Learning Under Moral Hazard with Instrumental Regression and Generalized Method of Moments

Shiliang Zuo
University of Illinois Urbana-Champaign

Abstract

Machine learning has become increasingly popular in informing data-driven policy-making. Policies influence behavior in individuals or populations, and ideally, through observational signals, policy-makers learn which policies are effective. However, in many settings, individual actions cannot be perfectly observed. This issue, known in economics as moral hazard, poses a significant challenge. In this work, we study the foundational multitasking principal–agent contract design problem and demonstrate how instrumental regression and the generalized method of moments (GMM) estimator can be used to estimate or learn a good contract. As a bonus result, we also give a uniformity characterization of the shape of the optimal contract.

1 INTRODUCTION

Machine learning has become a powerful tool for data-driven decision-making across many real-world applications, including ridesharing ([Qin et al., 2022]), education policy-making ([Hilbert et al., 2021]), healthcare policy-making ([Ashrafian and Darzi, 2018]), and credit scoring ([Fuster et al., 2022, Hurley and Adebayo, 2016]), among others. A central assumption in most machine learning methods is that the causal relationship between inputs and outcomes is well-specified and observable. However, in many real-world policy-making scenarios, the dependence between signals and outcomes is often complex and difficult to model. In particular, in economic and strategic environments, this assumption breaks down: data are generated by self-interested agents, whose underlying actions drive outcomes but remain hidden. As a result, the true causal links between observed signals and unobserved behavior are confounded, posing fundamental challenges for learning.

Consider an educational setting where we are interested in measuring students' skills to predict their future success (however defined). While some information is observable—such as educational background—many important features remain hidden, such as problem-solving or critical thinking skills. These latent skills can only be indirectly inferred through noisy signals like standardized test scores. In this case, the covariates of interest are not perfectly observed. Moreover, policymakers may wish to design better education policies based on students' data, but doing so requires addressing the fact that underlying behaviors and abilities are only imperfectly captured.

As another example, consider a vehicle insurance company seeking to design policies for its customers. A policyholder's driving behavior is influenced by the insurance plan: some individuals may become more cautious, while others may drive more recklessly, depending on the policy they participate in. The insurer's objective is to shape behavior in a way that maximizes profit. Yet, driving behavior cannot be directly observed; instead, the insurer only has access to noisy outcomes such as accident reports. Despite this limitation, the company must still rely on historical data to design effective insurance policies.

This type of challenge is known as moral hazard in economics ([Holmström, 1979]). Moral hazard arises when an agent's actions are hidden from the principal, making it difficult to align incentives. In this work, we initiate the study of data-driven methods for decision-making under moral hazard, where optimal policies must be learned from observational signals that depend on unobserved actions.

We focus on the multitasking principal–agent problem, a canonical model of contract design

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

capturing the salient feature of moral hazard ([Holmstrom and Milgrom, 1991]). Contract design is an important topic in economics: it captures how a principal can incentivize an agent to exert effort in working on tasks by linking payments to observable signals. In the multitasking principal-agent problem, the principal hires an agent to perform several tasks. The principal's utility depends on the agent's hidden action; however the effort the agent puts into each task can only be measured through a noisy signal.

In this work, we study how to learn effective contracts for the principal under such information constraints. Our approach combines tools from economics, econometrics, and machine learning, and in particular demonstrates how econometric methods such as instrumental regression and the generalized method of moments (GMM) can be adapted to address endogeneity and measurement error in this strategic environment.

Apart from our learning results, we also provide a fairness characterization of optimal contracts. In many real-world settings, the principal interacts with multiple agents under a common contractual framework; for instance, franchising arrangements [Bhattacharyya and Lafontaine, 1995] or revenue-sharing platforms such as YouTube, where content creators receive 55% of advertising revenue while the platform retains 45% [1]. While one might expect the principal to customize terms for different agents, we show that when the agent's cost function exhibits homogeneity, the optimal contract depends only on its degree of homogeneity. This implies that uniform contracts can simultaneously maximize the principal's utility and reinforce fairness across agents.

1.1 Contributions

In this work, we study multitasking principal-agent. We first give a *fairness characterization* of the optimal contract. We show that under a homogeneity assumption on the agent's cost function, the optimal linear contract depends only on the degree of homogeneity and is uniform across agents. This provides theoretical insight and practical justification for standardized contracts in settings like franchising and online platforms.

We then study the multitasking principal-agent problem from a *machine learning* perspective, we show how tools from econometrics, particularly instrumental regression and the generalized method of moments (GMM), can be applied in this setting in designing an effective contract. To our knowledge, this is the first work to examine the multitasking principal–agent problem through a machine learning lens. First, we identify contract learning under moral hazard as a regression problem with measurement error and show how instrumental regression, together with the generalized method of moments (GMM) estimator, can be used to recover the unknown parameters and thereby identify the optimal contract. Second, we demonstrate that when repeated signals are available and agents are sufficiently diverse, the principal can achieve significantly faster convergence rates.

[1] See [Google, 2024].

1.2 Related Work

Machine Learning in Strategic Contexts Machine learning increasingly operates in environments where data are generated by self-interested agents. Recent work introduced the study of strategic classification ([Hardt et al., 2016]) and performative prediction ([Perdomo et al., 2020]), showing how individuals may manipulate features in response to predictive models. Following this, researchers have highlighted endogeneity issues in strategic settings and developed instrumental regression approaches for regression with strategic responses ([Harris et al., 2022]), online learning and bandits ([Della Vecchia and Basu, 2023]), and reinforcement learning with hidden actions ([Yu et al., 2022]). More broadly, recent work emphasizes the causal perspective in machine learning systems ([Horowitz and Rosenfeld, 2023], [Miller et al., 2020]), motivated by real-world applications where decisions shape behavior and the data reflect these strategic interactions.

Contract Theory Contract theory studies a principal-agent problem where the principal must incentivize the agent to exert effort through contracts. Some early important works in the economics community include [Holmström, 1979, Holmstrom, 1982, Holmstrom and Milgrom, 1991]. The work by [Holmstrom and Milgrom, 1991] studies the multitask principal-agent problem, which serves as the starting point of the current work. Recently, computational aspects of contract design have been studied by many works in the computer science community. For example, [Dütting et al., 2019, Duetting et al., 2024, Dütting et al., 2022] study contract design from a combinatorial perspective; various other works [Ho et al., 2014, Zuo, 2024, Guruganesh et al., 2024, Zhu et al., 2022] study learning algorithms for contract design. Another line of work seek to understand the worst-case guarantee of contracts (e.g. [Carroll, 2015]), and show that linear contracts have remarkable worst-case performance guarantees. Apart from these, contract theory has also appeared in application domains such as signal processing ([Jain et al., 2023]).

