ASYMPTOTIC DEGRADATION OF LINEAR REGRESSION ESTIMATES WITH STRATEGIC DATA SOURCES[*]

Nicolas Gast
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
nicolas.gast@inria.fr

Patrick Loiseau
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
patrick.loiseau@inria.fr

Panayotis Mertikopoulos
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, Criteo AI Lab
panayotis.mertikopoulos@inria.fr

Benjamin Roussillon
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG
benjamin.roussillon@inria.fr

June 29, 2021

ABSTRACT

We consider the problem of linear regression from strategic data sources with a public good component, i.e., when data is provided by strategic agents who seek to minimize an individual provision cost for increasing their data's precision while benefiting from the model's overall precision. In contrast to previous works, our model tackles the case where there is uncertainty on the attributes characterizing the agents' data—a critical aspect of the problem when the number of agents is large. We provide a characterization of the game's equilibrium, which reveals an interesting connection with optimal design. Subsequently, we focus on the asymptotic behavior of the covariance of the linear regression parameters estimated via generalized least squares as the number of data sources becomes large. We provide upper and lower bounds for this covariance matrix and we show that, when the agents' provision costs are superlinear, the model's covariance converges to zero but at a slower rate relative to virtually all learning problems with exogenous data. On the other hand, if the agents' provision costs are linear, this covariance fails to converge. This shows that even the basic property of consistency of generalized least squares estimators is compromised when the data sources are strategic.

1 Introduction

Consider a linear regression problem consisting of n data points $(x_i, y_i)_{i\in\{1,\cdots,n\}}$, where x_i is a vector of independent variables and $y_i \in \mathbb{R}$ is the corresponding response variable. Assuming that these variables are linked by a linear model

$$y_i = \boldsymbol{\beta}^\top x_i + \varepsilon_i \tag{1}$$

where the ε_i are mutually independent noise variables, an analyst aims at estimating the parameter vector $\boldsymbol{\beta}$. If the variance of each ε_i is known and uniformly bounded in n (but not necessarily identical across i), the most widespread

[*] This work was supported by the French National Research Agency through the "Investissements d'avenir" program (ANR-15-IDEX-02) and through grant ANR-16-TERC0012; by the DGA; by the Alexander von Humboldt Foundation.

algorithm to estimate $\boldsymbol{\beta}$ is the famous generalized least squares (GLS) estimator, which is well-known to enjoy important statistical properties. In particular, for any fixed n, Aitken's theorem [2] shows that GLS is *BLUE*, i.e., best among linear unbiased estimators. GLS is also *consistent* (i.e., it converges in probability towards $\boldsymbol{\beta}$ as $n \to \infty$), and its covariance matrix decreases to zero at a rate $\Theta(1/n)$[2] [27, 22].

In a number of recent applications, however, the assumptions underlying those statistical properties do not hold because the data is provided by strategic agents who incur a cost for providing high-precision (i.e., low-variance) data. There can be multiple reasons. If the data is sensitive personal information (as in medical applications), revealing it with high precision entails a privacy cost that might incentivize individuals to decrease the disclosure precision [47, 19]. Also, producing high-precision data may require a certain amount of effort (possibly monetary): this is the case in crowdsourcing [16] or recommender systems [21, 4, 29] where providing content or feedback requires effort, or in applications where the data is produced by costly computations.

Considerations of this kind have become central in an emerging literature on learning with *strategic data sources*, i.e., when the precision of the provided data incurs a cost at the agent providing it. This literature mainly examines the design of monetary incentive mechanisms to optimize the model's error assuming that agents maximize their incentives minus their individual provision costs, see e.g., [8, 35, 48] and references therein. In many applications, however, the underlying model also has a *public good component*—i.e., the agents *also* benefit from the model's precision. This is the case in recommender systems (where users benefit from the overall service quality), medical applications (where individuals benefit from the data analysis through improved treatments or better healthcare advice), federated learning [49, 33, 25], etc. An additional issue in such applications is that the number of participating agents is typically large, so there is a commensurate degree of uncertainty regarding the state or incentives of other agents. In this context, the validity of standard results on linear regression are longer guaranteed: in particular, recent work has shown that the GLS estimator is no longer BLUE in the presence of strategic data sources [24]. Going deeper, this leads to the following open questions: *Does GLS remain consistent in the present of strategic agents? And, if so, does it still enjoy a $\Theta(1/n)$ convergence rate as in the non-strategic case?*

Our contributions. In this paper, we provide answers to these questions by means of a data provision model that accounts for both factors identified above: a public good component and uncertainty regarding the agents' types (their private data). Specifically, we propose a *linear regression game* in which the i-th agent's type is characterized by a d-dimensional *attribute vector* $x_i \in \mathbb{R}^d$ which is drawn i.i.d. across agents (but is otherwise assumed to be private information). This attribute vector is the primitive data associated to each agent and forms the basis of the linear model (1): each agent decides the precision of the response variable $y_i \in \mathbb{R}$ that is revealed to the analyst as a function of x_i. This choice is intended to balance the agent's data provision cost against a public good benefit that depends on the precision of the overall model; importantly, this choice is also made under uncertainty, because the players' attribute vectors are not assumed a priori known.

In this setting, we obtain the following general results:

1. We provide an explicit characterization of the game's equilibria for different families of data provision cost functions. Specifically, if the data provision costs are linear in the precision of the disclosed response, the equilibrium distribution of precisions over the space of attribute vectors corresponds to the solution of an optimal design problem. By contrast, this characterization is no longer valid if the data provision costs are superlinear.
2. Subsequently, we analyze the precision of the estimated model in the limit $n \to \infty$. In this asymptotic regime, our main result is that, for superlinear costs, the GLS estimator remains consistent, but its covariance decreases to zero at a rate *slower* than the standard $\Theta(1/n)$ rate. Surprisingly, as the data provision costs become approximately linear, this rate becomes progressively slower, to the point that the GLS estimator *fails to be consistent* if the data provision costs are linear.

Our analysis reveals that the key reason behind this asymptotic degradation of the GLS estimator is the following: as $n \to \infty$, the response provided by each agent at equilibrium tends to deteriorate because of the increasing free-riding effect inherent to public good games. When the data provision costs are linear, this decrease cannot be compensated by the increase in the number of data points, so the GLS estimator becomes inconsistent. In this regard, our results illustrate clearly how free-riding can disrupt even the most fundamental statistical properties of GLS estimators.

Related work. There is a growing body of works on scenarios where one wants to learn from data provided by sources that choose their effort when generating data [8, 36, 35, 48]. These works assume that the data sources maximize a monetary incentive minus effort exerted and look for mechanisms that minimize the model's error under the assumption that the analyst collecting data cannot see the effort exerted by the data sources. The data elicitation

[2] The notation $g(n) = \Theta(f(n))$ indicates that functions f and g grow at the same rate as n goes to infinity.

and crowdsourcing literatures contain similar mechanism design problems for cases where either the effort exerted or the data report (or both) are unverifiable [23, 16, 44, 34]. More broadly, there is an important literature on mechanism design for statistical estimation problems that assumes that the data sources are strategic in some way, notably for cases where agents may lie on their cost for revealing data [1, 10, 13] (see also related problems of mechanism design in the context of differential privacy [26, 20]).

Several works analyze mechanism design problems related to linear regression with strategic data sources, where the agents directly report their response variable y_i (or their input variable x_i) and may lie about it or strategically optimize it [42, 17, 9, 11, 5, 30, 45, 12] (see also similar problems in the context of classification [37, 28, 18, 39, 32, 50, 38, 46, 7]). In contrast, we assume that the agents choose the precision of the data provided. More importantly, the fundamental difference is that those works all assume that the agents are motivated by the accuracy or decision of the learned model in their own direction while we assume that agents equally benefit from the public good component.

Games with a public good component have been the subject of many studies in economics (see [40] and references therein); however, in all these works the public good is simply the sum of contributions from all agents. To the best of our knowledge, the only work that considers a public good component in the context of learning from strategic data sources is [24] (see also an earlier version of the same in [31] and [14] in the simple case of estimating a population average), which is perhaps the closest antecedent to our paper. The authors of [24] propose a game-theoretic model with a public good component and *common knowledge* of player types (i.e., the agents' attribute vectors x_i are public). We build on their model, but we introduce uncertainty on the agents' attributes (i.e., they are considered private)—which is crucial to frame our main questions (link to optimal design and asymptotic precision) rigorously. More importantly though, our analysis is of a different nature than that of [24], both technically and conceptually. The analysis of [24] only concerns the game's *price of stability* (PoS) and the validity of Aitken's theorem. Specifically, [24] shows that, for any given n, GLS may fail to be BLUE (i.e., Aitken's theorem fails), but it is otherwise "uniformly close" to optimal: the improvement ratio relative to any other linear unbiased estimator is independent of n, so GLS is still "order optimal" in their model. In contrast, our results show that GLS (and other linear unbiased estimator by proxy) does not even produce the same convergence *rate* as non-strategic regression. This means that characterizing this convergence rate is ultimately more important to the analyst than choosing an ad-hoc linear unbiased estimator which can only lead to constant-term multiplicative improvement. To ease the comparison of our work to [24], we include in Section 3 an adaptation of their results for our model with uncertainty, we then provide a detailed technical comparison in further sections that contain our new results. Note finally that our model with uncertainty enables uncovering an interesting connection to optimal design, which [24] does not touch upon. We also note that our game has the structure of an aggregative game in the sense of [15] which, however, does not offer any further insights.

2 Problem setup and assumptions

The linear regression game We consider a model of strategic data sharing in which a group of n agents wants to collectively learn a linear model $y \approx \boldsymbol{\beta}^\top x$. Here, x is a d-dimensional vector, y is a scalar and the vector $\boldsymbol{\beta}$ represents the weights of the linear model that agents want to estimate. Each agent $i \in N := \{1, \cdots, n\}$ is associated to an *attribute vector* x_i which is drawn i.i.d. across agents from a set of possible attribute profiles $\mathcal{X} \subseteq \mathbb{R}^d$. Then, based on these attributes, each agent can select a *precision level* $\lambda_i(x_i) \in \mathbb{R}_+$ and produce an unbiased estimate $\hat{y}_i$ of $\boldsymbol{\beta}^\top x_i$ with this precision.[3] More explicitly, the response variable reported by the i-th agent is

$$\hat{y}_i = \boldsymbol{\beta}^\top x_i + \varepsilon_i, \tag{2}$$

where ε_i is an error term of mean 0 and variance $1/\lambda_i(x_i)$. Agents send this estimate $\hat{y}_i$, along with their values of x_i and of the precision $\lambda_i(x_i)$ to an aggregator that publicly discloses an estimate $\hat{\boldsymbol{\beta}}$. The errors terms ε_i are assumed to be independent, but we do not make any further assumption on their distribution. We then posit that each agent tries to balance a trade-off between two components:

1. *Idiosyncratic cost:* The value $\hat{y}_i$ may be either sensitive or costly to produce. It is sensitive for example when it represents a disease likelihood, a total debt or any attribute that might hurt the agent if it is disclosed with full precision (e.g., by a potential increase in cost of health insurance): here, the agent possesses a value y_i but only reveals a noisy version of it $\hat{y}_i$. It is costly to produce when it is the result of a simulation involving heavy computations, or when it requires human work. We represent all these scenarios by assuming that releasing an estimator $\hat{y}_i$ with precision $\lambda_i(x_i)$ induces a cost $c_i(\lambda_i(x_i))$ to agent i. We refer to it as the ***(data) provision cost***.
2. *Public good benefit:* A key feature of our model is that all agents benefit from the learned model $\hat{\boldsymbol{\beta}}$. For example, in a medical context, agents would be interested to know that a given disease is correlated to their weight or cholesterol

[3] We can impose an upper bound λ_{max} on the precision that an agent can choose; our results would still hold for large-enough n as the constraint is never binding. In the sequel, we assume $\lambda_{\text{max}} = \infty$ to simplify the exposition.

level; in recommender systems, agents might be interested to know what affects the good rating of a restaurant; etc. We model this benefit as a *public good*, that is, we assume that each agent benefits equally from the estimated model's precision—which, in turn, depends on each agent's prescribed precision. As it is easier to maintain a cost-oriented perspective, we represent this by considering that each agents incurs a cost $C_{\mathrm{estim}}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = [\lambda_i]_{i\in N}$. We refer to it as the ***estimation cost***.

Remark. This model is particularly relevant in the context of federated learning [49]. There, each agent performs a local estimation and the estimations are combined to get a model. This paradigm can be used for reasons of efficiency (many agents, perform a local optimization [33]) or privacy (agents want to compute a joint representative model without explicitly having to share their personal data [25]). Our model can be viewed as an instance of both cases.

To proceed, we model the collective behavior of agents by considering a game in which each agent $i \in N$ chooses their strategy $\lambda_i : \mathcal{X} \to \mathbb{R}_+$ to minimize their cost $J_i(\lambda_i, \boldsymbol{\lambda}_{-i})$, defined here as

$$J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = \mathbb{E}\left[c_i(\lambda_i(x_i))\right] + C_{\mathrm{estim}}(\boldsymbol{\lambda}), \tag{3}$$

where the expectation is taken with respect to the law μ of the attribute vectors x_i. Here, $\boldsymbol{\lambda}_{-i}$ denotes the collection of strategies of all agents except the i-th one, and $\boldsymbol{\lambda} = (\lambda_i, \boldsymbol{\lambda}_{-i})$ will be called a *precision profile*. Note that, given that agent i chooses the function $\lambda_i : \mathcal{X} \to \mathbb{R}_+$, minimizing the expected provision cost $\mathbb{E}\left[c_i(\lambda_i(x_i))\right]$ as in (3) is equivalent to minimizing the cost for each value of x_i separately. On the other hand, the definition of C_{estim} is given below and involves an expectation on all agents' attributes, which models the uncertainty of an agent about other agents' attributes.

The setting described above defines a game that we refer to as the *linear regression game*. We emphasize that the strategy of each agent is a function $\lambda_i : \mathcal{X} \to \mathbb{R}_+$, i.e., each player's strategy space is the $|\mathcal{X}|$-dimensional orthant $\mathbb{R}_+^{\mathcal{X}}$. Throughout the paper, to avoid confusion, we will denote such functions with the greek letter λ and we will use the latin letter ℓ for scalar values such as $\lambda_i(x_i)$. In the sequel, we analyze the Nash equilibrium of this linear regression game. A precision profile $\boldsymbol{\lambda}^*$ is a Nash equilibrium of the game if for all $i \in N$, λ_i^* minimizes $J_i(\lambda_i, \boldsymbol{\lambda}_{-i}^*)$.

Generalized least squares, definition of C_{estim} The analyst receives the n triplets $(x_i, \hat{y}_i, \lambda_i(x_i))$ and uses them to produce an estimate $\hat{\boldsymbol{\beta}}$ that is then sent to the agents. We assume that the analyst computes this estimate using *generalized least squares* (GLS) and denote it $\hat{\boldsymbol{\beta}}_{\mathtt{GLS}}$. GLS is a generalization of the least squares regression to heteroscedastic data, that is, when the precisions $\lambda_i(x_i)$ of the different data points are different. It is one of the most widespread estimators for this scenario, in particular because by Aitken's theorem, GLS is optimal in that, for given precisions, it has the smallest covariance (in the positive semi definite sense) among all linear unbiased estimators [2]. The covariance of GLS is independent of $\hat{y}_i$ and equal to $\left(\sum_i \lambda_i(x_i) x_i x_i^\top\right)^{-1}$. Note that this quantity is well defined only if the matrix $\sum_i \lambda_i(x_i) x_i x_i^\top$, called the information matrix, is invertible; if not, the estimator $\hat{\boldsymbol{\beta}}_{\mathtt{GLS}}$ is not unique as the generalized least squares problem has infinitely many solutions.

