Truthful Linear Regression

Rachel Cummings Stratis Ioannidis Katrina Ligett

April 24, 2015

Abstract

We consider the problem of fitting a linear model to data held by individuals who are concerned about their privacy. Incentivizing most players to report their data to the analyst truthfully constrains our design to mechanisms that provide a privacy guarantee to the participants; we use differentially privacy to model individuals' privacy losses. This immediately poses a problem, as differentially private computation of a linear model necessarily produces a biased estimation, and existing approaches to design mechanisms to elicit data from privacy-sensitive individuals do not generalize well to biased estimators. We manage to overcome this this challenge, through appropriate design of the computation and payment scheme.

1 Introduction

Fitting a linear model is perhaps the most fundamental and basic learning task, with diverse applications from statistics to experimental sciences like medicine and sociology. In many settings, the data from which a model is to be learnt are not held by the analyst performing the regression task, but must be elicited from individuals. Such settings clearly include medical trials and census surveys, as well as mining online behavioral data, a practice currently happening at a massive scale.

If data are held by self-interested individuals, it is not enough to simply run a regression—the data holders may wish to influence the outcome of the computation, either because they could benefit directly from certain outcomes, or because they wish to mask their input due to privacy concerns. In this case, it becomes necessary to model the utility functions of the individuals, and to design mechanisms that provide proper incentives. Ideally, such mechanisms should still allow for accurate computation of the underlying regression. A tradeoff then emerges, between the accuracy of the computation, and the budget required to compensate participants.

In this paper, we focus on the problem posed by data holders who are concerned with their privacy. Our approach can easily be generalized to handle individuals who wish to manipulate the computation's outcome for other reasons, but, for clarity, we treat only privacy concerns. We consider a population of players, each holding private data, and an analyst who wishes to compute a linear model from their data. The analyst must design a mechanism (a computation he will do, and payments he will give the players) that incentivizes the players to provide information that will allow for accurate computation, while minimizing the payments the analyst must make.

We use a model of players' costs for privacy [Chen et al., 2013] that is based on the well-established notion of differential privacy [Dwork et al., 2006]. Incentivizing most players to report their data to the analyst truthfully constrains our design to mechanisms that are differentially private. This immediately poses a problem, as differentially private computation of a linear model necessarily produces a biased estimation; existing approaches [Ghosh et al., 2014] to design mechanisms to elicit data from privacy-sensitive individuals do not generalize well to biased estimators. Overcoming this challenge, through appropriate design of the computation and payment scheme, is the main technical contribution of the present work.

1.1 Our results

We study the above issues in the context of linear regression. We present a mechanism (Algorithm 1), which, under appropriate choice of parameters and fairly mild technical assumptions, satisfies the following properties: it is (a) *accurate* (Theorem 4), i.e., computes an estimator whose squared L_2 distance to the true linear model goes to zero as the number of individuals increases, (b) *asymptotically truthful* (Theorem 3), in that agents have no incentive to misreport their data, (c) it *incentivizes participation* (Theorem 5), as players receive positive utility, and (d) it requires an *asymptotically small budget* (Theorem 6), as total payments to agents go to zero as the number of individuals increases. Our assumptions are on how individuals experience privacy losses and on the distribution from which these losses are drawn. Accuracy of the computation is attained by establishing that the algorithm provides differential privacy (Theorem 2), and that it provides payments such that the vast majority of individuals are incentivized to participate and to report truthfully (Theorems 3 and 5). An informal statement appears in Theorem 1.

The fact that we find that our total budget can be made to decrease in the number of individuals in the population is an effect of the approach we use to eliciting truthful participation, which is based on the peer prediction technology (Section A.1), and of the model of agents' costs for privacy (Section 2.3). A similar effect was seen by Ghosh et al. [2014]. As they note, costs would no longer tend to zero if our model incorporated some fixed cost for interacting with each individual.

1.2 Related Work

Following Ghosh and Roth [2013], a series of papers have studied data acquisition problems from agents that have privacy concerns. The vast majority of this work [Fleischer and Lyu, 2012, Ligett and Roth, 2012, Nissim et al., 2014] operates in a model where agents cannot lie about their private information (their only recourse is to withhold it or perhaps to lie about their costs for privacy). A related thread of work [Ghosh and Roth, 2013, Nissim et al., 2012, Chen et al., 2013] explores models of costs for privacy, based on the notion of differential privacy [Dwork et al., 2006].

Our setting is closest to, and inspired by, Ghosh et al. [2014], who bring the technology of peer prediction to bear on the problem of incentivizing truthful reporting in the presence of privacy concerns. The peer prediction approach, of Miller et al. [2005], incentivizes truthful reporting (in the absence of privacy constraints) by, effectively, rewarding players for reporting information that is predictive of the reports of other agents. This allows the analyst to leverage correlations between players' information. Ghosh et al. [2014] adapt the peer prediction approach to overcome a number of challenges presented by privacy-sensitive individuals. The mechanism and analysis of Ghosh et al. [2014] was for the simplest possible statistic—the sum of a private binary type. In contrast, we regress a linear model over player data, a significantly more sophisticated learning task. In particular, to attain accurate, privacy-preserving linear regression, we are forced to contend with biased private estimators, which interferes with our ability to incentivize truth-telling, and hence to compute an accurate statistic.

Linear regression under strategic agents has been studied in a variety of different contexts. Dekel et al. [2010] consider an analyst that regresses a "consensus" model across data coming from multiple strategic agents; agents would like the consensus value to minimize a loss over their own data, and they show that, in this setting, empirical risk minimization is group strategy-proof. A similar result, albeit in a more restricted setting, is established by Perote and Perote-Pena [2004]. Regressing a linear model over data from strategic agents that can only manipulate their costs, but not their data, was studied by Horel et al. [2014] and Cai et al. [2014], while Ioannidis and Loiseau [2013] consider a setting without payments, in which agents receive a utility as a function of estimation accuracy. We depart from the above approaches by considering agents whose utilities depend on their loss of *privacy*, an aspect absent from the above works.

2 Model and Preliminaries

2.1 A Regression Setting

We consider a population where each player $i \in [n] \equiv \{1, \dots, n\}$ is associated with a vector $x_i \in \mathbb{R}^d$ (i.e., player i's *features*) and a variable $y_i \in \mathbb{R}$ (i.e., her *response* variable). We assume that responses are linearly related to the features; that is, there exists a $\theta \in \mathbb{R}^d$ such that,

$$y_i = \theta^\top x_i + z_i, \quad \text{for all } i \in [n], \tag{1}$$

where z_i are zero-mean noise variables.

An analyst wishes to infer a linear model from the players' data; that is, he wishes to estimate θ, e.g., by performing linear regression on the players' data. However, players incur a privacy cost from revelation of their data, and need to be properly incentivized to truthfully reveal it to the analyst. More specifically, as in Ioannidis and Loiseau [2013], we assume that a player i can manipulate her responses y_i *but not* her features x_i. This is indeed the case when features are measured directly by the analyst (e.g., are observed during a physical examination, or are measured in a lab) or are verifiable (e.g., features are extracted from a player's medical record, or are listed on her ID). A player may misreport her response y_i, on the other hand, which is unverifiable; this would be the case if, e.g., y_i is the answer the player gives to a survey question pertaining to her preferences or habits.

We assume that players are strategic, and may lie either to increase the payment they extract from the analyst, or to mitigate any privacy violation they incur by the disclosure of their data. To address such strategic behavior, the analyst will design a mechanism $\mathcal{M} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d \times \mathbb{R}^n_+$, that takes as input all player data, namely, the features x_i and possibly perturbed responses $\hat{y}_i$, and outputs an estimate $\hat{\theta}$ as well as a set of non-negative payments $\{\pi_i\}_{i\in[n]}$ to each player. Informally, we seek mechanisms that allow for *accurate* estimation of θ while requiring only asymptotically *small budget*. In order to ensure accurate estimation of θ, we will require that our mechanism *incentivize truthful participation* on the part of most players, which in turn will require that we provide an appropriate *privacy guarantee*. We discuss privacy in more detail in Section 2.2. Clearly, all of the above also depend on the players' rational behavior and, in particular, their utilities; we formally present our model of player utilities in Section 2.3.

Throughout our analysis, we assume that θ is drawn independently from a known distribution $\mathcal{F}$, the attribute vectors x_i are drawn independently from the uniform distribution on the d-dimensional unit ball, and the noise terms z_i are drawn independently from a known distribution $\mathcal{G}$. Thus θ, $\{x_i\}_{i\in[n]}$, and $\{z_i\}_{i\in[n]}$ are independent random variables, while responses $\{y_i\}_{i\in[n]}$ are determined by (1). Note that, as a result, responses are conditionally independent given θ.

We require some additional bounded support assumptions on these distributions. In short, these boundedness assumptions are needed to ensure the sensitivity of mechanism $\mathcal{M}$ is finite; however, it is also natural in practice that both features and responses take values in a bounded domain. More precisely, we assume that the distribution $\mathcal{F}$ has bounded support, such that $\|\theta\|_2^2 \leq B$ for some constant B; we also require the noise distribution $\mathcal{G}$ to have mean zero, finite variance σ^2, and bounded support: $\texttt{supp}(\mathcal{G}) = [-M, M]$ for some constant M. These assumptions together imply that $\left|\theta^\top x_i\right| \leq B$ and $|y_i| \leq B + M$.

2.2 Differential Privacy

Recall the classic definition of differential privacy by Dwork et al. [2006]:

Definition 1 (Differential Privacy [Dwork et al., 2006])**.** *A mechanism $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ is* ϵ-differentially private *if for every pair of databases $D, D' \in \mathcal{D}^n$ differing only in one element, and for every subset of possible outputs $\mathcal{S} \subseteq \mathcal{R}$,*

$$\Pr[\mathcal{M}(D) \in \mathcal{S}] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in \mathcal{S}].$$

Intuitively, a mechanism outputting the result of a computation over a database is differentially private if the probability mass it places on any outcome changes by no more than a $e^\epsilon \approx 1 + \epsilon$ multiplicative factor, if a single entry in the database changes. The parameter ϵ quantifies the privacy guarantee provided by the mechanism to individuals whose data is in the database: $\epsilon = 0$ provides perfect privacy, as the output becomes independent of the input.

Following, e.g., Kearns et al. [2014] and Ghosh et al. [2014], we depart from the classic differential privacy definition, quantifying privacy violation instead through *joint differential privacy* [Kearns et al., 2014]. Intuitively, full differential privacy requires, in our setting, that all output by the mechanism $\mathcal{M}$, including the payment it allocates to a player, is insensitive to every player's input. In settings like ours, however, it makes sense to assume that the payment to a player is also in some sense "private", in that it is shared neither publicly nor with other players. To that end, we assume that the estimate $\hat{\theta}$ computed by the mechanism $\mathcal{M}$ is a publicly observable output; in contrast, each payment π_i is observable *only by player* i. Hence, from the perspective of each player i, the mechanism output that is publicly released and that, in turn, might violate her privacy, is $(\hat{\theta}, \pi_{-i})$, where π_{-i} comprises all payments excluding player i's payment.

Definition 2 (Joint Differential Privacy [Kearns et al., 2014])**.** *Consider a mechanism $\mathcal{M} : \mathcal{D}^n \to \mathcal{O} \times \mathcal{R}^n$, for $\mathcal{D}, \mathcal{O}, \mathcal{R}$ arbitrary sets. For each $i \in [n]$, let $(\mathcal{M}(\cdot))_{-i} = (o, \pi_{-i}) \in \mathcal{O} \times \mathcal{R}^{n-1}$ denote the portion of the mechanism's output that is observable to outside observers and players $j \neq i$. A mechanism $\mathcal{M}$ is ϵ-jointly differentially private if, for every player i, every database $D \in \mathcal{D}^n$, every $d'_i \in \mathcal{D}$, and for every observable set of outcomes $\mathcal{S} \subseteq \mathcal{O} \times \mathcal{R}^{n-1}$:*

$$\Pr\left[(\mathcal{M}(D))_{-i} \in \mathcal{S}\right] \leq \exp(\epsilon) \Pr\left[(\mathcal{M}(d'_i, D_{-i}))_{-i} \in \mathcal{S}\right].$$

This relaxation of differential privacy is natural, but it is also necessary to incentivize truthfulness [Ghosh and Roth, 2013]. Requiring that a player's payment π_i be ϵ-differentially private implies that a player's unilateral deviation changes the distribution of her payment only slightly. Hence, under full differential privacy, a player's payment would remain roughly the same no matter what she reports, which intuitively cannot incentivize truthful reporting.

