Learning in the Presence of Strategic Data Sources: Models and Solutions

Benjamin Roussillon

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques Appliquées**

Arrêté ministériel : 25 mai 2016

Présentée par

Benjamin ROUSSILLON

Thèse dirigée par **Patrick LOISEAU**
et codirigée par **Panayotis MERTIKOPOULOS**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
et de l'École Doctorale **MSTII**

Apprentissage en présence de données stratégiques : modèles et solutions

Learning in the Presence of Strategic Data Sources: Models and Solutions

Thèse soutenue publiquement le **15 septembre 2021**,
devant le jury composé de :

Mathias HUMBERT
Chercheur, Cyber Defence Campus, Suisse, Rapporteur

Rachid EL AZOUZI
Professeur des universités, LIA, Université d'Avignon, France, Rapporteur

Marie-Christine ROUSSET
Professeur des universités, LIG, Université Grenoble-Alpes, France, Examinateur

Johanne COHEN
Directrice de recherche, LRI, CNRS, France, Examinateur

Yann CHEVALEYRE
Professeur des universités, LAMSADE, université Paris Dauphine-PSL, France, Examinateur

Panayotis MERTIKOPOULOS
Chargé de recherche, LIG, CNRS, France, Co-encadrant de thèse

Patrick LOISEAU
Chargé de recherche, LIG, CNRS, France, Directeur de thèse

Remerciements

(Acknowledgments)

Mes premiers remerciements vont à mes encadrants, Patrick Loiseau et Panayotis Mertikopoulos. Merci à Patrick de m'avoir présenté un sujet de stage et de thèse qui m'a plu dans un domaine que je ne connaissais pas bien à l'époque. Merci à Panayotis pour le recul que tu m'as apporté lors de certains moments difficiles de la thèse.

Je voudrais aussi remercier Mathias Humbert, Rachid El Azouzi, Marie-Christine Rousset, Johanne Cohen et Yann Chevaleyre d'avoir accepté de faire partie de mon jury et pour leurs questions et remarques pertinentes. Merci particulièrement à Mathias et Rachid pour leur lecture attentive du manuscrit et leur remarques qui donnent des perspectives de travaux futurs.

Je remercie également Nadia Brauner qui m'a fait découvrir le monde de la recherche pendant mes études et qui a toujours été disponible pour répondre à mes questionnements, notamment sur l'enseignement. J'en profite pour remercier aussi Carole Adam avec qui j'ai eu le plaisir de donner des cours d'informatique qui restent parmi les meilleurs souvenirs de ma thèse.

Une thèse n'est évidemment pas qu'un travail scientifique mais aussi trois années de vie et pour cela je voudrais remercier mes amis de m'avoir toujours accompagné. Merci à ceux que j'ai rencontré lors de mes études et de ma thèse. Merci à Quan, Clément et Vitalii. Merci particulièrement à Alexis avec les soirées sushi après l'escalade ou même les nombreuses discussions du vendredi après-midi. Merci aussi à mes amis du Lycée que je retrouve régulièrement à Lyon et qui ont toujours été là pour moi. Merci à Niels et Marion (et les vacances au ski), Nicolas que je connais maintenant depuis 15 ans. Merci à Alexandre et son enthousiasme sans limite pour des soirées toutes mémorables ainsi que pour les moments plus calmes à simplement jouer à la coinche. Merci à Anaëlle pour toutes les soirées et journées passées à discuter autour d'un verre ou de jeux, merci pour le soutien sans faille que tu m'as toujours apporté. Tant de bons moments passés avec vous (le voyage en Turquie restera un souvenir exceptionnel) qui m'ont donné la motivation nécessaire pour continuer.

Merci à ma famille, Denis, Sylvette, Roland, mes cousins Yohan, Joana et Mathias avec lesquels les réunions de famille (notamment au ski à coup de raclettes et de gâteaux au chocolat) ont toujours été un plaisir. Merci à mes grand-mères Monique (j'ai arrêté de m'accrocher à tous les portails qui passent) et Raymonde (je connais maintenant mes tables de multiplication sur le bout des doigts). Merci évidemment à mon frère Pierre et à mes parents, Éric et Laurence sans qui rien n'aurait été possible. Je remercie finalement Camille qui m'a accompagné au quotidien pendant ces trois années et qui m'a toujours soutenu dans les moments difficiles. Bientôt la libération pour toi aussi !

Abstract / Résumé

Abstract

In this thesis, we consider the problem of learning when data are strategically produced. This challenges the widely used assumptions in machine learning that test data are independent from training data which has been proved to fail in many applications where the result of the learning problem has a strategic interest to some agents. We study the two ubiquitous problems of *classification* and *linear regression* and focus on fundamental learning properties on these problems when compared to the classical setting where data are not strategically produced.

We first consider the problem of finding optimal classifiers in an adversarial setting where the class-1 data is generated by an attacker whose objective is not known to the defender—an aspect that is key to realistic applications but has so far been overlooked in the literature. To model this situation, we propose a Bayesian game framework where the defender chooses a classifier with no *a priori* restriction on the set of possible classifiers. The key difficulty in the proposed framework is that the set of possible classifiers is exponential in the set of possible data, which is itself exponential in the number of features used for classification. To counter this, we first show that Bayesian Nash equilibria can be characterized completely via functional threshold classifiers with a small number of parameters. We then show that this low-dimensional characterization enables us to develop a training method to compute provably approximately optimal classifiers in a scalable manner; and to develop a learning algorithm for the online setting with low regret (both independent of the dimension of the set of possible data). We illustrate our results through simulations and apply our training algorithm to a real bank fraud data set in a simple setting.

We then consider the problem of linear regression from strategic data sources. In the classical setting where the precision of each data point is fixed, the famous Aitken/Gauss-Markov theorem in statistics states that generalized least squares (GLS) is a so-called "Best Linear Unbiased Estimator" (BLUE) and is consistent (the model is perfectly learned when the amount of data grows). In modern data science, however, one often faces strategic data sources, namely, individuals who incur a cost for providing high-precision data. We model this as learning 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. 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 show that, in general, Aitken's theorem does not hold under strategic data sources, though it does hold if individuals have identical provision costs (up to a multiplicative factor). When individuals have non-identical costs, we derive a bound on the improvement of the equilibrium estimation cost that can be achieved by deviating from GLS, under mild assumptions on the provision cost functions and on the possible deviations from GLS. We also 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.

Résumé

Dans cette thèse, nous considérons le problème de l'apprentissage lorsque les données sont produites de manière stratégique. Cela remet en question l'hypothèse largement utilisée dans l'apprentissage automatique selon laquelle les données de test sont indépendantes des données d'apprentissage. Cette hypothèse est invalide lorsque le résultat de l'apprentissage a un intérêt stratégique pour certains agents. Nous étudions les deux problèmes omniprésents de *classification* et *régression linéaire* et plus particulièrement leur propriétés fondamentales par rapport aux modèles classiques où les données ne sont pas produites stratégiquement.

Nous considérons d'abord le problème de la classification dans un contexte antagoniste où les données de classe 1 sont générées par un attaquant dont l'objectif n'est pas connu du défenseur — un aspect qui est essentiel pour des applications réalistes mais qui a jusqu'à présent été négligé dans la littérature. Nous proposons un jeu bayésien où le défenseur choisit un classificateur sans restriction *a priori* sur l'ensemble des classificateurs possibles. La principale difficulté de ce modèle est que l'ensemble des classificateurs possibles est exponentiel dans l'ensemble des données possibles, qui est lui-même exponentiel dans le nombre de caractéristiques utilisées pour la classification. Pour contrer cela, nous montrons tout d'abord que les équilibres de Nash Bayésiens peuvent être caractérisés complètement via des classificateurs à seuils exprimés avec un faible nombre de paramètres. Nous montrons ensuite que cette caractérisation de faible dimension permet de développer une méthode d'apprentissage utilisant des données d'entraînement pour calculer des classificateurs approximativement optimaux avec de fortes garanties et de développer un algorithme d'apprentissage en ligne satisfaisant la propriété du "no-regret" (nos résultats sont indépendants de la dimension de l'ensemble de données possibles).

Nous considérons ensuite le problème de la régression linéaire à partir de sources de données stratégiques. Dans le cadre classique où la précision de chaque point de données est fixe, le théorème d'Aitken/Gauss-Markov en statistique énonce que l'estimateur des moindres carrés généralisés (GLS) est ce que l'on appelle le "meilleur estimateur linéaire sans biais" et est consistant. Dans les applications récentes, cependant, les données peuvent être stratégiques, c'est-à-dire que la production de données précises est coûteuse. Nous modélisons cela comme l'apprentissage en présence de données stratégiques avec un bien public, c'est-à-dire que les données sont fournies par des agents stratégiques qui minimisent un coût individuel de production de données précises tout en bénéficiant de la précision globale du modèle. Nous modélisons l'incertitude sur les données des agents - un aspect critique

du problème lorsque le nombre d'agents est important. Nous montrons qu'en général, le théorème d'Aitken n'est plus valide dans ce cadre, bien qu'il soit maintenu si les individus ont des coûts de provision identiques. Lorsque les individus ont des coûts non identiques, nous donnons une borne sur l'amélioration du coût d'estimation à l'équilibre qui peut être obtenu en s'écartant de GLS. Nous caractérisons aussi l'équilibre du jeu, révélant une connexion intéressante avec le problème de design optimal. Par la suite, nous étudions le comportement asymptotique de la covariance des paramètres de régression linéaire estimés par GLS. Nous fournissons des bornes pour cette covariance et montrons que, lorsque les coûts de production des agents sont super-linéaires, la covariance du modèle converge vers zéro mais à un rythme plus lent que les problèmes d'apprentissage classiques. En revanche, si les coûts de production des agents sont linéaires, cette covariance ne converge pas. Cela montre que même la propriété de base de consistance GLS est compromise lorsque les sources de données sont stratégiques.

Acronyms and Notations

Acronyms

BLUE	$\triangleq$	Best linear unbiased estimator,
BNE	$\triangleq$	Bayesian Nash equilibrium,
OGD	$\triangleq$	Online gradient descent,
OLS	$\triangleq$	Ordinary Least Squares estimator,
GLS	$\triangleq$	Generalized Least Squares estimator,
NE	$\triangleq$	Nash equilibrium,
ROC	$\triangleq$	receiver operating characteristic,
SAA	$\triangleq$	Sample average approximation.

Notations and Conventions

Γ	$\triangleq$	Strategic linear regression game,
Γ_L	$\triangleq$	Strategic linear regression game considering estimator L,
$\mathcal{G}$	$\triangleq$	Adversarial classification game,
$\Delta(S)$	$\triangleq$	The set of probability distribution over the set S.

Throughout this thesis we also use the following conventions:

- For any $m \in \mathbb{N}$, we denote the set of integers $\{1, \dots, m\}$ by $[\![1, m]\!]$
- We use bold symbols to denote vectors (all vectors are column vectors) and subscripts to denote in coordinate. Thus, $\boldsymbol{x} = (x_1, \dots, x_n)$ is a n-coordinate vectors with i-th coordinate x_i.
- We denote the indicator function of a set $S \subset A$ by $\mathbb{1}_{S(x)} = \begin{cases} 1 & \text{if } x \in E \\ 0 & \text{Otherwise.} \end{cases}$. By abuse of notation, we denote for a function f the indicator $\mathbb{1}_{\{x' \in A, f(x)=c\}}(x)$ as $\mathbb{1}_{f(x)=c}(x)$.
- For a set S, we denote 2^S its powerset (the set of all subsets of S) and sometimes identify this powerset with the set of all indicator functions of subset of S.

- For a random variable X taking values $x \in \mathcal{X}$ and with probability distribution p, we denote $p(x) = p(X = x)$ by abuse of notation. We also denote $\mathbb{E}_{x\sim\mathcal{D}}[x] = \int_{\mathcal{X}} xp(x)dx$ and omit the subscript on the expected value if the context is clear.
- In game models, we denote a player by an index i and the set of other players by the notation $-i$.
- We use the asymptotic notations of Bachman-Landau and Knuth. For any functions $f : S \subseteq \mathbb{R} \to \mathbb{R}$ and $g : S \subseteq \mathbb{R} \to \mathbb{R}$, we write:
  - $f(x) = O(g(x))$ if there exists $M > 0$ and $x_0 \in S$ such that $f(x) \leq Mg(x)$ for all $x \geq x_0$.
  - $f(x) = \Omega(g(x))$ if there exists $M > 0$ and $x_0 \in S$ such that $f(x) \geq Mg(x)$ for all $x \geq x_0$.
  - $f(x) = \Theta(g(x))$ if $f(x) = O(g(x))$ and $f(x) = \Omega(g(x))$.

Contents

1 Introduction

1.1 Context

Machine learning is an extensive field allowing us to exploit the ever increasing quantity of available data to automate tasks performed by humans (such as self-driving cars) or to analyze data sets too large and complex to be processed by humans when classical programs fail to produce satisfactory answers. In many scenarios, the learning process is done in two distinct steps. First, an analyst gathers data which can be labeled (e.g., characteristics of bank transactions and whether or not these transactions were fraudulent) or have a value of interest (e.g., medical data where characteristics of patients are gathered to understand under which conditions a disease is severe or not). Then, the analyst uses the gathered data to produce a result which can take varying forms such as a classifier (classifying transactions between fraudulent and non-fraudulent) or a regression (a model associating characteristics to a value in order to predict the value of interest in future data points).

1.1.1 Failure of Classical Learning Schemes

A common assumption in many such settings is that the data gathering process is in some sense independent from the following analysis and its result. In classification it usually takes the form of the following assumption: "The training data and the test data are drawn from the same distribution". This implies that whatever the analyst does will not change the way data is produced compared to what happened before the analysis. In regression, the variance of data points does not depend on the analysis and is supposed to be a parameter of the problem.

In many applications however, these assumptions do not hold as the result of the analysis has strategic value. In these settings, agents may modify either the way they produce data or directly the data they send to the analyst to achieve their goal. This can lead to conflict between the data generation process and the analysis which we call either *adversarial* if data generators have interests directly in conflict with the objectives of the analyst or simply *strategic* in which case their interests may or may not align with the analyst.

Adversarial settings arise typically when some agents are malicious. For example, in classification, the goal of the analyst might be to recognize fraudsters, such as fraudulent twitter accounts as studied in Thomas et al. [Tho+13]. It is clear that fraudsters have an interest in not being classified as fraud and this leads to an adaptation to evade classification. Such behavior however is not limited to fraud and can also be observed when detecting network intrusion or preventing DDoS attacks. It is well known that using standard classification algorithms for this task leads to poor performance. Attackers are able to avoid detection by adjusting the data that they generate while crafting their attacks (evasion attacks) or to alter the training data set so that the resulting classifier performs poorly against them (poisoning attacks). Nelson et al. [Nel+09] show that attackers can easily fool a spam filter with access to a small portion of spam used to train the spam filter. Goodfellow et al. [GSS15] explain why many machine learning models are vulnerable to carefully crafted attacks (such as slightly modifying a few pixels from a pictures and completely changing the label a neural network attributes to it) called adversarial examples. Sommer and Paxson [SP10] show that the network intrusion detection problem is fundamentally difficult to approach from a machine learning perspective due its adversarial nature. Wang et al. [Wan+14] study malicious crowdsourcing (called crowdturfing) systems where attackers pay users to carry a range of attacks and exhibit that, while such attacks can efficiently be detected by machine learning methods, these methods are also very vulnerable to evasion and poisoning. There is a vast literature on adversarial classification to patch this weakness (see Section 3.1 for a detailed discussion), but these works often propose ad-hoc defense methods optimized against specific attacks without fully modeling the attacker's adaptiveness. This leads to an arms race as classifiers adapt to a specific type of attack and attackers find ways to circumvent these defenses.

Strategic settings are encountered when users do not have malicious intents but have some interest in the result of the analysis which may or may not conflict with the result they should obtain. Such behavior is observed not only in specific applications involving experts but also in the general population. For example, home buyers in America may open multiple credit cards for the sole purpose of improving their credit score. This omnipresence of strategic behavior has led to famous laws such as Goodhart's law (see Goodhart [Goo75]) which states that "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes." and was later generalized by Strathern [Str97] in the form of "When a measure becomes a target, it ceases to be a good measure.". Among all possible examples of strategic behavior to manipulate a measure, some of the most commonly cited are teachers being under pressure to obtain good results from their students and

focusing on performing in tests rather than in learning and universities manipulating rankings by focusing their efforts on the specific metrics used for the ranking (see [ES07; ESE16]). It should be noted, however, that the range of application of such laws is broader than the examples we mentioned and includes varied settings such as stores manipulating the quantity of product they request as shown in Caro et al. [Car+10] or simply when agents do not wish to be perceived as outliers if the result of the analysis can impact their life as discussed by Perote and Perote-Pena [PP04]. These strategic behavior affect the data produced by agents which then affects the result of the analysis. In the university case for example, analysts learn the rank of each university depending on characteristics manipulated by universities. There is a growing body of work on these strategic considerations in particular in the context of linear regression (see Section 6.1 for a detailed discussion) but these works usually focus how to minimize the estimation error in various setting or how to guarantee strategy proof estimators (where agents reveal their data truthfully) and do not study the large array of possible statistical properties of classical estimators applied in strategic settings.

1.1.2 Overcoming the Dependency Between Data Generation and Analysis

In recent years, many game-theoretic models have emerged to circumvent the faulty assumption that the data generation process is independent from the analysis. Indeed, game-theory is a tool of choice to model such interactions as the main solution concept of equilibrium implies that both the data generators and the analyst act taking into account the strategic behavior of the other.

Several game-theoretic models of adversarial classification have emerged over the last decade pioneered by models such as the one of Dalvi et al. [Dal+04] (see Section 3.1 for a more detailed discussion). Most of them, however, have several crucial limitations. In particular, they are deeply tied to classical learning algorithms. Such algorithms rely on a reduction of the hypothesis class (i.e., the set of possible classifiers from which the defender chooses its defense) to obtain tractable (often convex for example) optimization problems to minimize the classification error. This tie is both a strength of their models as they may be more likely to be adopted by users and a weakness as, while classically used hypothesis classes usually represent some prior knowledge on the shape the optimal classifier may have, this prior knowledge has no reason to be valid when considering adversarially produced data and may even be exploited by attackers. These game theoretic models also assume

complete information about the attacker's objective,[1] which is often too strong in practice as suggested by Vorobeychik and Kantarcioglu [VK18]. Finally, while some of these models enjoy theoretical guarantees of errors (when considering restricted sets of classifiers), they are usually aimed at practical applications and lack a fundamental understanding of more general properties.

Considerations of strategic behavior have also become central in an emerging literature on learning with *strategic data sources* (see Section 6.1 for a more detailed discussion). In these settings agents are not necessarily antagonist to the analyst but have their own objectives. For example, we previously mentioned stores manipulating the quantity of product they request (see [Car+10]). In this case, stores optimize their own benefits while the main company behind them optimizes their global benefits. Such problems define a first category of models where agents strategize to obtain a desired outcome from the analysis. This leads to authors studying estimators which are strategy proof as in Perote and Perote-Pena [PP04], Chen et al. [Che+18b], and Dekel et al. [DFP10a]. These works focus on estimators having the desired property that no agent misreports their data. In some other settings, agents incur some cost to provide data and strategize on this cost as they may for example demand monetary compensation for their efforts. This is the case in crowdsourcing with the model Dasgupta and Ghosh [DG13] or recommender systems (see [ERK11; AZ97; Har+05]) where providing content or feedback requires effort, or in applications where the data is produced by costly computations. Costly data production may also come from 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 and add noise to it (hence decreasing the precision as seen in Warner [War65] and Duchi et al. [DJW13]). When providing high-quality data comes at a cost, it makes sense to consider strategic behavior among data sources. In particular, one should ask: why would strategic data sources provide any data at all? 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., Cai et al. [CDP15], Liu and Chen [LC16a], and Westenbroek et al. [Wes+20] 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 (see [Yan+19; Kon+16; GKN17]), etc. An additional issue in such applications is that

[1] with the exception [Gro+13], but on regression.

the number of participating agents is typically large, so there is a commensurate degree of uncertainty regarding the state or incentives of other agents.

1.1.3 Perspective of the Thesis

As our previous literature discussion shows, the problem of characterizing fundamental properties of learning algorithms when data is strategically or even adversarially produced is not sufficiently addressed. To the best of our knowledge, only Dritsoula et al. [DLM17], Cullina et al. [CBM18], and Bhagoji et al. [BCM19] characterize some classical learning properties in adversarial settings while Pinot et al. [Pin+20] and Meunier et al. [Meu+21] lay game theoretic foundations for further extension by showing that mixed Nash equilibrium exist in a model of practical importance. For strategic considerations (in particular in regression), authors usually consider how to minimize errors (see Abernethy et al. [Abe+15]) or guarantee that truthful data are obtained (see Chen et al. [CSZ20]) while the field of classical statistical properties to consider is much larger. Furthermore, the considered applications usually gather a very large amount of data from varied sources and characterizing the precise properties and goals of the different actors is unrealistic. For example, students competing against other students have no reasonable way to determine the exact characteristics of their competitors while networks face many different threats with varying degrees of seriousness.

This thesis studies these problems where results are lacking. In particular, we study learning problems in the presence of strategic data sources using a game theoretic approach and considering the solution concept of Nash equilibrium which finely models the adaptation of both the data producers and the analyst to the actions of the other. We are interested in fundamental learning properties of models applicable to settings where a large amount of data is gathered. This means that we should guarantee the scalability of our methods both in terms of computation power and information required. We aim to develop model and solution methods applicable to a large array of settings covering both adversarial and strategic settings going from bank fraud and network intrusion detection to learning with costly information as seen in federated learning or medical settings.

1.2 Contributions and Organization of the Thesis

In this thesis, we focus on learning problems where the data generation process is *not* independent from the result of the analysis and model such settings using game theory which allows us to consider the adaptation of both the data generation process and the analyst to the actions of the other party. Throughout the thesis, we focus on the two following key questions:

1. Do fundamental learning results still hold when the data generation process depends on the result of the analysis?

2. Can game theoretic models reasonably be applied in learning settings where there is a large quantity of available complex data?

We study the problems of adversarial classification and strategic linear regression through these questions. In particular, from a high level perspective, we show that fundamental learning results are challenged – optimal adversarial learning can be performed with simple classifiers and classical estimators resulting from linear regression are no longer optimal or even consistent when data are strategically produced.

This thesis is organized into three parts. Part I (Chapters 3, 4 and 5) is dedicated to the *adversarial classification* problem. Part II (Chapters 6, 7 and 8) presents our results on the *strategic linear regression* problem. Due to the differences between the two setting, we review in each part the corresponding literature. Finally, we draw our conclusions in Part III and discuss potential future work. We present a more detailed outline of this thesis and its contributions chapter by chapter in the following paragraphs.

Chapter 2 reviews technical concepts necessary for the results of this thesis. We introduce our learning problems (classification and linear regression) with their associated fundamental results. We define the VC-dimension for classification and the Generalized Least Squares estimator and its statistical properties. In particular, this estimator is consistent and is the Best Linear Unbiased Estimator (BLUE) in the sense that its covariance is minimum among linear unbiased estimators. We then present the game theoretic framework used to analyze these problems with the definition of a strategic (Bayesian) game, the notion of equilibrium and conditions guaranteeing the existence and the ease of computation of said equilibrium.

Chapter 3 introduces the adversarial classification game which is the primary focus of Part I. We first review literature related to adversarial classification, game theory

and the interface between these two fields. We then formulate our adversarial classification game. In this game a defender faces a mix of non-attackers who act according to a fixed probability distribution and attackers who choose how they act to maximize their utility. The defender is uncertain of the type of attackers they face but has a prior belief on the probability that attackers are of given types. The behaviors observed by the defender are translated into vectors of observable features for the classification task. The defender randomizes among all possible classifiers to minimize both the damage incurred by attackers and the false alarms due to the detection of non-attackers as attackers. In this setting, we study the Nash equilibrium of the game which corresponds to situations where both players act simultaneously. We show that the strategy of the defender can be expressed in a simpler way as a probability of detection function which describes the probability that each vector is detected. We conclude this chapter by a discussion on our assumptions and the settings on which our model is relevant to apply.

In Chapter 4, we characterize the equilibrium of the previously defined game. Our **first main contribution** is the characterization of the optimal defense strategy. We show that the defender optimally defends against attacks by deploying a strategy inducing a probability of detection function belonging to a specific parametrized class of functions. This result emphasizes the need for *randomization* in adversarial classification setting while showing that simple – but not necessarily classically used – classifiers lead to optimal defenses. More precisely, these probability of detection functions can be achieved with simple strategies randomizing uniformly against threshold classifiers where the thresholds are the previous parameters. We also prove that these parameters correspond to the minimizers of a piecewise linear convex function whose time required to evaluate at a given point depends on the size of the vector space. This vector space however may be large and the parameters necessary to evaluate this function (such as the probability distribution of non-attackers) may be hard to evaluate. Interestingly, the space of optimal probability of detection functions and threshold classifiers is not complex considering classical classification complexity definition; the former is of pseudo-dimension 1 and the latter of VC-dimension 1. We then characterize the strategy of the attacker as a best-response to the strategy of the defender. This gives some intuitive properties about the strategy of the attacker – they should attack with a behavior corresponding to relevant vectors with a probability which balances the false alarm loss of the defender and the loss due to attacks on this vector. We illustrate these results on a simple artificial game.

Chapter 5 focuses on efficient computation of the optimal defense parameters with limited information. To do so, we exploit our previous parametrization which shows

that finding approximately optimal parameters in our class of parametrized functions yields an approximately optimal defense. Our **second main contribution** provides an algorithm to compute approximately optimal parameters on an offline setting. We show that having access only to a labeled data set where the defender can evaluate the false alarm cost of each data point as well as the utility of attackers in case of detected or undetected attacks is sufficient to find approximately optimal parameters. The structure of our game even implies that the defender is able to find the optimal parameters with a probability growing potentially exponentially fast with the size of the data set. Importantly, we obtain approximation guarantees which *do not depend on the size of the vector set*. This is illustrated on both artificial games and a real bank fraud data set. We then focus on the online setting where the defender does not have access to historical data but must perform classification on incoming vectors and learn on the fly. We show that if the defender's objective is to minimize the *Stackelberg regret* (i.e. how much they lost compared to how much they would have lost when acting optimally considering that attackers adapt to their strategy), they can achieve low-regret (once again independent in the size of the vector set) by learning approximately optimal parameters. This is our **third main contribution**. We finally illustrate the online learning process on an artificial game with vectors consisting of a varying number of binary features.

Chapter 6 introduces our strategic linear regression game central to Part II. We first review existing work related to strategic linear regression. We then present our linear regression game. In this game, agents participate in a linear analysis by providing data points with a *chosen* precision. On one hand they benefit from a precise analysis in a *public-good* manner modeled by a cost depending on the covariance of the estimator obtained. On the other hand, producing precise data is costly either because of privacy concerns or because of computational challenges. This creates a trade-off between providing precise data to obtain a precise estimator and minimizing data provision cost.

In Chapter 7, we focus on *game-theoretic properties* of our linear regression game. In particular, we show that it is a potential game, reducing the problem of finding a Nash equilibrium to finding the minimum of a convex function. We then prove that our game is applicable even when agents have limited information about other agents when their data points are produced according to a common distribution. To do so, we show an equivalence between the complete information game where agents have complete knowledge of the data points of other agents and an average game where they are only aware of the common underlying distribution. We finally characterize the price of stability of the game – a measure of social efficiency which quantifies the impact of selfishness on social welfare by comparing the overall loss

of all agents to their overall loss if they had worked in coordination to minimize it. The price of stability exhibits a worst-case scenario when the cost of producing data increases linearly with the chosen precision.

Chapter 8 focuses on the *statistical properties* of the estimator resulting from the analysis. We show that the Generalized Least Squares estimator is no longer the Best Linear Unbiased Estimator (BLUE) but is approximately optimal among a class of estimators satisfying suitable statistical properties. This is our **fourth main contribution**. In our **fifth main contribution** we characterize the quality of the estimation at equilibrium from two perspectives. First, we show that agents with data provision costs linear in the precision lead to an optimal allocation of precision in the sense that the estimator obtained directly relates to what an optimal design would obtain. This is however not suitable in practice as we then show that in this case the Generalized Least Squares estimator is not even *consistent*. More precisely, we show that the presence of strategic agents always degrades the rate of convergence. 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. Additionally, 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).

Part III contains the conclusions of this work as well as potential future work. We defer some of our proofs and technical detail to the Appendix.

2 Background

We review in this chapter some important existing concepts which are necessary to introduce the results of this thesis. Section 2.1 introduces two ubiquitous machine learning problems and results quantifying the quality and complexity of learning in these problems. Section 2.2 presents some of the fundamental definitions of game theory with basic solution concepts. Note that this chapter exists for the sake of self sufficiency and is not a full introduction to the topics we present. We thus omit many fundamental results which are not necessary to the thesis. We however point the reader to relevant books in each section for in-depth introductions to the relevant topics.

2.1 Notions of Classification and Linear Regression

In this section, we introduce a subset of machine learning problems. We address the basics of classification and linear regression needed for the development of this thesis. Our notations for the general learning problem and definitions are drawn from Shalev-Shwartz and Ben-David [SB14] which we refer the reader to for a more in-depth introduction on machine learning. Our notations for the linear regression problem are inspired from Greene [Gre03] which contains all the results we present.

We aim to label objects drawn from a set $\mathcal{X}$ with labels from a label set $\mathcal{Y}$. To do so, we have access to training data: a finite sequence of labeled data points $((x_1, y_1), \ldots, (x_m, y_m))$ which we want to learn from. The learner wants to output a prediction $h : \mathcal{X} \to \mathcal{Y}$ from a set of *hypothesis class* $\mathcal{H}$ associating any possible input with a label. For this training to be possible we must define two major points: the data generation process and the objective of the learner. First, the data is generated by an unknown distribution D over $\mathcal{X} \times \mathcal{Y} = Z$. Then, we quantify the quality of an output of the learner through an average loss function $L_D(h) \equiv \mathbb{E}_{z\sim\mathcal{D}}\left[\ell(h, z)\right]$ where $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ is the loss of attributing label $h(x)$ to the data point $z = (x, y)$. Note that in many settings computing the average loss is not possible. To overcome this, it is often useful to consider the empirical risk $L_S(h) = \frac{\sum_{i=1}^{m} \ell(h,z_i)}{m}$. Its exact

properties however heavily depend on the setting and may be "optimal" such as with the least squares linear estimator or lead to overfitting of the model on the training data set.

Both classification and linear regression fit this general definition with different label sets $\mathcal{Y}$, hypothesis classes $\mathcal{H}$, data distributions $\mathcal{D}$ and loss functions ℓ. For this section, we consider that $\mathcal{X}$ is a subset of $\mathbb{R}^d$.

2.1.1 Classification

We first introduce some results on the binary classification problem where $\mathcal{Y} = \{0, 1\}$ and the loss of the learner is the $0 - 1$ loss $\ell(h, z) = \mathbb{1}_{h(x) \neq y}$. Note that what we present in this section can be generalized to multi-class classification where $\mathcal{Y} = \{1, \ldots, M\}$ but this is out of scope of this thesis. For this problem, we consider the agnostic probably approximately correct (PAC) framework defined as follows:

Definition 2.1. *A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon > 0$, $\delta \in (0, 1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples), we have:*

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon. \tag{2.1}$$

The notion of PAC learning patches an inherent difficulty of the learning framework we presented– the data generation process is random which implies that we can never truly be certain of our results. This definition allows us to prove that it is possible to approach the optimum hypothesis with certain probability and that improving either the quality of our solution or the probability that our solution is of the wanted quality can be done if we gather more training data. A very important notion for the PAC learning framework is the VC-dimension of a hypothesis class which quantifies the complexity of the learning task. This definition is based on the idea of shattering a set of points which determines if when the defender receives points $(c_1, \ldots, c_m)$ with arbitrary labels, it is always possible for them to correctly classify them using functions from their hypothesis class.

Definition 2.2 (Shattering). *Let $C = (c_1, \dots, c_m) \subset \mathcal{X}$ be a finite set. We say that $\mathcal{H}$ shatters C if for all assignment of labels $(y_1, \dots, y_m)$ to points of C, there exists $h \in \mathcal{H}$ such that $h(c_i) = y_i$ for all $i \in \{1, \dots, m\}$.*

Definition 2.3 (VC-dimension). *The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $VC\dim(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.*

The VC-dimension intuitively represents a notion of complexity of the hypothesis class considered and has a strong link with the notion of PAC learnable. In particular, the fundamental theorem of statistical learning shows that a hypothesis class $\mathcal{H}$ is PAC learnable if and only if it has finite VD-dimension and when it is PAC learnable, the number of sample points needed to obtain a δ-ϵ PAC approximation can be bounded by a function of δ, ϵ, the VC-dimension and constants and grows with the VC-dimension. This does however not mean that it is always suitable to select a low VC-dimension as this induces a bias in the solution. A simple hypothesis class may contain no suitable hypothesis for the model we want to learn. Indeed, while the VC-dimension bound guarantees that we will be able to find a hypothesis among the best in $\mathcal{H}$ easily, even the best hypothesis in our hypothesis class may fail to model what we want to learn in a satisfactory way. For example, it is trivial to "learn" the optimal hypothesis when the hypothesis class $\mathcal{H}$ is reduced to a single hypothesis but this does not exploit the information given by the training data. Such a trade-off is similar to the well known bias-variance trade-off in statistical analysis. We will use the notion of VC-dimension in Part I where one of the questions we address is what types of classifiers a defender should use when classifying in the presence of adversaries and how complex these classifiers should be.

2.1.2 Linear Regression

We then introduce the problem of linear regression. In this type of model, $\mathcal{D}$ defines a linear relation between the label $y \in \mathcal{Y} = \mathbb{R}$ and the "object" or regressor x: $y = \boldsymbol{\beta}^\top x + \epsilon$ where ϵ represents noise drawn from a distribution with null expected value and a given variance σ^2. The objective of the learner in this case is to minimize the squared error $\ell(h, z) = (h(x) - y)^2$. In this setting, the learner is restricted to the linear unbiased hypothesis class i.e. the class of hypothesis $\hat{\beta}$ such that there exists $L \in \mathbb{R}^{d \times n}$ such that $\hat{\beta} = L\boldsymbol{y}$ and which satisfies $\mathbb{E}\left[\hat{\beta}\right] = \boldsymbol{\beta}$. Biased estimators (such as the lasso) are also of interest in general but out of scope of this thesis as their theoretical properties are complex and hard to quantify in practice making them

challenging to incorporate in theoretical models. Note that the linear regression model also allows to model cases where y is an affine function of x by adding a fixed coordinate (often 1) to each vector.

In this section we refer to the hypothesis of the learner as an estimator to reflect the goal of estimating the linear relation β between the regressor and the label. When the learner minimizes the empirical risk associated to ℓ, the problem is solved in closed form with the ordinary least squares estimator (`OLS`) which enjoys strong theoretical properties (which are stated later in this section in a more general setting).

We are interested in a more general version of the linear regression problem where the data points $((x_1, y_1), \ldots, (x_m, y_m))$ of the training set are *not* identically distributed and we have instead $y_i = \beta^\top x_i + \epsilon_i$ where ϵ_i has null expected value and variance σ_i^2. For ease of notation, we will use the precision instead of the variance defined as $\lambda_i = 1/\sigma_i^2$. Assuming that the learner has access to the value of the variance σ_i^2, the previous empirical risk minimization is no longer suitable as the resulting `OLS` estimator is no longer optimal (in a sense precised later in Proposition 2.1).

The goal is then to find the unbiased estimator with "low" covariance $Cov(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (\hat{\boldsymbol{\beta}}_{\texttt{GLS}} - \boldsymbol{\beta})\right]$ which captures finely the uncertainty of the estimation of $\boldsymbol{\beta}$. To do so, we use the partial order on positive semi-definite matrices $\succeq$ where $A \succeq B$ if $A - B \in S_+^d$. This partial order is sufficient as the results state that there exists an estimator achieving minimum covariance among linear unbiased estimators. The solution to this problem is also known in closed form as the generalized least squares (`GLS`) estimator. To write this estimator clearly, let us first introduce some notations. Denote by $\boldsymbol{\lambda} = [\lambda_i]_{i \in N}$ the vector of precisions and by $\Lambda = \mathrm{diag}(\boldsymbol{\lambda})$ the diagonal matrix whose diagonal is given by vector $\boldsymbol{\lambda}$. $\boldsymbol{y} = [y_i]_{i \in N}$ is the n-dimensional vector of label variables, and $X = [\boldsymbol{x}_i^T]_{i \in N} \in \mathbb{R}^{n \times d}$ the $n \times d$ matrix whose rows comprise the transposed feature vectors. Then, the learner chooses hypothesis $h = \hat{\boldsymbol{\beta}}_{\texttt{GLS}}$ defined as follows:

$$\hat{\boldsymbol{\beta}}_{\texttt{GLS}} = (X^\top \Lambda X)^{-1} X^\top \Lambda \, \boldsymbol{y}. \tag{2.2}$$

This estimator has the following covariance $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\texttt{GLS}}) = (X^T \Lambda X)^{-1}$.

This estimator enjoys very strong properties with regards to the other linear unbiased estimators. In particular the following optimality result known as the Aitken (our Gauss-Markov when all variances are equal and the estimator is `OLS`) theorem holds:

Proposition 2.1 (Aitken [Ait35]). GLS *is the best linear unbiased estimator (BLUE) in the sense that its covariance is minimal in the semidefinite order among the covariances of all linear unbiased estimators.*

A simple proof of this theorem can be made by remarking that any linear unbiased estimator L can be written $L = (X^\top \Lambda X)^{-1} X^\top \Lambda + D$ with $DX = 0$. Then, computing the covariance of any estimator written this way shows that it is higher than the one of GLS. The GLS estimator is also *consistent* (i.e. it converges in probability to β when the number of data points grows) and is *asymptotically normally distributed* (i.e. we have $\sqrt{m}\left(\hat{\beta}_{\text{GLS}} - \beta\right) \xrightarrow{d} \mathcal{N}(0, Q^{*-1})$ where $Q^* = \operatorname{plim} \frac{1}{m} X^\top \Lambda X$). We will use these results in Part II where one of the questions we address is whether they still hold in a setting where data points are strategically produced.

Now that we have introduced the practical problems we study in this thesis, we will introduce the tools we use to analyze them in settings where data production depends on the algorithms used to analyze data. In particular, such interaction is well modeled using game-theory.

2.2 Notions of Game Theory

Game theory is a mathematical tool modeling situations where multiple agents (or players) make decisions whose outcome depends on the decisions of other agents. This field dates back to the 18th century with the analysis of a french card game ([Bel07]). Its most well-known application however dates from the 19th century with the Cournot duopoly and the field starts to formally exist on its own only with the work of Von Neumann [Von16]. Finally, the solution concepts we study in today's literature originated from Nash [Nas51] with the notion of equilibrium which then gave rise to a plethora of other solution concepts (such as Wardrop equilibrium of Wardrop [War52]) with different economical interpretations. In this section, we briefly introduce specific results needed for the thesis of this vast field. The notations and definitions are based on the book of Fudenberg and Tirole [FT91].

2.2.1 Strategic Game and Equilibrium

A game in strategic form is defined by the following elements:

- The set of players $i \in \mathcal{P} = \{1, \dots, I\}$.

- The pure strategy space S_i for each player i.

- The payoff functions $u_i(s)$ associating a strategy profile $s = (s_1, \dots, s_n)$ to a utility obtained by player i

We denote $\Gamma = (\mathcal{P}, (S_i)_{i\in\mathcal{P}}, (u_i)_{i\in\mathcal{P}})$ this strategic form. We say that players participate in a static complete information game if the following assumptions are verified: players are rational[1], have common knowledge[2] about all parameters of the game and choose their actions simultaneously and independently. Note that this definition is based on *utilities* i.e. what players gain. Authors sometimes define their games based on *costs* which is a strictly equivalent formulation (we can multiply utilities by -1 to obtain costs and vice-versa) and only changes the assumptions of concavity of utilities to convexity of costs when they are needed.

It is often useful to reason in terms of *best-response* of a player to the actions of the other players. We use the notation $-i$ to denote characteristics of every player except player i. For example, a strategy profile is often written as $s = (s_i, s_{-i})$. We then say that s_i^* is a best-response to the action of other players s_{-i} if $s_i^* \in \arg\max_{s_i\in S_i} u_i(s_i, s_{-i})$.

To obtain meaningful results on the interaction of players, we sometimes need to account for possible randomization among their actions. We thus extend the model to include mixed strategies which allows players to randomize among their pure strategies. We denote $\Sigma_i = \Delta(S_i)$ the set of probability distributions over S_i. This set is called the set of mixed strategies of player i and any $\sigma_i \in \Sigma_i$ is a mixed strategy. It is then straightforward to extend the definition of payoffs and best-responses to these mixed strategies.

Next, we present the notion of Nash equilibrium, one of the most important solution concepts in game theory. It defines a set of stable strategy-profiles relevant in many applications.

Definition 2.4 (Nash equilibrium)**.** *A mixed-strategy profile σ^* is a Nash equilibrium if, for all players i, we have:*

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(s_i, \sigma^* - i) \text{ for all } s_i \in S_i. \tag{2.3}$$

[1] See Osborne and Rubinstein [OR94] p.5 for an introduction on the definitions of Morgenstern and Von Neumann [MV53] and Savage [Sav72]

[2] While the notion of common knowledge can be grasped intuitively, its formal definition is complex. A great introduction to the topic is in Osborne and Rubinstein [OR94] p.67

A Nash equilibrium σ^* has the intuitive interpretation that no player has incentive to unilaterally deviate from the strategy profile σ^*. Another interpretation is that the strategies of each player i are best-responses to the strategies of each other player $-i$. This leads to the vision of a Nash equilibrium as a fixed-point of the best-response function which is often used to prove the existence of Nash equilibrium.

In general, a strategy profile achieving such a property is not guaranteed to exist and when it exists is not guaranteed to be unique. This is problematic as when several Nash equilibrium exist, if different agents act according to different equilibrium they can obtain an arbitrarily bad result and game theoretic assumptions prevent players from concerting over the choice of the equilibrium. Thus, game theoretic models often aim to show that there exists a unique equilibrium of the considered game or to show that all equilibrium are interchangeable. The next two results offer assumptions under which a Nash equilibrium always exists.

Theorem 2.1 (Nash [Nas51])**.** *Every finite strategic game has a mixed strategy Nash equilibrium.*

Theorem 2.2 (Debreu [Deb52], Glicksberg [Gli52], and Fan [Fan52])**.** *Consider a strategic-form game whose strategy spaces S_i are nonempty compact convex subsets of an Euclidean space. If the payoff functions u_i are continuous in s and quasi-concave in s_i, there exists a pure strategy Nash equilibrium.*

Note that many more results exist to guarantee that a game admits a Nash equilibrium under less different assumptions (such as non-continuous payoff) but are not necessary for the results of this thesis.

In many applications, however, all parameters are not necessarily common knowledge. In particular, it is frequent for players to have private information which influences the outcome of the game and can only be guessed by other players. To model this, we say that a player has a private type $\theta_i \in \Theta_i$ (we assume that Θ_i is finite in this thesis) and players type are drawn from a known distribution $p(\theta_1, \dots, \theta_I)$. We denote $p(\theta_i|\theta_{-i})$ the conditional probability of player i about the other players type. In this setting, the utility of each player depends on the type of all players and their action and is written as $u_i(s_1, \dots, s_I, \theta_1, \dots, \theta_I)$. The action of a player now depends on their type and we write $s_i(\theta_i)$ the action of player i when their type is θ_i. If player i knows their type θ_i and the strategies of other players $\sigma_{-i}(\cdot)$, they can optimize their expected gain using their prior $p(\theta_{-i}|\theta_i)$. This defines an "expanded" complete information game in which players optimize their gain

with respect to each of their possible types. We thus define the notion of Bayesian equilibrium through the notion of Nash equilibrium of this expanded game.