Recall that the values x_i are generated randomly according to a common distribution μ on $\mathcal{X}$. In what follows, we consider that the estimation cost is a function of the expected information matrix, that is:

$$C_{\mathrm{estim}}(\boldsymbol{\lambda}) = F\left(\left(\mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]\right)^{-1}\right), \tag{4}$$

where $F : S_+^d \to \mathbb{R}_+$ is a so-called *scalarization function* that maps the covariance of the estimator to a cost (we denote by S_+^d the set of positive semidefinite matrices of size $d \times d$). Scalarizations are standard in optimal design (see Section 4), and they include standard metrics of a model's quality (such as the mean squared error) as special cases—see details in Appendix B.

The public good component (4) is a function of the inverse of the expected information matrix. In particular, agent i is included in this expectation, so they minimize a function that includes their individual contribution x_i. In this regard, (4) can be seen as a "middle ground" between the approach of [24] (which assumes complete information of the x_i of each agent), and a Bayesian model where agents would optimize over $\mathbb{E}\left[F((\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top)^{-1})\right]$. The former is impractical in asymptotic settings while the latter introduces a series of modeling artifacts due to the nonzero probability of encountering an ill-defined linear regression problem when drawing vectors from a finite set.

Compared to the above, our model requires the same information as the Bayesian framework, but it does not face the same issues. In addition, it is possible to establish a precise equivalence between our game and the complete information game when the number of players is large—see Appendix D for the details. Note also that our model relies on GLS, which requires the analyst to know the precision of each data point. We stress here, however, that our main result also applies to the *ordinary least squares (OLS)* estimator, which does not require this information. We discuss this in detail in Section 5.

Technical assumptions Through our analysis, we make the following assumptions:

Assumption 1. The set $\mathcal{X}$ is finite, μ has full support on $\mathcal{X}$, and $\mathbb{E}\left[x_i x_i^\top\right]$ is positive definite.

Assumption 2. The scalarization $F : S_+^d \to \mathbb{R}_+$ is non-negative, increasing in the positive semidefinite order, and convex. F is homogeneous of degree q: $\forall a > 0, \forall V \in S_+^d$, $F(aV) = a^q F(V)$.

Assumption 3. The provision costs $c_i : \mathbb{R}_+ \to \mathbb{R}_+$ are non-negative, increasing, and convex.

Assumption 1 is a technical assumption that guarantees that the game is well defined and non-trivial. Specifically, the condition $\mathbb{E}\left[x_i x_i^\top\right] \succ 0$ simply implies that the information matrix is invertible for some $\boldsymbol{\lambda}$, while the finiteness of $\mathcal{X}$ avoids subtle compactness issues in the definition of an equilibrium. Assumption 2 ensures that the estimation cost is non-decreasing and convex, which is a very standard setting in game theory. Together with the homogeneity assumption, these conditions are satisfied by all the commonly used scalarizations in statistics and optimal design such as the trace ($q = 1$, A-optimal design), squared Frobenius norm ($q = 2$), or mean squared error ($q = 1$, I/V-optimal design).

Assumption 3 is also common. The convexity of the provision costs implies that there is a non-decreasing marginal cost to increase the precision of the data provided. This is reasonable and includes both the linear and superlinear cases, each being relevant. For instance, if data points represent an average over multiple measurements, the precision depends linearly on the number of measurements. If the precision depends on simulations (e.g., involving a discrete search of a continous space), obtaining a higher precision might require a polynomial increase in computation time.

3 Preliminaries and First Results

In this section, we discuss structural results for the linear regression game. In the spirit of [24], we show that our game is a potential game and provide a bound on its price of anarchy.

For a given precision profile, we define $\phi(\lambda_i, \boldsymbol{\lambda}_{-i})$

$$\phi(\boldsymbol{\lambda}) = \sum_{j=1}^{n} \mathbb{E}\left[c_j(\lambda_j(x))\right] + C_{\text{estim}}(\boldsymbol{\lambda}). \tag{5}$$

Using the form of $J_i(\lambda_i, \boldsymbol{\lambda}_{-i})$ in (3), a strategy λ_i minimizes $J_i(\lambda_i, \boldsymbol{\lambda}_{-i})$ over all possible strategies λ_i (for a fixed $\boldsymbol{\lambda}_{-i}$) if and only if it minimizes (5). Since function ϕ is independent of i, this shows that the game is a potential game [41] and ϕ is the potential of the game. As stated in the next proposition (whose proof is in Appendix C), expressing the game as a potential game simplifies the study of its Nash equilibria by transforming it into the easier problem of studying the minima of a convex function.

Proposition 1. *Under Assumptions 1, 2, and 3, a precision profile $\boldsymbol{\lambda}^*$ is a Nash equilibrium of the linear regression game if and only if it minimizes ϕ. Such an equilibrium exists. It is unique if all provision cost functions c_i are strictly convex. When there are multiple equilibria, the estimation cost $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ does not depend on the equilibrium.*

The price of anarchy (PoA) is a standard concept in game theory that characterizes the degradation of performance due to players' selfish behavior. It is the ratio between the total cost of the worst Nash equilibrium and the minimal achievable total cost. In our setting the total cost is $C_{\text{social}}(\boldsymbol{\lambda}) = \sum_i \mathbb{E}\left[c_i(\lambda_i(x))\right] + nC_{\text{estim}}(\boldsymbol{\lambda})$. Hence, denoting as $\texttt{NE} \subseteq \{\lambda : \mathcal{X} \to \mathbb{R}_+\}^n$ the set of Nash equilibria,

$$\texttt{PoA} = \frac{\max_{\boldsymbol{\lambda}^* \in \texttt{NE}} C_{\text{social}}(\boldsymbol{\lambda}^*)}{\min_{\boldsymbol{\lambda} \in \{\lambda : \mathcal{X} \to \mathbb{R}_+\}^n} C_{\text{social}}(\boldsymbol{\lambda})}.$$

Our linear regression game has the same PoA bound as that of [24] with a similar proof (see App. C):

Theorem 1. *In addition to Assumptions 1, 2 and 3, assume that there exist $p_{\min} \geq 1$ such that for all $i \in N, a > 1, \ell > 0$: $c_i(a\ell) \geq a^{p_{\min}} c_i(\ell)$. Then, the price of anarchy satisfies $\texttt{PoA} \leq n^{\frac{q}{p_{\min}+q}}$. Additionally, for all $\varepsilon > 0$, there exists a game such that $\texttt{PoA} \geq n^{\frac{q}{p_{\min}+q}}(1 - \varepsilon)$.*

Remark. We should note here that the above result bounds the game's price of anarchy whereas the study of [24] concerns the price of stability ($\texttt{PoS}$) of a suitable variant of our game without uncertainty. In contrast to the price of anarchy, the price of stability compares the social optimum cost to that of the *best* Nash equilibrium. In general, these two measures of selfishness can vary wildly, but in the linear regression game under study, they conincide; this is due to the fact that although we may have multiple equilibria, all equilibria have the same cost (from Proposition 1). This differs from [24] where there exists a unique non-trivial equilibrium, but there also exist trivial equilibria with infinite costs.

4 Characterization of the equilibrium

We now characterize how the attribute distribution μ affects the precision given by agents at equilibrium. We show that when provision costs are linear, the precision given by each agent can be mapped to the solution of an optimal design. This is no longer true when provision costs are not linear.

In optimal design [43, 3, 6], an analyst chooses the x_i's of the set of (non-strategic) data sources in order to maximize the quality of the linear model estimated via a scalarization of the covariance matrix. Formally, the optimal design problem for the scalarization F (see Appendix B for details) and the design space $\mathcal{X}$ is to find a probability measure ν^* that minimizes:

$$\nu^* \in \arg\min_{\nu} F\left(\left(\sum_{x\in\mathcal{X}} xx^\top \nu(x)\right)^{-1}\right). \tag{6}$$

In our linear regression game, the agents have an incentive to produce a useful information matrix to minimize the estimation cost but they are limited by the inherent allocation μ of attribute vectors and by the provision costs c_i. An equilibrium is a minimum of the potential (5) that contains the estimation cost $C_{\mathrm{estim}}(\boldsymbol{\lambda})$, which can be rewritten as:

$$C_{\mathrm{estim}}(\boldsymbol{\lambda}) = F\left(\left(\sum_{x\in\mathcal{X}} xx^\top \sum_{i\in N} \lambda_i(x)\mu(x)\right)^{-1}\right). \tag{7}$$

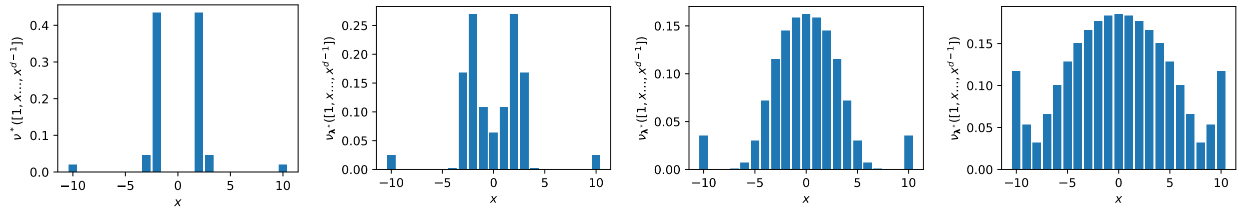
The similarity of (6) and (7) suggests a link between the Nash equilibria of the linear regression game and the solutions of the optimal design problem on $\mathcal{X}$ by interpreting $\sum_i \lambda_i(x)\mu(x)$ as a design $\nu(x)$:

Theorem 2. *Consider a linear regression game that satisfies Assumptions 1 and 2 and such that all provision costs are linear (*i.e.*, $c_i(\ell) = a_i\ell$ for all $i \in N$ and $\ell \in \mathbb{R}_+$, where a_i is a constant). Let $\boldsymbol{\lambda}^*$ be a Nash equilibrium and let $\nu_{\boldsymbol{\lambda}^*}$ be the measure such that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_{i\in N} \lambda_i^*(x)\mu(x)$ for all $x \in \mathcal{X}$. Then, the probability measure defined by $\nu_{\boldsymbol{\lambda}^*}(x)/\sum_{y\in\mathcal{X}} \nu_{\boldsymbol{\lambda}^*}(y)$ is an optimal design of* (6).

Sketch of proof. A detailed proof is given in Appendix C. The main idea is to see the minimization problem (6) as an optimization problem with constraint $\sum_{x\in\mathcal{X}} \nu(x) = 1$. When the provision costs are linear, the potential ϕ is a Lagrangian of this optimization problem with a dual variable $\min_{i\in N} a_i$. The fact that $\nu_{\boldsymbol{\lambda}^*}$ is proportional to an optimal design is then a consequence of the homogeneity of the scalarization (Assumption 2). □

While the shape of $\nu_{\boldsymbol{\lambda}^*}$ for an equilibrium $\boldsymbol{\lambda}^*$ is that of an optimal design, the total expected precision $\sum_{x\in\mathcal{X}} \nu_{\boldsymbol{\lambda}^*}(x)$ depends on the provision costs. Theorem 2 merely states that agents contribute proportionally to an optimal design but does not characterize how the total precision depends on the number of agents or on the agents' costs. We leave this discussion to Section 5 (in particular Theorem 3). This theorem also implies that with linear costs, agents that have data points which do not belong to an optimal design are pure free-riders. On the contrary, this is no longer the case with superlinear provision costs. We illustrate this in Figure 1 where we observe that the difference between the maximum and minimum precision given depending on the data point shrinks as the exponent of the cost grows.

The particular connection between optimal design and Nash equilibria exhibited in Theorem 2 is tightly connected to the linearity of provision costs. When costs are strictly convex, the allocation of precision across $\mathcal{X}$ at equilibrium is in general suboptimal. For instance, if an agent has a provision cost $c_i(\ell) = \ell^p$ with $p > 1$, then the derivative of this provision costs at 0 is zero, $c_i'(0) = 0$. In such a case, this agent will provide a positive precision, $\lambda_i(x) > 0$, for all attribute vectors $x \in \mathcal{X}$ even though the support of an optimal design might be smaller than $\mathcal{X}$. We illustrate the case of nonlinear costs in a polynomial regression setting that is an instance of our linear regression game as follows. Let $\mathcal{X} = [1, x, \cdots, x^{d-1}]^\top$ be the set of attribute vectors with $x \in \{-10 \ldots 10\}$. We compare in Figure 1 the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 2 to the optimal design for different monomial provision costs ($c(\ell) = \ell^p$). We set μ to the uniform distribution on $\mathcal{X}$, $d = 4$, $n = 10$ and the scalarization F is the trace. Other parameters give similar results (see App. G). We observe that when the provision costs are near-linear ($p = 1.01$), the precision function is similar to the optimal design yet different. When $p = 1.2$ or $p = 3$, however, the precision for the vector $[1, 0, \ldots, 0]$ is maximal whereas the optimal design sets a weight 0 to it. Intuitively, the convexity of provision costs yields a more spread-out allocation of precision than the optimal design. This shows that equilibrium can be different from optimal design, even when costs are close to linear.



(a) Optimal design ν^*

(b) $\nu_{\boldsymbol{\lambda}^*}$ when $c_i(\ell) = \ell^{1.01}$

(c) $\nu_{\boldsymbol{\lambda}^*}$ when $c_i(\ell) = \ell^{1.2}$

(d) $\nu_{\boldsymbol{\lambda}^*}$ when $c_i(\ell) = \ell^3$

Figure 1: Optimal design ν^* and allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$.

5 Asymptotic results

The previous section shows that linear provision costs drive agents to allocate their precision proportionally to an optimal design, while non-linear costs lead to a non-optimal allocation. In this section, we show that the situation is *radically different when considering the total model precision.*

The case of identical agents To gain intuition, we first consider agents with identical monomial costs. In this setting, the equilibrium for the n-agent game is obtained by scaling the solution of the optimization problem that would correspond to a single-agent game (the full proof is in Appendix C):

Proposition 2. *Consider a linear regression game satisfying Assumptions 1 and 2 and such that for all agent i and precision ℓ: $c_i(\ell) = \ell^p$ with $p \geq 1$. Let $\lambda_{single} = \arg\min_{\lambda \in \mathbb{R}_+^{|\mathcal{X}|}} \mathbb{E}\left[\lambda(x)^p\right] + C_{\text{estim}}(\lambda)$.*

(i) The precision profile $\boldsymbol{\lambda}^*$ with $\lambda_i^* = n^{-\frac{1+q}{p+q}} \lambda_{single}$ for all $i = 1, \ldots, n$ is a Nash equilibrium.

(ii) The estimation cost at equilibrium is $C_{\text{estim}}(\boldsymbol{\lambda}^*) = n^{-q\frac{p-1}{p+q}} C_{\text{estim}}(\lambda_{single})$.

Proposition $2(i)$ illustrates a major difference between the strategic and non-strategic settings. Indeed, in a non-strategic setting, each agent would provide data with a fixed precision, say $\lambda_{\text{ns}}(x) = \ell_{\text{ns}}$ for all x. By contrast, in the presence of strategic data sources, *the equilibrium precision given by each agent goes to* 0 *when the number of agents grows.* Moreover, the convergence rate is governed by the parameters p and q: when $p \to \infty$, the precision of each agent is almost constant, similar to the non-strategic case; instead, with linear costs ($p = 1$), the precision given by each agent goes to 0 at a $\Theta(1/n)$ rate.

