LEARNING AGAINST A STRATEGIC AGENT IN PRINCIPAL-AGENT GAMES

Raj Kiriti Velicheti,
Department of Electrical and Computer Engineering
University of Illinois Urbana Champaign
rkv4@illinois.edu

Subhonmesh Bose
Department of Electrical and Computer Engineering
University of Illinois Urbana Champaign
boses@illinois.edu

Tamer Başar
Department of Electrical and Computer Engineering
University of Illinois Urbana Champaign
basar1@illinois.edu

ABSTRACT

Principal-Agent interactions, studied within the framework of incentive design problems, deal with the Principal (P) designing strategies such that the Agent's (A's) actions would favor P's cost. It is well known that when A has more information, then P faces a loss in optimality, known as *information rent*. While a plethora of solutions seek to devise mechanisms to tackle information asymmetry in single-stage games, we consider here the scenario of a principal who learns. Via a prototype incentive design game with continuous types and action sets, we show that P can indeed overcome the information rent through repeated interaction via an explore-then-commit (ETC) incentive policy design, when A responds myopically. We illustrate that the story is more nuanced when the agent responds in a non-myopic fashion.

1 INTRODUCTION

The canonical principal-agent problem is one of inducement–principal P seeks to make an agent A act in such a way as to benefit P in optimizing her objective. More precisely, let P and A seek to minimize their own costs, $c_P(u, v, \theta)$ and $c_A(u, v, \theta)$ where $u \in \mathcal{U}, v \in \mathcal{V}$ are actions of P and A, respectively, and θ is a cost-relevant parameter. The classical question of incentive design considers the setup where P announces how she would react to an action v by A, i.e., she chooses $u = \gamma(v; \theta) \in \mathcal{U}$. She chooses γ in such a way that A's optimal response to γ minimizes P's cost c_P. Such an interaction lends itself to a Stackelberg game formulation and has been extensively studied in Zheng et al. (1984). One can associate the design of γ, the incentive policy, with crafting a *rule-book* that P responds to A's actions with. Naturally, the design of federal/state tax-codes and Pigouvian penalties for firms incurring environmental externalities are prime examples of incentive design problems which define a sub-class of mechanism design. See Bolton & Dewatripont (2004); Başar & Olsder (1998) for a detailed survey on the topic.

Starting with the seminal work of Akerlof (1978), the effect of information asymmetry in principal-agent interactions has taken center stage in economics. There have been two broad categories of problems considered in the literature: adverse selection (where A has private information) and moral hazard (where A's actions cannot be perfectly observed by P). Such settings have been widely studied across problems with specific cost structures, such as pricing in Mussa & Rosen (1978), designing incentives for a firm in Milgrom & Weber (1982), and auctions in Holmstrom & Milgrom (1994). Another form of inducement is information design Kamenica & Gentzkow (2011); Bergemann & Morris (2019); Sayin & Başar (2020); Velicheti et al. (2025a). There has been a recent surge of interest in combining this within the framework of incentive design, e.g., in Roesler & Szentes (2017); Bergemann et al. (2015; 2022); Velicheti et al. (2025b); Dahleh et al. (2024).

The classical literature on incentive design allows P to know all that is known to A, and possibly more. In other words, P seeks to design an incentive policy, fully knowing A's preferences. In our

notation, such a scenario arises when P knows θ if A knows θ. In a variety of practical situations, however, A can *know more* than P. For example, when designing a Pigouvian tax for environmental damages a factory might cause, the factory's operations can rely on proprietary information that cannot directly be monitored by an assessor; they can rather only obtain such information through the effects of such operations. When P does not have access to information private to A, it becomes untenable to extract a perfect inducement. The loss in performance can be viewed as an *information rent*. In games of incomplete information, there is a variety of static solutions one can pursue, e.g., offering a menu of contracts in Mussa & Rosen (1978) or allowing A to provide meaningful signals as Velicheti et al. (2025b; 2024). We consider a natural next step in this paper–can P learn away the impact of not knowing θ by repeatedly interacting with A? We study a prototype incentive design game with scalar types and quadratic cost functions, and consider a specific type of incentive policy sequence. Namely, we study explore-then-commit (ETC) type incentive policy design, where we fix the structure of the incentive policy γ that P announces in each round. The policies explore up to a horizon of length H, and uses the best estimate of θ after H rounds to compute the incentive policy over $T-H$ rounds. With suitable choices for $H \asymp \sqrt{T}$, we show that the regret of the ETC algorithm over T rounds grows sub-linearly, specifically at a $\sqrt{T}$-rate. The regret is measured against the baseline cost for P attained when P knows θ. Said otherwise, this analysis underscores that in the absence of knowledge about A's private information, P can effectively nullify that advantage by interacting repeatedly with A.