2.3 Player Utilities

As discussed in the related work section, starting from Ghosh and Roth [2013], a series of recent papers on strategic data revelation model player privacy costs as functions of the privacy parameter ϵ. We also adopt this modeling assumption. Having introduced the notion of joint differential privacy, we now present our model of player utilities. We assume that every player is characterized by a cost parameter $c_i \in \mathbb{R}_+$, determining her sensitivity to the privacy violation incurred by the revelation of her data to her analyst. In particular, each player has a privacy cost function $f_i(c_i, \epsilon)$ that describes the cost she incurs when her data is used in a ϵ-jointly differentially private computation. Players have quasilinear utilities, so if player i receives payment π_i for her report, and experiences cost $f(c_i, \epsilon)$ from her privacy loss, her utility is

$$u_i = \pi_i - f_i(c_i, \epsilon).$$

Following again recent work, we assume that f can be an arbitrary function, bounded by an increasing monomial of ϵ. In particular, we make the following assumption.

Assumption 1. *The privacy cost function of each player satisfies*

$$f_i(c_i, \epsilon) \leq c_i \epsilon^2.$$

The monotonicity in ϵ is intuitive, as smaller values imply stronger privacy properties, with $\epsilon = 0$ indicating the output is independent of player i's data. We note that the quadratic bound in Assumption 1 was introduced by Chen et al. [2013] and also adopted by Ghosh et al. [2014]; as noted by the above authors, the quadratic bound can be shown to hold for a broad class of natural cost functions f_i; we refer the reader to Appendix D for a formal description of this class.

We stress here that the notion of ϵ-joint differential privacy (and, thus, of player costs incurred due to privacy violation) depends on both y_i *and* x_i: in this sense, though a player can only manipulate y_i, both her response *and* her features are treated as "private" variables in our model, and both disclosures incur a privacy cost. Features should certainly be deemed private if, e.g., they are attributes in a player's medical record, or outcomes of a medical examination. Moreover, (1) implies a correlation between features and the response, which can be strong, for example, in the case where θ has small support; it is therefore reasonable to assume that, if the response is private, so should features correlated to this response.

Throughout our analysis, we assume that the privacy cost parameters are also random variables, sampled from a distribution $\mathcal{C}$. We allow c_i to depend on player i's data (x_i, y_i); however, we assume conditioned on (x_i, y_i), that c_i does not reveal any additional information about the costs or data of any other agents. Formally:

Assumption 2. *Given (x_i, y_i), (x_{-i}, y_{-i}, c_{-i}) is conditionally independent of c_i, i.e.,*

$$\Pr[(x_{-i}, y_{-i}, c_{-i})|(x_i, y_i), c_i] = \Pr[(x_{-i}, y_{-i}, c_i)|(x_i, y_i), c_i'] \text{ for all } (x_{-i}, y_{-i}, c_{-i}),\ (x_i, y_i),\ c_i,\ c_i'.$$

We also make the following additional technical assumption on the tail of $\mathcal{C}$.

Assumption 3. *The conditional marginal distribution satisfies for some constant $p > 1$,*

$$\min_{x_i, y_i} \left(\Pr_{c_j \sim \mathcal{C}|x_i, y_i} [c_j \leq \tau] \right) \geq 1 - \tau^{-p}.$$

Note, that Assumption 3 implies that $\Pr_{c_i \sim \mathcal{C}}[c_i \leq \tau] \geq 1 - \tau^{-p}$.

2.4 Mechanism Properties

We seek mechanisms that satisfy the following properties: (a) truthful reporting is an equilibrium, (b) the estimator computed under truthful reporting is highly accurate, (c) players are ensured non-negative utilities from truthful reporting, and (d) the budget required from the analyst to run the mechanism is small.

We present here standard definitions used in this paper. For the following definitions, consider a fixed regression mechanism $\mathcal{M}$. Let $\pi_i(x, y)$ and be the payment to player i and let $c_i(x, y)$ be the cost experienced by player i when (x, y) is the collection of reports to the regression mechanism by all players. For the purposes of these definitions, we will assume that in the non-private setting presented in Section 3, all privacy costs are zero.

We define a strategy profile $\sigma = (\sigma_1, \ldots, \sigma_n)$ to be a collection of strategies σ_i (one for each player), mapping from realized data (x_i, y_i) to reports $\hat{y}_i$. Under strategy σ_i, a player who has data (x_i, y_i) would report $\hat{y}_i = \sigma_i(x_i, y_i)$ to the regression mechanism.

Definition 3 (Bayes Nash equilibrium)**.** *A strategy profile σ forms an η-*approximate Bayes Nash equilibrium *if for every player i, for all realizable (x_i, y_i), and for every misreport $\hat{y}_i \neq y_i$,*

$$\mathbb{E}[\pi_i(x, \sigma(x, y)) - c_i(x, \sigma(x, y))] \geq \mathbb{E}[\pi_i(x, (\hat{y}_i, \sigma_{-i}(x_{-i}, y_{-i}))) - c_i(x, (\hat{y}_i, \sigma_{-i}(x_{-i}, y_{-i})))] - \eta.$$

Definition 4 (Accuracy)**.** *A regression is η-*accurate *if for all realizable parameters θ, it outputs an estimate $\hat{\theta}$ such that,*

$$\mathbb{E}[\|\hat{\theta} - \theta\|_2^2] \leq \eta.$$

Definition 5 (Individually Rational)**.** *A mechanism is* individually rational *(IR) if for every player i, and for all realizable (x, y),*

$$\mathbb{E}[\pi_i(x, y) - c_i(x, y)] \geq 0.$$

We will also be concerned with the total amount spent by the analyst in the mechanism. The *budget $\mathcal{B}$* of a mechanism is the sum of all payments made to players. That is, $\mathcal{B} = \sum_i \pi_i$.

Definition 6 (Asymptotically small budget)**.** *An* asymptotically small budget *has for all realizable (x, y),*

$$\mathcal{B} = \sum_{i=1}^{n} \pi_i(x, y) = o(1).$$

Algorithm 1 Truthful Regression Mechanism

```
Solicit reports $X \in (\mathbb{R}^d)^n$ and $\hat{y} \in \mathbb{R}^n$
Analyst computes $\hat{\theta}^L = (X^\top X)^{-1} X^\top \hat{y}$ and $\hat{\theta}^L_{-i} = (X^\top_{-i} X_{-i})^{-1} X^\top_{-i} \hat{y}_{-i}$ for each $i \in [n]$
Output estimator $\hat{\theta}^L$
Pay each player $i$, $\pi_i = B_{a,b}(x_i^\top \hat{\theta}^L_{-i}, x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i])$
```

2.5 Additional Background and Technical Preliminaries

For completeness, we provide a brief review of peer prediction, linear regression, and differential privacy in Appendix A.

3 Truthful Regression without Privacy Constraints

To illustrate the ideas we use in the rest of the paper, we present in this section a mechanism which incentivizes truthful reporting in the absence of privacy concerns. If the players do not have privacy concerns (i.e., $c_i = 0$ for all $i \in [n]$), the analyst can simply collect data, estimate θ using linear regression, and compensate players using a re-scaled version of the following scoring rule (c.f. Appendix A.1):

$$B_{a,b}(p, q) = a - b\left(p - 2pq + q^2\right).$$

The mechanism is formally presented in Algorithm 1. Intuitively, in the spirit of peer prediction, a player's payment depends on how well her reported $\hat{y}_i$ agrees with the predicted value of y_i, as constructed by the estimate $\hat{\theta}^L_{-i}$ of θ produced by all her peers. We now show that truthful reporting is a Bayes Nash equilibrium.

Lemma 1 (Truthfulness)**.** *For all $a, b > 0$, truthful reporting is a Bayes Nash equilibrium under Algorithm 1.*

Proof. Recall that, conditioned on x_i, y_i, the distribution x_{-i}, y_{-i} is independent of c_i. Hence, assuming all other players are truthful, player i's expected payment conditioned on her data (x_i, y_i) and her cost c_i, under her (deterministic) response $\hat{y}_i$ is

$$\mathbb{E}[\pi_i | x_i, y_i, c_i] = \mathbb{E}\left[B_{a,b}(x_i^\top \hat{\theta}^L_{-i}, x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i]) | x_i, y_i\right] = B_{a,b}\left(x_i^\top \mathbb{E}[\hat{\theta}^L_{-i} | x_i, y_i], x_i^\top \mathbb{E}[\theta | x_i, \hat{y}_i]\right),$$

by the linearity of $B_{a,b}$ in its first arguement, as well as the linearity of the inner product. Note that $B_{a,b}$ is uniquely maximized by reporting $\hat{y}_i$ such that $\mathbb{E}[\theta | x_i, \hat{y}_i]^\top x_i = \mathbb{E}[\hat{\theta}^L_{-i} | x_i, y_i]^\top x_i$. Since $\hat{\theta}^L$ is an unbiased estimator of θ, then $\mathbb{E}[\hat{\theta}^L_{-i} | x_i, y_i] = \mathbb{E}[\theta | x_i, y_i]$. Thus the optimal misreport is $\hat{y}_i$ such that $\mathbb{E}[\theta | x_i, \hat{y}_i]^\top x_i = \mathbb{E}[\theta | x_i, y_i]^\top x_i$, so truthful reporting is a Bayes Nash equilibrium. □

We note that truthfulness is essentially a consequence of (a) the fact that $B_{a,b}$ is a strictly proper scoring rule (as it is positive-affine in its first argument and strictly concave in its second argument), and (b), most importantly, the fact that $\hat{\theta}^L_{-i}$ is an unbiased estimator of θ. Moreover, as in the case of the simple peer prediction setting presented in Appendix A.1, truthfulness persists even if $\hat{\theta}^L_{-i}$ in Algorithm 1 is replaced by a linear regression estimator constructed over responses restricted to an arbitrary set $S \subseteq [n] \setminus i$.

Truthful reports enable accurate computation of the estimator.

Lemma 2 (Accuracy)**.** *Under truthful reporting, with probability at least $1 - d^{-t^2}$ and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the accuracy the estimator $\hat{\theta}^L$ in Algorithm 1 is $\mathbb{E}\left[\left\|\hat{\theta}^L - \theta\right\|_2^2\right] \leq \frac{\sigma^2}{(1-\xi)\frac{1}{d+2}n}$.*

Proof. Note that $\mathbb{E}\left[\left\|\hat{\theta}^L - \theta\right\|_2^2\right] = \text{trace}(\texttt{Cov}(\hat{\theta}^L)) \overset{(5)}{=} \sigma^2 \,\text{trace}\left((X^\top X)^{-1}\right)$. For i.i.d. features x_i, the spectrum of matrix $X^\top X$ can be asymptotically characterized by a theorem by Vershynin [2012] (c.f. Theorem 7 in Appendix A.2), and the lemma follows. □

Remark Note that individual rationality and a small budget can be trivially attained in the absence of privacy costs. To ensure individual rationality of Algorithm 1, payments π_i must be non-negative, but can be arbitrarily small. Thus payments can be scaled down to reduce the analyst's total budget. For example, setting $a = b(B + 2B(B + M) + (B + M)^2 - 1)$ and $b = \frac{1}{n^2}$ ensures $\pi_i \geq 0$ for all players i, and the total required budget is $\frac{1}{n}(2B + 4B(B + M) + (B + M)^2) = O(\frac{1}{n})$.

4 Truthful Regression with Privacy Constraints

As we saw in the previous section, in the absence of privacy concerns, it is possible to devise payments that incentivize truthful reporting. These payments compensate players based on how well their report agrees with a response predicted through a $\hat{\theta}^L$ estimated by other player's reports.

Players whose utilities depend on privacy raise several challenges. Recall that the parameters estimated by the analyst, and the payments made to players, need to satisfy joint differential privacy, and hence any estimate of θ revealed publicly by the analyst or used in a payment must be ϵ-differentially private. Unfortunately, the sensitivity of the linear regression estimator $\hat{\theta}^L$ to changes in the input data is, in general, unbounded; this is precisely because matrix $X^\top X$ may not be invertible. As a result, it is not possible to construct a non-trivial differentially private version of $\hat{\theta}^L$ by, e.g., adding noise to its output.

In contrast, differentially private versions of regularized estimators like the ridge regression estimator $\hat{\theta}^R$ can be constructed; indeed, recent techniques have been developed for precisely this purpose, not only for ridge regression but for the broader class of learning through (convex) empirical risk minimization [Chaudhuri et al., 2011, Bassily et al., 2014]. In short, the techniques by Chaudhuri et al. [2011] and Bassily et al. [2014] succeed precisely because, for $\gamma > 0$, the regularized loss (3) is *strongly convex*. This implies that the sensitivity of $\hat{\theta}^R$ is bounded, and a differentially private version of $\hat{\theta}^R$ can be constructed by adding noise of appropriate variance (see also Lemma 6), or though alternative techniques, like objective perturbation.

The above suggest that a possible approach to constructing a truthful, accurate mechanism in the presence of privacy-conscious players is to modify Algorithm 1 by replacing $\hat{\theta}^L$ with a ridge regression estimator $\hat{\theta}^R$, both with respect to the estimate released globally, as well as with respect to any estimates used in computing payments through the Brier scoring rule. Unfortunately, such an approach breaks truthfulness, because $\hat{\theta}^R$ is a *biased* estimator. The linear regression estimator $\hat{\theta}^L$ ensured that the Brier scoring rule $B_{a,b}$ was maximized precisely when players reported their response variable truthfully; however, in the presence of an expected bias $\mathtt{b}$, it can easily be seen that the optimal report of player i deviates from truthful reporting by a quantity proportional to $\mathtt{b}^T x_i$.