Thus, when aggregating the data from n non-strategic data sources, the estimation cost would be

$$C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}}) = n^{-q} C_{\text{estim}}(\lambda_{\text{ns}}) \tag{8}$$

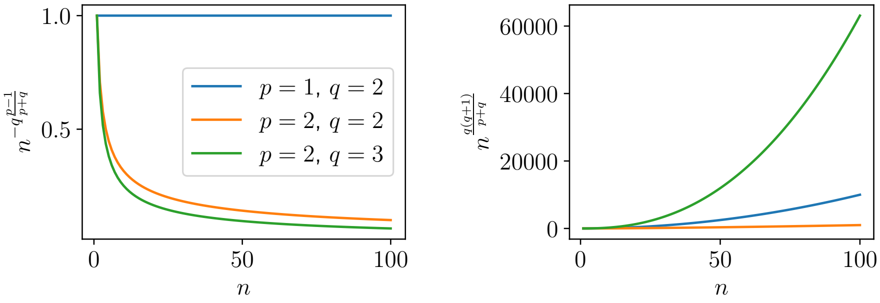
where $\boldsymbol{\lambda}_{\text{ns}} = (\lambda_{\text{ns}}, \cdots, \lambda_{\text{ns}})$ (which corresponds to the standard $1/n$ rate if $q = 1$). By contrast, when aggregating the data from n strategic data sources, Proposition $2(ii)$ shows that the rate of decrease is smaller, again governed by the parameters p and q. In the extreme, when the costs are linear ($p = 1$), *the estimation cost does not even go to* 0 *as* $n \to \infty$. This shows that GLS is not consistent in the presence of strategic data sources with linear provision costs: in this case, the estimator's covariance does not vanish as the number of data sources grows large.

To quantify how strategic considerations lead to a degradation of the GLS estimator, we can consider the ratio between the strategic and non-strategic estimation costs:

$$C_{\text{estim}}(\boldsymbol{\lambda}^*)/C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}}) = \Theta(n^{\frac{q(q+1)}{p+q}}). \tag{9}$$

This ratio goes to infinity for any possible value of the parameters, implying that strategic agents *always end up incurring an asymptotic degradation of the GLS estimator as* $n \to \infty$. In particular, higher values of q imply a more drastic degradation because the estimation cost is reduced in a neighborhood of 0, which makes agents less willing to exert effort. A high p implies a smaller degradation as agents are less sensitive to their provision costs as long as their precision is smaller than 1.

Figure 2 illustrates the convergence of the estimation cost and the degradation ratio for various values of p and q. Figure 2a pictures the convergence of the estimation cost (in $n^{-q\frac{p-1}{p+q}}$). It illustrates the inconsistency of GLS when provision costs are linear ($p = 1$) and the better convergence rate with larger p and q. In more detail, Figure 2b depicts the degradation of the estimation cost due to the presence of strategic agents. We observe that the relative position of the curves is different than in Figure 2a: the degradation ratio is higher for $(p = 2, q = 3)$ than for $(p = 1, q = 2)$, whereas the first case yields a consistent estimator and the second does not. This illustrates the dual impact of q on the linear regression game: a lower q implies a lower estimation cost but also implies a lower effort, making the estimation cost prohibitively high relative to the non-strategic setting.



(a) Estimation cost $C_{\text{estim}}(\boldsymbol{\lambda}^*)$

(b) Degradation ratio $C_{\text{estim}}(\boldsymbol{\lambda}^*)/C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}})$

Figure 2: Influence of p and q on (a) the estimation cost and (b) the degradation ratio.

Main result: Asymptotic degradation of estimation cost in the general case We are now ready to state the main result of the paper, which characterizes the asymptotic behavior of the estimation cost under non-identical and general provision costs. The next theorem provides upper and lower bounds on how the estimation cost decreases as $n \to \infty$.

Theorem 3. *Assume that Assumptions 1, 2 and 3 hold. Additionally, assume that there exist* $p_{\min}, p_{\max} \geq 1$ *and functions* $c_{\min}, c_{\max} : \mathbb{R}_+ \to \mathbb{R}_+$ *such that for all* $i \in N$ *and all* $a > 1, \ell > 0$*:* $a^{p_{\min}} c_i(\ell) \leq c_i(a\ell) \leq a^{p_{\max}} c_i(\ell)$ *and* $0 < c_{\min}(\ell) \leq c_i(\ell) \leq c_{\max}(\ell) < \infty$*. Then there exist constants* $d, D > 0$ *that do not depend on* n *and such that:*

$$dn^{-q\frac{p_{\min}-1}{p_{\min}+q}-\alpha} \leq C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq Dn^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \quad \textit{where} \quad \alpha = q\frac{(p_{\max} - p_{\min})(q+1)}{p_{\max}(q + p_{\min})}. \tag{10}$$

Sketch of proof. A full proof is given in Appendix C. To get the upper bound, we first obtain an upper bound of the potential ϕ by evaluating it on a well-chosen precision profile $\boldsymbol{\lambda}$ inspired by Proposition 2. Combining this with the assumption that $a^{p_{\min}} c_i(\ell) \leq c_i(a\ell), \forall \ell \in \mathbb{R}_+$ and with the homogeneity of the estimation cost gives the right-hand-side of (10). The lower bound is harder to get. We first exploit the previous upper bound to get an upper bound on the total provision cost (the left part of the potential (5)). Using the assumption that $c_i(a\ell) \leq a^{p_{\max}} c_i(\ell), \forall \ell \in \mathbb{R}_+$, we deduce an upper bound on the total precision. We then consider an optimal design scaled with this total precision and show, using the estimation cost homogeneity, that it gives the left-hand-side of (10). From this sketch of proof, observe that the constant d involves the estimation cost of an optimal design while the constant D involves the estimation cost of a non-strategic precision profile $C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}})$. □

Theorem 3 is our main result: it characterizes the decay of the GLS estimates covariance with strategic data sources for general data provision costs that satisfy a mild assumption governed by the two parameters $p_{\min}, p_{\max}$. This assumption roughly specifies that the provision costs grow faster than $\ell^{p_{\min}}$ and slower than $\ell^{p_{\max}}$; it is satisfied for instance by a sum of monomial terms with exponents between $p_{\min}$ and $p_{\max}$ and such that coefficients do not vanish or explode. Note that the result of Theorem 3 trivially implies that the precision of each agent goes to 0 when the number of agents grow. Although we do not formally prove it, it is clear that the result remains valid even when the assumptions on costs are valid only in a neighborhood near 0. We finally also note here that Theorem 3 remains valid when agents data point are not independent but produced by a joint distribution μ_{joint}. In such a case, the bounds would depend on $E_{\mu_{\text{joint}}}[\frac{1}{n}\sum_i x_i x_i^\top]$ (instead of $E_\mu[xx^\top]$) which captures precisely the impact of correlation on the estimation cost—we provide details on this in Appendix F.

In this degree of generality, it is no longer possible express the equilibrium precisions in closed form (as in Proposition 2). Nevertheless, Theorem 3 shows that we are able to provide precise bounds for the estimation cost. In particular, the upper bound in (10) shows that, as soon as $p_{\min} > 1$ (i.e., data provision costs are superlinear), the estimation cost converges to zero for any scalarization, meaning that the consistency property of GLS is preserved. If $p_{\min} = 1$ though, this is not guaranteed (and even guaranteed to fail if $p_{\min} = p_{\max} = 1$, i.e., for linear costs). Even when convergence to zero is guaranteed ($p_{\min} > 1$), the lower bound in (10) shows that the convergence rate is slower that the standard rate of $\Theta(n^{-q})$ (or $\Theta(1/n)$ for scalarizations with $q = 1$).

We immediately see that for the case $p_{\max} = p_{\min} = p$, the exponent α is equal to 0 and the exponents of the left-hand side and of the right hand-side of (10) coincide and are equal to the exponent of Proposition 2. When $p_{\min}$ and $p_{\max}$ are different, the bounds loosen. Intuitively, the upper bound then involves the parameter $p_{\min}$ because, when precisions are close to zero, the agents with exponent $p_{\min}$ are the ones that have the smallest precision at equilibrium due to larger marginal provision costs. The lower bound, however does not correspond exactly to the $n^{-q\frac{p_{\max}-1}{p_{\max}+q}}$ that one could expect (in fact it decreases faster than $n^{-q\frac{p_{\max}-1}{p_{\max}+q}}$). Whether this is a proof artifact or a consequence of our assumption on the provision costs (which is weak and allows for very diverse costs) remains an open question. We performed a numerical investigation of the result of Theorem 3 illustrating the lower and upper bounds—due to space constraints, the results are deferred to Appendix G.

Remark. We should also note here that our model formally relies on the GLS estimator—which is based on a principle of truthful revelation of data and of its precision to the analyst. This is a natural assumption to make for our envisioned applications where agents are motivated by the model's quality. However, there are other settings where strategic considerations might lead agents to act in a different manner: For instance, if the agents are rewarded as a function of the precision, they might be tempted to untruthfully disclose a higher precision; as another example, agents may be unable to properly quantify the precision of their data points. In such settings, an interesting alternative would be to consider the ordinary least squares (OLS) estimator instead of GLS, as OLS is oblivious to the disclosed precision of the data points. In this case, the conclusion of Theorem 3 would continue to hold; due to space limitations, the detailed statement and proof are relegated to Appendix E. Our analysis for OLS also reveals a potential shortfall of OLS: a single agent with a high provision cost can cause arbitrarily bad estimation cost (whereas GLS is robust to such agents). We discuss this in detail in Appendix E.

Comparison with Theorem 1 Theorem 3 and Theorem 1 both capture notions of efficiency of the game but they are hardly comparable because they characterize radically different types of inefficiencies. Theorem 3 characterizes the ratio of *estimation cost* (the analyst's viewpoint) between the case of strategic agents and a non-strategic scenario where each agent would give a fixed exogenous precision λ_{ns}. In contrast, the PoA of Theorem 1 is a bound of the *total cost* (the population viewpoint) and characterizes the inefficiency due to the self-interested agents by comparing the total cost at equilbrium and at social optimum. These two situations are radically different and the PoA result of Theorem 1 does not hint at the convergence issues addressed in Theorem 3, even in hindsight. For instance, in the case of linear provision costs ($p_{\min} = p_{\max} = 1$), GLS is inconsistent whereas Theorem 1 shows that the price of anarchy always grows sublinearly in n, even in this case where $\texttt{PoA} \leq n^{q/(q+1)}$.

In addition, the proofs of the two theorems are fundamentally different. The proof of Theorem 1 uses a scaling to transform the social optimum $\boldsymbol{\lambda}^{\texttt{opt}}$ into an equilibrium $\boldsymbol{\lambda}^*$. This approach works because $\boldsymbol{\lambda}^{\texttt{opt}}$ and $\boldsymbol{\lambda}^*$ are respectively the minimizers of the functions C_{social} and of the potential ϕ and because these two functions are tightly related. Such an approach cannot be adapted to start from $\boldsymbol{\lambda}_{\text{ns}}$ to obtain an equilibrium $\boldsymbol{\lambda}^*$ as $\boldsymbol{\lambda}_{\text{ns}}$ is not a minimizer. Conversely, the proof of Theorem 3 could be adapted to obtain a result in the spirit of Theorem 1 but would lead to a looser bound.

6 Concluding discussion

In this paper, we show that the precision of GLS estimates for linear regression problems in the presence of strategic data sources is significantly degraded compared to the standard case of non-strategic data sources. We characterize this degradation under mild assumptions and show that basic properties such as consistency no longer always hold with strategic data sources. This points out a necessity to take into account strategic agents in statistical learning. Our work is a stepping stone in this direction.

The objective in our model was to include in the simplest possible way two key elements of learning from strategic data sources: the public good component of the model's precision and the uncertainty about other agents' data. It could easily be extended to a case where agents have a (Bayesian) belief regarding other agents' provision costs as well. At the cost of heavier notation, such an extension would preserve the basic game's structure that leads to the convergence rates of Theorem 3.

References

[1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. Low-cost learning via active data procurement. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC)*, pages 619–636, 2015.

[2] Alexander Craig Aitken. On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1935.

[3] Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*. Oxford University Press New York, 2007.

[4] Christopher Avery and Richard Zeckhauser. Recommender systems for evaluating computer messages. *Commun. ACM*, 40(3):88–89, March 1997.

[5] Omer Ben-Porat and Moshe Tennenholtz. Regression equilibrium. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pages 173–191, 2019.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] Mark Braverman and Sumegha Garg. The role of randomness and noise in strategic classification. In *Proceedings of The Symposium on Foundations of Responsible Computing (FORC)*, 2020.

[8] Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 40.1–40.40, 2015.

[9] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. Truthful univariate estimators. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[10] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pages 27–44, 2018.

[11] Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pages 9–26, 2018.

[12] Yiling Chen, Yiheng Shen, and Shuran Zheng. Truthful data acquisition via peer prediction. In *In Advances in Neural Information Processing Systems (NIPS)*, 2020.

[13] Yiling Chen and Shuran Zheng. Prior-free data acquisition for accurate statistical estimation. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pages 659–677, 2019.

[14] Michela Chessa, Jens Grossklags, and Patrick Loiseau. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *Proceedings of the 28th IEEE Computer Security Foundations Symposium (CSF)*, 2015.

[15] Richard Cornes and Roger Hartley. Fully aggregative games. *Economics Letters*, 116(3):631–633, 2012.

[16] Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pages 319–330, 2013.

[17] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

[18] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pages 55–70, 2018.

[19] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pages 429–438, 2013.

[20] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, August 2014.

[21] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. Collaborative filtering recommender systems. *Found. Trends Hum.-Comput. Interact.*, 4(2):81–173, February 2011.

[22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[23] Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. Elicitation for aggregation. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*, 2015.

[24] Nicolas Gast, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon. Linear regression from strategic data sources. *ACM Trans. Econ. Comput.*, 8(2), May 2020.

[25] Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

[26] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC)*, pages 199–208, 2011.

[27] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[28] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS)*, pages 111–122, 2016.

[29] Maxwell F. Harper, Xin Li, Yan Chen, and Joseph A. Konstan. An economic model of user rating in an online recommender system. In *Proceedings of the 10th International Conference on User Modeling (UM)*, pages 307–316, 2005.

[30] Safwan Hossain and Nisarg Shah. Pure nash equilibria in linear regression. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[31] Stratis Ioannidis and Patrick Loiseau. Linear regression as a non-cooperative game. In *Proceedings of the 9th International Conference on Web and Internet Economics (WINE)*, pages 277–290, 2013.

[32] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, pages 825–844, 2019.

[33] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

[34] Yuqing Kong, Grant Schoenebeck, Biaoshuai Tao, and Fang-Yi Yu. Information elicitation mechanisms for statistical estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 2095–2102, Apr. 2020.

[35] Yang Liu and Yiling Chen. A bandit framework for strategic regression. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pages 1821–1829, 2016.

[36] Yuan Luo, Nihar B. Shah, Jianwei Huang, and Jean Walrand. Parametric prediction from parametric agents. In *Proceedings of the 10th Workshop on the Economics of Networks, Systems and Computation (NetEcon)*, pages 57–57, 2015.

[37] Reshef Meir, Ariel D. Procaccia, and Jeffrey S. Rosenschein. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

[38] John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[39] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 230–239, 2019.

[40] John Morgan. Financing public goods by means of lotteries. *Review of Economic Studies*, 67(4):761–84, October 2000.

[41] Abraham Neyman. Correlated equilibrium and potential games. *International Journal of Game Theory*, 26(2):223–227, June 1997.

[42] Javier Perote and Juan Perote-Pena. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47(2):153–176, 2004.

[43] Friedrich Pukelsheim. *Optimal design of experiments*, volume 50. Society for Industrial Mathematics, 2006.

[44] Nihar B. Shah and Dengyong Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *The Journal of Machine Learning Research*, 17(1):5725–5776, 2016.

[45] Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[46] Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, counterfactual explanations and strategic behavior. In *Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[47] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[48] Tyler Westenbroek, Roy Dong, Lillian J. Ratliff, and S. Shankar Sastry. Competitive statistical estimation with strategic data sources. *IEEE Transactions on Automatic Control*, 65(4):1537–1551, 2020.

[49] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[50] Hanrui Zhang, Yu Cheng, and Vincent Conitzer. When samples are strategically selected. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7345–7353, 2019.

A Notation table

To ease the reading, Table 1 summarizes the main notation introduced in the paper and used throughout the paper and the supplementary material.