Bandit Problems Bandit problems are a type of online learning problem with partial feedback. Typically, bandit algorithms need to balance exploration (experimenting with unknown actions) and exploitation (choosing actions whose reward is estimated to be higher). Research on bandit problems is too broad to cover here, but for an overview see [Lattimore and Szepesvári, 2020]. Recently, a number of papers have sought to understand the performance of the pure exploitation (i.e., greedy) algorithm in bandit problems. In particular, for linear bandits, the success of the greedy algorithm is explained under the framework of smoothed analysis [Bastani et al., 2021, Kannan et al., 2018, Sivakumar et al., 2022].

Instrumental Regression and Measurement Error Models. Measurement error models are a well-studied topic in econometrics and statistics, as noisy or imperfectly observed covariates can lead to biased and inconsistent estimates. A standard remedy is the use of instrumental variables, where an observed variable correlated with the true covariate but independent of the error serves as a proxy. The generalized method of moments (GMM) framework further offers a flexible way to construct consistent estimators in the presence of endogeneity. These methods are well established in econometrics; see [Greene, 2008, Fuller, 2009] for textbook treatments.

2 THE MULTITASKING PRINCIPAL-AGENT PROBLEM

In the multitasking principal-agent problem, there are two parties, a principal and an agent. The principal asks the agent to complete several tasks. The agent exerts effort across d dimensions, each representing a distinct task (so there are d tasks). We denote the agent's effort by the vector $a \subset (\mathbb{R}^+)^d$.

The agent incurs a private cost $c(a) \in \mathbb{R}^+$ when exerting the effort vector a. We will assume the cost function $c(a)$ is strictly increasing, strictly convex, and continuously differentiable. While the principal cannot directly observe or verify the exact effort level a, she can observe a signal $x \in \mathbb{R}^d$; the signal can be interpreted as a noisy measurement of the agent's true effort vector, in particular, the signal x_i can represent a noisy measurement for the true effort a_i in task i. We shall assume the signal is unbiased:

$$\mathbb{E}[x|a] = a.$$

The unbiasedness is a common assumption (e.g. [Holmstrom and Milgrom, 1991]).

The principal incentivizes the agent to exert effort through a contract. We focus on linear contracts, parameterized by $\beta \in (\mathbb{R}^+)^d$. Under the linear contract β, when the observed signal is x, the payment transferred from the principal to the agent is $\langle \beta, x \rangle$; in expectation, the transfer is equal to $\langle \beta, a \rangle$.

Agent's Response The agent's utility is his expected payment minus his private cost. The agent selects an action $a = a(\beta)$ that maximizes their expected utility given the contract terms:

$$a(\beta) = \arg\max_a \langle \beta, a \rangle - c(a).$$

Principal's Utility The agent's effort a gives a noisy private benefit $y(a)$ to the principal. We assume this private benefit takes a linear form

$$\mathbb{E}[y(a)|a] = \langle \theta^*, a \rangle.$$

The principal's expected utility is the expected private benefit minus the expected payment to the agent. In other words, when the contract offered is β and the agent best responds with the action a, the principal's expected utility is

$$u(\beta; a) = \langle \theta^*, a \rangle - \langle \beta, a \rangle.$$

We may also write $u(\beta) := u(\beta, a(\beta))$; here the variable a is suppressed and implicitly understood as the agent's best response to β. The principal's optimal contract β^* is then the contract maximizing her expected utility:

$$\beta^* \in \arg\max_\beta u(\beta).$$

Interaction The interaction protocol can be summarized as follows.

1. The principal posts a contract $\beta \in \mathbb{R}^d$
2. The agent best responds with a private effort (action) $a \in \mathbb{R}$, and incurs private cost $c(a)$
3. A noisy signal x is generated and that $\mathbb{E}[x|a] = a$
4. The agent's expected utility is $\langle \beta, a \rangle - c(a)$
5. The principal's private realized benefit is y with $\mathbb{E}[y|a] = \langle \theta^*, a \rangle$, the expected utility $u(\beta; a) = \langle \theta^*, a \rangle - \langle \beta, a \rangle$.

Note that from the principal's point of view, apart from the decision variable β, only the signal x and the realized private benefit y are observed; the agent's true effort a is not observed. The causal relation between the variables is summarized in Figure 1.

Before we present our main technical results, let us first clarify how our model captures the essential features of moral hazard. In its modern economic usage, moral hazard refers to settings where hidden actions create incentive misalignment between contracting parties. Our multitasking contract design problem exhibits the core characteristics of moral hazard: (1) the agent's action is unobservable and noncontractible, (2) actions affect outcomes but only noisy signals are observed, and (3) contracts must be designed based on observable signals to incentivize hidden actions.

These features are present in our motivating examples (Section 1). In the insurance setting, driving behavior is hidden from the insurer, who observes only noisy signals such as accident reports and claims costs; the insurer must design policies that incentivize safe driving to maximize profit. Similarly, in the education setting, student effort and true skills are latent, while policymakers observe only noisy test scores and long-term outcomes; they must design policies that incentivize appropriate effort allocation across different subjects or skill areas.

As a final remark, we note that the unbiasedness assumption on signals specifies the information structure but does not make actions observable or contractible. The moral hazard remains intact: the principal cannot directly observe or write contracts contingent on the agent's effort a. Although unbiasedness is a common assumption in the multitasking literature (e.g. [Holmstrom and Milgrom, 1991]), extending our results in this work to accommodate possibly biased signals is also a valuable future direction.

3 UNIFORMITY OF THE OPTIMAL LINEAR CONTRACT

We first give a characterization of the optimal linear contract. We present a uniformity result, which states that if the agent's cost function exhibits a certain degree of homogeneity (i.e. a consistent return to scale), then the optimal linear contract depends on the cost function only through its homogeneity degree.

Assumption 1. *The cost function of the agent is homogeneous of degree k. In other words, for any $\rho > 0$ and $a \in (\mathbb{R}^d)^+$, $c(\rho a) = \rho^k c(a)$.*

Remark 1. *The homogeneity assumption was also made in a related line of work in strategic classification [Shavit et al., 2020, Dong et al., 2018]. Another work on learning from revealed preferences in Stackelberg games also make a homogeneity assumption [Roth et al., 2016] (see the subsequent Remark 4).*

We show the optimal linear contract depends on the shape of $c(\cdot)$ only through the degree of homogeneity. The proof is very simple, yet to the best of our knowledge, it has not appeared in prior work.