We address this issue for large n using again the concentration result by Vershynin [2012] (c.f. Appendix A.2). This ensures that, for large n, the spectrum of $X^\top X$ should grow roughly linearly with n, with high probability. By (5), this implies that as long as γ grows more slowly than n, the bias term of $\hat{\theta}^R$ converges to zero, with high probability. Together, these statements ensure that, for an appropriate choice of γ, we attain approximate truthfulness for large n, while also ensuring that the output of our mechanism remains differentially private for all n. We exploit this intuition in proving that our mechanism presented in Section 4.1, based on ridge regression, indeed attains approximate truthfulness for large n, while also remaining jointly differentially private.

4.1 Private Regression Mechanism

We present our mechanism for private and truthful regression in Algorithm 2, which is a privatized version of Algorithm 1. We incorporate into our mechanism the the Output Perturbation algorithm from Chaudhuri et al. [2011], which first computes the ridge regression estimator and then adds noise to the output. This approach is used to ensure that our estimator $\hat{\theta}$ satisfies differential privacy.

The noise vector v will be drawn according to the following distribution P_L, which is a high-dimensional

Laplace distribution with parameter $\frac{4B+2M}{\gamma\epsilon}$:

$$P_L(v) \propto \exp\left(\frac{-\gamma\epsilon}{4B+2M}\left\|v\right\|_2\right).$$

Algorithm 2 Private Regression Mechanism

```
Solicit reports $X \in \left(\mathbb{R}^d\right)^n$ and $\hat{y} \in \mathbb{R}^n$
Randomly partition players into two groups, with respective data pairs $(X_0, \hat{y}_0)$ and $(X_1, \hat{y}_1)$
Analyst computes $\hat{\theta}^R = (\gamma I + X^\top X)^{-1} X^\top \hat{y}$ and $\hat{\theta}^R_j = (\gamma I + X_j^\top X_j)^{-1} X_j^\top \hat{y}_j$ for $j = 0, 1$
Independently draw $v, v_0, v_1 \in \mathbb{R}^d$ according to distribution $P_L$
Compute estimators $\hat{\theta}^P = \hat{\theta}^R + v$, $\hat{\theta}^P_0 = \hat{\theta}^R_0 + v_1$, and $\hat{\theta}^P_1 = \hat{\theta}^R_1 + v_1$
Output estimator $\hat{\theta}^P$
Pay each player $i$ in group $j$, $\pi_i = B_{a,b}((\hat{\theta}^P_{1-j})^\top x_i, \mathbb{E}[\theta | x_i, \hat{y}_i]^\top x_i)$ for $j = 0, 1$
```

Here we state an informal version of our main result. The formal version of this result is stated in Corollary 1, which aggregates and instantiates Theorems 2, 3, 4, 5, and 6, all presented in Section 5.

Theorem 1 (Main result (Informal))**.** *Under Assumptions 1, 2, and 3, there exists ways to set ϵ, γ, a, and b in Algorithm 2 to ensure that with high probability:*

1. *the output of Algorithm 2 is $o(\frac{1}{\sqrt{n}})$-jointly differentially private,*
2. *it is an $o\left(\frac{1}{n}\right)$-approximate Bayes Nash equilibrium for a $(1 - o(1))$-fraction of players to truthfully report their data,*
3. *the computed estimator $\hat{\theta}^P$ is $o(1)$-accurate,*
4. *it is individually rational for a $(1 - o(1))$-fraction of players to participate in the mechanism, and*
5. *the required budget from the analyst is $o(1)$.*

5 Analysis of Algorithm 2

In this section, we flesh out the claims made in Theorem 1. Due to space constraints, the proofs are deferred to Appendix B.

Theorem 2 (Privacy)**.** *The mechanism in Algorithm 2 is 2ϵ-jointly differentially private.*

Proof idea We first show that the estimators $\hat{\theta}^P$, $\hat{\theta}^P_0$, $\hat{\theta}^P_1$ together satisfy 2ϵ-differential privacy, by bounding the maximum amount that any player's report can affect the estimators. We then use the Billboard Lemma (Lemma 5) to show that the estimators, together with the vector of payments, satisfy 2ϵ-joint differential privacy.

Once we have a privacy guarantee, we can build on this to get truthful participation and hence accuracy. To do so, we first show that a symmetric threshold strategy equilibrium exists, in which all agents with cost c_i below some threshold τ participate and truthfully report their y_i. We will define $\tau_{\alpha,\beta}$ to be the cost threshold such that (1) with probability $1 - \beta$ (with respect to the prior from which costs are drawn), at least a $1 - \alpha$ fraction of players have cost coefficient $c_i \leq \tau_{\alpha,\beta}$, and (2) conditioned on her own data, each player i believes that with probability $1 - \alpha$, any other player j will have cost coefficient $c_j \leq \tau_{\alpha,\beta}$.

Definition 7 (Threshold $\tau_{\alpha,\beta}$)**.** *Fix a marginal cost distribution $\mathcal{C}$ on $\{c_i\}$, and let*

$$\tau^1_{\alpha,\beta} = \inf_{\tau} \left(\Pr_{c\sim\mathcal{C}} [|\{i : c_i \leq \tau\}| \geq (1-\alpha)n] \geq 1-\beta \right),$$

$$\tau^2_{\alpha} = \inf_{\tau} \left(\min_{x_i,y_i} \left(\Pr_{c_j\sim\mathcal{C}|x_i,y_i} [c_j \leq \tau] \right) \geq 1-\alpha \right).$$

Define $\tau_{\alpha,\beta}$ to be the larger of these thresholds:

$$\tau_{\alpha,\beta} = \max\{\tau^1_{\alpha,\beta}, \tau^2_{\alpha}\}.$$

We also define the threshold strategy σ_τ, in which a player reports truthfully if her cost c_i is below τ, and is allowed to misreport arbitrarily if her cost is above τ.

Definition 8 (Threshold strategy)**.** *Define the threshold strategy σ_τ as follows:*

$$\sigma_\tau(x_i, y_i, c_i) = \begin{cases} \textit{Report } \hat{y}_i = y_i & \textit{if } c_i \leq \tau \\ \textit{Report arbitrary } \hat{y}_i & \textit{otherwise} \end{cases}$$

We show that $\sigma_{\tau_{\alpha,\beta}}$ forms a symmetric threshold strategy equilibrium in the Private Regression Mechanism of Algorithm 2.

Theorem 3 (Truthfulness)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, and a desired confidence parameter β. Then under Assumptions 1 and 2, with probability $1-d^{t^2}$ and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an η-approximate Bayes-Nash equilibrium in Algorithm 2, for*

$$\eta = b\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

Proof idea There are three primary sources of error which cause the estimator $\hat{\theta}^P$ to differ from a player's posterior on θ. First, ridge regression is a biased estimation technique; second, Algorithm 2 adds noise to preserve privacy; third, players with cost c_i above threshold $\tau_{\alpha,\beta}$ are allowed to misreport their data. We show how to control the effects of these three sources of error, so that $\hat{\theta}^P$ is "not too far" from a player's posterior on θ. Finally, we use strong convexity of the payment rule to show that any player's payment from misreporting is at most η greater than her payment from truthful reporting.

Theorem 4 (Accuracy)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, and a desired confidence parameter β. Then under the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$, Algorithm 2 will output an estimator $\hat{\theta}^P$ such that with probability at least $1-\beta-d^{-t^2}$, and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$,*

$$\mathbb{E}[\|\hat{\theta}^P - \theta\|_2^2] = O\left(\left(\frac{\alpha n}{\gamma} + \frac{1}{\gamma\epsilon}\right)^2 + \left(\frac{\gamma}{n}\right)^2 + \left(\frac{1}{n}\right)^2 + \frac{\alpha n}{\gamma} + \frac{1}{\gamma\epsilon}\right).$$

Proof idea As in the proof of Theorem 3, we control the three sources of error in the estimator $\hat{\theta}^P$—the bias of ridge regression, the noise added to preserve privacy, and the error due to a fraction of players misreporting their data—this time measuring distance with respect to the expected L_2 norm difference.

We next see that players whose costs are below the threshold are incentivized to participate.

Theorem 5 (Individual Rationality)**.** *Under Assumption 1, the mechanism in Algorithm 2 is individually rational for all players with cost coefficients $c_i \leq \tau_{\alpha,\beta}$ as long as,*

$$a \geq \left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2,$$

regardless of the reports from players with cost coefficients above $\tau_{\alpha,\beta}$.

Proof idea A player's utility from participating in the mechanism is her payment minus her privacy cost. The parameter a in the payment rule is a constant offset that shifts each player's payment. We lower bound the minimum payment from Algorithm 2 and upper bound the privacy cost of any player with cost coefficient below threshold $\tau_{\alpha,\beta}$. If a is larger than the difference between these two terms, then any player with cost coefficient below threshold will receive non-negative utility.

Finally, we analyze the total cost of running the mechanism.

Theorem 6 (Budget)**.** *The total budget required by the analyst to run Algorithm 2 under threshold equilibrium strategy $\sigma_{\tau_{\alpha,\beta}}$ is*

$$\mathcal{B} = n\left[a + \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB)\right].$$

Proof idea The analyst's budget is the sum of all payments made to players in the mechanism. We upper bound the maximum payment to any player, and the total budget required is at most n times this maximum payment.

5.1 Formal Statement of Main Result

In this section, we present our main result, Corollary 1, which instantiates Theorems 2, 3, 4, 5, and 6 with a setting of all parameters, to get the bounds promised in Theorem 1. Before stating our main result, we first require the following lemma which asymptotically bounds $\tau_{\alpha,\beta}$ for an arbitrary bounded distribution. We will use this to control the asymptotic behavior of $\tau_{\alpha,\beta}$ under Assumption 3.

Lemma 3. *For a cost distribution $\mathcal{C}$ with conditional marginal CDF lower bounded by some function F: $\min_{x_i,y_i}\left(\Pr_{c_j\sim\mathcal{C}|x_i,y_i}[c_j \le \tau]\right) \ge F(\tau)$, then*

$$\tau_{\alpha,\beta} \le \max\{F^{-1}(1-\alpha\beta), F^{-1}(1-\alpha)\}.$$

We note that under Assumption 3, Lemma 3 implies that

$$\tau_{\alpha,\beta} \le \max\{(\alpha\beta)^{-1/p}, (\alpha)^{-1/p}\}.$$

Using this fact, we can state a formal version of our main result.

Corollary 1 (Main result (Formal))**.** *Under Assumptions 1, 2, and 3, setting $\alpha = n^{\frac{-p}{2p-1}}$, $\beta = n^{-0.01}$, $\epsilon = n^{-\frac{1.02}{2}}(\alpha\beta)^{\frac{1}{2p}}$, $\gamma = n^{-\frac{1}{2}-\frac{1.01}{3}-\frac{0.01-1.02p}{3p(2p-1)}}(\alpha\beta)^{\frac{1}{2p}}$, $a = (\alpha\beta)^{-1/p}$, $b = \frac{1}{n}$, $\xi = 1/2$, and $t = \sqrt{\frac{n}{4C(d+2)\log d}}$ in Algorithm 2 ensures that with probability $1 - d^{\Theta\left(\frac{n}{d\log d}\right)} - n^{-.01}$:*

1. *the output of Algorithm 2 is $O\left(n^{-\frac{1.02}{2}-\frac{1}{2(2p-1)}-\frac{0.01}{2p}}\right)$-jointly differentially private,*

2. *it is an $O\left(n^{-1.01}\right)$-approximate Bayes Nash equilibrium for a $1 - O\left(n^{\frac{-p}{2p-1}}\right)$ fraction of players to truthfully report their data,*

3. *the computed estimate $\hat{\theta}^P$ is $O\left(n^{-\frac{1.04}{3}-\frac{0.01-1.02p}{3p(2p-1)}}\right)$-accurate,*

4. *it is individually rational for a $1 - O\left(n^{\frac{-p}{2p-1}}\right)$ fraction of players to participate in the mechanism, and*

5. *the required budget from the analyst is $O\left(n^{-0.01}\right)$.*

This corollary follows immediately from instantiating Theorems 2, 3, 4, 5, and 6 with the specified parameters.

Remark Note that different settings of parameters can be used, to yield a different trade-off between approximation factors in the above result. For example, if the analyst is willing to supply a higher budget (say constant or increasing with n), he could improve on the accuracy guarantee.

References

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization, revisited. *CoRR*, abs/1405.7085, 2014.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1950.

Yang Cai, Constantinos Daskalakis, and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. *arXiv preprint 1408.2539*, 2014.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, July 2011.

Yiling Chen, Stephen Chong, Ian A. Kash, Tal Moran, and Salil Vadhan. Truthful mechanisms for agents that value privacy. In *Proceedings of the 14th ACM Conference on Electronic Commerce*, EC '13, pages 215–232, 2013.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, 2006.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the IEEE 51st Annual Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, 2010.