Table 1: Summary of the notation

Symbol	Meaning
$\mathcal{X}$	Finite set of possible attribute vectors x.
$\mu(\cdot)$	Probability distribution on attributes vectors x.
n	Number of agents.
$\lambda_i(x)$	Precision allocated to vector x by player i.
$\boldsymbol{\lambda}_{-i}(x)$	Vector of precisions allocated to vector x by every player except i.
$c_i(\ell)$	Data provision cost of agent i for a precision ℓ of the data provided.
$F(M)$	Scalarization mapping a (covariance) matrix M to a cost.
$C_{\text{estim}}(\boldsymbol{\lambda})$	Estimation cost $= F\left(\left(\mathbb{E}\left[\sum_{i\in N}\lambda_i(x_i)x_ix_i^T\right]\right)^{-1}\right)$.
$J_i(\lambda_i, \boldsymbol{\lambda}_{-i})$	Payoff of agent i considering the strategy profile $\boldsymbol{\lambda} = (\lambda_i, \boldsymbol{\lambda}_{-i})$.
$\phi(\boldsymbol{\lambda})$	Potential function of the linear regression game.
$\nu_{\boldsymbol{\lambda}^*}(x)$	Measure mapping a vector to its probability × the sum of precisions attributed by agents.
$p, p_{\min}, p_{\max}$	Homogeneity degrees of provision costs.
q	Homogeneity degree of a scalarization.
ν^*	Optimal design.

B Scalarizations

In this section, we detail some examples of usual matrix scalarizations mentioned briefly in the paper that fit Assumption 2 and are standard in optimal design. For further information on the subject, see [3] and the references therein.

B.1 Trace

The trace trivially satisfies Assumption 2 with $q = 1$. It is used in optimal design to minimize the average variance of the estimates of the regression coefficients and is known as the A-optimal design criterion.

B.2 Squared Frobenius norm

The squared Frobenius norm is defined on the set of matrices $V = [v_{ij}]$ of dimensions $d \times d$ as:

$$\begin{aligned} ||V||_F^2 &= \sum_{i=1}^{d}\sum_{j=1}^{d} v_{ij}^2 \\ &= \text{trace}(VV^T). \end{aligned}$$

It is easy to check that this scalarization satisfies Assumption 2 with $q = 2$.

B.3 Mean squared error

We define the mean squared error of *an estimator* $\hat{\boldsymbol{\beta}}$ estimating a linear model $\boldsymbol{\beta}$ as:

$$\texttt{MSE}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right]. \tag{B.1}$$

This mean squared error is simply the estimator's covariance matrix. It is a property of the estimator and it is a classical proxy to assess its quality.[4] In particular, in the linear regression setting, it does not depend on the realization of the values $\tilde{y}_i$ but only on the independent variables x_i and on the precisions of the response variables $\tilde{y}_i$ (unlike the empirical mean squared error).

[4] See Michel F. Dekking, Cornelis Kraaikamp, Paul H. Lopuhaä, and Ludolf Meester. A Modern Introduction to Probability and Statistics: Understanding why and how. Springer Science & Business Media (2005).

A similar definition can also be applied to the predicted value for a given data point x. In this case it is referred to as the mean squared error of the predictor:

$$\texttt{MSE}(\hat{\boldsymbol{\beta}}^T x) = \mathbb{E}\left[(\hat{\boldsymbol{\beta}}^T x - \boldsymbol{\beta}^T x)^2\right].$$

This quantity gives an indication on the average amount of error the estimator makes when predicting the value of the model on a given data point x. It is used in optimal design to define scalarizations by considering the average mean squared error made by the estimator on specific data points. To properly define these criteria, we first write this quantity in a more convenient form.

The mean squared error of the predictor of the linear model on a parameter x is:

$$\texttt{MSE}(\hat{\boldsymbol{\beta}}^T x) = \texttt{Var}(\hat{\boldsymbol{\beta}}^T x) + \texttt{Bias}(\hat{\boldsymbol{\beta}}^T x, \boldsymbol{\beta}^T x).$$

As $\hat{\boldsymbol{\beta}}$ is unbiased, we can rewrite the mean-squared error depending only on the variance. Let V be the covariance matrix of a linear unbiased estimator $\hat{\boldsymbol{\beta}}$. We then have:

$$\begin{aligned}\texttt{MSE}(\hat{\boldsymbol{\beta}}^T x) &= \texttt{Var}(\hat{\boldsymbol{\beta}}^T x)\\ &= xVx^T.\end{aligned}$$

We now define the two main design criteria (or scalarizations) that are based on this mean squared prediction error:

a. **The average mean squared error.** Given a set $\mathcal{V}$ and a probability distribution ρ on $\mathcal{V}$, we define the average mean-square error scalarization as:

$$F : V \to \int_{\mathcal{V}} xVx^T \rho(dx).$$

This scalarization is trivially convex, increasing in the positive semi-definite order and homogeneous of degree $q = 1$. It is known in the optimal design litterature as the I (integrated) optimal design criterion and is used to minimize the average prediction error. In our setting this scalarization can be directly applied with $\mathcal{V} = \mathcal{X}$ and $\rho = \mu$.

b. **The mean squared error over a set of specific points.** Given a finite set $\{x_1, \dots, x_m\}$ of possible attribute vectors, we define the mean-squared error on that specific set of points as:

$$F : V \to \sum_{i=1}^{m} x_i V x_i^T.$$

This scalarization is similar to the previous one and has the same properties but is used to minimize the prediction error only on a specific set of points of interest. It is known in the optimal design litterature as the V optimal design criteria.

C Proofs

C.1 Proof of Proposition 1

Recall that a strategy λ is a function from the finite set $\mathcal{X}$ to $\mathbb{R}_+$. Hence, a strategy λ is an element of the finite dimensional space $\mathbb{R}^{\mathcal{X}}$ and a precision profile $\boldsymbol{\lambda}$ is essentially a vector (of dimension $n|\mathcal{X}|$).

Step 1: The potential function is convex. By Assumption 3, the data provision costs are convex. Additionally, $C_{\text{estim}}(\boldsymbol{\lambda})$ is a composition of the function $\boldsymbol{\lambda} \to \mathbb{E}\left[\sum_i \lambda_i(x_i) x_i x_i^T\right]$, the matrix inverse function, and the scalarization. The matrix inverse function is a convex function. As the scalarization F is non-decreasing and convex (by Assumption 2), $C_{\text{estim}}(\boldsymbol{\lambda})$ is convex in $\boldsymbol{\lambda}$. This shows that the potential is convex; hence a strategy profile is a Nash equilibrium if and only if it is a minimum of the potential.

Step 2: The potential admits a minimum. Let $\phi(\mathbf{1}) = \sum_i \mathbb{E}\left[c_i(1)\right] + F((\mathbb{E}\left[\sum_i x_i x_i^T\right])^{-1})$. By Assumption 3, $\lim_{\ell \to +\infty} c_i(\ell) = +\infty$. Recall that μ has full support on $\mathcal{X}$ (Assumption 1). Then, for all $x \in \mathcal{X}$, $\lim_{\ell \to +\infty} c_i(\ell)\mu(x) = +\infty$. Hence, there exists $\ell_{\max}$ such that for all i and all x, $c_i(\ell_{\max})\mu(x) > \phi(\mathbf{1})$. This shows that if $\boldsymbol{\lambda}$ is a precision profile such that $\lambda_i(x) > \ell_{\max}$ for some i and x, then $\phi(\boldsymbol{\lambda}) \geq \phi(1)$.

As $\mathcal{X}$ is finite, the set of precision profile such that for all i, $\lambda_i : \mathcal{X} \to [0, \ell_{\max}]$ is a compact set. As ϕ is convex, it admits a minimum on this set. By definition of $\ell_{\max}$, this minimum is a global minimum. This concludes the proof that there exists an equilibrium. If in addition all data provision costs are strictly convex, then the potential is strictly convex; hence this minimum is unique and there exists a unique equilibrium.

Step 3: If different equilibria exist, they have the same estimation cost. As shown before, an equilibrium is a minimum of the potential function ϕ defined for all precision profiles $\boldsymbol{\lambda}$ as

$$\phi(\boldsymbol{\lambda}) = \sum_i \mathbb{E}\left[c_i(\lambda_i(x))\right] + C_{\mathrm{estim}}(\boldsymbol{\lambda}).$$

In the above equation, $C_{\mathrm{estim}}()$ is not necessarily strictly convex. Recall indeed that $C_{\mathrm{estim}}(\boldsymbol{\lambda})$ is defined as

$$C_{\mathrm{estim}}(\boldsymbol{\lambda}) = F\left(\left(\mathbb{E}\left[\sum_i \lambda_i(x_i)x_ix_i^T\right]\right)^{-1}\right).$$

If there exist $\boldsymbol{\lambda} \neq \boldsymbol{\lambda}'$ (which is the case for any linear regression game with $n \geq 2$ players) such that $\mathbb{E}\left[\sum_i \lambda_i(x_i)x_ix_i^T\right] = \mathbb{E}\left[\sum_i \lambda_i'(x_i)x_ix_i^T\right]$, then $C_{\mathrm{estim}}(\boldsymbol{\lambda}) = C_{\mathrm{estim}}(\boldsymbol{\lambda}') = C_{\mathrm{estim}}((\boldsymbol{\lambda}+\boldsymbol{\lambda}')/2)$ and $C_{\mathrm{estim}}()$ is not strictly convex.

Yet, we show below that $C_{\mathrm{estim}}(\cdot)$ is strictly convex when viewed as a function of $M(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_i \lambda_i(x_i)x_ix_i^T\right]$. Indeed F is an increasing convex function (by Assumption 2, see erratum in Appendix A) and $M \mapsto M^{-1}$ is a strictly convex function, the function $M \mapsto F(M^{-1})$ is a strictly convex function.

Assume that there exist two equilibria $\boldsymbol{\lambda}^*$ and $\tilde{\boldsymbol{\lambda}}^*$ and assume by contradiction that $\mathbb{E}\left[\sum_i \lambda_i^*(x_i)x_ix_i^T\right] \neq \mathbb{E}\left[\sum_i \tilde{\lambda}_i^*(x_i)x_ix_i^T\right]$. Let $\boldsymbol{\lambda}' = (\boldsymbol{\lambda}^* + \tilde{\boldsymbol{\lambda}}^*)/2$. The strict convexity of $M \mapsto F(M^{-1})$ implies that $C_{\mathrm{estim}}(\boldsymbol{\lambda}') < (C_{\mathrm{estim}}(\boldsymbol{\lambda}^*) + C_{\mathrm{estim}}(\tilde{\boldsymbol{\lambda}}^*))/2$. This implies that $\phi(\boldsymbol{\lambda}') < (\phi(\boldsymbol{\lambda}^*) + \phi(\tilde{\boldsymbol{\lambda}}^*))/2$, which contradicts the fact that $\boldsymbol{\lambda}^*$ and $\tilde{\boldsymbol{\lambda}}^*$ are minima of the potential function ϕ. Thus, if two different equilibria exist, they have the same information matrix and yield the same estimation cost.

C.2 Proof of Theorem 1

This proof relies on adapting the proof of [24] to our setting. For completeness, we redo this proof using the notations of our model.

Upper Bound. To simplify the notation, in this proof, we write p instead of $p_{\min}$; hence we show that $\texttt{PoA} \leq n^{\frac{q}{p+q}}$. Suppose that $\texttt{PoA} > n^{\frac{q}{p+q}}$. This implies that there exists an equilibrium $\boldsymbol{\lambda}^*$ such that

$$\begin{aligned}
C_{\mathrm{social}}(\boldsymbol{\lambda}^*) &\geq \sum_{i\in N} \mathbb{E}\left[c_i(\lambda_i^*(x_i))\right] + nC_{\mathrm{estim}}(\boldsymbol{\lambda}^*) \\
&> n^{\frac{q}{q+p}}(\sum_{i\in N} \mathbb{E}\left[c_i(\lambda_i^{\texttt{opt}}(x_i)\right] + nC_{\mathrm{estim}}(\boldsymbol{\lambda}^{\texttt{opt}})) \\
&= n^{\frac{q}{q+p}} C_{\mathrm{social}}(\boldsymbol{\lambda}^{\texttt{opt}}).
\end{aligned}$$

We will show that this implies that $\boldsymbol{\lambda}^*$ is not an equilibrium, which is a contradiction.

By using that $c_i(\lambda_i^*) \geq 0$ and dividing the above inequality by n, we obtain:

$$\begin{aligned}
\Phi(\boldsymbol{\lambda}^*) &= \sum_{i\in N} \mathbb{E}\left[c_i(\lambda_i^*(x_i))\right] + C_{\mathrm{estim}}(\boldsymbol{\lambda}^*) \\
&\geq \frac{1}{n}\left(\sum_{i\in N} \mathbb{E}\left[c_i(\lambda_i^*(x_i))\right] + nC_{\mathrm{estim}}(\boldsymbol{\lambda}^*)\right) && (= \frac{1}{n}C_{\mathrm{social}}(\boldsymbol{\lambda}^*)) \\
&> n^{-\frac{p}{q+p}} \sum_{i\in N} \mathbb{E}\left[c_i(\lambda_i^{\texttt{opt}}(x_i))\right] + n^{\frac{q}{p+q}} C_{\mathrm{estim}}(\boldsymbol{\lambda}^{\texttt{opt}}) && (= \frac{1}{n}n^{\frac{q}{q+p}} C_{\mathrm{social}}(\boldsymbol{\lambda}^{\texttt{opt}})) \\
&\geq \sum_{i\in N} \mathbb{E}\left[c_i\left(\frac{\lambda_i^{\texttt{opt}}(x_i)}{n^{\frac{1}{p+q}}}\right)\right] + C_{\mathrm{estim}}(\frac{\boldsymbol{\lambda}^{\texttt{opt}}}{n^{\frac{1}{p+q}}}) && (= \Phi(\boldsymbol{\lambda}^{\texttt{opt}}/(n^{1/(p+q)}))),
\end{aligned}$$

where we used the homogeneity assumptions for the last inequality.

To conclude the proof, we remark that $\frac{\boldsymbol{\lambda}^{\text{opt}}}{n^{1/(p+q)}}$ is a valid strategy profile. This would imply that $\boldsymbol{\lambda}^*$ is not a minimum of the potential function, which is a contradiction. Thus, we have $\texttt{PoA} \leq n^{\frac{q}{p+q}}$.

Lower Bound. Let $p, q \geq 1$. Consider the linear regression game where $\mathcal{X} = \{1\}$, $\mu(1) = 1$, $c_i(\ell) = \ell^p$ and $F(V) = \text{trace}(V)^q = V^q$. As μ is a deterministic measure, this game is also a valid game in the setting of [24]. It is straightforward to see that our game has a unique Nash equilibrium $\boldsymbol{\lambda}^*$ that corresponds to the unique non-trivial Nash equilibrium of the corresponding game of [24]. Hence, the price of anarchy of our game coincides with the price of stability of the corresponding game of [24]. Hence, the computation of [24] show that, for all ϵ, there exist n such that this game has a price of anarchy larger than $n^{q/p+q}(1-\epsilon)$.

C.3 Proof of Theorem 2

Recall that the provision cost of a player i is $c_i(\ell) = a_i\ell$ and assume without loss of generality that $a_1 \leq a_2 \leq \cdots \leq a_n$.