Theorem 1. *The principal's optimal contract is $\beta^* = \theta^*/k$.*

Proof. Suppose the principal posts the contract β. By the first-order condition of the agent's best response, we have:

$$\beta = \nabla c(a).$$

Then, the principal's utility function when posting the contract β, as a function of the best response a can be written as

$$\begin{aligned}\langle \theta^*, a\rangle - \langle \beta, a\rangle &= \langle \theta^*, a\rangle - \langle \nabla c(a), a\rangle \\ &= \langle \theta^*, a\rangle - kc(a),\end{aligned}$$

where we used Euler's theorem on homogeneous functions. Notice that the principal's utility function is a concave function in the agent's hidden effort a. Taking derivative with respect to a, we have

$$\theta^* = k\nabla c(a).$$

Combined with the agent's first-order condition, we have that at the optimum we must have

$$\beta^* = \theta^*/k. \qquad \square$$

Remark 2. *The uniformity result is particularly attractive from a practical perspective. It implies that the principal can achieve her maximum utility and ensure fairness simultaneously. In particular, there is no need to discriminate against agents whose productivity may differ (as long as their return to scale parameter k remains the same). This standardization is particularly relevant in industries like franchising or E-commerce, where the principal (the franchiser or the E-commerce platform) contracts with* multiple *agents (franchisees, E-commerce workers / content creators).*

Remark 3. *The uniformity result here is similar in flavor to [Bhattacharyya and Lafontaine, 1995]; the difference is that their work considers a double moral hazard setting with a single task. They also make a similar homogeneous degree assumption, but of course, in the single-dimensional case, a homogenous function can only take the polynomial form.*

Remark 4. *The recent work [Roth et al., 2016] in fact studied a very similar setup with the same homogeneity assumption. Their work studied optimization algorithms in Stackelberg games from revealed preferences. As a special case, they gave a discussion on how their algorithm can be applied to the multitask principal-agent problem (though their work did not explicitly identify their problem as such). They also make a homogeneity assumption; however, the uniformity result that we presented here seemed to have somehow escaped their analysis.*

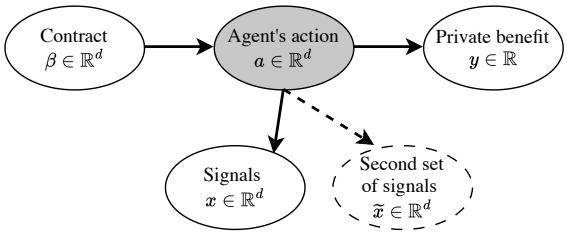


Figure 1: Causal relationship between variables. The agent's response is shaded indicating it is unobserved by the principal. The dashed item is a second set of observed signals, corresponding the the scenario studied in Section 5 when repeated observations are available.

4 ESTIMATING AND LEARNING THE OPTIMAL CONTRACT: AN INSTRUMENTAL REGRESSION APPROACH

In the previous section we gave a characterization of the optimal contract and showed that even for different agents with different cost functions, the optimal contract remains the same as long as the homogeneity degree remains the same. Now, we study the problem of estimating or learning the optimal contract when the parameter θ^* is unknown. Informally, when the parameter θ^* is unknown, the principal lacks precise knowledge of the "importance" of each task. For instance, when a firm's owner (the principal) hires a manager (the agent) to oversee multiple tasks, the owner may be uncertain about the exact contribution of each task to the firm's overall performance. In such cases, the principal must estimate or learn θ^* using observational data.

We assume the principal repeatedly interacts with potentially different agents, and denote the observed data in the t-th interaction as (β_t, x_t, y_t), representing the contract, observed signal, and private benefit, respectively; the agent's private effort a_t is unobserved by the principal. In a single round of interaction, the relationship between the variables is depicted in Figure 1. Similar to the previous section, it is assumed the observed signals for each task is unbiased:

$$\mathbb{E}[x_t|a_t] = a_t,$$

and that the total revenue satisfies a linear relation:

$$\mathbb{E}[y_t|a_t] = \langle \theta^*, a_t \rangle.$$

In repeated interactions, we allow the agent in each round to be of different types. In particular, different agents may have different cost functions and react differently to the same contract. However, we assume that the homogeneity degree parameter k is known and treated as a constant. In other words, the return to scale parameter is a constant for different agent types. [2] Informally, this means that even though agent's may have different "proficiency" in completing the tasks, the scaling behavior of the cost function is consistent.

Note that in general, in particular when the homogeneity Assumption 1 does not hold, recovering the unknown parameter θ^* will not inform the shape of the optimal contract, since the agent's cost function is unknown. However, thanks to the homogeneity assumption and the uniformity result, which establishes the optimal contract as given by $\beta^* = \theta^*/k$, obtaining an accurate estimate for θ^* is equivalent to identifying an approximately optimal contract.

We consider both the offline and online setting. In the offline setting, the principal has access to data in T time periods, with the observed data denoted as (β_t, x_t, y_t); the goal is then to obtain an accurate estimate of θ^*. In the online setting, the principal interacts with agents in a sequential manner, and in each round t the principal chooses a contract β_t based on historical observations; the goal here is to minimize cumulative utility loss.

We assume the contract β belongs to some bounded set $\mathcal{B}$, and that the optimal contract $\beta^* = \theta^*/k \in \mathcal{B}$. We refer to $\mathcal{B}$ as the feasible contract set; it can represent either external factors (such as certain law or regulations) which limit the set of contracts that can be enforced, or it can represent the fact that the principal has some limited prior knowledge that the optimal contract must belong to some bounded set (for example, he may at least have an lower and upper bound of each component of θ^*).

Finally, we assume the signals x_t and realized private benefits y_t are conditionally subgaussian.

Assumption 2. *Given the agent's action a_t, the random signal x_t and private benefit y_t are σ_0-subgaussian.*

We note that standard linear regression in fact fails in recovering the θ^*. Treating the problem as a regression problem, the true covariates a are not observed, and only a noisy measurement x is observed. Hence, our problem in fact has an errors-in-variables

[2] While one could also study the scenario where θ^* is known but k is unknown, this case seem less interesting from a technical perspective; since k is a single-dimensional parameter, one can apply a one-dimensional discretization-based algorithm and infer the value of k ([Kleinberg, 2004]). Thus, we focus on the case where the high-dimensional parameter θ^* is unknown while k is known and fixed, which we believe to be the more technically interesting problem.

problem, and must be solved using techniques from measurement error models. Below we show how the generalized method of moments estimator can be used in recovering θ.

4.1 Offline Setting

The causal relationship between (β_t, a_t, x_t, y_t) is summarized in Figure 1 and forms a measurement error model. It can be observed that the contract β_t acts as a valid instrumental variable. In particular, the following moment condition is satisfied

$$\mathbb{E}[\beta_t(y_t - \langle\theta^*, x_t\rangle)] = \mathbf{0}\ (\in \mathbb{R}^d). \tag{1}$$

Therefore an instrumental regression approach and the generalized method of moments estimator can be applied. When given a sequence of offline data, the generalized method of moments estimator takes the following form:

$$\hat{\theta}_T = (B_T^\top X_T)^{-1} B_T^\top Y_T. \tag{2}$$

Here $B_T \in \mathbb{R}^{T\times d}$ is a matrix with rows representing the posted contracts each round; $X_T \in \mathbb{R}^{T\times d}$ is a matrix with rows representing the signals each round; $Y_T \in \mathbb{R}^T$ is a column vector representing the outcomes each round.

