, Frank Cheng, Zheng Wu, and Wanqiao Xu. Pearl: A production-ready reinforcement learning agent. *CoRR*, abs/2312.03814, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Preliminaries

As the paper is heavy in notations, we summarize the frequently used notations with their meanings/definitions in Table 2.

Table 2: Description of notations frequently used in the document.

Notation	Description
$\mathcal{A}$	The set of all M agents.
$V^s_{a,t}$	The analytical design matrix of agent a at time t including samples from only self-play as per LsHon. Principal computes this using deterministic arm-play of LsHon.
$V^{s,\varepsilon}_{a,t}$	The actual design matrix of agent a at time t including samples from only self-play (with possible strategic deviation).
$B^{s,\varepsilon}_{a,t}$	The realized results matrix of agent a at time t including samples from only self-play (with possible strategic deviation).
$V^c_{a,\tau}$	The analytical aggregate design matrix at time τ including samples from self-play as per LsHon of all agents other than a. Principal computes this using $V^c_{a,\tau} = \sum_{b\in\mathcal{A}\setminus\{a\}} V^s_{b,\tau}$.
τ	Arbitrary time-step of communication.
$\theta^{\text{RR}}_{a,\tau}$	Agent a's ridge regression estimate of θ^* from statistics only as of preceding communication round τ—both from self-play and collaboration—as in Equation (3).
$\theta_{a,\tau}$	Agent a's true OLS estimate of θ^* from all locally available statistics—both from self-play and collaboration—at time τ, as in Equation (5).
$\theta^{\varepsilon}_{a,\tau}$	The estimate that agent a reports to the principal in a communication event at time τ (with possible strategic bias).
$B^{s,\varepsilon}_{a,\tau}$	The recovered results matrix of agent a at time τ including samples from only self-play (with possible deviation and bias) of all agents. Principal computes this in Equation (6).
$B^{c,\varepsilon}_{a,\tau}$	The aggregate of recovered results matrices at time τ from samples from self-play of all agents other than a (with possible deviation and bias).
$V^c_{a,\tau':\tau}$	$V^c_{a,\tau} - V^c_{a,\tau'}$. Analytical aggregate of design matrix from self-play as per LsHon of all agents other than a (with possible deviation and bias) in the period $[\tau'+1,\tau]$.
$B^{c,\varepsilon}_{a,\tau':\tau}$	$B^{c,\varepsilon}_{a,\tau} - B^{c,\varepsilon}_{a,\tau'}$. Aggregate of recovered results matrices from samples during $[\tau'+1,\tau]$ from self-play of all agents other than a (with possible deviation and bias).
$\theta^{c,\varepsilon}_{a,\tau':\tau}$	The recovered OLS estimate from self-play of agents other than a (with possible deviation and bias) since the preceding communication event τ'. Principal computes this in Equation (8)
$V_{a,t}$	The total design matrix of agent a at time t including samples from both self-play and collaboration as of preceding communication event τ. $V^s_{a,t} + V^c_{a,\tau}$
$B_{a,\tau}$	The total results matrix of agent a at time τ including samples from both self-play and collaboration as of preceding communication event τ. $B^{s,\varepsilon}_{a,\tau} + B^{c,\varepsilon}_{a,\tau}$
$P_{a,\tau}$	Penalty levied on agent a at communication event τ by the principal.
$V^{\mathcal{A}}_t$	The global/cumulative design of all agents at time t. $\sum_{a\in\mathcal{A}} V^s_{a,t}$.

A.1 Generalized Chi-squared distribution

Claim 1 (Probability density of Generalized $\chi^2(1)$)**.** *Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ be a univariate Gaussian random variable. The probability density function of $Y = X^2$ is given by*

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi y}} \exp\left(\frac{-1}{2} \cdot \frac{y+\mu^2}{\sigma^2}\right) \times \left[\exp\left(\frac{\sqrt{y}\mu}{\sigma^2}\right) + \exp\left(\frac{-\sqrt{y}\mu}{\sigma^2}\right)\right],$$

when $y \geq 0$, and 0 otherwise.

Proof. The PDF of X is given by $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$. The CDF of Y can be written as

$$F_Y(y) = \mathbb{P}\{Y \leq y\} = \mathbb{P}\{-\sqrt{y} \leq X \leq \sqrt{y}\} = F_X(\sqrt{y}) - F_X(-\sqrt{y}),$$

for any $y \geq 0$, and 0 otherwise (when $y < 0$). Then, the PDF of Y can be derived as follows:

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy}\left[F_X(\sqrt{y}) - F_X(-\sqrt{y})\right] \\
&= \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right] \\
&= \frac{1}{2\sqrt{y}}\left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{\sqrt{y}-\mu}{\sigma}\right)^2\right) + \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{-\sqrt{y}-\mu}{\sigma}\right)^2\right)\right] \\
&= \frac{1}{2\sigma\sqrt{2\pi y}}\left[\exp\left(\frac{-1}{2\sigma^2}\left(y+\mu^2-2\sqrt{y}\mu\right)\right) + \exp\left(\frac{-1}{2\sigma^2}\left(y+\mu^2+2\sqrt{y}\mu\right)\right)\right] \\
&= \frac{1}{2\sigma\sqrt{2\pi y}} \exp\left(\frac{-1}{2\sigma^2}\left(y+\mu^2\right)\right)\left[\exp\left(\frac{\sqrt{y}\mu}{\sigma^2}\right) + \exp\left(\frac{-\sqrt{y}\mu}{\sigma^2}\right)\right].
\end{aligned}$$

This completes the proof of the Claim. Additionally, the final expression can be concisely written using the hyperbolic cosine function as follows:

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi y}} \exp\left(\frac{-1}{2\sigma^2}\left(y+\mu^2\right)\right).\cosh\left(\frac{\sqrt{y}\mu}{\sigma^2}\right).$$

□

The corollary follows.

Corollary 2 (Probability density of $\chi^2(1)$)**.** *When $X \sim \mathcal{N}(0,1)$ is a standard normal random variable, $Y = X^2$ follows a chi-squared distribution of 1 degree of freedom whose probability density is given by*

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(\frac{-y}{2}\right),$$

when $y \geq 0$, and 0 otherwise.

A.2 Matrix Theory

In this subsection, we show some matrix properties/results that shall be used in the proofs of our Theorems.

Claim 2. *Let P, Q be two matrices such that $P \succ 0$ is symmetric positive definite, and $Q \not\succeq 0$ is not negative semi-definite. Then, the matrix $PQP \not\succeq 0$ is not negative semi-definite.*

Proof. As Q is not negative semi-definite, there exists some vector y such that

$$y^\top Q y > 0. \tag{10}$$

To show the Claim, it needs to be shown that there exists some vector z such that the following inequality holds:

$$z^\top (PQP) z > 0 \iff (P^\top z)^\top Q P z > 0 \overset{(a)}{\iff} (Pz)^\top Q (Pz) > 0, \tag{11}$$

where (a) uses the fact that P is symmetric. Choosing $z = P^{-1} y$ leads to

$$(Pz)^\top Q (Pz) = (PP^{-1}y)^\top Q (PP^{-1}y) = y^\top Q y \overset{(a)}{>} 0,$$

which shows Equation (11) and the Claim. Here, (a) is due to Equation (10).

□

Claim 3. *Let P, Q be two matrices such that $P \succ 0$ is symmetric positive definite, and $Q \succ 0$ is positive definite. Then, the matrix $PQP \succ 0$ is positive definite.*

Proof. For any vector z, we have

$$z^\top (PQP) z = (P^\top z)^\top Q (Pz) \overset{(a)}{=} (Pz)^\top Q (Pz) \overset{(b)}{>} 0, \tag{12}$$

where, (a) uses the condition that P is symmetric, (b) uses the condition that Q is positive definite. Then, Equation (12) shows the Claim. □

Claim 4. *For a positive definite matrix A such that $A - I \succ 0$ is also positive definite, some diagonal element of A is greater than* 1.

Proof. From $A - I \succ 0$, we have that $A \succ I$, and then $\det(A) > \det(I) = 1$. Let $\lambda_1, \ldots, \lambda_d$ be the eigenvalues of A, then, as the determinant value equals the product of the eigenvalues, we have $\Pi_{i \in [d]} \lambda_i > 1$, and then by AM-GM inequality, we have $\sum_{i \in [d]} \lambda_i > d$. As the trace (sum of diagonal elements) of a matrix equals the sum of eigenvalues, it follows that $\sum_{i \in [d]} A_{ii} > d$. So, it holds that $A_{ii} > 1$ for some $i \in [d]$. □

Lemma 5. *Let A, B, and ϵ be three matrices such that A and B are symmetric positive definite, ϵ is symmetric negative definite, and $A + \epsilon$ is positive definite, i.e., $A \succ 0, B \succ 0, \epsilon \prec 0, A + \epsilon \succ 0$.*

Then, $\Sigma = (A^{-1} + B^{-1})^{-1/2}((A + \epsilon)^{-1} + B^{-1})(A^{-1} + B^{-1})^{-1/2}$ has some diagonal element larger than 1.

Proof. As ϵ is negative definite, we start with

$$A + \epsilon \prec A \implies (A + \epsilon)^{-1} \succ A^{-1} \implies (A + \epsilon)^{-1} = A^{-1} + \bar{\epsilon} \tag{13}$$

for some $\bar{\epsilon} \succ 0$. Using this in the definition of Σ, we have

$$\begin{aligned}
\Sigma &= (A^{-1} + B^{-1})^{-1/2}(A^{-1} + \bar{\epsilon} + B^{-1})(A^{-1} + B^{-1})^{-1/2} \\
&= (A^{-1} + B^{-1})^{-1/2}(A^{-1} + B^{-1})(A^{-1} + B^{-1})^{-1/2} + (A^{-1} + B^{-1})^{-1/2}(\bar{\epsilon})(A^{-1} + B^{-1})^{-1/2} \\
&= I + (A^{-1} + B^{-1})^{-1/2}(\bar{\epsilon})(A^{-1} + B^{-1})^{-1/2} \\
\overset{(a)}{\implies} \Sigma &\succ I \implies \Sigma - I \succ 0.
\end{aligned} \tag{14}$$

Here, (a) uses Claim 3 with a choice of $P = (A^{-1} + B^{-1})^{-1/2} \succ 0, Q = \bar{\epsilon} \succ 0$. Then, by Claim 4, we have that Equation (14) implies that some diagonal element of Σ is greater than 1. □

A.3 Linear Regression and OLS

Notations. Denote by $\mathcal{N}(\mu, \Sigma)$ the multi-variate Gaussian distribution with mean μ and covariance Σ, and write short-hand $\mathcal{N}(\Sigma) = \mathcal{N}(0, \Sigma)$. Let $\eta(\Sigma)$ be a random vector that follows $\mathcal{N}(\Sigma)$.

Problem. Consider a linear regression problem on an unknown parameter $\theta^* \in \mathbb{R}^d$, with data points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where $y_i = \langle x_i, \theta^* \rangle + \eta_i$ for all i, where η_is are i.i.d. $\mathcal{N}(0, 1)$. Write $V = \sum_{i=1}^n x_i x_i^\top = X^\top X$ where $X \in \mathbb{R}^{n \times d}$ comprises of row vectors x_is, similarly, $B = \sum_{i=1}^n x_i y_i = X^\top Y$, where $Y \in \mathbb{R}^{n \times 1}$ comprises of rows of single rewards y_is. Then, $Y = X\theta^* + \mathcal{N}(0, I_n)$. Additionally denote by $\mathrm{proj}_{(A)}(v)$ the projection of vector v onto the subspace spanned by the columns of A.