Our analysis requires that A responds to an incentive policy γ from P *myopically*, i.e., A knowing θ chooses v to minimize his expected one-stage cost against P's chosen γ. If A knows that P will design her incentive policy sequence using ETC, can A do better? We explore the intricacies of learning in incentive design by considering such problems towards the end. In this work, we examine a fairly simple but illuminating scalar example with quadratic costs that illustrates the tension between learning in the presence of a myopic agent and a non-myopic agent.

2 A Prototype Incentive Game and Its Static Solution

Consider an incentive design problem between P and A, whose costs are given by

$$\begin{aligned} c_P(u, v; \theta) &= (\theta - u - v)^2 + 2u^2 + \beta v^2, \\ c_A(u, v; \theta) &= (\theta - u - v)^2 + v^2, \end{aligned} \tag{1}$$

parameterized by $\beta > 0$. The action sets are $\mathcal{U} = \mathcal{V} = \Theta = \mathbb{R}$. Consider first the case where θ is known. For this Stackelberg game, one can solve an equilibrium by first considering a strategy of P of a certain form, calculating an optimal response from A, and then optimizing over the strategy space of P, considering said optimal response from A. Such a route yields a functional optimization problem that can be challenging to solve, but the work in Başar (1984) comes to the rescue. It shows that P can solve the decentralized control problem, assuming the role of A in minimizing her (i.e., P's) own cost. Let the solutions be u_p, v_p. Then, an incentive policy of the form $\gamma(v; \theta) = u_p + Q^\star(v - v_p)$ can be computed where $Q^\star$ is such that A's optimal response to $\gamma(v; \theta)$ coincides with v_p. For this problem, the solution can be found in Başar (1984), given by

$$u_p = \frac{\beta\theta}{3\beta + 2},\ v_p = \frac{2\theta}{3\beta + 2},\ Q^\star = \frac{1 - \beta}{\beta}. \tag{2}$$

This *indirect mechanism* then automatically guarantees that P can achieve her optimal cost through an affine incentive policy of the form given above, since P shapes A's utility through γ in a way that A responds exactly as P would have liked to act on A's behalf.

When A knows θ, but P does not know θ, such a workflow to compute an incentive strategy breaks down. P cannot calculate what A must implement, not knowing what A knows. To tackle this problem with non-classical information structure, one can adopt the Bayesian route from the recent work Velicheti et al. (2025b). Let both players start with a common prior, $\theta \sim \mathcal{N}(z_0, \sigma_0^2)$, and P implement an *affine mean feedback* policy of the form

$$\gamma(v; z_0) = \frac{\beta z_0}{3\beta + 2} + \frac{1 - \beta}{\beta}\left(v - \frac{2z_0}{3\beta + 2}\right), \tag{3}$$

which takes the incentive policy for the known θ case and replaces it by the prior mean z_0. Given this incentive policy, A's best response, knowing θ is then given by

$$v^\star(\theta; z_0) = \frac{\beta}{\beta^2+1}\theta - \frac{\beta^2+2\beta-2}{(\beta^2+1)(3\beta+2)}z_0. \tag{4}$$

It is easy to verify that both players' equilibrium strategies collapse to the known θ case, upon replacing z_0 with θ.

One can compute the expected cost to P when θ is known versus when θ is not known. Naturally, P must pay for an *information rent*–a penalty for not knowing the unknown parameter, known to A. Indeed, as shown in Velicheti et al. (2023), this rent is non-negative and increases in σ_0^2, the uncertainty in not knowing θ. In what follows, we address the question: *can P do better by repeatedly engaging with A*?

3 The Repeated Game of Incentive Design

Leaving the domain of a static incentive design game, we now present a repeated version of the same quadratic-Gaussian game example as before, but with the added change that P does not observe A's action, $v(t)$, at each time step t; she rather observes $v(t)$ through a noisy channel $y(t) = v(t)+w(t)$, where $w \sim \mathcal{N}(0, \sigma_w^2)$'s are i.i.d. across time. Thus, at each time t, P announces an incentive policy of the form, $\Gamma(y(t); t)$ of how she would react to the observation $y(t)$ generated from A playing $v(t)$.