Lisa K. Fleischer and Yu-Han Lyu. Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 568–585, New York, NY, USA, 2012. ACM.

Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior*, 2013. Preliminary Version appeared un the Proceedings of the Twelfth ACM Conference on Electronic Commerce (EC 2011).

Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. Buying private data without verification. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 931–948, 2014.

Thibaut Horel, Stratis Ioannidis, and S. Muthukrishnan. Budget feasible mechanisms for experimental design. In Alberto Pardo and Alfredo Viola, editors, *LATIN 2014: Theoretical Informatics*, Lecture Notes in Computer Science, pages 719–730. 2014.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 21–30, 2014.

Stratis Ioannidis and Patrick Loiseau. Linear regression as a non-cooperative game. In Yiling Chen and Nicole Immorlica, editors, *Web and Internet Economics*, Lecture Notes in Computer Science, pages 277–290. 2013.

Michael Kearns, Mallesh Pai, Aaron Roth, and Jonathan Ullman. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 403–410, 2014.

Donald Knuth. *Seminumerical algorithms*, volume 2, pages 130–131. Addison-Wesley Publishing Company, 2 edition, 1981.

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 Internet and Network Economics*, WINE'12, pages 378–391, 2012.

Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *In Proceeding SIGMOD Conference*, pages 19–30, 2009.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Manage. Sci.*, 51(9):1359–1373, Sept 2005.

Kobbi Nissim, Claudio Orlandi, and Rann Smorodinsky. Privacy-aware mechanism design. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 774–789, 2012.

Kobbi Nissim, Salil Vadhan, and David Xiao. Is privacy compatible with truthfulness? In *Proceedings of the 4th Innovations in Theoretical Computer Science*, ITCS '14, 2014. To appear.

Javier Perote and Juan Perote-Pena. Strategy-proof estimators for simple regression. In *Mathematical Social Sciences 47*, pages 153–176, 2004.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, theory and applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

A Basics of peer prediction, linear regression, and privacy

A.1 Peer Prediction and the Brier Scoring Rule

Peer prediction [Miller et al., 2005] is a useful method of inducing truthful reporting among players that hold data generated by the same statistical model. In short, each player reports her data to an analyst, and is paid based on how well her report predicts the report of other players; tying each player's payment to how closely it predicts peer reports is precisely is exactly what induces truthfulness. Ghosh et al. [2014] illustrate these ideas in the context of privacy-sensitive individuals through the use of the Brier scoring rule [Brier, 1950] as a payment scheme among players holding a random bit. As we make use of the same technique, we review here how the Brier scoring rule can be used for basic peer prediction. Consider a set of n players, each holding a binary variable $b_i \in \{0, 1\}$. Assume that each of these variables is generated by i.i.d. Bernouli trials with parameter p, i.e., $\Pr(b_i = 1) = p$, for every $i \in [n]$. We assume here that p is itself a random variable generated from a known prior over $[0, 1]$. Each player reports a bit $\tilde{b}_i \in \{0, 1\}$ to the analyst, who wishes to estimate p through, e.g. $\frac{1}{n}\sum_{i\in[n]} \tilde{b}_i$. The analyst therefore wishes to incentivize truthful reporting of the bits b_i, through an appropriate payment scheme.

Let $\mathbb{E}[p \mid b]$ be expected value of p conditioned on observing that a player's bit is $b \in \{0, 1\}$; put differently, for every player whose bit is b, $\mathbb{E}[p \mid b]$ captures her belief on what value p takes, after she observes her own bit. Consider the following payment rule: to generate the payment for player i, the analyst selects a player j u.a.r. from $[n] \setminus i$ and pays player i:

$$B(\tilde{b}_j, \mathbb{E}[p \mid \tilde{b}_i]), \text{ where } B(q', q) = 1 - 2(q' - 2q' \cdot q + q^2). \tag{2}$$

The payment function $B(q', q)$ is the basic Brier scoring rule [Brier, 1950]; by design, it is *strictly proper*, i.e., it is uniquely maximized by truthful reporting. For completeness, we provide a proof.

Lemma 4. *[Miller et al., 2005] Under payments* (2)*, truthful reporting is a Bayes-Nash equilibrium.*

Proof. Observe that, for $q, q' \in [0, 1]$, $B(q', q)$ is positive, so payments (2) are individually rational. Moreover, for all $q' \in [0, 1]$, $B(q', q)$ is a strictly concave function of q maximized at $q' = q$. Moreover, $B(q', q)$ is an affine function of q'; hence, if user i's bit is b_i, and all other players report their bits truthfully (i.e., $\tilde{b}_j = b_j$ for all $j \neq i$), player i's expected payment is: $\mathbb{E}\left[B(b_j, \mathbb{E}[p \mid \tilde{b}_i]) \mid b_i\right] = B\left(\mathbb{E}[b_j \mid b_i], \mathbb{E}[p \mid \tilde{b}_i]\right) = B\left(\mathbb{E}[p \mid b_i], \mathbb{E}[p \mid \tilde{b}_i]\right)$. Hence, player i's payment is maximized when $\tilde{b}_i = b_i$. □

Informally, the payment scheme (2) induces truthfulness by paying a player the highest if the belief induced on p by her reported bit "agrees" with the belief induced by the bit of an arbitrary peer. We note that, instead of the bit of a peer selected u.a.r., any quantity whose expectation conditioned on b_i would be equal to $\mathbb{E}[p \mid b_i]$ would work as input to the Brier rule. For example, using the average value $\bar{b}_S = \frac{1}{|S|}\sum_{j \in S} \tilde{b}_j$ for any $S \subseteq [n] \setminus i$ as the first argument of B would also induce truthful reporting.

A.2 Linear Regression

We recall in this section the basic properties of linear regression, the method the analyst can employ to estimate the parameter vector $\theta \in \mathbb{R}^d$. Let $X = [x_i]_{i \in [n]} \in \mathbb{R}^{n \times d}$ denote the $n \times d$ matrix of features, and $y = [y_i]_{i \in [n]} \in \mathbb{R}^n$ the vector of responses. Estimating θ through *ridge regression* amounts to minimizing the following regularized quadratic loss function:

$$\mathcal{L}(\theta; X, y) = \sum_{i=1}^{n} \ell(\theta; x_i, y_i) = \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 + \gamma \|\theta\|_2^2 . \tag{3}$$

That is, the ridge regression estimator can be written as:

$$\hat{\theta}^R = \underset{\theta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (x_i^\top \theta - y_i)^2 + \gamma \|\theta\|_2^2 = (\gamma I + X^\top X)^{-1} X^\top y.$$

The parameter $\gamma > 0$, known as the regularization parameter, ensures that the loss function is *strongly convex* (see Appendix E) and, in particular, that the minimizer of (3) is unique. When $\gamma = 0$, the estimator is the standard *linear regression* estimator, which we denote by $\hat{\theta}^L = (X^\top X)^{-1} X^\top y$. The linear regression estimator is unbiased, i.e., under (1), it satisfies $\mathbb{E}[\hat{\theta}^L] = \theta$, which is not true when $\gamma > 0$: the general ridge regression estimator $\hat{\theta}^R$ is *biased.*

Nonetheless, in practice, $\hat{\theta}^R$ is preferable to $\hat{\theta}^L$ as it can achieve a desirable trade-off between *bias* and *variance.* In particular, consider the square loss error of the estimation $\hat{\theta}^R$, namely, $\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2]$. If we condition on the true parameter vector θ and the features X, this can be written as

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2 \,|] = \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2^2] + \|\mathbb{E}[\hat{\theta}^R] - \theta\|_2^2 = \mathrm{trace}(\mathtt{Cov}(\hat{\theta}^R)) + \|\,\mathtt{bias}(\hat{\theta}^R)\|_2^2 \tag{4}$$

where $\mathtt{Cov}(\hat{\theta}^R) = \mathbb{E}[(\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R])(\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R])^\top]$, $\mathtt{bias}(\hat{\theta}^R) = \mathbb{E}[\hat{\theta}^R] - \theta$ are the covariance and bias, respectively, of estimator $\hat{\theta}^R$. Assuming that the responses y follow (1)[1] and, again, conditioned on X and θ, these can be computed in closed form as:

$$\mathtt{Cov}(\hat{\theta}^R) = \sigma^2 (\gamma I + X^\top X)^{-1} X^\top X (\gamma I + X^\top X)^{-1}, \qquad \mathtt{bias}(\hat{\theta}^R) = -\gamma(\gamma I + X^\top X)^{-1}\theta, \tag{5}$$

where σ^2 is the variance of the noise variables in (1). It is easy to see that decreasing γ decreases the bias, but may significantly increase the variance. For example in the case where $\mathrm{rank}(X) < d$, the matrix $X^\top X$ is not invertible, and the trace of the covariance tends to infinity as γ tends to zero.

[1] i.e., under truthful reporting.

Whether trace($\texttt{Cov}(\hat{\theta}^R)$) is large and, therefore, whether regularizing the square loss is necessary, depends on largest eigenvalue (i.e., the *spectral norm*) of $(X^\top X)^{-1}$. Though for arbitrary X this can be infinite, if the x_i's are i.i.d. we expect that as n increases we get estimates of lower variance. Indeed, by the law of large numbers, we expect that, if we sample the features x_i independently from an isotropic distribution $\frac{1}{n}(X^\top X)$ should converge to the covariance of this distribution (namely cI). As such, for large n both the largest and smallest eigenvalues of $X^\top X$ should be of the order of n, leading to an estimation of ever decreasing variance even when $\gamma = 0$. The following theorem, which follows as a corollary of a result by Vershynin [2012] (see Appendix C), formalizes this notion, providing bounds on both the largest and smallest eigenvalue of $X^\top X$ and $\gamma I + X^\top X$.

Theorem 7. *Let $\xi \in (0,1)$, and $t \geq 1$. Let $\|\cdot\|$ denote the spectral norm. If $\{x_i\}_{i\in[n]}$ are i.i.d. and sampled uniformly from the unit ball, then with probability at least $1 - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, for some absolute constant C, then,*

$$\left\|X^\top X\right\| \leq (1+\xi)\frac{1}{d+2}n, \text{ and } \left\|(X^\top X)^{-1}\right\| \leq \frac{1}{(1-\xi)\frac{1}{d+2}n}, \text{ and}$$

$$\left\|\gamma I + X^\top X\right\| \leq \gamma + (1+\xi)\frac{1}{d+2}n, \text{ and } \left\|(\gamma I + X^\top X)^{-1}\right\| \leq \frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}.$$

Remark A generalization of Theorem 7 holds for $\{x_i\}_{i\in[n]}$ sampled from any distribution with a covariance Σ whose smallest eigenvalue is bounded away from zero (see Vershynin [2012]). We restrict our attention to the unit ball for simplicity and concreteness.

A.3 The Billboard Lemma

A very useful result regarding jointly differentially private mechanisms that we use in our analysis is the so-called "billboard-lemma":

Lemma 5 (Billboard Lemma [Hsu et al., 2014])**.** *Let $\mathcal{M} : \mathcal{D}^n \to \mathcal{O}$ be an ϵ-differentially private mechanism. Consider a set of n functions $f_i : \mathcal{D} \times \mathcal{O} \to \mathcal{R}$, for $i \in [n]$. Then, the mechanism $\mathcal{M}' : \mathcal{D}^n \to \mathcal{O} \times \mathcal{R}^n$ that computes $r = \mathcal{M}(D)$ and outputs $\mathcal{M}'(D) = (r, f_1(\Pi_2 D, r), \ldots, f_n(\Pi_n D, r))$, where π_i is the projection to player i's data, is ϵ-jointly differentially private.*

In short, the billboard lemma implies that if we can construct payments that depend on the data of individual players, as well as a universally observable output that is ϵ-differentially private (e.g., $\hat{\theta}$), the resulting mechanism will be ϵ-jointly differentially private.

B Proofs from Section 5

B.1 Privacy

We will now prove that the output $\hat{\theta}^P$ of this mechanism is ϵ-differentially private and that the payments π satisfy 2ϵ-joint differential privacy. First, we need the following lemma to bound the *sensitivity* of $\hat{\theta}^P$, formally defined in Definition 9, which is the maximum change in the output when a single player misreports her data. For vector-valued outputs, we measure this change with respect to the L_2 norm.

Definition 9 (Sensitivity)**.** *The* sensitivity *of a function $f : \mathcal{D} \to \mathcal{R}$ is the maximum L_2 norm of the function's output, when a single player changes her input:*

$$\textit{Sensitivity of } f = \max_{D,D',\ \textit{neighbors}} \|f(D) - f(D')\|_2$$

The following lemma follows from Chaudhuri et al. [2011]; a proof is provided for completeness.