Claim 6 (OLS estimate)**.** *The Ordinary Least Squares (OLS) estimator $\theta = V^\dagger B = (X^\top X)^\dagger X^\top Y$ follows the multi-variate Gaussian distribution with mean $\mathrm{proj}_{(X^\top X)}(\theta^*)$ and covariance $V^\dagger$. When $X^\top X$ spans $\mathbb{R}^d$, it follows the Gaussian distribution with mean θ^* and covariance V^{-1}.*

Proof. Note that the only source of randomness in θ estimate is the Gaussian noise in rewards y_is observed. By working our way upwards from there, we shall show the required distribution of θ as follows:

$$\begin{aligned}
\theta &= (X^\top X)^\dagger X^\top Y = (X^\top X)^\dagger X^\top (X\theta^* + \eta(I_n)) \\
&= (X^\top X)^\dagger X^\top X\theta^* + (X^\top X)^\dagger X^\top \eta(I_n) \\
&\overset{(a)}{=} \mathrm{proj}_{(X^\top X)}(\theta^*) + (X^\top X)^\dagger X^\top \eta(I_n) \\
&= \mathrm{proj}_{(X^\top X)}(\theta^*) + \eta\left((X^\top X)^\dagger X^\top \cdot I_n \cdot ((X^\top X)^\dagger X^\top)^\top\right) \\
&= \mathrm{proj}_{(X^\top X)}(\theta^*) + \eta\left((X^\top X)^\dagger X^\top X (X^\top X)^\dagger\right) \\
&\overset{(b)}{=} \mathrm{proj}_{(X^\top X)}(\theta^*) + \eta\left((X^\top X)^\dagger\right) \\
&\sim \mathcal{N}\left(\mathrm{proj}_{(X^\top X)}(\theta^*), (X^\top X)^\dagger\right) \\
&\overset{(c)}{=} \mathcal{N}\left(\mathrm{proj}_{(X^\top)}(\theta^*), (X^\top X)^\dagger\right). \qquad (15)
\end{aligned}$$

Here, (a) is from the following: for a positive semi-definite A, vector $A^\dagger Ax$ is the projection of vector x onto column space of A. Equality (b) is due to the 'weak inverse' property of Moore-Penrose pseudo-inverses. This completes the proof. Additionally, (c) is due to the fact that column space of $X^\top X$ equals column space of $X^\top$.

Continuing from Equation (15), if the matrix (of regressors) X has full rank with columns of $X^\top$ spanning $\mathbb{R}^d$, then $X^\top X$ is invertible and we have

$$\theta \sim \mathcal{N}\left(\theta^*, (X^\top X)^{-1}\right). \qquad (16)$$

Equations (15) and (16) show the Claim. □

B Related Work

In this section, we present a more elaborate survey of the related work, complementing Section 1.2 in the main paper.

Stochastic linear bandits. The stochastic multi-armed bandits problem has a rich history of development dating back to the work of Thompson (1933). The seminal result of Auer et al. (2002) shows that the UCB algorithm which is based on the principle of 'optimism in the face of uncertainty' achieves a $\tilde{O}(\sqrt{KT})$ regret bound for K-armed bandits. The linear bandits problem has also

received significant attention due to applications where each arm/action has side information which can be utilized in reward estimation (Wang et al., 2021; Moradipari et al., 2021). In the stochastic linear bandits problem, each action is embedded in a d dimensional space and the rewards are a linear (stochastic) function of these action embeddings. The principle of optimism has been extended to this setting to obtain a $\tilde{O}(\sqrt{dT \ln K})$ regret bound (Chu et al., 2011), and a $\tilde{\Theta}(d\sqrt{T})$ regret bound (Dani et al., 2008). The latter bound is independent of the number of actions K and allows for an infinite number of arms. Finally, Abbasi-Yadkori et al. (2011) proposed the OFUL algorithm (also loosely referred to as LinUCB) which is based on the same principle of optimism and achieves a bound of $\tilde{O}(d\sqrt{T})$ even when the (possibly infinite) set of actions changes over time.

Multi-agent settings. Multi-agent multi-armed bandits have similarly been studied under different settings. Motivated by applications in wireless communication, Kalathil et al. (2014) study a multi-agent bandits with 'collision' problem. In this, if several agents play the same arm at any given time (collide), then these agents receive no reward (or sometimes a shared reward). The agents are oblivious to each other's presence with no explicit/reliable communication channel, and perceive the presence of other agents only through such collisions from their observed rewards.

Many other works have considered variations of this basic setup, for example, allowing agents to enter and exit at different times (Rosenski et al., 2016), and allowing arms to be heterogeneous (with different reward means to different agents) (Bistritz & Leshem, 2018), explicitly informing agents of collision occurrence (Bubeck et al., 2020), and designed algorithms with regret guarantees.

Another rich line of work assumes agents are co-operative and communicate via an orchestrating Principal to share helpful information towards achieving a common goal such as regret minimization, best-arm identification etc. Hillel et al. (2013) introduced collaborative learning in stochastic bandits in the the best-arm-identification (or pure exploration) problem. With M agents who can get together only once to communicate (i.e., $R = 2$ rounds[3] of play), a $\sqrt{M}$ speed-up (in terms of sample complexity) per agent is achieved. Tao et al. (2019) formalize this setting further and generalize this result to achieve a speed-up of $M^{R/R-1}$ with R rounds of play. Wang et al. (2019) comprehensively cover the regret minimization problem in multi-agent (or distributed) linear bandits with communication guarantees. They show an arm elimination-based protocol for a fixed action set that achieves $\tilde{O}(d\sqrt{TM})$ regret (overall across all M agents). Further, similar to our setting, they also look at time-varying action sets and propose a Distributed Linear UCB (DisLinUCB) algorithm that extends the OFUL algorithm of Abbasi-Yadkori et al. (2011) and achieves a regret bound of $\tilde{O}(d\sqrt{MT})$. DisLinUCB is shown to have a communication cost of $O(M^{1.5}d^3)$ that is independent of T by using a specific dynamic condition to trigger communication events. We juxtapose our setting and results with very relevant and key pieces of works from the literature in Table 1.

In addition, there have been several variants and extensions of the communication model, for example, peer-peer messaging (instead of a full broad-cast) (Korda et al., 2016; Szorenyi et al., 2013), restricted message sizes (Sankararaman et al., 2019) which prohibits sharing information about *all* arms, and asynchronous communication (Li & Wang, 2022; He et al., 2022) where each agent independently and individually communicates with the Principal at different times based on their local information, as opposed to having a dedicated communication event for all agents.

Furthermore, several works have considered environments that are heterogeneous to different agents, i.e., where each agent, say a, faces a different bandit instance, like, different linear models θ_a^* (Moradipari et al., 2022; Ghosh et al., 2022) or sets of independent arms (Shi et al., 2021; Karpov & Zhang, 2022).

Finally, there has also been work on the collaborative multi-armed bandits in the presence of adversarial agents who actively try to disrupt the collaboration (Awerbuch & Kleinberg, 2008; Vial et al., 2021; Mitra et al., 2022). The work of Mitra et al. (2022) considers a linear bandit setup similar to ours. However, there are differences in motivation/techniques between our work and Mitra et al.

[3]Typically, the word time-step and round are used interchangeably to denote one arm play and observation. However, here, we reserve *round* to mean the series of time-steps between two communication events.

(2022): they design an algorithm that is robust to adversarial agents by computing robust reward statistics, while we consider the design of mechanisms for incentivizing strategic agents to perform sufficient exploration and report truthfully.

Strategic agents. A more recent theme in multi-armed bandits is to study strategic behavior of agents in a collaborative environment. In collaborative learning setups, the presence of such strategic agents brings about a notion of 'free-riding' where an agent gains from the effort of others and doesn't contribute back to the collaboration. Specifically, if an agent x is permitted to observe the actions, contexts, and rewards of other agents (all of whom are assumed to be running no-regret algorithms themselves) at every time-step, Jung et al. (2020) show that x can achieve $O(1)$ regret (w.r.t. T, but with an $1/\Delta$ factor otherwise) by estimating its own reward means from observations from other agents alone. They further argue that the knowledge of other agents' contexts are necessary for constant order regret by showing an $\Omega(\log T)$ lower bound in that case.

In the multi-agent bandits with 'collision' problem, if several agents play the same arm (collide), the reward from that arm is 'shared' among those agents. Xu et al. (2023); Boursier & Perchet (2020) show existence of ε-Nash Equilibrium behaviours with regret guarantees in such settings. Wei et al. (2023) suppose that every agent incurs some private cost to share information with the Principal and compares it against its utility from getting information from other agents and consequently decides if it should participate in data sharing or not. They show that there exists private cost sequences which make any data sharing mechanism of the Principal not individually rational to participate in. To overcome this, Wei et al. (2023; 2024) propose additional monetary payouts (that add to an agent's utility) based on honestly reported private costs of the agents to make all agents participate and achieve optimal $O\left(d/\sqrt{TM}\right)$ per-agent per-time-step regret rate.

Beyond multi-armed bandits, a closely related line of work (Karimireddy et al., 2022; Murhekar et al., 2024) considers mechanisms to incentivize data sharing in a supervised learning setting. However, these works consider the cost of acquiring and sharing data and designs mechanisms with payments to incentivize agents towards desired behaviour of acquiring and sharing data.

C Technical Lemmas

In this section, we restate some lemmas from the literature that we shall use.

Lemma 7 (Lemma 11 of Abbasi-Yadkori et al. (2011))**.** *Write* $V_t := \gamma I + \sum_{s=1}^{t-1} x_s x_s^\top$ *for all* $t \in [T]$*, where* $\|x_s\|_2 \leq 1$ *for all* $s \in [T]$*. Then,*

$$\sum_{t=1}^{T} \min\left(1, \|x_t\|^2_{V_t^{-1}}\right) \leq 2\log\left(\frac{\det(V_{T+1})}{\det(V_0)}\right) \leq 2d\log\left(\frac{d\gamma + T}{d\gamma}\right) = \widetilde{O}\left(d\log T\right).$$

Lemma 8 (Lemma 12 of Abbasi-Yadkori et al. (2011))**.** *Let* A, B*, and* C *be positive semi-definite matrices such that* $A = B + C$*. Then,*

$$\sup_{x \neq 0} \frac{x^\top A x}{x^\top B x} \leq \frac{\det(A)}{\det(B)}.$$

D Proof of Theorem 1

Proof outline. We want to show that when all other agents are honest, agent a enjoys the highest utility when he is honest in every single communication round. Towards this, we shall prove in Lemmas 10 and 17 that the expected penalty of an agent is $O(1)$ if he is honest and in Lemmas 13 and 19 that it doesn't converge (tends to $+\infty$) if he is dishonest or free-rides. This result primarily hinges on the principal's ability to distinguish between an honest and dishonest agent, probabilistically, from their reported parameter estimates $\theta^{\varepsilon}_{a,\tau}$s without actually knowing the true parameter value θ^*. Building on the intuition that a free-rider's parameter estimate $\theta^{\varepsilon}_{a,\tau}$ will be 'poorer' than

that of honest agents' as they don't perform the requisite exploration, the principal computes a specific statistical quantity $z_a \in \mathbb{R}^d$ (in line 9) by comparing the estimate of agent a with the aggregate estimate of all other agents' new statistics that is yet to be given to agent a. We shall show that this z_a doesn't depend on the unknown θ^*, and conditioned on the honesty of all other agents, z_a follows a 0-mean Gaussian distribution with all components having a marginal variance 1 if a was honest, but if a was dishonest and free-riding, some component's marginal variance goes strictly above 1 (as in Claim 14 and Claim 20). Taking advantage of this distinction, the principal computes the penalty of agent a as function of these components of z_a (in line 11) specifically designed such that it's expected value converges to a constant for variance 1 (for an honest agent) and doesn't converge for variance more than 1 (for a dishonest agent).

Theorem 1. *Let the principal follow* MonPen *in common knowledge and agents adhere to Assumption 2. Then, the strategy profile of all agents being honest, $s^* = ((\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0) : \forall \tau \in \mathcal{T})_{a\in\mathcal{A}}$, is a Nash Equilibrium of the principal-agents bandit game Γ.*

Proof of Theorem. We shall prove the Theorem by induction over the ordinality of communication events τ_is.