Proposition 1. *With probability at least $1-\delta$, the estimation error*

$$\left\|\theta^* - \hat{\theta}_T\right\|_2 \le \frac{\sqrt{dT\log\big(dT/\delta\big)}}{\sigma_{\min}(B_T^\top X_T)}.$$

Proof. Denote $\gamma_t = y_t - \langle\theta^*, x_t\rangle$, and collect values for $\gamma_{1:t-1}$ into the column vector Γ_t.

$$\begin{aligned}\hat{\theta}_T &= (B_T^\top X_T)^{-1} B_T^\top Y_T \\ &= (B_T^\top X_T)^{-1} B_T^\top (X_T\theta^* + \Gamma_T) \\ &= \theta^* + (B_T^\top X_T)^{-1} B_T^\top \Gamma_T.\end{aligned}$$

Hence

$$\left\|\theta^* - \hat{\theta}_T\right\|_2 \le \frac{\left\|B_T^\top \Gamma_T\right\|}{\sigma_{\min}(B_T^\top X_T)}.$$

By a standard concentration inequality (see Appendix A) the numerator can be upper bounded as $\sqrt{dT\log\big(dT/\delta\big)}$ with probability at least $1-\delta$; the proposition then follows. □

4.2 Online Setting

In the previous part, we developed an estimator based on GMM for estimating the parameter θ^*. This applies to the setting when offline observations are available. However, it does not directly give rise to a learning algorithm in the online setting, where the principal needs to *choose* the contract each round. To give some intuition, in this setting, the principal essentially faces an optimal design problem with an exploration-exploitation tradeoff. Specifically, the matrix B_t, consisting of contracts β_t, is the design, and the principal must choose this matrix such that the estimate $\hat{\theta}$ becomes accurate. At the same time, the principal must ensure that the chosen contracts β_t are close to the true optimal contract $\beta^* = \theta^*/k$ in order to guarantee a small cumulative utility loss. We define the cumulative utility loss as the different between the utility achieved with the posted contract and that of posting the optimal contract each round:

$$u_t(\beta^*) - u_t(\beta_t).$$

Algorithm 1 Use Contract as IV: Explore then Commit Algorithm

Input: Distribution $\mathcal{P}$ supported on the feasible contract space $\mathcal{B}$ satisfying Condition 1
Number of exploration rounds $\tau = d\sqrt{T}$
for $t = 1, \dots, \tau$ **do**
  Sample $\beta_t \sim \mathcal{P}$ as the contract
  Observe the signal x_t and realized private benefit y_t
end for
Record the data so far into B_τ, X_τ, Y_τ
Use the GMM estimator and compute the estimate:

$$\hat{\theta} = (B_\tau^\top X_\tau)^{-1} B_\tau Y_\tau.$$

for $t = \tau+1, \dots, T$ **do**
  Set $\beta_t = \hat{\theta}/k$ as the contract
end for

Upper Bound We show that a randomized exploration-type algorithm can achieve $O(d\sqrt{T})$ regret. We introduce the following condition.

Condition 1. *There exists a distribution $\mathcal{P}$ supported on the feasible contract space $\mathcal{B}$ and that there exists constant c_1, such that*

$$\sigma_{\min}(\mathbb{E}_{\beta\sim\mathcal{P}}[\beta\beta^T]) \ge c_1/d.$$

Further, fix any feasible agent's type and letting $a(\beta)$ denote this agent's best response function, we have

$$\sigma_{\min}(\mathbb{E}_{\beta\sim\mathcal{P}}[a(\beta)a(\beta)^T]) \ge c_1/d.$$

By sampling a contract from the distribution $\mathcal{P}$, it can be shown that in expectation the increase of $\sigma_{\min}(B_T^T X_T)$ by adding a new data point can be lower bounded, then by applying Proposition 1, the estimates will converge to the true θ^*.

Remark 5. *The above condition is in fact relatively mild. The first part requires the sampled contract*

comes from a diverse enough distribution. As an example, one can choose distribution $\mathcal{P}$ as the uniform distribution over d linearly independent vectors. The second part of the condition is also satisfied as long as the agent's cost function is well-behaved such that variating β gives sufficient variance along each direction in the agent's response.

Proposition 2. *Suppose Condition 1 holds. There exists an algorithm (Algorithm 1) achieving regret $\widetilde{O}(d\sqrt{T})$ with probability $1 - 1/T$.*

Instead of working with the actual utility loss, we will be bounding the following quantity, which is the square error of β_t each round:

$$\texttt{Reg} = \sum_{t=1}^{T} \|\beta^* - \beta_t\|_2^2.$$

The above can serve as a proxy for the actual cumulative utility loss: $\sum_{t=1}^{T} \big|u_t(\beta^*) - u_t(\beta_t)\big|$. In particular, as long as the function u_t has a bounded second derivative at the global maximum β^*, the cumulative utility loss can be upper bounded by $O(\texttt{Reg})$. Then, each term in Reg can be bounded by Proposition 1. The detailed proof can be found in Section B.

One can also consider a "pure-exploration" problem where, instead of measuring the cumulative utility loss, one can measure the immediate estimation error after T rounds. In this case, one can simply adapt Algorithm 1 to sample $\beta \sim \mathcal{P}$ in each of the T rounds, and then the estimation error can be bounded by the following.

Proposition 3. *Assume Condition 1 holds. Then there exists a pure-exploration algorithm, that after T rounds with probability $1 - \delta$ achieves:*

$$\left\|\theta^* - \hat{\theta}_T\right\|_2 \leq \widetilde{O}(\sqrt{d/T}).$$

Lower Bound Consider a setting where each agent has cost function $c(a) = \frac{1}{2}\|a\|_2^2$, and that this cost function is known to the principal. Then, the agent's best response under the contract β is exactly $a = \beta$. The principal's expected utility when a contract β is posted is then

$$u(\beta) = \langle \theta^*, \beta \rangle - \|\beta\|_2^2.$$

The optimal contract is then $\beta^* = \theta^*/2$ (this is also implied by the previous uniformity result Theorem 1, since the agent's cost function is quadratic and homogeneous of degree 2). The principal now essentially faces a zero-order optimization problem with a quadratic objective function. The contract β is the decision variable, and the realized total revenue is the observed variable. By a result in [Shamir, 2013], the tight regret bound for this problem is $\Theta(d\sqrt{T})$.