Lemma 6. *The sensitivity of $\hat{\theta}^R$ is $\frac{1}{\gamma}(4B+2M)$.*

Proof. Let (X, y) and (X', y') be two arbitrary neighboring databases, and let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ respectively denote the ridge regression estimators computed on (X, y) and (X', y'). Define $g(\theta)$ to be the change in loss when θ is used as an estimator for (X', y') and (X, y).

$$\begin{aligned} g(\theta) &= \mathcal{L}(\theta; X', y') - \mathcal{L}(\theta; X, y) \\ &= \left(\theta^\top x_i - y_i\right)^2 - \left(\theta^\top x_i' - y_i'\right)^2 \end{aligned}$$

Lemma 7 of Chaudhuri et al. [2011] says that if $\mathcal{L}(\theta; X, y)$ and $\mathcal{L}(\theta; X', y')$ are both Γ-strongly convex, then $\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2$ is bounded above by $\frac{1}{\Gamma} \cdot \max_\theta \|\nabla g(\theta)\|_2$. By Lemma 13, both $\mathcal{L}(\theta; X, y)$ and $\mathcal{L}(\theta; X', y')$ are 2γ-strongly convex, so $\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{1}{2\gamma} \cdot \max_\theta \|\nabla g(\theta)\|_2$. We now bound $\|\nabla g(\theta)\|_2$ for an arbitrary θ.

$$\begin{aligned} \|\nabla g(\theta)\|_2 &= 2\left\|(\theta^\top x_i - y_i)x_i - (\theta^\top x_i' - y_i')x_i'\right\|_2 \\ &\leq 4\left|\theta^\top x_i - y_i\right| \|x_i\|_2 \\ &\leq 4\left(\left|\theta^\top x_i\right| + |y_i|\right) \\ &\leq 4(2B + M) \end{aligned}$$

Since this bound holds for all θ, it must be the case that $\max_\theta \|\nabla g(\theta)\|_2 \leq 4(2B+M)$ as well. Then by Lemma 7 of Chaudhuri et al. [2011],

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{4}{2\gamma}(2B+M) = \frac{1}{\gamma}(4B+2M).$$

Since (X, y) and (X', y') were *any* two neighboring databases, this bounds the sensitivity of the computation, so changing the input of one player can change the ridge regression estimator (with respect to the L_2 norm) by at most $\frac{1}{\gamma}(4B+2M)$. □

We now prove that the output of Algorithm 2 satisfies 2ϵ-joint differential privacy.

Theorem 2 (Privacy)**.** *The mechanism in Algorithm 2 is 2ϵ-jointly differentially private.*

Proof. We begin by showing that the estimator $\hat{\theta}^P$ output by Algorithm 2 is differentially private.

Let h denote the PDF of $\hat{\theta}^P$ output by Algorithm 2, and ν denote the PDF of the noise vector v. Let (X, y) and (X', y') be any two databases that differ only in the i-th entry, and let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ respectively denote the ridge regression estimators computed on these two databases.

The output estimator $\hat{\theta}^P$ is the sum of the ridge regression estimator $\hat{\theta}^R$, and the noise vector v; the only randomness in the choice of $\hat{\theta}^P$ is the noise vector, because $\hat{\theta}^R$ is computed deterministically on the data. Thus the probability that Algorithm 2 outputs a particular $\hat{\theta}^P$ is equal to the probability that the noise vector is exactly the difference between $\hat{\theta}^P$ and $\hat{\theta}^R$. Fixing an arbitrary $\hat{\theta}^P$, let $\hat{v} = \hat{\theta}^P - \hat{\theta}^R$ and $\hat{v}' = \hat{\theta}^P - (\hat{\theta}^R)'$. Then,

$$\frac{h(\hat{\theta}^P | (X, y))}{h(\hat{\theta}^P | (X', y'))} = \frac{\nu(\hat{v})}{\nu(\hat{v}')} = \exp\left(\frac{-\gamma\epsilon}{8B+4M}(\|\hat{v}\|_2 - \|\hat{v}'\|_2)\right) = \exp\left(\frac{\gamma\epsilon}{8B+4M}(\|\hat{v}'\|_2 - \|\hat{v}\|_2)\right) \tag{6}$$

By definition, $\hat{\theta}^P = \hat{\theta}^R + \hat{v} = (\hat{\theta}^R)' + \hat{v}'$. Rearranging terms gives $\hat{\theta}^R - (\hat{\theta}^R)' = \hat{v}' - \hat{v}$. By Lemma 6 and the triangle inequality,

$$\|\hat{v}'\|_2 - \|\hat{v}\|_2 \leq \|\hat{v}' - \hat{v}\|_2 = \left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{1}{\gamma}(4B+2M)$$

Plugging this into Equation (6) gives the desired inequality,

$$\frac{h(\hat{\theta}^P|(X,y))}{h(\hat{\theta}^P|(X',y'))} \leq \exp\left(\frac{\gamma\epsilon}{4B+2M}\frac{1}{\gamma}(4B+2M)\right) = \exp(\epsilon).$$

Next, we show that the output $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1, \{\pi_i\}_i)$ of the mechanism satisfies joint differential privacy using the Billboard model. The estimators $\hat{\theta}^P_0$ and $\hat{\theta}^P_1$ are computed in the same way as $\hat{\theta}^P$, so $\hat{\theta}^P_0$ and $\hat{\theta}^P_1$ each satisfy ϵ-differential privacy. Since $\hat{\theta}^P_0$ and $\hat{\theta}^P_1$ are computed on disjoint subsets of the data, then by Theorem 4 of McSherry [2009], together they satisfy ϵ-differential privacy. The estimator a player should use to compute her payments depends only on the partition of players, which is independent of the data because it is chosen uniformly at random. Thus by the Composition Theorem in Dwork et al. [2006], the estimators $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1)$ together satisfy 2ϵ-differential privacy.

Each player's payment π_i is a function of only her private information — her report $(x_i, \hat{y}_i)$ and the estimator used to compute her payment — and the 2ϵ-differentially private vector of estimators $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1)$. Then by the Billboard Lemma 5, the output $(\hat{\theta}^P, \hat{\theta}^P_0, \hat{\theta}^P_1, \{\pi_i\}_i)$ of Algorithm 2 satisfies 2ϵ-joint differential privacy. □

B.2 Truthfulness

In order to show that $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes-Nash equilibrium, we require the following three lemmas. Lemma 7 bounds the expected number of players who will misreport under the strategy profile $\sigma_{\tau_{\alpha,\beta}}$. Lemma 8 bounds the norm of the expected difference of two estimators output by Algorithm 2 run on different datasets, as a function of the number of players whose data differs between the two datasets. Lemma 9 bounds the first two moments of the noise vector that is added to preserve privacy.

Lemma 7. *Under symmetric strategy profile $\sigma_{\tau_{\alpha,\beta}}$, each player expects that at most an α-fraction of other players will misreport, given Assumption 2.*

Proof. Let S_{-i} denote the set of players other than i who truthfully report under strategy $\sigma_{\tau_{\alpha,\beta}}$. From the perspective of player i, the cost coefficients of all other players are drawn independently from the posterior marginal distribution $\mathcal{C}|_{x_i,y_i}$. By the definition of $\tau_{\alpha,\beta}$, player i believes that each other player truthfully reports independently with probability at least $1-\alpha$. Thus $\mathbb{E}[|S_{-i}|\ |x_i, y_i] \geq (1-\alpha)(n-1)$. □

Lemma 8. *Let $\hat{\theta}^R$ and $(\hat{\theta}^R)'$ be the ridge regression estimators on two fixed databases that differ on the input of at most k players. Then*

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{k}{\gamma}(4B+2M)$$

Proof. Since the two databases differ on the reports of at most k players, we can define a sequence of databases $\mathcal{D}_0, \ldots, \mathcal{D}_k$, that each differ from the previous database in the input of at most one player, and $\mathcal{D}_0$ is the input that generated $\hat{\theta}^R$, and $\mathcal{D}_k$ is the input that generated $(\hat{\theta}^R)'$. Consider running Algorithm 2 on each database D_j in the sequence. For each D_j, let $\hat{\theta}^R_j$ be the ridge regression estimator computed on D_j. Note that $\hat{\theta}^R_0 = \hat{\theta}^R$ and $\hat{\theta}^R_k = (\hat{\theta}^R)'$.

$$\begin{aligned}
\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 &= \left\|\hat{\theta}^R_0 - \hat{\theta}^R_k\right\|_2 \\
&= \left\|\hat{\theta}^R_0 - \hat{\theta}^R_1 + \hat{\theta}^R_1 - \ldots - \hat{\theta}^R_{k-1} + \hat{\theta}^R_{k-1} - \hat{\theta}^R_k\right\|_2 \\
&\leq \left\|\hat{\theta}^R_0 - \hat{\theta}^R_1\right\|_2 + \left\|\hat{\theta}^R_1 - \hat{\theta}^R_2\right\|_2 + \ldots + \left\|\hat{\theta}^R_{k-1} - \hat{\theta}^R_k\right\|_2 \\
&\leq k \cdot \max_j \left\|\hat{\theta}^R_j - \hat{\theta}^R_{j+1}\right\|_2
\end{aligned}$$

For each j, $\hat{\theta}^R_j$ and $\hat{\theta}^R_{j+1}$ are the ridge regression estimators computed on databases that differ in the data of at most a single player. That means either the databases are the same, so $\hat{\theta}^R_j = \hat{\theta}^R_{j+1}$ and their normed difference is 0, or they differ in the report of exactly one player. In the latter case, Lemma 6 bounds $\|\hat{\theta}^R_j - \hat{\theta}^R_{j+1}\|_2$ above by $\frac{1}{\gamma}(4B+2M)$ for each j, including the j which maximizes the normed difference.

Combining this fact with the above inequalities gives,

$$\left\|\hat{\theta}^R - (\hat{\theta}^R)'\right\|_2 \leq \frac{k}{\gamma}(4B+2M).$$

□

Lemma 9. $\mathbb{E}[v] = \vec{0}$ *and* $\mathbb{E}[\|v\|_2^2] = 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2$ *and* $\mathbb{E}[\|v\|_2] = \frac{4B+2M}{\gamma\epsilon}$

Proof. For every $\bar{v} \in \mathbb{R}^d$, there exists $-\bar{v} \in \mathbb{R}^d$ that is drawn with the same probability, because $\|\bar{v}\|_2 = \|-\bar{v}\|_2$. Thus,

$$\mathbb{E}[v] = \int_{\bar{v}} \bar{v}\ \Pr(v = \bar{v}) d\bar{v} = \frac{1}{2}\int_{\bar{v}} (\bar{v} + -\bar{v})\ \Pr(v = \bar{v}) d\bar{v} = \vec{0}.$$

The distribution of v is a high dimensional Laplacian with parameter $\frac{4B+2M}{\gamma\epsilon}$ and mean zero. It follows immediately that $\mathbb{E}[\|v\|_2^2] = 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2$ and $\mathbb{E}[\|v\|_2] = \frac{4B+2M}{\gamma\epsilon}$. □

We now prove that symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes-Nash equilibrium in Algorithm 2.

Theorem 3 (Truthfulness)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, and a desired confidence parameter β. Then under Assumptions 1 and 2, with probability $1 - d^{t^2}$ and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ is an η-approximate Bayes-Nash equilibrium in Algorithm 2, for*

$$\eta = b\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

Proof. Suppose all players other than i are following strategy $\sigma_{\tau_{\alpha,\beta}}$. Let player i be in group $1-j$, so she is paid according to the estimator computed on the data of group j. Let $\hat{\theta}^P_j$ be the estimator output by Algorithm 2 on the reported data of group j under this strategy, and let $(\hat{\theta}^R_j)'$ be the ridge regression estimator computed within Algorithm 2 when all players in group j follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}^R_j$ be the ridge regression estimator that would have been computed within Algorithm 2 if all players in group j had reported truthfully. For ease of notation, we will suppress the subscripts on the estimators for the remainder of the proof.