Let $\boldsymbol{\lambda}^*$ be an equilibrium of the game and let ν^* be an optimal design. Recall that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_{i\in N} \lambda_i^*(x)\mu(x)$ for all $x \in \mathcal{X}$. Let $b = \sum_{x\in\mathcal{X}} \nu_{\boldsymbol{\lambda}^*}(x)$. Let λ_{ν^*} be the strategy such that $\lambda_{\nu^*}(x) = b\nu^*(x)/\mu(x)$ for all x and consider the precision profile $\boldsymbol{\lambda}_{\nu^*} = (\lambda_{\nu^*}, 0, \cdots, 0)$. We have:

$$\begin{aligned}
\phi(\boldsymbol{\lambda}^*) &= F((\sum_x xx^T\nu_{\boldsymbol{\lambda}^*}(x))^{-1}) + \sum_i a_i \sum_x \lambda_i^*(x)\mu(x) \\
&\geq F((\sum_x xx^T\nu_{\boldsymbol{\lambda}^*}(x))^{-1}) + a_1 b && \text{(C.1)} \\
&= b^{-q}F((\sum_x xx^T\nu_{\boldsymbol{\lambda}^*}(x)/b)^{-1}) + a_1 b && \text{(C.2)} \\
&\geq b^{-q}F((\sum_x xx^T\nu^*(x))^{-1}) + a_1 b && \text{(C.3)} \\
&= F((\sum_x xx^T\lambda_{\nu^*}(x)\mu(x))^{-1}) + a_1\sum_x \lambda_{\nu^*}(x)\mu(x) && \text{(C.4)} \\
&= \phi(\boldsymbol{\lambda}_{\nu^*}),
\end{aligned}$$

where the first inequality (C.1) is because $a_1 \leq a_i$ for all i, and the second inequality (C.3) is because ν^* is an optimal design. The equalities (C.2) and (C.4) are due to the homogeneity of F (Assumption 2 implies that $F((bM)^{-1}) = b^{-q}F(M^{-1})$), and in (C.4) we also use that by definition of λ_{ν^*} and since $\sum_x \nu^* = 1$ we have $\sum_x \lambda_{\nu^*}(x)\mu(x) = b$.

If $\nu_{\boldsymbol{\lambda}^*}/b$ was not an optimal design, the inequality (C.3) would be strict which would imply that $\phi(\boldsymbol{\lambda}^*) > \phi(\boldsymbol{\lambda}_{\nu^*})$ which would contradict the fact that $\boldsymbol{\lambda}^*$ is a minimum of the potential. This implies that (C.3) is an equality which means that $\nu_{\boldsymbol{\lambda}^*}(x)/b$ is an optimal design.

C.4 Proof of Proposition 2

An equilibrium is a minimum of the potential function ϕ. When all costs are identical, this function is symmetric. As ϕ is a convex function, this implies that there exists a minimum of ϕ that is symmetric. A symmetric precision profile $\boldsymbol{\lambda} = (\lambda, \ldots \lambda)$ is a Nash equilibrium if and only if it minimizes the potential ϕ. By symmetry, this potential can be rewritten as:

$$\phi(\lambda, \ldots, \lambda) = n\mathbb{E}\left[\lambda(x)^p\right] + C_{\text{estim}}(n\lambda)$$

Let us define the function $f : \mathbb{R}_+^{\mathcal{X}} \to \mathbb{R}_+$ that associates to a strategy λ, the quantity $f(\lambda) = \mathbb{E}\left[\lambda(x)^p\right] + C_{\text{estim}}(\lambda)$. Recall that λ_{single} is the minimum of f. For a given strategy λ, we have:

$$\begin{aligned}
\phi(n^{-\frac{q+1}{p+q}}\lambda, \ldots, n^{-\frac{q+1}{p+q}}\lambda) &= n\mathbb{E}\left[\lambda(x)^p n^{-\frac{q+1}{p+q}p}\right] + C_{\text{estim}}(nn^{-\frac{q+1}{p+q}}\lambda) \\
&= n^{q\frac{1-p}{p+q}}\mathbb{E}\left[\lambda(x)^p\right] + n^{-q\frac{p-1}{p+q}}C_{\text{estim}}(\lambda) \\
&= n^{-q\frac{p-1}{p+q}}C_{\text{estim}}(\lambda),
\end{aligned}$$

where we used the homogeneity of F, which implies that $C_{\text{estim}}(a\lambda) = a^{-q}C_{\text{estim}}(\lambda)$.

For any $n \in \{1, 2, \dots\}$, the function $\lambda \mapsto n^{-\frac{q+1}{p+q}}\lambda$ is a bijection from $\mathbb{R}_+^{\mathcal{X}}$ to $\mathbb{R}_+^{\mathcal{X}}$. Hence, λ is a minimum of f if and only if $(n^{-\frac{q+1}{p+q}}\lambda, \dots, n^{-\frac{q+1}{p+q}}\lambda)$ is a minimum of ϕ. Thus, the precision profile $\boldsymbol{\lambda}^*$ such that $\forall i: \lambda_i^* = n^{-\frac{1+q}{p+q}}\lambda_{\text{single}}$ is an equilibrium.

The second part of the proposition follows immediately from the homogeneity of F, which implies that for this equilibrium, $C_{\text{estim}}(\boldsymbol{\lambda}^*) = n^{-q\frac{p-1}{p+q}} C_{\text{estim}}(\lambda_{\text{single}})$. Moreover, all equilibria have the same estimation cost by Proposition 1.

C.5 Proof of Theorem 3

C.5.1 Upper bound

In this first step, we compute the value of the potential function for a particular constant strategy in which all players use the precision $\lambda(x) = n^{-\frac{q+1}{p_{\min}+q}}$ for all values of $x \in \mathcal{X}$. By abuse of notation, we denote this precision profile by $(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}})$. The value of the potential for this precision profile is

$$\begin{aligned}
\phi(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}}) &= \sum_{i=1}^{n} \mathbb{E}\left[c_i(n^{-\frac{q+1}{p_{\min}+q}})\right] + F((\sum_{i=1}^{n} \mathbb{E}\left[xx^T n^{-\frac{q+1}{p_{\min}+q}}\right])^{-1}) \\
&= \sum_{i=1}^{n} c_i(n^{-\frac{q+1}{p_{\min}+q}}) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}} \mathbb{E}\left[xx^T\right])^{-1}) \\
&\leq \sum_{i=1}^{n} n^{-p_{\min}\frac{q+1}{p_{\min}+q}} c_i(1) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}} \mathbb{E}\left[xx^T\right])^{-1}) \qquad \text{(C.5)} \\
&= n^{-p_{\min}\frac{q+1}{p_{\min}+q}} \sum_{i=1}^{n} c_i(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}\left[xx^T\right])^{-1}) \qquad \text{(C.6)} \\
&\leq n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}} c_{\max}(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}\left[xx^T\right])^{-1}) \qquad \text{(C.7)} \\
&= n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}} \left(c_{\max}(1) + F((\mathbb{E}\left[xx^T\right])^{-1})\right), \qquad \text{(C.8)}
\end{aligned}$$

where we use that $c_i(1) \geq a^{p_{\min}} c_i(1/a)$ with $a = n^{\frac{q+1}{p_{\min}+q}}$ (from the theorem's assumption) in (C.5), the homogeneity of F (Assumption 2) in (C.6), and the theorem's assumption, which implies that $c_i(1) \leq c_{\max}(1)$ for all i, in (C.7).

As $c_i(\ell) \geq 0$ and $\boldsymbol{\lambda}^*$ is a minimum of the potential, it holds that

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq \phi(\boldsymbol{\lambda}^*) \leq \phi(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}}).$$

Hence, the right-hand-side of (10) holds with $D = \left(c_{\max}(1) + F((\mathbb{E}\left[xx^T\right])^{-1})\right)$.

C.5.2 Lower bound

By (C.8), $\phi(\boldsymbol{\lambda}^*) \leq n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}}(c_{\max}(1) + F((\mathbb{E}\left[xx^T\right])^{-1}))$. Recall that all c_i are increasing convex and $\inf_i c_i(1) \geq c_{\min}(1) > 0$. This implies that $\lim_{\ell \to \infty} \inf_i c_i(\ell) = \infty$ as $\inf_i c_i(\ell) > \ell^{p_{\min}} c_{\min}(1)$. As $\boldsymbol{\lambda}^*$ is a minimum of the potential, this implies that there exists a value $\ell_{\max}$ independent of n such that $\lambda_i^*(x) \leq \ell_{\max}$.

We first obtain a bound on the total amount of precision given by all players. To do that we use Jensen's inequality for concave function in (C.9). Then we use that $c_i(\ell_{\max}) \leq (\ell_{\max}/\lambda_i(x))^{p_{\max}} c_i(\lambda_i(x))$ as $\ell_{\max}/\lambda_i(x) > 1$ to obtain (C.10) and $c_i(\ell_{\max}) \geq c_{\min}(\ell_{\max})$ to obtain (C.11):

$$\begin{aligned}
\left(\sum_{i=1}^{n} \frac{1}{n} \mathbb{E}\left[c_i(\lambda_i(x))\right]\right)^{\frac{1}{p_{\max}}} &\geq \sum_{i=1}^{n} \frac{1}{n} \mathbb{E}\left[(c_i(\lambda_i(x)))^{\frac{1}{p_{\max}}}\right] \qquad \text{(C.9)} \\
&\geq \sum_{i=1}^{n} \frac{1}{n} \mathbb{E}\left[((\lambda_i(x)/\ell_{\max})^{p_{\max}} c_i(\ell_{\max}))^{\frac{1}{p_{\max}}}\right] \qquad \text{(C.10)} \\
&\geq \frac{(c_{\min}(\ell_{\max}))^{\frac{1}{p_{\max}}}}{\ell_{\max}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\lambda_i(x)\right]. \qquad \text{(C.11)}
\end{aligned}$$

This shows that

$$\begin{aligned}\sum_{i=1}^{n}\mathbb{E}\left[\lambda_i^*(x)\right] &\leq \frac{n\ell_{\text{max}}}{(c_{\min}(\ell_{\text{max}}))^{1/p_{\max}}}\left(\sum_{i=1}^{n}\frac{1}{n}\mathbb{E}\left[c_i(\lambda_i^*(x))\right]\right)^{\frac{1}{p_{\max}}}\\ &\leq \frac{n\ell_{\text{max}}}{(c_{\min}(\ell_{\text{max}}))^{1/p_{\max}}}\left(\frac{1}{n}\phi(\boldsymbol{\lambda}^*)\right)^{1/p_{\max}}\\ &\leq \frac{n\ell_{\text{max}}}{(c_{\min}(\ell_{\text{max}}))^{1/p_{\max}}}\left(\frac{1}{n}n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}}\left(c_{\max}(\ell_{\text{max}})+F((\mathbb{E}\left[xx^T\right])^{-1})\right)\right)^{1/p_{\max}}, \qquad \text{(C.12)}\end{aligned}$$

where we used (C.11) for the first inequality, the fact that $C_{\text{estim}}(\boldsymbol{\lambda}) \geq 0$ for the second and (C.8) to obtain the last inequality.

Note that the exponent of n in (C.12) is

$$\begin{aligned}1-1/p_{\max}-\frac{q(p_{\min}-1)}{p_{\max}(p_{\min}+q)} &= \frac{p_{\max}(p_{\min}+q)-(p_{\min}+q)-q(p_{\min}-1)}{p_{\max}(p_{\min}+q)}\\ &= \frac{p_{\max}(p_{\min}+q)-p_{\min}(1+q)}{p_{\max}(p_{\min}+q)}\\ &= \frac{p_{\max}(p_{\min}-1)+(p_{\max}-p_{\min})(1+q)}{p_{\max}(p_{\min}+q)}\\ &= \frac{p_{\min}-1}{p_{\min}+q}+\alpha/q,\end{aligned}$$

where $\alpha = q\frac{(p_{\max}-p_{\min})(q+1)}{p_{\max}(q+p_{\min})}$ is the same α as in Theorem 3.

Plugging this into (C.12) yields the upper bound on the total amount of precision given by all players:

$$\sum_{i=1}^{n}\mathbb{E}\left[\lambda_i^*(x)\right] \leq \ell_{\text{max}}\left(1+\frac{F((\mathbb{E}\left[xx^T\right])^{-1})}{c_{\min}(\ell_{\text{max}})}\right)^{\frac{1}{p_{\max}}} n^{\frac{p_{\min}-1}{p_{\min}+q}+\alpha/q}. \qquad \text{(C.13)}$$

Recall that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_i \lambda_i(x)\mu(x)$. Following what we did in (C.3) with the notation $b = \sum_{x\in\mathcal{X}}\nu_{\boldsymbol{\lambda}^*}(x) = \mathbb{E}\left[\sum_i \lambda_i^*(x)\right]$, we have

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \geq \left(\mathbb{E}\left[\sum_i \lambda_i^*(x)\right]\right)^{-q} F\left(\left(\sum_{x\in\mathcal{X}} xx^T\nu^*(x)\right)^{-1}\right). \qquad \text{(C.14)}$$

Combining (C.14) and (C.13) shows that the right-hand-side of (10) holds with

$$d = F\left(\left(\sum_{x\in\mathcal{X}} xx^T\nu^*(x)\right)^{-1}\right)\ell_{\text{max}}\left(1+\frac{F((\mathbb{E}\left[xx^T\right])^{-1})}{c_{\min}(\ell_{\text{max}})}\right)^{-\frac{q}{p_{\max}}}.$$

D Equivalence

In this section, we show that our model is equivalent to the complete information model defined in [24], when the number of player goes to infinity. We consider a model with n agents in which the feature of agent i are chosen *i.i.d.*. The only difference between the two models is that:

- In our model, an agent i does not know the exact feature x_{-i} of the other individual but only knows the distribution μ from which they are drawn. As a result, an player i seeks to minimize

  $$J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = \mathbb{E}\left[c_i(\lambda_i(x_i))\right] + F\left(\left(\mathbb{E}\left[\sum_{i\in N}\lambda_i(x_i)x_ix_i^\top\right]\right)^{-1}\right),$$

  where $\lambda_i : \mathcal{X} \to \mathbb{R}^+$ is a function that associates to each possible feature $x \in \mathcal{X}$ a precision $\lambda_i(x)$.

- In the model of [24], a player i knows the exact features of other players. As a result, its cost function is

$$J_i^{\text{ci}}(\ell_i, \ell_{-i}, X) = c_i(\ell_i) + F((\sum_{i=1}^{n} \ell_i x_i x_i^T)^{-1}). \tag{D.1}$$

  In the above definition, we emphasize that the cost of a player depends on $X = (x_i)_{i \in N}$ which is the matrix of features of all players. In particular, the equilibrium of the complete information game is well defined only if $\sum_i x_i x_i^T \succ 0$. This assumption simply states that the data points held by the agents span $\mathbb{R}^d$ so that the corresponding linear regression is well defined. We refer to [24] for technical results regarding the existence of a Nash equilibrium. The complete information game is a potential game with the potential:

$$\phi^{\text{ci}}(\boldsymbol{\ell}, X) = \sum_{i=1}^{n} c_i(\ell_i) + F((\sum_{i=1}^{n} \ell_i x_i x_i^T)^{-1}), \tag{D.2}$$

  and the Nash equilibrium of the game is the minimum of the potential function.

In this section, we show that when $n \to +\infty$, the equilibrium of the complete information game, that we denote by $\boldsymbol{\ell}^{\text{ci}*}$, and the equilibrium of our linear regression $\boldsymbol{\lambda}^*$ are equivalent and can be exchanged.

Notations and assumptions: We assume the same assumptions as Theorem 3. In addition, we assume that there is a finite number T of provision cost functions and we denote by n_t the number of agents having provision cost c_t for $t \in \mathcal{T} := \{1 \dots T\}$.

D.1 Comparison of equilibrium

To formally compare the equilibrium of the complete information game to the equilibrium of our linear regression game, we will use need the following lemma. This lemma states that there always exists a symmetric equilibrium of the games considered. Note that if provision costs are strictly convex, the equilibrium is unique. If provision costs are linear, there might, however, exist an infinite number of equilibrium.

