TWO-PARTY PRIVACY GAMES:
HOW USERS PERTURB WHEN LEARNERS PREEMPT

JEFFREY PAWLIÇK AND QUANYAN ZHU

ABSTRACT. Internet tracking technologies and wearable electronics provide a vast amount of data to machine learning algorithms. This stock of data stands to increase with the developments of the internet of things and cyber-physical systems. Clearly, these technologies promise benefits. But they also raise the risk of sensitive information disclosure. To mitigate this risk, machine learning algorithms can add noise to outputs according to the formulations provided by differential privacy. At the same time, users can fight for privacy by injecting noise into the data that they report. In this paper, we conceptualize the interactions between privacy and accuracy and between user (input) perturbation and learner (output) perturbation in machine learning, using the frameworks of empirical risk minimization, differential privacy, and Stackelberg games. In particular, we solve for the Stackelberg equilibrium for the case of an averaging query. We find that, in equilibrium, either the users perturb their data before submission or the learner perturbs the machine learning output, but never both. Specifically, the learner perturbs if and only if the number of users is greater than a threshold which increases with the degree to which incentives are misaligned. Provoked by these conclusions - and by some observations from privacy ethics - we also suggest future directions. While other work in this area has studied privacy markets and mechanism design for truthful reporting of user information, we take a different viewpoint by considering both user and learner perturbation. We hope that this effort will open the door to future work in the area of differential privacy games.

Key words and phrases. Differential Privacy, Empirical Risk Minimization, Stackelberg Games, Behavioral Tracking, Machine Learning, Internet of Things, Cyber-Physical Systems, Privacy Economics.

This work is in part supported by an NSF IGERT grant through the Center for Interdisciplinary Studies in Security and Privacy (CRISSP) at New York University (NYU).

Author affiliations: Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, 2 MetroTech Center 10th Fl., Brooklyn, NY 11201. Email addresses: \{quanyan.zhu,jpawlick\}@nyu.edu.
TWO-PARTY PRIVACY GAMES: HOW USERS PERTURB WHEN LEARNERS PREEMPT

1. Introduction

In the modern digital ecosystem, users leave behind rich trails of behavioral information. On the internet, websites send user data to third-party trackers such as advertising agencies, social networking sites, and data analytic companies [22]. Persistent tracking technologies such as canvas fingerprinting and evercookies are able to near-uniquely identify users and “circumvent users’ deliberate attempts to start with a fresh profile” [3]. Even publicly available internet data can be used to infer sensitive user attributes. For example, the authors of [18] used Facebook “Likes” to used predict with 85% accuracy whether someone from the United States belonged to the republican party or the democratic party, and to predict with 95% accuracy whether a user was Caucasian American or African American.

Tracking is not limited, of course, to the internet. The internet of things (IoT) is a phenomenon that refers to the standardization and integration of communications between physical devices in a way that mimics the connection of computers on the internet. IoT devices such as smartwatches include accelerometers, heart rate sensors, and sleep trackers that measure and upload data about users’ physical and medical conditions [27]. IoT devices also include RFID tags that monitor business transactions and sensors that track resource usage [2]. A related phenomenon, cyber-physical systems (CPS) consist of integrated physical and informational components that interact with humans in novel ways [5]. CPS applications include medical devices, technologies to improve assisted living, and advanced automotive systems [19]. Data from these applications data can be used to improve product or service quality or to drive social change. Preceding Hurricane Frances in 2004, for example, WalMart stores adjusted inventory based on data analysis forecasts of an increase in strawberry Pop-Tarts sales - “like seven times their normal sales rate, ahead of a hurricane” [16]. In perhaps a more critical application of CPS, continuous glucose monitors can provide closed-loop blood glucose control for users with diabetes [1, 25]. The smart grid and renewable energy also stand to benefit from CPS [5].
1.1. Privacy in Machine Learning. While these technologies promise positive impacts, they also threaten privacy. Of course, privacy is a concept which takes some effort to define. (See different angles taken in [30], [31], [14], and [24].) But we are only concerned with one particular abstraction of privacy: a limit on access to information. The IoT and CPS involve new threats in the form of information access. Devices in these systems may directly collect sensitive information such as health and location data [2]. In addition, the pervasiveness of tracking and the development of analytics have enabled learners to infer habits and physical conditions over time. These inferences may run even to the granularity of “a user’s mood; stress levels; personality type; bipolar disorder; demographics” [26]. These are unprecedented degrees of access to user information. This access has prompted both qualitative and quantitative privacy research.

1.2. Differential Privacy for Private Release of Statistics. In the last two decades, scholars have developed a number of ways to quantify privacy. Well-tested framework frameworks include, e.g., $k$-anonymity [28] and $l$-diversity [21], while newer notions include empirical privacy [9] and membership privacy [20]. But our purpose in this paper is not to select one of these methods. It is rather to examine the larger trade-offs between user and learner perturbation and the resulting privacy and accuracy. Thus, we focus on one particular notion of privacy which has gained significant attention: differential privacy ([10, 12, 13]).

Proposed by Cynthia Dwork, differential privacy (DP) is a mathematical framework which gives probable limits on the disclosure risks that individuals incur by participating in a database. DP considers situations in which it is beneficial to release statistics of a database without revealing information about the individual users in the database. This framework would address, for instance, the task of releasing correlations between genes and diseases without giving away access to any one individual’s genetic record.

In general, statistics of a database do reveal something about individual data points - especially about those data points that are outliers, and especially for queries (such as maximum or minimum) that are
highly sensitive to worst-case changes in one data point. The output perturbation method of DP regulates the amount of information revealed by adding randomness to the output of a query \cite{13}. This randomness is proportional to the sensitivity of the query, and it takes into account the worst case data sets and individual data points. DP provides a measure of the privacy afforded by an algorithm in the form of a constant: $\epsilon_p$. Ideally, publishing this constant would inform users of the amount of information disclosure of the algorithm.

1.3. User Obfuscation Technologies. But DP seems to have gotten more attention from academia than industry, at least so far. Interestingly, scholars have recently developed various technologies that allow users to take privacy into their own hands. Finn and Nissenbaum describe two such examples: CacheCloak and TrackMeNot \cite{6}. CacheCloak provides a way for users to access location-based services without revealing their exact geographical positions \cite{23}. The app predicts multiple possibilities for the path of a user, and then retrieves location-based information for each path. This means that an adversary tracking the requests is left with many possible paths rather than a unique one. In a similar vein, TrackMeNot is a browser extension that generates decoy search queries in order to prevent trackers from assembling accurate profiles of its users \cite{17}. Both CacheCloak and TrackMeNot add randomness to a user’s behavioral data before it is made available to a learning algorithm.

1.4. Learner-User Interaction. Combining machine learning perturbation and user obfuscation, we have a two-party interaction. The learning algorithm can add randomness to the output of the computation, while the users can add randomization to the input. We depict this data flow in Fig. 1.1 (The notation is defined in Section 2). Both the users and the learner may be interested in the privacy and accuracy of the learning outcome. But these incentives are probably not aligned. Thus, the interaction is strategic - and aptly studied by game theory. We model the user-learner interaction as a two-step process in which the
learner first announces his output perturbation level, and then the users respond by implementing their input perturbation. This is a realistic assumption, since a critical aspect of DP is the ability to provide measurable privacy guarantees - quantities that can be announced. Knowing this protection, users can then decide whether to add their own perturbation in order to further protect their information. The dynamic, two-stage nature of this interaction suggests the framework of Stackelberg games [29, 4].