Statement. Let $H(i)$ denote the following: When all other agents are honest, i.e., follow strategy profile $((\varepsilon^\theta_{b,\tau} = 0, \varepsilon^V_{b,\tau} = 0) : \tau \in \mathcal{T})_{b\in\mathcal{A}\setminus\{a\}}$,

1. The results matrix recovered for all agents $b \in \mathcal{A} \setminus \{a\}$ (in Equation (6)) and other-agents OLS estimate for agent a computed (in Equation (8)) by the principal are accurate. That is,

$$B^{s,\varepsilon}_{b,\tau_i} = B^s_{b,\tau_i} \quad \text{and} \quad \theta^{c,\varepsilon}_{a,\tau_{i-1}:\tau_i} = \left(V^c_{a,\tau_{i-1}:\tau_i}\right)^{-1}\left(B^c_{a,\tau_{i-1}:\tau_i}\right).$$

2. It is the best response for agent a to follow a strategy in which, in the communication event τ_i, he is honest, i.e., $(\varepsilon^\theta_{a,\tau_i} = 0, \varepsilon^V_{a,\tau_i} = 0)$.

Write short-hands $\varepsilon_{-a} = 0$ to denote $((\varepsilon^\theta_{b,\tau} = 0, \varepsilon^V_{b,\tau} = 0) : \tau \in \mathcal{T})_{b\in\mathcal{A}\setminus\{a\}}$, and $\varepsilon_{a,\tau} = 0$ to denote $(\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0)$.

Base case proof. The first part of $H(1)$ is shown by the following claim:

Claim 9. *When all other agents are honest, i.e., $\varepsilon_{-a} = 0$, then, in the first communication event $\tau = \tau_1$, we have $B^{s,\varepsilon}_{b,\tau} = B^s_{b,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau} = \left(V^c_{a,\tau}\right)^{-1}\left(B^c_{a,\tau}\right)$.*

Proof. Next, for all other agents $b \in \mathcal{A} \setminus \{a\}$, the self-play statistics recovered by the principal in Equation (6) obeys

$$\begin{aligned} B^{s,\varepsilon}_{b,\tau} &= (V^s_{b,\tau} + V^c_{b,\tau'})\theta^\varepsilon_{b,\tau} - B^{c,\varepsilon}_{b,\tau'} \overset{(a)}{=} (V^s_{b,\tau})\theta^\varepsilon_{b,\tau} \overset{(b)}{=} (V^s_{b,\tau})\theta_{b,\tau} \\ &\overset{(c)}{=} (V^s_{b,\tau})\left(V^{s,\varepsilon}_{b,\tau} + V^c_{b,\tau'}\right)^\dagger\left(B^s_{b,\tau} + B^{c,\varepsilon}_{b,\tau'}\right) \\ &\overset{(d)}{=} (V^s_{b,\tau})\left(V^{s,\varepsilon}_{b,\tau}\right)^\dagger\left(B^s_{b,\tau}\right) \overset{(e)}{=} (V^s_{b,\tau})\left(V^s_{b,\tau}\right)^{-1}\left(B^s_{b,\tau}\right) = B^s_{b,\tau}, \end{aligned} \tag{17}$$

where (b), (e) respectively use $\theta^\varepsilon_{b,\tau} = \theta_{b,\tau}$ and $V^{s,\varepsilon}_{b,\tau} = V^s_{b,\tau}$ from lemma statement condition $\varepsilon_{-a} = 0$, and (a),(d) are due to $B^{c,\varepsilon}_{b,\tau'} = 0, V^c_{a,\tau'} = 0$ as there is no collaboration statistics yet ($\tau' = 0$ pedantically), and (c) is from Equation (5) with time-steps t, τ replaced with τ, τ' respectively.

Then, OLS estimate from the recovered data of other agents since the last communication round obeys

$$\theta^{c,\varepsilon}_{a,\tau':\tau} = \left(V^c_{a,\tau':\tau}\right)^{-1} B^{c,\varepsilon}_{a,\tau':\tau} \overset{\text{(a)}}{=} \left(V^c_{a,\tau}\right)^{-1} B^{c,\varepsilon}_{a,\tau} = \left(V^c_{a,\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s,\varepsilon}_{b,\tau}$$

$$\overset{\text{(b)}}{=} \left(V^c_{a,\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s}_{b,\tau} = \left(V^c_{a,\tau}\right)^{-1} \left(B^c_{a,\tau}\right) \tag{18}$$

$$\overset{\text{(c)}}{\sim} \mathcal{N}\left(\theta^*, (V^c_{a,\tau})^{-1}\right), \tag{19}$$

where, (a) is because τ is the first communication event (uses $\tau' = 0$), and (b) is due to Equation (17), and (c) uses Claim 6. Equations (17) and (18) show the Claim. □

To complete the proof for $H(1)$ (to show the second statement), what remains to be shown is that in the first communication event, when an agent a does *not* have any statistics from the other agents, it is the best response (in terms of utility maximization) for agent a to be honest, i.e, $\varepsilon_{a,\tau} = 0$, where $\tau = \tau_1$ is the first communication event.

To show this, we shall first establish, in Lemma 10, that the expected penalty of agent a converges to a small constant value when $\varepsilon_{a,\tau} = 0$, i.e., $\varepsilon^\theta_{a,\tau} = 0$ and $\varepsilon^V_{a,\tau} = 0$. Then, we shall establish, in Lemma 13, that agent a's expected penalty, however, fails to converge (and tends to infinity) when agent a unilaterally defects with $\varepsilon_{a,\tau} \neq 0$, i.e., $\varepsilon^\theta_{a,\tau} \neq 0$ or $\varepsilon^V_{a,\tau} \prec 0$.

Lemma 10 (Penalty of Honest Agent)**.** $\mathbb{E}\left[P_{a,\tau} \middle| (\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0), \varepsilon_{-a} = 0\right] = O(1)$.

Proof. We navigate through the sequence of steps in principal's penalty-levying procedure and arrive at the final expected penalty.

The agent a's reported OLS estimate obeys

$$\theta^\varepsilon_{a,\tau} = \theta_{a,\tau} + \varepsilon^\theta_{a,\tau} \overset{\text{(a)}}{=} \theta_{a,\tau} \overset{\text{(b)}}{=} \left(V^{s,\varepsilon}_{a,\tau} + V^c_{a,\tau'}\right)^\dagger \left(B^{s,\varepsilon}_{a,\tau} + B^{c,\varepsilon}_{a,\tau'}\right) \overset{\text{(c)}}{=} \left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} \left(B^s_{a,\tau} + B^{c,\varepsilon}_{a,\tau'}\right)$$

$$\overset{\text{(d)}}{=} \left(V^s_{a,\tau}\right)^{-1} \left(B^s_{a,\tau}\right) \overset{\text{(e)}}{\sim} \mathcal{N}\left(\theta^*, (V^s_{a,\tau})^{-1}\right), \tag{20}$$

where (a), (c) use lemma statement conditions $\varepsilon^\theta_{a,\tau} = 0$ and $\varepsilon^V_{a,\tau} = 0$ respectively, (b) is from Equation (5) with time-steps t, τ replaced with τ, τ' respectively, (d) is due to $B^{c,\varepsilon}_{a,\tau'} = 0, V^c_{a,\tau'} = 0$ as τ is the first communication event and there is no collaboration statistics yet ($\tau' = 0$), and finally, (e) uses Claim 6.

Claim 11. *Under the assumptions of the Lemma, the quantity z_a computed by principal in Equation (9) obeys $z_a \sim \mathcal{N}(0, I_d)$.*

Proof. The two OLS estimates $\theta^\varepsilon_{a,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau':\tau}$ are independent as the results matrices used $B^s_{a,\tau}$ and $B^c_{a,\tau}$ (as in Equations (20) and (19)) contain mutually exclusive observed rewards, and the observed rewards are themselves independent amongst each other. Thus, their difference also follows a Gaussian distribution of the following form:

$$\theta^\varepsilon_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau} \sim \mathcal{N}\left(0, (V^s_{a,\tau})^{-1} + (V^c_{a,\tau})^{-1}\right). \tag{21}$$

Subsequently, we have

$$\begin{aligned}
z_a &= \left(\left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} + \left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau}\right) \\
&= \left(\left(V^s_{a,\tau}\right)^{-1} + \left(V^c_{a,\tau}\right)^{-1}\right)^{-1/2}\left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau}\right) \\
&\overset{(a)}{\sim} \mathcal{N}\left(0, \left(\left(V^s_{a,\tau}\right)^{-1} + \left(V^c_{a,\tau}\right)^{-1}\right)^{-1/2}\left(\left(V^s_{a,\tau}\right)^{-1} + \left(V^c_{a,\tau}\right)^{-1}\right)\left(\left(V^s_{a,\tau}\right)^{-1} + \left(V^c_{a,\tau}\right)^{-1}\right)^{-1/2}\right) \\
&= \mathcal{N}\left(0, I_d\right),
\end{aligned}$$

where (a) uses Equation (21). □

Finally, we show in Claim 12 that such a z_a leads to the agent having constant penalty to complete the proof of the Lemma.

Claim 12. $\mathbb{E}\left[P_{a,\tau} | z_a \sim \mathcal{N}\left(0, I_d\right)\right] = O(1)$.

Proof. Write $z_a = (z^1, \ldots, z^d)$ as a vector of its components. As $z_a \sim \mathcal{N}\left(0, I_d\right)$, all these components, (for all $i \in [d]$) $z^i \sim \mathcal{N}\left(0, 1\right)$, are i.i.d. standard normal random variables, and consequently, $x = (z^i)^2$ (as in line 11) follows the chi-squared distribution with 1 degree of freedom. Denote by f_X the probability density function (p.d.f.) of x, and by $P(x)$ the penalty realised for a specific x (as in line 11). For any $i \in [d]$, the expected value of component-wise penalty P^i is given by

$$\begin{aligned}
\mathbb{E}\left[P^i \middle| z_a \sim \mathcal{N}\left(0, I_d\right)\right] &= \int_{-\infty}^{\infty} P(x) \cdot f_X(x)\, dx \\
&\overset{(a)}{=} \int_0^{\infty} \left(\frac{\sqrt{2\pi x}}{(x+1)^{1+c_2}} \exp\left(\frac{x}{2}\right)\right) \times \left(\frac{1}{\sqrt{2\pi x}} \exp\left(\frac{-x}{2}\right)\right) dx - \frac{1}{c_2} \\
&= \int_0^{\infty} \frac{1}{(x+1)^{1+c_2}}\, dx - \frac{1}{c_2} = 0 = O(1).
\end{aligned}$$

Here, (a) uses the component-wise penalty function in Line 11 (Algorithm 3) for $P(x)$, and uses the p.d.f. in Claim 1 for $f_X(x)$. Thus, the expectation of the total penalty $P_{a,\tau} = \frac{1}{d}\sum_{i\in[d]} P^i$ belongs to $O(1)$. □

This completes the proof of the Lemma. □

Lemma 13 (Penalty of Dishonest Agent)**.** $\mathbb{E}\left[P_{a,\tau} | \varepsilon_{a,\tau} \neq 0, \varepsilon_{-a} = 0\right]$ *does not converge/exist (tends to* $+\infty$*).*

Proof. We shall show this lemma in two steps. First, we shall obtain an expression the d-dimensional vector z_a (computed in line 9) and arrive at its covariance and bias based on the $\varepsilon^{\theta}_{a,\tau}, \varepsilon^{V}_{a,\tau}$ values in Claim 14. Second, we shall argue based on this covariance and bias that the component-wise penalty (computed in line 11) does not converge/tends to infinity for some component P^i, in Claim 15.