Definition 2.5 (Bayesian Nash equilibrium)**.** *A Bayesian Nash equilibrium in a game of incomplete information with a finite number of types θ_i for each player i, prior distribution p and pure strategy spaces S_i is a Nash equilibrium of the "expanded game" in which each player i's space of pure strategies is the set $S_i^{\Theta_i}$ of maps from Θ_i to S_i*

We will often use a more straightforward definition of Bayesian Nash equilibrium. It is easy to see that a mixed-strategy profile σ^* is a Bayesian Nash equilibrium if for all i and θ_i, we have:

$$\sigma_i^*(\theta_i) \in \arg \max_{\sigma_i(\theta_i) \in \Sigma_i} \sum_{\theta_{-i}} p(\theta_i|\theta_{-i}) u_i(\sigma_i(\theta_i), \sigma_{-i}^*(\theta_{-i}), (\theta_i, \theta_{-i})). \tag{2.4}$$

There exists other notions of equilibrium assuming different timing of actions or behavior of players. In particular, another natural equilibrium to consider in the classification setting is the Stackelberg equilibrium. In a Stackelberg game, a player acts first (the leader). All other players observe the action of the leader and act simultaneously. For simplicity's sake, we only present here the case of two-players Stackelberg games which are widely used in the literature. In Stackelberg games, the strategy of the follower depends on the strategy of the leader. We thus denote σ_1 the strategy of the leader and $f(\sigma_1)$ the strategy of the follower where $f : \Sigma_1 \to \Sigma_2$ represents the strategy of the follower as response to the strategy of the leader. The leader-follower dynamic naturally defines a notion of equilibrium similar to the notion of Nash equilibrium.

Definition 2.6 (Strong Stackelberg Equilibrium)**.** *Consider a Stackelberg game where player 1 is the leader and player 2 the follower. A pair of strategies $(\sigma_1^*, f(\sigma_1^*))$ is a Strong Stackelberg equilibrium if the following conditions hold:*

1. $u_1(\sigma^*, f(\sigma^*)) \geq u_1(\sigma, f(\sigma))$ *for all* $\sigma \in \Sigma_1$.
2. $u_2(\sigma_1, f(\sigma_1)) \geq u_1(\sigma_1, \sigma_2)$ *for all* $\sigma_1 \in \Sigma_1$ *and* $\sigma_2 \in \Sigma_2$.
3. $u_1(\sigma_1, f(\sigma_1)) \geq u_1(\sigma_1, \sigma_2)$ *for all* $\sigma_2 \in \arg\max_{\sigma_2'} u_1(\sigma_1, \sigma_2')$ *(The follower breaks ties in favor of the leader)*.

In this definition, we say that the follower breaks ties in favor of the leader. Other Stackelberg equilibrium exist such as the Weak Stackelberg equilibrium where the follower breaks ties unfavorably for the leader. In the type of classification settings

we consider however, the Strong Stackelberg equilibrium is the most interesting. Indeed, in classification it is natural to assume that the defender commits to their defense strategy and the attacker observes this strategy and acts. It is then possible for the defender to force the action of the attacker to be the most favorable outcome among their possible best responses (for example by defending slightly more against an unfavorable type of attack). Similarly to Nash equilibrium, every finite game admits a Strong Stackelberg equilibrium.

While the existence of (Bayesian) Nash equilibrium is well understood in the settings we consider, finding the equilibrium is in general a hard problem even for two player games (see Daskalakis et al. [DGP09] and Chen and Deng [CD06]). We see in the next section some specific types of games which have structural properties which makes computing their equilibrium less challenging.

2.2.2 Zero-Sum Game and Potential Game

The first type of game we introduce are the so-called *zero-sum* games where players compete and the gain of a player corresponds exactly to the loss of their adversaries.

Definition 2.7 (Zero-sum game)**.** *A game is called zero-sum if for all strategy profile s, we have:*

$$\sum_{i=1}^{I} u_i(s) = 0. \tag{2.5}$$

In general, it is challenging to compute the equilibrium of even zero-sum games (as long as the number of players is greater than two) as any two player game can be modeled as a three player zero-sum game where the payoff of the third player simply ensures the zero-sum property. We consider however in this thesis only the special case of two player zero-sum games. In this case, the gain of a player is exactly the loss of the other player. Combined with the best-response property of the equilibrium, it means that any of the two players should strive to maximize their minimum gain. This leads to players having incentives to find min-max strategies.

Definition 2.8 (min-max strategy)**.** *A min-max strategy for player i is a (mixed) strategy σ such that for all σ', we have:*

$$\min_{s_{-i}} u_i(\sigma_i, s_{-i}) \geq \min_{s_{-i}} u_i(\sigma', s_{-i}). \tag{2.6}$$

Two player zero-sum games reduce the problem of finding an equilibrium to finding a min-max of the utility function of each player. In particular, in the case of finite games, the equilibrium can be found through a linear program of size polynomial in the sizes of strategy spaces (this is no longer true even for 3 players zero-sum games). Another property of zero-sum game is that any combination of Nash equilibrium forms a Nash equilibrium thus bypassing the need for uniqueness of equilibrium. This type of game appears on Part I. Additionally, in two player zero-sum games, any Nash equilibrium is a Stackelberg equilibrium of the Stackelberg game where any of the two players is the leader and the other the follower.

Next, we define another type of game whose structure facilitates the computation of a Nash equilibrium: *potential* games.

Definition 2.9 (Potential game)**.** *A game Γ is a potential game if there exists a function $\phi : S \to \mathbb{R}$ such that $\forall i$, $\forall s_i \in S_i$, $\forall s_{-i} \in S_{-i}$, $\forall s'_i \in S_i$, we have:*

$$\phi(s_i, s_{-i}) - \phi(s'_i, s_{-i}) = u_i(s_i, s_{-i}) - u_i(s'_i, s_{-i}) \tag{2.7}$$

It is easy to see that any global minimum of the potential function is a Nash equilibrium. Additionally, if the potential function is strictly concave, there exists a unique Nash equilibrium which is the unique maximizer of ϕ. In these settings, the problem of finding an equilibrium is thus reduced to finding the maximum of a concave function which admits an extensive literature with practical algorithms. This type of game appears on Part II and appears otherwise in many different problems such as congestion games.

Part I

Adversarial classification

State of the Art and Model

3

In this part (Chapters 3, 4 and 5), we present our first example of learning with strategic, or in this case adversarial, data. More precisely, we tackle the problem of adversarial classification where a defender aims to classify incoming vectors from attacks and non-attacks and attackers aim to evade classification. We model this interaction with a game where the defender receives vectors from either attackers or non-attackers. Attackers choose the vector they use among a finite set of vectors $\mathcal{V}$ and non-attackers follow a fixed distribution on $\mathcal{V}$. The defender chooses a classifier without any a priori restriction on the set of possible classifiers. This raises two main issues. First, the number of possible classifiers is exponential in $|\mathcal{V}|$. Then, $\mathcal{V}$ itself is exponential in the number of features used to classify. Indeed, even considering the simplest case of k binary features leads to a vector set of size 2^k and a set of possible classifiers $2^{\mathcal{V}}$ of size 2^{2^k}. We focus on two key questions:

1. Which classifiers should the defender use at the equilibrium?

2. How can the defender compute optimal strategies in a scalable (in the number of features) manner?

We show that randomization is crucial to optimal defenses but, surprisingly, the defender can defend against attacks optimally using a class of classifiers with low complexity (of VC-dimension 1) using a parametrization depending on the gains of attackers. This parametrization in turns allows us to develop approximation methods which can generalize to unknown vectors both offline and online.

This first chapter reviews existing work broadly related to adversarial classification and to our model and introduces our adversarial classification game. In Chapter 4 we then characterize the equilibrium and show in particular that the optimal strategy of the defender can be parametrized with few parameters. This answers our first question. We then answer our second question in Chapter 5 where we show that our previous characterization can be used with approximate parameters to provide approximately optimal strategies both offline using an existing data set and online. Additionally, these methods require knowledge of only a few parameters of the game.

3.1 State of the Art

In this section, we review existing work relevant to this part organized as follows. The first 4 paragraphs review concepts broadly related to adversarial classification and learning in general. The last 2 paragraphs review game theoretic concepts related to our specific game and methods.

Adversarial learning: The literature on adversarial learning usually studies two types of attacks: 'poisoning attacks', where the attacker can alter the training set to tamper the classifier's training ([Dal+04; GR06; Bar+10; LL10; Hua+11; ZK14]); and 'evasion attacks', where the attacker tries to reverse engineer a fixed classifier to find a negative instance of minimal cost ([LM05; Nel+10; LV15]). This literature, however, does not fully model the attacker's adaptiveness, which often leads to an arms race. In recent years, the adversarial learning research focused on evasion attacks called adversarial examples that affect deep learning algorithms beyond attack detection applications ([GSS15; Pap+16; Pap+18]). These works, however, follow the same pattern.

Game-theoretic models of adversarial classification: A number of game-theoretic models of adversarial classification have been proposed, with various utility functions and hypotheses on the attacker's capabilities. Most of them, however, restrict a priori the possible classifiers: Zhou et al. [Zho+12] and Zhou and Kantarcioglu [ZK14] rely on kernel methods; Kantarcioglu et al. [KXC11] assumes that the defender uses a single type of classifier (though unspecified in the model); Dalvi et al. [Dal+04] focuses on naive Bayes classifiers (and only compute one-stage best responses); Brückner and Scheffer [BS11] and Brückner et al. [BKS12] constrain the classifier to a specific form and look for the (pure) equilibrium value of the parameters; Li and Vorobeychik [LV14] uses a different model but also restrict to linear classifiers; Dasgupta et al. [DCM20] restricts the defender to a set of adversarially trained classifiers of different strengths; Li and Vorobeychik [LV15] uses a more general classifier, but restricts for most results to a family of classifiers constructed on a given basis (their model of the attacker is also more constrained than ours); and Lisý et al. [LKP14] abstracts away the classifier through a ROC curve (attacker and defender only select thresholds). In contrast, the objective of our work is to derive the optimal form of the classifiers so we do not make any restriction a priori on the classifiers used. In a recent paper, Dritsoula et al. [DLM17] (see also Dritsoula et al. [DLM12] for a similar model specialized for network intrusion) propose a model

where the defender can select *any* classifier (i.e., function from the set of data to $\{0, 1\}$). A key difficulty lies in the exponential size of the resulting set of classifiers. The authors show that it is possible to restrict it to a small set of *threshold* classifiers on a function that appears in the attacker's payoff. The classifiers identified, however, have no parameter and their solution method is ad-hoc for the restrictive model chosen—with complete information and simplistic payoffs—, hence it cannot extend to more realistic scenarios. In realistic adversarial classification scenarios with uncertainty on the attacker's payoff, the leaves open the questions: *What classifiers should the defender use at equilibrium? And how to compute optimal classifiers in a scalable manner?*. Our work greatly generalized this work both from a modeling perspective and from the scope of our results which include practical algorithms (see Section 3.2.2).

At the exception of Lisý et al. [LKP14], Li and Vorobeychik [LV15], and Dritsoula et al. [DLM17], the aforementioned papers build deterministic classifiers while recent papers tend to advocate for randomization: Bulò et al. [Bul+16] introduces random strategies on top of Brückner et al. [BKS12] while Perdomo and Singer [PS19b] highlights the importance of randomized attacks and Pinot et al. [Pin+20] of randomized defenses (see also their Meunier et al. [Meu+21]). In our work, we completely characterize the equilibrium and naturally find that it must involve randomized attack and defense strategies.

It is important to understand that these works consider two main types of model. Brückner and Scheffer [BS11], Brückner et al. [BKS12], and Pinot et al. [Pin+20] study *adversarial learning* problems where the learning problem is defined even without attackers (e.g., image recognition), whereas Dalvi et al. [Dal+04], Lisý et al. [LKP14], Zhou et al. [Zho+12], Zhou and Kantarcioglu [ZK14], Kantarcioglu et al. [KXC11], and Li and Vorobeychik [LV14; LV15] study *adversarial classification* where the learning problem is to detect attacks and exists only because there are attackers (e.g., spam filtering). These models lead to different attack methods and defenses. Our work belongs to the second category, of adversarial classification problems.

Fundamental properties of adversarial learning: Some recent works generalize fundamental learning properties when facing an adversary. Cullina et al. [CBM18] extend PAC theory to adversarial settings and show that fundamental learning bounds can be extended to this setting and that the adversarial VC dimension can be either larger or smaller than the standard one. Bhagoji et al. [BCM19] show that there exists inherent lower bounds on the robustness a classifier can achieve in the presence of an attacker who can modify each vector to a neighbor vector

before the classification process. This robustness can be characterized by a distance between the classes the classifier should separate. Intuitively, the distance between classes depends on the number of "adversarially indistinguishable" pairs of vectors the attacker can make. It is noteworthy that in their setting the optimal defense is not randomized while their model is similar to Pinot et al. [Pin+20]. This comes from the combination of the fact that the model of Bhagoji et al. [BCM19] is closer to a Stackelberg setting where the defender is the leader and the fact that the attacker incurs no cost to modify a vector.

Interaction with other machine learning concepts: It is also important to mention that adversarial machine learning does not exist without machine learning and thus all concerns regarding machine learning algorithms are relevant to study in adversarial machine learning. In particular, there is a rapidly growing literature on fairness (Kusner et al. [Kus+17] and Jabbari et al. [Jab+17]) which aims to avoid discrimination with regards to sensitive attributes (with different notion of fairness) and differential privacy (Dwork and Lei [DL09] and Abadi et al. [Aba+16]) which aims to ensure that little information can be obtained about the training set by observing the machine learning model. The interaction between these notions and adversarial learning can be studied with two perspectives. The first one is how does the algorithm adapted for adversarial learning fare with regards to these? Such a question is studied for example in Milli et al. [Mil+19a] where strategic classification is shown to worsen disparities between sub-populations or in Phan et al. [Pha+19] where the generation of adversarial examples to train the model must be modified due to privacy concerns arising from the reuse of private training data for adversarial generation. The second one is the opposite – how does the mechanism to ensure wanted properties of the learning algorithm affect the adversarial setting? This question is studied for example in Ma et al. [MZH19] where differential privacy is used as a tool to prevent poisoning attacks as the output of the classifier does not depend too much on a single data point under differential privacy. On the contrary in Giraldo et al. [Gir+] differential privacy is shown to facilitate poisoning attacks in a setting where the privacy process happens before data is received by the learner and solutions specific to adversarial learning under differential privacy are studied. Such models, however, are recent and few.

Security games: Our game has similarities with *security resource allocation* games ([CL09; Kie+09; Boš+11; MTS12; FJT13; Bal+15; Sch+18; Bro+16]) used in applications such as airport security with the model of Pita et al. [Pit+09]. These works consider a defender with limited resources (guards, radars, etc.) to be

allocated to the defense of critical targets. In these settings, problems are at a relatively low scale and are usually entirely described via loss in case of attack of an undefended target. The challenge is then the management of the limited amount of resources, which produces NP-hard problems (see Korzhyk et al. [KCP10]) preventing these models to be transferred to very large scale settings. Our work studies a similar setting applied to classification, where targets would correspond to attack vectors in $\mathcal{V}$. In contrast to the security games literature, we do not impose limited resources (the defender self-restricts its detection to limit false alarm costs), which eliminates the combinatorial issue. We are then able to provide a very different characterization of the solutions with applicability to classification as well as to scale to very large sets $\mathcal{V}$, a problem that is never studied in classical security games.

Exponential zero-sum games: Our game reparametrization with 'randomized classifiers' to reduce the dimension of the set of classifiers from $2^{|\mathcal{V}|}$ to $|\mathcal{V}|$ borrows ideas classical in security games. This technique is also studied for more generic zero-sum games (see Immorlica et al. [Imm+11]); but with objectives and limitations similar to security games.

3.2 Model

We consider the following situation. A defender receives data examples that can be either attacks (class 1) or non-attacks (class 0) and wants to predict the class of incoming data. We assume that a data example is represented by a feature vector v that belongs to the same set $\mathcal{V}$ regardless of the class. This vector is typically a simplified representation of the actual attack/non-attack (e.g., spam/non-spam) in a feature space used to perform the classification. We assume that the probability that a data example is an attack, denoted p_a, is fixed.

Vectors corresponding to non-attacks follow a fixed probability distribution P_0 on $\mathcal{V}$ whereas vectors corresponding to attacks are generated by attackers. Attackers choose the vector they generate to maximize a utility function (see (3.1)) depending on the classification of the defender (this models adaptation to the defender's actions). To model the uncertainty of the defender, we assume that strategic attackers are endowed with a type $i \in [\![1, m]\!]$ that encodes their utility. The defender does not know the type of the attacker but holds a prior $(p_i)_{i\in[\![1,m]\!]}$ on the possible types.

The defender chooses a classifier in $\mathcal{C} = 2^{\mathcal{V}}$, that is a function mapping a vector to a predicted class. We assume that the defender maximizes a utility function that balances costs/gains in different cases as follows.

- A *false negative* incurs a loss $U_i^u(v)$ when facing a type-i attacker.
- A *true positive* incurs a gain $U_i^d(v)$ when facing a type-i attacker.
- A *false positive* incurs a false alarm cost $C_{\text{fa}}(v)$.
- A *true negative* incurs no cost.

We assume that the attacker's gain is the opposite of the defender's for false negatives and true positives (the classification outcomes involving the attacker).

Summarizing the above discussion, the utilities of the attacker and defender, when the attacker is of type i, are defined as follows:

$$\begin{aligned} U_i^A(v,c) =& U_i^u(v)\mathbb{1}_{c(v)=0} - U_i^d(v)\mathbb{1}_{c(v)=1}, \\ U_i^D(v,c) =& -p_a U_i^A(v,c) - (1-p_a)\sum_{v'\in\mathcal{V}} C_{\text{fa}}(v')P_0(v')\mathbb{1}_{c(v')=1}. \end{aligned} \tag{3.1}$$

We assume that $\mathcal{V}$ is finite and all functions of v are arbitrary. Our main result, however, extends to $\mathcal{V}$ compact assuming mild restrictions on the functions defining the payoffs (see Section 5.3).

The above primitives define a Bayesian game that we denote by $\mathcal{G}$. Note that we assume that all parameters of the game including p_a, P_0, and the utility functions (but not the attacker's type) are known to both players. (We will discuss later how to relax this assumption.) As we will see, in this game, equilibria exist only in mixed strategy (intuitively, both players have an incentive to be unpredictable). For the defender, a mixed strategy β is a probability distribution on $\mathcal{C}$. A mixed strategy of the attacker is a function $\alpha : [\![1, m]\!] \to \Delta(\mathcal{V})$ such that for all $i \in [\![1, m]\!]$, $\alpha_\cdot^i$ is a probability distribution over $\mathcal{V}$ chosen by a type-i attacker.

The defender's utility depends on the attacker they face. With the belief the defender holds on the probability of each attacker type, it is natural that the defender tries to maximize their average utility. The equilibrium is also described with the average utility of the different attacker types, but as the actions of different attacker types are unrelated it is equivalent to each type maximizing its own utility.

Finally, for all $i \in [\![1, m]\!]$, we define

$$\underline{G}_i = \max_{v \in \mathcal{V}} \left(-U_i^d(v)\right) \text{ and } \overline{G}_i = \max_{v \in \mathcal{V}} \left(U_i^u(v)\right),$$

which respectively represent the minimum possible gain of the attacker (even if all vectors are always detected they can gain this quantity) and their maximum possible gain.

3.2.1 Preliminary: reduction of dimensionality

The first difficulty of the model we study is the exponential size of $\mathcal{C}$ in $\mathcal{V}$. This issue is commonly found in resource allocation games (similar reparametrizations are found in other games such as dueling algorithms) and circumvented through the use of a probability of allocation function: only the probability that an abstract resource is allocated to a target is considered thus ignoring the actual allocation and removing combinatorial complexity (assuming that one can compute this function at equilibrium). In our case, in the spirit of Dritsoula et al. [DLM17], we define a probability of detection π, for any strategy β of the defender, as

$$\pi^\beta(v) = \sum_{c \in \mathcal{C}} \beta_c \mathbb{1}_{c(v)=1}.$$

This transformation exploits the fact that, as long as a vector is detected, the actual classifier used for the detection is not important. Thus, with this probability of detection function, we can rewrite the payoffs independently of classifiers:

$$\begin{aligned} U_i^A(\alpha, \beta) &= \sum_{v \in \mathcal{V}} \alpha_v^i \left[U_i^u(v) - \pi^\beta(v) \cdot \left(U_i^u(v) + U_i^d(v)\right)\right]; \\ U_i^D(\alpha, \beta) &= -p_a U_i^A(\alpha, \beta) - (1 - p_a) \sum_{v \in \mathcal{V}} C_{\text{fa}}(v) P_0(v) \pi^\beta(v). \end{aligned} \tag{3.2}$$

Any probability of detection function can be attained through simple threshold classifiers crafted for this function. To see this, consider the set of threshold classifiers $c(v) = \mathbb{1}_{\pi^\beta(v) \geq t}$ for some $t \in [0, 1]$. Then, picking a random threshold uniformly on $[0, 1]$ defines a strategy achieving detection probability $\pi^\beta(\cdot)$. Note that this is the main difference between our work resource allocation games in which computing a strategy achieving an allocation probability is generally NP-hard.

3.2.2 Model discussion

The main motivating scenarios for our model are detection of malicious behaviors such as spam (in emails, social media, etc.), fraud (e.g., bank or click fraud), or illegal intrusion. In such scenarios, the attacker is the spammer, fraudster or intruder while the non-attacker represents a normal user (e.g., non-spam message). The vector v is a representation of the observed behavior on which the classification is done. For spam filtering, it can be a simplified representation of the messages obtained by extracting features such as the number of characteristic words. The distribution P_0 represents the distribution over those features for normal messages (not chosen with any adversarial objective). In our basic model, we assume that it is known by both players. It is reasonable in applications where it can be estimated from observation of a large number of easily obtainable messages (e.g., in social media they are public). We relax it in Chapter 5 where we show that the defender can learn well without a priori knowledge of P_0, p_a and p_i.

In our model the defender is uncertain of its own utility as soon as they have uncertainty regarding the attacker they face. Although not the most classical setting, it is meaningful and well studied in Bayesian games (see Forges [For92])—recall that the defender maximizes an expectation of this utility at the BNE. It is well justified in our case. For instance, if a fraudster manages to get access to sensitive information or to an account, the amount of harm may differ depending on the skills and resources of the fraudster.

The interaction between classifier and attacker is often modeled as a Stackelberg game where the attacker observes and reacts to the defender's strategy. We focus on the (Bayesian) Nash equilibrium which makes sense if the attacker cannot have perfect information about the defender's strategy. More generally though, we will see that in our game the defender's strategy at BNE must be min-max; hence, any strategy of the defender in a strong Stackelberg equilibrium would have the same property. We use the Stackelberg model in the online setting where there would be a bigger difference. Note that the fact that the defender seeks min-max strategies also yields robustness.

Our payoff function generalizes that of Dritsoula et al. [DLM17] in a practically important way. In their model, a reward $R(v)$ is granted to an attack with vector v regardless of the outcome and a fixed detection cost c_{d} is paid if the attack is detected. This is unreasonable in many applications such as bank fraud. In contrast, here, we allow the utility in case of detected and undetected attacks to be arbitrary unrelated functions of v (which would be equivalent to letting the detection cost c_d

depend on v). This key generalization alone breaks the ad-hoc method of Dritsoula et al. [DLM17] to compute the equilibrium. On top of it, we also generalize to a Bayesian game (of which the complete information game is a particular case where $m = 1$), and consider training and online learning problems of practical importance.

3.2.3 Example of Games

In this section we present the 4 different games we use to illustrate our results. Games 1, 3, and 4 rely on artificial distributions where we define and know every parameter of the game. Game 2 is defined with a real bank fraud data set where P_0 and p_a are unknown. Game 3 is an artificial version of Game 2 where we know every parameter to estimate the efficiency of our training methods on the real data set.

The One Feature Game

Game 1 is a game with two possible types of attacker in which classification is based on a single feature. The attackers' strategy space consists of 101 attack $v_0, \cdots, v_{100}$. For Attacker 1, an undetected attack yields a utility $U_1^u(v_r) = r$. A detected attack incurs a cost $U_1^d(v_r) = 30 * (r \mod (10))$, (see Figure 4.1b). Attacker 2, whose strategy space is the same as Attacker 1, has $U_2^u(v_r) = 100 - r$ and bears a cost $U_2^d(v_r) = 300 - 30(r \mod (10))$ in case of detected attack. Their gain and cost functions mirror that of Attacker 1, being interested in low vectors while Attacker 1 is interested in high vectors. Hence, the defender faces two attackers with different interests. There is a proportion $p_a = 0.2$ of attackers. The defender bears a constant false alarm cost $C_{\text{fa}} = 140$. A non-attacker follows a binomial distribution, they play the vector v_r with probability $P_0(v_r) = \binom{100}{r}\theta_0^r(1-\theta_0)^{100-r}$ with $\theta_0 = 0.2$.

The Bank Fraud Data Set

Game 2 is a bank fraud setting where we do not know every parameter but have access to a labeled data set (In our case we use the data set of ULB [ULB]).

The dataset we consider ([ULB]) contains transactions made by European cardholders in September 2013. A data vector is composed of 31 features: the amount of the transaction (in €) denoted A, the time since the first transaction in the dataset, whether the transaction was malicious (i.e., the label), and 28 anonymized features

coming from a PCA (Principal Component Analysis). We instantiate our game with this static data set by replacing each attack in the data set by an abstract adaptative attack in our model. For simplicity, we focus only on the amount of the transaction and consider a single attacker type with the following gains: $U^u(v) = A$, $U^d(v) = 0$, and $C_{\text{fa}}(v) = \ell \times A$ for a given $\ell > 0$. This models an attacker that gains the transaction's amount if successful (and the bank loses it), but gains nothing if detected. On the other hand, when a valid transaction is blocked, the bank pays a fraction ℓ of the transaction as a false alarm cost. This choice of utility functions is meant to illustrate the equilibrium in a reasonable and simple scenario and not to represent a practical ready-to-implement setting. In the dataset, the fraction of attacks is $p_a = 0.00172$, the maximum transaction is $25,691.16$€ with an average of 88.35€. There are $N = 284,807$ transactions in total.

The Artificial Bank Fraud Game

Game 3 is a game similar to Game 2, but with features following an artificial (controlled) distribution. We consider vectors v of the form $v = A$ where A is the amount of the transaction discretized on integer amounts. We define as in Section 5.1.1, $U^u(v) = A$, $U^d(v) = 0$ and $C_{\text{fa}}(v) = \ell A$ for some $\ell > 0$. We prescribe the following non-attacker distribution: $P_0(v) = p(A)$ where the amount of the transaction of a user follows a binomial distribution between 0 and 25691 with a mean of 88.

The Binary Features Game

Game 4 aims to illustrate learning where there is no correlation between vectors and costs as well as the presence of multiple attackers. It is a game with four types of attackers with vectors of k binary features, for k up to 19 to be able to compute the exact optimum for comparison. For each attacker type i and vector v, $U_i^u(v)$, $U_i^d(v)$ and $C_{\text{fa}}(v)$ are assigned a random value uniformly between 10 and 20. We set $p_a = 0.1$ and generate p_i's randomly. The non-attacker distribution P_0 is drawn randomly and uniformly on the $|\mathcal{V}|$-dimensional unit simplex.

4 Equilibrium Characterization

In this chapter, we fully characterize the equilibrium of the adversarial classification game. In Section 4.1, we characterize the optimal *randomized* defense strategy. We exhibit a class of probability of detection function which can be achieved by threshold classifiers and which are sufficient to optimally defend against attacks. In particular, the probability of detection functions we consider can be expressed with few parameters and are of low pseudo-dimension, paving the way for our methods to compute optimal defense in a scalable fashion in Chapter 5. In Section 4.2, we characterize the optimal attack strategy in response to a defense strategy. Our result shows that the attacker's strategy should strike a simple balance between the false alarm costs the defender should endure to detect them and their gain. Finally, we illustrate the shape of the equilibrium in Section 4.3 on the simple Game 1.

4.1 Optimal Defense Classifiers Are Threshold Classifiers

In this section, we first characterize the optimal defense strategy of the defender. We show that it can be expressed with few parameters representing the gain of each type of attacker. We use this property and characterize the complexity of the optimal defense. Surprisingly, this complexity is low as we show that the class of optimal probability of detection functions has pseudo-dimension 1 and the class of optimal classifiers has VC-dimension 1.

Finding a Bayesian Nash equilibrium is often hard in general games. We thus first reduce this problem to finding a min-max problem on the probability of detection. To do so, we compile the key property that our game is essentially a zero-sum game with the action space reduction via the probability of detection.

Using the payoffs defined in (3.1), we can do the following transformation of payoffs without changing both players' equilibrium strategies. First, scale the payoff of the defender by a factor $1/p_a$. Then, add the false alarm term to the payoff of the attacker (this term is independent from the action of the attacker so does not change

their strategy). It is then easy to see that this defines an "average" zero-sum game where for all $i \in [\![1, m]\!]$, we have $U_i^A(\alpha, \beta) = -U_i^D(\alpha, \beta)$. Using the definition of a BNE, this implies that at equilibrium they maximize their minimum average gain and gives the following lemma (whose proof can be found in Appendix A.1):

Lemma 4.1. *Let (α^*, β^*) be a BNE. Then*

$$\beta^* \in \arg\max_{\beta} \min_{\alpha} \sum_i p_i U_i^D(\alpha, \beta). \tag{4.1}$$

Computing the min-max strategies of Lemma 4.1 can be done via a classical transformation to a linear program, but this "naive" program would be of size exponential in $|\mathcal{V}|$. Even by expressing it in terms of π^β, the program would remain of size $|\mathcal{V}|$, which may be too large. Instead, we will leverage the min-max property to show that the equilibrium can be described compactly using a small number of parameters $\boldsymbol{G} = (G_1, \cdots, G_m)$ that can be interpreted as the utility of the attacker for each type. Formally, we define:

Definition 4.1 (Optimal probability of detection)**.** *For any $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \ldots \times [\underline{G}_m, \overline{G}_m]$, let*

$$\pi_{\boldsymbol{G}}(v) = \max\left\{0, \max_i \left\{\frac{U_i^u(v) - G_i}{U_i^u(v) + U_i^d(v)}\right\}\right\}, \ \forall v \in \mathcal{V}. \tag{4.2}$$

As we will see, this quantity is the unique probability of detection that guarantees attacker utility below $\boldsymbol{G}$ while minimizing the false alarms, so it plays a key role in the BNE strategy. In particular, it allows us to express the strategy of the defender as the maximum of a concave function of $\boldsymbol{G}$:

Definition 4.2 (Minimum gain function U^D)**.** *For all $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \ldots \times [\underline{G}_m, \overline{G}_m]$, let*

$$U^D(\boldsymbol{G}) = -p_a \sum_i p_i G_i - (1 - p_a) \sum_{v \in \mathcal{V}} C_{fa}(v) P_0(v) \pi_{\boldsymbol{G}}(v).$$

This function represents the minimum utility of the defender assuming they use a probability of detection function $\pi_{\boldsymbol{G}}(\cdot)$ for some $\boldsymbol{G}$. It allows us to state our parametrization result which is the main tool we use to prove all our core results.

Proposition 4.1. *For any*

$$\boldsymbol{G}_{\max} \in \underset{\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \ldots \times [\underline{G}_m, \overline{G}_m]}{\arg\max} (U^D(\boldsymbol{G})),$$

any strategy of the defender that yields a probability of detection function $\pi(v) = \pi_{\boldsymbol{G}_{\max}}(v)$ *for all* $v \in \mathcal{V}$ *is a min-max strategy and* $\max_\beta \min_\alpha \sum_i p_i U_i^D(\alpha, \beta) = U^D(\boldsymbol{G}_{\max})$.

Sketch of proof. A proof of Proposition 4.1 can be found in Appendix A.2. The proof relies on the min-max property of the problem which implies that the defender must maximize their minimum gain. We show that for a given utility profile $\boldsymbol{G}$, the minimum gain of the defender as defined in (4.1) is at least $U^D(\boldsymbol{G})$. However, the key difficulty is that not all utility profiles $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times ... \times [\underline{G}_m, \overline{G}_m]$ are feasible and the set of feasible utility profiles needs not be convex due to our Bayesian game and arbitrary functions; hence $U^D(\boldsymbol{G})$ could be meaningless. Our proof bypasses this difficulty by showing that $\pi_{\boldsymbol{G}_{\max}}(\cdot)$ is a min-max strategy in any case and shows as a corollary that $\boldsymbol{G}_{\max}$ is a feasible utility profile. □

Proposition 4.1 essentially states that in order to find the equilibrium strategy, the defender should only find m parameters $(G_1, \cdots, G_m)$, corresponding to the maximum utility that they should let the attacker gain for each type. Then, from those parameters, the probability of detection function is naturally defined. This result has multiple consequences.

First, from this characterization we deduce that one does not need to know all the parameters of the problem to find a good strategy. Finding "good enough" parameters for the utility of the different attacker types allows the defender to fully define its strategy. This is the main tool allowing us to define strategies which can generalize to unknown vectors in Chapter 5. In particular, in Theorem 5.1 we prove that near-optimal (and even optimal with high probability) classifiers can be computed by training the model on a labeled dataset with very limited information. Note that this is a key difference between our work and security games where the probability of allocation is computed directly using a linear program. There, the lack of a simple expression for the allocation probability prevents the definition of strategies that can generalize. It is also worth noting that unlike linear programs, our method can be generalized to a continuous vector set—we refer to Section 5.3 for details about that.

Second, the result from Proposition 4.1 shows that the presence of strategic adversaries *simplifies* learning in our problem. Indeed, the class of real valued functions $\{\pi_{\boldsymbol{G}}\}$ which contains the optimal strategy is of low pseudo-dimension (e.g., if there exist v_1 (resp. v_0) of class 1 (resp 0) with $U^u(v_0) > U^u(v_1)$ and $U^d(v_0) < U^d(v_1)$, these two points cannot be shattered). This can be explained by the predictable aspect of adversaries acting according to their best-response. On the contrary, when

facing non-strategic adversaries the optimal strategy would be a cost-sensitive adaptation of the naive Bayes classifier, which can potentially be any arbitrary function of $2^{\mathcal{V}}$ (especially since we make no assumption on P_0). This can be explained by the predictable aspect of adversaries. As they act according to their best-response, the defender knows what to expect. On the contrary, when facing non-strategic adversaries the optimal strategy would be a cost-sensitive adaptation of the naive Bayes classifier. Such a classifier can potentially be any arbitrary function of $2^{\mathcal{V}}$, especially since we make no assumption on P_0. This simplification is also shown in Theorem 5.1 where we show that there exists training methods which are optimal with high probability. This is noteworthy as such a possibility was hinted at by Cullina et al. [CBM18] who show that, for adversaries who can modify vectors in some neighborhood, the adversarial VC dimension can be either lower or higher than the standard one—i.e., the complexity can either increase or decrease in the presence of adversaries. In our adversarial classification model, the complexity drastically decreases. This also suggests that classifiers relying on simply adapting classical training might be inefficient as they do not take into account the fundamental complexity differences between classical and adversarial learning.

With Proposition 4.1 describing the probability of detection function at equilibrium, we can deduce a characterization in terms of threshold classifiers.

Definition 4.3 (Generalized threshold classifiers)**.** *For all $\boldsymbol{G} \in \mathbb{R}^m$, we define*

$$\mathcal{C}^T_{\boldsymbol{G}} = \{c \in \mathcal{C} : c(v) = \mathbb{1}_{\pi_{\boldsymbol{G}}(v) \geq t}, \forall v \in \mathcal{V} \text{ for some } t \in [0,1]\}.$$

Theorem 4.1. *There exists $\boldsymbol{G} \in \mathbb{R}^m$ such that the defender can achieve equilibrium payoff using only classifiers from $\mathcal{C}^T_{\boldsymbol{G}}$.*

This theorem settles our first main question: "which classifiers should the defender use at the equilibrium?". These are threshold classifiers on a non-standard function with threshold t representing a probability of detection. A threshold t can be interpreted as classifying as an attack if, even when being detected with probability t, at least one type of attacker gains at least G_i on average. Interestingly, $\mathcal{C}^T_{\boldsymbol{G}}$ has a VC dimension of only 1 as the set composed of v_1 (resp. v_0) of class 1 (resp. 0) with $\pi_{\boldsymbol{G}}(v_1) < \pi_{\boldsymbol{G}}(v_0)$ cannot be shattered. This strengthens our previous remark on the complexity of adversarial classification. Efficient randomized classification for adversarial settings does not require high capacity classifiers but rather classifiers tailored to the players payoffs. Then, our threshold classifiers may be linear classifier if payoffs are linear as the condition $\pi_{\boldsymbol{G}}(v) \geq t$ can be rewritten as

$\max_i \left\{U_i^u(v) - G_i - t(U_i^u(v) + U_i^d(v))\right\} \geq 0$. This means that in the linear setting, our threshold classifiers correspond to the defender picking a linear classifier for each type of attacker and outputting class 1 if at least one of the linear classifiers outputs it. In general settings however, linear classifiers may perform sub optimally.

The fact that the defender uses specifically threshold classifiers is noteworthy as there is already a literature on the choice of threshold and on this choice in an adversarial setting as in Lisý et al. [LKP14]. However, the random choice of the threshold in our setting is surprisingly simple. By construction it is a threshold on the probability of detection and choosing a threshold uniformly over $[0, 1]$ gives the desired strategy. This emphasizes that randomization is necessary to defend against an adversary but also that the choice of the set of classifiers to use is crucial to obtain good results.

Finally, let us notice that the equilibrium characterization naively leads to a linear programming solution polynomial in $|\mathcal{V}|$ to compute an exact equilibrium—simply by observing that function U^D is piecewise linear. The result is presented in Proposition 4.2; note that a fairly similar program could be obtained without our equilibrium characterization. We give in Section 4.2 a linear program that allows computing the attacker's strategy in time polynomial in $|\mathcal{V}|$.

Proposition 4.2. *Maximizing $U^D(\boldsymbol{G})$ is equivalent to solving the following linear program:*

$$\begin{aligned} \underset{\pi,\boldsymbol{G}}{\textit{maximize}} \quad & -p_a \sum_{i=1}^{m} p_i G_i - (1-p_a) \sum_{v \in \mathcal{V}} C_{fa}(v) P_0(v) \pi_v \\ \textit{subject to:} \quad & G_i \geq U_i^u(v) - \pi_v (U_i^u(v) + U_i^d(v)), \forall i, \forall v \\ & \pi_v \leq 1, \forall v. \end{aligned}$$

4.1.1 The Complete Information Case: Insights on the Value of $G_{\max}$

While Theorem 4.1 characterizes the shape of the equilibrium, it does not give any insight on the choice of the optimal parameter $G_{\max}$. In the complete information case, however, we can state more precise results allowing us to have a better grasp on the strategy of the defender by characterizing precisely the minimum of $U^D(\cdot)$.

Function $U^D(\cdot)$ is piecewise linear but it is also concave and we can compute its super-gradient. for all $G \in]\underline{G}, \overline{G}[$ such that $G = U^u(v)$ for some $v \in \mathcal{V}$, we have:

$$\partial U^D(G) = [\, -p_a + (1-p_a) \sum_{v', U^u(v') > G} C_{\text{fa}}(v') \frac{P_0(v')}{U^u(v') + U^d(v')},$$
$$-p_a + (1-p_a) \sum_{v', U^u(v') \geq G} C_{\text{fa}}(v') \frac{P_0(v')}{U^u(v') + U^d(v')}].$$

We otherwise have, for all $G \in]\underline{G}, \overline{G}[$ such that $G \neq U^u(v)$ for all $v \in \mathcal{V}$:

$$\partial U^D(G) = \{-p_a + (1-p_a)v', \sum_{U^u(v') \geq G} C_{\text{fa}}(v') \frac{P_0(v')}{U^u(v') + U^d(v')}\}. \tag{4.3}$$

This is proved by observing that $U^D(\cdot)$ can be expressed as a minimum of functions $U^D_v(G) = -Gp_a - (1-p_a) \sum_{v', U^u(v') \geq U^u(v)} C_{\text{fa}}(v') \frac{U^u(v') - G}{U^u(v') + U^d(v')} P_0(v')$.

As $U^D(\cdot)$ is a concave function, its maximum is attained either at one end of the definition's interval (when the derivative is negative at $\underline{G}$ or positive at $\overline{G}$) or when 0 belongs to the super-gradient of the function. This allows us to gather some insights on the choice of $G_{\max}$ by the defender. Intuitively, the previous equations mean that the defender chooses the gain of the attacker to strike a balance between the probability that they face an attack p_a and the increase in risk associated to the detection of non-attackers $(1-p_a) \sum_{v', U^u(v') > G} C_{\text{fa}}(v') \frac{P_0(v')}{U^u(v') + U^d(v')}$.

This piecewise linearity of $U^D(\cdot)$ also trivially leads to the following proposition.

Proposition 4.3. *The maximum of the function $U^D(.)$ will always be attained at $G_{max} = \underline{G}$ or at $G_{max} = U^u(v)$ for some v. Thus, at the equilibrium there is always a vector that yields the maximum reward of the attacker that is either never detected or always detected.*

4.2 Attacking in Response to Optimal Defenses: Balancing the Defender's Risk

In this section, we characterize the strategy of the attacker as response to an optimal strategy of the defender. Our characterization cannot be used for an efficient computation of the attacker's equilibrium strategy as we do not believe a low-dimensional parametrization of their strategy to be possible and we instead focus on understanding the probability that attackers use different attacks.

First, as mentioned in Lemma 4.1, the game is equivalent to a zero-sum game meaning that it would be easy to compute the strategy of the attacker through a min-max linear program. This, however, only uses the basic properties of the game and does not yield any more insight. We note here that while Proposition 4.2 allows us to compute the strategy of the defender through a linear program, the dual of this linear program does not give the attacker's equilibrium strategy. Indeed, the dual is the following:

$$
\begin{array}{lll}
\underset{\alpha_v^i, \pi_v}{\text{minimize}} & \displaystyle\sum_{v \in \mathcal{V}} \sum_{i \in [\![1,m]\!]} \alpha_v^i U_i^u(v) + \sum_{v \in \mathcal{V}} \pi_v & \\
\text{subject to:} & \pi_v \geq 0, & \forall v \in \mathcal{V} \\
& \alpha_v^i \leq 0, & \forall i \in [\![1,m]\!], \forall v \in \mathcal{V} \\
& \sum_{v \in \mathcal{V}} \alpha_v^i = -p_i, & \forall i \in [\![1,m]\!] \\
& \sum_{i \in [\![1,m]\!]} \alpha_v^i (U_i^u(v) + U_i^d(v)) + \pi_v \geq -\frac{1-p_a}{p_a} C_{\text{fa}}(v) P_0(v), & \forall v \in \mathcal{V}
\end{array}
$$

The second and third constraints could make sense if we considered the variables $-\alpha_v^i/p_i$, but the first and fourth constraints do not correspond to the problem. Indeed, with these constraints π_v is unrestricted so it does not necessarily correspond to a probability of detection function. Additionally, α_v^i does not fit the characterization of the strategy of the attacker at equilibrium given in Lemma 4.2 below. While it may seem counter-intuitive that the dual of the linear program giving the min-max strategy of the defender does not output the min-max strategy of the attacker, recall that the min-max strategy of the defender was not computed with the standard linear program for min-max problems but through a linear program computing the maximum of a piecewise-linear function.

Instead of characterizing the strategy of the attackers at equilibrium as a standalone min-max linear program, we view it as a response to the optimal strategy of the defender. This yields the following characterization of the attacker's strategy at equilibrium:

Lemma 4.2. *Let (α^*, β^*) be a BNE of $\mathcal{G}$, then:*

$$
\begin{aligned}
&\forall v \in \mathcal{V} \text{ s.t. } 0 < \pi^{\beta^*}(v) < 1 : && p_a \sum_i p_i \alpha^{*i}_v (U_i^u(v) + U_i^d(v)) = (1-p_a) C_{fa}(v) P_0(v), \\
&\forall v \in \mathcal{V} \text{ s.t. } \pi^{\beta^*}(v) = 0 : && p_a \sum_i p_i \alpha^{*i}_v (U_i^u(v) + U_i^d(v)) \leq (1-p_a) C_{fa}(v) P_0(v), \\
&\forall v \in \mathcal{V} \text{ s.t. } \pi^{\beta^*}(v) = 1 \text{ and } v \in \mathcal{V} : && p_a \sum_i p_i \alpha^{*i}_v (U_i^u(v) + U_i^d(v)) \geq (1-p_a) C_{fa}(v) P_0(v).
\end{aligned}
$$

Intuitively, Lemma 4.2 states that the attacker's strategy strikes a balance between the risk $(1-p_a)C_{\text{fa}}(v)P_0(v)$ the defender takes to detect a vector v and the average gain associated with the detection of vectors $p_a \sum_i p_i \alpha_v^i (U_i^u(v) + U_i^d(v))$. The proof of this lemma can be found in Appendix A.4. It allows us to find the best-response of the attacker α^* to a min-max strategy β^* of the defender, hence allowing us to find a BNE (though we emphasize that finding the strategy of the attacker is not the focus of this section).