1.5. Content and Contributions. In this paper, we first describe the machine learning technique of empirical risk minimization (Section 2) and the framework of DP (Section 3). Then we employ these notions to quantify utility in a Stackelberg game in which a machine learning algorithm is the leader and a set of users (assumed to all perturb their data in the same manner) is the follower (Section 4). Empirical risk minimization and DP are suitable for a broad range of classification problems. In order to
get an intuition from a mathematically straight-forward problem, through, we apply these notions towards a simple query in which the learner releases the average of a set of scalar data points. It is possible to phrase this as an empirical risk minimization problem by formulating the risk for each data point as the squared difference between the point and the predicted average. For this problem, we find the equilibrium perturbations of the users and learner in Section 5. Then we discuss the importance of these results in Section 6. In this last installment, we also discuss variations on our model inspired by work in the ethics of privacy.

In summary, this paper presents the following contributions:

(1) We develop a game-theoretic model for machine learning interactions in which users and learners decide how much to perturb the input data and output statistic, respectively, based on the framework of DP.

(2) For the case of an averaging query, we obtain several analytical conclusions. We find that:

(a) the privacy-accuracy trade-off in the averaging query is demonstrated by a cross term between privacy and accuracy which completely divides out of the utility functions for the users and the learner.

(b) in equilibrium, either the users perturb their data before submission or the learner perturbs the machine learning output. It is never in equilibrium for both parties to perturb.

(c) the learner perturbs if and only if the number of users is greater than a threshold which increases with the degree to which incentives are misaligned.

(3) Finally, we consider the robustness of our results, and discuss multiple-user and incomplete information formulations of the game. We also base these ideas on ethical objections to obfuscation as a free rider problem and a characterization of the interaction as a public goods game.
1.6. **Related Work.** Our work benefits from significant achievements in research at the intersection of privacy and machine learning. The foundations of our work are in differential privacy in general \[10\] \[12\] \[13\], and differentially-private risk minimization - especially as studied by Chaudhuri and Monteleoni \[7\] - in particular. These frameworks, however, do not inherently take a strategic viewpoint; they do not require actions of users and learners to be incentive-compatible.

In order to address incentive-compatibility, a vein of research has arisen in *privacy markets*. Ghosh and Roth \[15\] contribute an important paper in this area. In their paper, a learner computes a sum of the private bits of a set of users and tries to either maximize accuracy or minimize cost. This paper assumes that users report their data truthfully but can misrepresent their individual valuation of their privacy. Later authors interchanged these assumptions; in \[32\], the type of each user is her private piece of information, and a mechanism is designed such that the user reports the type truthfully. This is closer to our work, since it studies perturbation of user data. In both \[15\] and \[32\], however, the game takes place between a set of users. Our formulation is different, because our game takes place between users and a learner.

Perhaps most similar to our work is a recent paper by Chessa *et al.* \[8\]. In this paper, users play a multiple person, prior-commitment game, which determines how much they perturb. Chessa *et al.* introduce a mechanism in which the learner defines a minimum level of precision with which users must report their data. If they do not comply with this minimum precision, the users are disallowed from participating. This paper also takes the users as the set of players\[1\]. It considers the adversary to be the learner. Our paper, on the contrary, considers a two-player game between a set of users and a learner, and considers the adversary to be a party that observes the output of the learning algorithm. In addition, our

\[1\]Because \[8\] considers a set of multiple users, it can also potentially help expand our work from a two-player game in which all of the users act together to an \(N + 1\) player game in which the users act individually. Here, though, we start with the two-player formulation.
work aims towards a framework for a broad range of queries by employing the tool set of empirical risk minimization.

2. Empirical Risk Minimization Model

Consider an interaction between a set of users $i \in S = \{1, \ldots, N\}$ and a learner $L$, in which users submit possibly-perturbed data to $L$, and $L$ releases a statistic or predictor of the data (hereafter, an output), possibly after adding its own perturbation. (We depict this data flow in Fig. I.I.) Assume that the data generating process is a random variable $Z$ with distribution $P$. The realized data is $z_i \overset{i.i.d.}{\sim} P, i \in S$.

Additionally, denote the value of the output by $\theta$, which belongs to some set $\Theta$. In the present section, we do not specify the domains of $Z$ and $\theta$, because we are interested in a framework applicable to multiple learning scenarios. In Section 5, we apply the framework specifically for a simple averaging query (in which $Z \in \mathbb{R}$). Here, we allow the variable definitions to be “overloaded” for the time being. Table I summarizes the notation for this section.

2.1. User Perturbation. Users may add noise to their actual data $z_i$ in order to increase their privacy. Here, we assume that all users perturb in the same fashion. Let $V$ be a random variable (of the same dimensions as $Z$) from which the perturbation $v_i, i \in S$ of each user is drawn. Following results in the DP literature [10, 12, 13], we consider $V$ in the form of a Laplace random variable with mean 0, variance $\sigma^2$, and appropriate dimension, denoted by $V = \text{LAP} (\sigma^2)$. We use $\tilde{z}_i = z_i + v_i$ to denote the perturbed value that each user reports to the learner, and $\tilde{Z} = P + V$ to denote the random variable from which these are drawn.

2.2. Output Computation. In empirical risk minimization, $L$ calculates a value of $\theta$ that minimizes the empirical risk, i.e., the total penalty due to imperfect classification of the realized data. Define the loss function $l(z_i, \theta)$, which expresses the penalty due to a single data point $z_i$ for the output $\theta$. $L$ obtains $\hat{\theta}$
Table 1. Notation employed in learning model

| Symbol | Meaning |
|--------|---------|
| $i \in S \in \{1, \ldots, N\}$ | User ($i$) and set of users ($S$) |
| $L$ | Learner |
| $Z$ | User data random variable |
| $z_i \overset{\text{i.i.d}}{\sim} P$ | Realized data ($z_i$) and distribution ($P$) |
| $\theta$ | Output of learning algorithm in general |
| $\theta = \theta^*$ | Minimizer of expected risk for unperturbed data |
| $V$ | User perturbation random variable |
| $v_i \overset{\text{i.i.d}}{\sim} \text{LAP} \left( \sigma^2_S \right)$ | Realized user perturbation with variance $\sigma^2_S$ |
| $\tilde{Z}$ | Perturbed data random variable |
| $\tilde{z}_i \overset{\text{i.i.d}}{\sim} P + \text{LAP} \left( \sigma^2_S \right)$ | Realized perturbed data ($\tilde{z}_i$) and distribution |
| $\theta$ | Output estimate from perturbed data |
| $W$ | Learner perturbation random variable |
| $w_i \overset{\text{i.i.d}}{\sim} \text{LAP} \left( \sigma^2_L \right)$ | Realized learner perturbation with variance $\sigma^2_L$ |
| $\hat{\theta}$ | Output reported after learner perturbation |
| $l(Z, \theta)$ | Loss function for data $Z$ and output $\theta$ |
| $\rho R(\theta)$ | Regularization term ($R(\theta)$) and weighting ($\rho$) |

given by Eq. 2.1, where $\rho \geq 0$ is a constant and $R(\theta)$ is a regularization term to prevent overfitting. (For example, $R(\theta)$ for the $L_2$-norm is $\frac{1}{2} \| \theta \|_2^2$.)