Lemma D.1. *There exists an equilibrium of the complete information game $\ell^{\text{ci}*}$ such that:*

$$\forall i, i', x_i = x_{i'} \text{ and } c_i = c_{i'} \Rightarrow \ell_i^{\text{ci}*} = \ell_{i'}^{\text{ci}*} \tag{D.3}$$

There exists an equilibrium of the linear regression game $\boldsymbol{\lambda}^*$ such that:

$$\forall i, i', c_i = c_{i'} \Rightarrow \forall x \in \mathcal{X}, \lambda_i^*(x) = \lambda_{i'}^*(x) \tag{D.4}$$

Proof. Consider an equilibrium $\boldsymbol{\ell}^{\text{ci}*}$ of the complete information game. We define the following strategy profile:

$$\forall i \in N, \ell_i = \sum_{i'=1}^{n} \mathbb{1}_{c_i = c_t \text{ and } x_{i'} = x_i} \frac{\ell_{i'}^{\text{ci}*}}{n_t^{x_i}},$$

where $n_t^{x_i}$ is the number of players with features x_i and cost type t.

This strategy profile is simply that each agent provides data with the precision being the average of the precision of similar agents in the equilibrium. It achieves the same estimation cost as the equilibrium and with our convexity assumptions achieves a lower total provision cost. This is thus a minimum of the potential and an equilibrium.

The proof for the linear regression game follows the same steps. □

As there is a symmetric equilbrium, this implies that instead of considering strategy profiles, we may restrict our attention to functions $\lambda_t(x)$ that associate a type of cost and a data point to a precision. This is true for the Bayesian game, in which $\lambda_i(x)$ is replaced by $\lambda_t(x)$ when $c_i = c_t$. This is also true for the complete information game, when ℓ_i is replaced by $\lambda_t(x_i)$ when $c_i = c_t$. We work with these functions for the rest of the section and by abuse of notation we redefine the potential of the games as follows:

$$\phi^{\text{ci}}(\boldsymbol{\lambda}, X) = \sum_{x \in \mathcal{X}} \sum_{t=1}^{T} c_t(\lambda_t(x)) n_t^x + F((\sum_{x \in \mathcal{X}} x x^T \sum_{t=1}^{T} \lambda_t(x) n_t^x)^{-1}) \tag{D.5}$$

$$\phi(\boldsymbol{\lambda}) = \sum_{x \in \mathcal{X}} \sum_{t=1}^{T} c_t(\lambda_t(x)) n_t \mu(x) + F((\sum_{x \in \mathcal{X}} x x^T \sum_{t=1}^{T} \lambda_t(x) n_t \mu(x))^{-1}), \tag{D.6}$$

where as before, n_t is the number of players having cost function c_t and n_t^x is the number of player having cost function c_t and features x in the complete information game.

By abuse of notation, we write $\boldsymbol{\lambda}^* = (\lambda_t^*)_{t\in\mathcal{T}}$ the equilibrium of our linear regression game and by $\boldsymbol{\lambda}^{\text{ci}*} = (\lambda_t^{\text{ci}*})_{t\in\mathcal{T}}$ the equilibrium of the complete information game. They are the minimum of (respectively) the potential functions (D.5) and (D.6).

D.2 Main equivalence result

The intuition behind the theorem is as follows. Equation (D.7) states that the minimum of the potentials are equivalent with high probability. Thus, computing the equilibrium of our linear regression game gives a general result on how large complete information games behave. Equations (D.8) and (D.9) state that the equilibrium are essentially equivalent. This means that agents can safely compute the equilibrium of the linear regression game without needing to acquire the information of all other agents. We remark that (D.7) applied with $p_{\max} = 1$ yields $\phi(\lambda^*) = \phi^{\text{ci}}(\lambda^{\text{ci}*})$. Finally, we emphasize that the complexity of Theorem 4 comes from the necessity to prove *equivalence* of potential to show that our results are also valid for the complete information game. Indeed, it is easy to show that both potential go to 0 as long as $p_{\min} > 1$. Thus, any result simply stating that the potential of the complete information game converges to the potential of our model is meaningless. With our result, however, it is easy to show that Theorem 3 is valid in the complete information setting with high probability.

Theorem 4. *Let $\boldsymbol{\lambda}^*$ be an equilibrium of the linear regression game and $\boldsymbol{\lambda}^{\text{ci}*}$ be an equilibrium of the complete information game. For all $0 < \epsilon < 1/2$, we have with probability at least $1 - |X| \sum_t 2\exp(-2n_t^{2\epsilon})$:*

$$\frac{1}{\max_{x,t}\left(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)}\right)^{p_{\max}-1}}\phi(\lambda^*) \le \phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X) \le \max_{x,t}\left(\frac{\mu(x)}{\mu(x) - n_t^{\epsilon-1/2}}\right)^{p_{\max}-1}\phi(\lambda^*), \tag{D.7}$$

$$\phi^{\text{ci}}(\boldsymbol{\lambda}^*, X) \le D_n \max_{x,t}(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)})^{p_{\max}-1}\phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X), \tag{D.8}$$

and

$$\phi(\boldsymbol{\lambda}^{\text{ci}*}) \le D'_n \max_{x,t}(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*); \tag{D.9}$$

where

$$D_n = \max(\max_{x,t}(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)n_t}), \frac{1}{(\min_{x,t}(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}))^q}) \quad \textit{and}$$

$$D'_n = \max(\max_{x,t}(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}), \frac{1}{(\min_{x,t}(\frac{\mu(x)}{\mu(x)+n_t^{\epsilon-1/2}}))^q}).$$

Proof. The equilibrium are defined as $\boldsymbol{\lambda}^{\text{ci}*} \in \arg\min(\phi^{\text{ci}}(\boldsymbol{\lambda}, X))$ and $\boldsymbol{\lambda}^* \in \arg\min(\phi(\boldsymbol{\lambda}))$, where the potential functions are defined in Equations (D.5) and (D.6).

We define $\tilde{\boldsymbol{\lambda}}^*(x) = \boldsymbol{\lambda}^*(x)\frac{\mu(x)n_t}{n_t^x}$. As $\boldsymbol{\lambda}^{\text{ci}*}$ attains the minimum of ϕ^{ci}, we have:

$$\begin{aligned}
\phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X) &\le \phi^{\text{ci}}(\tilde{\boldsymbol{\lambda}}^*, X)\\
&= \sum_x\sum_t c_t(\lambda_t^*(x)\frac{\mu(x)n_t}{n_t^x})n_t^x + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1})\\
&\le \sum_x\sum_t(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}}c_t(\lambda_t^*(x))n_t^x + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1}) && \text{(D.10)}\\
&= \sum_x\sum_t(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1}c_t(\lambda_t^*(x))n_t^x\mu(x) + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1})\\
&\le \max_{x,t}(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*), && \text{(D.11)}
\end{aligned}$$

where the inequality (D.10) comes from the assumption on the costs and the inequality (D.11) comes from the fact that $\max_x(\frac{\mu(x)n_t}{n_t^x}) \geq 1$ (Indeed, we have by definition $\sum_x n_t^x = n_t = \sum_x \mu(x)n_t$. Thus, there exists $x \in \mathcal{X}$ such that $n_t^x \geq \mu(x)n_t$).

We can prove similarly that:

$$\phi(\boldsymbol{\lambda}^*) \leq \max_{x,t}(\frac{n_t^x}{\mu(x)n_t})^{p_{\max}-1}\phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*})$$

We thus obtain that:

$$\frac{1}{\max_{x,t}(\frac{n_t^x}{\mu(x)n_t})^{p_{\max}-1}}\phi(\boldsymbol{\lambda}^*) \leq \phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X) \leq \max_{x,t}(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*) \tag{D.12}$$

High probability bound on $\frac{\mu(x)n_t}{n_t^x}$

Hoeffding inequality implies that for all t, x, we have $P(|n_t^x - n_t\mu(x)| \geq k) \leq 2\exp(-\frac{2k^2}{n_t^2})$. We apply this with $k = n_t^{1/2+\epsilon}$ for $0 < \epsilon < 1/2$ to obtain:

$$P(|n_t^x - n_t\mu(x)| \geq n_t^{1/2+\epsilon}) \leq 2\exp(-2n_t^{2\epsilon}) \tag{D.13}$$

We thus have $P(\cup_{t,x}(|n_t^x - n_t\mu(x)| \geq n_t^{1/2+\epsilon})) \leq |X|\sum_t 2\exp(-2n_t^{2\epsilon})$. We also note that if we have $|n_t^x - n_t\mu(x)| \leq n_t^{1/2+\epsilon}$, then:

$$\frac{\mu(x)n_t}{n_t\mu(x) + n_t^{1/2+\epsilon}} \leq \frac{\mu(x)n_t}{n_t^x} \leq \frac{\mu(x)n_t}{n_t\mu(x) - n_t^{1/2+\epsilon}},$$

which yields:

$$\frac{\mu(x)}{\mu(x) + n_t^{\epsilon-1/2}} \leq \frac{\mu(x)n_t}{n_t^x} \leq \frac{\mu(x)}{\mu(x) - n_t^{\epsilon-1/2}}. \tag{D.14}$$

Combined with (D.12), this shows that with probability at least $|X|\sum_t 2\exp(-2n_t^{2\epsilon})$, we have:

$$\frac{1}{\max_{x,t}(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)})^{p_{\max}-1}}\phi(\boldsymbol{\lambda}^*) \leq \phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X) \leq \max_{x,t}(\frac{\mu(x)}{\mu(x) - n_t^{\epsilon-1/2}})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*)$$

We conclude this proof by computing the value of the potential of the complete information game with the linear regression game equilibrium:

$$\begin{aligned}
\phi^{\text{ci}}(\boldsymbol{\lambda}^*, X) &= \sum_x\sum_t c_t(\lambda_t^*(x))n_t^x + F((\sum_x xx^\top\sum_t \lambda_t^*(x)n_t^x)^{-1}) \\
&= \sum_x\sum_t c_t(\lambda_t^*(x))\frac{n_t^x}{\mu(x)n_t}n_t\mu(x) + F((\sum_x xx^\top\sum_t \frac{n_t^x}{\mu(x)n_t}n_t\mu(x))^{-1}) \\
&\leq \max_{x,t}(\frac{n_t^x}{\mu(x)n_t})\sum_x\sum_c c_t(\lambda_t^*(x))n_t\mu(x) + \frac{1}{(\min_{x,t}(\frac{n_t^x}{\mu(x)n_t}))^q}F((\sum_x xx^\top\sum_t \lambda_t^*(x)n_t\mu(x))^{-1}) \\
&\leq D_n\phi(\boldsymbol{\lambda}^*),
\end{aligned}$$

where $D_n = \max(\max_{x,t}(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)n_t}), \frac{1}{(\min_{x,t}(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}))^q})$.

Combined with the previous result, we obtain:

$$\phi^{\text{ci}}(\boldsymbol{\lambda}^*, X) \leq D_n\max_{x,t}(\frac{\mu(x) + n_t^{\epsilon-1/2}}{\mu(x)})^{p_{\max}-1}\phi^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}, X).$$

We can show similarly that:

$$\phi(\boldsymbol{\lambda}_{\text{ci}}^*) \leq D_n'\max_{x,t}(\frac{\mu(x)}{\mu(x) - n_t^{\epsilon-1/2}})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*),$$

where $D_n' = \max(\max_{x,t}(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}), \frac{1}{(\min_{x,t}(\frac{\mu(x)}{\mu(x)+n_t^{\epsilon-1/2}}))^q})$.

□

E Ordinary least squares

In this section we present the model where the analyst uses the OLS estimator instead of the GLS estimator. We show that, while the use of the OLS estimator removes a strong assumption of our model (the knowledge of the variance of the data points by the analyst), the use of OLS might also highly degrade the estimation cost when agents are not identical. We show in particular that for any game using the OLS estimator, a single agent participating with prohibitively high provision cost can ruin the estimation.

Let us first define the strategic linear regression in the OLS setting. Formally, the analyst receives n couples $(x_i, \hat{y}_i)$ and uses them to produce an estimate $\hat{\boldsymbol{\beta}}$ that is then sent to the agents. Note that we do not assume in this setting that the analyst receives the precision associated to the data points as it is not needed for the estimation. In what follows, we assume that the analyst computes this estimate by using *ordinary least squares* (OLS) and we denote it by $\hat{\boldsymbol{\beta}}_{\texttt{OLS}}$. OLS is the least squares regression which is optimal in the case of homoskedastic data. It is however sub-optimal when data are heteroskedastic but still applicable. It is one of the most widespread estimators in general, in particular because, unlike GLS, it is easy to apply and does not require knowledge of the variance of the data points. The covariance of OLS is independent of $\hat{y}_i$ and is equal to $\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\lambda_i(x_i)} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}$. Note that this quantity is well defined only if each $\lambda_i(x_i)$ is strictly positive, unlike GLS that only requires the information matrix $(\sum_i \lambda_i(x_i) x_i x_i^\top)$ to be invertible.

In a system where data point $\hat{y}_i$ is revealed with precision ℓ_i, the covariance of $\hat{\boldsymbol{\beta}}_{\texttt{OLS}}$ is

$$\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\ell_i} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}$$

In our model, the values of x_i are generated randomly according to a common underlying distribution μ on $\mathcal{X}$. Hence, we define the OLS estimation cost as

$$C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}) = F\left(\mathbb{E}\left[\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\lambda_i(x_i)} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}\right]\right). \tag{E.1}$$

We denote $\Gamma_{\texttt{OLS}}$ (resp. $\Gamma_{\texttt{GLS}}$) an instance of the game where the analyst uses the OLS (resp. GLS) estimator. For a given precision profile, we define $\phi_{\texttt{OLS}}(\lambda_i, \boldsymbol{\lambda}_{-i})$

$$\phi_{\texttt{OLS}}(\boldsymbol{\lambda}) = \sum_{j=1}^{n} \mathbb{E}\left[c_j(\lambda_j(x))\right] + C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}). \tag{E.2}$$

We show that our main results still holds in this setting.

Proposition E.1. *Under Assumptions 1, 2, and 3, a precision profile $\boldsymbol{\lambda}^*$ is a Nash equilibrium of the OLS linear regression game if and only if it minimizes $\phi_{\texttt{OLS}}$. Such an equilibrium exists. It is unique if all provision cost functions c_i are strictly convex. When there are multiple equilibria, the estimation cost $C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}^*)$ does not depend on the equilibrium.*

The proof of this proposition is a trivial adaptation of Proof C.1. The game $\Gamma_{\texttt{OLS}}$ thus has the same basic properties as $\Gamma_{\texttt{GLS}}$ and we can now state our main result in this new model:

Proposition E.2. *Let $\Gamma_{\texttt{OLS}}$ be a game satisfying the Assumptions of Theorem 3. Then, with the same constants $d, D > 0$ as Theorem 3 that do not depend on n, we have that:*

$$dn^{-q\frac{p_{\min}-1}{p_{\min}+q}-\alpha} \leq C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}^*) \leq Dn^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \tag{E.3}$$

where $\alpha = q\frac{(p_{\max}-p_{\min})(q+1)}{p_{\max}(q+p_{\min})}$.

Proof. **Upper bound**

The proof of the upper bound is the same as in the proof of Theorem 3. Indeed, we compute the value of the potential function for a particular constant strategy in which all players use the precision $\lambda(x) = n^{-\frac{q+1}{p_{\min}+q}}$ for all values of $x \in \mathcal{X}$. It is then sufficient to observe that for such a strategy, we have homoskedasticity of the data points. Thus, the GLS estimator and the OLS estimator are the same and the algebra of the proof can trivially be applied.

Lower bound

It is sufficient to observe that for all $\boldsymbol{\lambda}$, we have $\mathbb{E}\left[\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\lambda_i(x_i)} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}\right] \succeq \left(\mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]\right)^{-1}$ by Aitken's theorem of optimality of GLS.

□

Differences between the asymptotic behavior of GLS and OLS

In this section, we show that, while our main result holds when the analyst uses the OLS estimator, $\Gamma_{\texttt{GLS}}$ and $\Gamma_{\texttt{OLS}}$ behave fundamentally differently when only subsets of agents satisfy our non-trivial assumptions.