Algorithm 2 Use Repeated Observation as IV: Pure Exploitation with Diversity

Parameters: time horizon T, number of tasks d, minimum eigenvalue λ_0, failure probability δ
Epoch e contains $|\tau_e| = \max\{d \cdot 2^e, 8K_0 \ln\big(d/\delta\big)/(\lambda_0), 4\sigma_0^2 d^2 \ln\big(d^2/\delta\big)/(\lambda_0)\}$ rounds
In the 0-th epoch with d rounds, post $\beta = \mathbf{e}_i$ for each $i \in [d]$ once and record $\widetilde{X}_1, X_1, Y_1$
for epoch $e = 1, 2, \cdots, \lceil \log\big(T/d\big) \rceil$ **do**
Compute $\hat{\theta}_e$ using the GMM estimator from data in previous epoch:

$$\hat{\theta}_e = (\widetilde{X}_e^\top X_e)^{-1} X_e Y_e$$

Post $\beta_t = \hat{\theta}_t/k$ for each round in this epoch
Record the data from this epoch into $\widetilde{X}_{e+1}, X_{e+1}, Y_{e+1}$
end for

Proposition 4. *For any (possibly randomized) strategy the principal can adopt for posting contracts, there exists some parameter θ^* such that the term* Reg *can be lower bounded by $\Omega(d\sqrt{T})$.*

5 FASTER CONVERGENCE WITH REPEATED OBSERVATIONS AND DIVERSITY

This section continues the discussion from the previous section, where the principal must learn or estimate θ^* using observational data. The previous section showed that using an instrumental regression approach and using the contract as the instrumental variable, the principal must balance exploration (ensuring that $\sigma_{\min}(B_t^\top X_t)$ grows as t increases) and exploitation (ensuring β_t is close to the true optimal contract). In this section, we show that, under some additional assumptions, the learning rate of the principal can be greatly improved. Informally, we show that when the agents are sufficiently 'diverse', and that when repeated observations are available, the principal can achieve a logarithmic regret, even if the principal is using a 'pure exploitation' algorithm. We state these as the following two assumptions.

Assumption 3. (Repeated observations are available.) *We assume that in each round, the principal can observe two sets of signals, denoted by x_t and $\widetilde{x}_t$. The signals are conditionally-σ_0^2-subgaussian vectors and satisfy the following:*

$$\mathbb{E}[x_t|a_t] = \mathbb{E}[\widetilde{x}_t|a_t] = a_t.$$

They are conditionally independent given the agent's hidden effort a_t:

$$x_t \perp \widetilde{x}_t | a_t.$$

The causal relationship between variables are summarized in Figure 1, with the dashed items included.

Remark 6. *The assumption requires repeated observations, which is often realistic in practical scenarios where the agent's action produces multiple measurable outcomes. We provide several examples where repeated observations naturally arise.* 1. Education: *When measuring student skills or abilities, multiple assessments may target the same underlying competencies (e.g., SAT retakes, midterm and final exams), yielding more than one signal per student.* 2. Gig economy and platform work: *A worker providing services may receive multiple customer ratings, each serving as a noisy measurement of their effort or quality.* 3. Lasting effects: *A content creator's effort on video quality may affect both initial view counts and long-term retention metrics, providing two distinct signals of the same underlying effort.*

We assume that agents are drawn from some distribution $\mathcal{D}$, each element in the support of $\mathcal{D}$ represents the cost function of the agent. It is assumed that for all elements in the support of $\mathcal{D}$, the cost function is homogeneous of degree k (i.e., it satisfies Assumption 1). Our second part of the assumption below states that agents are sufficiently diverse.

Assumption 4. (Agents are sufficiently diverse in their talent.) *Fix any $\beta \in \mathcal{B}$, we have that the following holds*

$$\lambda_{\min}\mathbb{E}_{c(\cdot)\sim\mathcal{D}}[a(\beta; c(\cdot))^\top a(\beta; c(\cdot))] \geq \lambda_0.$$

The above assumption is essentially requiring that there is sufficient diverity in the agent's ability to complete different tasks. We give a concrete example satisfying assumption below.

Example 1. *Consider a setting where agent's have different abilities in different tasks [Thiele, 2010]. Assume the cost functions of the agents have diagonal quadratic forms. Specifically, the cost function of the agent is parameterized by a vector $\kappa \in (\mathbb{R}^d)^+$, so that the cost function of the agent is $c(a) = \frac{1}{2}a^\top \mathtt{diag}(\kappa)^{-1}a$. Then, the agent's best response to a contract β is the maximizer of $\langle \beta, a\rangle - c(a)$, which is equal to*

$$a = \kappa \odot \beta.$$

We assume that the agent cost function in each round, parameterized by κ_t, is drawn from an unknown distribution. Assume the distribution satisfies the following.

$$\lambda_{\min}\mathbb{E}_{\kappa\sim\mathcal{D}}[\kappa\kappa^\top] \geq \lambda.$$

Then suppose each component of β can be lower bounded by some constant, we know that

$$\lambda_{\min}\mathbb{E}[a(\beta)a^\top(\beta)] = \Omega(\lambda).$$

Hence the diversity condition is satisfied.

In addition to the above two assumptions, we require a relatively mild boundedness assumption as below.

Assumption 5. *Fix any $\beta \in \mathcal{B}$. We have the following for any agent type: $\left\|a(\beta)\right\|_2^2 \leq K_0$. Here $a(\beta)$ is the best response to contract β.*

5.1 Offline Setting

The causal relationship is summarized in Figure 1 with the dashed items included. It can be observed that the second set of observations can serve as an instrumental variable for the other, and that we have:

$$\mathbb{E}[\widetilde{x}_t(y_t - \langle x_t, \theta\rangle)] = \mathbf{0}\ (\in \mathbb{R}^d).$$

Assume the principal is given some offline data as $((\beta_i, x_i, \widetilde{x}_i, y_i)_{i=1}^T)$. Then, the principal can use the following GMM estimator to obtain an estimate $\hat{\theta}$:

$$\hat{\theta}_T = (\widetilde{X}_T^\top X_T)^{-1}\widetilde{X}_T^\top Y_T. \tag{3}$$

Here $X_T, \widetilde{X}_T$ are a $T \times d$ matrices where row t is x_t or $\widetilde{x}_t$ respectively; Y_T is a vector with the t-th entry being y_t.

We state the convergence result for the GMM estimator above.

Proposition 5. *With probability $1-\delta$,*

$$\|\theta - \theta^*\|_2 \leq \frac{\sqrt{dT\log\big(dT/\delta\big)}}{\sigma_{\min}(\widetilde{X}_T^\top X_T)}.$$

The proof is straightforward and similar to that of *Proposition* 1.

5.2 Online Setting: Pure exploitation algorithm

We now turn to the online setting, where the principal must learn the optimal contract in real time. In this setting, the proposed algorithm relies directly on the estimator introduced above, without the need for explicit exploration. To reduce computational cost, the algorithm proceeds in epochs whose lengths grow geometrically. At the beginning of each epoch e, the principal updates the estimate $\hat{\theta}_e$ using data collected in previous epochs, and then posts contracts based on this estimate throughout the epoch (see Algorithm 2 for details).