The agent a's reported OLS estimate obeys

$$\begin{aligned}
\theta^{\varepsilon}_{a,\tau} = \theta_{a,\tau} + \varepsilon^{\theta}_{a,\tau} &\overset{(a)}{=} \left(V^{s,\varepsilon}_{a,\tau} + V^c_{a,\tau'}\right)^{\dagger}\left(B^{s,\varepsilon}_{a,\tau} + B^{c,\varepsilon}_{a,\tau'}\right) + \varepsilon^{\theta}_{a,\tau} \\
&\overset{(b)}{=} \left(V^{s,\varepsilon}_{a,\tau}\right)^{\dagger}\left(B^{s,\varepsilon}_{a,\tau}\right) + \varepsilon^{\theta}_{a,\tau} \overset{(c)}{\sim} \mathcal{N}\left(\mathrm{proj}_{\left(V^{s,\varepsilon}_{a,\tau}\right)}\left(\theta^*\right) + \varepsilon^{\theta}_{a,\tau}, \left(V^{s,\varepsilon}_{a,\tau}\right)^{\dagger}\right),
\end{aligned} \tag{22}$$

where (a) is from Equation (5) with time-steps t, τ replaced with τ, τ' respectively, (b) is due to $B^{c,\varepsilon}_{a,\tau'} = 0, V^c_{a,\tau'} = 0$ as τ is the first communication event and there is no collaboration statistics yet ($\tau' = 0$), and finally, (c) uses Claim 6.

With the agent's estimate $\theta_{a,\tau}$ behaving as such, we inquire about the consequent distribution of the the z_a quantity next.

Claim 14. *Under the assumption $\varepsilon_{a,\tau} \neq 0$ of the Lemma, i.e., either $\varepsilon^{\theta}_{a,\tau} \neq 0$ or $\varepsilon^{V}_{a,\tau} \prec 0$, the quantity z_a computed by principal in Equation (9) obeys $z_a \sim \mathcal{N}(\mu, \Sigma)$ such that either $\mu \neq 0$ or some diagonal element of Σ is larger than* 1.

Proof. The two OLS estimates $\theta^{\varepsilon}_{a,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau':\tau}$ are independent as the results matrices used $B^{s}_{a,\tau}$ and $B^{c}_{a,\tau}$ (as in Equations (22) and (19)) contain mutually exclusive observed rewards, and the observed rewards are themselves independent amongst each other. Thus, their difference also follows a Gaussian distribution of the following form:

$$\begin{aligned}\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau} &\sim \mathcal{N}\left(\mathrm{proj}_{(V^{s,\varepsilon}_{a,\tau})}(\theta^*) + \varepsilon^{\theta}_{a,\tau} - \theta^*, (V^{s,\varepsilon}_{a,\tau})^{\dagger} + (V^{c}_{a,\tau})^{-1}\right) \\ &\overset{(a)}{=} \mathcal{N}\left(\mu', (V^{s}_{a,\tau} + \varepsilon^{V}_{a,\tau})^{\dagger} + (V^{c}_{a,\tau})^{-1}\right), \end{aligned} \tag{23}$$

where in (a) we introduced $\mu' = \mathrm{proj}_{(V^{s,\varepsilon}_{a,\tau})}(\theta^*) + \varepsilon^{\theta}_{a,\tau} - \mathrm{proj}_{(V^{c}_{a,\tau})}(\theta^*)$. Observe that $\mu' \neq 0$ for any $\varepsilon^{\theta}_{a,\tau}$ as the agent cannot set his $\varepsilon^{\theta}_{a,\tau}$ as a function of the unknown θ^* (by our problem Assumption 2) to negate the $\mathrm{proj}_{(V^{s,\varepsilon}_{a,\tau})}(\theta^*) - \theta^*$ term. Thus, we have $\mu' = 0$ only if (columns of) $V^{s,\varepsilon}_{a,\tau}$ spans $\mathbb{R}^d$, in which case $V^{s,\varepsilon}_{a,\tau}$ is invertible. Subsequently, we have

$$\begin{aligned} z_a &= \left(\left(V^{s}_{a,\tau} + V^{c}_{a,\tau'}\right)^{-1} + (V^{c}_{a,\tau':\tau})^{-1}\right)^{-1/2} \left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau}\right) \\ &= \left(\left(V^{s}_{a,\tau}\right)^{-1} + (V^{c}_{a,\tau})^{-1}\right)^{-1/2} \left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau}\right) \\ &\overset{(a)}{\sim} \mathcal{N}\left(\mu'', \left(\left(V^{s}_{a,\tau}\right)^{-1} + (V^{c}_{a,\tau})^{-1}\right)^{-1/2} \left((V^{s}_{a,\tau} + \varepsilon^{V}_{a,\tau})^{-1} + (V^{c}_{a,\tau})^{-1}\right) \left(\left(V^{s}_{a,\tau}\right)^{-1} + (V^{c}_{a,\tau})^{-1}\right)^{-1/2}\right), \end{aligned} \tag{24}$$

where (a) uses Equation (23), and introduces $\mu'' = \left(\left(V^{s}_{a,\tau}\right)^{-1} + (V^{c}_{a,\tau})^{-1}\right)^{-1/2} \mu' \neq 0$ when $\varepsilon^{\theta}_{a,\tau} \neq 0$.

Write Σ to denote the covariance term of z_a as in Equation (24). Then, from Lemma 5 with a choice of $A = V^{s}_{a,\tau}, B = V^{c}_{a,\tau}$, and $\epsilon = \varepsilon^{V}_{a,\tau}$, it follows that some diagonal element of Σ is greater than 1. □

From Claim 14, it is seen that when an agent a is dishonest/free-rides, i.e., $\varepsilon^{\theta}_{a,\tau} \neq 0$ or $\varepsilon^{V}_{a,\tau} \prec 0$, the intermediate quantity z_a becomes biased (non-zero mean) or some component of z_a has a marginal variance greater than 1. Finally, we show in Claim 15 that such a z_a leads to the agent having an unbounded penalty.

Claim 15. $\mathbb{E}[P_{a,\tau} | z_a \sim \mathcal{N}(\mu, \Sigma)]$ *does not converge/exist (tends to $+\infty$) when $\mu \neq 0$ or Σ has some diagonal element larger than* 1.

Proof. Write $\mu = (\mu'')^i$ and $\sigma^2 = \Sigma^{ii}$ to be the mean and variance of component z^i for some $i \in [d]$ which satisfies $\mu \neq 0 \vee \sigma^2 > 1$ from Claim 14. Note that $x = (z^i)^2$ follows the generalized chi-squared distribution with 1 degree of freedom. Denote by f_X the probability density function (p.d.f.) of x and by $P(x)$ the penalty realised for a certain x (as in line 11). The expected component-specific penalty P^i is given by

$$\begin{aligned} &\mathbb{E}\left[P^i \middle| \left(\varepsilon^{\theta}_{a} \neq 0 \vee \varepsilon^{V}_{a} \prec 0\right), \varepsilon_{-a} = 0\right] = \int_{-\infty}^{\infty} P(x) f_X(x) dx \\ &\overset{(a)}{=} \int_{0}^{\infty} \left(\frac{\sqrt{2\pi x}}{(x+1)^{1+c_2}} \exp\left(\frac{x}{2}\right)\right) \cdot \left(\frac{1}{2\sigma\sqrt{2\pi x}} \exp\left(\frac{-x-\mu^2}{2\sigma^2}\right) \left[\exp\left(\frac{\sqrt{x}\mu}{\sigma^2}\right) + \exp\left(\frac{-\sqrt{x}\mu}{\sigma^2}\right)\right]\right) dx \\ &= \int_{0}^{\infty} \left[\frac{1}{2\sigma(x+1)^{1+c_2}}\right] \cdot \left[\exp\left(\frac{x}{2}\left(1 - \frac{1}{\sigma^2} - \frac{\mu^2}{x\sigma^2}\right)\right)\right] \cdot \left[\exp\left(\frac{\sqrt{x}\mu}{\sigma^2}\right) + \exp\left(\frac{-\sqrt{x}\mu}{\sigma^2}\right)\right] dx, \end{aligned} \tag{25}$$

where, (a) uses the component-wise penalty function in Line 11 (Algorithm 3), and uses the p.d.f. in Claim 1 for $f_X(x)$.

In Equation (25), the integral can be seen to diverge as $\sigma > 1$ or $\mu \neq 0$. Specifically, when $\sigma > 1$, the second term does not vanish for any value of μ. When $\sigma \leq 1$, any $\mu \neq 0$ makes the third term not vanish. Thus, the expectation of penalty $P_a = \frac{1}{d}\sum_i P^i$ does not converge. □

This completes the proof of the Lemma about the divergence of the expected penalty of an agent that is dishonest/free-rides. □

From Lemmas 10 and 13, we see that it is the best response for an agent to have $\varepsilon_{a,\tau} = 0$ compared to any other $\varepsilon_a \neq 0$. This completes the proof of the Base case.

Induction hypothesis. We assume $H(k)$ holds for some $k \in [q]$. One of its consequences is

$$B^{s,\varepsilon}_{b,\tau'} = B^{s}_{b,\tau'} \text{ for all } b \in \mathcal{A} \setminus \{a\}, \tag{26}$$

$$B^{c,\varepsilon}_{a,\tau'} = \sum_{b \in \mathcal{A}\setminus\{a\}} B^{s,\varepsilon}_{b,\tau'} = \sum_{b \in \mathcal{A}\setminus\{a\}} B^{s}_{b,\tau'} = B^{c}_{a,\tau'}, \tag{27}$$

where $\tau' = \tau_k$.

Induction step. All results in the induction step implicitly assume that $H(k)$ hold. We now proceed to show $H(k+1)$ holds. The line of proof is very similar to that of the Base case. Write $\tau = \tau_{k+1}$ and $\tau' = \tau_k$ to be time-steps of the current and preceding communication events respectively. The first part of $H(k+1)$ is shown by the following claim:

Claim 16. *When all other agents are honest, i.e., $\varepsilon_{-a} = 0$, then, in the communication event $\tau = \tau_{k+1}$, we have $B^{s,\varepsilon}_{b,\tau} = B^{s}_{b,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau':\tau} = \left(V^{c}_{a,\tau':\tau}\right)^{-1}\left(B^{c}_{a,\tau':\tau}\right)$.*

Proof. For all other agents $b \in \mathcal{A} \setminus \{a\}$, the self-play statistics recovered by the principal in Equation (6) obeys

$$\begin{aligned}
B^{s,\varepsilon}_{b,\tau} &\overset{(a)}{=} (V^{s}_{b,\tau} + V^{c}_{b,\tau'})\theta^{\varepsilon}_{b,\tau} - B^{c,\varepsilon}_{b,\tau'} \overset{(b)}{=} (V^{s}_{b,\tau} + V^{c}_{b,\tau'})\left(V^{s,\varepsilon}_{b,\tau} + V^{c}_{b,\tau'}\right)^{\dagger}\left(B^{s}_{b,\tau} + B^{c,\varepsilon}_{b,\tau'}\right) - B^{c,\varepsilon}_{b,\tau'} \\
&\overset{(c)}{=} (V^{s}_{b,\tau} + V^{c}_{b,\tau'})\left(V^{s}_{b,\tau} + V^{c}_{b,\tau'}\right)^{-1}\left(B^{s}_{b,\tau} + B^{c,\varepsilon}_{b,\tau'}\right) - B^{c,\varepsilon}_{b,\tau'} \\
&\overset{(d)}{=} (V^{s}_{b,\tau} + V^{c}_{b,\tau'})\left(V^{s}_{b,\tau} + V^{c}_{b,\tau'}\right)^{-1}\left(B^{s}_{b,\tau} + B^{c}_{b,\tau'}\right) - B^{c}_{b,\tau'} \\
&= \left(B^{s}_{b,\tau} + B^{c}_{b,\tau'}\right) - B^{c}_{b,\tau'} = B^{s}_{b,\tau}.
\end{aligned} \tag{28}$$

Here, (a), (c) use $\theta^{s,\varepsilon}_{b,\tau} = \theta^{s}_{b,\tau}$ and $V^{s,\varepsilon}_{b,\tau} = V^{s}_{b,\tau}$ respectively, for all $b \in \mathcal{A} \setminus \{a\}$ from the lemma statement condition ε_{-a}, (b) uses Equation (5) with time-steps t, τ replaced with τ, τ' respectively, (d) uses the induction hypothesis/assumption Equation (27) that result matrices were recovered correctly at τ', the preceding communication event.