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \rho R(\theta) + \frac{1}{N} \sum_{i=1}^{N} l(\tilde{z}_i, \theta),
\]

2.3. Learner Perturbation. $L$, however, may add perturbation to $\hat{\theta}$ before publishing the output. Let $W$ (of appropriate dimension) denote the perturbation added by $L$. Assume again that this is in the form of a Laplace random variable with variance $\sigma^2_L$, i.e., $W = \text{LAP} \left( \sigma^2_L \right)$. The actual value of $\theta$ that is output by $L$, then, is $\bar{\theta} = \hat{\theta} + W$.

2.4. Accuracy Quantification. Following [7], we use the expected loss of an output to obtain a measure of accuracy. The expected loss is different than the empirical loss, because it is computed as an expectation
over randomness in the data (and, in our case, in the added perturbations), while the empirical loss is tallied from realized samples. The expected loss of $\tilde{\theta}$, the published output, on the actual data $Z$ is

$$E \left\{ \rho R (\tilde{\theta}) + l (Z, \tilde{\theta}) \right\},$$

where the expectation is taken over randomness in $Z, V,$ and $W$. Let $\theta^\dagger$ denote the $\theta$ which would minimize the expected loss for unperturbed data:

$$\theta^\dagger = \arg \min_{\theta \in \Theta} E \left\{ \rho R (\theta) + l (Z, \theta) \right\}.$$ 

Here, the expectation is taken over randomness in $Z$. Let $\epsilon_g$ be a positive scalar that denotes the difference in expected loss between the perturbed and unperturbed scenarios. This quantity is given by

$$\epsilon_g = E \left\{ \rho R (\tilde{\theta}) + l (Z, \tilde{\theta}) \right\} - E \left\{ \rho R (\theta^\dagger) + l (Z, \theta^\dagger) \right\}.$$ 

We use this difference to formulate the accuracy component of utility in Section 4.

3. DP Model

Cynthia Dwork, who developed the framework of DP ([10] [12] [13]), explains that it guarantees that participation in a database only affects the probability of disclosure by a "small multiplicative factor... as a consequence, there is a nominally increased risk to the individual in participating, and only nominal gain to be had by concealing or misrepresenting one’s data...a bad disclosure can still occur, but our guarantee assures the individual that it will not be the presence of her data that causes it, nor could the disclosure be avoided through any action or inaction on the part of the user” [12].
What does it mean specifically to guarantee that participation in a database only increases a user’s risk of disclosure by a small multiplicative factor? Let $A(*)$ denote an algorithm and $D$ denote a database. Let $D'$ denote a database that differs from $D$ by only one entry (e.g., the entry of the user under consideration). Let $c$ be some measurable set among all possible sets $C$ in which the output of the algorithm $A$ may fall. Finally, let $\mu(*)\mid B$ be the measure on $C$ induced by the output $A(B)$ given a data set $B$. Then we have the following definition.

**Definition 1.** ($\epsilon_p$-DP) - An algorithm $A(B)$ taking values in a set $C$ provides $\epsilon_p$-DP if

$$\left(3.1\right) \sup_{c \in C, D, D'} \frac{\mu(c \mid B = D)}{\mu(c \mid B = D')} \leq e^{\epsilon_p},$$

where that ratio is automatically 1 if $c$ is a set of measure 0 under the conditional measures induced by $D$ and $D'$.

This definition is from [7], based on the original definition in [10]. The intuition behind Eq. 3.1 is that the likelihood of an algorithm’s output being included in a set is not too much different between a database in which one user is participating and the same database in which a user is not participating. Essentially, this is achieved by making the output of the algorithm somewhat random.

The degree of randomness determines the privacy level $\epsilon_p$. (Lower values of $\epsilon_p$ correspond to more privacy.) That randomness is attained through the noise added in the forms of $V$ and $W$, with variances $\sigma_S^2$ and $\sigma_L^2$, respectively. In the next section, we express the relationship between $\epsilon_p$, $\sigma_S^2$, and $\sigma_L^2$ as part of the quantification of utility due to privacy.
Table 2. Notation employed in utility formulation for the Stackelberg game

| Symbol | Meaning |
|--------|---------|
| $U_L (\sigma^2_S, \sigma^2_L)$ | Total utility for $L$ |
| $U_S (\sigma^2_S, \sigma^2_L)$ | Total utility for users |
| $\epsilon_p = \xi (\sigma^2_S, \sigma^2_L)$ | DP from noise variances $\sigma^2_S, \sigma^2_L$ |
| $\epsilon_g = \eta (\sigma^2_S, \sigma^2_L)$ | Accuracy from noise variances $\sigma^2_S, \sigma^2_L$ |
| $\beta^p_L, \beta^p_L, \beta^p_L, \beta^p_L$ | $U_L$ coefficients for the cross term, accuracy, privacy, and perturbation |
| $\beta^p_S, \beta^p_S, \beta^p_S, \beta^p_S$ | $U_S$ coefficients for the cross term, accuracy, privacy, and perturbation |
| $BR_S (\sigma^2_L)$ | Best response function of the users to learner perturbation variance $\sigma^2_L$ |
| $\sigma^2_S \in BR_S (\sigma^2_L)$ | Equilibrium strategy for $L$ |
| $\sigma^2_S \in BR_S (\sigma^2_L)$ | Equilibrium strategy for the users |

4. Stackelberg Game Model

In this Section, we incorporate the model of empirical risk minimization from Section 2 and the model of DP from section 3 into a game-theoretic model for the user-learner interaction. First, we define the utility functions, and then we describe the game structure and the relevant equilibrium concept. Table 2 summarizes our notation.

4.1. Utility Functions. Accuracy, privacy and a cost of perturbation affect utility. Define a function $\eta : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\epsilon_g = \eta (\sigma^2_S, \sigma^2_L)$ gives the difference in expected loss between the perturbed and unperturbed scenarios according to Eq. 2.4. Next, define $\xi : \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\epsilon_p = \xi (\sigma^2_S, \sigma^2_L)$ gives the $\epsilon_p$ for which some fixed algorithm is $\epsilon_p$-differentially private. Finally, assume that the component of utility due to perturbation cost is linear in the variance of the perturbation. Allowing for a cross term between accuracy and privacy, the cost functions can be written as

\begin{align}
U_L (\sigma^2_S, \sigma^2_L) &= \beta^p_L \eta (\sigma^2_S, \sigma^2_L) \xi (\sigma^2_S, \sigma^2_L) + \beta^p_S \eta (\sigma^2_S, \sigma^2_L) + \beta^p_L \xi (\sigma^2_S, \sigma^2_L) + \beta^0 L \sigma^2_L, \\
U_S (\sigma^2_S, \sigma^2_L) &= \beta^p_S \eta (\sigma^2_S, \sigma^2_L) \xi (\sigma^2_S, \sigma^2_L) + \beta^p_S \eta (\sigma^2_S, \sigma^2_L) + \beta^p_S \xi (\sigma^2_S, \sigma^2_L) + \beta^0 S \sigma^2_S.
\end{align}
4.2. Game Structure. The data analysis scenario that we have described is a strategic and dynamic interaction. Stackelberg games model dynamic, two-step interactions in which a leader first declares an action, and a follower chooses an action afterwards. Then both parties receive utility.