We will show that $\sigma_{\tau_{\alpha,\beta}}$ is an approximate Bayes-Nash equilibrium by bounding player i's incentive to deviate. We assume that $c_i \leq \tau_{\alpha,\beta}$ (otherwise there is nothing to show because player i would be allowed to submit an arbitrary report under $\sigma_{\tau_{\alpha,\beta}}$). We first compute the maximum amount that player i can increase her payment by misreporting to Algorithm 2. Consider the expected payment to player i from a fixed (deterministic) misreport, $\hat{y}_i = y_i + \delta$.

$$\begin{aligned}&\mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)|x_i,y_i] - \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i)|x_i,y_i]\\ &\quad = B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i) - B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i)\end{aligned}$$

The rule $B_{a,b}$ is a proper scoring rule, so it is uniquely maximized when its two arguments are equal. Thus any misreport of player i cannot yield payment greater than $B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)$, so the expression of interest is bounded above by the following.

$$\begin{aligned}
&B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i) - B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,y_i]^\top x_i)\\
&\quad = a - b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2 + (\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2\right)\\
&\qquad - a + b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)(\mathbb{E}[\theta|x_i,y_i]^\top x_i) + (\mathbb{E}[\theta|x_i,y_i]^\top x_i)^2\right)\\
&\quad = b\left((\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)^2 - 2(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i)(\mathbb{E}[\theta|x_i,y_i]^\top x_i) + (\mathbb{E}[\theta|x_i,y_i]^\top x_i)^2\right)\\
&\quad = b\left(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i - \mathbb{E}[\theta|x_i,y_i]^\top x_i\right)^2\\
&\quad = b\left(\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]^\top x_i\right)^2\\
&\quad \leq b(\|\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]\|_2^2 \|x_i\|_2^2)\\
&\quad \leq b\|\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]\|_2^2
\end{aligned}$$

We continue by bounding the term $\|\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]\|_2$.

$$\begin{aligned}
\|\mathbb{E}[\hat{\theta}^P - \theta|x_i,y_i]\|_2 &= \|\mathbb{E}[\hat{\theta}^P - \hat{\theta}^R + \hat{\theta}^R - \theta|x_i,y_i]\|_2\\
&= \|\mathbb{E}[(\hat{\theta}^R)' + v - \hat{\theta}^R + \hat{\theta}^R - \theta|x_i,y_i]\|_2\\
&= \|\mathbb{E}[v|x_i,y_i] + \mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i,y_i] + \mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i]\|_2\\
&\leq \|\mathbb{E}[v|x_i,y_i]\|_2 + \|\mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i,y_i]\|_2 + \|\mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i]\|_2
\end{aligned}$$

We again bound each term separately. In the first term, the noise vector is drawn independently of the data, so $\mathbb{E}[v|x_i,y_i] = \mathbb{E}[v]$, which equals $\vec{0}$ by Lemma 9. Thus $\|\mathbb{E}[v|x_i,y_i]\|_2 = 0$.

Jensen's inequality bounds the second term above by $\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2|x_i,y_i]$. The random variables $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ are the ridge regression estimators of two (random) databases that differ only on the data of players who misreported under threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. By Lemma 7, player i believes that at most αn players will misreport their $\hat{y}_j$,[2] so for all pairs of databases over which the expectation is taken, $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ differ in the input of at most αn players. By Lemma 8, their normed difference is bounded above by $\frac{\alpha n}{\gamma}(4B + 2M)$. Since this bound applied to every term over which the expectation is taken, it also bounds the expectation.

For the third term, $\mathbb{E}[\hat{\theta}^R - \theta|x_i,y_i] = \texttt{bias}(\hat{\theta}^R|x_i,y_i)$. Recall that $\hat{\theta}^R$ is actually $\hat{\theta}^R_j$, which is computed independently of player i's data, but is still correlated with (x_i,y_i) through the common parameter θ. However, conditioned on the true θ, the bias of $\hat{\theta}^R$ is independent of player i's data. That is, $\texttt{bias}(\hat{\theta}^R|x_i,y_i,\theta) = \texttt{bias}(\hat{\theta}^R|\theta)$. We now expand the third term using nested expectations.

$$\begin{aligned}
\mathbb{E}_{X,z,\theta}\left[\hat{\theta}^R - \theta|x_i,y_i\right] &= \mathbb{E}_\theta\left[\mathbb{E}_{X,z}[\hat{\theta}^R - \theta|x_i,y_i,\theta]\right]\\
&= \mathbb{E}_\theta\left[\texttt{bias}(\hat{\theta}^R|x_i,y_i,\theta)\right]\\
&= \mathbb{E}_\theta\left[\texttt{bias}(\hat{\theta}^R|\theta)\right]\\
&= \texttt{bias}(\hat{\theta}^R)\\
&= -\gamma(\gamma I + X^\top X)^{-1}\theta
\end{aligned}$$

[2] Lemma 7 promises that at most $\alpha(n-1)$ players will misreport. We use the weaker bound of αn for simplicity.

Then by Theorem 7, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the following holds with probability at least $1 - d^{-t^2}$.

$$\begin{aligned}\|\mathbb{E}[\hat{\theta}^R - \theta | x_i, y_i]\|_2 &= \| - \gamma(\gamma I + X^\top X)^{-1}\theta\|_2 \\ &\leq \gamma \|(\gamma I + X^\top X)^{-1}\|_2 \|\theta\|_2 \\ &\leq \gamma \left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n} \right) B \\ &= \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\end{aligned}$$

We will assume the above is true for the remainder of the proof, which will be the case except with probability at most d^{-t^2}. Thus with probability at least $1 - d^{-t^2}$, and when n is sufficiently large, the increase in payment from misreporting is bounded above by

$$b\|\mathbb{E}[\hat{\theta}^P - \theta | x_i, y_i]\|_2^2 \leq b \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} \right)^2.$$

In addition to an increased payment, a player may also experience decreased privacy costs from misreporting. By Assumption 1, this decrease in privacy costs is bounded above by $c_i \epsilon^2$. We have assumed $c_i \leq \tau_{\alpha,\beta}$ (otherwise player i is allowed to misreport arbitrarily under $\sigma_{\tau_{\alpha,\beta}}$, and there is nothing to show). Then the decrease in privacy costs for player i is bounded above by $\tau_{\alpha,\beta}\epsilon^2$.

Therefore player i's total incentive to deviate is bounded above by η, and the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$ forms an η-approximate Bayes Nash equilibrium for

$$\eta = b \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} \right)^2 + \tau_{\alpha,\beta}\epsilon^2.$$

□

B.3 Accuracy

In this section, we prove that the estimator $\hat{\theta}^P$ output by Algorithm 2 has high accuracy. We first require the following lemma, which uses the concentration inequalities of 7 to give high probability bounds on the distance from the ridge regression estimator to the true parameter θ.

Lemma 10. *Let $\hat{\theta}^R$ be the ridge regression estimator computed on a given database (X, y). Then with probability at least $1 - d^{-t^2}$, as long as $n \geq C(\frac{t}{\xi})^2(d+2)\log d$*

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] \leq \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} \right)^2 + \sigma^4 \left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2} \right)^2$$

and

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \leq \frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}.$$

Proof. Recall from Section A.2 that,

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] = \|\,\mathtt{bias}(\hat{\theta}^R)\|_2^2 + tr(\mathtt{Cov}(\hat{\theta}^R)),$$

and,

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] &= \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R] + \mathbb{E}[\hat{\theta}^R] - \theta\|_2] \\ &\leq \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] + \mathbb{E}[\|\mathbb{E}[\hat{\theta}^R] - \theta\|_2] \\ &= \mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] + \mathbb{E}[\|\,\mathtt{bias}(\hat{\theta}^R)\|_2]\end{aligned}$$

We now expand the remaining terms: $\|\,\texttt{bias}(\hat{\theta}^R)\|_2$ and $tr(\texttt{Cov}(\hat{\theta}^R))$ and $\mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2]$. For the remainder of the proof, we will assume the concentration inequalities in Theorem 7 hold, which will be the case, except with probability at most d^{-t^2}, as long as $n \geq C(\frac{t}{\xi})^2(d+2)\log d$.

$$\begin{aligned}
\|\,\texttt{bias}(\hat{\theta}^R)\|_2 &= \| -\gamma(\gamma I + X^\top X)^{-1}\theta\|_2 \\
&\leq \gamma\|\theta\|_2\|(\gamma I + X^\top X)^{-1}\|_2 \\
&\leq \gamma B\|(\gamma I + X^\top X)^{-1}\|_2 \\
&\leq \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

$$\begin{aligned}
tr(\texttt{Cov}(\hat{\theta}^R)) &= \|\,\texttt{Cov}(\hat{\theta}^R)\|_2^2 \\
&= \|\sigma^2(\gamma I + X^\top X)^{-1}X^\top X(\gamma I + X^\top X)^{-1}\|_2^2 \\
&\leq \sigma^4\|(\gamma I + X^\top X)^{-1}\|_2^2\|X^\top X\|_2^2\|(\gamma I + X^\top X)^{-1}\|_2^2 \\
&\leq \sigma^4\left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2\left((1+\xi)\frac{1}{d+2}n\right)^2\left(\frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 \\
&\leq \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{\left(\gamma + (1-\xi)\frac{1}{d+2}n\right)^2}\right)^2
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^R - \mathbb{E}[\hat{\theta}^R]\|_2] &= \mathbb{E}[\|\hat{\theta}^R - (\theta + \texttt{bias}(\hat{\theta}^R))\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top y - \theta + (\gamma I + X^\top X)^{-1}\gamma I\theta\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top(X\theta + z) - \theta + (\gamma I + X^\top X)^{-1}\gamma I\theta\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}(X^\top X + \gamma I)\theta - \theta + (\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&= \mathbb{E}[\|\theta - \theta + (\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&= \mathbb{E}[\|(\gamma I + X^\top X)^{-1}X^\top z\|_2] \\
&\leq \mathbb{E}[\|(\gamma I + X^\top X)^{-1}\|_2\|X^\top z\|_2] \\
&\leq \mathbb{E}[\|(\gamma I + X^\top X)^{-1}\|_2 Mn] \\
&\leq \frac{Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

Using these bounds, we see:

$$\mathbb{E}[\|\hat{\theta}^R - \theta\|_2^2] \leq \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 + \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2$$

and

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] &\leq \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + \frac{Mn}{\gamma + (1-\xi)\frac{1}{d+2}n} \\
&= \frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

□

We now prove the accuracy guarantee for the estimator $\hat{\theta}^P$ output by Algorithm 2.

Theorem 4 (Accuracy)**.** *Fix a participation goal $1-\alpha$, a privacy parameter ϵ, and a desired confidence parameter β. Then under the symmetric threshold strategy $\sigma_{\tau_{\alpha,\beta}}$, Algorithm 2 will output an estimator $\hat{\theta}^P$ such that with probability at least $1-\beta-d^{-t^2}$, and when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$,*

$$\mathbb{E}[\|\hat{\theta}^P-\theta\|_2^2] = O\left(\left(\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right)^2+\left(\frac{\gamma}{n}\right)^2+\left(\frac{1}{n}\right)^2+\frac{\alpha n}{\gamma}+\frac{1}{\gamma\epsilon}\right).$$

Proof. Let the data held by players be (X,y), and let $\hat{y}=y+\vec{\delta}$ be the reports of players under the threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. As in Theorem 3, let $\hat{\theta}^P$ be the estimator output by Algorithm 2 on the reported data under this strategy, and let $(\hat{\theta}^R)'$ be the ridge regression estimator computed Algorithm 2 when all players follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}^R$ be the ridge regression estimator that would have been computed within Algorithm 2 if all players had reported truthfully. Recall that v is the noise vector added in Algorithm 2.

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}^P-\theta\|_2^2] &= \mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R+\hat{\theta}^R-\theta\|_2^2]\\ &= \mathbb{E}\left[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2+\|\hat{\theta}^R-\theta\|_2^2+2\left\langle\hat{\theta}^P-\hat{\theta}^R,\hat{\theta}^R-\theta\right\rangle\right]\\ &\leq \mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|\hat{\theta}^R-\theta\|_2^2]+2\mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2\|\hat{\theta}^R-\theta\|_2]\end{aligned}$$

We start by bounding the first term. Recall that the estimator $\hat{\theta}^P$ is equal to the ridge regression estimator on the *reported* data, plus the noise vector v added by Algorithm 2.

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}^P-\hat{\theta}^R\|_2^2] &= \mathbb{E}[\|(\hat{\theta}^R)'+v-\hat{\theta}^R\|_2^2]\\ &= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|v\|_2^2]+2\mathbb{E}[\langle(\hat{\theta}^R)'-\hat{\theta}^R,v\rangle]\\ &= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+\mathbb{E}[\|v\|_2^2]+2\langle\mathbb{E}[(\hat{\theta}^R)'-\hat{\theta}^R],\mathbb{E}[v]\rangle\\ &= \mathbb{E}[\|(\hat{\theta}^R)'-\hat{\theta}^R\|_2^2]+2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2 \quad \text{(by Lemma 9)}\end{aligned}$$

The estimators $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ are the ridge regression estimators of two (random) databases that differ only on the data of players who misreported under threshold strategy $\sigma_{\tau_{\alpha,\beta}}$. The definition of $\tau_{\alpha,\beta}$ ensures us that with probability $1-\beta$, at most αn players will misreport their $\hat{y}_j$. For the remainder of the proof, we will assume that at most αn players misreported to the mechanism, which will be the case except with probability β.

Thus for all pairs of databases over which the expectation is taken, $(\hat{\theta}^R)'$ and $\hat{\theta}^R$ differ in the input of at most αn players, and by Lemma 8, their normed difference is bounded above by $\left(\frac{\alpha n}{\gamma}(4B+2M)\right)^2$. Since this bound applies to every term over which the expectation is taken, it also bounds the expectation.

Thus the first term satisfies the following bound:

$$\mathbb{E}[\|\hat{\theta}^P-\theta\|_2^2] \leq \left(\frac{\alpha n}{\gamma}(4B+2M)\right)^2+2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2.$$

By Lemma 10, with probability at least $1-d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, the second term is bounded above by

$$\mathbb{E}[\|\hat{\theta}^R-\theta\|_2^2] \leq \left(\frac{\gamma B}{\gamma+(1-\xi)\frac{1}{d+2}n}\right)^2+\sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma+(1-\xi)\frac{1}{d+2}n)^2}\right)^2.$$

We will also assume for the remainder of the proof that the above bound holds, which will be the case except with probability at most d^{-t^2}.