Then, OLS estimate (as in Equation (8)) from the recovered data of other agents since the last communication round obeys

$$\begin{aligned}
\theta^{c,\varepsilon}_{a,\tau':\tau} &= \left(V^c_{a,\tau':\tau}\right)^{-1} B^{c,\varepsilon}_{a,\tau':\tau} = \left(V^c_{a,\tau',\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s,\varepsilon}_{b,\tau':\tau} \\
&= \left(V^c_{a,\tau':\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} \left(B^{s,\varepsilon}_{b,\tau} - B^{s,\varepsilon}_{b,\tau'}\right) \\
&\overset{\text{(a)}}{=} \left(V^c_{a,\tau':\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} \left(B^{s}_{b,\tau} - B^{s}_{b,\tau'}\right) \\
&= \left(V^c_{a,\tau':\tau}\right)^{-1} \sum_{b\in\mathcal{A}\setminus\{a\}} B^{s}_{b,\tau':\tau} = \left(V^c_{a,\tau':\tau}\right)^{-1} \left(B^c_{a,\tau':\tau}\right) \qquad (29) \\
&\overset{\text{(b)}}{\sim} \mathcal{N}\left(\theta^*, (V^c_{a,\tau':\tau})^{-1}\right), \qquad (30)
\end{aligned}$$

where, (a) uses $B^{s,\varepsilon}_{b,\tau} = B^s_{b,\tau}$ from Equation (28) and $B^{s,\varepsilon}_{b,\tau'} = B^s_{b,\tau'}$ from the induction hypothesis Equation (26). And (b) uses Claim 6.

Equations (28) and (29) show the Claim. □

To complete the proof for $H(k+1)$ (to show the second statement), what remains to be shown is that, it is the best response (in terms of utility maximization) for agent a to be honest, i.e, $\varepsilon_{a,\tau} = 0$, where $\tau = \tau_{k+1}$ is the $k+1$th communication event.

To show this, we shall first establish, in Lemma 17, that the expected penalty of agent a converges to a small constant value when $\varepsilon_{a,\tau} = 0$, i.e., $\varepsilon^\theta_{a,\tau} = 0$ and $\varepsilon^V_{a,\tau} = 0$. Then, we shall establish, in Lemma 19, that agent a's expected penalty, however, fails to converge (and tends to infinity) when agent a unilaterally defects with $\varepsilon_{a,\tau} \neq 0$, i.e., $\varepsilon^\theta_{a,\tau} \neq 0$ or $\varepsilon^V_{a,\tau} \prec 0$.

Lemma 17 (Penalty of Honest Agent)**.** $\mathbb{E}\left[P_{a,\tau} \middle| (\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0), \varepsilon_{-a} = 0\right] = O(1)$.

Proof. We navigate through the sequence of steps in principal's penalty-levying procedure and arrive at the final expected penalty.

The agent a's reported OLS estimate obeys

$$\begin{aligned}
\theta^\varepsilon_{a,\tau} = \theta_{a,\tau} + \varepsilon^\theta_{a,\tau} &\overset{\text{(a)}}{=} \theta_{a,\tau} \overset{\text{(b)}}{=} \left(V^{s,\varepsilon}_{a,\tau} + V^c_{a,\tau'}\right)^\dagger \left(B^{s,\varepsilon}_{a,\tau} + B^{c,\varepsilon}_{a,\tau'}\right) \overset{\text{(c)}}{=} \left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} \left(B^s_{a,\tau} + B^{c,\varepsilon}_{a,\tau'}\right) \\
&\overset{\text{(d)}}{=} \left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} \left(B^s_{a,\tau} + B^c_{a,\tau'}\right) \overset{\text{(e)}}{\sim} \mathcal{N}\left(\theta^*, (V^s_{a,\tau} + V^c_{a,\tau'})^{-1}\right), \qquad (31)
\end{aligned}$$

where (a), (c) use lemma statement conditions $\varepsilon^\theta_{a,\tau} = 0$ and $\varepsilon^V_{a,\tau} = 0$ respectively, (b) is from Equation (5) with time-steps t, τ replaced with τ, τ' respectively, (d) is due to induction assumption Equation (27), and finally, (e) uses Claim 6.

Claim 18. *Under the assumptions of the Lemma, the quantity z_a computed by principal in Equation (9) obeys $z_a \sim \mathcal{N}(0, I_d)$.*

Proof. The two OLS estimates $\theta^\varepsilon_{a,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau':\tau}$ are independent as the results matrices used $B^s_{a,\tau} + B^c_{a,\tau'}$ and $B^c_{a,\tau':\tau}$ (as in Equations (31) and (30)) contain mutually exclusive observed rewards, and the observed rewards are themselves independent amongst each other. Thus, their difference also follows a Gaussian distribution of the following form:

$$\theta^\varepsilon_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau':\tau} \sim \mathcal{N}\left(0, (V^s_{a,\tau} + V^c_{a,\tau'})^{-1} + (V^c_{a,\tau':\tau})^{-1}\right). \qquad (32)$$

Subsequently, we have

$$
\begin{aligned}
z_a &= \left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\left(\theta^{\varepsilon}_{a,\tau}-\theta^{c,\varepsilon}_{a,\tau':\tau}\right)\\
&\overset{(a)}{\sim}\mathcal{N}\left(0,\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\right)\\
&\overset{(b)}{=}\mathcal{N}\left(0,I_d\right),
\end{aligned}
$$

where (a) uses Equation (32), and (b) uses that $V^s_{a,\tau}$ is full rank. □

Finally, from Claim 12, we have (identical to the Base case) that such a $z_a\sim\mathcal{N}(0,I_d)$ leads to the agent having constant penalty, to complete the proof of the Lemma. □

Lemma 19 (Penalty of Dishonest Agent)**.** $\mathbb{E}\left[P_{a,\tau}|\varepsilon_{a,\tau}\neq 0,\varepsilon_{-a}=0\right]$ *does not converge/exist (tends to* $+\infty$*).*

Proof. We shall show this lemma in two steps. First, we shall obtain an expression the d-dimensional vector z_a (computed in line 9) and arrive at its covariance and bias based on the $\varepsilon^{\theta}_{a,\tau},\varepsilon^{V}_{a,\tau}$ values in Claim 20. Second, we shall argue based on this covariance and bias that the component-wise penalty (computed in line 11) does not converge/tends to infinity for some component P^i using Claim 15.

The agent a's reported OLS estimate obeys

$$
\begin{aligned}
\theta^{\varepsilon}_{a,\tau}=\theta_{a,\tau}+\varepsilon^{\theta}_{a,\tau}&\overset{(a)}{=}\left(V^{s,\varepsilon}_{a,\tau}+V^c_{a,\tau'}\right)^{-1}\left(B^{s,\varepsilon}_{a,\tau}+B^{c,\varepsilon}_{a,\tau'}\right)+\varepsilon^{\theta}_{a,\tau}\\
&\overset{(b)}{=}\left(V^{s,\varepsilon}_{a,\tau}+V^c_{a,\tau'}\right)^{-1}\left(B^{s,\varepsilon}_{a,\tau}+B^c_{a,\tau'}\right)+\varepsilon^{\theta}_{a,\tau}\overset{(c)}{\sim}\mathcal{N}\left(\theta^*+\varepsilon^{\theta}_{a,\tau},(V^{s,\varepsilon}_{a,\tau}+V^c_{a,\tau'})^{-1}\right),
\end{aligned}
\tag{33}
$$

where (a) is from Equation (5) with time-steps t,τ replaced with τ,τ' respectively, then (b) is due to induction assumption Equation (27), and finally, (c) uses Claim 6.

With the agent's reported estimate $\theta^{\varepsilon}_{a,\tau}$ behaving as above, we inquire about the consequent distribution of the the z_a quantity next.

Claim 20. *Under the assumption* $\varepsilon_{a,\tau}\neq 0$ *of the Lemma, i.e., either* $\varepsilon^{\theta}_{a,\tau}\neq 0$ *or* $\varepsilon^{V}_{a,\tau}\prec 0$*, the quantity* z_a *computed by principal in Equation (9) obeys* $z_a\sim\mathcal{N}(\mu,\Sigma)$ *such that either* $\mu\neq 0$ *or some diagonal element of* Σ *is larger than* 1.

Proof. The two OLS estimates $\theta^{\varepsilon}_{a,\tau}$ and $\theta^{c,\varepsilon}_{a,\tau':\tau}$ are independent as the results matrices used $B^s_{a,\tau}+B^c_{a,\tau'}$ and $B^c_{a,\tau':\tau}$ (as in Equations (33) and (30)) contain mutually exclusive observed rewards, and the observed rewards are themselves independent amongst each other. Thus, their difference also follows a gaussian distribution of the following form:

$$
\begin{aligned}
\theta^{\varepsilon}_{a,\tau}-\theta^{c,\varepsilon}_{a,\tau':\tau}&\sim\mathcal{N}\left(\theta^*+\varepsilon^{\theta}_{a,\tau}-\theta^*,(V^{s,\varepsilon}_{a,\tau}+V^c_{a,\tau'})^{-1}+(V^c_{a,\tau':\tau})^{-1}\right)\\
&\overset{(a)}{=}\mathcal{N}\left(\varepsilon^{\theta}_{a,\tau},(V^s_{a,\tau}+\varepsilon^{V}_{a,\tau}+V^c_{a,\tau'})^{-1}+(V^c_{a,\tau':\tau})^{-1}\right),
\end{aligned}
\tag{34}
$$

where in (a) we expand $V^{s,\varepsilon}_{a,\tau}=V^s_{a,\tau}+\varepsilon^{V}_{a,\tau}$. Subsequently, we have

$$
\begin{aligned}
z_a&=\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\left(\theta^{\varepsilon}_{a,\tau}-\theta^{c,\varepsilon}_{a,\tau':\tau}\right)\\
&\overset{(a)}{\sim}\mathcal{N}\left(\mu'',\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\left(\left(V^s_{a,\tau}+\varepsilon^{V}_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)\left(\left(V^s_{a,\tau}+V^c_{a,\tau'}\right)^{-1}+\left(V^c_{a,\tau':\tau}\right)^{-1}\right)^{-1/2}\right),
\end{aligned}
\tag{35}
$$

where (a) uses Equation (34), and introduces $\mu'' = \left(\left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} + (V^c_{a,\tau':\tau})^{-1}\right)^{-1/2} \varepsilon^\theta_{a,\tau} \neq 0$ when $\varepsilon^\theta_{a,\tau} \neq 0$.

Write Σ to denote the covariance term of z_a as in Equation (35). Then, (similar to the Base case) Lemma 5 with a choice of $A = V^s_{a,\tau} + V^c_{a,\tau'}$, $B = V^c_{a,\tau':\tau}$, and $\epsilon = \varepsilon^V_{a,\tau}$ shows that some diagonal element of Σ is greater than 1. □

From Claim 20, it is seen that when an agent a is dishonest/free-rides, i.e., $\varepsilon^\theta_{a,\tau} \neq 0$ or $\varepsilon^V_{a,\tau} \prec 0$, the intermediate quantity z_a becomes biased (non-zero mean) or some component of z_a has a marginal variance greater than 1. Then, finally, we see from Claim 15 that such a z_a leads to the agent having an unbounded penalty (similar to the Base case). □

From Lemmas 17 and 19, we see that it is the best response for an agent to have $\varepsilon_{a,\tau} = 0$ compared to any other $\varepsilon_{a,\tau} \neq 0$. This completes the proof of the Induction step; $H(k+1)$ holds.