We take $L$ to be the leader and the users to be the followers. The reason is that we assume the policies of a company which publishes statistics of a database to be made public, and then the users to afterwards decide whether to obfuscate their data themselves. The action of $L$ is to choose an output perturbation variance $\sigma_L^2$; the (common) action of the users is to choose an input perturbation variance - e.g., by installing an application such as CacheCloak [23] or TrackMeNot [17] - given by $\sigma_S^2$. The learner and users then receive payouts $\mathcal{U}_L(\sigma_S^2, \sigma_L^2)$, $\mathcal{U}_S(\sigma_S^2, \sigma_L^2)$, respectively.

4.3. Stackelberg Equilibrium. Let $(\sigma_S^2, \sigma_L^2)$ denote an equilibrium strategy pair. Stackelberg equilibria are solved backwards; the first requirement is that for any learner action $\sigma_L^2$, the users respond optimally
according to a best response function $BR_S : \mathbb{R}_+ \rightarrow 2^{\mathbb{R}_+}$. The output can possibly be a set, because more than one follower action could be optimal for a given leader action. $BR_S$ is defined by

\[
BR_S (\sigma^2_S) = \arg\max_{\sigma^2_S \geq 0} U_S (\sigma^2_S, \sigma^2_L).
\]

The learner then chooses the action $\sigma^2_L$ which - when the users respond according to $BR_S (\sigma^2_L)$ - maximizes $U_L (\sigma^2_L, BR_S (\sigma^2_L)) :$

\[
(4.4) \quad \sigma^2_L \in \arg\max_{\sigma^2_L \geq 0} U_L (\sigma^2_L, BR_S (\sigma^2_L)).
\]

Finally, in the execution of the game, the users play a strategy $\sigma^2_L$ that belongs to the output set of their best response function $BR_S (\sigma^2_L) :$

\[
(4.5) \quad \sigma^2_L \in BR_S (\sigma^2^*_L).
\]

From this equilibrium strategy profile, the users receive $U_S (\sigma^2^*_S, \sigma^2^*_L)$, and the learner receives $U_L (\sigma^2^*_S, \sigma^2^*_L)$.

5. Analysis of Averaging Query

In this section, we apply the general model outlined in Sections 2 to a specific, simple but common problem: the computation of an average of scalar data points. Here, the support of $Z$ is some subset of $\mathbb{R}$, and $\theta \in \mathbb{R}$. The added noises $V$ of the users and the added noise $W$ of the learners also have support in $\mathbb{R}$. The loss function is $l : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$. We consider empirical risk minimization without regulariation, so $\rho = 0$. 

The expected loss in accuracy $\eta(\sigma^2_s, \sigma^2_L)$ is defined by the difference between the expected loss for the perturbed classifier and the expected loss for an unperturbed classifier; i.e., $(\sigma^2_s, \sigma^2_L) = \mathbb{E}\{ l(Z, \hat{\theta}) \} - \mathbb{E}\{ l(Z, \theta^1) \}$. For the averaging query, we show in Appendix A that $\eta(\sigma^2_s, \sigma^2_L) = \sigma^2_L + \frac{1}{N} \sigma^2_S$. Besides the accuracy, we also to quantify the DP level. Leaving the derivation to Appendix B, we state here that the privacy level is given by $\epsilon_p = \xi(\sigma^2_s, \sigma^2_L) = \frac{2r}{\sigma^2_S + N \sigma^2_L}$.

Substituting $\eta(\sigma^2_s, \sigma^2_L)$ and $\xi(\sigma^2_s, \sigma^2_L)$ into the utility functions for the users and $L$, we obtain

\begin{equation}
U_L(\sigma^2_s, \sigma^2_L) = \beta^p_L \frac{1}{N} \left( N \sigma^2_L + \sigma^2_s \right) + \beta^g_L \frac{1}{N} \left( N \sigma^2_L + \sigma^2_s \right) + \beta^p_L \left( \frac{2r}{N \sigma^2_L + \sigma^2_s} \right) + \beta^0_L \sigma^2_L,
\end{equation}

\begin{equation}
U_S(\sigma^2_s, \sigma^2_L) = \beta^p_S \frac{1}{N} \left( N \sigma^2_L + \sigma^2_s \right) + \beta^g_S \frac{1}{N} \left( N \sigma^2_L + \sigma^2_s \right) + \beta^p_S \left( \frac{2r}{N \sigma^2_L + \sigma^2_s} \right) + \beta^0_S \sigma^2_s.
\end{equation}

In equilibrium, the users best respond to the strategy $\sigma^2_L$ of the learner. Their best response function is given by

\begin{equation}
BR_S(\sigma^2_L) = \arg\max_{\sigma^2_s \geq 0} U_S(\sigma^2_s, \sigma^2_L) = \arg\max_{\sigma^2_s \geq 0} \frac{2 \beta^p_S r}{N} + \beta^g_S \sigma^2_L + \frac{\beta^2_S \sigma^2_s}{N} + \frac{2 \beta^p_S r}{N \sigma^2_L + \sigma^2_s} + \beta^0_S \sigma^2_s
\end{equation}

\begin{equation}
= \arg\max_{\sigma^2_s \geq 0} \sigma^2_s \left( \beta^0_S + \frac{\beta^g_S}{N} \right) + \frac{2 \beta^p_S r}{N \sigma^2_L + \sigma^2_s} + \beta^0_S \sigma^2_L.
\end{equation}

This maximization problem is depicted in Fig. 5.1. The objective function in Eq. 5.4 is plotted for various values of $\sigma^2_L$, and the arguments $\sigma^2_S$ which maximize the function for each $\sigma^2_L$ are marked with solid rectangles.
Figure 5.1. Maximization problem for the users. They choose perturbation variance $\sigma^2_S$ on the horizontal axis which maximizes their utility $U_S(\sigma^2_S, \sigma^2_L)$, depending on $\sigma^2_L$. The filled rectangles represent these maximizing $\sigma^2_S$ for the selected $\sigma^2_L$.

Expressing these optimal $\sigma^2_S$ as a function of $\sigma^2_L$, the best response for the users is given by

\[
BR_S(\sigma^2_L) = \begin{cases} 
\sqrt{\frac{2N\beta^p_L r}{N\beta^p_S + \beta^0_S}} - N\sigma^2_L, & \text{if } \sigma^2_L < \frac{1}{N}\sqrt{\frac{2N\beta^p_L r}{N\beta^p_S + \beta^0_S}}, \\
0, & \text{if } \sigma^2_L \geq \frac{1}{N}\sqrt{\frac{2N\beta^p_L r}{N\beta^p_S + \beta^0_S}}, 
\end{cases}
\]

where the radicand is positive because $\beta^p_S$, $\beta^0_S$, and $\beta^0_S$ are negative. This best response function is depicted in Fig. 5.2 for various values of $N$. The best response functions of the users decrease linearly in $\sigma^2_L$ for $0 \leq \sigma^2_L < \frac{1}{N}\sqrt{\frac{2N\beta^p_L r}{N\beta^p_S + \beta^0_S}}$. For $\sigma^2_L \geq \frac{1}{N}\sqrt{\frac{2N\beta^p_L r}{N\beta^p_S + \beta^0_S}}$, the users best respond by not perturbing at all.