To contrast how an algorithm might perform in this repeated context, consider the baseline obtained by repeatedly playing an affine mean feedback policy $\Gamma(y(t); t) = \gamma(y(t); \theta)$ with the notation γ defined in (equation 3). Notice that this affine incentive policy corresponds to the case that P and A both know θ, but P only observes A's action through a noisy channel to implement her own action. Given P's policy, A optimizes his response. A calculation similar to that for (equation 4) can be shown to yield the optimal response to be $v^\star(\theta; \theta) = 2\theta/(3\beta+2)$. With this response, the expected cost for P can be shown to be

$$J_P^{\text{baseline}} = \frac{3(\beta-1)^2}{\beta^2}\sigma_w^2 + \frac{2\beta}{3\beta+2}(z_0^2+\sigma_0^2). \tag{5}$$

In the above expression, the first term arises from the noise in the observation of A's action, and the second term coincides with the optimal cost with known θ and then taking an expectation over the Gaussian prior. P's expected cost from any learning algorithm can only hope to match this cost on average. In the next section, we will present and analyze the performance of a specific sequence of incentive policies through time and show that its average cost matches the above as $T \to \infty$.

4 Explore-Then-Commit Affine Mean Feedback Incentive Policy

At each time t, P announces an incentive policy $\Gamma(y(t); t)$, observes $y(t)$ generated from A playing $v(t)$, implements her action $u(t)$ according to that policy, and observes her realized cost, $c_P(u(t), v(t); \theta)$. Both $y(t)$ and $c_P(u(t), v(t); \theta)$ depend on θ, and hence reveal information about θ. In this section, we disregard the extra information that is possible to garner from the cost sequence and only utilize the observed noisy actions to design an incentive policy sequence.

For some $1 \le H < T$, let P use the explore-then-commit (ETC) incentive policy sequence

$$\Gamma(y(t); t) = \begin{cases} \gamma(y(t); z_0), & \text{if } 1 \le t \le H, \\ \gamma(y(t); z_H), & \text{otherwise,} \end{cases} \tag{6}$$

where z_H is calculated as follows. Upon using the same policy $\gamma(y(t); z_0)$ up to time H, A responds by playing $v(t) = v^\star(\theta; z_0)$ as defined in (equation 4), which generates observations, $y(t) = v^\star(\theta; z_0) + w(t)$ over $1 \le t \le H$. Notice that $y(t)$'s are linear in θ with a Gaussian noise. In particular, write it as $y(t) = A\theta + B + w(t)$. Then, the posterior distribution of θ, given $y(1), \ldots, y(H)$ is $\mathcal{N}(z_H, \sigma_H^2)$, where

$$\sigma_H^2 := \left(\frac{1}{\sigma_0^2} + \frac{HA^2}{\sigma_w^2}\right)^{-1}, \; z_H := \sigma_H^2 \left[\frac{z_0}{\sigma_0^2} + \frac{A}{\sigma_w^2}\sum_{t=1}^{H}(y_t - B)\right]. \tag{7}$$

Thus, with the ETC policy sequence, P collects observations with the prior-based affine mean feedback policy till time H (the exploration window), following which P uses the same incentive policy structure, but with a refined estimate of θ. We assume throughout this section that A keeps responding myopically to the affine mean feedback policy announced by P. In the next result, we characterize the performance of ETC.

When P implements $\gamma(y; z)$ and A responds by playing $v^\star(\theta; z)$, the expected cost to P is given by

$$\mathbb{E}_w\left[c_P(u, v; \theta)\right] = \frac{3(\beta-1)^2}{\beta^2}\sigma_w^2 + \frac{2\beta}{3\beta+2}\theta^2 + \frac{(\beta+1)\left(\beta^2+2\beta-2\right)^2}{(3\beta+2)\left(\beta^2+1\right)^2}(z-\theta)^2. \tag{8}$$

Call the expression on the right-hand side of the above equation to be $f(z)$. Then, the expected cost of the ETC algorithm is given by

$$\begin{aligned} J_P^{\text{ETC}}(T; H) &= \mathbb{E}_\theta[Hf(z_0) + (T-H)\mathbb{E}[f(z_H) \mid \theta]] \\ &= \frac{3(\beta-1)^2}{\beta^2}\sigma_w^2 + \frac{2\beta}{3\beta+2}(z_0^2+\sigma_0^2) \\ &\quad + \underbrace{\frac{(\beta+1)\left(\beta^2+2\beta-2\right)^2}{(3\beta+2)\left(\beta^2+1\right)^2}}_{:=\varphi(\beta)}\left[H\sigma_0^2 + (T-H)\sigma_H^2\right]. \end{aligned} \tag{9}$$