By mathematical induction, we have that when all other agents are honest, $\varepsilon_{-a} = 0$, then it is the best response for agent a to follow a strategy in which, for all communication events $\tau \in \mathcal{T}$, he is honest $\varepsilon_{a,\tau} = 0$. Thus, it is the best response for agent a to be honest $\varepsilon_a = 0$ (and not unilaterally deviate) when all other agents are $\varepsilon_{-a} = 0$. As this result holds for any agent a, it follows that $\left((\varepsilon^\theta_{a,\tau} = 0, \varepsilon^V_{a,\tau} = 0 : \tau \in \mathcal{T}\right)_{a\in\mathcal{A}}$ is a Nash Equilibrium. □

E Proof of Theorem 2

Proof outline. The key idea is that all (honest) agents shall explore and contribute equally to the collaboration. This holds due to the deterministic nature, across all agents, of the arm playing algorithm in LsHon between two communication events. As the learning only occurs in batches during the times of communication, all agents use the same confidence set $C_{a,\tau}$ for the entire period until the next communication event. Thus, we shall show (in Lemma 22) that choice of arm played (in line 4) is same across all agents leading to all agents maintain an identical design matrix $V_{a,t}$. This shall establish that all agents share the burden of exploration equally. We shall then quantify the benefit of collaboration among M agents as an improvement (reduction) in individual regret by a factor of $\sqrt{M}$ for all agents, by using this fact that all agents contribute equally to the collaboration. The remainder of the proof shall then broadly follows the line of argument in Abbasi-Yadkori et al. (2011) : First, we shall upper bound the instantaneous regret of an agent $r_{a,t}$ at time t, and shall then use it to bound the total regret to get the desired bound of $R_a(T)$.

We now establish some useful lemmas that are required in the proof of the Theorem.

Recollect from Equation (4) the confidence set of agent a at time t (where τ is the time of last communication event) is given by $C_{a,\tau} = \left\{\theta' \in \mathbb{R}^d : \|\theta' - \theta_{a,\tau}\|_{\lambda I + V_{a,\tau}} \leq \beta_t(\delta)\right\}$, where

$$\beta_t(\delta) = \sqrt{2\log\left(\frac{\det(V_{a,t})^{1/2}\det(\lambda I)^{1/2}}{\delta}\right)} + \sqrt{\lambda} = \widetilde{O}\left(\sqrt{d\log(Mt/\delta)}\right). \tag{36}$$

Lemma 21 ('Good event' occurs with high probability)**.** *For $\delta \in (0, 1/M)$, the event*

$$\mathcal{G} = \{\forall a \in \mathcal{A}, \forall t \in [T] : C_{a,t} \ni \theta^*\} \tag{37}$$

occurs with probability at least $1 - \delta M$.

Proof. We have in the single agent setting (by Theorem 2 of Abbasi-Yadkori et al. (2011)) that with probability $1-\delta$, $\theta^* \in C_{a,t}$ at all times t. By taking a union bound for all M agents over the probability of failure (i.e., $\theta^* \notin C_{a,t}$), we arrive at the lemma statement. □

Lemma 22 (Equality of design across agents)**.** *When the Principal follows* MONPEN *and all agents follow* LSHON*, then, for every time-step $t \in [T]$, the design matrix $V^s_{a,t}$ is the same across all agents $a \in \mathcal{A}$.*

Proof. We shall rely on the 'low-switching' aspect of the arm-play algorithm of LSHON to show this result. We shall argue using Claim 23 that after every communication event, all the agents possess identical statistics, and then shall argue that the arm-play is determined by this identical statistics—agnostic to reward noise since the last communication event—such that all agents play the same action in a given time-step, which leads to the design matrices being synchronized across agents always.

Firstly, as all agents follow LSHON and the principal follows MONPEN, the principal shares with every agent a, all other agents' statistics $V^c_{a,\cdot}, B^c_{a,\cdot}$ (note $B^{c,\varepsilon}_{a,\cdot} = B^c_{a,\cdot}$ as all agents are honest).

Consider any two agents $a, b \in \mathcal{A}$. Recall $\mathcal{T} = (\tau_1, \tau_2, \ldots, \tau_q)$ to be the time-steps of the q communication events that occur in the entire time horizon. Additionally, write $\tau_0 = 0, \tau_{q+1} = T$.

Claim 23. *After all communication steps $\tau \in \mathcal{T}$, for any two agents $a, b \in \mathcal{A}$, we have $B_{a,\tau} = B_{b,\tau}$ and $V_{a,\tau} = V_{b,\tau}$.*

Proof. Initially, $B_{a,0} = B_{b,0} = 0$. At any communication event τ, every agent a sends to the principal the true parameter estimate $\theta_{a,\tau}$ due to the Lemma assumption that all agents adhere to LSHON. The principal recovers the self-play result matrix $B^s_{a,\tau}$ accurately (as shown in Induction Statement 1 used in proof of Theorem 1) of every agent and shares back

$$B^c_{a,\tau} = \sum_{b \in \mathcal{A} \setminus \{a\}} B^s_{b,\tau_{k+1}}, \tag{38}$$

the sum of results of all other agents. Thus, every agent a has

$$B_{a,\tau} = B^s_{a,\tau} + B^c_{a,\tau} \overset{(a)}{=} B^s_{a,\tau_{k+1}} + \sum_{b \in \mathcal{A} \setminus \{a\}} B^s_{b,\tau} = \sum_{b \in \mathcal{A}} B^s_{b,\tau}, \tag{39}$$

where (a) uses Equation (38). Now, Equation (39) shows that $B_{a,\tau}$ is the same quantity independent of specific agent a. Thus, $B_{a,\tau} = B_{b,\tau}$. By an identical argument, we have that $V_{a,\tau} = V_{b,\tau}$ for all $\tau \in \mathcal{T}$. □

Next, we show that the designs of agents are identical at *all* time-steps $t \in [T]$.

Statement. Let $H(t)$ denote the hypothesis that $V^s_{a,t'} = V^s_{b,t'}$ for all $0 \le t' \le t$.

Base case. Note that initially $V^s_{a,0} = V^s_{b,0} = 0$. Thus, $H(0)$ is true.

Induction assumption. Assume $H(k)$ is true, we have $V^s_{a,0} = V^s_{b,0}, V^s_{a,1} = V^s_{b,1}, \ldots, V^s_{a,\tau} = V^s_{b,\tau}, \ldots, V^s_{a,k} = V^s_{b,k}$, where τ be the time of the most recent communication event.

Induction step. At time $k+1$, as per LSHON, we have that for every agent a, the confidence set $C_{a,\tau}$ is a function of the design $V_{a,\tau}$ and results $B_{a,\tau}$. We have that $V_{a,\tau} = V_{b,\tau}$ and $B_{a,\tau} = B_{b,\tau}$ from Claim 23. Thus, it follows that $C_{a,\tau} = C_{b,\tau}$. As LSHON's choice of arm to play (line 4) is a deterministic function of this confidence set $C_{a,\tau}$ as of last communication event τ, we have that at time $k+1$, the arm played by agents a and b are equal/same, i.e., $x_{a,k+1} = x_{b,k+1}$. Using this in (b) below, we have

$$V^s_{a,k+1} = V^s_{a,k} + x_{a,k+1}x^\top_{a,k+1} \overset{(a)}{=} V^s_{b,k} + x_{a,k+1}x^\top_{a,k+1} \overset{(b)}{=} V^s_{b,k} + x_{b,k+1}x^\top_{b,k+1} = V^s_{b,k+1}. \tag{40}$$

Here, (a) uses the induction assumption $V^s_{a,k} = V^s_{b,k}$. Then, (40) implies $H(k+1)$ is true. By induction, $H(T)$ is true, i.e., $V^s_{a,t} = V^s_{b,t}$ for all times $t \in [T]$.

As the above result holds for all pairs of agents a, b, this completes the proof of the Lemma. □

Recall $V^{\mathcal{A}}_t := \sum_{a\in\mathcal{A}} V^s_{a,t}$ to be the global/cumulative design at time t. The corollary follows.

Corollary 3. *For all agents a, it holds that $V^{\mathcal{A}}_t = M \cdot V^s_{a,t}$, at all times t.*

We are ready to prove the Theorem now.

Proof of Theorem. Set $\delta = 1/T^2M^2$. Then, event $\mathcal{G}$ (Equation (37)) fails to occur with probability at most $1/MT^2$ as per Lemma 21. As the cumulative/global regret is trivially upper bounded by MT, occurrence of this event only contributes a $O(T^{-1})$ summand to the expected cumulative/social regret, and consequently, to any single agent's expected regret. Thus, it is sufficient to analyse the regret of an agent when $\mathcal{G}$ occurs to show the required result.

The instantaneous regret at time t is given by

$$\begin{aligned}
r_{a,t} &= \langle x^*, \theta^*\rangle - \langle x_{a,t}, \theta^*\rangle \\
&\overset{(a)}{\leq} \max_{\theta\in C_{a,t}} \langle x^*, \theta\rangle - \langle x_{a,t}, \theta^*\rangle \\
&\overset{(b)}{\leq} \max_{\theta\in C_{a,t}} \langle x_{a,t}, \theta\rangle - \langle x_{a,t}, \theta^*\rangle \\
&\overset{(c)}{=} \langle x_{a,t}, \theta'\rangle - \langle x_{a,t}, \theta^*\rangle \\
&= \langle x_{a,t}, \theta' - \theta^*\rangle = \langle x_{a,t}, \theta' - \theta_{a,t}\rangle + \langle x_{a,t}, \theta_{a,t} - \theta^*\rangle\,.
\end{aligned}$$

Here, (a) is due to $\theta^* \in C_{a,t}$ as $\mathcal{G}$ occurs, (b) is due to $x_{a,t}$ being the optimistic (i.e., the expected reward-maximizing) choice w.r.t $C_{a,t}$. Finally, (c) introduces $\theta' := \arg\max_{\theta\in C_{a,t}} \langle x_{a,t}, \theta'\rangle$.

Continuing the analysis by applying Hölder's inequality with a choice of conjugates $\lambda I + V^{\mathcal{A}}_t$ and $\left(\lambda I + V^{\mathcal{A}}_t\right)^{-1}$, we get

$$\begin{aligned}
r_{a,t} \leq& \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} \|\theta' - \theta_{a,t}\|_{\lambda I+V^{\mathcal{A}}_t} + \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} \|\theta_{a,t} - \theta^*\|_{\lambda I+V^{\mathcal{A}}_t} \\
=& \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} \left(\|\theta' - \theta_{a,t}\|_{\lambda I+V^{\mathcal{A}}_t} + \|\theta_{a,t} - \theta^*\|_{\lambda I+V^{\mathcal{A}}_t}\right) \\
\overset{(a)}{\leq}& \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} \cdot \frac{\det\left(\lambda I + V^{\mathcal{A}}_t\right)}{\det(\lambda I + V_{a,\tau})} \left(\|\theta' - \theta_{a,t}\|_{\lambda I+V_{a,\tau}} + \|\theta_{a,t} - \theta^*\|_{\lambda I+V_{a,\tau}}\right) \\
\overset{(b)}{\leq}& \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} \cdot \frac{\det\left(\lambda I + V^{\mathcal{A}}_t\right)}{\det(\lambda I + V_{a,\tau})} . 2\beta_t(\delta) \\
\overset{(c)}{\leq}& \|x_{a,t}\|_{\left(\lambda I+V^{\mathcal{A}}_t\right)^{-1}} . 2(1+c_1)\beta_t(\delta) \\
\overset{(d)}{=}& \sqrt{\frac{1}{M}} \|x_{a,t}\|_{\left(\frac{\lambda}{M} I+V^s_{a,t}\right)^{-1}} . 2(1+c_1)\beta_t(\delta) = c_3 \frac{\beta_t(\delta)}{\sqrt{M}} . \|x_{a,t}\|_{\left(\frac{\lambda}{M} I+V^s_{a,t}\right)^{-1}} \\
\overset{(e)}{\leq}& c_3 \frac{\beta_t(\delta)}{\sqrt{M}} . \min\left\{1, \|x_{a,t}\|_{\left(\frac{\lambda}{M} I+V^s_{a,t}\right)^{-1}}\right\}. \qquad (41)
\end{aligned}$$

Here, (a) is due to Lemma 8 with matrices $\lambda I + V^{\mathcal{A}}_t \succeq \lambda I + V_{a,\tau}$. Then, (b) is due to the radius term used in construction of $C_{a,t} \ni \theta', \theta^*$. Then, (c) is from two things: first, a single agent a's design matrix is dominated by the design of all agents (including that agent a) put together, i.e., $V_{a,\tau} \succeq V^{\mathcal{A}}_\tau$; second, from the condition for principal to initiate a communication event (line 4). The (d) is from $V^{\mathcal{A}}_t = M.V^s_{a,t}$ as in Corollary 3. The (e) is by trivially upper bounding $r_{a,t} \leq 2$.