Based on this best response, the learner solves

\[
\sigma^2_L \in \arg\max_{\sigma^2_L \geq 0} U_L(\sigma^2_L, BR_S(\sigma^2_L)) = \arg\max_{\sigma^2_L \geq 0} \left[ \frac{2\beta^p_L r}{N} + \beta^0_L \sigma^2_L + \frac{\beta^0_S \sigma^2_S}{N} + \frac{2\beta^p_L r}{N\sigma^2_L + \sigma^2_S + \beta^0_L \sigma^2_L} \right]_{\sigma^2_S = BR_S(\sigma^2_L)}
\]
Figure 5.2. Best response function $BR_S(\sigma^2_L)$, plotted for a set of sample game parameters and $N = \{500, 1500, 2500, 3500, 4500\}$. As $N$ increases, the threshold above which $BR_S(\sigma^2_L) = 0$ decreases.

$$\sigma^2_L = \arg\max_{\sigma^2_L \geq 0} \left[ \sigma^2_L (\beta^0_L + \beta^g_L) + \frac{2\beta^p_L r}{N\sigma^2_L + \sigma^2_S} + \frac{\beta^g_L \sigma^2_S}{N} + \frac{2\beta^g p r}{N} \right]$$

Figure 5.3 depicts this maximization problem for $L$ : knowing the best response function $BR_S(\sigma^2_L)$, maximizes $U_L(\sigma^2_L, BR_S(\sigma^2_L))$. The filled squares in Fig. 5.3 indicate the maximum $\sigma^2_L^*$ for selected values of $N$. For low $N$, $\sigma^2_L^* = \max_{\sigma^2_L} U_L(\sigma^2_L, BR_S(\sigma^2_L))$ occurs at $\sigma^2_L = 0$, while for high $N$, the maximum occurs in the concave region on the right hand side of the figure (where the users do not perturb at all).

In Appendix C, we show specifically that

$$\sigma^2_L^* = \arg\max_{\sigma^2_L \geq 0} U_L(\sigma^2_L, BR_S(\sigma^2_L)) = \begin{cases} 0, & \text{if } N \leq \frac{\beta^g_S(\sigma^2_L + \sigma^2_S)}{\beta^p_L \beta^g_L}, \\ \sqrt{\frac{2\beta^p_L r}{N(\beta^p_L + \beta^g_L)}}, & \text{if } N > \frac{\beta^g_S(\sigma^2_L + \sigma^2_S)}{\beta^p_L \beta^g_L}. \end{cases}$$

Eq. 5.8 is depicted in Fig. 5.4.

In summary, Eq. 5.9 gives the optimal perturbation amounts for $L$ and for the users for different parameter regimes. Eq. 5.10 and Eq. 5.11 give the resulting equilibrium utilities. These equations demonstrate that there is a threshold for $N$ at which the equilibrium strategies flop from perturbation by the users and no perturbation by the learner to no perturbation by the users and perturbation by the
Figure 5.3. Depiction of the maximization problem for $L$ : knowing the best response function $BR_S(\sigma^2_L)$, he must maximize $U_L(\sigma^2_L, BR_S(\sigma^2_L))$. The utility $U_L(\sigma^2_L, BR_S(\sigma^2_L))$ decreases linearly for $0 \leq \sigma^2_L \leq \frac{1}{N} \sqrt{\frac{2N\beta^p_{Lr}}{N\beta^p_S+\beta^p_S}}$, and then becomes strictly concave for $\sigma^2_L > \frac{1}{N} \sqrt{\frac{2N\beta^p_{Lr}}{N\beta^p_S+\beta^p_S}}$. The filled squares indicate the maximum $\sigma^2_L^*$ for each value of $N$.

For the sample set of parameters, the critical $N$ around 1000 users. In the next section, we will discuss the importance of this and other characteristics of the results.

\[
\left(\sigma^2_L^*, \sigma^2_S^*\right) = \begin{cases} 
(0, \sqrt{\frac{2N\beta^p_{Lr}}{N\beta^p_S+\beta^p_S}}), & \text{if } N \leq \frac{\beta^p_S(\sigma^2_L + \beta^p_S)}{\beta^p_L + \beta^p_S} - \frac{\beta^p_S}{\beta^p_S} \\
\left(\sqrt{\frac{2\beta^p_{Lr}}{\left(N\beta^p_S+\beta^p_S\right)}}, 0\right), & \text{if } N > \frac{\beta^p_S(\sigma^2_L + \beta^p_S)}{\beta^p_L + \beta^p_S} - \frac{\beta^p_S}{\beta^p_S}
\end{cases}
\]
Figure 5.4. Optimal learner perturbation $\sigma_L^2$ versus number of users $N$ (given by Eq. 5.8). The optimal $\sigma_L^2$ is computed knowing the best response of the users, $BR_S(\sigma_L^2)$.

$$U_L(\sigma_L^2, \sigma_S^2) = \begin{cases} 2\beta_p^p \sqrt{\frac{N\beta_p^S + \beta_p^S}{2N\beta_p^S}} + \frac{\beta_p^S}{N} \sqrt{\frac{2N\beta_p^r}{N\beta_p^S + \beta_p^S}} + \frac{2\beta_p^p r}{N}, & \text{if } N \leq \frac{\beta_p^S \left( \beta_p^L + \beta_p^S \right)}{\beta_p^L \beta_p^S} - \frac{\beta_p^S}{\beta_p^S}, \\ \sqrt{\frac{8r\beta_p^p \left( \beta_p^L + \beta_p^S \right)}{N}} + \frac{2\beta_p^p r}{N}, & \text{if } N > \frac{\beta_p^S \left( \beta_p^L + \beta_p^S \right)}{\beta_p^L \beta_p^S} - \frac{\beta_p^S}{\beta_p^S}, \end{cases}$$

$$U_S(\sigma_L^2, \sigma_S^2) = \begin{cases} \sqrt{\frac{8r\beta_p^S \left( N\beta_p^S + \beta_p^S \right)}{N}} + \frac{2\beta_p^p r}{N}, & \text{if } N \leq \frac{\beta_p^S \left( \beta_p^L + \beta_p^S \right)}{\beta_p^L \beta_p^S} - \frac{\beta_p^S}{\beta_p^S}, \\ 2\beta_p^p r \sqrt{\frac{\beta_p^L + \beta_p^S}{2N\beta_p^S}} + \frac{\beta_p^L}{\beta_p^S} \sqrt{\frac{2\beta_p^p r}{N\left( \beta_p^L + \beta_p^S \right)}} + \frac{2\beta_p^p r}{N}, & \text{if } N \leq \frac{\beta_p^S \left( \beta_p^L + \beta_p^S \right)}{\beta_p^L \beta_p^S} - \frac{\beta_p^S}{\beta_p^S}, \end{cases}$$

6. Discussion

In this section, we discuss some principle conclusions from our model and analysis of the user-learner interaction. In order, we consider privacy-accuracy trade-offs, intuitive checks of the model results, and
more nuanced conclusions. Then, prompted by questions that arise from our analysis, and inspired by privacy ethics, we consider possible improvements to the model and areas in which more complex interactions can be studied.

6.1. Discussion of Averaging Query Results. First, when designing the utility functions, we explicitly included the cross terms $\beta^p_L \eta(\sigma_S^2, \sigma_L^2) \xi(\sigma_S^2, \sigma_L^2)$ and $\beta^p_S \eta(\sigma_S^2, \sigma_L^2) \xi(\sigma_S^2, \sigma_L^2)$ in order to consider the interaction between the privacy and accuracy of a learning algorithm. The basic mechanism of DP is to create uncertainty in the data underlying the output of an algorithm by adding randomness to the algorithm. Common wisdom suggests that this creates a trade-off between privacy and accuracy. Indeed, for the averaging query, the accuracy and privacy factors divide out perfectly:

$$\eta(\sigma_S^2, \sigma_L^2) \xi(\sigma_S^2, \sigma_L^2) = \frac{1}{N} \left( N\sigma_L^2 + \sigma_S^2 \right) \frac{2r}{N\sigma_L^2 + \sigma_S^2} = \frac{2r}{N}. $$

The variances of the perturbation by the users and by $L$ do not influence the magnitude of the cross term. Of course, for general empirical risk minimization queries, there is no reason to suppose that the factors cancel out completely.