The regret of using the ETC policy is given by

$$\begin{aligned} \text{Regret}^{\text{ETC}}(T) &= \min_{1\le H\le T} J_P^{\text{ETC}}(T; H) - TJ_P^{\text{baseline}} \\ &= \varphi(\beta)\cdot \min_{1\le H\le T}\left[H\sigma_0^2 + (T-H)\frac{\sigma_0^2\sigma_w^2}{\sigma_w^2 + HA^2\sigma_0^2}\right] \\ &= \varphi(\beta)\frac{2\sigma_w}{A^2}\left(\sqrt{\sigma_w^2 + A^2\sigma_0^2\,T} - \sigma_w\right), \end{aligned} \tag{10}$$

obtained with $H^\star(T) = \frac{\sigma_w}{A^2\sigma_0^2}\left(\sqrt{\sigma_w^2 + A^2\sigma_0^2\,T} - \sigma_w\right)$, ignoring integrality constraints. In effect, we have proven the following result.

Theorem 1 *Regret*$^{ETC}(T) \lesssim \sqrt{T}$.

P can therefore reduce the regret from not knowing θ through repeated interaction to be sub-linear in T. The average deviation from not knowing θ vanishes on average as $T \to \infty$. We believe that for this problem, regret for *any* algorithm $\gtrsim \sqrt{T}$, making ETC order-optimal. We also believe that the same result will continue to hold if ETC is replaced by a sequence of incentive policies, where at each time t, P updates the belief at t with the observed $y(t-1)$ to $\mathcal{N}(z_t, \sigma_t^2)$ and uses it to change $\Gamma(y(t); t) = \gamma(y(t), z_t)$.

5 The Case with a Non-Myopic Agent

The analysis of the ETC algorithm is premised on the assumption that A responds optimally in a stage-wise fashion to P's affine mean feedback policy. However, a forward-looking A might anticipate how his response will alter P's incentive policy sequence, and whether A can benefit by deviating from that myopic response to have lesser cumulative costs over T rounds. Consider the case where A knows that P will use an ETC design with a horizon length $H^\star$. For this analysis, let $z_0 = \sigma_0 = \sigma_w = 1$. When P announces $\gamma(y; z)$ and A plays $v^\star(\theta; z)$, then A incurs a cost of $g(z; \theta) := \mathbb{E}_w[c_A(u, v; \theta)]$, a quadratic function of z. For large T, we have $H^\star = \sqrt{T}/A$. If A plays v_0 up until time $H^\star$, P's estimated mean for θ at $t = H^\star$ becomes

$$z_{H^\star}(v_0) \approx \frac{1}{1+H^\star A^2}\left(1 + AH^\star(v_0 - B)\right) \approx \frac{\beta^2+1}{\beta}\left(v_0 + \frac{\beta^2+2\beta-2}{(\beta^2+1)(3\beta+2)}\right). \tag{11}$$

If A responds myopically within the exploration phase, A accrues a per-period expected cost of $g(z_{H^\star}(v^\star(\theta; 1)); \theta)$ beyond that phase. However, if A chooses $v_0 = 2/(3\beta+2)$, then $z_{H^\star}(v_0) \approx$

$z_0 = 1$, making P's estimated mean after $H^\star$ equal the prior mean. In this case, after $H^\star$, A accrues a per-period cost of $g(z_{H^\star}(2/(3\beta+2)); \theta)$. Then, A stands to gain

$$\begin{aligned} & g(z_{H^\star}(v^\star(\theta;1)); \theta) - g(z_{H^\star}(2/(3\beta+2)); \theta) \\ &= -\frac{\beta^2+2\beta-2}{(3\beta+2)^2}\left[(\beta^2+1)\left(\theta - \frac{\beta(\beta+1)}{\beta^2+1}\right)^2 + \frac{(\beta^2-1)^2}{\beta^2+1}\right] \end{aligned} \tag{12}$$

from fooling P during the exploration phase. This difference can be positive for $\beta < \sqrt{3}-1$. Since $H^\star \asymp \sqrt{T} \ll T$, A gains by this much per-stage post-exploration (the dominant contribution over T rounds) from fooling P. Moreover, P deduces a posterior mean $\approx$ the prior mean. The ETC sequence then produces linear regret for P. Thus, A's forward-looking behavior wipes out the advantage of repeated play from P in learning θ in this example.