Proposition 4.4. *Let $\beta^* \in \arg\max_\beta \min_\alpha \sum_i p_i U_i^D(\alpha, \beta)$. Then there exists a solution to the linear program $\text{find}_\alpha(\beta^*)$:*

$$
\begin{array}{lll}
\underset{\alpha_v^i}{max} & 0 & \\
s.t.: & 0 \leq \quad \alpha_v^i \leq 1, \forall i \in [\![1,m]\!], \forall v \in \mathcal{V} & , \\
& \sum_{v\in\mathcal{V}} \alpha_v^i = 1, \forall i \in [\![1,m]\!] & , \\
& \qquad \alpha_v^i = 0, \forall i \in [\![1,m]\!], \forall v \in \mathcal{V}, s.t. : U_i^u(v) - (U_i^u(v) + U_i^d(v))\pi^{\beta^*}(v) < G_i & , \\
& \sum_i p_i\, \alpha_v^i (U_i^u(v) + U_i^d(v)) = \frac{1-p_a}{p_a} C_{fa}(v) P_0(v), \forall v \in \mathcal{V}, s.t. : \pi^{\beta^*}(v) \in (0,1) & , \\
& \sum_i p_i\, \alpha_v^i (U_i^u(v) + U_i^d(v)) \leq \frac{1-p_a}{p_a} C_{fa}(v) P_0(v), \forall v \in \mathcal{V}, s.t. : \pi^{\beta^*}(v) = 0 & , \\
& \sum_i p_i\, \alpha_v^i (U_i^u(v) + U_i^d(v)) \geq \frac{1-p_a}{p_a} C_{fa}(v) P_0(v), \forall v \in \mathcal{V}, s.t. : \pi^{\beta^*}(v) = 1 & ,
\end{array}
$$

where $G_i = \max_v (U_i^u(v) - (U_i^u(v) + U_i^d(v))\pi^{\beta^*}(v))$. Additionally, for any solution α^* of $\text{find}_\alpha(\beta^*)$, (α^*, β^*) is a BNE.

Proof. Let β^* be a min-max strategy for the defender. Then, from the proof of Lemma 4.1, for any $\alpha^* \in \arg\max_\alpha \min_\beta -\sum_i p_i U_i^D(\alpha, \beta)$, (α^*, β^*) is a BNE. Thus, it satisfied the conditions of Lemma 4.2; using those, it is trivial to check that α^* is a solution of $\text{find}_\alpha(\beta^*)$. So this linear program admits a solution.

Next, any solution α^* of $\text{find}_\alpha(\beta^*)$ satisfies the conditions of Lemma 4.2, which implies that

$$\beta^* \in \arg\max_\beta \sum_i p_i U_i^D(\alpha^*, \beta). \tag{4.4}$$

Additionally, from the third constraint of the linear program, we observe that for any given i, by definition of G_i, for all $v \in \mathcal{V}$ s.t. $v \notin \arg\max_{v'}(U_i^u(v') - (U_i^u(v') + U_i^d(v'))\pi^{\beta^*}(v'))$, we have $\alpha_v^{*i} = 0$. Thus, we have

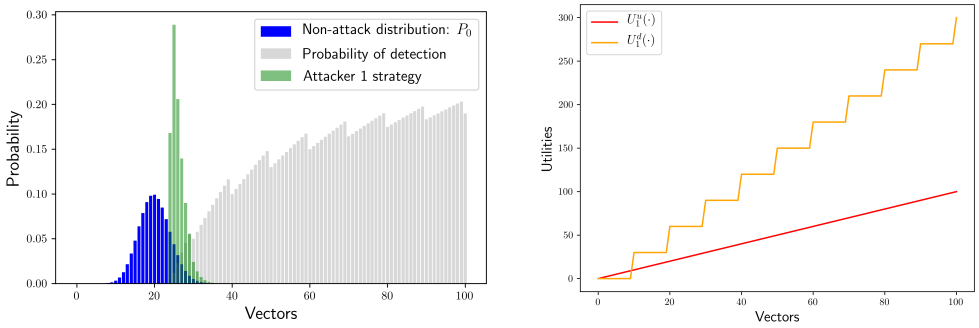
$$\alpha^* \in \arg\max_\alpha \sum_i p_i U_i^A(\alpha, \beta^*). \tag{4.5}$$

Combining (4.4) and (4.5), we conclude that (α^*, β^*) is a BNE. □

The linear program finds a valid strategy of the attacker which fits the conditions stated in Lemma 4.2 with the additional condition that each type of attacker must play their most rewarding vectors. While we had to resort to a linear program, finding the strategy of the attacker is still done in polynomial time in $m|\mathcal{V}|$ as we have a program with $m|\mathcal{V}|$ variables and $\mathcal{O}(m|\mathcal{V}|)$ constraints. It seems unlikely that a low-dimensional characterization similar to Section 4.1 can be found for the attacker's strategy as the defender's characterization heavily rely on the probability of detection function which removes the constraint to optimize on the simplex (the only constraint for π is that for all v in $\mathcal{V}$ we have $0 \leq \pi(v) \leq 1$).

4.3 Illustration of The Bayesian Nash Equilibrium

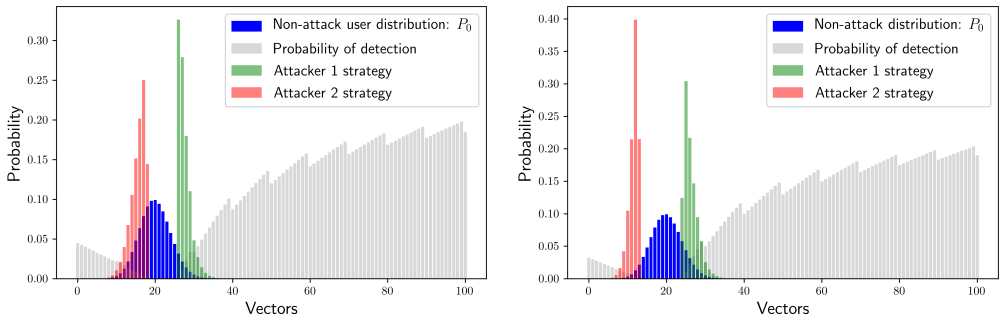
We provide here basic illustrations of our results concerning the structure of the BNE. We use Game 1 as it is the easiest to interpret the results. Figure 4.1a illustrates the behavior of both players at NE when Attacker 2 is not present ($p_2 = 0$). The attacker wants to play high vectors but must follow the distribution of the non-attacker over the vector they deem rewarding enough in regards to the defender's strategy to remain stealthy as stated in Lemma 4.2. The defender detects vectors with some spikes in the probability of detection function corresponding to the spikes in the cost incurred by detection. Indeed, at the equilibrium they make the attacker indifferent between some vectors and in order to do so, vectors which suffer from a sudden increase in cost incurred by detection can be detected less.



(a) (b)

Figure 4.1.: Game 1 illustration with only Attacker 1: (a) NE strategies; (b) parameters.

Figure 4.2 illustrates the impact of the presence of more than one type of attacker. In Figure 4.2a where both attackers are equally likely we observe that, compared to Figure 4.1a, Attacker 1 benefits from not being the only type of attacker as they play



(a) (b)

Figure 4.2.: Game 1 with both attackers: (a) BNE strategies $p_1 = p_1 = 0.5$; (b) BNE strategies $p_1 = 0.95, p_2 = 0.05$.

more rewarding vectors than when they were alone because the defender has less interest in detecting him. However, in Figure 4.2b where the Attacker 2 becomes less likely to appear, the situation of Attacker 1 gets closer to when they were the only attacker and they are reduced to playing less rewarding vectors. On the contrary, Attacker 2 benefits from being less likely as it is less interesting for the defender to detect them so they can play more rewarding vectors. Note that this situation is much better for Attacker 2 than for Attacker 1. Some of the most rewarding vectors for Attacker 2 are also used by non-attackers often so they can play them and remain stealthy while the most rewarding vectors of Attacker 1 are almost never used by the non-attackers so they are reduced to playing much less rewarding vectors to remain stealthy.

Summary: In this chapter, we characterized the strategies at the equilibrium of both the defender and the attacker. We showed that the optimal defense is random and can be parametrized using few parameters (in number independent in $|\mathcal{V}|$) representing the gain of each type of attacker. This parametrization allows us to compute the optimal strategy of the defender in time polynomial in $|\mathcal{V}|$, which is still not satisfactory as $|\mathcal{V}|$ is exponential in the number of features used to classify. This however builds the tools we need to provide efficient approximation of the optimal defense in the next chapter. We then characterized the strategy of the attacker as a response to the optimal defense strategy and showed that attackers should attack relevant vectors with probability tailored to strike a balance between the false alarm risk and the difference between detected and undetected gains. We finally illustrated the shape of the equilibrium on a simple artificial example.

5 Scalable Offline and Online Defense

Our previous results allow computing the equilibrium in time polynomial in $|\mathcal{V}|$. Yet, two major challenges remain: (i) $|\mathcal{V}|$ may be too large, in particular it grows exponentially with the number of features k; and (ii) computing the equilibrium requires knowledge of all parameters of the game and in particular of P_0, which can be hard to evaluate. We now answer our second main question "How can the defender compute optimal strategies in a scalable manner?". In this chapter, we provide two answers exploiting our low-dimensional characterization of the strategy of the defender depending on the data available to the defender. In Section 5.1, we tackle this problem from a stochastic optimization perspective assuming that the defender has access to historical data. We show that our problem is well suited to stochastic optimization techniques due to its polyhedral nature. In Section 5.2, we assume that the defender only has access to information on-line and provide an algorithm which is low-regret for the so-called Stackelberg regret. Both of these methods require limited knowledge of the parameters of the game (they notably do not require knowledge of P_0 and p_a). We finally show in Section 5.3 that both of these methods can be applied in a broader setting where $\mathcal{V}$ is no-longer finite but compact under mild technical assumptions. Finally, we extend our model to cover a mix of strategic and non-strategic attackers in Section 5.4. Our results hint that in this setting a mix between our low-dimensional characterization of the defender's strategy and classical learning algorithms may be suitable.

5.1 Scalable Offline Stochastic Optimization

In this section, we propose a training method that solves both issues by leveraging stochastic programming techniques. To do so, we first express $U^D(\boldsymbol{G})$ as an expected value as follows: $U^D(\boldsymbol{G}) = E[U^D(\boldsymbol{G}, \xi)]$ where $U^D(\boldsymbol{G}, \xi) = G_i$ with probability $p_a p_i$ and $U^D(\boldsymbol{G}, \xi) = C_{\text{fa}}(v)\pi_{\boldsymbol{G}}(v)$ with probability $(1-p_a)P_0(v)$ for all $v \in \mathcal{V}$. Leveraging the specific form of this stochastic function, we apply a stochastic programming

technique called sample average approximation (SAA, see [Sha03; KPH15; Ver+03; LSW06]) to obtain our training method, Algorithm 1.

Algorithm 1 Sample average approximation

Sample $\xi_1 \ldots, \xi_N$
Define $\tilde{U}^D(\boldsymbol{G}) = 1/N \sum_{i=1}^{N} U^D(\boldsymbol{G}, \xi_i)$
Maximize $\tilde{U}^D(\boldsymbol{G})$ on $[\underline{G}_1, \overline{G}_1] \times \ldots \times [\underline{G}_m, \overline{G}_m]$

The maximization step in Algorithm 1 can be done exactly through a linear program in the spirit of Proposition 4.1, in time polynomial in N since $\tilde{U}^D(\boldsymbol{G})$ is piecewise linear. Thus the complexity of this algorithm depends only on the sample size and not on the problem dimension. Additionally, very little information is required: the defender only needs to have access to N samples, which may correspond to a labeled dataset, as well as to the parameters $C_{\text{fa}}(v)$, $U^u(v)$, and $U^d(v)$ for non-attack samples. Yet the following theorem shows that Algorithm 1 outputs a very good approximation of the defender's min-max strategy.

Theorem 5.1. *Let $\hat{\mathcal{S}}$ be the set of maximizers of $\tilde{U}^D(\boldsymbol{G})$ from Algorithm 1 and $p_N = Pr[\hat{\mathcal{S}} \subseteq \arg\max U^D(\boldsymbol{G})]$. We have*

$$\limsup_{N \to \infty} \frac{1}{N} \log(1 - p_N) < 0.$$

Sketch of proof. A proof of Theorem 5.1 can be found in Appendix B.1. It relies on a strong result for sample average approximation (Theorem 15 of Shapiro [Sha03]), which fully exploits the structure of our problem as it requires the optimized stochastic function to be piecewise linear and to depend on random variables with finite support (extensions to continuous supports are possible under mild assumptions). This result is then enabled by the polyhedral structure of the problem. □

Theorem 5.1 states that Algorithm 1 will find an exact maximum of $U^D(\boldsymbol{G})$ with probability exponentially close to one (where the randomness is in the draw of the training set from unknown P_0, p_a and p_i). Then, from Theorem 4.1, this immediately gives an exact min-max strategy of the defender. The rate of the exponential convergence of p_N to 1 is not given by Theorem 5.1. It is possible to state a stronger result that gives the rate if the problem is "well conditioned"—which roughly means that $\arg\max U^D(\boldsymbol{G})$ is a singleton and the function is not flat around the optimum. More precisely, if there is a unique optimal solution we use the following bound to determine the conditioning of the problem. Denoting by $f'(x, d)$ the derivative of f in the direction d, there is a finite number of d_j s.t.

$U^{D\prime}(\mathbf{G}_{\max}, d_j) > 0$ and the event p_N happens if and only if $\tilde{U}^{D\prime}(\mathbf{G}_{\max}, d_j) > 0$. We then have $\lim_{N\to\infty} \frac{1}{N} \log(1 - p_N) < -\max_j \frac{(U^{D\prime}(\mathbf{G}_{\max}, d_j)^2}{2\mathrm{Var}(U^{D\prime}(\mathbf{G}_{\max}, \xi, d_j)]}$ ([Sha03], Ch. 6 Sec. 3.4).

This is however not guaranteed in any instance of our game, and such a result is anyways impractical because it depends on the true optimal value. From the high-probability result of Theorem 5.1, it is easy to derive that the output of Algorithm 1 is exponentially close to the true optimum since the function is bounded; although the exponential rate may be arbitrarily low if the problem is not well conditioned. In that case, though, worst case bounds show convergence of expected value at least in $N^{-1/2}$ and depending only on $Var[U^D(\boldsymbol{G}_{\max}, \xi)]$ (see Shapiro [Sha03]).

Theorem 5.1 combined with Theorem 4.1 shows that using SAA on top of our equilibrium characterization solves the key difficulties of our problem: we are able to compute an exact min-max strategy for the defender with high probability from a labeled training set without knowledge of P_0, p_a and p_i. It is remarkable that we do not need to estimate P_0 from the training set, this is automatically done within the stochastic approximation procedure. Other stochastic approximation algorithms (e.g., as stochastic gradient descent) could be used but without strong convexity property (which is our case since our function is piecewise linear), they only have convergence guarantees in $N^{-1/2}$.

5.1.1 Numerical illustration

We now illustrate our offline defense on a real bank fraud data set (Game 2).

Figure 5.1 represents the histogram of valid transaction amounts in $[0, 700]$ (where the majority of transactions occur) and the probability of detection function π_G obtained through our training for different values of ℓ (G_ℓ denotes the parameter trained on the dataset with false alarm cost factor ℓ). When ℓ is small, the defender classifies "aggressively" as fraud by accepting a high false alarm rate. When ℓ increases, the probability of detection functions show that the defender flags as fraud less often. For example, transactions of $700€$ are flagged with probability ~ 0.9 by the most aggressive strategy ($\ell = 0.006$) but only with probability ~ 0.1 for the least aggressive strategy ($\ell = 0.074$). The results presented here are computed through our training method in Algorithm 1 and may not be exact. We do not evaluate the quality of our approximation as we only have access to the empirical distribution P_0. We perform this investigation later in this section on games based on artificial distributions. The results suggest that the approximation is good even for much smaller training sets as hinted by the theoretical guarantee.

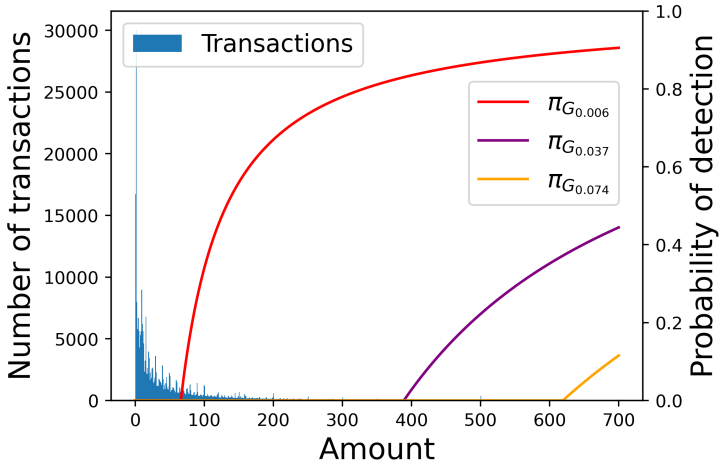


Figure 5.1.: Empirical distribution of transaction amounts and representation of defender min-max strategies for various l.

We now provide illustrations of the training process on Game 3 and Game 4. We use Game 3 to validate our previous experiments by showing that near-optimal defenses can be found on similar games with small data sets and that large data sets systematically yield a very good approximation. We use Game 4 to illustrate training in a setting with multiple attackers and no simple correspondence between features and costs.

First, Figure 5.2 shows the parameter G trained by the defender depending on ℓ (using the credit card fraud data set with $N = 284,807$). Recall that G corresponds to the gain of the attacker acting according to its best-response. Thus, a higher G means that the defender is willing to let the attacker gain more. This is compensated by the fact that when G increases, $\pi_G(v)$ decreases and so do the false alarms. We observe that the parameter G increases with the value of ℓ which corresponds to the fact that a defender facing higher false alarm costs is less willing to detect non-attacks. We also show computation time for the training on the data set for completeness. These are averaged over 10 run and plotted with error bars corresponding to one standard deviation and simply show that the training process can be applied with a medium sized data set with reasonable computation times. In the following figures, we evaluate the efficiency of our training process using a metric we call approximation ratio which is simply the ratio $100 * U^D(\boldsymbol{G}_{\max})/U^D(\tilde{\boldsymbol{G}})$ (expressed in percentage) where $\boldsymbol{G}_{\max}$ is a maximizer of U^D and $\tilde{\boldsymbol{G}}$ is the outcome of the sample average approximation algorithm. This value is always between 0 and 100 with 100 corresponding to a perfect training.

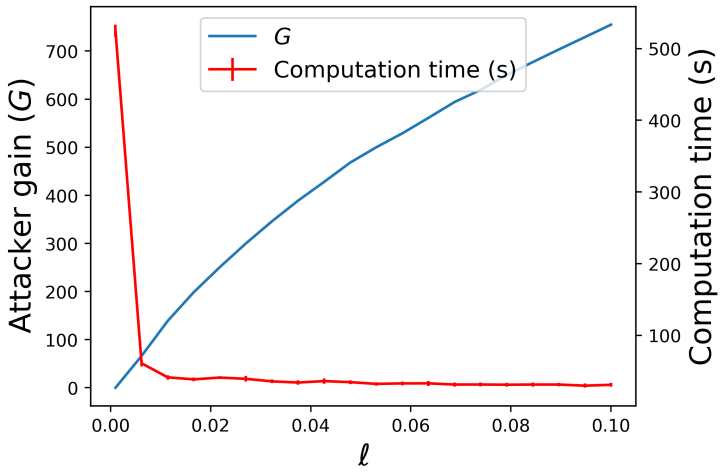


Figure 5.2.: Strategy of the defender training on the data set and computation times

Figure 5.3 illustrates training on Game 2 for different values of the parameter ℓ (0.001 and the parameters for which we plotted the corresponding probability of detection function in Figure 5.1). We plot the ratio between the loss of the defender using the optimal solution and the loss of the defender using the trained solution as well as the probability to obtain the optimal solution through training depending on the size of the data set. These are obtained by running our training algorithm on 300 different random training sets (each training set is generated i.i.d. with replacement). The approximation ratio is the average over the training set and is plotted with error bars corresponding to one standard deviation. The probability to obtain the optimal solution is computed on these random training sets. We observe that $\ell = 0.001$ is a best-case for p_N. This is caused by the fact that the equilibrium in this setting is trivial. The defender experiences such low false alarm that at equilibrium they classify all vectors as attacks. On the contrary, this is a worse case for the approximation ratio as a slight difference in the proportion of attackers in the training set can lead to a drastic change in strategy in this setting. We also observe that while p_N stays relatively low (but still significant) in all other experiments, the approximation ratio also reaches near 100% on average for data sets of size 5000. For comparison, we remind that Game 2 is defined with $|\mathcal{V}| = 25692$ vectors. Figure 5.4 shows the efficiency of the training process through the same metrics when we vary ℓ with a training set of the same size as the data set used in Section 5.1.1 ($N = 284807$). For each ℓ we performed experiments on 10 random data sets of size $N = 284807$. As previously we plot the averaged approximation ratio with error bars corresponding to one standard deviation. We observe that we systematically obtain the optimal solution, suggesting that a data set of this size is sufficient to correctly learn. Finally, Figure 5.5 shows training on Game 3 where for each training set size,

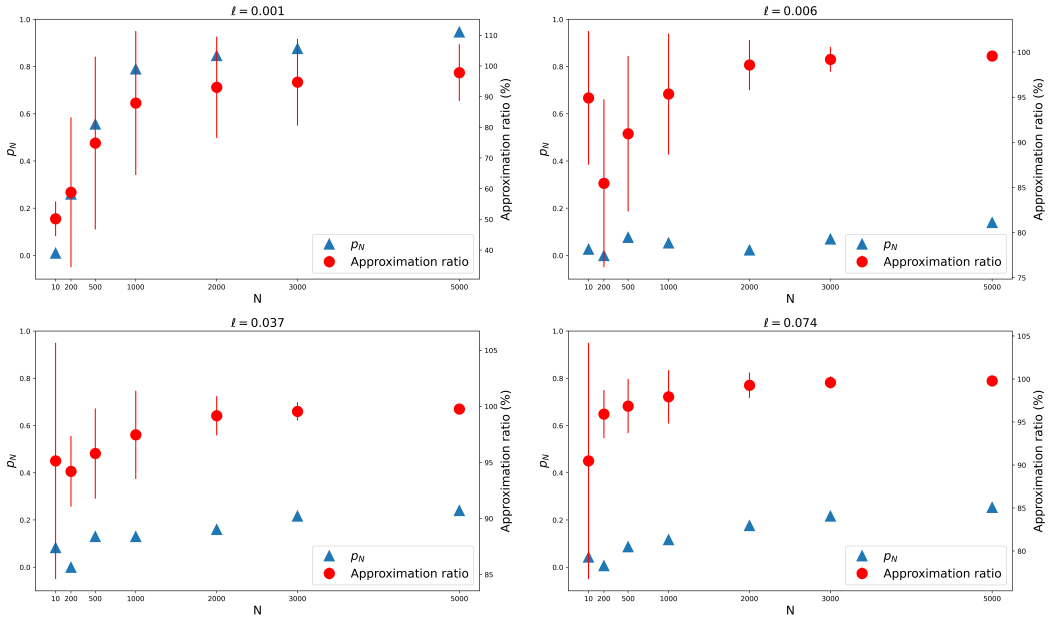


Figure 5.3.: Approximation ratio for Game 2.

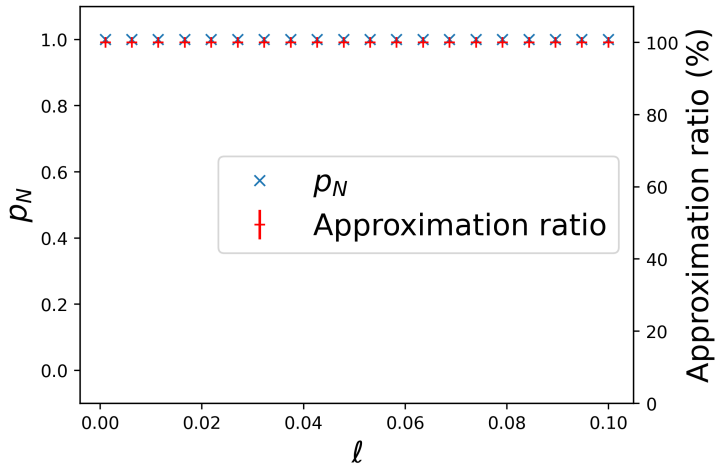


Figure 5.4.: Approximation ratio for $N = 284807$

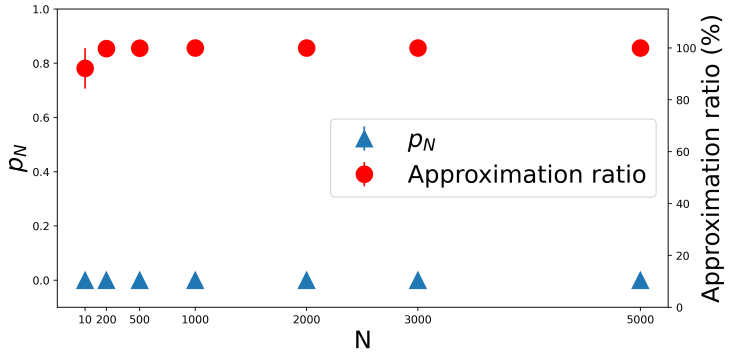


Figure 5.5.: Approximation ratio and p_N for Game 3.

we perform experiments 10 different random games with 4 types of attackers and $k = 19$. For each game, we perform experiments on 20 randomly generated training sets (generated iid with replacement) for a total of 200 experiments per data set size. We observe that the probability to obtain the optimal solution is always null. This is due to the linearity of the problem which makes the optimal values of $\boldsymbol{G}$ being equal to $U^u(v)$ for some v. As there is no relation between the costs of different vectors, with such small data sets, the probability that the vectors corresponding to the optimal parameters are present in the data set are very small and the linear program coming from the sample average approximation procedure cannot find the true optimal. We observe, however, that the approximation ratio is very good even for very small training sets. Note that the efficiency of the approximation is particularly striking in this case as we are able to obtain near-perfect results with training sets of size only 100 while the number of possible vectors is $|\mathcal{V}| = 2^{19}$. This illustrates well the independence in $|\mathcal{V}|$ of the training efficiency. Also note that our discussion in Section 5.1 about convergence of expected value in at least $N^{-1/2}$ rate translates directly to the approximation ratio. Our graphs suggest, however, that in many of our settings convergence happens at a faster rate.

5.2 Scalable Online Learning

In the previous section, we showed how the defender can compute an approximate min-max strategy from a training set. Yet, such historical data is not always available. We now show how our low-dimensional characterization of the min-max strategy also allows the defender to learn a good strategy *on-line*, without a priori knowledge of P_0, p_a and p_i, while incurring low loss as captured by the regret.

Specifically, we consider the following setting. At each time step $t = 1, \ldots, T$, the defender chooses a probability of detection function π_t and receives a vector v_t that is classified as an attack with probability $\pi_t(v_t)$. They incurs a loss $l(v_t)$ that is $C_{\text{fa}}(v_t)$ in case of false positive and 0 in case of true negative if facing a non-attacker; and $-U_i^d(v_t)$ and $U_i^u(v_t)$ in case of true positive and false negative respectively when facing a type i attacker. We assume that after classification, the defender can observe the type of attack (for convenience, we denote by type $i = 0$ non-attacks) and that they can compute $C_{\text{fa}}(v_t)$ and $U_i^u(v_t), U_i^d(v_t)$ for all i. Finally, as in Chen et al. [CLP20], we assume that attackers act according to best responses to π_t in a Stackelberg fashion, i.e., if the defender faces an attacker of type i at time t we have $v_t \in \arg\max_v \{U_i^u(v)(1 - \pi_t(v)) - U_i^d(v)\pi_t(v)\}$. The defender seeks to minimize the Stackelberg regret:

Definition 5.1 (Stackelberg regret)**.** *The average Stackelberg regret for a sequence of vectors* $(v_1, \ldots, v_T)$ *is:*

$$R(T) = \sum_{t=1}^{T} E_{\pi_t}[l(v_t)] - \min_{\pi} \sum_{t=1}^{T} E_{\pi}[l(v_t)].$$

The notion of Stackelberg regret has two implications. First, the sequence of vectors depends on the probabilities of detection used. In particular, $\min_\pi \sum_t E_\pi[l(v_t)]$ must be computed taking into account what would be the best response of the attacker to π. Second, it is key to remember that in our setting, the unknown quantities are P_0, p_a and p_i. There is no learning of the attacker's strategy as it is best-response to the utilities U_i^u, U_i^d assumed to be known.

It is possible to achieve low regret in T using naively the online gradient descent algorithm of Zinkevich [Zin03]—see Appendix B.2—to learn π directly. This gives, however, a bound on the Stackelberg regret of

$$R(T) \leq \frac{D^2\sqrt{T}}{2} + \left(\sqrt{T} - \frac{1}{2}\right) L^2, \tag{5.1}$$

with $L = \max(\max_v\{C_{\text{fa}}(v)\}, \max_{v,i}\{|U_i^u(v) + U_i^d(v)|\})$ (maximum gradient) and $D^2 = |\mathcal{V}|$ (maximum L_2 distance between two π functions)—see a proof in Appendix B.3. This bound is meaningless if the number of features k is large as $|\mathcal{V}| = \Omega(2^k)$. This approach is computationally infeasible as the full strategy π may not fit into memory.

Building on our characterization of the min-max strategy, we parametrize the defender's strategy by $\boldsymbol{G}$ to propose an alternate learning scheme as Algorithm 2 (where $\Pi_{\mathcal{S}}$ denotes the Euclidean projection on a set $\mathcal{S}$).

Algorithm 2 Efficient online gradient descent

Choose $\boldsymbol{G}_1 \in [\underline{G}_1, \overline{G}_1] \times ... \times [\underline{G}_m, \overline{G}_m]$ arbitrarily
for $t = 1, \ldots, T$ **do**
  Predict $\pi_{\boldsymbol{G}_t}$ and receive vector v_t and type i
  if v_t came from a non-attacker **then**
    grad $\in \partial(\pi_{\boldsymbol{G}_t}(v_t) C_{\text{fa}}(v_t))$
  else if v_t came from an attacker of type i **then**
    grad $= e_i$ (i^{th} vector of the canonical base of $\mathbb{R}^m$)
  end if
  $\boldsymbol{G}_{t+1} = \Pi_{[\underline{G}_1, \overline{G}_1] \times ... \times [\underline{G}_m, \overline{G}_m]}(\boldsymbol{G}_t - \frac{1}{\sqrt{t}}\text{grad})$
end for

Algorithm 2 exploits the fact that each attacker best responds to the defender's strategy, hence only strategies of the form $\pi_{\boldsymbol{G}}(\cdot)$ are worth using. Thus, instead of learning directly π, the defender learns the parameters $\boldsymbol{G}$. Note that this implies that the defender must be able to evaluate the bounds on the attackers gain they can impose. Algorithm 2 presents two major advantages: *First*, the defender's strategy is compactly represented with a small number m of parameters, independent of $|\mathcal{V}|$. *Second*, we get a much better regret bound:

Theorem 5.2. *Algorithm 2 gives Stackelberg regret bound* (5.1) *with*

$$L = \max\{1, \max_{v,i}\{\frac{C_{fa}(v)}{U_i^d(v) + U_i^u(v)}\}\} \text{ and } D = ||\overline{\boldsymbol{G}} - \underline{\boldsymbol{G}}||_2.$$

Sketch of proof. Theorem 5.2 is proved in Appendix B.4; the proof leverages our characterization of the min-max strategy with parameters $\boldsymbol{G}$. □

This result formalizes the intuition that learning $\boldsymbol{G}$ rather than π allows a much smaller regret (D is now independent in $|\mathcal{V}|$). Parameter L^2 now represents the change in false alarm cost one can expect at worst when changing parameters $\boldsymbol{G}$; which is different from L^2 in the naive procedure that corresponded to a gradient with respect to π. Note that we use the online gradient descent algorithm of Zinkevich [Zin03] for this theorem but similar results could be obtained using any low-regret online learning algorithm (with different constants) as our proof relies on applying online learning algorithms as black boxes to have low regret for the learning problem of $U^D(\boldsymbol{G})$ (a low dimensional function) which translate to

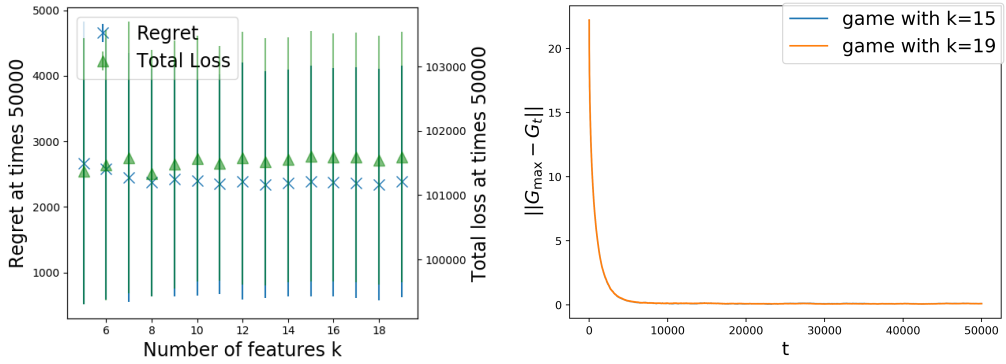
low Stackelberg regret for $U^D(\pi)$ (a high dimensional function). The choice of the learning algorithm to employ depends on the information available to the defender. In particular, we chose to present the version of Zinkevich [Zin03] as it does not require the computation of L (the maximum norm of the gradient of the function to optimize) which can be potentially hard to evaluate in our setting. We performed numerical experiments that illustrate the result of Theorem 5.2 in the next section (in particular the independence in $|\mathcal{V}|$). In addition, we observe that $\boldsymbol{G}_t$ converges (fast) towards $\boldsymbol{G}_{\max}$ (which is not implied by the no-regret property but is suggested by the similarity between online gradient descent and stochastic approximation algorithms in this particular setting).

5.2.1 Numerical Illustrations

We now illustrate Theorem 5.2 using Game 4.

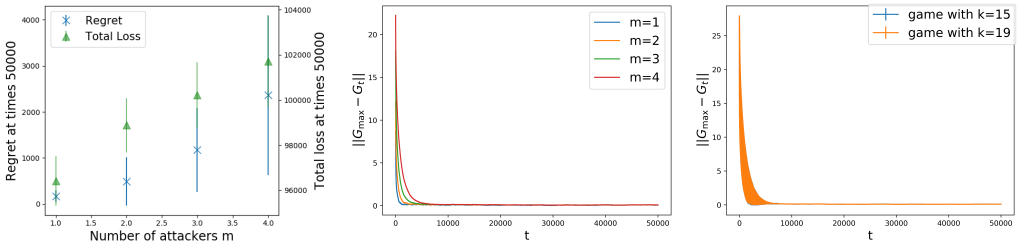
Figure 5.6a displays the regret the defender accumulates when learning with Algorithm 2 for different values of $k = \log_2(|\mathcal{V}|)$, at time $T = 50,000$. For each k, we average over 10 random games. In this experiment, D is close to 40 in all games which makes the regret bound of Theorem 5.2 at least $150,000$; the observed regret is significantly smaller (below $5,000$). We observe that increasing the number of features does not significantly impact the regret, in agreement with the bound of Theorem 5.2 that does not depend on the dimensionality of the problem, which illustrates the strength of our parametrization of the defender's strategy.

Figure 5.6b displays the distance between the parameters learned by Algorithm 2 and the optimal $\boldsymbol{G}_{\max}$ over time. For this experiment, we run the online algorithm on each game 10 different times with random starting point for the strategy of the defender. First, we observe that $\boldsymbol{G}_t$ converges towards $\boldsymbol{G}_{\max}$, hence the online strategy converges to the min-max strategy. This is interesting as it is not implied by the no-regret property. Second, it is remarkable to see that the convergence is fast (in $10,000$ steps even when $k = 19$). This can seem counter intuitive as one would be unable to learn P_0 in so few steps. However, what we need to learn is only the average false alarm cost associated to a strategy, this is learned fast through the update of the parameters $\boldsymbol{G}$. Figure 5.7 illustrates online learning on Game 3 with $k = 19$ and a number of possible attackers varying from 1 to 4. In Figure 5.7a, we observe that the regret accumulated by the defender increases with the number of attacker types. This illustrates the fact that our strategy is parametrized by the number of attackers and increasing this number increases the complexity of what we need to learn. Similarly, Figure 5.7b illustrates how far the learned strategy is



(a) Regret for different numbers of features. **(b)** Comparison BNE-online strategy.

Figure 5.6.: Regret and distance to the equilibrium



(a) **(b)** **(c)**

Figure 5.7.: Illustration of online learning on Game 2: (a) Regret for different numbers of attackers; (b) Distance to equilibrium for different numbers of attackers; (c) Distance to equilibrium for $m = 4$ with error bars.

from the equilibrium during learning. These two graphs can be contrasted with Figures 5.6a and 5.6b respectively, in which we observed that the regret and distance to equilibrium were not varying with the number of features. Thus, this illustrates the fact that our characterization is indeed independent from the number of vectors considered but depends only on the complexity of the characterization, i.e., the number of possible types of attackers. Finally, Figure 5.7c shows the same plot as Figure 5.6b (or Figure 5.7b with $m = 4$) with error bars that were omitted in the previous plots for readability.

5.3 Extension to Compact Vector Sets

Until now, we assumed that the set $\mathcal{V}$ of possible data is finite, but we make no other assumption on $\mathcal{V}$. That leaves a lot of flexibility; in particular it is possible to model situations where the features are categorical or boolean, or discrete numerical values (or a combination of those). Yet, some features are naturally continuous and it can

be convenient to model them as such instead of considering a discretization. One of the advantage of our characterization of the BNE is that is naturally extends to continuous feature spaces, as we sketch next.

To extend the model to continuous feature space, we assume that $\mathcal{V}$ is a compact metric set and that the defender's strategy space $\mathcal{D}$ is the set of M-Lipschitz continuous functions π from $\mathcal{V}$ to $[0,1]$. We use the same parameters and notation as in the finite game with the exception of $P_0(\cdot)$, which is now a continuous probability distribution. We assume that it has a density and denote it f_0. We then have the following payoffs:

$$\begin{aligned} U_i^A(v,\pi) =& U_i^u(v)(1-\pi(v)) - U_i^d(v)\pi(v), \\ U_i^D(v,\pi) =& -p_a U_i^u(v)(1-\pi(v)) + U_i^d(v)\pi(v) - (1-p_a)\int_{\mathcal{V}} cfa(v')f_0(v')\pi(v')\,dv'. \end{aligned}$$

Let us suppose that all functions are continuous and integrable. This ensures that the game is continuous and well defined, which, thanks to Glicksberg's theorem (see Glicksberg [Gli52]), ensures the existence of a BNE. Finally, we define $\underline{\boldsymbol{G}}$, $\overline{\boldsymbol{G}}$ and $\pi_{\boldsymbol{G}}$ as in the finite case. Let us assume that there exists $M \in \mathbb{R}$ such that, for all $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$ the function $\pi_{\boldsymbol{G}}(\cdot)$ is M-Lipschitz . This ensures that the optimal strategies of Definition 4.1 are available to the defender. Note that for this assumption to hold, it is sufficient to assume that $U_i^u(\cdot)$ is Lipschitz and that there exists $\epsilon \in \mathbb{R}$ such that $\forall v \in \mathcal{V}$, we have $|U_i^u(v) + U_i^d(v)| \geq \epsilon$. The first assumption simply implies that the variation of costs associated to undetected attacks should be bounded and the second one that there should always be at least a small difference between the reward an attacker gets with an undetected attack ($U_i^u(v)$) and a detected attack ($-U_i^d(v)$).

Note that in this continuous setting, we defined the game with strategies π directly for the defender (defining $\mathcal{C}$ for the discrete game was useful to get a finite game but this is irrelevant here) and therefore bypass many technical issues on the potential compactness of $\mathcal{C}$. We also see that the defender need not use mixed strategies as the payoff of any mixed strategy can be attained with a pure strategy π being the average of the functions in the mixed strategy. This is explained by the fact that π already represents a random classification, removing the need for further randomization.

We can then extend some of our results to the continuous case, in particular Theorem 4.1 that leads to the form of optimal classifiers. Again for the continuous case, intuitively, considering any equilibrium with attacker utility profile $\boldsymbol{G}$, the defender must have a probability of detection function $\pi(\cdot) = \pi_{\boldsymbol{G}}(\cdot)$ as it is the probability of

detection which gives gain G_i for each attacker while minimizing the false alarm cost. This gives the following proposition.

Proposition 5.1. *For all* $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$*, let*

$$U^D(\boldsymbol{G}) = -p_a \sum_{i=1}^{m} p_i G_i - (1 - p_a) \int_{\mathcal{V}} C_{fa}(v) f_0(v) \pi_{\boldsymbol{G}}(v)\, dv\,.$$

For all $\boldsymbol{G}_{\max} \in \arg\max_{\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]} (U^D(\boldsymbol{G}))$*,* $\pi_{\boldsymbol{G}_{\max}}(\cdot)$ *is a min-max strategy.*

Proposition 5.1 states that, as in the discrete case, finding a BNE in the continuous case amounts to finding the maximum of a concave function (function $U^D(\cdot)$ defined above). This problem can be solved using classical convex optimization tools, assuming that one can efficiently evaluate the function, in particular the integral that appears in the function. If one is unable to derive an exact formula for the integral stochastic optimization methods can still be applied. Theorem 5.1 however does not hold as it relies on the piecewise linearity of the function to optimize. Thus, only classical bounds in $N^{-1/2}$ hold. Note, however, that our results on online learning remain valid in the continuous setting as the proof of Theorem 5.2 does not rely on the assumption that $\mathcal{V}$ is finite.

5.4 Extension to Partially Strategic Attackers

We now extend our model to include non-strategic attackers. More precisely, we assume that attackers can either be strategic or non-strategic, with probability p_s and $(1 - p_s)$ respectively. Vectors generated by non-strategic attackers follow a fixed probability distribution P_{1n} on $\mathcal{V}$.

The defender chooses a classifier in $\mathcal{C} = 2^{\mathcal{V}}$, that is a function mapping a vector to a predicted class. We assume that the defender maximizes a utility function that balances costs/gains in different cases as follows.

- A false negative incurs a loss $U_i^u(v)$ (resp. $U_n^u(v)$) when facing a type i (resp. non-strategic) attacker.
- A true positive incurs a gain of $U_i^d(v)$ (resp. $U_n^d(v)$) when facing a type i (resp. non-strategic) attacker.
- A false positive incurs a false alarm cost $C_{\text{fa}}(v)$.

- A true negative incurs no cost.

Summarizing the above discussion, the utilities of the attacker and defender, when the attacker is of type i, are now defined as follows:

$$\begin{aligned} U_i^A(v,c) =& U_i^u(v)\mathbb{1}_{c(v)=0} - U_i^d(v)\mathbb{1}_{c(v)=1}, \\ U_i^D(v,c) =& -p_a p_s U_i^A(v,c) \\ & - (1-p_a)\sum_{v'\in\mathcal{V}} C_{\text{fa}}(v')P_0(v')\mathbb{1}_{c(v')=1} \\ & + p_a(1-p_s)\sum_{v'\in\mathcal{V}} U_n^d(v')P_{1n}(v')\mathbb{1}_{c(v')=1} \\ & - p_a(1-p_s)\sum_{v'\in\mathcal{V}} U_n^u(v')P_{1n}(v')\mathbb{1}_{c(v')=0}. \end{aligned} \tag{5.2}$$

For ease of notation, for all v, we define the quantities

$$\begin{aligned} C_d(v) &= (1-p_a)C_{\text{fa}}(v)P_0(v), \text{ and} \\ U_g(v) &= p_a(1-p_s)(U_n^u(v)+U_n^d(v))P_{1n}(v); \end{aligned}$$

We can then rewrite the payoffs of the players:

$$\begin{aligned} U_i^A(\alpha,\beta) =& \sum_{v\in\mathcal{V}} \alpha_v^i \left[U_i^u(v) - \left(U_i^u(v)+U_i^d(v)\right)\pi^\beta(v)\right]; \\ U_i^D(\alpha,\beta) =& -p_a p_s U_i^A(\alpha,\beta) \\ & + \sum_{v\in\mathcal{V}} \pi^\beta(v)\left(U_g(v)-C_d(v)\right) - U_g(v). \end{aligned} \tag{5.3}$$

We also define the set

$$\mathcal{V}_d = \{v\in\mathcal{V} : U_g(v)\geq C_d(v)\}$$

of vectors that are always more rewarding to detect for the defender. In a standard classification setting without a strategic attacker, the classification rule that maximizes the defender's utility (which corresponds to a Bayesian setting with prior on the classes) would be to simply classify vectors of $\mathcal{V}_d$ as 1 and others as 0.