Next, we note two sanity checks on the equilibrium utilities. As expected, we see that $U_L(\sigma_S^{2*}, \sigma_L^{2*})$ and $U_S(\sigma_S^{2*}, \sigma_L^{2*})$ are both increasing in $N$ and decreasing in $r$. As the number $N$ of data points increases, it becomes easier to attain a high privacy value, because the query output reveals progressively less about each individual data point. It also becomes easier to obtain a high accuracy value, as a greater number of random variables due to user input perturbation are added together and the net variance decreases. Both of these changes benefit the learner and the users - albeit unequally - so it is unsurprising if not a foregone conclusion that the equilibrium utilities improve for both.
Finally, we examine the regime change that occurs in the equilibrium solution. At low \(N\), the users carry out *all* of the perturbation while \(L\) reports his result exactly, while at high \(N\), the users report transparently while \(L\) carries out *all* of the perturbation. Rearranging the conditions on \(N\) in Eq. 5.9, 5.10, and 5.11, this regime change occurs at \(N_T\) such that\(^2\)

\[
N_T (-\beta^p_S) + (-\beta^g_S) = \frac{(-\beta^g_S)}{(-\beta^p_L)} \left[ (-\beta^o_L) + (-\beta^o_L) \right].
\]

Eq. 6.2 shows that \(N_T\) increases as the ratio \(\frac{-\beta^g_S}{-\beta^p_L}\) increases. This is the ratio between how much the users versus the learner care about privacy. \(N_T\) also increases if \((-\beta^g_S)\) increases and \((-\beta^o_L)\) decreases - *i.e.*, if \(L\) becomes more concerned with accuracy while the users become less concerned. It is likely that from the start the users are more concerned than learners with privacy, and the learner is more concerned than the users with the accuracy of the output. Thus, both of these movements could be considered increases in the misalignment of the system. As the misalignment increases, a higher number of users are necessary before it becomes incentive-compatible for a learning algorithm to perturb. Until then, the users carry the burden of perturbation. These results notwithstanding, we believe that much work remains to be done in improving and developing the DP games framework.

### 6.2. Robustness and Breadth of the Current Model

How can we improve the robustness of our model? First, consider what the model would predict in the absence of any learner perturbation. Considering only user perturbation, the accuracy term of \(U_S (\sigma^2_S, \sigma^2_L)\) is linear in \(\sigma^2_S\), and the privacy term, is proportional to \(\frac{1}{\sigma^2_S}\). Since these two costs are added, the privacy cost is dominant when \(\sigma^2_S \approx 0\). This is so because of the definition of DP. DP considers the output of an algorithm to be the measure or density function \(\mu (c)\) of the statistics or classifiers that it releases. For two databases \(D\) and \(D'\) that differ in at most

\(^2\)Since all of the \(\beta\) coefficients are \(< 0\), it becomes easier to examine their effect by looking at their absolute values, which in this case are their negations.
TWO-PARTY PRIVACY GAMES: HOW USERS PERTURB WHEN LEARNERS PREEMPT

One entry, DP requires $\mu(c \mid B = D)$ and $\mu(c \mid B = D')$ to be close at every possible $c$. For a deterministic output, the density function could change by an infinite factor with a change in the value of one user in the data set. Hence, deterministic outputs result in an infinite privacy penalty. This is unrealistic: plenty of users adopt services with no perturbation of statistics. Thus, greater robustness could be achieved through utility functions that limit the penalty for deterministic algorithms.

As a second point for discussion, note that the framework of empirical risk minimization that we described in Section 2 - Section 4 allows a broad range of queries, whereas we have only analytically solved the equilibrium for the averaging query. Future steps could include considering user perturbation in the calculation of the privacy and accuracy bounds in [7] in order to build these bounds into the utility functions of the Stackelberg game.

Finally, we are also inspired by recent work in the ethics of privacy that suggests the need to study more complex interactions. In [6], Brunton and Nissenbaum note several ethical objections that might be raised to user data obfuscation. The objections cited by these authors are rich in the language of economics, and hence prompt our considerations here. First, Brunton and Nissenbaum discuss the idea that user obfuscation is unethical because of a free rider problem. In one form of the argument, people who employ obfuscation when using a service which incorporates machine learning benefit from the service without providing it any useful data. In another form of the argument, users who obfuscate are actually free riding off of users who do not obfuscate. To address these questions quantitatively, we suggest adding multiple users - possibly with different privacy sensitivities to our model. An incomplete information game could model the situation in which the privacy sensitivities of the users are unknown to the learner. Second, the Brunton and Nissenbaum note that the term pollution is sometimes used when discussing obfuscation [6]. When users muddy the data that they submit, they are allegedly polluting a common resource. If this is true, then public goods games could provide a useful perspective on the user-learner interaction.
(This is also pointed out by [8].) In public goods games, players decide how much of a resource that they control to contribute to a shared pool. The utilities to all players are highest if everyone contributes, but it is typically not incentive-compatible for any individual player to contribute. The cost of anarchy, which describes the amount of decrease of social welfare due to competition, could become a relevant means of assessing grounds for legislation or other intervention that could be used to escape a detrimental equilibrium.

7. Conclusion

Privacy in machine learning will become increasingly important as the IoT and CPS continue to develop. Because sensitive behavioral data will be increasingly available, it will be crucial to be able to predict and improve interactions between users and machine learning algorithms that attempt to track and profile the users. On the one hand, the mechanism of output perturbation allows differentially-private mechanisms to provide provable bounds on information disclosure. On the other hand, users may actively protect their privacy by obfuscating their data before it is made available to a learner. In this paper, we have combined the tool sets of DP, empirical risk minimization, and Stackelberg games in order to study the user-learner interaction. After introducing each framework, we presented utility functions for a Stackelberg game in which DP quantifies the utility due to privacy, and expected loss quantifies the utility due to accuracy. Considering the learner as the Stackelberg game leader and the users as the follower, we derived the equilibrium strategy profile and utilities for the case of an averaging query. Our results indicate that, in equilibrium, either the users perturb their data before submission, or the learner perturbs the machine learning output, but that it is never in equilibrium for both parties to perturb. We showed that the learner perturbs if and only if the number of users exceeds a threshold which increases with the degree to which incentives are misaligned. After discussing our results, we considered the robustness of our model, and
proposed future work based on inspirations from the ethics of privacy. From both technical and ethical perspectives, privacy research poses challenging questions. Our work aims to provide a framework for a class of privacy-relevant interactions, and we look forward to future developments in DP games.

**APPENDIX A. DERIVATION OF ACCURACY OF AVERAGING QUERY**

The difference in expected loss between the perturbed average query and the unperturbed average query is given by

\[
E \left\{ l(Z, \hat{\theta}) \right\} - E \left\{ l(Z, \theta^\dagger) \right\} = E \left\{ (Z - \hat{\theta})^2 \right\} - E \left\{ (Z - \theta^\dagger)^2 \right\},
\]

where the expectation in the perturbed term is taken over \(Z \sim P, V \sim \text{LAP}(\sigma^2_S), W \sim \text{LAP}(\sigma^2_L)\), and the expectation in the unperturbed term is taken over \(Z \sim P\). Next, we solve for both terms of this expected loss difference.