A key distinction from the algorithms in the previous section is that here the principal does not encounter a conventional exploration–exploitation tradeoff. The inherent diversity across agents effectively provides the necessary exploration, allowing the algorithm to focus on exploitation while still ensuring efficient learning.

Theorem 2. *With probability* $1-1/T$*, the cumulative regret of the algorithm is* $\widetilde{O}(d/\lambda_0^2)$*.*

Proof Sketch. Consider an epoch e with length $|\tau_e|$. The minimum singular value of $(\widetilde{X}_e^\top X_e)$ is on the order of $\Omega(|\tau_e|\lambda_0)$. At the same time, the term $\left\|\widetilde{X}_e^\top Y_e\right\|$ can be upper bounded by $\widetilde{O}(\sqrt{d|\tau_e|})$. Hence, for each round in the epoch, the squared estimation error can be upper bounded by:

$$\left\|\hat{\theta}_e - \theta\right\|_2^2 \leq \widetilde{O}(d/(|\tau_e|\lambda_0^2)).$$

Therefore, the regret in each epoch is on the order of $\widetilde{O}(d/\lambda_0^2)$. There are a total of $O(\log T)$ epochs, therefore the total regret is $\widetilde{O}(d/\lambda_0^2)$. □

Remark 7. *Theorem 2 relies critically on the minimum eigenvalue condition in Assumption 4. In practice, one can verify whether this condition holds by examining the singular values of the matrix* $\widetilde{X}_T^\top X_T$ *constructed from observed data. By Proposition 5, if the minimum singular value is large, convergence is fast and pure exploitation performs well. Conversely, if the minimum singular value is small, the diversity condition may not hold in practice, and the principal should introduce explicit exploration into the algorithm. An interesting direction for future work is to design a fully adaptive algorithm that monitors the observed singular values and automatically transitions between exploration and exploitation. Such an algorithm would combine the efficiency of pure exploitation when agent diversity is sufficient with the robustness of explicit exploration when it is not.*

6 CONCLUSION

In this work, we studied the multitasking principal–agent problem under moral hazard. We first established a uniformity result, showing that the optimal contract depends only on the degree of homogeneity. We then demonstrated how an instrumental regression approach, leveraging the generalized method of moments (GMM) estimator, can be used to recover the unknown parameters and identify the optimal contract.

Beyond this specific setting, our results suggest that instrumental methods can be applied more broadly in interactive scenarios whenever moral hazard is present. In policy-making scenarios where hidden actions influence true outcomes, the policy itself may serve as a valid instrumental variable. Such interactions can also be modeled as Stackelberg games, where the leader is the policy-maker, and the follower is an individual or population of the policy target. This perspective opens the door to developing a more general framework for data-driven decision-making in the presence of moral hazard.

Finally, we highlight the broader connection between machine learning and econometrics. While both fields are fundamentally concerned with analyzing data, they have evolved with different emphases and developed complementary methodological toolkits. Our work demonstrates how classical econometric tools—specifically, instrumental variable regression and the generalized method of moments—can be integrated into online learning algorithms to address moral hazard.

References

[Ashrafian and Darzi, 2018] Ashrafian, H. and Darzi, A. (2018). Transforming health policy through machine learning. *PLoS Medicine*, 15(11):e1002692.

[Bastani et al., 2021] Bastani, H., Bayati, M., and Khosravi, K. (2021). Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349.

[Bhattacharyya and Lafontaine, 1995] Bhattacharyya, S. and Lafontaine, F. (1995). Double-sided moral hazard and the nature of share contracts. *The RAND Journal of Economics*, pages 761–781.

[Carroll, 2015] Carroll, G. (2015). Robustness and linear contracts. *American Economic Review*, 105(2):536–563.

[Della Vecchia and Basu, 2023] Della Vecchia, R. and Basu, D. (2023). Online instrumental variable regression: Regret analysis and bandit feedback. *arXiv preprint arXiv:2302.09357*.

[Dong et al., 2018] Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. (2018). Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70.

[Duetting et al., 2024] Duetting, P., Ezra, T., Feldman, M., and Kesselheim, T. (2024). Multi-agent combinatorial contracts. *arXiv preprint arXiv:2405.08260*.

[Dütting et al., 2022] Dütting, P., Ezra, T., Feldman, M., and Kesselheim, T. (2022). Combinatorial contracts. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 815–826. IEEE.

[Dütting et al., 2019] Dütting, P., Roughgarden, T., and Talgam-Cohen, I. (2019). Simple versus optimal contracts. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 369–387.

[Fuller, 2009] Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

[Fuster et al., 2022] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47.

[Google, 2024] Google (2024). Youtube partner earnings overview.

[Greene, 2008] Greene, W. H. (2008). *Econometric analysis*. Pearson/Prentice Hall, Upper Saddle River, N.J, 6th ed. edition.

[Guruganesh et al., 2024] Guruganesh, G., Kolumbus, Y., Schneider, J., Talgam-Cohen, I., Vlatakis-Gkaragkounis, E.-V., Wang, J. R., and Weinberg, S. M. (2024). Contracting with a learning agent. *arXiv preprint arXiv:2401.16198*.

[Hardt et al., 2016] Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. (2016). Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122.

[Harris et al., 2022] Harris, K., Ngo, D. D. T., Stapleton, L., Heidari, H., and Wu, S. (2022). Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pages 8502–8522. PMLR.

[Hilbert et al., 2021] Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., and Stachl, C. (2021). Machine learning for the educational sciences. *Review of Education*, 9(3):e3310.

[Ho et al., 2014] Ho, C.-J., Slivkins, A., and Vaughan, J. W. (2014). Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 359–376.

[Holmström, 1979] Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, pages 74–91.

[Holmstrom, 1982] Holmstrom, B. (1982). Moral hazard in teams. *The Bell journal of economics*, pages 324–340.

[Holmstrom and Milgrom, 1991] Holmstrom, B. and Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52.

[Horowitz and Rosenfeld, 2023] Horowitz, G. and Rosenfeld, N. (2023). Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pages 13233–13253. PMLR.

[Hurley and Adebayo, 2016] Hurley, M. and Adebayo, J. (2016). Credit scoring in the era of big data. *Yale JL & Tech.*, 18:148.

[Jain et al., 2023] Jain, S., Pattanayak, K., Krishnamurthy, V., and Berry, C. (2023). Adaptive eccm for mitigating smart jammers. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

[Kannan et al., 2018] Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. (2018). A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31.

[Kleinberg, 2004] Kleinberg, R. (2004). Nearly tight bounds for the continuum-armed bandit problem. *Advances in Neural Information Processing Systems*, 17.

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

[Miller et al., 2020] Miller, J., Milli, S., and Hardt, M. (2020). Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pages 6917–6926. PMLR.