6 Discussion and Conclusion

In this work, we have examined a principal-agent interaction where A possesses more information about his cost function, while P learns about it through sequential interaction. We have demonstrated that even in this basic setup, there are scenarios in which A benefits from preventing P from acquiring this additional information. Through explicit construction, we have shown that intuitive algorithms like ETC can be exploited by A to his advantage. Hence it is imperative to mathematically model and understand the capabilities of information rich player in such interactions.

Looking ahead, our goal is to further investigate what the analysis of this paper has revealed, particularly by exploring alternative learning algorithms for P, such as introducing stochasticity in the decision horizon before commitment or using adaptive methods like UCB and multiplicative weight updates. There is an inherent trade-off in these approaches: as the algorithm becomes more adaptive and P seeks to learn more, A gains the ability to influence the outcome to a greater extent. For instance, if P adopts a fixed incentive policy that remains unchanged, A has no incentive to manipulate it and would simply play his true best response. However, when P attempts to learn from interactions, it opens up the possibility for A to influence the learning process in his favor, creating an interesting meta-game. We intend to explore this dynamic in greater detail in future work.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We also thank Arda Guclu(UIUC) and Dr. Sina Sanjari (RMC, Canada) for their help with an initial version of this work. This work was partially supported by the grant NSF ECCS 2349418 from the US National Science Foundation.

References

George A Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in Economics*, pp. 235–251. Elsevier, 1978.

Tamer Başar. Affine incentive schemes for stochastic systems with dynamic information. *SIAM Journal on Control and Optimization*, 22(2):199–210, 1984.

Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 1998.

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

Dirk Bergemann, Benjamin Brooks, and Stephen Morris. The limits of price discrimination. *American Economic Review*, 105(3):921–957, 2015.

Dirk Bergemann, Tibor Heumann, and Stephen Morris. Screening with persuasion. *arXiv preprint arXiv:2212.03360*, 2022.

Patrick Bolton and Mathias Dewatripont. *Contract Theory*. MIT Press, 2004.

Munther A Dahleh, Thibaut Horel, and M Umar B Niazi. Mitigating information asymmetry in two-stage contracts with non-myopic agents. *IFAC-PapersOnLine*, 58(30):19–24, 2024.

Bengt Holmstrom and Paul Milgrom. The firm as an incentive system. *The American Economic Review*, pp. 972–991, 1994.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101 (6):2590–2615, 2011.

Paul R Milgrom and Robert J Weber. A theory of auctions and competitive bidding. *Econometrica: Journal of the Econometric Society*, pp. 1089–1122, 1982.

Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic Theory*, 18(2):301–317, 1978.

Anne-Katrin Roesler and Balázs Szentes. Buyer-optimal learning and monopoly pricing. *American Economic Review*, 107(7):2072–80, July 2017. doi: 10.1257/aer.20160145. URL https://www.aeaweb.org/articles?id=10.1257/aer.20160145.

Muhammed O Sayin and Tamer Başar. Persuasion-based robust sensor design against attackers with unknown control objectives. *IEEE Transactions on Automatic Control*, 66(10):4589–4603, 2020.

Raj Kiriti Velicheti, Melih Bastopcu, and Tamer Başar. Strategic information design in quadratic multidimensional persuasion games with two senders. In *2023 American Control Conference (ACC)*, pp. 1716–1722. IEEE, 2023.

Raj Kiriti Velicheti, Melih Bastopcu, S Rasoul Etesami, and Tamer Başar. Learning how to strategically disclose information. In *2024 American Control Conference (ACC)*, pp. 1604–1609. IEEE, 2024.

Raj Kiriti Velicheti, Melih Bastopcu, and Tamer Başar. Value of information in games with multiple strategic information providers. *IEEE Transactions on Automatic Control*, 70(7):4532–4547, 2025a.

Raj Kiriti Velicheti, Subhonmesh Bose, and Tamer Başar. Harnessing information in incentive design. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 7286–7291. IEEE, 2025b.

Ying-Ping Zheng, Tamer Başar, and Jose B Cruz. Stackelberg strategies and incentives in multiperson deterministic decision problems. *IEEE Transactions on Systems, Man, and Cybernetics*, (1): 10–24, 1984.