Next, we aggregate the total regret of an agent a as follows:

$$
\begin{aligned}
R_a(T) = \sum_{t=1}^{T} r_{a,t} &\overset{(a)}{\leq} \sqrt{T \sum_{t=1}^{T} r_{a,t}^2} \\
&\overset{(b)}{\leq} c_3.\beta_T(\delta)\sqrt{\frac{T}{M}} \cdot \sqrt{\sum_{t=1}^{T} \min\left\{1, \|x_{a,t}\|^2_{\left(\frac{\lambda}{M} I + V^s_{a,t}\right)^{-1}}\right\}} \\
&\overset{(c)}{\leq} c_3.\beta_T(\delta)\sqrt{\frac{T}{M}} . \sqrt{2d \log\left(\frac{d\lambda/M + T}{d\lambda/M}\right)} \\
&\overset{(d)}{\leq} c_3. \left(\sqrt{2\log\left(\frac{\det(V_{a,t})^{1/2}\det(\lambda I)^{1/2}}{\delta}\right)} + \lambda^{1/2}\right) \sqrt{\frac{T}{M}} . \sqrt{2d \log\left(\frac{d\lambda/M + T}{d\lambda/M}\right)} \\
&\overset{(e)}{\leq} c_4 . \sqrt{d\log(MT)} . \sqrt{\frac{T}{M}} . \sqrt{d\log(MT)} = O\left(d\sqrt{\frac{T\log(MT)}{M}}\right).
\end{aligned} \tag{42}
$$

Here, (a) uses Jensen's inequality (on concavity of the square root function) or Cauchy-Schwarz inequality, then (b) follows from Equation (41), then (c) is due to Lemma 7 with $\gamma = \lambda/M$. Then (d) uses Equation (36). Then (e) uses $\det(V_{a,t}) \leq \left(\frac{MT}{d}\right)^d$, $d < t$, $\delta = (T^2M^2)^{-1}$.

Equation (42) establishes the individual agent regret guarantee mentioned in the Theorem. The expected monetary penalty of an individual agent is $O(1)$ as already shown in Lemmas 10 and 17 when all agents are honest. This completes the proof of the Theorem. □

Theorem 2. *[Regret at Equilibrium] In the collaborative linear bandits problem, let the principal follow* MONPEN *and all agents follow* LSHON. *Then, the expected regret of every agent $a \in \mathcal{A}$ is*

$$
R_a(T) = O\left(d\sqrt{\frac{T\log(TM)}{M}}\right).
$$

And, the expected penalty $P_a(T)$ of every agent is $O(1)$.

F Proof of Theorem 3

As per Theorem 1, we have that the strategy of honesty $(\varepsilon^\theta_a = 0, \varepsilon^V_a = 0)_{a\in\mathcal{A}}$ is a Nash equilibrium strategy profile of game Γ. Note that we explicitly designed the mechanism (the principal's statistical tests and penalty assignments) such that the induced game admits this strategy profile as a Nash equilibrium. However, it is quite possible that this game induces other Nash equilibriums. In Theorem 3, we enumerate all possible Nash equilibriums by computing the best response strategy of an agent to an arbitrary strategy profile of other agents.

Theorem 3. *[All Nash Equilibria] Let the principal follow* MONPEN *in common knowledge. If there is one communication event in the principal-agent bandit game Γ, then the set of all Nash equilibria is*

$$
S \subseteq \left\{ \left(\varepsilon^\theta_a, \varepsilon^V_a\right)_{a\in\mathcal{A}} \;\middle|\; \varepsilon^V_a \not\prec 0 \text{ and } \exists\, \beta \in \mathbb{R}^d \text{ s.t. } \forall a \in \mathcal{A},\ \varepsilon^\theta_a = \beta \right\}
$$

In words, a strategy profile is a Nash equilibrium only if all agents apply the same identical bias $\beta \in \mathbb{R}^d$ to their reported estimates, and do not free-ride in it $(\varepsilon^V_a \not\prec 0)$.

Proof of Theorem. To reduce clutter, we drop the subscript τ. We recap the following. The principal's protocol MONPEN is designed in way that, as shown in Claims 12 and 15, the conditional

expectation of an agent's penalty

$$\mathbb{E}\left[P_a | z_a \sim \mathcal{N}\left(\mu, \Sigma\right)\right] = \begin{cases} 0 & \text{if } \mu = 0 \text{ and } \Sigma = I_d, \\ \text{diverges to } +\infty & \text{if } \mu \neq 0 \text{ or } \Sigma_{ii} > 1 \text{ for some } i \in [d]. \end{cases} \tag{43}$$

And, as originally stated in Equation (9),

$$z_a = \left(\left(V^s_{a,\tau} + V^c_{a,\tau'}\right)^{-1} + \left(V^c_{a,\tau',\tau}\right)^{-1} \right)^{-1/2} \left(\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau',\tau} \right). \tag{9}$$

is a function of not only agent a's strategy, but also depends on the other agents' strategies. Consider an arbitrary strategy profile $\left((\varepsilon^\theta_a, \varepsilon^V_a) : a \in \mathcal{A}\right)$ of the agents. To this, every agent a's best response is to choose his strategy $(\varepsilon^\theta_a, \varepsilon^V_a)$ such that in conjunction with the other agents' strategies $(\varepsilon^\theta_b, \varepsilon^V_b)_{b \neq a}$, the resultant z_a that the principal computes follows $z_a \sim \mathcal{N}\left(0, I_d\right)$.

Next, we explicitly calculate this best response in terms of the bias ε^θ_a and deviation ε^V_a.

$$\begin{aligned} \mathbb{E}\left[z_a\right] = 0 &\iff \mathbb{E}\left[\theta^{\varepsilon}_{a,\tau} - \theta^{c,\varepsilon}_{a,\tau',\tau}\right] = 0 \\ &\iff \left(\theta^* + \varepsilon^\theta_a\right) - \left(\frac{1}{M-1} \sum_{b \neq a} \theta^* + \varepsilon^\theta_b \right) = 0 \\ &\iff \varepsilon^\theta_a = \frac{1}{M-1} \sum_{b \neq a} \varepsilon^\theta_b. \end{aligned} \tag{44}$$

From Equation (44), we see that the best response of an agent a is to set his bias ε^θ_a to the average/arithmetic mean of the biases of all the other agents. By iterative convergence, we have that for any strategy profile to be a best response to itself, it is necessary that the bias $\varepsilon^\theta_a = \beta$ for all agents $a \in \mathcal{A}$ for some value $\beta \in \mathbb{R}^d$.

Next, for the best response in terms of deviation of design matrix, $\varepsilon^V_a \in \mathbb{R}^{d \times d}$, we consider the following mapping of strategies to a scalar $\mathbb{R}^+$, and perform the calculations in that space. Recall $V^{s,\varepsilon}_a = V^s_a + \varepsilon^V_a$, and define the 'slack/free-riding coefficient' for agent a as

$$f_a := \inf \left\{ f \in \mathbb{R}^+ : V^{s,\varepsilon}_a \succeq \frac{V^s_a}{f} \right\}.$$

One can observe that in this modified strategy space, $f_a > 1$ (sim. $f_a = 1$ and $f_a \in (0,1)$) corresponds to free-riding (sim. adhering to protocol and exploring more than protocol dictates).

We compute the best response in terms of deviation of design matrix so as to comply with $\mathrm{Var}\left[z_a\right] \preceq I_d$ as follows:

$$\begin{aligned} \mathrm{Var}\left[z_a\right] \preceq I_d &\implies \mathrm{Var}\left[\theta^{s,\varepsilon}_a - \theta^{c,\varepsilon}_a\right] \preceq \left(V^s_a\right)^{-1} + \left(V^c_a\right)^{-1} \\ &\implies \mathrm{Var}\left[\theta^{s,\varepsilon}_a\right] + \mathrm{Var}\left[\theta^{c,\varepsilon}_a\right] \preceq \frac{M}{M-1} \left(V^s_a\right)^{-1} \\ \implies \left(V^{s,\varepsilon}_a\right)^{-1} + \frac{1}{(M-1)^2} \sum_{b \neq a} \left(V^{s,\varepsilon}_b\right)^{-1} &\preceq \frac{M}{M-1} \left(V^s_a\right)^{-1} \\ &\implies \left(V^{s,\varepsilon}_a\right)^{-1} \preceq \frac{M}{M-1} \left(V^s_a\right)^{-1} - \frac{1}{(M-1)^2} \sum_{b \neq a} \left(V^{s,\varepsilon}_b\right)^{-1}. \end{aligned} \tag{45}$$

The second line uses $V^c_a = (M-1)V^s_a$ from Lemma 22, third line makes use of the variance of the estimates as in Claim 6. Expressing these variance quantities in terms of the controlling strategies f_as, we have $\left(V^{s,\varepsilon}_a\right)^{-1} \preceq f_a \left(V^s_a\right)^{-1}$ for all a. Then, for the bounds in Equation (45) to hold, the

agent a has to choose f_a such that

$$\begin{aligned}
f_a(V_a^s)^{-1} &\preceq \tfrac{M}{M-1}V^{-1} - \tfrac{1}{(M-1)^2}\sum_{b\neq a} f_b(V_b^s)^{-1} \\
\implies f_a &\leq \tfrac{M}{M-1} - \tfrac{1}{(M-1)^2}\sum_{b\neq a} f_b \\
&= \tfrac{1}{M-1}\left(M - \tfrac{1}{M-1}\sum_{b\neq a} f_b\right) \\
&= \tfrac{1}{M-1}\left(M - 1 + 1 - \tfrac{1}{M-1}\sum_{b\neq a} f_b\right) \\
&= 1 + \tfrac{1}{M-1} - \tfrac{1}{(M-1)^2}\sum_{b\neq a} f_b \\
\implies f_a &\leq 1 + \tfrac{1}{M-1}\left(1 - \tfrac{1}{M-1}\sum_{b\neq a} f_b\right). \qquad (46)
\end{aligned}$$

Here, Equation (46) has an interesting interpretation. It says that the deviation of f_a from 1 has to be lesser than $1/(M-1)$ fraction of the average deviation from 1 of all other agents combined, albeit in the opposite direction. This implies that the best response value of f_a is not farther from 1 than any other f_b is from 1. By iterative convergence, this implies that any strategy profile that is a best response to itself must necessarily have $f_a = 1$ for all a. In the original strategy space, this translates to

$$V_a^{s,\varepsilon} \succeq {}^{V_a^s}\!/_1 \implies V_a^s + \varepsilon_a^V \succeq V_a^s \implies \varepsilon_a^V \succeq 0.$$

$\varepsilon_a^V \succeq 0$ is a sufficient condition to say agent a does not free-ride ($\varepsilon_a^V \not\succ 0$). □

G Additional Experiments & Details

This section provides additional experiment details and explanations that were not a part of the main paper.