[Perdomo et al., 2020] Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

[Qin et al., 2022] Qin, Z. T., Zhu, H., and Ye, J. (2022). Reinforcement learning for ridesharing: An extended survey. *Transportation Research Part C: Emerging Technologies*, 144:103852.

[Roth et al., 2016] Roth, A., Ullman, J., and Wu, Z. S. (2016). Watch and learn: Optimizing from revealed preferences feedback. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 949–962.

[Shamir, 2013] Shamir, O. (2013). On the complexity of bandit and derivative-free stochastic convex optimization. In *Conference on Learning Theory*, pages 3–24. PMLR.

[Shavit et al., 2020] Shavit, Y., Edelman, B., and Axelrod, B. (2020). Causal strategic linear regression. In *International Conference on Machine Learning*, pages 8676–8686. PMLR.

[Sivakumar et al., 2022] Sivakumar, V., Zuo, S., and Banerjee, A. (2022). Smoothed adversarial linear contextual bandits with knapsacks. In *International Conference on Machine Learning*, pages 20253–20277. PMLR.

[Thiele, 2010] Thiele, V. (2010). Task-specific abilities in multi-task principal–agent relationships. *Labour Economics*, 17(4):690–698.

[Tropp, 2012] Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434.

[Yu et al., 2022] Yu, M., Yang, Z., and Fan, J. (2022). Strategic decision-making in the presence of information asymmetry: Provably efficient rl with algorithmic instruments. *arXiv preprint arXiv:2208.11040*.

[Zhu et al., 2022] Zhu, B., Bates, S., Yang, Z., Wang, Y., Jiao, J., and Jordan, M. I. (2022). The sample complexity of online contract design. *Proceedings of the 24th ACM Conference on Economics and Computation*.

[Zuo, 2024] Zuo, S. (2024). Harnessing the continuous structure: Utilizing the first-order approach in online contract design. *arXiv preprint arXiv:2403.07143*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A CONCENTRATION TOOLS

The below lemma gives a deviation bound on random vectors.

Lemma 1 (Adapted from [Kannan et al., 2018]). *Let $\gamma_1, \ldots, \gamma_T$ be independent σ-subgaussian random variables. Let $v_1, \ldots, v_T$ be vectors in $\mathbb{R}^d$ with each v_T chosen arbitrarily as a function of $(v_1, \gamma_1), \ldots, (v_{t-1}, \gamma_{t-1})$ subject to $\|v_t\| \leq C$. Then with probability at least $1 - \delta$,*

$$\left\| \sum_{t=1}^{T} \gamma_t v_t \right\| \leq \sqrt{2dC\sigma T \log\big(Td/\delta\big)}.$$

The below lemma bounds the eigenvalues of random matrices; it is adapted from [Tropp, 2012].

Lemma 2 (Matrix Chernoff, adapted from [Tropp, 2012]). *Consider a finite sequence $\{X_k\}$ of independent, random, symmetric matrix with dimension d. Assume that each matrix satisfies*

$$\lambda_{\min} X_k \geq 0, \lambda_{\max} X_k \leq R.$$

Define

$$\mu_{\min} := \lambda_{\min}(\sum_k \mathbb{E} X_k),$$
$$\mu_{\max} := \lambda_{\max}(\sum_k \mathbb{E} X_k).$$

Then

$$\Pr\left[\lambda_{\min}(\sum_k X_k) \leq (1-\delta)\mu_{\min}\right] \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}.$$

The following Corollary may be easier to use than the above Lemma.

Corollary 1. *Using the same setup as above, we have*

$$\Pr\left[\lambda_{\min}(\sum_k X_k) \leq (1-\delta)\mu_{\min}\right] \leq d \exp\Big(-\delta^2 \mu_{\min}/(2R)\Big).$$

B PROOF OF PROPOSITION 2

Proof. We focus on lower bounding the quantity $\sigma_{\min}(B_\tau^\top X_\tau)$. To start, note

$$\sigma_{\min}(B_\tau^\top X_\tau) \geq \sigma_{\min}(B_\tau)\sigma_{\min}(X_\tau).$$

Further,

$$\sigma^2_{\min}(B_\tau) = \lambda_{\min}(\sum_{t=1}^{\tau} \beta_t \beta_t^\top)$$
$$\sigma^2_{\min}(X_\tau) = \lambda_{\min}(\sum_{t=1}^{\tau} x_t x_t^\top)$$

By Corollary 1 and Condition 1, with probability $1-2\delta$, we have

$$\sigma^2_{\min}(B_\tau) > \Omega(\tau \log\big(1/\delta\big)/d),$$
$$\sigma^2_{\min}(X_\tau) > \Omega(\tau \log\big(1/\delta\big)/d).$$

Hence, we have

$$\sigma_{\min}(B_\tau^\top X_\tau) \geq \Omega(\tau/d) = \Omega(\sqrt{T}).$$

Therefore with probability $1-\delta$,

$$\left\|\hat{\theta} - \theta^*\right\| \leq \sqrt{d \log\big(1/\delta\big)/\sqrt{T}}.$$

□

C PROOF OF THEOREM 2

Recall the estimator in each epoch

$$\hat{\theta}_e = \left(\widetilde{X}_e^\top X_e\right)^{-1} \widetilde{X}_e Y_e.$$

Denote $\gamma_t = y_t - \langle x_t, \theta \rangle$, and collect γ_t into the column vector Γ_e. Recall that by Proposition 5, we have

$$\left\|\hat{\theta}_e - \theta\right\|_2 \leq \frac{\left\|\widetilde{X}_e \Gamma_e\right\|}{\sigma_{\min}(\widetilde{X}_e^\top X_e)}. \tag{4}$$

To bound the estimation error, the following will lower bound $\sigma_{\min}(\widetilde{X}_e^\top X_e)$ and upper bound $\left\|\widetilde{X}_e \Gamma_e\right\|$ separately.