We now bound the third term.

$$\begin{aligned}
&2\mathbb{E}[\|\hat{\theta}^P - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] = 2\mathbb{E}[\|(\hat{\theta}^R)' + v - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] \\
&\quad \leq 2\mathbb{E}[\left(\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2 + \|v\|_2\right)\|\hat{\theta}^R - \theta\|_2] \\
&\quad = 2\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] + 2\mathbb{E}[\|v\|_2\|\hat{\theta}^R - \theta\|_2] \\
&\quad = 2\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] + 2\mathbb{E}[\|v\|_2]\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \text{ (by independence)} \\
&\quad = 2\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] + 2\left(\frac{4B+2M}{\gamma\epsilon}\right)\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \text{ (by Lemma 9)}
\end{aligned}$$

We have assumed at most αn players misreported (which will occur with probability at most $1-\beta$), so for all pairs of databases over which the expectation in the first term is taken, Lemma 8 bounds $\|(\hat{\theta}^R)' - \hat{\theta}^R\|$ above by $\frac{\alpha n}{\gamma}(4B+2M)$. Thus we continue bonding the third term:

$$\begin{aligned}
&2\mathbb{E}[\|(\hat{\theta}^R)' - \hat{\theta}^R\|_2\|\hat{\theta}^R - \theta\|_2] + 2\left(\frac{4B+2M}{\gamma\epsilon}\right)\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \\
&\qquad \leq 2\mathbb{E}[\left(\frac{\alpha n}{\gamma}(4B+2M)\right)\|\hat{\theta}^R - \theta\|_2] + 2\frac{4B+2M}{\gamma\epsilon}\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \text{ (by Lemma 8)} \\
&\qquad = 2\left(\frac{\alpha n}{\gamma}(4B+2M)\right)\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] + 2\frac{4B+2M}{\gamma\epsilon}\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \\
&\qquad = 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\mathbb{E}[\|\hat{\theta}^R - \theta\|_2] \\
&\qquad \leq 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n} \text{ (by Lemma 10)}
\end{aligned}$$

We can now plug these terms back in to get our final accuracy bound. Taking a union bound over the two failure probabilities, with probability at least $1-\beta-d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}^P - \theta\|_2^2] \leq &\left(\frac{\alpha n}{\gamma}(4B+2M)\right)^2 + 2\left(\frac{4B+2M}{\gamma\epsilon}\right)^2 + \left(\frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n}\right)^2 \\
&+ \sigma^4\left(\frac{(1+\xi)\frac{1}{d+2}n}{(\gamma + (1-\xi)\frac{1}{d+2}n)^2}\right)^2 + 2\left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{4B+2M}{\gamma\epsilon}\right)\frac{\gamma B + Mn}{\gamma + (1-\xi)\frac{1}{d+2}n}
\end{aligned}$$

□

B.4 Individual Rationality and Budget

In this section we first characterize the conditions needed for individual rationality, and then compute the total budget required from the analyst to run the Private Regression Mechanism in Algorithm 2. Note that if we don't require individual rationality, it is easy to achieve a small budget: we can scale down payments as in the non-private mechanism from Section 3. However, once players have privacy concerns, they will no longer accept an arbitrarily small positive payment; each player must be paid enough to compensate for her privacy loss. In order to incentivize players to participate in the mechanism, the analyst will have to ensure that players receive non-negative utility from participation.

The first theorem that Algorithm 2 is individually rational for players with privacy costs below threshold. Note that because we allow cost coefficients to be unbounded, it's not possible to ensure individual rationality for all players while maintaining a finite budget.

Theorem 5 (Individual Rationality)**.** *Under Assumption 1, the mechanism in Algorithm 2 is individually rational for all players with cost coefficients $c_i \leq \tau_{\alpha,\beta}$ as long as,*

$$a \geq \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2,$$

regardless of the reports from players with cost coefficients above $\tau_{\alpha,\beta}$.

Proof. Let player i have privacy cost $c_i \leq \tau_{\alpha,\beta}$, and consider player i's utility from participating in the mechanism. Let player i be in group $1-j$, so she is paid according to the estimator computed on the data of group j. Let $\hat{\theta}_j^P$ be the estimator output by Algorithm 2 on the reported data of group j under this strategy, and let $(\hat{\theta}_j^R)'$ be the ridge regression estimator computed within Algorithm 2 when all players in group j follow strategy $\sigma_{\tau_{\alpha,\beta}}$. Let $\hat{\theta}_j^R$ be the ridge regression estimator that would have been computed within Algorithm 2 if all players in group j had reported truthfully. For ease of notation, we will suppress the subscripts on the estimators for the remainder of the proof.

$$\begin{aligned}
\mathbb{E}[u_i(x_i, y_i, \hat{y}_i)] &= \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i)|x_i, y_i] - \mathbb{E}[f_i(c_i, \epsilon)] \\
&\geq \mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i)|x_i, y_i] - \tau_{\alpha,\beta}\epsilon^2 \text{ (by Assump. 1)} \\
&= B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i, y_i]^\top x_i, \mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i) - \tau_{\alpha,\beta}\epsilon^2
\end{aligned}$$

We proceed by bounding the inputs to the payment rule, and thus lower-bounding the payment player i receives. The second input satisfies the following bound.

$$\mathbb{E}[\theta|x_i, \hat{y}_i]^\top x_i \leq \|\mathbb{E}[\theta|x_i, \hat{y}_i]\|_2 \|x_i\|_2 \leq B$$

We can also bound the first input to the payment rule as follows.

$$\begin{aligned}
\mathbb{E}[\hat{\theta}^P|x_i, y_i]^\top x_i &= \mathbb{E}[(\hat{\theta}^R)'|x_i, y_i]^\top x_i + \mathbb{E}[v|x_i, y_i]^\top x_i \\
&= \mathbb{E}[(\hat{\theta}^R)'|x_i, y_i]^\top x_i \\
&\leq \|\mathbb{E}[(\hat{\theta}^R)'|x_i, y_i]\|_2 \|x_i\|_2 \\
&\leq \|\mathbb{E}[(\hat{\theta}^R)' - \hat{\theta}^R|x_i, y_i]\|_2 + \|\mathbb{E}[\hat{\theta}^R - \theta|x_i, y_i]\|_2 + \|\mathbb{E}[\theta|x_i, y_i]\|_2 \\
&\leq \frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B \text{ (by Lemma 8 and Thm 7)}
\end{aligned}$$

Recall that the Brier payment rule is $B_{a,b}(p, q) = a - b\left(p - 2pq + q^2\right)$, which is bounded below by $a - b|p| - 2b|p|\,|q| - b|q|^2 = a - |p|(b + 2b|q|) - b|q|^2$. Using the bounds we just computed on the inputs to player i's payment rule, her payment is at least

$$\pi_i \geq a - \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) - bB^2.$$

Thus her expected utility from participating in the mechanism is at least

$$\mathbb{E}[u_i(x_i, y_i, \hat{y}_i)] \geq a - \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) - bB^2 - \tau_{\alpha,\beta}\epsilon^2.$$

Player i will be ensured non-negative utility as long as,

$$a \geq \left(\frac{\alpha n}{\gamma}(4B + 2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b + 2bB) + bB^2 + \tau_{\alpha,\beta}\epsilon^2.$$

□

The next theorem characterizes the total budget required by the analyst to run Algorithm 1.

Theorem 6 (Budget)**.** *The total budget required by the analyst to run Algorithm 2 under threshold equilibrium strategy $\sigma_{\tau_{\alpha,\beta}}$ is*

$$\mathcal{B} = n\left[a + \left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b+2bB)\right].$$

Proof. The total budget is the sum of payments to all players.

$$\begin{aligned}\mathcal{B} = \sum_{i=1}^{n}\mathbb{E}[\pi_i] &= \sum_{i=1}^{n}\mathbb{E}[B_{a,b}((\hat{\theta}^P)^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)|x_i,y_i] \\ &= \sum_{i=1}^{n} B_{a,b}(\mathbb{E}[\hat{\theta}^P|x_i,y_i]^\top x_i, \mathbb{E}[\theta|x_i,\hat{y}_i]^\top x_i)\end{aligned}$$

Recall that the Brier payment rule is $B_{a,b}(p,q) = a - b\left(p - 2pq + q^2\right)$, which is bounded above by $a + b|p| + 2b|p|\;|q| = a + |p|(b + 2b|q|)$. Using the bounds computed in the proof of Theorem 5, each player i receives payment at most,

$$\pi_i \geq a + \left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b+2bB).$$

Thus the total budget is at most:

$$\mathcal{B} = \sum_{i=1}^{n}\mathbb{E}[\pi_i] \leq n\left(a + \left(\frac{\alpha n}{\gamma}(4B+2M) + \frac{\gamma B}{\gamma + (1-\xi)\frac{1}{d+2}n} + B\right)(b+2bB)\right).$$

□

B.5 Bound on threshold $\tau_{\alpha,\beta}$

Lemma 3. *For a cost distribution $\mathcal{C}$ with conditional marginal CDF lower bounded by some function F: $\min_{x_i,y_i}\left(\Pr_{c_j\sim\mathcal{C}|x_i,y_i}[c_j \leq \tau]\right) \geq F(\tau)$, then*

$$\tau_{\alpha,\beta} \leq \max\{F^{-1}(1-\alpha\beta), F^{-1}(1-\alpha)\}.$$

Proof. We first bound $\tau^1_{\alpha,\beta}$.

$$\begin{aligned}\tau^1_{\alpha,\beta} &= \inf_{\tau}\left(\Pr_{c\sim\mathcal{C}}[|\{i : c_i \leq \tau\}| \geq (1-\alpha)n] \geq 1-\beta\right) \\ &= \inf_{\tau}\left(\Pr_{c\sim\mathcal{C}}[|\{i : c_i \geq \tau\}| \leq \alpha n] \geq 1-\beta\right) \\ &= \inf_{\tau}\left(1 - \Pr_{c\sim\mathcal{C}}[|\{i : c_i \geq \tau\}| \geq \alpha n] \geq 1-\beta\right) \\ &= \inf_{\tau}\left(\Pr_{c\sim\mathcal{C}}[|\{i : c_i \geq \tau\}| \geq \alpha n] \leq \beta\right)\end{aligned}$$

We continue by upper bounding the inner term of the expression.

$$\begin{aligned}\Pr_{c\sim\mathcal{C}}[|\{i : c_i \geq \tau\}| \geq \alpha n] &\leq \frac{\mathbb{E}[|\{i : c_i \geq \tau\}|}{\alpha n} \text{ (by Markov's inequality)} \\ &= \frac{n\;Pr[c_i \geq \tau]}{\alpha n} \text{ (by independence of costs)} \\ &= \frac{Pr[c_i \geq \tau]}{\alpha}\end{aligned}$$

From this bound, if $\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta$, then also $\Pr_{c \sim \mathcal{C}} \left[|\{i : c_i \geq \tau\}| \geq \alpha n \right] \leq \beta$. Thus,

$$\inf_{\tau} \left(\Pr_{c \sim \mathcal{C}} \left[|\{i : c_i \geq \tau\}| \geq \alpha n \right] \leq \beta \right) \leq \inf_{\tau} \left(\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta \right),$$

since the infimum in the first expression is taken over a superset of the feasible region of the latter expression. Then,

$$\begin{aligned}
\tau^1_{\alpha,\beta} &\leq \inf_{\tau} \left(\frac{Pr[c_i \geq \tau]}{\alpha} \leq \beta \right) \\
&= \inf_{\tau} \left(Pr[c_i \geq \tau] \leq \alpha\beta \right) \\
&= \inf_{\tau} \left(1 - Pr[c_i \leq \tau] \leq \alpha\beta \right) \\
&= \inf_{\tau} \left(C(\tau) \geq 1 - \alpha\beta \right) \\
&\leq \inf_{\tau} \left(F(\tau) \geq 1 - \alpha\beta \right) \\
&\quad \text{(since the extremal conditional marginal bounds the unconditioned marginal)} \\
&= \inf_{\tau} \left(\tau \geq F^{-1}(1 - \alpha\beta) \right) \\
&= F^{-1}(1 - \alpha\beta)
\end{aligned}$$

Thus under our assumptions, $\tau^1_{\alpha,\beta} \leq F^{-1}(1 - \alpha\beta)$.