G.1 Bandit Environment Setup

MovieLens environment. We simulate a realistic movie–recommendation scenario where multiple platforms share the burden of exploration while remaining robust to strategic behavior. The action sets $\{\mathcal{X}_t\}_t$ are generated by jointly embedding users and movies in a $d = 100$-dimensional latent space. Each $\mathcal{X}_t$ represents the set of movies available for recommendation to a given user, with $|\mathcal{X}_t| = 1682$. The ground-truth parameter $\theta^* \in \mathbb{R}^{100}$ is obtained as the minimizer of the least-squares error between the predicted and true ratings.

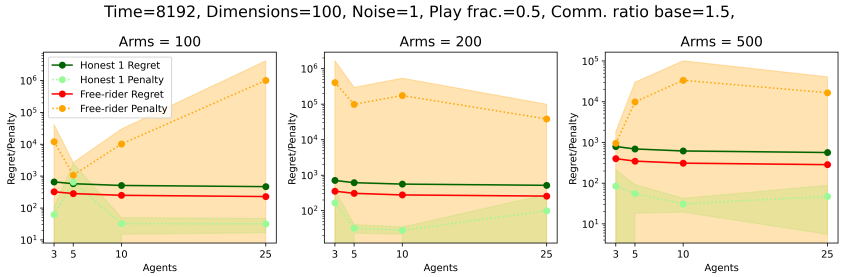
In Fig. 1, the reward for selecting an arm $x \in \mathcal{X}_t$ is $\langle x, \theta^* \rangle + \eta$, where $\eta \sim \mathcal{N}(0, 0.1)$ is drawn independently each round. We consider $M = 10$ agents and a time horizon of $T = 4096$, averaged over 3 independent runs. The regularization and penalty coefficients are set to $\lambda = 1$, $c_1 = 1.4^d$, and $c_2 = 0.1$. Our implementation extends the linear bandits module from the PEARL library (Zhu et al., 2023) to the multi-agent setting.

Synthetic environment. For the experiments in Figs. 2 and 3, we generate the true parameter θ^* and action sets $\{\mathcal{X}_t\}_t$ randomly: each coordinate of θ^* and of every $x \in \mathcal{X}_t$ is sampled uniformly at random from $[-1, 1]$ independent of all previous samples. The random seed used is 0. We use an embedding dimension $d = 50$, action sets of size $|\mathcal{X}_t| = 100$ for all t, and a total of $M = 25$ agents.

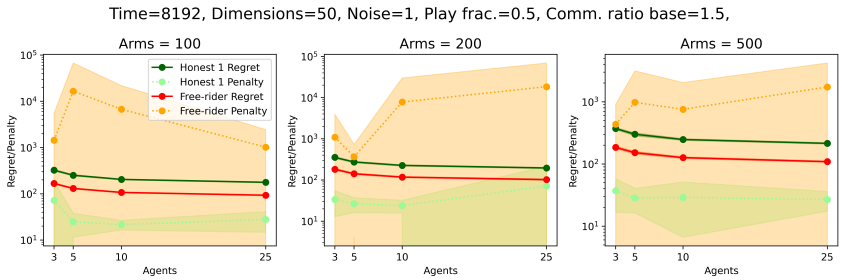
Noise variance and free-riding parameters are varied systematically. All results are averaged over 10 repetitions.

In other synthetic experiments, the data is similarly randomly sampled for the appropriate problem parameters such as agent count, action set size, and dimensionality.

G.2 Additional Results and Analysis



(a) Ambient dimension $d = 100$.

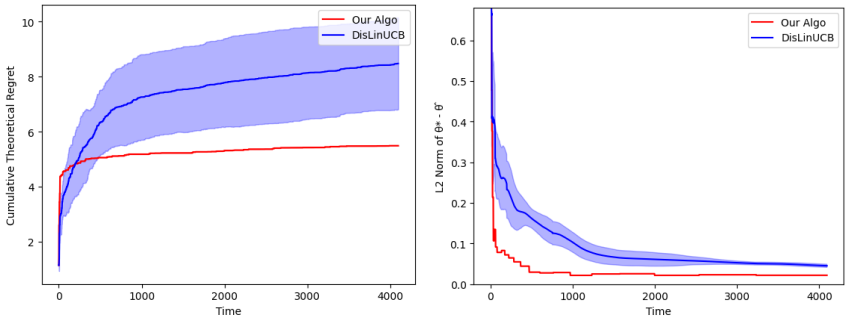


(b) Ambient dimension $d = 50$.

Figure 3: **Agent ablation:** A plot of how regret and penalties vary as a function of the total number of agents, fixing different action set size, dimension parameter values. A transposed visualization is given in Fig. 6 where the number of agents and dimension are fixed, and the variation with action set size is plotted.

Ablation on Agent Count, Action Set Size, and Dimensionality Figure 3 shows how the proposed mechanism MONPEN scales with the number of participating agents ($M \in \{3, 5, 10, 25\}$). In all cases, free-riders incur substantially higher penalties than honest agents across different dimensions and action-set sizes. Both honest and free-rider regrets decrease as M increases, since larger populations generate richer shared statistics. The same data are presented in a transposed format in Fig. 6 for comparison.

Comparison with Wang et al. (2019). For completeness, we compare the learning performance of our LSHON with DISLINUCB (Wang et al., 2019) in Fig. 4. The averages (with error bars) are computed across agents rather than repetitions. Under LSHON, all agents experience identical regret trajectories because arm plays are fully synchronized through the principal. Parameter estimates $\theta_{a,\tau}$ also coincide across agents, as updates occur only during synchronized communication rounds under the low-switching rule. In contrast, DISLINUCB exhibits variability across agents in both regret and parameter estimates. While LSHON achieves lower regret and faster convergence, these results are not directly comparable: DISLINUCB explicitly trades off performance for reduced communication, disseminating information more slowly by design.



(a) Ambient dimension $d = 100$. (b) Ambient dimension $d = 50$.

Figure 4: The cumulative regret and estimation error $\|\theta^* - \theta_{a,t}\|_2$ with all honest agents when run with DISLINUCB Wang et al. (2019) versus our LSHON. Note that estimate $\theta_{a,t}$ in LSHON is the estimate $\theta_{a,\tau}$ where τ is the time-step of the most recent communication event before t.

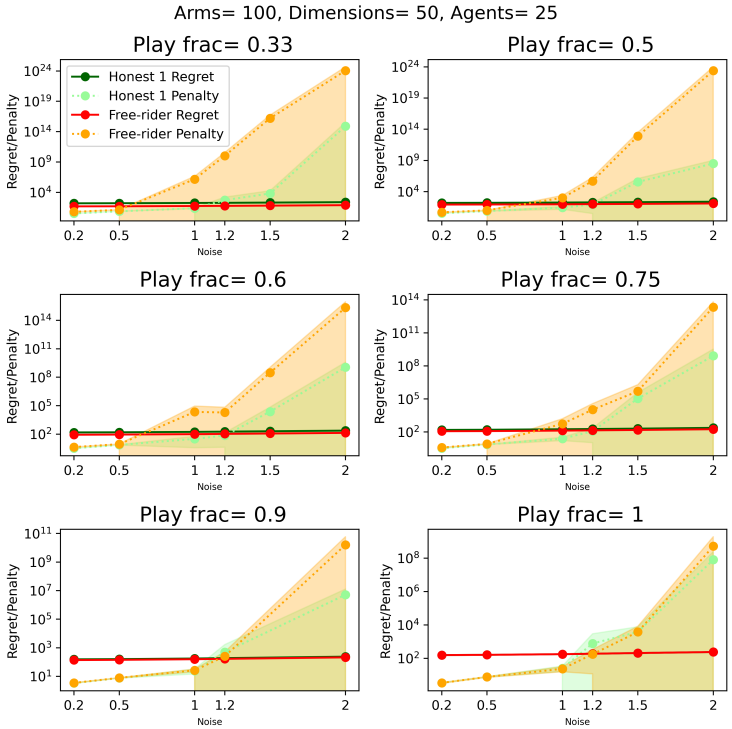
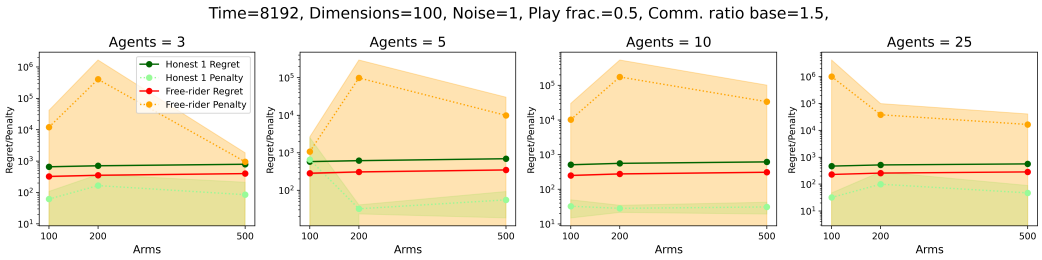


Figure 5: Transpose of Fig. 2 analysed in the main paper. We draw similar insights as given for the original plot.

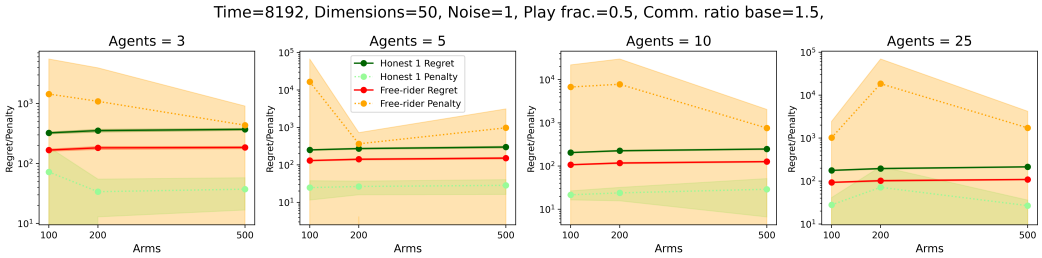
H Limitations

Limitations of Results. While our Theorem 1 guarantees that all agents adhering to LSHON is a Nash Equilibrium, it is *not* a dominant strategy. That is, it is not the best response for an agent to truthfully follow LSHON when other agents do not. This limitation stems from the principal's inability to recognize and appreciate an agent's honesty when he himself doesn't know what the true θ^* is, and having to settle for using approximations based on other agents' parameter estimates.

Our MONPEN is *not* ex post individually rational (only ex ante IR in equilibrium), as we only bound the a priori *expected* penalty, and not the final realized penalty. This is because, although the penalty



(a) Ambient dimension $d = 100$.



(b) Ambient dimension $d = 50$.

Figure 6: Transpose of Fig. 3. We draw similar insights as given for the original plot.

function is deterministic given the reported estimates of agents, the reported estimates themselves inherently contain the randomness from the noisy bandit feedback, and this affects the penalties.

Limitations of Setting. Firstly, our work involves monetary penalty in the design of mechanism. While it is possible to execute in a real-world environment, it would be interesting to see if one can design a payment-free mechanism. Secondly, we assume that the agents play the same bandit instance, that is, all agents have the same set of actions at each time-step. It would be interesting to allow for arbitrarily different actions sets for each agents. With arbitrarily different action sets, the very $\sqrt{M}$-fold benefit of collaboration, however, is not understood in the literature even when agents are not strategic. Lastly, our work does not consider competition between different agents which is a possibility in a real-world scenario.