In this setting, the structure of the game remains. Theorem 4.1 however does not hold due to the presence of vectors which should always be detected. We thus adapt the class of optimal probability of detection we consider:

Definition 5.2 (Optimal probability of detection)**.**

$$\pi_{\boldsymbol{G}}(v) = \begin{cases} \max_i \{\frac{U_i^u(v) - G_i}{U_i^d(v) + U_i^u(v)}\} & \text{if } v \notin \mathcal{V}_d, \\ 1 & \text{if } v \in \mathcal{V}_d. \end{cases} \tag{5.4}$$

Which in turn defines the minimum gain function:

Definition 5.3 (Minimum gain function U^D)**.** *For all* $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_i] \times \ldots \times [\underline{G}_m, \overline{G}_m]$*, let*

$$U^D(\boldsymbol{G}) = -p_a p_s \sum_{i=1}^{m} p_i G_i - \sum_{v \in \mathcal{V}} (C_d(v) - U_g(v)) \pi_{\boldsymbol{G}}(v).$$

We then have the following theorem, similar to Theorem 4.1

Theorem 5.3. *For any* $\boldsymbol{G}_{\max} \in \arg\max_{\boldsymbol{G}}(U^D(\boldsymbol{G}))$*, any strategy of the defender that yields a probability of detection function* $\pi(v) = \pi_{\boldsymbol{G}_{\max}}(v)$ *for all* $v \in \mathcal{V}$ *is an equilibrium strategy.*

With this theorem, we are now ready to adapt our main results in the presence of non-strategic attackers. To do so, we have two possible main choices.

1. Ignore the vectors in $\mathcal{V}_d$ and use the probability of detection function defined in Definition 4.1. It is then possible to adapt trivially our method (replacing $(1-p_a)C_{\text{fa}}(v)P_0(v)$ by $C_d(v) - U_g(v)$) to this setting at the cost of an imprecision equal to the loss of opportunity associated with the vectors in $\mathcal{V}_d$. More precisely, this adds the term $\sum_{v \in \mathcal{V}_d} U_g(v) - C_d(v)$ to the approximations we make.

2. Use classical learning to train a classifier which separates vectors belonging in $\mathcal{V}_d$ from other vectors. With this approach, our methods can also be trivially adapted. The imprecision term however depends on the accuracy of the classifier.

If non-strategic attackers are unlikely to be present, the first method might produce better results as it removes potential bias from the training of the classifier. On the contrary, if non-strategic attackers are likely to be present, the second method might produce better results as the added imprecision term $\sum_{v \in \mathcal{V}_d} U_g(v) - C_d(v)$ could be too high.

Summary: In this chapter, we showed that it is possible to efficiently compute an approximately optimal defense both offline and online with limited knowledge of the parameters of the game using our previous low-dimensional parametrization. In particular, the defender needs not to have access to P_0 and p_a which can be hard to evaluate precisely. Both our results guarantee approximations independent in the size of the vector set $|\mathcal{V}|$. Our offline method relies on sample average approximation and yields a non-null probability to find the optimal solution. Our online method relies on an adaptation to our parametrization of classical online learning algorithms such as gradient descent. We validated our results on both real data sets and artificial games. Finally, we show that our results extend to more general settings. We can consider compact vector sets at the cost of a slightly worse offline guarantee. We can also consider partially strategic attackers in which case our approach can be combined with classical learning approaches.

Part II

Strategic Linear Regression

6 State of the Art and Model

The model presented in this part is an extension of the model of Ioannidis and Loiseau [IL13a]. Some of the results and ideas presented in this part appear in our previously published article ([Gas+20]).

In this part (Chapters 6, 7 and 8), we present our second example of learning with strategic data. More precisely, we tackle the problem of linear regression when data comes from strategic data sources. Our main focus is the key question: *Do fundamental results on linear regression still hold when data are produced by strategic data sources?* We answer by the negative in general. In particular, estimators are no longer guaranteed to be *consistent* when the cost of producing data increases linearly with its precision. The `GLS` estimator is also no longer BLUE but is approximately optimal when considering a restricted family of estimators satisfying assumptions which are suitable in practice.

This first chapter reviews existing work broadly related to strategic linear regression and introduces our linear regression game. In Chapter 7 we then focus on game-theoretic properties of the linear regression game showing results fundamental for the rest of the analysis and for the applicability of our game in real-life settings. We finally answer our main question in Chapter 8 where we study the quality of the linear regression estimation in the strategic setting and the effect that different estimators have on the linear regression game.

6.1 State of the Art

There is a growing body of work on scenarios where one wants to learn from data provided by sources that choose their effort when generating data ([CDP15; Luo+15; LC16a; Wes+20]). 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 and crowdsourcing literature

contain similar mechanism design problems for cases where either the effort exerted or the data report (or both) are unverifiable ([FCK15; DG13; SZ16; Kon+20]). 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 as studied in Abernethy et al. [Abe+15], Chen et al. [Che+18a], and Chen and Zheng [CZ19] (see also related problems of mechanism design in the context of differential privacy with the work of Ghosh and Roth [GR11] and Dwork and Roth [DR14]).

A number of papers also consider data acquisition in sequential settings ([Abe+15; LC16b; Che+18a]). All this literature, however, considers agents that aim to maximize the payment received but are insensitive to the quality of the learning result. Moreover, agents aim to optimize payments while the learning algorithm is fixed; the only exceptions to the latter are Chorppath and Alpcan [CA13] and Caragiannis et al. [CPS16], which are restricted to the case of averaging and do not consider learning tasks such as regression. In contrast, in this thesis, we do not involve payments but assume that data sources benefit from the result of the learning algorithm.

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 [PP04; DFP10b; CPS16; Che+18b; BT19; HS20; SEA20; CSZ20] (see also similar problems in the context of classification [MPR12; Har+16; Don+18; Mil+19b; KR19; ZCC19; MMH20; TG20; BG20]). In particular, Dekel et al. [DFP10b] consider a broad class of regression problems in which data sources may misreport their private values, and determine loss functions under which empirical risk minimization is group strategyproof. The special case of linear regression is also treated, albeit in a more restricted setting, by Perote and Perote-Pena [PP04], who identify more general strategyproof mechanisms for the 2-dimensional case. More recently, Chen et al. [Che+18b] consider a similar setting and propose a family of group strategyproof regression mechanisms for any dimension, extending the results of both Dekel et al. [DFP10b] and Perote and Perote-Pena [PP04]. 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.

Closer to our setting, Hossain and Shah [HS19] consider the pure Nash equilibrium as a solution concept in regression games and investigate its efficiency, albeit in

a model closer to Dekel et al. [DFP10b] and Chen et al. [Che+18b] than to ours. Interestingly, this work considers the mean squared error, a standard quantity to measure a model's quality in linear regression, instead of our estimation cost based on the covariance matrix. Our estimation cost, however, includes a somewhat broader family of functions satisfying mild assumptions (see Assumption 6.3).

Our work is also related to different other fields of the literature which we quickly cover in the following paragraphs.

Data Perturbation for Privacy. Perturbing a dataset before submitting it as input to a data mining algorithm has a long history in privacy-preserving data-mining (see for example Vaidya et al. [VCZ06] and Domingo-Ferrer [Dom08]). Independent of an algorithm, early research focused on perturbing a dataset prior to its public release as in the works of Traub et al. [TYW84] and Duncan and Mukherjee [DM00]. Perturbations tailored to specific data mining tasks have also been studied in the context of, e.g., reconstructing the original distribution of the underlying data by Agrawal and Srikant [AS00], building decision trees by Agrawal and Srikant [AS00], clustering by Oliveira and Zaiane [OZ03], and association rule mining by Atallah et al. [Ata+99]. We approach such perturbation techniques via a non-cooperative setting, where individuals strategically choose the perturbation to their data.

The above setting differs from the framework of ϵ-differential privacy proposed in Dwork [Dwo06] and Kifer et al. [KST12], which has also been studied from the perspective of mechanism design by [NST12]. In differential privacy, noise is added to the *output* of a computation, which is subsequently publicly released. The analyst performing the computation is a priori trusted; as such, individuals submit unadulterated inputs. Several works study mechanisms incentivizing data disclosure under costs quantified by differential privacy (see [LR12; DFI14; GR11; CIL15]) whereby individuals are compensated for the privacy cost they incur. In contrast, we do not assume that the analyst is trusted, which motivates input perturbation. Such input perturbations also correspond to the more recently studied notion of local differential privacy of Duchi et al. [DJW13] and Kairouz et al. [KOV16], though such studies focus on the privacy/utility tradeoff, ignoring the strategic aspect of the input perturbation.

Experimental Design. In classic experimental design (see [Puk06; ADT07; BV04]), an analyst observes the public features of a set of experiments, and determines which experiments to conduct with the objective of learning a linear model (from non-strategic sources). The quality of an estimated model is quantified through

a scalarization of its variance as defined in Boyd and Vandenberghe [BV04]. As discussed in Section 6.2.1, many such scalarizations are used in the literature, including the so-called A-optimality, I-optimality, and V-optimality criteria we define in Section 6.2.2. We focus on non-negative scalarizations, to ensure meaningful notions of efficiency (as determined by the price of stability in Section 7.3). As we note in Section 6.2.1, convexity implies that the information gain (i.e., the cost reduction) due to new experiments is a submodular function. This has implications about mechanism design as well. For example, Horel et al. [HIM14] exploit this to produce a polytime mechanism with approximation guarantees for a version of the experimental design problem in which subjects report their private values truthfully, but may lie about the costs they require for their participation.

Public Good Provision Problems. We finally note that our model has analogies to models used in *public good* provision problems which have been the subject of many studies in economics (see, e.g., Morgan [Mor00] and references therein). Indeed, the variance reduction of estimators can be seen as a public good in that, when a source contributes data, all other sources in the game benefit. As is standard in such literature, our model assumes that the disclosure costs (corresponding to provision costs in public good problems) and the estimation cost (mapping to the public good benefit) are fully separable. 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 our previous publication [Gas+20] and the publications the model is based on (see Ioannidis and Loiseau [IL13b] and Chessa et al. [CGL15]). We also note that our game has the structure of an aggregative game in the sense of Cornes and Hartley [CH12]. This structure, however, does not offer any further insights (or help establish any of our results).

6.2 Models and Assumptions

6.2.1 The Linear Regression Game

We consider settings where globally successful data analysis may also provide a utility to the individuals from which the data is collected. This is evident in medical studies: an experiment may lead to the discovery of a treatment for a disease, from which an experiment subject may benefit. In the case of recommender systems, users may indirectly benefit from overall service improvements, as data disclosed may lead

to, e.g., improved product recommendations or better-targeted advertising. Similarly, open collaboration projects, by their nature, implicitly assume a common underlying utility, linked to the success of the collaboration. If such benefits outweigh associated privacy or effort costs, individuals may consent to the collection and analysis of high-quality data, e.g., by participating in a clinical trial, completing a survey, or disclosing their preferences in a recommender service.

We model this by 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.

We assume that an agent i can choose a *precision level* $\lambda_i(x_i) \in \mathbb{R}_+$ and produce an unbiased estimate $\hat{y}_i$ of $\boldsymbol{\beta}^\top x$ with this precision.Note that we can also impose an upper bound λ_{max} on the precision that an agent can choose; all 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. More explicitly, the response variable reported by the i-th agent for a data point x_i is

$$\hat{y}_i = \boldsymbol{\beta}^\top x_i + \epsilon_i, \tag{6.1}$$

where ϵ_i is an error term of mean 0 and variance $1/\lambda_i(x_i)$.

For each data point $(x_i, \hat{y}_i)$ they hold, agents send this estimate $\hat{y}_i$, along with the corresponding attribute x_i and of the precision $\lambda_i(x_i)$ to an aggregator (or analyst) that publicly discloses an estimate $\hat{\boldsymbol{\beta}}$. The error terms ϵ_i are assumed to be independent, but we do not make any further assumption on their distribution. We assume that regressors x_i are distributed with a n-dimensional joint distribution μ_{joint}.

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. Here, we call the ensemble $\boldsymbol{\lambda} = (\lambda_i, \boldsymbol{\lambda}_{-i})$ a *precision profile*. In what follows, we assume that the analyst computes this estimate by using *generalized least squares* (GLS) and denote it $\hat{\boldsymbol{\beta}}_{\text{GLS}}$. In this setting, we model the covariance of the resulting GLS estimator as:

$$V_{\text{GLS}}(\boldsymbol{\lambda}) = \left(\mathbb{E}\left[\sum_{i \in N} \lambda_i(x_i) x_i x_i^\top \right] \right)^{-1}. \tag{6.2}$$

Note that this quantity is well defined only if the matrix $M_{\text{GLS}}(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{i \in N} \lambda_i(x_i) x_i x_i^\top\right]$, called the information matrix, is invertible. If it is not, the estimator $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is not

unique as the generalized least squares problem has infinitely many solutions. We also sometimes refer to the setting where the analyst uses the *ordinary least squares* (OLS) estimator, in which case the definition of the game is simply obtained by replacing the covariance in (6.2) by the covariance of the OLS estimator:

$$V_{\texttt{OLS}}(\boldsymbol{\lambda}) = \mathbb{E}\left[\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\lambda_i(x_i)} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}\right]. \tag{6.3}$$

Note that the covariance we consider is a function of the expected information matrix $\mathbb{E}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]$ rather than the expected value of the inverse of the information matrix $\mathbb{E}\left[\left(\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right)^{-1}\right]$. We discuss this below (6.4) after defining the different costs the agents face.

We then posit that each agent tries to balance a trade-off between two components:

1. *Idiosyncratic cost:* The value $\hat{y}$ 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 but only reveals a noisy version of it $\hat{y}$. 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 $\hat{y}$ with precision $\lambda_i(x)$ induces a cost $c_i(\lambda_i(x))$ 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 $\hat{\beta}$. For example, in a medical context, agents would be interested to know that a given disease is correlated to their weight or cholesterol 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 an *estimation cost* defined as a function of the covariance of the obtained estimator:

$$C_{\text{estim}}(\boldsymbol{\lambda}) = F\left(\left(\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} \lambda_i(x_i) x_i x_i^\top\right]\right)^{-1}\right), \tag{6.4}$$

Where $F : S_+^d \to \mathbb{R}_+$ is a so-called *scalarization function* that maps the covariance of the estimator to a cost.

Scalarizations are standard in optimal design , and they include standard metrics of a model's quality (such as the mean squared error) as special cases (see Section 6.2.2). By convention, if the information matrix is not invertible and $V_{\texttt{OLS}}(\boldsymbol{\lambda})$ is not well defined, the estimation cost is infinite.

The public good component (6.4) 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, (6.4) can be seen as a slightly weaker Bayesian model where agents would optimize over $\mathbb{E}_{\mu_{\text{joint}}}\left[F((\sum_{i\in N}\lambda_i(x_i)x_ix_i^\top)^{-1})\right]$. The Bayesian model however 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. Our model in contrast can encompass uncertainty of agents over the data points of over agents similarly to Bayesian models (see Section 7.2) but does not encounter such issues and can be justified in several relevant settings. While this is not the most intuitive model (if data points are created according to a given distribution, each agent should hold only a single data point), there are several justification for it. First, from a modeling standpoint, it still makes sense if μ_{joint} represents a deterministic attribution of vectors (i.e. $\mu_{\text{joint}}(\boldsymbol{x}) = 1$ for some $\boldsymbol{x} \in (\mathbb{R}^d)^n$) as this is the complete information setting considered by Gast et al. [Gas+20]. Another relevant application of this model is in the context of federated learning as in Yang et al. [Yan+19]. 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 (there are many agents, each capable of a local optimization as in Konečny et al. [Kon+16]) or privacy (all agents want to compute a joint representative model without explicitly having to share their personal data in the spirit of Geyer et al. [GKN17]). Our model can be viewed as an instance of both cases. Finally, we show in Theorem 7.1 that the complete information setting is equivalent to the independent and identically distributed setting presented in Section 6.2.3 where the data point of each agent is generated through a common distribution μ. This latter model is relevant as it requires only knowledge of the common distribution rather than the knowledge of the data point of each agent which is unrealistic to assume when the number of agents grows.

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

$$J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = \mathbb{E}_{\mu_{\text{joint}}} \left[c_i(\lambda_i(x_i)) \right] + C_{\text{estim}}(\boldsymbol{\lambda}), \tag{6.5}$$

where the expectation is taken with respect to the law μ_{joint} of the attribute vectors $\boldsymbol{x}$. Note that this quantity can also be written exclusively using the marginal probability distribution of each agent i denoted μ^i_{marg}. This marginal distribution sometimes allows us to write results and assumptions in a cleaner way. Additionally, given that agent i chooses the function $\lambda_i : \mathcal{X} \to \mathbb{R}_+$, minimizing the expected data provision cost $\mathbb{E}_{\mu_{\text{joint}}} \left[c_i(\lambda_i(x_i)) \right]$ as in (6.5) is equivalent to minimizing the cost for each value of x_i separately.

The setting described above defines a game that we refer to as the *linear regression game* denoted Γ. 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 this part, 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 $\lambda_i(x_i)$.

We present all our results on the `GLS` estimator and point out when they are also valid for the `OLS` estimator (which changes the covariance matrix of the estimator but keeps the structure of the game). This is justified in two major ways. First, for the model to hold, the agents must believe that the analyst will use the estimator they specified. When the analyst knows the precision, `GLS` is BLUE and once the analyst receives data points, they have no reason to use another estimator. If the analyst does not know the precision, it makes sense to use the `OLS` model which does not use this information. Additionally, even if the analyst convinces the agents that another estimator will be used (for example through a trusted third party), we show in Section 8.1 that `GLS` is approximately optimal when considering a wide class of estimators which satisfy intuitive properties.

We note here that as 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, it is possible

to consider the ordinary least squares (OLS) estimator instead of GLS, as OLS is oblivious to the disclosed precision of the data points.

6.2.2 Assumptions

We first define assumptions which are necessary for our game to be properly defined.

Assumption 6.1. *The attribute set $\mathcal{X}$ is finite, $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n} x_i x_i^\top\right]$ is positive definite and for all vectors x of $\mathcal{X}$, there is an agent i such that $\mu^i_{marg}(x) > 0$.*

This first assumption states that we consider no useless vector in the model (we can reduce any model to this case by simply removing vectors which are present with probability 0) and that the attribute space is well explored. Indeed, the assumption $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n} x_i x_i^\top\right]$ is valid if and only if the family of vectors of $\mathcal{X}$ span $\mathbb{R}^d$. This assumption makes our model properly defined from a statistical point of view but we still need assumptions to ensure that our model is well defined from a game theory perspective.

Assumption 6.2. *The provision costs $c_i : \mathbb{R}_+ \to \mathbb{R}_+$ are non-negative, increasing, and convex.*

Assumption 6.3. *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, i.e., for all $a > 0$ and all $V \in S_+^d$, $F(aV) = a^q F(V)$.*

The monotonicity and convexity assumptions in Assumptions 6.2 and 6.3 are standard and natural. Increasing the precision λ_i leads to a higher disclosure cost. In contrast, increasing λ_i can only decrease the estimation cost: this is because decreasing the variance of an agent's provided perturbed variable also decreases the variance in the positive semidefinite sense (as the matrix inverse is a positive semidefinite decreasing function).

The convexity assumption in Assumption 6.3 is also standard and natural. Intuitively, the naturalness of Assumption 6.3 stems from the following observation: the convexity of F implies that the so called *information gain*, i.e., the relative reduction in F as a new label is collected, exhibits a diminishing returns property, as additional labels affect estimation quality less and less. Scalarizations of positive

semidefinite matrices and, in particular, of the covariance matrix $V(\boldsymbol{\lambda})$, are abundant in statistical inference literature in the context of experimental design, see Boyd and Vandenberghe [BV04], Pukelsheim [Puk06], and Atkinson et al. [ADT07] (also known as batch active learning). Similar to our setting, in experimental design an analyst has access to samples with known feature vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in N$, and wishes to conduct a limited number of k experiments, where $k \ll N$, to collect labels $y_i \in \mathbb{R}$ for a subset of these samples. Given a budget k, the experimental design problem amounts to determining which labels to collect. The standard approach is to accomplish this by minimizing a scalarization function of the covariance of the estimator applied to the labels selected (see detailed discussions in Boyd and Vandenberghe [BV04], Pukelsheim [Puk06], and Atkinson et al. [ADT07].). In what follows, we present several important scalarizations satisfying our assumptions.

The trace: The trace trivially satisfies Assumption 6.3 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.

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

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

It is easy to check that this scalarization satisfies Assumption 6.3 with $q = 2$.

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

$$\texttt{MSE}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right]. \tag{6.6}$$

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 (see Dekking et al. [Dek+05]). 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 precision of the response variables $\tilde{y}_i$ (unlike the empirical mean squared error).

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:

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

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:

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

As $\hat{\beta}$ 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 $\hat{\beta}$. We then have:

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

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

1. **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:

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

   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 $\rho = \mu$.

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

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

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 literature as the V optimal design criteria.

Finally, our main results hold for a class of provision cost functions which we call near-homogeneous and satisfy the following assumptions.

Assumption 6.4. *There exist* $1 \leq p_{\min} \leq p_{\max} \in \mathbb{R}_+ \cup \{+\infty\}$ *such that, for all* $i \in N$*, the disclosure cost* $c_i : \mathbb{R}_+ \to \mathbb{R}_+$ *satisfies:*

$$a^{p_{\min}} c_i(\lambda) \leq c_i(a\lambda) \leq a^{p_{\max}} c_i(\lambda), \qquad \textit{for all } \lambda \in \mathbb{R}_+ \textit{ and } a \geq 1. \tag{6.7}$$

Intuitively, Assumption 6.4 captures "near-homogeneity" of the provision cost functions. It is, for example, satisfied when all agents have monomial provision costs $c_i(\lambda) = r_i \lambda^{p_i}$, where r_i is a constant, with different exponents $p_i \in [p_{\min}, p_{\max}]$. Note that our asymptotic results such as Theorem 8.4 imply that the precision each agent provides tends to 0 when the number of agents grows. In practice, this means that only the behavior of the data provision costs near 0 matter and our results can thus often be extended when Assumption 6.4 is only verified locally around 0.

6.2.3 Examples Used in Proofs and Illustrations

Our negative results and our illustrations on this linear regression game often rely on simple settings for ease of understanding. In particular,

The complete information case: When there exists $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that $\mu_{\text{joint}}(\boldsymbol{x}) = 1$, there is no uncertainty on the vector each agent possesses. We refer to this case as the complete information game. In this setting, the choice of $\lambda_i(x_k)$ for $k \neq i$ does not change the cost of agents. We thus sometimes abuse notations in this case and denote the strategy of agent i by $\ell_i = \lambda_i(x_i)$. In this case, we use the matrix

notations of Chapter 2 to write the estimators. This is the model studied in Ioannidis and Loiseau [IL13a] and Chessa et al. [CGL15] and used for the experiments showed in Figure 8.3.

The mean estimation case: A special case of the previous complete information setting is when the set of possible vectors is reduced to a singleton (usually $\mathcal{X} = \{1\}$ or $\mathcal{X} = \{-1, 1\}$ for convenience). In this case, the linear regression corresponds to the estimation of the mean value of $\hat{y}$. This setting is useful to grasp intuition of the model while removing the need for linear algebra as in this case the cost of an agent is simply: $J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = \mathbb{E}_{\mu_{\text{joint}}}\left[c_i(\lambda_i(x_i))\right] + F(\frac{1}{\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i \in N} \lambda_i(x_i)\right]})$. We often consider this setting with F being the trace function which is the identity function in this 1-dimensional case. This model is used for the experiments showed in Figures 8.1, 8.2, 8.9, 8.10, and 8.11.

The independent and identically distributed case: In this setting, the probability distribution μ_{joint} can be expressed as a product of identical and independent probability distributions μ. Formally, for all $\boldsymbol{x} \in \mathcal{X}^n$, we have $\mu_{\text{joint}}(\boldsymbol{x}) = \mu(x_1) \times \cdots \times \mu(x_n)$. This model is of practical importance as we show in Theorem 7.1 that it is equivalent in some sense to the complete information case while requiring minimal information. Indeed, in this settings, agents only need to know the common distribution μ rather than the full distribution μ_{joint} which can be impractical to evaluate when the number of agents grow. This is the model studied in Gast et al. [Gas+20] and used for the experiments showed in Figures 8.4, 8.5, and 8.7.

7 Structural Results About the Game

In this chapter, we introduce structural results in our game which concern game theoretical aspects. In Section 7.1 we show that our game is a potential game. While this property is very common, it is crucial in our analysis of the statistical properties of the estimators in the next chapter. We then show in Section 7.2 that our assumptions allows us to consider two very important special cases which can be used to model interactions where there are few agents (the complete information model) of interactions where there are many agents and precise information about data points of other agents is hard to obtain (the independently and identically distributed model). Finally, in Section 7.3, we provide some bounds on the *price of stability* of our game which characterizes the social inefficiency of the linear regression game due to the selfishness of the agents.

7.1 The Linear Regression Game is a Potential Game

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

$$\phi(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{j=1}^{n} c_j(\lambda_j(x))\right] + C_{\text{estim}}(\boldsymbol{\lambda}). \tag{7.1}$$

Using the form of $J_i(\lambda_i, \boldsymbol{\lambda}_{-i})$ in (6.5), 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 (7.1). Since the function ϕ is independent of i, this shows that the game is a potential game as defined in Neyman [Ney97] and ϕ is the potential of the game. It is easy to see that this potential function is convex under Assumptions 6.1, 6.2, and 6.3 (and we show it rigorously in Appendix C.2). In this case, the set of Nash equilibria coincides with the set of local minima of the function ϕ. We note here that there may exist equilibria with infinite cost which we call trivial depending on. Intuitively, if no single agents can output data leading to a well-defined estimator, the situation where no agents participate is a Nash equilibrium with infinite cost as no estimator is obtained. For example, if $d \geq 2$ and $\mathcal{X} = \{[1,0]^\top, [0,1]^\top\}$ in the complete information setting

with $\mu_{\text{joint}}([1,0]^\top, [0,1]^\top) = 1$, the strategy profile $\boldsymbol{\lambda}^* = 0$ is a Nash equilibrium. We formalize this observation in the next proposition.

Proposition 7.1. *There exists a trivial Nash equilibrium of infinite cost if and only if there exists no agent i such that $\mathbb{E}_{\mu^i_{marg}}\left[x_i x_i^\top\right] \succ 0$.*

The proof is given in Appendix C.1. In this part, we will focus on non-trivial Nash equilibria. We note here that it is easy to guarantee that no trivial Nash equilibria exists. In particular, we can add non-strategic agents who produce data points with fixed precision and who guarantee that the estimators we consider are well defined. In practice, this would correspond to public data bases already available in many settings. Our model can be thought of as a limit the model including non-strategic agents when the precision of non-strategic points goes to 0. As stated in the next proposition, expressing the game as a potential game simplifies the study of its non-trivial Nash equilibria.

Proposition 7.2. *Under Assumptions 6.1, 6.2, and 6.3 a precision profile $\boldsymbol{\lambda}^*$ is a non-trivial 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 non-trivial equilibria, the estimation cost $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ does not depend on the equilibrium.*

The proof is given in Appendix C.2. We note here that this proof can be straightforwardly adapted to the `OLS` estimator. This result transforms the problem of studying the Nash equilibria into the easier problem of studying the minima of a convex function. Note that the main assumption that makes our linear regression game a potential game is that the estimation cost is independent of i, which is natural since the model's quality depends on a unique covariance matrix. It is also robust to variations such as introducing agent-dependent multiplicative factors, but we keep here the simplest formulation. This property is central to all of our proofs as we rely on exhibiting either approximate minimum of the potential (see the proof of Theorem 8.4) or constructing good parameters for the potential from an existing solution (see the proof of Theorem 7.3).

7.2 Equivalence Between Two Special Cases: the Complete Information Case and the Independent and Identically Distributed Case

In this section, we present two important special cases which can be modeled with our assumptions. We assume that data points are independently and identically distributed with a probability distribution μ. This assumption defines two possible games:

1. The most natural model is the complete information case where each agent knows the data point of each other agent. This is modeled in our case by setting $\mu_{\text{joint}}(\boldsymbol{x}) = 1$ for some $\boldsymbol{x} = (x_1, \dots, x_n)$ where $x_1, \dots, x_n$ are drawn independently according to distribution μ. This models intuitively the strategic linear regression setting but requires each agent to know the distribution μ_{joint} which means that each agent must have knowledge about the $n-1$ data points of others. In asymptotic settings, this is not realistic. We denote the potential function of this game $\phi^{\text{ci}}_{\boldsymbol{x}}$.

2. The second model is the independent and identically distributed case where the data point of each agent is independently produced from a common distribution μ (the joint distribution μ_{joint} is the product distribution of the independent and identically distributed random variables). This model requires only knowledge of the common distribution. We denote the potential function of this game ϕ.

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

Notations and assumptions: We assume that the provision costs functions satisfy Assumptions 6.2 and 6.4 and that the estimation cost satisfies Assumption 6.3. 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 $t \in \mathcal{T} := [\![1, T]\!]$. With these assumptions, we can state our theorem.

Theorem 7.1. *Let $\boldsymbol{\lambda}^*$ be a non-trivial equilibrium of the linear regression game and $\boldsymbol{\lambda}^{\text{ci}*}$ be a non-trivial equilibrium of the complete information game. For all $0 < \epsilon < 1/2$, we have with probability at least $1 - |\mathcal{X}| \sum_t 2\exp(-2n_t^{2\epsilon})$:*

$$\frac{1}{\max_{x,t}\left(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)}\right)^{p_{\max}-1}}\phi(\lambda^*) \leq \phi_{\boldsymbol{x}}^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}) \leq \max_{x,t}\left(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}\right)^{p_{\max}-1}\phi(\lambda^*), \tag{7.2}$$

$$\phi_{\boldsymbol{x}}^{\text{ci}}(\boldsymbol{\lambda}^*) \leq D_n \max_{x,t}\left(\frac{\mu(x)+n_t^{\epsilon-1/2}}{\mu(x)}\right)^{p_{\max}-1}\phi_{\boldsymbol{x}}^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}), \tag{7.3}$$

and

$$\phi(\boldsymbol{\lambda}^{\text{ci}*}) \leq D'_n \max_{x,t}\left(\frac{\mu(x)}{\mu(x)-n_t^{\epsilon-1/2}}\right)^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*); \tag{7.4}$$

where

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

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

The intuition behind the theorem is as follows. Equation (7.2) 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 (7.3) and (7.4) 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 (7.2) applied with $p_{\max} = 1$ yields $\phi(\lambda^*) = \phi_{\boldsymbol{x}}^{\text{ci}}(\lambda^{\text{ci}*})$. Finally, we emphasize that the complexity of Theorem 7.1 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 potentials go to 0 as long as $p_{\min} > 1$ (see Section 8.2).

The proof of this theorem can be found in Appendix C.3. It relies on using Hoeffding's inequality to bound with high probability the difference between number of agents having data point x and provision cost of type t and its average. We then exploit our homogeneity assumptions to show that these two games are similar by factorizing the ratio between the number of agents with data point x of type t and its average. This shows that the potentials can be written as a function of the potential of the other game times a factor which goes to 1 as the number of agents grows.

7.3 Price of Stability

We now turn our attention to issues of efficiency. We define the *social cost* function $C : \mathbb{R}^n \to \mathbb{R}_+$ as the sum of all agent costs, and say that a strategy profile $\boldsymbol{\lambda}^{\texttt{opt}}$ is *socially optimal* if it minimizes the social cost, i.e.,

$$C(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i \in N} c_i(\lambda_i)\right] + nC_{\text{estim}}(\boldsymbol{\lambda}), \qquad \text{and} \qquad \boldsymbol{\lambda}^{\texttt{opt}} \in \operatorname*{arg\,min}_{\boldsymbol{\lambda} \in [\mathcal{X} \to \mathbb{R}_+]^n} C(\boldsymbol{\lambda}).$$

Let $\texttt{opt} = C(\boldsymbol{\lambda}^{\texttt{opt}})$ be the minimal social cost. We define the *price of stability* as the ratio of the social cost of the best Nash equilibrium in Γ to $\texttt{opt}$, i.e.,

$$\texttt{PoS} = \min_{\boldsymbol{\lambda} \in \texttt{NE}} \frac{C(\boldsymbol{\lambda})}{\texttt{opt}},$$

where $\texttt{NE} \subseteq [\mathcal{X} \to \mathbb{R}_+]^n$ is the set of Nash equilibria of Γ. Intuitively, the price of stability represents the social inefficiency induced by the selfishness of agents. If they coordinated, they could aim for a common reduction of cost and experience a lesser cost on average. They instead take less socially-optimal decisions as other agents might engage in "free-riding", i.e. not being productive and only enjoying the result of the effort of other agents.

Note that there exists other measures of social efficiency such as the price of anarchy which compares the social cost to the worst Nash equilibrium in Γ to $\texttt{opt}$. In our case, however, recall that depending on the characteristics of the joint distribution μ_{joint}, trivial Nash equilibrium of infinite cost may exist making the price of anarchy infinite. When there exists no trivial Nash equilibrium, however, or when we ignore trivial Nash equilibrium (for example through modeling artifacts), Theorem 7.1 implies that all equilibrium achieve the same social cost and thus the price of stability and the price of anarchy have the same value.

The fact that our game admits a potential function has the following immediate consequence (see for example Schäfer [Sch11] and Sandholm [San10]):

Theorem 7.2. *Under Assumptions 6.2 and 6.3,* $\texttt{PoS} \leq n$.

The following result provides tighter bounds when the provision costs additionally satisfy Assumption 6.4.

Theorem 7.3. *In addition to Assumptions 6.2 and 6.3, assume that the provision cost functions satisfy Assumption 6.4 with* $p_{\min} \geq 1$ *and* $p_{\max} \in \mathbb{R} \cup \{\infty\}$. *Then, the*

price of stability satisfies PoS $\leq n^{\frac{q}{p_{\min}+q}}$. *Additionally, for all* $p_{\min}, q \geq 1$, *and all* $\varepsilon > 0$, *there exists a game in which the estimation cost and the provision costs satisfy Assumptions 6.4 and 6.3, respectively, such that* PoS $\geq n^{\frac{q}{p_{\min}+q}}(1-\varepsilon)$.

Note that, as the bound does not depend on $p_{\max}$, we can set $p_{\max} = \infty$ in Assumption 6.4, which is equivalent to replacing this assumption by

$$a^{p_{\min}} c_i(\lambda) \leq c_i(a\lambda), \qquad \text{for all } \lambda \in \mathbb{R}_+ \text{ and } a \geq 1.$$

The proof of the upper bound relies on deriving a "good" solution from the social optimum and showing that, if the PoS is too high, this "good" solution attains a lower potential than a Nash equilibrium (a contradiction). The proof of the lower bound in Theorem 7.3 relies on explicitly characterizing the socially optimal profile in a certain game class, and showing it equals the Nash equilibrium $\boldsymbol{\lambda}^*$ multiplied by a scalar. We note here that this proof can be straightforwardly adapted to the OLS estimator.

The theorem states that, among monomial provision costs and for any estimation cost satisfying Assumption 6.3, the largest PoS is $n^{\frac{q}{1+q}}$ and is attained for linear provision costs. Similarly, among all estimation costs satisfying Assumption 6.3 and all provision costs satisfying the assumptions presented in Theorem 7.3, the largest PoS is n; this is approached as q tends to infinity. We obtain a similar worst-case characteristic of linear provision cost regarding the estimation cost experienced by agents in Theorem 8.4. When q tends to infinity, however, it represents a best-case scenario for the estimation cost.

We note that a similar worst-case efficiency of linear functions among convex cost families has also been observed in the context of other games, including routing games as in Roughgarden and Tardos [RT02] and resource allocation games as in Johari and Tsitsiklis [JT04]. As such, Theorem 7.3 indicates that this behavior emerges in our linear regression game as well but only concerning the provision cost: We observe a worst-case efficiency of linear functions in this game for the provision cost but a worst-case efficiency of highly convex functions for the estimation cost.

Finally, note that we also see the worst-case efficiency of linear functions later for the convergence rate of the estimation cost in Theorem 8.4. In this case, however, highly convex functions for the estimation cost no-longer correspond to a worst-case efficiency but to a best-case efficiency. Intuitively, the role of the estimation cost in the price of stability is that when it is highly convex it shrinks quickly and thus does not incentivize agents to exert efforts leading to poor social cost. What Theorem 8.4 implies when compared to this theorem is that, purely from an estimation cost

perspective, the benefit of having an estimation cost which shrinks faster outweighs the loss due to the absence of incentive for agents to exert efforts.

Summary: In this chapter, we studied the game-theoretic aspect of the linear regression game. We showed that it is a potential game. This property will be central to our proofs. We also showed that our assumptions allow us to apply our model in a complete information setting while minimizing the information that needs to be available to each agent to only a single common distribution. Finally, we characterized the social inefficiency of this linear regression game and exhibited a higher inefficiency when provision costs are linear and when scalarizations are highly convex.

8 Properties of Linear Unbiased Estimators in the Strategic Setting

In this chapter, we study the statistical properties of linear estimators when data are strategically produced. We focus specifically on the *quality* of the estimation through the scalarization of the covariance and answer the two following key questions:

1. Does the `GLS` estimator produce the best-quality estimation? That is to say, is the estimation cost associated to `GLS` at equilibrium the lowest among estimation costs associated to any linear unbiased estimator at equilibrium?
2. Is the estimation still consistent in the presence of strategic agents? And if so, how does the convergence rate compare to the non-strategic case?

We first provide in Section 8.1 both a positive and a negative answer to question 1 – `GLS` produces an approximately best-quality estimation among a class of estimators satisfying suitable statistical properties. We then characterize in Section 8.2 how the presence of strategic agents degrades the quality of estimation and prove that the estimation may not even be consistent in this case.

8.1 Approximate Aitken's Theorem for Strategic Linear Regression

8.1.1 Extension of the Non-Cooperative Game to Linear Unbiased Estimators

Suppose now that the data analyst uses a linear unbiased estimator $\hat{\beta}_L$ which may depend on the data points and precision. Similarly to the model introduced in Section 6.2, we define a game Γ_L in which each agent i chooses her λ_i to minimize her cost; this time, however, the estimation cost depends on the variance of $\hat{\beta}_L$. A natural question to ask is the following: it is possible that, despite the fact that the

analyst uses an estimator that is "inferior" to $\hat{\beta}_{\texttt{GLS}}$ in the BLUE sense, an equilibrium reached under $\hat{\beta}_L$ is *better* than the equilibrium reached under $\hat{\beta}_{\texttt{GLS}}$ in terms of equilibrium estimation cost? If so, despite the Aitken theorem, the data analyst would have an incentive to use $\hat{\beta}_L$ instead and to inform the agents that she will use $\hat{\beta}_L$ and not $\hat{\beta}_{\texttt{GLS}}$.

In this section, we provide both a positive and a negative answer to this question, depending on specific assumptions on the disclosure costs and the class of estimators considered. Formally, we consider the game Γ_L defined as in Section 6.2, except that the estimation cost depends on the covariance of the estimator. This defines the following public cost: $C^L_{\text{estim}}(\boldsymbol{\lambda}) = F(\mathbb{E}_{\mu_{\text{joint}}}[V_L(\boldsymbol{\lambda})])$. We study a class of estimators satisfying the following assumptions:

Assumption 8.1. *$\hat{\beta}_L$ is linear and unbiased.*

Assumption 8.2. *$\mathbb{E}_{\mu_{\text{joint}}}[V_L(\boldsymbol{\lambda})]$ is convex and for all $a > 0$, we have $\mathbb{E}_{\mu_{\text{joint}}}[V_L(a\boldsymbol{\lambda})] = \frac{1}{a}\mathbb{E}_{\mu_{\text{joint}}}[V_L(\boldsymbol{\lambda})]$.*

The first assumption simply states that we consider the same type of estimators as GLS with the suitable properties of linearity of the absence of bias. Note that we consider only unbiased estimators here as our measure of quality is the covariance of the estimator. Biased estimators would allow us to trivially set this covariance to 0 by simply considering a constant estimator. We would then have to include the bias in our measure of quality, which is not practical as it is not possible to estimate the bias without access to the true parameters of the problem.

The second assumption ensures that the estimator used follows suitable statistical properties. In particular, the way its covariance depends on $\boldsymbol{\lambda}$ follows the same convexity and multiplicative properties as GLS. Intuitively, it means that the analyst does not break the structure of information they receive with their estimator. Indeed, recall that GLS is BLUE and its covariance thus represents an inherent quantity of information of the data set (it is a special case where the Cramér-Rao bound is tight and corresponds to the covariance of GLS). This quantity of information fits our Assumption 8.2 and an estimator exploiting the data set in a coherent fashion should keep this structure. These assumptions hold for the following examples:

1. The Generalized least squares estimator where

$$V_{\texttt{GLS}}(\boldsymbol{\lambda}) = \left(\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N}\lambda_i(x_i)x_i x_i^\top\right]\right)^{-1}. \tag{8.1}$$

2. The ordinary least squares estimator where

$$V_{\texttt{OLS}}(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\left(\sum_{i\in N} x_i x_i^\top\right)^{-1} \sum_{i\in N} \frac{x_i x_i^\top}{\lambda_i(x_i)} \left(\sum_{i\in N} x_i x_i^\top\right)^{-1}\right]. \tag{8.2}$$

3. Estimators which can be written $L = L_{\texttt{GLS}} + \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} \begin{bmatrix} d_1(x_1)^\top \\ \vdots \\ d_n(x_n)^\top \end{bmatrix}\right]$ with $d_i(x_i) \in \mathbb{R}^d$ and $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} d_i(x_i)^\top x_i\right] = 0$. In this case, we have

$$V_L(\boldsymbol{\lambda}) = V_{\texttt{GLS}}(\boldsymbol{\lambda}) + \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} \frac{d_i(x_i) d_i(x_i)^\top}{\lambda_i(x_i)}\right]. \tag{8.3}$$

Then, Γ_L is still a potential game with potential function $\phi_L(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{j=1}^n c_j(\lambda_j(x))\right] + C^L_{\text{estim}}(\boldsymbol{\lambda})$. Note that this assumes that agents believe that the analyst will use the estimator $\hat{\beta}_L$ as otherwise it will be common knowledge that the estimation cost is $C_{\text{estim}}(\boldsymbol{\lambda})$. This is not trivial as once the analyst holds the data point and the precision, the GLS estimator is always better than any other estimator. This can be done through trusted third parties for example but this issue emphasizes the need for the analyst to use "reasonable" estimators which justifies our assumptions.

This potential function has the same form as the potential of the original game, given by (7.1) and the same convexity properties by Assumption 8.2. Since the proof of Theorem 7.2 mostly relies on the convexity of the potential, a straightforward adaptation yields the following result.

Theorem 8.1. *Under Assumptions 6.1, 6.2 and 6.3, for any linear estimator L satisfying Assumptions 8.1 and 8.2, a precision profile $\boldsymbol{\lambda}^*$ is a non-trivial Nash equilibrium of the linear regression game if and only if it minimizes ϕ_L. 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 $C^L_{\text{estim}}(\boldsymbol{\lambda}^*)$ does not depend on the equilibrium.*

In what follows, we denote a non-trivial equilibrium of Γ_L by $\boldsymbol{\lambda}^*_L$ and we denote by $\boldsymbol{\lambda}^*_{\texttt{GLS}}$ a non-trivial equilibrium of the game Γ with the same parameters except that GLS is used as estimator.

8.1.2 Near Optimality of the Generalized Least Squares Estimator

For a given linear unbiased estimator L, the estimation cost at equilibrium is $C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L)$. We say that a linear estimator is efficient if it provides a small estimation cost at equilibrium. In the following theorem, we provide both a negative and a positive result about the efficiency of GLS: on the one hand, GLS is not always the most efficient estimator; on the other hand, under Assumptions 6.3, 6.4, 8.1, and 8.2, the ratio between the estimation cost at equilibrium of GLS and any other estimator is bounded by $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$; in order words, GLS is never too far from the most efficient estimator.