**A.1. Expected Loss of Unperturbed Query.** The expected loss of the unperturbed query is

\[
E \left\{ (Z - \theta^\dagger)^2 \right\} = E \left\{ Z^2 \right\} - 2E \left\{ Z \theta^\dagger \right\} E + E \left\{ \theta^{\dagger 2} \right\}.
\]

Here, \(\theta^\dagger\), is given by

\[
\theta^\dagger = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^N (z_i - \theta)^2 = \frac{1}{N} \sum_{i=1}^N z_i = \bar{Z}_N,
\]

where we have used \(\bar{Z}_N\) to denote the sample mean of \(Z\) with \(N\) data points. Eq. A.2 becomes

\[
E \left\{ (Z - \theta)^2 \right\} = E \left\{ Z^2 \right\} - 2E \left\{ Z \bar{Z}_N \right\} E + E \left\{ \bar{Z}_N^2 \right\}.
\]
Since \( Z \perp \bar{Z}_N \), and \( \mathbb{E}(\bar{Z}_N) = \mathbb{E}(Z) \), this reduces to

\[
(A.5) \quad \mathbb{E} \left\{ (Z - \theta)^2 \right\} = \text{VAR} \left\{ Z \right\} \left( 1 + \frac{1}{N} \right).
\]

**A.2. Expected Loss of Perturbed Query.** The expected loss of the perturbed query is

\[
(A.6) \quad \mathbb{E} \left\{ \left( Z - \bar{\theta} \right)^2 \right\} = \mathbb{E} \left\{ Z^2 \right\} - 2\mathbb{E} \left\{ Z \bar{\theta} \right\} \mathbb{E} + \mathbb{E} \left\{ \bar{\theta}^2 \right\}.
\]

Here, \( \bar{\theta} \) is the sum of the noise added by \( L \) and the estimate \( \hat{\theta} \) that \( L \) computes from the perturbed data that it receives from the users. That estimate \( \hat{\theta} \) is

\[
(A.7) \quad \hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} (\tilde{z}_i - \theta)^2 = \frac{1}{N} \sum_{i=1}^{N} \tilde{z}_i = \frac{1}{N} \sum_{i=1}^{N} (z_i + v_i) = \bar{Z}_N + \bar{V}_N,
\]

where \( \bar{V}_N \) is the sample mean of the perturbations added by the \( N \) users. To this, \( L \) adds noise \( W \):

\[
\bar{\theta} = \hat{\theta} + W = \bar{Z}_N + \bar{V}_N + W.
\]

Eq. (A.6) becomes

\[
(A.8) \quad \mathbb{E} \left\{ \left( Z - \bar{\theta} \right)^2 \right\} = \mathbb{E} \left\{ Z^2 \right\} - 2\mathbb{E} \left\{ Z (\bar{Z}_N + \bar{V}_N + W) \right\} + \mathbb{E} \left\{ (\bar{Z}_N + \bar{V}_N + W)^2 \right\}.
\]

After some simplification, the three terms become

\[
(A.9) \quad \mathbb{E} \left\{ \left( Z - \bar{\theta} \right)^2 \right\} = \mathbb{E} \left\{ Z^2 \right\} - 2\mathbb{E}^2 \left\{ Z \right\} + \left[ \text{VAR} \left\{ W \right\} + \frac{1}{N} (\text{VAR} \left\{ Z \right\} + \text{VAR} \left\{ V \right\}) + \mathbb{E}^2 \left\{ Z \right\} \right]
\]

\[
(A.10) \quad = \text{VAR} \left\{ Z \right\} \left( 1 + \frac{1}{N} \right) + \text{VAR} \left\{ W \right\} + \frac{1}{N} \text{VAR} \left\{ V \right\} = \sigma_L^2 + \frac{1}{N} \sigma_S^2.
\]
A.3. Difference in Expected Loss between Perturbed and Unperturbed Queries. Finally, from Eq. \[A.5\] and Eq. \[A.10\]

\[
\mathbb{E}\left\{ l\left( Z, \hat{\theta} \right) \right\} - \mathbb{E}\left\{ l\left( Z, \theta^\dagger \right) \right\} = \text{VAR}\{W\} + \frac{1}{N} \text{VAR}\{V\}.
\]

**Appendix B. Derivation of Privacy of Averaging Query**

Let \( S(A) \) denote the sensitivity of an algorithm \( A \), defined in [11] as the maximum difference in the output of \( A(D) \) versus the output of \( A(D') \) for two databases \( D \) and \( D' \) that differ in at most one entry. In [11], Dwork and coauthors show that for a non-private algorithm \( A(D) \), DP of \( \epsilon_p \) can be achieved by releasing \( A(D) + \text{LAP}\left( \frac{S(A)}{\epsilon_p} \right) \).

The non-private output of the averaging query is \( \theta^\dagger \). Since \( \forall i \in S, |z_i| \leq r \), the sensitivity of the averaging query \( A_{avg} \) is \( S(A_{avg}) = \frac{2r}{N} \). Thus, it suffices for \( \epsilon_p \)-DP to release

\[
\tilde{\theta} = \theta^\dagger + \text{LAP}\left( \frac{2r}{N\epsilon_p} \right).
\]

Now the perturbed output \( \tilde{\theta} \) is given by

\[
\tilde{\theta} = \hat{\theta} + W = \bar{Z}_N + \bar{V}_N + W = \theta^\dagger + \text{LAP}\left( \frac{1}{N} \sigma_S^2 + \sigma_L^2 \right),
\]

which is the sum of the non-private output and a Laplace random variable with the given noise. Solving for \( \epsilon_p \),

\[
\epsilon_p = \frac{2r}{\sigma_S^2 + N\sigma_L^2}.
\]
Appendix C. Derivation of Optimal Learner Perturbation Given User Best Response

Since $BR_S(\sigma^2_L)$ is piece-wise linear, we solve Eq. 5.7 in two restricted domains: the first in which $L$ plays $0 \leq \sigma^2_L < \frac{1}{N} \sqrt{\frac{2N\beta^p_S}{N\beta^p_S + \beta^g_S}}$, and the second in which he plays $\sigma^2_L \geq \frac{1}{N} \sqrt{\frac{2N\beta^p_S}{N\beta^p_S + \beta^g_S}}$

In the first domain, $BR_S(\sigma^2_L) = \frac{1}{N} \sqrt{\frac{2N\beta^p_S}{N\beta^p_S + \beta^g_S}} - N\sigma^2_L$, and Eq. 5.7 becomes

$$\text{(C.1)} \quad \argmax_{0 \leq \sigma^2_L < \frac{1}{N} \sqrt{\frac{2N\beta^p_S}{N\beta^p_S + \beta^g_S}}} \left[ \sigma^2_L \beta^0_L + 2\beta^p_L r \left( \sqrt{\frac{N\beta^0_S + \beta^g_S}{2N\beta^p_S r}} + \frac{\beta^0_L}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^0_S + \beta^g_S}} + \frac{2\beta^p_S r}{N} \right) \right].$$