We now bound τ^2_{α}.

$$\begin{aligned}
\tau^2_{\alpha} &= \inf_{\tau} \left(\min_{x_i, y_i} \left(Pr_{c_j \sim \mathcal{C} | x_i, y_i} [c_j \leq \tau] \right) \geq 1 - \alpha \right) \\
&\leq \inf_{\tau} \left(F(\tau) \geq 1 - \alpha \right) \\
&= \inf_{\tau} \left(\tau \geq F^{-1}(1 - \alpha) \right) \\
&= F^{-1}(1 - \alpha)
\end{aligned}$$

Finally,

$$\tau_{\alpha,\beta} = \max\{\tau^1_{\alpha,\beta}, \tau^2_{\alpha}\} \leq \max\{F^{-1}(1 - \alpha\beta), F^{-1}(1 - \alpha)\}.$$

□

C Proof of Theorem 7

Theorem 7. *Let $\xi \in (0, 1)$, and $t \geq 1$. Let $\|\cdot\|$ denote the spectral norm. If $\{x_i\}_{i \in [n]}$ are i.i.d. and sampled uniformly from the unit ball, then with probability at least $1 - d^{-t^2}$, when $n \geq C(\frac{t}{\xi})^2 (d + 2) \log d$, for some absolute constant C, then,*

$$\left\| X^\top X \right\| \leq (1 + \xi) \frac{1}{d + 2} n, \text{ and } \left\| (X^\top X)^{-1} \right\| \leq \frac{1}{(1 - \xi) \frac{1}{d+2} n}, \text{ and}$$

$$\left\| \gamma I + X^\top X \right\| \leq \gamma + (1 + \xi) \frac{1}{d + 2} n, \text{ and } \left\| (\gamma I + X^\top X)^{-1} \right\| \leq \frac{1}{\gamma + (1 - \xi) \frac{1}{d+2} n}.$$

Proof. We will first require Lemma 11, which characterizes the covariance matrix of the distribution on x.

Lemma 11. *The covariance matrix of x is $\Sigma = \frac{1}{d+2} I$.*

Proof. Let $z_1, \ldots, z_d \sim N(0,1)$, and let $u \sim U[0,1]$, all drawn independently. Define, $r = \sqrt{z_1^2 + \cdots + z_d^2}$ and $Z = (u^{1/d}\frac{z_1}{r}, \ldots, u^{1/d}\frac{z_d}{r})$. Then Z describes a uniform distribution over the d-dimensional unit ball Knuth [1981]. Recall that this is the same distribution from which the x_i are drawn. By the symmetry of the uniform distribution, $\mathbb{E}[Z] = \vec{0}$, and $Cov(Z)$ must be some scalar times the Identity matrix. Then to compute the covariance matrix of Z, it will suffice to compute the variance of some coordinate Z_i of Z. Since each coordinate of Z has mean 0, then $Var(Z_i) = \mathbb{E}[Z_i^2] + \mathbb{E}[Z_i]^2 = \mathbb{E}[Z_i^2]$.

$$\begin{aligned}
\sum_{i=1}^{d} \mathbb{E}[Z_i^2] &= \mathbb{E}\left[\sum_{i=1}^{d} Z_i^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} \left(u^{1/d}\frac{z_i}{r}\right)^2\right] \\
&= \mathbb{E}[u^{2/d}]\mathbb{E}\left[(\frac{1}{r})^2 \sum_{i=1}^{d} z_i^2\right] \\
&= \mathbb{E}[u^{2/d}] \\
&= \frac{d}{d+2}
\end{aligned}$$

By symmetry of coordinates, $\mathbb{E}[Z_i^2] = \mathbb{E}[Z_j^2]$ for all i, j. Then $\mathbb{E}[Z_i^2] = \frac{1}{d+2}$, and the covariance matrix of Z (and of x since both variables have the same distribution) is $\Sigma = \frac{1}{d+2}I$. □

From Corollary 5.52 in Vershynin [2012] and the calculation of covariance in Lemma 11, for any $\xi \in (0,1)$ and $t \geq 1$, with probability at least $1 - d^{-t^2}$,

$$\left\|\frac{1}{n}X^\top X - \frac{1}{d+2}I\right\| \leq \xi\frac{1}{d+2}, \tag{7}$$

when $n \geq C(\frac{t}{\xi})^2(d+2)\log d$, for some absolute constant C. We assume for the remainder of the proof that inequality (7) holds, which is the case except with probability at most d^{-t^2}, as long as n is sufficiently large. Then

$$\left\|X^\top X - \frac{1}{d+2}nI\right\| \leq \xi\frac{1}{d+2}n.$$

Let $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote respectively the maximum and minimum eigenvalues of a matrix A. By definition, $\lambda_{\max}(A) = \|A\|$.

Assume towards a contradiction that $\lambda_{\max}(X^\top X) = (1+\xi)\frac{1}{d+2}n + \delta$ for $\delta > 0$.

$$\begin{aligned}
\xi\frac{1}{d+2}n &\geq \left\|X^\top X - \frac{1}{d+2}nI\right\| \\
&= \|X^\top X\| - \frac{1}{d+2}n \\
&= \lambda_{\max}(X^\top X) - \frac{1}{d+2}n \\
&= (1+\xi)\frac{1}{d+2}n + \delta - \frac{1}{d+2}n \\
&= \xi\frac{1}{d+2}n + \delta
\end{aligned}$$

This implies $\delta \leq 0$, which is a contradiction. Thus $\lambda_{\max}(X^\top X) = \|X^\top X\| \leq (1+\xi)\frac{1}{d+2}n$.

Similarly, assume that $\lambda_{\min}(X^\top X) = (1-\xi)\frac{1}{d+2}n - \delta$ for some $\delta > 0$. Since all eigenvalues are positive, it must be the case that $\lambda_{\min}(X^\top X) \geq 0$.

$$\begin{aligned}
0 &\geq \lambda_{\min}(X^\top X - \frac{1}{d+2}nI) \\
&= \lambda_{\min}(X^\top X) - \frac{1}{d+2}n \\
&= (1-\xi)\frac{1}{d+2}n - \delta - \frac{1}{d+2}n \\
&= -\xi\frac{1}{d+2}n - \delta
\end{aligned}$$

This is also a contradiction, so $\lambda_{\min}(X^\top X) \geq (1-\xi)\frac{1}{d+2}n$. For any matrix A, $\lambda_{\max}(A^{-1}) = \frac{1}{\lambda_{\min}(A)}$. Thus,

$$\begin{aligned}
\lambda_{\min}(X^\top X) &= \frac{1}{\lambda_{\max}\left((X^\top X)^{-1}\right)} \\
&= \frac{1}{\|(X^\top X)^{-1}\|} \\
&\geq (1-\xi)\frac{1}{d+2}n \\
\implies \quad \|(X^\top X)^{-1}\| &\leq (1-\xi)\frac{1}{d+2}n
\end{aligned}$$

Using the fact that λ is an eigenvalue of a matrix A if and only if $(\lambda + c)$ is an eigenvalue of $(A + cI)$, we have the following inequalities to complete the proof:

$$\left\|\gamma I + X^\top X\right\| = \lambda_{\max}(\gamma I + X^\top X) \leq \gamma + (1+\xi)\frac{1}{d+2}n$$

$$\left\|(\gamma I + X^\top X)^{-1}\right\| = \frac{1}{\lambda_{\min}(\gamma I + X^\top X)} \leq \frac{1}{\gamma + (1-\xi)\frac{1}{d+2}n}$$

□

D Quadratically Bounded Privacy Penalty Costs

We will consider a particular functional form of $f_i(c_i, \epsilon)$, motivated by the model of privacy cost in the existing literature Chen et al. [2013]. In particular, we assume that the privacy cost function of each player i is upper-bounded by a function that depends on the effect of her input to a particular differentially private mechanism. This assumption leverages the functional relationship between player i's private data y_i, and the output of the mechanism. For example, if a particular mechanism ignores the input from player i, then her privacy cost should be 0 for participating in that computation, since her data is not used.

In order to formally state this assumption, we require the privacy cost function to take more inputs. Let $f_i(M, \hat{\theta}, (x_i, y_i), (x_{-i}, y_{-i}))$ denote the privacy cost to player i with observable attributes x_i for reporting y_i a mechanism $\mathcal{M}$ that takes in data vectors (x, y) and outputs an estimated parameter $\hat{\theta}$ when all other players have observable characteristics x_{-i} and report y_{-i}.

Assumption 4 (Chen et al. [2013], Privacy Cost Assumption[3]). *We assume that for any mechanism M that takes in data vectors (x, y) and outputs an estimated parameter $\hat{\theta}$, for all players i, estimates $\hat{\theta}$, and*

[3] The assumption proposed in Chen et al. [2013] allows privacy costs to be bounded by an arbitrary function of the log probability ratio that satisfies certain natural properties. We restrict to this particular functional form for simplicity, following Ghosh et al. [2014].

input data (x, y),

$$f_i(M, \hat{\theta}, (x_i, y_i), (x_{-i}, y_{-i})) \leq c_i ln \left(\max_{y_i', y_i''} \frac{Pr[M(x, y_i', y_{-i}) = \hat{\theta}]}{Pr[M(x, y_i'', y_{-i}) = \hat{\theta}]} \right).$$

Lemma 12 (Dwork et al. [2010], Chen et al. [2013], Composition Lemma)**.** *In settings that satisfy Assumption 4 and for mechanisms* M *that are* ϵ*-differentially private for* $\epsilon \leq 1$*, then for all players* i *with data* (x_i, y_i)*, for all data reports of other players* (x_{-i}, y_{-i})*, and for all possible misreports* y_i' *by player* i*,*

$$\mathbb{E}[f_i(M, M(x, y), (x_i, y_i), (x_{-i}, y_{-i}))] - \mathbb{E}[f_i(M, M(x, y_i', y_{-i}), (x_i, y_i), (x_{-i}, y_{-i}))] \leq 2c_i\epsilon(e^\epsilon - 1) \leq 4c_i\epsilon^2$$

Proof. (Sketch) The first inequality comes from Lemma 5.2 of Chen et al. [2013] and plugging in our specification of their "privacy-bound function" and replace statistical difference with the upper bound of $e^\epsilon - 1$. The second inequality comes from the bound $e^\epsilon \leq 1 + 2\epsilon$ for small ϵ. □

E Strong Convexity of Regularized Loss

Recall that we consider the loss function $\mathcal{L}(\theta, X, y)$ to be the sum of these individual loss functions plus a regularizing term:

$$\mathcal{L}(\theta; X, y) = \sum_{i=1}^{n} \ell(\theta; x_i, y_i) = \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2 + \gamma \|\theta\|_2^2,$$

where γ is a term that depends on n and will be defined later.

We now define strong convexity, which effectively states that the eigenvalues of the Hessian of a function are bounded away from zero, and prove that the loss function $\mathcal{L}$ is strongly convex.

Definition 10 (Strong Convexity)**.** *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is* m*-strongly convex if*

$$H(f(\chi)) - mI \text{ is positive semi-definite for all } \chi \in \mathbb{R}^d,$$

where $H(f(\chi))$ *is the Hessian*[4] *of* f*, and* I *is the* $d \times d$ *identity matrix.*

Notice that when f is a one-dimensional function ($d = 1$), strong convexity reduces to the requirement that $f''(\chi) \geq m > 0$ for all $\chi \in \mathbb{R}$. The following lemma proves that regularizing the quadratic loss $\mathcal{L}$ ensures it is strongly convex.

Lemma 13. $\mathcal{L}(\theta; X, y)$ *is* 2γ*-strongly convex in* θ*.*

Proof. We first compute the Hessian of $\mathcal{L}(\theta; X, y)$. For notational ease, we will suppress the dependence of $\mathcal{L}$ on X and y, and denote the loss function as $\mathcal{L}(\theta)$. We will use x_{ij} to denote the j-th coordinate of x_i, and θ_j to denote the j-th coordinate of θ.

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j} &= \sum_{i=1}^{n} \left[-2y_i x_{ij} + 2(\theta^\top x_i) x_{ij}\right] + 2\gamma\theta_j \\
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j \partial \theta_k} &= \sum_{i=1}^{n} \left[2(x_{ik}) x_{ij}\right] \text{ for } j \neq k \\
\frac{\partial \mathcal{L}(\theta)}{\partial \theta_j^2} &= \sum_{i=1}^{n} \left[2(x_{ij})^2\right] + 2\gamma
\end{aligned}$$

[4] The *Hessian* H of function f is a $d \times d$ matrix of its partial second derivatives, where

$$H(f(\chi))_{jk} = \frac{\partial^2 f(\chi)}{\partial \chi_j \partial \chi_k}.$$

A $d \times d$ matrix A is *positive semi-definite* (PSD) if for any $v \in \mathbb{R}^d$, $v^\top A v \geq 0$.

The Hessian of $\mathcal{L}$ is,

$$H(\mathcal{L}(\theta)) = \sum_{i=1}^{n} x_i x_i^\top + 2\gamma I,$$

where I is the identity matrix. Thus,

$$H(\mathcal{L}(\theta)) - 2\gamma I = \sum_{i=1}^{n} x_i x_i^\top,$$

which is positive semi-definite. To see this, let v be an arbitrary matrix in $\mathbb{R}^d$. Then for each i,$v(x_i x_i^\top)v^\top = (vx_i)^2 \geq 0$. The sum of PSD matrices is also PSD. Thus $\mathcal{L}(\theta)$ is 2γ-strongly convex. $\square$