Theorem 8.2. *Assume that the disclosure cost and scalarization functions satisfy Assumptions 6.2 and 6.3. Then: (i) There exists a game Γ such that* GLS *is not the most efficient estimator; i.e., there exists an unbiased linear estimator L such that, for these game parameters,*

$$C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L) < C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}).$$

(ii) For all games that additionally satisfy Assumptions 6.4, GLS *is $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$-optimal, i.e., for all unbiased estimators L satisfying Assumptions 8.1, and 8.2*

$$C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) \leq \frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})} C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L).$$

The proof is provided in Appendix D.1. Note that the bound in Theorem 8.2*(ii)* is clearly smaller than or equal to $p_{\max}/p_{\min}$. By remarking that it can be written as $(1+\frac{q}{p_{\min}})/(1+\frac{q}{p_{\max}})$, it is easy to see that it is also smaller than or equal to $1+q$. This shows that GLS is $p_{\max}/p_{\min}$-optimal for any q and $(1+q)$-optimal for any $p_{\min}, p_{\max}$. Note also that Theorem 8.2*(ii)* trivially implies the following:

Corollary 8.1. *Under Assumptions 6.2 and 6.3, Assumption 6.4 with $p_{\min} = p_{\max} = p$,* GLS *is the most efficient estimator among estimators satisfying Assumptions 8.1 and 8.2.*

Note that $p_{\min} = p_{\max} = p$, which literally translates to $c_i(a\lambda) = a^p c_i(\lambda)$ for all $i \in N$, $\lambda \in \mathbb{R}_+$ and $a \geq 1$, means that all agents have monomial costs functions with the same exponent. Put differently, for all i, there exists a constant $r_i > 0$ such that $c_i(\lambda_i) = r_i \lambda_i^p$. Theorem 8.2*(i)* may seem counter-intuitive as GLS is optimal in the case of non-strategic agents: by Aitken's theorem, if precision are fixed and known, then the best linear unbiased estimator is GLS, i.e., for all $\boldsymbol{\lambda}$: $C^L_{\text{estim}}(\boldsymbol{\lambda}) > C_{\text{estim}}(\boldsymbol{\lambda})$. Our result demonstrates that this is not the case with strategic agents.

Numerical Illustration of the Non-Optimality of GLS

The proof of the non-optimality of GLS (Theorem 8.2(i)) is a constructive proof that uses a counter-example with two agents in a one-dimensional model ($d = 1$) where both agents have the same public data. This raises the question of whether the sub-optimality of GLS arises in higher dimensions or, more generally, in more complicated scenarios. Although extending our analytical proof to more general cases appears to be difficult, in this section, we provide three numerical counter-examples that illustrate the gap of sub-optimality of GLS. In particular, our numerical counter-examples suggest that the sub-optimality of GLS is not limited to the simple counter-example of our analytical proof.

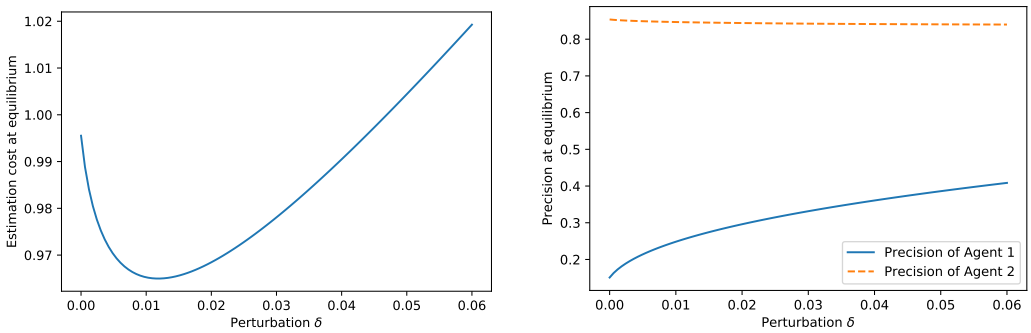
These counter-examples are constructed by using an estimator $L(\delta)$ equal to GLS plus a small perturbation term of the form δ times $D^\top$, i.e.,

$$L(\delta) = \texttt{GLS} + \delta D^\top \equiv (X^\top \Lambda X)^{-1} X^\top \Lambda + \delta D^\top,$$

for an appropriately selected D. The idea behind our counter-examples is that when using a perturbed estimator (with perturbation $\delta > 0$), that is less accurate than GLS under non-strategic agents, some agents will tend to choose a higher precision than under GLS at equilibrium. In all of our numerical examples, a small enough δ leads to an estimation cost at equilibrium smaller than the one of GLS because some agents will use a higher precision. When δ increases too much, the gain brought by the higher precision of agents is canceled by the loss of precision that is caused by using the estimator $L(\delta)$ that is less precise than GLS.

In all of our examples, the equilibrium costs of the estimators are very close to that of GLS and our examples are far from attaining the bound $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$ provided by Theorem 8.2. We believe that this bound is loose and can probably be refined.

We present three examples because each example is of independent interest. The first two involve 1-dimensional models ($d = 1$). In the first example, we use a perturbation term that affects all agents. For this example, we believe that GLS is sub-optimal only when the two exponents $p_{\min}$ and $p_{\max}$ are significantly different. In the second example, we use a perturbation that only affects two "less generous" agents. This allows us to build a counter-examples with similar disclosure costs (with exponents $p_{\min} = 1.01$ and $p_{\max} = 1.1$). Our third example includes several counter examples in settings for different values $d \geq 2$. This setting has d symmetrical agents and a single $(d+1)$-th agent whose public vector $\boldsymbol{x}_{d+1}$ is significantly different.



(a) Estimation cost (b) Precision of agents

Figure 8.1.: Counter-example 1: Estimation cost and precision of agents as a function of the perturbation δ.

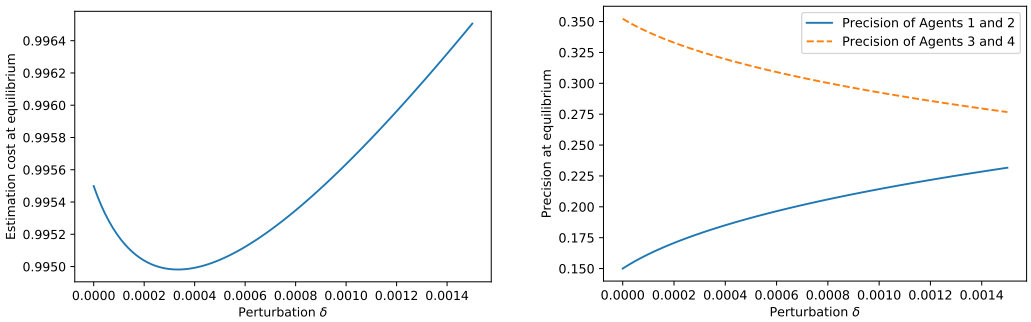
Example 1 (1-dimensional model with two agents). We consider a 1-dimensional model ($d = 1$) with two agents ($n = 2$) in which the public data of each agent is $x_i = 1$. For such a game, the estimator GLS is $(X^\top \Lambda X)^{-1} X^\top \Lambda \tilde{\boldsymbol{y}} = (\lambda_1 + \lambda_2)^{-1} \boldsymbol{\lambda}^\top \tilde{\boldsymbol{y}}$ and its covariance is $1/(\lambda_1 + \lambda_2)$. We consider a linear estimator $L(\delta)$ of the form

$$\texttt{GLS} + \begin{bmatrix} \sqrt{\delta} \\ -\sqrt{\delta} \end{bmatrix} = \begin{bmatrix} \lambda_1/(\lambda_1+\lambda_2)+\sqrt{\delta} \\ \lambda_2/(\lambda_1+\lambda_2)-\sqrt{\delta} \end{bmatrix}.$$

According to (8.3), its covariance is $1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$, where $\delta/\lambda_1 + \delta/\lambda_2$ is the loss of precision due to using a linear estimator that is less precise than GLS. We assume that the disclosure cost of Agent 1 is $c_1(\lambda) = \lambda^{1.01}$ ($p_{\min} = 1.01$) while the disclosure cost of Agent 2 is $c_2(\lambda) = \lambda^{20}$ ($p_{\max} = 20$). The scalarization function is the identity, which means that $C^{L(\delta)}_{\text{estim}}(\boldsymbol{\lambda}) = 1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$.

In Figure 8.1(a), we plot the estimation cost at equilibrium $C^{L(\delta)}_{\text{estim}}(\boldsymbol{\lambda}^*_{L(\delta)})$ as a function of δ. We observe that with GLS we get an estimation cost of approximately 0.99. When δ increases, the estimation cost at equilibrium decreases up to $\delta = 0.012$ for which it reaches approximately 0.96. This decrease is explained by the fact that for small δ, the gain due to a higher precision used by Agent 1 is larger than the loss of precision $\delta/\lambda_1 + \delta/\lambda_2$. When δ exceeds 0.012, this loss of precision is more important than the gain due to higher precision. This behavior is further illustrated in Figure 8.1(b), where we plot the precision released by the two agents. We observe that the precision of Agent 1 increases with δ while the precision of Agent 2 decreases (slightly).

Example 2 (1-dimensional model with four agents). We consider a 1-dimensional game with four agents in which the public data of each agent equals $x_i = 1$. Agents 1 and 2 have disclosure costs $c_i(\lambda) = \lambda^{1.01}$ while Agents 3 and 4 have disclosure



(a) Estimation cost (b) Precision of agents

Figure 8.2.: Counter-example 2: Estimation cost and precision of agents as a function of the perturbation δ.

costs $c_i(\lambda) = \lambda^{1.1}$. We consider a linear unbiased estimator that is equal to GLS plus a perturbation cost that only affects the first two agents: $D = [\sqrt{\delta}, -\sqrt{\delta}, 0, 0]$. Note that this perturbation is only applied to the most selfish agents as they are the ones we must incentivize to give more.

In Figure 8.2, we plot the estimation cost at equilibrium $C_{\text{estim}}^{L(\delta)}(\boldsymbol{\lambda}_{L(\delta)}^*)$ as a function of δ. With GLS ($\delta = 0$), we get an estimation cost of 0.9955, which is larger than the value 0.9950 that we obtain for $\delta = 3.10^{-4}$. As for Example 1, when δ increases, the precision used by the least generous agents (Agents 1 and 2) increase while the precision of the most generous agents decrease.

While the previous two counter-examples are in dimension 1 and with agents that all have $x_i = 1$, the sub-optimality of GLS is not limited to that case. To illustrate that, we consider in the next counter-example models in dimension d with $d \geq 2$. Note that, as we assume that matrix X has rank d, we need at least d players whose feature vectors $\boldsymbol{x}_i$'s span the d dimensions. Note also that, with d players in d dimensions, GLS is the only linear unbiased estimator. Indeed, as matrix X would then be invertible, the non-bias condition leads to $D^\top = 0$. In Example 3 below, we consider the simplest case of models with $d+1$ agents, though it is clear that one could construct similar counter examples with any number of agents larger than or equal to $d+1$.

Example 3 (d-dimensional models with $d+1$ agents). We consider a d-dimensional game with $d+1$ agents. The public data of the first d agents spans the d dimensions: $\boldsymbol{x}_i$ is a vector where all components equal 0 except the ith one that is equal to 1. All components of the public data of Agent $d+1$ are equal to $1/d$: $\boldsymbol{x}_{d+1} = [1/d, \cdots, 1/d]^\top$. We assume that the disclosure costs of the first d agents are $c_i(\lambda) =$

λ^{20} (for $i \in \{1, \cdots, d\}$), and the disclosure cost of the last agent is $c_{d+1}(\lambda) = \lambda^{1.5}$. $1/\sigma^2 = 1$.

The perturbation matrix D is a $(d+1)\times d$ matrix whose first column is $\sqrt{\delta}[1, \cdots, 1, -d]$, all other entries being 0. Hence, the public feature matrix X and the perturbation matrix D are the following $(d+1) \times d$ matrices:

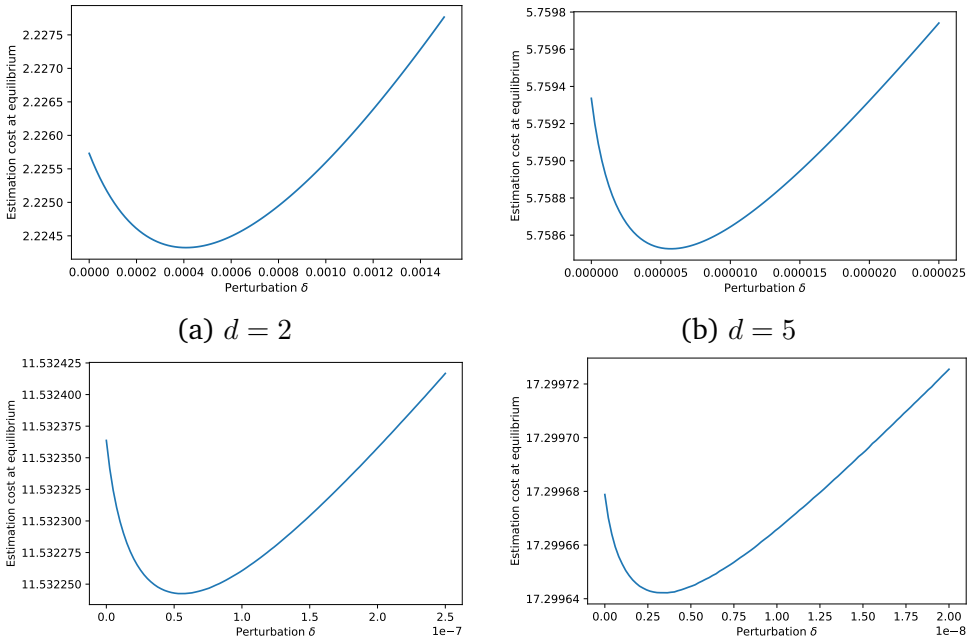
$$X = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ 1/d & \dots & 1/d \end{bmatrix}, \qquad D = \begin{bmatrix} \sqrt{\delta} & 0 & 0 & \dots \\ \vdots & 0 & 0 & \dots \\ \sqrt{\delta} & 0 & 0 & \dots \\ -d\sqrt{\delta} & 0 & 0 & \dots \end{bmatrix}. \tag{8.4}$$

It is easy to verify that $D^\top X = 0$, which implies $L(\delta) = \texttt{GLS} + D^\top$ is an unbiased estimator.

In Figure 8.3, we report the estimation cost at equilibrium $C_{\text{estim}}^{L(\delta)}(\boldsymbol{\lambda}^*_{L(\delta)})$ as a function of δ. We consider models of dimension $d \in \{2, 5, 10, 15\}$. We observe that for all dimensions d, the behavior is similar to the one observed in Figure 8.1(a) and 8.2(a): when δ is small enough, using the estimator $L(\delta)$ provides a higher precision at equilibrium (i.e., a lower equilibrium estimation cost as seen on the graphs). This comes from the fact that when δ increases, the precision at equilibrium provided by Agent $d+1$ increases with δ whereas the precision provided by Agents 1 to d is almost independent of δ. When δ increases too much, the estimation cost increases again because of the non-optimality of the estimator $L(\delta)$ (for given individual precisions). We also observe that the maximal gain that can be obtained by using an estimator other than `GLS` (and the perturbation δ for which it is achieved with our particular perturbation matrix D) seems to decrease when the dimension d increases.

Finally, although the public feature matrix X in (8.4) has a particular form, many d-dimensional models with $d+1$ agents can be cast in this model via an appropriate change of basis. In fact, we conjecture that for any matrix of public features X with at least $d+1$ agents, there exist disclosure costs such that `GLS` is not optimal.

Note that if we do note restrict the analyst to estimators satisfying Assumption 8.2, the linear regression game can produce arbitrary results as illustrated by the following examples. Consider a linear regression game with two agents, $\mathcal{X} = \{-1, 1\}$ and $\mu_{\text{joint}}(-1, 1) = 1$ (i.e. agent 1 holds data point -1 and agent 2 holds data point 2). Agent 1 has provision cost $c_1(\ell) = c_2(\ell) = \ell^p$. The scalarization is the trace



(a) $d = 2$ (b) $d = 5$

(c) $d = 10$ (d) $d = 15$

Figure 8.3.: Counter-example 3: Estimation cost and precision of agents as a function of the perturbation δ for models in dimension $d \geq 2$.

(corresponding to the identity function in this 1 dimensional case). The analyst uses the estimator $L = L_{\texttt{GLS}} + \begin{pmatrix} d(\lambda_1(-1), \lambda_1(1)) \\ d(\lambda_1(-1), \lambda_1(1)) \end{pmatrix}$ where

$$d(\lambda_1(-1), \lambda_1(1)) = \begin{cases} D & \text{if } \min(\lambda_1(-1), \lambda_2(1)) \leq M, \\ 0 & \text{otherwise.} \end{cases}$$

The potential in this game is then:

$$\phi_L(\boldsymbol{\lambda}) = \begin{cases} \lambda_1(-1)^p + \lambda_2(1)^p + \frac{1}{\lambda_1(-1)+\lambda_2(1)} & \text{if } \min(\lambda_1(-1), \lambda_2(1)) \geq M, \\ \lambda_1(-1)^p + \lambda_2(1)^p + \frac{1}{\lambda_1(-1)+\lambda_2(1)} + \frac{D^2}{\lambda_1(-1)} + \frac{D^2}{\lambda_2(1)} & \text{otherwise.} \end{cases}$$

It is then easy to see that for any M, there exists a value of D high enough such that the unique Nash equilibrium is $\lambda_1(-1) = \lambda_2(1) = M$. In this case, the estimation cost is $1/2M$ and the analyst can pick an estimator producing an arbitrarily low estimation cost. This, however, relates to our earlier discussion about the fact that the agents must believe that the analyst will use the estimator $\hat{\beta}_L$ which is impractical in this case. Indeed, in this example, it would mean that the analyst voluntarily ruins the estimation if one of the data points received is not precise enough.

8.2 Asymptotic Degradation of Estimates

Now that we established that the GLS estimator is approximately optimal among a class of estimators satisfying suitable statistical properties, we shift our focus from comparing estimators between them to assessing the quality of a single estimator. Specifically, we characterize the quality of an estimator by the convergence rate of its estimation cost. We work on the GLS estimator (and sometimes extend to OLS) as our near-optimality result implies that its convergence rate is optimal.

The quality of estimation is influenced by two aspects in our model. First, the total precision of data points gathered. If agents are willing to provide more precise data points the covariance will shrink faster. Then, the allocation of precision of data points: for a given total precision the quality of estimation may vary depending on which data points are precise and which are not. This is usually studied in optimal design and we draw a parallel with this field in this section.

We show that, surprisingly, in our model the cases corresponding to optimal allocation of precision also correspond to worst-case for the total precision of data points. We show that these worst-case scenarios are caused by linearity of the provision costs and that in these settings the GLS estimator even fails to be consistent. On the contrary, when provision costs are not linear, we retrieve the consistency of GLS but the convergence rate of the covariance is degraded with a degree of degradation depending on how close the provision costs are to being linear and how convex the scalarization is.

8.2.1 Link With Optimal Design

In optimal design (see Pukelsheim [Puk06], Atkinson et al. [ADT07], and Boyd and Vandenberghe [BV04] for a detailed introduction), 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 and the design space $\mathcal{X}$ is to find a probability measure ν^* that minimizes:

$$\nu^* \in \arg\min_{\nu} F\left(\left(\sum_{x\in\mathcal{X}} xx^\top \nu(x)\right)^{-1}\right). \tag{8.5}$$

Classical scalarizations used in optimal design are the trace (A-optimal design) or the mean squared error (I and V-optimal design) (refer to Boyd and Vandenberghe [BV04] Chapter 7 for nomenclature.).

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 (7.1) that contains the estimation cost $C_{\text{estim}}(\boldsymbol{\lambda})$, which can be rewritten as:

$$C_{\text{estim}}(\boldsymbol{\lambda}) = F\left(\left(\sum_{x\in\mathcal{X}} xx^\top \sum_{i\in N} \lambda_i(x)\mu(x)\right)^{-1}\right). \tag{8.6}$$

The similarity between (8.5) and (8.6) suggests a potential link between the Nash equilibria of the linear regression game and the solutions of the optimal design problem on $\mathcal{X}$ by interpreting $\sum_i \lambda_i(x)\mu^i_{\text{marg}}(x)$ as a design $\nu(x)$:

Theorem 8.3. *Consider a linear regression game that satisfies Assumptions 6.1 and 6.3 and such that all provision costs are linear (*i.e.*, $c_i(\ell) = a\ell$ for all $i \in N$ and $\ell \in \mathbb{R}_+$, where a is a constant). Let $\boldsymbol{\lambda}^*$ be a non-trivial Nash equilibrium and let $\nu_{\boldsymbol{\lambda}^*}$ be the measure such that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_{i\in N} \lambda^*_i(x)\mu^i_{marg}(x)$ 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* (8.5).

Sketch of proof. A detailed proof is given in Appendix D.2. The main idea is to see the minimization problem (8.5) 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 6.3). □

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 8.3 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 8.2.2 (in particular Theorem 8.4). Note that this theorem implies that there is a heavy free-riding aspect to this game. Indeed, if we consider the complete information setting, agents which have data points x which do not belong to an optimal design are pure free-riders. We illustrate in the next section however that this free-riding characteristic disappears when

data provision costs become superlinear. In this case, the difference between the maximum and minimum precision given depending on the data point shrinks.

General data provision costs

The particular connection between optimal design and Nash equilibria exhibited in Theorem 8.3 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 sub-optimal. 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 Theorem 8.3 in a polynomial regression setting that is an instance of our linear regression game as follows. Let $\mathcal{X} = \{[1, x, \cdots, x^{d-1}]^\top, x \in [\![-10 \dots 10]\!]\}$ be the set of attribute vectors. In these examples, we consider settings where the data point of each agent is independently produced from a common distribution μ. We compare in Figure 8.4 the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 8.3 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 below). 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 $[1, 0, \dots, 0]$ 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.

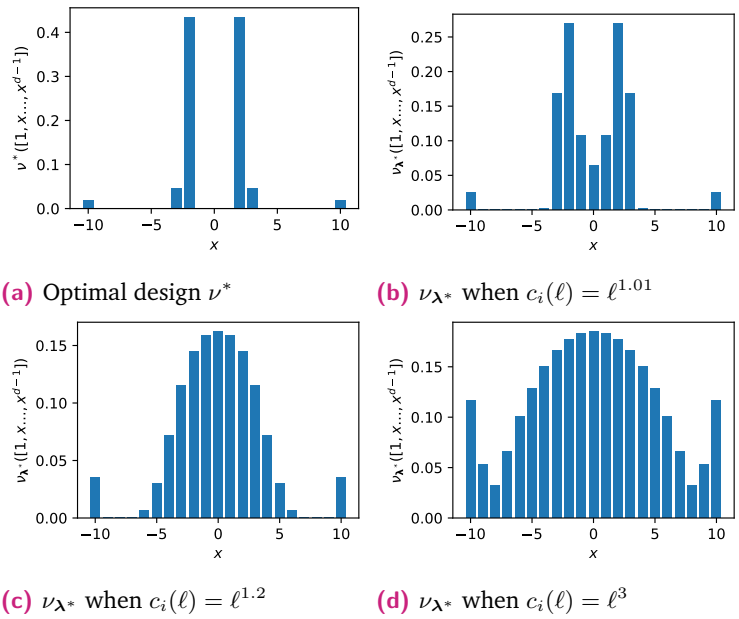
We now extend our previous observations to the following more general settings:

(a) In Figure 8.5, we vary the degree d of the polynomial regression (Figure 8.4 has $d = 4$).

(b) In Figure 8.6, we vary the distribution μ (Figure 8.4 has a uniform distribution that corresponds to the first row in Figure 8.6). Here, we fix $d = 4$ and we do not plot the optimal design as it does not depend on μ.

(c) In Figure 8.7, 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 8.5 illustrates the optimal design ν^* and the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 8.3 in the same setting as Figure 8.4 ($d = 4$)



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

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 8.6 illustrates the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 8.3 in the same setting as Figure 8.4 with various distributions μ of the agents' x_i vectors. The first row of graphs corresponds to the exact same setting as Figure 8.4 (uniform distribution) while the next rows show the results for other distributions. In addition to Figure 8.4, 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 8.4). 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 8.6 (compared to the first), we observe that $\nu_{\boldsymbol{\lambda}^*}([1, x, \cdots, x^{d-1}])$ shrinks for values of x close to 0.

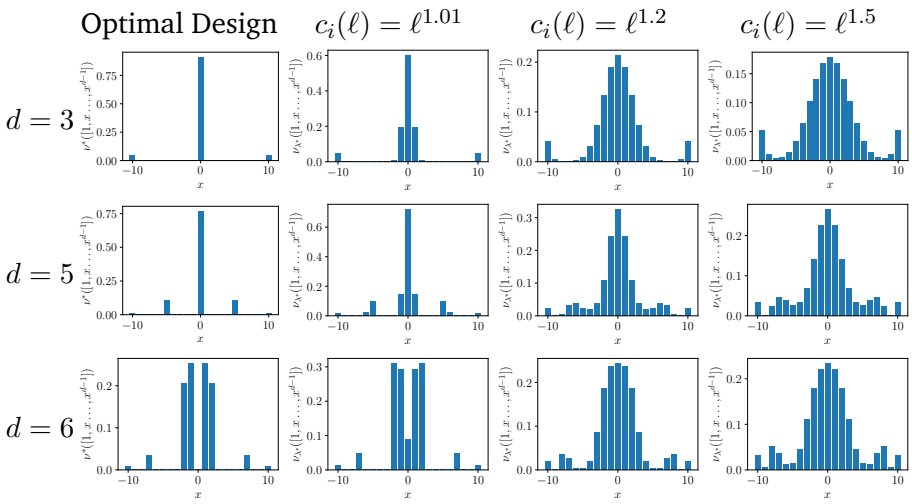


Figure 8.5.: Optimal design ν^* and allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with various degrees d of the polynomial regression (here μ is uniform and the scalarization is the trace as in Figure 8.4).

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 8.6 where $\nu_{\boldsymbol{\lambda}^*}$ has a shape similar to the first row with values skewed to the left where vectors have higher probability.

Figure 8.7 illustrates the optimal design ν^* and the allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ as defined in Theorem 8.3 in the same setting as Figure 8.4 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 8.7 with the squared Frobenius norm exhibits the same behaviors as discussed before on Figure 8.4 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 $[1, 0, \dots, 0]$ is maximal (instead of zero in the optimal design).

8.2.2 Bounds on the Estimation Cost

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*.

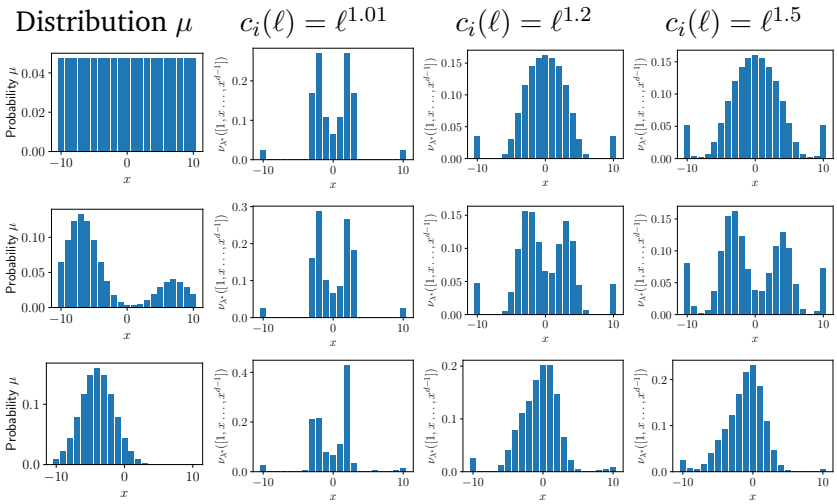


Figure 8.6.: Allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with various distributions μ (here $d = 4$ and the scalarization is the trace as in Figure 8.4). The optimal design ν^* does not depend on μ and is therefore the same as in Figure 8.4.

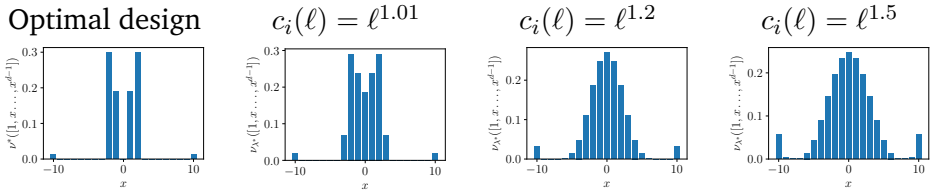


Figure 8.7.: Optimal design ν^* and allocation of precision at equilibrium $\nu_{\boldsymbol{\lambda}^*}$ with the squared Frobenius norm as a scalarization F (here μ is uniform and $d = 4$ as in Figure 8.4).

The Case of Identical Agents

To gain intuition, we begin with the case of agents with identical monomial costs and independently and identically distributed data points. In this simple 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:

Proposition 8.1. *Consider a linear regression game satisfying Assumptions 6.1 and 6.3 and such that for all agent $i \in N$ and precision $\ell \in \mathbb{R}_+$: $c_i(\ell) = \ell^p$ with $p \geq 1$. Assume that the data point of each agent is independently produced from a common distribution μ. Let $\lambda_{single} = \arg\min_{\lambda \in \mathbb{R}_+^{|x|}} \mathbb{E}_\mu [\lambda(x)^p] + C_{\text{estim}}(\lambda)$.*

(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 any non-trivial equilibrium is $C_{\text{estim}}(\boldsymbol{\lambda}^*) = n^{-q\frac{p-1}{p+q}} C_{\text{estim}}(\lambda_{single})$.

Sketch of proof. Identical agents have identical strategies at equilibrium. Thus we rewrite the potential as the potential of a single-agent game with scaled cost functions and scalarization. The full proof is given in Appendix D.3. □

Proposition 8.1(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, when the costs are linear ($p = 1$), the precision given by each agent goes to 0 at a $\Theta(1/n)$ rate.

As a consequence, when aggregating the data from n non-strategic data sources, the estimation cost would be

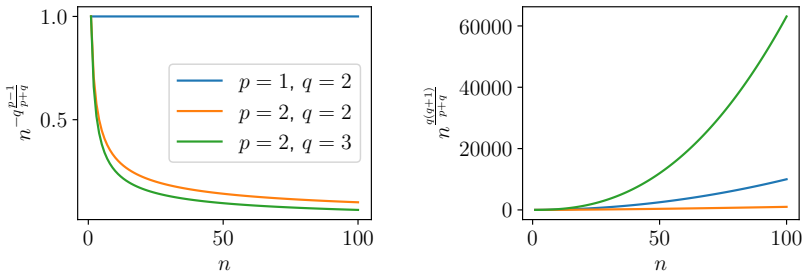
$$C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}}) = n^{-q} C_{\text{estim}}(\lambda_{\text{ns}}) \tag{8.7}$$

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 8.1(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:

$$C_{\text{estim}}(\boldsymbol{\lambda}^*)/C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}}) = \Theta(n^{\frac{q(q+1)}{p+q}}). \tag{8.8}$$

This ratio goes to infinity for any possible value of the parameters, implying in turn 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 thus 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. Note that this ratio follows variations similar to the price of anarchy (Theorem 7.3) when p and q vary. Intuitively, this corresponds to the fact that the agents' incentive to produce a high quality estimator diminishes when p diminishes (as it is costly to produce data) and when q increases (as the estimation cost shrinks



(a) Estimation cost. **(b)** Degradation ratio.

Figure 8.8.: Influence of p and q on (a) the estimation cost $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ and (b) the degradation ratio $C_{\text{estim}}(\boldsymbol{\lambda}^*)/C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}})$.

faster). Recall that the only difference between the potential and the social cost is the fact that the estimation cost has higher weight in the social cost as it is a public good. Thus, the price of anarchy and the degradation of the estimation cost *compared* to the non-strategic case are correlated. Note however that the convergence rate of the estimation cost does not follow the same variations (it increases with q) which shows that the loss of effort exerted by the agents when q is high is compensated by the more generous scalarization.

Figure 8.8 illustrates the convergence of the estimation cost and the degradation ratio for various values of p and q. Figure 8.8a 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 8.8b 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 8.8a: 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.

Asymptotic degradation of estimation cost in the general case

We are now ready to state our result characterizing 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 8.4. *Assume that Assumptions 6.1, 6.2, 6.3 and 6.4 hold. 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, for any non-trivial equilibrium $\boldsymbol{\lambda}^*$, we have:*

$$dn^{-q\frac{p_{\min}-1}{p_{\min}+q}-\alpha} \leq C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq Dn^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \tag{8.9}$$

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

Sketch of proof. A full proof is given in Appendix D.4. 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 8.1. 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 (8.9). 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 (7.1)). 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 (8.9). 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 $C_{\text{estim}}(\boldsymbol{\lambda}_{\text{ns}})$. □

First, let us note that Theorem 8.4 also holds for the `OLS` estimator. Indeed, our upper bound relies on a fixed homoskedastic strategy for which the `GLS` estimator corresponds to `OLS` and our lower bound relies on the comparison of the estimator to an optimal design which still holds for `OLS`. Our analysis for `OLS` reveals however 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 all this in detail in the Section 8.2.3. Theorem 8.4 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.

In this degree of generality, it is no longer possible to express the equilibrium precision in closed form (as in Proposition 8.1). Nevertheless, Theorem 8.4 shows that we are able to provide precise bounds for the estimation cost.

The bounds we obtained depend on several factors. First, the convergence rates depend only on $p_{\min}$, $p_{\max}$ and q. These convergence rates are discussed in detail in the next paragraphs. Then, the multiplicative factors depend on all parameters of the problem and in particular on the correlation between the different data points. More precisely, the multiplicative factor of the upper bound is

$$D = \left(c_{\max}(1) + F((\mathbb{E}_{\mu_{\text{joint}}} \left[\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right])^{-1}) \right).$$

This term depends on $F((\mathbb{E}_{\mu_{\text{joint}}} \left[\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right])^{-1})$ which precisely captures the impact on correlation on the estimation cost when agents allocate precision uniformly. Note that the correlation between data points may be arbitrarily bad leading to this term being arbitrarily high but there is a lower bound on its value given by the solution of the optimal design problem on $\mathcal{X}$. It also depends on the provision costs through the term $c_{\max}$ which intuitively represents the multiplicative factor of the provision costs. Indeed, if we consider a monomial cost $c(\ell) = a\ell^p$, the exponent of the monomial influences the convergence rate while the constant a influences the multiplicative factor of the estimation cost. The multiplicative factor of the lower bound is

$$d = F\left(\left(\sum_{x \in \mathcal{X}} x x^\top \nu^*(x) \right)^{-1} \right) \ell_{\max} \left(1 + \frac{F((\mathbb{E}_{\mu_{\text{joint}}} \left[\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right])^{-1})}{c_{\min}(\ell_{\max})} \right)^{-\frac{q}{p_{\max}}}.$$

Where ν^* is an optimal design on $\mathcal{X}$ considering the scalarization F and $\ell_{\max}$ is an upper bound on the precision given by a single agent. This factor depends on the parameter of the problem similarly to the factor of the upper bound with a new dependency on the optimal design. Intuitively, the optimal design appears in the bound as a limit to how efficient the allocation of precision by agents can be. Recall that in Section 8.2.1, we showed that agents allocate precision optimally when their provision cost is linear and more uniformly when their provision cost is highly convex. Thus, for linear provision cost, the dependency on ν^* is likely to be closer to the actual provision cost while for highly convex provision cost, the factor $F((\mathbb{E}_{\mu_{\text{joint}}} \left[\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right])^{-1})$ of the upper bound representing uniform precision allocation is likely to be more relevant.

An important special case to consider is the independently and identically distributed case where $\mathbb{E}_{\mu_{\text{joint}}} \left[\frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \right] = \mathbb{E}_{\mu} \left[x x^\top \right]$ is independent from n. In this case, the impact of the correlation between data points is constant and the only important parameters asymptotically are $p_{\min}$, $p_{\max}$ and q. We tailor the rest of the discussion to settings where data points are iid for clarity but our conclusions also apply when

there is only some kind of regularity to the distribution μ_{joint} when the number of agents grow so that $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$ does not vary too much. We believe that pathological cases where $\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i\in N} x_i x_i^\top\right]$ varies in such a way that GLS is no longer consistent of simply varies in unpredictable way are not relevant to the application we target and no result could be obtained without further assumptions.

The upper bound in (8.9) 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 (8.9) shows that the convergence rate is slower that the standard rate of $\Theta(n^{-q})$ (or $\Theta(1/n)$ for scalarizations with $q = 1$).

The consistency of GLS when $p_{\min} > 1$ can be interpreted economically through the first order conditions. At equilibrium, agents will invest so that their marginal benefit equals marginal cost. If the derivative of costs is 0 at 0 effort, then agents will invest positive effort even in the limit as n grows large, as long as the model isn't perfectly learned. We thus always learn at equilibrium when n grows. On the other hand, if the derivative of the cost function at 0 is positive, then after a certain amount of data is collected no agent will be willing to invest non-zero effort, since the costs for doing so outweigh the benefit.

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 (8.9) coincide and are equal to the exponent of Proposition 8.1. When $p_{\min}$ and $p_{\max}$ are different, the bounds loosen. Intuitively, the upper bound then involves the parameter $p_{\min}$ because, when precision 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 perform below a numerical investigation of the result of Theorem 8.4 illustrating the lower and upper bounds.

To illustrate our results, we consider a one-dimensional model with $\mathcal{X} = \{1\}$. The scalarization is the trace (which satisfies Assumption 6.3 with $q = 1$). This means that $C_{\text{estim}}(\boldsymbol{\lambda}) = (\sum_i \lambda_i(1))^{-1}$. Recall that Theorem 8.4 shows that

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

The goal of this section is to compare the upper and lower bounds of Theorem 8.4 to $C_{\text{estim}}(\boldsymbol{\lambda}^*)$, 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 $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ 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 8.4);

(b) $n^{-q\frac{p_{\min}-1}{p_{\min}+q}}$ (the rate of the upper bound of Theorem 8.4, 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 $c_i(\ell) = \ell^{p_{\max}}$).

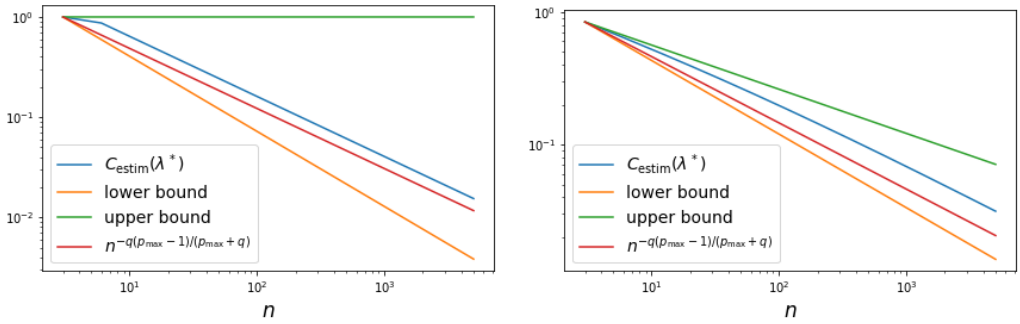
Note that (a) is the fastest convergence rate, followed by (c) and then by (b).

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 $C_{\text{estim}}(\boldsymbol{\lambda}^*)$).

Illustration: Heterogeneous Agents With Different Exponents

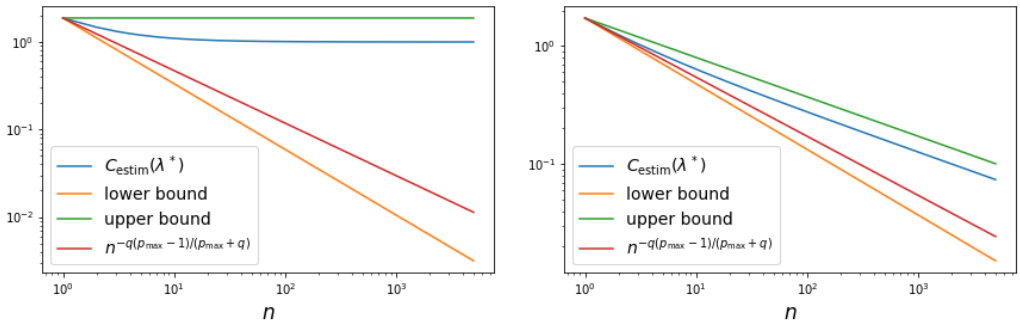
We first consider heterogeneous agents. For a given n, $n/3$ agents have provision costs $c_i(\ell) = \ell^{p_{\max}}$ and $2n/3$ agents have provision costs $c_i(\ell) = \ell^{p_{\min}}$. This setup satisfies the assumptions of Theorem 8.4 with the corresponding $p_{\min}$ and $p_{\max}$. We consider two setups: $(p_{\min}, p_{\max}) = (1, 4)$ and $(p_{\min}, p_{\max}) = (2, 3)$.

Figure 8.9 compares the convergence rate of $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ to the three bounds defined above. This figure suggests that the estimation cost behaves as when all players have an 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 $C_{\text{estim}}(\boldsymbol{\lambda}^*)$ is driven by agents having exponent $p_{\max}$.



(a) Comparison for $p_{\min} = 1$ and $p_{\max} = 4$ **(b)** Comparison for $p_{\min} = 2$ and $p_{\max} = 3$

Figure 8.9.: Comparison of the rate of convergence of the estimation cost with different bounds for agents with heterogeneous costs



(a) Comparison for $p_{\min} = 1$ and $p_{\max} = 4$ **(b)** Comparison for $p_{\min} = 2$ and $p_{\max} = 3$

Figure 8.10.: Comparison of the rate of convergence of the estimation cost with different bounds for agents with polynomial costs

Illustration: Agents With Polynomial Provision Costs

We then consider agents with polynomial provision costs. We assume that the n agents have the same provision cost $c_i(\ell) = \sum_{k=p_{\min}}^{p_{\max}} \ell^k$. Again, these provision costs satisfy the assumptions of Theorem 8.4 with the corresponding $p_{\min}$ and $p_{\max}$.

Figure 8.10 compares the convergence rate of the covariance to the upper and lower bounds of Theorem 8.4. 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 $p_{\min}$.

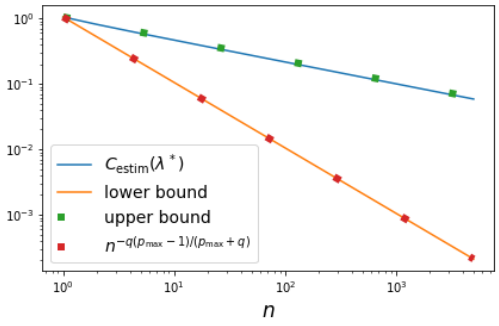


Figure 8.11.: Comparison of the rate of convergence of the estimation cost with the upper bound of Theorem 8.4 for agents with hyperbolic cosine costs.

Illustration: 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 costs 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 8.11 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 satisfies 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 8.11 suggests indeed that the convergence rate is close to this upper bound.

8.2.3 The OLS Estimator Suffers From a Single Arbitrarily Bad Provision Cost While the GLS Estimator does not

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

Proposition 8.2. *Assume that Assumptions 6.1, 6.3 and 6.2 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, for any non-trivial equilibrium, we have:*

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \leq D|S_N|^{-q\frac{p_{\min}-1}{p_{\min}+q}}, \tag{8.10}$$

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 Appendix D.4. □

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 $\boldsymbol{\lambda}, \lambda_{n+1}$, 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 $\{1, \dots, n\}$ and $c_{n+1}(\lambda) = (n+1)^2\lambda$. In the following game, we also consider the scalarization $F(\cdot)$ to be the trace which in this case is the identity function. We have in this game the following potential:

$$\phi_{\texttt{OLS}}(\boldsymbol{\lambda}) = \sum_{i=1}^n \lambda_i^p + (n+1)^2\lambda_{n+1} + \frac{1}{(n+1)^2}\sum_{i=1}^{n+1}\frac{1}{\lambda_i}. \tag{8.11}$$

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

$$C_{\text{estim}}^{\text{OLS}}(\boldsymbol{\lambda}^*) = \frac{1}{(n+1)^2} n (n+1)^{2/(p+1)} + 1. \tag{8.12}$$

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 8.2 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.