Lemma 3. *With high probability δ, $\lambda_{\min}(\sum_{t\in\tau_e} a_t a_t^\top) \geq |\tau_e|\lambda_0/2$.*

Proof. The matrices $\{W_t\}$, where $W_t = a_t a_t^\top$ satisfies the condition in the Lemma with $\lambda_{\max} W_t = \|a_t\|_2^2 \leq K_0$. By the diversity condition,

$$\lambda_{\min}(\mathbb{E} W_t) \geq \lambda_0.$$

By corollary 1,

$$\Pr\left[\lambda_{\min}(\sum_k W_k) \leq |\tau_e|\lambda_0/2\right] \leq d \cdot \exp\big(-|\tau_e|\lambda_0/(8K_0)\big)$$
$$\leq \delta$$

when

$$|\tau_e| \geq \frac{8K_0}{\lambda_0} \cdot \ln\big(d/\delta\big)$$

□

Lemma 4. *With probability $1-2\delta$, for each epoch e with length as specified in Algorithm 2, $\sigma_{\min}(\widetilde{X}_e^\top X_e) \geq |\tau_e|\lambda_0/4$.*

Proof. In the following denote $\varepsilon_t = x_t - a_t, \widetilde{\varepsilon}_t = \widetilde{x}_t - a_t$.

$$\begin{aligned}
\widetilde{X}_e^\top X_e &= \sum_{t\in\tau_e} x_t \widetilde{x}_t^\top \\
&= \sum_{t\in\tau_e} (a_t + \varepsilon_t)(a_t + \widetilde{\varepsilon}_t)^\top \\
&= \sum_{t\in\tau_e} a_t^\top a_t + \varepsilon_t^\top a_t + a_t \varepsilon_t + \varepsilon_t^\top \widetilde{\varepsilon}_t
\end{aligned}$$

The previous lemma showed that with probability $1-\delta$,

$$\lambda_{\min}(\sum_{t\in\tau_e} a_t^\top a_t) \geq |\tau_e|\lambda_0/2.$$

The remaining term

$$\sum_{t\in\tau_e} \varepsilon_t a_t^\top + a_t\widetilde{\varepsilon}_t^\top + \varepsilon\widetilde{\varepsilon}^\top$$

is a random matrix with mean $\mathbf{0}$. By a standard concentration inequality, the absolute value of each entry in the matrix can be upper-bounded as $\sqrt{|\tau_e|\sigma_0^2\ln\big(d^2/\delta\big)}$ with probability δ. Then, since the minimum singular value must be upper bounded by the Frobenius norm:

$$\sigma_{\min}\left(\sum_{t\in\tau_e} \varepsilon_t w_t^\top + w_t\widetilde{\varepsilon}_t^\top + \varepsilon\widetilde{\varepsilon}^\top\right) \leq \sqrt{|\tau_e|\sigma_0^2 d^2\ln\big(d^2/\delta\big)}.$$

The conclusion is that with probability $1-2\delta$,

$$\begin{aligned}\sigma_{\min}(\widetilde{X}_e^\top X_e) &\geq |\tau_e|\lambda_0/2 - \sqrt{|\tau_e|\sigma_0^2 d^2\ln\big(d^2/\delta\big)}\\ &\geq |\tau_e|\lambda_0/4.\end{aligned}$$

This finishes the proof. □

Lemma 5. *With probability* $1-\delta$, $\left\|\widetilde{X}_e\Gamma_e\right\| \leq \sqrt{2\sigma_0 K_0 d|\tau_e|\ln\big(|\tau_e|d/\delta\big)}$.

Proof. Recall $\gamma_t = y_t - \langle x_t, \theta\rangle$, and Γ_e is the column vector collecting the variables γ_t. Note that $\gamma_t\widetilde{x}_t$ is a zero-mean, $\sigma_0 K_0$-subgaussian random vector and independent conditioned on all previous steps. The result then follows from the standard concentration inequality (see Section A). □

Theorem 3. *With probability* $1-1/T$, *the total regret is* $\widetilde{O}(d/\lambda_0^2)$.

Proof. By Lemma 5, Lemma 4, and Equation (4), with probability $1-3\delta$, the regret in each epoch can be upper bounded as $|\tau_e|\cdot\widetilde{O}(d/(|\tau_e|\lambda_0^2)) = \widetilde{O}(d/\lambda_0^2)$. Choose some appropriate δ, e.g., $\delta < O(1/T^2)$. Since there are at most $\log T$ epochs, the total regret will be $\widetilde{O}(d/\lambda_0^2)$. □

D NUMERICAL EXPERIMENTS

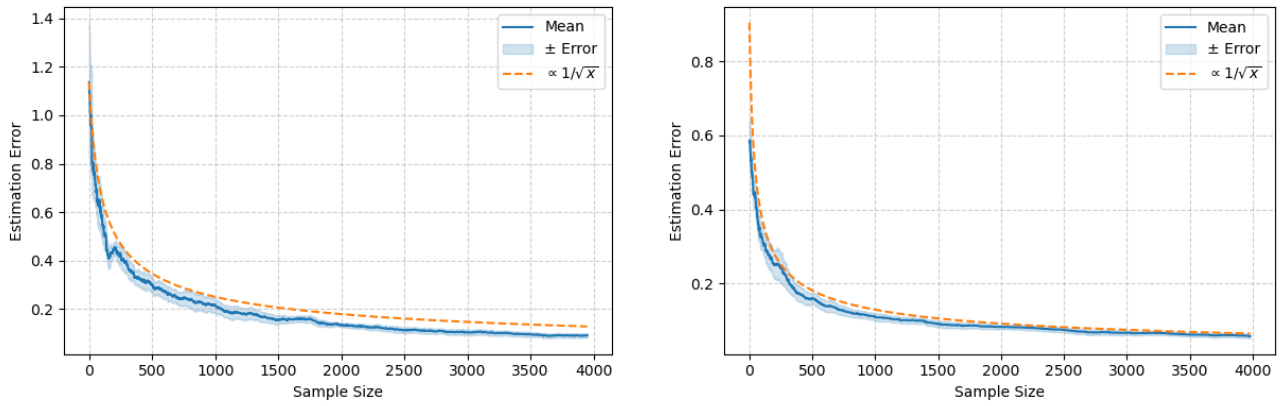
We perform numerical simulations that test the performance of the generalized method of moments based estimator.[3] We choose $d=5$ and $\theta^* = [1,2,3,4,5]$. We choose the agent's cost function as

$$c(a) = \sum_{i=1}^{d} \kappa_i a_i^2,$$

where each κ_i is sampled uniformly from the set $\{1,10\}$. We study how the estimation error decreases as more samples are received.

In the first setting, we use the contract as the instrumental variable and use the estimator as in Eq. Equation (2). In the second setting, we assume repeated observations are available and use the estimator as in Eq. Equation (3). We plot the relationship in Figure 2. As we can observe, the estimation error $\left\|\hat{\theta}-\theta^*\right\|$ decreases at a rate around $O(1/\sqrt{T})$.

[3]Code can be found at `https://anonymous.4open.science/r/GMM_learning-8E3F/README.md`.



(a) (b)

Figure 2: The estimation error $\left\|\hat{\theta} - \theta^*\right\|$ as sample size increases. (a): Using the contract as the instrumental variable and estimating $\hat{\theta}$ as $\hat{\theta} = (B_T X_T)^{-1} X_T Y_T$. (Proposition 1). (b): Using repeated observations as the instrumental variable and estimating $\hat{\theta}$ as $\hat{\theta} = (\widetilde{X}_T X_T)^{-1} \widetilde{X}_T Y_T$ (Proposition 5). In both cases, we observe the estimation error decreases at a rate around $\Theta(1/\sqrt{T})$ (where T is the sample size).