Proposition E.3. *Assume that Assumptions 1, 2 and 3 hold. Assume that for all $i \in N$, we have $c_i(0) = 0$. Additionally, assume that there exist $p_{\min} \geq 1$ a function $c_{\max} : \mathbb{R}_+ \to \mathbb{R}_+$ and $S_N \subseteq N$ such that for all $i \in S_N$ and all $a > 1, \ell > 0$: $a^{p_{\min}} c_i(\ell) \leq c_i(a\ell)$ and $c_i(\ell) \leq c_{\max}(\ell) < \infty$. Then there exists a constant $D > 0$ that does not depend on $|S_N|$ and such that:*

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq D|S_N|^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \tag{E.4}$$

Proof. We define the particular constant strategy

$$\lambda_i(x) = \begin{cases} |S_N|^{-\frac{q+1}{p_{\min}+q}} & \text{if } i \in S_N \\ 0 & \text{Otherwise.} \end{cases}$$

The algebra to obtain the bound is then exactly the same as in Section C.5.

□

This proposition states that for any subset of agents, the convergence rate of the estimation cost is at least as good as if only those agents participated. For example, if half a population suffers from linear provision cost $c_i(\lambda) = \lambda$ while the other half of the population has highly convex provision costs $c_i(\lambda) = \lambda^p$, the estimation cost will converge to 0 with rate at least $n^{-q\frac{p-1}{p+q}}$. This is significant as we have previously proved that if only agents with linear provision costs participate, GLS is not consistent and the estimation cost does not go to 0. This property is tightly linked to the GLS estimator. Indeed, GLS weights the data points according to their precision and low precision data points do not hinder the estimation. Formally, for any $\boldsymbol{\lambda}$, λ_{n+1}, we have $\sum_i \lambda_{i=1}^{n+1}(x_i) x_i x_i^\top \succeq \sum_{i=1}^{n} \lambda_i(x_i) x_i x_i^\top$ thus adding a data point can only improve the information matrix of the estimator. This is no longer true when using the OLS estimator as it gives the same weight to widely inaccurate data points as to very precise data points.

We show this difference on an example. We consider an OLS regression game where n agents are willing to give precise data (they have low provision cost) while one agent suffers from prohibitively high provision cost. Formally, let us consider $\Gamma_{\texttt{OLS}}$ the game where $\mathcal{X} = \{1\}$, $n+1$ agents participate, $c_i(\lambda) = \lambda^p$ for all i in $\{1, \ldots, n\}$ and $c_{n+1}(\lambda) = (n+1)^2\lambda$. In the following game, we also consider the scalariation $F(\cdot)$ to be the trace which in this case is the identity function. We have in this game the following potential:

$$\phi_{\texttt{OLS}}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i^p + (n+1)^2 \lambda_{n+1} + \frac{1}{(n+1)^2} \sum_{i=1}^{n+1} \frac{1}{\lambda_i} \tag{E.5}$$

It is then easy to show that at equilibrium, we have $\lambda_i^* = (n+1)^{-2/(p+1)}$ for all i in $\{1, \ldots, N\}$ and $\lambda_{n+1}^* = (n+1)^{-2}$. This implies that the equilibrium achieves the following estimation cost:

$$C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}^*) = \frac{1}{(n+1)^2} n (n+1)^{2/(p+1)} + 1 \tag{E.6}$$

This estimation cost does not converge to 0 when $n+1$ grows large. Also note that even if p grows large meaning that n of the $n+1$ agents almost do not suffer any cost for providing data, the estimation cost still does not converge to 0. In contrast, the cost functions we defined satisfy the assumptions of Proposition E.3 meaning that if the analyst used the GLS estimator, they would obtain a consistent estimator with convergence rate at least $n^{-q\frac{p-1}{p+q}}$. Alternatively, if the analyst refused the participation of agent $n+1$, they would also obtain a consistent estimator. This implies that designing a mechanism to control participants in the OLS model could greatly improve the estimation cost at equilibrium in some cases. This remains an open problem.

F Extension to joint distributions

In this section, we show how our main result can be extended to a setting where the data points x_i of agents are not independent and identically distributed but are distributed according to a joint distribution μ_{joint}. We make the following assumption on this joint distribution to ensure the non-triviality of the game:

Assumption 4. The set $\mathcal{X}$ is finite and $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} x_i x_i^\top\right]$ is positive definite.

For the rest of this section, we omit the subscript denoting that the expected value is taken with regard to the distribution μ_{joint}.

Having a joint distribution does not change the basic structure of the game. The game is still a potential game with potential

$$\phi(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{j=1}^{n} c_j(\lambda_j(x_j))\right] + C_{\text{estim}}(\boldsymbol{\lambda}).$$

Note that we still assume that each agent strategy is a function $\lambda_i : \mathcal{X} \to \mathbb{R}_+$ for ease of notation. We do not assume, however, that each agent holds each vector with non-zero probability. This implies that if an equilibrium exists, there exists an infinite number of equilibrium as agents may freely choose the precision of the data points that hold with probability zero (without changing their payoffs). As these precision are a simple modeling artifact without any impact on payoffs, we set them to 0 by convention.

Also note that there may now exist Nash equilibria $\boldsymbol{\lambda}^*$ for which $C_{\text{estim}}(\boldsymbol{\lambda}^*) = \infty$. For instance, if $d \geq 2$ and the joint distribution is such that $\mu_{\text{joint}}(\boldsymbol{x}) = 1$ for some $\boldsymbol{x} = (x_1, \ldots, x_n)$, then $\boldsymbol{\lambda}^* = 0$ is a Nash equilibrium. Indeed, in that case, no agent has an incentive to deviate since a single agent deviation still yields a non-invertible information matrix (recall that the information matrix is $\mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]$ and that the covariance is the inverse of this matrix). More generally, any profile $\boldsymbol{\lambda}$ such that the information matrix is non-invertible and remains non-invertible under unilateral deviations is an equilibrium. Following [24], we call Nash equilibria at which the estimation cost is infinite "trivial equilibria." These are not our focus as they can be avoided with model adjustments such as having d non-strategic agents with data points spanning $\mathbb{R}^d$ guaranteeing a finite covariance.

We claim that Proposition 1 (which states that the game has at least one equilibrium, and that if there are multiple equilibria, they have the same estimation cost) still holds for non-trivial equilibria under the extending model where the data points x_i are distributed according to a joint distribution μ_{joint} satisfying Assumption 4; with the following adapted proof. Note that the first step of this version of the proof is inspired from [24] to handle trivial equilibrium.

Proof. **Step 1: The potential function is convex.** The potential function $\phi(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{j=1}^{n} c_j(\lambda_j(x_j))\right] + C_{\text{estim}}(\boldsymbol{\lambda})$ takes values in the extended positive real numbers line $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$.

Recall that $C_{\text{estim}}(\boldsymbol{\lambda}) = F\left(\left(\mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]\right)^{-1}\right)$. We denote $V(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]^{-1}$ and $M(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]$. We have that $V(\boldsymbol{\lambda})$ is strictly convex and goes to infinity when $M(\boldsymbol{\lambda})$ goes to a non-invertible matrix (i.e., the largest eigenvalue of V goes to infinity for any sequence $\boldsymbol{\lambda}_n$ that converges to a $\boldsymbol{\lambda}$ such that $M(\boldsymbol{\lambda})$ is non-invertible). As F is convex and increasing, this shows that $C_{\text{estim}}(\boldsymbol{\lambda})$ is strictly convex and goes to $+\infty$ when $M(\boldsymbol{\lambda})$ goes to a non-invertible matrix, which then implies that $C_{\text{estim}}(\boldsymbol{\lambda}) : \mathbb{R}_+^n \to \bar{\mathbb{R}}_+$ is continuous. As the functions c_i are convex, we conclude that the potential function ϕ is strictly convex and continuous on $\bar{\mathbb{R}}_+$.

Step 2: The potential admits a minimum. We first consider the potential evaluated at an arbitrary value and show that this implies boundedness of agents precision at equilibrium. Let $\phi(\mathbf{1}) = \mathbb{E}\left[\sum_i c_i(1)\right] + F((\mathbb{E}\left[\sum_i x_i x_i^T\right])^{-1})$. By Assumption 3, $\lim_{\ell\to+\infty} c_i(\ell) = +\infty$. For all $x \in \mathcal{X}$, we denote $\mu_i(x)$ the the probability that agent i has data point x when data points are generated with the joint distribution μ_{joint}. If $\mu_i(x) = 0$, then the value of $\lambda_i(x)$ does not change the potential and we can set it to 0. Otherwise, $\lim_{\ell\to+\infty} c_i(\ell)\mu(x) = +\infty$. Hence, there exists ℓ_{max} such that for all i and all x, $c_i(\ell_{\text{max}})\mu(x) > \phi(\mathbf{1})$. This shows that if $\boldsymbol{\lambda}$ is a precision profile such that $\lambda_i(x) > \ell_{\text{max}}$ for some i and x, then $\phi(\boldsymbol{\lambda}) \geq \phi(\mathbf{1})$.

Let B be the subset of $\boldsymbol{\lambda}$ such that $\phi(\boldsymbol{\lambda}) \leq \phi(\mathbf{1})$. By continuity and convexity of ϕ, B is a non-empty convex and compact subset of $[0, \ell_{\text{max}}]^n$ on which $\phi(\boldsymbol{\lambda}) < \infty$. This implies that there ϕ admits a minimum and that all global minimum of ϕ are attained in B.

If different non-trivial equilibria exist, they have the same estimation cost. This step is strictly the same as the proof found in Section C.1. □

We are now ready to state our main result adapted to this setting. In the following theorem, $\boldsymbol{\lambda}^*$ denotes any non-trivial equilibrium.

Theorem 5. *Assume that Assumptions 2, 3, and 4 hold. Additionally, assume that there exist $p_{\min}, p_{\max} \geq 1$ and functions $c_{\min}, c_{\max} : \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $i \in N$ and all $a > 1, \ell > 0$: $a^{p_{\min}} c_i(\ell) \leq c_i(a\ell) \leq a^{p_{\max}} c_i(\ell)$ and $0 < c_{\min}(\ell) \leq c_i(\ell) \leq c_{\max}(\ell) < \infty$. Then there exist constants $d', D' > 0$ that depend on n only through $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$ and such that:*

$$d' n^{-q\frac{p_{\min}-1}{p_{\min}+q}-\alpha} \leq C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq D' n^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \tag{F.1}$$

where $\alpha = q\frac{(p_{\max}-p_{\min})(q+1)}{p_{\max}(q+p_{\min})}$.

Proof. In this first step, we compute the value of the potential function for a particular constant strategy in which all players use the precision $\lambda(x) = n^{-\frac{q+1}{p_{\min}+q}}$ for all values of $x \in \mathcal{X}$. By abuse of notation, we denote this precision profile by $(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}})$. The value of the potential for this precision profile is

$$\begin{aligned}
\phi(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}}) &= \mathbb{E}\left[\sum_{i=1}^{n} c_i(n^{-\frac{q+1}{p_{\min}+q}})\right] + F((\mathbb{E}\left[\sum_{i=1}^{n} x_i x_i^T n^{-\frac{q+1}{p_{\min}+q}}\right])^{-1}) \\
&= \sum_{i=1}^{n} c_i(n^{-\frac{q+1}{p_{\min}+q}}) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}}\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1}) \\
&\leq \sum_{i=1}^{n} n^{-p_{\min}\frac{q+1}{p_{\min}+q}} c_i(1) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}}\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1}) && \text{(F.2)} \\
&= n^{-p_{\min}\frac{q+1}{p_{\min}+q}} \sum_{i=1}^{n} c_i(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1}) && \text{(F.3)} \\
&\leq n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}} c_{\max}(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1}) && \text{(F.4)} \\
&= n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}} \left(c_{\max}(1) + F((\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1})\right), && \text{(F.5)}
\end{aligned}$$

where we use that $c_i(1) \geq a^{p_{\min}} c_i(1/a)$ with $a = n^{\frac{q+1}{p_{\min}+q}}$ (from the theorem's assumption) in (F.2), the homogeneity of F (Assumption 2) in (F.3), and the theorem's assumption, which implies that $c_i(1) \leq c_{\max}(1)$ for all i, in (F.4).

As $c_i(\ell) \geq 0$ and $\boldsymbol{\lambda}^*$ is a minimum of the potential, it holds that

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq \phi(\boldsymbol{\lambda}^*) \leq \phi(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}}).$$

Hence, the right-hand-side of (F.1) holds with $D = \left(c_{\max}(1) + F((\mathbb{E}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1})\right)$.

Lower bound. The lower bound is then simply obtained by plugging the new upper bound of the potential to the proof of the lower bound obtained in Section C.5. □

The main difference between Theorem 5 and Theorem 3 is that in Theorem 3, the constants d and D *do* not depend n whereas in Theorem 5, the constants d' and D' do depend on $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$. This is because in Theorem 5, we do not make any assumption on the joint distribution. We thus do not have any guarantee that the joint distribution will have some stable property when the number of agents grow. On the other hand, if $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$ is independent on n, the constants d' and D' will also not depend on n.

In fact, the multiplicative terms of Theorem 5 are simply obtained by replacing $\mathbb{E}\left[xx^T\right]$ with $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$ in the multiplicative terms of Theorem 3 (note that we retrieve Theorem 3 when data points are iid). This latter term captures precisely the impact of correlation on the estimation cost. For instance,if data points are highly correlated in a way that poorly represents the input space, $F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right])^{-1})$ can be arbitrarily large, leading to a commensurately large upper bound (the corresponding lower bound behavior is similar).

G Additional illustrations

G.1 Illustration of the equilibrium characterization

In this section, we provide additional illustrations on the equilibrium characterization (Section 4), which complement Figure 1 and show that the discussion on that figure in the paper continues to apply in different settings, namely:

a. In Figure 3, we vary the degree d of the polynomial regression (Figure 1 has $d = 4$).
b. In Figure 4, we vary the distribution μ (Figure 1 has a uniform distribution that corresponds to the first row in Figure 4). Here, we fix $d = 4$ and we do not plot the optimal design as it does not depend on μ.
c. In Figure 5, we use a different scalarization, the squared Frobenius norm ($F(M) = \sum_{ij} M_{ij}^2$), while keeping a uniform distribution μ and $d = 4$.

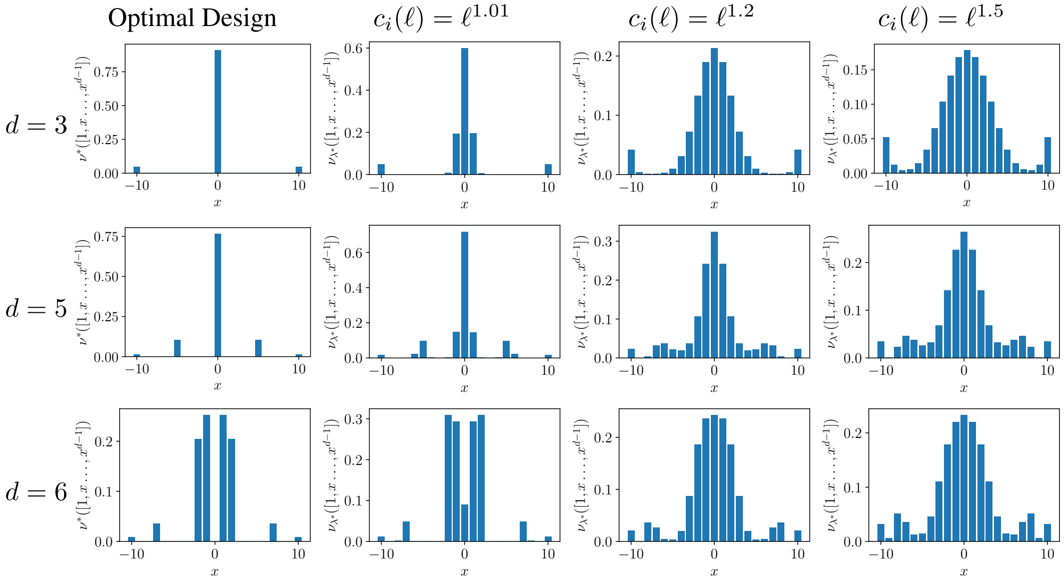


Figure 3: Optimal design ν^* and allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with various degrees d of the polynomial regression (here μ is uniform and the scalarization is the trace as in Figure 1).