Summary: In this chapter, we characterized some statistical properties of linear estimation when data points are strategically produced. We showed that, while Aitken's theorem does not hold in this setting, an approximate version holds among a class of linear unbiased estimators satisfying suitable assumptions. We then showed that when agents have linear provision cost, they allocate precision optimally among data points. We finally showed that the convergence rate of the estimation cost worsens when both the provision cost and the estimation cost become more linear. Importantly, the worst-case is attained for linear provision costs which make `GLS` no longer satisfy the key *consistency* property. There is thus a conflict between the *quantity* of precision produced by agents (which lessens when provision costs become linear) and the *quality* of the allocation of precision (which is optimal only if provision costs are linear).

Part III

Conclusions and Perspectives

9 Conclusion and Future Works

In this part, we present a high level overview of the results obtained in this thesis and then discuss several open challenges offering direction for future work related to the content of this thesis.

9.1 Conclusion

In this thesis, we choose to investigate learning problems when the data generation process may depend on the analysis and its result. In particular, we studied the *classification* problem and the *linear regression* problem with two key questions to answer:

1. Do fundamental learning results still hold when the data generation process depends on the result of the analysis?

2. Can game theoretic models reasonably be applied in learning settings where there is a large quantity of available complex data?

We answered Question 1 by the negative in general showing that learning in the presence of strategic data requires careful consideration of otherwise well known properties and parameters. For classification, we showed that one can optimally defend against attacks only through a *random* defense but using *simple* classifiers which may or may not belong to classical sets such as linear classifiers. Furthermore, this dependency does not depend on parameters studied in classical learning such as the distribution of non-attacks or the probability of attack but only on the game-theoretical metric of payoffs of detected and undetected attacks. This emphasizes that defenders could potentially hugely benefit from modeling attackers rather than applying well known learning algorithms on classical hypothesis classes such as linear or kernel classifiers. When considering linear regression problems, we showed that *some* results (approximately) hold such as the optimality of `GLS` while others are compromised in many settings – the linear regression process may not yield consistent estimators if participating agents are reluctant to produce precise data (which translates mathematically into linearity of data provision costs) and even

when they are consistent, their rate of convergence worsens. In practice, this means that analysts may underestimate the number of agents needed in an experiment to achieve a certain precision. Furthermore, our results showed that if analysts have a limit on the number of participants they can recruit, they should aim to avoid agents whose cost to produce data increases linearly with the precision. This is even more relevant when analysts do not have access to the precision of each data point and use the `OLS` estimator, in which case even a single strategic agent with high provision cost may ruin the estimation. The rate of convergence of estimators is also in direct conflict with the quality of the allocation of data as we showed that agents reluctant to provide precise data are particularly careful with their usage of resources leading to optimal designs while agents willing to provide data do so more uniformly regardless of how informative their data points are.

We provided training methods and approximation methods to answer question 2. More precisely, we showed that a defender could train near optimal classifiers simply by having access to parameters describing the cost of false alarm, true positive and false negative associated to a behavior. This training can be performed both online or offline with a data set and does not require access to difficult to evaluate parameters such as the behavior of non-attackers or even the probability of attack. In practice, this means that our model can be applied similarly to classical learning model using a data set or online information. We believe that our model could potentially be applied to bank fraud settings where taking into account the weighted classification problem (where misclassifying transaction has different consequences depending on the characteristics of the transaction) is crucial. Our experiments on a real bank fraud data set (albeit with a simplistic model) also show that reasonably sized data sets may be sufficient to train defenses robust to adaptive attackers. For linear regression, we showed that agents do not need to have complete information about the data points of other agents when data points are independently and identically distributed. We showed that it is sufficient to have access to the underlying data distribution which can be reasonably estimated in many settings with publicly available data. In particular, the two previously mentioned models yield equivalent (in terms of cost) and interchangeable (in terms of strategies) equilibrium.

9.2 Future Works

Many potential future work follow the line of work of this thesis of studying fundamental properties of learning algorithms when data are strategically produced. It should first be noted that our work relies on game theory which necessitates the

definition of payoffs for the participating agents. Instantiating these payoffs in real settings to match users' perception is challenging as they may encompass subjective considerations. Thus, a first line of work may be to close this gap between theoretical payoffs and real users' perception using tools such as conjoint analysis.

We divide the rest of this section following the organization of this thesis, starting with possible problems related to classification and ending with problems related to (linear) regression.

Classification. We studied the adversarial classification problem where a defender classifies behavior between malicious or non-malicious and a behavior is malicious if it comes from an attacker. In these settings, there is no preexisting learning problem in the absence of attackers. Such problems appear in fraud or network attacks but do not cover the whole field of machine learning. In particular, the important problem of picture recognition widely studied in the literature of adversarial examples (see Goodfellow et al. [GSS15] for example) does not fit our model well. Indeed, in such applications there is a preexisting learning problem defined theoretically by an unknown data distribution which may be modified to a certain extent by an attacker. We call such models *adversarial machine learning*. Preliminary work has been done by Meunier et al. [Meu+21] who show the existence of a Nash equilibrium and Bhagoji et al. [BCM19] who exhibit a bound on the possible robustness of classifiers. The former, however, do not characterize the equilibrium and the latter do not consider a game and consider that the attacker does not pay any cost to modify data points. Thus, results about the robustness of classification in a game-theoretic setting where the modification of data points incurs a cost to the attacker are missing. Additionally, in adversarial machine learning one can consider multi-class classification settings (for example classifying a picture of an animal between different possible species) which to the best of our knowledge have not been studied. This may lead to hardness results when increasing the number of possible classes in the spirit of results of hardness to compute robust attacks of Perdomo and Singer [PS19a]. Combining the previous two points, a promising line of work would be to consider a problem where there is an underlying distribution corresponding to a multi-class classification problem. An attacker can modify this distribution to a certain extent and a defender simultaneously chooses a classifier (this corresponds to a game where the notion of equilibrium considered is the Nash equilibrium but a Stackelberg setting where the defender is the leader could be interesting. Stackelberg settings where the attacker is the leader are less interesting as the optimal defense is simply a naive Bayes classifier). The goal would be to obtain bounds on the minimum error rate a defender can achieve depending on the similarity between the different classes as

well as to produce defenses which can achieve performances close to these bounds assuming access to a training set generated with the original distribution. Finally, while we considered adversarial settings in classification, there exists a literature on strategic classification where data producers have goals which do not align but also do not necessarily oppose the classifier. This is however less related to our line of work on classification.

Linear regression. We studied the linear regression model where agents strategically chose the precision of the data points they produce to minimize a mix of their individual provision cost and a cost linked to the precision of the result of the regression. In this line of work, several different directions are possible. First, we assumed that agents reported their precision truthfully which may not be. Thus, it would be beneficial to study models and mechanisms encouraging truthful revelation of data. Then, we studied a public good model where agents have incentive to obtain a precise estimator while other models study settings where agents have incentive to obtain a specific result from the estimation. To the best of our knowledge, however, adversarial settings where agents can manipulate their data points to a certain extent and aim to maximize the error of estimation do not exist in the literature. This raises modeling challenges as the estimation error depends on both bias and variance of estimation which may be hard to evaluate. Finally, we studied a subset of parametric regression with linear regression. We believe however that some of our results (in particular the asymptotic degradation of the precision of the estimator) may be possible to extend in non-parametric settings using of bounds on the quality non-parametric of estimation as presented in Györfi et al. [Gyö+02]. In particular, algorithms such as k-nearest neighbors have error guarantees under mild assumptions on the distribution of data and on the function to estimate. When data are strategically produced, we could observe a degradation of these guarantees and, specifically, this degradation could vary locally if some data points are inherently more costly to produce or more sensitive than others. Such behavior is not observed in our model as linear regression is a parametric model where each data point contains information allowing us to estimate the global parameters of the problem while k-nearest neighbor is a non-parametric model where data points contain information only on the local behavior of the function to estimate.

A Supplementary Materials for Chapter 4

A.1 Proof of Lemma 4.1

Proof of Lemma 4.1. Let (α^*, β^*) be a BNE. From the definition of a BNE, we have $\alpha^* \in \arg\max_\alpha \sum_i p_i U_i^A(\alpha, \beta^*)$; that is, α^* is a best response of the attacker to β^* (for each attacker type). By observing that the average gain of the defender can be written as

$$\sum_i p_i U_i^D(\alpha, \beta) = -p_a \sum_i p_i U_i^A(\alpha, \beta) + f(\beta)$$

where $f(\beta) = (1 - p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi^\beta(v)$ does not depend on α, we deduce that $\alpha^* \in \arg\min_\alpha \sum_i p_i U_i^D(\alpha, \beta^*)$. Then from the definition of a BNE again , we conclude that $\beta^* \in \arg\max_\beta \min_\alpha \sum_i p_i U_i^D(\alpha, \beta)$. □

Remark A.1. *Note that by symmetry we also have* $\alpha^* \in \arg\max_\alpha \min_\beta - \sum_i p_i U_i^D(\alpha, \beta)$. *Hence, for any* α^* *such that* $\alpha^* \in \arg\max_\alpha \min_\beta - \sum_i p_i U_i^D(\alpha, \beta)$ *and* β^* *such that* $\beta^* \in \arg\max_\beta \min_\alpha \sum_i p_i U_i^D(\alpha, \beta)$, (α^*, β^*) *is a BNE.*

A.2 Proof of Proposition 4.1

Proof of Theorem 4.1. In this proof, to simplify the exposition, we assimilate β and π^β and write by abuse of notation $U_i^A(\alpha, \pi)$ and $U_i^D(\alpha, \pi)$ to denote the attacker and defender payoff (3.2) for any defender strategy β such that $\pi^\beta = \pi$.

Let $\boldsymbol{G}_{\max} \in \arg\max_{\boldsymbol{G}} U^D(\boldsymbol{G})$ and let $\pi_{\boldsymbol{G}_{\max}}(.)$ be the associated probability of detection function (4.2). We show that $\pi_{\boldsymbol{G}_{\max}}(.)$ is a min-max strategy in two steps.

Step 1: Let $\boldsymbol{G}$ be any arbitrary vector in $[\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$ and let $\pi_{\boldsymbol{G}}(.)$ be the associated probability of detection function (4.2). By definition of $\pi_{\boldsymbol{G}}$, we have

$$\max_v \{U_i^u(v) - \pi_{\boldsymbol{G}}(v) \cdot (U_i^u(v) + U_i^d(v))\} \leq G_i, \quad \forall i \in [\![1, m]\!]; \tag{A.1}$$

that is, every type of attacker i can have at most G_i payoff if the defender uses strategy $\pi_{\boldsymbol{G}}$. From the definition of the utility (3.2), this implies that

$$\min_{\alpha} \sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}}) \geq -p_a \sum_i p_i G_i + (1-p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi_{\boldsymbol{G}}(v).$$

Finally, noting that the right hand side of the above inequality is exactly $U^D(\boldsymbol{G})$ and applying it to $\boldsymbol{G} = \boldsymbol{G}_{\max}$, we obtain

$$\min_{\alpha} \sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}_{\max}}) \geq U^D(\boldsymbol{G}_{\max}). \tag{A.2}$$

Step 2: Conversely, let π be any arbitrary probability of detection function and define $\boldsymbol{G}^\pi$ as the vector with components

$$G_i^\pi = \max_v \{U_i^u(v) - \pi(v) \cdot (U_i^u(v) + U_i^d(v))\}, \quad (i \in [\![1, m]\!]). \tag{A.3}$$

Again, from the definition of the utility (3.2), we have

$$\min_{\alpha} \sum_i p_i U_i^D(\alpha, \pi) = -p_a \sum_i p_i G_i^\pi - (1-p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi(v);$$

that is that the minimum payoff of the defender is achieved when each attacker type maximizes its gain. Using (A.3), we have, for all type i and vector v, $\pi(v) \geq \frac{U_i^u(v) - G_i^\pi}{U_i^u(v) + U_i^d(v)}$, hence $\pi(v) \geq \max\left\{0, \max_i \left\{\frac{U_i^u(v) - G_i^\pi}{U_i^u(v) + U_i^d(v)}\right\}\right\} = \pi_{\boldsymbol{G}^\pi}(v)$ for all v. Plugging this inequality in the above equation gives

$$\min_{\alpha} \sum_i p_i U_i^D(\alpha, \pi) \leq -p_a \sum_i p_i G_i^\pi - (1-p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi_{\boldsymbol{G}^\pi}(v) = U^D(\boldsymbol{G}^\pi).$$

Since $U^D(\boldsymbol{G}^\pi) \leq U^D(\boldsymbol{G}_{\max})$ for all π by definition of $\boldsymbol{G}_{\max}$ as a maximum of function U^D, we finally get

$$\min_{\alpha} \sum_i p_i U_i^D(\alpha, \pi) \leq U^D(\boldsymbol{G}_{\max}). \tag{A.4}$$

To conclude, observe that combining (A.2) and (A.4) gives that, for all π, we have $\min_\alpha \sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}_{\max}}) \geq \min_\alpha \sum_i p_i U_i^D(\alpha, \pi)$; hence $\pi_{\boldsymbol{G}_{\max}}$ is a min-max strategy. □

Remark A.2. *From the proof above, we observe that* $\min_\alpha(\sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}_{\max}})) = U^D(\boldsymbol{G}_{\max})$*, which implies that, for all* i *and any maximizer* $\boldsymbol{G}_{\max}$ *of the function* $U^D(\cdot)$*, we have for all* i

$$\max_v\{U_i^u(v) - \pi_{\boldsymbol{G}_{\max}}(v)\cdot(U_i^u(v)+U_i^d(v))\} = (G_{\max})_i;$$

that is, when the defender uses strategy $\pi_{\boldsymbol{G}_{\max}}$ *the attacker gets a payoff of exactly* $(G_{\max})_i$ *for all type. It is important to note that this is not obvious and it is not the case for all vectors* $\boldsymbol{G}$ *in* $[\underline{G}_1, \overline{G}_1]\times\cdots\times[\underline{G}_m, \overline{G}_m]$*. In particular, the set* $\mathcal{S} = \{\boldsymbol{G} : \exists\pi, \forall i, G_i = \max_v\{U_i^u(v) - \pi(v)\cdot(U_i^u(v)+U_i^d(v))\}\}$ *of all* $\boldsymbol{G}$ *that are "best response for each type" to a strategy* π *is not equal to* $[\underline{G}_1, \overline{G}_1]\times\cdots\times[\underline{G}_m, \overline{G}_m]$ *and may not even be convex. For a* $\boldsymbol{G}$ *outside this set* $\mathcal{S}$*, the maximum payoff of the attacker against strategy* $\pi_{\boldsymbol{G}}$ *will not be* G_i *for all* i*, hence the interpretation of* U^D *as the minimum utility of the defender no longer holds outside* $\mathcal{S}$*. On the other hand, maximizing* U^D *on* $\mathcal{S}$ *directly is not possible as it may not be convex. Our proof bypasses this difficulty by using inequality* (A.1) *that is valid for all* $\boldsymbol{G}$ *in* $[\underline{G}_1, \overline{G}_1]\times\cdots\times[\underline{G}_m, \overline{G}_m]$.

A.3 Proof of Proposition 4.2

Proof of Proposition 4.2. Let OPT be the optimal objective value of the linear program. First note that, for any $\boldsymbol{G}$, the parameters $\pi_{\boldsymbol{G}}$ and $\boldsymbol{G}$ form a valid solution of the linear program by definition. Thus, $\max_{\boldsymbol{G}} U^D(\boldsymbol{G}) \leq OPT$. Conversely, for any optimal solution of the linear program π^*, $\boldsymbol{G}^*$, we have $\pi_v^* = \pi_{\boldsymbol{G}^*}(v)$ as it is the probability of detection achieving utility profile $\boldsymbol{G}$ while minimizing false alarms. We also trivially have $G_i^* = \max_v\{U_i^u(v) - pi_v(U_i^u(v)+U_i^d(v))\} \in [\underline{G}_i, \overline{G}_i]$ thanks to the first constraint of the linear program and the fact that we want to minimize the objective function. Thus, $OPT \leq \max U^D(\boldsymbol{G})$.

Combining the two, we obtain $OPT = \max_{\boldsymbol{G}} U^D(\boldsymbol{G})$ and $\boldsymbol{G}^* \in \arg\max U^D(\boldsymbol{G})$. □

A.4 Proof of Lemma 4.2

Proof. In this proof, to simplify the exposition we assimilate β and π^β and write by abuse of notation $U_i^D(\alpha, \pi^\beta)$ to denote the defender payoff (3.2). Let (α, β) be a strategy profile; then we have for all $v \in \mathcal{V}$:

$$\frac{\partial \sum_i p_i U_i^D\left(\alpha, \pi^\beta\right)}{\partial \pi(v)} = p_a \sum_i \alpha_v^i p_i (U_i^u(v) + U_i^d(v)) - (1 - p_a) C_{\text{fa}}(v) P_0(v). \quad \text{(A.5)}$$

As (α^*, β^*) is a BNE, by definition, $\beta^* \in \arg\max_\beta \sum_i p_i U_i^D(\alpha^*, \beta)$. This implies that: For all $v \in \mathcal{V}$ such that $\pi^{\beta^*}(v) = 0$, we have:

$$\frac{\partial \sum_i p_i U_i\left(\alpha^*, \pi^{\beta^*}\right)}{\partial \pi(v)} = p_a \sum_i {\alpha^*}_v^i p_i (U_i^u(v) + U_i^d(v)) - (1 - p_a) C_{\text{fa}}(v) P_0(v) \leq 0.$$

For all $v \in \mathcal{V}$ such that $0 < \pi^{\beta^*}(v) < 1$, we have:

$$\frac{\partial \sum_i p_i U_i\left(\alpha^*, \pi^{\beta^*}\right)}{\partial \pi(v)} = p_a \sum_i {\alpha^*}_v^i p_i (U_i^u(v) + U_i^d(v)) - (1 - p_a) C_{\text{fa}}(v) P_0(v) = 0.$$

For all $v \in \mathcal{V}$ such that $\pi^{\beta^*}(v) = 1$, we have:

$$\frac{\partial \sum_i p_i U_i\left(\alpha^*, \pi^{\beta^*}\right)}{\partial \pi(v)} = p_a \sum_i {\alpha^*}_v^i p_i (U_i^u(v) + U_i^d(v)) - (1 - p_a) C_{\text{fa}}(v) P_0(v) \geq 0.$$

which directly concludes the proof. □

B Supplementary Materials for Chapter 5

B.1 Proof of Theorem 5.1

Proof of Theorem 5.1. We start by recalling the setup, assumptions and the main theorem we use from [Sha03]. Consider a stochastic optimization problem of the form

$$\min_{x \in X} \{f(x)\} = E[F(x, \xi)],$$

where ξ is a random vector with support Ξ, with the following assumptions:

(C1) The set X is a convex closed polyhedron;

(C2) For every $\xi \in \Xi$ the function $F(\cdot, \xi)$ is proper convex and lower semi continuous and piecewise linear on its domain;

(C3) The support Ξ of ξ is finite.

Then the following theorem holds:

Theorem B.1 ([Sha03])**.** *Let $\hat{\mathcal{S}}$ be the set of optimal solutions of the sample average approximation problem (Algorithm 1, with N samples) and $\mathcal{S}$ the set of optimal solution from the true problem. Let $p_N = Pr[\hat{\mathcal{S}} \subseteq \arg\max U^D(\boldsymbol{G})]$. We have*

$$\limsup_{N \to \infty} \frac{1}{N} \log(1 - p_N) < 0.$$

The result of Theorem 5.1 then immediately follows by observing that Assumptions **(C1)**-**(C3)** trivially hold for the function $U^D(\boldsymbol{G})$ defined in Definition 4.2 and written as an expectation as explained above Algorithm 1 (with $X = [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$ and Ξ is the set of all possible vectors in $\mathcal{V}$ and all possible attacker types, hence $|\Xi| = |\mathcal{V}| + m$). □

We then justify the fact that bounds in expected value are also relevant for our setting by proving the following lemma:

Lemma B.1. *Let* $\boldsymbol{G} \in [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$. *Then,* $\min_\alpha \sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}}) \geq U^D(\boldsymbol{G})$.

Proof. By definition of $\pi_{\boldsymbol{G}}$ (see (4.2)), we have for all $i \in [\![1, m]\!]$ and $v \in \mathcal{V}$, $\frac{U_i^d(v) - G_i}{U_i^u(v) + U_i^d(v)} \leq \pi_{\boldsymbol{G}}(v)$. This directly implies $U_i^A(v, \pi_{\boldsymbol{G}}) = U_i^d(v) - (U_i^u(v) + U_i^d(v))\pi_{\boldsymbol{G}}(v) \leq G_i$. We thus have:

$$\begin{aligned}
\min_\alpha \sum_i p_i U_i^D(\alpha, \pi_{\boldsymbol{G}}) &= -p_a \max_\alpha \sum_i p_i U_i^A(\alpha, \pi_{\boldsymbol{G}}) - (1 - p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi_{\boldsymbol{G}}(v) \\
&\geq -p_a \sum_i p_i G_i - (1 - p_a) \sum_v C_{\text{fa}}(v) P_0(v) \pi_{\boldsymbol{G}}(v) \\
&\geq U^D(\boldsymbol{G}).
\end{aligned}$$

□

Lemma B.1 simply shows that any approximate maximum of $U^D(\boldsymbol{G})$ also gives an approximate min-max strategy. Thus, any stochastic optimization algorithm which yields bounds in expected value for the minimization of $U^D(\boldsymbol{G})$ also yields the same expected values guarantees about the minimum gain of the defender.

B.2 Classical online gradient descent algorithm and associated regret bound

In this section, we present the online gradient descent algorithm (termed "greedy projection" in Zinkevich [Zin03]) and the associated regret bound from Zinkevich [Zin03]. Let $(c_1, \ldots, c_T)$ be convex functions defined on a convex set $\mathcal{S}$. Let $\Pi_{\mathcal{S}}$ to be the Euclidean projection on $\mathcal{S}$, and let η_t, $t = 1, \cdots, T$, be a sequence of learning rates. Let $\partial c_t(x)$ denote the set of sub-gradients of c_t at point x. The online gradient descent algorithm is as follows.

Algorithm 3 Online gradient descent (OGD)

Initialize $x_1 \in \mathcal{S}$ arbitrarily
for $t = 1 \ldots T$ **do**
  Observe c_t
  Select $x_{t+1} = \Pi_{\mathcal{S}}(x_t - \eta_t \nabla c_t(x_t))$ for any $\nabla c_t(x_t) \in \partial c_t(x_t)$
end for

Then, Zinkevich [Zin03] shows that we have the following regret bound.

Theorem B.2 (Zinkevich, 2003)**.** *Assume that*

$$\max_{x_1,x_2\in\mathcal{S}} ||x_2 - x_1||_2 \leq D$$

and

$$||\nabla c_t(x)||_2 \leq L, \quad \forall t \in \{1,\dots,T\}, x \in \mathcal{S}, \forall \nabla c_t(x) \in \partial c_t(x).$$

Let $(x_1,\dots,x_t)$ *be vectors in* $\mathcal{S}$ *selected by Algorithm 3 with* $\eta_t = \frac{1}{\sqrt{t}}$*. Then, the regret accumulated at time* T*, defined as* $R(T) \equiv \sum_{t=1}^{T} c_t(x_t) - \min_x \sum_{t=1}^{T} c_t(x)$*, is bounded by:*

$$R(T) \leq \frac{D^2\sqrt{T}}{2} + \left(\sqrt{T} - \frac{1}{2}\right) L^2.$$

Importantly, note that although the bounds L and D appear in the regret bound of Theorem B.2, it is not necessary to know them to run the online gradient descent algorithm and they are not used in the algorithm. Note also that our functions c_t may not be differentiable. As noted in footnote 3 of Zinkevich [Zin03], the algorithm works also in that case, using sub-gradients as presented above.

B.3 Proof of regret bound for the naive online learning algorithm

Assume that the defender uses Algorithm 3 directly with $\mathcal{S}$ being the set of probability distributions on $\mathcal{V}$ and with functions c_t such that $c_t(\pi_t) = E_{\pi_t}[l(v_t)]$. If at time t the defender faced an attacker of type i, we have $c_t(\pi_t) = \max_v[(1 - \pi_t(v))U_i^u(v) - \pi_t(v)U_i^d(v)]$. If the defender faced a non-strategic attacker, we have $c_t(\pi_t) = C_{\text{fa}}(v_t)\pi_t(v_t)$. It is easy to verify that $c_1,\cdots,c_T$ satisfy the conditions of Theorem B.2 with $L = \max(\max_v\{C_{\text{fa}}(v)\}, \max_{v,i}\{|U_i^u(v) + U_i^d(v)|\})$ and $D^2 = |\mathcal{V}|$, hence leading to the regret bound of Theorem B.2 with those constants.

B.4 Proof of Theorem 5.2

Proof of Theorem 5.2. Algorithm 2 corresponds to online gradient descent from Algorithm 3 applied with $\mathcal{S} = [\underline{G}_1, \overline{G}_1] \times \cdots \times [\underline{G}_m, \overline{G}_m]$ and with functions c_t defined as follows: $c_t(\boldsymbol{G}) = G_i$ if the defender faces an attacker of type i at time t

and $c_t(\boldsymbol{G}) = \pi_{\boldsymbol{G}}(v_t)C_{\text{fa}}(v_t)$ if the defender faced a non-attacker at time t. We first show that

$$\sum_t E_{\pi_{\boldsymbol{G}_t}}[l(v_t)] - \min_\pi E_\pi[\sum_t l(v_t)] \leq \sum_t c_t(\boldsymbol{G}_t) - \min_{\boldsymbol{G}} \sum_t c_t(\boldsymbol{G}), \tag{B.1}$$

in two steps.

Step 1: Let $\pi^* \in \arg\min_\pi E_\pi[\sum_t l(v_t)]$ and define

$$G_i^{\pi^*} = \max_v \{U_i^u(v) - \pi^*(v) \cdot (U_i^u(v) + U_i^d(v))\}. \tag{B.2}$$

By definition of $G_i^{\pi^*}$, we have for all $i \in [\![1, m]\!]$ and $v \in \mathcal{V}$, $\pi^*(v) \geq \frac{U_i^d(v) - G_i^{\pi^*}}{U_i^u(v) + U_i^d(v)}$; thus, $\pi^*(v) \geq \pi_{\boldsymbol{G}^{\pi^*}}(v)$.

Note that we have $E_{\pi^*}[l(v_t)] \geq c_t(\boldsymbol{G}^{\pi^*})$. Indeed, if at time t a non-attacker was encountered with vector v_t, we have

$$E_{\pi^*}[l(v_t)] = \pi^*(v_t)C_{\text{fa}}(v_t) \geq \pi_{\boldsymbol{G}^{\pi^*}}(v_t)C_{\text{fa}}(v_t) \geq c_t(\boldsymbol{G}^{\pi^*});$$

and if an attacker of type i was encountered, we have

$$E_{\pi^*}[l(v_t)] = \max_v \{U_i^u(v) - \pi^*(v) \cdot (U_i^u(v) + U_i^d(v))\} = G_i^{\pi^*} = c_t(\boldsymbol{G}^{\pi^*}).$$

We thus have $\min_\pi E_\pi[\sum_t l(v_t)] = E_{\pi^*}[\sum_t l(v_t)] \geq \sum_t c_t(\boldsymbol{G}^{\pi^*}) \geq \min_{\boldsymbol{G}} \sum_t c_t(\boldsymbol{G})$. Additionally, it is trivial to verify that for all t and $\boldsymbol{G}$, we have $E_{\pi_{\boldsymbol{G}}}[l(v_t)] \leq c_t(\boldsymbol{G})$. Combined with the previous inequality, we get:

$$\min_\pi E_\pi[\sum_t l(v_t)] = \min_{\boldsymbol{G}} c_t(\boldsymbol{G}). \tag{B.3}$$

Step 2: As stated above, it is trivial to verify that for all t and $\boldsymbol{G}$, we have $E_{\pi_{\boldsymbol{G}}}[l(v_t)] \leq c_t(\boldsymbol{G})$. In particular, this holds for $\boldsymbol{G} = \boldsymbol{G}_t$, which directly implies:

$$\sum_t E_{\pi_{\boldsymbol{G}_t}}[l(v_t)] \leq \sum_t c_t(\boldsymbol{G}). \tag{B.4}$$

Combining (B.3) and (B.4) immediately gives (B.1). To conclude the proof, we apply the regret bound of Theorem B.2 to the right-hand side of (B.1), noting that each loss function is convex and that it can be easily verified from the definitions of D and L that the conditions of Theorem B.2 are satisfied. □

C Supplementary Materials for Chapter 7

C.1 Proof of Proposition 7.1

Step 1: if there does not exist i such that $\mathbb{E}_{\mu^i_{\text{marg}}}\left[x_i x_i^\top\right] \succ 0$, there exists a trivial Nash equilibrium. Let $\boldsymbol{\lambda}^* = 0$. Then, for any agent i and any $\boldsymbol{\lambda} = (\lambda_i, \lambda^*_{-i})$, we have that $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{j=1}^n (\boldsymbol{\lambda})_j(x) x_j x_j^\top\right]$ is non-invertible. We thus have $C_{\text{estim}}(\boldsymbol{\lambda}) = \infty$ and $\phi(\boldsymbol{\lambda}) = \phi(\boldsymbol{\lambda}^*)$. Thus, $\boldsymbol{\lambda}^*$ is an equilibrium.

Step 2: if there exists i such that $\mathbb{E}_{\mu^i_{\text{marg}}}\left[x_i x_i^\top\right] \succ 0$, there exists no trivial Nash equilibrium. By contradiction, assume that there exists a trivial Nash equilibrium $\boldsymbol{\lambda}^*$ of infinite estimation cost. We will show that agent i can achieve a finite payoff. Let $\lambda_i(x) = 1$ for all $x \in \mathcal{X}$ and $\boldsymbol{\lambda} = (\lambda_i, \lambda^*_{-i})$. Then, we have $J_i(\lambda_i, \lambda^*_{-i}) \leq \mathbb{E}\left[c_i(\lambda_i(x))\right] + C_{\text{estim}}(\boldsymbol{\lambda})$. As by assumption we have $\mathbb{E}_{\mu^i_{\text{marg}}}\left[x_i x_i^\top\right] \succ 0$ the estimation cost $C_{\text{estim}}(\boldsymbol{\lambda})$ is finite and thus the individual cost of agent i, $J_i(\lambda_i, \lambda^*_{-i})$ is finite and strictly lower than infinity. Thus, $\boldsymbol{\lambda}^*$ is not an equilibrium.

Note that we work by contradiction for step 2 as this proof is completely separated from the proofs of existence of equilibrium. We thus do not assume that there exists any equilibrium in this proof.

C.2 Proof of Proposition 7.2

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

We denote $V(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{i\in N}\lambda_i(x_i)x_ix_i^\top\right]^{-1}$ and $M(\boldsymbol{\lambda}) = \mathbb{E}\left[\sum_{i\in N}\lambda_i(x_i)x_ix_i^\top\right]$. 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 $\phi(\mathbf{1}) = \mathbb{E}\left[\sum_i c_i(1)\right] + F((\mathbb{E}\left[\sum_i x_ix_i^T\right])^{-1})$. By Assumption 6.2, $\lim_{\ell\to+\infty} c_i(\ell) = +\infty$. 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. As shown before, a non-trivial equilibrium is a minimum of the potential function ϕ defined for all precision profiles $\boldsymbol{\lambda}$ as

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

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

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

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

Yet, we show below that $C_{\text{estim}}(\cdot)$ is strictly convex when viewed as a function of $M(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_i \lambda_i(x_i) x_i x_i^T\right]$. Indeed F is an increasing convex function (by Assumption 6.3) 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 non-trivial equilibria $\boldsymbol{\lambda}^*$ and $\tilde{\boldsymbol{\lambda}}^*$ and assume by contradiction that $\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_i \lambda_i^*(x_i) x_i x_i^T\right] \neq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_i \tilde{\lambda}_i^*(x_i) x_i x_i^T\right]$. Let $\boldsymbol{\lambda}' = (\boldsymbol{\lambda}^* + \tilde{\boldsymbol{\lambda}}^*)/2$. The strict convexity of $M \mapsto F(M^{-1})$ implies that $C_{\text{estim}}(\boldsymbol{\lambda}') < (C_{\text{estim}}(\boldsymbol{\lambda}^*) + C_{\text{estim}}(\tilde{\boldsymbol{\lambda}}^*))/2$. 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.3 Proof of Theorem 7.1

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

$$\forall i, i', \forall x \in \mathcal{X} \text{ and } c_i = c_{i'} \Rightarrow \boldsymbol{\lambda}_i^{\text{ci}*}(x) = \boldsymbol{\lambda}_{i'}^{\text{ci}*}(x) \tag{C.1}$$

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

$$\forall i, i', c_i = c_{i'} \Rightarrow \forall x \in \mathcal{X}, \lambda_i^*(x) = \lambda_{i'}^*(x) \tag{C.2}$$

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

$$\forall i \in N, \forall x \in \mathcal{X}, \lambda_i(x) = \sum_{i'=1}^{n} \mathbb{1}_{c_i = c_t} \frac{\lambda_{i'}^{\text{ci}*}(x)}{n_t^{x_i}},$$

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 equilibrium, 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 independently and identically distributed 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 $\lambda_i(x)$ is replaced by $\lambda_t(x)$ 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:

$$\phi^{\text{ci}}_{\boldsymbol{x}}(\boldsymbol{\lambda}) = \sum_{x\in\mathcal{X}}\sum_{t=1}^{T} c_t(\lambda_t(x))n_t^x + F((\sum_{x\in\mathcal{X}} xx^T \sum_{t=1}^{T}\lambda_t(x)n_t^x)^{-1}) \tag{C.3}$$

$$\phi(\boldsymbol{\lambda}) = \sum_{x\in\mathcal{X}}\sum_{t=1}^{T} c_t(\lambda_t(x))n_t\mu(x) + F((\sum_{x\in\mathcal{X}} xx^T \sum_{t=1}^{T}\lambda_t(x)n_t\mu(x))^{-1}), \tag{C.4}$$

where as before, n_t is the number of players having cost function c_t and n_t^x is the number of players 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 (C.3) and (C.4).

The equilibrium are defined as $\boldsymbol{\lambda}^{\text{ci}*} \in \arg\min(\phi^{\text{ci}}_{\boldsymbol{x}}(\boldsymbol{\lambda}))$ and $\boldsymbol{\lambda}^* \in \arg\min(\phi(\boldsymbol{\lambda}))$, where the potential functions are defined in Equations (C.3) and (C.4).

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:

$$\begin{aligned}
\phi^{\text{ci}}_{\boldsymbol{x}}(\boldsymbol{\lambda}^{\text{ci}*}) &\leq \phi^{\text{ci}}_{\boldsymbol{x}}(\tilde{\boldsymbol{\lambda}}^*) \\
&= \sum_x\sum_t c_t(\lambda_t^*(x)\frac{\mu(x)n_t}{n_t^x})n_t^x + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1}) \\
&\leq \sum_x\sum_t (\frac{\mu(x)n_t}{n_t^x})^{p_{\max}} c_t(\lambda_t^*(x))n_t^x + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1}) \qquad \text{(C.5)} \\
&= \sum_x\sum_t (\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1} c_t(\lambda_t^*(x))n_t^x\mu(x) + F((\sum_x xx^T\sum_t \lambda_t^*(x)\mu(x))^{-1}) \\
&\leq \max_{x,t}(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*), \qquad \text{(C.6)}
\end{aligned}$$

where the inequality (C.5) comes from the assumption on the costs and the inequality (C.6) 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:

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

We thus obtain that:

$$\frac{1}{\max_{x,t}(\frac{n_t^x}{\mu(x)n_t})^{p_{\max}-1}}\phi(\boldsymbol{\lambda}^*) \leq \phi_{\boldsymbol{x}}^{\text{ci}}(\boldsymbol{\lambda}^{\text{ci}*}) \leq \max_{x,t}(\frac{\mu(x)n_t}{n_t^x})^{p_{\max}-1}\phi(\boldsymbol{\lambda}^*) \tag{C.7}$$

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:

$$P(|n_t^x - n_t\mu(x)| \geq n_t^{1/2+\epsilon}) \leq 2\exp(-2n_t^{2\epsilon}) \tag{C.8}$$

We thus have $P\left(\cup_{t,x}\left(|n_t^x - n_t\mu(x)| \geq n_t^{1/2+\epsilon}\right)\right) \leq |X|\sum_t 2\exp(-2n_t^{2\epsilon})$. 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:

$$\frac{\mu(x)}{\mu(x) + n_t^{\epsilon-1/2}} \leq \frac{\mu(x)n_t}{n_t^x} \leq \frac{\mu(x)}{\mu(x) - n_t^{\epsilon-1/2}}. \tag{C.9}$$

Combined with (C.7), this shows that with probability at least $|\mathcal{X}|\sum_t 2\exp(-2n_t^{2\epsilon})$, we have:

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

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

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

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:

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

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

C.4 Proof of Theorem 7.3

Proof. To simplify the notation, in this proof, we write p instead of $p_{\min}$; hence we show that $\texttt{PoS} \leq n^{\frac{q}{p+q}}$.

Upper Bound. Recall that Assumption 6.4 implies that $\forall \lambda \in \mathbb{R}_+, \forall a \geq 1, a^p c_i(\lambda) \leq c_i(a\lambda)$. This implies, by rewriting the assumption with $\lambda' = a\lambda$, that $c_i(\frac{\lambda'}{a}) \leq a^{-p} c_i(\lambda')$ for all $a \geq 1$ and for all λ'.

Recall that we denote by $\boldsymbol{\lambda}^*$ the unique non-trivial Nash equilibrium. Suppose that $\texttt{PoS} > n^{\frac{q}{p+q}}$, that is

$$\sum_{i\in N} c_i(\lambda_i^*) + nC_{\text{estim}}(\boldsymbol{\lambda}^*) > n^{\frac{q}{q+p}}\left(\sum_{i\in N} c_i(\lambda_i^{\texttt{opt}}) + nC_{\text{estim}}(\boldsymbol{\lambda}^{\texttt{opt}})\right).$$

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:

$$\begin{aligned}
\sum_{i\in N} c_i(\lambda_i^*) + C_{\text{estim}}(\boldsymbol{\lambda}^*) &\geq \frac{1}{n}\left(\sum_{i\in N} c_i(\lambda_i^*) + nC_{\text{estim}}(\boldsymbol{\lambda}^*)\right) \\
&> n^{-\frac{p}{q+p}} \sum_{i\in N} c_i(\lambda_i^{\texttt{opt}}) + n^{\frac{q}{p+q}} C_{\text{estim}}(\boldsymbol{\lambda}^{\texttt{opt}}) \\
&\geq \sum_{i\in N} c_i\left(\frac{\lambda_i^{\texttt{opt}}}{n^{\frac{1}{p+q}}}\right) + C_{\text{estim}}\left(\frac{\boldsymbol{\lambda}^{\texttt{opt}}}{n^{\frac{1}{p+q}}}\right),
\end{aligned}$$

where for the last inequality, we used Assumption 6.4 and Assumption 6.3 with $a = n^{1/(p+q)}$.

This would imply that $\boldsymbol{\lambda}^*$ is not the minimum of the potential function which is a contradiction. Thus, we have $\texttt{PoS} \leq n^{\frac{q}{p+q}}$. □

Lower Bound. Fix $p \geq 1$ and $q \geq 1$. We consider a 1-dimensional model ($d = 1$) with $x_1 = 1$ and $\sigma^2 = (q/p)^{1/(p+q)}$. Let $c_i(\lambda_i) = \lambda_i^p$ for all i and $F(V) = \text{trace}(V)^q = V^q$ (the last equality holds because when $d = 1$, the covariance matrix is a scalar). Hence, the covariance matrix is $V(\boldsymbol{\lambda}) = (\sum_{i\in N} \lambda_i)^{-1}$.

As all agents are identical, and by uniqueness of the Nash equilibrium, the Nash equilibrium is a symmetric Nash equilibrium where all agents will give the same value λ^* where λ^* is the unique minimizer of the potential function:

$$n\lambda^p + (n\lambda)^{-q}.$$

The minimum of this function is attained when its derivative is equal to 0. This implies that $np\lambda^{p-1} = nq(n\lambda)^{-q-1}$ which implies that $\lambda^{p+q} = (q/p)n^{-1-q}$. This shows that $\lambda^* = ((q/p)n^{-1-q})^{1/(p+q)}$.

Similarly, the socially optimal $\boldsymbol{\lambda}^{\texttt{opt}}$ is also symmetric and is attained when all agents give $\lambda^{\texttt{opt}}$ the unique minimizer of the social cost:

$$n\lambda^p + n(n\lambda)^{-q}.$$

This implies that

$$\lambda^{\texttt{opt}} = (n(q/p)n^{-1-q})^{1/(p+q)} = n^{1/(p+q)}\lambda^*. \tag{C.10}$$

Hence, we get:

$$\begin{aligned}
\texttt{PoS} &= \frac{C(\boldsymbol{\lambda}^*)}{C(\boldsymbol{\lambda}^{\texttt{opt}})} = \frac{n(\lambda^*)^p + n(n\lambda^*)^{-q}}{n(\lambda^{\texttt{opt}})^p + n(n\lambda^{\texttt{opt}})^{-q}} \\
&= \frac{(\lambda^*)^p + (n\lambda^*)^{-q}}{(\lambda^{\texttt{opt}})^p + (n\lambda^{\texttt{opt}})^{-q}} \\
&= \frac{(n\lambda^*)-q}{(n\lambda^{\texttt{opt}})^{-q}} \frac{(\lambda^*)^{p+q} + 1}{(\lambda^{\texttt{opt}})^{p+q} + 1} \\
&= \left(\frac{\lambda^{\texttt{opt}}}{\lambda^*}\right)^q \frac{1 + (\lambda^*)^{p+q}}{1 + (\lambda^{\texttt{opt}})^{p+q}} \\
&= n^{q/(p+q)} \frac{1 + (q/p)n^{-1-q}}{1 + (q/p)n^{-q}},
\end{aligned}$$

where we use the expression in (C.10) for λ^* and $\lambda^{\texttt{opt}}$ in the last line. This shows that, for any ϵ, for large enough n, the price of stability is at least $n^{q/(p+q)}(1-\varepsilon)$. □

Supplementary Materials for Chapter 8

D

D.1 Proof of Theorem 8.2

Proof. **Proof of (i)**

We consider the same setting as Example 1, i.e., a 1-dimensional model ($d = 1$) with two agents in which the public data of each agent is $x_i = 1$. For such a game, the GLS estimator is $(X^T \Lambda X)^{-1} X^T \Lambda \tilde{\boldsymbol{y}} = (\lambda_1 + \lambda_2)^{-1} \boldsymbol{\lambda}^T \tilde{\boldsymbol{y}}$ and its covariance is $1/(\lambda_1 + \lambda_2)$. We consider a linear estimator $\hat{\boldsymbol{\beta}}(\delta)$ with $\delta \geq 0$ of the form $\hat{\boldsymbol{\beta}}_{\texttt{GLS}} + \boldsymbol{\delta}^T \tilde{\boldsymbol{y}}$ where $\boldsymbol{\delta} \in \mathbb{R}^2$ is a vector with coefficients $\delta_1 = -\delta_2 = \sqrt{\delta}$. Note that $\delta_1 = -\delta_2$ guarantees that this linear estimator is unbiased. We assume that the disclosure cost of Agent 1 is $c_1(\lambda) = \lambda^{p_1}$ while the disclosure cost of Agent 2 is $c_2(\lambda) = \lambda^{p_2}$. For a given δ, we denote the equilibrium of the game by $\boldsymbol{\lambda}^*(\delta)$.

Overall, this proof is decomposed in two steps:

> Step 1: We compute the derivative of the estimation cost at $\delta = 0$ to show that it is negative if and only if $\lambda_1^*(0)(2p_1 - p_2 - p_1 p_2) + \lambda_2^*(0)(2p_2 - p_1 - p_1 p_2) > 0$.
>
> Step 2: We show that there exists $x > 0$ such that the above inequality is satisfied for $p_1 = 1/x$ and $p_2 = 1 + x$.

We describe both steps in detail below.