Since $\beta^0_L < 0$, in this domain $\sigma^2_L^* = 0$; it is best for $L$ not to perturb at all. Here, $L$ obtains utility

$$\text{(C.2)} \quad U_L(0, BR_S(0)) = 2\beta^p_L r \sqrt{\frac{N\beta^0_S + \beta^g_S}{2N\beta^p_S r}} + \frac{\beta^0_L}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^0_S + \beta^g_S}} + \frac{2\beta^p_S r}{N}. $$

In the second domain, $L$ perturbs enough such that $BR_S(\sigma^2_L) = 0$. In that case, Eq. 5.7 becomes

$$\text{(C.3)} \quad \argmax_{\sigma^2_L \geq \frac{1}{N} \sqrt{\frac{2N\beta^p_S}{N\beta^p_S + \beta^g_S}}} \left[ \sigma^2_L \beta^0_L + 2\beta^p_L r \left( N\sigma^2_L + \frac{2\beta^p_S r}{N} \right) \right] = \begin{cases} \frac{1}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^p_S + \beta^g_S}}, & \text{if } N \leq \frac{\beta^p_S(\beta^0_L + \beta^g_L)}{\beta^p_L \beta^g_S} - \frac{\beta^0_S}{\beta^g_S} \ \text{and} \ \frac{2\beta^p_S r}{N} \geq Sg \beta^g_S, \\ \frac{2\beta^p_S r}{N(\beta^0_L + \beta^g_L)}, & \text{if } N > \frac{\beta^p_S(\beta^0_L + \beta^g_L)}{\beta^p_L \beta^g_S} - \frac{\beta^0_S}{\beta^g_S} \ \text{and} \ \frac{2\beta^p_S r}{N} < Sg \beta^g_S. \end{cases}$$

According to Eq. C.3, if $N$ is small, then within this range, it is best for $L$ to simply play the minimum perturbation. In this case, $L$ obtains utility

$$\text{(C.4)} \quad U_L \left( \frac{1}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^p_S + \beta^g_S}}, BR_S \left( \frac{1}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^p_S + \beta^g_S}} \right) \right) = 2\beta^p_L r \sqrt{\frac{N\beta^0_S + \beta^g_S}{2N\beta^p_S r}} + \frac{\beta^0_L + \beta^g_L}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^0_S + \beta^g_S}} + \frac{2\beta^p_S r}{N}. $$

Clearly, from Eq. C.2, this is just $U_L(0, BR_S(0)) + \frac{\beta^0_L}{N} \sqrt{\frac{2N\beta^p_S r}{N\beta^p_S + \beta^g_S}}$, and since $\beta^0_L < 0$, $L$ receives less through this perturbation then if he were to play $\sigma^2_L = 0$ in the lower domain. This cannot be a global maximum.
Again, according to Eq. (C.3) if $N$ is large, then within this range, $L$ maximizes his utility on the interior of the range. In this case, he receives

\begin{equation}
U_L \left( \sqrt{\frac{2\beta^p L r}{N (\beta^q_L + \beta^{gp}_L)}}, BR_S \left( \sqrt{\frac{2\beta^p L r}{N (\beta^q_L + \beta^{gp}_L)}} \right) \right) = \sqrt{8r \beta^p L (\beta^0_L + \beta^{gp}_L) N} + \frac{2\beta^{gp} r}{N}.
\end{equation}

Is this larger than $U_L (0, BR_S (0))$? We can see that as $N$ becomes large, $U_L (0, BR_S (0))$ is asymptotic to a non-zero (negative) value $\beta^p L r \sqrt{\frac{2\beta^p L}{\beta^{gp} L}}$, while $U_L \left( \sqrt{\frac{2\beta^p L r}{N (\beta^q_L + \beta^{gp}_L)}}, BR_S \left( \sqrt{\frac{2\beta^p L r}{N (\beta^q_L + \beta^{gp}_L)}} \right) \right)$ approaches 0. Hence, at large $N$, $L$ obtains a larger utility by perturbing at $\sigma^2_L = \sqrt{\frac{2\beta^p L}{N (\beta^q_L + \beta^{gp}_L)}}$ than $\sigma^2_L = 0$. This is depicted graphically in the text in Fig. 5.3.

REFERENCES

[1] Continuous glucose monitoring for diabetes. WebMD, [Online]. Available: http://www.webmd.com/diabetes/guide/continuous-glucose-monitoring.

[2] Internet of Things: Privacy and Security in a Connected World. Technical report, Federal Trade Commission, January 2015.

[3] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. pages 674–689. ACM Press, 2014.

[4] Tamer Baçar and Geert Jan Olsder. The Stackelberg Equilibrium Solution. In Dynamic Noncooperative Game Theory, volume 23 of Classics in Applied Mathematics. Academic Press, New York, 1999.

[5] R. Baheti and H. Gill. Cyber-physical systems. The impact of control technology, 12:161–166, 2011.

[6] Finn Brunton and Helen Nissenbaum. Obfuscation: A User’s Guide for Privacy and Protest. MIT Press, 2015.

[7] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. The Journal of Machine Learning Research, 12:1069–1109, 2011.

[8] Michela Chessa, Jens Grossklags, and Patrick Loiseau. A short paper on the incentives to share private information for population estimates. In Financial Cryptography and Data Security, pages 427–436. Springer, 2015.
[9] Graham Cormode, Cecilia M Procopiuc, Entong Shen, Divesh Srivastava, and Ting Yu. Empirical privacy and empirical utility of anonymized data. In *Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on*, pages 77–82. IEEE, 2013.

[10] Cynthia Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.

[12] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):8, 2008.

[13] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[14] Ruth Gavison. Privacy and the Limits of Law. *The Yale Law Journal*, 89(3):421, January 1980.

[15] Arpita Ghosh and Aaron Roth. Selling privacy at auction. *Games and Economic Behavior*.

[16] Constance L. Hays. What Wal-Mart Knows About Customers’ Habits. *The New York Times*, November 2004.

[17] Daniel C. Howe and Helen Nissenbaum. TrackMeNot: Resisting surveillance in web search. *Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society*, 23:417–436, 2009.

[18] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, April 2013.

[19] E. A. Lee. Cyber physical systems: Design challenges. In *Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on*, pages 363–369. IEEE, 2008.

[20] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: a unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 889–900. ACM, 2013.

[21] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[22] Jonathan R. Mayer and John C. Mitchell. Third-party web tracking: Policy and technology. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 413–427. IEEE, 2012.
Joseph Meyerowitz and Romit Roy Choudhury. Hiding stars with fireworks: location privacy through camouflage. In Proceedings of the 15th annual international conference on Mobile computing and networking, pages 345–356. ACM, 2009.

Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press, 2009.

Robert S Parker, Francis J Doyle III, Nicholas Peppas, et al. A model-based algorithm for blood glucose control in type i diabetic patients. Biomedical Engineering, IEEE Transactions on, 46(2):148–157, 1999.

Scott R. Peppet. Regulating the Internet of Things: First Steps toward Managing Discrimination, Privacy, Security and Consent. Tex. L. Rev., 93:85, 2014.

Melanie Swan. Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0. Journal of Sensor and Actuator Networks, 1(3):217–253, November 2012.

Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

Heinrich Von Stackelberg. Marktform und gleichgewicht. J. Springer, 1934.

Samuel D. Warren and Louis D. Brandeis. The Right to Privacy. Harvard Law Review, 4(5):193, December 1890.

Alan Westin. Privacy and freedom. Atheneum, 1967.

David Xiao. Is privacy compatible with truthfulness? In Proceedings of the 4th conference on Innovations in Theoretical Computer Science, pages 67–86. ACM, 2013.