Figure 3 illustrates the optimal design ν^* and the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 2 in the same setting as Figure 1 ($d = 4$) with different degrees for the polynomial regression ($d = 3, 5, 6$). We observe that for $d = 3$ and $d = 5$, the optimal design puts maximal weight on the central vector $[1, x, \cdots, x^{d-1}]$ with $x = 0$ while for $d = 4$ and $d = 6$, this vector does not belong to the support of the optimal design. We observe a similar property for the equilibrium of games with near-linear data provision cost. The allocations of precision at equilibrium for $p = 1.2$ and $p = 1.5$, however, are significantly different from the optimal design for all values of d (in particular with a maximum of precision for the central vector with $x = 0$), and they have a shape that does not significantly vary with the degree d.

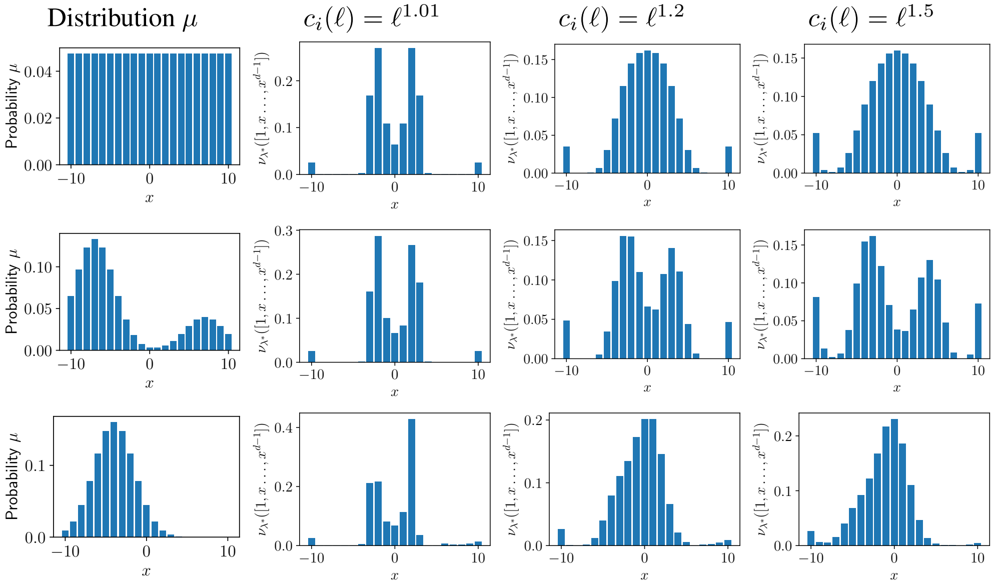


Figure 4: Allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with various distributions μ (here $d = 4$ and the scalarization is the trace as in Figure 1). The optimal design ν^* does not depend on μ and is therefore the same as in Figure 1.

Figure 4 illustrates the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 2 in the same setting as Figure 1 with various distributions μ of the agents' x_i vectors. The first row of graphs corresponds to the exact same setting as Figure 1 (uniform distribution) while the next rows show the results for other distributions. In addition to Figure 1, we plot the results for monomial costs of exponent $p = 1.5$, but we do not plot the optimal design ν^* as it is the same for all distributions (and shown on Figure 1). We first observe that, for all distributions, the allocation of precision at equilibrium is close to the optimal design (and hence almost independent of the distribution) for near-linear provision costs ($p = 1.01$). For more convex provision costs however, the allocation of precision at equilibrium varies with μ in non-trivial ways. In the second row of Figure 4 (compared to the first), we observe that $\nu_{\boldsymbol{\lambda}^*}([1, x, \cdots, x^{d-1}])$ shrinks for values of x close to 0. This is explained by two factors: i) vectors with x close to 0 have a low probability according to μ and ii) provision costs are superlinear meaning that the agent cannot compensate this probability by multiplying the precision attributed to this vector without prohibitively increasing its cost. We observe a similar behavior for the third row of Figure 4 where $\nu_{\boldsymbol{\lambda}^*}$ has a shape similar to the first row with values skewed to the left where vectors have higher probability.

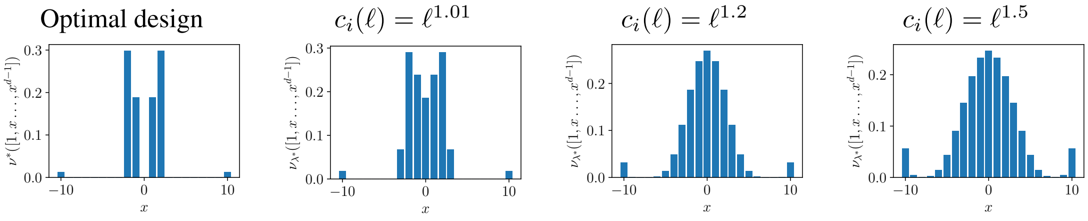
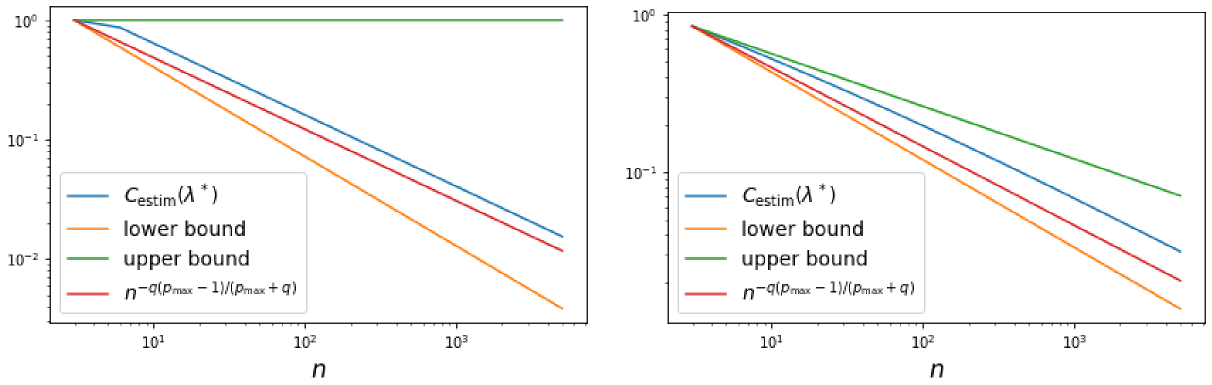


Figure 5: Optimal design ν^* and allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with the squared Frobenius norm as a scalarization F (here μ is uniform and $d = 4$ as in Figure 1).

Figure 5 illustrates the optimal design ν^* and the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 2 in the same setting as Figure 1 but when using the squared Frobenius norm as a scalarization to define the estimation cost instead of the trace. We observe that both figures show similar trends. In particular, Figure 5 with the squared Frobenius norm exhibits the same behaviors as discussed before on Figure 1 for the trace: the allocation of precision at equilibrium is close to the optimal design for $p = 1.01$ while it departs significantly for $p = 1.2$ and $p = 1.5$ where the precision for the vector $[1, 0, \ldots, 0]$ is maximal (instead of zero in the optimal design).



(a) Comparison for $p_{\min} = 1$ and $p_{\max} = 4$

(b) Comparison for $p_{\min} = 2$ and $p_{\max} = 3$

Figure 6: Comparison of the rate of convergence of the estimation cost with different bounds for agents with heterogeneous costs

G.2 Numerical exploration of Theorem 3

In this section, we explore the result of Theorem 3 through numerical simulations. We consider a one-dimensional model with $\mathcal{X} = \{1\}$. The scalarization is the trace (which satisfies Assumption 2 with $q = 1$). This means that $C_{\rm estim}(\boldsymbol{\lambda}) = (\sum_i \lambda_i(1))^{-1}$. Recall that Theorem 3 shows that

$$dn^{-q\frac{p_{\min}-1}{p_{\min}+1}-\alpha} \leq C_{\rm estim}(\boldsymbol{\lambda}^*) \leq Dn^{-q\frac{p_{\min}-1}{p_{\min}+1}}.$$

The goal of this section is to compare the upper and lower bounds of Theorem 3 to $C_{\rm estim}(\boldsymbol{\lambda}^*)$, to see if the true convergence rate is close to the lower or to the upper bound.

In the remaining of this subsection, we will display $C_{\rm estim}(\boldsymbol{\lambda}^*)$ as a function of n in loglog-scale and compare it to three possible convergence rates:

(a) $n^{-q\frac{p_{\min}-1}{p_{\min}+q}-\alpha}$ (the rate of the lower bound of Theorem 3);

(b) $n^{-q\frac{p_{\min}-1}{p_{\min}+q}}$ (the rate of the upper bound of Theorem 3, which is the convergence rate when all players have cost $c_i(\ell) = \ell^{p_{\min}}$);

(c) $n^{-q\frac{p_{\max}-1}{p_{\max}+q}}$ (the convergence rate when all players have cost $c_i(\ell) = \ell^{p_{\max}}$).

Note that (a) is the fastest convergence rate, followed by (c) and then by (b).

In all plots in this section, we normalize the values such that they all start at the same point for $n = 3$ ($n = 3$ is the smallest game for which we compute $C_{\rm estim}(\boldsymbol{\lambda}^*)$).

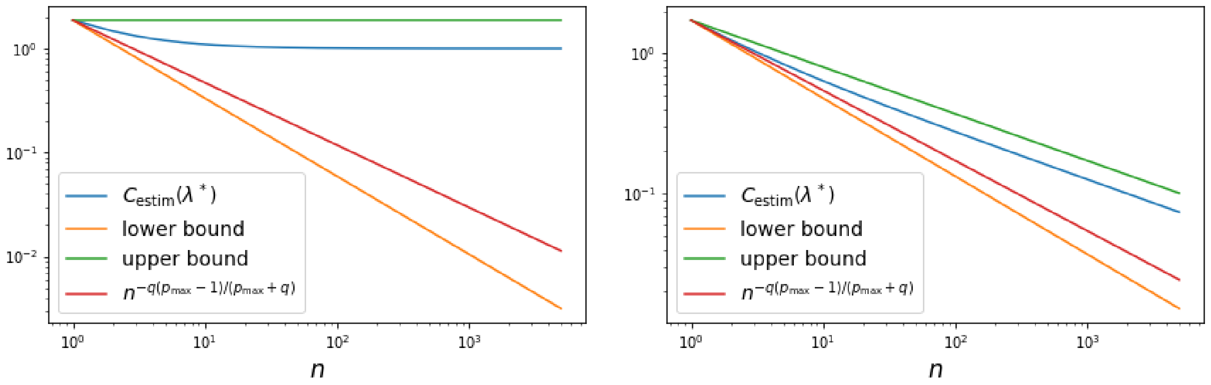
G.2.1 Heterogeneous agents with different exponents

We first consider heterogeneous agents. For a given n, $n/3$ agents have provision costs $c_i(\ell) = \ell^{p_{\max}}$ and $2n/3$ agents have provision costs $c_i(\ell) = \ell^{p_{\min}}$. This setup satisfies the assumptions of Theorem 3 with the corresponding $p_{\min}$ and $p_{\max}$. We consider two setups: $(p_{\min}, p_{\max}) = (1, 4)$ and $(p_{\min}, p_{\max}) = (2, 3)$.

Figure 6 compares the convergence rate of $C_{\rm estim}(\boldsymbol{\lambda}^*)$ to the three bounds defined above. This figure suggests that the estimation cost behaves as when all players have estimation cost $\ell^{p_{\max}}$. Intuitively, this is explained by the fact that in the game, an agent that has a cost $c_i(\ell) = \ell^{p_{\min}}$ will give a very small precision. Hence, the game will almost behave as if this agent was not in the game. This explains why the convergence rate of $C_{\rm estim}(\boldsymbol{\lambda}^*)$ is driven by agents having exponent $p_{\max}$.

G.2.2 Agents with polynomial provision costs

We then consider agents with polynomial provision costs. We assume that the n agents have the same provision costs $c_i(\ell) = \sum_{k=p_{\min}}^{p_{\max}} \ell^k$. Again, these provision cost satisfy the assumptions of Theorem 3 with the corresponding $p_{\min}$ and $p_{\max}$.



(a) Comparison for $p_{\min} = 1$ and $p_{\max} = 4$

(b) Comparison for $p_{\min} = 2$ and $p_{\max} = 3$

Figure 7: Comparison of the rate of convergence of the estimation cost with different bounds for agents with polynomial costs

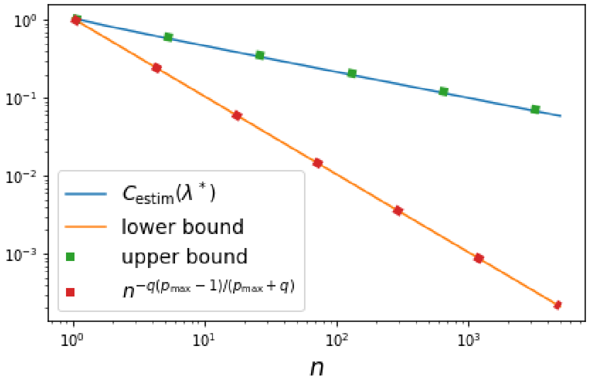


Figure 8: Comparison of the rate of convergence of the estimation cost with the upper bound of Theorem 3 for agents with hyperbolic cosine costs.

Figure 7 compares the convergence rate of the covariance to the upper and lower bounds of Theorem 3. We observe that the convergence rate is close to the upper bound $n^{(p_{\min}-1)/(p_{\min}+1)}$. This result is natural as polynomials are sums of monomials and it is logical to expect the convergence rate to be according to the ”worst” monomial of degree $p_{\min}$.

G.2.3 Agents with non-polynomial provision costs

This result on polynomial functions alongside the fact that the precision of each agent goes to 0 when the number of agents goes to infinity hints at the behavior of the estimation cost with more general provision costs. Indeed, if agents have provision cost which have a Taylor expansion at 0, their cost can be well approximated by a polynomial function. The previous figure then suggests that the convergence rate in this case is driven by the first non-null term of the Taylor expansion of the function of degree $p_{\min}$.

We illustrate this in Figure 8 where we consider homogeneous agents with provision costs $c_i(\ell) = \cosh(\ell) - 1$. Recall that $\cosh(\ell) - 1 = \sum_{k=1}^{\infty} \frac{\ell^{2k}}{(2k)!}$. This model therefore satisfy our assumptions with $p_{\min} = 2$ and $p_{\max} = \infty$. According to our previous observations, we expect the convergence rate in this case to be the upper bound $(p_{\min}-1)/(p_{\min}+1)$ with $p_{\min} = 2$. Note that in this case our lower bound and $n^{-q(p_{\max}-1)/(p_{\max}+1)}$ both represent convergence rates of n^{-q} corresponding to the non strategic setting. Figure 8 suggests indeed that the convergence rate is close to this upper bound.

H Hardware and software used for experiments

All experiments were run on a Dell xps-13 laptop with a Quad core Intel Core i7-8550U (-MT-MCP-) CPU under Ubuntu 18.04. Experiments were made using Python 3 code which is publicly available at `https://gitlab.inria.fr/broussil/linear-regression-with-strategic-data-sources`. The main libraries used are presented in README.md and the versions used for the experiments are available in requirements.txt.