Step 1. According to (8.3), the covariance of the estimator is $1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$, where $\delta/\lambda_1 + \delta/\lambda_2$ is the loss of precision due to using a linear estimator that is less precise than GLS. We assume that the scalarization function is the identity, which means that the estimation cost is

$$C^{\delta}_{\text{estim}}(\boldsymbol{\lambda}) = \frac{1}{\lambda_1 + \lambda_2} + \frac{\delta}{\lambda_1} + \frac{\delta}{\lambda_2}. \tag{D.1}$$

The equilibrium $\boldsymbol{\lambda}^*(\delta)$ is the minimum of the potential function $\Phi_\delta(\boldsymbol{\lambda}) = C^\delta_{\text{estim}}(\boldsymbol{\lambda}) + \lambda_1^{p_1} + \lambda_2^{p_2}$. The estimation cost at equilibrium is $C^\delta_{\text{estim}}(\boldsymbol{\lambda}^*(\delta))$. Our goal in this step is to compute the derivative of $C^\delta_{\text{estim}}(\boldsymbol{\lambda}^*(\delta))$ with respect to δ and to obtain a condition that ensures that it is negative at $\delta = 0$. Let use denote by $(\lambda^*)_i'(\delta) = d\lambda_i^*(\delta)/(d\delta)$ the derivative of $\lambda_i^*(\delta)$ with respect to δ. To simplify notation, we will omit the dependence on δ and simply denote $\lambda_i^* = \lambda_i^*(0)$ and $\lambda_i' = (\lambda^*)_i'(0)$ when it is not confusing. The derivative of the estimation cost evaluated at $\delta = 0$ is equal to

$$\frac{d}{d\delta}\left(C^\delta_{\text{estim}}(\boldsymbol{\lambda}^*(\delta))\right)\Big|_{\delta=0} = -\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^2} + \frac{1}{\lambda_1^*} + \frac{1}{\lambda_2^*} = -\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^2} + \frac{\lambda_1^* + \lambda_2^*}{\lambda_1^*\lambda_2^*}. \tag{D.2}$$

In particular, the above derivative is negative if an only if $\frac{\lambda_1'+\lambda_2'}{(\lambda_1^*+\lambda_2^*)^3}\lambda_1^*\lambda_2^* > 1$. In what follows, we compute the derivatives λ_i' as a function of the values of λ_i^* and p_i.

The equilibrium $\boldsymbol{\lambda}^*(\delta)$ is the minimum of the potential function $\Phi_\delta(\boldsymbol{\lambda}) = C^\delta_{\text{estim}}(\boldsymbol{\lambda}) + \lambda_1^{p_1} + \lambda_2^{p_2}$. By using the first order condition $\partial\Phi_\delta/\partial\lambda_i = 0$, this implies that for all $\delta \geq 0$:

$$-\frac{1}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^2} - \frac{\delta}{(\lambda_i^*(\delta))^2} + p_i(\lambda_i^*(\delta))^{p_i-1} = 0, \quad \text{for } i \in \{1, 2\}. \tag{D.3}$$

The derivative of $\lambda_i^*(\delta)$ with respect to δ exists thanks to the implicit function theorem. By differentiating (D.3) with respect to δ, we obtain

$$\begin{aligned} 0 &= \frac{d}{d\delta}\left(-\frac{1}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^2} - \frac{\delta}{(\lambda_i^*(\delta))^2} + p_i(\lambda_i^*(\delta))^{p_i-1}\right) \\ &= 2\frac{(\lambda^*)_1'(\delta) + (\lambda^*)_2'(\delta)}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^3} - \frac{1}{(\lambda_i^*(\delta))^2} + 2\delta\frac{(\lambda^*)_i'(\delta)}{(\lambda_i^*(\delta))^3} + p_i(p_i-1)(\lambda_i^*(\delta))^{p_i-2}(\lambda^*)_i'(\delta). \end{aligned} \tag{D.4}$$

Equation (D.3), evaluated at $\delta = 0$, shows that $p_i(\lambda_i^*)^{p_i-1} = \frac{1}{(\lambda_1^*+\lambda_2^*)^2}$. Evaluating Equation (D.4) at $\delta = 0$ and plugging this equality gives

$$\begin{aligned} 0 &= 2\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} - \frac{1}{(\lambda_i^*)^2} + p_i(p_i-1)(\lambda_i^*)^{p_i-2}\lambda_i' \\ &= 2\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} - \frac{1}{(\lambda_i^*)^2} + \frac{1}{(\lambda_1^* + \lambda_2^*)^2}\frac{p_i - 1}{\lambda_i^*}\lambda_i'. \end{aligned} \tag{D.5}$$

In order to isolate the term $\lambda_1' + \lambda_2'$, we multiply the above equation by $\lambda_i^*/(p_i - 1)$ and we sum over $i \in \{1, 2\}$. This gives:

$$\begin{aligned} 0 &= 2\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3}\left(\frac{\lambda_1^*}{p_1 - 1} + \frac{\lambda_2^*}{p_2 - 1}\right) - \frac{1}{\lambda_1^*(p_1 - 1)} - \frac{1}{\lambda_2^*(p_2 - 1)} + \frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^2} \\ &= \frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3}\left(\frac{2\lambda_1^*}{p_1 - 1} + \frac{2\lambda_2^*}{p_2 - 1} + \lambda_1^* + \lambda_2^*\right) - \frac{1}{\lambda_1^*(p_1 - 1)} - \frac{1}{\lambda_2^*(p_2 - 1)}. \end{aligned}$$

This shows that

$$\begin{aligned} \frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} &= \frac{\frac{1}{\lambda_1^*(p_1 - 1)} + \frac{1}{\lambda_2^*(p_2 - 1)}}{\frac{2\lambda_1^*}{p_1 - 1} + \frac{2\lambda_2^*}{p_2 - 1} + \lambda_1^* + \lambda_2^*} \\ &= \frac{1}{\lambda_1^*\lambda_2^*}\frac{\lambda_2^*(p_2 - 1) + \lambda_1^*(p_1 - 1)}{2\lambda_1^*(p_1 - 1) + 2\lambda_2^*(p_2 - 1) + (\lambda_1^* + \lambda_2^*)(p_1 - 1)(p_2 - 1)}. \end{aligned}$$

In particular, this implies that the derivative (D.2) is negative if and only if

$$\frac{\lambda_2^*(p_2 - 1) + \lambda_1^*(p_1 - 1)}{2\lambda_1^*(p_1 - 1) + 2\lambda_2^*(p_2 - 1) + (\lambda_1^* + \lambda_2^*)(p_1 - 1)(p_2 - 1)} > 1.$$

After some algebra, this gives

$$\lambda_1^*(2p_1 - p_2 - p_1p_2) + \lambda_2^*(2p_2 - p_1 - p_1p_2) > 0, \tag{D.6}$$

where, again, by abuse of notation we denote $\lambda_1^* = \lambda_1^*(0)$ and $\lambda_2^* = \lambda_2^*(0)$.

Step 2. We now consider $p_1 = 1+1/x$ and $p_2 = 1+x$ and $x \to \infty$. To emphasize the dependence in x, let us denote by $\boldsymbol{\lambda}^*(x) = (\lambda_1^*(x), \lambda_2^*(x))$ the value of the precision at equilibrium (for `GLS`) and $\Phi_x(\cdot)$ the potential of the game. By definition, $\boldsymbol{\lambda}^*(x)$ minimizes $\Phi_x(\boldsymbol{\lambda}) = 1/(\lambda_1 + \lambda_2) + \lambda_1^{1+1/x} + \lambda_2^{1+x}$. This implies that for all $\epsilon > 0$, $\Phi_x(\boldsymbol{\lambda}^*(x)) \leq \Phi_x(0, 1 - \varepsilon)$. As $\lim_{x\to\infty} \Phi_x(0, 1 - \varepsilon) = 1/(1 - \varepsilon)$ and because this is true for all ε, this implies that

$$\lim_{x\to\infty} \Phi_x(\boldsymbol{\lambda}^*(x)) = \lim_{x\to\infty}\left(\frac{1}{\lambda_1^*(x) + \lambda_2^*(x)} + (\lambda_1^*(x))^{1+1/x} + (\lambda_2^*(x))^{1+x}\right) \leq 1.$$

This implies that $\lim_{x\to\infty} \lambda_1^*(x) = 0$ and $\lim_{x\to\infty} \lambda_2^*(x) = 1$.

For our values of $p_1 = 1 + 1/x$ and $p_2 = 1 + x$, the left-hand side of (D.6) equals $\lambda_1(x)(1/x-2x-1)+\lambda_2(x)(x-2/x-1)$. As $\lim_{x\to\infty} \lambda_2(x) = 1$ and $\lim_{x\to\infty} \lambda_1(x) = 0$, this term is asymptotically equivalent to x and is therefore positive for x large enough. This implies that there exists a value x such that $d/(d\delta)C_{\text{estim}}^{\delta}(\boldsymbol{\lambda}^*(\delta)) < 0$. Hence,

for this x, there exists a perturbation value $\delta > 0$ such that $\hat{\beta}(\delta)$ is an estimator that is more efficient than GLS. □

Proof of (ii)

We will start by proving Lemma D.1. This lemma can easily be explained if we recall Assumption 6.4 and Assumption 6.3. Indeed, they dictate how the different components of the potential function behave when all agents multiply or divide the amount of information they give. If the sum of individual costs is too great compared to the common cost then dividing the amount that all agents give greatly reduces the individual costs while slightly augmenting the common cost, which is beneficial. The same goes the other way around where agents multiply the amount they give. This formalizes an intuition that one can have about this model: there is a balance between the individual costs paid to achieve the objective of reducing the common cost and the objective itself.

Lemma D.1. *Under Assumptions 6.2, 6.3, and 6.4, for any estimator L satisfying Assumption 8.2 the ratio between the sum of individual costs and the common cost is bounded. Formally, the equilibrium $\boldsymbol{\lambda}^*$ satisfies:*

$$\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] \leq \frac{q}{p_{\min}} C^L_{\text{estim}}(\boldsymbol{\lambda}^*) \qquad \textit{and} \qquad C^L_{\text{estim}}(\boldsymbol{\lambda}^*) \leq \frac{p_{\max}}{q}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right].$$

Proof. This proof mainly relies on the fact that $\boldsymbol{\lambda}^*$ is the minimum of the potential function. Let $\boldsymbol{\lambda}^*$ be the unique non-trivial equilibrium. Let $\kappa \in (0,1)$ be a multiplicative factor applied to the equilibrium profile. As $\boldsymbol{\lambda}^*$ is the minimum of the potential function, we have $\phi_L(\boldsymbol{\lambda}^*) \leq \phi_L(\kappa\boldsymbol{\lambda}^*)$ and $\phi_L(\boldsymbol{\lambda}^*) \leq \phi_L(\boldsymbol{\lambda}^*/\kappa)$. This implies that:

$$\begin{aligned}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + C^L_{\text{estim}}\boldsymbol{\lambda}^* &\leq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\kappa\lambda_i(x_i))\right] + C^L_{\text{estim}}(\kappa\boldsymbol{\lambda}^*) \\ &\leq \kappa^{p_{\min}}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + \kappa^{-q}C^L_{\text{estim}}(\boldsymbol{\lambda}^*),\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + C^L_{\text{estim}}(\boldsymbol{\lambda}^*) &\leq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i)/\kappa)\right] + C^L_{\text{estim}}(\boldsymbol{\lambda}^*/\kappa) \\ &\leq \kappa^{-p_{\max}}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + \kappa^{q}C^L_{\text{estim}}(\boldsymbol{\lambda}^*).\end{aligned}$$

Where we used Assumptions 6.3, 6.4, and 8.2 to obtain the last term of each inequality. The above equations imply that:

$$(1-\kappa^{p_{\min}})\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] \leq (\kappa^{-q}-1)C^L_{\text{estim}}(\boldsymbol{\lambda}^*), \qquad \text{and}$$
$$(1-\kappa^{q})C^L_{\text{estim}}(\boldsymbol{\lambda}^*) \leq (\kappa^{-p_{\max}}-1)\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right].$$

As $\kappa \in (0,1)$, we have $1-\kappa^{p_{\min}} > 0$ and $\kappa^{-p_{\max}} - 1 > 0$. Hence, the above equations imply that

$$\frac{1-\kappa^q}{\kappa^{-p_{\max}}-1} C^L_{\text{estim}}(\boldsymbol{\lambda}^*) \leq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] \leq \frac{\kappa^{-q}-1}{1-\kappa^{p_{\min}}} C^L_{\text{estim}}(\boldsymbol{\lambda}^*).$$

This inequality is valid for every $\kappa \in (0,1)$. As $\lim_{\kappa\to 1} \frac{1-\kappa^q}{\kappa^{-p_{\max}}-1} = q/p_{\max}$ and $\lim_{\kappa\to 1} \frac{1-\kappa^{-q}}{1-\kappa^{-p_{\min}}} = q/p_{\min}$, this gives

$$\frac{q}{p_{\max}} C^L_{\text{estim}}(\boldsymbol{\lambda}^*) \leq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] \leq \frac{q}{p_{\min}} C^L_{\text{estim}}(\boldsymbol{\lambda}^*).$$

□

We are now ready to prove Theorem 8.2(ii). Let $\Phi_L(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + C^L_{\text{estim}}(\boldsymbol{\lambda})$ be the potential function for any linear unbiased estimator and $\Phi_{\texttt{GLS}}(\boldsymbol{\lambda}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i(\lambda_i(x_i))\right] + C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda})$ be the potential function for GLS. Recall that $\boldsymbol{\lambda}^*_L$ and $\boldsymbol{\lambda}^*_{\texttt{GLS}}$ denote the non-trivial equilibria for the linear unbiased estimator and for GLS respectively. By optimality of GLS, for all $\boldsymbol{\lambda}$ we have $C^L_{\text{estim}}(\boldsymbol{\lambda}) \geq C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda})$. This implies that for all $\boldsymbol{\lambda}$, we have $\Phi_L(\boldsymbol{\lambda}) \geq \Phi_{\texttt{GLS}}(\boldsymbol{\lambda})$. Therefore

$$\Phi_L(\boldsymbol{\lambda}^*_L) = \min_{\boldsymbol{\lambda}} \Phi_L(\boldsymbol{\lambda}) \geq \Phi_{\texttt{GLS}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) = \min_{\boldsymbol{\lambda}} \Phi_{\texttt{GLS}}(\boldsymbol{\lambda}). \tag{D.7}$$

By applying the inequalities of Lemma D.1, we obtain:

$$\Phi_{\texttt{GLS}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i((\boldsymbol{\lambda}^*_{\texttt{GLS}})_i(x_i))\right] + C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) \geq \frac{q}{p_{\max}} C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) + C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}),$$

and

$$\Phi_L(\boldsymbol{\lambda}^*_L) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i\in N} c_i((\boldsymbol{\lambda}^*_L)_i(x_i))\right] + C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L) \leq \frac{q}{p_{\min}} C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L) + C^L_{\text{estim}}(\boldsymbol{\lambda}^*_L).$$

Combining the above two inequalities with (D.7), we conclude that

$$C^{\texttt{GLS}}_{\text{estim}}(\boldsymbol{\lambda}^*_{\texttt{GLS}}) \leq \frac{\frac{q}{p_{\min}}+1}{\frac{q}{p_{\max}}+1} C^{L}_{\text{estim}}(\boldsymbol{\lambda}^*_{L}) = \frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})} C^{L}_{\text{estim}}(\boldsymbol{\lambda}^*_{L}).$$

□

D.2 Proof of Theorem 8.3

Recall that the provision cost of a player i is $c_i(\ell) = a\ell$.

Let $\boldsymbol{\lambda}^*$ be an equilibrium of the game and let ν^* be an optimal design. Recall that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_{i\in N} \lambda_i^*(x)\mu^i_{\text{marg}}(x)$ 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 $\lambda_{\nu^*}(x) = b\nu^*(x)/\sum_{i\in N}\mu^i_{\text{marg}}(x)$ for all x (λ_{ν^*} is well defined as Assumption 6.1 ensures that $\sum_{i\in N}\mu^i_{\text{marg}}(x) > 0$) and consider the precision profile $\boldsymbol{\lambda}_{\nu^*} = (\lambda_{\nu^*}, 0, \cdots, 0)$. We have:

$$\begin{aligned}
\phi(\boldsymbol{\lambda}^*) &= F((\sum_x xx^T \nu_{\boldsymbol{\lambda}^*}(x))^{-1}) + \sum_i a \sum_x \lambda_i^*(x)\mu^i_{\text{marg}}(x) \\
&\geq F((\sum_x xx^T \nu_{\boldsymbol{\lambda}^*}(x))^{-1}) + ab && \text{(D.8)} \\
&= b^{-q} F((\sum_x xx^T \nu_{\boldsymbol{\lambda}^*}(x)/b)^{-1}) + ab && \text{(D.9)} \\
&\geq b^{-q} F((\sum_x xx^T \nu^*(x))^{-1}) + ab && \text{(D.10)} \\
&= F((\sum_x xx^T \lambda_{\nu^*}(x)\mu(x))^{-1}) + a_1 \sum_{i\in N}\sum_x (\lambda_{\nu^*}(x))_i \mu^i_{\text{marg}}(x) && \text{(D.11)} \\
&= \phi(\boldsymbol{\lambda}_{\nu^*}),
\end{aligned}$$

where the second inequality (D.10) is because ν^* is an optimal design. The equalities (D.9) and (D.11) are due to the homogeneity of F (Assumption 6.3 implies that $F((bM)^{-1}) = b^{-q}F(M^{-1})$), and in (D.11) 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 (D.10) 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 (D.10) is an equality which means that $\nu_{\boldsymbol{\lambda}^*}(x)/b$ is an optimal design.

D.3 Proof of Proposition 8.1

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 $\boldsymbol{\lambda} = (\lambda, \dots \lambda)$ is a Nash equilibrium if and only if it minimizes the potential ϕ. By symmetry, this potential can be rewritten as:

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

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

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

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 7.2.

D.4 Proof of Theorem 8.4

Proof. **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

$$\begin{aligned}
\phi(n^{-\frac{q+1}{p_{\min}+q}}, \dots, n^{-\frac{q+1}{p_{\min}+q}}) &= \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n} c_i(n^{-\frac{q+1}{p_{\min}+q}})\right] + F((\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n} xx^T n^{-\frac{q+1}{p_{\min}+q}}\right])^{-1}) \\
&= \sum_{i=1}^{n} c_i(n^{-\frac{q+1}{p_{\min}+q}}) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}}\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right])^{-1}) \\
&\leq \sum_{i=1}^{n} n^{-p_{\min}\frac{q+1}{p_{\min}+q}} c_i(1) + F((n^{\frac{p_{\min}-1}{p_{\min}+q}}\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right])^{-1}) \qquad \text{(D.12)} \\
&= n^{-p_{\min}\frac{q+1}{p_{\min}+q}} \sum_{i=1}^{n} c_i(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right])^{-1}) \qquad \text{(D.13)} \\
&\leq n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}} c_{\max}(1) + n^{\frac{q(1-p_{\min})}{p_{\min}+q}} F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right]))^{-1}) \qquad \text{(D.14)} \\
&= n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}}\left(c_{\max}(1) + F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right]))^{-1})\right), \qquad \text{(D.15)}
\end{aligned}$$

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 (D.12), the homogeneity of F (Assumption 6.3) in (D.13), and the theorem's assumption, which implies that $c_i(1) \leq c_{\max}(1)$ for all i, in (D.14).

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 (8.9) holds with $D = \left(c_{\max}(1) + F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right]))^{-1})\right)$.

Lower bound

By (D.15), $\phi(\boldsymbol{\lambda}^*) \leq n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}}(c_{\max}(1) + F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^T\right])^{-1}))$. 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 (D.16). 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 (D.17) and $c_i(\ell_{\max}) \geq c_{\min}(\ell_{\max})$ to obtain (D.18):

$$\left(\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\frac{1}{n}c_i(\lambda_i(x))\right]\right)^{\frac{1}{p_{\max}}} \geq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\frac{1}{n}\left(c_i(\lambda_i(x))\right)^{\frac{1}{p_{\max}}}\right] \tag{D.16}$$

$$\geq \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\frac{1}{n}((\lambda_i(x)/\ell_{\max})^{p_{\max}} c_i(\ell_{\max}))^{\frac{1}{p_{\max}}}\right] \tag{D.17}$$

$$\geq \frac{(c_{\min}(\ell_{\max}))^{\frac{1}{p_{\max}}}}{\ell_{\max}}\frac{1}{n}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\lambda_i(x)\right]. \tag{D.18}$$

This shows that

$$\begin{aligned}\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\lambda_i^*(x)\right] &\leq \frac{n\ell_{\max}}{(c_{\min}(\ell_{\max}))^{1/p_{\max}}}\left(\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\frac{1}{n}c_i(\lambda_i^*(x))\right]\right)^{\frac{1}{p_{\max}}} \\ &\leq \frac{n\ell_{\max}}{(c_{\min}(\ell_{\max}))^{1/p_{\max}}}\left(\frac{1}{n}\phi(\boldsymbol{\lambda}^*)\right)^{1/p_{\max}} \\ &\leq \frac{n\ell_{\max}}{(c_{\min}(\ell_{\max}))^{1/p_{\max}}}\left(\frac{1}{n}n^{-\frac{q(p_{\min}-1)}{p_{\min}+q}}\left(c_{\max}(\ell_{\max}) + F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T\right])^{-1})\right)\right)^{1/p}\end{aligned} \tag{D.19}$$

where we used (D.18) for the first inequality, the fact that $C_{\text{estim}}(\boldsymbol{\lambda}) \geq 0$ for the second and (D.15) to obtain the last inequality.

Note that the exponent of n in (D.19) 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 8.4.

Plugging this into (D.19) yields the upper bound on the total amount of precision given by all players:

$$\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_{i=1}^{n}\lambda_i^*(x)\right] \leq \ell_{\max}\left(1+\frac{F((\mathbb{E}_{\mu_{\text{joint}}}\left[\frac{1}{n}\sum_{i=1}^{n}x_i x_i^T\right])^{-1})}{c_{\min}(\ell_{\max})}\right)^{\frac{1}{p_{\max}}} n^{\frac{p_{\min}-1}{p_{\min}+q}+\alpha/q}. \tag{D.20}$$

Recall that $\nu_{\boldsymbol{\lambda}^*}(x) = \sum_i \lambda_i(x)\mu(x)$. Following what we did in (D.10) with the notation $b = \sum_{x\in\mathcal{X}}\nu_{\boldsymbol{\lambda}^*}(x) = \mathbb{E}_{\mu_{\text{joint}}}\left[\sum_i \lambda_i^*(x)\right]$, we have

$$C_{\text{estim}}(\boldsymbol{\lambda}^*) \geq \left(\mathbb{E}_{\mu_{\text{joint}}}\left[\sum_i \lambda_i^*(x)\right]\right)^{-q} F\left(\left(\sum_{x\in\mathcal{X}} xx^T\nu^*(x)\right)^{-1}\right). \tag{D.21}$$

Combining (D.21) and (D.20) shows that the right-hand-side of (8.9) holds with

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

□

E

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 will be made publicly available.

List of Figures

Bibliography

[Aba+16] Martin Abadi, Andy Chu, Ian Goodfellow, et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 2016, pp. 308–318. cit. on p. 26

[Abe+15] 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 '15)*. 2015, pp. 619–636. cit. on pp. 5, 62, A48

[ADT07] Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum experimental designs, with SAS*. Oxford University Press New York, 2007. cit. on pp. 63, 70, 92

[Ait35] A. C. Aitken. "On Least Squares and Linear Combinations of Observations". In: *Proceedings of the Royal Society of Edinburgh* 55 (1935), pp. 42–48. cit. on p. 15

[AS00] Rakesh Agrawal and Ramakrishnan Srikant. "Privacy-preserving Data Mining". In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. 2000, pp. 439–450. cit. on p. 63

[Ata+99] Mike Atallah, Elisa Bertino, Ahmed Elmagarmid, Mohamed Ibrahim, and Vassilios Verykios. "Disclosure limitation of sensitive rules". In: *Workshop on Knowledge and Data Engineering Exchange (KDEX'99)*. 1999, pp. 45–52. cit. on p. 63

[AZ97] Christopher Avery and Richard Zeckhauser. "Recommender Systems for Evaluating Computer Messages". In: *Commun. ACM* 40.3 (Mar. 1997), pp. 88–89. cit. on pp. 4, A47

[Bal+15] Maria-Florina Balcan, Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. "Commitment Without Regrets: Online Learning in Stackelberg Security Games". In: *Proceedings of EC*. 2015, pp. 61–78. cit. on p. 26

[Bar+10] Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. "The security of machine learning". In: *Machine Learning* 81.2 (2010), pp. 121–148. cit. on p. 24

[BCM19] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. "Lower Bounds on Adversarial Robustness from Optimal Transport". In: *In Advances in Neural Information Processing Systems (NIPS)*. 2019. cit. on pp. 5, 25, 26, 113, A48, A56

[Bel07] David Bellhouse. "The problem of Waldegrave". In: *Electronic Journal for the History of Probability and Statistics* 3.2 (2007), pp. 1–12. cit. on p. 15

[BG20] 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. cit. on p. 62

[BKS12] Michael Brückner, Christian Kanzow, and Tobias Scheffer. “Static Prediction Games for Adversarial Learning Problems”. In: *Journal of Machine Learning Research* 13 (2012), pp. 2617–2654. cit. on pp. 24, 25

[Boš+11] Branislav Bošanský, Viliam Lisý, Michal Jakob, and Michal Pěchouček. “Computing Time-dependent Policies for Patrolling Games with Mobile Targets”. In: *Proceedings of AAMAS*. 2011, pp. 989–996. cit. on p. 26

[Bro+16] Matthew Brown, Arunesh Sinha, Aaron Schlenker, and Milind Tambe. “One Size Does Not Fit All: A Game-Theoretic Approach for Dynamically and Effectively Screening for Threats”. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016, pp. 425–431. cit. on p. 26

[BS11] Michael Brückner and Tobias Scheffer. “Stackelberg games for adversarial prediction problems”. In: *Proceedings of ACM SIGKDD*. 2011, pp. 547–555. cit. on pp. 24, 25

[BT19] Omer Ben-Porat and Moshe Tennenholtz. “Regression Equilibrium”. In: *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*. 2019, pp. 173–191. cit. on p. 62

[Bul+16] Samuel Rota Bulò, Battista Biggio, Ignazio Pillai, Marcello Pelillo, and Fabio Roli. “Randomized prediction games for adversarial machine learning”. In: *IEEE transactions on neural networks and learning systems* 28.11 (2016), pp. 2466–2478. cit. on p. 25

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. cit. on pp. 63, 64, 70, 92, 93

[CA13] Anil Kumar Chorppath and Tansu Alpcan. “Trading privacy with incentives in mobile commerce: A game theoretic approach”. In: *Pervasive and Mobile Computing* 9.4 (2013), pp. 598–612. cit. on p. 62

[Car+10] Felipe Caro, Jeremie Gallien, Miguel Diaz, et al. “Zara uses operations research to reengineer its global distribution process”. In: *Interfaces* (2010). cit. on pp. 3, 4, A45, A46

[CBM18] Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. “PAC-learning in the presence of adversaries”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 31. 2018, pp. 230–241. cit. on pp. 5, 25, 36, A48

[CD06] Xi Chen and Xiaotie Deng. “Settling the complexity of two-player Nash equilibrium”. In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)*. IEEE. 2006, pp. 261–272. cit. on p. 19

[CDP15] Y. Cai, C. Daskalakis, and C. H. Papadimitriou. “Optimum statistical estimation with strategic data sources”. In: *Proceedings of the 28th Annual Conference on Learning Theory (COLT 2015)*. 2015, pp. 40.1–40.40. cit. on pp. 4, 61, A47

[CGL15] 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. cit. on pp. 64, 73

[CH12] Richard Cornes and Roger Hartley. "Fully aggregative games". In: *Economics Letters* 116.3 (2012), pp. 631–633. cit. on p. 64

[Che+18a] 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)*. 2018, pp. 27–44. cit. on p. 62

[Che+18b] 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)*. 2018, pp. 9–26. cit. on pp. 4, 62, 63, A47

[CIL15] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. "Truthful Linear Regression". In: *Proceedings of the 28th Annual Conference on Learning Theory (COLT 2015)*. Vol. 40. 2015, pp. 1–36. cit. on p. 63

[CL09] Lin Chen and Jean Leneutre. "A game theoretical framework on intrusion detection in heterogeneous networks". In: *IEEE Transactions on Information Forensics and Security* 4.2 (2009), pp. 165–178. cit. on p. 26

[CLP20] Yiling Chen, Yang Liu, and Chara Podimata. "Learning Strategy-Aware Linear Classifiers". In: *In Advances in Neural Information Processing Systems (NIPS)*. 2020. cit. on p. 50

[CPS16] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. "Truthful Univariate Estimators". In: *Proceedings of the 33rd International Conference on Machine Learning (ICML '16)*. 2016. cit. on p. 62

[CSZ20] Yiling Chen, Yiheng Shen, and Shuran Zheng. "Truthful Data Acquisition via Peer Prediction". In: *In Advances in Neural Information Processing Systems (NIPS)*. 2020. cit. on pp. 5, 62, A48

[CZ19] 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)*. 2019, pp. 659–677. cit. on p. 62

[Dal+04] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. "Adversarial Classification". In: *Proceedings of ACM KDD*. 2004, pp. 99–108. cit. on pp. 3, 24, 25, A46

[DCM20] Prithviraj Dasgupta, Joseph B Collins, and Michael McCarrick. "Improving Costs and Robustness of Machine Learning Classifiers Against Adversarial Attacks via Self Play of Repeated Bayesian Games". In: *The Thirty-Third International Flairs Conference*. 2020. cit. on p. 24

[Deb52] Gerard Debreu. "A social equilibrium existence theorem". In: *Proceedings of the National Academy of Sciences* 38.10 (1952), pp. 886–893. cit. on p. 17

[Dek+05] Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media, 2005. cit. on p. 70

[DFI14] Pranav Dandekar, Nadia Fawaz, and Stratis Ioannidis. "Privacy Auctions for Recommender Systems". In: *ACM Trans. Econ. Comput.* 2.3 (July 2014), 12:1–12:22. cit. on p. 63

[DFP10a] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. "Incentive compatible regression learning". In: *Journal of Computer and System Sciences* 76.8 (2010), pp. 759–777. cit. on pp. 4, A47

[DFP10b] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. "Incentive compatible regression learning". In: *Journal of Computer and System Sciences* 76.8 (2010), pp. 759–777. cit. on pp. 62, 63

[DG13] Anirban Dasgupta and Arpita Ghosh. "Crowdsourced Judgement Elicitation with Endogenous Proficiency". In: *Proceedings of the 22nd International Conference on World Wide Web (WWW '13)*. 2013, pp. 319–330. cit. on pp. 4, 62, A47

[DGP09] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. "The complexity of computing a Nash equilibrium". In: *SIAM Journal on Computing* 39.1 (2009), pp. 195–259. cit. on p. 19

[DJW13] J.C. Duchi, M.I. Jordan, and M.J. Wainwright. "Local Privacy and Statistical Minimax Rates". In: *Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. 2013, pp. 429–438. cit. on pp. 4, 63, A47

[DL09] Cynthia Dwork and Jing Lei. "Differential privacy and robust statistics". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, pp. 371–380. cit. on p. 26

[DLM12] Lemonia Dritsoula, Patrick Loiseau, and John Musacchio. "Computing the Nash equilibria of intruder classification games". In: *International Conference on Decision and Game Theory for Security*. Springer. 2012, pp. 78–97. cit. on p. 24

[DLM17] Lemonia Dritsoula, Patrick Loiseau, and John Musacchio. "A Game-Theoretic Analysis of Adversarial Classification". In: *IEEE Transactions on Information Forensics and Security* 12.12 (Dec. 2017), pp. 3094–3109. cit. on pp. 5, 24, 25, 29, 30, 31, A48

[DM00] George T Duncan and Sumitra Mukherjee. "Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise". In: *Journal of the American Statistical Association* 95.451 (2000), pp. 720–729. cit. on p. 63

[Dom08] Josep Domingo-Ferrer. "A survey of inference control methods for privacy-preserving data mining". In: *Privacy-preserving data mining*. Springer, 2008, pp. 53–80. cit. on p. 63

[Don+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)*. 2018, pp. 55–70. cit. on p. 62

[DR14] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3-4 (Aug. 2014), pp. 211–407. cit. on p. 62

[Dwo06] Cynthia Dwork. "Differential privacy". In: *International Colloquium on Automata, Languages and Programming (ICALP)*. 2006, pp. 1–12. cit. on p. 63

[ERK11] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. "Collaborative Filtering Recommender Systems". In: *Found. Trends Hum.-Comput. Interact.* 4.2 (Feb. 2011), pp. 81–173. cit. on pp. 4, A47

[ES07] Wendy Nelson Espeland and Michael Sauder. "Rankings and reactivity: How public measures recreate social worlds". In: *American journal of sociology* (2007). cit. on pp. 3, A45

[ESE16] Wendy Nelson Espeland, Michael Sauder, and Wendy Espeland. *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation, 2016. cit. on pp. 3, A45

[Fan52] Ky Fan. "Fixed-point and minimax theorems in locally convex topological linear spaces". In: *Proceedings of the National Academy of Sciences of the United States of America* 38.2 (1952), p. 121. cit. on p. 17

[FCK15] Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. "Elicitation for Aggregation". In: *Proceedings of the 29th Conference on Artificial Intelligence (AAAI '15)*. 2015. cit. on p. 62

[FJT13] Fei Fang, Albert Xin Jiang, and Milind Tambe. "Optimal Patrol Strategy for Protecting Moving Targets with Multiple Mobile Resources". In: *Proceedings of AAMAS*. 2013, pp. 957–964. cit. on p. 26

[For92] Françoise Forges. "Chapter 6 Repeated games of incomplete information: Non-zero-sum". In: *Handbook of Game Theory with Economic Applications*. Ed. by Robert Aumann and Sergiu Hart. Vol. 1. Elsevier, 1992, pp. 155–177. cit. on p. 30

[FT91] D. Fudenberg and J. Tirole. *Game Theory*. MIT press, 1991. cit. on p. 15

[Gas+20] Nicolas Gast, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon. "Linear Regression from Strategic Data Sources". In: *ACM Trans. Econ. Comput.* 8.2 (May 2020). cit. on pp. 61, 64, 67, 73

[Gir+] Jairo Giraldo, Alvaro A Cardenas, Murat Kantarcioglu, and Jonathan Katz. "Adversarial Classification Under Differential Privacy". In: (). cit. on p. 26

[GKN17] Robin C Geyer, Tassilo Klein, and Moin Nabi. "Differentially private federated learning: A client level perspective". In: *arXiv preprint arXiv:1712.07557* (2017). cit. on pp. 4, 67, A47

[Gli52] I. L. Glicksberg. “A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points”. In: *Proceedings of the American Mathematical Society* 3.1 (1952), pp. 170–174. cit. on pp. 17, 54

[Goo75] C. Goodhart. *Monetary Relationships: A View from Threadneedle Street*. University of Warwick Library Working Paper Collection. University of Warwick, 1975. cit. on pp. 2, A45

[GR06] A. Globerson and S. Roweis. “Nightmare at Test Time: Robust Learning by Feature Deletion”. In: *Proceedings of ICML*. 2006. cit. on p. 24

[GR11] Arpita Ghosh and Aaron Roth. “Selling Privacy at Auction”. In: *Proceedings of the 12th ACM Conference on Electronic Commerce (EC)*. 2011, pp. 199–208. cit. on pp. 62, 63

[Gre03] William H Greene. *Econometric analysis*. Pearson Education India, 2003. cit. on p. 11

[Gro+13] Michael Großhans, Christoph Sawade, Michael Brückner, and Tobias Scheffer. “Bayesian Games for Adversarial Regression Problems”. In: *Proceedings of ICML*. 2013, pp. III-55–III-63. cit. on pp. 4, A46

[GSS15] Ian Goodfellow, Jon Shlens, and Christian Szegedy. “Explaining and Harnessing Adversarial Examples”. In: *ICLR*. arXiv:1412.6572. 2015. cit. on pp. 2, 24, 113, A44, A56

[Gyö+02] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002. cit. on p. 114

[Har+05] 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)*. 2005, pp. 307–316. cit. on pp. 4, A47

[Har+16] 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 ’16)*. 2016, pp. 111–122. cit. on p. 62

[HIM14] Thibaut Horel, Stratis Ioannidis, and S Muthukrishnan. “Budget Feasible Mechanisms for Experimental Design”. In: *Proceedings of the 11th Latin American Theoretical INformatics Symposium (LATIN)*. 2014, pp. 719–730. cit. on p. 64

[HS19] Safwan Hossain and Nisarg Shah. “Pure Nash Equilibria in Linear Regression”. Preprint. 2019. cit. on p. 62

[HS20] 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. cit. on p. 62

[Hua+11] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. “Adversarial machine learning”. In: *Proceedings of ACM AISec*. 2011, pp. 43–58. cit. on p. 24

[IL13a] 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)*. 2013, pp. 277–290. cit. on pp. 61, 73

[IL13b] 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)*. 2013, pp. 277–290. cit. on p. 64

[Imm+11] Nicole Immorlica, Adam Tauman Kalai, Brendan Lucier, et al. “Dueling Algorithms”. In: *Proceedings of STOC*. 2011, pp. 215–224. cit. on p. 27

[Jab+17] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. “Fairness in reinforcement learning”. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1617–1626. cit. on p. 26

[JT04] Ramesh Johari and John N. Tsitsiklis. “Efficiency Loss in a Network Resource Allocation Game”. In: *Mathematics of Operations Research* 29.3 (2004), pp. 407–435. cit. on p. 80

[KCP10] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. “Complexity of Computing Optimal Stackelberg Strategies in Security Resource Allocation Games”. In: *Proceedings of AAAI*. 2010, pp. 805–810. cit. on p. 27

[Kie+09] Christopher Kiekintveld, Manish Jain, Jason Tsai, et al. “Computing Optimal Randomized Resource Allocations for Massive Security Games”. In: *Proceedings of AAMAS*. 2009, pp. 689–696. cit. on p. 26

[Kon+16] Jakub Konečny, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. “Federated optimization: Distributed machine learning for on-device intelligence”. In: *arXiv preprint arXiv:1610.02527* (2016). cit. on pp. 4, 67, A47

[Kon+20] 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)*. Apr. 2020, pp. 2095–2102. cit. on p. 62

[KOV16] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. “Extremal Mechanisms for Local Differential Privacy”. In: *Journal of Machine Learning Research* 17.17 (2016), pp. 1–51. cit. on p. 63

[KPH15] Sujin Kim, Raghu Pasupathy, and Shane G Henderson. “A guide to sample average approximation”. In: *Handbook of simulation optimization*. Springer, 2015, pp. 207–243. cit. on p. 44

[KR19] 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)*. 2019, pp. 825–844. cit. on p. 62

[KST12] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. “Private convex empirical risk minimization and high-dimensional regression”. In: *Proceedings of the 25th Annual Conference on Learning Theory (COLT 2012)*. 2012, pp. 25.1–25.40. cit. on p. 63

[Kus+17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness". In: *Advances in Neural Information Processing Systems*. 2017, pp. 4066–4076. cit. on p. 26

[KXC11] Murat Kantarcioglu, Bowei Xi, and Chris Clifton. "Classifier evaluation and attribute selection against active adversaries". In: *Data Mining and Knowledge Discovery* 22.1 (2011), pp. 291–335. cit. on pp. 24, 25

[LC16a] Yang Liu and Yiling Chen. "A Bandit Framework for Strategic Regression". In: *Advances in Neural Information Processing Systems 29 (NIPS)*. 2016, pp. 1821–1829. cit. on pp. 4, 61, A47

[LC16b] Yang Liu and Yiling Chen. "A Bandit Framework for Strategic Regression". In: *Advances in Neural Information Processing Systems 29 (NIPS)*. 2016, pp. 1821–1829. cit. on p. 62

[LKP14] Viliam Lisý, Robert Kessl, and Tomáš Pevný. "Randomized Operating Point Selection in Adversarial Classification". In: *Proceedings of ECML PKDD*. 2014, pp. 240–255. cit. on pp. 24, 25, 37

[LL10] Pavel Laskov and Richard Lippmann. "Machine learning in adversarial environments". In: *Machine Learning* 81.2 (2010), pp. 115–119. cit. on p. 24

[LM05] Daniel Lowd and Christopher Meek. "Adversarial Learning". In: *Proceedings of ACM KDD*. 2005, pp. 641–647. cit. on p. 24

[LR12] Katrina Ligett and Aaron Roth. "Take it or Leave it: Running a Survey when Privacy Comes at a Cost". In: *Proceedings of the 8th International Conference on Web and Internet Economics (WINE)*. 2012, pp. 378–391. cit. on p. 63

[LSW06] Jeff Linderoth, Alexander Shapiro, and Stephen Wright. "The empirical behavior of sampling methods for stochastic programming". In: *Annals of Operations Research* 142.1 (2006), pp. 215–241. cit. on p. 44

[Luo+15] 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 '15)*. 2015, pp. 57–57. cit. on p. 61

[LV14] Bo Li and Yevgeniy Vorobeychik. "Feature Cross-Substitution in Adversarial Classification". In: *Proceedings of NIPS*. 2014, pp. 2087–2095. cit. on pp. 24, 25

[LV15] Bo Li and Yevgeniy Vorobeychik. "Scalable Optimization of Randomized Operational Decisions in Adversarial Classification Settings". In: *Proceedings of AISTATS*. 2015. cit. on pp. 24, 25

[Meu+21] Laurent Meunier, Meyer Scetbon, Rafael Pinot, Jamal Atif, and Yann Chevaleyre. "Mixed Nash Equilibria in the Adversarial Examples Game". In: *arXiv preprint arXiv:2102.06905* (2021). cit. on pp. 5, 25, 113, A48, A56

[Mil+19a] 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*. 2019, pp. 230–239. cit. on p. 26

[Mil+19b] 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*)*. 2019, pp. 230–239. cit. on p. 62

[MMH20] 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. cit. on p. 62

[Mor00] John Morgan. "Financing Public Goods by Means of Lotteries". In: *Review of Economic Studies* 67.4 (Oct. 2000), pp. 761–84. cit. on p. 64

[MPR12] Reshef Meir, Ariel D. Procaccia, and Jeffrey S. Rosenschein. "Algorithms for strategyproof classification". In: *Artificial Intelligence* 186 (2012), pp. 123–156. cit. on p. 62

[MTS12] Janusz Marecki, Gerry Tesauro, and Richard Segal. "Playing Repeated Stackelberg Games with Unknown Opponents". In: *Proceedings of AAMAS*. 2012, pp. 821–828. cit. on p. 26

[MV53] Oskar Morgenstern and John Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953. cit. on p. 16

[MZH19] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. "Data poisoning against differentially-private learners: Attacks and defenses". In: *arXiv preprint arXiv:1903.09860* (2019). cit. on p. 26

[Nas51] John Nash. "Non-cooperative games". In: *Annals of mathematics* (1951), pp. 286–295. cit. on pp. 15, 17

[Nel+09] Blaine Nelson, Marco Barreno, Fuching Jack Chi, et al. "Misleading Learners: Co-opting Your Spam Filter". In: *Machine Learning in Cyber Trust: Security, Privacy, and Reliability*. Ed. by Philip S. Yu and Jeffrey J. P. Tsai. Springer, 2009. cit. on pp. 2, A44

[Nel+10] B. Nelson, B. I. P. Rubinstein, L. Huang, et al. "Near Optimal Evasion of Convex-Inducing Classifiers". In: *Proceedings of AISTATS*. 2010. cit. on p. 24

[Ney97] Abraham Neyman. "Correlated equilibrium and potential games". In: *International Journal of Game Theory* 26.2 (June 1997), pp. 223–227. cit. on p. 75

[NST12] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. "Approximately Optimal Mechanism Design via Differential Privacy". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*. 2012, pp. 203–213. cit. on p. 63

[OR94] Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994. cit. on p. 16

[OZ03] Stanley RM Oliveira and Osmar R Zaiane. "Privacy Preserving Clustering by Data Transformation." In: *SBBD*. 2003, pp. 304–318. cit. on p. 63

[Pap+16] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks". In: *Proceedings of IEEE S&P*. May 2016. cit. on p. 24

[Pap+18] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. "Towards the Science of Security and Privacy in Machine Learning". In: *Proceedings of IEEE EuroS&P*. Apr. 2018. cit. on p. 24

[Pha+19] NhatHai Phan, My T Thai, Ruoming Jin, Han Hu, and Dejing Dou. "Preserving differential privacy in adversarial learning with provable robustness". In: *arXiv preprint arXiv:1903.09822* (2019). cit. on p. 26

[Pin+20] Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. "Randomization matters. How to defend against strong adversarial attacks". In: *Proceedings of ICML*. 2020. cit. on pp. 5, 25, 26, A48

[Pit+09] James Pita, Manish Jain, Fernando Ordóñez, et al. "Using Game Theory for Los Angeles Airport Security". In: *AI Magazine* 30 (2009), pp. 43–57. cit. on p. 26

[PP04] Javier Perote and Juan Perote-Pena. "Strategy-proof estimators for simple regression". In: *Mathematical Social Sciences* 47.2 (2004), pp. 153–176. cit. on pp. 3, 4, 62, A45, A47

[PS19a] Juan C Perdomo and Yaron Singer. "Robust attacks against multiple classifiers". In: *arXiv preprint arXiv:1906.02816* (2019). cit. on pp. 113, A57

[PS19b] Juan C. Perdomo and Yaron Singer. "Robust Attacks against Multiple Classifiers". In: *CoRR* (2019). cit. on p. 25

[Puk06] Friedrich Pukelsheim. *Optimal design of experiments*. Vol. 50. Society for Industrial Mathematics, 2006. cit. on pp. 63, 70, 92

[RT02] Tim Roughgarden and Éva Tardos. "How bad is selfish routing?" In: *Journal of the ACM* 49.2 (Mar. 2002), pp. 236–259. cit. on p. 80

[San10] William H. Sandholm. *Population Games and Evolutionary Dynamics*. MIT Press, 2010. cit. on p. 79

[Sav72] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972. cit. on p. 16

[SB14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014. cit. on p. 11

[Sch+18] Aaron Schlenker, Omkar Thakoor, Haifeng Xu, et al. "Deceiving Cyber Adversaries: A Game Theoretic Approach". In: *Proceedings of AAMAS*. 2018, pp. 892–900. cit. on p. 26

[Sch11] Guido Schäfer. "Online Social Networks and Network Economics". Lecture notes, Sapienza University of Rome. 2011. cit. on p. 79

[SEA20] Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. "Causal Strategic Linear Regression". In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020. cit. on p. 62

[Sha03] Alexander Shapiro. "Monte Carlo sampling methods". In: *Handbooks in operations research and management science* 10 (2003), pp. 353–425. cit. on pp. 44, 45, A5

[SP10] Robin Sommer and Vern Paxson. “Outside the Closed World: On Using Machine Learning For Network Intrusion Detection”. In: *Proceedings of IEEE S&P*. 2010. cit. on pp. 2, A44

[Str97] Marilyn Strathern. “‘Improving ratings’: audit in the British University system”. In: *European Review* (1997). cit. on pp. 2, A45

[SZ16] Nihar B. Shah and Dengyong Zhou. “Double or nothing: Multiplicative incentive mechanisms for crowdsourcing”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5725–5776. cit. on p. 62

[TG20] 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. cit. on p. 62

[Tho+13] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. “Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse”. In: *Proceedings of USENIX Security*. 2013, pp. 195–210. cit. on pp. 2, A44

[TYW84] Joseph F Traub, Yechiam Yemini, and H Woźniakowski. “The statistical security of a statistical database”. In: *ACM Transactions on Database Systems (TODS)* 9.4 (1984), pp. 672–679. cit. on p. 63

[ULB] ULB. *Credit card fraud detection*. `https://www.kaggle.com/mlg-ulb/creditcardfraud/version/3`. cit. on p. 31

[VCZ06] Jaideep Vaidya, Christopher W. Clifton, and Yu Michael Zhu. *Privacy Preserving Data Mining*. Springer, 2006. cit. on p. 63

[Ver+03] Bram Verweij, Shabbir Ahmed, Anton J Kleywegt, George Nemhauser, and Alexander Shapiro. “The sample average approximation method applied to stochastic routing problems: a computational study”. In: *Computational Optimization and Applications* 24.2-3 (2003), pp. 289–333. cit. on p. 44

[VK18] Yevgeniy Vorobeychik and Murat Kantarcioglu. *Adversarial Machine Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool, 2018. cit. on pp. 4, A46

[Von16] John Von Neumann. “1. On the Theory of Games of Strategy”. In: *Contributions to the Theory of Games (AM-40), Volume IV*. Princeton University Press, 2016, pp. 13–42. cit. on p. 15

[Wan+14] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. “Man vs. Machine: Practical Adversarial Detection of Malicious Crowdsourcing Workers”. In: *Proceedings of USENIX Security*. 2014, pp. 239–254. cit. on pp. 2, A44

[War52] John Glen Wardrop. “Road paper. some theoretical aspects of road traffic research.” In: *Proceedings of the institution of civil engineers* 1.3 (1952), pp. 325–362. cit. on p. 15

[War65] Stanley L. Warner. “Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias”. In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69. cit. on pp. 4, A47

[Wes+20] Tyler Westenbroek, Roy Dong, Lillian J. Ratliff, and S. Shankar Sastry. “Competitive Statistical Estimation with Strategic Data Sources”. In: *IEEE Transactions on Automatic Control* 65.4 (2020), pp. 1537–1551. cit. on pp. 4, 61, A47

[Yan+19] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. “Federated machine learning: Concept and applications”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019), pp. 1–19. cit. on pp. 4, 67, A47

[ZCC19] Hanrui Zhang, Yu Cheng, and Vincent Conitzer. “When Samples Are Strategically Selected”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019, pp. 7345–7353. cit. on p. 62

[Zho+12] Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Bowei Xi. “Adversarial Support Vector Machine Learning”. In: *Proceedings of KDD*. 2012, pp. 1059–1067. cit. on pp. 24, 25

[Zin03] Martin Zinkevich. “Online convex programming and generalized infinitesimal gradient ascent”. In: *Proceedings of ICML*. 2003, pp. 928–936. cit. on pp. 50, 51, 52, A6, A7

[ZK14] Yan Zhou and Murat Kantarcioglu. “Adversarial Learning with Bayesian Hierarchical Mixtures of Experts”. In: *Proceedings of SIAM SDM*. 2014, pp. 929–937. cit. on pp. 24, 25

Résumé Détaillé en Français

F

F.1 Chapitre 1: Introduction

F.1.1 Contexte

L'apprentissage automatique est un vaste domaine qui nous permet d'exploiter la quantité toujours croissante de données disponibles pour automatiser des tâches effectuées par des humains (comme les voitures à conduite autonome) ou pour analyser des ensembles de données trop volumineux et complexes pour être traités par des humains lorsque les programmes classiques ne parviennent pas à produire des réponses satisfaisantes. Dans de nombreux scénarios, le processus d'apprentissage se fait en deux étapes distinctes. Tout d'abord, un analyste recueille des données qui peuvent être étiquetées (par exemple, les caractéristiques des transactions bancaires et le caractère frauduleux ou non de ces transactions) ou qui ont une valeur d'intérêt (par exemple, des données médicales où les caractéristiques des patients sont recueillies pour comprendre dans quelles conditions une maladie est grave ou non). Ensuite, l'analyste utilise les données recueillies pour produire un résultat qui peut prendre différentes formes, comme un classificateur (classification des transactions entre frauduleuses et non frauduleuses) ou une régression (modèle associant des caractéristiques à une valeur afin de prédire la valeur d'intérêt dans des points de données futurs).

Échec des schémas d'apprentissage classiques

Une hypothèse commune à de nombreux contextes de ce type est que le processus de collecte des données est, dans un certain sens, indépendant de l'analyse suivante et de son résultat. Dans le domaine de la classification, cela prend généralement la forme de l'hypothèse suivante : "Les données d'apprentissage et les données de test sont tirées de la même distribution". Cela implique que quoi que fasse l'analyste, cela ne changera pas la façon dont les données sont produites par rapport à ce qui s'est passé avant l'analyse. Dans la régression, la variance des points de données ne dépend pas de l'analyse et est censée être un paramètre du problème.

Cependant, dans de nombreuses applications, ces hypothèses ne tiennent pas car le résultat de l'analyse a une valeur stratégique. Dans ce contexte, les agents peuvent modifier soit la façon dont ils produisent les données, soit directement les données qu'ils envoient à l'analyste pour atteindre leur objectif. Cela peut conduire à un conflit entre le processus de génération de données et l'analyse, que nous appelons soit *antagoniste* si les générateurs de données ont des intérêts directement en conflit avec les objectifs de l'analyste, soit simplement *stratégique* dans le cas où leurs intérêts peuvent ou non s'aligner sur ceux de l'analyste.

Les contextes antagonistes surviennent généralement lorsque certains agents sont malveillants. Par exemple, dans la classification, l'objectif de l'analyste peut être de reconnaître les fraudeurs, tels que les comptes Twitter frauduleux étudiés dans Thomas et al. [Tho+13]. Il est clair que les fraudeurs ont un intérêt à ne pas être classés comme fraudeurs et cela conduit à une adaptation pour échapper à la classification. Ce comportement ne se limite toutefois pas à la fraude et peut également être observé lors de la détection d'intrusions dans un réseau ou de la prévention d'attaques DDoS. Il est bien connu que l'utilisation d'algorithmes de classification standard pour cette tâche conduit à des performances médiocres. Les attaquants sont en mesure d'éviter la détection en ajustant les données qu'ils génèrent lors de la conception de leurs attaques (attaques d'évasion) ou de modifier l'ensemble de données d'entraînement de sorte que le classificateur résultant soit peu performant contre eux (attaques d'empoisonnement). Nelson et al. [Nel+09] montrent que les attaquants peuvent facilement tromper un filtre anti-spam en ayant accès à une petite partie des spams utilisés pour entraîner le filtre anti-spam. Goodfellow et al. [GSS15] expliquent pourquoi de nombreux modèles d'apprentissage automatique sont vulnérables aux attaques soigneusement conçues (telles que la modification légère de quelques pixels d'une image et le changement complet de l'étiquette qu'un réseau neuronal lui attribue) appelées exemples antagonistes. Sommer and Paxson [SP10] montrent que le problème de la détection des intrusions dans les réseaux est fondamentalement difficile à aborder du point de vue de l'apprentissage automatique en raison de sa nature contradictoire. Wang et al. [Wan+14] étudient des systèmes de crowdsourcing malveillant (appelés crowdturfing) dans lesquels les attaquants paient les utilisateurs pour qu'ils mènent une série d'attaques et montrent que, si ces attaques peuvent être efficacement détectées par des méthodes d'apprentissage automatique, ces méthodes sont également très vulnérables à l'évasion et à l'empoisonnement. Il existe une vaste littérature sur la classification antagoniste pour pallier cette faiblesse (voir la section 3.1 pour une discussion détaillée), mais ces travaux proposent souvent des méthodes de défense ad-hoc optimisées contre des attaques spécifiques sans modéliser complètement la capacité

d'adaptation de l'attaquant. Cela conduit à une compétition constante, les classifieurs s'adaptant à un type d'attaque spécifique et les attaquants trouvant des moyens de contourner ces défenses.

Les contextes stratégiques sont rencontrés lorsque les utilisateurs n'ont pas d'intentions malveillantes mais ont un certain intérêt dans le résultat de l'analyse qui peut ou non entrer en conflit avec le résultat qu'ils devraient obtenir. Un tel comportement est observé non seulement dans des applications spécifiques impliquant des experts, mais aussi dans la population générale. Par exemple, les acheteurs de maison en Amérique peuvent ouvrir plusieurs cartes de crédit dans le seul but d'améliorer leur score de crédit. Cette omniprésence du comportement stratégique a conduit à des lois célèbres telles que la loi de Goodhart (voir Goodhart [Goo75]) qui stipule que "Toute régularité statistique observée aura tendance à s'effondrer dès qu'une pression sera exercée sur elle à des fins de contrôle" et a été généralisée plus tard par Strathern [Str97] sous la forme de "Lorsqu'une mesure devient une cible, elle cesse d'être une bonne mesure". Parmi tous les exemples possibles de comportement stratégique visant à manipuler une mesure, certains des plus couramment cités sont les enseignants qui subissent des pressions pour obtenir de bons résultats de leurs élèves et se concentrent sur les performances aux tests plutôt que sur l'apprentissage, et les universités qui manipulent les classements en concentrant leurs efforts sur les métriques spécifiques utilisées pour le classement (voir [ES07; ESE16]). Il convient toutefois de noter que le champ d'application de ces lois est plus large que les exemples que nous avons mentionnés et comprend des contextes variés tels que des magasins manipulant la quantité de produits qu'ils demandent comme le montre Caro et al. [Car+10] ou simplement lorsque les agents ne souhaitent pas se démarquer d'un résultat "normal" si le résultat de l'analyse peut avoir un impact sur leur vie comme le discute Perote and Perote-Pena [PP04]. Ces comportements stratégiques affectent les données produites par les agents, ce qui affecte ensuite le résultat de l'analyse. Dans le cas des universités par exemple, les analystes apprennent le rang de chaque université en fonction des caractéristiques manipulées par les universités. Il existe un nombre croissant de travaux sur ces considérations stratégiques, en particulier dans le contexte de la régression linéaire (voir la section 6.1 pour une discussion détaillée), mais ces travaux se concentrent généralement sur la manière de minimiser l'erreur d'estimation dans divers contextes ou sur la manière de garantir des estimateurs à l'épreuve de la stratégie (où les agents révèlent leurs données de manière véridique) et n'étudient pas le large éventail de propriétés statistiques possibles des estimateurs classiques appliqués dans des contextes stratégiques.

Surmonter la dépendance entre la génération de données et l'analyse

Ces dernières années, de nombreux modèles de théorie des jeux ont vu le jour pour contourner l'hypothèse erronée selon laquelle le processus de génération de données est indépendant de l'analyse. En effet, la théorie des jeux est un outil de choix pour modéliser de telles interactions car le concept principal de solution d'équilibre implique que les générateurs de données et l'analyste agissent en tenant compte du comportement stratégique de l'autre.

Plusieurs modèles de classification antagoniste basés sur la théorie des jeux ont vu le jour au cours de la dernière décennie, avec pour pionnier des modèles tels que celui de Dalvi et al. [Dal+04]. (voir la section 3.1 pour une discussion plus détaillée). La plupart d'entre eux présentent cependant plusieurs limites cruciales. En particulier, elles sont profondément liées aux algorithmes d'apprentissage classiques. Ces algorithmes reposent sur une réduction de la classe d'hypothèses (c'est-à-dire l'ensemble des classificateurs possibles parmi lesquels le défenseur choisit sa défense) afin d'obtenir des problèmes d'optimisation traitables (souvent convexes par exemple) pour minimiser l'erreur de classification. Ce lien est à la fois une force de leurs modèles, car ils peuvent être plus susceptibles d'être adoptés par les utilisateurs, et une faiblesse car, alors que les classes d'hypothèses utilisées classiquement représentent généralement une certaine connaissance préalable sur la forme que peut avoir le classificateur optimal, cette connaissance préalable n'a aucune raison d'être valide lorsqu'on considère des données produites de manière antagoniste et peut même être exploitée par les attaquants. Ces modèles de théorie des jeux supposent également une information complète sur l'objectif de l'attaquant,[1] ce qui est souvent trop fort en pratique comme le suggère Vorobeychik and Kantarcioglu [VK18]. Enfin, si certains de ces modèles présentent des garanties théoriques d'erreurs (lorsqu'on considère des ensembles restreints de classificateurs), ils sont généralement destinés à des applications pratiques et il leur manque une compréhension fondamentale de propriétés plus générales.

Les considérations sur le comportement stratégique sont également devenues centrales dans la littérature émergente sur l'apprentissage avec des sources de données stratégiques. (voir la section 6.1 pour une discussion plus détaillée). Dans ces contextes, les agents ne sont pas nécessairement antagonistes à l'analyste mais ont leurs propres objectifs. Par exemple, nous avons mentionné précédemment les magasins qui manipulent la quantité de produits qu'ils demandent (voir [Car+10]). Dans ce cas, les magasins optimisent leurs propres bénéfices tandis que la principale entreprise derrière eux optimise ses bénéfices globaux. De tels problèmes définissent

[1] à l'exception de [Gro+13], mais sur la régression.

une première catégorie de modèles où les agents élaborent des stratégies pour obtenir un résultat souhaité de l'analyse. Cela conduit les auteurs à étudier des estimateurs qui sont à l'épreuve de la stratégie comme dans Perote and Perote-Pena [PP04], Chen et al. [Che+18b], and Dekel et al. [DFP10a]. Ces travaux se concentrent sur des estimateurs ayant la propriété souhaitée qu'aucun agent ne fausse ses données. Dans d'autres contextes, les agents encourent un certain coût pour fournir des données et élaborent une stratégie en fonction de ce coût, car ils peuvent par exemple demander une compensation monétaire pour leurs efforts. C'est le cas dans le crowdsourcing avec le modèle Dasgupta and Ghosh [DG13] ou dans les systèmes de recommandation (voir [ERK11; AZ97; Har+05]) où fournir du contenu ou des commentaires demande un effort, ou dans les applications où les données sont produites par des calculs coûteux. La production de données coûteuses peut également provenir d'informations personnelles sensibles (comme dans les applications médicales), dont la divulgation avec une grande précision entraîne un coût de confidentialité qui pourrait inciter les individus à diminuer la précision de la divulgation et à y ajouter du bruit (d'où une diminution de la précision comme dans Warner [War65] and Duchi et al. [DJW13]). Lorsque la fourniture de données de haute qualité a un coût, il est logique d'envisager un comportement stratégique entre les sources de données. En particulier, il convient de se demander pourquoi les sources de données stratégiques fournissent des données, quelles qu'elles soient. Cette littérature examine principalement la conception de mécanismes d'incitation monétaire pour optimiser l'erreur du modèle en supposant que les agents maximisent leurs incitations moins leurs coûts de fourniture individuels, voir par exemple Cai et al. [CDP15], Liu and Chen [LC16a], and Westenbroek et al. [Wes+20] et ses références. Dans de nombreuses applications, cependant, le modèle sous-jacent comporte également une *composante de bien public*—-c'est-à-dire que les agents *également* bénéficient de la précision du modèle. C'est le cas dans les systèmes de recommandation (où les utilisateurs bénéficient de la qualité globale du service), les applications médicales (où les individus bénéficient de l'analyse des données grâce à l'amélioration des traitements ou à de meilleurs conseils en matière de santé), l'apprentissage fédéré (voir [Yan+19; Kon+16; GKN17]), etc. Un problème supplémentaire dans ce type d'applications est que le nombre d'agents participants est généralement important, ce qui entraîne un degré d'incertitude proportionnel concernant l'état ou les incitations des autres agents.

Perspective de la thèse

Comme le montre notre discussion bibliographique précédente, le problème de la caractérisation des propriétés fondamentales des algorithmes d'apprentissage lorsque les données sont produites de manière stratégique, voire adversariale, n'est pas suffisamment abordé. À notre connaissance, seuls Dritsoula et al. [DLM17], Cullina et al. [CBM18], and Bhagoji et al. [BCM19] caractérisent certaines propriétés classiques d'apprentissage dans des contextes antagonistes, tandis que Pinot et al. [Pin+20] and Meunier et al. [Meu+21] posent les bases de la théorie des jeux pour une extension ultérieure en montrant que des équilibres de Nash mixtes existent dans un modèle d'importance pratique. Pour les considérations stratégiques (en particulier dans la régression), les auteurs s'intéressent généralement à la manière de minimiser les erreurs (voir Abernethy et al. [Abe+15]) ou de garantir l'obtention de données véridiques (voir Chen et al. [CSZ20]) alors que le champ des propriétés statistiques classiques à considérer est beaucoup plus large. En outre, les applications considérées rassemblent généralement une très grande quantité de données provenant de sources variées et la caractérisation des propriétés et des objectifs précis des différents acteurs est irréaliste. Par exemple, les étudiants en compétition avec d'autres étudiants n'ont aucun moyen raisonnable de déterminer les caractéristiques exactes de leurs concurrents, tandis que les réseaux font face à de nombreuses menaces différentes avec des degrés de gravité variables.

Cette thèse étudie ces problèmes où les résultats font défaut. En particulier, nous étudions les problèmes d'apprentissage en présence de sources de données stratégiques en utilisant une approche de théorie des jeux et en considérant le concept de solution de l'équilibre de Nash qui modélise finement l'adaptation des producteurs de données et de l'analyste aux actions de l'autre. Nous nous intéressons aux propriétés d'apprentissage fondamentales des modèles applicables à des contextes où une grande quantité de données est collectée. Cela signifie que nous devons garantir l'évolutivité de nos méthodes tant en termes de puissance de calcul que d'informations requises. Nous visons à développer des modèles et des méthodes de solution applicables à un large éventail de contextes couvrant à la fois des contextes adverses et stratégiques, allant de la fraude bancaire et de la détection d'intrusion dans les réseaux à l'apprentissage avec des informations coûteuses comme on le voit dans l'apprentissage fédéré ou les contextes médicaux.

F.2 Contributions et organisation de la thèse

Dans cette thèse, nous nous concentrons sur les problèmes d'apprentissage où le processus de génération de données est *non* indépendant du résultat de l'analyse et nous modélisons de tels contextes en utilisant la théorie des jeux qui nous permet de considérer l'adaptation du processus de génération de données et de l'analyste aux actions de l'autre partie. Tout au long de la thèse, nous nous concentrons sur les deux questions clés suivantes :

1. Les résultats d'apprentissage fondamentaux sont-ils toujours valables lorsque le processus de génération de données dépend du résultat de l'analyse ?

2. Les modèles de la théorie des jeux peuvent-ils raisonnablement être appliqués dans des contextes d'apprentissage où il existe une grande quantité de données complexes disponibles ?

Nous étudions les problèmes de classification antagoniste et de régression linéaire stratégique à travers ces questions. En particulier, d'un point de vue large, nous montrons que les résultats fondamentaux de l'apprentissage sont remis en question : l'apprentissage antagoniste optimal peut être réalisé avec des classificateurs simples et les estimateurs classiques issus de la régression linéaire ne sont plus optimaux ni même consistants lorsque les données sont produites de manière stratégique.

Cette thèse est organisée en trois parties. La partie I (Chapitres 3, 4 et 5) est dédiée au problème de la *classification antagoniste*. La partie II (Chapitres 6, 7 et 8) présente nos résultats sur le problème *régression linéaire stratégique*. En raison des différences entre les deux cadres, nous passons en revue dans chaque partie la littérature correspondante. Enfin, nous tirons nos conclusions dans la partie III et discutons des travaux futurs potentiels.

F.3 Chapitre 2: Notions Essentielles

Nous passons en revue dans ce chapitre certains concepts importants existants qui sont nécessaires pour introduire les résultats de cette thèse. La section 2.1 présente deux problèmes d'apprentissage automatique omniprésents et des résultats quantifiant la qualité et la complexité de l'apprentissage dans ces problèmes. La section 2.2 présente certaines des définitions fondamentales de la théorie des jeux avec des concepts de solution de base. Notez que ce chapitre existe par souci d'autosuffisance et ne constitue pas une introduction complète aux sujets que nous

présentons. Nous omettons donc de nombreux résultats fondamentaux qui ne sont pas nécessaires à la thèse. Nous indiquons cependant au lecteur les livres pertinents dans chaque section pour une introduction approfondie aux sujets concernés.

F.4 Chapitre 3: Modèle et État de l'art (Classification)

Ce chapitre introduit la première partie des travaux de thèse sur le classification antagoniste. Dans cette partie (Chapitres 3, 4 et 5), nous présentons notre premier exemple d'apprentissage avec des données stratégiques, ou dans ce cas antagonistes. Plus précisément, nous abordons le problème de la classification antagoniste où un défenseur vise à classifier les vecteurs entrants d'attaques et de non-attaques et où les attaquants visent à échapper à la classification. Nous modélisons cette interaction par un jeu dans lequel le défenseur reçoit des vecteurs provenant soit d'attaquants, soit de non-attaquants. Les attaquants choisissent le vecteur qu'ils utilisent parmi un ensemble fini de vecteurs $\mathcal{V}$ et les non-attaquants suivent une distribution fixe sur $\mathcal{V}$. Le défenseur choisit un classificateur sans aucune restriction a priori sur l'ensemble des classificateurs possibles. Cela soulève deux problèmes principaux. Tout d'abord, le nombre de classificateurs possibles est exponentiel en $|Vset|$. Ensuite, $\mathcal{V}$ est lui-même exponentiel en le nombre de caractéristiques utilisées pour classifier. En effet, même en considérant le cas le plus simple de k caractéristiques binaires, on obtient un ensemble de vecteurs de taille 2^k et un ensemble de classificateurs possibles $2^{\mathcal{V}}$ de taille 2^{2^k}. Nous nous concentrons sur deux questions clés :

1. Quels classificateurs le défenseur devrait-il utiliser à l'équilibre ?
2. Comment le défenseur peut-il calculer des stratégies optimales de manière évolutive (en nombre de caractéristiques) ?

Nous montrons que l'aléatoire dans la stratégie est cruciale pour les défenses optimales mais, de manière surprenante, le défenseur peut se défendre contre les attaques de manière optimale en utilisant une classe de classificateurs de faible complexité (de VC-dimension 1) en utilisant une paramétrisation dépendant des gains des attaquants. Cette paramétrisation nous permet à son tour de développer des méthodes d'approximation qui peuvent se généraliser à des vecteurs inconnus à la fois hors ligne et en ligne.

Ce chapitre passe en revue les travaux existants liés de manière générale à la classification antagoniste et à notre modèle et présente notre jeu de classification

antagoniste. Dans le chapitre 4 nous caractérisons ensuite l'équilibre et montrons en particulier que la stratégie optimale du défenseur peut être paramétrée avec peu de paramètres. Ceci répond à notre première question. Nous répondons ensuite à notre deuxième question dans le Chapitre 5 où nous montrons que notre caractérisation précédente peut être utilisée avec des paramètres approximatifs pour fournir des stratégies approximativement optimales à la fois hors ligne en utilisant un ensemble de données existant et en ligne. De plus, ces méthodes ne nécessitent la connaissance que de quelques paramètres du jeu.

F.5 Chapitre 4: Caractérisation de l'équilibre

Dans ce chapitre, nous caractérisons complètement l'équilibre du jeu de classification adversatif. Dans la section 4.1, nous caractérisons la stratégie de défense optimale *aléatoire*. Nous exposons une classe de fonctions de probabilité de détection qui peuvent être réalisées par des classificateurs à seuil et qui sont suffisantes pour se défendre de manière optimale contre les attaques. En particulier, les fonctions de probabilité de détection que nous considérons peuvent être exprimées avec peu de paramètres et sont de faible pseudo-dimension, ce qui ouvre la voie à nos méthodes pour calculer la défense optimale de manière évolutive dans le chapitre 5. Dans la section 4.2, nous caractérisons la stratégie d'attaque optimale en réponse à une stratégie de défense. Notre résultat montre que la stratégie de l'attaquant doit trouver un équilibre simple entre les coûts de fausse alarme que le défenseur doit endurer pour les détecter et leur gain. Enfin, nous illustrons la forme de l'équilibre dans la Section 4.3 sur le jeu simple 1.

F.6 Chapitre 5: Défense Extensible en Ligne et Hors Ligne

Nous répondons dans ce chapitre à notre deuxième question principale : "Comment le défenseur peut-il calculer des stratégies optimales d'une manière extensible ?". Dans ce chapitre, nous fournissons deux réponses exploitant notre caractérisation basse dimensionnelle de la stratégie du défenseur en fonction des données dont il dispose. Dans la section 5.1, nous abordons ce problème sous l'angle de l'optimisation stochastique en supposant que le défenseur a accès à des données historiques. Nous montrons que notre problème est bien adapté aux techniques

d'optimisation stochastique en raison de sa nature polyédrique. Dans la section 5.2, nous supposons que le défenseur n'a accès qu'à des informations en ligne et nous fournissons un algorithme à faible regret pour le regret dit de Stackelberg. Ces deux méthodes nécessitent une connaissance limitée des paramètres du jeu (elles ne nécessitent notamment pas la connaissance de P_0 et p_a, respectivement la probabilité de distribution caractérisant les non-attaquants et la probabilité de présence d'un attaquant). Nous montrons enfin dans la Section 5.3 que ces deux méthodes peuvent être appliquées dans un cadre plus large où $\mathcal{V}$ n'est plus fini mais compact sous de légères hypothèses techniques. Enfin, nous étendons notre modèle pour couvrir un mélange d'attaquants stratégiques et non stratégiques dans la section 5.4. Nos résultats indiquent que dans ce contexte, un mélange entre notre caractérisation à faible dimension de la stratégie du défenseur et les algorithmes d'apprentissage classiques peut être approprié.

F.7 Chapitre 6: Modèle et État de l'art (Régression Linéaire)

Dans cette partie (Chapitres 6, 7 et 8), nous présentons notre deuxième exemple d'apprentissage avec des données stratégiques. Plus précisément, nous abordons le problème de la régression linéaire lorsque les données proviennent de sources de données stratégiques. Nous nous concentrons principalement sur la question clé : *Les résultats fondamentaux sur la régression linéaire sont-ils toujours valables lorsque les données sont produites par des sources de données stratégiques ?* Nous répondons par la négative en général. En particulier, il n'est plus garanti que les estimateurs soient *consistants* lorsque le coût de production des données augmente linéairement avec leur précision. De même, l'estimateur `GLS` n'est plus BLUE mais est approximativement optimal lorsqu'on considère une famille restreinte d'estimateurs satisfaisant des hypothèses convenables en pratique.

Ce premier chapitre passe en revue les travaux existants liés de manière générale à la régression linéaire stratégique et présente notre jeu de régression linéaire. Dans le chapitre 7, nous nous concentrons ensuite sur les propriétés théoriques du jeu de régression linéaire en montrant des résultats fondamentaux pour le reste de l'analyse et pour l'applicabilité de notre jeu dans des contextes réels. Nous répondons enfin à notre question principale dans le chapitre 8 où nous étudions la qualité de l'estimation de la régression linéaire dans le cadre stratégique et l'effet des différents estimateurs sur le jeu de régression linéaire.

F.8 Chapitre 7: Résultats Structuraux sur le Jeu

Dans ce chapitre, nous introduisons des résultats structurels de notre jeu qui concernent des aspects de la théorie des jeux. Dans la section 7.1, nous montrons que notre jeu est un jeu potentiel. Bien que cette propriété soit très courante, elle est cruciale dans notre analyse des propriétés statistiques des estimateurs dans le chapitre suivant. Nous montrons ensuite dans la Section 7.2 que nos hypothèses nous permettent de considérer deux cas particuliers très importants qui peuvent être utilisés pour modéliser des interactions où il y a peu d'agents (le modèle à information complète) ou des interactions où il y a beaucoup d'agents et où il est difficile d'obtenir des informations précises sur les points de données des autres agents (le modèle indépendamment et identiquement distribué). Enfin, dans la section 7.3, nous fournissons quelques limites sur le *prix de stabilité* de notre jeu qui caractérise l'inefficacité sociale du jeu de régression linéaire due à l'égoïsme des agents.

F.9 Chapitre 8: Propriétés de l'estimation Linéaire non Biaisée Dans un Cadre Stratégique

Dans ce chapitre, nous étudions les propriétés statistiques des estimateurs linéaires lorsque les données sont produites de manière stratégique. Nous nous concentrons spécifiquement sur la *qualité* de l'estimation par la scalarisation de la covariance et répondons aux deux questions clés suivantes :

1. L'estimateur `GLS` produit-il la meilleure qualité d'estimation ? En d'autres termes, le coût d'estimation associé à `GLS` à l'équilibre est-il le plus faible parmi les coûts d'estimation associés à tout estimateur linéaire sans biais à l'équilibre ?

2. L'estimation est-elle toujours consistante en présence d'agents stratégiques ? Et si oui, comment le taux de convergence se compare-t-il à celui du cas non stratégique ?

Nous fournissons d'abord dans la section 8.1 une réponse positive et négative à la question 1 – `GLS` produit une estimation approximativement de meilleure qualité parmi une classe d'estimateurs satisfaisant des propriétés statistiques appropriées. Nous caractérisons ensuite dans la Section 8.2 comment la présence d'agents

stratégiques dégrade la qualité de l'estimation et prouvons que l'estimation peut même ne pas être consistante dans ce cas.

F.10 Chapitre 9: Conclusion et Travaux Futurs

F.10.1 Conclusion

Dans cette thèse, nous avons choisi d'étudier les problèmes d'apprentissage lorsque le processus de génération de données peut dépendre de l'analyse et de son résultat. En particulier, nous avons étudié le problème de *classification* et le problème de *régression linéaire* avec deux questions clés auxquelles il faut répondre :

1. Les résultats fondamentaux de l'apprentissage sont-ils toujours valables lorsque le processus de génération des données dépend du résultat de l'analyse ?
2. Les modèles de la théorie des jeux peuvent-ils raisonnablement être appliqués dans des contextes d'apprentissage où il existe une grande quantité de données complexes disponibles ?

Nous avons répondu à la question 1 par la négative en général, en montrant que l'apprentissage en présence de données stratégiques nécessite une prise en compte attentive de propriétés et de paramètres par ailleurs bien connus. Pour la classification, nous avons montré que l'on peut se défendre de manière optimale contre les attaques uniquement par une défense *aléatoire* mais en utilisant des classificateurs *simple* qui peuvent ou non appartenir à des ensembles classiques tels que les classificateurs linéaires. De plus, cette dépendance ne dépend pas des paramètres étudiés dans l'apprentissage classique tels que la distribution des non-attaques ou la probabilité d'attaque mais uniquement de la métrique de la théorie des jeux des gains des attaques détectées et non détectées. Cela souligne le fait que les défenseurs pourraient potentiellement bénéficier énormément de la modélisation des attaquants plutôt que d'appliquer des algorithmes d'apprentissage bien connus sur des classes d'hypothèses classiques comme les classificateurs linéaires ou à noyau. En considérant les problèmes de régression linéaire, nous avons montré que *quelques* résultats tiennent (approximativement), comme l'optimalité de `GLS`, tandis que d'autres sont compromis dans de nombreux contextes - le processus de régression linéaire peut ne pas produire des estimateurs consistants si les agents participants sont réticents à produire des données précises (ce qui se traduit mathématiquement par la linéarité des coûts de fourniture de données) et même lorsqu'ils sont consistants, leur taux de convergence se dégrade. En pratique, cela signifie que les analystes peuvent

sous-estimer le nombre d'agents nécessaires dans une expérience pour atteindre une certaine précision. De plus, nos résultats ont montré que si les analystes ont une limite sur le nombre de participants qu'ils peuvent recruter, ils devraient chercher à éviter les agents dont le coût de production des données augmente linéairement avec la précision. Ceci est encore plus pertinent lorsque les analystes n'ont pas accès à la précision de chaque point de données et utilisent l'estimateur `OLS`, auquel cas même un seul agent stratégique avec un coût de fourniture élevé peut ruiner l'estimation. Le taux de convergence des estimateurs est également en conflit direct avec la qualité de l'allocation des données, car nous avons montré que les agents réticents à fournir des données précises sont particulièrement prudents dans leur utilisation des ressources, ce qui conduit à des allocations optimales, tandis que les agents disposés à fournir des données le font de manière plus uniforme, quel que soit le degré d'information de leurs points de données.

Nous avons fourni des méthodes d'entraînement et des méthodes d'approximation pour répondre à la question 2. Plus précisément, nous avons montré qu'un défenseur pouvait entraîner des classificateurs quasi optimaux simplement en ayant accès aux paramètres décrivant le coût des fausses alarmes, des vrais positifs et des faux négatifs associés à un comportement. Cet entraînement peut être effectué en ligne ou hors ligne avec un ensemble de données et ne nécessite pas l'accès à des paramètres difficiles à évaluer tels que le comportement des non-attaquants ou même la probabilité d'une attaque. En pratique, cela signifie que notre modèle peut être appliqué de manière similaire à un modèle d'apprentissage classique utilisant un ensemble de données ou des informations en ligne. Nous pensons que notre modèle pourrait potentiellement être appliqué à des contextes de fraude bancaire où la prise en compte du problème de classification pondérée (où une mauvaise classification de la transaction a des conséquences différentes selon les caractéristiques de la transaction) est cruciale. Nos expériences sur un ensemble de données réelles de fraude bancaire (bien qu'avec un modèle simpliste) montrent également que des ensembles de données de taille raisonnable peuvent être suffisants pour former des défenses robustes aux attaquants adaptatifs. Pour la régression linéaire, nous avons montré que les agents n'ont pas besoin d'avoir des informations complètes sur les points de données des autres agents lorsque les points de données sont distribués de manière indépendante et identique. Nous avons montré qu'il suffit d'avoir accès à la distribution sous-jacente des données, qui peut être raisonnablement estimée dans de nombreux contextes avec des données accessibles au public. En particulier, les deux modèles mentionnés précédemment donnent des équilibres équivalents (en termes de coût) et interchangeables (en termes de stratégies).

F.10.2 Futurs travaux

De nombreux travaux futurs potentiels suivent la ligne de travail de cette thèse qui consiste à étudier les propriétés fondamentales des algorithmes d'apprentissage lorsque les données sont produites de manière stratégique. Il faut d'abord noter que notre travail s'appuie sur la théorie des jeux qui nécessite la définition de gains pour les agents participants. L'instanciation de ces gains dans des contextes réels pour correspondre à la perception des utilisateurs est un défi car ils peuvent englober des considérations subjectives. Ainsi, une première ligne de travail pourrait consister à combler cet écart entre les gains théoriques et la perception réelle des utilisateurs en utilisant des outils tels que l'analyse conjointe.

Nous divisons le reste de cette section en suivant l'organisation de cette thèse, en commençant par les problèmes possibles liés à la classification et en terminant par les problèmes liés à la régression (linéaire).

Classification. Nous avons étudié le problème de la classification antagoniste où un défenseur classe un comportement entre malveillant et non-malveillant et où un comportement est malveillant s'il provient d'un attaquant. Dans ce contexte, il n'y a pas de problème d'apprentissage préexistant en l'absence d'attaquants. De tels problèmes apparaissent dans les attaques de fraude ou de réseau mais ne couvrent pas tout le champ de l'apprentissage automatique. En particulier, l'important problème de la reconnaissance d'images largement étudié dans la littérature des exemples adverses (voir Goodfellow et al. [GSS15] par exemple) ne correspond pas bien à notre modèle. En effet, dans de telles applications, il existe un problème d'apprentissage préexistant défini théoriquement par une distribution de données inconnue qui peut être modifiée dans une certaine mesure par un attaquant. Nous appelons ces modèles *adversarial machine learning*. Des travaux préliminaires ont été réalisés par Meunier et al. [Meu+21] qui montrent l'existence d'un équilibre de Nash et Bhagoji et al. [BCM19] qui exposent une limite sur la robustesse possible des classificateurs. Cependant, les premiers ne caractérisent pas l'équilibre et les seconds ne considèrent pas un jeu et considèrent que l'attaquant ne paie aucun coût pour modifier les points de données. Il manque donc des résultats sur la robustesse de la classification dans un cadre de théorie des jeux où la modification des points de données entraîne un coût pour l'attaquant. De plus, dans le cadre de l'apprentissage automatique contradictoire, on peut considérer des paramètres de classification multi-classes (par exemple, classer la photo d'un animal entre différentes espèces possibles) qui, à notre connaissance, n'ont pas été étudiés. Cela

peut conduire à des résultats de difficulté inhérente au problème lorsqu'on augmente le nombre de classes possibles dans l'esprit des résultats de difficulté pour calculer des attaques robustes de Perdomo and Singer [PS19a]. En combinant les deux points précédents, une piste de travail prometteuse serait de considérer un problème où il existe une distribution sous-jacente correspondant à un problème de classification multi-classes. Un attaquant peut modifier cette distribution dans une certaine mesure et un défenseur choisit simultanément un classificateur (ceci correspond à un jeu où la notion d'équilibre considérée est l'équilibre de Nash mais un cadre de Stackelberg où le défenseur est le leader pourrait être intéressant. Les configurations Stackelberg où l'attaquant est le leader sont moins intéressantes car la défense optimale est simplement un classificateur Bayes naïf). L'objectif serait d'obtenir des limites sur le taux d'erreur minimum qu'un défenseur peut atteindre en fonction de la similarité entre les différentes classes ainsi que de produire des défenses qui peuvent atteindre des performances proches de ces limites en supposant l'accès à un ensemble d'entraînement généré avec la distribution originale. Enfin, bien que nous ayons considéré la classification antagoniste, il existe une littérature sur la classification stratégique où les producteurs de données ont des objectifs qui ne s'alignent pas mais ne s'opposent pas nécessairement au classificateur. Cette littérature est toutefois moins liée à notre domaine de travail sur la classification.

Régression linéaire. Nous avons étudié le modèle de régression linéaire où les agents choisissent stratégiquement la précision des points de données qu'ils produisent pour minimiser un mélange de leur coût de fourniture individuelle et d'un coût lié à la précision du résultat de la régression. Dans cette ligne de travail, plusieurs directions différentes sont possibles. Premièrement, nous avons supposé que les agents ont déclaré leur précision de manière véridique, ce qui peut ne pas être le cas. Ainsi, il serait bénéfique d'étudier des modèles et des mécanismes encourageant la révélation véridique des données. Ensuite, nous avons étudié un modèle de bien public dans lequel les agents sont incités à obtenir un estimateur précis, tandis que d'autres modèles étudient des situations dans lesquelles les agents sont incités à obtenir un résultat spécifique de l'estimation. Cependant, à notre connaissance, il n'existe pas dans la littérature de situations antagonistes où les agents peuvent manipuler leurs points de données dans une certaine mesure et visent à maximiser l'erreur d'estimation. Cela pose des problèmes de modélisation, car l'erreur d'estimation dépend à la fois du biais et de la variance de l'estimation, qui peuvent être difficiles à évaluer.

Enfin, nous avons étudié un sous-ensemble de la régression paramétrique avec la régression linéaire. Nous pensons cependant que certains de nos résultats (en

particulier la dégradation asymptotique de la précision de l'estimateur) peuvent être étendus dans des cadres non-paramétriques en utilisant des bornes sur la qualité non-paramétrique de l'estimation telles que présentées dans Györfi et al. [Gyö+02]. En particulier, les algorithmes tels que k-plus proches voisins ont des garanties d'erreur sous des hypothèses légères sur la distribution des données et sur la fonction à estimer. Lorsque les données sont produites de manière stratégique, nous pourrions observer une dégradation de ces garanties et, plus précisément, cette dégradation pourrait varier localement si certains points de données sont intrinsèquement plus coûteux à produire ou plus sensibles que d'autres. Un tel comportement n'est pas observé dans notre modèle car la régression linéaire est un modèle paramétrique où chaque point de données contient des informations permettant d'estimer les paramètres globaux du problème alors que le k-plus proche voisin est un modèle non-paramétrique où les points de données contiennent des informations uniquement sur le comportement local de la fonction à estimer.

Abstract

In this thesis, we consider the problem of learning when data are strategically produced. This challenges the widely used assumptions in machine learning that test data are independent from training data which has been proved to fail in many applications where the result of the learning problem has a strategic interest to some agents. We study the two ubiquitous problems of *classification* and *linear regression* and focus on fundamental learning properties on these problems when compared to the classical setting where data are not strategically produced.

We first consider the problem of finding optimal classifiers in an adversarial setting where the class-1 data is generated by an attacker whose objective is not known to the defender—an aspect that is key to realistic applications but has so far been overlooked in the literature. To model this situation, we propose a Bayesian game framework where the defender chooses a classifier with no *a priori* restriction on the set of possible classifiers. The key difficulty in the proposed framework is that the set of possible classifiers is exponential in the set of possible data, which is itself exponential in the number of features used for classification. To counter this, we first show that Bayesian Nash equilibria can be characterized completely via functional threshold classifiers with a small number of parameters. We then show that this low-dimensional characterization enables us to develop a training method to compute provably approximately optimal classifiers in a scalable manner; and to develop a learning algorithm for the online setting with low regret (both independent of the dimension of the set of possible data). We illustrate our results through simulations and apply our training algorithm to a real bank fraud data set in a simple setting.

We then consider the problem of linear regression from strategic data sources. In the classical setting where the precision of each data point is fixed, the famous Aitken/Gauss-Markov theorem in statistics states that generalized least squares (GLS) is a so-called "Best Linear Unbiased Estimator" (BLUE) and is consistent (the model is perfectly learned when the amount of data grows). In modern data science, however, one often faces strategic data sources, namely, individuals who incur a cost for providing high-precision data. We model this as learning 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. 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 show that, in general, Aitken's theorem does not hold under strategic data sources, though it does hold if individuals have identical provision costs (up to a multiplicative factor). When individuals have non-identical costs, we derive a bound on the improvement of the equilibrium estimation cost that can be achieved by deviating from GLS, under mild assumptions on the provision cost functions and on the possible deviations from GLS. We also 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.